THE UNIVERSITY OF CHICAGO

MULTIPLE TESTING WITH PRIOR STRUCTURAL INFORMATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

ANG LI

CHICAGO, ILLINOIS
AUGUST 2017

Copyright © 2017 by Ang Li

All Rights Reserved

Dedicated to my beloved parents, Cheng Li and Liying Wang

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my advisor, Rina Foygel Barber, for her support and guidance during my time at the University of Chicago. Her vision and insights in research inspire me constantly in my study, which is invaluable to my experience throughout graduate school. I am also grateful for her clear and patient mentoring and teaching style. I would like to thank my committee members, John Lafferty and Matthew Stephens, for their helpful feedback and comments on my work, and their instructions on my graduate study. Thank you also to the faculty and staff of the Statistics Department, for the opportunity of excellent education here, and the friendly academic environment. Finally, I thank my parents Cheng Li and Liying Wang for their constant encouragement and love.

ABSTRACT

Multiple testing problems arise when we simultaneously test thousands or even millions of hypotheses. In many applications, the hypotheses have certain structures, based on prior studies or domain knowledge, which is a valuable source of information. We study how incorporating such information could improve the performance of multiple testing.

Specifically, we first consider the ordered testing problem in which the hypotheses are ranked from the one most likely to be signal to the least likely one. Given this ordered list of n hypotheses, the goal is to select a data-dependent cutoff k and declare the first k hypotheses to be statistically significant while bounding the false discovery rate (FDR). Generalizing existing methods, we develop a family of "accumulation tests" to choose a cutoff k that adapts to the amount of signal at the top of the ranked list. Our theoretical results prove that these methods control a modified FDR on finite samples, and characterize the power of the methods in the family. We apply the tests to simulated data, including a high-dimensional model selection problem for linear regression. We also compare accumulation tests to existing methods for multiple testing on a real data problem of identifying differential gene expression over a dosage gradient.

We then introduce the structure-adaptive Benjamini-Hochberg algorithm (SABHA). SABHA incorporates prior information about any pre-determined type of structure within the list of hypotheses, to reweight the p-values in a data-adaptive way. This raises the power by making more discoveries in regions where signals appear to be more common. Our main theoretical result proves that SABHA controls FDR at a level that is slightly higher than the target level, as long as the adaptive weights are not overfit to the data—interestingly, the excess FDR is related to the Rademacher complexity of the class from which we choose our data-adaptive weights. We apply this general framework to various structured settings, including ordered, grouped, low total variation structures, and structures from community models, and get the bounds on FDR for each setting. We also examine the empirical performance of SABHA on fMRI activity, gene/drug response data, and on simulated datasets.

CHAPTER 1

INTRODUCTION

In many modern applications of statistics, the availability of high-dimensional data sets allows simultaneous testing of a large number of potential hypotheses. Treating the many questions separately may lead to a large number of discoveries which are mostly spurious – if each hypothesis is tested with a significance threshold α, e.g. $\alpha = 0.05$, then in the scenario where the number of true signals among the n hypotheses is small, we can expect $\sim \alpha \cdot n$ false discoveries. This calls for methods that can test the hypotheses simultaneously, and the area of research on such methods is known as multiple hypotheses testing.

In multiple testing, the family-wise error rate (FWER) and the false discovery rate (FDR) are popular error measures. FWER is defined as the probability to make at least one false positive mistake,

$$\text{FWER} = \mathbb{P}\left\{\sum_i \mathbb{1}\{P_i \text{ rejected and } i \in \mathcal{H}_0\} \geq 1\right\}$$

and the FDR is defined as the expected false discovery proportion,

$$\text{FDR} = \mathbb{E}\left[\mathsf{FDP}\right] \text{ where } \mathsf{FDP} = \frac{\sum_i \mathbb{1}\{P_i \text{ rejected and } i \in \mathcal{H}_0\}}{1 \vee \sum_i \mathbb{1}\{P_i \text{ rejected}\}}.$$

Here $\mathcal{H}_0 \subseteq [n]$ is the (unknown) set of null hypotheses, i.e. the p-values P_i for $i \in \mathcal{H}_0$ correspond to testing hypotheses where no signal is present.

There is rich literature on the general multiple testing and FDR control. We begin by discussing three widely used methods: the Benjamini-Hochberg (BH) procedure Benjamini and Hochberg [1995], Storey [2002]'s modification of the BH procedure, and the empirical Bayes method (Efron et al. [2001], Efron and Tibshirani [2002]).

The BH procedure has been widely applied to examine large scale datasets, including microarray gene expression data (Reiner et al. [2003]), brain fMRI (Genovese et al. [2002],

Heller et al. [2006]), etc. Given a multiple testing problem consisting of p-values $P_1, \ldots, P_n$ corresponding to n hypotheses, the BH procedure proceeds by ordering the n hypotheses by ascending p-values $P_{(1)} \leq P_{(2)}, ..., \leq P_{(n)}$, and comparing $P_{(i)}$ to its Benjamini-Hochberg critical value, $\alpha \cdot \frac{i}{n}$, for $i = 1, ..., n$. The largest p-value that is lower than its BH critical value is then selected, and all hypotheses whose p-values falls under it are rejected.

The procedure can also be formulated as finding

$$\widehat{k} = \max \left\{ k \geq 1 : P_i \leq \alpha \cdot \frac{k}{n} \text{ for at least } k \text{ many p-values } P_i \right\},$$

or setting $\widehat{k} = 0$ if this set is empty, then rejecting any p-value P_i which satisfies $P_i \leq \alpha \cdot \frac{\widehat{k}}{n}$, for a total of $\widehat{k}$ many rejections. We can think of this as setting an adaptive rejection threshold, $\alpha \cdot \frac{\widehat{k}}{n}$, which lies between the naive threshold α and the Bonferroni adjusted threshold α/n.

In the setting where the null p-values are uniformly distributed, are mutually independent, and are independent of the non-null p-values, Benjamini and Hochberg [1995] prove that the BH procedure controls the FDR at the level

$$\text{FDR} = \alpha \cdot \frac{|\mathcal{H}_0|}{n} \leq \alpha.$$

To understand why BH procedure works intuitively, we sketch a coarse estimate for the false discovery proportion using this method. For some fixed k, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\left\{ i \text{ is null and } P_i \leq \alpha \frac{k}{n} \right\} \right] \leq \sum_{i=1}^{n} \alpha \frac{k}{n} = \alpha k,$$

where the inequality holds since $\mathbb{P}\{i \text{ is null}\} \leq 1$, and if P_i is a null p-value then it should be uniformly distributed. Therefore, when we reject k many p-values, we expect that there are at most αk many nulls among them, leading to a false discovery proportion that is at most α.

In the dependent case, Benjamini and Yekutieli [2001] showed that the BH method still

controls FDR if the joint distribution of statistics (or p-values) is PRDS on the set of true nulls. PRDS for I_0 stands for "positive regression dependence on a subset I_0", e.g. Gaussian variables with totally positive covariance matrix is PRDS. Its formal definition is that for any increasing set D (a set D is called increasing if $x \in D$ and $y \geq x$ imply that $y \in D$), and for each $i \in I_0$, $\mathbb{P}\{X \in D | X_i = x\}$ is nondecreasing in x. Benjamini and Yekutieli [2001] also proved that the BH procedure controls FDR for any joint distribution of test statistics, if it is conducted with level $\frac{\alpha}{\sum_{i=1}^{n} \frac{1}{i}}$ in place of the target FDR level α, where n is the total number of hypotheses.

Adapting to the proportion of nulls Looking at the above bound $\alpha \cdot \frac{|\mathcal{H}_0|}{n}$ on the FDR, we see that in situations where the number of signals is a substantial proportion of the total number of hypotheses, the Benjamini-Hochberg procedure will be overly conservative. To adapt to the proportion of signals, Storey [2002] proposes a two-stage method where first, the relative proportion of high and low p-values is used to estimate the proportion of nulls:

$$\widehat{\pi}_0 = \min\left\{1, \frac{\sum_i \mathbb{1}\{P_i > \tau\}}{n(1-\tau)}\right\}. \tag{1.1}$$

To understand this intuitively, if $\pi_0 = \frac{|\mathcal{H}_0|}{n}$ is the true proportion of nulls, and if the signals $i \notin \mathcal{H}_0$ are all strong with $\mathbb{P}\{P_i > \tau\} \approx 0$, then we should have approximately $\pi_0 n \cdot (1-\tau)$ many p-values which are $> \tau$.

Second, this estimated proportion is used to adjust the threshold in the Benjamini-Hochberg procedure: by running BH with $\frac{\alpha}{\widehat{\pi}_0}$, we now expect the FDR to be

$$\text{FDR} \approx \frac{\alpha}{\widehat{\pi}_0} \cdot \frac{|\mathcal{H}_0|}{n} \lessapprox \alpha,$$

where the first step holds by the FDR level of the BH procedure, while the second holds when $\widehat{\pi}_0$ is a good (over)estimate of $\pi_0 = \frac{|\mathcal{H}_0|}{n}$. Define $\widehat{\text{FDR}}_\tau(t) = \frac{\hat{\pi}_0(\tau) t n}{\sum_{i=1}^{n} \mathbb{1}\{P_i < t\} \vee 1}$ as the estimate of FDR when the p-value threshold is at t and $t_\alpha(\widehat{\text{FDR}}_\tau)$ as

$$t_\alpha(\widehat{\mathrm{FDR}}_\tau) = \sup\{0 \le t \le 1 : \widehat{\mathrm{FDR}}_\tau(t) \le \alpha\}$$

which is the largest t such that $\widehat{\mathrm{FDR}}_\tau(t) \le \alpha$. Define a more conservative $\widehat{\pi}_0$ as $\widehat{\pi}_0^*(\tau) = \frac{\sum_i \mathbb{1}\{P_i > \tau\}+1}{n(1-\tau)}$, and $\widehat{\mathrm{FDR}}_\tau^*(t)$ as the corresponding estimate of FDR at cutoff t, Storey et al. [2004, Theorem 3] proves that using $t_\alpha(\widehat{\mathrm{FDR}}_\tau^*)$ at the p-value threshold controls FDR:

$$\mathrm{FDR}(t_\alpha(\widehat{\mathrm{FDR}}_\tau^*)) \le \alpha.$$

The empirical Bayes method Efron et al. [2001] calculates the Bayesian FDR as the posterior probability of null conditioned on rejection, from the estimates of prior probabilities and densities of nulls and non-nulls. Efron et al. [2001] and Efron and Tibshirani [2002] also proposed local FDR, the probability of rejecting a null in a subset of the rejection region, and showed its application in the analysis of breast cancer microarray data. Mathematically, let p_1 be the probability that a hypothesis $\mathcal{H}$ is nonnull, and $p_0 = 1 - p_1$ be the probability that $\mathcal{H}$ is null, and $f_1(z), f_0(z)$ be the density of the observed data z from a nonnull and a null, respectively. Then

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

is the mixture density of the two populations, which can be estimated directly from data. By Bayes's Rule, the posteriori probability $p_0(z)$ that a hypothesis with data z is the null is

$$p_0(z) = p_0 f_0(z)/f(z).$$

This is the definition of local FDR. It is closely related to BH's FDR criterion, since for a collection of simultaneous hypothesis tests, FDR is the expected proportion of type I errors made.

Structured hypothesis testing In multiple testing, when hypotheses share a hierarchical structure, group dependence, or other types of structures, this information can be utilized to improve the power of testing. Here we describe several existing methods that adapt to particular types of structure in the signals and nulls.

First, Hu et al. [2012] consider a setting where the hypotheses are partitioned into disjoint blocks $\mathcal{B}_1, \ldots, \mathcal{B}_d \subseteq [n]$. The blocks may differ widely in terms of the proportion of nulls within the block, $\pi_0^{(k)} = \frac{|\mathcal{H}_0 \cap \mathcal{B}_k|}{|\mathcal{B}_k|}$. In their "Group Benjamini-Hochberg" procedure, after estimating these proportions with some $\widehat{\pi}_0^{(k)}$'s, the p-values are then reweighted to take these estimates into account before applying the BH procedure; this allows for higher power to make discoveries within those blocks where signals are prevalent. This can be viewed as an extension of the work of Storey [2002], which estimates the overall proportion of nulls π_0 without separating into blocks. Hu et al. [2012]'s main theoretical results show, firstly, that if the true $\pi_0^{(k)}$'s are known exactly, then FDR is controlled at the desired level α, and second, that if the $\pi_0^{(k)}$'s can instead be estimated consistently, then asymptotic FDR control is achieved.

While Hu et al. [2012]'s group BH procedure reweights the p-values in a data-adaptive way to account for group structure, an existing method by Genovese et al. [2006] handles *general* structures by reweighting the p-values in a *non-adaptive* way. Each p-value P_i is reweighted using prior information (i.e. the reweighting does not depend on the p-values themselves), then the BH procedure is applied to the reweighted p-values $\{P_i/w_i\}$. In this setting, the w_i's can be chosen to reflect any type of known structure, for instance, giving priority to certain hypotheses over others (like in the ordered testing setting) or to certain blocks of hypotheses, but the weights cannot be chosen as functions of the p-values themselves.

Second, multiple testing with spatial correlations or clustering has been extensively studied and arises in many applications, for instance, genomics, geophysical sciences, and astronomy. Chouldechova [2014] and Sun et al. [2015] develop a cluster-wise FDR, which treats discoveries at the cluster level rather than counting individual points. In application

on brain fMRI, Heller et al. [2006] grouped voxel units into clusters using previous correlation data, and applied the BH method at the cluster level. Their approach enjoys greater interpretability, as well as increased power.

Third, a hierarchical testing procedure is often applied in large multiple testing problems when the hypotheses are divided into families. In this procedure, families with evidence for true discoveries are first selected, then hypotheses within them are tested. It aims to control the expected average error-rate (e.g FDP) over the selected families, instead of controlling the expected error-rate globally for the combined set of discoveries. Benjamini and Bogomolov [2011] suggest this to be a more appropriate goal in many applications, and show that it can be achieved by controlling the expected error-rate in each selected family at a more stringent level: $\alpha \cdot$ the proportion of selected families over all families. The work of Benjamini and Yekutieli [2003] also falls into this category.

Besides hierarchical and grouped testing, there are several approaches in the literature to incorporate prior information into multiple testing. One approach, mostly applied in microarray analysis, uses prior information to exclude non-informative genes before the final selection step of significant ones, which enhances the power (Bourgon et al. [2010], McClintick and Edenberg [2006]). Also worth noting is the Bayesian mixture model approach to include previous knowledge in genome-wide linkage studies and association studies (Fridley et al. [2010]). Recently, Du and Zhang [2014] introduced a single-index modulated (SIM) procedure, which assumes the availability of a bivariate p-value (p_1, p_2) (where p_1 is the p-value from prior information, and p_2 is the main p-value reflecting current information), and project it into a single p-value combining p_1 and p_2 in some optimal direction for the final analysis. Scott and Berger [2006] explored a Bayesian hierarchical approach for multiple testing. The posterior probabilities, including the probability that hypothesis i is null given the data, are inferred through importance sampling. They discussed the choice of prior distributions on model parameters. This approach has been applied in disease mapping Catelan et al. [2010], and abnormal corporate performance identification Scott [2009].

Summary of proposed methods The existing literatures suggest that prior information on the structure of hypotheses is of great value if we can incorporate them into the testing procedure. Here we consider four types of structures that show up widely in applications and study how they can be utilized:

1. Ordered structure. This structure arises when we have a prior belief that certain hypotheses are more likely to contain signals than others, due to data from prior experiments, observations, or due to the way that these hypotheses are generated. For instance, in model selection procedures such as Lasso and forward stepwise selection, the model grows by adding predictors in sequence, until the full set of predictors is included. If we consider each predictor as a hypothesis, then the hypotheses added early into the list are more likely to be true signals than those added later. To leverage this ordered structure, we have developed two approaches:

   (a) rank the hypotheses from the one we believe to be most likely signal to the one believed to be least likely, and reject some initial stretch of the ranked list, based on experimental or observational data. In chapter 2, we formally define a family of methods for this problem, the "accumulation test", generalizing existing methods in the literature. We study its FDR control properties and statistical power performance, in theory and on a real problem from a drug dosage gene response experiment.

   (b) rank the hypotheses as before, but instead of restricting the rejection set to be an initial stretch, this time we only give priority to the earlier hypotheses to the extent that they have strictly lower prior probabilities to be nulls than the latter hypotheses. Such prior probabilities affect the testing results by making hypotheses with low priors easier to be rejected, and they can be estimated using the data and the structural constraint. In chapter 3, we formulate this approach mathematically, and provide a general framework for incorporating structural

information into multiple testing, called "structure-adaptive Benjamini-Hochberg procedure (SABHA)". The structures covered by SABHA are not limited to ordered structure type, as discussed below.

2. Block structure. In some applications, the hypotheses may come with some natural grouping, with true signals tending to co-occur in the same group. For instance, in genome-phenotype association studies, if each hypothesis corresponds to a gene, then the known gene pathways may form such a grouping. In this case, we might believe that each block has its own proportion of nulls and non-nulls, which can be estimated from data. This can be considered as an extension of Hu et al. [2012]'s group Benjamini-Hochberg approach, and can be formulated within the SABHA framework. The approach is demonstrated in the analysis of a fMRI activity data set in chapter 3.

3. Low total variation. In this setting, the hypotheses show local similarity. For instance, if each hypothesis corresponds to a test performed at some spatial location, then nearby hypotheses may have equal or similar probabilities of being null or non-null, and as a result the true signals are likely to appear in spatial clusters. This structure can be characterized mathematically as an upper bound on the total variation, and therefore can be incorporated into the SABHA framework straightforwardly. We study this setting in theory and simulations in chapter 3.

4. Structures from community models. In some situations, we are interested in the association between members in a community. If we formulate the community as a graph, and the members in it as individual nodes, then the hypotheses correspond to the edges of the graph (one hypothesis for each potential association between two nodes). It's natural here to assume that the prior of being null for a hypothesis follows the structure depicted in some community models, e.g. stochastic block model (SBM), degree-corrected SBM or general models, network with node attributes models, etc. We develop theories of SABHA with these structures, and apply them in simulations

to compare with other multiple testing methods in chapter 4.

Summary of theoretical and empirical results In Chapter 2, we prove that under the assumption that p-values are independent, the accumulation tests control the FDP with high probability; under an additional assumption that the null p-values are identically distributed, the accumulation tests control a modified FDR on finite samples, and controls FDR in asymptotic setting. In chapter 2 we also characterize the power of the accumulation methods, and provide theoretically-based guidance on how to choose an effective accumulation function. The accumulation tests outperform existing multiple testing methods consistently when it is applied to simulated data, including a high-dimensional model selection problem for linear regression, and a real data problem of identifying differential gene expression over a dosage gradient.

In Chapter 3, we proves that SABHA controls FDR at a level that is at most slightly higher than the target FDR level, as long as the prior probabilities are not overfitted to the data too much. We get the theoretical bounds on FDR for the ordered, grouped, and low total variation structures introduced above. The empirical performance of SABHA is examined on fMRI activity data and on gene/drug response data, as well as on simulated data.

In Chapter 4, we get the theoretical bounds of SABHA on FDR for the SBM, degree-corrected SBM (DC-SBM), and network with node attributes structures. We also study a degree-corrected general setting as an extension of any structured settings, and we show that the FDR bound of the DC general setting is just the FDR bound of the underlying structural setting, plus a penalty on the degree correction, which diminishes with increasing number of hypotheses. This greatly extends the potential applications of SABHA. We compare SABHA with other multiple testing methods in a simulation where the graph is generated based on a DC-SBM model, and in another simulation based on a node attributes model.

CHAPTER 2

ORDERED HYPOTHESIS TESTING WITH ACCUMULATION TESTS

In this chapter, we consider the multiple testing case where the hypotheses are accompanied by an ordering, where prior information leads us to believe that some hypotheses are more likely to contain a true signal than others, or where we want to test hypotheses in a certain order to respect the structure of the problem.

To describe the problem more precisely, suppose that we are interested in testing n hypotheses, denoted $H_1, \ldots, H_n$, and experimental data has yielded individual p-values for each of these hypotheses, which we write as $p_1, \ldots, p_n$. For a concrete example, we might be searching for a genetic cause for some particular disease, in which case we might test a hypothesis H_i that the ith SNP in our experiment is associated with the disease.

In many practical settings, the experiment that we would like to analyze has been carried out in the context of existing information from previous studies. At the same time, often this prior information cannot simply be treated as additional data in the statistical analysis. For instance, in the example above where p_i is a p-value testing the association between SNP i and some phenotype of interest, we might have prior information available from earlier experiments that may have:

- studied a different but related disease, or a different population of patients;
- used a different experimental protocol for genotyping the individuals;
- defined disease status differently, or measured the phenotype differently; or
- produced data that we believe may be unreliable.

In any of these scenarios, the data from the previous study cannot simply be integrated with our new experimental data, without violating the integrity of the statistical analysis.

However, this prior information is extremely valuable and can give us some power to detect signals in a very high-dimensional setting.

While the scenarios described above can correspond to many different forms of prior information about the n hypotheses being tested, in this chapter we focus on the specific problem of testing the hypotheses when the only prior information is a ranking of the list. Before performing the experiment, we use prior information to generate a ranked list of hypotheses $H_1, H_2, \ldots, H_n$, where H_1 is the hypothesis that we believe is most likely to correspond to a true signal, while H_n is the one believed to be least likely. After gathering new data, we then wish to test these hypotheses while taking this ordering into account.

Chapter outline[1]: In section 2.1, we formally introduce the ordered hypothesis testing problem considered here. In section 2.2, we give background on several existing methods, and develops our family of methods that generalizes these existing works. In section 2.3, we state and prove results for FDR control of the method in both finite-sample and asymptotic settings. In section 2.4, we calculate power of the method in an asymptotic setting. In section 2.5, we present experiments on simulated data to validate our theoretical results and provide an empirical comparison of various choices of the method within the general family. In section 2.6 we adapt our method to a dosage-response data set, where for a large set of genes, we would like to determine which genes respond to a low dose of a particular drug.

2.1 Ordered hypothesis testing problem

We begin by formally defining the problem that we consider here. Let $\mathcal{H}_0 \subseteq \{1, \ldots, n\}$ be a fixed set (the "null hypotheses"), and let $p_1, \ldots, p_n \in [0, 1]$ be random variables, such that the null p-values are independent from each other and independent from the non-null p-values. We do not require that the null p-values are exactly uniformly distributed, but instead allow for *conservative* nulls, with $p_i \succeq \mathsf{Uniform}[0, 1]$, meaning that p_i is stochastically

1. The work presented in this chapter is published in Li and Barber [2016a]

at least as large as a uniform random variable, i.e.

$$\mathbb{P}\{p_i \leq t\} \leq t \text{ for all } t \in [0,1], \text{ for all } i \in \mathcal{H}_0.$$

The null p-values may follow different distributions (although some of our theoretical results do require the nulls to be identically distributed).

Our method will construct a cutoff point $\widehat{k}$ that is adaptive to the data—formally, this cutoff is a function mapping the observed p-values $(p_1, \ldots, p_n)$ to some $\widehat{k} \in \{0, \ldots, n\}$. This cutoff $\widehat{k}$ is the output of our procedure, and should be interpreted as labeling the first $\widehat{k}$ hypotheses, i.e. $H_1, \ldots, H_{\widehat{k}}$, as "discoveries" (to use the terminology of hypothesis testing, we reject hypotheses $H_1, \ldots, H_{\widehat{k}}$ and do not reject $H_{\widehat{k}+1}, \ldots, H_n$).

Ideally, we would like to choose $\widehat{k}$ so that the selected list $H_1, \ldots, H_{\widehat{k}}$ contains only true signals and the remaining hypotheses $H_{\widehat{k}+1}, \ldots, H_n$ are all null. However, this may not be possible because our initial ranking may be imperfect—the ranked list $H_1, \ldots, H_n$ may contain signals and nulls interspersed with each other, meaning that there is no threshold that perfectly separates the signals from the nulls. However, the ranking is nonetheless informative if the signals are indeed concentrated towards the top of the ranked list, and we select $\widehat{k}$ with the goal of detecting as many signals as possible without too many false positives. To quantify this, we define the false discovery proportion (FDP) cumulatively along the list:

$$\mathsf{FDP}(k) = \frac{\mathsf{FalsePos}(k)}{k},$$

where $\mathsf{FalsePos}(k) = \#\{i \leq k : i \in \mathcal{H}_0\}$ is the number of false positives among the first k hypotheses. In other words, $\mathsf{FDP}(k)$ gives the proportion of false positives (i.e. null hypotheses) among the first k hypotheses in the list, i.e. $H_1, \ldots, H_k$. To agree with the definition of false discovery rate used in the literature, we define $\mathsf{FDP}(0) \coloneqq 0$ to cover the case that no discoveries are made; more formally, we can write $\mathsf{FDP}(k) = \frac{\mathsf{FalsePos}(k)}{\max\{1,k\}}$ to cover both cases $k = 0$ and $k \neq 0$. For ease of notation, we will omit this more precise definition and will

write $\frac{\mathsf{FalsePos}(k)}{k}$ with the understanding that $\frac{0}{0}$ is treated as 0.

Selecting a threshold $\widehat{k}$ involves a tradeoff: we would like a high $\widehat{k}$ to ensure that as many true signals as possible are captured in the selected list $H_1, \ldots, H_{\widehat{k}}$, but at the same time the proportion of false positives will generally increase farther down the list, since the signals will be concentrated towards the top of the ranking. In particular, we would like to bound the false discovery rate (FDR) Benjamini and Hochberg [1995], defined as the expectation of the FDP (where the expectation is taken with respect to the distribution of the p-values):

$$\text{False discovery rate} = \mathbb{E}\left[\mathsf{FDP}(\widehat{k})\right] = \mathbb{E}\left[\frac{\mathsf{FalsePos}(\widehat{k})}{\widehat{k}}\right] .$$

A note on the ranking In the setting described in the Introduction, where prior experience or preexisting data allows us to rank the hypotheses from most to least likely, this ranking is reflected in the indexing of the list. An implicit assumption is that this ranking takes place before the p-values are generated, that is, the p-values are independent from the process of ranking the hypotheses. For instance, we cannot use a data set to rank the hypotheses and then use the same data set to calculate p-values (unless the p-values are calculated in such a way that corrects for reusing the same data twice).

Application to high-dimensional regression In addition to situations where prior information, or data from related experiments, may provide a ranked list, this type of setting is also applicable to other statistical problems. As a key example, consider the problem of inference for sparse regression, where a response y depends linearly on some sparse subset of many possible features $X_1, \ldots, X_p$. When the sample size n is lower than the number of features p, classical methods for performing inference on the coefficients cannot be applied as the linear model is not identifiable (i.e. $X^\top X$ is rank-deficient). Recently, many approaches have been proposed for this inference problem, including a recent line of work by Taylor et al. [2014] that, when paired with the Lasso (penalized regression) or with a forward stepwise selection procedure, calculates p-values for each feature in the order that they are selected.

This provides a list of p-values with an inherent ordering, and therefore is an example of the ordered hypothesis testing problem we consider here.

2.2 Existing works and the accumulation test method

Existing methods Suppose that we would like to select a cutoff $\widehat{k}$ that is as large as possible, while bounding the FDR at some prespecified level α (e.g. $\alpha = 0.1$). Two approaches for the ordered hypothesis testing problem have been proposed recently in the literature. First, G'Sell et al. [2015] propose the ForwardStop method:

$$\widehat{k}_{\mathsf{ForwardStop}} = \max \left\{ k \in \{1, \dots, n\} : \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{1}{1-p_i} \right) \leq \alpha \right\} , \tag{2.1}$$

with the convention that if this set is empty, we set $\widehat{k}_{\mathsf{ForwardStop}} = 0$ and make no rejections. G'Sell et al. [2015, Theorem 1] prove that this procedure controls FDR at the level α, that is,

$$\mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{ForwardStop}}) \right] \leq \alpha \, .$$

A second existing method uses a similar summation to estimate the FDP, but with a discrete step function rule rather than continuous measure: given a parameter $C > 1$, the Sequential Step-up Procedure (SeqStep) of Barber and Candès [2015] sets

$$\widehat{k}_{\mathsf{SeqStep}(C)} = \max \left\{ k \in \{1, \dots, n\} : \frac{1}{k} \sum_{i=1}^{k} C \cdot \mathbb{1}\{p_i > 1 - 1/C\} \leq \alpha \right\} . \tag{2.2}$$

and Barber and Candès [2015, Theorem 3] prove that this procedure controls a modified form of the FDR:

$$\mathbb{E}\left[\frac{\#\{i \leq \widehat{k}_{\mathsf{SeqStep}(C)} : i \in \mathcal{H}_0\}}{C/\alpha + \widehat{k}_{\mathsf{SeqStep}(C)}} \right] \leq \alpha \, . \tag{2.3}$$

For exact FDR control, the same paper proposes a slightly more conservative variant, the

SeqStep+ method Barber and Candès [2015], defined by

$$\widehat{k}_{\mathsf{SeqStep}_+(C)} = \max\left\{k \in \{1,\ldots,n\} : \frac{1}{1+k}\left(C + \sum_{i=1}^{k} C \cdot \mathbb{1}\{p_i > 1 - 1/C\}\right) \leq \alpha\right\}, \tag{2.4}$$

with the guarantee that for any $C > 1$,

$$\mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{SeqStep}_+(C)})\right] \leq \alpha\,.$$

Another procedure known to us is the α-investing by Foster and Stine [2008], which controls the the ratio $\frac{\mathbb{E}[V]}{\mathbb{E}[R]+1}$ (where V = number of false positives and R = total number of discoveries), a criterion weaker than the FDR. It allows users to incorporate prior knowledge such as ordering and improve the power. However, α-investing shows lower power than ForwardStop in simulations carried out by G'Sell et al. [2015].

Ordered testing procedures have been shown to provide FDR control in regression models. Taylor et al. [2014] derived post-selection hypothesis tests at each step of the forward stepwise and LARS procedures. These tests yield an ordered list of p-values, corresponding to a nested sequence of models. With the ordered testing procedures, these sequential p-values can be transformed into a model selection procedure with FDR guarantee (G'Sell et al. [2015]), which we explore in Section 2.5.

A general family of methods Examining G'Sell et al. [2015]'s ForwardStop procedure and Barber and Candès [2015]'s SeqStep procedures above, we see that the two share a common structure:

$$\widehat{k} = \max\left\{k \in \{1,\ldots,n\} : \frac{1}{k}\sum_{i=1}^{k} h(p_i) \leq \alpha\right\}\,.$$

where $\int_{t=0}^{1} \mathsf{h}(t)\ \mathsf{d}t = 1, \mathsf{h} : [0,1] \mapsto [0,\infty)$ and is monotone nondecreasing. We now generalize these two procedures with a broader family:

Definition 1 (Accumulation test for ordered hypotheses)**.** Suppose we are given a ranked list of n hypotheses $H_1, \ldots, H_n$ with corresponding p-values $p_1, \ldots, p_n \in [0,1]$. Fix any monotone nondecreasing function $\mathsf{h} : [0,1] \mapsto [0,\infty)$ that satisfies $\int_{t=0}^{1} \mathsf{h}(t)\ \mathsf{d}t = 1$. Define the estimated FDP at each cutoff $k \in \{1, \ldots, n\}$ as

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) = \frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k} \ ,$$

and then select the adaptive cutoff

$$\widehat{k}_{\mathsf{h}} = \max\left\{k \in \{1, \ldots, n\} : \widehat{\mathsf{FDP}}_{\mathsf{h}}(k) \leq \alpha\right\} \ ,$$

where $\alpha \in (0,1)$ is a prespecified target FDR level. (We use the convention that $\widehat{k}_{\mathsf{h}} = 0$ if this set is empty.) We then reject the hypotheses $H_1, \ldots, H_{\widehat{k}_{\mathsf{h}}}$ and do not reject $H_{\widehat{k}_{\mathsf{h}+1}}, \ldots, H_n$.

We call h the "accumulation function", as for any fixed k, the sum $\sum_{i=1}^{k} \mathsf{h}(p_i)$ is the estimated "accumulation" of false positives by the time we have reached the kth position in the list:

$$\mathbb{E}\left[\sum_{i=1}^{k} \mathsf{h}(p_i)\right] \geq \mathbb{E}\left[\sum_{i \leq k, i \in \mathcal{H}_0} \mathsf{h}(p_i)\right] \geq \mathsf{FalsePos}(k) \ ,$$

where the last step holds because

$$\mathbb{E}\left[h(p_i)\right] \geq \mathbb{E}_{U \sim \mathsf{Uniform}[0,1]}\left[h(U)\right] = \int_{t=0}^{1} \mathsf{h}(t)\ \mathsf{d}t = 1$$

for each $i \in \mathcal{H}_0$, due to the fact that $p_i \succeq \mathsf{Uniform}[0,1]$ and $h(\cdot)$ is monotone nondecreasing.

For any accumulation function h, the data-dependent cutoff $\widehat{k}_{\mathsf{h}}$ should intuitively control the false discovery rate at the desired level, since for each k, the estimated false discovery proportion $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ is an overestimate of the actual FDP:

$$\mathbb{E}\left[\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)\right] = \frac{\mathbb{E}\left[\sum_{i=1}^{k} \mathsf{h}(p_i)\right]}{k} \geq \frac{\mathsf{FalsePos}(k)}{k} = \mathsf{FDP}(k) \ . \tag{2.5}$$

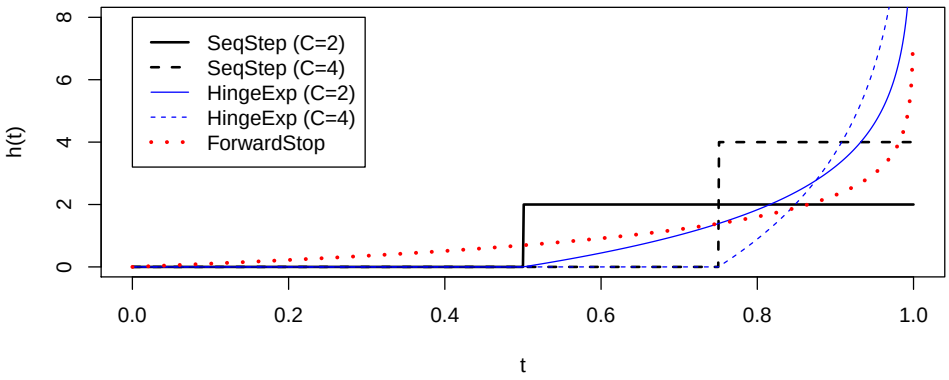

Figure 2.1: Illustration of several choices of the accumulation function $\mathsf{h} : [0,1] \to [0,\infty)$.

and $\widehat{k}$ is the last time that this overestimate is below the target FDR level.

Note that our notation implicitly treats the desired FDR bound α as fixed, while h appears in the subscript to emphasize that we may choose between many candidate accumulation functions h. In particular, choosing the accumulation function $\mathsf{h}_{\mathsf{ForwardStop}}(t) = \log\left(\frac{1}{1-t}\right)$ yields G'Sell et al. [2015]'s ForwardStop procedure, while $\mathsf{h}_{\mathsf{SeqStep}}(t) = C \cdot \mathbb{1}\{t > 1 - 1/C\}$ yields Barber and Candès [2015]'s SeqStep procedure (for any parameter $C > 1$). We will also study an additional choice for the accumulation function h, the "HingeExp" function,

$$\mathsf{h}_{\mathsf{HingeExp}}(t) = \begin{cases} C \cdot \log\left(\frac{1}{C \cdot (1-t)}\right), & \text{for } t > 1 - 1/C, \\ 0, & \text{for } t \leq 1 - 1/C. \end{cases} \tag{2.6}$$

This function combines the ideas of the step function used in SeqStep with the ForwardStop method; Figure 2.1 gives a visual comparison of these methods. The name "HingeExp" arises from the "hinge point" of the function at $t = 1 - 1/C$ (similar to the hinge loss function in machine learning), combined with the observation that, for a null p-value $p_i \sim \mathsf{Uniform}[0,1]$, we have $\mathsf{h}_{\mathsf{HingeExp}}(p_i)$ distributed as C times an $\mathsf{Exp}(1)$ random variable with probability $1/C$ (and equal to zero otherwise).

2.3 FDR control in finite-sample and asymptotic settings

2.3.1 Finite sample setting

We begin with concrete finite-sample results to bound the false discovery proportion (FDP) or false discovery rate (FDR) of the family of accumulation tests. Recall that, when a threshold $k \in \{1, \dots, n\}$ is selected, we are interested in the false discovery proportion

$$\mathsf{FDP}(k) = \frac{\mathsf{FalsePos}(k)}{k} = \frac{\#\{i \le k : i \in \mathcal{H}_0\}}{k}$$

(with the convention that $\mathsf{FDP}(0) = 0$).

We will derive two different finite-sample results for our methods. First, we will work under a more general setting, where the p-value are required to be independent but the nulls might not be identically distributed; in particular, we may have some "conservative" null p-values (where $\mathbb{P}\{p_i \le t\}$ is strictly smaller than t), and this distribution can vary across the different nulls.[2] In this setting, we give some bounds on the FDP. This result appears in two forms: first, for the accumulation test method as given, we bound the probability that the FDP exceeds $\alpha + \mathcal{O}\left(\sqrt{\log(\log(\widehat{k}_{\mathsf{h}}))/\widehat{k}_{\mathsf{h}}}\right)$. In a scenario where we expect many discoveries to be made, this essentially bounds the FDP very near to the desired level α. Second, for scenarios where a stricter bound is desired, we offer a more conservative method which bounds the FDP at α exactly, but will exhibit significantly lower power early in the sequence (i.e. if the cutoff value is low).

Before we proceed, we recall the definition of a subexponential random variable:

$$X \text{ is } (\sigma^2, b)\text{-subexponential if } \mathbb{E}\left[e^{\theta(X - \mathbb{E}[X])}\right] \le e^{\theta^2\sigma^2/2} \text{ for all } |\theta| \le \frac{1}{b}. \tag{2.7}$$

2. There is one minor caveat here: if the p-values are conservative then it's possible to have $\mathbb{P}\{p_i = 1\} > 0$; in this case, accumulation functions such as ForwardStop or HingeExp will fail, since $\mathsf{h}(1) = +\infty$. In practice, for any such choice of h, we would want to truncate h at some large but finite value, i.e. replace $\mathsf{h}(t)$ with $\mathsf{h}(t) \wedge C$ for some large positive C. Since this will reduce the expected value only slightly, i.e. $\int_{t=0}^{1} \mathsf{h}(t) \wedge C \, \mathsf{d}t$ is only slightly smaller than 1, the FDR and FDP bounds would only be slightly worse.

Note that a σ^2-subgaussian random variable, such as a $N(0, \sigma^2)$ variable, is $(\sigma^2, 0)$-subexponential trivially, and so the subexponential condition is weaker than the subgaussian condition.

We will say that an accumulation function h is (σ^2, b)-subexponential with respect to the p-values if $\mathsf{h}(p_i)$ is (σ^2, b)-subexponential for each $i = 1, \ldots, n$. In particular, note that the SeqStep function is bounded and therefore subexponential for any distribution of the p-values; the ForwardStop and HingeExp functions are both subexponential in the setting where the null p-values are uniform and the non-null p-values are stochastically no larger than uniform.

Theorem 1. *Suppose that the p-values are independent, and that* $p_i \succeq \mathsf{Uniform}[0,1]$ *stochastically for each* $i \in \mathcal{H}_0$. *Let* $\mathsf{h} : [0,1] \to [0,\infty)$ *be any monotone nondecreasing function with* $\int_{t=0}^{1} \mathsf{h}(t)\ \mathsf{d}t = 1$, *such that* h *is* (σ^2, b)*-subexponential with respect to the p-values, and let* $\alpha \in (0,1)$ *be some prespecified target FDR level. Fix any* $C > 0$ *and any* $\epsilon > 0$. *Define*

$$\widehat{k}_{\mathsf{h}} = \max\left\{ k \in \{1, \ldots, n\} : \frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k} \leq \alpha \right\}, \tag{2.8}$$

with the convention that $\widehat{k}_{\mathsf{h}} = 0$ *if this set is empty. Then*

$$\mathbb{P}\left\{ \mathsf{FDP}(\widehat{k}_{\mathsf{h}}) \leq \alpha + C_\epsilon \sqrt{\frac{16 \log(\frac{8 \log(2\widehat{k}_{\mathsf{h}})}{\epsilon})}{\widehat{k}_{\mathsf{h}}}} \right\} \geq 1 - \epsilon \tag{2.9}$$

where $C_\epsilon := \max\left\{\sigma, b\sqrt{3\log(8/\epsilon)}\right\}$. *Next, take any* $C \geq C_\epsilon$, *and define*

$$\widehat{k}_{\mathsf{h}}^{+C*} = \max\left\{ k \in \{1, \ldots, n\} : \frac{C\sqrt{16k \log\left(\frac{8\log(2k)}{\epsilon}\right)} + \sum_{i=1}^{k} \mathsf{h}(p_i)}{k} \leq \alpha \right\}, \tag{2.10}$$

with the same convention. Then

$$\mathbb{P}\left\{\mathsf{FDP}(\widehat{k}_{\mathsf{h}}^{+C*}) \leq \alpha\right\} \geq 1 - \epsilon. \tag{2.11}$$

The proof of this theorem follows from a lemma on random walks, which is an extension of Jamieson et al. [2014]'s recent finite-sample Law of the Iterated Logarithm result: we show that, with high probability, for each k, the accumulated sum over the first k hypotheses is a reliable estimate of the number of false positives at that point:

$$\sum_{i \leq k, i \in \mathcal{H}_0} \mathsf{h}(p_i) \geq \mathsf{FalsePos}(k) - \mathcal{O}\left(\sqrt{k \log(\log(k))}\right).$$

In particular, if this holds for the selected cutoff $k = \widehat{k}_{\mathsf{h}}$ or $k = \widehat{k}_{\mathsf{h}}^{+C*}$, then the resulting bounds on FDP will hold.

Note that, for the first bound (using the original cutoff $\widehat{k}_{\mathsf{h}}$), as $\epsilon \to 1$ the upper bound on FDP approaches $\alpha + C_\epsilon \sqrt{\frac{16 \log(8 \log(2\widehat{k}_{\mathsf{h}}))}{\widehat{k}_{\mathsf{h}}}}$; this is not a true "tail bound" in the sense of bounding FDP to be near α, but is instead a "modified tail bound" which is essentially equivalent when $\widehat{k}_{\mathsf{h}}$ is large but may be substantially different if the number of discoveries is small.

Next, we turn to our second main result, which examines FDR control under slightly more restrictive assumptions—we now require the null p-values to be identically distributed. (Note however that we can allow for "conservative" null p-values, which are stochastically larger than a uniform distribution, as long as they are i.i.d.). In addition to the FDR, we will also consider a modified form of the FDP, introduced in Barber and Candès [2015] for the SeqStep method, given by the expectation of the modified FDP,

$$\mathsf{mFDP}_c(k) = \frac{\mathsf{FalsePos}(k)}{c + k} = \frac{\#\{i \leq k : i \in \mathcal{H}_0\}}{c + k}.$$

Of course, when c is a constant while k is large, the modified FDP is nearly identical to the

FDP.

Our next result shows that the accumulation test controls the modified FDR (i.e. the expected value of the modified FDP). Furthermore, a slightly more conservative test, which is defined in the theorem, controls the original FDR (i.e. the expected value of the FDP).

Theorem 2. *Suppose that the p-values are independent, that the null p-values are identically distributed, and that* $p_i \succeq \mathsf{Uniform}[0,1]$ *stochastically for each* $i \in \mathcal{H}_0$. *Let* $\mathsf{h} : [0,1] \to [0,\infty)$ *be any monotone nondecreasing function with* $\int_{t=0}^{1} \mathsf{h}(t)\ \mathsf{d}t = 1$, *and let* $\alpha \in (0,1)$ *be some prespecified target FDR level. Fix any* $C > 0$. *Define*

$$\widehat{k}_{\mathsf{h}} = \max\left\{ k \in \{1,\dots,n\} : \frac{1}{k}\sum_{i=1}^{k} \mathsf{h}(p_i) \le \alpha \right\}\ , \tag{2.12}$$

with the convention that $\widehat{k}_{\mathsf{h}} = 0$ *if this set is empty, and define*

$$\widehat{k}_{\mathsf{h}}^{+C} = \max\left\{ k \in \{1,\dots,n\} : \frac{1}{1+k}\left(C + \sum_{i=1}^{k} \mathsf{h}(p_i)\right) \le \alpha \right\}\ , \tag{2.13}$$

with the same convention. Then, in the special case that $\max_{0\le t\le 1} \mathsf{h}(t) \le C$, *we have*

$$\mathbb{E}\left[\mathsf{mFDP}_{C/\alpha}(\widehat{k}_{\mathsf{h}})\right] \le \alpha \quad \textit{and} \quad \mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{h}}^{+C})\right] \le \alpha\ . \tag{2.14}$$

In the general case, with no restriction on the range of h, *we have*

$$\mathbb{E}\left[\mathsf{mFDP}_{C/\alpha}(\widehat{k}_{\mathsf{h}})\right] \le \frac{\alpha}{\int_{t=0}^{1} \mathsf{h}(t) \wedge C\ \mathsf{d}t} \quad \textit{and} \quad \mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{h}}^{+C})\right] \le \frac{\alpha}{\int_{t=0}^{1} \mathsf{h}(t) \wedge C\ \mathsf{d}t}\ , \tag{2.15}$$

where we use the notation $a \wedge b \coloneqq \min\{a,b\}$.

We pause here to give a brief sketch of the proof of this result. Treating only the case of

the cutoff $\widehat{k}_{\mathsf{h}}^{+C}$ here, we write

$$\mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{h}}^{+C})\right] = \mathbb{E}\left[\frac{\mathsf{FalsePos}(\widehat{k}_{\mathsf{h}}^{+C})}{\widehat{k}_{\mathsf{h}}^{+C} \vee 1}\right] \leq \mathbb{E}\left[\frac{1+\mathsf{FalsePos}(\widehat{k}_{\mathsf{h}}^{+C})}{1+\widehat{k}_{\mathsf{h}}^{+C}}\right]$$
$$= \mathbb{E}\left[\frac{1+\mathsf{FalsePos}(\widehat{k}_{\mathsf{h}}^{+C})}{C+\sum_{i=1}^{\widehat{k}_{\mathsf{h}}^{+C}}\mathsf{h}(p_i)} \cdot \frac{C+\sum_{i=1}^{\widehat{k}_{\mathsf{h}}^{+C}}\mathsf{h}(p_i)}{1+\widehat{k}_{\mathsf{h}}^{+C}}\right] \leq \alpha \cdot \mathbb{E}\left[\frac{1+\mathsf{FalsePos}(\widehat{k}_{\mathsf{h}}^{+C})}{C+\sum_{i=1}^{\widehat{k}_{\mathsf{h}}^{+C}}\mathsf{h}(p_i)}\right],$$

where the last step holds because $\frac{C+\sum_{i=1}^{\widehat{k}_{\mathsf{h}}^{+C}}\mathsf{h}(p_i)}{1+\widehat{k}_{\mathsf{h}}^{+C}} \leq \alpha$ by definition of the choice of $\widehat{k}_{\mathsf{h}}^{+C}$. The main portion of the proof, then, lies in proving that the remaining expectation term is bounded by 1; intuitively, this is plausible since $\mathbb{E}[\mathsf{h}(p_i)] \geq 1$ for each null, and so we would typically expect the denominator to be (approximately) as large as the numerator. The formal result is obtained via martingale theory, in the Supplementary Materials.

We also give a result specifically for the HingeExp function, which gives a tighter bound than that guaranteed by Theorem 2:

Lemma 1. *Suppose that the p-values are independent, and that* $p_i \sim \mathsf{Uniform}[0,1]$ *for all* $i \in \mathcal{H}_0$. *Let* $\mathsf{h}(t) = C \cdot \log\left(\frac{1}{C(1-t)}\right) \cdot \mathbb{1}_{t>1-1/C}$, *i.e. the HingeExp function with parameter* C. *Then, under the same definitions and assumptions as Theorem 2,*

$$\mathbb{E}\left[\mathsf{mFDP}_{2C/\alpha}(\widehat{k}_{\mathsf{h}})\right] \leq \alpha\ .$$

Comparison to existing results As discussed in Section 2.2, the accumulation test contains two existing procedures as special cases: SeqStep Barber and Candès [2015] with the step function $\mathsf{h}(t) = C \cdot \mathbb{1}\{t > 1 - 1/C\}$, and ForwardStop G'Sell et al. [2015] with $\mathsf{h}(t) = \log\left(\frac{1}{1-t}\right)$. Furthermore, the slightly altered accumulation test given in (2.13) contains as a special case the SeqStep+ procedure Barber and Candès [2015], again with $\mathsf{h}(t) = C \cdot \mathbb{1}\{t > 1 - 1/C\}$. For the special cases of SeqStep and SeqStep+, our main result, Theorem 2, obtains the same guarantee on modified and original FDP as proved in Barber and Candès [2015, Theorem 3]. For the special case of the ForwardStop procedure, the results obtained

in Theorem 2 and Lemma 1 are somewhat weaker than the guarantee given in G'Sell et al. [2015, Theorem 1], which proves that

$$\mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{h}})\right] \leq \alpha \ ,$$

that is, a guarantee on the FDP rather than the modified FDP (in the setting where the null p-values are i.i.d. $\mathsf{Uniform}[0,1]$ variables). As we will see in Theorem 3, however, asymptotically the same result is obtained.

2.3.2 Asymptotic setting

In our finite-sample results (Theorem 1 and Theorem 2), we proved that the accumulation test (using the original cutoff $\widehat{k}_{\mathsf{h}}$) gives a modified tail bound on the FDP, and controls a modified form of the FDR, which as discussed earlier, is nearly equal to the original FDP as long as the number of rejections $\widehat{k}_{\mathsf{h}}$ is large. Next, we show that the accumulation test controls FDP and FDR asymptotically as long as the number of rejections $\widehat{k}_{\mathsf{h}}$ tends to infinity.

Theorem 3. *Consider a sequence of ordered hypothesis testing problems, with $n = 1, 2, \ldots$, each satisfying the assumptions of Theorem 1, for some fixed α, σ^2, b. Suppose that there exists a sequence $m_n \in \mathbb{N}$ with $m_n \to \infty$ such that*

$$\mathbb{P}\left\{\widehat{k}_{\mathsf{h}} < m_n\right\} \to 0 \ \textit{as} \ n \to \infty \ .$$

Then for any $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{\mathsf{FDP}(\widehat{k}_{\mathsf{h}}) > \alpha + \delta\right\} = 0 \ .$$

In particular, this implies that

$$\lim_{n\to\infty} \mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{h}})\right] \leq \alpha \ .$$

Intuitively, this holds because, as $\widehat{k}_{\mathsf{h}} \to \infty$, the difference between the modified FDP and the (original) FDP becomes negligible; their denominators, $\widehat{k}_{\mathsf{h}} + C\alpha^{-1}$ as compared to $\widehat{k}_{\mathsf{h}} \vee 1$, are nearly identical once $\widehat{k}_{\mathsf{h}}$ grows large.

2.4 Calculation of power in an asymptotic setting

Up to this point, our discussion and theoretical analysis of accumulation tests has focused on controlling false discoveries. Of course, in practice we are interested in balancing the goals of reducing false positives (Type I error) while increasing the number of true discoveries (power). Formally, we will define the true positive proportion (TPP) of a cutoff $k \in \{1, \dots, n\}$ as the proportion of non-nulls that are discovered by the test,

$$\mathsf{TPP}(k) = \frac{\#\{i \le k : i \notin \mathcal{H}_0\}}{\#\{i : i \notin \mathcal{H}_0\}} ,$$

and will aim to choose the accumulation function h to maximize the asymptotic TPP of the accumulation test. More precisely, in an asymptotic setting, we will show that TPP converges in probability to a fixed value which is a function of the choice of accumulation function h. Formally, the power of a test is defined as the expected TPP; we will use the terms power and TPP somewhat interchangeably for the asymptotic setting, because if TPP converges to a constant value, then its expectation (the power) converges to this value as well.

Recall that $\widehat{k}_{\mathsf{h}}$ is defined as the largest k such that

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) = \frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k} \le \alpha .$$

In other words, high power (or, more precisely, high TPP) corresponds to a large value of $\widehat{k}$, which is possible only when the estimates $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ are low. Therefore, a good choice of the accumulation function h is one that allows for low estimates $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ of the false discovery

proportion along the sequence of hypotheses. Consider the expectation of this estimate at any fixed k,

$$\mathbb{E}\left[\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)\right] = \frac{\sum_{i=1}^{k}\mathbb{E}\left[\mathsf{h}(p_i)\right]}{k} = \frac{\sum_{i\le k, i\in\mathcal{H}_0}\mathbb{E}\left[\mathsf{h}(p_i)\right]}{k} + \frac{\sum_{i\le k, i\notin\mathcal{H}_0}\mathbb{E}\left[\mathsf{h}(p_i)\right]}{k}$$
$$\geq \mathsf{FDP}(k) + \frac{\sum_{i\le k, i\notin\mathcal{H}_0}\mathbb{E}\left[\mathsf{h}(p_i)\right]}{k},$$

where as before the last step holds because $\mathbb{E}\left[\mathsf{h}(p_i)\right] \geq 1$ for $i \in \mathcal{H}_0$, as shown before. Therefore, for a given choice of h, we see that $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ is an overestimate of $\mathsf{FDP}(k)$, with bias at least

$$\frac{\sum_{i\le k, i\notin\mathcal{H}_0}\mathbb{E}\left[\mathsf{h}(p_i)\right]}{k}.$$

Since the power (or TPP) will increase if the estimated false discovery proportion $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ is small, we see that a good choice of accumulation function h is one that minimizes $\mathbb{E}\left[\mathsf{h}(p_i)\right]$ for non-nulls $i \notin \mathcal{H}_0$ (while, of course, satisfying the requirements of accumulation functions).

In the following section, we will examine how $\mathbb{E}\left[\mathsf{h}(p_i)\right]$ affects the power by studying an asymptotic scenario. We will find the power of any accumulation function h can be characterized exactly by its expected value over the distributions of the null p-values and of the non-null p-values (Theorem 4). If all the null p-values are i.i.d. draws from some (uniform or conservative) null distribution $\mathcal{D}_0$, and all the non-null p-values are i.i.d. draws from some alternative distribution $\mathcal{D}_1$, we can think of these results as characterizing the power of an accumulation function h for testing the null hypothesis given by the null distribution $\mathcal{D}_0$ against the alternative hypothesis given by the distribution $\mathcal{D}_1$.

Of course, in practice, even if we assume that the null p-values are uniformly distributed (i.e. $\mathcal{D}_0 = \mathsf{Uniform}[0,1]$,) we will not always know the distribution $\mathcal{D}_1$ of the non-null p-values. In Section 2.4.2 we discuss the problem of determining a good choice of accumulation function h without prior knowledge of an alternate distribution.

2.4.1 Power calculation in an asymptotic setting

In this section, we show that we can calculate the asymptotic power of the accumulation methods, and compare between them, under the assumption that the proportion of non-nulls along the list is converging to a fixed function $f(\cdot)$, where $f(\cdot)$ must satisfy some mild conditions. Specifically, we consider an asymptotic scenario where

$$\max_{\log(n)<k\leq n} \left| \frac{\#\{i \leq k : i \notin \mathcal{H}_0\}}{k} - f\left(\frac{k}{n}\right) \right| \leq \epsilon_n \ , \tag{2.16}$$

and $\epsilon_n \to 0$ as $n \to \infty$. We assume that $f : [0,1] \to [0,1]$ is differentiable and satisfies, for some constants $\delta, \nu > 0$ and for the fixed target FDR level α,

$$\begin{cases} f \text{ is monotone nonincreasing over } t \in [0,1], \\ f(t) \leq f(t - \Delta_t) - \nu\Delta_t \text{ for all } t \text{ such that } f(t) \geq 1 - \alpha \text{ and } 0 \leq \Delta_t \leq \delta, \\ t \mapsto t \cdot f(t) \text{ is a nondecreasing function.} \end{cases} \tag{2.17}$$

In words, these conditions require that:

$$\begin{cases} \text{The proportion of true signals decreases as we move along the list;} \\ \text{The proportion is decreasing at a positive rate during the initial portion of the list; and} \\ \text{The } \textit{number} \text{ of true signals must of course increase as we move along the list.} \end{cases}$$

Then we have the following theorem:

Theorem 4. *Fix any $\sigma^2 > 0$ and $b \geq 0$, any target FDR level $\alpha \in (0,1)$, any $\mu_0 \geq 1$, and any $0 < \mu_1 < \mu_0$. Suppose that:*

- *The p-values are independent, with $p_i \overset{\text{iid}}{\sim} \mathcal{D}_0$ for all $i \in \mathcal{H}_0$ and $p_i \overset{\text{iid}}{\sim} \mathcal{D}_1$ for all $i \notin \mathcal{H}_0$ for some fixed distributions $\mathcal{D}_0 \succeq \mathsf{Uniform}[0,1]$ and $\mathcal{D}_1$;*
- *The function $\mathsf{h} : [0,1] \mapsto [0,\infty)$ is monotone nondecreasing, satisfies $\mathbb{E}_{p_i \sim \mathcal{D}_j}[h(p_i)] =$*

μ_j for $j = 0, 1$, and is (σ^2, b)-subexponential with respect to the p-values; and

- *The function $f : [0, 1] \to [0, 1]$ satisfies assumptions (2.16) and (2.17).*

Define[3]

$$T = \begin{cases} 0, & \text{if } \frac{\mu_0 - \alpha}{\mu_0 - \mu_1} \geq f(0), \\ f^{-1}\left(\frac{\mu_0 - \alpha}{\mu_0 - \mu_1}\right) \in (0, 1), & \text{if } f(1) < \frac{\mu_0 - \alpha}{\mu_0 - \mu_1} < f(0), \\ 1, & \text{if } \frac{\mu_0 - \alpha}{\mu_0 - \mu_1} \leq f(1). \end{cases}$$

Then the true positive proportion (TPP) converges as

$$\mathsf{TPP}(\widehat{k}_{\mathsf{h}}) = \frac{\left|\{1, \ldots, \widehat{k}_{\mathsf{h}}\} \backslash \mathcal{H}_0\right|}{n - |\mathcal{H}_0|} \to T \cdot \frac{f(T)}{f(1)},$$

where specifically this denotes convergence in probability as $n \to \infty$.

Remark 1. While the proof of this result is fairly technical, the intuition behind it is simple. As n grows large, for any fixed t, treating tn as an integer for simplicity, we would have

$$\begin{aligned} \widehat{\mathsf{FDP}}_{\mathsf{h}}(tn) \approx \mathbb{E}\left[\widehat{\mathsf{FDP}}_{\mathsf{h}}(tn)\right] &= \frac{\sum_{i=1}^{tn} \mathbb{E}\left[\mathsf{h}(p_i)\right]}{tn} \\ &\approx \frac{f(t) \cdot tn \cdot \mu_1 + (1 - f(t)) \cdot tn \cdot \mu_0}{tn} = \mu_0 - f(t) \cdot (\mu_0 - \mu_1), \end{aligned}$$

where the first approximation holds because, for large n and constant t, we expect $\widehat{\mathsf{FDP}}_{\mathsf{h}}(tn)$ to concentrate near its expectation, while the second approximation holds since we expect approximately $f(t) \cdot tn$ many non-nulls and $(1 - f(t)) \cdot tn$ many nulls among the first tn p-values.

Now, by definition of T, we see that $\mu_0 - f(t) \cdot (\mu_0 - \mu_1) \leq \alpha$ if and only if $t \leq T$; using the above approximation, this means that $\widehat{\mathsf{FDP}}_{\mathsf{h}}(tn) \leq \alpha$ if and only if $t \lessapprox T$. In other

3. Note that the second assumption in (2.17) ensures that T is uniquely defined in each of these cases, i.e. in the middle case where $f(1) < \frac{\mu_0 - \alpha}{\mu_0 - \mu_1} < f(0)$, the inverse f^{-1} is well-defined at the value where it is applied.

words, we will reject

$$\widehat{k}_{\mathsf{h}} = \max\{k : \widehat{\mathsf{FDP}}_{\mathsf{h}}(k) \leq \alpha\} \approx Tn$$

many hypotheses. Now let's look at the power of this procedure. There are roughly $f(1) \cdot n$ total non-nulls; roughly $f(T) \cdot Tn$ of them appear in among the first Tn hypotheses. This means that our TPP will be roughly equal to $Tf(T)/f(1)$, as claimed in the theorem.

Remark 2. Note that, when holding α, μ_0, and $f(\cdot)$ fixed, T is a nonincreasing function of $\mu_1 = \mathbb{E}_{p_i \sim \mathcal{D}_1}[\mathsf{h}(p_i)]$. Therefore, the (asymptotic) TPP is a nonincreasing function of μ_1 as well, since $T \mapsto T \cdot f(T)$ is nondecreasing. In the simple case where the nulls are uniform (and so $\mu_0 = 1$ for any choice of accumulation function), this means if $\mathbb{E}[\mathsf{h}_0(p_i)] \leq \mathbb{E}[\mathsf{h}(p_i)]$ for non-nulls $i \notin \mathcal{H}_0$, the accumulation test using $\mathsf{h}_0(\cdot)$ is asymptotically more powerful than that using $\mathsf{h}(\cdot)$.

2.4.2 Choosing the accumulation function

While our earlier results prove control of the (modified) FDR for a broad class of accumulation functions, without any knowledge about the presence or absence of true signals, our understanding of the power of these methods does depend on the "alternate hypothesis", i.e. the distribution of the non-null p-values. In particular, Theorem 4 shows that the power of some specific accumulation function h can be viewed, asymptotically, as a simple function of this alternate hypothesis. However, without knowledge of this alternate hypothesis, how can we proceed—that is, how can we choose an effective h?

Here we suggest two partial answers to this question. First, observe that in the proofs of our FDR results, Theorems 2 and 3, we do not actually assume that the p-values are independent, only that they are "unlabeled", meaning that we may know the values in the set $\{p_1, \ldots, p_n\}$ but do not know the order in which these values will appear along the sequence. Therefore, we can observe the *unordered* set of p-values, $\{p_i : i = 1, \ldots, n\}$ before choosing the accumulation function h, without negating the FDR control properties of the

method. In particular, we may then use an empirical Bayes method (e.g. see Efron [2010]) to estimate the distribution of the non-null p_i's, which can then guide us in selecting h.

Quite unexpectedly, in one special case it is possible to choose an optimal h without any knowledge of the alternate distribution: the case of bounded accumulation functions. The following result shows that the step function $\mathsf{h}(t) = C \cdot \mathbb{1}_{t>1-1/C}$, which is used in the SeqStep method of Barber and Candès [2015], is the optimal C-bounded accumulation function under a very mild assumption on the distribution of non-null p-values:

$$\begin{gathered}\text{The non-null p-value } p_i \text{ has a density } f_i : [0,1] \to [0,\infty), \\ \text{where } f_i \text{ is a nonincreasing function.}\end{gathered} \tag{2.18}$$

We say that p_i satisfies the assumption (2.18) *strictly* if its density f_i is a strictly decreasing function.

Lemma 2. *Consider any accumulation function* h *bounded by* $C \geq 1$*, that is,* $\mathsf{h} : [0,1] \to [0,C]$ *is monotone nondecreasing and satisfies* $\int_{t=0}^{1} \mathsf{h}(t)\, \mathsf{d}t = 1$*. Let* h_0 *be the step function with the same bound,* $\mathsf{h}_0(t) = C \cdot \mathbb{1}\{t > 1 - 1/C\}$*. Suppose that a non-null p-value* p_i *satisfies the assumption* (2.18)*. Then*

$$\mathbb{E}\left[\mathsf{h}(p_i)\right] \geq \mathbb{E}\left[\mathsf{h}_0(p_i)\right] \ .$$

Furthermore, the inequality is strict whenever p_i *satisfies* (2.18) *strictly, unless* $\mathsf{h}(t) = \mathsf{h}_0(t)$ *almost everywhere on* $t \in [0,1]$*.*

In other words, based on the discussion above, we expect the step function (the SeqStep method) to offer more power than any other accumulation function that maps to the same range, as long as the non-null p-values satisfy (2.18) (and the null p-values are uniform), which is natural because we expect non-null p-values to give evidence against the null hypothesis, placing more mass on low values (near 0) than on high values (near 1).

We expect that similar results may be possible under weaker assumptions on the function h (e.g. subexponential tails), and leave this question to future work.

2.5 Simulations for ordered hypothesis testing problem

In this section we evaluate the performance of various accumulation tests on two tasks with simulated data: a ranked hypothesis testing problem (Section 2.5.1), and a high-dimensional linear regression problem (Section 2.5.2). Code to reproduce the first simulated data experiment is available online.[4]

2.5.1 Simulations for the ranked hypothesis testing problem

Here we examine our accumulation test under four different simulation settings, and compare the performance of several accumulation test methods: SeqStep and SeqStep+ with parameter $C = 2$, ForwardStop, and the new HingeExp method with parameter $C = 2$ (see Section 2.2 for the method definitions). The sequences of hypotheses and of p-values in our simulations vary on the degree of separation between the nulls and the non-nulls (the extent to which non-null hypotheses concentrate early in the list), and on the signal strength of the non-nulls (the extent to which non-null p-values are visibly different from a uniform distribution).

Methods

To create the simulated data, we generate the sequence of p-values for $n = 1000$ hypotheses with 100 non-nulls, by the following steps:

1. First, we generate "prior information" for each hypothesis. We draw z-scores Z_i independently, with Z_i drawn from $N(0, 1)$ for nulls $i \in \mathcal{H}_0$ and from $N(\mu_{\text{sep}}, 1)$ for non-nulls $i \notin \mathcal{H}_0$. Here $\mu_{\text{sep}} > 0$ controls the extent of the separation between nulls and non-nulls.

2. Sort the z-scores in descending order according to magnitude: $|Z_{(1)}| > |Z_{(2)}| > \cdots >$

4. `http://www.stat.uchicago.edu/~rina/accumulationtests.html`

$|Z_{(n)}|$. Assign a new index to each hypothesis, according to its position in the sorted list.

3. Now we generate p-values for each hypothesis. We draw new z-scores Z_i^* independently for each hypothesis, with $Z_i^* \sim N(0,1)$ for nulls $i \in \mathcal{H}_0$ and $Z_i^* \sim N(\mu_{\text{sig}}, 1)$ for non-nulls $i \notin \mathcal{H}_0$. Here $\mu_{\text{sig}} > 0$ controls the strength of the true signals. Then we calculate p-values with a two-tailed z-test. Note that these p-values are independent from the process of ranking the hypotheses.

The ranking (and separation) of nulls and non-nulls is determined in steps 1 and 2, and controlled by $\mu_{\text{sep}} \in \{2, 3\}$. Larger μ_{sep} leads to better separation between nulls and non-nulls. The strength of non-null signals is specified in step 3, and controlled by $\mu_{\text{sig}} \in \{2, 3\}$. For settings with good separation of nulls and non-nulls and with strong signals, it is easier to achieve high power while keeping FDR controlled.

Under each simulation setting, we compare the performance of the four selected accumulation test methods. In each trial, the power and FDP for each accumulation function and rejection rule are recorded, over a range of target FDR levels $\alpha \in \{0.05, 0.075, \ldots, 0.25\}$. The performance results are averaged over 100 trials.

Results

Figure 2.2 shows the average power and average observed FDR of the four selected accumulation test methods, plotted against the target FDR level α. In the weak signal regime ($\mu_{\text{sig}} = 2$), the methods are mostly conservative in terms of FDR, with an observed FDR that is lower than the target level α, and thus power is low—this is expected, since the non-null p-values contribute to the estimated false discovery proportion. In contrast, with strong signal ($\mu_{\text{sig}} = 3$), the observed FDR levels are closer to α, and in fact HingeExp has slightly higher FDR than desired when separation is poor ($\mu_{\text{sep}} = 2$)—this is not unexpected, since HingeExp only guarantees the control of modified FDR (see Lemma 1). The

power of the methods improves with stronger signal (larger μ_{sig}) and with better separation (larger μ_{sep}). Across all settings, HingeExp consistently gives the highest average power and observed FDR, while SeqStep+ is generally the most conservative, with lowest average power and observed FDR.

We compare the estimated false discovery proportion along the list of hypotheses, $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ for $k = 1, \ldots, n$, for each of the four methods, and compare with the actual false discovery proportion, $\mathsf{FDP}(k)$. (For the SeqStep+ method we define $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ to agree with the method definition (2.4).) Figure 2.3 shows the results, averaged over 100 simulations. For settings with stronger signals (i.e. $\mu_{\text{sig}} = 3$), $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ is a good estimate of $\mathsf{FDP}(k)$, while for settings with weak signals (e.g. $\mu_{\text{sig}} = 2$), $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ overestimates $\mathsf{FDP}(k)$, as expected. Comparing across methods, the HingeExp method function yields the estimate $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ that approximates the actual FDP best.

2.5.2 Simulations for the least angle regression (LARS) path

Inference for high dimensional regression has been a problem of wide interest in many modern applications. Recently, Fithian et al. [2015] proposed the selective sequential test method for inference of LARS (closely related to the commonly used LASSO method for penalized sparse regression Tibshirani [1996]), which gives a p-value for each feature in the order of being selected in the LARS path. The ith p-value, p_i, is distributed as $\mathsf{Uniform}[0, 1]$, under the null hypothesis that all true signals are included in the active set at step $(i - 1)$; furthermore, the p-values are independent. This provides an ordered list of p-values that follows the assumptions of ordered hypothesis testing problem, and therefore, can be treated with the accumulation method. As the sequence corresponds to signal and noisy features in the LARS path, the test provides a stopping rule for LARS with guarantee on FDR level (here $\mathsf{FDP}(k)$ is the proportion of noise among all features included up to the kth LARS step). In this simulation, we compare the performance of the SeqStep (with parameter $C = 2$), SeqStep+ (with $C = 2$), ForwardStop, and HingeExp (with $C = 2$) accumulation

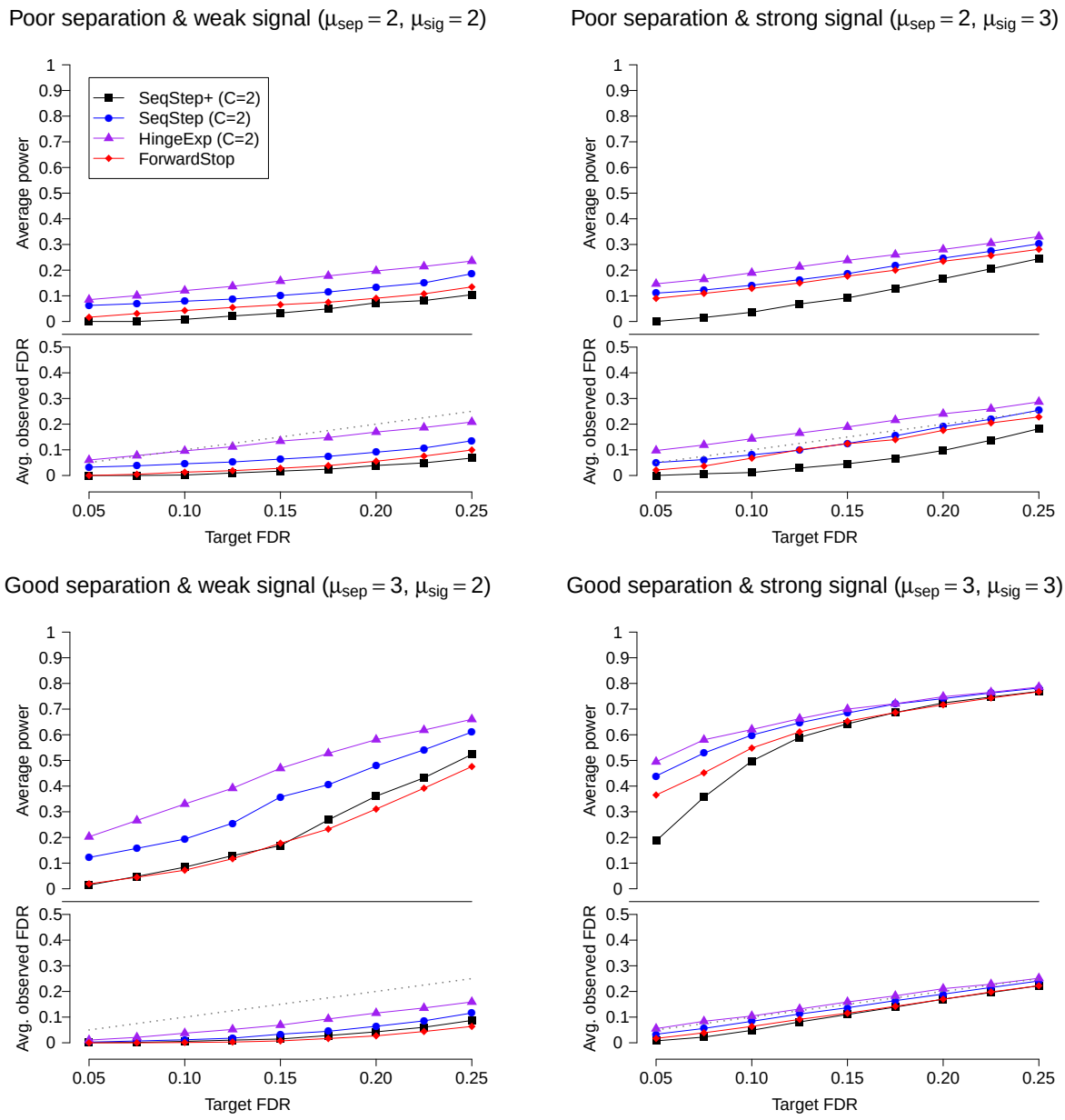

Figure 2.2: Power and observed FDR level of the SeqStep, SeqStep+, ForwardStop, and HingeExp methods, plotted against target FDR level α (averaged over 100 trials).

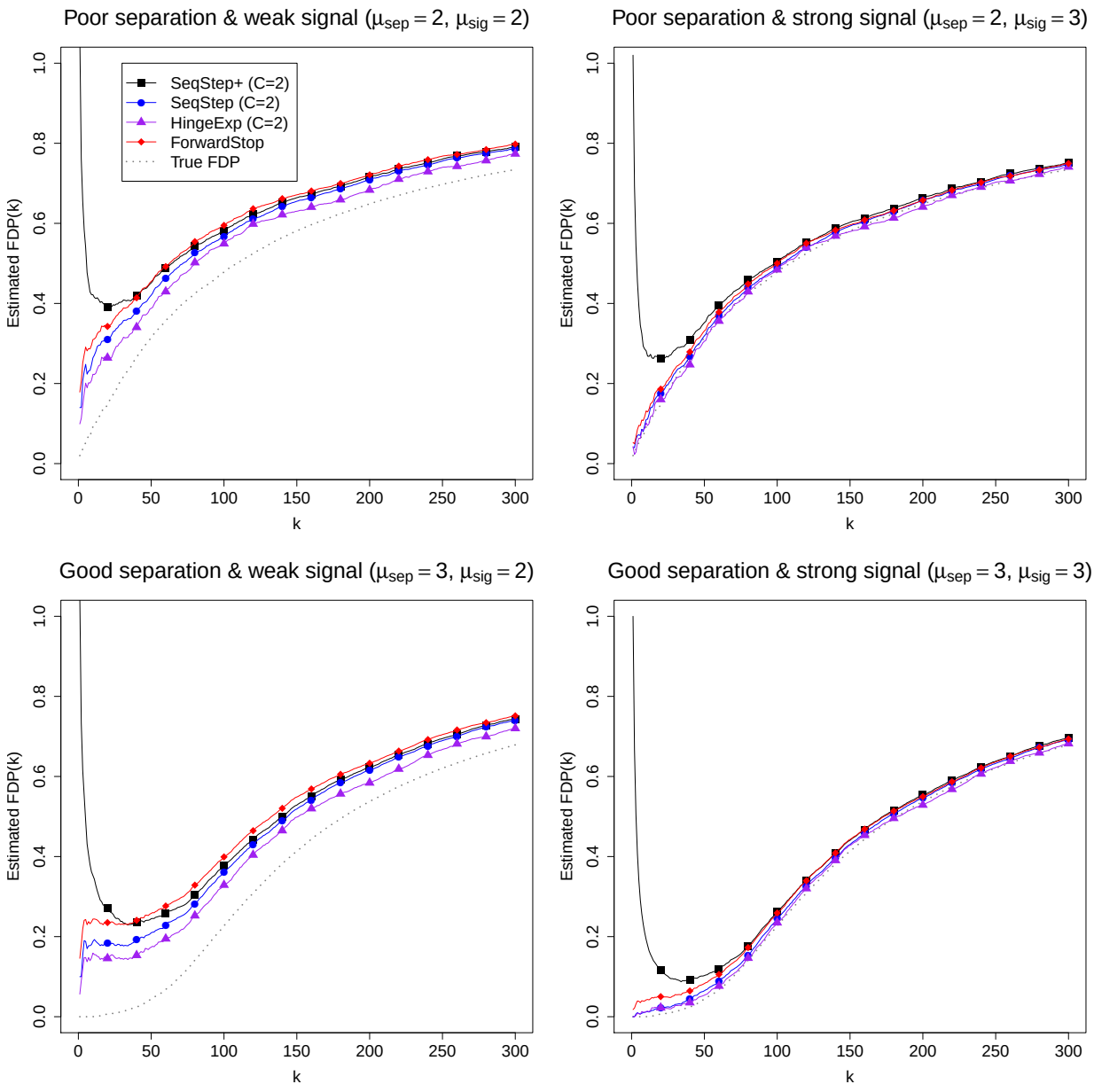

Figure 2.3: Estimated FDP with the SeqStep, SeqStep+, ForwardStop, and HingeExp methods, plotted against the true FDP, across $k = 1, \ldots, p$ (results are averaged over 100 trials).

test methods, under three settings of feature signal strength.

Methods

In all three simulation settings, there are $N = 200$ observations on $p = 100$ features, of which either $k^\star = 10$ or $k^\star = 20$ are true signals (nonzero coefficients). The design matrix X consists of i.i.d. standard normal entries. The nonzero signals β_j, for features $j = 1, \ldots, k^\star$, are taken to be equally spaced value from $2 \cdot \gamma$ to $\sqrt{2\log(p)} \cdot \gamma$, where $\gamma = 1, 5, 9$ in the three settings. This forms a gradient from weak signal to strong signal scenarios. The remaining coefficients are set as $\beta_{k^\star+1} = \cdots = \beta_p = 0$. The response is then generated as $y = X\beta + \epsilon$, where the entries of ϵ are also i.i.d. standard normal variables. Given the simulated data X, y, the LARS method and the spacing test are applied, yielding p-values for ordered testing. Note that the number of hypotheses is now given by p (one for each feature), rather than the former notation n.

Results

Figure 2.4 shows the average power and observed FDR of the four accumulation tests, averaged over 50 trials. When $\gamma = 1, 5$, all four methods successfully control FDR, and when $\gamma = 9$, SeqStep, SeqStep+ and ForwardStop control FDR well, while HingeExp slightly exceed the target FDR level α. In all settings, HingeExp ($C = 2$) attains the highest average power and FDR, while SeqStep+ (with $C = 2$) is extremely conservative for lower α values (due to the fact that, with few true signals, few discoveries are made overall, so the slightly conservative correction in this method (2.4) has a large effect).

We plot the estimated false discovery proportion, $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)$ over the first k steps of the LARS path ($k = 1, \ldots, p$), against the actual $\mathsf{FDP}(k)$. Figure 2.5 shows the results, averaged over 50 simulations. As expected, the estimated FDP levels increasingly overestimate the true FDP as signal strength γ decreases. SeqStep+ is quite conservative due to the correction term in the method's definition, while the other three methods show no consistent trend in

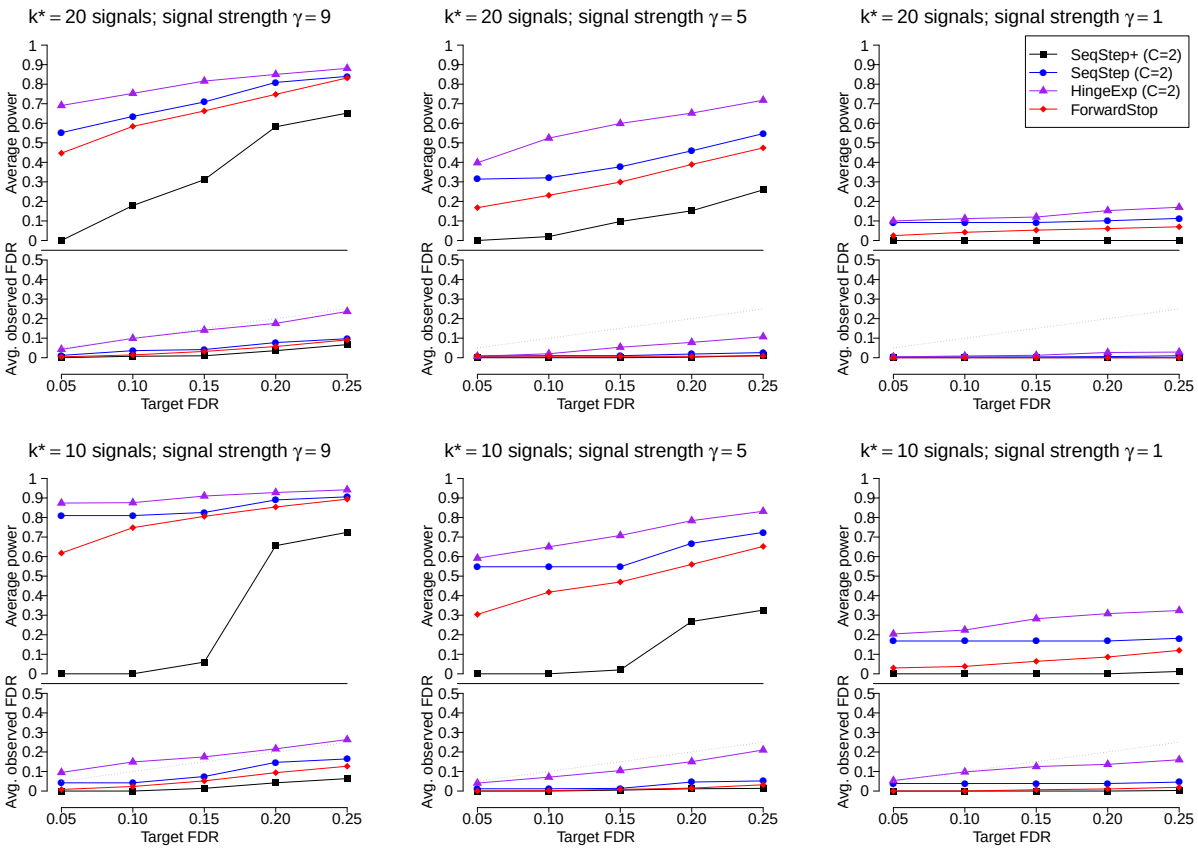

Figure 2.4: Power and observed FDR level of the SeqStep, SeqStep+, ForwardStop, and HingeExp methods for the LARS path (averaged over 50 trials).

terms of accurate estimation of $\mathsf{FDP}(k)$.

2.6 Application to dosage response data

We now show an application of our methods to the problem of identifying effects of drug dosage on gene expression levels. Code to reproduce this real data experiment is available online.[5]

Suppose that gene expression levels for genes $i = 1, \ldots, n$ are measured in $m = m_{\mathsf{C}} + m_{\mathsf{L}} + m_{\mathsf{H}}$ independent trials, where the trials $\{1, \ldots, m\}$ are partitioned into three sets $T_{\mathsf{C}} = \{1, \ldots, m_{\mathsf{C}}\}$, $T_{\mathsf{L}} = \{m_{\mathsf{C}}+1, \ldots m_{\mathsf{C}}+m_{\mathsf{L}}\}$, and $T_{\mathsf{H}} = \{m_{\mathsf{C}}+m_{\mathsf{L}}+1, \ldots, m_{\mathsf{C}}+m_{\mathsf{L}}+m_{\mathsf{H}}\}$, such that:

5. `http://www.stat.uchicago.edu/~rina/accumulationtests.html`

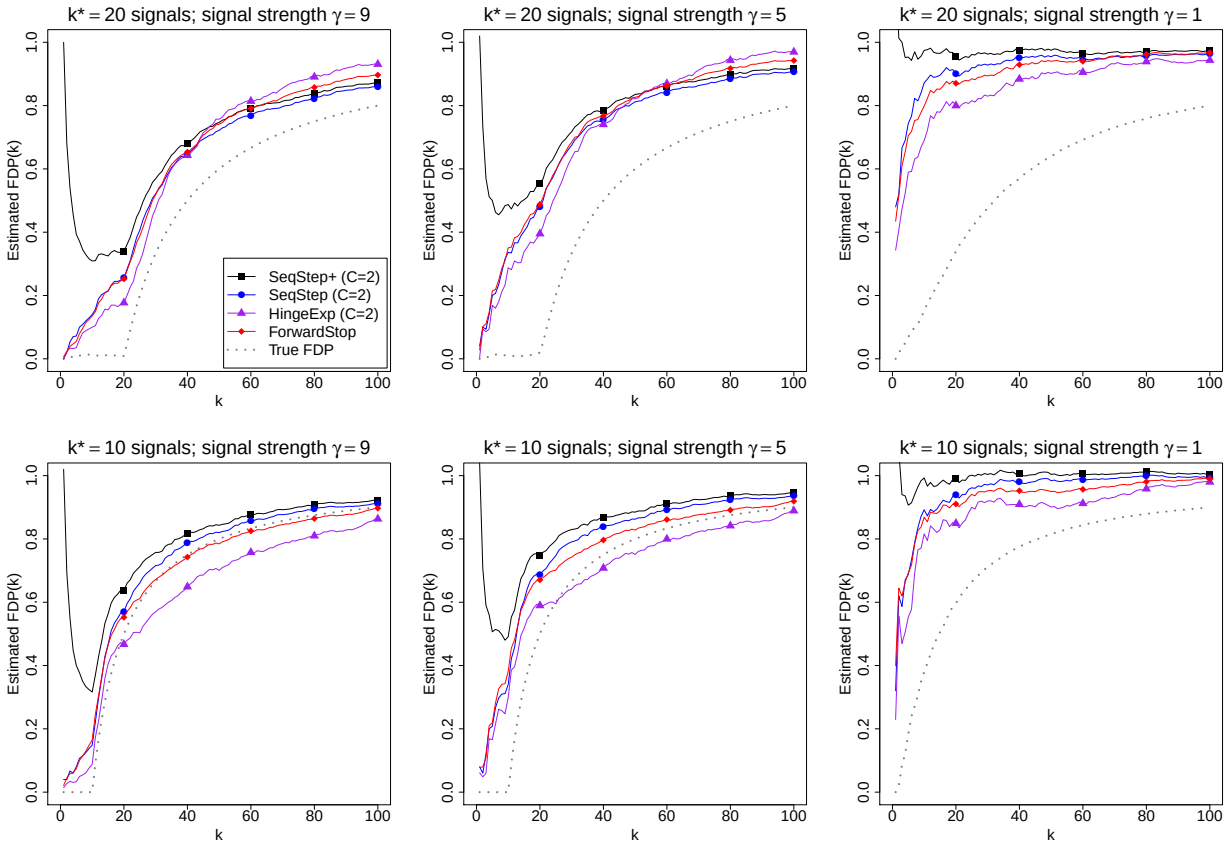

Figure 2.5: Estimated FDP with the SeqStep, SeqStep+, ForwardStop, and HingeExp methods for the LARS path, plotted against the true FDP, across $k = 1, \ldots, p$ (results are averaged over 50 trials).

- For $j \in T_{\mathsf{C}}$, trial j is carried out in the absence of the drug (control group);
- For $j \in T_{\mathsf{L}}$, trial j is carried out under a low drug dosage; and
- For $j \in T_{\mathsf{H}}$, trial j is carried out under a high drug dosage.

The variable X_{ij} gives the logarithm of the expression level of gene i in trial j.

Identifying genes that respond to the higher dosage will be easier than at the lower dosage, since the magnitude of the response will often depend on the dosage used. To be specific, for each gene i we are interested in testing the null hypothesis H_i, which states that the observations $\{X_{ij} : j \in T_{\mathsf{C}} \cup T_{\mathsf{L}}\}$ are i.i.d. (in other words, the low dose has no effect on the distribution for the gene expression level for gene i). For simplicity in the discussion and analysis below, we treat the measurements for each gene as though they were independent—of course, this does not hold in practice, and in future work we hope to develop results on FDR control of the accumulation tests under p-value dependence. For the present experiment, each of the methods we compare here comes with theoretical guarantees of FDR control only under the independence assumption. While several existing methods for multiple testing yield guarantees for FDR control even in the case of dependent p-values, the methods we are aware of are quite conservative and yield nearly zero power in this experiment. Specifically, Benjamini and Yekutieli's modification Benjamini and Yekutieli [2001] of the BH method, the Holm-Bonferroni method Holm [1979], and the Bonferroni correction each yielded, even at target FDR level $\alpha = 0.9$, no more than two discoveries for this gene expression experiment; in contrast, the accumulation test methods examined here yield thousands of discoveries, as we will see below.

Converting to an ordered hypothesis testing problem We now show how detecting differential gene expression levels can be converted to an ordered hypothesis testing problem, by making use of the high-dosage data.

For two sets of observations A and B, define $\mathsf{Pval}(A, B)$ to be the p-value produced by a two-sided two-sample t-test comparing the observations in A with the observations in B.

Define $\mathsf{Pval}_+(A, B)$ and $\mathsf{Pval}_-(A, B)$ analogously for one-sided two-sample t-tests, where $\mathsf{Pval}_+(A, B)$ tests for evidence that the mean of A's population is larger than the mean of B's population, and $\mathsf{Pval}_-(A, B)$ does the reverse.

We now follow these steps to reformulate the dosage/gene expression problem as an ordered hypothesis testing problem:

1. For each gene i, calculate p_i^{high} as

$$p_i^{\mathsf{high}} = \mathsf{Pval}\left(\{X_{ij} : j \in T_{\mathsf{H}}\}, \{X_{ij} : j \in T_{\mathsf{C}} \cup T_{\mathsf{L}}\}\right) .$$

   Record also $s_i \in \{+, -\}$, the sign of the estimated effect, i.e.

$$s_i = \text{sign}\left(\frac{1}{m_{\mathsf{H}}} \sum_{j \in T_{\mathsf{H}}} X_{ij} - \frac{1}{m_{\mathsf{C}} + m_{\mathsf{L}}} \sum_{j \in T_{\mathsf{C}} \cup T_{\mathsf{L}}} X_{ij}\right) .$$

2. We then use these high-dosage p-values to relabel the n genes, that is, reorder the genes so that

$$p_1^{\mathsf{high}} \leq p_2^{\mathsf{high}} \leq \cdots \leq p_n^{\mathsf{high}} .$$

3. Next, for each gene i, compute an initial p-value by comparing the low-dosage trials with the control trials. We use a one-sided two-sample t-test (determined by the sign s_i):

$$p_i^{\mathsf{init}} = \mathsf{Pval}_{s_i}\left(\{X_{ij} : j \in T_{\mathsf{L}}\}, \{X_{ij} : j \in T_{\mathsf{C}}\}\right) .$$

   We use a one-sided t-test because, if for instance we observe a positive response at the high dosage for gene i, then we are much more likely to see a positive (rather than negative) effect at the low dosage, as well. Therefore, a one-sided t-test is likely to achieve higher power than a two-sided test.

4. Now we transform to the final p-values using a permutation test. For each gene i, for

every permutation π on the trial labels $\{1, \ldots, m_{\mathsf{C}} + m_{\mathsf{L}}\}$, compute

$$p_i^{\pi} = \mathsf{Pval}_{s_i}\left(\{X_{i\pi(j)} : j \in T_{\mathsf{L}}\}, \{X_{i\pi(j)} : j \in T_{\mathsf{C}}\}\right) .$$

We then calculate the final p-value by comparing p_i^{init} with the empirical distribution $\{p_i^{\pi} : \text{all possible permutations } \pi\}$:

$$p_i = \frac{\#\{\text{Permutations } \pi \ : \ p_i^{\mathsf{init}} \leq p_i^{\pi}\}}{(m_{\mathsf{C}} + m_{\mathsf{L}})!} , \tag{2.19}$$

and perform the accumulation test on this sequence of p-values.

In fact, since p_i^{π} depends only on the partition of the $m_{\mathsf{C}} + m_{\mathsf{L}}$ many trial labels into two groups of size m_{C} and m_{L} (the control group and the low-dose group), we only need to calculate p_i^{π} for $\binom{m_{\mathsf{C}}+m_{\mathsf{L}}}{m_{\mathsf{C}}}$ many permutations.

Note that p_i^{high} and s_i depend on $\{X_{ij} : j \in T_{\mathsf{C}} \cup T_{\mathsf{L}}\}$, so we cannot use the t-test p-values p_i^{init} directly for the accumulation test. However, p_i^{high} and s_i are invariant to permutations of this input by definition of the two-sample t-test. In other words, even after conditioning on $(p_i^{\mathsf{high}}, s_i)$, the variables $\{X_{ij} : j \in T_{\mathsf{C}} \cup T_{\mathsf{L}}\}$ are exchangeable under the null hypothesis H_i. Therefore, even after reordering the genes according to the high-dosage p-values p_i^{high} and recording signs s_i, the final permutation test p-values p_i that we calculate are valid p-values for each true null hypothesis H_i. Our theory thus guarantees FDR control when the accumulation test is applied to these permutation test p-values (if we assume that the data for each gene is independent).

2.6.1 Empirical results

We now implement the methods described above on real data. All computations are carried out in R R Core Team [2014]. The data[6] Coser et al. [2003] measures differential expression

6. Data available at `http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2324` or via the `GEOquery` package Davis and Meltzer [2007] in R.

in response to estrogen in breast cancer cells. The data set consists of $n = 22283$ genes with 25 trials, with 5 trials each for the control group and for four different dosage levels. For our experiment, we use the $m_{\mathsf{C}} = 5$ control trials, the $m_{\mathsf{L}} = 5$ trials at the lowest dosage, and the $m_{\mathsf{H}} = 5$ trials at the highest dosage.

We compare the following methods, each with target FDR level $\alpha = 0, 0.01, 0.02, \ldots, 0.90$. First, we test four accumulation tests:

- The SeqStep and SeqStep+ methods Barber and Candès [2015] with parameter $C = 2$, the HingeExp method with parameter $C = 2$, and the ForwardStop method G'Sell et al. [2015].

We compare to several methods for controlling FDR under multiple testing, which do not use an ordered structure (and have no mechanism for incorporating the high-dosage data):

- The Benjamini-Hochberg procedure Benjamini and Hochberg [1995], using p-values that compare the low dosage trials with the control trials, via either a two-sided t-test,

$$p_i^{\mathsf{ttest}} = \mathsf{Pval}\left(\{X_{ij} : j \in T_{\mathsf{L}}\}, \{X_{ij} : j \in T_{\mathsf{C}}\}\right) , \tag{2.20}$$

  or a the permutation test on these t-tests,

$$p_i^{\mathsf{perm}} = \frac{\#\{\pi \ : \ p_i^{\mathsf{ttest}} \leq p_i^{\pi}\}}{(m_{\mathsf{C}} + m_{\mathsf{L}})!} \text{ where } p_i^{\pi} = \mathsf{Pval}\left(\{X_{i\pi(j)} : j \in T_{\mathsf{L}}\}, \{X_{i\pi(j)} : j \in T_{\mathsf{C}}\}\right) . \tag{2.21}$$

- Storey [2002]'s modification of the Benjamini-Hochberg procedure, applied to either the t-test p-values (2.20) or the permutation test p-values (2.21). We estimate the number of true nulls as $\widehat{m}_0 = 10 \cdot \#\{i : p_i^{\mathsf{ttest}} > 0.9\}$ or $\widehat{m}_0 = 10 \cdot \#\{i : p_i^{\mathsf{perm}} > 0.9\}$, respectively, for the two types of p-values.

For the various methods, Figure 2.6 displays the number of discoveries against the target FDR level α. We see that the accumulation tests far outperform the Benjamini-Hochberg

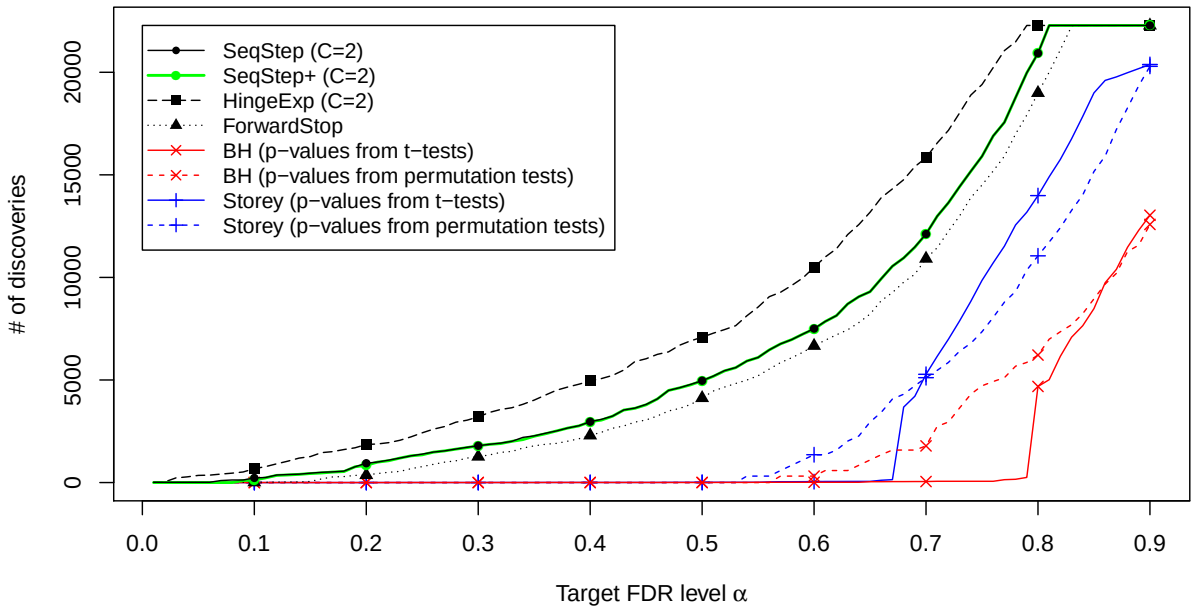

Figure 2.6: Results for the differential gene expression experiment: for each method, the plot shows the number of discoveries made (i.e. the number of genes selected as showing significant change in expression at the low drug dosage), at a range of target FDR values α. Note that the SeqStep and SeqStep+ methods are nearly indistinguishable in the plot.

and Storey procedures. At target FDR levels $\alpha \leq 0.5$, the Benjamini-Hochberg and Storey methods are unable to make more than a few discoveries, while the accumulation tests produce many discoveries.

Comparing the accumulation tests that are studied here, the SeqStep (with $C = 2$) and SeqStep+ (with $C = 2$) procedures are almost identical, showing that when the number of discoveries is high, the slight correction in the definition of the SeqStep+ method (2.4) has essentially no loss of power relative to SeqStep. The HingeExp method (with $C = 2$) yields substantially more discoveries than SeqStep (with $C = 2$) and SeqStep+ (with $C = 2$), which in turn yield more discoveries than ForwardStop.

Overall, the comparisons across the different accumulation tests examined here confirms the higher power attained by HingeExp compared to existing methods in the family (when HingeExp and SeqStep have the same C in their parameterizations). We also see substantial

power gain of the accumulation tests as compared to the Benjamini-Hochberg and Storey procedures, both of which do not use a sequential structure and do not make use of the high-dosage data, demonstrating the benefits of the ordered hypothesis testing approach.

2.7 Proofs and technical details

In this section, we prove all the theoretical results stated in the main paper.

2.7.1 Proof of Theorem 1 (finite-sample FDP control)

We begin by stating a preliminary lemma on maximal values for random walks:

Lemma 3. *Let $X_1, X_2, \ldots$ be independent random variables, and such that X_i is (σ^2, b)-subexponential (as defined in (2.7)) for all i. Then for any $\epsilon \in (0,1)$, with probability at least $1-\epsilon$, for all $t \geq 1$,*

$$\left| \sum_{i=1}^{t} X_i - \mathbb{E}\left[X_i\right] \right| \leq \max\left\{\sigma, b\sqrt{3\log\left(8/\epsilon\right)}\right\} \cdot \sqrt{16t \log\left(\frac{8\log(2t)}{\epsilon}\right)}\,.$$

This lemma is proved later in this section. With this result in place, we turn to the proof of Theorem 1.

Proof of Theorem 1. First, for any k, the sum $\sum_{i \leq k, i \in \mathcal{H}_0} \mathsf{h}(p_i)$ is a sum of $\mathsf{FalsePos}(k) \leq k$ many independent and (σ^2, b)-subexponential terms. By Lemma 3, then, with probability at least $1-\epsilon$, for all $k = 1, \ldots, n$,

$$\left| \sum_{i \leq k, i \in \mathcal{H}_0} \mathsf{h}(p_i) - \mathbb{E}\left[\mathsf{h}(p_i)\right] \right| \leq \max\left\{\sigma, b\sqrt{3\log\left(8/\epsilon\right)}\right\} \cdot \sqrt{16k \log\left(\frac{8\log(2k)}{\epsilon}\right)}\,.$$

From this point on, assume that this event holds. Since $p_i \geq \mathsf{Uniform}[0,1]$ for all $i \in \mathcal{H}_0$, by assumption on h we have $\mathbb{E}\left[\mathsf{h}(p_i)\right] \geq 1$ for all $i \in \mathcal{H}_0$, and so this implies that, for all

$k = 1, \ldots, n$,

$$\sum_{i \le k, i \in \mathcal{H}_0} \mathsf{h}(p_i) \ge \mathsf{FalsePos}(k) - \max\left\{\sigma, b\sqrt{3\log(8/\epsilon)}\right\} \cdot \sqrt{16k \log\left(\frac{8\log(2k)}{\epsilon}\right)}.$$

In particular, if $\widehat{k}_{\mathsf{h}} \neq 0$, this implies that

$$\sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i) \ge \mathsf{FalsePos}(\widehat{k}_{\mathsf{h}}) - \max\left\{\sigma, b\sqrt{3\log(8/\epsilon)}\right\} \cdot \sqrt{16\widehat{k}_{\mathsf{h}} \log\left(\frac{8\log(2\widehat{k}_{\mathsf{h}})}{\epsilon}\right)}.$$

Furthermore, by definition of $\widehat{k}_{\mathsf{h}}$, we know that

$$\frac{\sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i)}{\widehat{k}_{\mathsf{h}}} \le \frac{\sum_{i \le \widehat{k}_{\mathsf{h}}} \mathsf{h}(p_i)}{\widehat{k}_{\mathsf{h}}} \le \alpha,$$

proving that

$$\mathsf{FalsePos}(\widehat{k}_{\mathsf{h}}) - \max\left\{\sigma, b\sqrt{3\log(8/\epsilon)}\right\} \cdot \sqrt{16\widehat{k}_{\mathsf{h}} \log\left(\frac{8\log(2\widehat{k}_{\mathsf{h}})}{\epsilon}\right)} \le \alpha\widehat{k}_{\mathsf{h}}$$

or, equivalently,

$$\mathsf{FDP}(\widehat{k}_{\mathsf{h}}) \le \alpha + \max\left\{\sigma, b\sqrt{3\log(8/\epsilon)}\right\} \cdot \sqrt{\frac{16\log\left(\frac{8\log(2\widehat{k}_{\mathsf{h}})}{\epsilon}\right)}{\widehat{k}_{\mathsf{h}}}}.$$

Now we turn to $\widehat{k}_{\mathsf{h}}^{+C*}$. If this cutoff is nonzero, then by definition,

$$\frac{C\sqrt{16\widehat{k}_{\mathsf{h}}^{+C*} \log\left(\frac{8\log(2\widehat{k}_{\mathsf{h}}^{+C*})}{\epsilon}\right)} + \sum_{i \le \widehat{k}_{\mathsf{h}}^{+C*}, i \in \mathcal{H}_0} \mathsf{h}(p_i)}{\widehat{k}_{\mathsf{h}}^{+C*}} \le \alpha.$$

Furthermore, as above, we have

$$\sum_{i \leq \widehat{k}_{\mathsf{h}}^{+C*}, i \in \mathcal{H}_0} \mathsf{h}(p_i) \geq \mathsf{FalsePos}(\widehat{k}_{\mathsf{h}}^{+C*}) - \max\left\{\sigma, b\sqrt{3\log(8/\epsilon)}\right\} \cdot \sqrt{16\widehat{k}_{\mathsf{h}}^{+C*} \log\left(\frac{8\log(2\widehat{k}_{\mathsf{h}}^{+C*})}{\epsilon}\right)}.$$

Combining these two bounds, we see that

$$\mathsf{FDP}(\widehat{k}_{\mathsf{h}}^{+C*}) \leq \alpha,$$

as long as we choose $C \geq \max\left\{\sigma, b\sqrt{3\log(8/\epsilon)}\right\}$. □

Proof of Lemma 3 (bounds on random walks)

Proof of Lemma 3. Our proof follows the main steps laid out in the proof of the finite sample Law of Iterated Logarithms from Jamieson et al. [2014]:

- Step 1: Prove that $\left|\sum_{i=1}^{2^k} X_i - \mathbb{E}[X_i]\right| \lesssim \sqrt{2^k \log(\log(2^k))}$ for all $k \geq 0$, with high probability;
- Step 2: Prove that $\max_{t=1,\ldots,2^k-1} \left|\sum_{i=2^k+1}^{2^k+t} X_i - \mathbb{E}[X_i]\right| \lesssim \sqrt{2^k \log(\log(2^k))}$ for all $k \geq 0$, with high probability;
- Step 3: Combine Steps 1 and 2 to prove the lemma.

Step 1:

Without loss of generality, assume $\mathbb{E}[X_i] = 0$ for all i. Let $\tilde{\sigma} = \max\{\sigma, b\sqrt{3\log\left(\frac{8}{\epsilon}\right)}\}$. Since $\tilde{\sigma} \geq \sigma$, trivially each X_i is $(\tilde{\sigma}, b)$-subexponential. Then for all $\theta \in [0, \frac{1}{b}]$ and all $i \geq 1$,

$$\mathbb{E}\left[e^{\theta X_i}\right] \leq \exp\left\{\frac{\theta^2 \tilde{\sigma}^2}{2}\right\}.$$

Now taking any $k \geq 0$, and any $r > 0$ such that $\theta = \frac{r}{2^k \tilde{\sigma}^2} \leq \frac{1}{b}$,

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i=1}^{2^k} X_i \geq r\right\} &= \mathbb{P}\left\{\theta \sum_{i=1}^{2^k} X_i \geq \theta r\right\} \leq \mathbb{P}\left\{e^{\theta \sum_{i=1}^{2^k} X_i} \geq e^{\theta r}\right\} \\
&\leq \mathbb{E}\left[e^{\theta \sum_{i=1}^{2^k} X_i}\right] \cdot e^{-\theta r} \quad \text{(Markov inequality)} \\
&\leq \exp\left\{\frac{2^k \theta^2 \tilde{\sigma}^2}{2}\right\} \cdot e^{-\theta r} \\
&= \exp\left\{-\frac{r^2}{2^{k+1}\tilde{\sigma}^2}\right\} \quad \text{by taking } \theta = \frac{r}{2^k \tilde{\sigma}^2} \leq \frac{1}{b}\,.
\end{aligned}$$

By an identical argument, the same bound holds for $\mathbb{P}\left\{\sum_{i=1}^{2^k} X_i \leq -r\right\}$, and therefore, for all $k \geq 0$ and $r \leq \frac{2^k \tilde{\sigma}^2}{b}$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^{2^k} X_i\right| \geq r\right\} \leq 2\exp\left\{-\frac{r^2}{2^{k+1}\tilde{\sigma}^2}\right\}\,.$$

Setting $r = \tilde{\sigma}\sqrt{2^{k+1}\log\left(\frac{8(k+1)^2}{\epsilon}\right)}$, we can check that $\frac{r}{2^k\tilde{\sigma}^2} \leq \frac{1}{b}$ for all $k \geq 0$ and all $\epsilon \in (0,1)$ (by definition of $\tilde{\sigma}$), and so

$$\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{i=1}^{2^k} X_i\right| \geq \tilde{\sigma}\sqrt{2^{k+1}\log\left(\frac{8(k+1)^2}{\epsilon}\right)}\right\} \\
&\qquad\qquad \leq 2\exp\left\{-\frac{\left[\tilde{\sigma}\sqrt{2^{k+1}\log\left(\frac{8(k+1)^2}{\epsilon}\right)}\right]^2}{2^{k+1}\tilde{\sigma}^2}\right\} = \frac{\epsilon}{4(k+1)^2}.
\end{aligned}$$

Taking a union bound,

$$\mathbb{P}\left\{\left|\sum_{i=1}^{2^k} X_i\right| \leq \tilde{\sigma}\sqrt{2^{k+1}\log\left(\frac{8(k+1)^2}{\epsilon}\right)} \text{ for all } k \geq 0\right\} \geq 1 - \sum_{k\geq 0}\frac{\epsilon}{4(k+1)^2} \geq 1 - \frac{\epsilon}{2}.$$

Step 2:

For $k = 0$ the statement is trivial. Now fix any $k \geq 1$. We will use a result from Fan et al. [2014] (their Theorem 2.1) which, specialized to our setting, shows that, for a sequence of independent variables[7] $\xi_1, \ldots, \xi_n$ satisfying $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}\left[e^{\theta\xi_i - g\xi_i^2}\right] \leq 1 + f$ for all $i = 1, \ldots, n$ and for some fixed $f, g \geq 0$, for any $x, v > 0$,

$$\mathbb{P}\left\{\exists\, t \in \{1, \ldots, n\} \text{ with } \sum_{i=1}^{t} \xi_i \geq x \text{ and } \sum_{i=1}^{t} \xi_i^2 \leq v^2\right\} \leq \exp\left\{-\theta x + gv^2 + n\log(1+f)\right\}.$$

Now take $n = 2^k - 1$, and set $\xi_i = X_{2^k+i}$. Then $\mathbb{E}[\xi_i] = 0$ and, taking any θ with $0 \leq \theta \leq \frac{1}{b}$ and $g = 0$ and $f = e^{\theta^2\tilde{\sigma}^2/2} - 1$ so that

$$\mathbb{E}\left[e^{\theta\xi_i - g\xi_i^2}\right] = \mathbb{E}\left[e^{\theta\xi_i}\right] \leq e^{\theta^2\tilde{\sigma}^2/2} = 1 + f,$$

we have

$$\mathbb{P}\left\{\exists\, t \in \{1, \ldots, 2^k - 1\} \text{ with } \sum_{i=1}^{t} X_{2^k+i} \geq x \text{ and } \sum_{i=1}^{t} X_{2^k+i}^2 \leq v^2\right\} \leq \exp\left\{-\theta x + 2^k \cdot \theta^2\tilde{\sigma}^2/2\right\}.$$

Taking the limit as $v \to \infty$,

$$\mathbb{P}\left\{\max_{t=1,\ldots,2^k-1} \sum_{i=1}^{t} X_{2^k+i} \geq x\right\} \leq \exp\left\{-\theta x + 2^k \cdot \theta^2\tilde{\sigma}^2/2\right\}.$$

And, applying the same to the sequence given by $\xi_i = -X_{2^k+i}$, we see that

$$\mathbb{P}\left\{\max_{t=1,\ldots,2^k-1} \left|\sum_{i=1}^{t} X_{2^k+i}\right| \geq x\right\} \leq 2\exp\left\{-\theta x + 2^k \cdot \theta^2\tilde{\sigma}^2/2\right\}.$$

7. To convert their theorem to this statement, for their notation, define $V_i = 1$, $\lambda = \theta$, $f(\lambda) = f$, $g(\lambda) = g$, and note that requiring $\sum_{i=1}^{t} V_i \leq w$ for all $t = 1, \ldots, n$ is equivalent to setting $w = n$.

Now set

$$x = \tilde{\sigma}\sqrt{2^{k+1} \cdot \log\left(\frac{8k^2}{\epsilon}\right)} \text{ and } \theta = \frac{x}{2^k \tilde{\sigma}^2};$$

note that $0 \leq \theta \leq \frac{1}{b}$ for all $k \geq 1$ by definition of $\tilde{\sigma}$. Then we get

$$\mathbb{P}\left\{\max_{t=1,\ldots,2^k-1}\left|\sum_{i=1}^{t} X_{2^k+i}\right| \geq \tilde{\sigma}\sqrt{2^{k+1} \cdot \log\left(\frac{8k^2}{\epsilon}\right)}\right\} \leq 2\exp\left\{-\log\left(\frac{8k^2}{\epsilon}\right)\right\} = \frac{\epsilon}{4k^2}.$$

Finally, we can see that this statement holds for all $k \geq 1$ with probability at least

$$1 - \sum_{k\geq 1}\frac{\epsilon}{4k^2} \geq 1 - \frac{\epsilon}{2}.$$

Step 3:

With probability at least $1-\epsilon$, the results from Step 1 and Step 2 both hold for all $k \geq 0$. Now putting the above results together, for any integer $t \geq 1$, setting $k = \lfloor \log_2(t) \rfloor$ so that $2^k \leq t < 2^{k+1}$,

$$\left|\sum_{i=1}^{t} X_i\right| \leq \left|\sum_{i=1}^{2^k} X_i\right| + \left|\sum_{i=1}^{t-2^k} X_{2^k+i}\right| \leq \tilde{\sigma}\sqrt{2^{k+1}\log\left(\frac{8(k+1)^2}{\epsilon}\right)} + \tilde{\sigma}\sqrt{2^{k+1} \cdot \log\left(\frac{8k^2}{\epsilon}\right)}.$$

Since $2^{k+1} \leq 2t$ and $k+1 \leq \log_2(2t) \leq 2\log(2t)$, we can weaken this bound to

$$\left|\sum_{i=1}^{t} X_i\right| \leq \tilde{\sigma}\sqrt{8t\log\left(\frac{32\log^2(2t)}{\epsilon}\right)} \leq \tilde{\sigma}\sqrt{16t\log\left(\frac{8\log(2t)}{\epsilon}\right)},$$

which concludes the proof. □

2.7.2 Proof of Theorem 2 (finite-sample FDR control)

The key ingredient for the proof of Theorem 2 is the following lemma, proved below:

Lemma 4. *Let $a_1, \ldots, a_n \geq 0$ be any fixed thresholds, and let*

$$\widehat{k} = \max \left\{ k \in \{1, \ldots, n\} : \sum_{i=1}^{k} \mathsf{h}(p_i) \leq a_k \right\} , \tag{2.22}$$

with the convention that $\widehat{k} = 0$ if this set is empty. Then, under the assumptions of Theorem 2,

$$\mathbb{E}\left[\frac{1 + \#\{i \leq \widehat{k} : i \in \mathcal{H}_0\}}{C + \sum_{i \leq \widehat{k}, i \in \mathcal{H}_0} \mathsf{h}(p_i)} \right] \leq \frac{1}{\int_{t=0}^{1} \mathsf{h}(t) \wedge C \, \mathsf{d}t} .$$

To understand the role of this result in proving Theorem 2, first note that the definitions of $\widehat{k}_{\mathsf{h}}$ and $\widehat{k}_{\mathsf{h}}^{+C}$, given in (2.12) and (2.13), can each be rewritten as a threshold criterion of the form (2.22) as given in Lemma 4.

Essentially, Lemma 4 shows that, at $k = \widehat{k}_{\mathsf{h}}$ (or at $k = \widehat{k}_{\mathsf{h}}^{+C}$), we have $\sum_{i=1}^{k} \mathsf{h}(p_i) \gtrapprox \#\{i \leq k : i \in \mathcal{H}_0\}$, and thus, this result guarantees that the estimated FDP, $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) = \frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k}$, is a reliable (over)estimate of the actual FDP, $\mathsf{FDP}(k) = \frac{\#\{i \leq k : i \in \mathcal{H}_0\}}{k}$. Given this lemma, the proof of the bounds in Theorem 2 follows the arguments in Barber and Candès [2015, Theorem 3], as follows.

Proof of Theorem 2. First, note that the result (2.14) for bounded accumulation functions is simply a special case of the general result (2.15), since if the accumulation function h is bounded by C then

$$\int_{t=0}^{1} \mathsf{h}(t) \wedge C \, \mathsf{d}t = \int_{t=0}^{1} \mathsf{h}(t) \, \mathsf{d}t = 1 .$$

Therefore it suffices to prove (2.15). For the first bound in (2.15), treating $\mathsf{FDP}(\widehat{k}_{\mathsf{h}}^{+C})$, we

have

$$\begin{aligned}
\mathbb{E}\left[\mathsf{FDP}(\widehat{k}_{\mathsf{h}}^{+C})\right] &= \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}}^{+C} : i \in \mathcal{H}_0\}}{\widehat{k}_{\mathsf{h}}^{+C}} \cdot \mathbb{1}\left\{\widehat{k}_{\mathsf{h}}^{+C} > 0\right\}\right] \\
&\le \mathbb{E}\left[\frac{1 + \#\{i \le \widehat{k}_{\mathsf{h}}^{+C} : i \in \mathcal{H}_0\}}{1 + \widehat{k}_{\mathsf{h}}^{+C}}\right] \\
&= \mathbb{E}\left[\frac{1 + \#\{i \le \widehat{k}_{\mathsf{h}}^{+C} : i \in \mathcal{H}_0\}}{C + \sum_{i=1}^{\widehat{k}_{\mathsf{h}}^{+C}} \mathsf{h}(p_i)} \cdot \frac{C + \sum_{i=1}^{\widehat{k}_{\mathsf{h}}^{+C}} \mathsf{h}(p_i)}{1 + \widehat{k}_{\mathsf{h}}^{+C}}\right] \\
&\le \alpha \cdot \mathbb{E}\left[\frac{1 + \#\{i \le \widehat{k}_{\mathsf{h}}^{+C} : i \in \mathcal{H}_0\}}{C + \sum_{i=1}^{\widehat{k}_{\mathsf{h}}^{+C}} \mathsf{h}(p_i)}\right] \text{ by definition of } \widehat{k}_{\mathsf{h}}^{+C} \\
&\le \alpha \cdot \frac{1}{\int_{t=0}^{1} \mathsf{h}(t) \wedge C \, \mathsf{d}t} \text{ by Lemma 4.}
\end{aligned}$$

Turning to the first bound in (2.15), treating $\mathsf{mFDP}_{C/\alpha}(\widehat{k}_{\mathsf{h}})$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathsf{mFDP}_{C/\alpha}(\widehat{k}_{\mathsf{h}})\right] &= \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C/\alpha + \widehat{k}_{\mathsf{h}}}\right] \\
&= \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C + \sum_{i=1}^{\widehat{k}_{\mathsf{h}}} \mathsf{h}(p_i)} \cdot \frac{C + \sum_{i=1}^{\widehat{k}_{\mathsf{h}}} \mathsf{h}(p_i)}{C/\alpha + \widehat{k}_{\mathsf{h}}}\right] \\
&\le \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C + \sum_{i=1}^{\widehat{k}_{\mathsf{h}}} \mathsf{h}(p_i)} \cdot \frac{C + \widehat{k}_{\mathsf{h}} \cdot \alpha}{C/\alpha + \widehat{k}_{\mathsf{h}}}\right] \text{ by definition of } \widehat{k}_{\mathsf{h}} \\
&= \alpha \cdot \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C + \sum_{i=1}^{\widehat{k}_{\mathsf{h}}} \mathsf{h}(p_i)}\right] \\
&\le \alpha \cdot \frac{1}{\int_{t=0}^{1} \mathsf{h}(t) \wedge C \, \mathsf{d}t} \text{ by Lemma 4.}
\end{aligned}$$

□

Proof of Lemma 4

Next we turn to the proof of Lemma 4. We will use a result that treats the Bernoulli case specifically:

Lemma 5 (Adapted from [Barber and Candès, 2015, Lemma 1 in Suppl. Mat.])**.** *Let* $B_1, \dots, B_n \in \{0,1\}$ *be independent, with* $B_i \overset{\text{iid}}{\sim} \mathsf{Bernoulli}(\rho)$ *for all* $i \in \mathcal{H}_0$*. Let* $\{\mathcal{F}_k\}_{i=1,\dots,n}$ *be any filtration in reverse time (i.e.* $\mathcal{F}_k \supseteq \mathcal{F}_{k+1}$*) such that*

$$B_i \in \mathcal{F}_k \text{ for all } i \notin \mathcal{H}_0, \text{ and for all } i > k \text{ with } i \in \mathcal{H}_0, \tag{2.23}$$

$$\sum_{i \le k, i \in \mathcal{H}_0} B_i \in \mathcal{F}_k, \text{ and} \tag{2.24}$$

$$\{B_i : i \le k, i \in \mathcal{H}_0\} \text{ are exchangeable with respect to } \mathcal{F}_k, \tag{2.25}$$

for all $k = 1, \dots, n$*. Then*

$$M_k = \frac{1 + \#\{i \le k : i \in \mathcal{H}_0\}}{1 + \sum_{i \le k, i \in \mathcal{H}_0} B_i}$$

is a supermartingale (in reverse time) with respect to $\{\mathcal{F}_k\}$*, and* $\mathbb{E}[M_n] \le \frac{1}{\rho}$*.*

Since this is a modification of the result in Barber and Candès [2015], we give a proof here for completeness:

Proof of Lemma 5 (adapted from Barber and Candès [2015]). The proof that $\mathbb{E}[M_n] \le \frac{1}{\rho}$ is given in Barber and Candès [2015]. For the supermartingale, we first observe that since $\{B_i : i \le k+1, i \in \mathcal{H}_0\}$ are exchangeable with respect to $\mathcal{F}_{k+1}$, then for any k such that $k+1 \in \mathcal{H}_0$,

$$\mathbb{P}\{B_{k+1} = 1 \mid \mathcal{F}_{k+1}\} = \frac{\sum_{i \le k+1, i \in \mathcal{H}_0} B_i}{\#\{i \le k+1, i \in \mathcal{H}_0\}}.$$

We therefore have three cases:

- If $k+1 \notin \mathcal{H}_0$, then $M_k = M_{k+1}$ trivially.
- If $k+1 \in \mathcal{H}_0$ and $\sum_{i \le k+1, i \in \mathcal{H}_0} B_i = 0$, then $M_k \le M_{k+1}$ trivially.

- If $k+1 \in \mathcal{H}_0$ and $\sum_{i \le k+1, i \in \mathcal{H}_0} B_i > 0$,

$$\begin{aligned}
&\mathbb{E}\left[M_k \mid \mathcal{F}_{k+1}\right] = \mathbb{E}\left[\frac{1+\#\{i \le k : i \in \mathcal{H}_0\}}{1+\sum_{i \le k, i \in \mathcal{H}_0} B_i} \;\middle|\; \mathcal{F}_{k+1}\right] \\
&= \mathbb{E}\left[\frac{\#\{i \le k+1 : i \in \mathcal{H}_0\}}{1+\sum_{i \le k+1, i \in \mathcal{H}_0} B_i - B_{k+1}} \;\middle|\; \mathcal{F}_{k+1}\right] \\
&= \frac{\#\{i \le k+1 : i \in \mathcal{H}_0\}}{1+\sum_{i \le k+1, i \in \mathcal{H}_0} B_i - 1} \cdot \mathbb{P}\left\{B_{k+1} = 1 \mid \mathcal{F}_{k+1}\right\} \\
&\qquad + \frac{\#\{i \le k+1 : i \in \mathcal{H}_0\}}{1+\sum_{i \le k+1, i \in \mathcal{H}_0} B_i} \cdot \mathbb{P}\left\{B_{k+1} = 0 \mid \mathcal{F}_{k+1}\right\} \\
&= \frac{\#\{i \le k+1 : i \in \mathcal{H}_0\}}{1+\sum_{i \le k+1, i \in \mathcal{H}_0} B_i - 1} \cdot \frac{\sum_{i \le k+1, i \in \mathcal{H}_0} B_i}{\#\{i \le k+1, i \in \mathcal{H}_0\}} \\
&\qquad + \frac{\#\{i \le k+1 : i \in \mathcal{H}_0\}}{1+\sum_{i \le k+1, i \in \mathcal{H}_0} B_i} \cdot \left(1 - \frac{\sum_{i \le k+1, i \in \mathcal{H}_0} B_i}{\#\{i \le k+1, i \in \mathcal{H}_0\}}\right) \\
&= 1 + \frac{\#\{i \le k+1 : i \in \mathcal{H}_0\} - \sum_{i \le k+1, i \in \mathcal{H}_0} B_i}{1+\sum_{i \le k+1, i \in \mathcal{H}_0} B_i} \\
&= \frac{1+\#\{i \le k+1 : i \in \mathcal{H}_0\}}{1+\sum_{i \le k+1, i \in \mathcal{H}_0} B_i} \\
&= M_{k+1}.
\end{aligned}$$

This proves that M_k is a supermartingale, thus proving the lemma. □

Equipped with this result for the special case of Bernoulli variables, we turn to the proof of our key lemma, where we construct a coupling between the general case and the Bernoulli case.

Proof of Lemma 4. Define $V_i \overset{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$, independent from the p_i's. Define also

$$B_i = \mathbb{1}\{V_i \le \mathsf{h}(p_i)/C\} \ .$$

Write $p_{1:n}$ to denote $p_1, \ldots, p_n$. Note that, conditioning on $p_{1:n}$, B_i's have distributions

$$(B_i \mid p_{1:n}) \sim \mathsf{Bernoulli}\left(\frac{\mathsf{h}(p_i) \wedge C}{C}\right) \ . \tag{2.26}$$

Furthermore, marginally, we see that for all $i \in \mathcal{H}_0$, the B_i's are i.i.d. Bernoulli variables with

$$\mathbb{P}\{B_i = 1\} = \mathbb{E}\left[\mathbb{P}\{B_i = 1 \mid p_i\}\right] = \mathbb{E}\left[\frac{\mathsf{h}(p_i) \wedge C}{C}\right] := \rho \geq \frac{\int_{t=0}^{1} \mathsf{h}(t) \wedge C}{C},$$

where the next-to-last step uses the fact that the null p-values are identically distributed, while the last step uses the fact that they are conservative while h is nondecreasing.

Next, we would like to to bound $\mathbb{E}\left[\frac{1+\#\{i \leq \widehat{k} : i \in \mathcal{H}_0\}}{1+\sum_{i \leq \widehat{k}, i \in \mathcal{H}_0} B_i}\right]$. Let $\{\mathcal{F}_k\}$ be a filtration with $\mathcal{F}_1 \supseteq \cdots \supseteq \mathcal{F}_n$, where $\mathcal{F}_k$ is defined by knowing

$$\begin{cases} (p_i, V_i) \text{ for each } i > k \text{ or } i \notin \mathcal{H}_0, \\ \text{the unordered set } \{(p_i, V_i) : i \leq k, i \in \mathcal{H}_0\}. \end{cases}$$

Note that, since the null p-values are i.i.d., this means that the variables $\{B_i : i \leq k, i \in \mathcal{H}_0\}$ are exchangeable with respect to the σ-algebra $\mathcal{F}_k$. Therefore $\{\mathcal{F}_k\}$ satisfies the conditions of Lemma 5. Since $\widehat{k}$ is a stopping time (in reverse time) with respect to $\{\mathcal{F}_k\}$, we can apply Lemma 5 together with the Optional Stopping Time theorem to prove that

$$\mathbb{E}\left[\frac{1 + \#\{i \leq \widehat{k} : i \in \mathcal{H}_0\}}{1 + \sum_{i \leq \widehat{k}, i \in \mathcal{H}_0} B_i}\right] \leq \frac{1}{\rho} \leq \frac{1}{\frac{1}{C}\int_{t=0}^{1} \mathsf{h}(t) \wedge C \, \mathsf{d}t}. \tag{2.27}$$

Next, we calculate

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1+\#\{i\le\widehat{k}:i\in\mathcal{H}_0\}}{1+\sum_{i\le\widehat{k},i\in\mathcal{H}_0}B_i}\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\frac{1+\#\{i\le\widehat{k}:i\in\mathcal{H}_0\}}{1+\sum_{i\le\widehat{k},i\in\mathcal{H}_0}B_i}\;\middle|\;p_{1:n}\right]\right]\quad\text{by the tower rule of expectations}\\
&=\mathbb{E}\left[(1+\#\{i\le\widehat{k}:i\in\mathcal{H}_0\})\cdot\mathbb{E}\left[\frac{1}{1+\sum_{i\le\widehat{k},i\in\mathcal{H}_0}B_i}\;\middle|\;p_{1:n}\right]\right]\quad\text{as }\widehat{k}\text{ is a function of }p_{1:n}\\
&\ge\mathbb{E}\left[(1+\#\{i\le\widehat{k}:i\in\mathcal{H}_0\})\cdot\frac{1}{\mathbb{E}\left[1+\sum_{i\le\widehat{k},i\in\mathcal{H}_0}B_i\;\middle|\;p_{1:n}\right]}\right]\quad\text{by Jensen's inequality}\\
&=\mathbb{E}\left[(1+\#\{i\le\widehat{k}:i\in\mathcal{H}_0\})\cdot\frac{1}{1+\sum_{i\le\widehat{k},i\in\mathcal{H}_0}\frac{\mathsf{h}(p_i)\wedge C}{C}}\right]\quad\text{by (2.26)}\\
&\ge C\cdot\mathbb{E}\left[(1+\#\{i\le\widehat{k}:i\in\mathcal{H}_0\})\cdot\frac{1}{C+\sum_{i\le\widehat{k},i\in\mathcal{H}_0}\mathsf{h}(p_i)}\right].
\end{aligned}
$$

Combining this result with (2.27), we have proved the lemma. □

2.7.3 Proof of Lemma 1 (FDR control for the HingeExp function)

Proof of Lemma 1. First, by Jensen's inequality, drawing $E_{0,1}, E_{0,2} \overset{\text{iid}}{\sim} \mathsf{Exponential}(1)$ independently from the p-values $p_1, \ldots, p_n$,

$$
\begin{aligned}
\mathbb{E}\left[\frac{\#\{i\le\widehat{k}_{\mathsf{h}}:i\in\mathcal{H}_0\}}{2C\alpha^{-1}+\widehat{k}_{\mathsf{h}}}\right]&=\mathbb{E}\left[\frac{\#\{i\le\widehat{k}_{\mathsf{h}}:i\in\mathcal{H}_0\}}{C\alpha^{-1}\mathbb{E}\left[E_{0,1}+E_{0,2}\mid p_1,\ldots,p_n\right]+\widehat{k}_{\mathsf{h}}}\right]\\
&\le\mathbb{E}\left[\mathbb{E}\left[\frac{\#\{i\le\widehat{k}_{\mathsf{h}}:i\in\mathcal{H}_0\}}{C\alpha^{-1}(E_{0,1}+E_{0,2})+\widehat{k}_{\mathsf{h}}}\;\middle|\;p_1,\ldots,p_n\right]\right]\\
&=\mathbb{E}\left[\frac{\#\{i\le\widehat{k}_{\mathsf{h}}:i\in\mathcal{H}_0\}}{C\alpha^{-1}(E_{0,1}+E_{0,2})+\widehat{k}_{\mathsf{h}}}\right].
\end{aligned}
$$

Next,

$$\begin{aligned}
&\mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C\alpha^{-1}(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}}}\right] \\
&= \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C(E_{0,1}+E_{0,2}) + \sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i)} \cdot \frac{C(E_{0,1}+E_{0,2}) + \sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i)}{C\alpha^{-1}(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}}}\right] \\
&\le \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C(E_{0,1}+E_{0,2}) + \sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i)} \cdot \frac{C(E_{0,1}+E_{0,2}) + \sum_{i \le \widehat{k}_{\mathsf{h}}} \mathsf{h}(p_i)}{C\alpha^{-1}(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}}}\right] \\
&= \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C(E_{0,1}+E_{0,2}) + \sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i)} \cdot \frac{C(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}} \cdot \widehat{\mathsf{FDP}}_{\mathsf{h}}(\widehat{k}_{\mathsf{h}})}{C\alpha^{-1}(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}}}\right] \\
&\le \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C(E_{0,1}+E_{0,2}) + \sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i)} \cdot \frac{C(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}} \cdot \alpha}{C\alpha^{-1}(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}}}\right] \\
&= \alpha \cdot \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C(E_{0,1}+E_{0,2}) + \sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} \mathsf{h}(p_i)}\right] .
\end{aligned}$$

Next, note that $\mathsf{h}(p_i)$ is equal in distribution to $C \cdot B_i \cdot E_i$, where $B_i \sim \mathsf{Bernoulli}(1/C)$ and $E_i \sim \mathsf{Exponential}(1)$, for all $i \in \mathcal{H}_0$. (Here we assume that the variables B_i and E_i are all mutually independent). Therefore we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{C\alpha^{-1}(E_{0,1}+E_{0,2}) + \widehat{k}_{\mathsf{h}}}\right] \\
&\qquad\qquad \le \alpha \cdot \frac{1}{C} \cdot \mathbb{E}\left[\frac{\#\{i \le \widehat{k}_{\mathsf{h}} : i \in \mathcal{H}_0\}}{E_{0,1}+E_{0,2} + \sum_{i \le \widehat{k}_{\mathsf{h}}, i \in \mathcal{H}_0} B_i \cdot E_i}\right] \le \alpha \cdot \frac{1}{C} \cdot \mathbb{E}\left[M_{\widehat{k}_{\mathsf{h}}}\right],
\end{aligned}$$

where we define

$$M_k = \frac{1 + \#\{i \le k : i \in \mathcal{H}_0\}}{E_{0,1}+E_{0,2} + \sum_{i \le k, i \in \mathcal{H}_0} B_i \cdot E_i} .$$

Next, we prove that M_k is a supermartingale with $\mathbb{E}[M_n] \le C$, and that $\widehat{k}_{\mathsf{h}}$ is a stopping time.

Let $\mathcal{F}_k$ be the σ-algebra defined by knowing $E_{0,1}, E_{0,2}$, knowing (B_i, E_i) for all $i \notin \mathcal{H}_0$, knowing (B_i, E_i) for all $i > k$ with $i \in \mathcal{H}_0$, and knowing $\{(B_i, E_i) : i \le k, i \in \mathcal{H}_0\}$ (note that this is an unordered set, as before in e.g. the proof of Lemma 4). Let $\tilde{\mathcal{F}}_k$ be the

σ-algebra that additionally knows $B_1, \ldots, B_n$.

Now we show that $\mathbb{E}\left[M_k \mid \mathcal{F}_{k+1}\right] \leq M_{k+1}$. If $k+1 \notin \mathcal{H}_0$ or if $B_{k+1} = 0$, then $M_k \leq M_{k+1}$ trivially. Turning to the case where $k+1 \in \mathcal{H}_0$ and $B_{k+1} = 1$, we begin by conditioning on $B_1, ..., B_n$. In that case, we see that

$$E_{0,1} + E_{0,2} + \sum_{i \leq k, i \in \mathcal{H}_0} B_i \cdot E_i$$

is a sum of $(2 + \sum_{i \leq k, i \in \mathcal{H}_0} B_i)$ many $\mathsf{Exponential}(1)$ variables, which $\sim$ $\mathsf{Gamma}(2 + \sum_{i \leq k, i \in \mathcal{H}_0} B_i, 1)$, while

$$B_{k+1} \cdot E_{k+1}$$

is equal to another (independent) $\mathsf{Exponential}(1)$ variable, which $\sim \mathsf{Gamma}(1, 1)$. We will use the fact that

$$\text{If } X \sim \mathsf{Gamma}(k), Y \sim \mathsf{Gamma}(l), X \perp\!\!\!\perp Y, \text{then } \frac{X}{X+Y} \sim \mathsf{Beta}(k, l), \text{ and } \frac{X}{X+Y} \perp\!\!\!\perp X+Y \ .$$

Conditioning on $B_1, ..., B_n$,

$$\frac{E_{0,1} + E_{0,2} + \sum_{i \leq k, i \in \mathcal{H}_0} B_i \cdot E_i}{E_{0,1} + E_{0,2} + \sum_{i \leq k+1, i \in \mathcal{H}_0} B_i \cdot E_i} \sim \mathsf{Beta}(2 + \sum_{i \leq k, i \in \mathcal{H}_0} B_i, 1)$$

And conditioning on $\tilde{\mathcal{F}}_{k+1}$ yields the same result. Now, we will use the fact that

$$\text{If } X \sim \mathsf{Beta}(\alpha, \beta) \text{ and } \alpha > 1 \text{ then } \mathbb{E}\left[\frac{1}{X}\right] = \frac{\alpha + \beta - 1}{\alpha - 1} \ .$$

Therefore,

$$\mathbb{E}\left[\frac{E_{0,1} + E_{0,2} + \sum_{i \leq k+1, i \in \mathcal{H}_0} B_i \cdot E_i}{E_{0,1} + E_{0,2} + \sum_{i \leq k, i \in \mathcal{H}_0} B_i \cdot E_i} \;\middle|\; \tilde{\mathcal{F}}_{k+1}\right] = \frac{1 + \sum_{i \leq k+1, i \in \mathcal{H}_0} B_i}{1 + \sum_{i \leq k, i \in \mathcal{H}_0} B_i}.$$

We then have

$$
\begin{aligned}
&\mathbb{E}\left[M_k \mid \mathcal{F}_{k+1}\right] \\
&= \mathbb{E}\left[\frac{1+\#\{i\le k : i\in\mathcal{H}_0\}}{E_{0,1}+E_{0,2}+\sum_{i\le k, i\in\mathcal{H}_0} B_i\cdot E_i} \;\middle|\; \mathcal{F}_{k+1}\right] \\
&= \mathbb{E}\left[\frac{1+\#\{i\le k : i\in\mathcal{H}_0\}}{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i}\cdot\frac{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i}{E_{0,1}+E_{0,2}+\sum_{i\le k, i\in\mathcal{H}_0} B_i\cdot E_i} \;\middle|\; \mathcal{F}_{k+1}\right] \\
&= \mathbb{E}\left[\frac{(1+\#\{i\le k : i\in\mathcal{H}_0\})\cdot\mathbb{E}\left[\frac{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i}{E_{0,1}+E_{0,2}+\sum_{i\le k, i\in\mathcal{H}_0} B_i\cdot E_i} \;\middle|\; \tilde{\mathcal{F}}_{k+1}\right]}{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i} \;\middle|\; \mathcal{F}_{k+1}\right] \\
&\text{since here } k+1\in\mathcal{H}_0,\ \#\{i\le k : i\in\mathcal{H}_0\} = \#\{i\le k+1 : i\in\mathcal{H}_0\} - 1 \text{ is known, given } \tilde{\mathcal{F}}_{k+1} \\
&= \mathbb{E}\left[\frac{1+\#\{i\le k : i\in\mathcal{H}_0\}}{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i}\cdot\frac{1+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i}{1+\sum_{i\le k, i\in\mathcal{H}_0} B_i} \;\middle|\; \mathcal{F}_{k+1}\right] \\
&= \frac{1+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i}{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i}\cdot\mathbb{E}\left[\frac{1+\#\{i\le k : i\in\mathcal{H}_0\}}{1+\sum_{i\le k, i\in\mathcal{H}_0} B_i} \;\middle|\; \mathcal{F}_{k+1}\right] \\
&\le \frac{1+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i}{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i}\cdot\frac{1+\#\{i\le k+1 : i\in\mathcal{H}_0\}}{1+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i} \\
&= \frac{1+\#\{i\le k+1 : i\in\mathcal{H}_0\}}{E_{0,1}+E_{0,2}+\sum_{i\le k+1, i\in\mathcal{H}_0} B_i\cdot E_i} = M_{k+1}\ ,
\end{aligned}
$$

where the inequality in the next-to-last step comes from Lemma 5. This proves that M_k is

a supermartingale. Finally, we have

$$\begin{aligned}
\mathbb{E}\left[M_n\right] &= \mathbb{E}\left[\frac{1+|\mathcal{H}_0|}{E_{0,1}+E_{0,2}+\sum_{i\in\mathcal{H}_0} B_i\cdot E_i}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1+|\mathcal{H}_0|}{E_{0,1}+E_{0,2}+\sum_{i\in\mathcal{H}_0} B_i\cdot E_i}\;\middle|\;\sum_{i\in\mathcal{H}_0} B_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1+|\mathcal{H}_0|}{\mathsf{Gamma}(2+\sum_{i\in\mathcal{H}_0} B_i,1)}\;\middle|\;\sum_{i\in\mathcal{H}_0} B_i\right]\right] \\
&= \mathbb{E}\left[\frac{1+|\mathcal{H}_0|}{(2+\sum_{i\in\mathcal{H}_0} B_i)-1}\right] \\
&\leq C\,,
\end{aligned}$$

where the next-to-last step uses the mean of an inverse-gamma distribution, and where again we apply Lemma 5 for the last step. □

2.7.4 Proof of Theorem 3 (asymptotic FDR control)

Proof of Theorem 3. Define

$$\epsilon_n = \max\left\{\mathbb{P}\left\{\widehat{k}_{\mathsf{h}} < m_n\right\}, \frac{1}{m_n}\right\}.$$

Then $\epsilon_n \to 0$ by assumption. By Theorem 1, for each n,

$$\mathbb{P}\left\{\mathsf{FDP}(\widehat{k}_{\mathsf{h}}) > \alpha + \max\left\{\sigma, b\sqrt{3\log\left(8/\epsilon_n\right)}\right\}\cdot\sqrt{\frac{16\log\left(\frac{8\log(2\widehat{k}_{\mathsf{h}})}{\epsilon_n}\right)}{\widehat{k}_{\mathsf{h}}}}\right\} \leq \epsilon_n,$$

and therefore,

$$\mathbb{P}\left\{\mathsf{FDP}(\widehat{k}_{\mathsf{h}}) > \alpha + \max\left\{\sigma, b\sqrt{3\log\left(8/\epsilon_n\right)}\right\} \cdot \sqrt{\frac{16\log\left(\frac{8\log(2m_n)}{\epsilon_n}\right)}{m_n}}\right\} \leq \epsilon_n + \mathbb{P}\left\{\widehat{k}_{\mathsf{h}} < m_n\right\} \leq 2\epsilon_n.$$

Furthermore, by definition of ϵ_n, we see that

$$\max\left\{\sigma, b\sqrt{3\log\left(8/\epsilon_n\right)}\right\} \cdot \sqrt{\frac{16\log\left(\frac{8\log(2m_n)}{\epsilon_n}\right)}{m_n}} \leq \max\left\{\sigma, b\sqrt{3\log\left(8m_n\right)}\right\} \cdot \sqrt{\frac{16\log\left(8m_n\log(2m_n)\right)}{m_n}} \to 0.$$

So for any fixed δ,

$$\mathbb{P}\left\{\mathsf{FDP}(\widehat{k}_{\mathsf{h}}) > \alpha + \delta\right\} \leq \mathbb{1}\left\{\max\left\{\sigma, b\sqrt{3\log\left(8/\epsilon_n\right)}\right\} \cdot \sqrt{\frac{16\log\left(\frac{8\log(2m_n)}{\epsilon_n}\right)}{m_n}} > \delta\right\} + 2\epsilon_n,$$

and each term on the right-hand side is zero in the limit. □

2.7.5 Proof of Theorem 4 (asymptotic power calculation)

The proof of this theorem will again use Lemma 3, our result on bounding random walks.

Proof of Theorem 4. First, we define the asymptotic expected FDP estimate along the sequence of p-values,

$$\mathsf{E}(t) = \mu_0 - f(t) \cdot (\mu_0 - \mu_1) .$$

That is, at the cutoff $k = t \cdot n$, we expect that $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) \approx \mathsf{E}(t)$. Note that $\mathsf{E}(t)$ is monotone

nondecreasing, due to the assumption that $f(t)$ is nonincreasing.

Next, we prove that the approximation $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) \approx \mathsf{E}(t)$ is uniformly accurate. Fix any k with $\log(n) < k \leq n$. First we consider the expectation:

$$\mathbb{E}\left[\frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k}\right] = \mathbb{E}\left[\frac{k\mu_0 - \sum_{i=1}^{k}(\mu_0 - \mathsf{h}(p_i))}{k}\right] = \mu_0 - \frac{\#\{i \leq k : i \notin \mathcal{H}_0\}}{k} \cdot (\mu_0 - \mu_1)\,,$$

and so, applying (2.16),

$$\left|\mathbb{E}\left[\frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k}\right] - \mathsf{E}\left(\frac{k}{n}\right)\right| = \left|\frac{\#\{i \leq k : i \notin \mathcal{H}_0\}}{k} - f\left(\frac{k}{n}\right)\right| \cdot (\mu_0 - \mu_1)$$
$$\leq (\mu_0 - \mu_1)\epsilon_n\,. \quad (2.28)$$

Next, we apply Lemma 3 to prove that $\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) = \frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k} \approx \mathbb{E}\left[\frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k}\right]$ for all sufficiently large k. Applying Lemma 3 with $\epsilon = \frac{1}{\log(n)}$, we see that with probability at least $1 - \frac{1}{\log(n)}$,

$$\left|\frac{\sum_{i=1}^{k} \mathsf{h}(p_i)}{k} - \frac{\sum_{i=1}^{k} \mathbb{E}\left[\mathsf{h}(p_i)\right]}{k}\right| \leq$$
$$\max\left\{\sigma, b\sqrt{3\log\left(8\log(n)\right)}\right\} \cdot \sqrt{\frac{16\log\left(8\log(n)\log(2k)\right)}{k}} \text{ for all } k \geq 1\,, \quad (2.29)$$

and therefore, since we are restricting our attention to $k > \log(n)$, combining with our result in (2.28), we have

$$\max_{\log(n)<k\leq n}\left|\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) - \mathsf{E}\left(\frac{k}{n}\right)\right| <$$
$$\max\left\{\sigma, b\sqrt{3\log\left(8\log(n)\right)}\right\} \cdot \sqrt{\frac{16\log\left(8\log(n)\log(2\log(n))\right)}{\log(n)}} + (\mu_0 - \mu_1)\epsilon_n =: \beta_n\,. \quad (2.30)$$

Note that $\beta_n \to 0$. We also define $\tau_n = \frac{\beta_n}{(\mu_0 - \mu_1)\nu}$, where the constant ν is from assumption (2.17), and note that $\tau_n \to 0$ also.

Now we split into cases.

Case 1: α satisfies $f(1) < \frac{\mu_0-\alpha}{\mu_0-\mu_1} < f(0)$. In this case, we will prove that, if (2.29) holds, then

$$\begin{cases} \widehat{\mathsf{FDP}}_{\mathsf{h}}(k) \leq \alpha \text{ for all } \log(n) < k \leq n \cdot (T - \tau_n), \text{ and} \\ \widehat{\mathsf{FDP}}_{\mathsf{h}}(k) > \alpha \text{ for all } k > \max\{\log(n), n \cdot (T + \tau_n)\} \ . \end{cases} \tag{2.31}$$

Now consider n sufficiently large so that $n \cdot (T - \tau_n) > \log(n)$. If the above holds, then by definition of $\widehat{k}_{\mathsf{h}}$, this implies that

$$n \cdot (T - \tau_n) \leq \widehat{k}_{\mathsf{h}} \leq n \cdot (T + \tau_n) \ ,$$

and therefore,

$$\frac{\#\{i \leq n \cdot (T - \tau_n) : i \notin \mathcal{H}_0\}}{n - |\mathcal{H}_0|} \leq \mathsf{TPP}(\widehat{k}_{\mathsf{h}}) \leq \frac{\#\{i \leq n \cdot (T + \tau_n) : i \notin \mathcal{H}_0\}}{n - |\mathcal{H}_0|} \ .$$

Using assumption (2.16), therefore,

$$\frac{n \cdot (T - \tau_n) \cdot (f(T - \tau_n) - \epsilon_n)}{n \cdot (f(1) + \epsilon_n)} \leq \mathsf{TPP}(\widehat{k}_{\mathsf{h}}) \leq \frac{n \cdot (T + \tau_n) \cdot (f(T + \tau_n) + \epsilon_n)}{n \cdot (f(1) - \epsilon_n)} \ . \tag{2.32}$$

Since the limit of both sides is equal to $T \cdot \frac{f(T)}{f(1)}$, this proves the desired result. To be more precise, we have proved that the bound (2.32) holds with probability at least $1 - \frac{1}{\log(n)}$ (since this is a lower bound on the probability of the event (2.29)), which itself tends to 1. Therefore, the TPP of our procedure converges to the limit $T \cdot \frac{f(T)}{f(1)}$ in probability.

It remains to be shown that (2.29) implies (2.31). First, note that $f(T) = \frac{\mu_0-\alpha}{\mu_0-\mu_1} \geq 1 - \frac{\alpha}{\mu_0} > 1 - \alpha$, and so, since $\tau_n \to 0$ and f is continuous, we see that

$$f(T), f(T + \tau_n) \geq 1 - \alpha$$

for sufficiently large n. And, $\tau_n \leq \delta$, for large n. Therefore, by assumption (2.17),

$$f(T - \tau_n) \geq f(T) + \nu\tau_n.$$

Now take any k such that $\log(n) < k \leq n \cdot (T - \tau_n)$. Then

$$\begin{aligned}\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) &\leq \mathsf{E}\left(\frac{k}{n}\right) + \beta_n \leq \mathsf{E}(T - \tau_n) + \beta_n = \mu_0 - f(T - \tau_n) \cdot (\mu_0 - \mu_1) + \beta_n \\ &\leq \mu_0 - (f(T) + \tau_n \cdot \nu) \cdot (\mu_0 - \mu_1) + \beta_n = \mu_0 - f(T) \cdot (\mu_0 - \mu_1) = \alpha \ ,\end{aligned}$$

where the first inequality applies (2.30). This proves the first part of (2.31); the second part of (2.31) is proved similarly, using the fact that $f(T + \tau_n) \leq f(T) - \nu\tau_n$ by again applying (2.17).

Case 2: α satisfies $\frac{\mu_0 - \alpha}{\mu_0 - \mu_1} \geq f(0)$ For this case, we will show that power tends to zero. As in the first case, it will be sufficient to show that

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) > \alpha \text{ for all } k > \max\{\tau_n \cdot n, \log(n)\} \ . \tag{2.33}$$

To prove this, take any such k. Then, if the event (2.29) holds, we apply (2.30) to get

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) > \mathsf{E}\left(\frac{k}{n}\right) - \beta_n \geq \mathsf{E}(\tau_n) - \beta_n = \mu_0 - f(\tau_n) \cdot (\mu_0 - \mu_1) - \beta_n \ .$$

If $f(0) = \frac{\mu_0 - \alpha}{\mu_0 - \mu_1}$ exactly, then $f(0) \geq 1 - \frac{\alpha}{\mu_0} > 1 - \alpha$ and so $f(\tau_n) \geq 1 - \alpha$ for sufficiently large n. Therefore, applying assumption (2.17), $f(\tau_n) \leq f(0) - \tau_n\nu$ and then

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) > \mu_0 - (f(0) - \tau_n\nu) \cdot (\mu_0 - \mu_1) - \beta_n = \alpha \ .$$

Alternately, if $f(0) < \frac{\mu_0 - \alpha}{\mu_0 - \mu_1}$, then since f is continuous, for sufficiently large n we have $f(\tau_n) \leq \frac{\mu_0 - \alpha - \beta_n}{\mu_0 - \mu_1}$ and then the same bound holds.

Case 3: α satisfies $\frac{\mu_0-\alpha}{\mu_0-\mu_1} \leq f(1)$ For this case, we will show that power tends to 1. As in the previous cases, it will be sufficient to show that

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) < \alpha \text{ for all } \log(n) < k < n \cdot (1-\tau_n) \, .$$

To prove this, take any such k. Then, if the event (2.29) holds, we apply (2.30) to get

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) < \mathsf{E}\left(\frac{k}{n}\right) + \beta_n \leq \mathsf{E}(1-\tau_n) + \beta_n = \mu_0 - f(1-\tau_n) \cdot (\mu_0 - \mu_1) + \beta_n \, .$$

Since $f(1) \geq \frac{\mu_0-\alpha}{\mu_0-\mu_1} > 1-\alpha$, we see that $f(t) \geq 1-\alpha$ for all $t \in [1-\tau_n, 1]$ for sufficiently large n, and so $f(1-\tau_n) \geq f(1) + \tau_n \nu$ by assumption (2.17). Then,

$$\widehat{\mathsf{FDP}}_{\mathsf{h}}(k) < \mu_0 - (f(1) + \tau_n \nu) \cdot (\mu_0 - \mu_1) + \beta_n \leq \alpha \, .$$

□

2.7.6 Proof of Lemma 2 (bounded accumulation functions)

Proof of Lemma 2. We have

$$\begin{aligned}
\mathbb{E}\left[\mathsf{h}(p_i)\right] - \mathbb{E}\left[\mathsf{h}_0(p_i)\right] &= \int_{t=0}^{1} (\mathsf{h}(t) - \mathsf{h}_0(t)) \cdot f_i(t) \, \mathsf{d}t \\
&= \int_{t=0}^{1-1/C} (\mathsf{h}(t) - 0) \cdot f_i(t) \, \mathsf{d}t + \int_{t=1-1/C}^{1} (\mathsf{h}(t) - C) \cdot f_i(t) \, \mathsf{d}t \\
&= \int_{t=0}^{1-1/C} \mathsf{h}(t) \cdot f_i(t) \, \mathsf{d}t - \int_{t=1-1/C}^{1} (C - \mathsf{h}(t)) \cdot f_i(t) \, \mathsf{d}t \\
&\geq \int_{t=0}^{1-1/C} \mathsf{h}(t) \cdot f_i(1-1/C) \, \mathsf{d}t - \int_{t=1-1/C}^{1} (C - \mathsf{h}(t)) \cdot f_i(1-1/C) \, \mathsf{d}t \\
&= f_i(1-1/C) \cdot \left[\int_{t=0}^{1} \mathsf{h}(t) \, \mathsf{d}t - \int_{t=1-1/C}^{1} C \, \mathsf{d}t \right] \\
&= f_i(1-1/C) \cdot [1-1] = 0 \, ,
\end{aligned}$$

where the inequality is true since f_i is nonincreasing and since $\mathsf{h}(t) \geq 0$ and $C - \mathsf{h}(t) \geq 0$. Furthermore, if the inequality is not strict (i.e. is an equality), then we must have

$$\int_{t=0}^{1-1/C} \mathsf{h}(t) \cdot [f_i(t) - f_i(1-1/C)] \ \mathsf{d}t = 0$$

and

$$\int_{t=1-1/C}^{1} (C - \mathsf{h}(t)) \cdot [f_i(1-1/C) - f_i(t)] \ \mathsf{d}t = 0 \ .$$

Note that, in both integrals, the integrand is nonnegative. Therefore, in order for the integrals to equal zero, it must be true that the integrands are equal to zero almost everywhere. However, if f_i is strictly decreasing then the terms in square brackets are strictly positive in both integrals (except at endpoints). Therefore, in the first integral we must have $\mathsf{h}(t) = 0$ almost everywhere over $t \in (0, 1-1/C)$, and in the second integral we must have $C - \mathsf{h}(t) = 0$ almost everywhere over $t \in (1-1/C, 1)$. In other words, $\mathsf{h}(t) = \mathsf{h}_0(t)$ almost everywhere over $t \in [0, 1]$. $\square$

CHAPTER 3

MULTIPLE TESTING WITH THE STRUCTURE ADAPTIVE BENJAMINI-HOCHBERG ALGORITHM

Classical approaches to the multiple testing problem generally treat the hypotheses exchangeably; that is, the labels themselves play no role in the process of deciding which p-values to label as "discoveries". In many settings, however, we may have prior information or beliefs about the pattern of signals and nulls among the hypotheses being tested. For example, we may believe that certain hypotheses are more likely to contain signals than others (e.g. due to data from prior experiments); or that the true signals are likely to be clustered (e.g. if each hypothesis corresponds to a test performed at some spatial location, and the true signals are likely to appear in spatial clusters); or the hypotheses may come with some natural grouping, with true signals tending to co-occur in the same group (e.g. if each hypothesis corresponds to a gene, then known gene pathways may form such a grouping).

In this chapter, we give a general framework for incorporating this type of information when performing multiple testing. We introduce the *structure-adaptive Benjamini-Hochberg procedure*, a procedure which places data-adaptive weights on the p-values in order to adapt to the apparent patterns of signals and nulls in the data. This procedure offers increased power to detect signals by lowering the threshold for making discoveries in regions where the data suggests that signals are highly likely to occur. When run with a target FDR level α, the method offers a finite-sample guarantee of FDR control at a level that is only slightly higher than α as long as we restrict the extent to which the weights can adapt to the data, thus avoiding overfitting.

Chapter outline[1]: In section 3.1, we present the SABHA method and main theoretical results. In section 3.2, we give the details for applying SABHA in several different settings. In section 3.3, we prove the FDR control result. In section 3.4, we present empirical results

1. The work presented in this chapter is published in Li and Barber [2016b]

on simulated data and on several real data sets. In section 3.5, we give additional proofs, and the details of the SABHA algorithm.

3.1 The structure adaptive Benjamini-Hochberg procedure

We now define the structure-adaptive Benjamini-Hochberg algorithm.

Definition 2 (Structure-adaptive Benjamini-Hochberg algorithm (SABHA))**.** Given a target FDR level $\alpha \in [0,1]$, a threshold $\tau \in [0,1]$, and values $\widehat{q}_1, \ldots, \widehat{q}_n \in [0,1]$ (where $\widehat{q}_i$ represents an estimated probability that the ith test corresponds to a null), define

$$\widehat{k} = \max\left\{k \geq 1 : P_i \leq \left(\frac{\alpha}{\widehat{q}_i} \cdot \frac{k}{n}\right) \wedge \tau \text{ for at least } k \text{ many p-values } P_i\right\},$$

with the convention that $\widehat{k} = 0$ if this set is empty. Then the SABHA method rejects any p-value P_i satisfying

$$P_i \leq \left(\frac{\alpha}{\widehat{q}_i} \cdot \frac{\widehat{k}}{n}\right) \wedge \tau,$$

for a total of $\widehat{k}$ many rejections.

Note that, if we set $\widehat{q}_i = 1$ for all i, then this is exactly the original Benjamini-Hochberg procedure; if instead we set $\widehat{q}_i = \widehat{\pi}_0$ to be an estimate of the proportion of nulls (which is constant across all i), then this can give us Storey's modification of the BH procedure. Alternately, if we choose $\widehat{q}$ to be a fixed vector (i.e. not dependent on P), then this is equivalent to Genovese et al. [2006] p-value weighting method, their weights w_i are given by $1/\widehat{q}_i$ in our notation.

To understand our proposed method intuitively, we sketch a coarse estimate for the false discovery proportion of this method. We consider a random model where each p-value P_i has a probability q_i of being a null. Suppose that $\widehat{q}_i \approx q_i$ is an accurate approximation. For

some fixed k, we would like to know how many nulls satisfy $P_i \leq \left(\frac{\alpha}{\widehat{q}_i} \cdot \frac{k}{n}\right) \wedge \tau$. We have

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\left\{i \text{ is null and } P_i \leq \left(\frac{\alpha}{\widehat{q}_i} \cdot \frac{k}{n}\right) \wedge \tau\right\}\right] \approx \sum_{i=1}^{n} q_i \cdot \left[\left(\frac{\alpha}{q_i} \cdot \frac{k}{n}\right) \wedge \tau\right] \leq \alpha k,$$

where the approximation holds since $\widehat{q}_i \approx q_i$, and if P_i is a null p-value then it should be uniformly distributed. Therefore, when we reject $\widehat{k}$ many p-values, we expect that there are $\approx \alpha\widehat{k}$ many nulls among them, leading to a false discovery proportion that is $\approx \alpha$.

In fact, our theoretical results will show that this method has an interesting property: even without assuming that $\widehat{q}$ estimates some underlying random model over nulls and signals, we can nonetheless bound the false discovery rate of this method at a level that is not much higher than the target level α, provided that (1) $\widehat{q}$ is chosen from some low-complexity class of vectors, with a few additional mild constraints, and (2) $\widehat{q}$ is a function of only those p-values that lie above the threshold τ.

3.1.1 FDR control result

We begin by considering a fixed (and unknown) set of nulls, $\mathcal{H}_0 \subseteq [n]$, and choosing a fixed threshold $\tau \in (0, 1)$. In many applications, null p-values might not be exactly uniformly distributed (for instance the p-values may be discretized), so we instead require that their distribution is "τ-conservative"[2] as well as independent:

$$\begin{cases} \{P_i : i \in \mathcal{H}_0\} \text{ are mutually independent and are independent of } \{P_i : i \notin \mathcal{H}_0\}; \\ \mathbb{P}\{P_i \leq \tau\} \leq \tau \text{ for all } i \in \mathcal{H}_0; \text{ and} \\ \mathbb{P}\{P_i \leq t \mid P_i \leq \tau\} \leq t/\tau \text{ for all } t \in [0, \tau] \text{ and all } i \in \mathcal{H}_0 \text{ with } \mathbb{P}\{P_i \leq \tau\} > 0. \end{cases} \tag{3.1}$$

2. If desired, we can remove the last part of the τ-conservative assumption and simply require that $\mathbb{P}\{P_i \leq t\} \leq t$ for all $t \in [0, \tau]$ (i.e. nulls are conservative), at the cost of a slightly poorer constant in the excess FDR bound of our main theorem: in place of $\frac{1}{\tau(1-\tau)}$ we would have $\max_{i \in \mathcal{H}_0} \frac{1}{\mathbb{P}\{P_i \leq \tau\}\mathbb{P}\{P_i > \tau\}}$.

We also assume that, fixing a set $\mathcal{Q} \subseteq (0,1]^n$ which contains $\mathbf{1}_n$ (the vector with all 1's), the vector $\widehat{q} = \widehat{q}(P)$ satisfies

$$\begin{cases} \widehat{q}(P) \in \mathcal{Q} \text{ depends on the p-values } P \text{ only through } (P_i \cdot \mathbb{1}\{P_i > \tau\})_i; \text{ and} \\ \text{Either } \sum_{i=1}^n \frac{\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i(1-\tau)} \leq n \text{ or } \widehat{q}(P) = \mathbf{1}_n. \end{cases} \tag{3.2}$$

The first requirement of (3.2), that $\widehat{q}(P)$ cannot depend on p-values falling below τ, ensures that these low p-values (i.e. any p-values that we might reject, according to the SABHA method) are still "random" even after we compute $\widehat{q}$. To understand the intuition behind the second part of (3.2), we again consider the random model for nulls and non-nulls: suppose that P_i is uniformly distributed (i.e. null) with probability q_i, and the signal strength is such that the non-null p-values are almost never above the threshold τ. Then we would have

$$\mathbb{P}\{P_i > \tau\} \approx \mathbb{P}\{P_i > \tau \text{ and } P_i \text{ is a null}\} = q_i \cdot (1-\tau),$$

and therefore,

$$\mathbb{E}\left[\sum_{i=1}^n \frac{\mathbb{1}\{P_i > \tau\}}{q_i(1-\tau)}\right] \approx n.$$

Since this sum will typically concentrate strongly around its expectation, we see that the bound in (3.2) should be approximately true for the true probabilities q; requiring that this bound holds for $\widehat{q}$ essentially means that we are making sure that $\widehat{q}$ predicts an appropriate number of null hypotheses.

Of course, in practice we may find a data set where the bound $\sum_{i=1}^n \frac{\mathbb{1}\{P_i > \tau\}}{q_i(1-\tau)} \leq n$ is impossible to obtain for any $q \in \mathcal{Q}$, even if we minimize the sum by setting $q = \mathbf{1}_n$. This will occur when $\sum_i \mathbb{1}\{P_i > \tau\} > n(1-\tau)$. For example, if all the p-values are uniformly distributed, then $\sum_i \mathbb{1}\{P_i > \tau\} \sim \text{Binomial}(n, 1-\tau)$, which will exceed the bound roughly half of the time. In this setting, we have no evidence to believe that there are many signals present in the data set and it is intuitive to set $\widehat{q}_i = 1$ for all i.

Before stating our main result, we give one additional definition: for any set $\mathcal{A} \subseteq \mathbb{R}^n$, the "Rademacher width"[3] of the set $\mathcal{A}$ is defined as

$$\omega_{\mathrm{Rad}}(\mathcal{A}) = \mathbb{E}\left[\sup_{x \in \mathcal{A}} \langle x, B \rangle\right],$$

where the expectation is taken over a vector B of independent Rademacher variables, $B_i \overset{\text{iid}}{\sim} \text{Uniform}\{\pm 1\}$. It is known that Bartlett and Mendelson [2003]

$$\omega_{\mathrm{Rad}}(\mathcal{A}) \leq \sqrt{\frac{\pi}{2}} \cdot \omega_{\mathrm{Gaus}}(\mathcal{A}),$$

where $\omega_{\mathrm{Gaus}}(\mathcal{A})$ is the Gaussian width,

$$\omega_{\mathrm{Gaus}}(\mathcal{A}) = \mathbb{E}\left[\sup_{x \in \mathcal{A}} \langle x, g \rangle\right] \text{ for } g \sim \mathcal{N}(0, \mathbf{I}_n).$$

In our examples of different types of structure (i.e. different choices of $\mathcal{Q}$), we will bound the Rademacher width directly, but in other settings it may be easier to bound the Rademacher width via the Gaussian width.

We now turn to our main result, proving finite-sample FDR control for the SABHA procedure. The proof is deferred to Section 3.3.

Theorem 5. *Fix a target FDR level $\alpha \in [0,1]$, a threshold $\tau \in (0,1)$, and a set $\mathcal{Q} \subseteq (0,1]^n$ with $\mathbf{1}_n \in \mathcal{Q}$. Suppose that the vector of p-values $P \in [0,1]^n$ satisfies assumption (3.1) and $\widehat{q} = \widehat{q}(P)$ satisfies assumption (3.2). Then the false discovery rate of the SABHA procedure, run with parameters $\alpha, \tau, \widehat{q}(P)$ over the p-values P, is bounded as*

$$\mathrm{FDR} = \mathbb{E}\left[\mathsf{FDP}\right] \leq \alpha\left(1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{\omega_{\mathrm{Rad}}(\mathcal{Q}_{\mathrm{inv}})}{n}\right),$$

3. We use this rescaled version of the Rademacher complexity (see e.g. Bartlett and Mendelson [2003] for background) in order to agree with the standard definition of the Gaussian width.

where the expectation is taken with respect to the distribution of the p-values, and where

$$\mathcal{Q}_{\text{inv}} = \left\{ \left((q_1)^{-1}, \ldots, (q_n)^{-1} \right) : q \in \mathcal{Q} \right\}.$$

Note that, since this result holds for any fixed set of nulls $\mathcal{H}_0 \subseteq [n]$, it would also hold in expectation over any random model for the nulls and non-nulls.

To interpret the results of this theorem, we would typically choose a class $\mathcal{Q}$ which is sufficiently simple to ensure that $\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \ll n$; in this type of setting, the FDR bound would be only slightly higher than α. $\mathcal{Q}$ should reflect our beliefs about the structure of the problem, for instance, we may believe that the signals appear in clusters among our n hypotheses.

In contrast, if we do not sufficiently constrain $\mathcal{Q}$ then the bound on FDR can become meaningless—for instance, if we take $\mathcal{Q} = [\alpha, 1]^n$, then $\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) = n \cdot \frac{\alpha^{-1}-1}{2}$ and the upper bound on FDR is not even smaller than 1. This is not merely an artifact of the theory. For instance, suppose that $P_i \overset{\text{iid}}{\sim} \text{Uniform}[0,1]$ for all i, and so there is no signal in the data. If we choose $\widehat{q}_i = 1$ whenever $P_i > \tau$ and $\widehat{q}_i = \alpha$ whenever $P_i \leq \tau$, then whenever $\sum_i \mathbb{1}\{P_i \leq \tau\} \geq n\tau$ (which occurs roughly half the time), the SABHA procedure rejects *all* p-values $P_i \leq \tau$. The source of the problem is that the vector $\widehat{q}$ is drastically overfitting to the patterns in the binary vector $(\mathbb{1}\{P_i > \tau\})_i$. By instead choosing $\mathcal{Q}$ with a low Rademacher width, we ensure that $\widehat{q}$ cannot overfit too much to this data.

3.2 Application of SABHA to specific types of structure

We now consider the application of our main result to various specific structured settings. We will compare to existing work for the specific settings. For each case we also provide bounds on the Rademacher complexity of $\mathcal{Q}_{\text{inv}}$ for the relevant sets $\mathcal{Q}$, which allow us to apply our main FDR control result, Theorem 5; these results are proved in Section 3.5.

3.2.1 Ordered structure

In an ordered setting, we might believe that the hypotheses early in the list are more likely to contain true signals than those later in the list.

In this setting, one natural choice for $\widehat{q} = \widehat{q}(P)$ would be to require that this vector is nondecreasing. To avoid degeneracy, we also impose a lower bound $\epsilon > 0$, and define the set

$$\mathcal{Q} = \mathcal{Q}_{\text{ord}} = \{q : \epsilon \leq q_1 \leq \cdots \leq q_n \leq 1\}. \tag{3.3}$$

We could alternately choose to enforce that $\widehat{q}$ is a "step function" of the form $\widehat{q} = (\epsilon, \ldots, \epsilon, 1, \ldots, 1)$, in which case $\widehat{q} \in \mathcal{Q}_{\text{ord}}$ (in fact, $\mathcal{Q}_{\text{ord}}$ is the convex hull of such "step functions").

Now we examine the FDR control of the SABHA method for this ordered setting.

Lemma 6. *For $\mathcal{Q} = \mathcal{Q}_{\text{ord}}$ as defined above,*

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq (\epsilon^{-1} - 1) \cdot \sqrt{n}.$$

Therefore, applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{\tau}{1-\tau}} + \frac{\epsilon^{-1} - 1}{\tau(1-\tau)}\right]\right),$$

In this setting, we obtain meaningful bounds on FDR when $\epsilon \gg 1/\sqrt{n}$.

In this ordered scenario, the existing methods of G'Sell et al. [2015], Barber and Candès [2015], and and Li and Barber [2016a], mentioned earlier in chapter 2, propose algorithms for choosing an adaptive cutoff point k and labeling $P_1, \ldots, P_k$ as signals and $P_{k+1}, \ldots, P_n$ as nulls. Alternately, the Selective SeqStep method of Barber and Candès [2015] selects an adaptive cutoff k and then rejects P_i for $i \leq k$ only if $P_i \leq \tau$ for some predetermined threshold τ. In contrast, the ordered adaptive method proposed here does not have a firm cutoff; instead, the p-values early in the list are given higher priority, but rejections may

occur at any point in the list, e.g. if the last P_n is extremely low then it is likely to be rejected even if the overall number of rejections is small. The version of our method that is most comparable to these existing works, is when we choose $\widehat{q}$ to be a "step function", $\widehat{q} = (\epsilon, \ldots, \epsilon, 1, \ldots, 1)$; in this case, when k is the data-adaptive number of ϵ values in $\widehat{q}$, the first k p-values are given high priority for discovery while the remaining $n-k$ p-values are handled as in the original BH procedure, allowing for strong p-values to be rejected even if they appear late in the list. In this sense, this version of our method can be viewed as a relaxation of the ordered testing methods of G'Sell et al. [2015], Barber and Candès [2015], Li and Barber [2016a].

Choosing $\widehat{q} \in \mathcal{Q}$ To choose a monotone vector $\widehat{q}$ which satisfies $\epsilon \leq \widehat{q}_1 \leq \cdots \leq \widehat{q}_n \leq 1$ and reflects the patterns observed in the data, one approach would be to consider $\widehat{q}_i(1-\tau)$ as an estimate of $\mathbb{P}\{P_i > \tau\}$, and then maximize the likelihood of this model:

$$\widehat{q} = \underset{q \in \mathbb{R}^n}{\arg\max} \Bigg\{ \sum_i \mathbb{1}\{P_i > \tau\} \log\left(q_i(1-\tau)\right) + \mathbb{1}\{P_i \leq \tau\} \log\left(1 - q_i(1-\tau)\right) \\ : q \in \mathcal{Q}_{\text{ord}}, \sum_i \frac{\mathbb{1}\{P_i > \tau\}}{q_i(1-\tau)} \leq n \Bigg\}.$$

(If we have $\sum_i \mathbb{1}\{P_i > \tau\} > n(1-\tau)$ then we instead set $\widehat{q} = \mathbf{1}_n$.) We give an algorithm for solving this convex optimization problem in Section 3.5.4.

Alternately, instead of a likelihood-based approach for choosing a monotone vector, we may alternately wish to consider a simpler construction for $\widehat{q}$, given by a "step function" of the form

$$q^k = (\underbrace{\epsilon, \ldots, \epsilon}_{k \text{ times}}, \underbrace{1, \ldots, 1}_{n-k \text{ times}}) \tag{3.4}$$

as discussed above. The highest power (i.e. largest number of rejections) will be obtained by taking k as large as possible while still ensuring that the assumptions (3.2) hold. In this

case, we can simply set

$$K = \max\left\{k = 1, \dots, n : \sum_{i=1}^{k} \frac{\mathbb{1}\{P_i > \tau\}}{\epsilon(1-\tau)} + \sum_{i=k+1}^{n} \frac{\mathbb{1}\{P_i > \tau\}}{1-\tau} \le n\right\}, \tag{3.5}$$

or set $K = 0$ if this set is empty; we then set $\widehat{q} = q^K$ as in (3.4), which satisfies (3.2) by our choice of K.

Choosing $\widehat{q}$ to be a step function has some similarity to the recent Adaptive SeqStep method of Lei and Fithian [2016], which rejects all $P_i \le \epsilon$ for $i = 1, \dots, K$, where ϵ is fixed while K is determined adaptively by estimating the FDP of this set of discoveries using the indicators $\{P_i > \tau\}$ to estimate the proportion of nulls present.[4] To compare these two methods, Adaptive SeqStep finds an adaptive cutoff K and uses a *fixed* rejection threshold ϵ for all p-values that appear before the cutoff (i.e. $P_1, \dots, P_K$), while SABHA with $\widehat{q} = q^K$ also finds an adaptive cutoff K but uses an *adaptive* rejection threshold, and also is able to reject p-values after the cutoff (i.e. P_i for $i > K$) if they are extremely low.

3.2.2 Block structure

Suppose that the set of hypotheses is partitioned into blocks, $[n] = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_d$, with block sizes $n_1 + \cdots + n_d = n$, with the signals likely to appear together in these blocks according to some natural structure or prior information. Specifically, we might believe that each block has its own proportion of nulls and non-nulls, and might then estimate $\widehat{q}$ to be constant within each block but allow it to vary across blocks. Define

$$\mathcal{Q} = \mathcal{Q}_{\text{block}} = \left\{q : \epsilon \le q_i \le 1 \text{ for all } i, \text{ and } q_i = q_j \text{ whenever } i, j \text{ are in the same block}\right\}. \tag{3.6}$$

(We again impose a lower bound ϵ to avoid degeneracy.) For this choice of $\mathcal{Q}$, this is in fact equivalent to Hu et al. [2012]'s "Group Benjamini-Hochberg" method for block-wise

4. Our parameters ϵ, τ, K are equivalent to their notation $s, \lambda, \widehat{k}$.

reweighting, where the proportion of nulls in each block, $\pi_0^{(k)} = \frac{|\mathcal{H}_0 \cap \mathcal{B}_k|}{|\mathcal{B}_k|}$, is estimated and then used to recalibrate the BH procedure. To compare notation, in our setting since we choose $\widehat{q}$ to be block-wise constant, we can write $\widehat{q}_i = \widehat{\pi}_0^{(k)}$ for each $i \in \mathcal{B}_k$, for blocks $k = 1, \ldots, d$. Hu et al. [2012]'s theoretical results offer exact finite-sample FDR control in the oracle setting where the $\pi_0^{(k)}$'s are known rather than estimated, and asymptotic FDR control otherwise. We will now see that, if the partition of $[n]$ into blocks is not too fine, then our main result in Theorem 5 offers a finite-sample FDR guarantee when the $\pi_0^{(k)}$'s are estimated adaptively from the data, rather than known in advance.

Lemma 7. *For $\mathcal{Q} = \mathcal{Q}_{\text{block}}$ as defined above, if the blocks are of sizes $n_1, \ldots, n_d$, then*

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \frac{\epsilon^{-1} - 1}{2} \sum_{i=1}^{d} \sqrt{n_i}.$$

Therefore, applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{\epsilon^{-1} - 1}{2} \cdot \frac{\sum_{i=1}^{d} \sqrt{n_i}}{n} \right),$$

In this setting, to obtain meaningful bounds on FDR we need $\epsilon \gg \frac{\sum_{i=1}^{d} \sqrt{n_i}}{n}$. For example, if the d blocks are each of size $n_i = n/d$, then we require $\epsilon \gg \sqrt{d/n}$.

In summary, the SABHA method in this setting is an example of the procedure proposed by Hu et al. [2012] (where the estimated proportion of nulls within each group can be determined with any desired data-adaptive method). Hu et al. [2012]'s theoretical results for the adaptive setting prove asymptotic FDR control, while our theoretical results are able to give a strong finite-sample guarantee for the data-adaptive procedure.

Choosing $\widehat{q} \in \mathcal{Q}$ To choose a block-wise constant vector $\widehat{q}$, we simply need to choose a value $\widetilde{q}_k$ for each block $k = 1, \ldots, d$ and then set $\widehat{q}_i = \widetilde{q}_k$ for each $i \in \mathcal{B}_k$. Here $\widetilde{q}_k$ should represent our estimated proportion of nulls in block $\mathcal{B}_k$. We will maximize the likelihood

subject to the constraints that $\widehat{q}$ is block-wise constant and lies in the range $[\epsilon, 1]$. We take

$$\widetilde{q} = \operatorname*{arg\,max}_{q \in \mathbb{R}^d} \Bigg\{ \sum_k \left(\sum_{i \in \mathcal{B}_k} \mathbb{1}\{P_i > \tau\} \right) \log\left(q_k(1-\tau)\right) + \left(\sum_{i \in \mathcal{B}_k} \mathbb{1}\{P_i \leq \tau\} \right) \log\left(1 - q_k(1-\tau)\right) \\ : \epsilon \leq q_k \leq 1 \text{ for } k = 1, \ldots, d; \sum_k \frac{\sum_{i \in \mathcal{B}_k} \mathbb{1}\{P_i > \tau\}}{q_k(1-\tau)} \leq n \Bigg\},$$

and then define $\widehat{q} \in \mathbb{R}^n$ accordingly. (If we have $\sum_i \mathbb{1}\{P_i > \tau\} > n(1-\tau)$ then we simply set $\widehat{q} = \mathbf{1}_n$, as always.)

Note that, if each block satisfies

$$\frac{\sum_{i \in \mathcal{B}_k} \mathbb{1}\{P_i > \tau\}}{n_k(1-\tau)} \in [\epsilon, 1] \tag{3.7}$$

(that is, no block appears to have a proportion of nulls that is > 1 or $< \epsilon$), then the constrained maximum likelihood estimator is obtained by simply setting

$$\widetilde{q}_k = \frac{\sum_{i \in \mathcal{B}_k} \mathbb{1}\{P_i > \tau\}}{n_k(1-\tau)}$$

for each block $k = 1, \ldots, d$. This is equivalent to Storey [2002]'s method for estimating the proportion of nulls in a list of p-values, except applied separately to each block. When (3.7) is not satisfied, though, the optimization problem must be solved jointly over all blocks; in Section 3.5.4, we give an algorithm for this problem.

3.2.3 Low total variation

In some settings, the n hypotheses may exhibit some form of locally smooth or locally constant structure. For instance, if each hypothesis is associated with a spatial location, then it might be natural to assume that signals are spatially clustered, meaning that nearby hypotheses have equal or similar probabilities of being null or non-null. We might also have this sort of local similarity in other settings, for instance, similarity of individuals within a

data set.

To generalize this setting, consider a connected undirected graph $G = (V_G, E_G)$ on nodes $V_G = [n]$ with $e_G = |E_G|$ many undirected edges. We will search for a vector $\widehat{q}$ that is locally constant on this graph, meaning that $q_i - q_j = 0$ for most edges $(i, j) \in E_G$.

One choice for $\mathcal{Q}$ is given by

$$\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}} = \left\{ q \in [\epsilon, 1]^n : \sum_{(i,j)\in E} \mathbb{1}\{q_i \neq q_j\} \leq m \right\}, \tag{3.8}$$

where $\epsilon > 0$ bounds the values away from zero to avoid degeneracy, and m is some predetermined bound on the number of non-constant edges in the graph. However, $\mathcal{Q}_{\text{TV-sparse}}$ is not convex, and it may be difficult to choose a $\widehat{q}$ in this set. For this reason we may wish to consider a convex set,

$$\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1} = \left\{ q \in [\epsilon, 1]^n : \sum_{(i,j)\in E} |q_i - q_j| \leq m \right\}, \tag{3.9}$$

which is a strict relaxation of the set considered above—that is, $\mathcal{Q}_{\text{TV-sparse}} \subseteq \mathcal{Q}_{\text{TV-}\ell_1}$—since for $q \in [0, 1]^n$,

$$\sum_{(i,j)\in E} |q_i - q_j| \leq \sum_{(i,j)\in E} \mathbb{1}\{q_i \neq q_j\}.$$

In either case, the parameter m is a tuning parameter that should be specified in advance, perhaps reflecting prior knowledge about the amount of total variation in the underlying structure of the data; our theory will treat m as fixed rather than data-adaptive.

We will bound the relevant Rademacher complexities by making use of recent work by Hütter and Rigollet [2016]. First, following Hütter and Rigollet [2016], define the incidence matrix of the graph G, $D_G \in \{-1, 0, +1\}^{e_G \times n}$, which for each edge $(i, j) \in E_G$ has a corresponding row with ith entry $+1$, jth entry -1, and zeros elsewhere; define also the

quantity

$$\rho_G = \max_{k=1,\dots,e_G} \left\| (D_G^+)_k \right\|_2,$$

where $D_G^+ \in \mathbb{R}^{n \times e_G}$ is the pseudo-inverse of D_G and $(D_G^+)_k$ is its kth column.

Lemma 8. *For $\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}}$ as defined above,*

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \epsilon^{-1}\sqrt{n} + \epsilon^{-1} m \cdot \rho_G \cdot 2\sqrt{\log(n)}.$$

Therefore applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{\epsilon^{-1}\sqrt{n} + \epsilon^{-1} m \cdot \rho_G \cdot 2\sqrt{\log(n)}}{n} \right),$$

If we instead take $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$, then

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \epsilon^{-1}\sqrt{n} + \epsilon^{-2} m \cdot \rho_G \cdot 2\sqrt{\log(n)},$$

and thus

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{\epsilon^{-1}\sqrt{n} + \epsilon^{-2} m \cdot \rho_G \cdot 2\sqrt{\log(n)}}{n} \right),$$

For example, for a two-dimensional grid on an array of $\sqrt{n} \times \sqrt{n}$ nodes, it is shown in Hütter and Rigollet [2016, Proposition 4] that $\rho_G = \mathcal{O}\left(\sqrt{\log(n)}\right)$; therefore, we obtain nontrivial FDR guarantees for $\epsilon \gg \max\left\{\frac{1}{\sqrt{n}}, \frac{m\log(n)}{n}\right\}$ in the total variation sparsity setting ($\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}}$), and for $\epsilon \gg \sqrt{\frac{m\log(n)}{n}}$ in the total variation norm setting ($\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$). For the one-dimensional case, where G is a chain graph on n nodes, Hütter and Rigollet [2016, Section 3.1] calculate $\rho_G = \sqrt{n}$, and so to obtain a nontrivial FDR guarantee with $\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}}$ we need $\epsilon \gg \sqrt{\frac{m^2\log(n)}{n}}$; choosing $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$ we instead need $\epsilon \gg \sqrt[4]{\frac{m^2\log(n)}{n}}$.

To compare to methods such as those of Chouldechova [2014] and Sun et al. [2015],

which search for signals that appear in clusters (in a spatial domain or related setting), their work aims to make discoveries that form clusters and to measure error at the cluster-wise level as well; in contrast, our work sets the weights $\widehat{q}_i$ in a locally constant way so that the *threshold* for making a discovery is locally constant, but the p-values themselves may still lead to discoveries which do not form contiguous or well-defined clusters. In general, for our method, the pattern of discoveries that we might expect to see, would show some regions with high proportions of discoveries and other regions where this proportion is very sparse, with this gap being much wider than if we were to apply the BH procedure to the same data, since SABHA boosts our ability to make discoveries in high-signal regions. These different frameworks are related but are not directly comparable.

Choosing $\widehat{q} \in \mathcal{Q}$ As before, we can consider a constrained maximum likelihood approach to choosing $\widehat{q} \in \mathcal{Q}$, for either $\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}}$ or $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$:

$$\widehat{q} = \underset{q \in \mathbb{R}^n}{\arg\max} \Bigg\{ \sum_i \mathbb{1}\{P_i > \tau\} \log\left(q_i(1-\tau)\right) + \mathbb{1}\{P_i \leq \tau\} \log\left(1 - q_i(1-\tau)\right) \\ : q \in \mathcal{Q}, \sum_i \frac{\mathbb{1}\{P_i > \tau\}}{q_i(1-\tau)} \leq n \Bigg\}. \quad (3.10)$$

(If we have $\sum_i \mathbb{1}\{P_i > \tau\} > n(1-\tau)$ then we instead set $\widehat{q} = \mathbf{1}_n$, as always.) If we use $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$ then this is a convex optimization problem; we give an algorithm for this setting in Section 3.5.4. If instead we use $\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}}$ this is a highly nonconvex problem and may be very difficult to solve.

3.3 Proof of FDR control

In this section we will prove our main result, Theorem 5.

3.3.1 Complexity of a set: Rademacher width and cube width

We first develop a result on set complexity, which we will use in the proof of the theorem in Section 3.3.2. For any set $\mathcal{A} \subseteq \mathbb{R}^n$, recall that the Rademacher width of $\mathcal{A}$ is defined as

$$\omega_{\text{Rad}}(\mathcal{A}) = \mathbb{E}\left[\sup_{x \in \mathcal{A}} \langle x, B \rangle\right] \text{ where } B_i \overset{\text{iid}}{\sim} \text{Uniform}\{\pm 1\}.$$

We now define the "cube width", which is similar to the Rademacher width but uses arbitrary mean-zero product distributions on the cube $[-1, 1]^n$. First define $\mathfrak{C}$ to be the set of mean-zero product distributions on $[-1, 1]^n$, that is, each $\mathcal{D} \in \mathfrak{C}$ is a distribution of the form $\mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ where each $\mathcal{D}_i$ is a mean-zero distribution on $[-1, 1]$. Then define

$$\omega_{\text{cube}}(\mathcal{A}) = \sup_{\mathcal{D} \in \mathfrak{C}} \mathbb{E}_{Y \sim \mathcal{D}}\left[\sup_{x \in \mathcal{A}} \langle x, Y \rangle\right].$$

Note that, if we take each $\mathcal{D}_i$ to be the distribution placing probability 0.5 on $+1$ and on -1, then the Y_i's are independent Rademacher variables, i.e. Y has the same distribution as B above, and then we would have $\mathbb{E}_{Y \sim \mathcal{D}}[\sup_{x \in \mathcal{A}} \langle x, Y \rangle] = \omega_{\text{Rad}}(\mathcal{A})$; this proves that $\omega_{\text{cube}}(\mathcal{A}) \geq \omega_{\text{Rad}}(\mathcal{A})$. We now prove that the cube width is in fact equal to the Rademacher width—that is, the supremum over $\mathcal{D} \in \mathfrak{C}$ is attained by taking $\mathcal{D}_i = \text{Uniform}\{\pm 1\}$ for each i.

Lemma 9. *For any set $\mathcal{A} \subseteq \mathbb{R}^n$, the cube width satisfies $\omega_{\text{cube}}(\mathcal{A}) = \omega_{\text{Rad}}(\mathcal{A})$.*

Proof of Lemma 9. We know that $\omega_{\text{cube}}(\mathcal{A}) \geq \omega_{\text{Rad}}(\mathcal{A})$ trivially by choosing $\mathcal{D}_i = \text{Uniform}\{\pm 1\}$ for each i. Now we prove the reverse bound. Fix any $\mathcal{D} \in \mathfrak{C}$ and let $Y \sim \mathcal{D}$. Next, let $U_i \overset{\text{iid}}{\sim} \text{Uniform}[0, 1]$ be drawn independently of Y, and define $B \in \{\pm 1\}^n$ as

$$B_i = \begin{cases} +1, & U_i \leq \frac{1+Y_i}{2}, \\ -1, & U_i > \frac{1+Y_i}{2}. \end{cases}$$

Then we see that

$$\mathbb{E}\left[B_i \mid Y\right] = 1 \cdot \mathbb{P}\left\{U_i \leq \frac{1+Y_i}{2} \;\middle|\; Y\right\} + (-1) \cdot \mathbb{P}\left\{U_i > \frac{1+Y_i}{2} \;\middle|\; Y\right\} = Y_i,$$

and marginally, the B_i's are independent with

$$\mathbb{E}\left[B_i\right] = \mathbb{E}\left[\mathbb{E}\left[B_i \mid Y\right]\right] = \mathbb{E}\left[Y_i\right] = 0.$$

In other words, after marginalizing over Y, the B_i's are independent Rademacher variables.

We then have

$$\begin{aligned}
\mathbb{E}\left[\sup_{x\in\mathcal{A}} \langle x, Y\rangle\right] &= \mathbb{E}\left[\sup_{x\in\mathcal{A}} \langle x, \mathbb{E}\left[B \mid Y\right]\rangle\right] \leq \mathbb{E}\left[\mathbb{E}\left[\sup_{x\in\mathcal{A}} \langle x, B\rangle \;\middle|\; Y\right]\right] \quad \text{by Jensen's inequality} \\
&= \mathbb{E}\left[\sup_{x\in\mathcal{A}} \langle x, B\rangle\right] \quad \text{by the tower law of expectations} \\
&= \omega_{\mathrm{Rad}}(\mathcal{A}) \quad \text{by definition of Rademacher width.}
\end{aligned}$$

Since this holds for any choice $\mathcal{D} \in \mathfrak{C}$, this proves the desired bound, $\omega_{\mathrm{cube}}(\mathcal{A}) \leq \omega_{\mathrm{Rad}}(\mathcal{A})$.

□

3.3.2 Proof of FDR control

In this section we prove our main result, Theorem 5. Let $\mathcal{S} = \{i : P_i \leq \tau\}$, which is a random set. Below, we will prove that

$$\mathbb{E}\left[\mathsf{FDP} \mid \widehat{q}, \mathcal{S}\right] \leq \sum_{i\in\mathcal{H}_0} \frac{\alpha}{\widehat{q}_i n\tau} \cdot \mathbb{1}\{P_i \leq \tau\} \tag{3.11}$$

and that

$$\mathbb{E}\left[\sum_{i\in\mathcal{H}_0} \frac{\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i n\mathbb{P}\left\{P_i > \tau\right\}}\right] \leq 1 + \sqrt{\frac{\tau}{n(1-\tau)}}. \tag{3.12}$$

Taking these bounds as given for the time being, we are ready to bound the FDR. For

each $i \in \mathcal{H}_0$, define $a_i = \frac{\mathbb{P}\{P_i \le \tau\}}{\tau \cdot \mathbb{P}\{P_i > \tau\}}$. Note that

$$\frac{1+\tau a_i}{\tau} \cdot \mathbb{P}\left\{P_i \le \tau\right\} = a_i \text{ and } a_i \le \frac{1}{\mathbb{P}\left\{P_i > \tau\right\}}, \tag{3.13}$$

where the first part is a simple calculation from the definition of a_i while the second follows from the assumption that P_i is τ-conservative (3.1). Define also

$$Y_i = \begin{cases} (1-\tau)\cdot(1+\tau a_i)\cdot\left(\mathbb{1}\{P_i \le \tau\} - \mathbb{P}\left\{P_i \le \tau\right\}\right), & i \in \mathcal{H}_0, \\ 0, & i \notin \mathcal{H}_0. \end{cases}$$

We know that $a_i \le \frac{1}{1-\tau}$ by (3.1), meaning that $(1-\tau)\cdot(1+\tau a_i) \in [0,1]$. Thus we can see that $|Y_i| \le 1$ always, and trivially $\mathbb{E}[Y] = 0$. We then calculate

$$\begin{aligned}
&\alpha^{-1}\mathrm{FDR} = \alpha^{-1}\mathbb{E}\left[\mathsf{FDP}\right] = \alpha^{-1}\mathbb{E}\left[\mathbb{E}\left[\mathsf{FDP} \mid \widehat{q}, \mathcal{S}\right]\right] \\
&\le \mathbb{E}\left[\sum_{i\in\mathcal{H}_0} \frac{1}{\widehat{q}_i n\tau}\cdot\mathbb{1}\{P_i \le \tau\}\right] \quad \text{by applying (3.11)} \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{H}_0} \frac{a_i\cdot\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i n}\right] + \mathbb{E}\left[\sum_{i\in\mathcal{H}_0} \frac{\mathbb{1}\{P_i \le \tau\} - \tau a_i\cdot\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i n\tau}\right] \\
&\le \mathbb{E}\left[\sum_{i\in\mathcal{H}_0} \frac{\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i n\mathbb{P}\left\{P_i > \tau\right\}}\right] + \mathbb{E}\left[\sum_{i\in\mathcal{H}_0} \frac{1+\tau a_i}{\tau}\cdot\frac{\mathbb{1}\{P_i \le \tau\} - \mathbb{P}\left\{P_i \le \tau\right\}}{\widehat{q}_i n}\right] \quad \text{by (3.13)} \\
&= 1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)}\cdot\mathbb{E}\left[\sum_i \frac{Y_i}{\widehat{q}_i n}\right] \quad \text{by definition of } Y_i \text{ and by (3.12) above} \\
&\le 1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)}\cdot n^{-1}\cdot\mathbb{E}\left[\sup_{x\in\mathcal{Q}_{\text{inv}}}\langle x, Y\rangle\right] \\
&\quad \text{by definition of } \mathcal{Q}_{\text{inv}}\text{, since } \widehat{q}\in\mathcal{Q} \text{ by assumption (3.2)} \\
&\le 1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)}\cdot n^{-1}\cdot\omega_{\text{cube}}(\mathcal{Q}_{\text{inv}}),
\end{aligned}$$

where the last step holds by definition of the cube width, since Y lies in the cube $[-1,1]^n$

and is a mean-zero variable with independent components. Finally, by Lemma 9, we have $\omega_{\text{cube}}(\mathcal{Q}_{\text{inv}}) = \omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}})$. This proves the desired bound on FDR.

Now we turn to the proofs of the two main steps, (3.11) and (3.12).

Proof of (3.11) Fix any $i \in \mathcal{H}_0$ with $\mathbb{P}\{P_i \le \tau\} > 0$. We can define $k_i \in \{1, \dots, n\}$ which is the output of the SABHA procedure with the same $\widehat{q}$ and with P replaced with

$$P_{i\to 0} = (P_1, \dots, P_{i-1}, 0, P_{i+1}, \dots, P_n).$$

As is often used in proofs for the BH procedure (see e.g. Ferreira and Zwinderman [2006] for this type of proof technique), we observe that

$$P_i \text{ rejected } \Rightarrow \widehat{k} = k_i.$$

This implies that, if P_i is rejected, then P_i must satisfy $P_i \le \left(\frac{\alpha}{\widehat{q}_i} \cdot \frac{\widehat{k}}{n}\right) \wedge \tau = \left(\frac{\alpha}{\widehat{q}_i} \cdot \frac{k_i}{n}\right) \wedge \tau$. Now, conditioning on the event $\{P_i \le \tau\}$ and on $P_{i\to 0}$, we have

$$\mathbb{P}\{P_i \text{ rejected} \mid P_i \le \tau; P_{i\to 0}; \widehat{q}; \mathcal{S}\} = \mathbb{P}\left\{P_i \le \left(\frac{\alpha}{\widehat{q}_i} \cdot \frac{k_i}{n}\right) \wedge \tau \;\middle|\; P_i \le \tau; P_{i\to 0}; \widehat{q}; \mathcal{S}\right\} \le \frac{\frac{\alpha}{\widehat{q}_i} \cdot \frac{k_i}{n}}{\tau},$$

where the last step holds since P_i is τ-conservative (3.1) with $P_i \perp\!\!\!\perp P_{i\to 0}$, while $k_i, \widehat{q}, \mathcal{S}$ are functions of $P_{i\to 0}$ on the event $\{P_i \le \tau\}$. Therefore,

$$\mathbb{E}\left[\frac{\mathbb{1}\{P_i \text{ rejected}\}}{\widehat{k}} \;\middle|\; P_i \le \tau; P_{i\to 0}; \widehat{q}; \mathcal{S}\right] = \mathbb{E}\left[\frac{\mathbb{1}\{P_i \text{ rejected}\}}{k_i} \;\middle|\; P_i \le \tau; P_{i\to 0}; \widehat{q}; \mathcal{S}\right] \le \frac{\alpha}{\widehat{q}_i n \tau}.$$

Marginalizing over $P_{i\to 0}$, and using the fact that P_i cannot be rejected if $P_i > \tau$, we then have

$$\mathbb{E}\left[\frac{\mathbb{1}\{P_i \text{ rejected}\}}{\widehat{k}} \;\middle|\; \widehat{q}, \mathcal{S}\right] \le \frac{\alpha}{\widehat{q}_i n \tau} \cdot \mathbb{1}\{P_i \le \tau\}.$$

And, if $i \in \mathcal{H}_0$ with $\mathbb{P}\{P_i \le \tau\} = 0$, then the same bound holds trivially since P_i will never be rejected. Finally, to prove (3.11),

$$\mathbb{E}\left[\mathsf{FDP} \mid \widehat{q}, \mathcal{S}\right] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\{P_i \text{ rejected}\}}{\widehat{k}} \;\middle|\; \widehat{q}, \mathcal{S}\right] \le \sum_{i \in \mathcal{H}_0} \frac{\alpha}{\widehat{q}_i n \tau} \cdot \mathbb{1}\{P_i \le \tau\}.$$

Proof of (3.12) First, we have $\mathbb{P}\{P_i > \tau\} \ge 1 - \tau$ for all $i \in \mathcal{H}_0$ by (3.1), so we must have

$$\sum_{i \in \mathcal{H}_0} \frac{\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i n \mathbb{P}\{P_i > \tau\}} \le \sum_{i \in \mathcal{H}_0} \frac{\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i n (1-\tau)}.$$

Recall that, by our choice of $\widehat{q}$ satisfying (3.2), either this right-hand-side is bounded by 1 or we have $\widehat{q} = \mathbf{1}_n$. Therefore, taking the maximum of these two scenarios, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \in \mathcal{H}_0} \frac{\mathbb{1}\{P_i > \tau\}}{\widehat{q}_i n \mathbb{P}\{P_i > \tau\}}\right] &\le \mathbb{E}\left[\max\left\{1, \sum_{i \in \mathcal{H}_0} \frac{\mathbb{1}\{P_i > \tau\}}{n \mathbb{P}\{P_i > \tau\}}\right\}\right] \\
&\le 1 + \frac{1}{n}\mathbb{E}\left[\max\left\{0, \sum_{i \in \mathcal{H}_0}\left(\frac{\mathbb{1}\{P_i > \tau\}}{\mathbb{P}\{P_i > \tau\}} - 1\right)\right\}\right] \quad \text{since } |\mathcal{H}_0| \le n \\
&\le 1 + \frac{1}{n}\sqrt{\mathbb{E}\left[\left(\sum_{i \in \mathcal{H}_0}\left(\frac{\mathbb{1}\{P_i > \tau\}}{\mathbb{P}\{P_i > \tau\}} - 1\right)\right)^2\right]} \\
&= 1 + \frac{1}{n}\sqrt{\mathrm{Var}\left(\sum_{i \in \mathcal{H}_0} \frac{\mathbb{1}\{P_i > \tau\}}{\mathbb{P}\{P_i > \tau\}}\right)} \\
&= 1 + \frac{1}{n}\sqrt{\sum_{i \in \mathcal{H}_0} \frac{\mathbb{P}\{P_i \le \tau\}}{\mathbb{P}\{P_i > \tau\}}} \le 1 + \sqrt{\frac{\tau}{n(1-\tau)}},
\end{aligned}$$

where the last step holds since $\mathbb{P}\{P_i \le \tau\} \le \tau$ by (3.1), for each $i \in \mathcal{H}_0$. This proves (3.12), as desired.

3.4 Experiments

We now present empirical results on simulated and real data. For the simulated data, we will examine the low total variation setting described in Section 3.2.3, and in particular, will test the role of the total variation constraint on the power and FDR of the resulting method. Then, we will demonstrate applications of the other two types of structure described earlier on real data: ordered structure (as described in Section 3.2.1) with gene/drug response data, and grouped structure (as described in Section 3.2.2) with fMRI data. Code to reproduce all experiments is available online.[5]

3.4.1 Simulated data: low total variation

In our simulated experiments, the signals are arranged in a one-dimensional spatial array, with the pattern of signals and nulls exhibiting low total variation. Our estimate of $\widehat{q}$ uses the convex total variation norm as in (3.9). To generate the data, we create a list of $n = 500$ p-values, and assign a "true" underlying prior probability of each p-value being a null,

$$q = (\underbrace{0.1, \ldots, 0.1}_{250 \text{ times}}, \underbrace{0.9, \ldots, 0.9}_{250 \text{ times}}).$$

To generate the p-values themselves, we first draw $Z_i \overset{\perp\!\!\!\perp}{\sim} N(\mu_i, 1)$ where $\mu_i = 0$ for the nulls and $\mu_i = \mu_{\text{sig}} > 0$ for the non-nulls, for $\mu_{\text{sig}} \in \{0.5, 1, 1.5, \ldots, 3.5\}$ (with larger μ_{sig} indicating a stronger signal). Then we run two-sided z-tests, $P_i = 2(1 - \Phi(|Z_i|))$, where Φ is the CDF of the standard normal.

We implement SABHA on the one-dimensional chain graph, choosing $\tau = 0.5$ and defining $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$ as in (3.9) with the lower bound parameter set at $\epsilon = 0.1$ and with the total variation constraint $m \in \{0.25, 2, 20\}$. (Note that, for the "true" q, we have $\|q\|_{\text{TV-}\ell_1} = \sum_{i=1}^{n-1} |q_i - q_{i+1}| = 0.8$, that is, $q \in \mathcal{Q}$ for the two larger values of m.) We then compare

5. Available at `http://www.stat.uchicago.edu/~rina/sabha.html`.

SABHA with BH, and also with Storey's modification of BH ("Storey-BH") given in (1.1), implemented with parameter $\tau = 0.5$. Based on the data generating mechanism, we expect Storey-BH to estimate the proportion of nulls as $\widehat{\pi}_0 \approx 0.5$ when the signal is strong, and as $\widehat{\pi}_0 \approx 1$ when the signal is weak.

To fit $\widehat{q}$ for the SABHA method, i.e. to solve the optimization problem (3.10), details are given in Section 3.5.4. For comparison we also run an "oracle" version of SABHA where $\widehat{q} = q$, the true vector of prior probabilities of each p-value being a null. For all methods we choose the target FDR level $\alpha = 0.1$.

We then compare the observed FDR and power for each of the considered methods. As we can see in the left panel of Figure 3.1, for all methods, the average power and average observed FDR both increase with larger μ (stronger signal). The BH, Storey-BH, and oracle SABHA methods all control FDR at level $\alpha = 0.1$; the BH is in fact known to control FDR at the level $\alpha \cdot \frac{|\mathcal{H}_0|}{n}$, and is therefore quite conservative (i.e. FDR < 0.1) as we observe in our results, while the other two methods have FDR ≈ 0.1. For SABHA, as the signal grows stronger, the observed FDR grows very close to 0.1 for small $m = 0.25$ and medium $m = 2$; for large $m = 20$, the observed FDR exceeds the target level $\alpha = 0.1$, as $\widehat{q}$ is overfitting to the p-values. To compare the power, BH is most conservative (lowest power) followed by Storey-BH. Oracle SABHA shows the highest power, while SABHA shows power increasing as the total variation constraint m increases. Here, SABHA is consistently more powerful than BH and Storey-BH over the range of μ_{sig}.

In the right panel of Figure 3.1, we further compare the methods according to their estimated probabilities of a null, for one trial at signal strength $\mu_{\text{sig}} = 2.5$. For SABHA with $m = 0.25, 2, 20$, the plot displays the estimated $\widehat{q}$. We can compare against the true q, which is also the input to the oracle SABHA method. Finally, since the BH and Storey-BH methods are equivalent to taking $\widehat{q} = \mathbf{1}_n$ and $\widehat{q} = \widehat{\pi}_0 \cdot \mathbf{1}_n$ and proceeding with SABHA, we display these estimated $\widehat{q}$'s as well. As expected, for SABHA, we see that for $m = 0.25$ the method is unable to fit $\widehat{q}$ closely to the true q, as the true q has total variation norm

0.8, which is substantially larger. For $m = 2$ the fit is substantially better but becomes less locally smooth; for $m = 20$ the fit is extremely jagged and we see evidence of extreme overfitting (leading to the increased FDR observed previously).

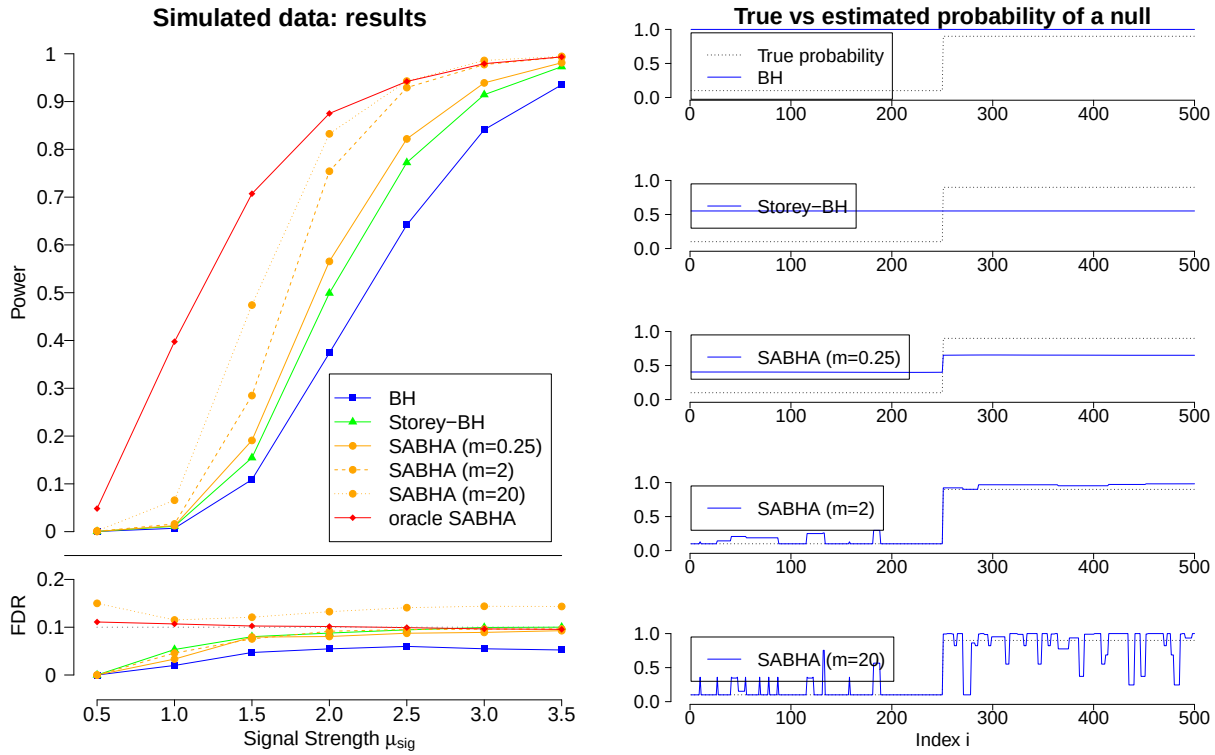

Figure 3.1: Power and observed FDR level of all procedures averaged over 10 trials (left), and true q vs. estimated $\widehat{q}$ for a single trial with $\mu_{\text{sig}} = 2.5$ (right); see Section 3.4.1 for details. The target FDR level is $\alpha = 0.1$.

3.4.2 Gene/drug response data: ordered structure

We next apply SABHA to an ordered testing problem, the gene dosage response application described in chapter 2 and Li and Barber [2016a], for identifying the response of $n = 22283$ genes' expression to drug dosages. As a reminder, for each gene $i = 1, \ldots, 22283$, we calculate a p-value P_i which evaluates evidence for a change in gene expression level between the control (no dose) setting and a low dose setting, using a one-sided permutation test. The genes are arranged in order, with the beginning of the list (low index i) corresponding to genes whose differential expression level at *high* dosage leads us to believe that this gene is more likely to

exhibit a nonzero response at *low* dosage; we run a one-sided test, for instance, a gene that showed increased expression at the high dosage is tested for increased expression at the low dosage. See chapter 2 and Li and Barber [2016a] for more details on the data set and on how the ordering and the p-values are calculated. While the data is not independent across the n genes, we treat it as independent for the purpose of the analysis.

We run SABHA using $\mathcal{Q} = \mathcal{Q}_{\text{ord}}$ (3.3) and choose $\widehat{q}$ to be a step function as in (3.4), with lower bound parameter $\epsilon = 0.1$ and threshold $\tau = 0.5$. We compare SABHA against BH, and against Storey-BH with threshold $\tau = 0.5$. We also compare against three methods for ordered hypothesis testing: the ForwardStop method G'Sell et al. [2015], the SeqStep method Barber and Candès [2015] (with parameter $C = 2$), the accumulation test method with the HingeExp function Li and Barber [2016a] (parameter $C = 2$), and the Adaptive SeqStep method Lei and Fithian [2016]; see Section 3.2.1 for some more details on this family of methods. We run each method with target FDR level $\alpha = 0.01, 0.02, \ldots, 0.5$.

In this data set, the sample size (number of measurements for each gene) at each of the three settings—control, low dose, and high dose—is 5. We repeat our experiment three times: first, using this full sample size; second, using a sample size of 2 for the high dose setting only, so as to reduce the quality of the ordering (that is, we will be less accurate in our assessment of which genes are most likely to show a change in expression level at the low dosage); and third, using a random ordering (that is, there is no information contained in the ordering; true signals are equally likely to appear anywhere on the list).

Results Figure 3.2 shows the results from all the methods across the range of α values, for all three settings for how the ordering is determined. The plotted outcome is the number of "discoveries", i.e. the number of genes selected as showing a significant difference between the low dose and no dose setting. Both the BH and Storey-BH methods, which do not use the information of the ordering, are not able to make more than a few discoveries at any level α below ≈ 0.3; since these methods do not use the ordering information, their outcomes are

identical across the three settings. Turning to the ordered testing methods (accumulation test/HingeExp, SeqStep, ForwardStop, Adaptive SeqStep), we see that these methods are able to recover a substantial number of genes even for low α when the ordering is highly informative, with a wide range of performance across the three methods, but make essentially no discoveries when the ordering is random (carries no information).

Next, we see that SABHA is able to perform well across the range of scenarios; while it's not the optimal method in any single scenario, it is the only method whose performance is adaptive while the existing methods are all specialized to one or the other extreme. For the first setting (sample size 5 used to obtain a highly informative ordering ordering), we see that SABHA is less powerful than the best ordered testing method but nonetheless gives strong performance even at low α values; for the second setting (sample size 2 used to obtain a moderately informative ordering), SABHA is now more comparable to the best ordered testing method across much of the range of α values; and for the third setting (a random i.e. completely uninformative ordering), SABHA continues to perform well, and is nearly as powerful as Storey-BH which makes the most discoveries in this setting, while the ordered testing methods now have effectively zero power except for Adaptive SeqStep at the high values of α. To summarize, SABHA is able to adapt to the amount of information carried in the ordering, achieving good performance relative to the best method in each setting. In practice, then, when we do not know ahead of time whether the ordered structure is informative or not, SABHA is a good choice as it can adapt to any scenarios.

3.4.3 fMRI data: grouped structure

We now show an application of our method in an analysis of fMRI data, in which the block structure of the problem can be exploited (see Section 3.2.2 for more details on the block structured setting). The data, gathered by Keller et al. [2001] and available online,[6] consists of fMRI measurements taken for subjects who are shown two stimuli in a sequence, a picture

6. Data available at `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/`

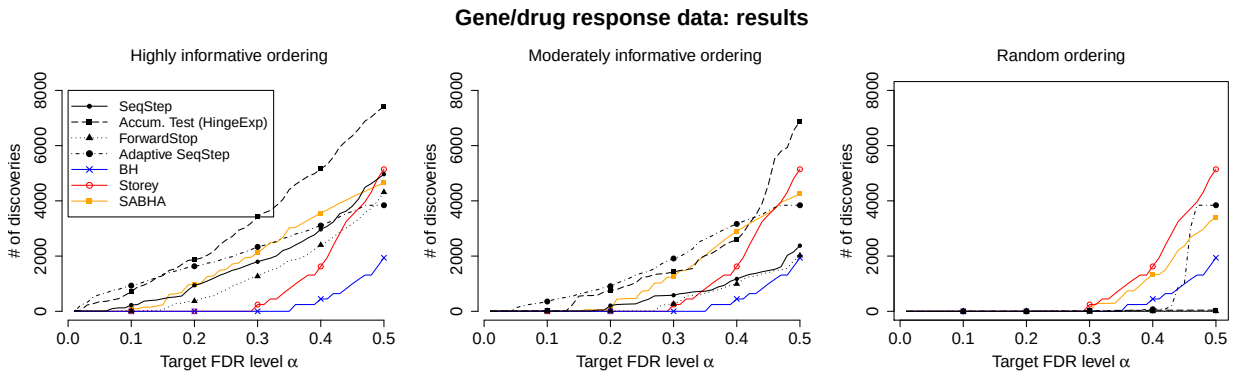

Figure 3.2: Results for the differential gene expression experiment in SABHA: for each method, the plot shows the number of discoveries made (i.e. the number of genes selected as showing significant change in expression at the low drug dosage as compared to the control), at a range of target FDR values α.

and a sentence in either order, with the task of determining whether the two agree. We restrict our attention to data from a single individual (Subject 04867), and to the 20 trials with the picture displayed first and the sentence second, where the procedure is of the form

$$\underbrace{\text{Picture (4 sec)} \;;\; \text{No stimulus (4 sec)}}_{\text{Picture phase (8 sec)}} \;;\; \underbrace{\text{Sentence (4 sec)} \;;\; \text{No stimulus (4 sec)}}_{\text{Sentence phase (8 sec)}}.$$

Measurements are taken every 0.5 seconds, for a total of 16 brain images each in the picture phase and in the sentence phase. Though only a fraction of the brain of each subject was imaged, these images reflect 3D activity information, as each images includes 8 two-dimensional slices. The 4691 measured voxels are grouped into 24 regions of interest (ROIs), anatomically or functionally distinct regions in the brain.[7]

We are interested in how the activity of different parts of the brain differs between the picture phase and the sentence phase, which can be formulated as a multiple testing problem, with each voxel in the brain corresponding to a hypothesis. For each voxel we compute the average activity level across the 16 images in each of the two phases, for each of the 20 trials, which leads to a p-value computed from a paired t-test with sample size 20. In the first few

7. While 4698 voxels are measured and 25 ROIs are defined, 7 voxels are not labeled with a ROI, and one ROI is not assigned to any voxels, so we remove these from the data.

columns of Figure 3.3, we display the partition into ROIs (Figure 3.3(a)), the original data averaged for the picture phase and the sentence phase (Figure 3.3(b)), and the calculated p-values (Figure 3.3(c)). Each column is a single 3D image of the brain, split into eight horizontal slices.

We then implement the SABHA method with block structure determined by the ROIs; we use $\mathcal{Q} = \mathcal{Q}_{\text{block}}$ as in (3.6), and choose parameters $\tau = 0.5$ and $\epsilon = 0.1$. (Recall that, for the block structure setting, this method is an example of the Group Benjamini-Hochberg method proposed by Hu et al. [2012], where the proportion of nulls within each group can be estimated with any data-adaptive method; here we specifically use our proposed estimate $\widehat{q}$ given in (3.6).) We compare against BH and against Storey-BH with threshold $\tau = 0.5$. For all methods we use target FDR level $\alpha = 0.2$.

Results The results from the fMRI experiment are displayed in Figure 3.3. The number of discoveries made by each method is:

Method	# discoveries
BH	931
Storey-BH	1217
SABHA	1234

In Figure 3.3(d) we see the estimated $\widehat{q}$ for the SABHA method, and Figure 3.3(e) displays the locations of the discoveries for each of the three methods.

To understand the performance of SABHA in greater detail, consider the estimated $\widehat{q}$ in Figure 3.3(d); we see that some of the ROIs are estimated to have a much lower proportion of nulls (those ROIs that appear darker in the figure), and these are the regions that show the greatest gains in the number of discoveries made by SABHA as compared to the other methods. To see the difference more quantitatively, in Figure 3.4 we display the proportion of discoveries made in each ROI by each of the three methods, compared to the estimated value for $\widehat{q}$ in this ROI. As expected, the greatest gains for SABHA are in those ROIs where $\widehat{q}$

is estimated to be lowest. In contrast, in ROIs where $\widehat{q}$ is estimated to be near 1, Storey-BH makes more discoveries. This is because Storey-BH effectively estimates a uniform (constant) $\widehat{q}$ across all ROIs (i.e. $\widehat{q} = \widehat{\pi}_0 \mathbf{1}_n$ where $\widehat{\pi}_0$ is the *overall* estimated proportion of nulls), while for SABHA this estimated proportion varies across ROIs, and will thus be lower for some ROIs and higher for others. Overall, using the block structure allows for more discoveries—and may perhaps be more accurate as it uses information that is more locally relevant for each ROI, although of course we cannot assess this without knowing the "ground truth".

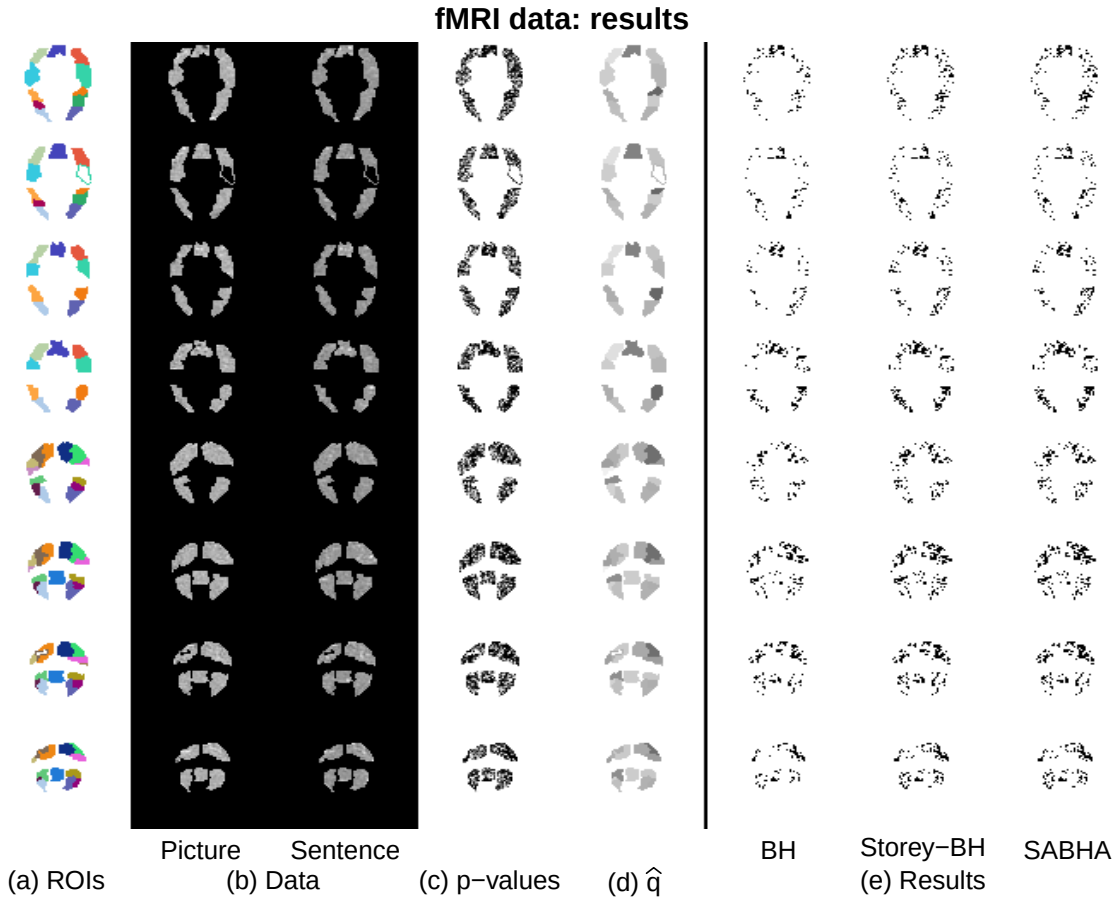

Figure 3.3: Results for the fMRI data; the eight images in each column are the horizontal slices of the brain. (a) The 24 ROIs defined in the data set, each pictured in a different color. (b) The average activity levels recorded in the experiment for the picture phase and for the sentence phase (white = highest activity level, black = zero activity level). (c) The p-values obtained at each voxel using the paired t-test (black = 0 = most significant, white = 1 = least significant). (d) The estimated vector $\widehat{q}$ for SABHA using the block structure defined by the ROIs (black = 0 = contains all signals, white = 1 = contains all nulls). (e) Results for the BH, Storey-BH, and SABHA methods (a black point indicates a voxel labeled as a discovery).

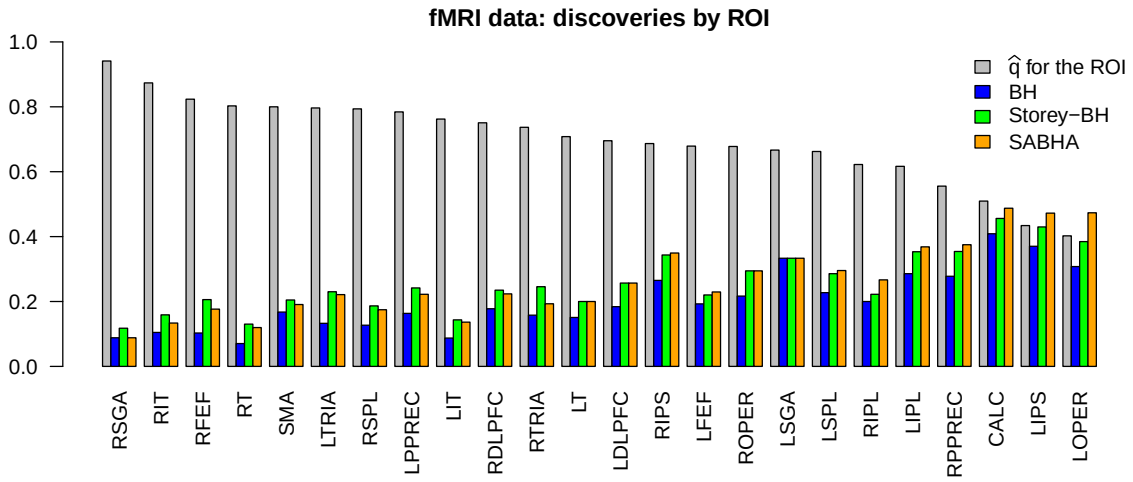

Figure 3.4: Proportion of discoveries within each ROI for the BH, Storey-BH, and SABHA methods, compared to the estimated proportion of nulls, $\widehat{q}$. The ROIs are sorted in decreasing order of the estimated $\widehat{q}$.

3.5 Proofs and technical details

In this section we prove the Rademacher complexity bounds for the three settings considered in Section 3.2.

3.5.1 Ordered structure

Recall that for ordered (i.e. monotone) vectors $\widehat{q}$ we defined the set

$$\mathcal{Q} = \mathcal{Q}_{\text{ord}} = \{q : \epsilon \leq q_1 \leq \cdots \leq q_n \leq 1\}.$$

Lemma 6. For $\mathcal{Q} = \mathcal{Q}_{\text{ord}}$ as defined above,

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq (\epsilon^{-1} - 1) \cdot \sqrt{n}.$$

Therefore, applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{\tau}{1-\tau}} + \frac{\epsilon^{-1} - 1}{\tau(1-\tau)}\right]\right),$$

Proof of Lemma 6. For this choice of $\mathcal{Q}$, we have

$$\mathcal{Q}_{\text{inv}} = \{x : 1 \leq x_n \leq \cdots \leq x_1 \leq \epsilon^{-1}\},$$

which is the convex hull of the set

$$\mathcal{A} = \{x^0, \ldots, x^n\} \text{ where } x^k = (\underbrace{\epsilon^{-1}, \ldots, \epsilon^{-1}}_{k \text{ times}}, \underbrace{1, \ldots, 1}_{n-k \text{ times}}).$$

In this setting, since Rademacher width is not increased by taking a convex hull, we have (for $B_i \overset{\text{iid}}{\sim} \text{Uniform}\{\pm 1\}$)

$$\begin{aligned}\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) = \omega_{\text{Rad}}(\mathcal{A}) &= \mathbb{E}\left[\max_{k=0,\ldots,n} \langle x^k, B\rangle\right] = \mathbb{E}\left[\max_{k=0,\ldots,n} \epsilon^{-1}\sum_{i=1}^{k} B_i + \sum_{i=k+1}^{n} B_i\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} B_i\right] + (\epsilon^{-1}-1)\cdot\mathbb{E}\left[\max_{k=0,\ldots,n}\sum_{i=1}^{k} B_i\right] = (\epsilon^{-1}-1)\cdot\mathbb{E}\left[\max_{k=0,\ldots,n}\sum_{i=1}^{k} B_i\right],\end{aligned}$$

since $\mathbb{E}[B_i] = 0$. This is the expected maximum of a simple random walk, and it is known that

$$\mathbb{E}\left[\max_{k=0,\ldots,n}\sum_{i=1}^{k} B_i\right] = \sum_{k=1}^{n} k^{-1}\mathbb{E}\left[\max\left\{0, \sum_{i=1}^{k} B_i\right\}\right]$$

by the Pollaczek-Spitzer identity (for example this equality is implied by Borovkov [1999, Section 11.8, Theorem 7]). For each $k = 1, \ldots, n$, we have

$$\mathbb{E}\left[\max\left\{0, \sum_{i=1}^{k} B_i\right\}\right] = \frac{1}{2}\mathbb{E}\left[\left|\sum_{i=1}^{k} B_i\right|\right] \leq \frac{1}{2}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{k} B_i\right)^2\right]} = \frac{1}{2}\sqrt{\text{Var}\left(\sum_{i=1}^{k} B_i\right)} = \frac{\sqrt{k}}{2},$$

where the first step holds since the distribution of $\sum_{i=1}^{k} B_i$ is symmetric. So,

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq (\epsilon^{-1}-1)\cdot\sum_{k=1}^{n} k^{-1}\cdot\frac{\sqrt{k}}{2} \leq (\epsilon^{-1}-1)\cdot\sqrt{n}.$$

□

3.5.2 Block structure

Recall that we defined

$$\mathcal{Q} = \mathcal{Q}_{\text{block}} = \left\{q : q_i \geq \epsilon \text{ for all } i, \text{ and } q_i = q_j \text{ whenever } i, j \text{ are in the same block}\right\}.$$

Lemma 7. For $\mathcal{Q} = \mathcal{Q}_{\text{block}}$ as defined above, if the blocks are of sizes $n_1, \ldots, n_d$, then

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \frac{\epsilon^{-1} - 1}{2} \sum_{i=1}^{d} \sqrt{n_i}.$$

Therefore, applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{\epsilon^{-1} - 1}{2} \cdot \frac{\sum_{i=1}^{d} \sqrt{n_i}}{n}\right),$$

Proof of Lemma 7. In this setting, each $q \in \mathcal{Q}$ is uniquely defined by choosing the constant value inside of each block, and so taking $B_i \overset{\text{iid}}{\sim} \text{Uniform}\{\pm 1\}$,

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) = \mathbb{E}\left[\sup_{x \in \mathcal{Q}_{\text{inv}}} \langle x, B\rangle\right] = \mathbb{E}\left[\sup_{y \in [1, \epsilon^{-1}]^d} \sum_{i=1}^{d} \sum_{j \in \mathcal{B}_i} y_i B_j\right] = \sum_{i=1}^{d} \mathbb{E}\left[\sup_{1 \leq y \leq \epsilon^{-1}} y \sum_{j \in \mathcal{B}_i} B_j\right]$$

$$= \sum_{i=1}^{d} \mathbb{E}\left[\sum_{j \in \mathcal{B}_i} B_j\right] + \sum_{i=1}^{d} \mathbb{E}\left[\sup_{0 \leq y \leq \epsilon^{-1} - 1} y \cdot \sum_{j \in \mathcal{B}_i} B_j\right] = \sum_{i=1}^{d} (\epsilon^{-1} - 1) \cdot \mathbb{E}\left[\max\left\{0, \sum_{j \in \mathcal{B}_i} B_j\right\}\right]$$

$$\overset{(*)}{=} \sum_{i=1}^{d} (\epsilon^{-1} - 1) \cdot \frac{1}{2} \mathbb{E}\left[\left|\sum_{j \in \mathcal{B}_i} B_j\right|\right] \leq \frac{\epsilon^{-1} - 1}{2} \sum_{i=1}^{d} \sqrt{\mathbb{E}\left[\left(\sum_{j \in \mathcal{B}_i} B_j\right)^2\right]}$$

$$= \frac{\epsilon^{-1} - 1}{2} \sum_{i=1}^{d} \sqrt{\text{Var}\left(\sum_{j \in \mathcal{B}_i} B_j\right)} = \frac{\epsilon^{-1} - 1}{2} \sum_{i=1}^{d} \sqrt{n_i},$$

where the step marked (*) holds since the distribution of $\sum_{j \in \mathcal{B}_i} B_j$ is symmetric. □

3.5.3 Low total variation

Recall that, for the low total variation setting, we considered two choices for $\mathcal{Q}$:

$$\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}} = \left\{ q \in [\epsilon, 1]^n : \sum_{(i,j)\in E} \mathbb{1}\{q_i \neq q_j\} \leq m \right\},$$

and

$$\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1} = \left\{ q \in [\epsilon, 1]^n : \sum_{(i,j)\in E} |q_i - q_j| \leq m \right\}.$$

Lemma 8. For $\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}}$ as defined above,

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \epsilon^{-1}\sqrt{n} + \epsilon^{-1} m \cdot \rho_G \cdot 2\sqrt{\log(n)}.$$

Therefore applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{\epsilon^{-1}\sqrt{n} + \epsilon^{-1} m \cdot \rho_G \cdot 2\sqrt{\log(n)}}{n} \right),$$

If we instead take $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$, then

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \epsilon^{-1}\sqrt{n} + \epsilon^{-2} m \cdot \rho_G \cdot 2\sqrt{\log(n)},$$

and thus

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{\tau}{n(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{\epsilon^{-1}\sqrt{n} + \epsilon^{-2} m \cdot \rho_G \cdot 2\sqrt{\log(n)}}{n} \right),$$

Proof of Lemma 8. First, for any $t_{\ell_2}, t_{\text{TV}} > 0$, define

$$\mathcal{T}_G(t_{\ell_2}, t_{\text{TV}}) = \{x \in \mathbb{R}^n : \|x\|_2 \leq t_{\ell_2}, \sum_{(i,j)\in E_G} |x_i - x_j| \leq t_{\text{TV}}\}.$$

Now take $B_i \overset{\text{iid}}{\sim} \text{Uniform}\{\pm 1\}$. Hütter and Rigollet [2016, Appendix B.1] calculates that

$$\max_{x \in \mathcal{T}_G(t_{\ell_2}, t_{\text{TV}})} \langle B, x \rangle \leq t_{\ell_2} \cdot \left\| \mathcal{P}^{\perp}_{\text{span}(D_G)}(B) \right\|_2 + t_{\text{TV}} \cdot \max_{k=1,\ldots,e_G} \left| (D_G^+)_k^\top B \right|,$$

where $\text{span}(D_G) \subseteq \mathbb{R}^n$ has co-dimension 1. Let u be a unit vector orthogonal to $\text{span}(D_G)$. Then

$$\mathbb{E}\left[\left\| \mathcal{P}^{\perp}_{\text{span}(D_G)}(B) \right\|_2 \right] = \mathbb{E}\left[|u^\top B| \right] \leq \sqrt{\mathbb{E}\left[(u^\top B)^2 \right]} = 1.$$

Next, since B is a subgaussian random vector with scale 1, $\max_{k=1,\ldots,e_G} \left| (D_G^+)_k^\top B \right|$ is the maximum of $e_G \leq \frac{n^2}{2}$ many centered subgaussian random variables each with scale bounded by $\rho_G = \max_k \left\| (D_G^+)_k \right\|_2$. This proves that

$$\omega_{\text{Rad}}(\mathcal{T}_G(t_{\ell_2}, t_{\text{TV}})) = \mathbb{E}\left[\max_{x \in \mathcal{T}_G(t_{\ell_2}, t_{\text{TV}})} \langle B, x \rangle \right] \leq t_{\ell_2} + t_{\text{TV}} \cdot \rho_G \cdot 2\sqrt{\log(n)}. \tag{3.14}$$

We now analyze the complexity of $\mathcal{Q}_{\text{inv}}$ in this setting. First take $\mathcal{Q} = \mathcal{Q}_{\text{TV-sparse}}$ and consider any $x \in \mathcal{Q}_{\text{inv}}$. Then

$$\sum_{(i,j) \in E} |x_i - x_j| = \sum_{(i,j) \in E} |(q_i)^{-1} - (q_j)^{-1}| \leq \epsilon^{-1} m,$$

since $|(q_i)^{-1} - (q_j)^{-1}| = 0$ whenever $q_i = q_j$ (i.e. for all but m edges $(i,j) \in E$), and $|(q_i)^{-1} - (q_j)^{-1}| \leq \epsilon^{-1}$ for all i, j. We also have $\|x\|_2 \leq \epsilon^{-1}\sqrt{n}$, and so

$$\mathcal{Q}_{\text{inv}} \subseteq \mathcal{T}_G(\epsilon^{-1}\sqrt{n}, \epsilon^{-1} m).$$

Applying (3.14), then,

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \epsilon^{-1}\sqrt{n} + \epsilon^{-1} m \cdot \rho_G \cdot 2\sqrt{\log(n)},$$

where ρ_G is defined as before.

Next consider $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$ In this case, $\mathcal{Q}_{\text{inv}}$ is actually substantially more complex: for any $x \in \mathcal{Q}_{\text{inv}}$, we again have $\|x\|_2 \leq \epsilon^{-1}\sqrt{n}$, but our bound on the total variation norm is larger:

$$\sum_{(i,j)\in E} |x_i - x_j| = \sum_{(i,j)\in E} |(q_i)^{-1} - (q_j)^{-1}| = \sum_{(i,j)\in E} \frac{|q_i - q_j|}{q_i q_j} \leq \epsilon^{-2} \sum_{(i,j)\in E} |q_i - q_j| \leq \epsilon^{-2} m.$$

In this setting, then,

$$\mathcal{Q}_{\text{inv}} \subseteq \mathcal{T}_G(\epsilon^{-1}\sqrt{n}, \epsilon^{-2}m).$$

Again applying (3.14),

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \epsilon^{-1}\sqrt{n} + \epsilon^{-2} m \cdot \rho_G \cdot 2\sqrt{\log(n)}.$$

□

3.5.4 Choosing $\widehat{q}$ via constrained maximum likelihood

In the different types of structure that we consider in Section 3.2, a common strategy for choosing $\widehat{q}$ is via the optimization problem

$$\widehat{q} = \arg\max_{q \in \mathbb{R}^n} \Bigg\{ \sum_i \mathbb{1}\{P_i > \tau\} \log\left(q_i(1-\tau)\right) + \mathbb{1}\{P_i \leq \tau\} \log\left(1 - q_i(1-\tau)\right) \\ : q \in \mathcal{Q}, \sum_i \frac{\mathbb{1}\{P_i > \tau\}}{q_i(1-\tau)} \leq n \Bigg\}. \quad (3.15)$$

(As before, if $\sum_i \mathbb{1}\{P_i > \tau\} > n(1-\tau)$ then we would instead set $\widehat{q} = \mathbf{1}_n$; from this point on, we proceed under the assumption that this is not the case.)

This is a constrained maximum likelihood problem, which is a convex optimization problem as long as $\mathcal{Q}$ is convex. We now give a general algorithm for this problem, implementing the Alternating Direction Method of Multipliers (ADMM) Boyd et al. [2011]. To implement

the ADMM method, we assume that the set $\mathcal{Q}$ can be characterized as follows:

$$\mathcal{Q} = \{q : Mq \in \mathcal{M}, \epsilon \leq q \leq 1\}$$

for some $\epsilon \geq 0$, some fixed matrix $M \in \mathbb{R}^{m \times n}$, and some convex set $\mathcal{M} \subseteq \mathbb{R}^m$ which has an easy-to-compute projection operator, i.e.

$$\text{Proj}_{\mathcal{M}}(z) = \underset{x \in \mathcal{M}}{\arg\min}\{\|x - z\|_2\}$$

is simple to compute for any $z \in \mathbb{R}^m$.

To make this concrete,

- For the ordered setting considered in Section 3.2.1, with $\mathcal{Q} = \mathcal{Q}_{\text{ord}}$, we take $\mathcal{M} = \{q : q_1 \leq \cdots \leq q_n\}$, and $M = \mathbf{I}_n$. The relevant projection operator can be computed via the Pool Adjacent Violators Algorithm (PAVA) Barlow et al. [1972].

- For the block-wise constant setting considered in Section 3.2.2, with $\mathcal{Q} = \mathcal{Q}_{\text{block}}$, we take $\mathcal{M} = \left\{q : q_i = q_j \text{ whenever } i, j \text{ are in the same block}\right\}$, and $M = \mathbf{I}_n$. The projection operator is very easy to compute: we simply take the average value within each block.

- For the bounded total variation norm setting considered in Section 3.2.3, suppose we want to work with $\mathcal{Q} = \mathcal{Q}_{\text{TV-}\ell_1}$, which is a convex set. Define

  $$M = D_G \in \{-1, 0, +1\}^{e_G \times n},$$

  the edge incidence matrix of the graph G defined in Section 3.2.3. Then define

  $$\mathcal{M} = \{z \in \mathbb{R}^{e_G} : \|z\|_1 \leq m\},$$

  a rescaled ℓ_1 unit ball; it is clear that $\mathcal{Q} = \{q : Mq \in \mathcal{M}, \epsilon \leq q \leq 1\}$, as desired. In

this case, projection to $\mathcal{M}$ can be computed via soft thresholding.

ADMM algorithm implementation

We now implement the algorithm as follows. First, our optimization problem is equivalent to calculating

$$\min_{q\in[\epsilon,1]^n, x\in\mathbb{R}^m, y\in\mathbb{R}^n} \Bigg\{ -\sum_i \left[\mathbb{1}\{P_i > \tau\} \log\left(q_i(1-\tau)\right) + \mathbb{1}\{P_i \le \tau\} \log\left(1 - q_i(1-\tau)\right)\right] \\ : x \in \mathcal{M}, \sum_i \frac{\mathbb{1}\{P_i > \tau\}}{y_i(1-\tau)} \le n, Mq = x, q = y \Bigg\}, \quad (3.16)$$

which then becomes

$$\min_{q\in[\epsilon,1]^n, x\in\mathbb{R}^m, y\in\mathbb{R}^n} \max_{u\in\mathbb{R}^m, v\in\mathbb{R}^n} L(q,x,y,u,v)$$

where

$$L(q,x,y,u,v) = \Bigg\{ -\sum_i \left[\mathbb{1}\{P_i > \tau\} \log\left(q_i(1-\tau)\right) + \mathbb{1}\{P_i \le \tau\} \log\left(1 - q_i(1-\tau)\right)\right] \\ +\delta(x \in \mathcal{M}) + \delta\left(\sum_i \frac{\mathbb{1}\{P_i > \tau\}}{y_i(1-\tau)} \le n\right) + \langle u, Mq - x\rangle + \frac{\alpha}{2}\|Mq - x\|_2^2 + \langle v, q - y\rangle + \frac{\beta}{2}\|q - y\|_2^2 \Bigg\},$$

where $\delta(\cdot)$ is the convex indicator function and where $\alpha, \beta > 0$ are fixed parameters.

The ADMM algorithm then iterates the steps:

$$\begin{cases} q_{t+1} = \arg\min_{q\in[\epsilon,1]^n}\{L(q, x_t, y_t, u_t, v_t)\} \\ (x_{t+1}, y_{t+1}) = \arg\min_{(x,y)\in\mathbb{R}^m\times\mathbb{R}^n}\{L(q_{t+1}, x, y, u_t, v_t)\} \\ u_{t+1} = u_t + \alpha(Mq_{t+1} - x_{t+1}),\ v_{t+1} = v_t + \beta(q_{t+1} - y_{t+1}) \end{cases}$$

Now we calculate formulas for the q, x, and y updates.

The q update step First, we modify the q update step slightly: we add a preconditioning term to the q update for easier computation,

$$q_{t+1} = \underset{q\in[\epsilon,1]^n}{\arg\min} \left\{ L(q, x_t, y_t, u_t, v_t) + \frac{\alpha}{2}(q - q_t)^\top(\eta \mathbf{I} - M^\top M)(q - q_t) \right\},$$

where $\eta \geq \|M\|^2$. Rearranging some terms, we are minimizing

$$-\sum_i \left[\mathbb{1}\{P_i > \tau\} \log\left(q_i(1-\tau)\right) + \mathbb{1}\{P_i \leq \tau\} \log\left(1 - q_i(1-\tau)\right)\right] + \frac{\alpha\eta + \beta}{2} \|q - w\|_2^2$$

over $q \in [\epsilon, 1]^n$, where w is the vector with entries

$$w = -\frac{M^\top(u_t + \alpha(Mq_t - x_t)) + (v_t - \beta y_t - \alpha\eta q_t)}{\alpha\eta + \beta}.$$

Note that this minimization separates over the entries q_i; this is the benefit of adding the preconditioning term. For $i = 1, \ldots, n$, the minimizer is given by

$$q_i = \begin{cases} \frac{w_i + \sqrt{w_i^2 + \frac{4}{\alpha\eta+\beta}}}{2} \wedge 1, & \text{if } P_i > \tau, \\ \frac{(w_i + \frac{1}{1-\tau}) - \sqrt{(w_i - \frac{1}{1-\tau})^2 + \frac{4}{\alpha\eta+\beta}}}{2} \vee \epsilon, & \text{if } P_i \leq \tau. \end{cases}$$

The x and y update step Since the x and y variables do not appear jointly in any of the terms of $L(q, x, y, u, v)$, their updates are calculated independently. Trivially the x update is computed as

$$x_{t+1} = \mathrm{Proj}_{\mathcal{M}}\left(Mq_{t+1} + u_t/\alpha\right)$$

and we assume that $\mathcal{M}$ is such that this step is easy to compute. The y update is given by

$$y_{t+1} = \mathrm{Proj}_{\mathcal{B}}\left(q_{t+1} + v_t/\beta\right)$$

where $\mathcal{B} = \left\{ y \in \mathbb{R}^n_+ : \sum_i \frac{\mathbb{1}\{P_i > \tau\}}{y_i(1-\tau)} \leq n \right\}$, and projection to this convex set is calculated as follows. Fix any $z \in \mathbb{R}^n_+$. If $\sum_i \frac{\mathbb{1}\{P_i > \tau\}}{z_i(1-\tau)} \leq n$ then trivially $\text{Proj}_{\mathcal{B}}(z) = z$. If not, then for each $\lambda > 0$, define the function $f_\lambda(x)$ as the unique solution $t > x$ to the cubic equation $t^3 - t^2 x = \lambda$ (which we can calculate in closed form by the cubic formula). $\text{Proj}_{\mathcal{B}}(z)$ is the vector $y \in \mathbb{R}^n_+$ minimizing $\frac{1}{2}\|y - z\|_2^2$ subject to $\sum_i \frac{\mathbb{1}\{P_i > \tau\}}{y_i} \leq n(1-\tau)$, so by the theory of Lagrangian multipliers, for some $\lambda > 0$ we have

$$(y - z) + \lambda \cdot (-\mathbb{1}\{P_i > \tau\} / y_i^2)_{1 \leq i \leq n} = 0.$$

In other words, for each i, y_i satisfies $y_i - z_i - \lambda \mathbb{1}\{P_i > \tau\} / y_i^2 = 0$, i.e.

$$y_i = \begin{cases} z_i, & \text{if } P_i \leq \tau, \\ f_\lambda(z_i), & \text{if } P_i > \tau. \end{cases}$$

Now let $y(\lambda) \in \mathbb{R}^n_+$ be defined in this way, and note that $(y(\lambda))_i$ is a nondecreasing function of λ, and is strictly increasing if $P_i > \tau$. Choosing λ_* as the unique value such that $\sum_i \mathbb{1}\{P_i > \tau\} / (y(\lambda_*))_i = n(1-\tau)$, then

$$\text{Proj}_{\mathcal{B}}(z) = y(\lambda_*).$$

CHAPTER 4

MULTIPLE TESTING ON EDGES OF A GRAPH WITH SABHA

The accumulation tests and SABHA method we introduced in previous chapters focus on detecting signals among a group of objects. In many applications, we are interested in discovering interacting objects or the relations between them. For example, the friendships in a social network, the collaborations in scientific communities, interactions among proteins in a metabolism process, the co-movement of stock prices in an industry sector, etc. Gaining such knowledge allows people to make better decisions in research and business.

The pairwise relationships among objects can be modeled by a graph, in which the nodes are the objects and the edges are the potential interactions we are interested in. Therefore, given data on the pairwise interactions or similarities (specifically, p-values), we can formulate this as a multiple testing problem. When the graph exhibits some specific structures, or we have prior knowledge on the underlying structure of the graph, utilizing such information may help to boost the true detections while controlling FDR, as chapter 2, 3 have shown.

In this chapter, we consider three specific types of graph structures – stochastic block structure, degree-corrected structure, and network with node attributes structure, all based on the community detection models. We develop theories of SABHA with these structures, and apply them in simulations to compare with other multiple testing methods. Our goal here is to recover the signals (the true interactions) among objects, instead of recovering the community structures.

Chapter outline: In Section 4.1, we give the details about the community detection models, and applying SABHA in each model setting. In Section 4.2, we present empirical results on a dataset simulated from the degree-corrected stochastic block model, and a dataset simulated from the network with node attributes model. In Section 4.3, we give details of proofs.

4.1 Application of SABHA to graph structures based on community detection models

We consider applying SABHA to various graph structure settings here. For each setting we provide bounds on the Rademacher complexity of $\mathcal{Q}_{\text{inv}}$, which allow us to apply the main FDR control result of SABHA (Chapter 3, Theorem 5). These results are proved in Section 4.3. For simplicity, all graphs $G = (V_G, E_G)$ discussed here are undirected and have n nodes and $\frac{1}{2}n(n-1)$ potential edges. The potential edges are the set of hypotheses.

4.1.1 Stochastic block model

The stochastic block model (SBM) has been widely employed in network and community detection studies, as one of the simplest models of a graph with communities (Karrer and Newman [2011]). In a stochastic block setting, we believe that the nodes are divided into different groups, and the probability of an edge between a pair of nodes is determined by the groups that they belong to. For example, the nodes from the same group may be more likely to connect than the ones from distinct groups, or the nodes from two specific groups are more possible to link than the general cases. In this setting, $\widehat{q}$ of the edges in graph G are from the set

$$\mathcal{Q}_{\text{SBM}} = \{q : q_{uv} = \frac{1}{1+\exp(w_{g(u)g(v)})}, w_{g(u)g(v)} \leq C, \forall u, v \in V_G\}.$$

in which $g(u), g(v)$ are the groups that nodes u, v belong to, respectively. $w_{g(u)g(v)} \leq C$ makes sure that $q_{uv} \geq \frac{1}{1+\exp(C)} > 0$. This is only one way to parameterize the model. It is natural because it is a logistic model of the connection probability, as it indicates $\text{logit}(\mathbb{P}\{\text{edge } uv\}) = w_{g(u)g(v)}$.

Since all connections between any two specific groups share the same $\widehat{q}$, it can be simplified as

$$\mathcal{Q}_{\text{SBM}} = \{q : q_{uv} = \tilde{w}_{g(u)g(v)}, \tilde{w}_{g(u)g(v)} \in [\epsilon, 1], \epsilon \in (0, 1)\}.$$

Therefore, this is a special case of the block structure that we have studied in chapter 3. Suppose there are d non-overlapping groups of nodes $g_1, \ldots, g_d$, which have sizes $n_1, \ldots, n_d$ ($n_1 + \cdots + n_d = n$), then based on previous calculations

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \frac{\epsilon^{-1} - 1}{2} \sum_{i \leq j} \sqrt{n_i n_j}.$$

Therefore, applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{\epsilon^{-1} - 1}{\tau(1-\tau)} \cdot \frac{\sum_{i \leq j} \sqrt{n_i n_j}}{n(n-1)} \right).$$

To obtain meaningful bounds on FDR here, we need $\epsilon \gg \frac{\sum_{i \leq j} \sqrt{n_i n_j}}{n(n-1)}$. For example, if the d blocks are each of size $n_i = n/d$, then we need $\epsilon \gg \frac{d}{n}$.

4.1.2 Degree-corrected stochastic block model

SBM, though simple, is not flexible enough to represent networks with structures similar to those found in most empirical network data (Karrer and Newman [2011]). As an extension of SBM, degree-corrected SBM (DC-SBM) has been introduced to include heterogeneity in the vertices. In DC-SBM, the nodes are again divided into different groups, yet the probability of connection between two nodes are not only determined by the groups of two nodes, but also by the two nodes themselves. The probabilities of being null over all the edges in graph G now take the form

$$\mathcal{Q}_{\text{DC-SBM}} = \{q : q_{uv} = \frac{1}{1 + \exp(\theta_u + \theta_v + w_{g_u g_v})}, w_{g_u g_v} \leq C, \theta_u, \theta_v \leq C', \forall u, v \in V_G\}.$$

where C, C' are positive constants. Now we examine the FDR control of the SABHA method for the DC-SBM setting.

Lemma 10. *For $\mathcal{Q} = \mathcal{Q}_{\text{DC-SBM}}$ as defined above,*

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \frac{e^{C+2C'}}{2} \sum_{i \leq j} \sqrt{n_i n_j}$$

Therefore, applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{6e^{C+2C'} \cdot n^{\frac{3}{2}}}{n(n-1)} \right) + \frac{\alpha}{\tau(1-\tau)} \cdot \frac{2e^{C+2C'} \sum_{i \leq j} \sqrt{n_i n_j}}{n(n-1)}.$$

Notice that the second and third terms in the FDR bound come from degree correction and the block model, respectively. To obtain meaningful bounds on FDR here, we need $\sum_{i \leq j} \sqrt{n_i n_j} \ll n(n-1)$. If the d groups are of equal size n/d, then we need $n \gg d$.

4.1.3 Degree-corrected general structure model

Now we extend DC-SBM to a more general setting, where any allowed structure on $\widehat{q}$ can be modified via degree correction. Suppose that for each pairs of nodes, the probability of connection is determined by the degrees of the two nodes as well as an edge factor, which has some specified structure – e.g, a group, ordered, or low total variation structure, based on prior knowledge. Suppose $\mathcal{Q}^*$ is the structured set s.t. $\log\left(\frac{1}{q^*} - 1\right) \leq C$, for all $q^* \in \mathcal{Q}^*$, we define the degree-corrected set as

$$\mathcal{Q}_{\text{DC}} = \{q : q_{uv} = \frac{1}{1 + \exp(\theta_u + \theta_v + w_{uv})}, w_{uv} = \log\left(\frac{1}{q^*_{uv}} - 1\right), q^* \in \mathcal{Q}^*, \theta_u, \theta_v \leq C', \forall u, v \in V_G\},$$

where C, C' are positive constants. Such a generalization would allow us to greatly expand the application of SABHA, if we can prove the FDR control of SABHA for this setting. As the lemma below shows, we find that the FDR is well controlled in the DC general structure setting, as long as the FDR is controlled in underlying structural setting.

Lemma 11. *For $\mathcal{Q} = \mathcal{Q}_{\text{DC}}$ as defined above,*

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq 3e^{C+2C'} \cdot n^{\frac{3}{2}} + 2e^{2C'} \cdot \omega_{\text{Rad}}(\mathcal{Q}^*_{\text{inv}})$$

applying Theorem 5,

$$\text{FDR} = \mathbb{E}\left[\mathsf{FDP}\right] \leq \alpha \left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{6e^{C+2C'} \cdot n^{\frac{3}{2}}}{n(n-1)}\right) + \frac{\alpha}{\tau(1-\tau)} \cdot \frac{4e^{2C'} \cdot \omega_{\text{Rad}}(\mathcal{Q}^*_{\text{inv}})}{n(n-1)}.$$

Since DC-SBM is a special case of the degree-corrected general structure model, Lemma 10 is based on Lemma 11.

Notice that the upper bound of $\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}})$ consists of two terms, one for the $\omega_{\text{Rad}}(\mathcal{Q}^*_{\text{inv}})$, and one for the degree correction, which $\ll n^2$. So to obtain meaningful bounds on FDR here, we just need $\omega_{\text{Rad}}(\mathcal{Q}^*_{\text{inv}}) \ll n^2$.

4.1.4 Network with node attributes model

Network models based on object (node) attributes assume that attributes information, such as users' social network profile, authors' publication histories, indicate the communities to which the nodes belong to (Yang et al. [2013]). Nodes with similar attributes are more likely to be connected, e.g. collaborating and interacting with each other. In the attributes model, suppose each node u in graph G has a normalized attributes vector $F_u \in R^r$, then $\widehat{q}$ of the edges are modeled as

$$\mathcal{Q}_{\text{attributes}} = \{q : q_{uv} = \exp\left(\min\left\{0, \frac{1}{2}||F_u - F_v||_2^2 - 1\right\}\right), F_u \in R^r, ||F_u||_2 = 1, \forall u \in V_G, \}$$

where $||F_u - F_v||_2^2$ measures the similarity between nodes u and v. Also, the higher dimension r, the more complex the attributes model. This is supported by the growing Rademacher complexity with r, as shown in the lemma below.

Lemma 12. *For $\mathcal{Q} = \mathcal{Q}_{\text{attributes}}$ as defined above,*

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq 9 \cdot \sqrt{r} \cdot n^{\frac{3}{2}},$$

applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{18 \cdot \sqrt{r}}{\tau(1-\tau)(\sqrt{n} - \frac{1}{\sqrt{n}})}\right)$$

.

To obtain meaningful bounds on FDR here, we need $r \ll n$.

While this model has nice theoretical guarantees, it's hard to project onto $\mathcal{Q}_{\text{attributes}}$ and therefore hard to compute the estimated probabilities $\widehat{q}$. As a simplification of the node attributes model, we also consider the setting where the probabilities of the edges are from

$$\mathcal{Q}_{\text{Sim-attributes}} = \{q : q_{uv} = \min\left\{1, \frac{1}{2}||F_u - F_v||_2^2\right\},$$
$$F_u \in R^r, ||F_u||_2 = 1, \forall u \in V_G, ||F_u - F_v||_2^2 \geq \epsilon > 0, \forall u, v \in V_G\},$$

and we prove that:

Lemma 13. *For* $\mathcal{Q} = \mathcal{Q}_{\text{Sim-attributes}}$ *as defined above,*

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \frac{12}{\epsilon^2} \cdot \sqrt{r} \cdot n^{\frac{3}{2}},$$

applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{\frac{24}{\epsilon^2} \cdot \sqrt{r}}{\tau(1-\tau)(\sqrt{n} - \frac{1}{\sqrt{n}})} \right).$$

Again we need $r \ll n$ to obtain meaningful bounds on FDR.

4.2 Experiments

4.2.1 Simulated data: degree-corrected models

Simulation In this setting, the graph has 20 nodes and 190 potential edges, when the graph is fully connected. The nodes belong to 4 groups, so there are 10 types of edges (6 of them are edges between nodes of different groups, and 4 of them are edges within a group). For each edge, the underlying prior probability of the edge uv being a null is generated as

$$q_{uv} = \frac{1}{1 + \exp(\theta_u + \theta_v + w_{g(u)g(v)})}$$

where θ_u, θ_v are the degrees of nodes u, v, and $g(u), g(v)$ are the groups that u, v belong to, respectively. $w_{g(u)g(v)}$ is the weight based on the edge type. The degrees of nodes are equally spaced within $[-1, 1]$, and the weights of edge types are equally spaced within $[-2, 2]$. This guarantees equal contributions to the signal from the node degrees and edge types.

To generate the p-value for the edge, we first draw $Z \overset{\perp\!\!\!\perp}{\sim} N(\mu, 1)$ where $\mu = 0$ for the nulls and $\mu = \mu_{\text{sig}} > 0$ for the non-nulls, for $\mu_{\text{sig}} \in \{0.5, 1, 1.5, \ldots, 3.5\}$ (with larger μ_{sig} indicating a stronger signal). Then we run two-sided z-tests, $P = 2(1 - \Phi(|Z|))$, where Φ is the CDF of the standard normal.

For the implementation of SABHA, we choose $\tau = 0.4, \epsilon = 0.1$, and three choices of model spaces: stochastic block model, degree only model, and the degree-corrected stochastic block model (SBM, DC, DC-SBM, for short, respectively). Obviously, SBM and DC are strictly smaller models than DC-SBM, and the true prior probability is based on DC-SBM. We then compare SABHA with BH, and also with Storey's modification of BH ("Storey-BH"), implemented with parameter $\tau = 0.4$.

To fit $\widehat{q}$ for SABHA, we apply the same ADMM framework developed in chapter 3 (**ref**). In the "x update step" of the algorithm, $M = I_{n_{\text{edges}}}$ and $\mathcal{M} = \{q_{uv} : q_{uv} = \frac{1}{1+\exp(w_{g(u)g(v)})}\}$ for SBM, $\mathcal{M} = \{q_{uv} : q_{uv} = \frac{1}{1+\exp(\theta_u+\theta_v)}\}$ for DC, and $\mathcal{M} = \{q_{uv} : q_{uv} = \frac{1}{1+\exp(\theta_u+\theta_v+w_{g(u)g(v)})}\}$ for DC-SBM. The projection is conducted by fitting the nonlinear equations parameterized in each model. For DC-SBM, an alternating direction optimization is applied, by altering between optimizing the degree parameters and the block parameters. For comparison we also run an "oracle" version of SABHA where $\widehat{q} = q$, the true vector of prior probabilities of each p-value being a null. For all methods we choose the target FDR level $\alpha = 0.1$.

As Figure 4.1 shows, the average power and observed FDR grow as the signal becomes stronger, and BH, oracle SABHA control FDR at level $\alpha = 0.1$. The FDR given by Storey-BH goes slightly over 0.1. SABHA is consistently more powerful than BH and Storey-BH over the range of μ_{sig}.

Among SABHA methods, SBM well controls FDR at low signal strength, and has similar FDR to that of Storey-BH at high signal strength. DC and DC-SBM show higher observed FDR across all signal strength levels. With respect to power, at low signal strength levels, DC-SBM $>$ DC $>$ SBM; at high levels, SBM catches up, and the power of these three becomes very close. The higher power and FDR observed when the model changes from SBM to DC, DC-SBM is due to the fact that DC and DC-SBM has more degrees of freedom than SBM, and DC-SBM is a strictly larger model than SBM and DC.

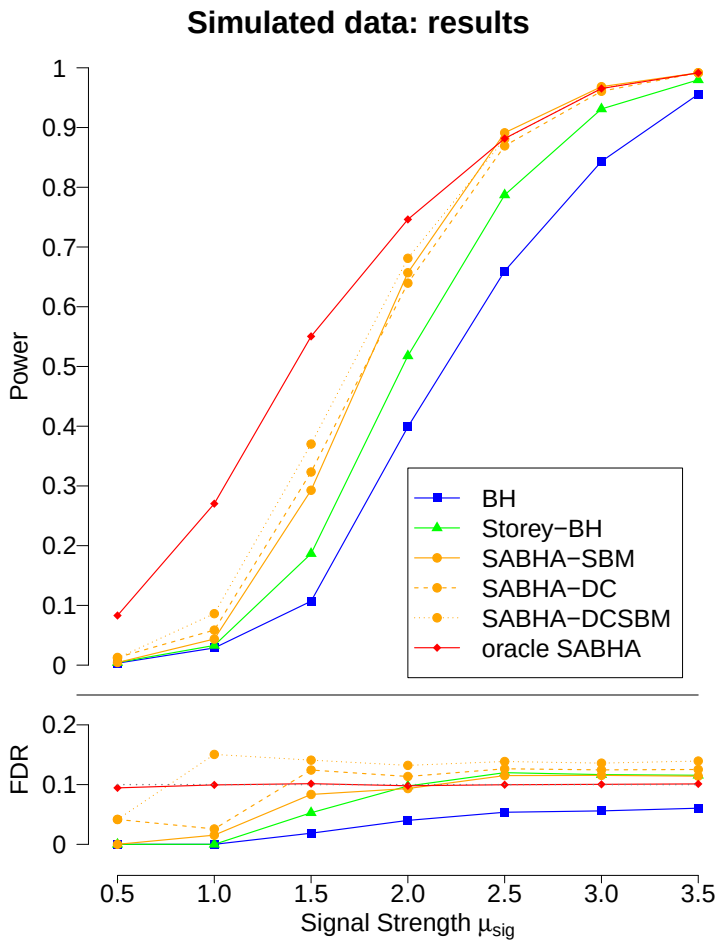

Figure 4.1: Power and observed FDR level of BH, Storey-BH, and SABHA (oracle, SBM, DC, DC-SBM) procedures averaged over 10 trials. The target FDR level is $\alpha = 0.1$.

4.2.2 Simulated data: simplified node attributes model

Simulation Here we simulate a graph with 35 nodes and 595 potential edges, for a fully-connected graph. The attributes of each node is generated from uniform distribution on the unit sphere in R^3, and the attributes are mutually independent.

For each edge, we create a p-value, and assign an underlying prior probability of the edge being a null, based on the attributes of the two nodes it connects:

$$q_{uv} = \min\{1, \frac{1}{2}||F_u - F_v||_2^2\}$$

where u, v are the nodes, and F_u, F_v are the attributes of u, v, respectively. $q_{uv} = 1$ when

F_u, F_v are orthogonal or negatively correlated. Note that this is slightly different from the definition in Section 4.1.4. The symmetric matrix Q_{mat}, whose (u, v) entry is q_{uv}, is full-rank but close to low-rank, as the sum of its first few eigenvalues is close to the sum of all eigenvalues.. The p-values are generated following the same way as in previous simulations on SABHA.

We implement SABHA on edges of the graph, choosing $\tau = 0.5$, the lower bound parameter of $\widehat{q}$ at $\epsilon = 0.2$, and the rank constraint $r \in \{1, 3, 5, 15\}$. We then compare SABHA with BH, and also with Storey's modification of BH ("Storey-BH"), implemented with parameter $\tau = 0.5$.

To fit $\widehat{q}$ for the SABHA method, we apply the same ADMM framework developed in chapter 3. In the "x update step" of the algorithm, $M = I_{n_{\text{edges}}}$ and $\mathcal{M} = \{\text{rank r matrices}\}$. The projection of Q_{mat} onto the rank r matrices space can be done by eigen-decomposition and keeping the first few eigenvalues and eigenvectors of it. For comparison we also run an "oracle" version of SABHA where $\widehat{q} = q$, the true vector of prior probabilities of each p-value being a null. For all methods we choose the target FDR level $\alpha = 0.1$.

We then compare the observed FDR and power for each of the considered methods. As we can see in Figure 4.2, for all methods, the average power and average observed FDR both increase with larger μ (stronger signal). The BH, Storey-BH, and oracle SABHA methods all control FDR at level $\alpha = 0.1$.

For SABHA, as the signal grows stronger, the observed FDR grows very close to 0.1 for $r = 1$; for $r = 3, 5, 15$, the observed FDR exceeds the target level $\alpha = 0.1$, as $\widehat{q}$ is overfitting to the p-values. To compare the power, BH is most conservative (lowest power) followed by Storey-BH. Oracle SABHA shows the highest power, while SABHA shows power increasing as the rank constraint r increases. Here, SABHA is consistently more powerful than BH and Storey-BH over the range of μ_{sig}.

In Figure 4.3, we compare the SABHA methods with different constraints according to their estimated probabilities of a null, for one trial at signal strength $\mu_{\text{sig}} = 2.5$. For

SABHA with $r = 1, 3, 5, 15$, the plot displays the histogram of the estimated $\widehat{q}$. We can compare against the true q, which is also the input to the oracle SABHA method.

As expected for SABHA, we see that for $r = 1$ the method is unable to fit $\widehat{q}$ closely to the true q. For $r = 3$ the fit is substantially better; for $r = 5, 15$ the fit shows evidence of overfitting, as the distribution is shifting to the lower bound ϵ. This leads to the increased power and FDR observed for $r = 5, 15$.

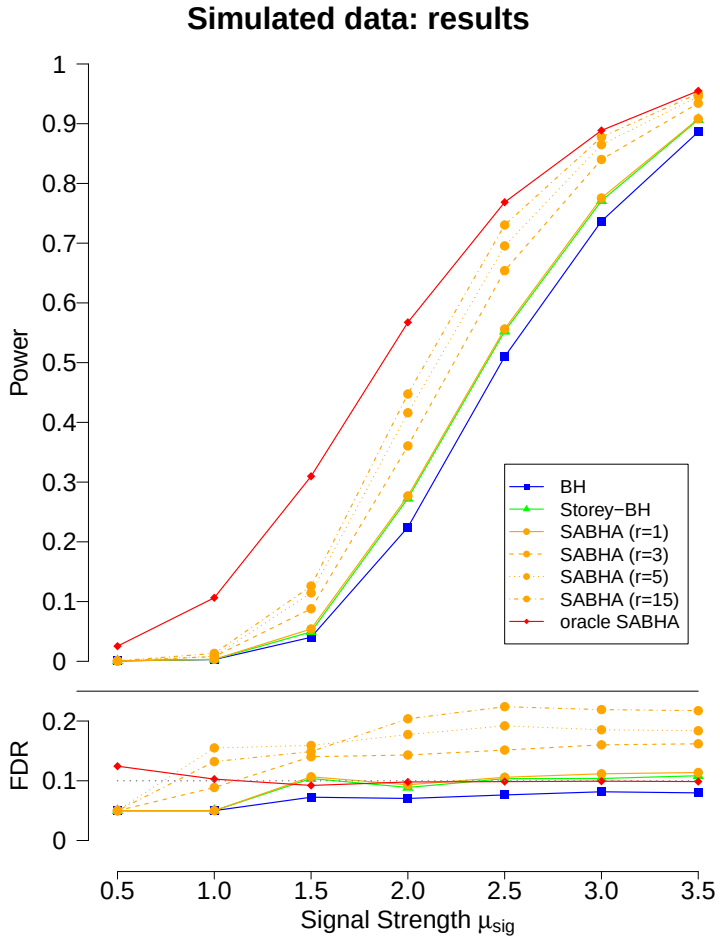

Figure 4.2: Power and observed FDR level of BH, Storey-BH, and SABHA (oracle, $r = 1, 3, 5, 15$) procedures averaged over 20 trials. The target FDR level is $\alpha = 0.1$.

4.2.3 S&P 500 data: stochastic block model

Real data experiment Here we consider a real data example publicly available from Yahoo Finance, which consists of the daily closing prices of 452 S&P 500 companies over 1258

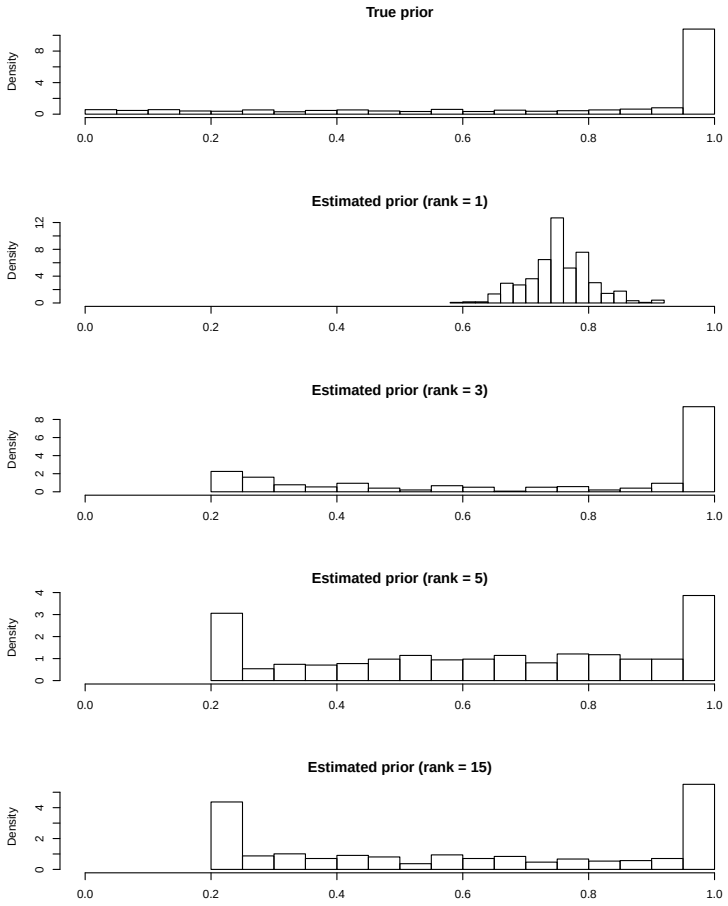

Figure 4.3: True q vs. estimated $\widehat{q}$ for a single trial with $\mu_{\text{sig}} = 2.5$, in the node attributes model simulation.

days. The 452 companies belong to 10 industry sectors, including Industrials, Financials, Health Care, Consumer discretionary, Information technology, Utilities, Materials, Consumer Staples, Telecommunications services, and Energy. We treat the S&P 500 companies as a graph, in which each individual stock is a node, and the conditional dependence relationship between each pair of stocks is an edge. The data is transformed into log-returns, and forms a matrix $X \in R^{1257\times 452}$ in which

$$X_{ij} = log\left(\frac{P_{j,i+1}}{P_{j,i}}\right)$$

where $P_{j,i}$ means the last price of stock j on day i.

To estimate the graph, we first produce a p-value for each pair of stocks using Pearson method, testing for their conditional independence given all other stocks. We then apply multiple testing methods including BH method, Storey's modification, and SABHA. For the implementation of SABHA, we choose $\tau = 0.5, \epsilon = 0.1$, and the stochastic block model space, with each industry sector as a block.

Figure 4.4 shows the estimated graph by BH, Storey's method, and SABHA, at target FDR level $\alpha = 0.01$, and 0.001. As we can see in Figure 4.4, SABHA detects more edges than BH and Storey's methods. At $\alpha = 0.01$, BH, Storey-BH, SABHA give 666, 697, and 1150 discoveries, respectively; at $\alpha = 0.001$, the three methods give 368, 376, 470 discoveries, respectively.

Figure 4.5 and Figure 4.6 show the $\hat{q}$ and detection rate (at $\alpha = 0.1$) for each pair of industry sectors, respectively. The Industrials sector has a low $\hat{q}$ (which means higher prior probability of connection) with all other sectors. The Telecommunications Services sector has a very low $\hat{q}$ with itself (0.16), and thus a high within block connection rate (0.2). These suggest that the sectors division here well-captures the structure within the S&P 500 stocks, and exploiting this structure allows us to make much more discoveries.

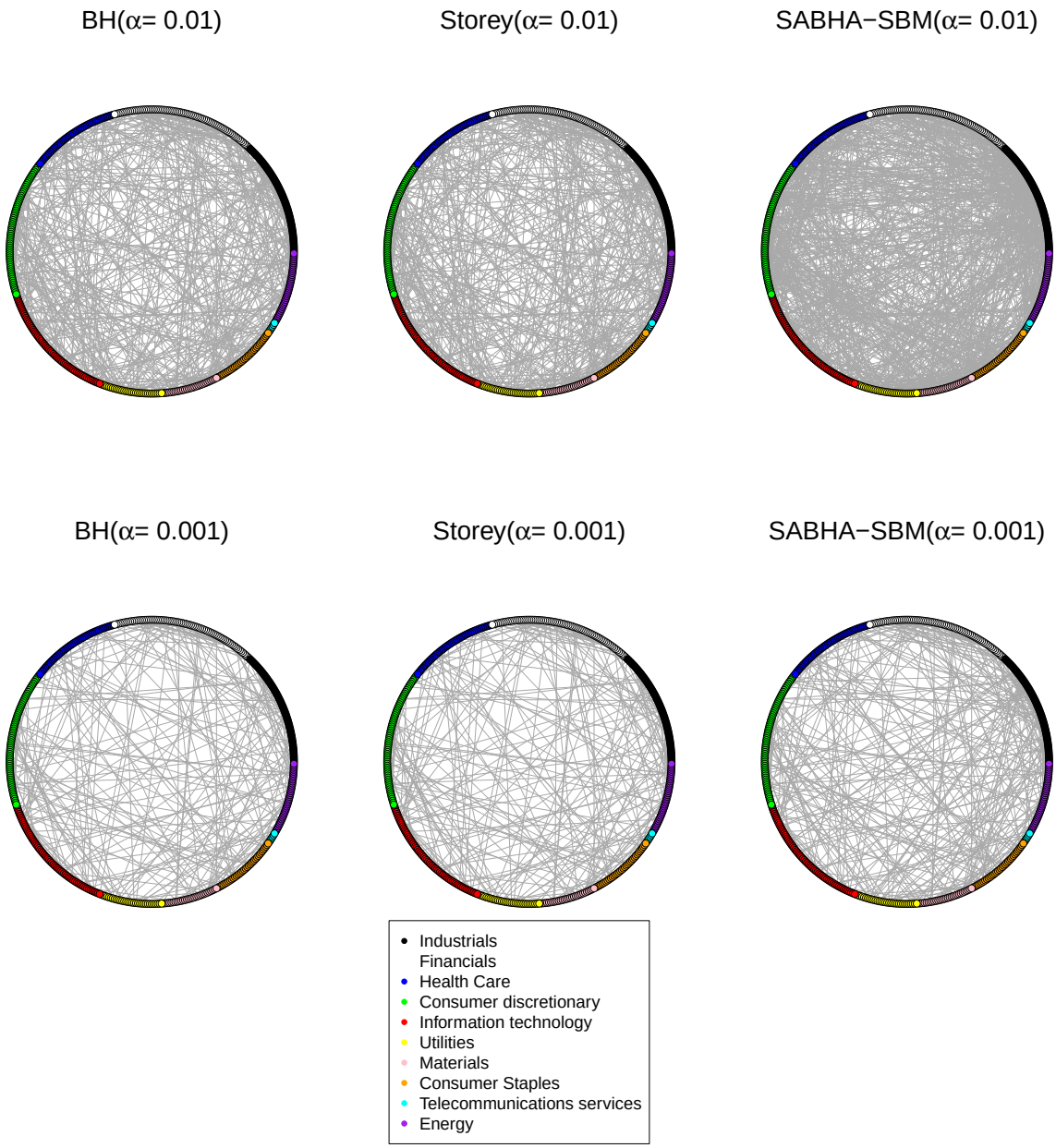

Figure 4.4: Estimated graph for the stock data, based on Pearson p-values and multiple testing methods including BH method, Storey's modification, and SABHA (with stochastic block model structure assumption). An edge is displayed for each pairwise conditional dependence at significance level 0.01 and 0.001, respectively. Graphs were drawn using the igraph package in R.

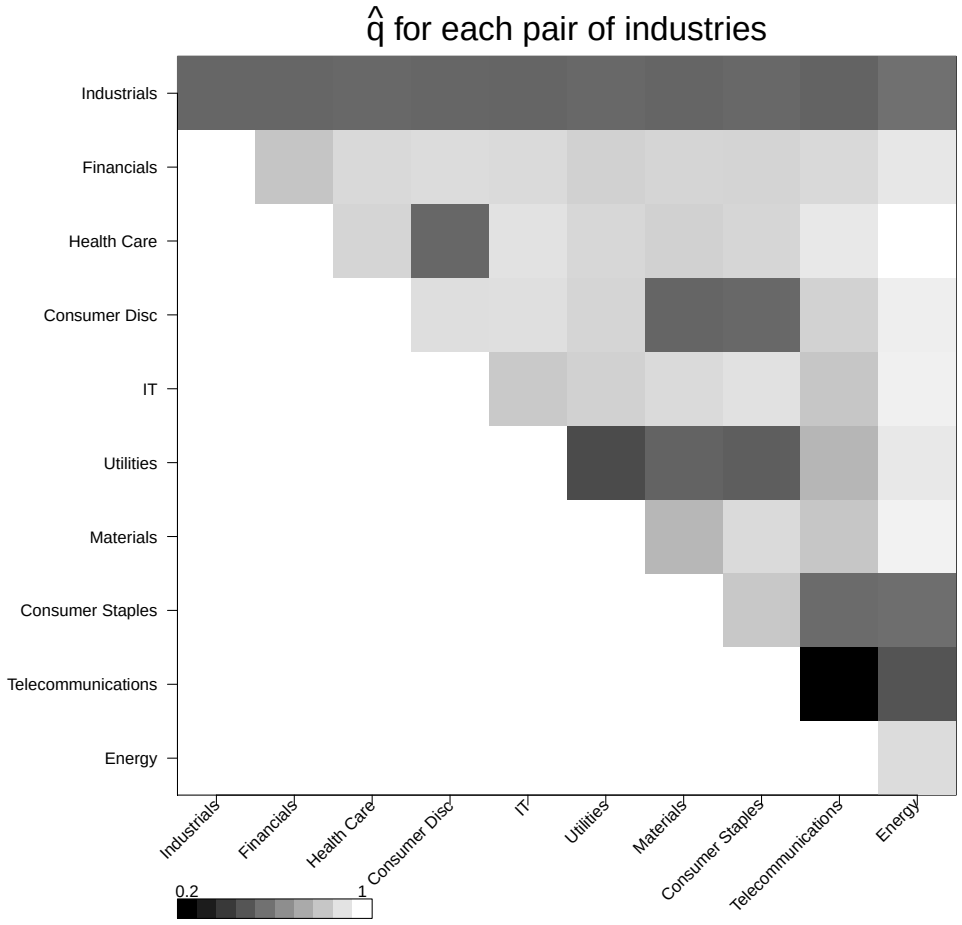

Figure 4.5: Estimated prior probability of null for each pair of industry sectors. The lower triangular part is left blank.

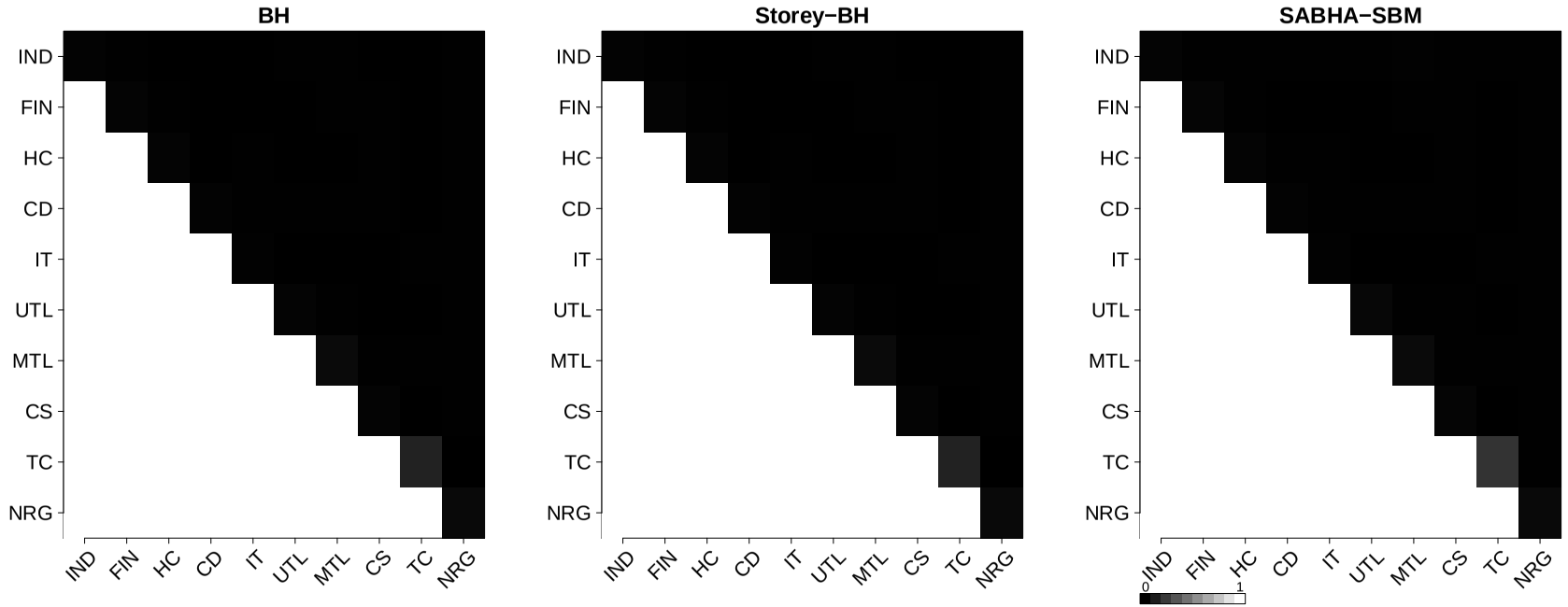

Figure 4.6: Detection rate for each pair of industry sectors, by BH method, Storey's modification, and SABHA (with stochastic block model structure assumption), at significance level 0.1. The lower triangular part is left blank. IND, FIN, HC, CD, IT, UTL, MTL, CS, TC, NRG stand for Industrials, Financials, Health Care, Consumer Discretionary, IT, Utilities, Materials, Consumer Staples, Telecommunications, Energy, respectively.

4.3 Proofs and technical details

Notation $||X||_\infty = \max_{ij} |X_{ij}|, ||X||_{2\to\infty} = \max_i |X_{i\cdot}|$ is the maximum row norm of X. $||X||_{\max} = \min_{X=UV'} ||U||_{2\to\infty}||V||_{2\to\infty}$.

4.3.1 DC general structure

Recall that for $\widehat{q}$ from DC-general we defined the set

$$\mathcal{Q}_{\mathrm{DC}} = \{q : q_{uv} = \frac{1}{1+\exp(\theta_u+\theta_v+w_{uv})}, w_{uv} = \log\left(\frac{1}{q^*_{uv}} - 1\right), q^* \in \mathcal{Q}^*, \\ w_{uv} \le C, \theta_u, \theta_v \le C', \forall u, v \in V_G\}$$

Lemma 11. For $\mathcal{Q}_{\mathrm{DC}}$ as defined above,

$$\omega_{\mathrm{Rad}}(\mathcal{Q}_{\mathrm{inv}}) \le 3e^{C+2C'} \cdot n^{\frac{3}{2}} + 2e^{2C'} \cdot \omega_{\mathrm{Rad}}(\mathcal{Q}^*_{\mathrm{inv}})$$

applying Theorem 5,

$$\mathrm{FDR} = \mathbb{E}\left[\mathsf{FDP}\right] \le \alpha\left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{1}{\tau(1-\tau)} \cdot \frac{6e^{C+2C'} \cdot n^{\frac{3}{2}}}{n(n-1)}\right) + \\ \frac{\alpha}{\tau(1-\tau)} \cdot \frac{4e^{2C'} \cdot \omega_{\mathrm{Rad}}(\mathcal{Q}^*_{\mathrm{inv}})}{n(n-1)}.$$

Proof of Lemma 11.

$$\omega_{\mathrm{Rad}}(\mathcal{Q}_{\mathrm{inv}}) = \mathbb{E}\left[\sup_{x\in\mathcal{Q}_{\mathrm{inv}}} \langle x, B\rangle\right] = \mathbb{E}\left[\sup_{\theta,w} \sum_{u<v} (1 + e^{\theta_u+\theta_v+w_{uv}})B_{uv}\right] \\ = \mathbb{E}\left[\sup_{\theta,w} \sum_{u<v} (e^{\theta_u+\theta_v+w_{uv}})B_{uv}\right] = \mathbb{E}\left[\sup_{\theta,w} \langle E, e^w \circ B\rangle\right]$$

in which E, B, e^w are the matrices whose u, v entries are $e^{\theta_u+\theta_v}, B_{uv}, e^{w_{uv}}$, respectively. Since

$$\sup_{\theta,w}\langle E, e^w \circ B\rangle \le e^{2C'} \sup_{w,||x||_\infty\le 1,||y||_\infty\le 1}\langle xy^T, e^w \circ B\rangle$$
$$\le e^{2C'} K_G \cdot \sup_{w,x\in\{\pm 1\}^n, y\in\{\pm 1\}^n}\langle xy^T, e^w \circ B\rangle$$

where the last inequality is due to the fact that $\{xy^T : ||x||_\infty \le 1, ||y||_\infty \le 1\} \subset K_G \cdot \text{conv}(\{xy^T : x \in \{\pm 1\}^n, y \in \{\pm 1\}^n\})$ (Corollary 2 in Srebro and Shraibman [2005], $1.67 < K_G < 1.79$)(max of convex combination is smaller than convex combination of max),

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \le e^{2C'} K_G \cdot \mathbb{E}\left[\sup_w \sup_{x,y\in\{\pm 1\}^n}\langle xy^T, e^w \circ B\rangle\right]$$
$$= e^{2C'} K_G \cdot \mathbb{E}\left[\sup_{x,y\in\{\pm 1\}^n} \sup_w \langle e^w, xy^T \circ B\rangle\right].$$

Let $f_{xy}(B) = \sup_w\langle e^w, xy^T \circ B\rangle$, as $xy^T \circ B \stackrel{d}{=} B$,

$$\mathbb{E}\left[f_{xy}(B)\right] = \omega_{\text{Rad}}(\mathcal{Q}^*_{\text{inv}}).$$

Since B_{uv} are independent and changing the realization of one entry of B will at most perturb $e^{-C} f_{xy}(B)$ by 1, by McDiarmid's inequality,

$$\mathbb{P}\left\{e^{-C}|f_{xy}(B) - \mathbb{E}\left[f_{xy}(B)\right]| \ge \lambda\right\} \le 2e^{-\frac{\lambda^2}{2n^2}}.$$

This suggests that $\frac{f_{xy}(B)-\mathbb{E}[f_{xy}(B)]}{ne^C}$ is 1-subgaussian. There are 2^{2n} many choices for $x, y \in \{\pm 1\}^n$, therefore,

$$\mathbb{E}\left[\sup_{x,y\in\{\pm1\}^n}\sup_w\langle e^w, xy^T\circ B\rangle\right] = \mathbb{E}\left[\sup_{x,y\in\{\pm1\}^n}(f_{xy}(B)-\mathbb{E}\left[f_{xy}(B)\right])\right] + \omega_{\mathrm{Rad}}(\mathcal{Q}^*_{\mathrm{inv}})$$
$$\leq e^C\cdot n\cdot\sqrt{2log2^{2n}} + \omega_{\mathrm{Rad}}(\mathcal{Q}^*_{\mathrm{inv}}) = 2\sqrt{log2}e^C\cdot n^{\frac{3}{2}} + \omega_{\mathrm{Rad}}(\mathcal{Q}^*_{\mathrm{inv}})$$

Thus,

$$\omega_{\mathrm{Rad}}(\mathcal{Q}_{\mathrm{inv}}) \leq 3e^{C+2C'}\cdot n^{\frac{3}{2}} + 2e^{2C'}\omega_{\mathrm{Rad}}(\mathcal{Q}^*_{\mathrm{inv}}).$$

□

4.3.2 Network with node attributes model

Recall that for $\widehat{q}$ from node attributes model we defined the set

$$\mathcal{Q}_{\mathrm{attributes}} = \{q : q_{uv} = \exp\left(\min\left\{0, \frac{1}{2}||F_u - F_v||_2^2 - 1\right\}\right), F_u \in R^r, ||F_u||_2 = 1, \forall u \in V_G,\}$$

Lemma 12. For $\mathcal{Q} = \mathcal{Q}_{\mathrm{attributes}}$ as defined above,

$$\omega_{\mathrm{Rad}}(\mathcal{Q}_{\mathrm{inv}}) \leq 9\cdot\sqrt{r}\cdot n^{\frac{3}{2}},$$

applying Theorem 5,

$$\mathrm{FDR} \leq \alpha\left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{18\cdot\sqrt{r}}{\tau(1-\tau)(\sqrt{n}-\frac{1}{\sqrt{n}})}\right)$$

.

Proof of Lemma 12. Let $M_{uv} = \frac{1}{2}||F_u - F_v||_2^2 - 1, M \in R^{n\times n}$. M has rank r, as $M_{uv} =$

$-\langle F_u, F_v\rangle = -\sum_{i=1}^{r} F_u^i F_v^i$.

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) = \mathbb{E}\left[\sup_{x\in\mathcal{Q}_{\text{inv}}} \langle x, B\rangle\right] \leq \mathbb{E}\left[\sup_{\text{rank}(M)\leq r, ||M||_\infty \leq 1} \sum_{u<v} e^{-\min\{0, M_{uv}\}} B_{uv}\right]$$
$$\leq e \cdot \mathbb{E}\left[\sup_{\text{rank}(M)\leq r, ||M||_\infty \leq 1} \sum_{u<v} M_{uv} B_{uv}\right]$$

where the last inequality applies the Lipschitz Rademacher theorem (Bartlett and Mendelson [2002]). Since $\{M : \text{rank}(M) \leq r, ||M||_\infty \leq 1\} \subset \{M : ||M||_{\max} \leq \sqrt{r}\} \subset \sqrt{r} K_G \cdot \text{conv}(\{xy^T : x \in \{-1,1\}^n, y \in \{-1,1\}^n\})$(the first step by Lemma 3 in Foygel and Srebro [2011], the second step by Corollary 2 in Srebro and Shraibman [2005]),

$$\sup_{\text{rank}(M)\leq r, ||M||_\infty \leq 1} \sum_{u<v} M_{uv} B_{uv} = \sup_{\text{rank}(M)\leq r, ||M||_\infty \leq 1} \langle B, M\rangle$$
$$\leq \sqrt{r} K_G \cdot \sup_{x\in\{-1,1\}^n, y\in\{-1,1\}^n} \langle B, xy^T\rangle$$

where B is the matrix of independent Rademacher variables $B_{uv}, u < v$ (note that only the upper diagonal part of B is nonzero). Based on the results in DC-SBM,

$$\mathbb{E}\left[\sup_{x\in\{-1,1\}^n, y\in\{-1,1\}^n} \langle B, xy^T\rangle\right] \leq 2\sqrt{log2} \cdot n^{\frac{3}{2}}$$

Putting these together, we have

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq 9 \cdot \sqrt{r} \cdot n^{\frac{3}{2}}$$

□

For the simplified node attributes model, we defined that

$$\mathcal{Q}_{\text{Sim-attributes}} = \{q : q_{uv} = \min\left\{1, \frac{1}{2}||F_u - F_v||_2^2\right\},$$
$$F_u \in R^r, ||F_u||_2 = 1, \forall u \in V_G, ||F_u - F_v||_2^2 \geq \epsilon > 0, \forall u, v \in V_G\},$$

Lemma 13. For $\mathcal{Q} = \mathcal{Q}_{\text{Sim-attributes}}$ as defined above,

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \frac{12}{\epsilon^2} \cdot \sqrt{r} \cdot n^{\frac{3}{2}},$$

applying Theorem 5,

$$\text{FDR} \leq \alpha \left(1 + \sqrt{\frac{2\tau}{n(n-1)(1-\tau)}} + \frac{\frac{24}{\epsilon^2} \cdot \sqrt{r}}{\tau(1-\tau)(\sqrt{n} - \frac{1}{\sqrt{n}})}\right).$$

Proof of Lemma 13. Let $M_{uv} = \frac{1}{2}||F_u - F_v||_2^2 - 1, M \in R^{n \times n}$, then

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) = \mathbb{E}\left[\sup_{x \in \mathcal{Q}_{\text{inv}}} \langle x, B\rangle\right] \leq \mathbb{E}\left[\sup_{\text{rank}(M) \leq r, ||M||_\infty \leq 1} \sum_{u<v} (\min\{1, M_{uv} + 1\})^{-1} B_{uv}\right]$$
$$\leq \frac{4}{\epsilon^2} \cdot \mathbb{E}\left[\sup_{\text{rank}(M) \leq r, ||M||_\infty \leq 1} \sum_{u<v} M_{uv} B_{uv}\right]$$

by the Lipschitz Rademacher theorem. Following the same proof procedure of Lemma 12, we can get that

$$\omega_{\text{Rad}}(\mathcal{Q}_{\text{inv}}) \leq \frac{12}{\epsilon^2} \cdot \sqrt{r} \cdot n^{\frac{3}{2}}$$

□

CHAPTER 5

SUMMARY

This thesis includes results for novel methods of multiple testing when prior information is available on the structure in the list of hypotheses $H_1, \ldots, H_n$. We now summarize our main results in each chapter.

In chapter 2, we proposed the family of accumulation tests, which generalizes existing methods for the ordered testing problem. The new HingeExp method within this family gives significantly higher power than the existing ForwardStop and SeqStep tests (when HingeExp and SeqStep have the same C in their parameterizations), while maintaining control of the modified FDR in a finite-sample setting, and asymptotic control of the FDR. Our theoretical results prove FDR control for methods in this family in general, and examine the power properties of the tests within this family. These methods are a natural fit for any multiple testing problem where there is an inherent ordering to the hypotheses, but many other settings can be framed in this way as well—our real data experiment, which uses measurements of gene expression level across a gradient of drug dosages, indicates that we can achieve strong power gains by framing the problem as an ordered hypothesis testing problem.

In chapter 3, we proposed SABHA, a unified approach to structured multiple testing which is adaptive to diverse structures in the patterns of signals and nulls among the hypotheses. SABHA re-weights the p-values of hypotheses in a way that is adaptive to both data and the structure. We proved that in finite sample SABHA controls FDR at a level that is slightly higher than target, as long as the Rademacher complexity of the set of inverse weights is well controlled. We calculated the FDR bound for three commonly observed structures – ordered, block, and low total variation structures, and demonstrated SABHA in three simulated and real data examples, each representing one of the structures. In these examples, SABHA leads to higher power than methods that do not exploit the known structures, while maintaining control of the FDR.

In chapter 4, we considered SABHA in the settings where the hypotheses can be formulated as the edges of a graph. Such multiple testing problems arise in applications where we are interested in pairwise relationship among a group of objects. We calculated the Rademacher and FDR bounds for three types of graph structures – SBM structure, degree-corrected general structure, and network with node attributes structure. We also examined SABHA in simulations where the data is generated based on a DC-SBM model, and a node attribute model. SABHA has better performance when compared with conventional multiple testing methods. The results here suggest that SABHA framework is general and can be customized to adapt to many desired type of structures in the pattern of signals and nulls, allowing for greater power in a potentially wide range of applications where multiple testing problems arise.

REFERENCES

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

R. E. Barlow, D. J. Bartholomew, J. Bremner, and H. D. Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley New York, 1972.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Y. Benjamini and M. Bogomolov. Adjusting for selection bias in testing multiple families of hypotheses. *arXiv preprint arXiv:1106.3670*, 2011.

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

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Y. Benjamini and D. Yekutieli. Hierarchical FDR testing of trees of hypotheses. Technical report, Department of Statistics and Operations Research, Tel Aviv University, 2003.

A. Borovkov. Probability theory, 1999.

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

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

D. Catelan, C. Lagazio, and A. Biggeri. A hierarchical Bayesian approach to multiple testing in disease mapping. *Biometrical Journal*, 52(6):784–797, 2010.

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

K. R. Coser, J. Chesnes, J. Hur, S. Ray, K. J. Isselbacher, and T. Shioda. Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in MCF7/BUS breast cancer cells by DNA microarray. *Proceedings of the National Academy of Sciences*, 100 (24):13994–13999, 2003.

S. Davis and P. Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 14:1846–1847, 2007.

L. Du and C. Zhang. Single-index modulated multiple testing. *The Annals of Statistics*, 42 (4):30–79, 2014.

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

B. Efron and R. Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

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

X. Fan, I. Grama, and Q. Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(1):1–22, 2014.

J. Ferreira and A. Zwinderman. On the Benjamini–Hochberg method. *The Annals of Statistics*, 34(4):1827–1849, 2006.

W. Fithian, J. Taylor, R. Tibshirani, and R. Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

D. P. Foster and R. A. Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *COLT*, pages 315–340, 2011.

B. L. Fridley, G. Jenkins, K. White, W. Bamlet, J. D. Potter, E. L. Goode, et al. Bayesian mixture models for the incorporation of prior knowledge to inform genetic association studies. *Genetic Epidemiology*, 34(5):418–426, 2010.

C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

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

M. G. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

R. Heller, D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini. Cluster-based analysis of fMRI data. *Neuroimage*, 33(2):599–608, 2006.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

J. X. Hu, H. Zhao, and H. H. Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 2012.

J.-C. Hütter and P. Rigollet. Optimal rates for total variation denoising. *arXiv preprint arXiv:1603.09388*, 2016.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. *Conference on Learning Theory (COLT)*, 2014.

B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

T. Keller, M. Just, and V. Stenger. Reading span and the time-course of cortical activation in sentence-picture verification. In *Annual Convention of the Psychonomic Society*, 2001.

L. Lei and W. Fithian. Power of ordered hypothesis testing. *arXiv preprint arXiv:1606.01969*, 2016.

A. Li and R. F. Barber. Accumulation tests for FDR control in ordered hypothesis testing. *Journal of the American Statistical Association*, 2016a.

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

J. N. McClintick and H. J. Edenberg. Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics*, 7(1):49, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

J. G. Scott. Nonparametric Bayesian multiple testing for longitudinal performance stratification. *The Annals of Applied Statistics*, pages 1655–1674, 2009.

J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.

N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

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

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

J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 267–288, 1996.

J. Yang, J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on*, pages 1151–1156. IEEE, 2013.