Controlling the false discovery rate with dynamic adaptive procedures and of grouped hypotheses

by

Peter W. MacDonald

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2018

© Peter W. MacDonald 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In the multiple testing problem with independent tests, the classical Benjamini-Hochberg (BH) procedure controls the false discovery rate (FDR) at level $\pi_0\alpha$, where π_0 is the proportion of true null hypotheses and α is the target FDR level.

Adaptive procedures can improve power by incorporating estimates of π_0, which typically rely on a tuning parameter. Fixed adaptive procedures set their tuning parameters before seeing the data and can be shown to control the FDR in finite samples. In Chapter 2 of this thesis, we develop theoretical results for dynamic adaptive procedures whose tuning parameters are determined by the data. We show that, if the tuning parameter is chosen according to a left-to-right stopping time rule, the corresponding dynamic adaptive procedure controls the FDR in finite samples. Examples include the recently proposed right-boundary procedure and the widely used lowest-slope procedure, among others. Simulation results show that the right-boundary procedure is more powerful than other dynamic adaptive procedures under independence and mild dependence conditions.

The BH procedure implicitly assumes all hypotheses are exchangeable. When hypotheses come from known groups, this assumption is inefficient, and power can be improved through a ranking of significance that incorporates group information. In Chapter 3 of this thesis, we define a general sequential framework for multiple testing procedures in the grouped setting. We develop a flexible grouped mirrored knockoff (GMK) procedure which approximates the optimal ranking of significance. We show that the GMK procedure controls the FDR in finite samples, and give a particular data-driven implementation using the expectation-maximization algorithm. Simulation and a real-data example demonstrate that the GMK procedure outperforms its competitors in terms of power and FDR control with independent tests.

Acknowledgements

I would like to thank my supervisor, Professor Kun Liang, who made this thesis possible through many hours of helpful meetings and many rounds of revisions. I would also like to thank Professor Yingli Qin for offering me my first opportunity to do research in statistics, as well as all of the professors in the Statistics department and otherwise that I have had the privilege of learning from during my time at the University of Waterloo.

I extend many thanks to my family and friends for their support and encouragement, especially my mother and father, Cameron, Kala, and of course Echo.

Dedication

This thesis is dedicated to my grandfather *Ronald W. Park*.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Hypothesis testing is a fundamental concept in frequentist statistics. In this context, the p-value is ubiquitous, and is a topic in any introductory statistics course. When testing a single null hypothesis H_0 against the alternative H_a, rejecting the null hypothesis for a p-value below $\alpha \in (0, 1)$ controls the probability of a type I error: under H_0, the probability of rejection is less than or equal to α (Casella and Berger, 2002). This property follows from the definition of the p-value; it is defined for the purpose of single hypothesis testing. However, suppose we instead seek to test m null hypotheses *simultaneously*, for $m \geq 1$. This setting, referred to as *multiple hypothesis testing* (or simply multiple testing), is the topic of this thesis.

In the classical multiple testing framework, we have null hypotheses $H_1, ..., H_m$ for $m \geq 1$, where $H_i = 1$ denotes that the ith hypothesis is true, and $H_i = 0$ that it is false. Let m_0 denote the total number of true null hypotheses, and m_1 the total number of false null hypotheses. Associated to each hypothesis $i = 1, ..., m$ is a p-value p_i. A multiple testing procedure will reject some subset of the hypotheses (the *rejection set*) based on the observed p-values. The results are typically summarized in a 2×2 table as follows (Benjamini and Hochberg, 1995; Storey et al., 2004).

	Not rejected	Rejected	Total
$H_i = 0$	U	V	m_0
$H_i = 1$	T	S	m_1
Total	W	R	m

Table 1.1: Classification of rejected hypotheses

Under this notation, classical multiple testing procedures aim to control the family-wise error rate (FWER), defined as $P(V \geq 1)$. We emphasise the concept of *control* with respect to a given error rate and nominal level $\alpha \in (0, 1)$. A multiple testing procedure is said to control the FWER (or any other error rate) at level α if

$$P(V \geq 1) \leq \alpha,$$

where V is based on the rejection set of the procedure, and the probability is taken with respect to the model which generates the p-values. If a multiple testing procedure satisfies the weaker condition

$$\lim_{m \to \infty} P(V \geq 1) \leq \alpha,$$

then it is said to control FWER asymptotically at level α. As a concrete example, consider the well-known Bonferroni procedure. The level-α Bonferroni procedure rejects

$$\{H_i : p_i \leq \alpha/m, i = 1, ..., m\},$$

and is known to control the FWER at level α for any $m_0 \leq m$, and with no restrictions on the dependence among the p-values.

1.1 False discovery rate

Despite the popularity of the Bonferroni procedure, it is highly conservative, and Benjamini and Hochberg (1995) remarked that it had been underused in applied research, in large part due to its low power. They went on to note that the FWER was perhaps too stringent of an error rate, and in many applications, an alternative was needed when the concern of inference was identifying a set made up mostly of false null hypotheses, and a single erroneous rejection was not costly. To this end they defined the *false discovery proportion*,

$$\text{FDP} = \frac{V}{R \vee 1}$$

where $a \vee b = \max\{a, b\}$. The FDP gives the proportion of total rejections (R) which are false rejections (V). Then Benjamini and Hochberg (1995) defined their error rate, the *false discovery rate*,

$$\text{FDR} = E[\text{FDP}],$$

where the expectation is taken with respect to the model which generates the p-values. Notice that when $R = 0$, it follows that FDP $= 0$. Furthermore, when $m_0 = m$, $R = V$, and

$$\text{FDR} = P(V \geq 1),$$

that is, it coincides with the FWER (Benjamini and Hochberg, 1995). This means that control of the FDR implies weak control of the FWER. Benjamini (2010) remarks that this relationship to FWER is why they chose to adopt the above definition of FDR in their initial paper, rather than related alternative quantities eventually termed *positive FDR* (pFDR) by Storey (2002), and *marginal FDR* by Genovese and Wasserman (2002):

$$\begin{aligned} \text{pFDR} &= E\left[\frac{V}{R}\middle| R > 0\right] = \frac{\text{FDR}}{P(R>0)}, \\ \text{mFDR} &= \frac{E[V]}{E[R]}. \end{aligned}$$

Along with this new error rate, Benjamini and Hochberg (1995) provided an FDR controlling procedure, inspired by work of Holm (1979) and Simes (1986) on sequential improvements of the Bonferroni procedure. Denote the ordered p-values by $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. Then their multiple testing procedure, which has become known as simply the *Benjamini-Hochberg (BH) procedure* rejects

$$\{H_i : p_{(i)} \leq p_{(k^*)}\},$$

where

$$k^* = \max\left\{i : p_{(i)} \leq \frac{i\alpha}{m}\right\}.$$

Benjamini and Hochberg (1995) went on to show that under the assumption that p-values follow the so-called *null independence model* (the true null p-values are independent of each other and the false null p-values), the BH procedure controls the FDR at level α. More precisely, the BH procedure controls the FDR at level $\pi_0\alpha$, where $\pi_0 = m_0/m$ is the proportion of true null hypotheses. These results have since been strengthened to a broader class of p-value dependence models (Benjamini and Yekutieli, 2001; Heesen and Janssen, 2015).

1.2 FDR estimation and control

By providing a procedure which can control the FDR at level α, Benjamini and Hochberg (1995) followed the traditional paradigm of the hypothesis testing literature: they fixed an error rate α and found a suitable rejection region $[0, p_{(k^*)}]$. The pioneering work of Storey (2002) and later Storey et al. (2004) proposed approaching this problem from the opposite

direction: fix a rejection region of the form $[0, \gamma]$ for $\gamma \in [0, 1]$, and estimate the FDR. Storey et al. (2004) viewed R, V, S, and FDR as empirical stochastic process indexed by the rejection threshold t. More precisely, for $t \in [0, 1]$, define the counting processes

$$\begin{aligned} R(t) &= |\{p_i : p_i \leq t\}|, \\ V(t) &= |\{p_i : p_i \leq t, H_i = 0\}|, \\ S(t) &= |\{p_i : p_i \leq t, H_i = 1\}|. \end{aligned}$$

Using this notation,

$$\text{FDR}(t) = E\left[\frac{V(t)}{R(t) \vee 1}\right].$$

$\text{FDR}(t)$ is the FDR of the multiple testing procedure with rejection region $[0, t]$. Under the null independence model, since the true null p-values are independent and identically distributed as Uniform$[0, 1]$ random variables, it follows that

$$V(t) \sim \text{Binomial}(m_0, t).$$

Furthermore, $R(t)$ is known given the observed p-values, and a natural estimator for $\text{FDR}(t)$ arises as

$$\widehat{\text{FDR}}(t) = \frac{\hat{E}[V(t)]}{R(t) \vee 1} = \frac{m\hat{\pi}_0 t}{R(t) \vee 1},$$

where $\hat{\pi}_0$ is some estimator of π_0, the proportion of true null hypotheses. For a tuning parameter $\lambda \in [0, 1)$, Storey (2002) proposed a widely used π_0-estimator

$$\hat{\pi}_0(\lambda) = \frac{m - R(\lambda)}{(1 - \lambda)m}.$$

The rationale behind $\hat{\pi}_0(\lambda)$ is that in the upper tail region $(\lambda, 1]$, most of the p-values correspond to true null hypotheses, so that

$$m - R(\lambda) \approx m_0 - V(\lambda).$$

Then once again using the binomial distribution of $V(t)$ given by the null independence model,

$$E[m_0 - V(\lambda)] = m_0(1 - \lambda).$$

Combining these two facts and rearranging gives that $\hat{\pi}_0(\lambda) \approx \pi_0$. Notice that

$$m - R(\lambda) = (m_0 - V(\lambda)) + (m_1 - S(\lambda)) \geq m_0 - V(\lambda)$$

will always hold, so this estimator will tend to overestimate π_0, which means that it is a conservative estimator. Using $\hat{\pi}_0(\lambda)$ in $\widehat{\text{FDR}}$ leads to

$$\widehat{\text{FDR}}_\lambda(t) = \frac{m\hat{\pi}_0(\lambda)t}{R(t) \vee 1}.$$

Liang and Nettleton (2012) later showed that $\widehat{\text{FDR}}_\lambda(t)$ is a conservative estimator of $\widehat{\text{FDR}}(t)$, that is

$$E[\widehat{\text{FDR}}_\lambda(t)] \geq \text{FDR}(t).$$

As it is good practice to bound $\hat{\pi}_0$ away from zero, Storey et al. (2004) proposed an asymptotically equivalent π_0-estimator

$$\hat{\pi}_0^*(\lambda) = \frac{m - R(\lambda) + 1}{(1-\lambda)m}.$$

Because $\hat{\pi}_0^*(\lambda) \geq \hat{\pi}_0(\lambda)$, using $\hat{\pi}_0^*(\lambda)$ in $\widehat{\text{FDR}}$ also leads to conservative estimation of the FDR.

The next major contribution of Storey et al. (2004) was to strengthen the relationship between FDR estimation and FDR control. They note that an estimator $\widehat{\text{FDR}}(t)$ of $\text{FDR}(t)$ naively motivates a plug-in multiple testing procedure which finds

$$t_\alpha(\widehat{\text{FDR}}) = \sup\{0 \leq t \leq 1 : \widehat{\text{FDR}}(t) \leq \alpha\},$$

and rejects all null hypotheses with $p_i \leq t_\alpha(\widehat{\text{FDR}})$. We will refer to multiple testing procedures of this type as *thresholding procedures*.

Lemma 1 of Storey et al. (2004) showed that the BH procedure is equivalent to the thresholding procedure with $\widehat{\text{FDR}}$ replaced by $\widehat{\text{FDR}}_{\lambda=0}$. This illuminated the close relationship between FDR estimation and control by demonstrating that the BH procedure, which is known to control FDR at level α, can be characterized as a thresholding procedure for a particular conservative estimate of the FDR process. Storey et al. (2004) then showed that this FDR control property extends to another class of thresholding procedures, where $\widehat{\text{FDR}}$ is replaced by

$$\widehat{\text{FDR}}_\lambda^*(t) = \begin{cases} \frac{m\hat{\pi}_0^*(\lambda)t}{R(t)\vee 1} & t \leq \lambda, \\ 1 & t > \lambda. \end{cases}$$

Note that the thresholding procedure based on this estimator will never reject any p-values greater than λ, but as long as λ is well chosen, this should rarely affect the procedure

in practice. Storey et al. (2004) motivated this fact in their Remark 1. The proof that the thresholding procedure with $\widehat{\text{FDR}}^*_\lambda$ controls the FDR is yet another highly influential contribution of Storey et al. (2004). It is the first application of martingale theory, in particular the optional stopping theorem (Karlin and Taylor, 1975), to the problem of FDR control. This proof will be discussed in more detail in Chapter 2.

When $\hat{\pi}^*_0(\lambda) < 1$ and $t_\alpha(\widehat{\text{FDR}}^*_\lambda) \leq \lambda$, Storey's thresholding procedure is more powerful than the BH procedure, since

$$\widehat{\text{FDR}}^*_\lambda(t) < \widehat{\text{FDR}}_{\lambda=0}(t) \ \forall t \in [0, \lambda],$$

which implies

$$t_\alpha(\widehat{\text{FDR}}^*_\lambda) \geq t_\alpha(\widehat{\text{FDR}}_{\lambda=0}),$$

and so it must reject at least as many p-values as the BH procedure. This increase in power comes from the adaptive nature of the procedure. Benjamini and Yekutieli (2001) showed that the BH procedure with nominal level α has FDR exactly equal to $\pi_0\alpha$. This implies that to control the FDR exactly at level α, the BH procedure should actually be run with nominal level $\alpha^* = \alpha/\pi_0$. Storey's thresholding procedure, by introducing an estimate of π_0, *adapts* to the proportion of true null hypotheses, achieves tighter control of the FDR, and is able to reject more null hypotheses when run at the same nominal level.

Since $\hat{\pi}^*_0(\lambda)$ is a conservative estimator of π_0, Storey's thresholding procedure still does not achieve exact control at level α. Its power depends on the estimation accuracy of $\hat{\pi}^*_0(\lambda)$, which is largely influenced by the choice of λ. In practice, λ should be chosen to balance a bias-variance trade-off. If $\lambda \approx 0$, then $\hat{\pi}^*_0(\lambda)$ has low variance, but high bias, since $m - R(\lambda)$ will tend to count many false null p-values. On the other hand, if $\lambda \approx 1$, then $\hat{\pi}_0(\lambda)$ has low bias, but high variance.

Storey et al. (2004) proposed a heuristic bootstrap method which attempts to minimize mean-squared error (MSE) to choose λ from a fixed and finite set of candidate values $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_k\}$. However, they admit that their theoretical results (conservative estimation and control of the FDR) apply only in the fixed λ case, not when λ is selected using the observed p-values. The setting where λ is selected *dynamically* from the data was studied by Liang and Nettleton (2012).

Liang and Nettleton (2012) worked with Storey's thresholding procedure, and a particular class of selection rules for λ, the *stopping time* rules. Motivated by the martingale arguments of Storey et al. (2004), they restrict the tuning parameter λ to be a stopping time with respect to the filtration generated by the p-values. These concepts from martingale theory will be defined in detail in Chapter 2. Put plainly, Liang and Nettleton (2012)

require that for all $s \in [0,1)$, it is possible to determine whether or not $\lambda \leq s$ without knowledge of the locations of the p-values in the upper tail region $(s,1]$. Under the null independence model, they were able to establish that $\hat{\pi}_0(\lambda)$ and $\widehat{\text{FDR}}_\lambda(t)$ are conservative estimators whenever λ is a stopping time and is bounded away from 0 and 1.

Under some weak dependence among true null p-values, Liang and Nettleton (2012) were also able to establish asymptotic control of the FDR whenever λ is selected dynamically from a fixed and finite set of candidates. However, like Storey et al. (2004) they were unable to show that any of the thresholding procedures with dynamic λ possess finite sample FDR control. Establishing this property for a class of λ-selection rules, *left-to-right selection* (LRS) rules, will be the focus of Chapter 2 of this thesis. The LRS rules comprise a particular subset of the stopping time rules of Liang and Nettleton (2012), for which the set of candidate values for λ is finite, but may be data-dependent. This class will be shown to be broad enough to encompass the most powerful dynamic adaptive procedures in the literature, including the *right-boundary procedure* (Liang and Nettleton, 2012).

1.3 Grouped hypotheses

The development to this point has operated under the implicit assumption of exchangeability of the null hypotheses. According to Remark 1 of Storey (2002), for two p-values p_1 and p_2 coming from exchangeable hypotheses, if p_2 is rejected and $p_1 \leq p_2$, then p_1 should also be rejected. This is why it is sufficient under the FDR estimation approach to restrict attention to rejection regions of the form $[0,t]$. However, without this exchangeability assumption, more flexible rejection regions need to be considered to maximize the power of an FDR controlling procedure.

In particular, suppose the null hypotheses and associated p-values have come from K known groups, and are exchangeable within each group. Denote the hypotheses and p-values by $H_{k,i}$ and $p_{k,i}$ for $i = 1,...,m_k$ and $k = 1,...,K$, so that $m = m_1 + \cdots m_K$. In this setting, the FDR can still be defined as in Section 1.1 based on total rejections R and total false rejections V summed across the K groups. While multiple testing procedures developed under the assumption of exchangeability will still maintain their theoretical guarantees of FDR control in this setting, it is possible to improve power by allowing different rejection thresholds for each of the K groups (Cai and Sun, 2009).

The typical approach in this grouped setting is to weight the p-values according to their group labels so that they can be pooled and treated as exchangeable. Genovese et al. (2006) demonstrated that when p-values are weighted according to some a priori

known weights, the BH procedure applied to the weighted p-values maintains control of the FDR. Roquain and Van De Wiel (2009) derived optimal weights when the distributions of false null p-values are known and the number of rejections is fixed. However, neither of these procedures have data-driven equivalents with weights derived from the data. Other hypothesis weighting methods with data-driven implementations only control the FDR asymptotically (Hu et al., 2010; Zhao and Zhang, 2014).

The aim of these p-value weighting methods is to construct a ranking of significance for the hypotheses which is more informative than the naive pooled ranking of the p-values. Working in the Bayesian setting, Cai and Sun (2009) derived the optimal ranking of significance to maximize power at a fixed level of the FDR. They gave an oracle multiple testing procedure that controls the FDR, but in the data-driven case, their procedure has only asymptotic control of the mFDR.

In order to construct a data-driven procedure that maximizes power while maintaining finite sample control of the FDR, we invoke the idea of *knockoffs* (Barber and Candès, 2015). The knockoff filter is designed in the context of variable selection in linear regression. It is able to control the FDR for selected variables, that is the expected proportion of falsely selected coefficients. Barber and Candès (2015) were able to guarantee control of the FDR under any dependence structure of the variables, through the construction of knockoff variables, and the careful censoring of available information. In the context of multiple hypothesis testing with general covariate information, Lei and Fithian (2018) applied the same concept in their *AdaPT* procedure. In Chapter 3, by combining both the optimal ranking of significance of Cai and Sun (2009) and the finite sample control property of Lei and Fithian (2018), we describe both oracle and data-driven versions of the so-called *grouped mirrored knockoff* (GMK) procedure for FDR control in the grouped multiple testing problem.

The remainder of this chapter will more extensively review the literature on FDR control and estimation, with a focus on the particular problems considered in Chapters 2 and 3. Chapter 2 will consider the FDR control properties of the dynamic adaptive procedures in the exchangeable case, and Chapter 3 will introduce a general framework for FDR controlling procedures in the grouped case, as well as a new method, theory, and implementation.

1.4 Review of Literature

1.4.1 Bayesian multiple hypothesis testing

The FDR as defined by Benjamini and Hochberg (1995), and the FDR control properties of their procedure are purely frequentist in nature, with the statuses of the hypotheses H_i fixed throughout. However, FDR also has a clear interpretation in the Bayesian setting, when the hypotheses are treated as random variables. The Bayesian interpretation of FDR has been illuminated in detail by Efron (2010). The typical assumption in the Bayesian setting is the *two-group model* for the p-values (Efron et al., 2001), which assumes

$$\begin{cases} H_i \sim \text{Bernoulli}(1-\pi_0), \\ p_i|H_i \sim H_i f_1 + (1-H_i) f_0 \end{cases}$$

for $i = 1, ..., m$, where the H_i are independent, and the p_i are conditionally independent given the H_i. f_0 is the uniform density on $[0, 1]$, and f_1 is a density supported on $[0, 1]$ which gives the distribution of the false null p-values. It is also common in the Bayesian framework to work with so-called z-values rather than p-values (Efron, 2004). z-values $\{z_i\}_{i=1}^m$, can be calculated from the observed test statistics such that z_i follows a standard normal distribution given $H_i = 0$. When working with z-values, the two-group model can be written as

$$\begin{cases} H_i \sim \text{Bernoulli}(1-\pi_0), \\ z_i|H_i \sim H_i g_1 + (1-H_i) \phi_0 \end{cases}$$

for $i = 1, ..., m$, where ϕ_0 is the standard normal density and g_1 is a density which gives the distribution of the false null z-values. Assuming the two-group model, each p-value has a marginal distribution given by the mixture density

$$f = (1-\pi_0) f_1 + \pi_0 f_0,$$

and each z-value has mixture density

$$g = (1-\pi_0) g_1 + \pi_0 \phi_0.$$

In this Bayesian setting, for a fixed rejection threshold t, the mFDR has an interpretation as the posterior probability that a null hypothesis is true given its p-value is in the rejection region. Under the two-group model it can be written as

$$\text{mFDR}(t) = \frac{E[V(t)]}{E[R(t)]} = P(H_1 = 0 | p_1 \leq t) = \frac{\pi_0 t}{F(t)}$$

by Bayes' rule, where F is the marginal CDF of each p-value. Efron (2008) refers to this quantity as the *Bayesian FDR*. The usual FDR still has an interpretation in the Bayesian setting, with

$$\text{FDR} = E\left[\frac{V}{R \vee 1}\right]$$

now unconditional on the hypotheses. That is, in the Bayesian setting, frequentist FDR is given by a conditional expectation,

$$\text{FDR} = E\left[\frac{V}{R \vee 1}\middle| H_1, ..., H_m\right].$$

The Bayesian setting also gives rise to the *local FDR* (Lfdr), the posterior probability a hypothesis is true. Assuming the two-group model, the Lfdr has a simple formulation as

$$\text{Lfdr}(p) = P(H_1 = 0|p_1 = p) = \frac{\pi_0 f_0(p)}{f(p)}, \text{ or}$$

$$\text{Lfdr}(z) = P(H_1 = 0|z_1 = z) = \frac{\pi_0 \phi_0(z)}{g(z)}.$$

Sun and Cai (2007) worked with the two-group model for the z-values, and analyzed Bayesian multiple hypothesis testing in the compound decision framework. They viewed the multiple testing problem as a classification problem with weighted 0-1 loss function

$$L(H, \delta) = \sum_{i=1}^{m} \{\lambda I(H_i = 0)\delta_i + I(H_i = 1)(1 - \delta_i)\},$$

where $\delta = (\delta_1, ..., \delta_m) \in \{0, 1\}^m$ is a general decision rule with $\delta_i = 1$ indicating the rejection of the ith hypothesis, and $\lambda > 0$ is a relative weight for the misclassification cost of a false positive. Sun and Cai (2007) showed that the optimal decision rule for the control of the mFDR is

$$\delta_i = I(\text{Lfdr}(z_i) < c^*),$$

where c^* is the maximal threshold such that the procedure controls the mFDR at level α. This optimality is in the sense that the procedure minimizes the *marginal false non-discovery rate* (mFNR), defined as

$$\text{mFNR}(t) = P(\delta_1 = 0|H_1 = 1),$$

the probability that a hypothesis is not rejected given that it is false (akin to frequentist type II error). Minimization of the mFNR is equivalent to maximization of power. Another

interpretation of this result of Sun and Cai (2007) is that it establishes the Lfdr as the optimal statistic to rank the significance of hypotheses.

The stochastic process view of R, V and FDR in the Bayesian setting has been treated in detail by Genovese and Wasserman (2002, 2004). Genovese and Wasserman (2004) established many properties of the FDP and FDR processes. Under the two-group model, Genovese and Wasserman (2004) showed that

$$\text{FDR}(t) = \frac{\pi_0 t}{F(t)} \cdot (1 - F(t))^m$$

and

$$\text{FDR}(t) = \text{mFDR}(t) + o(m^{-1/2}),$$

i.e. the asymptotic equivalence of the FDR and the mFDR. These results establish the FDR estimator of Storey (2002) as a conservative plug-in estimator of the FDR, which replaces $F(t)$ by the empirical CDF

$$\hat{F}(t) = \frac{1}{m} R(t),$$

π_0 by $\hat{\pi}_0(\lambda)$, and ignores the second factor $(1 - F(t))^m \approx 1$. Genovese and Wasserman (2004) also showed the asymptotic normality of the estimator $\hat{\pi}_0(\lambda)$, and that the limiting distribution of the FDP process is a Gaussian process. They derived asymptotic and exact confidence envelopes for FDP, random functions $\Gamma(t)$ for which

$$P(\text{FDP}(t) \leq \Gamma(t) \ \forall t) \geq 1 - q$$

for some confidence level $q \in (0, 1)$. By appealling to these results, they established asymptotic control of the FDR for plug-in thresholding procedures similar to those of Storey et al. (2004).

1.4.2 Adaptive procedures for FDR control

Storey's thresholding procedure is just one example within the class of *adaptive procedures* for FDR control. Adaptive procedures encompass all FDR controlling procedures which apply the BH procedure at level $\hat{\pi}_0 \alpha$ for some π_0-estimator $\hat{\pi}_0$. As a further distinction, procedures for which $\hat{\pi}_0$ has prespecified or fixed tuning parameters will be referred to as *fixed adaptive procedures*, while procedures for which all tuning parameters are chosen dynamically from the data will be referred to as *dynamic adaptive procedures*. Storey's

thresholding procedure with fixed $\lambda = 1/2$ (as Storey et al. (2004) chose in their paper) is an example of a fixed adaptive procedure, while Storey's thresholding procedure with their heuristic bootstrap method used to select λ is an example of a dynamic adaptive procedure. Since the original definition of the BH procedure, many authours have devised their own adaptive versions.

The first such adaptive BH procedure to appear was the dynamic adaptive procedure with the so-called *lowest-slope* estimator of Benjamini and Hochberg (2000). In this case π_0 is estimated by

$$\hat{\pi}_0^{(\mathrm{LSL})} = \frac{m - j + 1}{m(1 - p_{(j)})},$$

where $2 \leq j \leq m$ is the smallest index such that $\hat{\pi}_0(p_{(j)}) > \hat{\pi}_0(p_{(j-1)})$. The name "lowest-slope" comes from a motivating graphical interpretation given by Schweder and Spjøtvoll (1982). The details of this procedure will be discussed further in Chapter 2. In their original paper Benjamini and Hochberg (2000) were only able to demonstrate the FDR control of this procedure through simulation and were unable to prove it rigorously. In fact, according to Benjamini (2010), the non-adaptive BH procedure was presented as an alternative only because the control of this lowest-slope version could not be established.

In a later paper, Benjamini et al. (2006) develop two new adaptive procedures. Their first procedure is a dynamic adaptive procedure in which π_0 is estimated by

$$\hat{\pi}_0^{(\text{Q-k})} = \hat{\pi}_0(p_{(k)})$$

for some fixed k. In effect this is Storey's thresholding procedure, but with λ selected as a fixed quantile of the empirical distribution of the observed p-values. Benjamini et al. (2006) suggest choosing $k = \lfloor m/2 \rfloor$ so that $p_{(k)}$ is the median of the observed p-values. Their second procedure is a fixed adaptive procedure that operates in two-stages. From the definition of Benjamini et al. (2006) it is not immediately clear that this procedure is an adaptive procedure as defined above, but it is shown by Sarkar (2008) that it is equivalent to the fixed adaptive procedure with π_0-estimator

$$\hat{\pi}_0^{(\mathrm{BKY})}(\lambda) = \frac{m - R^{\mathrm{BH}}(\lambda)}{m(1 - \lambda)}$$

where $R^{\mathrm{BH}}(\lambda)$ is the number of rejections found using the BH procedure at level λ, which Benjamini et al. (2006) specify as $\lambda = \alpha/(1 + \alpha)$. Hence the characterization of their procedure as a two-stage procedure: at the first stage the BH procedure is run at level λ, and the number of rejections from this procedure is used to estimate the total number of true null hypotheses. This estimate is incorporated in the second stage, an adaptive BH

procedure. Benjamini et al. (2006) showed that both of their adaptive procedures control the FDR.

The conservative estimation results of Liang and Nettleton (2012) apply to any dynamic adaptive procedure which uses Storey's thresholding procedure with λ a stopping time, which includes the lowest-slope procedure of Benjamini and Hochberg (2000), and the fixed quantile procedure of Benjamini et al. (2006). Furthermore, they proposed a new dynamic adaptive procedure, the *right-boundary procedure*, which was initially used by Mosig et al. (2001) to estimate π_0 from the histogram of the p-values. It estimates π_0 using Storey's estimator $\hat{\pi}_0(\lambda)$, but with λ selected as the right boundary of the first histogram bin with bin count less than or equal to the tail average. Liang and Nettleton (2012) suggest choosing 20 equally spaced histogram bin boundaries on the interval $[0, 1]$.

More recently, Heesen and Janssen (2016) have proposed a class of dynamic adaptive procedures based on the so-called *generalized Storey estimator*

$$\hat{\pi}_0^*(\lambda, \gamma) = \frac{R(\gamma) - R(\lambda) + 1}{m(\gamma - \lambda)}$$

for two tuning parameters $0 < \lambda < \gamma \leq 1$. While the original π_0-estimator $\hat{\pi}_0^*(\lambda)$ of Storey et al. (2004) counts the proportion of p-values in the upper tail region $(\lambda, 1]$ for a single tuning parameter λ, the generalized Storey estimator counts the proportion of p-values in the central region $(\lambda, \gamma]$. The dynamic adaptive procedure of Heesen and Janssen (2016) estimates π_0 by

$$\sum_{i=1}^{k} \hat{\beta}_i \cdot \hat{\pi}_0^*(\lambda_i - 1, \lambda_i)$$

where $0 < \lambda_0 < \lambda_1 < \cdots < \lambda_k = 1$ form a grid of fixed tuning parameters, and $\hat{\beta}_i$ are non-negative data-dependent weights satisfying $\sum_{i=1}^{k} \hat{\beta}_i = 1$. They showed that under a measurability condition on the weights, this dynamic adaptive procedure controls the FDR. In particular their procedure requires that for each $i = 1, ..., k$, $\hat{\beta}_i$ can be calculated without knowledge of the locations of the p-values in the lower tail region $[0, \lambda_i)$.

1.4.3 Dependence

Much of the FDR control literature works under the null independence model for the p-values, originally introduced by Benjamini and Hochberg (1995). As mentioned above, the null independence model allows for arbitrary dependence among the false null p-values,

but require that the true null p-values are independent of each other and the false null p-values. FDR control under more relaxed assumptions has been studied extensively.

The first major work in this direction was by Benjamini and Yekutieli (2001), who showed that the BH procedure maintains control of the FDR under a dependence model they term *positive regression dependency on a subset* (PRDS). Their original characterization of the PRDS condition is difficult to interpret, but an equivalent, more instructive characterization was given by Sarkar (2008). Let $\psi : [0,1]^m \to \mathbb{R}$ be any coordinate-wise non-decreasing function. Then the PRDS condition holds if

$$E[\psi(p_1, ..., p_m)|p_i = u]$$

is a non-decreasing function of u, for each i such that $H_i = 0$. The null independence model trivially satisfies the PRDS condition. As a particular example, consider a vector of test statistics $(X_1, ..., X_m) \sim \text{MVN}(\mu, \Sigma)$ each testing the hypothesis $H_i : \mu_i = 0$ against the one-sided alternative $\mu_i > 0$, and suppose that Σ is a correlation matrix. Then the p-values are calculated as $p_i = 1 - \Phi_0(X_i)$ for $i = 1, ..., m$, where Φ_0 is the standard normal CDF. This model will be referred to as the *normal means* multiple testing problem. Then p-values arising from the normal means problem will satisfy the PRDS condition if each off-diagonal element of Σ is non-negative.

Sarkar (2008) also gives the slightly more relaxed dependence condition, *positive dependence*, which holds if

$$E[\psi(p_1, ..., p_m)|p_i \leq u]$$

is a non-decreasing function of u, for each i such that $H_i = 0$. Sarkar (2008) showed that the BH procedure also maintains control of the FDR under positive dependence. Benjamini and Yekutieli (2001) also proved that if the BH procedure is run at nominal level

$$\frac{\alpha}{\sum_{i=1}^{m} 1/i}$$

then it controls the FDR at level α under arbitrary dependence.

An alternative dependence model is introduced by Heesen and Janssen (2015) in both the Bayesian and frequentist settings, *reverse martingale dependence*. This model is motivated by the counting process view of Storey et al. (2004) and can be viewed as a mathematical construction, the widest class of dependence models under which their original proof will still hold. The reverse martingale condition requires that

$$\frac{P(p_i \leq t|p_i \leq s)}{t} = \frac{1}{s}$$

for all $0 \leq t \leq s \leq 1$, and all i such that $H_i = 0$. The reverse martingale model contains the null independence model. Heesen and Janssen (2015) showed that the BH procedure maintains control of the FDR under the reverse martingale model. They also presented a negative result, which demonstrated that there are PRDS and reverse martingale dependence models under which Storey's (fixed adaptive) thresholding procedure does not control the FDR.

In applications, it may be reasonable to assume dependence takes on a block structure: that hypotheses can be partitioned into disjoint blocks with dependence within blocks but not between them. Dependence models of this type have been studied by Guo and Sarkar (2016) and Heesen and Janssen (2015). In the block dependence setting, where each block satisfies the positive dependence condition, Guo and Sarkar (2016) defined a modified adaptive procedure which operates in two stages. At the first stage, an adaptive version of the BH procedure is run on the minimal p-value from each block. At the second stage, rejections are restricted to only the rejected blocks from the first stage. Their procedure maintains control of the FDR, but it can become overly conservative, especially as block size increases, such that it offers no power improvement over the BH procedure. They also give a similar non-adaptive two-stage procedure which maintains control of the FDR under arbitrary block dependence. Heesen and Janssen (2015) instead assumed that each block satisfies the reverse martingale condition. They modified the estimator in Storey's thresholding procedure, and demonstrated that when it is made sufficiently conservative, the fixed adaptive procedure maintains control of the FDR. Especially as the block size increases, the conservatism required of their π_0-estimator can make it likely that $\hat{\pi}_0 \geq 1$, and their procedure will offer no power improvement over the BH procedure.

The literature as a whole demonstrates that while the FDR control property of the BH procedure is relatively robust to dependence, valid adaptive modifications of the BH procedure are limited under dependence. It should also be noted that many adaptive BH procedures do possess asymptotic control of the FDR under dependence among the true null p-values. Liang and Nettleton (2012) showed that the dynamic adaptive version of Storey's thresholding procedure has asymptotic control of the FDR whenever λ is selected from a fixed and finite candidate set, and the p-values satisfy *weak dependence*. Weak dependence was originally introduced by Storey et al. (2004) and requires

- $\lim_{m\to\infty} \frac{V(t)}{m_0} = F_0(t) \leq t$ pointwise a.s.;
- $\lim_{m\to\infty} \frac{R(t)-V(t)}{m_1} = F_1(t)$ pointwise a.s.;
- $\lim_{m\to\infty} m_0/m$ exists.

which includes arbitrary block dependence as a special case (as long as the blocks have finite size). As a final note on dependence models, Fan and Han (2017), among others, have remarked that although many multiple testing procedures which are developed in the independent case can be shown to maintain control of the FDR under some dependence models, they will still suffer from a loss of efficiency if this dependence is not accounted for. Working with p-values arising from the normal means problem, Fan et al. (2012) derive a dependence-adjusted estimator of the FDP. Their estimator incorporates the first k principal components of Σ, the covariance matrix of the test statistics. When Σ is unknown, but can be estimated consistently by $\hat{\Sigma}$ (and subject to some regularity conditions), Fan and Han (2017) showed that their estimator is still conservatively consistent. Results like this that rely on the estimation of the dependence structure of the p-values naturally require repeated observations of test statistics.

1.4.4 Non-exchangeability

Very shortly after their initial paper, Benjamini and Hochberg (1997) extended the FDR and the BH procedure to the weighted case, where hypotheses are not exchangeable. Benjamini and Hochberg (1997) work with what they call *procedural weights*, which reflect a known ranking of importance of the hypotheses. Their canonical example is in meta-analysis, where hypotheses are ranked by the quality of the individual studies. Rather than working with the FDR, their procedure considers the *weighted FDR* (wFDR), defined as

$$\text{wFDR} = E\left[\frac{\sum_{i=1}^{m} w_i V_i}{\left(\sum_{i=1}^{m} w_i R_i\right) \vee 1}\right],$$

where V_i indicates whether hypothesis i is rejected and is a true null, and R_i indicates whether hypothesis i is rejected. $\{w_i\}_{i=1}^m$ are hypothesis weights satisfying $\sum_{i=1}^m w_i = m$. In this sense the wFDR penalizes the false rejection of some hypotheses more than others. Benjamini and Hochberg (1997) also developed a procedure similar to the BH procedure which controls the wFDR.

This idea of hypothesis weighting has remained important in the FDR literature, where the weights reflect the prior belief that a particular null hypothesis is false. Working with known weights $\{w_i\}_{i=1}^m$ satisfying $\sum_{i=1}^m w_i = m$, Genovese et al. (2006) showed that under the Bayesian setting and the two-group model, when the BH procedure is applied to the weighted p-values

$$p_i^* = p_i / w_i,$$

it controls the FDR conservatively at level α. They went on to show through simulation that when the weights are well chosen, that is when large weights correlate with a high probability of a false null hypothesis, there is potential for a substantial increase in power from their weighted procedure.

Continuing from the work of Genovese et al. (2006), Roquain and Van De Wiel (2009) sought out the optimal weights to maximize the power of the BH procedure, assuming a fully known p-value model. They operate under full independence of the p-values, and either the frequentist or Bayesian setting, which they term the conditional and unconditional models, respectively. Under their unconditional model, which is equivalent to the two-group model, each null hypothesis has a constant probability π_0 of being true. They also suppose that each false null p-value has a known continuous distribution F_i on $[0, 1]$, which may be different for each p-value, and has strictly decreasing density f_i. Then their main result is that for a fixed proportion of rejections u, and at nominal level α, the optimal p-value weighting is given by

$$w_i^*(u) = (\alpha u)^{-1} f_i^{-1}(y^*(u)) \cdot I(H_i = 1)$$

under the conditional model, and

$$w_i^*(u) = (\alpha u)^{-1} f_i^{-1}(y^*(u))$$

under the unconditional model, where $y^*(u)$ is a constant chosen so that the weights sum to m. Note that in the conditional case, true null hypotheses are given weight zero, so that $p_i^* = \infty$, and they are not rejected. Roquain and Van De Wiel (2009) gave an instructive example where they note that when the p-values arise from the normal means problem with $\Sigma = I$, so that $F_i \sim \mathrm{N}(\mu_i, 1)$, the optimal weights will have a bell-curve shape as a function μ_i. They will tend to be small when $\mu_i \approx 0$, increase to a maximum as μ_i increases, to effectively magnify these mid-sized signals, and then decay to zero as $\mu_i \to \infty$, since these hypotheses should have sufficiently small p-values to be identified as significant without the aid of large weights.

While these optimal weights are an interesting mathematical construction, they are unwieldy to implement, and Roquain and Van De Wiel (2009) gave no accompanying data-driven procedure. In practice, allowing each false null p-value to have a different distribution is far too flexible of a model to perform any inference. For this reason, data-driven p-value weighting has been investigated most extensively in the grouped setting. In this case, it is typically assumed that the hypotheses come from K groups, and are exchangeable within each group, so that the scope of the problem of p-value weighting is

reduced to simply finding one weight per group, $\{w_k\}_{k=1}^K$ and applying the BH procedure to the group-weighted p-values

$$p^*_{k,i} = p_{k,i}/w_k.$$

Hu et al. (2010) considered a weighted approach that incorporates the group-wise true null proportions. For $k = 1, ..., K$, they defined $m_{0,k}$ to be the number of true null hypotheses in group k, and

$$\pi_{0,k} = m_{0,k}/m_k,$$

the group-wise true null proportions. Then for each group their oracle *grouped Benjamini-Hochberg* (GBH) procedure sets p-value weights

$$w_k = \frac{1 - \pi_{0,k}}{\pi_{0,k}(1 - \pi_0)},$$

where the overall proportion of true null hypotheses π_0 can be recovered from the group-wise proportions via the relationship

$$\pi_0 = \frac{1}{m} \sum_{k=1}^{K} m_k \pi_{0,k}.$$

Hu et al. (2010) proved that their oracle GBH procedure controls the FDR under the PRDS model. Their data-driven GBH procedure operates by replacing each $\pi_{0,k}$ by some estimate $\hat{\pi}_{0,k}$, for instance the lowest-slope estimator (Benjamini and Hochberg, 2000). The data-driven GBH procedure controls the FDR asymptotically under weak dependence as long as each $\hat{\pi}_{0,k}$ is an asymptotically conservative estimator of $\pi_{0,k}$ (Hu et al., 2010).

When the distributions of false null p-values are constant across groups, Hu et al. (2010) gave a Bayesian argument to motivate the optimality of their oracle weighting procedure by noting that the ranking according to their (oracle) weighted p-values is equivalent to the ranking according to the Bayesian Lfdr. However, if the distributions of the false null p-values differ across groups, Zhao and Zhang (2014) demonstrated through simulation that the GBH procedure can be outperformed in terms of power by adaptive versions of the BH procedure that do not incorporate any group information. As a result they implemented a more flexible group-weighted procedure, the *weighted optimization* (WO) procedure. Under the WO procedure, the group-wise p-value weights are chosen to maximize an objective function. In the oracle case, suppose that $F_{1,k}$, the CDF of the false null p-values in group k, is known for $k = 1, ..., K$. Then the oracle objective function is the power function (written as a function of the group-wise weights),

$$\mathcal{O}(w) = \frac{1}{m} \sum_{k=1}^{K} m_k(1 - \pi_{0,k}) F_{1,k}(w_k \cdot \tilde{t}_M)$$

where $\tilde{t}_M = \max\{\tilde{t}_{BH}, \tilde{t}_H\}$, $\tilde{t}_{BH}$ is the rejection threshold given by Storey's (oracle) adaptive thresholding procedure, and $\tilde{t}_H$ is the rejection threshold for the weighted p-values given by the oracle GBH procedure (Hu et al., 2010). To ensure control of the FDR, $\mathcal{O}$ is maximized subject to the constraints

$$w_k \geq 0, k = 1, ..., K \quad \text{and} \quad \sum_{k=1}^{K} m_k \pi_{0,k} w_k = m.$$

The WO procedure can be thought of as a two-stage procedure: in the first stage, a weighted BH procedure is run to find $\tilde{t}_M$. Then the weights are chosen to optimize rejections at this threshold, and in the second stage, another weighted BH procedure is run with these new optimal weights. Zhao and Zhang (2014) proved that their oracle WO procedure has asymptotic control of the FDR, and has power performance no worse than that of the oracle GBH procedure. In the data-driven case, when the CDF's of the false null p-values are unknown, the power function is estimated using the empirical CDF of the p-values. The data-driven WO procedure chooses weights to optimize

$$\hat{\mathcal{O}}(w) = \frac{1}{m} \sum_{k=1}^{K} \sum_{i=1}^{m_k} I(p_{k,i} \leq w_k \cdot \hat{t}_M),$$

where analogously to the oracle case, $\hat{t}_M = \max\{\hat{t}_{BH}, \hat{t}_H\}$, $\hat{t}_{BH}$ is the rejection threshold given by Storey's adaptive thresholding procedure (with some estimate of π_0), and $\hat{t}_H$ is the rejection threshold for the weighted p-values given by the data-driven GBH procedure (Hu et al., 2010). Again, similar to the oracle case, w is chosen subject to the constraints

$$w_k \geq 0, k = 1, ..., K \quad \text{and} \quad \sum_{k=1}^{K} m_k \hat{\pi}_{0,k} w_k = m,$$

where now $\pi_{0,k}$ must be estimated by $\hat{\pi}_{0,k}$. Zhao and Zhang (2014) proved that their data-driven WO procedure has asymptotic control of the FDR as long as each $\hat{\pi}_{0,k}$ is an asymptotically conservative estimator of $\pi_{0,k}$. Furthermore, they showed that if each $\hat{\pi}_{0,k}$ is a consistent estimator of $\pi_{0,k}$, the power of their data-driven procedure is asymptotically equal to the power of the oracle WO procedure. The WO procedure uniformly improves upon the GBH procedure by taking into account the relative size of groups, as well as the group-wise distributions of false null p-values. However, in the data-driven case, it uses the p-values twice: once to optimize the weights, and again in the weighted BH procedure. Although Zhao and Zhang (2014) demonstrated asymptotic control of the FDR, in some

cases, their procedure can overfit to the observed p-values and lose control of the FDR in finite samples.

Operating under a similar Bayesian setting and compound decision framework to Sun and Cai (2007), Cai and Sun (2009) derived an optimal mutiple testing procedure for the grouped case. They worked with the z-values rather than p-values and define for each $k = 1, ..., K$,

$$\text{Lfdr}_k(z) = P(H_{k,1} = 0 | z_{k,1} = z) = \frac{\pi_{0,k}\phi_0(z)}{g_k(z)},$$

the group-wise Lfdr, where g_k denotes the marginal density of the p-values from group k. That is, they assume that the z-values for each group of hypotheses independently follow the two-group model. They showed that the optimal decision rule for grouped hypothesis testing is their so-called *CLfdr* procedure:

$$\delta_i = I(\text{Lfdr}_k(z_{k,i}) < c^*)$$

where δ_i corresponds to the rejection of the ith hypothesis, and c^* is the maximal threshold such that the procedure controls the overall FDR at level α. As in Sun and Cai (2007), this optimality is in the sense that it minimizes the overall FNR. The interpretation of this result is that the optimal pooled ranking of significance of the hypotheses is according to the group-wise Lfdr. Cai and Sun (2009) noted that c^* is typically difficult to calculate, and so they present an asymptotically equivalent procedure which estimates the overall FDR by the mean of the Lfdr's of the rejected hypotheses. Their data-driven procedure follows immediately by plugging in estimates of the group-wise Lfdr's. Cai and Sun (2009) proved that if the group-wise Lfdr's can be estimated consistently, their data-driven procedure asymptotically controls the mFDR. They also proved that it is asymptotically optimal, in the sense that it achieves mFNR equivalent to the oracle CLfdr procedure.

Several ingenious and flexible procedures for FDR control have also been considered in the case where hypotheses are accompanied by general covariate information, denoted by

$$\{p_i, x_i\}_{i=1}^m,$$

where x_i comes from some general space $\mathcal{X}$. Group information can be viewed as the case when this covariate is categorical with a finite number of levels, that is when $\mathcal{X} = \{1, ..., K\}$. In this more general setting, Ignatiadis et al. (2016) introduce a method, *independent hypothesis weighting* (IHW), which divides the hypotheses into G groups according to their covariate values, and chooses group-wise hypothesis weights $\{w_i\}_{i=1}^G$ to directly optimize the total number of rejections of the weighted BH procedure. As a regularization step, this optimization is done over a constrained subset which restricts the differences between

the weights. IHW also uses a cross-validation scheme to estimate the hypothesis weights: it splits the hypotheses into b folds, and for each fold, estimates the weights using only p-values from the remaining $b-1$ folds. Under the two-group model, Ignatiadis et al. (2016) showed that the BH procedure with these weights controls the FDR asymptotically.

This cross-validation idea of Ignatiadis et al. (2016) is one way to construct a highly flexible, data-driven procdure which maintains asymptotic control of the FDR. Another idea, introduced by Lei and Fithian (2018), is data masking. Their *AdaPT* procedure works in the general covariate setting. It operates sequentially, such that at each step t, the procedure rejects

$$\{H_i : p_i \leq s_t(x_i)\}$$

for some rejection curve $s_t : \mathcal{X} \to [0,1]$. Lei and Fithian (2018) noted that under the null independence model, a true null p-value p_i is equal in distribution to its mirror image $1-p_i$. Thus the number of p-values above the mirrored rejection curve $1-s_t(\cdot)$ can serve as knockoff p-values (Barber and Candès, 2015), and can be used to estimate the FDR of the current rejection region. If the estimated FDR is no greater than α, then AdaPT stops, and otherwise it updates $s_t \to s_{t+1}$. Lei and Fithian (2018) are able to show that under the null independence model, as long as s_t is updated without knowledge of the p-values below $s_t(\cdot)$ or above $1-s_t(\cdot)$, their procedure controls the FDR. Motivated by Cai and Sun (2009), Lei and Fithian (2018) show that in the Bayesian setting, the optimal choice of s_t will be a level curve of the conditional local FDR

$$\text{Lfdr}(p|x) = P(H_1 = 0 | p_1 = p, x_1 = x).$$

Chapter 2

Dynamic adaptive procedures that control the FDR

2.1 Introduction

In this chapter, we strive to prove the FDR control for a broad class of dynamic adaptive procedures, which include the right-boundary procedure (Liang and Nettleton, 2012) as a special case. The proof of FDR control is then extended to a second class of dynamic adaptive procedures which select λ from a data-driven candidate set, in particular the realized p-values. Examples include the lowest-slope procedure of Benjamini and Hochberg (2000) and the k-quantile procedure of Benjamini et al. (2006). The lowest-slope procedure is historically important in the field of multiple testing and especially in the FDR literature. As the earliest adaptive FDR procedure, the lowest-slope procedure is widely used, but its control of the FDR has not been theoretically established.

This chapter is organized as follows. In Section 2.2, we introduce notation and place assumptions on the p-value model, as well as briefly introducing some of the martingale terminology used in the theory to follow. Section 2.3 establishes finite sample FDR control for an initial class of dynamic adaptive procedures with fixed candidate set. In Section 2.4, this result is extended to a further class of dynamic adaptive procedures with data-driven candidate set. In Section 2.5, we report the results of simulations that demonstrate the advantages of dynamic adaptive procedures. Section 2.6 gives some brief discussion of identifiability and dependence. Technical proofs are postponed until the appendix.

2.2 Notation

Recall from Chapter 1 the multiple testing problem with m hypotheses $H_1, H_2, ..., H_m$. For $i = 1, ..., m$, $H_i = 1$ corresponds to a true null hypothesis and $H_i = 0$ to a false null hypothesis. Associated with each hypothesis H_i is a p-value p_i. Denote the ordered p-values by

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

Throughout this chapter, assume that the p-values arise from the null independence model, that is the true null p-values are independent and identically distributed as Uniform$[0, 1]$ random variables, and are independent of the false null p-values. Arbitrary dependence is allowed among the false null p-values. This is the same condition adopted by Benjamini and Hochberg (1995), Storey et al. (2004), and Liang and Nettleton (2012). This model is defined in the frequentist framework, thus the number of true nulls m_0 is fixed. An analogous model defined in the Bayesian framework with random m_0 is termed the *basic independence model* by Heesen and Janssen (2015); note that results in the frequentist model can be easily extended to the Bayesian case by conditioning on the hypotheses $H_1, H_2, ..., H_m$, and integrating.

Recall for $t \in [0, 1]$, the empirical process

$$\begin{aligned} R(t) &= |\{p_i : p_i \leq t\}|, \\ V(t) &= |\{p_i : p_i \leq t, H_i = 0\}|, \\ S(t) &= |\{p_i : p_i \leq t, H_i = 1\}|. \end{aligned}$$

Storey's thresholding procedure for FDR control uses the π_0-estimator

$$\hat{\pi}_0^*(\lambda) = \frac{m - R(\lambda) + 1}{(1 - \lambda)m}.$$

and the FDR estimator

$$\widehat{\text{FDR}}_\lambda^*(t) = \begin{cases} \frac{m\hat{\pi}_0^*(\lambda)t}{R(t) \vee 1} & t \leq \lambda, \\ 1 & t > \lambda. \end{cases}$$

for fixed $\lambda \in [0, 1)$. We refer to this thresholding procedure as the fixed adaptive procedure with tuning parameter λ. When λ is allowed to be data-dependent, we refer to the resulting procedure as the dynamic adaptive procedure, and λ is called a dynamic tuning parameter.

2.2.1 Martingales

Here we will briefly introduce some of the martingale terminology used in the sections to follow. The first notion required is that of a *filtration*. A filtration is an indexed collection of sigma algebras

$$\{\mathcal{S}_t\}_{t\in T}$$

for some subset $T \subseteq \mathbb{R}$, where $\mathcal{S}_s \subseteq \mathcal{S}_t$ for all $s \leq t$ (Karlin and Taylor, 1975). The index t can be thought of as a "time", with progressively more information being revealed as time progresses. In the context of FDR control, the typical filtration of interest specifies $T = (0, 1]$, and for $t \in (0, 1]$,

$$\mathcal{F}_t = \sigma(R(s) : 0 < s \leq t),$$

which gives the location of the p-values below t.

A (continuous-time) *martingale* with respect to a filtration $\{\mathcal{S}_t\}_{t\in T}$ is a stochastic process

$$\{X(t)\}_{t\in T}$$

which satifies the following conditions (Karlin and Taylor, 1975):

- For all $t \in T$, $X(t)$ is measurable with respect to $\mathcal{S}_t$.
- For all $t \in T$, $E[|X(t)|] < \infty$.
- For all $s \leq t$, $E[X(t)|\mathcal{S}_s] = X(s)$.

The first condition states that the martingale is *adapted* to the filtration. In the third condition, if the equality '=' is replaced by an inequality '$\leq$', then $\{X(t)\}_{t\in T}$ is said to be a *supermartingale*.

Some of the most useful results in martingale theory concern stopping times. A *stopping time* τ with respect to a filtration $\{\mathcal{S}_t\}_{t\in T}$ is a random variable such that for all $t \in T$, the event $\{\tau \leq t\}$ is contained in $\mathcal{S}_t$ (Karlin and Taylor, 1975). That is to say, at time t it is always possible to decide whether $\{\tau \leq t\}$ has occurred. In the context of FDR estimation, Liang and Nettleton (2012) work with tuning parameters λ that are stopping times with respect to $\{\mathcal{F}_t\}_{t\in(0,1]}$, that is the event $\{\lambda \leq t\}$ is known given the locations of the p-values below t.

2.3 Fixed grid dynamic adaptive procedures

In proving their results regarding conservative estimation of FDR, Liang and Nettleton (2012) consider any dynamic tuning parameter λ that is a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t\in(0,1]}$, and is bounded away from 0 and 1. However, establishing strong control of the FDR requires a better characterization of the selection behaviour, and so in this section we limit the focus to a subclass of stopping time rules which select λ from a fixed and finite grid of candidate values.

The canonical dynamic adaptive procedure of this type is the *right-boundary* procedure, but the results to follow will apply to a wide class of stopping time rules of which the right-boundary procedure is a special case. For $k \geq 1$, consider a fixed and finite λ candidate grid

$$\Lambda = \{\lambda_1, \ldots, \lambda_k\}$$

that divides the interval $(0,\, 1]$ into $k+1$ bins with boundaries at

$$\lambda_0 \equiv 0 < \lambda_1 < \cdots < \lambda_k < \lambda_{k+1} \equiv 1$$

such that the ith bin is $(\lambda_{i-1}, \lambda_i]$ for $i = 1, \ldots, k+1$. Then construct a sequence of π_0-estimators at candidate λ values as

$$\hat{\pi}_0^*(\lambda_i) = \frac{m - R(\lambda_i) + 1}{(1-\lambda_i)m}, \quad i = 1, \ldots, k.$$

The right-boundary procedure chooses the tuning parameter $\lambda = \lambda_j$, where

$$j = \min\{1 \leq i \leq k : \hat{\pi}_0^*(\lambda_i) \geq \hat{\pi}_0^*(\lambda_{i-1})\}$$

if this set is non-empty, and otherwise chooses $j = k$. That is, λ is chosen as the right boundary of the first bin where the π_0 estimate at its right boundary is larger or equal to that at its left boundary. Let N_i denote the number of p-values falling into the ith bin, i.e.,

$$N_i = R(\lambda_i) - R(\lambda_{i-1}) = |\{j : p_j \in (\lambda_{i-1}, \lambda_i]\}|.$$

Then it is straightforward to show that the right-boundary procedure is equivalent to stopping at the first bin where

$$\frac{N_i}{\lambda_i - \lambda_{i-1}} \leq \frac{m - R(\lambda_{i-1}) + 1}{1 - \lambda_{i-1}}.$$

Thus, another interpretation of the right-boundary procedure is to choose the first bin whose p-value density is less or equal to its tail average. This stopping rule compares the

current bin count N_i to a function of $R(\lambda_{i-1})$, which is the sum of past bin counts, i.e., $R(\lambda_{i-1}) = \sum_{j=1}^{i-1} N_j$.

In general, we require a fixed and finite candidate grid Λ, but allow the stopping rule to depend on any functions of the past bin counts and boundaries. More specifically, a λ-selection rule on a fixed candidate grid $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_k\}$ with $0 < \lambda_1 < \cdots < \lambda_k < 1$ is said to have the *left-to-right stopping time property* (and is called an LRS selection rule) if λ_i is selected when i is the smallest index in $\{1, \ldots, k\}$ such that $N_i \in C_i$, where C_i is a subset of $\{0, ..., m\}$ and can depend on the past bin counts. That is, C_1 can be any subset of $\{0, \ldots, m\}$, and for $2 \leq i \leq k$, $C_i = C_i(N_1, ..., N_{i-1}) \subseteq \{0, ..., m\}$. In order for the selection rule to always terminate, $C_k = \{0, ..., m\}$.

As a concrete example, the right-boundary procedure sets

$$C_i = \left\{ j \in \mathbb{Z} : 0 \leq j \leq (m - \sum_{j<i} N_j + 1) \frac{\lambda_i - \lambda_{i-1}}{1 - \lambda_{i-1}} \right\}$$

for $i = 1, \ldots, k-1$. Such a rule resembles a search from left to right which evaluates a stopping condition at each candidate value and stops the first time the condition is satisfied.

More precisely, we define the concept of an LRS selection rule as follows:

Definition 1 (LRS selection rule)**.** *A random variable $\lambda(\Lambda)$ is an LRS selection rule if for all finite grids $\Lambda \subset (0, 1)$,*

(i) $\lambda(\Lambda)$ takes values in Λ;
(ii) $\lambda(\Lambda)$ is a stopping time with respect to $\{\mathcal{A}_t\}_{t\in(0,1]}$, where

$$\mathcal{A}_t = \sigma(N_j : \lambda_j \leq t).$$

This definition puts into plain view the concept of an LRS selection rule as a restricted type of stopping time rule. While the general stopping rules of Liang and Nettleton (2012) must be stopping times with respect to $\{\mathcal{F}_t\}_{t\in(0,1]}$, Definition 1 states that LRS selection rules have the stronger requirement of being a stopping time with respect to the smaller filtration $\{\mathcal{A}_t\}_{t\in(0,1]}$ for every finite grid Λ.

2.3.1 Finite sample control of the FDR

Consider a fixed λ candidate grid $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_k\} \subseteq [\kappa, \tau]$ such that

$$0 < \kappa = \lambda_1 < \lambda_2 < \cdots < \lambda_{k-1} < \lambda_k = \tau < 1.$$

When λ is selected dynamically from Λ, define the following FDR estimator

$$\widehat{\mathrm{FDR}}^*_\lambda(t) = \begin{cases} \frac{m\hat{\pi}^*_0(\lambda)t}{R(t)\vee 1} & t \le \kappa, \\ 1 & t > \kappa. \end{cases}$$

Note that in the fixed adaptive case, the possible rejection thresholds are truncated at λ, while in the dynamic adaptive case they must be truncated at κ, the smallest candidate value. The definition of $\widehat{\mathrm{FDR}}^*_\lambda$ appears at first to be restrictive as it limits the possible range of rejection thresholds to $[0, \kappa]$. In practice, we can set κ not too small, say $\kappa = \alpha$. Following a justification of Storey et al. (2004), Remark 1, note that for a valid rejection threshold t^*,

$$\mathrm{mFDR}(t^*) \approx \mathrm{FDR}(t^*) \le \alpha$$

implies that

$$t^* \le \frac{(1-\pi_0)\alpha}{\pi_0(1-\alpha)}$$

since $E[R(t^*)] \le m_0 t^* + m_1$. Even for a relatively small $\pi_0 = 0.75$ and typical $\alpha \approx 0.1$, this gives $t^* \le \alpha$, and the truncation should rarely affect the outcome of the thresholding procedure. Recall the α-level thresholding functional defined as

$$t_\alpha(F) = \sup\{0 \le t \le 1 : F(t) \le \alpha\}.$$

Then an LRS selection rule applied to any fixed finite grid Λ leads to control of the FDR.

Theorem 1. *Under the null independence model, suppose λ is chosen using an LRS selection rule with fixed candidate grid $\Lambda = \{\lambda_1, ..., \lambda_k\}$. Then* $\mathrm{FDR}\{t_\alpha(\widehat{\mathrm{FDR}}^*_\lambda)\} \le \alpha$.

Theorem 1 immediately implies that the right-boundary procedure controls the FDR in finite samples. The details of the proof are left to the appendix, Section A.1, but a proof sketch is given here.

Proof sketch of Theorem 1

By applying martingale arguments similar to Storey et al. (2004), and by the definition of $\widehat{\mathrm{FDR}}^*_\lambda$, we first bound the FDR from above by

$$\mathrm{FDR}(t_\alpha(\widehat{\mathrm{FDR}}^*_\lambda)) \le \alpha E\left[\frac{1-\lambda}{m_0 - V(\lambda) + 1}\frac{V(\lambda_1)}{\lambda_1}\right].$$

It then suffices to show that the expectation on the right-hand side is bounded above by 1. By conditioning on the locations of the false null p-values, and invoking the LRS property of the λ selection rule, this summation over the candidate λ values can be re-indexed as a summation over the possible values of the true null bin counts, $(V_1, ..., V_k)$, which are defined analogously to $(N_1, ..., N_k)$ above, but only counting the true null p-values.

By the independence of the true null p-values in the null independence model, the binning process can be seen as a discretization of the uniform density, and hence, the vector of true null bin counts $(V_1, ..., V_k)$ will follow a known multinomial distribution with m_0 trials and probability vector $(\lambda_1, \lambda_2 - \lambda_1 ..., \lambda_k - \lambda_{k-1})$. Based on the distribution result of bin counts, we show that each term in the summation is equal to the probability of some combinations of bin counts, $(V_1, ..., V_k)$. Then the proof follows by showing those combinations are non-overlapping due to the sequential nature of the LRS selection rules.

2.4 Extension to other dynamic adaptive procedures

Although LRS selection rules encompass a large class of dynamic adaptive procedures, there are still some procedures prevalent in the literature which are not covered. In particular the lowest-slope procedure of Benjamini and Hochberg (2000) and the k-quantile procedure described in Benjamini et al. (2006) which selects $\lambda = p_{(k)}$ for some prespecified $1 \leq k \leq m$, are commonly used stopping time λ selection rules that are not strictly LRS. However, they very closely resemble LRS selection rules, as the grid of candidate values for λ remains finite in both cases.

The lowest-slope procedure can be interpreted as the right-boundary procedure applied to a particular data-driven grid. Under our notation, $i = R(p_{(i)})$, and the lowest-slope procedure operates by calculating for each $1 \leq i \leq m$ the slope

$$\sigma_i = \frac{1 - p_{(i)}}{m - R(p_{(i)}) + 1}$$

of the line from $(R(p_{(i)}), p_{(i)})$ to $(m+1, 1)$. The lowest-slope procedure stops at $p_{(j)}$, where

$$j = \min\{1 \leq i \leq k : \sigma_i < \sigma_{i-1}\},$$

the smallest index for which the slope decreases. The estimator of π_0 is then

$$\hat{\pi}_0^{\mathrm{LSL}} = \frac{1}{m\sigma_j} = \frac{m - R(p_{(j)}) + 1}{m(1 - p_{(j)})} = \hat{\pi}_0^*(p_{(j)}).$$

Note that for $1 \leq i \leq m$,

$$\frac{1}{m\sigma_i} = \hat{\pi}_0^*(p_{(i)}),$$

so that the stopping condition of the lowest-slope procedure satisfies

$$\min\{1 \leq i \leq k : \sigma_i < \sigma_{i-1}\} = \min\{1 \leq i \leq k : \hat{\pi}_0^*(p_{(i)}) > \hat{\pi}_0^*(p_{(i-1)})\}.$$

Then it is easy to see that the procedure is nearly equivalent to the right-boundary procedure with bins bounded by $\{p_{(1)}, p_{(2)}, ..., p_{(m)}\}$, where the only difference is the strictness of the inequality in the stopping condition. When the bins are constructed this way, there is exactly one p-value in each bin, so the bin counts become fixed while the bin boundaries are random. This is inverse to the LRS selection rules as defined above, for which the bin boundaries are fixed, and the bin counts random.

2.4.1 p-grid λ selection rules

In this section we consider a new class of dynamic adaptive procedures, under which the tuning parameter λ is selected from a data-driven finite grid. As in the previous section, κ and τ are fixed constants introduced to bound the tuning parameter away from 0 and 1. the so-called *p-grid LRS selection rules* are defined as follows.

Definition 2 (p-grid)**.** *For $0 < \kappa < \tau < 1$, the p-grid is defined as*

$$\Lambda^{(p)} := (\{p_{(1)}, ..., p_{(m)}\} \cap (\kappa, \tau)) \cup \{\tau\}.$$

The p-value grid (p-grid) is constructed by bounding the grid candidates between κ and τ. τ is added as a uniform last candidate in case the stopping condition is never satisfied previously.

Definition 3 (p-grid LRS selection rule)**.** *Fix $0 < \kappa < \tau < 1$ and let $\lambda^{LRS}(\Lambda)$ be an LRS selection rule. The p-grid LRS selection rule with underlying selection rule λ^{LRS} is defined as*

$$\lambda^*(p_1, ..., p_m) = \lambda^{LRS}(\Lambda^{(p)}) \in (\kappa, \tau].$$

A p-grid LRS selection rule is defined based on its underlying LRS selection rule, which is applied to the p-grid. The discussion in Section 2.4 shows that the lowest-slope procedure of Benjamini and Hochberg (2000), after slight modifications, can be characterized as the p-grid LRS selection rule when the underlying LRS selection rule is the right-boundary

procedure. The k-quantile procedure can also be characterized as a p-grid LRS selection rule, but with underlying LRS rule

$$\lambda^{LRS-k}(\Lambda) = \min\{\lambda \in \Lambda : R(\lambda) \geq k\}.$$

The proof of finite sample control of the FDR will rely on an application of Theorem 1, by approximating a p-grid LRS selection rule with a sequence of LRS rules on fixed and finite grids. In order for this approximation argument to hold, a further regularity assumption is required on the p-value model so that the p-grid is almost surely made up of distinct values. In particular, assume the *continuous null independence model*, which is the null independence model, plus the additional assumption that

$$p_i \,|\, (p_1, ..., p_{i-1}, p_{i+1}, ..., p_m)$$

is continuously distributed for all $1 \leq i \leq m$.

Furthermore, in order to construct the approximating sequence of LRS procedures, the following two mild regularity conditions are required on the p-grid LRS selection rules themselves.

Condition 1. *The p-grid LRS selection rule λ^* is a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t\in(\kappa,\tau]}$.*

Note that all fixed grid LRS rules are stopping time rules (see discussion following Definition 1). Since the aim is to approximate a p-grid LRS selection rule λ^* by a sequence of fixed grid LRS rules, it is natural to require that λ^* is also a stopping time. For the lowest-slope procedure, this condition is trivially satisfied, since the stopping decision for $p_{(i)}$ is made without knowledge of the locations of $p_{(i+1)}, ..., p_{(m)}$. The same is true for the k-quantile procedure. For other procedures, Condition 1 can be easily checked.

The next condition is a continuity condition such that the λ selected from a p-grid can be approximated arbitrarily closely by using a constructed grid that is close enough to the p-grid from the right-hand side.

Condition 2. *The underlying LRS selection rule λ^{LRS} is right continuous. That is, with probability 1, for every $\epsilon > 0$, there exists $\delta > 0$ such that*

$$|\lambda^{LRS}(\Lambda^{(p)}) - \lambda^{LRS}(\Lambda')| < \epsilon$$

for every grid Λ' of length $k^{(p)} := |\Lambda^{(p)}|$ with $\lambda'_j \in [\lambda_j^{(p)}, \lambda_j^{(p)} + \delta)$ for every $1 \leq j \leq k^{(p)}$, where λ'_j denotes the jth ordered element of Λ' and $\lambda_j^{(p)}$ the jth ordered element of the p-grid.

In Definitions 8 and 9 in the appendix, we will construct the grid Λ' using the right boundaries of non-empty bins such that it has the same length as the p-grid and can be made arbitrarily close to the p-grid. Condition 2 ensures that for almost all realizations of the null independence model, this convergence of grids is sufficient to conclude convergence of the selected tuning parameter. For any p-grid LRS rule to have finite sample FDR control, this right continuity condition is the primary condition which needs to be established.

2.4.2 Finite sample control of the FDR

Theorem 2. *Fix $0 < \kappa < \tau < 1$. Under the continuous null independence model, if λ is selected using a p-grid LRS selection rule which satisfies Conditions 1 and 2, then $\mathrm{FDR}(t_\alpha(\widehat{\mathrm{FDR}}^*_\lambda)) \leq \alpha$.*

The proof is technical and is left to the appendix, Section A.2, however a proof sketch is given at the end of this section.

Theorem 2 applies to a wide class of dynamic adaptive procedures with p-grid LRS selection rules, but its primary usefulness is its ability to show finite sample control for the commonly used lowest-slope and k-quantile procedures. Strictly speaking, the enforcement of the required regularity conditions means that some slight modifications are made to the lowest-slope procedure as it was originally defined by Benjamini and Hochberg (2000). This modified version bounds the tuning parameter away from 0 and 1, and only considers p-values in the open interval (κ, τ). The procedure simply selects τ if there are fewer than 2 p-values in this interval, since as it considers candidates from left to right, it never finds two points at which it can compare slopes, and thus never stops. It also selects τ if the slope comparison step never terminates. Benjamini and Hochberg (2000) are unclear on what to do in these two boundary situations, as when there are sufficiently many p-values, they are highly unlikely to occur in practice. As discussed in the beginning of this section, the original lowest-slope stopping condition uses a strict inequality (Benjamini and Hochberg, 2000), while the right-boundary procedure does not. Under the continuous null independence model, this modification will almost surely not affect the outcome of the procedure (see Lemma 4 in the appendix). Similar modifications are made to the k-quantile procedure of Benjamini et al. (2006), as κ and τ are introduced as upper and lower bounds respectively on the selected tuning parameter. More specifically, we define the modified procedures as follows:

Definition 4 (modified lowest-slope procedure)**.** *The modified lowest-slope procedure (LSL) is the p-grid LRS selection rule where the underlying LRS selection rule is the right-boundary procedure.*

Definition 5 (modified k-quantile procedure)**.** *The modified k-quantile procedure (*Q-k*) is the p-grid LRS selection rule where the underlying LRS selection rule is λ^{LRS-k}.*

Intuitively, bounding λ away from 0 and 1 can avoid high bias and high variance of the π_0-estimator, respectively. In practice, to control the FDR at level α, it is sensible to reject only p-values smaller than α, and it is reasonable to set $\kappa = \alpha$. Then τ can be set close to 1, say, $\tau = 0.95$. Overall, especially when the total number of tests m is sufficiently large, the modification is minor and keeps the most sensible region of the λ parameter.

The finite sample control of both of these procedures follows easily from Theorem 2, by showing that their underlying LRS selection rules are right continuous.

Corollary 1. *Under the continuous null independence model, suppose λ^{LSL} is selected using the modified lowest-slope procedure, then* $\text{FDR}(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^{\text{LSL}}})) \leq \alpha$.

Corollary 2. *Under the continuous null independence model, suppose $\lambda^{\text{Q-k}}$ is selected using the modified k-quantile procedure, then* $\text{FDR}(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^{\text{Q-k}}})) \leq \alpha$.

Corollary 1 is the first time the finite sample FDR control has been proven for the lowest-slope procedure of Benjamini and Hochberg (2000), and Corollary 2 is analogous to Theorem 2 in Benjamini et al. (2006). A lemma establishing the right continuity of the right-boundary procedure is given in the appendix. The underlying selection rule λ^{LRS-k} of the modified k-quantile procedure is defined above and can easily be shown to be right continuous, so its proof is omitted.

The generality of the class of p-grid LRS selection rules covered by Theorem 2 also makes it easy to show finite sample FDR control for other modifications of the above procedures. For instance, we could improve the lowest slope procedure by modifying it to check the stopping conditions less often, comparing slopes only at every Bth p-value, for some $B \geq 1$, or at some other fixed quantiles of the p-value empirical distribution. Since the lowest-slope procedure tests its stopping condition at every realized p-value, it tends to stop too early and return a smaller λ than the optimal λ, a feature which can be easily remedied by considering fewer stopping points. It is easy to show that such a procedure continues to satisfy the conditions of Theorem 2, and thus controls the FDR in finite samples.

Proof sketch of Theorem 2

We will bound the FDR of a p-grid LRS selection rule below α by showing it is the limit of a sequence of FDR's of so-called finite approximation LRS selection rules with progressively finer candidate grids. Theorem 1 ensures that each term of the sequence is bounded below α, which allows us to conclude that the limit is bounded below α, as desired.

More specifically, we aim to show pointwise convergence of a sequence of fixed grid LRS procedures to the p-grid LRS procedure. We first identify the set of p-value realizations for which this convergence is either difficult or impossible to prove, and demonstrate that this set of realizations comprises a null set under the continuous null independence model. Then, for any fixed realization of p-values outside this null set, we show that the tuning parameter λ, the π_0-estimator $\hat{\pi}_0^*(\lambda)$, and the proportion of false discoveries from the dynamic adaptive procedures defined using the finite approximation LRS selection rules all converge pointwise to the corresponding quantities using the p-grid LRS selection rule. The FDR for a given procedure is the expectation of the false discovery proportion, which is a bounded random variable. Hence, we invoke the bounded convergence theorem to show the convergence in expectation.

2.5 Simulation

Simulations are carried out to evaluate the FDR control, power and m_0 estimation properties of the dynamic adaptive procedures in the literature. The candidate procedures are

- `BH`, the original step-up procedure of Benjamini and Hochberg (1995);
- `ORC`, the oracle procedure by applying the BH procedure at level α/π_0, assuming known π_0;
- `RB20`, the right-boundary procedure with $\Lambda = \{0.05, 0.1, ..., 0.95\}$;
- `LSL`, the modified lowest-slope procedure (Definition 4);
- `RB20q`, the lowest-slope procedure that considers only 20 p-value stopping points at evenly spaced quantiles;
- `HJW`, the weight shifting dynamic adaptive procedure of Heesen and Janssen (2016).

The simulation settings are similar to those in Liang and Nettleton (2012). When true null p-values are independent, all the procedures considered are established to control the FDR in finite samples at level α. `BH` controls FDR conservatively at level $\pi_0\alpha$. The finite sample control of `RB20` is a consequence of Theorem 1, and the finite sample control of `LSL` and `RB20q` follows from Theorem 2. `HJW` is a particular example from a class of dynamic adaptive procedures shown to have finite sample control (Heesen and Janssen, 2016).

Simulations are based on $J = 10000$ replications, and the nominal FDR level is $\alpha = 0.05$. For each replication, $m = 10000$ one-sided tests of $H_0 : \mu = 0$ are performed, with standard normal true null statistics, and false null statistics having $N(\mu, 1)$ distribution. Effect sizes μ are set to $0.5, 1, 2$ and 4. For effect sizes larger than 4, the false null p-values are well separated from the true null p-values, and all procedures achieve full power relative to ORC.

2.5.1 Independent tests

Simulation for independent test statistics are reported in Figure 2.1. The first row plots average realized FDR, the second the power relative to `ORC`, and the third the log mean-squared error (MSE) of $\hat{m}_0 = \hat{\pi}_0 m$, defined as

$$\text{MSE} = \frac{1}{J}\sum_{j=1}^{J}(\hat{m}_0 - m_0)^2.$$

All procedures control the FDR below the nominal level 0.05, and see an increase in the FDR and relative power as the signal strength μ increases. `RB20` and `RB20q` provide the greatest relative power in all settings, and this is because they have the smallest MSE of $\hat{m}_0$. When the signal strength is larger, and the optimal λ may be smaller than $\lambda_1 = 0.05$, the minimal possible value from `RB20`, in which case the quantile-based bins of `RB20q` can provide a marginal improvement over `RB20` by considering smaller stopping points, similar to the RB20* procedure in Liang and Nettleton (2012). `HJW`, although similar in spirit to `RB20`, cannot achieve the same power performance since it restricts its estimation region to $[0.5, 1]$, and the right-to-left measurability condition placed on the weights forces it to sometimes over-weight the influence of smaller p-values in the estimation of π_0. Since it is known that `ORC` controls the FDR at exactly level α, all average realized FDR levels are corrected by the difference between in the FDR of `ORC` and the target FDR level α.

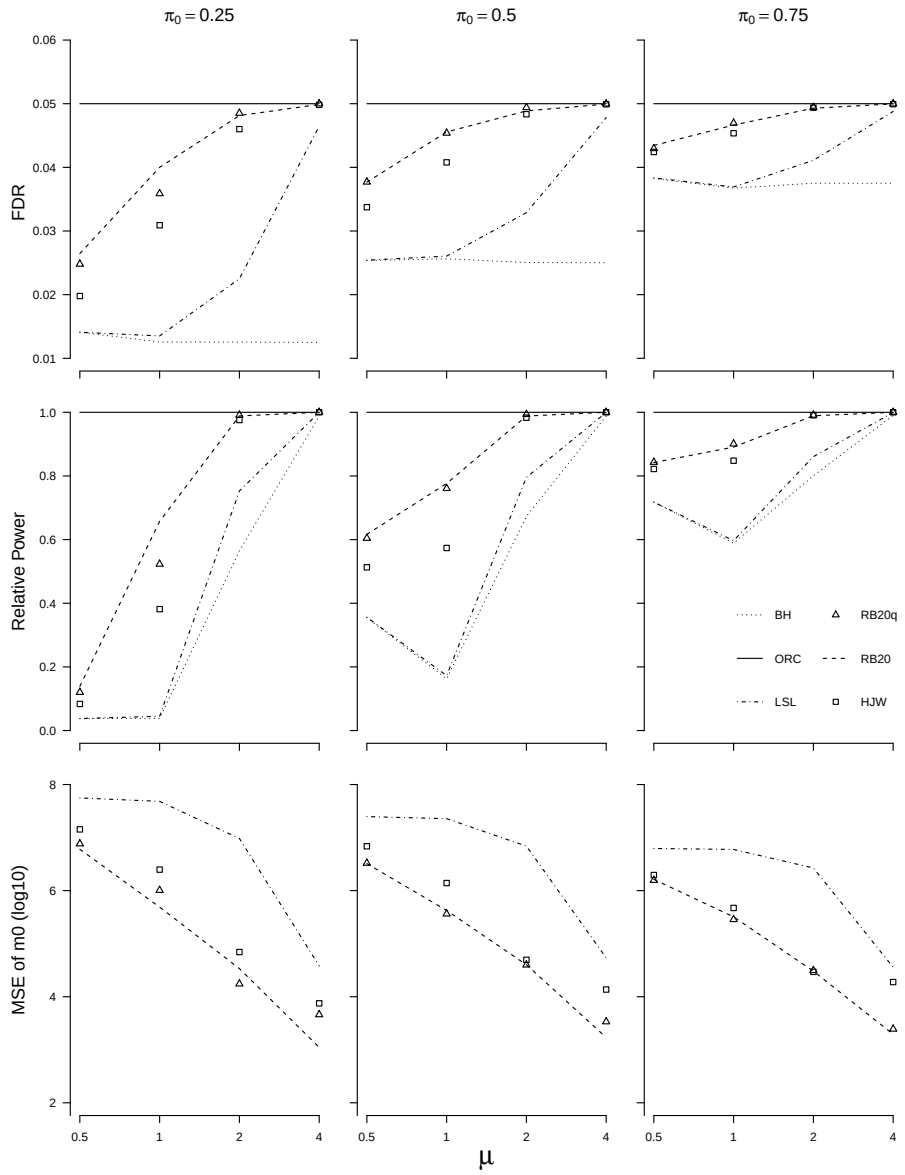

Figure 2.1: Simulation results for independent test statistics.

2.5.2 Dependent tests

Simulation was also performed with dependent test statistics. In particular, statistics have block auto-regressive order 1 correlation structure with block size 50 and correlation $\rho^{|i-j|}$

between the ith and jth elements in any block, and correlation coefficient $\rho = -0.9$. Block structure such as this has been used by Liang and Nettleton (2012), among others, to recreate the varying positive and negative correlations expected among genes in the same biological pathway. Results are reported in Figure 2.2. As above, all procedures control FDR below the nominal level 0.05, and increase in FDR and relative power as the signal strength increases. `RB20` and `RB20q` remain the best in terms of power. There is some evidence that all procedures, including `ORC`, become conservative in the small signal case, due to the dependence among the test statistics.

2.6 Discussion

2.6.1 Identifiability and purity

All of the results proven in this paper give only conservative control and estimation, rather than exact control or estimation. Heesen and Janssen (2015), among others, have shown that the original BH procedure has FDR exactly equal to $\pi_0\alpha$, but in the adaptive case in which we incorporate an estimate of π_0, identifiability issues manifest themselves, as discussed in Section 3.1 of Genovese and Wasserman (2004).

Under the Bayesian two-group model (Efron et al., 2001) where all p-values are independent, one can follow similar steps to the proof of Theorem 1 to show the bound

$$\mathrm{FDR}(t_\alpha(\widehat{\mathrm{FDR}}^*_\lambda)) \leq \alpha \cdot \sup_{\lambda \in \Lambda} P(H_i = 0 \mid p_i > \lambda).$$

There may in fact be no λ for which $P(H_i = 0 \mid p_i > \lambda) = 1$, a result of F_1, the distribution of the false null p-values, having a non-zero uniform component. This is termed *impurity* by Genovese and Wasserman (2004). Such purity issues are the reason that we cannot, without further assumptions on F_1, find an unbiased Storey-type estimator for π_0, and can only conclude conservatism. The effects of such bias carry through to the estimation of FDR, and the thresholding procedure, and can intrinsically bound the FDR of the procedure below the target level α.

2.6.2 Dependence

The results of this chapter are proven under the classical null independence model, but prior FDR control literature has considered estimation and control properties under dependence

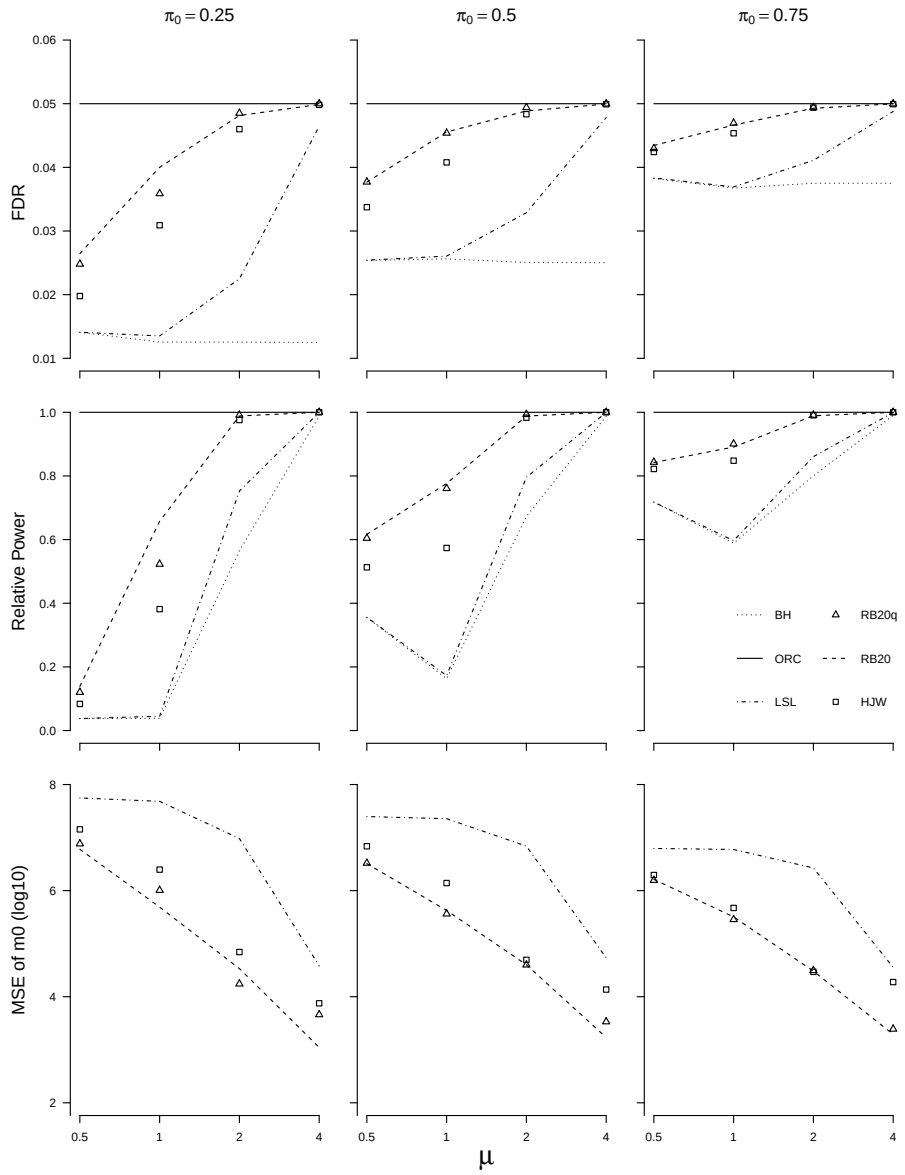

Figure 2.2: Simulation results for correlated test statistics, $\rho = -0.9$.

assumptions on the true null p-values, in particular the *positive regression dependence on a subset* (PRDS) condition in Benjamini and Yekutieli (2001) and the *reverse martingale dependence* (RMD) condition in Heesen and Janssen (2015). Proposition 6.2 of Heesen

and Janssen (2015) implies that finite sample control of the FDR will not hold under every PRDS or RMD model, even for the fixed adaptive procedure of Storey et al. (2004).

Chapter 3

Controlling the false discovery rate of grouped hypotheses

In this chapter we consider the case where hypotheses have a known group structure. In the grouped setting it is inefficient to treat hypotheses as exchangeable and Cai and Sun (2009) show that power can be improved through a ranking of significance that incorporates group information. First we define a general sequential framework for multiple testing procedures in the grouped setting. We develop the flexible grouped mirrored knockoff (GMK) procedure which approximates the optimal ranking of significance. Similar to Lei and Fithian (2018) and Barber and Candès (2015), when the available information at each step of the procedure is masked to avoid overfitting, the GMK procedure controls the FDR in finite samples.

This chapter is organized as follows. In Section 3.1 we introduce notation and place assumptions on the p-value model. Section 3.2 describes a highly general framework for grouped multiple testing procedures, through which we define the oracle GMK procedure in Section 3.3. In Section 3.4 we define the more general GMK procedure and state and prove the main theoretical result of the chapter. Section 3.5 gives details of implementation for the data-driven GMK procedure using the expectation-maximization (EM) algorithm. Section 3.6 gives simulation results which compare the GMK procedure to competing grouped procedures in the literature. Finally, Section 3.7 applies the new GMK procedure to real data from the adequate yearly progress (AYP) study of California high schools.

3.1 Notation

Recall from Chapter 1 the grouped multiple testing problem, with $m = m_1 + \cdots + m_K$ hypotheses $H_{k,i}$ from K fixed and known groups, where $k = 1, ..., K$ and $i = 1, ..., m_k$. Suppose $m_{0,k} \leq m_k$ hypotheses from each group are true nulls and denote $m_0 = m_{0,1} + \cdots + m_{0,K}$. For $k = 1, ..., K$, we call $\pi_{0,k} = m_{0,k}/m_k$ the *true null proportion* for the kth group. The overall true null proportion can be written as the weighted average

$$\pi_0 = \sum_{k=1}^{K} \left(\frac{m_k}{m} \right) \pi_{0,k} = \frac{m_0}{m}.$$

Denote the p-values by $p_{k,i}$, $k = 1, ..., K$ and $i = 1, ..., m_k$, and suppose they follow the null independence model. For ease of notation we assume throughout this chapter that all p-values are strictly between 0 and 1 and there are no ties.

Throughout the chapter we also make reference to the Bayesian two-group model. Recall the two-group model with K groups (Cai and Sun, 2009),

$$\begin{cases} H_{k,i} \sim \text{Bernoulli}(1 - \pi_{0,k}), \\ p_{k,i} | H_{k,i} \sim H_{k,i} f_{1,k} + (1 - H_{k,i}) f_0 \end{cases}$$

for $k = 1, ..., K$ and $i = 1, ..., m_k$, where the $H_{k,i}$ are independent, and the $p_{k,i}$ are conditionally independent given the $H_{k,i}$. Furthermore, recall that under the two-group model, the group-wise Lfdr can be written as

$$\text{Lfdr}_k(p) = \frac{\pi_{0,k}}{\pi_{0,k} + (1 - \pi_{0,k}) f_{1,k}(p)}.$$

3.2 A general framework for grouped multiple testing procedures

The GMK procedure, and the proof of its finite sample control of the FDR, relies on its characterization as a sequential procedure, with the potential set of rejections evolving through the steps of the procedure until it terminates. In this section we will formally define a sequential multiple testing procedure general enough that it encompasses the GMK procedure as well as its competitors in the literature (Cai and Sun, 2009; Hu et al., 2010; Zhao and Zhang, 2014). This sequential framework is a highly interpretable way

of analyzing the operational characteristics of different FDR controlling multiple testing procedures in the grouped setting.

First note that under the null independence model, the hypotheses within each group are exchangeable. Thus the rejection region for group k should be an interval of the form $[0, s_k)$ for some $s_k \in [0, 1]$ (Storey, 2002). While closed interval rejection regions have been used in the past, the GMK procedure operates by iteratively removing p-values from the rejection set, so we use a half-open interval to neaten the definitions to follow. Combining the group-wise rejection regions, the overall rejection set S can be described by a K-dimensional vector $s = (s_1, ..., s_K) \in [0, 1]^K$ as

$$S = \{H_{k,i} : p_{k,i} < s_k, k = 1, ..., K\}.$$

We will refer to s as a *rejection threshold vector*. Notice that these vectors have a natural element-wise ordering: if $s'_k \leq s_k$ for every $k = 1, ..., K$, then s will reject at least as many hypotheses as s'. This ordering motivates a sequential characterization of a typical multiple testing procedure.

Let t denote the step number. At a given step the vector $s^{(t)}$ of rejection thresholds defines a set of p-values to reject,

$$S^{(t)} = \{H_{k,i} : p_{k,i} < s_k^{(t)}, k = 1, ..., K\}.$$

The procedure then needs to decide whether to terminate and return the current rejection set, or proceed to step $t + 1$. If it proceeds to step $t + 1$ the procedure will decide how to update $s^{(t)}$ to $s^{(t+1)}$, where $s^{(t+1)} \leq s^{(t)}$ with respect to the element-wise ordering. Since the vector of rejection thresholds has finite dimension K, there is no loss of generality in assuming that each step of the procedure will strictly lower exactly one element of $s^{(t)}$.

Any multiple testing procedure with this structure is defined by an initialization point, and two decision rules. The *stopping rule* decides whether the procedure will terminate for a given rejection set and thus governs its FDR control. The *threshold updating rule* decides which group's threshold to lower at each update, and by how much, which governs the procedure's power. The following gives more formal notation for the general stages of any such procedure.

Sequential procedure for grouped hypotheses

1. Initialize a the rejection threshold vector $s^{(0)} = (s_1^{(0)}, ..., s_K^{(0)})$, and initialize $t = 0$.

2. Apply the stopping rule $\delta_1^{(t)}$ to the p-values and rejection thresholds $s^{(t)}$. If $\delta_1^{(t)}$ returns 1, the procedure terminates and returns the current rejection set $S^{(t)} = \{H_{k,i} : p_{k,i} < s_k^{(t)}, k = 1, ..., K\}$. If $\delta_1^{(t)}$ returns 0 then continue to stage 3. $\delta_1^{(t)}$ should be defined to always return 1 if all the current rejection set is empty.

3. Apply the threshold updating rule $\delta_2^{(t)}$ to $s^{(t)}$. $\delta_2^{(t)}$ lowers exactly one element of $s^{(t)}$ and leaves the remaining elements unchanged. Let $s^{(t+1)} = \delta_2^{(t)}(s^{(t)})$. Update $t \leftarrow t + 1$, and return to stage 2.

In the examples to follow, the stopping rule $\delta_1^{(t)}$ will estimate the overall FDR of the current rejection set and return 1 if and only if this estimate is no larger than the nominal level α. The threshold updating rule $\delta_2^{(t)}$ will lower the rejection threshold vector based on some pooled ranking of the significance of the hypotheses, removing the least significant hypothesis from the current rejection set. As long as the rejection set is empty for some finite step number, the procedure will always terminate. Although this is not the framework under which most of the procedures in the literature are defined, many can be written in this way. We present as an example the group-weighted Benjamini-Hochberg method of Hu et al. (2010). Our GMK procedure will serve as a second example.

Example 1 (Sequential representation of the GBH procedure)**.** *In stage 1, the rejection threshold vector is initialized as* $s^{(0)} = (1, 1, ..., 1)$. *Suppose the procedure is at step* t. *In stage 2, the stopping rule* δ_1 *checks whether an estimate of the FDR is below the nominal level* α. *In this case the estimate is given by*

$$\widehat{\mathrm{FDR}}_{GBH}(s^{(t)}) = \frac{m \cdot s_W^{(t)}}{|S^{(t)}| \vee 1},$$

where $s_W^{(t)}$ *is the overall rejection threshold on the weighted scale. This overall threshold can be recovered from the group-wise thresholds via the relationship*

$$s_W^{(t)} = \sum_{k=1}^{K} \left\{ \left(\frac{m_k}{m} \right) \pi_{0,k} \cdot \max_i \{ p_{k,i} : p_{k,i} < s_k^{(t)} \} \right\},$$

with the convention $\max \emptyset = 0$. *The true null proportions are taken as known in the oracle procedure, or replaced by their estimates in the data-driven case. The stopping rule is given by*

$$\delta_1(s^{(t)}) = I(\widehat{\mathrm{FDR}}_{GBH}(s^{(t)}) \leq \alpha).$$

In stage 3, define

$$(k^*, i^*) = \underset{(k,i)}{\text{argmax}}\{p_{k,i}/w_k : p_{k,i} < s_k^{(t)}\},$$

where for $k = 1, ..., K$,

$$w_k = \frac{\pi_{0,k}}{(1 - \pi_{0,k})(1 - \pi_0)}$$

are the group-wise weights in the oracle GBH procedure, with the true null proportions replaced by estimates in the data-driven procedure. Then $\delta_2^{(t)}$ *lowers the rejection threshold in group* k^* *to level* p_{k^*,i^*}*. That is,*

$$\delta_2(s^{(t)}) = (s_1^{(t)}, s_2^{(t)}, ..., p_{k^*,i^*}, ..., s_k^{(t)}).$$

Since the sequential procedure is written to reject all hypotheses with p-values strictly below $s^{(t)}$, this threshold updating rule will remove the largest group-weighted p-value from the current rejection set. The manner in which the rejection threshold vector is updated implies that at each step of the procedure, exactly one p-value is removed from the rejection set, so the procedure will terminate in a finite number of steps.

In this example, the threshold updating rule reflects a ranking of significance of the hypotheses, and at each step the least significant hypothesis with respect to this ranking is removed from the rejection set. In particular the GBH procedure ranks hypotheses according to their weighted p-values, where the weights incorporate the group-wise true null proportions. The stopping rule is based on an estimate of the FDR of the current rejection set, where the estimator is chosen to ensure either finite sample or asymptotic control of the FDR in the oracle or data-driven cases, respectively. These ideas motivate the GMK procedure.

3.3 Oracle grouped mirrored knockoff procedure

We now describe the oracle GMK procedure under the notation of the general sequential framework. Cai and Sun (2009) showed that under the two-group model, and assuming that $f_{1,k}$ is non-increasing for $k = 1, ..., K$, the group-wise Lfdr gives the optimal ranking of significance for the grouped multiple testing problem. Motivated by this result, In the oracle GMK procedure, the threshold updating rule $\delta_2^{(t)}$ is based on the group-wise Lfdr functions.

The stopping decision rule will be based on an estimate of the overall FDR, chosen so that the GMK procedure will maintain finite sample control of the FDR. A natural estimate

of the FDR in the single group case is defined in Storey et al. (2004). For $\lambda \in (0,1)$ and rejection cutoff s_0, this estimator is defined by

$$\widehat{\text{FDR}}_\lambda(s_0) = \frac{s_0}{1-\lambda} \cdot \frac{(m - \sum_{i=1}^m I(p_i \leq \lambda) + 1)}{\max\{\sum_{i=1}^m I(p_i \leq s_0), 1\}}.$$

Storey et al. (2004) show that in the single group case, the thresholding procedure with estimator $\widehat{\text{FDR}}_\lambda$ controls the FDR in finite samples. Suppose we were to specify $\lambda = 1 - s_0$, the *mirror* of the rejection threshold. Then s_0 in the numerator cancels with $1 - \lambda$ in the denominator, and the estimate is defined entirely by counting the ratio between the number of p-values above the mirror of the rejection threshold, and the total number of p-values below the rejection threshold. This motivates an extension to the grouped case, that is,

$$\widehat{\text{FDR}}_{\text{GMK}}(s) = \frac{1 + \sum_{k=1}^K \sum_{i=1}^{m_k} I(p_{k,i} > 1 - s_k)}{\max\{\sum_{k=1}^K \sum_{i=1}^{m_k} I(p_{k,i} < s_k), 1\}},$$

which analogously estimates the FDR as a ratio between the total number of p-values above the mirrors of the group-wise rejection thresholds and the number of p-values below the group-wise rejection thresholds. In this way, the p-values in the mirror image of the rejection region are used as knockoffs to estimate the number of false rejections. It will be made clear in the proof of Theorem 3 how this mirroring allows the application of martingale theory to prove finite sample control of the GMK procedure.

Based on the above development, the oracle GMK procedure can be defined under the framework of the general sequential procedure given in Section 3.2. In this case, $s^{(0)} = (1/2, 1/2, ..., 1/2)$. Suppose the procedure is at step t. Define

$$\delta_1(s^{(t)}) = I(\widehat{\text{FDR}}_{\text{GMK}}(s^{(t)}) \leq \alpha),$$

similar to the stopping rules of the GBH procedure. Furthermore define

$$(k^*, i^*) = \underset{(k,i)}{\text{argmax}}\{\text{Lfdr}_k(q_{k,i}) : q_{k,i} < s_k^{(t)}\},$$

where $q_{k,i} = \min\{p_{k,i}, 1 - p_{k,i}\}$ is either the p-value or its mirror image. Then the threshold updating rule is defined similarly to GBH procedure as

$$\delta_2(s^{(t)}) = (s_1^{(t)}, s_2^{(t)}, ..., q_{k^*,i^*}, ..., s_k^{(t)}).$$

In effect, the oracle GMK procedure updates rejection thresholds based on a pooled ranking with respect to the group-wise Lfdr. Since the estimator $\widehat{\text{FDR}}_{\text{GMK}}$ incorporates counts

below the rejection threshold vector and above its mirror image, thresholds have to be lowered more slowly, based on the values $q_{k,i}$ rather than $p_{k,i}$. It can be shown that the oracle GMK procedure controls the overall FDR at level α in finite samples. However, we omit this proof as it follows as a corollary of Theorem 3, the finite sample control of our more general GMK procedure. If we make the stronger assumption that the p-values follow the two-group model and the densities of the false null p-values are non-increasing, then the ranking of significance according to the Lfdr is optimal in terms of expected power (Cai and Sun, 2009; Lei and Fithian, 2018).

3.4 Grouped mirrored knockoff procedure

While in the oracle case, there is a clear optimal threshold updating rule, in practice the Lfdr functions for each group will not be known. We may still utilize the same stopping rule as in the oracle case, but the threshold updating rule will have to be estimated from the data. For this reason, we define the GMK procedure more generally, with no restriction on the form of $\delta_2^{(t)}$. However, in order to preserve the FDR control properties of the procedure, we shall restrict the information available to estimate $\delta_2^{(t)}$ at each step. In particular, at step t, with rejection threshold vector $s^{(t)}$, let

$$R_{t,k} = |\{i : p_{k,i} < s_k^{(t)}\}| \text{ and } A_{t,k} = |\{i : p_{k,i} > 1 - s_k^{(t)}\}|,$$

and let $R_t = \sum_{k=1}^K R_{t,k}$ and $A_t = \sum_{k=1}^K A_{t,k}$. Define

$$\tilde{p}_{t,k,i} = \begin{cases} p_{k,i}, & \text{if } s_k^{(t)} \leq p_{k,i} \leq 1 - s_k^{(t)}, \\ \{p_{k,i}, 1 - p_{k,i}\}, & \text{otherwise.} \end{cases}$$

In the first case we know the p-value and we say it is *unmasked* , whereas in the second case we know the p-value and its mirror image, but do not know which one is the true value; we say it is *masked*. The GMK procedure specifies that the threshold updating rule $\delta_2^{(t)}$ must be estimated using only information contained in the σ-algebra

$$\mathcal{K}_t = \sigma(\{\tilde{p}_{t,k,i}\}, A_t, R_t).$$

Notice that the collection $\{\mathcal{K}_t\}_{t \geq 0}$ is a filtration, as progressively more information is made available at each step of the procedure. More precisely,

Definition 6 (GMK procedure)**.** *The GMK procedure is defined as a multiple testing procedure that follows the general sequential stages given in Section 3.2, such that*

(i) $s^{(0)} = (1/2, 1/2, ..., 1/2)$;

(ii) $\delta_1(s^{(t)}) = I(\widehat{\text{FDR}}_{GMK}(s^{(t)}) \leq \alpha)$;

(iii) $\delta_2^{(t)}$ *is estimated using information in* $\mathcal{K}_t$.

Under the notation above, we can write

$$\widehat{\text{FDR}}_{\text{GMK}}(s^{(t)}) = \frac{1 + A_t}{\max\{R_t, 1\}}$$

so that step t of the GMK procedure only uses information in $\mathcal{K}_t$. The GMK procedure can be shown to control the overall FDR at the nominal level α in the finite sample case.

Theorem 3. *Assume that the p-values follow the null independence model. Then the data-driven GMK procedure as stated above controls the FDR at level* α.

The proof of Theorem 3 relies on an argument similar to Barber and Candès (2015) and Lei and Fithian (2018), themselves descendents of the martingale stopping time arguments first used by Storey et al. (2004). For a given time t in the sequential procedure, the FDP of the rejection set $S^{(t)}$ can be bounded above by

$$\text{FDP}_t \leq \widehat{\text{FDR}}_{\text{GMK}}(s^{(t)}) \cdot X_t$$

where X_t counts the ratio between the number of true nulls in the current rejection region and its mirror. By the null independence model, the true null p-values are uniformly distributed, so we expect $X_t \approx 1$. By construction, $\hat{t}$, the step when the GMK procedure stops, is a stopping time with respect to $\{\mathcal{K}_t\}_{t \geq 0}$, and satisfies

$$\text{FDP}_{\hat{t}} \leq \alpha \cdot X_{\hat{t}}.$$

It can be shown that $\{X_t\}_{t \geq 0}$ is a supermartingale with respect to a larger filtration $\{\mathcal{L}_t\}_{t \geq 0}$, so that

$$\text{FDR}_{\hat{t}} = E[\text{FDP}_{\hat{t}}] \leq \alpha \cdot E[X_0].$$

At step zero, all the true null p-values are masked. By the null independence model, we can use a binomial argument to bound $E[X_0] \leq 1$, thus completing the proof.

Proof. Without loss of generality, assume there are no ties in the p-values, so that at each step of the procedure, exactly one p-value is unmasked. Let $\hat{t}$ denote the step at which the

GMK procedure stops. Since each step t of the GMK procdure only uses information in $\mathcal{K}_t$, $\hat{t}$ is a stopping time with respect to the filtration $\{\mathcal{K}_t\}_{t\geq 0}$. By definition

$$\widehat{\mathrm{FDR}}_{\mathrm{GMK}}(s^{(\hat{t})}) \leq \alpha.$$

Define

$$V_{t,k} = |\{i : p_{k,i} < s_k^{(t)}, H_{k,i} = 0\}|, \text{ and } U_{t,k} = |\{i : p_{k,i} > 1 - s_k^{(t)}, H_{k,i} = 0\}|,$$

and let $V_t = \sum_{k=1}^K V_{t,k}$ and $U_t = \sum_{k=1}^K U_{t,k}$. Then

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

It only remains to show that $X_t := \frac{V_t}{1+U_t}$ is a supermartingale with respect to $\{\mathcal{L}_t\}_{t\geq 0}$, where

$$\mathcal{L}_t = \sigma(\{p_{k,i} : H_{k,i} = 1\}, \{\tilde{p}_{t,k,i} : H_{k,i} = 0\}, U_t, V_t),$$

which contains the information available in $\mathcal{K}_t$, plus the true location of any masked false null p-values. X_t is adapted to this filtration, so to show it is a supermartingale, it suffices to show that for all t

$$E[X_{t+1}|\mathcal{L}_t] \leq X_t.$$

If $\mathcal{L}_t = \mathcal{L}_{t+1}$, that is if a false null p-value is unmasked at this step of the filtration, then it follows that

$$\{X_{t+1}|\mathcal{L}_t\} = X_t$$

and hence the inequality holds. If $\mathcal{L}_t \subsetneq \mathcal{L}_{t+1}$, that is if a new true null p-value is unmasked at this step of the filtration, then we have

$$\{X_{t+1}|\mathcal{L}_t\} = \frac{V_t - B}{1 + U_t - (1 - B)}$$

where $B = 1$ if the revealed p-value is less than $1/2$, that is if it had previously been counted in V_t, and $B = 0$ otherwise. Since the true null p-values are independent and uniformly distributed, we have

$$P(B = 1|\mathcal{L}_t) = \frac{V_t}{V_t + U_t}.$$

Since a true null p-value is avaiable to be unmasked at this step of the filtration, we cannot have $V_t + U_t = 0$. We consider three cases. If $V_t = 0$ then $B = 0$ and it follows that

$$\{X_{t+1}|\mathcal{L}_t\} = X_t = 0.$$

If $U_t = 0$ then $B = 1$ and we have

$$E[X_{t+1}|\mathcal{L}_t] = V_t - 1 < X_t.$$

Finally if $V_t > 0$ and $U_t > 0$, then we have

$$\begin{aligned} E[X_{t+1}|\mathcal{L}_t] &= \left(\frac{V_t - 1}{U_t + 1}\right) \cdot \left(\frac{V_t}{V_t + U_t}\right) + \left(\frac{V_t}{U_t}\right) \cdot \left(\frac{U_t}{V_t + U_t}\right) \\ &= \frac{V_t^2 U_t + V_t U_t^2}{(V_t + U_t)(U_t + 1)U_t} \\ &= X_t. \end{aligned}$$

In each case the desired inequality holds, so it follows that X_t is a supermartingale with respect to $\{\mathcal{L}_t\}_{t\geq 0}$. Since $\mathcal{K}_t \subseteq \mathcal{L}_t$ and $\hat{t}$ is a stopping time with respect to $\{\mathcal{K}_t\}_{t\geq 0}$, $\hat{t}$ is a stopping time with respect to the larger filtration $\{\mathcal{L}_t\}_{t\geq 0}$. Thus by optional stopping theorem (Karlin and Taylor, 1975), we have

$$E\left[\frac{V_{\hat{t}}}{1 + U_{\hat{t}}}\right] = E[X_{\hat{t}}] \leq E[X_0].$$

Since all p-values are masked at time $t = 0$, all true null p-values are masked in the σ-algebra $\mathcal{L}_0$. Then by the independent and uniform assumptions on the true null p-values,

$$\{V_0|\mathcal{L}_0\} \sim \text{Binomial}(m_0, 1/2),$$

and $U_0 = m_0 - V_0$. Finally

$$\begin{aligned} E[X_0] &= E\left[\frac{V_0}{1 + m_0 - V_0}\right] \\ &= \sum_{i=1}^{m_0} \left(\frac{1}{2}\right)^{m_0} \cdot \frac{m_0!}{i!(m_0 - i)!} \frac{i}{1 + m_0 - i} \\ &= \sum_{i=1}^{m_0} \frac{m_0!}{(i-1)!(m_0 - (i-1))!} \cdot \left(\frac{1}{2}\right)^{m_0} \\ &= \sum_{j=0}^{m_0 - 1} \frac{m_0!}{j!(m_0 - j)!} \cdot \left(\frac{1}{2}\right)^{m_0} \\ &= 1 - \left(\frac{1}{2}\right)^{m_0} \\ &\leq 1, \end{aligned}$$

which completes the proof. This is because we have shown that under the GMK procedure,

$$\text{FDR} = E[\text{FDP}_{\hat{t}}] \leq \alpha E\left[\frac{V_{\hat{t}}}{1+U_{\hat{t}}}\right] \leq \alpha.$$

□

By noting that the oracle GMK procedure is a special case of the data-driven GMK procedure, we also have the following corollary.

Corollary 3. *Assume that the p-values follow the null independence model. Then the oracle GMK procedure defined in Section 3.3 controls the FDR at level α.*

The GMK procedure is defined for general $\delta_2^{(t)}$, but in practice the optimality result of Cai and Sun (2009) motivates defining $\delta_2^{(t)}$ to imitate the selection rule used in the oracle version of the procedure, through information-restricted estimates of the group-wise Lfdr functions.

3.5 Implementation

Implementation of the GMK procedure in the data-driven case requires a specific method to estimate $\delta_2^{(t)}$ at each step t. The method we provide in this section is designed to imitate the oracle GMK procedure by estimating the group-wise Lfdr functions with masked information using the expectation-maximization (EM) algorithm. Such a method is used in Sections 3.6 and 3.7 to define the threshold updating rules $\delta_2^{(t)}$. The GMK procedure that estimates $\delta_2^{(t)}$ in this way will be referred for the remainder of the chapter as the data-driven GMK procedure.

More precisely, for estimates $\{\widehat{\text{Lfdr}}_1, ..., \widehat{\text{Lfdr}}_K\}$ of the group-wise Lfdr functions, we define

$$(\hat{k}, \hat{i}) = \underset{(k,i)}{\text{argmax}}\{\widehat{\text{Lfdr}}_k(q_{k,i}) : q_{k,i} < s_k^{(t)}\},$$

and then define the threshold updating rule for the data-driven GMK procedure by

$$\delta_2(s^{(t)}) = (s_1^{(t)}, s_2^{(t)}, ..., q_{\hat{k},\hat{i}}, ..., s_K^{(t)}).$$

Estimation is done separately for each of the K groups, so for the remainder of this section, we consider p-values from a single group and for notational simplicity suppress the

group index. Suppose the p-values $\{p_i\}_{i=1}^m$ follow the two-group model. Then since the distribution of the true null p-values is known to be uniform on the interval $[0, 1]$, the Lfdr requires the estimation of the true null proportion π_0 and the density of false null p-values f_1. Both must be estimated with masked information: for some threshold $s_0 \leq 1/2$, we do not know the true location of the p-values below s_0 or above $1 - s_0$.

3.5.1 Estimation of π_0

A natural estimate of π_0 is the one defined in Storey (2002),

$$\hat{\pi}_0(\lambda) = \frac{\sum_{i=1}^m I(p_i > \lambda)}{m(1-\lambda)}$$

for a fixed tuning parameter $\lambda \in (0, 1)$. Notice that this relies on a count of the number of p-values above λ, a quantity that is unknown under masked information. However, we can construct an analogous estimator by observing that under masking we can count the number of p-values in any interval symmetric about $1/2$, and in any interval contained in $[s_0, 1 - s_0]$. Thus, for fixed tuning parameter $\lambda \in (0, 1/2)$, define the following limited-information estimator of π_0.

$$\hat{\pi}_0^{(\mathrm{LI})}(\lambda) = \begin{cases} \frac{\sum_{i=1}^m I(\lambda \leq p_i \leq 1-\lambda)}{m(1-2\lambda)}, & \text{if } s_0 \geq \lambda \\ \frac{\sum_{i=1}^m I(\lambda \leq p_i \leq 1-s_0)}{m(1-s_0-\lambda)}, & \text{if } s_0 < \lambda. \end{cases}$$

In the first case, many p-values are masked, and π_0 is estimated using the number of p-values in the symmetric interval $[\lambda, 1-\lambda]$. In the second case, more p-values are unmasked, and π_0 is estimated using the number of p-values in the fully unmasked interval $[\lambda, 1-s_0] \subseteq [s_0, 1-s_0]$. In our simulations, we used this estimator with fixed $\lambda \equiv 0.3$, although λ could also be chosen dynamically using a modified version of the right-boundary procedure (Liang and Nettleton 2012).

3.5.2 Estimation of f_1

We estimate f_1 parametrically using the EM algorithm, by treating the limited information as a missing data problem. As a parametric model, similar to Jin and Cai (2007), we assume the two-group model, and that the false null p-values are normally distributed (with variance 1) when transformed to the z-scale, that is

$$z_i = -\Phi^{-1}(p_i) \sim \begin{cases} \mathrm{N}(0, 1), & \text{if } H_i = 0 \\ \mathrm{N}(\theta, 1), & \text{if } H_i = 1 \end{cases}$$

where Φ^{-1} is the quantile function of the standard normal distribution. If a p-value is masked, then we do not know whether its true value is p_i or $1-p_i$. After transforming to the z-scale, we do not know whether the true value is z_i or $-z_i$. Hence, we define

$$y_i = |z_i|, \text{ and } B_i = I(z_i = y_i).$$

When a particular z-value z_i is masked, then we only know the absolute z-value y_i. However, if it is unmasked, then we also know B_i, which gives the sign of z_i, and the true z-value can be recovered. We apply the EM algorithm to estimate θ, where the full data is $\{y_i, B_i, H_i\}_{i=1}^m$. H_i is missing for $i = 1, ..., m$, and B_i is missing for the masked p_i's. Then the full-data likelihood for θ is

$$L(\theta) = \prod_{i=1}^{m} \phi_0(-y_i)^{(1-H_i)(1-B_i)} \phi_0(y_i)^{(1-H_i)B_i} \phi_\theta(-y_i)^{H_i(1-B_i)} \phi_\theta(y_i)^{H_i B_i}$$

where ϕ_θ denotes the density function of the $\mathrm{N}(\theta, 1)$ distribution. The maximum likelihood estimate for θ is found by maximizing the partial log-likelihood function,

$$\ell(\theta) = \sum_{i=1}^{m} H_i(1-B_i)\log(\phi_\theta(-y_i)) + H_i B_i \log(\phi_\theta(y_i)).$$

The EM algorithm produces a sequence of estimates $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, ...$ by alternating two steps, an E-step and an M-step, until some stopping criterion is met. For an arbitrary iteration $r+1$ of the EM algorithm, the E-step is derived by calculating the expected log-likelihood for each observation given the observed data, and an estimate $\hat{\theta}^{(r)}$ from the previous iteration of the algorithm. Denote the set of indices of masked p-values by $M^{(t)}$. For $i \notin M^{(t)}$, B_i is known, and the expected log-likelihood for that observation is

$$E[H_i | z_i; \hat{\theta}^{(r)}] \log(\phi_\theta(z_i))$$

where

$$\begin{aligned} E[H_i | z_i; \hat{\theta}^{(r)}] &= P(H_i = 1 | z_i; \hat{\theta}^{(r)}) \\ &= \frac{(1-\pi_0)\phi_{\hat{\theta}^{(r)}}(z_i)}{(1-\pi_0)\phi_{\hat{\theta}^{(r)}}(z_i) + \pi_0 \phi_0(z_i)} \end{aligned}$$

For $i \in M^{(t)}$, B_i is unknown and the expected log-likelihood is

$$E[H_i B_i | y_i; \hat{\theta}^{(r)}] \log(\phi_\theta(y_i)) + E[H_i(1-B_i) | y_i; \hat{\theta}^{(r)}] \log(\phi_\theta(-y_i)).$$

To calculate these two expectations, note that

$$\begin{aligned}
P(B_i = 1|y_i, H_i = 1; \hat{\theta}^{(r)}) &= \frac{\phi_{\hat{\theta}^{(r)}}(y_i)}{\phi_{\hat{\theta}^{(r)}}(y_i) + \phi_{\hat{\theta}^{(r)}}(-y_i)}, \\
P(B_i = 0|y_i, H_i = 1; \hat{\theta}^{(r)}) &= \frac{\phi_{\hat{\theta}^{(r)}}(-y_i)}{\phi_{\hat{\theta}^{(r)}}(y_i) + \phi_{\hat{\theta}^{(r)}}(-y_i)}, \\
P(H_i = 1|y_i; \hat{\theta}^{(r)}) &= \frac{(\phi_{\hat{\theta}^{(r)}}(y_i) + \phi_{\hat{\theta}^{(r)}}(-y_i))(1-\pi_0)}{(\phi_{\hat{\theta}^{(r)}}(y_i) + \phi_{\hat{\theta}^{(r)}}(-y_i))(1-\pi_0) + 2\phi_0(y_i)\pi_0}.
\end{aligned}$$

So that

$$\begin{aligned}
E[H_i B_i|y_i; \hat{\theta}^{(r)}] &= P(B_i = 1|y_i, H_i = 1; \hat{\theta}^{(r)}) \cdot P(H_i = 1|y_i; \hat{\theta}^{(r)}) \\
&= \frac{\phi_{\hat{\theta}^{(r)}}(y_i)(1-\pi_0)}{(\phi_{\hat{\theta}^{(r)}}(y_i) + \phi_{\hat{\theta}^{(r)}}(-y_i))(1-\pi_0) + 2 \cdot \phi_0(y_i)\pi_0} \\
E[H_i(1-B_i)|y_i; \hat{\theta}^{(r)}] &= P(B_i = 0|y_i, H_i = 1; \hat{\theta}^{(r)}) \cdot P(H_i = 1|y_i; \hat{\theta}^{(r)}) \\
&= \frac{\phi_{\hat{\theta}^{(r)}}(-y_i)(1-\pi_0)}{(\phi_{\hat{\theta}^{(r)}}(y_i) + \phi_{\hat{\theta}^{(r)}}(-y_i))(1-\pi_0) + 2 \cdot \phi_0(y_i)\pi_0}.
\end{aligned}$$

Denote $E[H_i|z_i; \hat{\theta}^{(r)}]$ by $w_i^{(r)}$, $E[H_i B_i|y_i; \hat{\theta}^{(r)}]$ by $w_{i+}^{(r)}$, and $E[H_i(1-B_i)|y_i; \hat{\theta}^{(r)}]$ by $w_{i-}^{(r)}$. Then the expected log-likelihood over all the observations is

$$Q(\theta, \hat{\theta}^{(r)}) = \sum_{i \notin M^{(t)}} w_i^{(r)} log(\phi_\theta(z_i)) + \sum_{i \in M^{(t)}} (w_{i+}^{(r)} log(\phi_\theta(y_i)) + w_{i-}^{(r)} log(\phi_\theta(-y_i))),$$

where the weights are calculated replacing π_0 by the limited-information estimator $\hat{\pi}_0^{(\mathrm{LI})}(\lambda)$ defined in Section 3.5.1.

The M-step updates the estimate of θ by maximizing the objective function Q over its first argument. Q is a weighted normal log-likelihood, so the optimal θ is a weighted average of the observations:

$$\hat{\theta}^{(r+1)} = \frac{\sum_{i \notin M^{(t)}} w_i^{(r)} z_i + \sum_{i \in M^{(t)}} (w_{i+}^{(r)} y_i + w_{i-}^{(r)}(-y_i))}{\sum_{i \notin M^{(t)}} w_i^{(r)} + \sum_{i \in M^{(t)}} (w_{i+}^{(r)} + w_{i-}^{(r)})},$$

which gives a closed form updating equation for $\hat{\theta}^{(r)}$. After initialization, this update step is iterated until the sequence of estimates meets a given stopping criterion. In our implementation of this EM algorithm, we iterated until consecutive estimates satisfied

$$|\hat{\theta}^{(r+1)} - \hat{\theta}^{(r)}| < \epsilon = 10^{-4}.$$

To initialize $\hat{\theta}^{(0)}$, we first estimate $\hat{m}_1 = \lfloor m(1 - \hat{\pi}_0^{(\mathrm{LI}(\lambda))}) \rfloor$, and then

$$\hat{\theta}^{(0)} = \frac{1}{\hat{m}_1} \cdot \sum_{i=m-\hat{m}_1+1}^{m} y_{(i)},$$

where $y_{(1)} \leq \cdots \leq y_{(m)}$ denote the order statistics of the absolute z-values.

Based on estimates $\hat{\theta}$ and $\hat{\pi}_0$ of θ and π_0, the Lfdr can be estimated as

$$\widehat{\mathrm{Lfdr}}(p) = \frac{\hat{\pi}_0 \phi_0(-\Phi^{-1}(p))}{\hat{\pi}_0 \phi_0(-\Phi^{-1}(p)) + (1 - \hat{\pi}_0)\phi_{\hat{\theta}}(-\Phi^{-1}(p))}.$$

Dividing numerator and denominator by $\phi_0(-\Phi^{-1}(p))$, it can be seen that this expression is equivalent to the expression for Lfdr given in Section 3.1. This EM method can be easily extended to other parametric models, for instance we could remove the assumption that the false null p-values have unit variance when transformed to the z-scale, or assume the false null p-values follow a beta distribution.

3.6 Simulation

In this section, we present a variety of simulation settings to demonstrate the performance of the oracle and data-driven GMK procedures (`GMK`) relative to other grouped p-value multiple testing procedures in the literature. Other procedures presented for comparison are the adaptive BH procedure (`MBH`) described in Storey et al. (2004), the group-weighted BH procedure (`GBH`) of Hu et al. (2010), and the optimally weighted BH procedure (`WO`) of Zhao and Zhang (2014). Note that Zhao and Zhang (2014) define two procedures, and we utilize their so-called "Pro2", which is asymptotically more powerful than their "Pro1". As `MBH` is designed for the exchangeable case, it ignores group labels altogether. The oracle *CLfdr* procedure of Cai and Sun (2009), which is theoretically optimal, is used as a benchmark level to assess the power of the four procedures. However, we did not include the data-driven CLfdr procedure in these simulations, as it aims to control the mFDR rather than the FDR (Cai and Sun, 2009).

All of the procedures considered have both oracle and data-driven versions. In the data-driven setting, group-wise null proportions $\pi_{0,k}$ are estimated using the lowest slope procedure (Benjamini and Hochberg, 2000). In the data-driven GMK procedure, rather than estimating new parameters for the Lfdr at every sequential step, the estimates are updated $G = 10$ times for each group as more information becomes available. Sensitivity

analysis to the choice of this "refresh rate" G showed that 10 updates is sufficient, and there is no significant power advantage to updating the Lfdr parameters more often.

Our simulation settings are similar to those in Cai and Sun (2009); and Zhao and Zhang (2014). In settings 1–4, p-values are calculated from normal statistics: true null statistics follow standard normal $\mathrm{N}(0,1)$, and false null statistics follow $\mathrm{N}(\theta,1)$ for some $\theta > 0$. We refer to θ as the signal strength. The p-values are calculated to test $H_{k,i} : \mu_{k,i} = 0$ against the one sided alternative $\mu_{k,i} > 0$. In these cases, data-driven `GMK` estimates the Lfdr from the correct parametric model. Settings 5–7 demonstrate the robustness of data-driven `GMK` to a misspecified parametric model by generating the alternative p-values from a $\mathrm{Beta}(\tau,1)$ distribution, while still estimating the Lfdr using the normal EM algorithm described in Section 3.5.2. Note that in the normal case, large values of θ correspond to strong signals, while in the beta case, small values of τ correspond to strong signals. In all settings the nominal significance level is $\alpha = 0.1$ and $J = 1000$ independent replications are performed.

1. $K = 2$ groups of sizes $m_1 = 3000$ and $m_2 = 1500$. Signal strengths are fixed at $\theta_1 = 2$ and $\theta_2 = 4$. The true null proportion for group 2 is fixed at $\pi_{0,2} = 0.9$, while $\pi_{0,1}$ varies from 0.71 to 0.99 with increment size 0.02.

2. $K = 2$ groups of sizes $m_1 = 3000$ and $m_2 = 1500$. The signal strength for group 2 is fixed at $\theta_2 = 4$, while θ_1 varies from 2.5 to 4.9 with increment size 0.2. True null proportions are fixed at $\pi_{0,1} = 0.8$ and $\pi_{0,2} = 0.9$.

3. $K = 2$ groups, with $m_2 = 1500$ fixed, while m_1 varies from 500 to 5000 with increment size 500. Signal strengths are fixed at $\theta_1 = 2$, $\theta_2 = 4$, and true null proportions are fixed at $\pi_{0,1} = 0.8$, $\pi_{0,2} = 0.9$.

4. $m = 5000$, and split into K equal-sized groups, where $K = 2, 4, 5, 8, 10, 15, 20$. The signal strengths θ_k are taken as an equally spaced sequence of length K between 2 and 6, and the true null proportions $\pi_{0,k}$ are taken as an equally spaced sequence of length K between 0.65 and 1.

5. $K = 2$ groups of sizes $m_1 = 3000$ and $m_2 = 1500$. False null p-values follow a $\mathrm{Beta}(\tau,1)$ distribution, with $\tau_1 = 1/4$ and $\tau_2 = 1/8$. The true null proportion for group 2 is fixed at $\pi_{0,2} = 0.9$, while $\pi_{0,1}$ varies from 0.71 to 0.99 with increment size 0.02.

6. $K = 2$ groups of sizes $m_1 = 3000$ and $m_2 = 1500$. False null p-values follow a $\mathrm{Beta}(\tau,1)$ distribution, with $\tau_2 = 1/8$ fixed, while τ_1^{-1} varies from 5 to 9.8 with increment size 0.4. True null proportions are fixed at $\pi_{0,1} = 0.8$ and $\pi_{0,2} = 0.9$.

7. $K = 2$ groups, with $m_2 = 1500$ fixed, while m_1 varies from 500 to 5000 with increment size 500. Signal strengths are fixed at $\tau_1 = 1/4$, $\tau_2 = 1/8$, and true null proportions are fixed at $\pi_{0,1} = 0.8$, $\pi_{0,2} = 0.9$.

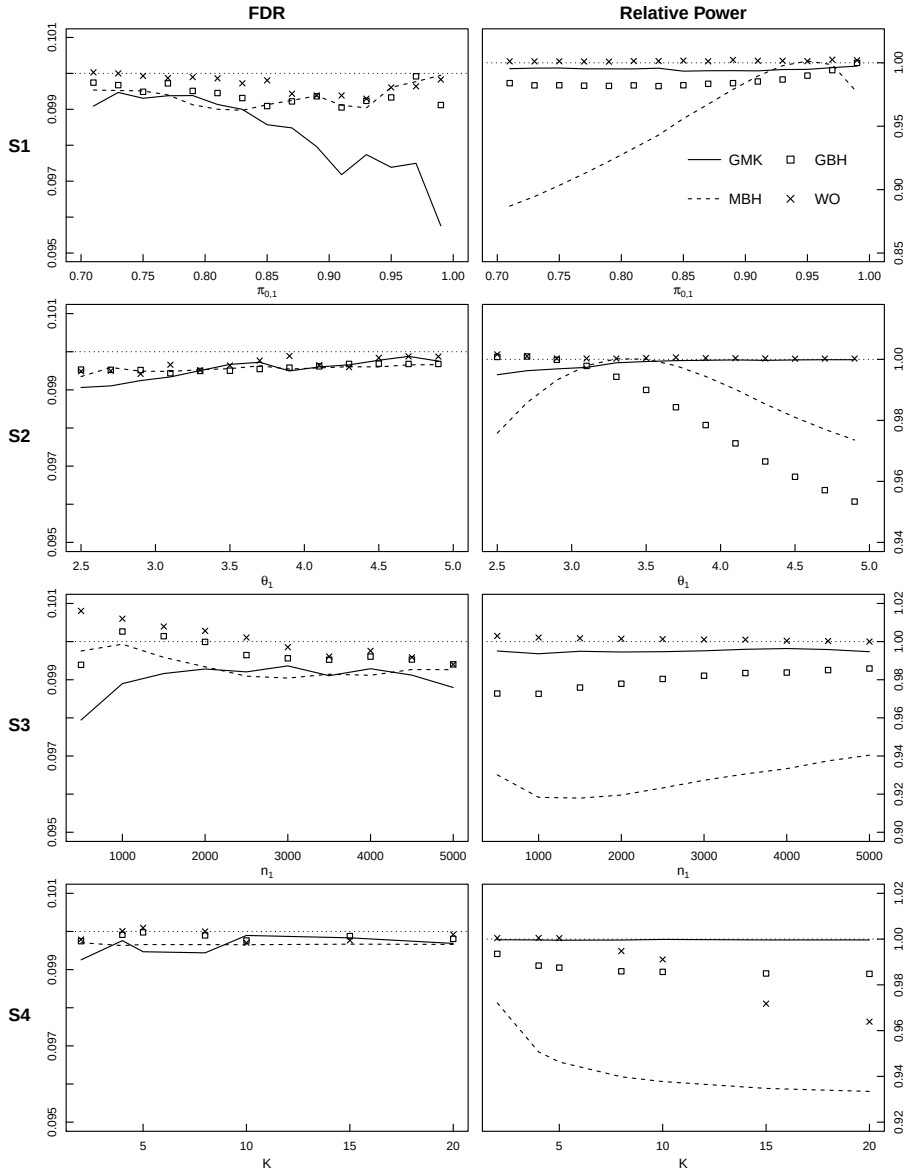

Figure 3.1: Realized FDR and relative power, oracle procedures, settings 1–4

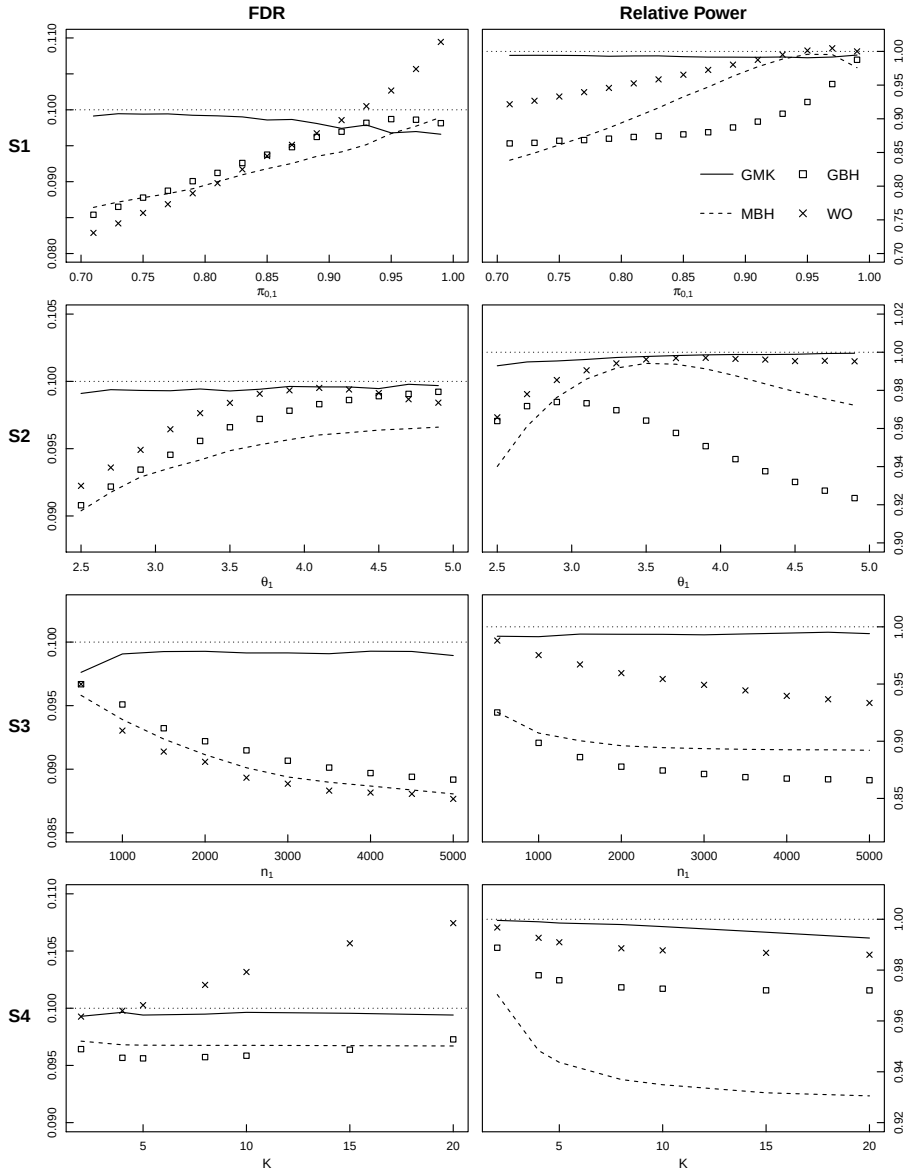

Figure 3.2: Realized FDR and relative power, data-driven procedures, settings 1–4

3.6.1 Results

Figure 3.1 presents the results of the oracle procedures for settings 1–4. In the oracle case, `MBH`, `GBH` and `GMK` have theoretical finite sample control of the FDR, while `WO` controls

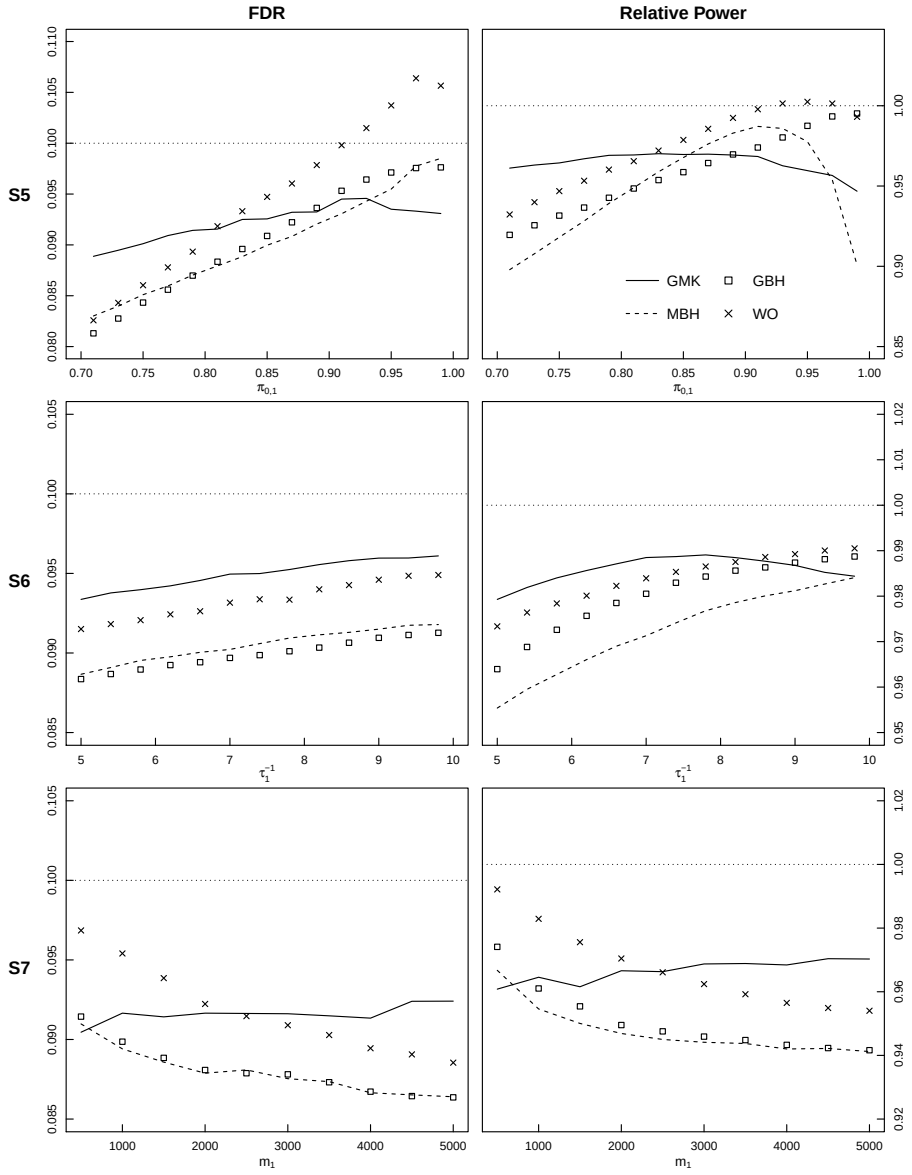

Figure 3.3: Realized FDR and relative power, data-driven procedures, settings 5–7

the FDR asymptotically. All procedures control the FDR very close to the nominal level $\alpha = 0.1$. In setting 1, `MBH` does not achieve the power of the other procedures, especially when the group null proportions are sufficiently different and group labeling becomes highly

informative. GBH achieves greater power than MBH, but does not achieve the power of GMK or WO since it does not incorporate the group signal strengths. Although the FDR of GMK is well below the nominal level when $\pi_{0,1}$ is close to 1, it still has near the optimal level of power. Setting 2 gives similar results for small θ_1. For large θ_1, where there is disagreement between the true null proportion and signal strength, the ignorance of GBH to the signal strength causes it not to achieve the power of MBH, which ignores group labels altogether. Setting 3 demonstrates the same relative performances as setting 1. In setting 4, it is demonstrated that oracle GMK is robust to a large number of groups. In all oracle cases, GMK achieves near to the optimal level of power.

Figure 3.2 presents the results of the data-driven procedures for settings 1–4. In the data-driven case, MBH and GMK have theoretical finite sample control of the FDR (based on Theorem 2 and Theorem 3, respectively), while GBH and WO control the FDR asymptotically. In setting 1, for low values of $\pi_{0,1}$, when the lowest-slope estimator is particularly conservative, MBH, GBH and WO control the FDR below the nominal level, and as a result, GMK is able to achieve greater power than the other three procedures. For values of $\pi_{0,1}$ close to 1, the lowest-slope estimator becomes relatively less conservative and the FDR levels of WO are above the nominal level. In setting 2, we see results similar to the oracle case, and GMK demonstrates its finite sample FDR control while achieving the greatest power of the four procedures. In setting 3, the conservatism of the lowest-slope estimator again leads to MBH, GBH and WO controlling the FDR well below the nominal level, allowing GMK to achieve the greatest power, especially when m_1 is large. Setting 4 demonstrates the robustness of GMK to the number of groups, while also demonstrating that for large K, WO can lose control of the FDR at the nominal level. In all cases, GMK maintains finite sample control, and is either comparable or significantly exceeds the power levels of all other procedures.

The data-driven setting 1 in particular demonstrates the advantage of GMK over its competitors. Because all other procedures incorporate conservative estimates of $\pi_{0,k}$ in their stopping rules δ_1, they control the FDR below the nominal level, and thus do not achieve the optimal level of power. On the other hand, if $\pi_{0,k}$ is not estimated conservatively enough, such as when $\pi_{0,1}$ is close to 1, it may cause GBH and WO to lose control of the FDR in a finite sample setting, even when m is large. GMK guarantees finite sample control of the FDR, and can almost exhaust the target FDR level.

Figure 3.3 presents the results of the data-driven procedures for settings 5– 7. Since they are meant to demonstrate robustness of the parametric EM estimation procedure for the Lfdr, we did not include oracle results for these settings, although the oracle CLfdr procedure was still run to provide a benchmark level for power. In setting 5, it is shown that while GMK is the most conservative of the four procedures when $\pi_{0,1}$ is close to 1, it has the best power for lower values of $\pi_{0,1}$. As in setting 1, WO loses control of FDR when

$\pi_{0,1}$ is close to 1. Setting 6 demonstrates that with beta distributed false null p-values, signal strength is less influential than in the normal case, and so when the group null proportions are similar, there is little room for improvement over `MBH`, and all procedures achieve over 95% of the oracle level of power. The misspecification of the p-value model hurts the power of `GMK` relative to the other procedures. However, it continues to maintain finite sample control, and in most cases produces power levels comparable to those of the other procedures. Finally, setting 7 demonstrates that `GMK` achieves the best relative performance in terms of FDR control and power as group size m_1 grows, similar to setting 3. In all three of these settings, `GMK` controls the FDR more conservatively than in the normal case. The impurity of the beta distribution (see Section 2.6.1) leads to identifiability issues in the estimation of $\pi_{0,k}$, and conservatively biases $\widehat{\text{FDR}}_{\text{GMK}}$. For comparison, we also implemented an EM algorithm using the correct beta model, but these results are not plotted. The correct model led to an improvement in power, but the effects of impurity are still present no matter the parametric form chosen for Lfdr. The normal model still performs admirably in some cases, demonstrating that the Lfdr functions can be closely approximated, even under an incorrect parametric model.

3.7 Application

We apply our proposed data-driven GMK procedure to a real dataset, the adequate yearly progress (AYP) study of California high schools for the year 2007. This dataset was analyzed in a similar fashion by Cai and Sun (2009) and Zhao and Zhang (2014). The dataset consists of observations of academic performance for 7867 California high schools, and the intention of this analysis is to apply grouped multiple testing procedures to identify "interesting" schools: those for which the relative performance of socioeconomically advantaged (SEA) students and socioeconomically disadvantaged (SED) students (in terms of success rate on math exams) differs from the typical amount. Previous analysis by Efron (2007, 2008), and Cai and Sun (2009) has shown that the typical relative performance of SEA and SED students is highly correlated with the school size, and a more informative list of "interesting" schools can be identified when the schools are grouped according to student population.

For $i = 1, .., m$, $m = 7687$, denote the number of successful SEA students at school i by X_i, out of a total of s_{xi} SEA students and the number of successful SED students at school i by Y_i, out of a total of s_{yi} SED students. Then the success rate among SEA students is $R_{xi} = X_i/s_{xi}$, and the success rate among SED students is $R_{yi} = Y_i/s_{yi}$. A summary

	α	0.01	0.025	0.04	0.055	0.07	0.085	0.10	0.115
Small	MBH	6	6	8	10	11	13	14	14
	GBH	6	6	9	10	12	14	14	15
	WO	6	6	9	10	11	13	10	10
	GMK	0	6	7	10	14	14	14	14
Medium	MBH	49	66	90	114	140	149	164	187
	GBH	46	61	82	99	116	132	148	153
	WO	49	59	87	107	140	149	167	197
	GMK	0	59	94	149	200	203	241	248
Large	MBH	32	37	42	51	60	63	67	70
	GBH	37	47	61	67	73	77	81	82
	WO	37	61	61	63	60	69	68	67
	GMK	0	38	52	67	73	73	77	79
Total	MBH	87	109	140	175	211	225	245	271
	GBH	89	114	152	176	201	223	243	250
	WO	92	126	157	180	211	231	245	274
	GMK	0	103	153	226	287	290	332	341

Table 3.1: Group-wise and total rejections, AYP data

statistic comparing SEA and SED performance at school i can be constructed as

$$Z_i = \frac{R_{xi} - R_{yi} - \gamma}{\sqrt{\frac{R_{xi}(1-R_{xi})}{s_{xi}} + \frac{R_{yi}(1-R_{yi})}{s_{yi}}}},$$

where $\gamma = \text{median}(R_{x1}, ..., R_{xm}) - \text{median}(R_{y1}, ..., R_{ym})$ is a centering constant. We group the data in the same way as Cai and Sun (2009) and Zhao and Zhang (2014), into a small group ($s_{xi} + s_{yi} \leq 120$), a medium group ($120 < s_{xi} + s_{yi} < 900$), and a large group ($s_{xi} + s_{yi} \geq 900$). For each group, the empirical null distribution is estimated using the method of Jin and Cai (2007), and p-values are calculated to test whether each Z_i comes from its group's null distribution. The same four data-driven procedures from Section 3.6 are applied to control the FDR at a range of nominal α levels. `GMK` estimates the parameters of the group-wise Lfdr functions using the normal EM algorithm described above. `MBH`, `GBH` and `WO` estimate $\pi_{0,k}$ using the lowest-slope procedure (Benjamini and Hochberg, 2000).

Figure 3.4 plots the results of the analysis for the four procedures, and rejections are also reported in Table 3.1. As the nominal level α is increased, all the procedures are able to identify more interesting schools. For sufficiently large α, `GMK` performs the best of all the procedures in terms of total rejections. When $\alpha = 0.01$, `GMK` returns zero rejections,

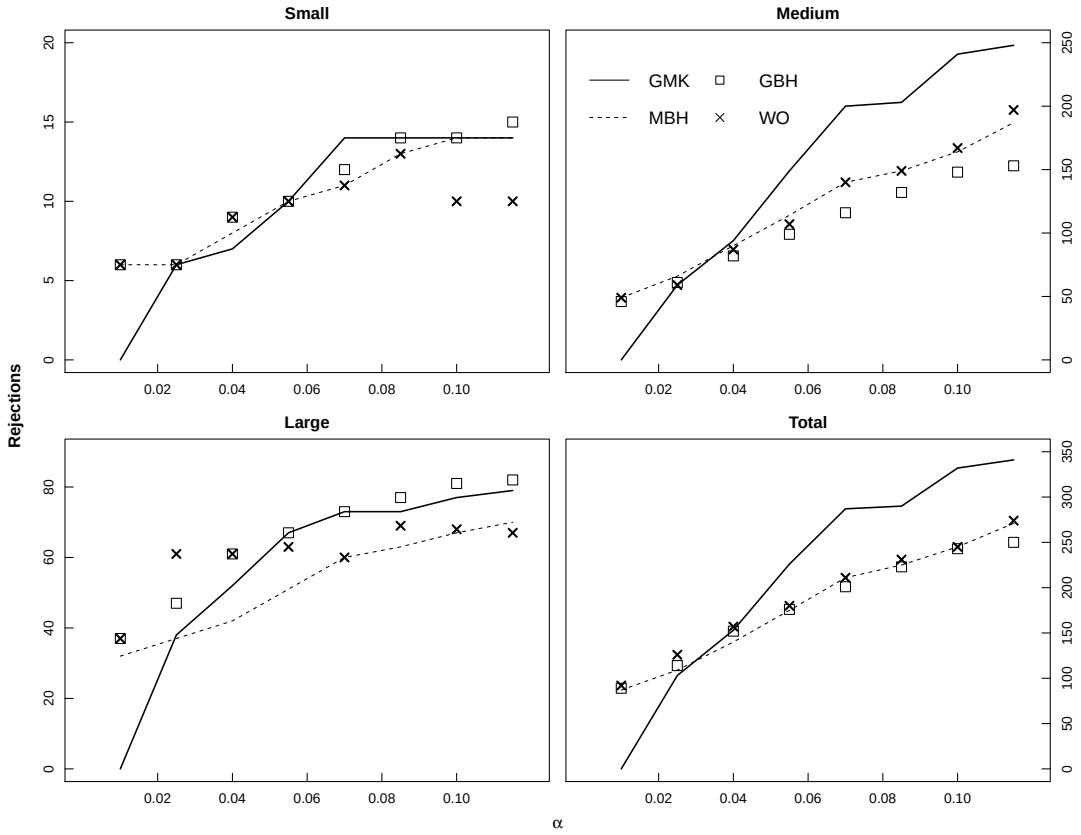

Figure 3.4: Group-wise and total rejections, AYP data

since its estimate of the FDR is never less than 0.01 at any step of the procedure. The discrete nature of the FDR estimator in the `GMK` stopping rule means that it can perform poorly if n is not large enough, especially for small values of α. Notice that the numerator of the estimate of the FDR used in the `GMK` stopping rule δ_1 is bounded below by 1, so it follows that `GMK` will never return fewer than $\lceil 1/\alpha \rceil$ rejections without returning zero rejections, which can limit its effectiveness when α is very small.

For small values of α, `GBH` outperforms `MBH` (which ignores group labels), but for larger values of α, `MBH` has better performance in terms of total rejections. `GBH` calculates its weights using only the estimated true null proportions, and ignores other factors like group size and signal strength that may impact total rejections. In this case, the estimates of $\pi_{0,k}$ encourage `GBH` to seek out more rejections in the large group, at a loss of total rejections due to the fact that most schools lie in the medium group. `WO` always performs the best out of

the three competing methods, as in the data-driven case it chooses its weights to maximize total rejections. Note that with the exception of `WO`, the methods have a "monotone" property: the number of rejections in each group always increases with α. However in `WO`, the optimization of weights for each specfic α-level can lead to a decrease in rejections for a particular group, despite increasing α, in exchange for more rejections in another.

Chapter 4

Conclusion

In this thesis, we have considered the multiple testing problem, with an aim towards powerful and flexible data-driven procedures that control the FDR in finite samples.

In Chapter 2, under the classical null independence model, we showed the novel result of finite sample control of the FDR for a broad class of dynamic adaptive procedures, namely those with LRS selection rules. This FDR control result is then extended to the class of dynamic adaptive procedures where λ is selected using an LRS selection rule on the set of p-values. It is demonstrated through simulation that the right-boundary procedure (`RB20`) and quantile-based right-boundary procedure (`RB20q`) outperform the competing dynamic adaptive procedures in terms of power and estimation accuracy of π_0, while maintaining control of the FDR at the nominal level. In similar simulation settings in Liang and Nettleton (2012), the RB20* procedure, which is a minor variation of `RB20`, was shown to be more powerful than many fixed adaptive procedures, such as $\lambda = 0.5$, the k-quantile and the two-stage procedures of Benjamini et al. (2006), and the two-stage step-down procedure of Blanchard and Roquain (2009). The simulation results thus far show that the right-boundary procedure is the most powerful adaptive procedure among all adaptive procedures with finite sample FDR control.

In Chapter 3, under a general framework for grouped multiple testing procedures, and under the classical null independence model, we have proposed the oracle grouped mirrored knockoff procedure, which ranks the significance of hypotheses using the group-wise Lfdr, while controlling the FDR; and the data-driven grouped mirrored knockoff procedure, which uses information-restricted estimates of the group-wise Lfdr to rank the significance of hypotheses while maintaining finite sample control of the FDR. Under the Bayesian two-group model, the threshold updating rule used by the oracle GMK procedure is optimal in

terms of expected power. It is demonstrated through simulation that the data-driven GMK procedure outperforms existing weighted p-value procedures in the literature, both in terms of power and maintenance of FDR control, and that it gives comparable performance even when the Lfdr model is misspecified. We also gave a real data example to demonstrate the performance of the GMK procedure in applications.

The unifying theme of this thesis is caution in the naive use of data-driven procedures for FDR control. While Storey's thresholding procedure with fixed λ has long been known to control the FDR in finite samples for each $\lambda \in (0, 1]$ (Storey et al., 2004), this does not immediately give the simultaneous control of the FDR for all λ. The results of Chapter 2 establish that finite sample control of the FDR is maintained for particular data-driven choices of λ, but these proofs are non-trivial and rely on an idea of data masking: the choice of λ is made with some information restriction, such that it remains a stopping time with respect to $\{\mathcal{F}_t\}_{t\in(0,1]}$. A similar data masking approach underlies the GMK procedure in Chapter 3. At each step of the procedure, the threshold updating decision is made subject to masking of a subset of the p-values. Simulations in Section 3.6 demonstrate that this data masking is essential to maintain finite sample control of the FDR. More aggressive procedures, like the data-driven WO procedure, that overuse the observed data to optimize tuning parameters can lose control of the FDR in finite samples, despite their asymptotic assurances.

4.1 Future work

The results of Chapter 2 strengthen the connection between the FDR estimation approach and the FDR control approach. With a conservative FDR estimator, we can use the thresholding procedure to find the largest p-value whose FDR estimate is below the target FDR level, and typically this thresholding procedure will control the FDR. This connection is most evident for fixed adaptive procedures through the work of Storey et al. (2004); and Liang and Nettleton (2012). It is further studied for certain dynamic adaptive procedures by Heesen and Janssen (2015). This thesis extends the connection to a further class of dynamic adaptive BH procedures. However, conservative estimation was established by Liang and Nettleton (2012) whenever λ is a stopping time with respect to $\{\mathcal{F}_t\}_{t\in(0,1]}$. This motivates that a proof of FDR control may still be possible for this broader class of procedures where λ is a stopping time, but is not necessarily an LRS selection rule.

It is typically difficult to establish FDR control results for adaptive procedures without assuming the null independence model. However, the simulation studies in Chapter 2 motivate that finite sample control may hold under certain types of block or autoregressive

dependence. The structure of the proof of Theorem 2 is such that if finite sample control can be shown under a particular dependence structure for fixed grid LRS selection rules, it can immediately be extended to right continuous p-grid LRS selection rules.

With respect to Chapter 3, there are several further research directions that could lead to improvement of the GMK procedure. The discussion in Section 3.7 of the issues for very small α-levels motivates an adjustment of the FDR estimate that allows the numerator to shrink to zero with the threshold, so that the procedure can return fewer than $\lceil 1/\alpha \rceil$ rejections.

Simulation results show that the GMK procedure is more variable in its rejection sets than the other weighted procedures, with the standard error of number of true rejections 50% - 60% higher on average than that of the WO procedure. This is due to high variability in A_t when the components of $s^{(t)}$ are very small, which leads to a highly variable stopping rule δ_1. It would be valuable to investigate whether this could be remedied by constructing a less variable estimate of the FDR that still maintains the martingale structure that allows us to prove finite sample control of the FDR.

Simulation settings 5–7 in Section 3.6 demonstrate that the power of the data-driven GMK procedure is sensitive to the parametric model chosen for Lfdr. Hence there is potential for improvement in power through a model selection step. However, if the GMK procedure is run to completion with various models and then a model is selected after the fact with respect to, for instance, total rejections, it will violate the information restrictions of the procedure. Hence, this model selection would have to be implemented within the steps of the procedure, by evaluating AIC or some similar criterion each time the Lfdr estimates are updated.

References

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

Y. Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society, Series B*, 72(4):405–416, 2010.

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*, 57(1):289–300, 1995.

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

Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

Y. Benjamini, A.M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

G. Blanchard and E. Roquain. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10:2837–2871, 2009.

T.T. Cai and W. Sun. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481, 2009.

G. Casella and R.L. Berger. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

B. Efron. Doing thousands of hypothesis tests at the same time. *Metron International Journal of Statistics*, 65(1):3–21, 2007.

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23 (1):1–22, 2008.

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

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.

J. Fan and X. Han. Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society, Series B*, 79(4):1143–1164, 2017.

J. Fan, X. Han, and W. Gu. Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035, 2012.

C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B*, 64(3):499–517, 2002.

C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061, 2004.

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

W. Guo and S. Sarkar. Adaptive controls of FWER and FDR under block dependence. 2016. Preprint.

P. Heesen and A. Janssen. Inequalities for the false discovery rate (FDR) under dependence. *Electronic Journal of Statistics*, 9(1):679–716, 2015.

P. Heesen and A. Janssen. Dynamic adaptive multiple tests with finite sample FDR control. *Journal of Statistical Planning and Inference*, 168:38–51, 2016.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

J.X. Hu, H. Zhao, and H.H. Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227, 2010.

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

J. Jin and T.T. Cai. Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478): 495–506, 2007.

S. Karlin and H.M. Taylor. *A first course in stochastic processes*. Academic Press, New York, 1975.

L. Lei and W. Fithian. AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society, Series B*, 2018. In press.

K. Liang and D. Nettleton. Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society, Series B*, 74(1): 163–182, 2012.

M.O. Mosig, E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller, and A. Friedmann. A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics*, 157(4):1683–1698, 2001.

E. Roquain and M.A. Van De Wiel. Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3:678–711, 2009.

S.K. Sarkar. On methods controlling the false discovery rate. *Sankhyā: The Indian Journal of Statistics, Series A*, 70(2):135–168, 2008.

T. Schweder and E. Spjøtvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.

R.J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 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*, 66(1):187–205, 2004.

W. Sun and T.T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

H. Zhao and J. Zhang. Weighted p-value procedures for controlling FDR of grouped hypotheses. *Journal of Statistical Planning and Inference*, 151–152:90–106, 2014.

APPENDICES

A.1 Proof of Theorem 1

For $k \geq 1$, consider a fixed λ candidate set $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ that divides the interval $(0, 1]$ into $k+1$ bins with boundaries at $\lambda_0 \equiv 0 < \lambda_1 < \ldots < \lambda_k < \lambda_{k+1} \equiv 1$ such that the ith bin is $(\lambda_{i-1}, \lambda_i]$ for $i = 1, \ldots, k+1$. For $1 \leq i \leq k$, let V_i denote the number of true null p-values falling into the ith bin,

$$V_i = V(\lambda_i) - V(\lambda_{i-1}) = \#\{p_j : j \in \mathcal{H}_0, p_j \in (\lambda_{i-1}, \lambda_i]\}.$$

Denote a single realization of a bin count by v_i for $1 \leq i \leq k$. Similarly, for $1 \leq i \leq k$, denote the number of false null p-values falling into the ith bin by S_i, with a particular realization denoted by s_i.

In general, $\widetilde{\cdot}$ applied to V_i, or v_i for $1 \leq i \leq k$ will denote the cumulative sum from 1 to i. For instance, $\widetilde{V_i} = \sum_{j=1}^{i} V_j = V_1 + ... + V_i$.

A.1.1 Distributional results

Under the null independence model, the bin counts of true null p-values have a multinomial distribution, that is

$$(V_1, V_2, ..., V_k) \sim \text{MULT}(m_0, \lambda_1, \lambda_2 - \lambda_1, ..., \lambda_k - \lambda_{k-1}).$$

By the properties of the multinomial distribution,

$$(V_1, ..., V_i) \sim \text{MULT}(m_0, \lambda_1, \lambda_2 - \lambda_1, ..., \lambda_i - \lambda_{i-1}) \text{ for } 1 \leq i \leq k$$

A.1.2 Lemmas for Theorem 1

Lemmas 1 and 2 below appear as Lemmas 3 and 4 respectively in Storey et al. (2004). Their statements have been slightly altered to correct an error in that paper, wherein under their original formulations the proposed martingale was not adapted to the proposed filtration. A similar error also appears in the filtrations of Liang and Nettleton (2012), but in both cases it has no effect on the validity of any results or proofs once it is redefined.

Let $\mathcal{H}_0$ be the set of index of all true null hypotheses. Define the filtration $\{\mathcal{G}_t\}_{t\in[0,1)}$ by $\mathcal{G}_t = \sigma((V(s), S(s)) : t \leq s \leq 1)$.

Lemma 1. *Under the null independence model, $V(t)/t$ for $t \in [0, 1)$ is a martingale with time running backwards with respect to the filtration $\{\mathcal{G}_t\}_{t\in[0,1)}$.*

Lemma 2. *When $\lambda \in \Lambda = \{\lambda_1, ..., \lambda_k\}$ is a selected by an LRS rule, the random variable $t_\alpha^\lambda := t_\alpha(\widehat{\text{FDR}}_\lambda^*)$ is a stopping time with respect to $\mathcal{G}_t^{\lambda_1} := \mathcal{G}_{t\wedge\lambda_1}$.*

Some additional definitions are required to state Lemma 3. Let

$$\mathcal{S} = \sigma(S(t) : 0 \leq t \leq 1)$$

be the information given by the locations of the false null p-values. In addition to the previous definition of left-to-right selection rules from Section 2.3, define the following.

Definition 7. *A random variable $\lambda(\Lambda)$ is a true null LRS selection rule if for all finite grids $\Lambda \subset (0, 1)$,*

(i) $\lambda(\Lambda)$ takes values in Λ.

(ii) $\lambda(\Lambda)$ is a stopping time with respect to $\{\mathcal{B}_t\}_{t\in(0,1]}$, where

$$\mathcal{B}_t = \sigma(V_j : \lambda_j \leq t).$$

Lemma 3. *If a selection rule is LRS on a fixed finite grid Λ, then conditional on $\mathcal{S}$, it is true null LRS on Λ.*

Proof. Suppose that under the conditioning, $(S_1, ..., S_k) = (s_1, ..., s_k)$. Since the original rule is LRS, let the selection sets be given by $C_1, ..., C_k \subseteq \{0, ..., m\}$. Then for $1 \leq i \leq k$, define

$$B_i = B_i(V_1, V_2, ..., V_{i-1}, s_1, ..., s_k) = \{w : 1 \leq w \leq m_0, w + s_i \in C_i\}.$$

Note that C_i depends on the past bin counts, which are given by

$$(V_1, ..., V_{i-1}, s_1, ..., s_{i-1}),$$

and also on s_i, so that B_i depends only on the false null bin counts and the past true null bin counts. We want to show that conditional on $\mathcal{S}$, the B_i give the λ selection behaviour of the procedure.

Fix some $1 \leq j \leq k$ and suppose that $\{\lambda = \lambda_j \,|\mathcal{S}\}$ occurs. This occurs if and only if

$$V_1 + s_1 \notin C_1, ..., V_{j-1} + s_{j-1} \notin C_{j-1}, V_j + s_j \in C_j$$

and $V_1 + \cdots + V_j \leq m_0$, which occurs if and only if

$$V_1 \notin B_1, ..., V_{j-1} \notin B_{j-1}, V_j \in B_j$$

and thus conditional on $(S_1, ..., S_k) = (s_1, ..., s_k)$, the selection rule is true null LRS with selection sets $B_1, ..., B_k \subseteq \{0, ..., m_0\}$. □

A.1.3 Proof of the theorem

Proof. By Lemma 1, we have that $\{\frac{V(t)}{t} : 0 \leq t < \lambda_1\}$ is a martingale with time running backwards with respect to the filtration $\mathcal{G}_t^{\lambda_1}$. Thus by Lemma 2,

$$E\left[\frac{V(t_\alpha^\lambda)}{t_\alpha^\lambda}\middle|\mathcal{G}_{\lambda_1}\right] = \frac{V(\lambda_1)}{\lambda_1},$$

and by following the proof steps of Theorem 3 of Storey et al. (2004), it is straightforward to show that

$$\mathrm{FDR}(t_\alpha(\widehat{\mathrm{FDR}}_\lambda^*)) \leq \alpha E\left[\frac{1-\lambda}{m - R(\lambda) + 1}\frac{V(\lambda_1)}{\lambda_1}\right].$$

Then to prove the theorem it is sufficient to show that the expectation on the right-hand side of the above inequality is bounded above by 1. Note that this is exactly the sufficient condition for finite sample control derived by Heesen and Janssen (2016), Proposition 1.

For any λ, $m = m_0 + m_1$, $R(\lambda) = V(\lambda) + S(\lambda)$, and $m_1 - S(\lambda) \geq 0$, so we may establish the bound

$$\begin{aligned} E\left[\frac{1-\lambda}{m - R(\lambda) + 1}\frac{V(\lambda_1)}{\lambda_1}\right] &= E\left[\frac{1-\lambda}{m_0 - V(\lambda) + 1 + (m_1 - S(\lambda))}\frac{V(\lambda_1)}{\lambda_1}\right] \\ &\leq E\left[\frac{1-\lambda}{m_0 - V(\lambda) + 1}\frac{V(\lambda_1)}{\lambda_1}\right] \end{aligned}$$

For $k = 1$, λ always takes on the fixed value λ_1, and $V(\lambda_1) \sim BIN(m_0, \lambda_1)$. The required bound in this fixed λ case is established by Storey et al. (2004). For $k \geq 2$, the proof is more technically involved but follows the same spirit. We will prove the desired bound conditional on $\mathcal{S}$, which is defined above as the locations of all false null p-values, then the result follows by integration.

Fix some realization of $\mathcal{S}$. Then by Lemma 3, since the original λ selection rule is LRS, it is true null LRS conditional on $\mathcal{S}$, and we may construct sets $B_1, ..., B_k \subseteq \{0, ..., m_0\}$ such that

$$\begin{aligned} P(\lambda = \lambda_j | \mathcal{S}) &= \sum_{v_1 \in \overline{B}_1} \cdots \sum_{v_{j-1} \in \overline{B}_{j-1}} \sum_{v_j \in B_j} P(V_1 = v_1, ..., V_j = v_j | \mathcal{S}) \\ &= \sum_{v_1 \in \overline{B}_1} \cdots \sum_{v_{j-1} \in \overline{B}_{j-1}} \sum_{v_j \in B_j} P(V_1 = v_1, ..., V_j = v_j). \end{aligned}$$

For the moment, we assume $k \geq 3$. Then

$$\begin{aligned} & E\left[\frac{1-\lambda}{m_0 - V(\lambda) + 1}\frac{V(\lambda_1)}{\lambda_1}\middle|\mathcal{S}\right] \\ = & \sum_{i=1}^{k} E\left[\frac{1-\lambda_i}{m_0 - V(\lambda_i) + 1}\frac{V(\lambda_1)}{\lambda_1}\middle|\lambda = \lambda_i, \mathcal{S}\right] P(\lambda = \lambda_i | \mathcal{S}) \\ = & \sum_{v_1 \in B_1} \frac{1-\lambda_1}{m_0 - v_1 + 1}\frac{v_1}{\lambda_1} P(V_1 = v_1) \\ & + \sum_{v_1 \in \overline{B}_1} \sum_{v_2 \in B_2} \frac{1-\lambda_2}{m_0 - \widetilde{v}_2 + 1}\frac{v_1}{\lambda_1} P(V_1 = v_1, V_2 = v_2) \\ & + \sum_{i=3}^{k} \sum_{v_1 \in \overline{B}_1} \sum_{v_2 \in \overline{B}_2} \cdots \sum_{v_i \in B_i} \frac{1-\lambda_i}{m_0 - \widetilde{v}_i + 1}\frac{v_1}{\lambda_1} P(V_1 = v_1, ..., V_i = v_i) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{v_1 \in B_1 \setminus \{0\}} \frac{1-\lambda_1}{m_0 - v_1 + 1} \frac{v_1}{\lambda_1} P(V_1 = v_1) \\
&\quad + \sum_{v_1 \in \overline{B}_1 \setminus \{0\}} \sum_{v_2 \in B_2} \frac{1-\lambda_2}{m_0 - \widetilde{v}_2 + 1} \frac{v_1}{\lambda_1} P(V_1 = v_1, V_2 = v_2) \\
&\quad + \sum_{i=3}^{k} \sum_{v_1 \in \overline{B}_1 \setminus \{0\}} \sum_{v_2 \in \overline{B}_2} \cdots \sum_{v_i \in B_i} \frac{1-\lambda_i}{m_0 - \widetilde{v}_i + 1} \frac{v_1}{\lambda_1} P(V_1 = v_1, ..., V_i = v_i) \\
&= \sum_{v_1 \in B_1 \setminus \{0\}} P(V_1 = v_1 - 1) + \sum_{v_1 \in \overline{B}_1 \setminus \{0\}} \sum_{v_2 \in B_2} P(V_1 = v_1 - 1, V_2 = v_2) \\
&\quad + \sum_{i=3}^{k} \sum_{v_1 \in \overline{B}_1 \setminus \{0\}} \sum_{v_2 \in \overline{B}_2} \cdots \sum_{v_i \in B_i} P(V_1 = v_1 - 1, ..., V_i = v_i) \\
&\leq 1 - P(V_1 = m_0) \\
&= 1 - \lambda_1^{m_0} \\
&\leq 1.
\end{aligned}
$$

In the second step, we rewrite the condition $\lambda = \lambda_i$ equivalently in terms of the values of V_i's, $i \leq k$. Because the sequential nature of the LRS procedure, the terms in the second step represent a sequential and complete partition of the probability space of the bin counts, $(V_1, \ldots, V_k)$. The fourth step follows easily by expanding the known multinomial probability mass function of $(V_1, ..., V_i)$, and cancelling like terms. Notice that the indexes of V_1 for all terms are effectively shifted down by 1. The third-to-last step is due to the observation that the probability of $V_1 = m_0$ is never tallied in the terms of the summation. The proof when $k = 2$ is identical to the $k \geq 3$ case, removing terms where necessary from the expressions above. □

A.2 Proof of Theorem 2

A.2.1 Outline of Proof of Theorem 2

In order to prove Theorem 2, we need to bound the FDR of the dynamic adaptive procedure with p-grid LRS selection rule λ^* below a constant α. We achieve this by showing the FDR is the limit of a sequence of FDR's of finite grid LRS procedures, in particular the finite approximation rules defined below (Definition 9), applied to the sequence of grids $\{\Lambda_k\}_{k=1}^{\infty}$ (Definition 8).

Definition 8. *Fix $0 < \kappa < \tau < 1$. For a positive integer k, define the λ candidate set* $\Lambda_k = \{\kappa\} \cup ((\kappa, \tau) \cap \{\frac{\ell}{k} : \ell = 1, \ldots, k\}) \cup \{\tau\}$.

Λ_k is defined such that the maximum distance between any two adjacent λ candidate values is at most $1/k$. As $k \to \infty$, the information captured by the bins bounded by Λ_k becomes progressively richer. By restricting to the right boundaries of the non-empty bins of Λ_k, for sufficiently large k we can approximate the elements of the p-grid in arbitrary precision. This idea motivates the following definition:

Definition 9. *For a finite candidate set $\Lambda = \{\lambda_1, ..., \lambda_r\} \subset (0, 1)$, take $\lambda_0 \equiv 0$ and define the bin counts $N_i, i = 1, ..., r$ as in Section 2.1. Define*

$$\Lambda^{(f)} = \{\lambda_j \in \Lambda : N_j > 0\} \cup \{\lambda_r\}.$$

Then the finite approximation to a p-grid LRS rule λ^* with underlying selection rule λ^{LRS} is defined as

$$\lambda^{(f)}(\Lambda) = \lambda^{LRS}(\Lambda^{(f)}).$$

We emphasize that $\lambda^{(f)}$ is still an LRS selection rule on the fixed and finite grid Λ, despite the fact that the intermediate grid $\Lambda^{(f)}$ adapts to the p-value locations by keeping only the right boundaries of the non-empty bins. Knowing all p-value locations, even only approximately, would invalidate the LRS property. It is better to view the construction of $\Lambda^{(f)}$ as a convenient way to simplify the definitions. In practice, there is no need to know all of the grid points in $\Lambda^{(f)}$ during the selection process. Instead, the construction of $\Lambda^{(f)}$ should be thought of as dynamically determined as the selection rule proceeds from κ to τ. Immediately after a new element of $\Lambda^{(f)}$ is added, which happens when arriving at the first candidate in Λ after passing a p-value, the underlying LRS rule is invoked to check if the stopping condition is satisfied. The addition of a new element to $\Lambda^{(f)}$ depends only on N_i, the current bin count of the finer grid Λ. Furthermore, by condition 1, the application of the underlying LRS rule to the currently known grid elements of $\Lambda^{(f)}$ only depends on the past and current bin counts and bin boundaries of Λ.

Thus, more precisely, the proof of Theorem 2 will show that the FDR when λ is selected using λ^* is the limit of the sequence of FDR's when λ is selected using $\{\lambda^{(f)}(\Lambda_k)\}_{k=1}^{\infty}$. For notational simplicity we will denote

$$\lambda^k = \lambda^{(f)}(\Lambda_k).$$

As described in the proof sketch in Section 2.3.1, we need to show the convergence as $k \to \infty$ of the proportion of false discoveries pointwise almost everywhere over the set of

realizations of the continuous null independence model. Next, we define the null set of p-value realizations that we want to exclude from our proof.

Definition 10. *Fix $m \geq 1$, $0 < \kappa < \tau < 1$, $\alpha \in (0,1)$, and $\lambda^* = \lambda^*(p_1, ..., p_m)$ a right continuous p-grid LRS selection rule. Define the following conditions for $i, j = 1, ..., m$ with $i \neq j$:*

(1) $p_i = p_j$
(2) $\hat{\pi}_0^*(p_i) = \hat{\pi}_0^*(p_j)$
(3) $\lim_{t \to p_i^+} \widehat{\mathrm{FDR}}^*_{\lambda^*}(t) = \alpha$
(4) $\lim_{t \to p_i^-} \widehat{\mathrm{FDR}}^*_{\lambda^*}(t) = \alpha$
(5) $p_i = \kappa$
(6) *The right continuity of λ^* does not hold for $(p_1, ..., p_m)$.*

Let $\Omega_0 := \Omega_0(m, \kappa, \tau, \alpha, \lambda^*) = \{(p_1, ..., p_m) \in (0,1)^m : (p_1, ..., p_m)$ satisfies one or more of (1) to (6).

Ω_0 as defined above contains all the realizations for which pointwise convergence is either difficult or impossible to prove, and in Lemma 4 we show it comprises a null set with respect to the measure induced by the continuous null independence model. Note that Ω_0 depends on m, α, κ, τ and λ^*, all of which we treat as fixed for the remainder of this section. Furthermore, note that right continuity as defined in Condtion 2 is a pointwise property that may not hold on a null set, hence the inclusion of (6) above.

The proof of pointwise convergence of false discovery proportion proceeds in several steps. For a fixed realization outside of Ω_0. Lemmas 5 and 6 show that the tuning parameter λ^k which is selected using the finite approximation procedure with grid Λ_k converges to the tuning parameter selected by the p-grid LRS procedure λ^*. Lemma 7 extends this convergence to the π_0-estimator evaluated at these tuning parameters, and Lemmas 8 and 9 extend it finally to the proportion of false discoveries below the rejection threshold produced by these procedures, the expectation of which is the FDR of the procedure.

Finally, we apply the bounded convergence theorem to extend this pointwise convergence to convergence in expectation to complete the proof of Theorem 2.

A.2.2 Lemmas for Theorem 2

Lemma 4. *Under the continuous null independence model,* $P(\Omega_0) = 0$.

Proof. We write $\Omega_0 = \Omega_1 \cup \cdots \cup \Omega_6$ where

$$\Omega_j = \{(p_1, ..., p_m) : (p_1, ..., p_m) \text{ satisfies } (j)\}.$$

First consider Ω_1. Note that

$$\begin{aligned} P(\Omega_1) &= P(\cup_{i<j}\{(p_1, ..., p_m) : p_i = p_j\}) \\ &\leq \sum_{i<j} P(\{(p_1, ..., p_m) : p_i = p_j\}). \end{aligned}$$

Then for any fixed $i_0 < j_0$, by the continuity condition on the p-values,

$$\begin{aligned} & P(\{(p_1, ..., p_m) : p_{i_0} = p_{j_0}\}) \\ = \; & E(I(p_{i_0} = p_{j_0})) \\ = \; & E(E(I(p_{i_0} = p_{j_0})|p_1, ..., p_{i_0-1}, p_{i_0+1}, ..., p_{j_0}, ..., p_m)) \\ = \; & 0 \end{aligned}$$

since by continuity, the inner expectation is zero for every fixed realization of the p-values without p_{i_0}, and there are countably many pairs, it follows $P(\Omega_1) = 0$.

Now consider Ω_2. Similarly, we have

$$\begin{aligned} & P(\Omega_2) \\ \leq \; & \sum_{i<j} P(\hat{\pi}_0^*(p_i) = \hat{\pi}_0^*(p_j)) \\ = \; & \sum_{i<j} P(\frac{m - R(p_i) + 1}{m(1 - p_i)} = \frac{m - R(p_j) + 1}{m(1 - p_j)}) \\ \leq \; & \sum_{i<j} \sum_{\ell_1=0}^{m-1} \sum_{\ell_2=0}^{m-1} P(\frac{m - \ell_1}{m(1 - p_i)} = \frac{m - \ell_2}{m(1 - p_j)}) \\ \leq \; & \sum_{i<j} \sum_{\ell_1=0}^{m-1} \sum_{\ell_2=0}^{m-1} P(p_i = \frac{m - \ell_1}{m - \ell_2} m(1 - p_j)). \end{aligned}$$

Then for fixed $i_0 < j_0$, ℓ_1 and ℓ_2, we have

$$\begin{aligned}
& P(p_{i_0} = \frac{m-\ell_1}{m-\ell_2} m(1-p_{j_0})) \\
= \;& E(I(p_{i_0} = \frac{m-\ell_1}{m-\ell_2} m(1-p_{j_0}))) \\
= \;& E(E(I(p_{i_0} = \frac{m-\ell_1}{m-\ell_2} m(1-p_{j_0}))|p_1, ..., p_{i_0-1}, p_{i_0+1}, ..., p_{j_0}, ..., p_m)) \\
= \;& 0
\end{aligned}$$

since again by continuity, the inner expectation is zero. It follows that $P(\Omega_2) = 0$. Now consider Ω_3. Note that since λ^* is a p-grid LRS selection rule, it will take values in the finite set $\{p_1, ..., p_m, \tau\}$. Thus

$$\begin{aligned}
& P(\Omega_3) \\
\leq \;& \sum_{i=1}^{m} P(\lim_{t \to p_i^+} \widehat{\mathrm{FDR}}^*_{\lambda^*}(t) = \alpha) \\
= \;& \sum_{i=1}^{m} P(\frac{m\hat{\pi}_0^*(\lambda^*) p_i}{R(p_i)} = \alpha) \\
= \;& \sum_{i=1}^{m} P(\frac{mp_i}{R(p_i)} \frac{m - R(\lambda^*) + 1}{m(1-\lambda^*)} = \alpha) \\
\leq \;& \sum_{i=1}^{m} \sum_{j=1}^{m} P(\frac{p_i}{R(p_i)} \frac{m - R(p_j) + 1}{1 - p_j} = \alpha) + \sum_{i=1}^{m} P(\frac{p_i}{R(p_i)} \frac{m - R(\tau) + 1}{1 - \tau} = \alpha) \\
\leq \;& \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{\ell_1=1}^{m} \sum_{\ell_2=0}^{m-1} P(\frac{p_i(m-\ell_2)}{\ell_1(1-p_j)} = \alpha) + \sum_{i=1}^{m} \sum_{\ell_1=1}^{m} \sum_{\ell_2=0}^{m-1} P(\frac{p_i(m-\ell_2)}{\ell_1(1-\tau)} = \alpha) \\
\leq \;& \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{\ell_1=1}^{m} \sum_{\ell_2=0}^{m-1} P(p_i = \frac{\alpha \ell_1 (1-p_j)}{m-\ell_2}) + \sum_{i=1}^{m} \sum_{\ell_1=1}^{m} \sum_{\ell_2=0}^{m-1} P(p_i = \frac{\alpha \ell_1 (1-\tau)}{m-\ell_2}).
\end{aligned}$$

Then for fixed $i_0 < j_0$, ℓ_1 and ℓ_2, we have

$$\begin{aligned} & P(p_{i_0} = \frac{\alpha\ell_1(1-p_{j_0})}{m-\ell_2}) \\ = \; & E(I(p_{i_0} = \frac{\alpha\ell_1(1-p_{j_0})}{m-\ell_2})) \\ = \; & E(E(I(p_{i_0} = \frac{\alpha\ell_1(1-p_{j_0})}{m-\ell_2})|p_1, ..., p_{i_0-1}, p_{i_0+1}, ..., p_{j_0}, ..., p_m)) \\ = \; & 0 \end{aligned}$$

again by continuity. A similar argument holds for the second term, replacing p_{j_0} by τ. Thus $P(\Omega_3) = 0$.

A similar argument can be used to bound $P(\Omega_4)$, and $P(\Omega_5) = 0$ is clear by continuity. $P(\Omega_6) = 0$ by definition of the right continuity condition. We conclude that $P(\Omega_0) = 0$. □

Lemma 5. *Let $\omega = (p_1, ..., p_m) \in \Omega_0^C$ be a fixed realization of the continuous null independence model. Then there exists $K \in \mathbb{N}$ such that for all $k \geq K$, each bin bounded by selection grid Λ_k contains at most one p-value.*

Proof. The lemma is vacuously true if $m = 1$, so we may assume $m \geq 2$. Since $\omega \in \Omega_0^C$, by Lemma 4 we have that $p_i \neq p_j$ for all $i, j = 1, ..., m, i \neq j$. Thus we may define the quantity

$$\epsilon_p = \min\{p_{(i)} - p_{(i-1)} : 2 \leq i \leq m\} > 0.$$

Let K be sufficiently large that $\frac{1}{K} < \epsilon_p$, then it is straightforward to see that, for all $k \geq K$, the bin width of Λ_k is at most $1/k < \epsilon_p$, and there cannot be two p-values in the same bin. □

Lemma 6. *For any realization of the null independence model $\omega = (p_1, ..., p_m) \in \Omega_0^C$, $\lambda^k \to \lambda^*$ as $k \to \infty$.*

Proof. Fix $\epsilon > 0$. Since $\omega \in \Omega_0^C$, by Lemma 5 there exists $K_p \in \mathbb{N}$ such that for all $k \geq K_p$, the bins bounded by Λ_k contain at most one p-value. Recall from Definition 3 that

$$\Lambda^{(p)} := (\{p_{(1)}, ..., p_{(m)}\} \cap (\kappa, \tau)) \cup \{\tau\}.$$

Since λ^* is right continuous for this realization, there exists $\delta > 0$ such that

$$|\lambda^{\text{LRS}}(\Lambda^{(p)}) - \lambda^{\text{LRS}}(\Lambda')| < \epsilon$$

for every grid $\Lambda' = \{\lambda'_1, ..., \lambda'_{k^{(p)}}\}$ of length $k^{(p)} := |\Lambda^{(p)}|$ such that $\lambda'_j \in [\lambda_j^{(p)}, \lambda_j^{(p)} + \delta)$ for every $1 \leq j \leq k^{(p)}$. Let K be sufficiently large that $K \geq K_p$, and $\frac{1}{K} < \delta$. Then for all $k \geq K$, since $k \geq K_p$, the finite grid $\Lambda_k^{(f)}$ is the same size as $\Lambda^{(p)}$. Denote this length by

$$L = |\Lambda^{(p)}| = |\Lambda_k^{(f)}|.$$

In addition, since $\frac{1}{k} < \delta$, the bin width of Λ_k is at most $\frac{1}{k} < \delta$, and thus for all $1 \leq j \leq L$,

$$[\Lambda_k^{(f)}]_j \in [\lambda_j^{(p)}, \lambda_j^{(p)} + \delta)$$

and so by the construction of δ, it follows that

$$|\lambda^{\text{LRS}}(\Lambda^{(p)}) - \lambda^{\text{LRS}}(\Lambda_k^{(f)})| < \epsilon$$

for all $k \geq K$. By definition of λ^* and λ^k, it follows that

$$|\lambda^* - \lambda^k| < \epsilon$$

for all $k \geq K$. □

Lemma 7. *Let $\omega = (p_1, ..., p_m) \in \Omega_0^C$ be a fixed realization of the continuous null independence model. Then $\hat{\pi}_0^*(\lambda^k) \to \hat{\pi}_0^*(\lambda^*)$ as $k \to \infty$.*

Proof. Recall that

$$\hat{\pi}_0^*(\lambda) = \frac{m - R(\lambda) + 1}{m(1 - \lambda)}.$$

Following the proof of Lemma 6, λ^k approaches λ^* from above, and thus there exists K such that for all $k \geq K$, λ^k is the right boundary of the bin containing λ^*, so that $R(\lambda^k) = R(\lambda^*)$.

The proof follows by noting that $\hat{\pi}_0^*(\lambda)$ is a continuous function of λ and $\lambda^k \to \lambda^*$ as $k \to \infty$ due to Lemma 6. □

Define $\widehat{\text{FDR}}^*_{\lambda^*}$ to be the estimated FDR function specifying $\lambda = \lambda^*$, and $\widehat{\text{FDR}}^*_{\lambda^k}$ to be the estimated FDR function specifying $\lambda = \lambda^k$. Define for both functions

$$\widehat{\text{FDR}}^*_{\cdot}(t_0^-) = \lim_{t \to t_0^-} \widehat{\text{FDR}}^*_{\cdot}(t)$$

and

$$\widehat{\text{FDR}}^*_{\cdot}(t_0^+) = \lim_{t \to t_0^+} \widehat{\text{FDR}}^*_{\cdot}(t).$$

Define n_1 as the ordered index of the smallest p-value in the interval (κ, τ). Then we have the following lemma.

Lemma 8. *Let $\omega = (p_1, ..., p_m) \in \Omega_0^C$ be a fixed realization of the continuous null independence model such that at least 1 p-value lies in the interval $(0, \kappa)$. Then there exists $K \in \mathbb{N}$ such that for all $k \geq K$, and $i = 1, ..., n_1 - 1$, $\widehat{\text{FDR}}^*_{\lambda^*}(p^{\pm}_{(i)}) < \alpha$ if and only if $\widehat{\text{FDR}}^*_{\lambda^k}(p^{\pm}_{(i)}) < \alpha$.*

Proof. Since $\omega \in \Omega_0^C$, we may define the quantities

$$\gamma_- = \min\{|\widehat{\text{FDR}}^*_{\lambda^*}(p^-_{(i)}) - \alpha| : i = 1, ..., n_1 - 1\} > 0$$

and

$$\gamma_+ = \min\{|\widehat{\text{FDR}}^*_{\lambda^*}(p^+_{(i)}) - \alpha| : i = 1, ..., n_1 - 1\} > 0.$$

Define $\gamma_F = \min\{\gamma_-, \gamma_+\}$.

For each $i = 1, ..., n_1 - 1$, define the continuous functions

$$g_i(x) = \frac{mp_{(i)}}{i}x$$

and

$$h_i(x) = \frac{mp_{(i)}}{(i-1) \vee 1}x.$$

By Lemma 7 and continuity of g_i and h_i, there exists K_i such that for all $k \geq K_i$,

$$|g_i(\hat{\pi}^*_0(\lambda^k)) - g_i(\hat{\pi}^*_0(\lambda^*))| < \gamma_F,$$

and

$$|h_i(\hat{\pi}^*_0(\lambda^k)) - h_i(\hat{\pi}^*_0(\lambda^*))| < \gamma_F.$$

Let $K = \max\{K_1, ..., K_{n_1-1}\}$, then it follows that K is as required. □

Lemma 9. *Let $\omega = (p_1, ..., p_m) \in \Omega_0^C$ be a fixed realization of the continuous null independence model. Then there exists $K \in \mathbb{N}$ such that for all $k \geq K$,*

$$\frac{V(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k}))}{R(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k})) \vee 1} = \frac{V(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^*}))}{R(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^*})) \vee 1}.$$

Proof. If there are no p-values in the interval $(0, \kappa)$, then V and R will always count 0 p-values, regardless of the selection of λ, and the result holds trivially for $K = 1$. Thus we may assume there is at least 1 p-value in the interval $(0, \kappa)$. Define n_1 as the ordered

index of the smallest p-value in the interval (κ, τ). By Lemma 8, there exists K such that for all $k \geq K$ and $i = 1, ..., n_1 - 1$,

$$\widehat{\text{FDR}}^*_{\lambda^*}(p^{\pm}_{(i)}) < \alpha$$

if and only if

$$\widehat{\text{FDR}}^*_{\lambda^k}(p^{\pm}_{(i)}) < \alpha.$$

Intuitively, for $k \geq K$, Lemma 8 implies that the thresholds $t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k})$ and $t_\alpha(\widehat{\text{FDR}}^*_{\lambda^*})$ lead to the same set of rejections and furthermore the same false discovery proportion. More specifically, consider $k \geq K$. Define

$$I_* = \{i : 1 \leq i \leq n_1 - 1, \widehat{\text{FDR}}^*_{\lambda^*}(p^+_{(i-1)}) < \alpha < \widehat{\text{FDR}}^*_{\lambda^*}(p^-_{(i)})\}$$

and

$$I_k = \{i : 1 \leq i \leq n_1 - 1, \widehat{\text{FDR}}^*_{\lambda^k}(p^+_{(i-1)}) < \alpha < \widehat{\text{FDR}}^*_{\lambda^k}(p^-_{(i)})\}$$

where we take $p^+_{(0)} \equiv 0$. By the selection of $k \geq K$, it follows that $I_* = I_k$. If $I_* = I_k = \emptyset$, then it follows that both $t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k})$ and $t_\alpha(\widehat{\text{FDR}}^*_{\lambda^*})$ are within the interval $(p_{(n_1-1)}, \kappa]$, and we reject the smallest $n_1 - 1$ p-values using both procedures. If $I_* = I_k \neq \emptyset$, then it follows that both $t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k})$ and $t_\alpha(\widehat{\text{FDR}}^*_{\lambda^*})$ are within the interval $(p_{(i_m-1)}, p_{(i_m)}]$, where $i_m = \max\{i : i \in I_*\} = \max\{i : i \in I_k\}$. This implies both R and V count the same number of total and null p-values respectively at either threshold. In summary, in both cases, we have

$$\frac{V(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k}))}{R(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k})) \vee 1} = \frac{V(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^*}))}{R(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^*})) \vee 1}$$

for all $k \geq K$. □

A.2.3 Proof of the theorem

Proof. Recall $\widehat{\text{FDR}}^*_{\lambda^k}$ as defined in Section 2.3.1. Since the tuning parameter for this estimate is selected using an LRS rule, by Theorem 1,

$$\text{FDR}(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k})) \leq \alpha$$

for all k. Notice that for all k and for any realization $\omega \in (0, 1)^m$ of the continuous null independence model, we have

$$\frac{V(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k}))}{R(t_\alpha(\widehat{\text{FDR}}^*_{\lambda^k})) \vee 1} \in [0, 1],$$

such that the sequence of random variables

$$\left\{ \frac{V(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^k}))}{R(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^k})) \vee 1} \right\}_{k=1}^{\infty}$$

is uniformly bounded. By Lemma 9, this sequence converges pointwise on Ω_0^C, and hence pointwise almost everywhere by Lemma 4, to the random variable

$$\frac{V(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^*}))}{R(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^*})) \vee 1}.$$

Then,

$$\begin{aligned}
\mathrm{FDR}(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^*})) &= E\left[\frac{V(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^*}))}{R(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^*})) \vee 1} \right] \\
&= E\left[\lim_{k\to\infty} \frac{V(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^k}))}{R(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^k})) \vee 1} \right] \\
&= \lim_{k\to\infty} E\left[\frac{V(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^k}))}{R(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^k})) \vee 1} \right] \\
&= \lim_{k\to\infty} \mathrm{FDR}(t_\alpha(\widehat{\mathrm{FDR}}^*_{\lambda^k})) \\
&\leq \alpha.
\end{aligned}$$

The third equality is due to the bounded convergence theorem, and the final inequality follows by properties of real sequences. □

A.3 Modified lowest-slope procedure

In this section the finite sample control of the modified lowest-slope procedure (Corollary 1) is justified by showing that the right boundary procedure is right continuous. Consider the following definition, which is similar to Definition 10.

Definition 11. *Fix $m \geq 1$, and $0 < \kappa < \tau < 1$. Define the following conditions for*

$i, j = 1, ..., m$ *with* $i \neq j$*:*

$$\begin{array}{ll} (1) & p_i = p_j \\ (2) & \hat{\pi}_0^*(p_i) = \hat{\pi}_0^*(p_j) \\ (3) & p_i \in \{\kappa, \tau\}. \end{array}$$

Let $B_0 := B_0(m, \kappa, \tau) = \{(p_1, ..., p_m) \in (0,1)^m : (p_1, ..., p_m)$ *satisfies one or more of (1),(2) and (3)*$\}$.

As it is a subset of Ω_0, by Lemma 4, B_0 is a null set under the continuous null independence model.

Lemma 10. *Fix* $0 < \kappa < \tau < 1$. *Under the continuous null independence model, the right-boundary procedure is right continuous.*

Proof. By definition we may exclude a null set of realizations, thus consider some fixed $\omega = (p_1, ..., p_m) \in B_0^C$. Fix $\epsilon > 0$. Denote $\Lambda^{(p)} = \{\lambda_1^{(p)}, ..., \lambda_{k^{(p)}}^{(p)}\}$ so that it has length $k^{(p)}$. Since $\omega \in B_0^C$, we have that $\hat{\pi}_0^*(p_i) \neq \hat{\pi}_0^*(p_j)$ for all $i, j = 1, ..., m, i \neq j$. Thus we may define the quantity

$$\gamma_\pi = \frac{1}{2} \min\{|\hat{\pi}_0^*(p_{(i)}) - \hat{\pi}_0^*(p_{(i-1)})| : 2 \leq i \leq m\} > 0.$$

For $j = 1, ..., k^{(p)}$, define the real function

$$f_j(x) = \frac{m - R(\lambda_j^{(p)}) + 1}{m(1-x)}.$$

Each f_j is a continuous function for $x \in (0,1)$. For each $j = 1, ..., k^{(p)}$ let $\epsilon_j > 0$ be sufficiently small that for all x with $|x - \lambda_j^{(p)}| < \delta_j$,

$$|f_j(x) - f_j(\lambda_j^{(p)})| = \left| \frac{m - R(\lambda_j^{(p)}) + 1}{m(1-x)} - \frac{m - R(\lambda_j^{(p)}) + 1}{m(1-\lambda_j^{(p)})} \right| < \gamma_\pi,$$

then define $\delta = \min\{\delta_1, ..., \delta_{k^{(p)}}\} > 0$. In addition, by Lemma 5 and since $\omega \in B_0^C$, choose δ sufficiently small that $[p_i, p_i + \delta)$ contains only one p-value for all $i = 1, ..., m$, and $[\tau, \tau + \delta)$ contains zero p-values. Finally choose δ sufficiently small that $\delta < \epsilon$.

Suppose we apply the right boundary procedure λ^{RBP} to a grid Λ' of length $k^{(p)}$, where $\lambda'_j \in [\lambda_j^{(p)}, \lambda_j^{(p)} + \delta)$ for all $j = 1, ..., k^{(p)}$. Note that since they are right boundaries and by the selection of δ,

$$\lambda_{j+1}^{(p)} > \lambda'_j \geq \lambda_j^{(p)}$$

for $j = 1, ..., k^{(p)} - 1$, and

$$R(\lambda'_{k^{(p)}}) = R(\tau) = R(\lambda_{k^{(p)}}^{(p)}),$$

which says that we must have $R(\lambda'_j) = R(\lambda_j^{(p)})$ for all $j = 1, ..., k^{(p)}$. Also note that $|\lambda'_j - \lambda_j^{(p)}| < \delta$ for all $j = 1, ..., k^{(p)}$. Combining these observations, we have

$$\hat{\pi}_0^*(\lambda'_j) \in (\hat{\pi}_0^*(\lambda_j^{(p)}) - \gamma_\pi, \hat{\pi}_0^*(\lambda_j^{(p)}) + \gamma_\pi).$$

The definition of γ_π then implies that $\pi_0^*(\lambda'_j)$ and $\pi_0^*(\lambda_j^{(p)})$ are sufficiently close to each other that comparisons of the estimator evaluated at consecutive grid points of $\Lambda^{(p)}$ will produce the same outcome as comparisons of the estimator evaluated at consecutive grid points of Λ', that is the sequential decisions of the right boundary procedure will be the same on both grids. Noting that $|\lambda'_j - \lambda_j^{(p)}| < \epsilon$ for all $j = 1, ..., k^{(p)}$, it follows that

$$|\lambda^{\text{RBP}}(\Lambda^{(p)}) - \lambda^{\text{RBP}}(\Lambda')| < \epsilon,$$

and thus the right boundary procedure is right continuous. □