False Discovery Rate Adjustments for Average Significance Level Controlling Tests

Timothy B. Armstrong*
University of Southern California

May 15, 2025

Abstract

Multiple testing adjustments, such as the Benjamini and Hochberg (1995) step-up procedure for controlling the false discovery rate (FDR), are typically applied to families of tests that control significance level in the classical sense: for each individual test, the probability of false rejection is no greater than the nominal level. In this paper, we consider tests that satisfy only a weaker notion of significance level control, in which the probability of false rejection need only be controlled on average over the hypotheses. We find that the Benjamini and Hochberg (1995) step-up procedure still controls FDR in the asymptotic regime with many weakly dependent p-values, and that certain adjustments for dependent p-values such as the Benjamini and Yekutieli (2001) procedure continue to yield FDR control in finite samples. Our results open the door to FDR controlling procedures in nonparametric and high dimensional settings where weakening the notion of inference allows for large power improvements.

1 Introduction

Consider testing m hypotheses $H_1, \dots, H_m$. Let $\mathcal{H}_0 \subseteq \{1, \dots, m\}$ denote the set of true null hypotheses. Given p-values $p_1, \dots, p_m$ for each of the hypotheses, we wish to form a multiple testing procedure which decides on a subset of hypotheses to reject. A common starting point for multiple testing procedures proposed in the literature is to assume that

*email: timothy.armstrong@usc.edu. Support from National Science Foundation Grant SES-2049765 is gratefully acknowledged.

the p-values are formed from tests that control significance level in the classical sense, which implies

$$\text{for all } i \in \mathcal{H}_0,\ P(p_i \leq t) \leq t. \tag{1}$$

One then adjusts the critical value so that some notion of multiple testing error, such as the false discovery rate (FDR), is controlled (see formal definitions below).

In this paper, we explore the possibility of forming FDR controlling multiple testing procedures from tests that satisfy a weaker *average significance level* control criterion. Such tests can be formed from confidence intervals (CIs) that weaken the classical definition of a CI by requiring coverage only on average over the reported CIs. Such CIs have been developed in a number of settings (Wahba, 1983; Nychka, 1988; Wasserman, 2007, Chapter 5.8; Cai et al., 2014; Armstrong et al., 2022), and are particularly appealing in high dimensional or nonparametric settings involving regularized estimation, where impossibility results (Low, 1997) severely restrict the scope for constructing classical tests and CIs. We ask: can such CIs still be used as an input to multiple testing procedures proposed in the literature, despite only satisfying a weaker notion of coverage?

We focus on multiple testing procedures designed to control the false discovery rate (FDR) of Benjamini and Hochberg (1995). We find that average significance level control is indeed sufficient to use a CI as an input to certain multiple testing procedures that guarantee FDR control. In particular, average significance level control is sufficient to guarantee FDR control of the Benjamini and Hochberg (1995) procedure in the asymptotic regime of weakly dependent p-values and many hypotheses ($m \to \infty$) and of the Benjamini and Yekutieli (2001) procedure with fixed m and arbitrary dependence among p-values. On the other hand, in contrast to the classical setting, we show by example that the Benjamini and Hochberg (1995) procedure does not in general have FDR control with fixed m and independent p-values, and that approaches that estimate the proportion of null hypotheses, such as the procedure of Storey (2002), can fail to control FDR even as $m \to \infty$.

Much of the literature on FDR controlling multiple testing procedures takes a family of p-values satisfying the classical significance level control condition (1) as a starting point. An important exception is the literature on knockoff based FDR controlling procedures (Barber and Candès, 2015), which instead rely on the construction of auxiliary random variables, called knockoffs. Constructing knockoffs typically requires modeling assumptions such as the "model-X" framework, in which the joint distribution of regression covariates is known or estimated with sufficient accuracy (Candès et al., 2018), or restricting the procedure to low dimensional settings. We view our results as complementary to this literature: our results allow for FDR controlling procedures based on average coverage intervals, which are available in nonparametric and high dimensional settings where the model-X framework is difficult to apply. Our results also complement the recent literature on covariate assisted FDR controlling procedures (see, among others, Lei and Fithian, 2018; Li and Barber, 2019; Ignatiadis and Huber, 2021): whereas this literature seeks to improve power using covariates X_i or other information associated with the p-value p_i, our results allow for more powerful FDR controlling procedures by directly basing such procedures on tests that achieve greater power than conventional tests by weakening the notion of significance level control.

The rest of this paper is organized as follows. Section 2 introduces the setup and provides an overview of results. Section 3 presents finite sample results. Section 4 presents results that are asymptotic in the number m of hypotheses being tested.

2 Setup and Overview of Results

To describe our results, consider intervals $CI_1(t), \ldots, CI_m(t)$ for parameters $\theta_1, \ldots, \theta_m$ with nominal coverage level $1-t$. Let $p_1, \ldots, p_m$ be p-values formed from testing hypotheses $H_i : \theta_i = \theta_{0,i}$ using these CIs: $p_i \leq t$ iff. $\theta_{0,i} \notin CI_i(t)$. The classical definition of a $100 \cdot (1-t)\%$ CI states that $P(\theta_i \notin CI(t)) \leq t$ for each i, which leads to p-values satisfying the classical significance level control condition (1). We are interested in p-values formed from CIs that satisfy the weaker *average coverage condition*

$$\frac{1}{m}\sum_{i=1}^{m} P(\theta_i \notin CI_i(t)) \leq t. \tag{2}$$

Papers that propose CIs that satisfy the average coverage condition (2) or related criteria (and which may not satisfy the classical CI coverage condition for each i) include Armstrong et al. (2022); Cai et al. (2014); Nychka (1988); Wahba (1983) and Wasserman (2007, Chapter 5.8)). Letting p_i be p-values formed from these CIs and letting $\mathcal{H}_0$ denote the set of indices i where the null hypothesis $H_i : \theta_i = \theta_{0,i}$ holds, we have $p_i \leq t$ iff. $\theta_{0,i} \notin CI_i(t)$ and $\frac{1}{m}\sum_{i\in\mathcal{H}_0} P(\theta_{0,i} \notin CI_i(t)) = \frac{1}{m}\sum_{i\in\mathcal{H}_0} P(\theta_i \notin CI_i(t)) \leq \frac{1}{m}\sum_{i=1}^{m} P(\theta_i \notin CI_i(t)) \leq t$. The p-values will therefore satisfy

$$\frac{1}{m}\sum_{i\in\mathcal{H}_0} P(p_i \leq t) \leq t. \tag{3}$$

We will refer to a family of p-values and their associated tests as having average significance level control (at level t) for the testing problem $(P, \mathcal{H}_0)$ when condition (3) holds. Note that the sum of false rejection probabilities in (3) is scaled by the total number of hypotheses m, whereas the classical condition (1) would allow one to replace m with $\#\mathcal{H}_0$. This is due to the fact that the average coverage criterion (2) only guarantees error bounds on average over all m CIs, and not over the subset for which some particular null hypothesis holds.

A multiple testing procedure is a function that maps the p-values $p_1, \ldots, p_m$ to a subset $\mathcal{R} = \mathcal{R}(p_1, \ldots, p_m) \subseteq \{1, \ldots, m\}$ of rejected null hypotheses. The false discovery proportion (FDP) of a procedure $\mathcal{R}$ is:

$$\mathrm{FDP}(\mathcal{R}, \mathcal{H}_0) = \frac{\#\left(\mathcal{R} \cap \mathcal{H}_0\right)}{\#\mathcal{R} \vee 1} \tag{4}$$

where $\#\mathcal{A}$ is the cardinality of $\mathcal{A}$ and $a \vee b$ denotes the maximum of a and b. The false discovery rate (FDR) of this procedure is the expectation of the FDP:

$$\mathrm{FDR}(\mathcal{R}, \mathcal{H}_0, P) = E_P\, \mathrm{FDP}(\mathcal{R}, \mathcal{H}_0) = E_P\left[\frac{\#\left(\mathcal{R} \cap \mathcal{H}_0\right)}{\#\mathcal{R} \vee 1}\right] \tag{5}$$

where E_P denotes expectation under the distribution P of the p-values. We say that $\mathcal{R}$ controls the false discovery rate at level q if $\mathrm{FDR}(\mathcal{R}, \mathcal{H}_0, P) \leq q$.

While some of our results are more general, our main focus is on the Benjamini and Hochberg (1995, BH) step-up procedure, and generalizations such as those considered by Benjamini and Yekutieli (2001), Storey (2002) and Blanchard and Roquain (2008). To describe these procedures, let

$$\mathcal{R}_t^{\text{fixed}}(p_1, \ldots, p_n) = \{i : p_i \leq t\}\,. \tag{6}$$

denote the fixed rejection region procedure with cutoff t. That is, we reject all hypotheses with p-value less than t. Let

$$V(t) = \sum_{i \in \mathcal{H}_0} I(p_i \leq t) = \#\left(\mathcal{R}_t^{\text{fixed}} \cap \mathcal{H}_0\right), \quad S(t) = \sum_{i \notin \mathcal{H}_0} I(p_i \leq t) = \#\left(\mathcal{R}_t^{\text{fixed}} \backslash \mathcal{H}_0\right)$$
$$\text{and } R(t) = V(t) + S(t) = \#\mathcal{R}_t^{\text{fixed}}. \tag{7}$$

The FDP of $\mathcal{R}_t^{\text{fixed}}$ is given by $V(t)/[R(t) \vee 1]$. The BH procedure can be motivated by noting that, while $V(t)$ cannot be observed, one can form a conservative estimate by replacing it with $m \cdot t$. This gives an estimate of the fixed rejection region FDR:

$$\widehat{\mathrm{FDR}}(t) = \frac{m \cdot t}{\#\mathcal{R}_t^{\mathrm{fixed}} \vee 1} = \frac{m \cdot t}{R(t) \vee 1}. \tag{8}$$

The BH procedure at nominal FDR level q uses a cutoff $\hat{t}_{\mathrm{BH},q}$ based on this estimate:

$$\mathcal{R}_{\mathrm{BH},q}(p_1, \ldots, p_m) = \{i : p_i \leq \hat{t}_{\mathrm{BH},q}\} \quad \text{where} \quad \hat{t}_{\mathrm{BH},q} = \max\{t : \widehat{\mathrm{FDR}}(t) \leq q\}. \tag{9}$$

A more general class of step-up procedures can be formed by using an estimate of the form $\pi m t$ for $V(t)$ and modifying the denominator using a nondecreasing function β, called a shape function:

$$\mathcal{R}_{\pi,\beta(\cdot),q}(p_1, \ldots, p_m) = \{i : p_i \leq \hat{t}_{\pi,\beta(\cdot),q}\} \quad \text{where} \quad \hat{t}_{\pi,\beta(\cdot),q} = \max\left\{t : \frac{\pi m t}{\beta(R(t))} \leq q\right\}. \tag{10}$$

Such procedures have been considered by, among others, Benjamini and Yekutieli (2001), Storey (2002) and Blanchard and Roquain (2008).

When the p-values satisfy the classical significance level control condition (1), these procedures are known to have the following properties.

(i) The BH procedure controls FDR when p-values are independent (Benjamini and Hochberg, 1995).

(ii) The estimate $\widehat{\mathrm{FDR}}(t)$ is upwardly biased for the FDR of the fixed rejection region procedure $\mathcal{R}_t^{\mathrm{fixed}}$ when p-values are independent (Storey et al., 2004).

(iii) The procedure $\mathcal{R}_{1,\beta(\cdot),q}$ (with $\pi = 1$) controls FDR under arbitrary dependence for the shape function $\beta(k) = k\left(\sum_{i=1}^m i^{-1}\right)^{-1}$ (Benjamini and Yekutieli, 2001) and, more generally, when $\beta(k) = \int_0^k x d\nu(x)$ for an arbitrary probability distribution ν on $(0, \infty)$ (Blanchard and Roquain, 2008).

(iv) The BH procedure controls FDR asymptotically (as $m \to \infty$) when the p-values satisfy a weak dependence condition (Storey et al., 2004; Genovese and Wasserman, 2004).

(v) The procedure $\mathcal{R}_{\hat{\pi},\beta(\cdot),q}$, where $\hat{\pi} = \sum_{i=1}^m I(p_i > \lambda)$ is an estimate of $\#\mathcal{H}_0/m$, controls FDR (a) under fixed m with independent p values (Storey et al., 2004) and (b) asymptotically as $m \to \infty$ when the p-values satisfy a weak dependence condition (Storey et al., 2004; Genovese and Wasserman, 2004).

Our results can be summarized as showing that, when the p-values only satisfy the weaker average significance level control condition (3), properties (ii), (iii) and (iv) continue to hold, but that properties (i) and (v)(a) and (v)(b) in general do not. Section 3.1 shows property (iii) and provides a counterexample to property (i). Section 3.2 shows property (ii). Section 4 shows property (iv).

3 Finite Sample Results

This section considers finite sample control of FDR for step-up procedures (Section 3.1) and point estimation of FDR of the fixed rejection region procedure $\mathcal{R}_t^{\text{fixed}}$ (Section 3.2).

3.1 FDR Control

Our results on FDR control for step-up procedures follows a corollary of a more general result that uses an invariance assumption on an oracle version of a multiple testing procedure. The basic idea is that, if the p-values satisfy the average significance level control condition (3), then one can form another multiple testing problem in which the classical condition (1) holds by randomly permuting the p-values of the true null hypotheses and multiplying them by $m/\#\mathcal{H}_0$. One can then apply results from the literature to this new setting.

To state our result, we explicitly introduce notation $\mathcal{R}(p_1, \dots, p_m; \mathcal{H}_0)$ for oracle procedures that depend on the set of true null hypotheses $\mathcal{H}_0$ (typically through the cardinality $\#\mathcal{H}_0$ of this set). We use a permutation invariance condition

$$i \in \mathcal{R}(p_1, \dots, p_m) \quad \text{iff.} \quad \sigma(i) \in \mathcal{R}(p_{\sigma(1)}, \dots, p_{\sigma(m)}) \tag{11}$$

for any permutation σ of the indices $1, \dots, m$ of the tests. This includes the class of step-up procedures (10), so long as π is either a fixed number or a permutation invariant function of the p-values.

Theorem 3.1. *Let $\mathcal{R}$ be a multiple testing procedure that satisfies the permutation invariance condition (11), and suppose that the oracle procedure $\widetilde{\mathcal{R}}(p_1, \dots, p_m; \mathcal{H}_0) = \mathcal{R}(p_1(m_0/m), \dots, p_m(m_0/m))$ (where $m_0 = \#\mathcal{H}_0$) controls FDR at level q for any $(P, \mathcal{H}_0)$ satisfying the classical significance level control condition (1). Then $\mathcal{R}$ controls FDR at level q for any $(P, \mathcal{H}_0)$ such that the average significance level control condition (3) holds.*

Proof. Given $(P, \mathcal{H}_0)$ such that (3) holds and $p_1, \dots, p_n$ drawn from P, define $\tilde{p}_i$ as follows. Let σ be a permutation of $\mathcal{H}_0$, taken at random from the set of all permutations of $\mathcal{H}_0$ with equal probability, independently of $p_1, \ldots, p_m$. Extend σ to a permutation on $\{1, \ldots, m\}$ by taking $\sigma(i) = i$ for $i \notin \mathcal{H}_0$. Let $\tilde{p}_i = (m/m_0)p_{\sigma(i)}$, where $m_0 = \#\mathcal{H}_0$. Then, for $i \in \mathcal{H}_0$,

$$P(\tilde{p}_i \leq \alpha) = \sum_{j \in \mathcal{H}_0} P(\sigma(i) = j) P(p_j(m/m_0) \leq \alpha | \sigma(i) = j) = \frac{1}{m_0} \sum_{j \in \mathcal{H}_0} P(p_j(m/m_0) \leq \alpha)$$

where we use independence of σ and p_j and the fact that $P(\sigma(i) = j) = 1/m_0$. Since $p_1, \ldots, p_m$ satisfy (3) under $(P, \mathcal{H}_0)$, this is bounded by $(m/m_0) \cdot \alpha m_0/m = \alpha$. Thus, letting $\tilde{P}$ denote the distribution of $\tilde{p}_1, \ldots, \tilde{p}_n$ under P, $(\tilde{P}, \mathcal{H}_0)$ satisfies the classical significance level control condition (1). It follows by the assumptions of the theorem that the oracle procedure $\widetilde{\mathcal{R}}(\tilde{p}_1, \ldots, \tilde{p}_m; \mathcal{H}_0) = \mathcal{R}(\tilde{p}_1(m_0/m), \ldots, \tilde{p}_m(m_0/m)) = \mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)})$ controls FDR at level q under $\mathcal{H}_0$ when $p_1, \ldots, p_m$ are drawn according to P. But by permutation invariance of $\mathcal{R}$ and the fact that σ maps $\mathcal{H}_0$ to itself, we have $\#\left(\mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)}) \cap \mathcal{H}_0\right) = \#\left(\mathcal{R}(p_1, \ldots, p_m) \cap \mathcal{H}_0\right)$. Also, $\#\mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)}) = \#\mathcal{R}(p_1, \ldots, p_m)$ by permutation invariance. Thus, the FDR of $\mathcal{R}(p_m, \ldots, p_m)$ is the same as the FDR of $\mathcal{R}(p_{\sigma(1)}, \ldots, p_{\sigma(m)})$, and is therefore bounded by q. $\square$

This immediately gives the following corollary.

Corollary 3.1. *Let β be a shape function such that the oracle step-up procedure (10) with $\pi = m_0/m$ where $m_0 = \#\mathcal{H}_0$ controls FDR at level q for any $(P, \mathcal{H}_0)$ that satisfy the classical significance level control condition (1). Then the conservative step-up procedure (10) with $\pi = 1$ and shape function β controls FDR at level q for any $(P, \mathcal{H}_0)$ that satisfy the average significance level control condition (3).*

As a special case, applying Proposition 2.7 and Lemma 3.2(iii) in Blanchard and Roquain (2008) gives the following.

Corollary 3.2. *The class of dependence controlling step-up procedures of Blanchard and Roquain (2008), given by (10) with $\pi = 1$ and $\beta(r) = \int_0^r x d\nu(x)$ for some probability measure ν, controls FDR at level q for any $(P, \mathcal{H}_0)$ such that the average significance level control condition (3) holds. In particular, the step-up procedure of Benjamini and Yekutieli (2001), which is given by (10) with $\pi = 1$ and $\beta(r) = r/\left(\sum_{i=1}^m 1/i\right)$, controls FDR at level q for any $(P, \mathcal{H}_0)$ such that the average significance level control condition (3) holds.*

Key requirements here are that the original procedure (a) controls FDR under arbitrary dependence and (b) can incorporate the m/m_0 adjustment through an oracle result. In particular, (b) rules out procedures of the form $\mathcal{R}_{\hat{\pi},\beta(\cdot),q}$ with $\hat{\pi}$ an estimate of m_0/m, as in Storey (2002). Clearly, ruling out estimates of m_0/m is necessary, since such estimates attempt to use a bound $m_0 \cdot t$ on $V(t)$, whereas average coverage only gives a bound of $m \cdot t$ on the expectation of $V(t)$. The following counterexample shows that (a) is necessary in general even if the original p-values are independent. In particular, the BH procedure need not control FDR under the average significance level control condition (3) and independent p-values.

Suppose $m \geq 2$ and $q < 2/3$. Let $P(p_1 \leq t) = t \cdot m$ for $0 \leq t \leq (3/2) \cdot (q/m)$, and let $P(p_2 \in [a, b]) = (b - a) \cdot m$ for any $(3/2) \cdot (q/m) \leq a \leq b \leq 2q/m$. We can then distribute the remaining probability mass of p_1, p_2 and $p_3, \ldots, p_m$ over the set $2q/m$ so that the condition $\frac{1}{m} \sum_{i=1}^{m} P(p_i \leq \alpha) = \alpha$ holds. It follows that the average significance level condition (3) holds, with $\mathcal{H}_0 = \{1, \ldots, m\}$. Now consider the FDR of the Benjamini-Hochberg procedure, which rejects all hypotheses i such that $p_i \leq q\hat{r}/m$ where $\hat{r}$ is the number of rejected hypotheses. The FDR is equal to the probability of at least one rejection in this case (since $\mathcal{H}_0 = \{1, \ldots, m\}$). Note that the event $p_1 \leq q/m$ implies that hypothesis 1 is rejected, and this has probability q. But the event $q/m < p_1 \leq (3/2) \cdot (q/m)$ and $p_2 \leq 2q/m$ has probability $(q/m) \cdot (1/2) \cdot (q/m) \cdot (1/2)$, and it is disjoint with the event $p_1 \leq q/m$. This gives a lower bound of $q + [(q/m) \cdot (1/2)]^2 > q$ for the FDR. Thus, the FDR is not controlled at level q.

3.2 Estimation of FDR for Fixed Rejection Region

We now consider using the BH cutoff as an estimate of the FDR for a fixed rejection region multiple testing procedure. Under independent p-values, it is known that $\widehat{\mathrm{FDR}}(t)$ is an upwardly biased estimate of $\mathrm{FDR}(\mathcal{R}_t^{\mathrm{fixed}})$ under the classical significance level control condition (1) (Storey et al., 2004). We now show that this property continues to hold under the weaker average significance level control condition (3). The result essentially follows from the same arguments as in the case where the p-values satisfy the classical significance level control condition.

Theorem 3.2. *Suppose that $(P, \mathcal{H}_0)$ satisfies the average significance level control condition (3). Then $E_P \widehat{\mathrm{FDR}}(t) \geq \mathrm{FDR}(\mathcal{R}_t^{\mathrm{fixed}}, \mathcal{H}_0, P)$.*

Proof. For $V(t)$ and $S(t)$ defined in (7), we have

$$E_P \widehat{\mathrm{FDR}}(t) - \mathrm{FDR}(\mathcal{R}_t^{\mathrm{fixed}}, \mathcal{H}_0, P) = E_P \frac{m \cdot t - V(t)}{[V(t) + S(t)] \vee 1} \geq E_P \frac{m \cdot t - V(t)}{[m \cdot t + S(t)] \vee 1}$$

(the last step follows by noting that replacing $V(t)$ with $m \cdot t$ in the denominator weakly decreases the denominator when the numerator is negative and weakly increases the denominator when the numerator is positive). The result then follows by noting that $S(t)$ and $V(t)$ are independent by the independence assumption on p-values, and that $E_P V(t) \leq m \cdot t$ by the assumption that the p-values satisfy the average significance level control condition (3). $\square$

4 Asymptotic Results

We now consider asymptotic FDR control, under a sequence $P = P^{(m)}$ of probability measures and $\mathcal{H}_0 = \mathcal{H}_0^{(m)}$ and $m \to \infty$. We suppress the dependence on m whenever it doesn't cause confusion, but we note that the p-values form a triangular array, since the distribution (and the set $\mathcal{H}_0$ of true null hypotheses) can change with m. Recall the definitions of $V(t)$, $S(t)$ and $R(t)$ in (7). If the average significance level control condition (3) holds, and the p-values do not exhibit too much statistical dependence, we will have

$$\frac{1}{m}V(t) \leq t + o_P(1) \text{ for all } t \in [0,1]. \tag{12}$$

For some results, we also assume a law of large numbers for the total rejections and rejected true nulls:

$$\frac{1}{m}V(t) \xrightarrow{p} G(t) \leq t \quad \text{and} \quad \frac{1}{m}R(t) \xrightarrow{p} F(t) \text{ for all } t \in [0,1]. \tag{13}$$

These assumptions are analogous to assumptions made for asymptotic FDR control under classical significance level control in the literature (e.g. Storey et al., 2004, Eq. (7)-(9)). The difference here is that the conditions are weaker, since the upper bound in (12) is given by t rather than $t\pi_0$ where π_0 is the limit of $\#\mathcal{H}_0/m$. As one might expect, this will lead to problems for "adaptive" procedures that attempt to estimate π_0. However, as we now show, it is not a problem for the Benjamini-Hochberg procedure, which uses the conservative upper bound of 1. We first show conservative consistency of the BH cutoff (8) for the FDR (and FDP) of the fixed rejection region procedure.

Theorem 4.1. *Let $\widehat{\mathrm{FDR}}(t)$ be the BH estimate, given in (8), of the FDR of the fixed rejection region procedure $\mathcal{R}_t^{fixed}$ given in (6) and suppose that (12) holds. Then, for any $\underline{t}$ such that* *there exists* $\eta > 0$ *with* $\frac{1}{m}\sum_{i=1}^n I(p_i \le \underline{t}) \ge \eta + o_P(1)$*, we have*

$$\inf_{t\in[\underline{t},1]} \left[\widehat{\mathrm{FDR}}(t) - \mathrm{FDP}(\mathcal{R}_t^{\textit{fixed}}, \mathcal{H}_0)\right] \ge o_P(1).$$

If, in addition, (13) holds for continuous functions G *and* F*, then, letting* $\mathrm{FDR}_\infty(t) = G(t)/F(t)$*, we have*

$$\sup_{t\in[\underline{t},1]} \left|\mathrm{FDP}(\mathcal{R}_t^{\textit{fixed}}, \mathcal{H}_0) - \mathrm{FDR}_\infty(t)\right| \xrightarrow{p} 0, \quad \sup_{t\in[\underline{t},1]} \left|\mathrm{FDR}(\mathcal{R}_t^{\textit{fixed}}, \mathcal{H}_0, P) - \mathrm{FDR}_\infty(t)\right| \to 0$$
$$\textit{and} \inf_{t\in[\underline{t},1]} \left[\widehat{\mathrm{FDR}}(t) - \mathrm{FDR}(\mathcal{R}_t^{\textit{fixed}}, \mathcal{H}_0, P)\right] \ge o_P(1).$$

Proof. We first note that (12) implies

$$\inf_{t\in[0,1]} [t - V(t)/m] \ge o_P(1) \tag{14}$$

and that (13) implies

$$\sup_{t\in[0,1]} |V(t)/m - G(t)| \xrightarrow{p} 0 \quad \text{and} \quad \sup_{t\in[0,1]} |R(t)/m - F(t)| \xrightarrow{p} 0 \tag{15}$$

(this follows by first replacing the set $[0,1]$ with $\{0, 1/K, \ldots, (K-1)/K, 1\}$ and then taking $K \to \infty$, using uniform continuity of $G(t)$ and $F(t)$ on $[0,1]$ and the fact that $F(t)$, $G(t)$, $V(t)/m$ and $R(t)/m$ are nondecreasing).

For any $\varepsilon > 0$, the event $\inf_{t\in[\underline{t},1]}\left[\widehat{\mathrm{FDR}}(t) - \mathrm{FDP}(\mathcal{R}_t^{\mathrm{fixed}}, \mathcal{H}_0)\right] < -\varepsilon$ implies that there exists $t \in [\underline{t}, 1]$ such that $t - V(t)/m < -\varepsilon(R(t) \vee 1)/m \le -\varepsilon(R(\underline{t}) \vee 1)/m$. This implies $\inf_{t\in[\underline{t},1]}[t - V(t)/m] \le -\varepsilon \cdot \eta/2$ on the event $R(\underline{t})/m = \frac{1}{m}\sum_{i=1}^m I(p_i \le \underline{t}) \ge \eta/2$, which hold with probability approaching one by assumption. The first statement now follows by noting that the probability of this event converges to zero by (14).

The first part of the second statement follows immediately from (15), using uniform continuity of the function $(a, b) \mapsto a/b$ over $b \in [F(\underline{t}), 1]$ since $F(\underline{t}) \ge \eta > 0$. The remaining parts of the second statement then follow immediatly from the dominated convergence theorem.

□

Next, we show asymptotic control of FDR for the BH procedure $\mathcal{R}_{\mathrm{BH},q}$ defined in (9).

Theorem 4.2. *Suppose Assumptions (12) and (13) hold for continuous functions* G *and* F

and that there exists $t^* > 0$ such that $F(t^*) > 0$ and $G(t^*)/F(t^*) < q$. Then

$$\mathrm{FDP}(\mathcal{R}_{\mathrm{BH},q}, \mathcal{H}_0) \leq q + o_P(1) \quad \textit{and} \quad \mathrm{FDR}(\mathcal{R}_{\mathrm{BH},q}, \mathcal{H}_0, P) \leq q + o(1).$$

Proof. We have

$$\mathrm{FDP}(\mathcal{R}_{\mathrm{BH},q}, \mathcal{H}_0) = \frac{V(\hat{t}_{\mathrm{BH},q})/m}{[R(\hat{t}_{\mathrm{BH},q}) \vee 1]/m} \leq \widehat{\mathrm{FDR}}(\hat{t}_{\mathrm{BH},q}) + I(\hat{t}_{\mathrm{BH},q} < t^*) + o_P(1)$$

using the fact that $\sup_{t \in [t^*,1]} \left[\frac{V(t)/m}{[R(t) \vee 1]/m} - \widehat{\mathrm{FDR}}(t) \right] \leq o_P(1)$ by Theorem 4.1. Since $\widehat{\mathrm{FDR}}(\hat{t}_{\mathrm{BH},q}) \leq q$ by construction, it suffices to show $P(\hat{t}_{\mathrm{BH},q} \geq t^*) \to 1$. But this follows since $\widehat{\mathrm{FDR}}(t^*) \leq q$ implies $\hat{t}_{\mathrm{BH},q} \geq t^*$, and $\widehat{\mathrm{FDR}}(t^*) \xrightarrow{p} G(t^*)/F(t^*) < q$ by (13). This shows that $\mathrm{FDP}(\mathcal{R}_{\mathrm{BH},q}, \mathcal{H}_0) \leq q + o_P(1)$, from which it also follows that $\mathrm{FDR}(\mathcal{R}_{\mathrm{BH},q}, \mathcal{H}_0, P) \leq q + o(1)$ by dominated convergence. □

References

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2022): "Robust Empirical Bayes Confidence Intervals," *Econometrica*, 90, 2567–2602.

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

Benjamini, Y. and Y. Hochberg (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Benjamini, Y. and D. Yekutieli (2001): "The Control of the False Discovery Rate in Multiple Testing under Dependency," *The Annals of Statistics*, 29, 1165–1188.

Blanchard, G. and E. Roquain (2008): "Two simple sufficient conditions for FDR control," *Electronic Journal of Statistics*, 2, 963–992.

Cai, T. T., M. Low, and Z. Ma (2014): "Adaptive Confidence Bands for Nonparametric Regression Functions," *Journal of the American Statistical Association*, 109, 1054–1070.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018): "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 551–577.

Genovese, C. and L. Wasserman (2004): "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061.

Ignatiadis, N. and W. Huber (2021): "Covariate powered cross-weighted multiple testing," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 720–751, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12411.

Lei, L. and W. Fithian (2018): "AdaPT: an interactive procedure for multiple testing with side information," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 649–679.

Li, A. and R. F. Barber (2019): "Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 45–74.

Low, M. G. (1997): "On nonparametric confidence intervals," *The Annals of Statistics*, 25, 2547–2554.

Nychka, D. (1988): "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134–1143.

Storey, J. D. (2002): "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 479–498.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004): "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 187–205.

Wahba, G. (1983): "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society: Series B (Methodological)*, 45, 133–150.

Wasserman, L. (2007): *All of Nonparametric Statistics*, New York: Springer.