Vol. 0 (0000)

Covariate-powered weighted multiple testing with false discovery rate control

Nikolaos Ignatiadis and Wolfgang Huber

Department of Statistics
Stanford University
e-mail: ignat@stanford.edu

European Molecular Biology Laboratory
Heidelberg, Germany
e-mail: whuber@embl.de

Abstract: Consider a multiple testing setup where we observe mutually independent pairs $((P_i, X_i))_{1 \leq i \leq m}$ of p-values P_i and covariates X_i, such that $P_i \perp X_i$ under the null hypothesis. Our goal is to use the information potentially available in the covariates to increase power compared to conventional procedures that only use the P_i, while controlling the false discovery rate (FDR). To this end, we recently introduced independent hypothesis weighting (IHW), a weighted Benjamini-Hochberg method, in which the weights are chosen as a function of the covariate X_i in a data-driven manner. We showed empirically in simulations and datasets from genomics and proteomics that IHW leads to a large power increase, while controlling the FDR. The key idea was to use hypothesis splitting to learn the weight-covariate function without overfitting. In this paper, we provide a survey of IHW and related approaches by presenting them under the lens of the two-groups model, when it is valid conditionally on a covariate. Furthermore, a slightly modified variant of IHW is proposed and shown to enjoy finite sample FDR control. The same ideas can also be applied for finite-sample control of the family-wise error rate (FWER) via IHW-Bonferroni.

1. Introduction

1.1. Motivation

Statistical analysis of modern high-throughput datasets invariably invokes the issue of multiple testing. The problem has been well studied, and many solutions have been proposed. The most commonly used approaches start with a list of p-values P_i, one for each hypothesis H_i, and reject all hypotheses with a p-value below a (possibly random, that is, data-dependent) threshold t^*. The goal is to control a measure of type-I error at level α. Traditionally, this measure has been the family-wise error rate (FWER), but for many applications this is too stringent, and over the last 20 years the false discovery rate (FDR) has become a popular choice [5], as it is more permissive and adaptive [4]. Nevertheless, practitioners often think of the multiple testing problem as a burden – a necessary price to pay for doing high-throughput exploratory work [44]. However, multiple testing also presents an unparalleled new opportunity: seeing the results from many tests simultaneously allows us to infer properties of our data that we could never learn from a single test. One might argue that the perception of multiple testing as a burden is exacerbated by shortcomings of existing statistical methods that do not make use of all the information available.

The case in point explored here is that often, beyond p-values, side information, represented by covariates X_i, is available for each hypothesis. This side-information may be related to the different power of the tests, or to different prior probabilities of the null hypothesis being true. Such covariates are often apparent to statisticians or to domain scientists [26]. Yet, existing methods working only with p-values ignore this information – possibly because this side information was irrelevant in the context of (frequentist) single hypothesis testing. Such covariate-adjusted FDR estimation and control had already been considered a decade ago [19,35], but has recently resurfaced as an active research topic.

In this paper we elaborate on an idea we recently introduced to the computational biology community [26]: We proposed indepent hypothesis weighting (IHW), a modification to the Benjamini-Hochberg procedure [5] (the most frequently used method for FDR control) that can increase power by incorporating such covariates X_i. Each X_i is assumed to be independent of the p-value P_i under the null hypothesis, yet informative about power. The basic idea is simple: assign a weight to each hypothesis based on the value of its covariate X_i by approximating the optimal decision boundary in a data-driven way. However, doing this in a naive way can result in overfitting and loss of FDR control. This is avoided by employing randomization in the form of hypothesis splitting into k-folds: Learn the weights of hypotheses in a given fold from the other $k-1$ folds. The ideas are widely applicable and can be generalized to other multiple testing procedures that can make use of weights, i.e., non-negative numbers that average to one, and that signal different priorities for different hypotheses. As an example for such generalization, we also show the applicability of our methods to the Bonferroni procedure.

1.2. Outline

In Section 2 we quickly review the Benjamini-Hochberg (BH) procedure in terms of the two-groups model. We also introduce the weighted BH procedure. In Section 3 the two-groups model is generalized to the situation in which we have covariates (conditional two-groups model) and a range of illustrative applications are mentioned. Subsequently, this model is used to motivate and describe IHW (Section 4). We prove finite sample FDR control of a further modification of this procedure in Section 5. Finally we provide a survey of related approaches in Section 6.

2. Multiple testing background

Consider testing m distinct hypotheses $H_1, \dots, H_m$ based on p-values $P_1, \dots, P_m$. We will write $H_i = 0$ for the null hypotheses and $H_i = 1$ for the alternatives. In this paper, the p-values are assumed to be uniformly distributed under the null (or stochastically larger than uniform), and jointly independent.

A multiple testing procedure will reject R hypotheses, and V of these will be nulls, i.e., it will commit V type-I errors. The generalized type-I error is usually the expectation of a function of V and R. For example, the family-wise error rate (FWER), is defined as FWER $:= \mathbb{P}[V \geq 1]$. Here we will mainly focus on the false discovery rate (FDR), defined as the expectation of the false discovery proportion (FDP):

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

2.1. The two-groups model and the Benjamini-Hochberg procedure

The two-groups model [17] starts with a Bayesian flavour: Instead of deterministic H_i, each hypothesis has the same prior probability $\pi_0 = \mathbb{P}[H_i = 0]$ of being null. In addition, p-values are uniformly distributed under the null, and alternative p-values have distribution F_{alt} that is stochastically smaller than uniform:

$$\begin{aligned} &H_i \sim \text{Bernoulli}(1-\pi_0) \\ &P_i \mid H_i = 0 \sim U[0,1] \\ &P_i \mid H_i = 1 \sim F_{\text{alt}} \end{aligned} \tag{1}$$

Under this setup, a natural quantity to control is the posterior probability of being a null. This is also sometimes referred to as the Bayesian false discovery rate, denoted by Fdr.

$$\text{Fdr}(t) = \mathbb{P}[H_i = 0 \mid P_i \leq t] = \frac{\pi_0 t}{F(t)}$$

$F(t) := \pi_0 t + (1-\pi_0) F_{\text{alt}}(t)$ is the marginal distribution of the p-values under the two-groups model.

The Benjamini-Hochberg procedure now proceeds as follows: First, conservatively estimate π_0 by 1, estimate $F(t)$ by the empirical CDF $\widehat{F}(t)$ and then estimate $\text{Fdr}(t)$ by plugging-in the above estimators, i.e., $\widehat{\text{Fdr}}(t) = \frac{t}{\widehat{F}(t)}$. In the second step, it chooses $\hat{t}^* \in [0,1]$ as large as possible, such that $\widehat{\text{Fdr}}(\hat{t}^*) \leq \alpha$, for some pre-specified $\alpha \in (0,1)$. Finally, it rejects all hypotheses with p-value $P_i \leq \hat{t}^*$.

When the two-groups model holds with $\pi_0 < 1$ (and reasonable alternative distribution), then asymptotically ($m \to \infty$), it controls the Bayesian Fdr and $\hat{t}^*$ converges almost surely to t^*, where $t^* := \sup\left\{ t \in [0,1] \mid \frac{t}{F(t)} \leq \alpha \right\}$ [45].

More importantly, the procedure also controls the frequentist FDR, even when the two-groups model does not hold, each alternative p-value has a different distribution and H_i are considered deterministic. This shows the robustness of the $\widehat{\text{Fdr}}(t)$ estimator and is related to the nonparametric properties of the empirical distribution function. In fact, it is a surprising result that FDR is controlled, despite the greedy choice of the threshold $\hat{t}^*$ [17].

2.2. The weighted Benjamini-Hochberg procedure

Despite the favourable properties of the BH procedure, one of its major shortcomings is that it ignores heterogeneity across hypotheses. For example, consider the following generalization of the two-groups model in which each hypothesis has its own prior probability $\pi_{0,i}$ and alternative distribution $F_{\text{alt},i}$.

$$\begin{aligned} &H_i \sim \text{Bernoulli}(1-\pi_{0,i}) \\ &P_i \mid H_i = 0 \sim U[0,1] \\ &P_i \mid H_i = 1 \sim F_{\text{alt},i} \end{aligned} \tag{2}$$

By ignoring this heterogeneity, the BH procedure pays a price in terms of power loss. However, we emphasize that type-I error control is still guaranteed.

While there are multiple ways to exploit such heterogeneity (Section 6), a very general approach is to assign a weight W_i to each hypothesis. The W_i satisfy $W_i \geq 0$ and $\sum_{i=1}^m W_i = m$. The weighted BH procedure (Algorithm 1, [21]) then simply applies the ordinary BH procedure to the weighted p-values P_i/W_i. Thus hypotheses with $W_i > 1$ get prioritized.

If the weights are chosen a-priori, i.e., without looking at the p-values, then the weighted BH procedure also controls the frequentist FDR [21]. In particular, most of

```
Algorithm 1: The weighted Benjamini-Hochberg method
Input: A nominal level $\alpha \in (0,1)$, a vector of p-values $P_1, \ldots, P_m$ and weights
       $W_1, \ldots W_m \geq 0$ with $\sum_{i=1}^m W_i = m$
Let $Q_i = \frac{P_i}{W_i}$ ($Q_i = 0$ for $P_i = 0$, $Q_i = \infty$ for $W_i = 0, P_i \neq 0$)
Let $Q_{(1)}, \ldots, Q_{(m)}$ be the order statistics of $Q_1, \ldots, Q_m$ and let $Q_{(0)} := 0$
Let $k^* = \max \left\{ k \mid Q_{(k)} \leq \frac{k\alpha}{m} \text{ , } k \geq 0 \right\}$
Reject all hypotheses with $Q_i \leq Q_{(k^*)}$
```

the literature to date only considers the case of deterministic weights [8,23,40]. These results are very valuable, since weighted multiple testing procedures have been shown to be robust to weight mispecification [21]: Choosing good weights can lead to huge increases in power, yet "bad" weights will only slightly decrease power compared to the unweighted procedure. This has lead to numerous papers heuristically suggesting weights for specific scientific applications (e.g. [12, 29, 36, 54]). As an example of principled prior guessing, in [14, 20] an elegant approach is developed for hypothesis weighting based on effect size information from a prior experiment.

In contrast, the major goal of this work is to allow the weights to depend also on the p-values in a data-driven way, while still controlling the FDR. Indeed, the weights will have to depend on the p-values if we want to develop a powerful procedure that is practical and can be used out-of-the-box by practitioners (e.g., computational biologists) without labourious modeling being a prerequisite (but see also Subsection 6.5 for a counter-point).

3. Conditional two-groups model

The starting point for developing a new procedure for data-driven hypothesis weighting is an extension of the two-groups model (1), so that it better approximates the very general model in (2), wherein each hypothesis has its own prior probability $\pi_{0,i}$ and alternative distribution $F_{\text{alt},i}$. The idea is that differences in $\pi_{0,i}$ and $F_{\text{alt},i}$ across hypotheses can be explained by an observed random covariate $X \in \mathcal{X}$ (see also [27]). For example, we model the prior probability π_0 as a function $\mathcal{X} \to [0,1]$. We call this the conditional two-groups model (3):

$$\begin{aligned} &X_i \sim \mathbb{P}^X \\ &H_i \mid X_i \sim \text{Bernoulli}(1 - \pi_0(X_i)) \\ &P_i \mid (H_i = 0, X_i) \sim U[0,1] \\ &P_i \mid (H_i = 1, X_i) \sim F_{\text{alt}|X_i} \end{aligned} \tag{3}$$

Marginalizing over H_i, we get that:

$$P_i \mid X_i = x \sim F(t|x) := \pi_0(x)t + (1 - \pi_0(x))F_{\text{alt}|X_i = x}(t)$$

Thus, in the conditional two-groups model, (P_i, X_i, H_i) are assumed to be exchangeable, and since we observe X_i we have more flexibility in our modeling than in the two-groups situation. Later we will model weights as a function of X_i.

Similarly to the two-groups model, we are just interested in it being an approximation to the truth. In practice, the critical component is the conditional independence of X_i and P_i under the null hypothesis. In addition, we can only expect power gains when indeed $F_{\text{alt}|X_i=x}$ and $\pi_0(x)$ are not constant as functions of x.

We now demonstrate that existence of such covariates X_i is a weak assumption, since these are available in multiple applications.

3.1. Domain-specific covariates: "Co-data"

In many scientific applications, such covariates are apparent to domain scientists: These covariates usually are related to the true effect sizes of the individual hypotheses or their prior probability of being true. This relation is the result of either a known causal relationship between the covariate and the hypotheses being tested or because previous, related studies have shown an association. In any case, for such domain-specific covariates to be independent of the p-values under the null hypothesis, the critical aspect is that they should not have been used in any way for the marginal hypothesis testing.

Here we mention some examples, see [26] for additional ones:

- In neuroscience, we are now able to simultaneously measure the activity of hundreds of neurons. A scientific question of interest is whether two neurons are in synchrony [43]. Here, we know that neurons which are in close proximity to each other are a-priori more likely to be interacting. Thus, the geometric distance between the neurons can be used as a covariate for rejecting the null hypothesis that the neurons are not interacting.
- There might exist p-values from a previous, but related experiment: For example, in [20] data from previous, independent genome-wide association studies (GWAS) for related diseases are used to increase the power of a longevity related GWAS study.

The widespread existence of such covariates was also observed in [53], who used the term "co-data" to describe these and developed a weighted ridge regression procedure, with data-driven weights based on the co-data.

3.2. Statistical covariates

In single hypothesis testing, classical theory often dictates which test statistics should be used under optimality considerations. All other information can essentially be discarded or should be conditioned on. In this section, we want to illustrate that such information, which is irrelevant in single hypothesis testing, can be embedded in the conditional two-groups framework and can help increase the power of the resulting multiple testing procedure; sometimes dramatically so.

3.2.1. Sample size N_i

A very generic covariate, with many applications, is the sample size N_i, when it differs across tests. Note that if the test statistic is continuous and the null hypothesis is simple, then the p-value under the null will still be uniformly distributed independently of N_i. It is also reasonable to assume that the prior probability of a hypothesis being true does not depend on N_i, i.e., $\pi_0(N_i) = \pi_0$. However, the alternative distribution does depend on N_i: For higher N_i we have more power. For example, consider a one-sided z-test in which we observe independent $X_1^i, \dots, X_{N_i}^i \sim \mathcal{N}(\mu H_i, 1)$ with $\mu > 0$ and use $P_i = 1 - \Phi\left(\sqrt{N_i}\,\overline{X^i}\right)$ as our statistic.

Then the conditional two-groups model applies with $\pi_0(N_i) = \pi_0$ and

$$F_{\text{alt}}(t \mid N_i) = 1 - \Phi(\Phi^{-1}(1-t) - \sqrt{N_i}\mu)$$

While this example is trivial, it is instructive: If one wants to maximize discoveries, then one expects that hypotheses with large sample sizes should be prioritized. The methods described here will be able to accomplish this automatically and hence increase power. Yet, in some cases, this might not necessarily be desirable: p-values are often criticized for reflecting sample size more than effect size, and optimal weights would amplify this effect.

Remark 1. Technically, optimal weighting automatically adjusts to very large sample sizes [22,34,37]: Hypotheses with a very large sample size (or effect size) should be down-weighted; this phenomenon is called size-investing. However, in many practically relevant situations, it will still be the case that larger sample size attracts larger optimal weight to a hypothesis.

3.2.2. Overall variance (independent of label) in two-sample tests

For a more interesting example, consider two-sample testing for equality of means. To simplify the discussion, assume that for the i-th hypothesis we observe

$$X_{i,1}, \ldots, X_{i,n} \sim \mathcal{N}(\mu_{X,i}, \sigma_i^2) \quad \text{and} \quad Y_{i,1}, \ldots, Y_{i,n} \sim \mathcal{N}(\mu_{Y,i}, \sigma_i^2)$$

(everything independent). We are interested in testing $H_i : \mu_{X,i} = \mu_{Y,i}$, and do not know σ_i. The optimal test statistic for this situation is the two-sample t-statistic:

$$T_i = \sqrt{\frac{n}{2}} \frac{\overline{X_i} - \overline{Y_i}}{\sqrt{\frac{S_{X,i}^2 + S_{Y,i}^2}{2}}}$$

Here $\overline{X_i}$ (resp. $\overline{Y_i}$) are the sample means of $X_{i,1}, \ldots, X_{i,n}$ (resp. $Y_{i,1}, \ldots, Y_{i,n}$). Similarly, $S_{X,i}^2$ (resp. $S_{Y,i}^2$) are the sample variances.

In addition, denote by $\hat{\mu}_i := \frac{1}{2}\left(\overline{X_i} + \overline{Y_i}\right)$ and S_i^2 the sample mean and sample variance after pooling all observations ($X_{i,1}, \ldots, X_{i,n}$ and $Y_{i,1}, \ldots, Y_{i,n}$) and forgetting their labels.

Now note that under the null hypothesis $\mu_{X,i} = \mu_{Y,i} =: \mu_i$ and we have $X_{i,1}, \ldots, X_{i,n}$, $Y_{i,1}, \ldots, Y_{i,n} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ i.i.d. Then, $(\hat{\mu}_i, S_i^2)$ is a complete sufficient statistic for the experiment, while T_i is ancillary. Thus, by Basu's theorem, when the i-th hypothesis is null ($H_i = 0$), $(\hat{\mu}_i, S_i^2)$ is independent of T_i and we can use it as a covariate. As mentioned above, while irrelevant in single hypothesis testing, $(\hat{\mu}_i, S_i^2)$ can be extremely useful in a multiple hypothesis testing.

First, consider S_i^2 and note that under the null it is distributed as a scaled χ^2-distribution. On the other hand, under the alternative, we expect S_i^2 to take larger values with high probability, especially for large values of $|\mu_{X,i} - \mu_{Y,i}|$. Therefore, if we are doing m t-tests, each with unknown variance σ_i^2 and if we assume $\sigma_i \sim G$, from a concentrated distribution G, then hypotheses with high S_i^2 are more likely to be true alternatives (and also likely to be alternatives with high power). Thus, the overall variance (ignoring sample labels) is independent of the p-values under the null hypothesis, yet informative about the alternative and can lead to a large power increase in simultaneous two-sample t-tests [10,26].

For a second example of the usefulness of $(\hat{\mu}_i, S_i^2)$ in this setting, consider the screening statistic $\frac{|\hat{\mu}_i|}{S_i}$. This can be interpreted as a statistic for the null hypothesis

$\mu_{X,i} = \mu_{Y,i} = 0$. If we believe a-priori that for many of the hypotheses i with $\mu_{X,i} = \mu_{Y,i}$ a sparsity condition holds, so that in fact $\mu_{X,i} = \mu_{Y,i} = 0$ [30], then large values of this statistic are more likely to correspond to alternatives. Note that we did not have to actually re-specify our null hypothesis from $\mu_{X,i} = \mu_{Y,i}$ to $\mu_{X,i} = \mu_{Y,i} = 0$. We instead just used the latter to derive our covariate and are still testing for $\mu_{X,i} = \mu_{Y,i}$.

3.2.3. Ratio of number observations in each group in two-sample tests

For yet another example, revisit the two-sample situation, but now assume that for the i-th hypothesis, we have $n_{1,i}$ observations of the first population and $n_{2,i}$ observations from the second population, such that $n_{1,i}+n_{2,i} = n_i$. Then $\frac{n_{1,i}(n_i-n_{1,i})}{n_i^2}$ is a statistic which is related to the alternative distribution, with values close to $\frac{1}{4}$ implying higher power [39]. This statistic is also related to the Minor Allele Frequency (MAF) in genome-wide association studies [9].

3.2.4. Sign of estimated effect size

As a final example of a statistical covariate, consider a two-sided test, where the null distribution is symmetric and the test-statistic is the absolute value of a symmetric statistic T_i. Then, the sign of T_i is independent of the p-value under the null hypothesis. However, we might a-priori believe that among the alternatives, more hypotheses have a true positive effect size, rather than a negative one or vice versa. Thus, the sign could also be used as an informative covariate. The idea of using the sign to improve power, while controlling the FDR, can actually be traced back to the early days of microarrays, where it was implemented in the SAM (significance analysis of microarrays) procedure [51].

3.3. p-value covariates: Multivariate p-values

In many situations, we have bivariate (or even multivariate) p-values for each hypothesis which are independent under the null. For a toy example, we revisit the two-sample t-test in 3.2.2, this time with known variance $\sigma_i^2 = 1$. For testing we again use T_i as our statistic (which is suboptimal in this case). Then, the overall variance (ignoring labels) can be converted into a second p-value for H_i, noting that it is χ^2-distributed under the null hypothesis [15]. In a more realistic setting, multivariate p-values are also available in heterogeneous multiomics experiments in biology [1] (e.g., a p-value based on transcriptomics measurements and one based on proteomics for each hypothesis).

In such a situation, the methods described here can be applied by considering one p-value as being primary and the rest as secondary (i.e., as covariates).

Nevertheless, for this situation, specialized methods have been developed [1, 15] and are preferable, since they consider the components of the multidimensional p-values in a symmetric way and can also profit from the additional distributional properties inherent to p-values.

4. Data-driven hypothesis weighting

4.1. An oracle procedure: IHWoracle

We will develop the IHW (Independent Hypothesis Weighting) procedure using an analogous motivation to that of the BH procedure. We consider a Bayesian oracle which knows the conditional two-groups model (3). Rather than returning a single threshold $t \in [0,1]$ and rejecting hypotheses with $P_i \leq t$, the new oracle returns a function $g : \mathcal{X} \to [0,1]$ in a function class $\mathcal{G}$ and rejects hypotheses with $P_i \leq g(X_i)$. In this setting, the posterior probability of being a null (conditionally on rejection is):

$$\mathbb{P}[H_i = 0 \mid H_i \text{ rejected}] = \mathbb{P}[H_i = 0 \mid P_i \leq g(X_i)] = \frac{\int_{\mathcal{X}} \pi_0(x) g(x) d\mathbb{P}^X}{\int_{\mathcal{X}} F(g(x)|x) d\mathbb{P}^X} \tag{4}$$

The BH-type oracle maximized t subject to a constraint on the posterior probability. This is equivalent to maximizing power ($F(t) = \mathbb{P}[P_i \leq t]$). Note that in the conditional two-groups model the power satisfies:

$$\mathbb{P}[P_i \leq g(X_i)] = \int_{\mathcal{X}} F(g(x)|x) d\mathbb{P}^X \tag{5}$$

This leads us to consider the threshold function g that maximizes power subject to a constraint on the posterior probability of making a type-I error:

$$\begin{aligned} &\underset{g \in \mathcal{G}}{\text{maximize}} && \int_{\mathcal{X}} F(g(x)|x) d\mathbb{P}^X \\ &\text{subject to} && \frac{\int_{\mathcal{X}} \pi_0(x) g(x) d\mathbb{P}^X}{\int_{\mathcal{X}} F(g(x)|x) d\mathbb{P}^X} \leq \alpha \end{aligned} \tag{6}$$

Replacing $\mathbb{P}^X$ by the empirical measure we get:

$$\begin{aligned} &\underset{g \in \mathcal{G}}{\text{maximize}} && \frac{1}{m} \sum_{i=1}^{m} F(g(X_i)|X_i) \\ &\text{subject to} && \frac{\frac{1}{m}\sum_{i=1}^{m} \pi_0(X_i) g(X_i)}{\frac{1}{m}\sum_{i=1}^{m} F(g(X_i)|X_i)} \leq \alpha \end{aligned} \tag{7}$$

As in the case of the BH procedure we would like a procedure that (a) approximates the oracle procedure when the assumptions of the conditional two-groups model are met and (b) is robust to misspecification of the conditional two-groups model. In particular, even given an adversary that manipulates our oracle, our procedure should only be affected in its power; but it should still control the FDR. The goal is, of course, that even if the conditional two-groups model is a coarse approximation to the truth and we have a good guess of that approximation, we will still be able to gain power compared to BH and control the FDR.

It turns out that to construct such a procedure (called IHWoracle henceforth, see Algorithm 2), we simply need to rescale the oracle threshold function from (7) and then apply the weighted BH procedure. We call the procedure IHWoracle since we assume we have some prior source (which could be the optimal oracle, but not necessarily) which informs us about $F(t \mid x)$ and $\pi_0(\cdot)$.

In particular, while motivated by the conditional two-groups model (3), the validity of this method will will depend on the following more general distributional assumption:

Algorithm 2: The IHWoracle procedure

```
Let $g$ be a solution of optimization problem (7)
For $i \in \{1,\dots,m\}$, set $W_i = 1$ if $g(X_i) = 0 \;\forall\, i$, otherwise set $W_i = \frac{m g(X_i)}{\sum_{i=1}^m g(X_i)}$
Apply the weighted BH procedure (Algorithm 1) to $((P_i, W_i))_i$
```

Assumption 1 (Distributional setting)**.** *Let (P_i, X_i), $i \in \{1,\dots,m\}$ be mutually independent (p-value, covariate) pairs and $\mathscr{H}_0 \subset \{1,\dots,m\}$ the index set of null hypotheses. Also assume that for $i \in \mathscr{H}_0$ it holds that $P_i \perp X_i$ (independent) and P_i is (super)uniform, i.e. $\mathbb{P}[P_i \leq t] \leq t$.*

To show the validity of the IHWoracle procedure, we first show:

Theorem 1. *Consider a measurable weighting function $\mathbf{W} = (W_1,\dots,W_m) : \mathcal{X}^m \to [0,m]$ that depends only on the covariates $\mathbf{X} = (X_1,\dots .X_m)$ such that $\sum_{i=1}^m W_i(\mathbf{X}) = m$ almost surely. Then, under Assumption 1, the weighted BH procedure (Algorithm 1) with p-values $\mathbf{P} = (P_1,\dots,P_m)$ and weights $\mathbf{W}(\mathbf{X})$ controls the* FDR *at the pre-specified level $\alpha \in (0,1)$.*

Proof. All the steps of the proof of Theorem 2 with $\tau = 1$ go through essentially unchanged. □

Corollary 1 (IHWoracle)**.** *Let $F(t \mid x)$ be an arbitrary (but deterministic) conditional distribution function and $\pi_0(\cdot)$ an arbitrary (deterministic) prior probability function, as specified in the conditional two-groups model (3). Then, under Assumption 1, the IHWoracle procedure controls the* FDR *at the nominal level α:*

$$\mathrm{FDR}_{IHWoracle} \leq \alpha$$

Proof. Just observe that the IHWoracle rule of assigning weights exactly fulfills the conditions specified in Theorem 1. □

We note that Theorem 1 is applicable more generally. For instance, let's revisit the two-sample t-test from Section 3.2.2, where for simplicity we assume $m = 2m', m' \in \mathbb{N}$. Now, consider the following procedure: Apply the BH procedure to the m' hypotheses with the highest overall sample variance (independent of sample label). This is equivalent to assigning weight $W_i = 2$ to hypotheses above that cutoff, $W_i = 0$ to the rest and applying the weighted BH procedure. Theorem 1 now gives a rigorous justification for the validity of such a procedure: something that has been observed in previous work [10, 49] and shown to lead to a strong power increase in applications [10].

4.2. Bird's eye view of IHW

So far, it might seem like we have not made a lot of progress: We have shifted the a-priori guessing from the weights [21] to guessing $\pi_0(x)$ and $F(t \mid x)$. However, the latter will enable us to develop a data-driven procedure.

The first approach that comes to mind is to replace $\pi_0(x)$ and $F(t \mid x)$ in the IHWoracle procedure (Algorithm 2) in a plug-in fashion by corresponding estimators. However, such a procedure will have bad finite sample properties, even in cases where it might be asymptotically consistent [26]. This is the case because we are overfitting

Algorithm 3: IHW (Independent Hypothesis Weighting)

```
(IHWsplit) Randomly split hypotheses into $K$ folds
           for $l \in \{1, \dots, K\}$ do
               Let $I_l \subset \{1, \dots, m\}$ be the index set of hypotheses in fold $l$
    (IHW1)     Estimate $\pi_0(x)$ and $F(t \mid x)$ using $\{(P_i, X_i) \mid i \in \{1, \dots, m\} \setminus I_l\}$
    (IHW2)     Let $g(\cdot)$ be a solution of optimization problem (7) plugging in the
               estimated $\pi_0(x)$ and $F(t \mid x)$ from the previous step and $\{X_i \mid i \in I_l\}$
    (IHW3)     For $i \in I_l$, set $W_i = 1$ if $g(X_i) = 0 \; \forall \, i \in I_l$, otherwise set
               $W_i = \frac{|I_l| g(X_i)}{\sum_{i \in I_l} g(X_i)}$
           end
     (wBH) Apply the weighted BH procedure (Algorithm 1) to $((P_i, W_i))_i$
```

by learning $\pi_0(x)$ and $F(t \mid x)$ from the same (P_i, X_i) pairs to which we are then applying the weighted BH procedure.

To overcome this, we introduce randomization to the IHWoracle procedure by means of hypothesis splitting. The resulting algorithm, called IHW, is presented in its general form in Algorithm 3. The subsections below will further elaborate on the individual components of IHW, as well as provide a discussion and guidelines for practical implementations.

4.3. Hypothesis splitting approach

As has been recently demonstrated in a plethora of papers [16,32,48,52], introducing external randomness to statistical procedures can often be the key to tractable inference for high dimensional problems. The randomization protects against overfitting and leads to increased power. The hypothesis splitting step (IHWsplit) in Algorithm 3 is in exactly this spirit. To further motivate the hypothesis splitting approach, we first consider a cross-validation procedure:

4.3.1. Cross-validation

Assume that rather than controlling the FDR, our goal is to estimate the true threshold function g, which is the minimizer of (6). Also assume that we have an estimator which depends on some tuning parameters λ. Then we could use cross-validation: Split the hypotheses into K-folds and then for each fold apply the estimator to the held-out (training) folds and use the test fold to estimate the quality of the estimator. For example, one could use the threshold function learned from the training folds to derive weights for the test fold as in step (IHW3) of Algorithm 2. Then one could apply the weighted BH procedure to the test fold and calculate the number of discoveries.

Finally one would choose the tuning parameter λ that across the K test folds led on average to the highest number of discoveries.

Such a procedure can lead to reasonable estimates of the weight function (see also Section 4.4.4). However, cross-validation is not related to our goal of controlling the FDR: Even with a cross-validated weight function, there is no guarantee that the resulting procedure controls the FDR and does not overfit.

4.3.2. Hypothesis splitting

This leads us to consider the hypothesis splitting approach: We proceed as in the cross-validation procedure and calculate weights for each hypothesis in each fold based on the threshold function learned from the other folds. However, then we do not evaluate the quality of the estimator on the test fold. Instead, we just immediately apply the weighted BH procedure to all hypotheses. Consequently, the weight of a given hypothesis only depends on the p-values in the other folds and on the covariates (Section 5).

This is a novel form of data splitting, suited to the multiple testing task, where the hypotheses are mutually independent. It is reminiscent of pre-validation [50], which follows a similar splitting approach in an attempt to compare predictors derived internally from a dataset to external predictors in a fair way. Note however that if we think of a $m \times n$ data-matrix from which we get our p-values by calculating the statistic in a row-wise fashion, then pre-validation splits by columns (samples), while IHW splits by rows.

4.4. Considerations for implementing IHW

The IHW procedure requires specifying an estimator for $\pi_0(\cdot)$, an estimator for the conditional distributions $F(t \mid x)$, as well as the class $\mathcal{G}$ of functions over which (7) should be optimized. Here there are two main goals guiding these design choices: First, the estimated threshold function should generalize to the held-out fold, so that the full procedure will be powerful. Second, the computation should be tractable, even for large-scale problems.

4.4.1. Convex optimization

A key observation in [26] was that optimization problem (7), in light of $\text{Fdr}(0) = 0$, can be equivalently written as follows:

$$\begin{aligned} \underset{g \in \mathcal{G}}{\text{maximize}} \quad & \frac{1}{m}\sum_{i=1}^{m} F(g(X_i)|X_i) \\ \text{subject to} \quad & \frac{1}{m}\sum_{i=1}^{m}(\pi_0(X_i)g(X_i) - \alpha F(g(X_i)|X_i)) \leq 0 \end{aligned} \tag{8}$$

This implies, that for many reasonable choices of $\mathcal{G}$, the problem will be convex if $F(\cdot \mid X_i)$ is concave for all i. For example, if we allow $\mathcal{G}$ to consist of all measurable functions $\mathcal{X} \to [0,1]$, then we just need to optimize over $t_i = g(X_i), t_i \in [0,1], i = \{1,\dots,m\}$, thus yielding a convex m-dimensional optimization problem.

Nevertheless, it is beneficial to impose additional conditions on $\mathcal{G}$, depending on the structure of the problem (cf. [13,28]). For example, we could impose a block-wise structure (i.e., enforce piecewise constant solutions within pre-specified intervals), low total-variation, monotonicity constraints, smoothness assumptions [9] (e.g., using regression or smoothing splines). All of these would still yield convex problems.

4.4.2. Conditional distribution function estimation

In principle we can use an arbitrary (e.g., kernel) estimator of the conditional distribution $F(t \mid x)$. However, as mentioned above, from a computational point of it is handy if the estimator is concave for all x. Beyond computational convenience, this assumption is also reasonable and commonly made in the multiple testing literature [21,46].

Given an arbitrary estimator of $t \mapsto F(t \mid x)$, we can make it concave by projecting onto the space of concave distributions. If $F(\cdot \mid x)$ also has a Lebesgue density, then one could project the estimated density onto the space of decreasing densities [31].

The projection step is in general very simple, and it consists of applying the pooled-adjacent violators (PAVA) algorithm. When applied to the empirical cumulative distribution function (ECDF), this estimator is also called the Grenander estimator, the least concave majorant of the ECDF.

An alternative is to use parametric models, such as the GLMs in [27].

4.4.3. Estimation of $\pi_0(\cdot)$

We estimate $\pi_0(x)$ by 1 for all x. We defer the discussion of adaptive $\pi_0(\cdot)$ estimation to Section 6.1.

4.4.4. Choice of tuning parameters

The estimator of $F(t \mid x)$ and the set $\mathcal{G}$ are allowed to depend on tuning parameters. These can be chosen by a cross-validation procedure as described in Section 4.3.1. Within IHW, this cross-validation should be nested within the loop iterating over folds. In other words, one would split the $K-1$ held out folds in step (IHW1) of Algorithm 3 into K' further folds and then apply the cross-validation procedure.

4.4.5. Specific IHW implementation in [26]

In [26], a concrete and practical recipe was given for implementing IHW. Here we describe this in terms of the general framework established above.

The covariate X is discretized into a covariate $\widetilde{X}$ with finitely many levels $1,\dots,J$. Estimate $\pi_0(x)$ by 1 and $F(t \mid x)$ by applying the Grenander estimator to p-values P_i with the same value of the discretized covariate. The class of functions being estimated is restricted by the total variation of the weights (rescaled thresholds), i.e., for fold l, $\lambda_l > 0$ (chosen by nested cross validation) set:

$$\mathcal{G} = \left\{ g : \{1,\dots,J\} \to [0,1] \mid \sum_{j=2}^{J} |g(j) - g(j-1)| \leq \lambda_l \sum_{i \in I_l} g(\widetilde{X}_i) \right\}$$

Because the Grenander estimator is concave and piecewise linear, the resulting optimization problem (7) is a linear program.

5. Finite sample results

In [26], IHW was shown to control FDR (under strong assumptions) asymptotically. Its excellent finite sample performance was evaluated through simulations. Towards understanding the finite sample properties of IHW, here we propose two variants: IHWc and IHW-Bonferroni. We show that these control the FDR and FWER respectively in finite sample situations. The key to both proofs will be based on the following lemma:

Lemma 1. *Let $((P_i, X_i))_i$ satisfy Assumption 1. Then for the IHW procedure (Algorithm 3) it holds for all $i \in \mathscr{H}_0$ that P_i is independent of W_i ($P_i \perp W_i$).*

Proof. Let $i \in \mathscr{H}_0$ and $\mathbf{P}_{-i} = (P_j)_{j \in \{1,\dots,m\} \setminus i}$ and $\mathbf{X} = (X_1, \dots, X_m)$. By construction of the IHW procedure (in particular the hypothesis splitting), it holds that W_i is a function of $\mathbf{P}_{-i}$, $\mathbf{X}$. On the other hand, P_i is independent of $(\mathbf{P}_{-i}, \mathbf{X})$ since the pairs $(P_j, X_j), 1 \leq j \leq m$ are mutually (jointly) independent and P_i is independent of X_i ($i \in \mathscr{H}_0$). This concludes the proof. □

5.1. IHWc controls the FDR

We describe IHWc (Independent Hypothesis Weighting with censoring), a modification of IHW with provable finite-sample FDR control. For this, let $\tau \in (0,1)$. Now consider the following two modifications to IHW, following [28]:

(IHWc1) In step (IHW1) of Algorithm 3, replace $\{(P_i, X_i)\}$ by $\{(P_i \; \mathbf{1}_{\{P_i > \tau\}}, X_i)\}$. In other words, during the weight learning of the IHWc algorithm, p-values $\leq \tau$ are set to 0.

(IHWc2) Instead of applying the weighted BH procedure (wBH) in Algorithm 3 use the following modification, which prohibits rejecting hypotheses with $P_i > \tau$: Reject all hypotheses with $P_i \leq W_i \hat{k} \frac{\alpha}{m} \wedge \tau$, where

$$\hat{k} = \max\left\{ k \in \mathbb{N} \mid P_i \leq \left(\frac{\alpha W_i k}{m} \right) \wedge \tau \text{ for at least } k \text{ p-values} \right\}$$

We emphasize that modification (IHWc1) only applies to the stage of learning the weight function (i.e., of learning $\pi_0(\cdot)$ and $F(t \mid x)$ within each fold). When the weighted BH procedures gets applied, the actual (non-censored) p-values are used.

Also note that modification (IHWc2) favours a choice of large τ, while (IHWc1) a choice of small τ. For choices of $\tau \geq \alpha$, IHWc2 will in most cases just be a technical assumption and therefore we recommend setting $\tau \approx \alpha$ when there are thousands of hypotheses being tested. In any case, IHWc will indeed lose power compared to IHW, as $F(t \mid x)$ will have to be estimated with less information. The exact extent of power loss will be assessed by simulations in future work. Nevertheless, at the very least, we can still get good estimates for the contribution of $\pi_0(\cdot)$ and thus apply informative weighting.

We are now ready to state the main result:

Theorem 2. *Let $((P_i, X_i))_i$ satisfy Assumption 1.*

Then the IHWc procedure controls the FDR *at the nominal level α:*

$$\text{FDR}_{IHWc} \leq \alpha$$

Proof. We defer the proof to Section 7.1.

□

The proof uses the beginning of the argument given in [28]. The authors of this paper then bound the FDR by a quantity which exceeds α and depends on m, τ and the Rademacher complexity of the class of weight functions considered. Here, because of the hypothesis-splitting approach, FDR is controlled exactly, independently of the complexity of the class of functions considered.

5.2. IHW-Bonferroni controls the FWER

All ideas presented here can be readily extended to guarantee control of the FWER, as shown in [26]. In particular the IHW procedure (Algorithm 3) can be modified as follows:

(IHWbonf1) Instead of solving the optimization problem (7) in step (IHW2), solve the following:

$$\begin{aligned} &\underset{g \in \mathcal{G}}{\text{maximize}} && \frac{1}{m}\sum_{i=1}^{m} F(g(X_i)|X_i) \\ &\text{subject to} && \sum_{i=1}^{m} g(X_i) \leq \alpha \end{aligned}$$

(IHWbonf2) Instead of applying the weighted BH (wBH) procedure, apply the weighted Bonferroni procedure (i.e., reject H_i if $P_i \leq \frac{\alpha W_i}{m}$).

Theorem 3. *Let $((P_i, X_i))_i$ satisfy Assumption 1. Then the IHW-Bonferroni procedure controls the FWER at the nominal level α:*

$$FWER_{IHW\text{-}Bonferroni} \leq \alpha$$

Proof. See Section 7.2. □

6. Connection to previous and related approaches

6.1. Estimation of $\pi_0(x)$

6.1.1. α-exhaustiveness

Under the two-groups model, the BH procedure actually controls the FDR at level $\pi_0 \alpha \leq \alpha$. Hence power can be increased by also estimating π_0 [6,45] by an estimator $\widehat{\pi}_0$ and then applying the BH procedure at level $\frac{\alpha}{\widehat{\pi}_0}$. Similarly, for a weighted BH procedure (assuming deterministic weight function $W(\cdot)$), one can see [23] that the FDR is controlled at approximately:

$$\alpha \pi_0' \leq \alpha$$

Here we defined:

$$\pi_0' := \frac{\sum_{i=1}^{m} \pi_0(X_i) W(X_i)}{\sum_{i=1}^{m} W(X_i)} \tag{9}$$

Hence, as in the unweighted case, one could improve power by applying the weighted BH method at level $\frac{\alpha}{\widehat{\pi_0}'}$, where $\widehat{\pi_0}'$ is an estimator of π_0'.

Note that another way of expressing this, is that the weight-budget for Algorithm 1 should actually be (rather than $\sum_{i=1}^{m} W(X_i) = m$) [56]:

$$\sum_{i=1}^{m} \pi_0(X_i) W(X_i) = m \tag{10}$$

For example, in Step 2 of the IHWoracle procedure (Algorithm 2), one could normalize by:

$$W_i = \frac{mg(X_i)}{\sum_{i=1}^m \pi_0(X_i)g(X_i)}$$

We leave the analysis of such α-exhaustive procedures to future work.

6.1.2. FDR-regression estimator of $\pi_0(\cdot)$

To implement an adaptive procedure, one could attempt to estimate the function $\pi_0(\cdot)$ by $\widehat{\pi_0}(\cdot)$ and plug it into equation (9) or (10).

For example, in the GBH (Group Benjamini Hochberg) procedure [25, 42], the covariate is categorical with a finite number of levels. In this case, for each level of the covariate, a standard π_0 estimator (e.g. [6, 7, 45]) can be used.

Boca and Leek [9] consider the more challenging task of estimating $\pi_0(\cdot)$ for arbitrary covariates. Their key observation is that ($\tau \in (0, 1)$ fixed):

$$\pi_0(x) \approx \mathbb{E}\left[\frac{\mathbf{1}_{\{P_i > \tau\}}}{1-\tau} \mid X_i = x\right]$$

Thus one can estimate $\pi_0(\cdot)$ by using an arbitrary nonparametric regression procedure (e.g. regression splines) of $\frac{\mathbf{1}_{\{P_i > \tau\}}}{1-\tau}$ onto X_i.

In [28], a constrained maximum likelihood approach ($\tau \in (0, 1)$ fixed again) is proposed instead:

$$\begin{aligned} &\underset{\pi_0 \in \mathcal{Q}}{\text{maximize}} \quad \sum_{i=1}^m \left[\mathbf{1}_{\{P_i > \tau\}} \log(\pi_0(X_i)(1-\tau)) + \mathbf{1}_{\{P_i \leq \tau\}} \log(1 - \pi_0(X_i)(1-\tau))\right] \\ &\text{subject to} \quad \sum_{i=1}^m \frac{\mathbf{1}_{\{P_i > \tau\}}}{\pi_0(X_i)(1-\tau)} \leq m \end{aligned} \tag{11}$$

Here $\mathcal{Q}$ is an appropriately chosen function class (cf. the discussion in Section 4.4), while the constraint protects against underestimating $\pi_0(\cdot)$.

6.1.3. Further uses of adaptive $\pi_0(\cdot)$ estimators and size-investing

Above we focused on using a $\pi_0(\cdot)$ estimator to exhaust the nominal α-level. Of course, there is a plethora of additional reasons why the estimation of $\pi_0(\cdot)$ is an interesting task [9]. The $\pi_0(\cdot)$ estimator can be useful not only in defining the weight budget, but also the weights themselves. For example, the weights suggested in [28] fulfill $W_i \propto \frac{1}{\pi_0(X_i)}$, while in [25] they fulfill $W_i \propto \frac{1-\pi_0(X_i)}{\pi_0(X_i)}$. Via the regression framework of [9] these could be used to get weights even for high-dimensional covariates, when the curse of dimensionality obstructs the estimation of conditional distributions. In addition, for IHW, the $\pi_0(\cdot)$ estimator can flow into the final weights via optimization problem (7).

We note that making the weights only depend on an estimate of $\pi_0(\cdot)$ and not also on the conditional (alternative) distributions has one major disadvantage: No size investing can occur [26]. Here, size investing refers to situations in which hypotheses with either very high or very low power should receive a low weight, so that most of the weight budget can be assigned to hypotheses with power in between, where the weighting will matter most [34, 37].

6.2. Connection to SABHA

The Structure Adaptive Benjamini Hochberg Algorithm (SABHA) [28] is another method which is very similar to IHW and IHWc, as evidenced by the fact that the proof of Theorem 2 closely follows the SABHA proof. The covariates X_i there are deterministic rather than random. In its current implementation SABHA solves the optimization problem in (11) using the p-values directly to estimate $\pi_0(\cdot)$. Then, each hypothesis gets assigned the weight $W_i = \frac{1}{\widehat{\pi}_0(X_i)}$, which fulfills an empirical version of the adaptive weight budget (10). The overfitting of the resulting procedure is dealt with by upper bounding the FDR by a quantity greater than α which depends on the Rademacher complexity of the class $\mathcal{Q}$ in (11). In contrast, IHW avoids overfitting by hypothesis splitting.

One of the biggest disadvantages of SABHA is that the weighting scheme is inadmissible even under oracle knowledge [27] and can be improved upon. For example, in a realistic situation with a binary covariate $X \in \{A, B\}$, for which $\pi_0(A) = 0.9$ and $\pi_0(B) = 1$, oracle SABHA weighting will only be able to mildly prioritize the hypotheses with $X_i = A$ (by a relative factor of 10/9). On the other hand, an optimal procedure should assign weight $W(B) = 0$.

6.3. Connection to AdaPT

AdaPT [27] is a recent and ingenious procedure, which is also very similar in spirit to IHW. AdaPT also uses the conditional two-groups model (3) and tries to approximate the Bayes decision boundary (6). Also, similarly to IHW, it enables flexibility in the modeling, yet robustness to misspecification, in the sense that the latter should only affect power and not jeopardize control of FDR.

The main difference is that, while IHW (and SABHA) use the FDR estimator of the BH procedure as the starting point, the authors of AdaPT use a Barber-Candès (BC) type FDR estimator [2, 3], which enables the elegance of their algorithm: To estimate the number of false discoveries, they do not use the fact that $P[P_i \leq t] \leq t$ under the null, but instead they use that $P_i \overset{d}{=} 1 - P_i$ under the null. The application of this estimator to the conditional setting (3) allows them to mask information, uncover it in a step-wise fashion and thus avoid overfitting. Thus, their iterative procedure ultimately has the same motivation as IHW's hypothesis splitting, but simultaneously respects the sufficiency principle.

Using a BH- versus a BC-type FDR estimator also has some further practical implications: The procedures proposed in this paper are valid under the standard multiple-testing conservativeness assumption (super-uniformity of the null p-values), while AdaPT requires a new, distinct notion, which the authors call mirror conservativeness (neither implies the other). In addition, AdaPT suffers from discretization issues due to the BC-estimator. For example, for $\alpha = 0.1$, AdaPT can only reject 10 or more hypotheses; or no hypotheses at all. IHW is not affected by this.

6.4. Connection to local false discovery rates

The idea that one loses power and interpretability by reducing each hypothesis to a single number (the p-value) has been prominent in multiple testing literature which employs the local false discovery rate (fdr) [35, 44]. In particular, multiple authors have considered variants of the conditional two-groups model (3) wherein p-values and covariates are available [11, 17–19, 33, 43, 47, 55]. Under the conditional two-groups model (3), the conditional local fdr is defined as follows (f being the Lebesgue density of the marginal p-value distribution F):

$$\text{fdr}(t \mid x) = \frac{\pi_0(x)}{f(t \mid x)}$$

In the present context, it has two important properties: It can be used to estimate the false discovery rate and the solution of optimization problem (7) has contours of equal fdr (under mild assumptions). See Appendix A for a more detailed discussion.

Most existing methods cited above make use of both of these properties: Their power is high and they are asymptotically consistent; however FDR control is often jeopardized for finite samples, especially when the conditional two-groups model is misspecified [26] or not estimated correctly. Methods such as AdaPT [27] and IHW provide a framework that enables use of fdr (and the flexibility this entails) while simultaneously guaranteeing finite-sample FDR control under broad conditions.

6.5. Connection to a-priori weighting

A parallel research avenue is to figure out a-priori weights, that is without using the p-values from the study at hand, but from a previous study [13, 14, 20, 38]. Such methods have the advantage of not having to mask information (e.g. as in AdaPT or by randomization) to avoid overfitting and instead control type-I error automatically by the results in [21]. From a practical point of view, these methods are particularly relevant when one expects at most a handful of discoveries even after optimally weighting (say when going from 0 discoveries with an unweighted procedure to 5 discoveries with weights). In such situations, the available signal might not be strong enough for successful modeling of the conditional distribution under masking.

6.6. Connection to data-splitting

One of the initial attempts at data-driven weights [41] also used randomization in the form of data-splitting: Again consider the setting where we start with a $m \times n$ data-matrix from which we get our p-values by calculating the statistic in a row-wise fashion. Then one can calculate "prior" p-values P_i'' based on n_1 columns and use the other n_2 ($n_1 + n_2 = n$) columns for the actual p-values P_i'. Hence, as in Subsection 6.5 one can derive prior weights based on $(P_i'')_i$ and apply the weighted procedure with $(P_i')_i$. However, the authors then show that in this case it is more powerful to simply use an unweighted procedure with p-values calculated based on the whole dataset, rather than a weighted procedure with data-splitting.

For IHW we instead randomize horizontally rather than vertically, and the final p-values stay the same (only the weights are randomized). Nevertheless, vertical randomization might still be a fruitful idea and we sketch an approach whose validity hinges on Corollary 1. We leverage the well-known "doubling" trick (see [24] for another application): Given $Z_1, Z_2 \sim \mathcal{N}(0,1)$ independent and $\alpha > 0$, then: $Z_1 + \alpha Z_2, Z_1 - \frac{1}{\alpha} Z_2$ are also independent. Now take a null p-value P_i with $P_i \sim U[0,1]$, then $\Phi^{-1}(P_i) \sim \mathcal{N}(0,1)$. Generate $Z_1, \ldots, Z_m$ i.i.d. $\mathcal{N}(0,1)$ and independent of $(\mathbf{P}, \mathbf{X})$. Consider the p-values (Φ is the standard Normal distribution function):

$$P_i' = \Phi\left(\frac{\Phi^{-1}(P_i) + \alpha Z_i}{\sqrt{1+\alpha^2}}\right), P_i'' = \Phi\left(\frac{\Phi^{-1}(P_i) - \frac{1}{\alpha} Z_i}{\sqrt{1 + \frac{1}{\alpha^2}}}\right)$$

Now use $(P''_i, X_i)_i$ to learn the weighting function, then apply the IHWoracle procedure with $(P'_i, X_i)_i$. $\alpha > 0$ is a parameter controlling the trade-off between information made available to the weighting procedure and power lost for the alternative p-values.

7. Main proofs

7.1. Theorem 2: IHWc controls the FDR

Proof of Theorem 2. We start by following closely the proof in [28], as adapted to the current setting. Let $\mathbf{W}$ be the weight vector and $\hat{k}$ the number of discoveries after applying IHWc at level α. Also write $\mathbf{X} = (X_1, \dots, X_m)$, $\mathbf{P} = (P_1, \dots, P_m)$ and $\mathbf{1}_{\{\mathbf{P} \leq \tau\}} = (\mathbf{1}_{\{P_1 \leq \tau\}}, \dots, \mathbf{1}_{\{P_m \leq \tau\}})$. For $i \in \{1, \dots, m\}$, denote by k_i the number of discoveries of IHWc when $\mathbf{P}$ gets replaced by $\mathbf{P}_{i \to 0} = (P_1, \dots, P_{i-1}, 0, P_{i+1}, P_m)$.

Note that because of (IHWc1), the weight vector $\mathbf{W}$ remains unchanged on the event $\{P_i \leq \tau\}$. Furthermore, because of (IHWc2), H_i rejected implies that $P_i \leq \left(\frac{\alpha W_i k}{m}\right) \wedge \tau$. In particular, the event $\{P_i \leq \tau\}$ holds. Hence, for any $k \geq \hat{k}$, counting the entries of $\mathbf{P}$, respectively $\mathbf{P}_{i \to 0}$, which are not greater than the corresponding entries of $\left(\frac{\alpha \mathbf{W} k}{m}\right) \wedge \tau$ must yield the same number.

We conclude that:

$$H_i \text{ rejected} \Rightarrow \hat{k} = k_i \geq 1$$

Therefore:

$$H_i \text{ rejected} \Rightarrow P_i \leq \frac{\alpha W_i k_i}{m} \wedge \tau$$

Note at this point that we can assume without loss of generality that $\mathbb{P}[P_i \leq \tau] > 0$ for all $i \in \mathscr{H}_0$. Otherwise, just set $\mathscr{H}_0' = \{i \in \mathscr{H}_0 \mid \mathbb{P}[P_i \leq \tau] > 0\}$ and all the steps below will go through essentially unchanged with $\mathscr{H}_0'$ replacing $\mathscr{H}_0$.

For $i \in \mathscr{H}_0$ and conditioning on the event $\{P_i \leq \tau\}$ and on the random vectors $\mathbf{W}, \mathbf{X}, \mathbf{P}_{i \to 0}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}$ we get:

$$\begin{aligned}
&\mathbb{P}[H_i \text{ rejected} \mid P_i \leq \tau, \mathbf{W}, \mathbf{X}, \mathbf{P}_{i \to 0}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}] \\
\leq &\mathbb{P}[P_i \leq \frac{\alpha W_i k_i}{m} \wedge \tau \mid P_i \leq \tau, \mathbf{W}, \mathbf{X}, \mathbf{P}_{i \to 0}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}] \\
\leq &\frac{\alpha W_i k_i}{m \mathbb{P}[P_i \leq \tau]}
\end{aligned}$$

This follows because for $i \in \mathscr{H}_0$ it holds that P_i is (super)uniform, $\mathbb{P}[P_i \leq \tau] > 0$ and P_i is independent of $(\mathbf{P}_{i \to 0}, \mathbf{X})$ and also because k_i, $\mathbf{W}$, $\mathbf{1}_{\{\mathbf{P} \leq \tau\}}$ are functions of $(\mathbf{P}_{i \to 0}, \mathbf{X})$ on the event $\{P_i \leq \tau\}$.

It then follows that:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbf{1}_{\{H_i \text{ rejected}\}}}{\hat{k} \vee 1} \mid P_i \leq \tau, \mathbf{W}, \mathbf{X}, \mathbf{P}_{i \to 0}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}\right] \\
= &\mathbb{E}\left[\frac{\mathbf{1}_{\{H_i \text{ rejected}\}}}{k_i \vee 1} \mid P_i \leq \tau, \mathbf{W}, \mathbf{X}, \mathbf{P}_{i \to 0}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}\right] \\
\leq &\frac{\alpha W_i}{m \mathbb{P}[P_i \leq \tau]}
\end{aligned}$$

Moreover, by marginalization over $\mathbf{P}_{i \to 0}$ and $\mathbf{X}$ (and noting again that $\mathbf{1}_{\{H_i \text{ rejected}\}} = 0$ for $\mathbf{1}_{\{P_i \leq \tau\}} = 0$), we get:

$$\mathbb{E}\left[\frac{\mathbf{1}_{\{H_i \text{ rejected}\}}}{\hat{k} \vee 1} \mid \mathbf{W}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}\right] \leq \frac{\alpha W_i}{m\mathbb{P}[P_i \leq \tau]}\mathbf{1}_{\{P_i \leq \tau\}}$$

In total we thus get:

$$\mathbb{E}[\text{FDP} \mid \mathbf{W}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}] = \mathbb{E}\left[\frac{\sum_{i \in \mathscr{H}_0} \mathbf{1}_{\{H_i \text{ rejected}\}}}{\hat{k} \vee 1} \mid \mathbf{W}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}\right] \leq \sum_{i \in \mathscr{H}_0} \frac{\alpha W_i}{m\mathbb{P}[P_i \leq \tau]}\mathbf{1}_{\{P_i \leq \tau\}}$$

At this point we can diverge from [28] and use the hypothesis splitting property (in the form of Lemma 1).

$$\begin{aligned}
\mathbb{E}[\text{FDP}] &= \mathbb{E}[\mathbb{E}[\text{FDP} \mid \mathbf{W}, \mathbf{1}_{\{\mathbf{P} \leq \tau\}}]] \\
&\leq \sum_{i \in \mathscr{H}_0} \mathbb{E}\left[\frac{\alpha W_i}{m\mathbb{P}[P_i \leq \tau]}\mathbf{1}_{\{P_i \leq \tau\}}\right] \\
&= \sum_{i \in \mathscr{H}_0} \frac{\alpha}{m\mathbb{P}[P_i \leq \tau]}\mathbb{E}\left[W_i\right]\mathbb{E}\left[\mathbf{1}_{\{P_i \leq \tau\}}\right] \\
&\leq \frac{\alpha}{m}\mathbb{E}\left[\sum_{i=1}^{m} W_i\right] \\
&= \alpha
\end{aligned}$$

Going from the second to the third line, we used that for $i \in \mathscr{H}_0$, P_i is independent of W_i.

□

7.2. Theorem 3: IHW-Bonferroni controls the FWER

Proof.

$$\begin{aligned}
\text{FWER}_{\text{IHW-Bonferroni}} &= \mathbb{P}\left[\bigcup_{i \in \mathscr{H}_0}\left\{P_i \leq \frac{\alpha W_i}{m}\right\}\right] \\
&\leq \sum_{i \in \mathscr{H}_0} \mathbb{P}\left[P_i \leq \frac{\alpha W_i}{m}\right] \\
&= \sum_{i \in \mathscr{H}_0} \mathbb{E}\left[\mathbb{P}\left[P_i \leq \frac{\alpha W_i}{m} \mid W_i\right]\right] \\
&\leq \sum_{i \in \mathscr{H}_0} \mathbb{E}\left[\frac{\alpha W_i}{m}\right] \\
&\leq \frac{\alpha}{m}\mathbb{E}\left[\sum_{i \in \mathscr{H}_0} W_i\right] \\
&\leq \alpha
\end{aligned}$$

Note that going from the third to the fourth line, we used the fact that for $i \in \mathscr{H}_0$ it holds that P_i is (super)uniform and also by a simple modification of Lemma 1 it also holds that P_i is independent of W_i. □

Acknowledgments

The authors would like to thank Edgar Dobriban, Michael Sklar and Britta Velten for helpful discussions and critical comments.

References

[1] Alishahi, K., Ehyaei, A. R. and Shojaie, A. (2016). A Generalized Benjamini-Hochberg Procedure for Multivariate Hypothesis Testing. *arXiv preprint arXiv:1606.02386.*

[2] Arias-Castro, E. and Chen, S. (2016). Distribution-free Multiple Testing. *arXiv preprint arXiv:1604.07520.*

[3] Barber, R. F., Candès, E. J. et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085.

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

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 289–300.

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

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

[8] Blanchard, G., Roquain, E. et al. (2008). Two simple sufficient conditions for FDR control. *Electronic journal of Statistics* **2** 963–992.

[9] Boca, S. M. and Leek, J. T. (2017). A regression framework for the proportion of true null hypotheses. *bioRxiv.*

[10] Bourgon, R., Gentleman, R. and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences* **107** 9546–9551.

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

[12] Cairns, J., Freire-Pritchett, P., Wingett, S. W., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.-M., Osborne, C., Fraser, P. et al. (2015). CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *bioRxiv* 028068.

[13] Dobriban, E. (2016). A general convex framework for multiple testing with prior information. *arXiv preprint arXiv:1603.05334.*

[14] Dobriban, E., Fortney, K., Kim, S. K. and Owen, A. B. (2015). Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika* **102** 753–766.

[15] Du, L. and Zhang, C. (2014). Single-index modulated multiple testing. *The Annals of Statistics* **42** 30–79.

[16] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science* **349** 636–638.

[17] Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction.* Cambridge University Press.

[18] Efron, B. and Zhang, N. R. (2011). False discovery rates and copy number variation. *Biometrika* **98** 251–271.

[19] Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. and Kong, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics* 714–735.

[20] Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B. and Kim, S. K. (2015). Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genetics* **11** e1005728.

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

[22] Habiger, J., Watts, D. and Anderson, M. (2015). Multiple Testing with Heterogeneous Multinomial Distributions. *arXiv preprint arXiv:1511.01400*.

[23] Habiger, J. D. (2014). Weighted Adaptive Multiple Decision Functions for False Discovery Rate Control. *arXiv preprint arXiv:1412.0645*.

[24] Harris, X. T. (2016). Prediction error after model search. *arXiv preprint arXiv:1610.06107*.

[25] Hu, J. X., Zhao, H. and Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association* **105**.

[26] Ignatiadis, N., Klaus, B., Zaugg, J. B. and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*.

[27] Lei, L. and Fithian, W. (2016). AdaPT: An interactive procedure for multiple testing with side information. *arXiv preprint arXiv:1609.06035*.

[28] Li, A. and Barber, R. F. (2016). Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *arXiv preprint arXiv:1606.07926*.

[29] Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., Lin, X. and Liang, L. (2013). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in Genetics* **4**.

[30] Liu, W. (2014). Incorporation of Sparsity Information in Large-scale Multiple Two-sample t Tests. *arXiv preprint arXiv:1410.4282*.

[31] Mammen, E., Marron, J. S., Turlach, B., Wand, M. et al. (2001). A general projection framework for constrained smoothing. *Statistical Science* **16** 232–248.

[32] Markovic, J. and Taylor, J. (2016). Bootstrap inference after using multiple queries for model selection. *arXiv preprint arXiv:1612.07811*.

[33] Ochoa, A., Storey, J. D., Llinás, M. and Singh, M. (2015). Beyond the E-Value: Stratified Statistics for Protein Domain Prediction. *PLoS Computational Biology* **11** e1004509.

[34] Peña, E. A., Habiger, J. D. and Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *The Annals of Statistics* **39** 556–583.

[35] Ploner, A., Calza, S., Gusnanto, A. and Pawitan, Y. (2006). Multidimensional local false discovery rate for microarray studies. *Bioinformatics* **22** 556–565.

[36] Roeder, K., Bacanu, S.-A., Wasserman, L. and Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics* **78** 243–252.

[37] Roeder, K., Devlin, B. and Wasserman, L. (2007). Improving power in genome-wide association studies: weights tip the scale. *Genetic Epidemiology*

31 741–747.

[38] Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical Science* **24** 398.

[39] Roquain, E. and Van De Wiel, M. (2008). Multi-weighting for FDR control. *arXiv preprint math.ST/0807.4081*.

[40] Roquain, E. and Van De Wiel, M. (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics* **3** 678–711.

[41] Rubin, D., Dudoit, S. and Van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology* **5**.

[42] Sankaran, K. and Holmes, S. (2014). structSSI: Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data. *Journal of statistical software* **59** 1.

[43] Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association* **110** 459-471.

[44] Stephens, M. (2016). False Discovery Rates: A New Deal. *Biostatistics*.

[45] Storey, J. D., Taylor, J. E. and Siegmund, D. (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.

[46] Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9** 303.

[47] Tansey, W., Koyejo, O., Poldrack, R. A. and Scott, J. G. (2014). False discovery rate smoothing. *arXiv preprint arXiv:1411.6144*.

[48] Tian, X., Bi, N. and Taylor, J. (2016). MAGIC: a general, powerful and tractable method for selective inference. *arXiv preprint arXiv:1607.02630*.

[49] Tibshirani, R. (2012). Screening and False Discovery Rates. https://normaldeviate.wordpress.com/2012/12/01/screening-and-false-discovery-rates/. [Online; accessed 16-June-2016].

[50] Tibshirani, R. J. and Efron, B. (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology* **1**.

[51] Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98** 5116–5121.

[52] Wasserman, L. (2014). Discussion: "A significance test for the lasso". *The Annals of Statistics* **42** 501–508.

[53] Wiel, M. A., Lien, T. G., Verlaat, W., Wieringen, W. N. and Wilting, S. M. (2016). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine* **35** 368–381.

[54] Xing, C., Cohen, J. C. and Boerwinkle, E. (2010). A weighted false discovery rate control procedure reveals alleles at FOXA2 that influence fasting glucose levels. *The American Journal of Human Genetics* **86** 440–446.

[55] Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M. and Thompson, W. K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* btu145.

[56] Zhao, H. and Zhang, J. (2014). Weighted p–value procedures for controlling FDR of grouped hypotheses. *Journal of Statistical Planning and Inference*.

Appendix A: Local false discovery rate methods

To introduce the local fdr, we revisit the two-groups model (1). Recall that the Bayesian Fdr was defined as the posterior probability $\mathbb{P}[H_i = 0 \mid P_i \leq t]$. In similar spirit, assuming that the marginal distribution has Lebesgue density $f(\cdot)$, we define:

$$\text{fdr}(t) := \mathbb{P}[H_i = 0 \mid P_i = t] = \frac{\pi_0}{f(t)}$$

The local fdr has the intuitive interpretation, that it assigns a posterior probability to each hypothesis, while the Bayesian Fdr instead assigns a posterior probability to the all hypotheses below a threshold. These two concepts however are related as follows:

$$\begin{aligned}
\mathbb{E}[\text{fdr}(P_i) \mid P_i \leq t] &= \mathbb{E}\left[\frac{\pi_0}{f(P_i)} \mid P_i \leq t\right] \\
&= \int_{[0,1]} \frac{\pi_0}{f(u)} \frac{f(u)\mathbf{1}_{\{u \leq t\}}}{F(t)} du \\
&= \frac{\pi_0}{F(t)} \int_{[0,1]} \mathbf{1}_{\{u \leq t\}} du \\
&= \frac{\pi_0 t}{F(t)} \\
&= \text{Fdr}(t) \\
&= \mathbb{P}[H_i = 0 \mid P_i \leq t]
\end{aligned}$$

Extending to the conditional two-groups model (3) for which the conditional distribution $F(t \mid x)$ has Lebesgue density $f(t \mid x)$ for all x, we analogously define the conditional local fdr:

$$\text{fdr}(t \mid x) = \frac{\pi_0(x)}{f(t \mid x)}$$

Analogously to above, we get for a threshold function g:

$$\mathbb{E}[\text{fdr}(P_i \mid X_i) \mid P_i \leq g(X_i)] = \frac{\int_{\mathcal{X}} \pi_0(x) g(x) d\mathbb{P}^X(x)}{\int_{\mathcal{X}} F(g(x)|x) d\mathbb{P}^X(x)} = \mathbb{E}[H_i = 0 \mid P_i \leq g(X_i)] \quad (12)$$

In addition, conditional local fdrs are related to the solution of optimization problem (7), when $\mathcal{G} = \{g : \mathcal{X} \to [0, 1] \text{ measurable}\}$. Under certain regularity conditions, all the conditional local fdrs must be equal in the optimal solution [33], i.e.:

$$\text{fdr}(g(X_i) \mid X_i) = \text{fdr}(g(X_j) \mid X_j) \quad \forall\, i, j \in \{1 \ldots, m\} \quad (13)$$

To show this, we provide a very informal Lagrange Multiplier argument (see [27] for a more rigorous treatment), starting from the equivalent form of the optimization problem in (8). Thus let $g(X_i) = t_i$ and define the Lagrangian:

$$L(t_1, \ldots, t_m, \lambda) = \sum_{i=1}^{n} F(t_i \mid X_i) + \lambda \sum_{i=1}^{m} (\pi_0(X_i) t_i - \alpha F(t_i \mid X_i))$$

Then:

$$\frac{\partial L}{\partial t_i} = f(t_i \mid X_i) + \lambda(\pi_0(X_i) - \alpha f(t_i \mid X_i))$$

Setting this equal to 0 and dividing by $f(t_i \mid X_i)$, we get:

$$\frac{\pi_0(X_i)}{f(t_i \mid X_i)} = \alpha - \frac{1}{\lambda}$$

i.e., $\forall\ i$ we have: $\text{fdr}(t_i \mid X_i) = \alpha - \frac{1}{\lambda}$

Now assume that the conditional two-groups model (3) holds with continuously differentiable conditional distributions and is known to an oracle. According to equation (13), we should rank the hypotheses by $\text{fdr}(P_i \mid X_i)$ rather than P_i. Equation (12) implies that we can estimate Fdr (and hence also FDR) of a procedure with decision threshold g by:

$$\widehat{\text{Fdr}}(g) = \frac{\sum_{i=1}^m \text{fdr}(P_i \mid X_i)\mathbf{1}_{\{P_i \leq g(X_i)\}}}{\sum_{i=1}^m \mathbf{1}_{\{P_i \leq g(X_i)\}}} \tag{14}$$

Putting these two ideas together, we get the oracle procedure in Algorithm 4.

Algorithm 4: The conditional local fdr procedure

```
Input: A nominal level $\alpha \in (0,1)$ and a vector of p-values $P_1, \dots, P_m$ and covariates
       $X_1, \dots, X_m$.
Let $\text{Cfdr}_i := \text{fdr}(P_i \mid X_i)$
Let $\text{Cfdr}_{(1)}, \dots, \text{Cfdr}_{(m)}$ be the order statistics of $\text{Cfdr}_1, \dots, \text{Cfdr}_m$ and let $\text{Cfdr}_{(0)} := 0$
Let $k^* = \max\left\{k \mid \frac{1}{k}\sum_{i=1}^k \text{Cfdr}_{(i)} \leq \alpha\ ,\ 1 \leq k \leq m\right\}$, if the latter set is empty, let $k^* = 0$.
Reject all hypotheses with $\text{Cfdr}_i \leq \text{Cfdr}_{(k^*)}$
```

Such a procedure indeed controls the FDR [11], when the conditional two-groups model (3) is true and the oracle has access to the true model. Data-driven approximations to this procedure can be developed by plugging in estimates of the conditional densities $f(t \mid x)$ and $\pi_0(\cdot)$ [11].

While such a procedure can be shown to be asymptotically consistent, unlike the result for the IHWoracle procedure in Corollary 1, this procedure is not robust towards misspecification of the conditional two-groups model and no finite-sample results are available. Even when (3) is true, the difficulty of estimating conditional densities, especially at the tails of the distribution, can make the estimator (14) an unreliable choice compared to estimators based only on the empirical cumulative distribution function.

Nevertheless, the local fdr, even if not estimated perfectly, can be used successfully within procedures such as IHW or AdaPT ([27]). Furthermore, given the Bayesian interpretation of the (conditional) local fdr and its appeal for ranking hypotheses, it is a very important task to develop practical estimators [44].