Covariate powered cross-weighted multiple testing

Nikolaos Ignatiadis[1]
Wolfgang Huber[2]

1 Department of Statistics, Stanford University, USA
ignat@stanford.edu

2 European Molecular Biology Laboratory, Heidelberg, Germany
wolfgang.huber@embl.org

Summary

A fundamental task in the analysis of datasets with many variables is screening for associations. This can be cast as a multiple testing task, where the objective is achieving high detection power while controlling type I error. We consider m hypothesis tests represented by pairs $((P_i, X_i))_{1 \leq i \leq m}$ of p-values P_i and covariates X_i, such that $P_i \perp X_i$ if H_i is null. Here, we show how to use information potentially available in the covariates about heterogeneities among hypotheses to increase power compared to conventional procedures that only use the P_i. To this end, we upgrade existing weighted multiple testing procedures through the Independent Hypothesis Weighting (IHW) framework to use data-driven weights that are calculated as a function of the covariates. Finite sample guarantees, e.g., false discovery rate (FDR) control, are derived from cross-weighting, a data-splitting approach that enables learning the weight-covariate function without overfitting as long as the hypotheses can be partitioned into independent folds, with arbitrary within-fold dependence. IHW has increased power compared to methods that do not use covariate information. A key implication of IHW is that hypothesis rejection in common multiple testing setups should not proceed according to the ranking of the p-values, but by an alternative ranking implied by the covariate-weighted p-values.

Keywords: Benjamini-Hochberg, Empirical Bayes, False Discovery Rate, Independent Hypothesis Weighting, Multiple Testing, p-value weighting

1 Introduction

Screening large datasets for interesting associations is a basic operation in statistical data analysis. A frequently taken approach is to enumerate all potential associations, set up a hypothesis test for each of them, summarize the results by the p-values P_i, and select as *discoveries* all hypotheses with a small enough p-value; typically, this is a small fraction of all hypotheses. More formally, for some cutoff $\hat{t}$:

$$\text{Reject hypothesis } i \iff P_i \leq \hat{t} \tag{1}$$

The choice of the cutoff $\hat{t}$ may be data-driven and is determined by a multiple testing procedure, such as those proposed by Bonferroni [1935] or Benjamini and Hochberg [1995], which compute a $\hat{t}$ that provides a defined level of protection against spurious discoveries. Common objectives are control of the family-wise error rate (FWER) or the false discovery rate (FDR).

These procedures operate solely on the list of p-values. Here, we consider situations in which beyond the p-value P_i, side information represented by a covariate X_i is available for each hypothesis. Such side-information reflects heterogeneity among the tests and may —more or less directly—carry information about their different power, or the different prior probabilities of their null hypothesis being true. Suitable covariates are often apparent to domain scientists or to statisticians. We will see that procedures that take into account such side information often have higher power, in the sense that they make more discoveries at the same level of type-I error.

To illustrate, we use a high-throughput genetics dataset by Grubert et al. [2015], who aimed to discover associations between genetic polymorphisms (SNPs) in the human genome and the activity of genomic regions (H3K27ac peaks). The main idea of the analysis of these data, which is presented in more detail in Section 6, is to carry out a hypothesis test for each pair of SNP and region on the same chromosome. On Chromosomes 1 and 2, $N_1 = 645452$ and $N_2 = 699343$ SNPs were recorded, and H3K27ac levels were measured in $K_1 = 12193$ and $K_2 = 11232$ regions, which amounts to nearly 16 billion ($N_1 K_1 + N_2 K_2$) tests. Figure 1 illustrates how the p-value distributions differ as a function of the genomic distance between SNP and region. These differences are consistent with biological domain knowledge: associations across shorter distances are a-priori more plausible and empirically more frequent. Methods that are able to take into account this heterogeneity among the tests should be able to discover more associations at the same FDR, compared to (1), which ignores such side information.

1.1 Independent Hypothesis Weighting

In this paper, we present Independent Hypothesis Weighting (IHW), a flexible framework that can leverage hypothesis heterogeneity to improve power, while retaining finite-sample type-I error control. To explain the method, consider testing m hypotheses $H_1, \ldots, H_m$ based on p-values $P_1, \ldots, P_m$ in the situation where we also have access to covariates $X_1, \ldots, X_m$ such that each X_i is independent of the p-value P_i if H_i is a null hypothesis; the codomain of the X_i can be any space (the same for all i). We propose to use a decision rule of the following form in place of (1):

$$\text{Reject hypothesis } i \iff P_i \leq \hat{t} \cdot \widehat{W}^{-\ell}(X_i) \qquad (\text{where } i \in I_\ell), \tag{2}$$

where I_ℓ, $\ell = 1, \ldots, K$ is a partition of the hypotheses into K disjoint folds, such that the (P_i, X_i) pairs are independent across folds.

There are two salient features to this rule: first, the decision boundary of hypothesis i does not only depend on its p-value P_i and the overall cutoff $\hat{t}$, but also on the weight function $\widehat{W}^{-\ell} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ of the covariate X_i, where $\mathcal{X}$ is the codomain of the X_i, and there is one such function for each fold I_ℓ. Second, the notation $\widehat{W}^{-\ell}$ is used to denote that each of these functions is learned from the data with the proviso that only p-values and covariates from the $K-1$ folds excluding I_ℓ are used. We call this proviso *cross-weighting*.

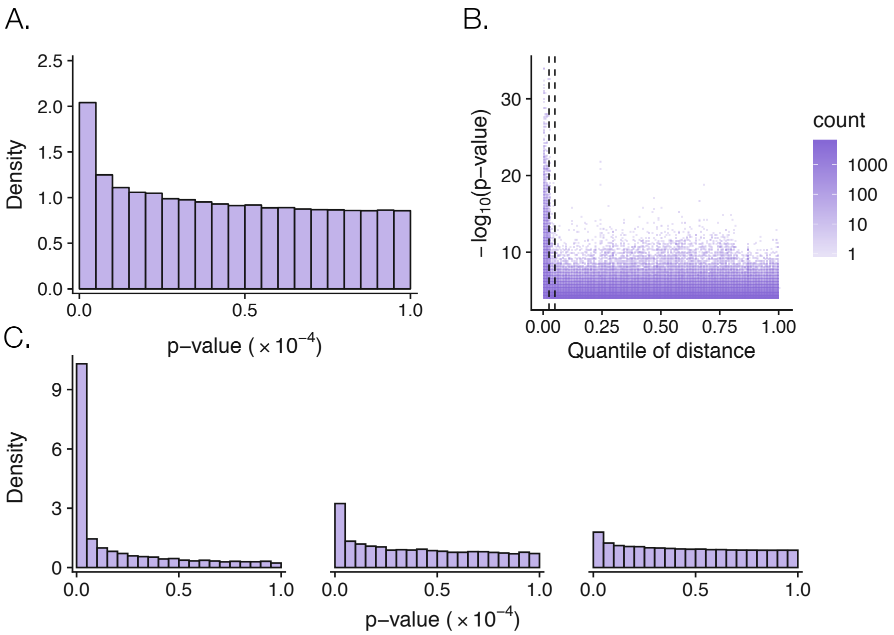


Figure 1: **Heterogeneous multiple hypothesis testing in a biological example.** For each hypothesis ($i = 1, \dots m$), a p-value P_i is provided as well as a covariate X_i, which here is the genomic distance between the two features tested for association: a single nucleotide polymorphism (SNP) and a biochemical chromatin modification. **A. Histogram of p-values**: We recognize the peak close to the origin, corresponding to enrichment of alternative hypotheses, and a near-uniform tail for larger p-values. Note that the displayed p-values are right-censored at 10^{-4}, as is further explained in Section 6, which provides more context on the data. **B. Two-dimensional heatmap of bin counts** of the joint empirical distribution $(-\log_{10} P_i, X_i)$: small p-values are enriched at lower distances. **C. Histograms of p-values stratified by the covariate:** Upon partitioning our hypotheses at the boundaries denoted by dashed lines in panel B, we observe that at small distances the signal (peak at the left of the histograms) is pronounced, while for larger distances the histogram is dominated by background (uniform distribution of p-values from true null hypotheses).

Conceptually, cross-weighting is related to cross-fitting [Schick, 1986], a method that has been successful in the fields of causal inference [Nie and Wager, 2020, Chernozhukov et al., 2017] and empirical Bayes [Ignatiadis and Wager, 2019] for estimation with high-dimensional nuisance parameters. Analogous to findings in the cross-fitting literature, we will show that naively using plug-in estimators to obtain the weight function tends to overfit, but cross-weighting salvages this at essentially no cost.

1.2 Related work

Previous work has shed light on optimal discovery thresholds in heterogeneous multiple testing. Similar to (2), these thresholds may take the form $\{P_i \leq \hat{t}w_i\}$ parametrized in terms of weights w_i that are optimal for controlling the family-wise error rate (FWER) [Roeder and Wasserman, 2009, Dobriban et al., 2015] or the false discovery rate (FDR) [Roquain and Van De Wiel, 2009, Durand, 2019]. Furthermore, in the case of FDR control, optimal decision thresholds are known to take the form of contours of equal local false discovery rate [Cai and Sun, 2009, Cai et al., 2019, Efron, 2010, Ferkingstad et al., 2008, Ochoa et al., 2015, Ploner et al., 2006]. Nevertheless, all of these optimal procedures are not implementable, as they depend on unknown properties of the data-generating mechanism. Instead, it has been proposed to apply a plug-in principle: the thresholds are estimated from the data at hand.

Such plug-in approaches however have no guarantees of type-I error control or only do so in an asymptotic limit, as the number of tested hypotheses goes to infinity [Cai and Sun, 2009, Cai et al., 2019, Durand, 2019, Ignatiadis et al., 2016]. More importantly, with finite samples, these plug-in methods often exceed the claimed type-I error; we will demonstrate this in Sections 2.1 and 5. This has motivated the provision of case-by-case, ad-hoc modifications, which however still do not provide finite-sample guarantees. For example, Durand [2017] recommends conducting a global test first and only proceeding with multiple testing if the global null hypothesis can be rejected. Cai, Sun, and Wang [2019] use a conservative modification of the density estimator employed by their (asymptotically valid) plug-in approach and show that this controls FDR in simulations with sparse signals. Furthermore, they suggest using the global screen of Durand [2017] first. Ignatiadis, Klaus, Zaugg, and Huber [2016] use cross-weighting (described above) as a heuristic to maintain FDR control in finite samples.

Dispensing with heuristics, several authors have recently provided procedures that are formally justified under full independence of the hypotheses: Li and Barber [2019] propose SABHA, a data-driven, weighted procedure for FDR control which directly confronts potential overfitting. The authors prove finite sample FDR control at an elevated level compared to the nominal α; i.e., at $(1+\varepsilon)\alpha$ for some $\varepsilon > 0$. However, their guarantee only applies for their specific weighting scheme, which furthermore is suboptimal even under knowledge of the data-generating process [Lei and Fithian, 2018]. Zhang, Xia, Zou, and Tse [2017] and Zhang, Xia, and Zou [2019] use a variant of hypothesis splitting to guarantee high-probability bounds on the false discovery proportion, however their proposals require a minimum number of rejections; otherwise an empty list of discoveries is declared. Closer to our approach is AdaPT [Lei and Fithian, 2018], which uses covariate information to learn covariate-modulated decision boundaries and provides finite sample FDR guarantees. Its construction is based on a variant of the optimal stopping theorem developed by Barber and Candès [2015], which provides the analyst with considerable flexibility in learning these boundaries from the data, while masking information that could lead to overfitting. However, AdaPT has no theoretical guarantees outside of full p-value independence, is tied to FDR control and suffers from a large variance of the false discovery proportion [Korthauer et al., 2019].

Here we propose a general and flexible framework that goes beyond these previous approaches. We formalize hypothesis weighting with weights as a function of covariates X_i and demonstrate that such weights can be learned from the data without overfitting (i.e., losing type-I error control) if we use cross-weighting as in (2). Hence we build upon the hypothesis-splitting idea of Ignatiadis et al. [2016] and demonstrate that it can be used not merely as a heuristic, but instead as a theoretically grounded and principled way of conducting multiple testing with side-information that has far reaching applications. The Independent Hypothesis Weighting method provides finite sample guarantees for multiple type-I measures, such as the FDR, the FWER and the k-FWER, unlike previous proposals that are tied to the FDR. IHW provides a clean way to deal with dependent settings, as it allows arbitrary dependence within folds. Finally, IHW provides the researcher with flexibility in choosing any weighting scheme that would be appropriate for the data at hand, but we also recommend a default scheme and provide a software implementation in the form of an R package.

1.3 Outline

In Section 2, we provide an overview of weighted multiple testing and explain our proposal in the context of FDR control under full independence of hypothesis tests. Section 3 extends the results to dependence, and to control of the k-FWER. Section 4 describes a framework for learning weighting rules. Section 5 provides simulation results, and Section 6 presents the high-throughput biology example from Figure 1. Section 7 discusses further relationships to previous work, and Section 8 concludes with a discussion.

2 Weighted and cross-weighted multiple testing

A multiple testing procedure operates on data for m hypotheses $H_1, \ldots, H_m$ and declares R hypotheses as rejections ("discoveries"). Among these, V will be nulls, i.e., the procedure will commit V type I errors. The goal is to make as many discoveries as possible while retaining (stochastic) guarantees that V is acceptable. Concretely, one possible objective is to control the family-wise error rate, defined as FWER $:= \mathbb{P}[V \geq 1]$, or the k-FWER $:= \mathbb{P}[V \geq k]$. In exploratory situations, a typically less stringent objective is to control the false discovery rate (FDR), i.e., the expectation of the false discovery proportion (FDP), namely FDR $:= \mathbb{E}[\text{FDP}] := \mathbb{E}\left[\frac{V}{R \vee 1}\right]$ [Benjamini and Hochberg, 1995].

Typically the data for each hypothesis are summarized into a single number, the p-value P_i, and a rule of form (1) is applied. However, in the presence of heterogeneity across tests, it might be suboptimal to use such a decision rule that treats all hypotheses exchangeably. Weighted multiple testing [Genovese, Roeder, and Wasserman, 2006] is a flexible way of encoding prior information and differentially prioritizing the hypotheses. Multiple testing weights are defined as non-negative numbers w_i such that $\sum_{i=1}^m w_i/m = 1$. Then, a weighted multiple testing decision rule takes the following form:

$$\text{Reject hypothesis } i \iff P_i \leq \min\{w_i \cdot \hat{t}, \tau\} \tag{3}$$

Here $\tau \in (0, 1]$ is a fixed number, of which more below, and as in (1), the cut-off $\hat{t}$ may be data-driven. A larger w_i implies that it is easier to reject hypothesis i. We first review two procedures for choosing $\hat{t}$.

Definition 1 (Weighted k-Bonferroni)**.** The k-FWER can be controlled at level $\alpha \in (0, 1)$ by applying the weighted k-Bonferroni procedure [Romano and Wolf, 2010], which takes the form (3) with deterministic cutoff $\hat{t} = k\alpha/m$ and $\tau = 1$. The case $k = 1$ is the weighted Bonferroni procedure proposed by Genovese et al. [2006].

Definition 2 (τ-censored, weighted Benjamini-Hochberg)**.** The FDR can be controlled at level $\alpha \in (0, 1)$ by applying the τ-censored, weighted Benjamini-Hochberg procedure, which takes the form (3) with $\tau \in (0, 1]$ fixed and data-driven cutoff $\hat{t}$ specified as:

$$\hat{t} = \frac{\alpha \hat{k}}{m}, \quad \hat{k} = \max\left\{k \in \mathbb{N}_{\geq 0} \mid P_i \leq \left(\frac{\alpha w_i k}{m}\right) \wedge \tau \text{ for at least } k \text{ p-values}\right\} \tag{4}$$

The weighted Benjamini-Hochberg (BH) procedure of Genovese, Roeder, and Wasserman [2006] is the special case $\tau = 1$. The more general form was proposed by Li and Barber [2019] and will be employed for our theoretical guarantees in the following. The number of rejections of τ-censored BH is non-decreasing in τ, so that a procedure with smaller τ will never make more discoveries. However, for large τ, say $\tau \geq 0.5$, the discovery set will be equal to that with $\tau = 1$, as long as weighted BH with $\tau = 1$ did not reject a p-value ≥ 0.5.

In decision rule (3), the weights w_i are denoted by lower-case letters. This reflects the fact that existing results treat these weights as deterministic—as prior knowledge that a researcher has to specify before seeing the p-values [Genovese, Roeder, and Wasserman, 2006, Blanchard and Roquain, 2008, Habiger, 2017, Roquain and Van De Wiel, 2009, Ramdas, Barber, Wainwright, and Jordan, 2019]. The main goal of this work is to let the weights depend on the data at hand—they are thus denoted as random variables W_i—while providing finite-sample guarantees. Such data-dependent weighting has been recognized as an important open problem [Benjamini, 2008, Roquain and Van De Wiel, 2009] that is essential for dealing with large scale multiple testing. To the best of our knowledge, no solution has been provided so far. Existing proposals for data-driven weighting either explicitly account for overfitting by establishing FDR control at an elevated level compared to nominal [Li and Barber, 2019] or only provide guarantees in the asymptotic limit [Hu, Zhao, and Zhou, 2010, Ignatiadis, Klaus, Zaugg, and Huber, 2016, Durand, 2019, Zhao and Zhang, 2014, Wang, 2018, Roeder, Devlin, and Wasserman, 2007].

2.1 Example: Group Benjamini-Hochberg with cross-weighting

We first provide a rudimentary version of our method that is applicable to situations with categorical (or suitably categorized) covariates $X_i \in \{1, \ldots, G\}$. This setting is called multiple testing with groups; each group consists of hypotheses whose covariate X_i takes on the same value. Our method builds upon the Group Benjamini-Hochberg (GBH) method proposed by Hu et al. [2010] to improve power compared to BH by using the group structure. GBH consists of first estimating the proportion of null hypotheses $\pi_0(g)$ in each group by $\widehat{\pi}_0(g)$, weighting each hypothesis proportionally to $(1-\widehat{\pi}_0(g))/\widehat{\pi}_0(g)$ and finally applying the weighted BH procedure. Algorithm 1 describes the method in detail[1], using the estimator of Storey et al. [2004] applied to the grouped setting, analogous to Sankaran and Holmes [2014].

Hu, Zhao, and Zhou [2010] provide the following guarantees for GBH: in the oracle situation where the $\pi_0(g)$ are known, GBH controls the FDR. In the asymptotic limit where the number of groups is fixed, the number of hypotheses in each group grows to infinity and $\text{plim}_{m\to\infty} \widehat{\pi}_0(g) \geq \pi_0(g)$ for all g, GBH controls the FDR. Furthermore, sufficient conditions are given so that asymptotically GBH is at least as powerful as BH. The asymptotics, however, do not necessarily apply for finite m/G, the number of hypotheses per group, as shown by simulations summarized in Figure 2. Intuitively, the reason is that some groups will randomly be enriched for smaller than expected p-values (and some for larger than expected ones), and the method further up-weights the former set of null p-values.

[1] A simplification is that in Algorithm 1, the weights are specified so that $\sum_i W_i = m$. In contrast, in the original GBH paper [Hu et al., 2010], the weights are less conservative and satisfy $\sum_i \widehat{\pi}_0(X_i) W_i = m$. This inflation ensures that in the oracle case of known $\pi_0(\cdot)$, the FDR of GBH is exactly equal to α. We return to the issue of null proportion adaptivity in Section 2.3 and Theorem 2; in the case of GBH it may be regained by employing the optional step in Algorithm 1.

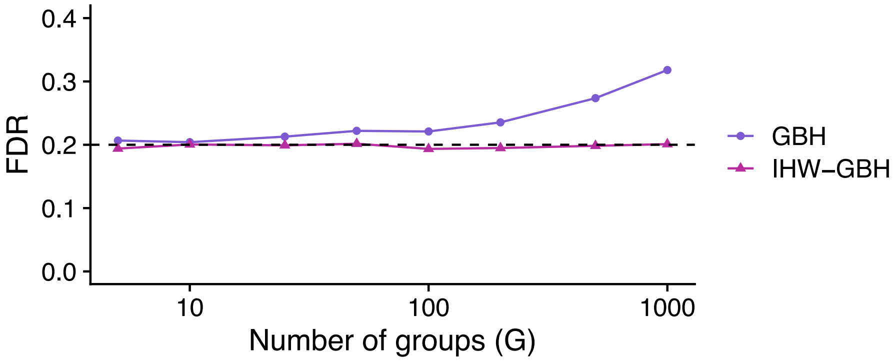


Figure 2: **The need for cross-weighting:** We simulated under the global null with $m = 10000$ independent $P_i \sim U[0,1]$ and $X_i \equiv i \pmod{G}$, where the number of groups G is a simulation parameter shown on the x-axis. Then we applied the GBH and IHW-GBH (with a random partition into 5 folds) methods (described in Algorithms 1 and 2, with $\tau = 0.5$ and without the null-proportion adaptivity step) at level $\alpha = 0.2$. The plot shows the FDR (obtained by averaging over 12000 Monte Carlo replicates) versus G. GBH does not control the FDR, and FDR increases as G increases, while IHW-GBH controls the FDR for all G.

Algorithm 1: The Group Benjamini-Hochberg (GBH) algorithm

```
Input: $(P_1, \dots, P_m) \in [0,1]^m$
       $(X_1, \dots, X_m) \in \{1, \dots, G\}^m$
       a nominal level $\alpha \in (0,1)$
       a censoring level $\tau \in (0,1)$
for $g = 1, \dots, G$ do
    $\widehat{\pi}_0(g) := \frac{1 + \sum_{i: X_i = g} \mathbf{1}(P_i > \tau)}{|\{i : X_i = g\}|(1-\tau)} \wedge 1$
end
for $i = 1, \dots, m$ do
    $W_i := \frac{1 - \widehat{\pi}_0(X_i)}{\widehat{\pi}_0(X_i)} \Big/ \sum_{i=1}^{m} \frac{1 - \widehat{\pi}_0(X_i)}{m \cdot \widehat{\pi}_0(X_i)}$
end
Optional (null prop. adaptivity):
    $\widehat{\pi}'_{0,W} := \frac{\max_{i=1,\dots,m} W_i + \sum_{i=1}^{m} W_i \mathbf{1}(P_i > \tau)}{m(1-\tau)}$
    Update $W_i := W_i / \widehat{\pi}'_{0,W}$
Apply weighted BH (Def. 2) with p-values $P_i$ and weights $W_i$.
```

Algorithm 2: The cross-weighted GBH (IHW-GBH) algorithm

```
Input: $(P_1, \dots, P_m) \in [0,1]^m$
       $(X_1, \dots, X_m) \in \{1, \dots, G\}^m$
       a partition $I_1, \dots, I_K$ of $\{1, \dots, m\}$
       a nominal level $\alpha \in (0,1)$
       a censoring level $\tau \in (0,1)$
for $\ell = 1, \dots, K$ do
    for $g = 1, \dots, G$ do
        $\widehat{\pi}_0^{-\ell}(g) := \frac{1 + \sum_{i \notin I_\ell: X_i = g} \mathbf{1}(P_i > \tau)}{|\{i \notin I_\ell : X_i = g\}|(1-\tau)} \wedge 1$
    end
    for $i \in I_\ell$ do
        $W_i := \frac{1 - \widehat{\pi}_0^{-\ell}(X_i)}{\widehat{\pi}_0^{-\ell}(X_i)} \Big/ \sum_{i \in I_\ell} \frac{1 - \widehat{\pi}_0^{-\ell}(X_i)}{|I_\ell| \cdot \widehat{\pi}_0^{-\ell}(X_i)}$
        Optional (null prop. adaptivity):
            $\widehat{\pi}'_{0,W,\ell} := \frac{\max_{i \in I_\ell} W_i + \sum_{i \in I_\ell} W_i \mathbf{1}(P_i > \tau)}{|I_\ell|(1-\tau)}$
            Update $W_i := W_i / \widehat{\pi}'_{0,W,\ell}$
    end
end
Apply the $\tau$-censored, weighted BH procedure (Def. 2) with p-values $P_i$ and weights $W_i$.
```

```
Algorithm 3: The general IHW algorithm
Input: $\mathbf{P} = (P_1, \ldots, P_m) \in [0,1]^m$
       $\mathbf{X} = (X_1, \ldots, X_m) \in \mathcal{X}^m$
       a partition $I_1, \ldots, I_K$ of $[m]$
       a nominal level $\alpha \in (0,1)$
for $\ell = 1, \ldots, K$ do
    Learn a weight function $\widehat{W}^{-\ell} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ from the pairs $(P_i, X_i)$, $i \notin I_\ell$
    for $i \in I_\ell$ do
        Let $W_i$ a suitable rescaling of $\widehat{W}^{-\ell}(X_i)$, typically
            $W_i := \frac{|I_\ell|\, \widehat{W}^{-\ell}(X_i)}{\sum_{i \in I_\ell} \widehat{W}^{-\ell}(X_i)}$, if $\sum_{i \in I_\ell} \widehat{W}^{-\ell}(X_i) > 0$, else $W_i := 1$
    end
end
Run a weighted multiple testing procedure with p-values $P_i$ and weights $W_i$.
```

Our solution is to use cross-weighting. We assign each hypothesis to one of K folds – randomly and independently of its p-value P_i and covariate X_i– and then calculate weights out-of-fold, as elaborated in Algorithm 2. With cross-weighting, a null p-value that is small by chance cannot lead to an upweighting of itself. FDR control is restored, as shown in Figure 2. On the other hand, if the weights are determined not just by noise, but by true signal, then IHW-GBH, just as GBH, has increased power compared to BH, as we show in a more comprehensive simulation study in Section 5.1. If G furthermore remains fixed as $m \to \infty$, then GBH and IHW-GBH are asymptotically equivalent (Corollary 2).

2.2 IHW: A family of multiple testing procedures

We now generalize the IHW-GBH procedure beyond categorical covariates, the GBH weighting scheme and the weighted BH procedure (Def. 2): we seek a general way of applying weighted multiple testing methods with *data-driven* weights W_i when covariates X_i—not necessarily categorical—are available. Our approach consists of two ingredients: first, we only consider weights that are functions of the covariates X_i, i.e., $W_i = W(X_i)$. The second ingredient is *cross-weighting*: We partition our m hypotheses into K disjoint folds[2] $I_1, \ldots, I_K$. Then, in determining the weight W_i for hypothesis $i \in I_\ell$, we set $W_i \propto \widehat{W}^{-\ell}(X_i)$, where the weight function $\widehat{W}^{-\ell}$ is learned from data *outside* fold I_ℓ and the weights are normalized, typically such that $\sum_{i \in I_\ell} W_i = |I_\ell|$. This overall framework is summarized in Algorithm 3.

In Sections 2.3 and 3.2 we provide formal guarantees of finite-sample type-I error control for the IHW algorithm, under the condition that the weighted multiple testing procedure is weighted BH with τ-censoring or weighted k-Bonferroni. We will discuss how to learn weight functions for general (non-categorical) covariates in Section 4.

2.3 Finite-sample FDR control with cross-weighting under independence

To derive formal guarantees for Algorithm 3, we set out with a sufficient distributional assumption that contains several independence relationships. In Section 3, we will consider more general dependence structures.

[2] Our baseline proposal is to construct the partition by splitting the set $[m] = \{1, \ldots, m\}$ into K (the default in the IHW software package is $K = 5$) equally sized folds randomly. Alternatively, domain specific knowledge can be used to derive folds that minimize across-fold dependence, cf. example in Section 6.

Assumption 1 (Distributional setting under independence)**.** Let (P_i, X_i), $i \in [m]$ be[3] (p-value, covariate) pairs and $\mathscr{H}_0 \subset [m]$ the index set of null hypotheses. We assume that:

(a_1) The null pairs $((P_i, X_i))_{i \in \mathscr{H}_0}$ are jointly independent.

(a_2) The null pairs $((P_i, X_i))_{i \in \mathscr{H}_0}$ are independent of the alternative pairs $((P_i, X_i))_{i \notin \mathscr{H}_0}$.

(b) For $i \in \mathscr{H}_0$, it holds that P_i is independent of X_i.

(c) For $i \in \mathscr{H}_0$, P_i is super-uniform, i.e., $\mathbb{P}[P_i \leq t] \leq t$ for all $t \in [0, 1]$.

To parse this assumption, let us first consider two important special cases: (i) marginalizing over the X_i, so that we only have access to p-values, and (ii) deterministic X_i. In both cases, Assumption 1 reduces to (a_1') $(P_i)_{i \in \mathscr{H}_0}$ are jointly independent, (a_2') independent of the alternative p-values $(P_i)_{i \notin \mathscr{H}_0}$ and (c). Of these, (a_1') and (a_2'), while admittedly strong, are a typical starting point for proving finite-sample results for multiple testing procedures, even in the absence of covariates: Liang and Nettleton [2012] call it the null independence assumption. In the setting with covariates, these are also assumptions made by Li and Barber [2019, Theorem 1] and Lei and Fithian [2018, Theorem 1]. Cai et al. [2019] also assume full independence of hypotheses. The super-uniformity; also called conservativeness, of the null p-values (c) is also a standard assumption in multiple testing [Blanchard and Roquain, 2008]. Li and Barber [2019] make a stronger assumption than (c).

The case of deterministic X_i is important, since for example the genomic distance between SNPs and peaks in our motivating example in Figure 1 is a deterministic covariate. See Supplement S6.1 for additional examples. Nevertheless, we formulate results for the more general case to also handle situations in which the covariate X_i is calculated from the same data that are used to calculate the p-value P_i. For instance, Cai et al. [2019] consider simultaneous two-sample testing, and construct an ancillary X_i that is independent of the t-statistic (and thus also the p-value) under the null hypothesis; we revisit their construction in the simulation study of Section 5.3. Assumption 1(b) is crucial in ensuring that knowledge of X_i does not influence the null distribution. Cai et al. [2019] call it a "principle for information extraction"; cf. Bourgon et al. [2010], Boca and Leek [2018] for further elaborations on this assumption and Supplement S6.2 for more examples of random covariates.

Next, we state two specifications on the weighting mechanism used. Unlike Assumption 1, the applicability of which depends on the generally unknown data-generating mechanism, these are entirely under the control of the analyst.

Specification 1 (Honest weighting)**.** Consider a partition of $[m]$ into K folds $I_1, \ldots, I_K$, i.e., $\bigcup_\ell I_\ell = [m]$ and $(I_\ell)_\ell$ are disjoint, and define $I_\ell^c = [m] \setminus I_\ell$. The partition is assigned independently of $((P_i, X_i))_{i \in [m]}$. Then, the data-driven weights $(W_i)_{i \in [m]}$ are honest with respect to the partition $I_1, \ldots, I_K$ if:

(a) W_i is a function of only $(P_j)_{j \in I_\ell^c}$ and $(X_j)_{j \in [m]}$ for all $\ell \in [K]$ and all $i \in I_\ell$.

(b) The weights in fold I_ℓ average to 1, i.e., $\sum_{i \in I_\ell} W_i = |I_\ell|$ for all $\ell \in [K]$.

(c) $W_i \geq 0$ for all i.

We call this specification "honest weighting", borrowing terminology from the honest tree construction of Wager and Athey [2018], who call a regression tree honest if the set of observations used to determine its structure is disjoint from the set of observations used for prediction in the leaves. Specification 1 encapsulates our idea of cross-weighting. Informally, it says that the weight W_i of hypothesis i should not depend on its p-value P_i. As already shown in Figure 2, without honesty it is easy to overfit the data. Part (b) of the definition encapsulates a fixed weighting budget [Genovese, Roeder, and Wasserman, 2006]. Instead of merely requiring $\sum_{i=1}^m W_i = m$, the budget is

[3] We use the notation $[m] = \{1, \ldots, m\}$.

restricted within each fold, to prevent information leakage across folds through the total magnitude of the weights.

Honesty suffices to guarantee type-I error control in some cases, for example for the weighted k-Bonferroni procedure (Section 3.2 and Theorem 3). However, for the τ-censored, weighted BH procedure with data-driven weights, we require one further condition on the weights, which was proposed by Li and Barber [2019] and states that the magnitude of p-values less than or equal to τ must be concealed from the weighting algorithm.

Specification 2 (τ-censored weighting)**.** The weights W_i are called τ-censored for $\tau \in (0, 1]$ if they depend on the p-values $(P_i)_{i\in[m]}$ only through $(P_i \, \mathbf{1}(P_i > \tau))_{i\in[m]}$.

We are ready to state the first result:

Theorem 1 (IHW-BH controls the FDR under honesty and τ-censored weighting)**.** Let $((P_i, X_i))_{i\in[m]}$ satisfy Assumption 1. Furthermore assume that we construct data-driven weights W_i that are honest (Specification 1) and τ-censored (Specification 2) for some $\tau \in (0, 1]$. Then the τ-censored, weighted BH procedure (Definition 2) with p-values P_i and weights W_i controls the FDR at the nominal level α.

The intuition for this theorem is the following: in the weighted BH algorithm (Definition 2), the rejection threshold of a null p-value P_i depends on its weight W_i and the total number of rejections R. Assumption 1 and honest weighting (Specification 1) ensure that a null p-value cannot influence its own weight. However, tests can coordinate adversarially by weighting each other in way that increases R and potentially leads to their own rejection. Supplement S1.3 provides an example of how such adversarial coordination can break FDR-control guarantees, even though honesty holds. However, under τ-censoring, the only p-values that can coordinate through weight assignment are the ones $> \tau$. These p-values are also excluded from being rejected and so FDR control is restored.

As a corollary, we get the following result:

Corollary 1 (IHW-GBH controls the FDR)**.** Let $((P_i, X_i))_{i\in[m]}$ satisfy Assumption 1, then the IHW-GBH procedure (without the null proportion adaptivity step) described in Algorithm 2 controls the FDR at the nominal level α.

Proof. By construction, the weights W_i of IHW-GBH are honest and τ-censored. □

A shortcoming of IHW-BH with weights that satisfy $\sum_{i=1}^m W_i = m$ is that FDR is controlled at $\pi'_{0,W}\alpha < \alpha$, where $\pi'_{0,W} := (\sum_{i\in\mathscr{H}_0} \mathbb{E}[W_i])/m$ and IHW-BH can thus be needlessly conservative. Motivated by null-proportion adaptive methods for unweighted BH [Storey, Taylor, and Siegmund, 2004] and weighted BH with deterministic weights [Habiger, 2017, Ramdas, Barber, Wainwright, and Jordan, 2019], we estimate $\pi'_{0,W}$ within fold I_ℓ by

$$\hat{\pi}'_{0,W,\ell} = \frac{\left(\max_{i\in I_\ell} W_i\right) + \sum_{i\in I_\ell} W_i \, \mathbf{1}(P_i > \tau')}{|I_\ell|(1-\tau')} \quad \text{with } \tau' \in [\tau, 1],^{4} \tag{5}$$

and use these estimates to inflate the weights W_i. We have the following result:

Theorem 2 (IHW-Storey controls the FDR under honesty and τ-censored weighting)**.** Assume that all assumptions of Theorem 1 are satisfied. Next let $\hat{\pi}'_{0,W,\ell}$ be defined as in (5) and define null-proportion adaptive weights as $W_i^{\text{Storey}} := W_i \,/\, \hat{\pi}'_{0,W,\ell}$ for $i \in I_\ell$. Then the τ-censored, weighted BH procedure (Definition 2) with p-values P_i and weights W_i^{Storey} controls the FDR at the nominal level α.

A direct application of this theorem is that the statement of Corollary 1 also holds for the null-proportion adaptive version of IHW-GBH (cf. Algorithm 2). This provides power gains in situations where the null proportion is substantially smaller than 1 at least in some regions of the covariate space, since then it will be the case that $\sum W_i^{\text{Storey}} > \sum W_i$, thus increasing the total weight budget.

[4] We suggest $\tau' = 0.5$ as a default choice.

2.4 FDR asymptotics with cross-weighting under independence

While the primary focus of this paper is on finite-sample guarantees and performance in simulations, in this section we provide asymptotic results for $m \to \infty$ that serve three purposes: they demonstrate how cross-weighting enables a streamlined proof of asymptotic FDR control under standard assumptions on (P_i, X_i) while dispensing of requirements on the class of weight functions; second, they show that in situations in which there is sufficient signal and the data-driven weight function has approached its asymptotic limit, no power is lost by using cross-weighting; third, they show that in an asymptotic regime, IHW-BH controls the FDR without a need for τ-censoring (Specification 2). On the other hand, our aim here is not to provide the sharpest asymptotics under the weakest conditions, but just to provide these conceptual insights.

We develop the asymptotics using the following Bayesian model [Ferkingstad et al., 2008, Lei and Fithian, 2018, Deb et al., 2018], which we call the conditional two-groups model and which extends the two-groups model of Storey [2003] and Efron, Tibshirani, Storey, and Tusher [2001]:

$$\begin{aligned} &X_i \sim \mathbb{P}^X, \;\; H_i \mid (X_i = x) \sim \text{Bernoulli}(1 - \pi_0(x)), \\ &P_i \mid (H_i = 0, X_i = x) \sim U[0,1], \;\; P_i \mid (H_i = 1, X_i = x) \sim F_{\text{alt}}(\cdot \mid X_i = x) \end{aligned} \tag{6}$$

We also define $F(t \mid X_i = x) = \pi_0(x)t + (1 - \pi_0(x))F_{\text{alt}}(t \mid X_i = x)$: the distribution of P_i given $X_i = x$. The distribution $F(t \mid X_i = x)$ can vary from test to test because of varying null probabilities $\pi_0(x)$ and/or alternative distributions $F_{\text{alt}}(\cdot \mid X_i = x)$, depending on the value of its covariate X_i.

Since m is a changing parameter in the asymptotics, it is useful to formalize what "learning a weight function" entails and use more involved notation:

Specification 3 (Weighting scheme)**.** A weighting scheme $\widehat{W}^{(\cdot)}$ is a mechanism that, for any finite subset $I \subset \mathbb{N}_{>0}$, uses samples $((P_i, X_i))_{i \in I}$ to learn a weight function $\widehat{W}^{(I)} : \mathcal{X} \to \mathbb{R}_{\geq 0}$. We assume that the learned weight function $\widehat{W}^{(I)}$ does not excessively upweight individual hypotheses, i.e., there exists $\Gamma < \infty$ such that

$$\int \widehat{W}^{(I)}(x)^2 d\mathbb{P}^X(x) \leq \Gamma \cdot \left(\int \widehat{W}^{(I)}(x) d\mathbb{P}^X(x) \right)^2 \qquad \text{for all subsets } I \subset \mathbb{N}. \tag{7}$$

Given m independent draws (P_i, X_i) from (6) and a weighting scheme (Specification 3), we seek to apply learned weights in conjunction with weighted BH (Definition 2). We consider two possibilities:

1. **Naive weighted BH:** We use all data $((P_i, X_i))_{i \in [m]}$ to learn $\widehat{W}^{([m])}$ and let $W_i \propto \widehat{W}^{([m])}(X_i)$ for $i = 1, \ldots, m$, such that the weights average to 1 (i.e., $\sum_{i=1}^m W_i = m$). Then we apply the weighted BH procedure with p-values P_i and weights W_i.
2. **IHW-BH:** We partition $[m]$ into K disjoint folds $I_1, \ldots, I_K$, independently of $((P_i, X_i))_{i \in [m]}$. Then we apply Algorithm 3 in conjunction with weighted BH, i.e., for each fold ℓ, we apply the weighting scheme on $[m] \setminus I_\ell$ and for $i \in I_\ell$ set weight $W_i \propto \widehat{W}^{([m] \setminus I_\ell)}(X_i)$ and such that the weights average to 1 in that fold (i.e., $\sum_{i \in I_\ell} W_i = 1$). Then we apply weighted BH with p-values P_i and weights W_i. We note that the data-driven weights W_i are honest (Specification 1) by construction. However, for the asymptotics, we do not require τ-censoring (Specification 2), but instead require the mild technical condition (7).

Proposition 1. Let (P_i, X_i) be i.i.d. from the conditional two-groups model (6) satisfying regularity Assumption 3 (in Supplement S2). If the partition satisfies $|I_\ell|/m \to \gamma_\ell \in (0,1)$ as $m \to \infty$ for all ℓ, then[5]:

(a) There exists a weighting scheme satisfying Specification 3, such that the naive weighted BH procedure asymptotically does not control the FDR.

(b) For any weighting scheme satisfying Specification 3, the IHW-BH procedure asymptotically controls the FDR.

[5] See Supplement S2 for the proof and formal statements.

(c) Consider a weighting scheme that converges in probability to a deterministic limiting weight function $W^* : \mathcal{X} \to \mathbb{R}_{\geq 0}$,

$$\left\| \widehat{W}^{([m])}(\cdot) - W^*(\cdot) \right\|_\infty \xrightarrow{\mathbb{P}} 0 \text{ as } m \to \infty, \quad \int W^*(x) d\mathbb{P}^X(x) = 1, \; \int W^*(x)^2 d\mathbb{P}^X(x) < \infty$$

Then, the naive weighted BH and IHW-BH procedures have the same power asymptotically.

Proof idea for (a) and (b): The proof of Storey et al. [2004] for asymptotic FDR control of BH argues that by the Glivenko-Cantelli theorem, $\sup_t \left| \frac{1}{m} \sum_{i=1}^m [\mathbf{1}(P_i \leq t) - \mathbb{P}[P_i \leq t]] \right| \xrightarrow{\mathbb{P}} 0$ and similarly for the subset of null hypotheses. A consequence is that the BH estimator of the false discovery rate is asymptotically uniformly conservative over all thresholds $\geq \delta > 0$, which in turn implies asymptotic FDR control. Extending this argument to the weighted case requires uniform convergence: $\sup_t \left| \frac{1}{m} \sum_{i=1}^m [\mathbf{1}(P_i \leq tW_i) - \mathbb{P}[P_i \leq tW_i]] \right| \xrightarrow{\mathbb{P}} 0$.

For data-driven weights, this can be achieved by learning the weight function from a suitably restricted class $\mathcal{W}$. Du and Zhang [2014], Ignatiadis et al. [2016], Durand [2019] all use $\mathcal{W}$ such that the functions $\{(p, x) \mapsto \mathbf{1}(p \leq tW(x)) \mid t \in (0, 1], W(\cdot) \in \mathcal{W}\}$ are $\mathbb{P}$-Glivenko-Cantelli [van der Vaart, 2000]. Similarly, Li and Barber [2019] consider $\mathcal{W}$ with low Rademacher complexity. On the other hand, if convergence is not uniform (e.g., if we are free to choose any weights satisfying Specification 3), then we can find regions of $\mathcal{X}$-space that are enriched for small p-values merely by chance, upweight them, and violate FDR control (cf. Figure 2).

Instead, through cross-weighting, the richness of $\mathcal{W}$ is irrelevant: upon conditioning on other folds, $P_i / \widehat{W}^{([m]\setminus I_\ell)}(X_i)$ in fold I_ℓ are i.i.d., and thus the one-dimensional Glivenko-Cantelli result applies. □

In words, while data-driven weights can lead to overfitting (a), cross-weighting universally alleviates this (b). A further upshot of (b) is that it dispenses with the requirement for τ-censored weights (Specification 2). Finally, the objection may be raised to cross-weighting that it drops data and should thus be less powerful than a procedure that uses all the data. However, (c) shows that asymptotically one loses no power by using cross-weighting if the weighting procedure is well-behaved, i.e., the weights asymptotically converge to a limit.

As a corollary of Proposition 1, we have that:

Corollary 2 (IHW-GBH asymptotics)**.** Under the assumptions of Proposition 1 with $\mathcal{X} = [G]$ for fixed $G \in \mathbb{N}$, the GBH and IHW-GBH procedures without null proportion adaptivity, described in Algorithms 1 and 2, have the same power asymptotically.

Proof. In Supplement S2.4, we verify (7) and the condition from part (c) of Proposition 1. □

At this point, we note that Durand [2019], motivated by a preprint version of this work, derived the following related and elegant result: in the setting with $\mathcal{X}$ a finite discrete space, Durand [2019, Theorem 7.1.] constructs a cross-weighted procedure that asymptotically controls the FDR and simultaneously achieves the power of the *optimal* weighted procedure.

3 Extension to dependence

3.1 The key assumption: Independence across folds, dependence within

Assumption 1 made the strong assumption of joint independence of all null p-values and was sufficient for the results presented in Section 2. Real data commonly deviate from this assumption. The consequences of such deviations on the applicability of results derived using independence assumptions are typically difficult to reason about. It is therefore desirable to construct guarantees that can be derived from weaker assumptions that are closer to realistic patterns of dependence.

Assumption 2 (Distributional setting with dependence)**.** Let (P_i, X_i), $i \in [m]$ be (p-value, covariate) pairs, $I_1, \ldots, I_K$ be folds of a partition of $[m]$ that is defined based on information independent of $((P_i, X_i))_{i \in [m]}$, and let $\mathscr{H}_0 \subset [m]$ the index set of null hypotheses. We assume that:

(a) The (p-value, covariate) pairs are independent across folds $I_1, \ldots, I_K$, but may be dependent within each fold. Formally, $(P_i, X_i)_{i \in I_\ell}, \ell \in [K]$ are jointly independent.

(b) For $i \in \mathscr{H}_0$, it holds that P_i is independent of $(X_i)_{i \in [m]}$.

(c) For $i \in \mathscr{H}_0$, P_i is super-uniform, i.e., $\mathbb{P}[P_i \leq t] \leq t$ for all t.

Let us compare Assumption 2 to Assumption 1. Parts 2(b, c) are mild. Part 2(c) is identical to 1(c) and standard in multiple testing. Part 2(b) is analogous to 1(b), albeit stronger, since we are conditioning on the full vector of X_i. Nevertheless, 2(b) is implied by 1(a,b). In the important case where the X_i are deterministic, 1(b) trivially holds. But it also allows for situations where, for instance, the X_i are random spatial locations. In this case, we may expect p-values with similar X_i to be correlated. Assumption 2(b) then means that knowing the locations X_i of all hypotheses provides no information about a *single* null p-value P_i.

The critical assumption is 2(a). Without covariates, the assumption implies that $I_1, \ldots, I_K$ is a partition of p-values into independent blocks. This is not an assumption typically encountered in the multiple testing literature, although it has appeared e.g., in Heesen and Janssen [2015], Guo and Sarkar [2019]. It is fundamental to the cross-weighting approach, the core idea of which is to avoid any dependence between each individual null p-value P_i and its data-driven weight W_i. Cross-weighting ensures that W_i is determined based on X_i and p-values from the other folds, but not P_i. This would not longer be true with dependence *across* folds. This observation is analogous to a similar phenomenon in cross-validation: in Chapter 7.1 of the *Elements of Statistical Learning*, Hastie, Tibshirani, and Friedman [2009] caution practitioners to split data into independent folds when evaluating a supervised learning method by cross-validation (CV): if the folds are not independent, the CV estimates of prediction error are not reliable.

From the application perspective, the assumption is practical: domain experts often have sufficient understanding of their data to find suitable partitions of the hypotheses into independent blocks. In the example from Figure 1, further detailed in Section 6, it is plausible to assume that the data for hypotheses located on different chromosomes are independent, or at least that any potential dependences are negligible. As another example, for covariates X_i that correspond to spatial or temporal positions, hypotheses that are sufficiently far away from each other will be independent if the dependences are mediated by spatial or temporal proximity.

We note that all other existing methods for multiple testing with covariates that provide FDR control assume either full independence [Lei and Fithian, 2018, Cai et al., 2019], weak dependence [Li and Barber, 2019] or the ability to consistently estimate the joint distribution of all hypotheses [Sun and Cai, 2009]. Thus, Assumption 2 is a practical starting point towards dealing with common patterns of dependence encountered in real data.

Next, we describe two multiple testing methods with data-driven weights that have provable type-I error guarantees under dependence.

3.2 k-FWER control with cross-weighting under dependence

k-FWER control is achieved by applying cross-weighting in conjunction with the weighted k-Bonferroni procedure of Definition 1. We are not aware of existing procedures with data-driven weights and finite-sample k-FWER control. Existing proposals provide asymptotic guarantees [Wang, 2018].

The proof is direct and without technical complications. We provide it here in the main text, since it shows the key idea behind cross-weighting: each null p-value P_i is independent of its weight W_i, and this protects against overfitting.

Theorem 3. Let $((P_i, X_i))_{i\in[m]}$ satisfy Assumption 2 (or Assumption 1) with respect to the partition $I_1, \ldots, I_K$. Furthermore assume that we construct data-driven weights W_i that are honest w.r.t. $I_1, \ldots, I_K$ (Specification 1). Then the weighted k-Bonferroni procedure (Definition 1) with p-values P_i and weights W_i controls the k-FWER at the nominal level α.

Proof. We first show that P_i is independent of W_i ($P_i \perp W_i$) for any $i \in \mathscr{H}_0$. Without loss of generality, $i \in \mathscr{H}_0 \cap I_\ell$. By honesty, W_i is a function only of the p-values in the other folds, $(P_i)_{i\in I_\ell^c}$ and all covariates $\mathbf{X} = (X_i)_{i\in[m]}$. It thus suffices to argue that P_i is independent of $((P_i)_{i\in I_\ell^c}, \mathbf{X})$. But this follows from Assumption 2 (resp. Assumption 1). We next bound the k-FWER.

$$\begin{aligned} k\text{-FWER} = \mathbb{P}[V \geq k] \leq \frac{1}{k}\mathbb{E}V = \frac{1}{k}\sum_{i\in\mathscr{H}_0} \mathbb{P}\left[P_i \leq \frac{k\alpha W_i}{m}\right] \\ = \frac{1}{k}\sum_{i\in\mathscr{H}_0} \mathbb{E}\left[\mathbb{P}\left[P_i \leq \frac{k\alpha W_i}{m} \mid W_i\right]\right] \overset{(*)}{\leq} \frac{1}{k}\sum_{i\in\mathscr{H}_0} \mathbb{E}\left[\frac{k\alpha W_i}{m}\right] = \frac{\alpha}{m}\mathbb{E}\left[\sum_{i\in\mathscr{H}_0} W_i\right] \leq \alpha\,. \end{aligned}$$

Note that in $(*)$, we used the fact that for $i \in \mathscr{H}_0$ it holds that P_i is super-uniform and P_i is independent of W_i. In the last step we used that honesty ensures that $\sum_i W_i = m$. □

3.3 FDR control with cross-weighting under dependence

We recall the basic procedure for controlling FDR with (deterministic) weights under arbitrary dependence:

Definition 3 (Weighted Benjamini-Yekutieli (wBY) [Benjamini and Yekutieli, 2001, Blanchard and Roquain, 2008])**.** Consider p-values $P_1, \ldots, P_m$ with arbitrary dependence such that the null p-values are super-uniform. Furthermore consider deterministic weights $w_i \geq 0$ such that $\sum_{i=1}^m w_i = m$. Then the FDR is controlled at level $\alpha \in (0, 1)$ by applying the weighted Benjamini-Yekutieli procedure at level α, i.e., the weighted Benjamini-Hochberg procedure (Definition 2) with $\tau = 1$ at level $\alpha / \sum_{k=1}^m \frac{1}{k}$.

We now show that applying the weighted BY procedure with cross-weighting controls the FDR under Assumption 2.

Theorem 4 (IHW-BY controls the FDR under honesty and independent folds)**.** Let $((P_i, X_i))_{i\in[m]}$ satisfy Assumption 2 with respect to the partition $I_1, \ldots, I_K$. Furthermore assume that we construct data-driven weights W_i that are honest w.r.t. $I_1, \ldots, I_K$ (Specification 1). Then the weighted BY procedure (Definition 3) with p-values P_i and weights W_i controls the FDR at the nominal level α.

To demonstrate that honesty is essential for the result of Theorem 4, we next describe two plausible candidate methods for FDR control with covariates that do not control FDR:

Example 1 (BY with arbitrary data-driven weights does not control FDR under Assumption 2)**.** Theorem 4 may appear as a consequence of Theorem 4.2. of Blanchard and Roquain [2008], who extended the results of Benjamini and Yekutieli [2001] and proved that the weighted BY procedure (Definition 3) controls the FDR for any choice of weights and any p-value distribution. However, their result holds only for deterministic weights and not for data-driven weights, as we now demonstrate.

Proof. We generate $((P_i, X_i))_{i\in[m]}$ satisfying Assumption 2 and under the global null as follows: Fix $m = 2m'$ for $m' \in \mathbb{N}$. We consider deterministic covariates $X_i = i$ and partitions $I_1 = \{1, \ldots, m'\}, I_2 = \{m' + 1, \ldots, m\}$. We first draw a permutation σ from the uniform measure on the permutation group of $\{1, \ldots, m'\}$. Next we independently draw: $U_i \sim U[(i-1)/m', i/m']$ for $i = 1, \ldots, m'$ and let $P_i = U_{\sigma(i)}$. Finally we draw independent $P_{m'+1}, \ldots, P_m \sim U[0, 1]$. Weights are chosen as follows: Let $i^* \in \text{argmin}_i \{P_i\}$ and then let $W_i = W(X_i) = m\mathbf{1}(X_i = i^*)$. Then the FDR of weighted BY at α is equal to 1 as soon as $m / \sum_{k=1}^m \frac{1}{k} > 2/\alpha$, as we now show:

Since the smallest p-value in I_1 is uniformly distributed on $U[0, 1/m']$, it follows that with probability 1, $P_{i^*} \leq 1/m'$ and hence $P_{i^*}/W_{i^*} \leq 2/m^2 < \frac{\alpha}{m\sum_{k=1}^{m}\frac{1}{k}}$. H_{i^*} gets rejected and so $\text{FDP} = 1$ almost surely. $\square$

In contrast, FDR control would be guaranteed, had we used weights derived through cross-weighting. BY with τ-censored weights (Specification 2) also does not control FDR, cf. Supplement S1.6.

Example 2 (AdaPT with BY correction does not control FDR under Assumption 2)**.** Lei and Fithian [2018] prove FDR control for AdaPT under full independence (cf. Assumption 1). Here we demonstrate that even with the Benjamini-Yekutieli correction, i.e., at level $\alpha/\sum_{k=1}^{m}\frac{1}{k}$ and τ-censoring (Specification 2), AdaPT does not control FDR under Assumption 2.

Proof. We generate $((P_i, X_i))_{i\in[m]}$ satisfying Assumption 2 and under the global null as follows: Again we fix $m = 2m', m' \in \mathbb{N}$ and partitions $I_1 = \{1, \dots, m'\}, I_2 = \{m'+1, \dots, m\}$. We take constant covariates $X_i = 1$ for all i and draw $P_1, P_{m'+1} \overset{\text{iid}}{\sim} U[0, 1]$. Finally we set $P_2, \dots, P_{m'} = P_1$ and $P_{m'+2}, \dots, P_m = P_{m'+1}$. We then run the AdaPT algorithm at level $\alpha/\sum_{k=1}^{m}\frac{1}{k}$ with the initialization specified in Lei and Fithian [2018]. Then $\text{FDR} \geq 0.2925$ as soon as $m/\sum_{k=1}^{m}\frac{1}{k} > 2/\alpha$, as we now show:

As specified in Section 4.4.1 of Lei and Fithian [2018], the AdaPT algorithm is initialized at threshold 0.45. Now call A the event that $\{P_1 \leq 0.45, P_{m'+1} < 0.55\}$. On the event A, on the first step of the algorithm, AdaPT estimates the FDP (cf. (13)) as $\left(1 + \sum_i \mathbf{1}\left(P_i > 1 - 0.45\right)\right) / \sum_i \mathbf{1}\left(P_i \leq 0.45\right)$, which is equal to $1/m'$ if $P_{m'+1} > 0.45$ and equal to $1/m$ otherwise. In both cases, the estimated FDP is less or equal than $1/m'$ and thus less than $\alpha/\sum_{k=1}^{m}\frac{1}{k}$ under our assumption on m, α. Thus AdaPT immediately terminates, rejecting all p-values in I_1, and so $\text{FDP} = 1$. Similarly $\text{FDP} = 1$ on the event $A' = \{P_1 < 0.55, P_{m'+1} \leq 0.45\}$ and $\text{FDR} \geq \mathbb{P}[A \cup A'] = 0.2925$. Finally, note that the above procedure is τ-censored with $\tau = 0.45$. $\square$

4 Learning powerful weighting rules

Sections 2 and 3 focused on sufficient conditions for type-I error control, but did not address power. These conditions leave considerable flexibility in the choice of the class of possible weight functions, and in the method of selecting (or "learning") these functions, given the data. This flexibility gives the analyst the opportunity to use domain-specific as well as statistical knowledge to make choices that have desirable type-II error properties. Nevertheless, it is useful to provide a default algorithm that works well across a range of settings. To this end, here we describe two schemes for learning weight functions, one for weighted k-Bonferroni and one for weighted BH. Both rely on positing the approximate applicability of model (6), estimating quantities appearing therein and solving a convex program to find a weight function that optimizes the expected number of discoveries.

4.1 Learning weights for IHW k-Bonferroni

The weighted k-Bonferroni procedure with weight function $W(\cdot)$ rejects hypotheses that satisfy $P_i \leq k\alpha/mW(X_i)$. Under Model (6), a weight function maximizing the expected number of discoveries is one that maximizes $\sum_i \mathbb{P}[P_i \leq k\alpha/mW(X_i) \mid X_i] = \sum_i F(k\alpha/mW(X_i) \mid X_i)$. To derive honest weights (Specification 1) that allow this objective, we learn $\widehat{W}^{-\ell}$ for each fold ℓ separately as follows: First we estimate $F(t \mid x)$ from Model (6) by $\widehat{F}^{-\ell}(t \mid x)$ using only p-values and covariates outside of fold ℓ. Next, identifying $\widehat{W}^{-\ell}(\cdot)$ with the function's values evaluated at the X_i, i.e. $W_i = \widehat{W}^{-\ell}(X_i)$, $i \in I_\ell$ we solve the $|I_\ell|$-dimensional problem with optimization variables $\mathbf{w} = (w_i)_{i\in I_\ell}$:

$$(W_i)_{i\in I_\ell} \in \underset{\mathbf{w}\in[0,\infty)^{|I_\ell|}}{\operatorname{argmax}} \left\{ \sum_{i\in I_\ell} \widehat{F}^{-\ell}\left(k\alpha/m \cdot w_i \;\middle|\; X_i\right) \;\middle|\; w_i \geq 0,\; \sum_{i\in I_\ell} w_i = |I_\ell| \right\}. \tag{8}$$

This setting allows for conditional distributions $\widehat{F}^{-\ell}(t \mid X_i)$ that are different for tests with different covariates X_i. We consider estimators of $\widehat{F}^{-\ell}(t \mid x)$ that are concave in t for all x. This has the advantage of turning (8) into a convex optimization program, which is often tractable. Concavity of the distribution of p-values is a reasonable assumption and often provides a good fit to multiple testing datasets [Strimmer, 2008b, Genovese, Roeder, and Wasserman, 2006]. However, the procedure works even when the concavity assumption does not hold: Given any (potentially non-concave) pilot estimator of the conditional distribution function $t \mapsto F(t \mid x)$, we can project it onto the set of concave distribution functions and solve the optimization problem with the projected distribution functions. We interpret the resulting procedure as a convex relaxation of (8) that makes computation tractable.

With this setup, we are ready to state a concrete weighting scheme, which proceeds in three steps: first, discretize the $X_i$into a finite number of bins defined, e.g., by quantile slicing or as the leaves of a tree. Second, estimate $\widehat{F}^{-\ell}(t \mid \text{bin})$ by the Grenander estimator [Grenander, 1956], i.e., the least concave majorant of the empirical cumulative distribution function of the p-values P_i with $i \in I_\ell^c$ and $X_i \in \text{bin}$. Third, solve (8) for each ℓ by linear programming. The reason that (8) may be expressed as a linear program is that the Grenander estimator is always concave in t and piecewise linear. We provide the details of the estimation and optimization procedures in Supplement S4.1; the computational complexity scales as $O(\log(m) \cdot m)$.

An alternative ansatz is to specify $\pi_0(x)$ and $F_{\text{alt}}(\cdot \mid X_i = x)$ in the conditional two-groups model (6) parametrically. For instance, we may consider for $X_i \in \mathbb{R}^p$

$$\begin{aligned} &\pi_0(x) = \text{expit}(a_0 + a^\top x), \qquad \text{where } \text{expit}(u) = \exp(u)/(1+\exp(u)) \\ &F_{\text{alt}}(\cdot \mid X_i = x) = \text{Beta}(\beta(x), 1), \quad \beta(x) = b_0 + b^\top x. \end{aligned} \tag{9}$$

Such a Beta-Uniform mixture model has been considered in the setting without covariates, e.g., by Allison et al. [2002], Klaus and Strimmer [2011] and with covariates by Lei and Fithian [2018]. In Supplement S4.2 we explain how to learn the parameters of the model using the expectation-maximization algorithm and how to optimize (8).

4.2 Learning weights for IHW Benjamini-Hochberg

Our starting point for deriving powerful weight functions for the weighted BH procedure (Definition 2) is again the conditional two-groups model (6). We seek a threshold function $s : \mathcal{X} \to [0,1]$, such that the multiple testing procedure that rejects hypotheses with $P_i \leq s(X_i)$ satisfies the following two properties: first, the marginal FDR, defined as $\text{mFDR}(s) := \mathbb{P}[H_i = 0 \mid P_i \leq s(X_i)]$ is bounded by α, i.e., $\text{mFDR}(s) \leq \alpha$ and second, the expected number of discoveries $\sum F(s(X_i) \mid X_i)$ is maximized[6]. Similarly to our Bonferroni construction, we learn the threshold function $\widehat{s}^{-\ell}$ for each fold ℓ separately. To this end, we estimate $\widehat{F}^{-\ell}(t \mid x)$ and $\widehat{\pi}_0^{-\ell}(x)$ out of fold. Noting that $\text{mFDR}(s) \leq \alpha$ is implied by $\sum_i \pi_0(X_i) s(X_i) \leq \alpha \sum_i F(s(X_i) \mid X_i)$, we propose solving:

$$\mathbf{t} = (t_i)_{i \in I_\ell} \in \underset{\mathbf{t} \in [0,1]^{|I_\ell|}}{\text{argmax}} \left\{ \sum_{i \in I_\ell} \widehat{F}^{-\ell}(t_i \mid X_i) \;\middle|\; t_i \geq 0,\; \sum_{i \in I_\ell} \widehat{\pi}_0^{-\ell}(X_i) t_i \leq \alpha \sum_{i \in I_\ell} \widehat{F}^{-\ell}(t_i \mid X_i) \right\} \tag{10}$$

As our goal is to apply the weighted BH procedure, we convert these thresholds t_i into weights W_i through normalization: for $i \in I_\ell$, set $W_i = |I_\ell| \cdot t_i / (\sum_{i \in I_\ell} t_i)$, unless the denominator is 0, in which case $W_i = 1$. A few remarks are in order: Similarly to optimization problem (8), (10) is also a convex program if $\widehat{F}^{-\ell}(t \mid x)$ is concave in t for all x, and may be expressed as a linear program if the Grenander estimator is used. We thus again suggest to discretize X_i and estimate distributions with the Grenander estimator. If the weights will be applied in conjunction with the weighted

[6] Such a Bayesian, Neyman-Pearson type procedure is motivated by the asymptotic equivalence between the frequentist FDR and the mFDR [Genovese and Wasserman, 2004, Sun and Cai, 2007, Cai and Sun, 2009, Cai et al., 2019].

BH algorithm, we suggest to simply set $\widehat{\pi_0}^{-\ell} \equiv 1$. This optimization and estimation scheme was proposed by Ignatiadis et al. [2016]. Alternatively, $\widehat{\pi_0}^{-\ell}(x)$ may be estimated by applying Storey's null proportion estimator [Storey et al., 2004] to all hypotheses outside fold I_ℓ that fall into the same bin as x. Details of the estimation and optimization procedures are provided in Supplement S4.1.

The weights W_i constructed above are honest (Specification 1). Yet, in view of Theorems 1 and 2, it might appear unsatisfying that W_i do not satisfy the τ-censored weights condition (Specification 2). In our experience, the proposed procedure with the Grenander estimator does not overfit and controls the FDR. This is corroborated by extensive simulations below and by the asymptotic guarantees of Proposition 1.

Our alternative proposal, which satisfies τ-censoring (Specification 2), is to fit the Beta-Uniform mixture model (9). The EM algorithm may be modified to accommodate for censored knowledge of $P_i \leq \tau$; cf. Markitsis and Lai [2010] in the setting without covariates. Furthermore, with model (9), the solution to problem (10) lies on a contour of equal conditional local fdr (cf. Theorem 2 in Lei and Fithian [2018]), and this fact facilitates the optimization. We describe the steps in more detail in Supplement S4.2.

Finally, we use the same framework to derive weights for the weighted Benjamini-Yekutieli procedure (Definition 3): we proceed as for weighted BH but solve (10) with α replaced by $\alpha / \sum_{k=1}^m \frac{1}{k}$. In this case, honesty suffices for FDR control (Theorem 4).

5 Numerical experiments

Our goal in this section is to corroborate through simulations of three important settings—grouped multiple testing, multiple testing with continuous covariates and simultaneous two-sample tests—the following claims: First, some methods with asymptotic FDR control guarantees do not control FDR in finite samples. Second, IHW is a flexible framework for multiple testing, its main advantage over other methods being finite sample error control (due to cross-weighting), while remaining competitive in terms of power. Throughout this section, we define power as

$$\text{Power} = \mathbb{E}\left[\frac{\sum_{i \notin \mathscr{H}_0} \mathbf{1}\left(i \text{ rejected}\right)}{\max\left\{1, m - |\mathscr{H}_0|\right\}}\right] \tag{11}$$

The expectation, just as the FDR, is evaluated through averaging over Monte Carlo replicates.

5.1 Grouped multiple testing

We first consider the multiple testing problem with groups, i.e, with categorical covariates $X_i \in [G]$. In each simulation we generate P_i, H_i, X_i as follows ($m = 20000$)

$$\tilde{X}_i = \lfloor 40 \cdot (i-1)/m \rfloor, \;\; X_i = \lceil \tilde{X}_i / 40 \cdot G \rceil$$

$$H_i \mid \tilde{X}_i \sim \text{Bernoulli}(1 - \pi_0(\tilde{X}_i)), \;\; \pi_0(\tilde{X}_i) = (0.2 + 0.8\tilde{X}_i/36) \cdot \mathbf{1}(\tilde{X}_i = 0 \text{ mod } 4) \; + \; \mathbf{1}(\tilde{X}_i \neq 0 \text{ mod } 4)$$

$$Z_i \mid H_i, \tilde{X}_i \sim \mathcal{N}(H_i \cdot \mu(\tilde{X}_i), 1), \;\; \mu(\tilde{X}_i) = 2.5 - 2\tilde{X}_i/36$$

$$P_i = 1 - \Phi(Z_i), \;\; \Phi \text{ is the standard Normal CDF}$$

In words, there are 40 latent groups defined by $\tilde{X}_i$, each with 500 hypotheses. A quarter of the groups has non-nulls, three quarters do not. The alternative signal strength $\mu(\cdot)$ and null proportion $\pi(\cdot)$ vary linearly across non-null groups. Parameters are chosen so that the overall proportion of nulls is 0.9. We then coarsen $\tilde{X}_i$ to $X_i = \lceil \tilde{X}_i/40 \cdot G \rceil$, with G varying across simulations; X_i is non-latent, i.e., visible to the algorithm. For example, for $G = 2$, X_i takes on only two levels (2 groups), while for $G = 40$, $X_i = \tilde{X}_i$ takes on all 40 levels. We also use the above configuration of covariates and simulate under the global null by drawing all p-values from the uniform distribution.

We compare the following seven methods:

1. The **Benjamini-Hochberg (BH)** method [Benjamini and Hochberg, 1995], which ignores the covariates X_i.

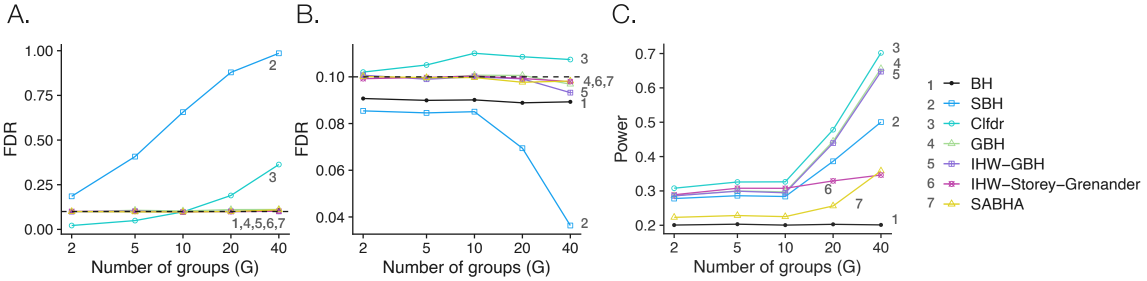


Figure 3: **Grouped multiple testing simulation: A. False discovery rate under the global null** in Model (5.1) (averaged over 10000 Monte Carlo replicates) for seven methods for multiple testing with groups. **B.,C. False discovery rate and power** in Model (5.1) (averaged over 200 Monte Carlo replicates) when there is signal (average null proportion is 0.9) for the same seven methods. The nominal α is equal to 0.1 throughout.

2. The **stratified BH procedure (SBH)** [Sun et al., 2006, Efron, 2008], wherein the BH procedure is applied G times separately to p-values corresponding to different levels of X_i.
3. The **Clfdr (conditional local fdr)** procedure of Cai and Sun [2009], which applies an optimal decision rule that rejects hypotheses with a low value of the group-wise local fdr (cf. Algorithm 4 in Supplement S3). We apply a data-driven version of the oracle rule by estimating local fdrs within each group with the `fdrtool` CRAN Package [Strimmer, 2008b], which estimates marginal densities with the Grenander estimator.
4. The **Group Benjamini-Hochberg (GBH)** procedure of Hu et al. [2010] with null-proportion adaptivity, as described in Algorithm 1 ($\tau = 0.5$).
5. The **IHW-GBH** procedure with null-proportion adaptivity, as described in Algorithm 2 ($\tau = 0.5$) with hypotheses randomly split into 5 folds.
6. The **IHW-Storey-Grenander** procedure: The IHW-Storey method (Theorem 2) with hypotheses randomly split into 5 folds and data-driven weights based on the Grenander estimator described in Section 4.2 and Supplement S4.1.
7. The **Structure Adaptive Benjamini-Hochberg (SABHA)** algorithm by Li and Barber [2019]: SABHA first estimates $\widehat{\pi}_0(\cdot)$ for each group by solving a joint convex optimization problem. Then, the τ-censored, weighted BH procedure is applied with weights $W_i = 1/\widehat{\pi}_0(X_i)$. We set the tuning parameters of group-wise SABHA to $\tau = 0.5$, $\varepsilon = 0.1$ following Section 7.1 of Li and Barber [2019].

All of the above methods provably control FDR asymptotically, as $m \to \infty$, the number of groups remains fixed and there is signal in the data, but only BH and IHW-GBH have provable finite-sample FDR control at α and SABHA at $\alpha(1 + 10\sqrt{G/m})$ [Li and Barber, 2019, Lemma 2].

Results are shown in Figure 3. Under the global null (Fig. 3A), SBH strongly overfits, since under the global null the FDR is equivalent to the FWER, so it would need to pay a Bonferroni correction to apply BH separately to each group. Clfdr has FDR much below nominal for a small number of groups (the oracle local fdr procedure would not reject anything under the global null), but as the number of groups increases, it no longer controls FDR. We further discuss this below. All other methods control FDR in this setting. For GBH, however, recall Fig. 2 for a situation where it does display a pronounced loss of FDR control.

For the simulations with signal (Fig. 3B, C) we make the following observations: As G increases, the covariates become more informative, hence in principle power can be increased. Indeed this is precisely what we observe (Fig. 3C) for the grouped methods methods that do not directly estimate the distribution in each group (all methods except Clfdr and IHW-Storey-Grenander). The power of BH remains constant. After BH, the least powerful procedure appears to be SABHA; the suboptimaliy of its weighting scheme has been previously pointed out [Lei and Fithian, 2018]. We also observe here that IHW-GBH matches the power of GBH and has the added advantage of provable finite-sample FDR control. Regarding the methods that estimate the distribution, when G is small relative to m, then the Grenander estimator can precisely estimate the distribution in each bin. This translates into the Clfdr procedure and IHW-Storey-Grenander having the largest power at small G; indeed Clfdr is provably asymptotically the most powerful procedure in this setting. However, as G increases and the amount of data in each group decreases, the distributions are not estimated as accurately. The consequence for Clfdr is loss of FDR control (Fig. 3B), while for IHW-Storey-Grenander this is only reflected in terms of power, which is lower than other methods, but still higher than the power of BH. This is a feature of its cross-weighting approach: poor estimation is translated to potentially small power, but not to loss of FDR control. In conclusion, in this set of simulations, IHW is the most powerful method of those that control FDR.

5.2 Multiple testing with continuous covariates

In this section we explore a setting with a two-dimensional, continuous covariate X_i. We seek to compare IHW, AdaPT and local fdr based methods with an emphasis on understanding behavior under model-misspecification (to be made precise momentarily). We simulate independent $(X_i, H_i, P_i), i = 1, \ldots, 10000$ from the conditional two-groups model (6) with the following choices for $\mathbb{P}^X, \pi_0(x)$ and $F_{\text{alt}}(\cdot \mid X_i = x)$:

$$\begin{aligned} &\mathbb{P}^X = U[0,1]^2, \;\; \pi_0(x) = 0.98 \cdot \mathbf{1}\left(x_1^2 + x_2^2 \leq 1\right) + 0.6 \cdot \mathbf{1}\left(x_1^2 + x_2^2 > 1\right), \;\; (\mathbb{E}[\pi_0(X_i)] \approx 0.9) \\ &F_{\text{alt}}(\cdot \mid X_i = x) = \text{Beta}(\beta(x), 1), \;\; \beta(x) = 1\big/\max\left\{1.3, \bar{\beta} \cdot (\sqrt{x_1} + \sqrt{x_2})\right\} \end{aligned} \tag{12}$$

$\bar{\beta} \in [1, 3]$ is a parameter that varies across simulation settings. The two-dimensional covariates X_i modulate both the null proportion $\pi_0(X_i)$ and the signal in the alternative density. We compare six methods.

1. The **Benjamini-Hochberg (BH)** [Benjamini and Hochberg, 1995] method ignoring X_i.
2. The **oracle Clfdr procedure (Clfdr-oracle)** that rejects hypotheses with a small conditional local fdr, $\text{fdr}(P_i|X_i) := \mathbb{P}[H_i = 0|X_i, P_i]$ with a threshold chosen through Algorithm 4 in Supplement S3. This procedure achieves an optimal trade-off between the false nondiscovery rate and the false discovery rate, cf. Sun and Cai [2007], Cai and Sun [2009]. Clfdr-oracle, however, would not be available to an analyst, as it assumes oracle knowledge of the components (12) in model (6).
3. The **IHW-BH-Grenander** procedure, similarly to the previous section, but without null-proportion adaptivity (i.e., with IHW-BH instead of IHW-Storey). The covariates $X_i \in [0, 1]^2$ are binned into 5×5 equal volume bins.

Furthermore, we compare three methods that fit Model (9) as a misspecified working model for the true model (12) using the EM algorithm (details in Supplement S4.2).

4. **Clfdr-EM:** this is the same as Clfdr-oracle, but instead of true quantities we use the ones estimated by maximum likelihood on the misspecified model (9). We employ the EM algorithm since the status $H_i \in \{0, 1\}$ is unknown.
5. **IHW-Storey-BetaMix:** this is the IHW-Storey method with hypotheses split randomly into 5 folds and weights derived from optimization problem (10) based on the (out-of-fold) estimated working model (9). Here the EM algorithm deals with both unknown H_i and unknown value of censored p-values $P_i \leq \tau$ with $\tau = 0.1$.
6. **AdaPT**, as implemented in the `adaptMT` CRAN package, wherein in each iteration the working model (9) is fitted. The EM algorithm deals with unknown H_i and for a subset of hypotheses ("masked hypotheses") the algorithm only has access to $\min\{P_i, 1 - P_i\}$ instead of P_i.

The results are shown in Fig. 4. As expected from theory, Clfdr-oracle controls the FDR and is most powerful. Clfdr-EM is also powerful, however because of misspecification in model (9), it does not control the FDR. All other algorithms control the FDR. Among these, AdaPT is most powerful, closely followed by IHW-Storey-BetaMix and then by IHW-BH-Grenander; all of these procedures improve substantially upon BH.

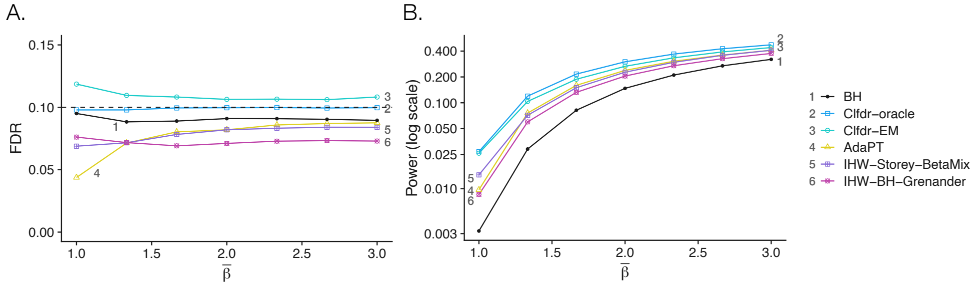


Figure 4: **Simulation for multiple testing with a continuous covariate: A. False discovery rate** in model (14) for six methods. The x-axis corresponds to a simulation parameter that is monotonically related to the strength of the signal for the alternatives. **B. Power** in model (14) for the same six methods. The nominal α is equal to 0.1 throughout and results are averaged over 400 Monte Carlo replicates.

Breaking AdaPT: Fig. 4 demonstrates that AdaPT is very powerful for multiple testing in model (12). Here we mention some caveats: AdaPT is a lot slower than e.g., IHW-Storey-BetaMix, since the model needs to be reestimated in every iteration (details below), while IHW-Storey-BetaMix only needs to estimate it 5 times, once for each fold. Furthermore, we next show that two small modifications suffice to diminish AdaPT's power (but not FDR control guarantees) even under independence. To this end, we provide a summary of how AdaPT works. In iteration j of AdaPT, a candidate rejection function $s_j : \mathcal{X} \to [0, 1]$ is maintained and hypotheses that satisfy $P_i \leq s_j(X_i)$ are in the provisional rejection set. The false discovery proportion at step j is estimated by the Barber and Candès [2015] estimator (cf. Arias-Castro and Chen [2017]):

$$\widehat{\mathrm{FDP}}_j = \frac{1 + \#\{i : P_i > 1 - s_j(X_i)\}}{\#\{i : P_i \leq s_j(X_i)\}} \tag{13}$$

If $\widehat{\mathrm{FDP}}_j \leq \alpha$, the algorithm terminates and returns the current rejection set. Otherwise the rejection region s_j is further shrunk to s_{j+1} with $s_{j+1}(x) \leq s_j(x)$ for all x. The iteration continues until either the stopping criterion is satisfied or the empty set is returned.

A first complication of (13) is that AdaPT must reject at least $1/\alpha$ hypotheses or none at all. For example, for $\alpha = 0.05$, if there are 19 very small p-values, AdaPT may not be able to reject them, even if BH could. Hence AdaPT has low power in situations with very sparse signals, where the best one could hope for is to detect a handful of hypotheses. This is apparent in Figure 4, in the lowest signal situation ($\bar{\beta} = 1.0$). There AdaPT has FDR substantially below the nominal α and furthermore has power a lot smaller than IHW-Storey-BetaMix.

A second complication is that AdaPT can be conservative when the null p-value distribution is strictly super-uniform instead of uniform, because the numerator in (13) will overestimate the false discoveries. In applications, a strictly super-uniform distribution is typically caused by discrete p-values or when the researcher is testing for a one-sided alternative using a test calibrated to effect size zero, but many nulls have an effect in the opposite direction. To explore such enrichment of large p-values, we repeat the previous simulation with $P_i \mid (H_i = 0) \sim (1 - \kappa)\, U[0, 1] + \kappa\, \mathrm{Beta}(1, 0.5)$, varying $\kappa \in [0, 0.1]$ and fixed $\bar{\beta} = 2$. Our previous simulations correspond to $\bar{p} = 0$, which yields the uniform null distribution. Fig. 5A shows the null density as κ varies, and panels B,C show the results of the simulation. We see that as κ increases, the FDR of AdaPT quickly drops below the nominal α and as a consequence power deteriorates.

5.3 Simultaneous two-sample testing

In this section we provide an example of a covariate X_i that is random and arises from statistical (rather than domain-specific) considerations. We study simultaneous two-sample testing for equality

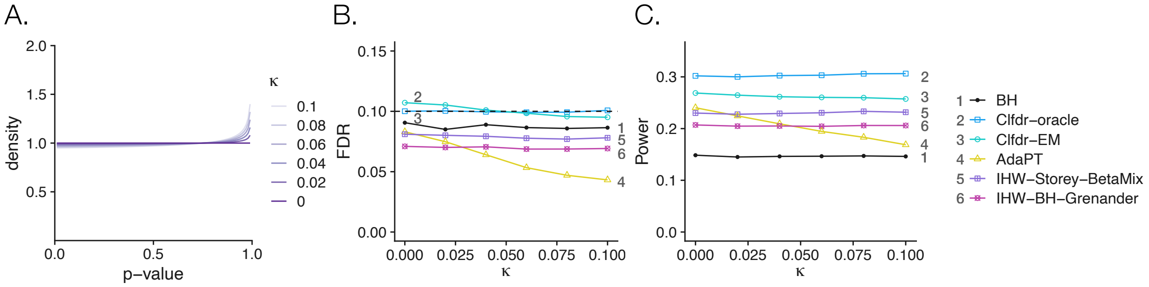


Figure 5: **Simulation for multiple testing with a strictly super-uniform null distribution: A. Density of null p-values** drawn from $(1-\kappa)U[0,1]+\kappa\text{Beta}(1,0.5)$ for varying κ. **B, C. Power and FDR control** under same simulation setting as Fig. 4, but with $\bar{\beta}=2$ fixed and κ varying (Fig. 4 corresponds to $\kappa=0$).

of means following Cai et al. [2019]. For the i-th hypothesis we observe

$$Y_{i,1},\dots,Y_{i,n} \sim \mathcal{N}(\mu_{Y,i},\sigma_i^2) \quad \text{and} \quad V_{i,1},\dots,V_{i,n} \sim \mathcal{N}(\mu_{V,i},\sigma_i^2) \tag{14}$$

(everything independent). We are interested in testing $H_i : \mu_{Y,i} = \mu_{V,i}$, $i = 1,\dots,m$ and assume the variances σ_i^2 are known[7]. The optimal test statistic (in single hypothesis testing [Lehmann and Romano, 2005]) for this situation is the two-sample z-statistic $Z_i := \sqrt{n/2}\left(\overline{Y_i}-\overline{V_i}\right)/\sigma_i$, where $\overline{Y_i}$ and $\overline{V_i}$ are the sample means in each group. The p-values can be calculated as $P_i = 2\left(1-\Phi(|Z_i|)\right)$, where Φ is the Standard Normal CDF. A basic multiple testing approach consists of applying BH to the p-values P_i.

In addition, denote by $\hat{\mu}_i := \frac{1}{2}\left(\overline{Y_i}+\overline{V_i}\right)$ the pooled average and let $X_i := \sqrt{2n}\hat{\mu}_i/\sigma_i$. A direct covariance calculation reveals that $\text{Cov}(X_i, Z_i) = 0$, and so X_i and Z_i are independent (note the joint normality). Hence we may apply the IHW framework with p-values P_i and covariates X_i.

In single hypothesis testing, there is nothing to be gained from X_i and its usefulness only emerges in the multiple testing setup. X_i is a test statistic for the null hypothesis $\mu_{Y,i} = \mu_{V,i} = 0$. If we believe a-priori that for many of the hypotheses i with $\mu_{Y,i} = \mu_{V,i}$, a sparsity condition holds, so that in fact $\mu_{Y,i} = \mu_{V,i} = 0$, then large absolute values of this statistic are more likely to correspond to alternatives. Note that we did not actually re-specify our null hypothesis from $\mu_{Y,i} = \mu_{V,i}$ to $\mu_{Y,i} = \mu_{V,i} = 0$. We just assumed properties of the null hypotheses to motivate a choice of covariate, and are still testing for $\mu_{Y,i} = \mu_{V,i}$.

In the simulation, which is similar to simulations in Cai et al. [2019], we generate data from model (14) with $m = 10000$, $n = 50$, $\sigma_i = 1$ for all i. Furthermore, we vary m_1, the number of alternatives and let

$$\mu_{Y,i} = \begin{cases} 0.5, & i = 1,\dots,m_1 \\ 0.25, & i = m_1+1,\dots,2m_1 \\ 0, & \text{otherwise} \end{cases}, \quad \mu_{V,i} = \begin{cases} 0, & i = 1,\dots,m_1 \\ 0.25, & i = m_1+1,\dots,2m_1 \\ 0, & \text{otherwise} \end{cases}$$

That is, only the first m_1 hypotheses are alternatives. The next $m - m_1$ hypotheses are nulls with the last $m - 2m_1$ also being nulls with respect to the the screening null $\mu_{Y,i} = \mu_{V,i} = 0$. We compare five methods.

1. The **Benjamini-Hochberg (BH)** procedure applied to P_i and ignoring X_i.
2. The **CARS** procedure (covariate-assisted ranking and screening) [Cai et al., 2019]: CARS is a multiple testing procedure designed specifically for simultaneous two-sample tests based on Z_i and X_i. At a high level, CARS learns a function $(z,x) \mapsto \hat{s}_{\text{CARS}}(z,x)$ and a threshold $\hat{t}_{\text{CARS}}$ and rejects all hypotheses such that $\hat{s}_{\text{CARS}}(Z_i, X_i) \leq \hat{t}_{\text{CARS}}$. Asymptotically, CARS controls

[7] The results extend to unequal sample sizes and to unknown variance. We refer the reader to Supplement S6.2.2 and Bourgon et al. [2010], Liu [2014], Cai et al. [2019].

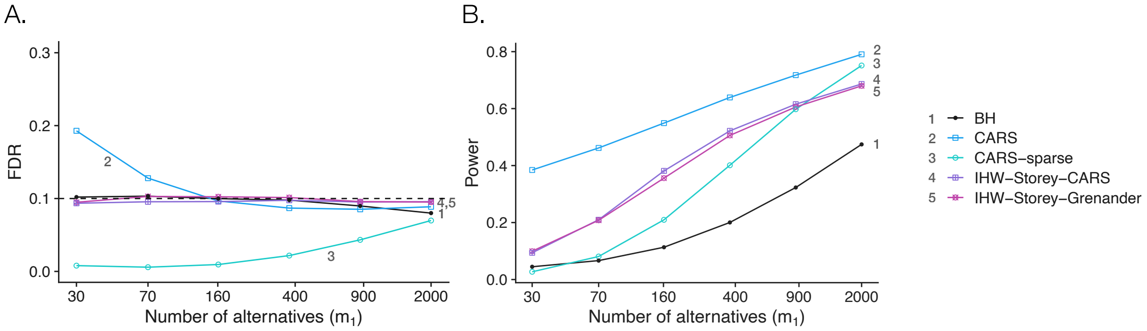


Figure 6: **Simulation for simultaneous two-sample testing: A. False discovery rate** and **B. Power** in model (14) for five methods for simultaneous two-sample testing. The nominal α is equal to 0.1 throughout, and results were averaged over 400 Monte Carlo replicates.

the FDR and learns the optimal decision boundary. We use the default settings of the `CARS` function (`option="regular"`) in the `CARS` R package.

3. **CARS-sparse**: a modification of CARS, also proposed by Cai et al. [2019], that is more conservative and empirically alleviates loss of FDR control in situations with sparse signals (`option="sparse"` in the `CARS` package).
4. **IHW-Storey-CARS:** we use IHW-Storey (Theorem 2) in conjunction with a honest (but not τ-censored) weighting heuristic based on CARS. We partition hypotheses randomly into 5 folds $I_1, \dots, I_5$. To choose weights for I_ℓ we proceed as follows: first, we run CARS on the remaining 4 folds and get $\hat{s}^{-\ell}_{\text{CARS}}(\cdot,\cdot)$ and $\hat{t}^{-\ell}_{\text{CARS}}$. Then, for $i \in I_\ell$, let t_i the smallest threshold at which H_i would get rejected,

$$t_i := \inf\left\{z \geq 0 : \; \hat{s}^{-\ell}_{\text{CARS}}(z, X_i) \leq \hat{t}^{-\ell}_{\text{CARS}}\right\}$$

   Then we let $\tilde{W}_i = 2\left(1 - \Phi(t_i)\right)$, $W_i = |I_\ell|\, \tilde{W}_i / \sum_{j \in I_\ell} \tilde{W}_j$ and finally apply the IHW-Storey procedure from Theorem 2.
5. **IHW-Storey-Grenander**, as in the grouped multiple testing simulations of Section 5.1; we discretize the covariate X_i into 10 groups with 1000 observations each.

The results are shown in Fig. 6. With sparse signal (small m_1), CARS fails to control the FDR. This observation had also been made by Cai et al. [2019], who therefore proposed an elaborate modification, CARS-sparse, which indeed controls FDR in our simulation, as do all other methods. On the other hand, IHW-Storey-CARS is easy to implement—using existing software for CARS—and turns out to have more power in the simulations than CARS-sparse. IHW-Storey-Grenander also has more power than CARS-sparse.

6 Application example: biological high-throughput data

Grubert et al. [2015] assayed cell lines derived from 75 human individuals for the status of their single nucleotide polymorphisms (SNPs, i.e., differences that exist between the genome sequences of individuals) and a biochemical modification of DNA-associated molecules called H3K27ac. We tested all within-chromosome associations by marginal regression of the quantitative readout from the ChiP-seq assay for H3K27ac on the polymorphisms, which are encoded as categorical variables with levels aa, ab, bb, using the software `Matrix eQTL` [Shabalin, 2012]. Here we restrict ourselves to associations in Chromosomes 1 and 2, for which Grubert et al. reported the status of $N_1 = 645452$ and $N_2 = 699343$ SNPs and the H3K27ac levels at $K_1 = 12193$ and $K_2 = 11232$ genomic positions ("peaks") on these chromosomes. This results in a total of approximately 16 billion hypotheses ($m = N_1 \times K_1 + N_2 \times K_2 \approx 1.6 \cdot 10^{10}$)[8]. Figure 1 shows the marginal histogram of the p-values and

[8] We note that computing and storing 16 billion p-values puts notable demands on computing infrastructure. There-

illustrates how these p-values are related to the genomic distance between SNP and H3K27ac peak. This covariate is motivated from biological domain knowledge: associations across shorter distances are a-priori more plausible and empirically more frequent.

We compare two different approaches of dealing with the multiplicity, while controlling the FDR:

1. The **Benjamini-Yekutieli (BY)** procedure on the m p-values (at level $\alpha = 0.01$): Such a conservative procedure is justified, since p-values for the same H3K27Ac peak and different, but genetically linked SNPs will be strongly dependent.
2. The **IHW-BY-Grenander** method (at level $\alpha = 0.01$) using as covariate the genomic distance between SNP and H3K27ac peak and weights based on the Grenander estimator after binning based on genomic distance; cf. Section 4.2 and Supplement S4.1 for a description of the algorithm and Supplement S5 for application-specific details. To satisfy Assumption 2 and hence have guaranteed FDR control by Theorem 4, we partition p-values into two folds corresponding to the different chromosomes. The data for these are, to sufficient approximation, independent.

The results are shown in Figure 7. IHW more than doubles the discoveries compared to the unweighted procedure while maintaining all formal guarantees of FDR control. Panel A shows the learned weight functions for the two folds. Upon applying the weighted BY procedure, the weights translate into thresholds for rejection: hypothesis i is rejected if $P_i \leq W_i \, \hat{t}^*_{\text{IHW}}$ for some common choice of $\hat{t}^*_{\text{IHW}}$ and hypothesis-dependent W_i (Panel D). In contrast, the BY procedure uses the same rejection threshold $\hat{t}^*_{\text{BY}}$ for all hypotheses (Panel C). As a consequence, the BY procedure had to be relatively stringent throughout, while IHW could be permissive at smaller and stringent only at higher distances.

There is another interpretation explaining why IHW increases power: it attempts to set thresholds in a way that balances the conditional local false discovery rate (fdr), at least among the non-zero thresholds. This is shown in Panel F. Indeed, under certain assumptions, the optimal decision boundary is one of constant local fdr, cf. Lei and Fithian [2018, Theorem 2]. On the other hand, since BY thresholds only depend on the p-values, the local fdr varies widely and increases as a function of genomic distance, as seen in Panel E.

Finally, we note that the estimation method for the local fdr in Panels E and F is the same that was used to derive the weights. The local fdr estimates appear to be noisy, while the learned weights do not. This is a key feature of IHW; even inaccurate estimates of the local fdr can lead to powerful weights (increase in number of discoveries). Furthermore, the frequentist guarantees of type-I error control are independent of and unaffected by (in)accuracies of the local fdr estimate.

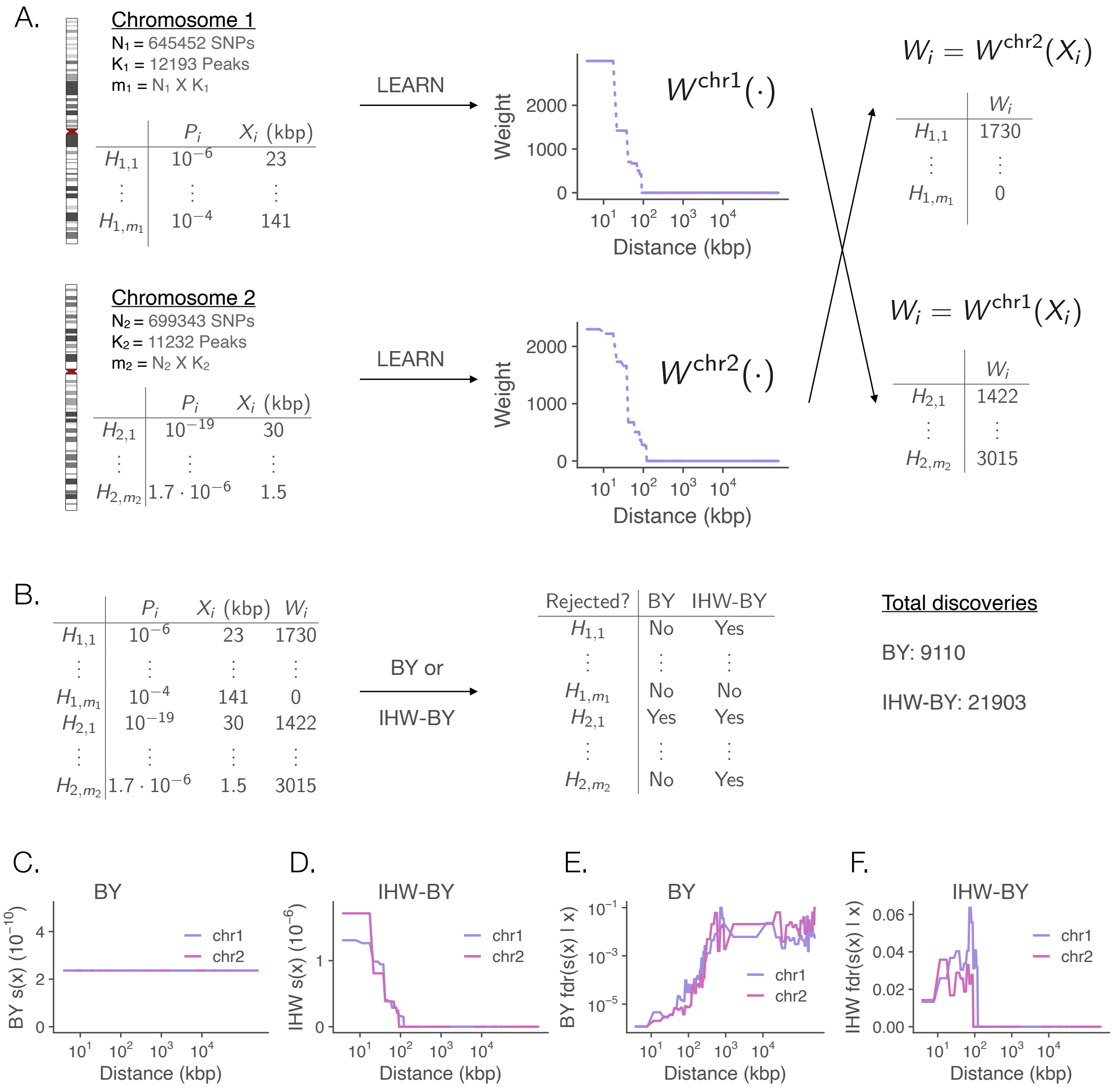
7 Further relations to previous work

Throughout this manuscript we have emphasized the relationship of the present research to previous work. In particular, in our numerical study in Section 5 we compared IHW to previously developed methods for grouped multiple testing, multiple testing with continuous covariates and simultaneous two-sample testing. In this section we provide some further connections of IHW to previous work.

7.1 Ignatiadis, Klaus, Zaugg, and Huber [2016]

The idea of cross-weighting for FDR control was introduced as one of three empirically promising heuristics by Ignatiadis, Klaus, Zaugg, and Huber [2016]; the other two heuristics being convex relaxations and regularization of the weights towards unity and/or low total variation. The contribution of this paper relative to Ignatiadis et al. [2016] is to clarify essential versus circumstantial concepts (e.g., Ignatiadis et al. [2016] only considered one possibility for weighting hypotheses through the Grenander estimator) and to establish formal, finite-sample FDR control for IHW-BH. We

fore, a common choice made by implementations such as Matrix eQTL [Shabalin, 2012] to reduce storage requirements is to only report p-values below some threshold (e.g., in this case, below 10^{-4}). Benjamini-Hochberg/Yekutieli and IHW-BH/BY can deal with this seamlessly by operating as if the right-censored p-values were equal to 1. In contrast, AdaPT critically depends on the large p-values to estimate the FDR, cf. (13).



	P_i	X_i (kbp)
$H_{1,1}$	10^{-6}	23
⋮	⋮	⋮
H_{1,m_1}	10^{-4}	141

	W_i
$H_{1,1}$	1730
⋮	⋮
H_{1,m_1}	0

	P_i	X_i (kbp)
$H_{2,1}$	10^{-19}	30
⋮	⋮	⋮
H_{2,m_2}	$1.7 \cdot 10^{-6}$	1.5

	W_i
$H_{2,1}$	1422
⋮	⋮
H_{2,m_2}	3015

	P_i	X_i (kbp)	W_i
$H_{1,1}$	10^{-6}	23	1730
⋮	⋮	⋮	⋮
H_{1,m_1}	10^{-4}	141	0
$H_{2,1}$	10^{-19}	30	1422
⋮	⋮	⋮	⋮
H_{2,m_2}	$1.7 \cdot 10^{-6}$	1.5	3015

Rejected?	BY	IHW-BY
$H_{1,1}$	No	Yes
⋮	⋮	⋮
H_{1,m_1}	No	No
$H_{2,1}$	Yes	Yes
⋮	⋮	⋮
H_{2,m_2}	No	Yes

Figure 7: **Biological data example revisited: A. Schematic representation of cross-weighting:** we consider a multiple testing situation with $m = m_1 + m_2$ hypotheses that can be partitioned into two independent folds (here: two chromosomes). Besides the p-value P_i, a covariate X_i is available for each hypothesis ($i = 1, \dots m$), which here is the genomic distance between SNP and peak. For each fold we learn the optimal weight function and assign weights to hypotheses from fold 1 using the function learned from the $(P_i, X_i)_i$ of fold 2, and vice versa. **B. Data-driven weighting increases power:** Upon merging the two tables of hypotheses, we apply the Benjamini-Yekutieli (BY) method at $\alpha = 0.01$ to the p-values, or the weighted BY method with the learned weights (IHW-BY). Each method returns a list of rejected hypotheses. IHW more than doubles the total number of discoveries. **C, D. Decision boundaries for BY and IHW-BY:** BY rejects all hypotheses with p-value P_i below a fixed threshold, while IHW-BY rejects hypotheses with $P_i \leq s_l(X_i)$, where $l \in \{1, 2\}$ denotes the fold, and the threshold depends on the covariate X_i. The threshold is more lenient for hypotheses with smaller genomic distance X_i. For larger X_i, the threshold becomes smaller (more stringent); in this example application, it reaches 0 for very large X_i. **E, F. Estimated conditional local fdr at the BY and IHW-BY rejection thresholds.** We observe that for BY the conditional local fdr varies widely, while for IHW-BY it is approximately balanced at the non-zero thresholds (note the different scales of the y-axis in panels E,F). The conditional density $f(t \mid x)$ is estimated by binning along X_i and applying the Grenander estimator within each bin. We set $f(0 \mid x) = \infty$, so that the conditional local fdr is 0 when $s(x) = 0$.

also show how the fundamental idea of cross-weighting applies beyond independence and introduce cross-weighted variants of the k-Bonferroni and BY procedures for k-FWER and FDR control under dependence.

7.2 Sample splitting

One of the initial attempts at data-driven weights [Rubin et al., 2006] used another form of data-splitting: consider the setting where we start with a $m \times n$ data-matrix from which we get our p-values P_i by calculating the test statistic in a row-wise fashion, say by applying a t-test for each row. Then one can calculate m "prior" p-values P_i'' based on $n_1 < n$ columns and derive prior weights W_i based on P_i''. The remaining $n - n_1$ columns are used to compute p-values P_i'. Finally, a weighted multiple testing procedure is applied with p-values P_i' and weights W_i. However, the authors then show that in this case it is more powerful to simply use an unweighted procedure with p-values P_i calculated based on the whole dataset, rather than a weighted procedure with sample-splitting. Habiger and Peña [2014] pursue a similar approach. For IHW, we instead split horizontally (on hypotheses) rather than vertically (on samples), and the p-values P_i are unaltered.

7.3 The weighted False Discovery Rate

In this work, we have studied heterogeneous multiple testing with the aim of increasing power, while controlling the k-FWER or the FDR. However, in light of non-exchangeability, the cost of a false discovery to the researcher may not be uniform, but vary across hypotheses; e.g., it may be equal to $a_i \geq 0$ for hypothesis H_i. Then it is of scientific interest to control the weighted FDR of Benjamini and Hochberg [1997] defined as

$$\text{wFDR}(\mathbf{a}) := \mathbb{E}\left[\frac{\sum_{i \in \mathscr{H}_0} a_i \mathbf{1}_{\{H_i \text{ rejected}\}}}{\sum_{i=1}^m a_i \mathbf{1}_{\{H_i \text{ rejected}\}}} \mathbf{1}\left(\sum_{i=1}^m a_i \mathbf{1}_{\{H_i \text{ rejected}\}} > 0\right)\right].$$

Similarly, the utility (benefit) b_i of a true discovery may vary across hypotheses. Then, instead of maximizing the expected number of discoveries (cf. Section 4), it may be more pertinent to maximize the expected total benefit. Basu et al. [2018] study optimal oracle procedures that achieve this optimization goal subject to control of wFDR($\mathbf{a}$), as well as data-driven procedures that achieve the same goal asymptotically. In future work it would be of interest to study whether cross-weighting may be applied to derive flexible and powerful procedures with finite-sample conrol of wFDR($\mathbf{a}$). We expect this to be tractable –for example by leveraging the results of Ramdas et al. [2019]– and useful if the utility b_i is a function of the covariates, i.e., $b_i = b(X_i)$.

8 Discussion

Despite the ubiquitous uptake by the natural sciences of the concepts of multiple testing (and in particular the FDR), and despite ever growing volumes of data and possible hypothesis tests, surprisingly little attention has been paid to systematic approaches to account for hypothesis heterogeneity in order to increase detection power. While this may be justifiable in situations where power is large anyway, in many cases the costs of the underlying experiments or studies are substantial and increase with sample size, and the question of power decides over success or failure. In such cases, an approach that increases power compared to a baseline analysis, at no cost and by purely computational means, should be of interest.

Our approach is an instance of the value of large scale data [Efron, 2010]: due to dataset size, modeling and inference opportunities open up that were previously irrelevant or impossible. In addition to the p-values P_i, our approach uses two further inputs: the covariates X_i and the fold assignment. These are different concepts and their construction is unrelated to each other. The X_i are informative about power and/or prior probability of the tests, but independent of P_i under the null hypothesis. Meanwhile, the folds are constructed as a device for the cross-weighting scheme, in order to achieve type-I error control: we want independence of folds so that the weights do not lead to overfitting. Their choice is unrelated to power. Random folds are an easy default, but to get independent folds, it is then necessary to require global independence (Assumption 1). When global independence cannot be assumed, the dependences are in many application scenarios—loosely speaking—"local" (under some suitable choice of metric on the set of hypotheses). This can be used to construct folds that are independent, at least to sufficient approximation. Making such loose speak more precise requires specification of individual application scenarios and the associated domain knowledge, as in the example of Section 6.

If, for a dataset at hand, independent folds cannot be achieved by any available fold-splitting scheme, it is possibly better not to try to address the dependences at the level of the multiple testing procedure, but upstream: strong, dataset-wide dependences often signal the need for a fundamental rethink of the analysis approach.

Sometimes, dataset-wide dependences are caused by so-called *batch effects*. They are undesirable, uninteresting with respect to the scientific question, and can be reduced or avoided by good experimental design [Leek et al., 2010]. Once they are a matter of fact, it is sometimes possible to remove them by mapping the data to a new set of properly "normalized" and "batch-corrected" variables [Leek and Storey, 2008, Stegle et al., 2010, Wang et al., 2017].

If avoiding dependence by modifying the analysis upstream of the multiple testing treatment is not possible, the analyst should also consider whether multiple marginal hypothesis tests are indeed more appropriate than, say, dimension reduction, or a multivariate model with FDR guarantees [Candès et al., 2018, Sesia et al., 2019, Ren and Candès, 2020].

Code availability and reproducibility

The study is made fully third-party reproducible, and we provide its code in Github under the repository `https://github.com/nignatiadis/IHWStatsPaper`. The Bioconductor package IHW (`http://bioconductor.org/packages/IHW`) provides a user-friendly implementation of IHW-BH/Storey based on the Grenander estimator.

Acknowledgments

We thank Judith Zaugg for making available data for the example in Section 6, and Edgar Dobriban, William Fithian, Susan Holmes, Lihua Lei, Michael Love, Stelios Serghiou, Michael Sklar, Youngtak Sohn, Oliver Stegle, Mark van de Wiel, Britta Velten and Stefan Wager for helpful discussions and critical comments on the manuscript. Michael Sklar proposed the counterexample from Supplement S1.3.

References

David B Allison, Gary L Gadbury, Moonseong Heo, José R Fernández, Cheol-Koo Lee, Tomas A Prolla, and Richard Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20, 2002.

Ery Arias-Castro and Shiyun Chen. Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1):1983–2001, 2017.

J. Kenneth Baillie, Andrew Bretherick, Christopher S. Haley, Sara Clohisey, Alan Gray, Lucile P. A. Neyton, Jeffrey Barrett, Eli A. Stahl, Albert Tenesa, Robin Andersson, J. Ben Brown, Geoffrey J. Faulkner, Marina Lizio, Ulf Schaefer, Carsten Daub, Masayoshi Itoh, Naoto Kondo, Timo Lassmann, Jun Kawai, Damian Mole, Vladimir B. Bajic, Peter Heutink, Michael Rehli, Hideya Kawaji, Albin Sandelin, Harukazu Suzuki, Jack Satsangi, Christine A. Wells, Nir Hacohen, Thomas C. Freeman, Yoshihide Hayashizaki, Piero Carninci, Alistair R. R. Forrest, and David

A. Hume and. Shared activity patterns arising at genetic susceptibility loci reveal underlying genomic and cellular architecture of human disease. *PLOS Computational Biology*, 14(3):e1005934, 2018.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Pallavi Basu, T Tony Cai, Kiranmoy Das, and Wenguang Sun. Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523): 1172–1183, 2018.

Yoav Benjamini. Comment: Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23(1):23–28, 2008.

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

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

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, pages 1165–1188, 2001.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic journal of Statistics*, 2:963–992, 2008.

Simina M Boca and Jeffrey T Leek. A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6:e6035, 2018.

Carlo E Bonferroni. *Il calcolo delle assicurazioni su gruppi di teste*. Studi in Onore del Professore Salvatore Ortu Carboni, Rome, Italy, 1935.

Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107 (21):9546–9551, 2010.

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

T Tony Cai, Wenguang Sun, and Weinan Wang. Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):187–234, 2019.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

Nabarun Deb, Sujayam Saha, Adityanand Guntuboyina, and Bodhisattva Sen. Two-component mixture model in the presence of covariates. *arXiv preprint arXiv:1810.07897*, 2018.

Edgar Dobriban, Kristen Fortney, Stuart K Kim, and Art B Owen. Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika*, 102(4):753–766, 2015.

Lilun Du and Chunming Zhang. Single-index modulated multiple testing. *The Annals of Statistics*, 42(4):30–79, 2014.

Guillermo Durand. Adaptive p-value weighting with power optimality. *arXiv preprint arXiv:1710.01094v1*, 2017.

Guillermo Durand. Adaptive p-value weighting with power optimality. *Electronic Journal of Statistics*, 13(2):3336–3385, 2019.

Bradley Efron. Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics*, pages 197–223, 2008.

Bradley Efron. *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction.* Cambridge University Press, 2010.

Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

Egil Ferkingstad, Arnoldo Frigessi, Håvard Rue, Gudmar Thorleifsson, and Augustine Kong. Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, pages 714–735, 2008.

Kristen Fortney, Edgar Dobriban, Paolo Garagnani, Chiara Pirazzini, Daniela Monti, Daniela Mari, Gil Atzmon, Nir Barzilai, Claudio Franceschi, Art B Owen, and Stuart K Kim. Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genetics*, 11 (12):e1005728, 2015.

Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, pages 1035–1061, 2004.

Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

Ulf Grenander. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(1): 70–96, 1956.

Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksandra Pekowska, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162(5): 1051–1065, 2015.

Wenge Guo and Sanat Sarkar. Adaptive controls of FWER and FDR under block dependence. *Journal of Statistical Planning and Inference*, 2019.

Joshua Habiger, David Watts, and Michael Anderson. Multiple testing with heterogeneous multinomial distributions. *Biometrics*, 73(2):562–570, 2017.

Joshua D Habiger. Adaptive false discovery rate control for heterogeneous data. *Statistica Sinica*, pages 1731–1756, 2017.

Joshua D Habiger and Edsel A Peña. Compound p-value statistics for multiple testing procedures. *Journal of multivariate analysis*, 126:153–166, 2014.

T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics. Springer, 2009. ISBN 9780387848587.

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

James X Hu, Hongyu Zhao, and Harrison H Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491), 2010.

Nikolaos Ignatiadis and Stefan Wager. Covariate-powered empirical Bayes estimation. In *Advances in Neural Information Processing Systems*, pages 9620–9632, 2019.

Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 2016.

Bernd Klaus and Korbinian Strimmer. Learning false discovery rates by fitting sigmoidal threshold functions. *Journal de la Société Française de Statistique*, 152(2):39–50, 2011.

Keegan Korthauer, Patrick K Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J Alm, and Stephanie C Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1):118, 2019.

Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10): 733–739, 2010.

EL Lehmann and JP Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 2005. ISBN 9780387988641.

Lihua Lei and William Fithian. AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

Ang Li and Rina Foygel Barber. Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1): 45–74, 2019.

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

Weidong Liu. Incorporation of sparsity information in large-scale multiple two-sample t tests. *arXiv preprint arXiv:1410.4282*, 2014.

Robin Lougee-Heimer. The common optimization interface for operations research: Promoting open-source software in the operations research community. *IBM Journal of Research and Development*, 47(1):57–66, 2003.

Anastasios Markitsis and Yinglei Lai. A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, 26(5):640–646, 2010.

Xinkun Nie and Stefan Wager. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika*, 09 2020. asaa076.

Alejandro Ochoa, John D Storey, Manuel Llinás, and Mona Singh. Beyond the E-value: Stratified statistics for protein domain prediction. *PLoS Computational Biology*, 11(11):e1004509, 11 2015.

Edsel A Peña, Joshua D Habiger, and Wensong Wu. Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *The Annals of Statistics*, 39(1):556–583, 2011.

Alexander Ploner, Stefano Calza, Arief Gusnanto, and Yudi Pawitan. Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, 22(5):556–565, 2006.

Aaditya K Ramdas, Rina F Barber, Martin J Wainwright, and Michael I Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47 (5):2790–2821, 2019.

Zhimei Ren and Emmanuel Candès. Knockoffs with side information. *arXiv preprint arXiv:2001.07835*, 2020.

R Tyrrell Rockafellar. *Convex analysis*. Number 28 in Princeton Landmarks in Mathematics and Physics. Princeton university press, 1970.

Kathryn Roeder and Larry Wasserman. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science*, 24(4):398, 2009.

Kathryn Roeder, Bernie Devlin, and Larry Wasserman. Improving power in genome-wide association studies: weights tip the scale. *Genetic Epidemiology*, 31(7):741–747, 2007.

Joseph P Romano and Michael Wolf. Balanced control of generalized error rates. *The Annals of Statistics*, 38(1):598–633, 2010.

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

Daniel Rubin, Sandrine Dudoit, and Mark Van der Laan. A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006.

Kris Sankaran and Susan Holmes. structSSI: Simultaneous and selective inference for grouped or hierarchically structured data. *Journal of statistical software*, 59(13):1, 2014.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

James G Scott, Ryan C Kelly, Matthew A Smith, Pengcheng Zhou, and Robert E Kass. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471, 2015.

Matteo Sesia, Chiara Sabatti, and Emmanuel J Candès. Gene hunting with knockoffs for hidden markov models. *Biometrika*, 106:1–18, 2019.

Andrey A Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, 6(5):e1000770, 2010.

John D Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, pages 2013–2035, 2003.

John D Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368, 2007.

John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

Korbinian Strimmer. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12):1461–1462, 2008a.

Korbinian Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9 (1):303, 2008b.

Lei Sun, Radu V Craiu, Andrew D Paterson, and Shelley B Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530, 2006.

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

Wenguang Sun and T Tony Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424, 2009.

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98 (9):5116–5121, 2001.

AW van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. ISBN 9781107268449.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Jingshu Wang, Qingyuan Zhao, Trevor Hastie, and Art B Owen. Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics*, 45(5):1863–1894, 2017.

Li Wang. Weighted multiple testing procedure for grouped hypotheses with k-FWER control. *Computational Statistics*, pages 1–25, 2018.

Mark A Wiel, Tonje G Lien, Wina Verlaat, Wessel N Wieringen, and Saskia M Wilting. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*, 35(3):368–381, 2016.

Martin J Zhang, Fei Xia, James Y Zou, and David Tse. NeuralFDR: Learning discovery thresholds from hypothesis features. In *Advances in Neural Information Processing Systems*, pages 1540–1549, 2017.

Martin J Zhang, Fei Xia, and James Zou. Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Communications*, 10(1):1–11, 2019.

Haibing Zhao and Jiajia Zhang. Weighted p–value procedures for controlling FDR of grouped hypotheses. *Journal of Statistical Planning and Inference*, 2014.

Supplement S1: Finite-sample results for FDR control of IHW

Throughout Supplementary Section S1, the weights W_i are considered random. Occasionally we explicitly condition on the weights; in which case we verify how the conditioning on (subsets of) weights influences conditional distributions.

S1.1 A preliminary lemma

They key property of IHW that enables finite-sample type-I error control is the following: cross-weighting makes the p-values and their weights independent of each other. This was already demonstrated in the beginning of the proof of Theorem 3 in Section 3.2. Here we formalize this result through the following Lemma:

Lemma 1. Let $(W_i)_{i\in[m]}$ be honest weights (Specification 1) w.r.t. the partition $I_1,\dots,I_K$ of $[m]$. If $((P_i,X_i))_{i\in[m]}$ satisfy Assumption 2, then:

(a) For all $\ell\in[K]$ and all $i\in\mathscr{H}_0\cap I_\ell$, P_i is independent of $(W_k)_{k\in I_\ell}$. In particular P_i is independent of W_i ($P_i\perp W_i$) for all $i\in\mathscr{H}_0$.

The conclusion may be strengthened if instead $((P_i,X_i))_{i\in[m]}$ satisfy Assumption 1:

(a') For all $\ell\in[K]$, $(P_i)_{i\in\mathscr{H}_0\cap I_\ell}$ is independent of $(W_i)_{i\in\mathscr{H}_0\cap I_l}$.

(b') For all $\ell\in[K]$, $(P_i)_{i\in\mathscr{H}_0\cap I_\ell}$ are jointly independent and super-uniform conditionally on $(W_i)_{i\in\mathscr{H}_0\cap I_\ell}$.

Proof. We prove (a); the other statements follow similarly. Fix $\ell\in[K]$ and let $i\in\mathscr{H}_0\cap I_\ell$. By definition of honesty (Specification 1), $(W_k)_{k\in I_\ell}$ is a function only of $(P_i)_{i\in I_\ell^c}$ and $\mathbf{X}=(X_i)_{i\in[m]}$. It thus suffices to argue that P_i is independent of $((P_i)_{i\in I_\ell^c},\mathbf{X})$. Writing the latter as $((P_i)_{i\in I_\ell^c},(X_i)_{i\in I_\ell^c},(X_i)_{i\in I_\ell})$ we conclude as a consequence of parts (a) and (b) of Assumption 2. □

S1.2 The IHW-BH procedure under independence: Proof of Theorem 1

Proof. Let $\mathbf{W}$ be the weights and $\hat{k}$ the number of discoveries after applying IHW-BH at level α and with censoring level τ. Also write $\mathbf{X}=(X_1,\dots,X_m)$, $\mathbf{P}=(P_1,\dots,P_m)$ and $\mathbf{1}(\mathbf{P}\leq\tau)=(\mathbf{1}(P_1\leq\tau),\dots,\mathbf{1}(P_m\leq\tau))$. Here $\mathbf{1}(P_i\leq\tau)$ is the indicator function that is 1 when $P_i\leq\tau$ and 0 otherwise.

We first give a high level idea regarding the proof. To bound the FDR we seek to bound expectations of $\mathbf{1}(H_i \text{ rejected})/(\hat{k}\vee 1)$, i.e., of $\mathbf{1}(P_i\leq\alpha W_i\hat{k}/m,\ P_i\leq\tau)/(\hat{k}\vee 1)$ where i is null.[9] If $W_i,\hat{k}$ were independent of P_i, then we could directly upper bound this expectation by $\mathbb{E}[(\alpha W_i\hat{k}/m)/(\hat{k}\vee 1)]\leq\mathbb{E}[\alpha W_i/m]$ from which FDR control would follow by summing over all i. Honesty (Specification 1) makes—in the way of Lemma 1—P_i and its weight W_i (for a single null i) independent. However, P_i directly influences $\hat{k}$. This is true also for unweighted BH and weighted BH with deterministic weights, yet here P_i also indirectly influences $\hat{k}$ through the weights $W_j,\ j\neq i$. Nevertheless, we will argue that the conclusion may still be salvaged: τ-censoring (Specification 2) ensures that on the event $\{P_i\leq\tau\}$ the exact value of P_i cannot influence weights $W_j,\ j\neq i$. Furthermore, it suffices to only consider the event $\{P_i\leq\tau\}$ (in turn for each null i), since i will never get rejected when $P_i>\tau$ (by Definition 2).

We make the above intuition rigorous using a leave-one-out argument as in the proof idea of Li and Barber [2019]. Let us first pay attention to a single index $i\in[m]$. We denote by k_i the number of discoveries of IHW-BH if $\mathbf{P}$ gets replaced by $\mathbf{P}_{i\mapsto 0}=(P_1,\dots,P_{i-1},0,P_{i+1},P_m)$. Note that because the weights are τ-censored (Specification 2) , the tuple of weights $\mathbf{W}$ remains unchanged by replacing P_i by 0 on the event $\{P_i\leq\tau\}$. Furthermore, by definition of the τ-censored weighted BH procedure (Definition 2), the rejection of H_i (by IHW-BH applied to $\mathbf{P}$) implies that

[9] We use the notation $a\vee b=\max\{a,b\}$, $a\wedge b=\min\{a,b\}$.

$P_i \leq \left(\frac{\alpha W_i \hat{k}}{m}\right) \wedge \tau$. In particular, the event $\{P_i \leq \tau\}$ holds. Furthermore, for any $k \geq \hat{k}$, counting the entries of $\mathbf{P}$, respectively $\mathbf{P}_{i \mapsto 0}$, that 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 $\mathbf{W}, \mathbf{X}, \mathbf{P}_{i \mapsto 0}, \mathbf{1}(\mathbf{P} \leq \tau)$, we get

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

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 $(\mathbf{P}_{i \mapsto 0}, \mathbf{X})$ and also because k_i, $\mathbf{W}$, $\mathbf{1}(\mathbf{P} \leq \tau)$ are functions of $(\mathbf{P}_{i \mapsto 0}, \mathbf{X})$ on the event $\{P_i \leq \tau\}$. It then follows that

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

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

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

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 diverge from the proof of Li and Barber [2019] and take advantage of the honesty assumptions (Specification 1) through Lemma 1.

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

Going from the second to the third line, we used that for $i \in \mathscr{H}_0$, P_i is independent of W_i, which holds from Lemma 1(a'). In the last step, we used Part (b) of the Honesty specification.

□

S1.3 Counterexample to demonstrate that honesty of weights does not suffice for FDR control

In this section, we provide a counterexample that the result of Theorem 1 no longer holds if we drop the assumption of τ-censored weighting. This is in contrast e.g., to the conclusion of Theorem 3 for k-Bonferroni, wherein honesty of the weights suffices (along with distributional assumptions on (P_i, X_i)).

Our agenda is as follows: for $m = 4$, we construct $((P_i, X_i))_{i \in [m]}$ under the global null such that Assumption 1 holds. Then we construct honest weights W_i (Specification 1) and finally we apply the weighted BH procedure at level $\alpha \in (0, 1)$ (Definition 2 with $\tau = 1$) with p-values P_i and weights W_i. We will show this procedure does not control the FDR at the nominal level.

We observe four independent and uniform (null) p-values P_1, P_2, P_3, P_4. Our covariates take values $X_i = i$. We partition the hypotheses into the folds $\{1, 2\}$ and $\{3, 4\}$. The (adversarial) honest weighting scheme is as follows: If $\frac{\alpha}{2} \leq P_1 \leq \alpha$, assign $W_3 = 2, W_4 = 0$. Otherwise assign $W_3 = 0, W_4 = 2$. Similarly if $\frac{\alpha}{2} \leq P_3 \leq \alpha$, then assign $W_1 = 2, W_2 = 0$ and otherwise $W_1 = 0, W_2 = 2$. These weights are honest; note that $W_i \geq 0$ for all i, $\sum_{i=1}^4 W_i = 4$.

To study the FDR of this procedure we partition the sample space according to the four possibilities for the weight assignment. Also note that due to the weighting scheme in the end we will be applying unweighted Benjamini-Hochberg to two hypotheses at level α. For notational convenience we will write $\text{BH}(P_i, P_j)$ for the event that BH applied to P_i, P_j at level α rejects at least one of these two p-values.

Case 1: Here we have $W_2 = W_4 = 2$ and $W_1 = W_3 = 0$. Thus we are just doing unweighted Benjamini-Hochberg on the p-values P_2 and P_4. Noting that occurence of this case depends only on P_1, P_3, we get by independence:

$$\mathbb{P}[\text{Case 1 occurs}, \text{BH}(P_2, P_4)] = \mathbb{P}[\text{Case 1 occurs}]\mathbb{P}[\text{BH}(P_2, P_4)] = \left(1 - \frac{\alpha}{2}\right)^2 \alpha\,.$$

Case 2: Now consider $W_1 = W_3 = 2$ and $W_2 = W_4 = 0$. In this case, we know that both $\frac{\alpha}{2} \leq P_1 \leq \alpha$ and $\frac{\alpha}{2} \leq P_3 \leq \alpha$. These in turn imply that $\text{BH}(P_1, P_3)$ also holds (in fact BH rejects both hypotheses). Thus:

$$\mathbb{P}[\text{Case 2 occurs}, \text{BH}(P_1, P_3)] = \mathbb{P}[\text{Case 2 occurs}] = \left(\frac{\alpha}{2}\right)^2\,.$$

Case 3: Now let $W_1 = W_4 = 2$ and $W_2 = W_3 = 0$. Then:

$$\begin{aligned}\mathbb{P}\left[\text{Case 3 occurs}, \text{BH}(P_1, P_4)\right] &= \mathbb{P}\left[P_1 \notin \left[\frac{\alpha}{2}, \alpha\right], \frac{\alpha}{2} \leq P_3 \leq \alpha, \text{BH}(P_1, P_4)\right] \\ &= \mathbb{P}\left[\frac{\alpha}{2} \leq P_3 \leq \alpha\right]\mathbb{P}\left[P_1 \notin \left[\frac{\alpha}{2}, \alpha\right], \text{BH}(P_1, P_4)\right] \\ &= \frac{\alpha}{2}\left[\frac{\alpha}{2} + \frac{\alpha}{2}(1 - \alpha)\right]\,.\end{aligned}$$

The latter is true since if $P_1 \notin [\frac{\alpha}{2}, \alpha]$, the only way BH will reject is if $P_1 < \frac{\alpha}{2}$ or $P_4 \leq \frac{\alpha}{2}$. Hence the event on the RHS can be written as the disjoint union of $\{P_1 < \alpha/2\}$ and $\{P_4 \leq \alpha/2, P_1 > \alpha\}$.

Case 4: By symmetry with Case 3, this contributes the same conditional probability.
Summing up all 4 cases, we see that

$$\text{FDR} = \text{FWER} = \alpha + \frac{\alpha^2}{4}(1 - \alpha) > \alpha$$

Hence FDR is not controlled at the nominal level α.

S1.4 The IHW-Storey procedure under independence: Proof of Theorem 2

Proof. Take $i \in I_\ell \cap \mathscr{H}_0$ and define the leave-one-out null proportion estimator (compare to Equation (5)):

$$\hat{\pi}_{0,I_\ell}^{-i} = \frac{\max\limits_{j \in I_\ell} W_j + \sum\limits_{j \in I_\ell \setminus \{i\}} W_j \mathbf{1}(P_j > \tau')}{|I_\ell|(1-\tau')} \,.$$

Now note that on the event $\{P_i \leq \tau\}$ (since $\tau' \geq \tau$) we have that:

$$\hat{\pi}_{0,I_\ell} = \hat{\pi}_{0,I_\ell}^{-i} \,. \tag{15}$$

Next, define

$$\widetilde{W}_i = \frac{W_i}{\hat{\pi}_{0,I_\ell}^{-i}} \,.$$

(15) implies that running the τ-censored, weighed BH procedure (Definition 2) with p-values P_i and weights $W_i / \hat{\pi}_{0,I_\ell}$ (i.e., the procedure whose FDR control we seek to prove) will have identical rejections if we replace the weights by $\widetilde{W}_i$. Hence we turn to study the procedure with weights $\widetilde{W}_i$. Proceeding as in the the leave-one-out argument of the proof of Theorem 1 we get

$$H_i \text{ rejected} \Rightarrow P_i \leq \frac{\alpha \widetilde{W}_i k_i}{m} \wedge \tau \,.$$

In fact, since $\hat{\pi}_{0,I_\ell}^{-i}$ does not depend on P_i (it depends on $\mathbf{P}_{i \mapsto 0}$), all arguments of the proof of Theorem 1 go through unchanged with $\widetilde{W}_i$ replacing W_i. The only step we need to pay attention to is the last line: it no longer holds that

$$\sum_{i=1}^m \widetilde{W}_i = m \text{ almost surely} \,.$$

Indeed we are hoping that this sum is greater than m so that we can gain power by the null-proportion adaptivity. Instead, it suffices to argue that

$$\sum_{i \in \mathscr{H}_0} \mathbb{E}\left[\widetilde{W}_i\right] \leq m \,.$$

And hence it also suffices to prove that for each fold ℓ the following holds

$$\sum_{i \in \mathscr{H}_0 \cap I_\ell} \mathbb{E}\left[\widetilde{W}_i\right] \leq |I_\ell| \,.$$

To prove this, we first recall from Lemma 1(a') that

$$(P_i)_{i \in \mathscr{H}_0 \cap I_\ell} \perp (W_i)_{i \in \mathscr{H}_0 \cap I_\ell} \,.$$

For notational convenience we write $\mathbf{W}_{\mathscr{H}_0 \cap I_\ell}$ for $(W_i)_{i \in \mathscr{H}_0 \cap I_\ell}$. Then:

$$
\begin{aligned}
\mathbb{E}\left[\widetilde{W}_i \;\middle|\; (W_i)_{i \in \mathscr{H}_0 \cap I_\ell}\right] &= \mathbb{E}\left[\frac{W_i}{\hat{\pi}^{-i}_{0,I_\ell}} \;\middle|\; \mathbf{W}_{\mathscr{H}_0 \cap I_\ell}\right] \\
&= W_i \mathbb{E}\left[\frac{1}{\hat{\pi}^{-i}_{0,I_\ell}} \;\middle|\; \mathbf{W}_{\mathscr{H}_0 \cap I_\ell}\right] \\
&= W_i \mathbb{E}\left[\frac{|I_\ell|(1-\tau')}{\max\limits_{j \in I_\ell} W_j + \sum\limits_{j \in I_\ell \setminus \{i\}} W_j \mathbf{1}(P_j > \tau')} \;\middle|\; \mathbf{W}_{\mathscr{H}_0 \cap I_\ell}\right] \\
&\leq W_i \left|I_\ell\right| (1-\tau') \mathbb{E}\left[\frac{1}{\max\limits_{j \in \mathscr{H}_0 \cap I_\ell} W_j + \sum\limits_{j \in \mathscr{H}_0 \cap I_\ell \setminus \{i\}} W_j \mathbf{1}(P_j > \tau')} \;\middle|\; \mathbf{W}_{\mathscr{H}_0 \cap I_\ell}\right] \\
&\leq W_i \left|I_\ell\right| (1-\tau') \frac{1}{(1-\tau') \sum\limits_{j \in \mathscr{H}_0 \cap I_\ell} W_j} \\
&= \frac{W_i \left|I_\ell\right|}{\sum\limits_{j \in \mathscr{H}_0 \cap I_\ell} W_j}.
\end{aligned}
$$

In the penultimate line we used the Inverse Binomial Lemma (Lemma 3 in Ramdas et al. [2019]), noting that conditionally on $\mathbf{W}_{\mathscr{H}_0 \cap I_\ell}$, the weights in folds ℓ may be treated as deterministic and by Lemma 1(b') the p-values $(P_i)_{i \in \mathscr{H}_0 \cap I_\ell}$ are jointly independent and super-uniform. We conclude our proof by iterated expectation and summing over $i \in \mathscr{H}_0 \cap I_\ell$. □

S1.5 The IHW-BY procedure under dependence: Proof of Theorem 4

Proof. We will equivalently prove that applying the weighted Benjamini-Hochberg procedure (without censoring, i.e., $\tau = 1$) at level α controls the FDR at level $\alpha \sum_{k=1}^m \frac{1}{k}$.

For a probability measure ν on $\mathbb{R}^+$, we define the reshaping function $\tilde{\beta} : \mathbb{R}^+ \to \mathbb{R}^+$ [Blanchard and Roquain, 2008, Ramdas et al., 2019]:

$$
\beta(r) = \int_0^r x d\nu(x)\,.
$$

Furthermore, let $\hat{k}$ be the number of rejections of the IHW-BH procedure applied at level α. Then for arbitrary $c > 0$, $i \in \mathscr{H}_0$ and on the event $\{W_i > 0\}$:

$$
\mathbb{E}\left[\frac{\mathbf{1}\left(P_i \leq \frac{c\alpha W_i}{m} \tilde{\beta}(\hat{k})\right)}{\hat{k} \vee 1} \;\middle|\; W_i\right] = \frac{c\alpha W_i}{m} \mathbb{E}\left[\frac{\mathbf{1}\left(P_i \leq \frac{c\alpha W_i}{m} \tilde{\beta}(\hat{k})\right)}{\frac{c\alpha W_i}{m}(\hat{k} \vee 1)} \;\middle|\; W_i\right] \leq \frac{c\alpha W_i}{m}\,. \tag{16}
$$

The inequality follows from Lemma 3.2. (iii) in Blanchard and Roquain [2008] (also Lemma 1(c) in Ramdas et al. [2019]), which we reproduce in a slightly modified form here for the reader's convenience:

Lemma 2. Let U a super-uniform random variable and $S > 0$ another random variable, then for all fixed $t > 0$:

$$
\mathbb{E}\left[\frac{\mathbf{1}(U \leq t\tilde{\beta}(S))}{tS}\right] \leq 1
$$

We recover (16) by applying Lemma 2 conditionally on W_i with $U = P_i$, $S = \hat{k} \vee 1$ and $t = \frac{c\alpha W_i}{m}$. To do so, note that we may treat $\frac{c\alpha W_i}{m}$ as a constant conditionally on W_i and that $P_i \mid W_i$ is super-uniform, since $P_i \perp W_i$ by Lemma 1(a) and P_i is unconditionally super-uniform.

Inequality (16) also holds true almost surely on the event $\{W_i = 0\}$, as the distribution of $P_i \mid W_i$ cannot have a point mass at 0, since this would contradict super-uniformity. Thus we also get unconditionally that

$$\mathbb{E}\left[\frac{\mathbf{1}\left(P_i \leq \frac{c\alpha W_i}{m}\tilde{\beta}(\hat{k})\right)}{\hat{k} \vee 1}\right] \leq \mathbb{E}\left[\frac{c\alpha W_i}{m}\right].$$

Now, consider the special case in which we use the measure $\nu(x) = \frac{1}{\sum_{k=1}^m \frac{1}{k}} \sum_{k=1}^m \frac{1}{k}\delta_k(x)$, where δ_k is the point mass at k. Then the reshaping function takes the form $\tilde{\beta}(r) = \frac{r}{\sum_{k=1}^m \frac{1}{k}}$. Applying the above result with this $\tilde{\beta}$ and $c = \sum_{k=1}^m \frac{1}{k}$ we get

$$\mathbb{E}\left[\frac{\mathbf{1}\left(P_i \leq \frac{\alpha W_i \hat{k}}{m}\right)}{\hat{k} \vee 1}\right] \leq \frac{\alpha \sum_{k=1}^m \frac{1}{k}}{m}\mathbb{E}[W_i]$$

We conclude by using that $\sum_{i=1}^m W_i = m$ almost surely as follows

$$\begin{aligned}
\mathbb{E}[\mathrm{FDP}] &= \sum_{i \in \mathscr{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\left(P_i \leq \frac{\alpha W_i \hat{k}}{m}\right)}{\hat{k} \vee 1}\right] \\
&\leq \frac{\alpha \sum_{k=1}^m \frac{1}{k}}{m} \sum_{i \in \mathscr{H}_0} \mathbb{E}\left[W_i\right] \\
&\leq \frac{\alpha \sum_{k=1}^m \frac{1}{k}}{m} \mathbb{E}\left[\sum_{i=1}^m W_i\right] \\
&= \alpha \sum_{k=1}^m \frac{1}{k}.
\end{aligned}$$

Note that the above proof extends to applying the weighted BH procedure with arbitrary reshaping function $\tilde{\beta}$ as in Blanchard and Roquain [2008], Ramdas et al. [2019].

□

S1.6 Counterexample to demonstrate that BY with τ-censored data-driven weights does not control FDR

For our counterexample, we consider the following τ-censored way of assigning data-driven weights: assign weight $W_i = 0$ to all hypotheses with p-value greater than τ and distribute the remaining weight equally across all hypotheses with p-value $\leq \tau$ in any given fold. This weighting procedure satisfies τ-censoring (Specification 2) as it only uses whether a p-value is below or above τ; however it does not satisfy honesty (Specification 1). Finally, we apply the weighted Benjamini-Yekutieli procedure with p-values P_i and weights W_i.

Proof. The result for this counterexample depends on m, τ, α. We make the following simplifying assumptions on these: first, to avoid issues with rounding, we assume that $\tau \in \mathbb{Q}$ and m is such that $m \cdot \tau \in \mathbb{N}$. Furthermore, we assume that $\alpha \leq \tau$.[10]

Below, we will construct a joint distribution on $((P_i, X_i))_{i \in [m]}$ such that Assumption 2 holds with one fold, i.e., $K = 1$ and $I_1 = [m]$. We discuss the case of two independent folds at the end of the proof.

[10] FDR control is also violated when $\alpha > \tau$: just replace τ by $\max\{\alpha, \tau\}$ in the following arguments.

First, we draw $X_i \overset{\text{iid}}{\sim} U[0,1]$ and independent of the p-values $(P_i)_{i\in[m]}$. The joint distribution of the p-values is constructed (details below) so that exactly $m\tau$ p-values are $\leq \tau$. This means that the $m\tau$ hypotheses with p-value $\leq \tau$ are assigned weights $m/(m\tau) = 1/\tau$ and so letting $\alpha_{BY} = \alpha\big/\sum_{j=1}^{m} \frac{1}{j}$, then weighted BY procedure will reject at least k hypotheses if:

$$
\begin{aligned}
& P_j \leq \frac{\alpha_{BY} \cdot k \cdot W_j}{m} \text{ for at least } k \text{ indices } \subset \{1, \ldots, m\} \\
\iff & P_j \leq \frac{\alpha_{BY} \cdot k \cdot \frac{1}{\tau}}{m} \wedge \tau \text{ for at least } k \text{ indices } \subset \{1, \ldots, m\}
\end{aligned}
$$

Next we define $q_k = \frac{\alpha_{BY} \cdot k}{m\tau}, k \geq 1$, $q_0 = 0$. Then the weighted BY will make at least k rejections (for $k \leq m\tau$) if:

$$
P_j \leq q_k \text{ for at least } k \text{ indices } \subset \{1, \ldots, m\} \tag{17}
$$

It remains to provide a distribution on p-values $(P_1, \ldots, P_m)$ such that Assumption 2 is satisfied and such that (17) with $k \geq 1$ occurs frequently enough so that FDR control is violated. To this end, we generate the m p-values hierarchically[11] as follows:

1. We draw a set of indices $\mathcal{T} \subset \{1, \ldots, m\}$ of cardinality $m\tau$ at random from $\{1, \ldots, m\}$.
2. For $i \notin \mathcal{T}$, we draw $P_i \sim U[\tau, 1]$.
3. For $i \in \mathcal{T}$, we instead proceed as follows:
   (a) We draw $\tilde{\kappa} \in \{0, \ldots, m\tau\}$ from the following distribution:

$$
\mathbb{P}[\tilde{\kappa} = k] = m\frac{q_k - q_{k-1}}{k} = \frac{\alpha_{BY}}{\tau k},\ k = 1, \ldots, m\tau,\ \ \mathbb{P}[\tilde{\kappa} = 0] = 1 - \frac{\alpha_{BY}}{\tau}\sum_{j=1}^{m\tau}\frac{1}{j}
$$

   (b) We draw a set of indices $\mathcal{S} \subset \mathcal{T}$ of cardinality $\tilde{\kappa}$ at random from $\mathcal{T}$.
   (c) For $i \in \mathcal{S}$, we draw $P_i \sim U[q_{\tilde{\kappa}-1}, q_{\tilde{\kappa}}]$.
   (d) For $i \in \mathcal{T} \setminus \mathcal{S}$, we draw $P_i \sim U[\alpha_{BY}, \tau]$.

Let us note that when $\tilde{\kappa} \geq 1$, then $|\mathcal{S}| = \tilde{\kappa}$ and so there will be $\tilde{\kappa}$ p-values in the interval $U[q_{\tilde{\kappa}-1}, q_{\tilde{\kappa}}]$, and so by (17) these p-values will be rejected leading to a FDP equal to 1. The only situation in which we will make no rejections is on the event that $\tilde{\kappa} = 0$ and so FDP $\geq \mathbf{1}(\tilde{\kappa} \geq 1)$. Thus:

$$
\text{FDR} \geq 1 - \mathbb{P}[\tilde{\kappa} = 0] = \frac{\alpha_{BY}}{\tau}\sum_{j=1}^{m\tau}\frac{1}{j} = \frac{\alpha}{\tau} \cdot \frac{\sum_{j=1}^{m\tau}\frac{1}{j}}{\sum_{j=1}^{m}\frac{1}{j}} \geq \frac{\alpha}{\tau} \cdot \frac{\log(m\tau + 1)}{\log(m) + 1} \tag{18}
$$

Note that for large enough m this approaches α/τ and so indeed, for $\tau < 1$, FDR is not controlled.

There remains one step to conclude the proof: we need to check that the p-values generated above indeed are all (marginally) uniform. Fix an arbitrary $i \in \{1, \ldots, m\}$. Note that conditionally on $K, \mathcal{S}, \mathcal{T}$, the distribution of the p-value P_i is as follows:

$$
P_i \sim \begin{cases}
U[0, q_1] & \text{if } i \in \mathcal{S},\ \tilde{\kappa} = 1 \\
U[q_1, q_2] & \text{if } i \in \mathcal{S},\ \tilde{\kappa} = 2 \\
\vdots & \\
U[q_{m\tau-1}, \alpha_{BY}] & \text{if } i \in \mathcal{S},\ \tilde{\kappa} = m\tau \\
U[\alpha_{BY}, \tau] & \text{if } i \in \mathcal{T} \setminus \mathcal{S} \\
U[\tau, 1] & \text{if } i \notin \mathcal{T}
\end{cases}
$$

[11] Our construction is a modification of an unpublished proof of the worst-case behavior of BH under dependence by Emmanuel Candès and Rina Foygel Barber. This proof has appeared in the STATS300C lecture notes of Emmanuel Candès, available at `https://statweb.stanford.edu/~candes/teaching/stats300c/`.

Let us compute the probabilities of the events above:

$$\mathbb{P}[i \in \mathcal{S},\ \tilde{\kappa} = k] = \mathbb{P}[i \in \mathcal{S} \,\big|\, \tilde{\kappa} = k]\, \mathbb{P}[\tilde{\kappa} = k] = \frac{k}{m} \cdot m \frac{q_k - q_{k-1}}{k} = q_k - q_{k-1}$$

$$\mathbb{P}[i \in \mathcal{T} \setminus \mathcal{S}] = \mathbb{P}[i \in \mathcal{T}] - \mathbb{P}[i \in \mathcal{S}] = \tau - \sum_{j=1}^{m\tau}(q_j - q_{j-1}) = \tau - q_{m\tau} = \tau - \alpha_{BY}$$

$$\mathbb{P}[i \notin \mathcal{T}] = 1 - \tau$$

This means that:

$$P_i \sim \begin{cases} U[0, q_1] & \text{with probability } q_1 \\ U[q_1, q_2] & \text{with probability } q_2 - q_1 \\ \vdots & \\ U[q_{m\tau-1}, \alpha_{BY}] & \text{with probability } \alpha_{BY} - q_{m\tau-1} \\ U[\alpha_{BY}, \tau] & \text{with probability } \tau - \alpha_{BY} \\ U[\tau, 1] & \text{with probability } 1 - \tau \end{cases}$$

This is precisely the uniform distribution, i.e., $P_i \sim U[0, 1]$.

Let us finally conclude by discussing how to extend this construction to the case of two independent folds. Let $m = 2m'$ for $m' \in \mathbb{N}$ and assume that $m'\tau \in \mathbb{N}$. Let us take the two folds to be $I_1 = [m']$ and $I_2 = [m] \setminus [m']$. We may apply the construction above independently to each fold. Now let A_ℓ, $\ell \in \{1, 2\}$ be the event of rejecting at least one hypothesis in fold ℓ. Then repeating the arguments leading up to (18), we find that $\mathbb{P}[A_\ell] \geq \alpha'/2$, where $\alpha' := \alpha/\tau \cdot \log(m'\tau + 1)/(\log(2m') + 1)$. Since $\text{FDR} \geq \mathbb{P}[A_1 \cup A_2]$ and the events A_1 and A_2 are independent, we find that $\text{FDR} \geq \alpha'/(\alpha' + 1)$. This is strictly larger than α, for example when $\tau < 1$, m' is large and α is small.

□

Supplement S2: Proofs for IHW-BH asymptotics

For our asymptotics, we make the following regularity assumption:

Assumption 3 (Regularity of conditional two-groups model)**.** The conditional two-groups model (6) satisfies:

(a) $F_{\text{alt}}(t \mid X_i = x)$ is $L(x)$-Lipschitz continuous in t for all $x \in \mathcal{X}$, i.e.,

$$|F_{\text{alt}}(t \mid X_i = x) - F_{\text{alt}}(t' \mid X_i = x)| \leq L(x)\, |t - t'| \text{ for all } t, t' \in [0, 1], x \in \mathcal{X}$$

$L(\cdot)$ satisfies $\int L^2(x) d\mathbb{P}^X(x) < \infty$ and furthermore $F_{\text{alt}}(0 \mid X_i = x) = 0$ for all x.

(b) $F_{\text{alt}}(t \mid X_i = x)$ is strictly concave in t for all x.

(c) There exists $t' \in (0, 1]$ such that $\frac{t'}{F(t' \mid X_i = x)} \leq \alpha'$ for an $\alpha' < \alpha$ and for all $x \in \mathcal{X}$.

Part (a) is a mild technical assumption restricting the smoothness of $F_{\text{alt}}(\cdot \mid X_i = x)$; it allows for the smoothness to vary as $x \in \mathcal{X}$ varies. Part (b) is a common assumption in multiple testing; see also the discussion and references in Section 4. The assumption (in the setting without covariates) appears for example in Lemma 1 and Theorem 2 of Genovese et al. [2006]. Part (c) is also an assumption made for FDR asymptotics without covariates (e.g., it appears in Theorem 4 of Storey, Taylor, and Siegmund [2004]). It is, however, less innocuous than Parts (a,b); for example it excludes the global null case in which $\pi_0(x) = 1$ for all x.

Some remarks on notation: In this section we use a different typeface for the weight function, i.e., we write $\mathscr{W} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and $\hat{\mathscr{W}}^{(I)}$ for the weight function learned based on data $(P_i, X_i), i \in I$. This ensures that the notation is unambiguous and not conflicting with the notation used in Supplement S1 for finite-sample results. We also use the notation $a_m = o(1)$ for a deterministic sequence a_m satisfying $a_m \to 0,$ as $m \to \infty$ and $Z_m = o_{\mathbb{P}}(1)$ for a sequence of random variables Z_m that converge to 0 in probability as $m \to \infty$.

S2.1 Proof of Proposition 1(a)

Proof. We first make a few assumptions on the data-generating mechanism (while making sure that Assumption 3 still holds): We first assume that $\mathcal{X}, \mathbb{P}^X$ are such that $X_1, \ldots, X_n$ are all unequal with probability 1; this is true for example when $\mathbb{P}^X$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^p$. Next we assume that for $\pi_1(x) = 1 - \pi_0(x)$ it holds that $\mathbb{E}\left[\pi_1(X_i)\right] < \delta$ for some $\delta > 0$; i.e., there are not too many alternative hypotheses. Finally we assume that we run weighted BH at $\alpha \in (0, 1/2)$.

Our application of naive weighted BH is as follows: We let $k_m = \lfloor \alpha m/2 \rfloor$ and $\mathcal{J}_m$ the index set of k_m hypotheses with smallest p-values. We will assign weight m/k_m to these and all other hypotheses will receive weight 0. This is equivalent to applying BH directly to the k_m smallest p-values (while ignoring their selection).

Formally, in terms of Specification 3, the weighting function takes the form:

$$\hat{\mathscr{W}}^{([m])}(x) = \mathbf{1}\left(x \notin \{X_j, j \in [m]\}\right) \; + \; \frac{m}{k_m} \mathbf{1}\left(x \in \{X_j, j \in \mathcal{J}_m\}\right);$$

This satisfies the conditions of Specification 3: $\int \hat{\mathscr{W}}^{([m])}(x) d\mathbb{P}^X(x) = 1$ almost surely for all m and second, $\sup_{x \in \mathcal{X}} \hat{\mathscr{W}}^{([m])}(x) = m/k_m = m/\lfloor \alpha m/2 \rfloor \leq 4/\alpha$ as soon as $\alpha m \geq 2$, which is stronger than the requirement on the growth of $\int \hat{\mathscr{W}}^{([m])}(x)^2 d\mathbb{P}^X(x)$.

Writing $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ for the order statistics of $P_1, \ldots, P_m$, consider the events $A_m = \left\{P_{(k_m)} \leq \alpha\right\}$ and $B_m = \left\{\sum_{i=1}^m H_i \leq 1.1 \cdot \delta \cdot m\right\}$.

On the event A_m, weighted BH will reject all hypotheses in $\mathcal{J}_m$, since by definition of A_m it holds that $P_{(k_m)} \leq \alpha = (k_m \cdot \alpha/m) \cdot (m/k_m) = (k_m \cdot \alpha/m) \cdot W_{(k_m)}$. On the other hand, on the event B_m, there will be at least $k_m - 1.1 \cdot \delta \cdot m$ false rejections (i.e., all rejections minus an upper bound on the number of alternative hypotheses). Thus on $A_m \cap B_m$ and for large enough m (we slightly enlarge $2.2 = 2 \cdot 1.1$ to 2.3 to account for rounding in the definition of k_m):

$$\text{FDP}_m \geq \frac{k_m - 1.1 \cdot \delta \cdot m}{k_m} \geq 1 - \frac{2.3\delta}{\alpha}$$

We will next argue that $\mathbb{P}\left[A_m\right], \mathbb{P}\left[B_m\right] \to 1$ as $m \to \infty$ and thus:

$$\liminf_{m \to \infty} \text{FDR}_m \geq 1 - \frac{2.3\delta}{\alpha}$$

The latter will in general be $> \alpha$ for small enough δ, so that naive weighted BH does not control FDR.

Let us prove the claims for A_m and B_m. For B_m, the result follows by noting that $\sum_{i=1}^m H_i \sim \text{Binomial}(m, \mathbb{E}\pi_1(X_i))$, as well as an application of Chernoff's bound. For A_m, we note that by Assumption 3(b), it follows that $P_{(k_m)}$ is stochastically smaller than $\tilde{P}_{(k_m)}$, defined as the k_m-th smallest order statistic of a sample of m i.i.d. uniform random variables $\tilde{P}_1, \ldots, \tilde{P}_m$. Note that $\tilde{P}_{(k_m)}$ is distributed as $\text{Beta}(k_m, m + 1 - k_m)$ which has expectation $k_m/(m+1) \leq \frac{\alpha}{2}$. Hence:

$$\mathbb{P}\left[A_m\right] \geq \mathbb{P}\left[\text{Beta}(k_m, m + 1 - k_m) \leq \alpha\right] \to 1 \text{ as } m \to \infty$$

The last convergence follows from concentration of a Beta random variable (say, by an application of Chebyshev's inequality.)

□

S2.2 Proof of Proposition 1(b)

Proof. We first give a sketch of the proof:

1. **Analysis for a single fold and a deterministic weighting function:** This serves as a warm-up. The analysis is very similar to asymptotics e.g., in Storey et al. [2004], adapted to the setting with covariates and a weighting function.
2. **Analysis for a single fold with data-driven weighting function learned out-of-fold**: Here we refine the analysis from Step 1 to account for the data-driven nature of the weighting function. The fundamental nature of the arguments however is the same as in Step 1.
3. **Aggregating results across folds:** We give an equivalent formulation of the IHW-BH rejection rule in terms of empirical processes. Then, by combining results shown in Step 2, we demonstrate FDR control.

Single fold, deterministic weighting function: We first study a single fold, say $I = I_\ell$ (that grows with m), and a deterministic weighting function with the following properties:

$$\mathscr{W} : \mathcal{X} \to \mathbb{R}_{\geq 0},\ \int \mathscr{W}(x) d\mathbb{P}^X(x) = 1,\ \int \mathscr{W}(x)^2 d\mathbb{P}^X(x) \leq \Gamma < \infty \tag{19}$$

We introduce notation for processes indexed by a threshold $t \in [0, 1]$, the weighting function $\mathscr{W}$ and the set $I \subset [m]$ indexing the hypotheses in the single fold under study.

$$\begin{aligned}
R(t, \mathscr{W}; I) &= \sum_{i \in I} \mathbf{1}(P_i \leq t\mathscr{W}(X_i)) \\
V(t, \mathscr{W}; I) &= \sum_{i \in I} \mathbf{1}\left(H_i = 0\right) \mathbf{1}(P_i \leq t\mathscr{W}(X_i)) \\
\widehat{V}^{\mathrm{BH}}(t, \mathscr{W}; I) &= \sum_{i \in I} \min\left\{t\mathscr{W}(X_i), 1\right\} \\
F(t, \mathscr{W}) &= \mathbb{P}\left[P_i \leq t\mathscr{W}(X_i)\right] \\
F_0(t, \mathscr{W}) &= \mathbb{P}\left[P_i \leq t\mathscr{W}(X_i); H_i = 0\right] \\
F_0^{\mathrm{BH}}(t, \mathscr{W}) &= \mathbb{E}\left[\min\left\{t\mathscr{W}(X_i), 1\right\}\right]
\end{aligned}$$

The goal will be to relate the empirical processes to their population counterparts through uniform (in t) laws of large numbers. We require one more definition to account for normalization of weights so that $\sum_{i \in I} W_i = |I|$

$$\hat{c}_{I,\mathscr{W}} = |I| \,/ \sum_{i \in I} \mathscr{W}(X_i) \tag{20}$$

Next pick a deterministic sequence $0 < \varepsilon_m = o(1)$ as $m \to \infty$ such that $\mathbb{P}\left[|\hat{c}_{I,\mathscr{W}} - 1| > \varepsilon_m\right] = o(1)$; such a sequence exists by the law of large numbers. Then for a $o_{\mathbb{P}}(1)$ term that is uniform in $t \in [0, 1]$, it holds that:

$$\begin{aligned}
R(\hat{c}_{I,\mathscr{W}} \cdot t, \mathscr{W}; I)\mathbf{1}\left(\hat{c}_{I,\mathscr{W}} < 1 + \varepsilon_m\right) / \,|I| &\overset{(i)}{\leq} R((1 + \varepsilon_m)t, \mathscr{W}; I) / \,|I| \\
&\overset{(ii)}{=} F((1 + \varepsilon_m)t; \mathscr{W}) + o_{\mathbb{P}}(1) \\
&\overset{(iii)}{\leq} F(t; \mathscr{W}) + o_{\mathbb{P}}(1) + \Gamma^{1/2}(1 + \int L(x)^2 d\mathbb{P}^X(x))^{1/2}\varepsilon_m
\end{aligned} \tag{21}$$

(i) follows by monotonicity of $R(t, \mathscr{W}; I)$ in t. (ii) follows from the Glivenko-Cantelli theorem applied to the i.i.d. $P_i/\mathscr{W}(X_i)$[12]. (iii) follows from Assumption 3(a), as follows: first note that $F(t \mid X_i = x)$

[12] We set the above to ∞ if $\mathscr{W}(X_i) = 0$.

must be $\max\{1, L(x)\}$ Lipschitz in t as it is a convex combination of a $L(x)$-Lipschitz function and a 1-Lipschitz function (the identity). Next

$$\begin{aligned}
|F((1+\varepsilon_m)t; \mathscr{W}) - F(t; \mathscr{W})| &= |\mathbb{E}\left[F((1+\varepsilon_m)t \cdot \mathscr{W}(X_i) \mid X_i) - F(t \cdot \mathscr{W}(X_i) \mid X_i)\right]| \\
&\leq \mathbb{E}\left[|F((1+\varepsilon_m)t \cdot \mathscr{W}(X_i) \mid X_i) - F(t \cdot \mathscr{W}(X_i) \mid X_i)|\right] \\
&\leq \mathbb{E}\left[\max\{1, L(X_i)\}\varepsilon_m t \mathscr{W}(X_i)\right] \\
&\leq \varepsilon_m t \mathbb{E}\left[\max\left\{1, L^2(X_i)\right\}\right]^{1/2} \mathbb{E}\left[\mathscr{W}^2(X_i)\right]^{1/2} \\
&\leq \Gamma^{1/2}(1 + \int L(x)^2 d\mathbb{P}^X(x))^{1/2}\varepsilon_m
\end{aligned}$$

Applying the same argument in the reverse direction we also get for the same (uniform in t) $o_{\mathbb{P}}(1)$ term:

$$R(\hat{c}_{I,\mathscr{W}} \cdot t, \mathscr{W}; I)\mathbf{1}\left(\hat{c}_{I,\mathscr{W}} > 1 - \varepsilon_m\right) / |I| \geq F(t, \mathscr{W}) - o_{\mathbb{P}}(1) - \Gamma^{1/2}(1 + \int L(x)^2 d\mathbb{P}^X(x))^{1/2}\varepsilon_m$$

Combining the two results, noting that $R(t, \hat{c}_{\mathscr{W}} \cdot \mathscr{W}; I) = R(\hat{c}_{\mathscr{W}} \cdot t, \mathscr{W}; I)$ and by choice of ε_m we conclude that:

$$\sup_{t\in[0,1]} |R(t, \hat{c}_{I,\mathscr{W}} \cdot \mathscr{W}; I)/|I| - F(t, \mathscr{W})| = o_{\mathbb{P}}(1) \tag{22}$$

We may analogously prove that:

$$\begin{aligned}
&\sup_{t\in[0,1]} |V(t, \hat{c}_{I,\mathscr{W}} \cdot \mathscr{W}; I)/|I| - F_0(t, \mathscr{W})| = o_{\mathbb{P}}(1) \\
&\sup_{t\in[0,1]} \left|\widehat{V}^{\text{BH}}(t, \hat{c}_{I,\mathscr{W}} \cdot \mathscr{W}; I)/|I| - F_0^{\text{BH}}(t, \mathscr{W})\right| = o_{\mathbb{P}}(1)
\end{aligned} \tag{23}$$

It also deterministically holds that $F_0(t; \mathscr{W}) \leq F_0^{\text{BH}}(t; \mathscr{W})$ for all $t, \mathscr{W}$ and so

$$\sup_{t\in[0,1]} \left(F_0(t, \mathscr{W}) - \widehat{V}^{\text{BH}}(t, \hat{c}_{I,\mathscr{W}} \cdot \mathscr{W}, I)/|I|\right) \leq o_{\mathbb{P}}(1) \tag{24}$$

Single fold, data-driven weighting function: Above we worked with a deterministic weighting function $\mathscr{W}$. However, for IHW we use the weighting function learned out-of-fold $\widehat{\mathscr{W}}^{([m]\setminus I)}$. It turns out that the conclusions hold verbatim, i.e.,

$$\begin{aligned}
&\sup_{t\in[0,1]} \left|R(t, \hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I)}; I)/|I| - F(t, \widehat{\mathscr{W}}^{([m]\setminus I)})\right| = o_{\mathbb{P}}(1) \\
&\sup_{t\in[0,1]} \left|V(t, \hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I)}; I)/|I| - F_0(t, \widehat{\mathscr{W}}^{([m]\setminus I)})\right| = o_{\mathbb{P}}(1) \\
&\sup_{t\in[0,1]} \left|\widehat{V}^{\text{BH}}(t, \hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I)}; I)/|I| - F_0^{\text{BH}}(t, \widehat{\mathscr{W}}^{([m]\setminus I)})\right| = o_{\mathbb{P}}(1) \\
&\sup_{t\in[0,1]} \left(F_0(t, \widehat{\mathscr{W}}^{([m]\setminus I)}) - \widehat{V}^{\text{BH}}(t, \hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I)}, I)/|I|\right) \leq o_{\mathbb{P}}(1)
\end{aligned} \tag{25}$$

To adapt the proof for deterministic $\mathscr{W}$ to a proof for data-driven $\widehat{\mathscr{W}}^{([m]\setminus I)}$ (where $\widehat{\mathscr{W}}^{([m]\setminus I)}$ depends on data outside of fold $I = I_\ell$, cf. Specification 3) we make the following observations:

1. We conduct the analysis conditionally on data in the other folds $\mathcal{D}_{[m]\setminus I} = ((P_i, X_i, H_i))_{i\in[m]\setminus I}$. For example, to show the first result in (25) it suffices to show (see arguments below) that for a sequence $\eta_m \to 0$:

$$\mathbb{P}\left[\sup_{t\in[0,1]} \left|R(t, \hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I)}; I)/|I| - F(t, \widehat{\mathscr{W}}^{([m]\setminus I)})\right| > \eta_m \;\middle|\; \mathcal{D}_{[m]\setminus I}\right] = o_{\mathbb{P}}(1) \tag{26}$$

Such a conditional convergence statement also implies unconditional convergence (cf. Lemma 6.1. in Chernozhukov et al. [2017]), i.e.,

$$\mathbb{P}\left[\sup_{t\in[0,1]}\left|R(t, \hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I)}; I)/\left|I\right| - F(t, \widehat{\mathscr{W}}^{([m]\setminus I)})\right| > \eta_m\right] = o(1)$$

The first result in (25) then follows.

2. It can be assumed without loss of generality that $\int \widehat{\mathscr{W}}^{([m]\setminus I)}(x)d\mathbb{P}^X(x) = 1$ for all m; otherwise we may redefine the weight function as $\widehat{\mathscr{W}}^{([m]\setminus I)}/\int \widehat{\mathscr{W}}^{([m]\setminus I)}(x)d\mathbb{P}^X(x)$. This is only a formal modification; the IHW-BH procedure applied remains the same, as the weights will subsequently be rescaled to sum to $|I|$ in fold I (this is captured here by the multiplication with $\hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}}$).

3. To establish (26), the argument used for a deterministic weighting function applies as long as we pay attention to controlling the two probabilistically negligible terms. In particular, we need to check that for (deterministic sequences) $\eta'_m, \eta''_m = o(1)$ that

$$\mathbb{P}\left[\left|\hat{c}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} - 1\right| > \eta'_m \;\middle|\; \mathcal{D}_{[m]\setminus I}\right] = o_{\mathbb{P}}(1)$$

and

$$\mathbb{P}\left[\sup_{t\in[0,1]}\left|R(t, \widehat{\mathscr{W}}^{([m]\setminus I)}; I)/\left|I\right| - F(t; \widehat{\mathscr{W}}^{([m]\setminus I)})\right| > \eta''_m \;\middle|\; \mathcal{D}_{[m]\setminus I}\right] = o_{\mathbb{P}}(1)$$

In the deterministic case, the corresponding results were a consequence of the law of large numbers, respectively the Glivenko-Cantelli theorem. In the conditional case we may establish these results directly. For the first one we note that by Chebyshev's inequality (conditionally on $\mathcal{D}_{[m]\setminus I}$) it holds almost surely for any $\delta > 0$ that

$$\mathbb{P}\left[\left|\hat{c}^{-1}_{I,\widehat{\mathscr{W}}^{([m]\setminus I)}} - 1\right| > \delta \;\middle|\; \mathcal{D}_{[m]\setminus I}\right] \leq \frac{\int \widehat{\mathscr{W}}^{([m]\setminus I)}(x)^2 d\mathbb{P}^X(x)}{\delta^2\left|I\right|} \leq \frac{\Gamma}{\delta^2\left|I\right|}$$

The conclusion follows. For the second result, we may replace the Glivenko-Cantelli theorem by an application of the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality conditionally on $\mathcal{D}_{[m]\setminus I}$.

Aggregating results across folds: Let us introduce some additional notation.

$$\widehat{\text{FDP}}^{\text{IHW}}(t) = \frac{\sum_{\ell=1}^{K}\widehat{V}^{\text{BH}}(t, \hat{c}_{I_\ell,\widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)/m}{\max\left\{1, \sum_{\ell=1}^{K} R(t, \hat{c}_{I_\ell,\widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)\right\}/m} \tag{27}$$

$$\hat{t}^{\text{IHW}} = \sup\left\{t \in [0,1] \mid \widehat{\text{FDP}}^{\text{IHW}}(t) \leq \alpha\right\} \tag{28}$$

This implies (the denominator of $\widehat{\text{FDP}}^{\text{IHW}}(t)$ is right-continuous and and non-decreasing in t with jumps, while the numerator is continuous) that

$$\widehat{\text{FDP}}^{\text{IHW}}(\hat{t}^{\text{IHW}}) \leq \alpha \text{ almost surely} \tag{29}$$

The quantities allow us to express IHW-BH from an empirical process viewpoint (cf. Storey et al. [2004])

$$\text{Reject } i \in I_\ell \iff P_i \leq \hat{t}^{\text{IHW}} \cdot \hat{c}_{I_\ell,\widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}(X_i) \tag{30}$$

Henceforth we make the additional assumption that α' in Assumption 3(c) further satisfies $\alpha' < \alpha/2$; this simplifies the step below but the proof goes through also for $\alpha' < \alpha$. With this simplification, we next argue that, for $t'' = t'/(4\Gamma)$ with t' defined in Assumption 3(c)

$$\mathbb{P}\left[\hat{t}^{\text{IHW}} < t''\right] = o(1) \text{ as } m \to \infty \tag{31}$$

Note that Assumption 3 implies that $F(t|X_i = x) \geq t$ for all $t \in (0, 1)$. Fixing a weighting function $\mathscr{W}$ as in (19)

$$\begin{aligned}
F(t, \mathscr{W}) &= \int F(t\mathscr{W}(x) \mid X_i = x) d\mathbb{P}^X(x) \\
&\geq \int \mathbf{1}\left(\mathscr{W}(x) > 1/2\right) F(t\mathscr{W}(x) \mid X_i = x) d\mathbb{P}^X(x) \\
&\geq \int \mathbf{1}\left(\mathscr{W}(x) > 1/2\right) F(t/2 \mid X_i = x) d\mathbb{P}^X(x) \\
&\geq t/2 \int \mathbf{1}\left(\mathscr{W}(x) > 1/2\right) d\mathbb{P}^X(x) \\
&\geq t/2 \cdot (1/4) \frac{(\int \mathscr{W}(x) d\mathbb{P}^X(x))^2}{\int \mathscr{W}(x)^2 d\mathbb{P}^X(x)} \\
&\geq t/(8 \cdot \Gamma)
\end{aligned}$$

In the penultimate step we used the Paley–Zygmund inequality. This lower bound holds uniformly over weighting functions satisfying (19). In the current setting this implies that for a constant $c > 0$

$$\sum_{\ell=1}^{K} |I_\ell| / m \cdot F(t''', \hat{\mathscr{W}}^{([m] \setminus I_\ell)}) \; > \; c \text{ for all } t''' \geq t'' \text{ almost surely for all } m$$

In conjunction with (25) this yields (we need the preceding claim to make sure the denominators in the expression below do not vanish)

$$\sup_{t \in [t'', 1]} \left| \widehat{\mathrm{FDP}}^{\mathrm{IHW}}(t) - \frac{\sum_{\ell=1}^{K} |I_\ell| / m \cdot F_0^{\mathrm{BH}}(t, \hat{\mathscr{W}}^{([m] \setminus I_\ell)})}{\sum_{\ell=1}^{K} |I_\ell| / m \cdot F(t, \hat{\mathscr{W}}^{([m] \setminus I_\ell)})} \right| = o_{\mathbb{P}}(1) \tag{32}$$

Fixing again a $\mathscr{W}$ as in (19)), we find using Cauchy-Schwartz and Markov's inequality, that

$$\begin{aligned}
\int \mathscr{W}(x) \mathbf{1}\left(\mathscr{W}(x) \leq 4\Gamma\right) d\mathbb{P}^X(x) &= 1 - \int \mathscr{W}(x) \mathbf{1}\left(\mathscr{W}(x) > 4\Gamma\right) d\mathbb{P}^X(x) \\
&\geq 1 - \Big(\int \mathscr{W}(x)^2 d\mathbb{P}^X(x)\Big)^{1/2} \Big(\int \mathbf{1}\left(\mathscr{W}(x) > 4\Gamma\right) d\mathbb{P}^X(x)\Big)^{1/2} \\
&\geq 1 - \Gamma^{1/2} 1/(4\Gamma)^{1/2} \\
&= \frac{1}{2}
\end{aligned} \tag{33}$$

Next, we find that:

$$\begin{aligned}
F_0^{\mathrm{BH}}(t'', \mathscr{W}) &\leq \int t'' \mathscr{W}(x) d\mathbb{P}^X(x) \\
&\overset{(i)}{\leq} 2 \int t'' \mathscr{W}(x) \mathbf{1}\left(\mathscr{W}(x) \leq 4\Gamma\right) d\mathbb{P}^X(x) \\
&= 2 \int (t'/4\Gamma) \mathscr{W}(x) \mathbf{1}\left(\mathscr{W}(x) \leq 4\Gamma\right) d\mathbb{P}^X(x) \\
&\overset{(ii)}{\leq} 2\alpha' \int \mathscr{W}(x)/(4\Gamma) \cdot F(t' \mid X_i = x) \mathbf{1}\left(\mathscr{W}(x) \leq 4\Gamma\right) d\mathbb{P}^X(x) \\
&\overset{(iii)}{\leq} 2\alpha' \int F(\mathscr{W}(x)/(4\Gamma) \cdot t' \mid X_i = x) \mathbf{1}\left(\mathscr{W}(x) \leq 4\Gamma\right) d\mathbb{P}^X(x) \\
&\leq 2\alpha' \int F(t'' \cdot \mathscr{W}(x) \mid X_i = x) d\mathbb{P}^X(x) \\
&= 2\alpha' F(t'', \mathscr{W})
\end{aligned} \tag{34}$$

Step (i) follows from (33), step (ii) follows by the definition of α' in Assumption 3(c) and (iii) from concavity of $F(\cdot \mid x)$ and the fact that $\mathscr{W}(x)/(4\Gamma) \leq 1$ on the stated event. Next, rearranging (34)

$$\frac{F_0^{\text{BH}}(t'', \mathscr{W})}{F(t'', \mathscr{W})} \leq 2\alpha'$$

Since this holds for an arbitrary weighting function (19), we also get that

$$\frac{\sum_{\ell=1}^K |I_\ell| / m \cdot F_0^{\text{BH}}(t'', \widehat{\mathscr{W}}^{([m]\setminus I_\ell)})}{\sum_{\ell=1}^K |I_\ell| / m \cdot F(t'', \widehat{\mathscr{W}}^{([m]\setminus I_\ell)})} \leq 2\alpha' < \alpha$$

And so along with (32) we see that (31) holds:

$$\mathbb{P}\left[\hat{t}^{\text{IHW}} \geq t''\right] \geq \mathbb{P}\left[\widehat{\text{FDP}}^{\text{IHW}}(t'') \leq \alpha\right] = 1 - o(1) \text{ as } m \to \infty$$

We are almost ready to prove FDR control. By (30), we see that the rejections of IHW-BH are precisely equal to:

$$R^{\text{IHW}} = \sum_{\ell=1}^K R(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)$$

The false rejections are equal to:

$$V^{\text{IHW}} = \sum_{\ell=1}^K V(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)$$

So:

$$\begin{aligned}
\text{FDP}^{\text{IHW}} \cdot \mathbf{1}(\hat{t}^{\text{IHW}} \geq t'') &= \frac{V^{\text{IHW}}}{\max\{1, R^{\text{IHW}}\}} \cdot \mathbf{1}(\hat{t}^{\text{IHW}} \geq t') \\
&= \frac{\sum_{\ell=1}^K V(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)}{\max\left\{1, \sum_{\ell=1}^K R(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)\right\}} \cdot \mathbf{1}(\hat{t}^{\text{IHW}} \geq t'') \\
&= \frac{\sum_{\ell=1}^K |I_\ell| \cdot F_0(\hat{t}^{\text{IHW}}, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)})}{\max\left\{1, \sum_{\ell=1}^K R(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)\right\}} \cdot \mathbf{1}(\hat{t}^{\text{IHW}} \geq t'') \; + \; o_{\mathbb{P}}(1) \\
&\leq \frac{\sum_{\ell=1}^K |I_\ell| \cdot F_0^{\text{BH}}(\hat{t}^{\text{IHW}}, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)})}{\max\left\{1, \sum_{\ell=1}^K R(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)\right\}} \cdot \mathbf{1}(\hat{t}^{\text{IHW}} \geq t'') \; + \; o_{\mathbb{P}}(1) \\
&\leq \frac{\sum_{\ell=1}^K \widehat{V}^{\text{BH}}(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)}{\max\left\{1, \sum_{\ell=1}^K R(\hat{t}^{\text{IHW}}, \hat{c}_{I_\ell, \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}} \cdot \widehat{\mathscr{W}}^{([m]\setminus I_\ell)}; I_\ell)\right\}} \cdot \mathbf{1}(\hat{t}^{\text{IHW}} \geq t'') \; + \; o_{\mathbb{P}}(1) \\
&= \widehat{\text{FDP}}^{\text{IHW}}(\hat{t}^{\text{IHW}}) \cdot \mathbf{1}(\hat{t}^{\text{IHW}} \geq t'') \; + \; o_{\mathbb{P}}(1) \\
&\leq \alpha \cdot 1 \; + \; o_{\mathbb{P}}(1)
\end{aligned}$$

In the last step we used (29). We carry along the constraint $\mathbf{1}(\hat{t}^{\text{IHW}} \geq t'')$ to emphasize that the denominator divided by m will remain bounded from below with probability converging to 1. We conclude with the dominated convergence theorem ($\text{FDR}^{\text{IHW}} = \mathbb{E}\left[\text{FDP}^{\text{IHW}}\right]$ and $\text{FDP}^{\text{IHW}} \in [0, 1]$) that

$$\limsup_{m \to \infty} \text{FDR}^{\text{IHW}} \leq \alpha$$

□

S2.3 Proof of Proposition 1(c)

Proof. Let us introduce the asymptotic threshold of both procedures;

$$t^*(\mathscr{W}^*) = \sup\left\{t \in [0,1] \;\middle|\; \frac{F_0^{\text{BH}}(t, \mathscr{W}^*)}{F(t, \mathscr{W}^*)} \leq \alpha\right\} \tag{35}$$

Assumption 3 ensures the existences of unique $t^*(\mathscr{W}^*) \in (t'', 1)$ for which in fact equality is attained, i.e., $F_0^{\text{BH}}(t^*(\mathscr{W}^*), \mathscr{W}^*)/F(t^*(\mathscr{W}^*), \mathscr{W}^*) = \alpha$.

We use (11) as our definition for power; the results for other notions such as $1 - \text{FNR}$ where FNR is the false nondiscovery rate are analogous. Our claim is that the power of both naive weighted BH and IHW-BH asymptotically is equal to

$$\lim_{m\to\infty} \text{Power}_m^{\text{IHW-BH}} = \lim_{m\to\infty} \text{Power}_m^{\text{Naive}} = \frac{F(t^*(\mathscr{W}^*), \mathscr{W}^*) - F_0(t^*(\mathscr{W}^*), \mathscr{W}^*)}{\int (1-\pi_0(x)) d\mathbb{P}^X(x)} \tag{36}$$

We start by analyzing Naive weighted BH. First, we may use continuity and Glivenko Cantelli arguments leading to (22) and (23), along with the assumption on uniform convergence (in probability) of $\widehat{\mathscr{W}}^{([m])}$, to show that

$$\begin{aligned}
&\sup_{t\in[0,1]} \left| R(t, \hat{c}_{[m], \widehat{\mathscr{W}}^{([m])}} \cdot \widehat{\mathscr{W}}^{([m])}; [m])/m - F(t, \mathscr{W}^*) \right| = o_{\mathbb{P}}(1) \\
&\sup_{t\in[0,1]} \left| V(t, \hat{c}_{[m], \widehat{\mathscr{W}}^{([m])}} \cdot \widehat{\mathscr{W}}^{([m])}; [m])/m - F_0(t, \mathscr{W}^*) \right| = o_{\mathbb{P}}(1) \\
&\sup_{t\in[0,1]} \left| \widehat{V}^{\text{BH}}(t, \hat{c}_{[m], \widehat{\mathscr{W}}^{([m])}} \cdot \widehat{\mathscr{W}}^{([m])}; [m])/m - F_0^{\text{BH}}(t, \mathscr{W}^*) \right| = o_{\mathbb{P}}(1)
\end{aligned} \tag{37}$$

The empirical process interpretation of naive weighted BH (analogous to (30) for IHW-BH) is as follows. Define

$$\widehat{\text{FDP}}^{\text{Naive}}(t) = \frac{V(t, \hat{c}_{[m], \widehat{\mathscr{W}}^{([m])}} \cdot \widehat{\mathscr{W}}^{([m])}; [m])}{\max\left\{1, R(t, \hat{c}_{[m], \widehat{\mathscr{W}}^{([m])}} \cdot \widehat{\mathscr{W}}^{([m])}; [m])\right\}} \tag{38}$$

$$\hat{t}^{\text{Naive}} = \sup\left\{t \in [0,1] \;\middle|\; \widehat{\text{FDP}}^{\text{Naive}}(t) \leq \alpha\right\} \tag{39}$$

Then the naive weighted BH procedure rejection rule takes the following form;

$$\text{Reject } i \in [m] \iff P_i \leq \hat{t}^{\text{Naive}} \cdot \hat{c}_{[m], \widehat{\mathscr{W}}^{([m])}} \cdot \widehat{\mathscr{W}}^{([m])}(X_i) \tag{40}$$

Our next step is to show that $\hat{t}^{\text{Naive}} = t^*(\mathscr{W}^*) + o_{\mathbb{P}}(1)$. Fix any $\delta \in (0, t^*(\mathscr{W}^*))$, then using Assumption 3 and (37) we deduce that:

$$\mathbb{P}\left[\left|\hat{t}^{\text{Naive}} - t^*(\mathscr{W}^*)\right| \leq \delta\right] \geq \mathbb{P}\left[\widehat{\text{FDP}}^{\text{Naive}}(t^*(\mathscr{W}^*) - \delta) < \alpha,\; \inf_{\delta' > \delta}\left\{\widehat{\text{FDP}}^{\text{Naive}}(t^*(\mathscr{W}^*) + \delta')\right\} > \alpha\right] = 1 - o(1)$$

Then, another application of (37) and continuity properties of $F(\cdot \mid X_i = x)$ demonstrates that:

$$R(\hat{t}^{\text{Naive}}, \widehat{\mathscr{W}}^{([m])}; [m])/m = F(t^*(\mathscr{W}^*), \mathscr{W}^*) + o_{\mathbb{P}}(1),\;\; V(\hat{t}^{\text{Naive}}, \widehat{\mathscr{W}}^{([m])}; [m])/m = F_0(t^*(\mathscr{W}^*), \mathscr{W}^*) + o_{\mathbb{P}}(1)$$

By the law of large numbers: $\sum_{i=1}^m H_i/m = \int (1-\pi_0(x)) d\mathbb{P}^X(x) + o_{\mathbb{P}}(1)$. By definition,

$$\text{Power}_m^{\text{Naive}} = \mathbb{E}\left[\frac{R(\hat{t}^{\text{Naive}}, \widehat{\mathscr{W}}^{([m])}; [m])/m - V(\hat{t}^{\text{Naive}}, \widehat{\mathscr{W}}^{([m])}; [m])/m}{\sum_{i=1}^m H_i/m}\right]$$

and so by dominated convergence (note the term within the expectation above is in $[0,1]$):

$$\text{Power}_m^{\text{Naive}} = \left(F(t^*(\mathscr{W}^*), \mathscr{W}^*) - F_0(t^*(\mathscr{W}^*), \mathscr{W}^*)\right) \Big/ \int (1-\pi_0(x)) d\mathbb{P}^X(x) + o(1)$$

The same argument also applies for IHW-BH, leveraging results proved already in part (b) of the proposition. In particular it follows as for naive weighted BH that also $\hat{t}^{\text{IHW}} = t^*(\mathscr{W}^*) + o_{\mathbb{P}}(1)$ by using (25) and Lipschitz continuity of $F(\cdot \mid x)$. We also note in passing that under the assumptions of part (c), we could have omitted the conditional analysis required in part (b) to prove (25). Instead, a more direct argument (along the lines of (21)) could be given by noting that the i.i.d. structure and convergence of the weighting mechanism imply that

$$\max_{\ell=1}^{K} \left\| \hat{\mathscr{W}}^{([m]\setminus I_\ell)}(\cdot) - W^*(\cdot) \right\|_\infty = o_{\mathbb{P}}(1)$$

Cross-weighting derives its flexibility from guarantees established in part (b), that hold even if the above convergence property of the learned weight function does not hold. □

S2.4 Proof of Corollary 2

Let $\nu(x) = \mathbb{P}[X_i = x]$ for $x \in [G]$. We assume without loss of generality that $\nu(x) > 0$ for all $x \in [G]$; otherwise it suffices to restrict the covariate space to $[G] \setminus \{x\}$.

In the setting with a categorical covariate, (7) is automatically satisfied for any weighting function $\mathscr{W}(x) \geq 0$. To see this, first note that $\int \mathscr{W}(x) d\mathbb{P}^X(x) \geq \max_{x \in [G]} \{\mathscr{W}(x)\nu(x)\}$. It also holds that

$$\int \mathscr{W}(x)^2 d\mathbb{P}^X(x) \leq G \cdot \max_{x\in[G]} \left\{\mathscr{W}(x)^2 \nu(x)\right\} \leq \left(G \Big/ \min_{x\in[G]} \{\nu(x)\} \right) \cdot \max_{x\in[G]} \mathscr{W}(x)^2 \nu(x)^2$$

Thus, (7) holds with $\Gamma = G \big/ \min_{x\in[G]} \{\nu(x)\}$. It remains to check part (c) of Proposition 1. Recall from algorithms 1, 2, that the weighting rules take the form, for $x \in [G]$:

$$\widehat{W}(x) \propto \frac{1 - \widehat{\pi}_0(x)}{\widehat{\pi}_0(x)}, \; \widehat{\pi}_0(x) := \frac{1 + \sum_{i: X_i = x} \mathbf{1}\left(P_i > \tau\right)}{|\{i : X_i = x\}|\,(1-\tau)}$$

We need to exhibit the weighting function towards which the aforementioned weighting function converges under Assumption 3. To this end, let us note that:

$$\mathbb{E}[\widehat{\pi}_0(x) \,\big|\, X_1, \dots, X_m] = \frac{1 + |\{i : X_i = x\}| \cdot \left[\pi_0(x)\cdot(1-\tau) + (1-\pi_0(x))\cdot\left(1 - F_{\text{alt}}(\tau \,\big|\, x)\right)\right]}{|\{i : X_i = x\}|\,(1-\tau)}$$

Next, define:

$$\pi_0^*(x) = \pi_0(x) + \frac{(1-\pi_0(x))\cdot\left(1 - F_{\text{alt}}(\tau \,\big|\, x)\right)}{1-\tau}$$

Note that for $\tau \in (0,1)$, by assumptions, $\pi_0(x) < 1$ and $F_{\text{alt}}(\tau \,\big|\, x) > \tau$ and so $\pi_0^*(x) < 1$. To avoid dealing with the (unlikely in multiple testing applications) situation that $\pi_0^*(x) = 0$ we further assume that either $\pi_0(x) > 0$ (i.e., there are at least some null hypotheses) or $F_{\text{alt}}(\tau \,\big|\, x) < 1$. Thus henceforth we assume that $\pi_0^*(x) \in (0,1)$.

The asymptotic weight function is $W^*(x)$ defined as

$$W^*(x) = \frac{1-\pi_0^*(x)}{\pi_0^*(x)} \Bigg/ \left(\sum_{g=1}^{G} \nu(g) \cdot \frac{1-\pi_0^*(g)}{\pi_0^*(g)} \right)$$

Notice that indeed $\int W^*(x) d\mathbb{P}^X(x) = \sum_{g=1}^G \nu(g) W^*(g) = 1$. By applications of the law of large numbers and the continuous mapping theorem, we may deduce that:

$$\widehat{\pi}_0(x) = \pi_0^*(x) + o_{\mathbb{P}}(1), \; \frac{|\{i : X_i = x\}|}{m} = \nu(x) + o_{\mathbb{P}}(1), \; \widehat{W}(x) = W^*(x) + o_{\mathbb{P}}(1)$$

We may conclude by noting that $\mathcal{X} = [G]$ is finite, and so $\widehat{W}(x) = W^*(x) + o_{\mathbb{P}}(1)$ for all $x \in [G]$ implies that

$$\left\| \widehat{W}(\cdot) - W^*(\cdot) \right\|_\infty = o_{\mathbb{P}}(1)$$

Supplement S3: Multiple testing with local false discovery rates

Consider the conditional two-groups model (6) and assume that $F(t \mid x)$ has Lebesgue density $f(t \mid x)$ for all x. Then define the conditional local fdr:

$$\mathrm{fdr}(t \mid x) = \frac{\pi_0(x)}{f(t \mid x)} \tag{41}$$

We make two observations: First, for any threshold function $g : \mathcal{X} \to [0, 1]$, one may show that

$$\mathbb{E}[\mathrm{fdr}(P_i \mid X_i) \mid P_i \leq g(X_i)] = \mathbb{P}[H_i = 0 \mid P_i \leq g(X_i)] \tag{42}$$

Equation (42) implies that we can estimate the FDR of a procedure with decision threshold g (i.e., of the procedure that rejects hypotheses that satisfy $P_i \leq g(X_i)$) by

$$\widehat{\mathrm{Fdr}}(g) = \frac{\sum\limits_{i=1}^{m} \mathrm{fdr}(P_i \mid X_i)\, \mathbf{1}_{\{P_i \leq g(X_i)\}}}{\sum\limits_{i=1}^{m} \mathbf{1}_{\{P_i \leq g(X_i)\}}} \tag{43}$$

Second, optimality considerations for multiple testing under model (6) dictate that hypotheses should be ranked by $\mathrm{fdr}(P_i|X_i)$ [Sun and Cai, 2007, Cai and Sun, 2009].

Putting these two ideas together, we arrive at the oracle multiple testing procedure in Algorithm 4.

Algorithm 4: The local fdr multiple testing procedure

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

Such a procedure indeed controls the FDR [Cai and Sun, 2009], if the conditional two-groups model (6) 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 $\pi_0(\cdot)$ [Cai and Sun, 2009]. Such a procedure can be shown to be asymptotically consistent, albeit no finite-sample results are available.

Supplement S4: Estimation and optimization of conditional two-groups model

S4.1 The nonparametric Grenander estimator

Our application of the Grenander estimator [Grenander, 1956] to estimating the conditional two-groups model begins by binning the covariate X_i; for example through quantile-slicing or as the leaves of a tree. Henceforth we will assume X_i is discrete and $X_i \in [G]$.

S4.1.1 Estimation

To estimate $\widehat{F}^{-\ell}(\cdot \mid g)$, $g \in [G]$ we first form the ECDF (empirical cumulative distribution function) of the p-values P_i with $i \notin I_\ell$ and $X_i = g$. Then we compute the least concave majorant of the ECDF. The latter operation can be computed fast through weighted isotonic regression; as implemented for example in the `gcmlcm` function of the R package `fdrtool` [Strimmer, 2008a]. Furthermore, the computational complexity of fitting the Grenander estimator in one group is $O(m_g \cdot \log(m_g))$, , where $m_g = \#\{i : X_i = g\}$, and so it is of order $O(m \cdot \log(m))$ for all the groups.

The estimated $\widehat{F}^{-\ell}(\cdot \mid g)$ is a piecewise-linear, concave function. In particular, for a finite index set[13] $\mathcal{J}_g$ and real numbers a_j^g, b_j^g for $j \in \mathcal{J}_G$ it holds that

$$\widehat{F}^{-\ell}(t \mid g) = \min_{j \in \mathcal{J}_g} \{a_j^g + b_j^g\, t\} \tag{44}$$

For applications to FDR control we also need to estimate $\widehat{\pi}_0^{-\ell}(g)$. We set it to $\widehat{\pi}_0^{-\ell}(g) = 1$ in all our experiments. An alternative would be to apply a π_0 estimator developed in the setting without groups (such as the estimator of Storey et al. [2004]) to the p-values P_i with $i \notin I_\ell$ and $X_i = g$. This would yield $\widehat{\pi}_0^{-\ell}(g)$.

S4.1.2 Optimization through linear programming

Using the Grenander estimator simplifies subsequent optimization in two ways: first, the optimization variable is G-dimensional –instead of $|I_\ell|$-dimensional– as all $i \in I_\ell$ with $X_i = g$ receive the same weight. Second, (44) enables us to cast the underlying convex optimization problems as linear programs by introducing additional variables $\bar{F}_g \in [0,1]$ $(g = 1, \ldots, G)$.

Optimization (8) **for k-Bonferroni:** Let $k_\alpha = \alpha k/m$. We then solve the following linear program (LP) with optimization variables $(w_g, \bar{F}_g),\ g = 1, \ldots, G$:

$$\begin{aligned} \text{maximize} \quad & \sum_{g=1}^{G} |\{i \in I_\ell : X_i = g\}| \cdot \bar{F}_g \\ \text{s.t.} \quad & \bar{F}_g \le a_j^g + b_j^g \cdot k_\alpha \cdot w_g,\ j \in \mathcal{J}_g,\ g = 1, \ldots, G \\ & \sum_{g=1}^{G} |\{i \in I_\ell : X_i = g\}| \cdot w_g = |I_\ell| \\ & w_g \ge 0,\ g = 1, \ldots, G \end{aligned} \tag{45}$$

In our implementation, we solve this LP problem with the open-source `SYMPHONY/Clp` solver of the COIN-OR project [Lougee-Heimer, 2003]. Hypothesis i in fold I_ℓ with covariate $X_i = g$ is then assigned weight w_g, where $(w_1, \ldots, w_G)$ is the optimal weight vector of the above optimization problem.

Optimization (10) **for BH:** The optimization here is similar with the difference that we optimize directly over the thresholds $t_g,\ g = 1, \ldots, G$ and we also enforce the plug-in FDR constraint. Concretely, we solve the linear program with optimization variables $(t_g, \bar{F}_g),\ g = 1, \ldots, G$:

$$\begin{aligned} \text{maximize} \quad & \sum_{g=1}^{G} |\{i \in I_\ell : X_i = g\}| \cdot \bar{F}_g \\ \text{s.t.} \quad & \bar{F}_g \le a_j^g + b_j^g \cdot t_g,\ j \in \mathcal{J}_g,\ g = 1, \ldots, G \\ & \sum_{g=1}^{G} |\{i \in I_\ell : X_i = g\}| \left(\widehat{\pi}_0^{-\ell}(g) \cdot t_g - \alpha \cdot \bar{F}_g\right) \ \le\ 0 \\ & 0 \le t_g \le 1,\ g = 1, \ldots, G \end{aligned} \tag{46}$$

Letting $(t_1, \ldots, t_G)$ the optimal threshold vector, we then let

$$w_g = |I_\ell| \cdot t_g \Big/ \left(\sum_{g=1}^{G} |\{i \in I_\ell : X_i = g\}|\, t_g \right),$$

unless all $t_g = 0$, in which case we set all $w_g = 1$. w_g is the weight assigned to hypotheses $i \in I_\ell$ with $X_i = g$.

[13] We omit the out-of-fold specification $(-\ell)$ from subsequent notation when it improves readability.

Convex constraints on the weights: For both linear programs (45) and (46) it is possible to incorporate additional linear constraints (so that the problems remain linear programs) that enforce weight functions of lower complexity. A concrete example is to enforce low total variation of the weight vector $(w_1, \ldots, w_G)$ i.e., to enforce $\sum_{g=2}^{G} |w_g - w_{g-1}| \leq \lambda$, for $\lambda \geq 0$. This may be directly incorporated into (45). We may also add this constraint to problem (46) in terms of $t_1, \ldots, t_G$ as follows

$$\sum_{g=2}^{G} |t_g - t_{g-1}| \leq \frac{\lambda}{|I_\ell|} \sum_{g=1}^{G} |\{i \in I_\ell : X_i = g\}| \, t_g \tag{47}$$

Throughout this work we always set $\lambda = \infty$ (i.e., we do not add the above total variation constraints), unless explicitly mentioned otherwise.

S4.1.3 Direct optimization

Here we describe an alternative optimization scheme that does not require the use of a linear programming solver and has strong computational complexity guarantees. For our numerical examples, however, the linear programming approach is fast enough.

We describe our algorithm for solving the k-Bonferroni objective (8); the steps for the BH objective (46) are similar.

Let (44) be the fitted Grenander estimator in group g. Let the non-zero slopes in group g be sorted as $b_1^g > \ldots > b_{|\mathcal{J}_g|}^g$=0 and let $s_j^g, j \in \mathcal{J}_g$ be the points at which the slope changes, i.e., the slope is equal to b_j^g in the interval (s_{j-1}^g, s_j^g). At the boundaries we define $s_0^g = 0$ and $s_{|\mathcal{J}_g|}^g = 1$. Further, consider the set:

$$\mathcal{B} = \left\{ b_j^g : \; j \in \mathcal{J}_g, \; g \in [G] \right\} \tag{48}$$

Algorithm 5 provides a computational routine for optimizing the objective (8) with computational complexity upper bounded by $O(m \cdot \log(m))$.
The following proof verifies the correctness of the algorithm above and the worst-case computational complexity.

Proof. We need to first check that the algorithm terminates, i.e., that there exists a λ^* so that $1 \in [\text{WeightBudget}_\ell(\lambda), \text{WeightBudget}_u(\lambda)]$ To this end, note that if we choose $\lambda = \max \mathcal{B}$, then all $\ell_g(\lambda) = 0$ and so $\text{WeightBudget}_\ell(\lambda) = 0$. On the other hand, letting $\lambda = 0$, then we can pick all $u_g(\lambda) = 1$, i.e., $\text{WeightBudget}_u(\lambda) = 1/k_\alpha > 1$. It remains to observe that for adjacent $\lambda_j < \lambda_{j+1}$ in $\mathcal{B}$, it holds that

$$\text{WeightBudget}_\ell(\lambda_{j+1}) = \text{WeightBudget}_u(\lambda_j),$$

and also that for all λ, $\text{WeightBudget}_\ell(\lambda) \leq \text{WeightBudget}_u(\lambda)$. As the algorithm terminates, we may now check computational complexity. First, note that $|\mathcal{B}| = O(m)$ since the Grenander estimator can only jump at support points of the per-group empirical distribution function. Thus the initial sorting step of $\mathcal{B}$ requires at most $O(m \log(m))$ operations. The 'while' loop of the algorithm proceeds by bisection of $\mathcal{B}$, hence will comprise of at most $O(\log(m))$ iterations and the cost of each iteration step is at most $O(m)$. Computation after the while loop is negligible ($O(G)$ operations). Thus, the total complexity of this algorithm is $O(m \log(m))$ at most.

Second, we need to check the Karush–Kuhn–Tucker (KKT) [Rockafellar, 1970] conditions for convex programming to verify the optimality of the weights returned by Algorithm 5. Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_G)$ the dual variables corresponding to the non-negativity constraint and λ the dual variable corresponding to the weight-budget constraint. The Lagrangian takes the form

$$\mathcal{L}(\mathbf{w}, \lambda, \boldsymbol{\nu}) = \sum_{g=1}^{G} \tilde{m}_g \cdot \widehat{F}^{-\ell}(w_g \cdot k_\alpha) + \boldsymbol{\nu}^\top \mathbf{w} - \lambda \cdot k_\alpha \left(\sum_{g=1}^{G} \tilde{m}_g w_g - \tilde{m} \right).$$

We seek to specify dual and primal optimal variables. We set the dual λ^* and primal w_g^* as described in the last steps of Algorithm 5. For the dual variables corresponding to the non-negativity constraints, we set $\nu_g^* = 0$ if $w_g^* > 0$ and $\nu_g^* = \tilde{m}_g \cdot k_\alpha (\lambda^* - b_1^g)$ if $w_g = 0$. **Complementary**

Algorithm 5: Optimization of the k-Bonferroni objective (8) when X-conditional distributions are estimated with Grenander's method.

```
Input : the number of type-I errors to protect against $k$
        the nominal level $\alpha$
        the total number of tests $m$
        the number of tests within fold $\ell$ and each group $g \in [G]$, $\tilde{m}_g = |\{i \in I_\ell : X_i = g\}|$
        the fitted Grenander estimator (44) with slope change points $s_j^g$ and slopes $b_j^g$
        the set $\mathcal{B}$ defined in (48)
---------------------------------------------------------------------------------------------
Sort the set $\mathcal{B}$.
Let $k_\alpha = \alpha \cdot k/m$ and $\tilde{m} = \sum_{g=1}^{G} \tilde{m}_g$.
while $|\mathcal{B}| > 1$ do
    Let $\lambda$ be a middle element in $\mathcal{B}$ (i.e., the median if $|\mathcal{B}|$ is odd).
    for $g \in [G]$ do
        if there exists $j \in \mathcal{J}_g$ such that $\lambda = b_j^g$ then
            let $\ell_g(\lambda) = s_{j-1}^g$, $u_g(\lambda) = s_j^g$.
        else if there exists $j \in \mathcal{J}_g$ such that $\lambda \in (b_{j+1}^g, b_j^g)$ then
            let $\ell_g(\lambda) = u_g(\lambda) = s_j^g$.
        else if $\lambda > b_j^g$ for all $j \in \mathcal{J}_g$ then
            let $\ell_g(\lambda) = u_g(\lambda) = 0$.
    end
    Compute: $\text{WeightBudget}_\ell(\lambda) = \sum_{g=1}^{G} (\tilde{m}_g/\tilde{m}) \ell_g(\lambda)/k_\alpha$ and
             $\text{WeightBudget}_u(\lambda) = \sum_{g=1}^{G} (\tilde{m}_g/\tilde{m}) u_g(\lambda)/k_\alpha$

    if $\text{WeightBudget}_\ell(\lambda) > 1$ then
        $\mathcal{B} \leftarrow \{b \in \mathcal{B} : b > \lambda\}$
    else if $\text{WeightBudget}_u(\lambda) < 1$ then
        $\mathcal{B} \leftarrow \{b \in \mathcal{B} : b < \lambda\}$
    else if $1 \in [\text{WeightBudget}_\ell(\lambda), \text{WeightBudget}_u(\lambda)]$ then
        $\mathcal{B} \leftarrow \{\lambda\}$
end
Let $\lambda^*$ the unique element remaining in $\mathcal{B}$.
Let $c = (\text{WeightBudget}_u(\lambda^*) - 1)/(\text{WeightBudget}_u(\lambda^*) - \text{WeightBudget}_\ell(\lambda^*))$.
Let $t_g = c\ell_g(\lambda^*) + (1-c)u_g(\lambda^*),\ g \in [G]$.
Return the weights $w_g^* = t_g/k_\alpha,\ g \in [G]$.
```

slackness thus holds by construction. Furthermore, note that when $w_g^* = 0$, then Algorithm 5 ensures that $\lambda^* \geq b_1^g$ and so $\nu_g^* \geq 0$. Hence for all g, $\nu_g^* \geq 0$ and so **dual feasibility** holds. **Primal feasibility**, i.e., $w_g^* \geq 0$ and $\sum \tilde{m}_g w_g^* = \tilde{m}$ also hold by construction.

It remains to check that **stationarity** holds. Let us take take the superdifferential of the Lagrangian along the g-th coordinate, where we keep $g \in [G]$ fixed.

$$\partial_g L(\mathbf{w}, \lambda, \nu) = \tilde{m}_g \cdot k_\alpha \cdot \left(\partial \widehat{F}^{-\ell}\left(\alpha \cdot k \cdot w_g/m \,\middle|\, g\right) - \lambda\right) + \nu_g$$

We next distinguish two cases for according to the value of w_g^*. **Case 1,** $w_g^* = 0$**:**. In this case, $b_1^g \in \partial \widehat{F}^{-\ell}\left(w_g^* \cdot k_\alpha \,\middle|\, g\right)$ and ν_g^* is defined precisely so that $0 \in \partial_g L(\mathbf{w}^*, \lambda^*, \nu^*)$.

Case 2, $w_g^* > 0$**:** First let us quickly study $\partial \widehat{F}^{-\ell}(t \,|\, g)$ for $t \in (0, 1)$. If $t = s_j^g$ for $j \in \mathcal{J}_g$, then $\partial \widehat{F}^{-\ell}(t \,|\, g) = [b_{j+1}^g, b_j^g]$ and if $t \in (s_{j-1}^g, s_j^g)$, then $\partial \widehat{F}^{-\ell}(t \,|\, g) = \{b_j^g\}$. In both cases, it holds that $\lambda^* \in \widehat{F}^{-\ell}(t \,|\, g)$ and so, since $\nu_g^* = 0$, it again follows that $0 \in \partial_g L(\mathbf{w}^*, \lambda^*, \nu^*)$. □

S4.2 Beta-uniform mixture GLM

In this section we consider the conditional two-groups model (6) with parametrization (9), where we assume throughout that $\beta(x) < 1, 0 < \pi_0(x) < 1$ hold strictly. We first explain how to estimate the parameters of the conditional two-groups model given access to X_i and censored p-values $(P_i \mathbf{1}(P_i > \tau))$ and then we explain the optimization procedure for deriving optimal weights.

As a preliminary step, we introduce explicit notation for the CDF and pdf of the Beta$(\beta, 1)$ distribution

$$F_\beta(t) = t^\beta,\ f_\beta(t) = \beta t^{\beta - 1}$$

S4.2.1 Estimation

In this section we let $Y_i = -\log(P_i)$. Our goal is estimation based on the censored data outside fold ℓ, i.e., $D_{-\ell}(\tau) = ((X_i, P_i \mathbf{1}(P_i > \tau)))_{i \in I_\ell^c}$[14]. We will proceed by maximum likelihood estimation and optimize the (non-convex) objective through the EM algorithm. The full-data (i.e., if we could observe $((X_i, P_i, H_i))_{i \in I_\ell^c}$) log-likelihood decouples into the sum of the log-likelihood of two generalized linear models (GLMs); a binomial GLM and a Gamma GLM (cf. (16) in Lei and Fithian [2018])

$$\begin{aligned} \ell(a, b; \mathbf{P}, \mathbf{X}, \mathbf{H}) = &\sum_{i \in I_\ell^c} \left(-H_i(a_0 + a^\top X_i) - \log\left[1 + \exp\left(-a_0 - a^\top X_i\right)\right]\right) \\ &+ \sum_{i \in I_\ell^c} \left(H_i\left[-Y_i(b_0 + b^\top X_i) + \log\left(b_0 + b^\top X_i\right)\right]\right) \end{aligned} \tag{49}$$

During the r-th iteration of EM, we keep track of the imputed data (E-step; more below) $\hat{Y}^{(r)}$, $\hat{H}^{(r)}$. Furthermore in principle we should keep track of the parameters $\hat{a}_0^{(r)}, \hat{a}^{(r)}, \hat{b}_0^{(r)}, \hat{b}^{(r)}$. Instead, we keep track of $\hat{\pi}_0{}^{(r)}(x) = \text{expit}(-\hat{a}_0^{(r)} - \hat{a}^{(r)\top} x)$ and $\hat{\beta}^{(r)}(x) = \hat{b}_0^{(r)} + \hat{b}^{(r)\top} x$ both evaluated at $X_i, i \in I_\ell^c$.

We now describe the details of the EM algorithm.

E-step: For the r-th E-step, we need to compute:

$$\mathbb{E}_{\hat{\pi}_0^{(r-1)}, \hat{\beta}^{(r-1)}}\left[\ell(a, b; \mathbf{P}, \mathbf{X}, \mathbf{H}) \mid D_{-\ell}(\tau)\right]$$

This boils down to computing

$$\hat{H}_i^{(r)} = \mathbb{E}_{\hat{\pi}_0^{(r-1)}, \hat{\beta}^{(r-1)}}\left[H_i \mid D_{-\ell}(\tau)\right] \text{ and } \hat{Y}_i^{(r)} = \mathbb{E}_{\hat{\pi}_0^{(r-1)}, \hat{\beta}^{(r-1)}}\left[Y_i \mid H_i = 1, D_{-\ell}(\tau)\right]$$

and plugging these into (49) in lieu of H_i, Y_i.

[14] The case $\tau = 0$ corresponds to no censoring. Without cross-weighting we use the data corresponding to all indices $i = 1, \ldots, m$.

- $\hat{H}_i^{(r)}$ update:

$$\hat{H}_i^{(r)} = \begin{cases} 1 - \frac{\hat{\pi}_0^{(r-1)} \cdot \tau}{\hat{\pi}_0^{(r-1)} \cdot \tau + (1 - \hat{\pi}_0^{(r-1)}) \cdot F_{\hat{\beta}^{(r-1)}(X_i)}(\tau)} & , \text{ if } P_i \leq \tau \\ 1 - \frac{\hat{\pi}_0^{(r-1)}}{\hat{\pi}_0^{(r-1)} + (1 - \hat{\pi}_0^{(r-1)}) \cdot f_{\hat{\beta}^{(r-1)}(X_i)}(P_i)} & , \text{ if } P_i > \tau \end{cases}$$

- $\hat{Y}_i^{(r)}$ update:

$$\hat{Y}_i^{(r)} = \begin{cases} \frac{1}{\hat{\beta}^{(r-1)}(X_i)} - \log(\tau) & , \text{ if } P_i \leq \tau \\ Y_i & , \text{ if } P_i > \tau \end{cases}$$

M-step: As already alluded to, the M-step consists of fitting two GLMs, a (quasi)binomial GLM ($\hat{H}_i^{(r)} \in [0, 1]$ takes on fractional values) and a weighted Gamma GLM. In R pseudocode, the M-step is as follows:

-
$$\hat{\pi}_0^{(r)} = \text{predict}(\text{glm}(1 - \hat{H}^{(r)} \sim 1 + X,\ \text{family=quasibinomial()}),\ \text{type=“response”})$$

-
$$\hat{\beta}^{(r)} = \text{predict}(\text{glm}(\hat{Y}^{(r)} \sim 1 + X,\ \text{family=Gamma(), weights=}\hat{H}^{(r)}),\ \text{type=“link”})$$

In this step we also seek to ensure $\beta(x) < 1, 0 < \pi_0(x) < 1$ (so that strict concavity of the estimated p-value distribution holds conditionally on all x). To this end we introduce parameters $\pi_{1,\text{min}}, \pi_{1,\text{max}}$ and β_{max} and clamp the $\beta(x), \pi_0(x)$ estimates ($\text{clamp}(x; a, b) = \max\{\min\{x, b\}, a\}$) to the above ranges. In our implementation we use $\pi_{0,\text{min}} = 0.1, \pi_{0,\text{max}} = 0.99$ and $\beta_{\text{max}} = 0.9$ (we have not needed to lower bound β in our experiments).

Initialization:

- $\hat{Y}^{(0)}$: We initialize $\hat{Y}_i$ by Y_i if P_i is not censored and by $-\log(\tau/2)$ otherwise.

$$\hat{Y}_i^{(0)} = Y_i \mathbf{1}(P_i > \tau) - \log(\tau/2) \mathbf{1}(P_i \leq \tau)$$

- $\hat{\pi}_0^{(0)}$: The $\hat{\pi}_0(X_i)$ are initialized through the procedure of Boca and Leek [2018]. First, let $\tau^{\text{BL}} \geq \tau$; in our simulations we use $\tau^{\text{BL}} = 0.5$. Then we fit a logistic regression of $\mathbf{1}\left(P_i \geq \tau^{\text{BL}}\right)$ onto X_i, let $\hat{\mathbb{P}}[P_i \geq \tau^{\text{BL}} \mid X_i]$ the fitted probabilities and finally we set

$$\hat{\pi}_0^{(0)}(X_i) = \text{clamp}(\ \hat{\mathbb{P}}[P_i \geq \tau^{\text{BL}} \mid X_i]/(1 - \tau);\ \pi_{0,\text{min}},\ \pi_{0,\text{max}})$$

- $\hat{H}^{(0)}$: We first compute the adjusted p-values adjP_i of the BH procedure applied to $P_i \vee \tau$ (i.e., in R pseudocode: `p.adjust(pmax(Ps, tau), method="BH")’’)` and then we set:

$$\hat{H}_i^{(0)} = 1 - \text{adj}P_i \cdot \hat{\pi}_0^{(0)}(X_i)$$

Output: Let r^* the final iteration of the EM algorithm, we keep $\hat{\beta}^{-\ell}(\cdot) = \hat{\beta}^{(r^*)}(\cdot)$ and $\hat{\pi}_0^{-\ell}(\cdot) = \hat{\pi}_0^{(r^*)}(\cdot)$. These fully specify the estimated conditional distribution

$$\widehat{F}^{-\ell}(t \mid x) = \hat{\pi}_0^{(r^*)}(x) \cdot t + (1 - \hat{\pi}_0^{(r^*)}(x)) \cdot F_{\hat{\beta}^{-\ell}(x)}(t)$$

.

S4.2.2 Optimization

The estimated conditional distributions and densities take the form:

$$\widehat{F}^{-\ell}(t \mid X_i = x) = \hat{\pi}_0^{-\ell}(x) \cdot t + (1 - \hat{\pi}_0^{-\ell}(x))t^{\hat{\beta}^{-\ell}(x)}$$

$$\hat{f}^{-\ell}(t \mid X_i = x) = \hat{\pi}_0^{-\ell}(x) + (1 - \hat{\pi}_0^{-\ell}(x)) \cdot \hat{\beta}^{-\ell}(x) \cdot t^{\hat{\beta}^{-\ell}(x)-1}$$

Optimization (8) **for** ***k*****-Bonferroni:** We seek to maximize $\sum_{i \in I_\ell} \widehat{F}^{-\ell}(k_\alpha \cdot w_i \mid X_i)$ subject to $w_i \geq 0$, $\sum_{i \in I_\ell} w_i = |I_\ell|$, where $k_\alpha = \alpha k/m$. This a convex optimization problem and furthermore strong duality is attained, e.g., by Slater's condition (also note that the program is feasible; take $w_i = 1$).

Assume momentarily that the optimizer satisfies $w_i > 0$ for all i. Let λ be the Lagrange multiplier corresponding to the constraint $\sum_{i \in I_\ell} w_i = |I_\ell|$. Then, differentiating the Lagrangian with respect to w_i, we see that it must hold that:

$$\hat{f}^{-\ell}(k_\alpha \cdot w_i \mid X_i)k_\alpha - \lambda \overset{!}{=} 0$$

So:

$$\hat{\pi}_0^{-\ell}(x) + (1 - \hat{\pi}_0^{-\ell}(x)) \cdot \hat{\beta}^{-\ell}(x) \cdot (k_\alpha \cdot w_i)^{\hat{\beta}^{-\ell}(x)-1} \overset{!}{=} \lambda/k_\alpha$$

Since $\hat{\beta}^{-\ell}(X_i) < 1$ and $\hat{\pi}_0^{-\ell}(X_i) < 1$ by our estimation procedure, we may solve the equation above analytically for $w_i > 0$. We call this solution $w_i(\lambda)$. Then we use bisection over λ to find λ^* such that the equality constraint is satisfied, i.e., $\sum_{i \in I_\ell} w_i(\lambda^*) = |I_\ell|$. Then the optimizing weights are $w_i = w_i(\lambda^*)$.

We may derive the computational complexity of the optimization step as follows: We can minimize the Lagrangian analytically in $O(m)$ operations. To find the optimal dual variable λ^* we need to use bisection. Thus, we need roughly $O(m \cdot \log(1/\delta))$ operations, where δ is a parameter controlling tolerance (accuracy).

Optimization (10) **for BH:** Here we seek to maximize $\sum_{i \in I_\ell} \widehat{F}^{-\ell}(t_i \mid X_i)$ over $t_i \geq 0$ subject to $\sum_{i \in I_\ell} \hat{\pi}_0^{-\ell}(X_i)t_i \leq \alpha \sum_{i \in I_\ell} \widehat{F}^{-\ell}(t_i \mid X_i)$. We may directly verify the conditions of Theorem 2 in Lei and Fithian [2018] (which ensures strong duality) and conclude that there exists $\lambda \in (0,1)$ such that at the optimal solution:

$$\text{fdr}^{-\ell}(t_i \mid X_i) \overset{!}{=} \lambda \text{ for all } i \in I_\ell$$

Here $\text{fdr}^{-\ell}(t_i \mid X_i)$ is defined as in (41) with population quantities replaced by estimated ones. Rearranging, this implies that:

$$\hat{f}^{-\ell}(t_i \mid X_i) \overset{!}{=} \hat{\pi}_0^{-\ell}(X_i)/\lambda$$

As already described for k-Bonferroni, for each fixed λ we may solve the above expression analytically for t_i, say by $t_i(\lambda) > 0$. Then it only remains to use bisection to find λ^* such that

$$\sum_{i \in I_\ell} \hat{\pi}_0^{-\ell}(X_i)t_i(\lambda^*) = \alpha \sum_{i \in I_\ell} \widehat{F}^{-\ell}\left(t_i(\lambda^*) \mid X_i\right).$$

Finally, hypothesis $i \in I_\ell$ is assigned weight $W_i = |I_\ell| \cdot t_i(\lambda^*) \big/ \sum_{j \in I_\ell} t_j(\lambda^*)$.

We note that here, just as for k-Bonferroni, the computational complexity scales as $O(m \cdot \log(1/\delta))$ operations, where δ is a parameter controlling tolerance (accuracy).

Supplement S5: More details on the data application of Section 6

For the hQTL example, we used the dataset described in Grubert et al. [2015] and looked for associations between SNPs and the histone modification mark (H3K27ac) on human Chromosomes 1 and 2. p-values for association were calculated as described in the original paper Grubert et al. [2015] using Matrix eQTL Shabalin [2012].

As a covariate we used the linear genomic distance between the SNP and the ChIP-seq signal, which we discretized using non-uniform binning: the bins corresponded to genomic segments of length 10 kb (kilobase) bins up to 300 kb (i.e., the categories were $0 - 10$ kb, $10 - 20$ kb, $\ldots$, $290 - 300$ kb), to segments of length 100 kb up to 1 Mb and finally to segments of length 10 Mb for the rest of the hypotheses. The longest genomic distance between SNPs and H3K27ac was approximately equal to 24 Mb.

For the application of IHW-BY (cf. Theorem 4), we split hypotheses into two folds corresponding to the two chromosomes. Honest weights are learned within each fold with the strategy described in Section 4.2 and Supplement S4.1 based on the Grenander estimator. Note that we set $\hat{\pi}_0^{-\ell} = 1$. Furthermore, we apply a mild constraint on the total variation of the learnd weights, i.e. by including the constraint (47) with $\lambda = 2000$ in the linear programming problem (46).

Supplement S6: Choice and examples of informative covariates

Covariates that can take the role of X_i in the conditional two-groups model (6) are available in many multiple testing applications of practical interest, and in this section we discuss a range of examples. We will group them into domain-specific and statistical covariates. Whereas the former derive from an understanding of the data-generating process, the latter reflect mathematical properties of the specific test procedure used to compute the p-values. Domain-specific covariates are often informative about prior probabilities (i.e., the function $\pi_0(x)$ depends on x), statistical covariates about the power of the test and thus the shape of the alternative distribution function $F_{\text{alt}}(\cdot \mid X_i = x)$. The categorization is informal, loose and partially overlapping.

For a given application, there will often be more than one possible choice of covariate. In our formulation of the conditional two-groups model (6), we assume for simplicity of notation that X_i is either one particular choice, or the combination of several original covariates into a single "effective" covariate, e.g., by taking the Cartesian product. The details of how to select or combine will depend on the application and the data and are beyond the scope of this paper.

S6.1 Domain-specific covariates

In many scientific applications, informative covariates are apparent to domain scientists due to mechanistic insight or prior experience. Examples include:

- **Genomic distance between SNPs and peaks.** This is the covariate in our motivating example in Figure 1 and Section 6. The p-values are from testing the association between SNPs and H3K27ac peak heights across different individuals from the human population. The choice of covariate is motivated by the expectation that many of the true instances where a DNA polymorphism affects a H3K27ac peak are short-range, so that π_0 for hypotheses with a short distance is smaller than for those where SNP and peak are far apart.

- **Physical distance between pairs of firing neurons**. It is now possible to simultaneously measure the activity of many neurons, and there is interest in determining whether two neurons are firing in synchrony [Scott et al., 2015]. We know that neurons in close proximity are a-priori more likely to be interacting, thus, the distance between neurons can be used as a covariate for association tests between pairs of neurons.

- **Gene expression patterns in nearby genetic variants**. Genome-wide association studies (GWAS) look for statistical associations between genetic variants in a population with prevalence of a disease. Once discovered, such an association can be the basis for a follow-up mechanistic study. Sample size and power tend to be limiting bottlenecks of many GWAS due to multiple testing and to the study's expense. Power can be increased by considering (phenotype-unrelated) gene expression patterns around the loci of the genetic variants [Baillie et al., 2018].

- **P-values from a distinct but related experiment**. For example, Fortney et al. [2015] used data from previous, independent GWAS for related diseases to increase the power of a GWAS study of a longevity phenotype.

In a different context—multivariate regression rather than hypothesis testing—the widespread existence of such covariates was observed by Wiel et al. [2016], who used the term "co-data" for them and developed a weighted ridge regression procedure, with data-driven penalization weights.

S6.2 Statistical covariates

In single hypothesis testing, classical theory [Lehmann and Romano, 2005] dictates that the whole dataset should be reduced to a sufficient statistic, which in turn can be used to derive the best test statistic under optimality considerations. Everything else, can be discarded or should be conditioned on. This data compression comes without any loss of statistical power.

However, the m resulting p-values for the individual tests are in general not able to capture how one should weigh the hypotheses relative to each other to arrive at an optimal multiple testing protocol [Storey, 2007]. The consequence is that information irrelevant for 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.

S6.2.1 Sample size

A generic covariate, likely to be useful whenever it differs across tests, is the sample size N_i. Note that if the test statistic is continuous and the null hypothesis is simple, then the p-value P_i under the null is uniformly distributed independently of N_i. Often, there is no reason to expect that the prior probability of a hypothesis being true depends on N_i. However, the alternative distribution will depend on N_i: for higher sample size, we have more power.

A simple, but generic and instructive example is as follows: consider a series of one-sided z-tests in which we observe independent $Y_1^i, \ldots, Y_{N_i}^i \sim \mathcal{N}(\mu_i, 1)$, where $\mu_i > 0$ if $H_i = 1$ and $\mu_i = 0$ otherwise. We can use $P_i = 1 - \Phi\left(N_i^{1/2}\, \overline{Y^i}\right)$ as our statistic, where $\overline{Y^i}$ is the sample average of $Y_1^i, \ldots, Y_{N_i}^i$. Then the alternative distribution of the i-th test is

$$F_{\text{alt},i}(t) = 1 - \Phi(\Phi^{-1}(1-t) - \sqrt{N_i}\,\mu_i). \tag{50}$$

Now consider the case in which $\pi_{0,i} = \pi_0$ and $\mu_i = \mu H_0\ \forall i$, i.e., a common prior probability and a common effect size. In this case, Equation (50) leads to the conditional two-groups model with covariate N_i and $F_{\text{alt},i}(t) = F_{\text{alt}}(t \mid N_i)$. Then, to maximize discoveries and thus power, hypotheses with large sample sizes N_i should be prioritized. The methods described here are able to accomplish this automatically.

Remark 1. At this point, readers might ask themselves whether this is desirable – since, in practice, different effect sizes μ_i may be present. Prioritizing hypotheses with large sample sizes N_i will lead to a trade-off where some discoveries with smaller N_i but higher μ_i are missed, for the benefit of making more discoveries with larger N_i but smaller μ_i. Yet, the former might be more valuable to us. In a way, one can draw analogies to the streetlight effect: if we have lost our keys during a walk at night and have no idea where it happened, it makes sense to start searching under the streetlight, where it is easiest to see. However, if we do have guesses where we might have dropped them, it makes sense to combine these guesses with the ease of seeing in each place to arrive at an optimal search schedule.

Remark 2. The optimal weights are not necessarily a monotonic function of the sample size. With IHW, it is possible that hypotheses with covariates associated with very large sample size (or effect size) are down-weighted relative to more intermediate hypotheses. This phenomenon is called *size-investing* [Roeder et al., 2007, Peña et al., 2011, Habiger et al., 2017, Ignatiadis et al., 2016]. The intuition is that higher weights should be preferentially allocated where they make most difference – and little to hypotheses that are anyway exceedingly easy or hard to reject.

S6.2.2 Overall variance (independent of label) in ANOVA tests

In Section 5.3 we demonstrated a covariate that can be used to improve power in the simultaneous two-sample testing problem for equality of means in the case of known variances. Here we extend the discussion to the case of unknown variances; cf. Cai et al. [2019] for a comprehensive treatment of more general forms of this problem.

Our data is drawn from model (14). We are interested in testing $H_i : \mu_{Y,i} = \mu_{V,i}$ and do not know σ_i. The optimal test statistic for this situation is the two-sample t-statistic:

$$T_i = \sqrt{n}\frac{\overline{Y_i} - \overline{V_i}}{\sqrt{S_{Y,i}^2 + S_{V,i}^2}}, \tag{51}$$

where $\overline{Y_i}$ and $\overline{V_i}$ are the sample means and $S_{Y,i}^2$ and $S_{V,i}^2$ the sample variances.

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

Now note that under the null hypothesis, $\mu_{Y,i} = \mu_{V,i} = \mu_i$ and $Y_{i,1}, \dots, Y_{i,n}$, $V_{i,1}, \dots, V_{i,n} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ i.i.d. Then, $(\hat{\mu}_i, S_i^2)$ is a complete sufficient statistic for the experiment, while T_i is ancillary for (μ_i, σ_i^2). Thus, by Basu's theorem, $(\hat{\mu}_i, S_i^2)$ is independent of T_i and we can use it as a covariate.

Now consider S_i^2 in particular 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 if $|\mu_{Y,i} - \mu_{V,i}|$ is large. 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 not only independent of the p-values under the null hypothesis, but also informative about the alternatives. Using it as a covariate can lead to a large power increase in simultaneous two-sample t-tests [Bourgon et al., 2010, Ignatiadis et al., 2016]. The result extends to more complex ANOVA settings.

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 $\mu_{Y,i} = \mu_{V,i} = 0$. If we believe a-priori that for many of the hypotheses i with $\mu_{Y,i} = \mu_{V,i}$ a sparsity condition holds, so that in fact $\mu_{Y,i} = \mu_{V,i} = 0$ [Liu, 2014], then large values of this statistic are more likely to correspond to alternatives. Note that we did not actually re-specify our null hypothesis from $\mu_{Y,i} = \mu_{V,i}$ to $\mu_{Y,i} = \mu_{V,i} = 0$. We just assumed properties of the alternatives to motivate a choice of covariate, and are still testing for $\mu_{Y,i} = \mu_{V,i}$.

Remark 3. In single hypothesis testing, there is nothing to be gained from $(\hat{\mu}_i, S_i^2)$. Its usefulness only emerges in the multiple testing setup.

S6.2.3 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 $n_{1,i}\, n_{2,i}/n_i^2$ is a statistic which is related to the alternative distribution, with values close to $\frac{1}{4}$ implying higher power [Roquain and Van De Wiel, 2009]. This statistic is also related to the Minor Allele Frequency (MAF) in genome-wide association studies [Boca and Leek, 2018].

S6.2.4 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 have one or the other sign of effect size. Thus, the sign can be used as an informative covariate. Previous uses of stratification by sign to improve power include the SAM (significance analysis of microarrays) procedure [Tusher et al., 2001].