Ignatiadis, Nikolaos; Huber, Wolfgang

**Article — Published Version**

# Covariate powered cross-weighted multiple testing

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

**Provided in Cooperation with:**
John Wiley & Sons

*Suggested Citation:* Ignatiadis, Nikolaos; Huber, Wolfgang (2021) : Covariate powered cross-weighted multiple testing, Journal of the Royal Statistical Society: Series B (Statistical Methodology), ISSN 1467-9868, Wiley, Hoboken, NJ, Vol. 83, Iss. 4, pp. 720-751, https://doi.org/10.1111/rssb.12411

This Version is available at:
https://hdl.handle.net/10419/284837

Received: 2 May 2018 | Accepted: 27 November 2020

DOI: 10.1111/rssb.12411

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

# Covariate powered cross-weighted multiple testing

**Nikolaos Ignatiadis**[1] | **Wolfgang Huber**[2]

[1]Department of Statistics, Stanford University, Stanford, CA, USA

[2]European Molecular Biology Laboratory, Heidelberg, Germany

**Correspondence**
Wolfgang Huber, European Molecular Biology Laboratory, Heidelberg, Germany.
Email: wolfgang.huber@embl.org

**Funding information**
German Federal Ministry of Education and Research, CompLS project MOFA under grant agreement number 031L0171A; Ric Weiland Graduate Fellowship

## Abstract

A fundamental task in the analysis of data sets 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, for example false discovery rate 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 data sets 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 data set by Grubert et al. (2015), who aimed to discover associations between genetic polymorphisms (single nucleotide 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, $N_1 = 645,452$ and $N_2 = 699,343$ SNPs were recorded, and H3K27ac levels were measured in $K_1 = 12,193$ and $K_2 = 11,232$ regions, which amounts to nearly 16 billion $(N_1K_1 + N_2K_2)$ 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 Equation (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):

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

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

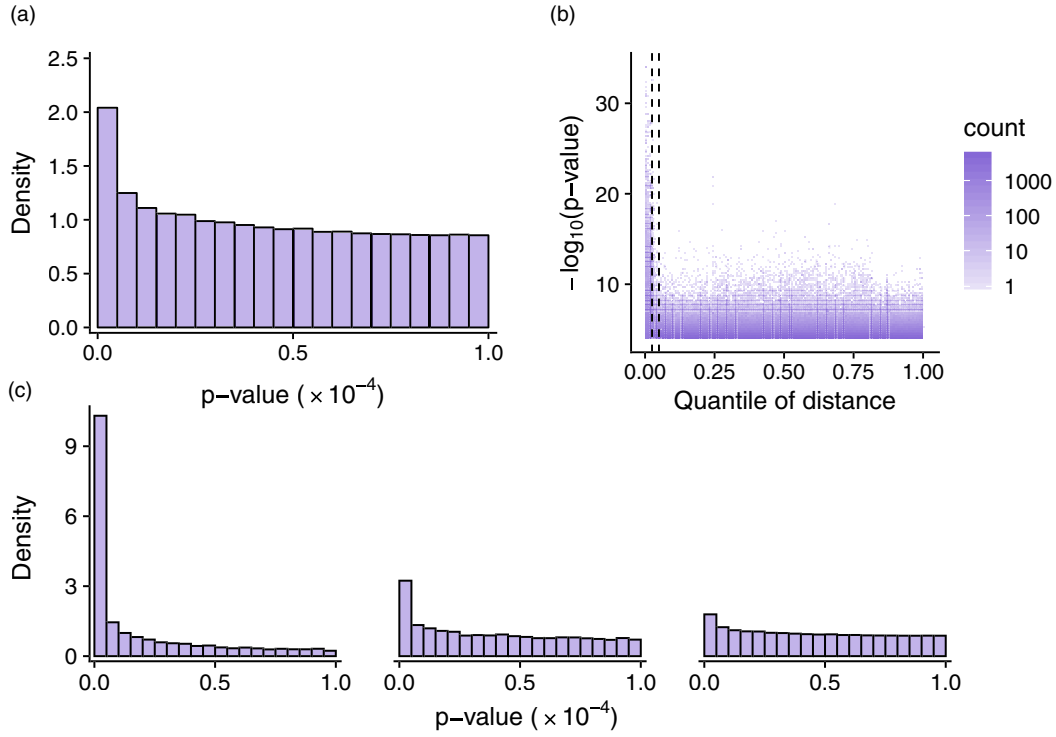


**FIGURE 1** **Heterogeneous multiple hypothesis testing in a biological example**: For each hypothesis, ($i = 1, \ldots, m$), a $p$-value $P_i$ is provided as well as a covariate $X_i$, which here is the genomic distance between the two features tested for association: a single nucleotide polymorphism (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 $\left(-\log_{10}P_i, X_i\right)$: 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). [Colour figure can be viewed at wileyonlinelibrary.com]

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 $I_\ell$. 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 from the $K-1$ folds excluding $I_\ell$ are used. We call this proviso *cross-weighting*.

Conceptually, cross-weighting is related to cross-fitting (Schick, 1986), a method that has been successful in the fields of causal inference (Chernozhukov et al., 2017; Nie & Wager, 2020) and empirical Bayes (Ignatiadis & 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 Equation (2), these thresholds may take the form $\{P_i \leq \hat{t} w_i\}$ parameterized in terms of weights $w_i$ that are optimal for controlling the family-wise error rate (FWER) (Dobriban et al., 2015; Peña et al., 2011; Roeder & Wasserman, 2009) or the FDR (Durand, 2019; Roquain & Van De Wiel, 2009). Furthermore, in the case of FDR control, optimal decision thresholds are known to take the form of contours of equal local fdr (Cai & Sun, 2009; Cai et al., 2019; Efron, 2010; Ferkingstad et al., 2008; Ochoa et al., 2015; Ploner et al., 2006; Scott et al., 2015). 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 & 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 et al. (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 et al. (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 $\alpha$; that is, 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 & Fithian, 2018). Zhang et al. (2017) and Zhang et al. (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 & 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 Equation (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 IHW 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, that is 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 $\text{FWER} := \mathbb{P}[V \geq 1]$, or the $k$-FWER $:= \mathbb{P}[V \geq k]$. In exploratory situations, a typically less stringent objective is to control the FDR, that is, the expectation of the false discovery proportion (FDP), namely $\text{FDR} := \mathbb{E}[\text{FDP}] := \mathbb{E}\left[\frac{V}{R \vee 1}\right]$ (Benjamini & 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 et al., 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:

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

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

**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 & Wolf, 2010), which takes the form (3) with deterministic cutoff $\widehat{t} = k\alpha/m$ and $\tau = 1$. The case $k$=1 is the weighted Bonferroni procedure proposed by Genovese et al. (2006).

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

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

The weighted BH procedure of Genovese et al. (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 $\tau$-censored BH is non-decreasing in $\tau$, so that a procedure with smaller $\tau$ will never make more discoveries. However, for large $\tau$, 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 $\geq 0.5$.

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 (Blanchard & Roquain, 2008; Genovese et al., 2006; Habiger, 2017; Ramdas et al., 2019; Roquain & Van De Wiel, 2009). 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 & 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 & Barber, 2019) or only provide guarantees in the asymptotic limit (Durand, 2019; Hu et al., 2010; Ignatiadis et al., 2016; Roeder et al., 2007; Wang, 2018; Zhao & Zhang, 2014).

## 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 et al. (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 $\alpha$. 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, cf. Ramdas et al. (2019).

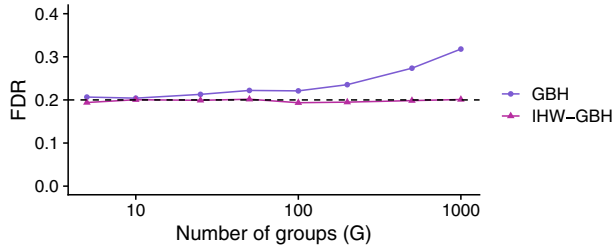


**FIGURE 2** **The need for cross-weighting**: We simulated under the global null with $m = 10{,}000$ independent $P_i \sim U[0,1]$ and $X_i \equiv i \pmod G$, where the number of groups $G$ is a simulation parameter shown on the $x$-axis. Then we applied the GBH and IHW-GBH (with a random partition into five 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 12,000 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 $G$. [Colour figure can be viewed at wileyonlinelibrary.com]

**Algorithm 1:** The Group Benjamini–Hochberg (GBH) algorithm

**Input :** $(P_1, \ldots, P_m) \in [0,1]^m$
$(X_1, \ldots, X_m) \in \{1, \ldots, G\}^m$

a nominal level $\alpha \in (0,1)$
a censoring level $\tau \in (0,1)$

**for** $g = 1, \ldots, G$ **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** $i = 1, \ldots, m$ **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):**

$$\hat{\pi}'_{0,W} := \frac{\max_{i=1,\ldots,m} W_i + \sum_{i=1}^{m} W_i \mathbf{1}(P_i > \tau)}{m(1-\tau)}$$

Update $W_i := W_i / \hat{\pi}'_{0,W}$

Apply weighted BH (Def. 2) with p-values $P_i$ and weights $W_i$.

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

**Input :** $(P_1, \ldots, P_m) \in [0,1]^m$
$(X_1, \ldots, X_m) \in \{1, \ldots, G\}^m$
a partition $I_1, \ldots, I_K$ of $\{1, \ldots, m\}$
a nominal level $\alpha \in (0,1)$
a censoring level $\tau \in (0,1)$

**for** $\ell = 1, \ldots, K$ **do**
  **for** $g = 1, \ldots, G$ **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):**

$$\hat{\pi}'_{0,W,\ell} := \frac{\max_{i \in I_\ell} W_i + \sum_{i \in I_\ell} W_i \mathbf{1}(P_i > \tau)}{|I_\ell|(1-\tau)}$$

  Update $W_i := W_i / \hat{\pi}'_{0,W,\ell}$

  **end**
**end**
Apply the $\tau$-censored, weighted BH procedure (Def. 2) 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. In contrast, 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

**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 $\widehat{W}^{-\ell} : \mathcal{X} \to \mathsf{R}_{\geq 0}$ from the pairs $(P_i, X_i),\ i \notin I_\ell$
  **for** $i \in I_\ell$ **do**
    Let $W_i$ be a suitable rescaling of $\widehat{W}^{-\ell}(X_i)$,

$$W_i := \frac{|I_\ell|\,\widehat{W}^{-\ell}(X_i)}{\sum_{i \in I_\ell} \widehat{W}^{-\ell}(X_i)},\ \text{if} \quad \sum_{i \in I_\ell} \widehat{W}^{-\ell}(X_i) > 0,\ \text{else } W_i := 1$$

  **end**
**end**
Run a weighted multiple testing procedure with p-values $P_i$ and weights $W_i$.

We now generalize the IHW-GBH procedure beyond categorical covariates, the GBH weighting scheme and the weighted BH procedure (Definition 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 $I_\ell$ 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 $\tau$-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.

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

[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. the example in Section 6.

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

($a_1$) The null pairs $((P_i, X_i))_{i\in\mathcal{H}_0}$ are jointly independent.
($a_2$) The null pairs $((P_i, X_i))_{i\in\mathcal{H}_0}$ are independent of the alternative pairs $((P_i, X_i))_{i\notin\mathcal{H}_0}$.
(b) For $i \in \mathcal{H}_0$, it holds that $P_i$ is independent of $X_i$.
(c) For $i \in \mathcal{H}_0$, $P_i$ is super-uniform, i.e., $\mathbb{P}[P_i \le t] \le t$ 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 ($a_1'$) $(P_i)_{i\in\mathcal{H}_0}$ are jointly independent, ($a_2'$) independent of the alternative $p$-values $(P_i)_{i\notin\mathcal{H}_0}$ 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 & 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 $I_1, \ldots, I_K$, that is, $\bigcup_\ell I_\ell = [m]$ and $\left(I_\ell\right)_\ell$ 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 $I_\ell$ average to 1, i.e., $\sum_{i\in I_\ell} W_i = |I_\ell|$ for all $\ell \in [K]$.
(c) $W_i \ge 0$ 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 et al., 2006). Instead of merely requiring $\sum_{i=1}^m W_i = m$, the budget is 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 $\tau$-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 $\tau$ must be concealed from the weighting algorithm.

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

**Theorem 1** *(IHW-BH controls the FDR under honesty and $\tau$-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 $\tau$-censored (Specification 2) for some $\tau \in (0,1]$. Then the $\tau$-censored, weighted BH procedure (Definition 2) with* p*-values $P_i$ and weights $W_i$ controls the FDR at the nominal level* $\alpha$.

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 a 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 $\tau$-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* $\alpha$.

*Proof* By construction, the weights $W_i$ of IHW-GBH are honest and $\tau$-censored.

A shortcoming of IHW-BH with weights that satisfy $\sum_{i=1}^{m} W_i = m$ is that FDR is controlled at $\pi'_{0,W}\alpha \le \alpha$, where $\pi'_{0,W} := (\sum_{i\in\mathcal{H}_0} \mathbb{E}[W_i])/m$ and IHW-BH can thus be needlessly conservative. Motivated by null-proportion adaptive methods for unweighted BH (Storey et al., 2004) and weighted BH with deterministic weights (Habiger, 2017; Ramdas et al., 2019), we estimate $\pi'_{0,W}$ within fold $I_\ell$ by[4]

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

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

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

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.

## 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: first, 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 $\tau$-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 (Deb et al., 2018; Ferkingstad et al., 2008; Lei & Fithian, 2018), which we call the conditional two-groups model and which extends the two-groups model of Storey (2003) and Efron et al. (2001):

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

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

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

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, that is for each fold $\ell$, 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 $\tau$-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 $\ell$, 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.

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

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

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, $\sup_t \left| \frac{1}{m} \sum_{i=1}^m \left[ \mathbf{1}(P_i \leq t) - \mathbb{P}[P_i \leq t] \right] \right| \xrightarrow{\mathbb{P}} 0$ and similarly for the subset of null hypotheses. A consequence is that the BH estimator of the FDR 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: $\sup_t \left| \frac{1}{m} \sum_{i=1}^m \left[ \mathbf{1}(P_i \leq t W_i) - \mathbb{P}[P_i \leq t W_i] \right] \right| \xrightarrow{\mathbb{P}} 0$.

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 $\{(p,x) \mapsto \mathbf{1}(p \leq t W(x)) \,|\, t \in (0,1], W(\cdot) \in \mathcal{W}\}$ 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 $I_\ell$ 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 $\tau$-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, that is, the weights asymptotically converge to a limit.

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

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 $\mathcal{H}_0 \subset [m]$ be 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 $i \in \mathcal{H}_0$, it holds that $P_i$ is independent of $(X_j)_{j\in[m]}$.
(c) For $i \in \mathcal{H}_0$, $P_i$ is super-uniform, that is $\mathbb{P}[P_i \leq t] \leq t$ for all $t \in [0,1]$.

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, for example 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 no 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 et al.

(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 (Cai et al., 2019; Lei & Fithian, 2018), weak dependence (Li & Barber, 2019) or the ability to consistently estimate the joint distribution of all hypotheses (Sun & 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 with respect to $I_1, \ldots, I_K$ (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* $\alpha$.

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

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

Note that in (*), we used the fact that for $i \in \mathcal{H}_0$ 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 & Yekutieli, 2001; Blanchard & 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 wBY procedure at level $\alpha$, that is the weighted BH procedure (Definition 2) with $\tau$=1 at level $\alpha / \sum_{k=1}^{m} \frac{1}{k}$.

We now show that applying the wBY 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 with respect to* $I_1, \ldots, I_K$ *(Specification 1). Then the wBY procedure (Definition 3) with* p*-values* $P_i$ *and weights* $W_i$ *controls the FDR at the nominal level* $\alpha$*.*

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 wBY 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 the partition $I_1 = \{1, \ldots, m'\}, I_2 = \{m'+1, \ldots, m\}$. We first draw a permutation $\sigma$ 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 $i^* \in \operatorname{argmin}_i \{P_i\}$ and then let $W_i = W(X_i) = m\mathbf{1}(X_i = i^*)$. Then the FDR of wBY at $\alpha$ is equal to 1 as long 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 $P_{i^*}/W_{i^*} \leq 2/m^2 < \alpha/(m \sum_{k=1}^{m} \frac{1}{k})$. $H_{i^*}$ gets rejected and so FDP=1 almost surely.

In contrast, FDR control would be guaranteed, had we used weights derived through cross-weighting. BY with $\tau$-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 BY correction, that is, at level $\alpha / \sum_{k=1}^{m} \frac{1}{k}$ and $\tau$-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: we fix $m = 2m', m' \in \mathbb{N}$ and consider the partition $I_1 = \{1, \ldots, m'\}, I_2 = \{m'+1, \ldots, m\}$. 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 $P_2, \ldots, P_{m'} = P_1$ and $P_{m'+2}, \ldots, P_m = P_{m'+1}$. 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 FDR$\geq$0.2925 as long 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. (14)) as $(1 + \sum_i \mathbf{1}(P_i \geq 1 - 0.45))/\sum_i \mathbf{1}(P_i \leq 0.45)$, 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 $m,\alpha$. Thus AdaPT immediately terminates, rejecting all $p$-values in $I_1$, and so FDP = 1. Similarly 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 $\tau$-censored with $\tau$=0.45.

# 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 $\sum_i \mathbb{P}[P_i \leq k\alpha/mW(X_i)\,|\,X_i] = \sum_i F(k\alpha/mW(X_i)\,|\,X_i)$. To derive honest weights (Specification 1) that approximately maximize this objective, we learn $\widehat{W}^{-\ell}$ for each fold $\ell$ separately as follows: first we estimate $F(t\,|x)$ from Model (6) by $\widehat{F}^{-\ell}(t\,|\,x)$ using only $p$-values and covariates outside of fold $\ell$. Next, identifying $\widehat{W}^{-\ell}(\cdot)$ with the function's values evaluated at the $X_i$, that is $W_i = \widehat{W}^{-\ell}(X_i), i \in I_\ell$ we solve the $|I_\ell|$-dimensional problem with optimization variables $\mathbf{w} = (w_i)_{i\in I_\ell}$:

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

This setting allows for conditional distributions $\widehat{F}^{-\ell}(t\,|\,X_i)$ that are different for tests with different covariates $X_i$. We consider estimators $\widehat{F}^{-\ell}(t\,|\,x)$ 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 data sets (Genovese et al., 2006; Strimmer, 2008b). 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, for example 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), that is 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 $\ell$ 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 $O(\log(\mathrm{m})\cdot m)$.

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 $X_i \in \mathbb{R}^p$

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

Such a beta-uniform mixture model has been considered in the setting without covariates, for example 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 $s\colon \mathcal{X} \to [0,1]$, such that the multiple testing procedure that rejects hypotheses with $P_i \le s(X_i)$ satisfies the following two properties: first, the marginal FDR, defined as $\text{mFDR}(s) := \mathbb{P}[H_i = 0 \mid P_i \le s(X_i)]$ is bounded by $\alpha$, i.e., $\text{mFDR}(s) \le \alpha$ and second, the expected number of discoveries $\sum F(s(X_i) \mid X_i)$ is large[6]. Similarly to our Bonferroni construction, we learn the threshold function $\widehat{s}^{-\ell}$ for each fold $\ell$ separately. To this end, we estimate $\widehat{F}^{-\ell}(t \mid x)$ and $\widehat{\pi}_0^{-\ell}(x)$ out of fold. Noting that $\text{mFDR}(s) \le \alpha$ is implied by $\in_i \pi_0(X_i) s(X_i) \le \alpha \in_i F(s(X_i) \mid X_i)$, we propose solving:

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

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

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

convex program if $\widehat{F}^{-\ell}(t\,|\,x)$ 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 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 $I_\ell$ 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 $\tau$-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 $\tau$-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, under 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 $\alpha$ 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

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

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, that is with categorical covariates $X_i \in [G]$. In each simulation, we generate $(P_i, H_i, X_i)$, $i = 1, \ldots, 20000$, independently, as follows:

$$\begin{aligned}
&\tilde{X}_i = \lfloor 40 \cdot (i-1)/m \rfloor, \;\; X_i = \lceil \tilde{X}_i / 40 \cdot G \rceil \\
&H_i | \tilde{X}_i \sim \text{Bernoulli}(1 - \pi_0(\tilde{X}_i)), \;\; \pi_0(\tilde{X}_i) = (0.2 + 0.8\tilde{X}_i/36) \cdot \mathbf{1}(\tilde{X}_i = 0 \bmod 4) + \mathbf{1}(\tilde{X}_i \neq 0 \bmod 4) \\
&Z_i | H_i, \tilde{X}_i \sim \mathcal{N}(H_i \cdot \mu(\tilde{X}_i), 1), \;\; \mu(\tilde{X}_i) = 2.5 - 2\tilde{X}_i/36 \\
&P_i = 1 - \Phi(Z_i), \;\; \Phi \text{ is the standard Normal CDF}
\end{aligned} \tag{12}$$

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, that is visible to the algorithm. For example, for $G = 2$, $X_i$ takes on only 2 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 **BH** method (Benjamini & Hochberg, 1995), which ignores the covariates $X_i$.
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, 2008a), 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 five folds.
6. The **IHW-Storey-Grenander** procedure: the IHW-Storey method (Theorem 2) with hypotheses randomly split into five 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 algorithm (SABHA)** by Li and Barber (2019): SABHA first estimates $\widehat{\pi}_0(\cdot)$ for each group by solving a joint convex optimization problem. Then, the $\tau$-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 $\alpha$ and SABHA at $\alpha(1 + 10\sqrt{G/m})$ (Li & Barber, 2019, Lemma 2).

Results are shown in Figure 3. Under the global null (Figure 3(a)), 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 Figure 2 for a situation where it does display a pronounced loss of FDR control.

For the simulations with signal (Figure 3(b), (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 (Figure 3(c)) for the grouped methods. The power of BH remains constant. After BH, the least powerful procedure appears to be SABHA; the suboptimality of its weighting scheme has been previously pointed out (Lei & Fithian, 2018). We also observe 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

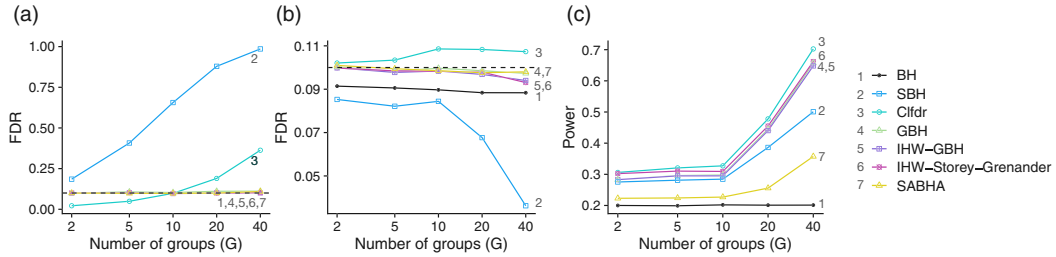


**FIGURE 3** **Grouped multiple testing simulation:** (a) False discovery rate under the global null in Model (12) (averaged over 10,000 Monte Carlo replicates) for seven methods for multiple testing with groups. (b), (c) False discovery rate and power in Model (12) (averaged over 200 Monte Carlo replicates) when there is signal (average null proportion is 0.9) for the same seven methods. The nominal $\alpha$ is equal to 0.1 throughout. [Colour figure can be viewed at wileyonlinelibrary.com]

Grenander estimator can precisely estimate the distribution in each bin. This translates into the Clfdr procedure and IHW-Storey-Grenander outperforming the other methods 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, while IHW-Storey-Grenander retains FDR control due to cross-weighting. 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 $F_{\text{alt}}(\cdot \mid X_i = x)$:

$$\begin{aligned} &\mathbb{P}^X = U[0,1]^2, \;\; \pi_0(x) = 0.98 \cdot \mathbf{1}(x_1^2 + x_2^2 \leq 1) + 0.6 \cdot \mathbf{1}(x_1^2 + x_2^2 > 1), \;\; (\mathbb{E}[\pi_0(X_i)] \approx 0.9) \\ &F_{\text{alt}}(\cdot \mid X_i = x) = \text{Beta}(\beta(x), 1), \;\; \beta(x) = 1/\max\{1.3, \overline{\beta} \cdot (\sqrt{x_1} + \sqrt{x_2})\} \end{aligned} \tag{13}$$

$\overline{\beta} \in [1, 3]$ 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 **BH** (Benjamini & Hochberg, 1995) method ignoring $X_i$.
2. The **oracle Clfdr procedure (Clfdr-oracle)** that rejects hypotheses with a small conditional local fdr, $\text{fdr}(P_i \mid X_i) := \mathbb{P}[H_i = 0 \mid X_i, P_i]$ 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 (13) 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 (13) using the EM algorithm (details in Supplement S4.2).

1. **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.
2. **IHW-Storey-BetaMix:** this is the IHW-Storey method with hypotheses split randomly into five 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 $\tau = 0.1$.
3. **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 Figure 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.

**Breaking AdaPT:** Figure 4 demonstrates that AdaPT is very powerful for multiple testing in model (13). However, under two conditions (more of which, below), AdaPT's power (but not FDR control guarantees) can be diminished, even under independence. To explain these two conditions, we first provide a summary of how AdaPT works. In iteration $j$ of AdaPT, a candidate rejection function $s_j \colon \mathcal{X} \to [0, 1]$ 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)):

$$\widehat{\mathrm{FDP}}_j = \frac{1 + |\{i \colon P_i \geq 1 - s_j(X_i)\}|}{|\{i \colon P_i \leq s_j(X_i)\}|}. \tag{14}$$

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.

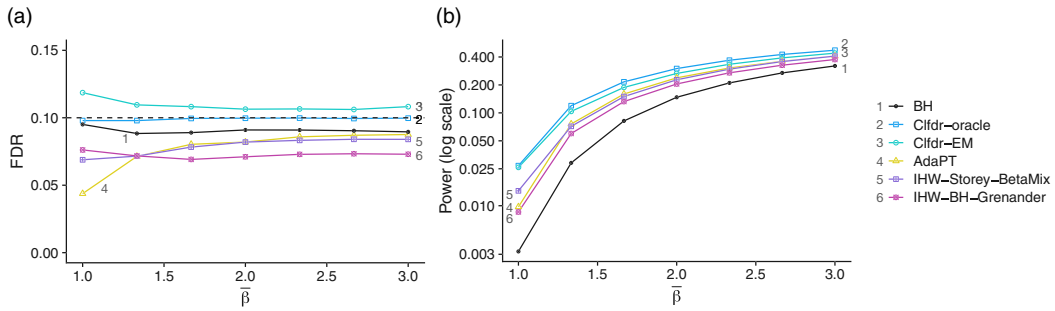


**FIGURE 4** **Simulation for multiple testing with a continuous covariate:** (a) False discovery rate in model (13) for six methods. The $x$-axis corresponds to a simulation parameter that is monotonically related to the strength of the signal for the alternatives. (b) Power in model (13) for the same six methods. The nominal $\alpha$ is equal to 0.1 throughout and results are averaged over 400 Monte Carlo replicates. [Colour figure can be viewed at wileyonlinelibrary.com]

The first complication of (14) 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 ($\overline{\beta} = 1.0$). There, AdaPT has FDR substantially below the nominal $\alpha$ and furthermore has lower power than IHW-Storey-BetaMix.

The 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 (14) 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 $P_i \,|\, (H_i = 0) \sim (1-\kappa)\, U[0,1] + \kappa\, \text{Beta}(1, 0.5)$, varying $\kappa \in [0,0.1]$ and fixed $\overline{\beta} = 2$. Our previous simulations correspond to $\kappa$=0, which yields the uniform null distribution. Figure 5(a) shows the null density as $\kappa$ varies, and panels (b),(c) show the results of the simulation. We see that as $\kappa$ increases, the FDR of AdaPT quickly drops below the nominal $\alpha$ 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 of means following Cai et al. (2019). For the $i$-th hypothesis, we observe

$$Y_{i,1}, \ldots, Y_{i,n} \sim \mathcal{N}(\mu_{Y,i}, \sigma_i^2) \quad \text{and} \quad V_{i,1}, \ldots, V_{i,n} \sim \mathcal{N}(\mu_{V,i}, \sigma_i^2) \tag{15}$$

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

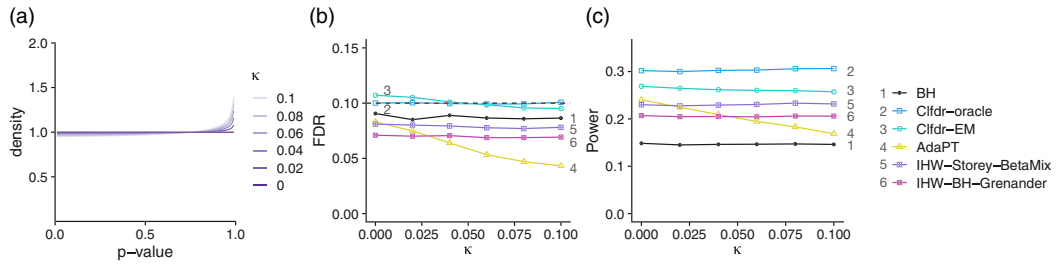


**FIGURE 5** **Simulation for multiple testing with a strictly super-uniform null distribution:** (a) Density of null $p$-values drawn from $(1-\kappa)U[0,1] + \kappa\text{Beta}(1,0.5)$ for varying $\kappa$. (b), (c) FDR control and power under same simulation setting as Figure 4, but with $\overline{\beta} = 2$ fixed and $\kappa$ varying (Figure 4 corresponds to $\kappa$=0). [Colour figure can be viewed at wileyonlinelibrary.com]

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

In addition, denote by $\hat{\mu}_i := \left(\overline{Y_i} + \overline{V_i}\right)/2$ the pooled average and let $X_i := \sqrt{2n}\hat{\mu}_i/\sigma_i$. A direct covariance calculation reveals that $\mathrm{Cov}(X_i, Z_i) = 0$ , 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 (15) with $m = 10{,}000$, $n = 50$, $\sigma_i = 1$ for all $i$. Furthermore, we vary $m_1$, the number of alternatives and let

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

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 screening null $\mu_{Y,i} = \mu_{V,i} = 0$. We compare five methods.

1. The **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 $(z,x) \mapsto \hat{s}_{\mathrm{CARS}}(z,x)$ and a threshold $\hat{t}_{\mathrm{CARS}}$ and rejects all hypotheses such that $\hat{s}_{\mathrm{CARS}}(Z_i, X_i) \leq \hat{t}_{\mathrm{CARS}}$. Asymptotically, CARS controls 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 $\tau$-censored) weighting heuristic based on CARS. We partition hypotheses randomly into five folds $I_1, \ldots, I_5$. To choose weights for $I_\ell$ we proceed as follows: first, we run CARS on the remaining four folds and get $\hat{s}^{-\ell}_{\mathrm{CARS}}(\cdot, \cdot)$ and $\hat{t}^{-\ell}_{\mathrm{CARS}}$. Then, for $i \in I_\ell$, we let $t_i$ be the smallest threshold at which $H_i$ would get rejected,

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

   Then we let $\tilde{W}_i = 2(1 - \Phi(t_i))$, $W_i = |I_\ell| \, \tilde{W}_i / \sum_{j \in I_\ell} \tilde{W}_j$ 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 Figure 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 a 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

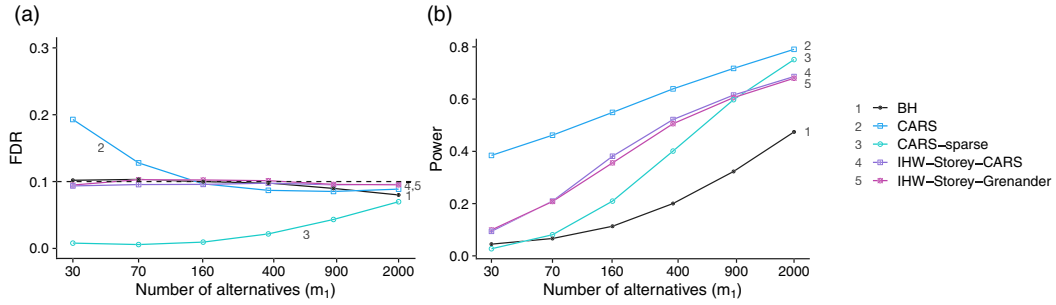


**FIGURE 6** **Simulation for simultaneous two-sample testing:** (a) False discovery rate and (b) Power in model (15) for five methods for simultaneous two-sample testing. The nominal $\alpha$ is equal to 0.1 throughout, and results were averaged over 400 Monte Carlo replicates. [Colour figure can be viewed at wileyonlinelibrary.com]

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 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 $N_1 = 645,452$ and $N_2 = 699,343$ SNPs and the H3K27ac levels at $K_1 = 12,193$ and $K_2 = 11,232$ 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 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 **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

[8]We note that computing and storing 16 billion $p$-values puts notable demands on computing infrastructure. Therefore, 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 $10^{-4}$). BH/BY 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 depends on the large $p$-values to estimate the FDR, cf. (14).

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 wBY procedure, the weights translate into thresholds for rejection: hypothesis $i$ is rejected if $P_i \leq W_i \, \widehat{t}^*_{\mathrm{IHW}}$ for some common choice of $\widehat{t}^*_{\mathrm{IHW}}$ and hypothesis-dependent $W_i$ (Panel (d)). In contrast, the BY procedure uses the same rejection threshold $\widehat{t}^*_{\mathrm{BY}}$ 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 fdr (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; 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 of IHW 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 the 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 et al. (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 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.

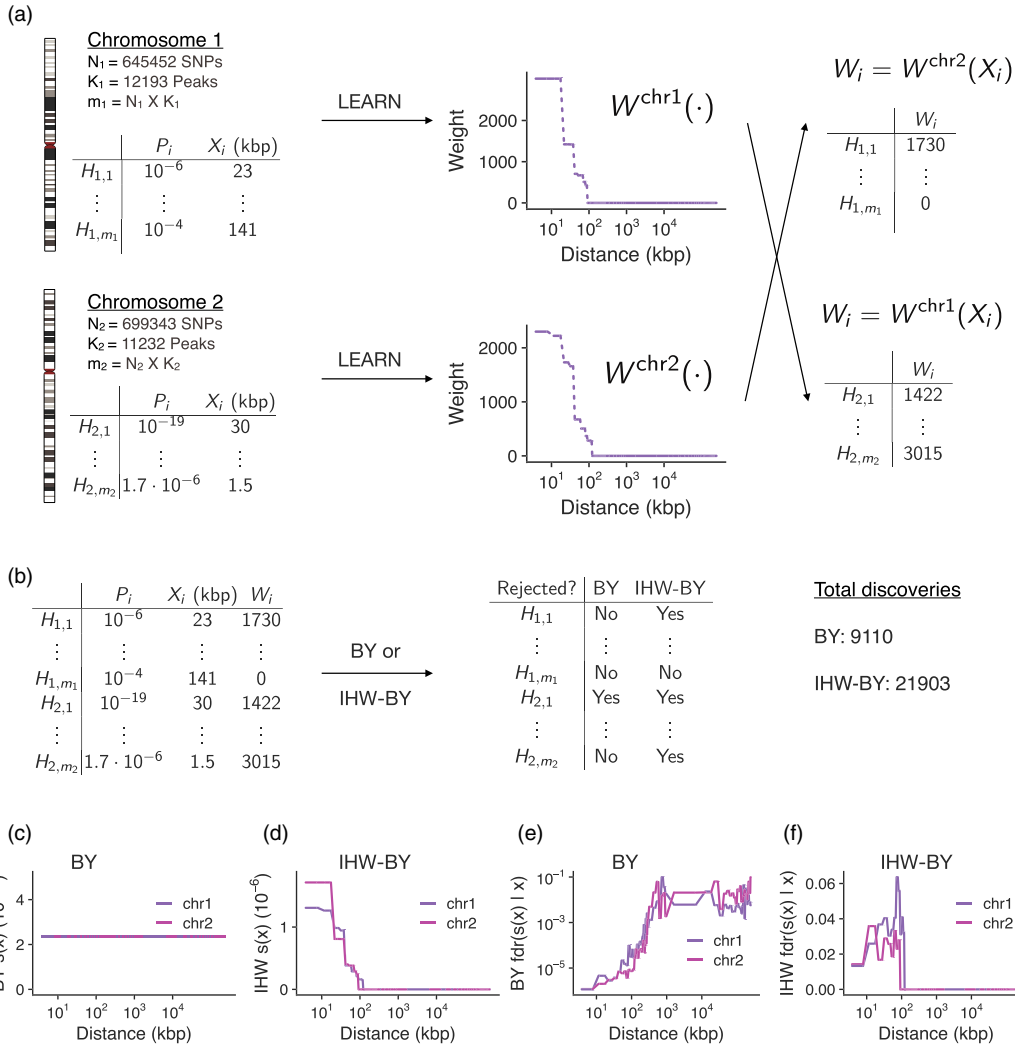


**FIGURE 7** **Biological data example revisited:** (a) Schematic representation of cross-weighting: we consider a multiple testing situation with $m = m_1 + m_2$ hypotheses that can be partitioned into two independent folds (here: two chromosomes). Besides the $p$-value $P_i$, a covariate $X_i$ is available for each hypothesis ($i = 1, \ldots m$), 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 $f(t|x)$ is estimated by binning along $X_i$ and applying the Grenander estimator within each bin. We set $f(0|x) = \infty$, so that the conditional local fdr is 0 when $s(x) = 0$. [Colour figure can be viewed at wileyonlinelibrary.com]

## 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 data set, 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 the light of non-exchangeability, the cost of a false discovery to the researcher may not be uniform, but vary across hypotheses; for example, 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

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

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 control 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 data set 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 data set 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, data set-wide dependences often signal the need for a fundamental rethink of the analysis approach.

Sometimes, data set-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 & 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; Ren & Candès, 2020; Sesia et al., 2019).

## CODE AVAILABILITY AND REPRODUCIBILITY

The study is made fully third-party reproducible, and we provide its code in Github under the link https://github.com/Huber-group-EMBL/covariate-powered-cross-weighted-multiple-testing. The Bioconductor package IHW (http://bioconductor.org/packages/IHW) provides a user-friendly implementation of IHW-BH/Storey based on the Grenander estimator.

### ACKNOWLEDGEMENTS

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, Gesthimani Roumpani, Stelios Serghiou, Michael Sklar, Youngtak Sohn, Oliver Stegle, Mark van de Wiel and Britta Velten for helpful discussions and critical comments on the manuscript. We thank Stefan Wager, an anonymous associate editor and two anonymous reviewers for feedback that motivated us to substantially improve the manuscript. Michael Sklar proposed the counterexample from Supplement S1.3. W.H. acknowledges support from the German Federal Ministry of Education and Research, Grant MOFA, under grant contract No. 031L0171A. N.I. acknowledges support from a Ric Weiland Graduate Fellowship. Open access funding enabled and organized by Projekt DEAL.

## REFERENCES


Allison, D.B.Gadbury, G.L., Heo, M., Fernández, J.R., Lee, C.-K., Prolla, T.A. et al. (2002) A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1), 1–20.

Arias-Castro, E. & Chen, S. (2017) Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1), 1983–2001.

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

Basu, P., Cai, T.T., Das, K. & Sun, W. (2018) Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523), 1172–1183.

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

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57, 289–300.

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

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

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

Boca, S.M. & Leek, J.T. (2018) A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6, e6035.

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

Bourgon, R., Gentleman, R. & Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(21), 9546–9551.

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

Cai, T.T., Sun, W. & Wang, W. (2019) 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.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. et al. (2017) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, C1–C68.

Deb, N., Saha, S., Guntuboyina, A. & Sen, B. (2021) Two-component mixture model in the presence of covariates. Journal of the American Statistical Association, pages 1–35.(https://urldefense.com/v3/__https://doi.org/10.1080/01621459.2021.1888739__;!!N11eV2iwtfs!8mLs1neu1RYPgLqn_fi-rPymhlgfSQADFWYU3z5MrJBvge5GStT-k0KzCSCxfrtS$)

Dobriban, E., Fortney, K., Kim, S.K. & Owen, A.B. (2015) Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika*, 102(4), 753–766.

Du, L. & Zhang, C. (2014) Single-index modulated multiple testing. *The Annals of Statistics*, 42(4), 30–79.

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

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

Efron, B. (2008) Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics*, 2, 197–223.

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

Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456), 1151–1160.

Ferkingstad, E., Frigessi, A, Rue, H., Thorleifsson, G. & Kong, A. (2008) Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 2, 714–735.

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

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

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

Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R. et al. (2015) Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162(5), 1051–1065.

Guo, W. & Sarkar, S. (2019) Adaptive controls of FWER and FDR under block dependence. *Journal of Statistical Planning and Inference*, 208, 13–24.

Habiger, J.D. (2017) Adaptive false discovery rate control for heterogeneous data. *Statistica Sinica*, 27, 1731–1756.

Habiger, J.D. & Peña, E.A. (2014) Compound *p*-value statistics for multiple testing procedures. *Journal of multivariate analysis*, 126, 153–166.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning: Data mining, inference, and prediction*, 2nd Edition. New York: Springer. Springer Series in Statistics, ISBN 9780387848587.

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

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

Ignatiadis, N. & Wager, S. (2019) Covariate-powered empirical Bayes estimation. In: *Advances in Neural Information Processing Systems*, pp. 9620–9632.

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

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

Korthauer, K., Kimes, P.K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M. et al. (2019) A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1), 118.

Leek, J.T. & Storey, J.D. (2008) A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48), 18718–18723.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E. et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 733–739.

Lehmann, E.L. & Romano, J.P. (2005) *Testing statistical hypotheses*. Berlin: Springer, Springer Texts in Statistics. ISBN 9780387988641.

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

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

Liang, K. & Nettleton, D. (2012) 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.

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

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

Nie, X. & Wager, S. (2020) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 09, asaa076.

Ochoa, A., Storey, J.D., Llinás, M. & Singh, M. (2015) Beyond the E-value: Stratified statistics for protein domain prediction. *PLoS Computational Biology*, 11(11), e1004509, 11.

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

Ploner, A., Calza, S., Gusnanto, A. & Pawitan, Y. (2006) Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, 22(5), 556–565.

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

Ren, Z. & Candès, E. (2020) Knockoffs with side information. *arXiv preprint arXiv:2001.07835*.

Roeder, K. & Wasserman, L. (2009) Genome-wide significance levels and weighted hypothesis testing. *Statistical Science*, 24(4), 398.

Roeder, K., Devlin, B. & Wasserman, L. (2007) Improving power in genome-wide association studies: Weights tip the scale. *Genetic Epidemiology*, 31(7), 741–747.

Romano, J.P. & Wolf, M. (2010) Balanced control of generalized error rates. *The Annals of Statistics*, 38(1), 598–633.

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

Rubin, D., Dudoit, S. & Laan, M. (2006) A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 5.

Sankaran, K. & Holmes, S. (2014) structSSI: Simultaneous and selective inference for grouped or hierarchically structured data. *Journal of Statistical Software*, 59(13), 1.

Schick, A. (1986) On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14, 1139–1151.

Scott, J.G., Kelly, R.C., Smith, M.A., Zhou, P. & Kass, R.E. (2015) False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510), 459–471.

Sesia, M., Sabatti, C. & Candès, E.J. (2019) Gene hunting with knockoffs for hidden Markov models. *Biometrika*, 106, 1–18.

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

Stegle, O., Parts, L., Durbin, R., & Winn, J. (2010) 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.

Storey, J.D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, 31, 2013–2035.

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

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

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

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

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

Sun, L., Craiu, R.V., Paterson, A.D. & Bull, S.B. (2006) Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6), 519–530.

van der Vaart, A.W. (2000) *Asymptotic statistics*. Cambridge: Cambridge University Press. Cambridge Series in Statistical and Probabilistic Mathematics. ISBN 9781107268449.

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

Wang, L. (2018) Weighted multiple testing procedure for grouped hypotheses with k-FWER control. *Computational Statistics*, 34, 1–25.

Wang, J., Zhao, Q., Hastie, T. & Owen, A.B. (2017) Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics*, 45(5), 1863–1894.

Zhang, M.J., Xia, F., Zou, J.Y. & Tse, D. (2017) NeuralFDR: Learning discovery thresholds from hypothesis features. In: *Advances in Neural Information Processing Systems*, pp. 1540–1549.

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

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

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Ignatiadis N, Huber W. Covariate powered cross-weighted multiple testing. *J R Stat Soc Series B*. 2021;83:720–751. https://doi.org/10.1111/rssb.12411