

MATHEMATISCHES INSTITUT
UNIVERSITÄT HEIDELBERG

MASTER THESIS

**Better multiple Testing: Using
multivariate co-data for hypothesis
weighting**

Daniel Fridljand

supervised by
Prof. Dr. Jan JOHANNES
Dr. Wolfgang HUBER

November 6, 2022

This master thesis was created at EMBL Heidelberg during the period 1/1/2022 to 7/1/2022.

Supervisor: Prof. Dr. Jan Johannes
Institute for Applied Mathematics
University of Heidelberg

Dr. Wolfgang Huber
Genome Biology Unit
EMBL Heidelberg

Ich erkläre hiermit ehrenwörtlich, die obenstehende Arbeit ohne Hilfe Dritter und ausschließlich unter Zuhilfenahme der von mir angegebenen Quellen angefertigt zu haben und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Ort, Datum _____ **Unterschrift** _____

Abstract

Consider a multiple testing task, where for each test we have access to its p -value and additional information represented by a univariate or multivariate covariate. The covariates may contain information on prior probabilities of null and alternative hypotheses and/or on the test's power. As per several recent proposals, the independent hypothesis weighting (IHW; Ignatiadis and Huber, 2021) framework capitalizes on these covariates for the multiple testing procedure. IHW partitions the covariate space into a finite number of bins and learns weights used to prioritize each bin a-priori based on the covariate. IHW guarantees false discovery rate (FDR) control, while increasing the proportion of correct discoveries (power) compared to unweighted methods such as the Benjamini-Hochberg procedure (BH).

Ignatiadis and Huber (2021) used per-covariate quantiles for the partition. Limitations of this are that the number of quantile combinations explodes with multiple covariates and they are inappropriate for heterogeneous covariates. Here we propose a random forest-based approach (IHW-Forest), where the leaves of the trees form a partition for the covariates. The objective function is chosen such that the splits are sensitive to the prior probability of a hypothesis being true. IHW-Forest scales well to high-dimensional covariates and can detect small regions with signal. IHW-Forest can deal with heterogeneous covariates and ignores uninformative covariates. The latter is useful in practice when the user does not know which covariates are relevant for the hypotheses under study. This extends the application of IHW by the automatic selection of the most relevant covariate. Lastly, IHW-Forest takes advantage of the p -values to construct the partition, yielding homogeneous bins and hence increasing power.

We demonstrate IHW-Forest's benefits in simulations and in a bioinformatics application. IHW's power vanishes with an increasing number of covariate dimensions, while IHW-Forest's power remains stable and well above BH and IHW. With the signal concentrated in a shrinking region, IHW-Forest outperforms BH, IHW and other competing methods in power. We apply IHW-Forest to an hQTL analysis, which looks for associations between genetic variation and histone modifications on the human chromosomes. This resulted in 16 billion tests on the first two chromosomes. We used 11 different covariates, among them the genomic distance. Due to an exponential increase of the number of per-covariate quantiles with the number of covariates, IHW is not applicable anymore, but IHW-Forest is. The updated package will be available from Bioconductor <http://www.bioconductor.org/packages/IHW>.

Zusammenfassung

Betrachte ein multiples Testproblem, bei dem wir für jeden Test Zugriff auf einen P-Wert sowie zusätzliche Information in Form einer ein- oder mehrdimensionalen Kovariate haben. Die Kovariate kann über die a-priori Wahrscheinlichkeit der Null-Hypothese und/oder die Trennschärfe des Tests informieren. Wie einige andere aktuelle Vorschläge, nutzt IHW (Ignatiadis und Huber, 2021) diese Kovariaten für das Verfahren für das multiple Testproblem. IHW zerlegt den Raum der Kovariaten in eine endliche Anzahl an Teilmengen, gewichtet jene individuell, und nutzt die Gewichte um die Hypothesen in den jeweiligen Teilmengen a-priori zu priorisieren. IHW kontrolliert die Falscherkennungsrate und erhöht den Anteil der korrekt erkannten Entdeckungen verglichen mit ungewichteten Methoden, wie Benjamini-Hochberg (BH).

Ignatiadis and Huber (2021) benutzen eine Zerlegung, die auf den Quantilen der einzelnen Kovariaten dimensionen basiert. Nachteile davon sind, dass die Anzahl der Kombinationen mit der Kovariaten dimension explodiert. Desweiteren ist die Methode ungeeignet für heterogene Kovariaten. Wir adressieren dieses Problem mit einem Random Forest (IHW-Forest), wo die Blätter der Bäume als Zerlegung für die Kovariaten dienen. Die Baumkonstruktion ist sensitiv bezüglich der a-priori Wahrscheinlichkeit einer Hypothese wahr zu sein. IHW-Forest skaliert gut zu hochdimensionalen Kovariaten und kann uninformativen Kovariaten ignorieren. Letzteres ist praktisch, wenn der/die Nutzer*in nicht weiß, welche Kovariaten für die Fragestellung relevant sind. Durch die automatische Auswahl der relevantesten Kovariaten wird die Anwendung von IHW deutlich erweitert. IHW-Forest nutzt die P-Werte um die Zerlegung zu konstruieren, wodurch Teilmengen der Zerlegung homogen werden und die Trennschärfe erhöht wird.

Wir demonstrieren die Vorteile von IHW-Forest in Simulationen und in einer Anwendung aus der Bioinformatik. IHWs Trennungsschärfe verschwindet mit steigender Anzahl an Kovariaten, während die von IHW-Forest stabil bleibt. Wenn das Signal in einer schrumpfenden Region konzentriert ist, übertrifft IHW-Forest vergleichbare Methoden. Wir wenden IHW-Forest auf eine hQTL Analyse an, welche nach Assoziationen zwischen genetischer Variation und Histonmodifikationen beim Menschen sucht. Dies führt zu 16 Milliarden Hypothesentests auf den ersten beiden Chromosomen. Wir benutzen 11 verschiedene Kovariaten, unter anderem den Abstand auf dem Genom. Wegen des exponentiellen Anstieges der Anzahl der Quantil Kombinationen und der Anzahl der Kovariaten, ist IHW nicht mehr anwendbar, aber IHW-Forest schon. Das aktualisierte Paket wird auf Bioconductor unter <http://www.bioconductor.org/packages/IHW> verfügbar gemacht.

Presentations

Some of the methods, concepts, analyses and results in this thesis have appeared previously in the following presentations:

- Workshop on adaptive designs and multiple testing procedures
Institute of Medical Biometry, Heidelberg, Germany
Jun 29, 2022
[Link 1](#), [Link 2](#)
- Statistical Data Analysis for Genome-Scale Biology summer school
Bressanone-Brixen, Italy
Jun 21, 2022
[Link](#)
- Young Statisticians Lecture Series: Modern Statistical Methods for Life Sciences
German Region of the International Biometric Society, virtual presentation
Jun 1, 2022
[Link 1](#), [Link 2](#)
- Bioinformatics and systems biology session
DAGStat Conference (German Statistical Society), University of Hamburg, Germany
Mar 29, 2022
[Link 1](#), [Link 2](#)
- Snapshots of Theory@EMBL Mini-Symposium
Virtual presentation
Jan 28, 2022
[Link](#)

Acknowledgments

First of all, I would like to thank my advisor, Dr. Wolfgang Huber for giving me the opportunity to work on this exciting topic as part of my M.Sc. thesis. I have spent the past nine months working in his lab at the European Molecular Biology Laboratory (EMBL) and he has been extremely supportive throughout the duration, taking the time to directly mentor me in scientific as well as personal matters. Many opportunities I have had and am going to have are due to him. I would also like to thank EMBL for providing such a stimulating research environment and I am honored to have been part of this amazing institute. Moreover, I would like to express my gratitude to everyone in the Huber group (past and current members), who made this time such a fun and educational experience. In particular, I would like to thank Constantin Ahlmann-Eltze, Tümay Capraz, Donnacha Fitzgerald, and Harald Vöhringer for helping me with the implementation sufficing the standards of BioConductor and the biological application in this thesis.

Next, I thank Nikolaos Ignatiadis, who together with Dr. Huber and the co-authors developed IHW. The resulting publications, software, as well as Ignatiadis' Bachelor and Master thesis form the basis for this thesis. He conceived the idea for IHW-Forest providing me with a detailed theoretical description before the project started. Furthermore, he assisted in the preparation of the material for the presentations. This whole project would not have been possible without his ingenuity and support.

I would also like to express my gratitude to Prof. Dr. Jan Johannes for co-supervising this thesis and adding mathematical aspect to the project. Indeed, Prof. Johannes has been my professor since my very first semester and I have had a total four semesters of course work during my Bachelor with him. He was the sole supervisor of my Bachelor thesis and me specialising in statistics is largely due to him. He has continuously supported me throughout my studies.

I thank Dr. Michael Love for suggestions for the biological application and providing insight on the hQTL analysis. I thank Cecile Le Sueur and Maren Hackenberg for fruitful discussions regarding the project and proposing new features for the IHW-Forest implementation. I acknowledge support from the Gerhard C Starck foundation.

Finally, I want to thank a long list of people for providing valuable feedback on different parts of drafts of this thesis: Yara Elshiaty, Erin Kim, Christian König, Eliane Kummer, Ann-Kristin Kunz, Benedikt Lauer, Maren Raus, Maximilian Siebel, Tom Stalljohann, Konstantinos Stefanidis, Henning Stein, Niklas Wünnel.

Contents

1	Introduction	9
1.1	Motivation	9
1.2	Context of this thesis	11
1.3	Contribution of this thesis	11
1.4	Code availability and reproducibility	12
1.5	Outline	12
2	Unweighted Benjamini-Hochberg	13
2.1	Single Hypothesis Testing	13
2.2	Multiple Hypothesis Testing	14
2.3	Two-group model	16
2.4	Benjamini-Hochberg	17
3	Weighted Benjamini-Hochberg	19
3.1	Setting and assumptions	19
3.2	Powerful weighting for BH	20
3.3	Breaking naive weighting	23
4	Crossweighted Benjamini-Hochberg	28
4.1	Definition of IHW	28
4.2	Empirical formulation of the FDR	29
4.3	Empirical formulation of IHW	38
4.4	Discussion	43
5	IHW-Forest	45
5.1	Optimization program	45
5.2	Stratification	46
5.3	Quantile slicing	47
5.4	Random forest	48
5.5	Alternative motivation	50
6	Application and simulation	53
6.1	Simulation	53
6.2	Application example: biological high-throughput data	55

A Proofs	61
B Full IHW algorithms	64

Chapter 1

Introduction

1.1 Motivation

In many scientific fields, ranging from genomics [DSB03; GS14], neuroimaging [GLN02], proteomics [CN08], psychology [Sha95; KCH02] to economics [BS12], one regularly needs to test thousands or millions of hypotheses simultaneously [LF18]. Particularly in biology, the availability of high-dimensional datasets has revolutionized the field. Microarrays [Sch+95], next-generation sequencing [SJ08], and high-throughput metabolomics [LNH11] enable researchers to simultaneously examine the association between a multitude of biomarkers and a specific exposure or outcome at the same time [BL18]. One can project that the availability of large data sets will continuously increase in the future.

These problems have the multiple testing task in common: a list of potential hypotheses is to be scanned for “significant hits” declared as discoveries. Large-scale multiple testing task can lead to a well-known problem: if m hypotheses are tested with the same significance threshold of, e.g. $\alpha = 0.05$, there is a high risk of making many false discoveries [LB17]. In the extreme case, where all m hypotheses are true nulls, we expect $\alpha \cdot m$ false discoveries.

This requires us to consider multiple testing procedures which have error control while maximizing the number of discoveries. Many error measures for multiple testing have been developed. For high-throughput biological experiments where investigators are willing to accept some false discoveries in exchange for a larger number of promising hits, the false discovery rate (FDR), i.e. the expected proportion of false discoveries (number of type-I errors), has found increasing popularity as error measure over the past decades [CRS10; Kor+19].

The first method devised to control the FDR, and possibly the most famous multiple-testing procedure of the modern era [LF18], is the Benjamini–Hochberg (BH) procedure [BH95]. The paper by Benjamini and Hochberg is one of the 100 most cited papers in all of science [VMN14]. BH records a p -value for each hypothesis and works solely with this list of p -values to return a list of discoveries.

BH implicitly assumes that data associated with each test are collected from identical experimental conditions, and hence hypotheses are exchangeable [CS09]. However, in many applications, the individual tests are known to differ in statistical properties such

as sample size, true effect size, signal-to-noise ratio, or prior probability of being false [Ign+16], and therefore hypotheses are intrinsically heterogeneous. This can lead to certain hypotheses being a-priori more powerful i.e. more likely to be true discoveries.

While ignoring this heterogeneity between hypotheses may be justified in instances where power is high anyway, in many cases, the costs of the underlying experiments are substantial and rise with the sample size [IH21]. In such cases, having a statistical method which increases the number of discoveries compared to the baseline BH analysis, while retaining FDR control, can decide over success or failure of the studies - at no additional cost and by purely computational means [IH21].

Motivated by this need for power, a new class of methods has been recently developed to exploit this variability across tests [IH21; BL18; CS09; LF18; Sco+15; CRS21; LB19a]. The set-up for those methods is that for each hypothesis H_i we have access to not only the standard p -value $P_i \in [0, 1]$, but also contextual information $X_i \in \mathcal{X} \subset \mathbb{R}^d$, referred to as an “informative covariate” [LF18]. Unlike P_i , X_i carries only indirect information about the hypotheses [LF18]. Its purpose is to distinguish the alternative from the null [LF18] and it can support the multiple testing procedures [Kor+19]. We shall see that a covariate can be used to prioritize individual or groups of tests [IH21; Kor+19]. If the covariate is informative and with multiple testing methods taking advantage of that, the overall power of the experiment can be increased substantially [IH21; Kor+19]. A restriction fundamental to all of those methods is that the covariate must be independent of the p -value under the null to guarantee FDR control.

Such covariates in fact exist in many, if not all, important multiple testing applications. For example, in a two-sample t-test the pooled variance can serve as such a covariate [BGH10; Ign+16; LF18]. Or, in a genome-wide association study, where samples are pooled across studies, the sample sizes at each loci can indicate a differing signal-to-noise ratio [Kor+19; BL18]. Lei and Fithian [LF18] used quantitative mass spectrometry [DG12] to study differences in protein abundance in yeast cells between different treatments. Here, the logarithm of the total number of peptides detected across all samples for each gene served as the covariate (see also [Ign+16]). In a similar vein, Ignatiadis and Huber [IH21] used the genomic distance between loci of genomic variation and peaks of histone modification on the genome to test for associated pairs as part of a histone quantitative trait loci (hQTL) analysis. This covariate is motivated by biological domain knowledge: associations across a short distance are a-priori more likely and empirically more frequent [IH21]. See section 6.2 for more details of the latter example.

For multiple testing tasks in presence of an informative covariate, Ignatiadis and Huber and other authors developed IHW (Independent Hypothesis Weighting) [Ign+16; IH21], which has been well received by the biological community. In a series of simulations and case studies of biological applications, Korthauer et al. [Kor+19] found IHW to be competitive and in some instances superior to other multiple testing methods which aim to increase power by including additional side information. More specifically, IHW reached the highest evaluation in the categories FDR control, applicability, consistency, and usability (see Fig. 6 in [Kor+19]). Apart from being a practical tool for applied

statisticians, the theory of IHW has sparked a lot of innovation and new work in the multiple testing field [Dur19; DR20].

1.2 Context of this thesis

Increasing the power in multiple testing by considering rejection regions which depend not only on the p -values but also on covariates has been the goal of Nikolaos Ignatiadis' bachelor thesis [Ign15]. However, the procedure developed in [Ign15] failed to control type-I error (FDR) and was computationally not scalable to multiple testing tasks with many hypotheses. Subsequently, the aim was to modify the procedure proposed in [Ign15] to establish good FDR properties and to make it computationally scalable to the requirements of high-throughput biology.

Three empirically promising heuristics were proposed to address those gaps: cross-weighting, convex relaxations and regularization of the weights. This salvaged FDR control and resulted in the novel method IHW, published in an article titled "Data-driven hypothesis weighting increases detection power in genome-scale multiple testing" [Ign+16]. The publication was targeted at computational biologists and demonstrated that IHW is a very practical and powerful approach for conducting multiple testing by showcasing its use for many relevant data-sets from high-throughput biology, as well as its strong performance in multiple numerical simulation.

Moreover, an open-source implementation of IHW was made available as a package in R/Bioconductor [Hub+15] (<http://www.bioconductor.org/packages/IHW>). Bioconductor is a software-development project covering a range of bioinformatic and statistical applications [Hub+15]. It has become a standard tool for researchers working in high-throughput data in genomics and molecular biology.

Subsequently, mathematical details for IHW were assembled in the technical supplement of Ignatiadis et al. [Ign+16] and Nikolaos Ignatiadis' master thesis [Ign16]. A plethora of mathematical results for IHW appeared in Ignatiadis and Huber [IH21]. Ignatiadis and Huber highlighted the importance of cross-weighting as compared to the other heuristics. Formally, finite-sample FDR control for IHW was established.

1.3 Contribution of this thesis

In this thesis, we revisit the asymptotic behaviour of IHW shown in [IH21]. We enrich the proofs with rigorous details not possible in a journal publication. Where ever possible and appropriate, our proofs can be followed top to bottom without the need to consult additional material and without gaps left for the reader.

A weakness ascribed to the current version of IHW is its performance with multivariate covariates. This is despite real interest from the biological community (feature request) and has been pointed out as major draw-back by competing methods (see [LF18]). However, IHW is a fairly general theoretical framework for multiple testing procedures without compromising FDR control. We fully embrace IHW's flexibility by marrying supervised

learning techniques to multiple testing procedures for a much richer, highdimensional setting.

Based on a random forest approach, we present a new version of IHW called IHW-Forest. IHW-Forest easily scales with more covariates, enabling practitioners to capture interactions from the rich resources of prior information available. We demonstrate IHW-Forest’s benefits in selected simulations and a bioinformatic application.

1.4 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/FridlDa/iHW-forest-paper>. To demonstrate IHW-Forest, we forked the IHW package and integrated the stratification method (<https://github.com/nignatiadis/IHW/pull/7>). The updated package will eventually be made available in the Bioconductor package IHW (<http://bioconductor.org/packages/IHW>) in a future release. The package is well documented and well tested.

The syntax for the new IHW version is userfriendly and extends the original IHW seamlessly. To run IHW with p -values “Ps”, covariates “Xs”, target FDR level “alpha” and the quantile stratification, we run

```
ihw_quantile <- ihw(Ps, Xs, alpha)
```

To run IHW-Forest, we just need to specify an additional argument.

```
ihw_forest <- ihw(Ps, Xs, alpha, stratification_method = "forest")
```

Several tuning parameters for the forest construction are available.

1.5 Outline

This thesis is organized as follows: First, we formally introduce the multiple testing task with type-I error control and the classic BH procedure in chapter 2. In chapter 3, we derive a natural extension of classic BH capturing heterogeneity between tests based on covariates and arrive at the procedure suggested in [Ign15]. Based on proofs in Ignatiadis and Huber [IH21], we show the shortcomings of the procedure suggested in [Ign15] in terms of FDR control. In chapter 4, we define full IHW and show asymptotic FDR control of IHW as presented in [IH21]. In chapter 5, we present IHW-Forest as extension of IHW. Chapter 6 concludes with a benchmark for IHW-Forest.

Chapter 2

Unweighted Benjamini-Hochberg

We begin with a quick overview of relevant concepts from the classical multiple testing literature. As this thesis heavily builds on previous work by Ignatiadis and Huber, we follow a very similar path as the introduction in [Ign15] and [Ign16].

2.1 Single Hypothesis Testing

Before we consider the multiple testing scenario, we first formally revisit the classic single hypothesis paradigm [LRC05]. In this paradigm we start with a measurable space $(\mathcal{Y}, \mathcal{Y})$ and a random variable $Y \in (\mathcal{Y}, \mathcal{Y})$ which is distributed according to an unknown probability distribution $Y \sim \mathbb{P}^Y$. The aim is to make inferences about $\mathbb{P}^Y$ based on a realization $y_1, \dots, y_n$ of this random variable Y . In particular, in hypothesis testing, we formulate null hypothesis $\mathcal{H}_0$ for $\mathbb{P}^Y$, while we refer to everything else as alternative hypothesis $\mathcal{H}_1$. Commonly, the null hypothesis corresponds to $\mathbb{P}^Y \in (\mathbb{P}_\theta)_{\theta \in \Omega}$ with $(\mathbb{P}_\theta)_{\theta \in \Omega}$ being a family of probability measures.

The role of $\mathcal{H}_0$ and $\mathcal{H}_1$ is asymmetric in hypothesis testing: based on our prior knowledge, we lean towards conservatively accepting the null hypothesis $\mathcal{H}_0$, unless there is strong enough evidence for the alternative $\mathcal{H}_1$. Or maybe the alternative $\mathcal{H}_1$ is rare, interesting, and/or worthwhile studying further, whereas the null $\mathcal{H}_0$ is a “catch-all” class for all the uninteresting rest [HH18].

A hypothesis test is now a measurable function: $\delta : (\mathcal{Y}, \mathcal{Y}) \rightarrow \{0, 1\}$. If $\delta = 1$, we reject the null hypothesis $\mathcal{H}_0$, otherwise we do not. We refer to a rejected hypothesis as a “discovery”. As motivated above, while we assume $\mathcal{H}_0$ to be true a-priori, we are more interested in discoveries.

Notice that this decision can lead to different types of error. In one case, the null hypothesis is rejected, despite it being true (i.e. $\mathbb{P}^Y \in (\mathbb{P}_\theta)_{\theta \in \Omega}$). This is called a type-I error. In contrast, a type-II error occurs when $\mathbb{P}^Y \notin (\mathbb{P}_\theta)_{\theta \in \Omega}$, but the null hypothesis is not rejected. The goal of minimizing these two errors is unfortunately contradictory. For this reason, the statistician proceeds as follows: (S)he decides on a significance level $\alpha \in (0, 1)$ and then constructs a hypothesis test δ , such that it controls the type-I error:

$$\sup_{\theta \in \Omega} \mathbb{P}_{\theta} [\delta(Y) = 1] \leq \alpha.$$

Let $Y_1, \dots, Y_n \sim \mathbb{P}^Y$ and let $T : (\mathcal{Y}, \mathcal{Z})^n \rightarrow ([0, 1], \mathcal{B}_{[0,1]})$ be measurable. Then the construction usually takes the form $\delta = \mathbb{1}(P \leq \alpha)$ with $P := T(Y_1, \dots, Y_n)$ and $P \sim \text{U}[0, 1]$ under the null $\mathcal{H}_0$. We call P p -value and the interpretation of a p -value is the following: It is the probability of observing the data $Y_1, \dots, Y_n$ if the null $\mathcal{H}_0$ is indeed true. It describes the smallest significance level at which we would reject the null hypothesis. p -values also provide a standardized way (i.e. with values in $[0, 1]$) of expressing how strong the evidence against the null hypothesis is, and they will also be the starting point for many multiple testing procedures. If T is a sufficient statistic (often: complete sufficient statistic), classical theory for single hypothesis testing [LRC05] shows that δ is a optimal testing procedure. All information apart from the test statistic T , can be discarded without any loss of power.

For multiple testing we will see that information irrelevant for single hypothesis testing can be leveraged to increase the power of the resulting multiple testing procedure. The reason is that a single test statistic for each hypothesis cannot capture how the hypotheses relate to each other.

2.2 Multiple Hypothesis Testing

In the single testing setting we test a single null hypothesis $\mathcal{H}_0$ against its alternative $\mathcal{H}_1$. Extending this to the classical multiple testing problem, now let $\mathcal{H} = \{1, \dots, m\}$ be an index set, with each index corresponding to a test. From these, m_0 tests have indices in $\mathcal{H}_0$, i.e. $|\mathcal{H}_0| = m_0$ are true nulls, while $m - m_0$ tests with indices in $\mathcal{H}_1$ are alternatives. Thus $\mathcal{H}_0 \dot{\cup} \mathcal{H}_1 = \mathcal{H}$. In the following, we denote the hypotheses with $H_1, \dots, H_m \in \{0, 1\}$ and will use the short-hand notation $i \in \mathcal{H}_0 \Leftrightarrow H_i = 0$ and $i \in \mathcal{H}_1 \Leftrightarrow H_i = 1$. We will model $H_1, \dots, H_m$ as random variables.

Having set-up the notation we can now formally define what a multiple testing procedure is: It is just a decision rule $\Delta = (\Delta_i)_{i \in \mathcal{H}}$ depending on the data and mapping it to $\{0, 1\}^m$ which encodes which hypotheses are rejected. Similar to the single hypothesis case, we let $\Delta_i = 1$ denote the rejection of the i -th null hypothesis, while $\Delta_i = 0$ denotes the opposite decision.

We were vague on purpose in declaring the “data” Δ depends on, because this will change through the course of thesis. To be informative about the true status of each hypothesis, the “data” will be correlated to the ground truth $H_1, \dots, H_m$. However, the “data” will never be measurable in $(H_1, \dots, H_m)$, since then the multiple testing would be obsolete. We can already say that the “data” will in particular encompass realizations of the p -values $p_1, \dots, p_m \in [0, 1]$ which are realizations of the random variables $P_1, \dots, P_m$.

Since the “data” is non-deterministic, Δ itself is associated with a random variable taking values in $\{0, 1\}^m$ and we will implicitly drop the dependence on the “data”. In multiple testing, one is usually interested in the total number of type-I errors (false discoveries) V , as well as the total number of discoveries/rejections R . These are defined as

follows:

$$V(\Delta, H) = \sum_{i \in \mathcal{H}_0} \Delta_i = \sum_{i \in \mathcal{H}} \Delta_i (1 - H_i),$$

$$R(\Delta, H) = \sum_{i \in \mathcal{H}} \Delta_i.$$

V and R take values in $\{1, \dots, m\}$. All the possible outcomes of multiple testing are summarized in Table 2.1.

In order to generalize the type-I error rate for single tests to the multiple testing scenario, a variety of error measures have been proposed. An error measure which has gained popularity over the past decades is the false discovery rate (FDR):

Definition 1. Let Δ be a multiple testing procedure taking values in $\{0, 1\}^m$. The false discovery proportion is defined as

$$\text{FDP}(\Delta, H) := \frac{V(\Delta, H)}{R(\Delta, H) \vee 1},$$

and the false discovery rate as

$$\text{FDR}(\Delta, H) := \mathbb{E}[\text{FDP}(\Delta, H)].$$

When it is unambiguous which multiple testing function Δ and which H is meant, we will drop the dependence and just write FDR instead. In high-throughput biology, experimenters are willing to accept some false discoveries in exchange for more overall discoveries [Kor+19]. Consequently, the FDR has become the main criterion of interest in that field.

Definition 2. Let $\alpha \in [0, 1]$. Consider a sequence of multiple testing functions $(\Delta^m)_{m \in \mathbb{N}}$ and hypotheses $(H^m)_{m \in \mathbb{N}}$ with $H^m \in \{0, 1\}^m$ for each $m \in \mathbb{N}$. We say $(\Delta^m)_{m \in \mathbb{N}}$ asymptotically controls the FDR at level α , if it holds

$$\limsup_{m \rightarrow \infty} \text{FDR}(\Delta^m, H^m) \leq \alpha.$$

We have defined the FDR as measure of error we aim to control beneath a certain threshold α . In contrary, we are interested in discovering as many true alternatives as possible. This is captured by definition 3.

Definition 3. Let Δ be a multiple testing procedure. Then the power of Δ is defined as

$$\text{Power}(\Delta, H) := \mathbb{E} \left[\frac{R(\Delta, H) - V(\Delta, H)}{m_1 \vee 1} \right].$$

Table 2.1: Outcomes of a multiple testing procedure applied to m hypotheses.

	Not-rejected hypotheses	Rejected hypotheses	Total
True nulls	U	V	m_0
False nulls	T	S	m_1
	$m - R$	R	m

2.3 Two-group model

As alluded to above, the p -values $P_1, \dots, P_m$ are informative about the hypotheses $H_1, \dots, H_m$. We now specify the typical dependency structure between them with Bayesian model:

$$\begin{aligned} P_i \mid H_i = 0 &\stackrel{\text{iid.}}{\sim} \text{U}[0, 1], \\ P_i \mid H_i = 1 &\stackrel{\text{independent}}{\sim} F_{\text{alt},i}. \end{aligned}$$

Here $F_{\text{alt},i}$ is a distribution with support on $[0, 1]$ which is stochastically smaller than uniform, i.e. $F_{\text{alt},i}(t) \geq t$. We embrace a Bayesian perspective and further assume a prior for the random indicator H_i with $\mathbb{P}[H_i = 0] = \pi_{0,i}$. As a consequence, our model yields that the i -th p -value comes from the a mixture:

$$\begin{aligned} H_i &\stackrel{\text{independent}}{\sim} \text{Bernoulli}(1 - \pi_{0,i}), \\ P_i \mid H_i = 0 &\stackrel{\text{iid.}}{\sim} \text{U}[0, 1], \\ P_i \mid H_i = 1 &\stackrel{\text{independent}}{\sim} F_{\text{alt},i}. \end{aligned} \tag{2.1}$$

Note that model (2.1) is very general: each hypothesis has an individual prior probability $\pi_{0,i}$ and alternative distribution $F_{\text{alt},i}$. Consequently, inference on model (2.1) is not tractable. To overcome this difficulty, the literature presents two ways forward: completely drop the heterogeneity between the hypotheses or find an approximative statistical model which is tractable, but captures heterogeneity. We discuss both approaches and resulting multiple testing procedures extensively.

The former approach is frequently encountered in the literature. It posits that all hypotheses have the same statistical properties, i.e. the hypotheses are “exchangeable”. Note that this is a very restrictive assumption. It leads to the celebrated “two-groups” model:

$$\begin{aligned} H_i &\stackrel{\text{iid.}}{\sim} \text{Bernoulli}(1 - \pi_0), \\ P_i \mid H_i = 0 &\stackrel{\text{iid.}}{\sim} \text{U}[0, 1], \\ P_i \mid H_i = 1 &\stackrel{\text{iid.}}{\sim} F_{\text{alt}}. \end{aligned} \tag{2.2}$$

Ignatiadis and Huber [IH21] find a compromise between the general (2.1) and the restrictive (2.2) by assuming that the differences in $\pi_{0,i}$ and $F_{\text{alt},i}$ across hypotheses can be explained by an observed random covariate $X \in \mathcal{X}$, where $\mathcal{X}$ is a generic space. In particular, Ignatiadis and Huber [IH21] model the prior probability π_0 as a function $\mathcal{X} \rightarrow [0, 1]$ arriving at the “conditional two group model” [Fer+08; LF18; IH21]:

$$\begin{aligned} X_i &\stackrel{\text{iid.}}{\sim} \mathbb{P}^X, \\ H_i \mid X_i &\stackrel{\text{independent}}{\sim} \text{Bernoulli}(1 - \pi_0(X_i)), \\ P_i \mid H_i = 0, X_i &\stackrel{\text{iid.}}{\sim} \text{U}[0, 1], \\ P_i \mid H_i = 1, X_i &\stackrel{\text{independent}}{\sim} F_{\text{alt}|X_i}. \end{aligned} \tag{2.3}$$

This extension of the two-groups model (2.2) serves as approximation of model (2.3), giving us more flexibility in our modelling via the additional variable X_i . While we do not have access to $(H_i)_{i \in [m]}$, we observe the pairs $((P_i, X_i))_{i \in [m]}$. Under equation (2.3), the covariate X_i is independent of the hypothesis H_i under the null hypothesis and informative under the alternative. The conditional independence of X_i and P_i under the null is the more critical component in practice. However, we can only expect a power increase, if $F_{\text{alt}|X_i=x}$ and $\pi_0(x)$ vary by x .

The covariates $(X_i)_{i \in [m]}$ are to be interpreted as side-information for each hypothesis. It is the starting point for developing a new multiple testing procedure for data-driven hypothesis weighting. Later we will model weights as a function of X_i .

2.4 Benjamini-Hochberg

One of the reasons the FDR has seen such wide-spread use, is the existence of a very simple multiple-testing procedure which can control the FDR at a prespecified level α . This procedure, named after Benjamini and Hochberg (BH; [BH95]) works as given in algorithm 1.

Algorithm 1: The Benjamini-Hochberg method (original formulation)

Input: A target level $\alpha \in (0, 1)$ and a vector of p -values $P_1, \dots, P_m$
 1 Let $P_{(1)}, \dots, P_{(m)}$ be the order statistics of $P_1, \dots, P_m$ and let $P_{(0)} := 0$
 2 Let $\hat{k}^* = \max \{k \mid P_{(k)} \leq \frac{k\alpha}{m}, k \geq 0\}$
Output: Decision $\Delta = (\mathbb{1}(P_i \leq P_{(\hat{k}^*)}))_{i \in \{1, \dots, m\}} \in \{0, 1\}^m$, where $\Delta_i = 1$ is a rejected hypothesis

From a geometric point of view, this corresponds to plotting the points $((k, P_{(k)}))_{k \in \{0, \dots, m\}}$ and the line $((t, \frac{\alpha}{m} \cdot t))_{t \in [1, m]}$. Both graphs start in the origin $(0, 0)$ and are monotonically increasing, but the former ends in $(m, \max_{k \in \{1, \dots, m\}} P_k)$ and the latter in (m, α) . We then reject all points on the left up to and including the last point that is below the line, i.e. declare them discoveries.

Storey et al. [Sto02] showed an alternative form of the BH algorithm, in terms of the ECDF $\hat{F}(t) = 1/m \sum_{i=1}^m \mathbb{1}(P_i \leq t)$, given in algorithm 2. The formulation in (2) sheds a new light on BH: BH chooses the rejection threshold t in the most greedy way, so that $\widehat{\text{Fdr}}(t) \leq \alpha$ and then rejects all hypotheses with p -value below it. One might ask oneself, if

Algorithm 2: The Benjamini-Hochberg method

Input: A target level $\alpha \in (0, 1)$ and a vector of p -values $P_1, \dots, P_m$
 1 Let $\hat{t}^* = \sup \left\{ t \mid \widehat{\text{Fdr}}(t) = \frac{t}{\hat{F}(t) \vee \frac{1}{m}} \leq \alpha, t \in [0, 1] \right\}$
Output: Decision $\Delta = (\mathbb{1}(P_i \leq \hat{t}^*))_{i \in \{1, \dots, m\}} \in \{0, 1\}^m$, where $\Delta_i = 1$ is a rejected hypothesis

this greedy choice of t leads to a valid multiple testing procedure. Theorem 1 shows that this is indeed the case.

Theorem 1 (BH controls FDR). *Let $(P_i)_{i \in \{1, \dots, m\}}$ come from the two-groups model (2.2). Then the BH method 1 controls the FDR at the pre-specified level $\alpha \in (0, 1)$:*

$$\text{FDR}_{BH} \leq \frac{m_0}{m} \alpha \leq \alpha.$$

Proof. See any of [BH95; BY01; STS04; BHJ15]. □

In the original publication [BH95], the BH procedure was shown to control the FDR for independent p-values. Later this result was generalized to diverse dependence structures among the tests [BY01].

Chapter 3

Weighted Benjamini-Hochberg

In chapter 2 we introduced the multiple testing task. The data was modelled with the two group model 2.1. We introduced the celebrated BH procedure, which is based on a simplified model with the strong assumption of exchangeability 2.2 of hypotheses. Here we instead want to work with the more general conditional two-group model 2.3 and derive the weighted BH procedure. We will motivate in section 3.2 that without oracle knowledge, powerful weights will be data-driven. This will naturally lead to the most straight-forward, naive weighted BH, where you simply plug-in all available data to learn the weights. However, we will show in 3.3 that this naive approach fails to control the FDR. We will deal with this problem in chapter 4.

3.1 Setting and assumptions

For hypotheses testing it is custom to compress the data into a single quantity, the p -value P_i (see section 2.1). The decision rule then takes form “Reject hypothesis $i \iff P_i \leq \hat{t}^*$ ” with a uniform rejection threshold $\hat{t}^*$. In particular, all hypotheses are treated equally and the underlying assumption is that they are exchangeable. However, in the presence of heterogeneity between hypotheses, this leads to sub-optimal decision rules. Recall the conditional two group model 2.3:

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

We want to derive a multiple testing procedure for this scenario, where we have access to the pair (P_i, X_i) for each hypothesis instead of only P_i . The side information X_i is meant to reflect the heterogeneity among the hypotheses. For examples of such covariates and a discussion, see the introduction 1.1. Marginalizing equation (2.3) over H_i , we see that:

$$P_i \mid X_i = x \stackrel{\text{iid.}}{\sim} F(t|x) := \pi_0(x)t + (1 - \pi_0(x))F_{alt|X_i=x}(t). \quad (3.1)$$

We will need following assumptions on F which are adapted from assumption 3 in the supplement of [IH21]:

Assumption 1. *Regularity of conditional two-groups model:*

The conditional two-groups model from equation 2.3 satisfies:

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

$$|F_{alt}(t|X_i = x) - F_{alt}(\tilde{t}|X_i = x)| \leq L(x) |t - \tilde{t}| \text{ for all } t, \tilde{t} \in [0, 1], x \in \mathcal{X}$$

$$L(\cdot) \text{ satisfies } \int L^2(x) d\mathbb{P}^X(x) < \infty.$$

(b) $F_{alt}(t|X_i = x)$ is strictly concave in t .

(c) There exists $t' \in (0, 1]$ such that $\frac{t'}{F(t'|X_i=x)} \leq \alpha'$ for an $\alpha' < \alpha/2$ and for all $x \in \mathcal{X}$.

(d) $F_{alt}(0|X_i = x) = 0$ and $F_{alt}(1|X_i = x) = 1$ for all x .

Part (a) bounds the variation of $F_{alt}(\cdot|X_i = x)$ by $L(x)$, where L is square integrable. To assume that the distribution of p -values is concave in Part (b) is mild and reasonable for real data. The assumption is common in multiple testing and is required in Lemma 1 and Theorem 2 of Genovese et al. [GRW06] or in Strimmer [Str08]. Part (c) is also a common assumption made for FDR asymptotics without covariates (e.g., it is required in Theorem 4 of [STS04]). It is, however, more restrictive than Parts (a,b); for instance it excludes the global null, i.e. $\pi_0(x) \equiv 1$, as:

$$\frac{t'}{F(t'|X_i = x)} = \frac{t'}{t'} = 1 > \frac{\alpha}{2}.$$

Part (d) implicitly follows from $P_i \in [0, 1]$ and is not included in [IH21], but we want to make it explicit here.

3.2 Powerful weighting for BH

The BH procedure records a p -value P_i for each hypothesis and treats them otherwise as exchangeable. Many multiple testing methods operate on just the list of p -values, also for other error measures than the FDR. However, we would like to incorporate the heterogeneity described in (2.3) in our testing procedure. Weighted BH [GRW06] is a flexible framework for this and is given in algorithm 3.

The weights $W_1, \dots, W_m$ encode prior information and enable us to differentially prioritize the hypotheses a-priori. We can think of $W_1, \dots, W_m$ as a weight budget so we do not waste power on less promising hypotheses. The weights $W_1, \dots, W_m$ will always be non-negative numbers. In this thesis we will only work with weights averaging to 1, i.e. satisfying $\sum_{i=1}^m W_i = m$. Alternatively, Ignatiadis and Huber [IH21] also considered weights where $\sum_{i=1}^m W_i \neq m$ which had the advantage of yielding less conservative multiple testing procedures. For uniform weights, i.e. $W_1 = \dots = W_m = 1$, the weighted BH

Algorithm 3: The weighted Benjamini-Hochberg method (wBH)**Input:**

weights $W_1, \dots, W_m \geq 0$ with $\sum_{i=1}^m W_i > 0$

p -values $P_1, \dots, P_m \in [0, 1]$

a target level $\alpha \in (0, 1)$

-
- 1 Let $\hat{t}^* = \sup \left\{ t \mid \frac{\sum_{i=1}^m W_i t}{\sum_{i=1}^m \mathbb{1}(P_i \leq W_i t)} \leq \alpha, t \in [0, 1] \right\}$
 - 2 Reject all hypotheses with $P_i \leq W_i \hat{t}^*$
-

procedure 3 is equivalent to the original BH procedure in algorithm 1. If $W_i > 1$, the i -th hypothesis is easier to reject. Since we want to maximize the number of rejections, we want to choose large weights for promising hypotheses. Though the theoretical framework presented in this work is applicable to a variety of weighting schemes, we want to derive a default which works well across a range of settings.

The wBH procedure in algorithm 3 can be cast as follows: Rather than returning a uniform threshold $t \in [0, 1]$ and rejecting hypotheses with $P_i \leq t$, wBH has individual rejection thresholds for each hypothesis. We will model this as function of the covariate $s : \mathcal{X} \rightarrow [0, 1]$ and our rejection rule has the form $P_i \leq s(X_i)$. In practice, we will fix a subset $\mathcal{S}$ of all functions from $\mathcal{X}$ to $[0, 1]$ and only consider $s \in \mathcal{S}$. We want s such that the resulting multiple testing procedure controls the type-I error and yields a large number of discoveries. As a proxy for the type-I error, we want the marginal FDR, defined as $\text{mFDR} := \mathbb{P}[H_i = 0 | H_i \text{ rejected}]$ to be bound by α . The rationale behind this is that the FDR and the mFDR are equivalent in the asymptotics. Under mild conditions we have $\text{mFDR} = \text{FDR} + O(m^{-1/2})$ [GW02], where m is the total number of tests.

We set out to find such a function s ultimately leading to a weight function by again considering the conditional two-groups model (2.3). The $\text{mFDR}(s)$ parametrized by the threshold function s can be generalized as follows:

$$\begin{aligned}
 \text{mFDR}(s) &= \mathbb{P}[H_i = 0 \mid H_i \text{ rejected}] \\
 &= \mathbb{P}[H_i \text{ rejected} \mid H_i = 0] \frac{\mathbb{P}[H_i = 0]}{\mathbb{P}[H_i \text{ rejected}]} \\
 &= \frac{\mathbb{P}[P_i \leq s(X_i) \mid H_i = 0] \mathbb{P}[H_i = 0]}{\mathbb{P}[H_i \text{ rejected}]} \\
 &= \frac{\int_{\mathcal{X}} \mathbb{P}[H_i = 0] \mathbb{P}[P_i \leq s(x) \mid H_i = 0] d\mathbb{P}^X}{\int_{\mathcal{X}} \mathbb{P}[P_i \leq s(x)] d\mathbb{P}^X} \\
 &= \frac{\int_{\mathcal{X}} \pi_0(x) s(x) d\mathbb{P}^X}{\int_{\mathcal{X}} F(s(x)|x) d\mathbb{P}^X}.
 \end{aligned} \tag{3.2}$$

The second line is Bayes theorem. In the penultimate step we used that X_i and P_i are independent under the null. In the last step we used $P_i \mid H_i = 0 \sim \text{U}[0, 1]$.

To maximize the number of discoveries, s needs to maximize:

$$\mathbb{P}[P_i \leq s(X_i)] = \int_{\mathcal{X}} F(s(x)|x) d\mathbb{P}^X. \tag{3.3}$$

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

$$\begin{aligned} & \underset{s \in \mathcal{S}}{\text{maximize}} && \int_{\mathcal{X}} F(s(x)|x) d\mathbb{P}^X, \\ & \text{subject to} && \frac{\int_{\mathcal{X}} \pi_0(x) s(x) d\mathbb{P}^X}{\int_{\mathcal{X}} F(s(x)|x) d\mathbb{P}^X} \leq \alpha. \end{aligned} \quad (3.4)$$

Now as usual, since we do not know $\mathbb{P}^X$, we substitute it by the empirical measure. Equally, we most often do not have access to F or π_0 either. The most natural thing to do is to replace them with estimators $\hat{F}$ and $\hat{\pi}_0$.

$$\begin{aligned} & \underset{s \in \mathcal{S}}{\text{maximize}} && \frac{1}{m} \sum_{i=1}^m \hat{F}(s(X_i)|X_i), \\ & \text{subject to} && \frac{\frac{1}{m} \sum_{i=1}^m \hat{\pi}_0(X_i) s(X_i)}{\frac{1}{m} \sum_{i=1}^m \hat{F}(s(X_i)|X_i)} \leq \alpha. \end{aligned} \quad (3.5)$$

Note that $\hat{F}$ and $\hat{\pi}_0$ typically depend on p -values P_i and covariates X_i . Hence, the weight will depend on $((P_i, X_i))_{i \in [m]}$. Since the number of hypotheses m is a running parameter in our study, we formalize what it means to “learn a weight function” and introduce more involved notation:

Definition 4. A **weighting scheme** $\mathscr{W}^{(\cdot)}$ is a mechanism that, for any subset $I \subset \mathbb{N}_{>0}$, uses the pairs $((P_i, X_i))_{i \in I}$ to learn a **weight function** $\hat{\mathscr{W}}^{(I)} : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$. Throughout this thesis, $I \subset \mathbb{N}_{>0}$ will be finite, i.e. we only have access to a finite sample.

Remark 1. Some remarks on notation: We will use different typefaces in the context of the weight function. We denote a weight function with emphasize on which data was used to learn it with $\hat{\mathscr{W}}^{(I)} : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$. This is to convey that $\hat{\mathscr{W}}^{(I)}$ itself is estimated. If the data set used is less central, we will simply write $\mathscr{W} : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$. The final (random) weights will be written as $W : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}_{\geq 0}$. This ensures that the notation is unambiguous. Typically, the relationship will be $W \propto \mathscr{W} \circ X = \hat{\mathscr{W}}^{(I)} \circ X$.

We therefore formulate the multiple testing procedure with data-driven weights in algorithm 4. We only consider weighting schemes which conditioned on the learning samples $\mathbb{P}$ -almost surely fulfill the requirements from specification 1 on its first two moments.

Specification 1. Let $\mathscr{W}^{(I)}$ be a weighting scheme. From now on we make the assumption:

$$\mathbb{E} \left[\mathscr{W}^{(I)}(X) \middle| ((P_i, X_i))_{i \in I} \right] = 1, \quad \mathbb{E} \left[\left(\mathscr{W}^{(I)}(X) \right)^2 \middle| ((P_i, X_i))_{i \in I} \right] \leq \Gamma,$$

with uniform $\Gamma < \infty$ for all subsets $I \subset \mathbb{N}_{>0}$. Note that this in particular deterministically implies $\mathbb{E} [\mathscr{W}^{(I)}(X)] = 1$ and $\mathbb{E} \left[\left(\mathscr{W}^{(I)}(X) \right)^2 \right] \leq \Gamma$ via iterated expectation.

This specification corresponds to equation (20) in the supplement of [IH21]. Note that the weights will subsequently be re-scaled in algorithm 5, s.t. $\sum_{i=1}^m W_i = m$.

Algorithm 4: Naive weighted BH (wBH)**Input:** p -values $P_1, \dots, P_m \in [0, 1]$ covariates $X_1, \dots, X_m \in \mathcal{X}$ a target level $\alpha \in (0, 1)$ a weighting scheme $\mathcal{W}^{(\cdot)}$ 1 Learn the weight function $\hat{\mathcal{W}}^{([m])} = \mathcal{W}$ from the pairs $((P_i, X_i))_{i \in [m]}$ 2 Let W_i be a suitable rescaling of $\mathcal{W}(X_i)$:

$$W_i := \frac{\mathcal{W}(X_i)}{\sum_{i \in [m]} \mathcal{W}(X_i)}, \text{ if } \sum_{i \in [m]} \mathcal{W}(X_i) > 0, \text{ else } W_i := 1$$

3 Apply the weighted BH (algorithm 3) with weights W_i , p -values P_i , level α

3.3 Breaking naive weighting

In section 3.2 we introduced the weighted BH procedure, which is able to incorporate heterogeneity between hypotheses. We derived a concrete weighting scheme which maximizes the number of rejections, while controlling the marginal version of the FDR. We motivated that the weighting itself depends on the data $((P_i, X_i))_{i \in [m]}$. Here we will show that the naive approach, where we plug-in all our observations into the weighting scheme $\hat{\mathcal{W}}^{(\cdot)}$, asymptotically fails to control the FDR and hence disqualifies as a valid multiple testing procedure. We follow the proof of Proposition 1(a) in S2.1 in [IH21].

We start with an elementary property which follows from assumption 1.

Lemma 1. *Let P be a p -value following the conditional two group model 2.3, where the marginal distribution F suffices assumption 1. Then P is stochastically smaller than the uniform distribution, i.e. $\mathbb{P}(P \leq t) \geq t$ for any $t \in (0, 1)$.*

Proof.

$$\begin{aligned}
\mathbb{P}(P \leq t) &= \mathbb{E}[\mathbb{P}(P \leq t | X)] \\
&= \mathbb{E}[F(t | X)] \\
&= \mathbb{E}[\pi_0(X)t + (1 - \pi_0(X))F_{alt|X}(t)] \\
&= \mathbb{E}[\pi_0(X)t + (1 - \pi_0(X))F_{alt|X}(t \cdot 1 + (1 - t) \cdot 0)] \\
&\geq \mathbb{E}[\pi_0(X)t + (1 - \pi_0(X))(t \cdot F_{alt|X}(1) + (1 - t) \cdot F_{alt|X}(0))] \\
&= \mathbb{E}[\pi_0(X)t + (1 - \pi_0(X))(t \cdot 1 + (1 - t) \cdot 0)] \\
&= t.
\end{aligned}$$

We used $F_{alt}(0 | X_i = x) = 0$ and $F_{alt}(1 | X_i = x) = 1$ for all x and concavity from assumption 1 and $0 \leq \pi_0(X) \leq 1$. $\square$

Next, let us establish a general connection between the uniform distribution and the beta distribution.

Lemma 2. Let $U_1, \dots, U_n \stackrel{iid}{\sim} \mathcal{U}[0, 1]$. Order the random variables, s.t. $U_{(1)} \leq \dots \leq U_{(n)}$. Then the k -th order statistic of the uniform distribution follows a Beta distribution, i.e. $U_{(k)} \sim \text{Beta}(k, n+1-k)$.

Proof. Let $\varepsilon > 0$ and $x \in (0, 1)$. Then

$$\begin{aligned}
\mathbb{P}(U_{(k)} \in [x, x + \varepsilon]) &= \mathbb{P}(\exists i \in [n] : U_i \in [x, x + \varepsilon], |\{j \in [n] | U_j < x\}| = k - 1) \\
&= \sum_{i=1}^n \mathbb{P}(U_i \in [x, x + \varepsilon], |\{j \in [n] | U_j < x\}| = k - 1) \\
&= \sum_{i=1}^n \mathbb{P}(U_i \in [x, x + \varepsilon], |\{j \in [n] \setminus \{i\} | U_j < x\}| = k - 1) \\
&= \sum_{i=1}^n \mathbb{P}(U_i \in [x, x + \varepsilon]) \cdot \mathbb{P}(|\{j \in [n] \setminus \{i\} | U_j < x\}| = k - 1) \\
&= \sum_{i=1}^n \varepsilon \cdot \binom{n-1}{k-1} \cdot x^{k-1} \cdot (1-x)^{n-k} \\
&= n\varepsilon \cdot \binom{n-1}{k-1} \cdot x^{k-1} \cdot (1-x)^{n-k}.
\end{aligned}$$

Derivation by ε yields:

$$f(x) = \frac{\partial}{\partial \varepsilon} \mathbb{P}(U_{(k)} \in [x, x + \varepsilon]) = n \cdot \binom{n-1}{k-1} \cdot x^{k-1} \cdot (1-x)^{n-k}.$$

This shows the claim. $\square$

For our main theorem it will be useful to study the beta distribution.

Lemma 3. Set $k_m := \lfloor \alpha m / 2 \rfloor$ and let $Y \sim \text{Beta}(k_m, m+1-k_m)$. Then as $m \rightarrow \infty$

$$\mathbb{P}[Y \leq \alpha] \rightarrow 1.$$

Proof. From Lemma 14 it follows that

$$\begin{aligned}
\mathbb{E}[Y] &= \frac{k_m}{(m+1)}, \\
\text{Var}[Y] &= \frac{k_m \cdot (m+1-k_m)}{(m+1)^2 \cdot (m+2)}.
\end{aligned}$$

In conjunction with Chebyshev's inequality from Lemma 17, this gives us the following concentration inequality:

$$\begin{aligned}
\mathbb{P}[Y \geq \alpha] &\leq \frac{\text{Var}[Y]}{(\alpha - \mathbb{E}[Y])^2} \\
&= \frac{\frac{k_m \cdot (m+1-k_m)}{(m+1)^2 \cdot (m+2)}}{(\alpha - k_m/(m+1))^2} \\
&= \frac{k_m \cdot (m+1-k_m)}{(m+1)^2 \cdot (m+2) \cdot (\alpha - k_m/(m+1))^2} \\
&= \frac{k_m \cdot (m+1-k_m)}{(m+2) \cdot (\alpha(m+1) - k_m)^2} \\
&= \frac{mk_m - k_m^2 + k_m}{(m+2) \cdot (\alpha(m+1) - k_m)^2}.
\end{aligned}$$

The denominator grows as m^3 , whereas the numerator only grows with m^2 rate. So we conclude $\mathbb{P}[Y \geq \alpha] \rightarrow 0$ and hence $\mathbb{P}[Y \leq \alpha] \rightarrow 1$. $\square$

We are now ready to proof that the naive IHW algorithm 4 can exceed the claimed type-I error. The idea behind the weight construction in the proof is the following: Storey et al. [STS04] proved asymptotic FDR control of BH by arguing by the Glivenko-Cantelli theorem, $\sup_t \left| \frac{1}{m} \sum_{i=1}^m [\mathbf{1}(P_i \leq t) - \mathbb{P}[P_i \leq t]] \right| \xrightarrow{\mathbb{P}} 0$ and similarly for the subset of null hypotheses [IH21]. A consequence is that the BH estimator of the (bayesian) FDR in algorithm 2 $\widehat{\text{Fdr}}(t) = \frac{t}{\widehat{F}(t) \vee \frac{1}{m}}$ is asymptotically uniformly conservative over all thresholds $\delta > 0$ which in turn implies asymptotic FDR control [IH21].

Extending this argument to the weighted case would require uniform convergence: $\sup_t \left| \frac{1}{m} \sum_{i=1}^m [\mathbf{1}(P_i \leq tW_i) - \mathbb{P}[P_i \leq tW_i]] \right| \xrightarrow{\mathbb{P}} 0$. However, if we are free to choose any data-driven weights satisfying specification 1, the convergence might not be uniform. A consequence is that we might find regions of $\mathcal{X}$ -space that are enriched for small p -values merely by chance, upweight them and violate FDR control [IH21].

Theorem 2. *Consider the conditional two-groups model 2.3. Assume $\mathbb{P}^X \ll \lambda$, i.e. the covariate distribution $\mathbb{P}^X$ is absolutely continuous w.r.t the Lebesgue measure λ . Let assumption 1 hold. Fix $\alpha \in (0, \frac{1}{2})$ and suppose there is a $\delta > 0$ with $\mathbb{E}[\pi_0(X_i)] > 1 - \delta$, i.e., there are “not too many alternatives”, and $1 - \frac{2.3\delta}{\alpha} > \alpha$. Then there exists a weighting scheme $\hat{\mathcal{W}}^{(\cdot)}$ (see definition 4) sufficing specification 1, such that wBH as in algorithm 4 does not control the FDR asymptotically (see definition 2).*

Proof. We closely follow the proof of Proposition 1(a) in S2.1 in [IH21]. Set $k_m := \lfloor \alpha m / 2 \rfloor$ and define $\mathcal{J}_m \subset [m]$ as the index set of the k_m smallest p -values. We first need to define the weighting scheme $\hat{\mathcal{W}}^{(\cdot)}$.

$$\hat{\mathcal{W}}^{([m])}(x) := \mathbb{1}(x \notin \{X_j, j \in [m]\}) + \frac{m}{k_m} \mathbb{1}(x \in \{X_j, j \in \mathcal{J}_m\}).$$

In words: We set $\hat{\mathcal{W}}^{([m])} \equiv 1$ outside of the covariates $X_1, \dots, X_m$, i.e. almost everywhere. We ignore all hypotheses with higher p -values i.e. outside of $\mathcal{J}_m$. Since those hypotheses are assigned zero weight, they cannot be rejected. So it is immediate from the weighting scheme that the number of rejections is smaller or equal than $|\mathcal{J}_m| = k_m$. We up-weight the hypotheses inside of $\mathcal{J}_m$, s.t. the weights average to 1, i.e. $\sum_{i=1}^m \hat{\mathcal{W}}^{([m])}(X_i) = m$.

This weighting scheme satisfies the conditions of specification 1: Since $\hat{\mathcal{W}}^{([m])} \equiv 1$ almost everywhere and $\mathbb{P}^X$ is absolutely continuous w.r.t the Lebesgue measure, we have $\mathbb{E}[\hat{\mathcal{W}}^{([m])}(X_i)] = 1$. Second, for m large enough we have $\sup_{x \in \mathcal{X}} \hat{\mathcal{W}}^{([m])}(x) = m/k_m = m/\lfloor \alpha m / 2 \rfloor \leq 4/\alpha$. Consequently, for m large enough

$$\begin{aligned} \int \widehat{W}^{([m])}(x)^2 d\mathbb{P}^X(x) &\leq \int \sup_{x \in \mathcal{X}} \hat{\mathcal{W}}^{([m])}(x)^2 d\mathbb{P}^X(x) \\ &\leq \int \frac{16}{\alpha^2} d\mathbb{P}^X(x) \\ &= \frac{16}{\alpha^2} \\ &= \Gamma. \end{aligned}$$

For the ordered p -values $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$ consider the events $A_m = \{P_{(k_m)} \leq \alpha\}$ and $B_m = \{\sum_{i=1}^m H_i \leq 1.1\delta m\}$. In the following, we want to operate on the event $A_m \cap B_m$.

We have already seen that the hypotheses rejected must be contained in $\mathcal{J}_m$. We argue that on the event A_m all hypotheses in $\mathcal{J}_m$ will be rejected. By the ordering we have $P_{(k_m)} = \max_{i \in \mathcal{J}_m} P_i$. Therefore, for $i \in \mathcal{J}_m$, we have

$$P_i \mathbb{1}(A_m) \leq P_{(k_m)} \mathbb{1}(A_m) \leq \alpha = (k_m \cdot \alpha/m) \cdot (m/k_m) = (k_m \cdot \alpha/m) \cdot \hat{\mathcal{W}}^{([m])}(X_i).$$

Recalling the definition of the weighted Benjamini-Hochberg procedure 3, this implies that the k_m hypotheses in $\mathcal{J}_m$ will be rejected on A_m . Hence, we have k_m rejections (corresponding to R in 2.1). On the other hand, on the event B_m at most $1.1\delta m$ of those rejections are correct (corresponding to an upper bound of S in 2.1). Combining both implies that at least $k_m - 1.1\delta m$ hypotheses are falsely rejected on the event $A_m \cap B_m$ (corresponding to a lower bound of $V = R - S$ in 2.1). Consequently, for large enough m the false discovery proportion on $A_m \cap B_m$ can be bounded as

$$\text{FDP}_m \mathbb{1}(A_m \cap B_m) \geq \frac{k_m - 1.1\delta m}{k_m} = 1 - \frac{1.1\delta m}{[\alpha m / 2]} \geq 1 - 1.1 \cdot \delta \cdot 2 \cdot \frac{m}{[\alpha m]} \geq 1 - \frac{2.3\delta}{\alpha}.$$

We are using $[a \cdot b] \geq [a] \cdot [b]$ for non-negative a, b and that $\frac{m}{[\alpha m]}$ gets arbitrarily close to $\frac{1}{\alpha}$ for $m \rightarrow \infty$. Since $1 - \frac{2.3\delta}{\alpha} > \alpha$, for large enough m we have $\text{FDP}_m \cdot \mathbb{1}(A_m \cap B_m) > \alpha$. To show that FDR_m is asymptotically not controlled, it suffices to show $\mathbb{P}[A_m], \mathbb{P}[B_m] \rightarrow 1$. Let $U_1, \dots, U_n$ be uniformly distributed on $[0, 1]$. Lemma 1 implies $\mathbb{P}[P_i \leq \alpha] \geq \mathbb{P}[U_i \leq \alpha]$ and thus $\mathbb{P}[P_{(k_m)} \leq \alpha] \geq \mathbb{P}[U_{(k_m)} \leq \alpha]$. Due to Lemma 2, we have $U_{(k_m)} \sim \text{Beta}(k_m, n + 1 - k_m)$. We see from Lemma 3, $\mathbb{P}[U_{(k_m)} \leq \alpha] \rightarrow 1$. We conclude $\mathbb{P}[A_m] = \mathbb{P}[P_{(k_m)} \leq \alpha] \geq \mathbb{P}[U_{(k_m)} \leq \alpha] \rightarrow 1$.

For B_m , the conditional two group model 2.3 implies $\sum_{i=1}^m H_i \sim \text{Binomial}(m, \mathbb{E}\pi_1(X_i))$. Then:

$$\begin{aligned} \mathbb{P}[B_m] &= \mathbb{P}\left[\sum_{i=1}^m H_i \leq 1.1\delta m\right] \\ &= \mathbb{P}\left[\frac{1}{m} \sum_{i=1}^m H_i - \mathbb{E}[H_1] \leq 1.1 \cdot \delta - \mathbb{E}[H_1]\right] \\ &= \mathbb{P}\left[\frac{1}{m} \sum_{i=1}^m H_i - \mathbb{E}[H_1] \leq 1.1 \cdot \delta - \mathbb{E}[\pi_1(X_i)]\right] \\ &\geq \mathbb{P}\left[\frac{1}{m} \sum_{i=1}^m H_i - \mathbb{E}[H_1] \leq 0.1 \cdot \delta\right] \\ &\xrightarrow{m \rightarrow \infty} 1. \end{aligned}$$

In the inequality, we used that we assumed $\mathbb{E}[\pi_1(X_i)] < \delta$. We conclude by using the dominated convergence theorem:

$$\begin{aligned} \limsup_{m \rightarrow \infty} \text{FDR}_m &= \limsup_{m \rightarrow \infty} \mathbb{E}[\text{FDP}_m] \\ &= \mathbb{E}[\limsup_{m \rightarrow \infty} \text{FDP} \cdot \mathbb{1}(A_m \cap B_m)] \\ &> \alpha. \end{aligned}$$

□

The intuition behind theorem 2 is the following: The decision rule $P_i \leq W_i \hat{t}^*$ for a null hypothesis H_i depends on its p -value P_i , the weight W_i and the total number of rejections R , where R itself is a function of the former two [SSS22]. In algorithm 4, the data-driven weight W_i can depend on the p -value P_i , i.e. a p -value can influence its own weight. Consequently, the effective rejection threshold $W_i \hat{t}^*$ is prone to over-fitting to the data pairs $((P_i, X_i))_{i \in [m]}$. To avoid this, Ignatiadis and Huber [IH21] use cross-weighting, as will be described in chapter 4.

Chapter 4

Crossweighted Benjamini-Hochberg

Chapter 3 presents a multiple testing procedure (algorithm 4) with the aim to maximize the number of discoveries. However, we showed that it does not control the FDR. In this chapter we set out to overcome this challenge by cross-folding - a concept related to cross-fitting [Sch86], which has been successful in the fields of causal inference [Che+18; NW21] and empirical Bayes for with high-dimensional nuisance parameters [IW19]. We will show that the derived procedure asymptotically controls the FDR. Compared to the proof in [IH21] we have divided the proof into a sequence of sections and lemmas. For the sake of completeness, we will also give a finite sample result without proof from [IH21]. We will discuss and contrast the mathematical results with the results in Ignatiadis and Huber [Ign+16; IH21]

4.1 Definition of IHW

In chapter 3 we saw that the plug-in approach in algorithm 4 asymptotically fails to control the FDR. This is not a problem unique to algorithm 4 [IH21]. More importantly, with finite samples, the FDR of such plug-in approaches is often above the target level α [IH21; Dur19; CS09]. The plug-in approaches exceed the claimed type-I error, because they suffer from an overfitting phenomenon: They use the same data $((P_i, X_i))_{i \in [m]}$ for the wBH decision rule from algorithm 3 $P_i \leq W_i \hat{t}^*$ twice: through the p -values $(P_i)_{i \in [m]}$ and the weights $(W_i)_{i \in [m]}$ [Dur19].

To overcome this, Ignatiadis et al. [Ign+16] introduce randomization to algorithm 4 by means of hypothesis splitting resulting in the IHW presented in its general form in algorithm 5. As recently demonstrated in a plethora of papers [Was14; Dwo+15; TBT16; Ign+16], introducing external randomness to statistical procedures can often be the key to tractable inference for high dimensional problems. The hypothesis splitting steps (step 1 and 2) in algorithm 5 are exactly in this spirit: The m hypotheses are partitioned into K disjoint folds $I_1, \dots, I_K$. Then, in determining the weight W_i for hypothesis $i \in I_\ell$, the weights are set $W_i \propto \mathcal{W}^{-\ell}(X_i)$, where the weight function $\mathcal{W}^{-\ell}$ is learned from data outside of the fold I_ℓ and the weights are normalized s.t. $\sum_{i \in I_\ell} W_i = |I_\ell|$.

Algorithm 5: IHW**Input:**

p -values $P_1, \dots, P_m \in [0, 1]$
 covariates $X_1, \dots, X_m \in \mathcal{X}$
 a target level $\alpha \in (0, 1)$
 a weighting scheme $\mathscr{W}^{\hat{(\cdot)}}$
 number of folds K

-
- 1 Partition $[m]$ into K disjoint folds $I_1, \dots, I_K$, independently of $((P_i, X_i))_{i \in [m]}$
 - 2 **for** $\ell = 1, \dots, K$ **do**
 - 3 Learn the weight function $\mathscr{W}^{\hat{([m] \setminus I_\ell)}} = \mathscr{W}^{-\ell}$ from the pairs $((P_i, X_i))_{i \notin I_\ell}$
 - 4 **for** $i \in I_\ell$ **do**
 - 5 Let W_i be a suitable re-scaling of $\mathscr{W}^{-\ell}(X_i)$,

$$W_i := \frac{|I_\ell| \mathscr{W}^{-\ell}(X_i)}{\sum_{i \in I_\ell} \mathscr{W}^{-\ell}(X_i)}, \text{ if } \sum_{i \in I_\ell} \mathscr{W}^{-\ell}(X_i) > 0, \text{ else } W_i := 1$$
 - 6 Apply the weighted BH (algorithm 3) with weights W_i , p -values P_i , level α

The cross-weighting approach ensures that W_i is determined based on $(X_i)_{i \in [m]}$ and the p -values from the other folds, but not P_i . Consequently, the information brought by a given p -value P_i is used only once instead of twice for the decision regarding hypothesis H_i . This core idea of IHW reduces overfitting and as we will see suffices to salvage FDR control at essentially no cost.

4.2 Empirical formulation of the FDR

The main result of this thesis is asymptotic FDR control of IHW 5. We first need to do some preparation and study IHW step by step. We want to formulate the FDR and our algorithm from a more empirical point of view. Recall that for a threshold t and weights W_i the weighted Benjamini-Hochberg decision rule 3 takes the form

$$\text{Reject hypothesis } i \iff P_i \leq W_i t.$$

In our case, W_i always takes the form $c_i \cdot \mathscr{W}(X_i)$ with the re-scaling factor c_i and the weight function $\mathscr{W}$. Therefore, we can write the number of rejections as

$$\sum_{i \in [m]} \mathbb{1}(P_i \leq t c_i \mathscr{W}(X_i)). \quad (4.1)$$

Equally, we can write the number of false rejections as

$$\sum_{i \in [m]} \mathbb{1}(H_i = 0) \mathbb{1}(P_i \leq t c_i \mathscr{W}(X_i)). \quad (4.2)$$

For the first part of our analysis we will only operate on single folds. In that spirit and with the number of rejections (4.1) in mind, we define following processes and

notations used throughout this chapter

$$\begin{aligned}\widehat{F}(\hat{t}^{\text{BH}}, \mathcal{W}; I) &:= \frac{1}{|I|} \sum_{i \in I} \mathbb{1}(P_i \leq t \mathcal{W}(X_i)), \\ \widehat{F}_0(t, \mathcal{W}; I) &:= \frac{1}{|I|} \sum_{i \in I} \mathbb{1}(H_i = 0) \mathbb{1}(P_i \leq t \mathcal{W}(X_i)), \\ \widehat{F}_0^{\text{BH}}(t, \mathcal{W}; I) &:= t \cdot \frac{1}{|I|} \sum_{i \in I} \mathcal{W}(X_i).\end{aligned}\tag{4.3}$$

The processes are parametrized by the threshold $t \in [0, 1]$, the weighting function $\mathcal{W}$ and the index set $I \subset [m]$ of the hypotheses in the fold under study. Furthermore, we define population counterparts of above empirical processes

$$\begin{aligned}F(t, \mathcal{W}) &:= \mathbb{P}[P_i \leq t \mathcal{W}(X_i)], \\ F_0(t, \mathcal{W}) &:= \mathbb{P}[P_i \leq t \mathcal{W}(X_i); H_i = 0], \\ F_0^{\text{BH}}(t, \mathcal{W}) &:= t \cdot \mathbb{E}[\mathcal{W}(X_i)].\end{aligned}\tag{4.4}$$

We will also require the following normalization factor:

$$\hat{c}_{I, \mathcal{W}} := \frac{|I|}{\sum_{i \in I} \mathcal{W}(X_i)}.\tag{4.5}$$

Remark 2. From specification 1 we see that $\mathbb{P}$ -almost surely $\mathbb{E}[\hat{c}_{I, \mathcal{W}}^{-1} | \mathcal{D}_{[m] \setminus I}] = 1$.

In particular, the normalization is constant within a fold. All combined, we can now write the number of rejections across all folds as

$$\sum_{i \in [m]} \mathbb{1}(P_i \leq t \cdot c_i \cdot \mathcal{W}(X_i)) = \sum_{l \in [K]} |I_l| \cdot \widehat{F}(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l),$$

and the number of false rejections across all folds

$$\sum_{i \in [m]} \mathbb{1}(H_i = 0) \mathbb{1}(P_i \leq t \cdot c_i \cdot \mathcal{W}(X_i)) = \sum_{l \in [K]} |I_l| \cdot \widehat{F}_0(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l).$$

We arrive at a new formulation for the FDP of IHW given a threshold t

$$\text{FDP}(t) := \frac{\sum_{l \in [K]} |I_l| \cdot \widehat{F}_0(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l)}{\max \left\{ 1, \sum_{l \in [K]} |I_l| \cdot \widehat{F}(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l) \right\}}.\tag{4.6}$$

The overall goal is to show that FDP and hence $\text{FDR} = \mathbb{E}[\text{FDP}]$ is asymptotically controlled beneath the target level α . Let us outline the main ideas of the proof. We will easily see from the definition of the Benjamini-Hochberg procedure that $\text{FDP}^{\text{BH}} \leq \alpha$ with

$$\text{FDP}^{\text{BH}}(t) := \frac{\sum_{l \in [K]} |I_l| \cdot \widehat{F}_0^{\text{BH}}(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l)}{\max \left\{ 1, \sum_{l \in [K]} |I_l| \cdot \widehat{F}(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l) \right\}},\tag{4.7}$$

We will bridge the gap from FDP to FDP^{BH} by going over their respective population counterparts

$$\begin{aligned}\widetilde{\text{FDP}}^{\text{BH}}(t) &= \frac{\sum_{l \in [K]} |I_l| \cdot F_0^{\text{BH}}(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l)}{\max \left\{ 1, \sum_{l \in [K]} |I_l| \cdot \widehat{F}(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l) \right\}}, \\ \widetilde{\text{FDP}}(t) &= \frac{\sum_{l \in [K]} |I_l| \cdot F_0(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l)}{\max \left\{ 1, \sum_{l \in [K]} |I_l| \cdot \widehat{F}(t, \hat{c}_{I_l, \mathcal{W}^{([m] \setminus I_l)}} \cdot \mathcal{W}^{([m] \setminus I_l)}; I_l) \right\}}.\end{aligned}\tag{4.8}$$

Note that we plugged in F_0 and F_0^{BH} instead of $\widehat{F}_0$ and $\widehat{F}_0^{BH}$ respectively. The aim of this section is to use uniform (in t) laws of large numbers to formulate relationships of the empirical processes from (4.3) and their population counterparts in (4.4). For the proofs, we closely follow S2.2 in [IH21].

The uniform convergence of the empirical cumulative distribution function F_n against the true cumulative distribution function F is given in the Glivenko–Cantelli theorem. We will require a concrete convergence rate stated in lemma 4.

Lemma 4. *Dvoretzky–Kiefer–Wolfowitz inequality:*

Let $X_1, \dots, X_n \stackrel{iid.}{\sim} F$ be real valued random variables with cumulative distribution function F . Let F_n denote the associated empirical distribution function defined by $F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq t\}}$. Then it holds for $\varepsilon > 0$:

$$\mathbb{P} \left(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2}.$$

Proof. A proof is given in [DKW56]. □

We will use the Landau notation 5 to denote a sequence of random variables X_m converging to 0 in probability as $m \rightarrow \infty$.

Definition 5. Landau Notation:

$$X_m = o_{\mathbb{P}}(1) \Leftrightarrow \forall \varepsilon > 0 : \lim_{m \rightarrow \infty} \mathbb{P}(|X_m| \geq \varepsilon) = 0,$$

$$X_m = O_{\mathbb{P}}(1) \Leftrightarrow \forall \varepsilon > 0 \exists m_0 \in \mathbb{N}, M > 0 \forall m \geq m_0 : \mathbb{P}(|X_m| \geq M) \leq \varepsilon.$$

Lemma 5. *With the Landau notation from definition 5 we have:*

$$o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1),$$

$$o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1),$$

$$(1 + o_{\mathbb{P}}(1))^{-1} = O_{\mathbb{P}}(1).$$

Proof. See section 2.2 in van der Vaart [Van00]. □

We begin by relating F to $\widehat{F}$ and F_0 to $\widehat{F}_0$.

Lemma 6. *Let $((P_i, X_i))_{i \in \mathbb{N}}$ be from the conditional two-groups model 2.3 satisfying regularity assumption 1. Let $\{I^m\}_{m \in \mathbb{N}}$ be a sequence of folds with $I^m \subset [m]$ and $\widehat{\mathcal{W}}^{([m] \setminus I^m)}$ a weight function learned out-of-fold on $[m] \setminus I^m$ satisfying specification 1. Suppose $|I^m| \rightarrow \infty$ as $m \rightarrow \infty$. Then we can relate empirical processes 4.3 and the population processes 4.4 uniformly over all thresholds t :*

$$\begin{aligned} \sup_{t \in [0,1]} \left| \widehat{F} \left(t, \widehat{c}_{I^m, \widehat{\mathcal{W}}^{([m] \setminus I^m)}} \cdot \widehat{\mathcal{W}}^{([m] \setminus I^m)}; I^m \right) - F(t, \widehat{\mathcal{W}}^{([m] \setminus I^m)}) \right| &= o_{\mathbb{P}}(1), \\ \sup_{t \in [0,1]} \left| \widehat{F}_0 \left(t, \widehat{c}_{I^m, \widehat{\mathcal{W}}^{([m] \setminus I^m)}} \cdot \widehat{\mathcal{W}}^{([m] \setminus I^m)}; I^m \right) - F_0(t, \widehat{\mathcal{W}}^{([m] \setminus I^m)}) \right| &= o_{\mathbb{P}}(1). \end{aligned}$$

Proof. We closely follow S2.2 in [IH21]. For ease of notation, we will write I and $\hat{\mathcal{W}}$, instead of I^m and $\hat{\mathcal{W}}^{([m]\setminus I^m)}$ despite them depending on m . We will only show the claim for $\hat{F}$. All steps work for $\hat{F}_0$ analogously. $\hat{\mathcal{W}}$ being learned out-of-fold means that it only depends on $\mathcal{D}_{[m]\setminus I} := ((P_i, X_i))_{i \in [m]\setminus I}$. For the first part of our analysis, we will condition on $\mathcal{D}_{[m]\setminus I}$. This enables us to treat $\hat{\mathcal{W}}^{([m]\setminus I)}$ as deterministic. We will resolve this later on with iterative expectation. Fix $\delta > 0$. We begin by showing that $A_m = o_{\mathbb{P}}(1)$ with

$$A_m := \mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F}(t, \hat{c}_{I, \hat{\mathcal{W}}} \cdot \hat{\mathcal{W}}; I) - F(t, \hat{\mathcal{W}}) \right| > 3\delta \middle| \mathcal{D}_{[m]\setminus I} \right]. \quad (4.9)$$

By the law of large numbers and remark 2, $\hat{c}_{I, \hat{\mathcal{W}}}$ (4.5) converges to 1 conditioned on $\mathcal{D}_{[m]\setminus I}$. It immediately follows that the random variable $|\hat{c}_{I, \hat{\mathcal{W}}} - 1| \mid \mathcal{D}_{[m]\setminus I}$ converges stochastically to 0, i.e. $|\hat{c}_{I, \hat{\mathcal{W}}} - 1| \xrightarrow{\mathbb{P}} 0$ with $\mathbb{P} = \mathbb{P}[\cdot \mid \mathcal{D}_{[m]\setminus I}]$. Lemma 15 gives us a positive null-sequence $\{\varepsilon_m\}_{m \in \mathbb{N}}$, s.t. $\mathbb{P} [|\hat{c}_{I, \hat{\mathcal{W}}} - 1| > \varepsilon_m \mid \mathcal{D}_{[m]\setminus I}] = o_{\mathbb{P}}(1)$. We separate 4.9 into two terms:

$$\begin{aligned} B_m &:= \mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F}(\hat{c}_{I, \hat{\mathcal{W}}} \cdot t, \hat{\mathcal{W}}; I) - \hat{F}(\hat{c}_{I, \hat{\mathcal{W}}} \cdot t, \hat{\mathcal{W}}; I) \mathbb{1}(|\hat{c}_{I, \hat{\mathcal{W}}} - 1| < \varepsilon_m) \right| > \delta \middle| \mathcal{D}_{[m]\setminus I} \right], \\ C_m &:= \mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F}(\hat{c}_{I, \hat{\mathcal{W}}} \cdot t, \hat{\mathcal{W}}; I) \mathbb{1}(|\hat{c}_{I, \hat{\mathcal{W}}} - 1| < \varepsilon_m) - F(t, \hat{\mathcal{W}}) \right| > 2\delta \middle| \mathcal{D}_{[m]\setminus I} \right]. \end{aligned}$$

The triangle inequality enables us to obtain $A_m \leq B_m + C_m$. Thus, it suffices to bound B_m and C_m to bound A_m . $A_m = o_{\mathbb{P}}(1)$ follows from $B_m = o_{\mathbb{P}}(1)$ and $C_m = o_{\mathbb{P}}(1)$. Let us first examine B_m :

$$\begin{aligned} B_m &= \mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F}(\hat{c}_{I, \hat{\mathcal{W}}} \cdot t, \hat{\mathcal{W}}; I) - \hat{F}(\hat{c}_{I, \hat{\mathcal{W}}} \cdot t, \hat{\mathcal{W}}; I) \cdot \mathbb{1}(|\hat{c}_{I, \hat{\mathcal{W}}} - 1| < \varepsilon_m) \right| > \delta \middle| \mathcal{D}_{[m]\setminus I} \right] \\ &= \mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F}(\hat{c}_{I, \hat{\mathcal{W}}} \cdot t, \hat{\mathcal{W}}; I) \right| \cdot \mathbb{1}(|\hat{c}_{I, \hat{\mathcal{W}}} - 1| \geq \varepsilon_m) > \delta \middle| \mathcal{D}_{[m]\setminus I} \right] \\ &\leq \mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F}(\hat{c}_{I, \hat{\mathcal{W}}} \cdot t, \hat{\mathcal{W}}; I) \right| > 0, \mathbb{1}(|\hat{c}_{I, \hat{\mathcal{W}}} - 1| \geq \varepsilon_m) > 0 \middle| \mathcal{D}_{[m]\setminus I} \right] \\ &\leq \mathbb{P} \left[\mathbb{1}(|\hat{c}_{I, \hat{\mathcal{W}}} - 1| \geq \varepsilon_m) > 0 \middle| \mathcal{D}_{[m]\setminus I} \right] \\ &\leq \mathbb{P} \left[|\hat{c}_{I, \hat{\mathcal{W}}} - 1| \geq \varepsilon_m \middle| \mathcal{D}_{[m]\setminus I} \right]. \end{aligned} \quad (4.10)$$

We want to bound the latter term. However, it is easier to work with $\hat{c}_{I, \hat{\mathcal{W}}}^{-1}$ than $\hat{c}_{I, \hat{\mathcal{W}}}$. To relate the $\hat{c}_{I, \hat{\mathcal{W}}}^{-1}$ and $\hat{c}_{I, \hat{\mathcal{W}}}$, we use the Lipschitz continuity of $x \mapsto \frac{1}{x}$ on $(\frac{1}{2}, \infty)$:

$$\begin{aligned} |\hat{c}_{I, \hat{\mathcal{W}}} - 1| &= |\hat{c}_{I, \hat{\mathcal{W}}} - 1| \cdot \mathbb{1}(\hat{c}_{I, \hat{\mathcal{W}}} \geq 2) + |\hat{c}_{I, \hat{\mathcal{W}}} - 1| \cdot \mathbb{1}(\hat{c}_{I, \hat{\mathcal{W}}} < 2) \\ &= |\hat{c}_{I, \hat{\mathcal{W}}} - 1| \cdot \mathbb{1}(\hat{c}_{I, \hat{\mathcal{W}}} \geq 2) + \hat{c}_{I, \hat{\mathcal{W}}} \cdot |1 - \hat{c}_{I, \hat{\mathcal{W}}}^{-1}| \cdot \mathbb{1}(\hat{c}_{I, \hat{\mathcal{W}}} < 2) \\ &< |\hat{c}_{I, \hat{\mathcal{W}}} - 1| \cdot \mathbb{1}\left(\hat{c}_{I, \hat{\mathcal{W}}}^{-1} \leq \frac{1}{2}\right) + 2 \cdot |1 - \hat{c}_{I, \hat{\mathcal{W}}}^{-1}| \\ &\leq |\hat{c}_{I, \hat{\mathcal{W}}} - 1| \cdot \mathbb{1}\left(|1 - \hat{c}_{I, \hat{\mathcal{W}}}^{-1}| \geq \frac{1}{2}\right) + 2 \cdot |1 - \hat{c}_{I, \hat{\mathcal{W}}}^{-1}|. \end{aligned} \quad (4.11)$$

Plugging in the observations from (4.11) into (4.10) implies

$$\begin{aligned}
 B_m &\leq \mathbb{P} \left[\left| \hat{c}_{I, \mathcal{H}} - 1 \right| \geq \varepsilon_m \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq \mathbb{P} \left[\left| \hat{c}_{I, \mathcal{H}} - 1 \right| \cdot \mathbb{1} \left(\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{1}{2} \right) + 2 \cdot \left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \varepsilon_m \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq \mathbb{P} \left[\left| \hat{c}_{I, \mathcal{H}} - 1 \right| \cdot \mathbb{1} \left(\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{1}{2} \right) \geq \frac{\varepsilon_m}{2} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\quad + \mathbb{P} \left[2 \cdot \left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{\varepsilon_m}{2} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq \mathbb{P} \left[\mathbb{1} \left(\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{1}{2} \right) > 0 \middle| \mathcal{D}_{[m] \setminus I} \right] + \mathbb{P} \left[2 \cdot \left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{\varepsilon_m}{2} \middle| \mathcal{D}_{[m] \setminus I} \right] \quad (4.12) \\
 &\leq \mathbb{P} \left[\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{1}{2} \middle| \mathcal{D}_{[m] \setminus I} \right] + \mathbb{P} \left[\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{\varepsilon_m}{4} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq 2\mathbb{P} \left[\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{\min(2, \varepsilon_m)}{4} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &= 2\mathbb{P} \left[\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq \frac{\varepsilon_m}{4} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq 2\mathbb{P} \left[\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq (4|I|^{1/4})^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right].
 \end{aligned}$$

The third inequality is $\mathbb{P}(X + Y > 2a) \leq \mathbb{P}(X > a) + \mathbb{P}(Y > a)$ for $X, Y \geq 0$. The penultimate step used that $\{\varepsilon_m\}_{m \in \mathbb{N}}$ is a null-sequence and w.l.o.g. $\varepsilon_m \leq 2$. The last line follows from $\varepsilon_m \geq \frac{1}{m^{1/4}} \geq |I|^{-1/4}$ by Lemma 15. It remains to bound $2\mathbb{P} \left[\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq (4|I|)^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right]$. By remark 2 we have $\mathbb{E} \left[\hat{c}_{I, \mathcal{H}}^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right] = 1$ $\mathbb{P}$ -almost surely. Chebyshev's inequality 17 conditionally on $\mathcal{D}_{[m] \setminus I}$ yields $\mathbb{P}$ -almost surely:

$$\begin{aligned}
 B_m &\leq 2\mathbb{P} \left[\left| 1 - \hat{c}_{I, \mathcal{H}}^{-1} \right| \geq (4|I|)^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &= 2\mathbb{P} \left[\left| \hat{c}_{I, \mathcal{H}}^{-1} - \mathbb{E} \left[\hat{c}_{I, \mathcal{H}}^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right] \right| > (4|I|)^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq 32(|I|^{1/4})^2 \mathbb{E} \left[\left(\hat{c}_{I, \mathcal{H}}^{-1} - \mathbb{E} \left[\hat{c}_{I, \mathcal{H}}^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right] \right)^2 \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &= 32|I|^{1/2} \mathbb{V}ar \left[\hat{c}_{I, \mathcal{H}}^{-1} \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &= 32|I|^{1/2} \frac{1}{|I|^2} \mathbb{V}ar \left[\sum_{i \in I} \mathcal{H}(X_i) \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &= 32|I|^{-3/2} \sum_{i \in I} \mathbb{V}ar \left[\mathcal{H}(X_i) \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &= 32|I|^{-3/2} |I| \mathbb{V}ar \left[\mathcal{H}(X_i) \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &= 32|I|^{-1/2} \mathbb{V}ar \left[\mathcal{H}(X_i) \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq 32|I|^{-1/2} \mathbb{E} \left[\mathcal{H}(X_i)^2 \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq 32|I|^{-1/2} \Gamma.
 \end{aligned}$$

In the third line we used the Jensen inequality 18. The last line follows from specification 1. In summary, we have found $B_m \leq 32|I|^{-1/2}\Gamma$. Next we turn to C_m . Recall that it has the form

$$C_m = \mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F} \left(\hat{c}_{I, \mathcal{H}} \cdot t, \mathcal{H}; I \right) \mathbb{1} \left(\left| \hat{c}_{I, \mathcal{H}} - 1 \right| < \varepsilon_m \right) - F \left(t, \mathcal{H} \right) \right| > 2\delta \middle| \mathcal{D}_{[m] \setminus I} \right].$$

We first study the term from an analytic point. We have $\mathbb{P}$ -almost surely:

$$\begin{aligned}
\widehat{F}(\hat{c}_{I,\mathcal{W}} \cdot t, \mathcal{W}; I) \mathbb{1}(\hat{c}_{I,\mathcal{W}} < 1 + \varepsilon_m) &= \frac{1}{|I|} \sum_{i \in I} \mathbb{1}(P_i \leq \hat{c}_{I,\mathcal{W}} \cdot t \mathcal{W}(X_i)) \cdot \mathbb{1}(\hat{c}_{I,\mathcal{W}} < 1 + \varepsilon_m) \\
&= \frac{1}{|I|} \sum_{i \in I} \mathbb{1}(P_i \leq \hat{c}_{I,\mathcal{W}} \cdot t \mathcal{W}(X_i), \hat{c}_{I,\mathcal{W}} < 1 + \varepsilon_m) \\
&\leq \frac{1}{|I|} \sum_{i \in I} \mathbb{1}(P_i \leq (1 + \varepsilon_m) \cdot t \mathcal{W}(X_i)) \\
&= \widehat{F}((1 + \varepsilon_m) \cdot t, \mathcal{W}; I).
\end{aligned}$$

The same argument in reverse direction gives us

$$\widehat{F}((1 - \varepsilon_m) \cdot t, \mathcal{W}; I) \leq \widehat{F}(\hat{c}_{I,\mathcal{W}} \cdot t, \mathcal{W}; I) \mathbb{1}(|\hat{c}_{I,\mathcal{W}} - 1| < \varepsilon_m) \leq \widehat{F}((1 + \varepsilon_m) \cdot t, \mathcal{W}; I). \quad (4.13)$$

Thus, we have found an interval $\widehat{F}(\hat{c}_{I,\mathcal{W}} \cdot t, \mathcal{W}; I) \mathbb{1}(|\hat{c}_{I,\mathcal{W}} - 1| < \varepsilon_m)$ must lay in, independent of $\hat{c}_{I,\mathcal{W}}$. We also want to find an interval for $|F((1 \pm \varepsilon_m)t, \mathcal{W}) - F(t, \mathcal{W})|$. Consider $F(t|X_i) = \mathbb{P}[P_i \leq t|X_i]$, the cumulative distribution function of P_i conditioned on X_i , not to be confused with $F(t, \mathcal{W}) = \mathbb{P}[P_i \leq t \mathcal{W}(X_i)]$ from (4.3). In the conditional two group model it takes the form $F(t|X_i = x) = \pi_0(x)t + (1 - \pi_0(x))F_{alt|X_i=x}(t)$, so $t \mapsto F(t|X_i = x)$ is the convex combination of $t \mapsto t$ and $t \mapsto F_{alt|X_i=x}(t)$. Since the identity is 1-Lipschitz and $F_{alt|X_i=x}$ is $L(x)$ -Lipschitz by Assumption 1, $t \mapsto F(t|X_i = x)$ must be $\max\{1, L(x)\}$ Lipschitz. We proceed by showing that we $\mathbb{P}$ -almost surely have:

$$\begin{aligned}
|F((1 \pm \varepsilon_m)t, \mathcal{W}) - F(t, \mathcal{W})| &= |\mathbb{E}_{X_i} [F((1 \pm \varepsilon_m)t \cdot \mathcal{W}(X_i)|X_i) - F(t \cdot \mathcal{W}(X_i)|X_i)]| \\
&\leq \mathbb{E}_{X_i} [|F((1 \pm \varepsilon_m)t \cdot \mathcal{W}(X_i)|X_i) - F(t \cdot \mathcal{W}(X_i)|X_i)|] \\
&\leq \mathbb{E}_{X_i} [\max\{1, L(X_i)\} |\varepsilon_m| t \mathcal{W}(X_i)] \\
&\leq \varepsilon_m t \mathbb{E}_{X_i} [\max\{1, L^2(X_i)\}]^{1/2} \mathbb{E}_{X_i} [\mathcal{W}^2(X_i)]^{1/2} \\
&\leq \Gamma^{1/2} \left(1 + \int L(x)^2 d\mathbb{P}^X(x)\right)^{1/2} \varepsilon_m.
\end{aligned} \quad (4.14)$$

The above inequalities used the triangle inequality, Lipschitz continuity and the Cauchy-Schwartz inequality in respective order. In the last inequality we used that $t \leq 1$ and $\mathbb{E}[\mathcal{W}^2(X_i)] \leq \Gamma$ irrespective of data used for training from specification 1, and $\max\{a, b\} \leq a + b$ for non-negative a, b . Since $\{\varepsilon_m\}_{m \in \mathbb{N}}$ is a null-sequence, we can assume without loss of generality that $\Gamma^{1/2} (1 + \int L(x)^2 d\mathbb{P}^X(x))^{1/2} \varepsilon_m \leq \delta$ for m large enough. In conjunction with equation (4.14) we $\mathbb{P}$ -almost surely obtain $|F((1 \pm \varepsilon_m)t, \mathcal{W}) - F(t, \mathcal{W})| \leq \delta$. This implies

$$F((1 + \varepsilon_m)t, \mathcal{W}) - \delta \leq F(t, \mathcal{W}) \leq F((1 - \varepsilon_m)t, \mathcal{W}) + \delta \quad \mathbb{P}\text{-a.s.} \quad (4.15)$$

Subtracting equation 4.15 from equation 4.13 yields $\mathbb{P}$ -almost surely

$$\begin{aligned}
 & \widehat{F}\left((1 - \varepsilon_m) \cdot t, \mathcal{W}; I\right) - F\left((1 - \varepsilon_m) t, \mathcal{W}\right) - \delta \\
 & \leq \widehat{F}\left(\widehat{c}_{I, \mathcal{W}} \cdot t, \mathcal{W}; I\right) \mathbb{1}\left(\left|\widehat{c}_{I, \mathcal{W}} - 1\right| < \varepsilon_m\right) - F(t, \mathcal{W}) \\
 & \leq \widehat{F}\left((1 + \varepsilon_m) \cdot t, \mathcal{W}; I\right) - F\left((1 + \varepsilon_m) t, \mathcal{W}\right) + \delta.
 \end{aligned}$$

Taking the absolute value in turn implies:

$$\begin{aligned}
 & \left| \widehat{F}\left(\widehat{c}_{I, \mathcal{W}} \cdot t, \mathcal{W}; I\right) \mathbb{1}\left(\left|\widehat{c}_{I, \mathcal{W}} - 1\right| < \varepsilon_m\right) - F(t, \mathcal{W}) \right| \\
 & \leq \max \left\{ \left| \widehat{F}\left((1 + \varepsilon_m) \cdot t, \mathcal{W}; I\right) - F\left((1 + \varepsilon_m) t, \mathcal{W}\right) \right|, \right. \\
 & \quad \left. \left| \widehat{F}\left((1 - \varepsilon_m) \cdot t, \mathcal{W}; I\right) - F\left((1 - \varepsilon_m) t, \mathcal{W}\right) \right| \right\} + \delta.
 \end{aligned} \tag{4.16}$$

We can now utilise equation (4.16) and the Dvoretzky–Kiefer–Wolfowitz inequality (see Lemma 4) to derive a concentration inequality for C_m :

$$\begin{aligned}
 C_m &= \mathbb{P} \left[\sup_{t \in [0, 1]} \left| \widehat{F}\left(\widehat{c}_{I, \mathcal{W}} \cdot t, \mathcal{W}; I\right) \mathbb{1}\left(\left|\widehat{c}_{I, \mathcal{W}} - 1\right| < \varepsilon_m\right) - F(t, \mathcal{W}) \right| > 2\delta \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq \mathbb{P} \left[\sup_{t \in [0, 1]} \max_{s \in \{-1, +1\}} \left| \widehat{F}\left((1 + s\varepsilon_m) \cdot t, \mathcal{W}; I\right) - F\left((1 + s\varepsilon_m) t, \mathcal{W}\right) \right| > \delta \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq \mathbb{P} \left[\sup_{t \in [0, 1]} \left| \widehat{F}\left((1 + \varepsilon_m) \cdot t, \mathcal{W}; I\right) - F\left((1 + \varepsilon_m) t, \mathcal{W}\right) \right| > \delta \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\quad + \mathbb{P} \left[\sup_{t \in [0, 1]} \left| \widehat{F}\left((1 - \varepsilon_m) \cdot t, \mathcal{W}; I\right) - F\left((1 - \varepsilon_m) t, \mathcal{W}\right) \right| > \delta \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq 4e^{-2|I|\delta^2}.
 \end{aligned} \tag{4.17}$$

We used $\mathbb{P}[\max(X, Y) > a] = \mathbb{P}[\{X > a\} \cup \{Y > a\}] \leq \mathbb{P}[X > a] + \mathbb{P}[Y > a]$ in the third line. Hence, combining equation (4.12) and equation (4.17), we get

$$\begin{aligned}
 A_m &= \mathbb{P} \left[\sup_{t \in [0, 1]} \left| \widehat{F}\left(t, \widehat{c}_{I, \mathcal{W}} \cdot \mathcal{W}; I\right) - F(t, \mathcal{W}([m] \setminus I)) \right| > 3\delta \middle| \mathcal{D}_{[m] \setminus I} \right] \\
 &\leq B_m + C_m \\
 &\leq 32|I|^{-1/2}\Gamma + 4e^{-2|I|\delta^2}.
 \end{aligned} \tag{4.18}$$

And thus

$$\begin{aligned}
 & \mathbb{P} \left[\sup_{t \in [0, 1]} \left| \widehat{F}\left(t, \widehat{c}_{I, \mathcal{W}} \cdot \mathcal{W}; I\right) - F(t, \mathcal{W}([m] \setminus I)) \right| > 3\delta \right] \\
 & \mathbb{E} \left[\mathbb{P} \left[\sup_{t \in [0, 1]} \left| \widehat{F}\left(t, \widehat{c}_{I, \mathcal{W}} \cdot \mathcal{W}; I\right) - F(t, \mathcal{W}([m] \setminus I)) \right| > 3\delta \middle| \mathcal{D}_{[m] \setminus I} \right] \right] \\
 & \leq 32|I|^{-1/2}\Gamma + 4e^{-2|I|\delta^2}.
 \end{aligned} \tag{4.19}$$

Recall that $\delta > 0$ is fixed and $I = I^m$ actually depends on m (omitted for ease of notation). We conclude the proof by noting that $|I^m| \rightarrow \infty$ as $m \rightarrow \infty$ by assumption. $\square$

We have connected F and F_0 to $\widehat{F}$ and $\widehat{F}_0$. We also want to connect $\widehat{F}_0^{BH}$ with F_0^{BH} .

Lemma 7. *Consider the processes from 4.3 and 4.4. Let $\mathcal{W}$ a weight function satisfying specification 1. Fix a single fold $I \subset [m]$. Then:*

$$\widehat{F}_0^{\text{BH}}(t, \hat{c}_{I, \mathcal{W}} \cdot \mathcal{W}; I) = F_0^{\text{BH}}(t, \mathcal{W}) = t.$$

Proof. By specification 1 we have $\mathbb{E}[\mathcal{W}(X_i)] = 1$. We observe

$$\begin{aligned} \widehat{F}_0^{\text{BH}}(t, \hat{c}_{I, \mathcal{W}} \cdot \mathcal{W}; I) &= t \cdot \frac{1}{|I|} \sum_{i \in I} \hat{c}_{I, \mathcal{W}} \cdot \mathcal{W}(X_i) \\ &= t \cdot \frac{|I|}{\sum_{i \in I} \mathcal{W}(X_i)} \cdot \frac{1}{|I|} \sum_{i \in I} \mathcal{W}(X_i) \\ &= t \\ &= t \cdot \mathbb{E}[\mathcal{W}(X_i)] \\ &= F_0^{\text{BH}}(t, \mathcal{W}). \end{aligned}$$

□

In summary, we have related F , F_0 and F_0^{BH} with $\widehat{F}$, $\widehat{F}_0$ and $\widehat{F}_0^{\text{BH}}$ in Lemma 6 and Lemma 7. Relating the empirical processes to their population counterparts guides us a way to go from $\widehat{\text{FDP}}$ to $\widehat{\text{FDP}}^{\text{BH}}$ and $\widehat{\text{FDP}}^{\text{BH}}$ to FDP^{BH} . To go from $\widehat{\text{FDP}}$ to FDP^{BH} we are only missing one piece.

Lemma 8. *Let $((P_i, X_i))_{i \in [m]}$ be from the conditional two-groups model 2.3. Consider a weight function $\mathcal{W}$. Then:*

$$\sup_{t \in [0, 1]} (F_0(t, \mathcal{W}) - F_0^{\text{BH}}(t, \mathcal{W})) \leq o_{\mathbb{P}}(1).$$

Proof. The proof is straightforward:

$$\begin{aligned} F_0(t, \mathcal{W}) &= \mathbb{P}[P_i \leq t\mathcal{W}(X_i), H_i = 0] \\ &= \mathbb{P}[P_i \leq t\mathcal{W}(X_i) | H_i = 0] \cdot \mathbb{P}[H_i = 0] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{P}[P_i \leq t\mathcal{W}(X_i) | H_i = 0] \mid X_i]] \cdot \mathbb{P}[H_i = 0] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{P}[P_i \leq t\mathcal{W}(X_i) \mid H_i = 0, X_i] \mid X_i]] \cdot \mathbb{P}[H_i = 0] \\ &= \mathbb{E}[\mathbb{E}[t\mathcal{W}(X_i) \mid X_i]] \cdot \mathbb{P}[H_i = 0] \\ &= t\mathbb{E}[\mathcal{W}(X_i)] \cdot \mathbb{P}[H_i = 0] \\ &\leq t\mathbb{E}[\mathcal{W}(X_i)] \\ &= F_0^{\text{BH}}(t, \mathcal{W}). \end{aligned}$$

We used $P_i \mid H_i = 0, X_i \sim \text{U}[0, 1]$ in the forth line. Hence, $F_0(t, \mathcal{W}) \leq F_0^{\text{BH}}(t, \mathcal{W})$. The claim immediatly follows. □

Lemma 9. *Let $I \subset \mathbb{R}$ and $\{Y_m^1\}_{m \in \mathbb{N}}, \{Y_m^2\}_{m \in \mathbb{N}}, \{Y_m^3\}_{m \in \mathbb{N}}, \{Y_m^4\}_{m \in \mathbb{N}}$ be sequences of stochastic processes with*

$$\begin{aligned}\sup_{t \in I} |Y_m^1(t) - Y_m^3(t)| &= o_{\mathbb{P}}(1) \text{ as } m \rightarrow \infty, \\ \sup_{t \in I} |Y_m^2(t) - Y_m^4(t)| &= o_{\mathbb{P}}(1) \text{ as } m \rightarrow \infty.\end{aligned}$$

and $\{Y_m^2\}_{m \in \mathbb{N}}, \{Y_m^4\}_{m \in \mathbb{N}}$ being bounded away from zero and bounded:

$$C \geq Y_m^2(t), Y_m^4(t) \geq c > 0 \text{ for } t \in I, m \in \mathbb{N} \quad \mathbb{P} - a.s.$$

Then we have

$$\sup_{t \in I} \left| \frac{Y_m^1(t)}{Y_m^2(t)} - \frac{Y_m^3(t)}{Y_m^4(t)} \right| = o_{\mathbb{P}}(1).$$

Proof. For $t \in I$ we have:

$$\begin{aligned}\left| \frac{Y_m^1(t)}{Y_m^2(t)} - \frac{Y_m^3(t)}{Y_m^4(t)} \right| &= \frac{1}{|Y_m^2(t)Y_m^4(t)|} |Y_m^1(t)Y_m^4(t) - Y_m^3(t)Y_m^2(t)| \\ &\leq c^{-2} |Y_m^1(t)Y_m^4(t) - Y_m^1(t)Y_m^2(t) + Y_m^1(t)Y_m^2(t) - Y_m^3(t)Y_m^2(t)| \\ &= c^{-2} |Y_m^1(t)Y_m^4(t) - Y_m^1(t)Y_m^2(t) + Y_m^1(t)Y_m^2(t) - Y_m^3(t)Y_m^2(t)| \\ &= c^{-2} |Y_m^1(t)(Y_m^4(t) - Y_m^2(t)) + Y_m^2(t)(Y_m^1(t) - Y_m^3(t))| \\ &\leq c^{-2} |Y_m^1(t)| |Y_m^4(t) - Y_m^2(t)| + c^{-2} |Y_m^2(t)| |Y_m^1(t) - Y_m^3(t)| \\ &\leq c^{-2} C |Y_m^4(t) - Y_m^2(t)| + c^{-2} C |Y_m^1(t) - Y_m^3(t)|.\end{aligned}$$

This concludes the proof. $\square$

Lemma 10. Let $((P_i, X_i))_{i \in \mathbb{N}}$ be from the conditional two-groups model 2.3 satisfying regularity Assumption 1. Let $\{I_\ell^m\}_{m \in \mathbb{N}, \ell \in [K]} \subset \mathbb{N}$ be a sequence of folds with $\dot{\cup}_{\ell=1}^K I_\ell^m = [m]$ for all $m \in \mathbb{N}$. Let $\hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}$ be a weight function learned out-of-fold on $[m] \setminus I_\ell^m$. Assume $|I_\ell^m| \rightarrow \infty$ as $m \rightarrow \infty$. Set $t'' := t'/(4\Gamma)$ where t' is defined in assumption 1 and Γ in specification 1. Then

$$\sup_{t \in [t'', 1]} \left| \text{FDP}^{\text{BH}}(t) - \frac{\sum_{\ell=1}^K |I_\ell| / m \cdot F_0^{\text{BH}} \left(t, \hat{\mathcal{W}}^{([m] \setminus I_\ell)} \right)}{\sum_{\ell=1}^K |I_\ell| / m \cdot F \left(t, \hat{\mathcal{W}}^{([m] \setminus I_\ell)} \right)} \right| = o_{\mathbb{P}}(1). \quad (4.20)$$

Proof. We closely follow S2.2 in [IH21]. We observe that the denominator of both quotients in the claims is bounded by 1 from above. We want to apply Lemma 9. In order to do that, we first need to bound the denominator from below and ensure that it does not

vanish. We can do this by bounding $F(t, \mathcal{W}^{([m] \setminus I_\ell^m)})$ away from zero:

$$\begin{aligned}
 F(t, \mathcal{W}^{([m] \setminus I_\ell^m)}) &= \mathbb{P} \left[P_i \leq t \mathcal{W}^{([m] \setminus I_\ell^m)}(X_i) \right] \\
 &= \int F \left(t \mathcal{W}^{([m] \setminus I_\ell^m)}(x) \middle| X_i = x \right) d\mathbb{P}^X(x) \\
 &\geq \int \mathbb{1}(\mathcal{W}^{([m] \setminus I_\ell^m)}(x) > 1/2) F \left(t \mathcal{W}^{([m] \setminus I_\ell^m)}(x) \middle| X_i = x \right) d\mathbb{P}^X(x) \\
 &\geq \int \mathbb{1}(\mathcal{W}^{([m] \setminus I_\ell^m)}(x) > 1/2) F(t/2 | X_i = x) d\mathbb{P}^X(x) \\
 &\geq t/2 \int \mathbb{1}(\mathcal{W}^{([m] \setminus I_\ell^m)}(x) > 1/2) d\mathbb{P}^X(x) \\
 &\geq t/2 \cdot (1/4) \frac{\left(\int \mathcal{W}^{([m] \setminus I_\ell^m)}(x) d\mathbb{P}^X(x) \right)^2}{\int \mathcal{W}^{([m] \setminus I_\ell^m)}(x)^2 d\mathbb{P}^X(x)} \\
 &\geq t/(8\Gamma).
 \end{aligned}$$

We used throughout that $\mathcal{W}^{([m] \setminus I_\ell^m)}$ is independent of X_i . We used monotonicity of $F(\cdot | X_i = x)$ from the third to the fourth line. The fifth line follows from $F(t | X_i = x) \geq t$ in Lemma 1. We used the Paley–Zygmund inequality from Lemma 19 in the penultimate step. The last line follows from iterated expectation and specification 1. For $t \geq t''$ we see

$$\sum_{\ell=1}^K \frac{|I_\ell|}{m} \cdot F \left(t, \mathcal{W}^{([m] \setminus I_\ell)} \right) \geq \sum_{\ell=1}^K \frac{|I_\ell|}{m} \cdot \frac{t}{8\Gamma} = \frac{t''}{8\Gamma}, \quad (4.21)$$

$\mathbb{P}$ -almost surely for all m . We used $\cup_{\ell=1}^K I_\ell^m = [m]$ by assumption. This yields a bound from below for the denominator of the subtraction in (4.20). The denominator of FDP^{BH} fulfills ≥ 1 by definition 4.7. By Lemma 8, the difference of the denominators of the terms in the claim is of order $o_{\mathbb{P}}(1)$ uniformly over t . By Lemma 6, the difference of the numerator of the terms in the claim is of order $o_{\mathbb{P}}(1)$ uniformly over t . In conjunction with Lemma 9, this concludes the proof. $\square$

4.3 Empirical formulation of IHW

Let us introduce some additional notation:

$$\hat{t}^{\text{BH}} := \sup \{ t \in [0, 1] \mid \text{FDP}^{\text{BH}}(t) \leq \alpha \}.$$

In words, $\hat{t}^{\text{BH}}$ is the highest threshold t , such that the FDP^{BH} is still controlled beneath α . Recall that $\hat{t}^* = \sup \left\{ t \in [0, 1] \middle| \frac{\sum_{i=1}^m W_i t}{\sum_{i=1}^m \mathbb{1}(\{P_i \leq W_i t\})} \leq \alpha \right\}$ from the BH algorithm 3. By choice of FDP^{BH} , we have $\hat{t}^{\text{BH}} = \hat{t}^*$. We can now formulate the IHW decision rule as follows:

$$\text{Reject } i \in I_\ell \iff P_i \leq \hat{t}^{\text{BH}} \cdot \hat{c}_{I_\ell, \mathcal{W}^{([m] \setminus I_\ell)}} \cdot \mathcal{W}^{([m] \setminus I_\ell)}(X_i).$$

The numerator of $\text{FDP}^{\text{BH}}(t)$ is continuous. The denominator of $\text{FDP}^{\text{BH}}(t)$ is right-continuous and piecewise constant in t . Thus, by construction of $\hat{t}^{\text{BH}}$ from (4.7) we have

$$\text{FDP}^{\text{BH}}(\hat{t}^{\text{BH}}) \leq \alpha \quad \mathbb{P}\text{-almost surely.} \quad (4.22)$$

To control the FDP^{BH} we need to make sure the denominator does not vanish. We want to work on the event $\{\hat{t}^{\text{BH}} \geq t''\}$ with t'' defined in Lemma 10. To legitimate that we can work on $\{\hat{t}^{\text{BH}} \geq t''\}$, we first need to verify that asymptotically the event $\{\hat{t}^{\text{BH}} \geq t''\}$ holds $\mathbb{P}$ -almost surely.

Lemma 11. *Let $((P_i, X_i))_{i \in \mathbb{N}}$ be from the conditional two-groups model 2.3 satisfying regularity Assumption 1. Let $\{I_\ell^m\}_{m \in \mathbb{N}, \ell \in [K]} \subset \mathbb{N}$ be a sequence of folds with $\dot{\cup}_{\ell=1}^K I_\ell^m = [m]$ for all $m \in \mathbb{N}$. Let $\hat{\mathcal{W}}$ be a weighting scheme 4. Assume $|I_\ell^m| \rightarrow \infty$ as $m \rightarrow \infty$ for all $\ell \in [K]$. Then*

$$\mathbb{P}[\hat{t}^{\text{BH}} \geq t''] = 1 - o(1) \text{ as } m \rightarrow \infty.$$

Proof. We closely follow S2.2 in [IH21]. By the definition of $\hat{t}^{\text{BH}}$ we have

$$\mathbb{P}[\hat{t}^{\text{BH}} \geq t''] \geq \mathbb{P}[\text{FDP}^{\text{BH}}(t'') \leq \alpha]. \quad (4.23)$$

It thus suffices to study FDP^{BH} . Lemma 10 equips us with a tool to study the latter through

$$\frac{\sum_{\ell=1}^K |I_\ell| / m \cdot F_0^{\text{BH}}(t, \hat{\mathcal{W}}([m] \setminus I_\ell))}{\sum_{\ell=1}^K |I_\ell| / m \cdot F(t, \hat{\mathcal{W}}([m] \setminus I_\ell))}.$$

We will first focus on a single weight function, i.e. learned on one fold. For ease of notation we will again write I_ℓ and $\hat{\mathcal{W}}$ instead of I_ℓ^m and $\hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}$ respectively. We first find that

$$\begin{aligned} \int \hat{\mathcal{W}}(x) \mathbb{1}(\hat{\mathcal{W}}(x) \leq 4\Gamma) d\mathbb{P}^X(x) &= \int \hat{\mathcal{W}}(x) (1 - \mathbb{1}(\hat{\mathcal{W}}(x) > 4\Gamma)) d\mathbb{P}^X(x) \\ &= \int \hat{\mathcal{W}}(x) d\mathbb{P}^X(x) \\ &\quad - \int \hat{\mathcal{W}}(x) \mathbb{1}(\hat{\mathcal{W}}(x) > 4\Gamma) d\mathbb{P}^X(x) \\ &= 1 - \int \hat{\mathcal{W}}(x) \mathbb{1}(\hat{\mathcal{W}}(x) > 4\Gamma) d\mathbb{P}^X(x) \\ &\geq 1 - \left(\int \hat{\mathcal{W}}(x)^2 d\mathbb{P}^X(x) \right)^{1/2} \\ &\quad \cdot \left(\int \mathbb{1}(\hat{\mathcal{W}}(x) > 4\Gamma) d\mathbb{P}^X(x) \right)^{1/2} \\ &\geq 1 - \Gamma^{1/2} (\mathbb{P}(\hat{\mathcal{W}}(X) > 4\Gamma))^{1/2} \\ &\geq 1 - \Gamma^{1/2} \left(\frac{\mathbb{E}[\hat{\mathcal{W}}(X)]}{4\Gamma} \right)^{1/2} \\ &= 1 - \Gamma^{1/2} \frac{1}{(4\Gamma)^{1/2}} \\ &= \frac{1}{2}, \end{aligned} \quad (4.24)$$

We used Cauchy-Schwartz's inequality for the first inequality. For the second inequality we used Markov's inequality 16. All other inequalities follow from specification 1. Rearranging

(4.24) gives us

$$\begin{aligned}
 F_0^{\text{BH}}(t'', \hat{\mathcal{W}}) &= t'' \mathbb{E}[\hat{\mathcal{W}}(X)] \\
 &= t'' \\
 &\leq t'' \cdot 2 \int \hat{\mathcal{W}}(x) \mathbb{1}(\hat{\mathcal{W}}(x) \leq 4\Gamma) d\mathbb{P}^X(x) \\
 &= (1/2\Gamma) \int t' \hat{\mathcal{W}}(x) \mathbb{1}(\hat{\mathcal{W}}(x) \leq 4\Gamma) d\mathbb{P}^X(x) \\
 &= (1/2\Gamma) \int_{\hat{\mathcal{W}}(x) \leq 4\Gamma} \frac{t'}{F(t' | X_i = x)} F(t' | X_i = x) \hat{\mathcal{W}}(x) d\mathbb{P}^X(x).
 \end{aligned}$$

With assumption 1 we obtain

$$\begin{aligned}
 F_0^{\text{BH}}(t'', \hat{\mathcal{W}}) &\leq (\alpha'/2\Gamma) \int_{\hat{\mathcal{W}}(x) \leq 4\Gamma} F(t' | X_i = x) \hat{\mathcal{W}}(x) d\mathbb{P}^X(x) \\
 &= 2\alpha' \int_{\hat{\mathcal{W}}(x) \leq 4\Gamma} \frac{\hat{\mathcal{W}}(x)}{4\Gamma} F(t' | X_i = x) d\mathbb{P}^X(x).
 \end{aligned}$$

We proceed by observing

$$\begin{aligned}
 F_0^{\text{BH}}(t'', \hat{\mathcal{W}}) &\leq 2\alpha' \int_{\hat{\mathcal{W}}(x) \leq 4\Gamma} F\left(\frac{\hat{\mathcal{W}}(x)}{4\Gamma} t' \middle| X_i = x\right) d\mathbb{P}^X(x) \\
 &= 2\alpha' \int_{\hat{\mathcal{W}}(x) \leq 4\Gamma} F(t'' \hat{\mathcal{W}}(x) | X_i = x) d\mathbb{P}^X(x) \\
 &\leq 2\alpha' \int F(t'' \cdot \hat{\mathcal{W}}(x) | X_i = x) d\mathbb{P}^X(x) \\
 &= 2\alpha' F(t'', \hat{\mathcal{W}}),
 \end{aligned}$$

where we used concavity of $F(\cdot | X_i = x)$ on $[0, 1]$ and $\frac{\hat{\mathcal{W}}(x)}{4\Gamma} \in [0, 1]$ on the event integrated over. Dividing both sides by $F(t'', \hat{\mathcal{W}})$ yields

$$\frac{F_0^{\text{BH}}(t'', \hat{\mathcal{W}})}{F(t'', \hat{\mathcal{W}})} \leq 2\alpha' < \alpha. \tag{4.25}$$

for every weight function $\hat{\mathcal{W}}$. We used $\alpha' < \alpha/2$ by assumption 1. We reintroduce the notation $\hat{\mathcal{W}}^{([m] \setminus I_\ell)}$ to distinguish between folds. We easily aggregate latter result across folds and see

$$\begin{aligned}
 \frac{\sum_{\ell=1}^K |I_\ell| / m \cdot F_0^{\text{BH}}(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell)})}{\sum_{\ell=1}^K |I_\ell| / m \cdot F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell)})} &= \frac{F_0^{\text{BH}}(t'', \hat{\mathcal{W}})}{F(t'', \hat{\mathcal{W}})} \cdot \frac{\sum_{\ell=1}^K |I_\ell| / m \cdot F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell)})}{\sum_{\ell=1}^K |I_\ell| / m \cdot F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell)})} \\
 &< \alpha.
 \end{aligned}$$

In conjunction with Lemma 10 and observation (4.23) this implies

$$\mathbb{P}[\hat{t}^{\text{BH}} \geq t''] \geq \mathbb{P}[\text{FDP}^{\text{BH}}(t'') \leq \alpha] = 1 - o(1) \text{ as } m \rightarrow \infty.$$

□

We are ready to prove the asymptotic FDR control.

Theorem 3. *Let $((P_i, X_i))_{i \in \mathbb{N}}$ be i.i.d. from the conditional two-groups model 2.3 satisfying regularity Assumption 1. Let $\{I_\ell^m\}_{m \in \mathbb{N}, \ell \in [K]} \subset \mathbb{N}$ be a sequence of folds with $\dot{\cup}_{\ell=1}^K I_\ell^m = [m]$ for all $m \in \mathbb{N}$. Let $\hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}$ be a weight function learned out-of-fold on $[m] \setminus I_\ell^m$. Assume the partition satisfies $|I_\ell^m|/m \rightarrow \gamma_\ell \in (0, 1)$ as $m \rightarrow \infty$ for all $\ell \in [K]$. For any weighting scheme sufficing specification 1, the IHW (algorithm 5) method asymptotically controls the FDR (cf. definition 2).*

Proof. We closely follow S2.2 in [IH21]. Set

$$G_m(t) := \max \left\{ 1, \sum_{\ell \in [K]} |I_\ell^m| \cdot \hat{F}(t, \hat{c}_{I_\ell^m, \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}} \cdot \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}; I_\ell^m) \right\} / m.$$

Note that we have

$$\begin{aligned} G_m(t) \mathbb{1}(t \geq t'') &\geq \sum_{\ell \in [K]} \frac{|I_\ell^m|}{m} \cdot \hat{F}(t, \hat{c}_{I_\ell^m, \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}} \cdot \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}; I_\ell^m) \mathbb{1}(t \geq t'') \\ &\geq \sum_{\ell \in [K]} \frac{|I_\ell^m|}{m} \cdot \hat{F}(t'', \hat{c}_{I_\ell^m, \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}} \cdot \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}; I_\ell^m) \\ &= \sum_{\ell \in [K]} \frac{|I_\ell^m|}{m} \cdot \left(F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}) + o_{\mathbb{P}}(1) \right) \\ &\geq \sum_{\ell \in [K]} \frac{|I_\ell^m|}{m} \cdot F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}) + \sum_{\ell \in [K]} o_{\mathbb{P}}(1) \\ &= \sum_{\ell \in [K]} \frac{|I_\ell^m|}{m} \cdot F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}) + o_{\mathbb{P}}(1). \end{aligned}$$

Next we see from our assumption $|I_\ell^m|/m \rightarrow \gamma_\ell \in (0, 1)$ that

$$\begin{aligned} G_m(t) \mathbb{1}(t \geq t'') &= \sum_{\ell \in [K]} (\gamma_\ell + o(1)) \cdot F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}) + o_{\mathbb{P}}(1) \\ &= \sum_{\ell \in [K]} \gamma_\ell \cdot F(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}) + o_{\mathbb{P}}(1) \\ &\geq \alpha^{-1} \sum_{\ell \in [K]} \gamma_\ell \cdot F_0^{\text{BH}}(t'', \hat{\mathcal{W}}^{([m] \setminus I_\ell^m)}) + o_{\mathbb{P}}(1) \\ &= \alpha^{-1} \sum_{\ell \in [K]} \gamma_\ell \cdot t'' + o_{\mathbb{P}}(1) \\ &\geq \alpha^{-1} t'' K \min_{\ell \in [K]} \gamma_\ell + o_{\mathbb{P}}(1). \end{aligned}$$

The second line used the monotonicity of $\hat{F}$ in the first argument. Line 3 follows from Lemma 6. The penultimate inequality is (4.25). Lemma 5 now gives us

$$G_m(t)^{-1} \mathbb{1}(t \geq t'') = O_{\mathbb{P}}(1). \quad (4.26)$$

We can use this knowledge to study $\text{FDP}(\hat{t}^{\text{BH}})$. As outlined before, we go from $\hat{F}_0$ to F_0 to F_0^{BH} to $\hat{F}_0^{BH}$:

$$\begin{aligned}
& \text{FDP}(\hat{t}^{\text{BH}}) \cdot \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot \hat{F}_0(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot \left(F_0(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) + o_{\mathbb{P}}(1) \right) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot F_0(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') \\
&\quad + \sum_{\ell \in [K]} |I_\ell^m|/m \cdot o_{\mathbb{P}}(1) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot F_0(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') \\
&\quad + \sum_{\ell \in [K]} (\gamma_\ell + o(1)) \cdot o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot F_0(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') + o_{\mathbb{P}}(1).
\end{aligned}$$

We used Lemma 6 in line 3, Lemma 8 in line 4 and Lemma 7 in line 5. Note that the last line correspond to $\widehat{\text{FDP}}^{\text{BH}}$. By carrying along the constraint $\mathbb{1}(\hat{t}^{\text{BH}} \geq t'')$ we ensured that the denominator will remain bounded away from 0. This was used in the penultimate step with equation (4.26). By using the rules for the Landau notation from Lemma 5, we were then able to pull out the $o_{\mathbb{P}}(1)$ from the fraction and absorb it into an added $o_{\mathbb{P}}(1)$ term. With similar arguments we get:

$$\begin{aligned}
& \text{FDP}(\hat{t}^{\text{BH}}) \cdot \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot F_0(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') + o_{\mathbb{P}}(1) \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot F_0^{BH}(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') + o_{\mathbb{P}}(1) \\
&= \sum_{\ell \in [K]} |I_\ell^m|/m \cdot \hat{F}_0^{BH}(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot G_m(t)^{-1} \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') + o_{\mathbb{P}}(1) \\
&= \frac{\sum_{\ell \in [K]} |I_\ell^m| \cdot \hat{F}_0^{BH}(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \cdot \mathbb{1}(\hat{t}^{\text{BH}} \geq t'')}{\max \left\{ 1, \sum_{\ell \in [K]} |I_\ell^m| \cdot \hat{F}(\hat{t}^{\text{BH}}, \hat{c}_{I_\ell^m, \mathcal{W}([m] \setminus I_\ell)} \cdot \hat{\mathcal{W}}([m] \setminus I_\ell); I_\ell^m) \right\}} + o_{\mathbb{P}}(1) \\
&= \text{FDP}^{\text{BH}}(\hat{t}^{\text{BH}}) \cdot \mathbb{1}(\hat{t}^{\text{BH}} \geq t'') + o_{\mathbb{P}}(1) \\
&\leq \alpha \cdot 1 + o_{\mathbb{P}}(1).
\end{aligned}$$

In the last inequality we used (4.22) and Lemma 11. Since we deterministically have

$\hat{F} \leq \hat{F}_0$, we see $0 \leq \text{FDP} \leq 1$ and we can apply the dominated convergence theorem:

$$\begin{aligned}
\limsup_{m \rightarrow \infty} \text{FDR} &= \limsup_{m \rightarrow \infty} \mathbb{E}[\text{FDP}(\hat{t}^{\text{BH}})] \\
&= \mathbb{E}[\limsup_{m \rightarrow \infty} \text{FDP}(\hat{t}^{\text{BH}})] \\
&= \mathbb{E}[\limsup_{m \rightarrow \infty} \text{FDP}(\hat{t}^{\text{BH}}) \mathbb{1}(\hat{t}^{\text{BH}} \geq t'')] \\
&\leq \alpha.
\end{aligned}$$

□

4.4 Discussion

Theorem 3 demonstrates that under some common assumptions from the multiple testing literature on the data generating model and the weighting scheme, algorithm 5 asymptotically controls the FDR. Theorem 3 corresponds to Ignatiadis and Huber [IH21] in 2021. Previously, Ignatiadis et al. [Ign+16] showed in 2016 asymptotic FDR control in theorem 4 under a different set of assumptions which is not typically encountered in the multiple testing literature. Beyond that, Ignatiadis and Huber [IH21] propose a variant of IHW involving censoring for the weighting scheme, called IHWc. With the censoring, IHWc is proven to enjoy finite-sample FDR control:

Theorem 4. *Finite sample control:*

Fix $m \in \mathbb{N}$. Let $((P_i, X_i))_{i \in [m]}$ be i.i.d. from the conditional two-groups model 2.3. Let $\{I_\ell\}_{\ell \in [K]} \subset [m]$ be folds with $\dot{\cup}_{\ell=1}^K I_\ell = [m]$. Consider the IHWc, that is IHW 5 with a weighting scheme $\hat{\mathcal{W}}^{([m] \setminus I_\ell)}$ that depends on the p -values $(P_i)_{i \in [m]}$ only through $(P_i \mathbb{1}(P_i > \tau))_{i \in [m]}$. Then the FDR is controlled in finite samples as

$$\text{FDR}(\text{IHWc}(((P_i, X_i))_{i \in [m]})) \leq \alpha.$$

Proof. This is theorem 1 in [IH21].

□

The censoring idea originates from the Structure Adaptive BH algorithm (SABHA, [LB19b]). However, using only large p -values to compute the weights is suboptimal: large p -values are likely to be associated with true nulls and to be uniformly distributed, so they won't be useful for the weights to properly discriminate between true nulls and true alternatives [Dur19]. This contradicts the goal to gain power compared to BH [Dur19]. This is why, in practice the IHW implementation given in <http://bioconductor.org/packages/IHW> does not incorporate τ -censoring.

Note that in theorem 3 and 4 the p -values $(P_i)_{i \in [m]}$ must be independent. Ignatiadis and Huber [IH21] demonstrate that the idea of cross-weighting applies beyond independence and introduce a variant called IHW-BY:

Theorem 5. *IHW-BY under dependence:*

Fix $m \in \mathbb{N}$ and $\{I_\ell\}_{\ell \in [K]} \subset [m]$ be folds with $\dot{\cup}_{\ell=1}^K I_\ell = [m]$. Let $((P_i, X_i))_{i \in [m]} \in [0, 1] \times \mathcal{X}$, s.t.

- the pairs (P_i, X_i) are independent across folds $\{I_\ell\}_{\ell \in [K]} \subset [m]$, but may be dependent within each fold I_ℓ
- under the null hypothesis, the p -values are uniform and independent of the covariate, i.e. $P_i \mid H_i = 0, X_i \sim \text{U}[0, 1]$

Consider IHW-BY, that is IHW with target level $\alpha / \sum_{k=1}^m \frac{1}{k}$ instead of α . Then the FDR is controlled in finite samples as

$$\text{FDR}(\text{IHW-BY}(((P_i, X_i))_{i \in [m]})) \leq \alpha.$$

Proof. This is theorem 4 in [IH21]. □

IHW-BY is more innocuous regarding dependence structures between p -values, but more conservative and hence leads to less discoveries. It is strongly related to the weighted Benjamini–Yekutieli (wBY) procedure [BY01; BR08].

Chapter 5

IHW-Forest

In this section, we highlight the role of a stratification method for the hypotheses based on the covariates $(X_i)_{i \in [m]}$. We discuss desirable features of the stratification method and first arrive at quantile slicing which is currently the status-quo. Shortcoming of quantile slicing are discussed and an alternative, based on a random forest, is presented.

5.1 Optimization program

Up to this point, we formulated IHW in algorithm 5 with a general weighting scheme. We focused on how cross-validation for the weighting rule protects against type-I error rather than talking about a concrete weighting scheme or addressing power. The mathematical framework leaves considerable flexibility in the choice of the input weighting scheme $\mathcal{W}^{(\cdot)}$. This flexibility gives the analyst the opportunity to use domain-specific as well as statistical knowledge. Nevertheless, it is useful to provide a default for the weighting scheme $\hat{\mathcal{W}}^{(\cdot)}$ that works well across a range of settings. Our starting point for deriving such a weighting scheme $\hat{\mathcal{W}}^{(\cdot)}$ is the equation (3.5). It relies on estimating quantities appearing in model (2.3) and solving an optimization program to find a weight function that optimizes the expected number of discoveries. For multiple testing, the underlying assumption is that $\mathbb{P}[H_i = 0] \approx 1$. Ignatiadis and Huber [IH21] therefore suggest to conservatively set $\hat{\pi}_0^{-\ell} \equiv 1$, as we will do in the following. In light of the cross-folding in IHW 5, the weighting function $\mathcal{W}^{-\ell}$ on fold I_ℓ are only allowed to depend on the out-of-fold data $(P_i, X_i)_{i \notin I_\ell}$. Consequently, the estimators $\hat{F}$ and $\hat{\pi}_0$ is only allowed to depend on $(P_i, X_i)_{i \notin I_\ell}$. We reflect this in our notation, by re-writing the optimization (3.5) to:

$$\begin{aligned} & \underset{s \in S}{\text{maximize}} \quad \frac{1}{m} \sum_{i=1}^m \hat{F}^{-\ell}(s(X_i)|X_i), \\ & \text{subject to} \quad \frac{\frac{1}{m} \sum_{i=1}^m s(X_i)}{\frac{1}{m} \sum_{i=1}^m \hat{F}^{-\ell}(s(X_i)|X_i)} \leq \alpha. \end{aligned} \tag{5.1}$$

We shall have established a full multiple testing procedure, if we specify, how $\hat{F}^{-\ell}$ can be estimated from the pairs $(P_i, X_i)_{i \notin I_\ell}$. One option is to specify $\pi_0(x)$ and $F_{\text{alt}}(\cdot|x)$ in the conditional two-groups model (2.3) parametrically (cf. section 4.1 in [IH21]). However, this suffers from model-misspecification. A non-parametric approach

is desirable. We make the Ansatz of using the empirical cumulative distribution function (ECDF) of the p -values P_i with $i \in I_\ell^c$ with the same covariate X_i :

$$\hat{F}^{-\ell}(t|x) := \frac{1}{|\{i \in I_\ell^c \mid X_i = x\}|} \sum_{\{i \in I_\ell^c \mid X_i = x\}} \mathbb{1}(P_i \leq t)$$

The ECDF is rigorously justified by the Glivenko-Cantelli theorem even in the non-parametric setting. Notably, we have already introduced a concrete convergence rate in Lemma 4. However, for $X_1, \dots, X_m \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$, $\{i \in I_\ell^c \mid X_i = x\}$ will be too small or empty. We need to stratify $X_1, \dots, X_m \in \mathbb{R}^d$ into a small number of bins. Formally, we are looking for a partition $U_1^{-\ell}, \dots, U_G^{-\ell}$ of the covariate space $\mathcal{X}$, s.t. $\dot{\cup}_{g=1}^G U_g^{-\ell} = \mathcal{X}$ for all folds $\ell \in [K]$. We then estimate the ECDF on each bin $U_g^{-\ell}$ separately. For $x \in U_g^{-\ell}$, we are left with a uniform $\hat{F}^{-\ell}(\cdot|x)$ for all tests, whose covariate fall into the same bin $U_g^{-\ell}$:

$$\hat{F}^{-\ell}(t|x) := \hat{F}^{-\ell}(t|U_g^{-\ell}) := \frac{1}{|\{i \in I_\ell^c \mid X_i \in U_g^{-\ell}\}|} \sum_{\{i \in I_\ell^c \mid X_i \in U_g^{-\ell}\}} \mathbb{1}(P_i \leq t). \quad (5.2)$$

By abuse of notation, we write $\hat{F}^{-\ell}(\cdot|U_g^{-\ell})$ for g in $[G]$ to denote the estimated cumulative distribution function on a bin $U_g^{-\ell}$. Equally by abuse of notation, we will write $F(t|U_g^{-\ell}) := \mathbb{E}[\mathbb{1}_{U_g^{-\ell}}(X_i)F(t|X_i)]$. Note that we have $\mathbb{E}[\hat{F}^{-\ell}(\cdot|U_g^{-\ell})] = F(t|U_g^{-\ell})$.

For the implementation, we will actually use the Grenander estimator [Gre56], that is the least concave majorant in t for all x of the ECDF. This has the advantage of turning (5.1) into a convex optimization program. Further, because the Grenander estimator is piecewise linear, (5.1) can be expressed as a linear program. We interpret the resulting procedure as a convex relaxation of (5.1) that makes computation tractable. Since the $L2$ distance between the Grenander estimator of $(t, x) \mapsto F(t|X=x)$ and the ECDF estimator from (5.2) is ideally negligible, we can work with the ECDF estimator for the following discussion.

5.2 Stratification

The weighting function $\mathcal{W}^{-\ell}$ hinges on the partition $U_1^{-\ell}, \dots, U_G^{-\ell}$ in (5.2). As motivated in section 3.2, the hypotheses in the same bin should share similar statistical properties with respect to the multiple testing procedure. The whole idea behind multiple testing procedures based on covariates is that the covariates $X_1, \dots, X_m \in \mathcal{X}$ carry valuable information. Thus, picking the right discretization is crucial.

5.3 Quantile slicing

The current status-quo for IHW is quantile slicing. The idea is to marginally divide the covariates into bins with the same number of observations based on the order on each dimension. Formally, for $X_1, \dots, X_m \in \mathcal{X} \subset \mathbb{R}^d$ we first find $(U_{g,s})_{g \in [G], s \in [d]} \subset \mathbb{R}$, s.t.

- $|\{i \in [m] \mid \text{proj}_{\mathbf{s}} \mathbf{X}_i \in U_{g,s}\}| \approx \text{const.}$ for all $g \in [G], s \in [d]$
- $\dot{\bigcup}_{g=1}^G U_{g,s} = \mathbb{R}$ for all $s \in [d]$
- $U_{1,s} \leq \dots \leq U_{G,s}$ elementwise for all $s \in [d]$

Then for $\mathbf{g} \in \{1, \dots, G\}^s$ define $U_{\mathbf{g}} := \times_{s=1}^d U_{\mathbf{g}_s, s}$.

Note that we are using the covariates from the hold-out fold I_ℓ to construct $U_{\mathbf{g}}$. Hence, the weight $\mathcal{W}^{-\ell}$ also depends on the covariates from the hold-out fold I_ℓ . This is technically not allowed in algorithm 5. However, because we do not use the p -values, we still have honest weights (specification 1 in [IH21]) and FDR control from theorem 1 in [IH21] still holds.

Quantile slicing is a very intuitive choice, because it ensures that $\{i \in [m] \mid X_i \in U_{\mathbf{g}}\}$ is of similar size across all partition bins $U_{\mathbf{g}}$. This is generally beneficial for ECDF (5.2). Furthermore, quantiles directly reflect the data-generating model $\mathbb{P}^X$ of the covariates. So it indirectly picks up heterogeneity among the tests.

However, we point out three caveats. First, the quantile bins $(U_{\mathbf{g}})_{\mathbf{g} \in \{1, \dots, G\}^s}$ do not take the p -values into account and hence cannot capture the heterogeneities among tests for complex data generating mechanism, e.g. when the relationship between P_i and X_i in the conditional two group model (2.3) is complicated.

Second, some covariate dimensions might be more informative than others. In an extreme situation, some coordinates might be not informative at all. This is plausible in practice: in a complex situation the user might not know which data is relevant for the hypothesis under study. Quantile slicing will treat all dimensions equally nonetheless. The partition will be equally granular in each dimension. Quantile slicing is therefore not appropriate for heterogenous covariates.

Lastly, the number of per-covariate quantile combinations explodes for high-dimensional covariates. Due to an exponential increase of the number of bins with the covariates' dimension, quantile slicing leads to many bins with few observations in each. This is bad for the convergence of the ECDF as can be seen in Lemma 4. This ‘‘curse of dimensionality’’ phenomenon is summarised in remark 3.

Remark 3. For $U_{\mathbf{g}}$ it is obvious that $|\{i \in [m] \mid X_i \in U_{\mathbf{g}}\}| \approx \frac{m}{G^d}$. $U_{\mathbf{g}}$ depends on the data $((P_i, X_i))_{i \in [m]}$. On the other hand, the random folds $I_1, \dots, I_K$ from algorithm 5 are not allowed to depend on the data $((P_i, X_i))_{i \in [m]}$. Consequently, $\{i \in I_\ell^c \mid X_i \in U_{\mathbf{g}}^{-\ell}\}$ might be empty for large d . Then the estimator in (5.2) and thus the whole algorithm 5 fails.

As a result, IHW currently does not support multivariate covariates, despite a real interest from the biological community. This has already been pointed out as major drawback by competing methods [LF18]. For a full overview over IHW based on quantile slicing, see algorithm 7.

5.4 Random forest

In summary, quantile slicing does not take full advantage of the data. We want a partition which is more adaptive to the data. To this end, we introduce a random forest based approach, where the leaves of the trees serve as partition bins. Let us first recap relevant concepts and terminology.

Regression trees [Bre+17] are a common non-parametric regression technique, suitable for high-dimensional data and providing a lot of flexibility. They fall into the category of supervised learning, which aims to construct a predictor $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ based on training data $(X_i, Y_i)_{i \in [m]}$ with the explanatory variable $X_i \in \mathcal{X} \subset \mathbb{R}^d$ and response $Y \in \mathcal{Y} \subset \mathbb{R}$. A regression tree first builds a tree structure θ by recursively partitioning the feature space $\mathcal{X}$. In each recursion step, the splitting rule takes the form of two sets $\{proj_s \mathbf{X} \leq v\}$ and $\{proj_s \mathbf{X} > v\}$ for a particular coordinate $s \in \{1, \dots, d\}$ and split point $v \in \mathbb{R}$. The split point v and coordinate s of each recursion step are greedily chosen, s.t. the associated predictor $f_{\hat{\theta}^n}$ of the resulting tree structure $\hat{\theta}^n$ minimizes the empirical risk function in each recursion step:

$$\hat{\theta}^n \in \arg \min_{\theta \text{ tree}} \left\{ \hat{R}_n(f_{\hat{\theta}^n}) \right\}, \quad \hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^n L(Y_i, f(X_i)), \quad (5.3)$$

where $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ is a loss function. Once a partition, controlled by a tuning parameter, becomes small enough, it becomes a leaf node and is no longer partitioned. Let $U_{1,\theta}, \dots, U_{G,\theta}$ denote the final partition bins induced by a tree structure θ . We can obtain them by taking an arbitrary feature variable $X \in \mathcal{X}$, traversing down the tree θ based on the split rules at each node and recording, which leaf node X lands in. If L suffices the triangle inequality and X and $\tilde{X}$ in the same partition bin $U_{j,\theta}$, the optimization in (5.3) indirectly enforces a small $L(f(X), f(\tilde{X}))$. This means that X and $\tilde{X}$ are similar with respect to the response Y . Consequently, regression trees can be viewed as a form of adaptive nearest-neighbor method [PL19]. We can capitalize on that. If the response Y_i is a proxy for a statistical property for hypothesis H_i relevant to the multiple testing procedure, the covariates X_i in the partition bins $U_{1,\theta}, \dots, U_{G,\theta}$ will be homogeneous - as desired. Note that we do not use the predictor f_θ directly and mainly care about the partition $U_{1,\theta}, \dots, U_{G,\theta}$.

It remains to specify, what our response Y_i and the loss function L is. One option proposed by Pospisil and Lee [PL19] is the conditional density $p(y | x)$ with a custom loss function. Here we instead focus on trees that can capture heterogeneity in $\pi_0(x) = \mathbb{P}[H_i = 0 | X_i = x]$, i.e., we define our objective as

$$L(\pi_0, \hat{\pi}_0) = \int_{\mathcal{X}} (\pi_0(x) - \hat{\pi}_0(x))^2 d\mathbb{P}^X(x).$$

Here $\hat{\pi}_0(x)$ is just Storey's [Sto02] estimator π_0 applied to all tests that land in the same leaf as x . It follows from the results of Boca and Leek [BL18] that we may implement the above using off-the-shelf trees for supervised regression problems as follows: let $Y_i = \frac{1(P_i \geq \tau)}{1-\tau}$ for some fixed $\tau \in (0, 1)$. Then we only need to regress $Y_i \sim X_i$.

Lemma 12. *Let $((P_i, X_i))_{i \in [m]}$ be i.i.d. from the conditional two-groups model 2.3. Let $\tau \in (0, 1)$ be the censoring level and set the scaled p -values threshold $Y_i = \frac{\mathbb{1}(P_i \geq \tau)}{1 - \tau}$. Then the estimator Y_i is conservative, i.e.*

$$\pi_0(X_i) \leq \mathbb{E}[Y_i | X_i] = \arg \min_{f \text{ measurable, in } L^2} \mathbb{E}[(Y_i - f(X_i))^2].$$

With equality if and only if $F_{alt|X_i}(\tau) = 1$ $\mathbb{P}$ -almost surely or $\pi_0(X_i) = 1$ $\mathbb{P}$ -almost surely, i.e. if the alternative distribution puts no mass on $(\tau, 1]$ or there are no alternatives.

Proof. The right equation is simply the projection property of the conditional expectation (see for example Korollar 8.17 in Klenke [Kle13]). For the inequality, we follow the proof of Theorem S1 in Boca and Leek [BL18]. Recalling the conditional two-group model (2.3) we have

$$\begin{aligned} (1 - \tau)\mathbb{E}[Y_i | X_i = x] &= \mathbb{P}[P_i > \tau | X_i = x] \\ &= 1 - F(\tau | x) \\ &= 1 - (\pi_0(x)\tau + (1 - \pi_0(x))F_{alt|X_i=x}(\tau)) \\ &= 1 - \pi_0(x)\tau - F_{alt|X_i=x}(\tau) + \pi_0(x)F_{alt|X_i=x}(\tau) \\ &= \pi_0(x) - \tau\pi_0(x) + 1 - \pi_0(x) - F_{alt|X_i=x}(\tau) + F_{alt|X_i=x}(\tau)\pi_0(x) \\ &= (1 - \tau)\pi_0(x) + (1 - F_{alt|X_i=x}(\tau))(1 - \pi_0(x)) \\ &\geq (1 - \tau)\pi_0(x). \end{aligned}$$

Dividing both sides of the remaining equation by $(1 - \tau)$ yields the claim. $\square$

Note that Y_i depends on the p -value P_i and thus the partition $U_{1,\theta}, \dots, U_{G,\theta}$ does as well. This poses the risk of leading to over-fitting (see 4.1). To overcome this, we grow regression trees $\theta^{-\ell}$ regressing $(Y_i^{-\ell})_{i \in I_\ell^c} \sim (X_i)_{i \in I_\ell^c}$ for $\ell \in [K]$ (see line 5 and 6 in algorithm 6).

Remark 4. Considerations for τ :

Large τ give $F_{alt|X_i}(\tau) \approx 1$, as required for equality in Lemma 12. On the other hand, large τ mean that only large p -values P_i are taken into account to construct the partition. As already discussed in section 4.4, this is undesirable, because the small p -values are the most interesting ones. In the extreme case, where $\tau \in [0, 1] \setminus [\min_{i \in [m]} P_i, \max_{i \in [m]} P_i]$, we have $Y_i = \text{const.}$ Then we will only have one partition bin U_1 and the IHW procedure reduces to BH. Boca and Leek [BL18] suggest τ between 0.8 and 0.95. Ultimately, τ which is adaptive to the p -values, without over-fitting would be ideal. This is accomplished in line 3 in algorithm 6, by taking quantiles from $\mathbb{P}^P$.

The construction of the regression tree θ hinges on the choice of τ . Remark 4 provides some guidelines for the choice of τ , but they are far from conclusive. We can avoid to fix a single τ for the algorithm and try different censoring thresholds $\tau_1, \dots, \tau_T$. This will lead to different trees $\theta_1^{-\ell}, \dots, \theta_T^{-\ell}$, different partitions $U_{g,\theta_1^{-\ell}}, \dots, U_{g,\theta_T^{-\ell}}$, different

weighting schemes $\mathcal{W}_{\theta_1^{-\ell}}^{(\cdot)}, \dots, \mathcal{W}_{\theta_T^{-\ell}}^{(\cdot)}$ and finally different weights $(W_{i,t})_{i \in [m], t \in [T]}$. We can average them out with $W_i = \frac{1}{T} \sum_{t=1}^T W_{i,t}$.

In general, simple regression trees as described before suffer from high variance [SSS22]. This means that if we randomly divide the training data into different parts, and grow separate regression trees on each part, the results that we get could be quite different [SSS22]. As will later be motivated in Lemma 13, it is beneficial to reduce the variance of the regression trees. This can be achieved by injecting randomization into the process, an approach called random forests [Bre01].

Specifically, randomization is introduced in two forms [IKL]. First, we can bootstrap, by taking repeated random sub-samples from the (single) training data set [SSS22] and grow a tree on each sub-sample. This is called bagging and has been demonstrated to give considerable improvements in accuracy [SSS22] by reducing the variance as defined above. Second, during the grow stage at each node of the tree, a randomly selected subset of variables is chosen as a candidate for splitting [SSS22]. This is called random feature selection and decorrelates the different trees.

We again average out the trees with $W_i = \frac{1}{T} \sum_{t=1}^T W_{i,t}$. Here $1, \dots, T$ is an enumeration of all trees, which we obtained by considering different censoring levels $\tau_1, \dots, \tau_T$ and the random forest approach. We omit the latter in the pseudo algorithm in 6. It is automatically provided by the “randomForestSRC” R package [IKK22] used in the implementation.

Random forests perform well for mixed-type unordered data and irrelevant features, while being robust to monotonic variable transformations [PL19]. The full IHW-Forest algorithm is given in algorithm 6.

5.5 Alternative motivation

We want to provide a measure for the partition $U_1, \dots, U_G$ in a simplified set-up. It can serve as alternative motivation for the partition method with some theoretical underpinning. The main purpose of the partition in the algorithm will be the estimator $\hat{F}^{-\ell}$ in (5.2), which will then be used in (5.1). We want to examine, how the partition affects the estimator $\hat{F}^{-\ell}$. Lemma 13 provides a criterion.

Lemma 13. *Fix a fold $I \subset [m]$. Let $U_1, \dots, U_G$ be a deterministic partition of $\mathcal{X}$. Consider the estimator $\hat{F}^{-\ell}$ from equation (5.2). Then*

$$\begin{aligned} \mathbb{E} \left[\left(F(t|X_i) - \hat{F}^{-\ell}(t|X_i) \right)^2 \right] &= \sum_{g=1}^G \mathbb{E} \left[\left(\mathbb{E}[\mathbb{1}_{U_g}(X_i)F(t|X_i)] - \hat{F}^{-\ell}(t|U_g) \right)^2 \right] \\ &\quad + \text{Var}[\mathbb{1}_{U_g}(X_i)F(t|X_i)]. \end{aligned}$$

Proof. We begin with a classical variance decomposition:

$$\begin{aligned}
 & \mathbb{E} \left[\left(F(t|X_i) - \hat{F}^{-\ell}(t|X_i) \right)^2 \right] \\
 &= \mathbb{E} \left[\sum_{g=1}^G \mathbb{1}_{U_g}(X_i) \left(F(t|X_i) - \hat{F}^{-\ell}(t|X_i) \right)^2 \right] \\
 &= \sum_{g=1}^G \mathbb{E} \left[\left(\mathbb{1}_{U_g}(X_i) F(t|X_i) - \mathbb{1}_{U_g}(X_i) \hat{F}^{-\ell}(t|U_g) \right)^2 \right] \\
 &= \sum_{g=1}^G \mathbb{E} \left[\left(\mathbb{1}_{U_g}(X_i) F(t|X_i) - \mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] + F(t|U_g) - \mathbb{1}_{U_g}(X_i) \hat{F}^{-\ell}(t|U_g) \right)^2 \right] \\
 &= \sum_{g=1}^G \mathbb{V}ar[\mathbb{1}_{U_g}(X_i) F(t|X_i)] \\
 &\quad + 2\mathbb{E} \left[\left(\mathbb{1}_{U_g}(X_i) F(t|X_i) - \mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] \right) \left(F(t|U_g) - \mathbb{1}_{U_g}(X_i) \hat{F}^{-\ell}(t|U_g) \right) \right] \\
 &\quad + \mathbb{E} \left[\left(F(t|U_g) - \mathbb{1}_{U_g}(X_i) \hat{F}^{-\ell}(t|U_g) \right)^2 \right].
 \end{aligned}$$

In the third line we used, that by abuse of notation $\mathbb{1}_{U_g}(X_i) \hat{F}^{-\ell}(t|X_i) = \mathbb{1}_{U_g}(X_i) \hat{F}^{-\ell}(t|U_g)$ (see 5.1). Define $\mathcal{D}^{-\ell} := ((P_i, X_i))_{i \in [m] \setminus I}$ again as in the proof of lemma 6. Since X_i is independent of $\mathcal{D}^{-\ell}$, we have

$$\mathbb{E} \left[\mathbb{1}_{U_g}(X_i) F(t|X_i) \middle| \mathcal{D}^{-\ell} \right] = \mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] \quad (5.4)$$

and $\hat{F}^{-\ell}(t|U_g)$ measurable in $\mathcal{D}^{-\ell}$. This gives us:

$$\begin{aligned}
 & \mathbb{E} \left[\left(\mathbb{1}_{U_g}(X_i) F(t|X_i) - \mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] \right) \left(F(t|U_g) - \hat{F}^{-\ell}(t|U_g) \right) \right] \\
 &= \mathbb{E} \left[\mathbb{E} \left[\left(\mathbb{1}_{U_g}(X_i) F(t|X_i) - \mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] \right) \left(\mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] - \hat{F}^{-\ell}(t|U_g) \right) \middle| \mathcal{D}^{-\ell} \right] \right] \\
 &= \mathbb{E} \left[\mathbb{E} \left[\left(\mathbb{1}_{U_g}(X_i) F(t|X_i) - F(t|U_g) \right) \middle| \mathcal{D}^{-\ell} \right] \left(\mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] - \hat{F}^{-\ell}(t|U_g) \right) \right] \\
 &= \mathbb{E} \left[\left(\mathbb{E} \left[\mathbb{1}_{U_g}(X_i) F(t|X_i) \middle| \mathcal{D}^{-\ell} \right] - F(t|U_g) \right) \left(\mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] - \hat{F}^{-\ell}(t|U_g) \right) \right] \\
 &= \mathbb{E} \left[0 \cdot \left(\mathbb{E}[\mathbb{1}_{U_g}(X_i) F(t|X_i)] - \hat{F}^{-\ell}(t|U_g) \right) \right] \\
 &= 0.
 \end{aligned}$$

Thus the middle term in the last line of (5.5) vanishes. This concludes the proof. $\square$

Note that Lemma 13 assumes a deterministic partition. For a data-driven partition, the proof would fail in equation (5.4). Nonetheless, Lemma 13 provides us with the means to develop a very intuitive criterion for the partition bins from a mathematical framework. This can be useful for the parameter choices in the actual implementation.

In conjunction with Lemma 4, it motivates us to fix a $m_{\min} \in \mathbb{N}$ and seek a partition with $|\{i \in [m] \setminus I \mid \mathcal{D}(X_i) = g\}| \geq m_{\min}$ for all $g \in [G], \ell \in [K]$. We can think of $m_{\min}$ as a budget. We therefore propose to solve the following optimization problem:

$$\begin{aligned}
& \underset{U_1, \dots, U_G}{\text{minimize}} && \int \sum_{g=1}^G \mathbb{V}ar[\mathbb{1}_{U_g}(X)F(t|X)] d\mathbb{P}^P(t), \\
& \text{subject to} && |\{i \in [m] \setminus I \mid X_i \in U_g\}| \geq m_{\min} \text{ for all } g \in [G] \text{ and } \ell \in [K], \\
& \text{and} && \dot{\bigcup}_{g=1}^G U_g = \mathcal{X}.
\end{aligned} \tag{5.5}$$

In other words: The partition bins should hold enough observations and should capture variation of $F(t|X)$ as effectively as possible. In particular, a regression tree should not split along covariates without variation in terms of $F(t|X)$.

Chapter 6

Application and simulation

To illustrate the merits of IHW-Forest and differences to IHW-Quantile and other published multiple testing procedures, we run simulations and a bioinformatic application. We compare the results based on FDR, power and number of rejections.

6.1 Simulation

Our goal in this section is to corroborate the following claims: First, IHW-Forest enjoys finite FDR control in practice. Note that the FDR control from theorem 3 is of asymptotic nature and theorem 4 requires further assumptions (“ τ -censoring”), which the actual IHW implementation does not fulfil. Second, IHW-Forest is still applicable and is powerful for high-dimensional covariates. Lastly, for multiple testing scenarios with sparse signal, IHW-Forest is competitive in terms of power with other published methods which improve power by including covariate information. Our main performance measures are the FDR and power. The expectation is evaluated through averaging over Monte Carlo replicates. We will do this through simulations of two important: multiple testing with high-dimensional and heterogenous covariate and multiple testing with sparse signal.

We compare the following five methods:

1. The **BH** [BH95] as in algorithm 1 solely operates on $(P_i)_{i \in [m]}$, ignoring the covariates $(X_i)_{i \in [m]}$.
2. The **Clfdr-EM** [CS09] method, which rejects hypotheses if they have a small conditional local fdr, $\text{fdr}(P_i|X_i) := \mathbb{P}[H_i = 0|X_i, P_i]$. A parametric model for the conditional two-group model is assumed and its parameters estimated by maximum likelihood.
3. **IHW-Quantile** as described in section 5.3 and in algorithm 7, where the the covariates are binned based on their per-dimension quantile and then the IHW procedure 5 is applied.
4. **IHW-Forest** as described in section 5.4 and algorithm 6, where the the covariates with a random forest approach and then the IHW procedure 5 is applied.

5. **AdaPT**, [LF18] as implemented in accompanying the CRAN package “adaptMT”. In each iteration a parametric approximation of the data generating model is fitted, a rejection threshold is proposed, and then a subset of the hypotheses is “masked”, i.e. the algorithm only has access to $\min\{P_i, 1 - P_i\}$ instead of P_i . AdaPT is not a wBH procedure, its whole philosophy is totally different and relies on symmetry properties of the true null distribution of the p -values by using an estimator of the FDP, different than the one implicitly used in BH-like methods [Dur19].

All of the above methods have ordinal covariates in common and do not assume oracle knowledge of the statistical model. We do not consider multiple testing procedures which are based on categorical covariates.

6.1.1 High-dimensional, heterogenous covariate

We first consider high-dimensional, heterogenous covariates $X \in \mathbb{R}^d$. In each simulation, we generate $(P_i, H_i, X_i), i = 1, \dots, 5000$, independently, as follows:

$$\begin{aligned} X_i &\sim \text{U}[0, 1]^d, \\ \pi_1(X) &= \mathbb{1}(X^1 \leq 0.5)0.9., \\ P_i \mid H_i = 0, X_i &\sim \text{U}[0, 1], \\ P_i \mid H_i = 1, X_i &\sim \text{Beta}(0.25, 1), \\ H_i \mid X_i &\sim \text{Bernoulli}(\pi_1(X_i)). \end{aligned} \tag{6.1}$$

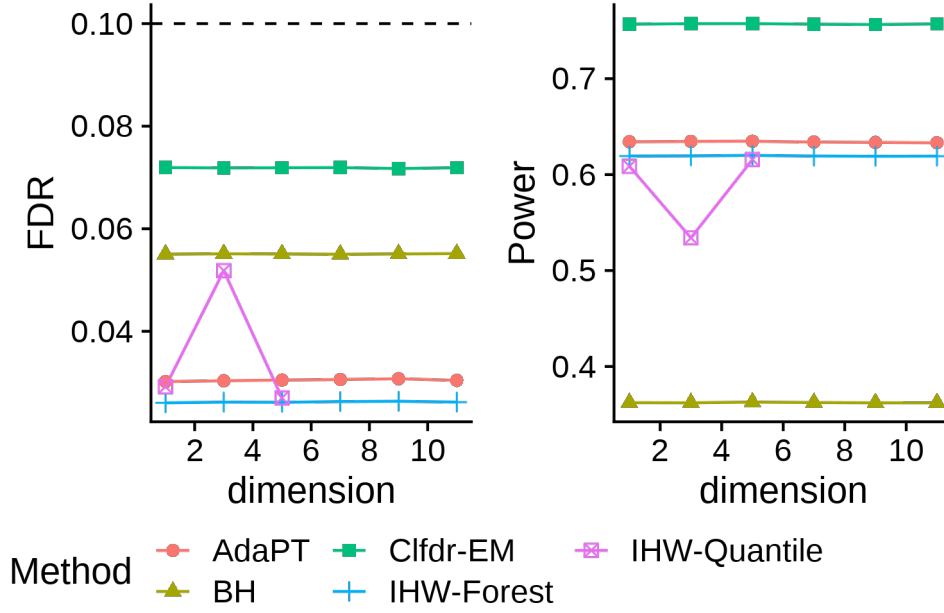
Here, X^1 refers to the projection of $X \in \mathbb{R}^d$ on the first coordinate component. In words, the covariates are uniformly distributed on the d -dimensional hypercube. However, because the hypotheses H_i only depend on the first component of X_i , the signal for the multiple testing procedure lives on an one-dimensional manifold. This is a common working hypotheses in machine learning for high-dimensional data [Cay05; FMN16]. All dimensions but the first are just noise. Therefore, the covariate dimensions are heterogeneous.

Results are shown in figure 6.1. All methods have FDR much below target. IHW-Quantile is no longer applicable as the number of dimensions increases, because the bins are too small (see remark 3). All other methods are applicable and have stable FDR and power. Among these, Clfdr-EM is most powerful, then AdaPT, closely followed by IHW-Forest. All of these procedures improve substantially upon BH.

6.1.2 Sparse signal

In this section, we explore a setting where the signal is in a very small region in terms of the covariate X_i . We seek to compare the methods with an emphasis on how they are able to detect this sparse signal. We will later motivate the relevance in a biological example. We simulate independent $(X_i, H_i, P_i), i = 1, \dots, 50000$ from the conditional two-groups model (2.3) with the following choices for $\mathbb{P}^X, \pi_0(x)$ and $F_{\text{alt}}(\cdot \mid X_i = x)$:

Figure 6.1: Noise sim



$$\begin{aligned} \mathbb{P}^X &\sim U[0, l], \quad \pi_1(X) = \mathbb{1}(X \leq 1)0.9, \\ F_{\text{alt}}(\cdot | X_i = x) &= \text{Beta}(0.25, 1). \end{aligned} \quad (6.2)$$

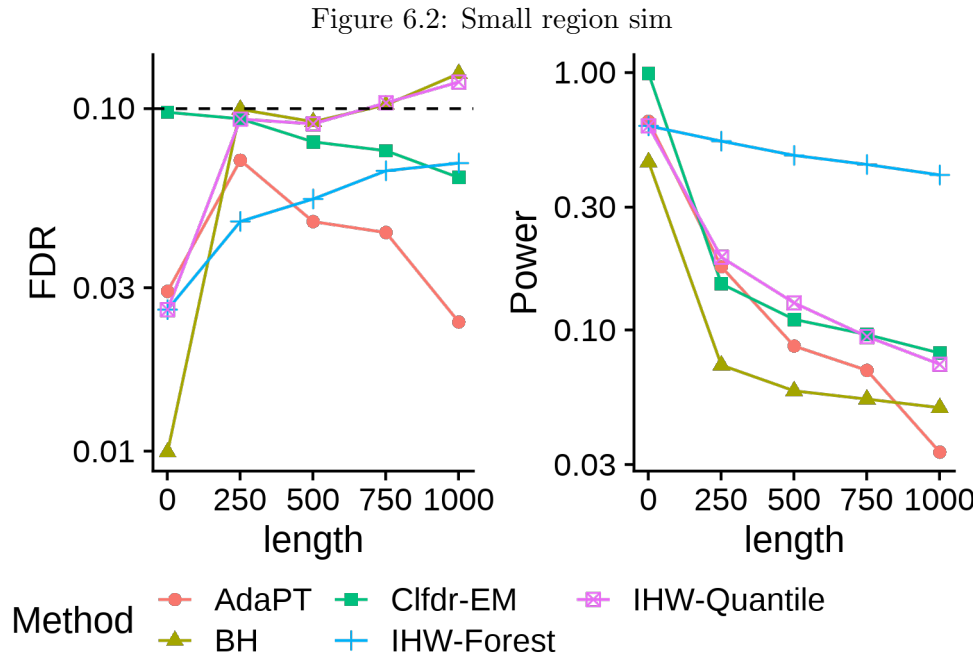
In words: The covariates X_i are uniformly drawn from an interval with length $l \in [1, 1001]$ varying across simulation settings. As the length l increases, the overall signal decreases with $\mathbb{E}[\pi_1(X)] = \frac{0.9}{l}$. Conditioned on this relatively shrinking region, we have a very strong signal $\mathbb{E}[\pi_1(X) | X \in [0, 1]] = 0.9$. Therefore we should upweight/ focus on $\{X \in [0, 1]\}$.

The results are shown in figure 6.2. All methods control FDR. We observe in terms of power that IHW-Forest outperforms all other methods by a large margin. Note the logarithmic scale. Hence, at the end with the sparsest signal, IHW-Forest has three-times the power compared to the next best. The power of all methods decreases as the signal region shrinks. BH, IHW-Forest, IHW-Quantile, Clfdr power decreases exponentially with the interval length l (again, logarithmic scale). However, we see that as l increases, AdaPT's power deteriorates, quickly dropping below BH.

Figure 6.1 shows that AdaPT is very powerful in this particular setting. Since AdaPT is known to be powerful in general [Kor+19], the behaviour in 6.2 may come as a surprise. However, there this phenomenon aligns with theoretical properties of the AdAPT algorithm. In each iteration step of the AdaPT algorithm, it must either reject $1/\alpha$ hypotheses or more or none at all [IH21]. So in situations with very small signal, AdaPT's power can be diminished.

6.2 Application example: biological high-throughput data

As alluded to in the introduction, large-scale multiple testing is omni-present in biology. Here we revisit the hQTL analysis from Ignatiadis and Huber [IH21]. The main points



of deviation are an automated choice of the partition and more covariates in the multiple testing procedure.

6.2.1 Biological context of hQTL-analysis

The human genome is a sequence of nucleic acids human, encoded as approximately 3.2 billion base-pairs of DNA in the nucleus of cells [Nur+22]. We refer to a particular position as locus (pl. loci). The DNA is composed of four nucleotides: adenine (A), guanine (G), cytosine (C) or thymine (T) [Alb+14]. The nucleotide at each locus is customarily represented by those four possible values (“G”, “A”, “T”, “C”).

The vast majority of loci of the human genome have the same nucleotide for most humans. As of 2015, the typical difference between an individual’s genome and the reference human genome was estimated at 20 million base pairs (or 0.6% of the total of 3.2 billion base pairs) [15].

Loci, with significant variations in a population, are called single nucleotide polymorphism (abbreviated SNP, pronounced snip). For example, at a specific locus, the A nucleotide may appear in a majority individuals, but in a few individuals, the locus is occupied by an G. This means that there is an SNP at this specific locus [Mon+17]. For a locus to be qualified as an SNP, the substitution at the locus is customarily required to be present in a sufficiently large fraction of the population (e.g. 1% or more) [15]. SNPs can have high functional importance [KTC08]. They can pinpoint differences in our susceptibility to a wide range of diseases [CGS14] and the severity of illness [HLP21]. The 1000 Genomes project identified 84.7 million SNPs [15].

Each human cell has about 1.8 meters of DNA if completely stretched out. However, the DNA is densely coiled into structures called chromosomes. To condense the DNA molecule, histones - a specific kind of protein- provide a higher order structure for the

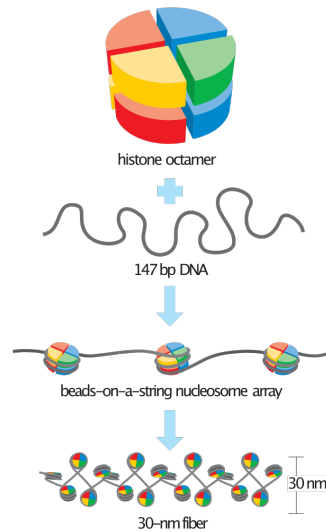


Figure 6.3: Adapted from Crosby (2007) [Cro07].

DNA inside the cell nuclei. They act as spools, serving as structural units for the DNA to wrap around [You06; Leh+05].

When wrapped around histones, the DNA length is reduced to about 90 micrometers (0.09 mm) of 30 nm diameter tightly packed chromatin fibers [Kie+98]. Histones make up the basic building blocks of chromosomes. They may be biochemically modified by bonding with other molecules. Those modifications can occur as a result of lifestyle and other environmental exposure. As an example, addiction can modify the abundance of specific histone modifications [MW10; HL14; RN11]. These effects of environmental factors are known as epigenetic changes. Histone modification can regulate gene expression, by altering the histone state and modifying DNA accessibility. [Car17]. One such biochemical modification of DNA-associated molecules is called H3K27ac.

H3K27ac is usually associated with the higher activation of transcription of genes and hence gene expression [SV12]. The presence or absence of a histone modification at each genomic location can be assessed with the ChIP-sequencing method [LS05]. The result of the ChIP-sequencing is to be understood as a local snapshot in time, since (as explained above) the level of histone modification may change over time. Regions with statistically significant levels of enrichment of histone modification are called “peaks” [Zha+08].

Recent evidence suggests that understanding the interaction between information encoded in the DNA sequence (such as SNPs) and epigenetic information “may hold the key to a better understanding of various pathological conditions” [CWR11]. This could lay the foundation for predicting disease susceptibility and drug response. Consequently, better understanding of both the genetic and epigenetic components of biological information may be an important tool to guide therapeutic intervention in the future of personalized medicine.

A “histone Quantitative Trait Locus” analysis (hQTL-analysis) aims to decipher the relationship between SNPs and histone modification [Gru+15]. For that, a hypothesis test for each pair of SNP and peak is carried out, yielding a large-scale multiple testing task.

Since associations across chromosomes are biologically unlikely, some hQTL analyses are restricted to SNPs and peaks within the same chromosomes [IH21]. However on a single chromosome, SNPs are often correlated, resulting in confounding variables in the hQTL analysis. Hence hypotheses for the same peak and different, but correlated SNPs will be strongly interdependent [IH21].

6.2.2 Previous work

Grubert et al. [Gru+15] analysed cell lines derived from 75 human individuals. For each individual they determined, which nucleotide is present at the SNP. The loci of the SNPs are from an independent database and are the same for all individuals. Furthermore, Grubert et al. called peaks of H3K27ac levels. Note that peaks and their genomic regions are specific for each individual and experiment. They merged peaks across all individuals, and extracted signal in each peak region for each individual. Ignatiadis and Huber [IH21] tested all within-chromosome associations in Chromosomes 1 and 2 using the software Matrix eQTL [Sha12] on the data from Grubert et al. [Gru+15]. For Chromosomes 1 and 2, Grubert et al. [Gru+15] reported the status of $N_1 = 645,452$ and $N_2 = 699,343$ SNPs and the H3K27ac levels at $K_1 = 12,193$ and $K_2 = 11,232$ peaks on these chromosomes. This amounts to nearly 16 billion tests ($m = N_1 \times K_1 + N_2 \times K_2 \approx 1.6 \cdot 10^{10}$) [IH21].

As explained above in subsection 6.2.1, the hypotheses in this multiple testing task are potentially highly dependent. Classical BH and IHW (5) require independence for guaranteed FDR control (1 3). As given in theorem 5, we can use IHW-BY, a variant of IHW, where only the last line in (5) is changed.

To increase power for this multiple testing task compared to wBY, [IH21] used IHW-BY. The genomic distance between the SNP and peak of the individual hypothesis served as covariate for IHW. The genomic distance ranged between 0 base pairs (bp) and $2.49 \cdot 10^8$ bp. This is as an ordered covariate. As explained in section 5.2, the covariate needs to be partitioned into a finite number of bins to apply IHW or IHW-BY. For that [IH21] manually specified intervals for the partition bins.

Since computing and storing 16 billion p -values puts notable demands on computing infrastructure, the implementation of Matrix eQTL [Sha12] only reports p -values below some threshold (e.g. in this case, below 10^{-4}) to reduce storage requirements. In this example, we effectively only have access to 0.015% of the p -values. Hence, the p -values available are right-censored and no longer uniformly distributed on $[0, 1]$ under the null, as required for BH/BY and IHW-BH/IHW-BY. To retain a valid multiple testing procedure, [IH21] separately calculates the genomic distance for all pairs of SNPs and peaks, counts the number of hypotheses m_{groups} in each of manually specified bins and feeds them to BH/BY and IHW-BH/IHW-BY.

6.2.3 new hQTL analysis - covariates

For the new hQTL analysis, we used a total of 11 different covariates instead of a single one.

IHW-BY-Quantile	IHW-BY-Forest	formula
12379	12903	pvalue $\sim$ dist
8781	9048	pvalue $\sim$ minor_allele_freq
9861	13249	pvalue $\sim$ dist + minor_allele_freq
—	13149	pvalue $\sim$ dist + minor_allele_freq + H3K27Ac_1 + H3K27Ac_2 + H3K27Ac_3 + H3K4me3_1 + H3K4me3_2 + H3K4me3_3 + kH3k4me1_1 + kH3k4me1_2 + kH3k4me1_3 + DNaseI

Table 6.1: Number of rejections for IHW with automated stratification using different subset of the available covariates. As explained in remark 3, IHW-BY-Quantile fails with too many covariates.

As [IH21], we use the genomic distance. This covariate is motivated from biological domain knowledge: associations across shorter distances are a priori more plausible and empirically more frequent [IH21]. This is because regions that are proximal in genomic distance are also more likely to physically interact [Gru+15].

For a fixed population, SNPs can be assigned a minor allele frequency (MAF)—the lowest allele frequency at a locus that is observed in a particular population [Zhu+15]. For example, at an SNP, the G nucleotide might appear in 90% of the population and the A nucleotide in 10% of the population. Then the MAF of that SNP would be 10%. We obtained the MAF for each SNP with rtracklayer [LGC09]. For SNPs, where the MAF was not available in rtracklayer, we conservatively used 0 as MAF. This was the case for 6.38% of the hypotheses on Chromosome 1 and 5.1% on Chromosome 2.

We obtained data from other ChIP-sequencing experiments via the UCSC Genome Browser [Ken+02]. We had three data sets for each of the histone modifications H3K27Ac, H3K4me3 and kH3k4me1. The data is published in [Ram+11]. For more details, see here. Each peak had a genomic range and a mean signal value across the interval. We mapped the peaks to the closest SNPs in terms of genomic distance using [Acr19] and annotated the corresponding hypotheses with the mean signal.

6.2.4 new hQTL analysis - results

We apply IHW-BY-Forest and IHW-BY-Quantile using different subsets of the covariates described in subsection 6.2.3. Compared to the analysis in [IH21], we are using more covariates and employ an automated choice of the partition instead of manually specifying partition bins. As described in 6.2.2, the p -values available are right censored and have $P_i | H_i = 0 \sim U[0, 10^{-4}]$. Ignatiadis and Huber [IH21] handle this by separately calculating the number of hypotheses in each bin m_{groups} on the manually specified bins. Due to limited storage capacity, this is not a feasible option for the integrated, automated stratification methods. Instead we artificially inflate the p -values and work with $10^4 \cdot P_i$, s.t. $10^4 \cdot P_i | H_i = 0 \sim U[0, 1]$. This does have the drawback that we are only working with a fraction of the available covariates and hence IHW has less input to learn the features of the data.

In [IH21] with separately specified m , BY yielded 9110 rejections. For the new analysis

with the inflated p -values the number of rejections are listed in table 6.1. IHW-Forest consistently yielded more rejections than IHW-Quantile. From biological domain knowledge, it can be assumed that the genomic distance is a more informative covariate than MAF. In line with that, both IHW-Quantile and IHW-Forest have more discoveries when provided the genomic distance than when provided with MAF. When both genomic distance and MAF are given, IHW-Forest achieves the highest number of rejections of all. This suggests that complex interactions between the information in the genomic distance and MAF are unveiled, when tackled with the right stratification method. On the other hand, quantile binning is not able to take advantage of those interactions appropriately and number of discoveries drops below just considering the genomic distance.

Appendix A

Proofs

We omitted some Lemmas in the main text, because they are very general statements and would disturb the flow of the proofs in this thesis.

Lemma 14. *Let $a, b > 0$ and $Y \sim \text{Beta}(a, b)$. Then:*

$$\begin{aligned}\mathbb{E}[Y] &= \frac{a}{a+b}, \\ \mathbb{V}ar[Y] &= \frac{ab}{(a+b)^2 \cdot (a+b+1)}.\end{aligned}$$

Proof. Recall the Beta distribution with parameters $a, b > 0$ has the density function

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a, b)},$$

for $x \in [0, 1]$ with

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1} dx.$$

and Γ being the Gamma function. The Gamma function fulfills the identity $\Gamma(t+1) = t\Gamma(t)$. For the first part of the claim we can use the identity $\Gamma(t+1) = t\Gamma(t)$:

$$\begin{aligned}\mathbb{E}[Y] &= \frac{\int_0^1 x \cdot x^{a-1}(1-x)^{b-1} dx}{B(a, b)} \\ &= \frac{B(a+1, b)}{B(a, b)} \\ &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{a}{a+b}.\end{aligned}\tag{A.1}$$

For the second part we first note

$$\begin{aligned}
 \mathbb{E}[Y^2] &= \frac{\int_0^1 x^2 \cdot x^{a-1}(1-x)^{b-1} dx}{B(a, b)} \\
 &= \frac{B(a+2, b)}{B(a, b)} \\
 &= \frac{B(a+2, b)}{B(a+1, b)} \cdot \frac{B(a+1, b)}{B(a, b)} \\
 &= \frac{a+1}{a+b+1} \cdot \frac{a}{a+b}.
 \end{aligned}$$

In the last step we used (A.1). Overall we deduce

$$\begin{aligned}
 \text{Var}[Y] &= \mathbb{E}[Y]^2 - \mathbb{E}[Y^2] \\
 &= \left(\frac{a}{a+b} \right)^2 - \frac{a+1}{a+b+1} \cdot \frac{a}{a+b} \\
 &= \left(\frac{a}{a+b} \right) \left(\frac{a}{a+b} - \frac{a+1}{a+b+1} \right) \\
 &= \left(\frac{a}{a+b} \right) \left(\frac{a(a+b+1) - (a+b)(a+1)}{(a+b)(a+b+1)} \right) \\
 &= \left(\frac{a}{a+b} \right) \left(\frac{a^2 + ab + a - a^2 - ab - a - b}{(a+b)(a+b+1)} \right) \\
 &= \frac{ab}{(a+b)^2 \cdot (a+b+1)}.
 \end{aligned}$$

□

Lemma 15. *Let $\{Y_m\}_{m \in \mathbb{N}}$ be a sequence of random variables stochastically converging to 0, i.e. $\forall \varepsilon, \delta > 0 \exists n_{\varepsilon, \delta} : \forall m \geq n_{\varepsilon, \delta} : \mathbb{P}(|Y_m| > \varepsilon) < \delta$. Then there exists a positive null-sequence $\{\varepsilon_m\}_{m \in \mathbb{N}}$, s.t. $\mathbb{P}(|Y_m| > \varepsilon_m) = o(1)$. Without loss of generality we can pick $\varepsilon_m \geq m^{-1/4}$.*

Proof. By definition of stochastic convergence, there exists a subsequence $\{m_k\}_{k \in \mathbb{N}}$ with $\mathbb{P}(|Y_m| > \frac{1}{k}) < \frac{1}{k}$ for $n \geq m_k$. $\{m_k\}_{k \in \mathbb{N}}$ is a non-decreasing sequence going to infinity. Consequently, for any $m \in \mathbb{N}$ we find exactly one k with $n \in \{m_k, \dots, m_{k+1} - 1\}$. Define $\{\varepsilon_m\}_{m \in \mathbb{N}}$ with $\varepsilon_m := \frac{1}{k}$ for $m \in \{m_k, \dots, m_{k+1} - 1\}$. Then $\{\varepsilon_m\}_{m \in \mathbb{N}}$ is a positive null-sequence. We want to show that $\mathbb{P}(|Y_m| > \varepsilon_m)$ is indeed a null-sequence. For $k \in \mathbb{N}$ and $m \geq m_k$, we find $j \geq k$ with $m \in \{m_j, \dots, m_{j+1} - 1\}$ and therefore $\mathbb{P}(|Y_m| > \varepsilon_m) < \frac{1}{j} \leq \frac{1}{k}$. Since $k \in \mathbb{N}$ was arbitrary, we see $\mathbb{P}(|Y_m| > \varepsilon_m) = o(1)$. Finally, choose $\tilde{\varepsilon}_m := \max(\varepsilon_m, m^{-1/4})$. Then we still have $\mathbb{P}(|Y_m| > \tilde{\varepsilon}_m) = \mathbb{P}(|Y_m| > \varepsilon_m, |Y_m| > m^{-1/4}) \leq \mathbb{P}(|Y_m| > \varepsilon_m) = o(1)$, but we also have $\tilde{\varepsilon}_m \geq m^{-1/4}$. □

Lemma 16. Markov's inequality:

Let X be a nonnegative random variable and $a > 0$. Then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. see for example Klenke [Kle13]. □

Lemma 17. Chebyshev's inequality:

Let X be a square-integrable random variable and $a > 0$. Then:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

Proof.

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) = \mathbb{P}((X - \mathbb{E}(X))^2 \geq a^2) \leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

We used the Markov's inequality 16 for the inequality □

Lemma 18. Jensen's inequality:

Let X be a integrable real random variable and φ a convex function. Then:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

Proof. see [Kle13]. □

Lemma 19. Paley–Zygmund inequality:

If $Z \geq 0$ $\mathbb{P}$ -almost surely is a random variable with finite variance, and if $0 \leq \theta \leq 1$, then

$$\mathbb{P}(Z > \theta \mathbb{E}[Z]) \geq (1 - \theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

Proof. We can separate $\mathbb{E}[Z]$ into

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}[Z \mathbb{1}(Z \leq \theta \mathbb{E}[Z])] + \mathbb{E}[Z \mathbb{1}(Z > \theta \mathbb{E}[Z])] \\ &\leq \theta \mathbb{E}[Z] + \mathbb{E}[Z^2]^{1/2} \mathbb{E}[\mathbb{1}(Z > \theta \mathbb{E}[Z])]^{1/2} \\ &= \theta \mathbb{E}[Z] + \mathbb{E}[Z^2]^{1/2} \mathbb{P}(Z > \theta \mathbb{E}[Z])^{1/2}. \end{aligned}$$

where we used the Cauchy-Schwarz inequality. Rearranging, we conclude

$$(1 - \theta) \mathbb{E}[Z] \leq \mathbb{E}[Z^2]^{1/2} \mathbb{P}(Z > \theta \mathbb{E}[Z])^{1/2}.$$

Since $0 \leq \theta \leq 1$ and $Z \geq 0$ $\mathbb{P}$ -almost surely, both sides are non-negative and squaring both sides preserves the inequality. Division by $\mathbb{E}[Z^2]$ then yields the desired inequality. □

Appendix B

Full IHW algorithms

Algorithm 6: IHW-Forest	
1	Randomly partition $[m]$ into $I_1, \dots, I_K$
2	for $l = 1, \dots, K$ do
3	Choose $\tau_1^{-\ell}, \dots, \tau_T^{-\ell} \in (0, 1)$ based on $(P_i)_{i \in I_\ell^c}$ (e.g. quantiles)
4	for $t = 1, \dots, T$ do
5	Set $Y_{i,t}^{-\ell} = \frac{\mathbb{1}(P_i \geq \tau_t^{-\ell})}{1 - \tau_t^{-\ell}}$ for $i \in I_\ell^c$
6	Grow regression tree $\theta_t^{-\ell}$ for $(Y_{i,t}^{-\ell})_{i \in I_\ell^c} \sim (X_i)_{i \in I_\ell^c}$
7	Let $G_{\theta_t^{-\ell}}$ be the number of leaves in $\theta_t^{-\ell}$
8	Set $U_{g, \theta_t^{-\ell}} = \{x \in \mathcal{X} \mid x \text{ lands in } g\text{-th leaf in } \theta_t^{-\ell}\}$
9	Set
	$\tilde{X}_{i, \theta_t^{-\ell}} = \begin{cases} 1 & \text{if } X_i \in U_{1, \theta_t^{-\ell}} \\ \dots & \\ G_{\theta_t^{-\ell}} & \text{if } X_i \in U_{G_{\theta_t^{-\ell}}, \theta_t^{-\ell}} \end{cases}$
10	for $g = 1, \dots, G_{\theta_t^{-\ell}}$ do
11	Estimate $F(\cdot \mid g) : t \mapsto \mathbb{P}(P_i \leq t \mid \tilde{X}_{i, \theta_t^{-\ell}} = g)$ by $\hat{F}_{\theta_t^{-\ell}}^{-\ell}(\cdot \mid g)$ based on
	$\{P_i \mid i \in I_\ell^c, \tilde{X}_{i, \theta_t^{-\ell}} = g\}$
12	Choose $(W_{i, \theta_t^{-\ell}})_{i \in I_\ell}$ solving optimization
	$\underset{w \in [0, \infty)^{ I_\ell }}{\operatorname{argmax}} \left\{ \sum_{i \in I_\ell} \hat{F}_{\theta_t^{-\ell}}^{-\ell} \left(k\alpha/m \cdot w_i \mid \tilde{X}_{i, \theta_t^{-\ell}} \right) : w_i \geq 0, \sum_{i \in I_\ell} w_i = I_\ell \right\}$
13	Set $W_i = \frac{1}{T} \sum_{t=1}^T W_{i,t}$
14	Reject H_i if $P_i \leq W_i t$ for some $t > 0$

The actual implementation deviates at regularisation, Grenander estimator instead of ECDF [Gre56], and feature subsampling [MR06].

Algorithm 7: IHW-Quantile

```

1 Randomly partition  $[m]$  into  $I_1, \dots, I_K$ 
2 for  $l = 1, \dots, K$  do
3   for  $s = 1, \dots, d$  do
4     Order the marginally covariates on dimension  $s$ , i.e.
        $(\text{proj}_s \mathbf{X})_{(1)} \leq \dots \leq (\text{proj}_s \mathbf{X})_{(m)}$ .
5     Divide the ordered covariates into similarly sized bins, i.e. set Set
       
$$\tilde{X}_{i,s} = \begin{cases} 1 & \text{if } \text{proj}_s \mathbf{X}_i \leq (\text{proj}_s \mathbf{X})_{(G)} \\ \dots & \\ G & \text{if } (\text{proj}_s \mathbf{X})_{(m-G)} \leq \text{proj}_s \mathbf{X}_i \end{cases}$$

6   Set  $\tilde{X}_i = \tilde{X}_{i,1} \times \dots \times \tilde{X}_{i,d}$ 
7   for  $g \in \{1, \dots, G\}^s$  do
8     Estimate  $F(\cdot \mid g) : t \mapsto \mathbb{P}(P_i \leq t \mid \tilde{X}_i = g)$  by  $\hat{F}^{-\ell}(\cdot \mid g)$  based on
        $\{P_i \mid i \in I_\ell^c, \tilde{X}_i = g\}$ 
9   Choose  $(W_i)_{i \in I_\ell}$  solving optimization
       
$$\underset{w \in [0, \infty)^{|I_\ell|}}{\operatorname{argmax}} \left\{ \sum_{i \in I_\ell} \hat{F}^{-\ell} \left( k\alpha/m \cdot w_i \mid \tilde{X}_i \right) : w_i \geq 0, \sum_{i \in I_\ell} w_i = |I_\ell| \right\}$$

10 Reject  $H_i$  if  $P_i \leq W_i t$  for some  $t > 0$ 

```

Bibliography

- [15] “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/> (visited on 05/19/2022).
- [Acr19] Constantin Ahlmann-Eltze, Stan Developers (Code from the Stan Math library is reused in 'cluster_interval.cpp'), and David Robinson (Code from the fuzzyjoin package is reused). *tidygenomics: Tidy Verbs for Dealing with Genomic Data Frames*. Version 0.1.2. Aug. 8, 2019. URL: <https://CRAN.R-project.org/package=tidygenomics> (visited on 05/23/2022).
- [Alb+14] B Alberts et al. “Chapter 4: DNA, chromosomes and genomes”. In: *Molecular biology of the cell, 6th edn.* Garland Science, New York/Abingdon, UK (2014).
- [BGH10] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. “Independent filtering increases detection power for high-throughput experiments”. In: *Proceedings of the National Academy of Sciences* 107.21 (2010), pp. 9546–9551.
- [BH95] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995). Publisher: [Royal Statistical Society, Wiley], pp. 289–300. ISSN: 0035-9246. URL: <https://www.jstor.org/stable/2346101> (visited on 02/15/2022).
- [BHJ15] Julia Benditkis, Philipp Heesen, and Arnold Janssen. “The false discovery rate (FDR) of multiple tests in a class room lecture”. In: *arXiv preprint arXiv:1511.07050* (2015).
- [BL18] Simina M. Boca and Jeffrey T. Leek. “A direct approach to estimating false discovery rates conditional on covariates”. In: *PeerJ* 2018.12 (2018). Publisher: PeerJ Inc. ISSN: 21678359. DOI: 10.7717/peerj.6035.
- [BR08] Gilles Blanchard and Etienne Roquain. “Two simple sufficient conditions for FDR control”. In: *Electronic journal of Statistics* 2 (2008), pp. 963–992.
- [Bre+17] Leo Breiman et al. *Classification and regression trees*. Routledge, 2017.
- [Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.
- [BS12] Pierre Bajgrowicz and Olivier Scaillet. “Technical trading revisited: False discoveries, persistence tests, and transaction costs”. In: *Journal of Financial Economics* 106.3 (2012), pp. 473–491.

- [BY01] Yoav Benjamini and Daniel Yekutieli. *The Control of the False Discovery Rate in Multiple Testing under Dependency*. Publication Title: Source: The Annals of Statistics Volume: 29 Issue: 4. 2001, pp. 1165–1188.
- [Car17] Gabriela Teixeira Carrasqueiro. “Epigenetic alterations in sepsis”. PhD thesis. Universidade do Algarve (Portugal), 2017.
- [Cay05] Lawrence Cayton. “Algorithms for manifold learning”. In: *Univ. of California at San Diego Tech. Rep* 12.1-17 (2005), p. 1.
- [CGS14] Bertrand Calippe, Xavier Guillonneau, and Florian Sennlaub. “Complement factor H and related proteins in age-related macular degeneration”. In: *Comptes Rendus Biologies* 337.3 (2014), pp. 178–184.
- [Che+18] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.
- [CN08] Hyungwon Choi and Alexey I Nesvizhskii. “False discovery rates and related statistical concepts in mass spectrometry-based proteomics”. In: *Journal of proteome research* 7.01 (2008), pp. 47–50.
- [Cro07] Meredith E Crosby. “Cell cycle: principles of control”. In: *The Yale journal of biology and medicine* 80.3 (2007), p. 141.
- [CRS10] James J Chen, Paula K Robeson, and Michael J Schell. “The false discovery rate: a key concept in large-scale genetic studies”. In: *Cancer Control* 17.1 (2010), pp. 58–62.
- [CRS21] Xiongzhai Chen, David G Robinson, and John D Storey. “The functional false discovery rate with applications to genomics”. In: *Biostatistics* 22.1 (2021), pp. 68–81.
- [CS09] T. Tony Cai and Wenguang Sun. “Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks”. In: *Journal of the American Statistical Association* 104.488 (Dec. 2009), pp. 1467–1481. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/jasa.2009.tm08415. URL: <http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.tm08415> (visited on 05/27/2022).
- [CWR11] Richard Chahwan, Sandeep N Wontakal, and Sergio Roa. “The multidimensional nature of epigenetic information and its role in disease”. In: *Discovery medicine* 11.58 (2011), pp. 233–243.
- [DG12] Noah Dephoure and Steven P Gygi. “Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast”. In: *Science signaling* 5.217 (2012), rs2–rs2.
- [DKW56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. “Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator”. In: *The Annals of Mathematical Statistics* 27.3 (Sept. 1956), pp. 642–669. ISSN: 0003-4851. DOI: 10.1214/aoms/1177728174. URL: <http://projecteuclid.org/euclid.aoms/1177728174> (visited on 12/16/2021).

- [DR20] Sebastian Döhler and Etienne Roquain. “Controlling the false discovery exceedance for heterogeneous tests”. In: *Electronic Journal of Statistics* 14.2 (Jan. 1, 2020). ISSN: 1935-7524. DOI: 10.1214/20-EJS1771. URL: <https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-14/issue-2/Controlling-the-false-discovery-exceedance-for-heterogeneous-tests/10.1214/20-EJS1771.full> (visited on 06/09/2022).
- [DSB03] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C Boldrick. “Multiple hypothesis testing in microarray experiments”. In: *Statistical Science* 18.1 (2003), pp. 71–103.
- [Dur19] Guillermo Durand. *Adaptive p-value weighting with power optimality*. Mar. 21, 2019. arXiv: 1710.01094[math,stat]. URL: <http://arxiv.org/abs/1710.01094> (visited on 06/08/2022).
- [Dwo+15] Cynthia Dwork et al. “The reusable holdout: Preserving validity in adaptive data analysis”. In: *Science* 349.6248 (2015), pp. 636–638.
- [Fer+08] Egil Ferkingstad et al. “Unsupervised empirical Bayesian multiple testing with external covariates”. In: *The Annals of Applied Statistics* 2.2 (2008), pp. 714–735.
- [FMN16] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. “Testing the manifold hypothesis”. In: *Journal of the American Mathematical Society* 29.4 (2016), pp. 983–1049.
- [GLN02] Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. “Thresholding of statistical maps in functional neuroimaging using the false discovery rate”. In: *Neuroimage* 15.4 (2002), pp. 870–878.
- [Gre56] Ulf Grenander. “On the theory of mortality measurement”. In: *Scandinavian Actuarial Journal* 1956.1 (Jan. 1956), pp. 70–96. ISSN: 0346-1238, 1651-2030. DOI: 10.1080/03461238.1956.10414936. URL: <http://www.tandfonline.com/doi/abs/10.1080/03461238.1956.10414936> (visited on 01/17/2022).
- [Gru+15] Fabian Grubert et al. “Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions”. In: *Cell* 162.5 (Aug. 2015), pp. 1051–1065. ISSN: 00928674. DOI: 10.1016/j.cell.2015.07.048. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0092867415009642> (visited on 05/06/2022).
- [GRW06] Christopher R. Genovese, Kathryn Roeder, and Larry Wasserman. “False discovery control with p-value weighting”. In: *Biometrika* 93.3 (Sept. 1, 2006), pp. 509–524. ISSN: 1464-3510, 0006-3444. DOI: 10.1093/biomet/93.3.509. URL: <http://academic.oup.com/biomet/article/93/3/509/380687/False-discovery-control-with-pvalue-weighting> (visited on 02/18/2022).
- [GS14] Jelle J Goeman and Aldo Solari. “Multiple hypothesis testing in genomics”. In: *Statistics in medicine* 33.11 (2014), pp. 1946–1978.

- [GW02] Christopher Genovese and Larry Wasserman. “Operating characteristics and extensions of the false discovery rate procedure”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 499–517.
- [HH18] SUSAN HUBER HOLMES and Wolfgang Huber. *Modern statistics for modern biology*. Cambridge University Press, 2018.
- [HL14] Leah N Hitchcock and K Matthew Lattal. “Histone-mediated epigenetics in addiction”. In: *Progress in molecular biology and translational science* 128 (2014), pp. 51–87.
- [HLP21] Mohammed Amir Husain, Benoit Laurent, and Mélanie Plourde. “APOE and Alzheimer’s Disease: From Lipid Transport to Physiopathology and Therapeutics”. In: *Frontiers in Neuroscience* 15 (Feb. 17, 2021), p. 630502. ISSN: 1662-4548. DOI: 10.3389/fnins.2021.630502. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7925634/> (visited on 05/19/2022).
- [Hub+15] Wolfgang Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. In: *Nature methods* 12.2 (2015), pp. 115–121.
- [Ign+16] Nikolaos Ignatiadis et al. “Data-driven hypothesis weighting increases detection power in genome-scale multiple testing”. In: *Nature Methods* 13.7 (June 29, 2016). Publisher: Nature Publishing Group, pp. 577–580. ISSN: 15487105. DOI: 10.1038/nmeth.3885.
- [Ign15] Nikolaos Ignatiadis. *Bachelorarbeit: Towards powerful FDR controlling procedures: Data-driven selection of asymmetric rejection regions by a greedy constrained maximization*. 2015.
- [Ign16] Nikolaos Ignatiadis. *Masterarbeit: False discovery rates, informative covariates and data-driven, weighted Benjamini-Hochberg procedures*. 2016.
- [IH21] Nikolaos Ignatiadis and Wolfgang Huber. “Covariate powered cross-weighted multiple testing”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 83.4 (Sept. 1, 2021). Publisher: John Wiley and Sons Inc, pp. 720–751. ISSN: 14679868. DOI: 10.1111/rssb.12411. arXiv: 1701.05179.
- [IKK22] Hemant Ishwaran, Udaya B Kogalur, and Maintainer Udaya B Kogalur. “Package ‘randomForestSRC’”. In: *breast* 6 (2022), p. 1.
- [IKL] Hemant Ishwaran, Udaya B. Kogalur, and Min Lu. *Getting starting with the randomForestSRC R-package for random forest analysis of regression, classification, survival and more*. URL: <https://luminwin.github.io/randomForestSRC/articles/getstarted.html> (visited on 03/19/2022).
- [IW19] Nikolaos Ignatiadis and Stefan Wager. “Covariate-powered empirical Bayes estimation”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [KCH02] HJ Keselman, Robert Cribbie, and Burt Holland. “Controlling the rate of Type I error over a large set of statistical tests”. In: *British Journal of Mathematical and Statistical Psychology* 55.1 (2002), pp. 27–39.

- [Ken+02] W James Kent et al. “The human genome browser at UCSC”. In: *Genome research* 12.6 (2002), pp. 996–1006.
- [Kie+98] GM Kiebert et al. “Quality of life after radiation therapy of cerebral low-grade gliomas of the adult: results of a randomised phase III trial on dose response (EORTC trial 22844)”. In: *European Journal of Cancer* 34.12 (1998), pp. 1902–1909.
- [Kle13] Achim Klenke. *Wahrscheinlichkeitstheorie*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. ISBN: 978-3-642-36017-6 978-3-642-36018-3. DOI: 10.1007/978-3-642-36018-3. URL: <http://link.springer.com/10.1007/978-3-642-36018-3> (visited on 06/14/2022).
- [Kor+19] Keegan Korthauer et al. “A practical guide to methods controlling false discoveries in computational biology”. In: *Genome Biology* 20.1 (June 4, 2019). Publisher: BioMed Central Ltd. ISSN: 1474760X. DOI: 10.1186/s13059-019-1716-1.
- [KTC08] Xiayi Ke, Martin S Taylor, and Lon R Cardon. “Singleton SNPs in the human genome and implications for genome-wide association studies”. In: *European Journal of Human Genetics* 16.4 (2008), pp. 506–515.
- [LB17] Ang Li and Rina Foygel Barber. “Accumulation Tests for FDR Control in Ordered Hypothesis Testing”. In: *Journal of the American Statistical Association* 112.518 (Apr. 3, 2017), pp. 837–849. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2016.1180989. URL: <https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1180989> (visited on 12/10/2021).
- [LB19a] Ang Li and Rina Foygel Barber. “Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.1 (2019), pp. 45–74.
- [LB19b] Ang Li and Rina Foygel Barber. “Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 81.1 (Feb. 1, 2019). Publisher: Blackwell Publishing Ltd, pp. 45–74. ISSN: 14679868. DOI: 10.1111/rssb.12298. arXiv: 1606.07926.
- [Leh+05] Albert L Lehninger et al. *Lehninger principles of biochemistry*. Macmillan, 2005.
- [LF18] Lihua Lei and William Fithian. *AdaPT: an interactive procedure for multiple testing with side information*. 2018.
- [LGC09] Michael Lawrence, Robert Gentleman, and Vincent Carey. “rtracklayer: an R package for interfacing with genome browsers”. In: *Bioinformatics* 25.14 (2009), pp. 1841–1842.
- [LNH11] John C Lindon, Jeremy K Nicholson, and Elaine Holmes. *The handbook of metabonomics and metabolomics*. Elsevier, 2011.

- [LRC05] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*. Vol. 3. Springer, 2005.
- [LS05] Fred van Leeuwen and Bas van Steensel. “Histone modifications: from genome-wide maps to functional insights”. In: *Genome biology* 6.6 (2005), pp. 1–6.
- [Mon+17] Isha Monga et al. “ASPSiRNA: a resource of ASP-siRNAs having therapeutic potential for human genetic disorders and algorithm for prediction of their inhibitory efficacy”. In: *G3: Genes, Genomes, Genetics* 7.9 (2017), pp. 2931–2943.
- [MR06] Nicolai Meinshausen and Greg Ridgeway. “Quantile regression forests.” In: *Journal of Machine Learning Research* 7.6 (2006).
- [MW10] Susan C McQuown and Marcelo A Wood. “Epigenetic regulation in substance use disorders”. In: *Current psychiatry reports* 12.2 (2010), pp. 145–153.
- [Nur+22] Sergey Nurk et al. “The complete sequence of a human genome”. In: *Science* 376.6588 (2022), pp. 44–53.
- [NW21] Xinkun Nie and Stefan Wager. “Quasi-oracle estimation of heterogeneous treatment effects”. In: *Biometrika* 108.2 (2021), pp. 299–319.
- [PL19] Taylor Pospisil and Ann B. Lee. “(f)RFCDE: Random Forests for Conditional Density Estimation and Functional Data”. In: (June 16, 2019). arXiv: 1906.07177. URL: <http://arxiv.org/abs/1906.07177>.
- [Ram+11] Oren Ram et al. “Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells”. In: *Cell* 147.7 (2011), pp. 1628–1639.
- [RN11] Alfred J Robison and Eric J Nestler. “Transcriptional and epigenetic mechanisms of addiction”. In: *Nature reviews neuroscience* 12.11 (2011), pp. 623–637.
- [Sch+95] Mark Schena et al. “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”. In: *Science* 270.5235 (1995), pp. 467–470.
- [Sch86] Anton Schick. “On asymptotically efficient estimation in semiparametric models”. In: *The Annals of Statistics* (1986), pp. 1139–1151.
- [Sco+15] James G Scott et al. “False discovery rate regression: an application to neural synchrony detection in primary visual cortex”. In: *Journal of the American Statistical Association* 110.510 (2015), pp. 459–471.
- [Sha12] A. A. Shabalín. “Matrix eQTL: ultra fast eQTL analysis via large matrix operations”. In: *Bioinformatics* 28.10 (May 15, 2012), pp. 1353–1358. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bts163. URL: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts163> (visited on 05/18/2022).
- [Sha95] Juliet Popper Shaffer. “Multiple hypothesis testing”. In: *Annual review of psychology* 46.1 (1995), pp. 561–584.

- [SJ08] Jay Shendure and Hanlee Ji. “Next-generation DNA sequencing”. In: *Nature biotechnology* 26.10 (2008), pp. 1135–1145.
- [SSS22] Fariha Sohail, Muhammad Umair Sohali, and Javid Shabbir. “An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, \$41.98, eISBN: 978-1-4614-7137-7”. In: *Statistical Theory and Related Fields* 6.1 (Jan. 2, 2022), pp. 87–87. ISSN: 2475-4269, 2475-4277. DOI: 10.1080/24754269.2021.1980261. URL: <https://www.tandfonline.com/doi/full/10.1080/24754269.2021.1980261> (visited on 06/13/2022).
- [Sto02] John D. Storey. “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (Aug. 2002), pp. 479–498. ISSN: 13697412, 14679868. DOI: 10.1111/1467-9868.00346. URL: <https://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00346> (visited on 06/15/2022).
- [Str08] Korbinian Strimmer. “A unified approach to false discovery rate estimation”. In: *BMC bioinformatics* 9.1 (2008), pp. 1–14.
- [STS04] John D. Storey, Jonathan E. Taylor, and David Siegmund. “Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1 (Feb. 2004), pp. 187–205. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.1467-9868.2004.00439.x. URL: <https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2004.00439.x> (visited on 02/06/2022).
- [SV12] Salvatore Spicuglia and Laurent Vanhille. “Chromatin signatures of active enhancers”. In: *Nucleus* 3.2 (2012), pp. 126–131.
- [TBT16] Xiaoying Tian, Nan Bi, and Jonathan Taylor. “MAGIC: a general, powerful and tractable method for selective inference”. In: *arXiv preprint arXiv:1607.02630* (2016).
- [Van00] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.
- [VMN14] Richard Van Noorden, Brendan Maher, and Regina Nuzzo. “The top 100 papers”. In: *Nature News* 514.7524 (2014), p. 550.
- [Was14] Larry Wasserman. “Discussion: “A significance test for the lasso””. In: *The Annals of Statistics* 42.2 (Apr. 2014), pp. 501–508. URL: <http://dx.doi.org/10.1214/13-AOS1175E>.
- [You06] Robert M Youngson. *Collins dictionary of human biology*. Collins, 2006.
- [Zha+08] Yong Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9 (2008), pp. 1–9.

- [Zhu+15] Zuobin Zhu et al. “Enrichment of minor alleles of common SNPs and improved risk prediction for Parkinson’s disease”. In: *PloS one* 10.7 (2015), e0133421.