Advances in interactive hypothesis testing

Boyan Duan

June 4th, 2021

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Aaditya Ramdas (Co-Chair)
Larry Wasserman (Co-Chair)
Sivaraman Balakrishnan
Peter Grünwald (CWI, Amsterdam)
William Fithian (UC, Berkeley)

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

Interactive testing is recently developed to allow to human experts participate in the hypothesis testing algorithms. Most testing methods are predefined algorithms that do not allow modifications after observing the data. However, in practice, analysts tend to choose a promising algorithm after observing the data; unfortunately, this violates the validity of the conclusion. In contrast, the interactive methods allow the algorithm to be much more flexible, such that a human (or a computer program) may adaptively design the algorithm in a data-dependent manner if they adhere to a particular protocol of "masking" and "unmasking". Interactive testing was first proposed for multiple hypothesis testing to control the false discovery rate (FDR). This thesis develops interactive tests in various problem settings.

Following the problem setting in multiple testing, Chapter 2 and Chapter 3 propose interactive tests with global type-I error control and familywise error rate (FWER) control, respectively. The interactive procedures can take advantage of covariates and repeated user guidance to focus on possible non-nulls, achieving high power in numerical experiments where the non-nulls are sparse and structured. In addition, we explore alternative forms of masking, which could be more robust to conservative nulls.

Moving outside of multiple testing with p-values, Chapter 4 studies the problem of comparing multiple samples. Classical nonparametric tests, such as the Wilcoxon test, are often based on the ranks of observations. We design an *interactive* rank test called i-Wilcoxon with type-I error control. The i-Wilcoxon test is first proposed for two-sample comparison with unpaired data, and then extended to paired data, multi-sample comparison, and sequential settings, thus also extending the Kruskal-Wallis and Friedman tests. As alternatives, we also numerically investigate (non-interactive) covariance-adjusted variants of the Wilcoxon test, and provide practical recommendations based on the anticipated population properties of the treatment effects.

Out of the participants in a randomized experiment with anticipated heterogeneous treatment effects, is it possible to identify which ones have a positive treatment effect, even though each has only taken either treatment or control but not both? While subgroup analysis has received attention, claims about individual participants are more challenging. Chapter 5 frame the problem in terms of multiple hypothesis testing: we think of each individual as a null hypothesis (the potential outcomes are equal, for example) and aim to identify individuals for whom the null is false (the treatment potential outcome stochastically dominates the control, for example). We develop a novel interactive algorithm that identifies such a subset, with nonasymptotic control of the false discovery rate (FDR). We also propose several extensions: (a) relaxing the null to nonpositive effects, (b) generalizing the setting to observational studies with heterogeneous and unknown propensity scores, (c) moving from unpaired to paired samples, and (d) subgroup identification.

Acknowledgments

I want to thank ...

Contents

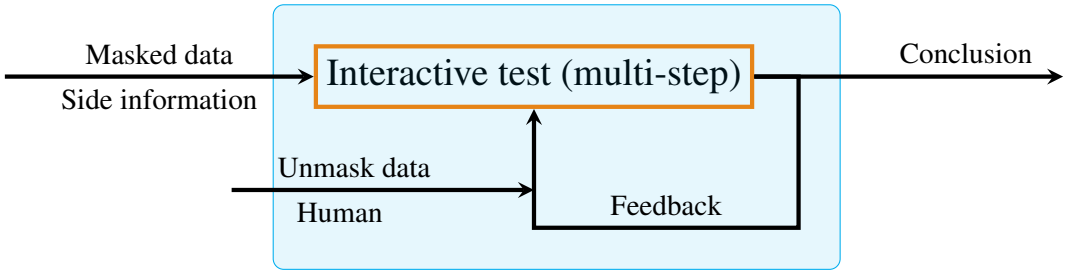
List of Figures

1 Introduction

There is increasing concern that many published results in various medical and life sciences are over-optimistic. Scientists tend to try out several statistical analyses on the same dataset until there is a significant ("positive") result. When a second group repeats the same experiments, the outcomes are often not as positive. The problem in reproducibility comes from the bias in selecting the analysis tool: researchers choose a promising method after observing the data, which violates the validity of the results. In seek of a framework that allows experts (scientists and statisticians) to work together with statistical models and machine learning algorithms to discover scientific insights with rigorous guarantees, we work on the idea of interactive testing.

Classical hypothesis testing follows a pipeline of specifying a hypothesis, choosing a test, collecting data, and running the test. Most tests follow a prespecified algorithm (or a fixed function of the data), such as ordering p-values in multiple hypothesis testing. Therefore, to utilize domain knowledge in various applications, each case might require designing a new test from scratch in order to optimally combines the data with prior knowledge or certain structural constraints. Interactive testing, instead, provides a simple and flexible framework to be customized to various sources of prior knowledge. Furthermore, interactive tests are iterative so that a human analyst is allowed to participate in the loop to modify the test progressively in a data-dependent manner.

(a) Classical hypothesis testing: a pipeline.



(b) Interactive testing: a loop.

Figure 1: Procedures of classical testing and interactive testing.

The key idea that permits human interaction while ensuring valid error control is "masking and unmasking". As an example, we present one form of masking p-values for multiple hypothesis testing, which is proposed in the first work of interactive testing. Given n hypotheses $H_1, \ldots, H_n$ and their corresponding p-values $P_1, \ldots, P_n$, each p-value P_i is decomposed into two parts,

$$h(p_i) = 2 \cdot \mathbb{1}\{P_i < 0.5\} - 1 \quad \text{and} \quad g(p_i) = \min\{P_i, 1 - P_i\}. \tag{1}$$

Here, $g(P_i)$ is called the *masked p-value*, while $h(P_i)$ is called the *missing bit* since it is either plus or minus one. The critical observation is that $h(P_i)$ and $g(P_i)$ are independent if H_i is null (P_i is uniformly distributed). Masking was introduced recently by Lei and Fithian [2018] in the context of false discovery rate (FDR) control, and further generalized and extended in Lei et al. [2020]. More forms of masking are also developed in this thesis for various problem settings. The underlying property of masking can be traced to the "knockoff" method by Arias-Castro and Chen [2017]; Barber and Candès [2015].

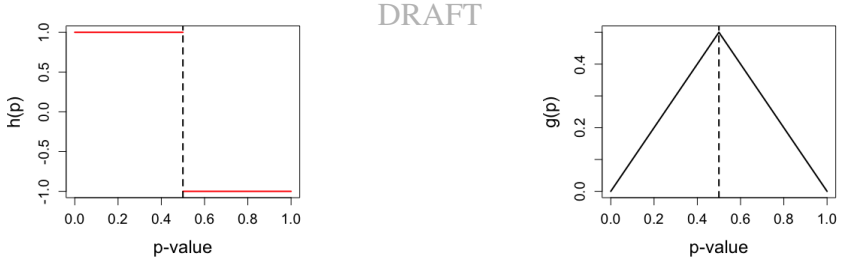


Figure 2: One form of masking p-values: missing bits h (left) and masked p-values g (right). For uniform p-values, $g(P)$ and $h(P)$ are independent.

To describe the interactive test in short, the analyst initially designs the algorithm by observing only the masked p-values. At each iteration, the missing bits are progressively unmasked (revealed) one at a time, and the analyst can modify the test statistic or any working model as needed. Note that even though a human is allowed to make subjective decisions at each step of the interaction, an algorithm can be deployed to act on the human's behalf. We remark that we do not wish to claim that our interactive tests are more powerful than prior work in any universal sense, but instead, attempt to expand the creative design space of new procedures that can involve a human in the loop and explore their potential benefits.

This thesis proposes interactive tests for several testing problems: multiple hypothesis testing (Chapter 2 and 3), nonparametric tests for multi-sample comparison (Chapter 4), and identification of positive treatment effect (Chapter 5). Different problem setups require different constructions of masking and the corresponding test statistics, where various techniques such as martingales and recent uniform concentration inequalities are involved. In the following chapters, we elaborate on each problem setup, detail the proposed interactive tests, and discuss their performances and extensions.

2 Interactive Martingale Tests for the Global Null

2.1 Introduction

This paper proposes new martingale-based methods for testing the global null corresponding to hypotheses $\{H_i\}_{i\in\mathcal{I}}$ using a corresponding set of p-values $\{p_i\}_{i\in\mathcal{I}}$ and possibly other covariates $\{x_i\}_{i\in\mathcal{I}}$, where the index set $\mathcal{I}$ can be finite or countably infinite. Global null testing corresponds to testing if all individual hypotheses are truly nulls (denoted as $H_i = 0$), against its complement:

$$\mathcal{H}_{\mathcal{G}_0} : H_i = 0 \text{ for all } i \in \mathcal{I}, \quad \mathcal{H}_{\mathcal{G}_1} : H_i = 1 \text{ for at least one } i \in \mathcal{I}.$$

As we review later in the introduction, this is a well-studied classical problem. We consider two settings, the batch setting and the online setting, and our proposed framework applies to both settings:

- Batch setting: we have access to a fixed batch of n hypotheses, thus $\mathcal{I} = \{1, \ldots, n\}$.
- Online setting: an unknown and potentially infinite number of hypotheses arrive sequentially in a stream, thus $\mathcal{I} = \{1, 2, \ldots, k, \ldots\}$.

Most common global null tests involve a one-step operation, comparing a single statistic with a critical value derived from its null distribution. Observing that many classical tests effectively use a martingale-type test statistic, we propose novel martingale analogs of these tests that are inherently sequential (multi-step) in nature, and thus naturally apply in the online setting, or in the batch setting if an ordering can be created using prior knowledge and/or the data. Intriguingly, the ordering may also be created *interactively*: this means that an analyst may adaptively create the ordering in a data-dependent manner if they adhere to a particular protocol of *masking* and *unmasking* (the definition is introduced later in equation (4)). In order to understand why our interactive martingale tests have desirable properties (both controlling type-I errors and having higher power in structured settings), it is necessary to present them last, after having derived the vanilla non-interactive martingale global null tests, which are also novel in their own right. Specifically, for the purposes of progressively developing intuition, our treatment follows three steps of increasing complexity:

- (Preordered setting, Section 2.2) In the batch setting, the analyst employs *prior* knowledge (data-independent) to preorder the hypotheses. In the online setting, an ordering of hypotheses is provided by nature.
- (Data-adaptive ordering, Section 2.3.1) In the batch setting, the hypotheses are unordered, but an adaptive data-dependent ordering is created based on "masked" p-values. In the online setting, nature orders hypotheses, but the analyst discards some hypotheses from the ordering based on their masked p-values. Even though the data-adaptive and preordered settings proceed sequentially and handle the p-values one at a time, the analyst plays no role *during* this sequential process, as all the rules for how to order the hypotheses are prespecified before the data is observed.
- (Interactive ordering, Section 2.3.2). The utility of masking to enable interaction with a human is most compelling in the batch setting, where in addition to the unordered hypotheses, we suppose that the analyst also has additional side information in the form of covariates, and perhaps prior knowledge in the form of structural constraints on the non-null set. Using these, and any working models of their choice, the analyst interactively creates an ordering by initially observing only masked p-values, and progressively unmasking them one at a time. The analyst can update their prior knowledge and/or structural constraints and/or working model in the middle of the process (when only some hypotheses have been ordered and their p-values unmasked), thus intervening to change the rest of the ordering. It is important to note that *even though an analyst is allowed to*

make subjective decisions at each step of the interaction, an algorithm can be deployed in place of the analyst.

Since all our tests proceed sequentially in nature, accumulating evidence from one hypothesis at a time, the type-I error guarantee we achieve is that

$$\mathbb{P}_0(\exists i \in \mathcal{I} : \text{the test stops and rejects } \mathcal{H}_{\mathcal{G}_0} \text{ after step } i) \leq \alpha,$$

where $\mathbb{P}_0$ is the probability under the global null $\mathcal{H}_{\mathcal{G}_0}$. They are judged based on their power,

$$\mathbb{P}_1(\exists i \in \mathcal{I} : \text{the test stops and rejects } \mathcal{H}_{\mathcal{G}_0} \text{ after step } i),$$

where $\mathbb{P}_1$ is the probability under some alternative in $\mathcal{H}_{\mathcal{G}_1}$. We remark that even though we formulate our tests in terms of a target type-I error level α, there is an equivalent formulation in terms of creating a sequential "always-valid" p-value for the global null that is valid at any arbitrary stopping time. Section 2.7 explicitly connects these two interpretations.

2.1.1 Assumptions

Instead of assuming that the marginal distribution of null p-values is exactly uniform, we relax it by allowing conservative p-values defined in two different ways. We either assume that (a) if the global null is true, all p-values are stochastically larger than uniform:

$$\text{If } \mathcal{H}_{\mathcal{G}_0} \text{ is true, } \mathbb{P}(p_i \leq t) \leq t \text{ for all } t \in [0,1], i \in \mathcal{I}. \tag{2}$$

or assume that (b) if the global null is true, all p-values are *mirror-conservative*:

$$\text{If } \mathcal{H}_{\mathcal{G}_0} \text{ is true, } f_i(a) \leq f_i(1-a) \text{ for all } 0 \leq a \leq 0.5, i \in \mathcal{I}, \tag{3}$$

where f_i is the probability mass function of p_i for discrete p-values or the density function otherwise. Neither of the aforementioned conditions implies the other, though the former is more commonly made. Examples of mirror-conservative p-values include permutation p-values and one-sided tests of univariate parameters with monotone likelihood ratio [Lei and Fithian, 2018]. In the majority of the paper, it may be easier for the reader to pretend that the null p-values are exactly uniform for simplicity. Later in the paper, we explicitly demonstrate the distinct advantages of our tests for conservative p-values. We also assume that if the global null is true, the null p-values are independent of each other:

$$\text{If } \mathcal{H}_{\mathcal{G}_0} \text{ is true, } \{p_i\}_{i \in \mathcal{I}} \text{ are jointly independent.}$$

This is also a common assumption; Fisher's test [Fisher, 1992] and Tukey's Higher Criticism [Donoho and Jin, 2015] are two other examples. Even though we are cognizant that independence is a strong assumption that only holds in some limited situations in practice (like meta-analysis), we wish to explore how much it can be exploited to design novel tests, for instance enabling the use of martingale techniques and "masking", as described soon.

We remark that all aforementioned assumptions on the null p-values only need to hold under the global null. If the global null is not true, we do not require the null p-values (or the non-nulls) to have any particular marginal distribution or to satisfy any independence assumptions.

2.1.2 Related work

Our paper builds on and connects three distinct lines of work: classical work on global null testing, modern ideas on permitting interaction using p-value masking, and recent ideas on uniform martingale concentration inequalities. We discuss these separately below.

Global null testing. Most previous tests for the global null have been designed to work in the batch setting, and it continues to be an active area of research [Kost and McDermott, 2002; Owen, 2009; Rüger, 1978; Rüschendorf, 1982; Vovk and Wang, 2020a]. Our work is most directly connected to tests which accumulate information as a sum, such as Fisher's and Stouffer's tests [Stouffer et al., 1949].

There are many other global null tests like the Bonferroni method, Simes' test [Simes, 1986], and Higher Criticism, and our techniques do not apply to these. Importantly, *we do not claim that our interactive martingale tests are more powerful than prior work in any universal sense, but instead, our goal is to expand the creative design space of new procedures that can involve a human in the loop and explore their potential benefits.*

Permitting interaction by masking the p-values. The motivation behind masking p-values is to permit interaction with an analyst, who may freely employ models, prior knowledge and intuition, without any risk of violating type-I error control. The main idea is to decompose each individual p-value p_i into two parts,

$$h(p_i) = 2 \cdot 1\{p_i < 0.5\} - 1 \quad \text{and} \quad g(p_i) = \min\{p_i, 1 - p_i\}. \tag{4}$$

Here, $g(p_i)$ is called the *masked p-value*, while $h(p_i)$ is called the *missing bit* since it is either plus or minus one. The critical observation is that $h(p_i)$ and $g(p_i)$ are independent if H_i is null (p_i is uniformly distributed). Masking was introduced recently by Lei and Fithian [2018] in the context of false discovery rate (FDR) control, and further generalized and extended in Lei et al. [2020] for FDR control under structural constraints, and then followed by work on FWER control [Duan et al., 2020a]. The underlying property of masking can be traced to the "knockoff" method by Arias-Castro and Chen [2017]; Barber and Candès [2015]. In this paper, we show that masking is also useful for global null testing in structured settings, and permitting interaction with an insightful analyst can improve power (but it is impossible for any analyst to violate type-I error control).

Uniform martingale concentration inequalities. All new test statistics in this paper are designed to be martingales under the global null. The type-I error control guarantees for our tests thus stem from utilizing *uniform* martingale concentration inequalities. These "boundary crossing" inequalities are high probability statements about the behavior of the entire trajectory of the martingale. In fact, several of our martingales have increments which are either fair coin flips (± 1) or standard Gaussians, which are some of the most well studied objects in sequential analysis, especially through their natural connections to Brownian motion [Siegmund, 1986]. In this paper, we care about nonasymptotic guarantees on the type-I error, and hence we use some recent line-crossing inequalities [Howard et al., 2020a] and new curve-crossing inequalities [Howard et al., 2020b] that are nonasymptotic generalizations of the law of the iterated logarithm, which goes back to the work by Robbins [Robbins, 1970] (see Appendix A.4 for a detailed comparison). For a martingale M_k, these boundaries are denoted $u_\alpha(k)$ and satisfy

$$\mathbb{P}(\exists k \in \mathbb{N} : M_k > u_\alpha(k)) \leq \alpha.$$

In the next section, we provide the exact expressions for the $u_\alpha(k)$ that we use, which are chosen because they have similar qualitative behavior but tighter constants than earlier work, references to which may be found within the aforementioned papers.

2.1.3 Outline

To progressively build intuition, the preordered martingale test is described in Section 2.2 followed by the adaptively ordered martingale test in Section 2.3.1. In Section 2.3.2, the general interactively ordered

martingale test is presented. For all these methods, the type-I error guarantees are presented immediately after the algorithms. However, power guarantees for all algorithms in the Gaussian sequence model are derived in Section 2.4. We then perform extensive simulations in Section 2.5. In Section 2.6, we examine the robustness of our test to conservative nulls. Section 2.7 explicitly describes how to interpret our tests as tracking an anytime-valid sequential p-value. Finally in Section 2.8, we discuss alternative ways of masking p-values. We end with a brief summary in Section 2.9, and defer all proofs and additional experiments to the Appendix.

2.2 The preordered martingale test

The preordered martingale test is not a single test, but instead, a general framework to extend the application of many classical methods that use the sum or product of transformed p-values, such as Stouffer's method [Stouffer et al., 1949] and Fisher's method [Fisher, 1992], from the batch setting to the online setting. In this section, the ordering of hypotheses is fixed in advance by nature, or by the analyst using prior knowledge to place potential/suspected non-nulls early in the ordering.

The general framework. Our test takes the following general form:

$$\text{Reject the null if } \sum_{i=1}^{k} f(p_i) \geq u_\alpha(k), \text{ for some } k \in \mathcal{I}, \tag{5}$$

where $f(\cdot)$ is some transformation of the p-value, and $\{u_\alpha(k)\}_{k\in\mathbb{N}}$ is a boundary sequence depending on the choice of f. The boundary is determined by first establishing that the sequence $\{\sum_{i=1}^{k} f(p_i)\}_{k\in\mathbb{N}}$ is a martingale under the global null (after appropriate centering if needed). We then characterize the tail behavior of the martingale increments $f(p_i)$ for a uniform p-value. Finally, to control the type-I error, we employ recent results [Howard et al., 2020a,b] which provide boundaries under parametric and nonparametric conditions on the increments, such that with high probability the entire trajectory of the martingale is contained within the boundary.

The preordered martingale test improves on its original batch version in two aspects. First, the applicability of the original test is extended from the batch setting to the online setting. Second, in the case of sparse non-nulls, the martingale version greatly improves the detection power if the non-nulls appear early on. As an example of converting a classic test to its martingale version, we develop the martingale Stouffer test below. Two more examples can be found in Appendix A.5 for a martingale Fisher test using $f(p_i) = -2\log p_i$, and Appendix A.6 for a martingale chi-square test using $f(p_i) = [\Phi^{-1}(1-p_i)]^2$.

An example: martingale Stouffer test (MST). The batch test by Stouffer et al. [1949] calculates $S_n = \sum_{i=1}^{n} \Phi^{-1}(1-p_i)$, where $\Phi(\cdot)$ denotes the standard Gaussian CDF. Since the distribution of S_n under the global null is $\mathcal{N}(0,n)$, the batch test rejects when $S_n > \sqrt{n}\Phi^{-1}(1-\alpha)$. To design the martingale test, simply observe that $\{S_k\}_{k\in\mathcal{I}}$ is a martingale whose increments $f(p_i) = \Phi^{-1}(1-p_i)$ are standard Gaussians under the global null. There are several types of uniform boundaries $u_\alpha(k)$ for a Gaussian increment martingale, and here we give two examples: linear and curved. The first boundary (transformed from equation (2.29) in Howard et al. [2020a]), which can be derived from the Gaussian sequential probability ratio test [Wald, 1945], grows linearly with time. Specifically, the test rejects the

global null if

$$\exists k \in \mathbb{N} : \sum_{i=1}^{k} \Phi^{-1}(1-p_i) \geq \sqrt{\frac{-\log \alpha}{2m}}k + \sqrt{\frac{-m \log \alpha}{2}}, \tag{6}$$

where $m \in \mathbb{R}_+$ is a tuning parameter that determines the time at which the bound is tightest: a larger m results in a lower slope but a larger offset, making the bound loose early on. We suggest a default value of $m = n/4$ if the number of hypotheses n is finite, but it should be chosen based on the time by which we expect to have encountered most non-nulls (if any). In contrast, the martingale Stouffer test with a curved boundary (equation (2) in Howard et al. [2020b]) rejects the global null if

$$\exists k \in \mathbb{N} : \sum_{i=1}^{k} \Phi^{-1}(1-p_i) \geq 1.7\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)}. \tag{7}$$

These bounds differ in the quota of error budget distributed to every step $k = 1, 2, \ldots$, which can influence the detection power of the martingale test as it is more likely to exceed a tighter bound. Curved bounds have a slower growth rate $O(\sqrt{k \log\log k})$ than the linear bounds, indicating a tighter bound for large enough k, but they are usually looser for small k. Comparisons of the test with several linear and curved boundaries are given in Appendix A.4. Generally, the linear bound is recommended for the batch setting, and the curved bound for the online setting.

The martingale Stouffer test with either boundary controls the type-I error, if under the global null the sum $\{\sum_{i=1}^{k} \Phi^{-1}(1-p_i)\}_{k\in\mathbb{N}}$ is stochastically upper bounded by a martingale with standard Gaussian increments, which holds under our assumption that the null p-values are stochastically larger than uniform, as stated below.

Theorem 1. *If the p-values are independent and stochastically larger than uniform under the global null, then the martingale Stouffer test with linear boundary (6) or curved boundary (7) controls the type-I error at level α.*

The next natural question is what we can prove about the detection power of the aforementioned tests. While this is treated more formally later in the paper, for now it suffices to say that the power of the martingale Stouffer test relies on a good preordering that places non-nulls up front. If such prior knowledge is not available (and say the preordering is completely random, or even adversarial), then the preordered martingale tests can have poor power. This motivates the development of methods based on data-adaptive orderings, as treated next.

2.3 Adaptive and interactive methods

To develop intuition progressively, we first introduce a martingale test whose ordering depends on the p-values in Section 2.3.1, and extend it in Section 2.3.2 to an interactive test, whose ordering can additionally depend on side information (covariates) and human interaction.

2.3.1 The adaptively ordered martingale test (AMT)

If we naively use the p-values to both determine the ordering as well as form the test statistic, the resulting "double-dipped" sequence of test statistics does not form a martingale under the global null. In order to allow using the p-value for determining the ordering, we use a recent idea called masking, as briefly mentioned in the introduction. Each p-value p_i is decomposed as

$$h(p_i) = 2 \cdot 1\{p_i < 0.5\} - 1, \qquad g(p_i) = \min\{p_i, 1-p_i\},$$

where $h(p_i)$ is called the missing bit, and $g(p_i)$ is called the masked p-value. The masked p-values are used to create the ordering (by placing smaller ones up front) while the test statistic just sums the missing bits $h(p_i)$ in that order. Since $h(p_i)$ and $g(p_i)$ are independent under the global null, sorting by the $g(p_i)$ values results in a uniformly random ordering, and the sum of $h(p_i)$ is just a random walk of independent coin flips. Formally, define the set M_k as the first k hypotheses ascendingly ordered by $g(p_i)$. Our test rejects $\mathcal{H}_{\mathcal{G}_0}$ if

$$\exists k \in \{1, \ldots, n\} : \sum_{i \in M_k} h(p_i) \geq u_\alpha(k),$$

where the upper bound $u_\alpha(k)$ is the same as for the martingale Stouffer test in equations (6) and (7), since the sequence of sums $\sum_{i \in M_k} h(p_i)$ is also a martingale with 1-subGaussian increments under the global null. The adaptively ordered martingale test in the batch setting is summarized below.

Algorithm 1 The adaptively ordered martingale test (batch setting)

Input: p-values $(p_i)_{i=1}^n$, target type-I error rate α;
Procedure: Initialize $M_0 = \emptyset$;
for $k = 1, \ldots, n$ **do**
  $M_k = M_{k-1} \cup \operatorname{argmin}_{i \notin M_{k-1}} g(p_i)$;
  if $\sum_{i \in M_k} h(p_i) > u_\alpha(k)$ **then**
    reject the global null and stop;
end

The adaptively ordered martingale test in the online setting proceeds slightly differently: it screens the hypotheses by $g(p)$ so that only promising non-nulls enter the set M_k. Specifically, given a threshold parameter c (such as 0.05), the set M_k expands at time t only if $g(p_t) < c$, as summarized below.

Algorithm 2 The adaptively ordered martingale test (online setting)

Input: target type-I error rate α, threshold parameter c;
Procedure: Initialize $M_0 = \emptyset$, size $k = 0$;
for $t = 1, \ldots,$ **do**
  p_t is revealed by nature;
  if $g(p_t) < c$ **then**
    $k \leftarrow k + 1$, $M_k = M_{k-1} \cup \{t\}$;
  if $\sum_{i \in M_k} h(p_i) > u_\alpha(k)$ **then**
    reject the global null and stop;
end

The adaptively ordered martingale test controls type-I error if under the global null, all p-values are *mirror-conservative* (3), as formally stated below.

Theorem 2. *If the p-values are independent and mirror-conservative under the global null, then the adaptively ordered martingale test controls the type-I error at level α.*

In the batch setting, the adaptive ordering (as realized by the nested sequence $\{M_k\}$) is fully determined at the start of the procedure by sorting the masked p-values. In the next section, we demonstrate that in the presence of independent covariates x_i for each hypothesis and side information such as structural constraints on potential rejected sets, it is actually beneficial to *interactively* determine the ordering one step at a time with a human-in-the-loop, who may be guided by the masked p-values as well as intuition and working models.

2.3.2 The interactively ordered martingale test (IMT)

The interactively ordered martingale test also applies to both batch and online settings. We first describe the method in the batch setting with side information and structural constraints, where the power of interactivity is more compelling.

To begin, first suppose that in addition to the p-values, the scientist also has some side information about each hypothesis available to them in the form of covariates x_i. For example, if the hypotheses are arranged in a rectangular grid, then x_i could be the coordinates on the grid for hypothesis i (examples in Section 2.5.1). We then suppose that the scientist also has some prior knowledge or intuition about what structural constraints the non-nulls would have, if the global null is false. For example, perhaps the scientist thinks that the non-nulls (if any) would be clustered on the grid, themselves forming a rectangular shape (of some size, at some location). Our main assumption about the covariates is:

$$\text{Under the global null, } x_i \perp p_j \text{ for all } i, j \in \mathcal{I}.$$

This is a common assumption for tests that incorporate covariate information, such as Independent Hypothesis Weighting [Ignatiadis et al., 2016], AdaPT [Lei and Fithian, 2018], and STAR [Lei et al., 2020]. In fact, because the aforementioned methods aim at error control of more stringent metrics such as FDR and FWER, their assumptions are stronger in the sense that the independence between x_i and p_i is required for the hypotheses that are truly null even when the global null is not true (i.e., there exist non-nulls). Our interactively ordered martingale test satisfies the following two properties: (a) if the global null is true, the type-I error is controlled, regardless of what the scientist thinks or acts, (b) if the global null is false, and the prior knowledge and/or structural constraints are accurate (or somewhat so), then the power of the test is high. The interactive test proceeds as follows:

- At the beginning, all covariates and masked p-values $(x_i, g(p_i))_{i\in\mathcal{I}}$ are revealed to the scientist, while only the missing bits $(h(p_i))_{i\in\mathcal{I}}$ remain hidden. We initialize $M_0 = \emptyset$.
- The scientist repeats the following at each time step $k \geq 1$: they choose a promising hypothesis $i_k^\star$ from $[n]\backslash M_{k-1}$, and update $M_k = M_{k-1} \cup \{i_k^\star\}$.
- On doing so, they learn $h(p_{i_k^\star})$, and thus keep track of $S_k := \sum_{i\in M_k} h(p_i)$. If $S_k > u_\alpha(k)$ for any k, they stop and reject the global null.

Type-I error control is essentially guaranteed because regardless of how the scientist acts at each step, if the global null is true, all the $g(p_i)$ values and the revealed $h(p_i)$ values do not provide any information about the still hidden missing bits, and thus S_k is a martingale.

When the global null is false, we expect the power to be high because of the following reasons. First, the scientist may use any working model of their choice (or none at all) to guide their choice at each step. For example, they can attempt to estimate the likelihood of being non-null for each hypothesis i at each step k, denoted as $\pi_i^{(k)}$ (posterior probability of being non-null). In fact, as they learn the missing bits at each step, they can change their model or update their prior knowledge based on the observed p-values thus far. The information available to the scientist at the end of step k is denoted by the filtration

$$\mathcal{F}_k := \sigma\left((x_i, g(p_i))_{i=1}^n, (p_i)_{i\in M_k}\right),$$

and thus the choice $i_k^\star$ is predictable, meaning it is measurable with respect to $\mathcal{F}_{k-1}$. The general interactive framework is summarized below as Algorithm 3.

Algorithm 3 The interactively ordered martingale test (batch setting)

Information available to the scientist: side covariate information and/or structural constraints, and masked p-values $\mathcal{F}_0 := \sigma((x_i, g(p_i))_{i=1}^n)$, target error α;
Procedure: Initialize $M_0 = \emptyset$;
for $k = 1, \ldots, n$ **do**
  Using $\mathcal{F}_{k-1}$, pick any $i_k^\star \in [n] \backslash M_{k-1}$. Update $M_k = M_{k-1} \cup \{i_k^\star\}$;
  Reveal $h(p_{i_k^\star})$ and update $\mathcal{F}_k := \sigma\left((x_i, g(p_i))_{i=1}^n, (p_i)_{i \in M_k}\right)$;
  if $\sum_{i \in M_k} h(p_i) > u_\alpha(k)$ **then**
    reject the global null and exit;
end

The interactively ordered martingale test in the online setting screens the hypotheses based on information in $\mathcal{F}_{t-1}$ such that p_t enters the set M_k only when it is a promising non-null, as described in Algorithm 4.

Algorithm 4 The interactively ordered martingale test (online setting)

Procedure: Input target error α. Initialize $M_0 = \emptyset$, size $k = 0$;
for $t = 1, \ldots,$ **do**
  Information available to the scientist: side covariate information and/or structural constraints, and (masked) p-values $\mathcal{F}_{t-1} := \sigma((x_i, g(p_i))_{i=1}^t, (p_i)_{i=1}^{t-1})$;
  Using $\mathcal{F}_{t-1}$, decide whether hypothesis t should be included in M_{k-1};
  if *include hypothesis* t **then**
    $k \leftarrow k + 1$, $M_k = M_{k-1} \cup \{t\}$;
  if $\sum_{i \in M_k} h(p_i) > u_\alpha(k)$ **then**
    reject the global null and stop;
end

The aforementioned algorithms (or frameworks) comes with the following error guarantee, regardless of the choices made by the scientist.

Theorem 3. *If under $\mathcal{H}_{\mathcal{G}_0}$, the p-values are mirror-conservative and are independent of each other and of the covariates x_i, then the interactively ordered martingale test controls the type-I error at level α.*

Note that there is no requirement whatsoever on the null or non-null p-values (i.e., p-values from the hypotheses that are truly non-null) when the global null is false. As before, note that under the global null, the missing bits are random fair coin flips, and the masked p-values are uniform on $[0, 0.5]$ and completely uninformative about the missing bit. However, under the alternative, the true signals have very small masked p-values (say 0.01, 0.003, etc.) and along with covariate information, one may be able to infer that the missing bit is more likely to be +1 and thus include it in the ordering. Continuing the grid example from the start of this section, by revealing all but one bit per p-value at the start of the procedure, the scientist can possibly notice if *small* masked p-values are randomly scattered or clustered on the grid.

Remark 1. *For any particular setup, like our example of a grid with a cluster of signals, it may be possible to design a better global null test that is perfectly suited for that setting. Hence, we do not claim that our interactive method is the right test to use in all problem setups. Its main advantage is its generality: instead of having to design a new test for each situation (trying to figure out how to optimally combine prior knowledge, structural constraints and covariates from scratch), our general framework provides a simple and flexible alternative.*

The correctness of the test (proof in Appendix A.1.2) hinges on one bit from each p-value being hidden from the scientist. Once this protocol has been run once, and all p-values have been unmasked, the procedure obviously cannot be run a second time from scratch. In other words, our interactive setup does not prevent these and related forms of p-hacking. This is similar to the traditional offline setup, where it is not allowed to pick the global null test after observing the p-values and guessing which test will have the highest power to reject, and if scientists do this anyway and report only the final finding, we would have no way to know whether such inappropriate double-dipping has occurred.

It is worth remarking on the main disadvantage of such a test, relative to (say) the martingale Stouffer test introduced earlier. The interactive test statistic is a sum of coin flips (missing bits) – no matter how strong the signal might be, the interactive test statistic can only increase by one at most. On the other hand, the martingale Stouffer test adds up Gaussians, and if there is a strong signal (very small p-value), it can stop very early. If a relatively good prior ordering is known to the scientist, the martingale Stouffer test should be preferred. However, if the prior knowledge is not in the form of an ordering, but some intuition about how the covariates and p-values may be related or what type of structure the non-nulls may have (if any), then the interactive test can be much more powerful.

The above framework leaves the specific strategy of expanding M_k unspecified, allowing much flexibility. Now, we give one example of how $i_k^\star$ can be chosen based on the available information $\mathcal{F}_k$. One straightforward choice for $i_k^\star$ is the hypothesis not in M_k with the highest posterior probability of being non-null, computed with the aid of a working model, like the Bayesian two groups model, where each p-value p_i is drawn from a mixture of a null distribution F_0 with probability $1-\pi_i$ and an alternative distribution F_1 with probability π_i:

$$p_i \sim (1-\pi_i)F_0 + \pi_i F_1. \tag{8}$$

For example, we can choose F_0 as a uniform and F_1 as a beta distribution. We may further posit a working model that treats π_i as a smooth function of x_i. The masked p-values $g(p_i)$ and the revealed missing bits in $\mathcal{F}_{k-1}$ can be used to infer the other missing bits using the EM algorithm (see Appendix A.7). The missing bits that are inferred to be more likely +1 should be chosen, potentially in accordance with other structural constraints. Importantly, the type-I error is controlled regardless of the correctness of the working model or any heuristics to expand M_k.

2.4 Power guarantees of non-interactive procedures

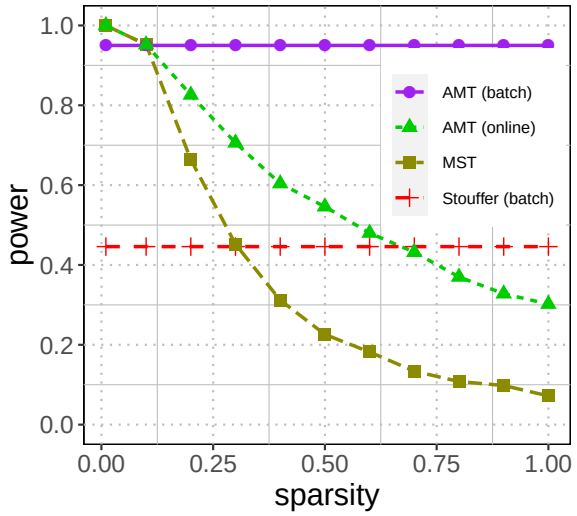
This section is devoted to an analysis of the power of the martingale Stouffer test and the adaptively ordered martingale test. It's hard to analyze the power for the interactively ordered martingale test due to its flexible framework offered to the user: it can have high power if the user specifies a good interactive algorithm, and vice versa. Nevertheless, to demonstrate the advantages of the interactively ordered martingale test, we present numerical results under structured non-nulls in the next section.

Our analysis includes power guarantees in the batch and online settings in a simple Gaussian setup. Specifically, we consider a simple multiple testing problem where each hypothesis is a one sided hypothesis on the mean value of a Gaussian. In this setting, the i-th null hypothesis is that a Gaussian has zero mean, and the alternative is that the Gaussian has a positive mean $\mu_i > 0$.

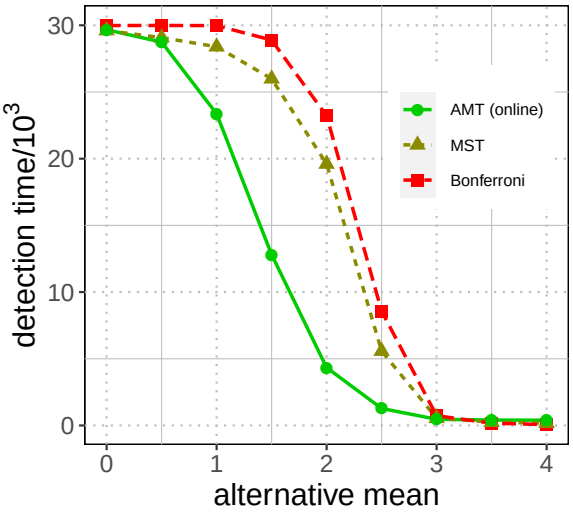
Setting 1. *We observe $Z_1, \ldots, Z_n$ where $Z_i \sim N(\mu_i, 1)$ and wish to distinguish the following hypotheses:*

$$\mathcal{H}_{\mathcal{G}_0} : \mu_i = 0 \textit{ for all } i \in \mathcal{I}, \quad \textit{versus}$$
$$\mathcal{H}_{\mathcal{G}_1} : \mu_i > 0 \textit{ for some } i \in \mathcal{I}.$$

In the remainder of this section, we let $r_i := I(\mu_i > 0)$ indicate the non-null hypotheses. Although we compare the power of various tests in this relatively simple setting, we emphasize that our tests are more broadly applicable to general settings where the p-values are mirror-conservative under the null.



(a) Power comparison in the batch setting when varying the prior ordering. Larger sparsity indicates a worse prior ordering. The AMT procedures (batch and online) adaptively alter the ordering and are more robust to bad quality of orderings than MST. Still, when the prior ordering is great, AMT has lower power than MST because the increments of AMT's test statistic are bounded by $+1$. This phenomenon is mathematically predicted by Theorem 4 and Theorem 5.



(b) Number of hypotheses needed to reject the global null (detection time) in the online setting when varying the alternative mean μ. Each hypothesis has the same probability of being non-null as 5%. The Bonferroni method cannot reject the global null unless μ is greater or equal than 3. AMT is the first to reject the global null when μ is small because it filters the hypotheses by masked p-values. This phenomenon is mathematically predicted by Theorem 6 and Theorem 7.

Figure 3: Illustrative simulations that compare the batch and online martingale Stouffer test (MST) and the adaptively ordered martingale test (AMT) under Setting 1. All plots in this paper present the averaged power (in the batch setting) and averaged rejection time (in the online setting) over 500 repetitions, and the type-I error is $\alpha = 0.05$.

With this setup in place, we now summarize the main results of this section.

- In Section 2.4.1, we focus on the batch setting. In Theorem 4, we compare the power of the martingale Stouffer test with its batch counterpart, showing that when a good a-priori ordering is used the martingale Stouffer test can have much higher power. Our next result, Theorem 5, studies the adaptively ordered martingale test in the batch setting. The adaptively ordered martingale test expands the testing set M_k based on masked p-values, and tests the global null using the missing bits $h(p_i)$. We show that in cases when the signal strength is high, re-ordering by the masked p-values can significantly improve power of the resulting test by ensuring that promising hypotheses are considered early on with high-probability.
- In Section 2.4.2 we turn our attention to the online setting. In Theorem 6, we study the power of a simple online Bonferroni test, and compare this in Theorem 7 with the power of the adaptively ordered martingale test. For the adaptively ordered martingale test, we study the role of the threshold parameter c in the power of the test, characterizing some of the tradeoffs involved in the

choice of this parameter.

Figure 3 visualizes the above power comparisons by two simple simulations in batch and online settings[1]. Details of the batch experiment appear next.

We simulate 10^4 hypotheses with 50 non-nulls ($\mu_i = 3$). The position of the non-nulls is encoded by a *sparsity* parameter: the non-nulls are uniformly distributed in the first *sparsity* $\cdot\, n$ hypotheses. Thus, larger sparsity indicates a poorer prior ordering (the non-nulls are more scattered), and it is expected to result in lower power for order-dependent methods. Indeed, we observe that: (1) two batch procedures (the adaptively ordered martingale test (AMT) in the batch version and Stouffer's test) get the p-values as a set, ignoring the prior ordering, and hence their power is a flat line; (2) the online AMT and the MST procedure uses p-values in the ordering provided to it, and their power degrades as the quality of the ordering degrades; (3) the online AMT is less sensitive to bad prior ordering than the MST because it discards possible nulls based on the masked p-values; but it could still let in many nulls if the discarding threshold is not tight and most nulls are in front, leading to lower power under a worse prior ordering; (4) overall, the AMT procedures (batch and online) are more robust to bad prior ordering than the MST because they adaptively alter the ordering.

Keep in mind that the simulations above and the power analysis below assume no prior knowledge, but the interactively ordered martingale test has higher power when taking advantage of the non-null structure, as shown in Section 2.5.

2.4.1 Power guarantees in the batch setting

We begin by studying the power of the batch, martingale and interactive martingale tests in the batch setting.

The batch Stouffer test and the martingale Stouffer test: The batch Stouffer test simply aggregates the observed $Z_1, \ldots, Z_n$ and compares this with an appropriate threshold. In contrast, the martingale Stouffer test *sequentially* compares partial aggregations with an appropriate threshold.

To state our result compactly, for a specified value γ, we define:

$$C_k^\gamma := 1.7\sqrt{\log\log(2k) + 0.72\log\frac{5.2}{\gamma}}, \tag{9}$$

which corresponds to the curved boundary in (7) divided by $\sqrt{k}$. This quantity grows very slowly with k (at the rate of $\sqrt{\log\log(k)}$) and for all practical purposes can be treated as a "constant". We have the following result:

Theorem 4. *(a)* **Batch Stouffer Test (necessary+sufficient):** *A necessary and sufficient condition for the batch Stouffer test with type-I error α to have at least $1-\beta$ power is that*

$$\sum_{i=1}^{n} r_i\mu_i \geq (Z_\alpha + Z_\beta)n^{1/2}, \tag{10}$$

where $Z_\alpha = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of a standard Gaussian.

(b) **Martingale Stouffer Test (sufficient):** *A sufficient condition for MST to have power at least $1-\beta$ is*

$$\exists\, k \in \{1,\ldots,n\}, \quad \sum_{i=1}^{k} r_i\mu_i \geq \left(C_k^\alpha + C_k^\beta\right)k^{1/2}. \tag{11}$$

[1] https://github.com/duanby/interactive-martingale has R code to reproduce all plots.

(c) **Martingale Stouffer Test (necessary):** *If $\alpha < 1 - \beta$, the power of MST is less than $1 - \beta$ whenever*

$$\forall\, k \in \{1, \ldots, n\}, \quad \sum_{i=1}^{k} r_i \mu_i \leq (C_k^{\alpha} - C_k^{1-\beta}) k^{1/2}.$$

We defer the proof of this result to Appendix A.2.1. Several remarks are in order.

- It is also possible to study the power of the Bonferroni test in the batch setting. A necessary condition for the power of the Bonferroni method to be at least $1 - \beta$ is:

  $$\exists\, k \in \{1, \ldots, n\}, \quad r_k \mu_k \geq Z_{\alpha/n} + Z_{\beta}.$$

  Comparing with the batch Stouffer test, we see that the Bonferroni method has high power when there is at least one large effect, but can have lower power in settings where there are many small non-null effects.
- Comparing condition (10) for the batch Stouffer test with its martingale counterpart (condition (11)), we observe that the batch test rejects when the average of *all* the effects is sufficiently large, while the martingale test rejects as long as *any* cumulative sum is sufficiently large. In cases where a good a-priori ordering is available, the martingale test can have much higher power.

The adaptively ordered martingale test: To ease our calculations, we assume that all the non-nulls have the same mean value, i.e. $\mu_i = \mu$ if $r_i = 1$. We denote the number of non-nulls by N_1 and the nulls by N_0. Let $Z(\nu)$ be a Gaussian random variable with unit variance and mean ν, then the non-nulls are $\{Z_j(\mu)\}$ for $j = 1, \ldots, N_1$ and we let $Z_{(j)}(\mu)$ be the j-th non-null after ordering by its absolute value so that

$$|Z_{(1)}(\mu)| \geq |Z_{(2)}(\mu)| \geq \ldots \geq |Z_{(N_1)}(\mu)|. \tag{12}$$

Suppose that $X \sim \text{Bin}(n, p)$. We let $t_\alpha(n, p)$ denote the α-upper quantile of the Binomial distribution $\text{Bin}(n, p)$, i.e. $\mathbb{P}(X \geq t_\alpha(n, p)) = \alpha$. Recall the definition of C_k^{γ} in equation (9). We define, for $j \in \{1, \ldots, N_1\}$,

$$q_j := \mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|),$$

to be a measure of signal strength. Roughly, the values q_j will be close to 0, if the signal strength μ is large.

Theorem 5. *The adaptively ordered martingale test with level α has at least $1 - \beta$ power if*

$$\exists\, j \in \{1, \ldots, N_1\} : \sum_{s=1}^{j} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right) \geq \left(C_n^{\alpha} + C_n^{\beta/2}\right) \left(j + t_{\beta/(2N_1)}(N_0, q_j)\right)^{1/2}. \tag{13}$$

We prove this result in Appendix A.2.2. Condition (13) gives a reasonably tight sufficient condition for the re-ordering based test to have high power (Figure 4). As expected, when the number of nulls increases (right columns) or the number of non-nulls decreases (bottom rows), the sufficient condition for the signal strength μ to guarantee high power grows.

The condition itself can be difficult to interpret as it depends on the distribution of Gaussian order statistics, as well as on the quantiles of a Binomial distribution. To build some intuition, we consider some simple cases.

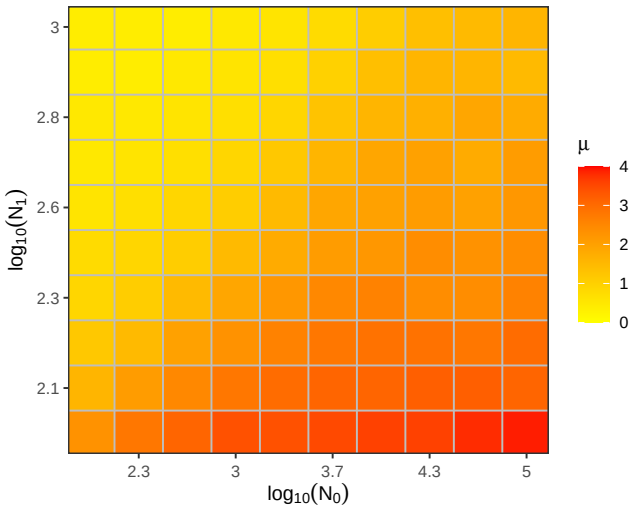


Figure 4: Sufficient signal strength μ for AMT to guarantee both type-I and type-II error control at 0.05 (derived from (13)), when varying the numbers of nulls $N_0 \in [10^2, 10^5]$ and non-nulls $N_1 \in [10^2, 10^3]$. The required signal strength grows when the number of nulls increases or the number of non-nulls decreases.

- In the extreme case, when the signal strength μ is quite large, the re-ordering will ensure that the non-nulls are placed early on with high-probability. In this case, the left-hand side in condition (13) grows linearly with j. On the other hand, if the signal strength is large then the probabilities q_j will be small and we can ignore the term $t_{\beta/(2N_1)}(N_0, q_j)$, so that the right-hand side grows at the rate of roughly $\sqrt{j}$ (ignoring $\log\log$ factors), ensuring that the condition will be satisfied even for a moderate number of non-nulls.
- We provide other conditions that suffice to ensure high power in Appendix A.2.3 by lower and upper bounding the left and right hand sides (respectively). We present one sufficient condition here. Suppose there are sufficient number of non-nulls such that $N_1 \geq 6\left(C_n^{\alpha} + C_n^{\beta/2}\right)^2$, and that the number of nulls is sufficiently large, i.e. that $N_0 > 0.1N_1^2$. A sufficient condition for the adaptively ordered martingale test to have $1-\beta$ power is

$$\mu \geq \sqrt{2\log\left(\frac{N_0}{N_1^2}\right) + 4\log\left(C_n^{\alpha} + C_n^{\beta/2}\right) + 3.45}. \tag{14}$$

  For comparison, the batch Stouffer test requires

$$\mu \geq (Z_{\alpha} + Z_{\beta})\sqrt{\frac{N_0}{N_1^2} + \frac{1}{N_1}}. \tag{15}$$

  Both conditions are stricter if the ratio $\frac{N_0}{N_1^2}$ is large, i.e. in the setting where there are many nulls and few non-nulls. However, the adaptively ordered martingale test requires a signal strength that only grows logarithmically with this ratio.

In Appendix A.2.3, we relate condition (14) to the detection threshold derived in the work of Donoho and Jin [Donoho and Jin, 2015] for the same setting of detecting sparse Gaussian mixtures.

To summarize our findings in the batch setting: the martingale Stouffer test and the adaptively ordered martingale test each require weaker conditions for the same power than the batch Stouffer

test. The martingale Stouffer test relies on a good pre-defined ordering, whereas the adaptively ordered martingale test relies on sufficiently large signal strength to ensure that re-ordering is helpful. We now turn our attention to the online setting.

2.4.2 Power guarantees in the online setting

When testing the global null, the natural test to compare to is the online Bonferroni method, which chooses a sequence of significance levels $\{\alpha_k\}_{k=1}^{\infty}$ such that $\sum_{k=1}^{\infty} \alpha_k = \alpha$, and rejects the global null if

$$\exists\, k \in \mathbb{N} : p_k \leq \alpha_k.$$

The following sections compare the power guarantee of the online Bonferroni method with the martingale Stouffer test and adaptively ordered martingale test. Specifically, we derive necessary conditions for the power of the online Bonferroni test, and compare it with sufficient conditions for the power of our proposed methods – revealing situation where the online Bonferroni has lower power than our proposed methods.

The online Bonferroni method versus the martingale Stouffer test: To better characterize the power of online Bonferroni, we consider two cases:

- Dense non-nulls: the number of non-nulls is infinite,

$$\sum_{k=1}^{\infty} r_k = \infty. \tag{16}$$

- Sparse non-nulls: the number of non-nulls is finite,

$$\sum_{k=1}^{\infty} r_k \leq M < \infty \text{ for some large constant } M. \tag{17}$$

The sparse case yields a stronger necessary condition when the sequence of significance levels satisfies a mild condition that $\{\alpha_k\}_{k=1}^{\infty}$ is nonincreasing.

Unlike previous methods, the online Bonferroni method does not aggregate p-values, so its power guarantee requires conditions on the individual means.

Theorem 6. *Suppose $\alpha \leq (1-\beta)/4$. In the case of dense non-nulls (16), a necessary condition for online Bonferroni to have at least $1-\beta$ power is*

$$\exists k \in \mathbb{N} : r_k \mu_k \geq 0.25 \left(\sqrt{2 \log \left(\frac{k^2}{\alpha} \right)} \right)^{-1}. \tag{18}$$

A stronger necessary condition can be derived for sparse non-nulls (17). If $\{\alpha_k\}_{k=1}^{\infty}$ is nonincreasing, then online Bonferroni can have at least $1-\beta$ power only if

$$\exists k \in \mathbb{N} : \begin{cases} r_k \mu_k \geq 0.4 \sqrt{\alpha_{k^*}}, & \text{if } k \leq k^*, \\ r_k \mu_k \geq \sqrt{\log\left(\frac{k}{4\alpha}\right)} - \sqrt{2 \log\left(\frac{M}{2(1-\beta-3\alpha)}\right)}, & \text{if } k > k^*, \end{cases} \tag{19}$$

where $k^* = M^2/\alpha$, and α_{k^*} is the k^*-th significance level.

In contrast, a sufficient condition for the martingale Stouffer test to have at least $1-\beta$ *power is*

$$\exists\, k \in \mathbb{N} : \sum_{i=1}^{k} \mu_i r_i \geq (C_k^{\alpha} + C_k^{\beta}) k^{1/2}. \tag{20}$$

Remarks:

- Condition (20) is (up to constants) necessary, because if $\alpha < 1-\beta$, the power of the martingale Stouffer test is less than $1-\beta$ whenever

$$\forall\, k \in \mathbb{N} : \sum_{i=1}^{k} r_i \mu_i \leq (C_k^{\alpha} - C_k^{1-\beta}) k^{1/2}.$$

- The necessary condition (18) under dense non-nulls requires a lower bound on $r_k \mu_k$ that decreases at the rate of $(\log k)^{-1/2}$. This lower bound is fairly tight: for an example of sequence $\{\alpha_k\}_{k=1}^{\infty}$ that decreases at the rate of $1/[k(\log k)^2]$, the power of the online Bonferroni test would be one if all hypotheses are non-null when $k > 1$ and the mean value decreases at a slower rate: $\mu_k = (\log k)^{-1/c}$ for any $c > 2$ (see Lemma 4 in Appendix A.3.1).
- The proof of Theorem 6 is in Appendix A.3.1. If asymptotically, the mean values are nonzero but fade as k grows at a fast rate, the online Bonferroni method has little power, but the martingale Stouffer test can have good power. For example, suppose all the hypotheses are non-nulls and $\mu_k = k^{-1/3}/10$. Controlling the type-I error α at 0.15, the online Bonferroni method has power less than 0.6 (by condition (18)) whereas the martingale Stouffer test has power that approaches 1 (by condition (20)).

The adaptively ordered martingale test: For clarity, we consider the same mean value for the non-nulls, $\mu_i = \mu$ if $r_i = 1$. Let a Z score for each hypothesis H_i be $Z_i = \Phi^{-1}(1-p_i)$. Our guarantee on the power for the adaptively ordered martingale test depends critically on the choice of the threshold parameter c (we consider Algorithm 2 with the filtering $\Phi^{-1}(1-g(p_t)) > c$, which is equivalent to $g(p_t) < c'$ for $c' = 1-\Phi(c)$). To concisely state our results, define the following quantities:

$$\begin{aligned}
A(\mu; c) &= \frac{5}{3} \frac{\sqrt{\Phi(-c)}}{\Phi(\mu - c) - \Phi(-\mu - c)}, \\
B(\mu; c) &= \frac{10(\Phi(\mu - c) + \Phi(-\mu - c) - 2\Phi(-c))}{9(\Phi(\mu - c) - \Phi(-\mu - c))^2} \vee \frac{25}{(\Phi(\mu - c) + \Phi(-\mu - c))^2}, \\
T(\beta; c) &= \frac{0.79 \log(15.57/\beta)\Phi^2(-c) + 0.4}{\Phi^4(-c)}.
\end{aligned}$$

For a reasonable choice of the threshold parameter, i.e., setting $c = \mu$ for instance, we note that the quantity $B(\mu; \mu)$ is upper bounded by a universal constant (when $\mu > 0$). On the other hand, the quantity $A(\mu; \mu)$ decays exponentially for large signal strength, i.e., when $\mu > 0.25$ we have:

$$A(\mu; \mu) \leq e^{-\mu^2/4}. \tag{21}$$

With these quantities in place, we now state our main result on the power of the adaptively ordered martingale test.

Theorem 7. *A sufficient condition for the adaptively ordered martingale test with type-I error α and threshold parameter c to have $1-\beta$ power is that:*

$$\exists\, k \geq T(\beta; c) : \sum_{i=1}^{k} r_i \geq A(\mu; c)\left(C_k^{\alpha} + C_k^{\beta/3}\right) k^{1/2} + B(\mu; c)\left(C_k^{\alpha} + C_k^{\beta/3}\right)^2 k^{-1/2}.$$

It is interesting to compare the above result with the necessary condition for the martingale Stouffer test: the power of MST is less than $1-\beta$ if

$$\forall\, k \in \mathbb{N} : \quad \sum_{i=1}^{k} r_i \leq \mu^{-1}\left(C_k^{\alpha} - C_k^{1-\beta}\right) k^{1/2}. \tag{22}$$

Both right-hand sides grow at the rate of $k^{1/2}$ (ignoring $\log\log$ factors), but the μ-dependent term $\exp(-\mu^2/4)$ for AMT (derived in bound (21) for $A(\mu;\mu)$) is much smaller than the corresponding $1/\mu$ term in condition (22) for MST. As a consequence, the adaptively ordered martingale test will have higher power when the non-nulls have sufficiently large mean values but are sparse.

To summarize the basic insights we derive in this section, we find that both in the batch setting and the online setting, the martingale Stouffer test and the adaptively ordered martingale test require weaker conditions than their classical counterparts to guarantee the same power when the non-nulls are sparse. The martingale Stouffer test relies on good prior knowledge to order the hypotheses, while the adaptively ordered martingale test uses masked p-values to generate a good ordering. The theoretical analyses in this section discuss the case with no prior knowledge, and the simulations in the next section delve deeper into the setting where the non-nulls are structured.

2.5 Numerical simulations

While the martingale Stouffer test can only use prior knowledge in the form of non-null probabilities for each hypothesis, the interactively ordered martingale test combines (a) side covariate information (which could include prior non-null probabilities in working model (8) as a component) with (b) structural constraints on the unknown non-null set, and (c) masked p-values, to infer whether a hypothesis is non-null and thus include it earlier in the ordering. Here, we demonstrate that prior structural constraints can help the interactively ordered martingale test attain a higher power than the martingale Stouffer test and some classical methods.

We first consider the batch setting and use two non-null structures as simple examples: a blocked structure within a grid and a hierarchical structure within a tree; and we discuss similar structures in the online setting. For each of these, we customize a heuristic strategy to expand M_k in the interactively ordered martingale test (recalling that type-I error is controlled regardless of the heuristic used, and only power is affected).

2.5.1 Clustered non-nulls in a grid of hypotheses

Consider the setting where the hypotheses are arranged in a rectangular grid, and if the null is false, then the non-nulls form a single coherent cluster. This is a common structure which, as a hypothetical example, is a reasonable belief when trying to detect if there is a tumor in a brain image. Here, the covariates x_i are simply the two-dimensional location of the hypothesis H_i on the grid. The blocked

non-null structure is utilized in specifying the posterior probability of being non-null using model (8) by constraining the prior non-null probabilities π_i to be a smooth function of x_i. Details can be found in Appendix A.7.

The block structure is also imposed in the strategy of interactively expanding M_k such that M_k forms a single connected component. The interactively ordered martingale test expands M_k by only including possible non-nulls that are on the boundary of M_k (see Figure 5 for example).

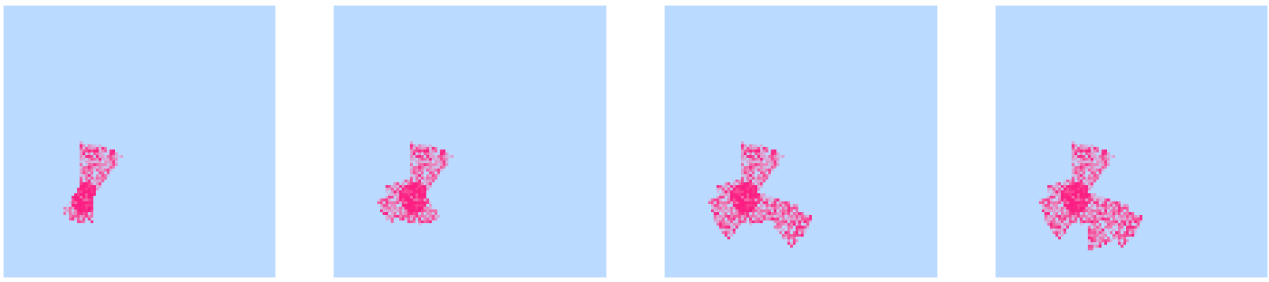
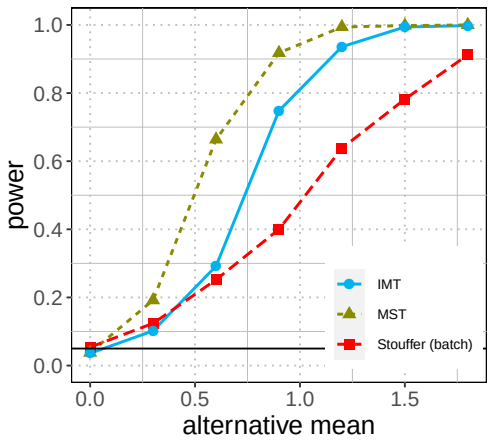


Figure 5: Visualization of the interactively ordered martingale test under the block structure: the hypotheses in M_k, which interactively expands (darker color indicates a lower p-value and possible non-null).

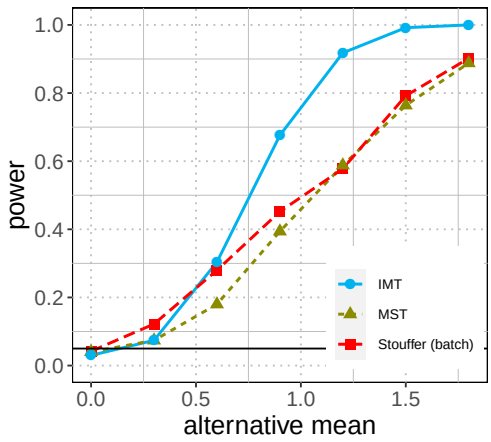
We compare the interactively ordered martingale test with the martingale Stouffer test and the batch Stouffer test. We use the martingale Stouffer test (MST) with a preordering that starts at the center of the grid, and the following hypotheses are included into the preordering in randomly chosen (data-independent) directions such that the hypotheses always form a single cluster. Our simulation has 10^4 hypotheses arranged in a 100×100 grid with a disc of about 150 non-nulls, placed either at the grid center and or at a corner of the grid. We use Setting 1 as defined in Section 2.4, where we varied the non-null mean as $(0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8)$.

The interactively ordered martingale test has high power for both positions of the non-null block, whereas the power of martingale Stouffer test drops quickly when the block is not at the center (Figure 6), which is because the martingale Stouffer test does not have information of the block position (its preordering starts from the center by default), whereas the interactively ordered martingale test uses masked p-values to learn the block position. It is worth noting that even with a bad preordering, the martingale Stouffer test does not do worse than the batch version, but has much higher power with a good preordering.

Remark 2. *As mentioned in the introduction, we do not intend to claim that the interactively ordered martingale test is in any sense the "best" test for this problem setting. It is possible, or even likely, that several other generic tests (Bonferroni, chi-squared, higher criticism, or many others) or specialized tests (scan statistics) might have higher power. We discuss the comparison with two recent methods: the adaptively weighted Fisher test [Fang et al., 2019; Huo et al., 2020; Li and Tseng, 2011] and the weighted Higher Criticism [Zhang et al., 2020] in Appendix A.8. Our goal in this section is to demonstrate the tradeoffs between the batch and martingale versions of the same test (Stouffer in this case), and the interactive versus preordered martingale tests. Also note that the power of our martingale tests depends crucially on the preordering, or on the model and heuristic used to form the ordering interactively, and perhaps better models/algorithms might further improve the power of our own tests. We chose settings that are easy to visualize for intuition, keeping in mind that our tests apply to any general covariates x_i, and prior knowledge or structural constraints, any working models, etc.*



(a) The power against non-null signal. The non-null block is in the grid center.



(b) The power against non-null signal. The non-null block is in the grid corner.

Figure 6: Testing the interactively ordered martingale test (IMT), the martingale Stouffer test (MST), and the batch Stouffer test with varying alternative mean under a block non-null structure (batch setting). The MST has lower power when the non-null is not in the center, whereas the IMT has high power in both cases. Type-I error corresponds to the power when the alternative mean value is zero. The horizontal line corresponds to the target type-I error level $\alpha = 0.05$.

2.5.2 A sub-tree of non-nulls in a tree of hypotheses

In applications such as wavelet decomposition, the hypotheses can have a hierarchical structure, where the child can be a non-null only if its parent is a non-null. The hierarchical structure is again encoded in modeling the posterior probability of being non-null (8) by adding a partial order constraint on π_i that

$$\pi_i \geq \pi_j, \quad \text{if } i \text{ is the parent of } j.$$

Also, the hierarchical structure is imposed in the strategy of update M_k such that M_k should keep as a sub-tree. Specifically, we compare the posterior probabilities of being non-null for all the leaf nodes of M_k and choose the highest one.

We compare the interactively ordered martingale test with the martingale Stouffer test and Stouffer's test, where the martingale Stouffer test order the hypotheses by level and from left to right within level. We simulate a tree of five levels (the root has twenty children and three children for each parent node after that) with over 800 nodes in total and 7 of them being non-nulls. Each node tests if a Gaussian is zero mean as described in Setting 1, where we vary the mean value for the non-nulls as $(0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4)$. The interactively ordered martingale test is implemented without modeling the posterior probabilities of being non-null for the sake of computational cost. The interactively ordered martingale test has a higher power especially when the signal is strong so that the masked p-values provide a better guide on the M_k update (Figure 7).

The interactively ordered martingale test with modeling is implemented on a smaller tree with 121 nodes (five levels and three children for each parent node) and 7 of them being non-nulls on a subtree. We consider two types of hierarchical non-null structure: one with the probability of being non-null decreasing down the tree, and one with increasing probability, which means the parent cannot be a non-null unless its children are non-nulls. The result is consistent with the above: the interactively ordered martingale test has higher power than the non-interactive martingale Stouffer test (Figure 8). Compared with decreasing probability of being non-null, both methods have lower power for the tree

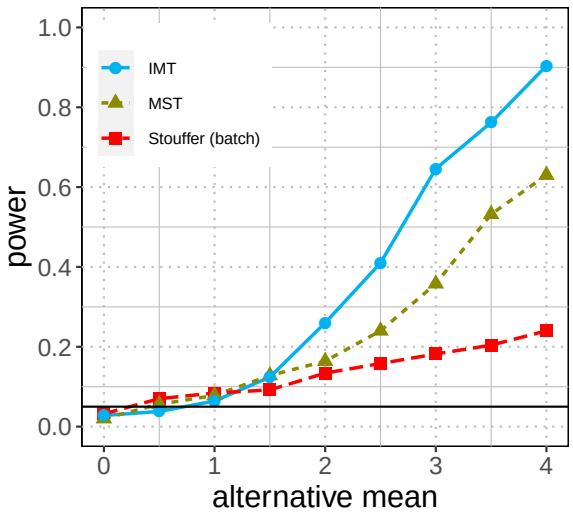
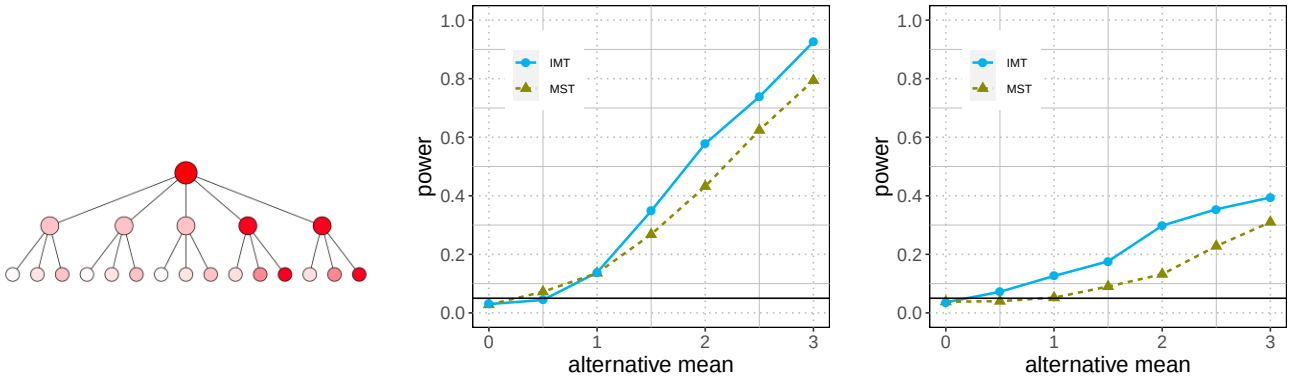


Figure 7: Power of the interactively ordered martingale test (IMT), the martingale Stouffer test (MST), and the batch Stouffer test under a hierarchical structure. Hypotheses form a fixed tree (batch setting) with non-nulls only on a sub-tree. When the alternative mean is big, masked p-values and the hierarchical non-null structure lead to a good ordering and hence high power for the IMT.



(a) Hypothesis tree with decreasing non-null probability, which is marked by fading red nodes.

(b) Power against alternative mean in a hypothesis tree with decreasing probability of being non-null.

(c) Power against alternative mean in a hypothesis tree with increasing probability of being non-null.

Figure 8: Hypothesis tree in the batch setting with decreasing/increasing probability of being non-null. Testing the interactively ordered martingale test (IMT) with a model for the posterior probability of being non-null, which has higher power than the martingale Stouffer test (MST) in both cases.

with an increasing probability of being non-null, because in the latter case, the non-nulls gathered at later generations where there are more nulls and the non-nulls are sparser.

2.5.3 Structures in the online setting

Recall that in the online setting, a potentially infinite number of hypotheses arrive, and the adaptively ordered martingale test and interactively ordered martingale test use some discarding rules to only allow promising non-nulls entering M_k. This section presents two examples of non-null structures in the online setting, and demonstrates the power of the interactive test as follows.

Blocks of non-nulls in a growing sequence of hypotheses. Suppose the non-nulls arrive as blocks. In other words, the next hypothesis is more likely to be a non-null if the last arrived hypothesis is truly non-null; and vise versa. Let the discarding rule in the interactively ordered martingale test be $g(p_t) > c_t$, where $c_t = c = 0.05$ by default. The interactively ordered martingale test adjusts c_t for $t > 10$ based on previous p-values: it alleviates the discarding rule by increasing c_t to $2c$ if the ten p-values prior to t $(p_{t-10}, \ldots, p_{t-1})$ are all less than 0.1; otherwise, it decreases c_t to $c/4$. For a fair comparison, the discarding threshold in the adaptively ordered martingale test is set to $c = 0.05$.

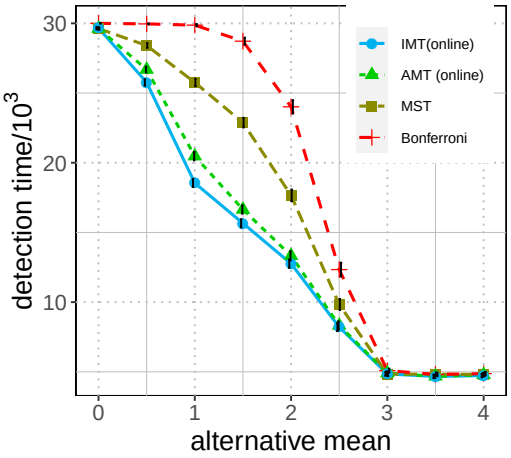


Figure 9: Number of hypotheses needed to reject the global null (detection time) in the online setting of the interactively ordered martingale test (IMT), the adaptively ordered martingale test (AMT), the martingale Stouffer test (MST), and the Bonferroni test when varying the alternative mean μ. The non-nulls arrive in blocks, and on average, every 10^4 hypotheses contain a block of 500 non-nulls. The length of the error bar is two standard error. The interactively ordered martingale test is the first to reject the global null because it incorporates the block structure and adjusts the discarding threshold based on past p-values.

The interactively ordered martingale test is the first to reject the global null since its discarding rule accounts for the block structure (see Figure 9). This advantage is more evident when the non-null signal is mild ($\mu < 3$), where the prefixed discarding rule in the adaptively ordered martingale test might be too strict or lenient, while the interactively ordered martingale test can adjust the rule accordingly. In practice, the adjustment on the discarding threshold can also utilize side information and prior knowledge, if provided.

A sub-tree of non-nulls in a growing tree of hypotheses. The online tree grows a new level at every step, with the probabilities of being non-null no bigger than their parents. For an arriving level k, the interactively ordered martingale test models the posterior probability of being non-null $\pi_j^{(k)}$ for the new hypothesis H_j by equation (8), where the prior probability of being non-null is the same as its direct parent H_i from the level $k-1$,

$$\pi_j^{(0)} = \pi_i^{(k-1)}, \quad \text{if } i \text{ is the parent of } j.$$

For simplicity, we set the discarding rule in the interactively ordered martingale test to be $\pi_i^{(k)} < c$ where $c = 0.6$ as a default. That is, hypothesis with $\pi_i^{(k)} < 0.6$ are omitted. We compare the interactively ordered martingale test with the martingale Stouffer test and a classical method, the online Bonferroni

method (with the sequence of significance levels $\{\alpha_k\}_{k=1}^{\infty}$ decreases at the rate of $1/[k(\log k)^2]$). In the online setting, their performances are assessed by the averaged number of hypotheses required to reject the global null (detection time); the smaller the better.

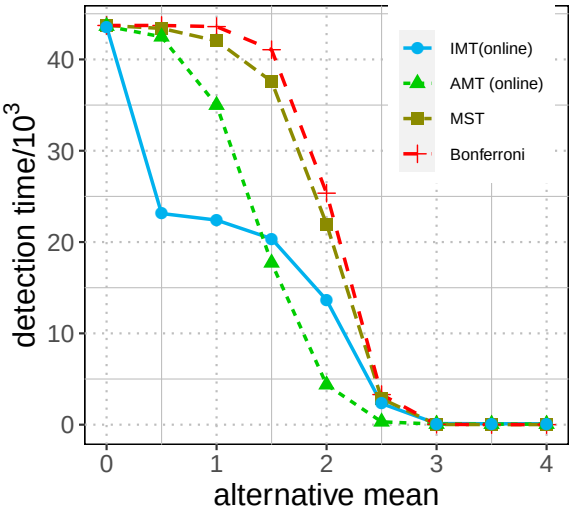


Figure 10: Number of hypotheses needed to reject the global null (detection time) in the online setting of the interactively ordered martingale test (IMT), the adaptively ordered martingale test (AMT), the martingale Stouffer test (MST), and the Bonferroni test when varying the alternative mean in a growing hypothesis tree (online setting). IMT incorporates the hierarchical structure of non-nulls, so it is the first to reject the global null when the non-null signal is mild ($\mu < 2$).

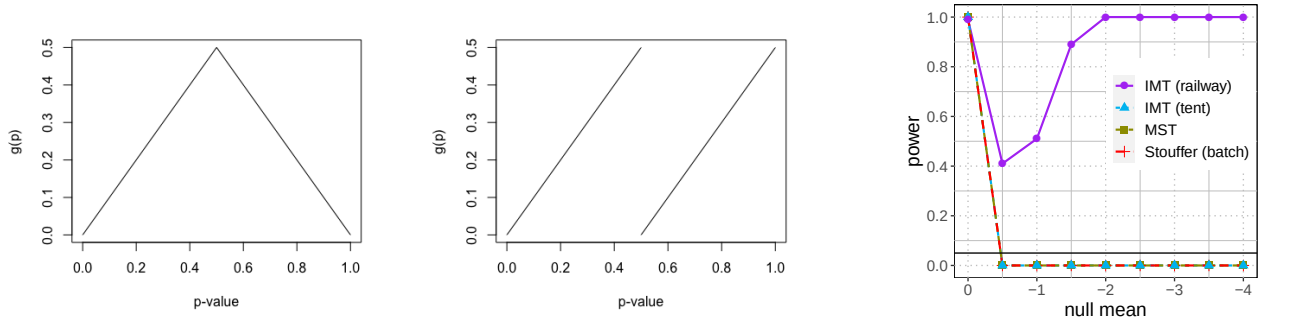
We simulate the online tree with forty children for the root node and three children for each parent node after that. The probability of being non-null for the first generation children is set to 0.1 for 30 children and 0.9 for the other 10 children. The ongoing three children of each node reduce the probability of being non-null as by a proportion of $100\%, 20\%, 0\%$. Each node tests if a Gaussian is zero mean as described in Setting 1, where we vary the mean value for the non-nulls as $(0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4)$. The interactively ordered martingale test needs much shorter time when the non-null signal is not strong ($\mu < 2$) because it incorporates the hierarchical structure and estimates the probability of an arriving hypothesis being non-null with the aid of the data from its ancestors (Figure 10). When the alternative mean is large, p-values themselves provide strong evidence of non-null, while the algorithm using the tree structure would treat all children from a non-null parent as promising non-nulls while at least one of them is null in our simulated example. Thus, the online AMT that uses only the p-value information can have better performance when the alternative mean is large.

Overall, both in the batch setting and the online setting, the interactively ordered martingale test has a higher detection power than the martingale Stouffer test, Stouffer's test, and the online Bonferroni method, provided with structured alternatives. We again remark the advantage of the interactively ordered martingale test in practice where prior knowledge often exists in various forms. The interactively ordered martingale test is highly flexible in that it allows modifications to the strategy of expanding M_k, at any step and with any form as a human analyst (or a program) wants to. The next section demonstrates one more advantage of the interactively ordered martingale test under the *conservative* nulls (see definition in the next section).

2.6 Robustness to conservative nulls

In all the above simulations, the nulls have uniformly distributed p-values, but in practice they could be stochastically larger than uniform (condition (2)) or mirror-conservative (condition (3)); both are henceforth referred to as "conservative nulls". For simplicity, this section focuses on the conservative null with an increasing density, which satisfies both descriptions in condition (2) and condition (3). Such conservative nulls diminish the detection power of many batch global null tests like Fisher's and Stouffer's methods. For example, each term in Stouffer's test is $\Phi^{-1}(1-p)$, whose value can be smaller than -2 if the p-value is bigger than 0.98; thus as the nulls grow more conservative and their p-values closer to one, its power can quickly drop to zero.

To examine the effect of conservative nulls on the interactively ordered martingale test, we first propose an alternative definition of a masked p-value as $\widetilde{g}(p) := \min(p, (p+\frac{1}{2}) \text{mod} 1)$. Recalling that $g(p) = \min(p, 1-p)$, we call g and $\widetilde{g}$ as the tent and railway functions respectively (see Figure 11a, Figure 11b). Note that if the p-value is exactly uniformly distributed, $\widetilde{g}(p)$ is still independent of $h(p)$, and $g(p)$ has the same distribution as $\widetilde{g}(p)$, and so all previous results still hold with the new masking function in place of the old one. (The error control when using the railway masking function can be found in Appendix A.1.3 for uniform and conservative p-values.) However, when the p-values are conservative, the new masking function has a clear advantage. To see this, consider a p-value of 0.99. The original masked p-value would be 0.01, thus causing the methods to potentially confuse this with a non-null masked p-value, but the new masked p-value would be 0.49, which the methods would easily exclude as being a null.



(a) The original masking function (tent).

(b) The modified masking function (railway).

(c) Power when varying the null means (negative values indicate conservative nulls).

Figure 11: Comparing the interactively ordered martingale test (IMT) with tent and railway masking functions, the martingale Stouffer test (MST), and Stouffer's test for the robustness to conservative nulls. The IMT with railway function is more robust.

As an example, we consider the simple case with no prior knowledge and simulate 1000 hypotheses with 100 non-nulls. Each hypothesis is a one sided hypothesis on whether a Gaussian is zero mean as described in Setting 1. The alternative mean values are set to 1.5. The mean values for nulls are negative so that the resulting null p-values are conservative. We tried nine values from 0 to -4 for the mean of nulls, with a smaller value indicating higher conservativeness. Figure 11c compares the power of the interactive martingale test with tent and railway functions, the martingale Stouffer test and Stouffer's test. The power of most tests drops sharply to zero, but the power of interactively ordered martingale test with the new railway function initially dips and then improves. The reason for the initial dip is that

the increasingly conservative nulls influence the interactive martingale test in two opposite directions: (a) more null $h(p)$ values are now equal to -1 (instead of being ± 1 with equal probability), and this hurts power because including a null $h(p)$ in the martingale almost always lowers its value (instead of increasing and lowering its value with equal probability), (b) as the p-value gets more conservative, $g(p)$ will approach 0.5 for nulls, allowing the tests to easily distinguish between the non-nulls and the nulls to increase the power. When the p-values are only slightly conservative, effect (a) dominates and hurts power, causing the initial dip in power in Figure 11c.

2.7 Anytime-valid p-values and safe e-values

In this paper, we defined the problem as testing the global null at a predefined level α. Instead, we could ask the test to output a sequential or anytime p-value for the global null, which is a sequence of p-values $\{\mathfrak{p}_t\}_{t=1}^{\infty}$ that are valid at any stopping time. We use $\mathfrak{p}_t$ to differentiate it from p_t — the latter is the input to our global null test, the former is the desired output of our global null test. Specifically, $\mathfrak{p}_t$ is a function of $p_1, \ldots, p_t$, such that if $p_1, \ldots, p_t$ are all null, then $\mathfrak{p}_t$ will be a valid p-value (its distribution will be stochastically larger than uniform), and this fact will be true uniformly over t.

Recall that all of the proposed procedures follow the same form; we reject the global null if

$$\exists k \in \{1, 2, \ldots\} \text{ s.t. } S_k > u_\alpha(k),$$

where S_k is a martingale under the global null and $u_\alpha(k)$ is a sequence of upper bounds at level α. The anytime p-value $\mathfrak{p}_t$ at time t is defined by the smallest level at which our test would have rejected the null at or before time t.

Definition 1. *The p-value $\mathfrak{p}_t$ can be defined as the smallest level α at which the test would have rejected at or before time t:*

$$\mathfrak{p}_t = \inf\{\alpha : \exists k \in \{1, \ldots, t\} \textit{ s.t. } S_k > u_\alpha(k)\}. \tag{23}$$

Viewing $u_\alpha(k)$ as a function of two variables k, α, we define an inverse function at a fixed k with respect to the level α as

$$u^{-1}(S; k) = \alpha \text{ iff } u_\alpha(k) = S,$$

which is unique for a given input S since the bound $u_\alpha(k)$ is continuous and strictly decreasing in α. Then the p-value at time t can be computed as

$$\mathfrak{p}_t = \min_{1 \le k \le t} \{u^{-1}(S_k; k)\}.$$

As one example, if $u_\alpha(k)$ is the linear bound as in test (6), its inverse is

$$u^{-1}(S; k) = \exp\left\{-2m\frac{S^2}{(k+m)^2}\right\}.$$

The p-value sequence $\{\mathfrak{p}_t\}_{t=1}^{\infty}$ has the following nice properties,

1. the anytime p-values decrease with time:

$$\mathfrak{p}_{t+j} \le \mathfrak{p}_t \text{ for all } j, t > 0.$$

2. $\inf_{t \in \mathcal{I}} \mathfrak{p}_t$ is also a valid p-value for the global null:

$$\mathbb{P}(\inf_{t \in \mathcal{I}} \mathfrak{p}_t \le x) \le x \;\equiv\; \mathbb{P}\{\exists t : \mathfrak{p}_t \le x\} \le x, \qquad \text{for all } x \in (0, 1).$$

In fact $\inf_{t \in \mathcal{I}} \mathfrak{p}_t$ is the global p-value: the smallest level α at which the test would ever reject:

$$\inf_{t \in \mathcal{I}} \mathfrak{p}_t = \inf\{\alpha : \exists k \in \{1, 2, \ldots\} \text{ s.t. } S_k > u_\alpha(k)\}.$$

3. for any arbitrary stopping time $\tau \in \mathcal{I}$, $\mathfrak{p}_\tau$ is a valid p-value:

$$\mathbb{P}(\mathfrak{p}_\tau \leq x) \leq x, \qquad \text{for all } x \in (0,1).$$

The second property implies that the p-value at any time t is a valid p-value. Recalling that fixed-sample p-values are dual to fixed-sample confidence intervals, it is also the case that anytime p-values are dual to anytime confidence intervals. These ideas are explored and explained in depth by Howard et al. [2020b]. An alternative to anytime p-values, called safe e-values, was recently proposed by Grünwald et al. [2019], and their relationship to confidence sequences, sequential tests and anytime p-values was detailed by Ramdas et al. [2020]. Specifically, optionally stopped nonnegative supermartingales, which underlie all our bounds, yield safe e-values. The main takeaway message for our current paper is that all aforementioned tests can be reformulated as calculating anytime p-values or safe e-values. To exactly recover our level α tests, we just stop and reject at the first time that $\mathfrak{p}_t \leq \alpha$ (or equivalently, the e-value exceeds $1/\alpha$).

2.8 Alternative masking functions

In most of this paper, we have considered one way of decomposing p-value as equation (4), but interactive tests can be developed for other decompositions. Shafer et al. [2011] discuss a class of *calibrators* (functions) for the p-values $f : [0,1] \to [0,\infty)$ such that f is non-increasing and $\int_0^1 f(p)dp \leq 1$. They consider a "product-martingale" $\prod_{i=1}^k f(p_i)$ and reject the null if

$$\exists k \in \mathbb{N} : \prod_{i=1}^{k} f(p_i) \geq \alpha^{-1},$$

which uses Ville's inequality (an infinite-horizon uniform extension of Markov's inequality). For each calibrator f, an interactive test can be developed by viewing $f(p)$ as the missing bit for inference and finding the corresponding masked p-value $g(p)$ for interactive ordering. Type-I error is controlled if the pair of $f(p)$ and $g(p)$ are *mean independent* under the null:

$$\mathbb{E}(f(p) \mid g(p)) = \mathbb{E}(f(p)). \tag{24}$$

Lei et al. [2020] provide a recipe to construct mean independent $g(p)$ given any calibrator. The interactive test given a pair of $f(p)$ and $g(p)$ follows the same procedure as Algorithm 3, with the rejection rule at each step k changed to

$$\prod_{i=1}^{M_k} f(p_i) \geq \alpha^{-1}. \tag{25}$$

or equivalently

$$\sum_{i=1}^{M_k} \log f(p_i) \geq \log(\alpha^{-1}).$$

We explore a class of calibrators f_c parameterized by a constant $c \in (0,1)$:

$$f_c(p) = cp^{c-1}. \tag{26}$$

In an interactive test, $\log f_c(p_i)$ is viewed as playing the role of the missing bit for inference (even though it is technically not one bit, we use the same terminology for simplicity). To calculate the

corresponding masked p-value, we define function $H_c(x) = x^c - x$ for $x \in [0, p_*]$, where p_* is the solution of $\log f_c(p) = 0$. The masked p-value is defined as

$$g_c(p_i) = \begin{cases} p_i, \text{ if } p_i \le p_* \\ s(p_i), \text{ otherwise,} \end{cases}$$

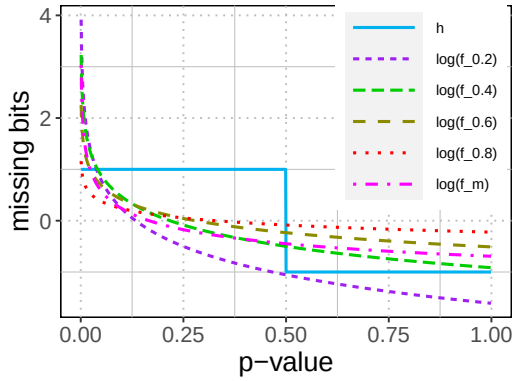
where for any $p_i > p_*$, we define $s(p_i)$ as the unique solution of $H_c(x) = H_c(p_i)$ within the range $[0, p_*]$. Both p_* and $s(p_i)$ can be obtained numerically by a simple binary search since $\log f_c(p)$ and $H_c(x)$ are monotonic. To compare different options of missing bits, Figure 12 shows the maps for original $h(p_i)$ (one bit) and the log term $\log(f_c(p_i))$, since they play similar roles in the interactive tests as forming cumulative sum statistics.

Different choices of missing bit and the corresponding masked p-value reflect a tradeoff between the information of p-values allocated for inference and interactive ordering. Compared with one bit h defined in equation (4), f_c maps small p-values to large value (Figure 12a), so that an evident non-null leads to a big increment in the test statistics and higher likelihood of being detected. In other words, f_c takes more information from p-values than h for inference. However, the corresponding masked p-value is less informative to suggest a good ordering. It's because a wider range of p-values that are bigger than 0.5 (from nulls) would have small masked p-value (Figure 12b), which mixes with the actual small p-values and makes it harder to select possible non-nulls. As c approaches zero, more information is allocated to inference and less for interactive ordering.

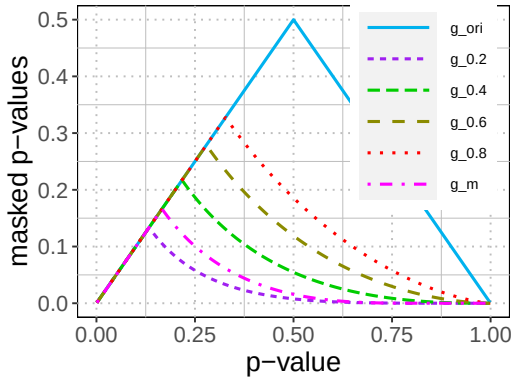
We also consider a mixture of f_c, denoted as f_m:

$$f_m(p) = \int_0^1 c p^{c-1} dc \equiv \frac{1 - p + p \log p}{p(\log p)^2}. \tag{27}$$

The corresponding masked p-value $g_m(p)$ can be calculated using the same formula as above except for a new definition of $H_m(x)$ as $\frac{x-1}{\log x} - x$. As shown in Figure 12, the amount of information that f_m takes for inference is between $f_{0.2}$ and $f_{0.4}$.



(a) Different maps from p-value to the missing bit.



(b) Corresponding maps from p-value to the masked p-value.

Figure 12: Different choices of missing bit and its corresponding masked p-value. When small p-values (possible non-nulls) are more evident when measured by one choice of the missing bit, they are less distinctive when looking at the corresponding masked p-values.

We compare the interactively ordered martingale tests using different missing bits: (a) the original one bit $h(p_i)$ defined in equation (4); (b) $f_c(p_i)$ where we vary parameter c as $(0.2, 0.4, 0.6, 0.8)$; and (c)

the mixed missing bit $f_m(p_i)$. Our simulation uses the structured hypotheses with a cluster of non-nulls (described in Section 2.5.1). The highest power comes from the test with the original definition of the missing bit: $h(p_i) = 2 \cdot 1\{p_i < 0.5\} - 1$ (Figure 13).

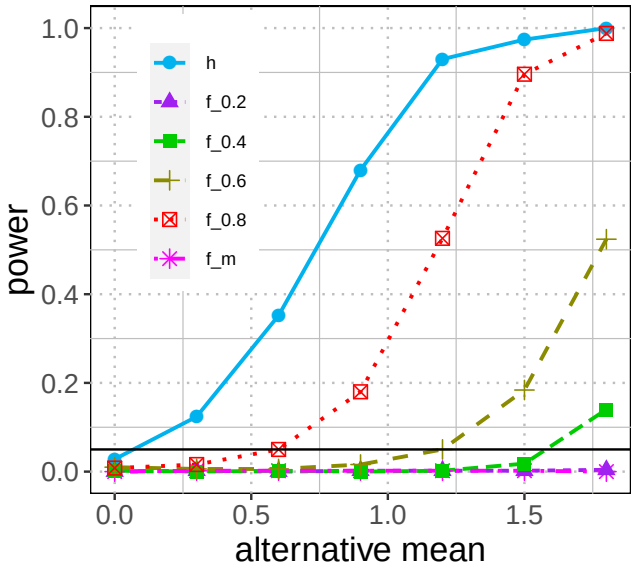


Figure 13: Power of interactive tests using different missing bits. Under the block structure of non-nulls as described in Section 2.5.1, the IMT with the original missing bit defined in equation (4) has the highest power.

However, given that there is a tradeoff between the information contained in the missing bit and the masked p-value, and that the masked p-value is used together with the prior knowledge for a good ordering, we conjecture that the performance of tests with different missing bits depends on the amount of prior knowledge. When the prior knowledge is informative to order the hypotheses, the test with most of the information in the missing bit has a higher power (an example is the martingale Stouffer test, which has the highest power in Figure 6a). We leave the following as an open question: under different types of prior knowledge, does there exist and can one determine an "optimal" p-value decomposition that leads to the highest power?

2.9 Summary

We have introduced martingale analogs of some classical global null tests, and used these to build adaptively ordered martingale tests through the idea of masking. These are further generalized to a protocol for interactively ordered martingale tests that possess the following interesting advantages:

- It is a general global null testing framework that can utilize any types of covariates, structural constraints, prior knowledge and repeated user interaction guided by a posited working model, all while provably controlling the type-I error.
- It permits the use of Bayesian modeling techniques while retaining frequentist error guarantees.
- It applies to both the batch and online settings.

- It is robust against conservative nulls.
- It has favorable theoretical power guarantees in simple settings, and performs well in simulations.

In fact, in most of this paper, we do not need to know the null distribution of the underlying test statistics and be tied to working with p-values as inputs. Given test statistics $T_i \in \mathcal{R}_n$ for each hypothesis H_i, the framework of the interactively ordered martingale test applies as long as there exits two functions $h : \mathcal{R}_n \to \{-1, 1\}$ and $g : \mathcal{R}_n \to \mathcal{R}$ such that

$$\mathbb{E}\left[h(T_i) \mid g(T_i)\right] \leq 0 \quad \text{for all } i \in \mathcal{I}. \tag{28}$$

As an example, if the distribution of the test statistic T_i is symmetric under the null (such as Gaussian with unknown covariance, a t distribution with unknown degrees of freedom, or a centered Cauchy), we can still use sign(T_i) and $|T_i|$ as $h(T_i)$ and $g(T_i)$ respectively. Indeed, type-I error control (Theorem 3) still holds in this setting, since $h(T_i)$ and $g(T_i)$ for the aforementioned decompositions are independent under the null.

Thus far, interactive tests are developed for FDR control and globel type-I error control. The next chapter considers another commonly used error metric: familywise error rate (FWER).

3 Familywise Error Rate Control by Interactive Unmasking

3.1 Introduction

Hypothesis testing is a critical instrument in scientific research to quantify the significance of a discovery. Recent work on testing focuses on a large number of hypotheses, referred to as *multiple testing*, driven by various applications in Genome-wide Association Studies, medicine, brain imaging, etc. (see [Farcomeni, 2008; Goeman and Solari, 2014] and references therein). In such a setup, we are given n null hypotheses $\{H_i\}_{i=1}^n$ and their p-values $P_1, \ldots, P_n$. A multiple testing method examines the p-values, possibly together with some side/prior information, and decides whether to reject each hypothesis (i.e., infers which ones are the non-nulls). Let $\mathcal{H}_0$ be the set of hypotheses that are truly null and $\mathcal{R}$ be the set of rejected hypotheses, then $V = |\mathcal{H}_0 \cap \mathcal{R}|$ is the number of erroneous rejections. This paper considers a classical error metric, *familywise error rate*:

$$\text{FWER} := \mathbb{P}(V \geq 1),$$

which is the probability of making any false rejection. Given a fixed level $\alpha \in (0, 1)$, a good test should have valid error control that FWER $\leq \alpha$, and high *power*, defined as the expected proportion of rejected non-nulls:

$$\text{power} := \mathbb{E}\left(\frac{|\mathcal{R} \backslash \mathcal{H}_0|}{|[n] \backslash \mathcal{H}_0|}\right),$$

where $[n] := \{1, \ldots, n\}$ denotes the set of all hypotheses.

Most methods with FWER control follow a prespecified algorithm (see, for instance, [Bretz et al., 2009; Goeman and Solari, 2011; Hochberg, 1988; Holm, 1979; Tamhane and Gou, 2018] and references therein). However, in practice, analysts tend to try out several algorithms or parameters on the same dataset until results are "satisfying". When a second group repeats the same experiments, the outcomes are often not as good. This problem in reproducibility comes from the bias in selecting the analysis tool: researchers choose a promising method after observing the data, which violates the validity of error control. Nonetheless, data would greatly help us understand the problem and choose an appropriate method if it were allowed. This motivates us to propose an interactive method called the *i-FWER test*, that (a) can use observed data in the design of testing algorithm, and (b) is a multi-step procedure such that a human can monitor the performance of the current algorithm and is allowed to adjust it at any step interactively; and still controls FWER.

The word "interactive" is used in many contexts in machine learning and statistics. Specifically, multi-armed bandits, active learning, online learning, reinforcement learning, differential privacy, adaptive data analysis, and post-selection inference all involve some interaction. Each of these paradigms has a different goal, a different model of interaction, and different mathematical tools to enable and overcome the statistical dependencies created by data-dependent interaction. The type of interaction proposed in this paper is different from the above. Here, the goal is to control FWER in multiple testing. The model of interaction involves "masking" of p-values followed by progressive unmasking (details in the next paragraph). The technical tools used are (a) for p-values of the true nulls (*null p-values*), the masked and revealed information are independent, (b) an empirical upper bound on the FWER that can be continually updated using the revealed information.

The key idea that permits interaction while ensuring FWER control is "masking and unmasking", proposed by Lei and Fithian [2018]; Lei et al. [2020]. In our method, it has three main steps and alternates between the last two (Figure 14):

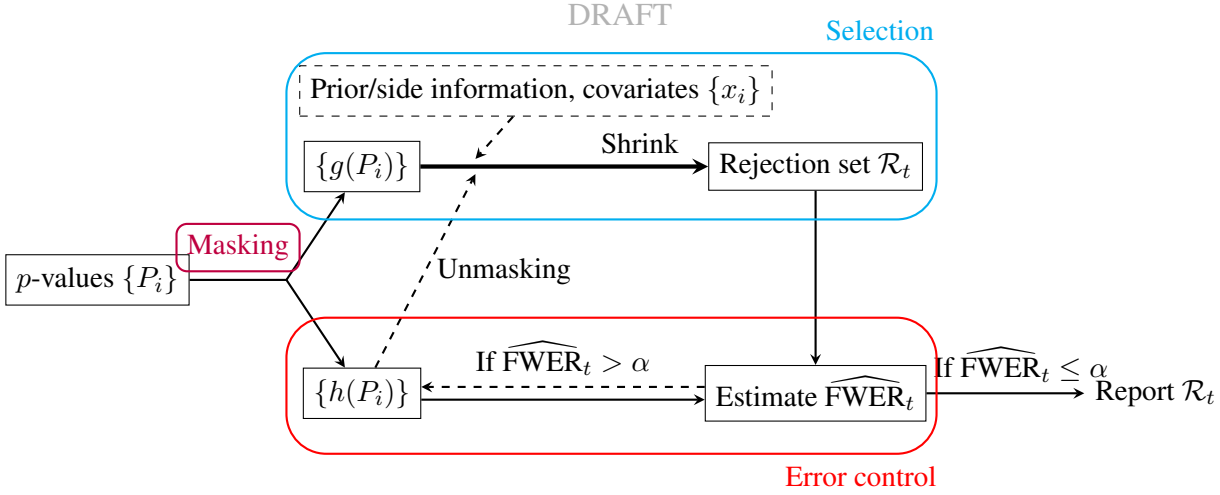


Figure 14: A schematic of the i-FWER test. All p-values are initially 'masked': all $\{g(P_i)\}$ are revealed to the analyst/algorithm, while all $\{h(P_i)\}$ remain hidden, and the initial rejection set is $\mathcal{R}_0 = [n]$. If $\widehat{\text{FWER}}_t > \alpha$, the analyst chooses a p-value to 'unmask' (observe the masked $h(P)$-value), effectively removing it from the proposed rejection set $\mathcal{R}_t$; importantly, using any available side information and/or covariates and/or working model, the analyst can shrink $\mathcal{R}_t$ in any manner. This process continues until $\widehat{\text{FWER}}_t \leq \alpha$ (or $\mathcal{R}_t = \emptyset$).

Figure 15: Functions for masking (29): missing bits h (left) and masked p-values g (right) when $p_* = 0.5$. For uniform p-values, $g(P)$ and $h(P)$ are independent.

1. **Masking**. Given a parameter $p_* \in (0, 1)$, each p-value P_i is decomposed into two parts by functions $h : [0, 1] \to \{-1, 1\}$ and $g : [0, 1] \to (0, p_*)$:

$$\begin{aligned} h(P_i; p_*) = \; & 2 \cdot \mathbb{1}\{P_i < p_*\} - 1; \\ \text{and } g(P_i; p_*) = \; & \min\left\{P_i, \frac{p_*}{1 - p_*}(1 - P_i)\right\}, \end{aligned} \tag{29}$$

   where $g(P_i)$, the *masked p-value*, is used to interactively adjust the algorithm, and $h(P_i)$, the *revealed missing bit*, is used for error control. Note that $h(P_i)$ and $g(P_i)$ are independent if H_i is null (P_i is uniformly distributed); this fact permits interaction with an analyst without any risk of violating FWER control.

2. **Selection**. Consider a set of candidate hypotheses to be rejected (rejection set), denoted as $\mathcal{R}_t$ for iteration t. We start with all the hypotheses included, $\mathcal{R}_0 = [n]$. At each iteration, the analyst excludes possible nulls from the previous $\mathcal{R}_{t-1}$, using all the available information (masked p-values, progressively unmasked $h(P_i)$ from step 3 and possible prior information). Note that our method does not automatically use prior information and masked p-values. The analyst is

free to use any black-box prediction algorithm or Bayesian working model that uses the available information, and orders the hypotheses possibly using an estimated likelihood of being non-null. This step is where a human is allowed to incorporate their subjective choices.

3. **Error control (and unmasking)**. The FWER is estimated using $h(P_i)$. If the estimation $\widehat{\text{FWER}}_t > \alpha$, the analyst goes back to step 2, provided with additional information: unmasked (reveal) $h(P_i)$ of the excluded hypotheses, which improves her understanding of the data and guides her choices in the selection step.

The rest of the paper is organized as follows. In Section 3.2, we describe the i-FWER test in detail. In Section 3.3, we implement the interactive test under a clustered non-null structure. In Section 3.4, we propose two alternative ways of masking p-values and explore their advantages.

3.2 An interactive test with FWER control

Interaction shows its power mostly when there is prior knowledge. We first introduce the *side information*, which is available before the test in the form of covariates x_i as an arbitrary vector (mix of binary, real-valued, categorical, etc.) for each hypothesis i. For example, if the hypotheses are arranged in a rectangular grid (such as when processing an image), then x_i could be the coordinate of hypothesis i on the grid. Side information can help the analyst to exclude possible nulls, for example, when the non-nulls are believed to form a cluster on the grid by some domain knowledge. Here, we state the algorithm and error control with the side information treated as fixed values, but side information can be random variables, like the bodyweight of patients when testing whether each patient reacts to a certain medication. Our test also works for random side information X_i by considering the conditional behavior of p-values given X_i.

The i-FWER test proceeds as progressively shrinking a candidate rejection set $\mathcal{R}_t$ at step t,

$$[n] = \mathcal{R}_0 \supseteq \mathcal{R}_1 \supseteq \ldots \supseteq \mathcal{R}_n = \emptyset,$$

where recall $[n]$ denotes the set of all the hypotheses. We assume without loss of generality that one hypothesis is excluded in each step. Denote the hypothesis excluded at step t as i_t^*. The choice of i_t^* can use the information available to the analyst before step t, formally defined as a filtration (sequence of nested σ-fields) [2]:

$$\mathcal{F}_{t-1} := \sigma\Big(\{x_i, g(P_i)\}_{i=1}^n, \{P_i\}_{i\notin\mathcal{R}_{t-1}}\Big), \tag{30}$$

where we unmask the p-values for the hypotheses that are excluded from the rejection set $\mathcal{R}_{t-1}$.

To control FWER, the number of false discoveries V is estimated using only the binary missing bits $h(P_i)$. The idea is to partition the candidate rejection set $\mathcal{R}_t$ into $\mathcal{R}_t^+$ and $\mathcal{R}_t^-$ by the value of $h(P_i)$:

$$\begin{aligned}\mathcal{R}_t^+ &:= \{i \in \mathcal{R}_t : h(P_i) = 1\} \equiv \{i \in \mathcal{R}_t : P_i < p_*\},\\ \mathcal{R}_t^- &:= \{i \in \mathcal{R}_t : h(P_i) = -1\} \equiv \{i \in \mathcal{R}_t : P_i \geq p_*\};\end{aligned}$$

recall that p_* is the prespecified parameter for masking (29). Instead of rejecting every hypothesis in $\mathcal{R}_t$, note that the test only rejects the ones in $\mathcal{R}_t^+$, whose p-values are smaller than p_* in $\mathcal{R}_t$. Thus, the number of false rejection V is $|\mathcal{H}_0 \cap \mathcal{R}_t^+|$ and we want to control FWER, $\mathbb{P}(V \geq 1)$. The distribution of $|\mathcal{H}_0 \cap \mathcal{R}_t^+|$ can be estimated by $|\mathcal{H}_0 \cap \mathcal{R}_t^-|$ using the fact that $h(P_i)$ is a (biased) coin flip. But $\mathcal{H}_0$ (the

[2] This filtration denotes the information used for choosing i_t^*. The filtration with respect to which the stopping time in Algorithm 5 is measurable includes the scale of R_t^-: $\mathcal{G}_{t-1} := \sigma\Big(\mathcal{F}_{t-1}, |i \in R_t : h(P_i) = -1|\Big)$.

Algorithm 5 The i-FWER test

Input: Side information and p-values $\{x_i, P_i\}_{i=1}^n$, target FWER level α, and parameter p_*;
Procedure:
Initialize $\mathcal{R}_0 = [n]$;
for $t = 1$ **to** n **do**
1. Pick any $i_t^* \in \mathcal{R}_{t-1}$, using $\{x_i, g(P_i)\}_{i=1}^n$ and progressively unmasked $\{h(P_i)\}_{i \notin \mathcal{R}_{t-1}}$;
2. Exclude i_t^* and update $\mathcal{R}_t = \mathcal{R}_{t-1} \backslash \{i_t^*\}$;

if $\widehat{\text{FWER}}_t \equiv 1 - (1 - p_*)^{|\mathcal{R}_t^-|+1} \leq \alpha$ **then**
Reject $\{H_i : i \in \mathcal{R}_t, h(P_i) = 1\}$ and exit;
end if
end for

set of true nulls) is unknown, so we use $|\mathcal{R}_t^-|$ to upper bound $|\mathcal{H}_0 \cap \mathcal{R}_t^-|$, and propose an estimator of FWER:

$$\widehat{\text{FWER}}_t = 1 - (1 - p_*)^{|\mathcal{R}_t^-|+1}. \tag{31}$$

Overall, the i-FWER test shrinks $\mathcal{R}_t$ until $\widehat{\text{FWER}}_t \leq \alpha$ and rejects only the hypotheses in $\mathcal{R}_t^+$ (Algorithm 5).

Remark 3. *The parameter p_* should be chosen in $(0, \alpha]$, because otherwise $\widehat{FWER}_t$ is always larger than α and no rejection would be made. In principle, because $|\mathcal{R}_t^-|$ only takes integer values, we should p_* such that $\frac{\log(1-\alpha)}{\log(1-p_*)}$ is an integer; otherwise, the estimated FWER at the stopping time, $\widehat{FWER}_\tau$, would be strictly smaller than α rather than equal. Our numerical experiments suggest that the power is relatively robust to the choice of p_*. A default choice can be $p_* \approx \alpha/2$ (see detailed discussion in Appendix B.7).*

Remark 4. *The above procedure can be easily extended to control k-FWER:*

$$k\text{-}FWER := \mathbb{P}(V \geq k), \tag{32}$$

by estimating k-FWER as

$$\widehat{k\text{-}FWER}_t = 1 - \sum_{i=0}^{k-1} \binom{|\mathcal{R}_t^-| + i}{i} (1 - p_*)^{|\mathcal{R}_t^-|+1} p_*^i.$$

The error control of i-FWER test uses an observation that at the stopping time, the number of false rejections is stochastically dominated by a negative binomial distribution. The complete proof is in Appendix B.2.

Theorem 8. *Suppose the null p-values are mutually independent and they are independent of the non-nulls, then the i-FWER test controls FWER at level α.*

Remark 5. *The null p-values need not be exactly uniformly distributed. For example, FWER control also holds when the null p-values have nondecreasing densities. Appendix B.1 presents the detailed technical condition for the distribution of the null p-values.*

Related work. The i-FWER test mainly combines and generalizes two sets of work: (a) we use the idea of masking from Lei and Fithian [2018]; Lei et al. [2020] and extend it to a more stringent error metric, FWER; (b) we use the method of controlling FWER from Janson and Su [2016] by converting a one-step

procedure in the context of "knockoff" statistics in regression problem to a multi-step (interactive) procedure in our context of p-values.

Lei and Fithian [2018] and Lei et al. [2020] introduce the idea of masking and propose interactive tests that control *false discovery rate* (FDR):

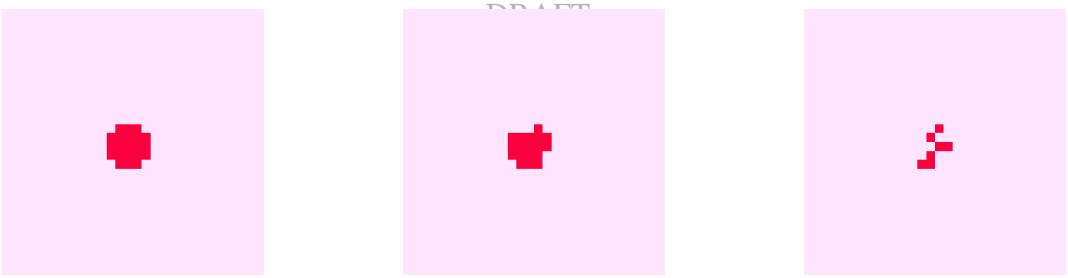
$$\text{FDR} := \mathbb{E}\left(\frac{V}{|\mathcal{R}| \vee 1}\right),$$

the expected proportion of false discoveries. It is less stringent than FWER, the probability of making *any* false discovery. Their method uses the special case of masking (29) when $p_* = 0.5$, and estimate V by $\sum_{i\in\mathcal{R}_t} \mathbb{1}\{h(P_i) = -1\}$, or equivalently $\sum_{i\in\mathcal{R}_t} \mathbb{1}\{P_i < 0.5\}$. While it provides a good estimation on the *proportion* of false discoveries, the indicator $\mathbb{1}\{P_i < 0.5\}$ has little information on the correctness of *individual* rejections. To see this, suppose there is one rejection, then FWER is the probability of this rejection being false. Even if $h(P_i) = 1$, which indicates the p-value is on the smaller side, the tightest upper bound on FWER is as high as 0.5. Thus, our method uses masking (29) with small p_*, so that $h(P_i) = 1$, or equivalently $P_i < p_*$, suggests a low chance of false rejection.

In the context of a regression problem to select significant covariates, Janson and Su [2016] proposes a one-step method with control on k-FWER; recall definition in (32). The FWER is a special case of k-FWER when $k = 1$, and as k grows larger, k-FWER is a less stringent error metric. Their method decomposes statistics called "knockoff" [Barber and Candès, 2015] into the magnitudes for ordering covariates (without interaction) and signs for estimating k-FWER, which corresponds to decomposing p-values into $g(P_i)$ and $h(P_i)$ when $p_* = 0.5$. However, the decomposition as magnitude and sign restricts the corresponding p-value decomposition with a single choice of p_* as 0.5, making the k-FWER control conservative and power low when $k = 1$; yet our method shows high power in experiments. Their error control uses the connection between k-FWER and a negative binomial distribution, based on which we propose the estimator $\widehat{\text{FWER}}_t$ for our multi-step procedure, and prove the error control even when interaction is allowed. As far as we know, this estimator viewpoint of the FWER procedure is also new in the literature.

Jelle Goeman (private communication) pointed out that the i-FWER test can be interpreted from the perspective of closed testing [Marcus et al., 1976]. Our method is also connected with the fallback procedure [Wiens and Dmitrienko, 2005], which allows for arbitrary dependence but is not interactive and combine covariate information with p-values to determine the ordering. See Appendix B.3 for details.

The i-FWER test in practice. Technically in a fully interactive procedure, a human can examine all the information in $\mathcal{F}_{t-1}$ and pick i_t^* subjectively or by any other principle, but doing so for every step could be tedious and unnecessary. Instead, the analyst can design an automated version of the i-FWER test, and still keeps the flexibility to change it at any iteration. For example, the analyst can implement an automated algorithm to first exclude 80% hypotheses (say). If $\widehat{\text{FWER}}_t$ is still larger than level α, the analyst can pause the procedure manually to look at the unmasked p-value information, update her prior knowledge, and modify the current algorithm. The next section presents an automated implementation of the i-FWER test that takes into account the structure on the non-nulls.



(a) True non-nulls (21 hypotheses).

(b) 18 rejections by the i-FWER test.

(c) 7 rejections by the Šidák correction

Figure 16: An instance of rejections by the i-FWER test and the Šidák correction [Šidák, 1967]. Clustered non-nulls are simulated from the setting in Section 2 with a fixed alternative mean $\mu = 3$.

3.3 An instantiation of an automated algorithm, and numerical experiments

One main advantage of the i-FWER test is the flexibility to include prior knowledge and human guidance. The analyst might have an intuition about what structural constraints the non-nulls have. For example, we consider two structures: (a) a grid of hypotheses where the non-nulls are in a cluster (of some size/shape, at some location; see Figure 16a), which is a reasonable prior belief when one wants to identify a tumor in a brain image; and (b) a tree of hypotheses where a child can be non-null only if its parent is non-null, as may be the case in applications involving wavelet decompositions.

3.3.1 An example of an automated algorithm under clustered non-null structure

We propose an automated algorithm of the i-FWER test that incorporates the structure of clustered non-nulls. Here, the side information x_i is the coordinates of each hypothesis i. The idea is that at each step of excluding possible nulls, we peel off the boundary of the current $\mathcal{R}_t$, such that the rejection set stays connected (see Figure 17).

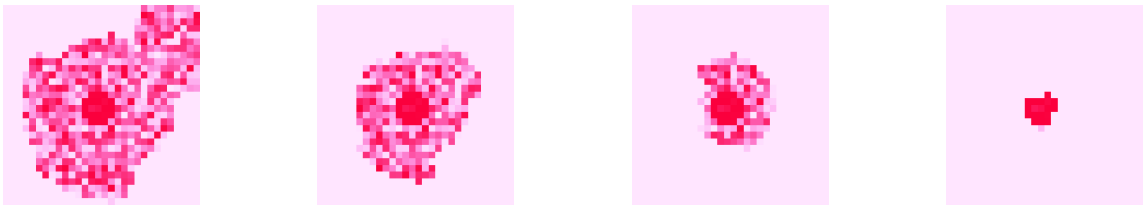


Figure 17: An illustration of $\mathcal{R}_t$ generated by the automated algorithm described in Section 3.3.1, at $t = 50, 100, 150$ and $t = 220$ when the algorithm stops. The p-values in $\mathcal{R}_t$ are plotted.

Suppose each hypothesis H_i has a score S_i to measure the likelihood of being non-null (*non-null likelihood*). A simple example is $S_i = -g(P_i)$ since larger $g(P_i)$ indicates less chance of being a non-null (more details on S_i to follow). We now describe an explicit fixed procedure to shrink $\mathcal{R}_t$. Given two parameters d and δ (eg. $d = 5, \delta = 5\%$), it replaces step 1 and 2 in Algorithm 5 as follows:

(a) Divide $\mathcal{R}_{t-1}$ from its center to d cones (like slicing a pizza); in each cone, consider a fraction δ of hypotheses farthest from the center, denoted $\mathcal{R}^1_{t-1}, \ldots, \mathcal{R}^d_{t-1}$;

(b) Compute $\bar{S}^j = \frac{1}{|\mathcal{R}^j_{t-1}|} \sum_{i \in \mathcal{R}^j_{t-1}} S_i$ for $j = 1, \ldots, d$;

(c) Update $\mathcal{R}_t = \mathcal{R}_{t-1} \backslash \mathcal{R}^k_{t-1}$, where $k = \operatorname{argmin}_j \bar{S}^j$.

The score S_i that estimates the non-null likelihood can be computed with the aid of a working statistical model. For example, consider a mixture model where each p-value P_i is drawn from a mixture of a null distribution F_0 (eg: uniform) with probability $1 - \pi_i$ and an alternative distribution F_1 (eg: beta distribution) with probability π_i, or equivalently,

$$P_i \overset{d}{=} (1 - \pi_i)F_0 + \pi_i F_1. \tag{33}$$

To account for the clustered structure of non-nulls, we may further assume a model that treats π_i as a smooth function of the covariates x_i. The hidden missing bits $\{h(P_i)\}_{i \in R_t}$ can be inferred from $g(P_i)$ and the unmasked $h(P_i)$ by the EM algorithm (see details in Appendix B.8). As R_t shrinks, progressively unmasked missing bits improve the estimation of non-null likelihood and increase the power. Importantly, the FWER is controlled regardless of the correctness of the above model or any other heuristics to shrink R_t.

The above algorithm is only one automated example and there are many possibilities of what we can do to shrink R_t.

1. A different algorithm can be developed for a different structure. For example, when hypotheses have a hierarchical structure and the non-nulls only appear on a subtree, an algorithm can gradually cut branches.
2. The score S_i for non-null likelihood is not exclusive for the above algorithm – it can be used in any heuristics such as directly ordering hypotheses by S_i.
3. Human interaction can help the automated procedure: the analyst can stop and modify the automated algorithm at any iteration. It is a common case where prior knowledge might not be accurate, or there exist several plausible structures. The analyst may try different algorithms and improve their understanding of the data as the test proceeds. In the example of clustered non-nulls, the underlying truth might have two clustered non-nulls instead of one. After several iterations of the above algorithm that is designed for a single cluster, the shape of $\mathcal{R}_t$ could look like a dumbbell, so the analyst can split $\mathcal{R}_t$ into two subsets if they wish.

Note that there is no universally most powerful test in nonparametric settings since we do not make assumptions on the distribution of non-null p-values, or how informative the covariates are. It is possible that the classical Bonferroni-Holm procedure [Holm, 1979] might have high power if applied with appropriate weights. Likewise, the power of our own test might be improved by changing the working model or choosing some other heuristic to shrink $\mathcal{R}_t$. The main advantage of our method is that it can accommodate structural and covariate information and revise the modeling on the fly (as p-values are unmasked) while other methods commit to one type of structure without looking at the data.

Next, we demonstrate via experiments that the i-FWER test can improve power over the Šidák correction, a baseline method that does not take side information into account[3]. We chose a clustered non-null structure for visualization and intuition, though our test can utilize any covariates, structural constraints, domain knowledge, etc.

3.3.2 Numerical experiments for clustered non-nulls

For most simulations in this paper, we use the setting below,

Setting 2. *Consider 900 hypotheses arranged in a* 30×30 *grid with a disc of 21 non-nulls. Each hypothesis tests the mean value of a univariate Gaussian, where the null hypothesis is nonpositive mean. The true nulls are generated from* $N(0, 1)$ *and non-nulls from* $N(\mu, 1)$*, where we varied* μ *as*

[3]In all experiments, the Hommel method has similar power to the Šidák correction, and was hence omitted.

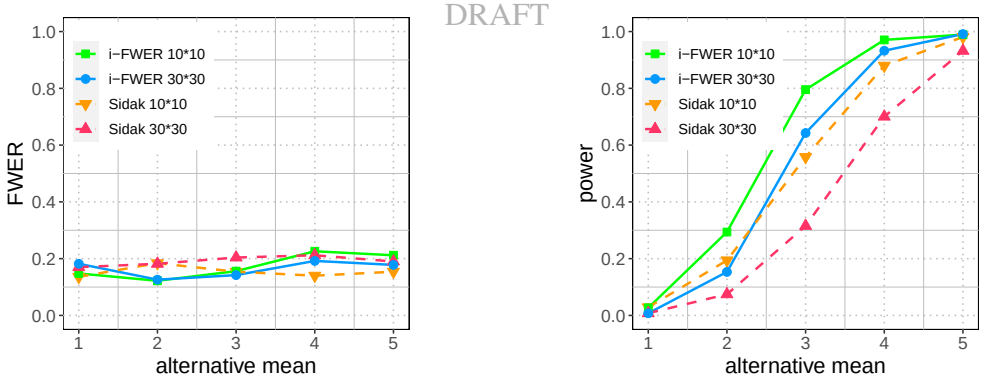


Figure 18: The i-FWER test versus Šidák for clustered non-nulls. The experiments are described in Section 2 where we tried two sizes of hypotheses grid: 10×10 and 30×30 (the latter is a harder problem since the number of nulls increases while the number of non-nulls remains fixed). Both methods show valid FWER control (left). The i-FWER test has higher power under both grid sizes (right).

$(1, 2, 3, 4, 5)$. *For all experiments in the paper, the FWER control is set at level* $\alpha = 0.2$*, and the power is averaged over 500 repetitions*[4].

The i-FWER test has higher power than the Šidák correction, which does not use the non-null structure (see Figure 18). It is hard for most existing methods to incorporate the knowledge that non-nulls are clustered without knowing the position or the size of this cluster. By contrast, such information can be learned in the i-FWER test by looking at the masked p-values and the progressively unmasked missing bits. This advantage of the i-FWER test is more evident as the number of nulls increases (by increasing the grid size from 10×10 to 30×30 with the number of non-nulls fixed). Note that the power of both methods decreases, but the i-FWER test seems less sensitive. This robustness to nulls is expected as the i-FWER test excludes most nulls before rejection, whereas the Šidák correction treats all hypotheses equally.

3.3.3 An example of an automated algorithm under a hierarchical structure of hypotheses

When the hypotheses form a tree, the side information x_i encodes the parent-child relationship (the set of indices of the children nodes for each hypothesis i). Suppose we have prior knowledge that a node cannot be non-null if its parent is null, meaning that the non-nulls form a subtree with the same root. We now develop an automated algorithm that prunes possible nulls among the leaf nodes of current $\mathcal{R}_t$, such that the rejection set has such a subtree shape. Like the algorithm for clustered non-nulls, we use a score S_i to choose which leaf nodes to exclude. For example, the score S_i can be the estimated non-null likelihood learned from model (33), where we account for the hierarchical structure by further assuming a partial order constraint on π_i that $\pi_i \geq \pi_j$ if $j \in x_i$ (i.e., i is the parent of j).

We simulate a tree of five levels (the root has twenty children and three children for each parent node after that) with 801 nodes in total and 7 of them being non-nulls. The non-nulls gather in one of the twenty subtrees of the root. Individual p-values are generated by the hypotheses of testing zero-mean Gaussian, same as for the clustered structure, where we varied the non-null mean values μ as $(1, 2, 3, 4, 5)$.

[4]Code can be found in https://github.com/duanby/i-FWER. It was tested on macOS using R (version 3.6.0) and the following packages: magrittr, splines, robustbase, ggplot2. The standard error of FWER and averaged power are less than 0.02, thus ignored from the plots in this paper.

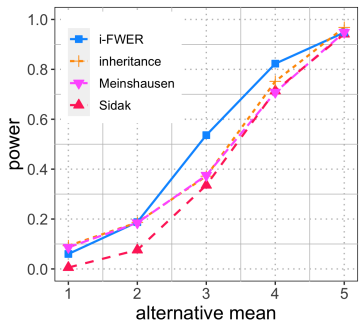


Figure 19: Power of the i-FWER test under a tree structure when varying the alternative mean value. It has higher power than inheritance procedure, Meinshausen's method, and the Sidak correction.

In addition to the Šidák correction, we compare the i-FWER test with two other methods for tree-structured hypotheses: Meinshausen's method [Meinshausen, 2008] and the inheritance procedure [Goeman and Finos, 2012], which work under arbitrary dependence. Their idea is to pass the error budget from a parent node to its children in a prefixed manner, whereas our algorithm picks out the subtree with non-nulls based on the observed data. In our experiments, the i-FWER test has the highest power (see Figure 19).

The above results demonstrate the power of the i-FWER test in one particular form where the masking is defined as (29). However, any two functions that decompose the null p-values into two independent parts can, in fact, be used for masking and fit into the framework of the i-FWER test (see the proofs of error control when using the following new masking functions in Appendix B.6). In the next section, we explore several choices of masking.

3.4 New masking functions

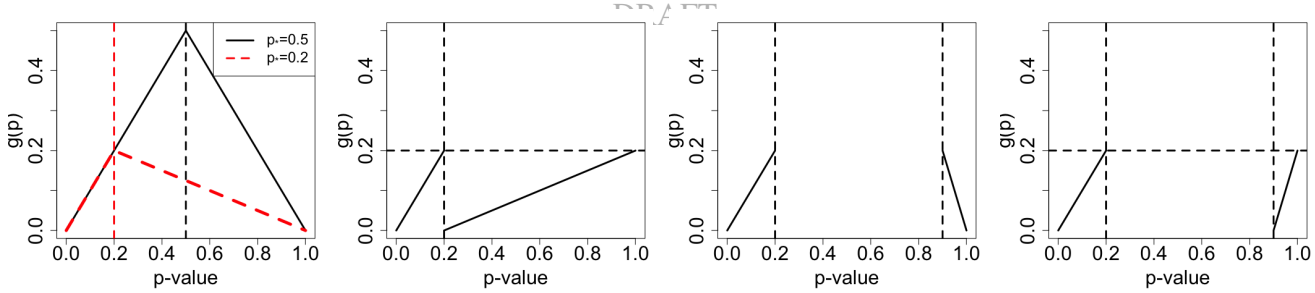
Recall that masking is the key idea that permits interaction and controls error at the same time, by decomposing the p-values into two parts: masked p-value $g(P)$ and missing bits $h(P)$. Such splitting distributes the p-value information for two different purposes, interaction and error control, leading to a tradeoff. More information in $g(P)$ provides better guidance on how to shrink $\mathcal{R}_t$ and improves the power, while more information in $h(P)$ enhances the accuracy of estimating FWER and makes the test less conservative. This section explores several ways of masking and their influence on the power of the i-FWER test. To distinguish different masking functions, we refer to masking (29) introduced at the very beginning as the "tent" function based on the shape of map g (see Figure 20a).

3.4.1 The "railway" function

We start with an adjustment to the tent function that flips the map g when $p > p_*$, which we call the "railway" function (see Figure 20b). It does not change the information distribution between $g(P)$ and $h(P)$, and yet improves the power when nulls are conservative, as demonstrated later. Conservative nulls are often discussed under a general form of hypotheses testing for a parameter θ:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 are two disjoint sets. Conservative nulls are those whose true parameter θ lies in the interior of Θ_0. For example, when testing whether a Gaussian $N(\mu, 1)$ has nonnegative mean where $\Theta_0 = \{\mu \leq 0\}$, the nulls are conservative when $\mu < 0$. The resulting p-values are biased toward



(a) Tent functions when p_* varies as $(0.5, 0.2)$. We need $p_* \leq \alpha$ for FWER control.

(b) The railway function when $p_* = 0.2$.

(c) The gap function when $(p_l, p_u) = (0.2, 0.9)$.

(d) The gap-railway function when $(p_l, p_u) = (0.2, 0.9)$.

Figure 20: Different masking functions leaves different amount of information to $g(P)$ (and the complement part to $h(P)$).

larger values, which compared to the uniform p-values from nonconservative nulls should be easier to distinguish from that of non-nulls. However, most classical methods do not take advantage of it, but the i-FWER test can, when using the railway function for masking:

$$h(P_i) = 2 \cdot \mathbb{1}\{P_i < p_*\} - 1;$$
$$\text{and } g(P_i) = \begin{cases} P_i, & 0 \leq P_i < p_*, \\ \frac{p_*}{1-p_*}(P_i - p_*), & p_* \leq P_i \leq 1. \end{cases} \tag{34}$$

The above masked p-value, compared with the tent masking (29), can better distinguish the non-nulls from the conservative nulls. To see this, consider a p-value of 0.99. When $p_* = 0.2$, the masked p-value generated by the originally proposed tent function would be 0.0025, thus causing potential confusion with a non-null, whose masked p-value is also small. But the masked p-value from the railway function would be 0.1975, which is close to 0.2, the upper bound of $g(P_i)$. Thus, it can easily be excluded by our algorithm.

We follow the setting in Section 2 for simulation , except that the alternative mean is fixed as $\mu = 3$, and the nulls are simulated from $N(\mu_0, 1)$, where the mean value μ_0 is negative so that the resulting null p-values are conservative. We tried μ_0 as $(0, -1, -2, -3, -4)$, with a smaller value indicating higher conservativeness, in the sense that the p-values are more likely to be biased to a larger value. When the null is not conservative ($\mu_0 = 0$), the i-FWER test with the railway function and tent function have similar power. As the conservativeness of nulls increases, while the power of the i-FWER test with the tent function decreases and the Šidák correction stays the same, the power of the i-FWER test with the railway function increases (see Figure 21).

3.4.2 The "gap" function

Another form of masking we consider maps only the p-values that are close to 0 or 1, which is referred to as the "gap" function (see Figure 20c) . The resulting i-FWER test directly unmasks all the p-values in the middle, and as a price, never rejects the corresponding hypotheses. Given two parameters p_l and p_u,

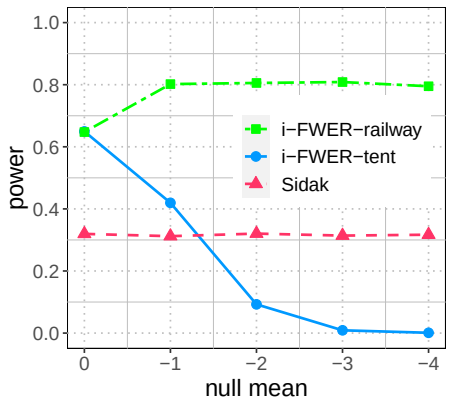


Figure 21: Power of the i-FWER test with the tent function and the railway function, where the nulls become more conservative as the null mean decreases in $(0, -1, -2, -3, -4)$. The i-FWER test benefits from conservative null when using the railway function.

the gap function is defined as

$$h(P_i) = \begin{cases} 1, & 0 \le P_i < p_l, \\ -1, & p_u < P_i \le 1; \end{cases}$$
$$\text{and } g(P_i) = \begin{cases} P_i, & 0 \le P_i < p_l, \\ \frac{p_l}{1-p_u}(1-P_i), & p_u < P_i \le 1. \end{cases} \tag{35}$$

All the p-values in $[p_l, p_u]$ are available to the analyst from the beginning. Specifically, let $\mathcal{M} = \{i : p_l < P_i < p_u\}$ be the set of skipped p-values in the masking step, then the available information at step t for shrinking $\mathcal{R}_{t-1}$ is

$$\mathcal{F}_{t-1} := \sigma\Big(\{x_i, g(P_i)\}_{i=1}^n, \{P_i\}_{\{i \notin \mathcal{R}_{t-1}\}}, \{P_i\}_{\{i \in \mathcal{M}\}}\Big).$$

The i-FWER test with the gap masking changes slightly. We again consider two subsets of $\mathcal{R}_t$:

$$\mathcal{R}_t^+ := \{i \in \mathcal{R}_t : h(P_i) = 1\} \equiv \{i \in \mathcal{R}_t : P_i < p_l\},$$
$$\mathcal{R}_t^- := \{i \in \mathcal{R}_t : h(P_i) = -1\} \equiv \{i \in \mathcal{R}_t : P_i > p_u\},$$

and reject only the hypotheses in $\mathcal{R}_t^+$. The procedure of shrinking $\mathcal{R}_t$ stops when $\widehat{\text{FWER}}_t \le \alpha$, where the estimation changes to

$$\widehat{\text{FWER}}_t = 1 - \left(1 - \frac{p_l}{p_l + 1 - p_u}\right)^{|\mathcal{R}_t^-|+1}. \tag{36}$$

To avoid the case that $\widehat{\text{FWER}}_t$ is always larger than α and the algorithm cannot make any rejection, the parameters p_l and p_u need to satisfy $\frac{1-\alpha}{\alpha} p_l + p_u < 1$. The above procedure boils down to the original i-FWER test with the tent function when $p_l = p_u = p_*$.

The "gap" function reveals more information to select out possible nulls and help the analyst shrink $\mathcal{R}_t$, leading to power improvement in numerical experiments. We present the power results of the i-FWER test using different masking functions after introducing a variant of the gap function.

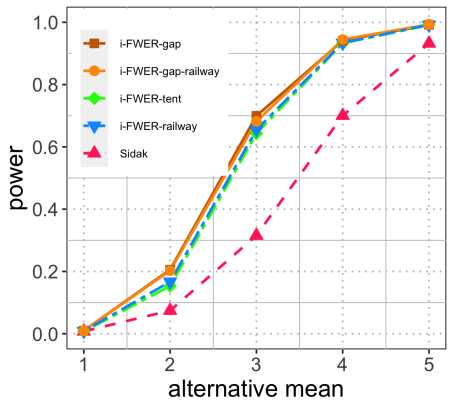


Figure 22: Power of the i-FWER test with the tent function ($p_* = 0.1$) and the gap function ($p_l = 0.1, p_u = 0.5$). The gap function leads to slight improvement in power. Simulation follows the setting in Section 2.

3.4.3 The "gap-railway" function

Combining the idea of the gap and railway functions, we develop the "gap-railway" function such that the middle p-values are directly unmasked and the map g for large p-values is an increasing function (see Figure 20d). Given parameters p_l and p_u, the gap-railway function is defined as

$$h(P_i) = \begin{cases} 1, & 0 \le P_i < p_l, \\ -1, & p_u < P_i \le 1; \end{cases}$$
$$\text{and } g(P_i) = \begin{cases} P_i, & 0 \le P_i < p_l, \\ \frac{p_l}{1-p_u}(P_i - p_u), & p_u < P_i \le 1. \end{cases} \tag{37}$$

Comparing with the tent function with $p_* = p_l$, the i-FWER test using the gap function additionally uses the entire p-values in $[p_l, p_u]$ for interaction, which leads to an increased power (see Figure 22). The same pattern is maintained when we flip the mappings for large p-values, shown in the comparison of the railway function and the gap-railway function[5]. This improvement also motivates why the i-FWER test progressively unmasks $h(P_i)$, in other words, to reveal as much information to the analyst as allowed at the current step. Unmasking the p-values even for the hypotheses outside of the rejection set can improve the power, because they help the joint modeling of all the p-values, especially when there is some non-null structure.

To summarize, we have presented four types of masking functions: tent, railway, gap, gap-railway (see Figure 20). Compared to the tent (gap) function, the railway (gap-railway) functions are more robust to conservative nulls. Compared with the tent (railway) function, the gap (gap-railway) function reveals more information to guide the shrinkage of $\mathcal{R}_t$. Note however that the railway or gap function is not *always* better than the tent function. We may favor the tent function over the railway function when there are less p-values close to one, and we may favor the tent function over the gap function when there is considerable prior knowledge to guide the shrinkage of $\mathcal{R}_t$.

The above discussion has explored specific non-null structures and masking functions. A large variety of masking functions and their advantages are yet to be discovered.

[5]The tests with the tent function and the railway function have similar power; and same for the gap function and the gap-railway function. As the null p-values follow an exact uniform distribution, so flipping the map g for large p-values does not change the power.

3.5 A prototypical application to genetic data

Below, we further demonstrate the power of the i-FWER test using a real 'airway dataset', which is analyzed by Independent Hypothesis Weighting (IHW) [Ignatiadis et al., 2016] and AdaPT [Lei and Fithian, 2018]; these are (respectively) adaptive and interactive algorithms with FDR control for independent hypotheses. We compare the number of rejections made by a variant of the IHW with FWER control and the i-FWER test using the tent function with the masking parameter p_* chosen as $\alpha/20, \alpha/10, \alpha/2$, when the targeted FWER level α varies in $(0.1, 0.2, 0.3)$.

The airway data is an RNA-Seq dataset targeting the identification of differentially expressed genes in airway smooth muscle cell lines in response to dexamethasone, which contains 33469 genes (hypotheses) and a univariate covariate (the logarithm of normalized sample size) for each gene[6]. The i-FWER test makes more rejections than the IHW for all considered FWER levels and choices of p_* (see Table 1).

Table 1: Number of rejections by IHW and i-FWER test under different FWER levels.

level α	IHW	i-FWER		
		$p_* = \alpha/2$	$p_* = \alpha/10$	$p_* = \alpha/20$
0.1	1552	1613	1681	1646
0.2	1645	1740	1849	1789
0.3	1708	1844	1925	1894

In hindsight, a small value for the masking parameter was more powerful in this dataset because over 1600 p-values are extremely small ($< 10^{-5}$), and these are highly likely to be the non-nulls. Thus, even when the masked p-values for all hypotheses are in a small range, such as $(0, 0.01)$ when $\alpha = 0.1$ and $p_* = \alpha/10$, the p-values from the non-nulls still stand out because they gather below 10^{-5}. At the same time, the smaller the p_*, the more accurate (less conservative) is our estimate of FWER in (31); the algorithm can stop shrinking $\mathcal{R}_t$ earlier since more hypotheses with negative $h(P)$ are allowed to be included in the final $\mathcal{R}_t$. In practice, the choice of masking parameter can be guided by the prior belief of the strength of non-null signals: if the non-nulls have strong signal and hence extremely small p-values (such as the mean value $\mu \geq 5$ when testing if a univariate Gaussian has zero mean), a small masking parameter is preferred; otherwise, we recommend $\alpha/2$ to leave more information for interactively shrinking the rejection set $\mathcal{R}_t$.

3.6 Summary

We proposed a multiple testing method with a valid FWER control while granting the analyst freedom of interacting with the revealed data. The masking function must be fixed in advance, but during the procedure of excluding possible nulls, the analyst can employ any model, heuristic, intuition, or domain knowledge, tailoring the algorithm to various applications.

The first two chapters focus on multiple testing, where we are given p-values and ignore the process of how individual p-values are generated. The next chapter switches the direction, and studies the problem of multi-sample comparison. In such a problem setting, we are given n subjects, each associated with an outcome of interest, a vector of covariates such as gender or age, and an indicator of which group this subject belongs to. In the next chapter, we describe how the masking idea can applied outside of the p-values and multiple testing.

[6]Data is collected by Himes et al. [2014] and available in R package `airway`. We follow Ignatiadis et al. [2016] and Lei and Fithian [2018] to analyze the data using `DEseq2` package [Love et al., 2014].

4 Which Wilcoxon should we use? An interactive rank test and other alternatives

4.1 Introduction

The problem of comparing two samples in a randomized experiment without parametric assumptions is frequently encountered in biology, medical research, and social sciences (see, for example, Calel and Dechezlepretre [2016]; Matsumoto and Hikosaka [2009]; Olive et al. [2009]). A classical nonparametric method is the Wilcoxon test (both rank-sum and signed-rank). However, the original Wilcoxon test does not adjust for covariates, but there have been several proposed extensions that do. For example, suppose we want to evaluate a medication by conducting a randomized trial and comparing the blood pressure (outcome) of subjects who take the medication (treatment) with that of subjects who do not (control). The blood pressure could be affected by the subject's gender, age, etc.—accounting for these would help increase power, especially when the medication only affects a subpopulation. In this paper, we discuss two classes of tests that take covariates into account. First, we propose a multi-step "interactive" test that allows an analyst to look at (partial) data and employ flexible working models to improve power. Second, we analyze several old and new (non-interactive) covariate-adjusted extensions of the Wilcoxon test and numerically examine how their powers are affected by the effects being one- or two-sided, dense or sparse, and the skewness of control outcomes, thus providing several practical insights along the way.

4.1.1 Problem setup

Consider a sample with n subjects. Let the outcome of subject i be Y_i, the covariates be X_i, and the treatment assignments be indicators A_i for $i \in [n] \equiv \{1, 2, \ldots, n\}$. The null hypothesis of interest is that there is no difference between treatment and control outcomes conditional on the covariates:

$$H_0 : (Y_i \mid A_i = 1, X_i) \stackrel{d}{=} (Y_i \mid A_i = 0, X_i) \text{ for all } i \in [n]. \tag{38}$$

Rejecting the above null means that there exist some subjects who respond differently when treated or not. We do not further identify which subject respond differently. Testing the above global null may appear in an exploratory analysis to see whether the treatment has any effect on any person, or as a building block within a closed testing procedure. For our interactive algorithm that we propose later to succeed in rejecting the global null, it must indeed *implicitly* learn which part of the covariate space exhibits this difference between treatment and placebo, and if the global null is rejected, one may use this information to design followup studies or analyses focused on other goals.

This paper deals with randomized experiments, and in particular we assume that

(i) the treatment assignments are independent and randomized:

$$\mathbb{P}(A_i = 1 \mid X_i) = 1/2 \text{ for all } i \in [n];$$

(ii) the outcome of one subject Y_{i_1} is independent of the assignment of another A_{i_2} for any $i_1 \neq i_2 \in [n]$.

To enable us to effectively adjust for covariates, we use the following "working model":

$$Y_i = \Delta(X_i)A_i + f(X_i) + U_i, \tag{39}$$

where $\Delta(X_i)$ is the treatment effect, $f(X_i)$ as the control outcome, and U_i is zero mean 'noise' (unexplained variance). When working with such a model, we effectively want to detect if $\Delta(X_i)$ is nonzero.

Importantly, model (39) only exists on the analyst's computer, and it need not be correctly specified or accurately reflect reality in order for the tests in this paper to be valid (but the more ill-specified or inaccurate the model is, the more test power may be hurt).

Notation. In the rest of the paper, capital letters are used to denote random variables. We use $\widehat{Z}^1(Z^2)$ to denote a prediction of Z^1 using Z^2 as input.

4.1.2 Rosenbaum's covariance-adjusted Wilcoxon rank-sum test

Recall that the original Wilcoxon rank-sum test (also referred to as the Mann–Whitney U-test) calculates

$$W^{\text{ori}} = \sum_{i=1}^{n} (2A_i - 1) \operatorname{rank}(Y_i),$$

where $\operatorname{rank}(Z_i)$ is the rank of Z_i amongst $\{Z_i\}_{i=1}^n$. When the treatment effect is large, the subjects receiving treatment $(A_i = 1)$ tend to have larger outcomes, and hence W^{ori} would be large. Note that there is another version of the Wilcoxon test called the signed-rank test[7], which differs slightly from the above one but usually has similar power; we examine this in detail in Section 4.3. The above Wilcoxon test ranks the outcomes, which may not be reliable evidence of the treatment effect, especially when the potential control outcome of different subjects is heterogeneous (varies with their covariates).

To increase power, Rosenbaum [2002] proposed the covariance-adjusted Wilcoxon test that considers the residuals of regressing the outcome Y_i on covariates X_i (without assignment A_i). Specifically, denote the residual for subject i as R_i:

$$R_i \equiv R_i(Y_i, X_i) := Y_i - \widehat{Y}(X_i), \tag{40}$$

where $\widehat{Y}(X_i)$ the prediction of Y_i using X_i via any modeling and R_i can be viewed as an approximation of the treatment effect after accounting for heterogeneous control outcome. The covariance-adjusted Wilcoxon test replaces the outcomes with the residuals:

$$W^{\text{CovAdj}} = \sum_{i=1}^{n} (2A_i - 1) \operatorname{rank}(R_i), \tag{41}$$

abbreviated as CovAdj Wilcoxon test in the rest of the paper. Note that the CovAdj Wilcoxon test improves power when the control outcome changes with covariates; however, it can have low power when the treatment effect is heterogeneous, as we show later in experiments.

A major merit of CovAdj Wilcoxon test is that its null distribution can be derived for any choice of the prediction model $\widehat{Y}$ without any parametric assumption on the outcomes, because under the null,

$$\mathbb{P}(A_i = 1 \mid Y_i, X_i) = 1/2 \quad \text{for all } i \in [n]. \tag{42}$$

In other words, the assignment A_i is independent of the outcome Y_i and the covariates X_i. In this paper, we build on the above observation and propose new tests that improve on the CovAdj Wilcoxon test by taking the heterogeneous treatment effect into consideration.

[7] Although the statistic W^{ori} for the rank-sum test appears to include a sign-like term $(2A_i - 1)$, this term is not the sign of Y_i, for which we calculate the rank, and hence the name of the rank-sum test to distinguish with the signed-rank test.

4.1.3 An interactive test

The tests we discuss can be broadly classified into two categories: (a) in contrast to one-step tests such as CovAdj Wilcoxon test, we propose a multi-step test involving human interaction to adjust its working model on the fly; and (b) we examine non-interactive variations of Rosenbaum's CovAdj Wilcoxon test that have complementary benefits. We focus on the first category since interactive testing is a recent idea that emerged in response to the growing practical needs of allowing human interaction in the process of data analysis. In practice, analysts tend to try several methods on the same dataset until the results are satisfying, but this violates the validity of standard statistical methods and causes reproducibility issues. The appealing advantage of an interactive test is (a) flexibility for the analyst to use combine (partial) data and prior knowledge in the design of the testing algorithm, and (b) the multi-step protocol during which the analyst can monitor the current algorithm's performance and is allowed to make adjustments to their working model at any step. Our proposed testing protocol always maintains valid type I error control.

The core idea that enables human interaction is to separate the information used for interactive algorithm design and that for testing, via "masking and unmasking" (Figure 23). Masking means we hide the information of treatment assignments $\{A_i\}_{i=1}^n$ from the analyst. The test considers the cumulative sums

$$S_t = \sum_{j=1}^{t} (2A_{\pi_j} - 1) \cdot w_j, \tag{43}$$

where $\{\pi_j\}_{j=1}^n$ denotes an ordering interactively decided by the analyst and w_j denotes a weight, both of which can be based on all the revealed information $\{Y_i, X_i\}_{i=1}^n$ and the true treatment assignments of all previous subjects in the ordering $A_{\pi_1}, \ldots, A_{\pi_{j-1}}$ (initially empty). The decision rule will involve rejecting the null when S_t is sufficiently large, meaning that it crosses some boundary $u_\alpha(t)$, which we specify later.

The above test retains validity amidst significant flexibility. For example, the analyst could employ any probabilistic working model or predictive machine-learning algorithm to guess the treatment assignments $\widehat{A}_i \in \{0, 1\}$, perhaps along with an associated level of confidence such as a posterior probability or a score $\nu_i \in [0, 1]$, for each subject i that have not yet been included in the ordering. Then, at step t of the algorithm, the next subject in the ordering π_t could be the one where the analyst is most confident, and w_t could equal $2\widehat{A}_t - 1 \in \{-1, 1\}$ or $2\nu_t - 1 \in [-1, 1]$. Regardless of the specific choices of the ordering and the weights, the fact that w_j is independent of A_{π_j} under the null (but not under the alternative), will allow us to provide a type-I error guarantee of the form $\mathbb{P}(\exists t \leq n : S_t > u_\alpha(t)) \leq \alpha$.

Formally, such bounds are not loose or overly conservative: a martingale property of S_t allows us to circumvent a naive union bound and use sophisticated maximal inequalities instead. To elaborate, under the null where A_i is independent of Y_i, X_i as described in (42), the increments behave like the sum of mutually independent (weighted) coin flips, regardless of how the order is determined and how we choose w_j. Such a process is a martingale with bounded increments, whose deviation up to time t can be controlled using $\sum_{j=1}^t w_j^2$. If the cumulative sum S_t deviates from this null behavior at any step t, we reject the null.

Intuitively, our algorithm tests whether there exist any non-nulls by examining whether we can succeed at guessing the treatment assignments better than random chance, and this is reflected by our ability to form a "smart, nonrandom" ordering that causes S_t to grow faster than a random walk, and cross the mentioned boundary as soon as possible. When all subjects are nulls and we have the independence property (42), we cannot distinguish subjects who are treated or not based on the outcomes and covariates, so each increment is $\pm w_j$ with equal probability, and no algorithm can result in S_t losing

this martingale property (and thus having controlled growth). In contrast, if the null is false, we hope that our algorithm will be able to correctly guess the treatment assignments, especially for subjects we are most confident about (ordered upfront), so that the cumulative sums $(S_t)_{t=1}^n$ are larger than the null case.

Interaction enters in the process of unmasking. Intuitively, to construct large S_t and reject the null, the analyst should guess whether a subject receives treatment while the assignments $\{A_i\}_{i=1}^n$ are hidden. She can guess the treatment assignments using the revealed data information $\{Y_i, X_i\}_{i=1}^n$ and $\{A_{\pi_j}\}_{j=1}^{t-1}$ (for the t-th iteration), and any prior knowledge, and she is free to use any algorithms or models. Even if the model chosen initially is inaccurate because of masking, the interactive test progressively reveals the assignments (of the first $t-1$ subjects at step t) to the analyst, so that she can improve her understanding of the data and update the model or heuristic for estimating the treatment assignments at any step.

We call our proposed procedure the i-Wilcoxon test, because (a) it exploits the same property as the classical CovAdj Wilcoxon test to ensure a valid error control: the independence between treatment assignment and other data information under the global null as mentioned in (42), and (b) the test statistic S_t in (43) shares a similar form as (41) for CovAdj Wilcoxon test. Our contribution is to demonstrate a new class of interactive multi-step algorithms that, by masking some of the data and progressively revealing it to the scientist, can combine the strengths of (automated) statistical modeling and (human-guided) scientific knowledge, in order to reject the global null while not suffering from any p-hacking or data-dredging concerns despite a great deal of flexibility provided to the scientist.

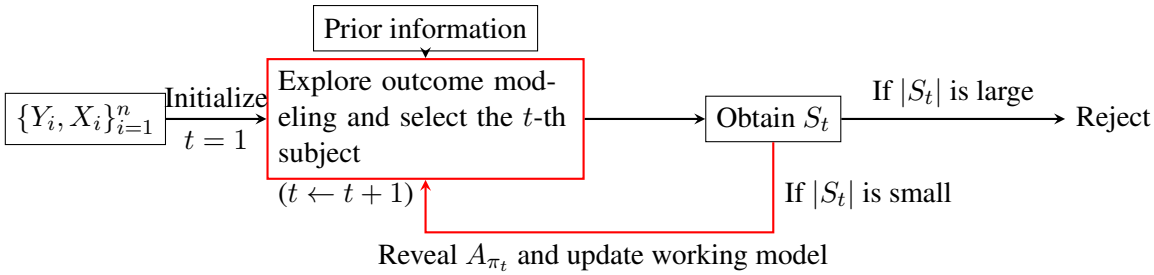


Figure 23: Schematics of the i-Wilcoxon test. At each step, a human analyst can freely explore and update models to guide the selection of the t-th subject (as the red box shows).

4.1.4 Related work

Interactive tests. The idea of interactive testing was recently proposed by Lei and Fithian [2018] and Lei et al. [2020], in the context of multiple testing problem to control FDR (the false discovery rate), followed by several works for other error metrics in multiple testing. Our interactive test for two-sample comparison relates most with the work of controlling the global type-I error [Duan et al., 2019], where the individual null hypothesis is zero effect for each subject, and the global null corresponds to the null of no treatment effect as null hypothesis (38). The main difference is that previous development of the interactive tests focused on generic multiple testing problems, which operates on the p-values, ignoring the process of generating p-values from data. Here, interactive testing is directly applied to the observed data, bringing another perspective to the potential of interactive tests.

Uniform martingale concentration inequalities. Type-I error control of the interactive test is based on the observation that, under the null, S_t is the cumulative sum of independent, fair coin flips; thus, the sequence of $S_1, S_2, \ldots$ forms a martingale. The rejection rule stems from utilizing *time-uniform*

boundary-cross inequalities for martingales. For a martingale M_t, the boundary is denoted as $u_\alpha(t)$ which satisfies

$$\mathbb{P}(\exists t \in \mathbb{N} : M_t > u_\alpha(t)) \leq \alpha, \tag{44}$$

for a constant $\alpha \in (0, 1)$. The martingale of fair coin flips is well studied in sequential analysis, especially through their natural connections to Brownian motion [Siegmund, 1986].

In this paper, we use a recent line-crossing inequality [Howard et al., 2020a]:

$$u_\alpha(t) = \sqrt{\frac{\log(1/\alpha)}{2m}} V_t + \sqrt{\frac{m \log(1/\alpha)}{2}}, \tag{45}$$

where $V_t = \sum_{j=1}^{t} w_j^2$ (for a simple example where $w_j \in \{-1, +1\}$, V_t equals t), and $m \in \mathbb{R}_+$ is a tuning parameter that determines the time at which the bound is tightest: a larger m results in a lower slope but a larger offset, making the bound loose early on. We suggest a default value of $m = n/4$, but it should be chosen based on the time by which we expect to have encountered most non-nulls (if any). One can also use curved boundaries [Howard et al., 2020b] that scale smaller than $O(V_t)$:

$$u_\alpha(t) = 1.7\sqrt{V_t \left(\log\log(2V_t) + 0.72 \log \frac{5.2}{\alpha} \right)} \text{ or} \tag{46}$$

$$u_\alpha(t) = \sqrt{(V_t + 0.13) \log\left(\tfrac{V_t + 0.13}{0.52\alpha^2}\right)}, \tag{47}$$

but these curved boundaries do not uniformly dominate the linear ones, and hence their powers are in general not comparable. We present the results using boundary (45) since it has a simple form and consistently resulted in reasonably good power.

Related problems. There are many works on related problems, but most of these tend to focus on a less strict null hypothesis H_0' than the global null H_0 in (38), which of course has pros and cons. These related methods would continue valid for the global null hypothesis of our interest, but they could have lower power especially when H_0' is true and H_0 is not true. Our strong global null is still sometimes of scientific interest, for example when certain quantiles of the distribution may be different under two different treatments (without the means differing), or one may be interested in the heavy-tailed case when the means may not even exist. We elaborate on the related work as follows.

While several works study treatment with multiple levels, we describe them in the case with two levels (treated or not) in our discussion. Akritas et al. [2000] assess the treatment effect by comparing the outcome CDF of treated and control group, denoted as $F_x^T(y)$ and $F_x^C(y)$ where x is the given covariate value. Let $G(x)$ be a prespecified distribution for the covariate or its empirical distribution. The null hypothesis concerns marginal CDF after averaging over the covariate:

$$H_0' : \int F_x^T(y) dG(x) = \int F_x^C(y) dG(x), \tag{48}$$

which is implied by the global null H_0 in our discussion. Fan and Zhang [2017] also study the above null hypothesis (48), and propose an alternative test statistic to incorporate covariates. Wang and Akritas [2006] consider several extensions in the type of null hypothesis and suggest the possibility of testing whether the conditional outcome CDF given the covariates is identical:

$$H_0' : F_x^T(y) = F_x^C(y) \text{ for every } x \text{ and } y, \tag{49}$$

which is equivalent to the global null H_0 in our discussion, but no explicit test is provided for this null hypothesis. Similar null hypotheses are discussed in the work of Edgar Brunner (such as Akritas et al. [1997]; Bathke and Brunner [2003]), which focus on factorial design and develop tests for the effect of one factor conditional on the level of the other factors. Thus, their methods can be used to test our global null H_0 when the covariate takes a finite number of values. Hettmansperger and McKean [2010] focus on testing the global null H_0 when the treatment effect is a linear function of the covariates, and discusses inference such as confidence intervals of the involved parameters. Along a different line of work, Thas et al. [2012] considers outcome Y and covariates Z (which include the treatment assignment A and other covariates X in our context) and let two instances (Y, Z) and (Y^*, Z^*) be independently distributed. The outcomes Y and Y^* are compared by estimating the probabilistic index $\mathbb{P}(Y > Y^* \mid Z, Z^*) + \frac{1}{2}\mathbb{P}(Y = Y^* \mid Z, Z^*)$. Their results imply a test for the null hypothesis of the probabilistic index being $1/2$, which can be used in our context:

$$\begin{aligned} H_0' :&\mathbb{P}(Y > Y^* \mid A = 1, A^* = 0, X = X^* = x) \\ &+ \tfrac{1}{2}\mathbb{P}(Y = Y^* \mid A = 1, A^* = 0, X = X^* = x) = 1/2 \text{ for all } x, \end{aligned}$$

which is true when our global null H_0 is true; hence, their method is valid for our problem of interest.

Aside from different target null hypotheses, several features distinguish our proposed algorithms from most existing work: (a) previous methods often commit to a single fixed procedure, while the i-Wilcoxon test we propose can employ arbitrary working models, and the working model can be changed by a human analyst at any iteration to improve power; (b) most other methods mentioned above guarantee type-I error asymptotically, whereas our interactive methods have exact type-I error control (without any parametric or model assumptions on the outcomes); (c) we demonstrate through numerical experiments that the advantage of our proposed methods is more evident when a treatment effect exists only for a few subjects, whereas the above methods do not specifically focus on such sparse effects.

4.1.5 Outline

The rest of the paper is organized as follows. In Section 4.2, we describe the i-Wilcoxon test in detail, followed by numerical experiments to demonstrate its advantage over standard methods. In Section 4.3, we discuss non-interactive tests that are variants of the Wilcoxon signed-rank test to improve its power under heterogeneous treatment effects. In Section 4.4, we discuss extensions of the i-Wilcoxon test to other settings, such as paired data, multiple treatments, and dynamic settings. Section 4.5 concludes the paper by a discussion on the potential of interactive rank tests.

4.2 An interactive Wilcoxon test with covariates (i-Wilcoxon)

To account for covariates through a flexible algorithm that involves human interaction, we propose the i-Wilcoxon test. In short, the analyst decides the ordering of subjects $\{\pi_j\}_{j=1}^n$ and the weights $\{w_j\}_{j=1}^n$ progressively: at step t, she selects the t-th subject from the to-be-ordered subjects $[n]\backslash\{\pi_j\}_{j=1}^{t-1}$ and decides the weight w_t, based on an increasing amount of data information starting from all the assignments $\{A_i\}_{i=1}^n$ masked and then gradually revealed. Mathematically, the data information available to the analyst at the end of step t is denoted by the filtration:

$$\mathcal{F}_t = \sigma\left(\{Y_i, X_i\}_{i=1}^n \cup \{A_{\pi_j}\}_{j=1}^t\right). \tag{50}$$

The choice of π_t and w_t are predictable (measurable) with respect to $\mathcal{F}_{t-1}$, while the analyst is allowed to explore and choose arbitrary models or heuristics to form the ordering and get the weights. After each

iteration of selecting π_t and choosing w_t, the test calculates

$$S_t = \sum_{j=1}^{t} (2A_{\pi_j} - 1) \cdot w_j, \tag{51}$$

and the iteration stops once $|S_t|$ reaches the boundary $u_{\alpha/2}(t)$ as defined in equation (45), or all the subjects are ordered. In other words, let the stopping time be

$$\tau := \min\{t \in [n+1] : |S_t| > u_{\alpha/2}(t) \text{ or } t = n+1\}, \tag{52}$$

where $S_{n+1} \equiv S_n$, and $\tau = n+1$ indicates $|S_t|$ never crosses the boundary. The null is rejected if $\tau \leq n$. Note that although ideally S_t should be large under the alternative, we monitor the absolute value $|S_t|$ because it is possible to guess the opposite assignments when all the assignments are hidden, making S_t goes in the opposite direction as intended and decreases to a smaller value than the null case. Still, the decreasing S_t can reflect difference from the behavior under the null, by monitoring the absolute value $|S_t|$ (see Appendix C.2 for a detailed explanation). We summarize the i-Wilcoxon test in Algorithm 6.

Algorithm 6 Framework for the interactive Wilcoxon test (i-Wilcoxon)

Input: Outcomes, treatment assignment, and covariates $\{Y_i, A_i, X_i\}_{i=1}^n$, target Type-I error rate α;
Procedure:
for $t = 1, \cdots, n$ **do**
  1. Using $\mathcal{F}_{t-1}$, pick any $\pi_t \in [n] \backslash \{\pi_j\}_{j=1}^{t-1}$ and obtain an arbitrary weight $w_t \in \mathbb{R}$;
  2. Reveal A_{π_t} and update $\mathcal{F}_t$;
  if $\left|\sum_{j=1}^{t} (2A_{\pi_j} - 1) \cdot w_j\right| > u_{\alpha/2}(t)$ **then**
    reject the null and stop;
end

Remark 6. *We defined the problem as testing the global null (38) of no treatment effect at a predefined level α. Instead, we could ask the test to output a sequential or anytime p-value for the global null, which is a sequence of p-values $\{p_t\}_{t=1}^{\infty}$ that are valid at any stopping time. Specifically, the stopping boundary $u_{\alpha/2}(t)$ as defined in (45) stems from applying Ville's (often attributed to Doob) maximal inequality [Ville, 1939] to an exponential supermartingale: $M_t := \exp(\lambda S_t - \frac{\lambda^2}{2} \sum_{j=1}^{t} w_j^2)$ for a particular choice of $\lambda = \sqrt{2\log(1/\alpha)/m}$. Indeed by Ville's inequality, $p_t = \inf_{s \leq t} 1/M_s$ is a p-value, and it is anytime valid in the sense that for arbitrary stopping time τ, p_τ is also a p-value. In another perspective, M_t is called a safe e-value, recently proposed by Grünwald et al. [2019]. Their relationship to confidence sequences, sequential tests and anytime p-values is detailed by Ramdas et al. [2020].*

Although more information is revealed to the analyst after each step, the error control is valid. It is because under the null, the increment A_{π_t} for testing is independent of the information for interaction:

$$\mathbb{P}\left(A_{\pi_t} = 1 \mid \mathcal{F}_{t-1}\right) = 1/2. \tag{53}$$

The complete proof is in Appendix C.1.

Theorem 9. *With the flexibility for an analyst to explore, examine, and update working models at any step t using the information in $\mathcal{F}_t$, the i-Wilcoxon test controls type-I error for null hypothesis (38) under assumptions (i),(ii) of randomized experiments.*

The i-Wilcoxon test allows the analyst to incorporate covariates and various types of domain knowledge for ordering and choosing weights. However, manually picking π_t for every step could be tedious and unnecessary. The analyst can instead design an automated algorithm for choosing π_t and w_t, such as the example we provide in the next section, and still keeps the flexibility to modify it at any step.

4.2.1 A concrete, automated, instantiation of i-Wilcoxon

We can infer the treatment assignments by exploring various models to fit the (partial) data. An example is to model the outcome as a mixture of the distributions for treatment and control groups:

$$Y_i \sim \begin{cases} N(\mu_i^1, 1), \text{ when } A_i = 1 \\ N(\mu_i^0, 1), \text{ when } A_i = 0 \end{cases} \quad \text{with } \mu_i^j = \theta_j(X_i) \text{ for } j = 0, 1, \tag{54}$$

where θ_j could be linear functions of the covariates and their second-order interaction terms. The masked treatment assignments can be viewed as missing values, and by the EM algorithm (details in Appendix C.4), we get an estimated posterior probability of receiving the treatment for each subject. The estimated probability of receiving treatment, denoted as $\widehat{q}_i$, provides an estimation of the assignment and an approach to select π_t. Recall that we hope to order upfront the subject whose estimated assignment we are most confident, which can be measured by $|\widehat{q}_i - 0.5|$, so we could select $\pi_t = \text{argmax}_{i\in[n]\backslash\{\pi_j\}_{j=1}^{t-1}}\{|\widehat{q}_i - 0.5|\}$. For the chosen subject, we weight the true assignment $2A_{\pi_j} - 1$ by the estimated assignment $2\widehat{A}_{\pi_j} - 1$:

$$S_t = \sum_{j=1}^{t} (2A_{\pi_j} - 1) \cdot (2\widehat{A}_{\pi_j} - 1), \tag{55}$$

where the estimated assignment $\widehat{A}_{\pi_j} := \mathbb{1}\{\widehat{q}_{\pi_j} > 0.5\}$ is a function of the estimated probability of receiving treatment. By design, the increment of S_t is $+1$ if the estimated assignment is consistent with the truth; and -1 otherwise. Ideally, when the null is false, we could guess most assignments correctly and order them upfront, leading to a larger S_t that could exceed the boundary $u_\alpha(t)$ [8]. We summarize this automated procedure in Algorithm 7.

Algorithm 7 An automated implementation of the i-Wilcoxon test

Input: Outcomes, treatment assignment, and covariates $\{Y_i, A_i, X_i\}_{i=1}^n$, target Type-I error rate α;
Procedure:
for $t = 1, \cdots, n$ **do**
  1. Estimate $\widehat{q}_i$ for subjects in $[n]\backslash\{\pi_j\}_{j=1}^{t-1}$ 2. Choose $\pi_t = \text{argmax}_{i\in[n]\backslash\{\pi_j\}_{j=1}^{t-1}}\{|\widehat{q}_i - 0.5|\}$;
  3. Reveal A_{π_t} and update $\mathcal{F}_t$;
  if $\left|\sum_{j=1}^{t} \left(2A_{\pi_j} - 1\right) \cdot (2\mathbb{1}\{\widehat{q}_{\pi_j} > 0.5\} - 1)\right| > u_{\alpha/2}(t)$ **then**
    reject the null and stop;
end

[8] Recall the discussion in the previous section, that S_t could decrease fast and have smaller value than the null case at the first few iterations (all the assignments are hidden). Thus, we propose an alternative choice of the weights to either make S_t decrease or increase based on the previous trend of S_t (we put it in Appendix C.3 for conciseness of the paper). This alternative strategy tends to result in slightly higher power in numerical experiments.

As test proceeds and more actual assignments get revealed for interaction, we refit the above model and update the estimation of posterior probabilities for every 100 steps (say). Keep in mind that the validity of the error control does not require model (54) to be correct. The analyst can choose other models such as logistic regression for θ_j if the revealed data or prior knowledge suggests so.

4.2.2 Numerical experiments

Simulation setup. To evaluate the performance of the automated algorithm, we simulate 500 subjects ($n = 500$). Suppose each subject is recorded with two binary attributes (e.g., female/male and senior/junior) and one continuous attribute (e.g., body weight), denoted as $X_i = (X_i(1), X_i(2), X_i(3)) \in \{0, 1\}^2 \times \mathbb{R}$. Among n subjects, the binary attributes are marginally balanced, and the subpopulation with $X_i(1) = 1$ and $X_i(2) = 1$ is of size m (see Table 2), where we set $m = 30$. The continuous attribute is independent of the binary ones and follows the distribution of a standard Gaussian.

Table 2: Size of the subpopulation in terms of two binary attributes.

	$X_i(1) = 0$	$X_i(1) = 1$	Totals
$X_i(2) = 0$	m	$n/2 - m$	$n/2$
$X_i(2) = 1$	$n/2 - m$	m	$n/2$
Totals	$n/2$	$n/2$	n

The outcomes are simulated as a function of the covariates X_i and the treatment assignment A_i following the generating model (39), where we vary the functions for the treatment effect Δ and the control outcome f to evaluate the performance of the i-Wilcoxon test. Recall that earlier, we used model (39) as a working model, which is not required to be correctly specified. Here, we generate data from such a model in simulation to provide various types of underlying truth for a clear evaluation of the considered methods[9].

Alternative tests for comparison. In addition to the CovAdj Wilcoxon test, we compare the i-Wilcoxon test with a semi-parametric test derived from the literature of estimating conditional average treatment effect (CATE), which we refer to as the linear-CATE-test. Here, the nonparametric testing problem is transformed into testing a parameter, potentially considering a less stringent null. Specifically, null hypothesis (38) implies that

$$\text{if } \mathbb{E}(Y_i \mid A_i = 1, X_i) - \mathbb{E}(Y_i \mid A_i = 0, X_i) = X_i^T \psi^*, \text{ then } \psi^* = \mathbf{0}. \tag{56}$$

Assume that the outcome difference is a linear function of covariates X_i, the method for CATE provides an asymptotic confidence interval for ψ^*, and the null is rejected if the confidence interval does not include zero (see Appendix C.5 for an explicit form of the test). Note that the test has valid error control even if the outcome difference is not linearly correlated with X_i, in which case, however, the power would be low.

The presented methods (the CovAdj Wilcoxon test, the linear-CATE-test, and the automated algorithm of the i-Wilcoxon test) all involve some working model of the outcomes, but the extent of flexibility varies. The linear-CATE-test requires us to specify the parametric model before looking at the data; the CovAdj Wilcoxon test allows model exploration given partial data $\{Y_i, X_i\}_{i=1}^n$ before testing; and the i-Wilcoxon test further permits the analyst to interactively change the model as the test proceeds and more assignments A_i become available for modeling.

[9] R code to fully reproduce all plots in the paper are available at https://github.com/duanby/interactive-rank.

Test performances when the default model is a good fit. Consider outcomes from the generating model (39) with the treatment effect Δ and the control outcome f specified as:

$$\Delta(X_i) = S_\Delta[X_i(1) \cdot X_i(2) + X_i(3)], \tag{57}$$
$$f(X_i) = 5[X_i(1) + X_i(2) + X_i(3)], \tag{58}$$

where S_Δ encodes the signal strength of the effect. Intuitively, all subjects have some Gaussian-distributed effect correlated with $X(3)$ and the subjects with $X(1) = 1$ and $X(2) = 1$ additionally have a constant positive effect. In such a setting, all the methods with their working models specified as linear functions should fit the data well.

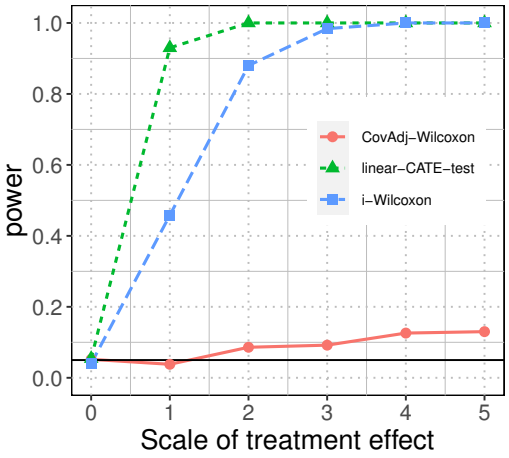


Figure 24: Power of the i-Wilcoxon test compared with the standard tests when varying the scale of the treatment effect, which is defined in (57). The linear model used in all the tests is a good fit for the underlying truth, and the linear-CATE-test (195) has higher power.

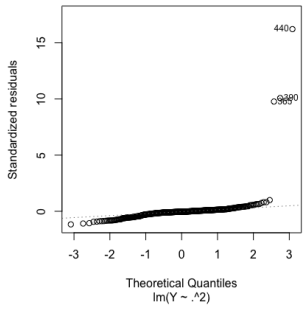
Under heterogeneous treatment effect, the CovAdj Wilcoxon test has low power because the positive effect cancels out with the negative effect in the sum statistics (41), while the linear-CATE-test and the i-Wilcoxon test can cumulate the effect of both signs. The linear-CATE-test has higher power as it targets the specific alternative of nonzero parameters in the linear model (56), although the i-Wilcoxon test also achieves comparable power (see Figure 24). Note that the three methods we compare (CovAdj Wilcoxon test, linear-CATE, i-Wilcoxon test) all have valid type-I error control for the same global null H_0, while these methods target alternatives in different directions. Since we do not make any assumptions on the distribution of non-nulls (ie, if some people do respond, no assumption is made on how they respond), or how informative the covariates are, it is well known that in such nonparametric settings, there is no universally most powerful test (Janssen [2000] discuss this phenomenon when testing goodness of fit). We argue through numerical experiments that our proposed methods could have higher power when the outcomes have a non-normal distribution as discussed below, among other situations.

Illustrations of adaptive modeling. One advantage of the interactive test is that it allows exploration of the working model using the revealed data. Here, we present two examples where model (54) might not fit the data well, but the i-Wilcoxon test can have higher power than the default automated algorithm because, before testing, the analyst explores and evaluates various models to find a reasonably good fit.

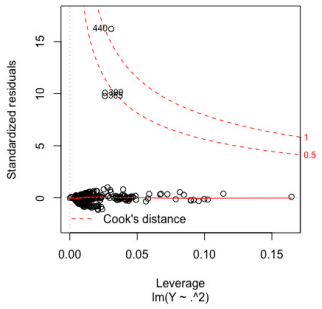
Suppose the control outcome is nonlinearly correlated with the attributes by specifying function f in the generating model (39) as

$$f(X_i) = 2\exp\{-2X_i(3)\}\mathbb{1}(X_i(3) < -2), \tag{59}$$

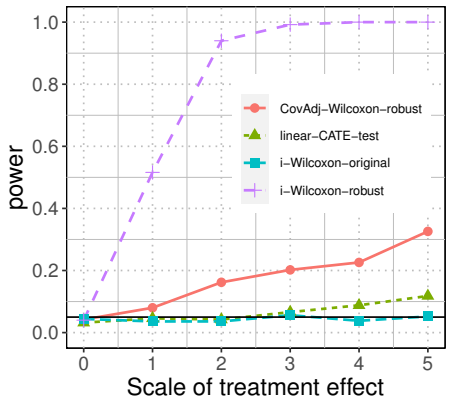
where the distribution of potential control outcomes is skewed (treatment effect Δ is the same as before in (57)). When we fit the default working model (54) with linear functions, the QQ-plot and Cook's distance indicate a poor fit because of possible outliers in the outcomes (see Figure 25a and 25b). An easy fix is to use robust linear regression, which leads to significant power improvement compared with the default algorithm (see Figure 25c). In practice, we recommend using the robust regression, since it keeps good power when the working model is correct while it improves power when the control outcome has a skewed distribution. The robust regression is also observed to improve power under heavy-tailed noise (see Appendix C.7).



(a) Eg: diagnose a misfit via QQ-plot for the original linear model before testing.



(b) Eg: diagnose a misfit via Cook's distance for the original linear model before testing.



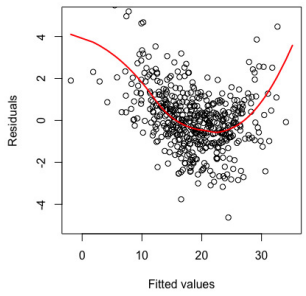
(c) Power under skewed control outcome (67).

Figure 25: Before ordering and testing, the analyst is allowed to explore and examine different working models using the revealed data $\{Y_i, X_i\}_{i=1}^n$. In the example with skewed control outcome, the QQ-plot and Cook's distance of the regular linear regression suggest outliers in the outcomes. The analyst can instead choose the robust linear regression, and the power is higher than that using the default model. For fair comparison, the CovAdj Wilcoxon test (41) is also implemented with robust linear regression. In plots of this section, the power is averaged over 500 repetitions and the error bar is omitted because its length is usually less than 0.02.

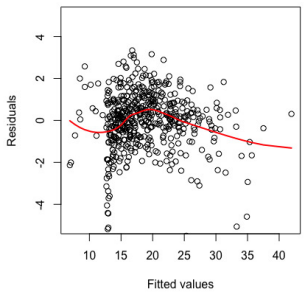
Another example considers the treatment effect as a quadratic function of the covariates, by specifying the function Δ in the generating model (39) as

$$\Delta(X_i) = S_\Delta \left[\frac{3}{5}\left(X_i^2(3) - 1\right)\right]. \tag{60}$$

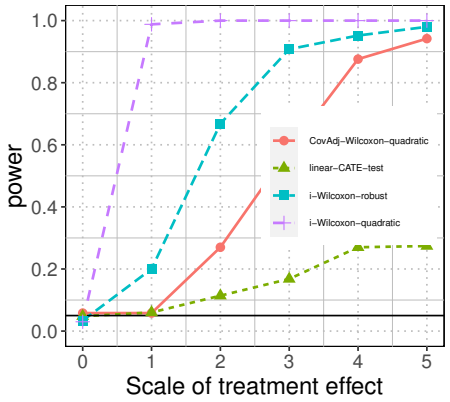
The control outcome is linearly correlated with the attributes as defined in (58). We observe that with the robust linear regression, the residuals have a nonlinear trend (see Figure 26a), indicating that the linear functions of covariates might not be accurate. If we add a quadratic term of $X_i^2(3)$ in the robust regression, the trend in residuals is less obvious, and the model fits better (see Figure 26b). As a result, the power is higher than the test using robust linear regression (see Figure 26c). Note that the presented



(a) Residual plot when using the robust linear regression.



(b) Residual plot when applying regression with a quadratic term.



(c) Power when the treatment effect is a nonlinear function of the covariates.

Figure 26: A second illustration of model exploration when the treatment effect is nonlinearly correlated with the attributes. The residuals show a quadratic pattern when using robust linear regression, and this trend is weakened by adding a quadratic term in the regression, suggesting the latter is a better modeling choice; this type of exploration using only $\{Y_i, X_i\}$ is permitted without violating error control, and can be repeated as $\{A_i\}$ are revealed one by one. The power can be improved using the adjusted (quadratic) model because the i-Wilcoxon test permits the analyst to explore models. For fair comparison, the CovAdj Wilcoxon test is also implemented with a quadratic term.

experiments use a large sample size ($n = 500$), and the results with a small sample size can be found in Appendix C.8. In short, the power of the i-Wilcoxon test decreases when the sample size is small but is still high when the treatment effect is nonlinear. Generally when the sample size is small, we recommend a Bonferroni correction of the i-Wilcoxon test and the non-interactive permutation tests we introduce in Section 4.3.

To summarize, the i-Wilcoxon test has valid error control without any parametric assumption on the outcomes and yet allows exploration of the working models so that the algorithm can adapt to different underlying data distribution. In practice, the working model can also be changed in the middle of the testing procedure, for example, if it fits the data worse as more treatment assignments get revealed. The flexibility of interactive data-dependent model design with the freedom of adjustment on the fly makes the i-Wilcoxon test with parametric working models practical and promising. One can also employ nonparametric working models, and infer the assignments based on nonparametric extensions of the EM algorithm (see Train [2008] without covariate information and Huang et al. [2013] for univariate covariate using kernel regression). To incorporate nonparametric modeling under various data types without involving advanced EM algorithms, we propose a variation of the i-Wilcoxon test in the next section.

4.2.3 A variation of the i-Wilcoxon test without parametric modeling

In the above automated algorithms, we use parametric working models for the outcomes because it enables us to use the EM algorithm to infer the posterior probabilities of receiving treatment when the actual treatment assignment is hidden. Below, we propose a variation if one prefers to use a nonparametric model such as random forest, and still get an estimated posterior probability of receiving treatment for ordering.

We randomly split the sample $D = \{Y_i, A_i, X_i\}_{i=1}^n$ into two parts by index (of equal size by default),

denoted as $D^{(1)}$ and $D^{(2)}$. First, use $D^{(1)}$ with the complete data information to train a classifier (e.g., random forest) for A_i using $\{Y_i, X_i\}$. With this initial model, we follow the procedure of the i-Wilcoxon test on $D^{(2)}$. That is, the assignment A_i in $D^{(2)}$ is masked, and we use the model trained by $D^{(1)}$ to estimate the probability of receiving treatment for subjects in $D^{(2)}$. The test statistic S_t cumulates A_i only for subjects in $D^{(2)}$ after ordering them based on the estimated probabilities. As the test proceeds, the actual assignments in $D^{(2)}$ are progressively revealed so that we obtain the complete data of more subjects, using which we can update the classifier at any step.

In this section, we presented the i-Wilcoxon test, which allows a human to guide the model or heuristic for ordering while keeping valid error control without parametric assumption on the underlying truth. As alternatives, we next introduce and compare several non-interactive nonparametric tests that are variants of the Wilcoxon signed-rank test in the following.

4.3 Options for adjusting Wilcoxon's signed-rank test for covariates

The Wilcoxon signed-rank test is a simple and efficient nonparametric test with a known null distribution. Of course, rank-based statistics have been explored in many directions: see Lehmann and D'Abrera [1975] for a review of classical methods. Recent work focuses on how to incorporate covariate information to improve power. Zhang et al. [2012] develop an optimal statistic to detect constant treatment effect; in multi-sample comparison, Ding et al. [2018] numerically compare rank statistics of outcomes or residuals from linear models; Rosenblum and Van Der Laan [2009] and Vermeulen et al. [2015] focus on related testing problems for conditional average effect and marginal effect; Rosenbaum [2010] and Howard and Pimentel [2020] use generalizations of rank tests for sensitivity analysis in observational studies. Here, we introduce variants of the signed-rank test for two-sample comparison in a randomized trial, which can improve the power of Rosenbaum's CovAdj Wilcoxon test under heterogeneous treatment effect.

The signed-rank test offers a general formula to construct tests for two-sample comparison. We note that the signed-rank test is perhaps more frequently used for paired data; but it can also be applied to unpaired data because the error control is also based on a decoupling between the sign and the rank. We discuss methods for the paired setting in Section 4.4.1. In the unpaired setting, for each subject $i \in [n]$, let E_i be any statistic that is larger when subject i has treatment effect. We compute

$$W = \sum_{i=1}^{n} \text{sign}(E_i)\text{rank}(|E_i|), \tag{61}$$

and the null is rejected when W is large. As an example, Rosenbaum [2002] proposed the covariance-adjusted signed-rank test by specifying E_i as

$$E_i^{R(X)} := (2A_i - 1)R_i, \tag{62}$$

where recall R_i is the residual of regressing Y_i on X_i without using A_i as a predictor. (The covariance-adjusted signed-rank test is slightly different from the covariance-adjusted Wilcoxon rank-sum test (41), but they had similar power in most of our experiments.) The null distribution of W depends on E_i, but one can use a permutation test that is valid for any choice of E_i. Recall that under the null, the assignment A_i is independent of other data information $\{Y_i, X_i\}$, as stated in (42). The permutation test estimates the null distribution of W by permuting the treatment assignments $\{A_i\}_{i=1}^n$, described as follows:

(i) calculate W using the observed data $\{Y_i, A_i, X_i\}_{i=1}^n$;

(ii) let $W^1 = W$ and for $b = 2, \ldots, B$, generate a random permutation of the treatment assignments $(A_1^b, \ldots, A_n^b)$; and calculate W^b using the permuted data $\{Y_i, A_i^b, X_i\}_{i=1}^n$;

(iii) obtain the p-value as $\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(W^b \geq W)$.

Ideally, statistic E_i should be designed to take larger value when subject i has larger treatment effect. In the following, we discuss the question of whether the original choice of $E_i = E_i^{R(X)}$ can be improved, and which choice of E_i should we prefer given different types of treatment effect.

4.3.1 Existing statistics and their drawbacks

Aside from Rosenbaum's design of E_i as $E_i^{R(X)}$, we can find several other alternatives to detect treatment effects in the causal inference literature. For example, one can construct a confidence interval for the ATE, which implies a test for zero ATE. However, the null of zero ATE is not the focus of this paper, as we are interested in the null of zero effect for any subpopulation. Lin [2013] suggests modeling Y_i by a linear function of A_i and X_i (recently extended in a preprint by Guo and Basse [2020] to other parametric models), and construct the estimator for ATE as an average over subjects:

$$\frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)(Y_i - \widehat{Y}(X_i; 1 - A_i)),$$

where $\widehat{Y}(\cdot;\cdot)$ denotes a fitted outcome using X_i, A_i and $\widehat{Y}(X_i; 1 - A_i)$ predicts using the *false* assignment.

This estimator provides a design of E_i that calculates the residual of predicting Y_i using covariates X_i and the false assignment $1 - A_i$ as follows:

$$E_i^{R(X,1-A)} := (2A_i - 1)(Y_i - \widehat{Y}(X_i; 1 - A_i)), \tag{63}$$

where $\widehat{Y}(X_i; 1 - A_i)$ can be the prediction via any black box algorithm, such as a random forest.

There is also a rich literature on doubly-robust methods (see, for example, Cao et al. [2009]; Chernozhukov et al. [2018]; Robins et al. [1994]; Robinson [1988]) to estimate ATE when the probability of receiving treatment varies with X_i. In a randomized experiment, the estimator boils down to

$$\frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)(Y_i - \widehat{Y}(X_i; 1)/2 - \widehat{Y}(X_i; 0)/2),$$

which suggests a design of E_i as $(2A_i - 1)(Y_i - \widehat{Y}(X_i; 1)/2 - \widehat{Y}(X_i; 0)/2)$. This design leads to similar power as $E_i^{R(X,1-A)}$ in most experiments and hence is omitted from this paper.

To examine the performance of tests using the statistics $E_i^{R(X)}$ and $E_i^{R(X,1-A)}$, we simulate outcomes from the generating model (39) where the function for treatment effect Δ and that for control outcome f are constructed with different features (e.g., dense/sparse effect and bell-shaped/skewed control outcome):

$$\Delta(X_i) = S_\Delta\left[1 - |\sin(3X_i(3))|\right] \quad \textbf{(dense and weak effect)}; \tag{64}$$

$$\Delta(X_i) = S_\Delta\left[2\exp\{X_i(3)\}\mathbb{1}\left(X_i(3) > 1.5\right)\right] \quad \textbf{(sparse and strong effect)}; \tag{65}$$

$$f(X_i) = 5[X_i(1) + X_i(2) + X_i(3)] \quad \textbf{(bell-shaped control outcome)}; \tag{66}$$

$$f(X_i) = 2\exp\{-2X_i(3)\}\mathbb{1}(X_i(3) < -2) \quad \textbf{(skewed control outcome)}. \tag{67}$$

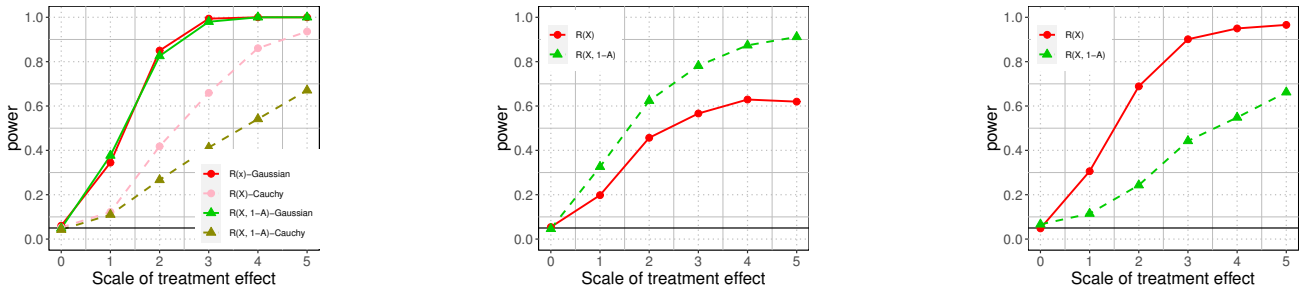
The dense (sparse) effect is set to be weak (strong) since otherwise all methods have power near one (zero).

We intentionally let the treatment effect and control outcome be nonlinear functions of the covariates because our discussion focuses on methods using nonparametric working models. In the rest of this paper, we employ random forests (with default parameters in the R package `randomForest`) as our working model since it usually generates good predictions for various data distributions [Breiman, 2001].

Although both methods have high power under a well-behaved distribution where the treatment effect is dense, the control outcome is bell-shaped, and the noise is standard Gaussian (solid lines in Figure 27a), they show different weak points when the effect is harder to detect—the test using $E_i^{R(X)}$ tends to have lower power when the treatment effect is sparse (Figure 27b); and the test using $E_i^{R(X,1-A)}$ tends to be less robust when the control outcome is skewed (Figure 27c). When the noise is heavy-tailed, both tests have lower power as expected, but the one using $E_i^{R(X,1-A)}$ appears to be more sensitive (Figure 27a). Broadly, the aforementioned pros and cons may be traced to two characteristics in the design of E_i:

(i) the prediction model that uses both X_i and A_i as in $E_i^{R(X,1-A)}$ accounts for heterogeneous treatment effect (by the interaction terms between X_i and A_i), leading to high power for sparse effects;

(ii) the residuals in $E_i^{R(X)}$ only uses X_i as predictors so that it effectively reduces the outcome variation that is *not* caused by the treatment, making the test robust under skewed control outcome.



(a) Power when the treatment effect is dense and the control outcome is bell-shaped, and the noise varies as Gaussian and Cauchy (heavy-tailed).

(b) Power when the treatment effect is sparse, the control outcome is bell-shaped, and the noise is Gaussian.

(c) Power when the treatment effect is dense, the control outcome is skewed, and the noise is Gaussian.

Figure 27: Power of the Wilcoxon test (61) using $E_i^{R(X)}$ and $E_i^{R(X,1-A)}$ as the scale of treatment effect S_Δ increases under different types of treatment effect, control outcome and noise. The test when using $E_i^{R(X,1-A)}$ tends to be more sensitive to heavy-tailed noise or skewed control outcome; and the test with $E_i^{R(X)}$ can have lower power when the treatment effect is sparse. Here and henceforth, we use 200 permutations, and the experiment is repeated 500 times.

Next, we propose other designs of E_i that combine the advantages of the above two characteristics.

4.3.2 Improve robustness under skewed control outcome by predicting residuals R_i

Because residuals R_i can downsize the noise caused by skewed control outcome, we propose to measure the treatment effect via a prediction on R_i. That is, we compute the statistic E_i by two steps of prediction:

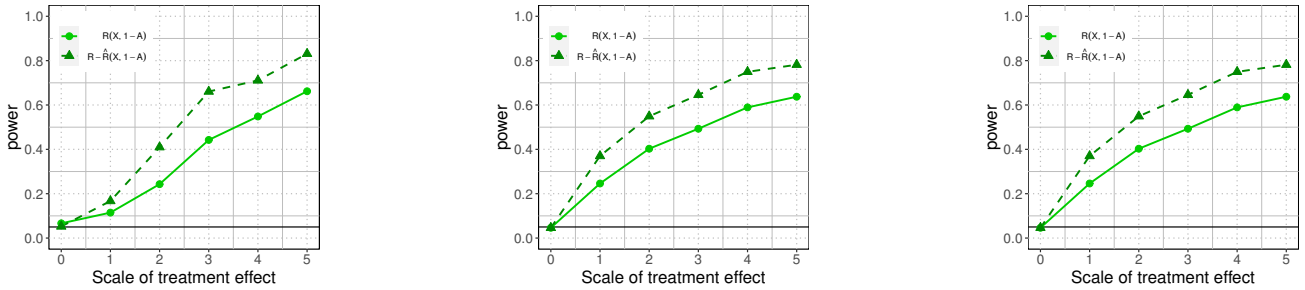
(i) obtain residuals R_i by predicting Y_i using X_i (without A_i);

(ii) fit a prediction model for R_i using X_i and A_i, denoted as $\widehat{R}(\cdot, \cdot)$;

(iii) get E_i from the prediction error of R_i using covariates X_i and the false assignment $1 - A_i$:

$$E_i^{R-\widehat{R}(X,1-A)} := (2A_i - 1)(R_i - \widehat{R}(X_i, 1 - A_i)). \tag{68}$$

Notice that $E_i^{R-\widehat{R}(X,1-A)}$ has a similar form as $E_i^{R(X,1-A)}$, where $\{R_i\}_{i=1}^n$ can be viewed as "denoised" outcomes: a large Y_i could stem from skewness in the control outcome, but a large R_i is more likely to indicate large treatment effect, and hence achieves higher robustness to skewed control outcome. Numerical experiments coincide with our intuition: the power of using $E_i^{R-\widehat{R}(X,1-A)}$ improves from that using $E_i^{R(X,1-A)}$ when the control outcome is skewed (see Figure 28).



(a) Power when the treatment effect is dense and weak.

(b) Power when the treatment effect is sparse and strong.

(c) Power when the treatment effect is sparse and strong.

Figure 28: Power of Wilcoxon test (61) using $E_i^{R(X,1-A)}$ and $E_i^{R-\widehat{R}(X,1-A)}$ as the treatment effect increases under skewed control outcome. The latter has higher power for both dense and sparse effects.

4.3.3 Improve robustness under heavy-tailed noise using difference in the prediction error

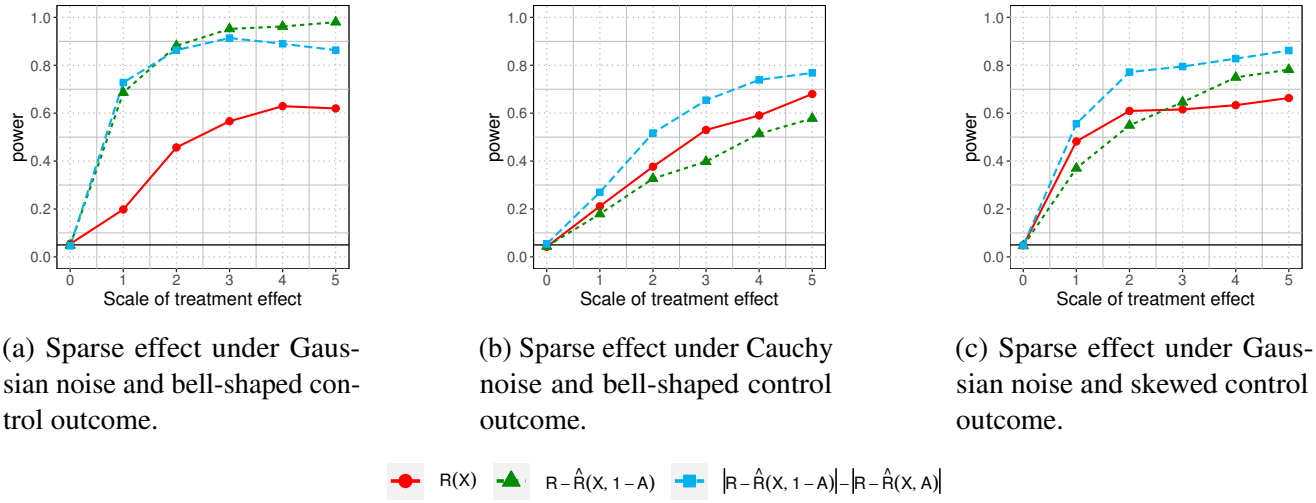
Treating residuals R_i as the pseudo outcomes is useful to account for variation in the control outcome, but R_i can still contain much irrelevant variation, such as when the random noise U_i in model (39) is Cauchy. Under heavy-tailed noise, the prediction model $\widehat{R}(\cdot, \cdot)$ in $E_i^{R-\widehat{R}(X,1-A)}$ could be inaccurate; and a large prediction error of using the false assignment as in $E_i^{R-\widehat{R}(X,1-A)}$ could result from heavy-tailed noise, while it is supposed to be evidence of large treatment effect.

So how to remove the large prediction error caused by poor modeling? We propose to consider the difference between the prediction error of using the false assignment $|\widehat{R}(X_i, 1 - A_i) - R(X_i)|$ and that using the true assignment $|\widehat{R}(X_i, A_i) - R(X_i)|$:

$$E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|} := |\widehat{R}(X_i, 1 - A_i) - R(X_i)| - |\widehat{R}(X_i, A_i) - R(X_i)|. \tag{69}$$

Intuitively, when the prediction model $\widehat{R}(\cdot, \cdot)$ is a good fit, the prediction error using true assignment $|\widehat{R}(X_i, A_i) - R(X_i)|$ should be close to zero, and the proposed statistic is similar to $E_i^{R-\widehat{R}(X,1-A)}$. The advantage shows when the modeling is poor, such as under heavy-tailed noise. Here, the prediction error is large using either true or false assignment, so taking their difference as in $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ can help rule out the variation caused by noise, letting the variation from treatment effect stand out. In

the experiment with sparse effect (65), the test using $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ has similar power as that using $E_i^{R-\widehat{R}(X,1-A)}$ when data is well-distributed (see Figure 29a), while it can achieve higher power under Cauchy noise or skewed control outcome (see Figure 29b and 29c), consistent with our intuition.

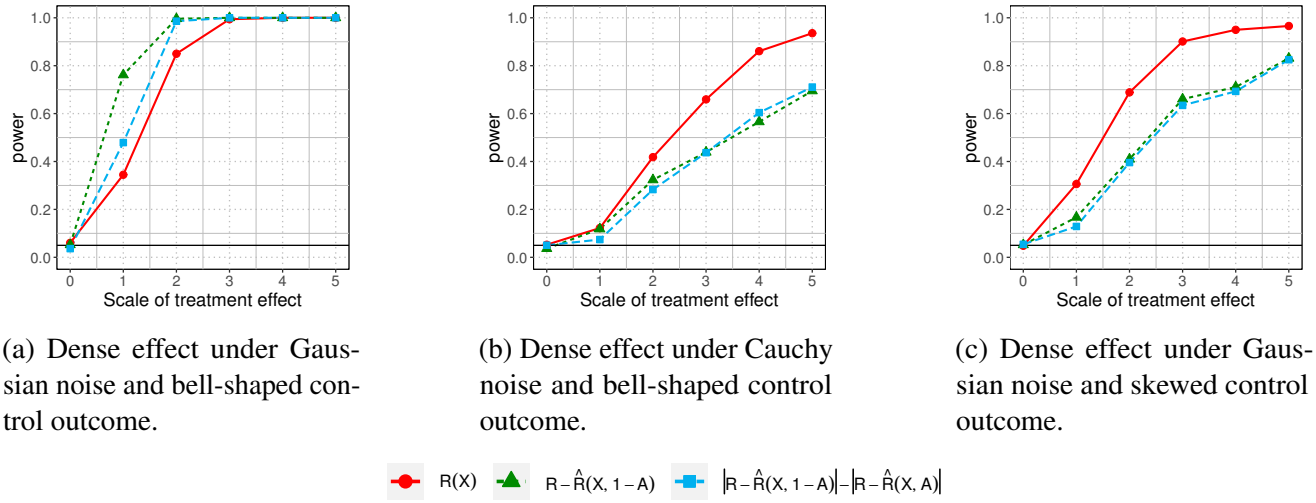


(a) Sparse effect under Gaussian noise and bell-shaped control outcome.

(b) Sparse effect under Cauchy noise and bell-shaped control outcome.

(c) Sparse effect under Gaussian noise and skewed control outcome.

Figure 29: The power of Wilcoxon test (61) using three statistics: $E_i^{R(X)}$, $E_i^{R-\widehat{R}(X,1-A)}$, and $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ under sparse treatment effect, with the noise varies as Gaussian and Cauchy, and the control outcome varies as a bell-shaped or skewed distribution. The test using $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ tends to have higher power especially under heavy-tailed noise or skewed control outcome.



(a) Dense effect under Gaussian noise and bell-shaped control outcome.

(b) Dense effect under Cauchy noise and bell-shaped control outcome.

(c) Dense effect under Gaussian noise and skewed control outcome.

Figure 30: The power of Wilcoxon test (61) using three statistics: $E_i^{R(X)}$, $E_i^{R-\widehat{R}(X,1-A)}$, and $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ under dense and weak treament effect, with the noise varies as Gaussian and Cauchy, and the control outcome varies as a bell-shaped or skewed distribution. Rosenbaum's Wilcoxon test using $E_i^{R(X)}$ can be more robust to heavy-tailed noise or skewed control outcome.

Remark 7. *Note that $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ leads to high power when we want to detect a sparse and strong effect. However, when the effect is dense and weak as in model (64), Rosenbaum's Wilcoxon*

test using $E_i^{R(X)}$ is more robust to peculiar noise or control outcomes (see Figure 30). It is because $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ uses a prediction model for R_i, which can be less informative for weak effect, especially when the noise is large. In practice, one may have some anticipation on the population properties of the treatment effect (density or strength), and choose the statistic accordingly. We summarize our recommendations under different settings in flowchart (74).

4.3.4 On one-sided versus two-sided effects

The statistic of difference in the prediction error leads to high power for two-sided effects. A major distinction between $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ and the statistics discussed previously is that it takes large value for both positive and negative effects. It is because the difference in the prediction error of using opposite assignments is large as long as the assignment is a significant predictor for the outcome, regardless of the direction of effect. Therefore, the test using $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ can cumulate effects of both signs while they cancel out in other statistics, leading to high power even when the average effect is close to zero. As some examples, we construct the following treatment effect:

$$\Delta(X_i) = S_\Delta \left[\exp\{X_i(3)\}\mathbb{1}\left(X_i(3) > 2\right) - X_i(1)/2\right] \tag{70}$$
(Sparse strong positive effect and dense weak negative effect);

$$\Delta(X_i) = S_\Delta \left[X_i^3(3)\mathbb{1}(|X_i(3)| > 1)\right] \tag{71}$$
(Sparse strong effect of both signs);

$$\Delta(X_i) = S_\Delta \left[\frac{2}{5}\sin(3X_i(3))\right] \tag{72}$$
(Dense weak effect of both signs).

In all examples, only the test using $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ has nontrivial power (see the first row in Figure 31). Such sensitivity may or may not be desirable depending on the problem context. For example, we would hope to reject the null when the positive effect is strong for a subpopulation as in (70). However, one might want to treat a weak effect in both directions (72) as noise and leave the null unrejected. Next, we propose a modification of $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ with such behavior.

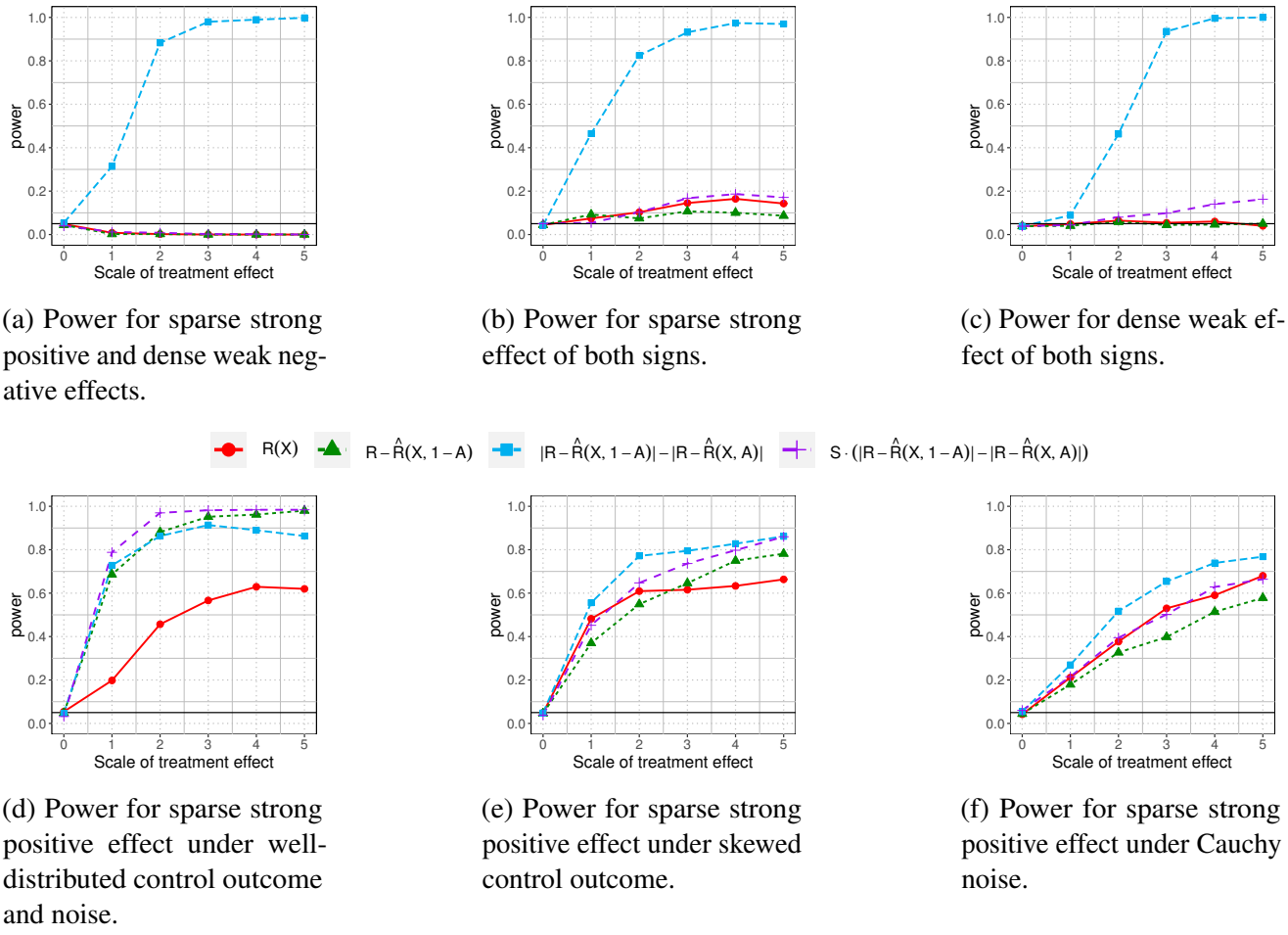
Targeting one-sided effects. To differentiate between positive and negative effects, we modify the statistic $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ by incorporating a sign that indicates the direction of the treatment effect. Consider the sign of two other statistics that approximate the treatment effect:

$$\begin{aligned}
S_i^1 &:= \mathbb{1}\{E_i^{R-\widehat{R}(X,1-A)} \geq 0\} \equiv \mathbb{1}\{(2A_i-1)(R_i - \widehat{R}(X_i, 1-A_i)) \geq 0\},\\
S_i^2 &:= \mathbb{1}\{(2A_i-1)(\widehat{R}(X_i, A_i) - \widehat{R}(X_i, 1-A_i)) \geq 0\}, \quad \text{and combine them to get}\\
S_i &:= \mathbb{1}\{S_i^1 > 0 \text{ or } S_i^2 > 0\}.
\end{aligned}$$

We then define

$$E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)} := (2S_i - 1)\cdot E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}, \tag{73}$$

which is large when the treatment effect is large and *positive*. We tried using only S_i^1 or S_i^2 for the sign, but the combined one is more robust in experiments. The essential idea is to construct S_i using



(a) Power for sparse strong positive and dense weak negative effects.

(b) Power for sparse strong effect of both signs.

(c) Power for dense weak effect of both signs.

(d) Power for sparse strong positive effect under well-distributed control outcome and noise.

(e) Power for sparse strong positive effect under skewed control outcome.

(f) Power for sparse strong positive effect under Cauchy noise.

Figure 31: Power of Wilcoxon test (61) using four statistics: $E_i^{R(X)}$, $E_i^{R-\widehat{R}(X,1-A)}$, $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ and $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$. In the first row where the treatment effect can be positive or negative, only the test using $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ has nontrivial power. In the second row, the treatment effect is sparse and positive, and the control outcome and noise varies. The test using $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ can have high power without being too sensitive to the weak effect in both directions (see subplot 31c).

some statistics that have a consistent sign with the treatment effect, while keeping the advantage of $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ under skewed control outcome and heavy-tailed noise.

As desired, the test using $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ is less sensitive to weak effect of both signs (Figure 31c) and keeps high power for sparse strong positive effect (Figure 31d). Note that the signed statistic is more sensitive to noise because the signs are generated from less robust statistics (Figures 31e, 31f). Nonetheless, among statistics that are insensitive to two-sided effect, $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ leads to high power for sparse effect, irrespective of whether the control outcome and the noise are well-distributed or have outliers.

4.3.5 Summarizing the observations made in this section

In this section, we proposed several variants of Rosenbaum's covariate adjusted Wilcoxon as follows:

(i) Instead of predicting the *outcomes*, using the prediction model $\widehat{R}(\cdot,\cdot)$ for *residuals* R_i can improve

power under skewed control outcome. This is because the residuals R_i, which are themselves obtained by regressing Y_i only on X_i (without A_i), can remove much variation caused by the control outcome, and in turn highlight the treatment effect (see Section 4.3.2).

(ii) The evidence of treatment effect can be measured by the prediction error using the false assignment, but large prediction error could also be a result of poorly fit model, such as when the noise is heavy-tailed. In contrast, the difference in the prediction error of using true and false assignments can eliminate most of the prediction error that is irrelevant to the treatment, including that from poorly fit models, and thus improve the power (see Section 4.3.3).

(iii) The difference in prediction error detects both positive and negative effects with no distinction, so it can arguably be *too sensitive* (if there is such a thing) to a weak effect in both directions. If one wishes to target one-sided effects while maintaining the robustness achieved by "difference in prediction error", we propose to multiply it with an estimated sign of the effect (see Section 4.3.4).

In summary, we recommend choosing one out of the three test statistics discussed in this section—$E_i^{R(X)}$, $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$, and $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$—depending on one's prior belief of the population properties of treatment effect (if one exists), as shown below:

$$\text{Nonzero effect} \begin{cases} \text{Effect of both signs} \to E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|} \\ \text{Positive effect} \begin{cases} \text{Sparse and strong effect} \to E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)} \\ \text{Dense and weak effect} \to E_i^{R(X)} \end{cases} \end{cases} \tag{74}$$

When there is little prior knowledge on the properties of treatment effect, we recommend using the Bonferroni correction of the above three Wilcoxon tests, which we call the Wilcoxon-Bonferroni test. This combination shows similar power as the recommended test in most simulations (see Appendix C.6 for simulation results). The presented experiments use a large sample size ($n = 500$), and the power comparison under a small sample size is similar (see the last paragraph of Appendix C.8). Note that the i-Wilcoxon test is not included here because its performance depends on the interaction and progressive updates to the initial working model made by the analyst based on revealed data. The flexibility makes the i-Wilcoxon test a potentially more robust and promising method compared with the aforementioned methods that also use a parametric (or semiparametric) working model.

4.4 Extensions

We have investigated several tests to account for heterogeneous treatment effect: (a) a new i-Wilcoxon test that allows human interaction; and (b) variants of the Wilcoxon signed-rank test. So far, the paper focuses on the setting of comparing *two* samples with *unpaired* data that is collected before testing as a *batch*. However, the proposed tests can be extended to other settings: (a) both the i-Wilcoxon test and the variants of the Wilcoxon signed-rank test can be applied to paired data; and (b) the i-Wilcoxon test can be extended to a multi-sample comparison for data with/without block structure (i.e., matching); and (c) the interactive test also works for two/multi-sample comparison with/without matching data in dynamic settings, where we obtain new data as the test proceeds.

4.4.1 Two-sample comparison with paired data

Suppose there are n pairs of subjects. Let the outcomes of subjects in the i-th pair be Y_{ij}, the treatment assignments be indicators A_{ij}, the covariates be vector X_{ij} for $j = 1, 2$ and $i \in [n]$. The null hypothesis of interest is that there is no difference between treatment and control outcomes conditional on covariates:

$$(Y_{ij} \mid A_{ij} = 1, X_{ij}) \overset{d}{=} (Y_{ij} \mid A_{ij} = 0, X_{ij}) \text{ for all } j = 1, 2 \text{ and } i \in [n]. \tag{75}$$

This paper deals with randomized experiments, and assume that

(i) the treatment assignments are independent across pairs, and randomized within each pair:

$$\mathbb{P}(A_{i1} = 1, A_{i2} = 0) = \mathbb{P}(A_{i1} = 0, A_{i2} = 1) = 1/2, \text{ for all } i \in [n];$$

(ii) the outcome of one subject Y_{i_1,j_1} is independent of the treatment assignment of another subject A_{i_2,j_2} for any $(i_1, j_1) \neq (i_2, j_2) \in [n] \times [2]$.

Under the null, observe that

$$\mathbb{P}(A_{i1} - A_{i2} = 1 \mid Y_{i1}, Y_{i2}, X_{i1}, X_{i2}) = 1/2 \quad \text{for all } i \in [n], \tag{76}$$

which is similar to the critical property (42) that guarantees the error control of all discussed methods. We can compress the paired data to an "unpaired" form, by treating the difference of paired assignments (after rescaling) $\widetilde{A}_i := (A_{i1} - A_{i2} + 1)/2$ as the pseudo treatment assignment, and the difference in the paired outcomes $\widetilde{Y}_i := Y_{i1} - Y_{i2}$ as the pseudo outcome, and the union of the covariates as the pseudo covariates $\widetilde{X}_i := \{X_{i1}, X_{i2}\}$. In such a way, observation (42) holds under the null with pseudo data $\{\widetilde{Y}_i, \widetilde{A}_i, \widetilde{X}_i\}_{i=1}^n$, and hence all the methods can be applied to paired data with valid error control. Meanwhile, under the alternative with positive (negative) effect, the outcome difference $\widetilde{Y}_i$ is positively (negatively) correlated with the (rescaled) assignment difference $\widetilde{A}_i$, so our proposed tests can have nontrivial power. For example, in the i-Wilcoxon test, the outcome difference $\widetilde{Y}_i$ can be used along with the union of covariates $\widetilde{X}_i$ to gather pairs with positive $\widetilde{A}_i$, as described in Algorithm 6 once we replace the input data with $\{\widetilde{Y}_i, \widetilde{A}_i, \widetilde{X}_i\}_{i=1}^n$.

Interestingly, we can derive another set of corresponding tests for the paired data from a different perspective. Rosenbaum [2002] and Howard and Pimentel [2020] consider the treatment-minus-control difference of the outcome, denoted as $D_i := (A_{i1} - A_{i2})(Y_{i1} - Y_{i2})$. Observe that under the null,

$$\mathbb{P}(\text{sign}(D_i) = 1 \mid |D_i|, X_{i1}, X_{i2}) = 1/2 \quad \text{for all } i \in [n], \tag{77}$$

because $(A_{i1} - A_{i2})$ has equal probability to be positive or negative as in (76). Note that here, we assume the outcomes are continuous to avoid nonzero probability of $\text{sign}(D_i) = 0$. Under the alternative, the treatment-minus-control difference D_i can bias to positive (or negative) value. Therefore, all the discussed methods can be applied to the data $\{|D_i|, \text{sign}(D_i), \widetilde{X}_i\}_{i=1}^n$ where $\text{sign}(D_i)$ is viewed as the pseudo treatment assignment (if rescaled), and $|D_i|$ as the pseudo outcome. In fact, using this design of pseudo data in the Wilcoxon signed-rank test (61) leads to the classical Wilcoxon test for paired sample.

Generally, we can derive nontrivial tests of similar forms for various problems, as long as we can find a binary statistic for each individual (subject or pair) that is independent of other data information under the null, but can be effectively inferred under the alternative. In the next section, we show that the i-Wilcoxon test can be further extended to using test statistics that are not binary.

4.4.2 Multi-sample comparison for data with/without block structure

Tests for data without block structure. In multi-sample comparison, the case where subjects are not matched is often referred to as data without block structure, for which a classical test is the Kruskal-Wallis test [Kruskal and Wallis, 1952] (details in Appendix C.9). We call the interactive test in this setting the i-Kruskal-Wallis test. Follow the notation of two-sample comparison with unpaired data, where the treatment assignment A_i now takes values in $[k] \equiv \{1, \ldots, k\}$ for k-sample comparison. The null hypothesis asserts that there is no difference between outcomes of any two treatments conditional on covariates:

$$(Y_i \mid A_i = a_1, X_i) \overset{d}{=} (Y_i \mid A_i = a_2, X_i) \text{ for all } i \in [n] \text{ and } a_1, a_2 \in [k]. \tag{78}$$

In a randomized experiment, we assume that

(i) the treatment assignments are independent and randomized

$$\mathbb{P}(A_i = a \mid X_i) = 1/k \text{ for all } i \in [n] \text{ and } a \in [k];$$

(ii) the outcome of one subject Y_{i_1} is independent of the assignment of another A_{i_2} for any $i_1 \neq i_2 \in [n]$.

Observe that under the null,

$$\mathbb{P}(A_i = a \mid Y_i, X_i) = 1/k \quad \text{for all } a \in [k] \text{ and } i \in [n], \tag{79}$$

similar to two-sample comparison. In other words, A_i is independent of $\{Y_i, X_i\}$ with a known distribution. A difference from comparing two treatment is that under the alternative, the association between the outcome Y_i and the treatment A_i can have various patterns depending on the underlying truth. Here, we consider an example of the i-Kruskal-Wallis test that targets a specific type of alternative.

Given three treatments ($k = 3$), suppose we wish to target the alternative of decreasing outcomes:

$$(Y_i \mid A_i = 1, X_i) \succeq (Y_i \mid A_i = 2, X_i) \succeq (Y_i \mid A_i = 3, X_i), \tag{80}$$

where $Y^1 \succeq Y^2$ means that Y^1 stochastically dominants Y^2. We can define the pseudo assignment $\widetilde{A}_i$ as

$$\widetilde{A}_i = \begin{cases} 1, & \text{if } A_i = 1, \\ 0, & \text{if } A_i = 2, \\ -1, & \text{if } A_i = 3, \end{cases}$$

such that $\widetilde{A}_i$ is larger for larger outcomes under the targeted alternative. The i-Kruskal-Wallis test can then use $\{Y_i, X_i\}$ to infer and gather $\widetilde{A}_i$ with larger values, and reject the null. The complete procedure follows Algorithm 6 where the input data is replaced by $\{Y_i, \widetilde{A}_i, X_i\}$.

Note that the error control uses boundary $u_{\alpha/2}(t)$ for fair coin flips although $\widetilde{A}_i$ is not binary, because here the null distribution of $|S_t|$ is stochastically dominated by the sum of weighted coin flips (given that the null distribution of $\widetilde{A}_i$ is discrete uniform in $\{-1, 0, 1\}$). We can also use tighter boundaries for cumulative sums of discrete uniforms, which are well-studied by Howard et al. [2020a] and Howard et al. [2020b]. Keep in mind that the above design of $\widetilde{A}_i$ is an example to target the specific alternative (80) for three treatments; similar tests can be developed for other alternatives or comparing more treatments.

Tests for data with block structure. Suppose we want to compare k treatments with n blocks of data; a "block" is a group of k subjects each of whom receives a different treatment (each treatment is assigned to exactly one subject). A classical test is the Friedman test [Friedman, 1937] (see Appendix C.10 for details), and we call the interactive test as the i-Friedman test. For block $i \in [n]$ and subject $j \in [k]$, denote the outcome as Y_{ij}, the treatment assignment as A_{ij}, and the covariates as X_{ij}. The null hypothesis states that there is no difference between outcome of any two treatments conditional on covariates:

$$(Y_{ij} \mid A_{ij} = a_1, X_{ij}) \stackrel{d}{=} (Y_{ij} \mid A_{ij} = a_2, X_{ij}) \text{ for all } j \in [k] \text{ and } i \in [n] \text{ and } a_1, a_2 \in [k]; \tag{81}$$

We focus on randomized experiments, and in particular assume that

(i) the treatment assignment A_{ij} takes value $1, \ldots, k$ such that (a) $\{A_{i1}, \ldots, A_{ik}\}$ is equally likely to be any permutation of $\{1, \ldots, k\}$, and (b) the treatment assignments are independent across blocks;

(ii) the outcome of one subject Y_{i_1,j_1} is independent of the assignment of another subject A_{i_2,j_2} for any $(i_1, j_1) \neq (i_2, j_2) \in [n] \times [k]$.

Consider the vector of treatment assignments within each block i ordered by the outcomes, denoted as $\mathbf{A}_i = \big(A_{i,(1)}, \ldots, A_{i,(k)}\big)$, where $Y_{i,(1)} \geq \ldots \geq Y_{i,(k)}$. Because the assignments are independent of the outcomes under the null, we claim that

$$\mathbb{P}(\mathbf{A}_i = \mathbf{a} \mid \{Y_{ij}, X_{ij}\}_{j=1}^k) = 1/k! \quad \text{for all } \mathbf{a} \in \text{permute}([k]) \text{ and } i \in [n], \tag{82}$$

where permute$([k])$ denotes the set of all possible permutations of $[k]$. Under the alternative, the conditional distribution of $\mathbf{A}_i$ can bias to a certain ordering depending on the underlying truth.

As an example to compare three treatments ($k = 3$), suppose we wish to detect the following alternative:

$$(Y_{ij} \mid A_{ij} = 1, X_{ij}) \succeq (Y_{ij} \mid A_{ij} = 2, X_{ij}) \succeq (Y_{ij} \mid A_{ij} = 3, X_{ij}), \tag{83}$$

in which case $\mathbf{A}_i$ are more likely to be $(1, 2, 3)$. To develop an interactive test, which uses cumulative sums as test statistics, we encode the vector of assignments by a scalar (pseudo assignment $\widetilde{A}_i$) such that it takes larger value when $\mathbf{A}_i$ is more "similar" to the ideal permutation $(1, 2, 3)$. Specifically, the similarity (distance) between $\mathbf{A}_i$ and $(1, 2, 3)$ can be measured by the number of exchange operations needed to convert $\mathbf{A}_i$ to $(1, 2, 3)$. We define $\widetilde{A}_i$ as:

$$\widetilde{A}_i = \begin{cases} 1, & \text{if } \mathbf{A}_i = (1,2,3), & (84) \\ 1, & \text{if } \mathbf{A}_i = (2,1,3), & (85) \\ 1, & \text{if } \mathbf{A}_i = (1,3,2), & (86) \\ -1, & \text{if } \mathbf{A}_i = (3,1,2), & (87) \\ -1, & \text{if } \mathbf{A}_i = (2,3,1), & (88) \\ -1, & \text{if } \mathbf{A}_i = (3,2,1), & (89) \end{cases}$$

where the ordered assignments (85) and (86) need one exchange operation to be converted to $(1, 2, 3)$; (87) and (88) need two; and (89) is the opposite of the ideal permutation, which needs three exchange operations. This design of $\widetilde{A}_i$ takes binary values, but it can also take different values for each ordering of $\mathbf{A}_i$. We present above definition because it has a simple form and leads to relatively high power for a broad range of alternatives in simple simulations.

With the above transformation from a vector of assignments to a scalar $\widetilde{A}_i$ for each block i, we can view the blocks as individuals in the interactive test. That is, we use the pseudo assignment $\widetilde{A}_i$ for testing while ordering the blocks using the revealed data $\{Y_{ij}, X_{ij}\}_{i=1,j=1}^{i=n,j=k}$ and the actual assignments $\{A_{ij}\}_{j=1}^{k}$ once block i is ordered. In other words, let the pseudo assignment $\widetilde{A}_i$ be defined in (84)-(89), the pseudo outcome be the union within each block, $\widetilde{Y}_i = \{Y_{ij}\}_{j=1}^{k}$, and same for the pseudo covariates $\widetilde{X}_i = \{X_{ij}\}_{j=1}^{k}$. The i-Friedman test follows Algorithm 6 with the input data replaced by $\{\widetilde{Y}_i, \widetilde{A}_i, \widetilde{X}_i\}_{i=1}^{n}$.

4.4.3 Sample comparison in dynamic settings

We have proposed interactive tests for two/multi-sample comparison with unpaired/paired data, all of which are in the batch setting where the sample size is fixed before testing. Nonetheless, in many applications, one hopes to monitor the null of zero treatment effect as more subjects are collected, so that the experiment can stop once there is enough evidence to reject the null. In this section, we consider an sequential setting where an unknown and potentially infinite number of subjects (or pairs) arrive sequentially in a stream and introduce the sequential interactive tests.

First, we propose the seq-Wilcoxon test for two-sample comparison with unpaired data. Because the subjects arrive one by one, it is hard to order them on the fly, and we instead propose to filter the subjects to be cumulated in the sum S_t. At time $t+1$ when a new subject arrives, the analyst can interactively decide whether to add A_{t+1} to current S_t. Denote the decision by an indicator I_{t+1}, and the sum is

$$S_t = \sum_{i=1}^{t} I_i(2A_i - 1) \cdot w_i. \tag{90}$$

The available information to decide I_{t+1} and weight w_{t+1} includes the complete data information of the first t subjects and the revealed data of the $(t+1)$-th subject, denoted by the filtration:

$$\mathcal{G}_t = \sigma\left(\{Y_i, A_i, X_i, I_i\}_{i=1}^{t} \cup \{Y_{t+1}, X_{t+1}\}\right), \tag{91}$$

where the complete data $\{Y_i, A_i, X_i\}_{i=1}^{t}$ can be used for modeling and guide the decision of I_{t+1}. Under the null, we have

$$\mathbb{P}(A_i = 1 \mid I_i = 1) = 1/2, \tag{92}$$

so the sum S_{t+1} behaves as the sum of $\sum_{i=1}^{t+1} I_i$ number of coin flips (see details in Appendix C.11). The algorithm stops and rejects the null when $|S_t|$ reaches the boundary $u_{\alpha/2}(v)$ where $v = \sum_{i=1}^{t} I_i$. Equivalently, we can define a stopping time as

$$\tau := \min\left\{t \in \mathbb{N} : |S_t| > u_{\alpha/2}\left(\sum_{i=1}^{t} I_i\right)\right\}, \tag{93}$$

and the null is rejected if $\tau < \infty$. Recall in definition (44), the boundary $u_{\alpha/2}(v)$ is valid uniformly for any $v \in \mathbb{N}$, so the test has valid error control even in the sequential setting where $\sum_{i=1}^{t} I_i$ can potentially be infinite. The seq-Wilcoxon test is summarized in Algorithm 8.

In practice, to get a reasonably good model for our filtering process, we can first collect 50 subjects (say) and reveal their complete data $\{Y_i, A_i, X_i\}$ for modeling and then apply the seq-Wilcoxon test from the 51-th subject. Note that Algorithm 8 also applies to the sequential setting with paired data

Algorithm 8 Framework of the sequential Wilcoxon test (seq-Wilcoxon)

Input: First sample $\{Y_1, A_1, X_1\}$, target type-I error rate α;
Procedure: for $t = 1, 2, \ldots,$ **do**
  1. Using $\mathcal{G}_{t-1}$ to decide I_t, that is whether to include the t-th subject;
  2. Reveal A_t and update $\mathcal{F}_t$;
  if $\left|\sum_{i=1}^{t} I_i(2A_i - 1)w_i\right| > u_{\alpha/2}\left(\sum_{i=1}^{t} I_i\right)$ **then**
    reject the null and stop;
  else
    Collect the $(t+1)$-th sample $\{Y_{t+1}, A_{t+1}, X_{t+1}\}$;
  end
end

or multi-sample comparison when we replace the input data by pseudo sample $\{\widetilde{Y}_t, \widetilde{A}_t, \widetilde{X}_t\}$ defined in previous sections.

Another dynamic setting of practical interest lies in the middle of the batch setting and the sequential setting. That is what we call the mini-batch setting, where small batches of subjects arrive sequentially. Let $\mathcal{B}_t$ be the set of subjects arrive at time t. The interactive test can compute the cumulative sum S_t by progressively selecting subjects from the current pool of subjects $\bigcup_{i=1}^{t} \mathcal{B}_i$, but not necessarily ordering each subject. For example, we can order the subjects collected so far if their estimated posterior probabilities of receiving treatment are higher than a threshold, say 0.7; then, we wait for the next mini-batch to see if there are new subjects with higher posterior probabilities of receiving treatment. We remark that in both the sequential and mini-batch settings, human interaction is allowed to design and change the algorithm of filtering or ordering the subjects; and the interactive tests still have valid error control.

4.5 Summary

We have discussed two types of tests for sample comparison in a randomized trial. First is i-Wilcoxon test that incorporates the recent idea of allowing human interaction via the procedure of "masking" and "unmasking". A second type is non-interactive variants of the Wilcoxon signed-rank test with different intermediate statistics E_i that improve the power of detecting heterogeneous treatment effect. The latter offer good options when one does not want to impose any parametric model, possibly because the data is messy and we want to avoid potential power loss caused by misspecification of the working model. We recommend choosing E_i from three candidates $E_i^{R(X)}$, $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$, and $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ based on the prior beliefs or anticipated population properties of the treatment effect. In contrast, the interactive tests encourage the analyst to explore various working models before and during the testing procedure, so that the test can integrate the observed data information with prior knowledge of various types and even a human's subjective belief in a highly flexible manner.

The interactive rank test is generalized to two/multi-sample comparison with unpaired/paired data in the batch setting (with fixed sample size) or a dynamic setting (with subjects or mini-batches of subjects arrive sequentially). These extensions can be combined, following Algorithm 6 for the batch setting and Algorithm 8 for the sequential setting, where the input data is the union of pseudo samples from different settings. As an example of mixed data from several settings, Kapelner and Krieger [2014] propose a dynamic matching procedure that pairs the subjects on the fly, which generates a mixture of paired and unpaired data that arrives sequentially; and the interactive test can be applied to the generated

dataset as the matching proceeds.

An alternative perspective of unpaired two-sample comparison is causal inference. Testing the global null can be interpreted as testing whether there is *any* subject having non-zero treatment effect. A potentially more practically interesting question is to identify *which* subjects with positive treatment effect with rigorous error control. By connecting the identification of positive effect with hypothesis testing problem, we are able to propose interactive algorithms that identify positive treatment effects with FDR control, as discussed in the next chapter.

5 Interactive identification of individuals with positive treatment effect while controlling false discoveries

5.1 Introduction

Subgroup identification has been a major topic in the clinical trial community and the causal literature (see Lipkovich et al. [2017]; Loh et al. [2019]; Powers et al. [2018] and references therein). Typically, the treatment effect in the investigated population varies by the subject's gender, age, and other covariates. Identifying subjects with positive effects can help guide follow-up research and provide medication guidance. However, most existing methods do not have an error control guarantee at the level of the individual — it is possible that most subjects in the identified subgroup do not have positive effects. For example, an identified subgroup could be defined as "female subjects with age less than 40", but only 10% of them with age between 18 and 20 may truly have a positive treatment effect.

We propose to upper bound the proportion of falsely identified subjects (whose potential outcomes under treatment and control are equal, for example) via the language of hypothesis testing. As formalized in the next section, we interpret the problem of identifying subjects with positive effects as one of multiple hypothesis testing, where each subject corresponds to one null hypothesis, and the proportion of false identifications corresponds to a standard error metric, the false discovery rate (FDR). In this context, we propose algorithms with two appealing properties. First, they achieve a finite sample guarantee on FDR control. Second, our proposed algorithms identify subjects with positive effects through an *interactive* procedure according to a particular protocol — an analyst is allowed to look at an initially "masked" dataset (that is progressively "unmasked" over rounds of interaction), and combines available covariates with prior knowledge via flexible Bayesian (or black box machine learning) working models to improve power. In summary, she can combine the strengths of (automated) statistical modeling and (human-guided) scientific knowledge, all while avoiding selection bias and guaranteeing valid FDR control.

5.1.1 Problem setup

Suppose we have n subjects in the data. Each subject i has potential control outcome Y_i^C, potential treated outcome Y_i^T, and the treatment indicator A_i for $i \in [n] \equiv \{1, 2, \ldots, n\}$. Our results allow the potential outcomes to either be viewed as random variables or fixed. The treatment effect of subject i is defined as $Y_i^T - Y_i^C$ and the observed outcome is $Y_i = Y_i^C(1 - A_i) + Y_i^T A_i$ under the standard causal assumption of consistency ($Y_i = Y_i^T$ when $A_i = 1$ and $Y_i = Y_i^C$ when $A_i = 0$). Person i's covariate is denoted as X_i. This paper focuses on Bernoulli randomized experiments without interference:

(i) conditional on covariates, treatment assignments are independent coin flips:

$$\mathbb{P}[(A_1, \ldots, A_n) = (a_1, \ldots, a_n) \mid X_1, \ldots, X_n] = \prod_{i=1}^{n} \mathbb{P}(A_i = a_i) = (1/2)^n, \tag{94}$$

for any $(a_1, \ldots, a_n) \in \{0, 1\}^n$.

(ii) conditional on covariates, the outcome of one subject Y_i is independent of the assignment A_j of another subject, for any $i \neq j$:

$$Y_i \perp A_j \mid \{X_1, \ldots, X_n\} \text{ for } i \neq j, \tag{95}$$

which is implied by (94) when the potential outcomes are viewed as fixed values.

We do not assume the observed data (Y_i, A_i, X_i) are identically distributed. We consider heterogeneous effects in the sense that the distribution of $Y_i^T - Y_i^C$ varies, and aim at identifying those individuals with a positive treatment effect. (If the covariates X_i are not informative about the heterogeneity in $Y_i^T - Y_i^C$, our identification power could be low, and this is to be expected.) To formalize and frame the problem in terms of multiple hypothesis testing, we first define the null hypothesis for subject i as having zero treatment effect:

$$H_{0i}^{\text{zero}} : (Y_i^T \mid X_i) \stackrel{d}{=} (Y_i^C \mid X_i), \tag{96}$$

or equivalently, $H_{0i}^{\text{zero}} : (Y_i \mid A_i = 1, X_i) \stackrel{d}{=} (Y_i \mid A_i = 0, X_i)$. Alternatively, we can treat the potential outcomes and covariates as fixed, and frame the null hypothesis as

$$H_{0i}^{\text{zero}} : Y_i^T = Y_i^C. \tag{97}$$

A last, hybrid, version (e.g., Howard and Pimentel [2020]) is to treat the two potential outcomes as random with joint distribution $(Y_i^T, Y_i^C) \mid X_i \sim P_i$, and the null posits

$$H_{0i}^{\text{zero}} : Y_i^T = Y_i^C \text{ almost surely-}P_i, \tag{98}$$

meaning that P_i is supported on $\{(x, y) : x = y\}$. Our work handles any interpretation.

In later sections, we describe an extension where we relax the null as those with a nonpositive effect, defined by stochastic dominance $(Y_i^T \mid X_i) \preceq (Y_i^C \mid X_i)$, meaning that $\mathbb{P}(Y_i^T \le y \mid X_i) \le \mathbb{P}(Y_i^C \le y \mid X_i)$, or simply $Y_i^T \le Y_i^C$ if the potential outcomes are fixed.

Our algorithms control the error of falsely identifying subjects whose null hypothesis is true (i.e., having zero effect), and aim at correctly identifying subjects with positive effects. Let $\succ$ denote stochastic dominance, as above. We say a subject has a *positive effect* if

$$(Y_i^T \mid X_i) \succ (Y_i^C \mid X_i). \tag{99}$$

When treating the potential outcomes and covariates as fixed, we simply write $Y_i^T > Y_i^C$.

The output of our proposed algorithms is a set of identified subjects, denoted as $\mathcal{R}$, with a guarantee that the expected proportion of falsely identified subjects is upper bounded. Specifically, denote the set of subjects that are true nulls as $\mathcal{H}_0 := \{i \in [n] : H_{0i}^{\text{zero}} \text{ is true}\}$. Then the number of false identifications is $|\mathcal{R} \cap \mathcal{H}_0|$. The expected proportion of false identifications is a standard error metric, the false discovery rate (FDR):

$$\text{FDR} := \mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{|\mathcal{R}|, 1\}}\right]. \tag{100}$$

Given $\alpha \in (0, 1)$, we propose algorithms that guarantee FDR $\le \alpha$, and have reasonably high *power*, which is defined as the expected proportion of correctly identified subjects:

$$\text{power} := \mathbb{E}\left[\frac{|\mathcal{R} \cap \text{Pos}|}{\max\{|\text{Pos}|, 1\}}\right],$$

where $\text{Pos} := \left\{i : (Y_i^T \mid X_i) \succ (Y_i^C \mid X_i)\right\}$ or $\text{Pos} := \{i : Y_i^T > Y_i^C\}$ is the set of subjects with positive effects.

5.1.2 Related work: error control in subgroup identification

We note that our problem setup is not exactly the same as most work in subgroup identification, such as Foster et al. [2011]; Imai and Ratkovic [2013]; Zhao et al. [2012]. The identified subgroups are usually defined by functions of covariates, rather than a subset of the investigated subjects as in our paper. While defining the subgroup by a function of covariates makes it easy to generalize the finding in the investigated sample to a larger population, it does not seem straightforward to nonasymptotically control the error of false identifications using the former definition, which is a major distinction between previous studies and our work. Most existing work does not have an error control guarantee (see an overview in Lipkovich et al. [2017], Table XV), except a few discussing error control on the level of subgroups as opposed to the level of individuals in our paper. The difference between FDR control at a subgroup level and at an individual level is detailed below.

Subgroup FDR control. Gu and Shen [2018]; Karmakar et al. [2018]; Xie et al. [2018] discuss FDR control at a subgroup level, where the latter two have little discussion on incorporating continuous covariates and require parametric assumptions on the outcomes. Thus, we follow the setup in Karmakar et al. [2018] to compare the FDR control at a subgroup level (in their paper) and individual level (in our paper). Let the subgroups be non-overlapping sets $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$. The null hypothesis for a subgroup $\mathcal{G}_g$ is defined as:

$$\mathcal{H}_{0g} : H_{0i}^{\text{zero}} \text{ is true for all } i \in \mathcal{G}_g,$$

or equivalently, $\mathcal{H}_{0g} : \mathcal{G}_g \subseteq \mathcal{H}_0$ (recall $\mathcal{H}_0$ is the set of subjects with zero effect). Let D_g be the $0/1$-valued indicator function for whether $\mathcal{H}_{0g}$ is identified or not. The FDR at a subgroup level is defined as the expected proportion of falsely identified subgroups:

$$\text{FDR}^{\text{subgroup}} := \mathbb{E}\left[\frac{|\{g \in [G] : \mathcal{G}_g \subseteq \mathcal{H}_0, D_g = 1|}{\max\{|\{g \in [G] : D_g = 1|, 1\}}\right], \tag{101}$$

which collapses to the FDR at an individual level as defined in (100) when each subgroup has exactly one subject. Although our interactive procedure is designed for FDR control at an individual level, we propose extensions to FDR control at a subgroup level in Section 5.9. As a brief summary, Karmakar et al. [2018] propose to control $\text{FDR}^{\text{subgroup}}$ by constructing a p-value for each subgroup and apply the classical BH method [Benjamini and Hochberg, 1995]. While their method has many orthogonal benefits (e.g., handling observational studies), it is not trivially applicable to control FDR at an individual level, because their p-values would only take value $1/2$ or 1 when each subgroup has exactly one subject, leading to zero identification power following the BH procedure. In other cases where subgroups have more than one subject, the above error control does not imply whether subjects within a rejected subgroup are mostly non-nulls, or if many are nulls with zero effect. Our paper appears to be the first to propose methods for identifying subjects having positive effects with (finite sample) FDR control.

Other related error control at a subgroup level. Cai et al. [2011] and Athey and Imbens [2016] develop confidence intervals for the averaged treatment effect within subgroups, where the former assumes the size of each subgroup to be large, and the latter requires a separate sample for inference. These intervals can potentially be used to generate a p-value for each subgroup and control FDR at a subgroup level via standard multiple testing procedures, but no explicit discussion is provided. Lipkovich et al. [2011], Lipkovich and Dmitrienko [2014], Sivaganesan et al. [2011] and Berger et al. [2014] propose methods with control on a different error metric: the global type-I error, which is the probability of identifying any subgroup when no subject has nonzero treatment effect (i.e., H_{0i}^{zero} is true for all

subjects). Our FDR control guarantee implies valid global type-I error, and FDR control is more informative on the correctness of the identified subgroups/subjects when there exist subjects having nonzero effects.

5.1.3 An overview of our procedure

As discussed, it appears to be new and practically interesting to provide FDR control guarantees at an individual level. Another merit of our proposed method is that it allows a human analyst and an algorithm to interact, in order to better accomplish the goal.

Interactive testing is a recent idea that emerged in response to the growing practical needs of allowing human interaction in the process of data analysis. In practice, analysts tend to try several methods or models on the same dataset until the results are satisfying, but this violates the validity of standard testing methods (e.g., invalid FDR control). In our context of identifying positive effects, the appealing advantages of an interactive test include that (a) an analyst is allowed to use (partial) data, together with prior knowledge, to design a strategy of selecting subjects potentially having positive effects, and (b) it is a multi-step iterative procedure during which the analyst can monitor performance of the current strategy and make adjustments at any step (at the cost of not altering earlier steps). Despite the flexibility of an analyst to design and alter the algorithm using (partial) data, our proposed procedure always maintains valid FDR control. We name our proposed algorithm as I^3 (I-cube), for interactive identification of individual treatment effects.

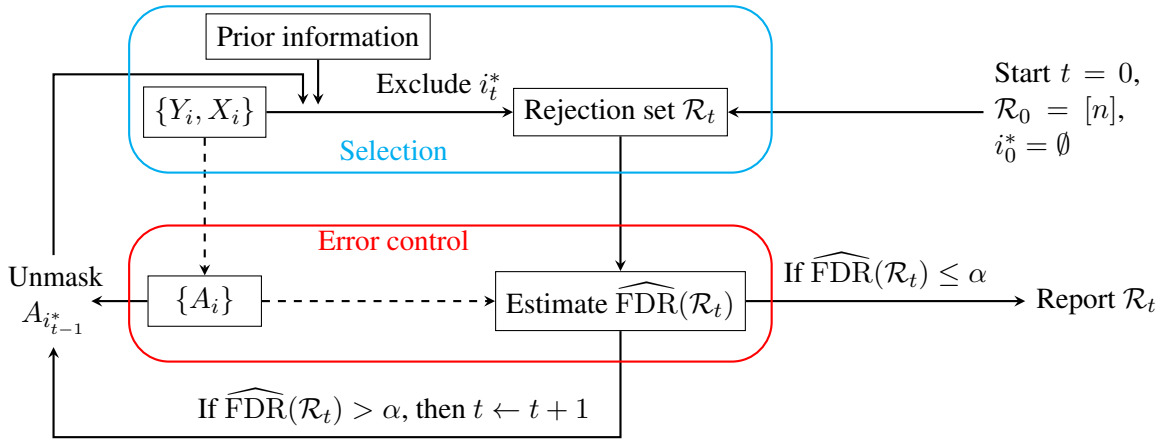


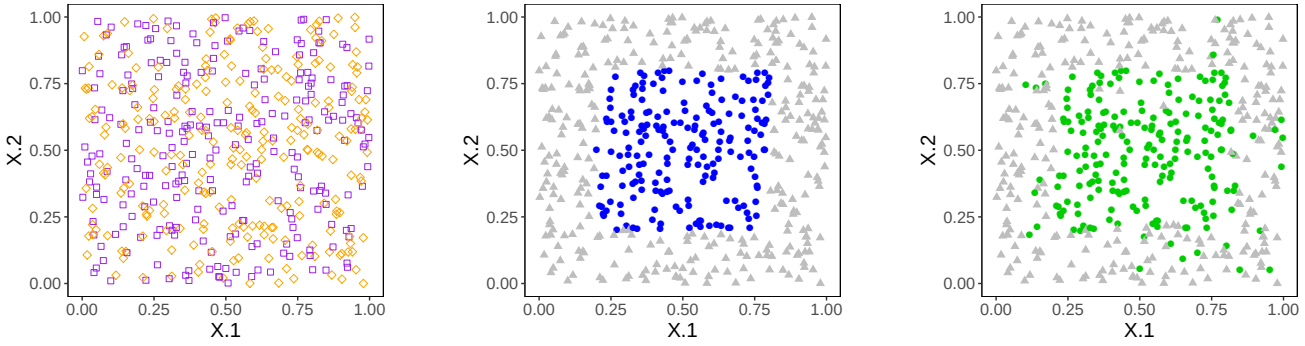
Figure 32: A schematic of the I^3 algorithm. All treatment assignments are initially kept hidden: only $(Y_i, X_i)_{i\in[n]}$ are revealed to the analyst, while all $\{A_i\}$ remain 'masked'. The initial candidate rejection set is $\mathcal{R}_0 = [n]$ (thus no subject is excluded initially and $i_0^* = \emptyset$). The false discovery proportion $\widehat{\mathrm{FDR}}$ of the current candidate set $\mathcal{R}_t$ is estimated by the algorithm (dashed lines), and reported to the analyst. If $\widehat{\mathrm{FDR}}(\mathcal{R}_t) > \alpha$, the analyst chooses a subject i_t^* to remove it from the proposed rejection set $\mathcal{R}_t = \mathcal{R}_{t-1}\backslash\{i_t^*\}$, whose assignment $A_{t_t^*}$ is then 'unmasked' (revealed). Importantly, using any available prior information, covariates and working model, the analyst can choose subject i_t^* and shrink $\mathcal{R}_t$ in any manner. This process continues until $\widehat{\mathrm{FDR}}(\mathcal{R}_t) \le \alpha$ (or $\mathcal{R}_t = \emptyset$).

The core idea that enables human interaction is to separate the information used for selecting subjects with positive effects and that for error control, via "masking and unmasking" (Figure 32). In short,

masking means we hide $\{A_i\}_{i=1}^n$ from the analyst. The algorithm alternates between two steps — selection and error control — until a simple stopping criterion introduced later is reached.

1. **Selection**. Consider a set of candidate subjects to be identified as having a positive effect (whose null to be rejected), denoted as rejection set $\mathcal{R}_t$ for iteration t. We start with all the subjects included, $\mathcal{R}_0 = [n]$. At each iteration, the analyst excludes possible nulls (i.e., subjects that are unlikely to have positive effects) from the previous $\mathcal{R}_{t-1}$, using all the available information (outcomes Y_i and covariates X_i for all subjects $i \in [n]$, and progressively unmasked A_i from the step of error control, and possible prior information). Note that our method does not automatically use prior information and the revealed data. The analyst is free to use any black-box prediction algorithm that uses the available information, and evaluates the subjects possibly using an estimated probability of having a positive treatment effect. This step is where a human is allowed to incorporate her subjective choices.
2. **Error control (and unmasking)**. The algorithm uses the complete data $\{Y_i, A_i, X_i\}$ to estimate FDR of the current candidate rejection set $\widehat{\mathrm{FDR}}(\mathcal{R}_t)$, as a feedback to the analyst. If the estimated FDR is above the target level $\widehat{\mathrm{FDR}}(\mathcal{R}_t) > \alpha$, the analyst goes back to the step of selection, along with additional information: the excluded subjects ($i \notin \mathcal{R}_t$) have their A_i unmasked (revealed), which could improve her understanding of the data and guide her choices in the next selection step.

The algorithms we propose in the main paper build on and modify the above procedure to achieve reasonably high power and develop various extensions. An illustration of the identifications made by the Crossfit-I^3 (our central algorithm) is in Figure 33.



(a) Subjects in treated group and control group separated by two shapes and colors.

(b) Blue round dots represent subjects with true positive effect (unknown ground truth).

(c) Green round dots represent subjects identified by the Crossfit-I^3.

Figure 33: An illustrative example with 1000 subjects, each has two covariates that are uniform in $[0, 1]$. The Crossfit-I^3 identifies most subjects with positive effects, although about half of them did not receive treatment.

Outline. The rest of the paper is organized as follows. In Section 5.2, we describe an interactive algorithm wrapped by a cross-fitting framework, which identifies subjects with positive effects with FDR control. We evaluate our proposed algorithm numerically in Section 5.3, and provide theoretical power analysis in simple settings in Section 5.4. We point out several extensions of the proposed algorithm from Section 5.5 to 5.9, and present a prototypical application to a real data set in Section 5.10. Section 5.11 concludes the paper with a discussion on the potential of our proposed interactive procedures.

5.2 An interactive algorithm with FDR control

To enable us to effectively infer the treatment effect, we use the following *working model*:

$$Y_i^C = f(X_i) + U_i \text{ and } Y_i^T = \Delta(X_i) + f(X_i) + U_i, \tag{102}$$

where U_i is zero-mean noise (unexplained variance) that is independent of A_i. When working with such a model, we effectively want to identify subjects with a positive treatment effect $\Delta(X_i)$. Importantly, model (102) needs not be correctly specified or accurately reflect reality in order for the algorithms in this paper to have a valid FDR control (but the more ill-specified or inaccurate the model is, the more power may be hurt).

To identify subjects with positive effects, we first introduce an estimator of the treatment effect $\Delta(X_i)$ following the working model (102). Denote the expected outcome given the covariates as $m(X_i) := \mathbb{E}(Y_i \mid X_i)$, and let $\widehat{m}(X_i)$ be an arbitrary estimator of $m(X_i)$ using the outcomes and covariates $\{Y_i, X_i\}_{i=1}^n$. Define the *residual* as $E_i := Y_i - \widehat{m}(X_i)$, and an estimator of $\Delta(X_i)$ is

$$\widehat{\Delta}_i := 4(A_i - 1/2) \cdot E_i, \tag{103}$$

which, under randomized experiments, is equivalent to the nonparametric estimator of the conditional treatment average effect $\mathbb{E}(Y_i^T \mid X_i) - \mathbb{E}(Y_i^C \mid X_i)$ in several recent papers [Kennedy, 2020; Nie and Wager, 2020], and can be traced back to the semiparametrics literature with Robinson [1988]. A critical property of $\widehat{\Delta}_i$ that later leads to FDR control is that[10]

$$\mathbb{P}(\widehat{\Delta}_i > 0 \mid \{Y_j, X_j, E_j\}_{j=1}^n) \le 1/2, \tag{104}$$

under H_{0i}^{zero} (for any definition in (96), (97), or (98)), because H_{0i}^{zero} implies $A_i \perp\!\!\!\perp \{Y_i, X_i\}$ and $\mathbb{P}(A_i - 1/2 > 0) = 1/2$. Recall in Figure 32, the treatment assignments A_i are hidden from the analyst in the selection step, which is reflected in (104) as A_i omitted from the condition. The above property indicates that the estimated effect $\widehat{\Delta}_i$ is no more likely to be positive than negative if the selected subject has zero effect, regardless of how the analyst decides which subject to select. Therefore, the sign of $\widehat{\Delta}_i$ can be used to estimate the number of false identifications and achieve FDR control, which we elaborate next.

5.2.1 An interactive algorithm with valid FDR control

This section presents the I^3 with valid FDR control. We introduce a modification based on cross-fitting that improves identification power in the next section.

The I^3 proceeds as progressively shrinking a candidate rejection set $\mathcal{R}_t$ at iteration t,

$$[n] = \mathcal{R}_0 \supseteq \mathcal{R}_1 \supseteq \ldots \supseteq \mathcal{R}_n = \emptyset,$$

where recall $[n]$ denotes the set of all subjects. We assume without loss of generality that one subject is excluded in each step. Denote the subject excluded at iteration t as i_t^*. The choice of i_t^* can use the information available to the analyst before iteration t, formally defined as a filtration (sequence of nested σ-fields):

$$\mathcal{F}_{t-1} = \sigma\left(\{Y_j, X_j\}_{j\in\mathcal{R}_{t-1}}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}}, \sum_{j\in\mathcal{R}_{t-1}} \mathbb{1}\{\widehat{\Delta}_j > 0\}\right), \tag{105}$$

[10]Note that property (104) uses the fact that outcome estimator $\widehat{m}(X_i)$ is independent of A_i, so it is important that the estimation of $\widehat{m}$ does not use the assignments $\{A_i\}_{i=1}^n$; however, it should not affect the estimation much because $m(X_i) \equiv \mathbb{E}(Y_i \mid X_i)$ is not a function of A_i.

where we unmask (reveal) the treatment assignments A_j for subjects excluded from $\mathcal{R}_{t-1}$, and the sum $\sum_{i\in\mathcal{R}_{t-1}} \mathbb{1}\{\widehat{\Delta}_i > 0\}$ is mainly used for FDR estimation as we describe later. The above available information include arbitrary functions of the revealed data, such as the residuals $\{E_j\}_{j=1}^n$ defined above equation (103). Similar to property (104), for each candidate subject $i \in \mathcal{R}_{t-1}$, we have

$$\mathbb{P}(\widehat{\Delta}_i > 0 \mid \{Y_j, X_j\}_{j\in\mathcal{R}_{t-1}}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}}) \le 1/2, \tag{106}$$

which ensures the FDR control as we explain next.

To control FDR, the number of false identifications is estimated by (106). The idea is to partition the candidate rejection set $\mathcal{R}_t$ into $\mathcal{R}_t^+$ and $\mathcal{R}_t^-$ by the sign of $\widehat{\Delta}_i$:

$$\mathcal{R}_t^- := \{i \in \mathcal{R}_t : \widehat{\Delta}_i \le 0\}, \quad \mathcal{R}_t^+ := \{i \in \mathcal{R}_t : \widehat{\Delta}_i > 0\}.$$

Notice that our proposed procedure only identifies the subjects whose estimated effect is positive, i.e., those in $\mathcal{R}_t^+$. Thus, the FDR is $\mathbb{E}\left[\frac{|\mathcal{R}_t^+\cap\mathcal{H}_0|}{\max\{|\mathcal{R}_t^+|,1\}}\right]$ by definition, where recall $\mathcal{H}_0$ is the set of true nulls. Intuitively, the number of false identifications $|\mathcal{R}_t^+ \cap \mathcal{H}_0|$ can be approximately upper bounded by $|\mathcal{R}_t^- \cap \mathcal{H}_0|$, since the number of positive signs should be no larger than the number of negative signs for the falsely identified nulls, according to property (104). Note that the set of true nulls $\mathcal{H}_0$ is unknown, so we use $|\mathcal{R}_t^-|$ to upper bound $|\mathcal{R}_t^- \cap \mathcal{H}_0|$, and propose an estimator of FDR for the candidate rejection set $\mathcal{R}_t$:

$$\widehat{\mathrm{FDR}}(\mathcal{R}_t) = \frac{|\mathcal{R}_t^-| + 1}{\max\{|\mathcal{R}_t^+|, 1\}}. \tag{107}$$

Overall, the I^3 shrinks $\mathcal{R}_t$ until time $\tau := \inf\{t : \widehat{\mathrm{FDR}}(\mathcal{R}_t) \le \alpha\}$ and identifies only the subjects in $\mathcal{R}_\tau^+$, as summarized in Algorithm 9. We state the FDR control of I^3 in Theorem 10 and the proof can be found in Appendix D.2.1.

Theorem 10. *In a randomized experiment with assumptions (94) and (95), and for any analyst that updates their working model(s) at any iteration t using the information in $\mathcal{F}_{t-1}$, the set R_τ^+ rejected by the* I^3 *algorithm has FDR controlled at level α, meaning that*

$$\mathbb{E}\left[\frac{|\mathcal{R}_\tau^+ \cap \mathcal{H}_0|}{\max\{|\mathcal{R}_\tau^+|, 1\}}\right] \le \alpha,$$

for any definition of the null hypothesis (96), (97) or (98). For the last definition, FDR control also holds conditional on the covariates and potential outcomes.

Consider a simple case where model (102) is accurate for every subject with a constant treatment effect $\Delta(X_i) = \delta > 0$. If δ is larger than the maximum noise, we have $\mathcal{R}_0^+ = [n]$, and the algorithm can stop at the very first step identifying all subjects. At the other extreme, if the effect δ is too small, the algorithm may also return an empty set, and this makes sense because while small *average* treatment effects can be learned using a large population, larger treatment effects are needed for *individual-level* identification.

Related work. Testing procedures that allow human interaction are first proposed by Lei and Fithian [2018] and Lei et al. [2020] for the problem of FDR control in multiple testing, followed by several

Algorithm 9 The I^3 (interactive identification of individual treatment effect) procedure.

Initial state: Explorer (E) knows covariates and outcomes $\{X_i, Y_i\}_{i=1}^n$.
Oracle (O) knows the treatment assignments $\{A_i\}_{i=1}^n$.
Target FDR level α is public knowledge.
Initial exchange: Both players initialize $\mathcal{R}_0 = [n]$ and set $t = 1$.
1. E builds a prediction model $\widehat{m}$ from X_i to Y_i.
2. E informs O about residuals $E_i \equiv Y_i - \widehat{m}(X_i)$.
3. O estimates the treatment effect as $\widehat{\Delta}_i \equiv 4(A_i - 1/2)E_i$.
4. O then divides $\mathcal{R}_t$ into $\mathcal{R}_t^- := \{i \in \mathcal{R}_t : \widehat{\Delta}_i \leq 0\}$ and $\mathcal{R}_t^+ := \{i \in \mathcal{R}_t : \widehat{\Delta}_i > 0\}$.
5. O reveals only $|\mathcal{R}_t^+|$ to E (who infers $|\mathcal{R}_t^-|$).
Repeated interaction: 6. E checks if $\widehat{\mathrm{FDR}}(\mathcal{R}_t) \equiv \frac{|\mathcal{R}_t^-|+1}{\max\{|\mathcal{R}_t^+|,1\}} \leq \alpha$.
7. If yes, E sets $\tau = t$, reports $\mathcal{R}_\tau^+$ and exits.
8. Else, E picks any $i_t^* \in \mathcal{R}_{t-1}$ using everything E currently knows.
(E tries to pick an i_t^* that E thinks is null, i.e. E hopes that $\widehat{\Delta}_{i_t^*} \leq 0$.)
9. O reveals $A_{i_t^*}$ to E, who also infers $\widehat{\Delta}_{i_t^*}$ and its sign.
10. E updates $\mathcal{R}_{t+1} = \mathcal{R}_t \backslash \{i_t^*\}$, and also $|\mathcal{R}_{t+1}^+|$ and $|\mathcal{R}_{t+1}^-|$.
11. Increment t and go back to Step 6.

works for other error metrics [Duan et al., 2019, 2020a]. These papers focus on generic multiple testing problems, which operate on the p-values and ignore the process of generating p-values from data. In contrast, Duan et al. [2020b] applies the idea of interactive testing to observed data, to which our paper relates most. Both works propose tests for treatment effect, and the difference is that Duan et al. [2020b] test whether *any* subject has nonzero effect with type-I error control, whereas our proposed algorithm aims at identifying subjects having positive effects with FDR control. While the former may appear in an exploratory analysis to see whether the treatment has any effect on any person, the latter is useful to characterize the population where the treatment has an effect.

We end the section with a remark. In step 8 of Algorithm 9, we hope to exclude subjects that are unlikely to have positive effects, based on the revealed data in $\mathcal{F}_{t-1}$. In other words, we should guess the sign of treatment effect $\widehat{\Delta}_i$, which depends on both the revealed data $\{Y_i, X_i\}$ and the hidden assignment A_i. However, notice that at the first iteration, we may learn/guess the opposite signs for all the subjects; when all assignments $\{A_i\}_{i=1}^n$ are hidden at $t = 1$, the likelihood of $\{A_i\}_{i=1}^n$ being the true values (leading to all correct signs for $\widehat{\Delta}_i$) is the same as the likelihood of all opposite values (leading to all opposite signs for $\widehat{\Delta}_i$), no matter what working model we use. Consequently, the subjects with large positive effects could be guessed as having large negative effects, causing them to be excluded from the rejection set. To improve power, we propose to wrap around the I^3 by a cross-fitting framework as described in the next section.

5.2.2 Improving stability and power with Crossfit-I^3

Cross-fitting refers to the idea of splitting the samples into two halves. We perform the I^3 on each half separately, so that for each half, the complete data (including the assignments) of the other half is revealed to the analyst to help infer the sign of treatment effect, addressing the issue of learning the opposite signs and improving the identification power.

Specifically, split the subjects randomly into two sets of equal size, denoted as $\mathcal{I}$ and $\mathcal{II}$ where $\mathcal{I} \cup \mathcal{II} = [n]$. The I^3 (Algorithm 9) is implemented on each set separately: at the start of I^3 on set $\mathcal{I}$,

the analyst has access to the complete data for all subjects in set $\mathcal{II}$, and tries to identify subjects with positive effects in set $\mathcal{I}$ with FDR control at level $\alpha/2$; similar is the I^3 on set $\mathcal{II}$. Mathematically, let the candidate rejection set of implementing the I^3 on set $\mathcal{I}$ be $\mathcal{R}_t(\mathcal{I})$, where the initial set is $\mathcal{R}_0(\mathcal{I}) = \mathcal{I}$. The available information at iteration t is defined as:

$$\mathcal{F}_{t-1}(\mathcal{I}) = \sigma\left(\{Y_i, X_i\}_{i\in\mathcal{R}_{t-1}(\mathcal{I})}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}(\mathcal{I})}, \sum_{i\in\mathcal{R}_{t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i > 0\}\right), \tag{108}$$

which includes the complete data $\{Y_j, A_j, X_j\}$ for $j \in \mathcal{II}$ at any iteration $t \geq 0$ [11]. Similarly, we define $\mathcal{R}_t(\mathcal{II})$ and $\mathcal{F}_{t-1}(\mathcal{II})$ for the I^3 implemented on set $\mathcal{II}$. The final rejection set is the union of rejections in $\mathcal{I}$ and $\mathcal{II}$ (see Algorithm 10). We call this algorithm the Crossfit-I^3.

Algorithm 10 The Crossfit-I^3.

Input: Covariates, outcomes, treatment assignments $\{Y_i, A_i, X_i\}_{i=1}^n$, target level α;
Procedure:
1. Randomly split the sample into two subsets of equal size, denoted as $\mathcal{I}$ and $\mathcal{II}$;
2. Implement Algorithm 9 at level $\alpha/2$, where E initially knows $\{Y_k, X_k\}_{k=1}^n \cup \{A_j\}_{j\in\mathcal{II}}$ and sets $\mathcal{R}_0(\mathcal{I}) = \mathcal{I}$, getting a rejection set $\mathcal{R}_\tau^+(\mathcal{I}) \subseteq \mathcal{I}$;
3. Implement Algorithm 9 at level $\alpha/2$, where E initially knows $\{Y_k, X_k\}_{k=1}^n \cup \{A_j\}_{j\in\mathcal{I}}$ and sets $\mathcal{R}_0(\mathcal{II}) = \mathcal{II}$, getting a rejection set $\mathcal{R}_\tau^+(\mathcal{II}) \subseteq \mathcal{II}$;
4. Combine two rejection sets as the final rejection set, $\mathcal{R}_\tau^+ = \mathcal{R}_\tau^+(\mathcal{I}) \cup \mathcal{R}_\tau^+(\mathcal{II})$.

As long as the I^3 on two sets do not exchange information, Algorithm 10 has a valid FDR control (see the proof in Appendix D.2.2).

Theorem 11. *Under assumption (94) and (95) of randomized experiments, $\mathcal{R}_\tau^+$ rejected by the Crossfit-I^3 has FDR controlled at level α for any of the null hypotheses (96), (97) or (98). For the last case, FDR control also holds conditional on the covariates and potential outcomes.*

In addition to addressing the issue of learning the opposite $\widehat{\Delta}_i$ in the original I^3, another benefit of using the crossing-fitting framework is that with the complete data revealed for at least half of the sample, the analyst does not have to deal with the problem of inferring missing data (the assignment A_i), which probably needs some parametric probabilistic modeling and the EM algorithm. Instead, because the assignments are revealed for subjects not in the candidate rejection set (at least half of the sample), their signs of $\widehat{\Delta}_j$ can be correctly calculated and used as "training data". The analyst can then employ a black-box prediction model, such as a random forest, to predict the signs of $\widehat{\Delta}_i$ for the subjects whose assignments are masked (hidden). As an example, we propose an automated strategy as follows to select a subject at step 8 in Algorithm 9.

[11] For notational clarity, we use i to denote candidate subjects $i \in \mathcal{R}_t(\mathcal{I})$, and j for non-candidate subjects $j \notin \mathcal{R}_t(\mathcal{I})$, while k is used to index all subjects $k \in [n]$.

Algorithm 11 An automated heuristic to select i_t^* in the Crossfit-I^3.

Input: Current rejection set $\mathcal{R}_{t-1}(\mathcal{I})$, and available information for selection $\mathcal{F}_{t-1}(\mathcal{I})$;
Procedure:
1. Train a random forest classifier where the label is $\mathrm{sign}(\widehat{\Delta}_j)$ and the predictors are Y_j, X_j and the residuals E_j, using non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;
2. Estimate the probability of $\widehat{\Delta}_i$ being positive as $\widehat{p}(i,t)$ for subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$;
3. Select $i_t^* = \mathrm{argmin}\{\widehat{p}(i,t) : i \in \mathcal{R}_{t-1}(\mathcal{I})\}$.

We remark that in practice, the analyst can interactively change the prediction model, such as exploring parametric models to see which fits the data better. In principle, the analyst can perform any exploratory analysis on data in $\mathcal{F}_{t-1}(\mathcal{I})$ to decide a heuristic or score for selecting subject i_t^*; and the FDR control is valid as long as she does not use the assignments A_i for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$. For computation efficiency, we usually update the prediction models (or their parameters) once every 100 iterations (say).

To summarize, the Crossfit-I^3 described in Algorithm 10 involves two rounds of the I^3 (Algorithm 9), where step 8 of selecting a subject is allowed to involve human interaction; alternatively, step 8 can be an automated heuristic as presented in Algorithm 11. Recall the illustrative example in Figure 33, where we implement the Crossfit-I^3 with the above automated strategy to select subjects. Each subject, recorded with two covariates in $[0,1]$, has a constant positive effect when the covariate values are around 0.5 (see Figure 33b). Even though half of the subjects with positive effects do not receive treatment (hence we do not know their potential treated outcomes), the Crossfit-I^3 correctly identifies most of them (see Figure 33c). Next, we demonstrate through repeated numerical experiments and theoretical analysis that the Crossfit-I^3 has reasonably high power.

5.3 Numerical experiments

To assess our proposed procedure, we first describe a baseline method, which calculates a p-value for each subject under the assumption of linear models, and applies the classical BH method [Benjamini and Hochberg, 1995]. We call this method the linear-BH procedure.

5.3.1 A baseline: the BH procedure under linear assumptions

For the treated group and control group, we first separately learn a linear model to predict Y_i using X_i, denoted as $\widehat{l}^T$ and $\widehat{l}^C$. By imputing the unobserved potential outcomes, we get estimators of the potential outcomes $\widetilde{Y}_i^T = Y_i \mathbb{1}\{A_i = 1\} + \widehat{l}^T(X_i)\mathbb{1}\{A_i = 0\}$ and $\widetilde{Y}_i^C = \widehat{l}^C(X_i)\mathbb{1}\{A_i = 1\} + Y_i \mathbb{1}\{A_i = 0\}$, and the treatment effect for subject i can be estimated as $\widehat{\Delta}_i^{\mathrm{BH}} := \widetilde{Y}_i^T - \widetilde{Y}_i^C$. If the potential outcomes are linear functions of covariates with standard Gaussian noises (which we refer to as the linear assumption), the estimated treatment effect asymptotically follows a Gaussian distribution. For each subject $i \in [n]$, we calculate a p-value for the zero-effect null (96) as

$$P_i = 1 - \Phi\left(\widehat{\Delta}_i^{\mathrm{BH}} \Big/ \sqrt{\widehat{\mathrm{Var}}(\widehat{\Delta}_i^{\mathrm{BH}})}\right), \tag{109}$$

where the estimated variance is $\widehat{\mathrm{Var}}(\widehat{\Delta}_i^{\mathrm{BH}}) = \widehat{\mathrm{Var}}(\widetilde{Y}_i^T) + \widehat{\mathrm{Var}}(\widetilde{Y}_i^C)$, and Φ denotes the CDF of a standard Gaussian. To identify subjects having positive effects with FDR control, we apply the BH procedure to the above p-values. Notice that the error control would not hold when the linear assumption is violated (see Appendix D.2.4 for the formal FDR control guarantee).

5.3.2 Numerical experiments and power comparison

We run a simulation with 500 subjects ($n = 500$). Each subject is recorded with two binary attributes (eg. female/male and senior/junior) and one continue attribute (eg. body weight), denoted as a vector $X_i = (X_i(1), X_i(2), X_i(3)) \in \{0, 1\}^2 \times \mathbb{R}$. Among n subjects, the binary attributes are marginally balanced, and the subpopulation with $X_i(1) = 1$ and $X_i(2) = 1$ is of size 30. The continuous attribute is independent of the binary ones and follows the distribution of a standard Gaussian.

The outcomes are simulated as a function of the covariates X_i and the assignment A_i following the generating model (102). Recall that we previously used model (102) as a working model, which is not required to be correctly specified. Here, we generate data from such a model in simulation for a clear evaluation of the considered methods. We specify the noise U_i as a standard Gaussian, and the expected control outcome as $f(X_i) = 5(X_i(1) + X_i(2) + X_i(3))$, and the treatment effect as

$$\Delta(X_i) = S_\Delta \cdot [5X_i^3(3)\mathbb{1}\{X_i(3) > 1\} - X_i(1)/2], \tag{110}$$

where $S_\Delta > 0$ encodes the scale of the treatment effect. In this setup, around 15% subjects have positive treatment effects with a large scale, and 43% subjects have a mild negative effect [12]. More experiments can be found in Appendix D.5.1.

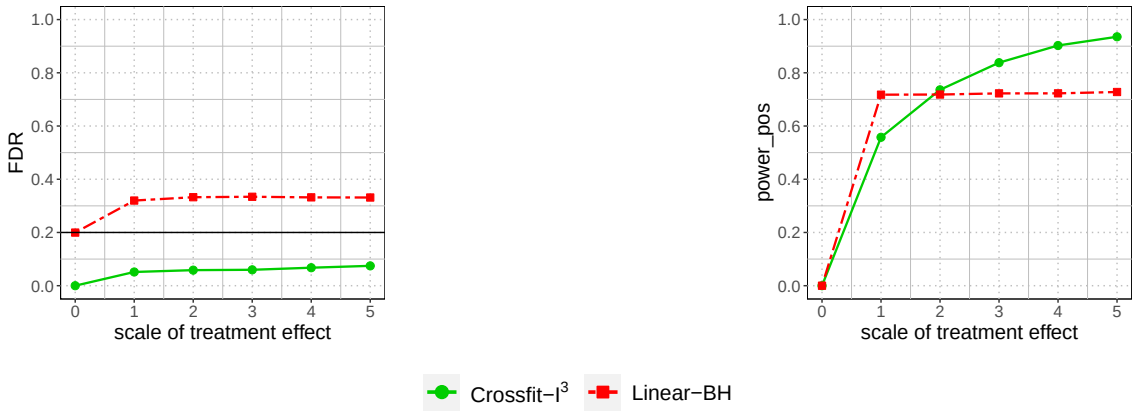


Figure 34: FDR (left) and power (right) of the Crossfit-I^3 compared with the linear-BH procedure, with the treatment effect specified as model (110) and the scale S_Δ varying in $\{0, 1, 2, 3, 4, 5\}$. The FDR control level is 0.2, marked by a horizontal line in error control plots. For all plots in this paper, the FDR and power are averaged over 500 repetitions. The linear-BH procedure does not have valid FDR control because the treatment effect is nonlinear, whereas the Crossfit-I^3 controls FDR and can achieve high power.

For the Crossfit-I^3, we use random forests (with default parameters in `R`) to compute $\widehat{m}$, and use the automated selection strategy Algorithm 11 to select a subject at step 8 in Algorithm 9. The linear-BH procedure results in a substantially higher FDR than desired because the linear assumption does not hold in the underlying truth (110) (see Figure 34), whereas our proposed Crossfit-I^3 controls FDR at the target level as expected. At the same time, the Crossfit-I^3 appears to have comparable power as the linear-BH procedure to correctly identify subjects with true positive effects.

[12] R code to fully reproduce all plots in the paper are available at `https://github.com/duanby/I-cube`.

5.4 Asymptotic power analysis in simple settings

In addition to the numerical experiments, we provide a theoretical power analysis in some simple cases to understand the advantages and limitations of our proposed Crossfit-I^3.

First, consider the case without covariates. Our analysis is inspired by the work of Arias-Castro and Chen [2017]; Rabinovich et al. [2020], who study the power of methods with FDR control under a sparse Gaussian sequence model. Let there be n hypotheses, each associated with a test statistic V_i for $i = 1, \ldots, n$. They consider a class of methods called *threshold procedures* such that the final rejection set $\mathcal{R}$ is in the form $\mathcal{R} = \{i : V_i \geq \tau(V_1, \ldots, V_n)\}$, for some threshold $\tau(V_1, \ldots, V_n)$; they discuss two types of thresholds; see Appendix D.4 for details of their results. An example of the threshold procedure is the BH procedure. Our proposed I^3 can also be simplified to a threshold procedure when using an automated selection strategy at step 8 of Algorithm 9: at each iteration, we exclude the subject with the smallest absolute value of the estimated treatment effect $|\widehat{\Delta}_i|$ (note that this strategy satisfies our requirement of not using assignments since $|\widehat{\Delta}_i| = |4(A_i - 1/2)(Y_i - \widehat{m}(X_i))| = 2|Y_i - \widehat{m}(X_i)|$). The resulting (simplified and automated) I^3 is a threshold procedure where $V_i = \widehat{\Delta}_i$. Note that our original interactive procedure is highly flexible, making the power analysis less obvious, so we discuss the power of Crossfit-I^3 with the above simplified selection strategy.

To contextualize our power analysis, we paraphrase one of the results in Arias-Castro and Chen [2017]; Rabinovich et al. [2020]. Assume the test statistics $V_i \sim N(\mu_i, 1)$ are independent, with $\mu_i = 0$ under the null and $\mu_i = \mu > 0$ otherwise. Denote the number of non-nulls as n_1 and the *sparsity* of the non-nulls is parameterized by $\beta \in (0, 1)$ such that $n_1/n = n^{-\beta}$. Let the signal μ increase with n as $\mu = \sqrt{2r \log n}$, where the *signal strength* is encoded by $r \in (0, 1)$. Their power analysis is characterized by the signal r and sparsity β, which are also critical parameters to characterize the power in our context as we state later. These authors effectively prove that *for any fixed FDR level $\alpha \in (0, 1)$, no threshold procedure can have nontrivial power if $r < \beta$, but there exist threshold procedures with asymptotic power one if $r > \beta$.*

Our analysis differs from theirs in the non-null distribution of the test statistics. Given n subjects, suppose the potential outcomes for subject i are distributed as: $Y_i^C \sim N(0, 1)$ and $Y_i^T \sim N(\mu_i, 1)$, where the alternative mean is $\mu_i = 0$ if subject i is null, or $\mu_i = \mu > 0$ if i is non-null. Thus, the observed outcome of a null is $N(0, 1)$, and that of a non-null is a *mixture* of $N(\mu, 1)$ and $N(0, 1)$ (depending on the treatment assignment), instead of a shift of the null distribution as assumed in Arias-Castro and Chen [2017], and the proof of the following result thus involves some modifications on their proofs (see Appendix D.4.1).

Theorem 12. *Given a fixed FDR level $\alpha \in (0, 1)$ and let the number of subjects n go to infinity. When there is no covariate, the automated Crossfit-I^3 and the linear-BH procedure have the same power asymptotically: if $r < \beta$, their power goes to zero; if $r > \beta$, their power goes to $1/2$. Further, among the treated subjects, their power goes to one.*

Remark 8. *Power of both methods cannot converge to a value larger than $1/2$ because without covariates, we cannot differentiate between the subjects with zero effect (whose outcome follows standard Gaussian regardless of treated or not) and the subjects with positive effects that are not treated (which also follows standard Gaussian). And the proportion of untreated subjects among those with positive effects is $1/2$ because of the assumed randomization.*

The above theorem discusses the case where there are no covariates to help guess which untreated subjects have positive effects. Next, we consider the case with an "ideal" covariate X_i: its value corresponds to whether a subject is a non-null (having positive effect) or not, $X_i = \mathbb{1}\{\mu_i > 0\}$. Here, we design the selection strategy (for step 8 of Algorithm 9) as a function of the covariates, because

we hope that subjects with the similar covariates have similar treatment effects. Specifically, for the I^3 implemented on $\mathcal{I}$, we learn a prediction of $\widehat{\Delta}_j$ by X_j using non-candidate subjects $j \in \mathcal{I}\mathcal{I}$: $\text{Pred}(x) = \frac{1}{|\mathcal{I}\mathcal{I}|}\sum_{i\in\mathcal{I}\mathcal{I}} \widehat{\Delta}_j \mathbb{1}\{X_j = x\}$, where $x = \{0, 1\}$. Then for candidate subjects $i \in \mathcal{I}$, we exclude the ones whose $\text{Pred}(X_i)$ are lower. As we integrate information among subjects with the same covariate value, all non-null subjects (i.e., those with $X_i = 1$) would excluded after the nulls (with probability tending to one), regardless of whether they are treated or not; hence we achieve power one.

Theorem 13. *Given a fixed FDR level $\alpha \in (0, 1)$ and let the number of subjects n go to infinity. With a covariate $X_i = \mathbb{1}\{\mu_i > 0\}$, the power of the automated Crossfit-I^3 converges to one for any fixed $r \in (0, 1)$ and $\beta \in (0, 1)$. In contrast, the power of the linear-BH procedure goes to zero if $r < \beta$. (When $r > \beta$, power of both methods converges to one.)*

Here is a short informal argument for why our power goes to one. Since the nulls can be excluded before the non-nulls, we focus on the test statistics of the non-nulls. Let $\xrightarrow{d}$ denote convergence in distribution. The estimated effect $\widehat{\Delta}_i \xrightarrow{d} N(\mu, 1)$ for each non-null (since in the notation of Algorithm 9, $\widehat{m}(X_i = 1)$ converges to $\mu/2$ for the non-nulls, and thus, $E_i \xrightarrow{d} N(\mu/2, 1)$ for those with $A_i = 1$, and $E_i \xrightarrow{d} N(-\mu/2, 1)$ for those with $A_i = 0$.) Hence, at the time t_0 right after all the nulls are excluded (and all the non-nulls are in $\mathcal{R}_{t_0}$), the proportion of positive estimated effects $|\mathcal{R}^+_{t_0}|/|\mathcal{R}_{t_0}|$ converges to $\Phi(\mu)$, where Φ denotes the CDF of a standard Gaussian. We can stop before t_0 and identify subjects in $\mathcal{R}^+_{t_0}$ if $\widehat{\text{FDR}}(\mathcal{R}_{t_0})$, as a function of $|\mathcal{R}^+_{t_0}|/|\mathcal{R}_{t_0}|$, is less than α, which holds when $\Phi(\mu) > \frac{1}{1+\alpha}$. The power goes to one because μ grows to infinity for any fixed $r \in (0, 1)$, so that for large n, we stop before t_0 and the proportion of rejected non-nulls $|\mathcal{R}^+_{t_0}|/|\mathcal{R}_{t_0}|$ (which converges to $\Phi(\mu)$ as argued above) also goes to one. In short, the power guarantee does not depend on the sparsity β because of the designed selection strategy that incorporates covariates.

We note that our theoretical power analysis discusses two extreme cases, one with no covariate to assist the testing procedure (Theorem 12), and the other with a single "ideal" covariate that equals the indicator of non-nulls (Theorem 13). The numerical experiments in Section 5.3 consider more practical settings, where the analyst is provided with a mixture of covariates informative about the heterogeneous effect ($X_i(1)$ and $X_i(3)$ in our example) and some uninformative ones; still, the Crossfit-I^3 tends to have reasonably high power. In the following sections, we turn to present extensions of the Crossfit-I^3 in various directions.

5.5 Extension I: FDR control of nonpositive effects

The Crossfit-I^3 controls the false identifications of subjects with zero treatment effect, as defined in the null hypothesis (96), (97) or (98). In this section, we develop a modification to additionally control the error of falsely identifying subjects with nonpositive treatment effects, by defining a different null hypothesis.

Problem setup. We define the null hypothesis for subject i as nonpositive effect:

$$H_{0i}^{\text{nonpositive}} : (Y_i^T \mid X_i) \preceq (Y_i^C \mid X_i), \tag{111}$$

or equivalently, $H_{0i}^{\text{nonpositive}} : (Y_i \mid A_i = 1, X_i) \preceq (Y_i \mid A_i = 0, X_i)$. As before, our algorithm applies to two alternative definitions of the null hypothesis. In the context of treating the potential outcomes and covariates as fixed, the null hypothesis is

$$H_{0i}^{\text{nonpositive}} : Y_i^T \leq Y_i^C, \tag{112}$$

and in the hybrid version where the potential outcomes are random with joint distribution $(Y_i^T, Y_i^C) \mid X_i \sim P_i$, the null posits

$$H_{0i}^{\text{nonpositive}} : Y_i^T \leq Y_i^C \text{ almost surely-}P_i. \tag{113}$$

Note that the nonpositive-effect null is less strict than the zero-effect null. Thus, an algorithm with FDR control for $H_{0i}^{\text{nonpositive}}$ must have valid FDR control for H_{0i}^{zero}, but the reverse needs not be true. Indeed, we observe in numerical experiments (Figure 35b) that the Crossfit-I^3 does not control FDR for the nonpositive-effect null. This section presents a variant of Crossfit-I^3 that controls false identifications of nonpositive effects, possibly more practical when interpreting the identified subjects. For example, when controlling FDR for the nonpositive-effect null at level $\alpha = 0.2$, we are able to claim that the expected proportion of identified subjects with positive effects is no less than 80%.

An interactive procedure with FDR control of nonpositive effects. Recall that the FDR control of the Crossfit-I^3 is based on property (104) that when the null hypothesis is true for subject i, we have $\mathbb{P}\left(\widehat{\Delta}_i \mid \{Y_j, X_j, E_j\}_{j=1}^n\right) \leq 1/2$, but this statement no longer holds when the null hypothesis is defined as $H_{0i}^{\text{nonpositive}}$ in (111). Fortunately, this issue can be fixed by making the condition in (104) coarser and removing the outcomes:

$$\mathbb{P}\left(\widehat{\Delta}_i \mid \{X_j\}_{j=1}^n\right) \leq 1/2,$$

which is reflected in the interactive procedure as reducing the available information for selecting subject i_t^* (at step 8 of Algorithm 9) — we additionally mask (hide) the outcome Y_i of the candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$ when implementing the I^3 on set $\mathcal{I}$. We call the resulting interactive algorithm MaY-I^3, as it masks the outcomes.

Specifically, the MaY-I^3 modifies Crossfit-I^3 where we define the available information to select subjects when implementing Algorithm 9 on set $\mathcal{I}$ as

$$\mathcal{F}_{t-1}^{-Y}(\mathcal{I}) = \sigma\left(\{X_i\}_{i\in\mathcal{R}_{t-1}(\mathcal{I})}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}(\mathcal{I})}, \sum_{i\in\mathcal{R}_{t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i > 0\}\right). \tag{114}$$

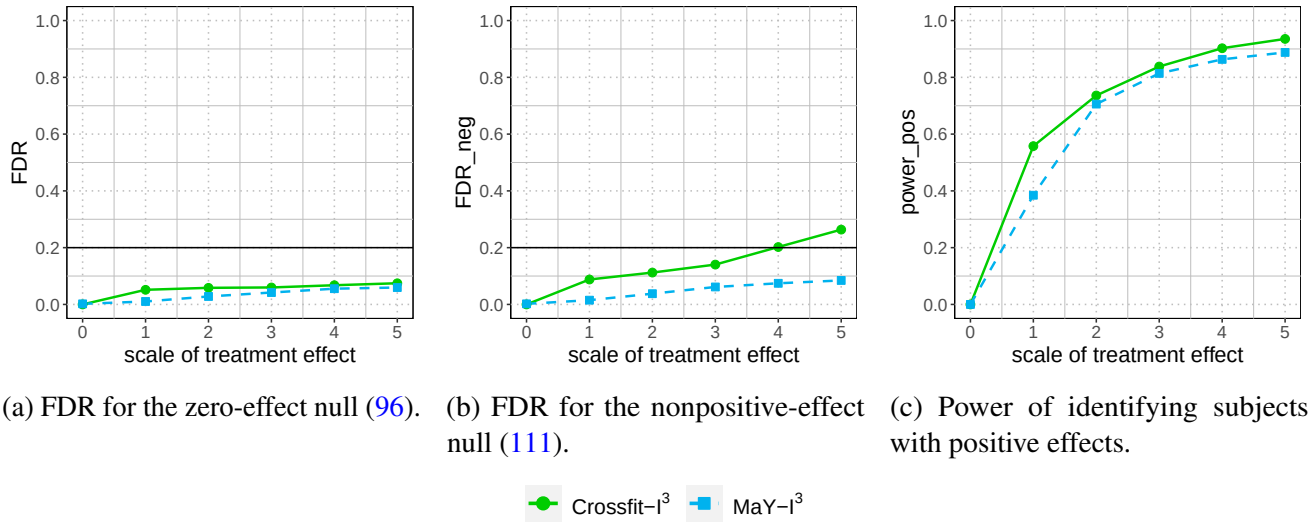
To calculate $\widehat{\Delta}_i$ at $t = 0$ when Y_i for all $i \in \mathcal{I}$ are masked, let $\widehat{m}^{-\mathcal{I}}(X_i)$ be an estimator of $\mathbb{E}(Y_i \mid X_i)$ that is learned using data from non-candidate subjects $\{Y_j, X_j\}_{j\notin\mathcal{I}}$, and let the residuals be $E_i^{-\mathcal{I}} := Y_i - \widehat{m}^{-\mathcal{I}}(X_i)$. Define $\Delta_i^{-\mathcal{I}} := 4(A_i - 1/2) \cdot E_i^{-\mathcal{I}}$, and similar to property (104) for the zero-effect null, we have

$$\mathbb{P}\left(\widehat{\Delta}_i^{-\mathcal{I}} > 0 \mid \{X_j\}_{j\in\mathcal{I}} \cup \{Y_j, X_j, E_j^{-\mathcal{I}}\}_{j\notin\mathcal{I}}\right) \leq 1/2, \tag{115}$$

under $H_{0i}^{\text{nonpositive}}$, leading to valid FDR control for nonpositive effects. Overall, the MaY-I^3 follows Algorithm 10, except the estimated treatment effect $\widehat{\Delta}_i$ replaced by $\widehat{\Delta}_i^{-\mathcal{I}}$, and the available information for selection $\mathcal{F}_{t-1}(\mathcal{I})$ replaced by $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$. See Appendix D.2.3 for the proof of FDR control.

Theorem 14. *Under assumption* (94) *and* (95) *of randomized experiments, the MaY-*I^3 *has a valid FDR control at level* α *for the nonpositive-effect null hypothesis under any of definitions* (111)*,* (112) *or* (113)*. For the last definition, FDR control also holds conditional on the covariates and potential outcomes.*

Similar to Algorithm 11 for the Crossfit-I^3, we can design an automated algorithm for the MaY-I^3 to select a subject in step 8 of Algorithm 9, but the available information $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$ no longer includes the outcomes of candidate subjects. We defer the details of the automated selection strategy in Appendix D.1.2.



(a) FDR for the zero-effect null (96).

(b) FDR for the nonpositive-effect null (111).

(c) Power of identifying subjects with positive effects.

Figure 35: Performance of two interactive methods, Crossfit-I^3 and MaY-I^3, with the treatment effect specified as model (110) and the scale S_Δ varying in $\{0, 1, 2, 3, 4, 5\}$. The MaY-I^3 controls FDR for a more relaxed null (nonpositive effects) than the Crossfit-I^3, while the Crossfit-I^3 has slightly higher power than the MaY-I^3.

Numerical experiments. We compare the Crossfit-I^3 and MaY-I^3 using the same experiment as Section 5.3.2. In terms of the error control, both the Crossfit-I^3 and MaY-I^3 control FDR for the zero-effect null at the target level (Figure 35a). When the null is defined as having a nonpositive effect, the Crossfit-I^3 can violate the error control (Figure 35b), whereas the MaY-I^3 preserves valid FDR control. In terms of the power, the Crossfit-I^3 has slightly higher power since the analyst can select subjects using information defined by $\mathcal{F}_{t-1}(\mathcal{I})$ in (105), which is richer compared to $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$ in (114) for the MaY-I^3.

To summarize, the error control of the MaY-I^3 is more strict than the Crossfit-I^3, controlling false identifications of both zero effects and negative effects, while its power is slightly lower. We recommend the Crossfit-I^3 if one only concerns the error of falsely identifying subjects with zero effect. Alternatively, we recommend the MaY-I^3 when it is desired to control the error of falsely identifying subjects with nonpositive effects.

5.6 Extension II: heterogeneous propensity scores with known bounds

Often in practice, different subjects might have a different probability of receiving treatment, possibly depending on their demographics and possibly unknown to the analysts. Following standard terminology, we refer to the probability of receiving treatment as the *propensity score*, denoted as

$$\pi_i = \mathbb{P}(A_i \mid X_1, \ldots, X_n). \tag{116}$$

Note that we allow the propensity score for subject i to depend on covariates of other subjects. This chapter extends the Crossfit-I^3 and MaY-I^3 from the setting where the propensity scores are independent of the covariates and equals $1/2$ for all subjects, to the setting with heterogeneous propensity scores that can depend on covariates and can be unknown.

To enable inference on the potential outcomes and treatment effect, we consider two common assumptions:

(iii) the propensity scores are bounded away from 0 and 1:

$$0 < \pi_{\min} \leq \pi_i \leq \pi_{\max} < 1 \text{ for all } i \in [n], \tag{117}$$

and the bounds $\pi_{\min}$ and $\pi_{\max}$ are known; and

(iv) conditional on the covariates, probabilities of receiving treatment are mutually independent:

$$\mathbb{P}[(A_1, \ldots, A_n) = (a_1, \ldots, a_n) \mid X_1, \ldots, X_n] = \prod_{i=1}^{n} \mathbb{P}(A_i = a_i \mid X_1, \ldots, X_n), \tag{118}$$

for any $(a_1, \ldots, a_n) \in \{0, 1\}^n$.

FDR error control The Crossfit-I^3 and the MaY-I^3 have valid FDR control if we modify the FDR estimator $\widehat{\mathrm{FDR}}_t$ as

$$\widehat{\mathrm{FDR}}^{\pi}(\mathcal{R}_t(\mathcal{I})) := \left(\frac{1}{\min\{\pi_{\min}, 1 - \pi_{\max}\}} - 1 \right) \frac{|\mathcal{R}_t^-(\mathcal{I})| + 1}{|\mathcal{R}_t^+(\mathcal{I})| \vee 1}, \tag{119}$$

when conducting the I^3 on set $\mathcal{I}$; and similarly for set $\mathcal{II}$. As the bounds get closer to zero or one, the above estimator takes larger value, potentially leading to a more conservative FDR control. An alternative FDR estimator is introduced in Appendix D.6, which can be less conservative when the propensity scores have extreme bounds while most are close to $1/2$ (yet, it can be more conservative when there is only mild heterogeneity in propensity scores.) We call the interactive algorithms using the above estimator $\widehat{\mathrm{FDR}}^{\pi}(\mathcal{R}_t(\mathcal{I}))$ as Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\pi^*}$, which has valid error control for the zero-effect null and the nonpositive-effect null, respectively.

Theorem 15. *Consider a randomized experiment with heterogeneous propensity scores where assumption* (117)*,* (118) *and* (95) *holds. The Crossfit-*I$^3_{\pi^*}$ *has FDR controlled at level* α *for any of the null hypotheses* (96)*,* (97) *or* (98)*. The MaY-*I$^3_{\pi^*}$ *has a valid FDR control at level* α *for the nonpositive-effect null hypothesis under any of definitions* (111)*,* (112) *or* (113)*. For the last definition of the zero-effect null* (98) *and the nonpositive-effect null* (113)*, FDR control also holds conditional on the covariates and potential outcomes.*

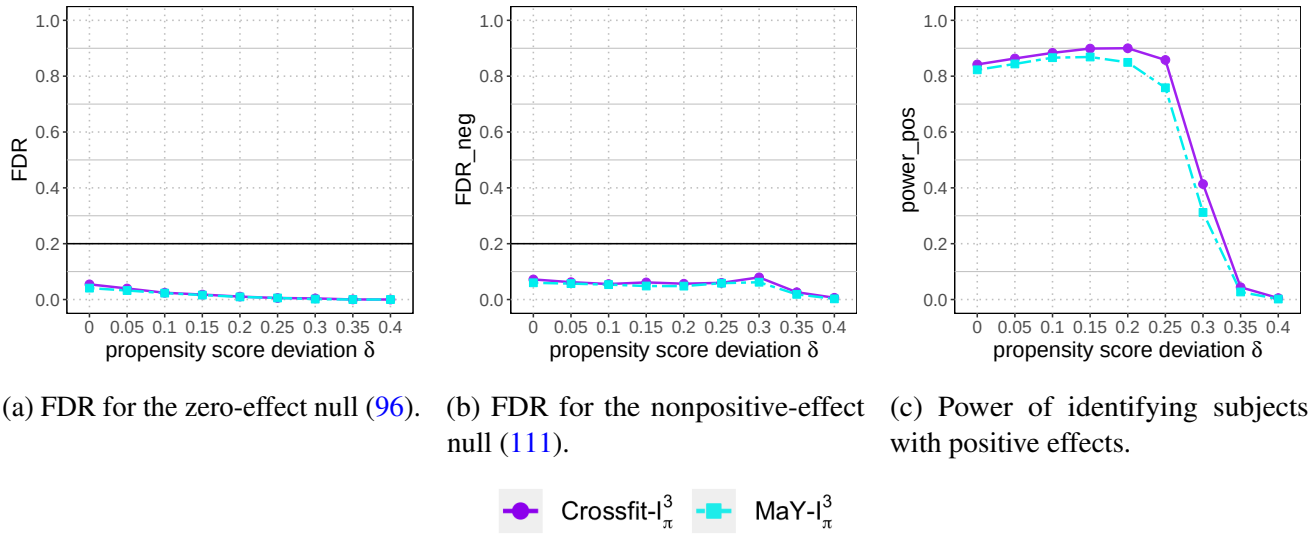
Numerical experiments. We follow the simulation setting in Section 5.3.2, except different propensity scores specified as a function of covariates. Let the the treatment effect be

$$\Delta(X_i) = 15X_i^3(3)\mathbb{1}\{X_i(3) > 1\} - 3X_i(1)/2, \tag{120}$$

which is the treatment effect in (110) with $S_\Delta = 3$. Consider the case where subjects with positive effects coincides with those having higher propensity scores:

$$\pi_i = \pi(X_i) = (1/2 + \delta)\mathbb{1}\{\Delta(X_i) > 0\} + 1/2\mathbb{1}\{\Delta(X_i) = 0\} + (1/2 - \delta)\mathbb{1}\{\Delta(X_i) < 0\}, \tag{121}$$

where $\delta \in (0, 0.5)$ denotes the deviation of the propensity score bounds to 1/2. Both the Crossfit-I$^3_{\pi^*}$ and the MaY-I$^3_{\pi^*}$ have valid error control for their target null hypotheses respectively (see Figure 36). Compared with the simple setting with $\pi_i = 1/2$ for all $i \in [n]$ (i.e., $\delta = 0$), power of both methods first increase as the deviation δ increase (to the point where $\pi_{\max} = 0.7$ and $\pi_{\min} = 0.3$). It is because with larger δ, more subjects potentially having positive treatment effects get treated, so that they show larger



(a) FDR for the zero-effect null (96).

(b) FDR for the nonpositive-effect null (111).

(c) Power of identifying subjects with positive effects.

Figure 36: Performance of of Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\pi^*}$ with knowledge of the true propensity scores, when the treatment effect specified as model (120) and the propensity score deviates from 1/2 by δ where δ varies in $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. Both Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\pi^*}$ control FDR, and have similar power. As δ increases, power first slightly increases and then decreases down to zero.

outcomes, making them easier to be identified. Nonetheless, when the deviation δ continue to increase, the power decreases to zero because the FDR estimator (119) becomes too conservative. For example, when $\delta = 0.25$, the FDR estimator is three times the FDR estimation if π_i were all $1/2$; and nine times when $\delta = 0.4$.

The Crossfit-I$^3_{\pi^*}$ and the MaY-I$^3_{\pi^*}$ achieve valid error control when we have the oracle knowledge of the bounds on the propensity scores, which might not be available in practice. Nonetheless, we can easily extend both methods to estimate the propensity scores thanks to the cross-fitting framework. The next section discusses their (asymptotic) FDR control guarantees under unknown bounds.

5.7 Extention III: heterogeneous propensity scores with unknown bounds

When the bounds for propensity scores $\pi_{\min}$ and $\pi_{\max}$ are unknown, we can estimate them using revealed data and follow the algorithms under heterogeneous propensity scores described in the previous section. Specifically in the first part of the cross-fitting framework, all the data information for subjects in $\mathcal{II}$ is revealed and we aim at finding a rejection set in $\mathcal{I}$. Prior to implementing the I^3, we estimate the bounds for the propensity scores as $\widehat{\pi_{\min}}(\mathcal{I})$ and $\widehat{\pi_{\max}}(\mathcal{I})$ by the complete data in $\mathcal{II}$ (see the first step in blue text of Algorithm 12). For example, we can estimate individual propensity scores by a logistic regression on covariates X_j using the complete data from non-candidate subjects $j \in \mathcal{II}$. Then, the analyst can conduct the I^3 with the FDR estimator defined as

$$\widehat{\mathrm{FDR}}_t^{\widehat{\pi}}(\mathcal{R}_t(\mathcal{I})) := \left(\frac{1}{1 - \max\{1 - \widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}} - 1\right) \frac{|\mathcal{R}_t^-(\mathcal{I})| + 1}{|\mathcal{R}_t^+(\mathcal{I})| \vee 1}; \tag{122}$$

and similarly for the procedure on set $\mathcal{II}$. We call the resulting algorithms Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$, to control FDR for the zero-effect null and the nonpositive-effect null, respectively.

Algorithm 12 The I^3 (implemented on $\mathcal{I}$) in Crossfit-$\mathrm{I}^3_{\widehat{\pi}}$ with unknown bounds of propensity scores.

Initial state: Explorer (E) knows covariates and outcomes $\{X_k, Y_k\}_{k=1}^n \cup \{A_j\}_{j\in\mathcal{II}}$.
Oracle (O) knows the treatment assignments $\{A_i\}_{i\in\mathcal{I}}$.
Target FDR level α is public knowledge.
Initial exchange: Both players initialize $\mathcal{R}_0 = [\mathcal{I}]$ and set $t = 1$.
1. E estimates the bounds of propensity scores $\widehat{\pi_{\min}}(\mathcal{I})$ and $\widehat{\pi_{\max}}(\mathcal{I})$.
2. E builds a prediction model $\widehat{m}$ from X_i to Y_i.
3. E informs O about residuals $E_i \equiv Y_i - \widehat{m}(X_i)$.
4. O estimates the treatment effect as $\widehat{\Delta}_i \equiv 4(A_i - 1/2)E_i$.
5. O then divides $\mathcal{R}_t$ into $\mathcal{R}_t^- := \{i \in \mathcal{R}_t : \widehat{\Delta}_i \le 0\}$ and $\mathcal{R}_t^+ := \{i \in \mathcal{R}_t : \widehat{\Delta}_i > 0\}$.
6. O reveals only $|\mathcal{R}_t^+|$ to E (who infers $|\mathcal{R}_t^-|$).
Repeated interaction:
7. E checks if $\widehat{\mathrm{FDR}}_t^{\widehat{\pi}}(\mathcal{R}_t(\mathcal{I})) \equiv \left(\frac{1}{1-\max\{1-\widehat{\pi_{\min}}(\mathcal{I}),\widehat{\pi_{\max}}(\mathcal{I})\}} - 1\right)\frac{|\mathcal{R}_t^-(\mathcal{I})|+1}{|\mathcal{R}_t^+(\mathcal{I})|\vee 1} \le \alpha$.
8. If yes, E sets $\tau = t$, reports $\mathcal{R}_\tau^+$ and exits.
9. Else, E picks any $i_t^* \in \mathcal{R}_{t-1}$ using everything E currently knows.
(E tries to pick an i_t^* that E thinks is null, i.e. E hopes that $\widehat{\Delta}_{i_t^*} \le 0$.)
10. O reveals $A_{i_t^*}$ to E, who also infers $\widehat{\Delta}_{i_t^*}$ and its sign.
11. E updates $\mathcal{R}_{t+1} = \mathcal{R}_t \backslash \{i_t^*\}$, and also $|\mathcal{R}_{t+1}^+|$ and $|\mathcal{R}_{t+1}^-|$.
12. Increment t and go back to Step 6.

5.7.1 Asymptotic FDR control

With estimated propensity scores, the FDR control can be achieved asymptotically if certain statistics are well-estimated, as we discuss in the following.

Asymptotic FDR control when the bounds of propensity scores are well-estimated. Because the bounds of propensity scores are estimated, the Crossfit-$\mathrm{I}^3_{\widehat{\pi}}$ cannot guarantee the FDR control exactly at the target level. Still, we can show that small error in the propensity score estimation would not inflate FDR dramatically.

Theorem 16. *Suppose there are n samples. In the cross-fitting framework, let $\widehat{\pi_{\min}}(\mathcal{I})$ and $\widehat{\pi_{\max}}(\mathcal{I})$ be the estimated lower and upper bound of the propensity scores based on $\mathcal{F}_0(\mathcal{I})$ (data initially known to the explorer). Define the estimation error as*

$$\epsilon_n^\pi(\mathcal{I}) \equiv \max\{|\widehat{\pi_{\min}}(\mathcal{I}) - \pi_{\min}(\mathcal{I})|, |\widehat{\pi_{\max}}(\mathcal{I}) - \pi_{\max}(\mathcal{I})|\}, \tag{123}$$

where $\pi_{\min}(\mathcal{I})$ and $\pi_{\max}(\mathcal{I})$ are minimum and maximum propensity score among subjects in $\mathcal{I}$; and define $\epsilon_n^\pi(\mathcal{II})$ similarly. The FDR is upper bounded:

$$\mathbb{E}\left[\mathrm{FDP}_\tau^{\widehat{\pi}}\right] \le \alpha\left(1 + \frac{\left(\mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\epsilon_n^\pi(\mathcal{I})\right] + \mathbb{E}_{\mathcal{F}_0(\mathcal{II})}\left[\epsilon_n^\pi(\mathcal{II})\right]\right)}{\max\{1-\pi_{\min}, \pi_{\max}\}(1-\max\{1-\pi_{\min}, \pi_{\max}\})}\right), \tag{124}$$

in a Bernoulli randomized experiment with heterogeneous propensity scores where assumption (117), (118) and (95) holds, for Crossfit-$\mathrm{I}^3_{\widehat{\pi}}$ under the zero-effect null in any of the definitions (96), (97) or (98), and for MaY-$\mathrm{I}^3_{\widehat{\pi}}$ under the nonpositive-effect null in any of the definitions (111), (112) or (113).

Corollary 1. *The Crossfit-*$\text{I}^3_{\widehat{\pi}}$ *(and MaY-*$\text{I}^3_{\widehat{\pi}}$*) has asymptotic FDR control for the zero-effect null (and the nonpositive-effect null) when the estimation of propensity score bounds is consistent in the sense that* $\mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\epsilon^{\pi}_n(\mathcal{I})\right]$ *and* $\mathbb{E}_{\mathcal{F}_0(\mathcal{I}\mathcal{I})}\left[\epsilon^{\pi}_n(\mathcal{I}\mathcal{I})\right]$ *goes to zero as sample size* n *goes to infinity.*

The Crossfit-$\text{I}^3_{\widehat{\pi}}$ (and MaY-$\text{I}^3_{\widehat{\pi}}$) would have a larger FDR than the target level if the propensity score estimation is inconsistent, and this inflation increases as the true bounds of propensity score get close to 0 and 1. Nonetheless, in the perspective where the outcomes are treated as random variables, the MaY-$\text{I}^3_{\widehat{\pi}}$ can still have asymptotic FDR control with inconsistent estimation on the propensity score bounds, if $\widehat{m}^{-\mathcal{I}}(X_i)$ is a good estimation of the expected outcomes $\mathbb{E}(Y_i \mid \{X_i\}_{i=1}^n)$. In other words, the FDR control of the MaY-$\text{I}^3_{\widehat{\pi}}$ may be considered as doubly robust.

Doubly robust asymptotic FDR control for MaY-$\text{I}^3_{\widehat{\pi}}$ when the outcomes are treated as random variables. FDR control for MaY-I^3 holds even when the bounds of true propensity scores reach 0 or 1:

(v) the propensity scores are bounded away from 0 and 1:

$$0 \leq \pi_{\min} \leq \pi_i \leq \pi_{\max} \leq 1 \text{ for all } i \in [n]. \tag{125}$$

Recall that in the simple setting where $\pi_i = 1/2$ for all subjects, FDR control guarantee is based on the probability of the sign of $\widehat{\Delta}_i$:

$$q_i(\mathcal{I}) := \mathbb{P}\left(\widehat{\Delta}_i \equiv (A_i - 1/2) \cdot (Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})\right). \tag{126}$$

Intuitively, the FDR control is close to the target level when the upper bound on $q_i(\mathcal{I})$ is close to our estimation (for example, the bound equals $1/2$ if $\pi_i = 1/2$). In MaY-$\text{I}^3_{\widehat{\pi}}$, the estimated upper bound of $q_i(\mathcal{I})$ depends on two estimations: the estimated propensity scores denoted as $\widehat{\pi}_i(\mathcal{I})$, and the estimator $\widehat{m}^{-\mathcal{I}}(X_i)$ for the conditional expected outcome $\mathbb{E}(Y_i \mid \{X_i\}_{i=1}^n)$.

To formalize the FDR control, we use the above notion of error in propensity score estimation $\epsilon^{\pi}_n(\mathcal{I})$ in (123); and introduce a characterization for error in outcome estimation. We define a "centered" CDF Φ for the conditional distribution of outcome Y_i given covariates:

$$\Phi_i(\epsilon) := \mathbb{P}(Y_i - \mathbb{E}(Y_i \mid \{X_i\}_{i=1}^n) \leq \epsilon \mid \{X_i\}_{i=1}^n), \tag{127}$$

which measure the deviation of Y_i from $\mathbb{E}(Y_i \mid \{X_i\}_{i=1}^n)$. Given a fixed ϵ, denote the lower and upper bounds as $\Phi_{\min}(\epsilon) := \min_{i\in[n]} \Phi_i(\epsilon)$ and $\Phi_{\max}(\epsilon) := \max_{i\in[n]} \Phi_i(\epsilon)$. In our characterization, we define the error of outcome estimation by $\widehat{m}^{-\mathcal{I}}(X_i)$ as

$$\epsilon^{Y}_n(\mathcal{I}) := \max_{i\in\mathcal{I}}\{|\mathbb{E}(Y_i \mid \{X_i\}_{i=1}^n) - \widehat{m}^{-\mathcal{I}}(X_i)|\}, \tag{128}$$

where recall that $\widehat{m}^{-\mathcal{I}}(X_i))$ is the expected outcome estimated using $\mathcal{F}_0^{-Y}(\mathcal{I})$, and such error is shown in the deviation of estimating q_i by the above centered CDF function Φ. Intuitively, if the estimation error $\epsilon^{Y}_n(\mathcal{I})$ is small, the centered CDF $\Phi_i\left[\epsilon^{Y}_n(\mathcal{I})\right]$ is close to $1/2$ by definition (when the outcome distribution is symmetric and continuous [13]), leading to less FDR inflation as we describe soon.

With the above notion of estimation error in the propensity score bounds $\epsilon^{\pi}_n(\mathcal{I})$ and in the outcome $\epsilon^{Y}_n(\mathcal{I})$, we can quantify two statistics that determine the FDR control:

[13] Readers might notice that the outcome distribution need not be symmetric for the centered CDF to be around $1/2$, if we replace the expected outcomes with median. Detailed discussion on the algorithm using median can be found in Appendix D.7, which leads to more robust error control in certain cases but tend to have lower power.

- an upper bound for $q_i(\mathcal{I})$ in (126), denoted as $q_{\max}(\mathcal{I})$,

$$q_{\max}(\mathcal{I}) := \min\{\max\{\pi_{\max}(\mathcal{I}), 1-\pi_{\min}(\mathcal{I})\}, \max\left\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1-\Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\right\}\}, \tag{129}$$

  which is close to 1/2 (the ideal case) when *either* the true propensity score is close to $1/2$ *or* the error of the expected outcome estimation $\epsilon_n^Y(\mathcal{I})$ is small;

- an upper bound for the estimation error $q_{\max}(\mathcal{I}) - \widehat{q}_{\max}(\mathcal{I})$, where $\widehat{q}_{\max}(\mathcal{I}) = \max\{1-\widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}$ in MaY-I$^3_{\widehat{\pi}}$[14], denoted as $\epsilon_n^q(\mathcal{I})$:

$$\epsilon_n^q(\mathcal{I}) := \\ \epsilon_n^\pi(\mathcal{I}) - \max\left\{0, \max\{\pi_{\max}(\mathcal{I}), 1-\pi_{\min}(\mathcal{I})\} - \max\left\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1-\Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\right\}\right\}, \tag{130}$$

  which is small if *either* the propensity score bounds are well-estimated so that $\epsilon_n^\pi(\mathcal{I})$ is small, *or* the expected outcomes are well-estimated so that $\epsilon_n^Y(\mathcal{I})$ is small and true propensity score bounds $\pi_{\min}$ and $\pi_{\max}$ are away from 1/2.

Similar error terms can be derived for the procedure on set $\mathcal{II}$.

Theorem 17. *The FDR of MaY-*I$^3_{\widehat{\pi}}$ *is upper bounded:*

$$\mathbb{E}\left[\mathrm{FDP}_\tau^{\widehat{\pi}}\right] \le \alpha\left\{1 + \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\frac{\epsilon_n^q(\mathcal{I})}{q_{\max}(\mathcal{I})(1-q_{\max}(\mathcal{I}))}\right] + \mathbb{E}_{\mathcal{F}_0(\mathcal{II})}\left[\frac{\epsilon_n^q(\mathcal{II})}{q_{\max}(\mathcal{II})(1-q_{\max}(\mathcal{II}))}\right]\right\},$$

in a Bernoulli randomized experiment with heterogeneous propensity scores where assumption (117), (118) *and* (95) *holds, for the zero-effect null in the two definitions* (96) *or* (98) *that treats the outcomes as random variables.*

Corollary 2. *As sample size n goes to infinity, the MaY-*I$^3_{\widehat{\pi}}$ *has asymptotic FDR control for the zero-effect null* (96) *or* (98) *when either*

1. *(a) the propensity score estimation is consistent in that $\mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\epsilon_n^\pi(\mathcal{I})\right]$ and $\mathbb{E}_{\mathcal{F}_0(\mathcal{II})}\left[\epsilon_n^\pi(\mathcal{II})\right]$ goes to zero; and (b) the true propensity scores are bounded away from 0 and 1: $0 < \pi_{\min} \le \pi_{\max} < 1$; or*
2. *(a) the expected outcome estimation is consistent in that $\epsilon_n^Y(\mathcal{I})$ goes to zero almost surely over the conditional distribution given $\mathcal{F}_0(\mathcal{I})$ (and same for set $\mathcal{II}$); and (b) the difference between bounds on true propensity scores and $1/2$ is larger than its estimation error: $\max\{\pi_{\max}(\mathcal{I}), 1-\pi_{\min}(\mathcal{I})\} - 1/2 \ge \epsilon_n^\pi(\mathcal{I})$ almost surely over the conditional distribution given $\mathcal{F}_0(\mathcal{I})$ (and same for set $\mathcal{II}$); and (c) the distribution of outcome Y_i is symmetric given the covariates for $i \in [n]$.*

Hence double robustness. We remark that there is one case where the FDR inflation could be large: the actual propensity scores are $1/2$ for all subjects but unknown, and the propensity score estimation is poor.

The above theorem states the FDR guarantee for the zero-effect null, and the error control for the nonpositive-effect null is discussed in Appendix D.3.3. The condition to ensure asymptotic error control is the same as above except a different definition of the outcome estimation error $\widetilde{\epsilon}_n^Y(\mathcal{I}) := \max_{i\in\mathcal{I}}\{|\mathbb{E}(Y_i^C \mid \{X_i\}_{i=1}^n) - \widehat{m}^{-\mathcal{I}}(X_i)|\}$ and a different definition of the centered CDF $\widetilde{\Phi}_i(\epsilon) :=$

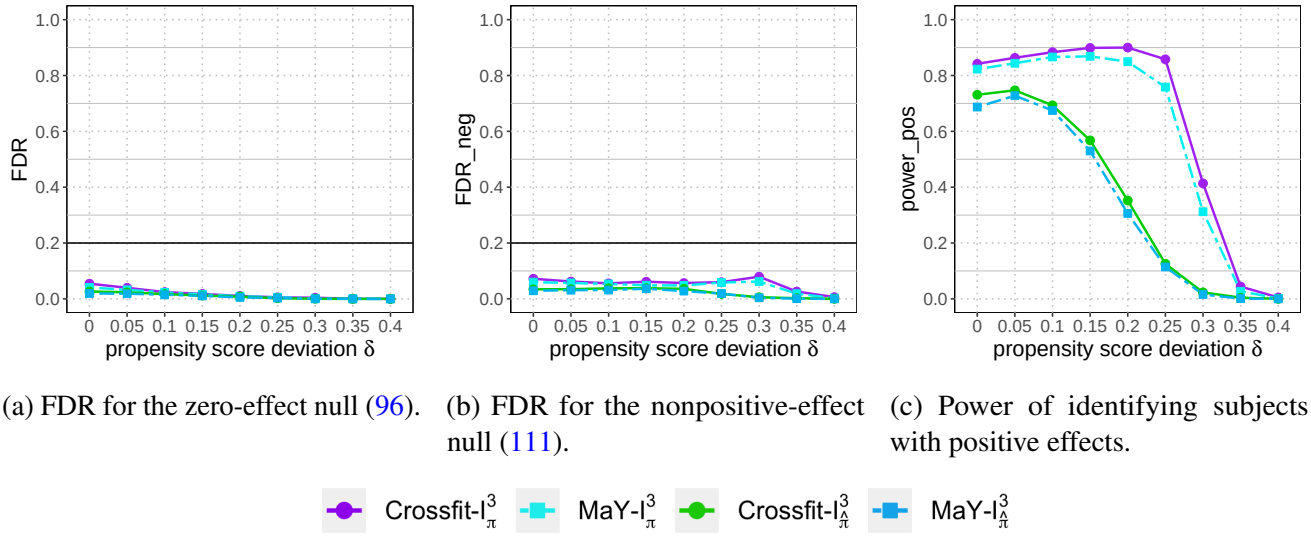
[14]Note that it is possible to design alternative estimation $\widehat{q}_{\max}(\mathcal{I})$, which lead to different $\epsilon_n^q(\mathcal{I})$, and in turn different levels of robustness for FDR control and power. See Appendix D.8 for details, and the presented MaY-I$^3_{\widehat{\pi}}$ has reasonably high power and double robustness.

$\mathbb{P}(Y_i^C - \mathbb{E}(Y_i^C|\{X_i\}_{i=1}^n) \leq \epsilon \mid \{X_i\}_{i=1}^n)$, where Y_i^C is the potential control outcome. Note that it could be less likely to take advantage of the double robustness when the propensity scores are poorly estimated, because $\widetilde{\epsilon}_n^Y(\mathcal{I})$ tends to be not small ($\widehat{m}^{-\mathcal{I}}(X_i)$ is an estimator for $\mathbb{E}(Y_i \mid \{X_i\}_{i=1}^n)$, and can be very different from the expected *control* outcome). Additionally, we comment that the above theorem provides an upper bound of the FDR in terms of the maximum estimation error over all subjects, which could potentially be conservative, and in practice we expect the FDR to be close to the target level when the the estimation error is small for most subjects.

5.7.2 Numerical experiments

Follow the same simulation setting as in Section 5.6, we explore several approaches when the propensity scores are unknown: estimating the propensity scores as in Crossfit-$\text{I}^3_{\widehat{\pi}}$ and MaY-$\text{I}^3_{\widehat{\pi}}$; and falsely treating all propensity scores as $1/2$ and implement the original Crossfit-I^3 and MaY-I^3. We are interested in the sensitivity of the latter approach because we might assume propensity scores to be $1/2$ while they differ in practice.

Crossfit-$\text{I}^3_{\widehat{\pi}}$ and MaY-$\text{I}^3_{\widehat{\pi}}$ with estimated propensity scores appear to control FDR at the target level for their corresponding null hypotheses, respectively (see Figure 37). They have less power compared with the Crossfit-$\text{I}^3_{\pi^*}$ and MaY-$\text{I}^3_{\pi^*}$, which is expected since the latter two methods make use of the true propensity scores.



(a) FDR for the zero-effect null (96).

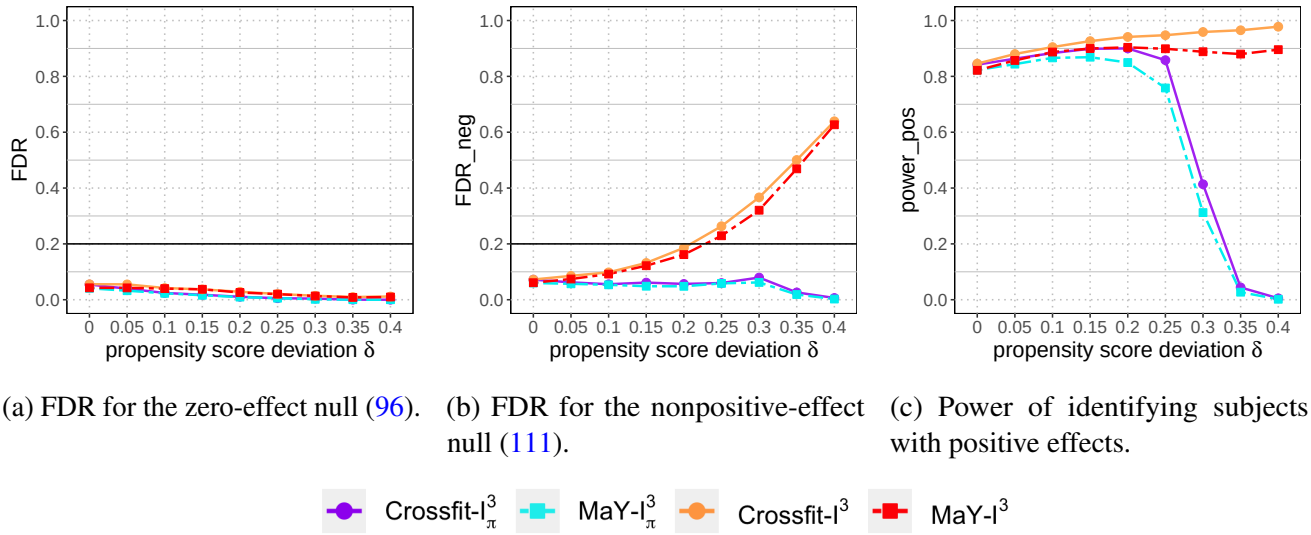
(b) FDR for the nonpositive-effect null (111).

(c) Power of identifying subjects with positive effects.

Figure 37: Performance of Crossfit-$\text{I}^3_{\widehat{\pi}}$ and MaY-$\text{I}^3_{\widehat{\pi}}$, which estimate the propensity scores, compared with Crossfit-$\text{I}^3_{\pi^*}$ and MaY-$\text{I}^3_{\widehat{\pi}}$, which use the knowledge of the true propensity scores, when the treatment effect specified as model (120) and the propensity score deviates from 1/2 by δ where δ varies in $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. Both Crossfit-$\text{I}^3_{\widehat{\pi}}$ and MaY-$\text{I}^3_{\widehat{\pi}}$ appears to control FDR, and have similar power. Their power are lower than Crossfit-$\text{I}^3_{\pi^*}$ and MaY-$\text{I}^3_{\widehat{\pi}}$ because the latter additionally use the true propensity scores.

When all propensity scores are falsely treated as $1/2$, we can implement Crossfit-I^3 and MaY-I^3 (see Figure 38). In our experiments, the FDR for the zero-effect null seems to be controlled below the target level even when the true propensity scores are extreme (with $\pi_{\min} = 0.1$ and $\pi_{\max} = 0.9$ when $\delta = 0.4$). It coincides with our claim on doubly robust FDR control once noticing that MaY-I^3 is equivalent to MaY-$\text{I}^3_{\widehat{\pi}}$ when $\widehat{\pi}_{\min} = \widehat{\pi}_{\max} = 1/2$. In such a case, the propensity scores are poorly

estimated $|\widehat{\pi}_{\min} - \pi_{\min}| = |\widehat{\pi}_{\max} - \pi_{\max}| = 0.4$, but FDR can be small when the expected outcome $\mathbb{E}(Y_i \mid \{X_i\}_{i=1}^n)$ is well-estimated by $\widehat{m}^{-\mathcal{I}}(X_i)$. The FDR for the nonpositive-effect null can exceed the target level when the deviation δ is large and the propensity score estimation is poor, corresponding to the case where $\pi_{\min} \leq 0.25$ and $\pi_{\max} \geq 0.75$. The power of Crossfit-I^3 and MaY-I^3 does not follow the same trend as Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\pi^*}$ when δ grows large, because their FDR estimator does not suffer from conservativeness introduced by extreme propensity scores.



(a) FDR for the zero-effect null (96). (b) FDR for the nonpositive-effect null (111). (c) Power of identifying subjects with positive effects.

Figure 38: Performance of Crossfit-I^3 and MaY-I^3, which falsely treat all propensity scores as $1/2$, compared with Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\widehat{\pi}}$, which use the true propensity scores, when the treatment effect specified as model (120) and the propensity score deviates from 1/2 by δ where δ varies in $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. Power of the Crossfit-I^3 and MaY-I^3 increases because they do not suffer from conservative FDR estimator as δ increases. Although FDR for the nonpositive-effect null grows to exceed the target level when δ is larger than 0.2, FDR control for the zero-effect null seems to hold even when the true propensity scores are vastly different from $1/2$.

5.7.3 Adjustment in the case with a few extreme propensity scores.

We have seen in the above experiments that the FDR estimator accounting for heterogeneous propensity scores can be rather conservative. Nonetheless, we hope the procedure to still be powerful for most subjects when only a few have extreme propensity scores. A simple solution is to exclude the ones with extreme propensity scores before implement I^3. Data of these excluded subjects can be revealed at the beginning for identification on the rest subjects (see Algorithm 13 where the changes from Crossfit-I$^3_{\pi^*}$ are marked by blue text; similar changes applies to MaY-I$^3_{\pi^*}$).

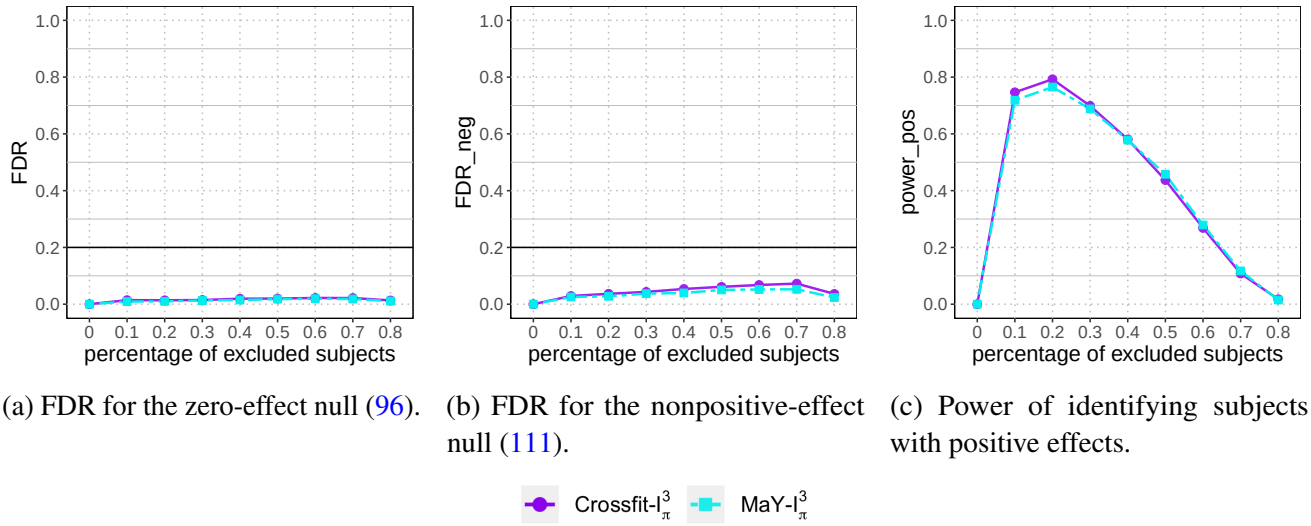
In the previous experiments, the highest propensity score corresponds to all subjects with positive effect that we hope to identify, so excluding them would lead to low power. Here, we implement the above procedure with only a few extreme propensity scores which possibly do not have positive effect. Specifically, let $e \sim N(0, 1)$ be a random noise adding to the original propensity scores. Define

$$\pi_i = \begin{cases} 0.5 + \delta + 0.05e, & \text{if } \Delta(X_i) > 0; \\ 0.5 + 0.2e, & \text{if } \Delta(X_i) = 0; \\ 0.5 - \delta + 0.05e, & \text{if } \Delta(X_i) < 0; \end{cases} \tag{131}$$

which compared with (121) add a small Gaussian noise with 0.05 standard deviation if a subject's covariates indicate nonzero effect; and a larger noise with 0.2 standard deviation otherwise. We bound π_i in $[0,1]$: $\pi_i = \pi_i \mathbb{1}\{0 \le \pi \le 1\} + 1\mathbb{1}\{\pi_i > 1\}$ for it to be a well-defined propensity score. Here, we set $\delta = 0.1$ and explore the performance of several excluding rules.

Algorithm 13 The Crossfit-I$^3_{\pi^*}$ that excludes extreme propensity scores.

Input: Covariates, outcomes, treatment assignments $\{Y_i, A_i, X_i\}_{i=1}^n$, target level α;
Procedure:
1. Randomly split the sample into two subsets of equal size, denoted as $\mathcal{I}$ and $\mathcal{II}$;
2. Select subjects in $\mathcal{I}$ with extreme propensity scores, denoted as $\mathcal{E}(\mathcal{I})$.
3. Implement Algorithm 9 with FDR estimator $\widehat{\mathrm{FDR}}^{\pi}$ at level $\alpha/2$, where E initially knows $\{Y_k, X_k\}_{k=1}^n \cup \{A_j\}_{j \in \mathcal{II} \cup \mathcal{E}(\mathcal{I})}$ and sets $\mathcal{R}_0(\mathcal{I}) = \mathcal{I} \backslash \mathcal{E}(\mathcal{I})$, getting a rejection set $\mathcal{R}_\tau^+(\mathcal{I}) \subseteq \mathcal{I} \backslash \mathcal{E}(\mathcal{I})$;
4. Select subjects in $\mathcal{II}$ with extreme propensity scores, denoted as $\mathcal{E}(\mathcal{II})$.
5. Implement Algorithm 9 with FDR estimator $\widehat{\mathrm{FDR}}^{\pi}$ at level $\alpha/2$, where E initially knows $\{Y_k, X_k\}_{k=1}^n \cup \{A_j\}_{j \in \mathcal{I} \cup \mathcal{E}(\mathcal{II})}$ and sets $\mathcal{R}_0(\mathcal{II}) = \mathcal{II} \backslash \mathcal{E}(\mathcal{II})$, getting a rejection set $\mathcal{R}_\tau^+(\mathcal{II}) \subseteq \mathcal{II} \backslash \mathcal{E}(\mathcal{II})$;
4. Combine two rejection sets as the final rejection set, $\mathcal{R}_\tau^+ = \mathcal{R}_\tau^+(\mathcal{I}) \cup \mathcal{R}_\tau^+(\mathcal{II})$.



(a) FDR for the zero-effect null (96).

(b) FDR for the nonpositive-effect null (111).

(c) Power of identifying subjects with positive effects.

Figure 39: Performance of Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\pi^*}$ when the treatment effect specified as model (120) and the propensity score deviates from 1/2 by 0.1, where we vary the percentage of excluded subjects with most extreme propensity scores in $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Excluding 20% of the subjects seems to lead to the highest power.

For an example in implementation, the set of excluded subjects $\mathcal{E}$ can be defined as subjects whose propensity score deviation $|\pi_i - 1/2|$ is larger than a q-upper quantile. That is, q-percentage of the most extreme propensity scores would be excluded. We try the excluding procedure with varying percentage parameter q in $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Note that the procedure when $q = 0$ collapse to the original Crossfit-I$^3_{\pi^*}$ (and MaY-I$^3_{\pi^*}$). While excluding too small proportion of the extreme propensity scores might leave FDR estimator still being conservative, excluding too many subjects could also lead to power loss because the excluded subjects cannot be identified even if they have positive effects. In our

experiments, excluding 20% seems a good choice leading to high power. The adjustment of excluding extreme propensity scores can also be applied to Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$ where the propensity scores are estimated, after which we can decide which subjects to exclude. However, because the estimated propensity scores often tends to be less extreme than the true ones, the benefit of excluding is not as evident; thus omitted from the main paper.

5.8 Extension IV: paired samples

Problem setup. Our discussion has focused on the case where samples are not paired, and the proposed algorithms can be extended to the paired-sample setting. Suppose there are n pairs of subjects. Let outcomes of subjects in the i-th pair be Y_{ij}, treatment assignments be indicators A_{ij}, covariates be X_{ij} for $j = 1, 2$ and $i \in [n]$. We deal with randomized experiments without interference, and assume that

(i) conditional on covariates, the treatment assignments are independent coin flips:

$$\mathbb{P}[(A_{11}, \ldots, A_{n1}) = (a_1, \ldots, a_n) \mid X_1, \ldots, X_n] = \prod_{i=1}^{n} \mathbb{P}(A_i = a_i) = (1/2)^n, \text{ and}$$

$$A_{i1} + A_{i2} = 1 \text{ for all } i \in [n].$$

(ii) conditional on covariates, the outcome of one subject Y_{i_1,j_1} is independent of the treatment assignment of another subject A_{i_2,j_2} conditional on A_{i_1,j_1}, for any $(i_1, j_1) \neq (i_2, j_2)$.

As before, we can develop interactive algorithms for two types of error control (only the definitions when treating the potential outcomes as random variables are presented, but the FDR control still applies to all versions of the null):

$$H_{0i}^{(\text{zero, paired})} : (Y_{ij}^T \mid X_{ij}) \overset{d}{=} (Y_{ij}^C \mid X_{ij}) \text{ for both } j = 1, 2; \tag{132}$$

$$H_{0i}^{(\text{nonpositive, paired})} : (Y_{ij}^T \mid X_{ij}) \preceq (Y_{ij}^C \mid X_{ij}) \text{ for both } j = 1, 2. \tag{133}$$

Here, we present the extension of Crossfit-I^3 for FDR control of zero effect, and defer the extension of MaY-I^3 for FDR control of nonpositive effect to Appendix D.1.3.

Interactive algorithms for paired samples. With the pairing information, the treatment effect can be estimated without involving $\widehat{m}$ as in (103):

$$\widehat{\Delta}_i^{\text{paired}} := (A_{i1} - A_{i2})(Y_{i1} - Y_{i2}), \tag{134}$$

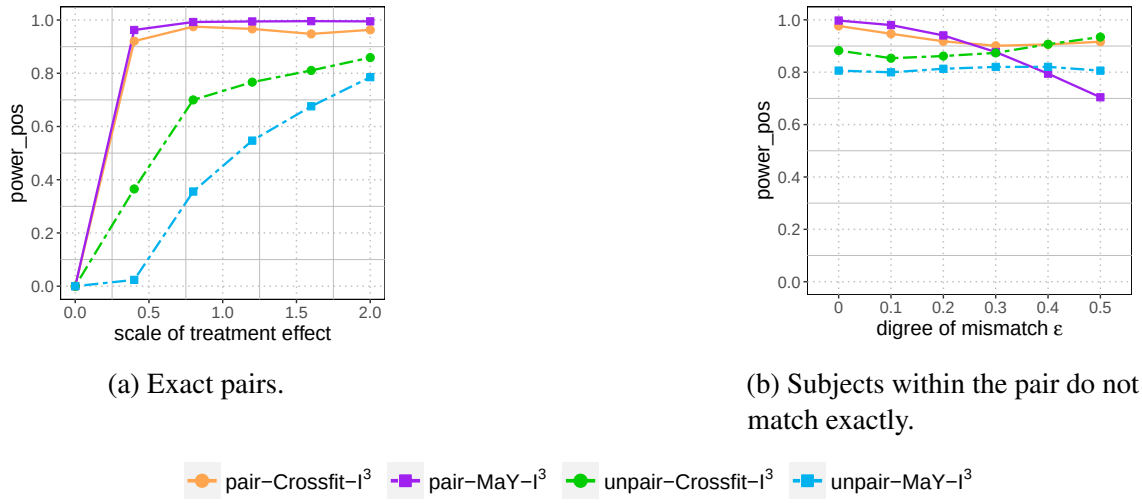
as used by Rosenbaum [2002] and Howard and Pimentel [2020], among others. The above estimation satisfies the critical property to guarantee FDR control: for a null pair i of two subjects with zero effects in (132), we have

$$\mathbb{P}(\widehat{\Delta}_i^{\text{paired}} > 0 \mid \{Y_{j1}, Y_{j2}, X_{j1}, X_{j2}\}_{j=1}^n) \leq 1/2. \tag{135}$$

Thus, the Crossfit-I^3 (Algorithm 10) with $\widehat{\Delta}_i$ replaced by $\widehat{\Delta}_i^{\text{paired}}$ has valid FDR control for the zero-effect null (132), where the analyst excludes pairs using the available information, including $\{Y_{i1}, Y_{i2}, X_{i1}, X_{i2}\}$ for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$, and $\{Y_{j1}, Y_{j2}, A_{j1}, A_{j2}, X_{j1}, X_{j2}\}$ for non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$, and the sum $\sum_{i \in \mathcal{R}_{>t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i^{\text{paired}} > 0\}$ for FDR estimation. An automated strategy exclude pair i_t^* (at step 8 of Algorithm 9) under paired samples is the same as Algorithm 11, except $\widehat{\Delta}_i$ being replaced by $\widehat{\Delta}_i^{\text{paired}}$.

Numerical experiments. We compare the power of the interactive procedures with and without the pairing information, using the same experiments as previous. When the subjects within each pair have the same covariate values, the power under paired samples is higher than treating them as unpaired (see Figure 40a), because the noisy variation in the observed outcomes that results from the potential control outcomes can be removed by taking the difference in outcomes within each pair.

The advantage of procedures under paired samples becomes less evident when the subjects within a pair do not match exactly. We simulate unmatched pairs by introducing a parameter ϵ such that for each pair i, the covariates of the two subjects within satisfy: $\mathbb{P}(X_{i1}(1) \neq X_{i2}(1)) = \epsilon, \mathbb{P}(X_{i1}(2) \neq X_{i2}(2)) = \epsilon, X_{i1}(3) = X_{i2}(3) + U(0, 2\epsilon)$, where $U(0, 2\epsilon)$ is uniformly distributed between 0 and 2ϵ, and a larger ϵ leads to a larger degree of mismatch. As ϵ increases, the power of procedures using the pairing information decreases (see Figure 40b), because the estimated treatment effect $\widehat{\Delta}_i^{\text{paired}}$ becomes less accurate for the mismatching setting. We further investigate the power decrease in Appendix D.5.2.



(a) Exact pairs.

(b) Subjects within the pair do not match exactly.

Figure 40: Power under paired samples with treatment effects specified by model (110) when our proposed algorithms (Crossfit-I^3 and MaY-I^3) utilize the pairing information, which is higher than treating all subjects as unpaired. The advantage is less evident when the subjects within each pair are not exactly matched to have the same covariate values.

5.9 Extension V: FDR control at a subgroup level

Our proposed interactive methods control FDR on *individual* level, which means upper bounding the proportion of falsely identified subjects. In this section, we show that the idea of interactive testing can be extended to control FDR on *subgroup* level, where we aim at identifying multiple subgroups with positive effects and upper bounding the proportion of falsely identified subgroups. Recall that FDR control at a subgroup level is studied by Karmakar et al. [2018] as we review in Section 5.1.2 of the main paper.

Problem setup. Let there be G non-overlapping subgroups $\mathcal{G}_g$ for $g \in [G] \equiv \{1, \ldots, G\}$. The null hypothesis for each subgroup is defined as zero effect for all subjects within:

$$\mathcal{H}_{0g} : H_{0i}^{\text{zero}} \text{ is true for all } i \in \mathcal{G}_g, \tag{136}$$

or equivalently, $\mathcal{H}_{0g} : \mathcal{G}_g \subseteq \mathcal{H}_0$ (recall that $\mathcal{H}_0$ is the set of true null subjects). Let D_g be the decision function receiving the values 1 or 0 for whether $\mathcal{H}_{0g}$ is rejected or not rejected, respectively, and the FDR at a subgroup level is defined as:

$$\text{FDR}^{\text{subgroup}} := \mathbb{E}\left[\frac{|\{g \in [G] : \mathcal{G}_g \subseteq \mathcal{H}_0, D_g = 1|}{\max\{|\{g \in [G] : D_g = 1|, 1\}}\right].$$

Same as the algorithms at an individual level, the algorithms we propose at a subgroup level can be applied to samples that are paired or unpaired. For simple notation, we use $\{Y_i, A_i, X_i\}$ to denote the observed data for subject i when the samples are unpaired, and for pair i when the samples are paired (where $Y_i = \{Y_{i1}, Y_{i2}\}$ and similarly for A_i and X_i).

An interactive algorithm to identify subgroups. We first follow the same steps of Karmakar et al. [2018] to define subgroups and generate the p-value for each subgroup. Specifically, the subgroups $\mathcal{G}_g$ for $g \in [G]$ is defined using the outcomes and covariates $\{Y_j, X_j\}_{j=1}^n$ (by an arbitrary algorithm or strategy, such as grouping subjects with the same covariates). For each subgroup $\mathcal{G}_g$, we can compute a p-value P_g by the classical Wilcoxon test (or using a permutation test, which obtains the null distribution by permuting the treatment assignment $\{A_i\}_{i=1}^n$).

The interactive procedure we propose differs from Karmakar et al. [2018] by how we process the p-values of the subgroups. We adopt the work of Lei et al. [2020] that proposes an interactive procedure with FDR control for generic multiple testing problems. The key property that allows human interaction while guaranteeing valid FDR control is similar to that in the I^3: the independence between the information used for selection and that used for FDR control. Here with the p-values of subgroups, the two independent parts are

$$P_g^1 := \min\{P_g, 1 - P_g\},$$

which is revealed to the analyst for selection and

$$P_g^2 := 2 \cdot \mathbb{1}\{P_g < \tfrac{1}{2}\} - 1,$$

which is masked (hidden) for FDR control. Notice that for a null subgroup with a uniform p-value, (P_g^1, P_g^2) are independent, and we have that

$$\mathbb{P}(P_g^2 = 1 \mid P_g^1, [Y_i, X_i]_{i \in \mathcal{G}_g}) \leq 1/2, \tag{137}$$

because the p-values obtained by permutating assignments is uniform when conditional on the outcomes and covariates. We remark that the above property is similar to property (104) and (115) in main paper that lead to valid FDR control at an individual level.

Similar to the proposed methods at an individual level, the interactive procedure for subgroups progressively excludes subgroups and recursively estimates the FDR. Let the candidate rejection set $\mathcal{R}_t$ be a set of selected subgroups, starting from all subgroups included $\mathcal{R}_0 = [G]$. We interactively shrink $\mathcal{R}_t$ using the available information:

$$\mathcal{F}_{t-1}^{\text{subgroup}} = \sigma\left(\{P_g^1, [Y_i, X_i]_{i \in \mathcal{G}_g}\}_{g \in \mathcal{R}_{t-1}}, \{P_g, [Y_j, A_j, X_j]_{j \in \mathcal{G}_g}\}_{g \notin \mathcal{R}_{t-1}}, \sum_{g \in \mathcal{R}_{t-1}} P_g^2\right),$$

Algorithm 14 An interactive procedure for subgroup identification.

Initial state: Explorer (E) knows the covariates, outcomes $\{Y_i, X_i\}_{i=1}^n$.
Oracle (O) knows the treatment assignments $\{A_i\}_{i=1}^n$.
Target FDR level α is public knowledge.
Initial exchange: Set $t = 1$.
1. E defines subgroups $\{\mathcal{G}_g\}_{g=1}^G$ using $\{Y_i, X_i\}_{i=1}^n$.
2. Both players initialize $\mathcal{R}_0 = [G]$, and E informs O about the subgroup division.
3. O compute the p-value for each subgroup $\{P_g\}_{g=1}^G$, and decompose each p-value as $P_g^1 := \min\{P_g, 1 - P_g\}$ and $P_g^2 := 2 \cdot \mathbb{1}\{P_g < \frac{1}{2}\} - 1$.
4. O then divides $\mathcal{R}_t$ into $\mathcal{R}_t^- := \{g \in \mathcal{R}_t : P_g^2 \leq 0\}$ and $\mathcal{R}_t^+ := \{g \in \mathcal{R}_t : P_g^2 > 0\}$.
5. O reveals $\{P_g^1\}_{g=1}^G$, $|\mathcal{R}_t^-|$ and $|\mathcal{R}_t^+|$ to E.
Repeated interaction: 6. E checks if $\widehat{\text{FDR}}^{\text{subgroup}}(\mathcal{R}_t) \equiv \frac{|\mathcal{R}_t^-|+1}{\max\{|\mathcal{R}_t^+|,1\}} \leq \alpha$.
7. If yes, E sets $\tau = t$, reports $\mathcal{R}_\tau^+$ and exits;
8. Else, E picks any $g_t^* \in \mathcal{R}_{t-1}$ using everything E currently knows.
(E tries to pick an g_t^* that they think is null; E hopes that $P_g^2 \leq 0$.)
9. O reveals $\{A_i\}_{i \in \mathcal{G}_g}$ to E, who also infers P_g^2.
10. E updates $\mathcal{R}_{t+1} = \mathcal{R}_t \backslash \{g_t^*\}$, and also $|\mathcal{R}_{t+1}^+|$ and $|\mathcal{R}_{t+1}^-|$;
11. Increment t and go back to Step 6.

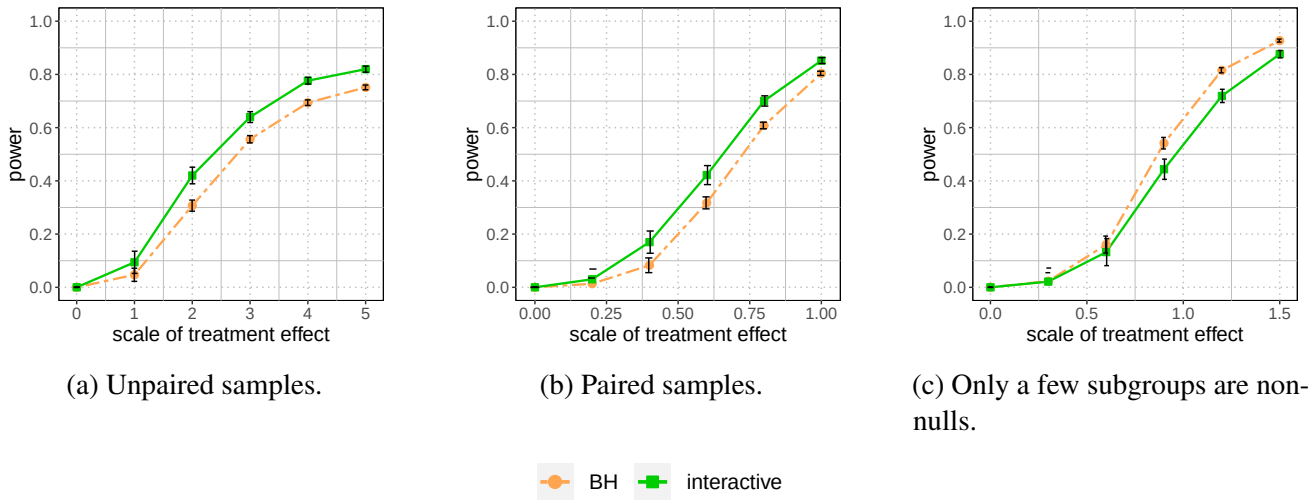
which masks (hides) the partial p-value P_g^2 and the treatment assignment A_i for candidate subgroups in $\mathcal{R}_{t-1}$; and the sum $\sum_{g \in \mathcal{R}_{t-1}} P_g^2$ is mainly provided for FDR estimation. Similar to our previously proposed interactive procedures, the FDR estimator is defined as:

$$\widehat{\text{FDR}}^{\text{subgroup}}(\mathcal{R}_t) = \frac{|\mathcal{R}_t^-| + 1}{\max\{|\mathcal{R}_t^+|, 1\}}, \tag{138}$$

with $\mathcal{R}_t^+ = \{g \in \mathcal{R}_t : P_g^2 = 1\}$ and $\mathcal{R}_t^- = \{g \in \mathcal{R}_t : P_g^2 = -1\}$. The algorithm shrinks $\mathcal{R}_t$ until time $\tau := \inf\{t : \widehat{\text{FDR}}^{\text{subgroup}}(\mathcal{R}_t) \leq \alpha\}$, and identifies only the subgroups in $\mathcal{R}_\tau^+$, as summarized in Algorithm 14. Details of strategies to select subgroup based on the revealed p-value and covariates can be found in Lei et al. [2020]. As a comparison, Karmakar et al. [2018] use the same set of p-values $\{P_g\}_{g \in [G]}$, and control FDR by the classical BH procedure.

Numerical experiments. We compare the performance of our proposed interactive procedure for subgroup identification with the method proposed by Karmakar et al. [2018], following an experiment in their paper. Suppose each subject is recorded with two discrete covariates $X_i = \{X_i(1), X_i(2)\}$ where $X_i(1) \in \{1, \ldots, 40\}$ takes 40 levels with equal probability, and $X_i(2)$ is binary with equal probability (for example, $X_i(1)$ could encode the city subject i lives in, and $X_i(2)$ the gender). The treatment effect $\Delta(X_i)$ is a constant δ if $X_i(1)$ is even, and we vary δ in six levels. We conduct the above experiment in two cases: unpaired samples ($n = 2000$) with independent covariates and paired samples ($n = 1000$) whose covariate values are the same for subjects within each pair.

Recall that the subgroups can be defined by covariates and outcomes. Here, since the covariates are discrete, we define subgroups by different values of $(X_i(1), X_i(2))$, resulting in 80 subgroups. The interactive procedure tends to have higher power than the BH procedure (see Figure 41a and Figure 41b) because it focuses on the subgroups that are more likely to be the non-nulls using the excluding process, and utilizes the covariates together with the p-values to guide the algorithm (see



(a) Unpaired samples.

(b) Paired samples.

(c) Only a few subgroups are non-nulls.

Figure 41: Performance of methods to identify subgroups with positive effects: the BH procedure and the interactive procedure (for 80 subgroups defined by the distinct values of covariates). We vary the scale of treatment effect under unpaired or paired samples. In both cases, the interactive procedure can have higher power than the BH procedure. When the number of non-null subgroups is too small (less than 20), the BH procedure can have higher power. The error bar marks two standard deviations from the center.

details in Appendix D.1.1). Meanwhile, the BH procedure does not account for covariates once the p-values are calculated. Nonetheless, the interactive procedure can have lower power when the total number of subgroups that are truly non-null is small. We simulate the case where a subject has a positive effect δ if $X_i(1)$ a multiplier of 4 (i.e., $X_i(1)/2$ is even), so that there are 20 non-null subgroups in total (previously 40 non-nulls). The power of the interactive procedure is lower than the BH procedure (see Figure 41c) because the FDR estimator in (138) can be conservative when $|\mathcal{R}^+|$ is small due to a small number of true non-nulls (for example, with FDR control at $\alpha = 0.2$, we need to shrink $\mathcal{R}_t$ until $|\mathcal{R}^-| < 3$ when $|\mathcal{R}^+|$ is around 20).

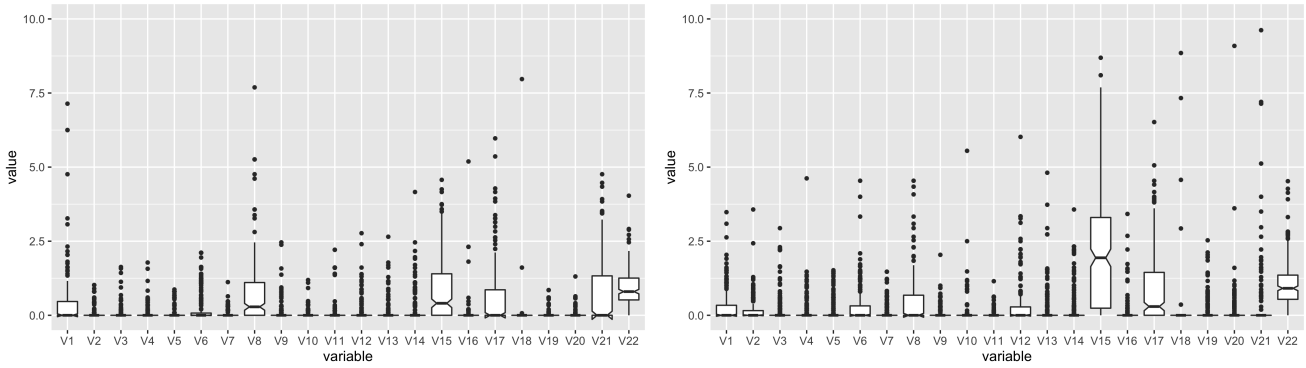
A side note is that we define the subgroups by distinct values of the covariates, whereas Karmakar et al. [2018] suggest forming subgroups by regressing the outcomes on covariates using a tree algorithm. In their experiments and several numerical experiments we tried, we find that the number of subgroups defined by the tree algorithm is usually less than ten. However, we think the FDR control is less meaningful when the total number of subgroups is small. To justify our comment, note that an algorithm with valid FDR control at level α can make zero rejection with probability $1 - \alpha$ and reject all subgroups with probability α, which can happen when the total number of subgroups is small. In contrast, with a large number of subgroups, a reasonable algorithm is unlikely to jump between the extremes of making zero rejection and rejecting all n subgroups; and thus, controlling FDR indeed informs that the proportion of false identifications is low for the evaluated algorithm.

5.10 A prototypical application to ACIC challenge dataset

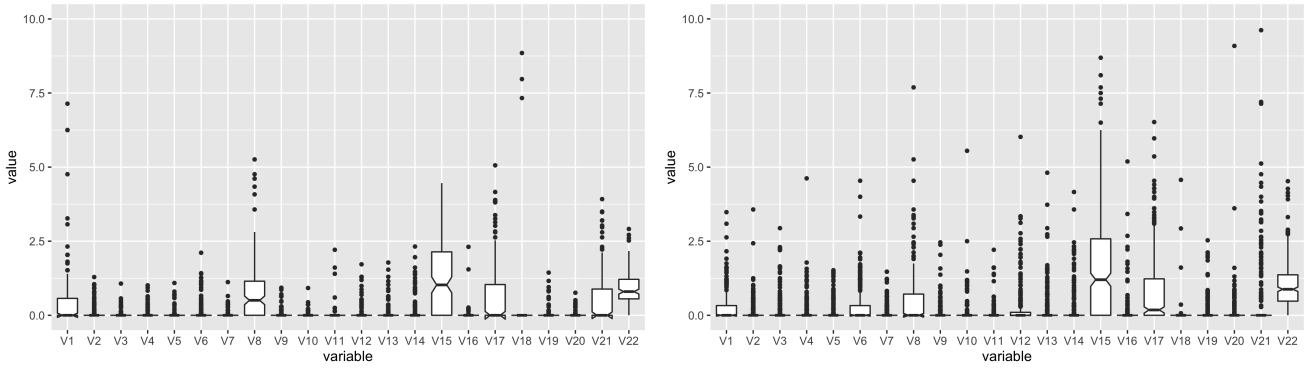
We implement our proposed methods on datasets generated by Atlantic Causal Inference Conference (ACIC), which intend to evaluate methods for average treatment effect(ATE) estimation and uses real data covariates and modified outcomes to simulate cases with heterogeneous treatment effect, heterogeneous propensity scores, etc. We take an example dataset with 500 subjects, each of which is recorded with 22

continuous covariates. The proportion of treated subjects is 0.7, indicating that the propensity scores might not be $1/2$ as in a standard randomized experiment. The actual ATE is 0.1, rather small compared to the outcomes range $[14, 76]$, but the treatment effect could be positive and large for a subgroup of subjects and our proposed algorithms can be used to identify them.

Four of our proposed methods are implemented with FDR control at level $\alpha = 0.2$: the Crossfit-I^3 and MaY-I^3 which assume the propensity scores to be $1/2$ for all subjects, and Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$ which estimate the propensity scores. The numbers of identifications by Crossfit-I^3, MaY-I^3, Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$ are $446, 429, 238, 162$. Among them, 234 subjects are commonly identified by Crossfit-I^3 and Crossfit-I$^3_{\widehat{\pi}}$, which control the expected proportion of falsely identifying subjects with zero effect (approximately if the propensity scores are not $1/2$); and 158 subjects are commonly identified by MaY-I^3 and MaY-I$^3_{\widehat{\pi}}$, which control the expected proportion of falsely identifying subjects with nonpositive effect (approximately if the propensity scores are not $1/2$). Compared with the rest subjects, the ones identified as having positive effect tend to have larger values for covariate 8, 21 and smaller values for covariate 6, 15, 17 (see Figure 42).



(a) Boxplot of covariates for subjects identified as nonzero effect (left) versus those not being identified (right).



(b) Boxplot of covariates for subjects identified as positive effect (left) versus those not being identified (right).

Figure 42: Characteristics of identified subjects: they tend to have larger value for variable 8, 21 and smaller value for variable 6, 15, 17, compared with not identified subjects.

5.11 Summary

We discuss the problem of identifying subjects with positive effects. Most existing methods identify *subgroups* with positive treatment effects, and they cannot upper bound the proportion of falsely identified *subjects* within an identified subgroup. In contrast, we propose Crossfit-I^3 with finite-sample FDR

control (i.e., the expected proportion of subjects with zero effect is no larger than α among the identified subjects). One advantage of the Crossfit-I^3 is allowing human interaction — an analyst (or an algorithm) can incorporate various types of prior knowledge and covariates using any working model; she can also adjust the model at any step, potentially improving the identification power. Despite this flexibility, the Crossfit-I^3 achieves valid FDR control. Notably, because Crossfit-I^3 incorporates covariates, it can identify subjects with positive effects, including those not treated.

Our proposed interactive procedure was extended to various settings: from FDR control of zero effects to FDR control of nonpositive effects, from equal and known propensity scores to heterogeneous and unknown propensity scores, from unpaired samples to paired samples, and from FDR control at an individual level to FDR control at a subgroup level.

6 Discussion

Our interactive methods should be contrasted with data-splitting approaches, and have been called "data-carving" to drive home the difference Lei and Fithian [2018]; Lei et al. [2020]. We remark that no test, interactive or otherwise, can be run twice from scratch (with a tweak made the second time to boost power) after the entire data has been examined; this amounts to p-hacking. Our interactive tests are one step towards enabling experts (scientists and statisticians) to work together with statistical models and machine learning algorithms in order to discover scientific insights with rigorous guarantees.

The error control for our interactive procedures is based on the independence properties between the data used for error control and the revealed data for interaction. Such independence may either implied by a specific null hypothesis or be constructed by decomposing or transforming the observed data. Examples include the "railway" masking function in (34) for p-values in multiple testing; and sign and absolute value decomposition in (77) for paired-sample comparison; and transformation of the independence between treatment assignments and observed outcomes in property (104) for identification of positive treatment effects.

Substantial potentials are to be explored in the idea of masking and interactive testing. For example, while this thesis focuses on developing interactive tests via masking in nonparametric settings, we recently realized that these ideas could also be powerful for parametric analysis. We can define decomposition and masked data for many types of canonical distribution such as Gaussian, Beta, Gamma distribution (which include special cases like Exponential distribution and Chi-square distribution), and Bernoulli, Binomial, Poisson distribution, and so on. We anticipate that these ideas might have downstream applications such as model selection, post-selection inference, data privacy, creation of fake datasets, evaluating or comparing machine learning algorithms, and so on.

A Appendix for "Interactive Martingale Tests for the Global Null"

A.1 Error control

This section proves the type-I error control for our proposed methods: the martingale Stouffer test and the interactively ordered martingale test.

A.1.1 Proof of Theorem 1

Proof. Under the global null, because p-values are independent and stochastically larger than the uniform, the transformed p-values $\Phi^{-1}(1-p_i)$ are independent and stochastically smaller than a standard Gaussian. Thus given the uniform bound for a Gaussian increment martingale $u_\alpha(k)$,

$$\begin{aligned} & \mathbb{P}_0\left(\exists k \in \mathbb{N} : \sum_{i=1}^{k} \Phi^{-1}(1-p_i) \geq u_\alpha(k)\right) \\ \leq\ & \mathbb{P}\left(\exists k \in \mathbb{N} : \sum_{i=1}^{k} G_i \geq u_\alpha(k)\right) \\ \leq\ & \alpha, \end{aligned}$$

where G_i for $i \in \mathcal{I}$ are i.i.d. standard Gaussians. By definition the above argument proves the type-I error control. □

A.1.2 Proof of Theorem 3

This proof also implies Theorem 2 since the adaptively ordered martingale test is a special case of the interactively ordered martingale test.

Proof. **Batch setting.** We argue that the sum $\{\sum_{i\in M_k} h(p_i)\}_{k\in\mathcal{I}}$ is a supermartingale with respect to the filtration $\{\mathcal{F}_{k-1}\}_{k\in\mathcal{I}}$. First, the sum $\sum_{i\in M_k} h(p_i)$ is measurable with respect to $\mathcal{F}_{k-1}$ because the random set $M_k = M_{k-1} \cup \{i_k^*\}$ has its distribution defined with respect to $\mathcal{F}_{k-1}$.

Second, we prove that

$$\mathbb{E}(\sum_{i\in M_k} h(p_i) \mid \mathcal{F}_{k-1}) \leq \sum_{i\in M_{k-1}} h(p_i), \tag{139}$$

Because $\mathbb{E}(\sum_{i\in M_k} h(p_i) \mid \mathcal{F}_{k-1}) = \sum_{i\in M_{k-1}} h(p_i) + \mathbb{E}(h(p_{i_k^*}) \mid \mathcal{F}_{k-1})$, condition (139) boils down to proving

$$\mathbb{E}(h(p_{i_k^*}) \mid \mathcal{F}_{k-1}) \leq 0.$$

Since i_k^* and M_{k-1} are $\mathcal{F}_{k-1}$ measurable, and $i_k^* \notin M_{k-1}$, we see that

$$\mathbb{E}(h(p_{i_k^*}) \mid \mathcal{F}_{k-1}) \leq \max_{i\notin M_{k-1}} \mathbb{E}(h(p_i) \mid \mathcal{F}_{k-1}) = \max_{i\notin M_{k-1}} \mathbb{E}(h(p_i) \mid g(p_i)),$$

where the last equation is because the p-values are assumed to be independent of each other and of the covariates x_i under the global null; and thus, $h(p_i) \mid \mathcal{F}_{k-1}$ has the same distribution as $h(p_i) \mid g(p_i)$.

The proof is completed if

$$\mathbb{E}(h(p_i) \mid g(p_i)) \leq 0, \tag{140}$$

for any $i \notin M_{k-1}$. In this case, the sum $\{\sum_{i\in M_k} h(p_i)\}_{k\in\mathcal{I}}$ is a martingale. Also, the increment is stochastically smaller than a Rademacher and following the same argument in Section A.1.1, so the test

using a bound for a Gaussian increment martingale controls the type-I error (because a Rademacher is subGaussian).

We have an intermediate result: the interactively ordered martingale test has type-I error control for any $h(p)$ and $g(p)$ such that condition (140) holds. For a mirror-conservative p-value, the missing bit $h(p_i)$ conditioned on its corresponding masked p-value $g(p_i)$ is stochastically smaller than a fair coin flip:

$$\begin{aligned}&\mathbb{P}_0(h(p_i) = -1 \mid g(p_i) = x) = \frac{f_i(1-x)}{f_i(1-x) + f_i(x)}\\ \geq\ &\frac{f_i(x)}{f_i(1-x) + f_i(x)} = \mathbb{P}_0(h(p_i) = 1 \mid g(p_i) = x),\end{aligned}$$

for any $x \in [0, 0.5]$ (i.e., the range of $g(p_i)$), which implies condition (140) and thus completes the proof.

Online setting. Let the index of the hypothesis that enters the rejection set M_{k-1} be t_k^*. Notice that t_k^* is a stopping time with respect to $\mathcal{F}_{t-1}$ (that is, $\{t_k^* = t\}$ is measurable with respect to $\mathcal{F}_{t-1}$ because we decide whether to include p_t based on $\mathcal{F}_{t-1}$). For a clear notation, define a filtration indexed by k as

$$\mathcal{G}_{k-1} := \mathcal{F}_{t_k^*-1}, \tag{141}$$

denoting all the information available prior to the k-th entered hypothesis. We argue that the sum $\{\sum_{i \in M_k} h(p_i)\}_{k \in \mathcal{I}}$ is a supermartingale with respect to the filtration $\{\mathcal{G}_{k-1}\}_{k \in \mathcal{I}}$. The proof is similar to the above batch setting, where we prove that

$$\mathbb{E}(h(p_{t_k^*}) \mid \mathcal{G}_{k-1}) \leq 0.$$

Since t_k^* is a stopping time with respect to $\mathcal{F}_{t_k^*-1}$, we see that

$$\begin{aligned}&\mathbb{E}(h(p_{t_k^*}) \mid \mathcal{G}_{k-1}) = \mathbb{E}(h(p_{t_k^*}) \mid \mathcal{F}_{t_k^*-1})\\ \leq\ &\max_t \mathbb{E}(h(p_t) \mid \mathcal{F}_{t-1}) = \max_t \mathbb{E}(h(p_t) \mid g(p_t)),\end{aligned}$$

where the last equation is because the p-values are assumed to be independent of each other and of the covariates x_i under the global null; and thus, $h(p_i) \mid \mathcal{F}_{k-1}$ has the same distribution as $h(p_i) \mid g(p_i)$.

The rest of the proof is the same as the batch setting where we show condition (140) holds:

$$\mathbb{E}(h(p_t) \mid g(p_t)) \leq 0,$$

for mirror-conservative p-values. Thus, the sum $\{\sum_{i \in M_k} h(p_i)\}_{k \in \mathcal{I}}$ is a supermartingale with respect to the filtration $\{\mathcal{G}_{k-1}\}_{k \in \mathcal{I}}$. Recall that the increment is stochastically smaller than a Rademacher. Following the same argument in Section A.1.1, the interactively ordered martingale test in the online setting using bound for a Gaussian increment martingale controls the type-I error. □

A.1.3 Error control of the interactively ordered martingale test with railway masking function in Section 2.6

Let the masked p-values defined by the railway function in Section 2.6 be:

$$\widetilde{g}(p) := \min(p, (p + \tfrac{1}{2}) \text{mod} 1)$$

The corresponding interactively ordered martingale test has a valid error control when the p-values have nondecreasing densities under the global null.

Theorem 18. *If under $\mathcal{H}_{\mathcal{G}_0}$, the p-values have nondecreasing densities and are independent of each other and of the covariates x_i, then the interactively ordered martingale test using $\widetilde{g}(p)$ in place of $g(p)$ controls the type-I error at level α.*

Proof. Recall that in Appendix A.1.2, we have an intermediate result: the interactively ordered martingale test has type-I error control for any $h(p)$ and $g(p)$ such that condition (140) holds. For a p-value with a nondecreasing density, the missing bit $h(p_i)$ conditioned on its corresponding masked p-value $\widetilde{g}(p_i)$ is stochastically smaller than a fair coin flip:

$$\begin{aligned}\mathbb{P}_0(h(p_i) = -1 \mid \widetilde{g}(p_i) = x) &= \frac{f_i(x+0.5)}{f_i(x+0.5) + f_i(x)} \\ \geq \frac{f_i(x)}{f_i(x+0.5) + f_i(x)} &= \mathbb{P}_0(h(p_i) = 1 \mid \widetilde{g}(p_i) = x),\end{aligned}$$

for any $x \in [0, 0.5]$ (i.e. the range of $\widetilde{g}(p_i)$), which implies condition (140) and thus completes the proof. □

Remark 9. *The above proof implies that the error control holds as long as under the global null, the p-values satisfy:*

$$f_i(a) \leq f_i(a+0.5) \text{ for all } 0 \leq a \leq 0.5, i \in \mathcal{I},$$

where f_i is the probability mass function of p_i for discrete p-values or the density function otherwise. This condition can be viewed as a third definition of conservativeness in addition to condition (2) and (3) in the main paper. It is not a consequence of condition (2) (take $f(a) = \mathbb{1}(a \leq 0.5) + 4(a - 0.5)\mathbb{1}(a > 0.5)$) or condition (3) (take $f(a) = 4\min(a, 1-a)$), and it does not imply condition (2) and (3) (take $f(a) = 4(0.5-a)\mathbb{1}(a < 0.5) + 4(1-a)\mathbb{1}(0.5 \leq a < 1) + 4\mathbb{1}(a = 1)$). For simplicity, we focus on the p-values with increasing densities in Section 2.6, which are considered as conservative p-values in all three definitions.

A.2 Power guarantees in the batch setting

This section presents the proofs of power guarantees in the batch setting for (1) the batch Stouffer test, (2) the martingale Stouffer test and (3) the interactively ordered martingale test.

A.2.1 Proof of Theorem 4

We divide the proof into two subsections for the batch Stouffer test and the martingale Stouffer test.

The batch Stouffer test

Proof. Define the Z-score for each hypothesis H_i as $Z_i = \Phi^{-1}(1 - p_i)$. Under setting 1 in the main paper of testing Gaussian mean, the Z-score is a Gaussian $Z_i \sim N(\mu_i, 1)$, or written as $N(r_i\mu_i, 1)$ to separate the true nulls from the true non-nulls. Thus, the sum $S_n = \sum_{i=1}^n Z_i$ is also a Gaussian $S_n \sim N\left(\sum_{i=1}^n r_i\mu_i, n\right)$. The power of the batch Stouffer test is

$$\begin{aligned}\mathbb{P}_1\left(\frac{S_n}{\sqrt{n}} \geq \Phi^{-1}(1-\alpha)\right) &= \mathbb{P}_1\left(\frac{S_n - \sum_{i=1}^n r_i\mu_i}{\sqrt{n}} \geq \Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^n r_i\mu_i}{\sqrt{n}}\right) \\ &= 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^n r_i\mu_i}{\sqrt{n}}\right).\end{aligned}$$

A power of at least $1-\beta$ is is equivalent to

$$1-\Phi\left(\Phi^{-1}(1-\alpha)-\frac{\sum_{i=1}^{n} r_i\mu_i}{\sqrt{n}}\right)\geq 1-\beta,$$

which can be rewritten as

$$\sum_{i=1}^{n} r_i\mu_i \geq (\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta))n^{1/2},$$

which is the condition in Theorem 4. □

The martingale Stouffer test

Proof. Following the same proof for $S_n \sim N(r_i\mu_i, 1)$ in Section A.2.1, for any $k = 1,\ldots,n$, $S_k \sim N\left(\sum_{i=1}^{k} r_i\mu_i, k\right)$. The power of the martingale Stouffer test is

$$\begin{aligned}
&\mathbb{P}_1\left(\exists k\in\{1,\ldots,n\} : S_k \geq u_\alpha(k)\right)\\
=\ &\mathbb{P}_1\left(\exists k\in\{1,\ldots,n\} : S_k-\sum_{i=1}^{k} r_i\mu_i \geq u_\alpha(k)-\sum_{i=1}^{k} r_i\mu_i\right),
\end{aligned}$$

The power of martingale Stouffer test is at least $1-\beta$ if

$$\exists k^*\in\{1,\ldots,n\} : u_\alpha(k^*)-\sum_{i=1}^{k^*} r_i\mu_i \leq -u_\beta(k^*) \quad \text{(a sufficient condition)},$$

since under such condition,

$$\begin{aligned}
&\mathbb{P}_1\left(\exists k\in\{1,\ldots,n\} : S_k-\sum_{i=1}^{k} r_i\mu_i \geq u_\alpha(k)-\sum_{i=1}^{k} r_i\mu_i\right)\\
\geq\ &\mathbb{P}_1\left(S_{k^*}-\sum_{i=1}^{k^*} r_i\mu_i \geq u_\alpha(k^*)-\sum_{i=1}^{k^*} r_i\mu_i\right)\\
\geq\ &\mathbb{P}_1\left(S_{k^*}-\sum_{i=1}^{k^*} r_i\mu_i \geq -u_\beta(k^*)\right)\\
\geq\ &\mathbb{P}_1\left(\forall k\in\{1,\ldots,n\} : S_k-\sum_{i=1}^{k} r_i\mu_i \geq -u_\beta(k)\right)\geq 1-\beta.
\end{aligned}$$

The last step holds because Gaussian increment martingale is symmetric so that $-u_\beta(k)$ is a uniform lower bound.

The power of martingale Stouffer test is less than $1-\beta$ if

$$\forall k\in\{1,\ldots,n\} : u_\alpha(k)-\sum_{i=1}^{k} r_i\mu_i \geq u_{1-\beta}(k) \quad \text{(a necessary condition)},$$

since

$$\mathbb{P}_1\left(\exists k \in \{1,\dots,n\} : S_k - \sum_{i=1}^{k} r_i\mu_i \geq u_\alpha(k) - \sum_{i=1}^{k} r_i\mu_i\right)$$
$$\leq \mathbb{P}_1\left(\exists k \in \{1,\dots,n\} : S_k - \sum_{i=1}^{k} r_i\mu_i \geq u_{1-\beta}(k)\right) \leq 1-\beta.$$

Thus, we find a sufficient condition and a necessary condition for the martingale Stouffer test to have $1-\beta$ power. The proof completes by plugging the curved bound in test (7) in the main paper into the conditions. If without further explanation, $u_\alpha(k)$ in rest of the proofs denotes the curved bound. □

A.2.2 Proof of Theorem 5

The adaptively ordered martingale test uses the missing bits $h(p_i)$ for testing, and under no prior knowledge, uses the masked p-values $g(p_i)$ to order the hypotheses. We divide the proof into three steps: (1) derive the power guarantee given a fixed order in Lemma 1; (2) quantify the effect of ordering by masked p-values in Lemma 2, and (3) derive the power guarantee for the adaptively ordered martingale test (Theorem 5).

The power of adaptively ordered martingale test given a fixed order

Lemma 1. *Given a fixed sequence of $\{M_k\}_{k=1}^n$ with the size $|M_k| = k$, the adaptively ordered martingale test with type-I error control α has power at least $1-\beta$ if*

$$\exists k \in \{1,\dots,n\} : \sum_{i\in M_k} (r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1)) \geq \left(C_k^\alpha + C_k^\beta\right) k^{\frac{1}{2}}.$$

where $S_i(1) = \mathbb{P}(h(p_i) = 1 \mid r_i = 1, \{M_k\}_{k=1}^n)$ is a measurement of the "signal strength" from the non-nulls and $S_i(0) = \mathbb{P}(h(p_i) = 1 \mid r_i = 0, \{M_k\}_{k=1}^n)$ is from the nulls. Meanwhile the power is less than $1-\beta$ if

$$\forall k \in \{1,\dots,n\} :$$
$$\sum_{i\in M_k} (r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1)) \leq \left(C_k^\alpha - C_k^{1-\beta}\right) k^{\frac{1}{2}}.$$

Proof. Consider the re-scaled increment $(h(p_{i_k^*}) + 1)/2 \mid \mathcal{F}_k$, which follows a Bernoulli:

$$\frac{h(p_{i_k^*}) + 1}{2} \sim r_i \text{Ber}(S_{i_k^*}(1)) + (1-r_i)\text{Ber}(S_{i_k^*}(0)).$$

So the cumulative sum S_k is a martingale with sub-Gaussian increments after centering, with expected value $\sum_{i\in M_k} (r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1))$. So the power of adaptively ordered martingale test is

$$\mathbb{P}_1\left(\exists k \in \{1,\dots,n\} : S_k \geq u_\alpha(k)\right)$$
$$= \mathbb{P}_1\left(\exists k \in \{1,\dots,n\} : S_k - \sum_{i\in M_k} [r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1)]\right.$$
$$\left. \geq u_\alpha(k) - \sum_{i\in M_k} [r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1)]\right).$$

The proof can be completed by following similar steps in the proof for martingale Stouffer test (Section A.2.1). □

The effect of ordering Define the Z-score as $Z_i = \Phi^{-1}(1-p_i)$ for each hypothesis H_i. Under setting 1 in the main paper, Z_i is a Gaussian with unit variance and mean value μ_i. We consider the simple case where for all the non-nulls $\mu_i = \mu$. The adaptively ordered martingale test orders the hypotheses increasingly by $g(p_i)$, which is equivalent to ordering decreasingly by $|Z_i|$. Following definition (12), the Z-scores for non-nulls have the same distribution as $Z(\mu)$, and $Z_{(j)}(\mu)$ is the Z-score of j-th non-null when they are ordered decreasingly by $|Z_i|$. We describe the effect of ordering by the size of the set M_k right after the j-th non-null enters, denoted as $M(j)$.

Lemma 2. *The size of $M(j)$ follows a Binomial distribution (up to a constant):*

$$|M(j)| \sim j + \text{Bin}\left(N_0, \mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)\right).$$

The size $|M(j)|$ is uniformly upper bounded:

$$\mathbb{P}_1\left(\forall j \in 1,\dots,N_1 : |M(j)| \le j + t_{\beta/N_1}(N_0, q_j)\right) \ge 1-\beta,$$

where $t_{\beta/N_1}(N_0,q_j)$ is β/N_1-th upper quantile of $\text{Bin}\left(N_0, \mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)\right)$.

Remark 10. *Denote $P(\mu) = \mathbb{P}(|Z(0)| \ge |Z(\mu)|)$. The quantile $t_{\beta/N_1}(N_0,q_j)$ is upper bounded by a ratio of $P(\mu)N_0$ (when $P(\mu)N_0 > 1$):*

$$t_{\beta/N_1}(N_0,q_j) \le \frac{2+2\sqrt{2\log(N_1/\beta)}}{N_1\left[\frac{N_1+1-j}{N_1} - P(\mu)\right]^2}\max\{P(\mu)N_0, 1\},$$

for $j = 1,\dots,\lfloor N_1(1-P(\mu))+1\rfloor$.

Proof. In $M(j)$, the number of non-nulls is known as j and the number of nulls is random. The nulls in $M(j)$ should have a higher absolute Z-score than $|Z_{(j)}(\mu)|$. Note that the Z-scores of the nulls are i.i.d. standard Gaussians, so the probability of a null to be in front of the j-th non-null is $\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)$ for any nulls. Thus the number of nulls before the j-th non-null follows a binomial distribution:

$$\sum_{i:r_i=0} 1(|Z_i(0)| > |Z_{(j)}(\mu)|) \sim \text{Bin}\left(N_0, \mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)\right).$$

Thus, the size of $M(j)$ is distributed as

$$|M(j)| \sim j + \text{Bin}\left(N_0, \mathbb{P}(|Z(0)| > |Z(\mu_{\pi_j})|)\right).$$

By the Bonferroni correction, with high probability $|M(j)|$ is upper bounded by

$$\mathbb{P}_1\left(\forall j \in 1,\dots,N_1 : |M(j)| \le j + t_{\beta/N_1}(N_0, q_j)\right) \ge 1-\beta,$$

where $t_{\beta/N_1}(N_0,q_j)$ is β/N_1-th upper quantile of $\text{Bin}\left(N_0, \mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)\right)$.

We further characterize the Binomial quantile $t_{\beta/N_1}(N_0,q_j)$ (proof of Remark 10). The quantile is upper bounded (by Chernoff inequality):

$$\begin{aligned} t_{\beta/N_1}(N_0,q_j) &\le \mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)N_0 + \sqrt{2\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)N_0\log(\tfrac{N_1}{\beta})} \\ &\le (1+\sqrt{2\log(\tfrac{N_1}{\beta})})\max\{\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)N_0, 1\}. \end{aligned}$$

The proof completes by showing that the probability term $\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)$ is upper bounded:

$$\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|) \leq \frac{2P(\mu)}{N_1 \left[\frac{N_1+1-j}{N_1} - P(\mu)\right]^2}. \tag{142}$$

The above bound (142) holds because the event $|Z(0)| > |Z_{(j)}(\mu)|$ can be viewed as comparing the absolute value of $Z(0)$ with N_1 Gaussians $\{Z^i(\mu)\}_{i=1}^{N_1}$ with the same distribution as $Z(\mu)$, and $|Z(0)|$ is bigger than $N_1 - j + 1$ of them. The number of $Z^i(\mu)$ that $|Z(0)| > |Z^i(\mu)|$ follows a binomial distribution, with probability $\mathbb{P}\left(|Z(0)| > |Z(\mu)|\right) := P(\mu)$. Let X be $\mathrm{Bin}(N_1, P(\mu))$ and bound (142) holds because

$$\begin{aligned}
&\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|) = \mathbb{P}(X > N_1 - j + 1) \\
\leq\ & \exp\left\{-\frac{[N_1(1-P(\mu)) - j + 1]^2}{2N_1P(\mu)(1-P(\mu))}\right\} \leq \exp\left\{-\frac{N_1\left[\frac{N_1+1-j}{N_1} - P(\mu)\right]^2}{2P(\mu)}\right\} \\
\leq\ & \frac{2P(\mu)}{N_1\left[\frac{N_1+1-j}{N_1} - P(\mu)\right]^2},
\end{aligned}$$

for $j = 1, \ldots, \lfloor N_1(1-P(\mu)) + 1 \rfloor$. The proof of Remark 10 is completed by plugging bound (142) in the upper bound for $t_{\beta/N_1}(N_0, q_j)$. □

Proof of Theorem 5

Proof. Lemma 1 provides a condition for adaptively ordered martingale test to have at least $1-\beta$ power given any choice of $\{M_k\}_{k=1}^n$, thus when $\{M_k\}_{k=1}^n$ is random, the power is at least $1-\beta$ if

$$\begin{aligned}
&\exists k \in \{1, \ldots, n\}: \\
&\sum_{i \in M_k} (r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1)) \geq \left(C^{\alpha}_{|M_k|} + C^{\beta}_{|M_k|}\right)(|M_k|)^{1/2},
\end{aligned} \tag{143}$$

where $S_i(0)$ and $S_i(1)$ as the probabilities conditioning on M_k are random. Whether the above condition holds is not determinant, and Theorem 5 provides a sufficient condition such that the above condition holds with high probability.

First, for all the nulls,

$$\begin{aligned}
S_i(0) &= \mathbb{P}(h(p_i) > 0 | r_i = 0, \{M_k\}_{k=1}^n) \\
&\overset{(a)}{=} \mathbb{P}(Z_i > 0 | r_i = 0, \{M_k\}_{k=1}^n) \\
&\overset{(b)}{=} \mathbb{P}(Z_i > 0 | r_i = 0) = 0.5,
\end{aligned}$$

where (a) is because by the definition of the Z-score, $h(p_i) > 0$ is equivalent to $Z_i > 0$; and (b) is because $\{M_k\}_{k=1}^n$ is determined by $|Z_i|$ which is independent of $\mathbb{1}(Z_i > 0)$ when $r_i = 0$. Thus, $(2S_i(0)-1)(1-r_i) = 0$ and in the above condition the sum on the left-hand side only increases when a non-null enters M_k. Therefore, the above condition is satisfied if and only if it is satisfied when a non-null enters M_k:

$$\exists j \in \{1, \ldots, N_1\}: \sum_{i \in M(j)} r_i(2S_i(1)-1) \geq \left(C^{\alpha}_{|M(j)|} + C^{\beta}_{|M(j)|}\right)(|M(j)|)^{1/2}.$$

Second, the non-nulls in $M(j)$ are the ones with j highest absolute Z-scores, whose Z-scores are $Z_{(1)}(\mu), \ldots, Z_{(j)}(\mu)$. Thus, $\sum_{i \in M(j)} r_i S_i(1)$ can be expressed as $\sum_{s=1}^{j} \mathbb{P}(Z_{(s)}(\mu) > 0)$, and the above condition can be rewritten as

$$\exists j \in \{1, \ldots, N_1\} : \sum_{s=1}^{j} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right) \geq \left(C^{\alpha}_{|M(j)|} + C^{\beta}_{|M(j)|}\right) (|M(j)|)^{1/2}.$$

The above condition holds with probability at least $1 - \beta$ if

$$\exists j \in \{1, \ldots, N_1\} : \sum_{s=1}^{j} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right) \geq \left(C^{\alpha}_{n} + C^{\beta}_{n}\right) (j + t_{\beta/N_1}(N_0, q_j))^{\frac{1}{2}}, \tag{144}$$

where $C^{\alpha}_{n} + C^{\beta}_{n} \geq C^{\alpha}_{|M(j)|} + C^{\beta}_{|M(j)|}$ and $j + t_{\beta/N_1}(N_0, q_j)$ is the uniform upper bound of $|M(j)|$ by Lemma 2.

Overall when condition (144) as above holds, the probability of failing to reject is less than the sum of (a) the probability that $|M(j)|$ exceeds its upper bound, which is less than β; and (b) the probability of not rejecting when condition (143) is satisfied, which is also less than β; thus the power is at least $1 - 2\beta$. The proof of theorem 5 completes after replacing all β in condition (144) with $\beta/2$. □

A.2.3 Proof of condition (14) in the main paper

Proof. Let $j = N_1/2$ in Theorem 5, the power of adaptively ordered martingale test is at least $1 - \beta$ if

$$\sum_{s=1}^{N_1/2} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right) \geq \left(C^{\alpha}_{n} + C^{\beta/2}_{n}\right) \left(N_1/2 + t_{\beta/(2N_1)}\left(N_0, q_{N_1/2}\right)\right)^{1/2}. \tag{145}$$

First, the left-hand side can be lower bounded by

$$\sum_{s=1}^{N_1/2} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right) \geq N_1/2 \cdot (2\Phi(\mu) - 1) = N_1\Phi(\mu) - N_1/2,$$

since the term $\frac{1}{j}\sum_{s=1}^{j} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right)$ decreases in j and is minimum at $j = N_1$, whose value is

$$\begin{aligned}
\frac{1}{N_1}\sum_{s=1}^{N_1} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right) &= \frac{1}{N_1}\sum_{s=1}^{N_1} \left(2\mathbb{E}(\mathbb{1}(Z_{(s)}(\mu) > 0)) - 1\right) \\
&= \frac{1}{N_1}\left(2\mathbb{E}\left(\sum_{s=1}^{N_1} \mathbb{1}(Z_{(s)}(\mu) > 0)\right) - N_1\right) \\
&= \frac{1}{N_1}\left(2N_1\mathbb{E}\left(\mathbb{1}(Z(\mu) > 0)\right) - N_1\right) = 2\Phi(\mu) - 1.
\end{aligned}$$

Second on the right-hand side, $t_{\beta/(2N_1)}\left(N_0, q_{N_1/2}\right)$ can be upper bounded (by Chernoff inequality):

$$\begin{aligned}
t_{\beta/(2N_1)}\left(N_0, q_{N_1/2}\right) \leq\ & \mathbb{P}(|Z(0)| > |Z_{(N_1/2)}(\mu)|)N_0 \\
& + \sqrt{2\mathbb{P}(|Z(0)| > |Z_{(N_1/2)}(\mu)|)N_0 \log(2N_1/\beta)},
\end{aligned}$$

in which the probability term $\mathbb{P}(|Z(0)| > |Z_{(N_1/2)}(\mu)|)$ can be further upper bounded by

$$\mathbb{P}(|Z(0)| > |Z(\mu_{\pi_{N_1/2}})|) \leq 2 - 2\Phi(\mu),$$

since

$$\begin{aligned}\mathbb{P}(|Z(0)| > |Z(\mu_{\pi_{N_1/2}})|) &\overset{(a)}{\leq} \frac{2P(\mu)}{N_1\left(1 - P(\mu) - \frac{N_1/2-1}{N_1}\right)^2} \\ &\overset{(b)}{\leq} P(\mu) \overset{(c)}{\leq} 2 - 2\Phi(\mu),\end{aligned}$$

where (a) is in the proof of Remark 10 in Section A.2.2; (b) holds because of the condition $N_1 \geq 6\left(C_n^\alpha + C_n^{\beta/2}\right)^2$ and $\mu > 2$ (an assumption we visit later); and (c) is because $P(\mu) = \mathbb{P}(|Z(0)| \geq |Z(\mu)|) = 2\mathbb{P}(Z(0) \geq |Z(\mu)|)$, which is less than $2\mathbb{P}(Z(0) \geq Z(\mu))$.

Plugging the lower bound of the left-hand side and the upper bound of the right-hand side, condition (145) is implied by

$$\begin{aligned}(\Phi(\mu) - \tfrac{1}{2})^2 \geq\ & \left(C_n^\alpha + C_n^{\beta/2}\right)^2 \frac{4\max\{(1-\Phi(\mu))N_0, \sqrt{(1-\Phi(\mu))N_0\log(\frac{2N_1}{\beta})}\}}{N_1^2} \\ &+ \left(C_n^\alpha + C_n^{\beta/2}\right)^2 \frac{N_1/2}{N_1^2}.\end{aligned}$$

Given $\mu > 2$ and $N_1 \geq 6\left(C_n^\alpha + C_n^{\beta/2}\right)^2$, the above condition holds if

$$\begin{aligned}&\frac{1}{(1-\Phi(\mu))} \\ \geq\ & \left(C_n^\alpha + C_n^{\beta/2}\right)^2 \left(\frac{28N_0}{N_1^2}\right) \max\left(1, \left(C_n^\alpha + C_n^{\beta/2}\right)^2 \left(\frac{28\log(\frac{2N_1}{\beta})}{N_1^2}\right)\right).\end{aligned}$$

Given $\mu > 2$ and $N_1 \geq 6\left(C_n^\alpha + C_n^{\beta/2}\right)^2$, indicating $1 - \Phi(\mu) \leq e^{-\mu^2/2}/2$ and $\log(2N_1/\beta) < \frac{N_1}{5}$, we have a sufficient condition of the above condition:

$$2e^{\mu^2/2} \geq \frac{28}{\sqrt{2\pi}}\left(C_n^\alpha + C_n^{\beta/2}\right)^2 \left(\frac{N_0}{N_1^2}\right),$$

which can be written as a condition on μ:

$$\mu \geq \sqrt{2\log\left(\frac{N_0}{N_1^2}\right) + 4\log\left(C_n^\alpha + C_n^{\beta/2}\right) + 3.45}.$$

Finally we complete the proof by noting that the above condition implies the assumption $\mu \geq 2$ when $N_0 > 0.1N_1^2$. □

Remark 11. *Condition (14) in the main paper falls within the "detectable region" derived in the work of Donoho and Jin Donoho and Jin [2015]: for any test for the problem of detecting sparse Guassian mean ($N_1 \leq n^{1/2}$), type-I error α and type-II error β would be big such that $\alpha + \beta \to 1$ when $n \to \infty$ unless*

$$\mu \geq \sqrt{\log\left(\frac{n}{N_1^2}\right)}, \qquad \textit{when } n^{1/4} \leq N_1 \leq n^{1/2}, \tag{146}$$

$$\mu \geq \sqrt{2}(\sqrt{\log n} - \sqrt{\log N_1}), \qquad \textit{when } 1 < N_1 < n^{1/4}. \tag{147}$$

Proof. First note that condition (14) in the main paper indicates

$$\mu \geq \sqrt{2\log\left(\frac{n}{N_1^2}\right)},$$

for any $N_1 \leq n^{1/2}$, since

$$\begin{aligned}
&\sqrt{2\log\left(\frac{N_0}{N_1^2}\right) + 4\log\left(C_n^\alpha + C_n^{\beta/2}\right) + 3.45} \\
\geq\ &\sqrt{2\log\left(\frac{N_0}{N_1^2}\right) + 4\log\left(C_1^1 + C_1^1\right) + 3.45} = \sqrt{2\log\left(\frac{n}{N_1^2} - \frac{1}{N_1}\right) + 8.6} \\
\geq\ &\sqrt{2\log\left(\frac{n}{2N_1^2}\right) + 8.6} \geq \sqrt{2\log\left(\frac{n}{N_1^2}\right)},
\end{aligned}$$

when $2 \leq N_1 \leq n^{1/2}$ and it is obvious when $N_1 = 1$. So when $n^{1/4} \leq N_1 \leq n^{1/2}$, condition (14) is a subset in the detectable region (146).

When $1 < N_1 < n^{1/4}$, denote $N_1 = n^a$ where $0 < a < 1/4$. The detectable region (147) can be written as

$$\mu \geq (1 - \sqrt{a})\sqrt{2\log n},$$

which is implied by condition (14), since

$$\sqrt{2\log\left(\frac{n}{N_1^2}\right)} = \sqrt{1-2a}\sqrt{2\log n} \geq (1-\sqrt{a})\sqrt{2\log n},$$

when $a < 1/4$. Hence condition (14) is a subset of the detectable region (146) and (147). □

A.3 Power guarantees in the online setting

This section proves the power guarantees in the online setting for three methods: the martingale Stouffer test, the adaptively ordered martingale test, and a benchmark, the online Bonferroni method.

A.3.1 Proof of Theorem 6

The power guarantee for the martingale Stouffer test in the online setting follows the same steps as that in the batch setting (Section A.2.1), except that the range of k is changed from $\{1, \ldots, n\}$ to $\{1, 2, \ldots\}$. We present the proof of the power guarantee for the online Bonferroni method as follows.

First, we derive an upper bound on the power of the online Bonferroni test. Recall the Z-score $Z_k = \Phi^{-1}(1 - p_k)$, which follows a Gaussian distribution $Z_k \sim N(r_k\mu_k, 1)$. The power of rejecting the k-th hypothesis at α_k is

$$\mathbb{P}(p_k < \alpha_k) = \mathbb{P}(Z_k > \Phi^{-1}(1-\alpha_k)) = 1 - \Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k],$$

and the overall power of the online Bonferroni is upper bounded by a union of rejecting individual hypotheses:

$$\mathbb{P}(\exists k \in \mathbb{N} : p_k < \alpha_k) \leq \sum_{k=1}^{\infty} \mathbb{P}(p_k < \alpha_k) = \sum_{k=1}^{\infty} 1 - \Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k]. \tag{148}$$

To upper bound the overall power, we claim the following upper bound on individual power of any hypothesis k, which is in the ratio of the individual significance level α_k.

Lemma 3. *Given any constant $C \in (e^{1/4}, 1)$, if the alternative mean is upper bounded:*

$$r_k \mu_k \leq \frac{1}{4\Phi^{-1}(1-\alpha_k)}, \tag{149}$$

the power of rejecting individual hypothesis k is upper bounded:

$$1 - \Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \leq C \cdot \alpha_k,$$

for large k such that $\alpha_k < a(C)$, where the threshold $a(C)$ increases in C. For example, $a(2) > 0.3$.

Proof. Consider the ratio of individual power over α_k:

$$\frac{1 - \Phi\left[\Phi^{-1}(1-\alpha_k) - \frac{1}{4\Phi^{-1}(1-\alpha_k)}\right]}{\alpha_k},$$

which converges to $e^{1/4}$ as $\alpha_k \to 0$ by L'Hospital's rule:

$$\begin{aligned}
&\lim_{\alpha_k \to 0} \frac{1 - \Phi\left[\Phi^{-1}(1-\alpha_k) - \frac{1}{4\Phi^{-1}(1-\alpha_k)}\right]}{\alpha_k} \\
= &\lim_{\alpha_k \to 0} \frac{\phi\left[\Phi^{-1}(1-\alpha_k) - \frac{1}{4\Phi^{-1}(1-\alpha_k)}\right]}{\phi\left[\Phi^{-1}(1-\alpha_k)\right]} \left(1 + \frac{1}{4\left(\Phi^{-1}(1-\alpha_k)\right)^2}\right) = e^{1/4}.
\end{aligned}$$

We observe through simulations that the threshold $a(C) \geq 0.3$ when $C \geq 2$. □

In the following, we derive sufficient conditions for the power of the online Bonferroni to be less than $1-\beta$ (i.e., the complement of necessary conditions to have at least $1-\beta$ power), separately under the case of dense non-nulls and sparse non-nulls.

Proof of Theorem 6. **Dense non-nulls.** First, consider the dense case where the number of non-nulls are infinite, $\sum_{k=1}^{\infty} r_k = \infty$. The power of the online Bonferroni is less than $1-\beta$ when

$$\sum_{k=1}^{\infty} 1 - \Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \leq 1 - \beta,$$

which holds if for each individual hypothesis k with a positive error budget (i.e., $\alpha_k > 0$), the power of rejection is bounded

$$1 - \Phi\left[\Phi^{-1}(1-\alpha_k) - r_k\mu_k\right] \leq \frac{1-\beta}{\alpha}\alpha_k, \tag{150}$$

where the upper bound $\frac{1-\beta}{\alpha}\alpha_k$ is chosen to satisfy two conditions: (a) the overall power is less than $1-\beta$: $\sum_{k=1}^{\infty} \frac{1-\beta}{\alpha}\alpha_k \leq 1-\beta$ and (b) individual power bound is larger than the corresponding error control level, $\frac{1-\beta}{\alpha}\alpha_k > \alpha_k$, so that the above condition is not trivially satisfied in the case of a null: $r_k\mu_k = 0$. By Lemma 3, the above bound on individual power holds when $r_k\mu_k$ satisfy condition (149) and $\alpha_k < 0.3$ (Notice that here the constant in the lemma is $C = \frac{1-\beta}{\alpha} \geq 4$, so threshold $a\,(C) > 0.3$).

To further characterize condition (149) on $r_k\mu_k$, we consider a baseline sequence where $\alpha_k^* = (6/\pi^2)\alpha/k^2$, which sums to α. For an arbitrary sequence $\{\alpha_k\}_{k=1}^{\infty}$ that sums to α, apply the condition for

the baseline sequence, $r_k\mu_k \le \frac{1}{4\Phi^{-1}(1-\alpha_k^*)}$, and the power for each hypothesis k is still upper bounded. Particularly, this upper bound differs by whether $\alpha_k \le \alpha_k^*$ or $\alpha_k > \alpha_k^*$:

$$\begin{aligned}
&1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right]\\
\le\ &1-\Phi\left[\Phi^{-1}(1-\alpha_k^*)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \le C\alpha_k^*, \qquad \text{if } \alpha_k \le \alpha_k^*;\\
&1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right]\\
\le\ &1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k)}\right] \le C\alpha_k, \qquad \text{if } \alpha_k > \alpha_k^*,
\end{aligned}$$

for k such that $\max\{\alpha_k, \alpha_k^*\} \le a(C)$, and hence,

$$1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \le C\max\{\alpha_k^*, \alpha_k\} \le C(\alpha_k^*+\alpha_k).$$

Choose the constant $C = \frac{1-\beta}{2\alpha}$ (with $a(C) > 0.3$), and the overall power is upper bounded by $1-\beta$:

$$\sum_{k=1}^{\infty} 1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \le \frac{1-\beta}{2\alpha}(2\alpha) = 1-\beta,$$

if (a) the significance levels are small: $\max\{\alpha_k, \alpha_k^*\} \le 0.3$ for all $k = 1, 2, \ldots$, which holds since $\alpha \le (1-\beta)/4 \le 0.25$; and (b) the alternative mean $r_k\mu_k$ satisfies condition (149) for the baseline sequence, which holds when

$$r_k\mu_k \le 0.25\left(\sqrt{2\log\left(\frac{k^2}{\alpha}\right)}\right)^{-1},$$

where the bound decreases at the rate of $\left(\sqrt{\log k}\right)^{-1}$.

Sparse non-nulls. Suppose the sequence $\{\alpha_k\}_{k=1}^{\infty}$ is nonincreasing. A stronger necessary condition can be derived if the non-nulls are sparse in the sense that there exists an upper bound M such that $\sum_{k=1}^{\infty} r_k \le M < \infty$. We separately discuss the set of nulls $\{k : r_k = 0\}$, and the set of small and large α_k. Let $k^* = M^2/\alpha$, and define the sets of large and small α_k as $L(k^*) := \{k \le k^* : r_k = 1\}$ and $S(k^*) := \{k > k^* : r_k = 1\}$. The power would be less than $1-\beta$ if

$$\sum_{r_k=0} 1-\Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \le \alpha, \text{ and} \tag{151}$$

$$\sum_{k\in L(k^*)} 1-\Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \le 2\alpha, \text{ and} \tag{152}$$

$$\sum_{k\in S(k^*)} 1-\Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \le 1-\beta-3\alpha. \tag{153}$$

Power bound (151) for the nulls ($r_k = 0$) holds because individual power equals α_k and $\sum_{r_k=0}\alpha_k \le \alpha$. Power bound (152) for large α_k holds if we bound the power of each individual hypothesis $k \in L(k^*)$:

$$1-\Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \le 2\alpha_k,$$

which can be rewritten as

$$r_k\mu_k \leq \Phi^{-1}(1-\alpha_k) - \Phi^{-1}(1-2\alpha_k).$$

Note that the above bound on $r_k\mu_k$ decreases in α_k and that the set of α_k for $k \in L(k^*)$ is lower bounded because $L(k^*)$ has finite number of hypotheses. Thus, the above condition holds if for $k \in L(k^*)$, all $r_k\mu_k$ are smaller than the bound corresponding to the smallest significance level in $L(k^*)$, which is α_{k^*}:

$$r_k\mu_k \leq \Phi^{-1}\left(1-\alpha_{k^*}\right) - \Phi^{-1}\left(1-2\alpha_{k^*}\right),$$

where $k^* = M^2/\alpha$. Notice that $\Phi^{-1}\left(1-x\right)$ is a convex function and its derivative is $-\left(\phi(\Phi^{-1}\left(1-x\right)\right)^{-1}$, so we have

$$\Phi^{-1}\left(1-\alpha_{k^*}\right) - \Phi^{-1}\left(1-2\alpha_{k^*}\right) \geq \left(\phi(\Phi^{-1}\left(1-2\alpha_{k^*}\right)\right)^{-1}\alpha_{k^*} \geq 0.4\sqrt{\alpha_{k^*}},$$

and power bound (152) for large α_k holds when $r_k\mu_k \leq 0.4\sqrt{\alpha_{k^*}}$.

For small α_k, a sufficient condition for the power bound (153) is

$$1 - \Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \leq \frac{1-\beta-3\alpha}{M},$$

for all $k \in S(k^*)$ using the fact that the number of hypotheses in $S(k^*)$ is smaller than M. The above condition can be rewritten as

$$r_k\mu_k \leq \Phi^{-1}(1-\alpha_k) - \Phi^{-1}\left(1-\frac{1-\beta-3\alpha}{M}\right).$$

To characterize the rate of the above bound, recall that the sequence $\{\alpha_k\}_{k=1}^{\infty}$ decreases and sums to α, so $\alpha_k \leq \alpha/k$ for any $k = 1, 2, \ldots$. Thus, the above condition on $r_k\mu_k$ holds when

$$r_k\mu_k \leq \sqrt{\log\left(\frac{k}{4\alpha}\right)} - \sqrt{2\log\left(\frac{M}{2(1-\beta-3\alpha)}\right)},$$

where the threshold increases at the rate of $\sqrt{\log k}$. We note that the above threshold is positive for $k \in S(k^*)$, since $k > k^*$ and $\frac{k}{4\alpha} > \frac{M^2}{4\alpha^2} \geq \frac{M^2}{4(1-\beta-3\alpha)^2}$, so that the condition on $r_k\mu_k$ is nontrivial.

□

We also demonstrate that the necessary condition for dense non-nulls is fairly tight when all the hypotheses are non-null.

Lemma 4. *Suppose the sequence $\{\alpha_k\}_{k=1}^{\infty}$ decreases at a slow rate,*

$$\alpha_1 = 0 \text{ and } \alpha_k = A/[k(\log k)^2] \text{ for } k > 1,$$

with constant $A = \alpha/\left(\sum_{k=2}^{\infty} 1/[k(\log k)^2]\right)$ such that $\sum_{k=1}^{\infty}\alpha_k = \alpha$. The power of the online Bonferroni test is one if all hypotheses are non-null for $k > 1$ and the mean value decreases: $\mu_k = (\log k)^{-1/c}$ for any $c > 2$.

Proof. Let $Z_k = \Phi^{-1}(1-p_k) \sim N(\mu_k, 1)$ and $X_k = Z_k - \mu_k \sim N(0,1)$. The power of the online Bonferroni test is

$$\begin{aligned}\mathbb{P}(\exists k \in \mathbb{N} : Z_k \geq \Phi^{-1}(1-\alpha_k)) &= \mathbb{P}(\exists k \in \mathbb{N} : X_k \geq \Phi^{-1}(1-\alpha_k) - \mu_k) \\ &= 1 - \prod_{k=1}^{\infty}\Phi\left[\Phi^{-1}\left(1-\alpha_k\right) - \mu_k\right]. \end{aligned} \tag{154}$$

Intuitively, the power would not converge to one when $\Phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right] \gtrapprox (1-\alpha_k)$ (the case with $\mu_k = 0$) since $1-\prod_{k=1}^{\infty}(1-\alpha_k) \leq \sum_{k=1}^{\infty}\alpha_k \leq \alpha$, but could be one when $\Phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right] \ll 1-\alpha_k$. To quantify this comparison, we consider the following ratio:

$$b_k := \frac{1-\Phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right]}{\alpha_k},$$

and the power could be one when b_k is large. Indeed, we claim that b_k increases at a rate faster than $\log k$, or equivalently, $(\log k)/b_k \to 0$. It can be verified by L'Hospital's rule:

$$\begin{aligned}\lim_{k\to\infty}(\log k)/b_k = & \lim_{k\to\infty}\frac{\alpha_k \log k}{1-\Phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right]} \\ = & \lim_{k\to\infty}\frac{\phi\left[\Phi^{-1}\left(1-\alpha_k\right)\right]}{\phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right]}\frac{\log k+\frac{\alpha_k}{k}\Big/\frac{\partial\alpha_k}{\partial k}}{1+\phi\left[\Phi^{-1}\left(1-\alpha_k\right)\right]\frac{\partial\mu_k}{\partial k}\Big/\frac{\partial\alpha_k}{\partial k}},\end{aligned}$$

where for large k, we have $\Phi^{-1}\left(1-\alpha_k\right) \geq \sqrt{\log k}$ and

$$\begin{aligned}\frac{\phi\left[\Phi^{-1}\left(1-\alpha_k\right)\right]}{\phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right]} \leq & \ 2\exp\{-(\log k)^{1/2-1/c}\}; \\ \log k+\frac{\alpha_k}{k}\Big/\frac{\partial\alpha_k}{\partial k} \leq & \ 2\log k; \\ 1+\phi\left[\Phi^{-1}\left(1-\alpha_k\right)\right]\frac{\partial\mu_k}{\partial k}\Big/\frac{\partial\alpha_k}{\partial k} \geq & \ 1.\end{aligned}$$

Thus, $\lim_{k\to\infty}(\log k)/b_k \leq \lim_{k\to\infty}\frac{4\log k}{\exp\{(\log k)^{1/2-1/c}\}} = 0$ for any $c > 2$. In other words, we have proved that $b_k/\log k \to \infty$.

The power (154) is one if $\prod_{k=1}^{\infty}\Phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right] = 0$, or equivalently,

$$\sum_{k=1}^{\infty}\log\Phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right] = -\infty, \tag{155}$$

where for large k, we have

$$\begin{aligned}& \log\Phi\left[\Phi^{-1}\left(1-\alpha_k\right)-\mu_k\right] \\ = & \ \log(1-b_k\alpha_k) \leq -b_k\alpha_k \\ \leq & \ -A\log k/[k(\log k)^2] = -A/(k\log k).\end{aligned}$$

Condition (155) holds because $\sum_{k=1}^{\infty}-A/(k\log k) = -\infty$; and thus, we prove that the power of the online Bonferroni test is one.

□

A.3.2 Proof of Theorem 7

Theorem 7 is a simplified version of the following Theorem 19 (by Claim 1). Before stating Theorem 19, we first define the distinction measure $D(c)$ as

$$D(c) = \frac{\mathbb{P}(|Z(\mu)| > c)}{\mathbb{P}(|Z(0)| > c)},$$

where c is the screening parameter in the online adaptively ordered martingale test. Bigger $D(c)$ indicates bigger distinction. Further denote $N_1(k) = \sum_{i=1}^{k} r_i$ as the number of non-nulls after k hypotheses arrive and $N_0(k) = \sum_{i=1}^{k} 1 - r_i$ as for the nulls.

Theorem 19. *The adaptively ordered martingale test with type-I error α and threshold c guarantees $1 - \beta$ power if*

$$\exists k \in \mathbb{N} : (2S(\mu, c) - 1) \left(N_1(k) - \frac{C_k^{\beta/3} \sqrt{N_1(k)}}{2\mathbb{P}(|Z(\mu)| > c)} \right) \geq \frac{C_k^{\alpha} + C_k^{\beta/3}}{\mathbb{P}^{1/2}(|Z(\mu)| > c)} \left[N_1(k) + D^{-1}(c) N_0(k) + \frac{C_k^{\beta/3} k^{1/2}}{2\mathbb{P}(|Z(\mu)| > c)} \right]^{1/2},$$

where $S(\mu; c) = \mathbb{P}(Z(\mu) > 0 \mid |Z(\mu)| > c)$.

Proof. Denote M_k as the set of hypotheses that pass screening ($|Z_i| > c$) after k hypotheses arrive. By extending Lemma 1 from $k = 1, \ldots, n$ to $k = 1, 2, \ldots$, the power of adaptively ordered martingale test is at least $1 - \beta$ if

$$\exists k \in \mathbb{N} : \sum_{i \in M_k} (r_i(2S_i(1) - 1) + (1 - r_i)(2S_i(0) - 1)) \geq \left(C^{\alpha}_{|M_k|} + C^{\beta}_{|M_k|} \right) (|M_k|)^{1/2}, \tag{156}$$

where for the passed non-nulls, $S_i(1) = \mathbb{P}(h(p_i) = 1 \mid r_i = 1, i \in M_i)$, which can be written in terms of Z_i as $\mathbb{P}(Z_i > 0 \mid r_i = 1, |Z_i| > c) = S(\mu, c)$, and for passed the nulls, $S_i(0) = \mathbb{P}(Z_i > 0 \mid r_i = 0, |Z_i| > c) = \mathbb{P}(Z(0) > 0 \mid |Z(0)| > c) = 0.5$. By the lemmas presented below, the right-hand side is upper bounded by

$$|M_k| \leq \mathbb{P}(|Z(\mu)| > c) \left(N_1(k) + D^{-1}(c) N_0(k) \right) + \frac{C_k^{\beta}}{2} k^{1/2},$$

with probability $1 - \beta$ (Lemma 5). The left-hand side is lower bounded by

$$\sum_{i \in M_k} (2S_i(1) - 1) r_i = (2S(\mu, c) - 1) \sum_{i \in M_k} r_i \geq (2S(\mu, c) - 1) \left(\mathbb{P}(|Z(\mu)| > c) N_1(k) - \frac{C_k^{\beta}}{2} \sqrt{N_1(k)} \right),$$

with probability $1 - \beta$ (Lemma 6). The condition in Theorem 19 results from plugging the bounds of both sides into condition (156).

Overall, when the condition in Theorem 19 holds, the probability of failing to reject is less than the sum of (a) the probability that the upper bound for the right-hand side is violated, which is less than $\beta/3$; (b) the probability that the lower bound for the left-hand side is violated, which is less than $\beta/3$; and (c) the probability of not rejecting when condition (156) is satisfied, which is less than $\beta/3$; thus the power is at least $1 - \beta$. □

Lemma 5. *The size of M_k in the online setting is uniformly upper bounded:*

$$\mathbb{P}_1 \left(\forall k \in \mathbb{N} : |M_k| - \mathbb{E}(|M_k|) \leq \frac{C_k^{\beta}}{2} k^{1/2} \right) \geq 1 - \beta,$$

where

$$\mathbb{E}(|M_k|) = \mathbb{P}(|Z(\mu)| > c)\left(N_1(k) + D^{-1}(c)N_0(k)\right).$$

Proof. The probability of a hypothesis H_i passing screening is $\mathbb{P}(|Z(\mu)| > c)$ when H_i is a non-null, and $\mathbb{P}(|Z(0)| > c)$ when H_i is a null. Denote X_i as the indicator of whether H_i passes the screening, then $|M_k| = \sum_{i=1}^{k} X_i$. Because X_i are independent and each X_i is a mixture of two Bernoullis (of value $\{0, 1\}$), the size $|M_k|$ is a martingale with $\frac{1}{4}$-subGaussian increment. Therefore,

$$\mathbb{P}_1\left(\forall k \in \mathbb{N} : |M_k| - \mathbb{E}(|M_k|) \leq \frac{u_\beta(k)}{2}\right) \geq 1 - \beta,$$

where $u_\beta(k)$ is the upper bound for Gaussian increment martingale as test (7) in the main paper. The expected value is

$$\begin{aligned}\mathbb{E}(|M_k|) = &\sum_{i=1}^{k} r_i\mathbb{P}(Z(\mu)| > c) + (1 - r_i)\mathbb{P}(|Z(0)| > c)\\ = &\ \mathbb{P}(|Z(\mu)| > c)\left(N_1(k) + D^{-1}(c)N_0(k)\right),\end{aligned}$$

which completes the proof. □

Lemma 6. *The number of non-nulls in M_k is uniformly lower bounded:*

$$\mathbb{P}_1\left(\forall k \in \mathbb{N}, \quad \sum_{i\in M_k} r_i - \mathbb{E}(\sum_{i\in M_k} r_i) \geq -\frac{C_k^\beta}{2}(N_1(k))^{1/2}\right) \geq 1 - \beta,$$

where

$$\mathbb{E}(\sum_{i\in M_k} r_i) = \mathbb{P}(|Z(\mu)| > c)N_1(k).$$

The proof follows the same steps as in Lemma 5, by considering only the non-nulls, or equivalently assuming $r_i = 1$ for all i.

Claim 1. *The condition of adaptively ordered martingale test to have $1 - \beta$ power in Theorem 7 implies that in Theorem 19.*

Proof. First, the condition in Theorem 19 can be written as a quadratic inequality on $N_1(k)$,

$$\begin{aligned}&\exists k \in \mathbb{N} : (2S(\mu, c) - 1)^2[0.9N_1(k)]^2\\ &\quad \geq \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\mathbb{P}(|Z(\mu)| > c)}\left((1 - D^{-1}(c))N_1(k) + D^{-1}(c)k + \frac{C_k^{\beta/3}k^{1/2}}{2\mathbb{P}(|Z(\mu)| > c)}\right),\end{aligned}$$

by noting that $N_1(k) - \frac{C_k^{\beta/3}\sqrt{N_1(k)}}{2\mathbb{P}(|Z(\mu)|>c)} \geq 0.9N_1(k)$ since the condition in Theorem 7 guarantees $N_1(k) \geq \left(\frac{C_k^{\beta/3}}{0.2\mathbb{P}(|Z(\mu)|>c)}\right)^2$ (a claim we visit later).

Solve the quadratic inequality for $N_1(k)$ to get a sufficient condition of the above one:

$$2N_1(k) \geq \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)}(1 - D^{-1}(c)) + \left\{ \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^4}{\widetilde{S}^2(\mu,c)} \left(1 - D^{-1}(c)\right)^2 + 4\frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)} D^{-1}(c)k + \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)} \frac{C_k^{\beta/3}}{2\mathbb{P}(|Z(\mu)| > c)} k^{1/2} \right\}^{1/2},$$

where $\widetilde{S}(\mu,c) = [0.9(2S(\mu,c) - 1)]^2\mathbb{P}(|Z(\mu)| > c)$ and $D^{-1}(c) = \frac{2\Phi(-c)}{\Phi(\mu-c)+\Phi(-\mu-c)}$. Note that under the square root, the last two terms involving k is upper bounded by

$$\begin{aligned} & 4\frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)} D^{-1}(c)k + \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)} \frac{C_k^{\beta/3}}{2\mathbb{P}(|Z(\mu)| > c)} k^{1/2} \\ = & \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)(\Phi(\mu-c) + \Phi(-\mu-c))} \left(8\Phi(-c)k + \frac{C_k^{\beta/3}}{2} k^{1/2}\right) \\ \leq & \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)(\Phi(\mu-c) + \Phi(-\mu-c))} 9\Phi(-c)k = \frac{9\left(C_k^\alpha + C_k^{\beta/3}\right)^2 D^{-1}(c)}{2\widetilde{S}(\mu,c)} k, \end{aligned}$$

when $k \geq \left(\frac{C_k^{\beta/3}}{2\Phi(-c)}\right)^2$. By the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$, an upper bound on the right-hand side is

$$2\frac{1 - D^{-1}(c)}{\widetilde{S}(\mu,c)} \left(C_k^\alpha + C_k^{\beta/3}\right)^2 + 3(C_k^\alpha + C_k^{\beta/3}) \frac{\sqrt{D^{-1}(c)/2}}{\widetilde{S}^{1/2}(\mu,c)} k^{1/2}.$$

Thus, the above condition on $N_1(k)$ is implied by

$$\exists k \geq \left(\frac{C_k^{\beta/3}}{2\Phi(-c)}\right)^2 : N_1(k) \geq \widetilde{B}(\mu;c) \left(C_k^\alpha + C_k^{\beta/3}\right)^2 + A(\mu;c)(C_k^\alpha + C_k^{\beta/3}) k^{1/2},$$

where $A(\mu;c) = 3/2\frac{\sqrt{D^{-1}(c)/2}}{\widetilde{S}^{1/2}(\mu,c)}$ and $\widetilde{B}(\mu;c) = \frac{1-D^{-1}(c)}{\widetilde{S}(\mu,c)}$.

Finally we review the assumptions made throughout the proof: (a) we assume $N_1(k) \geq \left(\frac{C_k^{\beta/3}}{0.2\mathbb{P}(|Z(\mu)|>c)}\right)^2$, which is implied if $\widetilde{B}(\mu,c)$ is adjusted to $B(\mu,c)$ as defined in Theorem 7; and (b) we assume $k \geq \left(\frac{C_k^{\beta/3}}{2\Phi(-c)}\right)^2$, which holds when $k \geq T(\beta;c)$; adjusting for these assumptions results in the condition in Theorem 7. □

A.4 Choices for the uniform bounds in the martingale Stouffer test

The martingale Stouffer test has the general form:

$$\exists k \in \mathbb{N} : \sum_{i=1}^{k} \Phi^{-1}(1-p_i) \geq u_\alpha(k),$$

where $u_\alpha(k)$ is the uniform bound for a martingale with standard Gaussian increment. We present four bounds from the work of Howard et al. Howard et al. [2020a,b],

1. a linear bound
$$u_\alpha(k) = \sqrt{\frac{-\log\alpha}{2m}}k + \sqrt{\frac{-m\log\alpha}{2}}, \tag{157}$$
where $m \in \mathbb{R}_+$ is a tuning parameter that determines the time at which the bound is tightest: a larger m results in a lower slope but a larger offset, making the bound loose early on.
2. a curved bound from polynomial stitching method
$$u_\alpha(k) = 1.7\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)}. \tag{158}$$
3. a curved bound from discrete mixture method
$$u_\alpha(k) = \inf\left\{s \in \mathcal{R} : \sum_{i=0}^{\infty} \omega_i \exp\{\lambda_i s - \psi(\lambda_i)k\} \geq 1/\alpha\right\}, \tag{159}$$
where $\lambda_i = 1.1^{-(i+1/2)}\lambda_{\max}$ and $\omega_i = 1.1^{-(i+1)}\lambda_{\max} f(1.05\lambda_i)/10$, in which $\lambda_{\max} = \sqrt{2\log\alpha^{-1}}$ and $f(x) = 0.4\frac{\mathbf{1}_{0\leq x\leq\lambda_{\max}}}{x\log^{1.4}(e\lambda_{\max}/x)}$.
4. a curved bound from inverted stitching method (for finite time)
$$u_\alpha(k) = 2.42\sqrt{k\log\log(ek) + 4.7}, \quad k = 1, 2, \ldots, 10^4, \tag{160}$$
where the time limit 10^4 is chosen as the number of hypotheses in the following simulation.

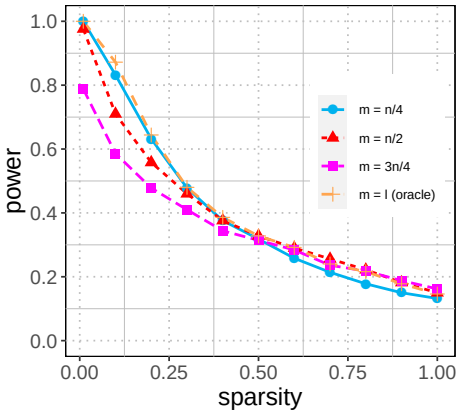
We use simulations to explore two choices in the martingale Stouffer test: (1) the choice of parameter m in the linear bound (157); and (2) the choice among the above four types of bound.

Choice of the parameter m in the linear bound A good choice of parameter m should make the bound tight at where most non-nulls appear; thus, it depends on how the non-nulls distribute. A smaller m results in a faster slope but a tighter bound at front, so it is desired when the non-nulls are gathered at front; and vice versa.

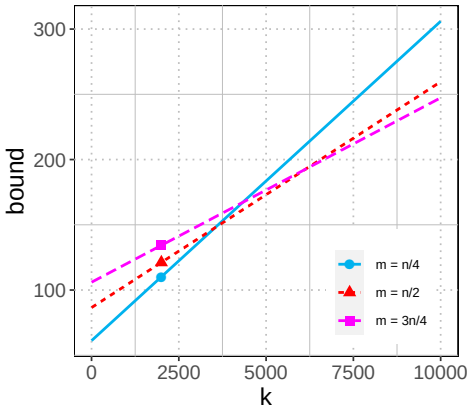
We seek for a robust value of m such that the resulting test has relatively high power under different non-null sparsity. The following constructed simulation is used for exploring bounds in both the martingale Stouffer test and the martingale Fisher test (introduced in Appendix A.5).

Setting 3. *Consider the hypothesis of testing if a Gaussian has zero mean as in Setting 1 in the main paper. In total $n = 10^4$ samples are simulated, with 100 from the non-null distribution $N(1.5, 1)$ and the rest from the null $N(0, 1)$. The non-null sparsity varies by restricting the range where the non-nulls randomly distribute. The non-null range is set as H_1 to H_l and we test values $l = 100, 10^3, 2\times10^3, \ldots, 10^4$. We define the non-null sparisty as $\frac{l}{n}$ and a bigger value indicates a more sparse non-null distribution.*

We compare three choices of $m = n/4, n/2, 3n/4$, with an oracle benchmark of $m = l$ (whose corresponding bound is the tightest right after all the non-nulls appear). The choice of $m = n/4$ leads to the highest power, which is also close to the oracle benchmark (see Figure 43a).



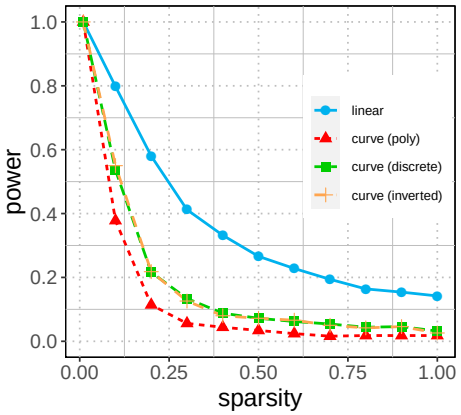
(a) Power of the martingale Stouffer test using the linear bound with different choices of parameter m.



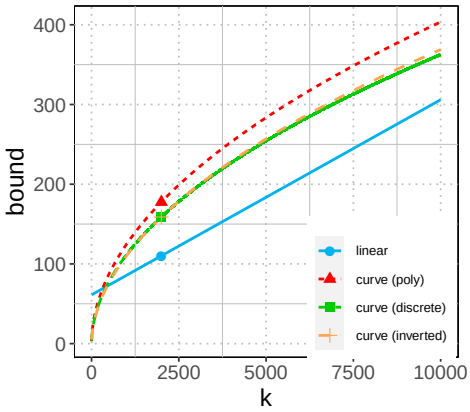
(b) Plot of the linear bound with different choices of parameter m.

Figure 43: Testing martingale Stouffer test using linear bound (157) with different choices of parameter m across varying non-null sparsity. The choice $m = n/4$ leads to the highest power.

Choice of the uniform bound The four bounds presented above can be generally classified as two types: linear and curved. Curved bounds have a slower increasing rate $O(\sqrt{k \log\log(k)})$ than the linear bound, indicating a tighter bound for large enough k, but they are usually looser for small k (Figure 44b).



(a) Power of the martingale Stouffer test with varying non-null sparsity.



(b) Plot of four bounds. The linear bound is much tighter than the curved bounds for most $k \leq 10^4$.

Figure 44: Comparison of the aforementioned four bounds (157)-(160) for the martingale Stouffer test.

Under the batch setting where the number of hypotheses n is finite, we use the simulation setting 3, and the linear bound (157) (with $m = n/4$) results in the highest power (Figure 44a). Similar to tuning the parameter m in the linear bound, we explored to tune the implicit parameters in the curved bound, and yet the linear bound still has the highest power. However, under the online setting where new hypotheses keep arriving, the tests with curved bounds are expected to need less time (number of hypotheses) on average to reach rejection.

A.5 Martingale Fisher test

The batch test by Fisher Fisher [1992] calculates $S_n = -2\sum_{i=1}^{n} \log p_i$. Since the distribution of S_n under the global null is χ^2_{2n} (chi-square with $2n$ degree of freedom), the batch test rejects when S_n is bigger than the $1-\alpha$ quantile for χ^2_{2n}. To design the martingale test, simply observe that $\{S_k\}_{k\in\mathcal{I}}$ is a martingale whose increments $f(pi) = -2\log p_i$ are χ^2_2 under the global null (after centering as $S_k - 2k$). Similar to the martingale Stouffer test, there are several types of uniform boundaries $u_\alpha(k)$ for chi-square increment martingales from the work of Howard et al. Howard et al. [2020a,b]. We present two types: a sub-exponential (linear) boundary, and a sub-Gamma (curved) boundary. The general form of the martingale Fisher test rejects the global null if

$$\exists k \in \mathbb{N} : -2\sum_{i=1}^{k} \log p_i - 2k \geq u_\alpha(k), \tag{161}$$

where examples of $u_\alpha(k)$ include

1. a sub-exponential linear boundary

$$u_\alpha(k) = \left(\left(\frac{1.41m}{x_{m,\alpha}} + 2\right)\log\left(1 + \frac{1.41x_{m,\alpha}}{m}\right) - 2\right)(k-m) + 2.82x_{m,\alpha}, \tag{162}$$

   where $x_{m,\alpha} = \min\left\{x : \exp\left\{-0.71x + \frac{m}{2}\log(1+\frac{1.41x}{m})\right\} \leq \alpha\right\}$; and

2. a sub-Gamma curved boundary

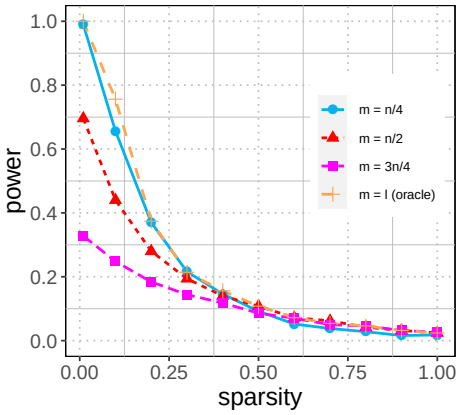
$$\begin{aligned} u_\alpha(k) = \; & 4.07\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)} \\ & + 9.66\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right). \end{aligned} \tag{163}$$

The linear bound contains a parameter m with the same interpretation as m in the linear bound (6) for martingale Stouffer test (in the main paper): it determines the time at which the bound is tightest — a larger m results in a lower slope but a larger offset, making the bound loose early on. Based on the simulation results in Figure 45a, we suggest a default value of $m = n/4$ if the number of hypotheses n is finite, but it should be chosen based on the time by which we expect to have encountered most non-nulls (if any).

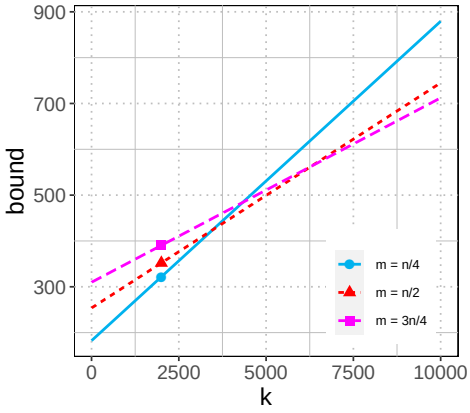
The power of the martingale Fisher test using linear and curved bounds are compared under different non-null sparsity (using simulation setting 3). The curve bound loses power quickly when non-null is rather sparse (see Figure 46a), consistent with the comparison between linear and curved bounds for the martingale Stouffer test in Appendix A.4.

A.6 Martingale chi-squared test

The chi-squared test calculates $S_n = \sum_{i=1}^{n}\left[\Phi^{-1}(1-p_i)\right]^2$. Since the distribution of S_n under the global null is χ^2_n (a chi-square with n degrees of freedom), the batch test rejects when S_n is bigger than the $1-\alpha$ quantile for χ^2_n. To design the martingale test, simply observe that $\{S_k - k\}_{k\in\mathcal{I}}$ is a martingale, whose increment $\left[\Phi^{-1}(1-p_i)\right]^2 - 1$ is distributed as χ^2_1 (minus one) under the global null. Similar to the martingale Stouffer test and martingale Fisher test (in Appendix A.4 and A.5), there are several linear and

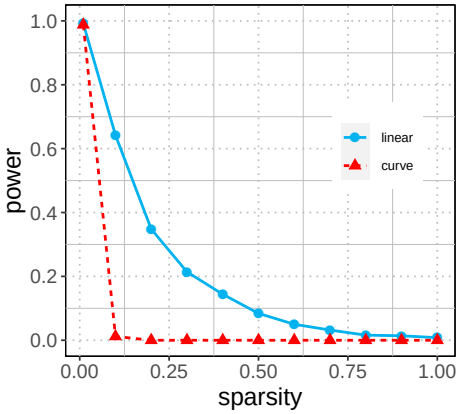


(a) Power of martingale Fisher test using the linear bound with different choices of parameter m.

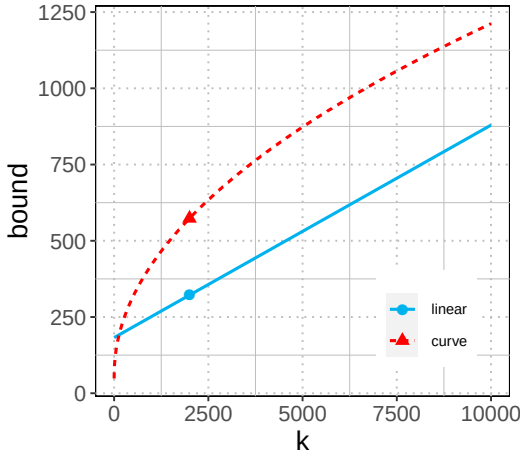


(b) Plot of the linear bound with different choices of parameter m.

Figure 45: Testing the martingale Fisher test using the linear bound (162) with different choices of parameter m across varying non-null sparsity. The choice $m = n/4$ leads to the highest power.



(a) Power of the martingale Fisher test with varying sparsity score.



(b) Plot of two bounds. The linear bound ($m = n/4$) is tighter for most $k \leq 10^4$.

Figure 46: Comparison of the aforementioned two bounds (162) and (163) for the martingale Fisher test.

curved boundaries $u_\alpha(k)$ for chi-square increment martingales from the work of Howard et al. Howard et al. [2020a,b]. We present two types: a sub-exponential (linear) boundary, and a sub-Gamma (curved) boundary. The general form of the martingale chi-square test rejects the global null if

$$\exists k \in \mathbb{N} : \sum_{i=1}^{k} \left[\Phi^{-1}(1-p_i)\right]^2 - k \geq u_\alpha(k), \tag{164}$$

where examples of $u_\alpha(k)$ include

1. a sub-exponential linear boundary

$$u_\alpha(k) = \left(\left(\frac{m}{2x_{m,\alpha}} + 1\right)\log\left(1 + \frac{2x_{m,\alpha}}{m}\right) - 1\right)(k-m) + 2x_{m,\alpha}, \tag{165}$$

where $x_{m,\alpha} = \min\left\{x : \exp\left\{-\frac{x}{2} + \frac{m}{4}\log(1+\frac{2x}{m})\right\} \leq \alpha\right\}$; and

2. a sub-Gamma curved boundary

$$u_\alpha(k) = 3.42\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)} + 9.66\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right). \tag{166}$$

We expect the discussions on parameter m in the linear bound and on the comparison between the linear and curved bounds to be similar to that in the martingale Stouffer test (Appendix A.4) and the martingale Fisher test (Appendix A.5). If testing the martingale chi-squared test by the same numerical experiment in Setting 3, $m = n/4$ should lead to high power for various degrees of sparsity; and the linear bound should be tighter than the curved bound for most time $k \leq 10^4$, and hence lead to higher power when non-null is rather sparse.

A.7 Bayesian modeling for the posterior probability of being non-null

Modeling the posterior probabilities of being non-null Define the Z-score for hypothesis H_i be $Z_i = \Phi^{-1}(1 - p_i)$. Instead of modeling the p-values, we choose to model the Z-scores since under setting 1 in the main paper they are distributed as a Gaussian either under the null or the alternative:

$$H_0 : Z_i \sim N(0, 1) \text{ versus } H_1 : Z_i \sim N(\mu, 1),$$

where μ is the mean value for all the non-nulls. We model Z_i by a mixture of Guassians:

$$Z_i \sim (1 - q_i)N(0, 1) + q_i N(\mu, 1), \text{ with } q_i \sim \text{Bernoulli}(\pi_i),$$

where q_i is the indicator of whether the hypothesis H_i is a true non-null.

The non-null structures are imposed by the constraints on non-null probability π_i. In our examples, the blocked non-null structure is encoded by fitting non-null probabilities π_i as a smooth function of the hypothesis position (covariates) x_i, specifically as a logistic regression model on a spline basis:

$$\pi_i = \pi_\beta(x_i) = \frac{1}{1 + \exp(-\beta\phi(x_i))}, \tag{167}$$

where $\phi(x_i)$ is a spline basis. The hierarchical structure is imposed by a partial ordering constraint on π_i:

$$\pi_i \geq \pi_j, \quad \text{if } i \text{ is the parent of } j, \tag{168}$$

when the probability of being non-null decreases down the tree ($\pi_i \geq \pi_j$ if the probability increases).

An EM framework for the posterior probabilities of being non-null An EM algorithm is used to train the model because masked p-values are modeled. Specifically, we treat p-values as the hidden variables, and the masked p-values $g(p)$ as observed. In terms of the Z-score Z_i, Z_i is a hidden variable and the observed variable $\widetilde{Z}_i$ is its absolute value $|Z_i|$ (if p_i is masked).

Define a sequence of hypothetical labels $w_i = \mathbb{1}(Z_i = \widetilde{Z}_i)$, and the likelihood of data $(\widetilde{Z}_i, w_i, q_i)$ is

$$\begin{aligned} l(\widetilde{Z}_i, w_i, q_i) = {} & w_i q_i \log(\pi_i \phi(\widetilde{Z}_i - \mu)) + w_i(1 - q_i)\log((1 - \pi_i)\phi(\widetilde{Z}_i)) \\ & + (1 - w_i)q_i \log(\pi_i \phi(-\widetilde{Z}_i - \mu)) \\ & + (1 - w_i)(1 - q_i)\log((1 - \pi_i)\phi(-\widetilde{Z}_i)), \end{aligned}$$

where $\phi(\cdot)$ is the PDF of a standard Gaussian.

The E-step updates w_i, q_i. Notice that w_i and q_i are not independent, so we update the joint distribution of (w_i, q_i), namely parameters

$$w_i q_i =: a_i, \quad w_i(1-q_i) =: b_i, \quad (1-w_i)q_i =: c_i, \quad (1-w_i)(1-q_i) =: d_i,$$

where $a_i + b_i + c_i + d_i = 1$. For a simple expression of the updates, we define

$$\begin{aligned} L\left(\widetilde{Z}_i, \mu, \pi_i\right) := \; & \pi_i \phi(\widetilde{Z}_i - \mu)) + (1-\pi_i)\phi(\widetilde{Z}_i) \\ & + \pi_i \phi(-\widetilde{Z}_i - \mu) + (1-\pi_i)\phi(-\widetilde{Z}_i). \end{aligned}$$

For hypothesis i whose p-value is masked, the updates are

$$\begin{aligned} a_{i,\text{new}} = \; & \mathbb{E}[w_i q_i \mid \widetilde{Z}_i] = \frac{\pi_i \phi(\widetilde{Z}_i - \mu)}{L\left(\widetilde{Z}_i, \mu, \pi_i\right)}; \\ b_{i,\text{new}} = \; & \mathbb{E}[w_i(1-q_i) \mid \widetilde{Z}_i] = \frac{(1-\pi_i)\phi(\widetilde{Z}_i)}{L\left(\widetilde{Z}_i, \mu, \pi_i\right)}; \\ c_{i,\text{new}} = \; & \mathbb{E}[(1-w_i)q_i \mid \widetilde{Z}_i] = \frac{\pi_i \phi(-\widetilde{Z}_i - \mu)}{L\left(\widetilde{Z}_i, \mu, \pi_i\right)}; \\ d_{i,\text{new}} = \; & \mathbb{E}[(1-w_i)(1-q_i) \mid \widetilde{Z}_i] = \frac{(1-\pi_i)\phi(-\widetilde{Z}_i)}{L\left(\widetilde{Z}_i, \mu, \pi_i\right)}. \end{aligned}$$

If the p-value is unmasked for hypothesis i, we have $w_i = 1$ and the updates are

$$\begin{aligned} a_{i,\text{new}} = \; & \left(1 + \frac{(1-\pi_i)\phi\left(\widetilde{Z}_i\right)}{\pi_i \phi\left(\widetilde{Z}_i - \mu\right)}\right)^{-1}; \\ b_{i,\text{new}} = \; & 1 - a_{i,\text{new}}; \quad c_{i,\text{new}} = 0; \quad d_{i,\text{new}} = 0. \end{aligned}$$

In the M-step, parameters μ and π_i are updated. The update for μ is

$$\mu_{\text{new}} = \operatorname*{argmin}_{\mu} \sum_i l(\widetilde{Z}_i) = \frac{\sum (a_i - c_i)\widetilde{Z}_i}{\sum (a_i + c_i)}.$$

The update for π_i depends on the non-null structure, which encodes different constraints on π_i. Under the block non-null structure, updating π_i corresponds to updating β in model (167) for $\pi_\beta(x_i)$. The update is equivalent to fitting $a_i + c_i$ by a logistic regression:

$$\begin{aligned} (\beta_{\text{new}}) = \; & \operatorname*{argmax}_{\beta} \sum_i (a_i + c_i) \log \pi_\beta(x_i) + (b_i + d_i) \log(1 - \pi_\beta(x_i)) \\ = \; & \operatorname*{argmax}_{\beta} \sum_i (a_i + c_i) \log \pi_\beta(x_i) + (1 - a_i - c_i) \log(1 - \pi_\beta(x_i)), \end{aligned}$$

and $\pi_{i,\text{new}} = \pi_{\beta_{\text{new}}}(x_i)$. Under the hierarchical structure, updating π_i is equivalent to fitting a partial isotonic regression on $a_i + c_i$ (Barlow Barlow and Brunk [1972], Theorem 3.1 and Robertson Robertson et al. [1988], Theorem 1.5.1):

$$\begin{aligned}(\pi_{i,\text{new}}) = & \underset{\text{partial ordered}\{\pi_i\}}{\text{argmax}} \sum_i (a_i + c_i)\log\pi_i + (1 - a_i - c_i)\log(1 - \pi_i) \\ = & \underset{\text{partial ordered}\{\pi_i\}}{\text{argmin}} \sum_i (a_i + c_i - \pi_i)^2,\end{aligned}$$

where the partial ordering is defined in statement (168).

Suppose we wish to model the alternative mean μ differently for individual hypotheses. In that case, we can think of the alternative mean as a parametric function of the covariates: $\mu_i = \mu_\gamma(x_i)$ where the vector γ denotes the parameters. A simple example is a linear function: $\mu_\gamma(x_i) = \gamma^T x_i$. The updates in the E-step is the same as above with μ replaced by $\mu_\gamma(x_i)$. In the M-step, the update for μ_i corresponds to the update for γ:

$$(\gamma_{\text{new}}) = \underset{\gamma}{\text{argmax}} \sum_i a_i \left(\widetilde{Z}_i - \mu_\gamma(x_i)\right)^2 + c_i \left(-\widetilde{Z}_i - \mu_\gamma(x_i)\right)^2,$$

which is equivalent to the solution of a least square regression to a set of pseudo responses $\{\widetilde{Z_1}, \ldots, \widetilde{Z_n}, -\widetilde{Z_1}, \ldots - \widetilde{Z_n}\}$ with weights $\{a_1, \ldots, a_n, c_1, \ldots, c_n\}$. We use the EM algorithm with constant μ for the experiments in our paper, because it tends to be robust to heterogeneous alternative mean values in simulations.

A.8 Comparison with alternative methods

We compared the interactive test with the adaptive weighted Fisher test (AW-Fisher) and weighted Higher Criticism (weighted-HC) in the example of a grid of hypotheses. Our simulation considers a small grid (10×10) because the AW-Fisher test has a very high computational cost. We used the R package `AWFisher` by Huo et al. (2020) Huo et al. [2020], which refers to a base library of null distributions for cases with less than 100 hypotheses; it took 6373.5 CPU hours using AMD Opteron(tm) Processor (1.4GHz) to complete the base library. Without such a base library, the computational complexity of the AW-Fisher test is $\mathcal{O}(2^n)$, and roughly $\mathcal{O}(n\log(n))$ for our interactive test.

As described in Section 2.5.1, we simulated a non-null cluster is in the center of the hypothesis grid. The weights in HC use the oracle information of the non-null position and is set to 1 for the non-nulls and 0.5 for others. Since we have included several simulations to compare the interactively ordered martingale test with martingale Stouffer test and Stouffer's test in Section 2.5, above in Figure 47, we only focus on the comparison among the interactive test, AW-Fisher and weighted-HC. Although the AW-Fisher test achieves similar power as the interactively ordered martingale test, it has very high computational cost as described above. Also, we remark that one main advantage of the interactive test we propose is that it can incorporate various types of prior knowledge and covariates in a data-dependent way. Meanwhile, most existing methods require the analyst to commit to one structure or prior knowledge before observing the p-values. For example, the weighted-HC might achieve higher power with a different set of weights, but the weights need to be specified ahead of time.

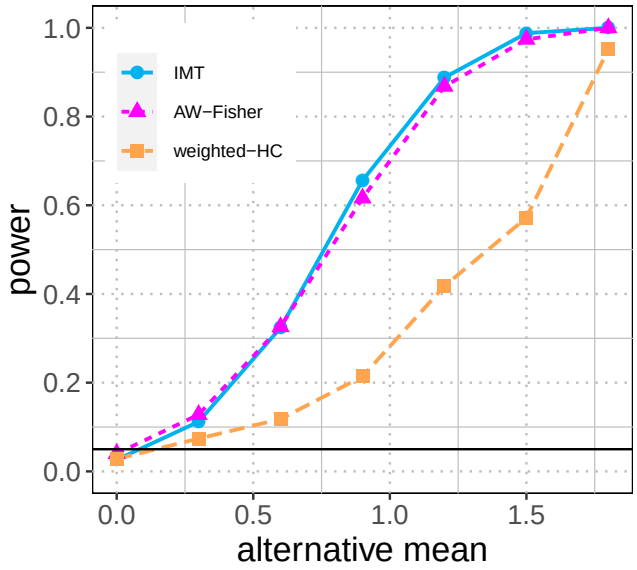


Figure 47: Power of the interactively ordered martingale test (IMT), AW-Fisher, and weighted-HC when the non-null cluster is in the center of a 10×10 grid. IMT and AW-Fisher both have high power, but the AW-Fisher has a high computational cost.

B Appendix for "Familywise Error Rate Control by Interactive Unmasking"

B.1 Distribution of the null p-values

With tent masking, error control holds for null p-values whose distribution satisfies a property called *mirror-conservativeness*:

$$f(a) \leq f\left(1 - \frac{1-p_*}{p_*}a\right), \quad \text{for all } 0 \leq a \leq p_*, \tag{169}$$

where f is the probability mass function of P for discrete p-values or the density function otherwise, and p_* is the parameter in Algorithm 5 (see proof in Appendix B.2). The mirror-conservativeness is first proposed by Lei and Fithian [2018] in the case of $p_* = 0.5$. A more commonly used notion of conservativeness is that p-values are stochastically larger than uniform:

$$\mathbb{P}(P \leq a) \leq a, \quad \text{for all } 0 \leq a \leq 1,$$

which neither implies nor is implied by the mirror-conservativeness.

A sufficent condition of the mirror-conservativeness is that the p-values have non-decreasing densities. For example, consider a one-dimensional exponential family and the hypotheses to test the value of its parameter θ:

$$H_0 : \theta \leq \theta_0, \quad \text{versus} \quad H_1 : \theta > \theta_0,$$

where θ_0 is a prespecified constant. The p-value calculated from the uniformly most powerful test is shown to have a nondecreasing density [Zhao et al., 2019]; thus, it satisfies the mirror-conservativeness. The conservative nulls described in Section 3.4.1 also fall into the above category where the exponential family is Gaussian, and the parameter is the mean value. Indeed, when the p-values have non-decreasing

densities, the i-FWER test also has a valid error control using alternative masking functions as proposed in Section 3.4 (see proof in Appendix B.6).

B.2 Proof of Theorem 8

The main idea of the proof is that the missing bits $h(P_i)$ of nulls are coin flips with probability p_* to be heads, so the number of false rejections (i.e. the number of nulls with $h(P_i) = 1$ before the number of hypotheses with $h(P_i) = -1$ reaches a fixed number) is stochastically dominated by a negative binomial distribution. There are two main challenges. First, the interaction uses unmasked p-value information to reorder $h(P_i)$, so it is not trivial to show that the reordered $h(P_i)$ preserve the same distribution as that before ordering. Second, our procedure runs backward to find the first time that the number of hypotheses with negative $h(P_i)$ is below a fixed number, which differs from the standard description of a negative binomial distribution.

B.2.1 Missing bits after interactive ordering

We first study the effect of interaction. Imagine that Algorithm 5 does not have a stopping rule and generates a full sequence of $\mathcal{R}_t$ for $t = 0, 1, \dots n$, where $\mathcal{R}_0 = [n]$ and $\mathcal{R}_n = \emptyset$. It leads to an ordered sequence of $h(P_i)$:

$$h(P_{\pi_1}), h(P_{\pi_2}), \dots, h(P_{\pi_n}),$$

where π_n is the index of the first excluded hypothesis and π_j denotes the index of the hypothesis excluded at step $n - j + 1$, that is $\pi_j = \mathcal{R}_{n-j} \backslash \mathcal{R}_{n-j+1}$.

Lemma 7. *Suppose the null p-values are uniformly distributed and all the hypotheses are nulls, then for any $j = 1, \dots, n$,*

$$\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right)\right] = p_*,$$

and $\{\mathbb{1}\left(h(P_{\pi_j}) = 1\right)\}_{j=1}^n$ are mutually independent.

Proof. Recall that the available information for the analyst to choose π_j is $\mathcal{F}_{n-j} = \sigma\Big(\{x_i, g(P_i)\}_{i=1}^n, \{P_i\}_{i \notin \mathcal{R}_{n-j}}\Big)$. First, consider the conditional expectation:

$$\begin{aligned}
& \mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right)\middle|\mathcal{F}_{n-j}\right] \\
= & \sum_{i\in[n]} \mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right)\middle|\pi_j = i, \mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i|\mathcal{F}_{n-j}\right) \\
\overset{(a)}{=} & \sum_{i\in\mathcal{R}_{n-j}} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right)|\pi_j = i, \mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i|\mathcal{F}_{n-j}\right) \\
\overset{(b)}{=} & \sum_{i\in\mathcal{R}_{n-j}} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right)|\mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i|\mathcal{F}_{n-j}\right) \\
\overset{(c)}{=} & \sum_{i\in\mathcal{R}_{n-j}} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right)\right] \mathbb{P}\left(\pi_j = i|\mathcal{F}_{n-j}\right) \\
= & \ p_* \sum_{i\in\mathcal{R}_{n-j}} \mathbb{P}\left(\pi_j = i|\mathcal{F}_{n-j}\right) = p_*,
\end{aligned} \tag{170}$$

where equation (a) narrows down the choice of i because $\mathbb{P}(\pi_j = i \mid \mathcal{F}_{n-j}) = 0$ for any $i \notin \mathcal{R}_{n-j}$; equation (b) drops the condition of $\pi_j = i$ because π_j is measurable with respect to $\mathcal{F}_{n-j}$; and equation (c)

drops the condition $\mathcal{F}_{n-j}$ because by the independence assumptions in Theorem 8, $h(P_i)$ is independent of $\mathcal{F}_{n-j}$ for any $i \in \mathcal{R}_{n-j}$.

Therefore, by the law of iterated expectations, we prove the claim on expected value:

$$\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}\right]\right] = p_*.$$

For mutual independence, we can show that for any $1 \le k < j \le n$, $\mathbb{1}\left(h(P_{\pi_k})=1\right)$ is independent of $\mathbb{1}\left(h(P_{\pi_j})=1\right)$. Consider the conditional expectation:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_k})=1\right)\middle|\mathbb{1}\left(h(P_{\pi_j})=1\right)\right] \\
=\ &\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_k})=1\right)\middle|\mathcal{F}_{n-k}, \mathbb{1}\left(h(P_{\pi_j})=1\right)\right]\middle|\mathbb{1}\left(h(P_{\pi_j})=1\right)\right] \\
&\text{(note that } \mathbb{1}\left(h(P_{\pi_j})=1\right) \text{ is measurable with respect to } \mathcal{F}_{n-k}) \\
=\ &\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_k})=1\right)|\mathcal{F}_{n-k}\right]\middle|\mathbb{1}\left(h(P_{\pi_j})=1\right)\right] \\
&\text{(use equation (170) for the conditional expectation)} \\
=\ &\mathbb{E}\left[p_* \mid \mathbb{1}\left(h(P_{\pi_j})=1\right)\right] = p_*.
\end{aligned}$$

It follows that $\mathbb{1}\left(h(P_{\pi_k})=1\right) \mid \mathbb{1}\left(h(P_{\pi_j})=1\right)$ is a Bernoulli with parameter p_*, same as the marginal distribution of $\mathbb{1}\left(h(P_{\pi_k})=1\right)$; thus, $\mathbb{1}\left(h(P_{\pi_k})=1\right)$ is independent of $\mathbb{1}\left(h(P_{\pi_j})=1\right)$ for any $1 \le k < j \le n$ as stated in the Lemma. □

Corollary 3. *Suppose the null p-values are uniformly distributed and there may exist non-nulls. For any* $j = 1, \ldots, n$,

$$\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\left\{\mathbb{1}\left(h(P_{\pi_k})=1\right)\right\}_{k=j+1}^{n}, \left\{\mathbb{1}\left(\pi_k \in \mathcal{H}_0\right)\right\}_{k=j+1}^{n}, \pi_j \in \mathcal{H}_0\right] = p_*,$$

where $\{\pi_k\}_{k=j+1}^{n}$ *represents the hypotheses excluded before* π_j.

Proof. Denote the condition $\sigma\left(\left\{\mathbb{1}\left(h(P_{\pi_k})=1\right)\right\}_{k=j+1}^{n}, \left\{\mathbb{1}\left(\pi_k \in \mathcal{H}_0\right)\right\}_{k=j+1}^{n}\right)$ as $\mathcal{F}_{n-j}^{h}$. The proof is similar to Lemma 7. First, consider the expectation conditional on $\mathcal{F}_{n-j}$:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \\
=\ &\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \quad \text{(since } \mathcal{F}_{n-j}^{h} \text{ is a subset of } \mathcal{F}_{n-j}) \\
=\ &\sum_{i\in[n]} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right) \mid \pi_j = i, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) \\
=\ &\sum_{i\in\mathcal{R}_{n-j}\cap\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right) \mid \pi_j = i, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right] \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) \\
=\ &\sum_{i\in\mathcal{R}_{n-j}\cap\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i)=1\right)|\mathcal{F}_{n-j}\right] \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) \\
=\ &p_* \sum_{i\in\mathcal{R}_{n-j}\cap\mathcal{H}_0} \mathbb{P}(\pi_j = i \mid \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}) = p_*,
\end{aligned} \tag{171}$$

where we use the same technics of proving equation (170).

Thus, by the law of iterated expectations, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0\right] \\
=\ &\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j})=1\right)\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \mathcal{F}_{n-j}\right]\middle|\mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0\right] = p_*,
\end{aligned}$$

which completes the proof. □

Corollary 4. *Suppose the null p-values can be mirror-conservative as defined in (169) and there may exist non-nulls, then for any $j = 1, \ldots, n$,*

$$\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \{\mathbb{1}\left(h(P_{\pi_k}) = 1\right)\}_{k=j+1}^{n}, \{\mathbb{1}\left(\pi_k \in \mathcal{H}_0\right)\}_{k=j+1}^{n}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right] \leq p_*,$$

where $\{g(P_{\pi_k})\}_{k=1}^{n}$ denotes $g(P)$ for all the hypotheses (excluded or not).

Proof. First, we claim that a mirror-conservative p-value P satisfies that

$$\mathbb{E}\left[\mathbb{1}\left(h(P) = 1\right) \mid g(P)\right] \leq p_*, \tag{172}$$

since for every $a \in (0, p_*)$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P) = 1\right) \mid g(P) = a\right] \\
=\ & \frac{p_* f(a)}{p_* f(a) + (1 - p_*) f\left(1 - \frac{1-p_*}{p_*} a\right)} \\
=\ & \frac{p_*}{p_* + (1 - p_*) f\left(1 - \frac{1-p_*}{p_*} a\right) / f(a)} \leq p_*,
\end{aligned}$$

where recall that f is the probability mass function of P for discrete p-values or the density function otherwise. The last inequality comes from the definition of mirror-conservativeness in (169). The rest of the proof is similar to Corollary 3, where we first condition on $\mathcal{F}_{n-j}$:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right] \\
=\ & \sum_{i \in \mathcal{R}_{n-i} \cap \mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right) \mid \mathcal{F}_{n-j}\right] \mathbb{P}\left(\pi_j = i \middle| \mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right) \\
\overset{(a)}{=}\ & \sum_{i \in \mathcal{R}_{n-i} \cap \mathcal{H}_0} \mathbb{E}\left[\mathbb{1}\left(h(P_i) = 1\right) \mid g(P_i)\right] \mathbb{P}\left(\pi_j = i \middle| \mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right) \\
\leq\ & p_* \sum_{i \in \mathcal{R}_{n-i} \cap \mathcal{H}_0} \mathbb{P}\left(\pi_j = i \middle| \mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right) = p_*,
\end{aligned}$$

where equation (a) simplify the condition of $\mathcal{F}_{n-j}$ to $g(P_i)$ because for any $i \in \mathcal{R}_{n-i} \cap \mathcal{H}_0$, $h(P_i)$ is independent of other information in $\mathcal{F}_{n-j}$.

Then, by the law of iterated expectations, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right] \\
=\ & \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(h(P_{\pi_j}) = 1\right) \middle| \mathcal{F}_{n-j}, \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right] \middle| \mathcal{F}_{n-j}^{h}, \pi_j \in \mathcal{H}_0, \{g(P_{\pi_k})\}_{k=1}^{n}\right] \leq p_*,
\end{aligned}$$

thus the proof is completed. □

B.2.2 Negative binomial distribution

In this section, we discuss several procedures for Bernoulli trials (coin flips) and their connections with the negative binomial distribution.

Lemma 8. *Suppose $A_1, \ldots, A_n$ are i.i.d. Bernoulli with parameter p_*. For $t = 1, \ldots, n$, consider the sum $M_t = \sum_{j=1}^{t} A_j$ and the filtration $\mathcal{G}_t^o = \sigma\left(\{A_j\}_{j=1}^t\right)$. Define a stopping time parameterized by a constant $v(\geq 1)$:*

$$\tau^o = \min\{0 < t \leq n : t - M_t \geq v \text{ or } t = n\}, \tag{173}$$

then M_{τ^o} is stochastically dominated by a negative binomial distribution:

$$M_{\tau^o} \preceq \mathrm{NB}(v, p_*).$$

Proof. Recall that the negative binomial $\mathrm{NB}(v, p_*)$ is the distribution of the number of success in a sequence of independent and identically distributed Bernoulli trials with probability p_* before a predefined number v of failures have occurred. Imagine the sequence of A_j is extended to infinitely many Bernoulli trials: $A_1, \ldots, A_n, A'_{n+1}, \ldots$, where $\{A'_j\}_{j=n+1}^{\infty}$ are also i.i.d. Bernoulli with parameter p_* and they are independent of $\{A_j\}_{j=1}^n$. Let U be the number of success before v-th failure, then by definition, U follows a negative binomial distribution $\mathrm{NB}(v, p_*)$. We can rewrite U as a sum at a stopping time: $U \equiv M_{\tau'}$, where $\tau' = \min\{t > 0 : t - M_t \geq v\}$. By definition, $\tau^o \leq \tau'$ (a.s.), which indicates $M_{\tau^o} \leq M_{\tau'}$ because M_t is nondecreasing with respect to t. Thus, we have proved that $M_{\tau^o} \preceq \mathrm{NB}(v, p_*)$. □

Corollary 5. *Following the setting in Lemma 8, we consider the shrinking sum $\widetilde{M}_t = \sum_{j=1}^{n-t} A_j$ for $t = 0, 1, \ldots, n-1$. Let the filtration be $\widetilde{\mathcal{G}}_t = \sigma\left(\widetilde{M}_t, \{A_j\}_{j=n-t+1}^n\right)$. Given a constant $v(\geq 1)$, we define a stopping time:*

$$\widetilde{\tau} = \min\{0 \leq t < n : (n-t) - \widetilde{M}_t < v \text{ or } t = n-1\}, \tag{174}$$

then it still holds that $\widetilde{M}_{\widetilde{\tau}} \preceq \mathrm{NB}(v, p_*)$.

Proof. We first replace the notion of time t by $n-s$, and let time runs backward: $s = n, n-1, \ldots, 1$. The above setting can be rewritten as $\widetilde{M}_t (= \sum_{j=1}^{n-t} A_j) \equiv M_{n-t} \equiv M_s$ and $\widetilde{\mathcal{G}}_t = \sigma\left(M_s, \{A_j\}_{j=s+1}^n\right) =: \mathcal{G}_s^b$. Define a stopping time:

$$\tau^b = \max\{0 < s \leq n : s - M_s < v \text{ or } s = 1\}, \tag{175}$$

which runs backward with respect to the filtration $\mathcal{G}_s^b$. By definition, we have $n - \widetilde{\tau} \equiv \tau^b$, and hence $\widetilde{M}_{\widetilde{\tau}} \equiv M_{\tau^b}$.

Now, we show that $M_{\tau^b} \equiv M_{\tau^o}$ for τ^o defined in Lemma 8. First, consider two edge cases: (1) if $t - M_t < v$ holds for every $0 < t \leq n$, then $\tau^b = n = \tau^o$, and thus $M_{\tau^b} = M_{\tau^o}$; (2) if $t - M_t \geq v$ holds for every $0 < t \leq n$, then $\tau^b = 1 = \tau^o$, and again $M_{\tau^b} = M_{\tau^o}$. Next, consider the case where $t - M_t < v$ for some t, and $t - M_t \geq v$ for some other t. Note that by definition, $\tau^b + 1$ is a stopping time with respect to $\mathcal{G}_t^o$, and $\tau^b + 1 = \tau^o$. Also, note that by the definition of τ^o, we have $A_{\tau^o} = 0$, so $M_{\tau^o - 1} = M_{\tau^o}$. Thus, $M_{\tau^b} = M_{\tau^o - 1} = M_{\tau^o}$. Therefore, by Lemma 8, $\widetilde{M}_{\widetilde{\tau}} \equiv M_{\tau^b} \equiv M_{\tau^o} \preceq NB(v, p_*)$, as stated in the above Corollary. □

Corollary 6. *Consider a weighted version of the setting in Corollary 5. Let the weights $\{W_j\}_{j=1}^n$ be a sequence of Bernoulli, such that (a) $\sum_{j=1}^n W_j = m$ for a fixed constant $m \leq n$; and (b) $A_j \mid \sigma\left(\{A_k, W_k\}_{k=j+1}^n, W_j = 1\right)$ is a Bernoulli with parameter p_*. Consider the sum $M_t^w = \sum_{j=1}^{n-t} W_j A_j$.*

Given a constant $v(\geq 1)$, we define a stopping time:

$$\begin{aligned}\tau^w = &\ \min\{0 \leq t < n : \sum_{j=1}^{n-t} W_j(1-A_j) < v \text{ or } t = n-1\} \\ \equiv &\ \min\{0 \leq t < n : \sum_{j=1}^{n-t} W_j - M_t^w < v \text{ or } t = n-1\},\end{aligned} \tag{176}$$

then it still holds that $M_{\tau^w}^w \preceq \mathrm{NB}(v, p_*)$.

Proof. Intuitively, adding the binary weights should not change the distribution of the sum $M_{\tau^w}^w = \sum_{j=1}^{n-\tau^w} W_j A_j$, since by condition (b), A_j is still a Bernoulli with parameter p_* when it is counted in the sum. We formalize this idea as follows.

Let $\{B_l\}_{l=1}^m$ be a sequence of i.i.d. Bernoulli with parameter p_*, and denote the sum $\sum_{l=1}^{m-s} B_l$ as $\widetilde{M_s}(B)$. Let $T(t) = m - \sum_{j=1}^{n-t} W_j$, then the stopping time τ^w can be rewritten as

$$\tau^w \equiv \min\{0 \leq t < n : m - T(t) - \widetilde{M_{T(t)}}(B) < v \text{ or } t = n-1\}, \tag{177}$$

because $m - T(t) = \sum_{j=1}^{n-t} W_j$ by definition, and

$$\widetilde{M_{T(t)}}(B) = \sum_{l=1}^{m-T(t)} B_l \overset{d}{=} \sum_{j=1}^{n-t} W_j A_j = M_t^w. \tag{178}$$

For simple notation, we present the reasoning of equation (178) when $t = 0$ (for arbitrary t, consider the distributions conditional on $\{A_k, W_k\}_{k=n-t+1}^n$). That is, we show that $\mathbb{P}(\sum_{l=1}^m B_l = x) = \mathbb{P}(\sum_{j=1}^n W_j A_j = x)$ for every $x \geq 0$. Let $\{b_l\}_{j=1}^m \in \{0,1\}^m$, then we derive that

$$\mathbb{P}(\sum_{l=1}^m B_l = x) = \sum_{\sum b_l = x} \mathbb{P}(B_l = b_l \text{ for } l = 1, \ldots, n) = \sum_{\sum b_l = x} \prod_{l=1}^m f^B(b_l),$$

where f^B is the probability mass function of a Bernoulli with parameter p_*. Let $\{a_k\}_{k=1}^{n-m} \in \{0,1\}^{n-m}$, then for the weighted sum,

$$\begin{aligned}&\mathbb{P}(\sum_{j=1}^n W_j A_j = x) \\ = &\sum_{\sum b_l = x} \sum_{\sum w_j = m} \sum_{a_k} \mathbb{P}(A_j = b_l \text{ if } w_j = 1; A_j = a_k \text{ if } w_j = 0; W_j = w_j \text{ for } i = 1, \ldots, n) \\ = &\sum_{\sum b_l = x} \prod_{l=1}^m f^B(b_l) \underbrace{\sum_{\sum w_j = m} \sum_{\sum a_k} \prod_{w_j = 0} \mathbb{P}(A_j = a_k \mid \sigma\left(\{A_k, W_k\}_{k=j+1}^n, W_j = 0\right) \prod_{j=1}^n \mathbb{P}(W_j = w_j \mid \{A_k, W_k\}_{k=j+1}^n)}_{C \quad \text{(a constant with respect to } x)} \\ = &\ C \sum_{\sum b_l = x} \prod_{l=1}^m f^B(b_l) = C\mathbb{P}(\sum_{l=1}^m B_l = x),\end{aligned}$$

for every possible value $x \geq 0$, which implies that $\mathbb{P}(\sum_{l=1}^m B_l = x)$ and $\mathbb{P}(\sum_{j=1}^n W_j A_j = x)$ have the same value; and hence we conclude equation (178). It follows that the filtration for both the stopping

time τ^w and the sum $M^w_{t^w}$, denoted as $\sigma\left(\sum_{j=1}^{n-t} W_j, M^w_{t^w}, \{A_j, W_j\}_{j=n-t+1}^n\right)$, has the same probability measure as $\sigma\left(m - T(t), \widetilde{M_{T(t)}}(B), \{A_j, W_j\}_{j=n-t+1}^n\right)$. Thus, the sums at the stopping time have the same distribution, $M^w_{\tau^w} \overset{d}{=} \widetilde{M_{T(\tau^w)}}(B)$. The proof completes if $\widetilde{M_{T(\tau^w)}}(B) \preceq \mathrm{NB}(v, p_*)$. It can be proved once noticing that stopping rule (177) is similar to stopping rule (174) except $T(t)$ is random because of W_j, so we can condition on $\{W_j\}_{j=1}^n$ and apply Corollary 5; and this concludes the proof. □

Corollary 7. *In Corollary 6, consider A_j with different parameters. Suppose $A_j \mid \sigma\left(\{A_k, W_k\}_{k=j+1}^n, W_j = 1\right)$ is a Bernoulli with parameter $p\left(\{A_k, W_k\}_{k=j+1}^n\right)$ for every $j = 1, \ldots, n$. Given a constant $p_* \in (0,1)$, if the parameters satisfy that $p\left(\{A_k, W_k\}_{k=j+1}^n\right) \leq p_*$ for all $j = 1, \ldots, n$, then it still holds that $M^w_{\tau^w} \preceq \mathrm{NB}(v, p_*)$.*

Proof. We first construct Bernoulli with parameter p_* based on A_j by an iterative process. Start with $j = n$. Let C_n be a Bernoulli independent of $\{A_k\}_{k=1}^n$ with parameter $\frac{p_* - p_n}{1 - p_n}$, where $p_n = \mathbb{E}(A_n \mid W_n = 1)$. Construct

$$B_n = A_n \mathbb{1}\left(A_n = 1\right) + C_n \mathbb{1}\left(A_n = 0\right), \tag{179}$$

which thus satisfies that $\mathbb{E}(B_n \mid W_n = 1) = p_*$, and that $B_n \geq A_n$ (a.s.). Now, let $j = j - 1$ where we consider the previous random variable. Let C_j be a Bernoulli independent of $\{A_k\}_{k=1}^j$, with parameter

$$\frac{p_* - \widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right)}{1 - \widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right)}, \tag{180}$$

where $\widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right) = \mathbb{E}\left[A_j \mid \sigma\left(\{B_k, W_k\}_{k=j+1}^n, W_j = 1\right)\right]$ (note that the parameter for C_j is well-defined since $\widetilde{p}\left(\{B_k, W_k\}_{k=j+1}^n\right) \leq p_*$ by considering the expectation further conditioning on $\{A_k\}_{k=j+1}^n$). Then, we construct B_j as

$$B_j = A_j \mathbb{1}\left(A_j = 1\right) + C_j \mathbb{1}\left(A_j = 0\right), \tag{181}$$

which thus satisfies that $\mathbb{E}\left[B_j \mid \sigma\left(\{B_k, W_k\}_{k=j+1}^n, W_j = 1\right)\right] = p_*$, and that $B_j \geq A_j$ (a.s.).

Now, consider two procedures for $\{A_j\}_{j=1}^n$ and $\{B_j\}_{j=1}^n$ with the same stopping rule (176) in Corollary 6, where the sum of A_j is denoted as $M^w_t(A)$ and the stopping time as τ^w_A (and the similar notation for B_j). Since construction (181) ensures that $B_j \geq A_j$ for every $j = 1, \ldots, n$, we have $M^w_t(B) \geq M^w_t(A)$ for every t; and hence, $\tau^w_A \geq \tau^w_B$. It follows that

$$M^w_{\tau^w_A}(A) \leq M^w_{\tau^w_B}(A) \leq M^w_{\tau^w_B}(B) \preceq \mathrm{NB}(v, p_*),$$

where the first inequality is because M^w_t is nonincreasing with respect to t, and the last step is the conclusion of Corollary 6; this completes the proof. □

B.2.3 Proof of Theorem 8.

Proof. We discuss three cases: (1) the simplest case where all the hypotheses are null, and the null p-values are uniformly distributed; (2) the case where non-nulls may exist, and the null p-values are uniformly distributed; and finally (3) the case where non-nulls may exist, and the null p-values can be mirror-conservative.

Case 1: nulls only and null p-values uniform. By Lemma 7, $\{\mathbb{1}\left(h\left(P_{\pi_j}\right)=1\right)\}_{j=1}^{n}$ are i.i.d. Bernoulli with parameter p_*. Observe that the stopping rule in Algorithm 5, $\widehat{\text{FWER}}_t \equiv 1-(1-p_*)^{|\mathcal{R}_t^-|+1} \le \alpha$, can be rewritten as $|\mathcal{R}_t^-|+1 \le v$ where

$$v = \left\lfloor \frac{\log(1-\alpha)}{\log(1-p_*)} \right\rfloor, \tag{182}$$

which is also equivalent as $|\mathcal{R}_t^-| < v$. We show that the number of false rejections is stochastically dominated by $\text{NB}(v, p_*)$ by Corollary 5. Let $A_j = \mathbb{1}\left(h\left(P_{\pi_j}\right)=1\right)$ and $\widetilde{M}_t = \sum_{j=1}^{n-t} \mathbb{1}\left(h\left(P_{\pi_j}\right)=1\right)$. The stopping time is $\widetilde{\tau} = \min\{0 \le t < n : |\mathcal{R}_t^-| = (n-t) - \widetilde{M}_t < v \text{ or } t = n-1\}$. The number of rejections at the stopping time is

$$|\mathcal{R}_{\widetilde{\tau}}^+| \equiv \sum_{j=1}^{n-\widetilde{\tau}} \mathbb{1}\left(h\left(P_{\pi_j}\right)=1\right) \equiv \widetilde{M}_{\widetilde{\tau}} \preceq \text{NB}(v, p_*),$$

where the last step is the conclusion of Corollary 5. Note that we assume all the hypotheses are null, so the number of false rejections is $|\mathcal{R}_{\widetilde{\tau}}^+ \cap \mathcal{H}_0| = |\mathcal{R}_{\widetilde{\tau}}^+| \preceq \text{NB}(v, p_*)$. Thus, FWER is upper bounded:

$$\mathbb{P}(|\mathcal{R}_{\widetilde{\tau}}^+ \cap \mathcal{H}_0| \ge 1) \le 1-(1-p_*)^v \le \alpha, \tag{183}$$

where the last inequality follows by the definition of v in (182). Thus, we have proved FWER control in Case 1.

Remark: This argument also provides some intuition on the FWER estimator (31): $\widehat{\text{FWER}}_t = 1-(1-p_*)^{|\mathcal{R}_t^-|+1}$. Imagine we run the algorithm for one time without any stopping rule until time t_0 to get an instance of $\widehat{\text{FWER}}_{t_0}$, then we run the algorithm on another independent dataset, which stops once $\widehat{\text{FWER}}_t \le \widehat{\text{FWER}}_{t_0}$. Then in the second run, FWER is controlled at level $\widehat{\text{FWER}}_{t_0}$.

Case 2: non-nulls may exist and null p-values are uniform. We again argue that the number of false rejections is stochastically dominated by $\text{NB}(v, p_*)$, and in this case we use Corollary 6. Consider $A_j = \mathbb{1}\left(h\left(P_{\pi_j}\right)=1\right)$ and $W_j = \mathbb{1}\left(\pi_j \in \mathcal{H}_0\right)$, which satisfies condition (b) in Corollary 6 according to Corollary 3. Let $m = |\mathcal{H}_0|$, then $\sum_{j=1}^{n} W_j = m$, which corresponds to condition (a). Imagine an algorithm stops once

$$\sum_{j=1}^{n-t} \mathbb{1}\left(h\left(P_{\pi_j}\right)=-1 \cap \pi_j \in \mathcal{H}_0\right) = \sum_{j=1}^{n-t} W_j(1-A_j) < v, \tag{184}$$

and we denote the stopping time as τ^w. By Corollary 6, the number of false rejections in this imaginary case is

$$\sum_{j=1}^{n-\tau^w} \mathbb{1}\left(h\left(P_{\pi_j}\right)=1 \cap \pi_j \in \mathcal{H}_0\right) = \sum_{j=1}^{n-t} W_j A_j = M_{\tau^w}^w \preceq \text{NB}(v, p_*).$$

Now, consider the actual i-FWER test which stops when $|R_t^-| = (n-t) - \sum_{j=1}^{n-t} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right) < v$, and denote the true stopping time as τ_T^w. Notice that at the stopping time, it holds that

$$\begin{aligned} & \sum_{j=1}^{n-\tau_T^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = -1 \cap \pi_j \in \mathcal{H}_0\right) \\ \leq & \sum_{j=1}^{n-\tau_T^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = -1\right) \\ = & (n-\tau_T^w) - \sum_{j=1}^{n-\tau_T^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1\right) < v, \end{aligned}$$

which means that stopping rule (184) is satisfied at τ_T^w. Thus, $\tau_T^w \geq \tau^w$ and $M_{\tau_T^w}^w \leq M_{\tau^w}^w$ (because M_t^w is nonincreasing with respect to t). It follows that the number of false rejections is

$$|\mathcal{R}_{\tau_C^w}^+ \cap \mathcal{H}_0| \equiv \sum_{j=1}^{n-\tau_C^w} \mathbb{1}\left(h\left(P_{\pi_j}\right) = 1 \cap \pi_j \in \mathcal{H}_0\right) \equiv M_{\tau_C^w}^w \leq M_{\tau^w}^w \preceq \mathrm{NB}(v, p_*).$$

We then prove FWER control using a similar argument as (183):

$$\mathbb{P}(|\mathcal{R}_{\tau^w}^+ \cap \mathcal{H}_0| \geq 1) \leq 1 - (1-p_*)^v \leq \alpha,$$

which completes the proof of Case 2.

Case 3: non-nulls may exist and null p-values can be mirror-conservative. In this case, we follow the proof of Case 2 except additionally conditioning on all the masked p-values, $\{g(P_{\pi_k})\}_{k=1}^n$. By Corollary 4 and Corollary 7, we again conclude that the number of false rejections is dominated by a negative binomial:

$$|\mathcal{R}_{\tau_C^w}^+ \cap \mathcal{H}_0| \preceq \mathrm{NB}(v, p_*),$$

if given $\{g(P_{\pi_k})\}_{k=1}^n$. Thus, FWER conditional on $\{g(P_{\pi_k})\}_{k=1}^n$ is upper bounded:

$$\mathbb{P}\left(|\mathcal{R}_{\tau^w}^+ \cap \mathcal{H}_0| \geq 1 \middle| \{g(P_{\pi_k})\}_{k=1}^n\right) \leq 1 - (1-p_*)^v \leq \alpha,$$

which implies the FWER control by the law of iterated expectations. This completes the proof of Theorem 8. □

B.3 An alternative perspective: closed testing

This section summarizes the comments from Jelle Goeman, who kindly points out the connection between our proposed method and the *closed testing* [Marcus et al., 1976]. Closed testing is a general framework that generates a procedure with FWER control given any test with Type 1 error control. Specifically, we reject H_i if all possible sets of hypotheses involving H_i, denoted as $U \ni i$, can be rejected by a "local" test for hypotheses in U with Type 1 error control at level α.

The i-FWER test we propose shares some commonalities with the *fallback procedure* [Wiens and Dmitrienko, 2005], which can be viewed as a shortcut of a closed testing procedure. We briefly describe the commonalities and differences next. Let v be a prespecified positive integer. The fallback procedure

orders the hypotheses from most to least interesting, and proceeds to test them one by one at level α/v until it has failed to reject v hypotheses. The hypothesis ordering is allowed to be data-dependent as long as the ordering is independent of the p-values, corresponding to ordering by the side information x_i in our language. This procedure is essentially also what the i-FWER test does except (a) the i-FWER test uses the Šidák correction instead of the Bonferroni correction; (b) we are interested in whether rejecting each hypothesis instead of adjusting individual p-values, so the ordering only needs to be independent of reject/non-reject status instead of on the full p-values, which allows us to split each p-value into $h(P_i)$ and $g(P_i)$; (c) under the assumption of independent null p-values, we are allowed to use the p-values excluded from the candidate rejection set $\mathcal{R}_t$ as independent information to create the ordering. The latter two differences enable the i-FWER test to be interactive based on a considerably large amount of data information.

B.3.1 Alternative proof of Theorem 8

The above observation leads to a simple proof of the error control guarantee without involving any martingales or negative binomial distributions, once we rewrite the i-FWER test in the language of closed testing.

Proof. For simplicity, we consider the nulls with only uniform p-values. Let v be a prespecified positive integer, and define $p_* = 1 - (1-\alpha)^{1/v}$. Imagine that the i-FWER test does not have a stopping rule and let $\pi_n, \dots, \pi_1$ be the order in which the hypotheses are chosen by an analyst, where each choice π_t can base on all the information in $\mathcal{F}_{n-t}$.

Here, we construct a closed testing procedure by defining a local test with Type 1 error control for an arbitrary subset $U \in [n]$ of size $|U|$. Sort the hypotheses in U according to the analyst-specified ordering from the last π_n to the first chosen π_1. If the number of hypotheses in U is larger than v, define U_v as the subset of U of size v corresponding to the hypotheses in U that are chosen last. For example, if $U = [n]$, we have $U_v = \{\pi_v, \dots, \pi_1\}$. If $|U| \le v$, define $U_v = U$. We reject the subset U if $h(P_i) = 1$ (i.e., $P_i \le p_*$) for at least one $i \in U_v$. This is a valid local test, since it controls the Type 1 error when all the hypotheses in U are null. To verify the error control, notice that $h(P_i)$'s are independent and follows Bernoulli(p_*), and U_v is independent of $\{h(P_i)\}_{i \in U_v}$ by the construction of sequence $\pi_1, ..., \pi_n$, so the Type 1 error satisfy

$$\mathbb{P}(\exists i \in U_v : h(P_i) = 1) \le 1 - (1-p_*)^v,$$

which is less than α by the definition of v and p_*. Indeed, the local test corresponds to a Šidák correction for v number of hypotheses. Through closed testing, this local test leads to a valid test with FWER control.

Next, we show that the rejection set from the i-FWER test, $\mathcal{R}_\tau^+$, is included in the rejection set from the above closed testing procedure. Choose any hypothesis $j \in \mathcal{R}_\tau^+$ and any set $W \ni j$. If H_j is among the last v hypotheses last chosen in W (or if $|W| \le v$), the local test for W reject the null since $P_j \le p_*$ by the definition of $\mathcal{R}_\tau^+$. Otherwise, the v hypotheses last chosen in W are all chosen after H_j. Since $j \in \mathcal{R}_\tau^+$ and by the definition of τ, we have $|\mathcal{R}_\tau^-| \le v - 1$. That is, there can be at most $v-1$ hypotheses among these v such that $h(P_i) = -1$, so set W is rejected by the local test as described in the previous paragraph. It follows from the definition of FWER and the error control of the larger (or equivalent) rejection set from the closed testing procedure that $\mathcal{R}_\tau^+$ has FWER control. □

B.3.2 Improvement on an edge case

From the closed testing procedure constructed in the above proof, we observe that the local tests do not exhaust the α-level for intersections of less than v hypotheses. This suboptimality can be remedied, but it will only improve power for rejecting all hypotheses given that almost all are already rejected (i.e., most subsets U with $|U| > v$ are rejected by the local test). In the i-FWER test, such a case potentially corresponds to the case where the initial rejection set has less than v hypotheses with negative $h(P_i)$, so the algorithm stops before shrinking $\mathcal{R}_0$, and reject all the hypotheses with positive $h(P_i)$. However, we might not fully use the error budget because $\widehat{\text{FWER}}_0 < \alpha$. However, we might not fully use the error budget because $\widehat{\text{FWER}}_0 < \alpha$. To improve power and efficiently use all the error budget, we propose randomly rejecting the hypotheses with a negative $h(P_i)$ if the algorithm stops at step 0.

Algorithm 15 The adjusted i-FWER test

Input: Side information and p-values $\{x_i, P_i\}_{i=1}^n$, target FWER level α, and parameter p_*;
Procedure:
Initialize $\mathcal{R}_0 = [n]$;
if $\widehat{\text{FWER}}_0 \equiv 1 - (1 - p_*)^{|\mathcal{R}_0^-|+1} \le \alpha$ **then**
  Obtain n independent indicators from a Bernoulli distribution with probability $1 - (1 - \alpha + \widehat{\text{FWER}}_0)^{1/|\mathcal{R}_0^-|}$, denoted as $\{I_i\}_{i\in[n]}$;
  Reject $\{H_i : i \in [n], h(P_i) = 1 \text{ or } I_i = 1\}$ and exit;
else
  for $t = 1$ **to** n **do**
    1. Pick any $i_t^* \in \mathcal{R}_{t-1}$, using $\{x_i, g(P_i)\}_{i=1}^n$ and $\{h(P_i)\}_{i\notin\mathcal{R}_{t-1}}$;
    2. Exclude i_t^* and update $\mathcal{R}_t = \mathcal{R}_{t-1}\backslash\{i_t^*\}$;
    if $\widehat{\text{FWER}}_t \equiv 1 - (1 - p_*)^{|\mathcal{R}_t^-|+1} \le \alpha$ **then**
      Reject $\{H_i : i \in \mathcal{R}_t, h(P_i) = 1\}$ and exit;
    end if
  end for
end if

Recall that the number of negative $h(P_i)$ is $|\mathcal{R}_0^-|$. For each hypothesis with a negative $h(P_i)$, we independently decide to reject it with probability $1 - (1 - \alpha_{\text{re}})^{1/|\mathcal{R}_0^-|}$, where $\alpha_{\text{re}} := \alpha - \widehat{\text{FWER}}_0$ denotes the remaining error budget after rejecting all the hypotheses with positive $h(P_i)$'s. We summarize the adjusted i-FWER test in Algorithm 15. To see the error control guarantee of this improved algorithm, notice that

$$\begin{aligned}
&\mathbb{P}(\exists i \in \mathcal{H}_0 : H_i \text{ is rejected}) \\
\le\ &\mathbb{P}(\exists i \in \mathcal{H}_0 : h(P_i) = 1) + \mathbb{P}(\exists i \in \mathcal{H}_0 : h(P_i) = -1 \text{ and } H_i \text{ is rejected}) \\
\le\ &\widehat{\text{FWER}}_0 + \mathbb{P}(\exists i \in \mathcal{R}_0^- : H_i \text{ is rejected}) \\
\le\ &\widehat{\text{FWER}}_0 + \alpha_{\text{re}} = \alpha,
\end{aligned}$$

where $\mathbb{P}(\exists i \in \mathcal{H}_0 : h(P_i) = 1) \le \widehat{\text{FWER}}_0$ follows the argument using negative binomial distribution as in the proof of the original algorithm; and $\mathbb{P}(\exists i \in \mathcal{R}_0^- : H_i \text{ is rejected}) \le \alpha_{\text{re}}$ is the result of a Šidák correction.

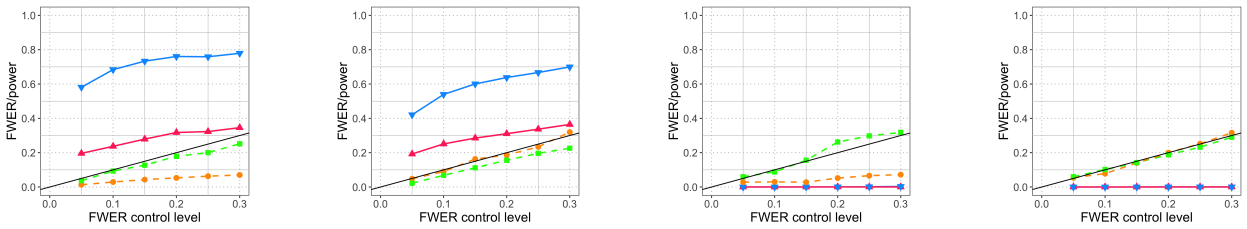
B.4 Sensitivity analysis

The i-FWER test is proved to have valid error control when the nulls are mutually independent and independent of the non-nulls. In this section, we evaluate the performance of the i-FWER test under correlated p-values. Our numerical experiments construct a grid of hypotheses as described in the setting in Section 2. The p-values are generated as

$$P_i = 1 - \Phi(Z_i), \text{where } Z = (Z_1, \ldots, Z_n) \sim N(\mu, \Sigma), \tag{185}$$

where $\mu = 0$ for the nulls and $\mu = 3$ for the non-nulls. The covariance matrix Σ, which is identity matrix in the main paper, is now set to an equi-correlated matrix:

$$\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}. \tag{186}$$



(a) Positively correlated case where $\rho = 0.5$ in the covariance matrix (186). The non-null mean value is 3.

(b) Negatively correlated case where $\rho = -0.5/n$ in the covariance matrix (186). Non-null mean value is 3.

(c) Negatively correlated case where $\rho = -0.5/n$ in the covariance matrix (186). All hypotheses are nulls.

(d) Negatively correlated case where $\rho = -0.5/n$ in the covariance matrix (186). All hypotheses are nulls.

Figure 48: FWER and power of the i-FWER test and the Šidák correction for dependent p-values generated by Gaussians as in (185) with covariance matrix (186) when the targeted level of FWER control varies in $(0.05, 0.1, 0.15, 0.2, 0.25, 0.3)$. The i-FWER test appears to control FWER below the targeted level and has relatively high power.

Under both the positively correlated case ($\rho = 0.5$) and the negatively correlated case ($-\rho = 0.5/n$ to guarantee that Σ is positive semi-definite), the i-FWER test seems to maintain the FWER control at most target levels even when all the hypotheses are nulls (see Figure 48c and 48d), and has higher power than the Šidák correction (see Figure 48a and 48b).

B.5 More results on the application to genetic data

Section 3.5 presents the number of rejections of the i-FWER test when the masking uses the tent function. We evaluate the i-FWER test when using the other three masking functions under the same experiments, but for simplicity, we only present the result when the FWER control is at level $\alpha = 0.2$ (see Table 3). Overall, the gap function leads to a similar number of rejections as the tent function, consistent with the

numerical experiments. However, the railway (gap-railway) function leads to fewer rejections than the tent (gap) function, which seems counterintuitive. Upon a closer look at the p-values, we find that the null p-values are not uniform or have an increasing density (see Figure 49). As a result, when using the tent function, there are fewer masked p-values from the nulls that could be confused with those of the non-nulls (with huge p-values), compared with using the railway function where the masked p-values of the confused nulls are those close to the masking parameter (around 0.02).

Table 3: Number of rejections by i-FWER test using different masking functions when $\alpha = 0.2$. The tent function and the gap function leads to more rejections compared with the railway function and the gap-railway function. The parameters in the gap and gap-railway function are set to $p_l = p_*$ and $p_u = 0.5$, and we need $p_l < \alpha/2$ for the test to make any rejection under level α.

Masing function	$p_* = \alpha/2$	$p_* = \alpha/10$	$p_* = \alpha/20$
Tent	1752	1848	1794
Railway	1778	1463	1425
Gap	NA	1802	1846
Gap-railway	NA	1764	1788

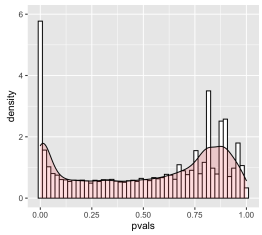


Figure 49: Histogram of p-values in the airway dataset. The number of p-values that are close to one is less than those that are close to the cutting point of the masking function (say 0.02). Consequently, the tent (gap) function leads to more rejections than the railway (gap-railway) function.

B.6 Error control for other masking functions

The proof in Appendix B.2 is for the i-FWER test with the original tent masking function. In this section, we check the error control for two new masking functions introduced in Section 3.4.

B.6.1 The railway function

We show that the i-FWER test with the "railway" function (34) has FWER control, if the null p-values have non-decreasing densities. We again assume the same independence structure as in Theorem 8 that the null p-values are mutually independent and independent of the non-nulls.

The proof in Appendix B.2 implies that under the same independence assumption, the FWER control is valid if the null p-values satisfy condition (172). When using the railway masking function,

condition (172) is indeed satisfied if the null p-values have nondecreasing f since

$$\begin{aligned}\mathbb{P}(h(P)=1 \mid g(P)=a) &= \frac{p_* f(a)}{p_* f(a)+(1-p_*) f(\frac{1-p_*}{p_*}a+p_*)} \\ &= \frac{p_*}{p_*+(1-p_*) f(\frac{1-p_*}{p_*}a+p_*)/f(a)} \\ &\leq p_*,\end{aligned}$$

for every $a \in (0, p_*)$. Then, we can prove the FWER control following the same argument as Appendix B.2.

B.6.2 The gap function

The essential difference of using the gap function instead of the tent function is that here, $\mathbb{1}(h(P)=1)$ for the nulls follow a Bernoulli distribution with a different parameter, $\widetilde{p} = \mathbb{P}(h(P)=1 \mid P < p_l \text{ or } P > p_u) = \frac{p_l}{p_l+1-p_u}$. Specifically, we replace replace condition (172) by

$$\mathbb{P}(h(P)=1 \mid g(P)=a) \leq \widetilde{p}, \tag{187}$$

for every $a \in (0, p_l)$, which holds for p-values with non-decreasing densities because

$$\begin{aligned}\mathbb{P}(h(P)=1 \mid g(P)=a) &= \frac{p_l f(a)}{p_l f(a)+(1-p_u) f(1-\frac{1-p_u}{p_l}a)} \\ &= \frac{p_l}{p_l+(1-p_u) f(1-\frac{1-p_u}{p_l}a)/f(a)} \\ &\leq \widetilde{p}.\end{aligned}$$

We also replace all p_* by $\widetilde{p}$ and get a the new FWER estimator $\widehat{\text{FWER}}_t$ as defined in (36), and the error control can be proved following Appendix B.2.

B.6.3 The gap-railway function

The proof is the same as that for the gap function except condition (187) is verified for p-values with non-decreasing densities differently as follow:

$$\begin{aligned}\mathbb{P}(h(P)=1 \mid g(P)=a) &= \frac{p_l f(a)}{p_l f(a)+(1-p_u) f(\frac{1-p_u}{p_l}a+p_u)} \\ &= \frac{p_l}{p_l+(1-p_u) f(\frac{1-p_u}{p_l}a+p_u)/f(a)} \\ &\leq \widetilde{p},\end{aligned}$$

for every $a \in (0, p_l)$.

B.7 Varying the parameters in the presented masking functions

We first discuss the original tent masking (29), which represents a class of masking functions parameterized by p_*. Similar to the discussion in Section 3.4, varying p_* also changes the amount of p-value

information distributed to $g(P)$ for interaction (to exclude possible nulls) and $h(P)$ for error control (by estimating FWER), potentially influencing the test performance. On one hand, the masking function with smaller p_* effectively distributes less information to $g(P)$, in that a larger range of big p-values is mapped to small $g(P)$ (see Figure 20a). In such a case, the true non-nulls with small p-values and small $g(P)$ are less distinctive, making it difficult to exclude the nulls from $\mathcal{R}_t$. On the other hand, the rejected hypotheses in $\mathcal{R}_t^+$ must satisfy $P < p_*$, so smaller p_* leads to less false rejections given the same $\mathcal{R}_t$.

Experiments show little change in power when varying the value of p_* in $(0, \alpha)$ as long as it is not near zero, as it would leave little information in $g(P)$. Our simulations follow the setting in Section 2, where the alternative mean value is fixed at $\mu = 3$. We tried seven values of p_* as $(0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2)$, and the power of the i-FWER test does not change much for $p_* \in (0.05, 0.2)$. This trend also holds when varying the mean value of non-nulls, the size of the grid (with a fixed number of non-nulls), and the number of non-nulls (with a fixed size of the grid). In general, the choice of p_* does not have much influence on the power, and a default choice can be $p_* = \alpha/2$.

There are also parameters in two other masking functions proposed in Section 3.4. The railway function flips the tent function without changing the distribution of p-value information, hence the effect of varying p_* should be similar to the case in the tent function. The gap function (35) has two parameters: p_l and p_u. The tradeoff between information for interaction and error control exhibits in both values of p_l and p_u: as p_l decreases (or p_u increases), more p-values are available to the analyst from the start, guiding the procedure of shrinking $\mathcal{R}_t$, while the estimation of FWER becomes less accurate. Whether revealing more information for interaction should depend on the problem settings, such as the amount of prior knowledge.

B.8 Mixture model for the non-null likelihoods

Two groups model for the p-values. Define the Z-score for hypothesis H_i as $Z_i = \Phi^{-1}(1 - P_i)$, where Φ^{-1} is the inverse function of the CDF of a standard Gaussian. Instead of modeling the p-values, we choose to model the Z-scores since when testing the mean of Gaussian, Z-scores are distributed as a Gaussian either under the null or the alternative:

$$H_0 : Z_i \overset{d}{=} N(0, 1) \quad \text{versus} \quad H_1 : Z_i \overset{d}{=} N(\mu, 1),$$

where μ is the mean value for all the non-nulls. We model Z_i by a mixture of Gaussians:

$$Z_i \overset{d}{=} (1 - q_i)N(0, 1) + q_i N(\mu, 1), \text{ with } q_i \overset{d}{=} \text{Bernoulli}(\pi_i),$$

where q_i is the indicator of whether the hypothesis H_i is truly non-null.

The non-null structures are imposed by the constraints on π_i, the probability of being non-null. In our examples, the blocked non-null structure is encoded by fitting π_i as a smooth function of the hypothesis position (coordinates) x_i, specifically as a logistic regression model on a spline basis $B(x) = (B_1(x), \ldots, B_m(x))$:

$$\pi_\beta(x_i) = \frac{1}{1 + \exp(-\beta^T B(x_i))}, \tag{188}$$

EM framework to estimate the non-null likelihoods. An EM algorithm is used to train the model. Specifically we treat the p-values as the hidden variables, and the masked p-values $g(P)$ as observed. In terms of the Z-scores, Z_i is a hidden variable and the observed variable $\tilde{Z}_i$ is

$$\tilde{Z}_i = \begin{cases} Z_i, & \text{if } Z_i > \Phi^{-1}(1 - p_*), \\ t(Z_i), & \text{otherwise,} \end{cases}$$

where $t(Z_i)$ depends on the form of masking. The updates needs values of its inverse function $t^{-1}(\tilde{Z}_i)$ and the derivative of $t^{-1}(\cdot)$, denoted as $\left(t^{-1}\right)'(\tilde{Z}_i)$, whose exact forms are presented below.

1. For tent masking (29),

$$\begin{aligned} t(Z_i) &= \Phi^{-1}\left[1-\frac{p_*}{1-p_*}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) &= \Phi^{-1}\left[\frac{1-p_*}{p_*}\left(1-\Phi(\tilde{Z}_i)\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) &= -\frac{1-p_*}{p_*}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right), \end{aligned}$$

   where $\phi(\cdot)$ is the density function of standard Gaussian.

2. For railway masking (34),

$$\begin{aligned} t(Z_i) &= \Phi^{-1}\left[1-p_*+\frac{p_*}{1-p_*}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) &= \Phi^{-1}\left[\frac{1-p_*}{p_*}\left(\Phi(\tilde{Z}_i)-1+p_*\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) &= \frac{1-p_*}{p_*}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right). \end{aligned}$$

3. For gap masking (35),

$$\begin{aligned} t(Z_i) &= \Phi^{-1}\left[1-\frac{p_l}{1-p^u}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) &= \Phi^{-1}\left[\frac{1-p_u}{p_l}\left(1-\Phi(\tilde{Z}_i)\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) &= -\frac{1-p_u}{p_l}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right). \end{aligned}$$

   if $Z_i < \Phi^{-1}(1-p_u)$. If $\Phi^{-1}(1-p_u) \le Z_i \le \Phi^{-1}(1-p_l)$, which corresponds to the skipped p-value between p_l and p_u, then $\tilde{Z}_i = Z_i$.

4. For gap-railway masking (37),

$$\begin{aligned} t(Z_i) &= \Phi^{-1}\left[1-\frac{p_l}{1-p^u}\Phi(Z_i)\right]; \\ t^{-1}(\tilde{Z}_i) &= \Phi^{-1}\left[\frac{1-p_u}{p_l}\left(\Phi(\tilde{Z}_i)-1+p_l\right)\right]; \\ \left(t^{-1}\right)'(\tilde{Z}_i) &= \frac{1-p_u}{p_l}\phi\left(\tilde{Z}_i\right)\Big/\phi\left(t^{-1}(\tilde{Z}_i)\right). \end{aligned}$$

   if $Z_i < \Phi^{-1}(1-p_u)$. If $\Phi^{-1}(1-p_u) \le Z_i \le \Phi^{-1}(1-p_l)$, which corresponds to the skipped p-value between p_l and p_u, then $\tilde{Z}_i = Z_i$.

Define two sequences of hypothetical labels $w_i = \mathbb{1}\{Z_i = \tilde{Z}_i\}$ and $q_i = \mathbb{1}\{H_i = 1\}$, where $H_i = 1$ means hypothesis i is truly non-null ($H_i = 0$ otherwise). The log-likelihood of observing $\tilde{Z}_i$ is

$$\begin{aligned} l(\tilde{Z}_i) = {} & w_i q_i \log\left\{\pi_i\phi\left(\tilde{Z}_i-\mu\right)\right\} + w_i(1-q_i)\log\left\{(1-\pi_i)\phi\left(\tilde{Z}_i\right)\right\} \\ & + (1-w_i)q_i\log\left\{\pi_i\phi\left(t^{-1}(\tilde{Z}_i)-\mu\right)\right\} + (1-w_i)(1-q_i)\log\left\{(1-\pi_i)\phi\left(t^{-1}(\tilde{Z}_i)\right)\right\}. \end{aligned}$$

The E-step updates w_i, q_i. Notice that w_i and q_i are not independent, and hence we update the joint distribution of (w_i, q_i), namely

$$\mathbb{E}[w_i q_i] =: a_i, \quad \mathbb{E}[w_i(1-q_i)] =: b_i, \quad \mathbb{E}[(1-w_i)q_i] =: c_i, \quad \mathbb{E}[(1-w_i)(1-q_i)] =: d_i,$$

where $a_i + b_i + c_i + d_i = 1$. To simplify the expression for updates, we denote

$$L_i := \pi_i \phi\left(\tilde{Z}_i - \mu\right) + (1-\pi_i)\phi\left(\tilde{Z}_i\right) + \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right| \pi_i \phi\left(t^{-1}(\tilde{Z}_i) - \mu\right) + \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right| (1-\pi_i)\phi\left(t^{-1}(\tilde{Z}_i)\right).$$

For the hypothesis i whose p-value is masked, the updates are

$$\begin{aligned}
a_{i,\text{new}} &= \mathbb{E}[w_i q_i \mid \tilde{Z}_i] = \pi_i \phi\left(\tilde{Z}_i - \mu\right) \Big/ L_i; \\
b_{i,\text{new}} &= \mathbb{E}[w_i(1-q_i) \mid \tilde{Z}_i] = (1-\pi_i)\phi\left(\tilde{Z}_i\right) \Big/ L_i; \\
c_{i,\text{new}} &= \mathbb{E}[(1-w_i)q_i \mid \tilde{Z}_i] = \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right| \pi_i \phi\left(t^{-1}(\tilde{Z}_i) - \mu\right) \Big/ L_i; \\
d_{i,\text{new}} &= \mathbb{E}[(1-w_i)(1-q_i) \mid \tilde{Z}_i] = \left|\left(t^{-1}\right)'(\tilde{Z}_i)\right| (1-\pi_i)\phi\left(t^{-1}(\tilde{Z}_i)\right) \Big/ L_i.
\end{aligned}$$

If the p-value is unmasked for i, the updates are

$$\begin{aligned}
a_{i,\text{new}} &= \left(1 + \frac{(1-\pi_i)\phi\left(\tilde{Z}_i\right)}{\pi_i \phi\left(\tilde{Z}_i - \mu\right)}\right)^{-1}; \\
b_{i,\text{new}} &= 1 - a_{i,\text{new}}; \quad c_{i,\text{new}} = 0; \quad d_{i,\text{new}} = 0.
\end{aligned}$$

In the M-step, parameters μ and β (in model (188) for π_i) are updated. The update for μ is

$$\mu_{\text{new}} = \operatorname*{argmax}_{\mu} \sum_i l(\tilde{Z}_i) = \frac{\sum a_i \tilde{Z}_i + c_i t^{-1}(\tilde{Z}_i)}{\sum a_i + c_i}.$$

The update for β is

$$\beta_{\text{new}} = \operatorname*{argmax}_{\beta} \sum_i (a_i + c_i) \log \pi_\beta(x_i) + (1 - a_i - c_i) \log(1 - \pi_\beta(x_i)),$$

where $\pi_\beta(x_i)$ is defined in equation (188). It is equivalent to the solution of GLM (generalized linear model) with the logit link function on data $\{a_i + c_i\}$ using covariates $\{B(x_i)\}$.

C Appendix for "Which Wilcoxon should we use? An interactive rank test and other alternatives"

C.1 Proof of Theorem 9

Proof. We argue that the sum $\{S_t\}_{t=1}^n$ is a martingale with respect to the filtration $\{\mathcal{F}_{t-1}\}_{t=1}^n$. First, the sum S_t is measurable with respect to $\mathcal{F}_{t-1}$, because $S_t = \sum_{j=1}^{t-1}(2A_{\pi_j} - 1)w_j + (2A_{\pi_t} - 1)w_t$, where $\sum_{j=1}^{t-1}(2A_{\pi_j} - 1)w_j$ and the t-th selected subject π_t and its weight w_t are all $\mathcal{F}_{t-1}$-measurable.

Second, we show that $\mathbb{E}(S_t \mid \mathcal{F}_{t-1}) = S_{t-1}$. Note that $\mathbb{E}(S_t \mid \mathcal{F}_{t-1}) = S_{t-1} + \mathbb{E}((2A_{\pi_t} - 1)w_t \mid \mathcal{F}_{t-1})$, so $\mathbb{E}(S_t \mid \mathcal{F}_{t-1}) = S_{t-1}$ holds when $\mathbb{E}((2A_{\pi_t} - 1)w_t \mid \mathcal{F}_{t-1}) = 0$, which is implied when $\mathbb{P}(A_{\pi_t} = 1 \mid \mathcal{F}_{t-1}, w_t) = \mathbb{P}(A_{\pi_t} = 0 \mid \mathcal{F}_{t-1}, w_t) = 1/2$. Note that assignment A only takes two values $\{0, 1\}$ and w_t is $\mathcal{F}_{t-1}$-measurable, so proving $\mathbb{P}(A_{\pi_t} = 1 \mid \mathcal{F}_{t-1}, w_t) = 1/2$ is equivalent to proving

$$\mathbb{E}(2A_{\pi_t} - 1 \mid \mathcal{F}_{t-1}) = 0. \tag{189}$$

Let the set of subjects ordered before t be $\mathcal{C}_{t-1} = \{\pi_j\}_{j=1}^{t-1}$. Claim (189) follows because

$$\mathbb{E}(2A_{\pi_t} - 1 \mid \mathcal{F}_{t-1}) \le \max_{i \notin \mathcal{C}_{t-1}} \mathbb{E}(2A_i - 1 \mid \mathcal{F}_{t-1}) = \mathbb{E}(2A_i - 1) = 0,$$

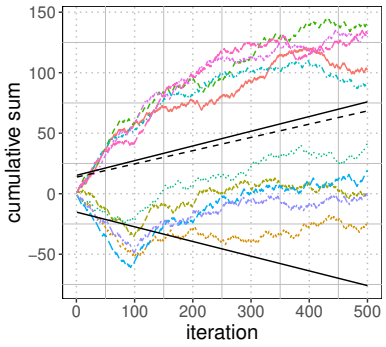
where the last equation is because A_i is independent of each other and of the covarites and outcomes under the global null; and thus, $2A_i - 1 \mid \mathcal{F}_{t-1}$ has the same distribution as $2A_i - 1$. Similarly, we have $\mathbb{E}(2A_{\pi_t} - 1 \mid \mathcal{F}_{t-1}) \ge 0$. Thus, we conclude that $\{S_t\}_{t=1}^n$ is a martingale.

Note that the increment $2A_{\pi_t} - 1$ conditional on $\mathcal{F}_{t-1}$ takes value in $\{\pm 1\}$. Combining with Claim (189) that the increment has zero mean under the null, it follows that the increment is 1 or -1 with the same probability. Therefore, boundary $u_\alpha(t)$ as defined in (45) for the sum of independent, fair coin flips is a time-uniform upper boundary for S_t. Note that S_t is symmetric around zero under the null, so $u_{\alpha/2}(t)$ is a two-sided boundary at level α: $\mathbb{P}\left(\exists t \in [n] : |S_t| > u_{\alpha/2}(t)\right) \le \alpha$, under the null. Recall the stopping time τ as defined in (52). The above inequality implies that the probability of $\tau \le n$, which corresponds to the case of rejection, is less α when the null hypothesis is true; thus, we have proved the error control. □

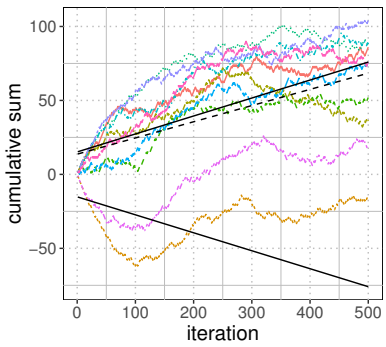
C.2 Comparison between monitoring S_t and its absolute value

We describe in Section 4.1.3 that the test statistic S_t is constructed such that it can grow fast under the alternative hypothesis, and we can reject the null when S_t is larger than a boundary $u_\alpha(t)$. However, when testing the null, we use a two-sided test which compares the absolute value of S_t with the boundary $u_{\alpha/2}(t)$ at level $\alpha/2$. It is because, at the first iteration, we could learn the opposite assignments for all the subjects using any modeling. In other words, when all assignments $\{A_i\}_{i=1}^n$ are hidden at $t = 1$, the likelihood of $\{\widehat{A}_i\}_{i=1}^n$ being the true values (treated/untreated) is the same as the likelihood of all opposite values (untreated/treated), no matter what working model we use. Consequently, the increment would take the opposite value and S_t decreases fast. This phenomenon is evident and not rare especially when we only update the estimation of assignments every 100 iterations (say) in practice for computational consideration (see Figure 50). Thus, we cannot reject the null when S_t decreases fast if we only monitor whether S_t is larger than $u_\alpha(t)$, although the decreasing S_t is also evidence of null being false (because the increment under the null would take $\pm w_j$ with equal probability).

Indeed, the decreasing S_t can be used to reject the null when we monitor the absolute value of S_t. Specifically, we reject the null when $|S_t|$ exceed the boundary $u_{\alpha/2}(t)$, which includes the case of S_t



(a) Linear effect (57) with $S_\Delta = 2$.



(b) Nonlinear effect (60) with $S_\Delta = 0.8$.

Figure 50: Instances of cumulative sums S_t under two types of effect. The solid lines are two-sided boundaries $-u_{\alpha/2}(t)$ and $u_{\alpha/2}(t)$, and the dashed line is the one-sided boundary $u_\alpha(t)$. Under linear treatment effect, about half of the instances reject the null by crossing the lower boundary, which is consistent with the power comparison (0.96 when using the two-sided test, and 0.65 using the one-sided test). Similar behavior can be found under nonlinear treatment effect.

being smaller than $-u_{\alpha/2}(t)$. We observe in numerical experiments that monitoring $|S_t|$ leads to higher power than monitoring S_t (see Figure 50).

C.3 An alternative strategy to choose weight w_j

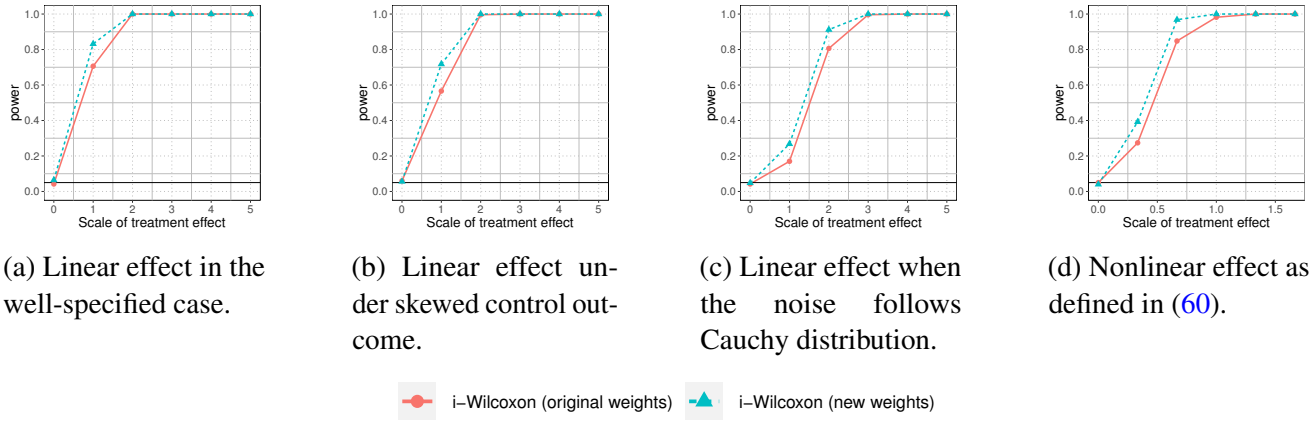
Recall that the cumulative sum S_t could decrease at early iterations, and we can still reject the null by examining whether S_t becomes smaller than a lower bound $-u_{\alpha/2}(t)$. The decreasing S_t results from guessing all the assignments falsely when all assignments are hidden. Yet, S_t would start increasing when we reestimate the assignments when some true assignments are revealed (see Figure 50 where we reestimate the assignments after 100 iterations). However, the direction change of S_t trajectory could potentially diminish the difference from the null behavior — S_t would not cross either the lower or the upper boundary because S_t is not small enough when it decreases and cannot be large enough when slowing increasing from a small value — and we could miss the chance of correctly rejecting the null. Based on the above observation, we propose an adjustment to our default automated interactive test (Algorithm 7), where the weight is constructed to make S_t either decrease or increase based on the previous trend.

Recall that the originally choice $w_j = 2\mathbb{1}\{\widehat{q}_{\pi_j} > 0.5\} - 1$ would make S_t increase when $\widehat{q}_{\pi_j}$ is a good estimation. Suppose we update the estimation $\widehat{q}_{\pi_j}$ every K iterations (eg. $K = 20$). After K iterations, the updated $\widehat{q}_{\pi_j}$ is likely to be a good estimation thanks to the revealed assignments. Thus, we propose use the opposite weight $w_j = 2\mathbb{1}\{\widehat{q}_{\pi_j} < 0.5\} - 1$ if $t \geq K$ and $S_{t-1} < 0$ (indicating that the initial estimation leads to small S_t and it is more likely to cross the lower boundary), such that S_t continues to decrease even after most assignments are no longer falsely estimated (see Figure 51). We also observe in the numerical experiments we discuss in the main paper, that this new strategy of choosing weights leads to slightly higher power (see Figure 52).

(a) Original weights.

(b) New weights.

Figure 51: Instances of the cumulative sums S_t with two types of weighting: original weight $w_j = 2\mathbb{1}\{\widehat{q}_{\pi_j} > 0.5\} - 1$, and new weight based on previous trend of S_t: $w_j = 2\mathbb{1}\{\widehat{q}_{\pi_j} < 0.5\} - 1$ if $t \geq K$ and $S_{t-1} < 0$ where K is the number of iterations after which we update the estimation $\widehat{q}_i$. We set $K = 20$ and simulate linear treatment effect (57) with $S_\Delta = 2$ under Cauchy noise. The solid lines are two-sided boundaries $-u_{\alpha/2}(t)$ and $u_{\alpha/2}(t)$ at level $\alpha/2$. The trajectories of S_t tends to have a consistent direction throughout the procedure, making it easy to cross the lower or the upper boundary and reject the null.



(a) Linear effect in the well-specified case.

(b) Linear effect under skewed control outcome.

(c) Linear effect when the noise follows Cauchy distribution.

(d) Nonlinear effect as defined in (60).

Figure 52: Power of the i-Wilcoxon test using two weighting strategies when varying the scale of the treatment effect under various situations of the outcome distribution. The new strategy that decide weights based on previous S_t trend usually leads higher power than the original strategy $w_j = 2\mathbb{1}\{\widehat{q}_{\pi_j} < 0.5\} - 1$ in the main paper.

C.4 Estimation of the posterior probability of receiving treatment

Under model (54), we view the treatment assignments of to-be-ordered subjects as hidden variables and apply the EM algorithm. At step t, the hidden variables are A_i for subjects $i \notin \{\pi_j\}_{i=1}^{t-1}$. And the rest of the complete data $\{Y_i, A_i, X_i\}_{i=1}^n$ is the observed data, denoted by σ-field $\mathcal{F}_{t-1}$ as defined in (50). In the working model (54), the log-likelihood of $\{Y_i, A_i, X_i\}_{i=1}^n$ is

$$l\left(\{Y_i, A_i, X_i\}_{i=1}^n\right) = \sum_{i\in[n]} \left[A_i \log \phi\left(Y_i - \theta_1(X_i)\right) + (1 - A_i) \log \phi\left(Y_i - \theta_0(X_i)\right) + g(X_i)\right],$$

where $\phi(\cdot)$ is the density of standard Gaussian and $g(\cdot)$ denotes the density of the covariates. In the E-step, we update the hidden variable A_i for $i \notin \{\pi_j\}_{i=1}^{t-1}$ as

$$A_i^{\text{new}} = \mathbb{E}(A_i \mid \mathcal{F}_{t-1}) = \frac{\phi\left(Y_i - \theta_1(X_i)\right)}{\phi\left(Y_i - \theta_1(X_i)\right) + \phi\left(Y_i - \theta_0(X_i)\right)}.$$

In the M-step, we update the (parametric) functions θ_0 and θ_1 as

$$\theta_0^{\text{new}} = \operatorname{argmax} l\left(\{Y_i, A_i, X_i\}\right) = \operatorname{argmin} \sum_{i\in[n]} (1 - A_i)(Y_i - \theta_0(X_i))^2,$$
$$\theta_1^{\text{new}} = \operatorname{argmax} l\left(\{Y_i, A_i, X_i\}\right) = \operatorname{argmin} \sum_{i\in[n]} A_i(Y_i - \theta_1(X_i))^2,$$

which are least square regressions with weights. The posterior probability of receiving treatment is estimated as $\mathbb{E}(A_i \mid \mathcal{F}_{t-1})$ for $i \notin \{\pi_j\}_{i=1}^{t-1}$.

C.5 The linear-CATE-test

We first describe the general framework of CATE without specifying the working model (see Vansteelandt and Joffe [2014] for a review). Suppose ψ^* is a vector of parameters, and a pre-defined function h satisfies $h(\psi^*, x) = 0$ if $\psi^* = 0$, for which a standard choice is a linear function of the covariates, $h(\psi^*, x) = x^T\psi^*$. CATE assumes that the difference in conditional expectations satisfy

$$\mathbb{E}(Y_i \mid X_i, A_i = 1) - \mathbb{E}(Y_i \mid X_i, A_i = 0) = h(\psi^*, X_i). \tag{190}$$

Thus, a valid test for null hypothesis (38) can be developed by testing $\psi^* = 0$. Note that the test is model-free (regardless of the correctness of h) since $\psi^* = 0$ is implied by null hypothesis (38) for any function h specified as above. The inference on ψ^* uses an observation that for any function g of the covariates and the assignment, we have

$$\mathbb{E}\{[g(X_i, A_i) - \mathbb{E}(g(X_i, A_i) \mid X_i)] \cdot [Y_i - \mathbb{E}(Y_i \mid X_i, A_i)]\} = 0, \tag{191}$$

where $\mathbb{E}(Y_i \mid X_i, A_i) = A_i \cdot h(\psi^*, X_i) + \mathbb{E}(Y_i \mid X_i, A_i = 0)$ because of (190). To get an estimation of ψ^*, we need to specify function h and g, and estimate $\mathbb{E}(g(X_i, A_i) \mid X_i)$ and $\mathbb{E}(Y_i \mid X_i, A_i = 0)$. Notice that in a randomized experiment, $\mathbb{E}(g(X_i, A_i) \mid X_i)$ is known given g, which guarantee that equation (191) holds regardless of whether $\mathbb{E}(Y_i \mid X_i, A_i = 0)$ is correctly specified (double robustness). In the following, we choose functions h, g and estimation of $\mathbb{E}(Y_i \mid X_i, A_i = 0)$ without concerns on the validity of equation (191). After getting the estimator of ψ^*, we present the test for $\psi^* = 0$ in the end.

For fair comparison with the i-Wilcoxon test that uses linear model by default, we set h to be a linear function of the covariates and their second-order interaction terms. Let X_i' be the vector of covariates X_i and the interaction terms, then $h = (X_i')^T\psi^*$. In such as case, a good choice of function g is $X_i' \cdot A_i$ [Vansteelandt and Joffe, 2014]. Because other methods in our comparison use linear models by default, we estimate $\mathbb{E}(Y_i \mid X_i, A_i = 0)$ by a linear model of X_i', denoted as $(X_i')^T\widehat{\beta}$ (note that $\widehat{\beta}$ can be learned by regressing Y_i on X_i without involving A_i since under the null, $\mathbb{E}(Y_i \mid X_i, A_i = 0) = \mathbb{E}(Y_i \mid X_i, A_i = 1) = \mathbb{E}(Y_i \mid X_i)$). With the above choices, equation (191) can be written as

$$\mathbb{E}\left[\underbrace{(A_i - 1/2)(Y_i - (X_i')^T\widehat{\beta})X_i'}_{b_i}\right] = \mathbb{E}\left[\underbrace{\left(X_i'^T A_i(A_i - 1/2)X_i'\right)}_{B_i}\psi^*\right], \tag{192}$$

which is denoted as $\mathbb{E}(b_i) = \mathbb{E}(B_i)\psi^*$ for simplicity. Let $\mathbb{P}_n b$ be the sample average of $\{b_i\}_{i=1}^n$ and $\mathbb{P}_n B$ be the sample average of $\{B_i\}_{i=1}^n$. A consistent estimator of ψ^* is

$$\begin{aligned}\widehat{\psi} = & \ (\mathbb{P}_n B)^{-1} \mathbb{P}_n b \\ = & \ \left(\frac{1}{n}\sum_{j=1}^n X_j'^T A_j(A_j - 1/2)X_j'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n (A_i - 1/2)(Y_i - (X_i')^T\widehat{\beta})X_i'\right).\end{aligned} \tag{193}$$

To derive the distribution of $\widehat{\psi}$, notice that its asymptotic variance is $B^{-1}\text{Var}(b)(B^{-1})^T$, for which a consistent estimator is

$$\widehat{\text{Var}}(\psi) = (\mathbb{P}_n B)^{-1} \widehat{\text{Var}}(b) \left[(\mathbb{P}_n B)^{-1}\right]^T, \tag{194}$$

where $\widehat{\text{Var}}(b)$ denotes the sample covariance of $\{b_i\}_{i=1}^n$. Under the null, we have

$$\widehat{\psi}^T[\widehat{\text{Var}}(\psi)]^{-1}\widehat{\psi} = (\mathbb{P}_n b)^T[\widehat{\text{Var}}(b)]^{-1}(\mathbb{P}_n b) \to \chi^2_p,$$

where p is the dimension of X_i'. The linear-CATE-test rejects the null if

$$(\mathbb{P}_n b)^T[\widehat{\text{Var}}(b)]^{-1}(\mathbb{P}_n b) > \chi^2_p(1-\alpha), \tag{195}$$

where b_i is defined in (192); and $\mathbb{P}_n$ and $\widehat{\text{Var}}$ denotes sample average and sample covariance matrix; and $\chi^2_p(1-\alpha)$ is the $1-\alpha$ quantile of a Chi-squared distribution with p degrees of freedom.

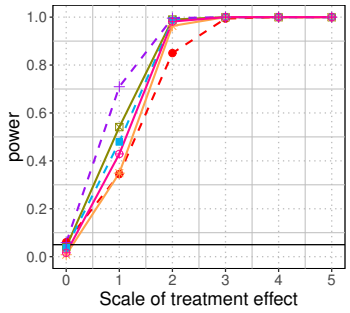
C.6 Bonferroni correction of the candidate Wilcoxon tests

In Section 4.3, we recommend choosing from three candidate Wilcoxon tests using different E_i ($E_i^{R(X)}$, $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$, $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$), whose p-values are denoted as p_1, p_2, p_3 respectively, depending on the property of treatment effect. In the case without prior knowledge of treatment effect, we could combine these candidates by a meta test and have high power under various underlying truth. We recommend using the Bonferroni correction of three tests according to the simulation results as follows.

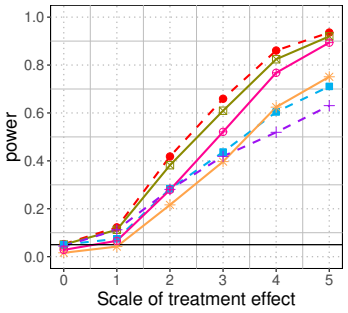
We tried four classical combinations [Vovk and Wang, 2020b]:

- arithmetic mean $2(p_1 + p_2 + p_3)/3$, and
- geometric mean $e(p_1 \cdot p_2 \cdot p_3)^{1/3}$, and
- harmonic mean $e\log(3) \cdot 3/(p_1^{-1} + p_2^{-1} + p_3^{-1})$, and
- Bonferroni correction $3\min(p_1, p_2, p_3)$.

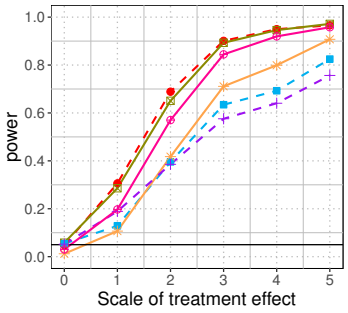
Note that because p-values from the Wilcoxon tests can have arbitrary dependence, we cannot use some alternative meta methods such as Fisher's or Simes' methods. Nonetheless, the above-presented combinations are valid under arbitrary dependence.



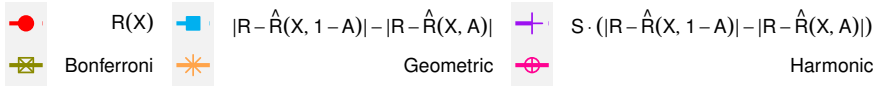
(a) Dense effect under Gaussian noise and bell-shaped control outcome.

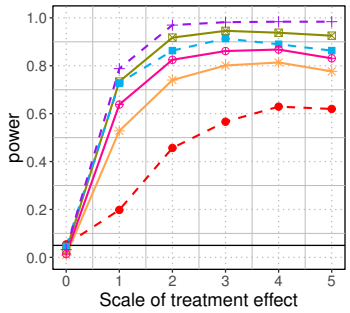


(b) Dense effect under Cauchy noise and bell-shaped control outcome.

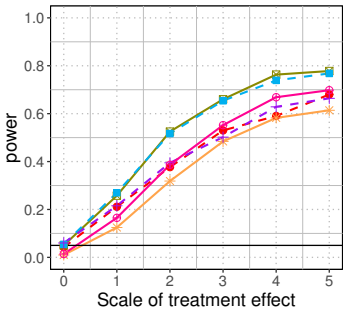


(c) Dense effect under Gaussian noise and skewed control outcome.

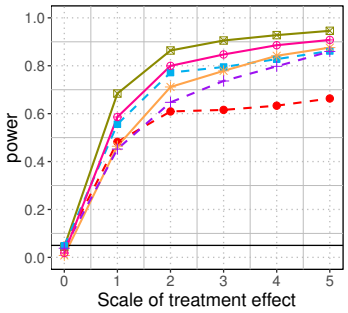




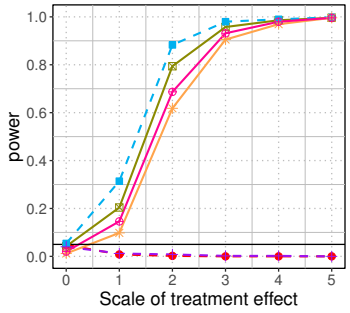
(d) Sparse effect under Gaussian noise and bell-shaped control outcome.



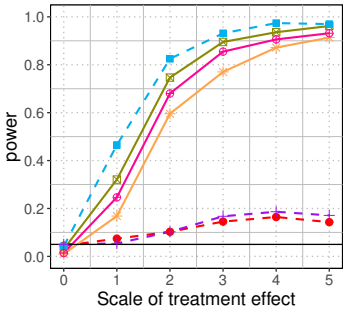
(e) Sparse effect under Cauchy noise and bell-shaped control outcome.



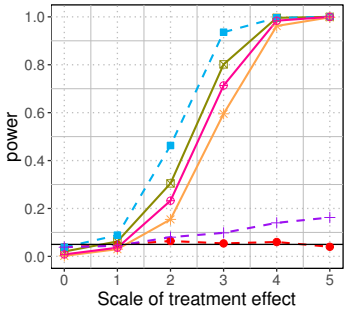
(f) Sparse effect under Gaussian noise and skewed control outcome.



(g) Sparse strong positive and dense weak negative effects.



(h) Sparse strong effect of both signs.



(i) Dense weak effect of both signs.

Figure 53: Power of the Wilcoxon test using $E_i^{R(X)}$, $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$, $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ and four meta tests that combine these three tests (p-values) by arithmetic mean (not shown due to low power), geometric mean, harmonic mean, Bonferroni correction, under different types of treatment effect with the scale of treatment effect S_Δ increases. In all simulations, the Bonferroni correction leads to similar power as the recommended test ($E_i^{R(X)}$ for dense effect in the first row, $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ for sparse effect in the second row, and $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ for two-sided effect in the third row).

We follow the same simulation settings in the main paper. In all simulations, the Bonferroni correction leads to similar power as the recommended test (See Figure 53 for simulation results). Thus,

we would recommend the Bonferroni correction when there is little prior knowledge of the treatment effect type.

One may notice that the power of the Bonferroni correction is even higher than all of the candidate Wilcoxon tests when the treatment effect is sparse and the control outcome is skewed (see Figure 53f). While this may initially seem impossible, it is actually possible when different tests perform well on different instances. To elaborate, this effect arises because the power curves are averaged over many repetitions, and in each repetition different Wilcoxon tests have lower p-values (higher chance of rejection) under different data realization. The low p-values from different tests are all captured by taking the minimum as in the Bonferroni correction. Specifically, we found that although the test with $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ has the highest power among three Wilcoxon tests, there are more than 10% of repeated experiments such that the p-value of the recommended one is larger than α (no rejection using this individual test), while the minimum of p-values from the other two is smaller than $\alpha/3$ (rejection using the Bonferroni correction).

C.7 Experiments for the i-Wilcoxon test under heavy-tailed noise

In the automated algorithm of i-Wilcoxon test, we recommend using the robust regression because it is less sensitive to skewed control outcomes, as shown by Figure 25c. Here, we show that the robust regression also makes the i-Wilcoxon test more robust to heavy-tailed noise (see Figure 54).

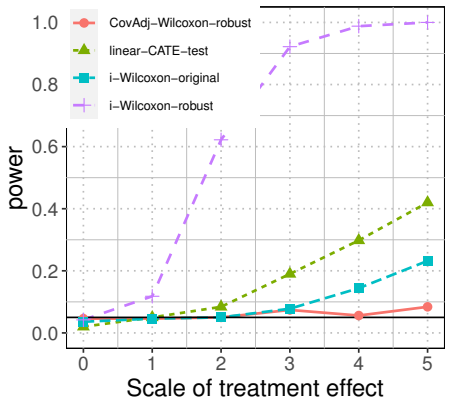
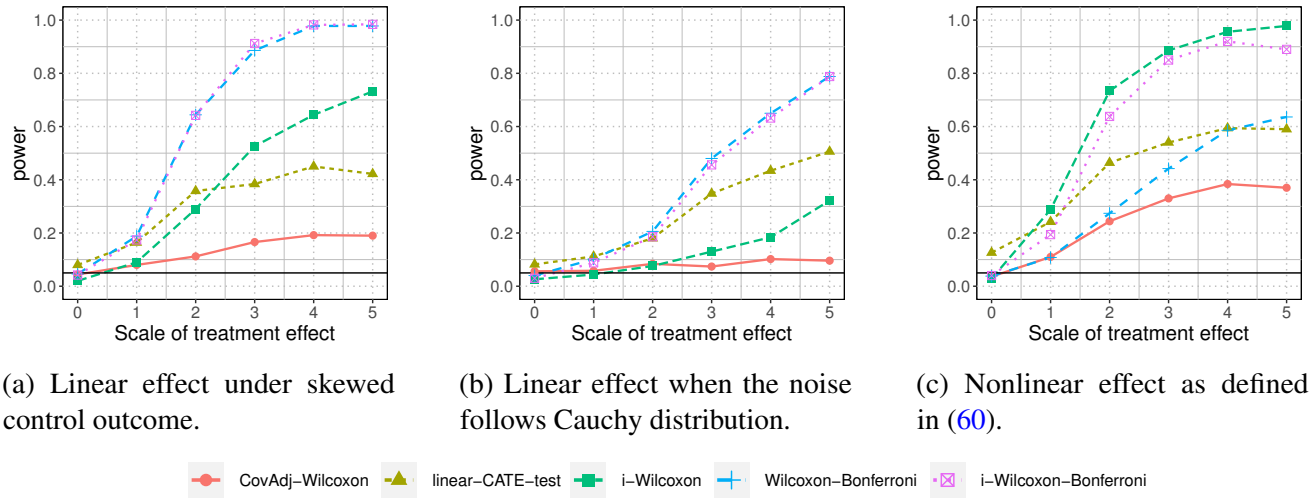


Figure 54: Power of the i-Wilcoxon test using regular linear regression and robust linear regression compared with standard methods. The outcome simulates from (39), where the function of treatment effect Δ and the function of control outcome f are linear as defined in (57) and (58). Instead of Gaussian noise in Section 4.2.2, the noise U_i is now simulated from a Cauchy distribution. The i-Wilcoxon test with robust linear regression has higher power than that using regular linear regression under heavy-tailed noise. For fair comparison, the CovAdj Wilcoxon test is also implemented with robust linear regression.

C.8 Numerical experiments under small sample sizes

The experiments in the main paper generates 500 samples, and here, We present the results of the same experiments with a smaller sample size $n = 50$ (the size of signals is doubled for a clear power comparison). We focus on checking whether the power comparison among different methods is consistent with the results under the large sample size in the main paper.

Numerical experiments for the i-Wilcoxon test. For the discussion of the i-Wilcoxon test, one of the comparison methods, the linear-CATE-test, no longer has type-I error control as its validity only holds asymptotically (see Figure 55). The power of the i-Wilcoxon test decreases compared to the cases with a large sample size in the main paper. Upon a closer examination, it appears that this decrease in power is primarily because the potential outcomes are not well estimated by our interactive algorithm; this cannot hurt validity but it does hurt power, as explained next. Recall that in the i-Wilcoxon test, we infer the treatment assignments by fitting a mixture model (54) on the outcomes, and use the EM algorithm to alternate between estimating the function for potential outcomes (θ_0, θ_1) and predicting the assignments A_i. We observe that poor estimation of the potential outcomes often leads to a poor prediction of A_i, and hence the resulting ordering tends to include more subjects with incorrect $\widehat{A}_i$ at the front instead of the correct ones. As a result, the cumulative sum S_t in (52) would grow slower, so it becomes harder to exceed the boundary $u_{\alpha/2}(t)$ in order to reject. Take an example of the experiment under Cauchy noise. In our original experiment with $n = 500$ (Figure 54), we often need to accumulate more than 100 subjects for S_t to exceed the boundary and reject the null (see Figure 56a). Thus, such a difference is not evident enough when we can at most accumulate 50 subjects when $n = 50$ (see Figure 56b). As a comparison, if an oracle knows the true potential outcomes $\theta_0(X_i) = f(X_i)$, $\theta_1 = \Delta(X_i) + f(X_i)$, where Δ and f are in model (39) to generate the data in our experiment, and uses the same mixture model (54) to infer assignments and order subjects, the resulting power under small sample size improves significantly (see Figure 56d), because it allows for the treated subjects to be better gathered earlier in the ordering (see Figure 56c).



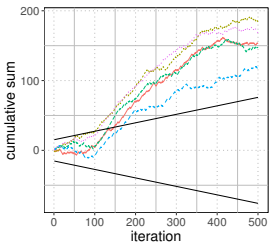
(a) Linear effect under skewed control outcome.

(b) Linear effect when the noise follows Cauchy distribution.

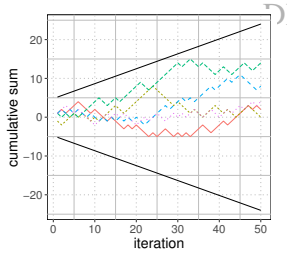
(c) Nonlinear effect as defined in (60).

Figure 55: Power of the i-Wilcoxon test compared with the standard tests (the linear-CATE-test and the CovAdj Wilcoxon test) when varying the scale of the treatment effect under various situations of the outcome distribution (the small-sample-size version of experiments in Figure 25, Figure 26, and Figure 54). The sample size is set to be as small as $n = 50$. As a result, the linear-CATE-test does not have valid type-I error control, and the power of i-Wilcoxon test decreases, but its power still tends to be higher than others when the effect is nonlinear. Additionally, we recommend the Bonferroni correction of the i-Wilcoxon test and the permutation-based Wilcoxon tests (i-Wilcoxon-Bonferroni).

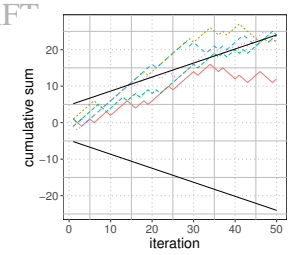
Still, when the treatment effect is a nonlinear function of the covariates, the difference in the cumulative sums is usually detected before including 50 subjects in our original experiment when $n = 500$, so the i-Wilcoxon test can preserve its advantage and have high power under a small sample size $n = 50$. Moreover, we find that the Wilcoxon-Bonferroni test (our proposed Wilcoxon tests



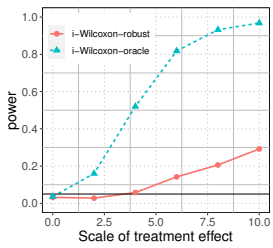
(a) The cumulative sums of five instances under large sample size ($n = 500$).



(b) The cumulative sums of five instances under small sample size ($n = 50$).



(c) The cumulative sums of five instances under small sample size ($n = 50$) in an oracle case where the potential outcomes are known.



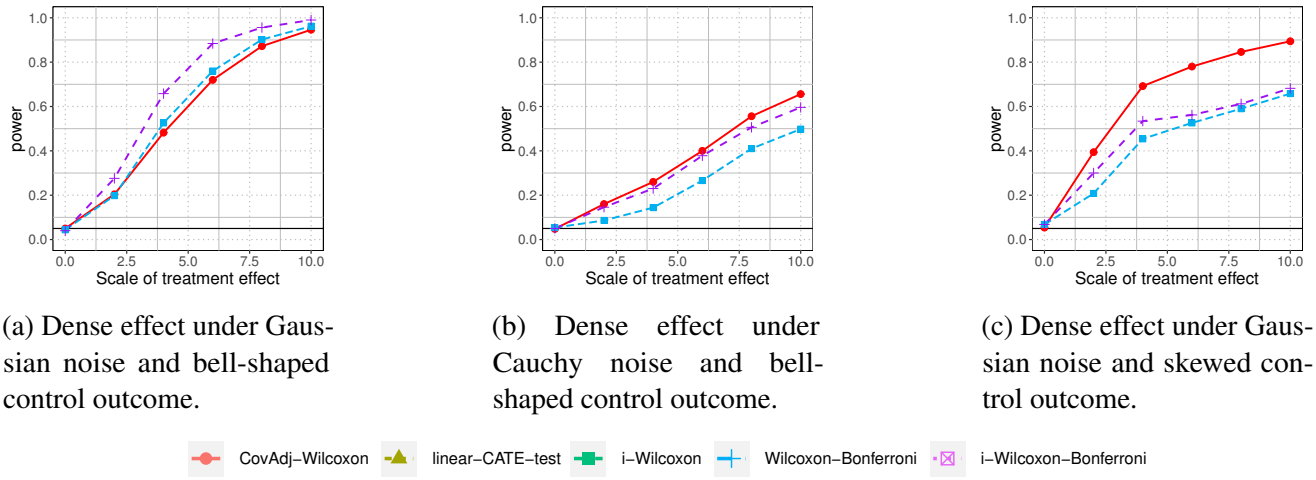
(d) Power of the interactive test with or without oracle knowledge on potential outcomes.

Figure 56: Diagnostics of the low power of the interactive test under small sample size when the noise follows Cauchy distribution. Because heavy-tailed noise makes it harder to learn the potential outcomes, the cumulative sum of treatment assignments usually exceeds the boundaries (black lines) after including 100 subjects; thus not detectable when the total number of subjects is small ($n = 50$).

combined by Bonferroni correction as discussed in Appendix C.6) can have higher power in several cases (see Figure 55). Specifically, recall that the Wilcoxon tests we propose in (61) involve the residual R_i and two types of its estimation $\widehat{R}(X, A),\ \widehat{R}(X, 1-A)$, which can be obtained from an arbitrary regression model. Here, we use robust linear regression for a fair comparison because other methods also use linear regression as a default (note that the validity of all methods is not affected by whether the linear model is the underlying truth). The Wilcoxon-Bonferroni test leads to higher power than other methods when the noise is Cauchy or the control outcome is skewed. In the case where the treatment effect is nonlinear, its advantage is less evident compared to the interactive test with quadratic regression because the former uses linear regression. This comparison under the nonlinear effect also indicates the advantage of the interactive test we demonstrate in Figure 26: the interactive test can explore data information to decide which model to use (e.g., correctly choose the quadratic regression in this example), whereas most existing methods including the Wilcoxon-Bonferroni test commit to a pre-specified model (e.g., linear model in this example).

Overall, we would recommend the Bonferroni correction of the i-Wilcoxon test and the non-interactive permutation-based tests for experiments with small sample sizes (the computation cost of the permutation tests is also not heavy with small sample sizes).

Numerical experiments for the Wilcoxon tests. For the discussion of permutation tests in Section 4.3, the power comparison has the same pattern as the case with a large sample size (see Figure 57). As summarized in flowchart (74), we still recommend using the Wilcoxon test with $E_i^{R(X)}$ for dense effect, $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ for sparse effect, and $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ for two-sided effect.



(a) Dense effect under Gaussian noise and bell-shaped control outcome.

(b) Dense effect under Cauchy noise and bell-shaped control outcome.

(c) Dense effect under Gaussian noise and skewed control outcome.

(d) Sparse effect under Gaussian noise and bell-shaped control outcome.

(e) Sparse effect under Cauchy noise and bell-shaped control outcome.

(f) Sparse effect under Gaussian noise and skewed control outcome.

(g) Sparse strong positive and dense weak negative effects.

(h) Sparse strong effect of both signs.

(i) Dense weak effect of both signs.

Figure 57: Power of the candidate Wilcoxon test using three choices of E_i under different types of treatment effect with the scale of treatment effect S_Δ increases. The sample size is set to be small as $n = 50$, but the power comparison is similar to the previous experiments with $n = 500$: we recommend using the Wilcoxon test with $E_i^{R(X)}$ for dense effect (the first row), $E_i^{S\cdot(|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|)}$ for sparse effect (the second row), and $E_i^{|\widehat{R}(X,1-A)-R|-|\widehat{R}(X,A)-R|}$ for two-sided effect (the third row).

C.9 The Kruskal-Wallis test for multi-sample comparison without block structure

The Kruskal-Wallis test considers the ranks of all observations. For subjects with treatment a, let the sample size be $N_a = \sum_{i=1}^n \mathbb{1}(A_i = a)$ and the average rank be $\overline{RK(a)} = \frac{1}{N_a}\sum_{i=1}^n \text{rank}(Y_i)\mathbb{1}(A_i = a)$. Denote the overall averaged rank as $\overline{RK} = \frac{1}{n}\sum_{i=1}^n \text{rank}(Y_i)$. The test statistic is

$$H = (n-1)\frac{\sum_{a=1}^k N_a \left(\overline{RK(a)} - \overline{RK}\right)^2}{\sum_{i=1}^n \left(\text{rank}(Y_i) - \overline{RK}\right)^2}, \tag{196}$$

which measures the relative variation across blocks and is expected to be large under the alternative. Thus, the Kruskal-Wallis test rejects the null if H is larger than a threshold. The threshold is obtained from the null distribution of H, which can be derived if the sample size is small; otherwise, it is approximated by a chi-squared distribution.

C.10 The Friedman test for multi-sample comparison with block structure

The Friedman test considers the ranks *within* each block $\{Y_{i1}, \ldots, Y_{ik}\}$, denoted as $\text{rank}(Y_{ij})$. Let the rank of the subjects with treatment a averaged over n blocks be $\overline{RK(a)} = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^k \text{rank}(Y_{ij})\mathbb{1}\left(A_{ij} = a\right)$, and its expected value under the null is $\frac{1+k}{2}$. Under the alternative, the outcomes for one of the treatment could be larger (or smaller) than those for other treatments and the averaged rank would be higher (or lower). The Friedman test computes:

$$F = \sum_{a=1}^k \left(\overline{RK(a)} - \frac{1+k}{2}\right)^2,$$

and reject the null if F is larger than a threshold obtained by the null distribution of F, which is approximated by a Chi-square when n or k is large.

C.11 Error control of the seq-Wilcoxon test

In the sequential setting, we try to argue that even though we can filter the subjects to enter the sum in a data-dependent manner, denoted by decision I_i, it does not affect the behavior of the cumulative sum (sum of independent (weighted) coin flips). Intuitively, it is because the decision I_{t+1} is based on the σ-field $\mathcal{G}_t$, which is independent of A_{t+1} that we potentially would cumulate. We formalize this intuition as follows.

By definition, only when $I_t = 1$, the sum S_t changes its value and the boundary $u_{\alpha/2}\left(\sum_{i=1}^t I_i\right)$ updates, so the algorithm can only stop at τ when there is a new increment ($I_\tau = 1$). Thus, we can measure the time of the martingale sequence different from t by ignoring the subjects that are filtered out ($I_i = 0$). Let the new "time" be $v = 1, 2, \ldots$ and define a random time T_v in terms of I_i's:

$$T_v := \min\left\{t \in \mathbb{N} : \sum_{i=1}^t I_i \geq v + 1\right\} - 1. \tag{197}$$

In words, we count time v only before there is a new increment, which comes from the $(T_v + 1)$-th subject. Consequently, we have $\sum_{i=1}^{T_v} I_i = v$. Let the sum be $\widetilde{S}_v := S_{T_v} \equiv \sum_{i=1}^{T_v} I_i(2A_i - 1) \cdot w_i$, and

by definition, there are v number of nonzero increment in $\widetilde{S}_v$. Under this notation, the stopping time τ for rejection can be equivalently defined as $\tau \equiv T_\nu$ where

$$\nu := \min\left\{v \in \mathbb{N} : |\widetilde{S}_v| > u_{\alpha/2}(v)\right\}. \tag{198}$$

The test rejects the null if $T_\nu < \infty$, which is equivalent as $\nu < \infty$. Thus, the proof of error control boils down to proving that under the null, $\mathbb{P}\left(\exists v \in \mathbb{N} : |\widetilde{S}_v| > u_{\alpha/2}(v)\right) \leq \alpha$, or equivalently that $\{\widetilde{S}_v\}$ is a martingale with weighted coin flips as increments. Define the filtration as $\widetilde{\mathcal{G}}_v := \sigma\left(\mathcal{G}_{T_v} \cup \{T_1, \ldots, T_v\}\right)$, we prove that $\{\widetilde{S}_{v+1}\}$ is a martingale with respect to the filtration $\{\widetilde{\mathcal{G}}_v\}$.

Proof. We first argue that $\widetilde{S}_{v+1}$ is measurable with respect to $\widetilde{\mathcal{G}}_v$. By definition, the last nonzero increment in $\widetilde{S}_{v+1}$ comes from the (T_v+1)-th subject, so $\widetilde{S}_{v+1} \equiv S_{T_v+1}$. And S_{T_v+1} is measurable with respect to $\widetilde{\mathcal{G}}_v$ because I_{T_v+1} has its distribution with respect to $\widetilde{\mathcal{G}}_v$. Next, we show that $\mathbb{E}(\widetilde{S}_{v+1} \mid \widetilde{\mathcal{G}}_v) = \widetilde{S}_v$, which boils down to the claim that

$$\mathbb{E}\left(2A_{T_v+1} - 1 \mid \widetilde{\mathcal{G}}_v\right) = 0, \tag{199}$$

because $\widetilde{S}_{v+1} = S_{T_v+1}$ and $I_{T_v+1} = 1$, and that w_{T_v+1} is $\widetilde{\mathcal{G}}_v$-measurable. Note that conditional on T_v, the information in $\widetilde{\mathcal{G}}_v$ is independent of A_{T_v+1}. Thus, we derive that $mathbbE\left(2A_{T_v+1} - 1 \mid \widetilde{\mathcal{G}}_v\right) = \mathbb{E}\left(2A_{T_v+1} - 1 \mid T_v\right)$. For any $t \in \mathbb{N}$, we claim that

$$\mathbb{E}\left(2A_{T_v+1} - 1 \mid T_v = t\right) = 0,$$

because

$$\begin{aligned}
&\mathbb{E}\left(2A_{T_v+1} - 1 \mid T_v = t\right) = \mathbb{E}\left(2A_{t+1} - 1 \mid T_v = t\right) \\
\overset{(a)}{=}\ &\mathbb{E}\left[\mathbb{E}\left(2A_{t+1} - 1 \middle| I_1, \ldots, I_t, \sum_{i=1}^{t} I_i = v, I_{t+1} = 1\right) \middle| T_v = t\right] \\
\overset{(b)}{=}\ &\mathbb{E}\left[\mathbb{E}\left(\mathbb{E}\left(2A_{t+1} - 1 \mid \mathcal{G}_t\right) \middle| I_1, \ldots, I_t, \sum_{i=1}^{t} I_i = v, I_{t+1} = 1\right) \middle| T_v = t\right] \overset{(c)}{=} 0,
\end{aligned}$$

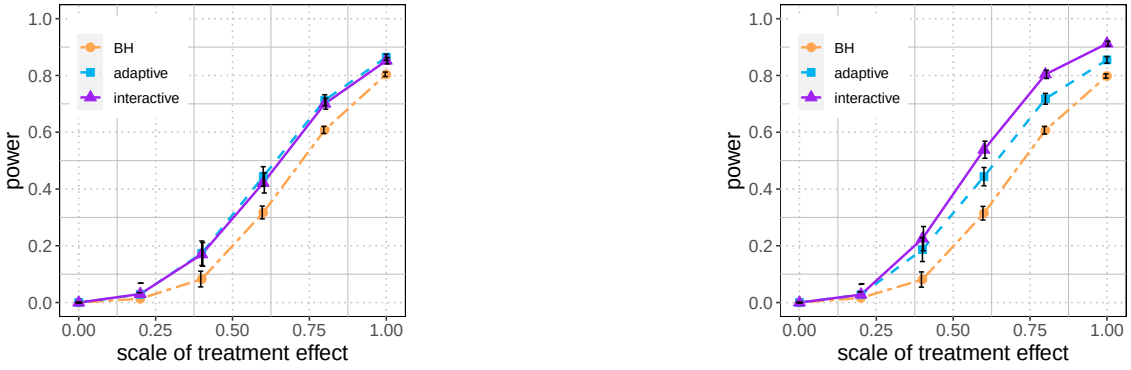
where (a) holds because $\{T_v = t\}$ is implied by $\bigcup_{i=1}^{t+1}\{I_i\}$ that satisfy $\sum_{i=1}^{t} I_i = v$ and $I_{t+1} = 1$; and (b) is because I_{i+1} is measurable with respect to $\mathcal{G}_i$ for each $i \in [t]$, and $\mathcal{G}_1 \subseteq \ldots \subseteq \mathcal{G}_t$; to see (c), notice that under the null, A_{t+1} is independent of $\mathcal{G}_t$ and $\mathbb{E}(A_{t+1}) = 0$; thus, we prove that $\mathbb{E}\left(2A_{T_v+1} - 1 \mid T_v = t\right) = 0$. Notice that the increment of $\widetilde{S}_v$ takes value in $\{\pm 1\}$ with zero mean value, so its distribution is $\{\pm 1\}$ with equal probability. Thus, boundary $u_{\alpha/2}(v)$ for the sum of independent, fair coin flips leads to valid error control for Algorithm 8. □

D Appendix for "Interactive identification of individuals with positive treatment effect while controlling false discoveries"

D.1 Details in the extensions of the I^3

D.1.1 FDR control at a subgroup level

Here, we provide explanation of the higher power achieved by the interactive procedure. Although the interactive procedure and the BH procedure define the same set of subgroups and corresponding p-values, the interactive procedure has two properties that potentially improve the power from the BH procedure: (a) it excludes possible null subgroups so that it can be less sensitive to a large number of nulls, whereas the BH procedure considers all the subgroups at once; (b) the interactive procedure additionally uses the covariates. We can separately evaluate the effect of the above two properties by implementing two versions of Algorithm 14, which differ in the strategy to select subgroups in step a. Specifically, the adaptive procedure selects the subgroup whose revealed (partial) p-value P_g^1 is the smallest (not using the covariates); and the interactive procedure selects the subgroup by an estimated probability of the P_g^2 to be positive (using the revealed P_g^1, the covariates, and the outcomes).



(a) Effect as discrete function of covariates.

(b) Effect as a simpler function of covariates.

Figure 58: Power of two methods for subgroup identification: the BH procedure proposed by Karmakar et al. [2018], the adaptive procedure, and the interactive procedure under different types of treatment effect (we define 80 subgroups by discrete values of the covariates). Our proposed interactive procedure tends to have higher power than the BH procedure because (1) it excludes possible nulls (shown by higher power of the adaptive procedure than the BH procedure in both plots); and (2) it additionally uses the covariates (shown when the treatment effect can be well learned as a function of covariates in the right plot).

To see if both properties of Algorithm 14 contribute to the improvement of power from the BH procedure, we tested the methods under two simulation settings. Recall that the previous experiment defines a positive treatment effecte when the discrete covariate $X_i(1) \in \{1, \ldots, 40\}$ is even. Here, we add another case where the treatment effect is positive when $X_i(1) \leq 20$, so that the density of subgroups with positive effects is the same as previous, but the treatment effect is a simpler function of the covariates. Hence in the latter case, we would expect the interactive procedure learn this function of covariates rather accurately, and have higher power than the adaptive procedure which does not use the covairates; as confirmed in Figure 58b. In the former simulation setting where the treatment effect is

not a smooth function of the covariates and hard to be learned, the adaptive procedure and interactive procedure have similar power (Figure 58a). Still, they have higher power than the BH procedure because they exclude possible null subgroups.

D.1.2 An automated algorithm with FDR control on nonpositive effects

We have proposed the MaY-I^3 to guarantee a valid FDR control for the nonpositive-effect null in (111), by reducing the available information for selecting subjects from $\mathcal{F}_{t-1}(\mathcal{I})$ for the Crossfit-I^3 to $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$ (recall it no longer includes the outcomes of candidate subjects). Here, we present an automated algorithm to select a subject for the MaY-I^3.

One naive strategy is to follow Algorithm 11 in the main paper, which is designed for the Crossfit-I^3, with the outcomes removed from the predictors; however, it appears to result in less accurate prediction of the effect signs, and in turn rather low power (numerial results are in the next paragraph). Here, we take a different approach by predicting the treatment effect instead of their signs, because the treatment effect might be better predicted as a function of the covariates (without outcomes) than a binary sign, especially when the treatment effect is indeed a smooth and simple function of the covariates. Specifically, we first estimate the treatment effect for the non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$ using a well-studied doubly-robust estimator (see Kennedy [2020] and references therein):

$$\Delta_j^{\mathrm{DR}} = 4(A_j - 1/2) \cdot (Y_j - \widehat{\mu}_A(X_j)) + \widehat{\mu}_1(X_j) - \widehat{\mu}_0(X_j), \tag{200}$$

where $(\widehat{\mu}_0, \widehat{\mu}_1)$ are random forests trained to predict the outcomes for the control and treated group, respectively. Using the provided covariates X_i, we can predict Δ_i^{DR} for the candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$. The subject with the smallest prediction of Δ_i^{DR} is then excluded. This automated strategy is described in Algorithm 16.

Algorithm 16 An automated heuristic to select i_t^* in the MaY-I^3.

Input: Current rejection set $\mathcal{R}_{t-1}(\mathcal{I})$, and available information for selection $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$;
Procedure:
1. Estimate the treatment effect for non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$ as Δ_j^{DR} in (200);
2. Train a random forest where the label is the estimated effect Δ_j^{DR} and the predictors are the covariates X_j, using non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;
3. Predict Δ_i^{DR} for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$ via the above random forest, denoted as $\widehat{\Delta}_i^{\mathrm{DR}}$;
4. Select i_t^* as $\mathrm{argmin}\{\widehat{\Delta}_i^{\mathrm{DR}} : i \in \mathcal{R}_{t-1}(\mathcal{I})\}$.

To summarize, we have presented two types of strategy for selecting subjects: the Crossfit-I^3 chooses the one with the smallest predicted probability of a positive $\widehat{\Delta}_i$ (see Algorithm 11 in the main paper), which we denote here as the *min-prob strategy*; and the MaY-I^3 chooses the one with the smallest prediction of estimated effect Δ_j^{DR} (see Algorithm 16), which we denote here as the *min-effect strategy*. Note that the proposed interactive algorithm can use arbitrary strategy as long as the available information for selection is restricted. That is, the Crossfit-I^3 can use the same min-effect strategy, and the MaY-I^3 can use the min-prob strategy (after removing the outcomes from the predictors, which we elaborate in the next paragraph). However, we observe in numerical experiments that both interactive procedures have higher power when using their original strategies, respectively (see Figure 59).

Before details of the experiment results, We first describe the min-prob strategy for the MaY-I^3, where the available information $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$ does not include the outcomes for candidate subjects. Similar

to the min-prob strategy in Algorithm 11 of the main paper, we hope to use the outcome Y_i and residual $E_i = Y_i - \widehat{m}(X_i)$ as predictors, and predict the sign of treatment effect for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$, but Y_i and E_i for the candidate subjects are not available in $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$. Thus, we propose algorithm 17, where we first estimate Y_i and E_i using the covariates (see step 1-2); and step 3-5 are similar to Algorithm 11, which obtain the probability of having a positive treatment effect.

Algorithm 17 The min-prob strategy to select i_t^* in the MaY-I^3.

Input: Current rejection set $\mathcal{R}_{t-1}(\mathcal{I})$, and available information for selection $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$;
Procedure:
1. Predict the outcome Y_k of each subject $k \in [n]$ by covariates, denoted as $\widehat{Y}^{-\mathcal{I}}(X_k)$, where $\widehat{Y}^{-\mathcal{I}}$ is learned using non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;
2. Predict the residual $E_k = Y_k - \widehat{m}(X_K)$ of each subject $k \in [n]$ by covariates, denoted as $\widehat{E}^{-\mathcal{I}}(X_k)$, where $\widehat{E}^{-\mathcal{I}}$ is learned using non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;
3. Train a random forest classifier where the label is $\text{sign}(\widehat{\Delta}_j)$ and the predictors are $\left(\widehat{Y}^{-\mathcal{I}}(X_j), X_j, \widehat{E}^{-\mathcal{I}}(X_j)\right)$, using non-candidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;
4. Predict the probability of $\widehat{\Delta}_i$ being positive as $\widehat{p}(i,t)$ for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$;
5. Select $i_t^* = \text{argmin}\{\widehat{p}(i,t) : i \in \mathcal{R}_{t-1}(\mathcal{I})\}$.

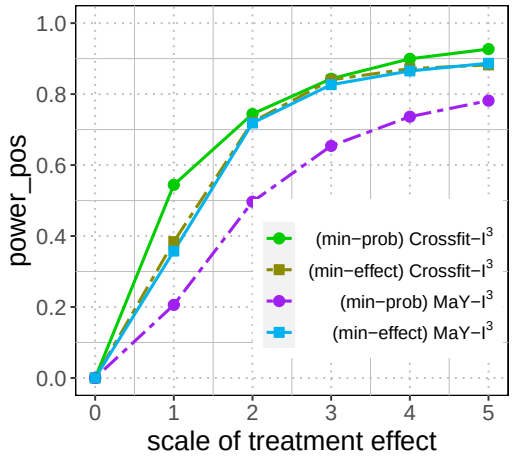


Figure 59: Power of the Crossfit-I^3 and MaY-I^3 with two strategies to select subjects: the min-prob strategy and the min-effect strategy, under the treatment effect defined in (110) of the main paper with the scale S_Δ varies in $\{0, 1, 2, 3, 4, 5\}$. The Crossfit-I^3 tends to have higher power when using the min-prob strategy, and the MaY-I^3 tends to have higher power when using the min-effect strategy.

The Crossfit-I^3 has higher power when using the min-prob strategy than the min-effect strategy because the former additionally uses the outcome as a predictor. For the MaY-I^3, the min-effect strategy leads to higher power because the estimated treatment effect Δ_j^{DR} in (200) can provide reliable evidence of which subjects have a positive effect. If using the min-prob strategy, it could be harder to learn an accurate prediction by Algorithm 17 where two of the predictors $\widehat{Y}^{-\mathcal{I}}(X_j)$ and $\widehat{E}^{-\mathcal{I}}(X_j)$ are obtained by estimation, increasing the complexity in modeling. Therefore, we present the Crossfit-I^3 and MaY-I^3 with the min-prob and min-effect strategies, respectively, as preferred in numerical experiments. Nonetheless, we remark that our proposed interactive frameworks for the Crossfit-I^3 and MaY-I^3 allow

arbitrary strategies to select subjects, and an analyst can design her own strategy based on her domain knowledge.

D.1.3 FDR control of nonpositive effects for paired samples

Recall the nonpositive-effect null under paired samples:

$$H_{0i}^{(\text{nonpositive, paired})} : (Y_{ij} \mid A_{ij} = 1, X_{ij}) \preceq (Y_{ij} \mid A_{ij} = 0, X_{ij}) \text{ for both } j = 1, 2,$$

and we observe that

$$\mathbb{P}(\widehat{\Delta}_i^{\text{paired}} > 0 \mid \{X_{j1}, X_{j2}\}_{j=1}^n) \leq 1/2, \tag{201}$$

where $\widehat{\Delta}_i^{\text{paired}}$ is defined in (134) of the main paper. Thus, the MaY-I^3 with $\widehat{\Delta}_i$ replaced by $\widehat{\Delta}_i^{\text{paired}}$ has valid FDR control for the nonpositive-effect null, where the analyst progressively excludes pairs using the available information:

$$\mathcal{F}_{t-1}^{-Y,\text{paired}} = \sigma\left(\{X_{i1}, X_{i2}\}_{i\in\mathcal{R}_{t-1}}, \{Y_{j1}, Y_{j2}, A_{j1}, A_{j2}, X_{j1}, X_{j2}\}_{j\notin\mathcal{R}_{t-1}}, \sum_{i\in\mathcal{R}_{t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i^{\text{paired}} > 0\}\right).$$

We can also implement an automated version of the MaY-I^3 where the selection of the excluded subject follows a similar procedure as Algorithm 16. The difference is that in step 1, we estimate the treatment effect for non-candidate subjects $j \notin \mathcal{R}(\mathcal{I})$ directly as $\widehat{\Delta}_i^{\text{paired}} \equiv (A_{i1} - A_{i2})(Y_{i1} - Y_{i2})$ instead of Δ_i^{DR} to avoid estimating outcomes in $(\widehat{\mu}_0, \widehat{\mu}_1)$.

D.2 Proof of FDR control with 1/2 propensity scores

The proofs are based on an optional stopping argument, as a variant of the ones presented in Lei and Fithian [2018], Lei et al. [2020], Li and Barber [2017] and Barber and Candès [2015].

Lemma 9 (Lemma 2 of Lei and Fithian [2018])**.** *Suppose that, conditionally on the σ-field $\mathcal{G}_{-1}$, $b_1, \ldots, b_n$ are independent Bernoulli random variables with*

$$\mathbb{P}(b_i = 1 \mid \mathcal{G}_{-1}) = \rho_i \geq \rho > 0, \text{ almost surely.}$$

Let $(\mathcal{G}_t)_{t=0}^{\infty}$ be a filtration with $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots$ and suppose that $[n] \supseteq \mathcal{C}_0 \supseteq \mathcal{C}_1 \supseteq \ldots$, with each subset $\mathcal{C}_{t+1}$ measurable with respect to $\mathcal{G}_t$. If we have

$$\mathcal{G}_t = \sigma\left(\mathcal{G}_{-1}, \mathcal{C}_t, (b_i)_{i\notin\mathcal{C}_t}, \sum_{i\in\mathcal{C}_t} b_i\right), \tag{202}$$

and τ is an almost-surely finite stopping time with respect to the filtration $(\mathcal{G}_t)_{t\geq 0}$, then

$$\mathbb{E}\left[\frac{1 + |\mathcal{C}_\tau|}{1 + \sum_{i\in\mathcal{C}_\tau} b_i}\middle| \mathcal{G}_{-1}\right] \leq \rho^{-1}.$$

D.2.1 Proof of theorem 10

Proof. We show that the I^3 controls FDR by Lemma 9, where

$$b_i := \mathbb{1}\{(A_i - 1/2) \cdot E_i \leq 0\} \text{ and } \mathcal{G}_{-1} := \sigma\left(\{Y_j, X_j\}_{j=1}^n\right) \text{ and } \mathcal{C}_t := \mathcal{R}_t \cap \mathcal{H}_0,$$

for $t = 0, 1, \ldots$. The assumptions in Lemma 9 are satisfied: (a) $\mathbb{P}(b_i = 1 \mid \mathcal{G}_{-1}) \geq 1/2$ for subjects with zero effect $i \in \mathcal{H}_0$:

$$\begin{aligned} & \mathbb{P}\left((A_i - 1/2) \cdot E_i \leq 0 \mid \mathcal{G}_{-1}\right) \\ = \; & \mathbb{P}(A_i = 1)\mathbb{1}(E_i \leq 0 \mid \mathcal{G}_{-1}) + \mathbb{P}(A_i = 0)\mathbb{1}(E_i \geq 0 \mid \mathcal{G}_{-1}), \\ & \text{because } A_i \text{ is independent of } \mathcal{G}_{-1} \\ = \; & 1/2\left[\mathbb{1}(E_i \leq 0 \mid \mathcal{G}_{-1}) + \mathbb{1}(E_i \geq 0 \mid \mathcal{G}_{-1}))\right] \geq 1/2; \end{aligned}$$

and (b) the filtration in our algorithm satisfies $\mathcal{F}_t \subseteq \mathcal{G}_t$, so the time of stopping the algorithm $\widehat{t} := \min\{t : \widehat{\text{FDR}}(\mathcal{R}_t) \leq \alpha\}$ is a stopping time with respect to $\mathcal{G}_t$; and (c) $\mathcal{C}_{t+1}$ is measurable with respect to $\mathcal{G}_t$. Thus, by Lemma 9, expectation, we have

$$\mathbb{E}\left[\frac{1 + |\mathcal{R}_{\widehat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\widehat{t}} \cap \mathcal{H}_0|} \,\middle|\, \mathcal{G}_{-1}\right] \leq 2,$$

By definition, the FDR conditional on $\mathcal{G}_{-1}$ at the stopping time $\widehat{t}$ is

$$\begin{aligned} & \mathbb{E}\left[\frac{|\mathcal{R}^+_{\widehat{t}} \cap \mathcal{H}_0|}{\max\{|\mathcal{R}^+_{\widehat{t}}|, 1\}} \,\middle|\, \mathcal{G}_{-1}\right] = \mathbb{E}\left[\frac{1 + |\mathcal{R}^-_{\widehat{t}} \cap \mathcal{H}_0|}{\max\{|\mathcal{R}^+_{\widehat{t}}|, 1\}} \cdot \frac{|\mathcal{R}^+_{\widehat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\widehat{t}} \cap \mathcal{H}_0|} \,\middle|\, \mathcal{G}_{-1}\right] \\ \leq \; & \mathbb{E}\left[\widehat{\text{FDR}}(\mathcal{R}_{\widehat{t}}) \cdot \frac{|\mathcal{R}^+_{\widehat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\widehat{t}} \cap \mathcal{H}_0|} \,\middle|\, \mathcal{G}_{-1}\right] \leq \alpha\mathbb{E}\left[\frac{|\mathcal{R}^+_{\widehat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\widehat{t}} \cap \mathcal{H}_0|} \,\middle|\, \mathcal{G}_{-1}\right] \\ = \; & \alpha\mathbb{E}\left[\frac{1 + |\mathcal{R}_{\widehat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\widehat{t}} \cap \mathcal{H}_0|} - 1 \,\middle|\, \mathcal{G}_{-1}\right] \leq \alpha, \end{aligned}$$

and the proof completes by applying the tower property of conditional expectation.

Notice that when the potential outcomes are treated as fixed, the same proof applies to the null defined as $Y_j^T = Y_j^C$, because the independence between A_i and $\mathcal{G}_{-1}$ still holds for the nulls. In the hybird version of the null (98) in the main paper, the above proof applies with $\mathcal{G}_{-1} := \sigma\left(\{Y_j, Y_j^T, Y_j^C, X_j\}_{j=1}^n\right)$. Thus, FDR is controlled at level α conditional on the potential outcomes and covariates $\{Y_j^T, Y_j^C, X_j\}_{j=1}^n$. □

D.2.2 Proof of theorem 11

Proof. Let the set of false rejections in $\mathcal{R}(\mathcal{I})$ be $\mathcal{V}(\mathcal{I})$. We conclude that the FDR of the I^3 implemented on set $\mathcal{I}$ is controlled at level $\alpha/2$:

$$\mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I})|}{\max\{|\mathcal{R}(\mathcal{I})|, 1\}} \,\middle|\, \mathcal{G}_{-1}\right] \leq \alpha/2,$$

following the error control of the I^3 in Section D.2.1, where the initial candidate rejection set is $R_0 = \mathcal{I}$, and thus, $C_0 = \mathcal{I} \cap \mathcal{H}_0$. Similarly, the FDR of the I^3 implemented on set $\mathcal{I}\mathcal{I}$ is also controlled at level

$\alpha/2$. Therefore, the FDR of the combined set $\mathcal{R}(\mathcal{I}) \cup \mathcal{R}(\mathcal{II})$ is controlled at level α as claimed:

$$\begin{aligned}
& \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I}) \cup \mathcal{V}(\mathcal{II})|}{|\mathcal{R}(\mathcal{I}) \cup \max\{|\mathcal{R}(\mathcal{II})|, 1\}} \,\middle|\, \mathcal{G}_{-1}\right] \\
\leq\ & \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I})|}{|\mathcal{R}(\mathcal{I}) \cup \max\{|\mathcal{R}(\mathcal{II})|, 1\}} \,\middle|\, \mathcal{G}_{-1}\right] + \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{II})|}{|\mathcal{R}(\mathcal{I}) \cup \max\{|\mathcal{R}(\mathcal{II})|, 1\}} \,\middle|\, \mathcal{G}_{-1}\right] \\
\leq\ & \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I})|}{\max\{|\mathcal{R}(\mathcal{I})|, 1\}} \,\middle|\, \mathcal{G}_{-1}\right] + \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{II})|}{|\max\{|\mathcal{R}(\mathcal{II})|, 1\}} \,\middle|\, \mathcal{G}_{-1}\right] \leq \alpha,
\end{aligned}$$

the proof completes for the null (96) in the main paper after applying the tower property of conditional expectation. The FDR control also applies to the other two definitions of the null (97) and (98) in the main paper, following the same arguments as the end of Section D.2.1. □

D.2.3 Proof of theorem 14

Proof. We prove that the FDR control holds for the I^3 implemented on $\mathcal{I}$, and the same conclusion applies to $\mathcal{II}$, so the overall FDR control is guaranteed following the proof of theorem 11 in Section D.2.2.

We first present the proof when the potential outcomes are viewed as random variables. Define $\mathcal{G}_{-1} := \sigma\left(\{X_i\}_{i=1}^n, \{Y_i, A_i\}_{i\notin\mathcal{I}}\right)$, and $\mathcal{G}'_t = \sigma\left(\mathcal{G}_{-1}, \mathcal{C}_t, (Y_i, A_i)_{i\notin\mathcal{C}_t}, \sum_{i\in\mathcal{C}_t} b_i\right)$, which contains more information than $\mathcal{G}_t$ as defined in (202). We claim that Lemma 9 holds when we replace $\mathcal{G}_t$ by $\mathcal{G}'_t$, because the distribution of b_i conditional on $\mathcal{G}_t$ is the same as on $\mathcal{G}'_t$ for any $t = 0, \ldots, n$. Similar to the proof of Theorem 10 in Section D.2.1, we check that the assumption in Lemma 9 are satisfied: (a) the filtration in our algorithm satisfies $\mathcal{F}_t \subseteq \mathcal{G}'_t$, so the time of stopping the algorithm $\widehat{t} := \min\{t : \widehat{\mathrm{FDR}}(\mathcal{R}_t) \leq \alpha\}$ is a stopping time with respect to $\mathcal{G}'_t$; and (b) $\mathcal{C}_{t+1}$ is measurable with respect to $\mathcal{G}'_t$; and (c) for subjects with nonpositive effect $i \in \mathcal{H}_0^{\text{nonpositive}}$:

$$\mathbb{P}\left((A_i - 1/2) \cdot E_i^{-\mathcal{I}} \leq 0 \mid \mathcal{G}_{-1}\right) \geq 1/2. \tag{203}$$

To see that the last assumption holds, notice that

$$\begin{aligned}
& \mathbb{P}\left((A_i - 1/2) \cdot (Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) \leq 0 \mid \mathcal{G}_{-1}\right) \\
=\ & \mathbb{P}(Y_i^C \geq \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1})\mathbb{P}(A_i = 0) + \mathbb{P}(Y_i^T \leq \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1})\mathbb{P}(A_i = 1); \text{ and} \\
& \mathbb{P}\left((A_i - 1/2) \cdot (Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) > 0 \mid \mathcal{G}_{-1}\right) \\
=\ & \mathbb{P}(Y_i^C < \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1})\mathbb{P}(A_i = 0) + \mathbb{P}(Y_i^C > \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1})\mathbb{P}(A_i = 1).
\end{aligned}$$

For any potential outcomes of the nulls such that $(Y_i^T \mid X_i) \preceq (Y_i^C \mid X_i)$, it holds that

$$\mathbb{P}(Y_i^C \geq D \mid X_i) \geq \mathbb{P}(Y_i^T > D \mid X_i), \text{ and } \mathbb{P}(Y_i^T \leq D \mid X_i) \geq \mathbb{P}(Y_i^C < D \mid X_i),$$

for any constant D, so

$$\begin{aligned}
& \mathbb{P}(Y_i^C \geq \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1}) \geq \mathbb{P}(Y_i^T > \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1}), \text{ and} \\
& \mathbb{P}(Y_i^T \leq \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1}) \geq \mathbb{P}(Y_i^C < \widehat{m}^{-\mathcal{I}}(X_i) \mid \mathcal{G}_{-1}),
\end{aligned}$$

because $m^{-\mathcal{I}}(X_i)$ is fixed given $\mathcal{G}_{-1}$. Because $\mathbb{P}(A_i = 1)$ is $1/2$ for all subjects, we have

$$\mathbb{P}\left((A_i - 1/2) \cdot (Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) \leq 0 \mid \mathcal{G}_{-1}\right) \geq \mathbb{P}\left((A_i - 1/2) \cdot (Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) > 0 \mid \mathcal{G}_{-1}\right),$$

which proves Claim (203) and in turn the FDR control of the MaY-I^3.

When the potential outcomes are treated as fixed, the above proof applies to the null defined as $Y_i^T \leq Y_i^C$ in (112) of the main paper, in which case $\mathbb{P}(Y_i^C \geq D \mid \mathcal{G}_{-1})$ is zero or one, and the above arguments still hold. For the hybird version of the null (113) in the main paper, the above proof applies with $\mathcal{G}_{-1} := \sigma\left(\left\{Y_i^T, Y_i^C, X_i\right\}_{i=1}^n, \{Y_i, A_i\}_{i\notin\mathcal{I}}\right)$. Thus, FDR is controlled at level α conditional on the potential outcomes and covariates $\{Y_j^T, Y_j^C, X_j\}_{j=1}^n$. □

D.2.4 Error control guarantee for the linear-BH procedure

Theorem 20. *Suppose the outcomes follow a linear model: $Y_i = l^\Delta(X_i)A_i + l^f(X_i) + U_i$, where l denotes a linear function, and U_i is standard Gaussian noise. The linear-BH procedure controls FDR of the nonpositive-effect null in (111) of the main paper asymptotically as the sample size n goes to infinity.*

Note that the error control would not hold when the linear assumption is violated. For example, if the expected treatment effect $\mathbb{E}(Y_i^T - Y_i^C \mid X_i)$ is some nonlinear function of the covariates, the estimated treatment effect $\widehat{\Delta}_i^{\text{BH}}$ would not be consistent; in turn, for the null subjects with zero effect, the p-values would not be valid (i.e., not stochastically equal or larger than uniform). Hence, the linear-BH procedure would not guarantee the desired FDR control, as we show in the numerical experiments in Section 5.3 of the main paper.

Proof. For simplicity, we treat all the covariates as fixed values and denote them as the covariance matrix $\mathbb{X}_a = (X_i : A_i = a)^T$ for $a \in \{T, C\}$, where we temporarily use $A_i = T$ to denote the case of being treated $A_i = 1$. Under the linear assumption, the estimated outcome $\widehat{l^a}$ asymptotically follows a Gaussian distribution, whose expected value is $l^\Delta(X_i)\mathbb{1}\{a = T\} + l^f(X_i)$. Its variance can be estimated as

$$\widehat{\text{Var}}(\widehat{l^a}(X_i)) = \widehat{\sigma}_a^2(X_i^T(\mathbb{X}_a^T\mathbb{X}_a)^{-1}X_i^T),$$

where the variance from noise is estimated as

$$\widehat{\sigma}_a^2 = \sum_{A_i=a}(Y_i - \widehat{l^a}(X_i))^2 \Bigg/ \left(\sum_i \mathbb{1}\{A_i = a\} - d - 1\right),$$

and d is the number of covariates. Note that the observed outcome also follows a Gaussian distribution $N(l^\Delta(X_i)\mathbb{1}\{a = T\} + l^f(X_i), \sigma_a^2)$. Note that in each estimated effect $\widehat{\Delta}_i^{\text{BH}}$, the observed outcome Y_i^a is independent of the estimated potential outcome $Y_i^{\overline{a}}$, where $\overline{a}$ is the complement of a: $\overline{a} \cup a = \{T, C\}$. Thus, the estimated effect asymptotically follow a Gaussian distribution whose expected value is $l^\Delta(X_i)$ (nonpositive under the null) and the variance is $\text{Var}(\widehat{\Delta}_i^{\text{BH}}) = \text{Var}(\widetilde{Y}_i^T) + \text{Var}(\widetilde{Y}_i^C)$, where an estimation is $\widehat{\text{Var}}(\widetilde{Y}_i^a) = \widehat{\sigma}_a^2\mathbb{1}\{A_i = a\} + \widehat{\text{Var}}(\widehat{l^a}(X_i))\mathbb{1}\{A_i = \overline{a}\}$. Therefore, the resulting p-value P_i as defined in (109) of the main paper is asymptotically valid (uniform or stochastically larger) if subject i is a null, and hence the BH procedure leads to asymptotic FDR control [Fan et al., 2007]. □

D.3 Proof of FDR control under heterogeneous propensity score

Lemma 10. *Let q_i be the conditional probability of a positive estimated sign:*

$$q_i := \mathbb{P}\left[(A_i - 1/2) \cdot E_i > 0 \mid \mathcal{F}_0(\mathcal{I})\right],$$

and the maximum be $q_{\max}(\mathcal{I}) := \max_{i\in\mathcal{I}} q_i$. Denote an estimation of $q_{\max}(\mathcal{I})$ using information in $\mathcal{F}_0(\mathcal{I})$ be $\widehat{q}_{\max}(\mathcal{I})$, and the (one-sided) estimation error be $\epsilon_n^q(\mathcal{I}) = \max\left\{q_{\max}(\mathcal{I}) - \widehat{q}_{\max}(\mathcal{I}), 0\right\}$. For the I^3

on set $\mathcal{I}$, redefine the FDR estimator as

$$\widehat{\mathrm{FDR}}(\mathcal{R}_t(\mathcal{I})) \equiv \left(\frac{1}{1-\widehat{q_{\max}}(\mathcal{I})} - 1\right) \frac{|\mathcal{R}_t^-(\mathcal{I})| + 1}{|\mathcal{R}_t^+(\mathcal{I})| \vee 1},$$

and stop the algorithm at $\tau := \inf\{t : \widehat{\mathrm{FDR}}(\mathcal{R}_t(\mathcal{I})) \leq \alpha/2\}$. Then, FDR for $\mathcal{R}_\tau^+(\mathcal{I})$ on set $\mathcal{I}$ is bounded:

$$\mathbb{E}\left[\frac{|\mathcal{R}_\tau^+(\mathcal{I}) \cap \mathcal{H}_0|}{|\mathcal{R}_\tau^+(\mathcal{I})| \vee 1}\middle| \mathcal{F}_0(\mathcal{I})\right] \leq \alpha/2 \left\{1 + \epsilon_n^q(\mathcal{I}) \cdot \frac{1}{q_{\max}(\mathcal{I})(1-q_{\max}(\mathcal{I}))}\right\}.$$

Same conclusion applies to the MaY-I^3 with residual $E_i^{-\mathcal{I}}$ and σ-field $\mathcal{F}_0^{-Y}(\mathcal{I})$.

Proof. By Lemma 9 where

$$b_i := \mathbb{1}\{(A_i - 1/2) \cdot E_i \leq 0\} \quad \text{and} \quad \mathcal{C}_t := \mathcal{R}_t \cap \mathcal{H}_0,$$

we have

$$\mathbb{E}\left[\frac{1 + |\mathcal{R}_\tau(\mathcal{I}) \cap \mathcal{H}_0|}{1 + |\mathcal{R}_\tau^-(\mathcal{I}) \cap \mathcal{H}_0|}\middle| \mathcal{F}_0(\mathcal{I})\right] \leq \frac{1}{1 - q_{\max}(\mathcal{I})}.$$

The FDR at τ is upper bounded:

$$\begin{aligned}
&\mathbb{E}\left[\frac{|\mathcal{R}_\tau^+(\mathcal{I}) \cap \mathcal{H}_0|}{|\mathcal{R}_\tau^+(\mathcal{I})| \vee 1}\middle| \mathcal{F}_0(\mathcal{I})\right] \\
=\ &\mathbb{E}\left[\frac{1 + |\mathcal{R}_\tau^-(\mathcal{I}) \cap \mathcal{H}_0|}{|\mathcal{R}_\tau^+(\mathcal{I})| \vee 1} \cdot \frac{|\mathcal{R}_\tau^+(\mathcal{I}) \cap \mathcal{H}_0|}{1 + |\mathcal{R}_\tau^-(\mathcal{I}) \cap \mathcal{H}_0|}\middle| \mathcal{F}_0(\mathcal{I})\right] \\
\leq\ &\alpha/2 \left(\frac{1}{1-\widehat{q_{\max}}(\mathcal{I})} - 1\right)^{-1} \mathbb{E}\left[\frac{|\mathcal{R}_\tau^+(\mathcal{I}) \cap \mathcal{H}_0|}{1 + |\mathcal{R}_\tau^-(\mathcal{I}) \cap \mathcal{H}_0|}\middle| \mathcal{F}_0(\mathcal{I})\right] \\
\leq\ &\alpha/2 \left(\frac{1}{1-\widehat{q_{\max}}(\mathcal{I})} - 1\right)^{-1} \left(\frac{1}{1-q_{\max}(\mathcal{I})} - 1\right).
\end{aligned}$$

By Taylor expansion, FDR is close to the target level when $q_{\max}(\mathcal{I}) - \widehat{q_{\max}}(\mathcal{I})$ is small:

$$\begin{aligned}
\mathbb{E}\left[\frac{|\mathcal{R}_\tau^+(\mathcal{I}) \cap \mathcal{H}_0|}{|\mathcal{R}_\tau^+(\mathcal{I})| \vee 1}\middle| \mathcal{F}_0(\mathcal{I})\right] &\leq \alpha/2 \left\{\left(\frac{1}{1-q_{\max}(\mathcal{I})} - 1\right)^{-1} + \epsilon_n^q(\mathcal{I}) \left(\frac{1}{q_{\max}(\mathcal{I})}\right)^2\right\} \left(\frac{1}{1-q_{\max}(\mathcal{I})} - 1\right) \\
&= \alpha/2 \left\{1 + \epsilon_n^q(\mathcal{I}) \frac{1}{q_{\max}(\mathcal{I})(1-q_{\max}(\mathcal{I}))}\right\}.
\end{aligned}$$

Same proof applies to the MaY-I^3 with residual $E_i^{-\mathcal{I}}$ and σ-field $\mathcal{F}_0^{-Y}(\mathcal{I})$; thus complete the proof. □

D.3.1 Proof of theorem 15

This section provides proof of the Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\pi^*}$ under heterogeneous propensity scores with known bounds.

Proof. The following arguments applies to both the Crossfit-I$^3_{\pi^*}$ and MaY-I$^3_{\pi^*}$, with their corresponding definition of the filtration and null hypothesis , as described in Appendix D.2.3.

The probability of probability of a positive sign of the estimated treatment effect q_i is

$$\begin{aligned}
q_i :=\ &\mathbb{P}\left((A_i - 1/2) \cdot E_i > 0 \mid \mathcal{G}_{-1}\right) \\
=\ &\pi_i \mathbb{P}(E_i > 0 \mid \mathcal{G}_{-1}) + (1-\pi_i)\mathbb{P}(E_i < 0 \mid \mathcal{G}_{-1}) \leq \max\{1-\pi_{\min}, \pi_{\max}\}.
\end{aligned}$$

We prove FDR control by Lemma 10, where $q_{\max}(\mathcal{I}) = \widehat{q}_{\max}(\mathcal{I}) = \max\{1-\pi_{\min}, \pi_{\max}\}$ and $\epsilon_n^q(\mathcal{I}) = 0$. □

D.3.2 Proof of theorem 16

Proof. By Lemma 10 where $q_{\max}(\mathcal{I}) = \max\{1-\pi_{\min}, \pi_{\max}\}$ and $\widehat{q}_{\max}(\mathcal{I}) = \max\{1-\widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}$, and $\epsilon_n^q(\mathcal{I}) = q_{\max}(\mathcal{I}) - \widehat{q}_{\max}(\mathcal{I})$, we have

$$\mathbb{E}\left[\mathrm{FDP}_{\hat{t}}^{\widehat{\pi}}(\mathcal{I})\right] \leq \alpha/2 \left\{1 + \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\epsilon_n^q(\mathcal{I}) \cdot \frac{1}{q_{\max}(\mathcal{I})(1-q_{\max}(\mathcal{I}))}\right]\right\}.$$

□

D.3.3 Proof of theorem 17

Proof. Denote the individual propensity score as $\mathbb{P}(A_i = 1 \mid X_i) = \pi_i$. For the nulls, the probability of probability of a positive sign of the estimated treatment effect as q_i:

$$\begin{aligned} q_i(\mathcal{I}) :=\ & \mathbb{P}\left((A_i - 1/2) \cdot (Y_i - \widehat{m}(X_i)) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})\right) \\ =\ & \pi_i \mathbb{P}(Y_i - \widehat{m}(x_i) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})) + (1-\pi_i)\mathbb{P}(Y_i - \widehat{m}(x_i) < 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})) \\ \leq\ & \min\left\{\max\{\pi_i, 1-\pi_i\}, \max\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1 - \Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\}\right\} \end{aligned}$$

where $\mathbb{P}(Y_i - \widehat{m}(x_i) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I}))$ can be separated from $\mathbb{P}(A_i - 1/2 > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I}))$ because they are independent for zero-effect nulls. Let the estimator of q_i be

$$\widehat{q_i} := \max\{\widehat{\pi}_i, 1 - \widehat{\pi}_i\}.$$

To describe the resulting estimation error of q_i, we define a difference $d_i(\mathcal{I}) := \max\{\pi_i, 1-\pi_i\} - \max\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1 - \Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\}$, which takes large value if the propensity score deviates from $1/2$ (smaller value if the outcome probability deviates from $1/2$). The true q_i is upper bounded by estimated $\widehat{q_i}$ plus some estimation error that depends on $d_i(\mathcal{I})$:

$$\begin{aligned} q_i - \widehat{q_i} \leq\ & \epsilon_i^{\pi}(\mathcal{I}) && \text{if } d_i(\mathcal{I}) \leq 0; \\ q_i - \widehat{q_i} \leq\ & \epsilon_i^{\pi}(\mathcal{I}) - d_i(\mathcal{I}) && \text{if } d_i(\mathcal{I}) > 0, \end{aligned}$$

where $\epsilon_i^{\pi}(\mathcal{I}) = \pi_i - \widehat{\pi}_i(\mathcal{I})$; it can be written in one line as

$$q_i - \widehat{q_i} \leq \epsilon_i^{\pi}(\mathcal{I}) - \max\left\{0, \max\{\pi_i, 1-\pi_i\} - \max\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1 - \Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\}\right\}.$$

Thus, the estimation error for $q_{\max}$ is upper bounded as

$$\begin{aligned} & \max_{i \in \mathcal{I}}\{\epsilon_i^{\pi}(\mathcal{I}) - \max\left\{0, \max\{\pi_i, 1-\pi_i\} - \max\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1 - \Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\}\right\}\} \\ \leq\ & \epsilon_n^{\pi}(\mathcal{I}) - \max\left\{0, \max\{\pi_{\max}, 1-\pi_{\min}\} - \max\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1 - \Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\}\right\} =: \epsilon_n^q(\mathcal{I}). \end{aligned} \tag{204}$$

(205)

By Lemma 10, we have

$$\mathbb{E}\left[\mathrm{FDP}_{\hat{t}}^{\widehat{\pi}}(\mathcal{I})\right] \leq \alpha/2 \left\{1 + \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\epsilon_n^q(\mathcal{I}) \left(\frac{1}{q_{\max}(\mathcal{I})(1-q_{\max}(\mathcal{I}))}\right)\right]\right\},$$

where we relax the bound using $q_{\max} \leq \min\{\max\{\pi_{\max}, 1-\pi_{\min}\}, \max\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1-\Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\}\}$.

□

Remark 12. *If we consider nulls as nonpositive effects, we have*

$$\begin{aligned} q_i(\mathcal{I}) :=\ & \mathbb{P}\left((A_i - 1/2)\cdot(Y_i - \widehat{m}(X_i)) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})\right) \\ =\ & \pi_i \mathbb{P}(Y_i^T - \widehat{m}(x_i) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})) + (1-\pi_i)\mathbb{P}(Y_i^C - \widehat{m}(x_i) < 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})) \\ \le\ & \pi_i \mathbb{P}(Y_i^C - \widehat{m}(x_i) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})) + (1-\pi_i)\mathbb{P}(Y_i^C - \widehat{m}(x_i) < 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})) \\ \le\ & \min\left\{\max\{\pi_i, 1-\pi_i\}, \max\{\Phi_{\max}\left[\epsilon_n^Y(\mathcal{I})\right], 1 - \Phi_{\min}\left[-\epsilon_n^Y(\mathcal{I})\right]\}\right\}, \end{aligned}$$

where $\Phi_{\max}(c) := \max_{i\in\mathcal{I}} \mathbb{P}(Y_i^C - \mathbb{E}(Y_i^C|X_i) \le c \mid X_i)$, $\Phi_{\min}(c) := \min_{i\in\mathcal{I}} \mathbb{P}(Y_i^C - \mathbb{E}(Y_i^C|X_i) \le c \mid X_i)$ *and* $\epsilon_n^Y(\mathcal{I}) = \max_{i\in\mathcal{I}} |\widehat{m}(X_i) - \mathbb{E}(Y_i^C|X_i)|$*, and then, we can make the same claim as above. However, it is harder to have robust FDR control when the propensity scores are poorly estimated, because* $\widehat{m}(x_i)$ *is an estimator for* $\mathbb{E}(Y_i \mid X_i)$*, which can be very different from the expected control outcome for a subject with negative effect. (We can design an algorithm where* $\widehat{m}(x_i)$ *is an estimation of the expected control outcome, but when the estimation is well, it would have zero power to detect positive effect if the subject is not treated.)*

D.4 Proof of power analysis

Our proof of the power analysis mainly uses the results in Arias-Castro and Chen [2017], who consider the setup with n hypotheses, each associated with a test statistic V_i for $i \in [n]$. Assume the test statistics are independent with the survival function $\mathbb{P}(V_i \ge x) = \Psi_i(x)$, which equals $\Psi(x - \mu_i)$ where $\mu_i = 0$ under the null and $\mu_i > 0$ otherwise. They focus on a class of distribution called asymptotically generalized Gaussian (AGG), whose survial function satisfies:

$$\lim_{x\to\infty} x^{-\gamma} \log \Psi(x) = -1/\gamma, \tag{206}$$

with a constant $\gamma > 0$. For example, a normal distribution is AGG with $\gamma = 2$. They discuss a class of multiple testing methods called *threshold procedure*: the final rejection set $\mathcal{R}$ is in the form

$$\mathcal{R} = \{i : V_i \ge \tau(V_1, \ldots, V_n)\}, \tag{207}$$

for some threshold $\tau(V_1, \ldots, V_n)$, and separately study two types of thresholds: the BH procedure [Benjamini and Hochberg, 1995] with threshold:

$$\tau_{\text{BH}} = V_{(\iota_{\text{BH}})}, \quad \iota_{\text{BH}} := \max\{i : V_{(i)} \ge \Psi^{-1}(i\alpha/n)\}, \tag{208}$$

where $V_{(1)} \ge \ldots \ge V_{(n)}$ are ordered statistics; and the Barber-Cand'es (BC) procedure [Barber and Candès, 2015] with a threshold on the absolute value of V_i:

$$\tau_{\text{BC}} = \inf\{\nu \in |\mathbf{V}| : \widehat{\text{FDP}}(\nu) \le \alpha\}, \tag{209}$$

where $|\mathbf{V}| := \{|V_i| : i \in [n]\}$ is the set of sample absolute values, and

$$\widehat{\text{FDP}}(\nu) := \frac{|\{i : V_i \le -\nu\}| + 1}{\max\{|\{i : V_i \ge \nu\}|, 1\}},$$

and the final rejection set is those with positive V_i and value larger than τ_{BC}. The stopping rule for the BC procedure is similar to our proposed algorithms, as detailed next.

Recall in Section 5.4 of the main paper, we consider a simplified automated version of the I^3 that exclude the subject with the smallest absolute value of the estimated treatment effect $|\widehat{\Delta}_i|$. Thus, the automated I^3 is a BC procedure where the test statistic of interest is $V_i = \widehat{\Delta}_i = 4(A_i - 1/2)(Y_i - \widehat{m}_n(X_i))$. Following the above notations and let Φ be the CDF for standard Gaussian, we denote the survival function for the nulls as

$$\Psi_n^{\text{null}}(x) = \frac{1}{2}(1 - \Phi(x + \widehat{m}_n)) + \frac{1}{2}(1 - \Phi(x - \widehat{m}_n)),$$

which is a mixture of two Gaussians, with $\widehat{m}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, and $\widehat{m}_n \overset{a.s.}{\to} 0$ by the strong law of large numbers. For the non-nulls, the survival function is

$$\Psi_n^{\text{non-null}}(x) = \frac{1}{2}(1 - \Phi(x + \widehat{m}_n - \mu)) + \frac{1}{2}(1 - \Phi(x - \widehat{m}_n)).$$

Note that our setting is slightly different from the discussion in Arias-Castro and Chen [2017] because the non-nulls differ from the nulls by a shift on one of the Gaussian component (rather than a shift in the overall survival function Ψ_n^{null}). Similar to the characterization by AGG in (206), both survival functions Ψ_n^{null} and $\Psi_n^{\text{non-null}}$ asymptotically satisfy a tail property that for any $x_n \to \infty$ as $n \to \infty$:

$$\lim_{n\to\infty} x_n^{-\gamma} \log \Psi_n(x_n) = -1/\gamma, \tag{210}$$

with probability one and $\gamma = 2$, which we later refer to as asymptotic AGG. Conclusions in our paper basically follows the proofs in Arias-Castro and Chen [2017] with the test statistics V_i specified as the estimated treatment effect $\widehat{\Delta}_i$.

D.4.1 Proof of theorem 12

We first present the proof for the power of the automated Crossfit-I^3, and the power of the linear-BH is proven similarly as shown later.

Proof. **Zero power when $r < \beta$.** The argument of zero power indeed applies to any threshold procedure as defined in (207): $\mathcal{R} = \{i : \widehat{\Delta}_i \geq d\}$, for some $d \in \mathbb{R}$. Following the proof of Theorem 1 in Arias-Castro and Chen [2017], we argue that the FDR control cannot be satisfied for any $\alpha \in (0, 1)$ unless the threshold d is large enough such that $d > \mu + \delta_n$ with $\delta_n = \log\log n$; but in this case, most non-nulls cannot be included in the rejection set, and thus the power goes to zero.

First, we claim that when $d \leq \mu + \delta_n$, the false discovery proportion (FDP) goes to one in probability. By the proof of Theorem 1 in Arias-Castro and Chen [2017], we have that FDP goes to one in probability if $\frac{(n-n_1)\Psi_n^{\text{null}}(\mu+\delta_n)}{n_1} \to \infty$ with probability one, where n_1 is the number of non-nulls. Their proof also verifies that $\frac{(n-n_1)\Psi_n^{\text{null}}(\mu+\delta_n)}{n_1} \to \infty$ because Ψ_n^{null} satisfies property (210) with probability one.

Next, we show that when $d > \mu + \delta_n$, the power goes to zero. Notice that power can be equivalently defined as $\mathbb{E}(1 - \text{FNR})$, where FNR (false negative rate) is defined as the proportion of non-nulls not identified. Again by the proof of Theorem 1 in Arias-Castro and Chen [2017], we have that the FNR converge to one in probability if $\Psi_n^{\text{non-null}}(\mu + \delta_n)$ goes to zero with probability one, which is true because $\delta_n \to \infty$.

Combining the above two arguments, we conclude that for any threshold procedure whose rejection set is in the form of (207), the power goes to zero for any FDR control $\alpha \in (0, 1)$ when $r < \beta$.

Note that the above proof assumes that the test statistics $V_i = \widehat{\Delta}_i$ are mutually independent. For simplicity, we design the Crossfit-I^3 where $\widehat{m}_n$ for the I^3 implemented on $\mathcal{I}$ is computed using data in

$\mathcal{II}$, to ensure the above mutual independence. Thus, the above proof applies to the I^3 implemented on each half, $\mathcal{I}$ and $\mathcal{II}$. The overall power behaves the same asymptotically, since $\mathcal{I}$ and $\mathcal{II}$ result from a random split of all subjects $[n]$. For all cases hereafter, we prove the power claim for the I^3 implemented on $\mathcal{I}$ conditional on data in $\mathcal{II}$, and the same claim holds for the overall power as reasoned above.

Half power when $r > \beta$. We first prove the limit inferior of the power is at least $1/2$, and then the limit superior is at most $1/2$, mainly using the proof of Theorem 3 in Arias-Castro and Chen [2017].

They consider a sequence of thresholds $d_n^* = (\gamma r^* \log n)^{1/\gamma}$ for some $r^* \in (\beta, r \wedge 1)$. We first claim that the FDR estimator at d_n^* is less than any $\alpha \in (0,1)$ for large n, or mathematically $\widehat{\text{FDR}}(d_n^*) \le \alpha$. It can be verified by the proof of Theorem 3 in Arias-Castro and Chen [2017] where the survival function of $\widehat{\Delta}_i$ is $G(d_n^*) = (1-\epsilon)\Psi_n^{\text{null}}(d_n^*) + \epsilon\Psi_n^{\text{non-null}}(d_n^*)$ with $\epsilon = n^{-\beta}$, and the fact that $\Psi_n^{\text{non-null}}(d_n^*) \to 1/2$ and $\Psi_n^{\text{null}}(d_n^*) \to n^{-r^*}$ (by property (210)) with probability one. It follows that the true stopping threshold τ_n satisfies $\tau_n \le d_n^*$. Also, by Lemma 1 in Arias-Castro and Chen [2017], we have that the proportion of correctly identified non-nulls at threshold d_n^* is $\frac{1}{n_1}\sum_{i \notin \mathcal{H}_0} \mathbb{1}\{\widehat{\Delta}_i \ge d\} = \Psi_n^{\text{non-null}}(d_n^*) + o_{\mathbb{P}}(1)$, where $\Psi_n^{\text{non-null}}(d)$ decreases in d and converges to $1/2$ when $d = d_n^*$. Recall that the true stopping threshold is no larger than d_n^*, so the limit inferior of the power is at least $1/2$.

The power converges to $1/2$ once we show that the limit superior of the power is at most $1/2$. Consider a positive constant $d^0 \in (0, \infty)$, and we claim that the actual stopping threshold $\tau_n \ge d^0$ for large n because the FDR estimator goes to one, following similar arguments in the proof of Theorem 3 in Arias-Castro and Chen [2017]. Specifically,

$$\widehat{\text{FDR}}(d^0) \equiv \frac{|\{i \in [n] : \widehat{\Delta}_i \le -d^0\}| + 1}{\max\{|\{i \in [n] : \widehat{\Delta}_i \ge d^0\}|, 1\}} = \frac{1 + n(1 - \widehat{G}_n(-d^0))}{\max\{n\widehat{G}_n(d^0), 1\}},$$

where $\widehat{G}_n(d^0) = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}(\widehat{\Delta}_i \ge d^0)$ denotes the empirical survival function. Use the fact that $G_n(d^0) = (1-\epsilon)\Psi_n^{\text{null}}(d^0) + \epsilon\Psi_n^{\text{non-null}}(d^0) \to 1 - \Phi(d^0)$ and $G_n(-d^0) \to \Phi(d^0)$ almost surely, we observe that $\widehat{\text{FDR}}(d^0) \to 1$ with probability one. Also, the proportion of correctly identified non-nulls at threshold d^0 is $\Psi_n^{\text{non-null}}(d^0) + o_{\mathbb{P}}(1)$ (recall in the previous paragraph), where $\Psi_n^{\text{non-null}}(d^0) \to 1/2 + 1/2(1 - \Phi(d^0))$. Thus, the power for large n is smaller than $1/2 + 1/2(1 - \Phi(d^0))$ for all $d^* \in (0, \infty)$; in other words, the limit superior of the power is smaller than $\inf_{d \in (0,\infty)} 1/2 + 1/2(1 - \Phi(d^*)) = 1/2$.

With the limit inferior and superior of the power bounded by $1/2$, we conclude that the power converges to $1/2$. In fact, the above proof implies that the power of identifying non-null subjects that are treated is one (notice that $\Psi_n^{\text{non-null, treated}}(d_n^*) = 1 - \Phi(d_n^* + \widehat{m}_n - \mu) \to 1$ with probability one, so the limit inferior of the power for treated non-nulls is at least 1). □

Proof for the linear-BH procedure The power of the linear-BH procedure when there is no covariates can be proved following similar steps as above, and using intermediate results of Theorem 2 in Arias-Castro and Chen [2017]. In their notation, the linear-BH procedure uses $V_i = \widehat{\Delta}_i^{\text{BH}}$ as the test statistics, and we separately discuss power among the treated group and the control group in ensure the independence among V_i. The survival functions for the nulls and non-nulls in the treated group are

$$\Psi_n^{\text{null}}(x) = 1 - \Phi\left(\frac{x}{\sqrt{1 + 1/n^C}}\right) \text{ and } \Psi_n^{\text{non-null}}(x) = 1 - \Phi\left(\frac{x - \mu}{\sqrt{1 + 1/n^C}}\right),$$

where n^C is the number of untreated subjects. For the control group, the survival functions are

$$\Psi_n^{\text{null}}(x) = 1 - \Phi\left(\frac{x + \widehat{Y}^T}{\sqrt{1 + 1/n^T}}\right) \text{ and } \Psi_n^{\text{non-null}}(x) = 1 - \Phi\left(\frac{x + \widehat{Y}^T}{\sqrt{1 + 1/n^T}}\right),$$

where $\widehat{Y}^T = \sum_{A_i=1} Y_i \xrightarrow{a.s.} 0$, and n^T is the number of treated subjects. Since the above distributions converge to a Gaussian, these survival functions satisfy the AGG property asymptotically as defined in (210).

Proof. First, we claim that the power goes to zero when $r < \beta$, following the proof in Section D.4.1 for any threshold procedure (separately for the treated group conditional on control group). Then, we prove that power converges to $1/2$ when $r > \beta$: the power among untreated subjects is asymptotically zero because the survival functions for the nulls and non-nulls are the same; the power among treated subjects is asymptotically one following the proof of Theorem 2 in Arias-Castro and Chen [2017] as detailed next.

Again consider the sequence of thresholds $d_n^* = (\gamma r^* \log n)^{1/\gamma}$ for some $r^* \in (\beta, r \wedge 1)$. We first claim that the FDR estimator at d_n^* is less than any $\alpha \in (0,1)$ for large n, or mathematically $\widehat{\text{FDR}}(d_n^*) \leq \alpha$. It can be verified by the proof of Theorem 2 in Arias-Castro and Chen [2017] where $G(d_n^*) = (1-\epsilon)\Psi_n^{\text{null}}(d_n^*) + \epsilon\Psi_n^{\text{non-null}}(d_n^*)$ with $\epsilon = n^{-\beta}$, and the fact that $\Psi_n^{\text{non-null}}(d_n^*) \to 1$ for the treated group and $\Psi_n^{\text{null}}(d_n^*) \to n^{-r^*}$ (by property (210)) with probability one. It follows that the true stopping threshold τ_n satisfies $\tau_n \leq d_n^*$. Also, by Lemma 1 in Arias-Castro and Chen [2017], we have that the proportion of correctly identified non-nulls at threshold d_n^* is $\frac{1}{n_1}\sum_{i \notin \mathcal{H}_0} \mathbb{1}\{\widehat{\Delta}_i^{\text{BH}} \geq d\} = \Psi_n^{\text{non-null}}(d_n^*) + o_{\mathbb{P}}(1)$, where $\Psi_n^{\text{non-null}}(d)$ for the treated group decreases in d and converges to 1 when $d = d_n^*$. Recall that the true stopping threshold is no larger than d_n^*, so the limit inferior of the power among the treated subjects is at least 1. Therefore, the overall power converges to $1/2$.

□

D.4.2 Proof of theorem 13

We first consider the power when all subjects are non-nulls ($\beta = 0$).

Lemma 11. *Given any fixed FDR control level $\alpha \in (0,1)$ and let the number of subjects n goes to infinity. When all subjects are non-nulls, the stopping time $\tau = 0$ with probability tending to one if $\mu > \Phi^{-1}(\frac{1}{1+\alpha})$, and in this case the power converges to $\Phi(\mu)$.*
E.g., when $\alpha = 0.2$, the asymptotic power of the automated I^3 is larger than 0.8 if $\mu \geq 1$.

Proof. The stopping time $\tau = 0$ if and only if the FDR control is satisfied when all the subjects are included: $\widehat{\text{FDR}}_n(\mathcal{R}_0) = \frac{|\mathbb{R}_0^-|+1}{\max\{|\mathbb{R}_0^+|,1\}} \leq \alpha$, or equivalently, $\frac{|\mathbb{R}_0^+|}{n} \geq \frac{1+\frac{1}{n}}{1+\alpha}$. Notice that the proportion of positive $\widehat{\Delta}_i$ converges to a constant: $\frac{|\mathbb{R}_0^+|}{n} \xrightarrow{a.s.} \Phi(\mu)$, because $\widehat{\Delta}_i$ of each non-null follows a Gaussian distribution with mean μ and variance less than 2. Thus, if $\Phi(\mu) > \frac{1}{1+\alpha}$, for any $\epsilon \in (0,1)$, there exists N such that for all $n \geq N$, we have that (a) $\left|\frac{|\mathbb{R}_0^+|}{n} - \Phi(\mu)\right| < \epsilon$ with probability at least $1-\epsilon$; and (b) $\widehat{\text{FDR}}_n(\mathcal{R}_0) = \frac{|\mathbb{R}_0^-|+1}{\max\{|\mathbb{R}_0^+|,1\}} \leq \alpha$ (hence $\tau = 0$) with probability at least $1-\epsilon$. (Notice that the threshold N can be chosen as not depending on μ, which is useful in the next proof.) In such a case, the power is no less than $(1-\epsilon)(\Phi(\mu) - \epsilon)$ when $n \geq N$; and the power is no larger than $\Phi(\mu) - \epsilon$; so the power converges to $\Phi(\mu)$. The proof completes once notice that the condition $\Phi(\mu) > \frac{1}{1+\alpha}$ is equivalent to $\mu > \Phi^{-1}(\frac{1}{1+\alpha})$. □

Proof of Theorem 13. **Power of the Crossfit-I^3.** Recall that the I^3 implemented on $\mathcal{I}$ exclude subjects based on the averaged estimated effect on $\mathcal{II}$: $\text{Pred}(x) = \overline{\widehat{\Delta}_i(X_i = x)}$, which converges to μ almost surely when $x = 1$ (the non-nulls), and 0 almost surely when $x = 0$ (the nulls). Thus, no non-nulls in $\mathcal{I}$ would be excluded before excluding all the nulls in $\mathcal{I}$ (with probability going to one) for any fixed $\mu > 0$.

Combined with Lemma 11, we have that if $\mu > \Phi^{-1}(\frac{1}{1+\alpha})$, for any $\epsilon \in (0,1)$, there exists $N(\epsilon,\alpha)$ such that for all $n \geq N$, the power of the I^3 is higher than $\Phi(\mu) - \epsilon$. Also, the limit of power increases to one for any $r > 0$ (where the signal μ increases): there exists $N'(\epsilon)$ such that for all $n \geq N'(\epsilon)$, $\Phi(\mu) \geq 1 - \epsilon$. Therefore, for any $\epsilon \in (0, \frac{1}{1+\alpha})$, we have that for all $n \geq \max\{N'(\epsilon), N(\epsilon,\alpha)\}$, the power of I^3 implemented on $\mathcal{I}$ is no less than $1 - 2\epsilon$; thus completes the proof.

Power of the linear-BH procedure. As before, we separately argue that the power for the treated group and the control group converges to zero when $r < \beta$, and converges to one when $r > \beta$. For a subject in the treated group with $X_i = x$ where $x \in \{0,1\}$, the estimated effect is a Gaussian $\widehat{\Delta}_i^{\text{BH}} \sim N(0, 1 + \frac{1}{\sum_i \mathbb{1}(X_i=x, A_i=0)})$ for the nulls and $\widehat{\Delta}_i^{\text{BH}} \sim N(\mu, 1 + \frac{1}{\sum_i \mathbb{1}(X_i=x, A_i=0)})$ for the non-nulls. For a subject in the control group with $X_i = x$ where $x \in \{0,1\}$, the estimated effect is a Gaussian $\widehat{\Delta}_i^{\text{BH}} \sim N(0, 1 + \frac{1}{\sum_i \mathbb{1}(X_i=x, A_i=1)})$ for the nulls and $\widehat{\Delta}_i^{\text{BH}} \sim N(\mu, 1 + \frac{1}{\sum_i \mathbb{1}(X_i=x, A_i=1)})$ for the non-nulls.

The power of the linear-BH procedure directly results from Theorem 2 in Arias-Castro and Chen [2017] because in both the treated and control group, (a) the linear-BH procedure is the BH procedure where the random variable of interest is $V_i = \widehat{\Delta}_i^{\text{BH}}$; and (b) $\widehat{\Delta}_i^{\text{BH}}$ of non-nulls and nulls differ by a shift μ; and (c) the survival function of $\widehat{\Delta}_i^{\text{BH}}$ is asymptotically AGG (recall definition in (210)) since it converges to a Gaussian distribution). □

D.5 Additional numerical experiments

D.5.1 Identification of individual positive effect

We have seen the numerical results of the proposed methods in the main paper where the treatment effect is defined in (110) with sparse and strong positive effect, and dense and weak negative effect. This section presents three more examples of the treatment effect.

Linear effect. Let the treatment effect be

$$\Delta(X_i) = S_\Delta \cdot [2X_i(1)X_i(2) + 2X_i(3)], \tag{211}$$

where $S_\Delta > 0$. In this case, all subjects have treatment effects, and the scale correlates with the covariates (with interaction terms) linearly. Thus, the linear-BH procedure has valid error control as shown in Figure 60a (unlike other cases with nonlinear treatment effect).

Sparse and strong effect that is positive. Let the treatment effect be

$$\Delta(X_i) = S_\Delta \cdot [5X_i^3(3)\mathbb{1}\{X_i(3) > 1\}], \tag{212}$$

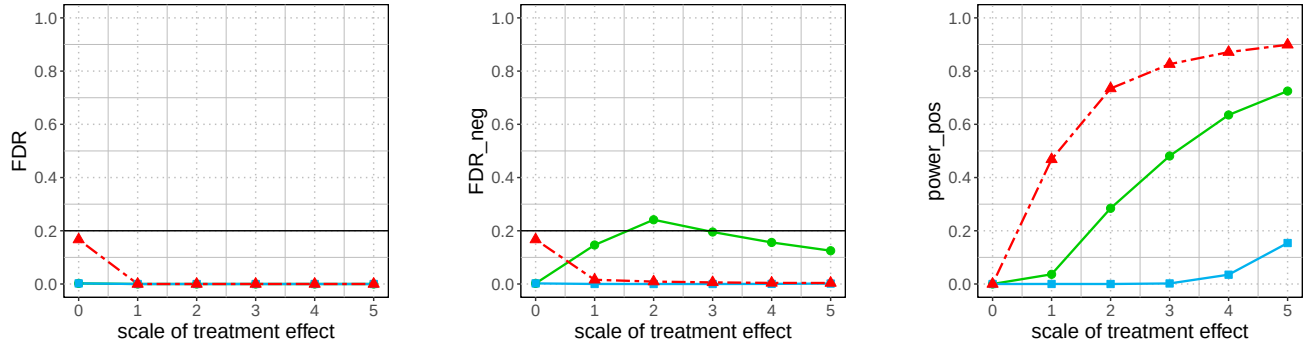
where $S_\Delta > 0$. Here, the subjects with $X_i(3) > 1$ have positive treatment effects. Although linear-BH procedure seems to have high power, its FDR is largely inflated since the assumption of linear correlation does not hold (see Figure 60b). In contrast, our proposed methods have valid FDR control.

Sparse and strong effect in both directions. Let the treatment effect be

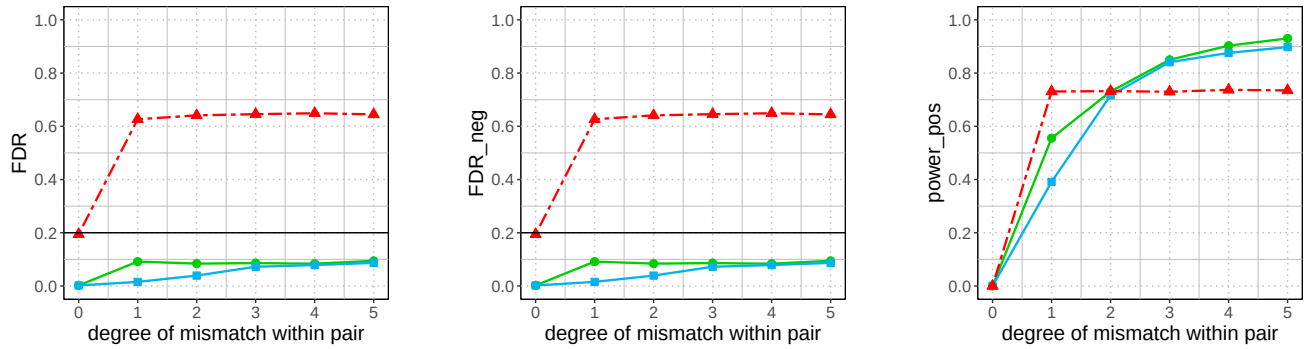
$$\Delta(X_i) = S_\Delta \cdot [5X_i^3(3)\mathbb{1}\{|X_i(3)| > 1\}], \tag{213}$$

where $S_\Delta > 0$. Here, the subjects with $X_i(3) > 1$ have positive treatment effects and those with $X_i(3) < -1$ have negative treatment effects; the scale and proportion of effects in both directions are

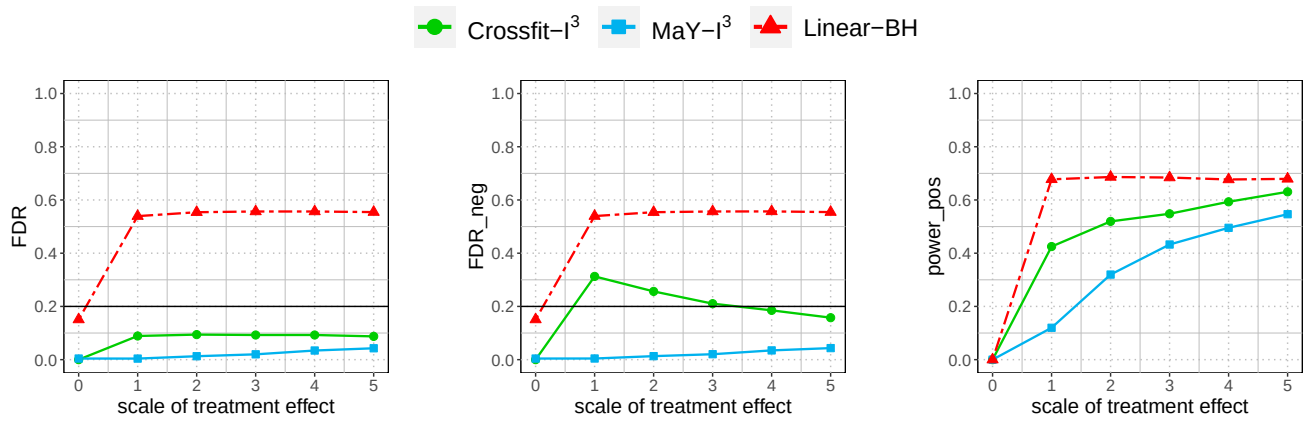
the same. The power comparison is similar to the previous setting with only positive effect, except the power for the methods with valid FDR control are lower since there is addtionally negative effect in this example (see Figure 60c).



(a) Dense two-sided effect (linear) as in model (211).



(b) Sparse and strong effect that is positive (nonlinear) in model (212).



(c) Sparse and strong effect in both directions (nonlinear) in model (213).

Figure 60: FDR for the zero-effect null (96) in the main paper (the first column), and FDR for the nonpositive-effect null (111) in the main paper (the second column), and power (the third column) of three methods: linear-BH procedure, Crossfit-I^3, MaY-I^3, under three types of treatment effect when varying the scale of treatment effect S_Δ in $\{0, 1, 2, 3, 4, 5\}$. When the linear assumption holds as in the first row, the linear-BH procedure has valid FDR control and high power, but its FDR is large when the treatment is a nonlinear function of the covariates as in the latter two rows. In contrast, the Crossfit-I^3 and MaY-I^3 have valid FDR control for their target null hypotheses, respectively.

D.5.2 Paired samples

We have presented in Section 5.8 of the main paper that the interactive algorithms can be applied to paired samples, and have discussed their power in two cases where the samples within each pair either matched exactly or mismatch to some degree. A side observation is that the algorithms not using the pairing information seem to have a small change in power when the degree of mismatch varies. Here, we increase the degree of mismatch to show a more clear pattern of this change.

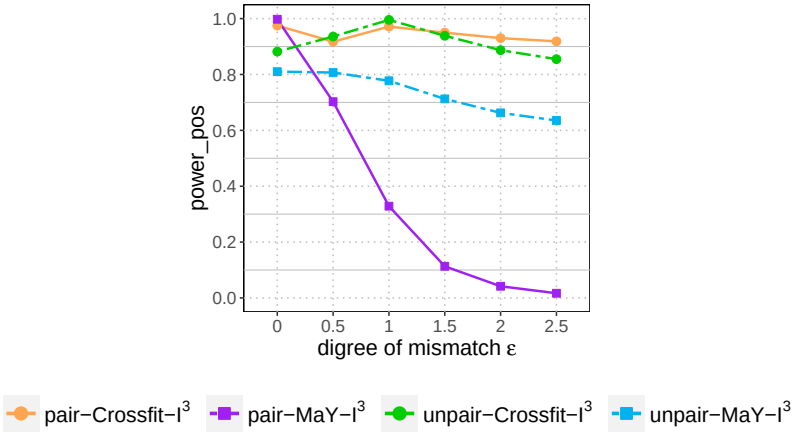


Figure 61: Power of identifying subjects with positive effects of the proposed algorithms (Crossfit-I^3 and MaY-I^3) with or without pairing information, when the scale of treatment effect is fixed at 2 and the degree of mismatch ϵ varies. The power of algorithms without pairing information first increase and then decrease as ϵ becomes larger.

We extend the definition of mismatch for $\epsilon \in (0, 1)$ to a larger ϵ: $\mathbb{P}(X_{i1}(1) \neq X_{i2}(1)) = \min\{\epsilon, 1\}$ and $\mathbb{P}(X_{i1}(2) \neq X_{i2}(2)) = \min\{\epsilon, 1\}$ and $X_{i1}(3) = X_{i2}(3) + U(0, 2\epsilon)$, where $U(0, 2\epsilon)$ is uniformly distributed between 0 and 2ϵ, and a larger ϵ leads to a larger degree of mismatch. As ϵ increases, the power under the unpaired samples first increases (see Figure 61). It is because the treatment effect is positive when $X_i(3) > 1$, which only takes 15% proportion if without mismatching; thus, the pattern of treatment effect is not easy to learn. In contrast, when there is a positive shift on $X_i(3)$ as designed in the mismatching setting above, more subjects have positive effects so that the algorithm can more easily learn the effect pattern and hence increase the power. The power can slightly decrease when the degree of mismatch is too large ($\epsilon > 1$), because there are fewer subjects without treatment effect, also affecting the estimation of treatment effect.

D.6 An alternative FDR estimator

Let π_i be the known propensity score for each subject $i \in [n]$. We introduce $q_i = \max\{\pi_i, 1 - \pi_i\}$, to measure the "worst-case bias". The FDR estimator we propose in the main paper can be written as

$$\widehat{\mathrm{FDR}}^{\mathrm{AdaPT}}(\mathcal{R}_t) \equiv \frac{1}{1 - \max_i q_i} \cdot \frac{(\max_i q_i) \cdot \left(1 + \sum_{i \in \mathcal{R}_t} \mathbb{1}\{\widehat{\Delta}_i \leq 0\}\right)}{\max\left(1, \sum_{i \in \mathcal{R}_t} \mathbb{1}\{\widehat{\Delta}_i > 0\}\right)}, \tag{214}$$

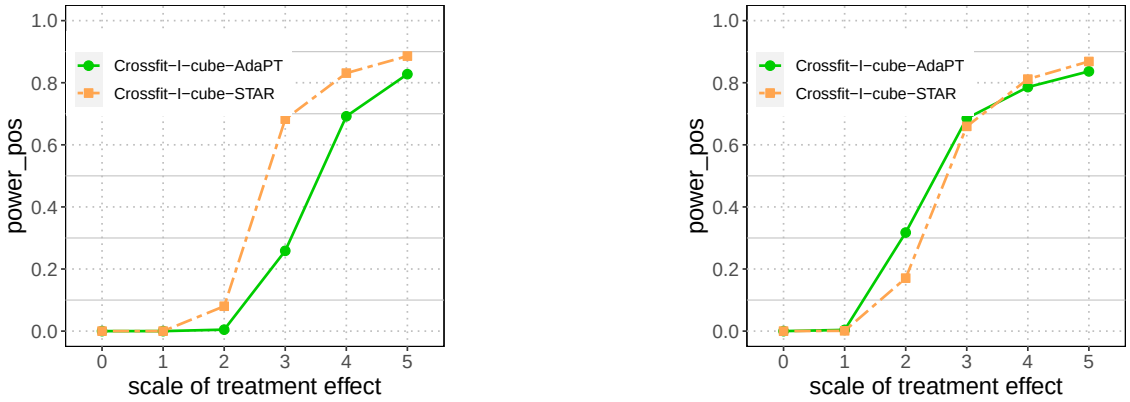
which counts each nonpositive $\widehat{\Delta}_i$ by the same weight, and adjusts for the heterogeneous propensity score by multiplying a common factor $\max_i q_i$ (considering the worst case).

Another FDR estimation [Lei et al., 2020] puts heterogeneous weights on the nonpositive $\widehat{\Delta}_i$, depending on each subject's propensity score: we are less punished for a wrongly included subject if its propensity score is close to $1/2$. Specifically, the FDR estimator is defined as

$$\widehat{\mathrm{FDR}}^{\mathrm{STAR}}(\mathcal{R}_t) := \frac{1}{1-\max_i q_i} \cdot \frac{1+\sum_{i\in\mathcal{R}_t} \frac{1-\max_i q_i}{1-q_i} \mathbb{1}\{\widehat{\Delta}_i \le 0\}}{1+|\mathcal{R}_t|}. \tag{215}$$

For example, when $\max_i q_i = 0.9$, a nonpositive sign with $q_i = 0.5$ leads to an increment of 0.2 in the denominator; for comparison, the increment is 0.9 if using $\widehat{\mathrm{FDR}}^{\mathrm{AdaPT}}(\mathcal{R}_t)$.

However, we do not claim that $\widehat{\mathrm{FDR}}^{\mathrm{STAR}}(\mathcal{R}_t)$ is uniformly better than $\widehat{\mathrm{FDR}}^{\mathrm{AdaPT}}(\mathcal{R}_t)$, especially when the propensity scores do not vary much among the investigated subjects. For example, when all scores have the same deviation from $1/2$ such that $\max_i q_i = q_i = q$ for all $i \in [n]$, the ratio of $\widehat{\mathrm{FDR}}^{\mathrm{AdaPT}}(\mathcal{R}_t)$ over $\widehat{\mathrm{FDR}}^{\mathrm{STAR}}(\mathcal{R}_t)$ is approximately $q\frac{|\mathcal{R}_t|}{|\mathcal{R}_t^+|}$, which indicates smaller FDR estimation using the AadPT method when q is not too extreme and the signal is strong such that $\frac{|\mathcal{R}_t|}{|\mathcal{R}_t^+|}$ is close to one (note their corresponding rejection sets are also different).



(a) $\pi_i = 0.5$ for all subjects except one being as large as $\pi_i = 0.9$.

(b) $\pi_i = 0.6$ for all subjects.

Figure 62: Power of the Crossfit-I^3 with two FDR estimators in (214) and (215), when under a simple treatment effect $\Delta(X_i) = S_\Delta[2X_i(1) - 1]$ with the scale S_Δ varying in $\{1, 2, 3, 4, 5\}$. The "STAR" estimation (214) leads to higher power when the propensity scores have a few outliers, and the "AdaPT" version seems to be better when there is not much heterogeneity in propensity scores.

D.7 Effect estimator using median

Estimating the median instead of mean improves the FDR control when the propensity scores are poorly estimated (FDR bound is in theorem 17 with "centered" CDF Φ defined as $\Phi(c) := \mathbb{P}(Y_i - \mathrm{median}(Y_i \mid X_i) \le c \mid X_i)$, where $\Phi(\epsilon)$ is close to 1/2 for any continuous distribution when ϵ is small, whereas the algorithm in the main paper has good performance for symmetric distributions). However, it can lead to lower power in well-specified case, where the outcomes have a symmetric distribution (mean is the same as the median) and the experiment is randomized.

The I^3 using median estimator leads to higher power when the noise is right-skewed (absolute-Cauchy) and the effect is dense. Recall that we select subjects with positive estimated effect sign:

$$\mathrm{sign}[(A_i - 1/2) \cdot (Y_i - \widehat{m}(X_i))],$$

where $\widehat{m}$ can be estimator of $\mathbb{E}(Y_i \mid X_i)$ or $\text{median}(Y_i \mid X_i)$. Ideally when there is no random noise, we have

$$Y_i(X_i, A_i = 0) < m(X_i) < Y_i(X_i, A_i = 1),$$

for the subjects with positive effect, leading to a positive estimated effect sign, consistent with the underlying truth.

When the outcomes have noise, the correctness of the estimated effect sign depends on two factors: the outcome estimation $\widehat{m}(X_i)$ and the noise in Y_i. Even if $\widehat{m}(X_i)$ is correctly learned such that

$$\mathbb{E}[Y_i(X_i, A_i = 0)] < \widehat{m}(X_i) < \mathbb{E}[Y_i(X_i, A_i = 1)],$$

when $\widehat{m}$ estimates the conditional expectation, or

$$\text{median}[Y_i(X_i, A_i = 0)] < \widehat{m}(X_i) < \text{median}[Y_i(X_i, A_i = 1)],$$

when $\widehat{m}$ estimates the conditional median, the estimated sign can be the opposite of the truth when the outcome have large variance.

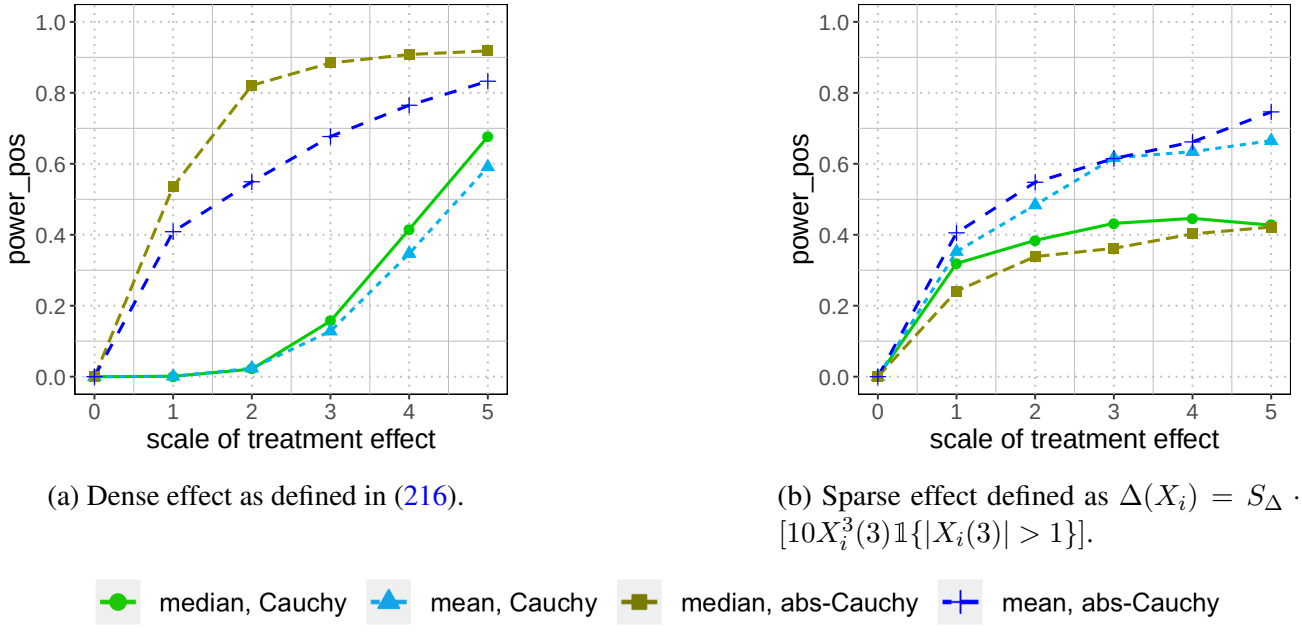
To see how these two factor influences the power of I^3, we show that the methods using median has comparable power when the treatment effect is dense:

$$\Delta(X_i) = 2S_\Delta, \tag{216}$$

and the control outcome is simple

$$f(X_i) = 0.$$

we vary the random noise as Cauchy (heavy-tail on both side) or the absolute value of Cauchy (right-skewed). We expect that in terms of the factor of estimation $\widehat{m}(X_i)$, the median estimator is better; in terms of the factor of noise in Y_i: the Cauchy noise would cause larger damage than the absolute Cauchy noise.



(a) Dense effect as defined in (216).

(b) Sparse effect defined as $\Delta(X_i) = S_\Delta \cdot [10X_i^3(3)\mathbb{1}\{|X_i(3)| > 1\}]$.

Figure 63: Power of the MaY-I^3 when using a median estimator and mean estimator under Cauchy noise or absolute-Cauchy noise in an randomized experiment with dense or sparse effect.

First consider a simple case where all subjects have a constant effect. When the noise is Cauchy, the factor of larger outcome variance plays a larger role than the factor of $\widehat{m}$ estimation, so the power is low either using the median estimator or the mean estimator (see blue and green line in Figure 63a). Nonetheless, when the noise is absolute-Cauchy, the robustness of the median estimator shows its advantage, so the method using median estimator has higher power than using the mean estimator (see olive and purple line in Figure 63a).

The difference in power caused by Cauchy noise and absolute-Cauchy noise vanishes when the effect is sparse (see Figure 63b), because the effect sizes are set to be large for both methods to have nontrivial power, making the noise less influential. In such a case, the I^3 using the mean estimator has higher power. The reason is from the factor of $\widehat{m}$ estimation: for subjects with true treatment effect, the median tends to underestimate the conditional outcome level $(Y_i \mid X_i)$, especially when the number of subjects with positive effect is small and the estimator is downsized because of the outcomes from subjects with nonpositive treatment effect.

To summarize, the I^3 using median estimator is beneficial when the effect is dense and the noise is one-sided heavy-tail; however, it leads to lower power when the effect is sparse (and the effect size is large).

D.8 Alternative notions of robustness in FDR control

We propose three alternative ways of defining FDR estimator by introducing two estimators for the maximum probability of positive estimated effect $q_{\max}(\mathcal{I})$, which leads to three "degrees" of robustness (the MaY-$\mathrm{I}^3_{\widehat{\pi}}$ proposed in the main paper falls into one of the robustness degree). Recall for subject $i \in \mathcal{I} \cap \mathcal{H}_0$,

$$\begin{aligned} q_i :=&\ \mathbb{P}\left((A_i - 1/2) \cdot (Y_i - \widehat{m}(X_i)) > 0 \mid \mathcal{F}_0^{-Y}(\mathcal{I})\right) \\ \le&\ \min\left\{\max\{1-\pi_{\min}, \pi_{\max}\}, \max\{1-p_{\min}, p_{\max}\}\right\} =: q_{\max}(\mathcal{I}), \end{aligned}$$

where $\pi_{\min}(\mathcal{I}), \pi_{\max}(\mathcal{I})$ are bounds on the true propensity scores and $p_{\min}(\mathcal{I}), p_{\max}(\mathcal{I})$ are bounds on the outcome probability

$$p_i = \mathbb{P}(Y_i > \widehat{m}(X_i) \mid \mathcal{F}_0^{-Y}(\mathcal{I})) \tag{217}$$

for $i \in \mathcal{I} \cap \mathcal{H}_0$. For each $\widehat{q_{\max}}(\mathcal{I})$ we propose later, the FDR estimator is defined as

$$\left(\frac{1}{1-\widehat{q_{\max}}(\mathcal{I})} - 1\right) \frac{|\mathcal{R}^-_{\widehat{t}}| + 1}{|\mathcal{R}^+_{\widehat{t}}| \vee 1},$$

and by Lemma 10, the resulting FDR of MaY-I^3 is upper bounded as

$$\mathrm{FDR} \le \alpha \left\{1 + \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\epsilon^q_n(\mathcal{I})\left(\frac{1}{q_{\max}(\mathcal{I})(1-q_{\max}(\mathcal{I}))}\right)\right] + \mathbb{E}_{\mathcal{F}_0(\mathcal{I}\mathcal{I})}\left[\epsilon^q_n(\mathcal{I}\mathcal{I})\left(\frac{1}{q_{\max}(\mathcal{I}\mathcal{I})(1-q_{\max}(\mathcal{I}\mathcal{I}))}\right)\right]\right\},$$

where $\epsilon^q_n(\mathcal{I}) = \max\{q_{\max}(\mathcal{I}) - \widehat{q_{\max}}(\mathcal{I}), 0\}$ is (one-sided) estimation error of $q_{\max}(\mathcal{I})$. Intuitively, a less conservative estimator $\widehat{q_{\max}}(\mathcal{I})$ (one that is close to $1/2$) lead to a less conservative FDR estimator so that the identification power could be higher; however, the estimation error $\epsilon^q_n(\mathcal{I})$ tend to be larger, making the FDR upper bound looser (less robust).

As a preparation, we estimate two quantities involved in q_i: the propensity score π_i and the outcome probability $p_i(\mathcal{I}) := \mathbb{P}(Y_i > \widehat{m}(X_i) \mid \mathcal{F}_0^{-Y}(\mathcal{I}))$, where both estimations use $D(\mathcal{I}\mathcal{I})$, denoted as $\widehat{\pi}_i(\mathcal{I})$ and $\widehat{p}_i(\mathcal{I})$. And let the minimum and maximum be $\widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})$ and $\widehat{p_{\min}}(\mathcal{I}), \widehat{p_{\max}}(\mathcal{I})$. The estimation error is denoted as $\epsilon^\pi_n(\mathcal{I}) = \max\{1-\pi_{\min}, \pi_{\max}\} - \max\{1-\widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}$ and $\epsilon^p_n(\mathcal{I}) = \max\{1-p_{\min}, p_{\max}\} - \max\{1-\widehat{p_{\min}}(\mathcal{I}), \widehat{p_{\max}}(\mathcal{I})\}$, both of which are measurable random variables with respect to $\mathcal{F}_0(\mathcal{I})$.

"0.5" robustness. Define

$$\widehat{q_{\max}}(\mathcal{I}) := \min\{\max\{1-\widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}, \max\{1-\widehat{p_{\min}}(\mathcal{I}), \widehat{p_{\max}}(\mathcal{I})\}\},$$

which is less conservative as the deviation of $\widehat{q_{\max}}(\mathcal{I})$ from $1/2$ (desired value) is the minimum of the deviation for estimated propensity score and $\mathbb{P}(Y_i > \widehat{m}(X_i) \mid \mathcal{F}_0(\mathcal{I}))$. However, the FDR upper bound can be loose because the estimation error is bounded by the maximum:

$$\epsilon_n^q(\mathcal{I}) = \max\{q_{\max}(\mathcal{I}) - \widehat{q_{\max}}(\mathcal{I}), 0\} \leq \max\{\epsilon_n^\pi(\mathcal{I}), \epsilon_n^p(\mathcal{I}), 0\},$$

and thus, FDR is close to the target level when both estimation error, $\epsilon_n^\pi(\mathcal{I})$ and $\epsilon_n^p(\mathcal{I})$, are small.

"2" robustness. Define

$$\widehat{q_{\max}}(\mathcal{I}) := \max\{\max\{1-\widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}, \max\{1-\widehat{p_{\min}}(\mathcal{I}), \widehat{p_{\max}}(\mathcal{I})\}\},$$

which is more conservative as the deviation of $\widehat{q_{\max}}(\mathcal{I})$ from $1/2$ (desired value) is the maximum of the deviation for estimated propensity score and $\mathbb{P}(Y_i > \widehat{m}(X_i) \mid \mathcal{F}_0(\mathcal{I}))$. However, the FDR upper bound can be tighter because the estimation error is bounded by the maximum:

$$\epsilon_n^q(\mathcal{I}) = \max\{q_{\max}(\mathcal{I}) - \widehat{q_{\max}}(\mathcal{I}), 0\} \leq \max\{\min\{\epsilon_n^\pi(\mathcal{I}), \epsilon_n^p(\mathcal{I})\}, 0\},$$

and thus, FDR is close to the target level when either estimation error, $\epsilon_n^\pi(\mathcal{I})$ or $\epsilon_n^p(\mathcal{I})$, is small.

"1.5" robustness when estimating p_i Define

$$\widehat{q}_i(\mathcal{I}) := \max\{\widehat{p}_i(\mathcal{I}), 1-\widehat{p}_i(\mathcal{I})\} \quad \text{and } \widehat{q_{\max}}(\mathcal{I}) = \max_{i\in\mathcal{I}} \widehat{q}_i(\mathcal{I}), \tag{218}$$

whose conservativeness depends on the estimation of the outcome probability p_i for all $t \in [|\mathcal{I}|]$. To describe the resulting estimation error of q_i, we define a difference $d_i(\mathcal{I}) := \max\{\pi_i, 1-\pi_i\} - \max\{p_i, 1-p_i\}$, which takes large value if the propensity score deviates from $1/2$ (smaller value if the outcome probability deviates from $1/2$). The true q_i is upper bounded by estimated $\widehat{q}_i$ plus some estimation error that depends on $d_i(\mathcal{I})$:

$$\begin{aligned} q_i - \widehat{q}_i &\leq \epsilon_i^p(\mathcal{I}) && \text{if } d_i(\mathcal{I}) \geq 0; \\ q_i - \widehat{q}_i &\leq \epsilon_i^p(\mathcal{I}) + d_i(\mathcal{I}) && \text{if } d_i(\mathcal{I}) < 0, \end{aligned}$$

where $\epsilon_i^p(\mathcal{I}) = p_i - \widehat{p}_i(\mathcal{I})$; it can be written in one line as

$$q_i - \widehat{q}_i \leq \epsilon_i^p(\mathcal{I}) + \min\left\{0, \max\{\pi_i, 1-\pi_i\} - \max\{p_i, 1-p_i\}\right\}.$$

Thus, the estimation error for $q_{\max}$ is upper bounded as

$$\epsilon_n^q(\mathcal{I}) \leq \max_{i\in\mathcal{I}}\{\epsilon_i^p(\mathcal{I}) + \min\left\{0, \max\{\pi_i, 1-\pi_i\} - \max\{p_i, 1-p_i\}\right\}\}, \tag{219}$$

which indicates that the FDR control would be close to the desired level if either the outcome probability estimation has small error (i.e., small $\epsilon_i^p(\mathcal{I})$) or the true propensity score is close to $1/2$ while the outcome probability deviates from $1/2$. (There is one case where the FDR inflation could be large: the outcome probability is $1/2$, which occurs if the estimated outcome $\widehat{m}$ is the conditional mean $\mathbb{E}(Y_i \mid X_i)$ for all subjects, but poorly estimated).

"1.5" robustness when estimating π_i Similarly, when we use the originally proposed MaY-I$^3_{\widehat{\pi}}$ where

$$\widehat{q}_i(\mathcal{I}) := \max\{\widehat{\pi}_i(\mathcal{I}), 1 - \widehat{\pi}_i(\mathcal{I})\} \quad \text{and } \widehat{q_{\max}}(\mathcal{I}) = \max_{i\in\mathcal{I}} \widehat{q}_i(\mathcal{I}),$$

the resulting estimation error for $q_{\max}$ is upper bounded as

$$\epsilon^q_n(\mathcal{I}) \leq \max_{i\in\mathcal{I}}\{\epsilon^\pi_i(\mathcal{I}) - \max\{0, \max\{\pi_i, 1-\pi_i\} - \max\{p_i, 1-p_i\}\}\},$$

which indicates that the FDR control would be close to the desired level if either the propensity score estimation has small error (i.e., small $\epsilon^\pi_i(\mathcal{I})$) or the true propensity score is deviates from $1/2$ while the the outcome probability close to $1/2$, which occurs if the estimated outcome $\widehat{m}$ is a good estimation of the conditional mean $\mathbb{E}(Y_i \mid X_i)$. (There is one case where the FDR inflation could be large: the true propensity score is $1/2$ but poorly estimated).

Note that, however, using estimated outcome probability p_i might not be practically powerful because we cannot tell which ones are the nulls, and the outcome probability for the non-nulls would be close to zero or one. Consequently, the power of the methods with "2" robustness or "1.5" robustness with estimated p_i have almost zero power, and the method with "0.5" robustness has similar power as the MaY-I$^3_{\widehat{\pi}}$ proposed in the main paper.

References

Akritas, M. G., S. F. Arnold, and E. Brunner (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association 92*(437), 258–265. 4.1.4

Akritas, M. G., S. F. Arnold, and Y. Du (2000). Nonparametric models and methods for nonlinear analysis of covariance. *Biometrika 87*(3), 507–526. 4.1.4

Arias-Castro, E. and S. Chen (2017). Distribution-free multiple testing. *Electronic Journal of Statistics 11*(1), 1983–2001. 1, 2.1.2, 5.4, D.4, D.4, D.4, D.4.1, D.4.2

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*, 7353–7360. 5.1.2

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055–2085. 1, 2.1.2, 3.2, D.2, D.4

Barlow, R. E. and H. D. Brunk (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association 67*(337), 140–147. A.7

Bathke, A. and E. Brunner (2003). A nonparametric alternative to analysis of covariance. In *Recent advances and trends in nonparametric statistics*, pp. 109–120. Elsevier. 4.1.4

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300. 5.1.2, 5.3, D.4

Berger, J. O., X. Wang, and L. Shen (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics 24*(1), 110–129. 5.1.2

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32. 4.3.1

Bretz, F., W. Maurer, W. Brannath, and M. Posch (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine 28*(4), 586–604. 3.1

Cai, T., L. Tian, P. H. Wong, and L. Wei (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics 12*(2), 270–282. 5.1.2

Calel, R. and A. Dechezlepretre (2016). Environmental policy and directed technological change: evidence from the european carbon market. *Review of Economics and Statistics 98*(1), 173–191. 4.1

Cao, W., A. A. Tsiatis, and M. Davidian (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*(3), 723–734. 4.3.1

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. NBER working paper No. 23564. 4.3.1

Ding, P., L. Keele, et al. (2018). Rank tests in unmatched clustered randomized trials applied to a study of teacher training. *The Annals of Applied Statistics 12*(4), 2151–2174. 4.3

Donoho, D. and J. Jin (2015). Special Invited Paper: Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects. *Statistical Science*, 1–25. 2.1.1, 2.4.1, 11

Duan, B., A. Ramdas, S. Balakrishnan, and L. Wasserman (2019). Interactive martingale tests for the global null. *arXiv preprint arXiv:1909.07339*. 4.1.4, 5.2.1

Duan, B., A. Ramdas, and L. Wasserman (2020a). Familywise error rate control by interactive unmasking. In *International Conference on Machine Learning (accepted)*. 2.1.2, 5.2.1

Duan, B., A. Ramdas, and L. Wasserman (2020b). Which Wilcoxon should we use? An interactive rank test and other alternatives. *arXiv preprint arXiv:2009.05892*. 5.2.1

Fan, C. and D. Zhang (2017). Rank repeated measures analysis of covariance. *Communications in Statistics-Theory and Methods 46*(3), 1158–1183. 4.1.4

Fan, J., P. Hall, and Q. Yao (2007). To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied? *Journal of the American Statistical Association 102*(480), 1282–1288. D.2.4

Fang, Y., S. Tang, Z. Huo, G. C. Tseng, and Y. Park (2019). Properties of adaptively weighted Fisher's method. *arXiv preprint arXiv:1908.00583*. 2

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical methods in medical research 17*(4), 347–388. 3.1

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in Statistics*, pp. 66–70. Springer. 2.1.1, 2.2, A.5

Foster, J. C., J. M. Taylor, and S. J. Ruberg (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine 30*(24), 2867–2880. 5.1.2

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32*(200), 675–701. 4.4.2

Goeman, J. J. and L. Finos (2012). The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical applications in genetics and molecular biology 11*(1), 1–18. 3.3.3

Goeman, J. J. and A. Solari (2011). Multiple testing for exploratory research. *Statistical Science 26*(4), 584–597. 3.1

Goeman, J. J. and A. Solari (2014). Multiple hypothesis testing in genomics. *Statistics in medicine 33*(11), 1946–1978. 3.1

Grünwald, P., R. de Heide, and W. Koolen (2019). Safe testing. *arXiv preprint arXiv:1906.07801*. 2.7, 6

Gu, J. and S. Shen (2018). Oracle and adaptive false discovery rate controlling methods for one-sided testing: theory and application in treatment effect evaluation. *The Econometrics Journal 21*(1), 11–35. 5.1.2

Guo, K. and G. Basse (2020). The Generalized Oaxaca-Blinder Estimator. *arXiv preprint arXiv:2004.11615*. 4.3.1

Hettmansperger, T. P. and J. W. McKean (2010). *Robust nonparametric statistical methods*. CRC Press. 4.1.4

Himes, B. E., X. Jiang, P. Wagner, R. Hu, Q. Wang, B. Klanderman, R. M. Whitaker, Q. Duan, J. Lasky-Su, and C. Nikolos (2014). RNA-Seq transcriptome profiling identifies crispld2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PloS one 9*(6), e99625. 6

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 75*(4), 800–802. 3.1

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70. 3.1, 3.3.1

Howard, S. R. and S. D. Pimentel (2020). The uniform general signed rank test and its design sensitivity. *Biometrika*. asaa072. 4.3, 4.4.1, 5.1.1, 5.8

Howard, S. R., A. Ramdas, J. McAuliffe, and J. Sekhon (2020a). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys 17*, 257–317. 2.1.2, 2.2, 2.2, 4.1.4, 4.4.2, A.4, A.5, A.6

Howard, S. R., A. Ramdas, J. McAuliffe, and J. Sekhon (2020b). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics (accepted)*. 2.1.2, 2.2, 2.2, 2.7, 4.1.4, 4.4.2, A.4, A.5, A.6

Huang, M., R. Li, and S. Wang (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association 108*(503), 929–941. 4.2.2

Huo, Z., S. Tang, Y. Park, and G. Tseng (2020). P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. *Bioinformatics 36*(2), 524–532. 2, A.8

Ignatiadis, N., B. Klaus, J. B. Zaugg, and W. Huber (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods 13*(7), 577. 2.3.2, 3.5, 6

Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics 7*(1), 443–470. 5.1.2

Janson, L. and W. Su (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics 10*(1), 960–975. 3.2

Janssen, A. (2000). Global power functions of goodness of fit tests. *Annals of Statistics 28*(1), 239–253. 4.2.2

Kapelner, A. and A. Krieger (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics 70*(2), 378–388. 4.5

Karmakar, B., R. Heller, and D. S. Small (2018). False discovery rate control for effect modification in observational studies. *Electronic Journal of Statistics 12*(2), 3232–3253. (document), 5.1.2, 5.1.2, 5.9, 5.9, 5.9, 5.9, 5.9, 58

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*. 5.2, D.1.2

Kost, J. T. and M. P. McDermott (2002). Combining dependent p-values. *Statistics & Probability Letters 60*(2), 183–190. 2.1.2

Kruskal, W. H. and W. A. Wallis (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association 47*(260), 583–621. 4.4.2

Lehmann, E. L. and H. J. D'Abrera (1975). *Nonparametrics: statistical methods based on ranks*. Holden-day. 4.3

Lei, L. and W. Fithian (2018). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 649–679. 1, 2.1.1, 2.1.2, 2.3.2, 3.1, 3.2, 3.5, 6, 4.1.4, 5.2.1, 6, B.1, D.2, 9

Lei, L., A. Ramdas, and W. Fithian (2020). STAR: A general interactive framework for FDR control under structural constraints. *Biometrika (accepted)*. 1, 2.1.2, 2.3.2, 2.8, 3.1, 3.2, 4.1.4, 5.2.1, 5.9, 5.9, 6, D.2, D.6

Li, A. and R. F. Barber (2017). Accumulation tests for FDR control in ordered hypothesis testing. *Journal of the American Statistical Association 112*(518), 837–849. D.2

Li, J. and G. C. Tseng (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics 5*(2A), 994–1019. 2

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics 7*(1), 295–318. 4.3.1

Lipkovich, I. and A. Dmitrienko (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of Biopharmaceutical Statistics 24*(1), 130–153. 5.1.2

Lipkovich, I., A. Dmitrienko, and R. B D'Agostino Sr (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine 36*(1), 136–196. 5.1, 5.1.2

Lipkovich, I., A. Dmitrienko, J. Denne, and G. Enas (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine 30*(21), 2601–2621. 5.1.2

Loh, W.-Y., L. Cao, and P. Zhou (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9*(5), e1326. 5.1

Love, M. I., W. Huber, and S. Anders (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology 15*(12), 550. 6

Marcus, R., P. Eric, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*(3), 655–660. 3.2, B.3

Matsumoto, M. and O. Hikosaka (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature 459*, 837–841. 4.1

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika 95*(2), 265–278. 3.3.3

Nie, X. and S. Wager (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*. asaa076. 5.2

Olive, K. P., M. A. Jacobetz, C. J. Davidson, A. Gopinathan, D. McIntyre, D. Honess, B. Madhu, M. A. Goldgraben, M. E. Caldwell, D. Allard, et al. (2009). Inhibition of Hedgehog signaling enhances delivery of chemotherapy in a mouse model of pancreatic cancer. *Science 324*(5933), 1457–1461. 4.1

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *The Annals of Statistics 37*(6B), 3867–3892. 2.1.2

Powers, S., J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine 37*(11), 1767–1787. 5.1

Rabinovich, M., A. Ramdas, M. I. Jordan, and M. J. Wainwright (2020). Optimal rates and trade-offs in multiple testing. *Statistica Sinica 30*, 741–762. 5.4

Ramdas, A., J. Ruf, M. Larsson, and W. Koolen (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*. 2.7, 6

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics 41*(5), 1397–1409. 2.1.2

Robertson, T., F. Wright, and R. Dykstra (1988). Order restricted statistical inference. A.7

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866. 4.3.1

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the*

Econometric Society 56(4), 931–954. 4.3.1, 5.2

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science 17*(3), 286–327. 4.1.2, 4.3, 4.4.1, 5.8

Rosenbaum, P. R. (2010). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association 105*(490), 692–702. 4.3

Rosenblum, M. and M. J. Van Der Laan (2009). Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics 65*(3), 937–945. 4.3

Rüger, B. (1978). Das maximale Signifikanzniveau des Tests: "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika 25*(1), 171–178. 2.1.2

Rüschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability 14*(3), 623–632. 2.1.2

Shafer, G., A. Shen, N. Vereshchagin, and V. Vovk (2011). Test martingales, Bayes factors and p-values. *Statistical Science 26*(1), 84–101. 2.8

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633. (document), 16

Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *The Annals of Statistics*, 361–404. 2.1.2, 4.1.4

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*(3), 751–754. 2.1.2

Sivaganesan, S., P. W. Laud, and P. Müller (2011). A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in Medicine 30*(4), 312–323. 5.1.2

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr (1949). The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. 2.1.2, 2.2, 2.2

Tamhane, A. C. and J. Gou (2018). Advances in p-value based multiple test procedures. *Journal of biopharmaceutical statistics 28*(1), 10–27. 3.1

Thas, O., J. D. Neve, L. Clement, and J.-P. Ottoy (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 623–671. 4.1.4

Train, K. E. (2008). EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling 1*(1), 40–69. 4.2.2

Vansteelandt, S. and M. Joffe (2014). Structural nested models and G-estimation: the partially realized promise. *Statistical Science 29*(4), 707–731. C.5, C.5

Vermeulen, K., O. Thas, and S. Vansteelandt (2015). Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment. *Statistics in Medicine 34*(6), 1012–1030. 4.3

Ville, J. (1939). *1ère thèse: Etude critique de la notion de collectif; 2ème thèse: La transformation de Laplace*. Ph. D. thesis, Gauthier-Villars & Cie. 6

Vovk, V. and R. Wang (2020a, 06). Combining p-values via averaging. *Biometrika*. asaa027. 2.1.2

Vovk, V. and R. Wang (2020b). Combining p-values via averaging. *Biometrika*. asaa027. C.6

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics 16*(2),

117–186. 2.2

Wang, L. and M. G. Akritas (2006). Testing for covariate effects in the fully nonparametric analysis of covariance model. *Journal of the American Statistical Association 101*(474), 722–736. 4.1.4

Wiens, B. L. and A. Dmitrienko (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics 15*(6), 929–942. 3.2, B.3

Xie, Y., N. Chen, and X. Shi (2018). False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 5.1.2

Zhang, K., M. Traskin, and D. S. Small (2012). A powerful and robust test statistic for randomization inference in group-randomized trials with matched pairs of groups. *Biometrics 68*(1), 75–84. 4.3

Zhang, M., S. Gelfman, J. McCarthy, M. B. Harms, C. A. Moreno, D. B. Goldstein, and A. S. Allen (2020). Incorporating external information to improve sparse signal detection in rare-variant gene-set-based analyses. *Genetic Epidemiology 44*(4), 330–338. 2

Zhao, Q., D. S. Small, and W. Su (2019). Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association 114*(527), 1291–1304. B.1

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association 107*(499), 1106–1118. 5.1.2