Electronic Journal of Statistics
Vol. 14 (2020) 4489–4551
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1790

Interactive martingale tests for the global null

Boyan Duan and Aaditya Ramdas and Sivaraman Balakrishnan and Larry Wasserman

Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213,
e-mail: boyand@stat.cmu.edu; aramdas@stat.cmu.edu; siva@stat.cmu.edu; larry@stat.cmu.edu

Abstract: Global null testing is a classical problem going back about a century to Fisher's and Stouffer's combination tests. In this work, we present simple martingale analogs of these classical tests, which are applicable in two distinct settings: (a) the online setting in which there is a possibly infinite sequence of p-values, and (b) the batch setting, where one uses prior knowledge to preorder the hypotheses. Through theory and simulations, we demonstrate that our martingale variants have higher power than their classical counterparts even when the preordering is only weakly informative. Finally, using a recent idea of "masking" p-values, we develop a novel interactive test for the global null that can take advantage of covariates and repeated user guidance to create a data-adaptive ordering that achieves higher detection power against structured alternatives.

MSC2020 subject classifications: Primary 60G10, 62M07.
Keywords and phrases: Interactive testing, global null, data carving.

Received December 2019.

Contents

1 Introduction 4490
1.1 Assumptions 4492
1.2 Related work 4492
1.3 Outline 4494
2 The preordered martingale test 4494
3 Adaptive and interactive methods 4496
3.1 The adaptively ordered martingale test (AMT) 4496
3.2 The interactively ordered martingale test (IMT) 4497
4 Power guarantees of non-interactive procedures 4501
4.1 Power guarantees in the batch setting 4503
4.2 Power guarantees in the online setting 4506
5 Numerical simulations 4508
5.1 Clustered non-nulls in a grid of hypotheses 4509
5.2 A sub-tree of non-nulls in a tree of hypotheses 4511
5.3 Structures in the online setting 4512
6 Robustness to conservative nulls 4515

7 Anytime-valid p-values and safe e-values 4516
8 Alternative masking functions 4517
9 Summary 4520
A Error control 4521
A.1 Proof of Theorem 1 4521
A.2 Proof of Theorem 3 4521
A.3 Error control of the interactively ordered martingale test with railway masking function in Section 6 4523
B Power guarantees in the batch setting 4524
B.1 Proof of Theorem 4 4524
B.1.1 The batch Stouffer test 4524
B.1.2 The martingale Stouffer test 4525
B.2 Proof of Theorem 5 4526
B.3 Proof of condition (13) in the main paper 4529
C Power guarantees in the online setting 4532
C.1 Proof of Theorem 6 4532
C.2 Proof of Theorem 7 4537
D Choices for the uniform bounds in the martingale Stouffer test . . . 4540
E Martingale Fisher test 4543
F Martingale chi-squared test 4544
G Bayesian modeling for the posterior probability of being non-null . . 4545
H Comparison with alternative methods 4548
Acknowledgements 4549
References 4549

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 (3)). 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) 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 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 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 7 explicitly connects these two interpretations.

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:

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

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

$$\text{If } \mathcal{H}_{\mathcal{G}0} \text{ is true, } f_i(a) \leq f_i(1-a) \text{ for all } 0 \leq a \leq 0.5, i \in \mathcal{I}, \tag{2}$$

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 [14]. 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 [7] and Tukey's Higher Criticism [4] 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.

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 [22, 23, 13, 18, 29]. Our work is most directly connected to tests which accumulate information as a sum, such as Fisher's and Stouffer's tests [27].

There are many other global null tests like the Bonferroni method, Simes' test [26], 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,

$$h(p_i) = 2 \cdot 1\{p_i < 0.5\} - 1 \quad \text{and} \quad g(p_i) = \min\{p_i, 1 - p_i\}. \tag{3}$$

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 [14] in the context of false discovery rate (FDR) control, and further generalized and extended in Lei, Ramdas and Fithian [15] for FDR control under structural constraints, and then followed by work on FWER control [5]. The underlying property of masking can be traced to the "knockoff" method by Barber and Candès [2, 1]. 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 [25]. In this paper, we care about nonasymptotic guarantees on the type-I error, and hence we use some recent line-crossing inequalities [9] and new curve-crossing inequalities [10] that are nonasymptotic generalizations of the law of the iterated logarithm, which goes back to the work by Robbins [20] (see Appendix D 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.

1.3. Outline

To progressively build intuition, the preordered martingale test is described in Section 2 followed by the adaptively ordered martingale test in Section 3.1. In Section 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 4. We then perform extensive simulations in Section 5. In Section 6, we examine the robustness of our test to conservative nulls. Section 7 explicitly describes how to interpret our tests as tracking an anytime-valid sequential p-value. Finally in Section 8, we discuss alternative ways of masking p-values. We end with a brief summary in Section 9, and defer all proofs and additional experiments to the Appendix.

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 [27] and Fisher's method [7], 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:

$$\text{Reject the null if } \sum_{i=1}^{k} f(p_i) \geq u_\alpha(k), \text{ for some } k \in \mathcal{I}, \tag{4}$$

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 [9, 10] 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 E for a martingale Fisher test using $f(p_i) = -2\log p_i$, and Appendix F for a martingale chi-square test using $f(p_i) = \left[\Phi^{-1}(1-p_i)\right]^2$.

An example: martingale Stouffer test (MST). The batch test by Stouffer [27] 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. [9]), which can be derived from the Gaussian sequential probability ratio test [30], grows linearly with time. Specifically, the test rejects the global null if

$$\exists k \in \mathbb{N} : \sum_{i=1}^{k} \Phi^{-1}(1-p_i) \geq \sqrt{\frac{-\log\alpha}{2m}}\,k + \sqrt{\frac{-m\log\alpha}{2}}, \tag{5}$$

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. [10]) rejects the global null if

$$\exists k \in \mathbb{N} : \sum_{i=1}^{k} \Phi^{-1}(1-p_i) \geq 1.7\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)}. \tag{6}$$

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 D. 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* (5) *or curved boundary* (6) *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.

3. Adaptive and interactive methods

To develop intuition progressively, we first introduce a martingale test whose ordering depends on the p-values in Section 3.1, and extend it in Section 3.2 to an interactive test, whose ordering can additionally depend on side information (covariates) and human interaction.

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 (5) and (6), 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* (2), 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.

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 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 [12], AdaPT [14], and STAR [15]. 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.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:

$$p_i \sim (1-\pi_i)F_0 + \pi_i F_1. \tag{7}$$

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 G). 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.

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:*

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

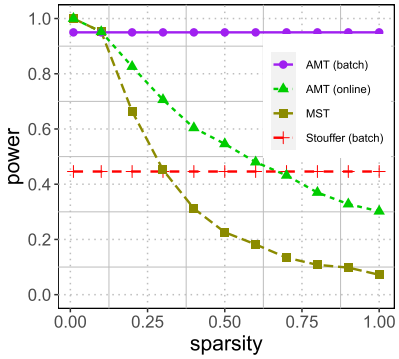
In the remainder of this section, we let $r_i := I(\mu_i > 0)$ indicate the non-null hypotheses. Although we compare the power of various tests in this relatively simple setting, we emphasize that our tests are more broadly applicable to general settings where the p-values are mirror-conservative under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

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

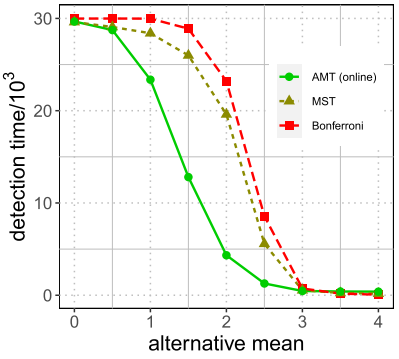
- In Section 4.1, we focus on the batch setting. In Theorem 4, we compare the power of the martingale Stouffer test with its batch counterpart, showing that when a good a-priori ordering is used the martingale Stouffer test can have much higher power. Our next result, Theorem 5, studies the adaptively 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 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 1 visualizes the above power comparisons by two simple simulations in batch and online settings[1]. Details of the batch experiment appear next.

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



(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.

FIG 1. *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* $\alpha = 0.05$.

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 5.

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:

$$C_k^\gamma := 1.7\sqrt{\log\log(2k) + 0.72\log\frac{5.2}{\gamma}}, \tag{8}$$

which corresponds to the curved boundary in (6) 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*

$$\sum_{i=1}^{n} r_i\mu_i \geq (Z_\alpha + Z_\beta)n^{1/2}, \tag{9}$$

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*

$$\exists\ k \in \{1,\ldots,n\}, \quad \sum_{i=1}^{k} r_i\mu_i \geq \left(C_k^\alpha + C_k^\beta\right)k^{1/2}. \tag{10}$$

(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 B.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 (9) for the batch Stouffer test with its martingale counterpart (condition (10)), 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

$$|Z_{(1)}(\mu)| \geq |Z_{(2)}(\mu)| \geq \ldots \geq |Z_{(N_1)}(\mu)|. \tag{11}$$

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 (8). 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*

$$\exists\, j \in \{1, \ldots, N_1\} : \sum_{s=1}^{j} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1\right) \geq \left(C_n^\alpha + C_n^{\beta/2}\right) \left(j + t_{\beta/(2N_1)}(N_0, q_j)\right)^{1/2}. \tag{12}$$

We prove this result in Appendix B.2. Condition (12) gives a reasonably tight sufficient condition for the re-ordering based test to have high power (Figure 2). 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.

- 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 (12) grows linearly with j. On the other hand, if the signal strength is large then the prob-

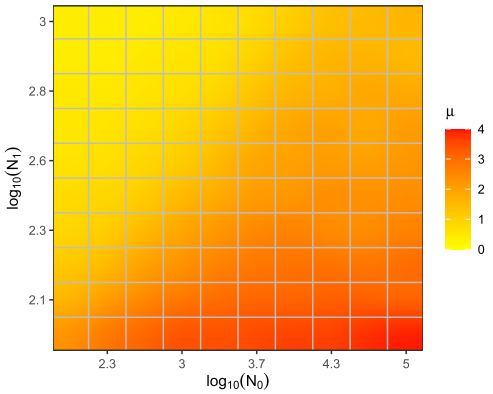


FIG 2. *Sufficient signal strength μ for AMT to guarantee both type-I and type-II error control at 0.05 (derived from (12)), when varying the numbers of nulls $N_0 \in [10^2, 10^5]$ and non-nulls $N_1 \in [10^2, 10^3]$. The required signal strength grows when the number of nulls increases or the number of non-nulls decreases.*

abilities 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 B.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

$$\mu \geq \sqrt{2\log\left(\frac{N_0}{N_1^2}\right) + 4\log\left(C_n^{\alpha} + C_n^{\beta/2}\right) + 3.45}. \tag{13}$$

  For comparison, the batch Stouffer test requires

$$\mu \geq (Z_{\alpha} + Z_{\beta})\sqrt{\frac{N_0}{N_1^2} + \frac{1}{N_1}}. \tag{14}$$

  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 B.3, we relate condition (13) to the detection threshold derived in the work of Donoho and Jin [4] 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.

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,

$$\sum_{k=1}^{\infty} r_k = \infty. \tag{15}$$

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

$$\sum_{k=1}^{\infty} r_k \leq M < \infty \text{ for some large constant } M. \tag{16}$$

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 (15), a necessary condition for online Bonferroni to have at least $1-\beta$ power is*

$$\exists k \in \mathbb{N} : r_k \mu_k \geq 0.25 \left(\sqrt{2 \log \left(\frac{k^2}{\alpha} \right)} \right)^{-1}. \tag{17}$$

A stronger necessary condition can be derived for sparse non-nulls (16). *If* $\{\alpha_k\}_{k=1}^{\infty}$ *is nonincreasing, then online Bonferroni can have at least* $1-\beta$ *power only if*

$$\exists k \in \mathbb{N}: \begin{cases} r_k\mu_k \geq 0.4\sqrt{\alpha_{k^*}}, & \text{if } k \leq k^*, \\ r_k\mu_k \geq \sqrt{\log\left(\frac{k}{4\alpha}\right)} - \sqrt{2\log\left(\frac{M}{2(1-\beta-3\alpha)}\right)}, & \text{if } k > k^*, \end{cases} \tag{18}$$

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*

$$\exists\ k \in \mathbb{N}: \sum_{i=1}^{k} \mu_i r_i \geq (C_k^{\alpha} + C_k^{\beta})k^{1/2}. \tag{19}$$

Remarks:

- Condition (19) 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 (17) 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 C.1).
- The proof of Theorem 6 is in Appendix C.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 (17)) whereas the martingale Stouffer test has power that approaches 1 (by condition (19)).

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:

$$A(\mu; c) = \frac{5}{3}\frac{\sqrt{\Phi(-c)}}{\Phi(\mu-c) - \Phi(-\mu-c)},$$

$$B(\mu;c) = \frac{10(\Phi(\mu-c)+\Phi(-\mu-c)-2\Phi(-c))}{9(\Phi(\mu-c)-\Phi(-\mu-c))^2} \vee \frac{25}{(\Phi(\mu-c)+\Phi(-\mu-c))^2},$$
$$T(\beta;c) = \frac{0.79\log(15.57/\beta)\Phi^2(-c)+0.4}{\Phi^4(-c)}.$$

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:

$$A(\mu;\mu) \le e^{-\mu^2/4}. \tag{20}$$

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:*

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

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

$$\forall\ k \in \mathbb{N} : \quad \sum_{i=1}^{k} r_i \le \mu^{-1}\left(C_k^{\alpha}-C_k^{1-\beta}\right)k^{1/2}. \tag{21}$$

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 (20) for $A(\mu;\mu)$) is much smaller than the corresponding $1/\mu$ term in condition (21) 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.

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

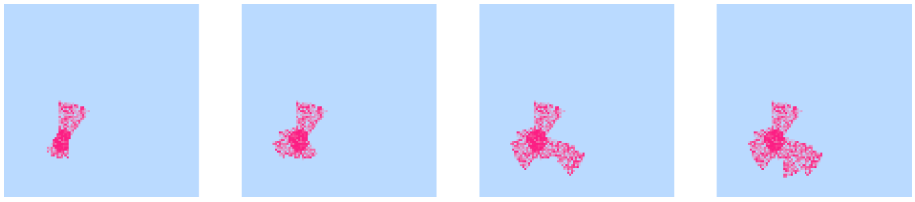


FIG 3. *Visualization of the interactively ordered martingale test under the block structure: the hypotheses in M_k, which interactively expands (darker color indicates a lower p-value and possible non-null).*

test combines (a) side covariate information (which could include prior non-null probabilities in working model (7) 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.

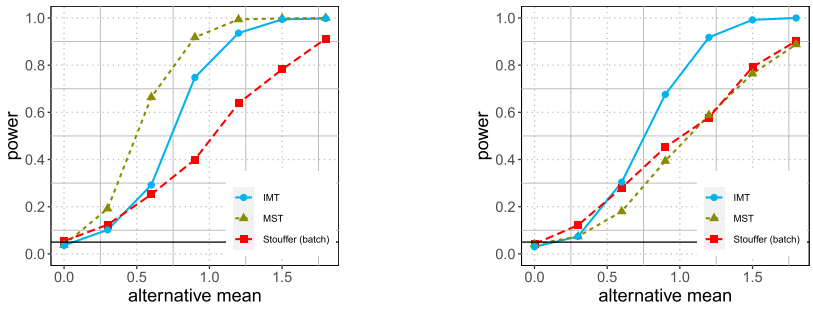
We first consider the batch setting and use two non-null structures as simple examples: a blocked structure within a grid and a hierarchical structure within a tree; and we discuss similar structures in the online setting. For each of these, we customize a heuristic strategy to expand M_k in the interactively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

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 (7) by constraining the prior non-null probabilities π_i to be a smooth function of x_i. Details can be found in Appendix G.

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 3 for example).

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



(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.

FIG 4. *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* $\alpha = 0.05$.

grid. We use Setting 1 as defined in Section 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 4), 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 [16, 6, 11] and the weighted Higher Criticism [31] in Appendix H. 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.*

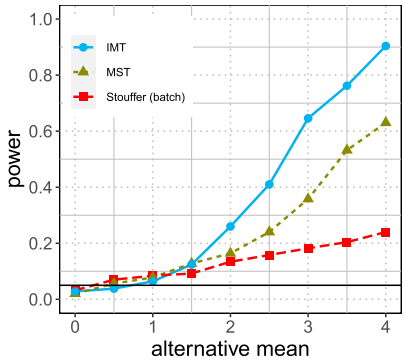


FIG 5. *Power of the interactively ordered martingale test (IMT), the martingale Stouffer test (MST), and the batch Stouffer test under a hierarchical structure. Hypotheses form a fixed tree (batch setting) with non-nulls only on a sub-tree. When the alternative mean is big, masked p-values and the hierarchical non-null structure lead to a good ordering and hence high power for the IMT.*

5.2. A sub-tree of non-nulls in a tree of hypotheses

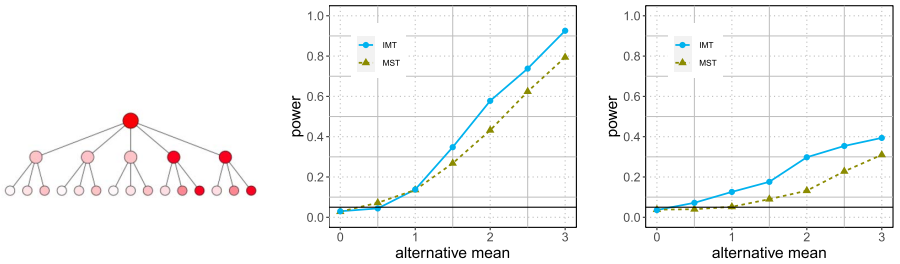
In applications such as wavelet decomposition, the hypotheses can have a hierarchical structure, where the child can be a non-null only if its parent is a non-null. The hierarchical structure is again encoded in modeling the posterior probability of being non-null (7) by adding a partial order constraint 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 5).

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



(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.

FIG 6. *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.*

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 6). Compared with decreasing probability of being non-null, both methods have lower power for the tree 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.

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 $c = 0.05$.

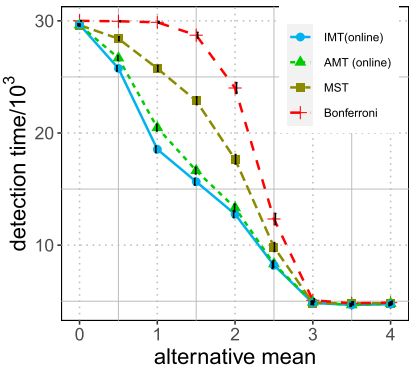


FIG 7. *Number of hypotheses needed to reject the global null (detection time) in the online setting of the interactively ordered martingale test (IMT), the adaptively ordered martingale test (AMT), the martingale Stouffer test (MST), and the Bonferroni test when varying the alternative mean μ. The non-nulls 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 7). 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 (7), 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.

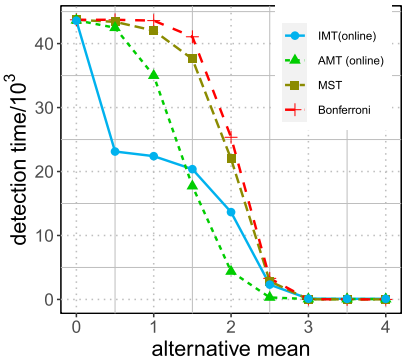
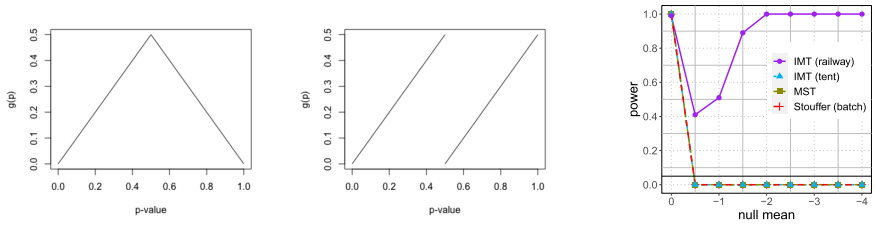


FIG 8. *Number of hypotheses needed to reject the global null (detection time) in the online setting of the interactively ordered martingale test (IMT), the adaptively ordered martingale test (AMT), the martingale Stouffer test (MST), and the Bonferroni test when varying the alternative mean in a growing 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 8). 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).



(a) The original masking function (tent).

(b) The modified masking function (railway).

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

FIG 9. *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.*

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 (1)) or mirror-conservative (condition (2)); 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 (1) and condition (2). 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 $\widetilde{g}(p) := \min(p, (p+\frac{1}{2}) \bmod 1)$. Recalling that $g(p) = \min(p, 1-p)$, we call g and $\widetilde{g}$ as the tent and railway functions respectively (see Figure 9a, Figure 9b). 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.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.

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 9c 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 9c.

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:*

$$\mathfrak{p}_t = \inf\{\alpha : \exists k \in \{1, \ldots, t\} \text{ s.t. } S_k > u_\alpha(k)\}. \tag{22}$$

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 (5), 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:

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

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

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

   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:

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

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. [10]. An alternative to anytime p-values, called safe e-values, was recently proposed by Grünwald et al. [8], and their relationship to confidence sequences, sequential tests and anytime p-values was detailed by Ramdas et al. [19]. 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 $1/\alpha$).

8. Alternative masking functions

In most of this paper, we have considered one way of decomposing p-value as equation (3), but interactive tests can be developed for other decompositions. Shafer et al. [24] 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:

$$\mathbb{E}(f(p) \mid g(p)) = \mathbb{E}(f(p)). \tag{23}$$

Lei et al. [15] 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

$$\prod_{i=1}^{M_k} f(p_i) \geq \alpha^{-1}, \tag{24}$$

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 $c \in (0, 1)$:

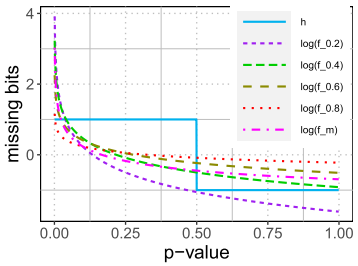
$$f_c(p) = cp^{c-1}. \tag{25}$$

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

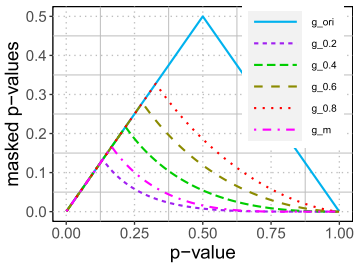
$$g_c(p_i) = \begin{cases} p_i, & \text{if } p_i \leq p_* \\ s(p_i), & \text{otherwise,} \end{cases}$$

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 10 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 (3), f_c maps small p-values to large value (Figure 10a), 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)



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



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

FIG 10. *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.*

would have small masked p-value (Figure 10b), 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:

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

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 10, the amount of information that f_m takes for inference is between $f_{0.2}$ and $f_{0.4}$.

We compare the interactively ordered martingale tests using different missing bits: (a) the original one bit $h(p_i)$ defined in equation (3); (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 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 11).

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 4a). 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?

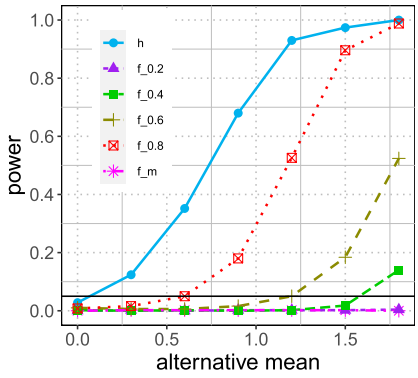


FIG 11. *Power of interactive tests using different missing bits. Under the block structure of non-nulls as described in Section 5.1, the IMT with the original missing bit defined in equation (3) has the highest power.*

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

$$\mathbb{E}\left[h(T_i) \mid g(T_i)\right] \leq 0 \quad \text{for all } i \in \mathcal{I}. \tag{26}$$

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 $\text{sign}(T_i)$ 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.

We believe that interactive testing protocols are only beginning to be explored in the literature, and constitute both an intellectually fascinating direction for further exploration, as well as a potentially powerful one. Masking (and progressive unmasking) is a promising technique that permits interaction, and it deserves further scrutiny and generalization to other settings.

Appendix A: 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. 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.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

$$\mathbb{E}\Big(\sum_{i\in M_k} h(p_i) \mid \mathcal{F}_{k-1}\Big) \leq \sum_{i\in M_{k-1}} h(p_i). \tag{27}$$

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 (27) boils down to proving

$$\mathbb{E}(h(p_{i_k^*}) \mid \mathcal{F}_{k-1}) \le 0.$$

Since i_k^* and M_{k-1} are $\mathcal{F}_{k-1}$ measurable, and $i_k^* \notin M_{k-1}$, we see that

$$\mathbb{E}(h(p_{i_k^*}) \mid \mathcal{F}_{k-1}) \le \max_{i\notin M_{k-1}} \mathbb{E}(h(p_i) \mid \mathcal{F}_{k-1}) = \max_{i\notin M_{k-1}} \mathbb{E}(h(p_i) \mid g(p_i)),$$

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

$$\mathbb{E}(h(p_i) \mid g(p_i)) \le 0, \tag{28}$$

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, so the test using a bound for a Gaussian increment martingale controls the type-I error (because a Rademacher is sub-Gaussian).

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 (28) 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:

$$\begin{aligned}\mathbb{P}_0(h(p_i) = -1 \mid g(p_i) = x) &= \frac{f_i(1-x)}{f_i(1-x) + f_i(x)} \\ \ge\ \frac{f_i(x)}{f_i(1-x) + f_i(x)} &= \mathbb{P}_0(h(p_i) = 1 \mid g(p_i) = x),\end{aligned}$$

for any $x \in [0, 0.5]$ (i.e., the range of $g(p_i)$), which implies condition (28) 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

$$\mathcal{G}_{k-1} := \mathcal{F}_{t_k^*-1}, \tag{29}$$

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

$$\mathbb{E}(h(p_{t_k^*}) \mid \mathcal{G}_{k-1}) \le 0.$$

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 (28) 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, the interactively ordered martingale test in the online setting using bound for a Gaussian increment martingale controls the type-I error. □

A.3. Error control of the interactively ordered martingale test with railway masking function in Section 6

Let the masked p-values defined by the railway function in Section 6 be:

$$\widetilde{g}(p) := \min(p, (p + \tfrac{1}{2}) \mathrm{mod} 1)$$

The corresponding interactively ordered martingale test has a valid error control when the p-values have nondecreasing densities under the global null.

Theorem 8. *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.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 (28) 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 (28) and thus completes the proof. □

Remark 3. *The above proof implies that the error control holds as long as under the global null, the p-values satisfy:*

$$f_i(a) \le f_i(a+0.5) \text{ for all } 0 \le a \le 0.5, i \in \mathcal{I},$$

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 (1) and (2) in the main paper. It is not a consequence of condition (1) (take $f(a) = \mathbb{1}(a \le 0.5) + 4(a-0.5)\mathbb{1}(a > 0.5)$) or condition (2) (take $f(a) = 4\min(a, 1-a)$), and it does not imply condition (1) and (2) (take $f(a) = 4(0.5-a)\mathbb{1}(a < 0.5) + 4(1-a)\mathbb{1}(0.5 \le a < 1) + 4\mathbb{1}(a = 1)$). For simplicity, we focus on the p-values with increasing densities in Section 6, which are considered as conservative p-values in all three definitions.

Appendix B: 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.

B.1. Proof of Theorem 4

We divide the proof into two subsections for the batch Stouffer test and the martingale Stouffer test.

B.1.1. 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

$$\begin{aligned}\mathbb{P}_1\left(\frac{S_n}{\sqrt{n}} \ge \Phi^{-1}(1-\alpha)\right) &= \mathbb{P}_1\left(\frac{S_n - \sum_{i=1}^n r_i\mu_i}{\sqrt{n}} \ge \Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^n r_i\mu_i}{\sqrt{n}}\right) \\ &= 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^n r_i\mu_i}{\sqrt{n}}\right).\end{aligned}$$

A power of at least $1-\beta$ is is equivalent to

$$1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^n r_i\mu_i}{\sqrt{n}}\right) \ge 1-\beta,$$

which can be rewritten as

$$\sum_{i=1}^n r_i\mu_i \ge (\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))n^{1/2},$$

which is the condition in Theorem 4. □

B.1.2. The martingale Stouffer test

Proof. Following the same proof for $S_n \sim N(r_i\mu_i, 1)$ in Section B.1.1, 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

$$\begin{aligned} & \mathbb{P}_1\left(\exists k \in \{1, \ldots, n\} : S_k - \sum_{i=1}^k r_i\mu_i \geq u_\alpha(k) - \sum_{i=1}^k r_i\mu_i\right) \\ \leq \; & \mathbb{P}_1\left(\exists k \in \{1, \ldots, n\} : S_k - \sum_{i=1}^k r_i\mu_i \geq u_{1-\beta}(k)\right) \leq 1 - \beta. \end{aligned}$$

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 (6) in the main paper into the conditions. If without further explanation, $u_\alpha(k)$ in rest of the proofs denotes the curved bound. □

B.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:

$$\frac{h(p_{i_k^*}) + 1}{2} \sim r_i \mathrm{Ber}(S_{i_k^*}(1)) + (1-r_i)\mathrm{Ber}(S_{i_k^*}(0)).$$

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

$$\begin{aligned}
&\mathbb{P}_1\left(\exists k \in \{1,\dots,n\} : S_k \geq u_\alpha(k)\right) \\
=\ &\mathbb{P}_1\left(\exists k \in \{1,\dots,n\} : S_k - \sum_{i\in M_k} [r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1)]\right. \\
&\left. \geq u_\alpha(k) - \sum_{i\in M_k} [r_i(2S_i(1)-1) + (1-r_i)(2S_i(0)-1)]\right).
\end{aligned}$$

The proof can be completed by following similar steps in the proof for martingale Stouffer test (Section B.1.2). □

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 (11), 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 4. *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 4). The quantile is upper bounded (by Chernoff inequality):

$$t_{\beta/N_1}(N_0, q_j) \leq \mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)N_0 + \sqrt{2\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)N_0 \log(\tfrac{N_1}{\beta})}$$
$$\leq (1 + \sqrt{2\log(\tfrac{N_1}{\beta})}) \max\{\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)N_0, 1\}.$$

The proof completes by showing that the probability term $\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|)$ is upper bounded:

$$\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|) \leq \frac{2P(\mu)}{N_1 \left[\frac{N_1+1-j}{N_1} - P(\mu)\right]^2}. \tag{30}$$

The above bound (30) 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 $\text{Bin}(N_1, P(\mu))$ and bound (30) holds because

$$\begin{aligned}
&\mathbb{P}(|Z(0)| > |Z_{(j)}(\mu)|) = \mathbb{P}(X > N_1 - j + 1) \\
\leq\ & \exp\left\{-\frac{[N_1(1-P(\mu)) - j + 1]^2}{2N_1 P(\mu)(1-P(\mu))}\right\} \leq \exp\left\{-\frac{N_1\left[\frac{N_1+1-j}{N_1} - P(\mu)\right]^2}{2P(\mu)}\right\} \\
\leq\ & \frac{2P(\mu)}{N_1\left[\frac{N_1+1-j}{N_1} - P(\mu)\right]^2},
\end{aligned}$$

for $j = 1, \ldots, \lfloor N_1(1 - P(\mu)) + 1 \rfloor$. The proof of Remark 4 is completed by plugging bound (30) 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

$$\exists k \in \{1, \ldots, n\} :$$
$$\sum_{i \in M_k} (r_i(2S_i(1) - 1) + (1 - r_i)(2S_i(0) - 1)) \geq \left(C^{\alpha}_{|M_k|} + C^{\beta}_{|M_k|}\right)(|M_k|)^{1/2}, \tag{31}$$

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

$$\exists j \in \{1, \ldots, N_1\} : \sum_{s=1}^{j} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1 \right) \geq \left(C^{\alpha}_{n} + C^{\beta}_{n} \right) (j + t_{\beta/N_1}(N_0, q_j))^{\frac{1}{2}}, \tag{32}$$

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 (32) 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 (31) 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 (32) with $\beta/2$. □

B.3. Proof of condition (13) 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

$$\sum_{s=1}^{N_1/2} \left(2\mathbb{P}(Z_{(s)}(\mu) > 0) - 1 \right) \geq \left(C^{\alpha}_{n} + C^{\beta/2}_{n} \right) \left(N_1/2 + t_{\beta/(2N_1)} \left(N_0, q_{N_1/2} \right) \right)^{1/2}. \tag{33}$$

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 4 in Section B.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 (33) 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 5. *Condition (13) in the main paper falls within the "detectable region" derived in the work of Donoho and Jin [4]: 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*

$$\mu \geq \sqrt{\log\left(\frac{n}{N_1^2}\right)}, \qquad \text{when } n^{1/4} \leq N_1 \leq n^{1/2}, \tag{34}$$

$$\mu \geq \sqrt{2}(\sqrt{\log n} - \sqrt{\log N_1}), \qquad \text{when } 1 < N_1 < n^{1/4}. \tag{35}$$

Proof. First note that condition (13) 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 $2 \le N_1 \le n^{1/2}$ and it is obvious when $N_1 = 1$. So when $n^{1/4} \le N_1 \le n^{1/2}$, condition (13) is a subset in the detectable region (34).

When $1 < N_1 < n^{1/4}$, denote $N_1 = n^a$ where $0 < a < 1/4$. The detectable region (35) can be written as

$$\mu \ge (1 - \sqrt{a})\sqrt{2 \log n},$$

which is implied by condition (13), since

$$\sqrt{2 \log \left(\frac{n}{N_1^2} \right)} = \sqrt{1 - 2a}\sqrt{2 \log n} \ge (1 - \sqrt{a})\sqrt{2 \log n},$$

when $a < 1/4$. Hence condition (13) is a subset of the detectable region (34) and (35). □

Appendix C: 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.

C.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 B.1.2), 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:

$$\mathbb{P}(\exists k \in \mathbb{N} : p_k < \alpha_k) \le \sum_{k=1}^{\infty} \mathbb{P}(p_k < \alpha_k) = \sum_{k=1}^{\infty} 1 - \Phi[\Phi^{-1}(1 - \alpha_k) - r_k \mu_k]. \quad (36)$$

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:*

$$r_k \mu_k \le \frac{1}{4\Phi^{-1}(1 - \alpha_k)}, \quad (37)$$

the power of rejecting individual hypothesis k is upper bounded:

$$1 - \Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \le C \cdot \alpha_k,$$

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 $a(C) \ge 0.3$ when $C \ge 2$. □

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

$$\sum_{k=1}^{\infty} 1 - \Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \le 1 - \beta,$$

which holds if for each individual hypothesis k with a positive error budget (i.e., $\alpha_k > 0$), the power of rejection is bounded

$$1 - \Phi\left[\Phi^{-1}(1-\alpha_k) - r_k\mu_k\right] \le \frac{1-\beta}{\alpha}\alpha_k, \tag{38}$$

where the upper bound $\frac{1-\beta}{\alpha}\alpha_k$ is chosen to satisfy two conditions: (a) the overall power is less than $1-\beta$: $\sum_{k=1}^{\infty} \frac{1-\beta}{\alpha}\alpha_k \le 1-\beta$ 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 (37) and $\alpha_k < 0.3$ (Notice that here the constant in the lemma is $C = \frac{1-\beta}{\alpha} \ge 4$, so threshold $a(C) > 0.3$).

To further characterize condition (37) 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, $r_k\mu_k \leq \frac{1}{4\Phi^{-1}(1-\alpha_k^*)}$, and the power for each hypothesis k is still upper bounded. Particularly, this upper bound differs by whether $\alpha_k \leq \alpha_k^*$ or $\alpha_k > \alpha_k^*$:

$$\begin{aligned} & 1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \\ \leq\ & 1-\Phi\left[\Phi^{-1}(1-\alpha_k^*)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \leq C\alpha_k^*, \quad \text{if } \alpha_k \leq \alpha_k^*; \\ & 1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \\ \leq\ & 1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k)}\right] \leq C\alpha_k, \quad \text{if } \alpha_k > \alpha_k^*, \end{aligned}$$

for k such that $\max\{\alpha_k, \alpha_k^*\} \leq a(C)$, and hence,

$$1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \leq C\max\{\alpha_k^*, \alpha_k\} \leq C(\alpha_k^* + \alpha_k).$$

Choose the constant $C = \frac{1-\beta}{2\alpha}$ (with $a(C) > 0.3$), and the overall power is upper bounded by $1-\beta$:

$$\sum_{k=1}^{\infty} 1-\Phi\left[\Phi^{-1}(1-\alpha_k)-\frac{1}{2\Phi^{-1}(1-\alpha_k^*)}\right] \leq \frac{1-\beta}{2\alpha}(2\alpha) = 1-\beta,$$

if (a) the significance levels are small: $\max\{\alpha_k, \alpha_k^*\} \leq 0.3$ for all $k = 1, 2, \ldots$, which holds since $\alpha \leq (1-\beta)/4 \leq 0.25$; and (b) the alternative mean $r_k\mu_k$ satisfies condition (37) for the baseline sequence, which holds when

$$r_k\mu_k \leq 0.25\left(\sqrt{2\log\left(\frac{k^2}{\alpha}\right)}\right)^{-1},$$

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 $\sum_{k=1}^{\infty} r_k \leq M < \infty$. 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 $L(k^*) := \{k \leq k^* : r_k = 1\}$ and $S(k^*) := \{k > k^* : r_k = 1\}$. The power would be less than $1-\beta$ if

$$\sum_{r_k=0} 1-\Phi[\Phi^{-1}(1-\alpha_k) - r_k\mu_k] \leq \alpha, \text{ and} \tag{39}$$

$$\sum_{k\in L(k^*)} 1-\Phi[\Phi^{-1}(1-\alpha_k)-r_k\mu_k] \le 2\alpha, \text{ and} \tag{40}$$

$$\sum_{k\in S(k^*)} 1-\Phi[\Phi^{-1}(1-\alpha_k)-r_k\mu_k] \le 1-\beta-3\alpha. \tag{41}$$

Power bound (39) for the nulls ($r_k = 0$) holds because individual power equals α_k and $\sum_{r_k=0}\alpha_k \le \alpha$. Power bound (40) 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

$$r_k\mu_k \le \Phi^{-1}(1-\alpha_k)-\Phi^{-1}(1-2\alpha_k).$$

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^*}:

$$r_k\mu_k \le \Phi^{-1}\left(1-\alpha_{k^*}\right)-\Phi^{-1}\left(1-2\alpha_{k^*}\right),$$

where $k^* = M^2/\alpha$. Notice that $\Phi^{-1}(1-x)$ is a convex function and its derivative is $-\left(\phi(\Phi^{-1}(1-x)\right)^{-1}$, so we have

$$\Phi^{-1}\left(1-\alpha_{k^*}\right)-\Phi^{-1}\left(1-2\alpha_{k^*}\right) \ge \left(\phi(\Phi^{-1}\left(1-2\alpha_{k^*}\right)\right)^{-1}\alpha_{k^*} \ge 0.4\sqrt{\alpha_{k^*}},$$

and power bound (40) for large α_k holds when $r_k\mu_k \le 0.4\sqrt{\alpha_{k^*}}$.

For small α_k, a sufficient condition for the power bound (41) is

$$1-\Phi[\Phi^{-1}(1-\alpha_k)-r_k\mu_k] \le \frac{1-\beta-3\alpha}{M},$$

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

$$r_k\mu_k \le \Phi^{-1}(1-\alpha_k)-\Phi^{-1}\left(1-\frac{1-\beta-3\alpha}{M}\right).$$

To characterize the rate of the above bound, recall that the sequence $\{\alpha_k\}_{k=1}^{\infty}$ decreases and sums to α, so $\alpha_k \le \alpha/k$ for any $k = 1, 2, \ldots$. Thus, the above condition on $r_k\mu_k$ holds when

$$r_k\mu_k \le \sqrt{\log\left(\frac{k}{4\alpha}\right)}-\sqrt{2\log\left(\frac{M}{2(1-\beta-3\alpha)}\right)},$$

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 $\frac{k}{4\alpha} > \frac{M^2}{4\alpha^2} \ge \frac{M^2}{4(1-\beta-3\alpha)^2}$, 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

$$\begin{aligned}\mathbb{P}(\exists k \in \mathbb{N} : Z_k \geq \Phi^{-1}(1-\alpha_k)) = &\ \mathbb{P}(\exists k \in \mathbb{N} : X_k \geq \Phi^{-1}(1-\alpha_k) - \mu_k) \\ = &\ 1 - \prod_{k=1}^{\infty} \Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right]. \qquad (42)\end{aligned}$$

Intuitively, the power would not converge to one when $\Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right] \gtrapprox (1-\alpha_k)$ (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 $\Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right] \ll 1 - \alpha_k$. To quantify this comparison, we consider the following ratio:

$$b_k := \frac{1 - \Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right]}{\alpha_k},$$

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:

$$\begin{aligned}\lim_{k\to\infty} (\log k)/b_k = &\ \lim_{k\to\infty} \frac{\alpha_k \log k}{1 - \Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right]} \\ = &\ \lim_{k\to\infty} \frac{\phi\left[\Phi^{-1}(1-\alpha_k)\right]}{\phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right]} \frac{\log k + \frac{\alpha_k}{k} \Big/ \frac{\partial \alpha_k}{\partial k}}{1 + \phi\left[\Phi^{-1}(1-\alpha_k)\right] \frac{\partial \mu_k}{\partial k} \Big/ \frac{\partial \alpha_k}{\partial k}},\end{aligned}$$

where for large k, we have $\Phi^{-1}(1-\alpha_k) \geq \sqrt{\log k}$ and

$$\begin{aligned}\frac{\phi\left[\Phi^{-1}(1-\alpha_k)\right]}{\phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right]} \leq &\ 2\exp\{-(\log k)^{1/2-1/c}\}; \\ \log k + \frac{\alpha_k}{k} \Big/ \frac{\partial \alpha_k}{\partial k} \leq &\ 2\log k; \\ 1 + \phi\left[\Phi^{-1}(1-\alpha_k)\right] \frac{\partial \mu_k}{\partial k} \Big/ \frac{\partial \alpha_k}{\partial k} \geq &\ 1.\end{aligned}$$

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 (42) is one if $\prod_{k=1}^{\infty} \Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right] = 0$, or equivalently,

$$\sum_{k=1}^{\infty} \log \Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right] = -\infty, \tag{43}$$

where for large k, we have

$$\begin{aligned} & \log \Phi\left[\Phi^{-1}(1-\alpha_k) - \mu_k\right] \\ = \; & \log(1 - b_k\alpha_k) \leq -b_k\alpha_k \\ \leq \; & -A \log k/[k(\log k)^2] = -A/(k \log k). \end{aligned}$$

Condition (43) 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. □

C.2. Proof of Theorem 7

Theorem 7 is a simplified version of the following Theorem 9 (by Claim 1). Before stating Theorem 9, 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 9. *The adaptively ordered martingale test with type-I error α and threshold c guarantees $1-\beta$ power if*

$$\begin{aligned} \exists k \in \mathbb{N} : & (2S(\mu, c) - 1)\left(N_1(k) - \frac{C_k^{\beta/3}\sqrt{N_1(k)}}{2\mathbb{P}(|Z(\mu)| > c)}\right) \\ & \geq \frac{C_k^{\alpha} + C_k^{\beta/3}}{\mathbb{P}^{1/2}(|Z(\mu)| > c)}\left[N_1(k) + D^{-1}(c)N_0(k) + \frac{C_k^{\beta/3}k^{1/2}}{2\mathbb{P}(|Z(\mu)| > c)}\right]^{1/2}, \end{aligned}$$

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

$$\begin{aligned} \exists k \in \mathbb{N} : \; & \sum_{i \in M_k} (r_i(2S_i(1) - 1) + (1 - r_i)(2S_i(0) - 1)) \\ & \geq \left(C_{|M_k|}^{\alpha} + C_{|M_k|}^{\beta}\right)(|M_k|)^{1/2}, \end{aligned} \tag{44}$$

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

$$\begin{aligned}\sum_{i \in M_k}(2S_i(1) - 1)r_i = &\ (2S(\mu, c) - 1)\sum_{i \in M_k} r_i \\ \geq &\ (2S(\mu, c) - 1)\left(\mathbb{P}(|Z(\mu)| > c)N_1(k) - \frac{C_k^\beta}{2}\sqrt{N_1(k)}\right),\end{aligned}$$

with probability $1 - \beta$ (Lemma 6). The condition in Theorem 9 results from plugging the bounds of both sides into condition (44).

Overall, when the condition in Theorem 9 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 (44) 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 (6) 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 9.*

Proof. First, the condition in Theorem 9 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:

$$\begin{aligned}2N_1(k) \geq\ & \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)}(1-D^{-1}(c))\\ &+\left\{\frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^4}{\widetilde{S}^2(\mu,c)}\left(1-D^{-1}(c)\right)^2 + 4\frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)}D^{-1}(c)k\right.\\ &\left.+\frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)}\frac{C_k^{\beta/3}}{2\mathbb{P}(|Z(\mu)|>c)}k^{1/2}\right\}^{1/2},\end{aligned}$$

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

$$4\frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)}D^{-1}(c)k + \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)}\frac{C_k^{\beta/3}}{2\mathbb{P}(|Z(\mu)|>c)}k^{1/2}$$

$$= \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)(\Phi(\mu-c)+\Phi(-\mu-c))}\left(8\Phi(-c)k + \frac{C_k^{\beta/3}}{2}k^{1/2}\right)$$
$$\leq \frac{\left(C_k^\alpha + C_k^{\beta/3}\right)^2}{\widetilde{S}(\mu,c)(\Phi(\mu-c)+\Phi(-\mu-c))}9\Phi(-c)k = \frac{9\left(C_k^\alpha + C_k^{\beta/3}\right)^2 D^{-1}(c)}{2\widetilde{S}(\mu,c)}k,$$

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. □

Appendix D: 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. [9, 10],

1. a linear bound

$$u_\alpha(k) = \sqrt{\frac{-\log\alpha}{2m}}k + \sqrt{\frac{-m\log\alpha}{2}}, \tag{45}$$

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

$$u_\alpha(k) = 1.7\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)}. \tag{46}$$

3. a curved bound from discrete mixture method

$$u_\alpha(k) = \inf\left\{s \in \mathcal{R} : \sum_{i=0}^{\infty} \omega_i \exp\{\lambda_i s - \psi(\lambda_i)k\} \geq 1/\alpha\right\}, \tag{47}$$

   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 $f(x) = 0.4\frac{\mathbf{1}_{0<x\leq\lambda_{\max}}}{x\log^{1.4}(e\lambda_{\max}/x)}$.
4. a curved bound from inverted stitching method (for finite time)

$$u_\alpha(k) = 2.42\sqrt{k\log\log(ek) + 4.7}, \quad k = 1, 2, \ldots, 10^4, \tag{48}$$

   where the time limit 10^4 is chosen as the number of hypotheses in the following simulation.

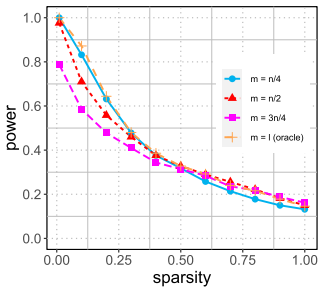
We use simulations to explore two choices in the martingale Stouffer test: (1) the choice of parameter m in the linear bound (45); and (2) the choice among the above four types of bound.

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.

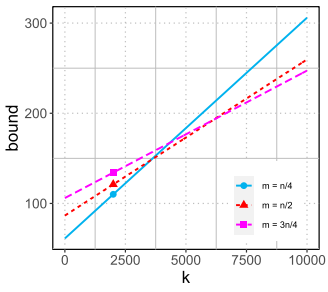
We seek for a robust value of m such that the resulting test has relatively high power under different non-null sparsity. The following constructed simulation is used for exploring bounds in both the martingale Stouffer test and the martingale Fisher test (introduced in Appendix E).

Setting 2. *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 12a).

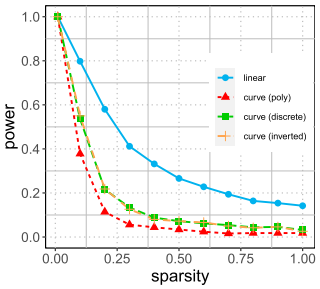


(a) Power of the martingale Stouffer test using the linear bound with different choices of parameter m.

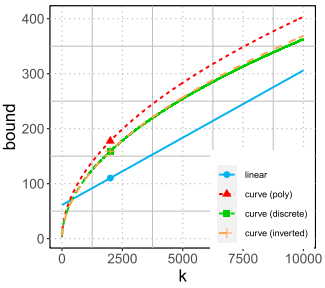


(b) Plot of the linear bound with different choices of parameter m.

FIG 12. *Testing martingale Stouffer test using linear bound* (45) *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 Stouffer test with varying non-null sparsity.



(b) Plot of four bounds. The linear bound is much tighter than the curved bounds for most $k \leq 10^4$.

FIG 13. *Comparison of the aforementioned four bounds* (45)-(48) *for the martingale Stouffer test.*

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 13b).

Under the batch setting where the number of hypotheses n is finite, we use the simulation setting 2, and the linear bound (45) (with $m = n/4$) results in the highest power (Figure 13a). 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.

Appendix E: Martingale Fisher test

The batch test by Fisher [7] 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. [9, 10]. 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

$$\exists k \in \mathbb{N} : -2\sum_{i=1}^{k} \log p_i - 2k \geq u_\alpha(k), \tag{49}$$

where examples of $u_\alpha(k)$ include

1. a sub-exponential linear boundary

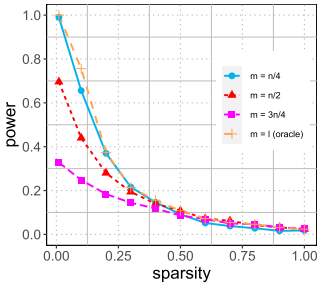
$$u_\alpha(k) = \left(\left(\frac{1.41m}{x_{m,\alpha}} + 2\right)\log\left(1 + \frac{1.41x_{m,\alpha}}{m}\right) - 2\right)(k-m) + 2.82x_{m,\alpha}, \tag{50}$$

   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

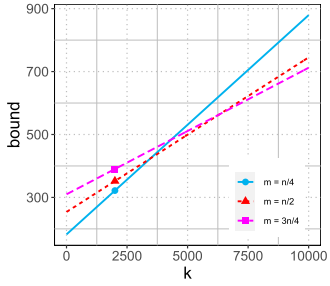
$$u_\alpha(k) = 4.07\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)} + 9.66\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right). \tag{51}$$

The linear bound contains a parameter m with the same interpretation as m in the linear bound (5) 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 14a, 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 2). The curve bound loses power quickly when non-null is rather sparse (see Figure 15a), consistent with the comparison between linear and curved bounds for the martingale Stouffer test in Appendix D.

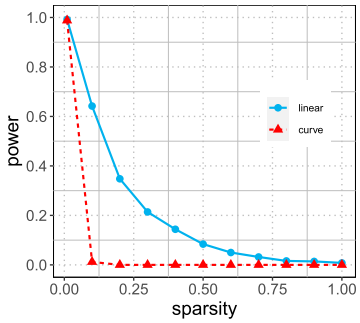


(a) Power of martingale Fisher test using the linear bound with different choices of parameter m.

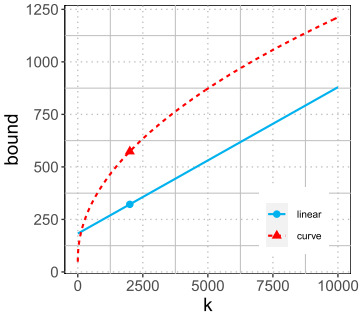


(b) Plot of the linear bound with different choices of parameter m.

FIG 14. *Testing the martingale Fisher test using the linear bound* (50) *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 $k \leq 10^4$.

FIG 15. *Comparison of the aforementioned two bounds* (50) *and* (51) *for the martingale Fisher test.*

Appendix F: 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 D and E), there are several linear and curved boundaries $u_\alpha(k)$ for chi-square increment martingales from the work of Howard et al. [9, 10]. 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

$$\exists k \in \mathbb{N} : \sum_{i=1}^{k} \left[\Phi^{-1}(1-p_i)\right]^2 - k \geq u_\alpha(k), \tag{52}$$

where examples of $u_\alpha(k)$ include

1. a sub-exponential linear boundary

$$u_\alpha(k) = \left(\left(\frac{m}{2x_{m,\alpha}} + 1\right)\log\left(1 + \frac{2x_{m,\alpha}}{m}\right) - 1\right)(k-m) + 2x_{m,\alpha}, \tag{53}$$

   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

$$\begin{aligned} u_\alpha(k) = \; & 3.42\sqrt{k\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right)} \\ & + 9.66\left(\log\log(2k) + 0.72\log\frac{5.2}{\alpha}\right). \end{aligned} \tag{54}$$

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 D) and the martingale Fisher test (Appendix E). If testing the martingale chi-squared test by the same numerical experiment in Setting 2, $m = n/4$ 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.

Appendix G: 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:

$$\pi_i = \pi_\beta(x_i) = \frac{1}{1+\exp(-\beta\phi(x_i))}, \tag{55}$$

where $\phi(x_i)$ is a spline basis. The hierarchical structure is imposed by a partial ordering constraint on π_i:

$$\pi_i \geq \pi_j, \quad \text{if } i \text{ is the parent of } j, \tag{56}$$

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

$$\begin{aligned} a_{i,\text{new}} = \; & \mathbb{E}[w_i q_i \mid \widetilde{Z_i}] = \frac{\pi_i \phi(\widetilde{Z_i} - \mu)}{L\left(\widetilde{Z_i}, \mu, \pi_i\right)}; \\ b_{i,\text{new}} = \; & \mathbb{E}[w_i(1-q_i) \mid \widetilde{Z_i}] = \frac{(1-\pi_i)\phi(\widetilde{Z_i})}{L\left(\widetilde{Z_i}, \mu, \pi_i\right)}; \end{aligned}$$

$$c_{i,\text{new}} = \mathbb{E}[(1-w_i)q_i \mid \widetilde{Z_i}] = \frac{\pi_i \phi(-\widetilde{Z_i} - \mu)}{L\left(\widetilde{Z_i}, \mu, \pi_i\right)};$$

$$d_{i,\text{new}} = \mathbb{E}[(1-w_i)(1-q_i) \mid \widetilde{Z_i}] = \frac{(1-\pi_i)\phi(-\widetilde{Z_i})}{L\left(\widetilde{Z_i}, \mu, \pi_i\right)}.$$

If the p-value is unmasked for hypothesis i, we have $w_i = 1$ and the updates are

$$a_{i,\text{new}} = \left(1 + \frac{(1-\pi_i)\phi\left(\widetilde{Z_i}\right)}{\pi_i \phi\left(\widetilde{Z_i} - \mu\right)}\right)^{-1};$$

$$b_{i,\text{new}} = 1 - a_{i,\text{new}}; \quad c_{i,\text{new}} = 0; \quad d_{i,\text{new}} = 0.$$

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 (55) 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 [3], Theorem 3.1 and Robertson [21], Theorem 1.5.1):

$$\begin{aligned}(\pi_{i,\text{new}}) &= \operatorname*{argmax}_{\text{partial ordered}\{\pi_i\}} \sum_i (a_i + c_i)\log \pi_i + (1 - a_i - c_i)\log(1 - \pi_i) \\ &= \operatorname*{argmin}_{\text{partial ordered}\{\pi_i\}} \sum_i (a_i + c_i - \pi_i)^2,\end{aligned}$$

where the partial ordering is defined in statement (56).

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:

$$(\gamma_{\text{new}}) = \operatorname*{argmax}_{\gamma} \sum_i a_i \left(\widetilde{Z_i} - \mu_\gamma(x_i)\right)^2 + c_i \left(-\widetilde{Z_i} - \mu_\gamma(x_i)\right)^2,$$

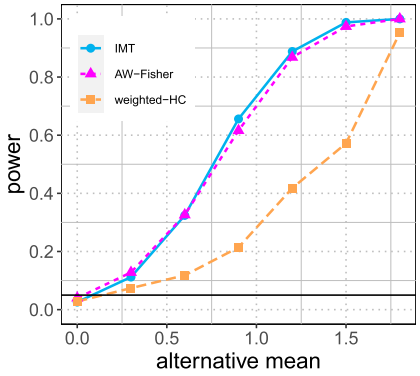


FIG 16. *Power of the interactively ordered martingale test (IMT), AW-Fisher, and weighted-HC when the non-null cluster is in the center of a* 10×10 *grid. IMT and AW-Fisher both have high power, but the AW-Fisher has a high computational cost.*

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.

Appendix H: 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) [11], 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 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 5, above in Figure 16, 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.

Acknowledgements

We thank the anonymous reviewers for their helpful suggestions. AR acknowledges support from NSF DMS 1916320, and NSF CAREER 1945266. SB acknowledges support from NSF DMS 1713003, and CCF 1763734. LW acknowledges support from NSF DMS 1713003. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) [28], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system [17], which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

References

[1] Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electronic Journal of Statistics* **11** 1983–2001. MR3651021

[2] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085. MR3375876

[3] Barlow, R. E. and Brunk, H. D. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association* **67** 140–147. MR0314205

[4] Donoho, D. and Jin, J. (2015). Special Invited Paper: Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects. *Statistical Science* 1–25. MR3317751

[5] Duan, B., Ramdas, A. and Wasserman, L. (2020). Familywise Error Rate Control by Interactive Unmasking. In *International Conference on Machine Learning (accepted)*.

[6] Fang, Y., Tang, S., Huo, Z., Tseng, G. C. and Park, Y. (2019). Properties of adaptively weighted Fisher's method. *arXiv preprint arXiv:1908.00583*.

[7] Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in Statistics* 66–70. Springer. MR0346954

[8] Grünwald, P., de Heide, R. and Koolen, W. (2019). Safe testing. *arXiv preprint arXiv:1906.07801*.

[9] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys* **17** 257–317. MR4100718

[10] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics (accepted)*.

[11] Huo, Z., Tang, S., Park, Y. and Tseng, G. (2020). P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. *Bioinformatics* **36** 524–532.

[12] Ignatiadis, N., Klaus, B., Zaugg, J. B. and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* **13** 577.
[13] Kost, J. T. and McDermott, M. P. (2002). Combining dependent P-values. *Statistics & Probability Letters* **60** 183–190. MR1945440
[14] Lei, L. and Fithian, W. (2018). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 649–679. MR3849338
[15] Lei, L., Ramdas, A. and Fithian, W. (2020). STAR: A general interactive framework for FDR control under structural constraints. *Biometrika (accepted)*.
[16] Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5** 994–1019. MR2840184
[17] Nystrom, N. A., Levine, M. J., Roskies, R. Z. and Scott, J. R. (2015). Bridges: a uniquely flexible HPC resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure* 1–8.
[18] Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *The Annals of Statistics* **37** 3867–3892. MR2572446
[19] Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.
[20] Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* **41** 1397–1409. MR0277063
[21] Robertson, T., Wright, F. and Dykstra, R. (1988). Order restricted statistical inference. MR0961262
[22] Rüger, B. (1978). Das maximale Signifikanzniveau des Tests: "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika* **25** 171–178. MR0526476
[23] Rüschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability* **14** 623–632. MR0665297
[24] Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science* **26** 84–101. MR2849911
[25] Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *The Annals of Statistics* 361–404. MR0840504
[26] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872
[27] Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. and Williams Jr, R. M. (1949). The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.
[28] Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R. and Wilkins-Diehr, N. (2014).

XSEDE: Accelerating Scientific Discovery. *Computing in science & engineering* **16** 62–74.

[29] Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*. asaa027. MR4186488

[30] Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* **16** 117–186. MR0013275

[31] Zhang, M., Gelfman, S., McCarthy, J., Harms, M. B., Moreno, C. A., Goldstein, D. B. and Allen, A. S. (2020). Incorporating external information to improve sparse signal detection in rare-variant gene-set-based analyses. *Genetic Epidemiology* **44** 330–338.