Familywise Error Rate Control by Interactive Unmasking

Boyan Duan [1] **Aaditya Ramdas** [1 2] **Larry Wasserman** [1 2]

Abstract

We propose a method for multiple hypothesis testing with familywise error rate (FWER) control, called the i-FWER test. Most testing methods are predefined algorithms that do not allow modifications after observing the data. However, in practice, analysts tend to choose a promising algorithm after observing the data; unfortunately, this violates the validity of the conclusion. The i-FWER test allows much flexibility: a human (or a computer program acting on the human's behalf) may adaptively guide the algorithm in a data-dependent manner. We prove that our test controls FWER if the analysts adhere to a particular protocol of masking and unmasking. We demonstrate via numerical experiments the power of our test under structured non-nulls, and then explore new forms of masking.

1. Introduction

Hypothesis testing is a critical instrument in scientific research to quantify the significance of a discovery. For example, suppose an observation $Z \in \mathbb{R}$ follows a Gaussian distribution with mean μ and unit variance. We wish to distinguish between the following null and alternative hypotheses regarding the mean value:

$$H_0 : \mu \leq 0 \quad \text{versus} \quad H_1 : \mu > 0. \tag{1}$$

A test decides whether to reject the null hypothesis, usually by calculating a *p-value*: the probability of observing an outcome at least as extreme as the observed data under the null hypothesis. In the above example, the p-value is $P = 1 - \Phi(Z)$, where Φ is the cumulative distribution function (CDF) of a standard Gaussian. When the true mean μ is exactly zero, the p-value is uniformly distributed; when $\mu < 0$, it has nondecreasing density. A low p-value suggests evidence to reject the null hypothesis.

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

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

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

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

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

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

[1]Department of Statistics and Data Science, Carnegie Mellon University [2]Machine Learning Department, Carnegie Mellon University. Correspondence to: Boyan Duan <boyand@stat.cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

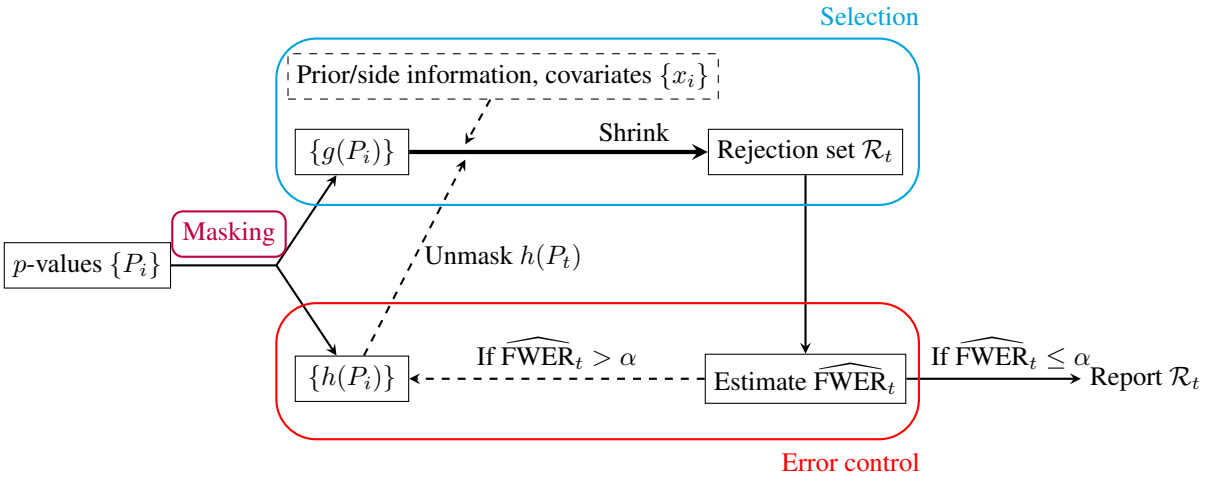


Figure 1. A schematic of the i-FWER test. All p-values are initially 'masked': all $\{g(P_i)\}$ are revealed to the analyst/algorithm, while all $\{h(P_i)\}$ remain hidden, and the initial rejection set is $\mathcal{R}_0 = [n]$. If $\widehat{\text{FWER}}_t > \alpha$, the analyst chooses a p-value to 'unmask' (observe the masked $h(P)$-value), effectively removing it from the proposed rejection set $\mathcal{R}_t$; importantly, using any available side information and/or covariates and/or working model, the analyst can shrink $\mathcal{R}_t$ in any manner. This process continues until $\widehat{\text{FWER}}_t \leq \alpha$ (or $\mathcal{R}_t = \emptyset$).

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

The key idea that permits interaction while ensuring FWER control is "masking and unmasking", proposed by Lei & Fithian (2018); Lei et al. (2020). In our method, it has three main steps and alternates between the last two (Figure 1):

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

$$h(P_i; p_*) = 2 \cdot \mathbb{1}\{P_i < p_*\} - 1;$$
$$\text{and } g(P_i; p_*) = \min\left\{P_i, \frac{p_*}{1 - p_*}(1 - P_i)\right\}, \quad (2)$$

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

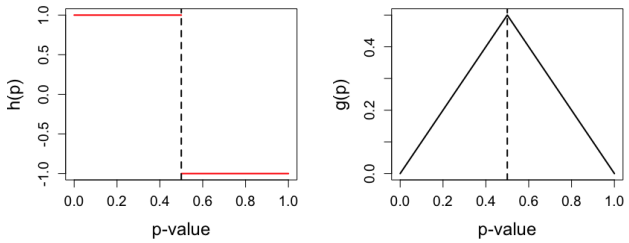


Figure 2. Functions for masking (2): missing bits h (left) and masked p-values g (right) when $p_* = 0.5$. For uniform p-values, $g(P)$ and $h(P)$ are independent.

2. **Selection**. Consider a set of candidate hypotheses to be rejected (rejection set), denoted as $\mathcal{R}_t$ for iteration t. We start with all the hypotheses included, $\mathcal{R}_0 = [n]$. At each iteration, the analyst excludes possible nulls from the previous $\mathcal{R}_{t-1}$, using all the available information (masked p-values, progressively unmasked $h(P_i)$ from step 3 and possible prior information). Note that our method does not automatically use prior information and masked p-values. The analyst is free to use any black-box prediction algorithm or Bayesian working model that uses the available information, and orders the hypotheses possibly using an estimated likelihood of being non-null. This step is where a human is allowed to incorporate their subjective choices.

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

The rest of the paper is organized as follows. In Section 2, we describe the i-FWER test in detail. In Section 3, we implement the interactive test under a clustered non-null structure. In Section 4, we propose two alternative ways of masking p-values and explore their advantages.

2. An interactive test with FWER control

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

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

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

where recall $[n]$ denotes the set of all the hypotheses. We assume without loss of generality that one hypothesis is excluded in each step. Denote the hypothesis excluded at step t as i_t^*. The choice of i_t^* can use the information available to the analyst before step t, formally defined as a filtration (sequence of nested σ-fields) [1]:

$$\mathcal{F}_{t-1} := \sigma\Big(\{x_i, g(P_i)\}_{i=1}^n, \{P_i\}_{i\notin\mathcal{R}_{t-1}}\Big), \qquad (3)$$

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

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

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

recall that p_* is the prespecified parameter for masking (2). Instead of rejecting every hypothesis in $\mathcal{R}_t$, note that the test only rejects the ones in $\mathcal{R}_t^+$, whose p-values are smaller than p_* in $\mathcal{R}_t$. Thus, the number of false rejection V is $|\mathcal{H}_0 \cap \mathcal{R}_t^+|$ and we want to control FWER, $\mathbb{P}(V \geq 1)$. The distribution of $|\mathcal{H}_0 \cap \mathcal{R}_t^+|$ can be estimated by $|\mathcal{H}_0 \cap \mathcal{R}_t^-|$ using the fact that $h(P_i)$ is a (biased) coin flip. But $\mathcal{H}_0$ (the set of true nulls) is unknown, so we use $|\mathcal{R}_t^-|$ to upper bound $|\mathcal{H}_0 \cap \mathcal{R}_t^-|$, and propose an estimator of FWER:

$$\widehat{\text{FWER}}_t = 1 - (1-p_*)^{|\mathcal{R}_t^-|+1}. \qquad (4)$$

Overall, the i-FWER test shrinks $\mathcal{R}_t$ until $\widehat{\text{FWER}}_t \leq \alpha$ and rejects only the hypotheses in $\mathcal{R}_t^+$ (Algorithm 1).

Algorithm 1 The i-FWER test

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

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

Remark 2. *The above procedure can be easily extended to control k-FWER:*

$$k\text{-}FWER := \mathbb{P}(V \geq k), \qquad (5)$$

by estimating k-FWER as

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

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

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

Theorem 1. *Suppose the null p-values are mutually independent and they are independent of the non-nulls, then the i-FWER test controls FWER at level α.*

Remark 3. *The null p-values need not be exactly uniformly distributed. For example, FWER control also holds when the null p-values have a convex CDF or nondecreasing probability mass function (for discrete p-values or the density function otherwise). Appendix 1 presents the detailed technical condition for the distribution of the null p-values.*

Related work. The i-FWER test mainly combines and generalizes two sets of work: (a) we use the idea of masking from Lei & Fithian (2018); Lei et al. (2020) and extend it to a more stringent error metric, FWER; (b) we use the method of controlling FWER from Janson & Su (2016) by converting a one-step procedure in the context of "knockoff" statistics in regression problem to a multi-step (interactive) procedure in our context of p-values.

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

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

the expected proportion of false discoveries. It is less stringent than FWER, the probability of making *any* false discovery. Their method uses the special case of masking (2) when $p_* = 0.5$, and estimate V by $\sum_{i\in\mathcal{R}_t} \mathbb{1}\{h(P_i) = -1\}$, or equivalently $\sum_{i\in\mathcal{R}_t} \mathbb{1}\{P_i < 0.5\}$. While it provides a good estimation on the *proportion* of false discoveries, the indicator $\mathbb{1}\{P_i < 0.5\}$ has little information on the correctness of *individual* rejections. To see this, suppose there is one rejection, then FWER is the probability of this rejection being false. Even if $h(P_i) = 1$, which indicates the p-value is on the smaller side, the tightest upper bound on FWER is as high as 0.5. Thus, our method uses masking (2) with small p_*, so that $h(P_i) = 1$, or equivalently $P_i < p_*$, suggests a low chance of false rejection.

In the context of a regression problem to select significant covariates, Janson & Su (2016) proposes a one-step method with control on k-FWER; recall definition in (5). The FWER is a special case of k-FWER when $k = 1$, and as k grows larger, k-FWER is a less stringent error metric. Their method decomposes statistics called "knockoff" (Barber & Candès, 2015) into the magnitudes for ordering covariates (without interaction) and signs for estimating k-FWER, which corresponds to decomposing p-values into $g(P_i)$ and $h(P_i)$ when $p_* = 0.5$. However, the decomposition as magnitude and sign restricts the corresponding p-value decomposition with a single choice of p_* as 0.5, making the k-FWER control conservative and power low when $k = 1$; yet our method shows high power in experiments. Their error control uses the connection between k-FWER and a negative binomial distribution, based on which we propose the estimator $\widehat{\text{FWER}}_t$ for our multi-step procedure, and prove the error control even when interaction is allowed. As far as we know, this estimator viewpoint of the FWER procedure is also new in the literature.

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

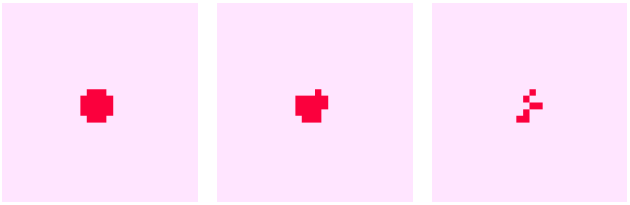
The i-FWER test in practice. Technically in a fully interactive procedure, a human can examine all the information in $\mathcal{F}_{t-1}$ and pick i_t^* subjectively or by any other principle, but doing so for every step could be tedious and unnecessary. Instead, the analyst can design an automated version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

3. An instantiation of an automated algorithm, and numerical experiments

One main advantage of the i-FWER test is the flexibility to include prior knowledge and human guidance. The analyst might have an intuition about what structural constraints the non-nulls have. For example, we consider two structures: (a) a grid of hypotheses where the non-nulls are in a cluster (of some size/shape, at some location; see Figure 3a), which is a reasonable prior belief when one wants to identify a tumor in a brain image; and (b) a tree of hypotheses where a child can be non-null only if its parent is non-null, as may be the case in applications involving wavelet decompositions.

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

We propose an automated algorithm of the i-FWER test that incorporates the structure of clustered non-nulls. Here, the side information x_i is the coordinates of each hypothesis i. The idea is that at each step of excluding possible nulls,



(a) True non-nulls (21 hypotheses). (b) 18 rejections by the i-FWER test. (c) 7 rejections by the Šidák correction.

Figure 3. An instance of rejections by the i-FWER test and the Šidák correction (Šidák, 1967). Clustered non-nulls are simulated from the setting in Section 3.2 with a fixed alternative mean $\mu = 3$.

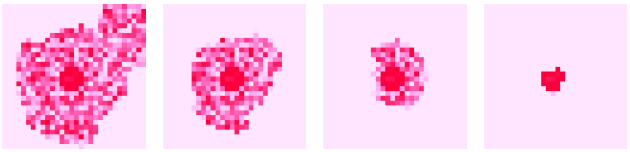


Figure 4. An illustration of $\mathcal{R}_t$ generated by the automated algorithm described in Section 3.1, at $t = 50, 100, 150$ and $t = 220$ when the algorithm stops. The p-values in $\mathcal{R}_t$ are plotted.

we peel off the boundary of the current $\mathcal{R}_t$, such that the rejection set stays connected (see Figure 4).

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

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

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

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

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

$$P_i \overset{d}{=} (1 - \pi_i)F_0 + \pi_i F_1. \tag{6}$$

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

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

1. A different algorithm can be developed for a different structure. For example, when hypotheses have a hierarchical structure and the non-nulls only appear on a subtree, an algorithm can gradually cut branches.

2. The score S_i for non-null likelihood is not exclusive for the above algorithm – it can be used in any heuristics such as directly ordering hypotheses by S_i.

3. Human interaction can help the automated procedure: the analyst can stop and modify the automated algorithm at any iteration. It is a common case where prior knowledge might not be accurate, or there exist several plausible structures. The analyst may try different algorithms and improve their understanding of the data as the test proceeds. In the example of clustered non-nulls, the underlying truth might have two clustered non-nulls instead of one. After several iterations of the above algorithm that is designed for a single cluster, the shape of $\mathcal{R}_t$ could look like a dumbbell, so the analyst can split $\mathcal{R}_t$ into two subsets if they wish.

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

Next, we demonstrate via experiments that the i-FWER test can improve power over the Šidák correction, a baseline method that does not take side information into account[2]. We chose a clustered non-null structure for visualization and intuition, though our test can utilize any covariates, structural constraints, domain knowledge, etc.

3.2. Numerical experiments for clustered non-nulls

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

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

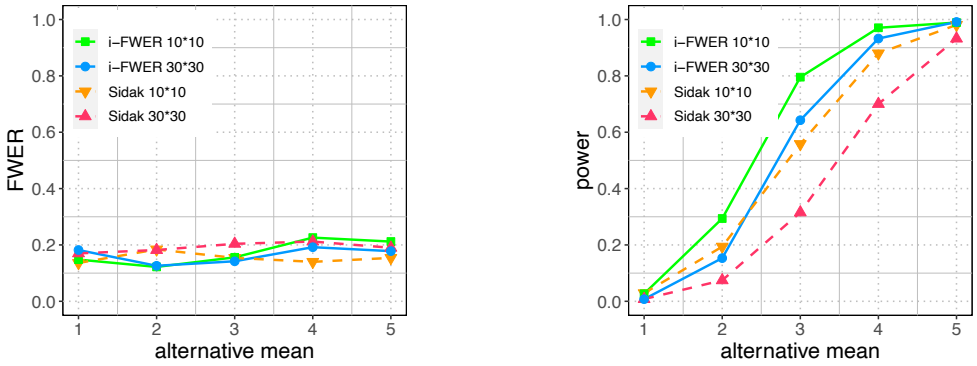


Figure 5. The i-FWER test versus Šidák for clustered non-nulls. The experiments are described in Section 3.2 where we tried two sizes of hypotheses grid: 10×10 and 30×30 (the latter is a harder problem since the number of nulls increases while the number of non-nulls remains fixed). Both methods show valid FWER control (left). The i-FWER test has higher power under both grid sizes (right).

Setting. *Consider 900 hypotheses arranged in a* 30×30 *grid with a disc of 21 non-nulls. Each hypothesis tests the mean value of a univariate Gaussian as described in* (1)*. The true nulls are generated from* $N(0, 1)$ *and non-nulls from* $N(\mu, 1)$*, where we varied* μ *as* $(1, 2, 3, 4, 5)$*. For all experiments in the paper, the FWER control is set at level* $\alpha = 0.2$*, and the power is averaged over 500 repetitions*[3].

The i-FWER test has higher power than the Šidák correction, which does not use the non-null structure (see Figure 5). It is hard for most existing methods to incorporate the knowledge that non-nulls are clustered without knowing the position or the size of this cluster. By contrast, such information can be learned in the i-FWER test by looking at the masked p-values and the progressively unmasked missing bits. This advantage of the i-FWER test is more evident as the number of nulls increases (by increasing the grid size from 10×10 to 30×30 with the number of non-nulls fixed). Note that the power of both methods decreases, but the i-FWER test seems less sensitive. This robustness to nulls is expected as the i-FWER test excludes most nulls before rejection, whereas the Šidák correction treats all hypotheses equally.

3.3. An example of an automated algorithm under a hierarchical structure of hypotheses

When the hypotheses form a tree, the side information x_i encodes the parent-child relationship (the set of indices of the children nodes for each hypothesis i). Suppose we have prior knowledge that a node cannot be non-null if its parent is null, meaning that the non-nulls form a subtree with the same root. We now develop an automated algorithm that prunes possible nulls among the leaf nodes of current $\mathcal{R}_t$, such that the rejection set has such a subtree shape. Like the algorithm for clustered non-nulls, we use a score S_i to

[3]The standard error of FWER and averaged power are less than 0.02, thus ignored from the plots in this paper.

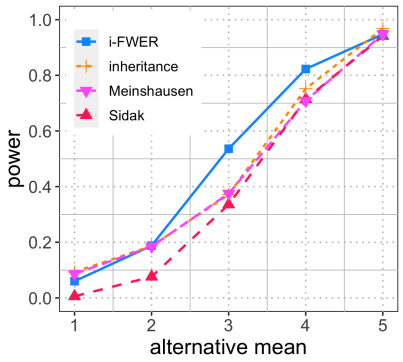


Figure 6. Power of the i-FWER test under a tree structure when varying the alternative mean value. It has higher power than inheritance procedure, Meinshausens method, and the Sidak correction.

choose which leaf nodes to exclude. For example, the score S_i can be the estimated non-null likelihood learned from model (6), where we account for the hierarchical structure by further assuming a partial order constraint on π_i that $\pi_i \geq \pi_j$ if $j \in x_i$ (i.e., i is the parent of j).

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

In addition to the Šidák correction, we compare the i-FWER test with two other methods for tree-structured hypotheses: Meinshausens method (Meinshausen, 2008) and the inheritance procedure (Goeman & Finos, 2012), which work under arbitrary dependence. Their idea is to pass the error budget from a parent node to its children in a prefixed manner, whereas our algorithm picks out the subtree with

non-nulls based on the observed data. In our experiments, the i-FWER test has the highest power (see Figure 6).

The above results demonstrate the power of the i-FWER test in one particular form where the masking is defined as (2). However, any two functions that decompose the null p-values into two independent parts can, in fact, be used for masking and fit into the framework of the i-FWER test (see the proofs of error control when using the following new masking functions in Appendix 6). In the next section, we explore several choices of masking.

4. New masking functions

Recall that masking is the key idea that permits interaction and controls error at the same time, by decomposing the p-values into two parts: masked p-value $g(P)$ and missing bits $h(P)$. Such splitting distributes the p-value information for two different purposes, interaction and error control, leading to a tradeoff. More information in $g(P)$ provides better guidance on how to shrink $\mathcal{R}_t$ and improves the power, while more information in $h(P)$ enhances the accuracy of estimating FWER and makes the test less conservative. This section explores several ways of masking and their influence on the power of the i-FWER test. To distinguish different masking functions, we refer to masking (2) introduced at the very beginning as the "tent" function based on the shape of map g (see Figure 7a).

4.1. The "railway" function

We start with an adjustment to the tent function that flips the map g when $p > p_*$, which we call the "railway" function (see Figure 7b). It does not change the information distribution between $g(P)$ and $h(P)$, and yet improves the power when nulls are conservative, as demonstrated later. Conservative nulls are often discussed under a general form of hypotheses testing for a parameter θ:

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

where Θ_0 and Θ_1 are two disjoint sets. Conservative nulls are those whose true parameter θ lies in the interior of Θ_0. For example, when testing whether a Gaussian $N(\mu, 1)$ has nonnegative mean in (1) where $\Theta_0 = \{\mu \le 0\}$, the nulls are conservative when $\mu < 0$. The resulting p-values are biased toward larger values, which compared to the uniform p-values from nonconservative nulls should be easier to distinguish from that of non-nulls. However, most classical methods do not take advantage of it, but the i-FWER test can, when using the railway function for masking:

$$h(P_i) = 2 \cdot \mathbb{1}\{P_i < p_*\} - 1;$$
$$\text{and } g(P_i) = \begin{cases} P_i, & 0 \le P_i < p_*, \\ \frac{p_*}{1-p_*}(P_i - p_*), & p_* \le P_i \le 1. \end{cases} \quad (7)$$

The above masked p-value, compared with the tent masking (2), can better distinguish the non-nulls from the conservative nulls. To see this, consider a p-value of 0.99. When $p_* = 0.2$, the masked p-value generated by the originally proposed tent function would be 0.0025, thus causing potential confusion with a non-null, whose masked p-value is also small. But the masked p-value from the railway function would be 0.1975, which is close to 0.2, the upper bound of $g(P_i)$. Thus, it can easily be excluded by our algorithm.

We follow the setting in Section 3.2 for simulation , except that the alternative mean is fixed as $\mu = 3$, and the nulls are simulated from $N(\mu_0, 1)$, where the mean value μ_0 is negative so that the resulting null p-values are conservative. We tried μ_0 as $(0, -1, -2, -3, -4)$, with a smaller value indicating higher conservativeness, in the sense that the p-values are more likely to be biased to a larger value. When the null is not conservative ($\mu_0 = 0$), the i-FWER test with the railway function and tent function have similar power. As the conservativeness of nulls increases, while the power of the i-FWER test with the tent function decreases and the Šidák correction stays the same, the power of the i-FWER test with the railway function increases (see Figure 8).

4.2. The "gap" function

Another form of masking we consider maps only the p-values that are close to 0 or 1, which is referred to as the "gap" function (see Figure 7c) . The resulting i-FWER test directly unmasks all the p-values in the middle, and as a price, never rejects the corresponding hypotheses. Given two parameters p_l and p_u, the gap function is defined as

$$h(P_i) = \begin{cases} 1, & 0 \le P_i < p_l, \\ -1, & p_u < P_i \le 1; \end{cases}$$
$$\text{and } g(P_i) = \begin{cases} P_i, & 0 \le P_i < p_l, \\ \frac{p_l}{1-p_u}(1 - P_i), & p_u < P_i \le 1. \end{cases} \quad (8)$$

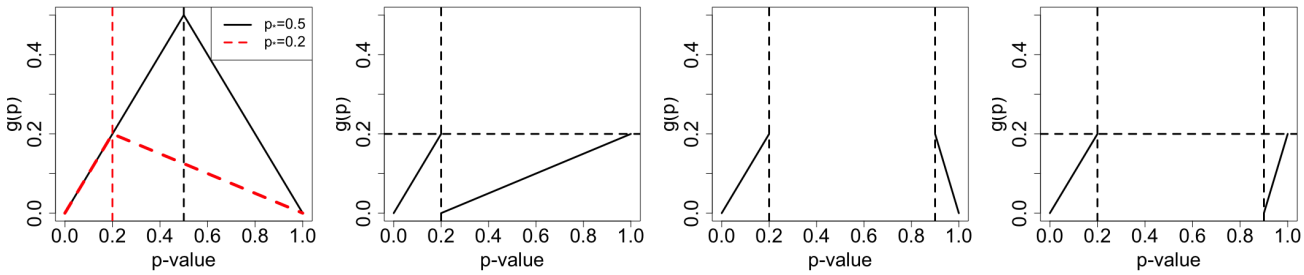
All the p-values in $[p_l, p_u]$ are available to the analyst from the beginning. Specifically, let $\mathcal{M} = \{i : p_l < P_i < p_u\}$ be the set of skipped p-values in the masking step, then the available information at step t for shrinking $\mathcal{R}_{t-1}$ is

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

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

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

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



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

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

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

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

Figure 7. Different masking functions leaves different amount of information to $g(P)$ (and the complement part to $h(P)$).

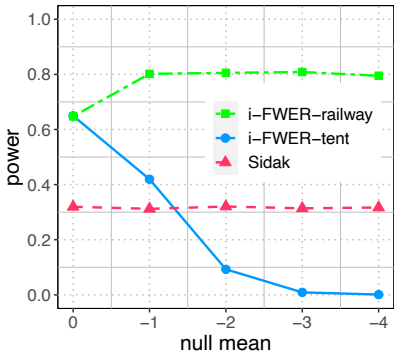


Figure 8. Power of the i-FWER test with the tent function and the railway function, where the nulls become more conservative as the null mean decreases in $(0, -1, -2, -3, -4)$. The i-FWER test benefits from conservative null when using the railway function.

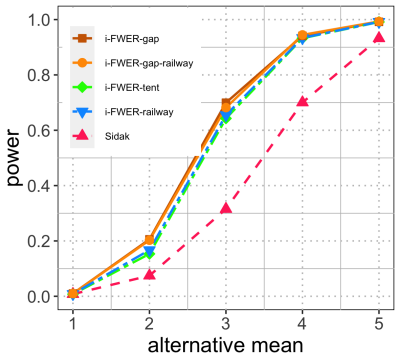


Figure 9. Power of the i-FWER test with the tent function ($p_* = 0.1$) and the gap function ($p_l = 0.1, p_u = 0.5$). The gap function leads to slight improvement in power. Simulation follows the setting in Section 3.2.

changes to

$$\widehat{\text{FWER}}_t = 1 - \left(1 - \frac{p_l}{p_l + 1 - p_u}\right)^{|\mathcal{R}_t^-|+1}. \quad (9)$$

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

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

4.3. The "gap-railway" function

Combining the idea of the gap and railway functions, we develop the "gap-railway" function such that the middle p-values are directly unmasked and the map g for large p-values is an increasing function (see Figure 7d). Given parameters p_l and p_u, the gap-railway function is defined as

$$h(P_i) = \begin{cases} 1, & 0 \le P_i < p_l, \\ -1, & p_u < P_i \le 1; \end{cases}$$
$$\text{and } g(P_i) = \begin{cases} P_i, & 0 \le P_i < p_l, \\ \frac{p_l}{1-p_u}(P_i - p_u), & p_u < P_i \le 1. \end{cases} \quad (10)$$

Comparing with the tent function with $p_* = p_l$, the i-FWER test using the gap function additionally uses the entire p-values in $[p_l, p_u]$ for interaction, which leads to an increased power (see Figure 9). The same pattern is maintained when we flip the mappings for large p-values, shown in the comparison of the railway function and the gap-railway function[4]. This improvement also motivates why the i-FWER test progressively unmasks $h(P_i)$, in other words, to reveal

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

as much information to the analyst as allowed at the current step. Unmasking the p-values even for the hypotheses outside of the rejection set can improve the power, because they help the joint modeling of all the p-values, especially when there is some non-null structure.

To summarize, we have presented four types of masking functions: tent, railway, gap, gap-railway (see Figure 7). Compared to the tent (gap) function, the railway (gap-railway) functions are more robust to conservative nulls. Compared with the tent (railway) function, the gap (gap-railway) function reveals more information to guide the shrinkage of $\mathcal{R}_t$. Note however that the railway or gap function is not *always* better than the tent function. We may favor the tent function over the railway function when there are less p-values close to one, and we may favor the tent function over the gap function when there is considerable prior knowledge to guide the shrinkage of $\mathcal{R}_t$.

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

5. A prototypical application to genetic data

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

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

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

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

In hindsight, a small value for the masking parameter was more powerful in this dataset because over 1600 p-values are extremely small ($< 10^{-5}$), and these are highly likely to be the non-nulls. Thus, even when the masked p-values for all hypotheses are in a small range, such as $(0, 0.01)$ when $\alpha = 0.1$ and $p_* = \alpha/10$, the p-values from the non-nulls still stand out because they gather below 10^{-5}. At the same time, the smaller the p_*, the more accurate (less conservative) is our estimate of FWER in (4); the algorithm can stop shrinking $\mathcal{R}_t$ earlier since more hypotheses with negative $h(P)$ are allowed to be included in the final $\mathcal{R}_t$. In practice, the choice of masking parameter can be guided by the prior belief of the strength of non-null signals: if the non-nulls have strong signal and hence extremely small p-values (such as the mean value $\mu \geq 5$ when testing if a univariate Gaussian has zero mean), a small masking parameter is preferred; otherwise, we recommend $\alpha/2$ to leave more information for interactively shrinking the rejection set $\mathcal{R}_t$.

6. Discussion

We proposed a multiple testing method with a valid FWER control while granting the analyst freedom of interacting with the revealed data. The masking function must be fixed in advance, but during the procedure of excluding possible nulls, the analyst can employ any model, heuristic, intuition, or domain knowledge, tailoring the algorithm to various applications. Although the validity requires an independence assumption, our method is a step forward to fulfilling the practical needs of allowing interactive human guidance to automated large-scale testing using ML in the sciences.

The critical idea that guarantees the FWER control is "masking and unmasking". A series of interactive tests are developed following the idea of masking: Lei & Fithian (2018) and Lei et al. (2020) proposed the masking idea and an interactive test with FDR control; Duan et al. (2019) developed an interactive test for the global null; this work presents an interactive test with FWER control. At a high level, masking-based interactive testing achieves rigorous conclusions in an exploratory framework, giving this broad technique much appeal and potential.

Code and Data

Code can be found in https://github.com/duanby/i-FWER. It was tested on macOS using R (version 3.6.0) and the following packages: magrittr, splines, robustbase, ggplot2.

Data in Section 5 is collected by (Himes et al., 2014) and available in R package `airway`. We follow (Ignatiadis et al., 2016) and (Lei & Fithian, 2018) to analyze the data using `DEseq2` package (Love et al., 2014).

Acknowledgements

We thank Jelle Goeman for his insightful comments on the connection between our proposed method and closed testing. We thank Will Fithian for related discussions. Eugene Katsevich, Ian Waudby-Smith, Jinjin Tian and Pratik Patil are acknowledged for their feedback on an early draft, and anonymous reviewers for helpful suggestions.

References

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

Bretz, F., Maurer, W., Brannath, W., and Posch, M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine*, 28(4):586–604, 2009.

Duan, B., Ramdas, A., Balakrishnan, S., and Wasserman, L. Interactive martingale tests for the global null. *arXiv preprint arXiv:1909.07339*, 2019.

Farcomeni, A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical methods in medical research*, 17 (4):347–388, 2008.

Goeman, J. J. and Finos, L. The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical applications in genetics and molecular biology*, 11(1): 1–18, 2012.

Goeman, J. J. and Solari, A. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

Goeman, J. J. and Solari, A. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

Himes, B. E., Jiang, X., Wagner, P., Hu, R., Wang, Q., Klanderman, B., Whitaker, R. M., Duan, Q., Lasky-Su, J., and Nikolos, C. RNA-Seq transcriptome profiling identifies crispld2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PloS one*, 9(6):e99625, 2014.

Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7): 577, 2016.

Janson, L. and Su, W. Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1):960–975, 2016.

Lei, L. and Fithian, W. AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679, 2018.

Lei, L., Ramdas, A., and Fithian, W. STAR: A general interactive framework for FDR control under structural constraints. *Biometrika (accepted)*, 2020.

Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.

Marcus, R., Eric, P., and Gabriel, K. R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Meinshausen, N. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008.

Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

Tamhane, A. C. and Gou, J. Advances in p-value based multiple test procedures. *Journal of biopharmaceutical statistics*, 28(1):10–27, 2018.

Wiens, B. L. and Dmitrienko, A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*, 15(6):929–942, 2005.