Tree-based Multiple Hypothesis Testing with General FWER Control

David Kim

March 4, 2025

Contents

1. Mathematical Formulation

Here I consider a testing procedure without data splitting, and only about a case on the 'multilplicative trees'.

In the tree-structured multiple hypotheses setting, we say $\boldsymbol{X}_t$ is a data set to construct each tth p-value.

Let $p_t a$ be the p-value of the parent node and p_{t+1} be the p-value of the subsequent child of that parent node. When we fix some $\alpha \in (0, 1]$, our basic observation is

$$P(p_{t+1} > \alpha | p_t \leq \alpha_1) \leq P(p_{t+1} > \alpha | p_t \leq \alpha_2),$$

whenever $\alpha_1 \leq \alpha_2$. This is because, we can understand our procedure is actually using

$$p'_{t+1} = \max(p_1, ...p_t, p_{t+1}),$$

as a p-value implemented in the $t+1$th node, where p_i is all other ancestors of the p_{t+1}. This can be another (maybe equivalent) understanding of the **monotonicity condition** in Goeman and Solari (2010).

From this we can recall the **positive regression dependency on each one from a subset I_0(or PRDS on I_0)** from Benjamini and Yekutieli (2001), which states when D is a non-decreasing set(so whenever $x \in D$ and $x \leq y$, then $y \in D$),

$$P(\boldsymbol{X} \in D | X_i = X)$$

is non-decreasing in x.

We also provide some notation for the tree-structured hypotheses. Let us have a tree with depth of d where for some fixed integer k, k is a number of children for every parent node (so we consider *multiplicative trees*). Then we say N is the number of total nodes and n as a number of total clusters, where

$$N = \sum_{r=1}^{d} k^{r-1}$$

$$n = 1 + \sum_{r=1}^{d-1} k^{r-1}.$$

2. Weak FWER Control

Here we consider about the weak FWER control, so when we assume all of the nulls are true. We first impose an assumption, called 'conditional super-uniformity', following Robertson et al. (2023). For the p-values p_i, we define $R_i = \mathbb{1}(p_i \leq \alpha_i)$, where $\mathbb{F}^t = \sigma(R_1, ..., R_t)$ and we let α_t as a $\mathbb{F}^{t-1}$-measurable function of $(R_1, ..., R_{t-1})$.

Definition 1. The null p-values are said to be **conditionally super-uniform** if $P(p_t \leq \alpha_t | \mathbb{F}^{t-1}) \leq \alpha_t$ for any $\mathbb{F}^{t-1}$-measurable α_t.

We also give additional notation before proposing our theorem on the weak FWER control. We say C_i is an ordered set of parents of the ith node on the tree, and d_i is the depth of the ith node. Then we see $|C_i| = d_i + 1$. Denote $C_{i,j}$ as the jth element of C_i. For example, consider a binary tree and the case $i = 8$. Then $C_8 = \{1, 2, 4\}$, and $C_{8,2} = 4$. We are also able to check $|C_8| = 3$, and $d_8 = 2$.

Now we state our theorem.

Theorem 1. If the null p-values are conditionally super-uniform, then the procedure with the critical value functions for every tth node

$$\alpha_t = \begin{cases} \frac{\alpha}{1+k\alpha}, \text{ if all of the } i\text{th tests where } i \in C_t \text{ are rejected,} \\ 0, \text{ otherwise,} \end{cases}$$

implies that the weak FWER is controlled at level α.

Proof. Assume that the all of the nulls are true. First, note that for any ith node and $\alpha \in (0, 1)$, since we assumed conditional super-uniformity on the null p-values,

$$\begin{aligned} P(p_i \leq \alpha) &= P(p_i \leq \alpha \text{ and } \forall j \in C_i,\ p_j \leq \alpha) + P(p_i \leq \alpha \text{ and } \exists j \in C_i \text{ such that } P_j > \alpha) \\ &= P(p_i \leq \alpha \text{ and } \forall j \in C_i,\ p_j \leq \alpha) \\ &= P(p_1 \leq \alpha) P(p_{C_{i,2}} \leq \alpha | p_1 \leq \alpha) ... P(p_i \leq \alpha | p_1 \leq \alpha, ..., p_{C_{i,d_i}} \leq \alpha) \\ &\leq \alpha^{d_i+1} \end{aligned}$$

holds. Therefore,

$$\begin{aligned} P(V \geq 1) &= P\Big((\cap_{i=1}^N \{p_i > \alpha_i\})^c\Big) \\ &= P\Big(\cup_{i=1}^N \{p_i \leq \alpha_i\}\Big) \\ &\leq \sum_{i=1}^N P(p_i \leq \alpha_i) \\ &= \underbrace{P(p_1 \leq \alpha_1)}_{p\text{-value on the root node}} + \underbrace{P(p_2 \leq \alpha_2) + ... + P(p_{1+k} \leq \alpha_{1+k})}_{\text{k } p\text{-values on the second level}} + \underbrace{P(p_{2+k} \leq \alpha_{2+k}) + P(p_{1+k+k^2} \leq \alpha_{1+k+k^2})}_{k^2\ p\text{-values on the third level}} + ... \\ &\quad + \underbrace{P(p_{N-k^{d-1}+1} \leq \alpha_{N-k^{d-1}+1}) + ... + P(p_N \leq \alpha_N)}_{k^{d-1}\ p\text{-values on the final level}} \\ &\leq \sum_{j=1}^d k^{j-1} \alpha_j^j \\ &\leq \sum_{j=1}^\infty k^{j-1} \Big(\frac{\alpha}{1+k\alpha}\Big)^j = \alpha, \end{aligned}$$

which gives us a desired conclusion. □

3. General FWER/FDR Control

Now we assume that, whenever j is a child node of i in the tree structure, $\boldsymbol{X}_j \subseteq \boldsymbol{X}_i$. Note that this is a much more realistic than the conditional super-uniformity assumption.

Here we give one definition and proposition:

Definition 2. Let $\boldsymbol{X} = \{X_i\}_{i \in I_0}$. We say that **PRDS property holds on $\boldsymbol{X}$** when, for any increasing set D and for each $i \in I_0$, $P(\boldsymbol{X} \in D | X_i = x)$ is nondecreasing in x.

Proposition 1 (Selection Bias on the Pooled p-values). For two data sets X_1 and X_2, assume that $X_2 \subseteq X_1$, and p_i is a p-value coming from each X_i. Then for any $\alpha, \alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1 \leq \alpha_2$,

$$P(p_2 \leq \alpha | p_1 \leq \alpha_2) \leq P(p_2 \leq \alpha | p_1 \leq \alpha_1)$$

holds.

Then, this is the argument that I will prove:

Proposition 2. Assume that previous proposition is true, and we follow the tree strucutre procedure of hypothesis testing. Then the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to true null hypotheses. That is, the Benjamini-Hochberg procedure controls the FDR.

Consider the simple case that there are only two nodes where p_1 is a parent p-value and p_2 is a child p-value. Further, assume that the nulls on each node are all true. Then we see, our actual testing procedure is using p_1 at the first node and $\max(p_1, p_2)$ at the second node. Then if the Proposition 1 is true, we see for any $\alpha, \alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1 \leq \alpha_2$,

$$P(p_1 > \alpha, \max(p_1, p_2) > \alpha | p_i \leq \alpha_1) \leq P(p_1 > \alpha, \max(p_1, p_2) > \alpha | p_i \leq \alpha_2)$$

for each $i \in \{1, 2\}$.

4. Miscellanea

My idea is that,

- **Since $p_1, ..., p_n$ (so the p-values after local adjustments) are still has a PRDS property, we can try to directly apply Benjamini and Yekutieli (2001)'s approach to get general FDR or FWER control.**
- **We can also consider which kind of local adjustments give valid and more powerful tests.**

I will further check about,

- **Idea of data splitting used in the recent simulation.**
- **Prove explicitly the validness of locally adjusted p-values from different methods.**

REFERENCES

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 38:6:3782–3810, 2001.

Jell J. Goeman and Aldo Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38:6:3782–3810, 2010.

David S. Robertson, James M. S. Wason, and Aaditya Ramdas. Online multiple hypothesis testing. *Statistical Science*, 38:4:557–575, 2023.