An Interactive Framework for Structured Multiple Testing

Lihua Lei, Aaditya Ramdas and William Fithian

Setup

- Hypotheses $H_{0,i}, i \in [n]$ with $\mathcal{H}_0 = \{i : H_{0,i} \text{ is null}\}$.
- p_i: p-values, x_i: side information.
- Prior: $\mathcal{X}^* \triangleq \{x_i : i \notin \mathcal{H}_0\}$ has certain "combinatorial" structure:

  x_i: location in $\mathbb{R}^2$ or $\mathbb{R}^3$, $\mathcal{X}^*$: a convex set.
  x_i: node of a tree, $\mathcal{X}^*$: a rooted subtree.
  x_i: node of a DAG, $\mathcal{X}^*$: satisfying a heredity principle.
- Goal: to reject a subset of hypotheses satisfying the given structure while controlling FDR at a predefined level.

Accumulation Test

Accumulation Test (Li and Barber, 2015): reject $H_{0,i}, i \in [\hat{k}]$ where

$$\hat{k} = \max\{k : \widehat{\text{FDP}}_k \le \alpha\}, \quad \widehat{\text{FDP}}_k = \frac{C + \sum_{i=1}^k h(p_i)}{1+k},$$

and h can be any non-negative function with $\int_0^1 h(p)dp = 1$. Then

$$\text{FDR} \le \frac{\alpha}{\mathbb{E}_{P\sim U([0,1])}\{h(P) \wedge C\}}.$$

- SeqStep (Barber and Candès, 2014): $h(p) = \frac{I(p \ge p^*)}{1-p^*}$
- ForwardStop (G'sell et al, 2013): $h(p) = -\log(1-p)$
- HingeExp (Li and Barber, 2015): $h(p) = \frac{\log(1-p^*) - \log(1-p)}{1-p^*} \cdot I(p \ge p^*)$

Interactive Accumulation Test

- Observation: AT only uses partial information $h(p)$.
- Idea: use the "leftover information" $g(p)$ to guide the procedure.

Interactive Accumulation Test (IAT):

Step 1. Given $(h(\cdot), C)$ and *auxiliary function* $g(\cdot)$. Initialize

$$\mathcal{F}_{-1} = \sigma\left((x_i, g(p_i))_{i\in[n]}, \sum_{i=1}^n h(p_i)\right), \quad \mathcal{R}_{-1} = [n]$$

Step 2. In step t, update $\mathcal{R}_t$, the rejection set considered at step t, with ***an arbitrary method*** such that

$$\mathcal{R}_t \subset \mathcal{R}_{t-1}, \quad \mathcal{R}_t \in \mathcal{F}_{t-1}$$

Step 3. Observe the p-values in $\mathcal{R}_{t-1} \setminus \mathcal{R}_t$ and update $\mathcal{F}_{t-1}$ accordingly:

$$\mathcal{F}_t = \sigma\left((x_i, g(p_i))_{i\in[n]}, (p_i)_{i\notin\mathcal{R}_t}, \sum_{i\in\mathcal{R}_t} h(p_i)\right).$$

Step 4. Estimate FDP by

$$\widehat{\text{FDP}}_t = \frac{C + \sum_{i\in\mathcal{R}_t} h(p_i)}{1 + |\mathcal{R}_t|}$$

Step 5. Repeat Step 2-4 until $\widehat{\text{FDP}}_t \le \alpha$.

Theorem 1. *Assume that the null p-values are independent of each other and of the non-null p-values. Then IAT controls FDR at level $\rho^{-1}\alpha$ where*

$$\rho = \inf_{i\in\mathcal{H}_0} \inf_q \mathbb{E}\left(h(p_i) \wedge C \mid g(p_i) = q\right).$$

Choice of $g(p)$

Given $(h(\cdot), C)$, we derive the optimal $g(\cdot)$ is of the form

$$g(p) = \min\{p, s(p)\},$$

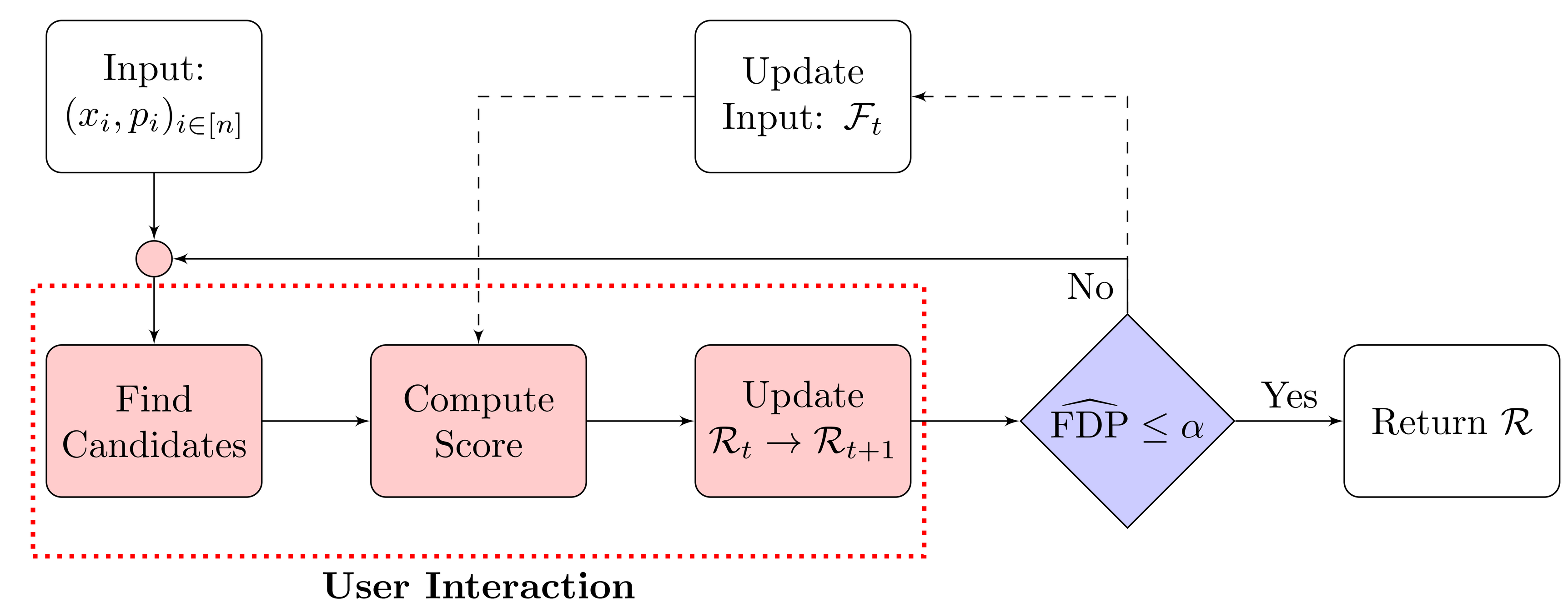
where $s(p)$ is the decreasing solution of an ordinary differential equation. This choice yields

$$\rho = \rho_{\max} = \mathbb{E}_{P\sim U([0,1])}\{h(P) \wedge C\}.$$

Examples:

- Interactive SeqStep (ISS): $C = \frac{1}{1-p^*}, \rho = 1, s(p) = \frac{p^*}{1-p^*}(1-p)$
- Interactive ForwardStop (IFS): $C = 4.605, \rho = 0.99$
- Interactive HingeExp (IHE): $C = \frac{4.605}{1-p^*}, \rho = 0.99$

Implementation



- Find all candidate (subsets of) hypotheses to reject:
  - Sets of hypotheses removing which does not break the structure
- Compute a "score" for each candidate:
  - Naive score: $g(p)$ (taking average for a set)
  - Model-assist score: case-specific
- Update the rejection set by "revealing" some of candidates:
  - Default: remove the candidate with the least favorable score

Applications

Ask the presenter for the fun animations!

Convex Region Detection

- **Setup**: $x_i \in \mathbb{R}^p$
- **Goal**: detect a convex region/axis-parallel box on $\mathbb{R}^p$
- **Candidates**: $\mathcal{C}_\mathcal{L} = \{i : x_i \text{ is above } \mathcal{L}\}$ for all hyperplanes $\mathcal{L}$ s.t. $|\mathcal{C}_\mathcal{L}| = 0.01n$
- **Score**: Fitting a non-parametric Beta-GLM (Lei and Fithian, 2016)
- **Real Applications**: 1) bump hunting, 2) change point detection

Hierarchical Testing

- **Setup**: $H_{0,i}$ are placed on a tree $\mathcal{T}$, x_i represents the node
- **Goal**: detect a subtree from $\mathcal{T}$ with same root
- **Candidates**: all leaf hypotheses from the masking subtree $\mathcal{R}_t$
- **Real Applications**: 1) pruning CART, 2) phylogeny, 3) microarray data, 4) wavelet thresholding

Selection Under Heredity Principle on DAG

- **Setup**: $H_{0,i}$ are placed on a DAG $\mathcal{G}$, x_i represents the node
- **Goal**: detect a subgraph under strong/weak heredity principles
- **Candidates**: restricted in the masking subgraph $\mathcal{R}_t$:
  - (strong) all leaf hypotheses
  - (weak) hypotheses not being the only parent of other nodes
- **Real Applications**: 1) factorial design, 2) gene ontology

Some Simulation Results

Hierarchical Testing:

- $\mathcal{T}$ is a binary tree with $|\mathcal{T}| = 2000$
- $x_i \sim N(\mu_i, 1)$ with $\mu_i = 2I(i \in \mathcal{H}_0^c)$
- $\mathcal{H}_0^c = \{1, \ldots, 100\}$ under 1) the breadth-first ordering (BFS) and the depth-first ordering (DFS)