Interactive algorithms for multiple hypothesis testing

Aaditya Ramdas
Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

(Always looking for postdocs and funding)
IS THERE A REPRODUCIBILITY CRISIS?

- 52% Yes, a significant crisis
- 38% Yes, a slight crisis
- 3% No, there is no crisis
- 7% Don’t know

1,576 researchers surveyed

©nature
My opinion: one major issue is researcher degrees of freedom

Data scientists want to interact with data, and use their intuition, priors to exploit structure in the data.

However, many classical statistical methods were not built to handle interactivity.

“double-dipping” selection bias

There is a need for principled statistical methods that work in sequential and interactive settings.
Classical testing is a unidirectional pipeline

Choose a test → Collect data (& side info) → Conclusion

Wait... perhaps I should use some other model/methods

Conclusion not reproducible

Test statistics cannot be altered after observing data
A hypothetical data scientists’ wish-list:

- use prior knowledge and intuition
- incorporate structure, soft or hard constraints
- interactive exploration with human-in-the-loop
- employ flexible probabilistic modeling tools
- robust to unknown dependence

The challenge: correct “statistical inference”

This talk: some progress within multiple testing
Interactive testing is allows “loops”

1. Collect data

2. Mask Data (& side info)

3. Interactive test (multi-step)

4. Progressively Unmask Data

5. Not rejected

6. Conclusion

Test can be customized and revised after observing data.
- False discovery proportion \( \text{FDP} = \frac{V}{R} \)
- False discovery rate \( \text{FDR} = \mathbb{E}[\text{FDP}] \)
- Many notions of power, e.g.: \( \text{Power} = \frac{\mathbb{E}[R - V]}{|\mathcal{H}_1|} \)
- We desire procedures that have low FDR and high power
Interactive multiple testing

**Lei, Ramdas, Fithian** (Biometrika’21)
STAR algorithm (i-FDR) enforces structure, generalized masking

**Duan, Ramdas, Wasserman** (ICML’20) handles conservative nulls, better, flexible masking (i-FWER)

**Duan, Ramdas, Wasserman** (arXiv) moves away from p-values (i-FDR-causal)

---

Time

**Lei & Fithian** (JRSSB’18)
conceptualized masking, designed AdaPT algorithm (i-FDR)

**Duan, Ramdas, Balakrishnan, Wasserman** (EJS’21) martingale tests (i-global null)

**Duan, Ramdas, Wasserman** (arXiv) interactive Wilcoxon and Friedman tests (i-rank tests)

**Yang, Lei, Ho, Fithian** (arXiv) Bonus algorithm for multivariate settings (i-FDR)

Boyan Duan (PhD 2021; currently, Google)
Thesis: Advances in interactive inference.
The rest of this talk:
Involving a human-in-the-loop

**STAR**: an interactive framework for structured multiple testing (Selectively Traversed Accumulation Rules)

**I^3**: interactive identification of individual treatment effects (causal inference)
For the first part of this talk:

Each hypothesis will have some associated side information in the form of covariates (features).

**Assumption:** null p-values are conditionally independent of each other, and of the non-nulls, given the covariates.

We will return to this assumption.
First, a new interactive framework for structured multiple testing: STAR

- use prior knowledge and intuition
- incorporate structure, soft or hard constraints
- interactive exploration with human-in-the-loop
- employ flexible probabilistic modeling tools
- robustness to unknown dependence

Knockoffs + Accumulation Tests + AdaPT + more = STAR
Before presenting STAR, 3 pieces of intuition:

**Intuition part 1:** “masking the p-values” enables exploration while avoiding selection bias.

\[
g(P_i) = \min(P_i, 1 - P_i) \quad \text{“masked p-value”}
\]

Used for selection

\[
h(P_i) = 2I(P_i > 1/2) \quad \text{“missing bit”}
\]

Used for inference

(Independent under the null)
\[ Z_i \sim N(\mu_i, 1) \]
\[ P_i = 1 - \Phi(Z_i) \]
\[ \Phi : \text{Gaussian CDF} \]
\[ g(P_i) := \min\{P_i, 1 - P_i\}. \]

\[ \mu_i = 0 \text{ for pink pixels} \]
\[ \mu_i = 2 \text{ for red pixels} \]

Reveal masked p-values \( g(P_i) \).

Scientist “discovers” set \( R \).

\[ h(P_i) := 2I(P_i > 1/2). \]

\[ \hat{\text{FDP}} \approx \frac{\sum_{i \in R} h(P_i)}{|R|} \]
Before presenting STAR, 3 pieces of intuition:

**Intuition part 2**: masking permits free use of modeling, even if assumed models are wrong.
\( g(P_i) = \min\{P_i, 1 - P_i\} \)

**Idea:** Assume an underlying model, and impute missing bit.

\[
\hat{\text{FDP}} \approx \frac{\sum_{i \in \mathcal{R}} h(P_i)}{|\mathcal{R}|}
\]

\[
h(P_i) = 2I(P_i > 1/2)
\]

Selection still based only on masked p-values!
Before presenting STAR, 3 pieces of intuition:

**Intuition part 3**: progressive unmasking enables interaction with the scientist.
Progressive unmasking allows interaction

\[ \widehat{\text{FDP}} \approx \frac{\sum_{i \in \mathcal{R}} h(P_i)}{|\mathcal{R}|} \]

\[ h(P_i) \perp g(P_i) \text{ for nulls } i. \]

\[ \text{FDP} \perp P_i \text{ for } i \notin \mathcal{R} \]

- After selection, we can reveal all un-rejected p-values!
- Use new information to *contract* selected set (freely).
- When we recurse, we obtain the **STAR algorithm**.
Scientist chooses target FDR $\alpha$, arbitrary constraints $\mathcal{K}_0$, covariates $x_i$ for hypothesis $i$.

**Step 1** Initialize $\mathcal{R}_0 = [n]$, and reveal the information

$$\mathcal{F}_0 = \sigma \left( (x_i, g(P_i))_{i=1}^{n}, \sum_{i \in \mathcal{R}_0} h(P_i) \right)$$

**Step 2** Stop and reject $\mathcal{R}_t$ if $\mathcal{R}_t \in \mathcal{K}_t$ and $\widehat{\text{FDP}}_t \leq \alpha$, where

$$\widehat{\text{FDP}}_t = \frac{2 + \sum_{i \in \mathcal{R}_t} h(P_i)}{1 + |\mathcal{R}_t|}$$

**Step 3** Choose $\mathcal{R}_t \subset \mathcal{R}_{t-1}$, $\mathcal{R}_t \in \mathcal{F}_{t-1}$, $\mathcal{K}_t \in \mathcal{F}_{t-1}$, and update human-in-the-loop, or automated procedure.

$$\mathcal{F}_t = \sigma \left( (x_i, g(P_i))_{i=1}^{n}, (P_i)_{i \notin \mathcal{R}_t}, \sum_{i \in \mathcal{R}_t} h(P_i) \right)$$

(go to step 2)
Main Theorem

Assume that null p-values are independent of each other and of the non-null p-values.

Then \( \text{STAR} \) controls FDR at level \( \alpha \).

If \( \mathcal{K}_t \) are “predictable” constraints, and
\[
\tau = \min\{t \geq 0 : \mathcal{R}_t \in \mathcal{K}_t, \hat{\text{FDP}}_t \leq \alpha\},
\]
then \( \mathbb{E} [\text{FDP}(\mathcal{R}_\tau)] \leq \alpha \).
Other applications in the paper:

- Tree structured hypotheses
- Directed acyclic graphs
- Bump-hunting
- Knockoffs with structured covariates
For the second part of this talk:

Each hypothesis (person) will have some associated side information in the form of covariates (features).

**Assumption:** each person is assigned treatment or control at random, and their corresponding potential outcome is observed (no interference).
Motivating example
- covariates $X_i$ (eg. age, body weight, gender...)
- i.i.d. $\text{Ber}(1/2)$ treatment assignment $A_i \in \{0, 1\}$
- potential control and treated outcome $(Y_i^C, Y_i^T)$
- observe $Y_i = A_i Y_i^T + (1 - A_i) Y_i^C$ (consistency)

Q: which subjects have positive treatment effects $Y_i^T > Y_i^C$?
(individual level inference)
Motivating example

- covariates $X_i$ (e.g., age, body weight, gender...)
- i.i.d. $\text{Ber}(1/2)$ treatment assignment $A_i \in \{0,1\}$
- potential control and treated outcome $(Y_i^C, Y_i^T)$
- observe $Y_i = A_i Y_i^T + (1 - A_i) Y_i^C$ (consistency)

Goal: identify subjects with positive effect $Y_i^T > Y_i^C$, denoted as set $R$, with error control on the expected proportion of false identifications:

$$
\mathbb{E} \left[ \frac{|\{i : Y_i^T \leq Y_i^C \} \cap R|}{|R| \lor 1} \right] .
$$

We choose to formalize the problem using the language of hypothesis testing.
Problem setup

Define a hypothesis testing problem for each subject \( i \in \{1, \ldots, n\} \):

**Definition 1** treating potential outcomes as random variables

\[
\begin{align*}
H_{i0}^*: F(Y_i^T \mid X_i) &= F(Y_i^C \mid X_i) \\
H_i : F(Y_i^T \mid X_i) &> F(Y_i^C \mid X_i)
\end{align*}
\]

zero effect (equal in distribution)  
positive effect (stochastic dominance)

Alternative Definitions eg. treating potential outcomes and covariates as fixed values

We deal with randomized experiments without interference, assuming:

(A1) assignments are independent coin flips:

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

(A2) conditional on the covariates, the outcome of one subject \( Y_{i_1} \) is independent of the assignment of another \( A_{i_2} \) for any \( i_1 \neq i_2 \).

\[\star\] can be extended to error control for nonpositive effects:

\[
\widehat{H}_{i0} : F(Y_i^T \mid X_i) \leq F(Y_i^C \mid X_i).
\]
Problem setup
Let the set of nulls (subjects with zero effect) be $\mathcal{H}_0 = \{i : H_{i0} \text{ is true}\}$, the set of rejected (identified) subjects be $R$. False discovery rate: expected proportion of false identifications

$$\text{FDR} := \mathbb{E} \left( \frac{|\mathcal{H}_0 \cap R|}{|R| \lor 1} \right).$$

Output of proposed $I^3$ (for interactive identification of individual effects): a set of identified subjects $R$ with FDR $\leq \alpha$ and reasonably high power.

Subjects (with two covariates) randomly assigned to treated and control group. True positive effects (blue) are unknown ground truth. Identified subjects (green) contain most true positives, regardless of treated or not.
Algorithm description of $I^3$

**Explorer**

1. Let $\hat{Y}$ be any estimator of $\mathbb{E}(Y_i \mid X_i)$, and compute residuals $E_i = Y_i - \hat{Y}(X_i)$

**Oracle**

2. Compute a treatment effect estimator: $\hat{\Delta}_i := 4(A_i - 1/2)E_i$

3. Divide $R_t$ into
   - $R_t^+ := \{i \in R_t : \hat{\Delta}_i > 0\}$ and $R_t^- := \{i \in R_t : \hat{\Delta}_i \leq 0\}$
   - (to construct FDR estimator)

Repeat step 4-8 for $t = 1, \ldots, n$
Estimated effect $\hat{\Delta}_i$

Define

$$\hat{\Delta}_i := 4(A_i - 1/2)(Y_i - \hat{Y}(X_i))$$

where $\hat{Y}(\cdot)$ is an arbitrary estimator of $\mathbb{E}(Y_i | X_i)$ without using $\{A_i\}_{i=1}^n$.

Reasonable
- used in recent papers such as Nie and Wager (2020), Kennedy (2020); and can be traced back to Robinson (1988).
- recover true effect in simple cases. Eg. Suppose $Y_i^C = c$ and $Y_i^T = c + \delta$ for all $i$.
  
  If $\hat{Y}$ is correctly learned: $\hat{Y}(X_i) = c + \delta/2$, then $\hat{\Delta}_i = \delta$.

Error control by the sign property:

$$\mathbb{P}\left(\hat{\Delta}_i > 0 \mid \{Y_j, X_j\}_{i=1}^n\right) \leq 1/2,$$

if subject $i$ has zero effect (null hypothesis is true), under assumption (A1) and (A2) of randomized experiments;

Error control

Selection
Algorithm description of $\mathcal{I}^3$

Start with $R_0 = \{1, \ldots, n\}$ and $t = 0$

**Prior knowledge** \{\(Y_i, X_i\)\}_{i=1}^{n}

**Explorer**

1. Learn $\hat{Y}$ and compute $E_i := Y_i - \hat{Y}(X_i)$

2. Set $\Delta_i := 4(A_i - 1/2)E_i$

3. Divide $R_t$ into $R_t^+$ and $R_t^-$

4. Check $\overline{\text{FDR}}(R_t) := \frac{|R_t^-| + 1}{|R_t^+| \lor 1} \leq \alpha$

   - If no

5. Pick $i_t^* \in R_t$ (hopefully a null and $\Delta_{i_t^*} \leq 0$) using explorer's current knowledge

6. Reveal $A_{i_t^*}$

7. Calculate $\Delta_{i_t^*}$ (and its sign).

8. Update $R_{t+1} = R_t \backslash \{i_t^*\}$ also $|R_{t+1}^+|, |R_{t+1}^-|$

Repeat step 4-8

**Oracle**

Stop at $\tau = t$

Report $R_{\tau}^+$

Yes

Stop at $\tau = t$

Report $R_{\tau}^+$

Theorem

\(\text{FDR}(R_{\tau}^+) \leq \alpha\)

FDR control using binary variables:

- Barber and Candes (2015)
- Lei, Fithian (2018)
- Lei, Ramdas, Fithian (2021)
An example of automated algorithm to shrink $R_t$

1. Using non-candidate subjects $j \notin R_t$ with complete data, train a random forest classifier where the label is the sign

$$1 \left( \hat{\Delta}_j > 0 \right) \equiv 1 \left( (A_j - 1/2)(Y_j - \hat{Y}(X_j)) > 0 \right)$$

and the predictors are $Y_j, X_j$ and $Y_j - \hat{Y}(X_j)$.

2. For candidate subjects $i \in R_t$ without $A_i$ hence without sign (label), predict probability of $\hat{\Delta}_i$ being positive, denoted as $\hat{p}(i, t)$.

3. Find $i^*_t = \arg\min\{\hat{p}(i, t) : i \in R_t(I)\}$ and $R_{t+1}(I) = R_t(I) \setminus \{i^*_t\}$.

**Note:** The explorer can choose to **update** or **change** the working modeling (could lead to higher power if correct) at any step.
Numerical experiments

Generating model:

\[ X_i = (X_i(1), X_i(2), X_i(3)) \in \{0,1\}^2 \times \mathbb{R} \text{ and } \epsilon_i \sim N(0,1) \]

\[ Y_i^C = 5(X_i(1) + X_i(2) + X_i(3)) + \epsilon_i \]

\[ Y_i^T = \Delta(X_i) + 5(X_i(1) + X_i(2) + X_i(3)) + \epsilon_i \]

Positive-biased effect:

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

- 15% positive effects with size 20S_{\Delta} and
- 45% negative effects with size 0.5S_{\Delta}.

- Crossfit-\(I^3\) guarantees FDR control; the parametric method (Linear-BH) does not.
- Less than 25% identified by Crossfit-\(I^3\) are nonpositive; more than 50% by Linear-BH.
- Crossfit-\(I^3\) has good power to identify true positive effect.
Talk Summary

Important and advantageous to have a human-in-the-loop

- (STAR) Interaction can be enabled by masking of p-values
- (STAR) Arbitrary structural constraints on the rejected set
- (STAR) Use flexible models, retain frequentist guarantee

In randomized experiments, hide the treatment assignment

- (I-cubed) Enables scientist and algorithm to work together
- (I-cubed) Discover +ve effects despite unobserved outcome
- (I-cubed) Non-asymptotic FDR control
Interactive algorithms for multiple hypothesis testing

Aaditya Ramdas
Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

(Always looking for postdocs and funding)