Detecting Interference in A/B Testing with Increasing Allocation

Shuangning Li

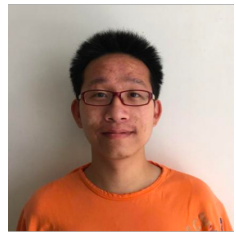
Harvard University

LinkedIn

Joint work with

Kevin Han

Jialiang Mao

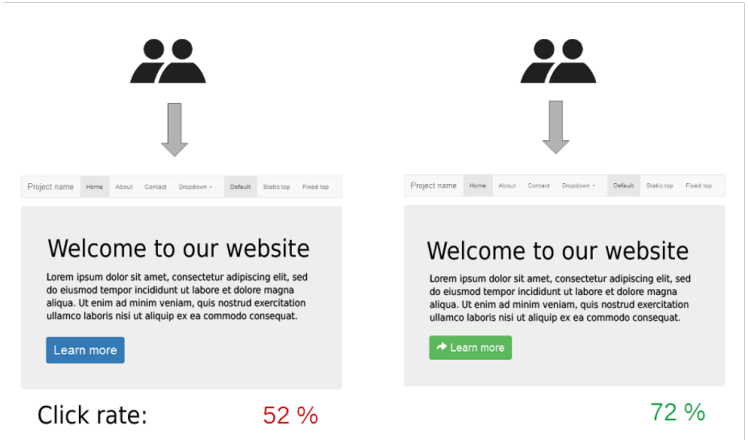


Han Wu

Han, Kevin, Shuangning Li, Jialiang Mao, and Han Wu. **Detecting Interference in Online Controlled Experiments with Increasing Allocation.** *KDD*. 2023.

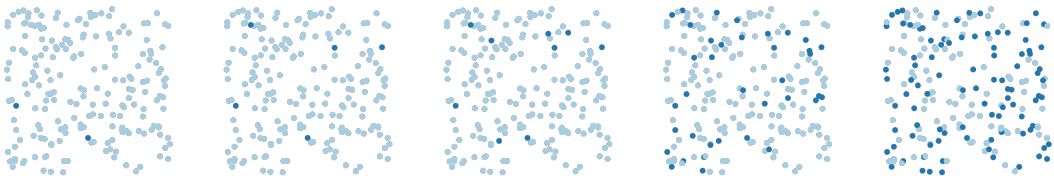
A/B Testing

- A/B testing has been adopted by the technology industry to guide product development and make business decisions.



A/B Testing with Increasing Allocation

- Dynamic phase release framework: a new treatment (such as a new product feature) is gradually released to an increasing number of units in the target population through a sequence of randomized experiments.

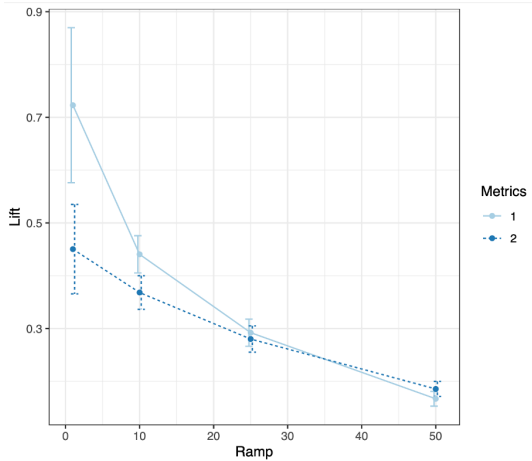


Difference in Mean Estimator

- $\hat{\tau}$ = Average(treatment group) - Average(control group)
- Under Stable Unit Treatment Value Assumption (SUTVA), $\hat{\tau}$ shouldn't change much when we increase the treatment probability.
- But sometimes we see this:

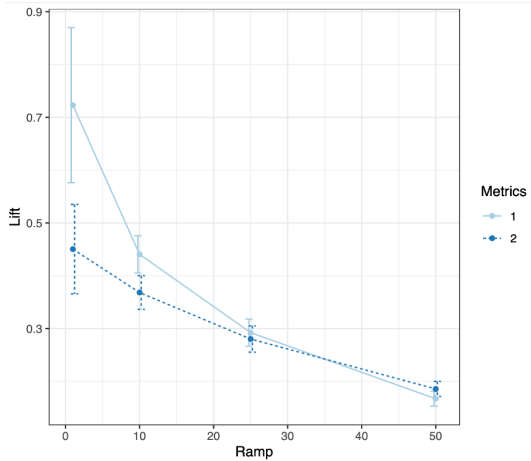
Difference in Mean Estimator

- $\hat{\tau}$ = Average(treatment group) - Average(control group)
- Under Stable Unit Treatment Value Assumption (SUTVA), $\hat{\tau}$ shouldn't change much when we increase the treatment probability.
- But sometimes we see this:
- $\hat{\tau}$ decreases with treatment probability!



Difference in Mean Estimator

- What happens?
  - Randomness?
  - Interference?
- Interference examples:
  - Marketplace cannibalization (decreasing)
  - Social network (increasing)
- What to do in practice?
  - Need a formal statistical way to "decide" whether interference exists.



Our Contribution

Sequential A/B testing

Filter

No interference

Standard statistical analysis

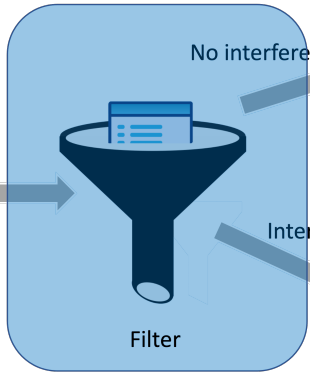
Interference exists

Cluster randomized trials
Budget-split platform
...

Platforms that carefully deal with interference

Our Contribution

Sequential A/B testing



Standard statistical analysis

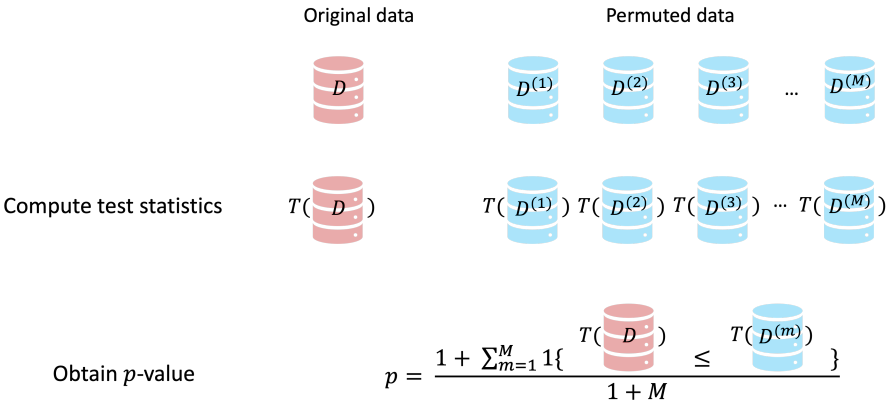
Cluster randomized trials
Budget-split platform
…

Platforms that carefully deal with interference

- Scalable, parallelable, “almost for free!”
- Agnostic to interference mechanism

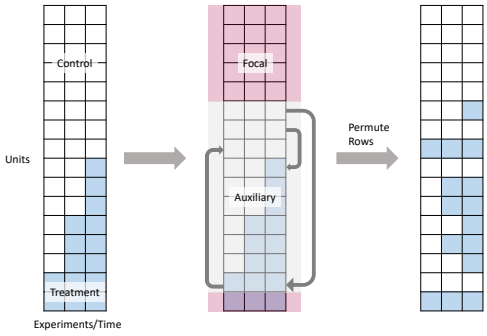
Permutation Test

- We develop methods that test for the null hypothesis that there is no cross-unit interference.
- We consider permutation tests:



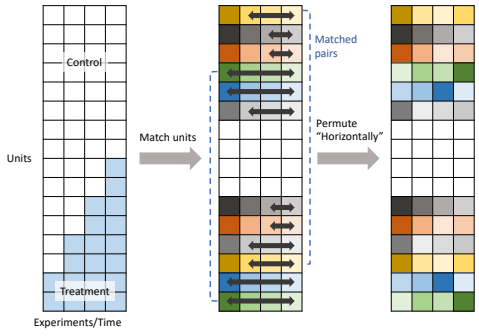
Two permutation tests

General assumption



- No modeling assumption
- Need to construct graph: helps understand the underlying interference mechanism

Time fixed effect assumption



- Low computational complexity
- No need to construct graph
- More powerful

Problem Setup

- There are K experiments on a population of n units.
- Let π_k be the marginal treatment probability of the k^{th} experiment with

$$\pi_1 < \pi_2 < \cdots < \pi_K.$$

- For each experiment $k \in \{1, \ldots, K\}$ and each unit $i \in \{1, \ldots, n\}$, let

$$W_{i,k} := \text{treatment of unit } i \text{ assigned in the } k^{\text{th}} \text{ experiment} \in \{0, 1\},$$
$$Y_{i,k} := \text{outcome of unit } i \text{ in the } k^{\text{th}} \text{ experiment} \in \mathbb{R}.$$

- Let $X_i \in \mathbb{R}^d$ be the observed covariates of unit i that do not change over the course of the experiments.

Problem Setup

- In the first experiment, each unit i is randomly assigned a treatment $W_{i,1}$, where

$$W_{i,1} \sim \text{Bernoulli}(\pi_1) \text{ independently.} \tag{1}$$

- In the subsequent experiments, conditioning on the previous treatments, each $W_{i,k}$ is sampled from the following distribution independently:

$$\begin{cases} W_{i,k} \sim \text{Bernoulli}\left((\pi_k - \pi_{k-1})/(1-\pi_{k-1})\right), & \text{if } W_{i,k-1} = 0; \\ W_{i,k} = 1, & \text{if } W_{i,k-1} = 1. \end{cases} \tag{2}$$

- This formulation guarantees that if we look at the k^{th} experiment alone, then the treatments $W_{i,k}$'s are i.i.d. Bernoulli(π_k).

Hypothesis

No cross-unit interference hypothesis

$Y_{i,k}(w_{1:n,1:K}) = Y_{i,k}(\tilde{w}_{1:n,1:K})$ if $w_{i,1:K} = \tilde{w}_{i,1:K}$.

- The hypothesis states that the outcomes of unit i depend only on the treatments of unit i and not on the treatments of others.

Testing under General Assumptions

- **Candidate Exposure**: captures the potential form of interference.
- Social network setting: number of friends treated, proportion of friends being treated.
- Marketplace setting: number of treated competitors.
- We denote the candidate exposure by $H_{i,k}$.

Testing under General Assumptions

Algorithm 1 Testing for interference effect (one experiment).

1. Randomly split the data into two folds: focal units and auxiliary units.
2. Run a linear regression of $Y_{\text{foc}} \sim W_{\text{foc}} + X_{\text{foc}} + H_{\text{foc}}$, extract the coefficient of H_{foc}, and take the test statistic $T^{(0)}$ to be the absolute value of the coefficient.
3. **For** $b = 1, \dots B$:
   Regenerate treatments for auxiliary units.
   Recompute the candidate exposure H for focal units.
   Recompute the test statistic: $T^{(b)}$ with the newly generated H.
   End For

Output: The p-value

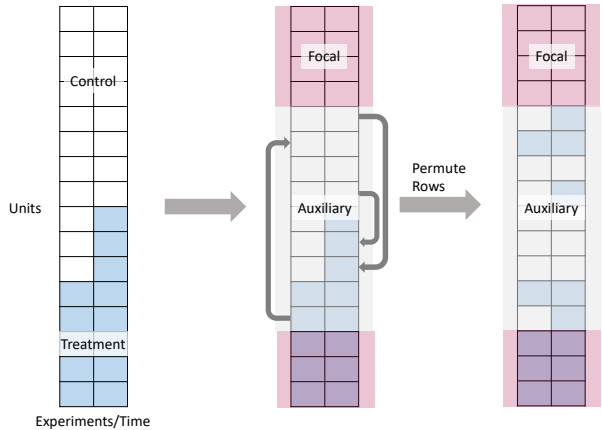
$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T^{(0)} \leq T^{(b)}\right\}\right).$$

Athey et al. [2018]

Testing under General Assumptions

Utilize two experiments?

- Choice of focal units
- Permutation instead of regenerating
- Use $Y_{i,2} - Y_{i,1}$ to reduce variance



Testing with a Time Fixed Effect Model

- The method on the previous slides essentially contrasts different units based on their values of H_i. Some units possess higher values of H_i, while others have lower values.
- Can we contrast units at different times instead?
- Need additional assumptions!

Assumption (No temporal interference)

$Y_{i,k}(w_{1:n,1:K}) = Y_{i,k}(\tilde{w}_{1:n,1:K})$ if $w_{1:n,k} = \tilde{w}_{1:n,k}$.

Assumption (Time fixed effect)

$Y_{i,k}(w_{1:n}) = \alpha_i(w_{1:n}) + u_k + \epsilon_{i,k}(w_{1:n})$.

Testing with a Time Fixed Effect Model

Hypothesis'

$$Y_{i,k}(w_{i,k}) = \alpha_i(w_{i,k}) + u_k + \epsilon_{i,k}(w_{i,k}).$$

- Two units i and j: i has been in the treatment group the whole time, while j has been in the control group the whole time.
- Under Hypothesis',
  - For the first experiment,
    $Y_{i,1} - Y_{j,1} = (\alpha_i(1) + u_1 + \epsilon_{i,1}(1)) - (\alpha_j(0) + u_1 + \epsilon_{j,1}(0)) = \alpha_i(1) - \alpha_j(0) + \epsilon_{i,1}(1) - \epsilon_{j,1}(0).$
  - For the second experiment,
    $Y_{i,2} - Y_{j,2} = (\alpha_i(1) + u_2 + \epsilon_{i,2}(1)) - (\alpha_j(0) + u_1 + \epsilon_{j,1}(0)) = \alpha_i(1) - \alpha_j(0) + \epsilon_{i,2}(1) - \epsilon_{j,2}(0).$

    $\Rightarrow Y_{i,1} - Y_{j,1} = \alpha_i(1) - \alpha_j(0) + \epsilon_{i,1}(1) - \epsilon_{j,1}(0) \stackrel{d}{=} \alpha_i(1) - \alpha_j(0) + \epsilon_{i,2}(1) - \epsilon_{j,2}(0) = Y_{i,2} - Y_{j,2}.$
  - $Y_{i,1} - Y_{j,1}$ and $Y_{i,2} - Y_{j,2}$ have the same distribution under Hypothesis'!

Testing with a Time Fixed Effect Model

Hypothesis'

$$Y_{i,k}(w_{i,k}) = \alpha_i(w_{i,k}) + u_k + \epsilon_{i,k}(w_{i,k}).$$

- Two units i and j: i has been in the treatment group the whole time, while j has been in the control group the whole time.
- Under alternative hypothesis,
  - Consider a simple alternative model:

$$Y_{i,k} = W_{i,k} H_{i,k} + \epsilon_{i,k},$$

    where $H_{i,k}$ is the fraction of neighbors of unit i treated in experiment k.
  - $Y_{i,1} - Y_{j,1} = H_{i,1} + \epsilon_{i,1} - \epsilon_{j,1}$ and $Y_{i,2} - Y_{j,2} = H_{i,2} + \epsilon_{i,2} - \epsilon_{j,2}$.
  - When the number of neighbors of unit i is large, by law of large numbers, we have $H_{i,1} \approx \pi_1$ and $H_{i,2} \approx \pi_2$.
  - $Y_{i,1} - Y_{j,1}$ and $Y_{i,2} - Y_{j,2}$ are different!

Testing with a Time Fixed Effect Model

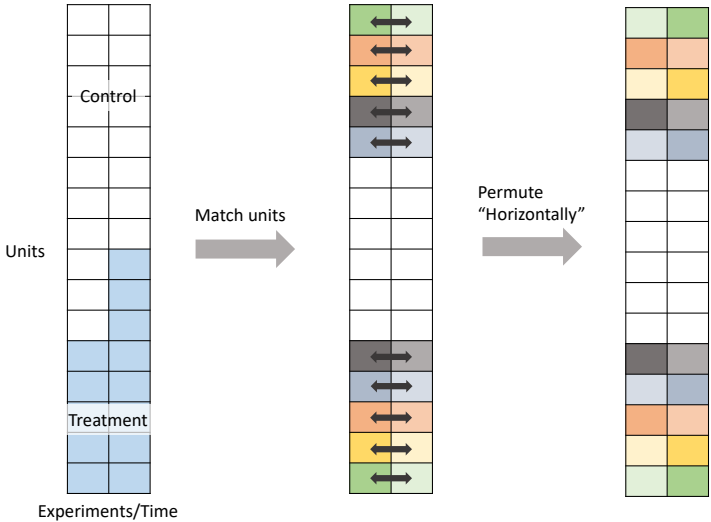
Algorithm 3 Testing for interference effect (two experiments, time fixed effect model).

1. Let $\mathcal{I}_0 = \{i : W_{i,1} = W_{i,2} = 0\}$ and $\mathcal{I}_1 = \{i : W_{i,1} = W_{i,2} = 1\}$.
2. For each i in $\mathcal{I}_1$, match an index $j \in \mathcal{I}_0$ to i (with no repeat)
3. For each $k \in \{1, 2\}$, compute $Y^{\text{diff}}_{\mathcal{I}_1,k} = \left(Y_{i,k} - Y_{m(i),k}\right)_{i \in \mathcal{I}_1}$. Compute a test statistic $T^{(0)} = |\operatorname{mean}(Y^{\text{diff}}_{\mathcal{I}_1,2}) - \operatorname{mean}(Y^{\text{diff}}_{\mathcal{I}_1,1})|$.
4. **For** $b = 1, \ldots B$:
   For each $i \in \mathcal{I}_1$:
   Randomly permute outcomes across experiments.
   End For
   Recompute $\widetilde{Y}^{\text{diff},(b)}_{\mathcal{I}_1,k} = (\widetilde{Y}^{(b)}_{i,k} - \widetilde{Y}^{(b)}_{m(i),k})_{i \in \mathcal{I}_1}$.
   Recompute the test statistic: $T^{(b)}$.
   End For

Output: The p-value

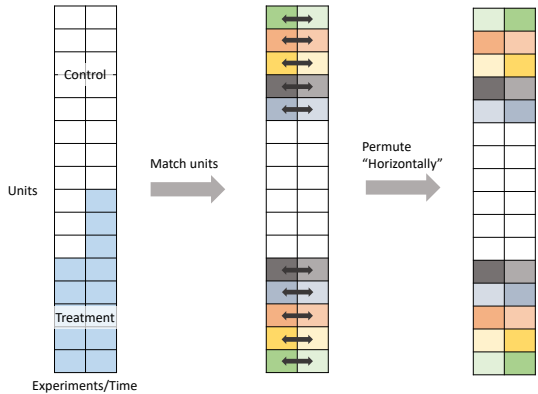
$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left\{T^{(0)} \leq T^{(b)}\right\}\right).$$

Testing with a Time Fixed Effect Model

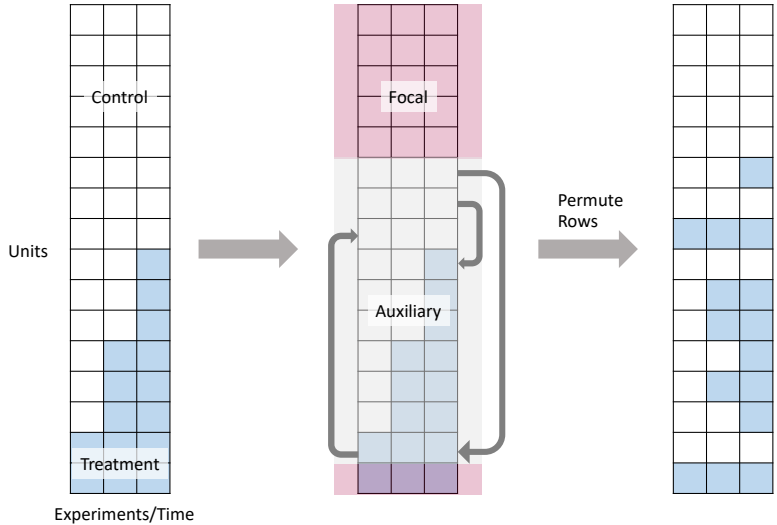


Testing with a Time Fixed Effect Model

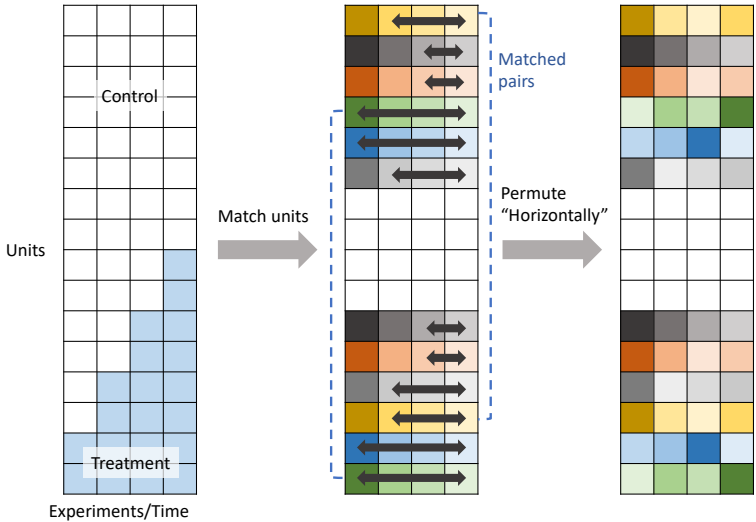
- "Difference-in-differences" test statistic: formalizes the intuition of our motivating example.
- Incorporate covariates.
- Matching: help reduce variance.



Extension to three or more experiments

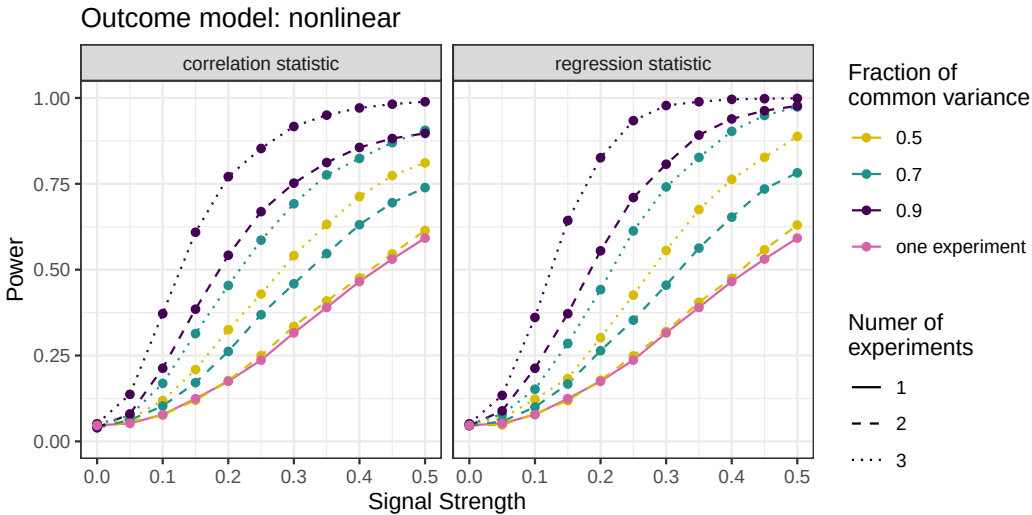


Extension to three or more experiments



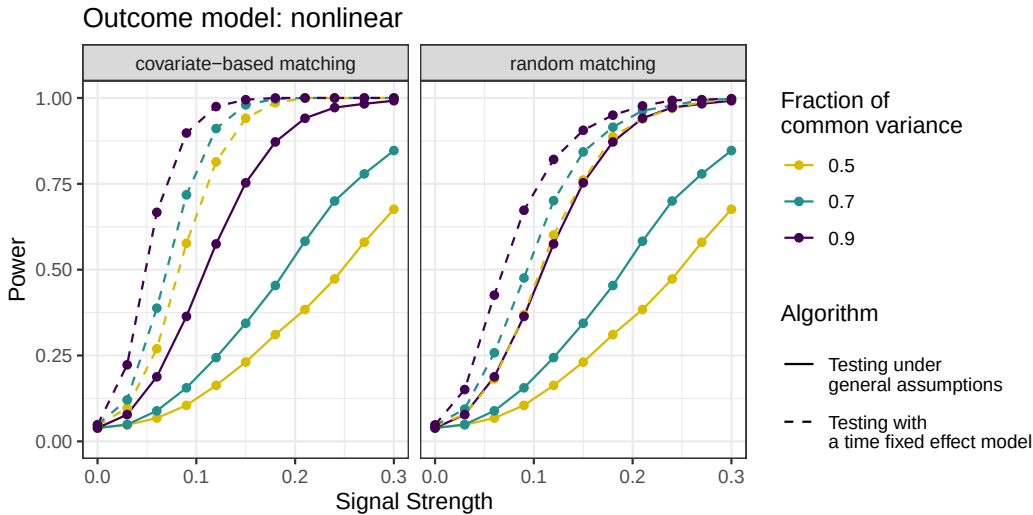
Simulations

- "Vertical Permutation"



Simulations

- "Vertical Permutation" vs "Horizontal Permutation"



Simulations

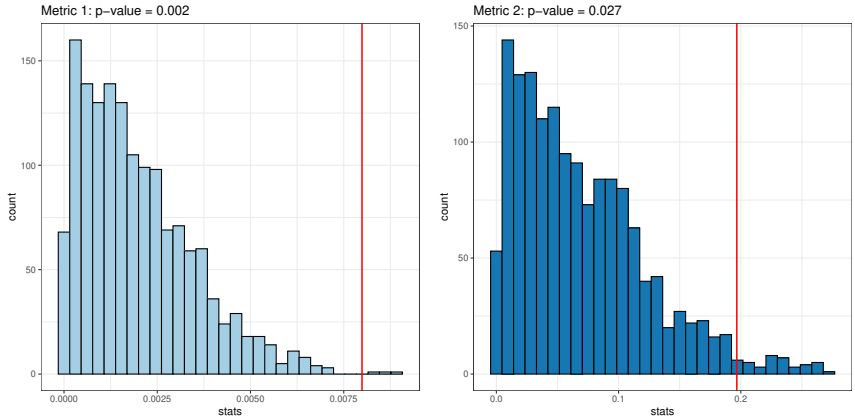
- Why is "Horizontal Permutation" more powerful than "Vertical Permutation"?
- Due to the nature of A/B tests, there is more variability in treatment allocation across experiments than across units.
- For example, assume that all units have around n_{ngb} neighbors in the social network. Looking at the fraction of neighbors in the treatment group, we find that the variation of this quantity across units is of scale $1/\sqrt{n_{\text{ngb}}}$, whereas the variation of this quantity across experiments is of constant scale.
  - By permuting over data points that are more different, "Horizontal Permutation" gains extra power.

Applications

- As an illustration, we test the method on a setting where we believe interference exists.
- Treatment: a new feature that improves the quality of LinkedIn members' attribute for ads targeting.
- Members as the randomization units.
- Interference effect is expected in these experiments:
  - When the allocation percentage is small, only a small set of members have the updated attributes, making them easier to be targeted by ad campaigns. Thus, when comparing metrics such as total ad impressions, these members tend to have larger average results than members in the control group.
  - When the treatment allocation increases, more members get the improved attributes. Since the total ad budget does not increase much, the average difference between treatment and control units becomes smaller.

Applications

- Horizontal permutation with 10% and 25% iterations.



Thank you!