ML-assisted Randomization Tests for Complex Treatment Effects

Wenxuan Guo*[1] JungHo Lee*[2] Panos Toulis[1]

[1]University of Chicago, Booth School of Business [2]Carnegie Mellon University

Highlights

We propose a new approach for testing complex treatment effects that combines **modern machine learning (ML) tools** with **randomization tests**.

1. **Complex effects:** We test complex effects such as heterogeneous effects and spillovers.
2. **ML test statistic:** We leverage ML-based test statistics, harnessing their predictive power.
3. **Finite-sample validity:** Randomization-based testing framework offers finite-sample validity.

Setup

Data

- $Z = (Z_1, \dots, Z_n) \sim P$ as binary treatments.
- $Y = (Y_1, \dots, Y_n)$ as vector of outcomes.
- $X_1, \dots, X_n$ as length-p vector of covariates.
- G: $n \times n$ adjacency matrix.

General model

$$Y_i = f_{base}(X_i) + \tau Z_i + f_{het}(X_i) Z_i + f_{sp}(X_i, Z_{-i}) + \varepsilon_i$$

f_{base}, f_{het}, and f_{sp} capture **baseline, heterogeneous, and spillover effects**, respectively.

Hypotheses of interest

1. Global null of no treatment effect

$$H_0^{glob}: \tau = 0, f_{het} = 0, f_{sp} = 0,$$
$$H_1^{glob}: \tau \neq 0, f_{het} \neq 0, f_{sp} \neq 0.$$

2. Heterogeneous treatment effects

$$H_0^{het}: \tau \neq 0, f_{het} = 0, f_{sp} = 0,$$
$$H_1^{het}: \tau \neq 0, f_{het} \neq 0, f_{sp} = 0.$$

3. Spillover effects

$$H_0^{sp}: \tau \neq 0, f_{het} \neq 0, f_{sp} = 0,$$
$$H_1^{sp}: \tau \neq 0, f_{het} \neq 0, f_{sp} \neq 0.$$

Backbone: Randomization inference

Classical Fisher randomization test (FRT)

1. **Compute observed value** $t^{\text{obs}} = t_n(Z, Y, X)$
2. **Draw** $Z' \sim P$ **and impute** $Y = Y(Z')$
3. **Obtain p-value**

$$p = \mathbb{E}[\mathbb{1}\{t_n(Z', Y, X) > t^{\text{obs}}\}],$$

Expectation is with respect to randomization distribution P.

Conditional FRT

e.g., Basse et al. (2019), Athey et al. (2018)

- Performed on a subset called "focal units"
- Select focal units $\mathcal{I}$, draw $Z' \sim P$ with $Z'_i = Z_i$ for $i \in \mathcal{I}$, and apply classical FRT

$\rightarrow$ Exact test under interference.

ML-FRT procedure

Procedure

1. **Fit ML models:** Fit two ML models, with (*full*) and without (*reduced*) complex effect.
2. **Compute CV-statistic:** Define test statistic as difference in cross-validated errors:
$$t_n = CV_{full} - CV_{reduced}.$$
3. **Obtain p-value:** Apply randomization test.

Under each hypothesis, ML-FRT procedure reduces to:

Global null: Classical FRT with test statistic

$$\text{CV}(Y; X, Z) - \text{CV}(Y; X).$$

$\rightarrow$ Captures variation explained by Z.

Heterogeneity: Let $p(\tau)$: p-value from global null with augmented outcome $Y^\tau = Y - \tau Z$. Define

$$p_\gamma := \sup_{\tau \in CI_\gamma} p(\tau) + \gamma, \quad \gamma \in (0, \alpha),$$

CI_γ: $(1-\gamma)$-significant confidence interval for ATE.

$\rightarrow$ Captures variation explained by f_{het}.

Interference: Conditional FRT with test statistic

$$\text{CV}(Y; Z, [GZ]_{\mathcal{I}}, X) - \text{CV}(Y; Z, [G]_{\mathcal{I}} Z_{\mathcal{I}}, X).$$

$\rightarrow$ Captures variation explained by non-focal units.

Validity & Power

Validity

Under H_0^h with $h \in \{glob, het, sp\}$, p-value of ML-FRT satisfies

$$\mathbb{P}(\text{pval} \leq \alpha) \leq \alpha,$$

for any $\alpha \in [0, 1]$ and any $n > 0$.

Power analysis

- Type-II error under H_1^{glob}.
- Applicable to general models with function class $\mathcal{F}$.

Suppose the data are i.i.d. and a "boundedness" assumption holds. If $\Delta > 0$, our type II error satisfies

$$\mathbb{P}(\text{pval} > \alpha) = \mathcal{O}\left(k \exp\left(-\frac{0.003 n \Delta^2}{k M^4}\right)\right).$$

- k: number of folds in cross-validation.
- M: boundedness constant.
- Δ: **"signal-to-noise"** difference

$$:= \underbrace{\inf_{f \in \mathcal{F}} \mathbb{E}(Y_i - f(X_i, Z'_i))^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(Y_i - f(X_i, Z_i))^2}_{\text{improvement in prediction}} - \underbrace{8\mathcal{R}_{n-n/k}(\mathcal{F})}_{\text{estimation error}}$$

- $\mathcal{R}_n(\mathcal{F})$: Rademacher complexity

$$:= \frac{1}{n} \mathbb{E}_{X,Z,\sigma}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(X_i, Z_i) \right| \right)$$

$\rightarrow$ measures the "size" of $\mathcal{F}$.

better prediction $\Rightarrow$ larger $\mathbf{\Delta}$ $\Rightarrow$ higher power!

	τ^S	0.2	0.4	0.5	1.0
$\mathcal{F}_{ML}$	Power	0.200	0.850	0.950	1.000
	$\widehat{\Delta}$	0.053	0.234	0.379	1.939
$\mathcal{F}_{LM}$	Power	0.075	0.225	0.150	0.325
	$\widehat{\Delta}$	0.005	0.037	0.046	0.371

Table 1. Power and $\widehat{\Delta}$ under different alternatives.

Example 1: Testing for heterogeneity

Bernoulli design with $n = 100$, $p = 5$, $X_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$, where Σ is a randomly generated correlation matrix. Assume no interference ($f_{sp} = 0$) and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. We test heterogeneity by varying τ^S.

- **Simple:** $\tau = 0$,

$$f_{base}(X_i) = -0.05 X_i^\top \beta_0, \qquad \beta_0 \sim \text{U}([1,5]^d),$$
$$f_{het}(X_i) = 0.5 \tau^S X_i^\top \beta_1, \qquad \beta_1 \sim \text{U}([1,5]^d).$$

- **Complex:** $\tau = 1$,

$$f_{base}(X_i) = -0.5(2\mathbb{1}\{X_{i1} < 0.5\} - 3\mathbb{1}\{X_{i2} > -0.5\}),$$
$$f_{het}(X_i) = \tau^S(2\mathbb{1}\{X_{i1} < 0.5\} - 3\mathbb{1}\{X_{i2} > -0.5\}).$$

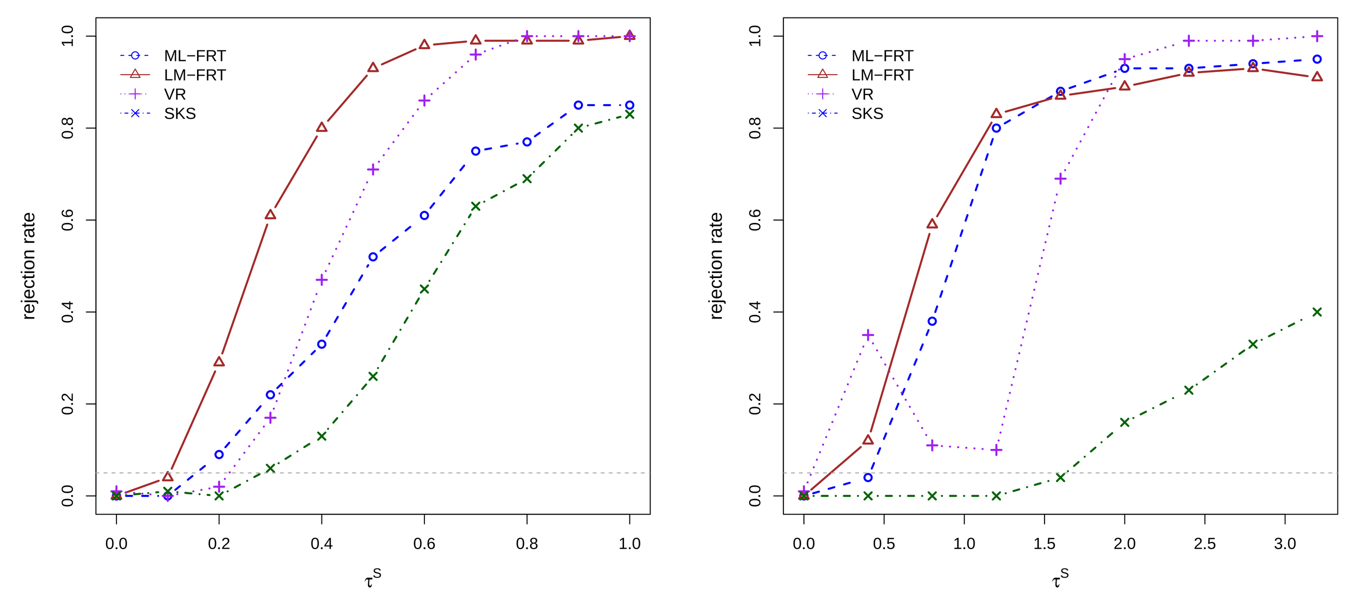

Figure 1. (left) Rejection rates under the "simple" DGP (right) rejection rates under the "complex" DGP. We compare our methods to the variance ratio ("VR") and shifted KS statistic ("SKS") from Ding et al. (2016).

- All methods achieve Type-I error control.
- ML-FRT showcases highest power under complex heterogeneous effects.

Example 2: Testing for spillovers

Two-stage experiment from Basse and Feller (2018) under clustered interference with $n = 300$ and 20 clusters. Consider $p = 2$ and $X_i \overset{iid}{\sim} \mathcal{N}(0, I_2)$.

Cluster-level potential outcomes:

$Y_{c_i,0} \sim \mathcal{N}(2, 0.1^2)$ and $Y_{c_i,2} \sim \mathcal{N}(Y_{c_i,0} + 1.5, 0.1^2)$.

Individual-level potential outcomes:

- **Simple: Constant spillover effects**

$Y_i(0) \sim \mathcal{N}(Y_{c_i,0}, 0.5^2), \quad Y_i(1) = Y_i(0) + \tau^S, \quad Y_i(2) \sim \mathcal{N}(Y_{c_i,1}, 0.5^2).$

- **Complex: Nonlinear spillover effects**

$Y_i(0) \sim \mathcal{N}(Y_{c_i,0}, X_{i1}^2/3^2),$
$Y_i(1) = Y_i(0) + \tau^S(3\mathbb{1}\{X_{i2} > -0.5\} - 2\mathbb{1}\{X_{i1} < 0.5\}),$
$Y_i(2) \sim \mathcal{N}(Y_{c_i,2}, X_{i2}^2/2^2).$

$\rightarrow f_{base}(X_i) = 2$, $\tau = 1.5$, $f_{het} = 0$, with different f_{sp}.

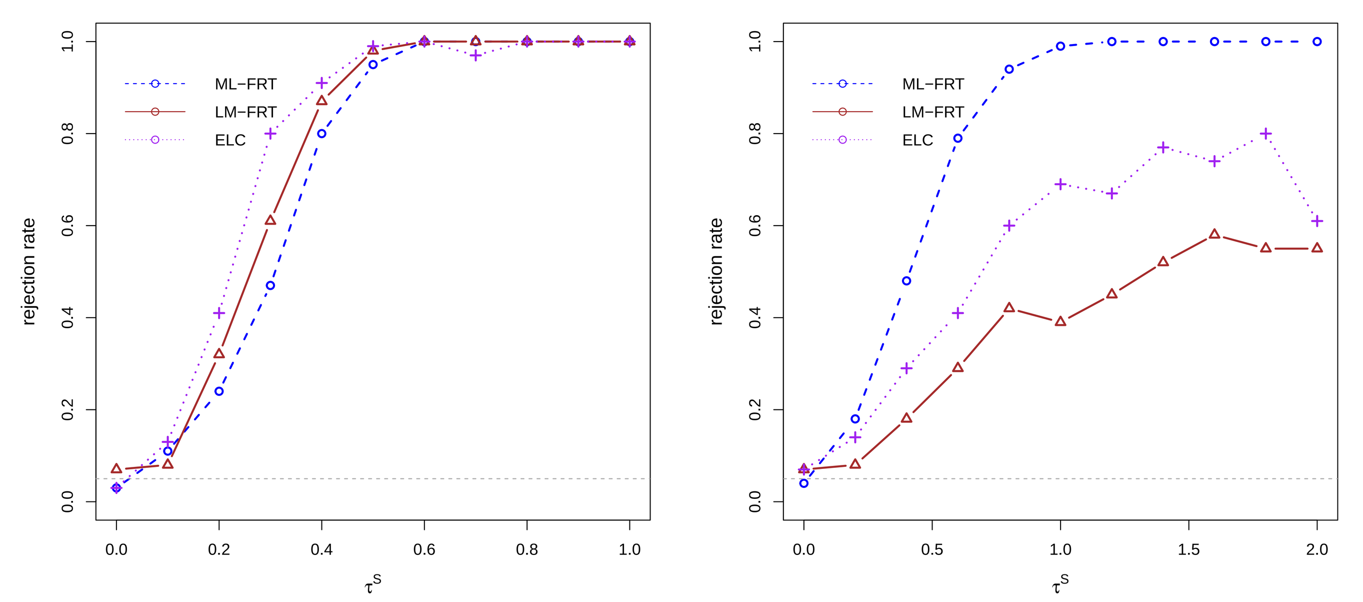

Figure 2. (left) Rejection rates under the "simple" DGP (right) rejection rates under the "complex" DGP. We compare our methods to the edge-level constrast statistic ("ELC") from Athey et al. (2018).

- All methods display similar power under "simple".
- Only ML-FRT maintains power under "complex".

Athey et al., Exact *p*-values for network interference, 2018.
Basse et al., Randomization tests of causal effects under interference, 2019.
Basse and Feller, Analyzing two-stage experiments in the presence of interference, 2018.
Ding et al., Randomization inference for treatment effect variation, 2016.