FisherBART: integrating causal machine learning and randomization test

JungHo Lee[1] David Puelz[2] Panos Toulis[1]

[1]University of Chicago [2]University of Texas at Austin

Highlights

We propose a new approach for assessing ***causal estimates*** from machine learning models using a ***randomization*** testing framework.

1. **Testing for Weak Nulls:** We model weak null hypotheses as a collection of sharp nulls of ITEs.
2. **Causal ML:** We produce estimates of ITEs conditional on the weak null.
3. **Randomization Test:** We test the weak null using a randomization-based method.

Setup

Data

$Z = (Z_1, \dots, Z_N)$ as binary population treatment;
$Y = (Y_1, \dots, Y_N)$ as population vector of outcomes;
$X = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$ as $N \times p$ matrix of covariates.

Potential outcome of unit i under z_i: $Y_i(z_i)$.

No interference: $Y_i(z_i)$ depends only on z_i.
⇒ Only two potential outcomes, $Y_i(0), Y_i(1), \forall\ i$.

Individual treatment effect: $\boldsymbol{\tau} \in \mathbb{R}^N$,

$$\tau_i = Y_i(1) - Y_i(0), \quad \forall i.$$

Weak null hypothesis

$$\mathbf{H_0} : g(\boldsymbol{\tau}, \theta) = 0 \text{ where } \theta \text{ is a scalar,}$$
$$g \text{ is a known function.}$$

Examples:

1. Sample average treatment effect (SATE):

$$g(\boldsymbol{\tau}, \theta) = \frac{1}{N}\sum_{i=1}^{N} \tau_i - \theta$$

2. No treatment heterogeneity:

$$g(\boldsymbol{\tau}, \theta) = \boldsymbol{\tau} - \theta\mathbf{1}$$

$\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^N$.

Fisher randomization test

Choose a test statistic $t(z, y, X) : \{0,1\}^N \times \mathbb{R}^N \times \mathbb{R}^{N \times p} \mapsto \mathbb{R}$. Suppose it is imputable under H_0, i.e., $t(z, Y(z), X) = t(z, Y(z'), X)$ for all $z, z' \in \{z \in \{0,1\}^N \mid P(z) > 0\}$.

1. **Draw** $Z^{\text{obs}} \sim P(Z^{\text{obs}})$, **observe** $Y = Y(Z^{\text{obs}})$
2. **Compute test statistic** $T^{\text{obs}} = t(Z^{\text{obs}}, Y^{\text{obs}}, X)$
3. **Obtain p-value**

$$p = \mathrm{E}[\mathbb{1}\{t(Z, Y^{\text{obs}}, X) > T^{\text{obs}})\}]$$

where the expectation is with respect to the randomization distribution $P(Z^{\text{obs}})$.

We formalize FRT as a **function** from $\mathbb{R}^N$ to $[0,1]$ dependent on data $D = (Z^{\text{obs}}, Y^{\text{obs}}, X)$ and t,

$$\mathrm{FRT}(\boldsymbol{\tau} \mid D, t) = p.$$

Note: depending on the data (experimental or observational) and the weak null, the design will either be given or need to be estimated. The test statistic must also be chosen carefully (see **Example 2**).

FisherBART procedure

Method:

1. **Fit a causal ML model:** Obtain posterior $p(\boldsymbol{\tau} \mid D)$ using a causal machine learning model
2. **Compute hypothesis conditional distribution:** $p(\boldsymbol{\tau} \mid H_0, D)$, given H_0
3. **Define confidence set:** C^g_β, a $1 - \beta$ confidence set for $\boldsymbol{\tau} \sim p(\boldsymbol{\tau} \mid H_0, D)$
4. **Obtain p-value:**

$$p^g_\beta := \sup_{\boldsymbol{\tau} \in C^g_\beta} \mathrm{FRT}(\boldsymbol{\tau} \mid D, t) + \beta, \quad \beta \in (0, \alpha).$$

Remarks:

- One could infer posterior $p(\boldsymbol{\tau} \mid D)$ from Bayesian models, e.g. from **Hahn** et al. (2020), **Hill** (2011).
- Or potentially non-Bayesian, **Athey** et al. (2019).
- We use the procedure of **Berger and Boos** (1994) to produce a valid, aggregate p-value.
- Computing the hypothesis conditional distribution $p(\boldsymbol{\tau} \mid H_0, D)$ in general remains a challenge.
- Other randomization-based testing for weak nulls:
  - Wu and Ding (2021), Fogarty (2019). Utilizing studentized test statistics ⇒ relies on asymptotics.
  - Ding et al. (2016), Nolen and Hudgens (2011). Given an appropriate nuisance parameter, computing a p-value as the supremum over the nuisance parameter space ⇒ underpowered when nuisance is high-dimensional.
  - Basse et al. (2019), Athey et al. (2018). Restricting tests to a subset of units and assignments ⇒ problem structure specific.

Why is this valid?

First, define $\boldsymbol{\tau_0}$ be the true but unknown vector of ITEs.

We want to show $E[\mathbb{1}\{p^g_\beta \le \alpha\} \mid H_0] \le \alpha$. Dropping the conditioning on H_0 for simpler notation,

Proof:

$$\begin{aligned} E[\mathbb{1}\{p^g_\beta \le \alpha\}] &= P(p^g_\beta \le \alpha) \\ &\le P(p^g_\beta \le \alpha \mid \boldsymbol{\tau_0} \in C^g_\beta) + P(\boldsymbol{\tau_0} \notin C^g_\beta) \\ &\le P(\mathrm{FRT}(\boldsymbol{\tau_0} \mid D, t) + \beta \le \alpha) + \beta \\ &= P(\mathrm{FRT}(\boldsymbol{\tau_0} \mid D, t) \le \alpha - \beta) + \beta \\ &\le \alpha - \beta + \beta = \alpha \end{aligned}$$

The second inequality comes from the fact that with $\boldsymbol{\tau_0} \in C^g_\beta$, $\mathrm{FRT}(\boldsymbol{\tau_0} \mid D, t) \le p^g_\beta$ by definition of p^g_β. This demonstrates the procedure is a **valid** method for testing the weak null hypothesis $\mathbf{H_0}$.

Example 1: Testing for SATE

We simulate a scenario for testing a hypothesis on SATE. The DGP is given by: $x_i \sim \text{unif}(0,1)$, $Z_i \sim \text{Ber}(0.5)$, $Y_i(0) \sim N(2,1)$, and $\tau_i \sim N(1, 4x_i^2\tau^2_{\text{spread}})$.

We generate 2000 data sets from this DGP with $N = 100$ and $\tau_{\text{spread}} = 0.5$. For each data set, we estimate SATE using the proposed procedure (new) and Bayesian causal forests (bcf). We use the difference-in-means statistic for FRT. An estimate and confidence interval are produced by inverting our test.

method	bias	rmse	cover	IL
new (oracle)	-0.002	0.211	0.972	0.959
new (bcf)	-0.003	0.211	0.961	0.896
bcf	-0.085	0.240	0.925	0.870

Table 1. All metrics are computed with respect to the true sample average treatment effect and are averaged across 2000 realization of DGP

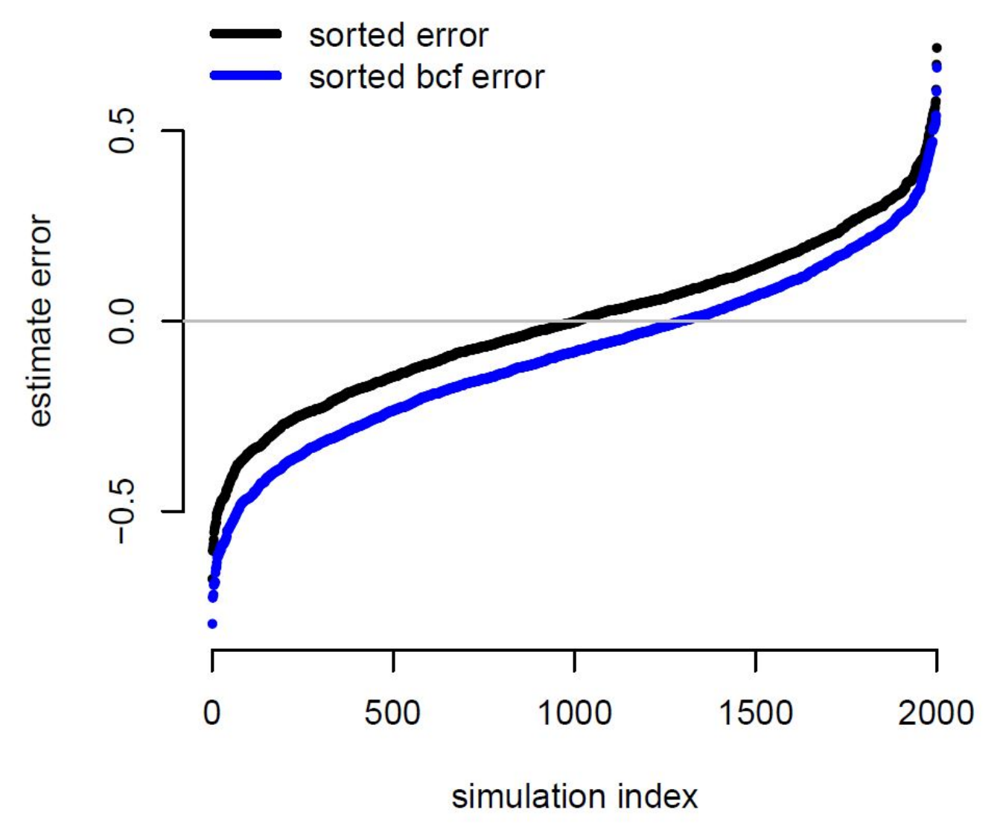

Figure 1. Visual depiction of the bias results for the “new (oracle)” (black) and “bcf” (blue) methods.

The new (oracle) method that uses ITE moments has the smallest bias, conservative coverage, and the largest interval length. The new (bcf) method with moments estimated from the bcf posterior improves greatly upon bcf.

Example 2: Testing for heterogeneity

Here, we test the hypothesis that asks $g(\boldsymbol{\tau}, \theta) = \boldsymbol{\tau} - \mathbf{1}\theta$ equals zero for all $\theta \in \mathbb{R}$.

We generate 200 data sets from the same DGP from Example 1 varying τ_{spread}. Larger values of τ_{spread} should result in us rejecting the weak null more frequently. We plot “model-based” distributions of the p-values and also the percentage of rejections (power). A function providing an estimate of the variance of the ITEs is used as the test statistic.

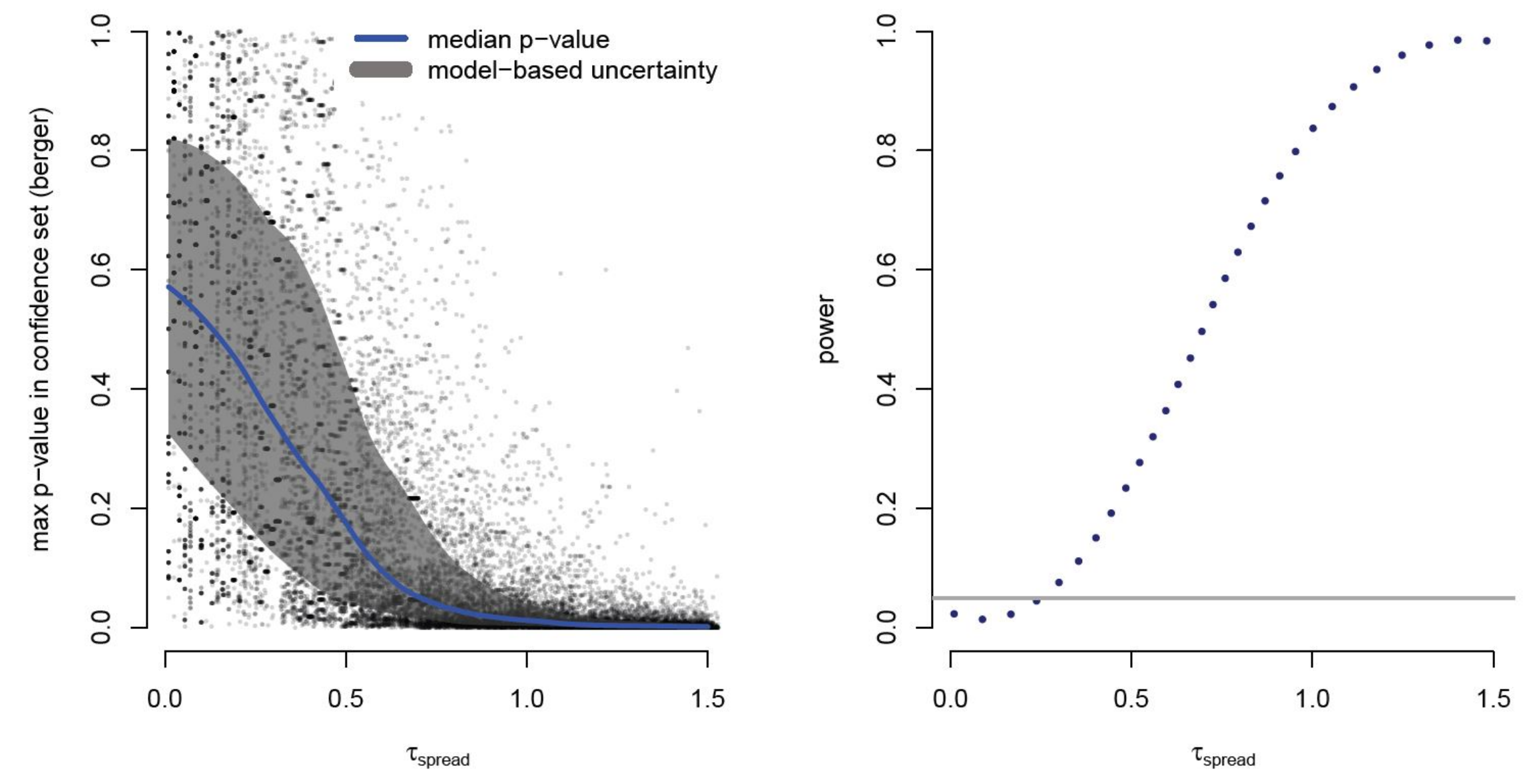

Figure 2. (left) Model-based distributions of the test’s p-values for varying values of τ_{spread}. (right) Power of the test.

As hoped, the test fails to reject for small values of τ_{spread} and rejects for large values. But,

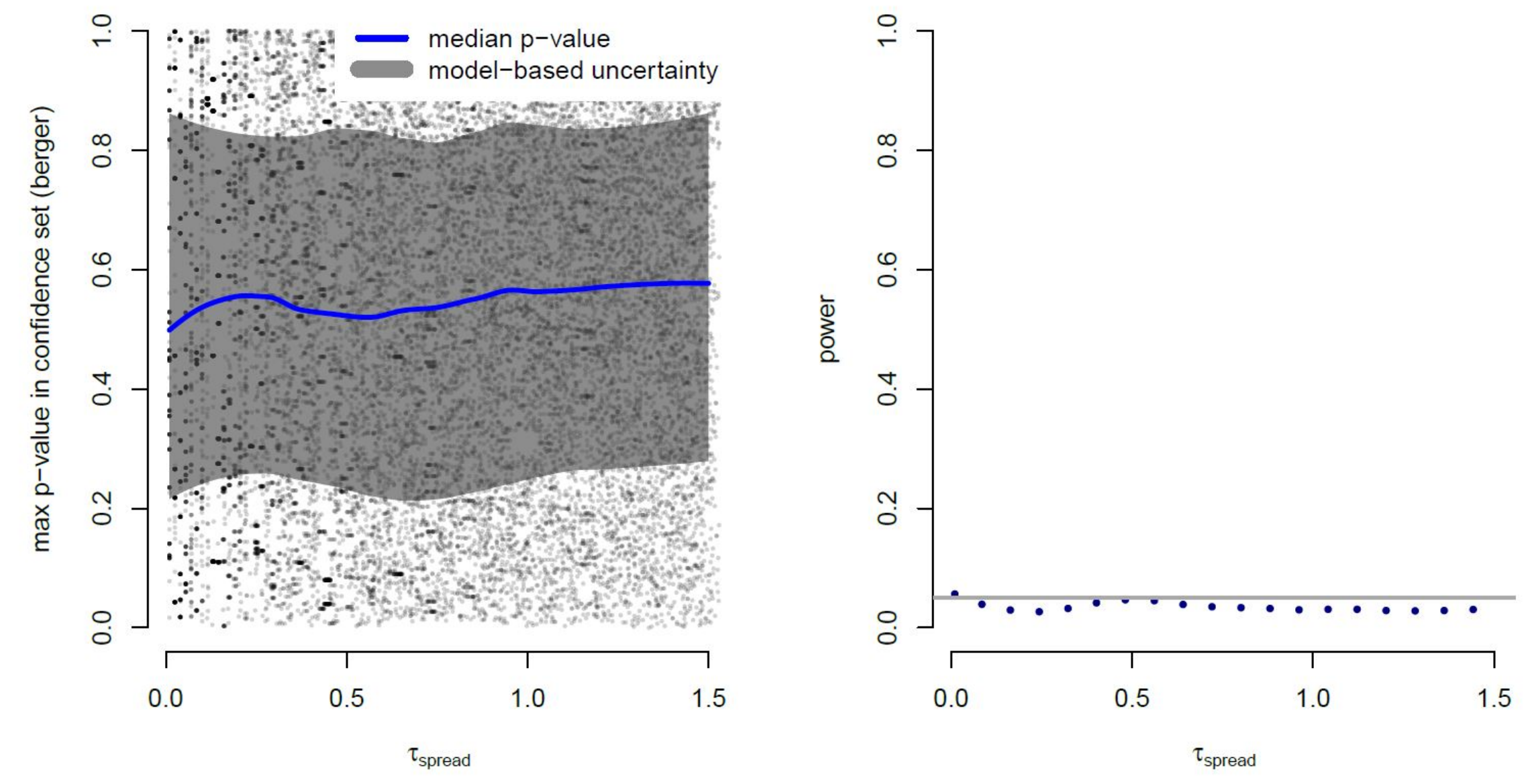

Figure 3. (left) Model-based distributions of the test’s p-values for varying values of τ_{spread}. (right) Power of the test. Difference-in-means as the test statistic for the FRT instead.

Difference-in-means, *not* a function of the covariates, results in a powerless test of the weak null! ⇒ Choosing a proper test statistic is **crucial** when treatment effect variation is derived from the covariates.