JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2025, VOL. 43, NO. 1, 256–268
https://doi.org/10.1080/07350015.2024.2358909

Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments

Kosuke Imai[a] and Michael Lingzhi Li[b]

[a]Department of Government and Department of Statistics, Harvard University, Cambridge, MA; [b]Technology and Operations Management, Harvard Business School, Boston, MA

ABSTRACT

Researchers are increasingly turning to machine learning (ML) algorithms to investigate causal heterogeneity in randomized experiments. Despite their promise, ML algorithms may fail to accurately ascertain heterogeneous treatment effects under practical settings with many covariates and small sample size. In addition, the quantification of estimation uncertainty remains a challenge. We develop a general approach to statistical inference for heterogeneous treatment effects discovered by a generic ML algorithm. We apply the Neyman's repeated sampling framework to a common setting, in which researchers use an ML algorithm to estimate the conditional average treatment effect and then divide the sample into several groups based on the magnitude of the estimated effects. We show how to estimate the average treatment effect within each of these groups, and construct a valid confidence interval. In addition, we develop nonparametric tests of treatment effect homogeneity across groups, and rank-consistency of within-group average treatment effects. The validity of our methodology does not rely on the properties of ML algorithms because it is solely based on the randomization of treatment assignment and random sampling of units. Finally, we generalize our methodology to the cross-fitting procedure by accounting for the additional uncertainty induced by the random splitting of data.

ARTICLE HISTORY
Received May 2023
Accepted May 2024

KEYWORDS
Causal heterogeneity; Causal inference; Conditional average treatment effect; Cross-fitting; Randomization inference; Sample splitting

1. Introduction

A growing number of researchers are turning to machine learning (ML) algorithms to uncover causal heterogeneity in randomized experiments. ML algorithms are appealing because in many applications the structure of heterogeneous treatment effects is unknown. Despite their promise, however, relatively little theoretical properties have been established for many of these algorithms. In addition, the choice of tuning parameter values remains to be often difficult and consequential in practice. As a result, ML algorithms may fail to ascertain heterogeneous treatment effects under common settings with many covariates and small sample size. Furthermore, one major challenge is the quantification of statistical uncertainty when estimating heterogeneous treatment effects using ML algorithms.

In this article, we develop a general approach to statistical inference for heterogeneous treatment effects estimated through the application of a generic ML algorithm to experimental data. We apply the Neyman's (1923) repeated sampling framework to a common setting, in which researchers use ML algorithms to estimate the conditional average treatment effect (CATE) given pre-treatment covariates and then divide the sample into several groups based on the magnitude of these estimated effects. We show how to obtain a consistent estimate of the average treatment effect within each of these groups—the sorted group average treatment effect (GATES; Chernozhukov et al. (2023))—and construct an asymptotically valid confidence interval.

We also propose two nonparametric tests of treatment effect heterogeneity that are of interest to applied researchers. First, we test whether there exists any treatment effect heterogeneity across groups. Second, we develop a statistical test of the rank-consistency of GATES. If an ML algorithm produces a reasonable scoring rule, the rank ordering of GATES based on their magnitude should be monotonic. To accommodate the use of various ML algorithms, we make no assumption about their properties. Specifically, ML algorithms do not have to be consistent or unbiased. This is possible because the validity of our confidence intervals and nonparametric tests solely depends on the randomization of treatment assignment and random sampling of units. Thus, our approach imposes only a minimal set of assumptions on the underlying data generating process.

We first consider the setting, in which an external dataset is used to estimate the CATE. For example, researchers may apply an ML algorithm to an observational dataset. Alternatively, an experimental dataset may be split into the training and validation datasets where an ML algorithm is first applied to the training data to estimate the CATE, and the validation data is

CONTACT Michael Lingzhi Li *mili@hbs.edu* Technology and Operations Management, Harvard Business School, Boston, MA 02163.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

then used to estimate the GATES. Here, we treat the estimated CATE function as fixed and do not account for the uncertainty that arises from its estimation.

To incorporate this additional source of uncertainty, we further generalize our methodology to the cross-fitting procedure, which randomly splits the data into multiple folds. Each fold is used as the validation data to estimate the GATES while the remaining folds serve as the corresponding training data to estimate the CATE. After repeating this for each fold, we aggregate the GATES estimates to the entire sample. Unlike the sample-splitting case where we condition on the split, we account for additional uncertainty induced by the randomness of its cross-fitting procedure. This directly addresses the fact that when the sample size is small the GATES estimate may vary considerably due to the random splitting of data.

Related Literature. The proposed methodology builds on the existing literature about statistical inference for heterogeneous treatment effects. In an early work, Crump et al. (2008) propose nonparametric tests of treatment effect heterogeneity. The authors rely on the consistency of sieve methods under the assumption that heterogeneous treatment effects are a smooth function of covariates. In contrast, our methodology does not require the consistent estimation of the CATE by ML algorithms. Moreover, while Crump et al. assume the continuous differentiability of the CATE, we only require its continuity.

Ding, Feller, and Miratrix (2016) propose an alternative approach based on Fisher's randomization test. Similar to our proposed methodology, this test neither requires modeling assumptions nor imposes restrictive assumptions on data generating process. In fact, their test yields conservative p-values without asymptotic approximation whereas other approaches including ours are only valid in large samples. The authors, however, test restrictive sharp null hypotheses. For example, Ding, Feller, and Miratrix (2016) consider a null hypothesis that the individual treatment effect is constant within each group and the effect only varies across groups. In contrast, we focus on the null hypotheses about average treatment effects that may vary within and across groups under the Neyman's repeated sampling framework. While our tests are valid only asymptotically, our simulation studies show that they perform reasonably well in small samples. In addition, Ding, Feller, and Miratrix (2019) use the Neyman's repeated sampling framework to explore treatment effect heterogeneity like we do, but rely entirely on the linear regression and does not allow for the use of more flexible ML algorithms.

More recently, Chernozhukov et al. (2023) study the settings that are identical to the ones considered in this article. Similar to our methodology, the authors do not impose strong assumptions on the properties of ML algorithms that are used to estimate the CATE. However, to incorporate the additional uncertainty of the cross-fitting procedure, Chernozhukov et al. (2023) propose to repeat the procedure many times and aggregate the resulting p-values. We avoid such a computationally intensive procedure and instead use the Neyman's repeated sampling framework to conduct valid statistical inference under cross-fitting. In simulation studies reported elsewhere (Imai and Li 2023a), we show that our confidence intervals are less conservative than those proposed by Chernozhukov et al. (2023) in finite samples.

Other researchers also have considered GATES and related quantities. For example, Yadlowsky et al. (2021) establish the asymptotic properties for a related general class of metrics that summarize the effect of treatment prioritization rules. In addition to the different focus, the authors assume that a treatment prioritization rule of interest is fixed and do not consider the uncertainty that arises from its estimation. Dwivedi et al. (2020) also estimate the GATES to explore treatment effect heterogeneity and develop calibration methods. However, they do not derive the asymptotic distribution of GATES and hence stop short of providing formal statistical methods.

Finally, Imai and Li (2023b) show how to evaluate an individualized treatment rule derived from the application of a generic ML algorithm in general settings including the one based on cross-fitting. We build on this work and derive the asymptotic properties of the GATES estimator. Imai and Li (2023c) further extends the methodology proposed in this article and develop uniform asymptotic confidence bands. This allows researchers to choose, with a statistical guarantee, a group of individuals who are predicted to benefit from or be harmed by the treatment, using the estimated CATE based on a generic ML algorithm. They do not, however, consider the estimation uncertainty of the CATE.

2. The Proposed Methodology

We start by developing our methodology in a setting where the conditional average treatment effect (CATE) function is estimated using a separate dataset, but is considered fixed when estimating the sorted group average treatment effect (GATES) and conducting statistical tests. For instance, the estimated CATE might come from an external, possibly observational, dataset. An alternative is *sample splitting*, where the sample is divided randomly into training and evaluation sets. The training data is used for CATE estimation via a machine learning algorithm, and the evaluation data for GATES estimation. In this section, we do not account for the uncertainty in estimating the CATE. In Section 3, we extend our methodology to cross-fitting, incorporating this estimation uncertainty.

2.1. Setup

Suppose that we have an iid sample of n units from a super-population $\mathcal{P}$. Let T_i represent the treatment assignment indicator variable, which is equal to 1 if unit i is assigned to the treatment condition and is equal to 0 otherwise, that is, $T_i \in \mathcal{T} = \{0, 1\}$. For each unit, we observe the outcome variable $Y_i \in \mathcal{Y}$ and a vector of pre-treatment covariates, $\boldsymbol{X}_i \in \mathcal{X}$, where $\mathcal{Y}$ and $\mathcal{X}$ represent the support of the outcome variable and that of the pre-treatment covariates, respectively.

We require the standard causal inference assumptions of consistency and no interference between units, denoting the potential outcome for unit i under the treatment condition $T_i = t$ as $Y_i(t)$ for $t = 0, 1$ (e.g., Neyman 1923; Holland 1986; Rubin 1990). The observed outcome is given by $Y_i = Y_i(T_i)$. For notation simplicity, we assume that the treatment assignment

is completely randomized with exactly n_1 units assigned to the treatment condition though the extensions to other experimental designs and unconfounded observational designs are possible. We formally state these assumptions below.

Assumption 1 (*No Interference between Units*). The potential outcomes for unit i do not depend on the treatment status of other units. That is, for all $t_1, t_2, \ldots, t_n \in \{0,1\}$, we have, $Y_i(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n) = Y_i(T_i = t_i)$.

Assumption 2 (*Random Sampling of Units*). Each of n units, represented by a three-tuple consisting of two potential outcomes and pre-treatment covariates, is assumed to be independently sampled from a super-population $\mathcal{P}$, that is,

$$(Y_i(1), Y_i(0), \boldsymbol{X}_i) \overset{\text{iid}}{\sim} \mathcal{P}$$

Assumption 3 (*Complete Randomization*). For any $\boldsymbol{t} \in \{0,1\}^n$ such that $\sum_{i=1}^n \boldsymbol{t}_i = n_1$, the treatment assignment probability is given by,

$$\Pr(\boldsymbol{T} = \boldsymbol{t} \mid \{Y_{i'}(1), Y_{i'}(0), \boldsymbol{X}_{i'}\}_{i'=1}^n) = \frac{1}{\binom{n_1}{n}}.$$

Suppose that a researcher applies an ML algorithm to a training dataset and estimate the CATE. As noted earlier, this training dataset can be obtained through the sample splitting or it may be an external dataset. The CATE is defined as,

$$\tau(\boldsymbol{x}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}),$$

for any $\boldsymbol{x} \in \mathcal{X}$. The ML algorithm produces the following scoring rule,

$$s : \mathcal{X} \longrightarrow \mathcal{S} \subset \mathbb{R} \tag{1}$$

where a greater score indicates a higher priority to receive the treatment. Without loss of generality, we assume that the scoring rule is bijective, that is, $s(\boldsymbol{x}) \neq s(\boldsymbol{x}')$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ with $\boldsymbol{x} \neq \boldsymbol{x}'$. Note that one can always redefine $\mathcal{X}$ to satisfy this assumption.

As noted earlier, we assume almost nothing about the properties of this scoring rule derived by the ML algorithm. In particular, the scoring rule does not have to be a consistent estimate of CATE. In fact, the scoring rule need not even be an estimate of CATE so long as it satisfies the definition given in (1).

2.2. Estimation and Inference

Given the setup introduced above, we first consider the estimation and inference for the sorted group average treatment effect (GATES), which is a common quantity of interest in applied research and is studied by Chernozhukov et al. (2023). The idea is that researchers sort units into a total of K groups based on the quantile of the scoring rule, and then estimate the average treatment effect within each group. For simplicity, we assume that the number of treated and control units, that is, n_1 and n_0, are multiples of K. The formal definition of GATES is given by,

$$\tau_k = \mathbb{E}(Y_i(1) - Y_i(0) \mid c_{k-1}(s) < s(\boldsymbol{X}_i) \leq c_k(s)) \tag{2}$$

for $k = 1, 2, \ldots, K$ where c_k represents the cutoff between the $(k-1)$th and kth groups and is defined as,

$$c_k(s) = \inf\{c \in \mathbb{R} \mid \Pr(s(\boldsymbol{X}_i) \leq c) \geq k/K\},$$

for $k = 1, 2, \ldots, K$, with $c_0 = -\infty$. Equivalently, GATES can be seen as a special case of the rank-weighted average treatment effect (RATE) with $\alpha(u) = \mathbf{1}\{\frac{k-1}{K} < u \leq \frac{k}{K}\}$ (Yadlowsky et al. 2021).

Thus, units that belong to the Kth group, for example, represent those who are likely to have the greatest treatment effect according to the ML algorithm whereas those in the first group are likely to have the least treatment effect. However, we do not assume that the GATES is monotonic, that is, $\tau_k \leq \tau_{k'}$ for all $k < k'$. This is important because we want to impose as little restriction on the underlying scoring rule as possible. Indeed, if the scoring rule is not a good estimate of CATE, such an assumption may be violated. To address this problem, we later develop a statistical test of this monotonicity assumption.

We consider the following estimator of GATES using the experimental data,

$$\hat{\tau}_k = \frac{K}{n_1} \sum_{i=1}^n Y_i T_i \hat{f}_k(\boldsymbol{X}_i) - \frac{K}{n_0} \sum_{i=1}^n Y_i (1 - T_i) \hat{f}_k(\boldsymbol{X}_i), \tag{3}$$

for $k = 1, 2, \ldots, K$ where $\hat{f}_k(\boldsymbol{X}_i) = \mathbf{1}\{s(\boldsymbol{X}_i) > \hat{c}_{k-1}(s)\} - \mathbf{1}\{s(\boldsymbol{X}_i) > \hat{c}_k(s)\}$, and $\hat{c}_k(s) = \inf\{c \in \mathbb{R} : \sum_{i=1}^n \mathbf{1}\{s(\boldsymbol{X}_i) \leq c\} \geq nk/K\}$ is the estimated cutoff. First, we derive the bias bound and exact variance of the GATES estimator.

Theorem 1 (*Bias Bound and Exact Variance of the GATES Estimator*). Under Assumptions 1–3, the bias of the proposed estimator of GATES given in (3) can be bounded as follows,

$$\begin{aligned}
&\mathbb{P}(|\mathbb{E}\{\hat{\tau}_k - \tau_k \mid \hat{c}_k(s), \hat{c}_{k-1}(s)\}| \geq \epsilon) \\
&\leq 1 - B\left(\frac{k}{K} + \gamma_k(\epsilon), \frac{nk}{K}, n - \frac{nk}{K} + 1\right) \\
&\quad + B\left(\frac{k}{K} - \gamma_k(\epsilon), \frac{nk}{K}, n - \frac{nk}{K} + 1\right) \\
&\quad - B\left(\frac{k-1}{K} + \gamma_{k-1}(\epsilon), \frac{n(k-1)}{K}, n - \frac{n(k-1)}{K} + 1\right) \\
&\quad + B\left(\frac{k-1}{K} - \gamma_{k-1}(\epsilon), \frac{n(k-1)}{K}, n - \frac{n(k-1)}{K} + 1\right),
\end{aligned}$$

for any given constant $\epsilon > 0$ where $B(\epsilon, \alpha, \beta)$ is the incomplete beta function (if $\alpha \leq 0$ and $\beta > 0$, we set $B(\epsilon, \alpha, \beta) := H(\epsilon)$ for all ϵ where $H(\epsilon)$ is the Heaviside step function), and

$$\gamma_k(\epsilon) = \frac{\epsilon}{K \max_{c \in [c_k(s) - \epsilon,\, c_k(s) + \epsilon]} \mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = c)}.$$

The variance of the estimator is given by,

$$\mathbb{V}(\hat{\tau}_k) = K^2 \left(\frac{\mathbb{E}(S_{k1}^2)}{n_1} + \frac{\mathbb{E}(S_{k0}^2)}{n_0}\right) + \frac{(n-K)\kappa_{k11}}{n-1} - \kappa_{k1}^2,$$

where $S_{kt}^2 = \sum_{i=1}^n (Y_{ki}(t) - \overline{Y_k(t)})^2/(n-1)$, $\kappa_{kt} = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{f}_k(\boldsymbol{X}_i) = t)$, and $\kappa_{ktt} = \mathbb{E}[(Y_i(1) - Y_i(0))(Y_j(1) - Y_j(0)) \mid \hat{f}_k(\boldsymbol{X}_i) = \hat{f}_k(\boldsymbol{X}_j) = t]$ for $i \neq j$ with $Y_{ki}(t) = \hat{f}_k(\boldsymbol{X}_i) Y_i(t)$, and $\overline{Y_k(t)} = \sum_{i=1}^n Y_{ki}(t)/n$, for $t = 0, 1$.

Proof is given in supplementary Appendix S1.

When compared to the standard variance estimator, there are additional two terms. These terms result from the fact that the cutoff points are estimated, yielding a cross-unit correlation in terms of $\hat{f}_k(\boldsymbol{X}_i)Y_i(t)$. Since exactly n/K data points are taken to have $\hat{f}_k(\boldsymbol{X}_i) = 1$, the value of this function is generally negatively correlated across units, that is, $\text{corr}(\hat{f}_k(\boldsymbol{X}_i), \hat{f}_k(\boldsymbol{X}_j)) < 0$.

The variance can be consistently estimated by replacing each unknown parameter with its sample analogue:

$$\widehat{\mathbb{E}(S_{kt}^2)} = \frac{1}{n_t - 1}\sum_{i=1}^{n} \mathbf{1}\{T_i = t\}(Y_{ki} - \overline{Y}_{kt})^2,$$

$$\hat{\kappa}_{kt} = \frac{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i}{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i} - \frac{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i}{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)},$$

$$\hat{\kappa}_{ktt} = \frac{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i^2}{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i}$$

$$+\frac{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i^2}{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)}$$

$$-2\frac{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i][\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i]}{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)][\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i]}.$$

for $t = 0, 1$ where $Y_{ki} = \hat{f}_k(\boldsymbol{X}_i)Y_i$ and $\overline{Y}_{kt} = \sum_{i=1}^n \mathbf{1}\{T_i = t\}Y_{ki}/n_t$. The expression of $\hat{\kappa}_{ktt}$ above enables the calculation in $O(n)$ rather than $O(n^2)$ time. The details of the derivation is given in Appendix S2.

We can further derive the asymptotic sampling distribution of the GATE estimator by requiring the following continuity assumption and moment conditions:

Assumption 4 (Continuity of CATE at the Thresholds). Let $F(c) = \Pr(s(\boldsymbol{X}_i) \leq c)$ represent the cumulative distribution function of the scoring rule and define its pseudo-inverse $F^{-1}(p) = \inf\{c : F(c) \geq p\}$ for $p \in [0, 1]$. The CATE function $\mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(p))$ is assumed to be left-continuous with bounded variation on any interval $(\theta, 1-\theta)$ with $\theta > 0$, and continuous in p at $p = 1/K, \ldots, (K-1)/K$.

Assumption 5 (Moment Conditions). For each $t = 0, 1$, we have

1. $\mathbb{V}(Y_i(t)) > 0$;
2. $\mathbb{E}(Y_i(t)^3) < \infty$.

Assumption 4 is similar to the assumption commonly used in the literature that the CATE is continuous in the covariates $\boldsymbol{X}_i$ (e.g., Künzel et al. 2018; Wager and Athey 2018), but we only require continuity at the thresholds, $1/K, \ldots, (K-1)/K$ and bounded variation everywhere else. We will show in Proposition 1 that Assumption 4 is among the weakest assumptions necessary for our asymptotic results. In particular, this assumption requires that the scoring rule cannot be discontinuous at the thresholds unless the CATE is constant in the scoring rule, that is $\mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(p)) = \mathbb{E}(Y_i(1) - Y_i(0))$ for all p.

We now present the asymptotic sampling distribution of GATES estimator.

Theorem 2 (Asymptotic Sampling Distribution of GATES Estimator). Under Assumptions 1–5, we have,

$$\frac{\hat{\tau}_k - \tau_k}{\sqrt{\mathbb{V}(\hat{\tau}_k)}} \xrightarrow{d} N(0, 1)$$

for $k = 1, \ldots, K$ where $\mathbb{V}(\hat{\tau}_k)$ is given in Theorem 1.

Proof is given in supplementary Appendix S3. We emphasize that Theorem 2 does not impose a strong assumption about the properties of the ML algorithm used to generate the scoring rule s.

In fact, the continuity of the CATE at the thresholds (Assumption 4) is among the weakest assumptions that can ensure the validity of Theorem 2. To see this, consider an alternative assumption that there exists a threshold at which CATE is bounded but discontinuous, slightly relaxing Assumption 4. The following proposition shows that this assumption is not sufficient for Theorem 2.

Proposition 1 (Insufficiency of Bounded Variation). Suppose Assumptions 1–3 and 5 hold. Further assume that the there exists a threshold k/K, such that $\mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(p))$, is discontinuous (but bounded) at $p = k/K$. Then, there exist a scoring rule s and a population $\mathcal{P}$ such that as $n \to \infty$ with $0 < n_1/n < 1$ staying constant, we have,

$$\mathbb{E}\left(\frac{\hat{\tau}_k - \tau_k}{\sqrt{\mathbb{V}(\hat{\tau}_k)}}\right) \not\to 0.$$

Proof is given in supplementary Appendix S4. Proposition 1 shows that if the CATE is mildly discontinuous at a threshold, then we cannot sufficiently control the bias in estimating the boundary points, $c_k(s)$. Under this scenario, the bias decays at the rate of $n^{-1/2}$, which is not fast enough for the application of the central limit theorem.

2.3. Nonparametric Test of Treatment Effect Heterogeneity

In many applications, heterogeneous treatment effects are imprecisely estimated. Researchers may wish to know whether the treatment effect heterogeneity discovered by ML algorithms represents signal rather than noise. In addition, checking the statistical significance of each GATES suffers from multiple testing problems. To address these challenges, we develop a nonparametric test of treatment effect heterogeneity. In particular, we consider the following null hypothesis that all GATEs are equal to one another,

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K. \tag{4}$$

This null hypothesis is equivalent to $\tau_k = \tau$ for any k where $\tau = \mathbb{E}(Y_i(1) - Y_i(0))$ represents the overall average treatment effect (ATE). Thus, we consider the following test statistic,

$$\hat{\boldsymbol{\tau}} = (\hat{\tau}_1 - \hat{\tau}, \ldots, \hat{\tau}_K - \hat{\tau})^\top,$$

where

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} Y_iT_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - T_i).$$

To derive the asymptotic reference distribution of this test statistic,

Imai and Li (2023b) derive the bias bound and the exact variance of this PAPE estimator. Leveraging those results, the following theorem shows that we can use a χ^2 distribution as an asymptotic approximation to the reference distribution when testing treatment effect heterogeneity.

Theorem 3 (Nonparametric Test of Treatment Effect Heterogeneity). Suppose Assumptions 1–5 hold. Under H_0 defined in (4) and against the alternative $H_1 : \mathbb{R}^K \setminus H_0$, as $n \to \infty$ with $0 < n_1/n < 1$ stays constant, we have,

$$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2_K$$

where the entries of the covariance matrix $\boldsymbol{\Sigma}$ are defined as follows,

$$\begin{aligned}
\boldsymbol{\Sigma}_{kk} &= K^2 \left[\frac{\mathbb{E}(S^{*2}_{k1})}{n_1} + \frac{\mathbb{E}(S^{*2}_{k0})}{n_0} \right. \\
&\qquad + \frac{1}{K^3} \left\{ (K-2) \left(\frac{n-K}{n-1} \kappa_{kk11} - \kappa^2_{k1} \right) \right. \\
&\qquad \left. \left. - \frac{2n(K-1)}{(n-1)} \kappa_{kk01} + 2\kappa_{k1}\kappa_{k0} \right\} \right], \\
\boldsymbol{\Sigma}_{kk'} &= K^2 \left\{ \frac{\mathbb{E}(S^{*2}_{kk'1})}{n_1} + \frac{\mathbb{E}(S^{*2}_{kk'0})}{n_0} \right\} \\
&\quad + \frac{1}{K} \left\{ (K-2) \left(\kappa_{kk'11} - \kappa_{k1}\kappa_{k'1} \right) \right. \\
&\qquad - \frac{Kn-n-1}{n-1} \left(\kappa_{kk'10} + \kappa_{kk'01} \right) \\
&\qquad \left. + \kappa_{k1}\kappa_{k'0} + \kappa_{k0}\kappa_{k'1} \right\},
\end{aligned}$$

for $k, k' \in \{1, \dots K\}$ and $k \neq k'$ where $S^{*2}_{kt} = \sum_{i=1}^n (Y^*_{ki}(t) - \overline{Y^*_k(t)})^2/(n-1)$, $S^{*2}_{kk't} = \sum_{i=1}^n (Y^*_{ki}(t) - \overline{Y^*_k(t)})(Y^*_{k'i}(t) - \overline{Y^*_{k'}(t)})/(n-1)$, $\kappa_{kt} = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{f}_k(\boldsymbol{X}_i) = t)$, and $\kappa_{kk'ts} = \mathbb{E}[(Y_i(1) - Y_i(0))(Y_j(1) - Y_j(0)) \mid \hat{f}_k(\boldsymbol{X}_i) = t, \hat{f}_{k'}(\boldsymbol{X}_i) = s]$ for $i \neq j$ with $Y^*_{ki}(t) = (\hat{f}_k(\boldsymbol{X}_i) - 1/K) Y_i(t)$, and $\overline{Y^*_k(t)} = \sum_{i=1}^n Y^*_{ki}(t)/n$, for $t = 0, 1$.

Proof is given in supplementary Appendix S5. Similar to Theorem 1, there is an additional third term in the variance beyond the two standard terms, induced by the fact that $\hat{f}_k(\boldsymbol{X}_i)$ is negatively correlated across units. In practice, we replace the entries of $\boldsymbol{\Sigma}$ with their sample analogues, which result in a consistent estimator $\widehat{\boldsymbol{\Sigma}}$. By Slutsky's lemma, the asymptotic distribution is not affected by this substitution.

2.4. Nonparametric Test of Rank-Consistent Treatment Effect Heterogeneity

To evaluate the quality of the scoring rule produced by an ML algorithm, we can test whether or not the rank of estimated GATES is consistent with that of the true GATES. The relevant null hypothesis is given by,

$$H^*_0 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K. \tag{5}$$

Unlike the null hypothesis for treatment effect heterogeneity given in (4), this is a composite null hypothesis.

To characterize the sampling distribution under this null hypothesis H^*_0, we consider the following optimization problem,

$$\boldsymbol{\mu}^*(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \boldsymbol{x}\|_2^2 \quad \text{subject to } \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K,$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_K)^\top$ and $\boldsymbol{x} \in \mathbb{R}^K$. If $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$, the following test statistic has a mixture of appropriately weighted χ^2 distribution with K degrees of freedom, called chi-bar-squared distribution (Shapiro 1988),

$$(\boldsymbol{x} - \boldsymbol{\mu}^*(\boldsymbol{x}))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}^*(\boldsymbol{x})) \sim \bar{\chi}^2_K.$$

Using this fact, the next theorem derives a nonparametric test of rank-consistent treatment effect heterogeneity that is asymptotically uniformly most powerful.

Theorem 4 (Nonparametric Test of Rank-Consistent Treatment Effect Heterogeneity). Suppose that Assumptions 1–5 hold. Then, as $n \to \infty$ and $0 < n_1/n < 1$ stays constant, an asymptotically uniformly most powerful test of size α for the null hypothesis H^*_0 defined in (5) against the alternative $H^*_1 : \mathbb{R}^K \setminus H^*_0$ has the following critical region,

$$\{\hat{\boldsymbol{\tau}} \in \mathbb{R}^K \mid \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}})\right)^\top \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}})\right) > C_\alpha\},$$

for some constant C_α that only depends on α. The expression of $\boldsymbol{\Sigma}$ is given in Theorem 3. Under H^*_0 and as $n \to \infty$, we have,

$$\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})\right)^\top \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})\right) \xrightarrow{d} \bar{\chi}^2_K.$$

Proof is given in supplementary Appendix S6. In practice, we use Monte Carlo simulations to approximately compute the critical values.

While our test is the asymptotically most powerful test of its type, it is likely to be conservative as we control the critical value based on the worst-case scenario among all the distributions consistent with the null hypothesis. In the literature on statistical tests of moment inequalities, scholars have developed subsampling and moment selection techniques that can improve their statistical power (see e.g., Andrews and Guggenberger 2009; Andrews and Soares 2010; Canay 2010; Chernozhukov, Chetverikov, and Kato 2019). Canay, Illanes, and Velez (2023) provides an up-to-date review.

3. Generalization to Cross-Fitting

In this section, we generalize our methodology to *cross-fitting*, in which researchers use the same experimental data to first generate the scoring rule using an ML algorithm and then estimate the GATES based on the resulting scoring rule. In comparison with *sample splitting* discussed in Section 2 where they are done on separate samples, cross-fitting could potentially be much more efficient. The key challenge, however, is the incorporation of additional uncertainty due to the random splitting of the data. We show how to overcome this under the Neyman's repeated sampling framework.

3.1. Estimation and Inference

Under cross-fitting, we randomly divide the experimental data into $L \geq 2$ folds of equal size $m = n/L$ where for the sake of simplicity we assume n is a multiple of L, and each fold contains m_1 treated units with m_0 control units, that is, $m = m_0 + m_1$. We maintain Assumptions 1–3 introduced in Section 2.1. Then, for each $\ell = 1, 2, \ldots, L$, we use the ℓth fold as a validation dataset $\mathcal{Z}_\ell = \{\boldsymbol{X}_i^{(\ell)}, T_i^{(\ell)}, Y_i^{(\ell)}\}_{i=1}^m$ to conduct statistical tests and estimate the GATES. We use the remaining folds, $\mathcal{Z}_{-\ell} = \{\boldsymbol{X}_i^{(-\ell)}, T_i^{(-\ell)}, Y_i^{(-\ell)}\}_{i=1}^{n-m}$, as the training dataset to estimate the scoring rule with an ML algorithm.

Suppose that we define a generic ML algorithm as a deterministic map from the space of training data $\boldsymbol{\mathcal{Z}}_{\text{train}}$ to the space of scoring rules $\mathcal{S}$:

$$F : \boldsymbol{\mathcal{Z}}_{\text{train}} \to \mathcal{S}.$$

Then, for a given training dataset $\mathcal{Z}_{\text{train}}$ of size $n - m$, the estimated scoring rule is given by,

$$\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}} = F(\mathcal{Z}_{\text{train}}^{n-m}). \tag{6}$$

We now generalize the definition of the GATES to the cross-fitting case,

$$\begin{aligned}&\tau_k(F, n-m)\\ &= \mathbb{E}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid c_{k-1}(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}) \leq \hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}(\boldsymbol{X}_i) \leq c_k(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}})\}],\end{aligned} \tag{7}$$

where the inner expectation is taken over the distribution of $\{\boldsymbol{X}_i, Y_i(0), Y_i(1)\}$ among the units who belong to the kth group, and the outer expectation is taken over all possible training sets of size $n - m$ from $\mathcal{Z}_{\text{train}}^{n-m}$ the population $\mathcal{P}$.

This generalized GATES is not a function of fixed scoring rule. Rather, it is a function of ML algorithm F itself (as well as the sample size of training data, $n - m$). Intuitively, it represents the average of GATES based on all observations that score between $(k-1)/K \times 100$th percentile and $k/K \times 100$th percentile under the ML algorithm F across all possible training datasets of size $n-m$. Alternatively, the cross-fitted GATE can be seen as a weighted average of GATEs that are specific to scoring rules where weights are determined by the training data and the particular ML algorithm.

We describe estimation and inference for $\tau_k(F, n-m)$. For each fold ℓ, we first estimate a scoring rule s by applying an ML algorithm F to the training data $\mathcal{Z}_{-\ell}$,

$$\hat{s}_\ell = F(\mathcal{Z}_{-\ell}). \tag{8}$$

We then estimate the GATES based on the validation data $\mathcal{Z}_\ell$, using the following estimator that is analogous to the one defined in (3),

$$\begin{aligned}\hat{\tau}_k^\ell(F, n-m) = K\Bigg[&\frac{1}{m_1}\sum_{i=1}^m Y_i^{(\ell)} T_i^{(\ell)} \hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)})\\ &+ \frac{1}{m_0}\sum_{i=1}^m Y_i^{(\ell)}(1 - T_i^{(\ell)})\left\{1 - \hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)})\right\}\\ &- \frac{1}{m_0}\sum_{i=1}^m Y_i^{(\ell)}(1 - T_i^{(\ell)})\Bigg],\end{aligned}$$

Algorithm 1 Estimation of the Sorted Group Average Treatment Effects (GATES) under Cross-fitting

Input: Data $\mathcal{Z} = \{\boldsymbol{X}_i, T_i, Y_i\}_{i=1}^n$, Machine learning algorithm F, Estimator $\hat{\tau}_k$, Number of folds L

Output: Estimated GATES $\{\hat{\tau}_k(F, n-m)\}_{k=1}^K$

Split the data $\mathcal{Z}$ into L random subsets of equal size $\{\mathcal{Z}_1, \ldots, \mathcal{Z}_L\}$
Set $m \leftarrow n/L$ and $\ell \leftarrow 1$
while $\ell \leq L$ **do**
  $\mathcal{Z}_{-\ell} = \{\mathcal{Z}_1, \cdots, \mathcal{Z}_{\ell-1}, \mathcal{Z}_{\ell+1}, \ldots, \mathcal{Z}_L\}$
    ▷ Create the training dataset
  $\hat{s}_{-\ell} = F(\mathcal{Z}_{-\ell})$
    ▷ Estimate the scoring rule s by applying F to $\mathcal{Z}_{-\ell}$
  $\hat{\tau}_k^\ell = \hat{\tau}_k(\mathcal{Z}_\ell)$ for each $k = 1, 2, \ldots, K$
    ▷ Calculate the GATES estimator using $\mathcal{Z}_\ell$
  $\ell \leftarrow \ell + 1$
end while
return $\hat{\tau}_k(F, n-m) = \frac{1}{L}\sum_{\ell=1}^L \hat{\tau}_k^\ell$ for each $k = 1, 2, \ldots, K$

where $\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = \mathbf{1}\{\hat{s}_\ell(\boldsymbol{X}_i^{(\ell)}) \geq \hat{c}_{k-1}^\ell(\hat{s}_\ell)\} - \mathbf{1}\{\hat{s}_\ell(\boldsymbol{X}_i^{(\ell)}) \geq \hat{c}_k^\ell(\hat{s}_\ell)\}$, and $\hat{c}_k^\ell(\hat{s}_\ell) = \inf\{c \in \mathbb{R} : \sum_{i=1}^m \mathbf{1}\{\hat{s}_\ell(\boldsymbol{X}_i^{(\ell)}) > c\} \leq mk/K\}$ represents the estimated cutoff in the ℓth subsample. Repeating this for each fold and averaging the results gives us the final GATES estimator,

$$\hat{\tau}_k(F, n-m) = \frac{1}{L}\sum_{\ell=1}^L \hat{\tau}_k^\ell \tag{9}$$

for $k = 1, 2, \ldots, K$. Algorithm 1 summarizes this estimation procedure.

We extend our bias and variance results under sample splitting (Theorem 1) to the cross-fitting case by incorporating the additional randomness induced by the cross-fitting procedure.

Theorem 5 (Bias Bound and Exact Variance of the GATES Estimator under Cross-fitting). Under Assumptions 1–3, the bias of the proposed GATES estimator given in (9) can be bounded as follows,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{P}\left(\left|\mathbb{E}\left\{\hat{\tau}_k(F, n-m) - \tau_k(F, n-m) \,\middle|\, \hat{c}_k(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}),\right.\right.\right.\\ &\qquad\left.\left.\left.\hat{c}_{k-1}(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}})\right\}\right| \geq \epsilon \,\middle|\, \mathcal{Z}_{\text{train}}^{n-m}\right)\right]\\ &\leq 1 - B\left(\frac{k}{K} + \gamma_k(\epsilon), \frac{nk}{K}, n - \frac{nk}{K} + 1\right)\\ &\quad + B\left(\frac{k}{K} - \gamma_k(\epsilon), \frac{nk}{K}, n - \frac{nk}{K} + 1\right)\\ &\quad - B\left(\frac{k-1}{K} + \gamma_{k-1}(\epsilon), \frac{n(k-1)}{K}, n - \frac{n(k-1)}{K} + 1\right)\\ &\quad + B\left(\frac{k-1}{K} - \gamma_{k-1}(\epsilon), \frac{n(k-1)}{K}, n - \frac{n(k-1)}{K} + 1\right),\end{aligned}$$

for any given constant $\epsilon > 0$ where $B(\epsilon, \alpha, \beta)$ is the incomplete beta function (if $\alpha \leq 0$ and $\beta > 0$, we set $B(\epsilon, \alpha, \beta) := H(\epsilon)$ for

all ϵ where $H(\epsilon)$ is the Heaviside step function), and

$$\gamma_k(\epsilon) = \frac{\epsilon}{K\mathbb{E}\{\max_{c\in[c_k(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}(\boldsymbol{X}_i))-\epsilon,\, c_k(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}(\boldsymbol{X}_i))+\epsilon]} \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}(\boldsymbol{X}_i) = c)\}}.$$

The variance of the estimator is given by,

$$\begin{aligned}\mathbb{V}(\hat{\tau}_k(F, n-m)) = {} & K^2\left(\frac{\mathbb{E}(S_{Fk1}^2)}{m_1} + \frac{\mathbb{E}(S_{Fk0}^2)}{m_0}\right) \\ & + \frac{(n-K)\mathbb{E}_\ell(\kappa_{k11}^\ell)}{n-1} - \mathbb{E}_\ell[(\kappa_{k1}^\ell)^2] \\ & + \mathbb{V}\left(\kappa_{k1}^\ell\right) - \frac{L-1}{L}\mathbb{E}(S_{Fk}^2),\end{aligned}$$

where $S_{Fkt}^2 = \sum_{i=1}^m (Y_{ki}^\ell(t) - \overline{Y_k^\ell(t)})^2/(m-1)$, $S_{Fk}^2 = \sum_{\ell=1}^L (\hat{\tau}_k^{(\ell)} - \hat{\tau}_k(F, n-m))^2/(L-1)$, $\kappa_{kt}^\ell = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{f}_k^\ell(\boldsymbol{X}_i) = t)$, and $\kappa_{ktt}^\ell = \mathbb{E}[(Y_i(1) - Y_i(0))(Y_j(1) - Y_j(0)) \mid \hat{f}_k^\ell(\boldsymbol{X}_i) = \hat{f}_k^\ell(\boldsymbol{X}_j) = t]$ for $i \neq j$ with $Y_{ki}^\ell(t) = \hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)})Y_i^{(\ell)}(t)$, and $\overline{Y_k^\ell(t)} = \sum_{i=1}^m Y_{ki}^\ell(t)/n$, for $t = 0, 1$.

Proof is given in supplementary Appendix S7. When compared to Theorem 1, although the bias bound is of a similar form, the variance expression implies two additional terms. The first additional term, $\mathbb{V}\left(\kappa_{k1}^\ell\right)$, accounts for the variation across training datasets. The second negative term, $-(L-1)\mathbb{E}(S_{Fk}^2)/L$, represents the efficiency gain of the cross-fitting procedure. As expected, when $L = 1$, the expression reduces to the sample splitting case (see Theorem 1).

The estimation of $\mathbb{E}(S_{Fkt}^2)$, $\mathbb{E}\{(\kappa_{kt}^\ell)^2\}$, $\mathbb{E}\{(\kappa_{ktt}^\ell)\}$ and $\mathbb{V}(\kappa_{kt}^\ell)$ is straightforward and based on their sample analogues:

$$\widehat{\mathbb{E}(S_{Fkt}^2)} = \frac{1}{(m-1)L}\sum_{\ell=1}^L\sum_{i=1}^m \mathbf{1}\{T_i^{(\ell)} = t\}(Y_{ki}^\ell - \overline{Y_{kt}^\ell})^2,$$

$$\widehat{\mathbb{E}\left\{(\kappa_{kt}^\ell)^2\right\}} = \frac{1}{L}\sum_{\ell=1}^L \left(\hat{\kappa}_{kt}^\ell\right)^2,$$

$$\widehat{\mathbb{V}(\kappa_{kt}^\ell)} = \frac{1}{L-1}\sum_{\ell=1}^L (\hat{\kappa}_{kt}^\ell - \overline{\hat{\kappa}_{kt}^\ell})^2,$$

where $Y_{ki}^\ell = \hat{f}_k^\ell(\boldsymbol{X}_i)Y_i^{(\ell)}$, $\overline{Y_{kt}^\ell} = \sum_{i=1}^m \mathbf{1}\{T_i = t\}Y_{ki}^{(\ell)}/m$, $\overline{\hat{\kappa}_{kt}^\ell} = \sum_{\ell=1}^L \hat{\kappa}_{kt}^\ell/L$ and

$$\hat{\kappa}_{kt}^\ell = \frac{\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}T_i^{(\ell)}Y_i^{(\ell)}}{\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}T_i^{(\ell)}} - \frac{\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}(1-T_i^{(\ell)})Y_i^{(\ell)}}{\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}(1-T_i^{(\ell)})},$$

$$\hat{\kappa}_{ktt}^\ell = \frac{[\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}T_i^{(\ell)}Y_i^{(\ell)}]^2 - \sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}T_i^{(\ell)}(Y_i^{(\ell)})^2}{[\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}T_i^{(\ell)}]^2 - \sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}T_i}$$

$$- \frac{[\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}(1-T_i^{(\ell)})Y_i^{(\ell)}]^2 - \sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}(1-T_i^{(\ell)})(Y_i^{(\ell)})^2}{[\sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}(1-T_i^{(\ell)})]^2 - \sum_{i=1}^m \mathbf{1}\{\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = t\}(1-T_i^{(\ell)})},$$

In contrast, the estimation of $\mathbb{E}(S_{Fk}^2)$ requires care. In particular, although it is tempting to estimate $\mathbb{E}(S_{Fk}^2)$ using a realization of S_{Fk}^2, this estimate is highly variable especially when L is small. As a result, it often yields a negative overall variance estimate. We address this problem by applying Lemma 1 from Nadeau and Bengio (2000) to $\hat{\tau}_k(F, n-m)$, which gives,

$$\mathbb{V}(\hat{\tau}_k(F, n-m)) \geq \mathbb{E}(S_{Fk}^2).$$

Since Theorem 5 implies:

$$\begin{aligned}&\mathbb{V}(\hat{\tau}_k(F, n-m)) \\ &\leq K^2\left(\frac{\mathbb{E}(S_{Fk1}^2)}{m_1} + \frac{\mathbb{E}(S_{Fk0}^2)}{m_0}\right) + \frac{(n-K)\mathbb{E}_\ell[\kappa_{k11}^\ell]}{n-1} - \mathbb{E}_\ell[(\kappa_{k1}^\ell)^2] + \mathbb{V}\left(\kappa_{k1}^\ell\right),\end{aligned}$$

this inequality suggests the following estimator of $\mathbb{E}(S_{Fk}^2)$,

$$\begin{aligned}\widehat{\mathbb{E}(S_{Fk}^2)} = \min\Bigg(& S_{Fk}^2, K^2\left(\frac{\widehat{\mathbb{E}(S_{Fk1}^2)}}{m_1} + \frac{\widehat{\mathbb{E}(S_{Fk0}^2)}}{m_0}\right) \\ & + \frac{(n-K)\mathbb{E}_\ell[\kappa_{k11}^\ell]}{n-1} - \mathbb{E}_\ell[(\kappa_{k1}^\ell)^2] + \widehat{\mathbb{V}\left(\kappa_{k1}^\ell\right)}\Bigg). \quad (10)\end{aligned}$$

Although this yields a conservative estimate of $\mathbb{V}(\hat{\tau}_k(F, n-m))$ in finite samples, the bias appears to be relatively small in practice (see Section 4). In Appendix S8, we show that the estimator is consistent as L goes to infinity and sufficiently large m.

To establish the asymptotic sampling distribution of our cross-fitting GATES estimator, we first extend our CATE continuity condition (Assumption 4) by assuming that each CATE given a training dataset is continuous and the average CATE (over all possible training datasets) is bounded.

Assumption 6 (Continuity of CATE at the Thresholds under Cross-Fitting). Let $F_{\mathcal{Z}_{\text{train}}^{n-m}}(c) = \Pr(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}(\boldsymbol{X}_i) \leq c)$ represent the cumulative distribution function of the scoring rule under training set $\mathcal{Z}_{\text{train}}^{n-m}$ and define its pseudo-inverse as $F_{\mathcal{Z}_{\text{train}}^{n-m}}^{-1}(p) = \inf\{c : F_{\mathcal{Z}_{\text{train}}^{n-m}}(c) \geq p\}$ for $p \in [0, 1]$. Then, for all but asymptotically measure-zero set of possible training sets $\mathcal{Z}_{\text{train}}^{n-m}$ of size $n-m$, the CATE function $\tau_{\mathcal{Z}_{\text{train}}^{n-m}}(p) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}(\boldsymbol{X}_i) = F_{\mathcal{Z}_{\text{train}}^{n-m}}^{-1}(p))$ is left-continuous with bounded variation on any interval $(\epsilon, 1-\epsilon)$ with $0 < \epsilon < 1/2$, and continuous in p at $p = 1/K, \ldots, (K-1)/K$. Furthermore, we assume $\lim_{n\to\infty} \mathbb{E}_{\mathcal{Z}_{\text{train}}^{n-m}}[\max_{p\in[0,1]} \tau_{\mathcal{Z}_{\text{train}}^{n-m}}(p)] < \infty$.

In addition, we require the ML algorithm F to be stable.

Assumption 7 (ML Algorithm Stability). Let $\mathcal{Z}_{\text{train}}^n$ be a training dataset of size n and $\hat{s}_{\mathcal{Z}_{\text{train}}^n} = F(\mathcal{Z}_{\text{train}}^n)$ represent the estimated scoring rule that results from the application of an ML algorithm F to the training dataset. Then, as $m \to \infty$ (with L fixed), for any a, b with $a < b$:

$$\|\mathbb{E}[Y_i(1) - Y_i(0) \mid a \leq \hat{s}_{\mathcal{Z}_{\text{train}}^n}(\boldsymbol{X}_i) \leq b]\|_2 = o\left(m^{-1}\right).$$

The expectation is taken over the distribution of $\{\boldsymbol{X}_i, Y_i(0), Y_i(1)\}$ among those units in the population $\mathcal{P}$ who belong to the group defined by the conditioning set. The outer norm is computed across the random sampling of training dataset of size n from the same population. Assumption 7 implies that as the size of training data approaches infinity, L_2 norm of the resulting scoring rule $\hat{s}_{\mathcal{Z}_{\text{train}}^n}$ stabilizes sufficiently quickly at a rate faster than $O(m^{-1})$. The required rate is

consistent with the asymptotic conditions needed for other related cross-validation settings (e.g., Austern and Zhou 2020). Importantly, we do not assume that the ML algorithm converges to the true CATE.

Finally, the next theorem established the asymptotic distribution of GATES estimator under cross-fitting.

Theorem 6 (Asymptotic Sampling Distribution of GATES Estimator under Cross-Fitting). Suppose L is fixed. Then, under Assumptions 1–3, 5–7, we have, as m goes to infinity,

$$\frac{\hat{\tau}_k(F, n-m) - \tau_k(F, n-m)}{\sqrt{\mathbb{V}(\hat{\tau}_k(F, n-m))}} \xrightarrow{d} N(0,1)$$

where the expression of $\mathbb{V}(\hat{\tau}_k(F, n-m))$ is given in Theorem 5.

Proof is given in supplementary Appendix S9, and is similar to the proof of Theorem 2.

3.2. Nonparametric Tests of Treatment Effect Heterogeneity

We now extend the nonparametric tests of treatment effect heterogeneity and its rank-consistency introduced in Sections 2.3 and 2.4 to the cross-fitting setting. Similar to Chernozhukov et al. (2023), we account for the additional uncertainty due to random splitting. Unlike their method, however, the proposed tests do not require a computationally intensive resampling procedure.

Our first null hypothesis of interest is that the GATES are all equal to the ATE,

$$H_{F0} : \tau_1(F, n-m) = \tau_2(F, n-m) = \cdots = \tau_K(F, n-m). \tag{11}$$

This null hypothesis depends on the ML algorithm F whereas the null hypothesis given in (4) depends on the (fixed) scoring rule.

The following theorem generalizes the result of Theorem 3 to cross-fitting.

Theorem 7 (Nonparametric Test of Treatment Effect Heterogeneity Under Cross-fitting). Suppose L is fixed. Then, under Assumptions 1–3, 5–7, and the null hypothesis H_{F0} defined in (11) and against the alternative $H_{F1} : \mathbb{R}^K \setminus H_{F0}$, as $m \to \infty$, and $0 < m_1/m < 1$ stays constant, we have,

$$\hat{\boldsymbol{\tau}}_F^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}}_F \xrightarrow{d} \chi^2_K$$

where $\hat{\boldsymbol{\tau}}_F = (\hat{\tau}_1(F, n-m) - \hat{\tau}, \ldots, \hat{\tau}_K(F, n-m) - \hat{\tau})$, and $\boldsymbol{\Sigma}$ is defined as for $k, k' \in \{1, \ldots K\}$:

$$\begin{aligned}
\boldsymbol{\Sigma}_{kk} = {} & K^2 \left(\frac{\mathbb{E}(S^{*2}_{Fk1})}{m_1} + \frac{\mathbb{E}(S^{*2}_{Fk0})}{m_0} \right) - \frac{L-1}{L} \mathbb{E}(S^2_{Fk}) + \mathbb{V}\left(\kappa^\ell_{k1}\right) \\
& + \frac{1}{K} \mathbb{E}_\ell \left\{ (K-2) \left(\frac{n-K}{n-1} \kappa^\ell_{kk11} - (\kappa^\ell_{k1})^2 \right) \right. \\
& \left. - \frac{2n(K-1)}{(n-1)} \kappa^\ell_{kk01} + 2\kappa^\ell_{k1}\kappa^\ell_{k0} \right\} \\
\boldsymbol{\Sigma}_{kk'} = {} & K^2 \left(\frac{\mathbb{E}(S^{*2}_{Fkk'1})}{m_1} + \frac{\mathbb{E}(S^{*2}_{Fkk'0})}{m_0} \right) - \frac{L-1}{L} \mathbb{E}(S^2_{Fkk'}) \\
& + \operatorname{cov}\left(\kappa^\ell_{k1}, \kappa^\ell_{k'1}\right) + \frac{1}{K} \mathbb{E}_\ell \left\{ (K-2) \left(\kappa^\ell_{kk'11} - \kappa^\ell_{k1}\kappa^\ell_{k'1} \right) \right. \\
& \left. - \frac{Kn-n-1}{n-1} \left(\kappa^\ell_{kk'10} + \kappa^\ell_{kk'01} \right) + \kappa^\ell_{k1}\kappa^\ell_{k'0} + \kappa^\ell_{k0}\kappa^\ell_{k'1} \right\}
\end{aligned}$$

where $S^{*2}_{Fkt} = \sum_{i=1}^m (Y^{*\ell}_{ki}(t) - \overline{Y^{*\ell}_k(t)})^2/(m-1)$, $S^{*2}_{Fkk't} = \sum_{i=1}^m (Y^{*\ell}_{ki}(t) - \overline{Y^{*\ell}_k(t)})(Y^{\ell}_{k'i}(t) - \overline{Y^{*\ell}_{k'}(t)})/(m-1)$, $S^2_{Fkk'} = \sum_{\ell=1}^L (\hat{\tau}^\ell_k(F, n-m) - \hat{\tau}_k(F, n-m))(\hat{\tau}^\ell_{k'}(F, n-m) - \hat{\tau}_{k'}(F, n-m))/(L-1)$, $\kappa^\ell_{kt} = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{f}^\ell_k(\boldsymbol{X}_i) = t)$ and $\kappa^\ell_{kk'ts} = \mathbb{E}[(Y_i(1) - Y_i(0))(Y_j(1) - Y_j(0)) \mid \hat{f}^\ell_k(\boldsymbol{X}_i) = t, \hat{f}^\ell_{k'}(\boldsymbol{X}_i) = s]$with $Y^{*\ell}_{ki}(t) = (\hat{f}^\ell_k(\boldsymbol{X}^{(\ell)}_i) - 1/K) Y^{(\ell)}_i(t)$, and $\overline{Y^{*\ell}_k(t)} = \sum_{i=1}^m Y^{*\ell}_{ki}(t)/m$, for $t = 0, 1$.

Proof is given in supplementary Appendix S10. Compared to Theorem 3, the only difference appears in the expression of the covariance matrix $\boldsymbol{\Sigma}$ due to the efficiency gains of the cross-validation procedure. Similar to Theorem 5, the estimation of $\mathbb{E}(S^2_{Fkk'})$ for $k = k'$ requires care, and we use the consistent estimator as identified in (10). If the resulting covariance matrix estimate is not positive definite, we find the nearest positive definite matrix in the L_2 norm by using the algorithm of Higham (2002).

Finally, we extend the nonparametric test of rank-consistent treatment effect heterogeneity (Theorem 4) to cross-fitting. The null hypothesis is given by,

$$H^*_{F0} : \tau_1(F, n-m) \le \tau_2(F, n-m) \le \cdots \le \tau_K(F, n-m). \tag{12}$$

Now, we present the result.

Theorem 8 (Nonparametric Test of Rank-Consistent Treatment Effect Heterogeneity Under Cross-Fitting). Suppose L is fixed. Then, under Assumptions 1–3, 5–7, as $m \to \infty$ and $0 < m_1/m < 1$ stays constant, the uniformly most powerful test of size α for the null hypothesis H^*_{F0} defined in (12) against the alternative $H^*_{F1} : \mathbb{R}^K \setminus H^*_{F0}$ has the following critical region,

$$\{\hat{\boldsymbol{\tau}}_F \in \mathbb{R}^K \mid \left(\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F)\right)^\top \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F)\right) > C_\alpha\},$$

for some constant C_α that only depends on α. Furthermore, under H_{F0} and as $n \to \infty$, we have,

$$\left(\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F)\right)^\top \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F)\right) \xrightarrow{d} \bar{\chi}^2_K,$$

where $\hat{\boldsymbol{\tau}}_F$ and $\boldsymbol{\Sigma}$ are defined in Theorem 7.

Proof directly follows from the fact by Theorem 7, $\boldsymbol{\Sigma}^{-1/2} \hat{\boldsymbol{\tau}}_F$ is asymptotically normally distributed with variance-covariance matrix $\boldsymbol{I}$, which is an identity matrix of size $K \times K$. Then, following the same steps as those in supplementary Appendix S6 immediately establishes the result.

4. A Simulation Study

We undertake a simulation study to examine the finite sample performance of the proposed methodology. We consider both sample-splitting and cross-fitting cases. For the estimation of GATES, we evaluate the bias and variance of the proposed

estimators as well as the coverage of their confidence intervals. For hypothesis tests, we examine the actual power and size of the proposed tests. We show that the proposed methodology performs well even when the sample size is as small as 100.

4.1. The Setup

We use the data generating process from the 2016 Atlantic Causal Inference Conference (ACIC) Competition. We briefly describe its simulation setting here and refer interested readers to Dorie et al. (2019) for additional details. The focus of this competition was the inference of average treatment effect in observational studies. There are a total of 58 pre-treatment covariates $\boldsymbol{X}$, including three categorical, 5 binary, 27 count data, and 13 continuous variables. The data were taken from a real-world study with the sample size $n = 4802$.

In our simulation, we assume that the empirical distribution of these covariates represent the population $\mathcal{P}$ and obtain each simulation sample via bootstrap. We consider small and moderate sample sizes: $n = 100$, 500, and 2500. For the sample-splitting case, the models are pre-trained on the original dataset from the 2016 ACIC data challenge, and the sample size n refers to the number of testing samples. For the cross-validation case, n refers to the total dataset size, which we then conduct 5-fold cross-validation, $L = 5$. One important change from the original competition is that instead of using a propensity model to determine T, we assume that the treatment assignment is completely randomized, that is, $\Pr(T_i = 1) = 1/2$, and the treatment and control groups are of equal size, that is, $n_1 = n_0 = n/2$.

To generate the outcome variable, we use one of the settings from the competition, which is based on the generalized additive model with polynomial basis functions. The model represents a setting, in which there exists a substantial amount of treatment effect heterogeneity. The formula for this outcome model is reproduced here:

$$
\begin{aligned}
&\mathbb{E}(Y_i(t) \mid \boldsymbol{X}_i) \\
&= 1.60 + 0.53 \times x_{29} - 3.80 \times x_{29}(x_{29} - 0.98)(x_{29} + 0.86) \\
&- 0.32 \times \mathbf{1}\{x_{17} > 0\} \\
&+ 0.21 \times \mathbf{1}\{x_{42} > 0\} - 0.63 \times x_{27} + 4.68 \times \mathbf{1}\{x_{27} < -0.61\} \\
&- 0.39 \times (x_{27} + 0.91)\mathbf{1}\{x_{27} < -0.91\} \\
&+ 0.75 \times \mathbf{1}\{x_{30} \leq 0\} - 1.22 \times \mathbf{1}\{x_{54} \leq 0\} \\
&+ 0.11 \times x_{37}\mathbf{1}\{x_4 \leq 0\} - 0.71 \times \mathbf{1}\{x_{17} \leq 0, t = 0\} \\
&- 1.82 \times \mathbf{1}\{x_{42} \leq 0, t = 1\} + 0.28 \times \mathbf{1}\{x_{30} \leq 0, t = 0\} \\
&+ \{0.58 \times x_{29} - 9.42 \times x_{29}(x_{29} - 0.67)(x_{29} + 0.34)\} \\
&\times \mathbf{1}\{t = 1\} + (0.44 \times x_{27} - 4.87 \times \mathbf{1}\{x_{27} < -0.80\}) \\
&\times \mathbf{1}\{t = 0\} - 2.54 \times \mathbf{1}\{t = 0, x_{54} \leq 0\}.
\end{aligned}
$$

Throughout, we set $K = 5$ so that observations are sorted into five groups based on the magnitude of the CATE. For the case of sample-splitting, we can directly compute the true values of GATES using the outcome model and evaluate each quantity based on the entire original dataset. For the cross-validation case, however, the exact calculation of GATES true values would require averaging over all combinations of training datasets from the original dataset. Since this is computationally prohibitive, we obtain their approximate true values by independently sampling 10,000 training datasets. For each training dataset, we train an ML algorithm F using 5-fold cross-validation. Then, we use the sample mean of each estimated causal quantity across the 10,000 simulated datasets as our approximate truth.

We evaluate Bayesian Additive Regression Trees (BART) (see Chipman et al. 2010; Hill 2011; Hahn et al. 2020) and Causal Forest (Wager and Athey 2018), and LASSO (Tibshirani 1996). The number of trees were tuned through the 5-fold cross-validation for both algorithms. For implementation, we use R 3.6.3 with the following packages: bartMachine (version 1.2.6) for BART, grf (version 2.0.1) for Causal Forest, and glmnet (version 4.1-2) for LASSO. The number of trees was tuned through 5-fold cross-validation for BART and Causal Forest. The regularization parameter was tuned similarly for LASSO.

4.2. Finite-Sample Performance of the Proposed Estimators

Table 1 presents the results for the estimation of GATES in the sample-splitting case. According to this simulation setup, Causal Forest and BART appear to identify treatment effect heterogeneity better than LASSO. For example, for BART, the largest and smallest GATES are 5.89 and 2.09, respectively. In contrast, the gap between the corresponding quantities is much smaller for the LASSO, roughly equaling 2 points.

For each sample size, we conducted 1000 simulation trials. For all three algorithms, the estimated biases of the proposed GATES estimators are negligibly small, accounting for less than 5% of their estimated standard deviation in almost all cases. The bias also generally decreases as the sample size grows. We also find that the empirical coverage of the confidence intervals is close to the theoretical 95% value even when the sample size is as small as $n = 100$.

We obtain similar findings for the cross-fitting case. Table 2 shows the results for Causal Forest and LASSO. Unfortunately, BART is too computationally intensive to include for this simulation. For the results of Causal Forest and LASSO, we use 1000 trials as before. As seen in the sample-splitting case, the estimated biases of the proposed GATES estimators are relatively small even when $n = 100$ and becomes negligible when $n = 500$.

Recall that under the 5-fold cross-fitting, for example, $n = 500$ implies the evaluation sample of size 100 for each fold. And, yet, combining the five folds leads to a much lower standard deviation than the sample-splitting case with the $n = 100$ case in Table 1. The results are similar when comparing the $n = 2500$ cross-fitting case with the $n = 500$ sample-splitting case. Indeed, in some cases, the reduction in standard deviation is more than 50%. This experimentally demonstrates the efficiency gain from using a cross-fitting approach. We further find that although Theorem 5 implies that the proposed variance estimate is conservative, the results show only the slight overcoverage of the confidence intervals. In Imai and Li (2023a) we show that the methodology proposed in Chernozhukov et al. (2023) leads to more significant overcoverage of the confidence intervals.

Table 1. The finite sample performance of the GATES estimators under sample-splitting.

		$n_{\text{test}} = 100$			$n_{\text{test}} = 500$			$n_{\text{test}} = 2500$		
Estimator	Truth	Bias	SD	Coverage	Bias	SD	Coverage	Bias	SD	Coverage
Causal Forest										
$\hat{\tau}_1$	2.164	0.034	2.486	93.8%	0.041	1.071	95.0%	0.007	0.467	96.0%
$\hat{\tau}_2$	4.001	0.011	2.551	93.7	−0.060	1.183	94.4	−0.002	0.510	95.3
$\hat{\tau}_3$	4.583	−0.018	2.209	94.0	−0.003	0.956	96.4	0.020	0.421	95.8
$\hat{\tau}_4$	4.931	−0.077	2.500	94.6	0.001	1.138	94.3	0.003	0.517	95.6
$\hat{\tau}_5$	5.728	−0.058	3.332	96.0	−0.010	1.499	95.0	−0.009	0.661	95.2
BART										
$\hat{\tau}_1$	2.092	0.016	3.188	94.0%	−0.014	1.402	96.2%	0.009	0.626	95.8%
$\hat{\tau}_2$	3.913	0.127	2.918	95.1	0.028	1.388	94.0	−0.003	0.618	95.3
$\hat{\tau}_3$	4.478	−0.077	2.218	94.3	−0.041	0.968	95.0	−0.001	0.425	95.1
$\hat{\tau}_4$	5.042	−0.154	2.366	94.2	0.014	1.106	95.8	0.015	0.495	95.4
$\hat{\tau}_5$	5.881	−0.019	2.510	94.7	−0.019	1.104	94.4	−0.000	0.489	95.0
LASSO										
$\hat{\tau}_1$	3.243	0.028	2.507	94.1%	0.049	1.119	95.1%	0.003	0.769	95.1%
$\hat{\tau}_2$	3.817	−0.012	1.848	93.6	−0.013	0.834	94.5	−0.000	0.684	95.4
$\hat{\tau}_3$	4.318	−0.013	2.095	94.2	−0.002	0.930	94.5	0.010	0.516	95.0
$\hat{\tau}_4$	4.788	−0.041	2.475	94.0	−0.015	1.101	94.6	−0.001	0.480	94.6
$\hat{\tau}_5$	5.241	−0.046	3.921	94.4	0.021	1.739	95.1	0.002	0.505	95.3

NOTE: The table presents the estimated bias and standard deviation of the GATES estimators as well as the empirical coverage of their 95% confidence intervals for Causal Forest, BART, and LASSO. The machine learning algorithms are trained on the original dataset from the 2016 ACIC data challenge.

Table 2. The finite sample performance of the GATES estimators under cross-fitting.

	$n = 100$				$n = 500$				$n = 2500$			
Estimator	Truth	Bias	SD	Coverage	Truth	Bias	SD	Coverage	Truth	Bias	SD	Coverage
Causal Forest												
$\hat{\tau}_1$	3.976	−0.053	2.971	94.0%	2.900	−0.007	1.572	95.6%	2.210	−0.007	0.594	97.7%
$\hat{\tau}_2$	4.173	−0.061	2.584	95.9	4.112	−0.038	1.075	98.2	4.057	0.011	0.541	98.6
$\hat{\tau}_3$	4.286	−0.012	2.560	96.7	4.510	−0.054	1.058	97.7	4.545	0.019	0.465	98.1
$\hat{\tau}_4$	4.400	−0.119	2.865	97.4	4.799	0.066	1.149	97.9	4.951	−0.009	0.509	98.6
$\hat{\tau}_5$	4.569	0.140	3.447	94.1	5.086	0.001	1.620	96.0	5.643	−0.006	0.620	98.3
LASSO												
$\hat{\tau}_1$	4.191	−0.125	3.196	97.6%	4.017	−0.025	1.488	96.0%	3.752	−0.004	0.669	96.0%
$\hat{\tau}_2$	4.205	0.036	2.281	97.5	4.137	−0.069	1.027	97.9	4.028	−0.019	0.590	98.9
$\hat{\tau}_3$	4.268	−0.126	2.354	96.6	4.291	−0.019	1.000	97.9	4.323	0.037	0.488	97.5
$\hat{\tau}_4$	4.334	−0.003	2.536	96.8	4.430	0.035	1.174	96.8	4.571	0.033	0.642	97.2
$\hat{\tau}_5$	4.406	0.111	3.615	96.2	4.530	0.047	1.811	95.0	4.732	0.022	0.697	95.3

NOTE: The table presents the estimated bias and standard deviation of the proposed GATES estimators as well as the empirical coverage of their 95% confidence intervals for Causal Forest and LASSO.

4.3. Finite-Sample Performance of the Proposed Hypothesis Tests

We next examine the finite sample performance of the proposed hypothesis tests. Due to the aforementioned computational intensity of BART, we focus on Causal Forest and LASSO. For each simulated dataset, we conduct hypothesis tests of two null hypotheses of interest: treatment effect homogeneity (see (4) and (11) for sample-splitting and cross-fitting, respectively) and rank-consistency of the GATES (see (5) and (12) for sample-splitting and cross-fitting cases, respectively).

According to the true values shown in Tables 1 and 2, the null hypothesis of treatment effect homogeneity is false while the rank-consistency null hypothesis is correct. This suggests that the proposed test should reject the former hypothesis more frequently as the sample size increases whereas it should reject the latter hypothesis no more frequently than the specified size of the test, which we set to 5% throughout.

We first consider the sample-splitting setting based on 500 simulation trials. Table 3 presents the rejection rate and median p-value for each scenario across different training and testing data sizes, denoted by n_{train} and n_{test}, respectively. We find that for Causal Forest, the training data of size 400 and the test data of size 2000 are required to reject the null hypothesis of treatment effect homogeneity with a high probability. This highlights the difficulty of identifying treatment effect heterogeneity in randomized experiments. For the hypothesis test of the rank-consistency of GATES, we find that if trained with a small sample ($n_{\text{train}} = 100$), Causal Forest might falsely reject the null hypothesis but this false rejection rate is less than the size of the test regardless of the size of the test data. This reflects the conservative nature of our test as discussed at the end of Section 2.

We obtain similar findings for LASSO, where small training data leads to low rejection rates for the treatment effect homogeneity hypothesis and some false rejection of the rank consistency hypothesis. As before, the false rejection rates are approximately 5% or lower (the small number of simulations induce some noise). Interestingly, the proposed test is much less powerful for LASSO than for Causal Forest. Even when the size of training data is 2000 and the test data size is 2500, the rejection rate is only slightly above 25%. This is consistent with

Table 3. The finite sample performance of the hypothesis tests for treatment effect homogeneity and rank-consistency of GATES under sample-splitting.

	$n_{\text{test}} = 100$		$n_{\text{test}} = 500$		$n_{\text{test}} = 2500$	
	Rejection rate	Median p-value	Rejection rate	Median p-value	Rejection rate	Median p-value
Causal Forest						
H_0:*Treatment effect homogeneity*						
$n_{\text{train}} = 100$	5.2%	0.504	7.4%	0.529	19.6%	0.361
$n_{\text{train}} = 400$	9.0	0.459	22.0	0.254	74.4	0.002
$n_{\text{train}} = 2000$	13.0	0.367	40.4	0.092	96.0	0.000
H_0^*: *Rank consistency of GATES*						
$n_{\text{train}} = 100$	4.0%	0.583	2.2%	0.624	2.2%	0.704
$n_{\text{train}} = 400$	2.8	0.687	0.2	0.820	0.2	0.907
$n_{\text{train}} = 2000$	1.2	0.707	0.2	0.852	0.0	0.967
LASSO						
H_0: *Treatment effect homogeneity*						
$n_{\text{train}} = 100$	5.8%	0.496	5.2%	0.544	9.6%	0.516
$n_{\text{train}} = 400$	7.0	0.557	4.0	0.578	10.4	0.468
$n_{\text{train}} = 2000$	6.2	0.489	9.4	0.519	26.2	0.249
H_0^*: *Rank consistency of GATES*						
$n_{\text{train}} = 100$	4.6%	0.525	3.0%	0.584	5.4%	0.596
$n_{\text{train}} = 400$	6.0	0.494	1.8	0.600	2.4	0.687
$n_{\text{train}} = 2000$	3.2	0.608	1.4	0.698	1.2	0.851

NOTE: The results are based on Causal Forest and LASSO. The table presents the percent of 500 simulation trials where each null hypothesis is rejected using the 5% test size. In addition, the median p-value across all trials is also shown. The results are presented for different training data sizes n_{train} and different test data sizes n_{test}.

Table 4. The finite sample performance of the hypothesis tests for treatment effect homogeneity and rank-consistency of GATES under cross-fitting.

	$n_{\text{test}} = 100$		$n_{\text{test}} = 500$		$n_{\text{test}} = 2500$	
	Rejection rate	Median p-value	Rejection rate	Median p-value	Rejection rate	Median p-value
Causal Forest						
Homogeneous treatment effects	1.4%	0.790	4.6%	0.712	51.4%	0.041
Consistent treatment effects	1.4%	0.702	0.8%	0.845	0.0%	0.976
LASSO						
Homogeneous treatment effects	0.6%	0.880	1.8%	0.850	9.0%	0.664
Consistent treatment effects	1.0%	0.722	0.6%	0.769	0.2%	0.889

NOTE: The results are based on Causal Forest and LASSO. The table presents the percent of 500 simulation trials where each null hypothesis is rejected using the 5% test size and also the median p-value across all trials.

the finding in Section 4.2 that LASSO discovers less treatment effect heterogeneity than Causal Forest.

We also examine the performance of our hypothesis tests under cross-fitting, again using 500 simulation trials. Table 4 presents the rejection rate and median p-value across different sample sizes. We use $L = 5$ fold cross-fitting for all simulations. Note that the $n = 500$ case under cross-fitting is analogous in the size of training and testing data to the ($n_{\text{train}} = 400, n_{\text{test}} = 100$) case for sample splitting. Similarly, the $n = 2500$ case under cross-fitting corresponds to the ($n_{\text{train}} = 2000, n_{\text{test}} = 500$) case under sample-splitting.

For both Causal Forest and LASSO, the rejection rate of the homogeneous treatment effect hypothesis is lower in the $n = 500$ case compared with the corresponding sample-splitting case, reflecting the additional uncertainty due to the sampling of training data (under sample-splitting, the scoring rule is regarded as fixed). However, when the sample size is $n = 2500$, for both algorithms the rejection rate of homogeneous treatment effects is higher under cross-fitting than sample-splitting, demonstrating that the efficiency gain of cross-fitting outweigh its additional sampling uncertainty. For the hypothesis test of rank-consistency, we find that the rejection rate under cross-fitting is significantly lower than the nominal test size for all cases.

5. An Empirical Application

To demonstrate the applicability of the proposed framework, we use the experimental data from the male sub-sample of the National Supported Work Demonstration (NSW) (LaLonde 1986; Dehejia and Wahba 1999). NSW was a temporary employment program to help disadvantaged workers by providing them with work experience and counseling in a sheltered environment. Specifically, qualified applicants were randomly assigned to the treatment and control groups, where the workers in the treatment group were given a guaranteed job for 9–18 months. The primary outcome of interest is the annualized earnings in 1978, 36 months after the program. The data contains a total of $n = 722$ observations, with $n_1 = 297$ participants assigned to the treatment group and $n_0 = 425$ participants in the control group. There are seven available pre-treatment covariates $\boldsymbol{X}$ that records the demographics and pre-treatment earnings of the participants.

We evaluate Causal Forest, BART, and LASSO under the two settings considered in this article. For sample-splitting, we randomly select 67% of the data (484 observations) to serve as a training dataset. We use the remaining 238 samples to estimate the GATES and conduct the proposed hypothesis tests. For cross-fitting, we first randomly split the data into 3-fold, that is, $L = 3$. We use each fold once as a testing set, while

Table 5. The Estimated GATES and their 95% confidence intervals based on Causal Forest, BART, and LASSO under sample-splitting and cross-fitting.

	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
Sample-splitting					
Causal Forest	3.40	0.13	−0.85	−1.91	7.21
	[−1.29,3.40]	[−5.37,5.63]	[−5.22, 3.52]	[−5.16,1.34]	[1.22,13.19]
BART	2.90	−0.73	−0.02	3.25	2.57
	[−2.25,8.06]	[−5.05,3.58]	[−3.47,3.43]	[−1.53,8.03]	[−3.82,8.97]
LASSO	1.86	2.62	−2.07	1.39	4.17
	[−3.59, 7.30]	[−1.69,6.93]	[−5.39,1.26]	[−2.95,5.73]	[−2.30,10.65]
Cross-fitting					
Causal Forest	−3.72	1.05	5.32	−2.64	4.55
	[−6.52,−0.93]	[−2.28,4.37]	[2.63,8.01]	[−5.07,−0.22]	[1.14,7.96]
BART	0.40	−0.15	−0.40	2.52	2.19
	[−3.79,4.59]	[−2.54,2.23]	[−3.37,2.56]	[−0.99,6.03]	[−0.73,5.11]
LASSO	0.65	0.45	−2.88	1.32	5.02
	[−3.65,4.94]	[−3.28,4.18]	[−5.38,−0.38]	[−1.83,4.48]	[−0.14,10.18]

NOTE: The estimated GATES based on quintiles are reported in units of 1000 US dollars. Sample-splitting is done using 67% of the sample as the training data and 33% of the sample as the evaluation data. For cross-fitting, we use 3-fold of equal size.

Table 6. The results of the proposed hypothesis tests under sample-splitting and cross-fitting using Causal Forest, BART, and LASSO.

	Causal Forest		BART		LASSO	
	stat	*p*-value	stat	*p*-value	stat	*p*-value
Sample-splitting						
Homogeneous treatment effects	9.78	0.082	2.76	0.737	5.26	0.362
Rank-consistent treatment effects	3.07	0.323	1.13	0.657	3.14	0.302
Cross-fitting						
Homogeneous treatment effects	30.29	0.000	2.32	0.803	10.79	0.056
Rank-consistent treatment effects	0.06	0.691	0.04	0.885	0.45	0.711

NOTE: The values of test statistics and *p*-values are presented. We test the null hypotheses of treatment effect homogeneity and rank-consistency of the GATES.

the remaining two folds are the training set. The number of trees was tuned through 5-fold cross-validation for BART and Causal Forest within each training dataset. The regularization parameter was tuned similarly for LASSO.

We focus on the quintile GATES ($K = 5$). Table 5 presents the results (reported in 1000 U.S. dollars) under the sample-splitting and cross-fitting settings. We find that Causal Forest is able to produce statistically significantly positive GATES for the highest quintile group ($\hat{\tau}_5$) under both sample-splitting and cross-fitting. Thus, unlike the other two algorithms, Causal Forest can identify a 20% subset that benefits significantly from the temporary employment program.

Two additional observations are worth noting. First, the confidence intervals are generally narrower in the cross-fitting case compared to the sample-splitting case. This finding is consistent with the fact that cross-fitting is more efficient than sample-splitting. Second, the three algorithms failed to produce any statistically significant positive GATES for the remaining groups. This may be because there are few additional workers who benefit from the program. Alternatively, it is also possible that such workers exist but the algorithms are unable to identify them.

To formally evaluate the statistical significance of several GATES estimates, we must account for the potential multiple testing problem. Thus, we apply the proposed hypothesis tests to evaluate the null hypotheses of treatment effect homogeneity and rank-consistency of the GATES. Table 6 presents the resulting values of test statistics and *p*-values. We find that under sample-splitting, only Causal Forest is able to reject the null hypothesis of treatment effect homogeneity at the 10% level. However, under cross-fitting, both Causal Forest and LASSO algorithms can reject the null hypothesis at the 10% level, with Causal Forest being able to reject the hypothesis at even the 0.1% level. In contrast, BART fails to reject the treatment effect homogeneity hypothesis under both sample-splitting and cross-fitting. The results with Causal Forest suggest that the identification of a statistically significant GATES estimate for one subgroup under cross-fitting is able to grant enough power to reject the null hypothesis that the average treatment effects are homogeneous across all subgroups. Finally, we find that all three algorithms fail to reject the null hypothesis of the rank-consistency of GATEs. Thus, under our conservative tests, there is no strong statistical evidence that these algorithms are producing unreliable GATES.

6. Concluding Remarks

Many randomized experiments have a limited sample size and the resulting treatment effect estimates are often small and noisy. Yet, applied researchers often use machine learning algorithms to estimate heterogeneous treatment effects. Therefore, it is important to statistically distinguish signal from noise. We have developed the framework that does not impose a strong assumption on machine learning algorithms and hence is applicable to a wide range of situations. The proposed methodology allows researchers to construct confidence intervals on the estimated average treatment effects within a group identified by any machine learning algorithm. We also show how to conduct formal hypothesis tests about heterogeneous treatment effects. Our method solely relies upon the randomization of treatment assignment and the random sampling of units, and hence, yields

reliable statistical inference even when the sample size is relatively small and machine learning algorithms are not performing well.

Supplementary Materials

All proofs to the theorems in the article are contained in the supplementary materials available online.

Acknowledgments

We also thank Kevin Du, Lucas Janson, and Yash Nair for their feedback on an earlier version of this article.

Data Availability Statement

The proposed methodology is implemented through an open-source R package, evalITR, which is freely available for download at the Comprehensive R Archive Network (CRAN; *https://CRAN.R-project.org/package=evalITR*).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

We thank the Sloan foundation (Economics Program; 2020–13946) for partial support.

ORCID

Kosuke Imai http://orcid.org/0000-0002-2748-1022
Michael Lingzhi Li http://orcid.org/0000-0002-2456-4834

References

Andrews, D. W., and Guggenberger, P. (2009), "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669–709. [260]

Andrews, D. W., and Soares, G. (2010), "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. [260]

Austern, M., and Zhou, W. (2020), "Asymptotics of Cross-Validation," arXiv preprint arXiv:2001.11111. [263]

Bhattacharya, P. K. (1974), "Convergence of Sample Paths of Normalized Sums of Induced Order Statistics," *The Annals of Statistics*, 2, 1034–1039.

Canay, I., Illanes, G., and Velez, A. (2023), "A User's Guide to Inference in Models Defined by Moment Inequalities," Technical Report, National Bureau of Economic Research. [260]

Canay, I. A. (2010), "El Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," *Journal of Econometrics*, 156, 408–425. [260]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2019), "Inference on Causal and Structural Parameters Using Many Moment Inequalities," *The Review of Economic Studies*, 86, 1867–1900. [260]

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2023), "Fisher-Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments," Technical Report, arXiv:1712.04802. [256,257,258,263,264]

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010), "Bart: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [264]

Cramér, H., and Wold, H. (1936), "Some Theorems on Distribution Functions," *Journal of the London Mathematical Society*, 1, 290–294.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008), "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90, 389–405. [257]

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American statistical Association*, 94, 1053–1062. [266]

Ding, P., Feller, A., and Miratrix, L. (2016), "Randomization Inference for Treatment Effect Variation," *Journal of the Royal Statistical Society*, Series B, 78, 655–671. [257]

Ding, P., Feller, A., and Miratrix, L. (2019), "Decomposing Treatment Effect Variation," *Journal of the American Statistical Association*, 114, 304–317. [257]

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019), "Automated Versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition," *Statistical Science*, 34, 43–68. [264]

Dwivedi, R., Tan, Y. S., Park, B., Wei, M., Horgan, K., Madigan, D., and Yu, B. (2020), "Stable Discovery of Interpretable Subgroups via Calibration in Causal Studies," *International Statistical Review*, 88, S135–S178. [257]

Hahn, P. R., Murray, J. S., Carvalho, C. M., et al. (2020), "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," *Bayesian Analysis*, 15, 965–1056. [264]

Higham, N. J. (2002), "Computing the Nearest Correlation Matrix—A Problem from Finance," *IMA journal of Numerical Analysis*, 22, 329–343. [263]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [264]

Holland, P. W. (1986), "Statistics and Causal Inference," (with Discussion), *Journal of the American Statistical Association*, 81, 945–960. [257]

Imai, K., and Li, M. L. (2023a), "Comment on 'Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments'," *Econometrica*, Forthcoming. [257,264]

——— (2023b), "Experimental Evaluation of Individualized Treatment Rules," *Journal of the American Statistical Association*, 118, 242–256. [257,260]

——— (2023c), "Statistical Performance Guarantee for Selecting Those Predicted to Benefit Most from Treatment," Technical Report, arXiv preprint 2310.07973. [257]

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2018), "Meta-Learners for Estimating Heterogeneous Treatment Effects Using Machine Learning," Technical Report, arXiv:1706.03461. [259]

LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 76, 604–620. [266]

Nadeau, C., and Bengio, Y. (2000), "Inference for the Generalization Error," in *Advances in Neural Information Processing Systems*, pp. 307–313. [262]

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9. (translated in 1990)," *Statistical Science*, 5, 465–480. [256,257]

Rubin, D. B. (1990), "Comments on 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9' by J. Splawa-Neyman Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed," *Statistical Science*, 5, 472–480. [257]

Shapiro, A. (1988), "Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis," *International Statistical Review/Revue Internationale de Statistique*, 56, 49–62. [260]

Shorack, G. R. (1972), "Functions of Order Statistics," *The Annals of Mathematical Statistics*, 43, 412–427.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via LASSO," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [264]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [259,264]

Wellner, J. A. (1977), "A Glivenko-Cantelli Theorem and Strong Laws of Large Numbers for Functions of Order Statistics," *The Annals of Statistics*, 5, 473–480.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021), "Evaluating Treatment Prioritization Rules Via Rank-Weighted Average Treatment Effects," arXiv preprint arXiv:2111.07966. [257,258]