Trustworthy selection of heterogeneous treatment effect estimator

Zijun Gao *

October 16, 2024

1 Formulation and background

1.1 Potential outcome model

We follow the Neyman-Rubin potential outcome model with a treatment condition and a control condition. For unit i, there is a d-dimensional covariate vector X_i, a binary treatment assignment $W_i \in \{0, 1\}$, and two potential outcomes: $Y_i(0)$ under the control condition and $Y_i(1)$ under the treatment condition. We denote the observed outcome by Y_i, which equals $Y_i(1)$ if the unit is under treatment, and $Y_i(0)$, otherwise. We use $Z_i = (X_i, W_i, Y_i)$ to denote all the observed data of unit i.

We make the conventional assumptions of SUTVA, overlap, and unconfoundedness.

Assumption 1.1 (Stable unit treatment value assumption (SUTVA))**.** *The potential outcomes for any unit do not depend on the treatments assigned to other units. There are no different versions of each treatment level.*

Assumption 1.2 (Unconfoundedness)**.** *The assignment mechanism does not depend on potential outcomes:*

$$(Y_i(1), Y_i(0)) \perp W_i \mid X_i.$$

Assumption 1.3 (Overlap)**.** *There is a positive probability of receiving treatment and control for all individuals.*

To define our causal estimand, we introduce the super-population. We assume individuals are sampled i.i.d. from a super-population, denoted by $\mathbb{P}$. In particular, the covariates X_i are sampled from an unknown distribution $\mathbb{P}_X$. Given the covariates X_i, a binary group assignment $W_i \in \{0, 1\}$ is generated from a Bernoulli distribution with mean $e(X_i)$ (also known as the propensity score). Echoing with Assumption 1.3, we assume there exists $\eta > 0$ such that $\eta < e(x) < 1 - \eta$. The potential outcomes are modeled by

$$Y_i(1)|X_i = \mu_1(X_i) + \epsilon_i,$$
$$Y_i(0)|X_i = \mu_0(X_i) + \epsilon_i,$$

where $\mu_0(x)$, $\mu_1(x)$ represent the conditional mean function of the control and the treatment group respectively, and ϵ_i denotes the error term, assumed to be zero-mean and independent of X_i. The estimand HTE is defined as

$$\tau(x) = \mathbb{E}\left[Y(1) - Y(0) \mid X = x\right] = \mu_1(x) - \mu_0(x),$$

which can also be expressed as the difference of the two group conditional mean functions.

*Marshall School of Business, University of Southern California, USA

1.2 Absolute error and relative error

In this paper, we focus on the evaluation and comparison of the HTE estimators using a test dataset of size n drawn from the super-population $\mathbb{P}$. We use $\hat{\tau}_1(x)$, $\hat{\tau}_2(x)$ to denote HTE estimators derived independently of the test dataset. When there is only one HTE estimator, we drop the subscript and use $\hat{\tau}(x)$. We highlight that the HTE estimators are provided to us and the HTE estimation problem itself is not the focus of this paper.

To quantify the accuracy of an HTE estimator $\hat{\tau}(x)$, the absolute error is defined as

$$\phi(\hat{\tau}(x)) := \mathbb{E}\left[(\hat{\tau}(X) - \tau(X))^2\right], \tag{1}$$

where the expectation is evaluated at the distribution $\mathbb{P}_X$ of the covariates. A smaller absolute error suggests a more accurate HTE estimator, and a zero error implies $\hat{\tau}(X) = \tau(X)$ with probability one. The relative error of two estimators $\hat{\tau}_1(x)$ and $\hat{\tau}_2(x)$ is quantified as the difference between their absolute errors

$$\delta(\hat{\tau}_1, \hat{\tau}_2) := \phi(\hat{\tau}_1(x)) - \phi(\hat{\tau}_2(x)) = \mathbb{E}\left[\hat{\tau}_1^2(X) - \hat{\tau}_2^2(X) - 2(\hat{\tau}_1(X) - \hat{\tau}_2(X))\tau(X)\right]. \tag{2}$$

A negative $\delta(\hat{\tau}_1, \hat{\tau}_2)$ indicates that $\hat{\tau}_1(x)$ is more accurate; otherwise, $\hat{\tau}_2(x)$ is more accurate.

2 Absolute error estimation based on influence functions

Consider a sequence of m HTE estimators $\hat{\tau}_j(x)$, $1 \le j \le m$. For the j-th HTE estimator $\hat{\tau}_j(x)$, we adopt the following influence function for $\phi(\hat{\tau}_j)$,

$$\begin{aligned}\psi(\phi(\hat{\tau}_j); Z) :=& \left((\mu_1(X) - \mu_0(X)) - \hat{\tau}(X)\right)^2 \\ &+ 2\left((\mu_1(X) - \mu_0(X)) - \hat{\tau}_j(X)\right) \cdot \left(\frac{W(Y - \mu_1(X))}{e(X)} - \frac{(1-W)(Y - \mu_0(X))}{1 - e(X)}\right) - \phi(\hat{\tau}_j).\end{aligned} \tag{3}$$

The one-step correction estimator associated with (3) is

$$\begin{aligned}\hat{\phi}(\hat{\tau}_j) :=& \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}(\phi(\hat{\tau}_j); Z_i) = \frac{1}{n}\sum_{i=1}^{n} \left((\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)) - \hat{\tau}_j(X_i)\right)^2 \\ &+ 2\left((\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)) - \hat{\tau}_j(X_i)\right) \cdot \left(\frac{W_i(Y_i - \tilde{\mu}_1(X_i))}{\tilde{e}(X_i)} - \frac{(1-W_i)(Y_i - \tilde{\mu}_0(X_i))}{1 - \tilde{e}(X_i)}\right).\end{aligned} \tag{4}$$

Here $\tilde{\mu}_0(x)$, $\tilde{\mu}_1(x)$, and $\tilde{e}(x)$ are estimators of the nuisance functions $\mu_0(x)$, $\mu_1(x)$, and $e(x)$ obtained on the test dataset[1]. The covariance matrix of the estimated evaluation errors, denoted by Σ, can be estimated by the empirical covariance matrix of $\hat{\phi}_i(\hat{\tau}_j)$, denoted by $\hat{\Sigma}$. We emphasize that the nuisance function estimators are shared by the absolute error estimators for all $\hat{\tau}_j$, which induces dependence among the estimated absolute errors.

The theorem below characterize the asymptotic distribution of the estimated absolute errors.

Theorem 2.1. *Assume the following conditions.*

(a) Y is bounded, $\eta < e(X) < 1 - \eta$ for some $\eta > 0$.

(b) The nuisance function estimators $\tilde{\mu}_0(x)$, $\tilde{\mu}_1(x)$, $\tilde{e}(x)$ obtained from the test data[2] satisfy $\mathbb{E}[(\tilde{\mu}_1(X) - \mu_1(X))^2]^{1/2}$, $\mathbb{E}[(\tilde{\mu}_0(X) - \mu_0(X))^2]^{1/2}$, $\mathbb{E}[(\hat{e}(X) - e(X))^2]^{1/2} = o_p(n^{-1/4})$, $1 \le k \le K$.

(c) The true absolute error $\phi(\hat{\tau}_j) > 0$, $1 \le j \le m$.

Then $\hat{\phi}(\hat{\tau})$, $\hat{\Sigma}(\hat{\phi}(\hat{\tau}))$ of Algorithm 1 satisfy

$$\frac{1}{\sqrt{n}}\left(\hat{\phi}(\hat{\tau}) - \phi(\hat{\tau})\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma(\hat{\phi}(\hat{\tau})\right), \quad \hat{\Sigma}(\hat{\phi}(\hat{\tau})) \xrightarrow{p} \Sigma(\hat{\phi}(\hat{\tau})).$$

In addition, the estimators $\hat{\phi}(\hat{\tau}_j)$, $1 \le j \le m$ are semi-parametrically efficient for $\phi(\hat{\tau}_j)$, respectively regarding the nonparametric model.

[1] We use cross-fitting (**chernozhukov2018double**) to ensure the independence of $\tilde{\mu}_0(x)$, $\tilde{\mu}_1(x)$, and $\tilde{e}(x)$ used by $\hat{\phi}_i(\hat{\tau})$ are independent of Y_i, W_i, X_i therein when computing $\hat{\phi}(\hat{\tau})$.

[2] When cross-fitting is used to compute the absolute error, we require (b) to hold for all $\tilde{\mu}_0^{-k}(x)$, $\tilde{\mu}_1^{-k}(x)$, $\tilde{e}^{-k}(x)$, $1 \le k \le K$.

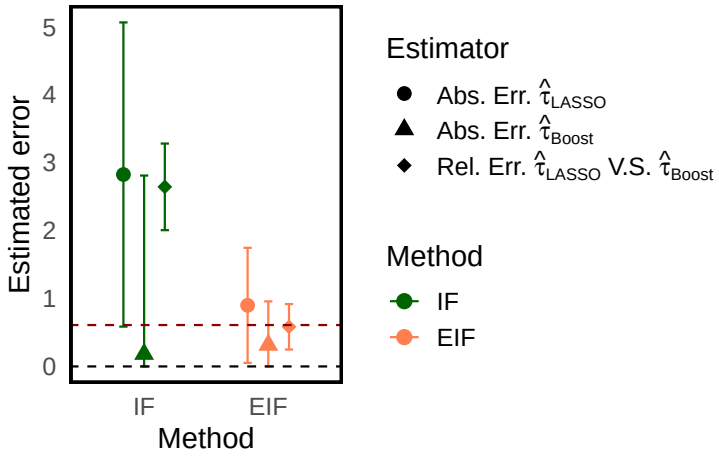

Figure 1: Relative error evaluation estimators are more effective in determining the better one of two similar HTE estimators. We consider two similar HTE estimators using LASSO for nuisance function estimation which only differ in the regularization hyperparameter. We consider the IF (**alaa2019validating**), EIF (our proposal) absolute, relative error confidence intervals. The 90% absolute error confidence intervals of both IF and EIF are too wide to distinguish the two HTE estimators, while the confidence interval of the relative error estimator is significantly narrower and find the better estimator $\hat{\tau}_2$ (the true relative error is indicated in dark red and $\hat{\tau}_2$ has a smaller error). We remark that even if IF's relative error confidence interval stays above zero, it does not contain the true relative error.

3 Winner selection

Given m estimators and their absolute error estimates as in Section 2, a natural choice for the best HTE estimator is the one with the smallest estimated absolute error. The challenge lies in rigorously testing whether the chosen HTE estimator is indeed associated with the smallest true absolute error.

3.1 Baseline approach

For the absolute error estimator in Section 2, the individual $1-\alpha$ confidence interval can be constructed as

$$[\underline{\hat{\phi}}(\hat{\tau}_j; 1-\alpha),\ \overline{\hat{\phi}}(\hat{\tau}_j; 1-\alpha)] := \left[\hat{\phi}(\hat{\tau}_j) - q_{1-\alpha/2}\hat{\Sigma}_{jj}^{1/2}(\hat{\phi}(\hat{\tau})),\ \hat{\phi}(\hat{\tau}_j) + q_{1-\alpha/2}\hat{\Sigma}_{jj}^{1/2}(\hat{\phi}(\hat{\tau}))\right]. \tag{5}$$

We can conclude with confidence $1-\alpha$ that $\hat{\tau}_j$ is the best estimator if $\overline{\hat{\phi}}(\hat{\tau}_j; 1-\alpha/m) < \underline{\hat{\phi}}(\hat{\tau}_{j'}; 1-\alpha/m)$ for all $j' \neq j$. Here we use Bonferroni correction and adopt the more conservative level α/m due to the presence of m individual confidence intervals. We call this selection method the baseline approach.

Note that the estimated absolute errors of different HTE estimators are typically positively correlated, as they are based on the same validation data and share the complex nuisance function estimators. The baseline approach, which holds for the worst-case dependence structure across the confidence intervals Theorem 2.1, fails to leverage this favorable positive correlation, and may therefore be conservative. To utilize the dependence structure, for two HTE estimators $\hat{\tau}_1(x)$, $\hat{\tau}_2(x)$, we demonstrated in a previous work that directly estimating and conducting inference for their relative error $\psi(\delta(\hat{\tau}_1, \hat{\tau}_2); Z)$ is more powerful and robust, both theoretically and empirically (see Figure 1 for a numerical example). In the following section, we extend the relative-error-based approach for selecting between two HTE estimators to the more general case of choosing the best estimator among m candidates.

3.2 Proposal

In this section, we design a test based on the asymptotic distribution of the absolute errors outlined in Theorem 2.1, particularly using the covariance matrix $\Sigma(\hat{\phi}(\hat{\tau}))$ to leverage the dependence structure.

Mathematically, we use J to denote the HTE estimator with the smallest true absolute error, and $\hat{J}$ to denote the HTE estimator with the smallest estimated true absolute error. The alternative hypothesis we consider is $H_1 : \hat{J} = J$, and the associated null hypothesis is $H_0 : \hat{J} \neq J$. We aim to construct a test which controls the conditional type I error $\mathbb{P}(H_0 \text{ rejected} \mid \hat{J}) \leq \alpha$, a stronger notion of the type I error $\mathbb{P}(H_0 \text{ rejected})$ marginalized over $\hat{J}$. Without loss of generality, below we assume $\hat{J} = 1$[3].

We express H_0 given $\hat{J} = 1$ using the true absolute errors $\phi(\hat{\tau}_j)$ and the estimated absolute errors $\hat{\phi}(\hat{\tau}_j)$,

$$\begin{aligned} & H_0 : \phi(\hat{\tau}_1) \geq \phi(\hat{\tau}_j) \text{ for some } j \neq 1, \text{ given that } \hat{\phi}(\hat{\tau}_1) < \hat{\phi}(\hat{\tau}_j) \text{ for all } j \neq 1. \\ \Leftrightarrow\ & H_0 : \cup_{j \neq 1} H_{1 \geq j,0} \text{ given that } \hat{\phi}(\hat{\tau}_1) < \hat{\phi}(\hat{\tau}_j) \text{ for all } j \neq 1, \ H_{1 \geq j,0} := \{\phi(\hat{\tau}_1) \geq \phi(\hat{\tau}_j)\}. \end{aligned}$$

The null hypothesis is composite and is the union of $m-1$ pairwise comparisons $\phi(\hat{\tau}_1) \geq \phi(\hat{\tau}_j)$. We decompose the task of testing H_0 into testing a series of simpler nulls $H_{1 \geq j,0}$ and then aggregating the individual p-values:

1. For all $j \neq 1$, construct a p-value, denoted by $P_{1 \geq j}$, for the simpler null $H_{1 \geq j,0}$ given that $\hat{\phi}(\hat{\tau}_1) < \hat{\phi}(\hat{\tau}_j)$.
2. Compute the aggregated p-value for H_0 as $P = \max_{j \neq 1} P_{1 \geq j}$.

The p-value P controls the conditional type I error as long as each $P_{1 \geq j}$ controls the conditional type I error for $H_{1 \geq j,0}$. Below we elaborate the construction of $P_{1 \geq j}$.

Note that for $H_{1 \geq j,0}$ given that $\hat{\phi}(\hat{\tau}_1) < \hat{\phi}(\hat{\tau}_j)$ for all $j \neq 1$, the conditioning set involves $\hat{\phi}(\hat{\tau}_{j'})$ for $j' \neq j, 1$, but the null hypothesis does not specify their true values $\phi(\hat{\tau}_{j'})$. To ensure that the null distribution of the test statistic is computable under the sub-null $H_{1 \geq j,0}$, we condition on a sufficient statistic of the nuisance parameters $\phi(\hat{\tau}_{j'})$. Explicitly, we define the relative error estimator as $\hat{\delta}(\hat{\tau}_1, \hat{\tau}_j) := \hat{\phi}(\hat{\tau}_1(x)) - \hat{\phi}(\hat{\tau}_j(x))$, and denote its asymptotic covariance matrix by Σ_δ. Let $A_j = I_{m-1} - \Sigma_\delta e_j (e_j^\top \Sigma_\delta e_j)^{-1} e_j^\top$, where $e_j \in \mathbb{R}^{m-1}$ is the vector with a one in the j-th position and zeros elsewhere. We decompose $\hat{\delta}(\hat{\tau}_1, \hat{\tau}) = \Sigma_\delta e_j (e_j^\top \Sigma_\delta e_j)^{-1} e_j^\top \hat{\delta}(\hat{\tau}_1, \hat{\tau}) + A_j \hat{\delta}(\hat{\tau}_1, \hat{\tau})$. Asymptotically, $A_j \hat{\delta}(\hat{\tau}_1, \hat{\tau})$ and $\Sigma_\delta e_j (e_j^\top \Sigma_\delta e_j)^{-1} e_j^\top \hat{\delta}(\hat{\tau}_1, \hat{\tau}) = \Sigma_\delta e_j (e_j^\top \Sigma_\delta e_j)^{-1} \hat{\delta}(\hat{\tau}_1, \hat{\tau}_j)$ are uncorrelated and thus independent. We consider the test statistics

$$\hat{\delta}(\hat{\tau}_1, \hat{\tau}_j) \mid \hat{\delta}(\hat{\tau}_1, \hat{\tau}_j) < 0, \ j \neq 1, \ A_j \hat{\delta}(\hat{\tau}_1, \hat{\tau}), \tag{6}$$

where we further condition on the value of $M-1$ random variables $A_j \hat{\delta}(\hat{\tau}_1, \hat{\tau})$. According to Lee et al. 2016, for any $c \in \mathbb{R}^{m-1}$, the constraint $\hat{\delta}(\hat{\tau}_1, \hat{\tau}_{-1}) < 0$ given that $A_j \hat{\delta}(\hat{\tau}_1, \hat{\tau}) = c$ is equivalent to $\hat{\delta}(\hat{\tau}_1, \hat{\tau}_j) \in (l_j(c), u_j(c))$ for some constants $l_j(c)$, $u_j(c) \in \mathbb{R} \cup \{\pm\infty\}$, and the conditional distribution of the test statistic (6) given $A_j \hat{\delta}(\hat{\tau}_1, \hat{\tau}) = c$ is a truncated normal,

$$\hat{\delta}(\hat{\tau}_1, \hat{\tau}_j) \mid \hat{\delta}(\hat{\tau}_1, \hat{\tau}_j) < 0, \ j \neq 1, \ A_j \hat{\delta}(\hat{\tau}_1, \hat{\tau}) = c, \sim \text{TN}\left(\delta(\hat{\tau}_1, \hat{\tau}_j), \Sigma_{\delta,jj}, l_j(c), u_j(c)\right).$$

Let $F_{\mu,\sigma^2}^{[a,b]}$ denote the CDF of a $N(\mu, \sigma^2)$ random variable truncated to the interval $[a, b]$. We define the p-value for the sub-null $H_{1 \geq j,0}$ as

$$P_{1 \geq j} := F_{0,\Sigma_{\delta,jj}}^{(l(c),u(c))}(\hat{\delta}(\hat{\tau}_1, \hat{\tau}_j)) \mid \hat{J} = 1, \tag{7}$$

which controls the conditional type I error. When there are only $m = 2$ hypotheses, $A_j = 0$, no additional conditioning is enforced, and $P_{1 \geq j}$ reduces to the p-value based on the relative error estimator in the previous work.

4 Simulation

[ZG: TODO].

[3]The procedure applies to $\hat{J} = j$ for any $1 \leq j \leq m$.

Algorithm 1 Absolute error

Input: A sequence of HTE estimator $\hat{\tau}_j(x)$, $1 \leq j \leq m$, test data $Z_i = (X_i, W_i, Y_i)$, $1 \leq i \leq n$, methods of estimating nuisance functions $\mu_0(x)$, $\mu_1(x)$, $e(x)$, number of folds K for cross-fitting, confidence level $1 - \alpha$.

Randomly split the test dataset into K folds of approximately equal size. Denote the k-th fold by D_k.

for $k = 1, \ldots, K$ **do**

  Apply the nuisance function estimators to test folds $\cup_{k' \neq k} D_{k'}$ and obtain $\tilde{\mu}_0^{-k}(x)$, $\tilde{\mu}_1^{-k}(x)$, and $\tilde{e}^{-k}(x)$.

end for

Compute the one-step correction estimator based on Equation (4). For $i \in D_k$, $1 \leq j \leq m$,

$$\begin{aligned}\hat{\psi}_i^+(\phi(\hat{\tau}_j); Z_i) := &\left((\tilde{\mu}_1^{-k}(X_i) - \tilde{\mu}_0^{-k}(X_i)) - \hat{\tau}_j(X_i)\right)^2 \\ &+ 2\left((\tilde{\mu}_1^{-k}(X_i) - \tilde{\mu}_0^{-k}(X_i)) - \hat{\tau}_j(X_i)\right) \cdot \left(\frac{W_i(Y_i - \tilde{\mu}_1^{-k}(X_i))}{\tilde{e}^{-k}(X_i)} - \frac{(1 - W_i)(Y_i - \tilde{\mu}_0^{-k}(X_i))}{1 - \tilde{e}^{-k}(X_i)}\right),\end{aligned}$$

and take the average

$$\hat{\phi}(\hat{\tau}_j) := \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in D_k} \hat{\psi}_i^+(\phi(\hat{\tau}_j); Z_i).$$

Here $\hat{\psi}_i^+(\phi(\hat{\tau}_j); Z_i)$ and $\hat{\psi}_i(\phi(\hat{\tau}_j); Z_i)$ differ in a constant.

Compute the estimator of the covariance matrix of $\hat{\phi}(\hat{\tau})$,

$$\hat{\Sigma}(\hat{\phi}(\hat{\tau})) := \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\psi}_i^+(\phi(\hat{\tau}); Z_i) - \hat{\phi}(\hat{\tau})\right) \left(\hat{\psi}_i^+(\phi(\hat{\tau}); Z_i) - \hat{\phi}(\hat{\tau})\right)^\top.$$

Output: Estimated errors $\hat{\phi}(\hat{\tau}_j)$, the estimator of its variance $\hat{\Sigma}(\hat{\phi}(\hat{\tau}))$.