Trustworthy assessment of heterogeneous treatment effect estimator via analysis of relative error: Supplementary Materials

Notations. For a causal estimand ϕ, we use $\text{IF}(\phi)$ to denote an influence function of ϕ. We derive the influence functions assuming the covariates are discrete as in Kennedy 2022. We use $\mathbb{P}_n$ to denote the distribution of n test data points, and $\mathbb{E}_n$ be the corresponding expectation (the average over n test data points).

A. PROOFS

Proof of Theorem 1. We first derive the efficient influence function for the absolute error (1). We next prove the asymptotic distribution of the absolute error estimator (4).

Derivation of the efficient influence function. We consider non-parametric models, where the tangent space contains the entire Hilbert space of mean-zero, finite-variance functions (Tsiatis 2006). In this case, the influence function is unique, and it must be efficient. Below we derive this influence function by applying differentiation rules and using putative influence functions for common causal estimands (Kennedy 2022).

By the linearity of influence functions,

$$\text{IF}(\phi(\hat{\tau})) = \underbrace{\text{IF}\left(\mathbb{E}\left[\hat{\tau}^2(X_i)\right]\right)}_{:=(\text{a})} - 2\underbrace{\text{IF}\left(\mathbb{E}\left[\hat{\tau}(X_i)\tau(X_i)\right]\right)}_{:=(\text{b})} + \underbrace{\text{IF}\left(\mathbb{E}\left[\tau^2(X_i)\right]\right)}_{:=(\text{c})}.$$

We deal with (a) to (c) one by one. For (a), $\hat{\tau}(x)$ is a known function of the covariates, and

$$(\text{a}) = \hat{\tau}^2(X_i) - \mathbb{E}[\hat{\tau}^2(X_i)]. \tag{8}$$

For (b), $\mathbb{E}\left[\hat{\tau}(X_i)\tau(X_i)\right]$ is a weighted average treatment effect with weights $\hat{\tau}(X_i)$, and

$$(\text{b}) = \hat{\tau}(X_i) \cdot \left(\frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} + \mu_1(X_i) - \frac{(1-W_i)(Y_i - \mu_0(X_i))}{1-e(X_i)} - \mu_0(X_i)\right) - \mathbb{E}[\hat{\tau}(X_i)\tau(X_i)]. \tag{9}$$

For (c), we rewrite the second moment as a finite sum, and

$$\begin{aligned}
(\text{c}) &= \sum_x \text{IF}(\tau^2(x))\mathbb{P}_X(dx) + \sum_x \tau^2(x)\text{IF}(\mathbb{P}_X(dx)) \quad \text{(Linearity, product rule of influence functions)} \\
&= \sum_x 2\tau(x)\text{IF}(\tau(x))\mathbb{P}_X(dx) + \sum_x \tau^2(x)\text{IF}(\mathbb{P}_X(dx)) \quad \text{(Chain rule of influence functions)} \\
&= 2(\mu_1(X_i) - \mu_0(X_i))\left(\frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1-W_i)(Y_i - \mu_0(X_i))}{1-e(X_i)}\right) \\
&\quad + (\mu_1(X_i) - \mu_0(X_i))^2 - \mathbb{E}[\tau^2(X_i)]. \quad \text{(Influence function of } \tau(x), \mathbb{P}_X(dx).)
\end{aligned} \tag{10}$$

Finally, combing (a) to (c) and we have finished the derivation of the influence function for the absolute error (1).

Asymptotic distribution of the one-step correction estimator. As in Kennedy 2022, we decompose the error into three terms,

$$\begin{aligned}
\hat{\phi}(\hat{\tau}) - \phi(\hat{\tau}) = &\underbrace{\mathbb{E}_n[\psi(\phi(\hat{\tau}); Z_i)] - \mathbb{E}[\psi(\phi(\hat{\tau}); Z_i)]}_{:=\text{S}} + \underbrace{\phi(\hat{\tau}) - \tilde{\phi}(\hat{\tau}) + \mathbb{E}[\hat{\psi}(\hat{\tau}; Z_i)]}_{:=R} \\
&+ \underbrace{\mathbb{E}_n[\hat{\psi}(\phi(\hat{\tau}); Z_i) - \psi(\phi(\hat{\tau}); Z_i)] - \mathbb{E}[\hat{\psi}(\phi(\hat{\tau}); Z_i) - \psi(\phi(\hat{\tau}); Z_i)]}_{:=N},
\end{aligned}$$

where $\tilde{\phi}(\hat{\tau})$ denote the absolute error regarding the estimated nuisance functions. We deal with S, N, R one by one. For S, the U-statistic term, by the central limit theorem,

$$\sqrt{n}S \xrightarrow{d} \mathcal{N}(0, V(\hat{\phi}(\hat{\tau}))). \tag{11}$$

For N, the empirical process term, by the boundedness assumption (Y_i is bounded, $e(X_i)$ is bounded away from zero and one, and the nuisance function estimators $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$ are bounded, $\hat{e}(x)$ is bounded away from zero and one), and that nuisance function estimators are consistent,

$$n\mathsf{Var}(N) = \mathsf{Var}\left(\hat{\psi}(\phi(\hat{\tau}); Z_i) - \psi(\phi(\hat{\tau}); Z_i)\right) \to 0. \tag{12}$$

By Chebyshev's inequality, we have $\sqrt{n}N \xrightarrow{p} 0$. For R, the remainder term, we decompose it into the remainder term of (a), (b), and (c) in (8), (9), and (10), respectively. R(a) is exactly zero. R(b) is the remainder term of a weighted average treatment effect, and satisfies $o_p(n^{-1/2})$ under the boundedness assumption and the assumption that $\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$, and $\hat{e}_1(x)$ converge at rate $o_p(n^{-1/4})$ in L_2 norm. For R(c),

$$\begin{aligned} R(\mathrm{c}) &= -\mathbb{E}\left[((\mu_1(X_i) - \mu_0(X_i)) - (\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)))^2\right] + 2\mathbb{E}\left[(\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i))r(X_i)\right], \\ r(x) &:= \left(\frac{e(x)}{\tilde{e}(x)} - 1\right)(\mu_1(x) - \tilde{\mu}_1(x)) - \left(\frac{1-e(x)}{1-\tilde{e}(x)} - 1\right)(\mu_0(x) - \tilde{\mu}_0(x)). \end{aligned} \tag{13}$$

Again we use the boundedness assumption, and the $o_p(n^{-1/4})$ convergence assumption of $\tilde{\mu}_1(x)$, $\tilde{\mu}_0(x)$, and $\tilde{e}(x)$,

$$\begin{aligned} \mathbb{E}\left[|r(X_i)|\right] &\le \mathbb{E}\left[\left|\left(\frac{1-e(X_i)}{1-\tilde{e}(X_i)} - 1\right)(\mu_0(X_i) - \tilde{\mu}_0(X_i))\right|\right] + \mathbb{E}\left[\left|\left(\frac{1-e(X_i)}{1-\tilde{e}(X_i)} - 1\right)(\mu_0(X_i) - \tilde{\mu}_0(X_i))\right|\right] \\ &\le \mathbb{E}^{1/2}\left[\left(\frac{1-e(X_i)}{1-\tilde{e}(X_i)} - 1\right)^2\right]\mathbb{E}^{1/2}\left[(\mu_0(X_i) - \tilde{\mu}_0(X_i))^2\right] \\ &\quad + \mathbb{E}^{1/2}\left[\left(\frac{1-e(X_i)}{1-\tilde{e}(X_i)} - 1\right)^2\right]\mathbb{E}^{1/2}\left[(\mu_0(X_i) - \tilde{\mu}_0(X_i))^2\right] = o_p(n^{-1/2}). \end{aligned}$$

By the boundedness assumption, there exists $C > 0$ such that

$$\mathbb{E}\left[|(\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i))r(X_i)|\right] \le C \cdot \mathbb{E}\left[|r(X_i)|\right] = o_p(n^{-1/2}).$$

Combining S, N, R and we have proved $\sqrt{n}(\hat{\phi}(\hat{\tau}) - \phi(\hat{\tau})) \xrightarrow{d} \mathcal{N}(0, V(\hat{\phi}(\hat{\tau})))$. The above derivation considers pre-fixed nuisance function estimators. As for cross-fitting in Algorithm 1, we apply the argument in Chernozhukov et al. 2018.

For the variance estimator, by the boundedness assumption and the law of large number,

$$\hat{V}(\hat{\phi}(\hat{\tau})) \xrightarrow{p} V(\hat{\phi}(\hat{\tau})).$$

By the assumption that $\mathbb{P}(\hat{\tau}(X_i) \neq \tau(X_i)) > 0$, we have $V(\hat{\phi}(\hat{\tau})) > 0$. Finally, by Slutsky's lemma, we have finished the proof of Theorem 1.

Proof of Theorem 2. Similar to the proof of Theorem 1, we first derive the efficient influence function for the relative error (2). We next prove the asymptotic distribution of the absolute error estimator (7).

Derivation of the efficient influence function. Note that $\delta(\hat{\tau}_1, \hat{\tau}_2) = \phi(\hat{\tau}_1) - \phi(\hat{\tau}_2)$. By the linearity of influence functions, the influence function for $\delta(\hat{\tau}_1, \hat{\tau}_2)$ is simply the difference between the influence function of $\phi(\hat{\tau}_1)$ and that of $\phi(\hat{\tau}_2)$ derived in Theorem 1. Alternatively, recognize that the estimand $\delta(\hat{\tau}_1, \hat{\tau}_2)$ is a weighted average treatment effect, for which the influence function is a well-established result. Both methods lead to the influence function (6).

Asymptotic distribution of the one-step correction estimator. Given that the estimand $\delta(\hat{\tau}_1, \hat{\tau}_2)$ is a weighted average treatment effect, the asymptotic distribution of the one-step correction estimator, under the assumptions of Theorem 2, follows from the standard result (Kennedy 2022). Regarding cross-fitting, we apply the analysis in Chernozhukov et al. 2018.

Additionally, the degeneracy assumption $\mathbb{P}(\hat{\tau}_1(X_i) \neq \hat{\tau}_2(X_i)) > 0$ is required to ensure that the one-step correction estimator, when scaled by $\sqrt{n}$, has a non-zero variance. This also arises in the proof of Theorem 1. □

Proposition 1. *There exists $C > 0$ such that*

$$V(\hat{\delta}(\hat{\tau}_1, \hat{\tau}_2)) \leq C \cdot \mathbb{E}[(\hat{\tau}_1(X) - \hat{\tau}_2(X))^2].$$

This together with Theorem 2 implies the width of the confidence interval of the relative error estimator (7) *based on Theorem 2 is of order* $\sqrt{\mathbb{E}[(\hat{\tau}_1(X) - \hat{\tau}_2(X))^2]} \cdot n^{-1/2}$.

Proof of Proposition 1. By Cauchy-Schwarz inequality,

$$\begin{aligned} V(\hat{\delta}(\hat{\tau}_1, \hat{\tau}_2)) &\leq \mathbb{E}\left[(\hat{\tau}_2(X_i) - \hat{\tau}_1(X_i))^2\right] \cdot \mathbb{E}\left[\left(\frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} + \mu_1(X_i) - \frac{(1 - W_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} - \mu_0(X_i)\right)^2\right] \\ &\leq C \cdot \mathbb{E}\left[(\hat{\tau}_2(X_i) - \hat{\tau}_1(X_i))^2\right]. \quad \text{(Boundedness assumption)} \end{aligned}$$

□

Corollary 1 (DINA)**.** *Theorem 1, Theorem 2 hold for the causal estimands in Section 4.3.*

Proof of . To derive the influence functions for the causal estimand in Section 4.3, we can apply the chain rule of influence functions to the proofs of Theorem 1, Theorem 2.

For the asymptotic distribution of the one-step correction estimator, we can still apply the arguments in Theorem 1 and Theorem 2. □

B ADDITIONAL EMPIRICAL EXPERIMENTS

We provide additional demonstrations of the hypothetical real data analysis in Section 5. The datasets, HTE estimators, and evaluation methods remain the same. In Section B.1, we present the widths of the confidence intervals, which can serve as an indication of the statistical power (shorter confidence interval suggests higher power). In Section B.2, we report the estimation errors for both the absolute and relative error estimators, that is $|\hat{\phi}(\hat{\tau}) - \phi(\hat{\tau})|$ and $|\hat{\delta}(\hat{\tau}_1, \hat{\tau}_2) - \delta(\hat{\tau}_1, \hat{\tau}_2)|$.

B.1 Widths of confidence intervals

In Figure 5, we report the widths of the estimated absolute (LASSO, Boosting) and relative (LASSO vs. Boosting) error 90% confidence intervals across three methods (plug-in, IF, EIF) over four scenarios from the ACIC competition data ((a) to (d)). Shorter widths indicate higher power of identifying the better HTE estimator.

We make the following observations:

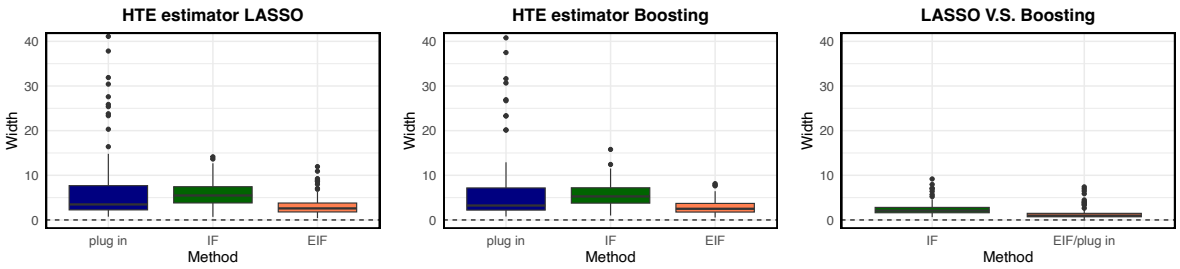
- Across all settings and methods, the confidence intervals for the relative error estimators are significantly shorter than those for the absolute error estimators.
- For the absolute error estimators, our proposal (EIF) performs significantly better than the estimator in Alaa and Van Der Schaar 2019 (IF) in scenario (d), and the two methods are comparable in scenarios (a) to (c). The plug-in estimator (plug in) produces the widest intervals across all settings.

B.2 Error of estimated errors

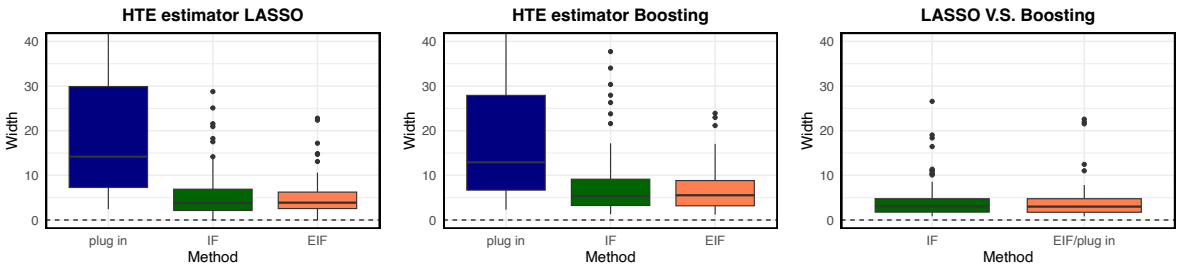
In Figure 6, we report the error of the estimated absolute (LASSO, Boosting)/relative (LASSO V.S. Boosting) error across three methods (plug in, IF, EIF) over four scenarios of the ACIC competition data ((a) to (d)). Error of the estimated error is defined as the absolute difference between the estimated error and the corresponding oracle value. Smaller errors suggest that the estimator has higher accuracy and is thus more favorable.

We make the following observations.

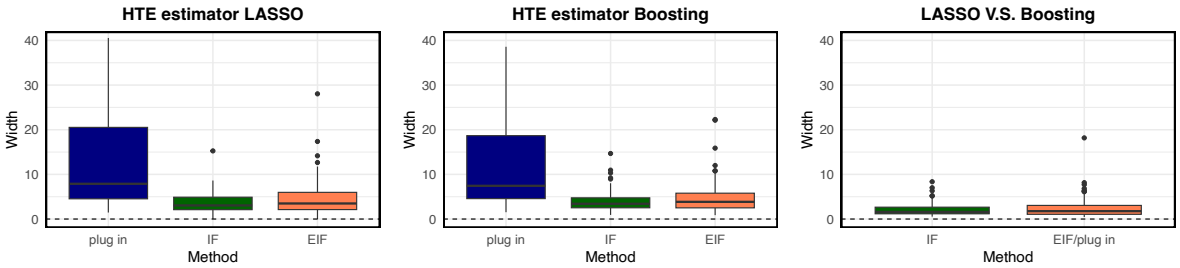
- Across all settings and methods, the errors of the relative error estimators are significantly smaller than those for the absolute error estimators.
- For the absolute error estimators, our proposal (EIF) performs significantly better than the estimator in Alaa and Van Der Schaar 2019 (IF) and the plug-in estimator (plug in) in scenario (d). The plug-in estimator is unfavorable from scenarios (b) to (d).



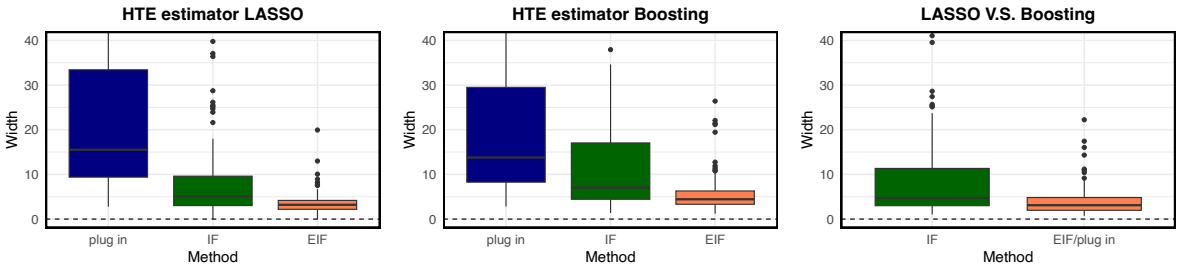
(a) Linear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$



(b) Nonlinear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$

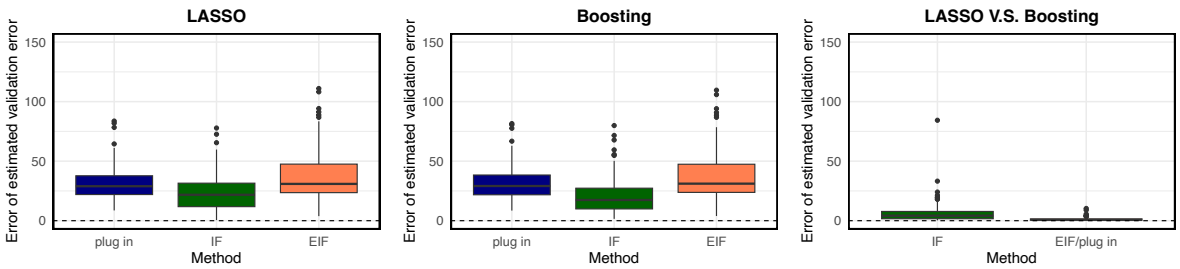


(c) Linear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$

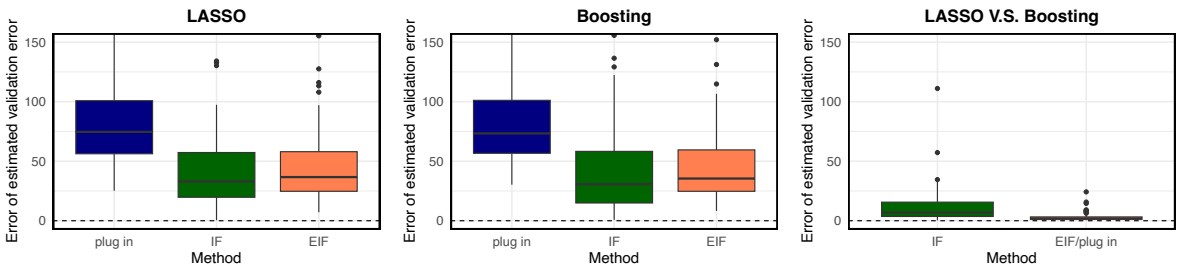


(d) Nonlinear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$

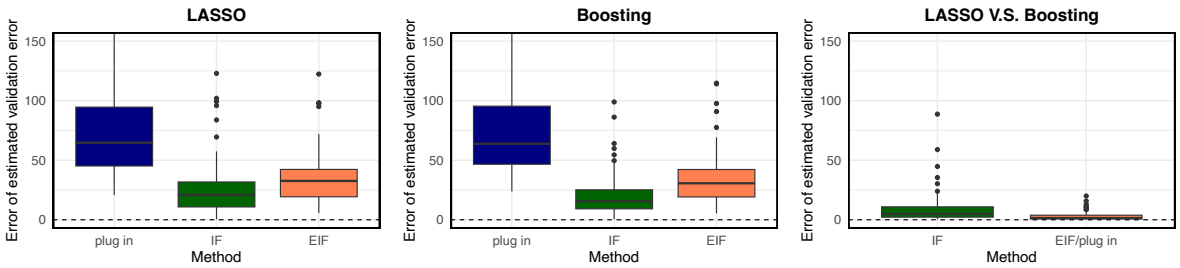
Figure 5: Width of the estimated absolute (LASSO, Boosting)/relative (LASSO V.S. Boosting) error's 90% confidence intervals across three methods (plug in, IF, EIF) over four scenarios of the ACIC competition data ((a) to (d)).



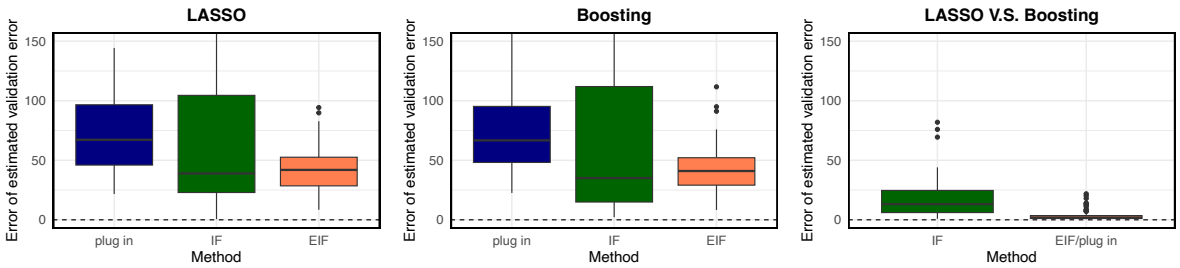
(a) Linear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$



(b) Nonlinear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$

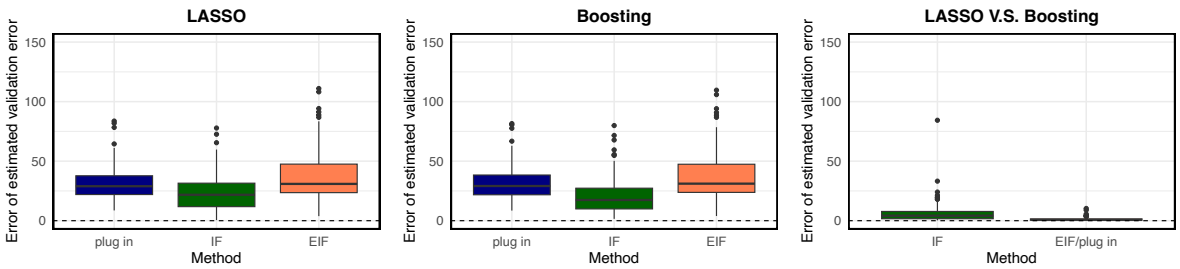


(c) Linear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$

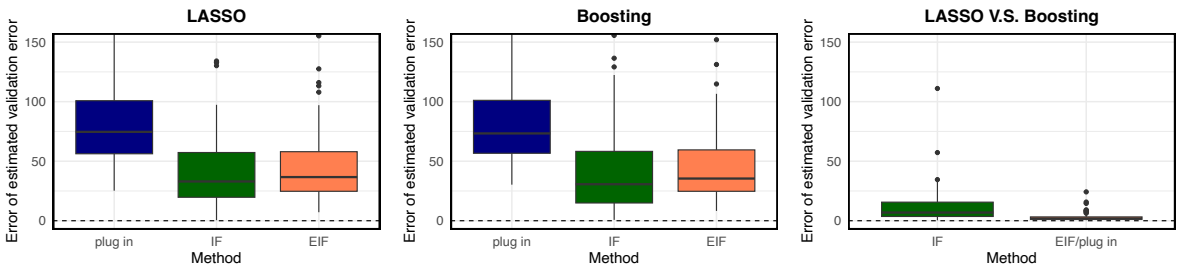


(d) Nonlinear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$

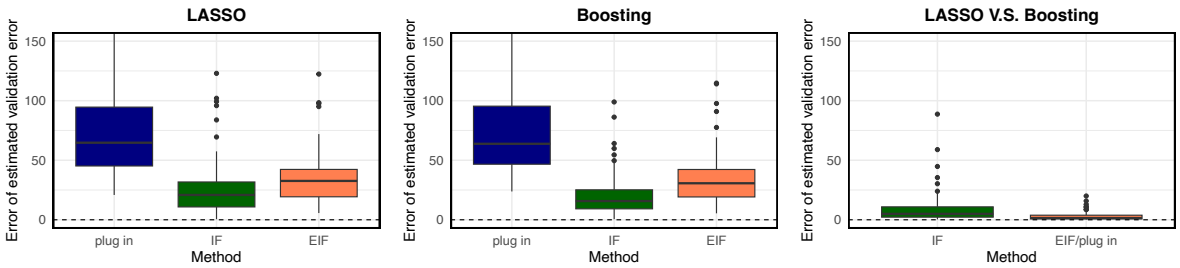
Figure 6: Error of the estimated absolute (LASSO, Boosting)/relative (LASSO V.S. Boosting) error across three methods (plug in, IF, EIF) over four scenarios of the ACIC competition data ((a) to (d)). Error of the estimated error is defined as the absolute difference between the estimated error and the corresponding oracle value.



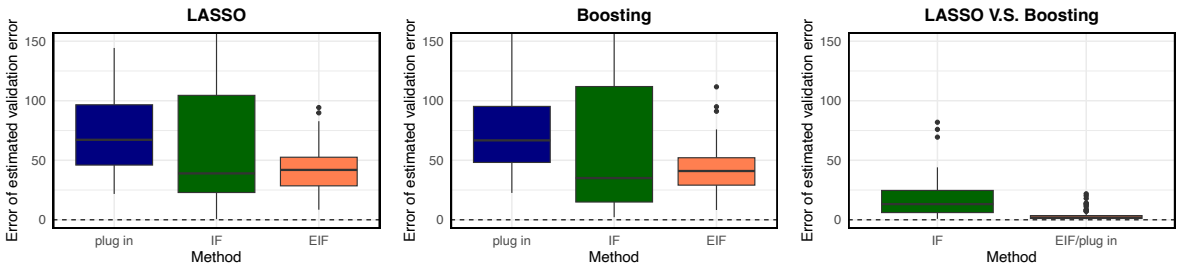
(a) Linear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$



(b) Nonlinear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$



(c) Linear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$



(d) Nonlinear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$

Figure 6: Error of the estimated absolute (LASSO, Boosting)/relative (LASSO V.S. Boosting) error across three methods (plug in, IF, EIF) over four scenarios of the ACIC competition data ((a) to (d)). Error of the estimated error is defined as the absolute difference between the estimated error and the corresponding oracle value.