Calibration Error for Heterogeneous Treatment Effects

Yizhe Xu, PhD[1] and Steve Yadlowsky, PhD[2]

[1] Stanford Center for Biomedical Informatics Research Institute, Stanford University; [2] Google Research, Brain Team; Correspondence: yizhex@stanford.edu

Introduction

Recently, many researchers have advanced data-driven methods for modeling heterogeneous treatment effects (HTEs). Even still, estimation of HTEs is a difficult task: these methods frequently over- or under-estimate the treatment effects, leading to poor calibration of the resulting models. While many methods exist for evaluating the calibration of prediction and classification models, formal approaches to assess the calibration of HTE models are limited to the calibration slope. In this paper, we define an analogue of the (l_2) expected calibration error for HTEs and propose a robust estimator. Our approach is motivated by doubly robust treatment effect estimators, making it unbiased, and resilient to confounding, overfitting, and high-dimensionality issues. Furthermore, our method is straightforward to adapt to many structures under which treatment effects can be identified, including randomized trials, observational studies, and survival analysis. We illustrate how to use our proposed metric to evaluate the calibration of learned HTE models through the application to the CRITEO-UPLIFT Trial.

Problem Setup

Consider the problem of learning a heterogeneous treatment effect model in the form of the conditional average treatment effect (CATE), $\tau(x) = \mathrm{E}[Y(1) - Y(0) \mid X = x]$ of a binary treatment $W \in \{0,1\}$ on a scalar outcome $Y \in \mathbb{R}$, in the presence of fully observed confounding variables $\mathrm{X} \in \mathbb{R}^d$ that affect treatment assignment and the outcome. The calibration of an estimator $\hat{\tau}(x)$ is

$$\gamma_{\hat{\tau}}(\delta) = \mathrm{E}[Y(1) - Y(0) \mid \hat{\tau}(X) = \delta],$$

where $\hat{\tau}(X)$ and $\gamma_{\hat{\tau}}(\delta)$ are the predicted and observed CATEs, respectively, and $\hat{\tau}(X)$ is *miscalibrated* if $\gamma_{\hat{\tau}}(\delta) \neq \delta$.

We study the l_p -Expected Calibration Error for predictors of Treatment Heterogeneity (l_p -ECETH), defined as

$$\theta = \mathrm{E}[|\gamma_{\hat{\tau}}(\Delta) - \Delta|^p],$$

where $\Delta = \hat{\tau}(X)$.

Funding Support: Partially supported by R01 HL144555 from the National Heart, Lung, and Blood Institute (NHLBI).

References

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evalu- ating treatment prioritization rules via rank- weighted average treatment effects. arXiv preprint arXiv:2111.07966, 2021.

Methods

Calibration Function Estimation

Given the independent and identically distributed sample of observations $(Y_i, W_i, X_i)_{i=1}^n$ that is used to learn a CATE model for $\hat{\tau}(.)$, we can estimate the calibration function $\gamma_{\hat{\tau}}(\delta)$ by taking advantage of carefully constructed "scores" that are a surrogate for the CATE. Let Γ_i be some function of (Y_i, W_i, X_i) such that

$$\mathrm{E}[\Gamma_i | X = x] = \tau(x) \quad (1)$$

Then, we partition the range of predictions R into K equally sized bins and apply the histogram model as

$$\hat{\gamma}_{\hat{\tau}}(\delta) = \frac{1}{|I_k|}\sum_{i \in I_k} \Gamma_i, \quad (2)$$

for any $\delta \in R_k$. $|I_k|$ indices the number of predictions in bin I_k.

Augmented Inverse Propensity Weighted (AIPW) Score

Given the propensity of getting treated π and an outcome prediction model $\hat{\mu}(\mathrm{x}, \mathrm{w})$, the AIPW score is

$$\Gamma_i^{aipw} = \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0) + \frac{W_i - \pi}{\pi(1-\pi)}(Y_i - \hat{\mu}(X_i, W_i))$$

Remarks

a) In randomized trials, the AIPW score helps to reduce the variance of Γ_i as compared to an IPW score

b) In observational studies, π is replaced with the estimated propensity score $\hat{\pi}$ for both scores, and the AIPW score is more robust to confounding than the IPW score

c) When data are from observational studies or are right-censored, under unconfoundedness and non-informative censoring, respectively, $\hat{\Gamma}_i^{aipw}$ almost satisfies condition (1) with an approximation error e_i, which goes to 0 at an appropriate rate.

l_2-ECETH Estimation

Given the estimate $\hat{\gamma}_{\hat{\tau}}(\delta)$ from (2), we propose the robust (i.e., debiased) estimator for HTE calibration error as

$$\hat{\theta}_{robust} = \frac{1}{n}\sum_{i=1}^n (\hat{\Gamma}_i - \Delta_i)(\hat{\gamma}_{\hat{\tau}}(\Delta_i) - \Delta_i), \quad (3)$$

where $\Delta_i = \hat{\tau}(X_i)$.

Remarks

a) We show that $\hat{\theta}_{robust}$ is asymptotically linear under appropriate regularity conditions

b) We show that $\hat{\theta}_{robust}$ results in less bias than the plug-in estimator $\hat{\theta}_{plug} = \frac{1}{n}\sum_{i=1}^n (\hat{\gamma}_{\hat{\tau}}(\Delta_i) - \Delta_i)^2$

c) To ensure the approximation error e_i is negligible, we let $\hat{\gamma}_{\hat{\tau}}(.)$ and $\hat{\Gamma}_i$ to be independent by making the leave-one-out (LOO) correction for $\hat{\gamma}_{\hat{\tau}}(\delta)$ in (2) as $\hat{\gamma}_{\hat{\tau}}^{-i} = \frac{1}{|I_k|-1}\sum_{j \neq i, i \in I_k} \hat{\Gamma}_j$

d) One may conduct a practical hypothesis testing of $H_0: \theta > \varepsilon$, for some $\varepsilon > 0$. Rejecting this hypothesis suggests that there is enough evidence to believe that the model is calibrated enough for practical use.

Simulation Study

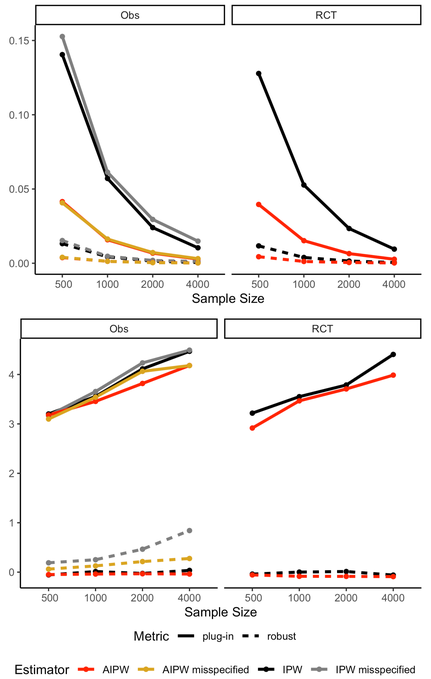

FIGURE 1: Estimation Performance of Different Calibration Error Estimators. The top and bottom panels show results of mean squared error and standardized bias, respectively. Overall, using the robust estimator with AIPW scores resulted in the best calibration error estimates.

Table 1: Performance of the robust ECETH estimator under the high-dimensional settings. The robust estimator performed reasonably well in low dimension situations

N	P	Bias	S.E.	S.bias	MSE
500	50	0.0053	0.0702	0.0759	0.0050
1000	50	0.0001	0.0366	0.0037	0.0013
	100	0.0060	0.0412	0.1454	0.0017
2000	50	0.0017	0.0228	0.0738	0.0005
	100	0.0044	0.0232	0.1919	0.0006
	200	0.0073	0.0257	0.2854	0.0007
4000	50	0.0008	0.0134	0.0616	0.0002
	100	0.0035	0.0136	0.2569	0.0002
	200	0.0070	0.0150	0.4639	0.0003
	400	0.0102	0.0166	0.6134	0.0004

Simulation Scheme

- Simulation scenarios: 1) RCT and observational study designs, 2) three levels of miscalibration, 3) model misspecification: whether the treatment model is misspecified, and 4) low- and high-dimensional covariates
- We compared plug-in and robust l_2-ECETH estimators where the calibration function is estimated using IPW or AIPW scores
- Evaluation metrics: 1) Bias, 2) standard error, 2) standardized bias, 4) mean squared error
- Main results:
  - The plug-in estimator underperformed the robust estimator across all scenarios
  - Under RCTs, the AIPW score yielded much smaller MSEs than the IPW score
  - Under observational studies, the AIPW score was more robust to misspecified treatment model than the IPW score
  - The robust estimator showed decent performance under high-dimensional settings

Data Application

- CRITEO is a large-scale trial where a portion of users are randomly prevented from being targeted by advertising.
- Binary outcome: whether a user visited/converted on the advertiser website during the test period
- We randomly sampled 640,000 users, split the data into training and testing sets, and derived two CATE models: causal forest and S-learner of generalized random forest. Implementations are via R package grf
- We evaluated l_2-ECETH using the $\hat{\theta}_{robust}$ estimator, where the calibration function was estimated using an AIPW score in 10 bins. 95% CIs were computed via bootstrapping.

Table 2: Calibration Errors of Two CATE Models Derived Using CRITEO

HTE Model	Estimate	95% CI
Causal forest	0	$(0, 8.9 \times 10^{-7})$
Random forest	4.2×10^{-6}	$(2.1 \times 10^{-6}, 6.8 \times 10^{-6})$

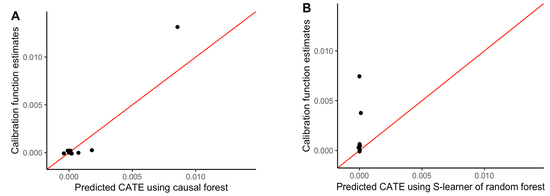

FIGURE 2: Calibration Plot between Predicted and Observed CATEs in CRITEO. The red line at 45 degrees indicates perfect calibration. The CATE estimates from causal forest are better calibrated than those from the S-learner of random forest.

Conclusions

We propose a general calibration metric, an AIPW score based robust ECETH estimator, for evaluating the calibration error of CATE estimates. Given a large amount of interest in HTE estimation and many proposals of novel statistical methods, it is crucial to compare available approaches and choose the top-performing one for deployment at clinical sites.