5. Automated Causal Inference: A Case Study

As envisioned in Figure 2, practitioners can use our validation procedure to select the best causal inference method for a given dataset. Unlike pervasive "expert-driven" modeling practices (Rubin, 2010), this *automated* and data-driven approach to model selection enables confident deployment of (black-box) machine learning-based methods, and safeguards against naïve modeling choices.

In this Section, we demonstrate the practical significance of influence function-based validation by assessing its utility in model selection. In particular, we assemble a pool of models — comprising all methods published recently in ICML, NeurIPS and ICLR — and use our validation procedure to predict the best performing model on 77 benchmark datasets from a recent causal inference competition.

5.1. Experimental Setup

■ **Influence function-based validation.** We implement a stratified P-fold version of our validation procedure as follows. First, we randomly split the training data into P mutually exclusive subsets, with $\mathcal{Z}_p$ being the set of indexes of data points in the p^{th} subset, and $\mathcal{Z}_{-p}$ its complement. In the p^{th} fold, the model being evaluated is trained on the data in $\mathcal{Z}_{-p}$, and issues a CATE estimate $\widehat{T}_{-p}$. For validation, we execute our two-step procedure as follows:

Step 1: Plug-in estimation (Section 3.1)

Using the dataset indexed by $\mathcal{Z}_{-p}$, we fit the plug-in model $\tilde{\theta}_{-p} = \{\tilde{\mu}_{-p,0}, \tilde{\mu}_{-p,1}, \tilde{\pi}_{-p}\}$ as explained in Section 3.1. We use two XGBoost regression models for $\tilde{\mu}_{-p,0}$ and $\tilde{\mu}_{-p,1}$, and then calculate $\widetilde{T}_{-p} = \tilde{\mu}_{-p,1} - \tilde{\mu}_{-p,0}$. For $\tilde{\pi}_{-p}$, we use an XGBoost classifier. Our choice of XGBoost is motivated by its minimax optimality (Linero & Yang, 2018), which is required by Theorem 1.

Step 2: Unplugged validation (Section 3.2)

Given $\tilde{\theta}_{-p}$, we estimate the model's PEHE on the held-out sample $\mathcal{Z}_p$ using the estimator in (9) with $m = 1$, i.e.,

$$\hat{\boldsymbol{\ell}}_p^{(1)} = \sum_{i \in \mathcal{Z}_p} \left[(\widehat{T}_{-p}(X_i) - \widetilde{T}_{-p}(X_i))^2 + \dot{\boldsymbol{\ell}}_{\tilde{\theta}_{-p}}^{(1)}(Z_i; \widehat{T}_{-p}) \right],$$

where $\dot{\boldsymbol{\ell}}_\theta^{(1)}(.)$ is given by Theorem 2. (Here, the first order U-statistic $\mathbb{U}_1$ in (9) reduces to a sample average.)

The final PEHE estimate is given by the average PEHE estimates over the P validation folds, i.e., $\hat{\boldsymbol{\ell}}_n^{(1)} = n^{-1} \sum_p \hat{\boldsymbol{\ell}}_p^{(1)}$. In all experiments, we set $m = 1$ since higher order influence terms did not improve the results. This may be either because the condition $m \geq d/2(\alpha_0 \wedge \alpha_1)$ (Theorem 1) is satisfied with $m = 1$, or because we use approximate higher order influence (Appendix G). We defer investigations into the utility of higher order influence terms to future work.

Method name	Reference	% Winner
BNN*	Johansson et al. (2016)	3 %
CMGP‡	Alaa et al. (2017)	12 %
TARNet*	Shalit et al. (2017)	8 %
CFR Wass.*	Shalit et al. (2017)	12 %
CFR MMD*	Shalit et al. (2017)	9 %
NSGP*	Alaa et al. (2018)	17 %
GAN-ITE$^\diamond$	Yoon et al. (2018)	7 %
SITE‡	Yao et al. (2018)	7 %
BART	Hill (2011)	15 %
Causal Forest	Wager et al. (2017)	10 %
Factual	—	53 %
IPTW	—	54 %
Plug-in	—	65 %
IF-based	—	72 %
Random	—	10 %
Supervised	—	84 %

Table 1. **Comparison of baselines over all datasets.**
Refer to Appendix H for description of each method. (*ICML, ‡NeurIPS, $^\diamond$ICLR.)

■ **Automated causal inference.** Using our validation procedure, we validate a set of candidate models for a given dataset, and then pick the model with smallest $\hat{\boldsymbol{\ell}}_n^{(1)}$. Our candidate models include all methods published in ICML, NeurIPS and ICLR conferences from 2016 to 2018. This comprises a pool of 8 models, with modeling approaches ranging from Gaussian processes to generative adversarial networks. In addition, we included two other key models developed in the statistics community (BART and causal forests). All candidate models are presented in Table 1.

■ **Data description.** We conducted extensive experiments on benchmark datasets released by the "Atlantic Causal Inference Competition" (Hill, 2016), a data analysis competition that compared models of treatment effects. The competition involved 77 semi-synthetic datasets: all datasets shared the same data on features X, but each dataset had its own simulated outcomes and assignments (W, Y). Features were extracted from a real-world observational study, whereas outcomes and assignments were simulated via data generating processes that were carefully designed to mimic real data. Each of the 77 datasets had a unique data generating process encoding varying properties (e.g., levels of treatment effect heterogeneity, dimensionality of the relevant feature space, etc.) Detailed explanation of the data generating processes was published by the organizers of the competition in (Dorie et al., 2017).

The feature data shared by all datasets was extracted from the Collaborative Perinatal Project (Niswander, 1972), a study conducted on a cohort of pregnant women to identify causes of infants' developmental disorders. The treatment was a child's birth weight ($W = 1$ if weight $< 2.5\ kg$), and outcome was the child's IQ after a given follow-up period. The study contained 4,802 data points with 55 features (5 are binary, 27 are count data, and 23 are continuous).