The Importance of Hyperparameter Tuning in Causal Effect Estimation

Damian Machlanski

d.machlanski@essex.ac.uk

Department of Computer Science and Electronic Engineering

ESRC Research Centre on Micro Social Change

University of Essex, UK

Work with:

Spyridon Samothrakis (IADS) and Paul Clarke (ISER)

Motivation

- Current situation
  - Many causal effect estimators[1]
  - Highly flexible base learners (regression and classification; ML)[2]
  - Model selection and tuning (non-trivial)[3]
- Questions
  - How **really** important is
    - The choice of causal effect estimators?
    - The choice of base learners?
    - The choice of model selection performance metrics?
  - Can common causal estimators achieve SotA performance if tuned properly?

1. (Guo et al., 2020; Yao et al., 2020)
2. (Samothrakis et al., 2022)
3. (Alaa & Schaar, 2019; Nie & Wager, 2021; Rolling & Yang, 2014; Schuler et al., 2018)

Preliminaries

- Assumptions (strong ignorability and SUTVA)
- Background covariates X, treatment t={0,1}, outcome $\mathcal{Y}^t$
- Conditional Average Treatment Effect (CATE)

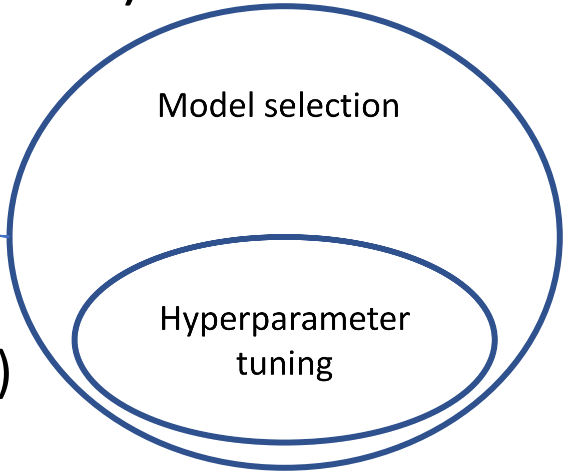
$$\tau(x) = \mathbb{E}[\mathcal{Y}^1 | X = x] - \mathbb{E}[\mathcal{Y}^0 | X = x]$$

$$PEHE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}(x_i) - \tau(x_i))^2}$$

PEHE – Precision in Estimation of Heterogeneous Effect
SUTVA – Stable Unit Treatment Value Assumption

Preliminaries

- No single learner is universally better than others (no free lunches)
- Different model - meaning
  - Meta learners (*X-Learner, Doubly Robust, etc.*)
  - Base learner(s) (*OLS, Decision Tree, etc.*)
  - Hyperparameters (*regularisation strength, max depth, etc.*)

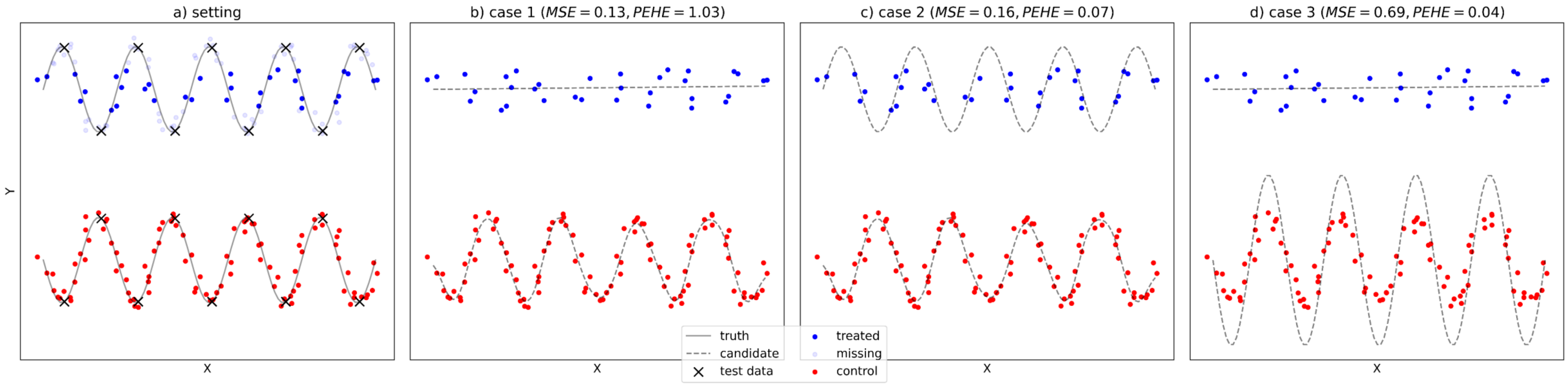


- Select a winning model w.r.t. performance metric (*MSE*, R^2, *etc.*)
- Train/validation or cross-validation
- **Problem**
  - Model selection – minimise errors on outcomes $\mathcal{Y}^t$(e.g. MSE)
  - Estimation target – minimise errors on CATEs (e.g. PEHE)
  - Better fit -> better CATEs?

Causal Model Selection

Worse MSE -> better PEHE?!

MSE	0.13	<	0.16	<	0.69
PEHE	1.03	>	0.07	>	0.04



MSE - evaluation metric on observed data (validation set/cross-validation) used for model selection purposes (accessible with real datasets; lower is better). PEHE - evaluation metric on unobserved test data (not accessible with real datasets due to missing counterfactuals; lower is better).

Experimental Setup

True CATEs

- Datasets: *IHDP[1], Jobs[2], Twins[3]*
- Model selection metrics: MSE, R^2, $\mathcal{R}_{pol}$, Plugin, Matching, R-Score[4]
- Oracle: *always picks the best model (optimal choice; inaccessible otherwise)*
- CATE estimators: *S-, T- and X-Learners, Doubly Robust, Double ML, IPSW, Causal Forest*
- Base learners: *OLS L1 and L2, Decision Tree, Random Forest, Extreme Trees, Kernel Ridge, LightGBM, CatBoost, Neural Networks*
- Hyperparameters: *search space defined for each base learner (exhaustive search)*
- Miscellaneous: *10-fold cross-validation, 10 iterations per dataset*

1. (Brooks-Gunn et al., 1992; Hill, 2011)
2. (A. Smith & E. Todd, 2005; Dehejia & Wahba, 2002; LaLonde, 1986)
3. (Almond et al., 2005; Louizos et al., 2017)
4. Based on R-Loss (Nie & Wager, 2021)

Results – CATE Estimators

Model selection = “oracle”

Search space = base learners X hyperparameters

Current SotA

method	IHDP ϵ_{ATE}	IHDP $PEHE$	Jobs ϵ_{ATT}	Jobs $\mathcal{R}_{pol}$	Twins ϵ_{ATE}	Twins $PEHE$
S-Learner	$.001 \pm .001$	1.205 ± 0.561	$.003 \pm .001$	$.158 \pm .011$	$.000 \pm .000$	$.317 \pm .002$
T-Learner	$.000 \pm .000$	0.621 ± 0.200	$.000 \pm .000$	$.128 \pm .012$	$.000 \pm .000$	$.318 \pm .002$
IPSW	$.001 \pm .000$	1.204 ± 0.560	$.024 \pm .019$	$.158 \pm .013$	$.000 \pm .000$	$.317 \pm .002$
Doubly Robust	$.002 \pm .001$	1.275 ± 0.581	$.001 \pm .001$	$.149 \pm .010$	$.001 \pm .000$	$.318 \pm .002$
Double ML	$.007 \pm .003$	1.679 ± 0.830	$.016 \pm .005$	$.193 \pm .012$	$.001 \pm .000$	$.317 \pm .002$
X-Learner	$.009 \pm .007$	1.067 ± 0.409	$.003 \pm .002$	$.153 \pm .013$	$.013 \pm .001$	$.318 \pm .002$
Causal Forest	$.198 \pm .131$	2.290 ± 1.186	$.053 \pm .022$	$.204 \pm .017$	$.063 \pm .000$	$.323 \pm .002$
S-Learner NN	$.104 \pm .038$	0.925 ± 0.224	$.023 \pm .019$	$.162 \pm .010$	$.001 \pm .000$	$.317 \pm .002$
T-Learner NN	$.000 \pm .000$	0.641 ± 0.190	$.010 \pm .010$	$.132 \pm .009$	$.000 \pm .000$	$.328 \pm .004$
TARNet	$.280 \pm .010$	0.950 ± 0.020	$.110 \pm .040$	$.210 \pm .010$	-	$.315 \pm .003$
CFR-WASS	$.270 \pm .010$	0.760 ± 0.020	$.090 \pm .030$	$.210 \pm .010$	-	$.313 \pm .008$
SITE	-	0.656 ± 0.108	-	$.219 \pm .009$	-	-
GANITE	-	2.400 ± 0.400	-	$.140 \pm .010$	-	$.297 \pm .016$
CEVAE	$.460 \pm .020$	2.600 ± 0.100	$.030 \pm .010$	$.260 \pm .000$	-	-

Table 1: Model selection with access to oracle across base learners, but within causal estimators.

Mean +- standard error. Lower is better. NN – Neural Network.

TARNet and CFR-WASS (Shalit et al., 2017), SITE (Yao et al., 2018), GANITE (Yoon et al., 2018), CEVAE (Louizos et al., 2017). Numbers as in papers.

Results – Base Learners

Model selection = "oracle"

Search space =
CATE estimators
X
hyperparameters

method	IHDP ϵ_{ATE}	IHDP $PEHE$	Jobs ϵ_{ATT}	Jobs $\mathcal{R}_{pol}$	Twins ϵ_{ATE}	Twins $PEHE$
OLS L1	$.046 \pm .016$	$1.643 \pm .884$	$.032 \pm .022$	$.197 \pm .013$	$.022 \pm .000$	$.317 \pm .002$
OLS L2	$.161 \pm .112$	$1.603 \pm .841$	$.061 \pm .023$	$.199 \pm .014$	$.030 \pm .001$	$.318 \pm .002$
Decision Tree	$.000 \pm .000$	$1.890 \pm .932$	$.002 \pm .001$	$.142 \pm .010$	$.000 \pm .000$	$.317 \pm .002$
Random Forest	$.020 \pm .011$	$1.529 \pm .811$	$.011 \pm .006$	$.155 \pm .015$	$.000 \pm .000$	$.317 \pm .002$
Extra Trees	$.003 \pm .002$	$1.582 \pm .915$	$.014 \pm .008$	$.148 \pm .013$	$.000 \pm .000$	$.318 \pm .002$
Kernel Ridge	$.001 \pm .001$	$0.653 \pm .195$	$.000 \pm .000$	$.141 \pm .009$	$.000 \pm .000$	$.317 \pm .002$
CatBoost	$.004 \pm .002$	$0.893 \pm .386$	$.025 \pm .011$	$.156 \pm .011$	$.023 \pm .001$	$.318 \pm .002$
LightGBM	$.016 \pm .008$	$1.326 \pm .562$	$.009 \pm .003$	$.174 \pm .011$	$.011 \pm .000$	$.317 \pm .002$
Neural Net	$.000 \pm .000$	$0.641 \pm .190$	$.010 \pm .010$	$.131 \pm .009$	$.000 \pm .000$	$.317 \pm .002$
TARNet	$.280 \pm .010$	$0.950 \pm .020$	$.110 \pm .040$	$.210 \pm .010$	-	$.315 \pm .003$
CFR-WASS	$.270 \pm .010$	$0.760 \pm .020$	$.090 \pm .030$	$.210 \pm .010$	-	$.313 \pm .008$
SITE	-	$0.656 \pm .108$	-	$.219 \pm .009$	-	-
GANITE	-	$2.400 \pm .400$	-	$.140 \pm .010$	-	$.297 \pm .016$
CEVAE	$.460 \pm .020$	$2.600 \pm .100$	$.030 \pm .010$	$.260 \pm .000$	-	-

Table 2: Model selection with access to oracle across causal estimators, but within base learners.

Mean +- standard error. Lower is better.
TARNet and CFR-WASS (Shalit et al., 2017), SITE (Yao et al., 2018), GANITE (Yoon et al., 2018), CEVAE (Louizos et al., 2017). Numbers as in papers.

Results – Model Selection Metrics

Search space =
CATE estimators
X
base learners
X
hyperparameters

method	IHDP		Jobs		Twins	
	ϵ_{ATE}	$PEHE$	ϵ_{ATT}	$\mathcal{R}_{pol}$	ϵ_{ATE}	$PEHE$
MSE*	$.188 \pm .097$	$0.786 \pm .189$	$.077 \pm .025$	$.245 \pm .013$	$.039 \pm .000$	$.319 \pm .002$
R^2*	$.352 \pm .146$	$0.922 \pm .237$	$.072 \pm .022$	$.257 \pm .019$	$.039 \pm .000$	$.319 \pm .002$
$\mathcal{R}_{pol}$	–	–	$.300 \pm .090$	$.220 \pm .016$	–	–
Plugin	$.209 \pm .048$	$1.341 \pm .518$	$.085 \pm .023$	$.244 \pm .015$	$.047 \pm .001$	$.320 \pm .002$
Matching	$.164 \pm .067$	$0.718 \pm .239$	$.080 \pm .028$	$.235 \pm .014$	$.077 \pm .000$	$.326 \pm .002$
R-Score	$.535 \pm .207$	$1.389 \pm .498$	$.075 \pm .019$	$.224 \pm .017$	$.026 \pm .004$	$.320 \pm .003$
Oracle	$.000 \pm .000$	$0.585 \pm .198$	$.000 \pm .000$	$.121 \pm .011$	$.000 \pm .000$	$.317 \pm .002$

Table 3: Effectiveness of model selection methods. *Includes only S-Learner, T-Learner and IPSW.

Mean +- standard error. Lower is better.

Summary

- **If we make optimal choices w.r.t. model selection ("oracle") then:**
  - The choice of CATE estimators and base learners might be less important (use what you like)
  - Common CATE estimators can be very competitive (even reach SotA)
- **Can we make such optimal choices today?**
  - Not quite (table 3)
  - More research on causal model selection is needed
  - Include priors, carefully consider model selection
- **Future work**
  - Less focus on new CATE estimators (they are great already; tables 1 and 2)
  - More emphasis on causal model selection (appears more important)
- **Code**
  - https://github.com/misoc-mml/hyperparam-sensitivity

References

A. Smith, J., & E. Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125(1–2), 305–353.

Alaa, A., & Schaar, M. V. D. (2019). Validating Causal Inference Models via Influence Functions. Proceedings of the 36th International Conference on Machine Learning, 191–201. https://proceedings.mlr.press/v97/alaa19a.html

Almond, D., Chay, K. Y., & Lee, D. S. (2005). The Costs of Low Birth Weight. The Quarterly Journal of Economics, 120(3), 1031–1083. https://doi.org/10.1093/qje/120.3.1031

Brooks-Gunn, J., Liaw, F. R., & Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. The Journal of Pediatrics, 120(3), 350–359. https://doi.org/10.1016/s0022-3476(05)80896-0

Dehejia, R. H., & Wahba, S. (2002). Propensity Score-Matching Methods For Nonexperimental Causal Studies. The Review of Economics and Statistics, 84(1), 151–161.

Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2020). A Survey of Learning Causality with Data: Problems and Methods. ACM Computing Surveys, 53(4), 75:1-75:37. https://doi.org/10.1145/3397269

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. Journal of Computational and Graphical Statistics, 20(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. The American Economic Review, 76(4), 604–620. JSTOR.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal Effect Inference with Deep Latent-Variable Models. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/94b5bde6de888ddf9cde6748ad2523d1-Abstract.html

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. Biometrika, 108(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Rolling, C. A., & Yang, Y. (2014). Model selection for estimating treatment effects. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(4), 749–769. https://doi.org/10.1111/rssb.12043

Samothrakis, S., Matran-Fernandez, A., Abdullahi, U., Fairbank, M., & Fasli, M. (2022). Grokking-like effects in counterfactual inference. 2022 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN55064.2022.9891910

Schuler, A., Baiocchi, M., Tibshirani, R., & Shah, N. (2018). A comparison of methods for model selection when estimating individual treatment effects. ArXiv:1804.05146 [Cs, Stat]. http://arxiv.org/abs/1804.05146

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. International Conference on Machine Learning, 3076–3085. http://proceedings.mlr.press/v70/shalit17a.html

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., & Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2638–2648.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2020). A Survey on Causal Inference. ArXiv:2002.02770 [Cs, Stat]. http://arxiv.org/abs/2002.02770

Yoon, J., Jordon, J., & Schaar, M. van der. (2018, February 15). GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. International Conference on Learning Representations. https://openreview.net/forum?id=ByKWUeWA-

Thank you!