Causal model selection : is my metric adapted ? which metric in the absence of perfect overlap or oracles?

Matthieu Doutreligne [1 2], Gaël Varoquaux [2]

[1]Haute Autorité de Santé, Saint-Denis
[2]Inria Paris Saclay

Motivation: Flexible predictive models are useful for causal inference

In extensive simulations, flexible models of the outcome have proven efficient for ATE and CATE estimation [Dorie et al., 2018].

Question: How to select between different models for Conditional Average Treatment Effect (CATE) estimation ?

Neyman Rubin potential outcomes [Imbens and Rubin, 2015]

Data: $(Y(1), Y(0), X, A) \sim p(y(1), y(0), x, a)$

Target quantities (estimands):

- ATE (population): $\tau = \mathbb{E}_{Y(1),Y(0)}[Y(1) - Y(0)]$
- Conditional Average Treatment Effect (CATE):

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

Identifiable under strong ignorability [Rubin, 2005]

Problem Statement: select the best causal model [Schuler et al., 2018]

Given candidate models of the outcome $f \in \mathcal{F} : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$
Each model induces a CATE with $\hat{\tau}_f(x) = f(x, 1) - f(x, 0)$

Select the best with mean squared error on the true CATE, called τ-risk:

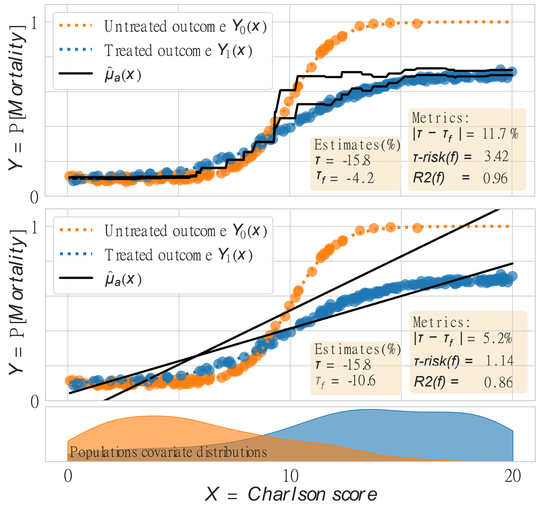
$$f^* = argmin_f \mathbb{E}_{Y,X,A}[(\tau(x) - \hat{\tau}_f(x))^2]$$

Mean Squared Error on the outcome is not what you need !

Figure 1. Toy example: On the first row, a random-forest model with high regression performance (high R^2) yielding poor ATE estimation (large error between true effect τ and predicted $\hat{\tau}_{\hat{f}}$),

On the second row, a linear model with smaller regression performance leading to better ATE and CATE estimations.

Take away: Mean squared error does not capture the inhomogeneity between treated and controls.



Better causal metrics ?

Idea: Select the model with the smallest τ-risk.
Difficulty: τ-risk is an oracle quantity, we need feasible (finite samples and observable) metrics.
It is possible to do better than the usual mean squared error μ-risk(f) [Schuler et al., 2018, Alaa and Schaar, 2019].

Metric	Equation
τ-risk(f)	$\mathbb{E}_{x\sim p(x)}[(\tau(x) - \hat{\tau}_f(x))^2]$
μ-risk(f)	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[(y - f(x; a))^2\right]$
μ-risk$_{IPW}(w, f)$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left(\frac{a}{e(x)} + \frac{1-a}{1-e(x)}\right)(y - f(x; a))^2\right]$
R-risk $= \tau$-risk$_R$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left((y_i - m(x_i)) - (a_i - e(x_i))\tau_f(x_i)\right)^2\right]$

R-risk [Nie and Wager, 2017] is more complex since it requires a model of the outcome, $m(x) = \mathbb{E}[Y|X = x]$.
Question: Why is it a pertinent risk and how well does it perform ?

R-risk appears as a reweighted oracle.

R-risk as reweighted τ-risk:

$$R\text{-risk}^*(f) = \int_x e(x)(1 - e(x))\left(\tau(x) - \tau_f(x)\right)^2 p(x)dx + \tilde{\sigma}^2_B(1) + \tilde{\sigma}^2_B(0)$$

Intuition: R-risk targets the oracle for good overlap settings.

Empirical Study on simulations and three semi-synthetic datasets.

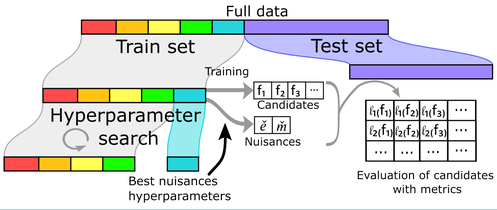
Simulation: 1000 instances using Gaussian-distributed covariates and random functional basis to control the complexity of the response surfaces and the overlap.

Candidates models of the outcome $\mathcal{F}$:
120 candidate ridge regressions with 6 choices of regularization parameter $\alpha \in [10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2]$, coupled with a T-Learner or a Sft-Learner (a shared random basis and 2 predictors), 10 different random seeds.

Semi-synthetic datasets. Real world covariates, simulated outcomes and treatments : ACIC 2016 (770 instances) [Dorie et al., 2018], ACIC 2018 (432 instances) [Shimoni et al., 2018] , Twins (1000 instances) [Louizos et al., 2017].

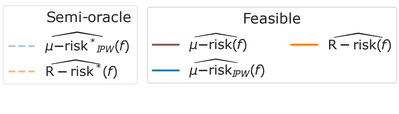
Candidates models of the outcome $\mathcal{F}$:
18 candidate gradient boosting with S-learner, learning rate in $\{0.01, 0.1, 1\}$, and maximum number of leaf nodes in $\{25, 27, 30, 32, 35, 40\}$.

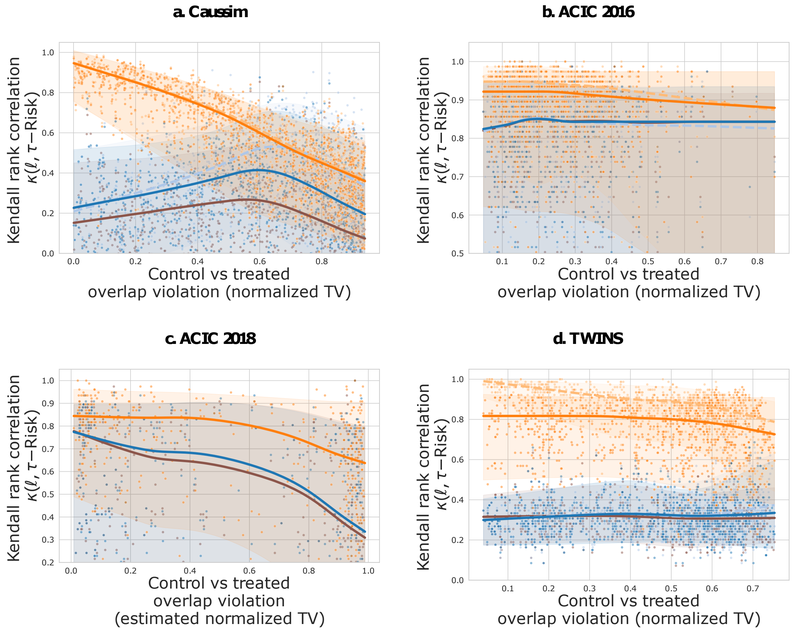
Figure 2. Selection procedure.



Empirical results confirm that R-risk is more performant in all settings.

Figure 3. Rank correlation of candidate outcome models between the oracle τ-risk and the other metrics





Take home messages

R-risk dominates in all settings: Including an estimate of the outcome model into the risk seems always beneficial.

High **lack of overlap hurts** model selection performance: Overlap should be measured and controlled.

There is not a large performance gap between metrics using **oracle versus learned nuisances** $(e(x), m(x))$: plugin-nuisances fitted with complex models such as boosting trees gives good causal metrics.

Variability between dataset instances is high: Empirical results should be reported on numerous simulations.

References

Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. *International Conference on Machine Learning*, pages 191–201, 2019.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv:1707.02641 [stat]*, July 2018.

Guido W. Imbens and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv:1705.08821 [cs, stat]*, 2017.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv:1712.04912 [econ, math, stat]*, 2017.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv:1804.05146 [cs, stat]*, 2018.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv:1802.05046 [cs, stat]*, 2018.