Measuring Variable Importance in Individual Treatment Effect Estimation with High Dimensional Data

https://arxiv.org/abs/2408.13002

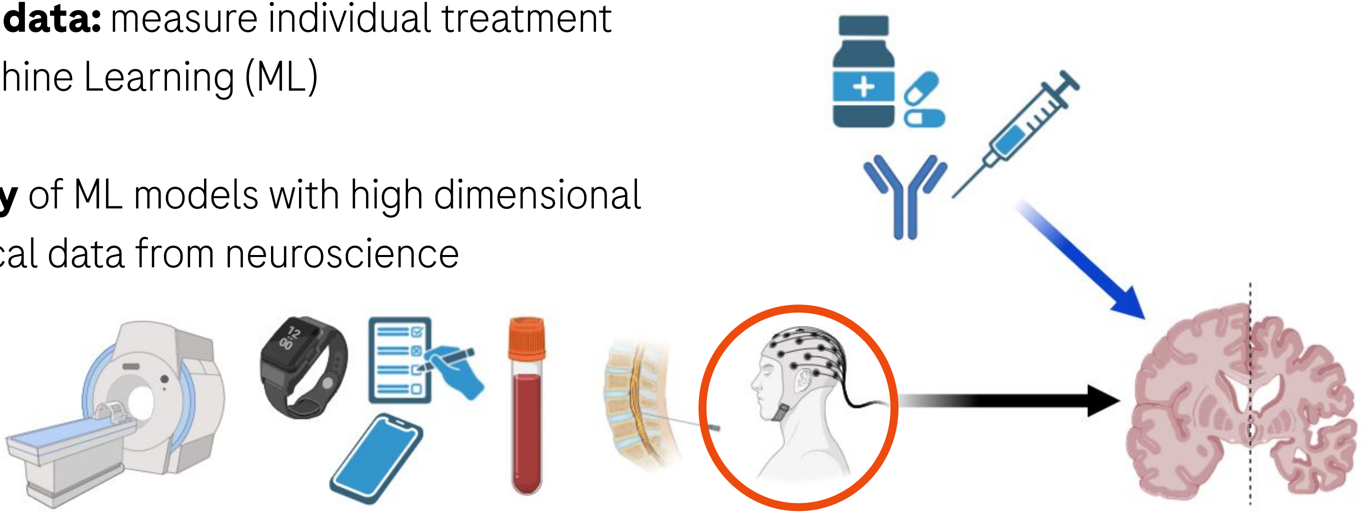
Joseph Paillard, 1st year PhD student
Supervisors: Bertrand Thirion (Inria, Saclay)
Denis Engemann (Roche, Basel)

Workshop on statistical inference for complex data - December 2024

Understanding treatment effects using neuroimaging data

Challenges

- **Interventional data:** measure individual treatment effects with Machine Learning (ML)
- **Interpretability** of ML models with high dimensional multimodal clinical data from neuroscience

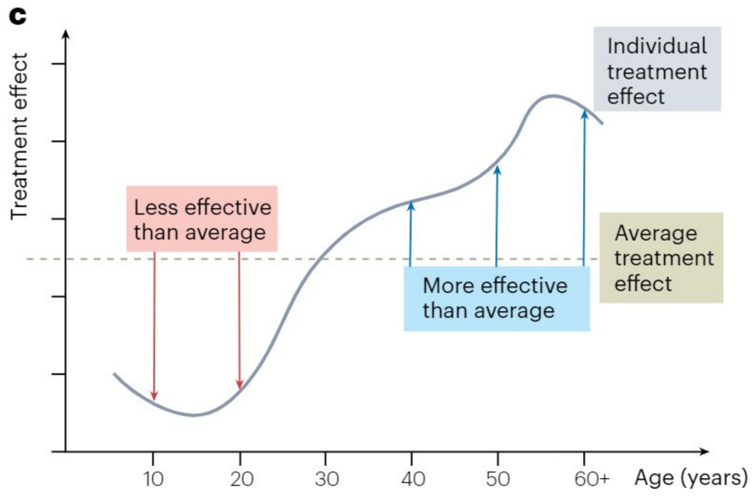


Conditional Average Treatment effect (CATE)

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

Y: outcome, X: Covariate, binary treatment

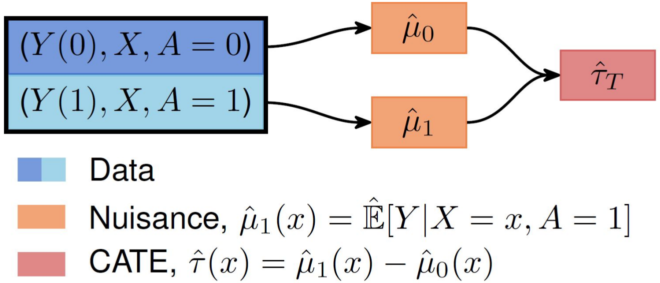
- Studying heterogeneity in individual treatment effect
- Involves reasoning with counterfactual quantities: $Y(1)$ and $Y(0)$ cannot be observed simultaneously



1.Feuerriegel, S. *et al.* Causal machine learning for predicting treatment outcomes. *Nature Medicine*. 2024

Meta-learners for estimating the CATE

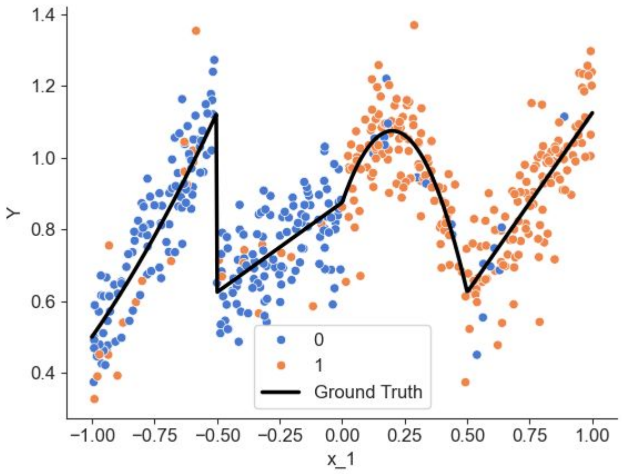
T-learner



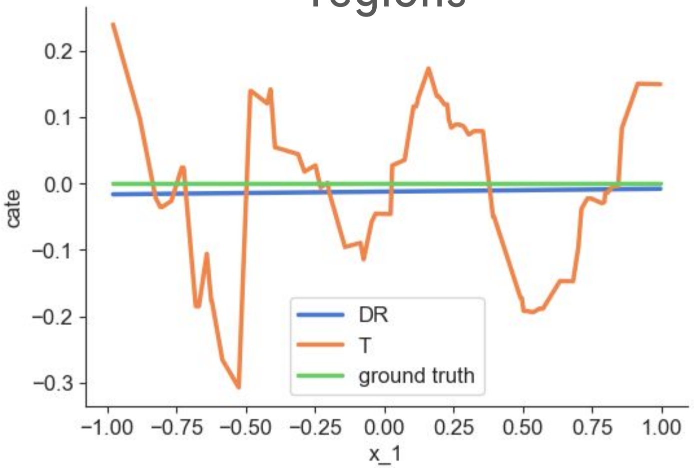
2.Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics. 2023

Failure case of the “simple” T-learner

Weak overlap scenario



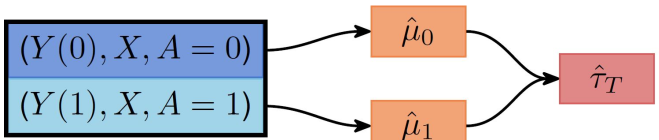
Poor estimation in low density regions



2.Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics. 2023

Meta-learners for estimating the CATE

T-learner

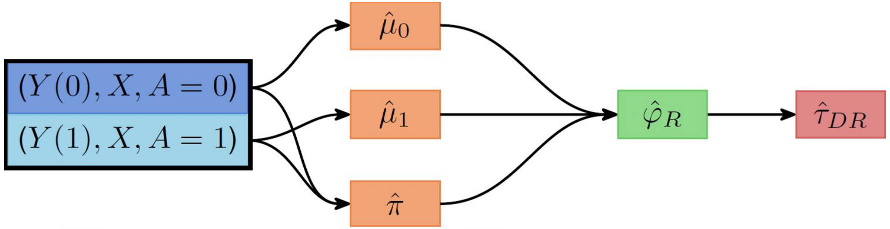


Data

Nuisance, $\hat{\mu}_1(x) = \hat{\mathbb{E}}[Y|X = x, A = 1]$

CATE, $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

DR-learner



Nuisance parameters

Pseudo outcomes

$\hat{\pi}(x) = \mathbb{P}(A = 1|X = x)$

$$\varphi(z) = \frac{(y - \mu_a(x))(a - \pi(x))}{\pi(x)(1 - \pi(x))} + \mu_1(x) - \mu_0(x)$$

$$z = (y, z, x)$$

2.Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics. 2023

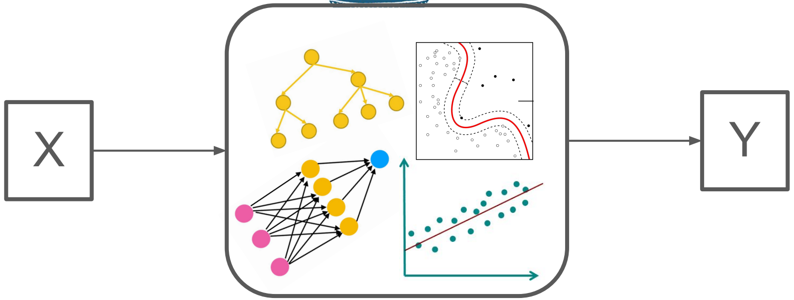
Interpreting model prediction: variable importance

Desired properties:

- Be model agnostic
- Provide global interpretability (opposed to local: around an observation)

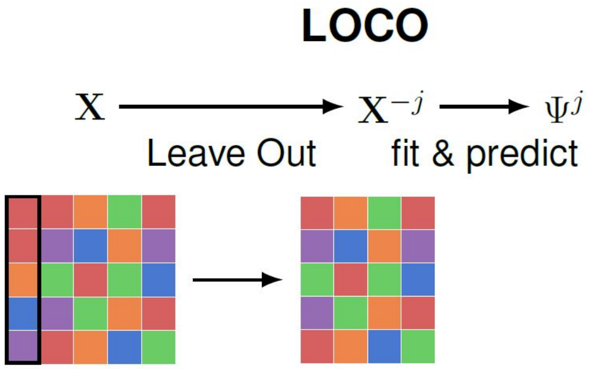
Interpretability
Explainability

- Shapley value
- Leave One Covariate Out
- Feature permutation
- Knockoffs
- ...



3.Chamma, A. Statistical interpretation of high-dimensional complex prediction models for biomedical data. (HAL open science, 2024)

Leave One Covariate Out (LOCO)



Re-fit the model on a subset of covariates

$$\mu_{-j}(x) = \mathbb{E}[Y|X^{-j} = x^{-j}]$$

Measure the performance drop

$$\Psi^j = \mathbb{E}[(Y - \mu_{-j}(X))^2] - \mathbb{E}[(Y - \mu(X))^2]$$

Estimates the total Sobol index

4.Verdinelli, I. & Wasserman, L. Feature Importance: A Closer Look at Shapley Values and LOCO. arXiv (2023)

LOCO application to heterogeneous causal effects

Open questions:

- How would other methods compare ?

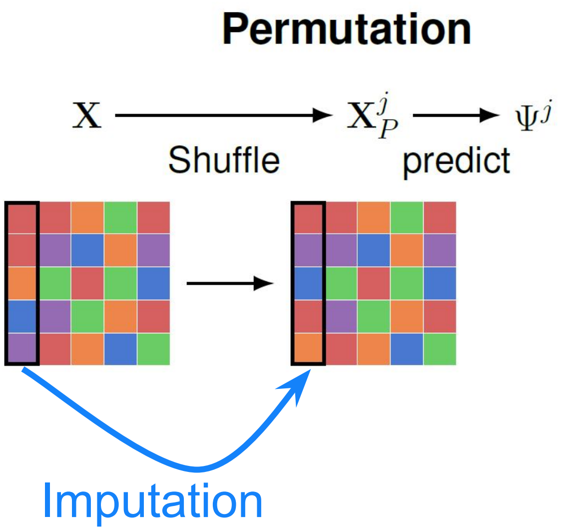
- How would that scale with dimension ?
  In O.Hines et al, 2022: d=6 with linear treatment effect

- Which risk to use ?
  The precision of estimating heterogeneous effects (PEHE) is not feasible!

$$\Psi^{j}_{LOCO} = R\left(\tau, X, A, Y\right) - R\left(\tau^{-j}, X^{-j}, A, Y\right)$$

5.Hines, O., Diaz-Ordaz, K. & Vansteelandt, S. Variable importance measures for heterogeneous causal effects. arXiv (2022)
6.Doutreligne, M. & Varoquaux, G. How to select predictive models for causal inference? 2023

Conditional Permutation Importance (CPI)



Estimate a (group) covariate from the others

$$\hat{\nu_j} = \widehat{\mathbb{E}}[X^j | X^{-j}]$$

Compute and shuffle the remainder

$$r_j = X^j - \hat{\nu_j}$$

Accounting for the correlation structure allows to control the False Discovery Rate

Estimates the total Sobol index

7.Chamma, A., Engemann, D. A. & Thirion, B. Statistically Valid Variable Importance Assessment through Conditional Permutations, NeurIPS 2023

Contribution

- Study the finite sample variance of LOCO and CPI
- Apply to CATE estimation in high-dimensional, non-linear settings

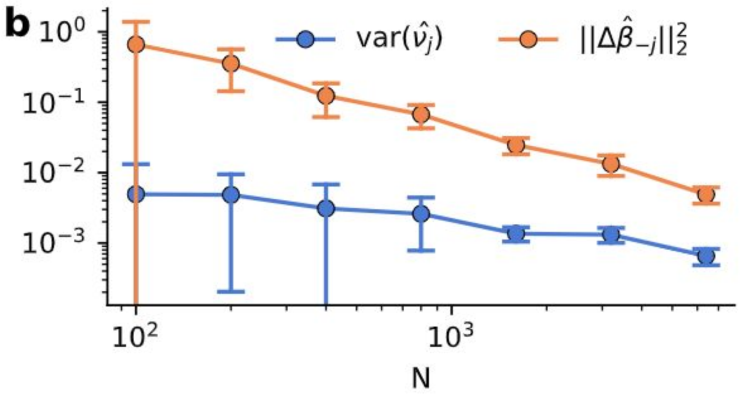
The impact of finite sample estimation error on variance

Linear CATE scenario: $\tau(x) = \sum_j \beta_j x_j + \varepsilon$

CPI

$$\text{var}\left(\widehat{\Psi^j}_{CPI}\right) = 2\hat{\beta}_j^4\left(1 + \boxed{\text{var}(\hat{\nu}_j)}\right)^2$$

$$\hat{\nu_j} = \widehat{\mathbb{E}}[X^j | X^{-j}]$$



LOCO

$$\text{var}\left(\widehat{\Psi^j}_{LOCO}\right) = 2\left(\hat{\beta}_j^2 + \boxed{||\Delta\hat{\beta}_{-j}||_2^2}\right)^2$$

$\Delta\hat{\beta}_{-j}$:The difference between the coefficients estimated on the full set of covariates and the coefficient of the model fitted on the subset of covariates

The impact of finite sample estimation error on variance

Linear CATE scenario: $\tau(x) = \sum_j \beta_j x_j + \varepsilon$

CPI
PermuCATE

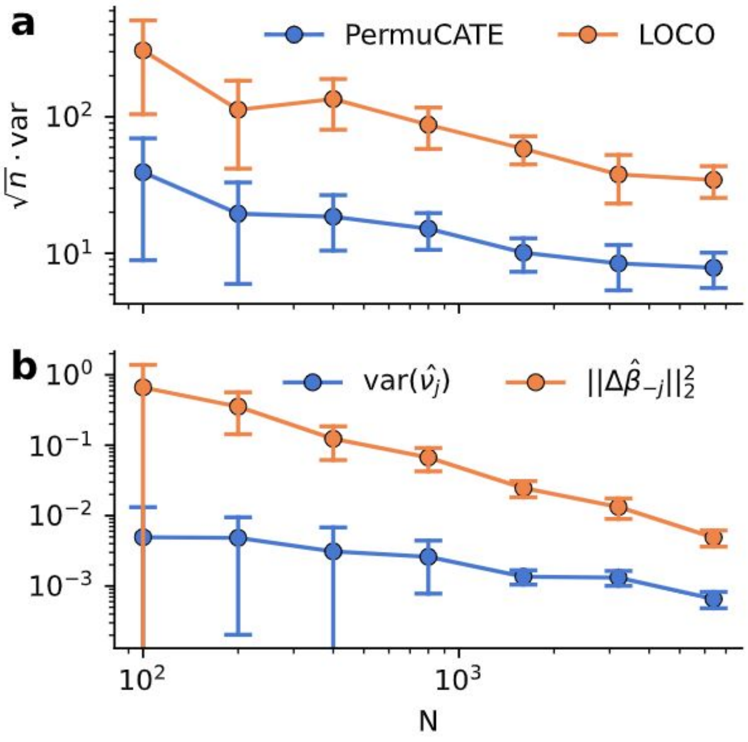
$$\boxed{\mathrm{var}\left(\widehat{\Psi^j}_{CPI}\right) = 2\hat{\beta}_j^4\left(1 + \mathrm{var}(\hat{\nu}_j)\right)^2}$$

$$\hat{\nu_j} = \widehat{\mathbb{E}}[X^j | X^{-j}]$$

LOCO

$$\boxed{\mathrm{var}\left(\widehat{\Psi^j}_{LOCO}\right) = 2\left(\hat{\beta}_j^2 + ||\Delta\hat{\beta}_{-j}||_2^2\right)^2}$$

$\Delta\hat{\beta}_{-j}$:The difference between the coefficients estimated on the full set of covariates and the coefficient of the model fitted on the subset of covariates



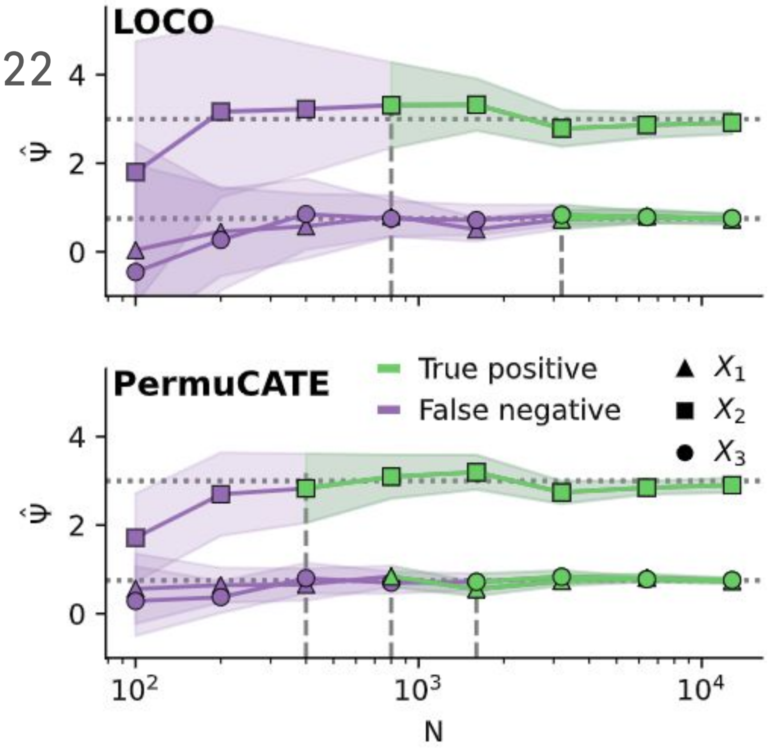
Larger variance leads to a decrease in statistical power

- Using the simulation from O.Hines et al, 2022

$$\tau(X) = X_1 + 2X_2 + X_3$$

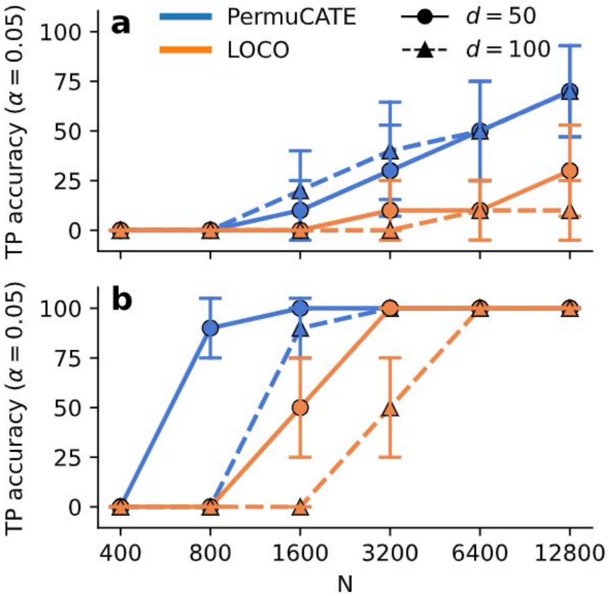
- Important variables are identified as such with smaller sample size, N

- This trend might increase with stochastic optimization

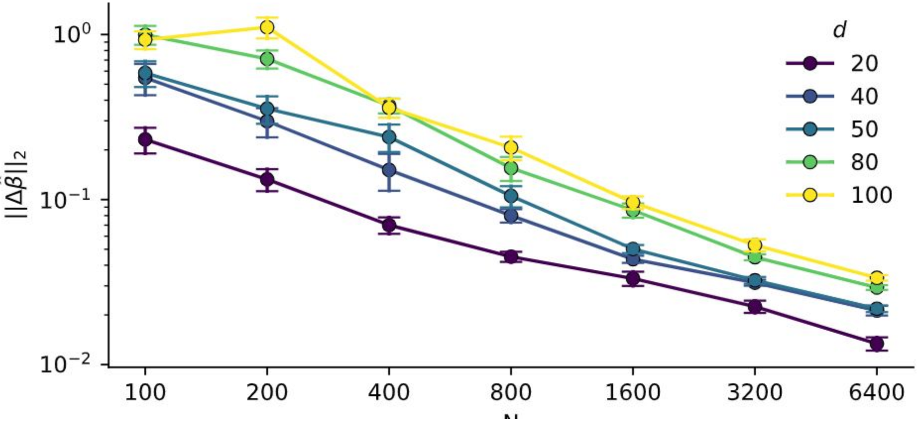


High dimension and non-linear scenario

PermuCATE: conditional permutation importance for CATE estimation



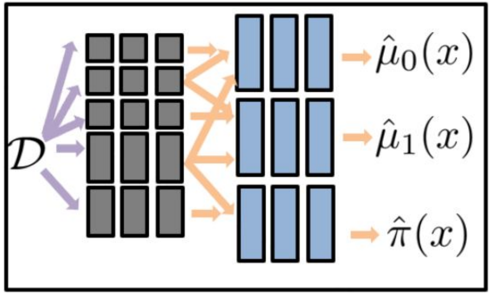
ThePermuCATEmethod identified more im-portant variables in high-dimensional, linear, and com-plex scenarios



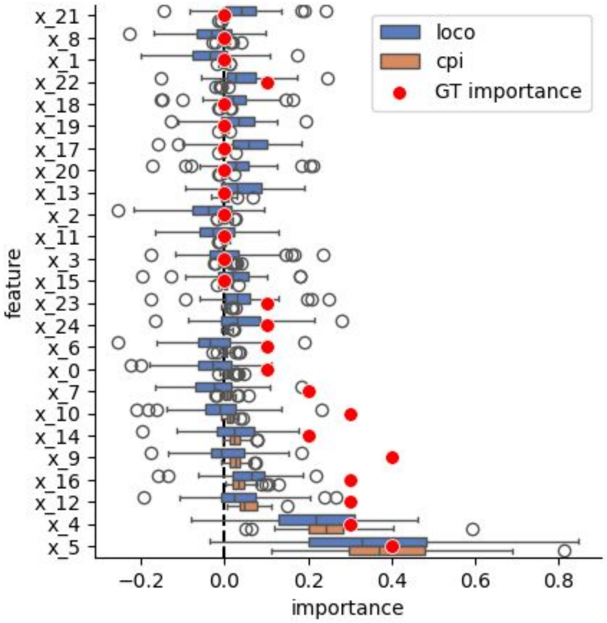
The finite sample fitting error term grows with dimensionality of the data

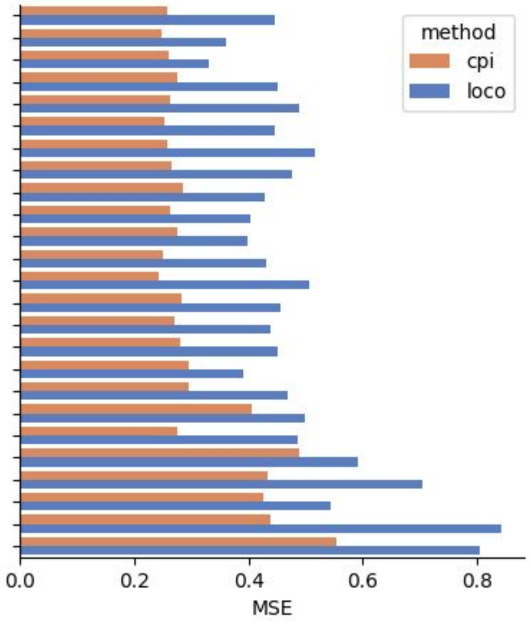
IHDP example using DL models

Infant Health and Development Program (IHDP): estimating treatment effect using CATENets



SNet





7. Curth, A. & Schaar, M. van der. Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms, AISTATS 2021

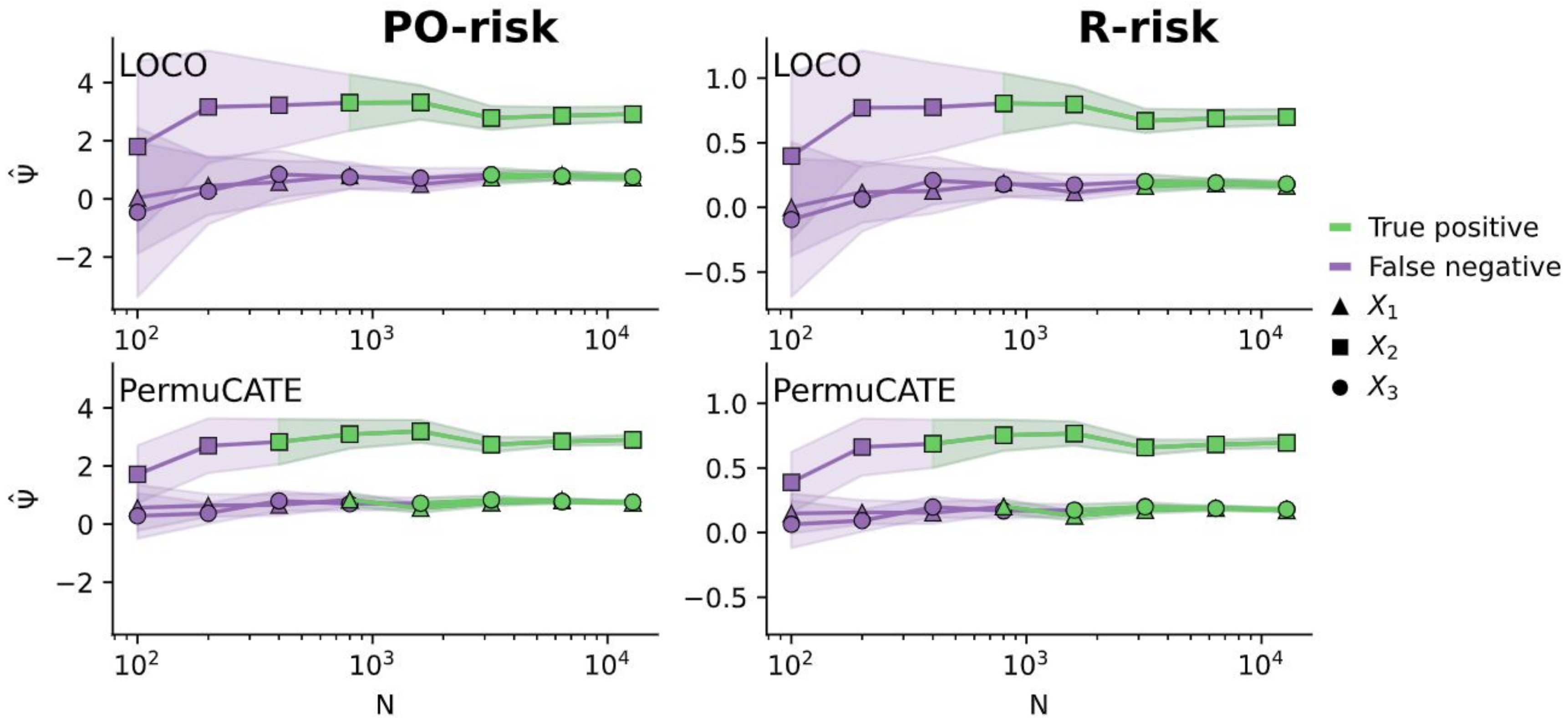
Conclusion and open challenges

- CPI benefits from computational benefits
- CPI empirical error terms are lesser than LOCO
- This leads to reduced variance and increased statistical power
- Simulation of causal effects in high dimensional & non-linear settings offers a new benchmark
- Feasible causal risks are well suited for variable importance inference

Open questions

- Real world data application
- Importance vanishing problem under very high correlation
- General theory supporting the comparison CPI / LOCO

Questions



PO-risk
R-risk
LOCO
PermuCATE
$\hat{\Psi}$
N
True positive
False negative
X_1
X_2
X_3

Potential outcome

$$\varphi(z) = \frac{(y - \mu_a(x))(a - \pi(x))}{\pi(x)(1 - \pi(x))} + \mu_1(x) - \mu_0(x)$$

R-risk

$$((Y - m(X)) - A - \pi(X))\tau(X))^2$$