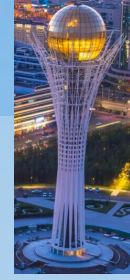
unicef | for every child



Implementing Causal Machine Learning in

Tuesday: Causal Machine Learning & Optimal Policy

August / September / October 2024

Michael Lechner
Professor of Econometrics | University of St. Gallen | Switzerland

unicef | for every child

The workshop series

Astana Workshop September 30 to October 4, 2024 (10-13, 14:00-17:30)

- Monday
  - Morning: Identification with experiments & selection on observables
  - Afternoon: Discussion of potential programmes to be evaluated
- **Today: Causal Machine Learning (theory) (ends at 16:00)**
- Wednesday
  - Empirical examples: Active labour market programmes in Flanders
  - The mcf package – how to use it & how to interpret the results
- Thursday: Empirical study in groups with the data introduced in online workshop 4
- Friday: Discussion of programmes to be evaluated continued (core team only)

unicef | for every child

Today

Causal machine learning estimators

- **M**achine **L**earning vs **C**ausal **M**achine **L**earning
- **D**ouble **D**ebiased **M**achine **L**earning (Chernozhukov et al., 2018)
- Causal Trees & Forests
  - The seminal **C**ausal **F**orest of Wager & Athey (2018)
  - **G**eneralized **R**andom **F**orest (Athey, Tibshirani, Wager, 2019)
  - **M**odified **C**ausal **F**orest (Lechner, 2018)
- Pro's & con's of these estimators (Lechner & Mareckova, 2024)
  - Asymptotics
  - Finite sample properties

Optimal Policy

unicef | for every child

1 | Introduction

2 | Machine Learning vs Causal Machine Learning

3 | Double Debiased Machine Learning

4 | Causal Trees & Forests

5 | A comparison of Comprehensive Causal Machine Learners

6 | Optimal Policy & Algorithmic Decision Making

7 | Conclusions & outlook

Already discussed in online workshop 3

Machine learning meets econometrics

ML & econometrics have different goals

Goal of machine learning

y = f(x) + unobservables

- Get best guess of *y* once you know *x* → *learn* (=estimate) *f(x)*
- Carefully ensure that estimated *f(x)* does not depend on unobservables (i.e. avoid in-sample overfitting)

Goal of econometrics

- Learn *parameters* capturing how a change in *x* changes *y*
  - Get respective uncertainty measures for these causal parameters
  - Examples: regression coefficients, average treatment effects, other 'structural' parameters

Already discussed in online workshop 3

How to compare estimators? *SL/ML*

Goal: Reliable prediction (& classification)

- Limited interest in inference

How to: Use many reference data sets

- Split data in estimation data (e.g., used for estimation)
- Compare predictive performance relative to true values in test data
  - Test data has not been used at all for estimation
- Many references to such data sets can be found at
  - https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Already discussed in online workshop 3

How to compare estimators? *Econometrics (causal analysis)*

Goal: Estimation of a parameter

Comparison of prediction with true parameter is impossible in real data

- Therefore, statistical guarantees are important (valid whatever the data & true effect)
- Therefore, evidence of finite sample performance important
  - Usually done with simulation studies in which true effect is known (i.e. specified by researcher)
  - Reference data sets exist, but they also contain simulated components (as true effect is unknown)
- Therefore, in a particular application, inference is most important to understand relevance of estimated effect

unicef | for every child

Already discussed in online workshop 3

This table oversimplifies (a bit)

Classical Econometrics & Machine Learning

Issue	Classical econometrics	Supervised statistical learning
Target of interest (θ)	Structural & causal parameters (*low dimensional*)	Prediction (*Ey|x)* or classification of *y*
Sample analogue of θ	--	*y*
Judging quality of estimation	*Indirect* (fit, …), in-sample	*Direct* ($\widehat{y}$ vs *y),* out-of-sample
Inference & theoretical properties	Very important	Less important (irrelevant?)
Sample size (*N*)	Large *N* is nice to have	Large *N* may be required
# of variables (k)	Much smaller than *N*	Smaller or larger than *N*
Preferred model complexity	Simple, likely to be parametric (linearity popular)	Complicated (overparametrized; nonlinear)
Names of methods	Boring	Cool

unicef | for every child

Already discussed in online workshop 3

Notation | Binary treatment

Binary treatment as simplification

D Treatment

Y^0 Potential outcome for *D=0*

Y^1 Potential outcome for *D=1*

$Y = D\,Y^1 + (1-D)\,Y^0$

Propensity score: $p(x) = P(D=1 \mid X=x)$

Outcome regression: $\mu(d,x) = E(Y \mid D=d, X=x)$

X All confounders & heterogeneity variables

Z Specific heterogeneity variables (low dimensional; included in *X*)

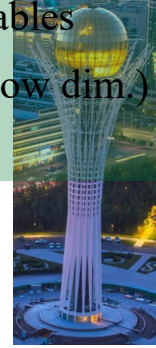
Sample contains i.i.d. realisations of *D*, *Y*, *X*

unicef | for every child

D: Treatment (0 or 1) X: Confounder & heterogeneity variables

Y^1: Outcome when $D=1$ Z: Specific heterogeneity variables (low dim.)

Y^0: Outcome when $D=0$ Observable: $D, Y, X;\ Z \subset X$



Estimands of mean causal effects at different aggregation levels | 1

Individualized (**C**onditional) **A**verage **T**reatment **E**ffect (*IATE(x)*, *CATE(x)*)

$$IATE(x) = E(Y^1 - Y^0 \mid X = x) = \mu(1,x) - \mu(0,x)$$

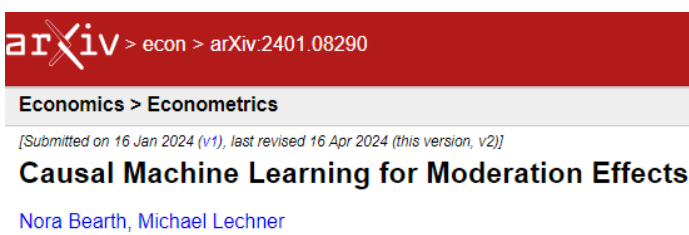
$$\mu(d,x) = E(Y \mid D = d, X = x)$$

Group **A**verage **T**reatment **E**ffect (*GATE(z),CATE(z)*)

$$GATE(z) = E_{X|Z=z} IATE(X)$$

Balanced Group **A**verage **T**reatment **E**ffect (B*GATE(z)*)

$$BGATE(z) = E_{\tilde{X}} E_{X|Z=z,\tilde{X}=\tilde{x}} IATE(x)$$



Average **T**reatment **E**ffect (*ATE*)

$$ATE = E_X IATE(X)$$

GATEs & ATEs on the treated / non-treated

unicef | for every child

The estimation problem | 1

$$IATE(x) = E(Y \mid X = x, D = 1) - E(Y \mid X = x, D = 0)$$
$$GATE(z) = E_{Z=z} IATE(x)$$
$$ATE = E_X IATE(x)$$

Naïve ML estimator

- ML estimation of *E(Y|X=x, D=d)* in treated (*d=1*) & non-treated subsample (*d=0*)
  - *IATE(x)*: Predictions of *Y* for treated - predictions of *y* for non-treated
  - *GATE(h)* : Average *IATE(x_i) with h_i same/similar to h*
  - *ATE*: Average *IATE(x_i)*

Naïve estimator may be a bad idea

- Predictions of ML estimators are usually biased
  - MSE optimal-prediction of *E(Y|X=x, D=d)* but inference may not work (too much bias)
- Estimating a difference well is different from estimating its components well
  - Only difference of estimation errors matters

unicef | for every child

The Econometrics Journal
Econometrics Journal (2021), volume **24**, pp. 134–161.
doi: 10.1093/ectj/utaa014

Machine learning estimation of heterogeneous causal effects: empirical Monte Carlo evidence

MICHAEL C. KNAUS, MICHAEL LECHNER AND ANTHONY STRITTMATTER

Estimators | 1 (see Knaus, Lechner, Strittmatter, '21)

Generic approaches (à la Knaus, Lechner, Strittmatter, 2021)

- Modify outcome & use standard ML (Section 4)
  - 1st: Signorovitch (2007), PhD thesis Harvard
- Modify covariates & use standard ML (Section 4)
  - 1st: Tian, Alizadeh, Gentles, Tibshirani (2014, JASA)
- Modify ML algorithms directly
  - CART → Causal Trees
    - Su, Tsai, Wang, Nickerson, Li (2009, *Journal of Machine Learning Research*)
    - Athey & Imbens (2016, *PNAS*)
  - Random Forests → Causal Forests (Section 3)
    - Introduced by Wager & Athey (2018, *JASA*);
    - **M**odified **C**ausal **F**orest: Lechner (2018), Lechner, Mareckova (2024) for theoretical guarantees
- Combine specific moment conditions & ML
  - **D**ouble/debiased **M**achine **L**earning: Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018, *Econometrics Journal*) (Section 2)
  - **G**eneralized **R**andom **F**orests: Athey, Tibshirani, Wager (2019, *Annals of Statistics*) (Section 3)

unicef | for every child

General approaches to effect estimation

Parameter specific estimation methods

- Use ‘best’ method for each aggregation level (levels of granularity)
  - May work better for specific aggregation levels
  - Difficult to understand & monitor performance of different estimators
    - Many tuning parameters, …
  - Internal consistency not guaranteed: Aggregates of heterogeneous effects may differ from average effects

Comprehensive Causal Machine Learning (CCML)

- Unified approach for estimating of all parameters of interest
- CCML estimators for different effects share important components (estimated by machine learning)

unicef | for every child

Comprehensive **C**ausal **M**achine **L**earning

Double/debiased **M**achine **L**earning

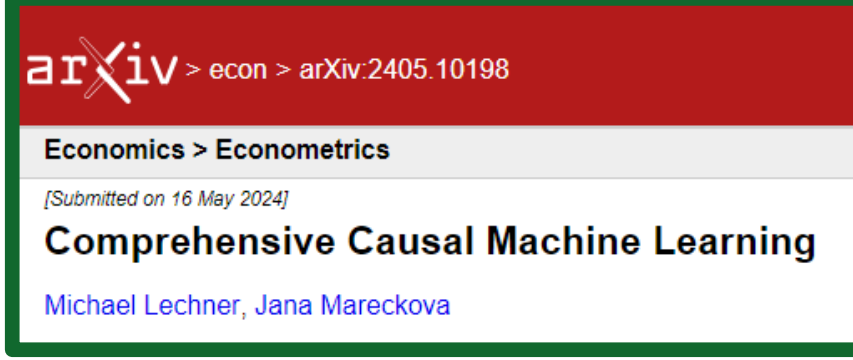
- Common element: **D**oubly **R**obust **S**core

Generalized **R**andom **F**orest

- Common element: Generalized Random Forest structure

Modified **C**ausal **F**orest

- Common element: Causal Forest
- Aggregate effects obtained by aggregation of finer-grained effects



Econometrics Journal (2018), volume **21**, pp. C1–C68.
doi: 10.1111/ectj.12097

Double/debiased machine learning for treatment and structural parameters

VICTOR CHERNOZHUKOV[†], DENIS CHETVERIKOV[‡], MERT DEMIRER[†], ESTHER DUFLO[†], CHRISTIAN HANSEN[§], WHITNEY NEWEY[†] AND JAMES ROBINS[‖]

GENERALIZED RANDOM FORESTS

BY SUSAN ATHEY*, JULIE TIBSHIRANI[†] AND STEFAN WAGER*

unicef | for every child

unicef | for every child

1 | Introduction

2 | Machine Learning vs Causal Machine Learning

3 | Double Debiased Machine Learning

4 | Causal Trees & Forests

5 | A comparison of Comprehensive Causal Machine Learners

6 | Optimal Policy & Algorithmic Decision Making

7 | Conclusions & outlook

unicef

The Econometrics Journal

Econometrics Journal (2018), volume **21**, pp. C1–C68.
doi: 10.1111/ectj.12097

Double/debiased machine learning for treatment and structural parameters

VICTOR CHERNOZHUKOV[†], DENIS CHETVERIKOV[‡], MERT DEMIRER[†], ESTHER DUFLO[†], CHRISTIAN HANSEN[§], WHITNEY NEWEY[†] AND JAMES ROBINS[‖]

Nice technical survey

Semiparametric Doubly Robust Targeted Double Machine Learning: A Review *†

Edward H. Kennedy
Department of Statistics & Data Science
Carnegie Mellon University

The Econometrics Journal

Econometrics Journal (2018), volume **21**, pp. C1–C68.
doi: 10.1111/ectj.12097

Comprehensive methodologies | 1

Double/debiased machine learning for treatment and structural parameters

VICTOR CHERNOZHUKOV[†], DENIS CHETVERIKOV[‡], MERT DEMIRER[†], ESTHER DUFLO[†], CHRISTIAN HANSEN[§], WHITNEY NEWEY[†] AND JAMES ROBINS[‖]

Double / debiased machine learning (DML): Theory

- Main idea
  - Use specific moment condition that fulfils **N**eyman **O**rthogonality **C**ondition
    - **Here**: Dependence of moment condition on propensity score ($P(D=1|X=x)$ & outcome equations ($EY|X=x, D=d$)) is such that small errors in those (*nuisance*) functions do not affect distribution of estimator
  - 1st step: Use ML to estimate nuisance functions
  - 2nd step: Solve moment conditions given estimated nuisance functions (X-fitting)
- This principle is applicable to many estimation problems
- It is related to *Double Robustness* (in treatment effect estimation)
  - DR in parametrics: OK if propensity score *or* outcome equations are misspecified
  - DR in CML: OK if propensity score & outcome equations are approximated with small enough errors

An alternative modelling approach | Notation

Common in this literature to use (slightly) more structure

$$Y = g_0(D,X) + U, \qquad E(U \mid X = x, D = d) = 0$$
$$D = m_0(X) + E, \qquad E(E \mid X = x) = 0$$

Assumption very similar to CIA as defined before

Without putting additional structure on *m(X)* & *g(D,X)*, the model is very general (& allows for almost general heterogeneity)

$$ATE = E\left[g_0(1,X) - g_0(0,X)\right] = E\left[\mu(1,X) - \mu(0,X)\right]$$
$$ATE(d) = E\left[g_0(1,X) - g_0(0,X) \mid D = d\right] = E\left[\mu(1,X) - \mu(0,X) \mid D = d\right]$$

unicef | for every child

$$Y = g_0(D,X) + U, \qquad E(U \mid X = x, D = d) = 0$$
$$D = m_0(X) + E, \qquad E(E \mid X = x) = 0$$
$$ATE = E\left[g_0(1,X) - g_0(0,X)\right], \quad ATE(d) = E\left[g_0(1,X) - g_0(0,X) \mid D = d\right]$$

Double machine learning for treatment effects | 1

Basic idea

- Step1 : Estimate unknown functions $g_0(.)$ **&** $m_0(.)$ by standard predictive ML
- Combine these estimates so that the estimator for the treatment effect has good properties
  - Based on moment conditions that fulfil the *Neyman orthogonality condition*
    - $W=(Y,D,X)$;
    - Assume there is a **low**-dimensional vector of *structural* parameters θ_0
    - Vector η_0 of *nuisance* parameters may be **high** dimensional
    - Let $\psi(W, \theta, \eta)$ be a function
    - Moment condition: $E\psi(W,\theta_0,\eta_0) = 0$
    - Neyman orthogonality condition (NOC): $\partial_\eta E\psi(W,\theta_0,\eta)\,|_{\eta=\eta_0} = 0$

Double machine learning | 2

$$E\psi(W,\theta_0,\eta_0)=0$$
$$\partial_\eta E\psi(W,\theta_0,\eta_0)\,|_{\eta=\eta_0}=0$$

A *Gateaux derivative* is a directional derivative

Suppose X and Y are locally convex topological vector spaces (for example, Banach spaces), $U \subset X$ is open, and $F : X \to Y$. The Gateaux differential $dF(u;\psi)$ of F at $u \in U$ in the direction $\psi \in X$ is defined as

$$dF(u;\psi) = \lim_{\tau\to 0} \frac{F(u+\tau\psi)-F(u)}{\tau} = \frac{d}{d\tau}F(u+\tau\psi)\Big|_{\tau=0}$$

If the limit exists for all $\psi \in X$, then one says that F is Gateaux differentiable at u.

Source: Wikipedia, May, 17, 2019

The NOC implies robustness of the resulting estimator of θ to small estimation errors in η

- Thus, a 2-stage estimator is possible that
  - (1) obtains sufficiently precise estimates of η (by machine learning)
  - (2) solves the moment condition for θ given the estimated η

The influence functions derived by Hahn (1998) fulfil these 2 conditions → leads to efficient estimation for *ATE, ATE(d)*

- Note: All semi-parametrically efficient scores share the NOC, but not all scores that have the NOC, are efficient

Fisher, A., E. H. Kennedy (2019): Visually Communicating and Teaching Intuition for Influence Functions, arXiv:1810.03260v2

Double machine learning | 3

ATE

$$\psi(W,\theta,g,m) = g(1,X) - g(0,X) + \frac{(Y - g(1,X))D}{m(X)} - \frac{(Y - g(0,X))(1-D)}{1-m(X)} - ATE$$

$$g_0(d,x) = \mu(d,x); \quad m_0(x) = p(x)$$

- Analogous expression for the *ATE(1), ATE(0)*

The double robustness property (leading to the NOC) turns out to be particularly useful in this context

- Small estimation errors for $\mu(d,x)$ & $p(x)$ can be ignored

Double machine learning | 4

Estimator suggested by CCDDHNR (2018):

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_{-i}(1,x_i) - \hat{\mu}_{-i}(0,x_i) + d_i \frac{y_i - \hat{\mu}_{-i}(1,x_i)}{\hat{p}_{-i}(x_i)} - (1-d_i)\frac{y_i - \hat{\mu}_{-i}(0,x_i)}{1-\hat{p}_{-i}(x_i)}$$

- The nuisance functions $\hat{\mu}_{-1}(1,x_i), \hat{\mu}_{-1}(1,x_i), \hat{p}_{-1}(x_i)$ are estimated without using observation '*i*'
  - cross-fitting (to avoid overfitting)
  - simplifies asymptotics, as observed & predicted quantities are uncorrelated

$$\sqrt{N}\frac{\widehat{ATE} - ATE_0}{E\left[\psi(Y,X,D,ATE_0,g_0,m_0)^2\right]} \xrightarrow{d} N(0,1) \qquad \text{(efficient, Hahn, 1998)}$$

The Econometrics Journal

Econometrics Journal (2022), volume **00**, pp. 1–26.
https://doi.org/10.1093/ectj/utac015

$$\mu(d,x) = E(Y \mid D = d, X = x)$$
$$p(x) = P(D = 1 \mid X = x)$$

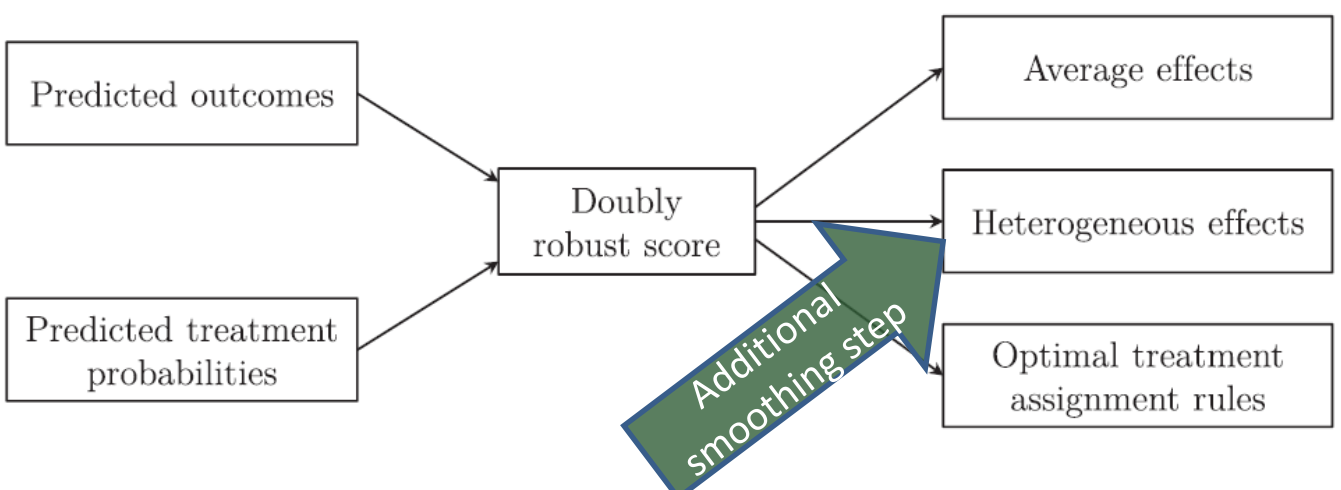
Double machine learning-based programme evaluation under unconfoundedness

MICHAEL C. KNAUS

DML for binary treatment

The central role of the (cross-fitted) **D**oubly **R**obust **S**core

$$DRS(y_i, d_i, x_i, \hat{\mu}_{-i}(\cdot), \hat{p}_{-i}(\cdot)) =$$
$$\hat{\mu}_{-i}(1, x_i) - \hat{\mu}_{-i}(0, x_i) + \frac{[y_i - \hat{\mu}_{-i}(1, x_i)]d_i}{\hat{p}_{-i}(x_i)} - \frac{[y_i - \hat{\mu}_{-i}(0, x_i)](1 - d_i)}{1 - \hat{p}_{-i}(x_i)}$$



DML for binary treatment | Averages of DR score

ATE

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} DRS(y_i, d_i, x_i, \hat{\mu}_{-i}(\cdot), \hat{p}_{-i}(\cdot))$$

$$\sqrt{N}\frac{\widehat{ATE} - ATE_0}{E\left[DRS(Y, D, X, g_0(\cdot), m_0(\cdot))^2\right]} \xrightarrow{d} N(0,1) \qquad \text{(as. efficient)}$$

GATE (z-discrete)

$$\widehat{GATE(z)} = \frac{1}{N_z}\sum_{i=1}^{N} 1(z_i = z) DRS(y_i, d_i, x_i, \hat{\mu}_{-i}(\cdot), \hat{p}_{-i}(\cdot)) \qquad N_z = \sum_{i=1}^{N} 1(z_i = z)$$

$$\sqrt{N}\frac{\widehat{GATE(z)} - GATE_0(z)}{E\left[DRS(Y, D, X, g_0(\cdot), m_0(\cdot))^2 \mid Z = z\right]} \xrightarrow{d} N(0,1)$$

- Identical estimator: OLS with z-specific dummy variables
  - Regular OLS standard errors valid

DML for binary treatment | DR learner

GATE (z continuous) & IATE(x)

- If *f(x)* parametrically specified & estimation by OLS

$$\widehat{CATE(x)} = \underset{f(x)}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \left[DRS(y_i, d_i, x_i, \hat{\mu}_{-i}(\cdot), \hat{p}_{-i}(\cdot)) - f(x) \right]^2$$

$$\sqrt{N} \frac{\widehat{CATE(x)} - f_0(x)}{E\left[\left[\widehat{CATE(x)} - f_0(x)\right]^2\right]} \xrightarrow{d} N(0,1)$$

  - Drawback: Not robust to misspecification of *f(x)*
    - *If linear & OLS: Best linear predictor of CATE(x)*
- If Z continuous, or of high-dimension & *f(x)* is nonparametrically estimated, or by ML
  - Estimator is conjectured to be consistent
  - Inference known only in special cases (e.g., continuous Z → nonparametric rate → Lechner, Zimmer, 2024)

unicef | for every child

Double machine learning | 6

Pro's

- Theoretical properties known & attractive
- Each ML estimators may converge at a speed of only $N^{1/4}$
  - Chose ML estimator known to work 'best' for particular prediction problem
  - Estimators with slower & faster speeds than $N^{1/4}$ can be combined
- Additional uncertainty due to sample splitting during cross-validation can be accounted for by repeating the estimation with different splits & analysing the resulting distribution of estimated ATEs
  - This uncertainty may be accounted for when constructing confidence intervals

Con's

- Weighting by *p(x)* may be unstable if estimated *p(x)* gets close to 0 or 1
  - If ML estimators predicts *p(x)* well so as to (almost) separate treated & controls
  - If (real) common support is weak (leading to first point)
- Limited theoretical guarantees high-dimensional *CATE(x)* rather unexplored (DR-Learner)

unicef | for every child

Extensions of DML theory (examples)

Econometrica, Vol. 90, No. 3 (May, 2022), 967–1027

AUTOMATIC DEBIASED MACHINE LEARNING OF CAUSAL AND STRUCTURAL EFFECTS

VICTOR CHERNOZHUKOV
Department of Economics, Massachusetts Institute of Technology

WHITNEY K. NEWEY
Department of Economics, Massachusetts Institute of Technology and NBER

RAHUL SINGH
Department of Economics, Massachusetts Institute of Technology

Econometrica, Vol. 90, No. 4 (July, 2022), 1501–1535

LOCALLY ROBUST SEMIPARAMETRIC ESTIMATION

VICTOR CHERNOZHUKOV
Department of Economics, MIT

JUAN CARLOS ESCANCIANO
Department of Economics, Universidad Carlos III de Madrid

HIDEHIKO ICHIMURA
Department of Economics, University of Arizona and Department of Economics, University of Tokyo

WHITNEY K. NEWEY
Department of Economics, MIT and NBER

JAMES M. ROBINS
Epidemiology, School of Public Health, Harvard University

Automatic Debiased Machine Learning for Dynamic Treatment Effects

Victor Chernozhukov MIT
Whitney Newey MIT
Rahul Singh MIT
Vasilis Syrgkanis Microsoft Research

Published as a conference paper at ICLR 2022

RIESZNET AND FORESTRIESZ: AUTOMATIC DEBIASED MACHINE LEARNING WITH NEURAL NETS AND RANDOM FORESTS

Victor Chernozhukov MIT
Whitney K. Newey MIT
Victor Quintas-Martinez MIT
Vasilis Syrgkanis Microsoft Research

Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression*

Victor Chernozhukov *MIT*
Whitney K. Newey *MIT*
Victor Quintas-Martinez *MIT*
Vasilis Syrgkanis *Microsoft Research*

A Simple and General Debiased Machine Learning Theorem with Finite Sample Guarantees

Victor Chernozhukov MIT Economics vchern@mit.edu
Whitney K. Newey MIT Economics wnewey@mit.edu
Rahul Singh MIT Economics rahul.singh@mit.edu

unicef | for every child

unicef | for every child

unicef | for every child

Estimation of $IATE(x) = \mu(1;x) - \mu(0;x)$

Causal problem is transformed into prediction problem(s)

- Machine learning (ML) methods are good with prediction problem(s)

Goal is to estimate a *difference of* ***2*** conditional expectations in ***2*** (nonrandom) *populations* defined by observed treatment status

- Nonstandard ML problem
- 'Ground truth' not observable ('true' dependent variable unobservable)

Selection into these (sub-)populations by 'selection process'

- *Propensity score,* $p_m(x)$*,* affects properties of estimators by changing the …
  - observed joint distribution of *X* & *Y* in the subpopulations (confounding)
  - sample sizes of subpopulations (bias & precision)

unicef | for every child

General remarks on predicting IATE(x)

The ML *goodness-of-fit perspective* on this problem

$$\widehat{IATE}(x) = \hat{\mu}(1;x) - \hat{\mu}(0;x); \qquad IATE(x) = \mu(1;x) - \mu(0;x)$$

$$\begin{aligned} MSE(\widehat{IATE}(x)) &= E(\widehat{IATE}(x) - IATE(x))^2 \\ &= E(\hat{\mu}(1;x) - \mu(1;x))^2 + E(\hat{\mu}(0;x) - \mu(0;x))^2 \\ &\qquad - 2E(\hat{\mu}(1;x) - \mu(1;x))(\hat{\mu}(0;x) - \mu(0;x)) \\ &= MSE(\hat{\mu}(1;x)) + MSE(\hat{\mu}(0;x)) - 2MCE(\hat{\mu}(1;x), \hat{\mu}(0;x)). \end{aligned}$$

The last term (**M**ean **C**orrelated **E**rror) cannot be easily estimated

- If x is continuous, y_i's in treatments *1* **&** 0 with same value of x will not be observed
- Different estimation methods approximate MCE() differently

unicef | for every child

Remarks on this prediction problem | 3

$$MSE\left[\widehat{IATE}(x)\right] = MSE\left[\hat{\mu}(1;x)\right] + MSE\left[\hat{\mu}(0;x)\right] - 2MCE\left[\hat{\mu}(1;x), \hat{\mu}(0;x)\right]$$

The 2 estimators of the 2 regression curves are tied together

- Positively correlated estimation errors of the regression curves matter less

MCE leads to biased estimation of the 2 regression curves

- If estimation errors are correlated across treatments, this will be a better estimator for the difference despite being a worse estimator for the 2 regression curves

ML principles may be modified

- to account for this 'tying together' of the regression curves
  - Different estimators use different approximations of the MCE()
- to remove bias of the estimator (needed for inference)

unicef | for every child

Recursive partitioning for heterogeneous causal effects

Susan Athey[a,1] and Guido Imbens[a]

[a]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Basic principles of *Causal Trees*

Homogeneous-in-X strata → no-selection-bias → difference of *Y* between treated & controls is 'good' estimator

Approximation of *min(MSE)* used: *max(Var(IATE(x)))*

Estimator of *IATEs(x)*: Difference of (weighted) within leaf treatment-control outcome means

Honesty

- Different data to build forest & to populate the final leaves with y_i
- Avoidance of overfitting
- Debiasing & allowing for inference

Currently, rarely used for estimation

- Trees tend to be unstable & to have too much variance
- Outperformed by Causal Forests

unicef | for every child

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager & Susan Athey

Basic principles of *Causal Forests* | 1

Use deep causal trees

- Low bias, but high variance

Use Random-Forest-like splitting algorithm with adapted splitting rule

- Different Causal Forests use different splitting rules that try to approximate the MSE of the IATE
  - Maximise treatment effect heterogeneity: CF (Wager, Athey, 2018), ***grf*** (Athey, Tibshirani, Wager, 2019)
  - Approximate estimation errors directly: *mcf* (Lechner, 2018)

Honest estimation to avoid bias & getting inference

- Different data to build forest & to populate the final leaves with y_i

unicef | for every child

The Annals of Statistics
2019, Vol. 47, No. 2, 1148–1178
https://doi.org/10.1214/18-AOS1709
© Institute of Mathematical Statistics, 2019

GENERALIZED RANDOM FORESTS

BY SUSAN ATHEY*, JULIE TIBSHIRANI† AND STEFAN WAGER*

Generalized Random Forests | 1 | Estimation of IATE

Extension of Random Forest to nonparametric local GMM

Local moment conditions: $E\left(\psi^{grf}(Y,X,D;\theta^0(X),\eta^{grf,0}(X))\middle|X=x\right)=0$; $\theta^0(x)$: True value of parameter of interest

$\eta^{grf,0}(x)$: True value of nuisance parameters

$$\widehat{IATE}^{grf}(x)=\hat{\theta}^{grf}(x)=\left(\sum_{i=1}^{N}w_i^{grf}(x)\left(d_i-\bar{d}_w\right)\left(d_i-\bar{d}_w\right)'\right)^{-1}\left(\sum_{i=1}^{N}w_i^{grf}(x)\left(d_i-\bar{d}_w\right)\left(y_i-\bar{y}_w\right)\right),$$

$$\bar{d}_w=\sum_{i=1}^{N}w_i^{grf}(x)d_i$$

$$\bar{y}_w=\sum_{i=1}^{N}w_i^{grf}(x)y_i$$

- Weights obtained from gradient-based splitting rule → maximise treatment effect heterogeneity
- Honesty (instead of cross-fitting)

unicef | for every child

Generalized Random Forests | 2 | Estimation ATE & GATE

ATEs & GATEs are estimated through additional regressions using *grf* score function

- *dml*-<u>like</u> score function:

$$\hat{\Gamma}^{grf}(o_i;\hat{\eta}) = \hat{\theta}(x_i) + \frac{d_i(y_i - \hat{\mu}(1;x_i))}{\hat{p}(x_i)} + \frac{(1-d_i)(y_i - \hat{\mu}(0;x_i))}{1-\hat{p}(x_i)}$$

- ATE:

$$\widehat{ATE^{grf}}(m,0) = \hat{\theta}^{grf} = \frac{1}{N}\sum_{i=1}^{N}\hat{\Gamma}^{grf}(o_i;\hat{\eta})$$

$$\widehat{Var}(\widehat{ATE^{grf}}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\Gamma}^{grf}(o_i;\hat{\eta})^2$$

- GATEs
  - Regress score on Z
  - Heteroscedasticity-robust standard errors of OLS regression

unicef | for every child

The Applied Researchers' Wish List

Properties of estimator to be attractive in practice

Point estimators & inference for all effects

- All relevant aggregation levels
- Multiple treatments

Desirable statistically properties for reasonable *N*

- RMSE better than difference-of-regressions estimator
- Good basis for reliable inference: Smallish bias & normally distributed estimator
  - Maybe at the cost of some efficiency loss

Theoretical asymptotic guarantees for *point estimator* & *inference*

- Consistency, asymptotic normality for all parameters of interest
- Consistent estimator of standard errors / valid inference

Computationally not too demanding

- No separate estimations for different aggregation levels

unicef | for every child

IZA DP No. 12040

Modified Causal Forests for Estimating Heterogeneous Causal Effects

Michael Lechner

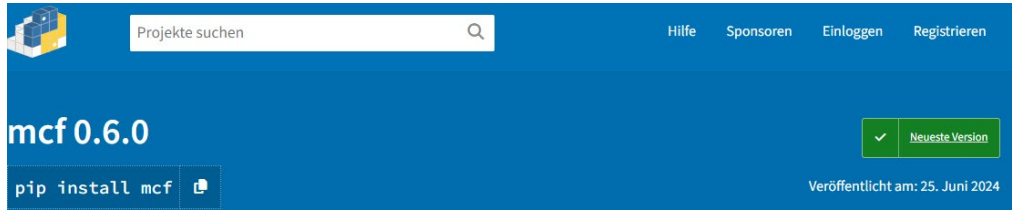
DECEMBER 2018

Modified **C**ausal **F**orest | Main ideas

The *mcf* is a modification of the Causal Forest by Wager & Athey (2018)

Important modification

- Splitting rule based on alternative proxy for MSE
  - Nearest neighbour approximation of the MCE
- Penalty function to favour split that lead to more extreme propensity scores in leaves
  - to further reduce selection bias
- Representation of causal effect as weighted mean of y_i's is used for inference
  - Requires different implementation of sample splitting
- *ATE* & *GATE(z)* are obtained by aggregating *IATE(x)*'s



Objective function | MCE | 2

Optimize sample splits in training sample w.r.t.

$$MSE\left[\hat{\mu}(1;x)\right]+MSE\left[\hat{\mu}(0;x)\right]-2MCE\left[\hat{\mu}(1;x),\hat{\mu}(0;x)\right]$$

$$\widehat{MSE}\left[\hat{\mu}_d(x)\right]=\frac{1}{N_{S_x}^d}\sum_{i=1}^{N}\underline{1}(x_i\in S_x)\underline{1}(d_i\in d)\left[\overline{y}_{S_x,d}-y_i\right]^2,\quad x\in S_x,\qquad \tilde{y}_{(i,m)}=\begin{cases} y_i & if \quad d_i=m \\ y_{(i,m)} & \quad\quad d_i\neq m \end{cases}$$

$$\widehat{MCE}(x)=\frac{1}{N}\sum_{i=1}^{N}\underline{1}(x_i\in S_x)\left[\overline{y}_{S_x,1}-\tilde{y}_{(i,1)}\right]\left[\overline{y}_{S_x,0}-\tilde{y}_{(i,0)}\right]$$

Unknown counterfactuals substituted by nearest neighbours using main diagonal of Mahalanobis distance

- Matching on the propensity score(s) is a bad idea: MCE is not (directly) related to selection bias
- Matching on predictive scores works: Computationally more expensive than Mahalanobis matching as predictive scores need to be estimated by ML

Objective function | Penalty term

Main idea

- Add penalty term that penalizes splits with similar treatment probabilities
  - Reduces selection bias effectively

Example of a penalty function

$$penalty(x', x'') = \lambda \left\{ 1 - \frac{1}{M} \sum_{d=0}^{M-1} (P(D = d \mid X \in leaf(x') - P(D = d \mid X \in leaf(x''))^2 \right\}, \quad \lambda \geq 0$$

- = *0* if split leads to a perfect prediction of the probabilities in the daughter leaves
- = λ if treatment probabilities in daughter leaves are the same
- Cheap to compute as it only counts the treatments in any potential split

unicef | for every child

Weights for aggregation & inference | 1

Estimators coming from RF/CF can be expressed as weighted outcome means

Estimators for intermediate levels can be obtained by aggregating weights

- Computationally convenient

Inference

- Weight-based inference
  - Same principle for all aggregation levels
  - Sample splitting required

unicef | for every child

Background

Inference | General weight-based estimators | 1

Structure of estimator: $\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i; \qquad Var(\hat{\theta}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i\right).$

Law of total variance (*Var(Y)=E[Var(Y|X)]+Var[E(Y|X)]*):

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i\right) = E_{\hat{W}} Var\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i \mid \hat{w}_1,...,\hat{w}_N\right) + Var_{\hat{W}} E\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i \mid \hat{w}_1,...,\hat{w}_N\right)$$

i.i.d. sampling:

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i \mid \hat{w}_1,...,\hat{w}_N\right) = \frac{1}{N^2}\sum_{i=1}^{N} \hat{w}_i^2 Var(y_i \mid \hat{w}_1,...,\hat{w}_N)$$

$$E\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i \mid \hat{w}_1,...,\hat{w}_N\right) = \frac{1}{N}\sum_{i=1}^{N} \hat{w}_i E(y_i \mid \hat{w}_1,...,\hat{w}_N)$$

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i y_i\right) = E_{\hat{W}}\left(\frac{1}{N^2}\sum_{i=1}^{N} \hat{w}_i^2 Var(y_i \mid \hat{w}_1,...,\hat{w}_N)\right) + Var_{\hat{W}}\left(\frac{1}{N}\sum_{i=1}^{N} \hat{w}_i E(y_i \mid \hat{w}_1,...,\hat{w}_N)\right).$$

RF: $\hat{w}_i = \hat{w}(x_i, \vec{X}^T, \vec{Y}^T)$ $(\vec{X}^T, \vec{Y}^T)$ are the training data to build RF

Background

Inference | General weight-based estimators | 2

If i does not belong to training data, and observations are i.i.d., y_i is independent of $\hat{w}_j$

$$E(y_i \mid \hat{w}_1, \ldots, \hat{w}_N) = E(y_i \mid \hat{w}_i) = \mu_{Y|W}(\hat{w}_i)$$

$$Var(y_i \mid \hat{w}_1, \ldots, \hat{w}_N) = Var(y_i \mid \hat{w}_i) = \sigma^2_{Y|W}(\hat{w}_i)$$

This leads to a simpler formula for the variance:

$$Var\left(\frac{1}{N}\sum_{i=1}^{N}\hat{w}_i y_i\right) = E_{\hat{W}}\left(\frac{1}{N^2}\sum_{i=1}^{N}\hat{w}_i^2\sigma^2_{Y|\hat{W}}(\hat{w}_i)\right) + Var_{\hat{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{w}_i\mu_{Y|\hat{W}}(\hat{w}_i)\right).$$

Inference | General weight-based estimators | 3

$$Var(\hat{\theta}) = E\left(\left[Var(\hat{\theta} \mid \hat{W}\right]\right) + Var\left[E(\hat{\theta} \mid \hat{W})\right] \qquad \text{Law of total variance}$$

… suggests to use the following estimator for the variance:

$$\widehat{Var(\hat{\theta})} = \frac{1}{N^2}\sum_{i=1}^{N} \hat{w}_i^2 \hat{\sigma}^2_{Y|\hat{W}}(\hat{w}_i) + \frac{1}{N(N-1)}\sum_{i=1}^{N}\left[\hat{w}_i \hat{\mu}_{Y|\hat{W}}(\hat{w}_i) - \frac{1}{N}\sum_{i=1}^{N}\hat{w}_i \hat{\mu}_{Y|\hat{W}}(\hat{w}_i)\right]^2$$

These are 2 1-dimensional problems. Therefore, standard non-parametric estimators are fast & accurate for $\hat{\mu}_{Y|\hat{W}}(\hat{w}_i)$ & $\hat{\sigma}^2_{Y|\hat{W}}(\hat{w}_i)$

- Some estimation noise will be washed out by averaging these non-parametric estimators

unicef | for every child

Background

Inference | General weight-based estimators | 4

Implementation

- *k*-nearest neighbour estimators (*k* increases slowly with *N*)
  - Computationally convenient (fast)
  - Alternatively, Nadaraya-Watson estimation
- Bodory, Camponovo, Huber, & Lechner (2020) investigate *k*-nearest neighbour estimators to obtain these weights
  - they found good results in a binary treatment setting for the ATET
- Finite sample properties of different version not yet investigated

Clustering can also 'easily' be incorporated

Shortcoming: Lack of theory for choosing smoothing parameters

unicef | for every child

Background

Inference | General weight-based estimators | 5

Final remarks on inference

- Would be nice to have some asymptotic theory for the estimators of conditional means & variances to get consistency conditions (problem a bit more non-standard as weights come from a machine learner)
- The more spread-out the weights are (as for ATE & GATEs), the better the inference, as it may not require consistent estimation of $\hat{\mu}_{Y|\hat{W}}(\hat{w}_i)$, $\hat{\sigma}^2_{Y|\hat{W}}(\hat{w}_i)$
- With the less spread-out weights of IATE, the estimator will much more benefit from good estimation of $\hat{\mu}_{Y|\hat{W}}(\hat{w}_i)$, $\hat{\sigma}^2_{Y|\hat{W}}(\hat{w}_i)$
- Open issue: Are the choices for *k* also reasonable for IATE (that removes less noise by averaging), or do we need larger or smaller *k*?

unicef | for every child

Background

Estimator of (G)ATE as function of estimator of IATE

$$\widehat{GATE}(m,l;z,\Delta) = \frac{1}{N^{z,\Delta}} \sum_{i=1}^{N} \underline{1}(z_i = z, d_i \in \Delta) \widehat{IATE}(m,l;x_i)$$

$$= \frac{1}{N^{z,\Delta}} \sum_{i=1}^{N} \underline{1}(z_i = z, d_i \in \Delta) \frac{1}{N} \sum_{j=1}^{N} \hat{w}_j^{IATE(m,l;x_i)} y_j$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N^{z,\Delta}} \sum_{j=1}^{N} \underline{1}(z_j = z, d_j \in \Delta) \hat{w}_i^{IATE(m,l;x_j)} \right) y_i = \frac{1}{N} \sum_{i=1}^{N} \hat{w}_i^{GATE(m,l;z,\Delta)} y_i;$$

$$\hat{w}_i^{GATE(m,l;z,\Delta)} = \frac{1}{N^{z,\Delta}} \sum_{j=1}^{N} \underline{1}(z_j = z, d_j \in \Delta) \hat{w}_j^{IATE(m,l;x_i)}; \qquad N^{z,\Delta} = \sum_{i=1}^{N} \underline{1}(z_i = z, d_i \in \Delta).$$

$$\widehat{ATE}(m,l;\Delta) = \frac{1}{N^{z}} \sum_{i=1}^{N} \underline{1}(d_i \in \Delta) \widehat{IATE}(m,l;x_i) = \frac{1}{N} \sum_{i=1}^{N} \hat{w}_i^{ATE(m,l;\Delta)} y_i;$$

$$\hat{w}_i^{ATE(m,l;z,\Delta)} = \frac{1}{N^{\Delta}} \sum_{j=1}^{N} \underline{1}(d_j \in \Delta) \hat{w}_j^{IATE(m,l;x_i)}; \qquad N^{\Delta} = \sum_{i=1}^{N} \underline{1}(d_i \in \Delta).$$

unicef | for every child

Background

Implementation of estimation & inference | 1

1) Split the estimation sample randomly into two parts of equal size (sample A and sample B)

2) Estimate the trees that define the random forest in sample A

- *Modified Causal Forest*: Estimate the same forest for all treatment states jointly. Before building the first tree, for each observation in each treatment state, find a close 'neighbour' in every other treatment state and save its outcome (to estimate MCE). The splitting rule is based on minimising the overall MSEs, taking account of all MCEs & and penalty term

unicef | for every child

Background

Implementation of estimation & inference |2

3) Apply the sample splits obtained in sample A to all subsamples (by treatment state) of sample B and take the mean of the outcome in the respective leaf as the prediction that comes with this Causal Forest

4) Obtain the weights from the estimated Forest by counting how many times an observation in sample B is used to predict *IATE(x)* for a particular value of *x*.

5) Aggregate the IATEs to GATEs by taking the average over observations in sample B with the same value of *z* and treatment group Δ. Do the same aggregation with the weights to obtain the weights for GATEs.

6) Do the same steps as in 5) to obtain the ATEs, but average over all observations in treatment group Δ.

unicef | for every child

Background

Implementation of estimation & inference |3

7) Compute weight-based standard errors. Use these estimated standard errors together with the quantiles from the normal distribution to obtain critical values (p-values).

unicef | for every child

Modified Causal Forest | 3 | Asymptotic properties

ATE & GATE & IATE

- Consistent, asymptotically normal

ATE:
$$\frac{\hat{\theta}^{mcf} - \theta^0}{\sqrt{Var\left(\hat{\theta}^{mcf}\right)}} \xrightarrow{d} N(0,1)$$

GATE(z):
$$\frac{\hat{\theta}^{mcf}(z) - \theta^0(z)}{\sqrt{Var\left(\hat{\theta}^{mcf}(z)\right)}} \xrightarrow{d} N(0,1)$$

IATE(x):
$$\frac{\hat{\theta}^{mcf}(x) - \theta^0(x)}{\sqrt{Var\left(\hat{\theta}^{mcf}(x)\right)}} \xrightarrow{d} N(0,1)$$

arXiv > econ > arXiv:2405.10198

Economics > Econometrics

[Submitted on 16 May 2024]

Comprehensive Causal Machine Learning

Michael Lechner, Jana Mareckova

unicef | for every child

unicef | for every child

1 | Introduction

2 | Machine Learning vs Causal Machine Learning

3 | Double Debiased Machine Learning

4 | Causal Trees & Forests

5 | A comparison of Comprehensive Causal Machine Learners

6 | Optimal Policy & Algorithmic Decision Making

7 | Conclusions & outlook

unicef | for every child

Summary of theoretical properties

ATE

- All estimators are consistent, asymptotically normal
- *dml* & *grf* are efficient
- *mcf*: ATE is sum of IATEs

GATE

- All estimators are consistent, asymptotically normal (*Z* discrete)
- *mcf*: GATE is sum of IATEs in subsamples defined by values of *Z*

IATEs (finite dimensional *X*)

- *mcf* & *grf* are consistent, asymptotically normal
- *dml (DR-learner)* is consistent, asymptotically normal for parametric approximations (e.g., best linear predictors)

unicef | for every child

Goal | 1

Comprehensive analysis of behaviour of *grf*, *dml*, *mcf* (& in different versions, OLS) in finite (but larger) samples

- Quality of point estimates **&** inference procedures

Computationally very demanding

- Many scenarios needed to see if some method breaks down in certain situations

unicef | for every child

General considerations | 1

Simulations (Econometrics) vs. *benchmark studies* (Machine learning)

- Difficult to use for inference (requires replications) ➔ *simulations*

Empirical Monte Carlo Study (EMCS) vs. *synthetic* design

- EMCS somewhat more realistic for special case
- We want to vary very many characteristics → *synthetic design*
  - Type & quantity of covariates, # of treatments , effect heterogeneity, degree of selectivity, sample size, functional forms, influence of covariates on outcomes & heterogeneity, share of treated

Fixed or *random* features (*X*)

- Only fixed-*X* design allows to analyse distributions of *IATEs(x)* for given *x*
- Random-*X* design better reflects distribution theory ➔ *fixed X*

unicef | for every child

Simulation design | 1

Repeat R times

- Draw 2 data sets of size *N*
  - Data set 1 (to train forest): contains *D*, Y, X
  - Data set 2 (to predict effects): contains *IATE(x)*, X
- Estimate effects & standard errors
- Save results

Compute performance measures

- 1st to 4th moments, bias, MSE, absolute error, bias of standard error, coverage probability

unicef | for every child

Goal | 1

Compare finite sample behaviour of *dml*, *grf*, *mcf*

- Quality of point estimates **&** inference procedures (out-of-training sample)

Synthetic Monte Carlo study with different data generating processes

- **Sample size**: N=2'500 & 10'000
- **Covariates**
  - # of covariates in data: p = 10, 20, 50
  - # of covariates in DGP (sparsity): k = 0.2 p, 0.5 p, 0.8 p
  - Type of covariates: Normal, uniform, dummy (pure & mixtures of different types)
  - Z: Continuous covariate split into 5, 10, 20, 40 groups
- **Treatment**
  - # of treatments: 2, 4
  - Strength of selection (R^2 = 0, 10%, 42%)
  - Treatment shares: 25%, 50%
- **Potential outcomes**
  - Y^0: Sine function of linear index, R^2 = 0, 10%, 45%
  - Y^1: Y^0 + IATE + noise
  - IATE: 0, Linear, quadratic, logistic, step functions

unicef | for every child

Summary | 1

dml

- Dominates when # of groups is small & selectivity is not too large
- Performance may drastically deteriorate for a larger # of groups or stronger selectivity
- Not useful for IATEs

grf (pre-centred)

- Similar good than *dml* when # of groups is small & selectivity is not too large
- Performance drastically deteriorate for a larger # of groups (& somewhat for strong selectivity)
- Good performance for IATE

unicef | for every child

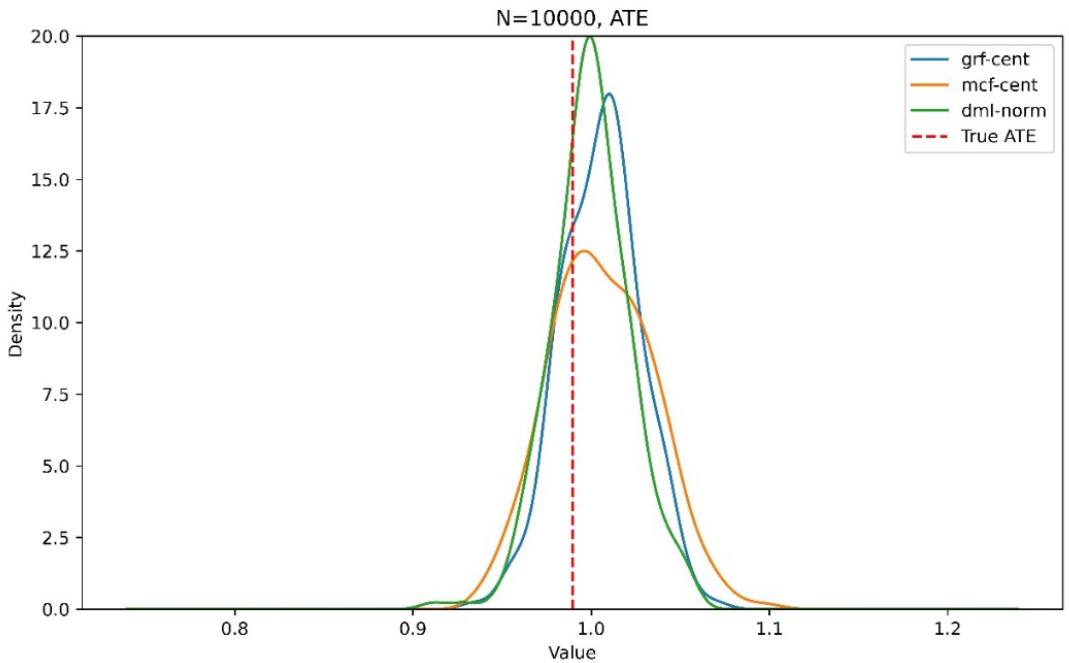
Summary | 2

mcf (pre-centred)

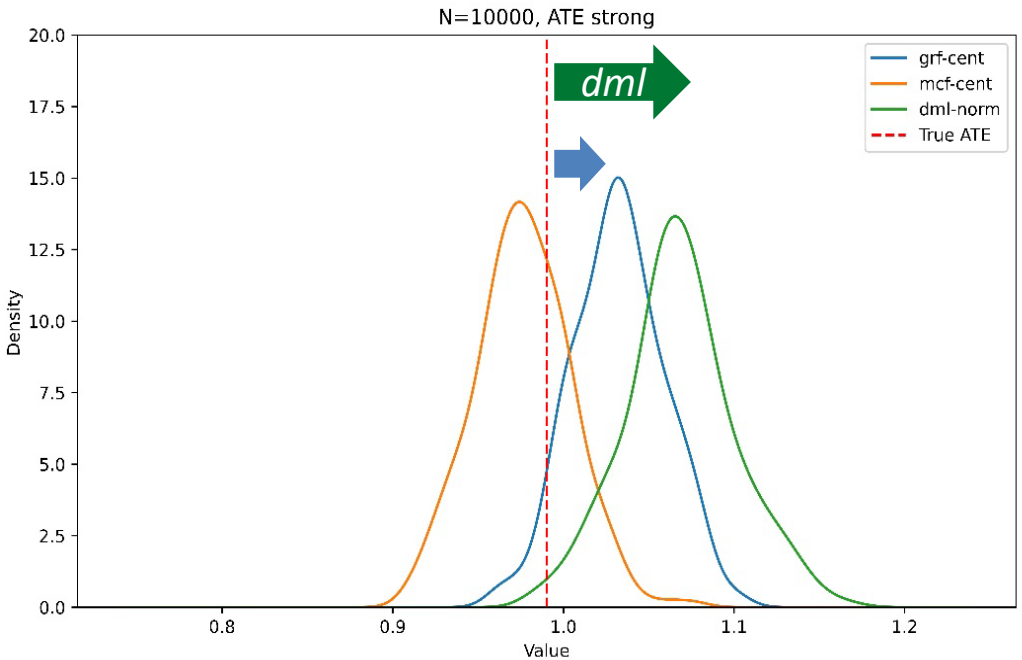
- Less efficient when # of groups is small & selectivity is not too large
- Much less affected by larger # of groups & stronger selectivity
- Overall robust (& competitive) behaviour

ATE estimator may be sensitive to different levels of selectivity

Medium selectivity (R^2=10%)



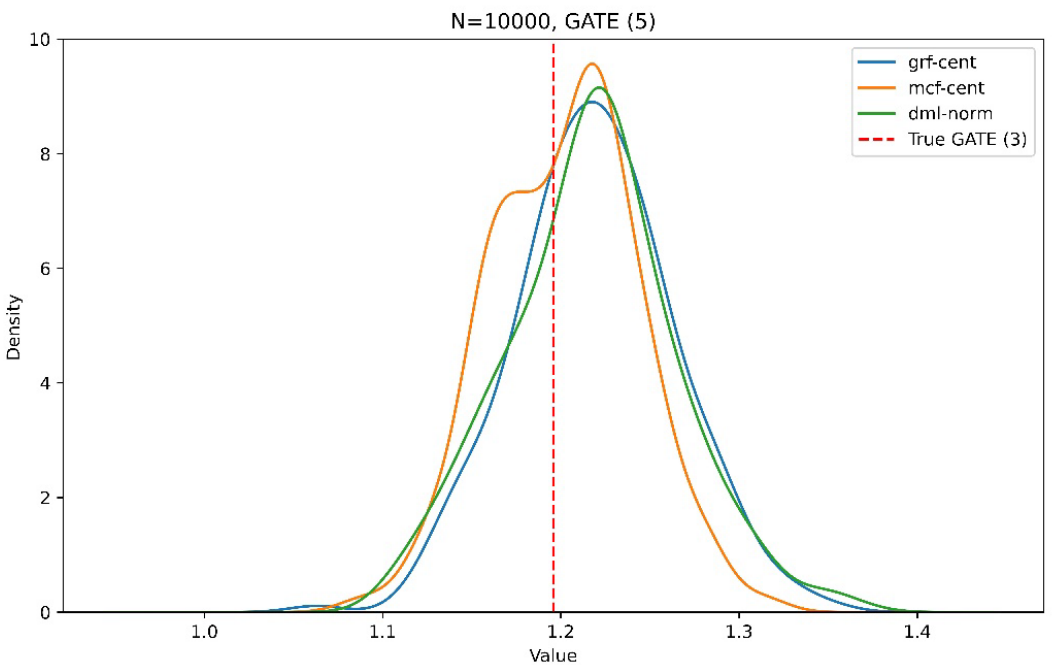
Strong selectivity (R^2=42%)



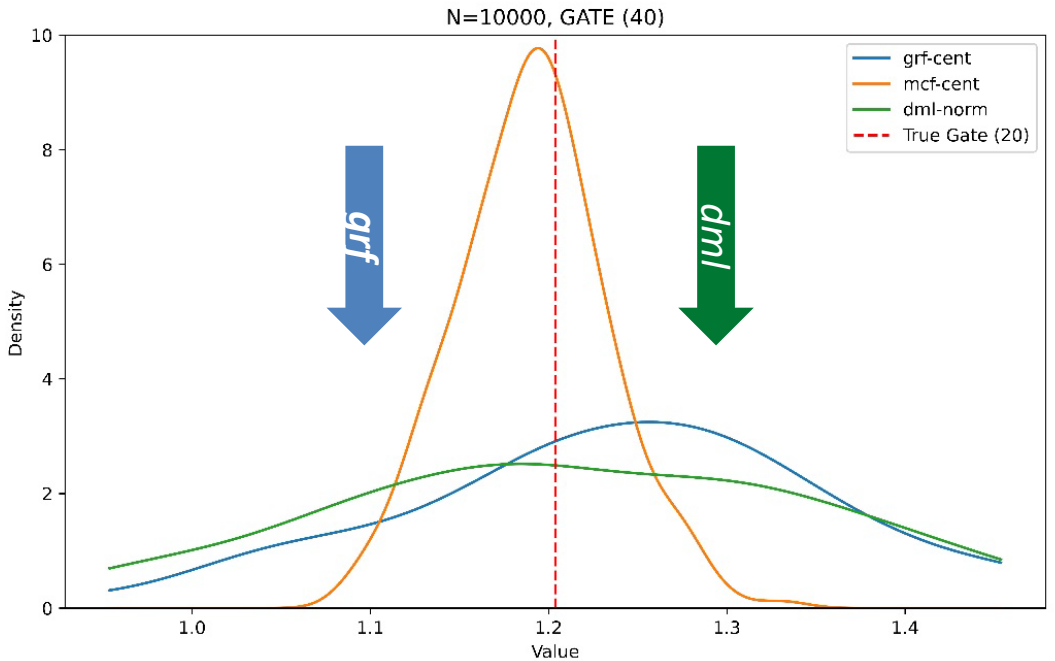
Bias (& MSE) is even larger for smaller sample

GATE estimator may be sensitive w.r.t. different # of groups

GATE with 5 groups



GATE with 40 groups



Variance (MSE) is even larger for smaller sample

unicef | for every child

Recommendations for empirical work

Large sample available (standard error reduction not a major concern)

- *mcf*

Smaller sample (standard error reduction is important)

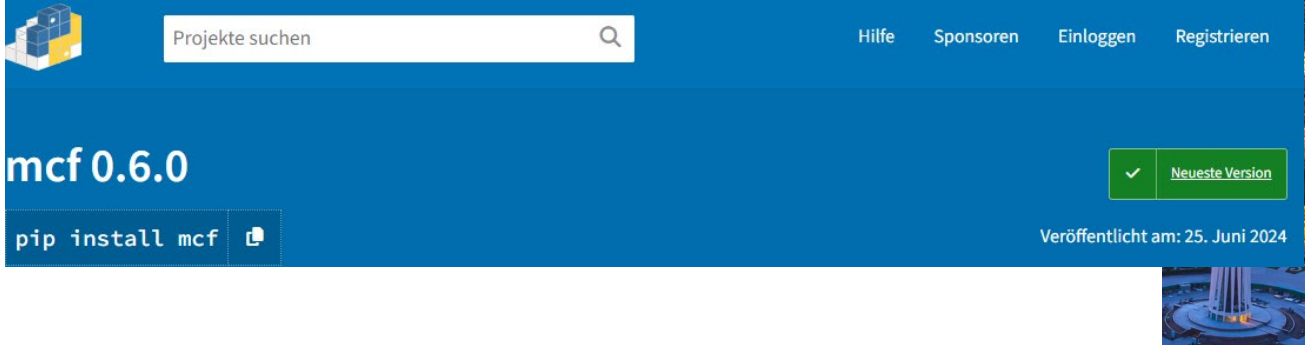
- ATE, GATE (few groups): *dml* or *grf* (robustness check with more robust mcf)
- GATE (many groups): *mcf*
- IATEs: *grf* or *mcf*

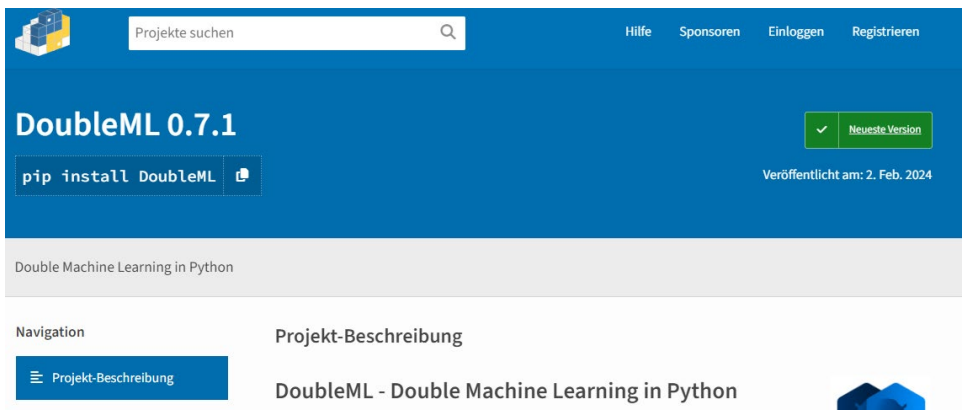
A note on computation

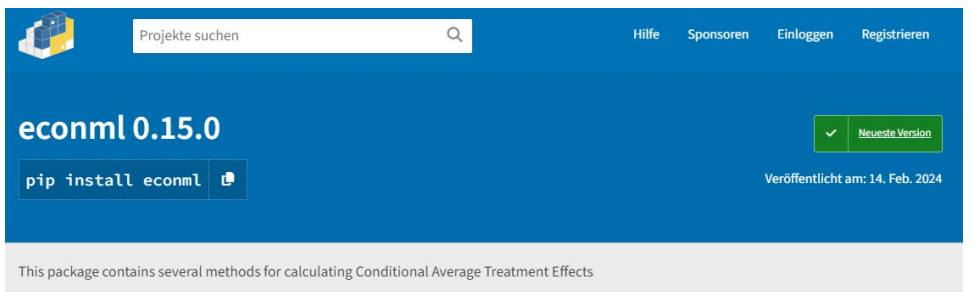
- User friendly packages available in *Python* & *R* (& Stata for *dml*)

unicef | for every child

Some packages







DoubleML: Double Machine Learning in R

Implementation of the double/debiased machine learning framework of Chernozhukov et a
regression models. 'DoubleML' allows estimation of the nuisance parts in these models by
based on the 'R6' package is very flexible. More information available in the publication ir

Version:	1.0.0
Depends:	R (≥ 3.5.0)
Imports:	R6 (≥ 2.4.1), data.table (≥ 1.12.8), stats, checkmate, mlr3 (≥ 0.5.0), m
Suggests:	knitr, rmarkdown, testthat, covr, patrick (≥ 0.1.0), paradox (≥ 0.4.0), d
Published:	2024-02-15
Author:	Philipp Bach [aut, cre], Victor Chernozhukov [aut], Malte S. Kurz [au

grf: Generalized Random Forests

Forest-based statistical estimation and inference. GRF provides non-parametric methods for heterogeneous tre
regression, and survival regression, all with support for missing covariates.

Version:	2.3.2
Depends:	R (≥ 3.5.0)
Imports:	DiceKriging, lmtest, Matrix, methods, Rcpp (≥ 0.12.15), sandwich (≥ 2.4-0)
LinkingTo:	Rcpp, RcppEigen
Suggests:	DiagrammeR, MASS, rdd, survival (≥ 3.2-8), testthat (≥ 3.0.4)
Published:	2024-02-25
Author:	Julie Tibshirani [aut], Susan Athey [aut], Rina Friedberg [ctb], Vitor Hadad [ctb], Davi
Maintainer:	Erik Sverdrup <erik.sverdrup at monash.edu>

unicef | for every child

arXiv > econ > arXiv:2405.10198

Economics > Econometrics

[Submitted on 16 May 2024]

Comprehensive Causal Machine Learning

Michael Lechner, Jana Mareckova

unicef | for every child

unicef | for every child

Optimal policy & individualized treatment rules | 2

Holy grail & final culmination of Causal Machine Learning

- So far: We estimated heterogeneous causal effects
- Now: We use this knowledge to improve decisions
  - This may also lead to changes in the estimation of the heterogenous effects
  - New & rapidly developing literature in all fields

This lecture focuses mainly on the following papers

- *Manski (2004):* Statistical Treatment Rules for Heterogeneous Populations, *Econometrica, 72*
- *Athey, Wager (2020):* Policy Learning with Observational Data, *Econometrica*
- *Kitagawa, Tetenov (2018):* Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice, *Econometrica*, 86, 591–616
- *Zhou, Wager, Athey (2019):* Offline Multi-Action Policy Learning: Generalization & Optimization, arXiv

unicef | for every child

Optimal policy & individualized treatment rules | 3

Economics

- *Active labour market policy*: Which type of unemployed should be assigned to which type of active labour market programme (subject to budget constraint)?
- *Regional policy*: Which region gets which type of subsidized infrastructure investment?

Business

- *Private investment decisions*: In which type of machinery should we invest to maximise profits?

Medicine

- *Personalized medicine*: Which patient gets which type of treatment?

…

unicef | for every child

Optimal policy & individualized treatment rules | 4

Naïve Rule

- Estimate IATEs by some good estimator & allocate treatment with highest estimated potential outcome

Naïve rule may have limitations

- Estimation problem for IATE & optimal policy (allocation) rule differ
  - MSE minimisation vs. classification type problems
- Does not respect possible budget constraints

Algorithms with proven properties are mainly available for experiments & selection-on-observables

- Baseline settings in rich data environments
- Other identification strategies not much trusted outside economics (IV, …)
- Other IS may identify effects only for (unknown) compliers (IV, RDD) or treated (DiD)

Notation

$D \in \{0,1,...,M-1\}$	M exclusive treatments
Y^m	M potential outcome (PO) of treatment m
$Y = \sum_{m=0}^{M} \underline{1}(d=m)Y(m)$	observed outcomes
$X = (X^1,...,X^P)$	covariates (exogenous)
$p_m(x) = P(D=m \mid X=x)$	propensity scores for given value of x
$\mu(m,x) = E(Y(m) \mid X=x)$	expected PO of m for given value of x
$IATE^{m,l}(x) = \mu(m,x) - \mu(l,x)$	individualised average treatment effect for x
$\pi(x) \in \{0,1,...,M-1\}$	assignment (policy) rule for given value of x
$Y(\pi(x))$	PO of policy rule $\pi(x)$ for given value of x
$Q(\pi) = E\left[Y(\pi(X))\right]$	policy value function ($\forall x \in \chi$)
Π	class of allowed policies

unicef | for every child

Optimal policy & minimum regret | 1

$$\pi^* = \arg\max_{\pi \in \Pi} EY(\pi(X)), \forall x \in \chi \qquad \text{optimal policy}$$

$$R(\pi) = Q(\pi^*) - Q(\pi), \pi \in \Pi, \forall x \in \chi \quad \text{regret}$$

Remarks

- Maximising *Q(π)* is equivalent to minimising *R(π) (minimal regret)*
- Minimizing regret depends on high quality estimates of PO
- Nevertheless, this is closer to classification than to regression problems
- Example: $x = a: \quad Y(1,a) = Y(0,a); \qquad x = b: \quad Y(1,b) \ll Y(0,a)$
  - Optimal policy: **Estimation error for *x=a* does not matter**, for *x=b* it matters hugely
  - Regression-type: They both matter (by how much, depends on distribution of *X*)

unicef | for every child

Minimax regret

Usually, optimal allocation such that it minimizes the largest possible regret over all possible data generating processes

$$\pi^{MM*} = \inf_{\pi \in \Pi} \sup_{f(X,Y)} R(\pi)$$

Usually, this is the best possible property that can be provenly attributed to a certain policy allocation algorithm

However, difficult (impossible?) to achieve with feasible estimators

- Recently advances by considering restricted policy classes (reducing complexity of policy rules)
  - More on such approaches in next section

unicef | for every child

Optimal policy & individualized treatment rules | 1

Further remarks

- Major challenges are
  - Estimating expected potential outcomes (or the elements of an alternative policy score) precisely
  - Finding best policy within potentially huge numbers of possible allocations of treatments
    - Could be particularly demanding (statistically & computationally) when there are …
      - » many covariates
      - » many treatments
      - » ‘difficult’ constraints
      - » dynamics, i.e. sequences of treatments

unicef | for every child

Online vs offline policy learning

Offline policy learning

- Experimental or non-experimental data is already available
- Use this data to learn good policy rules to apply in future
- Good approach if data is large & informative & world is stationary

Online policy learning

- Gather new data (usually experimental)
- Allocate such as to …
  - get good policy (exploitation)
  - get better estimates (possibly suboptimal allocation in smallish groups; exploration)
  - This trade-off needs to be optimized to get good long-term results
    - Large & growing literature (Bandits) on how to best do this

unicef | for every child

Background

Kitagawa-Tetenov (2018) | 1

Binary treatment, known propensity score

Empirical welfare maximisation method

- Maximise sample analogue of average social welfare (sum of potential outcomes)
- Features
  - Constraints on policy
    - Ethical etc. constraints
    - Budget → units per treatment
  - Simple rules based on few 'scores'

Properties for known propensity score

- √N - convergence, *minimax optimality* under fairly general conditions (& unconfoundedness)

Background

Kitagawa-Tetenov (2018) | 2

Structure of estimator

Let $\pi(x)$ be the set of values of x for which the treatment is assigned

- This set has lower dimension than x (restricted policies)

The *Empirical Welfare* of a policy is given by

$$W(\pi) = E\left[\frac{DY}{p(X)}\underline{1}(X \in \pi) + \frac{(1-D)Y}{1-p(X)}\underline{1}(X \notin \pi)\right] = EY^0 + E\left[\left(\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\right)\underline{1}(X \in \pi)\right]$$

Optimal policy: $$\pi^* = \arg\max_{\pi \in \Pi} W(\pi) = \arg\max_{\pi \in \Pi} E\left\{\left[\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\right]\underline{1}(X \in \pi)\right\}$$

unicef | for every child

Background

Kitagawa-Tetenov (2018) | 3

Estimation: Solve empirical analogue

$$\hat{\pi} = \arg\max_{\pi \in \Pi} W(\pi) = \arg\max_{\pi \in \Pi} \frac{1}{N} \sum_{i=1}^{N} \left[\frac{d_i y_i}{p(x_i)} \underline{1}(x_i \in \pi) + \frac{(1-d_i) y_i}{1-p(x_i)} \underline{1}(x_i \notin \pi) \right]$$

If the link of *x* to the policy is not substantially restricted, this estimation poses huge computational (as well as statistical) challenges (Mixed Integer Linear Programming)

unicef | for every child

Athey & Wager (2020) | 1

Binary treatment (essential)

Selection-on-observables (essential)

Double machine learning-type-of-argument incorporated

Restricted class of policies

- Linear-in-x-rules (not too many x) & (shallow) decision trees
- Leads to much better theoretical guarantees than more complex policies
- Simple rules that depend on a limited number of covariates may be more easily accepted by decision makers ('interpretable AI')

DML & restricted class of policies leads to good theoretical guarantees

Athey & Wager (2020) | 2

AW2019 propose two-step procedure

1) Estimate doubly robust policy score

$$\hat{\Gamma}_i = \hat{\mu}(1, x_i) - \hat{\mu}(0, x_i) + \frac{d_i - \hat{p}(x_i)}{\hat{p}(x_i)\left[1 - \hat{p}(x_i)\right]}\left[y_i - \hat{\mu}(d_i, x_i)\right]$$

2) Find best allocation by minimizing regret given the score

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[2\pi(x_i) - 1\right] \hat{\Gamma}_i \right\}$$

unicef | for every child

$$\hat{\Gamma}_i = \hat{\mu}(1, x_i) - \hat{\mu}(0, x_i) + \frac{d_i - \hat{p}(x_i)}{\hat{p}(x_i)\left[1 - \hat{p}(x_i)\right]}\left[y_i - \hat{\mu}(d_i, x_i)\right]$$

Athey & Wager (2020) | Step 1

Conditions

- The 3 consistent 1st stage estimators converge at least at rate $N^{1/4}$
  - Many ML procedures achieve this rate (RF, Lasso, some NN, etc.)
- Observation *i* must not be used in 1st stage → cross-fitting
  - May be implemented via k-fold cross-validation

This is double machine learning used for policy analysis

unicef | for every child

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \left\{ \frac{1}{N} \sum_{i=1}^{N} [2\pi(x_i) - 1] \hat{\Gamma}_i \right\}$$

Athey & Wager (2020) | Step 2

Use policy score to find best allocation in restricted class

- Not a convex problem ➔ may be computationally very demanding, depending on the policy class considered (& type & dim of X)
- However it has an interpretation as a weighted classification problem with sample weights λ_i & response variable H_i

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \left\{ \frac{1}{N} \sum_{i=1}^{N} \lambda_i H_i \left[2\pi(x_i) - 1 \right] \right\}, \quad \lambda_i = \left| \hat{\Gamma}_i \right|, \quad H_i = sign(\hat{\Gamma}_i)$$

- Thus, AW19 propose to train a (weighted) classifier by standard ML methods

Zhou-Athey-Wager (2022) | 1

Extension of AW19 to multiple treatments

1) Estimate doubly robust policy score (with cross-fitting)

$$\hat{\Gamma}_i = \frac{d_i\left[y_i - \hat{\mu}(d_i, x_i)\right]}{\hat{p}(x_i)} + \begin{bmatrix} \hat{\mu}(0, x_i) \\ \vdots \\ \hat{\mu}(M-1, x_i) \end{bmatrix}$$

2) Find best allocation by minimizing regret given the vector of scores

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\langle \hat{\Gamma}_i, \pi(x_i) \right\rangle \right\}$$

unicef | for every child

Zhou-Athey-Wager (2022) | Step 1

Conditions

- The 2 *M+1* consistent 1st stage estimators converge at least at rate $N^{1/4}$
  - Many ML procedures achieve this rate (RF, Lasso, etc.)
- Observation *i* must not be used in 1st stage → cross-fitting
  - May be implemented via k-fold cross-validation

This is double machine learning used for policy analysis

unicef | for every child

$$\hat{\Gamma}_i = \frac{d_i\left[y_i - \hat{\mu}(d_i, x_i)\right]}{\hat{p}(x_i)} + \begin{bmatrix} \hat{\mu}(0, x_i) \\ \vdots \\ \hat{\mu}(M-1, x_i) \end{bmatrix}$$

Zhou-Athey-Wager (2022) | Step 2

Use policy score to find best allocation in restricted class

- Not a convex problem ➔ may be computationally very demanding, depending on the policy class considered
- Interpretation as weighted classification problem seems not to work
- They provide 2 algorithms for decision trees
  - Exact solution via mixed integer programming
    - Very, very computational intensive
  - Approximate solutions via a limited number of trees
    - Computationally (**sort-of**) efficient, but does not allow for constraints
    - Cox, Lechner, Boolens (2019) modify their algorithm to allow for capacity constraints

unicef | for every child

Practical problems | 1

Overfitting

- Optimal policy too closely targeted to sample at hand → may not generalize
  - Split sample randomly → see whether same recommendations emerge in all sample splits (cross-validation type arguments)

Incorporating 'difficult' constraints

- In particular for trees, it seems difficult to introduce them without forgoing efficiency
  - See Cox, Lechner, Boolens (2022) for a proposal

unicef | for every child

Extensions & discussion points

Issues

- Explainability
- Selection of decision variables
- Fairness
- Computational efficiency
  - Policy trees are NP-hard

Extensions

- Large literature on dynamic treatments & continuous treatments

unicef | for every child

Background

Estimation of IATE and optimal policy: Some 'after-thoughts'

Decision making & effect estimation are different tasks (classification vs. regression)

- Thus, they have different objective functions & which should give better results for the task they are optimized for
- But, if we know potential outcomes (or estimate them without error), assigning treatments according to potential outcomes is optimal
- Biases in the policy scores (or consistent estimation of IATEs or potential outcomes) do not matter if these biases do not change the order of the potential outcomes
- Thus, in finite samples it *could* be that …
  - Longer data (with some missing confounders) may be preferable to wider data with all confounders (unclear however when this is true)
  - Using precisely estimated potential outcomes (like from the MCF or similar) could give better results than using the DML based policy scores → to be investigated further

Nice discussion in Fernández-Loría, Carlos, Foster Provost (2021): Causal Decision Making and Causal Effect Estimation Are Not the Same... and Why It Matters, arXiv:2104.04103, *INFORMS Journal of Data Science*

unicef | for every child

1 | Introduction

2 | Machine Learning vs Causal Machine Learning

3 | Double Debiased Machine Learning

4 | Causal Trees & Forests

5 | A comparison of Comprehensive Causal Machine Learners

6 | Optimal Policy & Algorithmic Decision Making

7 | Conclusions & outlook

unicef | for every child

Conclusions

Comprehensive causal machine learning has many advantages in applied work

- Workflow monitoring
- Internal consistency of causal effects obtained for different aggregation levels

The **m**odified **c**ausal **f**orest is robust estimator for all aggregation levels

- Python module available

The *mcf* Python module provides also methods for optimal policy / algorithmic decision making

unicef | for every child

unicef | for every child

Tomorrow morning
Application of these methods to the evaluation of training programmes in Flanders

Michael Lechner

Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen | Switzerland