Computational Data Analytics
for Economists

Lecture 4

Causal Machine Learning

Helge Liebert Anthony Strittmatter

Outline

1. Estimation Targets
2. Double Selection Procedure
3. Methods Adapting the Data
    - Modified Outcome Method
    - Modified Covariate Method
    - R-learning
4. Methods Adapting the ML Algorithm
    - Causal Tree and Forest
5. Performance Comparison
6. IV with many Instruments

Literature

- Athey (2018): ”The Impact of Machine Learning on Economics”, *The Economics of Artificial Intelligence: An Agenda*, editors: Agrawal, Gans, and Goldfarb, University of Chicago Press, download.
- Belloni, Chernozhukov, and Hansen (2014): ”High-Dimensional Methods and Inference on Structural and Treatment Effects”, Journal of Economic Perspectives, 28 (2), pp. 29-50, download.
- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017): ”Double/Debiased/Neyman Machine Learning of Treatment Effects”, American Economic Review, P&P, 107 (5), pp. 261-265, download.
- Tian, Alizadeh, Gentles, and Tibshirani (2014): ”A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates”, Journal of the American Statistical Association, 109 (508), pp. 1517-1532, download.
- Athey and Imbens (2016): ”Recursive Partitioning for Heterogeneous Causal Effects”, Proceedings of the National Academy of Science, 113 (27), pp. 7353-7360, download.

1. Estimation Targets

Notation:

- D_i binary treatment dummy (e.g., assignment to training program)
- $Y_i(1)$ potential outcome under treatment (e.g., earnings under participation in training)
- $Y_i(0)$ potential outcome under non-treatment (e.g., earnings under non-participation in training)

Infeasible parameter:

- Individual causal effect: $\delta_i = Y_i(1) - Y_i(0)$

Feasible parameters:

- Average Treatment Effect (ATE): $\delta = E[Y_i(1) - Y_i(0)] = E[\delta_i]$
- Average Treatment Effect on the Treated (ATET): $\rho = E[\delta_i | D_i = 1]$
- Local Average Treatment Effect (LATE): $\gamma = E[\delta_i | Compliers]$

Conditional Average Treatment Effect (CATE)

$$\delta(x) = E[Y_i(1) - Y_i(0)|X_i = x] = E[\delta_i|X_i = x]$$

- X_i exogenous pre-treatment covariates/features
- X_i includes not only confounders but also other covariates which are potentially responsible for effect heterogeneity
- CATEs are often called individualised or personalised treatment effects
- CATEs can differ from CATET, $\rho(x)$, and CLATE, $\gamma(x)$

Aggregation of CATEs

- ATEs:

$$\delta = E[E[Y_i(1) - Y_i(0)|X_i = x]] = E[\delta(x)]$$

- Group Average Treatment Effects (GATEs):

$$\delta(g) = E[\delta(x)|G_i = g]$$

  where the groups g can be defined based on exogenous or endogenous variables

- Examples of GATEs:
  - Aggregate by gender: $\delta(m) = E[\delta(x)|Male]$ and $\delta(f) = E[\delta(x)|Female]$
  - Aggregate by earnings-quantile-range $[y_{floor}(\tau), y_{ceil}(\tau)]$:

$$\delta(\tau) = E[\delta(x)|y_{floor}(\tau) \leq Y_i < y_{ceil}(\tau)]$$

Identifying Assumptions for ATE and CATE

- **Stable Unit Treatment Value Assumption (SUTVA):**

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

- **Exogeneity of Covariates:**

$$X_i(1) = X_i(0)$$

- **No Support Problems:**

$$\varepsilon < Pr(D_i = 1|X_i = x) = p(x) < 1 - \varepsilon$$

for some small $\varepsilon \geq 0$ and all x in the support of X_i

- **Conditional Independence Assumption (CIA):**

$$Y_i(1), Y_i(0) \perp\!\!\!\perp D_i|X_i = x$$

for all x in the support of X_i

What are the Potential Advantages of ML for Causal Inference?

- **Average effects (low dimensional):**
  - ML enables us to incorporate (very) many covariates which can make the exclusion restriction more credible
  - Some ML approaches make little functional form assumptions
  - Can improve out-of-sample predictions of nuisance parameters which we use as input-factor (e.g., first stage of IV)

- **Heterogeneous effects (functionals):**
  - We can predict CATEs (potentially) based on (very) many covariates
  - Principled approach makes it less likely to overlook important heterogeneity

Limits of Causal Machine Learning (CML)

- CML can estimate causal effects (if at all) only for a few variables
- We will not obtain the (complete) structural model automatically
- CML will not select the relevant causal parameters automatically
- ML cannot distinguish between causation and correlation
  $\rightarrow$ we have to provide some structure to the CML algorithm
- We should resist the temptation to interpret prediction models
- Identifying assumptions do not change, no matter if we use ML methods or not

Application of CML Methods

- Davis and Heller (2017) investigate the heterogeneous effects of summer jobs on the probability to commit a violent crime (experimental study)
- Ascarza (2018) investigate the heterogeneous effects of contacting customers pro-actively on the churn probability (experimental study)
- Taddy, Gardner, Chen, and Draper (2016) investigate the heterogeneous effects of A/B-experiments in online-auctions (EBay) on customer responses (experimental study)
- Bertrand, Crépon, Marguerie, and Premand (2017) estimate the heterogeneous effects of a work experience program in Côte d'Ivoire on post-participation employment and wages (experimental study)
- Knaus, Lechner, and Strittmatter (2017) estimate the heterogeneous employment effects of a job search program in Switzerland (observational study)
- Knaus (2018) estimates the effects of musical practice on student's skills and selects confounders with ML methods (observational study)

2. Double Selection Procedure

- **Partial Linear Model:**

$$Y_i = D_i\delta + g(Z_i) + U_i$$
$$D_i = m(Z_i) + V_i$$

with $E[U_i|D_i, Z_i] = 0$ and $E[V_i|Z_i] = 0$

- **Approximation with Linear Model:**

$$Y_i = D_i\delta + X_i\beta_g + r_{gi} + U_i$$
$$D_i = X_i\beta_m + r_{mi} + V_i$$

with $X_i = p(Z_i)$

- Columns p of X_i can be much larger than sample size N ($p \gg N$)
- r_{gi} and r_{mi} are approximation errors of functions $g(\cdot)$ and $m(\cdot)$, respectively

Reference: Belloni, Chernozhukov, and Hansen (2014)

Types of Covariates

Relation between covariates and outcome (for some $s_g > 0$):

- $|\beta_{gj}| > s_g$: covariate X_j has a **strong association** with Y_i
- $0 < |\beta_{gj}| \leq s_g$: covariate X_j has a **weak association** with Y_i
- $\beta_{gj} = 0$: covariate X_j has a **no association** with Y_i

Relation between covariates and treatment (for some $s_m > 0$):

- $|\beta_{mj}| > s_m$: covariate X_j has a **strong association** with D_i
- $0 < |\beta_{mj}| \leq s_m$: covariate X_j has a **weak association** with D_i
- $\beta_{mj} = 0$: covariate X_j has a **no association** with D_i

$\rightarrow$ Approximate sparsity means (roughly) that covariates with $|\beta_{gj}| \leq s_g$ and $|\beta_{mj}| \leq s_m$ are not important

Types of Covariates (cont.)

	$\beta_{gj} = 0$	$0 < \lvert\beta_{gj}\rvert \leq s_g$	$\lvert\beta_{gj}\rvert > s_g$
$\beta_{mj} = 0$	Irrelevant	Irrelevant	Irrelevant
$0 < \lvert\beta_{mj}\rvert \leq s_m$	Irrelevant	Approx. Sparsity	Weak Confounder
$\lvert\beta_{mj}\rvert > s_m$	Irrelevant	Weak Confounder	Strong Confounder

- $|\beta_{gj}| > s_g$ and $0 < |\beta_{mj}| \leq s_m$: "Weak Outcome Confounder"
- $|\beta_{mj}| > s_m$ and $0 < |\beta_{gj}| \leq s_g$: "Weak Treatment Confounder"

Naive Approaches

- Apply LASSO to structural model

$$\min_{\beta_g} E[(Y_i - D_i\delta - X_i\beta_g)^2] + \lambda \|\beta_g\|_1$$

  without a penalty on δ

  - Covariates that are highly correlated with D_i are probably not selected, even though this can be ”strong confounders”
  - ”Weak treatment confounders” are less likely selected

- Apply LASSO to selection model

$$\min_{\beta_m} E[(D_i - X_i\beta_m)^2] + \lambda \|\beta_m\|_1$$

  - ”Weak outcome confounders” are less likely selected

Double Selection Procedure

1. Apply LASSO to the reduced form models

$$\min_{\tilde{\beta}_g} E[(Y_i - X_i\tilde{\beta}_g)^2] + \lambda \|\tilde{\beta}_g\|_1 \tag{1}$$

$$\min_{\beta_m} E[(D_i - X_i\beta_m)^2] + \lambda \|\beta_m\|_1 \tag{2}$$

with $\tilde{\beta}_g = \delta\beta_m + \beta_g$

   - "Strong confounders" and "weak treatment confounders" are likely selected in (2)
   - $\tilde{\beta}_{gj} \approx \beta_g$ when $\beta_{gj} \approx 0$, such that "weak outcome confounders" are likely selected in (1)
   - Possibly, we additionally select less important variables in (1)

2. Take the union of all covariates $\tilde{X}_i$ with estimated LASSO coefficients of either $\hat{\beta}_{gj} \neq 0$ **or** $\hat{\beta}_{mj} \neq 0$ and estimate the OLS model

$$Y_i = D_i\delta + \tilde{X}_i\beta_g^* + u_i$$

Asymptotic Results

(Main) regularity conditions:

- Approximate sparsity
- Sparse eigenvalues $\rightarrow$ restriction on the correlation structure between covariates

Asymptotic results of the double selection procedure:

- Asymptotic normality

$$\sqrt{N}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \sigma)$$

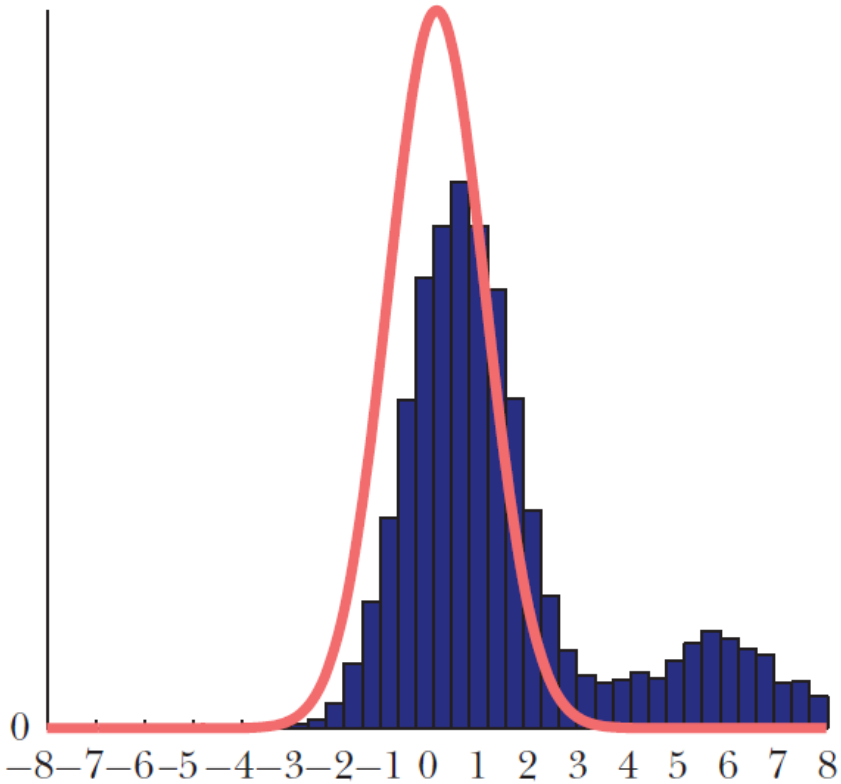
- Model selection step is asymptotically negligible for building confidence intervals
- Optimal penalty parameter $\lambda^* = 2c \cdot \Phi^{-1}(1 - \gamma/2p)/\sqrt{N}$ (e.g., $c = 1.1$ and $\gamma \leq 0.05$) for "Feasible LASSO"

$$\min_{\beta} E[(Y_i - X_i\beta)^2] + \lambda^* \|\beta\|_1$$

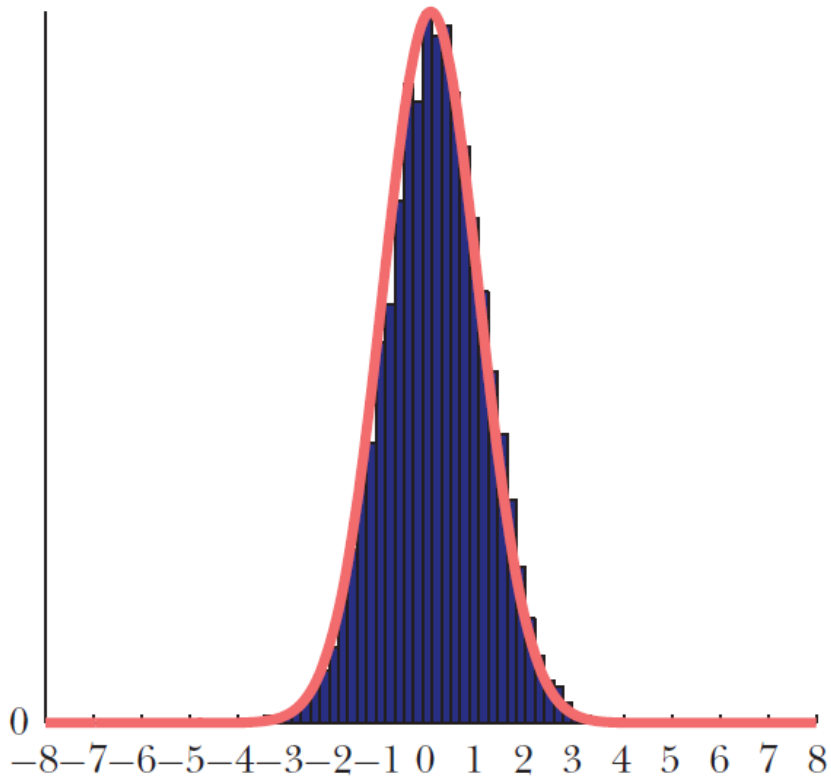
Simulation Exercise

Distribution of Estimators

Naive Single-Post-Selection on Structural Model



Double-Post-Selection



Source: Belloni, Chernozhukov, and Hansen (2014)

Example: Effect of Abortion on Crime

	Crime Type					
	Violent		Property		Murder	
	Effect	Std. err.	Effect	Std. err.	Effect	Std. err.
Donohue and Levitt (2001)	-.157***	0.034	-.106***	0.021	-.218***	0.068
284 controls	0.071	0.284	-.161	0.106	-1.327	0.932
Double-selection	-.171	0.117	-.061	0.057	-.189	0.177

Source: Belloni, Chernozhukov, and Hansen (2014), $N = 600$

Summary Double Selection Procedure

Advantages:

- Asymptotic results available
- Standard inference
- Computationally fast

Disadvantages:

- Effect homogeneity
- Restrictive assumptions required
- Potentially too many covariates selected

Some General Thoughts

- Partial Linear Model:

$$Y_i = D_i\delta + g(X_i) + U_i \text{ and } D_i = m(X_i) + V_i$$

- Split sample in partitions S and S^c with sample sizes $n = N/2$
- Use ML to estimate $\hat{g}(X_i)$ in sample S^c
- Estimate $\hat{\delta}$ in sample S

$$\hat{\delta} = \left(\frac{1}{n}\sum_{i\in S} D_i^2\right)^{-1} \frac{1}{n}\sum_{i\in S} D_i(Y_i - \hat{g}(X_i))$$

- Regularisation bias

$$\sqrt{n}(\hat{\delta} - \delta) = \left(E[D_i^2]\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in S} D_i\left(U_i + (g(X_i) - \hat{g}(X_i))\right)$$

- $\hat{g}(X_i)$ converges to $g(X_i)$ at rate $n^{-\varphi_d}$, with $\varphi_d < 1/2$ for ML methods
- $\hat{\delta}$ has a convergence rate below $\sqrt{n}$: $|\sqrt{n}(\hat{\delta} - \delta)| \xrightarrow{p} \infty$

Some General Thoughts (cont.)

- Orthogonalised regressor: $\hat{V}_i = D_i - \hat{m}(X_i)$
- Estimate $\hat{\delta}$ in sample S

$$\hat{\delta} = \left(\frac{1}{n} \sum_{i \in S} \hat{V}_i D_i \right)^{-1} \frac{1}{n} \sum_{i \in S} \hat{V}_i (Y_i - \hat{g}(X_i))$$

- Estimation error

$$\sqrt{n}(\hat{\delta} - \delta) = \left(E[V_i^2] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in S} (V_i U_i + (m(X_i) - \hat{m}(X_i))(g(X_i) - \hat{g}(X_i))) + c^*$$

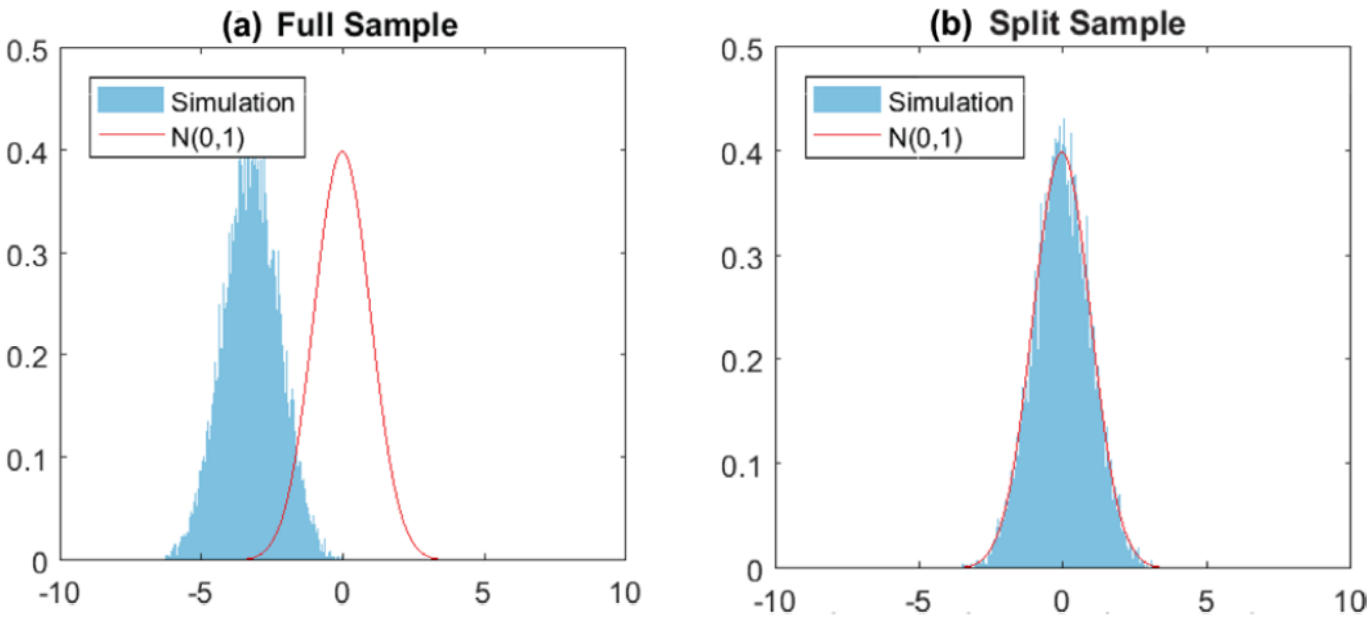
- $\hat{m}(X_i)$ converges to $m(X_i)$ at rate $n^{-\varphi_m}$
- The regularisation bias will vanish at $\sqrt{n}$-rate when $\varphi_g + \varphi_m \geq 1/2$
- Double-robustness property

Role of Sample Splitting

- Remainder term:

$$c^* = \left(E[V_i^2]\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in S} \left(U_i(m(X_i) - \hat{m}(X_i)) + V_i(g(X_i) - \hat{g}(X_i))\right)$$

- Vanishes because of sample splitting



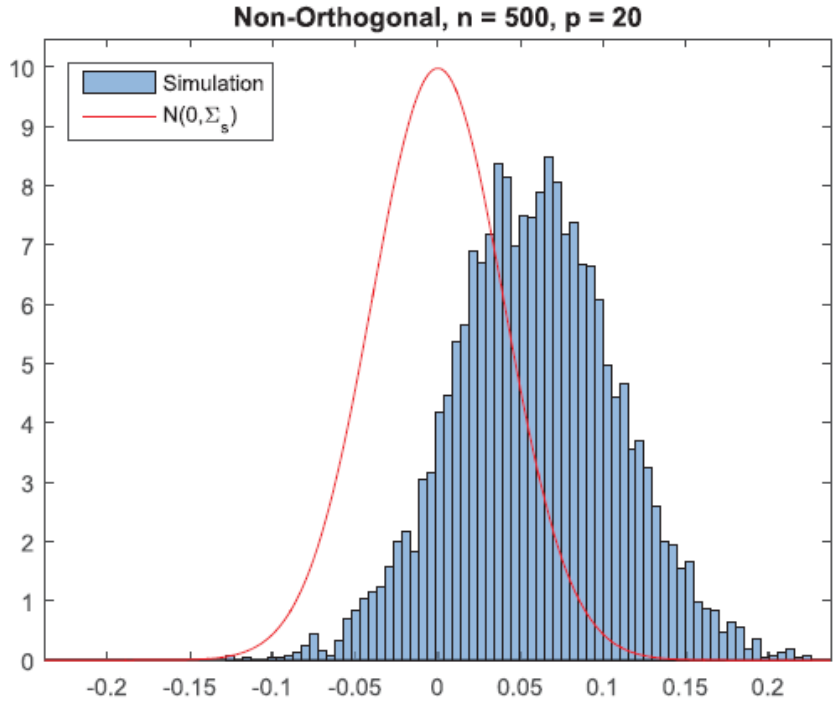
- Loss of efficiency because of sample splitting $\rightarrow$ cross-fitting

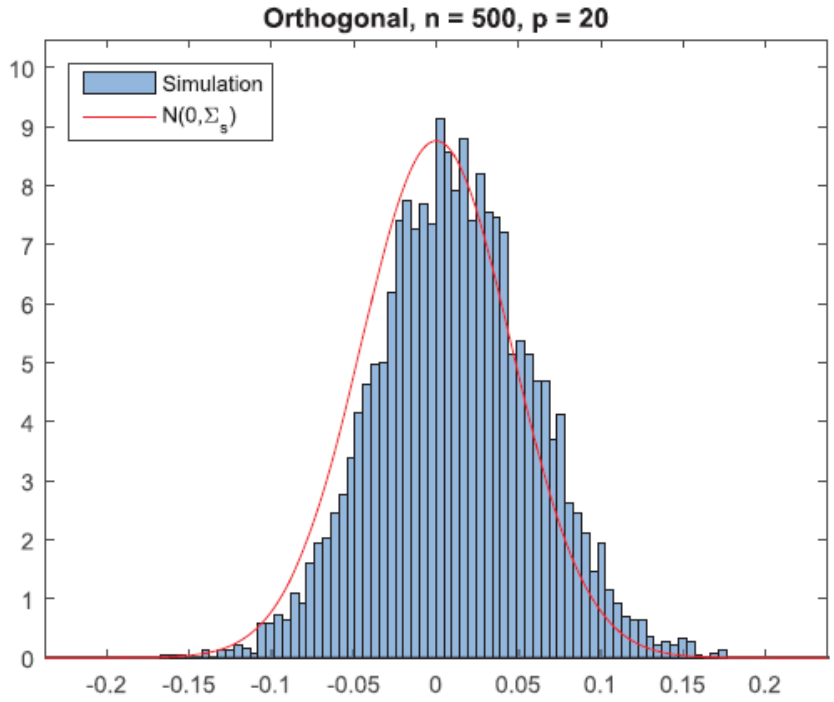
Source: Chernozhukov et al. (2018)

Neyman-Orthogonality

- **General Condition:**
  - Moment Condition:
$$\frac{1}{n}\sum_{i\in S}\psi(W;\hat{\delta}_0,\hat{\eta}_0)=0$$
  - Gateaux derivative:
$$\partial_\eta E[\psi(W;\delta_0,\eta_0][\eta-\eta_0]=0$$
- **Example OLS without orthogonalisation:**
  - Score: $\psi_i = D_i(Y_i - D_i\hat{\delta} - X_i\hat{\beta}_g)$
  - Jacobian: $-E[D_iX_i] \neq 0$
- **Example OLS with orthogonalisation:**
  - Score: $\psi_i = \hat{V}_i(Y_i - D_i\hat{\delta} - X_i\hat{\beta}_g)$
  - Jacobian: $-E[\hat{V}_iX_i] = 0$

Simulation Exercise





Lessons learned from general thoughts:

- → Sample splitting is important
- → Orthogonalisation is important

Source: Chernozhukov et al. (2018)

3. Methods Adapting the Data

Inverse Probability Weighting (IPW):

$$
\begin{aligned}
\delta(x) &= E\left[Y_i(1)|X_i = x\right] - E\left[Y_i(0)|X_i = x\right] \\
&\overset{CIA}{=} E\left[Y_i(1)|D_i = 1, X_i = x\right] - E\left[Y_i(0)|D_i = 0, X_i = x\right] \\
&= E\left[Y_i|D_i = 1, X_i = x\right] - E\left[Y_i|D_i = 0, X_i = x\right] \\
&\overset{Bayes}{=} E\left[\frac{D_i Y_i}{p(x)}\middle| X_i = x\right] - E\left[\frac{(1-D_i)Y_i}{1-p(x)}\middle| X_i = x\right] \\
&= E\left[\frac{D_i Y_i}{p(x)} - \frac{(1-D_i)Y_i}{1-p(x)}\middle| X_i = x\right] \\
&= E\left[\frac{D_i - p(x)}{p(x)(1-p(x))} Y_i\middle| X_i = x\right]
\end{aligned}
$$

with $p(x) = Pr(D_i = 1|X_i = x)$

Reference: Horvitz and Thompson (1952)

Modified Outcome Method

- $Y^*_{i,IPW} = W_i Y_i$ with $W_i = (D_i - p(x))/(p(x)(1-p(x)))$
- $Y^*_{i,IPW}$ is a crude approximation of the causal effect, because $\delta(x) = E[Y^*_{i,IPW}|X_i = x]$ and $\delta = E[Y^*_{i,IPW}]$
- We can use standard ML methods to estimate $\hat{\delta}(x)$ and/or $\hat{p}(x)$ (possibly in different samples using cross-fitting)
- **Advantages:**
  - Generic approach
- **Disadvantages:**
  - Potentially omitting ”weak outcome confounders” (sparsity assumption on selection equation)
  - Shows weak performance in simulations and applications
  - Moments are not Neyman-orthogonal

Orthogonal Score

$$
\begin{aligned}
\delta(x) =& E\left[\mu_1(x) - \mu_0(x) + \frac{D_i(Y_i - \mu_1(x))}{p(x)} - \frac{(1-D_i)(Y_i - \mu_0(x))}{1-p(x)} \middle| X_i = x\right] \\
=& E\left[\frac{D_i - p(x)}{p(x)(1-p(x))} Y_i + \frac{(D_i - p(x))\mu_1(x)}{p(x)} - \frac{(D_i - p(x))\mu_0(x)}{1-p(x)} \middle| X_i = x\right] \\
=& E\left[\frac{D_i - p(x)}{p(x)(1-p(x))} Y_i \middle| X_i = x\right] + \frac{E[D_i - p(x)|X_i = x]}{p(x)}\mu_1(x) \\
& - \frac{E[D_i - p(x)|X_i = x]}{1-p(x)}\mu_0(x) \\
=& E\left[\frac{D_i - p(x)}{p(x)(1-p(x))} Y_i \middle| X_i = x\right] = E\left[Y_i(1)|X_i = x\right] - E\left[Y_i(0)|X_i = x\right]
\end{aligned}
$$

with $\mu_1 = E[Y_i(1)|X_i = x]$ and $\mu_0 = E[Y_i(0)|X_i = x]$

Reference: Robins and Rotnitzki (1995)

Double/Debiased Machine Learning (DML)

- $Y^*_{i,DML} = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1-D_i)(Y_i - \mu_0(X_i))}{1-p(X_i)}$
- We can use standard ML methods to estimate $\hat{\delta}(x)$, $\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$, and $\hat{p}(x)$ (possibly in different samples using cross-fitting)
- **Advantages:**
  - Generic approach
  - Neyman orthogonality
  - Double robustness properties
  - We know asymptotic properties for ATE, ATET, and LATE
  - More robust than IPW when $p(x)$ is close to zero or one
- **Disadvantages:**
  - Asymptotic results for CATEs under investigation (e.g., Lee, Okui, and Whang (2017))

DML Cross-Fitting Algorithm

1. Split data in samples S^A and S^B
2. Estimate the nuisance parameters $\mu_1^A(x)$,$\mu_0^A(x)$, and $p^A(x)$ in S^A; and $\mu_1^B(x)$,$\mu_0^V(x)$, and $p^B(x)$ in S^B with ML
3. Construct the efficient scores

$$Y_{i,DML}^{A*} = \mu_1^B(X_i) - \mu_0^B(X_i) + \frac{D_i(Y_i - \mu_1^B(X_i))}{p^B(X_i)} - \frac{(1-D_i)(Y_i - \mu_0^B(X_i))}{1-p^B(X_i)}$$

$$Y_{i,DML}^{B*} = \mu_1^A(X_i) - \mu_0^A(X_i) + \frac{D_i(Y_i - \mu_1^A(X_i))}{p^A(X_i)} - \frac{(1-D_i)(Y_i - \mu_0^A(X_i))}{1-p^A(X_i)}$$

4. Calculate ATE,

$$\hat{\delta} = \frac{1}{2}\left\{\hat{E}[Y_{i,DML}^{A*}] + \hat{E}[Y_{i,DML}^{B*}]\right\},$$

or use another ML estimator to estimate CATEs,

$$\hat{\delta}(x) = \frac{1}{2}\left\{\hat{E}[Y_{i,DML}^{A*}|X_i = x] + \hat{E}[Y_{i,DML}^{B*}|X_i = x]\right\},$$

Asymptotic Results for ATE

- ATE (and other group averages) can be estimated $\sqrt{N}$-consistently

$$\sqrt{N}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \sigma)$$

with $\sigma^2 = Var(Y^*_{i,DML})$

- Split sample estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{2}\left(\hat{\sigma}_A^2 + (\hat{\delta}_A - \hat{\delta})^2\right) + \frac{1}{2}\left(\hat{\sigma}_B^2 + (\hat{\delta}_B - \hat{\delta})^2\right)$$

for $\hat{\delta} = 1/2(\hat{\delta}_A + \hat{\delta}_B)$

ATET

- Selection on Observables (Chernozhukov et al., 2018):

$$\rho = E\left[\frac{D_i(Y_i - \mu_0(x))}{p} - \frac{p(x)(1-D_i)(Y_i - \mu_0(x))}{p(1-p(x))}\right]$$

with $p = Pr(D_i = 1)$

- Difference-in-Differences (Zimmert, 2018):

$$\rho = E\left[\frac{T - p_t}{p_t(1-p_t)}\frac{D_i - p(x)}{p(1-p(x))}(Y_i - \theta_0(x,t))\right]$$

with $\theta_0(x,t) = E[Y_i|D = 0, T = t, X = x]$ and $p_t = Pr(T = 1)$

LATE

- LATE for binary instrument $Z_i \in \{0,1\}$ (Chernozhukov et al., 2018):
  - First Stage:

$$\gamma_F = E\left[\nu_1(x) - \nu_0(x) + \frac{Z_i(D_i - \nu_1(x))}{e(x)} - \frac{(1-Z_i)(D_i - \nu_0(x))}{1-e(x)}\right]$$

  - Second Stage:

$$\gamma_S = E\left[\omega_1(x) - \omega_0(x) + \frac{Z_i(Y_i - \omega_1(x))}{e(x)} - \frac{(1-Z_i)(Y_i - \omega_0(x))}{1-e(x)}\right]$$

with $e(x) = Pr(Z_i = 1|X_i = x)$, $\nu_1(x) = E[D_i|Z_i = 1, X_i = x]$, $\nu_0(x) = E[D_i|Z_i = 0, X_i = x]$, $\omega_1(x) = E[Y_i|Z_i = 1, X_i = x]$, and $\omega_0(x) = E[Y_i|Z_i = 0, X_i = x]$

$\rightarrow$ Apply Wald-estimator $\gamma = \gamma_S/\gamma_F$

Other Efficient Scores

- Multiple treatments $d \in \{1,2,3,\ldots,\}$ (e.g., Farrell, 2015) :

$$E[Y(d)] = E\left[\frac{1\{D_i = d\}(Y_i - \hat{\mu}_d(x))}{Pr(D_i = d|X_i = x)} + \hat{\mu}_d(x)\right]$$

- Continuous treatments see, e.g., Graham and Pinto (2018)
- Mediation analysis see Tchetgen Tchetgen and Shpitser (2012)
- Synthetic control group method see, e.g., Arkhangelsky et al. (2018)

A Simple Model for Effect Heterogeneity

- 50% randomly assigned to treatment and 50% to control group
- Fully interacted model,

$$Y_i = X_i\beta + D_i X_i \delta_v + \varepsilon_i$$

with $E[Y_i(0)|X_i = x] = x\beta$ and $\delta(x) = E[Y_i(1) - Y_i(0)|X_i = x] = x\delta_v$
(δ_v is a vector of coefficients)

- Transformation of participation dummy, $T_i = 2D_i - 1$, such that $T_i/2 \in \{-0.5, 0.5\}$

$$Y_i = X_i\alpha + \frac{T_i}{2} X_i \delta_v + u_i$$

with $E[Y_i|X_i = x] = x\alpha$ and $\delta(x) = E[Y_i(1) - Y_i(0)|X_i = x] = x\delta_v$

- $Cov(X_{ij}, T_i X_{ik}) = Cov(X_{ij}, X_{ik})E[T_i] = 0$ for all $j,k \in \{1,...,p\}$
- $Cov(X_{ij}, D_i X_{ik}) = Cov(X_{ij}, X_{ik})E[D_i] \geq 0$ for all $j,k \in \{1,...,p\}$

Modified Covariate Method

50% Randomisation:

$$\min_{\delta} E\left[\left(Y_i - \frac{T_i}{2} X_i \delta_v\right)^2\right]$$

Generalisation to selection on observables:

$$\min_{\delta} E\left[T_i \frac{D_i - p(X_i)}{p(X_i)(1-p(X_i))}\left(Y_i - \frac{T_i}{2}\delta(X_i)\right)^2\right]$$

With efficiency augmentation (EA):

$$\min_{\delta} E\left[T_i \frac{D_i - p(X_i)}{p(X_i)(1-p(X_i))}\left(Y_i - \mu(X_i) - \frac{T_i}{2}\delta(X_i)\right)^2\right]$$

for $\mu(x) = E[Y_i | X_i = x]$

Reference: Chen, Tian, Cai, and Yu (2017)

Advantages and Disadvantages

Advantages:

- Explicitly estimates CATEs
- Efficiency augmentation available (orthogonalisation)
- Oracle properties for LASSO are known
- Suitable for non-linear models

Disadvantages:

- Asymptotic properties unknown
- Currently only used for linear model $\delta(X_i) = X_i \delta_v$

Algorithm for Linear MCM with EA

1. Split data in samples S^A and S^B
2. Estimate the nuisance parameters $\mu^A(x)$ and $p^A(x)$ in S^A; and $\mu^B(x)$ and $p^B(x)$ in S^B with ML
3. Estimate the LASSO (or other linear) model

$$\min_{\delta_v^A} \sum_{i \in S^A} T_i \frac{D_i - \hat{p}^B(X_i)}{\hat{p}^B(X_i)(1-\hat{p}^B(X_i))} \left(Y_i - \hat{\mu}^B(X_i) - \frac{T_i}{2} X_i \hat{\delta}_v^A \right)^2 + \lambda_A \| \hat{\delta}_v^A \|_1$$

   and correspondingly for $\hat{\delta}_v^B$

4. The the fitted values are the CATEs $\hat{\delta}(X_i) = \hat{Y}_i$ and the ATEs are $\hat{\delta} = E[\hat{Y}_i]$

Non-Linear Models

- 50% randomly assigned to treatment and 50% to control group
- Define $X_i^* = \frac{T_i}{2} X_i$
- **Binary response model:**
  - Logistic log-likelihood function:

$$L = \sum_{i=1}^{N} (Y_i X_i^* \delta_v - log(1 + exp(X_i^* \delta_v)))$$

  - Logit-LASSO:

$$\min_{\delta_v} \{-L + \lambda \| \delta_v \|_1\}$$

  - Chen, Tian, Cai, and Yu (2017) propose extensions for observational studies (using similar weights as before)

- Additional extension for survival and Poisson models are available (see Tian, Alizadeh, Gentles, and Tibshirani, 2014)

Application: Swiss Job Search Program

- Content: Learn how to search and apply for a job
- Goal: Improve matching process
- Duration approximately 3 weeks
- Class room training
- Private providers
- Participants should continue active job search during the programme
- Yearly expenditures approximately 100 million CHF

Reference: Knaus, Lechner, and Strittmatter (2017)

Linked Unemployed-Caseworker Data

- Combination of social security, caseworker questionnaire, and regional data
- All registered unemployed in the year 2003 (12,000 participants and 72,000 controls)
- Outcome: Months employed after participation begins
- Treatment: First participation in a job search programme during the first six months of unemployment
- Caseworkers can assign unemployed persons to job search programmes (mostly) based on subjective measures

$\Rightarrow$ Non-random selection into participation

Rich Controls

- Unemployed: residence status, qualification, gender, education, language skills, employment history, profession, job position, industry of last job, desired occupation and industry, subjective employability rate by their caseworker, etc.
- Caseworker: age, gender, tenure, education, employment history, cooperativeness, etc.
- Region: language, population size of municipalities, the cantonal unemployment rate, etc.

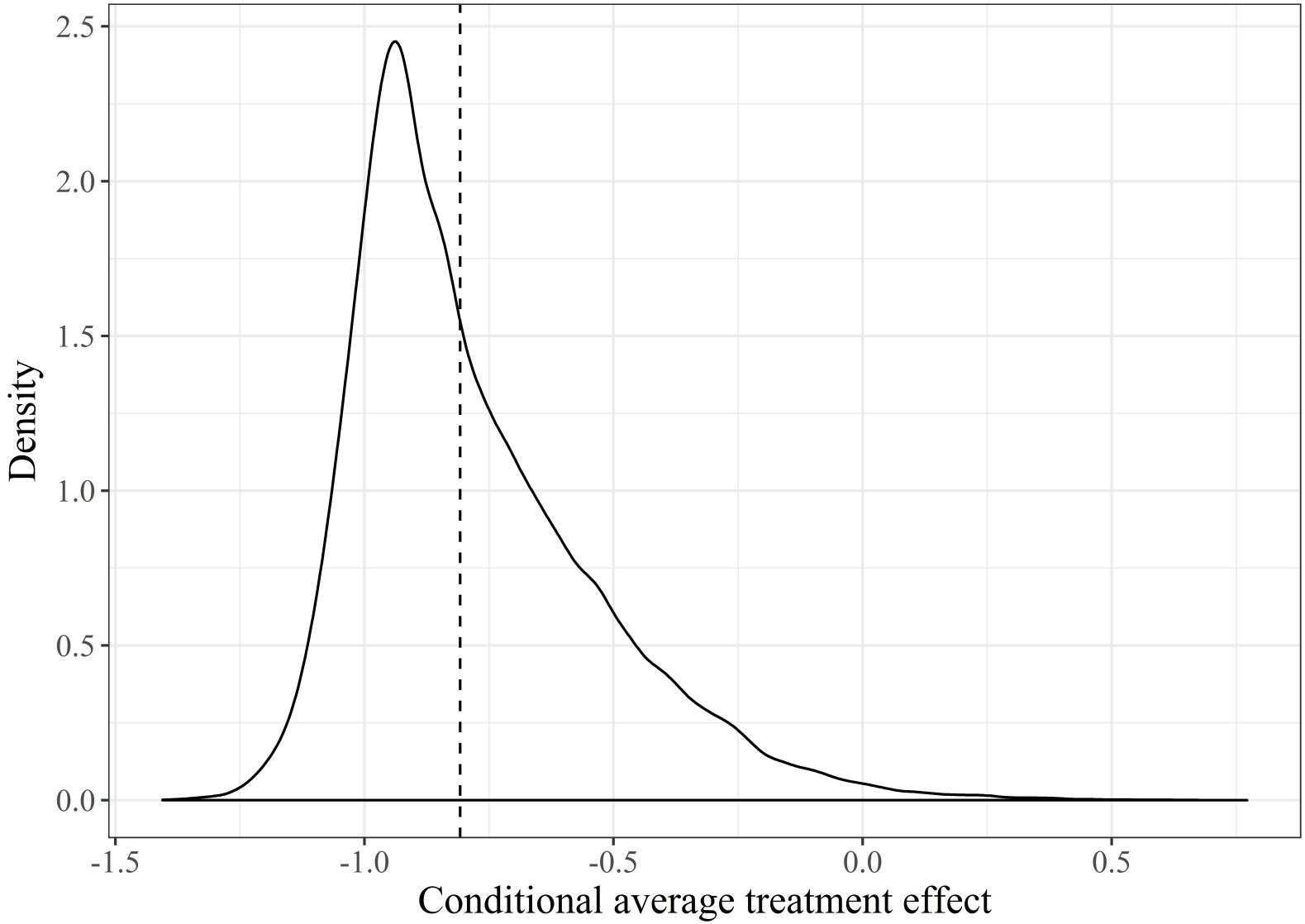
$\rightarrow$ Approximately 120 covariates and additionally interactions and polynomials

Aggregated Average Effects

Months employed since start of participation	ATE		ATET	
	Coef.	S.E.	Coef.	S.E.
	(1)		(2)	
During first 6 months	-0.80***	(0.02)	-0.82***	(0.02)
During first 12 months	-1.10***	(0.05)	-1.13***	(0.04)
During first 31 months	-1.14***	(0.14)	-1.20***	(0.13)
During months 25-31	-0.007	(0.03)	-0.011	(0.03)

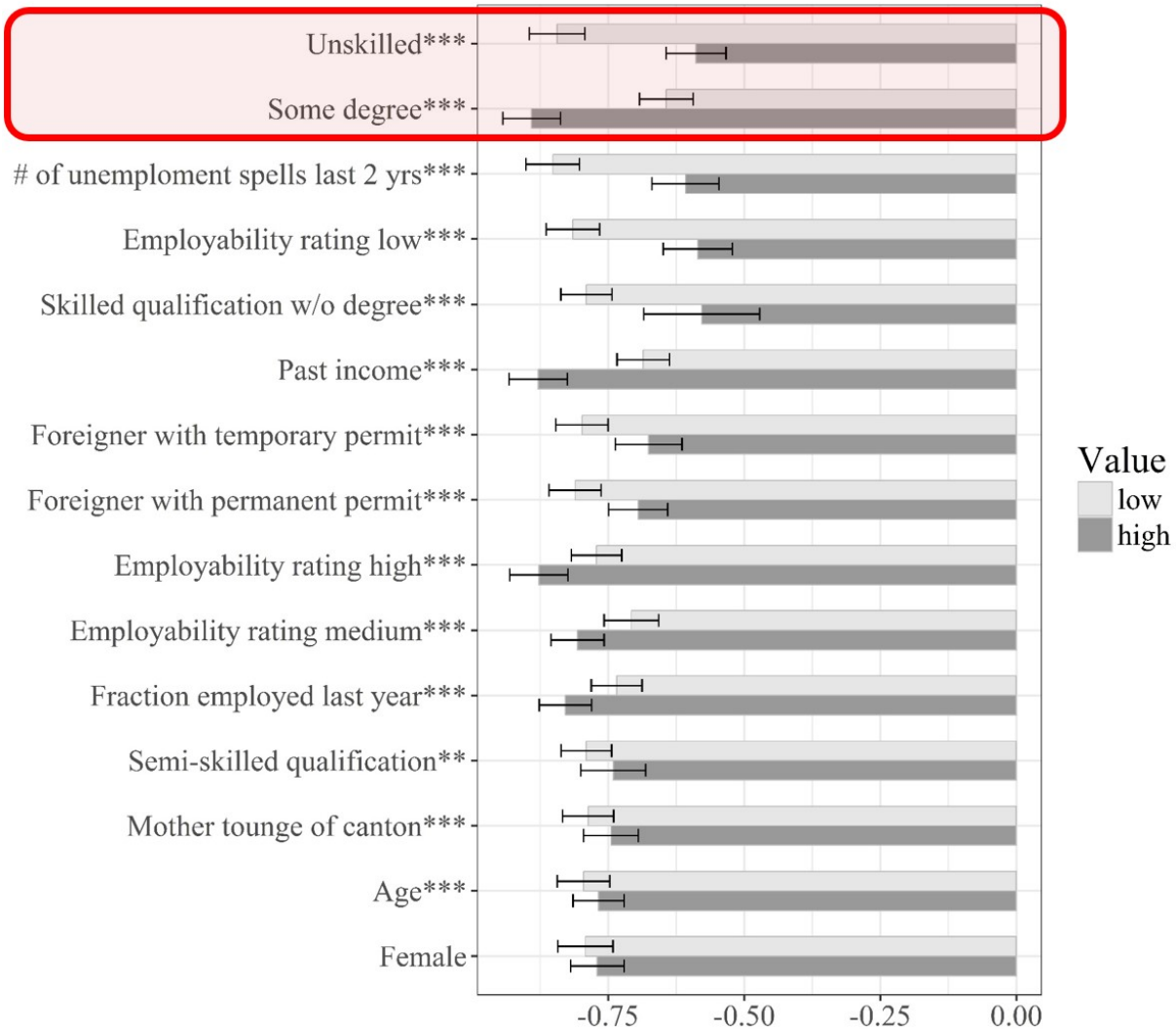
We obtain standard errors (S.E.) from a clustered bootstrap at caseworker level with 4,999 replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

Distribution of Predicted CATEs



Kernel smoothed distribution of average predicted individual effects. Gaussian kernel with bandwidth 0.02, chosen by Silverman's rule-of-thumb. The dashed vertical line shows the ATE.

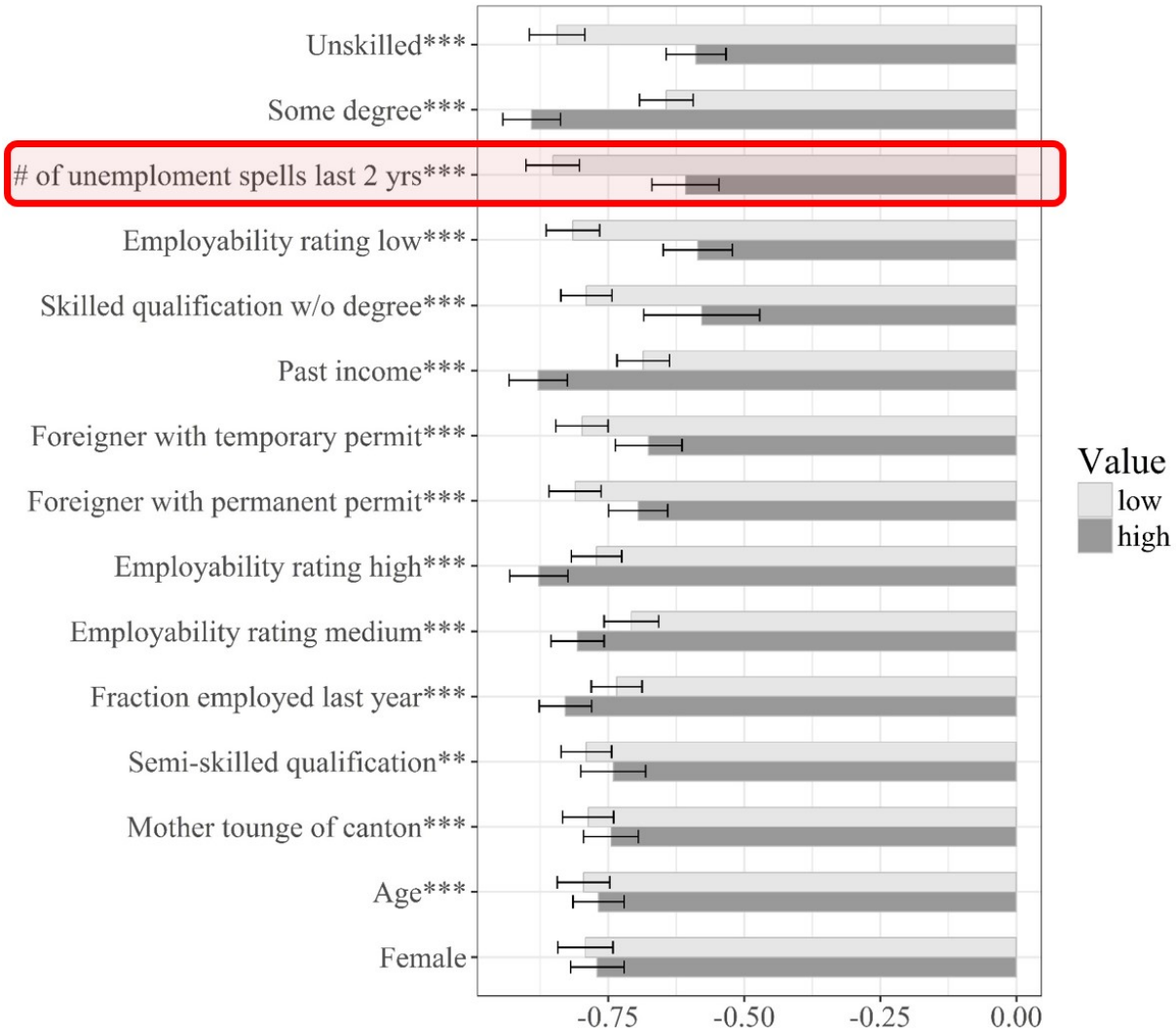
Characteristics of Unemployed Persons



Large heterogeneity by education:
Highly educated suffer much more

CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

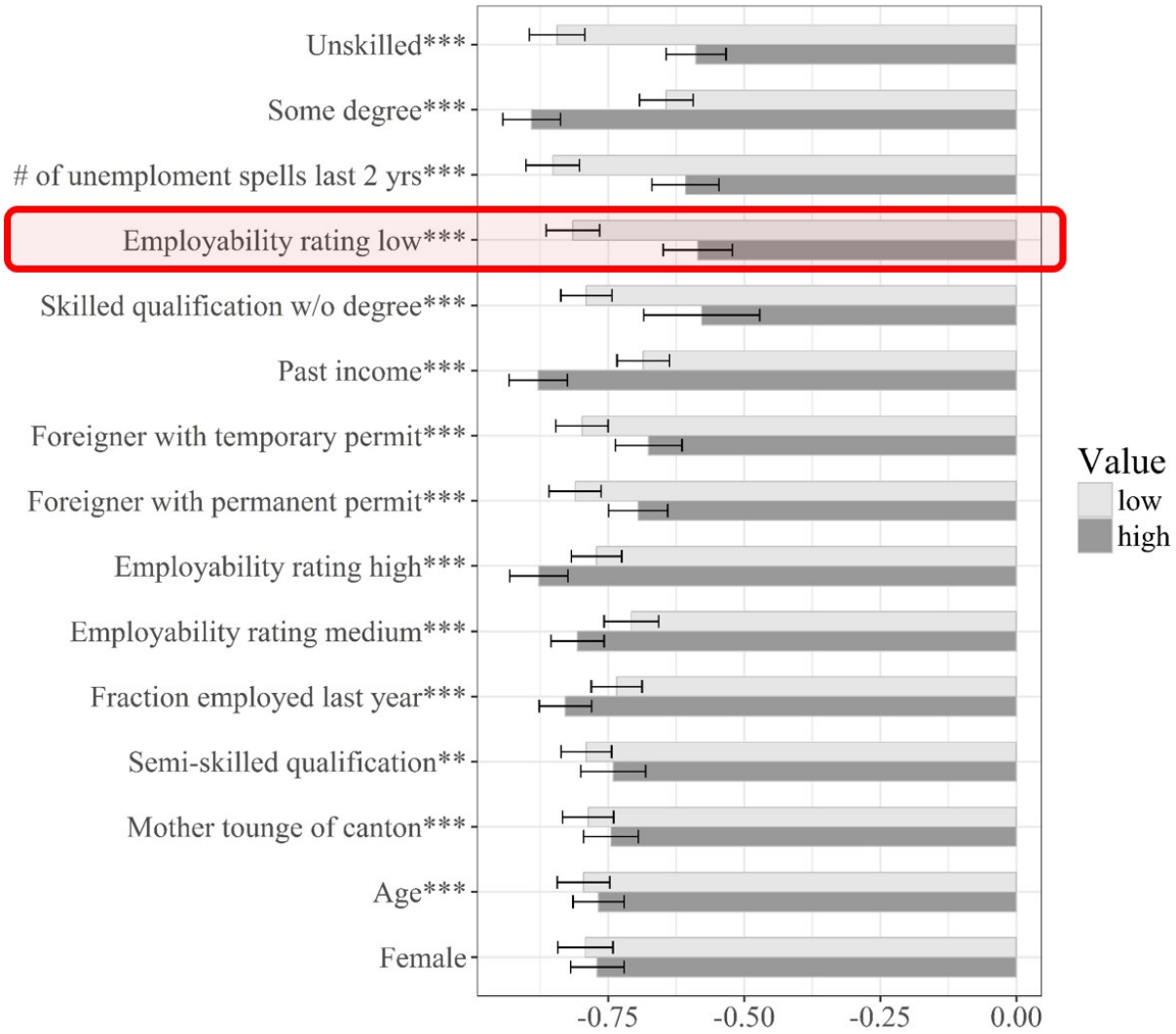
Characteristics of Unemployed Persons



Large heterogeneity by previous labor market success:
Never unemployed suffer much more

CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

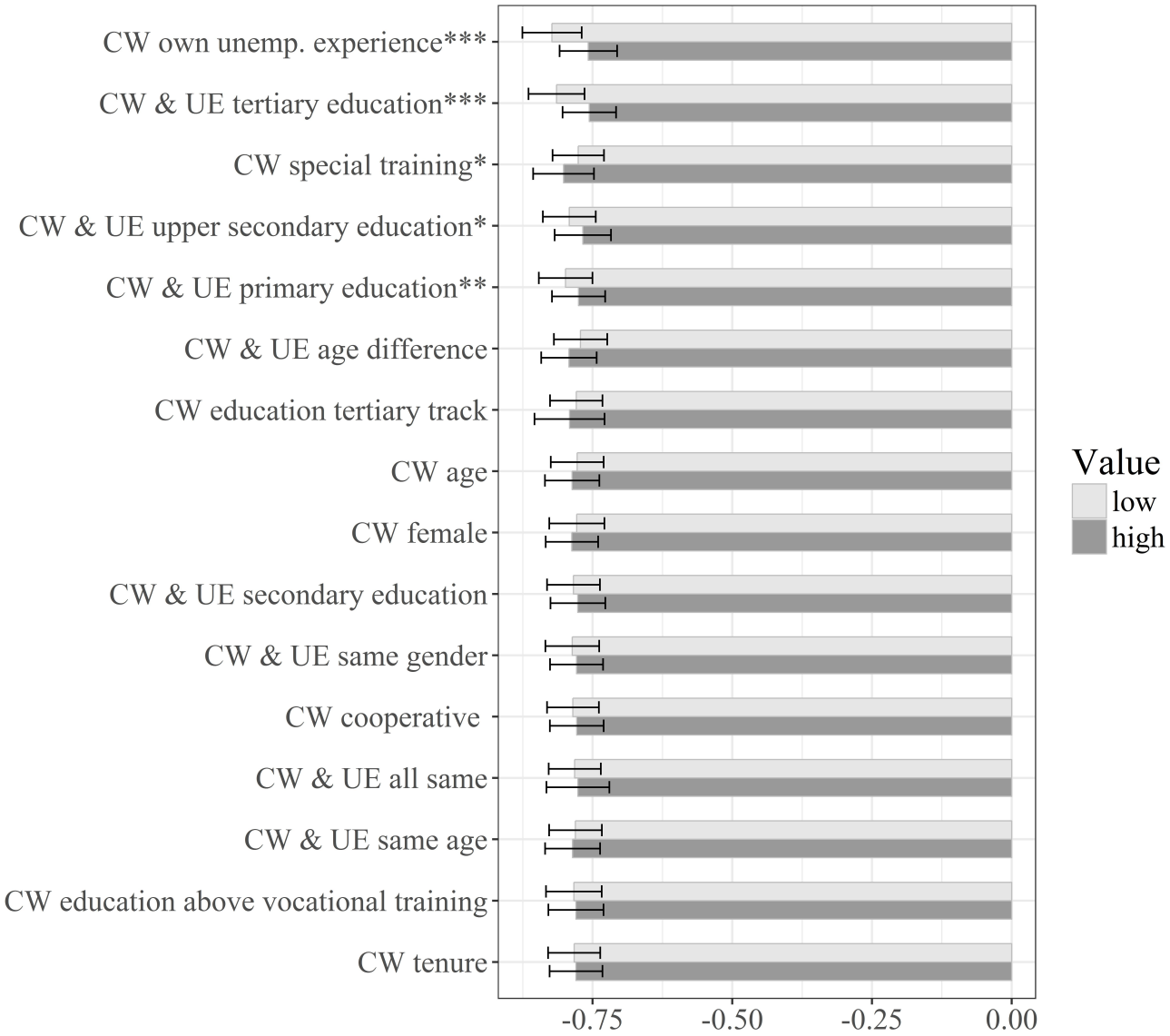
Characteristics of Unemployed Persons



Large heterogeneity by employability rating: Unemployed with low employability rating suffer much less

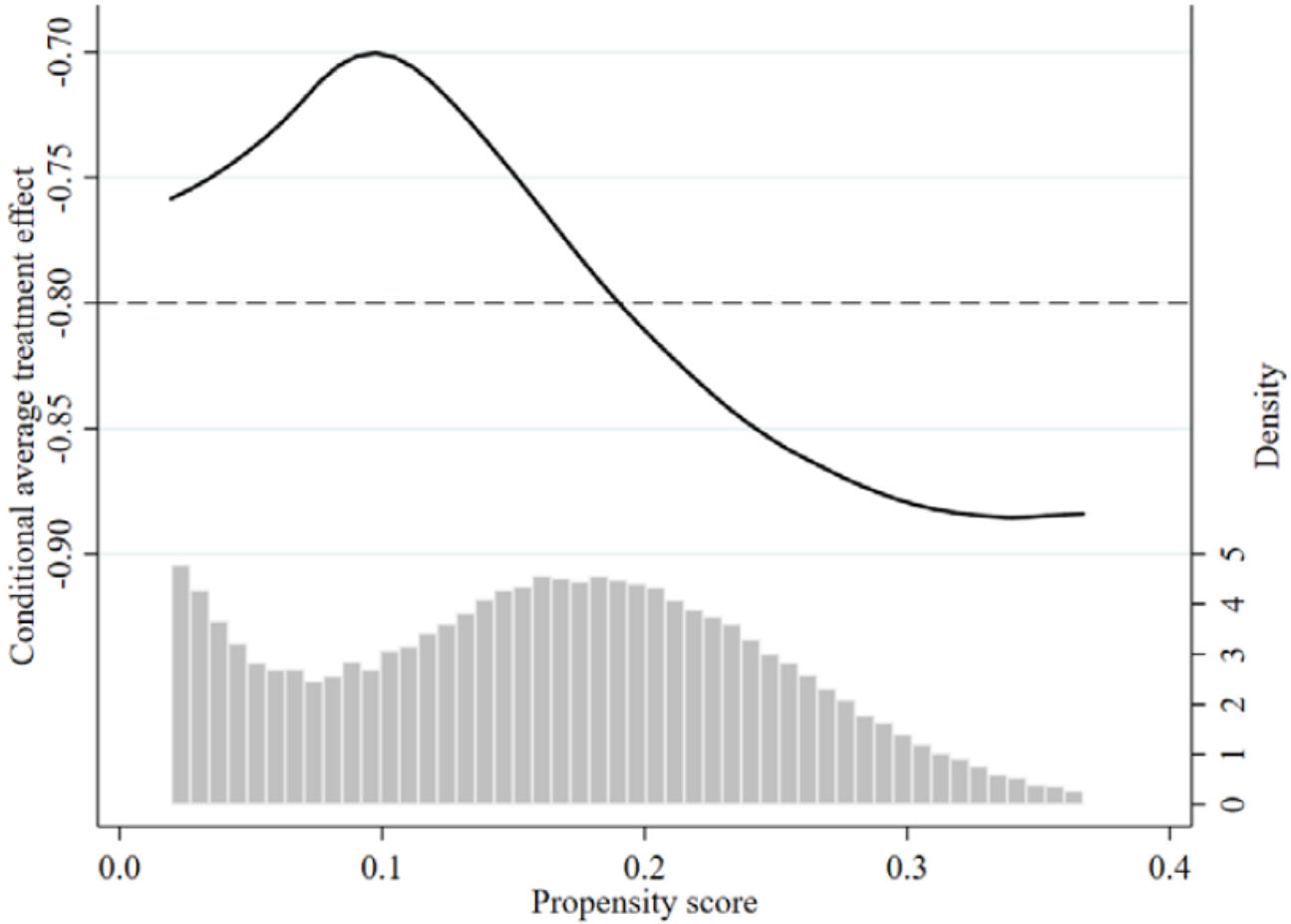
CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

Caseworker Characteristics



CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary.

Correlation of the propensity score and the CATEs



Kernel smoothed regression of propensity score and CATEs. Local constant kernel regression used with Gaussian kernel and bandwidth 0.02. The dashed horizontal line shows the ATE. The grey bars show the histogram of the propensity score.

R-Learning

Objective function:

$$\min_{\delta} E\left[\left(Y_i - \mu(X_i) - (D_i - p(X_i))\delta(X_i)\right)^2\right]$$

with $\mu(X_i) = E[Y_i | X_i = x]$

$\rightarrow$ Based on the partially linear model of Robinson (1988)

$\rightarrow$ Similar to modified covariate method when $p(x) = 0.5$, but different weighting scheme otherwise

- Sometimes called A-learning (e.g., Chen, Tian, Cai, and Yu, 2017)

Reference: Nie and Wager (2018)

Advantages and Disadvantages

Advantages:

- Estimates CATEs explicitly
- Oracle properties known
- Neyman orthogonality

Disadvantages:

- Currently only used for linear model $\delta(X_i) = X_i \delta_v$

Algorithm for Linear R-Learning

1. Split data in samples S^A and S^B
2. Estimate the nuisance parameters $\mu^A(x)$ and $p^A(x)$ in S^A; and $\mu^B(x)$ and $p^B(x)$ in S^B with ML
3. Estimate the LASSO (or other linear) model

$$\min_{\delta_v^A} \sum_{i \in S^A} \left(Y_i - \mu^B(X_i) - (D_i - p^B(X_i)) X_i \hat{\delta}_v^A \right)^2 + \lambda \| \hat{\delta}_v^A \|_1$$

   and correspondingly for $\hat{\delta}_v^B$
4. The the fitted values are the CATEs $\hat{\delta}(X_i) = \hat{Y}_i$ and the ATEs are $\hat{\delta} = E[\hat{Y}_i]$

General Framework

$$\min_{\delta} E\left[W_i \left(Y_i^* - \delta(X_i)\right)^2\right]$$

Approach	W_i	Y_i^*
MOM IPW	1	$Y_{i,IPW}^*$
MOM DML	1	$Y_{i,DML}^*$
MCM	$T_i \frac{D_i - p(X_i)}{4p(X_i)(1-p(X_i))}$	$2T_iY_i$
MCM with EA	$T_i \frac{D_i - p(X_i)}{4p(X_i)(1-p(X_i))}$	$2T(Y_i - \mu(X_i))$
R-learning	$(D_i - p(X_i))^2$	$\frac{Y_i - \mu(X_i)}{D_i - p(X_i)}$

Reference: Knaus, Lechner, and Strittmatter (2018)

4. Adapting the ML Algorithm

- CATE: $\delta(x) = \mu_1(x) - \mu_0(x)$
- Criterion to estimate $\hat{\delta}(x)$:

$$\begin{aligned} MSE(\hat{\delta}(x)) =& E\left[(\hat{\delta}(x) - \delta(x))^2\right], \\ =& E\left[(\hat{\mu}_1(x) - \mu_1(x))^2\right] + E\left[(\hat{\mu}_0(x) - \mu_0(x))^2\right] \\ & - 2E\left[(\hat{\mu}_1(x) - \mu_1(x))\right] E\left[(\hat{\mu}_0(x) - \mu_0(x))\right], \\ =& MSE(\hat{\mu}_1(x)) + MSE(\hat{\mu}_0(x)) - 2MCE(\hat{\mu}_1(x), \hat{\mu}_0(x)), \end{aligned}$$

with mean-correlation-error $MCE = E\left[(\hat{\mu}_1(x) - \mu_1(x))\right] E\left[(\hat{\mu}_0(x) - \mu_0(x))\right]$

$\rightarrow$ The main complication is the estimation of MCE

Reference: Lechner (2018)

Simple Solution

- Estimate $\hat{\mu}_1(x)$ in the sample with $D_i = 1$ and $\hat{\mu}_0(x)$ in the sample with $D_i = 0$ using two separate ML optimisation criteria
- For example, optimise π_d for the tree estimator using the criteria

$$\widehat{MSE}(\hat{\mu}_d(x)) = \frac{1}{N_X^d} \sum_{i=1}^{N} 1\{D_i = d\} 1\{X_i \in l_j(x, d, \pi)\} \cdot (\hat{\mu}_d(X_i) - Y_i)^2$$

with

$$\hat{\mu}_d(X_i) = \frac{1}{N_X^d} \sum_{i=1}^{N} 1\{D_i = d\} 1\{X_i \in l_j(x, d, \pi)\} \cdot Y_i$$

and $N_X^d = \sum_{i=1}^{N} 1\{D_i = d\} 1\{X_i \in l_j(x, d, \pi)\}$

- Estimate CATEs with $\hat{\delta}(x) = \hat{\mu}_1(x, \pi_1^*) - \hat{\mu}_0(x, \pi_0^*)$
- Implicit assumption: $MCE = const$

Nearest Neighbour Fix

- Use nearest neighbour to approximate the MCE

$$\widehat{MCE}(\hat{\mu}_d(x)) = \frac{1}{N_X}\sum_{i=1}^{N} 1\{X_i \in l_j(x,d,\pi)\} \cdot \left(\hat{\mu}_1(X_i) - \tilde{Y}_i^1\right)\left(\hat{\mu}_0(X_i) - \tilde{Y}_i^0\right)$$

with

$$\tilde{Y}_i^d = \begin{cases} Y_i & \text{if } D_i = d \\ \tilde{Y}_i^{NN,1-d} & \text{if } D_i = 1-d \end{cases}$$

and $N_X = \sum_{i=1}^{N} 1\{X_i \in l_j(x,d,\pi)\}$

- Select the splitting rule π_δ minimising
$\widehat{MSE}(\hat{\delta}(x)) = \widehat{MSE}(\hat{\mu}_1(x)) + \widehat{MSE}(\hat{\mu}_0(x)) - 2\widehat{MCE}(\hat{\mu}_1(x), \hat{\mu}_0(x))$
- Splits can be based on D_i
- Estimate CATEs with $\hat{\delta}(x) = \hat{\mu}_1(x, \pi_\delta^*) - \hat{\mu}_0(x, \pi_\delta^*)$

Reference: Lechner (2018)

Causal Tree

- S^{te} is the test sample and S^{tr} is the training sample
- Optimisation criteria for δ:

$$
\begin{aligned}
MSE_{\delta}(S^{te}, S^{tr}, \pi) &= \frac{1}{N^{te}} \sum_{i \in S^{te}} \left\{ (\hat{\delta}(X_i; S^{tr}, \pi) - \delta_i)^2 + \delta_i^2 \right\}, \\
&= \frac{1}{N^{te}} \sum_{i \in S^{te}} \left\{ \hat{\delta}^2(X_i; S^{tr}, \pi) - 2 \cdot \delta_i \cdot \hat{\delta}(X_i; S^{tr}, \pi) \right\}, \\
&= \frac{1}{N^{te}} \sum_{i \in S^{te}} \left\{ \hat{\delta}^2(X_i; S^{tr}, \pi) - 2 \cdot \hat{\delta}(X_i; S^{te}, \pi) \cdot \hat{\delta}(X_i; S^{tr}, \pi) \right\},
\end{aligned}
$$

with $E_{S^{te}}[\delta_i | i \in S^{te} : i \in l_j(x, d, \pi)] = E[\hat{\delta}(X_i; S^{te}, \pi)]$

- In-sample approximation of optimisation criteria:

$$
\widehat{MSE}_{\delta}(S^{tr}, S^{tr}, \pi) = -\frac{1}{N^{tr}} \sum_{i \in S^{tr}} \left\{ \hat{\delta}^2(X_i; S^{tr}, \pi) \right\},
$$

Reference: Athey and Imbens (2016)

Extended Optimisation Criteria

- S^{est} is the estimation sample
- Athey and Imbens (2016) propose to expand the MSE for honest causal trees:

$$\widehat{EMSE}_{\delta}(\pi) = -\frac{1}{N^{tr}} \sum_{i \in S^{tr}} \left\{ \hat{\delta}^2(X_i; S^{tr}, \pi) \right\} + \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}} \right) \sum_{j=1}^{\#(\pi)} \left(\frac{\sigma_j^2(S^{tr}_{treat})}{p} + \frac{\sigma_j^2(S^{tr}_{control})}{1-p} \right)$$

where $\sigma_j^2(S^{tr})$ is the within-leaf variance and $p = Pr(D_i = 1)$

Generalised Causal Forest (GCF)

- Consider the random effects model $Y_i = D_i\delta(X_i) + \varepsilon_i$ with

$$\delta(x) = Var(D_i|X_i = x)^{-1} Cov(Y_i, D_i|X_i = x)$$

- Estimator

$$\hat{\delta}(x) = \left(\sum_{i=1}^{N} \alpha_i(x)(D_i - \bar{D}_\alpha)^2 \right)^{-1} \sum_{i=1}^{N} \alpha_i(x)(D_i - \bar{D}_\alpha)(Y_i - \bar{Y}_\alpha)$$

  with $\bar{D}_\alpha = \sum_{i=1}^{N} \alpha_i(x) D_i$ and $\bar{Y}_\alpha = \sum_{i=1}^{N} \alpha_i(x) Y_i$

- The weights $\alpha_i(x)$ are obtained with the generalised causal forest algorithm
  - Weights for causal tree g:

$$\alpha_{ig}(x) = \frac{1\{X_i \in l_j(x, d, \pi_g)\}}{\sum_{i=1}^{N} 1\{X_i \in l_j(x, d, \pi_g)\}}$$

  - Causal forest weights:

$$\alpha_i(x) = \frac{1}{G} \sum_{i=1}^{G} \alpha_{ig}(x)$$

Reference: Athey, Tibshirani, and Wager (2018)

Algorithm

- Optimal optimisation criteria:

$$\max \Delta(C_1, C_2) = \frac{N_{C_1} N_{C_2}}{N_P} \left(\hat{\delta}_{C_1} - \hat{\delta}_{C_2} \right)^2$$

  $\rightarrow$ Requires the estimation of $\hat{\delta}_{C_1}$ and $\hat{\delta}_{C_2}$ at each candidate split

- Computational efficient optimisation algorithm

  (1) **Labelling step:** Calculate $\hat{\delta}_P$, $A_P = Var_P(D_i)$, and the pseudo-outcome

$$p_i = A_P^{-1}(D_i - \bar{D}_P)(Y_i - \bar{Y}_P - (D_i - \bar{D}_P)\hat{\delta}_P)$$

  (2) **Regression step:**

$$\max \tilde{\Delta}(C_1, C_2) = \frac{1}{N_{C_1}} \left(\sum_{i:X_i \in C_1} p_i \right)^2 + \frac{1}{N_{C_2}} \left(\sum_{i:X_i \in C_2} p_i \right)^2$$

  (3) Relabel child nodes to parent nodes and repeat (1) and (2) until stopping criteria of tree is reached

  (4) Calculate weights $\alpha_{ig}(x)$ and build forest by repeating (1)-(3) with different subsamples and covariates

Pseudo-Outcome

- Assume D_i is binary and randomly assigned and denote $p = \bar{D}_P = Pr(D_i = 1) = Pr(D_i = 1|C_j)$

$$
\begin{aligned}
E[p_i|C_j] =& E\left[\frac{D_i - p}{p(1-p)}(Y_i - \bar{Y}_P - (D_i - p)\hat{\delta}_P)\middle| C_j\right] \\
=& E\left[\frac{1}{p}(Y_i - \bar{Y}_P - (1-p)\hat{\delta}_P)\middle| C_j, D_i = 1\right] p \\
& - E\left[\frac{1}{(1-p)}(Y_i - \bar{Y}_P + p\hat{\delta}_P)\middle| C_j, D_i = 0\right](1-p) \\
=& E\left[Y_i - \bar{Y}_P - (1-p)\hat{\delta}_P\middle| C_j, D_i = 1\right] - E\left[Y_i - \bar{Y}_P + p\hat{\delta}_P\middle| C_j, D_i = 0\right] \\
=& E\left[Y_i(1) - \bar{Y}_P\middle| C_j, D_i = 1\right] - E\left[Y_i(0) - \bar{Y}_P\middle| C_j, D_i = 0\right] - \hat{\delta}_P \\
=& E\left[Y_i(1) - Y_i(0)|C_j\right] - \hat{\delta}_P
\end{aligned}
$$

- Difference between approximated causal effect at child and parent node
- Pseudo-outcome is updated at each parent node

Local Centering

- Generalised causal forest is targeted to find maximum heterogeneity in pseudo-outcome
- But not specifically designed to account for selection into treatment (even though deep causal forests correct automatically for some extent of selection)
- Define centred variables

$$\tilde{Y}_i = Y_i - \hat{\mu}(X_i)$$

and

$$\tilde{D}_i = Y_i - \hat{p}(X_i)$$

with $\hat{\mu}(x) = \hat{E}[Y_i|X_i = x]$ and $\hat{p}(x) = \hat{E}[D_i|X_i = x]$

- Apply generalised causal forest algorithm to $\tilde{Y}_i$ and $\tilde{D}_i$ instead of Y_i and D_i

$\rightarrow$ Orthogonalisation

Local Centring (cont.)

MSE from Simulation

Confounding	Heterogeneity	K	N	GCF	Centred GCF
No	No	10	800	0.85	0.87
No	No	10	1,600	0.58	0.59
No	No	20	800	0.92	0.93
No	No	20	1,600	0.52	0.52
Yes	No	10	800	1.12	0.27
Yes	No	10	1,600	0.80	0.20
Yes	No	20	800	1.17	0.17
Yes	No	20	1,600	0.95	0.11
Yes	Yes	10	800	1.92	0.91
Yes	Yes	10	1,600	1.51	0.62
Yes	Yes	20	800	1.92	0.93
Yes	Yes	20	1,600	1.55	0.57

Note: K is the number of covariates in the simulation

Source: Athey, Tibshirani, and Wager (2018)

Asymptotic Properties for Causal Forest

- Minimum subsample size S is scaled N^{β} with $\beta_{min} < \beta < 1$
- CATEs are consistent and asymptotically normal

$$(\hat{\delta}(x) - \delta(x))/\sqrt{Var(\delta(x))} \xrightarrow{d} N(0,1)$$

- Infinitesimal Jackknife
  - $Q_{ig} = 1$ when observation i is used to built tree g and $Q_{ig} = 0$ otherwise
  - Calculate the covariance $Cov(\hat{\delta}(x), Q_{ig})$ across all trees $g = 1, ..., G$
  - Variance estimator:

$$\hat{V}(x) = \frac{N-1}{N}\left(\frac{N}{N-S}\right)^2 \sum_{i=1}^{N} Cov(\hat{\delta}(x), Q_{ig})^2$$

  - $(N/(N-S)))^2$ is a finite sample correction for subsampling
  - $\hat{V}(x) \xrightarrow{p} Var(\delta(x))$

Reference: Wager and Athey (2018)

Advantages and Disadvantages of Causal Forest

Advantages:

- D_i can be binary or continuous
- Asymptotic properties even for CATEs available
- Variance estimator available
- Extensions to IV and quantiles available

Disadvantages:

- Best suited for experiments
- We have to assume that subsample sizes do not get too small, because otherwise asymptotic results break down

Application: Summer Jobs

- Chicago's One Summer Plus (OSP) program conducted in 2012 and 2013
- OSP provides disadvantaged youth ages 14-22 with 25 hours a week of employment, an adult mentor, and some other programming
- Participants are paid Chicago's minimum wage ($8.25 at the time)
- Previous study finds on average 43 percent reduction in violent-crime arrests in the 16 months after random assignment (Heller, 2014)
- **Outcomes**:
  - Violent-crime arrests within two years of random assignment
  - Employment during the six quarters after the program
- **Covariates:**
  age, gender, education, ethnicity, criminal history, employment history, regional unemployment rate, regional median income

Reference: Davis and Heller (2017)

Application: Summer Jobs (cont.)

Estimation target GATEs:

$$\hat{\delta}(+) = E[\hat{\delta}(X_i)|\hat{\delta}_{tr}(X_i) > 0] \text{ and } \hat{\delta}(-) = E[\hat{\delta}(X_i)|\hat{\delta}_{tr}(X_i) \leq 0]$$

	No. arrests	Employment
	In-sample	
$\hat{\delta}(+)$	0.22	0.19
	(0.05)	(0.03)
$\hat{\delta}(-)$	-0.05	-0.14
	(0.02)	(0.03)
H_0: p-val	0.00	0.00
	Out-of-sample	
$\hat{\delta}(+)$	-0.01	0.08
	(0.05)	(0.03)
$\hat{\delta}(-)$	-0.02	-0.01
	(0.02)	(0.03)
H_0: p-val	0.77	0.02

5. Performance Comparison

- Empirical Monte Carlo Study: Knaus, Lechner, and Strittmatter (2018)

	Random Forest	Lasso	Cross-fitting
Infeasible benchmark	x	x	
Conditional mean regression	x	x	
MOM IPW	x	x	x
MOM DR	x	x	x
MCM		x	x
MCM with EA		x	x
R-learning		x	x
Causal Forest	x		
Causal Forest with local centering	x		x

Results

- MOM DR, MCM with EA, R-learning, and Causal Forest with local centering show approximately equally good finite sample performance
  $\rightarrow$ all estimators incoporate some sort of orthogonalisation
- Convergence rates for CATEs are fairly below $\sqrt{N}$, but increase for GATEs and ATEs
- LASSO estimates have a higher variance than Forest estimates, particularly because of heavy tails in the smaller samples ($N = 1,000$)
- Forests estimates approximate normal distribution even in smaller samples well

Reference: Knaus, Lechner, and Strittmatter (2018)

6. IV with Many Instruments

- Belloni, Chen, Chernozhukov, and Hansen (2012): ”Sparse Models and Methods for Instrumental Regression, with an Application to Eminent Domain”, Econometrica, 80 (6), 2369-2429, download
- Hansen and Kozbur (2014): ”Instrumental Variables Estimation with Many Weak Instruments Using Regularised JIVE”, Journal of Econometrics, 182(2), 290-308, download
- Breunig, Mammen, and Simoni (2018): ”Ill-posed Estimation in High-Dimensional Models with Instrumental Variables”, download”