Double ML: Causal Inference based on ML

Part III: AutoDML

uai2022, August 1, 2022, Eindhoven

Philipp Bach, Martin Spindler (UHH & Economic AI)

Collaborators: Victor Chernozhukov (MIT), Malte Kurz (TUM)

Outline

- Present a simple, yet general framework for learning and bounding causal effects, that utilizes machine learning (aka adaptive statistical learning methods)
- List of examples all for *general, nonseparable* models:
  - (weighted) average potential outcomes; e.g. policy values
  - average treatment effects, including subgroup effects such as for the treated,
  - (weighed) average derivatives
  - average effects from transporting covariates
  - distributional changes in covariates

Many other examples fall in or extend this framework (mediators, surrogates, dynamic effects)

Outline

- Using machine learning is great, because we can learn regression functions very well
- However, since ML has to "shrink, chop, and throw out" variables to perform prediction well in high-dimensional settings, the learners are biased. These biases transmit into estimation of main causal effects
- The biases can be eliminated by using carefully crafted -- Neyman orthogonal -- score functions. Additional overfitting biases are dealt with by using cross-fitting
- Another source of biases is presence of *unobserved confounders*. These biases can not be eliminated, but we can *bound* the biases and perform inference on the size of the bias under the hypotheses that limit the strength of confounding

Set-up: Causal Inference via Regression

The set-up uses potential outcomes framework (Imbens and Rubin, 2015). Let $Y(d)$ denote the potential outcome in policy state d. The chosen policy D is assumed to be independent of potential outcomes conditional on controls X and A:

$$Y(d) \perp D \mid X, A.$$

The observed outcome Y is generated via

$$Y := Y(D).$$

Under the conditional exogeneity condition,

$$E[Y(d) \mid X, A] = E[Y \mid D = d, X, A] =: g(d, X, A),$$

that is the *conditional average potential outcome coincides with the regression function*

Running Example

The key examples of causal parameters include the average causal effect (ACE):

$$\theta = E[Y(1) - Y(0)] = E[g(1, X, A) - g(0, X, A)]$$

for the case of the binary $\boldsymbol{d}$, and the average causal derivative (ACD), for the case of continuous $\boldsymbol{d}$:

$$\theta = E[\partial_d Y(d) \mid_{d=D}] = E[\partial_d g(D, X, A)]$$

More Examples

- Average Incremental Effect (AIE):

$$\theta = E[Y(D+1) - Y(D)] = E[g(D+1, X, A)] - EY.$$

- Average Policy Effect (APEC) from covariate Shift:

$$\theta = \int E[Y(d) \mid X = x]\mathrm{d}(F_1(x) - dF_0(x)).$$

$$= \int E[g(d, x, A)]\mathrm{d}(F_1(x) - dF_0(x))$$

- See others in Chernozhukov et al. (2018b), Chernozhukov et al. (2020), Chernozhukov et al. (2021a), Chernozhukov et al. (2022), Singh (2021)

Case I: No Unobserved Confounders

Let $W := (D, X, A)$ be all observed variables.

Assumption Target Parameter The target parameter can be expressed as a continuous linear functional of long regression:

$$\theta := Em(W; g),$$

where $g \mapsto Em(W; g)$ is continuous in g with respect to the $L^2(P)$ norm

In working examples above

- $m(W, g(W)) = g(1, X, A) - g(0, X, A)$ for ACE and
- $m(W, g(W)) = \partial_d g(D, X, A)$ for ACD.

Weak overlap conditions are required to make the continuity hold

Case I: No Unobserved Confounders

> Lemma: Riesz Representation (Cherozhukov et al. 2021a, 2021b)
>
> There exist unique square integrable random variables $\alpha(W)$ such that
>
> $$Em(W,g) = Eg(W)\alpha(W),$$
>
> for all square-integrable g

Partially linear models

(Frisch-Waugh) For partially linear models,

$$g(W) = \theta D + f(X,A),$$

then for either ACE or ACD we have that

$$\alpha(W) = \frac{D - E[D \mid X,A]}{E(D - E[D \mid X,A])^2},$$

Case I: No Unobserved Confounders

General nonparametric models:

- (Horwitz-Thomposon). In the case of ACE,

$$\alpha(W) = \frac{1(D=1)}{P(D=1 \mid X, A)} - \frac{1(D=0)}{P(D=0 \mid X, A)}.$$

- (Powell-Stock-Stocker) For the case of ACD,

$$\alpha(W) = -\partial_d \log f(D \mid X, A)$$

Case I: No Unobserved Confounders

- It turns out, we don't need closed form solutions for RRs for each new problem. We can obtain them automatically

> Lemma: Auto Characterization for Representers (Chernozhukov et al. 2021a, 2021b)
>
> $$\alpha = \arg\min_{a(W)} E[a^2(W) - 2m(W, a)].$$

- Chernozhukov et al. (2021a) and Chernozhukov et al. (2018c)] employ this formulation to learn the Riesz representer without knowing the functional form
- Chernozhukov et al. (2020) and Chernozhukov et al. (2018b) employ adversarial method of moments to learn the representer without knowing the functional form

Case I: No Unobserved Confounders

Three potential representations for the target parameter, θ:

$$\theta = Em(W, g) = EY\alpha = Eg\alpha.$$

the "regression matching", "propensity score", and "mixed" approaches respectively

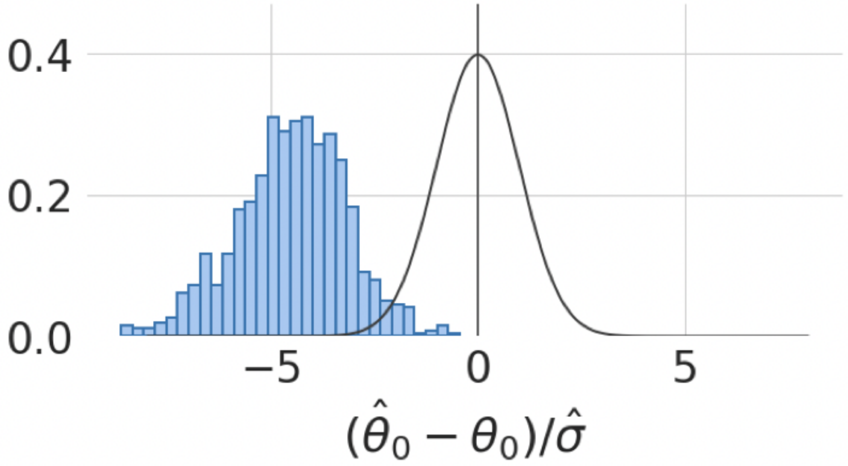
Which one should we use?

- In parametric models, wide path, because can use parametric learning of g or α and use expression above
- In low-dimensional nonparametric models, still a wide path using flexible parametric approximations (series and sieve methods)

Case I: No Unobserved Confounders

What about modern high-dimensional nonparametric problems, when we are forced to use machine learning to learn g or learn α?

- Problem: shrinkage biases create inferential problems for either one of the approaches



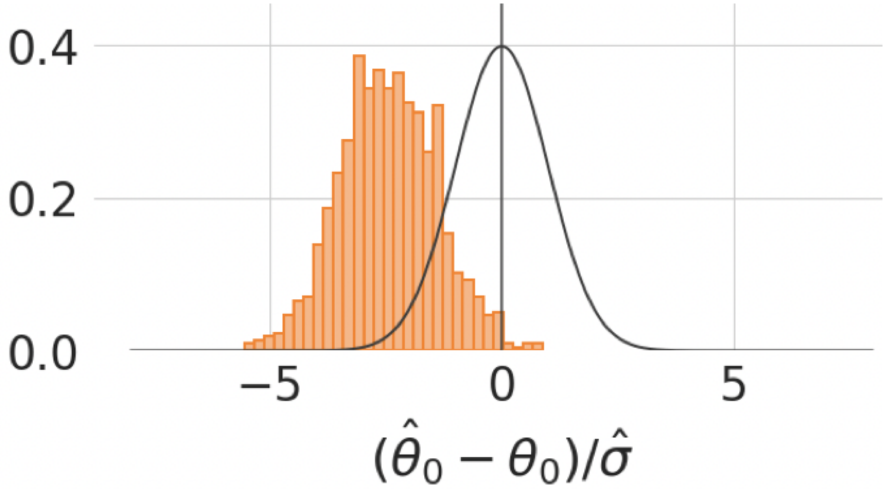
Case I: No Unobserved Confounders

Debiased Learning: Narrow the Path

Narrow the Path: Use all three of the learning approaches:

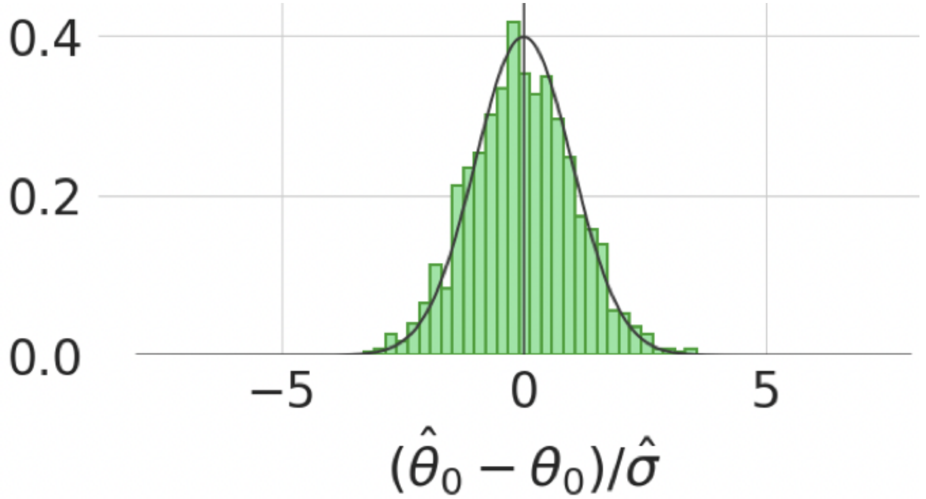
$$\theta = Em(W, g) + EY\alpha - Eg\alpha.$$

(Intuitively, each part corrects the bias in the other.)



Case I: No Unobserved Confounders

Narrow the Path Even More: Use Cross-Fitting to Eliminate Overfitting Biases (Entopy Bias)



Big Picture

Debiased machine learning is a generic recipe that isolates the narrow path. It is a

- method-of-moments estimator
- that utilizes any debiased/orthogonal moment scores
- together with cross-fitting
- automatic learning of representers aids the construction
- Delivers standard approximately normal inference on main parameters

Applies more broadly than the setting discussed here, for example for economic models identified through conditional method of moments (Chamberlain, 1987), albeit much more work is needed in this area.

References

References

- Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.
- Victor Chernozhukov et al. 'Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression'. In: arXiv preprint arXiv:2104.14737 (2021a).
- Victor Chernozhukov et al. 'Adversarial estimation of riesz representers'. In: arXiv preprint arXiv:2101.00009 (2020).
- Rahul Singh. 'A Finite Sample Theorem for Longitudinal Causal Inference with Machine Learning: Long Term, Dynamic, and Mediated Effects'. In: arXiv preprint arXiv:2112.14249 (2021).
- Victor Chernozhukov, Whitney Newey, and Rahul Singh. 'De-biased machine learning of global and local parameters using regularized Riesz representers'. In: arXiv preprint arXiv:1802.08667 (2018b).
- Victor Chernozhukov et al. 'Automatic Debiased Machine Learning for Dynamic Treatment Effects'. In: arXiv preprint arXiv:2203.13887 (2022).

- Victor Chernozhukov et al. RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests (2021b).

- Victor Chernozhukov, Whitney K Newey, and Rahul Singh. 'Automatic debiased machine learning of causal and structural effects'. In: arXiv preprint arXiv:1809.05224 (2018c).

- G. Chamberlain. 'Asymptotic Efficiency in Estimation with Conditional Moment Restrictions'. In: Journal of Econometrics 34 (1987), pp. 305–334.
- Victor Chernozhukov et al. 'Double/debiased machine learning for treatment and structural parameters'. In: The Econometrics Journal (2018a). ArXiv 2016; arXiv:1608.00060.

Appendix

Theoretical Details

For debiased machine learning we use representations:

$$\theta = E[m(W, g) + (Y - g)\alpha],$$

We have that

$$\theta - E[m(W, \bar g) + (Y - \bar g)\bar\alpha] = -E(\bar g - g)(\bar\alpha - \alpha).$$

Therefore, this representation has the Neyman orthogonality property:

$$ \partial{\bar g, \bar \alpha} E [m(W, \bar g) + (Y- \bar g) \bar \alpha] \Big |{\bar \alpha = \alpha, \bar g = g} =0, $$

where ∂ is the Gateaux (pathwise derivative) operator.

Theoretical Details

Therefore the estimators are defined as

$$\hat{\theta} := DML(\psi_\theta);$$

for the score:

$$\psi_\theta(Z;\theta;\alpha,g) := \theta - m(W,g) + (Y-g)\alpha(W);$$

Generic DML is a method-of-moments estimator that utilizes any Neyman orthogonal score, together with cross-fitting.

Theoretical Details

Definition: DML (ψ)

Input the Neyman-orthogonal score $\psi(Z;\beta,\eta)$, where $\eta=(g,\alpha)$ are nuisance parameters and β is the target parameter. Input random sample $(Z_i := (Y_i, D_i, X_i, A_i))_{i=1}^{n}$. Then

- Randomly partition $\{1,\ldots,n\}$ into folds $(I_\ell)_{\ell=1}^{L}$ of approximately equal size. For each ℓ, estimate $\hat{\eta}_\ell=(\hat{g}_\ell,\widehat{\alpha}_\ell)$ from observations excluding I_ℓ.
- Estimate β as a root of:

$$0 = n^{-1}\sum_{l=1}^{L}\sum_{i\in I_l}\psi(\beta, Z_i;\hat{\eta}_l)$$

Output $\widehat{\beta}$ and the estimated scores $\widehat{\psi}^{o}(Z_i)=\psi(\widehat{\beta}, Z_i;\hat{\eta}_\ell)$

Theoretical Details

R notebook, http://www.kaggle.com/r4hu15in9h/auto-dml

Lasso Learner for Nuisance Parameters:

- Regression Learner: Over a subset of data excluding I_ℓ:

$$\min \sum_{i \notin I_l} (Y_i - g(W_i))^2 + \mathrm{pen}(g) :$$

$$g(W_i) = b(W_i)'\gamma; \;\; \mathrm{pen}(g) = \lambda_g \sum_j |\gamma_j|,$$

where $b(W_i)$ is dictionary of transformations of W_i, for example polynomials and interactions, and λ_g is penalty level.

Theoretical Details

- Representer Learner: Over a subset of data excluding I_ℓ:

$$\min \sum_{i \in I_l^c} a^2(W_i) - 2m(W_i, a) + \text{pen}(a) :$$

$$a(W_i) = b(W_i)'\rho; \;\; \text{pen}(a) = \lambda_a \sum_j |\rho_j|,$$

where λ_a is penalty level.

- Can use any high-quality regression learner in place of lasso.
- Can use random forest and neural network learners of RR. See Chernozhukov et al. (2018c)

Theoretical Details

We say that an estimator $\hat{\beta}$ of β is asymptotically linear and Gaussian with the centered influence function $\psi^o(Z)$ if

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi^o(Z_i) + o_P(1) \rightsquigarrow N(0, E\psi_0^2(Z)).$$

The application of the results in Chernozhukov et al. (2018a) for linear score functions yields the following result.

> **Theorem: DML for CEs**
>
> Suppose that we can learn g and α sufficiently well, at $o_P(n^{-1/4})$ rates in $L^2(P)$ norm. Then the DML estimator $\hat{\theta}$ is asymptotically linear and Gaussian with influence functions:
>
> $$\psi_\theta^o(Z) := \psi_\theta(Z; \theta, g, \alpha),$$
>
> evaluated at the true parameter values. Efficiency follows from Newey (1994). The covariance of the scores can be estimated by the empirical analogues using the covariance of the estimated scores.

Case II: Subset A of Confounders is not observed

We often do not observe $\boldsymbol{A}$, and therefore we can only identify the short regression:

$$g_s(D, X) := E[Y \mid D, X] = E[g(D, X, A)|D, X].$$

Given this short regression, we can compute "short" parameters (or approximation) $\boldsymbol{\theta_s}$ for $\boldsymbol{\theta}$: for ACE

$$\theta_s = E[g_s(1, X) - g_s(0, X)],$$

and for ACD,

$$\theta_s = E[\partial_d g_s(D, X)].$$

Our goal therefore is to provide bounds on the omitted variable bias (OMVB):

$$\theta_s - \theta,$$

under the assumptions that limit strength of confounding, and provide DML inference on its size.

For this kind of work we refer to Chernozhukov et al. (2021c).