Identifiable Energy-based Representations: An Application to Estimating Heterogeneous Causal Effects

AISTATS 2022

Yao Zhang*, Jeroen Berrevoets*, Mihaela van der Schaar

van_der_Schaar \ LAB
vanderschaar-lab.com

UNIVERSITY OF CAMBRIDGE

jb2384@cam.ac.uk
@J_Berrevoets
linkedin.com/in/jeroenberrevoets

We collect many variables to identify the causal effect

Consider the conditional average treatment effect (CATE):

$$\mathbb{E}[Y(1) - Y(O)|X]$$

Where $X \in \mathbb{R}^d$ are an individual's covariates, and $Y(W) \in \mathbb{R}$ is the potential outcome, with $W \in \{O, 1\}$ the treatment.

Before we can predict the CATE, we have to assume that *all* confounding variables are in X. Otherwise, we may be predicting a spurious (or biased) treatment effect.

In practice, people will collect a lot of variables to make sure all of the confounders are actually measured

While it does help avoiding spurious estimates, it also results in poor estimates due to *the curse of dimensionality*, especially in low samples.

We could reduce dimensions…

Of course, one could reduce dimensions using: PCA, AEs, …

… but they have downsides:
- Linear
- Not consistent

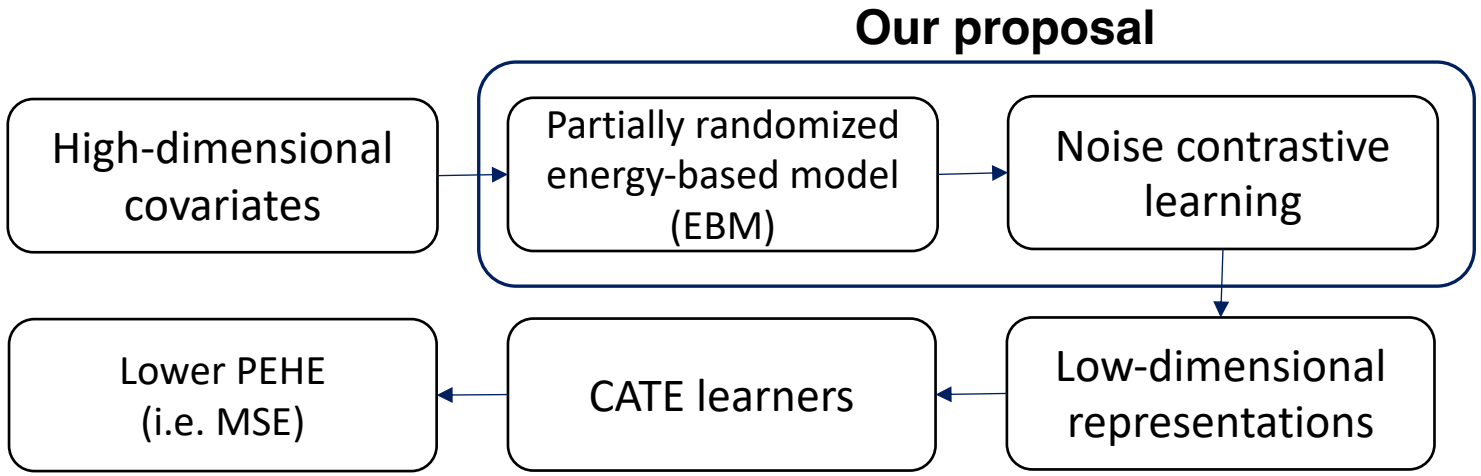
With linear dim-reduction, we loose valuable non-linear interactions: not good

With non-consistent methods, we throw away some of the CATE-literature's advances: also not good

We introduce a method that is both non-linear, and consistent

We propose an identifiable EBM to reduce dimensions

Let's have a high-level look at our proposal



We propose an identifiable EBM to reduce dimensions

Specifically…

- **Partially randomized EBM p_θ is a combination of multiple EBMs $p_{\theta,j}$ with a shared representation $f_\theta(x)$ and β's randomized and orthogonalized:**

$$p_\theta = (p_{\theta,j} : j \in [k])$$

$$p_{\theta,j}(x) = Z_{\theta,j}^{-1} \exp\left[-\beta_j^\top f_\theta(x)\right], \ \forall j \in [k]$$

- **With $f_\theta(x)$ parameterized by a neural network, we show the partially randomized p_θ has two properties:**
  - ✓ **$f_\theta(x)$ is identifiable up to some universal constants.**
  - ✓ **universal capacity for approximating any density functions**
- **Split the clean samples between $p_{\theta,j}$, then generate corrupted samples conditionally,**

$$\bar{X}_i = (X_i, \tilde{X}_{i1}, \ldots, \tilde{X}_{ib}) \sim p_X(x) \prod_{a=1}^{b} p_{\tilde{X}|X}(\tilde{x} \mid x)$$

- **Train $p_{\theta,j}$ jointly so the shared f_θ goes through noise contrastive training on full samples.**
- **We show f_θ converges to limits that are different by some universal constants**

Some results on prediction

Methods		X-Learner		DR-Learner		T-Learner		R-Learner	
	EBM	✗	✓	✗	✓	✗	✓	✗	✓
d	n	*Synth. data with increasing sample size and increasing dimensions*							
50	100	2.309 ±.00	**1.994** ±.02	4.594 ±.56	**2.017** ±.04	2.441 ±.00	**1.993** ±.01	3.194 ±.26	**1.982** ±.04
100	250	2.779 ±.00	**2.018** ±.01	4.056 ±.32	**2.154** ±.39	2.838 ±.00	**2.019** ±.01	3.702 ±.23	**2.018** ±.01
150	500	2.618 ±.00	**2.000** ±.01	3.030 ±.12	**2.001** ±.01	2.641 ±.00	**2.000** ±.01	2.877 ±.08	**2.000** ±.01
200	1k	2.185 ±.00	**1.940** ±.01	2.283 ±.02	**1.941** ±.01	2.189 ±.00	**1.939** ±.01	2.271 ±.01	**1.940** ±.01
250	1.5k	2.267 ±.00	**1.949** ±.02	2.427 ±.01	**1.976** ±.00	2.271 ±.00	**1.948** ±.01	2.436 ±.02	**1.949** ±.02
	n	*Synth. data with increasing sample size and dimensions fixed at* $d = 100$							
	100	2.134 ±.00	**1.927** ±.01	24.61 ±9.9	**2.096** ±.09	2.279 ±.00	**1.929** ±.01	3.192 ±.13	**1.925** ±.01
	250	2.779 ±.00	**2.018** ±.01	4.056 ±.32	**2.154** ±.39	2.838 ±.00	**2.019** ±.01	3.702 ±.23	**2.018** ±.01
	500	2.155 ±.00	**2.056** ±.02	2.334 ±.07	**2.273** ±.67	2.166 ±.00	**2.053** ±.02	2.271 ±.05	**2.056** ±.02
	1k	2.059 ±.00	**1.964** ±.02	2.105 ±.01	**2.016** ±.16	2.061 ±.00	**1.964** ±.02	2.086 ±.01	**1.965** ±.02
	1.5k	2.013 ±.00	**1.998** ±.02	2.043 ±.01	**1.998** ±.02	2.014 ±.00	**1.998** ±.02	2.024 ±.01	**1.991** ±.02
	n	*Twins* ($d = 48$) *with increasing sample size*							
	500	0.214 ±.00	**0.144** ±.00	0.236 ±.04	**0.182** ±.05	0.221 ±.00	**0.145** ±.00	0.222 ±.02	**0.145** ±.00
	1k	0.294 ±.00	**0.162** ±.00	0.348 ±.12	**0.173** ±.03	0.301 ±.00	**0.162** ±.01	0.532 ±.11	**0.161** ±.00
	1.5k	0.165 ±.00	**0.154** ±.00	0.189 ±.06	**0.159** ±.01	0.165 ±.00	**0.154** ±.00	0.172 ±.01	**0.154** ±.00
	2k	0.167 ±.00	**0.156** ±.00	0.197 ±.03	**0.159** ±.00	0.167 ±.00	**0.156** ±.00	0.222 ±.05	**0.157** ±.00
	2.5k	0.297 ±.00	**0.153** ±.00	0.390 ±.19	**0.156** ±.00	0.297 ±.00	**0.153** ±.00	0.358 ±.22	**0.153** ±.00

Methods		PCA	FA	SE	Isomap	K-PCA	AE	EBM
	n	*Twins* ($d = 48$) *with increasing sample size*						
	500	1.092 ±.11	1.758 ±1.1	1.011 ±.00	1.006 ±.00	1.015 ±.00	0.580 ±.03	**0.145** ±.00
	1k	1.015 ±.00	0.963 ±.00	1.010 ±.00	1.004 ±.00	1.010 ±.00	0.549 ±.04	**0.161** ±.00
	1.5k	1.014 ±.00	0.965 ±.00	1.005 ±.00	1.006 ±.00	1.012 ±.00	0.546 ±.04	**0.154** ±.00
	2k	1.013 ±.00	0.957 ±.00	1.009 ±.00	1.007 ±.00	1.013 ±.00	0.579 ±.03	**0.157** ±.00
	2.5k	1.007 ±.00	0.951 ±.00	1.002 ±.00	1.006 ±.00	1.006 ±.00	0.542 ±.04	**0.153** ±.00

Lower PEHE across learners

Lower PEHE than other dimensionality reduction methods

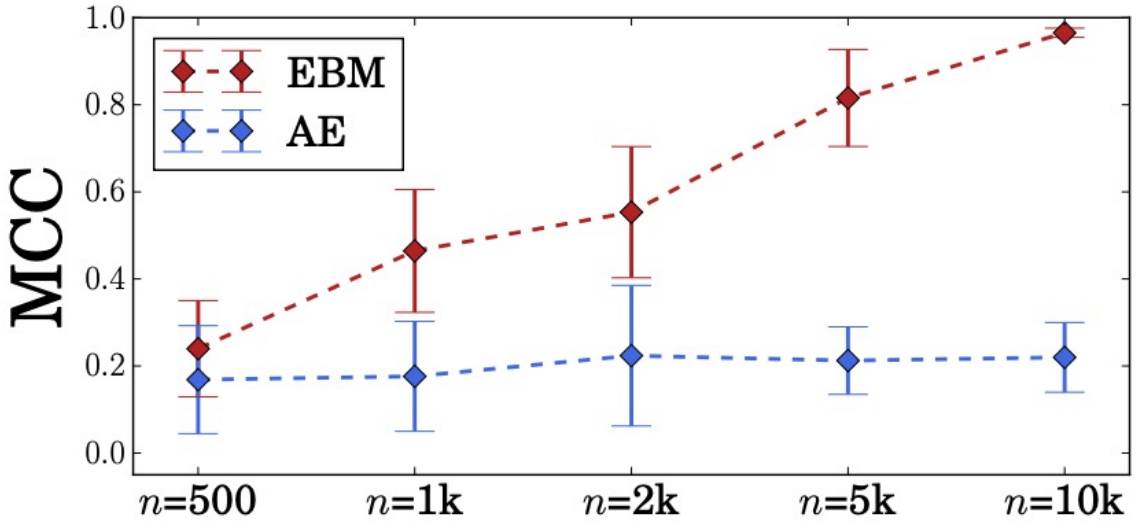
van_der_Schaar \ LAB

vanderschaar-lab.com

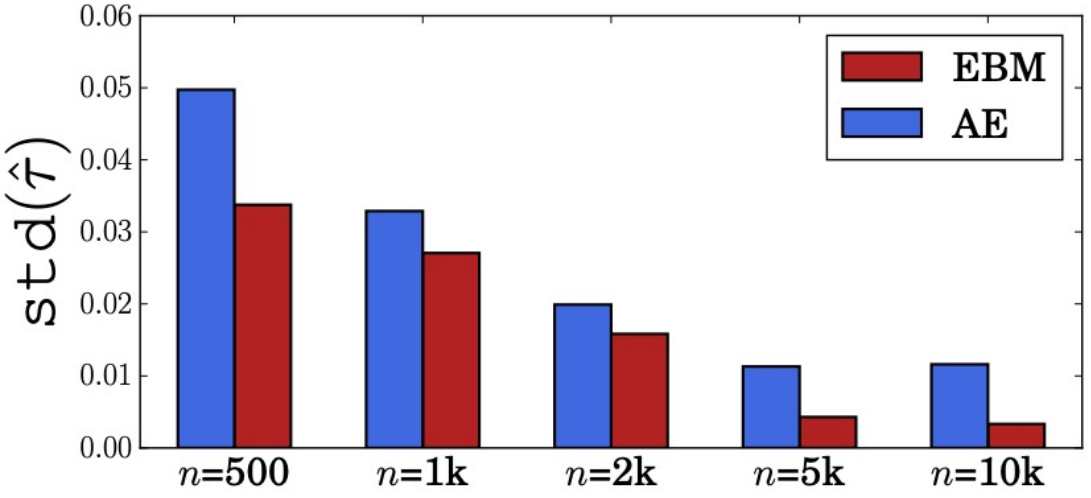
UNIVERSITY OF CAMBRIDGE

Some results on Identifiability

Converging Representations and CATE estimates



Increasing Mean Correlation Coefficient (MCC) with sample size



Decreasing standard error on CATE-estimates