Invariant Representation Learning for Treatment Effect Estimation

Claudia Shi Victor Veitch David M. Blei

Feburary 9th 2024

Presented by Yuqi Li

Section 1: Introduction

Challenge in causal inference from observational data

- The presence of confounders, needs adjustment
- Identify confounders that are safe to adjust for
- Avoid risk of including “bad-controls”, variables that induce bias when conditioned

Purpose

- Learn a **representation** of the covariates, that strips out bad controls but preserves sufficient information to adjust for confounding.
- Adjust for the learned representation, rather than the covariates themselves
- Provide valid causal estimation (estimation of treatment effect)

Section 2: Background

Consider the following causal inference problem:

Estimate the effect of sleeping pills on lung disease using EHR from multiple hospitals

- For each hospital e and patient i:

  observe drug administration T_i^e, outcome Y_i^e, covariates X_i^e

- The covariates include comprehensive information, there is no unobserved confounders
- The distribution of X^e is different across the datasets.
- The causal mechanism between sleeping pills T^e and lung disease Y^e remains the same

Main question:

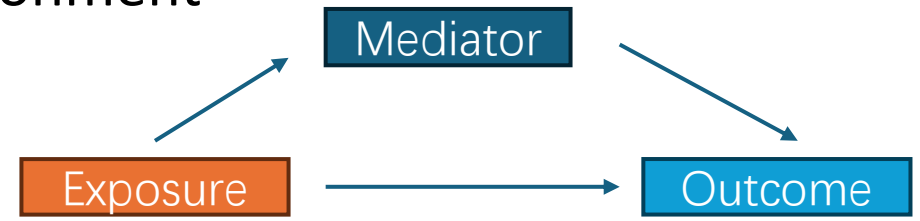
- How to use the multiple environments to find a representation of the covariates for valid causal estimation?
- Develop nearly invariant causal estimation (NICE)

Section 2: Background

NICE applies **Invariant Risk Minimization (IRM)** for causal adjustment.

Invariant Risk Minimization (IRM) (Arjovskyet al., 2019)

- produce a predictor that is robust to changes in the deployment domain
- learn an invariant representation $\Phi(T, X)$, a function such that the outcome Y and $\Phi(T, X)$ have the same relationship in each environment



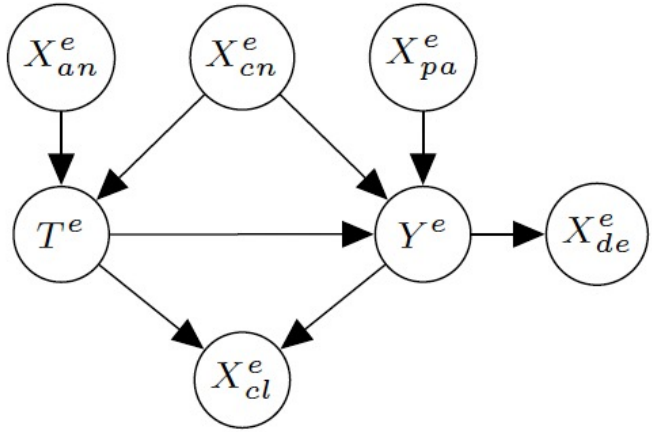
The main insight that enables NICE:

- the IRM invariant representation also suffices for causal adjustment
- a representation is invariant if and only if it is informationally equivalent to the causal parents of the outcome Y (assuming no mediators)
- The principle of invariance: if a relationship between X and Y is causal, then it is invariant to perturbations that changes the distributions of X.

Section 3: Methods 1) causal estimation

Assumptions:

- Multiple datasets (environments) , for each e, $\{X_i^e, T_i^e, Y_i^e\}\ iid \sim P^e$
- Enough information in X_i^e to estimate the causal effect, but not sure the composition
- Causal mechanism relating Y to T and X is the same in each environment



Goal: estimate **Average Treatment effect on the Treated (ATT)** in each environment

$$\psi^e = E[Y^e|do(T^e = 1), T^e = 1] - E[Y^e|do(T^e = 0), T^e = 1]$$

Admissible representation:

$\Phi(T, X)$, a function of full set of covariates but one that captures the confounding factors and excludes the bad controls, i.e. the descendants of the outcome.

Section 3: Methods 2) Invariant Risk Minimization (IRM)

IRM: a framework for learning predictors that perform well across many environments.

Definition 3.1 Valid Environment

Consider a causal graph G and a distribution $P(X, T, Y)$ respecting G.

Let G_e denote the graph under an intervention and $P^e = (X^e,\ T^e,\ Y^e)$.

An intervention is valid if :

1. $E_{P_e}[Y^e|Pa(Y)] = E_P[Y|Pa(Y)]$
2. $V[Y^e|Pa(Y)]$ is finite.

An environment is valid with respect to (G, P) if it can be created by a valid intervention.

Section 3: Methods 2) Invariant Risk Minimization (IRM)

Definition 3.2 Invariant Representation

A representation $\Phi(T, X)$ is invariant with respect to environments E if and only if

$$E[Y^{e1}|\Phi(T^{e1}, X^{e1}) = \pi] = E[Y^{e2}|\Phi(T^{e2}, X^{e2}) = \pi] \quad \text{for all } e_1, e_2 \in E$$

Definition 3.3 Invariant Representation via Predictor

$\Phi: X \to H$ elicits an invariant predictor across E if a classifier $\omega: H \to Y$ is optimal for all $e \in E$

$$w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min}\, R^e(\bar{w} \circ \Phi) \quad \text{for all } e \in \mathcal{E},$$

In this case, we can find an invariant predictor $Q^{inv} = \omega \circ \Phi(T^e, X^e) = E[Y|\Phi(T, X)]$ by solving the equation above for both ω and Φ.

Section 3: Methods 2) Invariant Risk Minimization (IRM)

Definition 3.4 IRMv1

$$\hat{\Phi} = \arg\min_{\Phi} \sum_{e \in \mathcal{E}} R^e(1.0 \cdot \Phi) + \lambda \parallel \nabla_{w|w=1.0} R^e(w \cdot \Phi) \parallel^2 .$$

- Simplest choice of classifier: $\omega = 1.0$
- The gradient norm penalizes model deviations from the optimal classifier in each environment
- Parameterize Φ with a neural network that takes $\{t_i^e, x_i^e\}$ as input and outputs a real number

$$\hat{R}^e(Q) = \frac{1}{n_e} \sum_i \ell(y_i^e, Q(t_i^e, x_i^e)).$$

- $\hat{Q}^{inv} = 1.0 \cdot \hat{\Phi}$ is an empirical estimate of $E[Y|\Phi(T, X)]$

Section 3: Methods 3) Nearly Invariant Causal Estimation (NICE)

Invariant representation across all valid environments :

- $Pa(Y)$ is the minimal information required for invariance
- A representation that is invariant over all valid environments will be minimal

Thus, an invariant representation must capture only the parents of Y

$$\psi = \mathbb{E}[\mathbb{E}\left[Y \mid T = 1, \mathrm{Pa}(Y) \setminus \{T\}\right] - \mathbb{E}\left[Y \mid T = 0, \mathrm{Pa}(Y) \setminus \{T\}\right)] \mid T = 1]$$

$$\text{Since } \mathbb{E}\left[Y|T, \mathrm{Pa}(Y) \setminus \{T\}\right] = \mathbb{E}\left[Y|\Phi(T, X)\right],$$

$$\psi = \mathbb{E}\left[\mathbb{E}\left[Y \mid \Phi(1, X)\right] - \mathbb{E}\left[Y \mid \Phi(0, X)\right] \mid T = 1\right].$$

Section 3: Methods 3) Nearly Invariant Causal Estimation (NICE)

Recall $Q^{inv}(T,X) = \omega \circ \Phi(T^e, X^e) = E[Y|\Phi(T,X)]$

The NICE procedure is:

1. Input multiple datasets $\mathcal{D}_e := \{(X_i^e, Y_i^e, T_i^e)\}_{i=1}^{n_e}$
2. Estimate the invariant predictor $\hat{Q}^{inv} = 1.0 \cdot \hat{\Phi}$ using an invariant objective, e.g.IRMv1
3. Compute $\hat{\psi}^e = \frac{1}{\sum_i t_i^e} \sum_{i:t_i^e=1} \hat{Q}^{\text{inv}}(1, x_i^e) - \hat{Q}^{\text{inv}}(0, x_i^e)$

Section 4: Justification of NICE

Theorem 4.2 Unbiased

Let L be a loss function such that the minimizer of the associated risk is a conditional expectation and let Φ be a representation that elicits a predictor that Q^{inv} is invariant for all valid environments.

Assuming X^e does not contain mediators between the treatment and the outcome, then

$$\psi^e = \mathbb{E}\left[Q^{\mathrm{inv}}(1, X^e) - Q^{\mathrm{inv}}(0, X^e) | T^e = 1\right].$$

Theorem 4.3 Overlap

Suppose $\epsilon \leq P(T^e = 1|X^e) \leq 1 - \epsilon$ with probability 1,

then $\epsilon \leq P(T^e = 1|\Phi(X^e)) \leq 1 - \epsilon$ with probability 1.

Even when the representation does not exclude the bad controls, invariance may remove at leastsome (if not all) collider dependence.

Section 5: Empirical Studies

Causal Estimands & Evaluation metrics

1. Sample Average Treatment effect on the Treated (SATT):

$$\psi_s = \frac{1}{\sum_i t_i} \sum_{i:t_i=1} (Q(1, Z(x_i)) - Q(0, Z(x_i)))$$

where $Z(x)$ is an admissible subset of X.

2. Mean Absolute Error (MAE):

$$\epsilon_{att} = |\hat{\psi}_s - \psi_s|$$

3. Comparison:

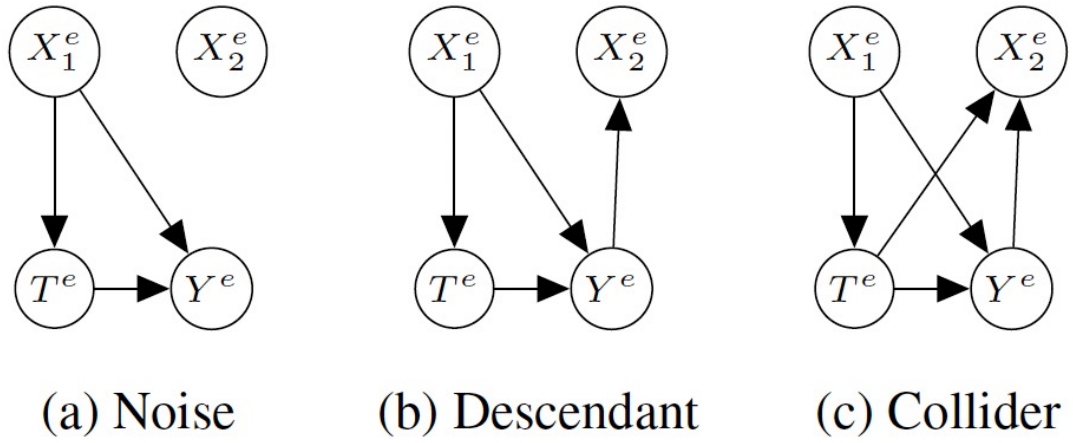
1) Adjusting for all covariates; 2) NICE;

3) Unadjusted; 4) Causal Discovery

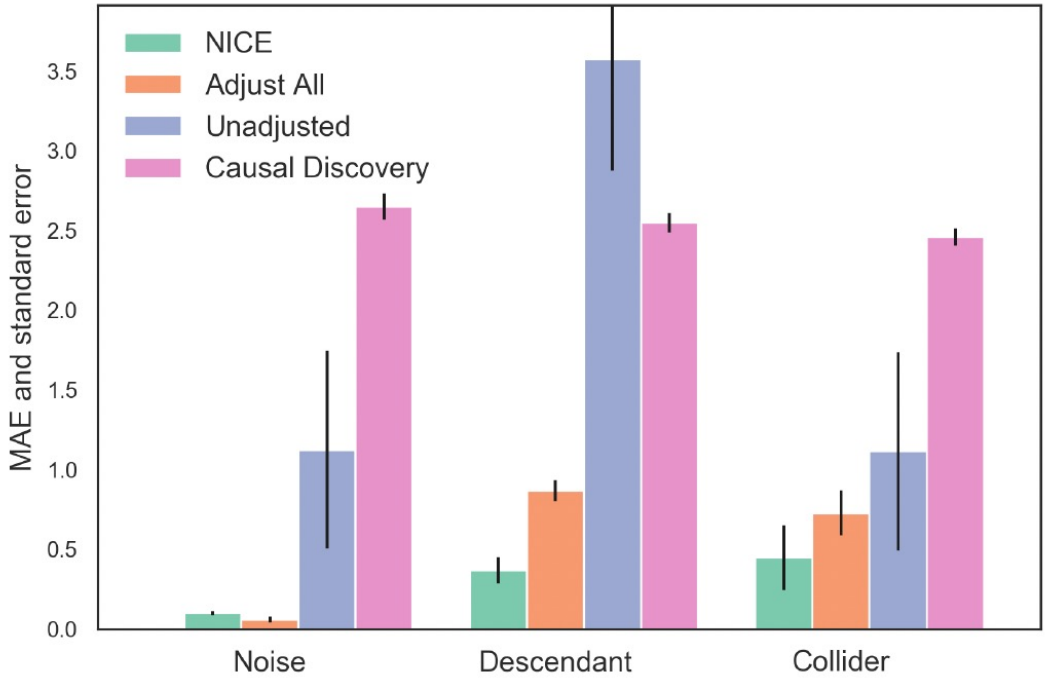
Section 5: Empirical Studies

Experiment 1: Linear setting datasets

- Theoretically guaranteed to strip out bad controls
- Predictor $\hat{Q}^{inv}$: OLS-2, linear regression with two separate regressors for the treated and the control population
- $X_e = (X_1^e, X_2^e)$, where X_1^e is a 5-dimensional confounder, X_2^e is either noise, a descendant, or a collider.



(a) Noise (b) Descendant (c) Collider



Section 5: Empirical Studies

Experiment 2: Non-linear setting datasets

- Predictor $\hat{Q}^{inv}$:

  2 neural network models
- Benchmark dataset:

  4 modifications of SpeedDating
- Simulate bad controls
- Metrics:

  average MAE and bootstrap std of SATT.

Valid Adjustment — ϵ_{att}

		Mod1	Mod2	Mod3	Mod4
Low	Adjust All	$.04 \pm .08$	$.05 \pm .09$	$.07 \pm .09$	$.01 \pm .01$
	NICE	$.07 \pm .03$	$.02 \pm .01$	$.09 \pm .03$	$.04 \pm .02$
Med	Adjust All	$.07 \pm .10$	$.05 \pm .05$	$.04 \pm .04$	$.07 \pm .08$
	NICE	$.05 \pm .02$	$.04 \pm .03$	$.05 \pm .03$	$.03 \pm .02$
High	Adjust All	$.07 \pm .07$	$.06 \pm .05$	$.06 \pm .07$	$.04 \pm .04$
	NICE	$.02 \pm .01$	$.06 \pm .03$	$.04 \pm .02$	$.07 \pm .04$

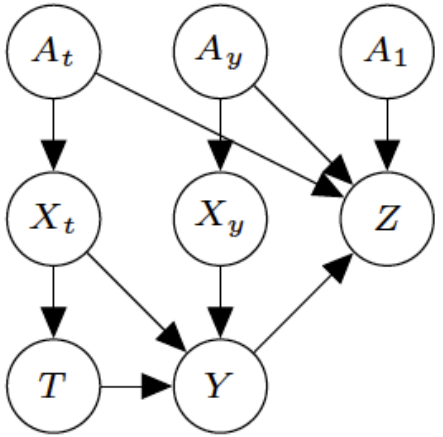
Bad Controls in Adjustment Set — ϵ_{att}

		Mod1	Mod2	Mod3	Mod4
low	Adjust All	$.26 \pm .09$	$.42 \pm .03$	$.34 \pm .08$	$.46 \pm .09$
	NICE	$.09 \pm .07$	$.03 \pm .01$	$.11 \pm .04$	$.08 \pm .04$
med	Adjust All	$.38 \pm .10$	$.35 \pm .06$	$.40 \pm .17$	$.3 \pm .09$
	NICE	$.06 \pm .03$	$.06 \pm .03$	$.06 \pm .02$	$.03 \pm .03$
high	Adjust All	$.32 \pm .14$	$.38 \pm .09$	$.42 \pm .05$	$.28 \pm .05$
	NICE	$.05 \pm .03$	$.11 \pm .03$	$.16 \pm .05$	$.11 \pm .05$

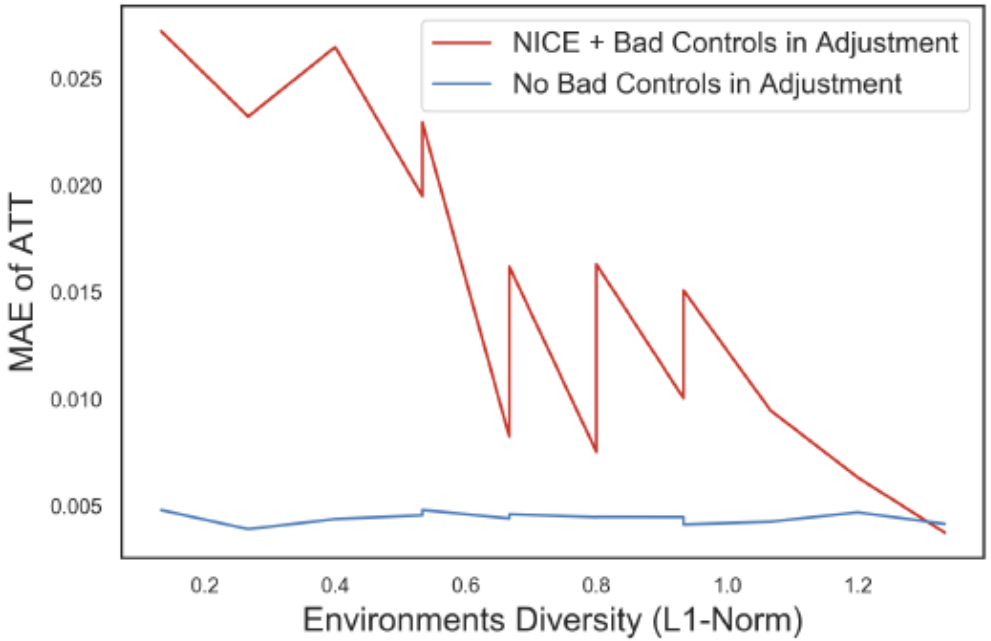
Section 5: Empirical Studies

Experiment 3: effect of environment variations

- Simulate non-linear data using causal graph, where adjustment set $\{X, A\}$ is valid, $\{X, A, Z\}$ is not valid.
- First draw 3 source environments $\{P^{e1}, P^{e2}, P^{e3}\}$, then construct new environments by draw different proportions from source environments.



- The more diverse the environments, the more likely that NICE can strip out bad controls and reduce bias.



Section 6: Conclusion and Discussion

Conclusions

- demonstrating how representation learning ideas can be harnessed to improve causal estimation
- NICE can reduce bias induced by the bad controls
- When there are no bad controls, NICE does not hurt the estimation quality.
- Whether NICE can strip out bad controls depends on the diversity of environments.

Comments

- NICE is essentially an application of robust estimation
- Lack of interpretability for causal inferences
- Assumptions: comprehensive, multiple environments, overlap…
- Adjusting for all covariates generally does not affect treatment effect estimates

References

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz (2019). "'**Invariant Risk Minimization**.'" In: arXiv preprint arXiv:1907.02893.

- Model generalization
- Algorithms for invariant risk minimization
- Invariance, causality and generalization

- Concludes with a Socratic dialogue discussing directions for future research.

6 Looking forward: a concluding dialogue

[ERIC and IRMA are two graduate students studying the Invariant Risk Minimization (IRM) manuscript. Over a cup of coffee at a café in Palais-Royal, they discuss the advantages and caveats that invariance brings to Empirical Risk Minimization (ERM).]

IRMA: I have observed that predictors trained with ERM sometimes absorb biases and spurious correlations from data. This leads to undesirable behaviours when predicting about examples that do not follow the distribution of the training data.

ERIC: I have observed that too, and I wonder what are the reasons behind such phenomena. After all, ERM is an optimal principle to learn predictors from empirical data!

IRMA: It is, indeed. But even when your hypothesis class allows you to find the empirical risk minimizer efficiently, there are some assumptions at play. First, ERM assumes that training and testing data are identically and independently distributed according to the same distribution. Second, generalization bounds require that the ratio between the capacity of our hypothesis class and the number of training examples n tends to zero, as $n \to \infty$. Third, ERM achieves zero test error only in the realizable case —that is, when there exists a function in our hypothesis class able to achieve zero error. I suspect that violating these assumptions leads ERM into absorbing spurious correlations, and that this is where invariance may prove useful.