Symmetry as intervention;

an analysis of causal effect estimation using outcome invariant data augmentation

Uzair Akbar[1], Niki Kilbertus[2,3], Hao Shen[2,4], Krikamol Muandet[5], Bo Dai[1,6]

[1]Georgia Tech [2]TU Munich [3]Helmholtz AI [4]Fortiss GmbH [5]RI Lab., CISPA [6]Google DeepMind

in/uzair25
@uzairakbar25
uzairakbar.com
arXiv:2510.25128

Can knowledge of **symmetries in data generation**—often used implicitly in certain regularizers—be repurposed to improve **causal effect estimation** given only observational (X, Y) data?

Statistical vs. Causal Estimation

Empirical Risk Minimisation (ERM):

Treatment X, outcome Y are generated as

$$Y = f(X) + \xi, \qquad \mathbb{E}[\xi] = 0, \qquad (\star)$$

statistical inference entails recovering the *optimal predictor* $\mathbb{E}[Y \mid X = \mathbf{x}]$ via minimiser h_{erm} of the risk:

$$R_{\text{erm}}(h) := \mathbb{E}\left[\left\| Y - h(X) \right\|^2 \right],$$

over hypotheses $h \in \mathscr{H}$, for a rich enough class $\mathscr{H}$.

Problem: For $\mathbb{E}[\xi \mid X = \mathbf{x}] \neq 0$, biased estimator of f.

Causal Estimation via Intervention:

Remove correlation b/w X, ξ, via an *intervention*—assign X independently sampled $X' \sim \mathbb{P}_X$ in data generation [1], represented by $\text{do}(X := X')$ or $\text{do}(X)$ as shorthand.

$$R^{\text{do}(X)}_{\text{erm}}(h) = R^{\text{do}(X:=X')}_{\text{erm}}(h), \qquad \text{s.t. } X' \sim \mathbb{P}_X.$$

Also define *causal estimation error (CER)* [2]:

$$\text{CER}(h) := R^{\text{do}(X)}_{\text{erm}}(h) - R^{\text{do}(X)}_{\text{erm}}(f).$$

Problem: Data generating process is unavailable.

Instrumental Variable (IV) Regression:

To workaround interventions, use *instrument* Z satisfying

i. exclusion $Z \perp\!\!\!\perp Y \mid X$, **ii.** un-confoundedness $Z \perp\!\!\!\perp \xi$,

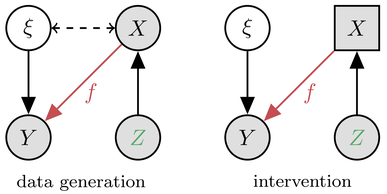
iii. X relevance: $Z \not\perp\!\!\!\perp X$. **iv.** Y relevance: $Y \not\perp\!\!\!\perp Z$,

Condition $(\star)$ on Z to get $\mathbb{E}[Y \mid Z] = \mathbb{E}[f(X) \mid Z]$.

Which can be solved for f by minimizing the risk

$$R_{\text{iv}}(h) := \mathbb{E}\left[\left\| Y - \mathbb{E}[h(X) \mid Z] \right\|^2 \right].$$

Problem: IVs are scarce in most application domains.



Known Data Symmetries and Invariance

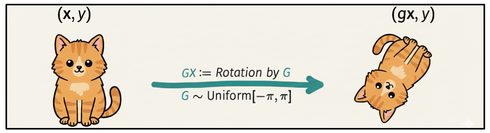
Data Augmentation (DA):

For finite n samples $\mathscr{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=0}^n$, *regularization* techniques are used to mitigate estimation *variance*.

E.g., *data augmentation (DA)* achieves this via multiple random augmentations $(G\mathbf{x}_i, \mathbf{y}_i)$ per sample in the risk

$$R_{\text{da+erm}}(h) := \mathbb{E}\left[\left\| Y - h(GX) \right\|^2 \right], \quad G \sim \mathbb{P}_G.$$

DA = Model Symmetries:



Assume G takes values in $\mathscr{G}$ such that f is $\mathscr{G}$*-invariant*:

$$f(\mathbf{x}) = f(\mathbf{gx}), \qquad \forall\ (\mathbf{x}, \mathbf{g}) \in \mathscr{X} \times \mathscr{G}.$$

Example: classifying images $\mathbf{x} \in \mathscr{X}$ of cats vs. dogs, the true classifier f is invariant to random image rotations $G\mathbf{x}$.

Causal Estimation w/ Data Augmentation

Question: For $\mathscr{G}$*-invariant* f, can DA improve CER?

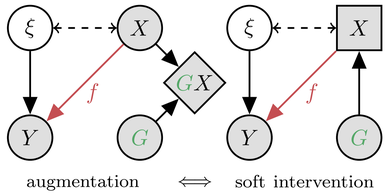
Why bother? Robust prediction under shifts in $\mathbb{P}_X$ [3].

First, we specify a linear, Gaussian model for our analysis:

$$Y = \mathbf{f}^\top X + \xi, \quad \mathbf{f} \in \mathbb{R}^m, \quad (X, \xi) \sim \mathcal{N}(\mathbf{0}, \cdot), \qquad (\dagger)$$

And the following DA model with *strength* $\gamma > 0$.

$$GX := X + \gamma \cdot G \qquad G \in \text{null}(\mathbf{f}). \qquad (\ddagger)$$



DA = Soft Intervention:

For $\mathscr{G}$-invariant f, (Y, GX) follows the data generation:

$$Y = f(GX) + \xi.$$

⇒ Soft intervention on the treatment!

For the linear, Gaussian setting of $(\dagger), (\ddagger)$, we show:

⇒ DA+ERM *dominates* ERM on causal estimation:

$$\text{CER}(\mathbf{h}_{\text{da+erm}}) \leq \text{CER}(\mathbf{h}_{\text{erm}}),$$

equality iff $\mathbb{E}[GX \mid G] \perp_{\text{a.s.}} \mathbb{E}[X \mid \xi]$.

- Strictly better if DA perturbs spurious features of X.
- But never performs worse than ERM otherwise.

DA = Relaxed IVs:

For $\mathscr{G}$-invariant f, DA parameters G are *IV-like (IVL)*—meaning they have IV properties **(i)—(iii)** by construction.

Such a relaxation renders IV regression **ill-posed**, so we suggest an ERM regulariser in the form of *IVL regression*:

$$R_{\text{ivl}}(h) := R_{\text{iv}}(h) + \alpha \cdot R_{\text{erm}}(h)$$

Use G as an IV

ERM regulariser for ill-posed IV reg.

For the linear, Gaussian setting of $(\dagger), (\ddagger)$, we show:

⇒ The composition DA+IVL simulates a *worst-case, or adversarial* application of DA.

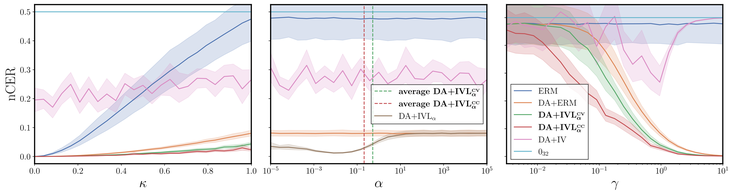
⇒ DA+IVL *dominates* DA+ERM on causal estimation:

$$\text{CER}(\mathbf{h}_{\text{da+ivl}}) \leq \text{CER}(\mathbf{h}_{\text{da+erm}}),$$

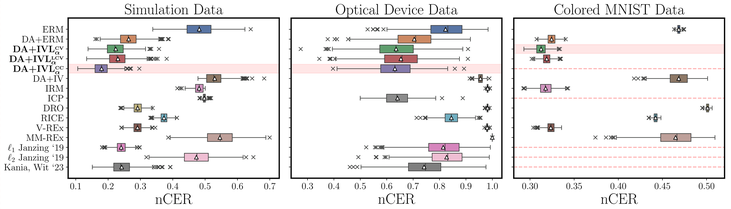
equality iff $\mathbb{E}[GX \mid G] \perp_{\text{a.s.}} \mathbb{E}[X \mid \xi]$.

Experiments

Normalized CER (nCER) $= 0$ for true f and 1 for pure confounding.



Simulation exp. for $(\dagger), (\ddagger)$ with *confounding strength* $\kappa > 0$.



Comparison with select causal regularization and domain generalisation baselines. All methods provided only (X, Y) data and DA transformations G—Gausian noise in optical device, and hue, saturation, contrast, random translations in colored-MNIST.

Necessity + practicality of prior knowledge

Untestable assumptions are fundamental for causal claims from observational data. To justify the ones we make, we compare with IVs:

$$\underbrace{\mathscr{G}\text{-invariance}}_{\text{popular model for DA}} + \underbrace{\text{spurious targets}}_{\text{benign failure if violated}} \iff \underbrace{\text{un-confoundedness} + \text{exclusion}}_{\text{un-testable IV/IVL assumptions}}$$

un-testable DA assumptions

reframing [4]

References

[1] J. Peter, 2017.
[2] C. Vankadar, 2022.
[3] D. Rothenhäusler, 2020.
[4] R. Christianse, 2022