Symmetry as Intervention; Causal Estimation with Data Augmentation

Uzair Akbar * UZAIR.AKBAR@GATECH.EDU
Georgia Institute of Technology

Niki Kilbertus NIKI.KILBERTUS@TUM.DE
Technical University of Munich & Helmholtz AI

Hao Shen SHEN@FORTISS.ORG
Technical University of Munich & Fortiss GmbH

Krikamol Muandet MUANDET@CISPA.DE
Rational Intelligence Lab, CISPA

Bo Dai BODAI@GOOGLE.COM
Georgia Institute of Technology & Google DeepMind

Code: https://github.com/uzairakbar/causal-data-augmentation
Project Page: https://uzairakbar.com/causal-data-augmentation

Abstract

To our knowledge, we provide the first analysis of causal estimation under hidden confounding using only observational (X, Y) data and knowledge of symmetries in data generation via data augmentation (DA) transformations. We show that such DA is equivalent to interventions on the treatment X, mitigating bias from hidden confounding, and that framing DA as a relaxation of instrumental variables (IVs)—sources of X randomization that are conditionally independent of the outcome Y—can further improve causal estimation beyond simple DA.

Keywords: Causal Inference, Intervention, IV Regression, Invariance, Data Augmentation

1. Preliminaries

For treatment $X \in \mathcal{X} \subseteq \mathbb{R}^m$, outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}^l$ in the *structural equation model (SEM)* $\mathfrak{M}$

$$X = \tau(Y, Z, C, N_X), \quad Y = f(X) + \epsilon(C) + N_Y, \quad \text{s.t.} \quad \boxed{\xi := Y - f(X), \quad \mathbb{E}[\xi] = 0,} \tag{1}$$

where Z, C, N_X, N_Y are exogenous, we want to estimate $f \in \mathcal{H} := \{h : \mathcal{X} \to \mathcal{Y}\}$ from $\mathbb{P}^{\mathfrak{M}}_{X,Y}$.[1]

When $X \perp\!\!\!\perp \xi$, we estimate f via *empirical risk minimization (ERM)* given a convex loss ℓ,

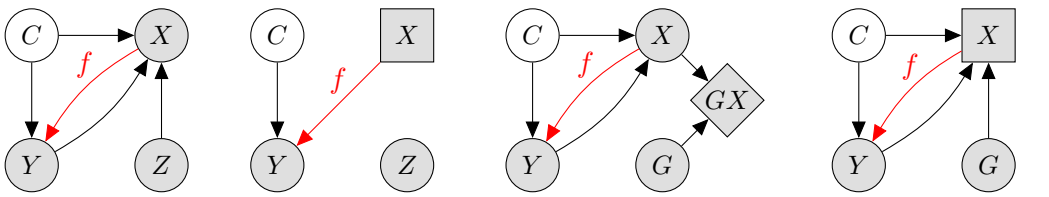
$$R^{\mathfrak{M}}_{\text{ERM}}(h) := \mathbb{E}^{\mathfrak{M}}[\ell(Y, h(X))], \qquad \hat{h}^{\mathfrak{M}}_{\text{ERM}} := \text{argmin}_{h \in \mathcal{H}} R^{\mathfrak{M}}_{\text{ERM}}(h). \tag{2}$$

For finite n samples $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=0}^n$, *data augmentation (DA)* is used to reduce estimation variance (Lyle et al., 2020) via multiple random augmentations $(G\mathbf{x}_i, \mathbf{y}_i)$ per sample in the risk

$$R^{\mathfrak{M}}_{\text{DA}_G+\text{ERM}}(h) := \mathbb{E}^{\mathfrak{M}}[\ell(Y, h(GX))], \qquad G \sim \mathbb{P}_G. \tag{3}$$

* Part of this work was done while at Max Planck Institute for Intelligent Systems and TU Munich.
. This work is a non-archival summary of Akbar et al. (2025), for discussion at the NeurReps Workshop.
1. Assume all SEMs under discussion entail unique observational distributions $\mathbb{P}^{\mathfrak{M}}_{X,Y}$. Details in Appendix B.

© 2025 U. Akbar, N. Kilbertus, H. Shen, K. Muandet & B. Dai.



(a) $\mathfrak{M}$. (b) $\mathfrak{M}; \mathrm{do}(X)$. (c) $\mathfrak{A}$ post DA. (d) $\mathfrak{A}; \mathrm{do}(\tau := G\tau)$.

Figure 1: Graphs of respective SEMs; (a) Z is an IV w.r.t. confounded (X, Y). (b) Graph obtained via intervention on X in $\mathfrak{M}$; IV regression simulates this intervention with only observational data. (c) Graph for DA. (d) Graph for soft intervention. Observational distributions of (GX, Y, G, C) in (c) and (X, Y, G, C) in (d) are identical.

However, generally $X \not\perp\!\!\!\perp \xi$ due to which the ERM minimizer is biased (Pearl, 2009). This bias is known as the *confounding bias* and X, Y are said to be confounded. Confounding is removed via an *intervention* $\mathrm{do}(X := X')$ that sets X to some i.i.d. X', now yielding the *causal risk*

$$R^{\mathfrak{M}}_{\mathrm{CR}}(h) := R^{\mathfrak{M};\mathrm{do}(X)}_{\mathrm{ERM}}(h) = R^{\mathfrak{M};\mathrm{do}(X:=X')}_{\mathrm{ERM}}(h), \qquad \text{s.t.} \qquad X' \sim \mathbb{P}^{\mathfrak{M}}_X. \tag{4}$$

Where $\mathrm{do}(X)$ denotes such interventions. Minimizers of Eq. (4) identify f and are robust predictors to $\mathbb{P}^{\mathfrak{M}}_X$ shifts over $\mathrm{supp}(\mathbb{P}^{\mathfrak{M}}_X)$ (Christiansen et al., 2022). Define *causal excess risk (CER)*

$$\mathrm{CER}_{\mathfrak{M}}(h) := R^{\mathfrak{M}}_{\mathrm{CR}}(h) - R^{\mathfrak{M}}_{\mathrm{CR}}(f),$$

to capture estimation error by removing irreducible noise from Eq. (4), so that $\mathrm{CER}_{\mathfrak{M}}(f) = 0$.

In practice, interventions are often unavailable. A common workaround is to use auxiliary variables. One approach is that of *instrumental variable (IV) regression* (Belsley, 1988), where an instrument Z satisfies: (i) **treatment relevance** $Z \not\perp\!\!\!\perp X$, (ii) **exclusion** $Z \perp\!\!\!\perp Y^{\mathfrak{M};\mathrm{do}(X)}$, (iii) **un-confoundedness** $Z \perp\!\!\!\perp \xi$, and (iv) **outcome relevance** $Y \not\perp\!\!\!\perp Z$. Now, Eq. (1) gives

$$\mathbb{E}^{\mathfrak{M}}[Y \mid Z] = \mathbb{E}^{\mathfrak{M}}[f(X) \mid Z]. \tag{5}$$

which admits consistent estimation of f and can be solved by minimizing the following risk

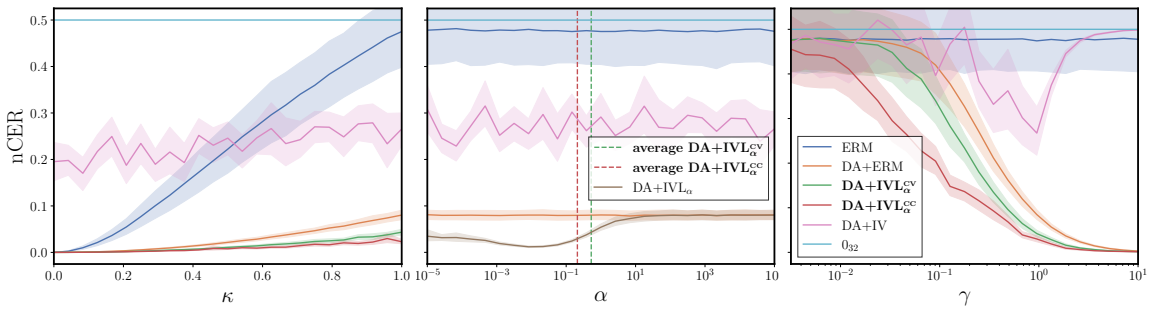
$$R^{\mathfrak{M}}_{\mathrm{IV}}(h) := \mathbb{E}^{\mathfrak{M}}\Big[\ell\Big(Y,\ \mathbb{E}^{\mathfrak{M}}[h(X) \,|\, Z]\Big)\Big]. \tag{6}$$

2. Causal Estimation with Data Augmentation

Problem setup. We discuss the following SEM $\mathfrak{A}$ for exogenous C, N_X, N_Y and $X \not\perp\!\!\!\perp \xi$,

$$X = \tau(Y, C, N_X), \qquad Y = f(X) + \epsilon(C) + N_Y, \qquad \text{s.t.} \qquad \boxed{\xi := Y - f(X), \qquad \mathbb{E}[\xi] = 0.} \tag{7}$$

Consider also a data augmentation with respect to which f is invariant (Chen et al., 2020). The action of a group $\mathcal{G}$ is a mapping $\delta : \mathcal{X} \times \mathcal{G} \to \mathcal{X}$ compatible with the group operation. We write $\mathbf{gx} := \delta(\mathbf{x}, \mathbf{g})$ as shorthand and say that f is *$\mathcal{G}$-invariant* if $f(\mathbf{gx}) = f(\mathbf{x})$, $\forall\, (\mathbf{g}, \mathbf{x}) \in \mathcal{G} \times \mathcal{X}$. We refer to such a map $\mathbf{gx}$, henceforth assumed to be continuous in $\mathbf{x}$, as a valid *outcome-invariant* DA transformation parameterized by the vector $\mathbf{g} \in \mathcal{G}$. Let $\mathcal{G}$ have a (unique) normalized Haar measure and $\mathbb{P}^{\mathfrak{A}}_G$ the corresponding distribution defined over it.



(a) $\gamma = 1$ (b) $\gamma, \kappa = 1$ (c) $\kappa = 1$

Figure 2: Simulation experiment for the linear Gaussian SEM in Example 1. κ and γ control the amount of confounding and *strength* of DA respectively. α is the IVL regularization parameter. Each data-point averages nCER over 32 trials with a 95% CI.

The task. Given samples for only $(X, Y) \sim \mathbb{P}^{\mathfrak{A}}_{X,Y}$ and a valid outcome invariant DA parameterized by $G \sim \mathbb{P}^{\mathfrak{A}}_{G}$, we want to improve estimation of f compared to standard ERM.

Now, take a *soft* intervention on $\mathfrak{A}$ where we replace the mechanism τ of X with $G\tau$. Abusing notation, we represent this SEM by $\mathfrak{A}; \mathrm{do}(\tau := G\tau)$, its graph depicted in Fig. 1(*d*).[2] Comparing the DA mechanism in $\mathfrak{A}$ (Fig. 1(*c*)) and the intervention $\mathfrak{A}; \mathrm{do}(\tau := G\tau)$ (Fig. 1(*d*)):

Observation 1 (soft intervention with DA) $\mathbb{P}^{\mathfrak{A}}_{GX,Y,G,C}$ *and* $\mathbb{P}^{\mathfrak{A};\mathrm{do}(\tau:=G\tau)}_{X,Y,G,C}$ *are identical.*

We can hence treat samples generated from $\mathfrak{A}$ via DA as if they were instead generated from $\mathfrak{A}; \mathrm{do}(\tau := G\tau)$ by intervening on X. This allows us to rewrite the risk from Eq. (3) as $R^{\mathfrak{A}}_{\mathrm{DA}_G+\mathrm{ERM}}(h) = R^{\mathfrak{A};\mathrm{do}(\tau:=G\tau)}_{\mathrm{ERM}}(h)$, to emphasize that DA is equivalent to a (soft) intervention and as such can mitigate confounding bias when estimating f, as shown in the next example.

Example 1 (a linear Gaussian DA example) *For $\kappa, \sigma > 0$, non-zero $\mathbf{\Gamma}, \mathbf{T} \in \mathbb{R}^{*\times m}$ and $\boldsymbol{\tau}^\top, \mathbf{f}, \boldsymbol{\epsilon} \in \mathbb{R}^m$ such that $\mathbf{f}^\top \boldsymbol{\tau}^\top \neq \kappa^{-1}$ so that the following SEM $\mathfrak{A}$ is solvable in (X, Y)*[3]

$$X = \kappa \cdot \boldsymbol{\tau}^\top Y + \mathbf{T}^\top C + \sigma N_X, \qquad Y = \mathbf{f}^\top X + \kappa \cdot \boldsymbol{\epsilon}^\top C + \sigma N_Y, \qquad GX := X + \gamma \cdot \mathbf{\Gamma}^\top G,$$

where G, C, N_X, N_Y are conformable, centered Gaussian vectors, κ determines how much (X, Y) are confounded and $\mathrm{range}(\mathbf{\Gamma}^\top) \subseteq \mathrm{null}(\mathbf{f}^\top)$ *to make GX a valid outcome invariant DA.*

We evaluate an estimate $\hat{\mathbf{h}}^{\mathcal{D}}$ using CER. For squared loss and covariance $\mathbf{\Sigma}^{\mathfrak{A}}_X$ in Example 1,

$$\mathrm{CER}_{\mathfrak{A}}\left(\hat{\mathbf{h}}^{\mathcal{D}}\right) = \left\|\hat{\mathbf{h}}^{\mathcal{D}} - \mathbf{f}\right\|^2_{\mathbf{\Sigma}^{\mathfrak{A}}_X}. \tag{8}$$

Theorem 1 (causal estimation with DA+ERM) *For SEM $\mathfrak{A}$ in Example 1, we have*

$$\mathrm{CER}_{\mathfrak{A}}\left(\hat{\mathbf{h}}^{\mathfrak{A}}_{DA_G+ERM}\right) \leq \mathrm{CER}_{\mathfrak{A}}\left(\hat{\mathbf{h}}^{\mathfrak{A}}_{ERM}\right), \quad \textit{equality iff} \quad \mathbb{E}^{\mathfrak{A}}[GX \,|\, G] \perp \mathbb{E}^{\mathfrak{A}}[X \,|\, \xi] \quad \textit{a.s.}$$

Proof See Appendix H.3 for the proof. ■

2. For any $\mathfrak{A}$ with unique distribution, $\mathfrak{A}; \mathrm{do}(\tau := G\tau)$ also has a unique distribution (proof in Appendix H.2).
3. See Appendix B, Lemma 3 for details on solving for and sampling (X, Y) in such linear, simultaneous SEMs.

That is, DA improves causal estimation iif it targets spurious features of X. Domain knowledge may therefore be needed to design such DA. Still, with outcome invariance, DA is never worse than ERM; allowing regularization at worst, and mitigating confounding bias at best.

We once again point our attention to the graph of $\mathfrak{A}; \mathrm{do}(\tau := G\tau)$ from Fig. 1(*d*) to see:

Observation 2 (IV-like DA parameters) *In SEM $\mathfrak{A}; \mathrm{do}(\tau := G\tau)$, the DA parameters G satisfy IV properties (i) through (iii). We refer to such an IV relaxation as* IV-like (IVL).

This IV relaxation may render an ill-posed Eq. (5), so we suggest the regularization $R^{\mathfrak{M}}_{\mathrm{IVL}_\alpha}(h) := R^{\mathfrak{M}}_{\mathrm{IV}}(h) + \alpha R^{\mathfrak{M}}_{\mathrm{ERM}}(h)$ as *IVL regression*, discussed separately in Appendix E. When composed with DA in $\mathfrak{A}$ now gives $R^{\mathfrak{A}}_{\mathrm{DA}_G+\mathrm{IVL}_\alpha}(h) = R^{\mathfrak{A};\mathrm{do}(\tau:=G\tau)}_{\mathrm{IVL}_\alpha}(h)$. The next results follow.

Corollary 1 (worst-case DA with DA+IVL regression) *For SEM $\mathfrak{A}$ in Example 1,*

$$\hat{\mathbf{h}}^{\mathfrak{A}}_{DA_G+IVL_\alpha} \in \operatorname*{argmin}_{\mathbf{h}} \max_{\mathbf{g}\in\mathcal{G}_\alpha} R^{\mathfrak{A}}_{DA_{\mathbf{g}}+ERM}(\mathbf{h}), \quad s.t. \quad \mathcal{G}_\alpha := \left\{ \mathbf{g} \,\middle|\, \mathbf{\Gamma}^\top \mathbf{g}\mathbf{g}^\top \mathbf{\Gamma} \preccurlyeq \left(\frac{1}{\alpha}+1\right) \mathbf{\Gamma}^\top \mathbf{\Sigma}^{\mathfrak{A}}_G \mathbf{\Gamma} \right\}.$$

Proof The result follows from Observation 1, Observation 2 and Theorem 2. ■

Corollary 2 (causal estimation with DA+IVL regression) *In Example 1, $\alpha, \gamma < \infty$,*

$$\mathrm{CER}_{\mathfrak{A}}\left(\hat{\mathbf{h}}^{\mathfrak{A}}_{DA_G+IVL_\alpha}\right) \leq \mathrm{CER}_{\mathfrak{A}}\left(\hat{\mathbf{h}}^{\mathfrak{A}}_{DA_G+ERM}\right), \quad \textit{equality iff} \quad \mathbb{E}^{\mathfrak{A}}[GX\,|\,G] \perp \mathbb{E}^{\mathfrak{A}}[X\,|\,\xi] \quad a.s.$$

Proof The result follows directly from Theorem 3 and Observation 2. ■

Using DA parameters as IVL therefore simulates a worst-case, or adversarial application of DA within a set of transforms $\mathcal{G}_\alpha$. Of course Corollary 1 can also be viewed as a predictor that generalizes to treatment interventions encoded by $\mathcal{G}_\alpha$. As is intuitive, such a worst-case intervention improves causal estimation so long as the features of X intervened along include some that are spurious (Corollary 2). DA and IVL regression may therefore be used in composition if the application can benefit from regularization and/ or better prediction generalization across DA-induced interventions, with a "bonus" of lower confounding bias if the DA also augments any spurious features of X. The Appendix covers limitations and related work.

3. Experiments

We empirically evaluate DA's effectiveness in reducing hidden confounding bias in the finite-sample regime. Since our focus is on generalizing across interventions rather than i.i.d. generalization, we fix augmented data size to match the original throughout all experiments.

We compare against standard ERM, DA, IV regression, and re-purposed *domain generalization* methods including DRO, IRM, ICP, RICE, V-REx, MM-REx, and causal regularization approaches. For methods requiring additional variables, we replace these with DA parameters G (see Appendix G for implementation details and detailed analysis).

For better interpretability of results, we evaluate using normalized CER (nCER): $\mathrm{nCER}_{\mathfrak{M}}(h) = \frac{\mathrm{CER}_{\mathfrak{M}}(h)}{\mathrm{CER}_{\mathfrak{M}}(h)+\mathrm{CER}_{\mathfrak{M}}(h_0)} \in [0,1]$, where h_0 represents null treatment effect. This has the property that nCER $= 0$ for ground-truth causal solution but 1 under pure confounding.

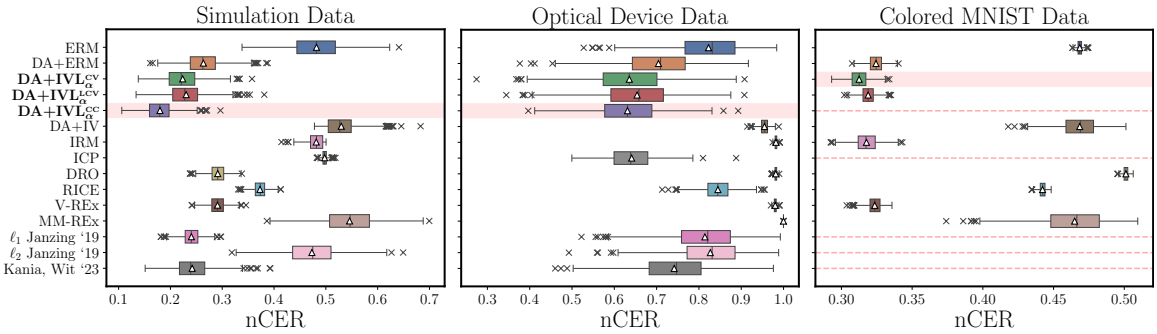


Figure 3: Experiment results; common domain generalisation benchmarks compared against the ERM, DA+ERM and DA+IV baselines, including DA+IVL.

Simulation experiment. Using the linear SEM from our theory with $m = 32$, $n = 2048$ samples across 32 experiments, we find: (1) ERM degrades with increasing confounding κ, (2) DA alone improves performance, (3) DA+IVL achieves best results while DA+IV is unstable. The cross validation approaches of CC, CV and LCV are explained in Appendix E.

Optical device dataset. On 1000 samples across 12 datasets where hidden confounders affect both webcam-captured images and photo-diode readings, DA+ERM improves over ERM, with DA+IVL outperforming other baselines.

Colored MNIST. Where training labels spuriously correlate with color but correlation flips at test, DA via perturbations to hue/brightness helps reduce confounding. DA+ERM provides substantial gains over ERM, with DA+IVL achieving competitive performance

4. Conclusion

We conclude that re-purposing the widely used variance reduction tool of data augmentation (DA) for reducing hidden confounding bias can be effective under outcome invariance. Crucially, it offers a "no-regret" choice for practitioners; improving causal estimation when targeting spurious features, yet performing no worse than the ERM baseline otherwise. Such mitigation of hidden confounding has direct positive implications for the downstream tasks of robust prediction across shifts in $\mathbb{P}_X^{\mathfrak{A}}$ (Reddy et al., 2025), tighter bounds in partial identification (Kilbertus et al., 2020), and more informative sensitivity analyses De Bartolomeis et al. (2024).

References

Uzair Akbar, Niki Kilbertus, Hao Shen, Krikamol Muandet, and Bo Dai. An analysis of causal effect estimation using outcome invariant data augmentation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ahmed Aloui, Juncheng Dong, Cat Phuoc Le, and Vahid Tarokh. CATE estimation with potential outcome imputation from local regression. In *Conference on Uncertainty in Artificial Intelligence*, volume 286, 2025.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. arXiv:1907.02893.

Núria Armengol Urpí, Marco Bagatella, Marin Vlastelica, and Georg Martius. Causal action influence aware counterfactual data augmentation. In *International Conference on Machine Learning*, volume 235, 2024.

David A. Belsley. Two-or three-stage least squares? *Computer Science in Economics and Management*, 1:21–30, 1988. doi: 10.1007/BF00435200.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009. doi: 10.1515/9781400831050.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, second edition, 2009.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5), 2021. doi: 10.1214/21-AOS2064.

Peter Bühlmann and Dominik Cevid. Deconfounding and causal regularisation for stability and external validity. *International Statistical Review*, 88(S1):S114–S134, 2020. doi: 10.1111/insr.12383.

Ahsan J. Cheema, Katherine L. Marks, Hamzeh Ghasemzadeh, Jarrad H. Van Stan, Robert E. Hillman, and Daryush D. Mehta. Characterizing vocal hyperfunction using ecological momentary assessment of relative fundamental frequency. *Journal of Voice*, 2024. ISSN 0892-1997. doi: 10.1016/j.jvoice.2024.10.025.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Carl F. Christ. The Cowles Commission's contributions to econometrics at Chicago, 1939-1955. *Journal of Economic Literature*, 32(1):30–59, 1994.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2022. doi: 10.1109/TPAMI.2021.3094760.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *International Conference on Artificial Intelligence and Statistics*, volume 54, 2017.

Hugh Dance and Benjamin Bloem-Reddy. Counterfactual cocycles: A framework for robust and coherent counterfactual transports, 2025. arXiv:2405.13844.

Piersilvio De Bartolomeis, Javier Abad Martinez, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR, 2024.

Y. Dubois et al. Lossy compression for lossless prediction. In *Advances in Neural Information Processing Systems*, 2021.

Mordecai Ezekiel. The Cobweb theorem. *The Quarterly Journal of Economics*, 52(2), 1938. doi: 10.2307/1881734.

A. Fawzi and P. Frossard. Manitest: Are classifiers really invariant? In *British Machine Vision Conference*, 2015.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David Blei. Data augmentations for improved (large) language model generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

John Fox. Simultaneous equation models and two-stage least squares. *Sociological Methodology*, 10:130–150, 1979. doi: 10.2307/270769.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Alastair R. Hall. Generalized method of moments. In *A Companion to Theoretical Econometrics*, chapter 11, pages 230–255. Wiley, 2003. doi: 10.1002/9780470996249.ch12.

Seunghyup Han, Osama Waqar Bhatti, Woo-Jin Na, and Madhavan Swaminathan. Reinforcement learning applied to the optimization of power delivery networks with multiple voltage domains. In *2023 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO)*, 2023. doi: 10.1109/NEMO56117.2023.10202224.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018. doi: doi:10.1515/jci-2017-0016.

Tom Heskes. Bias-variance decompositions: The exclusive privilege of Bregman divergences, 2025. arXiv:2501.18581.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

Antti Hyttinen, Frederick Eberhardt, and Patrik O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439, 2012.

Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*, volume 139, 2021.

Dominik Janzing. Causal regularization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.

Sravan Jayanthi, Letian Chen, Nadya Balabanska, Van Duong, Erik Scarlatescu, Ezra Ameperosa, Zulfiqar Haider Zaidi, Daniel Martin, Taylor Keith Del Matto, Masahiro Ono, and Matthew Gombolay. DROID: Learning from offline heterogeneous demonstrations via reward-policy distillation. In *Conference on Robot Learning*, volume 229. PMLR, 2023.

John Johnston. *Econometric Methods*. McGraw-Hill, second edition, 1971.

Lucas Kania and Ernst Wit. Causal regularization: On the trade-off between in-sample risk and out-of-sample risk guarantees, 2023. arXiv:2205.01593.

Niki Kilbertus, Matt J. Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning*, 2021.

Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Conference on Uncertainty in Artificial Intelligence*, pages 366–374. AUAI Press, 2008.

Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):321–348, 2002. doi: 10.1111/1467-9868.00340.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments, 2018. arXiv:1803.07164.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks, 2020. arXiv:2005.00178.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, 2021.

Arash Mastouri, Yuhang Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, volume 139, 2021.

O. Montasser et al. Transformation-invariant learning and theoretical guarantees for OOD generalization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Joris M. Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, volume 28, 2013.

John F. Muth. Rational expectations and the theory of price movements. *Econometrica*, 29 (3):315–335, 1961.

Michael Oberst, Nikolaj Thams, Jonas Peters, and David Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, volume 139, 2021.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press, 2017.

M. Petrache and S. Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems*, 2023.

Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. MoCoDA: Model-based counterfactual data augmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Abbavaram Gowtham Reddy, Celia Rubio-Madrigal, Rebekka Burkholz, and Krikamol Muandet. When shift happens - confounding is to blame, 2025. arXiv:2505.21422.

Tongzheng Ren, Haotian Sun, Antoine Moulin, Arthur Gretton, and Bo Dai. Spectral representation for causal estimation with hidden confounders. In *International Conference on Artificial Intelligence and Statistics*, 2025.

D. Romero and S. Lohit. Learning partial equivariances from data. In *Advances in Neural Information Processing Systems*, 2022.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

H. Shao et al. A theory of PAC learnability under transformation invariances. In *Advances in Neural Information Processing Systems*, 2022.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chennuru Vankadara, Luca Rendsburg, Ulrike von Luxburg, and Debarghya Ghoshdastidar. Interpolation and regularization for causal learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.00047.

S. Wong et al. Understanding data augmentation for classification: When to warp? In *Digital Image Computing: Techniques and Applications*, 2016.

Jefrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data.* The MIT Press, 2010.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

Fanny Yang, Zuowen Wang, and Christina Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jianhao Yuan, Francesco Pinto, Adam Davies, and Philip Torr. Not Just Pretty Pictures: Toward interventional data augmentation using text-to-image generators. In *International Conference on Machine Learning*, 2024.

Arnold Zellner and H. Theil. Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica*, 30(1):54–78, 1962. doi: 10.2307/1911287.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1), 2023. doi: 10.1515/jci-2022-0073.

S. Zhu et al. Understanding the generalization benefit of model invariance from a data perspective. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Appendix Contents

A Confounding Bias **13**

B Simultaneity as Cyclic Structures in Equilibrium **15**

C Related Work **18**

D IV Regression **20**

E IV-like Regression **22**

F Limitations **24**

G Experiment Supplement **25**
G.1 Simulation experiment 25
G.2 Optical device experiment 26
G.3 Colored-MNIST experiment 28

H Proofs **31**
H.1 Proof of Proposition 1 – IVL regression closed form solution in the linear case 31
H.2 Proof of Proposition 2 – Existence of interventional distribution far a DA . . 32
H.3 Proof of Theorem 1 – Causal estimation with DA+ERM 34
H.4 Proof of Theorem 2 – Robust prediction with IVL regression 35
H.5 Proof of Theorem 3 – Causal estimation with IVL regression 37
H.6 Miscellaneous supporting lemmas 39

List of Symbols

The notation is largely borrowed from Peters et al. (2017), with overloading where necessary.

$\mathbb{R}^n$	n–dimensional Euclidean space.
$\mathbb{R}^{n\times *}$	$n \times *$ Euclidean space; dimension $*$ conformal with & inferred from context.
x	Scalar.
$\mathbf{x}$	Vector. When $\mathbf{x}^\top$ is described as a vector, it means $\mathbf{x}$ is a flat $1 \times *$ matrix.
$\mathbf{X}$	Matrix.
$\mathcal{X}$	Set.
X	Random vector.
$\mathfrak{M}$	SEM.
$X^{\mathfrak{M}}$	Random vector X with its SEM $\mathfrak{M}$ specified when unclear from context.
$\mathbb{P}_X^{\mathfrak{M}}$	Distribution of X entailed by $\mathfrak{M}$. Superscript dropped if clear from context.
$\mathbb{E}^{\mathfrak{M}}[\,X\,]$	Expected value of X under distribution $\mathbb{P}_X^{\mathfrak{M}}$.
$\mathbf{\Sigma}_X^{\mathfrak{M}}$	Variance–covariance matrix of X under distribution $\mathbb{P}_X^{\mathfrak{M}}$.
$\mathbf{\Sigma}_{X,Y}^{\mathfrak{M}}$	Cross–covariance matrix of X and Y under distribution $\mathbb{P}_{X,Y}^{\mathfrak{M}}$.
$\mathrm{do}(X := \mathbf{x})$	Intervention — X is set to $\mathbf{x}$.
$\mathrm{do}(X)$	Shorthand for $\mathrm{do}(X := X')$ where $X' \sim \mathbb{P}_X^{\mathfrak{M}}$ is i.i.d. to X.
$\mathfrak{M}; \mathrm{do}(X := \mathbf{x})$	Intervention SEM.
$\mathfrak{M}_{X=\mathbf{x}}$	SEM with mechanisms of $\mathfrak{M}$, but exogenous noise distribution $\mathbb{P}_{N\vert X=\mathbf{x}}^{\mathfrak{M}}$.
$\mathfrak{M}_{Y=\mathbf{y}}; \mathrm{do}(X := \mathbf{x})$	Counterfatual SEM—intervention SEM of $\mathfrak{M}_{Y=\mathbf{y}}$.
$X \perp\!\!\!\perp Y$	Random vectors X, Y are statistically independent, i.e. $\mathbb{P}_{Y\vert X}^{\mathfrak{M}} = \mathbb{P}_Y^{\mathfrak{M}}$.
$\mathbf{x} \perp \mathbf{y}$	$\mathbf{x}, \mathbf{y}$ are perpendicular, i.e. $\mathbf{x}^\top\mathbf{y} = 0$. For random vectors, $X^\top Y = 0$ a.s.
$\hat{h}^{\mathfrak{M}}$	Population/ infinite-sample estimate based on distribution $\mathbb{P}^{\mathfrak{M}}$.
$\hat{h}^{\mathcal{D}}$	Finite-sample estimate based on samples in the dataset $\mathcal{D}$.

Appendix A. Confounding Bias

Statistical vs. causal inference. The target estimand for the statistical risk in Eq. (2) is the Bayes optimal predictor $\mathbb{E}^{\mathfrak{M}}[Y \mid X = \mathbf{x}]$. And the target estimand for the causal risk in Eq. (4) is the average treatment effect (ATE) $\mathbb{E}^{\mathfrak{M};\mathrm{do}(X:=\mathbf{x})}[Y \mid X = \mathbf{x}] = f(\mathbf{x})$. As such, *statistical inference* is concerned with *predictions* of outcome Y, whereas *causal inference* is concerned with *estimating* $f(\mathbf{x})$.

Statistical vs. confounding bias. Both types of inference are subject to bias. *Statistical bias* arises due to miss-specification of the hypothesis class $\mathcal{H}$, whereas confounding bias arises due to how the data are generated. The former is therefore a property of the estimator while the later is a property of the data itself. For an estimator $\hat{h}^{\mathcal{D}}$ with the expected value $\bar{h}(\cdot) = \mathbb{E}_{\mathcal{D}}^{\mathfrak{M}}\left[\hat{h}^{\mathcal{D}}(\cdot)\right]$, we define these as

$$\begin{aligned}\text{Statistical bias} &:= \mathbb{E}^{\mathfrak{M}}[Y \mid X = \cdot] - \bar{h}(\cdot),\\ \text{Confounding bias} &:= f(\cdot) - \mathbb{E}^{\mathfrak{M}}[Y \mid X = \cdot].\end{aligned}$$

Bias-variance decomposition of the causal risk. Because the treatment X and residual ξ are not correlated under $\mathfrak{M};\mathrm{do}(X)$ in Eq. (1), for any loss function ℓ that admits a 'clean' or 'additive' bias-variance decomposition Heskes (2025), the causal risk in Eq. (4) also admits a bias-variance decomposition. Using squared loss as an example, we have for some hypothesis $\hat{h}^{\mathcal{D}}$,

$$\begin{aligned}
&\Rightarrow R_{\mathrm{CR}}^{\mathfrak{M}}\left(\hat{h}^{\mathcal{D}}\right)\\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}(X)}\left[\left\|Y - \hat{h}^{\mathcal{D}}(X)\right\|^2\right],\\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}(X)}\left[\left\|f(X) + \xi - \hat{h}^{\mathcal{D}}(X)\right\|^2\right], && \text{(Structural eq. of } Y.)\\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}(X)}\left[\|\xi\|^2\right] + \mathbb{E}^{\mathfrak{M};\mathrm{do}(X)}\left[\left\|f(X) - \hat{h}^{\mathcal{D}}(X)\right\|^2\right], && \text{(Cross term is 0 as } \xi \perp\!\!\!\perp X^{\mathfrak{M};\mathrm{do}(X)}.)\\
&= \underbrace{\mathbb{E}^{\mathfrak{M};\mathrm{do}(X)}\left[\|\xi\|^2\right]}_{\text{irreducible noise}} + \underbrace{\mathbb{E}^{\mathfrak{M}}\left[\left\|f(X) - \hat{h}^{\mathcal{D}}(X)\right\|^2\right]}_{\text{estimation error, } \mathrm{CER}_{\mathfrak{M}}\left(\hat{h}^{\mathcal{D}}\right)=}. && (\mathbb{P}_X^{\mathfrak{M}},\ \mathbb{P}_X^{\mathfrak{M};\mathrm{do}(X)} \text{ identical by construction.})
\end{aligned}$$

We can show by following standard procedure that

$$\mathbb{E}_{\mathcal{D}}^{\mathfrak{M}}\left[\mathrm{CER}_{\mathfrak{M}}\left(\hat{h}^{\mathcal{D}}\right)\right] = \underbrace{\mathbb{E}_X^{\mathfrak{M}}\left[\|f(X) - \bar{h}(X)\|^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}_{\mathcal{D}}^{\mathfrak{M}}\left[\mathbb{E}_X^{\mathfrak{M}}\left[\left\|\bar{h}(X) - \hat{h}^{\mathcal{D}}(X)\right\|^2\right]\right]}_{\text{variance}}.$$

Since for any population estimate $\hat{h}^{\mathfrak{M}}(X) = \bar{h}(X)$, the CER equals the average (squared) bias in estimation

$$\mathrm{CER}_{\mathfrak{M}}\left(\hat{h}^{\mathfrak{M}}\right) = \mathbb{E}_X^{\mathfrak{M}}\left[\left\|f(X) - \hat{h}^{\mathfrak{M}}(X)\right\|^2\right] = \mathbb{E}_X^{\mathfrak{M}}\left[\|f(X) - \bar{h}(X)\|^2\right].$$

For a rich enough hypothesis class, the ERM estimate coincides with the Bayes optimal predictor $\hat{h}^{\mathfrak{M}}_{\mathrm{ERM}}(\cdot) = \mathbb{E}^{\mathfrak{M}}[Y \,|\, X = \cdot]$ and the CER exactly equals the (average squared) confounding bias as we define it above. For a general estimate $\hat{h}^{\mathcal{D}}$, however, the CER also contains statistical bias. Nevertheless, our claims of "better causal estimation via reducing confounding bias" rest on the fact that we are essentially manipulating the data via DA and/or using treatment randomization sources in the form of IVLs. And recall that confounding bias is a property of the data.

Appendix B. Simultaneity as Cyclic Structures in Equilibrium

Linear cyclic assignments

SEMs with cyclic structures have been well studied both in the linear case Lauritzen and Richardson (2002); Lacerda et al. (2008); Hyttinen et al. (2012), as well as the non-linear case Mooij et al. (2011); Bongers et al. (2021). Here we briefly provide a causal interpretation to linear simultaneous equations as SEMs with cyclic assignments.

Consider a square matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ and the SEM

$$W = \mathbf{M}W + N \ , \tag{9}$$

where random noise vector N is exogenous and $\mathbf{M}$ allows for a cyclic structure. We enforce $(\mathbf{I}_d - \mathbf{M})$ to be invertible so that the above equation has a unique solution W for any given N. Re-writing the *structural form* in Eq. (9) into a *reduced form*, the distribution over W is defined by

$$W = (\mathbf{I}_d - \mathbf{M})^{-1} N \ . \tag{10}$$

One way we can present a causal interpretation of the above solution is to view it as a stationary point to the following sequence of random vectors W_t

$$W_t = \mathbf{M}W_{t-1} + N \ ,$$

which converges if $\mathbf{M}$ has a spectral norm strictly smaller than one so that $\mathbf{M}^t \to 0$ as $t \to \infty$. The structural form Eq. (9) essentially describes the iterative application of this operation. And in the limit the distribution of $\lim_{t\to\infty} W^t$ will be the same as the reduced form Eq. (10). Although equivalent, reduced form of a cyclic SEM (if one exists) obscures the causal relations in the data generation process.

Furthermore, we restrict our models to not have any "self-cycles" (an edge from a vertex to itself). So, e.g., the matrix $\mathbf{M}$ in Eq. (9) has all zero diagonal entries. This not only simplifies our analysis by providing a simple and intuitive interpretation for our definition of DA in **??**, but it also ensures that non-linear SEMs entail unique, well-defined distributions under mild assumptions Bongers et al. (2021); Lacerda et al. (2008).

Similarly we can write the example SEM $\mathfrak{M}$ from Example 2 in this (block matrix) form as

$$\underbrace{\begin{bmatrix} X \\ Y \end{bmatrix}}_{W} = \underbrace{\begin{bmatrix} \mathbf{0}_{m\times m} & \boldsymbol{\tau}^\top \\ \mathbf{f}^\top & \mathbf{0}_{1\times 1} \end{bmatrix}}_{\mathbf{M}} \underbrace{\begin{bmatrix} X \\ Y \end{bmatrix}}_{W} + \underbrace{\begin{bmatrix} \boldsymbol{\Gamma}^\top \\ \mathbf{0}_{1\times k} \end{bmatrix} Z + \begin{bmatrix} \mathbf{T}^\top \\ \boldsymbol{\epsilon}^\top \end{bmatrix} C + \sigma \cdot \begin{bmatrix} N_X \\ N_Y \end{bmatrix}}_{N},$$

For this simple case, $\left(\mathbf{I}_{(m+1)} - \mathbf{M}\right)$ is always invertible so long as $\mathbf{f}^\top \boldsymbol{\tau}^\top \neq 1$ from Lemma 3. Or we can also restrict $\left|\mathbf{f}^\top \boldsymbol{\tau}^\top\right| < 1$ to ensure that the spectral norm of $\mathbf{M}$ is strictly smaller than 1. We sample from this SEM by first sampling all of the exogenous variables Z, C, N_X, N_Y and then solving the above system for each sample of X, Y via the reduced form in Lemma 3.

A motivating example

Cyclic SEMs were first discussed in the econometrics literature Christ (1994) to model various observational phenomena, and often solved via 2SLS based IV regression Fox (1979) since it is computationally less costly compared to solving the entire system Belsley (1988). A classic example from economics Ezekiel (1938); Muth (1961) is that of a *supply and demand model* $\mathfrak{M}$ where the relation of price P of a good with quantity Q of demand can be thought of as a cyclic feed-back loop where producers adjust their price in response to demand of the good and consumers change their demand in response to price of a good. In contrast, a change in consumer tastes or preferences would be an exogenous change on the demand curve and can therefore be used as an IV Z.

$$\begin{aligned} \text{consumer demand:} \quad & Q = \tau \cdot P + \gamma \cdot Z + N_Q \ , \\ \text{producer price:} \quad & P = f \cdot Q + N_P \ . \end{aligned}$$

Where scalars f, τ are such that $|f \cdot \tau| < 1$ so that the system converges to an equilibrium. We say that the measurements made for P and Q are at the equilibrium state of the market[4] with zero mean measurement noise N_P, N_Q respectively.

Mitigating simultaneity bias for causal effect estimation. If we now want to *estimate* the effect of demand on price f, standard regression will produce a biased estimate $\hat{f}^{\mathfrak{M}}_{\text{ERM}} = f + \frac{\text{Cov}(Q,N_P)}{\mathbb{V}[Q]}$ because of the simultaneity causing Q and N_P to be correlated (to see this, substitute model of P into the model of Q). We can now use IV regression to get an unbiased estimate of the effect of demand on price in the market as $\hat{f}^{\mathfrak{M}}_{\text{IV}} = f$.

Mitigating spurious correlations for robust prediction. Similarly, if the producer wants to *predict* the effect on demand if price is changed (i.e. intervened on), naive ERM will not be a good choice because it will also capture the spurious correlation from $Q \to P$. We therefore use three-stage-least-squares (3SLS) Zellner and Theil (1962); Belsley (1988) (or similar methods) to estimate the ATE $\hat{\tau}^{\mathfrak{M}}_{\text{3SLS}} = \mathbb{E}^{\mathfrak{M};\text{do}(P:=.)}[Q \,|\, P = .]$ where we use the first two stages to estimate $\hat{f}^{\mathfrak{M}}_{\text{IV}}$, followed by ERM to regress from the residuals $\hat{N}_P := P - \hat{f}^{\mathfrak{M}}_{\text{IV}} \cdot Q$ to Q in the third stage.

Implications for independence of causal mechanisms

Here we clarify how the equilibrium assumption/interpretation of cyclic SEMs is not at odds with the classic independent causal mechanism (ICM) principle Peters et al. (2017). Note that our SEM formulation in Eq. (1) is a direct instantiation of the ICM principle as described by Peters et al. Peters et al. (2017). The two equations represent the autonomous mechanisms, and their independence is captured by the mutual independence of the exogenous noise terms N_X, N_Y. The simultaneity in our model is not a violation of ICM, but rather the equilibrium state resulting from the interaction of these two independent mechanisms. Assuming the existence of this equilibrium is a statement about the scope of systems under analysis, and not about the nature of the mechanisms themselves. Indeed, surgically changing

4. In fact, such a feed-back model of supply and demand was initially developed to understand the irregular fluctuations of prices/quantities that are observed in some markets when not at equilibrium Ezekiel (1938).

τ to some τ', for example, does not in itself alter f and vice versa. And precisely because of the ICM, this may or may not make the system unstable depending on the nature of τ'. Nevertheless, in our setting, Proposition 2 (Appendix H.2) shows that soft interventions induced by outcome-invariant DA are *always* stable.

Appendix C. Related Work

Causal regularization is perhaps the most appropriate classification for this work. These methods aim for more robust prediction by mitigating the upstream problem of confounding bias in a more accessible way than is required for full identification. This is done, for example, by relaxing properties of auxiliary variables Kania and Wit (2023); Bühlmann and Cevid (2020); Oberst et al. (2021); Rothenhäusler et al. (2021), as we have done via our IVL approach. Most relevant, however, are methods that re-purpose common regularizers, canonically used for estimation variance reduction and i.i.d. prediction generalization, for confounding bias mitigation. Of note is Janzing (2019), where a certain linear modeling assumption allows the estimation of $\|\mathbf{f}\|^2$ from observational (X, Y) data, which is then used to develop a cross-validation scheme for ℓ_1, ℓ_2 regularization. Vankadara et al. (2022) conducted a similar theoretical analysis for the min-norm interpolator. To the best of our knowledge, we are the first to study the same for DA—re-purposing yet another ubiquitous regularizer to mitigate confounding bias.

Domain generalization (DG) Muandet et al. (2013) methods aim for prediction generalization to unseen test domains via *robust optimization (RO)* Ben-Tal et al. (2009) over a perturbation set $\mathcal{P}$ of possible test domains $\rho \in \mathcal{P}$ as

$$R^{\mathcal{P}}_{\mathrm{RO}}(h) \coloneqq \max_{\rho \in \mathcal{P}} R^{\rho}_{\mathrm{ERM}}(h),$$

Since generalizing to arbitrary test domains is impossible, the choice of perturbation set encodes one's assumptions about which test domains might be encountered. Instead of making such assumptions a priori, it is often assumed to have access to data from multiple training domains which can inform one's choice of perturbation set. This setting is explored in group distributionally robust optimization (DRO) Sagawa et al. (2020). Variations have been used to mitigate confounding bias and subsequently generalize to treatment interventions when used with interventional data Peters et al. (2016); Heinze-Deml et al. (2018), confounder information (i.e. entire graph) Krueger et al. (2021); Lu et al. (2022); Dance and Bloem-Reddy (2025) or some proxy thereof in the form of environments Arjovsky et al. (2019); Cheema et al. (2024); Han et al. (2023); Krueger et al. (2021). We, however, do not assume access to any of these and instead synthesize interventions via DA.

Counterfactual DA strategies have been the primary lens for causal analyses of DA Ilse et al. (2021); Yuan et al. (2024); Feder et al. (2023); Pitis et al. (2022); Armengol Urpí et al. (2024); Mahajan et al. (2021); Aloui et al. (2025). These aim for prediction robustness to treatment interventions via DA simulated *counterfactuals*.[5] As with counterfactual reasoning more broadly, this requires strong assumptions—such as access to the full SEM Yuan et al. (2024); Feder et al. (2023), auxiliary variables Ilse et al. (2021); Feder et al. (2023); Mahajan et al. (2021); Aloui et al. (2025), or causal graphs Pitis et al. (2022); Armengol Urpí et al. (2024). By contrast, we show that outcome invariance of DA suffices for treatment intervention robustness without invoking counterfactuals. Moreover, prior work has largely overlooked causal effect estimation, often assuming reverse-causal settings where the ATE becomes

5. Representing an SEM with exogenous noise distribution conditioned on some variable $Y = \mathbf{y}$ by $\mathfrak{A}_{Y=\mathbf{y}}$, the counterfactual SEM $\mathfrak{A}_{Y=\mathbf{y}}; \mathrm{do}(X \coloneqq \mathbf{x})$ is an intervention $\mathrm{do}(X \coloneqq \mathbf{x})$ on $\mathfrak{A}_{Y=\mathbf{y}}$. The resulting *counterfactual distribution* then captures questions like: "After observing $Y = \mathbf{y}$, what would have been had $X = \mathbf{x}$ been true."

trivial Ilse et al. (2021); Feder et al. (2023); Yuan et al. (2024). Ours is the first framework to study ATE estimation via DA with minimal structural assumptions.

Invariant prediction based methods aim to make predictions based on statistical relationships that remain stable across all domains in $\mathcal{P}$. A common assumption, for instance, is that $\mathbb{P}_{Y|X}$ is invariant across $\mathcal{P}$, with only the marginal $\mathbb{P}_X$ being allowed to vary. Invariance is also closely linked to causal discovery—following the classic ICM principle Peters et al. (2017), causal mechanisms remain stable under interventions on inputs Christiansen et al. (2022); Reddy et al. (2025). This connection has inspired approaches that enforce invariance conditions to recover causal structures Peters et al. (2016); Heinze-Deml et al. (2018). IV regression can also be viewed as one such method, where the goal is to learn predictors whose residuals are invariant to the instruments Zhang et al. (2023); Singh et al. (2019); Muandet et al. (2020); Xu et al. (2021); Ren et al. (2025). More broadly, the principle of invariance, whether motivated by causality or otherwise, has proven useful for improving prediction generalization across heterogeneous settings Rothenhäusler et al. (2021); Arjovsky et al. (2019); Dai et al. (2017); Oberst et al. (2021); Montasser et al. (2024); Yang et al. (2019); Jayanthi et al. (2023); Wang et al. (2022); Muandet et al. (2013).

Appendix D. IV Regression

Two-stage estimators. Minimizing risk of the form Eq. (6) is known as two-stage IV regression. Another approach for two-stage IV regression is to minimize the risk Mastouri et al. (2021); Rothenhäusler et al. (2021)

$$R^{\mathfrak{M}}_{\mathrm{IV}_{\mathrm{LB}}}(h) := \mathbb{E}^{\mathfrak{M}}\Big[\Big\|\mathbb{E}^{\mathfrak{M}}[Y\,|\,Z] - \mathbb{E}^{\mathfrak{M}}[h(X)\,|\,Z]\Big\|^2\Big], \tag{11}$$

which can be shown to lower-bound (hence the subscript LB) the surrogate risk in Eq. (6) Mastouri et al. (2021) under squared loss.

$$\begin{aligned}
&\Rightarrow R^{\mathfrak{M}}_{\mathrm{IV}}(h)\\
&= \mathbb{E}\Big[\|Y - \mathbb{E}[h(X)\,|\,Z]\|^2\Big],\\
&= \mathbb{E}\Big[\|(Y - \mathbb{E}[Y\,|\,Z]) + (\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z])\|^2\Big], \quad \text{(Adding and subtracting } \mathbb{E}[Y\,|\,Z].)\\
&= \mathbb{E}\Big[\|Y - \mathbb{E}[Y\,|\,Z]\|^2\Big] + \mathbb{E}\Big[\|\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z]\|^2\Big] \qquad \text{(Expand squared norm.)}\\
&\qquad\qquad + 2\mathbb{E}\Big[(Y - \mathbb{E}[Y\,|\,Z])^\top(\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z])\Big],\\
&= \mathbb{E}\Big[\|Y - \mathbb{E}[Y\,|\,Z]\|^2\Big] + \mathbb{E}\Big[\|\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z]\|^2\Big], \qquad (12)\\
&= \mathbb{E}\Big[\|\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z]\|^2\Big] + \mathbb{E}\Big[\mathbb{E}\Big[(Y - \mathbb{E}[Y\,|\,Z])^2\,\Big|\,Z\Big]\Big],\\
&\qquad\qquad \text{(Tower rule and scalar } Y.)\\
&= \mathbb{E}\Big[\|\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z]\|^2\Big] + \mathbb{E}[\mathbb{V}[Y\,|\,Z]] = R^{\mathfrak{M}}_{\mathrm{IV}_{\mathrm{LB}}}(h) + \mathbb{E}[\mathbb{V}[Y\,|\,Z]], \qquad (13)
\end{aligned}$$

where Eq. (13) follows from the definition of conditional variance and we get Eq. (12) by setting the cross term to zero since

$$\begin{aligned}
&\Rightarrow \mathbb{E}\Big[(Y - \mathbb{E}[Y\,|\,Z])^\top(\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z])\Big]\\
&= \mathbb{E}\Big[\mathbb{E}\Big[(Y - \mathbb{E}[Y\,|\,Z])^\top(\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z])\,\Big|\,Z\Big]\Big], \qquad \text{(Tower rule.)}\\
&= \mathbb{E}\Big[\mathbb{E}\Big[(Y - \mathbb{E}[Y\,|\,Z])^\top\,\Big|\,Z\Big](\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z])\Big], \qquad (14)\\
&= \mathbb{E}\Big[(\mathbb{E}[Y\,|\,Z] - \mathbb{E}[Y\,|\,Z])^\top(\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z])\Big],\\
&= \mathbb{E}\Big[\mathbf{0}^\top(\mathbb{E}[Y\,|\,Z] - \mathbb{E}[h(X)\,|\,Z])\Big] = 0,
\end{aligned}$$

where Eq. (14) follows from the "taking out what is known" rule, i.e.,

$$\mathbb{E}[g(B)A\,|\,B] = g(B)\mathbb{E}[A\,|\,B]. \tag{15}$$

Generalized method of moments. The IV regression in our colored-MNIST experiment uses the popular *generalized methods of moments (GMM)* Hall (2003); Bennett et al. (2019); Lewis and Syrgkanis (2018), or equivalently the *conditional moment restriction*

(CMR) Mastouri et al. (2021) framework which tries to directly solve for the fact that in Eq. (1) with scalar Y

$$\mathbb{E}^{\mathfrak{M}}[\,\xi\,|\,Z\,] = \mathbb{E}^{\mathfrak{M}}[\,Y - f(X)\,|\,Z\,] = 0,$$

which holds as a direct consequence of un-confoundedness of Z. For any $q : \mathcal{Z} \to \mathbb{R}$, it then follows

$$\mathbb{E}^{\mathfrak{M}}\big[\big(Y - f(X)\big) \cdot q(Z)\big] = 0\,.$$

The GMM-IV estimate of f therefore tries to enforce this condition Hall (2003); Bennett et al. (2019); Lewis and Syrgkanis (2018) by minimizing the risk

$$R^{\mathfrak{M}}_{\mathrm{IV}_{\mathrm{GMM}}}(h) \coloneqq \sum_{i=1}^{\mu} \mathbb{E}^{\mathfrak{M}}\big[\big(Y - h(X)\big) \cdot q_i(Z)\big]^2 = \Big\|\mathbb{E}^{\mathfrak{M}}[\,(Y - h(X)) \cdot \mathbf{q}(Z)\,]\Big\|^2,$$

where $\mathbf{q}(\cdot) \in \mathbb{R}^{\mu}$ represents a vector form of the set of μ arbitrary real-valued functions q_i. A more general form of the above GMM based IV risk is to weight the norm by some SPD $\mathbf{W}$ Johnston (1971); Hall (2003); Bennett et al. (2019)

$$R^{\mathfrak{M}}_{\mathrm{IV}_{\mathrm{GMM}-\mathbf{W}}}(h) \coloneqq \Big\|\mathbb{E}^{\mathfrak{M}}[\,(Y - h(X)) \cdot \mathbf{q}(Z)\,]\Big\|^2_{\mathbf{W}},$$

which gives the most statistically efficient estimator, minimizing the asymptotic variance, for $\mathbf{W} = \mathbf{\Sigma}_Z^{-1}$ Johnston (1971); Hall (2003); Bennett et al. (2019). We use the same for our colored-MNIST experiments, together with the identity function $\mathbf{q}(Z) = Z$. This gives us the final loss of the form

$$R^{\mathfrak{M}}_{\mathrm{IV}_{\mathrm{GMM}-\mathbf{\Sigma}_Z^{-1}}}(h) = \Big\|\mathbb{E}^{\mathfrak{M}}[\,Z \cdot (Y - h(X))\,]\Big\|^2_{\mathbf{\Sigma}_Z^{-1}}.$$

And the empirical version of which can be written as follows

$$R^{\mathcal{D}}_{\mathrm{IV}_{\mathrm{GMM}-\mathbf{\Sigma}_Z^{-1}}}(h) \coloneqq \Big(\hat{\mathbf{y}} - \mathbf{h}\Big(\hat{\mathbf{X}}\Big)\Big)^{\top} \hat{\mathbf{Z}}\hat{\mathbf{Z}}^{\dagger}\Big(\hat{\mathbf{y}} - \mathbf{h}\Big(\hat{\mathbf{X}}\Big)\Big), \tag{16}$$

where for dataset samples $(\mathbf{x}_i, y_i, \mathbf{z}_i) \in \mathcal{D}$, we construct the vector $\hat{\mathbf{y}} \coloneqq [y_0, \cdots, y_n]^{\top}$, matrices $\hat{\mathbf{X}} \coloneqq [\mathbf{x}_0^{\top}, \cdots, \mathbf{x}_n^{\top}]^{\top}$, $\hat{\mathbf{Z}} \coloneqq \begin{bmatrix}\mathbf{z}_0 & \cdots & \mathbf{z}_n\end{bmatrix}^{\top}$ with pseudo-inverse $\hat{\mathbf{Z}}^{\dagger}$ and define $\mathbf{h}\Big(\hat{\mathbf{X}}\Big) \coloneqq [h(\mathbf{x}_0), \cdots, h(\mathbf{x}_n)]^{\top}$.

Appendix E. IV-like Regression

Faithfulness and outcome-relevance in IVs

Consider the SEM $\mathfrak{M}$ from Sec. 1. The distribution $\mathbb{P}^{\mathfrak{M}}_{X,Y,Z,C}$ is said to be *faithful* to the graph of $\mathfrak{M}$ if it only exhibits independences implied by the graph Peters et al. (2017); Koller and Friedman (2009).[6] This standard assumption in IV settings renders outcome-relevance implicit and therefore rarely mentioned. In this section we discuss the case where only the first three IV properties are satisfied, i.e. outcome-relevance may not hold. Since such a Z may not be a valid IV, therefore identifiability of f is not possible in general as the problem in Eq. (5) can now be misspecified, having multiple, potentially infinitely many solutions when $Y \perp\!\!\!\perp Z$. Nevertheless, we shall refer to such a Z as *IV-like (IVL)* to emphasize that while Z may not be an IV, it may still be "instrumental" for reducing confounding bias when estimating f compared to the standard ERM baseline.

ERM regularized IV regression. Despite problem miss-specification for a IVL Z, the target function f remains a minimizer for the IV risk in Eq. (6). Albeit, potentially not unique – for example, a linear h with squared loss leads to an under-determined problem in Eq. (6). We therefore propose a regularized version of the IV risk for such an IVL setting,

$$R^{\mathfrak{M}}_{\mathrm{IVL}_\alpha}(h) := R^{\mathfrak{M}}_{\mathrm{IV}}(h) + \alpha R^{\mathfrak{M}}_{\mathrm{ERM}}(h), \tag{17}$$

where $\alpha > 0$ is the regularization parameter. The ERM risk as a penalty allows our estimations to have good predictive performance while the IV risk encourages solution search within a subspace where we know f to be present. We refer to minimising Eq. (17) as *IVL regression*.

Note that the motivation behind IVL regression is not the identifiability of f, but rather potentially better estimation of f by reducing confounding bias. We provide an example.

Example 2 (a linear Gaussian IVL example) *For $\sigma > 0$, non-zero $\mathbf{\Gamma}, \mathbf{T} \in \mathbb{R}^{*\times m}$ and $\boldsymbol{\tau}^\top, \mathbf{f}, \boldsymbol{\epsilon} \in \mathbb{R}^m$ such that $\mathbf{f}^\top \boldsymbol{\tau}^\top \neq 1$ so that the following SEM $\mathfrak{M}$ is solvable in (X, Y)*

$$X = \boldsymbol{\tau}^\top Y + \mathbf{\Gamma}^\top Z + \mathbf{T}^\top C + \sigma N_X, \qquad Y = \mathbf{f}^\top X + \boldsymbol{\epsilon}^\top C + \sigma N_Y,$$

where Z, C, N_X, N_Y are conformable, centered Gaussian random vectors and Z is IVL.

Theorem 2 (robust prediction with IVL regression) *For SEM $\mathfrak{M}$ in Example 2,*

$$\hat{\mathbf{h}}^{\mathfrak{M}}_{IVL_\alpha} \in \operatorname*{argmin}_{\mathbf{h}} \max_{\boldsymbol{\zeta} \in \mathcal{P}_\alpha} R^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot) := \boldsymbol{\zeta}\right)}_{ERM}(\mathbf{h}), \quad s.t. \quad \mathcal{P}_\alpha := \left\{ \boldsymbol{\zeta} \,\middle|\, \boldsymbol{\zeta}\boldsymbol{\zeta}^\top \preccurlyeq \left(\frac{1}{\alpha} + 1\right) \mathbf{\Gamma}^\top \mathbf{\Sigma}^{\mathfrak{M}}_Z \mathbf{\Gamma} \right\}.$$

Proof See Appendix H.4 for the proof. ■

Theorem 3 (causal estimation with IVL regression) *In Example 2, for $\alpha < \infty$,*

$$\mathrm{CER}_{\mathfrak{M}}\left(\hat{\mathbf{h}}^{\mathfrak{M}}_{IVL_\alpha}\right) \leq \mathrm{CER}_{\mathfrak{M}}\left(\hat{\mathbf{h}}^{\mathfrak{M}}_{ERM}\right), \qquad \textit{equality iff} \qquad \mathbb{E}^{\mathfrak{M}}[X \,|\, Z] \perp \mathbb{E}^{\mathfrak{M}}[X \,|\, \xi] \qquad a.s.$$

Proof See Appendix H.5 for the proof. ■

6. Also known as *stability* in some texts (Pearl, 2009, p. 48).

Theorem 2 shows that IVL regression achieves optimal predictive performance across treatment interventions within the perturbation set $\mathcal{P}_\alpha$ defined by α. Theorem 3 further states that this strictly reduces confounding bias in estimation of f iff the perturbations align with spurious features of X, as indicated by the equality condition (also necessary for identifiability in linear IV settings Wooldridge (2010); Christiansen et al. (2022)).

Closed form solution in the linear case.

The following result gives us a way to compute a closed-form solution to the IVL_α regression problem in the linear Gaussian case. An empirical version of this is used for our linear experiments.

Proposition 1 (IVL_α closed form solution) *For SEM $\mathfrak{M}$ in Example 2, $\hat{\mathbf{h}}^{\mathfrak{M}}_{IVL_\alpha}$ is the closed form linear OLS solution between*

$$X' \coloneqq aX + b\mathbb{E}[X \,|\, Z], \qquad Y' \coloneqq aY + b\mathbb{E}[Y \,|\, Z],$$

where

$$a \coloneqq \sqrt{\alpha}, \qquad b \coloneqq \sqrt{1+\alpha} - \sqrt{\alpha}.$$

Proof See Appendix H.1 for the proof. ■

For the empirical version of Proposition 1 we fit a closed-form OLS regressor between

$$X' \coloneqq \sqrt{\alpha}X + \left(\sqrt{1+\alpha} - \sqrt{\alpha}\right)\hat{\mathbf{Z}}\hat{\mathbf{Z}}^\dagger X, \qquad Y' \coloneqq \sqrt{\alpha}Y + \left(\sqrt{1+\alpha} - \sqrt{\alpha}\right)\hat{\mathbf{Z}}\hat{\mathbf{Z}}^\dagger Y,$$

where $\hat{\mathbf{Z}}, \hat{\mathbf{Z}}^\dagger$ are as defined in Eq. (16).

Choice of regularization parameter.

We try the following approaches to select the parameter α.

Cross validation (CV), or any variation thereof. We specifically use the following two in our experiments; (i) vanilla CV with 20% samples held-out for validation (ii) *level cross validation (LCV)* for when Z is discrete, where hold-out data corresponding to 20% of the levels of Z for validation.

Confounder correction (CC), where in a linear setting we follow an approach similar to Janzing (2019) by estimating the length of the true solution f from the observational data $\mathcal{D}$. We then chose α such that the length of $\hat{h}^{\mathcal{D}}_{\text{DA+IVL}_\alpha}$ is closest to the estimated length of the ground truth solution.

Appendix F. Limitations

Necessity and practicality of prior knowledge. As discussed in Sec. 2, outcome invariance alone does not suffice to lower confounding bias and practitioners may need domain knowledge to construct DA that targets spurious features as well. Alternatively, one can also take a 'carpet bombing' approach by exhausting all available outcome invariant DA in hope that some may align with spurious features. Nevertheless, under outcome invariance, our methods should perform no worse than standard ERM.

Fundamentally, causal estimation from purely observational data is impossible without untestable assumptions. For instance, the IV (or IVL) assumptions of un-confoundedness and exclusion restriction are inherently untestable and must be justified through domain knowledge. Moreover, the requirement of alignment with spurious features in Theorem 3 is not an artifact of our IVL relaxation—it is a rephrasing of the exclusion principle that underlies identifiability in IV regression. If an IV does not influence Y through the spurious features of X, the corresponding causal components of f cannot be identified Christiansen et al. (2022). IVLs, being relaxations of IVs, inherit these same untestable premises.

Viewed through the lens of IVs/IVLs (Observation 2), our assumptions on DA are arguably more modest than they may initially seem, especially since a symmetry-based DA model has well-established precedent in the literature (Lyle et al., 2020; Chen et al., 2020; Montasser et al., 2024; Shao et al., 2022; Fawzi and Frossard, 2015; Dubois et al., 2021; Petrache and Trivedi, 2023; Romero and Lohit, 2022; Zhu et al., 2021; Wong et al., 2016). This correspondence can be summarized as follows:

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

In this light, our framework may in fact be quite practical in domains where valid IVs (or other auxiliary variables) are scarce, but plausible outcome-invariances—i.e., data augmentations—are abundant.

Finally, we recognize the hesitation in committing to strict notions of outcome invariance in practice and leave a more thorough exploration of approximate or even violated invariance to future work.

Choice of α. Selecting the IVL regularization parameter α in finite-sample settings is not straightforward. As outlined in Appendix E, we propose several strategies that work well empirically, though some may appear less principled since α is tuned via cross-validation within the same distribution, even though the task concerns OOD generalization. This challenge is not unique to IVL, but rather a broader limitation common to DG methods Gulrajani and Lopez-Paz (2021).

Appendix G. Experiment Supplement

We began by presenting results in the infinite-sample setting to emphasize that mitigating confounding bias is fundamentally not a sample size issue, i.e., not solvable through traditional regularization alone. In this section, we turn to the finite-sample regime and empirically evaluate the effectiveness of DA in reducing hidden confounding bias. Importantly, we do not use DA for its conventional purpose of augmenting data to improve i.i.d. generalization. Since our focus is on generalizing across interventions, we fix the number of samples in the augmented dataset to match that of the original dataset throughout all experiments.

Finding baselines for evaluating our results is however a challenge – reducing the bias due to hidden confounding in regression estimates having only access to the treatment X and outcome Y is a non-trivial problem. Nevertheless, for the sake of completeness we make an effort to re-purpose existing methods from domain-generalization, invariance learning and causal inference literature to be used as baselines. These methods often require access to additional variables (e.g. IVs, confounders, domains/environments, etc.), and to maintain fairness we will replace these with DA parameters G. Such a comparison is conceptually valid since by virtue of being DG methods, they are essentially solving a robust loss of a similar form as in Corollary 1, giving us meaningful baselines for DA+IVL.

In addition to standard ERM, DA and IV regression, our baselines include DRO Sagawa et al. (2020), invariant risk minimization (IRM) Arjovsky et al. (2019), invariant causal prediction (ICP) Peters et al. (2016), regularization with invariance on causal essential set (RICE) Wang et al. (2022), variance risk extrapolation (V-REx) and minimax risk extrapolation (MM-REx) Krueger et al. (2021). We also compare against causal regularization methods, including Kania and Wit Kania and Wit (2023) and the ℓ_1, ℓ_2 approaches by Janzing Janzing (2019). We discretise G if the method accepts only discrete variables. For IVL regression, we select the regularization parameter α in a variety of ways, including vanilla cross validation (CV), level-based cross validation (LCV) and confounder correction (CC) as described in Appendix E. Other implementation details are provided in Appendix G.

For the methods that use *stochastic gradient descent (SGD)*, we use a learning rate of 0.01, batch size of 256 for 16 epochs. For baselines that require a discrete domains/environments, we uniformly discretise each dimension of G into 2 bins. Higher discretisation bins renders most baselines ineffective since each domain/environment rarely has more than 1 sample. To keep the comparison fair, however, we also discretize G for IVL_α regression when using LCV. For the colored MNIST experiment, all CV implementations including baselines use 5-folds for a random search over an exponentially distributed regularization parameter with rate parameter of 1. Same is the case for simulation and optical device experiments, except that DA+IVL methods use a log-uniform distributed regularization parameter over $[10^{-4}, 1]$. Since RICE Wang et al. (2022) grows the dataset size by augmenting each sample T times, we provide it a $1/T$ sub-sample of the original data for fair comparison.

G.1. Simulation experiment

For the finite sample results of the linear SEM $\mathfrak{A}$ from Example 1, by taking $m = 32$, $k = 31$ (dimension of G), $\sigma = 0.1$ and fixing $\boldsymbol{\tau}^\top = \mathbf{0}$, we sample a new $\mathbf{f}, \boldsymbol{\epsilon}$ and $\mathbf{T} \in \mathbb{R}^{m \times m}$ from a standard normal distribution for each of the 32 experiments for every combination of κ and γ. Each time we construct a $\boldsymbol{\Gamma} := \mathbf{V}_0$ with k rows as orthonormal basis of null($\mathbf{f}$), such that

the SVD of $\mathbf{f}$ is

$$\mathbf{f} = \begin{bmatrix}\mathbf{u} & \mathbf{U}_0\end{bmatrix} \begin{bmatrix} \sigma & \mathbf{0}_{1\times(m-1)} \\ \mathbf{0}_{(m-1)\times 1} & \mathbf{0}_{(m-1)\times(m-1)} \end{bmatrix} \begin{bmatrix}\mathbf{v}^\top \\ \mathbf{V}_0^\top\end{bmatrix}.$$

Although this construction of $\mathbf{\Gamma}$ relies on direct knowledge of $\mathbf{f}$ (which is unavailable in practice), we include it here purely for illustrative purposes. We treat access to $\mathbf{\Gamma}$ as our prior structural knowledge about the invariance properties of $\mathbf{f}$, noting that this information alone is insufficient to recover $\mathbf{f}$.

We then generate $n = 2048$ samples of (X, Y) for each experiment. For ERM we use a closed form linear OLS solution, for DA+IV, we make use of linear 2SLS. Finally, DA+IVL$_\alpha$ was implemented using a closed form linear OLS solution between empirical versions (see Proposition 1) of

$$X' := \sqrt{\alpha}X + \big(\sqrt{1+\alpha} - \sqrt{\alpha}\big)\mathbb{E}[X\,|\,Z], \qquad Y' := \sqrt{\alpha}Y + \big(\sqrt{1+\alpha} - \sqrt{\alpha}\big)\mathbb{E}[Y\,|\,Z].$$

Our first experimental result in Fig. 2(*a*) compares the different estimation methods across varying levels of confounding $\kappa \in [0, 1]$. As expected, ERM performance degrades with increasing confounding. Applying DA alone already brings us closer to the causal solution, while DA+IVL achieves even better performance. DA+IV regression is unstable and generally performs poorly as it is under-determined.

In the second experiment (Fig. 2(*b*)), we fix the confounding and DA strengths at $\kappa = \gamma = 1$, and sweep over the regularization parameter $\alpha \in [10^{-5}, 10^5]$ for DA+IVL$_\alpha$. The results show that optimal performance is achieved for intermediate values of α, confirming that arbitrarily small values of α, while beneficial in the population setting (as suggested by Theorem 3), are suboptimal in finite samples.[7] We also find that both CV and CC strategies effectively select reasonable values of α.

Finally, we examine sensitivity to the DA strength $\gamma \in [10^{-2.5}, 10]$, fixing $\kappa = 1$. As expected, stronger DA results in stronger interventions on X, which improves causal effect estimation. However, we also observe diminishing returns; when the variation induced by DA is either too small or too large, DA+IVL$_\alpha$ does not yield significant improvements over the DA+ERM baseline.

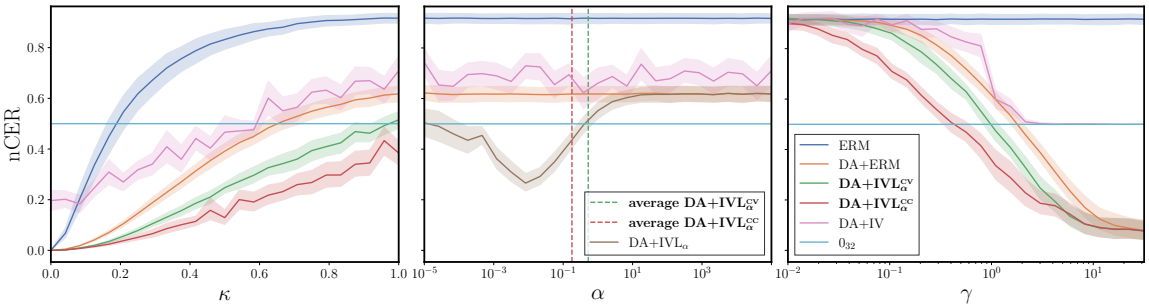
For completeness, we also benchmark our approach against other baseline methods on 16 distinct simulation SEMs with 2048 samples each. Aggregated results are presented in Fig. 3 (left most).

For the parameter sweep experiments of Fig. 2, we generate a treatment of dimension $m = 32$, but for the OOD baseline comparison experiment in Fig. 3 we use $m = 16$. Furthermore, for the OOD baseline comparison experiment in Fig. 3, we randomly pick each basis of null($\mathbf{f}$) with a probability 2/3 to construct $\mathbf{\Gamma}$ (i.e., we know only some, but not all symmetries of $\mathbf{f}$).

G.2. Optical device experiment

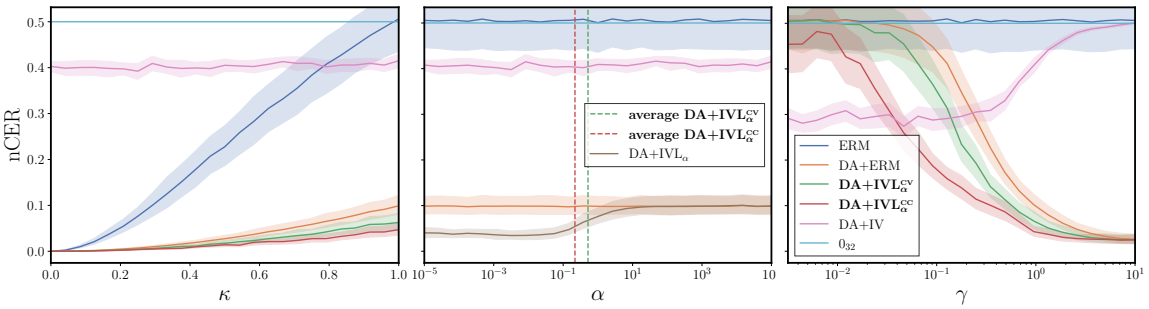
The dataset from Janzing and Schölkopf (2018) consists of 3×3 pixel images X displayed on a laptop screen that cause voltage readings Y across a photo-diode. A hidden confounder

7. We conjecture that this is due to outcome invariance not holding exactly in practice.



(a) $\gamma = 1$ (b) $\gamma, \kappa = 1$ (c) $\kappa = 1$

Figure 4: Simulation of the linear Gaussian SEM of Example 1 with the same setting as Fig. 2, but $\boldsymbol{\tau}^\top, \mathbf{f}$ sampled uniformly over a unit sphere, representing a cyclic structure.



(a) $\gamma = 1$ (b) $\gamma, \kappa = 1$ (c) $\kappa = 1$

Figure 5: Same experiment as Fig. 2, but with $\mathbf{\Gamma}$ constructed by randomly selecting each basis of $\text{null}(\mathbf{f}^\top)$ with a probability of $2/3$, so that we can simulate the effect of knowing only some symmetries of $\mathbf{f}$.

C controls two LEDs; one affects the webcam capturing X, the other affects the photo-diode measuring Y. The ground-truth predictor $\mathbf{f}$ is computed by first regressing Y on $(\phi(X), C)$, where $\phi(X)$ are polynomial features of X with degree $d \in \{1, \cdots, 5\}$ that best explains the data. The component corresponding to C is then removed to recover $\mathbf{f}$. We add Gaussian noise $G \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_X/10)$ for DA and evaluate methods from **??** on $n = 1000$ samples across 12 datasets. Figure 3 (middle) shows that DA+ERM improves over ERM, and DA+IVL performs even better, outperforming other baselines.

In this experiment we fit a linear function $h(.) := \mathbf{h} \in \mathbb{R}^m$ for a squared loss in all of our risk metrics. For IVL$_\alpha$ regression, we use the closed-form OLS solution from Appendix E. We also use a closed-form solution for ERM, DA+ERM and DA+IV (2SLS) baselines. The rest of the baselines (other than ICP) use SGD.

Most of the datasets in the optical device dataset were best explained by polynomial features of degree 2. We use the same ground-truth degree to fit each of the methods listed in Fig. 3. This is important so as to avoid statistical bias from model miss-specification as our analysis squarely focuses on confounding bias.

G.3. Colored-MNIST experiment

We evaluate on Colored MNIST Arjovsky et al. (2019), where labels are spuriously correlated with image color during training, but this correlation is flipped at test time. We use the same neural architecture and parameters as Arjovsky et al. (2019) across all baselines, training with the IV-based objective described in the Appendix D. DA is implemented via small perturbations to hue, brightness, contrast, saturation, and translation, each parameterized by $G \sim \boldsymbol{\beta}(2,2)$. Although these do not directly manipulate color, the actual spurious feature, they still help reduce confounding. Results in Fig. 3 (rightmost) show that ERM underperforms, DA+ERM provides substantial gains, and DA+IVL$_\alpha$ performs competitively with the best DG baselines, with DA+IVL$_\alpha^{\text{CV}}$ achieving the best overall performance.

We use the same 3-layer neural network (NN) architecture for h across all methods comprising of a fully-connected input layer of input dimension m, hidden layer of input/output dimension 256 and output classification layer with a Sigmoid function. Each layer is separated by an intermediary *rectified linear unit* activation function. For the IV risk, we use the empirical version of the GMM based risk from Eq. (16).

Colored-MNIST as a cyclic SEM – From invariant prediction to estimating causal effects

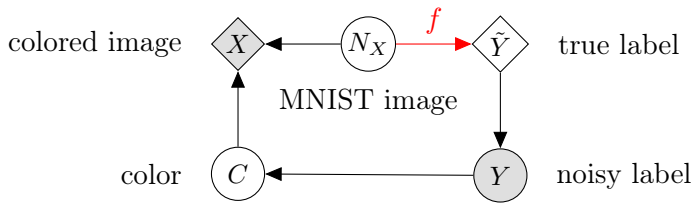


Figure 6: The data generation DAG for colored-MNIST as discussed by the original authors Arjovsky et al. (2019). They aim to learn a predictor $h : \mathcal{X} \to \mathcal{Y}$ such that it is invariant to changes in $\mathbb{P}_{X|Y}$. We argue that this DAG view of colored-MNIST does not make it obvious how the true labeling function $f(\mathbf{x})$ is related to the ATE $\mathbb{E}^{\mathfrak{M};\text{do}(X:=\mathbf{x})}[Y \,|\, X = \mathbf{x}]$, which we believe is because it is virtually equivalent to the reduced form of our structural form presented in Fig. 7.

In this section we give a cyclic SEM perspective of the colored-MNIST experiment from Arjovsky et al. (2019). The task is binary classification of colored images X from the MNIST dataset into low digits ($y = 0$ for digits from 0 to 4) and high digits ($y = 1$ for digits from 5 to 9). The difficulty of the task arises from there being a higher spurious correlation between the color C of the images ($c = 0$ for blue and $c = 1$ for green) and (noisy) labels Y as compared to the correlation between the digits in the image and the label.

Consider the following cyclic SEM in Fig. 7.

$$
\begin{aligned}
\mathbf{n}_X &\sim \mathbb{P}_{N_X}, n_Y \sim \mathbb{B}(0.25), n_c \sim \mathbb{B}(e) && \text{sample all exogenous variables} \\
X &= \texttt{colour}(C, \mathbf{n}_X) && \text{apply color } C \text{ to the image} \\
\tilde{Y} &= f(X) && \text{generate ground-truth label with true labeling function} \\
Y &= \texttt{xor}\Big(\tilde{Y}, n_Y\Big) && \text{flip the label with probability 0.25} \\
C &= \texttt{xor}(Y, n_C) && \text{generate color by flipping } Y \text{ with probability } e,
\end{aligned}
$$

where we first randomly sample an un-colored MNIST image $\mathbf{n}_X$, and some Bernoulli distributed label noise $n_Y \sim \mathbb{B}(0.25)$ and color noise $n_C \sim \mathbb{B}(e)$ which is different for each environment $e \in \{0.1, 0.2\}$. Then for some initial arbitrary values $\mathbf{x}_0$, $\tilde{y}_0$, y_0 and c_0 respectively for the observed colored image X, the ground-truth label $\tilde{Y}$, the observed noisy label Y and the image color C, we iteratively apply the following assignments from the SEM

$$
\begin{aligned}
\mathbf{x}_t &= \texttt{colour}(c_{t-1}, \mathbf{n}_X) && \text{apply color } C \text{ to the image} \\
\tilde{y}_t &= f(\mathbf{x}_{t-1}) && \text{generate ground-truth label with true labeling function} \\
y_t &= \texttt{xor}(\tilde{y}_{t-1}, n_Y) && \text{flip the label with probability 0.25} \\
c_t &= \texttt{xor}(y_{t-1}, n_C) && \text{generate color by flipping } Y \text{ with probability } e,
\end{aligned}
$$

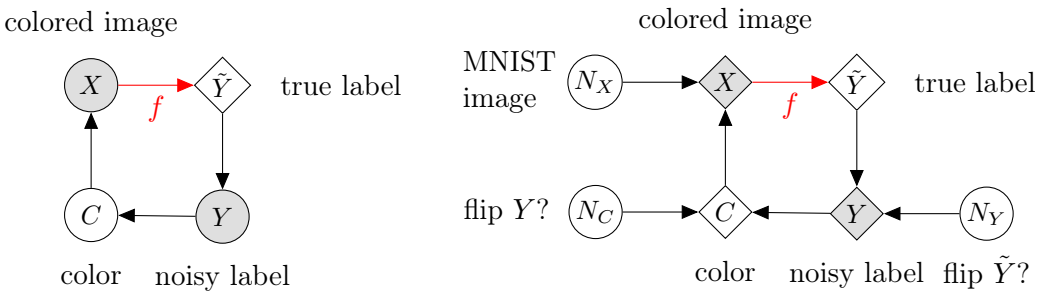
until they converge while keeping all sampled exogenous variables $\mathbf{n}_X, n_Y, n_C$ fixed. It is straightforward to show that this SEM will converge after a maximum of $t = 5$ iterations[8] due to the invariance of f to the color of the image C. Furthermore, this stationary-point will be uniquely determined by our exogenous samples $\mathbf{n}_X, n_Y, n_C$. And this is how we generate one sample $(\mathbf{x}, y)$ for our colored-MNIST experiment. We repeat this process to generate a sample $(\mathbf{x}, y)$ for each of n samples $\mathbf{n}_X, n_Y, n_C$.

Note that the ground-truth labeling function f can only correctly predict the labels 75% of the time. At test time we flip the correlation between the label Y and the image color C by setting $e = 0.9$. Also, the above cyclic SEM for colored-MNIST produces the same distribution for (X, Y) as Arjovsky et al. (2019).

The above cyclic SEM perspective of colored-MNIST is interesting because it makes it clear that colored-MNIST is essentially a causal effect estimation task. Specifically, we can estimate the true labeling function f by estimating the ATE $\mathbb{E}^{\mathfrak{M};\text{do}(X:=\mathbf{x})}[Y \,|\, X = \mathbf{x}]$ since

$$
\begin{aligned}
\mathbb{E}^{\mathfrak{M};\text{do}(X:=\mathbf{x})}[Y \,|\, X = \mathbf{x}] &= \mathbb{E}^{\mathfrak{M};\text{do}(X:=\mathbf{x})}[\texttt{xor}(f(X), N_Y) \,|\, X = \mathbf{x}], \\
&= \mathbb{E}^{\mathfrak{M}}[\texttt{xor}(f(\mathbf{x}), N_Y)], && (N_Y \perp\!\!\!\perp X^{\mathfrak{M};\text{do}(X:=\mathbf{x})}.) \\
&= \mathbb{E}^{\mathfrak{M}}[f(\mathbf{x}) + N_Y - 2f(\mathbf{x})N_Y], && (\text{Definition of } \texttt{xor}.) \\
&= f(\mathbf{x}) + \mathbb{E}^{\mathfrak{M}}[N_Y] - 2f(\mathbf{x})\mathbb{E}^{\mathfrak{M}}[N_Y], \\
&= \Big(1 - 2\mathbb{E}^{\mathfrak{M}}[N_Y]\Big) f(\mathbf{x}) + \mathbb{E}^{\mathfrak{M}}[N_Y], \\
&= 0.5 f(\mathbf{x}) + 0.25\,. && (N_Y \sim B(0.25).)
\end{aligned}
$$

8. Following the mechanisms $c_0 \to \mathbf{x}_1 \to \tilde{y}_2 \to y_3 \to c_4 \to \mathbf{x}_5$, we see that $(\mathbf{x}_4, y_4, c_4) = (\mathbf{x}_5, y_5, c_5)$ (same for $\tilde{y}_4 = \tilde{y}_5$).



(a) Graph for generating colored-MNIST data.

(b) Augmented graph – exogenous variables explicitly shown.

Figure 7: A cyclic SEM perspective of the colored-MNIST data – an MNIST image N_X is assigned color C to produce a colored-MNIST image X. This is then passed through the ground-truth labeling function f to produce the true label $\tilde{Y}$. We flip this with probability 0.25 to produce the observed label Y, which in turn is flipped with probability e (at train time $e \in \{0.1, 0.2\}$ and $e = 0.9$ at test time) to produce the color C. These assignments are iteratively applied for any joint sample of the exogenous variables N_X, N_Y, N_C starting at arbitrary values of endogenous variables until convergence to the unique stationary point X, Y, C (and $\tilde{Y}$).

Because this is a binary classification task, we have

$$\texttt{round}\Big(\mathbb{E}^{\mathfrak{M};\mathrm{do}(X:=\mathbf{x})}[Y \,|\, X = \mathbf{x}]\Big) = f(\mathbf{x}).$$

This is in contrast to the original DAG perspective of colored-MNIST shown in Fig. 6, where the connection to the estimation of the causal mechanism f is not immediately obvious. We argue that this is because the DAG in Fig. 6 is virtually equivalent to the reduced form of our structural form presented in Fig. 7.

Appendix H. Proofs

H.1. Proof of Proposition 1 – IVL regression closed form solution in the linear case

The OLS solution for (X', Y') minimizes the following ERM risk

$$
\begin{aligned}
&\Rightarrow \mathbb{E}\left[\left\|Y' - \mathbf{h}^\top X'\right\|^2\right] \\
&= \mathbb{E}\left[\left\|aY + b\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top(aX + b\mathbb{E}[X\,|\,Z])\right\|^2\right], \quad \text{(Substitute in definitions of } X', Y'.) \\
&= \mathbb{E}\left[\left\|a\Big(Y - \mathbf{h}^\top X\Big) + b\Big(\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big)\right\|^2\right], \quad \text{(Distribute the subtraction.)} \\
&= a^2\mathbb{E}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right] + b^2\mathbb{E}\left[\left\|\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\right\|^2\right] \quad \text{(Expand squared norm.)} \\
&\qquad + 2ab\mathbb{E}\left[\Big(Y - \mathbf{h}^\top X\Big)^\top\Big(\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big)\right]. \qquad (18)
\end{aligned}
$$

First we note that from the definitions of a, b we have

$$
a^2 = \sqrt{\alpha}, \qquad b^2 + 2ab = \left(\sqrt{1+\alpha} - \sqrt{\alpha}\right)^2 + 2\sqrt{\alpha}\left(\sqrt{1+\alpha} - \sqrt{\alpha}\right) = 1. \qquad (19)
$$

Now we evaluate the cross term in Eq. (18)

$$
\begin{aligned}
&\Rightarrow \mathbb{E}\left[\Big(Y - \mathbf{h}^\top X\Big)^\top\Big(\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\Big(Y - \mathbf{h}^\top X\Big)^\top\Big(\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big)\,\middle|\,Z\right]\right], \quad \text{(Law of iterated expectation.)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\Big(Y - \mathbf{h}^\top X\Big)^\top\,\middle|\,Z\right]\Big(\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big)\right] \\
&\qquad\qquad \text{(Taking out what is known; Eq. (15).)} \\
&= \mathbb{E}\left[\Big(\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big)^\top\Big(\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big)\right] \\
&= \mathbb{E}\left[\left\|\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\right\|^2\right].
\end{aligned}
$$

Substituting this back in Eq. (18) we get

$$
\begin{aligned}
&\Rightarrow \mathbb{E}\left[\left\|Y' - \mathbf{h}^\top X'\right\|^2\right] \\
&= a^2\mathbb{E}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right] + (b^2 + 2ab)\mathbb{E}\left[\left\|\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\right\|^2\right], \\
&= \alpha\mathbb{E}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right] + \mathbb{E}\left[\left\|\mathbb{E}[Y\,|\,Z] - \mathbf{h}^\top\mathbb{E}[X\,|\,Z]\right\|^2\right], \quad \text{(From Eq. (19).)} \\
&= \alpha R^{\mathfrak{M}}_{\mathrm{ERM}}(\mathbf{h}) + R^{\mathfrak{M}}_{\mathrm{IV}}(\mathbf{h}) - \mathbb{E}[\mathbb{V}[Y\,|\,Z]], \quad \text{(From Eq. (13).)} \\
&= R^{\mathfrak{M}}_{\mathrm{IVL}_\alpha}(\mathbf{h}) - \mathbb{E}[\mathbb{V}[Y\,|\,Z]].
\end{aligned}
$$

■

H.2. Proof of Proposition 2 – Existence of interventional distribution far a DA

Proposition 2 (unique stationary interventional distribution) *In SEM $\mathfrak{A}$ from Eq. (7), given any $(\mathbf{g}, \mathbf{c}, \mathbf{n}_X, \mathbf{n}_Y) \sim P^{\mathfrak{A}}_{G,C,N_X,N_Y}$, if for all $(\mathbf{x}_0, \mathbf{y}_0) \in \mathcal{X} \times \mathcal{Y}$ the unique limits*

$$\mathbf{x}^{\mathfrak{A}} := \lim_{t\to\infty} \mathbf{x}^{\mathfrak{A}}_t = \lim_{t\to\infty} \tau\Big(\mathbf{y}^{\mathfrak{A}}_{t-1}, \mathbf{c}, \mathbf{n}_X\Big),$$
$$\mathbf{y}^{\mathfrak{A}} := \lim_{t\to\infty} \mathbf{y}^{\mathfrak{A}}_t = \lim_{t\to\infty} f\Big(\mathbf{x}^{\mathfrak{A}}_{t-1}\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y$$

exist, then in $\mathfrak{A}; \mathrm{do}(\tau := \mathbf{g}\tau)$ the unique limits

$$\mathbf{x}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)} := \lim_{t\to\infty} \mathbf{x}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_t = \lim_{t\to\infty} \mathbf{g}\tau\Big(\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_{t-1}, \mathbf{c}, \mathbf{n}_X\Big) = \mathbf{g}\mathbf{x}^{\mathfrak{A}},$$
$$\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)} := \lim_{t\to\infty} \mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_t = \lim_{t\to\infty} f\Big(\mathbf{x}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_{t-1}\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y = \mathbf{y}^{\mathfrak{A}}$$

also exist.

Proof First we try to show that

$$\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_t = \mathbf{y}^{\mathfrak{A}}_t. \tag{20}$$

For the base case, we have by construction

$$\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_0 := \mathbf{y}_0 =: \mathbf{y}^{\mathfrak{A}}_0.$$

For the step case, assuming that $\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_t = \mathbf{y}^{\mathfrak{A}}_t$, we have[9],

$$\begin{aligned}
\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_{t+2} &= f\Big(\mathbf{x}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_{t+1}\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y, \\
&= f\Big(\mathbf{g}\tau\Big(\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_t, \mathbf{c}, \mathbf{n}_X\Big)\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y, \\
&= f(\tau\Big(\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_t, \mathbf{c}, \mathbf{n}_X\Big)) + \epsilon(\mathbf{c}) + \mathbf{n}_Y, && \text{(Invariance of } f \text{ to } \mathbf{g}.) \\
&= f\Big(\tau\Big(\mathbf{y}^{\mathfrak{A}}_t, \mathbf{c}, \mathbf{n}_X\Big)\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y, && \text{(Assumption } \mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_t = \mathbf{y}^{\mathfrak{A}}_t.) \\
&= f\Big(\mathbf{x}^{\mathfrak{A}}_{t+1}\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y, \\
&= \mathbf{y}^{\mathfrak{A}}_{t+2}.
\end{aligned}$$

Hence, we have shown that Eq. (20) holds for all even t. For odd t, we simply replace $t = 0$ with $t = 1$ in the base case

$$\begin{aligned}
\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_1 &= f\Big(\mathbf{x}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_0\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y, \\
&= f\Big(\mathbf{x}^{\mathfrak{A}}_0\Big) + \epsilon(\mathbf{c}) + \mathbf{n}_Y, && \text{(Definitions } \mathbf{x}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}_0 := \mathbf{x}_0 =: \mathbf{x}^{\mathfrak{A}}_0.) \\
&= \mathbf{y}^{\mathfrak{A}}_1,
\end{aligned}$$

9. Note that here the step size for proof by induction would be $\Delta t = 2$ since $\mathbf{y}_t$ precedes $\mathbf{y}_{t+2}$. Similar is the case for $\mathbf{x}_t$ as well.

We have now finally shown that Eq. (20) holds for all $t \geq 0$.

Next, it is now relatively straightforward to show that for any $t > 0$, we have

$$\begin{aligned}
\mathbf{x}_t^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)} &= \mathbf{g}\tau\Big(\mathbf{y}_{t-1}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)}, \mathbf{c}, \mathbf{n}_X\Big), \\
&= \mathbf{g}\tau\Big(\mathbf{y}_{t-1}^{\mathfrak{A}}, \mathbf{c}, \mathbf{n}_X\Big), && \text{(Follows from Eq. (20).)} \\
&= \mathbf{g}\mathbf{x}_t^{\mathfrak{A}}. && (21)
\end{aligned}$$

Finally, by applying limit as $t \to \infty$ to both sides of Eq. (20) and Eq. (21), we get

$$\begin{aligned}
\mathbf{y}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)} &= \lim_{t\to\infty} \mathbf{y}_t^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)} = \lim_{t\to\infty} \mathbf{y}_t^{\mathfrak{A}} = \mathbf{y}^{\mathfrak{A}}, \\
\mathbf{x}^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)} &= \lim_{t\to\infty} \mathbf{x}_t^{\mathfrak{A};\mathrm{do}(\tau:=\mathbf{g}\tau)} = \lim_{t\to\infty} \mathbf{g}\mathbf{x}_t^{\mathfrak{A}} = \mathbf{g} \lim_{t\to\infty} \mathbf{x}_t^{\mathfrak{A}} = \mathbf{g}\mathbf{x}^{\mathfrak{A}}, && (22)
\end{aligned}$$

where the limit can be moved past $\mathbf{g}$ in Eq. (22) because $\mathbf{g}$ is assumed continuous in its domain.

■

H.3. Proof of Theorem 1 – Causal estimation with DA+ERM

$$
\begin{aligned}
&\Rightarrow \left\|\hat{\mathbf{h}}^{\mathfrak{A}}_{\mathrm{DA}_G+\mathrm{ERM}} - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X} = \left\|\mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1}\mathbb{E}\Big[(GX)Y^\top\Big] - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X},\\
&= \left\|\mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1}\mathbb{E}\Big[(GX)\Big(\mathbf{f}^\top X + \xi\Big)^\top\Big] - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X}, && \text{(Structural eq. of } Y.)\\
&= \left\|\mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1}\mathbb{E}\Big[(GX)\Big(\mathbf{f}^\top (GX) + \xi\Big)^\top\Big] - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X}, && \text{(Using } \mathcal{G}\text{-invariance of } \mathbf{f}.)\\
&= \left\|\left(\mathbf{f} + \mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1}\mathbb{E}\Big[(GX)\xi^\top\Big]\right) - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X},\\
&= \left\|\mathbb{E}\Big[(GX)(GX)^\top\Big]^{-1}\mathbb{E}\Big[(GX)\xi^\top\Big]\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X},\\
&= \left\|\mathbb{E}\Big[\Big(X+\tilde{G}\Big)\Big(X+\tilde{G}\Big)^\top\Big]^{-1}\mathbb{E}\Big[\Big(X+\tilde{G}\Big)\xi^\top\Big]\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X}, && (\tilde{G} \coloneqq \mathbb{E}[GX\,|\,G] = \gamma\cdot\mathbf{\Gamma}^\top G.)\\
&= \left\|\Big(\mathbb{E}\Big[XX^\top\Big] + \mathbb{E}\Big[\tilde{G}\tilde{G}^\top\Big]\Big)^{-1}\mathbb{E}\Big[X\xi^\top\Big]\right\|_{\mathbf{\Sigma}^{\mathfrak{A}}_X}, && \text{(Using } \tilde{G} \perp\!\!\!\perp X, \xi.)\\
&= \left\|\Big(\mathbf{S}^\top\mathbf{S} + \mathbf{S}^\top\mathbf{D}\mathbf{S}\Big)^{-1}\mathbb{E}\Big[X\xi^\top\Big]\right\|_{\mathbf{S}^\top\mathbf{S}}, && \text{(Lemma 2.)}\\
&= \left\|\mathbf{S}^{-1}(\mathbf{I}_m + \mathbf{D})^{-1}\mathbf{S}^{-\top}\mathbb{E}\Big[X\xi^\top\Big]\right\|_{\mathbf{S}^\top\mathbf{S}}, && (\mathbf{S}, \mathbf{S}^\top \text{ invertible.})\\
&= \left\|\mathbf{S}\mathbf{S}^{-1}(\mathbf{I}_m + \mathbf{D})^{-1}\mathbf{S}^{-\top}\mathbb{E}\Big[X\xi^\top\Big]\right\|, && \text{(Switch to } \ell_2 \text{ norm.)}\\
&= \left\|(\mathbf{I}_m + \mathbf{D})^{-1}\mathbf{S}^{-\top}\mathbb{E}\Big[X\xi^\top\Big]\right\| \leq \left\|\mathbf{S}^{-\top}\mathbb{E}\Big[X\xi^\top\Big]\right\|, && (23)\\
&= \left\|\mathbf{S}\mathbf{S}^{-1}\mathbf{S}^{-\top}\mathbb{E}\Big[X\xi^\top\Big]\right\|, && \text{(Substitute in } \mathbf{I}_m = \mathbf{S}\mathbf{S}^{-1}.)\\
&= \left\|\mathbf{S}^{-1}\mathbf{S}^{-\top}\mathbb{E}\Big[X\xi^\top\Big]\right\|_{\mathbf{S}^\top\mathbf{S}}, && \text{(Back to weighted norm.)}\\
&= \left\|\mathbb{E}\Big[XX^\top\Big]^{-1}\mathbb{E}\Big[X\xi^\top\Big]\right\|_{\mathbf{\Sigma}^{\mathfrak{M}}_X}, && (\text{Substitute in } \mathbf{\Sigma}^{\mathfrak{M}}_X \coloneqq \mathbb{E}^{\mathfrak{M}}\big[XX^\top\big] = \mathbf{S}^\top\mathbf{S}.)\\
&= \left\|\mathbf{f} + \mathbb{E}\Big[XX^\top\Big]^{-1}\mathbb{E}\Big[X\xi^\top\Big] - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{M}}_X}, && \text{(Add and subtract } \mathbf{f}.)\\
&= \left\|\mathbb{E}\Big[XX^\top\Big]^{-1}\Big(\mathbb{E}\Big[XX^\top\Big]\mathbf{f} + \mathbb{E}\Big[X\xi^\top\Big]\Big) - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{M}}_X}, && (\text{Use } \mathbf{I}_m = \mathbb{E}\big[XX^\top\big]^{-1}\mathbb{E}\big[XX^\top\big].)\\
&= \left\|\mathbb{E}\Big[XX^\top\Big]^{-1}\mathbb{E}\Big[X\Big(\mathbf{f}^\top X + \xi\Big)^\top\Big] - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{M}}_X}, && \text{(Linearity of expectation.)}\\
&= \left\|\mathbb{E}\Big[XX^\top\Big]^{-1}\mathbb{E}\Big[XY^\top\Big] - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{M}}_X} = \left\|\hat{\mathbf{h}}^{\mathfrak{A}}_{\mathrm{ERM}} - \mathbf{f}\right\|_{\mathbf{\Sigma}^{\mathfrak{M}}_X}, && \text{(Structural eq. of } Y.)
\end{aligned}
$$

where inequality Eq. (23) holds because $\mathbf{D}$ is non-negative diagonal. Also, inequality Eq. (23) only holds with equality iff $\mathbf{S}^{-\top}\mathbb{E}\big[X\xi^\top\big]$ is in kernel of $\mathbf{D}$. Or equivalently, iff $\mathbb{E}\big[X\xi^\top\big]$ is in the kernel of $\mathbf{S}^\top\mathbf{D}\mathbf{S} = \mathbf{\Sigma}_{\tilde{G}}$, which from Lemma 1 holds iff $\mathbb{E}^{\mathfrak{M}}[GX\,|\,G] \perp \mathbb{E}^{\mathfrak{M}}[X\,|\,\xi]$ a.s. ■

H.4. Proof of Theorem 2 – Robust prediction with IVL regression

Write X in terms of the exogenous variables C, Z, N_X, N_Y using the reduced form from Lemma 3 as

$$X = \tilde{Z} + \tilde{C} + \tilde{N}, \tag{24}$$

where for readability we represent

$$\tilde{Z} \coloneqq \mathbf{M}_{m\times m}\mathbf{\Gamma}^\top Z, \qquad \tilde{C} \coloneqq \mathbf{M}\begin{bmatrix}\mathbf{T}^\top \\ \boldsymbol{\epsilon}^\top\end{bmatrix} C, \qquad \tilde{N} \coloneqq \sigma \cdot \mathbf{M}\begin{bmatrix}N_X \\ N_Y\end{bmatrix},$$

with

$$\mathbf{M} \coloneqq \begin{bmatrix}\mathbf{M}_{m\times m} & \mathbf{M}_{m\times 1} \\ \mathbf{M}_{1\times m} & \mathbf{M}_{1\times 1}\end{bmatrix} = \begin{bmatrix}\mathbf{I}_m & -\boldsymbol{\tau}^\top \\ -\mathbf{f}^\top & 1\end{bmatrix}^{-1}.$$

Now, we start by writing the ERM objective under the intervention $\mathrm{do}\big(\mathbf{\Gamma}^\top(\cdot) \coloneqq \boldsymbol{\zeta}\big)$ as

$$\begin{aligned}
&\Rightarrow R_{\mathrm{ERM}}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}(\mathbf{h}) \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right], \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|\xi + (\mathbf{f}-\mathbf{h})^\top\left(\tilde{Z} + \tilde{C} + \tilde{N}\right)\right\|^2\right], \qquad (Y \text{ structural form \& Eq. (24).}) \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|\xi + (\mathbf{f}-\mathbf{h})^\top\left(\mathbf{M}_{m\times m}\boldsymbol{\zeta} + \tilde{C} + \tilde{N}\right)\right\|^2\right], \quad (\tilde{Z} \text{ \& intervention definition.}) \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|\xi + (\mathbf{f}-\mathbf{h})^\top\left(\tilde{C} + \tilde{N}\right) + (\mathbf{f}-\mathbf{h})^\top\mathbf{M}_{m\times m}\boldsymbol{\zeta}\right\|^2\right], \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|\xi + (\mathbf{f}-\mathbf{h})^\top\left(\tilde{C} + \tilde{N}\right) + \mathbf{h'}^\top\boldsymbol{\zeta}\right\|^2\right], \quad (\text{Define } \mathbf{h'}^\top \coloneqq (\mathbf{f}-\mathbf{h})^\top\mathbf{M}_{m\times m}.) \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|\xi + (\mathbf{f}-\mathbf{h})^\top\left(\tilde{C} + \tilde{N}\right)\right\|^2\right] + \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|\mathbf{h'}^\top\boldsymbol{\zeta}\right\|^2\right], \\
&\qquad\qquad (\text{Follows from exogeneity of } \boldsymbol{\zeta} \text{ under intervention, } \Rightarrow \text{cross term zeros-out.}) \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\mathbf{0}_m\right)}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right] + \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\boldsymbol{\zeta}\right)}\left[\left\|\mathbf{h'}^\top\boldsymbol{\zeta}\right\|^2\right], \qquad (25) \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\mathbf{0}_m\right)}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right] + \left\|\mathbf{h'}^\top\boldsymbol{\zeta}\right\|^2, \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\mathbf{0}_m\right)}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right] + \mathrm{tr}\left(\boldsymbol{\zeta}^\top\mathbf{h'}\mathbf{h'}^\top\boldsymbol{\zeta}\right), \\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot)\coloneqq\mathbf{0}_m\right)}\left[\left\|Y - \mathbf{h}^\top X\right\|^2\right] + \mathrm{tr}\left(\mathbf{h'}^\top\boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathbf{h'}\right). \qquad (26)
\end{aligned}$$

Now, note that the maximum of the trace term over $\boldsymbol{\zeta} \in \mathcal{P}_\alpha$ gives

$$\Rightarrow \max_{\boldsymbol{\zeta}\in\mathcal{P}_\alpha} \mathrm{tr}\left(\mathbf{h'}^\top\boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathbf{h'}\right),$$

$$
\begin{aligned}
&= \left(\frac{1}{\alpha}+1\right)\ \mathrm{tr}\Big(\mathbf{h}'^\top\Big(\mathbf{\Gamma}^\top\mathbb{E}^{\mathfrak{M}}\Big[ZZ^\top\Big]\mathbf{\Gamma}\Big)\mathbf{h}'\Big), && \text{(Linearity of trace and definition of } \mathcal{P}_\alpha.\text{)}\\
&= \left(\frac{1}{\alpha}+1\right)\ \mathbb{E}^{\mathfrak{M}}\Big[\mathrm{tr}\Big(\mathbf{h}'^\top\mathbf{\Gamma}^\top ZZ^\top\mathbf{\Gamma}\mathbf{h}'\Big)\Big], && \text{(Linearity of expectation.)}\\
&= \left(\frac{1}{\alpha}+1\right)\ \mathbb{E}^{\mathfrak{M}}\Big[\mathrm{tr}\Big(Z^\top\mathbf{\Gamma}\mathbf{h}'\mathbf{h}'^\top\mathbf{\Gamma}^\top Z\Big)\Big], && \text{(Cyclic property of trace.)}\\
&= \left(\frac{1}{\alpha}+1\right)\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|\mathbf{h}'^\top\mathbf{\Gamma}^\top Z\Big\|^2\Big],\\
&= \left(\frac{1}{\alpha}+1\right)\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|(\mathbf{f}-\mathbf{h})^\top\mathbf{M}_{m\times m}\mathbf{\Gamma}^\top Z\Big\|^2\Big], && \text{(Substitute in definition of } \mathbf{h}'^\top.\text{)}\\
&= \left(\frac{1}{\alpha}+1\right)\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|(\mathbf{f}-\mathbf{h})^\top\tilde{Z}\Big\|^2\Big]. && \text{(Definition of } \tilde{Z}.\text{)}
\end{aligned}
$$

We can now substitute this in while maximizing both sides of Eq. (26) over interventions $\boldsymbol{\zeta}\in\mathcal{P}_\alpha$ as

$$
\begin{aligned}
&\Rightarrow \max_{\boldsymbol{\zeta}\in\mathcal{P}_\alpha} R_{\mathrm{ERM}}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot):=\mathbf{0}_m\right)}(\mathbf{h})\\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot):=\mathbf{0}_m\right)}\Big[\Big\|Y-\mathbf{h}^\top X\Big\|^2\Big] + \max_{\boldsymbol{\zeta}\in\mathcal{P}_\alpha}\ \mathrm{tr}\Big(\mathbf{h}'^\top\boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathbf{h}'\Big), && \text{(First term does not have } \boldsymbol{\zeta}.\text{)}\\
&= \mathbb{E}^{\mathfrak{M};\mathrm{do}\left(\mathbf{\Gamma}^\top(\cdot):=\mathbf{0}_m\right)}\Big[\Big\|Y-\mathbf{h}^\top X\Big\|^2\Big] + \left(\frac{1}{\alpha}+1\right)\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|(\mathbf{f}-\mathbf{h})^\top\tilde{Z}\Big\|^2\Big],\\
&= \mathbb{E}^{\mathfrak{M}}\Big[\Big\|Y-\mathbf{h}^\top X\Big\|^2\Big] + \frac{1}{\alpha}\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|(\mathbf{f}-\mathbf{h})^\top\tilde{Z}\Big\|^2\Big], && \text{(Inverse step of Eq. (25).)}\\
&= \mathbb{E}^{\mathfrak{M}}\Big[\Big\|Y-\mathbf{h}^\top X\Big\|^2\Big] + \frac{1}{\alpha}\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|(\mathbf{f}-\mathbf{h})^\top\mathbb{E}[X\,|\,Z]\Big\|^2\Big],\\
& && \text{(From conditional exp. of Eq. (24).)}\\
&= \mathbb{E}^{\mathfrak{M}}\Big[\Big\|Y-\mathbf{h}^\top X\Big\|^2\Big] + \frac{1}{\alpha}\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|\mathbb{E}\Big[\mathbf{f}^\top X\,\Big|\,Z\Big]-\mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big\|^2\Big],\\
& && \text{(Linearity of expectation.)}\\
&= \mathbb{E}^{\mathfrak{M}}\Big[\Big\|Y-\mathbf{h}^\top X\Big\|^2\Big] + \frac{1}{\alpha}\ \mathbb{E}^{\mathfrak{M}}\Big[\Big\|\mathbb{E}[Y\,|\,Z]-\mathbf{h}^\top\mathbb{E}[X\,|\,Z]\Big\|^2\Big], && \text{(Inverse step of Eq. (25).)}\\
&= R_{\mathrm{ERM}}^{\mathfrak{M}}(\mathbf{h}) + \frac{1}{\alpha}\Big(R_{\mathrm{IV}}^{\mathfrak{M}}(\mathbf{h}) - \mathbb{E}[\mathbb{V}[Y\,|\,Z]]\Big), && \text{(From Eq. (13).)}\\
&= \frac{1}{\alpha}\Big(R_{\mathrm{IVL}_\alpha}^{\mathfrak{M}}(\mathbf{h}) - \mathbb{E}[\mathbb{V}[Y\,|\,Z]]\Big).
\end{aligned}
$$

■

H.5. Proof of Theorem 3 – Causal estimation with IVL regression

For $\hat{\mathbf{h}}^{\mathfrak{M}}_{\mathrm{IVL}_\alpha}$, we have from Proposition 1

$$\left\|\hat{\mathbf{h}}^{\mathfrak{M}}_{\mathrm{IVL}_\alpha} - \mathbf{f}\right\|^2_{\mathbf{\Sigma}^{\mathfrak{M}}_X} = \left\|\mathbb{E}\left[X'X'^\top\right]^{-1}\mathbb{E}\left[X'Y'^\top\right] - \mathbf{f}\right\|^2_{\mathbf{\Sigma}^{\mathfrak{M}}_X}.$$

Note that we have

$$\begin{aligned}
&\Rightarrow \mathbb{E}\left[X'Y'^\top\right] \\
&= \mathbb{E}\left[X'(aY + b\mathbb{E}[Y\,|\,Z])^\top\right], \\
&= \mathbb{E}\left[X'\Big(aY + b\mathbb{E}\Big[\mathbf{f}^\top X + \xi\,\Big|\,Z\Big]\Big)^\top\right], \\
&= \mathbb{E}\left[X'\Big(aY + b\mathbf{f}^\top\mathbb{E}[X\,|\,Z]\Big)^\top\right], && \text{(Dy definition } Z \perp\!\!\!\perp \xi.) \\
&= \mathbb{E}\left[X'\Big(a\mathbf{f}^\top X + a\xi + b\mathbf{f}^\top\mathbb{E}[X\,|\,Z]\Big)^\top\right], \\
&= \mathbb{E}\left[X'\Big(\mathbf{f}^\top X' + a\xi\Big)^\top\right], && \text{(Substituting in } X' \coloneqq aX + b\mathbb{E}[X\,|\,Z].) \\
&= \mathbb{E}\left[X'X'^\top\mathbf{f} + aX'\xi^\top\right], \\
&= \mathbb{E}\left[X'X'^\top\right]\mathbf{f} + a\mathbb{E}\left[X'\xi^\top\right], \\
&= \mathbb{E}\left[X'X'^\top\right]\mathbf{f} + a^2\mathbb{E}\left[X\xi^\top\right], && (Z \perp\!\!\!\perp \xi\text{, therefore } \mathbb{E}\left[X'\xi^\top\right] = a\mathbb{E}\left[X\xi^\top\right].) \\
&= \mathbb{E}\left[X'X'^\top\right]\mathbf{f} + \alpha\mathbb{E}\left[X\xi^\top\right], && (27)
\end{aligned}$$

We also see that

$$\begin{aligned}
&\Rightarrow \mathbb{E}\left[X'X'^\top\right] \\
&= \mathbb{E}\left[(aX + b\mathbb{E}[X\,|\,Z])(aX + b\mathbb{E}[X\,|\,Z])^\top\right], \\
&= \mathbb{E}\left[\left(aX + b\tilde{Z}\right)\left(aX + b\tilde{Z}\right)^\top\right], && \text{(Set } \tilde{Z} \coloneqq \mathbb{E}[X \mid Z] \text{ for brevity.)} \\
&= a^2\mathbb{E}\left[XX^\top\right] + b^2\mathbb{E}\left[\tilde{Z}\tilde{Z}^\top\right] + ab\mathbb{E}\left[X\tilde{Z}^\top\right] + ab\mathbb{E}\left[\tilde{Z}X^\top\right], \\
&= a^2\mathbb{E}\left[XX^\top\right] + \left(b^2 + 2ab\right)\mathbf{\Sigma}_{\tilde{Z}}, && \text{(Because } \mathbb{E}\left[X\tilde{Z}^\top\right] = \mathbf{\Sigma}_{\tilde{Z}}.) \\
&= \alpha\mathbb{E}\left[XX^\top\right] + \mathbf{\Sigma}_{\tilde{Z}}, && (28)
\end{aligned}$$

where we substituted in Eq. (19) in Eq. (28).

Finally, we now have

$$\Rightarrow \left\|\hat{\mathbf{h}}^{\mathfrak{M}}_{\mathrm{IVL}_\alpha} - \mathbf{f}\right\|^2_{\mathbf{\Sigma}^{\mathfrak{M}}_X}$$

$$
\begin{aligned}
&= \left\| \mathbb{E}\left[X'X'^\top\right]^{-1}\mathbb{E}\left[X'Y'^\top\right] - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, \\
&= \left\| \mathbb{E}\left[X'X'^\top\right]^{-1}\Big(\mathbb{E}\left[X'X'^\top\right]\mathbf{f} + \alpha\mathbb{E}\left[X\xi^\top\right]\Big) - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, && \text{(Substituting in Eq. (27).)} \\
&= \left\| \mathbf{f} + \alpha\mathbb{E}\left[X'X'^\top\right]^{-1}\mathbb{E}\left[X\xi^\top\right] - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, \\
&= \left\| \alpha\mathbb{E}\left[X'X'^\top\right]^{-1}\mathbb{E}\left[X\xi^\top\right]\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, \\
&= \left\| \alpha\Big(\alpha\mathbb{E}\left[XX^\top\right] + \mathbf{\Sigma}_{\tilde{Z}}\Big)^{-1}\mathbb{E}\left[X\xi^\top\right]\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, && \text{(Substituting in Eq. (28).)} \\
&= \left\| \left(\mathbf{S}^\top\mathbf{S} + \frac{1}{\alpha}\mathbf{S}^\top\mathbf{D}\mathbf{S}\right)^{-1}\mathbb{E}\left[X\xi^\top\right]\right\|^2_{\mathbf{S}^\top\mathbf{S}}, && \text{(Using Lemma 2.)} \\
&= \left\| \mathbf{S}^{-1}\left(\mathbf{I}_m + \frac{1}{\alpha}\mathbf{D}\right)^{-1}\mathbf{S}^{-\top}\mathbb{E}\left[X\xi^\top\right]\right\|^2_{\mathbf{S}^\top\mathbf{S}}, && \text{($\mathbf{S}$ is invertible.)} \\
&= \left\| \left(\mathbf{I}_m + \frac{1}{\alpha}\mathbf{D}\right)^{-1}\mathbf{S}^{-\top}\mathbb{E}\left[X\xi^\top\right]\right\|^2, && \text{(Switch to ℓ_2 norm.)} \\
&\leq \left\| \mathbf{S}^{-\top}\mathbb{E}\left[X\xi^\top\right]\right\|^2, && (29) \\
&= \left\| \mathbf{S}\mathbf{S}^{-1}\mathbf{S}^{-\top}\mathbb{E}\left[X\xi^\top\right]\right\|^2, && \text{(Substituting $\mathbf{I} = \mathbf{S}\mathbf{S}^{-1}$.)} \\
&= \left\| \mathbf{S}^{-1}\mathbf{S}^{-\top}\mathbb{E}\left[X\xi^\top\right]\right\|^2_{\mathbf{S}^\top\mathbf{S}}, && \text{(Back to weighted norm.)} \\
&= \left\| \mathbb{E}\left[XX^\top\right]^{-1}\mathbb{E}\left[X\xi^\top\right]\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, && \text{(Substituting $\mathbf{\Sigma}_X^{\mathfrak{M}} := \mathbb{E}^{\mathfrak{M}}\left[XX^\top\right] = \mathbf{S}^\top\mathbf{S}$.)} \\
&= \left\| \mathbf{f} + \mathbb{E}\left[XX^\top\right]^{-1}\mathbb{E}\left[X\xi^\top\right] - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, && \text{(Adding and subtracting $\mathbf{f}$.)} \\
&= \left\| \mathbb{E}\left[XX^\top\right]^{-1}\Big(\mathbb{E}\left[XX^\top\right]\mathbf{f} + \mathbb{E}\left[X\xi^\top\right]\Big) - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, \\
& && \text{(Substituting $\mathbf{I} = \mathbb{E}\left[XX^\top\right]^{-1}\mathbb{E}\left[XX^\top\right]$.)} \\
&= \left\| \mathbb{E}\left[XX^\top\right]^{-1}\mathbb{E}\left[X\Big(\mathbf{f}^\top X + \xi\Big)^\top\right] - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, && \text{(Linearity of expectation.)} \\
&= \left\| \mathbb{E}\left[XX^\top\right]^{-1}\mathbb{E}\left[XY^\top\right] - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}} = \left\| \hat{\mathbf{h}}^{\mathfrak{M}}_{\text{ERM}} - \mathbf{f}\right\|^2_{\mathbf{\Sigma}_X^{\mathfrak{M}}}, && \text{(Substituting $Y = \mathbf{f}^\top X + \xi$.)}
\end{aligned}
$$

where inequality Eq. (29) holds because $\mathbf{D}$ is non-negative diagonal. Additionally, inequality Eq. (29) holds with equality iff $\mathbf{S}^{-\top}\mathbb{E}\left[X\xi^\top\right]$ is in the kernel of $\mathbf{D}$. Equivalently, iff $\mathbb{E}\left[X\xi^\top\right]$ is in kernel of $\mathbf{S}^\top\mathbf{D}\mathbf{S} = \mathbf{\Sigma}_{\tilde{Z}}$, which from Lemma 1 holds iff $\mathbb{E}^{\mathfrak{M}}[X\,|\,Z] \perp \mathbb{E}^{\mathfrak{M}}[X\,|\,\xi]$ a.s. ■

H.6. Miscellaneous supporting lemmas

Lemma 1 (Gaussian conditional orthogonality lemma) *Let $X, Y, Z \in \mathbb{R}^n$ be zero-mean jointly Gaussian random vectors with covariance matrices $\mathbf{\Sigma}_X = \mathbb{E}[XX^\top]$, $\mathbf{\Sigma}_Z = \mathbb{E}[ZZ^\top]$, and cross-covariance $\mathbf{\Sigma}_{Y,Z} = \mathbb{E}[YZ^\top]$. Define the conditional expectation*

$$\mathbb{E}[Y \mid Z] \coloneqq \left(\mathbb{E}\Big[ZZ^\top\Big]^{-1}\mathbb{E}\Big[ZY^\top\Big]\right)^\top Z = \mathbf{\Sigma}_{Y,Z}\mathbf{\Sigma}_Z^{-1}Z.$$

Then the following are equivalent:

$$X \perp \mathbb{E}[Y \mid Z] = 0 \quad a.s. \qquad \Longleftrightarrow \qquad \mathbf{\Sigma}_X\mathbf{\Sigma}_{Y,Z} = \mathbf{0}.$$

Proof Since X, Y, Z are jointly Gaussian, $\mathbb{E}[Y \mid Z] = \mathbf{M}Z$ with $\mathbf{M} \coloneqq \mathbf{\Sigma}_{Y,Z}\mathbf{\Sigma}_Z^{-1}$. The scalar random variable

$$S \coloneqq X^\top\mathbb{E}[Y \mid Z] = X^\top\mathbf{M}Z$$

is Gaussian with mean zero. Hence,

$$S = 0 \quad \text{a.s.} \qquad \Longleftrightarrow \qquad \mathrm{Var}(S) = 0.$$

Compute the variance:

$$\mathrm{Var}(S) = \mathbb{E}\big[S^2\big] = \mathbb{E}\Big[(X^\top\mathbf{M}Z)^2\Big] = \mathbb{E}\Big[Z^\top\mathbf{M}^\top XX^\top\mathbf{M}Z\Big].$$

Using independence and zero-mean assumptions,

$$\mathrm{Var}(S) = \mathrm{tr}\Big(\mathbf{M}^\top\mathbf{\Sigma}_X\mathbf{M}\mathbf{\Sigma}_Z\Big).$$

Since covariance matrices are positive semidefinite, $\mathrm{Var}(S) = 0$ iff

$$\mathbf{\Sigma}_X^{1/2}\mathbf{M}\mathbf{\Sigma}_Z^{1/2} = \mathbf{0} \implies \mathbf{\Sigma}_X\mathbf{M}\mathbf{\Sigma}_Z = \mathbf{0}.$$

Substituting $\mathbf{M} = \mathbf{\Sigma}_{Y,Z}\mathbf{\Sigma}_Z^{-1}$ gives

$$\mathbf{\Sigma}_X\mathbf{\Sigma}_{Y,Z} = \mathbf{0},$$

completing the proof. ■

Lemma 2 (SPD and PSD simultaneous denationalization via congruence) *For any $n \times n$ matrices $\mathbf{A} \succ \mathbf{0}$, $\mathbf{B} \succcurlyeq \mathbf{0}$, there exists an invertible $\mathbf{S} \in \mathbb{R}^{n\times n}$ and non-negative diagonal $\mathbf{D} \in \mathbb{R}^{n\times n}$ such that*

$$\mathbf{A} = \mathbf{S}^\top \mathbf{S}, \qquad \mathbf{B} = \mathbf{S}^\top \mathbf{D}\mathbf{S}.$$

Proof This is similar to Theorem 7.6.4 in (Horn and Johnson, 1985, p. 465) for two SPD matrices. We proceed similarly; Since $\mathbf{A}$ is SPD, it admits a unique SPD square root $\mathbf{A}^{1/2}$. Define

$$\mathbf{C} := \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2},$$

which is SPD. By the spectral theorem, there exists an orthogonal matrix $\mathbf{U}$ such that

$$\mathbf{C} = \mathbf{U}^\top \mathbf{D}\mathbf{U},$$

where $\mathbf{D}$ is diagonal with non-negative entries (the eigenvalues of $\mathbf{C}$). Set

$$\mathbf{S} := \mathbf{U}\mathbf{A}^{1/2}.$$

Then

$$\mathbf{S}^\top \mathbf{S} = \mathbf{A}^{1/2}\mathbf{U}^\top \mathbf{U}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{I}\mathbf{A}^{1/2} = \mathbf{A},$$

and

$$\mathbf{S}^\top \mathbf{D}\mathbf{S} = \mathbf{A}^{1/2}\mathbf{U}^\top \mathbf{D}\mathbf{U}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2} = \mathbf{B}.$$

Since $\mathbf{A}^{1/2}$ and $\mathbf{U}$ are invertible, $\mathbf{S}$ is invertible, completing the proof. ■

Lemma 3 (solvability of simultaneous SEM) *The SEM $\mathfrak{M}$ in Example 2 is solvable iff $\mathbf{f}^\top \boldsymbol{\tau}^\top \neq 1$, in which case the following solution defines the reduced form of the SEM.*

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \mathbf{I}_m & -\boldsymbol{\tau}^\top \\ -\mathbf{f}^\top & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} \boldsymbol{\Gamma}^\top \\ \mathbf{0}_{1\times k} \end{bmatrix} Z + \begin{bmatrix} \mathbf{T}^\top \\ \boldsymbol{\epsilon}^\top \end{bmatrix} C + \sigma \cdot \begin{bmatrix} N_X \\ N_Y \end{bmatrix} \right),$$

Similarly, SEM $\mathfrak{A}$ in Example 1 solves for $\mathbf{f}^\top \boldsymbol{\tau}^\top \neq \kappa^{-1}$.

Proof We re-state the SEM $\mathfrak{M}$ in the following block form

$$\begin{aligned} \begin{bmatrix} X \\ Y \end{bmatrix} &= \begin{bmatrix} \mathbf{0}_{m\times m} & \boldsymbol{\tau}^\top \\ \mathbf{f}^\top & \mathbf{0}_{1\times 1} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Gamma}^\top \\ \mathbf{0}_{1\times k} \end{bmatrix} Z + \begin{bmatrix} \mathbf{T}^\top \\ \boldsymbol{\epsilon}^\top \end{bmatrix} C + \sigma \cdot \begin{bmatrix} N_X \\ N_Y \end{bmatrix}, \\ \Rightarrow \begin{bmatrix} \mathbf{I}_m & -\boldsymbol{\tau}^\top \\ -\mathbf{f}^\top & 1 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \end{bmatrix} &= \begin{bmatrix} \boldsymbol{\Gamma}^\top \\ \mathbf{0}_{1\times k} \end{bmatrix} Z + \begin{bmatrix} \mathbf{T}^\top \\ \boldsymbol{\epsilon}^\top \end{bmatrix} C + \sigma \cdot \begin{bmatrix} N_X \\ N_Y \end{bmatrix} \end{aligned}$$

solving for (X, Y) involves inverting the block matrix on the LHS. The result immediately follows from Proposition 2.8.7 in (Bernstein, 2009, p. 108), via the Schur complement formula for block matrix inversion. ■