A Measurement Perspective on Causal Representations

The Third Pillar of Causal Analysis?

Shimeng Huang
with Dingling Yao, Riccardo Cadei, Kun Zhang, and Francesco Locatello

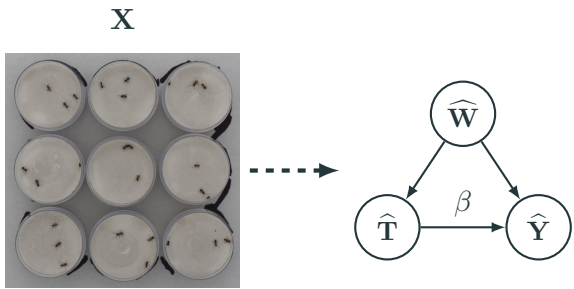
June 6, 2025
@ TGIF Seminar · UNIVERSITY *of* WASHINGTON

Background on Causal Representation Learning

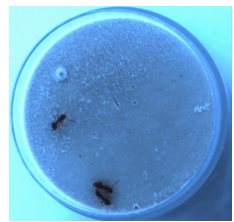
The Idea of Causal Representation Learning (CRL)

An ecological experiment:

ISTAnt consists of *video recordings* of ant triplets where some ants are exposed to a pathogen (treatment) which may change their grooming behavior (outcome). The goal is to estimate the average treatment effect (ATE).



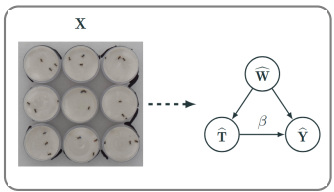
Grooming ants.



This Talk in One Slide

Causal Representation Learning (CRL)

A New Measurement Perspective

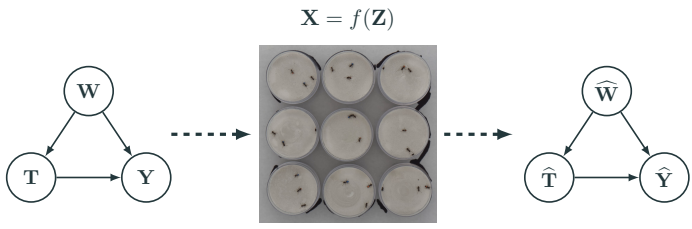


Causal Validity

A New Evaluation Metric

A Brief Overview of CRL Methods

CRL assume that the complex observable $\mathbf{X}$ is generated based on a set of latent variables via a mixing function, where the latent variables are causally related.

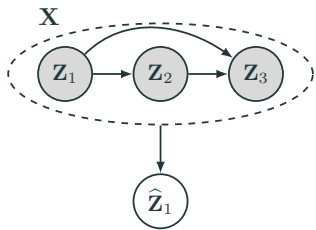


The latent variables $\mathbf{Z} := (\mathbf{W}, \mathbf{T}, \mathbf{Y}) \in \mathbb{R}^N$ are transformed by a function $f : \mathbb{R}^N \to \mathbb{R}^d$, and CRL is then formulated as aiming to recover the latent variables.

A Brief Overview of CRL Methods

It is in general impossible to recover the latent variables and the causal graph without further assumptions on the data generating process (e.g., Locatello et al., 2019).

- Identifiability results have been developed under various assumptions;
- A latent variable Z is considered as *identified* if one can learn from $\mathbf{X}$ a bijective transformation of Z (e.g., von Kügelgen et al., 2021; Yao et al., 2025)

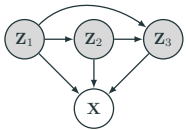


$\mathbf{Z}_1$ is identified by $\widehat{\mathbf{Z}}_{A_1}$ if
$\exists$ a bijective function g s.t. $\widehat{\mathbf{Z}}_{A_1} = g(\mathbf{Z}_1)$.

A Measurement Perspective of CRL

Representing the Causal Representations

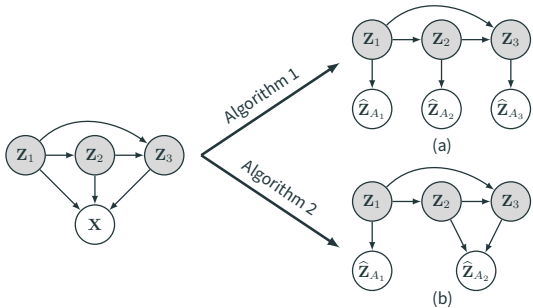
How should we represent CRL methods?



Left: A fully entangled observation X.

Representing the Causal Representations

How should we represent CRL methods? Recall that different CRL methods gives different identifiability results under different assumptions.



Left: A fully entangled observation X. Right: Representation resulted from two CRL algorithms. Observed variables are in gray.

A Measurement Model Framework

Measurement Model

Let $\mathbf{V} = (\mathbf{Z}, \widehat{\mathbf{Z}})$ contain two sets of variables: a set of latent causal variables $\mathbf{Z} = \{\mathbf{Z}_1, \cdots, \mathbf{Z}_N\}$ with $\mathbf{Z}_i$ taking values in $\mathbb{R}$, and a set of observed measurement variables $\widehat{\mathbf{Z}} = \{\widehat{\mathbf{Z}}_{A_1}, \cdots, \widehat{\mathbf{Z}}_{A_M}\}$ where $\widehat{\mathbf{Z}}_{A_j}$ takes values in $\mathbb{R}^{D_j}$ with $D_j \in \mathbb{N}$, $\widehat{\mathbf{Z}} \cap \mathbf{Z} = \emptyset$.

A Measurement Model Framework

Measurement Model

Let $\mathbf{V} = (\mathbf{Z}, \widehat{\mathbf{Z}})$ contain two sets of variables: a set of latent causal variables $\mathbf{Z} = \{\mathbf{Z}_1, \cdots, \mathbf{Z}_N\}$ with $\mathbf{Z}_i$ taking values in $\mathbb{R}$, and a set of observed measurement variables $\widehat{\mathbf{Z}} = \{\widehat{\mathbf{Z}}_{A_1}, \cdots, \widehat{\mathbf{Z}}_{A_M}\}$ where $\widehat{\mathbf{Z}}_{A_j}$ takes values in $\mathbb{R}^{D_j}$ with $D_j \in \mathbb{N}$, $\widehat{\mathbf{Z}} \cap \mathbf{Z} = \emptyset$.
A measurement model $\mathcal{M} = \langle \mathbf{Z}, \widehat{\mathbf{Z}}, \{h_j\}_{j=1}^{M} \rangle$ specifies that

$$\left\{ \widehat{\mathbf{Z}}_{A_j} \coloneqq h_j(\mathbf{Z}_{\mathsf{pa}(\widehat{\mathbf{Z}}_{A_j})}) \right\}_{j=1}^{M},$$

where $\mathsf{pa}(\widehat{\mathbf{Z}}_{A_j}) \subseteq [N]$ for all $j \in [M]$. The functions h_j for all $j \in [M]$ are called the measurement functions. If for some $j \in [M]$, $|\mathsf{pa}(\widehat{\mathbf{Z}}_{A_j})| = 1$ and the function h_j is the identity map, then the causal variable $\mathbf{Z}_{\mathsf{pa}(\widehat{\mathbf{Z}}_{A_j})}$ is said to be directly measured.

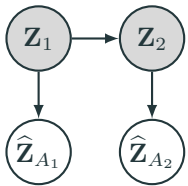
What Can We Do with Causal Representations?

Causal Validity of the Causal Representations

It was mentioned by von Kügelgen et al. (2024) that identification of a latent causal variable up to a bijective transformation is not sufficient for estimating the ATE if either the treatment or outcome is a latent causal variable.

Causal Validity of the Causal Representations

It was mentioned by von Kügelgen et al. (2024) that identification of a latent causal variable up to a bijective transformation is not sufficient for estimating the ATE if either the treatment or outcome is a latent causal variable. E.g.,



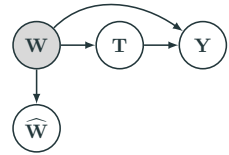
$\mathbf{Z}_2 \coloneqq a \cdot \mathbf{Z}_1 + \epsilon, \epsilon \sim P_\epsilon$ with $\mathbb{E}[\epsilon] = 0$ and $\epsilon \perp\!\!\!\perp \mathbf{Z}_1$. $\widehat{\mathbf{Z}}_{A_i}$ measures $\mathbf{Z}_i$ through a non-linear bijection for both $i = 1, 2$.

True ATE$(\mathbf{Z}_1 \to \mathbf{Z}_2) = \cdots = a$.
Suppose $\widehat{\mathbf{Z}}_{A_1} = \alpha_1 \cdot \mathbf{Z}_1$, $\widehat{\mathbf{Z}}_{A_2} = \alpha_2 \cdot \mathbf{Z}_2$, $\alpha_1, \alpha_2 \neq 0$. Then

$$\text{ATE}(\widehat{\mathbf{Z}}_{A_1} \to \widehat{\mathbf{Z}}_{A_2}) = \cdots = \frac{\alpha_2}{\alpha_1} a.$$

Causal Validity of the Causal Representations

It is, however, possible to estimate the ATE if the measurement variables are only measuring the confounders or instruments.



ATE($\mathbf{T} \to \mathbf{Y}$) is invariant under bijective transformations of the confounder $\mathbf{W}$.

Causally Valid Measurement Model

A measurement model is *"causally valid"* with respect to a estimand g, if the measurement $\widehat{\mathbf{Z}}$ is a drop-in replacement in g for the true causal variables $\mathbf{Z}$, i.e., $g(\mathbf{Z}) = g(\widehat{\mathbf{Z}})$.

Other estimands being invariant up to bijective transformations of the causal variables include mutual information based causal strength (Janzing et al., 2013).

Evaluating Causal Representations

Problems of Current Evaluation Metrics

Given the identifiability results of a CRL method, how should we evaluate empirically if the guarantees are met?

Current evaluation of CRL models usually based on

- R^2 score for predicting the latent variables from the measurement variables;
- Mean coefficient correlations;

However, neither of these metric actually serves the purpose.

Problems of Current Evaluation Metrics: R^2

Suppose that the latent variables follow the following structural causal model:

$$\mathbf{Z}_2 \coloneqq a \cdot \mathbf{Z}_1 + \epsilon,$$

and $\widehat{\mathbf{Z}}_A$ is a measurement variable.

Problems of Current Evaluation Metrics: R^2

Suppose that the latent variables follow the following structural causal model:

$$\mathbf{Z}_2 \coloneqq a \cdot \mathbf{Z}_1 + \epsilon,$$

and $\widehat{\mathbf{Z}}_A$ is a measurement variable. The R^2 score satisfies that

$$R^2(\mathbf{Z}_2, \widehat{\mathbf{Z}}_A) = \cdots = \frac{a^2 \mathbb{V}(\mathbf{Z}_1)}{a^2 \mathbb{V}(\mathbf{Z}_1) + \mathbb{V}(\epsilon)} R^2(\mathbf{Z}_1, \widehat{\mathbf{Z}}_A).$$

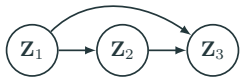
Other problems of R^2: the measurement function (relationship between $\mathbf{Z}$ and $\widehat{\mathbf{Z}}$) can be highly non-linear.

Another Naive Attempt

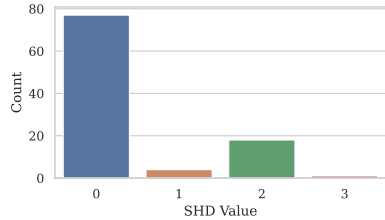
The problems seems to be that the previous metrics ignores the dependencies between the latent variables. A naive attempt to fix this is to look at the *Structural Hamming Distance* (SHD) between the true latent graph and the graph one may find using causal discovery algorithms.

Another Naive Attempt

The problems seems to be that the previous metrics ignores the dependencies between the latent variables. A naive attempt to fix this is to look at the *Structural Hamming Distance* (SHD) between the true latent graph and the graph one may find using causal discovery algorithms.



$$\widehat{\mathbf{Z}}_{A_1} = \gamma_1 \cdot \mathbf{Z}_1 + \gamma_{21} \cdot \mathbf{Z}_2$$
$$\widehat{\mathbf{Z}}_{A_2} = \gamma_2 \cdot \mathbf{Z}_2$$
$$\widehat{\mathbf{Z}}_{A_3} = \gamma_3 \cdot \mathbf{Z}_3$$

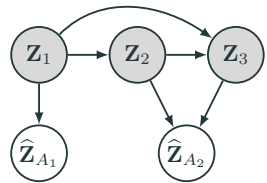


The measurement variables are mixtures of latent variables, yet the graphs discovered from them seems “pretty good”.

Exclusivity of Measurements

Recall that identification means that there is a bijection between a measurement variable and its target latent variable(s), this measurement variable must contain exclusively the information of its targets.

These relationships are encoded in the measurement model. Specifically, a measurement variable $\widehat{\mathbf{Z}}_{A_j}$ exclusively measures $\mathbf{Z}_{\mathsf{pa}(\widehat{\mathbf{Z}}_{A_j})}$.



$$\widehat{\mathbf{Z}}_{A_1} = h_1(\mathbf{Z}_1)$$
$$\widehat{\mathbf{Z}}_{A_2} = h_2(\mathbf{Z}_2, \mathbf{Z}_3)$$

The Test-based Measurement EXclusivity (T-MEX) Score

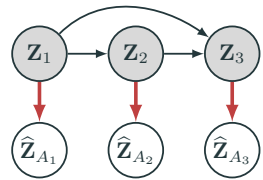
We propose to evaluate a learned CRL model by checking the exclusivity criteria, specifically, by testing $\mathcal{H}_0(i,j) : \widehat{\mathbf{Z}}_{A_j} \perp\!\!\!\perp \mathbf{Z}_i | \mathbf{Z}_{[N]\setminus\{i\}}$ for all $i \in [N]$ and $j \in [M]$.

Test-based Masurement EXclusivity (T-MEX)

Let $V \in \{0,1\}^{N\times M}$ such that $V_{ji} = 1$ if a causal variable $\mathbf{Z}_i$ is a causal parent of a measurement variable $\widehat{\mathbf{Z}}_{A_j}$ according to the measurement model, and $V_{ji} = 0$ otherwise.
Let $\widehat{W} \in \{0,1\}^{N\times M}$ such that $\widehat{W}_{ji} = 1$ if $\mathcal{H}_0(i,j)$ is rejected, and $\widehat{W}_{ji} = 0$ otherwise. Then T-MEX is defined as the Hamming distance between V and $\widehat{W}$:

$$\text{T-MEX}(V, \widehat{W}) := \sum_{j=1}^{M} \sum_{i=1}^{N} \mathbb{1}(V_{ji} \neq \widehat{W}_{ji}),$$

where $\mathbb{1}$ denotes the indicator function.



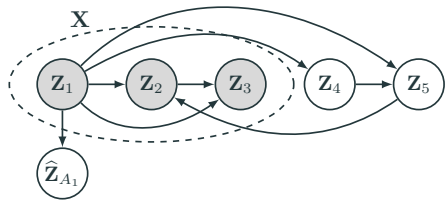
This measurement model gives $V = \text{diag}(3)$.

Experiments

Simulated Experiments (Setup)

We train three CRL models based on a CRL algorithm proposed by Yao et al. (2024):

- **Model A**: a sufficiently trained model, we expect the learned representation $\widehat{\mathbf{Z}}^A_{A_1}$ to *exclusively measure* $\mathbf{Z}_1$;
- **Model B**: an insufficiently trained model, unclear latent-measurement correspondence;
- **Model C**: a corrupted version of Model A, the representation $\widehat{\mathbf{Z}}^C_{A_1}$ is defined as a linear mixing of the identified representation $\widehat{\mathbf{Z}}^A_{A_1}$ and $\mathbf{Z}_2, \mathbf{Z}_3$.



Measurement model according to the identification guarantees of the CRL method under this setting.

Simulated Experiments (Results)

We report the T-MEX scores of the three models (and compare them with R^2), as well as the bias of ATE$(\mathbf{Z}_4 \rightarrow \mathbf{Z}_5)$ when $\widehat{\mathbf{Z}}_{A_1}$ was adjusted for.

Model	T-MEX (↓)
A	0.1200 ± 0.3283
B	1.1800 ± 0.3881
C	2.0000 ± 0.0000

Simulated Experiments (Results)

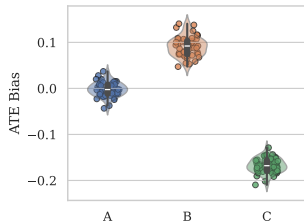
We report the T-MEX scores of the three models (and compare them with R^2), as well as the bias of ATE$(\mathbf{Z}_4 \to \mathbf{Z}_5)$ when $\widehat{\mathbf{Z}}_{A_1}$ was adjusted for.

Model	T-MEX (↓)	$\mathbf{R^2}$		
		$\mathbf{Z}_1$	$\mathbf{Z}_2$	$\mathbf{Z}_3$
A	0.1200 ± 0.3283	0.9984 ± 0.0001	0.7516 ± 0.0064	0.8001 ± 0.0006
B	1.1800 ± 0.3881	0.6665 ± 0.0078	0.8305 ± 0.0032	0.8707 ± 0.0027
C	2.0000 ± 0.0000	0.9394 ± 0.0016	0.5421 ± 0.0096	0.6627 ± 0.0084

Simulated Experiments (Results)

We report the T-MEX scores of the three models (and compare them with R^2), as well as the bias of ATE$(\mathbf{Z}_4 \to \mathbf{Z}_5)$ when $\widehat{\mathbf{Z}}_{A_1}$ was adjusted for.

Model	T-MEX (↓)	$\mathbf{R^2}$		
		$\mathbf{Z}_1$	$\mathbf{Z}_2$	$\mathbf{Z}_3$
A	0.1200 ± 0.3283	0.9984 ± 0.0001	0.7516 ± 0.0064	0.8001 ± 0.0006
B	1.1800 ± 0.3881	0.6665 ± 0.0078	0.8305 ± 0.0032	0.8707 ± 0.0027
C	2.0000 ± 0.0000	0.9394 ± 0.0016	0.5421 ± 0.0096	0.6627 ± 0.0084



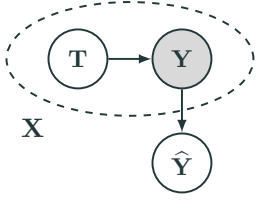
T-MEX aligns with the absolute bias of the ATE estimates of $\mathbf{Z}_4$ on $\mathbf{Z}_5$ where $\widehat{\mathbf{Z}}_1$ is conditioned on as the back door adjustment.

A Real-World Ecological Experiment (Setup)

In the ISTAnt experiment, we have

- A large amount of video recordings of ant triplets where some triplets are exposed to chemicals (treatment $\mathbf{T}$ is available, but outcome $\mathbf{Y}$ is not) $\rightarrow$ we can extract $\widehat{\mathbf{Y}}$ from the video by ML;
- A small amount of samples where $\mathbf{Y}$ is hand-labeled, and $\mathbf{T}$ was randomized $\rightarrow$ we can obtain an unbiased estimate of the ATE

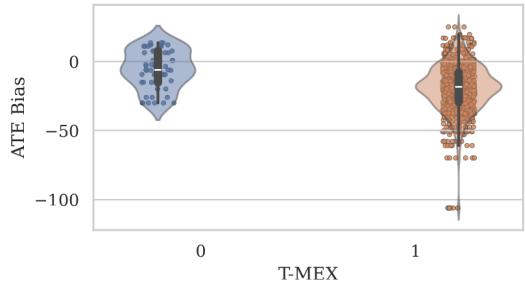
A frame in the videos of the ISTAnt experiment.



A measurement model where $\widehat{\mathbf{Y}}$ measures $\mathbf{Y}$ exclusively.

A Real-World Ecological Experiment (Results)

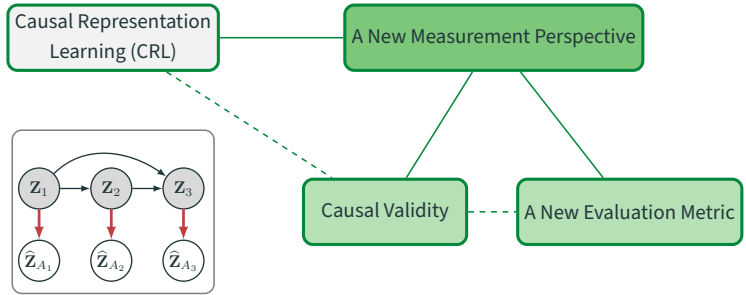
We compute the T-MEX score for 2400 different models, and compare T-MEX with both classification accuracy and ATE bias.



Models with lower T-MEX center their ATE bias near zero with reduced variance.

Conclusion and Future Directions

In this work:



Future works:

- What we can and cannot do with CRL models in the lens of measurement models?
- Can we categorize statistical estimands that are invariant up to bijective transformations of variables?
- Could CRL improve statistical properties of estimators? E.g., Christgau and Hansen (2024)

Contact and Links

Please feel free to reach out: shimeng.huang@ist.ac.at

Link to the preprint

References i

A. M. Christgau and N. R. Hansen. Efficient adjustment for complex covariates: Gaining efficiency with dope. *arXiv preprint arXiv:2402.12980*, 2024.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

A. R. Lundborg, I. Kim, R. D. Shah, and R. J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851–2878, 2024.

J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34:16451–16467, 2021.

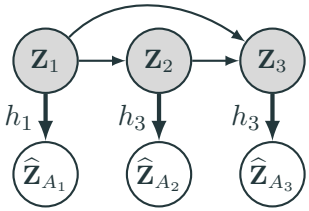
References ii

J. von Kügelgen, M. Besserve, L. Wendong, L. Gresele, A. Kekić, E. Bareinboim, D. Blei, and B. Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024.

D. Yao, D. Rancati, R. Cadei, M. Fumero, and F. Locatello. Unifying causal representation learning with the invariance principle. *The Thirteenth International Conference on Learning Representations*, 2025.

Appendix: Measurement Models and DAGs



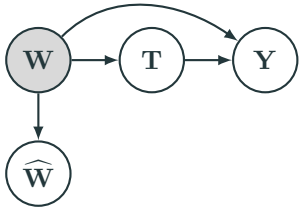
A DAG $\mathcal{G}$ induced by an SCM over $\mathbf{Z}$ and the measurement functions $\{h_j\}_{j\in[3]}$ [a].

[a] In fact, there does not have to exist a DAG among $\mathbf{Z}$, our measurement model framework focuses only on the relationship between $\mathbf{Z}$ and $\widehat{\mathbf{Z}}$

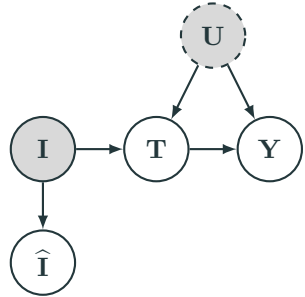
Let P be the joint distribution based on the SCM, and suppose the measurement functions h_1 is a bijection, then

- P is still Markov w.r.t. $\mathcal{G}$;
- P is not faithful to $\mathcal{G}$: e.g., $\mathbf{Z}_1 \perp\!\!\!\perp \mathbf{Z}_2 | \widehat{\mathbf{Z}}_{A_1}$ but $\mathbf{Z}_1 \not\perp\!\!\!\perp_{\mathcal{G}} \mathbf{Z}_2 | \widehat{\mathbf{Z}}_{A_1}$

Appendix: Measurement Models that are Causally Valid w.r.t ATE



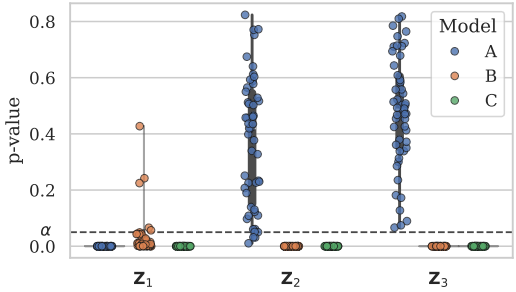
ATE remains invariant under bijective transformation of confounders. The treatment $\mathbf{T}$ and outcome $\mathbf{Y}$ are directly measured (i.e., observed) whereas confounder $\mathbf{W}$ is measured by $\widehat{\mathbf{W}}$ through a non-linear bijection.



ATE remains invariant under bijective transformation of instruments. $\widehat{\mathbf{I}}$ measures the instrument variable $\mathbf{I}$ through a non-linear bijection. The treatment $\mathbf{T}$ and outcome $\mathbf{Y}$ are directly measured (i.e., observed), and $\mathbf{U}$ denotes unobserved confounding.

Appendix: Additional Results in Experiments

We employ the Projected Covariance Measure (PCM, Lundborg et al., 2024) test when computing the T-MEX scores in our experiments.



Violin plots of p-values from testing the conditional independencies $\widehat{\mathbf{Z}}_{A_1} \perp\!\!\!\perp \mathbf{Z}_i | \mathbf{Z}_{[5]\setminus i}$ for $i \in [3]$ based on the PCM tests (Lundborg et al., 2024). The black dashed line is at the significance level $\alpha = 0.05$. A p-value $< \alpha$ for $\mathbf{Z}_i$ means there is an edge from $\mathbf{Z}_i$ to the measurement $\widehat{\mathbf{Z}}_{A_1}$.