Identifying Causal Direction via Variational Bayesian Compression

Quang-Duy Tran Bao Duong Phuoc Nguyen Thin Nguyen

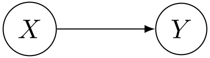
Deakin Applied Artificial Intelligence Initiative, Deakin University, Geelong, Australia

42nd International Conference on Machine Learning

July 2025

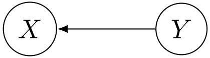
Problem Setup: Cause-Effect Identification from Observational Data

Given X and Y, there are 4 possible cases for their relationship



X causes Y

$(X \to Y)$

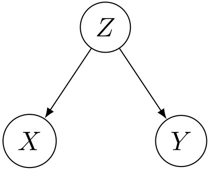


Y causes X

$(Y \to X)$

X and Y are independent

$(X \perp\!\!\!\perp Y)$

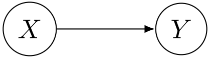


X and Y have a hidden confounder Z $(X \leftarrow Z \rightarrow Y)$

Problem Setup: Cause-Effect Identification from Observational Data

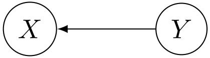
DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

Given X and Y, there are 4 possible cases for their relationship



X causes Y

$(X \to Y)$

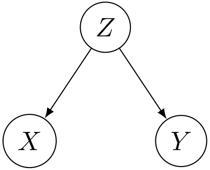


Y causes X

$(Y \to X)$

X and Y are independent

$(X \perp\!\!\!\perp Y)$



X and Y have a hidden confounder Z $(X \leftarrow Z \to Y)$

Causal Asymmetry Interpretation via Kolmogorov Complexity

DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

1. Algorithmic independence of conditionals[1]

$$K\left(p\left(X, Y\right)\right) \stackrel{+}{=} K\left(p\left(X\right)\right) + K\left(p\left(Y \mid X\right)\right)$$

where $K\left(\cdot\right)$ is the Kolmogorov/algorithmic complexity[2]

2. Causal asymmetry in terms of Kolmogorov complexity[3]

$$\underbrace{K\left(p\left(X\right)\right) + K\left(p\left(Y \mid X\right)\right)}_{\Delta_{X \rightarrow Y}} \stackrel{+}{\leq} \underbrace{K\left(p\left(Y\right)\right) + K\left(p\left(X \mid Y\right)\right)}_{\Delta_{Y \rightarrow X}}$$

where $\Delta_{X \rightarrow Y}$ and $\Delta_{Y \rightarrow X}$ are causal indicator scores of $X \rightarrow Y$ and $Y \rightarrow X$ respectively

[1] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic Markov condition," *IEEE Trans. Inf. Theory*, 2010.

[2] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2019.

[3] J. M. Mooij et al., "Probabilistic latent variable models for distinguishing between cause and effect," in *NIPS*, 2010.

Incomputability of Causal Indicator Scores

Issues of the Kolmogorov complexity-based scores[4]:

1. The ground-truth distributions are not available
   - Estimate the joint complexity $K\left(x, p_X\right)$ and $K\left(y, p_{Y|X} \mid x\right)$[5]
2. The Kolmogorov complexity is not computable in practice
   - Approximate with the MDL principle[6] by choosing a model from **a predefined class** $\mathcal{M}$ that minimizes

$$L_M^{\text{2-p}}(D) := \underbrace{L_1(D \mid M)}_{\text{fitness}} + \underbrace{L_2(M)}_{\text{complexity}}, M \in \mathcal{M}$$

$$\hat{\Delta}_{X \to Y}^{\text{2-p}} := L_{M_X^*}^{\text{2-p}}(X) + L_{M_{Y|X}^*}^{\text{2-p}}(Y \mid X)$$

where M_X^* and $M_{Y|X}^*$ are models that minimize the corresponding codelengths.

[4] D. Kaltenpoth and J. Vreeken, "Causal discovery with hidden confounders using the algorithmic Markov condition," in *UAI*, 2023.

[5] A. Marx and J. Vreeken, "Formally justifying MDL-based inference of cause and effect," in *AAAI ITCI'22 Workshop*, 2022.

[6] P. D. Grünwald, *The minimum description length principle*. The MIT Press, 2007.

Classes of Models for Modeling the Conditionals

Classes of models in related cause-effect identification methods:

- A set of basis functions[7]
- Splines[8]
- GP[9] or GPLVM[10]
- Neural networks?
  - Bayesian Compression of Neural Networks for Identifying Causal Direction (COMIC)

[7] A. Marx and J. Vreeken, "Telling cause from effect by local and global regression," *Knowl. Inf. Syst.*, 2019.

[8] A. Marx and J. Vreeken, "Identifiability of cause and effect using regularized regression," in *KDD*, 2019.

[9] J. M. Mooij et al., "Probabilistic latent variable models for distinguishing between cause and effect," in *NIPS*, 2010.

[10] A. Dhir et al., "Bivariate causal discovery using Bayesian model selection," in *ICML*, 2024.

COMIC: Variational Bayesian Codelengths of Neural Networks

- Variational Bayesian (VB) codelength of a neural network[11,12,13]:

$$L^{\text{2-p}}_{p(\boldsymbol{\theta})}\left(y^{(1:N)} \mid x^{(1:N)}\right) := \underbrace{-\log p\left(y^{(1:N)} \mid x^{(1:N)}, \boldsymbol{\theta}\right)}_{\text{fitness}} \underbrace{-\log p\left(\boldsymbol{\theta}\right)}_{\text{complexity}}$$

$$\begin{aligned} L^{\text{var}}_{q_\phi(\boldsymbol{\theta})}\left(y^{(1:N)} \mid x^{(1:N)}\right) &:= \mathbb{E}_{q_\phi(w)}\left[L^{\text{2-p}}_{p(\boldsymbol{\theta})}\left(y^{(1:N)} \mid x^{(1:N)}, \boldsymbol{\theta}\right)\right] - H\left(q_\phi\left(\boldsymbol{\theta}\right)\right) \\ &:= \underbrace{-\mathbb{E}_{q_\phi(w)}\left[\log p\left(y^{(1:N)} \mid x^{(1:N)}, \boldsymbol{\theta}\right)\right]}_{\text{fitness}} + \underbrace{\mathrm{KL}\left(q_\phi\left(\boldsymbol{\theta}\right) \,||\, p\left(\boldsymbol{\theta}\right)\right)}_{\text{complexity}} \end{aligned}$$

- The VB codelength is the upper bound of the marginal Bayesian codelength:

$$L^{\text{Bayes}}_{p(\boldsymbol{\theta})}\left(y^{(1:N)} \mid x^{(1:N)}\right) := -\log \int p\left(y^{(1:N)} \mid x^{(1:N)}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}$$

11 A. Honkela and H. Valpola, "Variational learning and bits-back coding: An information-theoretic view to Bayesian learning," *IEEE Trans. Neural. Netw.*, 2004.

12 C. Louizos et al., "Bayesian compression for deep learning," in *NeurIPS*, 2017.

13 L. Blier and Y. Ollivier, "The description length of deep learning models," in *NeurIPS*, 2018.

COMIC: Identifying Causal Direction via the VB Codelengths

- Likelihood distribution for the conditionals $p\left(y \mid x, \boldsymbol{\theta}\right)$:

$$p\left(y^{(1:N)} \mid x^{(1:N)}, \boldsymbol{\theta}\right) := \prod_{i=1}^{N} \mathcal{N}\left(y^{(i)} \mid \mu\left(x^{(i)}; \boldsymbol{\theta}\right), \sigma^2\left(x^{(i)}; \boldsymbol{\theta}\right)\right)$$

where $\mu\left(\cdot; \boldsymbol{\theta}\right)$ and $\sigma\left(\cdot; \boldsymbol{\theta}\right)$ are modeled via a neural network $\mathbf{f}_Y : \mathbb{R} \times \boldsymbol{\Theta} \to \mathbb{R}^2$ as

$$\mu\left(\cdot; \boldsymbol{\theta}\right) = f_{Y,1}\left(\cdot; \boldsymbol{\theta}\right) \text{ and } \sigma\left(\cdot; \boldsymbol{\theta}\right) = \zeta\left(f_{Y,2}\left(\cdot; \boldsymbol{\theta}\right)\right), \zeta : \mathbb{R} \to (0, +\infty)$$

- Distribution for encoding the marginals $p\left(x\right)$:

$$p\left(x\right) := \mathcal{N}\left(x \mid 0, 1\right), L_{\mathcal{N}}\left(x^{(1:N)}\right) := -\sum_{i=1}^{N} \log p\left(x^{(i)}\right)$$

COMIC: Causal Identifiability

DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

- Approximated causal indicator score:

$$\hat{\Delta}^{\mathrm{var}}_{X\to Y}\left(\mathcal{D}^N\right) := L_{\mathcal{N}}\left(x^{(1:N)}\right) + L^{\mathrm{var}}_{q_{\phi^*}(\boldsymbol{\theta})}\left(y^{(1:N)} \mid x^{(1:N)}\right)$$

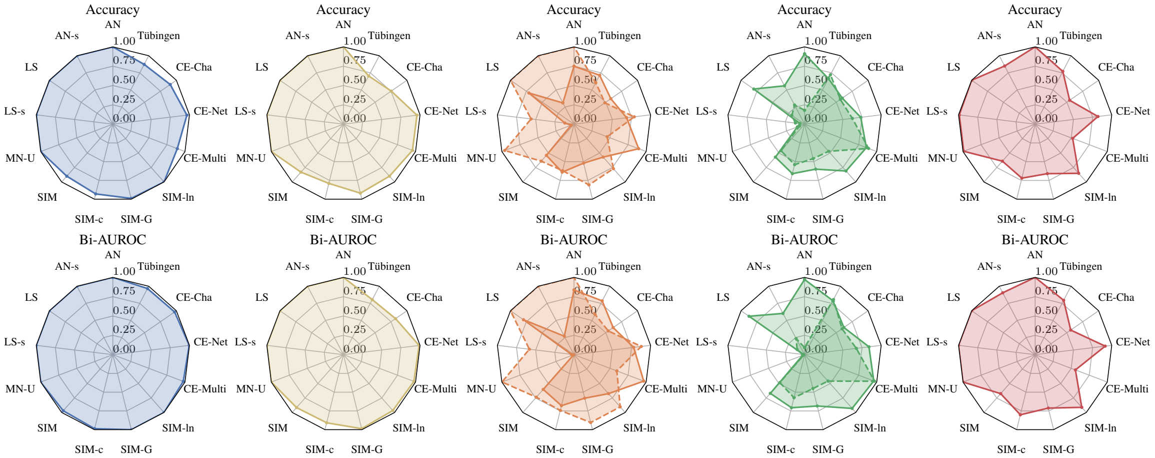
- If $q_{\phi^*}(\boldsymbol{\theta})$ converges to $p\left(\boldsymbol{\theta} \mid x^{(1:N)}, y^{(1:N)}\right)$ and $\mathcal{N}(0,1)$ is the ground-truth distribution of $p(X)$, $\hat{\Delta}^{\mathrm{var}}_{X\to Y}\left(\mathcal{D}^N\right)$ will become

$$\begin{aligned}\Delta^{\mathrm{Bayes}}_{X\to Y}\left(\mathcal{D}^N\right) &:= L_{\mathcal{N}}\left(x^{(1:N)}\right) + L^{\mathrm{Bayes}}_{p(\boldsymbol{\theta})}\left(y^{(1:N)} \mid x^{(1:N)}\right) \\ &:= -\log \underbrace{p\left(\mathcal{D}^N \mid M_{X\to Y}\right)}_{\text{marginal likelihood}}\end{aligned}$$

- *The identifiability of COMIC is verifiable via the identifiability results of marginal likelihoods*[14]

[14] A. Dhir et al., "Bivariate causal discovery using Bayesian model selection," in *ICML*, 2024.

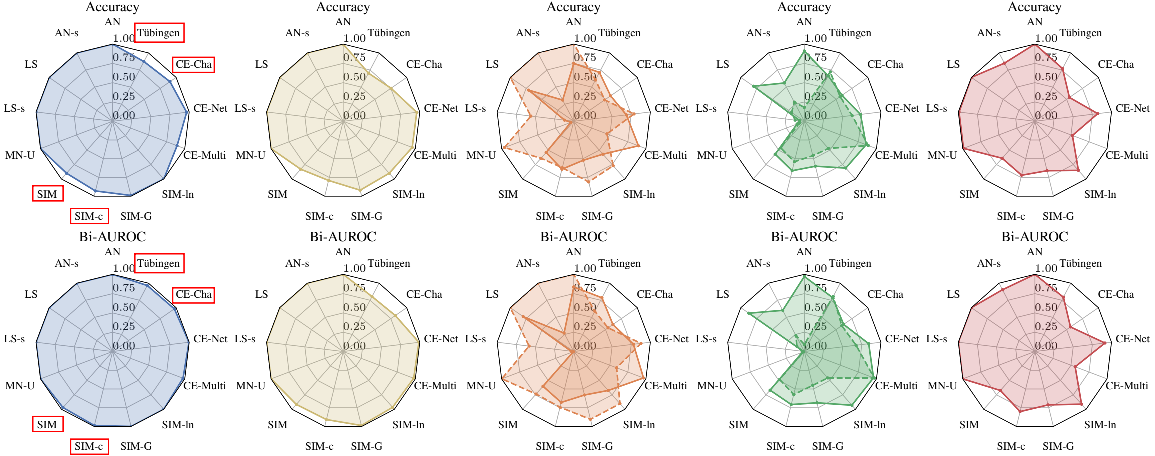
Empirical Performance on Synthetic & Real-World Benchmarks I



(a) **COMIC (Ours)** (b) GPLVM (c) SLOPPY (d) SLOPE (e) QCCD

(——: AIC, – –: BIC) (——: SLOPER, – –: SLOPE)

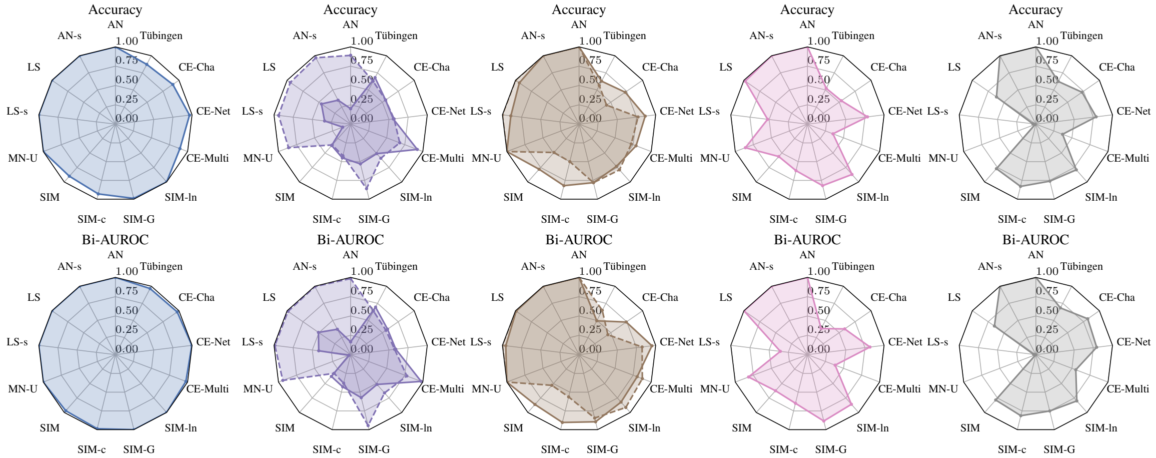
Empirical Performance on Synthetic & Real-World Benchmarks I



(a) **COMIC (Ours)** (b) GPLVM (c) SLOPPY (d) SLOPE (e) QCCD

(——: AIC, – –: BIC) (——: SLOPER, – –: SLOPE)

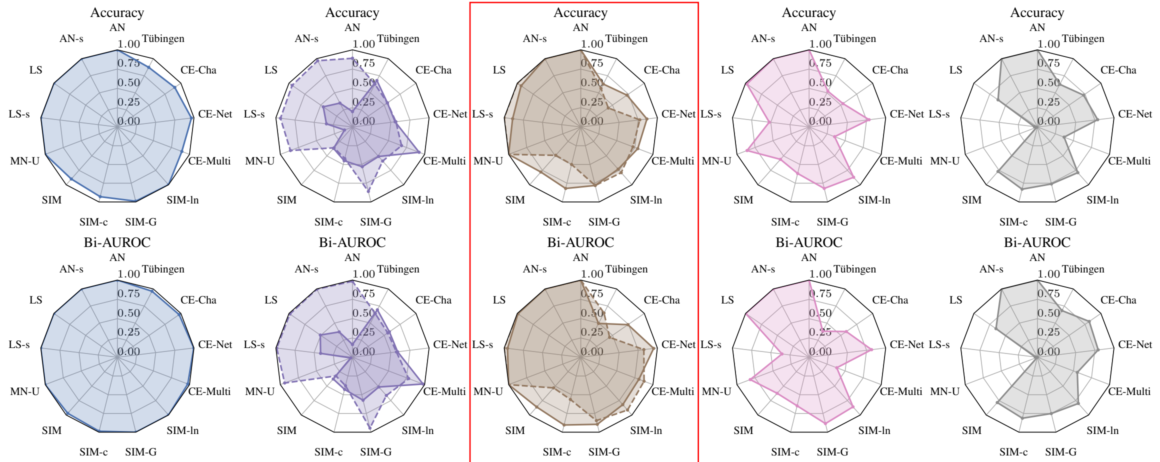
Empirical Performance on Synthetic & Real-World Benchmarks II



(a) **COMIC (Ours)** (f) IGCI (g) LOCI (h) CAM (i) RESIT

(——: unif., – –: Gauss.) (——: HSIC, – –: likelihood)

Empirical Performance on Synthetic & Real-World Benchmarks II



(a) **COMIC (Ours)** (f) IGCI (g) LOCI (h) CAM (i) RESIT

(f) IGCI: (——: unif., – –: Gauss.)

(g) LOCI: (——: HSIC, – –: likelihood)

Summary

- We introduce COMIC—an approach based on variational Bayesian compression of neural networks for cause-effect identification
- The variational Bayesian codelengths in COMIC allow the identifiability to be justified from the marginal likelihood perspective
- The effectiveness of COMIC is demonstrated through comprehensive experiments on 13 benchmarks, which deliver promising results compared to related methods

Thank you!

DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

- Paper: Identifying Causal Direction via Variational Bayesian Compression
- Contact: Quang-Duy Tran <q.tran@deakin.edu.au>
- Code: `https://github.com/quangdzuytran/COMIC`

Preprint

Code

References I

[1] L. Blier and Y. Ollivier, "The description length of deep learning models," in *NeurIPS*, 2018.

[2] A. Dhir, S. Power, and M. van der Wilk, "Bivariate causal discovery using Bayesian model selection," in *ICML*, 2024.

[3] P. D. Grünwald, *The minimum description length principle*. The MIT Press, 2007.

[4] A. Honkela and H. Valpola, "Variational learning and bits-back coding: An information-theoretic view to Bayesian learning," *IEEE Trans. Neural. Netw.*, 2004.

[5] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic Markov condition," *IEEE Trans. Inf. Theory*, 2010.

[6] D. Kaltenpoth and J. Vreeken, "Causal discovery with hidden confounders using the algorithmic Markov condition," in *UAI*, 2023.

[7] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2019.

[8] C. Louizos, K. Ullrich, and M. Welling, "Bayesian compression for deep learning," in *NeurIPS*, 2017.

References II

[9] A. Marx and J. Vreeken, “Identifiability of cause and effect using regularized regression,” in *KDD*, 2019.

[10] A. Marx and J. Vreeken, “Telling cause from effect by local and global regression,” *Knowl. Inf. Syst.*, 2019.

[11] A. Marx and J. Vreeken, “Formally justifying MDL-based inference of cause and effect,” in *AAAI ITCI’22 Workshop*, 2022.

[12] J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, and Schölkopf, “Probabilistic latent variable models for distinguishing between cause and effect,” in *NIPS*, 2010.