A Compression-Based Dependence Measure for Causal Discovery by Additive Noise Models

Hans Jarett Ong[1], Brian Godwin Lim[1], Renzo Roel Tan[1,2], Kazushi Ikeda[1]

[1]Nara Institute of Science and Technology, [2]Ateneo de Manila University

Introduction

Graphical causality [1, 2] addresses key limitations in machine learning, such as out-of-distribution generalization, by modeling the data-generating process instead of individual data points. Causal discovery methods aim to learn the causal graph from data, which is essential for causal inference and enables interventional and counterfactual predictions. Many causal discovery methods rely on testing dependence relations from data, which require effective measures of dependence. This work explores data compression as a novel way to measure dependence for causal discovery, inspired by its use in estimating entropy and mutual information (e.g., PNG file size for images [3]) and in text classification [4].

Additive Noise Models

Additive noise models (ANMs) are expressed as:

$$X_i := f_i(\mathsf{PA}_i) + N_i, \tag{1}$$

where f_i represents a deterministic function of the parent variables PA_i and N_i denotes the noise term that is typically assumed to be normally distributed.

Causal Additive Models (CAMs) are a subtype of ANMs that represent the output variable X_i as a sum of nonlinear functions of its parent variables:

$$X_i := \sum_{j \in \mathsf{PA}_i} f_{i,j}(X_j) + N_i, \tag{2}$$

Linear Non-Gaussian Acyclic Models (LiNGAM) represent another significant subtype of ANMs that operate under the assumption of linear relationships among variables combined with non-Gaussian noise. LiNGAM is formulated as a system of linear equations:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{3}$$

where $\mathbf{B}$ is a strictly lower triangular adjacency matrix, and the elements of $\mathbf{e}$ are continuous non-Gaussian variables with zero mean and nonzero variance.

Normalized Compression Distance

Normalized compression distance (NCD), initially proposed by [5], is defined as

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}, \tag{4}$$

where $C(x)$, $C(y)$, and $C(xy)$ denote the compressed sizes of x, y, and the concatenation of x and y, respectively. The NCD offers a practical realization of the information distance using real-world compressors like `gzip`, `bzip2`, `PPMZ`, etc. In this work, we propose a modified version of NCD tailored for continuous data as the dependence measure for RESIT.

Methods

We propose a compression-based dependence metric, denoted as DM_{comp}, inspired by the normalized compression distance (NCD) but tailored for continuous or floating-point data. The key adaptation involves redefining the compression operation $C(\cdot)$ (from eq. 4)

Redefining $C(x)$ for Continuous Data. We redefine $C(x)$ by transforming the input vector through the following steps:

1. Obtain the pairwise difference matrix $\mathbf{D}$, where $D_{ij} = |x_i - x_j|$.
2. Flatten $\mathbf{D}$, where "flatten" refers to concatenating all rows of a matrix into a single vector.
3. Convert each element of flatten($\mathbf{D}$) from floating point to its string representation and concatenate them using some delimiter (e.g., `|`).

Redefining $C(xy)$ for Continuous Data. To redefine $C(xy)$, we replace the difference matrix $\mathbf{D}$ with the element-wise maximum difference matrix $\mathbf{M}$, where $M_{ij} = \max\{|x_i - x_j|, |y_i - y_j|\}$.

Aggregating the Dependence Measurements. RESIT requires one-to-many comparisons for its dependence metric. The proposed metric, DM_{comp}, currently operates in a pairwise manner and does not inherently satisfy this requirement. To address this, we introduce a simplifying assumption that the one-to-many dependence can be approximated by averaging the pairwise dependence measures. Formally, we define:

$$DM_{comp}((X_1, X_2, \ldots, X_N), Y) = \frac{1}{N} \sum_{i=1}^{N} DM_{comp}(X_i, Y). \tag{5}$$

Results and Discussion

Performance on Simulated Data. Here we present the results of our simulations, using boxplots to show the ordering error,

$$E_o = \frac{2r}{m(m-1)}, \tag{6}$$

for ANM, CAM, and LiNGAM simulations. These simulations test the performance of RESIT [6] with DM_{comp} against existing baselines.

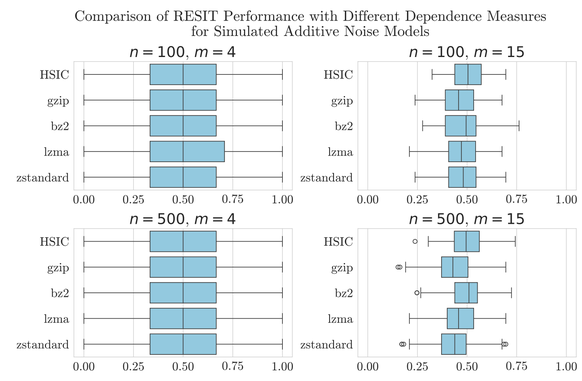

Figure 1: Ordering errors E_o for the ANM simulations using RESIT with random forest regressors and various dependence metrics.

Figure 1 shows that for $m = 15$ some compressors, praticularly `gzip`, performs slightly better than the baseline, Hilbert-Schmidt Independence Criterion (HSIC), but the difference is small.

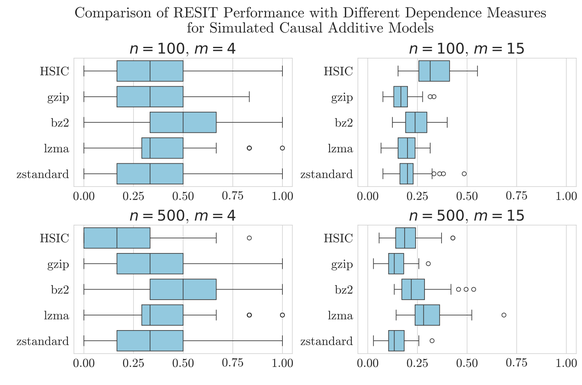

Figure 2: Ordering errors E_o for the CAM simulations using RESIT with random forest regressors and various dependence metrics.

Meanwhile, Figure 2 shows a more pronounced difference between HSIC and the compressors. For $m = 4$, HSIC still performs better or at least as well as the compressors. However, for $m = 15$, the compressors tend to outperform HSIC, with the difference being more apparent for smaller sample sizes ($n = 100$). This suggests a possible strength of using DM_{comp} for smaller sample sizes.

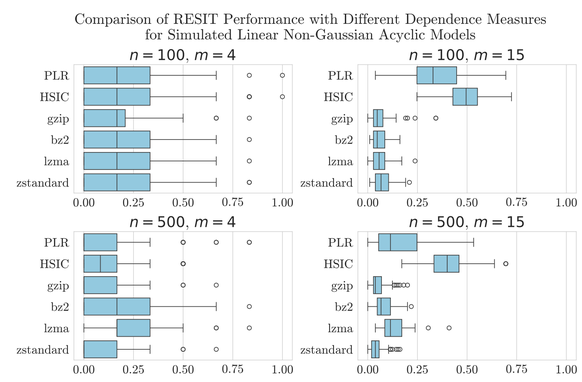

Figure 3: Ordering errors E_o for the LiNGAM simulations. PLR refers to DirectLiNGAM with pairwise likelihood ratios, while the rest are RESIT with linear regressors.

Finally, Figure 3 shows that the compressors consistently outperform the baseline for $m = 15$. Similar to the CAM case (Figure 2), the difference is more pronounced for the smaller sample size ($n = 100$). This indicates that our method can perform well even with small sample sizes, while the baseline methods, PLR DirectLiNGAM and HSIC RESIT, need larger sample sizes to perform better.

Summary and Limitations

We introduced a compression-based dependence metric, DM_{comp}, for causal discovery and demonstrated its effectiveness through simulations with ANM, CAM, and LiNGAM models. DM_{comp} performed competitively with existing metrics like HSIC and PLR, especially with smaller sample sizes. However, one limitation of DM_{comp} is its sensitivity to data scale, similar to methods like NOTEARS and GOLEM [7]. Despite this limitation, our work shows the potential of integrating data compression with causal discovery. Future research could focus on developing scale-invariant versions of DM_{comp}.

References

[1] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: a primer*. Chichester, West Sussex: Wiley, 2016.

[2] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*, ser. Adaptive computation and machine learning series. Cambridge, Massachuestts: The MIT Press, 2017.

[3] M. Zbili and S. Rama, "A quick and easy way to estimate entropy and mutual information for neuroscience," *Frontiers in Neuroinformatics*, vol. 15, 06 2021.

[4] Z. Jiang, M. Yang, M. Tsirlin, R. Tang, Y. Dai, and J. Lin, ""low-resource" text classification: A parameter-free classification method with compressors," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6810–6828. [Online]. Available: https://aclanthology.org/2023.findings-acl.426

[5] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[6] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *Journal of Machine Learning Research*, vol. 15, no. 58, pp. 2009–2053, 2014. [Online]. Available: http://jmlr.org/papers/v15/peters14a.html

[7] A. Reisach, C. Seiler, and S. Weichwald, "Beware of the simulated dag! causal discovery benchmarks may be easy to game," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 772–27 784, 2021.