Proceedings of Machine Learning Research 323:1–22, 2026 5th Conference on Causal Learning and Reasoning

Bivariate Causal Discovery Using Rate-Distortion MDL: An Information Dimension Approach

Tiago Brogueira TIAGO.BROGUEIRA@TECNICO.ULISBOA.PT
Mário A. T. Figueiredo MARIO.FIGUEIREDO@TECNICO.ULISBOA.PT
Instituto de Telecomunicações
Instituto Superior Técnico
Universidade de Lisboa, Portugal

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Approaches to bivariate causal discovery based on the *minimum description length* (MDL) principle approximate the (uncomputable) Kolmogorov complexity of the models in each causal direction, selecting the one with the lower total complexity. The premise is that nature's mechanisms are *simpler* in their true causal order. Inherently, the description length (complexity) in each direction includes the description of the cause variable and that of the causal mechanism. In this work, we argue that current state-of-the-art MDL-based methods do not correctly address the problem of estimating the description length of the cause variable, effectively leaving the decision to the description length of the causal mechanism. Based on *rate-distortion theory*, we propose a new way to measure the description length of the cause, corresponding to the minimum rate required to achieve a distortion level representative of the underlying distribution. This distortion level is deduced using rules from histogram-based density estimation, while the rate is computed using the related concept of information dimension, based on an asymptotic approximation. Combining it with a traditional approach for the causal mechanism, we introduce a new bivariate causal discovery method, termed *rate-distortion* MDL (RDMDL). We show experimentally that RDMDL achieves competitive performance on the Tübingen dataset. All the code and experiments are publicly available at `github.com/tiagobrogueira/Causal-Discovery-In-Exchangeable-Data`.

Keywords: Causal discovery, cause-effect pairs, minimum description length, information dimension, rate-distortion theory.

1. Introduction: Bivariate Causal Discovery

Science seeks to identify cause-and-effect relationships, not just to understand the world, but to intervene on it and answer 'what if' questions (Pearl, 2009). Traditionally, causal relationships are inferred via experiments, either deliberate (e.g., randomized trials) (Guo et al., 2020) or natural (Leatherdale, 2019). When such experiments are infeasible, impractical, unavailable, or unethical (Glymour et al., 2019), as is often the case, *causal discovery* aims at learning causal structure from purely observational data (Eberhardt, 2017; Peters et al., 2017; Zanga et al., 2022).

In causal discovery, a key distinction arises between time-series data, where temporal ordering provides causal cues, versus exchangeable data (usually assumed to be independent, identically distributed – i.i.d.), where no such temporal information exists (Hasan et al., 2023; Guo et al., 2024). In the latter, when considering only two variables (X, Y), both causal directions ($X \to Y$ and $Y \to X$) belong to the same Markov equivalence class (MEC), which is the smallest causal graph that can be deduced from i.i.d. data with no further assumptions (Spirtes et al., 2001). The bivariate

© 2026 T. Brogueira & M.A.T. Figueiredo.

case, also known as the cause-effect problem, has emerged as the quintessential causal discovery problem and has received considerable attention (Guyon et al., 2019). While general causal discovery seeks to recover entire causal graphs, often validated on synthetic or pseudo-real datasets due to the scarcity of real-world benchmarks, bivariate causal discovery often relies on stricter assumptions on the underlying causal mechanisms, allowing for stronger empirical validation (Eberhardt, 2017; Chickering, 2002; Mooij et al., 2016; Guyon et al., 2019; Peters et al., 2017; Zanga et al., 2022).

A common assumption in bivariate causal discovery (which we adopt) is *causal sufficiency*, *i.e.*, the absence of hidden confounders, thus excluding the possibility of hidden confounders being responsible for the observed dependency between the two variables (Guyon et al., 2019; Tran et al., 2025). The most common approaches to tackle the cause-effect problem are score-based, casting the problem as optimizing an objective function. This objective function embodies desired properties of the causal system (Zanga et al., 2022), and usually is either: a measure of independence between system components, *e.g.*, the cause and the causal mechanism (Zhang and Hyvärinen, 2010; Figueiredo and Oliveira, 2023); a measure of fitness to certain model assumptions (Goudet et al., 2018; Blöbaum et al., 2018); a measure of model complexity, *e.g.*, the distance to a non-informative distribution (Chen et al., 2014) or via the *minimum description length* (MDL) principle (Marx and Vreeken, 2017; Tagasovska et al., 2020)).

2. Background and Contributions

2.1. Independent Causal Mechanism, Kolmogorov Complexity, and MDL

The *independent causal mechanism* (ICM) principle states that the causal generative process is composed of autonomous modules that do not inform or influence each other. It is a basic assumption underlying the possibility of causal discovery from observations (Papineau, 2022; Pearl, 2009; Peters et al., 2017). The ICM principle can be formalized using the notion of (Kolmogorov) *algorithmic independence* as stating that, if random variable X causes Y, then the probability distribution of the cause, $P(X)$, and the causal mechanism $P(Y|X)$ are *algorithmically independent* (Janzing and Scholkopf, 2010), i.e. knowing $P(X)$ does not enable a *shorter description* of $P(Y|X)$ or vice versa. It has been shown (Stegle et al., 2010; Janzing and Scholkopf, 2010) that this algorithmic independence condition implies an inequality between the Kolmogorov complexities (KC) of the two possible causal directions,

$$K(P(X)) + K(P(Y|X)) \stackrel{+}{\leq} K(P(Y)) + K(P(X|Y)), \tag{1}$$

where $\stackrel{+}{\leq}$ stands for "less or equal up to a constant". Since KC is uncomputable (Li and Vitányi, 2008), the criterion in Eq. (1) can not be used directly, and some alternative is needed. For example, Janzing et al. (2012) proposed an information-geometric approach to avoid the uncomputable KC.
Note: in the next paragraph, we will overload the notation X and Y, which will also be used to denote sets of N i.i.d. paired samples from random variables X and Y, respectively. That is $X = (x_1, ..., x_N)$ and $Y = (y_1, ..., y_N)$.

Marx and Vreeken (2021) proposed using the MDL principle as a proxy for the uncomputable KC, making the following two concessions: 1) while KC is not restricted to a specific family of models, MDL requires a well-defined class of models; 2) whereas KC measures the shortest description of the underlying probability distribution, MDL computes the shortest description of the data (set

of samples). They argue that the approximation is valid if the family of models is sufficiently expressive to capture the underlying causal mechanisms and the sampled data is representative of the underlying distribution. The MDL principle for the cause-effect problem corresponds to choosing the $X \to Y$ direction if $L(X \to Y) < L(Y \to X)$, where (and reciprocally for $L(Y \to X)$),

$$L(X \to Y) = L(M) + L(X|M) + L(Y|M, X), \tag{2}$$

where $L(\cdot)$ denotes *description length* (*e.g.*, in bits) and M denotes the model. $L(X|M)$ is the description length of the samples of the cause variable, and $L(Y|X, M)$ is that of the samples of the effect, given the corresponding cause samples. Score-based approaches using the MDL principle include two of the best performing methods: *bivariate quantile causal discovery* (bCQD, Tagasovska et al. (2020)) and SLOPE (Marx and Vreeken, 2017).

Recall that MDL was originally proposed as a criterion to select a model in some model class $\mathcal{M}$, according to

$$\widehat{M} = \arg\min_{M \in \mathcal{M}} L(M) + L(D|M), \tag{3}$$

where D is the observed data (Rissanen, 1978). Intuitively, a more complex model (higher $L(M)$) describes the data more precisely (lower $L(D|M)$), while a simpler model (lower $L(M)$) leaves more structure unexplained in the data (higher $L(D|M)$). MDL elegantly formalizes this trade-off, offering a non-arbitrary criterion based on information theory (Rissanen, 1978; Grünwald, 2007).

2.2. Computing Conditional Description Lengths

In this section, we discuss how several MDL-based approaches to bivariate causal discovery define the encoding length of the conditional distribution, i.e. $L(Y|X) = L(Y|X, M) + L(M)$. Since this paper focuses on continuous variables, we only address in detail methods for this type of data. Nevertheless, we mention that Budhathoki and Vreeken (2017) deal solely with discrete data and use *stochastic complexity* (Grünwald, 2007; Rissanen, 2007) to define the required description lengths. Stochastic complexity is the negative logarithm of the *normalized maximum likelihood* (NML) distribution, and it is minimax optimal: even if the true distribution is not in the model class, the NML-based stochastic complexity is optimal relative to that class (Grünwald, 2007).

For continuous variables, to the best of our knowledge, there are four methods that use MDL for bivariate causal discovery: bQCD (Tagasovska et al., 2020), COMIC (Bayesian COMpression-based approach to identifying the Causal direction, by Tran et al. (2025)), LCUBE (Hlavackova-Schindler and Marsela, 2025), and SLOPE (Marx and Vreeken, 2017) and variations thereof.

We start by examining how these methods compute $L(Y|X, M)$. SLOPE uses a combination of global and local regression, both minimizing the sum of squared errors, implicitly corresponding to maximizing the likelihood under Gaussian noise. The residuals are thus encoded under this Gaussian assumption, yielding $L(Y|X, M) \propto \hat{\sigma}^2 \, [P(Y|X, M)]$, the estimated variance of the conditional distribution $P(Y|X, M)$. Since the objective is to minimize the encoding length, encoding the residuals as a Gaussian distribution is only optimal when these follow a Gaussian distribution (Grünwald, 2007). Similarly, LCUBE also minimizes the sum of squared errors, encodes the residuals under a Gaussian assumption, with the key difference with respect to SLOPE being the choice of cubic regression splines as model functions.

Tagasovska et al. (2020) use the *pinball loss* (a.k.a. *quantile loss*, which is an asymmetric weighted ℓ_1 loss denoted as S)) to fit the model and encode the residuals using the asymmetric

Laplace distribution,

$$L(Y|X, M) \propto \sum_{j=1}^{m} w_{\tau_j} \sum_{i}^{N} S(Y_i, q_{X,\tau_j}) \tag{4}$$

where τ_j indicates the j-th quantile, q_{X,τ_j} indicates the value of the estimated j-th quantile of Y based on (X, τ_j), and m is the total number of quantiles in consideration.

Finally, COMIC (Tran et al., 2025) uses a Bayesian formulation to express the model,

$$L(Y|X, M) = -\log \int p(Y|X, \theta)p(\theta)d\theta, \tag{5}$$

where $p(\theta)$ is estimated using mean-field variational inference.

Regarding $L(M)$, COMIC and bQCD both abstain from computing it. In COMIC, $p(\theta)$ (the model) is computed from the data, under a Gaussian approximation, so there is no model to encode. For bQCD, they argue that both causal directions require encoding the same number of quantiles, so it is a constant that can be ignored.

LCUBE assumes parameter independence, adopting the classic choice (Grünwald, 2007)

$$L(M) = L(\theta) = (k/2)\log_2(N), \tag{6}$$

where θ is the k-dimensional vector of model parameters and N the number of samples.

Lastly, SLOPE computes the size of the model description by adding the combinatorial encoding representing the choice of models, both for the global and each of the local regressions, with a fixed *a priori* precision of each parameter within the model.

2.3. Analysis of $L(X)$

This subsection describes how each state-of-the-art MDL method computes $L(X)$, pointing out their restrictive assumptions. As described in Subsection 2.1, for MDL to approximate well the Kolmogorov complexity, the chosen family of models must be representative of the true distribution. As we argue next, bQCD, COMIC, LCUBE, and SLOPE do not do so when defining $L(X)$.

LCUBE assumes a uniform prior over a normalized distribution (both for X and Y), thus explicitly setting $L(X) = L(Y)$. SLOPE assumes a discretized uniform distribution over the minimum required resolution, thus ignoring any relevant information about the distribution. COMIC measures $L(X)$ by encoding X as the residuals of a Gaussian distribution with the mean estimated from the samples; as SLOPE and LCUBE, this ignores any information regarding the shape of the distribution. Finally, bQCD computes multiple estimates at its quantiles, thus assuming a necessarily discrete underlying distribution.

GPI (Stegle et al., 2010) also seeks to approximate Kolmogorov complexity, but using Bayesian theory instead of MDL. Both approaches balance the complexity of the cause and conditional distributions. Thus, an analogy can be made where the negative log-likelihood of the cause distribution in Bayesian theory corresponds to MDL's cause distribution encoding length (and similarly for the conditional distribution). For the latter, the conditional log-likelihood in GPI takes up a complex formula that we decided not to dive into. However, given this work's emphasis on $L(X)$, it is important to mention how GPI computes the log-likelihood of the observed cause distribution. First, it estimates a Gaussian mixture model from the data, and then measures the log-likelihood of the observed points according to the minimum message length principle for Gaussian mixtures defined

by Figueiredo and Jain (2002). Consequently, it measures the distance of the observed distribution to the best-fitted Gaussian mixture. Intuitively, it states that the more similar the observed distribution is to a Gaussian mixture, the smaller its complexity. Additionally, throughout this paper, even though GPI is not an MDL-based method, we will often refer to it as such, alongside the other four analysed (bQCD, COMIC, LCUBE, and SLOPE). This will be useful, since GPI shares MDL's rationale of measuring the complexity of both the cause and the causal mechanism. Furthermore, for GPI, $L(X)$ can be defined as $-\log p(X)$.

In all, none of these methods attempt to capture features of $P(X)$, or to approximate in a meaningful way the real distribution. Instead, they assume X follows a concrete distribution (e.g., Uniform, Gaussian, discrete Uniform, or mixture of Gaussians[1].), resulting in poor approximations of the Kolmogorov complexity of $P(X)$, when these narrow assumptions do not hold.

More generally, note that the use of MDL (or Bayesian theory) in the problem of causal discovery (see Equation (2)) is unorthodox, as the need to estimate and compare the complexity of the cause distribution is uncommon. For example, consider the use of MDL to perform model selection in a regression problem; *e.g.*, selecting the order when fitting a polynomial to a collection of (x, y) pairs, simply written as (X, Y). In this case, the MDL criterion would read

$$\widehat{M} = \arg\min_{M \in \mathcal{M}} L(M) + L(X) + L(Y|X, M) = \arg\min_{M \in \mathcal{M}} L(M) + L(Y|X, M), \tag{7}$$

where, here, M specifies only the regression model (*e.g.*, the polynomial order), thus $L(X)$ does not depend on M. The second equality, i.e., dropping $L(X)$, is justified by the fact that, in this case, it is a constant, since the description length of the independent variable does not depend on the regression model. In contrast, in the bivariate causal discovery problem, we are comparing two models in which X and Y play different roles; consequently, when using MDL to select between the two possible causal directions, the description lengths of the cause $L(X)$ (in the $X \to Y$ model) and $L(Y)$ (in the $Y \to X$ model) do not cancel each other out.

Although this difference between bivariate causal discovery and more standard applications of MDL partly explains why no method uses an informative estimate of $L(X)$, it still means that none are faithfully following the MDL framework to the causal discovery problem. By relying almost only on $L(Y|X, M)$, these methods are close to score-based methods that simply compute the causal direction that best fits a specific regression model (e.g., RECI, *regression error based causal inference*, proposed by Blöbaum et al. (2018)). The difference then lies in MDL-based approaches measuring goodness of fit using an MDL formalism, similarly to the regression problem defined in Equation (7). In turn, this deviates from the original motivation behind MDL for causal discovery: approximating Kolmogorov complexity in the two directions (as explained in Subsection 2.1).

Dropping the dependency on the model assumptions M for notational simplicity, MDL in causal discovery compares $L(X)+L(Y|X)$ with $L(Y)+L(X|Y)$, choosing the direction that corresponds to a smaller encoding. This can be written as computing a score,

$$s = (L(X) - L(Y)) + (L(Y|X) - (L(X|Y)), \tag{8}$$

and choosing $X \to Y$ if it is $s < 0$, and $Y \to X$ otherwise. Thus, MDL can be seen as the sum of two differences, one representing the difference in encoding length between the cause distribution

1. Although a Gaussian mixture is theoretically capable of approximating any probability distribution arbitrarily well given infinitely many observations, in practice, since each example contains few datapoints, the approximation only holds when the true underlying distribution is itself a mixture of Gaussians.

and another representing the difference in the encoding lengths of the two conditional distributions. In turn, each of these two differences can be seen as an individual score, each attempting to estimate the true causal direction, and MDL resulting in the combination of these two estimators.

Taking a step back, and noting that the use of MDL in causal discovery stems from the ICM principle, the lengths in the true causal direction will not encode overlapping information over the joint distribution. In contrast, the same cannot be said in the wrong causal direction. More relevant to this work, the length of the effect (Y) is expected to encode information over the conditional distribution of $X|Y$, while the opposite is not true. Therefore, even though MDL succeeds due to the complementary nature of these two estimators, each alone should be expected to perform significantly above average. Given that observation, the corresponding $(L(X) - L(Y))$ estimators of four out of the five analysed causal discovery methods were evaluated, using the *area under the receiver operating characteristic* (AUROC), on the Tübingen benchmark (Mooij et al., 2016) (LCUBE was excluded since it defines $L(X) = L(Y)$). Table 1 shows that, for all methods, $(L(X) - L(Y))$ by itself has significantly lower AUROC than the corresponding full methods. In turn, this further confirms the notion that current MDL methods for causal discovery fail to make accurate estimates over the encoding length of the cause distribution.

Table 1: AUROC (in %) on the Tübingen dataset, achieved by $(L(X) - L(Y))$ alone as a prediction score versus the full method, for bQCD, COMIC, GPI, and SLOPE.

	$L(X) - L(Y)$	full method
bQCD	54.7	71.1
COMIC	55.6	76.0
GPI	49.0	68.1
SLOPE	53.7	80.5

2.4. Contributions and Overview

In this work, we started by arguing (Subsection 2.3) that the five previously mentioned state-of-the-art MDL-based methods for the cause-effect problem with continuous variables (bQCD, COMIC, GPI, LCUBE, SLOPE) do not adequately define $L(X)$, thus not following the MDL recipe for causal discovery. In the next section, we propose a new approach based on rate-distortion theory to define $L(X)$ and couple it with a standard approach to compute $L(Y|X)$, yielding the new causal discovery method RDMDL. Subsequently, we overview the main assumptions underlying our approach and report experimental results on the Tübingen and on synthetic benchmarks, showing that RDMDL performs competitively with the best methods.

3. Proposed Approach: RDMDL

This section describes the proposed MDL-based approach for bivariate causal discovery. The key contribution is a new way to define the description length of the cause, $L(X)$, based on rate-distortion theory. We then couple it with a standard choice to compute $L(Y|X, M)$, leading to the proposed RDMDL. A detailed description of the algorithm is present in Appendix A. We begin by reviewing the concepts from rate-distortion theory that are needed to describe RDMDL.

3.1. The Rate-distortion Function and Information Dimension

The *rate-distortion* (R-D) function is central in the theory of lossy compression (Cover and Thomas, 2005; Geiger and Koch, 2019). For a random variable $X \in \mathbb{R}$ and using mean squared error (MSE) as the distortion measure, the R-D function is defined as

$$R(X, D) \triangleq \inf_{P_{\hat{X}|X}:\, E[(X-\hat{X})^2] \leq D} I(X; \hat{X}), \tag{9}$$

where $I(X; \hat{X}) = H(X) - H(X|\hat{X}) = H(\hat{X}) - H(\hat{X}|X)$ is the mutual information between X and a reconstruction/approximation $\hat{X}$, $H(X)$ and $H(X|\hat{X})$ denote Shannon differential entropy and conditional entropy, and D is an upper bound on the allowed distortion. Essentially, the R-D function is the minimum amount of information (in bits, if mutual information is computed with base-2 logarithms) that any $\hat{X}$ needs to have about X to satisfy the distortion upper bound.

The *rate-distortion dimension* (RDD) of a source is defined as

$$\dim_R(X) = 2 \lim_{D \downarrow 0} \frac{R(X, D)}{-\log D} \tag{10}$$

and measures how the number of bits required to encode a vector scales with the maximum allowed distortion, as that allowed distortion approaches zero (Rezagah et al., 2016). For example, any source with an absolutely continuous distribution has $\dim_R(X) = 1$, while any discrete source has $\dim_R(X) = 0$. Intuitively, for a discrete source, even with a countably infinite alphabet, each outcome lies in an isolated point of the space, so achieving small MSE does not require specifying any continuous degree of freedom. In contrast, a source taking values in a continuum requires resolving a continuous dimension, and describing it to vanishing distortion is reflected in its RDD.

Kawabata and Dembo (1994) proved that the RDD coincides with the so-called *information dimension* (ID, Rényi (1961)),

$$\dim_R(X) = \dim_I(X) = \lim_{\varepsilon \to 0} H_\varepsilon(X) / \log(1/\varepsilon), \tag{11}$$

where $H_\varepsilon(X)$ is the (discrete) entropy of a uniformly quantized version of X with resolution ε.

3.2. Setting $L(X)$

We propose to approximate $L(X)$ by $R(X, D)$, for a distortion level D that is representative of the relevant sampled information from the underlying probability distribution. To do so, two questions must be answered: 1) what does it mean for a distortion to be representative of the underlying probability distribution? 2) How does the theoretical lower bound provided by the R-D combines with the MDL approach to approximate the KC of the underlying distribution?

Starting with the second question, MDL approximates the KC by adopting a model, encoding the residuals, and measuring the description length of X as that of the residuals plus the model. Therefore, it approximates the theoretical lower bound of KC by measuring the shortest description length for a specific set of models. As mentioned in Subsection 2.1, this requires the following two assumptions: the samples are representative of the underlying probability distribution; and the set of models approximates well the real distribution. By using R-D theory to measure $L(X)$, the second assumption is no longer required, as we directly measure a non-parametric lower bound estimate of the encoding length of X.

The other major question is how to choose the distortion level so that $\hat{X}$ is representative of the underlying probability distribution. As is well known, $\lim_{D\to 0} R(X, D) = \infty$, since we are dealing with continuous variables. Conversely, under MSE distortion and assuming finite variance, $\text{var}(X) < \infty$, then $D \geq \text{var}(X) \Rightarrow R(X, D) = 0$, which corresponds to setting $\hat{X} = \mathbb{E}(X)$ (Cover and Thomas, 2005). Between these two extremes, there is a middle ground, where D is small enough such that $\hat{X}$ "accurately" represents X (and consequently $R(X, D)$ is a good approximation of $L(X)$), while not being so small such that it forces $\hat{X}$ to encode all the randomness inherent to the sampling process. This question is akin to that of optimal bin-size selection for histogram-based non-parametric density estimation (Scott, 1992). If the quantization bin width δ is too large (distortion is large), the histogram has high bias, ignoring genuine structural information about the distribution. Conversely, if δ (and D) is too small, the histogram has high variance, and we are no longer just encoding the density, but also irrelevant fluctuation of individual samples, thus effectively "wasting bits" to store noise rather than structure.

For histogram-based density estimation, there are several well-established rules for optimal bin sizes. We will consider four of these rules:

Freedman-Diaconis' rule: $\delta = 2\ \text{IQR}(X)\ N^{-1/3}$, where $\text{IQR}(X) = Q3 - Q1$ is the interquartile range of the samples in X, *i.e.*, the difference between the 75th percentile and the 25th percentile (Freedman and Diaconis, 1981).

Sturges' rule: $\delta = \text{range}(X)/(1 + \log_2(N))$, where $\text{range}(X)$ is the range of the samples in X (Sturges, 1926; Scott, 2009).

Scott's rule: $\delta \simeq 3.5\ \hat{\sigma}(X)\ N^{-1/3}$, where $\hat{\sigma}(X)$ is the sample standard deviation (Scott, 1979).

Rice University rule: $\delta = 0.5\ \text{range}(X)\ N^{-1/3}$ (Rubia, 2024).

Since the Freedman-Diaconis' rule is the only non-parametric one, it stands out as the most fitting to our approach, as thus, it will be considered as the default implementation of RDMDL. Howerver, unlike in histogram-based density estimation, here the goal is not to approximate the probability density function, but the samples themselves. Consequently, using the well-known high-resolution approximation, the MSE resulting for quantizing with bin size δ is simply $D = \delta^2/12$, which we will adopt, where δ is set using one of the four rules just mentioned.

Given a choice of distortion level D, we compute $L(X)$ from $R(X, D)$. To do so, we first estimate the ID, which is known to be equivalent to the RDD. Then, from the desired distortion and the RDD, we estimate the rate required to encode X. We assume that we are in the high-resolution (small ϵ and small D) asymptotic regime, thus the leading terms of both $R(X, D)$ and $H_\epsilon(X)$ are dominant. Under this assumption, Equations (10) and (11) can be transformed into, respectively,

$$H_\epsilon(X) = -\ \dim_I(X)\ \log(\epsilon)\ +\ o(\log(1/\epsilon)), \tag{12}$$

$$R(X, D) = -\ \frac{\dim_R(X)}{2}\ \log(D)\ +\ o(\log(1/D)), \tag{13}$$

where $o(\log(1/\epsilon))$ and $o(\log(1/D))$ are stricter upper bounds of any unknown additive terms.

To estimate the ID, we begin by considering a set of bin sizes $\{\epsilon_m\}_{m=1}^{M}$ (small enough to be in the high-resolution regime) and compute the entropy at each scale,

$$H_{\epsilon_m}(X) = -\sum_{j=1}^{N_m} p_j^{(\epsilon_m)} \log p_j^{(\epsilon_m)}, \tag{14}$$

where $p_j^{(\epsilon_m)}$ is the probability of X falling into the j-th bin of size ϵ_m, and N_m is the number of bins at that scale (e.g., if $X \in [0,\, 1]$, then $N_m = 1/\epsilon_m$). The information dimension is then obtained by performing a least-squares fit of $H_{\epsilon_m}(X)$ against $\log(1/\epsilon_m)$:

$$\dim_I(X) = \underset{d}{\operatorname{argmin}} \sum_{m=1}^{M} \left(H_{\epsilon_m}(X) - d \, \log \frac{1}{\epsilon_m} \right)^2. \tag{15}$$

Afterwards, since $\dim_R = \dim_I$ (Kawabata and Dembo, 1994), it is possible to compute $R(X, D)$ for the chosen D directly from Equation (13). Finally, $L(X)$ is set to the length needed to encode the N samples in X, thus

$$L(X) = N \cdot R(X, D) = N \, \frac{\dim_I(X)}{2} \log\left(\frac{1}{D}\right). \tag{16}$$

3.3. Setting $L(Y|X)$ and $L(M)$

The central goal of this paper is to highlight and motivate the focus on $L(X)$ when using MDL for causal discovery. Consequently, a very standard approach is used for $L(Y|X, M)$ and $L(M)$. Following the standard assumptions in MDL, and specifically those used in SLOPE and LCUBE, we used least squares to fit the model, and Gaussian encodings to represent the residuals, ensuring coherence between both operations. This choice yields

$$L(Y|X, M) = \frac{N}{2} \log_2\!\left(2\pi \, \hat{\sigma}^2\right) + \frac{N}{2 \log 2}, \tag{17}$$

where $\hat{\sigma}^2$ is the empirical residual variance. This term corresponds to the negative log-likelihood (in bits) under a Gaussian model with variance $\hat{\sigma}^2$.

As model, we use a single global regressor from the following three parametric families:

- *Polynomial models*, with degree up to 5.
- *Reciprocal models:* $\hat{y} = a/x^p + b$, with $p \in \{1, 2\}$.
- *Exponential and logarithmic models:* $\hat{y} = a \, e^x + b$ and $\hat{y} = a \, \log(x) + b$.

The model description length $L(M)$ is set to the standard choice in Equation (6), where k is the number of parameters. This choice minimizes the total conditional MDL ($L(Y|X)$), if the model parameters are independent. Finally, for each family of models, all admissible parameterizations are fitted, and their total code length $L(Y|X, M) + L(M)$ is evaluated, and

$$L(Y|X) = \min_{M} L(Y|X, M) + L(M). \tag{18}$$

3.4. Overview of RDMDL's assumptions

All bivariate causal discovery methods share the same goal. What differentiates them is the underlying assumptions of each method. For greater clarity and to facilitate future research, we consider that our method is supported by the following non-trivial assumptions/claims:

1. There is an interval of distortions D such that $R(X, D)$ is a good approximation of $L(X)/N$ and histogram-based density estimation rules can compute distortions D that fall within this interval. This claim is supported by the good performance of all RDMDL's variations.

2. The samples are representative of the underlying probability distributions. It corresponds to the first assumption in Subsection 2.1, and it is necessary to ensure the equivalence between the description length of the samples and that of the underlying joint distribution.

3. The family of models used to encode $L(Y|X)$ is representative enough to capture the underlying causal mechanisms. This is the second criterion in Subsection 2.1.

4. Results

In this section, we start by analyzing how RDMDL compares with other bivariate causal discovery methods and, afterwards, we will study how our approach to computing $L(X)$ contributes to the performance of RDMDL. As described in Section 3.2, we consider four different rules to estimate the optimal bin size (and with it the desired distortion). Consequently, we report the results obtained by each implementation, named RDMDL-FD (using the Freedman-Diaconis rule), RDMDL-R (Rice University rule), RDMDL-Sc (Scott's rule), and RDMDL-St (Sturges rule). Additionally, remember that the FD rule is considered the default implementation of RDMDL, given its non-parametric approach.

Table 2 shows the results on the Tübingen dataset of Mooij et al. (2016), the only real-world bivariate causal discovery benchmark. In turn, method performance was measured using three evaluation metrics: AUROC, accuracy, and AUDRC (area under the decision rate curve (Immer et al., 2023)). The table includes results for a comprehensive set of other open-source bivariate causal discovery methods. In fact, to the best of our knowledge, all causal discovery methods with open-source and retrievable implementations, in Python or R, are included: ANM (Peters et al., 2010), bQCD, CAM (Bühlmann et al., 2014), CDCI (Duong and Nguyen, 2022), CDS (Fonollosa, 2016), CGNN (Goudet et al., 2018), FOM (Cai et al., 2020), GPI, HECI (Xu et al., 2022), IGCI (Mian et al., 2023), LCUBE, LOCI (Immer et al., 2023), NNCL (Wang and Zhou, 2021), RECI (Blöbaum et al., 2018), ROCHE (Tran et al., 2024), SLOPE, and SLOPPY (Marx and Vreeken, 2019)[2]. For methods without publicly available implementations, the reported results are present in Appendix B.1. The only method with an open source implementation that we excluded is GPLVM (Dhir et al., 2023), because it involves training on the same dataset on which it is evaluated. Therefore, it is not a pure causal discovery method, as the model is able to learn the causal relations present in the dataset under evaluation.

The results in Table 2 show that all versions of RDMDL outperform most methods on the Tübingen dataset across all metrics. The only exceptions are SLOPE, SLOPPY, and RECI, which are still outperformed by RDMDL-FD in terms of AUROC and AUDRC. Moreover, comparing to other results reported in the literature without publicly available implementations, only GRCI, COMIC, and GPLVM outperform RDMDL-R, RDMDL-Sc, and RDMDL-St (while scoring worse than RDMDL-FD). However, similarly to GPLVM, both COMIC and GRCI also require training on the benchmark dataset, and so, their results should be taken with a grain of salt.

Additionally, the Tübingen dataset is known for having more discrete causes than effects. Controlling for such, and scoring only the examples whose number of unique values per variable is at least a third of the total number of points, RDMDL-FD still obtains an AUROC of 84.5% and the

2. For bQCD, CAM, CDCI, FOM, GPI, HECI, LCUBE, LOCI, NNCL, ROCHE, SLOPE, and SLOPPY, the original implementations provided by the authors were used. For ANM, CDS, CGNN, IGCI, and RECI, the code from the CausalDiscoveryToolbox of Kalainathan and Goudet (2019) was used. Finally, since the original implementation of CGNN was too slow, the hyperparameters were adapted to ensure no example took longer than 30 seconds to score.

$(L(X) - L(Y))$ estimator alone an AUROC of 73%. This accounts for a total weight of 39% of the original dataset, without a significant dip in RDMDL's performance. Therefore, the proposed measure of $L(X)$ captures a notion of complexity of X that goes beyond a mere notion of how "discrete" it is. Instead, it appears to follow the initial intuition of measuring complexity by assessing how rugged the distribution becomes when using smaller and smaller bins (see subsection 3.2).

Table 2: Performance of the 4 RDMDL versions and 17 other methods on the Tübingen benchmarks. The best results in each column are in bold; the second best is underlined.

Method	AUROC (%)	Accuracy (%)	AUDRC (%)
ANM	61.9	60.4	62.9
bQCD	71.1	69.6	70.1
CAM	46.6	52.3	43.1
CDCI	60.7	61.5	55.0
CDS	59.0	60.5	57.8
CGNN	61.6	61.5	69.6
FOM	45.1	45.5	40.9
GPI	68.1	68.7	67.0
HECI	75.3	70.5	79.2
IGCI	69.8	60.9	73.3
LCUBE	58.0	58.9	70.1
LOCI	57.4	61.5	49.5
NNCL	65.3	55.2	63.4
RECI	76.4	70.5	75.7
ROCHE	56.7	53.0	53.7
SLOPE	80.5	**73.3**	86.4
SLOPPY	79.3	72.6	85.3
RDMDL-FD (ours)	**82.0**	71.9	**86.6**
RDMDL-R (ours)	74.7	68.0	79.1
RDMDL-Sc (ours)	76.3	67.4	81.7
RDMDL-St (ours)	75.3	68.0	80.1

Table 3 shows the results of RDMDL, bQCD, GPI, LCUBE, and SLOPE on three synthetic benchmarks: CE (Guyon et al., 2019), ANLSMN (Tagasovska et al., 2020), and SIM (Mooij et al., 2016). The performance of all other implemented methods is in Appendix B.2. It is clear that all versions of RDMDL perform significantly above average in most synthetic datasets, confirming it is a strong method. However, it is also notable that RDMDL does not achieve extremely high performance in these datasets, unlike bQCD or LCUBE. Usually, causal discovery methods are designed with certain assumptions (*e.g.*, additive noise or uniform distributions), which can more easily be matched with the generation process of synthetic datasets. In contrast, RDMDL is more agnostic; while it achieves very good results on the real-world benchmark, it does not seem to take as much advantage of explicit causal assumptions present in synthetic datasets.

Table 3: Results of the four versions of RDMDL in three collections of synthetic datasets: CE, ANLSMN, and SIM, evaluated using AUROC and accuracy (in parentheses in each cell).The best results for each dataset are in bold and the second-best are underlined.

Method	CE-Cha	CE-Gauss	CE-Multi	CE-Net	AN	AN-s	LS	LS-s	MN-U	SIM	SIM-c	SIM-G	SIM-ln
RDMDL-FD	62.7 (60.7)	64.7 (58.7)	**97.0** (**89.3**)	75.2 (68.3)	51.2 (49.0)	43.3 (45.0)	49.0 (47.0)	13.4 (17.0)	0.9 (6.0)	53.6 (47.0)	59.6 (57.0)	59.4 (55.0)	87.2 (77.0)
RDMDL-R	64.2 (60.3)	64.5 (59.0)	96.5 (88.3)	81.1 (72.0)	84.5 (74.0)	52.3 (51.0)	77.0 (68.0)	6.9 (13.0)	1.3 (7.0)	61.9 (56.0)	66.9 (65.0)	66.2 (57.0)	91.4 (79.0)
RDMDL-Sc	63.7 (59.3)	69.4 (62.7)	96.9 (89.0)	77.5 (69.3)	58.6 (52.0)	34.7 (34.0)	53.5 (51.0)	4.8 (12.0)	0.8 (4.0)	58.1 (53.0)	63.9 (62.0)	60.1 (54.0)	88.6 (78.0)
RDMDL-St	64.0 (60.3)	65.1 (59.0)	96.5 (88.3)	81.7 (72.3)	86.8 (76.0)	53.7 (51.0)	78.5 (69.0)	5.8 (13.0)	1.4 (7.0)	62.8 (56.0)	67.6 (66.0)	66.4 (58.0)	91.6 (79.0)
bQCD	56.3 (55.0)	48.1 (54.3)	53.2 (49.0)	**90.7** (81.3)	**100.0** (**100.0**)	90.5 (83.0)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	69.9 (68.0)	80.8 (77.0)	71.7 (67.0)	91.9 (**87.0**)
GPI	**64.9** (**62.4**)	**89.1** (**83.2**)	72.2 (71.5)	88.8 (81.3)	79.2 (85.5)	62.4 (62.5)	72.8 (83.6)	73.8 (70.9)	85.7 (80.7)	**88.3** (**83.7**)	**92.4** (**85.6**)	**90.1** (**80.4**)	92.8 (84.5)
LCUBE	55.7 (52.5)	31.8 (31.3)	43.5 (33.1)	87.8 (**82.1**)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	71.4 (62.3)	78.1 (66.7)	88.7 (76.8)	**95.4** (86.0)
SLOPE	59.3 (57.0)	73.3 (67.3)	96.9 (88.7)	67.1 (62.3)	9.2 (18.0)	23.4 (28.0)	12.6 (21.0)	11.0 (17.0)	1.1 (7.0)	47.9 (45.0)	57.2 (54.0)	45.1 (46.0)	44.5 (47.0)

Additionally, it is also important to note that the four RDMDL versions achieve comparable performance across both the Tübingen and most synthetic benchmarks, suggesting its performance is not overly sensitive to the choice of distortion.

Figure 1 displays the AUDRC for the four versions of RDMDL, compared to other methods on the Tübingen benchmark. In it, it is clear that RDMDL performs consistently above average across all decision rates. This is an even stronger result than the values of AUROC, accuracy, and AUDRC shown in Table 2, showing that, regardless of the decision rate, RDMDL is a leading method for causal discovery.

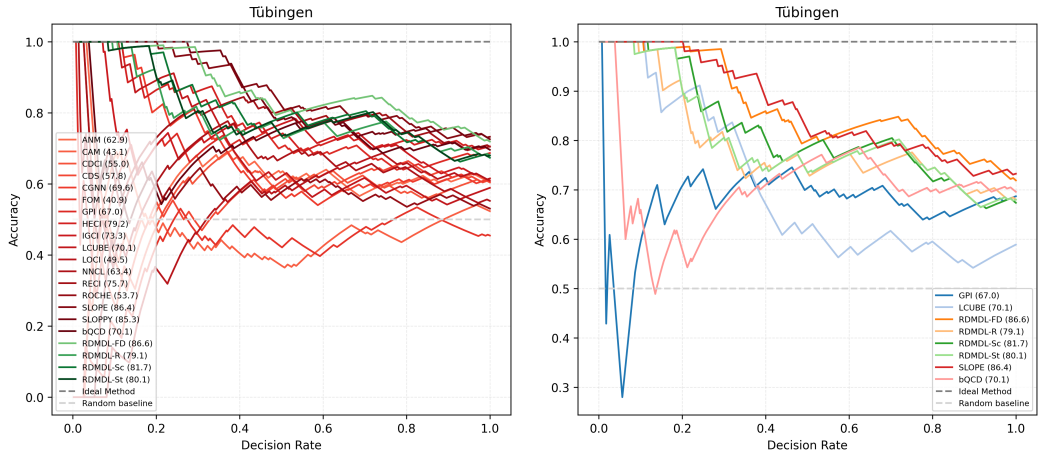

Figure 1: AUDRC curves in the Tübingen benchmark. Left: RDMDL (green) compared to all other implemented methods (shades of red); right: analysed MDL causal discovery methods.

Furthermore, given that one of key contributions of this work is to propose an MDL method for bivariate causal discovery that performs an adequate estimate of $L(X)$, we now analyze the relevance of $L(X)$ in the overall performance of RDMDL. Table 4 studies how informative $L(X)$ and $L(Y|X)$ are individually, and how they interact to yield the result on the Tübingen benchmark. Understanding this behaviour is crucial, as it reflects the principle that MDL seeks to exploit (see Equation (8)). Each entry in both tables contains two values: on the left, the percentage of weighted guesses where the difference $(L(X) - L(Y))$ dominates the decision (higher absolute value), while on the right are those where $(L(Y|X) - L(X|Y))$ dominates.

The first clear conclusion is that both differences are individually good predictors. The second one is that there are a few differences between the four versions [3] of RDMDL. The third, and maybe most crucial, is that both differences are responsible for a significant part of the decisions, thus both contribute actively and positively to the predictions.

Lastly, although each individual predictor performs well and contributes equally to the output, they do not appear to reinforce one another. Recall that a central principle of MDL is that a joint distribution producing a negative prediction in one estimator should correspondingly generate an

3. Note that the computation using Sturges and Rice methods yields the same results. This is not, however, necessarily always true; their difference is not strong enough to meaningfully cause any difference.

even stronger positive prediction in the other. However, this effect is not observed in practice. This would manifest as the correct predictor outnumbering the wrong one in the off-diagonal entries.

Table 4: Confusion matrix: in each cell, on the left, the result when taken by $(L(X) - L(Y))$ and on the right when $(L(Y|X) - L(X|Y))$ dominates. Green indicates a correct decision and red an incorrect decision. Top-left table: RDMDL-FD; top-right table: RDMDL-R; bottom-left table: RDMDL-Sc; bottom-right table: RDMDL-St.

Predictors-FD		$L(Y\|X)$		**Marginals**
		correct	**incorrect**	
$L(X)$	**correct**	38.2 / 16.2	15.8 / 4.8	70.2 / 4.8
	incorrect	8.1 / 1.7	8.3 / 6.9	1.7 / 23.3
Marginals		56.1 / 8.1	15.8 / 20.0	

Predictors-R		$L(Y\|X)$		**Marginals**
		correct	**incorrect**	
$L(X)$	**correct**	23.3 / 21.5	11.3 / 7.6	56.1 / 7.6
	incorrect	7.4 / 12.0	4.8 / 12.2	12.0 / 24.4
Marginals		56.8 / 7.4	11.3 / 24.6	

Predictors-Sc		$L(Y\|X)$		**Marginals**
		correct	**incorrect**	
$L(X)$	**correct**	27.1 / 26.4	10.6 / 4.8	64.1 / 4.8
	incorrect	7.4 / 3.2	10.0 / 10.5	7.4 / 27.9
Marginals		56.7 / 7.4	10.6 / 25.3	

Predictors-St		$L(Y\|X)$		**Marginals**
		correct	**incorrect**	
$L(X)$	**correct**	23.3 / 21.5	11.3 / 7.6	56.1 / 7.6
	incorrect	7.4 / 12.0	4.8 / 12.2	12.0 / 24.4
Marginals		56.8 / 7.4	11.3 / 24.6	

5. Conclusions and Future Work

This work revisited the fundamental rationale behind MDL for bivariate causal discovery. A detailed analysis of existing methods showed that the estimation of $L(X)$, the description length of the cause variable, has been largely overlooked, with decisions often effectively relying only on the complexity of the causal mechanism. Motivated by this gap, we proposed a novel approach to estimate $L(X)$ via rate-distortion theory, combined with a traditional MDL-based approach to computing $L(Y|X)$. The computation of $L(X)$ is based on an informative choice of distortion, stemming from histogram-based density estimation rules, and an approximation of the rate-distortion dimension under the asymptotic assumptions. This yields our new method, rate-distortion MDL (RDMDL), which achieves competitive performance on the Tübingen and most synthetic datasets.

A major avenue for future work is to extend the rate-distortion estimation, or a similar theoretically grounded lower bound on Kolmogorov complexity, to measure $L(Y|X)$. This would not only provide a more coherent integration with the estimation of $L(X)$, but also hopefully better performance on real data. Additionally, the deduction of specific scenarios where RDMDL is guaranteed to be correct would contribute to reinforce the validity of the current approach.

Acknowledgments

Work partially funded by *Fundação para a Ciência e a Tecnologia* (Portugal), project doi.org/10.54499/UID/50008/2025. We also acknowledge support from *Feedzai Lda* (Portugal).

References

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 900–909, 2018.

Kailash Budhathoki and Jilles Vreeken. MDL for causal inference on discrete data. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 751–756, 2017.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. 2014.

Ruichu Cai, Jincheng Ye, Jie Qiao, Huiyuan Fu, and Zhifeng Hao. FOM: Fourth-order moment based causal direction identification on the heteroscedastic data. *Neural Networks*, 124:193–201, 2020.

Zhitang Chen, Kun Zhang, Laiwan Chan, and Bernhard Schölkopf. Causal discovery via reproducing kernel Hilbert space embeddings. *Neural computation*, 26(7):1484–1517, 2014.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2005.

Anish Dhir, Samuel Power, and Mark van der Wilk. Bivariate causal discovery using Bayesian model selection. *arXiv preprint arXiv:2306.02931*, 2023.

Bao Duong and Thin Nguyen. Bivariate causal discovery via conditional divergence. In *Conference on Causal Learning and Reasoning*, pages 236–252. PMLR, 2022.

Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3:81–91, 2017.

Mario Figueiredo and Catarina Oliveira. Distinguishing cause from effect on categorical data: The uniform channel model. In *Conference on Causal Learning and Reasoning*, pages 122–141. PMLR, 2023.

Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

José AR Fonollosa. Conditional distribution variability measures for causality detection. *arXiv preprint arXiv:1601.06680*, 2016.

David Freedman and Persi Diaconis. On the histogram as a density estimato: l_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete (Probability Theory and Related Fields)*, 57:433–476, 1981.

Bernhard C Geiger and Tobias Koch. On the information dimension of stochastic processes. *IEEE Transactions on Information Theory*, 65(10):6496–6518, 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning*, pages 39–80, 2018.

Peter Grünwald. *Information and Complexity in Statistical Modeling*. MIT Press, 2007.

Peter D Grünwald. *The Minimum Description Length Principle*. MIT press, 2007.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Siyuan Guo, Chi Zhang, Karthika Mohan, Ferenc Huszár, and Bernhard Schölkopf. Do Finetti: On causal effects for exchangeable data. In *Advances in Neural Information Processing Systems*, 2024.

Isabelle Guyon, Alexander Statnikov, and Berna Bakir Batu. *Cause Effect Pairs in Machine Learning*. Springer, 2019.

Uzma Hasan, Emam Hossain, and Md. Osman Gani. A survey on causal discovery methods for i.i.d. and time series data. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=YdMrdhGx9y.

Daniel Hernandez-Lobato, Pablo Morales-Mombiela, David Lopez-Paz, and Alberto Suarez. Nonlinear causal inference using gaussianity measures. *Journal of Machine Learning Research*, 17 (28):1–39, 2016.

Katerina Hlavackova-Schindler and Suzana Marsela. Identifying causal direction via dense functional classes. *arXiv preprint arXiv:2509.00538*, 2025.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332. PMLR, 2023.

Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 2010.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

Tsutomu Kawabata and Amir Dembo. The rate-distortion dimension of sets and measures. *IEEE transactions on information theory*, 40(5):1564–1572, 1994.

Scott T. Leatherdale. Natural experiment methodology for research: a review of how different methods can support real-world research. *International Journal of Social Research Methodology*, 22:19–35, 2019.

Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 3rd edition, 2008.

Alexander Marx and Jilles Vreeken. Telling cause from effect using MDL-based local and global regression. In *IEEE International Conference on Data Mining (ICDM)*, pages 307–316, 2017.

Alexander Marx and Jilles Vreeken. Identifiability of cause and effect using regularized regression. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 852–861, 2019.

Alexander Marx and Jilles Vreeken. Formally justifying MDL-based inference of cause and effect. *arXiv preprint arXiv:2105.01902*, 2021.

Osman Mian, Michael Kamp, and Jilles Vreeken. Information-theoretic causal discovery and intervention detection over multiple environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9171–9179, 2023.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

David Papineau. The statistical nature of causation. *The Monist*, 105:247–275, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 597–604. JMLR Workshop and Conference Proceedings, 2010.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

Farideh Rezagah, Shirin Jalali, Elza Erkip, and Vincent Poor. Rate-distortion dimension of stochastic processes. In *IEEE International Symposium on Information Theory*, pages 2079–2083, 2016.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Jorma Rissanen. *The minimum description length principle*. Springer, 2007.

José Moral De La Rubia. Rice University rule to determine the number of bins. *Open Journal of Statistics*, 14:119–149, 2024.

David Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

David Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

David Scott. Sturges' rule. *WIREs Computational Statistics*, 1:303–306, 2009.

Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Artificial intelligence and statistics*, pages 847–855. PMLR, 2015.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2001.

Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *Advances in neural information processing systems*, 23, 2010.

Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 72:102099, 2023.

Herbert Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21: 65–66, 1926.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, pages 9311–9323. PMLR, 2020.

Quang-Duy Tran, Bao Duong, Phuoc Nguyen, and Thin Nguyen. Robust estimation of causal heteroscedastic noise models. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 788–796. SIAM, 2024.

Quang-Duy Tran, Bao Duong, Phuoc Nguyen, and Thin Nguyen. Identifying causal direction via variational Bayesian compression. In *International Conference on Machine Learning*, 2025.

Bingling Wang and Qing Zhou. Causal network learning with non-invertible functional relationships. *Computational Statistics & Data Analysis*, 156:107141, 2021.

Sascha Xu, Osman A Mian, Alexander Marx, and Jilles Vreeken. Inferring cause and effect in the presence of heteroscedastic noise. In *International Conference on Machine Learning*, pages 24615–24630. PMLR, 2022.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment*, pages 157–164. PMLR, 2010.

Appendix A. Full Algorithm

This appendix provides the complete description of our implementation. Algorithm 1 outlines the RDMDL procedure. The initial step involves scaling both the input data X and Y, using either min-max scaling or normalization (subtracting the mean and dividing by the standard deviation). Subsequently, the distortion parameter D is selected according to the criteria detailed in Subsection 3.2. The algorithm then computes the respective Minimum Description Length (MDL) encoding lengths for both directions: $L(X)$, $L(Y \mid X)$, $L(Y)$, and $L(X \mid Y)$, utilizing Algorithms 2 and 5. Finally, it calculates the total encoding length for both directions, and outputs their difference, which is then divided by N. This normalization serves to express the score as the average difference in the number of bits per observation pair, providing a better interpretation and facilitating the use of rank-based metrics (such as AUROC and AUDRC) for assessing the confidence level of RDMDL. Consequently, the output score represents the average difference in the number of bits per pair required to encode the data under the causal assumption in each direction.

Algorithm 1 RDMDL

Input: X, Y

Output: $S_{X,Y}$

1 Scale X and Y using min-max scaling or normalization.

2 Choose the distortion parameter D.

3 Compute the total encoding lengths for the models in each direction using Algorithms 2 and 5:

$$L_{X \to Y} = L_X(X) + L_{Y|X}(X, Y)$$

$$L_{Y \to X} = L_X(Y) + L_{Y|X}(Y, X).$$

4 Compute the final score as the average encoding length per data point:

$$S_{X,Y} \leftarrow \frac{L_{Y \to X} - L_{X \to Y}}{N_X}.$$

A.1. $L(X)$ Algorithm

This subsection details the computation of the encoding length $L(X)$. In general, the procedure first calculates the information dimension $\dim_X$, and then uses it to compute the total encoding length based on Equation (16). The function used to compute the information dimension is described in Algorithm 3, while the entropy calculation via binning is detailed in Algorithm 4.

Algorithm 2 L_X (from RDMDL)

Input: X, D

Output: L_X

Compute the Information Dimension using Algorithm 3:

$$\dim_X \leftarrow \dim(X).$$

Compute the total encoding length for the data:

$$L_X \leftarrow N_X \frac{\dim_X}{2} \log\left(\frac{1}{D}\right).$$

In Algorithm 2, D is the chosen distortion parameter and N_X is the number of observations in X.

Algorithm 3 Information Dimension $\dim(X)$ (from RDMDL)

Input: X, $H_X(\cdot)$, $\epsilon_{\min}$, $\epsilon_{\max}$, and N_ϵ

Output: $\dim_X$

E = logspace$(\epsilon_{\min}, \epsilon_{\max}, N_\epsilon)$.

$H_i \leftarrow H_X(\epsilon_i), \forall\, \epsilon_i \in E$. ▷ In accordance with Algorithm 4.

Fit a linear model to the points $(\log \epsilon_i, H_i)$: ▷ Using Numpy's polyfit.

$$H_i \approx a \log \epsilon_i + b$$

with least squares regression.

Compute the information dimension:

$$\dim_X \leftarrow -a.$$

In Algorithm 3, the function logspace stands for the standard Numpy function.

Algorithm 4 Entropy H (from RDMDL)

Input: X, ϵ_i

Output: H_{X,ϵ_i}

$b \leftarrow$ linspace$(\min(X), \max(X), \frac{1}{\epsilon_i} + 1)$ ▷ Define bin edges

$c \leftarrow$ histogram(X, b)

$p \leftarrow \frac{c_{c>0}}{N_X}$ ▷ Compute non-zero probabilities

$H_{X,\epsilon_i} \leftarrow -\sum p_i \cdot \log(p_i)$

In Algorithm 4, both linspace and histogram refer to the standard Numpy functions.

A.2. $L(Y \mid X)$ Algorithm

The general procedure for computing the conditional encoding length $L(Y|X)$ involves fitting three different sets of functional models (polynomial, inversely proportional, and exponential/logarithmic),

and then selecting the model that yields the shortest description length. For each model set, the process simultaneously determines the best least squares fit, by minimizing the total encoding length for both the residuals and the model parameters.

Algorithm 5 $L_{Y|X}$ (from RDMDL)

Input: X, Y
Output: $L_{Y|X}$

15 $L_p = \text{fit}_{\text{polynomial}}(X, Y)$ ▷ Check Algorithm 6
16 $L_i = \text{fit}_{\text{inversely}}(X, Y)$
17 $L_e = \text{fit}_{\text{explog}}(X, Y)$
18 Return the encoding length of the best-fitting model:

$$L_{Y|X} \leftarrow \min\left(L_p, L_i, L_e\right)$$

In Algorithm 5, L_p, L_i, and L_e stand for the minimum obtained length when encoding the causal mechanism with polynomial, inversely proportional, and exponential/logarithmic functions respectively. The three different lengths are computed in accordance with Algorithm 6, by passing a different sets of possible families of functions in the argument (F).

Algorithm 6 fit (from RDMDL)

Input: $X, Y, F = \{f_1, \ldots, f_m\}$
Output: L

19 For each function f in the set of functions F, compute:

$$r_f, \theta_f \leftarrow \text{LS}_f(X, Y),$$

where LS_f denotes the best least squares fit for the function f, yielding the residuals r_f and model parameters θ_f.

20 Then, for each function, compute the encoding length of the residuals using a Gaussian distribution assumption:

$$L(r_f) \leftarrow \frac{N_r}{2} \log\left(2\pi \frac{\sum r_f^2}{N_r}\right) + \frac{N_r}{2\,\log(2)}.$$

21 Compute the encoding length for the model parameters:

$$L(\theta_f) \leftarrow \frac{N_{\theta_f}}{2} \log(N_r).$$

22 Lastly, the overall length L is the minimum of the sum of the residual and model parameter lengths obtained by all tested functions:

$$L \leftarrow \min_f (L(r_f) + L(\theta_f)).$$

In Algorithm 6, N_r is the length of the residual vector ($N_r = N_X = N_Y$) and N_{θ_f} is the number of parameters in model f.

Appendix B. Additional results

B.1. Unimplemented results

In this Appendix, the results of 9 different unimplemented causal discovery methods are reported: COMIC (Tran et al., 2025), EMD (Chen et al., 2014), GPLVM (Dhir et al., 2023), GRCI (Strobl and Lasko, 2023), LiNGAM (Shimizu et al., 2006), PNL (Zhang and Hyvärinen, 2010), CURE (Sgouritsa et al., 2015), and GRAN (Hernandez-Lobato et al., 2016). These methods were not implemented since either 1) no publicly available implementation was found or 2) the method required training. Nevertheless, the reported results of each method (alongside its sources) are reported in Tables 5, 6, and 7. There are results reported for the real-world Tübingen dataset (Mooij et al., 2016) and the synthetic datasets of Guyon et al. (2019) (CE), Mooij et al. (2016) (SIM), and Tagasovska et al. (2020) (ANLSMN).

Table 5: Accuracy (in %) obtained by the different unimplemented methods for all evaluated datasets.

Method	Source	AN	AN-s	CE-Cha	CE-Multi	CE-Net	LS	LS-s	MN-U	Multi	Net	SIM	SIM-G	SIM-c	SIM-ln	Tübingen
COMIC	COMIC	100	100	43	79	78	100	100	100			57	78	54	89	67
EMD	bQCD	36	33				60	42	83			45	58	40	52	68
GPLVM	GPLVM, COMIC	100	100				100	100	100			83	92	79	90	
GRCI	LOCI	100	94	70			98	87	88	77	85	77	70	77	77	82
LiNGAM	bQCD	0	4				7	3	0			42	25	53	31	42
PNL	bQCD	96	63				91	44	66			70	64	65	61	73
CURE	bQCD															61
GRAN	bQCD	5	6				11	20	50			48	37	44	43	50

Table 6: AUROC (in %) obtained by the different unimplemented methods for all evaluated datasets.

Method	Source	AN	AN-s	CE-Cha	CE-Gauss	CE-Multi	CE-Net	LS	LS-s	MN-U	SIM	SIM-G	SIM-c	SIM-ln	Tübingen
COMIC	COMIC	100	100	47		91	84	100	100	100	58	85	59	95	75
EMD	bQCD														43
GPLVM	GPLVM, COMIC	100	100	82	89	98	99	100	100	100	92	99	91	97	78
LiNGAM	bQCD														30
PNL	bQCD														70
CURE	bQCD														64
GRAN	bQCD														47

Table 7: AUDRC (in %) obtained by the different unimplemented methods for all evaluated datasets.

Method	Source	AN	AN-s	CE-Cha	LS	LS-s	MN-U	Multi	Net	SIM	SIM-G	SIM-c	SIM-ln	Tübingen
GRCI	LOCI	100	100	71	100	95	97	74	96	90	88	92	92	73

B.2. Synthetic results

In this Appendix, the evaluation of the seventeen implemented bivariate causal discovery methods, alongside the four implementations of RDMDL, is reported for the following three sets of synthetic cause-effect benchmarks: the CE datasets of Guyon et al. (2019), the SIM datasets of Mooij et al. (2016), and the ANLSMN datasets of Tagasovska et al. (2020).

Table 8: Performance of the sixteen implemented causal discovery methods, alongside the four versions of RDMDL, in three sets of synthetic benchmarks, evaluated using AUROC (accuracy). The value in bold corresponds to the best result in each column, and those underlined to the second best.

Method	CE-Cha	CE-Gauss	CE-Multi	CE-Net	AN	AN-s	LS	LS-s	MN-U	SIM	SIM-c	SIM-G	SIM-ln
RDMDL-FD	62.7 (60.7)	64.7 (58.7)	**97.0** (89.3)	75.2 (68.3)	51.2 (49.0)	43.3 (45.0)	49.0 (47.0)	13.4 (17.0)	0.9 (6.0)	53.6 (47.0)	59.6 (57.0)	59.4 (55.0)	87.2 (77.0)
RDMDL-R	64.2 (60.3)	64.5 (59.0)	96.5 (88.3)	81.1 (72.0)	84.5 (74.0)	52.3 (51.0)	77.0 (68.0)	6.9 (13.0)	1.3 (7.0)	61.9 (56.0)	66.9 (65.0)	66.2 (57.0)	91.4 (79.0)
RDMDL-Sc	63.7 (59.3)	69.4 (62.7)	96.9 (89.0)	77.5 (69.3)	58.6 (52.0)	34.7 (34.0)	53.5 (51.0)	4.8 (12.0)	0.8 (4.0)	58.1 (53.0)	63.9 (62.0)	60.1 (54.0)	88.6 (78.0)
RDMDL-St	64.0 (60.3)	65.1 (59.0)	96.5 (88.3)	81.7 (72.3)	86.8 (76.0)	53.7 (51.0)	78.5 (69.0)	5.8 (13.0)	1.4 (7.0)	62.8 (56.0)	67.6 (66.0)	66.4 (58.0)	91.6 (79.0)
ANM	34.3 (40.0)	43.3 (45.3)	48.7 (51.0)	56.3 (59.0)	52.8 (51.2)	52.9 (55.8)	59.4 (57.8)	52.3 (57.8)	63.5 (62.2)	49.1 (46.3)	47.8 (43.6)	27.5 (37.2)	60.7 (56.2)
bQCD	56.3 (55.0)	48.1 (54.3)	53.2 (49.0)	90.7 (81.3)	**100.0** (**100.0**)	90.5 (83.0)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	69.9 (68.0)	80.8 (77.0)	71.7 (67.0)	91.9 (**87.0**)
CAM	58.6 (53.3)	30.0 (25.0)	38.3 (35.3)	81.6 (78.3)	**100.0**(**100.0**)	**100.0**(**100.0**)	99.9 (98.0)	36.9 (52.0)	81.2 (86.0)	62.3 (56.0)	64.3 (62.0)	88.2 (82.0)	86.0 (**87.0**)
CDCI	**72.2** (66.7)	**91.7** (83.0)	95.7 (90.0)	**91.8** (80.7)	**100.0**(**100.0**)	99.1 (95.0)	**100.0**(**100.0**)	98.9 (95.0)	99.5 (95.0)	87.2 (81.0)	84.1 (72.0)	81.2 (73.0)	85.0 (76.0)
CDS	69.2 (**71.0**)	90.6 (**84.0**)	41.2 (43.7)	88.6 (78.3)	**100.0**(99.0)	99.9 (99.0)	84.7 (76.0)	1.3 (5.0)	71.7 (70.0)	82.1 (71.0)	84.7 (76.0)	79.7 (73.0)	78.5 (65.0)
CGNN	55.5 (56.0)	37.4 (42.3)	80.1 (68.0)	60.0 (56.3)	77.9 (67.0)	89.8 (81.0)	78.8 (70.0)	86.6 (75.0)	2.7 (8.0)	42.0 (42.0)	51.3 (53.0)	84.9 (73.0)	77.7 (68.0)
FOM	42.0 (44.3)	56.1 (53.0)	10.4 (20.0)	27.0 (35.0)	94.9 (86.0)	**100.0**(**100.0**)	92.9 (93.0)	99.3 (99.0)	**100.0**(**100.0**)	51.1 (52.0)	46.2 (45.0)	25.1 (33.0)	36.9 (43.0)
GPI	64.9 (62.4)	89.1 (83.2)	72.2 (71.5)	88.8 (81.3)	79.2 (85.5)	62.4 (62.5)	72.8 (83.6)	73.8 (70.9)	85.7 (80.7)	**88.3** (**83.7**)	**92.4** (**85.6**)	90.1 (80.4)	92.8 (84.5)
HECI	55.6 (56.7)	51.9 (48.3)	97.5 (90.7)	79.4 (72.3)	99.9 (98.0)	58.4 (55.0)	99.1 (92.0)	62.4 (55.0)	22.9 (33.0)	53.8 (49.0)	59.4 (55.0)	63.5 (56.0)	78.9 (65.0)
IGCI	55.9 (55.0)	16.0 (21.3)	78.0 (68.0)	57.2 (57.0)	98.5 (89.0)	99.5 (97.0)	98.6 (95.0)	98.7 (94.0)	92.9 (86.0)	32.2 (36.0)	37.8 (42.0)	**94.4** (**86.0**)	65.3 (59.0)
LCUBE	55.7 (52.5)	31.8 (31.3)	43.5 (33.1)	87.8 (**82.1**)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	**100.0** (**100.0**)	71.4 (62.3)	78.1 (66.7)	88.7 (76.8)	**95.4** (86.0)
LOCI	56.9 (57.3)	82.7 (73.0)	71.4 (70.0)	89.9 (77.7)	**100.0**(**100.0**)	99.9 (99.0)	91.1 (85.0)	58.4 (58.0)	99.7 (96.0)	82.7 (72.0)	79.6 (68.0)	81.9 (73.0)	81.3 (74.0)
NNCL	49.6 (50.0)	48.5 (48.7)	57.6 (55.7)	79.9 (66.7)	93.4 (84.0)	26.8 (33.0)	86.2 (84.0)	17.6 (19.0)	78.7 (60.0)	67.1 (64.0)	68.5 (59.0)	71.7 (62.0)	69.6 (54.0)
RECI	58.8 (56.0)	71.1 (64.3)	94.8 (85.3)	65.8 (60.3)	8.4 (18.0)	24.8 (35.0)	15.2 (22.0)	39.0 (44.0)	5.6 (13.0)	45.2 (44.0)	55.7 (53.0)	43.7 (39.0)	38.1 (44.0)
ROCHE	54.7 (51.3)	21.7 (28.3)	86.2 (73.7)	73.9 (69.0)	**100.0**(**100.0**)	**100.0**(**100.0**)	**100.0**(**100.0**)	99.0 (99.0)	**100.0** (**100.0**)	43.0 (45.0)	47.6 (46.0)	83.7 (72.0)	81.7 (71.0)
SLOPE	59.3 (57.0)	73.3 (67.3)	96.9 (88.7)	67.1 (62.3)	9.2 (18.0)	23.4 (28.0)	12.6 (21.0)	11.0 (17.0)	1.1 (7.0)	47.9 (45.0)	57.2 (54.0)	45.1 (46.0)	44.5 (47.0)
SLOPPY	58.5 (54.0)	68.3 (59.0)	96.6 (**90.3**)	70.1 (62.3)	11.2 (17.0)	15.4 (20.0)	14.3 (22.0)	6.6 (10.0)	0.5 (5.0)	47.7 (46.0)	56.9 (53.0)	46.3 (46.0)	56.9 (54.0)