The Effect of Noise Level on the Accuracy of Causal Discovery Methods with Additive Noise Models

Benjamin Kap[0000−0002−9230−9341], Marharyta Aleksandrova[0000−0002−1863−0129], and Thomas Engel[0000−0002−7374−3927]

University of Luxembourg
2, avenue de l'Université
L-4365 Esch-sur-Alzette
{benjamin.kap, marharyta.aleksandrova, thomas.engel}@uni.lu

Abstract. In recent years a lot of research was conducted within the area of causal inference and causal learning. Many methods were developed to identify the cause-effect pairs. These methods also proved their ability to successfully determine the direction of causal relationships from observational real-world data. Yet in bivariate situations, causal discovery problems remain challenging. A class of methods, that also allows tackling the bivariate case, is based on Additive Noise Models (ANMs). Unfortunately, one aspect of these methods has not received much attention until now: *what is the impact of different noise levels on the ability of these methods to identify the direction of the causal relationship?* This work aims to bridge this gap with the help of an empirical study. We consider a bivariate case and two specific methods *Regression with Subsequent Independence Test* and *Identification using Conditional Variances*. We perform a set of experiments with an exhaustive range of ANMs where the additive noises' levels gradually change from 1% to 10000% of the causes' noise level (the latter remains fixed). Additionally, we consider several different types of distributions as well as linear and non-linear ANMs. The results of the experiments show that these causal discovery methods can fail to capture the true causal direction for some levels of noise.

Keywords: Causal Learning · Additive Noise Models · Noise Level.

1 Introduction

Thanks to the technological and computational advances during the last decades, scientists were able to tackle successfully non-trivial problems from different research areas, with causality being a prominent example. One of the fundamental problems of causality theory is to determine the causal relationship between two

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-93842-0_7.

or more variables. This problem is known as *causal discovery*, *causal identification* or *structure learning* [8, 27]. For example, given altitude and temperature, we want to answer the question if the temperature has an effect on altitude, or if altitude has an effect on temperature. This is of particular interest since if such a causal relationship is known then one can predict the effects on a system in case of an intervention or a perturbation.

Controlled experimentation, or A/B tests, are considered to be a golden standard for causal discovery [11, 34]. In such experiments, there are two identical groups with only one variation. The only variable that is varied (intervened on) is the potential cause. This procedure allows estimating the causal effect of this variable in a given system. A/B tests are widely used in practical applications. For example, testing the efficacy of medications is usually done with A/B tests, see [32] for an example. In this case, the first group, also known as *control group*, receives no medication or a placebo, and the second group, known as *intervention group*, receives the real medication. The results show the true effect (if any) of the medication on human health. However, such tests are often too expensive, unethical, or even technically impossible to execute. For example, to test the effect of smoking on health with this approach, one needs two non-smoker groups. Next, the members of one group should be forced to smoke, and the others not do so. Therefore, it is of great interest to determine causal relationships from observational data only.

There exist many methods which are able to determine causal relationships from observational data. One particular group of such methods is based on *Additive Noise Models* (ANMs). These methods, as the name suggests, exploit the additivity of the random hidden noise. ANMs received a lot of attention as they are well established and yielded many good results [12]. Despite all the research in the past years, one small but nonetheless important aspect of causal discovery with ANMs has not received much attention: how do different noise *levels* of the additive noise impact the correctness of these methods? In the real world, it can occur that noise levels change drastically from cause to effect. It can happen, for example, if the data collection process has a lot of interference like in outer space.

In this work, we aim to bridge this research gap with an empirical study. For our analysis, we selected two specific methods: *Regression with Subsequent Independence Test (Resit)* [20] and *Identification using Conditional Variances (Uncertainty Scoring)* [17]. We chose Resit, as it is known to produce reliable results [15]. However, this method is not capable to identify the correct causal direction in the case both the cause and the noise are Gaussian. In fact, this case was only recently successfully tackled by the Uncertainty Scoring method. That is why we chose the latter one as well. We perform a set of experiments with an exhaustive range of ANMs where the additive noises' levels gradually change from 1% to 10000% of the causes' noise level (the latter remains fixed). We also consider several types of distributions as well as linear and non-linear data. The results of the experiments show that these causal discovery methods can fail to capture the true causal direction for some levels of noise.

This paper is organized as follows. In Section 2 we introduce related work. Next, in Section 3 we describe the chosen causal discovery methods. In Section 4 and Section 5 we discuss the experimental setup and the experimental results respectively. Lastly, in Section 6 we draw conclusions and present possible future work.

2 Related Work

Structure learning is the procedure of determining causal relationship directions from observational data only and representing these as a (causal) graph. The basic idea emerged from [33] as *path analysis*.

Judea Pearl presented in his work [8] a comprehensive theory of causality and unified the probabilistic, manipulative, counterfactual, and structural approaches to causation. From this work we have the following key point. If there is a statistical association, e.g. two variables X and Y are dependent, then one of the following is true: 1) there is a causal relationship, either X has an effect on Y or Y has an effect on X; 2) there is a common cause (*confounder*) that has an effect on both X and Y; 3) there is a possibly unobserved common effect of X and Y that is conditioned upon data acquisition (selection bias); or 4) there can be a combination of these. From there on, a lot of research has been conducted to develop theoretical approaches and methods for structure learning. In the rest of this section, we first introduce the common concept behind all these approaches, and then we present some major works related to additive noise models.

In general, all methods for structure learning exploit the complexity of the marginal and conditional probability distributions in some way, see [1–7, 9, 13, 14, 16, 18–25, 27–30, 35]. Under certain assumptions, these methods are then able to solve the task of causal discovery. Let C denote the cause and E the effect. Then their joint density can be expressed with $p_{C,E}(c, e)$. This joint density can be factorized into either (1) $p_C(c) \cdot P_{E|C}(e|c)$ or (2) $p_E(e) \cdot P_{C|E}(c|e)$. The idea is then that (1) gives models of lower total complexity than (2) and this allows us to conclude the causal relationship direction. Intuitively, this makes sense, because the effect contains information from the cause but not vice-versa (of course, under the assumption that there are no cycles aka feedback loops). Therefore, (2) has at least as much complexity as (1). However, the definition of complexity is ambiguous. For example, one can say that "p_C *contains no information about* $P_{E|C}(e|c)$" and then draw partial conclusions about the causal direction in a given system. This complexity question is often colloquially referred to as *breaking the symmetry*, that is $p_C(c) \cdot P_{E|C}(e|c) \neq p_E(e) \cdot P_{C|E}(c|e)$.

As it was already mentioned, causal discovery based on ANMs was widely studied in the research literature. Silva et al. introduced in [26] a method for learning the structure of linear latent variable models. The main assumption in their work is that each variable is a linear function of its parents plus an additive error term of positive finite variance. Hoyer et al. generalized the linear framework of additive noise models to the nonlinear case [4]. Earlier works often assumed linear models for continuous variables. The authors showed that if

data contains non-Gaussian variables, then this can help in distinguishing the causal directions and identifying the causal graph. Mooij et al. introduced Resit[1] method in [13]. This method is based on the idea of minimizing the statistical dependence between the regressors and residuals[2]. The authors demonstrated that if the residuals are no longer dependent on the input, then regression can successfully model the causal dependence. This method does not need to assume a particular distribution of the noise because any form of regression can be used (e.g., Linear Regression), and it is well suited for the task of causal inference in additive noise models. Next, Mooij et al. introduced a method to determine the causal relationship in cyclic additive noise models and showed that such models are generally identifiable in the bivariate, Gaussian-noise case [14]. Their method works for continuous data and can be seen as a special case of nonlinear independent component analysis. Later, Peters and Bühlmann proved in [19] *full identifiability*[3] of linear Gaussian structural equation models if all the noise variables have the same variance. In the next work, Peters et at. proposed a method that can identify the directed acyclic graph from the distribution under mild conditions [20]. In contrast, previous methods assumed faithfulness and could only identify the Markov equivalence class of the graph[4]. Finally, the authors of [1, 18] proved that linear Gaussian models with different error variance can be also identifiable. In their method, referred to as Uncertainty Scoring[5], this is done by ordering variables according to the law of total variances and then performing independence tests between them. Park extended this result to additive noise models in [17].

As we can see, many researchers contributed to the development of ANMs-based causal discovery methods and widened our understanding of their application cases. However, no previous research work analyzed how the level of noise variance relative to that of the cause variance can impact the accuracy of these methods. This question forms the basis of the current study.

3 Causal Discovery Methods

In this section, we introduce notations and then describe to two analyzed causal discovery methods: *Regression with Subsequent Independence Test (Resit)* [20], see Section 3.2, and *Identification using Conditional Variances (Uncertainty Scoring)* [17], see Section 3.3.

[1] Resit method is described in Section 3.2.
[2] The residuals are defined as the difference between the actual output and the predicted output.
[3] *Full identifiability* means that not only the skeleton of the causal graph is recoverable but also the arrows are.
[4] *Markov equivalence class* refers to the class of graphs in which all graphs have the same skeleton.
[5] Uncertainty Scoring method is described in Section 3.3.

3.1 Notations

In the following text, we give a short definition of additive noise models for the bivariate case. For more details and multivariate cases, please refer to [4, 20].

Let $X, Y \in \mathbb{R}$ be the cause and the effect respectively. Let there also be m latent (hidden) causes $U = (U_1, \ldots, U_m) \in \mathbb{R}^m$. Then the causal relationship can be modeled as follows.

$$\begin{cases} Y = f(X, U_1, \cdots, U_m) \\ X \perp\!\!\!\perp U \end{cases}, \text{ with } X \sim p_X(x) \text{ and } U \sim p_U(u_1, \cdots, u_m),$$

where $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ is a linear or nonlinear function, and $p_X(x)$ and $p_U(u_1, \cdots, u_m)$ are the joint densities of the observed cause X and the latent causes U. We assume that there is no confounding, no selection bias, and no feedback loops between X and Y. In this case, X and U are independent, which is denoted by $X \perp\!\!\!\perp U$. Since the latent causes U are unobserved, their influence can be summarized with a single noise variable $N_y \in \mathbb{R}$, and the model can be rewritten as follows:

$$\begin{cases} Y = f(X, N_y) \\ X \perp\!\!\!\perp N_y \end{cases}, \text{ with } X \sim p_X(x) \text{ and } N_y \sim p_{N_y}(n_y).$$

In our experiments, we are considering both linear and nonlinear additive noise models:

$$Y = \beta X + N_y \text{ with } \beta \in \mathbb{R}, \text{ for the linear case}$$

and

$$Y = \beta X^\alpha + N_y \text{ with } \beta, \alpha \in \mathbb{R}, \text{ for the nonlinear case.}$$

Also, X and N_y can be drawn from one of the following three distributions: the normal distribution denoted by the calligraphic letter $\mathcal{N}$, the uniform distribution denoted by the calligraphic letter $\mathcal{U}$, or the Laplace distribution denoted by the calligraphic letter $\mathcal{L}$. For example, throughout this work "*X is drawn from a normal distribution*" is denoted by $X \sim \mathcal{N}$ or $X \sim \mathcal{N}(\mu_x, \sigma_x)$ with μ_x standing for the mean and σ_x for the standard deviation.

3.2 Regression with Subsequent Independence Test (Resit)

We implement Resit following Algorithm 1 from [15]. This algorithm requires the following inputs: X and Y, a regression method, and a score estimator $\hat{C} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$; it outputs *dir* (casual relationship **dir**ection). The idea is to regress Y on X, predict $\hat{Y}$, and then calculate residuals $Y_{res} = \hat{Y} - Y$. Y_{res} and X are then used to calculate $\hat{C}_{X \to Y}$, a score for the assumed case $X \to Y$. Similarly, to test the other causal direction ($Y \to X$), we regress X on Y, calculate residuals $X_{res} = \hat{X} - X$ and estimate $\hat{C}_{Y \to X}$. In our experiments, the generated data always follows $X \to Y$. This verifies the **assumption** that only

Algorithm 1 General procedure to decide whether $p(x, y)$ satisfies Additive Noise Model $X \to Y$ or $Y \to X$.

Input:
- I.i.d. sample data X and Y
- Regression method
- Score estimator $\hat{C} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$

Output:
- *dir*

1: $reg_1 \leftarrow$ Regress Y on X
2: $reg_2 \leftarrow$ Regress X on Y

3: $Y_{res} \leftarrow reg_1.predict(X) - Y$
4: $X_{res} \leftarrow reg_2.predict(Y) - X$

5: $\hat{C}_{X\to Y} \leftarrow \hat{C}(X, Y_{res})$
6: $\hat{C}_{Y\to X} \leftarrow \hat{C}(Y, X_{res})$

$$\textbf{return } dir = \begin{cases} X \to Y & \text{if } \hat{C}_{X\to Y} < \hat{C}_{Y\to X}, \\ Y \to X & \text{if } \hat{C}_{X\to Y} > \hat{C}_{Y\to X}, \\ ? & \text{if } \hat{C}_{X\to Y} = \hat{C}_{Y\to X}. \end{cases} \tag{1}$$

one direction in our data is correct (and not both). Under this assumption, we can compare both scores directly to decide on the cause-effect direction, and we do not need to determine the value of α for the independence tests, see Eq. (1). Additionally, we can also use entropy estimators to estimate the score $\hat{C}$.

In Algorithm 1, it is possible to split the data into training and test parts. In this case, the training data is used to fit the regression model and the test data is used to calculate the value of $\hat{C}$. This procedure is referred to as *decoupled estimation* [12]. The advantage of splitting the data lies in the reduction of the computational time for calculating independence estimates $\hat{C}$. However, in this work, we use *coupled estimation*. This means that the entire data-set is used for both the regression and the independence estimation steps. The latter approach tends to produce more accurate results for independence estimation.

In our work, we use Linear Regression as a regression algorithm. If an appropriate transformation of coordinates is applied, Linear regression can be used in the non-linear cases as well. In our experiments, we used six different independence tests and six different entropy measures for calculating $\hat{C}$. In general, for the independence tests we have:

$$\hat{C}(X_{Test}, Y_{res}) = I(X_{Test}, Y_{res}),$$

with $I(\cdot, \cdot)$ being any independence test. In the case of entropy estimators we have:

$$\hat{C}(X_{Test}, Y_{res}) = H(X_{Test}) + H(Y_{res}),$$

with $H(\cdot)$ being any entropy measure. The entropy-based estimator score is derived from Lemma 1 in [12].

The following estimators were used in this work. The implementation of estimators with numbers 2 - 12 was taken from the *information theoretical estimators* toolbox [31]. Here we briefly introduce every estimator. Mathematical formulas for each of them can be found in the Appendix.

1. *HSIC*: Hilbert-Schmidt Independence Criterion with RBF Kernel [6].
2. *HSIC_IC*: Hilbert-Schmidt Independence Criterion using incomplete Cholesky decomposition[7].
3. *HSIC_IC2*: Same as HSIC_IC but with lower precision.
4. *DISTCOV*: Distance covariance estimator using pairwise distances.
5. *DISTCORR*: Distance correlation estimator using pairwise distances. It is simply the standardized version of the distance covariance.
6. *HOEFFDING*: Hoeffding's Phi.
7. *SH_KNN*: Shannon differential entropy estimator using kNNs (k-nearest neighbors) where $k = 3$.
8. *SH_KNN_2*: Same as SH_KNN but with different search method.
9. *SH_KNN_3*: Same as SH_KNN but with $k = 5$.
10. *SH_MAXENT1*: Maximum entropy distribution-based Shannon entropy estimator.
11. *SH_MAXENT2*: Same as SH_MAXENT1 with minor changes.
12. *SH_SPACING_V*: Shannon entropy estimator using Vasicek's spacing method.

3.3 Identification using Conditional Variances (Uncertainty Scoring)

The Uncertainty Scoring method is composed of Algorithm 2 and Algorithm 3 from [17]. It consists of two parts: 1) ordering and 2) conditional independence testing.

For the first step, ordering, we used *backward step-wise selection* (Algorithm 2), as it is more convenient for implementation. The algorithm starts with a set S which contains all variables represented as nodes in a causal graph. Next, we iterate over S, and for each node, we calculate its conditional variance given all other remaining nodes. Then, we select the node with the highest conditional variance, append it to the ordering π, and also remove it from the set S. With the updated set S, we repeat this process until S is empty. Lastly, the *reverse* of the ordering π is returned. The first node to be appended to the ordering is the last one in the ordering, which is reflected in the name "*backward step-wise selection*".

In the second step, we perform uncertainty scoring using Algorithm 3. This algorithm iterates over the ordering π. For every node j, it performs conditional

[6] Source: https://github.com/amber0309/HSIC

[7] Low rank decomposition of Gram matrices, which permits an accurate approximation to HSIC as long as the kernel has a fast decaying spectrum.

Algorithm 2 Backward step-wise selection

Input: All variables from an ANM: $X = (x_1, x_2, \ldots, x_n)$
Output: Estimated ordering $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$

Set $S = \{1, 2, \ldots, n\}$
List $\pi = [\ ]$
for $m = 1 \ldots n$ **do**
    for $j \in S$ **do**
        Estimate the conditional variance x_j given $\{x_1, \ldots, x_n\} \backslash x_j$, $\sigma^2_{j|S\backslash \mathrm{j}}$
    end
    Append $\pi_m = argmax_j \sigma^2_{j|S\backslash j}$ to π
    Update $S = S \backslash \pi_m$
end
return Reversed list π

independence tests conditioning on every other node l appearing before the node j in the ordering π. If a node l is dependent on j, then it is added to the set of parents of j, denoted as $Pa(j)$. In this algorithm, the first node in the ordering never has parents, so the procedure starts with the second node. *Fisher's z-transform of the partial correlation,* is used for the conditional independence testing.

Algorithm 3 Uncertainty Scoring

Input: All variables from an ANM: $X = (x_1, x_2, \ldots, x_n)$
Output: Dictionary with estimated parents for all variables: $G = \{Pa(x_1) : [\ldots], Pa(x_2) : [\ldots], \ldots, Pa(x_n) : [\ldots]\}$

Get ordering from backward step-wise selection: $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$
$G = \{\}$
for $m = 2 \ldots n$ **do**
    $Pa(\pi_m) = [\ ]$
    for $j = 1 \ldots m - 1$ **do**
        Conditional independence test between π_m and π_j given $\{\pi_1, \ldots, \pi_{m-1}\} \backslash \pi_j$
        If dependent, include π_j into $Pa(\pi_m)$
    end
    Insert $Pa(\pi_m)$ into G
end
return G

4 Experimental setup

Generation of synthetic data. For all empirical tests, we assume X to be a cause of Y, that is $X \to Y$. In the sense of additive noise models, we use the

following equations: $Y = X + N_y$ for the linear case, and $Y = X^3 + N_y$ for the non-linear case, where

$$X \sim \begin{cases} \mathcal{N}(0,1) & \text{or} \\ \mathcal{U}(-1,1) & \text{or} \\ \mathcal{L}(0,1) \end{cases} \quad \textit{and} \quad N_y \sim \begin{cases} \mathcal{N}(0,1\cdot i) & \text{or} \\ \mathcal{U}(-1\cdot i, 1\cdot i) & \text{or} \\ \mathcal{L}(0,1\cdot i) \end{cases}$$

with i being a scaling factor for the noise level in N_y. The goal is to analyze how different standard deviations (boundaries for the uniform case) in the noise term N_y relative to the standard deviations (or boundaries for the uniform case) in the X term impact the ANM methods.

To cover various dependencies between the distributions of X and N_y, we generate 199 different i factors:

$$i \in \{0.01, 0.02, \ldots, 1.00\} \cup \{1, 2, \ldots, 100\}.$$

For each i, every linear and non-linear combination with different distributions is tested. Totally, we have 18 combinations corresponding to the general structures $Y = X + N_y$ and $Y = X^3 + N_y$, where X and N_y are drawn from the three different distributions, $\mathcal{N}$, $\mathcal{U}$ or $\mathcal{L}$.

$$\begin{aligned} Y = X \sim \mathcal{N} + N_y \sim \mathcal{N}, \\ Y = X \sim \mathcal{N} + N_y \sim \mathcal{U}, \\ Y = X \sim \mathcal{N} + N_y \sim \mathcal{L}, \\ \vdots \\ Y = X \sim \mathcal{L}^3 + N_y \sim \mathcal{L}. \end{aligned}$$

Note that $\mathcal{L}^3$ here signifies the non-linear case $Y = X^3 + N_y$.

Evaluation. For each of the 18 combinations, we perform 100 tests. In every test, we generate 1000 new samples for X and N_y and attempt to identify the direction of the causal relationship[8] using one of the two algorithms presented in Section 3. Lastly, we simply calculate the fraction of successful tests and define this ratio as our accuracy measure.

5 Experimental Results

Since we used a large range for the values of i-factor, several different combinations of distributions, linear and non-linear data, we have too many results to show them all in detail in this paper. Therefore, we discuss several representative cases and provide a summary of all results. The latter shows for which values of i-factor the models are consistently identifiable. For the detailed analysis, we refer to the document [10]. Alternatively, all the results and source codes can be accessed from the relative repository[9].

[8] The true direction of the causal relationship is known as we generate synthetic data.
[9] https://gitlab.com/Shinkaiika/noise-level-causal-identification-additive-noise-models

5.1 Resit

We start with the analysis of Resit method. In this set of experiments, we are interested in which ranges of i-factor allow causal identifiability and how it is related to the functional model and the chosen independence estimator. Fig. 1 shows the detailed results for the following 4 linear combinations and their non-linear counterparts: $Y = \mathcal{N} + \mathcal{U}$, $Y = \mathcal{U} + \mathcal{N}$, $Y = \mathcal{U} + \mathcal{L}$, and $Y = \mathcal{L} + \mathcal{L}$. The y-axis shows the accuracy of causal discovery ($\frac{\#\text{successful tests}}{100}$), and the x-axis corresponds to i-factor. Different colors encode 12 estimators used in this work. The value of accuracy close to 0.5 means that Resit outputs the correct causal direction in only 50% of the tests thus indicating **unidentifiability.** The values close to 1 signify very good/consistent **identifiability**. In the following text, we analyze the results for individual models.

Fig. 1a shows the linear model $Y = \mathcal{N} + \mathcal{U}$. We can see, that all estimators reach an accuracy close to 100% inside the interval $i \in [0.8; 5]$. However, for smaller or larger i-factors the accuracy of all estimators start to drop until they reach unidentifiability (~ 0.5). Not all estimators perform the same. For example, HISC with Incomplete Cholesky decomposition performs worse for decreasing i-factors compared to all other estimators. SH_SPACING_V performs the best among all estimators for this linear model. Fig. 1b shows the non-linear model $Y = \mathcal{N}^3 + \mathcal{U}$. The non-linear version shows much better results. With $i \in [0.2; 100]$, we have accuracy close to 100% for all estimators. Only a few estimators drop towards unidentifiability for $i < 0.2$.

Fig. 1c shows the linear model $Y = \mathcal{U} + \mathcal{N}$. For $i \in [0.1; 1]$ this model is identifiable. However, for larger values of i-factor, the accuracy of many estimators drop quickly. In this range, SH_SPACING_V remains above 90%, most other estimators drop between 60% and 80% but HSIC_IC and HSIC_IC2 drop to 50% accuracy demonstrating complete unidentifiability. Fig. 1d shows the results for the non-linear version of this model. For $i \leq 1$, all estimators remain above 90% accuracy, with the exceptions now being HSIC_IC and HSIC_IC2. For i-factors larger than 1, estimators behave differently. SH_KNN, SH_KNN_2, SH_KNN_3, DISTCOV, DISCORR and HOEFFDING remain above 90% accuracy up to $i = 100$. SH_MAXENT1 remains between 80% and 90%, HSIC and SH_MAXENT2 between 60% and 80%, and HSIC_IC and HSIC_IC2 become unidentifiable.

Fig. 1e shows the linear case $Y = \mathcal{U} + \mathcal{L}$ and Fig. 1f shows the non-linear case $Y = \mathcal{U}^3 + \mathcal{L}$. The demonstrated results are quite similar to the two cases discussed above. This indicates that models with the same type of distribution for X behave similarly.

Fig. 1g shows the linear case $Y = \mathcal{L} + \mathcal{L}$. For $i \in [0.1; 10]$ most estimators are above 90%, except SH_KNN, SH_KNN_2 and SH_KNN_3 which are above 90% for $i \in [0.4; 2]$. For larger values of i-factor, all estimators drop quickly to unidentifiability. Finally, Fig. 1h shows the non-linear case $Y = \mathcal{L}^3 + \mathcal{L}$. Similarly to the model $Y = \mathcal{N}^3 + \mathcal{U}$ presented in Fig. 1b, this model demonstrates that non-linearity generally helps in identifying causal relationships. For $i \in [0.15; 100]$ all estimators are above 90% accuracy, often reaching 100%.

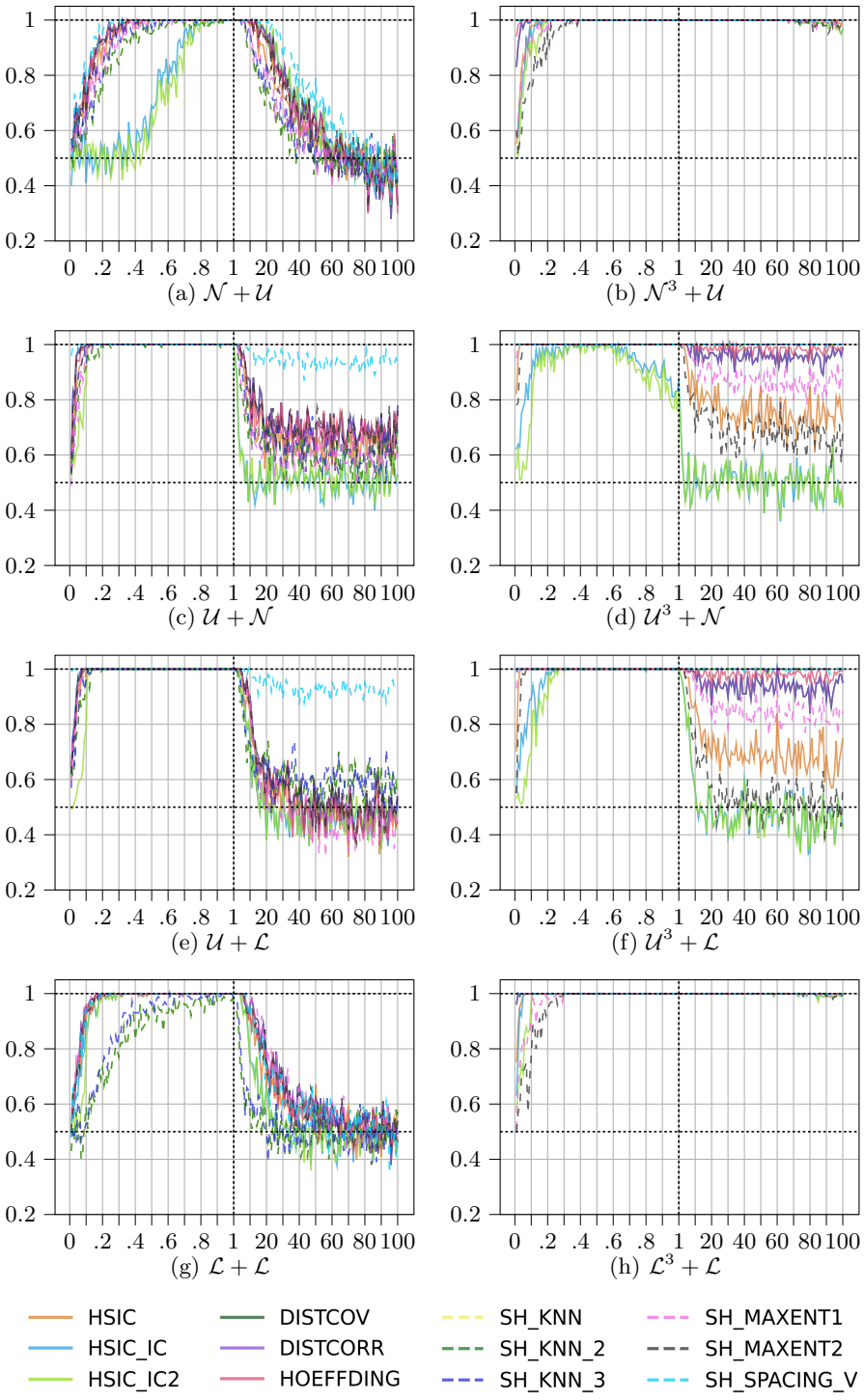

Fig. 1: Several selected detailed results for Resit. *x-axis shows the values of i-factor and y-axis shows the accuracy of causal identification.*

Table 1: Summary for Resit with linear models. The numbers reflect the ranges of i-factor that allow identifiability with accuracy around or above 90%.

Equation	$\mathcal{N}+\mathcal{N}$	$\mathcal{N}+\mathcal{U}$	$\mathcal{N}+\mathcal{L}$	$\mathcal{U}+\mathcal{N}$	$\mathcal{U}+\mathcal{U}$	$\mathcal{U}+\mathcal{L}$	$\mathcal{L}+\mathcal{N}$	$\mathcal{L}+\mathcal{U}$	$\mathcal{L}+\mathcal{L}$
HSIC		0.17 - 18	0.13 - 8	0.05 - 6	0.06 - 16	0.04 - 7	0.1 - 7	0.12 - 23	0.1 - 13
HSIC_IC		0.65 - 26	0.31 - 7	0.04 - 3	0.06 - 15	0.04 - 5	0.1 - 4	0.14 - 26	0.1 - 8
HSIC_IC2		0.7 - 26	0.33 - 7	0.1 - 3	0.14 - 15	0.11 - 5	0.1 - 4	0.14 - 26	0.12 - 8
DISTCOV		0.16 - 23	0.13 - 7	0.04 - 7	0.05 - 21	0.04 - 10	0.1 - 7	0.1 - 25	0.08 - 15
DISTCORR		0.16 - 23	0.13 - 7	0.04 - 7	0.05 - 21	0.04 - 10	0.1 - 7	0.1 - 25	0.08 - 15
HOEFFDING		0.16 - 25	0.13 - 8	0.04 - 7	0.05 - 21	0.04 - 8	0.1 - 7	0.1 - 25	0.1 - 10
SH_KNN		0.32 - 12	0.76 - 1	0.08 - 4	0.07 - 12	0.09 - 4	0.61 - 1	0.27 - 12	0.37 - 3
SH_KNN_2		0.32 - 12	0.76 - 1	0.08 - 4	0.07 - 12	0.09 - 4	0.61 - 1	0.27 - 12	0.37 - 3
SH_KNN_3		0.24 - 12	0.51 - 1	0.05 - 5	0.07 - 14	0.05 - 5	0.37 - 3	0.21 - 15	0.32 - 4
SH_MAXENT1		0.23 - 12	0.12 - 10	0.06 - 4	0.1 - 12	0.04 - 8	0.07 - 13	0.11 - 24	0.07 - 17
SH_MAXENT2		0.15 - 22	0.13 - 7	0.03 - 7	0.05 - 17	0.04 - 8	0.1 - 7	0.11 - 23	0.1 - 13
SH_SPACING_V		0.13 - 33	0.17 - 5	0.01 - 100	0.03 - 40	0.01 - 100	0.14 - 6	0.11 - 33	0.09 - 13

Table 2: Summary for Resit with non-linear data. The numbers reflect the ranges of i-factor that allow identifiability with accuracy around or above 90%.

Equation	$\mathcal{N}^3+\mathcal{N}$	$\mathcal{N}^3+\mathcal{U}$	$\mathcal{N}^3+\mathcal{L}$	$\mathcal{U}^3+\mathcal{N}$	$\mathcal{U}^3+\mathcal{U}$	$\mathcal{U}^3+\mathcal{L}$	$\mathcal{L}^3+\mathcal{N}$	$\mathcal{L}^3+\mathcal{U}$	$\mathcal{L}^3+\mathcal{L}$
HSIC	0.04 - 100	0.08 - 100	0.04 - 100	0.02 - 6	0.03 - 16	0.03 - 7	0.02 - 100	0.04 - 100	0.02 - 100
HSIC_IC	0.04 - 83	0.06 - 100	0.04 - 70	0.1 - 0.92	0.14 - 13	0.1 - 4	0.03 - 100	0.05 - 100	0.03 - 100
HSIC_IC2	0.08 - 83	0.08 - 100	0.09 - 70	0.12 - 0.91	0.17 - 13	0.17 - 4	0.7 - 100	0.07 - 100	0.09 - 100
DISTCOV	0.02 - 100	0.02 - 100	0.02 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100
DISTCORR	0.02 - 100	0.02 - 100	0.02 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100
HOEFFDING	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100
SH_KNN	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100
SH_KNN_2	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100
SH_KNN_3	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100
SH_MAXENT1	0.05 - 100	0.06 - 100	0.05 - 100	0.01 - 100	0.02 - 90	0.01 - 88	0.1 - 100	0.17 - 100	0.1 - 100
SH_MAXENT2	0.11 - 98	0.16 - 100	0.1 - 100	0.03 - 4	0.04 - 12	0.04 - 5	0.14 - 100	0.15 - 100	0.15 - 100
SH_SPACING_V	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100	0.01 - 100

The experimental results for Resit with linear and non-linear models are summarized in Tables 1 and 2 respectively. The rows correspond to different estimators, and columns correspond to structural equation models. The values in the cells show on what range of i a particular estimator *can* reach over 90% accuracy. Estimators have some variance in the results and thus on some intervals they fall below 90% accuracy. The limits in the cells were chosen as follows: the lower limit designates where an estimator reaches 90% or higher for the first time, and the upper limit designates for which value of i it was observed for the last time. In between, most of the time estimators remain above 90% or rarely fall below, but never below 80% accuracy. An empty cell means that the corresponding estimator never resulted in accuracy $\geq 90\%$.

As the results show, different noise levels do have an impact on the identifiability performance of Resit. In general, the linear equation models are more fragile than the non-linear ones. This is explained by the fact that the non-linear

relationships tend to break the symmetry between the variables easier, see [4]. The only structural equation which always remains unidentifiable is $Y = \mathcal{N} + \mathcal{N}$, see [24]. For all other cases, all estimators reach an accuracy of over 90% for some values of i-factor. For example, all estimators perform perfectly when the noise level of the X term is comparable to the noise level of the corresponding noise term (N_y), that is $i = 1$. For other values of i, there are differences between linear and non-linear equations. Generally, the accuracy for linear cases drops if $i > 7$. However, most non-linear cases retain accuracy over 90% for much larger values of i-factor, even up to 100. Similar results are observed for the decreasing i-factors.

We can also observe differences between estimators in terms of accuracy. For example, HSIC is overall the best performing independence estimator while HSIC_IC and HSIC_IC_2 perform the worst. SH_SPACING_V is the best performing entropy estimator while SH_MAXENT1 and SH_MAXENT2 perform the worst. Some estimators show better performance for particular structural causal models, for example, SH_SPACING_V for $Y = \mathcal{U} + \mathcal{N}$; others are particularly unsuitable for some structural equations, for example, HSIC_IC and HSIC_IC2 for $Y = \mathcal{N} + \mathcal{U}$. For all non-linear equation models, SH_SPACING_V and the three Shannon kNN estimators result in accuracy close to 100% for all values of i. SH_SPACING_V also keeps its good performance in the case of linear equation models. As for independence measures, HSIC, DISTCOV, DISTCORR, and HOEFFDING perform quite similarly and are good overall. Note again, that these results are based on the assumption that in our bivariate structure only one direction of the causal relationship is present, namely $X \to Y$. Without this assumption, we cannot compare the estimates directly but rather need to compare the estimate to a derived p-value given some significance level α.

5.2 Uncertainty Scoring

Fig. 2 shows the results for the Uncertainty Scoring algorithm. Recall that for these experiments we use only one estimator, the Fisher's conditional independence test. Therefore, we use different colors and styles of lines to encode structural equation models. The colours of the lines correspond to the distribution type of the noise variable N_y with the following coding: blue for $N_y \sim \mathcal{N}$, green for $N_y \sim \mathcal{U}$, and red for $N_y \sim \mathcal{L}$. The type of the lines encodes the distribution type of the cause X as follows: solid line for $X \sim \mathcal{N}$, dashed line for $X \sim \mathcal{U}$, and dotted line for $X \sim \mathcal{L}$. As in the previous experiment, the x-axis shows the values of i-factor and the y-axis shows the accuracy of causal identification. However, the results should be interpreted differently. The Uncertainty Scoring method generates a set of parents for every variable. This set can be empty or can contain cause variables. Therefore, only one structure of this result is correct and thus the y-axis of the plots in Fig. 2 shows consistent identifiability at 1, and consistent unidentifiability at 0.

We proceed to the analysis of the results. First, we can notice that the linear Gaussian model $Y = \mathcal{N} + \mathcal{N}$ is now identifiable, as it was demonstrated by the

authors of this method [17]. Interestingly, for this method, the linear cases perform better than the non-linear as opposed to Resit. Only the non-linear cases where the cause X is drawn from the Uniform distribution $\mathcal{U}$ show the same performance as the linear cases. This group of models demonstrates good identifiability for $i < 1$, however the accuracy drops fast for $i > 1$. The reason for accuracy degradation lies within step 2 of the method, the conditional independence test. If noise levels are significantly different, then the independence test fails to capture the correlation between the two nodes and therefore concludes that the nodes are independent (Type II Error). However, for any given i, the ordering step always performs correctly[10].

We can also notice that models with similar structures have similar performance. For example, in Fig. 2b we can clearly identify 3 groups: 1) the group of dashed lines representing models with $X \sim \mathcal{U}$ show the best performance for $i < 1$ and the worst performance for $i > 1$; 2) the group of dotted lines corresponding to models with $X \sim \mathcal{L}$ demonstrate the worst accuracy for $i < 1$ and the best accuracy for $i > 1$; finally 3) the group of solid lines that represent the models with $X \sim \mathcal{N}$ lie in the middle. A similar observation was done for Resit as well, that is the type of the distribution of the cause variable affects the accuracy of causal discovery. If we analyze the linear cases from Fig. 2a in the same way, we can notice that here the type of the distribution of the noise variable N_y probably has more impact. Indeed, the lines overlap, but they are now grouped more by colors than by line type. Again, we can observe 3 groups: 1) the group of green lines corresponding to the models with $N_y \sim \mathcal{U}$ show worse performance for $i < 1$ and better performance for $i > 1$; 2) the group of red lines representing the models with $N_y \sim \mathcal{L}$ have better performance for $i < 1$ and worse accuracy for $i > 1$; 3) and the group of blue lines corresponding to $N_y \sim \mathcal{N}$ lies in between.

The results obtained for the Uncertainty Scoring method are summarized in Table 3. Here, each row corresponds to a combination of distribution types. The second and the third columns show the results for linear or non-linear models respectively. The values inside the table are encoded in the same way as it was done for Table 1; that is they show the ranges where the method has an accuracy around or above 90%.

6 Conclusions

The results from the experiments showed that two analyzed causal discovery methods, Resit and Uncertainty Scoring, are affected by different noise scales. For significantly small noise levels in the disturbance term N_y, or significantly high noise levels, these causal discovery methods fail to capture the true causal relationship of the given structural equation model. Recall that *significantly* here depends on the model. For example, for some models, if the noise level was

[10] A quick test in python shell, with $i = 57$, $X \sim \mathcal{L}$ and $N_y \sim \mathcal{U}$ and 100 repetitions showed that in these runs the ordering was always correct but only in 35 runs (from the 100 repetitions) the independence tests were correct.

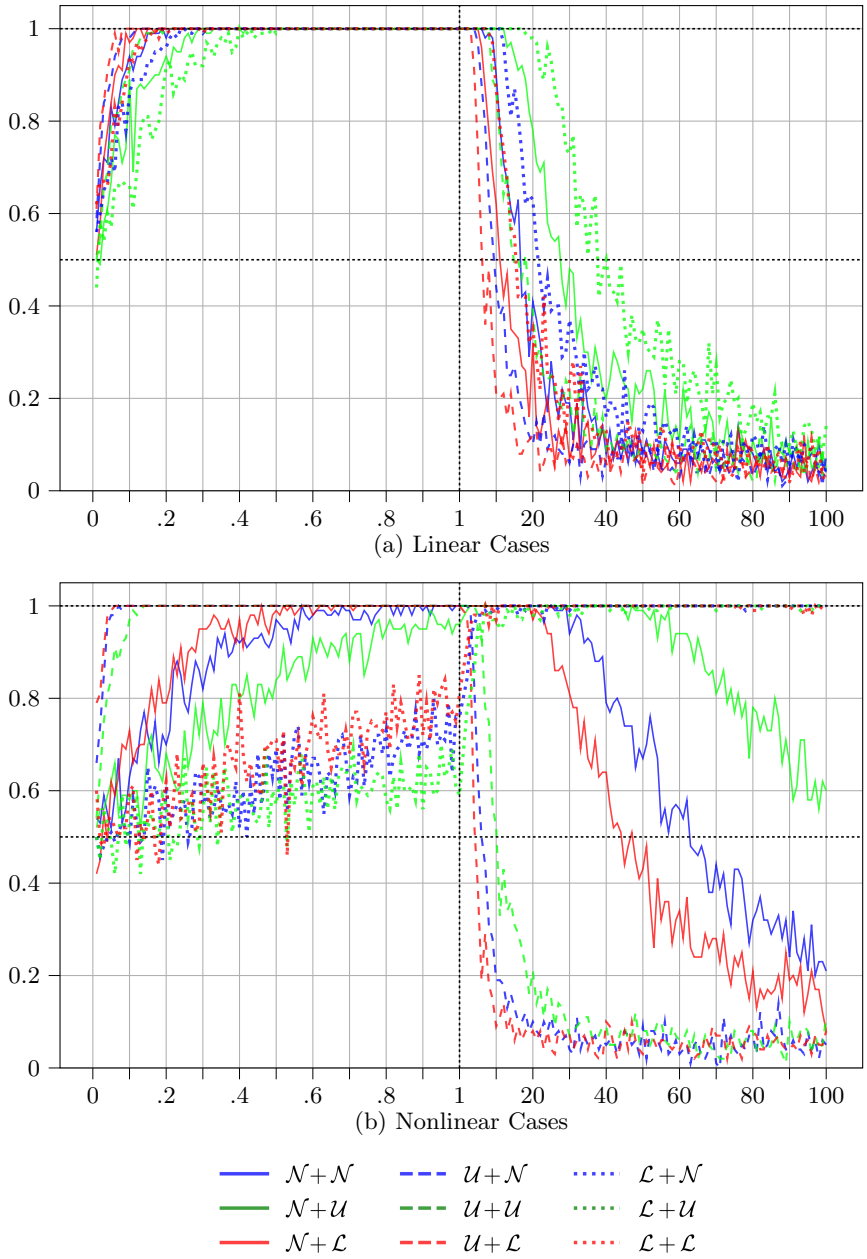

Fig. 2: Results of the Uncertainty Scoring algorithm. *x-axis shows the values of i-factor and y-axis shows the accuracy of causal identification.*

Table 3: Summary for Uncertainty Scoring. The numbers reflect the ranges of i-factor that allow identifiability with accuracy around or above 90%.

Equation	**Linear**	**Non-Linear**
$\mathcal{N}+\mathcal{N}$	0.08 - 10	0.33 - 37
$\mathcal{N}+\mathcal{U}$	0.16 - 10	0.52 - 67
$\mathcal{N}+\mathcal{L}$	0.05 - 6	0.23 - 25
$\mathcal{U}+\mathcal{N}$	0.04 - 5	0.04 - 4
$\mathcal{U}+\mathcal{U}$	0.1 - 8	0.05 - 6
$\mathcal{U}+\mathcal{L}$	0.03 - 3	0.03 - 3
$\mathcal{L}+\mathcal{N}$	0.14 - 13	4 - 100
$\mathcal{L}+\mathcal{U}$	0.19 - 26	5 - 100
$\mathcal{L}+\mathcal{L}$	0.1 - 10	2 - 100

already twice larger then the methods failed to determine the causal direction consistently. Other models remained identifiable with 100 times higher noise levels. The range of different noise levels analyzed in this work is quite exhaustive and realistically speaking having noise levels 100 times higher than the potential cause variable is very rare. Additionally, with very high noise levels the effect of the cause variable is very likely negligible anyways. However, the discovered relationships can be useful to guide researchers in practical applications. We also observed different behavior for different distribution types (e.g., Gaussian or Uniform).

For both methods, we observed that if the variance of the noise term is smaller than that of the cause, then models remained identifiable. The opposite relationship is observed when the variance of the noise term is larger. For example, often when the standard deviation of the noise term was only half of that of the cause, the model was still identifiable. However, in several cases, if the standard deviation of the noise term was already twice larger than the standard deviation of the cause, then the model became unidentifiable. We also tested linear and non-linear models and our results show that non-linear models were still identifiable in situations where the linear models are not. For example, some non-linear models, where the noise term's variance was 100 times higher than that of the cause, were still perfectly identifiable while their linear counterparts were not.

Lastly, for Resit we used several estimators: 6 independence estimators and 6 entropy estimators. Our results show differences in terms of performance depending on which estimator is used. We observed that Hilbert-Schmidt Independence Criterion with RBF Kernel was the best independence estimator, and Shannon entropy estimator using Vasicek's spacing method was the best entropy estimator.

In our experiments, we tested only two particular methods and three different distribution types. However, similar results are expected for other methods of

causal discovery with additive noise models, as their common failing point lies in the independence estimation.

Future work. In reality, observed data does not always strictly follow a certain distribution type. As there are many different possible combinations, it would be interesting to generalize the impact of different noise levels on any distribution by using the different properties an observed distribution exhibits. Furthermore, this work does not formalize mathematically the effect of different noise levels in ANM causal discovery methods. This could be done in future work.

7 Acknowledgments

This work was partially supported by the European Union Horizon 2020 research programme within the project CITIES2030 "Co-creating resilient and sustainable food towards FOOD2030", grant 101000640.

Appendix

Detailed description of estimators

1. **HSIC**: Hilbert-Schmidt Independence Criterion with RBF Kernel [11]

$$I_{HSIC}(x,y) := ||C_{xy}||^2_{HS}$$

where C_{xy} is the cross-covariance operator and HS the squared Hilbert-Schmidt norm.
2. **HSIC_IC**: Hilbert-Schmidt Independence Criterion using incomplete Cholesky decomposition (low rank decomposition of the Gram matrices, which permits an accurate approximation to HSIC as long as the kernel has a fast decaying spectrum) which has $\eta = 1 * 10^{-6}$ precision in the incomplete cholesky decomposition.
3. **HSIC_IC2**: Same as HSIC_IC but with $\eta = 1 * 10^{-2}$.
4. **DISTCOV**: Distance covariance estimator using pairwise distances. This is simply the L^2_w norm of the characteristic functions φ_{12} and $\varphi_1\varphi_2$ of input x, y:

$$\varphi_{12}(\boldsymbol{u}^1, \boldsymbol{u}^2) = \mathbb{E}[e^{i\langle \boldsymbol{u}^1, \boldsymbol{x}\rangle + i\langle \boldsymbol{u}^2, \boldsymbol{y}\rangle}],$$
$$\varphi_1(\boldsymbol{u}^1) = \mathbb{E}[e^{i\langle \boldsymbol{u}^1, \boldsymbol{x}\rangle}],$$
$$\varphi_2(\boldsymbol{u}^2) = \mathbb{E}[e^{i\langle \boldsymbol{u}^2, \boldsymbol{y}\rangle}].$$

With $i = \sqrt{-1}$, $\langle \cdot, \cdot \rangle$ the standard Euclidean inner product and $\mathbb{E}$ the expectation. Finally, we have:

$$I_{dCov}(x,y) = ||\varphi_{12} - \varphi_1\varphi_2||_{L^2_w}$$

5. **DISTCORR**: Distance correlation estimator using pairwise distances. It is simply the standardized version of the distance covariance:

$$I_{dCor}(x,y) = \begin{cases} \frac{I_{dCov}(x,y)}{\sqrt{I_{dVar}(x,x) I_{dVar}(y,y)}}, & \text{if } I_{dVar}(x,x) I_{dVar}(y,y) > 0 \\ 0, & \text{otherwise,} \end{cases}$$

with

$$I_{dVar}(x,x) = ||\varphi_{11} - \varphi_1\varphi_1||_{L^2_w},\ I_{dVar}(y,y) = ||\varphi_{22} - \varphi_2\varphi_2||_{L^2_w}$$

(see characteristic functions under 4. DISTCOV)
6. **HOEFFDING**: Hoeffding's Phi

$$I_{\Phi}(x,y) = I_{\Phi}(C) = \left(h_2(d) \int_{[0,1]^d} [C(\boldsymbol{u}) - \Pi(\boldsymbol{u})]^2 d\boldsymbol{u} \right)^{\frac{1}{2}}$$

with C standing for the copula of the input and Π standing for the product copula.

[11] Source: https://github.com/amber0309/HSIC

7. **SH_KNN**: Shannon differential entropy estimator using kNNs (k-nearest neighbors)

$$H(\boldsymbol{Y}_{1:T}) = log(T-1) - \psi(k) + log(V_d) + \frac{d}{T}\sum_{t=1}^{T} log(\rho_k(t))$$

with T standing for the number of samples, $\rho_k(t)$ - the Euclidean distance of the k^{th} nearest neighbour of $\boldsymbol{y}_t$ in the sample $\boldsymbol{Y}_{1:T}\backslash\{\boldsymbol{y}_t\}$ and $V \subseteq \mathbb{R}^d$ a finite set.
8. **SH_KNN_2**: Same as SH_KNN but using kd-tree for quick nearest-neighbour lookup
9. **SH_KNN_3**: Same as SH_KNN but with $k = 5$
10. **SH_MAXENT1**: Maximum entropy distribution-based Shannon entropy estimator

$$H(\boldsymbol{Y}_{1:T}) = H(n) - \left[k_1\left(\frac{1}{T}\sum_{t=1}^{T} G_1(y'_t)\right)^2 + k_2\left(\frac{1}{T}\sum_{t=1}^{T} G_2(y'_t) - \sqrt{\frac{2}{\pi}}\right)^2\right] + log(\hat{\sigma}),$$

with

$$\hat{\sigma} = \hat{\sigma}(\boldsymbol{Y}_{1:T}) = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(y_t)^2},$$

$$y'_t = \frac{y_t}{\hat{\sigma}}, (t = 1, \ldots, T)$$

$$G_1(z) = ze^{\frac{-z^2}{2}},$$

$$G_2(z) = |z|,$$

$$k_1 = \frac{36}{8\sqrt{3} - 9},$$

$$k_2 = \frac{1}{2 - \frac{6}{\pi}},$$

11. **SH_MAXENT2**: Maximum entropy distribution-based Shannon entropy estimator, same as SH_MAXENT1 with the following changes:

$$G_2(z) = e^{\frac{-z^2}{2}},$$

$$k_2 = \frac{24}{16\sqrt{3} - 27},$$

12. **SH_SPACING_V**: Shannon entropy estimator using Vasicek's spacing method.

$$H(\boldsymbol{Y}_{1:T}) = \frac{1}{T}\sum_{t=1}^{T} log\left(\frac{T}{2m}[y_{(t+m)} - y_{(t-m)}]\right)$$

with T number of samples, the convention that $y_{(t)} := y_{(1)}$ if $t < 1$ and $y_{(t)} := y_{(T)}$ if $t > T$ and $m = \lfloor\sqrt{T}\rfloor$.

Bibliography

[1] Chen, W., Drton, M., Wang, Y.S.: On causal discovery with an equal-variance assumption. Biometrika **106**(4), 973–980 (2019)

[2] Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., Schoelkopf, B.: Inferring deterministic causal relations (2012), http://arxiv.org/abs/1203.3475

[3] Friedman, N., Nachman, I.: Gaussian process networks. CoRR **abs/1301.3857** (2013), http://arxiv.org/abs/1301.3857

[4] Hoyer, P., Janzing, D., Mooij, J.M., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. Advances in neural information processing systems **21**, 689–696 (2009)

[5] Hyvärinen, A., Smith, S.M.: Pairwise likelihood ratios for estimation of non-gaussian structural equation models. Journal of Machine Learning Research **14**(Jan), 111–152 (2013)

[6] Janzing, D., Hoyer, P.O., Schoelkopf, B.: Telling cause from effect based on high-dimensional observations (2009), http://arxiv.org/abs/0909.4386

[7] Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., Schölkopf, B.: Information-geometric approach to inferring causal directions. Artificial Intelligence **182**, 1–31 (2012)

[8] Judea, P.: Causality: models, reasoning, and inference. Cambridge University Press. ISBN 0 **521**(77362), 8 (2000)

[9] Kano, Y., Shimizu, S.: Causal inference using nonnormality. In: Proceedings of the international symposium on science of modeling, the 30th anniversary of the information criterion. pp. 261–270 (2003)

[10] Kap, B.: The effect of noise level on causal identification with additive noise models (2021), https://arxiv.org/abs/2108.11320

[11] Kohavi, R., Longbotham, R.: Online controlled experiments and a/b testing. Encyclopedia of machine learning and data mining **7**(8), 922–929 (2017)

[12] Kpotufe, S., Sgouritsa, E., Janzing, D., Schölkopf, B.: Consistency of causal inference under the additive noise model. In: International Conference on Machine Learning. pp. 478–486. PMLR (2014)

[13] Mooij, J., Janzing, D., Peters, J., Schölkopf, B.: Regression by dependence minimization and its application to causal inference in additive noise models. In: Proceedings of the 26th annual international conference on machine learning. pp. 745–752 (2009)

[14] Mooij, J.M., Janzing, D., Heskes, T., Schölkopf, B.: On causal discovery with cyclic additive noise models. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. pp. 639–647 (2011)

[15] Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research **17**(1), 1103–1204 (2016)

[16] Nowzohour, C., Bühlmann, P.: Score-based causal learning in additive noise models. Statistics **50**(3), 471–485 (2016)

[17] Park, G.: Identifiability of additive noise models using conditional variances. Journal of Machine Learning Research **21**(75), 1–34 (2020)
[18] Park, G., Kim, Y.: Identifiability of gaussian structural equation models with homogeneous and heterogeneous error variances (2019), http://arxiv.org/abs/1901.10134
[19] Peters, J., Bühlmann, P.: Identifiability of Gaussian structural equation models with equal error variances. Biometrika **101**(1), 219–228 (11 2013). https://doi.org/10.1093/biomet/ast043, https://doi.org/10.1093/biomet/ast043
[20] Peters, J., Mooij, J., Janzing, D., Schölkopf, B.: Causal discovery with continuous additive noise models. Journal of Machine Learning Research **15**(1), 2009–2053 (2014)
[21] Rebane, G., Pearl, J.: The recovery of causal poly-trees from statistical data. CoRR **abs/1304.2736** (2013), http://arxiv.org/abs/1304.2736
[22] Sgouritsa, E., Janzing, D., Hennig, P., Schölkopf, B.: Inference of cause and effect with unsupervised inverse regression. In: Artificial intelligence and statistics. pp. 847–855. PMLR (2015)
[23] Shimizu, S.: Lingam: Non-gaussian methods for estimating causal structures. Behaviormetrika **41**(1), 65–98 (2014)
[24] Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A., Jordan, M.: A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research **7**(10) (2006)
[25] Shimizu, S., Hyvarinen, A., Kawahara, Y.: A direct method for estimating a causal ordering in a linear non-gaussian acyclic model (2014), http://arxiv.org/abs/1408.2038
[26] Silva, R., Scheines, R., Glymour, C., Spirtes, P., Chickering, D.M.: Learning the structure of linear latent variable models. Journal of Machine Learning Research **7**(2) (2006)
[27] Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, vol. 81. Springer Science & Business Media (2012)
[28] Stegle, O., Janzing, D., Zhang, K., Mooij, J.M., Schölkopf, B.: Probabilistic latent variable models for distinguishing between cause and effect. Advances in neural information processing systems **23**, 1687–1695 (2010)
[29] Sun, X., Janzing, D., Schölkopf, B.: Causal inference by choosing graphs with most plausible markov kernels. In: Ninth International Symposium on Artificial Intelligence and Mathematics (AIMath 2006). pp. 1–11 (2006)
[30] Sun, X., Janzing, D., Schölkopf, B.: Causal reasoning by evaluating the complexity of conditional densities with kernel methods. Neurocomputing **71**(7-9), 1248–1256 (2008)
[31] Szabó, Z.: Information theoretical estimators toolbox. Journal of Machine Learning Research **15**, 283–287 (2014)
[32] Thase, M.E., Parikh, S.V., Rothschild, A.J., Dunlop, B.W., DeBattista, C., Conway, C.R., Forester, B.P., Mondimore, F.M., Shelton, R.C., Macaluso, M.: Impact of pharmacogenomics on clinical outcomes for patients taking medications with gene-drug interactions in a randomized controlled trial. The Journal of clinical psychiatry **80**(6), 0–0 (2019)

[33] Wright, S.: Correlation and causation. Journal of Agricultural Research **20**, 557–580 (1921)
[34] Young, S.W.: Improving library user experience with a/b testing: Principles and process. Weave: Journal of Library User Experience **1**(1) (2014)
[35] Zhang, K., Hyvarinen, A.: On the identifiability of the post-nonlinear causal model (2012), http://arxiv.org/abs/1205.2599