Causal Identification with Additive Noise Models: Quantifying the Effect of Noise

Benjamin Kap, Marharyta Aleksandrova, Thomas Engel
University of Luxembourg, 2 avenue de l'Université, L-4365 Esch-sur-Alzette, Luxembourg
{benjamin.kap.001, marharyta.aleksandrova, thomas.engel}@uni.lu

Abstract

In recent years, a lot of research has been conducted within the area of causal inference and causal learning. Many methods have been developed to identify the cause-effect pairs in models and have been successfully applied to observational real-world data to determine the direction of causal relationships. Yet in bivariate situations, causal discovery problems remain challenging. One class of such 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 test *Regression with Subsequent Independence Test* using an exhaustive range of models where the level of additive noise gradually changes from 1% to 10000% of the causes' noise level (the latter remains fixed). Additionally, the experiments in this work consider several different types of distributions as well as linear and non-linear models. The results of the experiments show that ANMs methods can fail to capture the true causal direction for some levels of noise.

1 Introduction & Related Work

Causal identification is the procedure of determining causal relationship direction from observational data only and representing these as a (causal) graph. This problem is closely related to *structure learning of Bayesian networks*, as static causal graphs are often represented and studied as Bayesian networks.

The basic idea of structure learning emerged from Wright (1921) as *path analysis*. In his work, Wright made a distinction between three possible types of causal substructures that were allowed in a directed acyclic graph:

$$X \to Y \to Z, \text{ or } X \leftarrow Y \to Z, \text{ or } X \to Y \leftarrow Z.$$

Later, Rebane and Pearl (1987) developed an algorithm to recover directed acyclic graphs from statistical data, which relied on the distinction of these substructures. Spirtes et al. (2000) used Bayes networks to axiomatize the connection between causal structure and probabilistic independence and formalized under what assumptions one could draw causal knowledge from observational data only. Furthermore, they also formalized how incomplete causal knowledge could be used for causal intervention. Judea Pearl presented in his work (Judea 2000) a comprehensive theory of causality and unified the probabilistic, manipulative, counterfactual, and structural approaches to causation. From the work of Judea (2000) we have the following key point: if there is a statistical association, e.g. two variables X, Y are dependent, then one of the following is true:

1. there is a causal relationship, either X affects Y or Y affects X;
2. there is a common cause (*confounder*) that affects both X and Y;
3. there is a possibly unobserved common effect of X and Y that is conditioned upon data acquisition (*selection bias*);
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 identifying causal relationships from observational data.

In general, all these methods exploit the complexity of the marginal and conditional probability distributions in some way (e.g., Janzing et al. (2012); Sgouritsa et al. (2015)), and under certain assumptions these methods are then able to solve the task of causal identification. Let C denote the cause and E the effect. In a system with two or more variables we might have cause-effect pairs and then their joint density can be expressed with $p_{C,E}(c, e)$. This joint density can be factorized in either of the following ways:

$$p_{C,E}(c, e) = p_C(c) \cdot P_{E|C}(e|c), \text{ or} \tag{1}$$

$$p_{C,E}(c, e) = p_E(e) \cdot P_{C|E}(c|e). \tag{2}$$

The idea is then that Eq. (1) gives models of lower total complexity than Eq. (2), and this allows us to draw conclusions about 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, Eq. (2) has at least as much complexity as Eq. (1). This unequal distribution of complexity is often colloquially referred to as "*breaking the symmetry*", that is $p_C(c) \cdot P_{E|C}(e|c) \neq p_E(e) \cdot P_{C|E}(c|e)$.

In recent years, numerous approaches were proposed for structure learning. Friedman and Nachman (2000) addressed the problem of learning the structure of a Bayesian network in domains that contain continuous variables. Kano and Shimizu (2003) developed a model for causal inference using non-normality of observed data and improved path analysis proposed by Wright (1921). Shimizu et al. (2006) proposed a method to determine the complete causal graph of continuous data under three assumptions: the data generating process is linear, no unobserved confounders, and noise variables have non-Gaussian distributions of non-zero variances. This method was not scale-invariant, but later work by Shimizu et al. (2009) addressed this problem. Sun, Janzing, and Schölkopf (2006) introduced a method based on comparing the conditional distributions of variables given their direct causes for all hypothetical causal directions and choosing the most plausible one (Markov kernels). Sun, Janzing, and Schölkopf (2008) continued the work on kernels by using the concept of reproducing kernel Hilbert spaces.

A group of well-known and well-established methods is based on the *Additive Noise Models* (ANMs), that yield many good results (Kpotufe et al. 2014). In these models, the effect is a function of the cause and some random and non-observed additive noise term. These methods received a lot of attention from researchers in the past years. Hoyer et al. (2009) generalized the linear framework of additive noise models to nonlinear models. Mooij et al. (2009) introduced a method that minimizes the statistical dependence between the regressors and residuals. 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 is well suited for the task of causal inference in additive noise models. Mooij et al. (2011) introduced a method to determine the causal relationship in cyclic additive noise models and stated that such models are generally identifiable in the bivariate, Gaussian-noise case. Their method works for continuous data and can be seen as a special case of nonlinear independent component analysis. Hyvärinen and Smith (2013) proposed a method which is based on the likelihood ratio under the linear non-Gaussian acyclic model known as LiNGAM (Shimizu 2014). This method does not resort to independent component analysis algorithm as previous methods did.

As indicated in the name, ANMs are heavily based on the presence of noise. However, despite all the research in the past years one small but nonetheless important aspect of causal discovery methods with ANMs has not received much attention: *can different noise levels have an impact on the correctness of these methods?* In the real world, observational data often differ in terms of the noise level. Usually, these levels do not differ significantly but it can occur that noise levels change drastically from cause to effect. For example, if the data collection process has a lot of interference (e.g., in outer space) then the related noise levels can differ a lot. In this work, we aim to bridge this research gap. We perform an empirical study with a well-established method of the ANMs group *Regression with Subsequent Independence Test (RESIT)* (Peters et al. 2014). This method yields good results and can be used even when variables have different distribution types. In our experimental evaluation, we aim to quantify the impact of different noise levels on the performance of RESIT.

The rest of the paper is organized as follows. In Section 2 we describe RESIT and discuss its functioning. Section 3 presents experimental setup followed by results analysis in Section 4. Finally, we conclude our work and summarize our findings in Section 5.

2 Model

2.1 RESIT

The RESIT method is based on the fact that for each node X_i the corresponding noise variable N_i is independent of all non-descendants of X_i. For example, if we have $Y = X_1 + N_1$ then $X_1 \perp\!\!\!\perp N_1$. RESIT works in both bivariate and multivariate cases, see Peters et al. (2014). We restrict our experiments to bivariate cases to reduce runtimes. In our experiments, we have two variables, X and Y, and the task is to determine whether X causes Y ($X \to Y$) or Y causes X ($Y \to X$).

We apply the same algorithm as Algorithm 1 from Mooij et al. (2016) which requires inputs X and Y, a regression method, and a score estimator $\hat{C} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$. The algorithm outputs dir (casual relationship **dir**ection). First, the data is split into training data and test data. Kpotufe et al. (2014) refers to this as *decoupled estimation*[1]. The training data is used to fit the regression model and the test data is used to calculate the value of the estimator. The idea is to regress Y on X with the training data, predict $\hat{Y}$ with the test data and then calculate residuals $Y_{res} = \hat{Y} - Y_{Test}$. Y_{res} and X_{Test} are then used to calculate the score for the assumed case $X \to Y$: $\hat{C}_{X\to Y}$. Similarly, to test the other case ($Y \to X$), we regress X on Y, calculate residuals $X_{res} = \hat{X} - X_{Test}$ and estimate $\hat{C}_{Y\to X}$. If only one direction in our data is correct (and not both), we can compare estimates directly. Otherwise, we need to determine the value of α for the independence tests.

2.2 Estimators

Both *independence tests* and *entropy measures* can be used to calculate the scores $\hat{C}_{X\to Y}$ and $\hat{C}_{Y\to X}$. 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 estimator score for entropy is derived from Lemma 1 from Kpotufe et al. (2014).

[1] As opposed to *decoupled estimation*, in *coupled estimation* the data is not split into training and test data, see Kpotufe et al. (2014); Mooij et al. (2016).

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

Input:
  1) I.i.d. sample data X and Y
  2) Regression method
  3) Score estimator $\hat{C} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$
Output:
  dir

Procedure
1) Split data into training and test data:
  $X_{Train}, X_{Test} \leftarrow X$
  $Y_{Train}, Y_{Test} \leftarrow Y$

2) Train regression models
  $reg_1 \leftarrow$ Regress Y_{Train} on X_{Train}
  $reg_2 \leftarrow$ Regress X_{Train} on Y_{Train}

3) Calculate Residuals:
  $Y_{res} = reg_1.predict(X_{Test}) - Y_{Test}$
  $X_{res} = reg_2.predict(Y_{Test}) - X_{Test}$

4) Calculate Scores:
  $\hat{C}_{X \to Y} = \hat{C}(X_{Test}, Y_{res})$
  $\hat{C}_{Y \to X} = \hat{C}(Y_{Test}, X_{res})$

5) Output direction *dir*:

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

The following 6 independence tests and 6 entropy measures were used as estimators in this work. The implementation of all estimators except **HSIC** was taken from the *information theoretical estimators* toolbox (Szabó 2014):[2]

1. **HSIC**: Hilbert-Schmidt Independence Criterion with RBF Kernel[3]:

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

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) with $\eta = 1 * 10^{-6}$ 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:

$$I_{dCor}(x, y) = \frac{I_{dCov}(x, y)}{\sqrt{I_{dVar}(x, x) I_{dVar}(y, y)}}$$

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 DISTCOV). If $I_{dVar}(x, x) I_{dVar}(y, y) \leq 0$, then $I_{dCor}(x, y) = 0$.

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.

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**: Shannon differential entropy estimator using kNNs with $k = 3$ and kd-tree for quick nearest-neighbour lookup.

9. **SH_KNN_3**: Shannon differential entropy estimator using kNNs 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**: Same as SH_MAXENT1 with the following changes:

$$G_2(z) = e^{\frac{-z^2}{2}}, k_2 = \frac{24}{16\sqrt{3} - 27}.$$

[2] See the documentation of the toolbox for more details.
[3] Source: https://github.com/amber0309/HSIC

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

3 Experimental Setup

For all experiments, we generate artificial data using linear and non-linear functions. While both linear and non-linear data can be identifiable in causal models, non-linearity helps in identifying the causal direction as was shown by Hoyer et al. (2009). In all experiments we use the equation $Y = X + N_Y$ for the linear cases and $Y = X^3 + N_Y$ for the non-linear cases. These two structural causal models have been selected arbitrarily for simplicity. For the consistency of the identifiability of linear and non-linear data in additive noise models, the reader is referred to Kpotufe et al. (2014); Shimizu et al. (2006); Hoyer et al. (2009); Zhang and Hyvarinen (2009). 80% of the generated data is used for training a regression model, and the rest 20% is used to calculate the values of estimators $\hat{C}_{X\to Y}$ and $\hat{C}_{Y\to X}$.

In all our tests, we assume X to be a cause of Y, that is $X \to Y$. X and N_Y can be drawn from one of the following distributions: the normal distribution denoted by $\mathcal{N}$, the uniform distribution denoted by $\mathcal{U}$, or the laplace distribution denoted by $\mathcal{L}$. The parameters of the distributions for X and N_Y are defined by the equations below:

$$X \sim \begin{cases} \mathcal{N}(0,1) & \text{or} \\ \mathcal{U}(-1,1) & \text{or} \\ \mathcal{L}(0,1) \end{cases}$$

$$N_Y \sim \begin{cases} \mathcal{N}(0, 1\cdot i) & \text{or} \\ \mathcal{U}(-1\cdot i, 1\cdot i) & \text{or} \\ \mathcal{L}(0, 1\cdot i) \end{cases}$$

with i being a scaling factor for the noise level in N_Y, i-factor for short. By varying the value of i, we can analyze how different values of standard deviations (boundaries for the uniform case) in the noise term N_Y relative to the standard deviation (or boundaries for the uniform case) in the X term impact the accuracy of RESIT method.

In our experiments, we consider 199 different i-factors:

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

The values $i < 1$ correspond to the cases when deviation of N_Y is less than that of X, and the values $i > 1$ correspond to the cases when the deviation of N_Y is larger than that of X. The deviation of noise ranges from 1% (for $i = 0.01$) to 10000% (for $i = 100$) of the deviation of X. For each value of i, we have 18 different combination of models: two general structures $Y = X + N_Y$ and $Y = X^3 + N_Y$ where X and N_Y are drawn from one of the three different distributions: $\mathcal{N}$,$\mathcal{U}$ or $\mathcal{L}$. To represent the models, we use the notations like $Y = L^3 + \mathcal{U}$, that signifies a nonlinear model with $X \sim L$ and $N_Y \sim \mathcal{U}$. For each of the 18 combinations, for a single test, we generate 1000 samples from the relative distributions. Next, we perform causal identification according to the procedure described in Section 2.1 using one of the estimators presented in Section 2.2. These tests are repeated 100 times. Finally, we calculate the fraction of successful tests for each combination of a model and an estimator, and define this ratio as our accuracy measure.

For the regression, we used Linear Regression with an appropriate coordinates transformation for the non-linear cases.

4 Experimental Results

Figs. 1 and 2 show the results for different estimators obtained for liner and nonlinear models respectively. In these figures, the y-axis shows the accuracy ($\frac{\#\text{successful tests}}{100}$) for different estimators, and the x-axis shows the range of the i-factor. The results for independence estimators are presented with solid lines and the results for entropy estimators are shown with dashed lines. The values of the estimators close to 0.5 indicate that in 50% of the tests the algorithm chose the correct direction and vice versa 50% chose the wrong one. Such cases are **unidentifiable**. The values of accuracy closer to 1 mean very good or consistent **identifiability**. Also, the plots for DISTCOV (dark green) and DISTCORR (medium purple) often overlap (more than in 90% of cases), resulting in a dark purple line.

Additionally, in Table 1 and Table 2 we summarize our experimental result. The values in the cells show on what range of i-factor the estimators *can* reach over 90%. 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 shows where estimators reach the first time 90% or higher, and the upper limit shows the last time where it reaches 90% or higher. In between, most of the time estimators remain above 90% or rarely fall below, but not more than 10% of the cases. An empty cell in the tables means that for the relevant model and estimator the accuracy never reached 90%. An open range from one side, for example, "– 5" or "5 –", or from both sides, such as " – ", indicates an unbounded interval with one or two missing bounds.

4.1 Linear Models: $Y = X + N_Y$

We start with the analysis of the results for the linear models presented in Fig. 1 and Table 1. First, we consider the models with the independent variable distributed normally, $X \sim \mathcal{N}$. Fig. 1a shows the only case where we never achieve identifiability. This is the well-known linear Gaussian structural causal model $Y = N + N$. Only recently it has been tackled successfully by Chen, Drton, and Wang (2019); Park and Kim (2019). However, we do not consider their approach in this work. Fig. 1b shows the linear model with $Y = \mathcal{N} + \mathcal{U}$. SH_SPACING_V performs the best with the accuracy of 100% for $i \in [0.55; 7]$. HSIC_IC and HSIC_IC2 perform the worst here. The associated accuracy reaches 90% only for $i \in [3; 7]$. All other estimators perform mediocre with an accuracy above 80% for $i \in [0.5; 7]$.

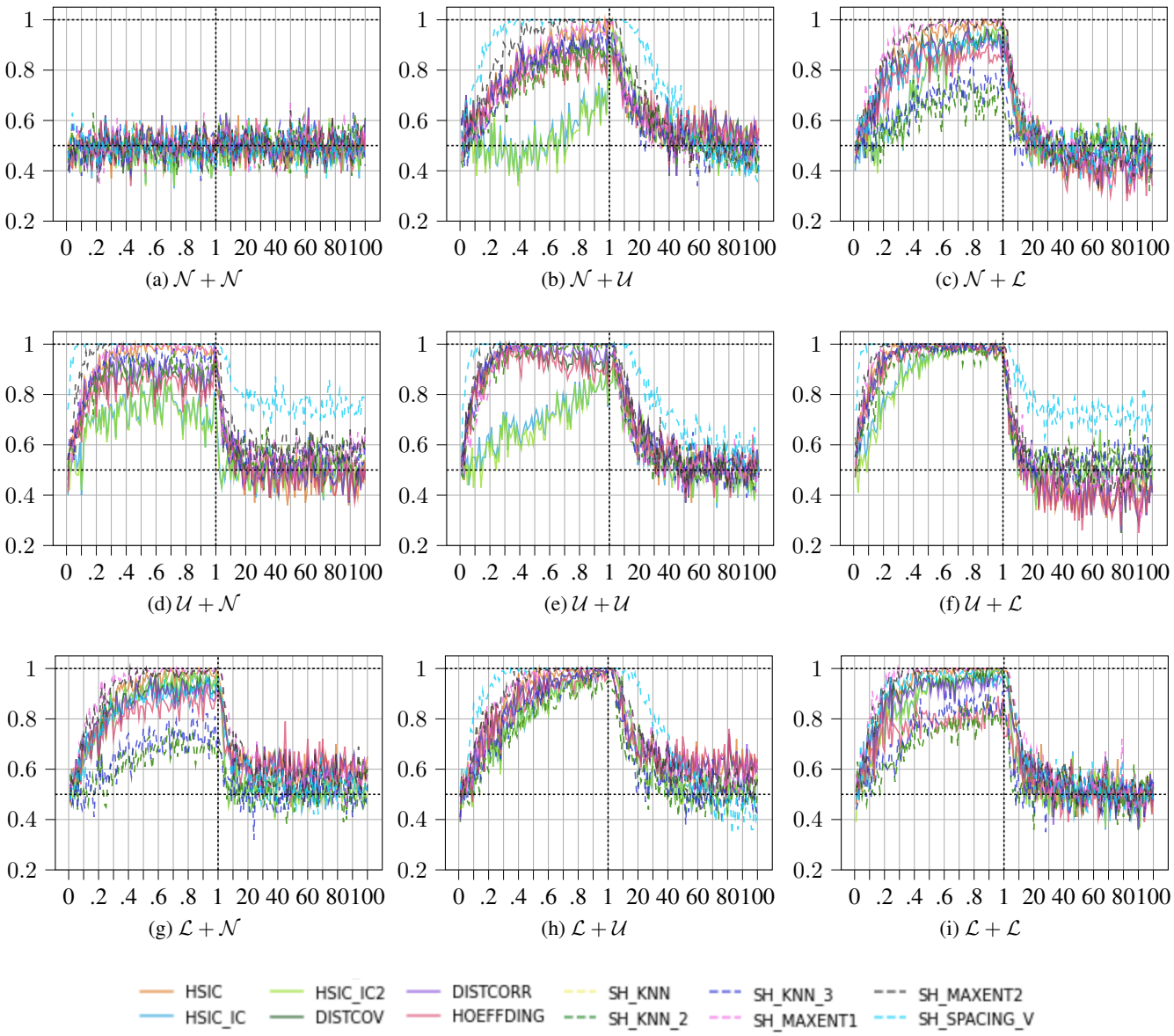

Figure 1: Accuracy of RESIT for linear models as a function of i-factor.

Table 1: Summary for linear models. The numbers reflect the ranges of noise that allow identifiability with accuracy $\approx 90\%$.

Estimator	$\mathcal{N}+\mathcal{N}$	$\mathcal{N}+\mathcal{U}$	$\mathcal{N}+\mathcal{L}$	$\mathcal{U}+\mathcal{N}$	$\mathcal{U}+\mathcal{U}$	$\mathcal{U}+\mathcal{L}$	$\mathcal{L}+\mathcal{N}$	$\mathcal{L}+\mathcal{U}$	$\mathcal{L}+\mathcal{L}$
HSIC		0.60 – 4	0.38 – 2	0.21 – 1	0.18 – 6	0.15 – 1	0.32 – 3	0.40 – 8	0.30 – 5
HISC_IC		3 – 7	0.75 – 2		3	0.35 – 1	0.36 – 3	0.60 – 7	0.40 – 4
HSIC_IC2		3 – 7	0.75 – 2		3	0.35 – 1	0.36 – 3	0.60 – 7	0.40 – 4
DISTCOV		1	0.80 – 1	0.25 – 1	0.20 – 3	0.20 – 1	0.33 – 2	0.40 – 5	0.23 – 4
DISTCORR		1	0.87 – 1	0.25 – 1	0.20 – 3	0.20 – 1	0.33 – 2	0.40 – 5	0.23 – 4
HOEFFDING				1	0.20 – 3	0.15 – 1	0.57 – 1	0.40 – 5	
SH_KNN				0.26 – 1	0.20 – 3	0.21 – 1		0.53 – 4	
SH_KNN_2				0.26 – 1	0.20 – 2	0.21 – 1		0.53 – 4	
SH_KNN_3		0.85 – 4		0.20 – 1	0.20 – 3	0.16 – 1		0.51 – 5	0.60 – 1
SH_MAXENT1		0.65 – 5	0.30 – 3	0.20 – 1	0.23 – 3	0.12 – 3	0.21 – 4	0.32 – 10	0.20 – 6
SH_MAXENT2		0.40 – 8	0.33 – 3	0.10 – 4	0.12 – 8	0.10 – 3	0.21 – 4	0.32 – 10	0.17 – 5
SH_SPACING_V		0.20 – 22	0.82 – 1	0.04 – 9	0.05 – 21	0.03 – 8	0.49 – 4	0.16 – 27	0.17 – 4

Fig. 1c shows the linear model $Y = \mathcal{N} + \mathcal{L}$. The best estimators are SH_MAXENT1 and SH_MAXENT2 with accuracy around 90% for $i \in [0.30; 3]$. HSIC also performs good with accuracy over 90% for $i \in [0.38; 2]$. The worst estimators are the three Shannon differential entropy estimators using kNNs which never remain consistently above 80% accuracy. The remaining estimators lie within the range $90\% \pm 8\%$ accuracy for $i \in [0.3; 3]$.

Next, we consider models with $X \sim U$. Fig. 1d shows the linear model $Y = \mathcal{U} + \mathcal{N}$. Here all estimators differ stronger than in the previous cases. First, SH_SPACING_V performs the best with 100% accuracy for $i \in [0.08; 2]$. With $i = 1$ all other estimators remain above 90%, expect HOEFFDING ($\sim$88%) and HSIC_IC and HSIC_IC2 (both $\sim 75\%$). After i becomes larger than 2, all estimators drop drastically towards 50% accuracy except SH_SPACING_V which remains above 70%. For $i \in [0.2; 1]$ some estimators remain between 80% and 95% while HSIC is above 95%. HSIC_IC and HSIC_IC2 perform worse than all other estimators. Fig. 1e, shows the linear model $Y = \mathcal{U} + \mathcal{U}$. In this case, HSIC_IC and HSIC_IC2 reach accuracy above 90% only around $i = 3$, see the 3d column in Table 1. All other estimator perform quite good with $i \in [0.2; 3]$, however HOEFFDING and DISTCOV drop slightly below 90% accuracy for $i = 1$. SH_SPACING_V has 100% accuracy for $i \in [0.12; 10]$ and has on the remaining values of i better accuracy than all other estimators. In general, we can observe that identifiability is much better for $i < 1$, that is when the range for noise term is less than the range of X. For $i > 50$ the accuracy of most of the estimators is around 50%, indicating that the predicted direction is wrong in half of the tests. Fig. 1f shows the model $Y = \mathcal{U} + \mathcal{L}$. For $i \in [0.3; 1]$ all estimators perform well with 90% or higher accuracy, except HSIC_IC and HSIC_IC2 which remain above 90% accuracy only after $i = 0.45$. After $i = 1$ each estimator drops drastically and all converge towards 50% accuracy. The only exception is SH_SPACING_V which remains with a mean of 70% accuracy longer than other estimators. For $i \in [0.08; 1]$ SH_SPACING_V also has accuracy 100%. For $i < 0.3$ all other estimators drop fast towards 50%.

Now we proceed to the analysis of the remaining cases where the independent variable X is distributed according to the Laplace distribution, $X \sim \mathcal{L}$. Fig. 1g shows the model $Y = \mathcal{L} + \mathcal{N}$. For $i \in [0.3; 1]$ HSIC, SH_MAXENT1 and SH_MAXENT2 have accuracy greater than 90%. SH_SPACING_V, HSIC_IC, HSIC_IC2, DISTCOV and DISTCORR lie between 85% and 95% accuracy for $i \in [0.4; 1]$ and HOEFFDING remains between 80% and 90%. Again, the three Shannon kNN estimators never reach an accuracy higher than 80%. After i reaches the value of 1, all estimators drop fairly fast towards unidentifiability. Fig. 1h shows the linear model $Y = \mathcal{L} + \mathcal{U}$. SH_SPACING_V performs the best of all estimators and has an accuracy of 100% for $i \in [0.5; 5]$. All other estimators slowly climb towards good identifiability and for $i \in [0.7; 7]$ they remain above 90% accuracy. Afterward, all other estimators drop with a similar pace towards unidentifiability. Fianlly, Fig. 1i shows the linear model $Y = \mathcal{L} + \mathcal{L}$. Here we can observe the following. For $i \in [0.4; 2]$ SH_MAXENT1 and SH_MAXENT2 have accuracy close to 100%. Next, for $i \in [0.4; 1]$ HSIC, HSIC_IC, HSIC_IC2, SH_SPACING_V, DISTCOV and DISTCORR remain above 90% accuracy. HOEFFDING, and the three Shannon kNN estimators never reach an accuracy above 90%. After $i = 1$ all estimators drop fast towards 50%.

4.2 Nonlinear Models: $Y = X^3 + N_Y$

The results for nonlinear models are grouped in the same way as for linear models and are presented in Fig. 2 and Table 2. In general, we can notice much better identifiability in the nonlinear case.

Similar to the linear case, We start with the analysis of the modes with $X \sim \mathcal{N}$. Fig. 2a shows the nonlinear model $Y = \mathcal{N}^3 + \mathcal{N}$. Here all estimators perform very good with $i \in [0.4; 25]$ having an accuracy of almost 100%. With $i < 0.3$ most estimators drop fast below 90% accuracy. With $i \in [20; 100]$ all estimators remain above 90% accuracy, except for HSIC_IC, HSIC_IC2, SH_MAXENT1 and SH_MAXENT2 which drop below 90% after $i = 45$. DISTCOV, SH_SPACING_V and the three Shannon kNN estimators remain close to 100% in $i \in [0.01; 100]$. Fig. 2b shows the model $Y = \mathcal{N}^3 + \mathcal{U}$. In this case, for $i \in [0.45; 80]$ we have 90% or higher accuracy for all estimators. DISTCOV, SH_SPACING_V and the three Shannon kNN estimators remain close to 100% in $i \in [0.01; 100]$. Fig. 2c shows the nonlinear model $Y = \mathcal{N}^3 + \mathcal{L}$. For this model, all estimators perform very good with $i \in [0.4; 30]$ having an accuracy close to 100%. With $i < 0.25$ estimator drop rapidly and for $i > 30$ HSIC_IC, HSIC_IC2, SH_MAXENT1 and SH_MAXENT2 drop below 90% accuracy. All others remain over 90% accuracy while HSIC remains around 90%.

Now we proceed to the analysis of the models with $X \sim \mathcal{U}$. Fig. 2d shows the model $Y = \mathcal{U}^3 + \mathcal{N}$. All estimators, except HSIC_IC and HSIC_IC2, remain above 95% for $i \in [0.05; 1]$. For $i \in [1; 100]$ the estimators converge differently. All three Shannon differential entropy measures with kNNs and SH_SPACING_V remain above 95% accuracy. DISTCOV, DISCORR and HOEFFDING keep a mean of $\sim$85% accuracy. HSIC and SH_MAXENT1 remain above 60% accuracy. SH_MAXENT2 is pretty much unidentifiable. Finally, HSIC_IC and HSIC_IC2 are unidentifiable for all values of i. Fig. 2e shows the nonlinear model $Y = \mathcal{U}^3 + \mathcal{U}$. For $i \in [0.09; 1]$ all estimators except HSIC_IC and HSIC_IC2 remain above 95% accuracy, while SH_KNN, SH_KNN_2, and SH_SPACING_V continue to do so for $i \in [1; 100]$. DISTCOV, DISCORR and HOEFFDING remain between 80% and 90%. HSIC and SH_MAXENT1 drop to $\approx$ 60% after $i = 20$ and remain above 60% for $i < 100$. SH_MAXENT2, HSIC_IC and HSIC_IC2 drop to 50% for $i \in [20; 100]$. Fig. 2f shows the model $Y = \mathcal{U}^3 + \mathcal{L}$. The behaviour of different estimators is almost the same as for $Y = \mathcal{U}^3 + \mathcal{N}$. The only differences are that HSIC_IC performs slightly better for $i \in [0.2; 1]$ and DISTCOV, DISTCORR, HSIC and SH_MAXENT2 perform worse.

Lastly, we analyze the 3 remaining models with $X \sim L$. Fig. 2g shows the model $Y = \mathcal{L}^3 + \mathcal{N}$. For $i \in [0.1; 100]$ all estimators (except SH_MAXENT1 and SH_MAXENT2) have an accuracy of 90% or higher, SH_SPACING_V and

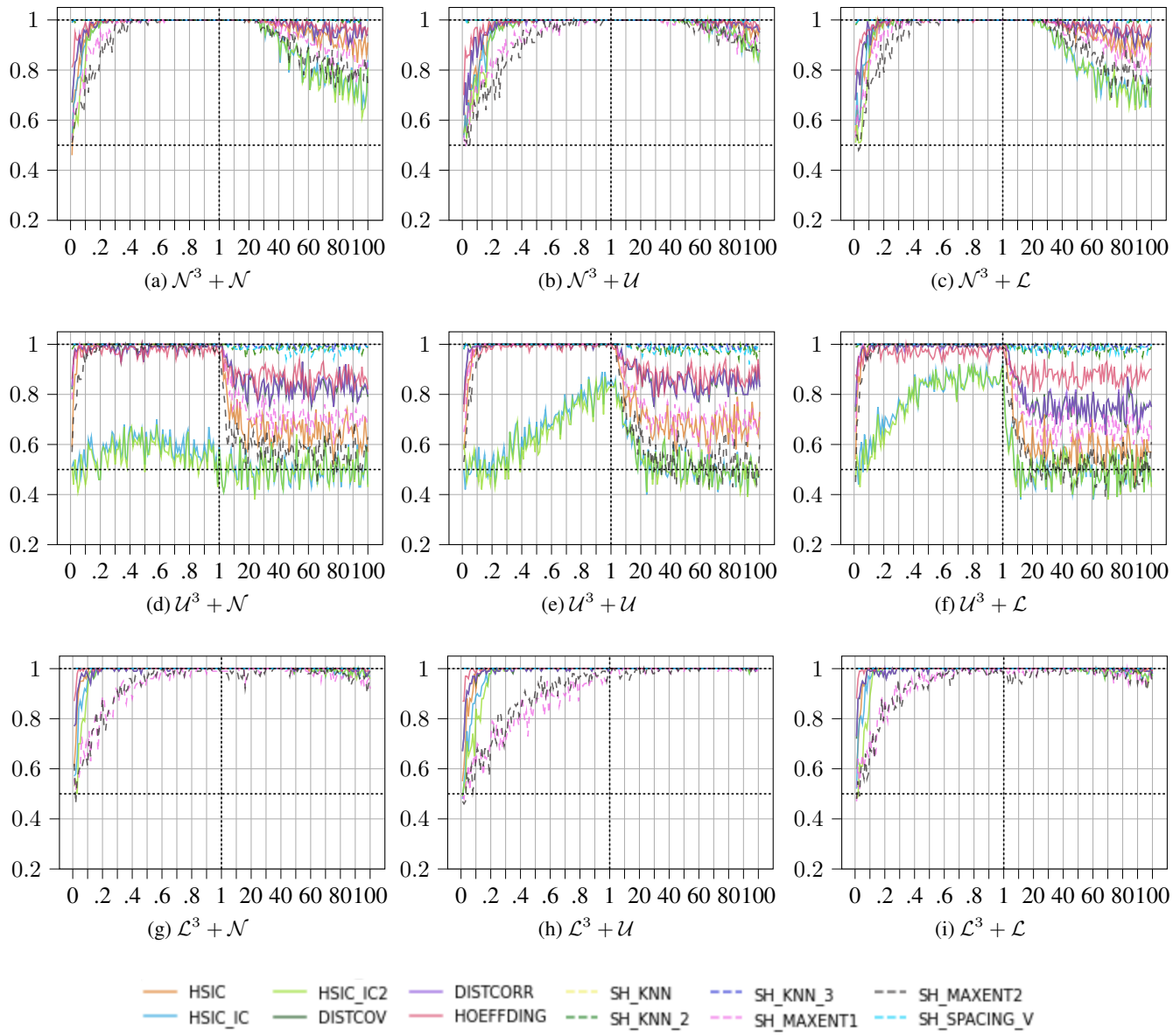

Figure 2: Accuracy of RESIT for nonlinear models as a function of i-factor.

Table 2: Summary for nonlinear models. The numbers reflect the ranges of noise that allow identifiability with accuracy $\approx 90\%$.

Estimator	$\mathcal{N}^3 + \mathcal{N}$	$\mathcal{N}^3 + \mathcal{U}$	$\mathcal{N}^3 + \mathcal{L}$	$\mathcal{U}^3 + \mathcal{N}$	$\mathcal{U}^3 + \mathcal{U}$	$\mathcal{U}^3 + \mathcal{L}$	$\mathcal{L}^3 + \mathcal{N}$	$\mathcal{L}^3 + \mathcal{U}$	$\mathcal{L}^3 + \mathcal{L}$
HSIC	0.09 – 68	0.16 –	0.11 – 70	0.04 – 3	0.05 – 6	0.05 – 3	0.04 –	0.07 –	0.05 –
HISC_IC	0.10 – 40	0.16 – 88	0.10 – 35			0.70 – 1	0.07 –	0.12 –	0.06 –
HSIC_IC2	0.10 – 40	0.17 – 88	0.10 – 35			0.70 – 1	0.10 –	0.15 –	0.10 –
DISTCOV	0.07 –	0.14 –	0.08 – 88	–85	0.03 – 20	0.02 – 5	0.03 –	0.04 –	0.03 –
DISTCORR	0.07 –	0.14 –	0.05 – 88	–85	0.03 – 20	0.02 – 5	0.03 –	0.04 –	0.03 –
HOEFFDING	0.02 –	0.05 –	0.04 –	–95	0.03 –	0.02 –	0.02 –	0.02 –	–
SH_KNN	–	–	–	–	–	–	–	–	–
SH_KNN_2	–	–	–	–	–	–	–	–	–
SH_KNN_3	–	–	–	–	–	–	–	–	–
SH_MAXENT1	0.20 – 45	0.33 – 85	0.18 – 67	0.02 – 4	0.05 – 7	0.03 – 4	0.28 –	0.53 –	0.26 –
SH_MAXENT2	0.28 – 45	0.43 – 91	0.25 – 65	0.08 – 2	0.09 – 5	0.06 – 3	0.30 –	0.43 –	0.28 –
SH_SPACING_V	–	–	–	–	–	–	–	–	–

the three Shannon kNN estimators have an accuracy of 100% for all values of i-factor. Only SH_MAXENT1 and SH_MAXENT2 perform badly at the beginning but still have an accuracy of 90% or higher for $i \in [0.35; 100]$. Fig. 2h shows the nonlinear model $Y = \mathcal{L}^3 + \mathcal{U}$. It is very similar to the previous case. For $i \in [0.15; 100]$ all estimators (except SH_MAXENT1 and SH_MAXENT2) have an accuracy of 90% or higher, and SH_SPACING_V, and the three Shannon kNN estimators have an accuracy of 100% for all values of i. As in the previous case, SH_MAXENT1 and SH_MAXENT2 perform badly at the beginning but still have an accuracy of 90% or higher for $i \in [0.7; 100]$. For $i \geq 1$ all estimators are very close to 100% accuracy. Finally, Fig. 2i shows the nonlinear model $Y = \mathcal{L}^3 + \mathcal{L}$. This model allows the best identifiability of all. For $i \geq 0.1$ all estimators except SH_MAXENT1 and SH_MAXENT2 have an accuracy 90% or higher. SH_SPACING_V and the three Shannon kNN estimators have an accuracy of 100% for all values of i. Only SH_MAXENT1 and SH_MAXENT2 perform badly at the beginning but still have an accuracy of 90% or higher for $i \geq 0.35$.

4.3 Summary

As the results show, different noise levels do have an impact on the identifiability performance in RESIT methods. In general, the linear equation models are more fragile in RESIT than the nonlinear equation models because nonlinear relationships tend to break the symmetry between the variables easier (Hoyer et al. 2009). Furthermore, in all cases the test results themselves have a standard deviation between 0.05 to 0.1 as one can see in the sharp wiggles in the plots.

We can notice some similarities in the models depending on how distributed their components X and N_Y. It is visually visible in a matrix of plots in Figs. 1 and 2. The plots on the main diagonals, Figs. 1a, 1e, 1i, 2a, 2e and 2i, represent the models for which both X and N_Y are drawn from the same type of distribution. In the case of nonlinear models, the diagonal plots demonstrate 3 distinct behaviors of estimators, as presented in Figs. 2a, 2e and 2i. We can also clearly see the similarity of plots in the same rows. It means that the models with the same type of distribution for the independent variable X have common characteristics. We observe that all models with $X \sim \mathcal{L}$, see Figs. 2g to 2i allow very good and consistent identifiability by all estimators for $i \geq 1$. For the values $i < 1$, many estimators, except SH_MAXENT1 and SH_MAXENT2, also perform well with accuracy $\approx 90\%$. The group of models with $X \sim \mathcal{N}$ allow all estimators achieve almost perfect identifiability for $0.8 < i < 20$. The accuracy then reduces for larger and smaller values of i. Finally, the group of models with $X \sim N$ allow the worst identifiability, see Figs. 2d to 2f. For $i > 20$ several estimators have accuracy of $50\% - 60\%$. However, SH_KNN estimators allow consistent identifiability for all values of i even in this case, see Table 2. Similar but much less prominent row-wise similarity can be observed for linear models as well, see Fig. 1. This indicates that the type of distribution of the independent variable X impacts the accuracy of different estimators.

Looking now only at the best estimation function and assuming a strong identifiability of $\geq 90\%$ accuracy, we can observe that for linear models and $i \notin [0.5; 5]$ the accuracy is usually below 90%. This looks different for the nonlinear cases. Such estimators as SH_KNN, SH_KNN_2, SH_KNN_3, and SH_SPACING_V allow consistent identifiability for all nonlinear modes. At the same time, the models with $X \sim \mathcal{L}$ are identifiable with accuracy $\geq 90\%$ by all estimators on almost all range of values of i, see Table 2.

Some estimators perform differently depending on the setup. For example, for all nonlinear cases, the three Shannon differential entropy estimators with kNNs always perform above 90% accuracy for all values of i, see Table 2. The associated accuracy even reaches 100% for all i in the case of nonlinear models with $X \sim \mathcal{L}$, see Figs. 2g to 2i. In case of linear models, these estimatros perform relatively poor, sometimes never reaching 90% accuracy, see linear models with $X \sim N$, $Y = \mathcal{L} + \mathcal{N}$, and $Y = \mathcal{L} + \mathcal{L}$ in Table 1.

Overall, SH_SPACING_V performs the best in almost all cases, and is only outperformed by SH_MAXENT1 and SH_MAXENT2 for the following three linear models: $Y = \mathcal{N} + \mathcal{L}$, $Y = \mathcal{L} + \mathcal{L}$, and $Y = \mathcal{L} + \mathcal{N}$, see Figs. 1c, 1g and 1i and Table 1. Some independence tests lose some of the accuracy while entropy estimators retain accuracy over 90%. This is observed for nonlinear models with $X \sim \mathcal{U}$, see Figs. 2d to 2f and Table 2. Additionally, it is worth mentioning that entropy estimators are less computationally demanding than independence tests but can be quite sensitive to discretization effects (Mooij et al. 2016). However, entropy estimators can only be used with the prior assumption we made: *there is only one causal direction and it is present in the model*.

5 Conclusions

In this paper, we study the performance of a well-known causal discovery method RESIT which falls in a group of additive noise models. While RESIT was widely studied in the literature before, previous research paid little attention to the effect of noise level on the accuracy of this approach. This work aims to fill this gap by means of an empirical study. In our experiments, we tested a liner model $Y = X + N_Y$ and a nonlinear model $Y = X^3 + N_Y$ with X and N_Y being drawn from one of the following distributions: Normal $\mathcal{N}$, Uniform $\mathcal{U}$ or Laplace $\mathcal{L}$. We also used 12 different estimators (6 independence estimators and 6 entropy estimators). The results from our experiments show that the effect of noise is not negligible and can impact the model's identifiability. For significantly small noise levels in the disturbance term N_Y or significantly large noise levels, this causal discovery method fails to capture the true causal relationship of the given structural equation model. *Significantly* here depends on the model. For example, on some models if the noise level is already twice larger than the variation of the independent variable, then the model becomes unidentifiable. Other models remained identifiable with 100 times larger noise levels, see Section 4 for details.

The range of different noise levels in our experiments is quite exhaustive, changing from 100 times less to 100 times larger than the variance of the causal variable X. Some of

these cases can be very rare in practice, however, the discovered relationships can be useful for the practitioners and researchers. In general, if the standard deviation of the noise term is smaller than the standard deviation of the cause, then models remained identifiable more often as opposed to the case when the standard deviation 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.

Our results also show differences in terms of the performance of the analyzed estimators. In our experiments, Hilbert-Schmidt Independence Criterion with RBF Kernel is the best independence estimator, and Shannon entropy with Vasicek's spacing method is the best entropy estimator. Comparing the performance on linear and non-linear models, our results show that non-linear models are still identifiable in situations where linear models are not. For example, some non-linear models with the noise term's standard deviation of 100 times higher than that of the cause, are perfectly identifiable while their linear counterparts are not. Finally, our experiments show different behavior for different distribution types (e.g., Gaussian, Uniform. or Laplace). Generally, models with the causal variable drawn from Laplace distribution $X \sim \mathcal{L}$ allow better identifiability.

In our experiments, we tested only two particular models and three different distribution types. Similar results are expected with other methods for causal discovery with additive noise models, as the failing point is the independence estimation (or entropy estimation). Therefore, methods relying on these estimations are generally prone to errors for some levels of noise. This work also does not formalize the effect of different noise levels in ANM causal discovery methods but it could be done in 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 properties exhibited by an observed distribution.

6 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.

References

Chen, W.; Drton, M.; and Wang, Y. S. 2019. On causal discovery with an equal-variance assumption. *Biometrika* 106(4): 973–980. URL https://arxiv.org/abs/1807.03419.

Friedman, N.; and Nachman, I. 2000. Gaussian Process Networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI2000)*, 211–219. URL https://arxiv.org/abs/1301.3857.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems* 21: 689–696. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.144.4921&rep=rep1&type=pdf.

Hyvärinen, A.; and Smith, S. M. 2013. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research* 14(Jan): 111–152. URL https://www.jmlr.org/papers/volume14/hyvarinen13a/hyvarinen13a.pdf.

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

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

Kano, Y.; and Shimizu, S. 2003. Causal inference using nonnormality. In *Proceedings of the international symposium on science of modeling, the 30th anniversary of the information criterion*, 261–270. URL http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/papers/aic30_web2.pdf.

Kpotufe, S.; Sgouritsa, E.; Janzing, D.; and Schölkopf, B. 2014. Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning*, 478–486. PMLR. URL http://proceedings.mlr.press/v32/kpotufe14.html.

Mooij, J.; Janzing, D.; Peters, J.; and Schölkopf, B. 2009. 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*, 745–752. URL http://www.machinelearning.org/archive/icml2009/papers/279.pdf.

Mooij, J. M.; Janzing, D.; Heskes, T.; and Schölkopf, B. 2011. On causal discovery with cyclic additive noise models. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 639–647. URL https://repository.ubn.ru.nl/bitstream/handle/2066/92140/92140.pdf.

Mooij, J. M.; Peters, J.; Janzing, D.; Zscheischler, J.; and Schölkopf, B. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research* 17(1): 1103–1204. URL https://www.jmlr.org/papers/volume17/14-518/14-518.pdf.

Park, G.; and Kim, Y. 2019. Identifiability of Gaussian Structural Equation Models with Homogeneous and Heterogeneous Error Variances. *arXiv preprint* URL https://arxiv.org/abs/1901.10134.

Peters, J.; Mooij, J.; Janzing, D.; and Schölkopf, B. 2014. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research* 15(1): 2009–2053. URL https://www.jmlr.org/papers/volume15/peters14a/peters14a.pdf.

Rebane, G.; and Pearl, J. 1987. The recovery of causal polytrees from statistical data. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, 222–228. URL https://arxiv.org/abs/1304.2736.

Sgouritsa, E.; Janzing, D.; Hennig, P.; and Schölkopf, B. 2015. Inference of cause and effect with unsupervised inverse regression. In *Artificial intelligence and statistics*, 847–855. PMLR. URL http://proceedings.mlr.press/v38/sgouritsa15.pdf.

Shimizu, S. 2014. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika* 41(1): 65–98. URL http://www.cox-associates.com/CausalAnalytics/LiNGAMShimuzi2014.pdf.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; Kerminen, A.; and Jordan, M. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7(10). URL https://www.jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf.

Shimizu, S.; Hyvärinen, A.; Kawahara, Y.; and Washio, T. 2009. A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 506–513. URL https://arxiv.org/abs/1408.2038.

Spirtes, P.; Glymour, C.; Scheines, R.; et al. 2000. Causation, Prediction, and Search. *MIT Press Books* .

Sun, X.; Janzing, D.; and Schölkopf, B. 2006. Causal inference by choosing graphs with most plausible Markov kernels. In *Ninth International Symposium on Artificial Intelligence and Mathematics (AIMath 2006)*, 1–11. URL https://pure.mpg.de/rest/items/item_1791171/component/file_3158882/content.

Sun, X.; Janzing, D.; and Schölkopf, B. 2008. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing* 71(7-9): 1248–1256. URL https://www.sciencedirect.com/science/article/pii/S092523120800060X.

Szabó, Z. 2014. Information Theoretical Estimators Toolbox. *Journal of Machine Learning Research* 15: 283–287. URL https://www.jmlr.org/papers/volume15/szabo14a/szabo14a.pdf.

Wright, S. 1921. Correlation and causation. *Journal of Agricultural Research* 20: 557–580.

Zhang, K.; and Hyvarinen, A. 2009. On the Identifiability of the Post-Nonlinear Causal Model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI2009)*, 647–655. URL https://arxiv.org/abs/1205.2599.