Identifiability of Cause and Effect using Regularized Regression

Alexander Marx
Max Planck Institute for Informatics,
Saarbrücken, Germany
amarx@mpi-inf.mpg.de

Jilles Vreeken
CISPA Helmholtz Center for Information Security,
Saarbrücken, Germany
jv@cispa.saarland

ABSTRACT

We consider the problem of telling apart cause from effect between two univariate continuous-valued random variables X and Y. In general, it is impossible to make definite statements about causality without making assumptions on the underlying model; one of the most important aspects of causal inference is hence to determine under which assumptions are we able to do so.

In this paper we show under which general conditions we can identify cause from effect by simply choosing the direction with the best regression score. We define a general framework of identifiable regression-based scoring functions, and show how to instantiate it in practice using regression splines. Compared to existing methods that either give strong guarantees, but are hardly applicable in practice, or provide no guarantees, but do work well in practice, our instantiation combines the best of both worlds; it gives guarantees, while empirical evaluation on synthetic and real-world data shows that it performs at least as well as the state of the art.

CCS CONCEPTS

• **Mathematics of computing → Causal networks**; *Robust regression.*

KEYWORDS

causal inference, identifiability, regression

ACM Reference Format:
Alexander Marx and Jilles Vreeken. 2019. Identifiability of Cause and Effect using Regularized Regression. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3292500.3330854

1 INTRODUCTION

Determining cause from effect is one of the most fundamental questions in science [19]. Unlike standard associative models, causal models give insight in the true data generating process and allow for answering *what-if* questions, which make them both robust and transparent. Determining causality cannot, however, be done without making assumptions on the causal model [19]. To ensure that we discover causation rather than association we need to assume a model that is *identifiable.* That is, models that under specific conditions and given infinite data are *guaranteed* to infer the correct causal direction. After all, unless we know we can unambiguously distinguish cause from effect given infinitely many samples there is little point in trying it on fewer samples. A lot of research in causal inference is therefore focused on identifiability results, figuring out the conditions under which models are identifiable.

Here, we consider the problem of inferring the causal direction between two statistically dependent continuous-valued random variables. Given a sample from their joint distribution $P(X, Y)$ we want to determine whether X causes Y, or Y causes X. We assume there is no hidden confounder Z that causes both X and Y, that is, we assume causal sufficiency. This problem is particularly important because not only do we often only have two variables—and hence conditional independence tests are inapplicable—but also as it allows us to orient any edge in the Markov-equivalent causal skeletons that constraint-based methods discover [26].

The first approach proposed for this setting was the Additive Noise Model (ANM). The main assumption of the ANM is that *effect* is generated as a function of *cause* with additive noise independent of the *cause*, i.e. $Y = f(X) + N_X$ with $X \perp\!\!\!\perp N_X$. Shimizu et al. [25] showed that that the true causal direction is identifiable if f is linear and N_X is non-Gaussian, as then there does not exist a function g for which $X = g(Y) + N_Y$ such that $Y \perp\!\!\!\perp N_Y$. It has since been shown that this holds for a broad range of settings [8, 9, 21, 22, 31], but the ANM remains a rather strict assumption on how the world works; in practice we often find functions and independent noise for both directions, which puts us back at square one.

The second main line of research in causal inference is based on the algorithmic Markov condition [11]. This postulate by Janzing and Schölkopf states that the true causal model coincides with the factorization of the joint distribution with the lowest Kolmogorov complexity. In other words, the causal model is the simplest explanation of the joint distribution. This model is highly general, as Kolmorogov complexity can capture any physical process [5], but also unpractical as Kolmogorov complexity is not computable. We can instantiate it, however, with other notions of complexity. Although methods based on this postulate typically do not come with strong identifiability results, they tend to perform very well in practice [10, 16, 24, 28].

Recently, Blöbaum et al. [3] gave a set of assumptions under which the true causal direction is identifiable via regression; they show that it is possible to identify cause from effect simply by selecting the direction of minimal residual error. That is, they fit both $Y = f(X) + N_X$ and $X = g(X) + N_Y$ minimizing the respective residual errors N_X and N_Y, and then directly compare the sum of squared errors. They show that for models of the complexity of the true model, the one in the causal direction will achieve the lower

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330854

error. This method is not very practical, however, as then we do not know the complexity of the true model and will be comparing residuals of arbitrarily under- or overfit models.

In this paper we extend these results, and show under which conditions cause and effect are identifiable via *regularized* regression. That is, we do not have to assume the true complexity of the causal model, but rather can compare models of different complexity. The key assumption that makes this possible we derive from Kolmogorov's structure function, and states that the best anti-causal model requires at least as many parameters as the causal model. We show that a large class of L_0 based regularized regression functions are identifiable, and as a proof of concept instantiate this general framework using spline-based regression.

Through experiments on synthetic and real-world data it turns out that this instantiation, performs very well in practice; it outperforms identifiable methods, such as RESIT [21], IGCI [10], and performs either on-par or better than existing non-identifiable methods such as Slope [16], CAM [4], and QCCD [28]. Important for practical usage, and unlike its competitors, it shows a strong correlation between confidence and accuracy; essentially, if it is confident we can trust it, and if it's not, we should refrain from using it.

The roadmap of this paper is as follows. First, in Sec. 2 we cover the preliminaries and give a short introduction to the main concepts of RECI [3] that we build upon. We then in Sec. 3 show how we can derive the key assumption to our approach. Based on this assumption, we show that the class of *identifiable regression-based scoring functions* is identifiable, and show how to instantiate it. Related work is discussed in Sec. 5. We empirically evaluate our method in Sec. 6, and round up with conclusions in Sec. 7

2 PRELIMINARIES

In this section, we first introduce the notation, then briefly explain the main idea behind RECI [3], its limitations and how we want to solve them.

2.1 Notation

We consider causal inference from two correlated random variables X and Y and assume causal sufficiency—i.e. there exists no confounding variable Z. In particular, we use capital letters X for a random variables and lowercase letters x to values from the domain $\mathcal{X}$ of X. We write β_f for the set of parameters of a function f and denote with $||\beta_f||_0$ to the number of non-zero parameters. Unless explicitly stated, we use log to refer to the logarithm with base 2 and follow the convention that $0 \log 0 = 0$.

2.2 A brief Introduction to RECI

The general idea behind RECI [3] is that we can infer cause from effect simply by comparing the regression error of the best fitting model for the causal and anti-causal direction. In particular, they formulate a set of assumptions under which they can differentiate between cause and effect with certainty. Formally, if ϕ is the function that minimizes the least-squared error when predicting the effect Y from the cause X and vice versa ψ the function minimizing the error when predicting the cause from the effect, Blöbaum et al. [3] formulate a set of assumptions, under which

$$\mathbb{E}[(Y_\alpha - \phi(X))^2] \leq \mathbb{E}[(X - \psi(Y_\alpha))^2] \qquad (1)$$

always holds. In other words, when their assumptions hold, Blöbaum et al. [3] proof that we can *identify* the true causal model using Eq. (1). That is, if the assumptions below are fulfilled we know with certainty that Eq. (1) holds. Hence, we can use the asymmetry in the regression error to infer the causal direction between two random variables. Identifiability is an important concept in causal inference, as we can only make statements about the true causal model, when we can guarantee identifiability. As this cannot be done in general, the goal is to proof identifiability under a set of assumptions that are as lightweight and as general as possible. The main assumptions for RECI, can be summarized as follows [3].

Assumption 1 (Causal model). *We can write the effect as*

$$Y_\alpha := f(X) + \alpha N \, ,$$

with noise term N and parameter α restricting the noise level.

Assumption 2 (Unbiased noise). *The noise term N is unbiased and has unit variance.*

Assumption 3 (Compact supports). *The distribution of X has compact support and w.l.o.g. X attains values between 0 and 1, which can be achieved by normalizing X. Further, the distribution of N has compact support and there exist values $n_+ > 0 > n_-$ such that for each value $x \in \mathcal{X}$, $[n_-, n_+]$ is the smallest interval that containing the support for the conditional density of N given x. Hence, we know that $[\alpha n_-, 1 + \alpha n_+]$ is the smallest interval containing the support of the density of Y_α and rescale it to*

$$\tilde{Y}_\alpha := \frac{Y_\alpha - \alpha n_-}{1 + \alpha n_+ - \alpha n_-} \, .$$

$\tilde{Y}_\alpha$ has the same scale as X and attains values between 0 and 1.

Based on Assumptions 1-3, Blöbaum et al. [3] show that their approach works under the additive noise assumption, that is, the cause X is independent of the noise term. In addition, their framework allows slight violations of this assumptions, as it also works when there is a low dependence between the noise and the cause. In particular, they show that Eq. (1) holds for strictly monotonically increasing and twice differentiable ϕ and trivially holds for non-invertible functions, as there is an information loss in the anti-causal direction. Last, they show that Eq. (1) holds with equality iff ϕ is a linear function, which means that we cannot identify the causal direction for linear functions.

In general, RECI provides a solid framework to identify cause from effect only based on regression error, which is easy to obtain. Also, Assumption 3 is not very restrictive, as we can achieve it by normalizing the data, if we have a sufficient number of samples. The problem is, however, that in practice we do not know the true functions. Hence, we need to restrict ourselves to comparing functions of the same type, i.e. polynomials of degree three or six, but cannot compare across functions of different complexities. The goal of this paper is to solve exactly this issue, while conserving the identifiability guarantees.

2.3 Main Idea

The key idea to solve the limitations of RECI is simple. Instead of only comparing the regression error, we use regularized regression and compare the regularized scores of those functions for which they are minimal. To illustrate this, consider the following example.

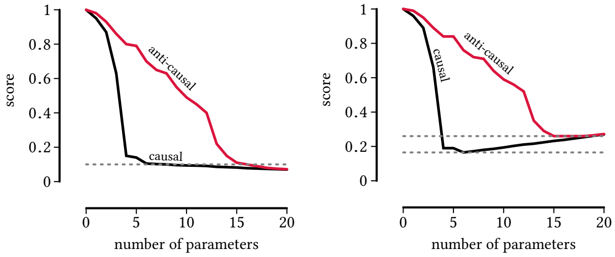


Figure 1: Left: Error for the best fitting function in the causal and anti-causal direction, when restricting the number of parameters. Right: Error plus L_0 penalty on the number of parameters for the same data.

Example 1. *Assume we are given a sample from the joint distribution of X and Y, and we know the true causal direction. Now we use our go to algorithm to fit a regression function in both the causal and the anti-causal direction. In Figure 1, we plot the minimum regression error for both directions, where we gradually allow the model to fit more parameters. If we allow a sufficient number of parameters, we can reduce the regression error for both directions to approximately the same level. As a consequence, comparing purely the regression errors of the best fitting model does not identify the correct causal direction. However, we observe that for the true causal direction, we can find a much simpler function, which attains approximately the same regression error; in contrast to the anti-causal direction. When we compare the scores of those functions minimizing the regression error plus an L_0 penalty over the parameters (right plot), we can identify the correct model, as there is a clear difference between both scores.*

Of course, we do not want to rely on a proof by an artificial example, but from Example 1 we get our motivation. What we completely forgot about for a second, is the identifiability. It is known that we can use regularized regression to fit functions, but does this also result in scoring functions that are identifiable?

In the following, we will show that it does. In particular, we define a class of scoring functions for regularized regression that are identifiable under the assumption that the mechanism mapping the cause to the effect is simpler than the anti-causal one. We derive and justify this assumption from the algorithmic model of causality using Kolmogorov's structure function.

3 PRINCIPLED REGULARIZATION FOR CAUSALITY

In order to define our new inference rule, we need to introduce one more assumption, that is, we assume that true causal model has a lower complexity than the anti-causal model. This claim might not be too intuitive and hence we are going to carefully justify our assumption in the following from the algorithmic model of causality, which is formulated in terms of Kolmogorov complexity. Before we introduce Kolmogorov complexity, note that in this context, a lower-case letter x will refer to the binary string representation of X and not to a value from the domain of X.

Kolmogorov complexity. The Kolmogorov complexity of a finite binary string x is the length of the shortest binary program p^* for a universal Turing machine $\mathcal{U}$ that generates x, and then halts [13, 14]. Formally, we have

$$K(x) = \min_p \{|p| \mid p \in \{0,1\}^*, \mathcal{U}(p) = x\} .$$

That is, program p^* is the most succinct *algorithmic* description of x, or in other words, the ultimate lossless compressor for that string. Hence, $K(x)$ is the length of this compressor and therefore the perfect measure of complexity. Further, the conditional Kolmogorov complexity is defined as

$$K(x \mid y) = \min_q \{|q| \mid q \in \{0,1\}^*, \mathcal{U}(y, q) = x\} \leq K(x) ,$$

which is again the length of the shortest binary program p^* that generates x, and halts, but now given y as input for free.

We next introduce the algorithmic Markov condition, which builds upon the algorithmic model of causality (AMC) [11]. The latter is defined over a general causal directed acyclic graph and states that in a causal network, each node can be generated from its parents and an additional noise term. In the following, we simplified the definition to the graph with only X and Y and a single edge, that is $X \to Y$.

Postulate 1 (Algorithmic Model of Causality). *Let x and y be two strings and let further $x \to y$ be a causal graph, where no latent confounder exists. Then y is computable by a program q with length $O(1)$ from x and an additional input n. We formally write*

$$y = q(x, n) ,$$

meaning that the Turing machine computes y from the inputs x, n using the additional program q and halts.

In other words, the effect can be generated from a program that takes the cause and a noise term as input. This program can in theory model every physical process [5], which includes functional relationships. Hence, it also supports the causal model that we assume in this paper (Assumption 1). Under the algorithmic model of causality, Janzing and Schölkopf postulate that the symmetry of information—i.e.

$$K(P(X,Y)) \overset{+}{=} K(P(X)) + K(P(Y \mid X)) \overset{+}{=} K(P(Y)) + K(P(X \mid Y)) ,$$

where $\overset{+}{=}$ denotes equality up to an additive constant, does not hold [11]. In particular, they postulate that if $X \to Y$,

$$K(P(X)) + K(P(Y \mid X)) \overset{+}{\leq} K(P(Y)) + K(P(X \mid Y)) . \tag{2}$$

That is, we infer that direction, which provides the simplest factorization of the joint distribution of X and Y. In theory, we could infer the causal direction for any physical process—if only we could compute Kolmogorov complexity. One way to approximate Kolmogorov complexity is to split it into the complexity of the meaningful information that can be efficiently represented by a short program and the complexity of the irreducible noise that cannot be modelled efficiently. A sound theoretical concept that differentiates between those quantities is described by Kolmogorov's structure function.

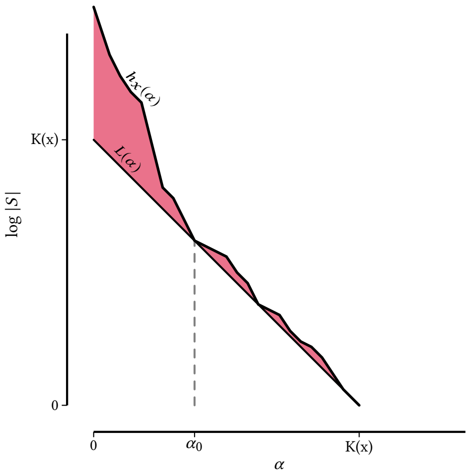


Figure 2: Shown is the sufficiency line $L(\alpha) = K(x) - \alpha$ and the structure function h_x. At α_0 corresponding to the minimum sufficient statistic, h_x hits the sufficiency line for the first time. For all $\alpha > \alpha_0$ where $h_x(\alpha) = L(\alpha)$, the corresponding set S is a sufficient statistic for x. Marked in red it the gap between $K(x)$ and $h_x(\alpha) + \alpha$.

Kolmogorov's Structure Function. Although there is no written publication of Kolmogorov about the structure function, it has found its way into research [29]. The key concept we need is that of a model $S \ni x$, that is, a model is a set of binary strings of which x is a member. Given such a set S and no further input, we will need $\log |S|$ bits to look up x in S. Simple models, i.e. those with low Kolmogorov complexity $K(S)$, will consist of many possible strings and hence it will take us relatively many bits to identify x in S. If we increase the budget for $K(S)$ we can contemplate more complex models that consist of fewer possible strings, and for these it will cost much fewer bits to single out x. In the most extreme case, where we set $K(S) = K(x)$, we can have $S = \{x\}$. Formally, we can describe this relationship as Kolmogorov's structure function

$$h_x(\alpha) = \min_S \{\log |S| \colon S \ni x, K(S) \leq \alpha\}$$

with S being a contemplated model for x and α a non-negative number bounding the complexity of the contemplated S's [29]. There exists complexity threshold's α for which $\alpha + h_x(\alpha) = K(x) + O(1)$. For these, the associated model S is called an *optimal set* for x. Its description of up to α bits is called sufficient statistic for x. Moreover, for a sufficient statistic S it holds that $K(x) \leq K(S) + \log |S| \leq K(x) + O(1)$. If we consider all sufficient statistics S for x, we call that S which is associated with the smallest α—i.e. α_0, the *minimal sufficient statistic* for x. That is, the minimum sufficient statistic S contains all meaningful information about x and the associated term $h_x(\alpha_0)$ measures the complexity of the irreducible noise contained in x. Further, it holds that $h_x(\alpha_0) + \alpha_0 = K(x)$. In Figure 2 we visualize this concept as suggested by Vereshchagin and Vitányi [29]. We see that for α_0 the structure function h meets the sufficiency line, that is defined as $L(\alpha) = K(x) - \alpha$, which is optimal and hence $\alpha_0 + h_x(\alpha_0) = K(x)$. For $\alpha < \alpha_0$, $h_x(\alpha)$ can be arbitrarily far above the sufficiency line and for $\alpha > \alpha_0$, $h_x(\alpha)$ is within a constant term above the sufficiency line.

Similar to conditional Kolmogorov complexity, we define the conditional structure function as

$$h_x(i \mid y) = \min_S \{\log |S| \colon S \ni x, K(S \mid y) \leq i\} \,.$$

We will need this conditional version as we will be considering functional relationships from X to Y and vice versa.

Now let us consider Eq. (2) again and let x and y correspond to the binary string representations of X and Y. Further, be i_0^x the complexity level of the minimum sufficient statistic of x conditioned on y, and accordingly i_0^y the complexity level of the minimum sufficient statistic for y given x. We can rewrite Eq. (2) as

$$K(x) + i_0^y + h_y(i_0^y \mid x) \leq K(y) + i_0^x + h_x(i_0^x \mid y) \,.$$

In the following, we explain the above inequality given that Assumption 1-3 hold. As i_0^x contains all meaningful information of x given y, $h_x(i_0^x \mid y)$ relates to $K(C - \psi(E_\alpha))$ and $h_y(i_0^y \mid x)$ relates to $K(E_\alpha - \phi(C)) = K(N)$. Now assume that ϕ is an invertible function, we find that according to Postulate 1 both i_0^x and i_0^y must have constant complexity. If the variance of the noise term goes to zero, that is, the function is near deterministic, Blöbaum et al. [3] showed that the expected error for the causal model is smaller or equal to the error in the anti-causal direction—i.e. $h_y(i_0^y \mid x) \leq h_x(i_0^x \mid y)$. As a consequence, purely judging from the algorithmic Markov condition, comparing only the expected least squared errors for both directions, as done in RECI [3], can only be true if we assume that $K(x) \leq K(y)$. Conceptually, by standardizing or normalizing X and Y, we can achieve that $K(x) \overset{+}{=} K(y)$. If we standardize the data and assume a Gaussian distribution, we can describe $K(x)$ and $K(y)$ with a zero mean and unit variance, which leads to (approximately) the same complexity. If we normalize, we can assume a uniform distribution or prior and achieve the same effect. When $K(x) = K(y)$, we can infer that $X \to Y$, if

$$i_0^y + h_y(i_0^y \mid x) \leq i_0^x + h_x(i_0^x \mid y) \,. \tag{3}$$

Note that this inequality also holds if the function ϕ is not invertible and there does not exist an inverse function ψ. This follows from the fact that there is an information loss in the anti-causal direction and we cannot efficiently use the information about x to derive y. In addition, we can see from Eq. (3) that if we only consider the regression error, it is important to know the true functions. If we do not, and overfit in e.g. the anti-causal direction we fit noise and obtain lower errors than are true, which can lead to wrong inferences. Formally, if we allow for a complexity level $i^x > i_0^x$, it is possible and for large i^x will eventually happen that $h_x(i^x \mid y) < h_x(i_0^x \mid y)$. If, we also consider the complexity of the function, however, we have that $K(x \mid y) \leq i^x + h_x(i^x \mid y) \leq K(x \mid y) + O(1)$ and hence $i^x + h_x(i^x \mid y) \geq i_0^x + h_x(i_0^x \mid y)$. In other words, we are resistant against making wrong inferences due to overfitting. Further, we are also resistant to underfitting as for $i^x < i_0^x$ we have that $K(x \mid y) \leq i^x + h_x(i^x \mid y)$ [29].

Hence, we need to include the complexity of the model into our score, without breaking the identifiability results. Judging purely algorithmically, we know from Postulate 1 that both i_0^x and i_0^y are constant if ϕ is an invertible function. If not, i_0^x could be larger. As a consequence, we have that $i_0^y \overset{+}{\leq} i_0^x$. As we cannot compute Kolmogorov complexity, we need to formalize this idea differently. In essence, if the causal mechanism has a lower complexity than the anti-causal one, the true causal function ϕ should need at most as many parameters or degrees of freedom as the reverse function ψ. We formulate this in Assumption 4.

Assumption 4 (Simplicity). *Let Y_α be generated as in Assumption 1 [3]. Further, let ϕ be the function minimizing the expected least-squared error for predicting the effect Y from the cause X and ψ be the function minimizing the expected least-squared error in the anti-causal direction. We assume that ψ has at least as many parameters as ϕ, i.e. $\|\beta_\phi\|_0 \leq \|\beta_\psi\|_0$.*

While we cannot show that Assumption 4 holds in general, there are strong indications that it holds in many real-world settings. For example, if we know that ϕ consists of a linear combination of basis functions that are linearly independent of each other, we cannot find an inverse function that has fewer degrees of freedom. Moreover, Kilbertus et al. [12] recently considered the problem of anti-causal learning and give indications on why it is harder than learning the causal direction. In particular, they give various examples, why it is simpler to learn the causal direction, from which we selected a few. As for low degree polynomials, it is easy to see that it is not possible to formulate an inverse with less parameters as the original function, the Abel-Ruffini theorem states that general polynomial equations of degree greater than 4 do not have an algebraic solution [1]. Further, it is known that some elementary transcendental functions as $x + \sin(x)$ do not have an elementary inverse. In addition, in cryptography there exist the concept of a one-way function [1]. Those are functions, that are easy to obtain in one direction but almost impossible to reverse.

Utilizing Assumption 4, we can finally connect all the dots and introduce our new framework.

4 IDENTIFIABLE REGULARIZED REGRESSION

In the following, we show how we can design scoring functions, which 1) allow to identify the true causal direction under Assumptions 1-4, 2) help to identify the true functions ϕ and ψ and 3) are more robust w.r.t. overfitting. To this end, we define below a Identifiable Regression-based Scoring Function, or short IRSF and show that an IRSF fulfils the claims listed above.

Definition 1 (Identifiable Regression-based Scoring Functions). *Given two random variables X and Y and a regression function ϕ that maps X to Y. Further, we are given a scoring function $S : \mathbb{R}_{\geq 0} \times \mathbb{N} \mapsto \mathbb{R}$ that takes as input the expected least-squared error $\mathbb{E}[(Y - \phi(X))^2]$ and the number of parameters of ϕ, $\|\beta_\phi\|_0$. We call such a scoring function*

$$S(Y \mid X, \phi) := \gamma(\mathbb{E}[(Y - \phi(X))^2]) + \lambda(\|\beta_\phi\|_0)$$

an Identifiable Regression-based Scoring Function (IRSF), if both $\gamma : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}$ and $\lambda : \mathbb{N} \mapsto \mathbb{R}$ are strictly monotonically increasing.

It is easy to see that the number of parameters corresponds to the complexity of the function and hence $\lambda(\|\beta_\phi\|_0)$ could be instantiated such that it approximates i_0^y. Further, under Assumptions 1-2, we can see that $\gamma(\mathbb{E}[(Y - \phi(X))^2])$ can be formulated to approximate $h_y(i_0^y \mid x)$. Hence, if we instantiate γ and λ correctly, comparing $S(Y \mid X, \phi)$ to $S(X \mid Y, \psi)$ is an approximation of comparing $K(y \mid x)$ to $K(x \mid y)$ and hence sufficient to infer the causal direction, if we preprocess the data, e.g. via standardization or normalization, such that $K(x) \overset{+}{=} K(y)$. However, the question that remains is, can we identify this model and under which conditions. This we formalize in our main theorem.

Theorem 1. *Let Assumptions 1-4 hold, where ϕ denotes the function that minimizes the expected least-squared error when predicting the effect Y from the cause X and ψ be the function minimizing the expected least-squared error for predicting X from Y—i.e. $\phi(x) = \mathbb{E}[Y|x]$ and vice versa $\psi(y) = \mathbb{E}[X|y]$. Further, let S be an IRSF according to Definition 1. The following limit always holds*

$$\lim_{\alpha \to 0} \frac{S(\mathbb{E}[(\tilde{Y}_\alpha - \phi(X))^2], \|\beta_\phi\|_0)}{S(\mathbb{E}[(X - \psi(\tilde{Y}_\alpha))^2], \|\beta_\psi\|_0)} \geq 1\,,$$

with equality if and only if ϕ is linear.

Proof. We know from Blöbaum et al. [3] that under Assumptions 1-3 the following always holds

$$(*) = \lim_{\alpha \to 0} \frac{\mathbb{E}[(\tilde{Y}_\alpha - \phi(X))^2]}{\mathbb{E}[(X - \psi(\tilde{Y}_\alpha))^2]} \geq 1\,. \tag{4}$$

As S is an IRSF, we can write it as $S(a, b) := \gamma(a) + \lambda(b)$, where γ is a strictly monotonically increasing function. Hence, the statement does not change by applying γ to the nominator and denominator in Eq. (4). Based on Assumption 4 we know that $\|\beta_\phi\|_0 \leq \|\beta_\psi\|_0$. Hence,

$$\frac{\gamma(\mathbb{E}[(\tilde{Y}_\alpha - \phi(X))^2]) + \|\beta_\phi\|_0}{\gamma(\mathbb{E}[(X - \psi(\tilde{Y}_\alpha))^2]) + \|\beta_\psi\|_0} \geq (*)\,,$$

with equality if and only if $\|\beta_\phi\|_0 = \|\beta_\psi\|_0$. As λ is strictly monotonically increasing, applying it to $\|\beta_\phi\|_0$ and $\|\beta_\psi\|_0$ will not change this statement. □

4.1 Specifying γ and λ

Theorem 1 holds independently of how we exactly specify γ and λ. The problem, however, is that we do not know ϕ nor ψ beforehand. If we knew those functions, we could also apply the inference rule that is used in RECI [3]. The advantage of our score is that it not only identifies the true causal direction, when given ϕ and ψ, but also if specified correctly, can help to find exactly those functions and hence reduces the probability to overfit and underfit.

The perfect definition of γ and λ would be such that the minimum value of S is attained when the function we find approximates the minimum sufficient statistic and no further structure can be exploited, leaving γ to be the cost function over the irreducible noise. Therefore, it is important to specify S s.t. it approximates the Kolmogorov complexity of the conditional. If S gives too much weight to γ, we prioritize minimizing the error, which will lead to overfitting. On the other hand, if we define λ such that it grows too fast, we over-penalize complexity and underfit.

To illustrate this, consider Example 1 again. If we assign too little weight to the complexity of the function, we could probably train a deep neural network for the anti-causal direction that has a similar regression error as the simple causal model. Luckily, model selection is not a new topic and there already exist model selection criteria that try to avoid overfitting and aim at recovering the true function [2, 7, 23]. Interesting for us are only those that can be specified as an IRSF. We provide a selection of those below.

The most well-known scoring functions that we can write as an IRFS according to Definition 1 are the Akaike information criterion [2] (AIC) and the Bayesian information criterion [23] (BIC).

AIC. For the causal direction AIC can be written as

$$n \log(\mathbb{E}[(\tilde{Y}_\alpha - \phi(X))^2]) + 2\|\beta_\phi\|_0 + c\,,$$

where c is a constant term independent of the model. As the sample size n is the same for the causal and the anti-causal direction, we can consider it as a parameter of the function γ and write down an IRSF with $\gamma(a) := n\log(a)$ and $\lambda(b) = 2b + c$.

BIC. The Bayesian information criterion for scoring the causal direction is equal to

$$n \log(\mathbb{E}[(\tilde{Y}_\alpha - \phi(X))^2]) + \log(n) \cdot \|\beta_\phi\|_0$$

and can similar to AIC be written as an IRSF. Hence, both scores can be used in Theorem 1. One detail that we have to consider for AIC and BIC is that log is not defined for 0 and is negative for values between 0 and 1. Hence, it is necessary to adjust both scores by taking $\log(\mathbb{E}[(\tilde{Y}_\alpha - \phi(X))^2] + 1)$.

MDL. One well defined way to balance the complexity of the model and the data given the model is to use two-part Minimum Description Length (MDL) codes [7]. Simply put, when we want to measure the complexity of data D, or its description length, we restrict ourselves to a model class $\mathcal{M}$ for which we know how to compute the description length—i.e. this could be the class of regression functions. Then we find that model $M \in \mathcal{M}$ for which

$$L(D, M) := L(D \mid M) + L(M)$$

is minimal. In other words, we jointly minimize the complexity of the data given the model and the complexity of the model. In addition, there is a close connection between MDL and Kolmogorov's structure function. In particular, $L(D \mid M)$ corresponds to the value of the structure function h and $L(M)$ to the complexity level α. Further, when the model class $\mathcal{M}$ contains the true model, the model M minimizing $L(D, M)$ describes the minimum sufficient statistic [29].

Defining an optimal encoding for continuous data without making any assumptions is a hard problem. One approach that utilizes two-part MDL to approximate the algorithmic Markov condition for continuous data is Slope [17]. In Slope, the main assumption is that the error is Gaussian distributed. Crudely speaking, the score used in Slope can be written as $\gamma(\mathbb{E}[(\tilde{Y}_\alpha - \phi(X))^2])$, where γ is based on the negative log likelihood, plus a function ρ over the parameters. As this function ρ does not purely consider the number of parameters, but assigns different weights according to the value, of the parameter, the corresponding scoring function is not an IRSF and hence Theorem 1 does not apply for Slope. If we loosen the encoding from Slope slightly and "forget" about the exact values of the parameters but encode each parameter with the same constant number of bits, we arrive at an IRFS and Theorem 1 can be applied. In this case, the encoding would be called *lossy* as we do not encode all the information available to us.[1]

4.2 Instantiation

In theory, there are many possible ways to instantiate our framework, as we can use every function learning algorithm that minimizes the regression error and allows to control the number of parameters. During our empirical evaluation, we evaluate two possible ways, one using basis functions and one splines.

We refer to our method as Sloppy. We name it such both because it is partially inspired by Slope, because it is the first instantiation of the IRSF framework, but primarily because from an information-theoretic perspective the notion of a constant penalty per parameter can be inefficient (too high), as well as lossy (too low), and hence, sloppy. In practice, we consider the following two variants,

(1) **Sloppy**$_B$**:** We find the best linear combination according to the given score function S from a set of basis functions that include polynomials up to a degree of six, an exponential and logarithmic basis function as well as reciprocal up to the degree of two. This can be done with an algorithm following the standard forward-backward selection scheme.

(2) **Sloppy**$_S$**:** We fit a cubic spline, where we control the degrees of freedom and find that selection of splines, for which S is minimal. Even, when we do this exhaustively, Sloppy$_S$ is still very fast in practice.

For our experiments, we use AIC and BIC as scoring function S.

Inference. Before applying Sloppy, we standardize X and Y to zero mean and unit variance or normalize them between zero and one. Hence, we have that $K(x) \overset{+}{=} K(y)$ and can infer the causal direction according to Theorem 1, as described in the previous section. Then given an IRSF S, we use Sloppy to compute those functions ϕ and ψ that minimize $S(Y \mid X, \phi)$ and $S(X \mid Y, \psi)$. We decide that $X \to Y$ if $S(Y \mid X, \phi) < S(X \mid Y, \psi)$, that $Y \to X$ if $S(Y \mid X, \phi) > S(X \mid Y, \psi)$ and do not decide in case of equality.

4.3 Confidence

The authors of RECI [3] showed that in empirical evaluations we can use the minimum of the error terms for both directions divided by the maximum as a confidence measure. We do so accordingly and define the confidence of a decision as

$$C(X, Y) := 1 - \frac{\min\{S(Y \mid X, \phi), S(X \mid Y, \psi)\}}{\max\{S(Y \mid X, \phi), S(X \mid Y, \psi)\}}\,.$$

The higher $C(X, Y)$, the more certain we are that our decision is correct. This allows to order decisions across different inferences by their confidence. In addition, we can set a threshold t such that we require $C(X, Y) \geq t$ and otherwise do not decide for a direction as we are not confident enough about the decision.

5 RELATED WORK

Recently, causal inference for the bivariate setting assuming no confounder has attracted a lot of attention [3, 9, 16, 20, 28]. Traditional constraint based approaches, such as conditional independence

[1] Such an encoding could also mean that we assume that all parameters are drawn from the same distribution and hence use a fixed amount of bits to encode them.

tests, require at least three random variables and hence cannot be used to identify the causal direction in the bivariate setting [19, 30]; unlike those approaches that we are going to discuss. In this section, we restrict ourselves to state of the art methods for continuous data and those that are strongly connected to our work.

The first approaches with strong identifiability guarantees are those that are based on the Additive Noise Model (ANM) [20, 25] where we assume that Y was generated as a function of X with additive noise N_X independent of X, i.e. $Y = f(X) + N_X$ with $X \perp\!\!\!\perp N_X$. It turns out that for various settings [21, 22] this model is identifiable as there does not exist an ANM in the anti-causal direction; it is impossible to find a function $X = g(Y) + N_Y$ where $Y \perp\!\!\!\perp N_Y$ holds. This is the case for linear functions f and non-Gaussian noise N_X [25], nonlinear functions and additive noise [8], post-nonlinear models [31], as well as for mixtures of multiple additive noise models [9].

A limiting factor of these approaches is that the results strongly rely on the used independence test and the fitting algorithm [18]. Problems can arise when the functions overfit. In addition, it is hard to derive a meaningful confidence score from the corresponding p-values, as they are highly dependent on the sample size [16].

Similar to RECI, Janzing et al. [10] also developed an approach for the low noise setup. In particular, they infer the causal direction based on the Shannon entropy of the marginals. A problem with this approach is, that it is hard to accurately estimate the entropy for continuous data.

More closely related to our approach are methods based on the algorithmic Markov condition [11]. The first step towards this direction was a method based on Minimum Message Length (MML) [27], however, it did not outperform the competing methods based on the additive noise setup. More recent approaches on the other hand have shown good performance on real world as well as synthetic data, outperforming state of the art methods that try to maximize the independence between cause and noise distribution [16, 28]. A very recent proposal, QCCD [28], approximates the algorithmic Markov condition using non-parametric conditional quantile estimation. Although performing well in practice, QCCD lacks strong identifiability guarantees.

As already described in Section 4, the most related methods to this work are RECI [3] and Slope [16] as both approaches base their inference rules on the regression error. However, RECI does not employ model selection and Slope has no strong identifiability results. A third method that uses the regression error is CAM [4], which was designed to find a general causal graph. For the bivariate case, CAM decides for the causal direction using regularized log-likelihood scores.

6 EXPERIMENTS

In this section, we empirically evaluate Sloppy and benchmark it against competing state of the art methods. To represent additive noise models, we select RESIT [21] using the Hilbert Schmidt Independence Criterion to measure the independence between cause and noise distribution [6]. A recent study shows that the overall performance of RESIT on simulated and real-world data is on par if not better than competing methods of this type [28]. In addition, we compare against IGCI [10] representing methods for the low-noise setup and QCCD [28] as it is to the best of our knowledge the method with the best overall performance. Note that we can configure Slope and CAM such that we obtain similar results as those obtained with Sloppy. We provide this information as well as the configurations for all methods in the supplemental material.[2]

We first show the overall performance over synthetic and real-world benchmark data sets and then go a bit more into details. For all experiments, we applied both Sloppy$_S$ and Sloppy$_B$. As it turned out that the results differ only marginally, we here show only the results for Sloppy$_S$, which for conciseness we will refer to as Sloppy. For completeness, we provide the results for Sloppy$_B$ in the appendix. As the default scoring criterium, we used BIC for our experiments. We only applied AIC for the real-world benchmark data set, as there we found that BIC was too restrictive and mainly fitted linear models.

All experiments were performed single threaded and Sloppy took only up to a couple of seconds for a single pair. For research purposes and to make our results reproducible, we make the code for Sloppy available online.[3]

6.1 Benchmarking

In order to benchmark Sloppy against RESIT, QCCD and IGCI, we applied them to ten benchmark data sets and reported their accuracies. We took five data sets from Mooij et al. [18]. Those consist of four simulated data sets generated using a Gaussian process: *SIM* (without confounder), *SIM-ln* (with low noise), *SIM-G* (with distributions close to Gaussian) and *SIM-c* (with confounder). The fifth one is a collection of 99 real-world bivariate continuous cause effect pairs, known as the *Tübingen* benchmark data set (version from December 17), for which we weigh the pairs as recommended. The remaining five data sets were taken from Tagasovska et al. [28]. These consist of nonlinear functions with additive noise (*AN*), sigmoidal functions with additive noise (*AN-s*), nonlinear and sigmoidal location scale functions (*LS* and *LS-s*), where

$$Y = f(X) + g(X) \cdot N_Y$$

and sigmoid functions with multiplicative uniform noise (*MN-U*)—i.e. we can write the effect Y as

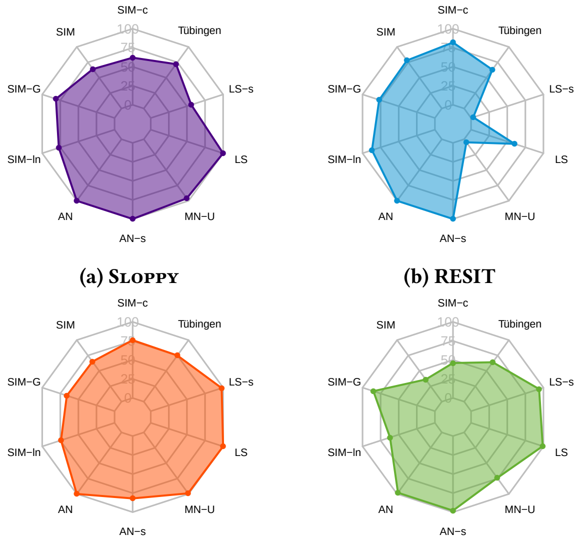
$$Y = f(X) \cdot N_Y \ .$$

All simulated data sets consist of 100 cause-effect pairs with 1 000 samples per pair.

In Figure 3 we show the accuracies for Sloppy, RESIT, QCCD and IGCI on all data sets. On average, Sloppy has an accuracy of 81%. If we consider only those data sets for which our assumptions hold, those are *AN* and *AN-s*, we have an accuracy of 100%. The reason why we could not achieve this for the *SIM* data sets is because they also contain pairs for which the function is close to linear, have high noise or are sampled from mixture models. Taking this under consideration, Sloppy still performs very well on these data sets. The only data set where we have a poor performance is *LS-s*, which violates our assumption w.r.t. the generating model. This is also the

[2] As RECI can be instantiated with any regression method and the authors do not provide a sample implementation, we cannot fairly compare against it but do give an intuition in the appendix.

[3] https://eda.mmci.uni-saarland.de/sloppy/



(a) Sloppy **(b) RESIT**

(c) QCCD **(d) IGCI**

Figure 3: Accuracy of Sloppy, RESIT, QCCD and IGCI over all synthetic data set and the *Tübingen* benchmark data set.

only data set where we clearly lose to QCCD. In turn, we perform better than QCCD on *AN-s* and are on par for the remaining data sets. Overall, RESIT and IGCI have more problems than Sloppy with those data sets that do not follow their assumptions.

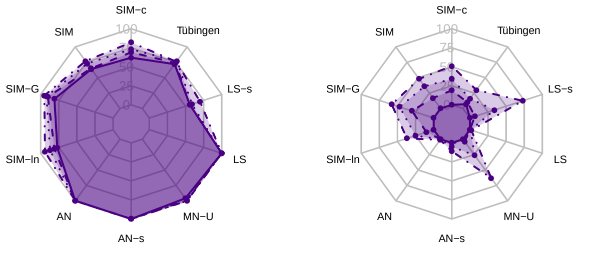
6.2 Setting a Confidence Threshold

In this experiment, we consider the same data sets as above and look at the confidence of Sloppy. In particular, we show in Figure 4 how the accuracy of Sloppy improves when we only consider those decisions with a confidence greater or equal than $\{0, 0.01, 0.05, 0.1\}$. We can observe that setting a threshold of 0.1 improves the average accuracy over all data sets from 81% to 89%, which clearly shows that we assign low confidence values to bad decisions. In addition, we show the percentage of pairs that do not reach the corresponding threshold. We undoubtedly see that this number is higher for those data sets that do not fulfil our assumptions, whereas for those data sets that do, the number of pairs where we do not decide remains low, even for a cut-off of 0.1.

6.3 Decision Rates

Highly related to confidence values are decision rates. In particular, we obtain a decision rate, if we order a set of decisions by their confidence values and report for each percentage k the accuracy over the top k% of the decisions. In Figure 5 we report the decision rates of Sloppy for each tested data set. Importantly, we observe that for all data sets, even for *LS-s*, the first 10% of our decisions are correct. Then, depending on the overall accuracy that we achieve on the corresponding data set, the accuracy slowly drops after considering more and more decisions with lower confidence values.

In addition, we show in Figure 6 the decision rates for Sloppy, RESIT, QCCD and IGCI for the real-world benchmark data set. Although the overall performance of all methods does not differ too much, we can clearly see that Sloppy has the best decision rate. In particular, for the first 31% of all decisions, we only get one decision wrong and only drop below 95% accuracy after considering more than 40% of all decisions. In comparison, the competing approaches more frequently assign high confidence values to wrong decisions.



(a) accuracy [%] **(b) undecided [%]**

Figure 4: Accuracy of Sloppy (left), for those decisions that have a higher confidence than $\{0, 0.01, 0.05, 0.1\}$ and right the corresponding percentage of pairs where we did not decide.

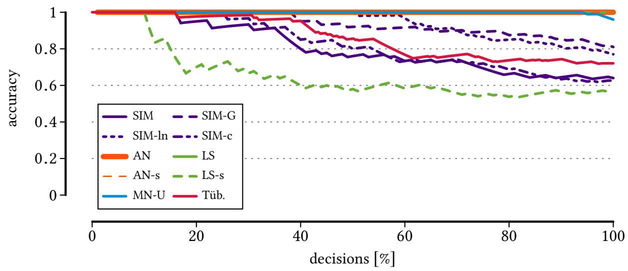


Figure 5: [Higher is better] Decision rates of Sloppy for every tested data set. As we obtain 100% accuracy for *AN*, *AN-s* and *LS*, those curves lie above each other.

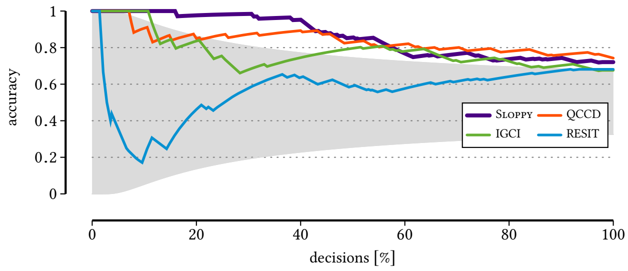
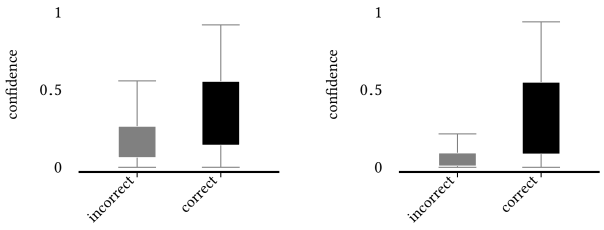


Figure 6: [Higher is better] Decision rates for Sloppy, RESIT, QCCD and IGCI on the *Tübingen* benchmark data set. The gray area marks the 95% confidence interval of a random coin flip.

6.4 Confidence Distribution

In the last experiment, we consider the distribution of the confidence values for correct and incorrect decisions for the real-world pairs and the simulated pairs, as shown in Figure 7. It is encouraging to see that there is a clear difference in the distribution and



(a) Tübingen **(b) Simulated**

Figure 7: Distribution of confidence values for correct and incorrect decisions for the *Tübingen* benchmark data set (left) and the simulated data sets (right).

higher confidence values are assigned to correct decisions. For the simulated data, the first quartile for the incorrect decisions (0.094) is approximately on the same level as the third quartile for the correct decisions (0.085), which means that we could almost separate the correct form the incorrect decisions using a threshold in this region. For the real-world data the distributions overlap a bit more, however, when applying the Wilcoxon-Mann-Whitney test [15], we get that the confidence values for the incorrect directions are smaller than for the correct directions with a p-value $< 10^{-4}$.

7 CONCLUSION

We considered causal inference between two continuous random variables X and Y without hidden confounders. In this setup, we showed under which conditions we can use regularized regression to identify cause from effect with guarantees.

As a possible instantiation of our framework, we introduced Sloppy—which finds the best fitting function for the causal and anticausal direction according to a given IRFS. In practice, we model functions using either a set of basis functions or cubic splines and use AIC or BIC as scoring function. Our results show that Sloppy outperforms the state of the art algorithms with identifiability guarantees on synthetic and real world data and is on par with methods that do not have such guarantees. We note, however, that Sloppy is just a first instantiation and are quite certain it is possible to define—and are looking forward to see—instantiations of IRFS that will outperform our method in practice.

For future work, we would like to extend our framework to causal discovery, and are particularly interested how an instantiation based on a regularized deep neural network would perform.

ACKNOWLEDGMENTS

Alexander Marx is supported by the International Max Planck Research School for Computer Science (IMPRS-CS). Both authors are supported by the Cluster of Excellence on “Multimodal Computing and Interaction” within the Excellence Initiative of the German Federal Government.

REFERENCES

[1] Niels Henrik Abel. 1826. Démonstration de l‘impossibilité de la résolution algébrique des équations générales qui passent le quatrieme degré. *Journal für die reine und angewandte Mathematik* 1 (1826), 65–96.

[2] Hirotugu Akaike. 1983. Information measures and model selection. *Int Stat Inst* 44 (1983), 277–291.

[3] Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. 2018. Cause-Effect Inference by Comparing Regression Errors. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. 900–909.

[4] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* 42, 6 (2014), 2526–2556.

[5] David Deutsch. 1985. Quantum theory, the Church–Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A* 400, 1818 (1985), 97–117.

[6] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. 2008. A kernel statistical test of independence. 585–592.

[7] Peter Grünwald. 2007. *The Minimum Description Length Principle.* MIT Press.

[8] PO. Hoyer, D. Janzing, JM. Mooij, J. Peters, and B. Schölkopf. 2009. Nonlinear causal discovery with additive noise models. 689–696.

[9] Shoubo Hu, Zhitang Chen, Vahid Partovi Nia, Laiwan CHAN, and Yanhui Geng. 2018. Causal Inference and Mechanism Clustering of A Mixture of Additive Noise Models. In *Proceedings of the 32th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 5212–5222.

[10] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. 2012. Information-geometric approach to inferring causal directions. 182-183 (2012), 1–31.

[11] D. Janzing and B. Schölkopf. 2010. Causal Inference Using the Algorithmic Markov Condition. *IEEE Transactions on Information Technology* 56, 10 (2010), 5168–5194.

[12] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. 2018. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524* (2018).

[13] A.N. Kolmogorov. 1965. Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii* 1, 1 (1965), 3–11.

[14] M. Li and P. Vitányi. 1993. *An Introduction to Kolmogorov Complexity and its Applications.* Springer.

[15] Alexander Marx, Christina Backes, Eckart Meese, Hans-Peter Lenhof, and Andreas Keller. 2016. EDISON-WMW: Exact Dynamic Programming Solution of the Wilcoxon-Mann-Whitney Test. *Genomics, Proteomics & Bioinformatics* (2016).

[16] Alexander Marx and Jilles Vreeken. 2017. Telling Cause from Effect using MDL-based Local and Global Regression. IEEE, 307–316.

[17] Alexander Marx and Jilles Vreeken. 2018. Telling cause from effect by local and global regression. *Knowledge and Information Systems* (2018).

[18] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research* 17, 32 (2016), 1–102.

[19] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York, NY, USA.

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

[21] J. Peters, JM. Mooij, D. Janzing, and B. Schölkopf. 2014. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research* 15 (2014), 2009–2053.

[22] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. 2011. Identifiability of Causal Graphs Using Functional Models. In *Proceedings of the 27nd International Conference on Uncertainty in Artificial Intelligence (UAI).* AUAI Press, 589–598.

[23] Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6, 2 (1978), 461–464.

[24] Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. 2015. Inference of Cause and Effect with Unsupervised Inverse Regression. 38 (2015), 847–855.

[25] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* 7 (2006), 2003–2030.

[26] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. 2000. *Causation, prediction, and search.* MIT press.

[27] Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. 2010. Probabilistic latent variable models for distinguishing between cause and effect. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)* 26 (2010), 1687–1695.

[28] Natasa Tagasovska, Thibault Vatter, and Valérie Chavez-Demoulin. 2018. *Nonparametric Quantile-Based Causal Discovery.* Technical Report 1801.10579. arXiv.

[29] N.K. Vereshchagin and P.M.B. Vitányi. 2004. Kolmogorov’s Structure functions and model selection. *IEEE Transactions on Information Technology* 50, 12 (2004), 3265– 3290.

[30] Thomas Verma and Judea Pearl. 1991. Equivalence and Synthesis of Causal Models. 255–270.

[31] Kun Zhang and Aapo Hyvärinen. 2009. On the Identifiability of the Post-nonlinear Causal Model. AUAU Press, 647–655.

A APPENDIX

In this section, we first clarify how exactly we applied the competing methods in our experiments. Then we explain in which configuration we applied Sloppy and analyze the differences between Sloppy, RECI, CAM and Slope.

A.1 Configuration for Competing Methods

For RESIT and QCCD, we used the default configurations as recommended by the authors [21, 28]. Before we applied IGCI to the synthetic data sets, we standardized X and Y to have zero mean and unit variance. As for all of the simulated data sets the cause was generated as a Gaussian or near Gaussian distributed random variable this preprocessing step led to better results than standardizing the data. However, when we applied IGCI to the Tübingen data set, we found that normalizing the data between zero and one led to better results, hence we reported those results.

A.2 Configuration of Sloppy

We implemented Sloppy using cubic splines, as described in Section 4.2. Equivalent to the preprocessing that we did for IGCI, we standardized X and Y to have zero mean and unit variance for the simulated data sets, as for those we knew that the cause was generated with a Gaussian or near Gaussian distribution. Since we did not know the distributions for the real-world data sets beforehand, we choose to use a uniform prior and normalized the data between zero and one for the Tübingen data set. As scoring function we used BIC for the simulated data pairs. For the normalized real-world data sets, however, BIC was too restrictive and mainly fitted linear models. Hence, we used AIC for these.

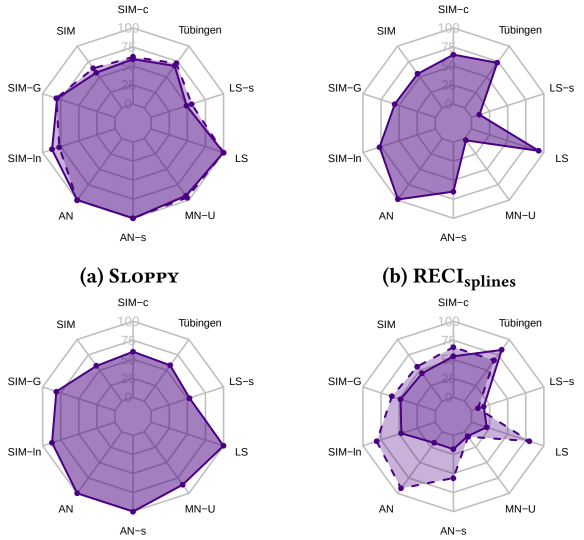
When we apply Sloppy$_B$ with the same configuration, we obtain similar results, as shown in Figure 8.

A.3 Comparison to RECI, CAM and Slope

In the following, we are going to explain how we needed to configure Sloppy to obtain similar results to RECI, CAM and Slope and briefly discuss the differences to the results that we presented in the main part of the paper.

A.3.1 RECI. As RECI assumes that the true functions are known, it is hard to do a fair comparison without preselecting for a suitable regressor [3]. To provide an impression of the results, we preprocessed the data by normalizing it between zero and one (as suggested by the authors) and then applied Sloppy$_S$ with zero penalty for the parameters. First of all, we observe that the splines strongly overfit, where the average number of degrees of freedom is over 140, in contrast to Sloppy, where the average number of degrees of freedom is 5. Nonetheless, the results on the synthetic and benchmark data are still reasonable, as shown in Figure 8. The overall average performance, however, drops from 81% to 60%. Since the authors of RECI suggest to only fit low degree polynomials [3], splines are, however, a sub-optimal choice.

A.3.2 CAM. In the bivariate setting, CAM is very related to Sloppy. CAM also uses regularized splines and standardizes the data, where they maximize the log-likelihood for both directions [4]. Therefore, it is not surprising, that the results obtained with CAM are very similar to the results we get with Sloppy, as shown in Figure 8. However, as mentioned in Section 4, when CAM was developed, the authors only showed consistency of their method and did not have strong identifiability results. In addition, when we compare the decision rates of CAM and Sloppy on the *Tübingen* benchmark data set (see Figure 9), we see that Sloppy clearly outperforms CAM.



(a) Sloppy

(b) RECI$_{\text{splines}}$

(c) CAM

(d) Slope

Figure 8: Accuracy of Sloppy (solid: Sloppy$_B$, dashed: Sloppy$_S$), RECI (using cubic splines), CAM and Slope (solid: Slope allowing for non-deterministic functions, dashed: Slope using a mixture of deterministic basis functions) over all synthetic data set and the *Tübingen* benchmark data set.

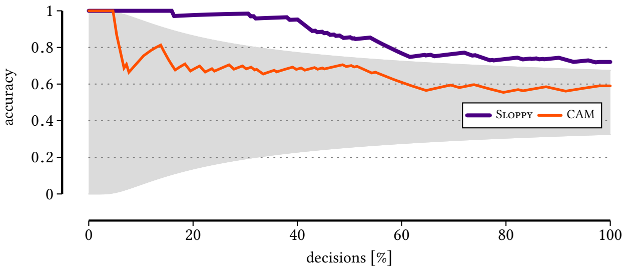


Figure 9: Decision rates for Sloppy and CAM on the *Tübingen* benchmark data set.

A.3.3 Slope. For Slope, the authors also standardize the data between zero and one. In particular, there exist two versions: Slope using a deterministic function and allowing for non-deterministic functions [16] and Sloper using a set of basis functions, without fitting non-deterministic functions [17]. Apart from the exact score and the preprocessing, Sloper comes close to Sloppy$_B$. When we look at the results over all data sets (Figure 8), we see that Sloper performs similar to RECI using cubic splines. On average, Slope performs much worse than Sloper and only has a better performance on the *Tübingen* benchmark data set.