Analysis of cause-effect inference by comparing regression errors

Patrick Blöbaum[1], Dominik Janzing[2], Takashi Washio[1], Shohei Shimizu[3] and Bernhard Schölkopf[2]

[1] Osaka University, Osaka, Japan
[2] MPI for Intelligent Systems, Tübingen, Germany
[3] Shiga University, Shiga, Japan

ABSTRACT

We address the problem of inferring the causal direction between two variables by comparing the least-squares errors of the predictions in both possible directions. Under the assumption of an independence between the function relating cause and effect, the conditional noise distribution, and the distribution of the cause, we show that the errors are smaller in causal direction if both variables are equally scaled and the causal relation is close to deterministic. Based on this, we provide an easily applicable algorithm that only requires a regression in both possible causal directions and a comparison of the errors. The performance of the algorithm is compared with various related causal inference methods in different artificial and real-world data sets.

Subjects Artificial Intelligence, Data Mining and Machine Learning
Keywords Causality, Causal discovery, Cause-effect inference

Submitted 18 September 2018
Accepted 16 December 2018
Published 21 January 2019

Corresponding author
Patrick Blöbaum,
bloebaum@ar.sanken.osaka-u.ac.jp

Academic editor
Charles Elkan

Additional Information and Declarations can be found on page 26

DOI **10.7717/peerj-cs.169**

INTRODUCTION

Causal inference (*Spirtes, Glymour & Scheines, 2000*; *Pearl, 2009*) is becoming an increasingly popular topic in machine learning. The results are often not only of interest in predicting the result of potential interventions, but also in general statistical and machine learning applications (*Peters, Janzing & Schölkopf, 2017*). While the causal relationship between variables can generally be discovered by performing specific randomized experiments, such experiments can be very costly, infeasible or unethical[1]. In particular, the identification of the causal direction between two variables without performing any interventions is a challenging task. However, recent research developments in causal discovery allow, under certain assumptions, inference of the causal direction between two variables purely based on observational data (*Kano & Shimizu, 2003*; *Comley & Dowe, 2003*; *Shimizu et al., 2006*; *Sun, Janzing & Schölkopf, 2006*; *Zhang & Hyvärinen, 2009*; *Hoyer et al., 2009*; *Janzing, Sun & Schölkopf, 2009*; *Daniušis et al., 2010*; *Peters, Janzing & Schölkopf, 2011*; *Janzing et al., 2012*; *Sgouritsa et al., 2015*; *Mooij et al., 2016*; *Marx & Vreeken, 2017*). In regard to the present work, we further contribute to the causal discovery in an unconfounded bivariate setting based on observational data, where one variable is the cause and the other variable is the effect. That is, given observed data X,Y that are drawn from a joint distribution $p_{X,Y}$, we are interested in inferring whether X caused Y or Y caused X. In this sense, we define X as the cause and Y as the effect if intervening on X

How to cite this article Blöbaum P, Janzing D, Washio T, Shimizu S, Schölkopf B. 2019. Analysis of cause-effect inference by comparing regression errors. *PeerJ Comput. Sci.* **5**:e169 http://doi.org/10.7717/peerj-cs.169

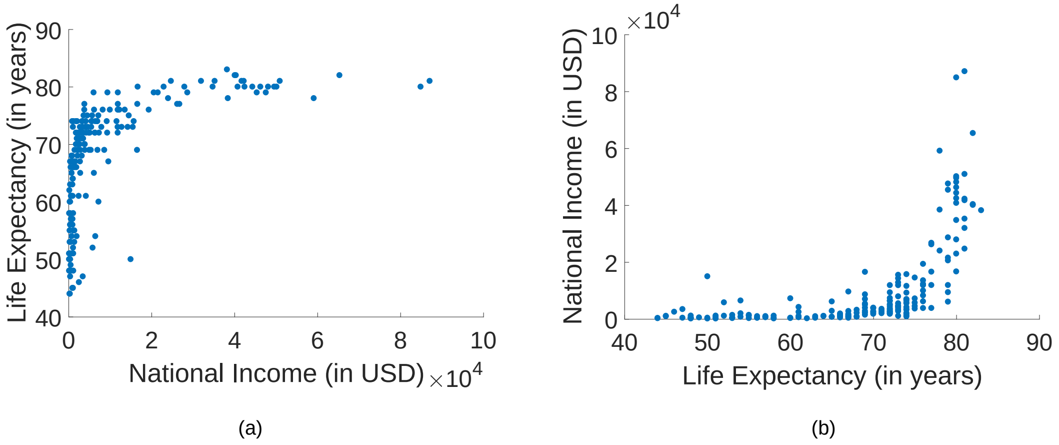


Figure 1 A comparison of the national income of 194 countries and the life expectancy at birth. (A) The national income on the x-axis and the life expectancy on the y-axis. (B) The life expectancy on the x-axis and the national income on the y-axis.

Full-size DOI: 10.7717/peerjcs.169/fig-1

changes the distribution of Y. In the following, we use the term ’causal inference’ to refer to the identification of the true causal direction.

A possible application is the discovery of molecular pathways, which relies on the identification of causal molecular interactions in genomics data (*Statnikov et al., 2012*). Other examples in biomedicine where observational data can be used for causal discovery are discussed in the work by *Ma & Statnikov (2017)*. An example for a bivariate relationship is provided in Fig. 1, where the national income of countries are compared with the life expectancy at birth[2]. Here, a clear statement about the causal relationship is not obvious. It has been argued that richer countries have a better health care system than poorer countries. Hence, a higher national income leads to a higher life expectancy (*Mooij et al., 2016*). Based on the plots, this causal relationship is not clear at all. Nevertheless, we provide a way to correctly determine the causal direction by only using these data points.

Conventional approaches to causal inference rely on conditional independences and therefore require at least three observed variables. Given the observed pattern of conditional dependences and independences, one infers a class of directed acyclic graphs (DAGs) that is compatible with the respective pattern (subject to Markov condition and faithfulness assumption (*Spirtes, Glymour & Scheines, 2000*; *Pearl, 2009*)). Whenever there are causal arrows that are common to all DAGs in the class, conditional (in)dependences yield definite statements about causal directions. In a bivariate setting, however, we rely on asymmetries between cause and effect that are already apparent in the bivariate distribution alone.

One kind of asymmetry is given by restricting the structural equations relating cause and effect to a certain function class: For linear relations with non-Gaussian independent noise, the linear non-Gaussian acyclic model (LiNGAM) (*Shimizu et al., 2006*) provides a method to identify the correct causal direction. For nonlinear relations, the additive noise model (ANM) (*Hoyer et al., 2009*) and its generalization to post-nonlinear models (PNL) (*Zhang & Hyvärinen, 2009*) identify the causal direction by assuming an independence

[1] Further discussions about ethics in randomized experiments, especially in the context of clinical trials, can be found in *Rosner (1987)*.

[2] The data is taken from https://webdav.tuebingen.mpg.de/cause-effect/ and further discussed in *Mooij et al. (2016)*.

between cause and noise, where, apart from some exceptions such as bivariate Gaussian, a model can only be fit in the correct causal direction such that the input is independent of the residual.

Further recent approaches for the bivariate setting are based on an *informal* independence assumption stating that the distribution of the cause (denoted by p_C) contains no information about the conditional distribution of the effect given the cause (denoted by $p_{E|C}$). Here, the formalization of 'no information' is a challenging task. For the purpose of foundational insights (rather than for practical purposes), *Janzing & Schölkopf (2010)* and *Lemeire & Janzing (2012)* formalize the idea via *algorithmic information* and postulate that knowing p_C does not enable a shorter description of $p_{E|C}$ and vice versa. Using algorithmic information theory, one can, for instance, show that the algorithmic independence of p_C and $p_{E|C}$ implies

$$K(p_C)+K(p_{E|C}) \leq K(p_E)+K(p_{C|E}), \tag{1}$$

if K denotes the description length of a distribution in terms of its Kolmogorov complexity (for details see Section 4.1.9 in *Peters, Janzing & Schölkopf (2017)*). In this sense, appropriate independence assumptions between p_C and $p_{E|C}$ imply that $p_{E,C}$ has a simpler description in causal direction than in anticausal direction. An approximation of (1) is given by the SLOPE algorithm in the work by *Marx & Vreeken (2017)*, where regression is utilized to estimate and compare the approximated Kolmogorov complexities. For this, a logarithmic error is used, which is motivated by a minimum description length perspective. Another work that is inspired by the independence assumption is the information-geometric approach for causal inference (IGCI) (*Janzing et al., 2012*). IGCI provides a method to infer the causal direction in deterministic nonlinear relationships subject to a certain independence condition between the slope of the function and the distribution of the cause. A related but different independence assumption is also used by a technique called unsupervised inverse regression (CURE) (*Sgouritsa et al., 2015*), where the idea is to estimate a prediction model of both possible causal directions in an unsupervised manner, i.e., only the input data is used for the training of the prediction models. With respect to the above independence assumption, the effect data may contain information about the relation between cause and effect that can be employed for predicting the cause from the effect, but the cause data alone does not contain any information that helps the prediction of the effect from the cause (as hypothesized in *Schölkopf et al. (2013)*). Accordingly, the unsupervised regression model in the true causal direction should be less accurate than the prediction model in the wrong causal direction.

For our approach, we address the causal inference problem by exploiting an asymmetry in the mean-squared error (MSE) of predicting the cause from the effect and the effect from the cause, respectively, and show, that under appropriate assumptions and in the regime of almost deterministic relations, the prediction error is smaller in causal direction. A preliminary version of this idea can be found in *Blöbaum, Washio & Shimizu (2017)* and *Blöbaum, Shimizu & Washio (2017)* but in these works the analysis is based on a simple heuristic assuming that the regression of Y on X and the regression of X on Y yield functions that are inverse to each other, which holds approximately in the limit of small

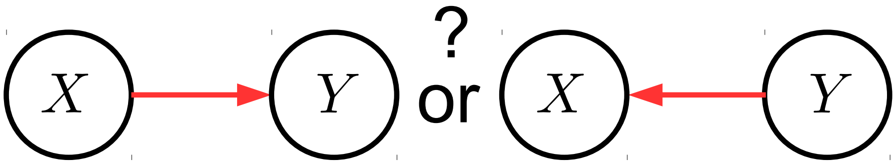


Figure 2 An illustration of the goal of our proposed method. It aims to identify the causal DAG of two variables, where either X causes Y or Y causes X.

Full-size DOI: 10.7717/peerjcs.169/fig-2

noise. Moreover, the analysis is also based on the assumption of an additive noise model in causal direction and on having prior knowledge about the functional relation between X and Y, which makes it impractical for generic causal inference problems.

In this work, we aim to generalize and extend the two aforementioned works in several ways: (1) We explicitly allow a dependency between cause and noise. (2) We give a proper mathematical proof of the theory that justifies the method subject to clear formal assumptions. (3) We perform extensive evaluations for the application in causal inference and compare it with various related approaches. The theorem stated in this work might also be of interest for general statistical purposes. A briefer version of this work with less extensive experiments, lesser details and without detailed proofs can be found in *Blöbaum et al. (2018)*.

This paper is structured as follows: In 'Preliminaries', we define the problem setting and introduce the used notations and assumptions, which are necessary for the main theorem of this work stated in 'Theory'. An algorithm that utilizes this theorem is proposed in Algorithm and evaluated in various artificial and real-world data sets in 'Experiments'.

PRELIMINARIES

In the following, we introduce the preliminary problem setting, notations and assumptions.

Problem setting and notation

In this work, we use the framework of structural causal models (*Pearl, 2009*) with the goal of correctly identifying cause and effect variables of given observations from X and Y. As illustrated in Fig. 2, this can be described by the problem of identifying whether the causal DAG of $X \rightarrow Y$ or $X \leftarrow Y$ is true. Throughout this paper, a capital letter denotes a random variable and a lowercase letter denotes values attained by the random variable. Variables X and Y are assumed to be real-valued and to have a joint probability density (with respect to the Lebesgue measure), denoted by $p_{X,Y}$. By slightly abusing terminology, we will not further distinguish between a distribution and its density since the Lebesgue measure as a reference is implicitly understood. The notations p_X, p_Y, and $p_{Y|X}$ are used for the corresponding marginal and conditional densities, respectively. The derivative of a function f is denoted by f'.

General idea

As mentioned before, the general idea of our approach is to simply compare the MSE of regressing Y on X and the MSE of regressing X on Y. If we denote cause and effect by $C, E \in \{X, Y\}$, respectively, our approach explicitly reads as follows. Let ϕ denote the function that minimizes the expected least squares error when predicting E from C, which implies that ϕ is given by the conditional expectation $\phi(c) = \mathbb{E}[E|c]$. Likewise, let ψ be the minimizer of the least squares error for predicting C from E, that is, $\psi(e) = \mathbb{E}[C|e]$. Then we will postulate assumptions that imply

$$\mathbb{E}[(E-\phi(C))^2] \leq \mathbb{E}[(C-\psi(E))^2] \tag{2}$$

in the regime of almost deterministic relations. This conclusion certainly relies on some kind of scaling convention. For our theoretical results we will assume that both X and Y attain values between 0 and 1. However, in some applications, we will also scale X and Y to unit variance to deal with unbounded variables. Equation (2) can be rewritten in terms of conditional variance as

$$\mathbb{E}[\mathrm{Var}[E|C]] \leq \mathbb{E}[\mathrm{Var}[C|E]].$$

Assumptions

First, recall that we assume throughout the paper that either X is the cause of Y or vice versa in an unconfounded sense, i.e., there is no common cause. Therefore, the general structural equation is defined as

$$E = \zeta(C, \tilde{N}), \tag{3}$$

where $C \perp\!\!\!\perp \tilde{N}$. For our analysis, we first define a function ϕ to be the conditional expectation of the effect given the cause, i.e.,

$$\phi(c) := \mathbb{E}[E|c]$$

and, accordingly, we define a noise variable N as the residual

$$N := E - \phi(C). \tag{4}$$

Note that (4) implies that $\mathbb{E}[N|c] = 0$. The function ϕ is further specified below. Then, to study the limit of an almost deterministic relation in a mathematically precise way, we consider a family of effect variables E_α by

$$E_\alpha := \phi(C) + \alpha N, \tag{5}$$

where $\alpha \in \mathbb{R}^+$ is a parameter controlling the noise level and N is a noise variable that has some (upper bounded) joint density $p_{N,C}$ with C. Note that N here does not need to be statistically independent of C (in contrast to ANMs), which allows the noise to be non-additive. Therefore, (5) does not, a priori, restrict the set of possible causal relations, because for any pair (C, E) one can always define the noise N as (4) and thus obtain $E_{\alpha=1} = E$ for any arbitrary function ϕ[3].

For this work, we make use of the following assumptions:

1. **Invertible function:** ϕ is a strictly monotonically increasing two times differentiable function $\phi : [0,1] \rightarrow [0,1]$. For simplicity, we assume that ϕ is monotonically increasing with $\phi(0) = 0$ and $\phi(1) = 1$ (similar results for monotonically decreasing functions follow by reflection $E \rightarrow 1 - E$). We also assume that $\phi^{-1'}$ is bounded.

[3] Note that although the form of (5) is similar to that of ANM, the core assumption of ANM is an independence between cause and noise, which we do not need in our approach. Therefore, we assume the general structural equation defined in (3), whereas ANM assumes a more restrictive structural equation of the form $E = \zeta(C) + \tilde{N}$ with $C \perp\!\!\!\perp \tilde{N}$.

2. **Compact supports:** The distribution of C has compact support. Without loss of generality, we assume that 0 and 1 are, respectively, the smallest and the largest values attained by C. We further assume that the distribution of N has compact support and that there exist values $n_+ > 0 > n_-$ such that for any c, $[n_-, n_+]$ is the smallest interval containing the support of $p_{N|c}$. This ensures that we know $[\alpha n_-, 1+\alpha n_+]$ is the smallest interval containing the support of p_{E_α}. Then the shifted and rescaled variable
$$\tilde{E}_\alpha := \frac{1}{1+\alpha n_+ - \alpha n_-}(E_\alpha - \alpha n_-) \tag{6}$$
attains 0 and 1 as minimum and maximum values and thus is equally scaled as C.
3. **Unit noise variance:** The expected conditional noise variance is $\mathbb{E}[\mathrm{Var}[N|C]] = 1$ without loss of generality, seeing that we can scale the noise arbitrary by the parameter α and we are only interested in the limit $\alpha \to 0$.
4. **Independence postulate:** While the above assumptions are just technical, we now state the essential assumption that generates the asymmetry between cause and effect. To this end, we consider the unit interval $[0,1]$ as probability space with uniform distribution as probability measure. The functions $c \mapsto \phi'(c)$ and $c \mapsto \mathrm{Var}[N|c]p_C(c)$ define random variables on this space, which we postulate to be uncorrelated, formally stated as
$$\mathrm{Cov}[\phi', \mathrm{Var}[N|c]p_C] = 0. \tag{7}$$
More explicitly, (7) reads:
$$\int_0^1 \phi'(c)\mathrm{Var}[N|c]p_C(c)dc - \int_0^1 \phi'(c)dc \int_0^1 \mathrm{Var}[N|c]p_C(c)dc = 0. \tag{8}$$

The justification of (7) is not obvious at all. For the special case where the conditional variance $\mathrm{Var}[N|c]$ is a constant in c (e.g., for ANMs), (7) reduces to

$$\mathrm{Cov}[\phi', p_C] = 0, \tag{9}$$

which is an independence condition for deterministic relations stated in *Schölkopf et al. (2013)*. Conditions of similar type as (9) have been discussed and justified in *Janzing et al. (2012)*. They are based on the idea that ϕ contains no information about p_C. This, in turn, relies on the idea that the conditional $p_{E|C}$ contains no information about p_C.

To discuss the justification of (8), observe first that it *cannot* be justified as stating some kind of 'independence' between p_C and $p_{E|C}$. To see this, note that (8) states an uncorrelatedness of the two functions $c \mapsto \phi'(c)$ and $c \mapsto \mathrm{Var}[N|c]p_C(c)$. While ϕ' depends only on the conditional $p_{E|C}$ and not on p_C, the second function depends on both $p_{C|E}$ and p_E, since $\mathrm{Var}[N|c]$ is a property of $p_{E|C}$. Nevertheless, to justify (8) we assume that the function ϕ represents a law of nature that persists when p_C and N change due to changing background conditions. From this perspective, it becomes unlikely that they are related to the background condition at hand. This idea follows the general spirit of 'modularity and autonomy' in structural equation modeling, that some structural equations may remain unchanged when other parts of a system change (see Chapter 2 in *Peters, Janzing & Schölkopf (2017)* for a literature review)[4]. To further justify (7), one could think of a scenario where someone changes ϕ independently of $p_{N,C}$, which then results in vanishing correlations. Typically, this assumption would be violated if ϕ is adjusted to $p_{N,C}$ or vice versa. This could happen due to an intelligent design by, for instance, first

[4] Note, however, that the assignment (5) is not a structural equation in a strict sense, because then C and N would need to be statistically independent.

observing $p_{N,C}$ and then defining ϕ or due to a long adaption process in nature (see Janzing et al. (2012) for further discussions of possible violations in a deterministic setting).

A simple implication of (8) reads

$$\int_0^1 \phi'(c)\mathrm{Var}[N|c]p_C(c)dc = 1, \tag{10}$$

due to $\int_0^1 \phi'(c)dc = 1$ and $\int_0^1 \mathrm{Var}[N|c]p_C(c)dc = \mathbb{E}[\mathrm{Var}[N|C]] = 1$.

In the following, the term *independence postulate* is used to refer to the aforementioned postulate and the term *independence* to a statistical independence, which should generally become clear from the context.

THEORY

As introduced in 'General idea', we aim to exploit an inequality of the expected prediction errors in terms of $\mathbb{E}[\mathrm{Var}[E|C]] \leq \mathbb{E}[\mathrm{Var}[C|E]]$ to infer the causal direction. In order to conclude this inequality and, thus, to justify an application to causal inference, we must restrict our analysis to the case where the noise variance is sufficiently small, since a more general statement is not possible under the aforementioned assumptions. The analysis can be formalized by the ratio of the expectations of the conditional variances in the limit $\alpha \to 0$.

We will then show

$$\lim_{\alpha\to 0} \frac{\mathbb{E}[\mathrm{Var}[C|\tilde{E}_\alpha]]}{\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]} \geq 1.$$

Error asymmetry theorem

For our main theorem, we first need an important lemma:

Lemma 1 (Limit of variance ratio) *Let the assumptions 1–3 in 'Assumptions' hold. Then the following limit holds:*

$$\lim_{\alpha\to 0} \frac{\mathbb{E}[\mathrm{Var}[C|\tilde{E}_\alpha]]}{\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]} = \int_0^1 \frac{1}{\phi'(c)^2}\mathrm{Var}[N|c]p_C(c)dc \tag{11}$$

Proof: We first give some reminders of the definition of the conditional variance and some properties. For two random variables Z and Q the conditional variance of Z, given q is defined by

$$\mathrm{Var}[Z|q] := \mathbb{E}[(Z - \mathbb{E}[Z|q])^2|q],$$

while $\mathrm{Var}[Z|Q]$ is the random variable attaining the value $\mathrm{Var}[Z|q]$ when Q attains the value q. Its expectation reads

$$\mathbb{E}[\mathrm{Var}[Z|Q]] := \int \mathrm{Var}[Z|q]p_Q(q)dq.$$

For any $a \in \mathbb{R}$, we have

$$\mathrm{Var}\left[\frac{Z}{a}|q\right] = \frac{\mathrm{Var}[Z|q]}{a^2}.$$

For any function h, we have

$$\mathrm{Var}[h(Q)+Z|q] = \mathrm{Var}[Z|q],$$

which implies $\mathrm{Var}[h(Q)|q] = 0$. Moreover, we have

$$\mathrm{Var}[Z|h(q)] = \mathrm{Var}[Z|q],$$

if h is invertible.

To begin the main part of the proof, we first observe

$$\mathbb{E}[\mathrm{Var}[E_\alpha|C]] = \mathbb{E}[\mathrm{Var}[\phi(C)+\alpha N|C]] = \alpha^2 \underbrace{\mathbb{E}[\mathrm{Var}[N|C]]}_{=\,1\text{ (Assumpt. 3))}} = \alpha^2. \tag{12}$$

Moreover, one easily verifies that

$$\lim_{\alpha\to 0}\frac{\mathbb{E}[\mathrm{Var}[C|\tilde{E}_\alpha]]}{\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]} = \lim_{\alpha\to 0}\frac{\mathbb{E}[\mathrm{Var}[C|E_\alpha]]}{\mathbb{E}[\mathrm{Var}[E_\alpha|C]]}, \tag{13}$$

due to (6) provided that these limits exist. Combining Eqs. (12) and (13) yields

$$\lim_{\alpha\to 0}\frac{\mathbb{E}[\mathrm{Var}[C|\tilde{E}_\alpha]]}{\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]} = \lim_{\alpha\to 0}\frac{\mathbb{E}[\mathrm{Var}[C|E_\alpha]]}{\alpha^2} = \lim_{\alpha\to 0}\mathbb{E}\left[\mathrm{Var}\left[\frac{C}{\alpha}|E_\alpha\right]\right]. \tag{14}$$

Now, we can rewrite (14) as

$$\begin{aligned}\lim_{\alpha\to 0}\mathbb{E}\left[\mathrm{Var}\left[\frac{C}{\alpha}|E_\alpha\right]\right] &= \lim_{\alpha\to 0}\mathbb{E}\left[\mathrm{Var}\left[\frac{\phi^{-1}(E_\alpha-\alpha N)}{\alpha}|E_\alpha\right]\right]\\ &= \lim_{\alpha\to 0}\int_{\phi(0)+\alpha n_-}^{\phi(1)+\alpha n_+}\mathrm{Var}\left[\frac{\phi^{-1}(e-\alpha N)}{\alpha}|e\right]p_{E_\alpha}(e)de\\ &= \lim_{\alpha\to 0}\int_{\phi(0)}^{\phi(1)}\mathrm{Var}\left[\frac{\phi^{-1}(e-\alpha N)}{\alpha}|e\right]p_{E_\alpha}(e)de. \end{aligned}\tag{15}$$

In the latter step, αn_+ and $-\alpha n_-$ vanishes in the limit seeing that the function

$$e \mapsto \mathrm{Var}\left[\phi^{-1}(e-\alpha N)/\alpha\,|e\right]p_{E_\alpha}(e)$$

is uniformly bounded in α. This is firstly, because ϕ^{-1} attains only values in $[0,1]$, and hence the variance is bounded by 1. Secondly, $p_{E_\alpha}(e)$ is uniformly bounded due to

$$\begin{aligned}p_{E_\alpha}(e) = \int_{n_-}^{n_+} p_{\phi(C),N}(e-\alpha n, n)dn = \int_{n_-}^{n_+} p_{C,N}(\phi^{-1}(e-\alpha n),n)\phi^{-1'}(e-\alpha n)dn\\ \leq \|\phi^{-1'}\|_\infty\|p_{C,N}\|_\infty(n_+-n_-).\end{aligned}$$

Accordingly, the bounded convergence theorem states

$$\lim_{\alpha\to 0}\int_{\phi(0)}^{\phi(1)}\mathrm{Var}\left[\frac{\phi^{-1}(e-\alpha N)}{\alpha}|e\right]p_{E_\alpha}(e)de = \int_{\phi(0)}^{\phi(1)}\lim_{\alpha\to 0}\left(\mathrm{Var}\left[\frac{\phi^{-1}(e-\alpha N)}{\alpha}|e\right]p_{E_\alpha}(e)\right)de.$$

To compute the limit of

$$\mathrm{Var}\left[\frac{\phi^{-1}(e-\alpha N)}{\alpha}|e\right],$$

we use Taylor's theorem to obtain

$$\phi^{-1}(e-\alpha n)=\phi^{-1}(e)-\alpha n\phi^{-1'}(e)-\frac{\alpha^2 n^2\phi^{-1''}(E_2(n,e))}{2}, \tag{16}$$

where $E_2(n,e)$ is a real number in the interval $(e-\alpha n,e)$. Since (16) holds for every $n\in[-\frac{e}{\alpha},\frac{1-e}{\alpha}]$ (note that ϕ and ϕ^{-1} are bijections of $[0,1]$, thus $e-\alpha n$ lies in $[0,1]$) it also holds for the random variable N if $E_2(n,e)$ is replaced with the random variable $E_2(N,e)$ (here, we have implicitly assumed that the map $n\mapsto e_2(n,e)$ is measurable). Therefore, we see that

$$\lim_{\alpha\to 0}\mathrm{Var}\left[\frac{\phi^{-1}(e-\alpha N)}{\alpha}\middle|e\right]=\lim_{\alpha\to 0}\mathrm{Var}\left[-N\phi^{-1'}(e)-\frac{\alpha N^2\phi^{-1''}(E_2(N,e))}{2}\middle|e\right]$$
$$=\phi^{-1'}(e)^2\mathrm{Var}[N|e]. \tag{17}$$

Moreover, we have

$$\lim_{\alpha\to 0}p_{E_\alpha}(e)=p_{E_0}(e). \tag{18}$$

Inserting Eqs. (18) and (17) into (15) yields

$$\begin{aligned}\lim_{\alpha\to 0}\mathbb{E}\left[\mathrm{Var}\left[\frac{C}{\alpha}\middle|E_0\right]\right]&=\int_{\phi(0)}^{\phi(1)}\phi^{-1'}(e)^2\mathrm{Var}[N|e]p_{E_0}(e)de\\&=\int_0^1\phi^{-1'}(\phi(c))^2\mathrm{Var}[N|\phi(c)]p_C(c)dc\\&=\int_0^1\frac{1}{\phi'(c)^2}\mathrm{Var}[N|c]p_C(c)dc,\end{aligned}$$

where the second equality is a variable substitution using the deterministic relation $E_0=\phi(C)$ (which implies $p_{E_0}(\phi(c))=p_C(c)/\phi'(c)$ or, equivalently, the simple symbolic equation $p_{E_0}(e)de=p_C(c)dc$). This completes the proof due to (14). □

While the formal proof is a bit technical, the intuition behind this idea is quite simple: just think of the scatter plot of an almost deterministic relation as a thick line. Then $\mathrm{Var}[E_\alpha|c]$ and $\mathrm{Var}[C|E_\alpha=\phi(c)]$ are roughly the squared widths of the line at some point $(c,\phi(c))$ measured in vertical and horizontal direction, respectively. The quotient of the widths in vertical and horizontal direction is then given by the slope. This intuition yields the following approximate identity for small α:

$$\mathrm{Var}[C|\tilde{E}_\alpha=\phi(c)]\approx\frac{1}{(\phi'(c))^2}\mathrm{Var}[\tilde{E}_\alpha|C=c]=\alpha^2\frac{1}{(\phi'(c))^2}\mathrm{Var}[N|c]. \tag{19}$$

Taking the expectation of (19) over C and recalling that Assumption 3 implies $\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]=\alpha^2\mathbb{E}[\mathrm{Var}[N|C]]=\alpha^2$ already yields (11).

With the help of Lemma 1, we can now formulate the core theorem of this paper:

Theorem 1 (Error Asymmetry) *Let the assumptions 1–4 in 'Assumptions' hold. Then the following limit always holds*

$$\lim_{\alpha\to 0}\frac{\mathbb{E}[\mathrm{Var}[C|\tilde{E}_\alpha]]}{\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]}\geq 1,$$

with equality only if the function stated in Assumption 1 is linear.

Proof: We first recall that Lemma 1 states

$$\lim_{\alpha\to 0}\frac{\mathbb{E}[\mathrm{Var}[C|\tilde{E}_\alpha]]}{\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]}=\int_0^1\frac{1}{\phi'(c)^2}\mathrm{Var}[N|c]p_C(c)dc.$$

We then have

$$\begin{aligned}
&\int_0^1\frac{1}{\phi'(c)^2}\mathrm{Var}[N|c]p_C(c)dc\\
&=\int_0^1\frac{1}{\phi'(c)^2}\mathrm{Var}[N|c]p_C(c)dc\cdot\underbrace{\int_0^1\mathrm{Var}[N|c]p_C(c)dc}_{=\,1\text{ (Assumpt. 3)}}\\
&=\int_0^1\overline{\sqrt{\left(\frac{1}{\phi'(c)}\right)^2\mathrm{Var}[N|c]}}^{\,2}p_C(c)dc\cdot\int_0^1\overline{\sqrt{\mathrm{Var}[N|c]}}^{\,2}p_C(c)dc\\
&\geq\left(\int_0^1\sqrt{\left(\frac{1}{\phi'(c)}\right)^2\mathrm{Var}[N|c]}\sqrt{\mathrm{Var}[N|c]}p_C(c)dc\right)^2\\
&=\left(\int_0^1\frac{1}{\phi'(c)}\mathrm{Var}[N|c]p_C(c)dc\right)^2,
\end{aligned}\tag{20}$$

where the inequality is just the Cauchy Schwarz inequality applied to the bilinear form $f,g\mapsto\int f(c)g(c)p_C(c)dc$ for the space of functions f for which $\int f^2(c)p_c(c)dc$ exists. Note that if ϕ is linear, (20) becomes 1, since $\phi'=1$ according to Assumpt. 1. We can make a statement about (20) in a similar way by using (10) implied by the independence postulate and using Cauchy Schwarz:

$$\begin{aligned}
&\int_0^1\frac{1}{\phi'(c)}\mathrm{Var}[N|c]p_C(c)dc\\
&=\int_0^1\frac{1}{\phi'(c)}\mathrm{Var}[N|c]p_C(c)dc\cdot\underbrace{\int_0^1\phi'(c)\mathrm{Var}[N|c]p_C(c)dc}_{=\,1\text{ (10)}}\\
&=\int_0^1\overline{\sqrt{\frac{1}{\phi'(c)}\mathrm{Var}[N|c]}}^{\,2}p_C(c)dc\cdot\int_0^1\overline{\sqrt{\phi'(c)\mathrm{Var}[N|c]}}^{\,2}p_C(c)dc\\
&\geq\left(\int_0^1\sqrt{\frac{1}{\phi'(c)}\mathrm{Var}[N|c]}\sqrt{\phi'(c)\mathrm{Var}[N|c]}p_C(c)dc\right)^2\\
&=\left(\underbrace{\int_0^1\mathrm{Var}[N|c]p_C(c)dc}_{=\,1\text{ (Assumpt. 3)}}\right)^2=1.
\end{aligned}\tag{21}$$

Combining Eqs. (20) and (21) with Lemma 1 completes the proof. □

Remark

Theorem 1 states that the inequality holds for all values of α smaller than a certain finite threshold. Whether this threshold is small or whether the asymmetry with respect to regression errors already occurs for large noise cannot be concluded from the theoretical insights. Presumably, this depends on the features of ϕ, p_C, $p_{N|C}$ in a complicated way. However, the experiments in 'Experiments' suggest that the asymmetry often appears even for realistic noise levels.

If the function ϕ is non-invertible, there is an information loss in anticausal direction, since multiple possible values can be assigned to the same input. Therefore, we can expect that the error difference becomes even higher in these cases, which is supported by the experiments in 'Simulated cause–effect pairs with strong dependent noise'.

ALGORITHM

A causal inference algorithm that exploits Theorem 1 can be formulated in a straightforward manner. Given observations X, Y sampled from a joint distribution $p_{X,Y}$, the key idea is to fit regression models in both possible directions and compare the MSE. We call this approach Regression Error based Causal Inference (RECI) and summarize the algorithm in Algorithm 1.

Although estimating the conditional expectations $\mathbb{E}[Y|X]$ and $\mathbb{E}[X|Y]$ by regression is a standard task in machine learning, we should emphasize that the usual issues of over- and underfitting are critical for our purpose (like for methods based on ANMs or PNLs), because they under- or overestimate the noise levels. It may, however, happen that the method even benefits from underfitting: if there is a simple regression model in causal direction that fits the data quite well, but in anticausal relation the conditional expectation becomes more complex, a regression model with underfitting increases the error even more for the anticausal direction than for the causal direction.

Algorithm 1 The proposed causal inference algorithm.

```
function RECI($X$, $Y$)                                ▷ $X$ and $Y$ are the observed data.
    $(X, Y) \leftarrow$ RescaleData$(X, Y)$
    $f \leftarrow$ FitModel$(X, Y)$                    ▷ Fit regression model $f: X \rightarrow Y$
    $g \leftarrow$ FitModel$(Y, X)$                    ▷ Fit regression model $g: Y \rightarrow X$
    MSE$_{Y|X} \leftarrow$ MeanSquaredError$(f, X, Y)$
    MSE$_{X|Y} \leftarrow$ MeanSquaredError$(g, Y, X)$
    if MSE$_{Y|X} <$ MSE$_{X|Y}$ then
        return $X$ causes $Y$
    else if MSE$_{X|Y} <$ MSE$_{Y|X}$ then
        return $Y$ causes $X$
    else
        return No decision
    end if
end function
```

This speculative remark is related to (1) and somehow supported by our experiments, where we observed that simple models performed better than complex models, even though they probably did not represent the true conditional expectation.

Also, an accurate estimation of the MSE with respect to the regression model and appropriate preprocessing of the data, such as removing isolated points in low-density regions, might improve the performance. While Algorithm 1 only rejects a decision if the error is equal, one could think about utilizing the error difference as a rejection criteria of a decision. For instance, if the error difference is smaller than a certain threshold, the algorithm returns ’no decision’. This idea is further evaluated in ‘Error ratio as rejection criterion’.

EXPERIMENTS

In this section, we compare our algorithm with five different related methods for inferring the causal direction in various artificially generated and observed real-world data sets. In each evaluation, observations of two variables were given and the goal was to correctly identify cause and effect variable.

Causal inference methods for comparison

In the following, we briefly discuss and compare the causal inference methods which we used for the evaluations.

LiNGAM. The model assumptions of LiNGAM (*Shimizu et al., 2006*) are

$$E = \beta C + N,$$

where $\beta \in \mathbb{R}$, $C \perp\!\!\!\perp N$ and N is non-Gaussian. While LiNGAM is especially suitable for linear functional relationships with non-Gaussian noise, it performs poorly if these assumptions are violated. The computational cost is, however, relatively low.

For the experiments, we used a state-of-the-art implementation of LiNGAM that utilizes an entropy based method for calculating the likelihood ratios of the possible causal directions, instead of an independent component analysis based algorithm as in the original version (*Hyvärinen & Smith, 2013*). For this, Eq. (3) in *Hyvärinen & Smith (2013)* is used in order to estimate the likelihood ratio Eq. (2) in *Hyvärinen & Smith (2013)*.

ANM. The ANM (*Hoyer et al., 2009*) approach assumes that

$$E = f(C) + N,$$

where f is nonlinear and $C \perp\!\!\!\perp N$. An asymmetry between cause and effect is achieved by the assumption of an independence between cause and residual. Therefore, this method requires fitting a regression function and performing an additional evaluation of the relation between input and residual, which lead to a high computational cost. Note that the choice of the evaluation method is crucial for the performance.

We used an implementation provided by *Mooij et al. (2016)*, which uses a Gaussian process regression for the prediction and provides different methods for the evaluation of the causal direction. For the experiments, we chose different evaluation methods; *HSIC* for statistical independence tests, an entropy estimator for the estimation of the mutual information between input and residuals (denoted as *ENT*) and a Bayesian model comparison that assumes Gaussianity (denoted as *FN*). Implementation details and parameters can be found in Table 2 of *Mooij et al. (2016)*.

PNL. Post non-linear models (*Zhang & Hyvärinen, 2009*) are a generalization of ANMs. Here, it is assumed that

$$E = g(f(C) + N),$$

where g is nonlinear and $C \perp\!\!\!\perp N$. Due to the additional nonlinearity coming from g, this allows a non-additive influence of the noise as in contrast to an ANM. For inferring the causal direction, the idea remains roughly the same; fit a PNL in both possible directions and check for independence between input and disturbance. However, the disturbance here is different from the regression residual and fitting a PNL model is a significantly harder problem than fitting an ANM.

In the experiments, we used an implementation provided by the authors *Zhang & Hyvärinen (2009)*, where a constrained nonlinear independent component analysis is utilized for estimating the disturbances and HSIC for statistical independence tests.

IGCI. The IGCI (*Janzing et al., 2012*) approach is able to determine the causal relationship in a deterministic setting

$$E = f(C),$$

under the 'independence assumption' $\mathrm{Cov}[\log f', p_C] = 0$, i.e., the (logarithmic) slope of the function and the cause distribution are uncorrelated. The causal direction can then be inferred if the Kullback Leibler divergence between a reference measure and p_X is bigger or smaller than the Kullback Leibler divergence between the same reference measure and p_Y, respectively. The corresponding algorithm has been applied to noisy causal relations with partial success (and some heuristic justifications (*Janzing et al., 2012*)), but generalizations of IGCI for non-deterministic relations are actually not known and we consider Assumption 4 in 'Assumptions' as first step towards a possibly more general formulation. The computational cost depends on the utilized method for estimating the information criterion, but is generally low. Therefore, IGCI is the fastest of the methods.

For the experiments, we also used an implementation provided by *Mooij et al. (2016)*, where we always tested all possible combinations of reference measures and information estimators. These combinations are denoted as IGCI-ij, where i and j indicate:

- $i = \mathrm{U}$: Uniform reference measure (normalizing X and Y)
- $i = \mathrm{G}$: Gaussian reference measure (standardizing X and Y)
- $j = 1$: Entropy estimator using Eq. (12) in *Daniušis et al. (2010)*
- $j = 2$: Integral approximation of Eq. (13) in *Daniušis et al. (2010)*
- $j = 3$: Integral approximation of Eq. (22) in *Mooij et al. (2016)*

CURE. CURE (Sgouritsa et al., 2015) is based on the idea that an unsupervised regression of E on C by only using information from p_C performs worse than an unsupervised regression of C on E by only using information from p_E. CURE implements this idea in a Bayesian way via a modified Gaussian process regression. However, since CURE requires the generation of Markov-Chain-Monte-Carlo (MCMC) samples, the biggest drawback is a very high computational cost.

An implementation of CURE by the authors has been provided for our experiments. Here, we used similar settings as described in Section 6.2 of Sgouritsa et al. (2015), where 200 data samples were used and 10000 MCMC samples were generated. The number of internal repetitions depends on the experimental setting.

SLOPE. The SLOPE approach by Marx & Vreeken (2017) is essentially motivated by (1) and compares an estimation of $(K(p_X)+K(p_{Y|X}))/(K(p_X)+K(p_Y))$ with an estimation of $(K(p_Y)+K(p_{X|Y}))/(K(p_X)+K(p_Y))$ based on the minimum description length principle (Rissanen, 1978). This approach uses a global and multiple local regression models to fit the data, where the description length of the fitted regression models and the description length of the error with respect to the data can be used to approximate $K(p_{Y|X})$ and $K(p_{X|Y})$, respectively. Seeing that multiple regression models need to be fit depending on the structure of the data, the computational costs can vary between data sets.

For our experiments, we used the implementation provided by the authors with the same parameters as used in their experiments with real-world data.

RECI. Our approach addresses non-deterministic nonlinear relations and, in particular, allows a dependency between cause and noise. Since we only require the fitting of a least-squares solution in both possible causal directions, RECI can be easily implemented. It does not rely on any independence tests and has, depending on the regression model and implementation details, a low computational cost.

In the experiments, we have always used the same class of regression function for the causal and anticausal direction to compare the errors, but performed multiple experiments with different function classes. For each evaluation, we randomly split the data into training and test data, where we tried different ratios and selected the best performing model on the test data. The used percentage of training data were 70%, 50% or 30%, where the remaining data served as test data. In each run, we only randomly split the data once. The utilized regression models were:

- a logistic function (LOG) of the form $a+(b-a)/(1+\exp(c\cdot(d-x)))$
- shifted monomial functions (MON) of the form ax^n+b with $n\in[2,9]$
- polynomial functions (POLY) of the form $\sum_{i=0}^{k} a_i x^i$ with $k\in[1,9]$
- support vector regression (SVR) with a linear kernel
- neural networks (NN) with different numbers of hidden neurons: 2, 5, 10, 20, 2-4, 4-8, where '-' indicates two hidden layers

The logistic and monomial functions cover rather simple regression models, which are probably not able to capture the true function ϕ in most cases. On the other hand, support vector regression and neural networks should be complex enough to capture ϕ.

The polynomial functions are rather simple too, but more flexible than the logistic and monomial functions.

We used the standard Matlab implementation of these methods and have always chosen the default parameters, where the parameters of LOG, MON and POLY were fitted by minimizing the least-squares error.

During the experiments, we observed that the MSE varied a lot in many data sets due to relatively small sample sizes and the random selection of training and test data. Therefore, we averaged the MSE over all performed runs within the same data set first before comparing them, seeing that this should give more accurate estimations of $\mathbb{E}[\mathrm{Var}[Y|X]]$ and $\mathbb{E}[\mathrm{Var}[X|Y]]$ with respect to the class of the regression function. Although the choice of the function class for each data set is presumably a typical model selection problem, we did not optimize the choice for each data set individually. Therefore, we only summarize the results of the best performing classes with respect to the experimental setup in the following. The estimated MSE in each data set were averaged over all performed runs. A detailed overview of all results, including the performances and standard deviations of all function classes when estimating the MSE in single and multiple runs, can be found in the supplements. For the normalization of the data we used

$$\hat{C} := \frac{C - \min(C)}{\max(C) - \min(C)}$$
$$\hat{E} := \frac{E - \min(E)}{\max(E) - \min(E)}$$

and for the standardization we used

$$\hat{C} := \frac{C - \mathbb{E}[C]}{\sqrt{\mathrm{Var}[C]}}$$
$$\hat{E} := \frac{E - \mathbb{E}[E]}{\sqrt{\mathrm{Var}[E]}}.$$

General Remark. Each evaluation was performed in the original data sets and in preprocessed versions where isolated points (low-density points) were removed. For the latter, we used the implementation and parameters from *Sgouritsa et al. (2015)*, where a kernel density estimator with a Gaussian kernel is utilized to sort the data points according to their estimated densities. Then, data points with a density below a certain threshold (0.1 in our experiments) are removed from the data set. In this way, outliers should have a smaller impact on the performance. It also shows how sensitive each approach is to outliers. However, removing outliers might lead to an underestimation of the noise in a heavy tail noise distribution and is, therefore, not always the best choice as a preprocessing step. Note that CURE per default uses this preprocessing step, also in the original data. In all evaluations, we forced a decision by the algorithms, where in case of ANM the direction with the highest score of the independence test was taken.

Except for CURE, we averaged the performances of each method over 100 runs, where we uniformly sampled 500 data points for ANM and SVR if the data set contains more than 500 data points. For CURE, we only performed four internal repetitions in the artificial and eight internal repetitions in the real-world data sets due to the high computational cost. As performance measure, we consider the accuracy of correctly identifying the true causal direction. Therefore, the accuracy is calculated according to

$$\text{accuracy} = \frac{\sum_{m=1}^{M} w_m \delta_{\hat{d}_m, d_m}}{\sum_{m=1}^{M} w_m}, \tag{22}$$

where M is the number of data sets, w_m the weight of data set m, d_m the correct causal direction and $\hat{d}_m$ the inferred causal direction of the corresponding method. Note that we consider $w_m = 1$ for all artificial data sets, while we use different weights for the real-world data sets. Since we use all data points for SLOPE, IGCI and LiNGAM, these methods have a consistent performance over all runs. An overview of all utilized data sets with their corresponding number of cause–effect pairs and data samples can be found in Table 7 in the supplements.

Artificial data

For experiments with artificial data, we performed evaluations with simulated cause–effect pairs generated for a benchmark comparison in *Mooij et al. (2016)*. Further, we generated additional pairs with linear, nonlinear invertible and nonlinear non-invertible functions where input and noise are strongly dependent.

Simulated benchmark cause–effect pairs

The work of *Mooij et al. (2016)* provides simulated cause–effect pairs with randomly generated distributions and functional relationships under different conditions. As pointed out by *Mooij et al. (2016)*, the scatter plots of these simulated data look similar to those of real-world data. We took the same data sets as used in *Mooij et al. (2016)* and extend the reported results with an evaluation with SLOPE, CURE, LiNGAM, RECI and further provide results in the preprocessed data.

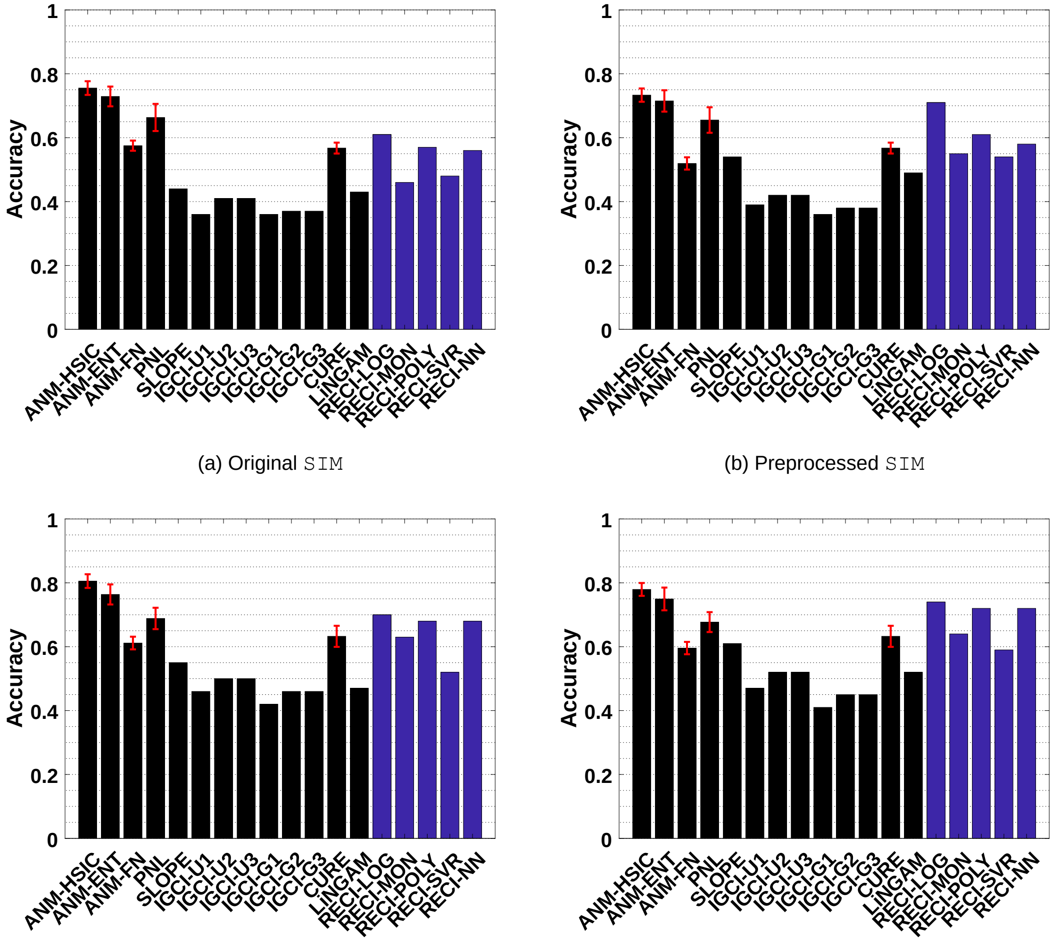
The data sets are categorized into four different categories:

- `SIM`: Pairs without confounders. The results are shown in Figs. 3A–3B
- `SIM-c`: A similar scenario as `SIM`, but with one additional confounder. The results are shown in Figs. 3C–3D
- `SIM-ln`: Pairs with low noise level without confounder. The results are shown in Figs. 4A-4B
- `SIM-G`: Pairs where the distributions of C and N are almost Gaussian without confounder. The results are shown in Figs. 4C–4D.

The general form of the data generation process without confounder but with measurement noise[5] is

$$\begin{aligned} C' &\sim p_C, N \sim p_N \\ N_C &\sim \mathcal{N}(0, \sigma_C), N_E \sim \mathcal{N}(0, \sigma_E) \\ C &= C' + N_C \\ E &= f_E(C', N) + N_E \end{aligned}$$

[5] Note, however, that adding noise to the cause (as it is done here) can also be considered as a kind of confounding. Actually, C' is the cause of E in the below generating model, while the noisy version C is not the cause of E. Accordingly, C' is the hidden common cause of C and E. Here we refer to the scenario as an unconfounded case with measurement noise as in *Mooij et al. (2016)*.



(a) Original `SIM`

(b) Preprocessed `SIM`

(c) Original `SIM-c`

(d) Preprocessed `SIM-c`

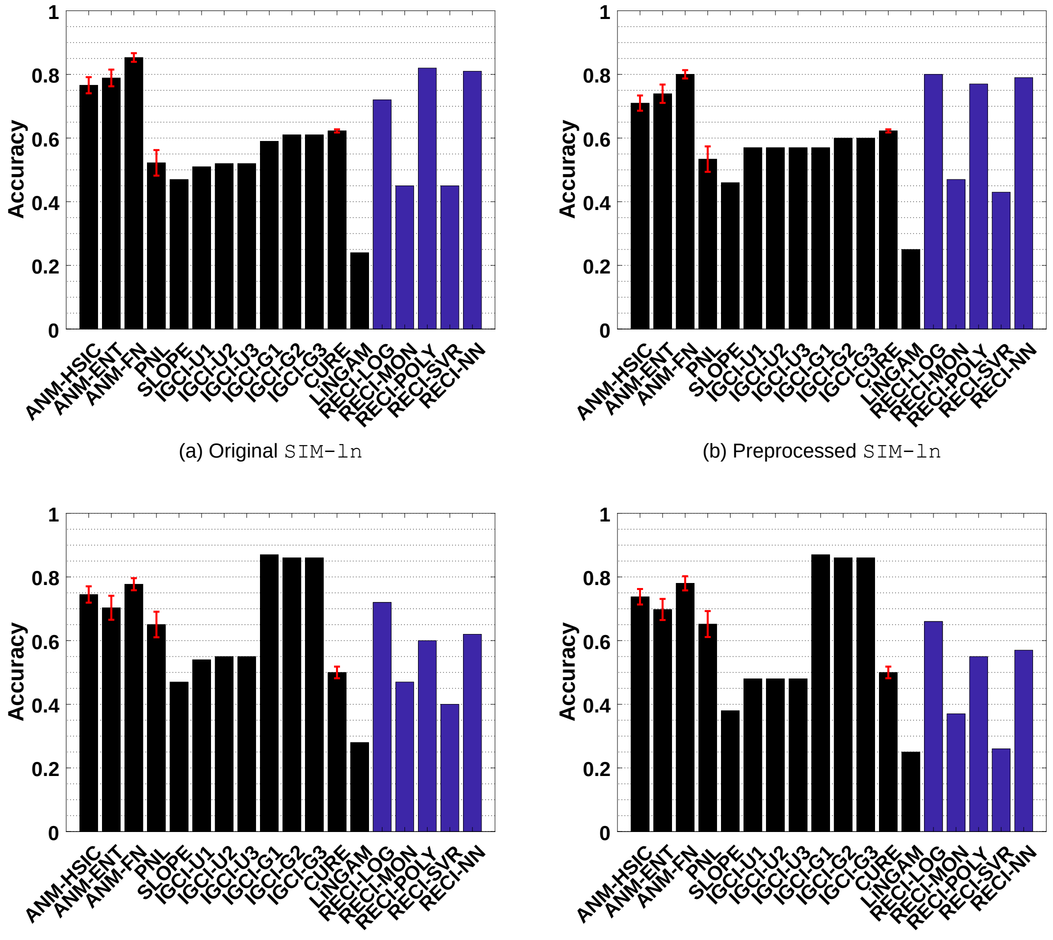
Figure 3 Evaluation results of all methods in the `SIM` and `SIM-c` data sets. (A) and (C) show the results of the evaluations in the original data and (B) and (D) the results in the preprocessed versions where low-density points were removed.

Full-size DOI: 10.7717/peerjcs.169/fig-3

and with confounder

$$\begin{aligned} C' &\sim p_C, N \sim p_N, Z \sim p_Z \\ C'' &= f_C(C', Z) \\ N_C &\sim \mathcal{N}(0,\sigma_C), N_E \sim \mathcal{N}(0,\sigma_E) \\ C &= C'' + N_C \\ E &= f_E(C'', Z, N) + N_E, \end{aligned}$$

where N_C, N_E represent independent observational Gaussian noise and the variances σ_C and σ_E are chosen randomly with respect to the setting. Note that only N_E is Gaussian, while the regression residual is non-Gaussian due to the nonlinearity of f_E and non-Gaussianity of N, Z. Thus, the noise in `SIM`, `SIM-c` and `SIM-G` is non-Gaussian. More details can be found in Appendix C of *Mooij et al. (2016)*.



(a) Original `SIM-ln`

(b) Preprocessed `SIM-ln`

(c) Original `SIM-G`

(d) Preprocessed `SIM-G`

Figure 4 Evaluation results of all methods in the `SIM-ln` and `SIM-G` data sets. (A) and (C) show the results of the evaluations in the original data and (B) and (D) the results in the preprocessed versions where low-density points were removed.

Full-size DOI: 10.7717/peerjcs.169/fig-4

Generally, ANM performs the best in all data sets. However, the difference between ANM and RECI, depending on the regression model, becomes smaller in the preprocessed data where isolated points were removed. According to the observed performances, removing these points seems to often improve the accuracy of RECI, but decrease the accuracy of ANM. In case of the preprocessed `SIM-G` data sets, the accuracy of RECI is decreased seeing that in these nearly Gaussian data the removal of low-density points leads to an underestimation of the noise distribution.

In all data sets, except for `SIM-G`, RECI always outperforms SLOPE, IGCI, CURE and LiNGAM if a simple logistic or polynomial function is utilized for the regression. However, in the `SIM-G` data set, our approach performs comparably poor, which could be explained by the violation of the assumption of a compact support. In this nearly Gaussian setting, IGCI performs the best with a Gaussian reference measure. However, we also evaluated RECI with standardized data in the `SIM-G` data sets, which is equivalent to a Gaussian reference measure for IGCI. A summary of the results can be found in Figures 1(c)–1(d) in the supplements and more detailed results in Table 4 and Table 5 in the supplements. These results are significantly better than normalizing the data in this case. However, although our theorem only justifies a normalization, a different scaling, such as standardization, might be a reasonable alternative.

Even though Theorem 1 does not exclude cases of a high noise level, it makes a clear statement about low noise level. Therefore, as expected, RECI performs the best in `SIM-ln`, where the noise level is low. In all cases, LiNGAM performs very poorly due to the violations of its core assumptions. Surprisingly, although PNL is a generalization of ANM, we found that PNL performs generally worse than ANM, but better than SLOPE, CURE and LiNGAM.

ANM and RECI require a least-squares regression, but ANM additionally depends on an independence test, which can have a high computational cost and a big influence on the performance. Therefore, even though RECI does not outperform ANM, it represents a competitive alternative with a lower computational cost, depending on the regression model and MSE estimation. Also, seeing that RECI explicitly allows both cases, a dependency and an independency between C and N and ANM only the latter, it can be expected that RECI performs significantly better than ANM in cases where the dependency between C and N is strong. This is evaluated in 'Simulated cause–effect pairs with strong dependent noise'. In comparison with PNL, SLOPE, IGCI, LiNGAM and CURE, RECI outperforms in almost all data sets. Note that *Mooij et al. (2016)* performed more extensive experiments and showed more comparisons with ANM and IGCI in these data sets, where additional parameter configurations were tested. However, they reported no results for the preprocessed data.

Simulated cause–effect pairs with strong dependent noise

Since the data sets of the evaluations in 'Simulated benchmark cause–effect pairs' are generated by structural equations with independent noise variables, we additionally performed evaluations with artificial data sets where the input distribution and the noise distribution are strongly dependent. For this, we considered a similar data generation process as described in the work by *Daniušis et al. (2010)*. We generated data with various cause and noise distributions, different functions and varying values for $\alpha \in [0,1]$. In order to ensure a dependency between C and N, we additionally introduced two unobserved source variables S_1 and S_2 that are randomly sampled from different distributions. Variables C and N then consist of a randomly weighted linear combination of S_1 and S_2. The general causal structure of these data sets is illustrated in Fig. 5. Note that S_1 and S_2 can be seen as hidden confounders affecting both C and E.

Apart from rather simple functions for ϕ, *Daniušis et al. (2010)* proposed to generate more general functions in the form of convex combinations of mixtures of cumulative Gaussian distribution functions $\psi(C|\mu_i,\sigma_i)$:

$$s_n(C) = \sum_{i=1}^{n} \beta_i \psi(C|\mu_i,\sigma_i),$$

where $\beta_i, \mu_i \in [0,1]$ and $\sigma_i \in [0, 0.1]$. For the experiments, we set $n = 5$ and chose the parameters of $s_5(C)$ randomly according to the uniform distribution. Note that $\psi(C|\mu_i,\sigma_i)$

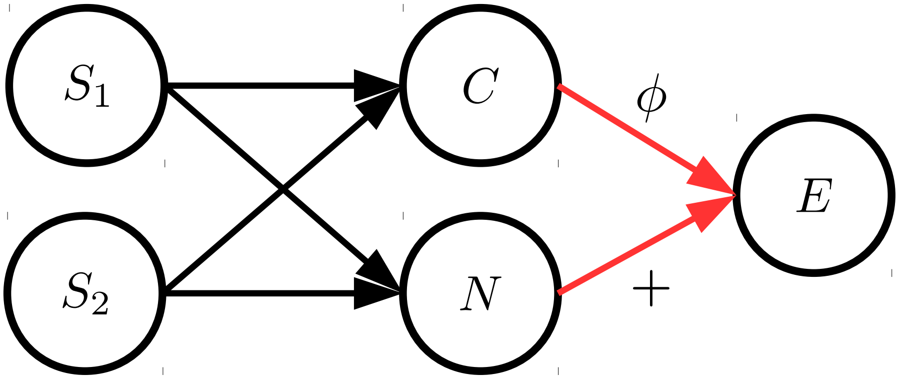


Figure 5 The general structure of the data generation process where C and N are dependent. In order to achieve this, cause and noise consist of a mixture of two sources S_1 and S_2.

Full-size DOI: 10.7717/peerjcs.169/fig-5

is always monotonically increasing and thus $s_5(C)$ can have an arbitrary random shape while being monotonically increasing.

Cause-effect pairs were then generated in the following way

$$\begin{aligned}
w_1, w_2 &\sim U(0,1)\\
S_1, S_2 &\sim p_S\\
S_1 &= S_1 - \mathbb{E}[S_1]\\
S_2 &= S_2 - \mathbb{E}[S_2]\\
C' &= w_1 \cdot f_1(S_1) + (1 - w_1) \cdot f_2(S_2)\\
N' &= w_2 \cdot f_3(S_1) + (1 - w_2) \cdot f_4(S_2)\\
C &= \text{normalize}(C')\\
N &= \alpha \cdot \text{standardize}(N')\\
E &= \phi(C) + N,
\end{aligned}$$

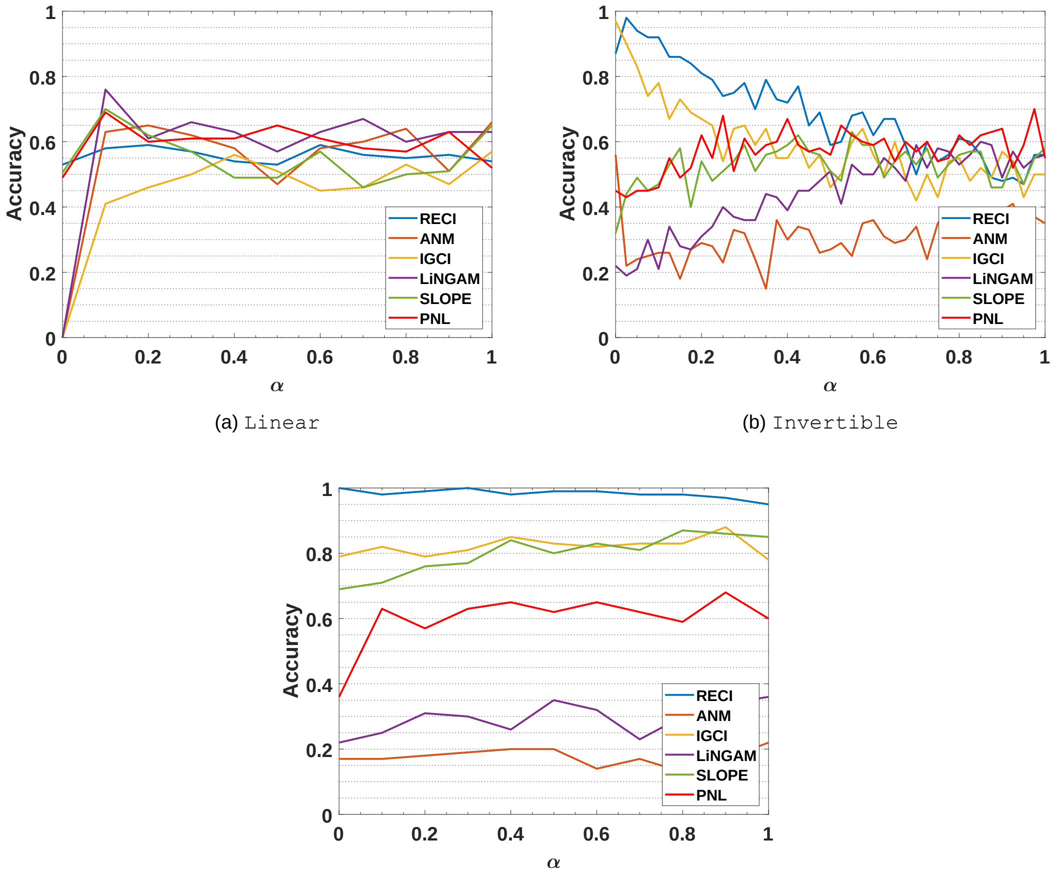
where the distributions of S_1 and S_2 and the functions f_1, f_2, f_3, f_4 were chosen randomly from p_S and f in Table 1, respectively. Note that S_1 and S_2 can follow different distributions. The choice of ϕ depends on the data set, where we differentiated between three data sets:

- `Linear`: Only the identity function $\phi(C) = C$
- `Invertible`: Arbitrary invertible functions $\phi(C) = s_5(C)$
- `Non-invertible`: Functions that are not invertible on the respective domain

In total, we generated 100 data sets for each value of parameter α, which controls the amount of noise in the data. In each generated data set, we randomly chose different distributions and functions. For `Linear` and `Non-invertible` the step size of α is 0.1 and for `Invertible` 0.025. Here, we only performed one repetition on each data set for all algorithms. Figs. 6A–6C summarize all results and Table 6 in the supplements shows the best performing functions and parameters of the different causal inference methods. Note that we omitted experiments with CURE in these data sets due to the high computational cost.

Table 1 All distributions p_S, functions ϕ and functions f that were used for the generation of the `Linear`, `Non-invertible` and `Invertible` data sets. In case of the functions for `Non-invertible`, rescale $(X,-n,n)$ denotes a rescaling of the input data X on $[-n,n]$. $GM_{\mu,\sigma}$ denotes a Gaussian mixture distribution with density $p_{GM_{\mu,\sigma}}(c)=\frac{1}{2}(\varphi(c|\mu_1,\sigma_1)+\varphi(c|\mu_2,\sigma_2))$ and Gaussian pdf $\varphi(c|\mu,\sigma)$.

Data set	$\phi(C)$		
`Linear`	C	p_S	$f(X)$
`Invertible`	$s_5(C)$	$U(0,1)$	X
`Non-invertible`	$\mathrm{rescale}(C,-2,2)^2$	$\mathcal{N}(0,\sigma^2)$	$\exp(X)$
	$\mathrm{rescale}(C,-2,2)^4$	$\mathcal{N}(0.5,\sigma^2)$	$s_5(X)$
	$\sin(\mathrm{rescale}(C,-2\cdot\pi,2\cdot\pi))$	$\mathcal{N}(1,\sigma^2)$	
		$GM_{[0.3,0.7]^{\mathrm{T}},[0.1,0.1]^{\mathrm{T}}}$	



(a) `Linear`

(b) `Invertible`

(c) `Non-invertible`

Figure 6 Evaluation results of all methods in the (A) `Linear`, (B) `Invertible` and (C) `Non-Invertible` data sets. The parameter α controls the amount of noise in the data.

Full-size DOI: 10.7717/peerjcs.169/fig-6

`Linear`: As expected, ANM, PNL, SLOPE, IGCI and RECI perform very poorly, since they require nonlinear data. In case of RECI, Theorem 1 states an equality of the MSE if the functional relation is linear and, thus, the causal direction can not be inferred. While LiNGAM performs well for $\alpha=0.1$, and probably for smaller values of α too, the performance drops if α increases. The poor performances of LiNGAM and ANM can also be explained by the violation of its core assumption of an independence between cause and input.

`Invertible`: In this data set, IGCI performs quite well for small α, since all assumptions approximately hold, but the performance decreases when the noise becomes stronger and violates the assumption of a deterministic relation. In case of RECI, we made a similar observation, but it performs much better than IGCI if $\alpha < 0.5$. Aside from the assumption of linear data for LiNGAM, the expected poor performance of LiNGAM and ANM can be explained by the violation of the independence assumption between C and N. In contrast to the previous results, PNL performs significantly better in this setting than ANM, although, likewise LiNGAM and ANM, the independence assumption is violated.

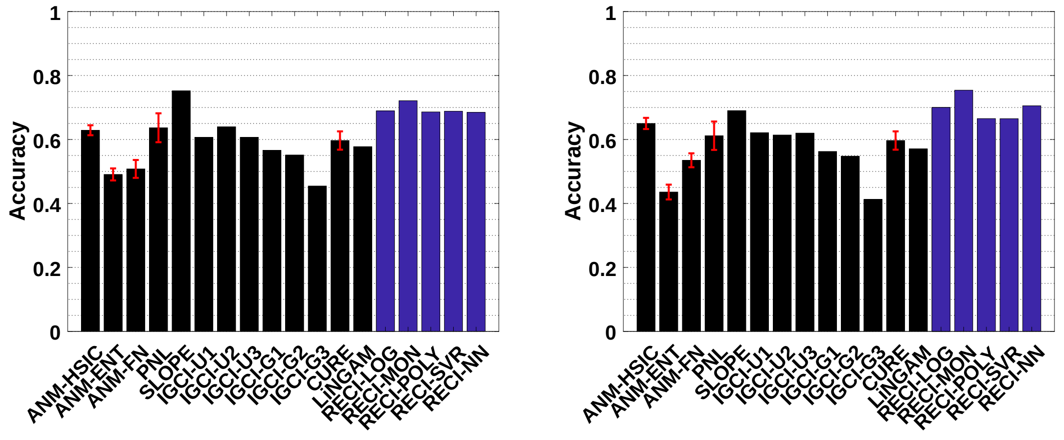
`Non-invertible`: These results seem very interesting, since it supports the argument that the error asymmetry becomes even clearer if the function is not invertible due to an information loss of regressing in anticausal direction. Here, IGCI and SLOPE perform reasonably well, while ANM and LiNGAM perform even worse than a baseline of just guessing. Comparing ANM and PNL, PNL has a clear advantage, although the overall performance is only slightly around 60% in average. The constant results of each method can be explained by the rather simple and similar choice of data generating functions.

While the cause and noise also have a dependency in the `SIM-c` data sets, the performance gap between ANM and RECI is vastly greater in `Invertible` and `Non-invertible` than in `SIM-c` due to a strong violation of the independent noise assumption. Therefore, RECI might perform better than ANM in cases with a strong dependency between cause and noise.

Real-world data

In real-world data, the true causal relationship generally requires expert knowledge and can still remain unclear in cases where randomized controlled experiments are not possible. For our evaluations, we considered the commonly used cause–effect pairs (`CEP`) benchmark data sets for inferring the causal direction in a bivariate setting. These benchmark data sets provided, at the time of these evaluations, 106 data sets with given cause and effect variables and can be found on https://webdav.tuebingen.mpg.de/cause-effect/. However, since we only consider a two variable problem, we omit six multivariate data sets, which leaves 100 data sets for the evaluations. These data sets consist of a wide range of different scenarios, such as data from time dependent and independent physical processes, sociological and demographic studies or biological and medical observations. An extensive analysis and discussion about the causal relationship of the first 100 data sets can be found in the work by *Mooij et al. (2016)*. Each data set comes with a corresponding weight determined by expert knowledge. This is because several data sets are too similar to consider them as independent examples, hence they get lower weights. Therefore, the weight w_m in Eq. (22) depends on the corresponding data set. The evaluation setup is the same as for the artificial data sets, but we doubled the number of internal repetition of CURE to eight times in order to provide the same conditions as in *Sgouritsa et al. (2015)*.

Figures 7A and 7B shows the results of the evaluations in the original and preprocessed data, respectively. In all cases, SLOPE and RECI perform significantly better than ANM,



(a) Original `CEP`

(b) Preprocessed `CEP`

Figure 7 **Evaluation results of all methods in the real-world CEP data sets.** (A) shows the result of the evaluations in the original data and (B) the results in the preprocessed versions where low-density points were removed.

Full-size DOI: 10.7717/peerjcs.169/fig-7

PNL, IGCI, CURE and LiNGAM. While SLOPE performs slightly better in the original data sets than RECI, RECI performs better overall in the preprocessed data[6]. Further, the performance gap even increases in the preprocessed data with removed low-density points. Surprisingly, as Table 1 in the supplements indicates, the simple shifted monomial function ax^2+c performs the best, even though it is very unlikely that this function is able to capture the true function ϕ. We obtained similar observations in the artificial data sets, where the simple logistic function oftentimes performs the best.

In order to show that RECI still performs reasonably well under a different scaling, we also evaluated RECI in the real-world data set with standardized data. These results can be found summarized in Figures 1(a)–1(b) in the supplements and more detailed in Table 3 and Table 4 in the supplements. While standardizing the data improves the performance in the SIM-G data, it slightly decreases the performance in the real-world data as compared to a normalization of the data, but still performs reasonably well. This shows some robustness with respect to a different scaling.

Error ratio as rejection criterion

It is not clear how a confidence measure for the decision of RECI can be defined. However, since Theorem 1 states that the correct causal direction has a smaller error, we evaluated the idea of utilizing the error ratio for a confidence measure in terms of:

$$\text{confidence} = 1 - \frac{\min(\mathbb{E}[\mathrm{Var}[X|Y]], \mathbb{E}[\mathrm{Var}[Y|X]])}{\max(\mathbb{E}[\mathrm{Var}[X|Y]], \mathbb{E}[\mathrm{Var}[Y|X]])}, \tag{23}$$

The idea is that, the smaller the error ratio, the higher the confidence of the decision, due to the large error difference. Note that we formulated the ratio inverse to Theorem 1 in order to get a value on $[0,1]$. Algorithm 1 can be modified in a straight forward manner to utilize this confidence measure. The modification is summarized in Algorithm 2[7].

[6] The work of *Mooij et al. (2016)* provides further evaluations of ANM and IGCI in the original CEP data set with parameter configurations that reached slightly higher accuracies than the presented results in this work. Regarding CURE, we had to use a simplified implementation due to the high computational cost, which did not perform as well as the results reported in *Sgouritsa et al. (2015)*.

[7] Algorithm 1 and Algorithm 2 are equivalent if $t=0$.

Algorithm 2 Causal inference algorithm that uses Eq. (23) as rejection criterion.

```
function RECI($X, Y, t$) ▷ $X$ and $Y$ are the observed data and $t \in [0,1]$ is the confidence threshold for rejecting a decision.
    $(X,Y) \leftarrow$ RescaleData$(X,Y)$
    $f \leftarrow$ FitModel$(X,Y)$                ▷ Fit regression model $f: X \rightarrow Y$
    $g \leftarrow$ FitModel$(Y,X)$                ▷ Fit regression model $g: Y \rightarrow X$
    $\text{MSE}_{Y|X} \leftarrow$ MeanSquaredError$(f,X,Y)$
    $\text{MSE}_{X|Y} \leftarrow$ MeanSquaredError$(g,Y,X)$
    $\xi \leftarrow 1 - \frac{\min(\text{MSE}_{X|Y},\text{MSE}_{Y|X})}{\max(\text{MSE}_{X|Y},\text{MSE}_{Y|X})}$
    if $\xi \geq t$ then
        if $\text{MSE}_{Y|X} < \text{MSE}_{X|Y}$ then
            return $X$ causes $Y$
        else
            return $Y$ causes $X$
        end if
    else
        return No decision
    end if
end function
```

We re-evaluated the obtained results by considering only data sets where Algorithm 2 returns a decision with respect to a certain confidence threshold. Figs. 8A–8D show some examples of the performance of RECI if we use Eq. (23) to rank the confidence of the decisions. A decision rate of 20%, for instance, indicates the performance when we only force a decision on 20% of the data sets with the highest confidence. In this sense, we can get an idea of how useful the error ratio is as rejection criterion. While Figs. 8A, 8C and 8D support the intuition that the smaller the error ratio, the higher the chance of a correct decision, Fig. 8B has a contradictive behavior. In Fig. 8B, it seems that a small error ratio (big error difference) is rather an indicator for an uncertain decision (probably caused by over- or underfitting issues). Therefore, we can not generally conclude that Eq. (23) is a reliable confidence measure in all cases, but it seems to be a good heuristic approach in the majority of the cases. More plots can be found in the supplements, where Fig. 2 shows the plots in the original data, Fig. 3 in the preprocessed data and Fig. 4 in the standardized data. Note that a deeper analysis of how to define an appropriate confidence measure for all settings is beyond the scope of this paper and we rather aim to provide some insights of utilizing the error ratio for this purpose.

Run-time comparison

In order to have a brief overview and comparison of the run-times, we measured the execution durations of each method in the evaluation of the original CEP data sets. All measures were performed with a Intel Xeon E5-2695 v2 processor. Table 2 summarizes the measured run-times, where we stopped the time measurement of CURE after 9999 s. As the table indicates, IGCI is the fastest method, followed by LiNGAM and RECI. The

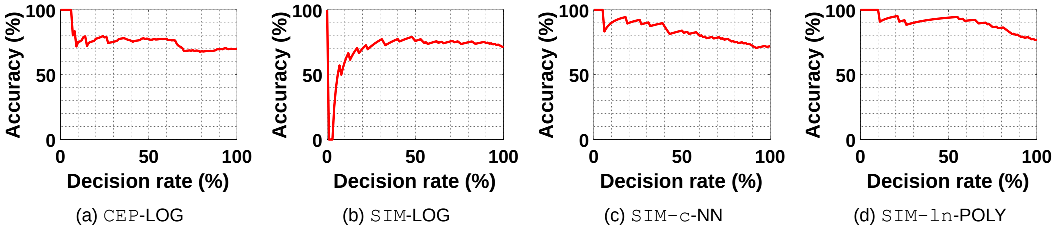


Figure 8 The exemplary performance of RECI if a certain decisions rate is forced. (A) CEP-LOG, (B) SIM-LOG, (C) SIM-c-NN, (D) SIM-ln-POLY. Here, the decisions are ranked according to the confidence measure defined in Eq. (23).

Full-size DOI: 10.7717/peerjcs.169/fig-8

Table 2 The average run-time in seconds of each method using all CEP data sets.

Method	Time (s)
ANM-HSIC	2551.8 ± 161.5
ANM-ENT	2463.1 ± 39
ANM-FN	2427.6 ± 40.2
PNL	6019.7 ± 118.49
SLOPE	1654.9 ± 10.01
IGCI-U1	0.0385 ± 0.0028
IGCI-U2	0.0384 ± 0.0024
IGCI-U3	0.5843 ± 0.0327
IGCI-G1	0.0414 ± 0.0025
IGCI-G2	0.0429 ± 0.0028
IGCI-G3	0.5866 ± 0.0329
CURE	> 9999
LiNGAM	0.1459 ± 0.0053
RECI-LOG	63.16 ± 3.35
RECI-MON	4.65 ± 0.28
RECI-POLY	2.78 ± 0.1
RECI-SVR	87.33 ± 33.43
RECI-NN	46.62 ± 0.28

ranking of ANM, PNL, SLOPE and RECI is not surprising; ANM and PNL need to evaluate the independence between input and residual on top of fitting a model. In case of SLOPE, multiple regression models need to be fitted depending on a certain criterion that requires to be evaluated. Therefore, by construction, RECI can be expected to be faster than ANM, PNL and SLOPE.

Discussion

Due to the greatly varying behavior and the choice of various optimization parameters, a clear rule of which regression function is the best choice for RECI remains an unclear and difficult problem. Overall, it seems that simple functions are better in capturing the error asymmetries than complex models. However, a clear explanation for this is still lacking. A possible reason for this might be that simple functions in causal direction already achieve a small error, while in anticausal direction, more complex models are required to achieve a small error. To justify speculative remarks of this kind raises deep questions about the foundations of causal inference. According to Eq. (1), it is possible to conclude that the joint distribution has a simpler description in causal direction than in anticausal direction. Seeing this, a model selection based on the regression performance and model complexity considered in a dependent manner might further improve RECI's practical applicability. Regarding the removal of low-density points, the performance of methods that are based on the Gaussiantiy assumption, such as FN and IGCI with Gaussian reference measure, seems not to be influenced by the removal. On the other hand, the performance of HSIC, ENT, and IGCI with uniform measure is negatively affected, while the performance of LiNGAM and RECI increases. In case of RECI, this can be explained by a better estimation of the true MSE with respect to the regression function class.

Regarding the computational cost, we want to emphasize again that RECI, depending on the implementation details, can have a significantly lower computational cost than ANM, SLOPE and CURE, while providing comparable or even better results. Further, it can be easily implemented and applied.

CONCLUSION

We presented an approach for causal inference based on an asymmetry in the prediction error. Under the assumption of an independence among the data generating function, the noise, and the distribution of the cause, we proved (in the limit of small noise) that the conditional variance of predicting the cause by the effect is greater than the conditional variance of predicting the effect by the cause. For instance, in the example shown in Fig. 1, the regression error in the true causal direction is smaller than the error in the anticausal direction. In our work, the additive noise is not assumed to be independent of the cause (in contrast to so-called additive noise models). The stated theorem might also be interesting for other statistical applications.

We proposed an easily implementable and applicable algorithm, which we call RECI, that exploits this asymmetry for causal inference. The evaluations show supporting results and leave room for further improvements. By construction, the performance of RECI depends on the regression method. According to our limited experience so far, regression with simple model classes (that tend to underfit the data) performs reasonably well. To clarify whether this happens because the conditional distributions tend to be simpler—in a certain sense—in causal direction than in anticausal direction has to be left for the future.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by JST CREST Grant Number JPMJCR1666 and JSPS KAKENHI Grant Number JP17K00305, Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
ST CREST: JPMJCR1666.
JSPS KAKENHI: JP17K00305.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Patrick Blöbaum conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, performed the computation work, authored or reviewed drafts of the paper, approved the final draft.
- Dominik Janzing and Takashi Washio conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.
- Shohei Shimizu and Bernhard Schölkopf conceived and designed the experiments, contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Real-world data can be found at: https://webdav.tuebingen.mpg.de/cause-effect/
Artificial data can be found at: http://jmlr.org/papers/v17/14-518.html.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.169#supplemental-information.

REFERENCES

Blöbaum P, Janzing D, Washio T, Shimizu S, Schölkopf B. 2018. Cause-effect inference by comparing regression errors. In: *Proceedings of the 21st international conference on artificial intelligence and statistics (AISTATS 2018)*.

Blöbaum P, Shimizu S, Washio T. 2017. A novel principle for causal inference in data with small error variance. In: *European symposium on artificial neural networks*. Louvain-la-Neuve, Belgium: i6doc, 347–352.

Blöbaum P, Washio T, Shimizu S. 2017. Error asymmetry in causal and anticausal regression. *Behaviormetrika* **44(2)**:491–512.

Comley JW, Dowe DL. 2003. General Bayesian networks and asymmetric languages. In: *Proceedings of the second Hawaii international conference on statistics and related fields*.

Daniušis P, Janzing D, Mooij J, Zscheischler J, Steudel B, Zhang K, Schölkopf B. 2010. Inferring deterministic causal relations. In: *Proceedings of the 26th conference on uncertainty in artificial intelligence*. Corvallis: AUAI Press, 143–150.

Hoyer P, Janzing D, Mooij J, Peters J, Schölkopf B. 2009. Nonlinear causal discovery with additive noise models. In: *Advances in neural information processing systems 21*. Red Hook: Curran Associates, Inc., 689–696.

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

Janzing D, Mooij J, Zhang K, Lemeire J, Zscheischler J, Daniušis P, Steudel B, Schölkopf B. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* **182**:1–31 DOI 10.1016/j.artint.2012.01.002.

Janzing D, Schölkopf B. 2010. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory* **56(10)**:5168–5194 DOI 10.1109/TIT.2010.2060095.

Janzing D, Sun X, Schölkopf B. 2009. Distinguishing cause and effect via second order exponential models. eprint http://arxiv.org/abs/0910.5561.

Kano Y, Shimizu S. 2003. Causal inference using nonnormality. In: *Proceedings of the international symposium on science of modeling, the 30th anniversary of the information criterion*. Tokyo, 261–270.

Lemeire J, Janzing D. 2012. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines* **23(2)**:227–249 DOI 10.1007/s11023-012-9283-1.

Ma S, Statnikov A. 2017. Methods for computational causal discovery in biomedicine. *Behaviormetrika* **44(1)**:165–191 DOI 10.1007/s41237-016-0013-5.

Marx A, Vreeken J. 2017. Telling cause from effect using MDL-based local and global regression. In: *2017 IEEE international conference on data mining (ICDM)*. Piscataway: IEEE, 307–316 DOI 10.1109/ICDM.2017.40.

Mooij J, Peters J, Janzing D, Zscheischler J, Schölkopf B. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research* **17(32)**:1–102.

Pearl J. 2009. *Causality: models, reasoning and inference*. 2nd edition. New York: Cambridge University Press.

Peters J, Janzing D, Schölkopf B. 2011. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33(12)**:2436–2450 DOI 10.1109/TPAMI.2011.71.

Peters J, Janzing D, Schölkopf B. 2017. Elements of causal inference—foundations and learning algorithms. Cambridge: MIT Press.

Rissanen J. 1978. Modeling by shortest data description. *Automatica* **14(5)**:465 – 471 DOI 10.1016/0005-1098(78)90005-5.

Rosner F. 1987. The ethics of randomized clinical trials. *The American Journal of Medicine* **82(2)**:283–290 DOI 10.1016/0002-9343(87)90069-6.

Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. 2013. Semi-supervised learning in causal and anticausal settings. In: Schölkopf B, Luo Z, Vovk V, eds. *Empirical inference. Festschrift in Honor of Vladimir Vapnik*, Berlin, Heidelberg: Springer-Verlag, 129–141 DOI 10.1007/978-3-642-41136-6_13.

Sgouritsa E, Janzing D, Hennig P, Schölkopf B. 2015. Inference of cause and effect with unsupervised inverse regression. In: *Artificial intelligence and statistics*. San Diego: PMLR, 847–855.

Shimizu S, Hoyer P, Hyvärinen A, Kerminen A. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7**:2003–2030.

Spirtes P, Glymour C, Scheines R. 2000. *Causation, prediction, and search*. Cambridge: MIT press.

Statnikov A, Henaff M, Lytkin NI, Aliferis CF. 2012. New methods for separating causes from effects in genomics data. *BMC Genomics* **13**(**8**):S22 DOI 10.1186/1471-2164-13-S8-S22.

Sun X, Janzing D, Schölkopf B. 2006. Causal inference by choosing graphs with most plausible Markov kernels. In: *Proceedings of the 9th international symposium on artificial intelligence and mathematics*. Fort Lauderdale, FL, 1–11.

Zhang K, Hyvärinen A. 2009. On the identifiability of the post-nonlinear causal model. In: *Proceedings of the 25th conference on uncertainty in artificial intelligence*. Arlington: AUAI Press, 647–655.