Cause-Effect Inference by Comparing Regression Errors

Patrick Blöbaum
Osaka University
Japan

Dominik Janzing
MPI for Intelligent Systems
Tübingen, Germany

Takashi Washio
Osaka University
Japan

Shohei Shimizu
Osaka University, Shiga University
Japan

Bernhard Schölkopf
MPI for Intelligent Systems
Tübingen, Germany

Abstract

We address the problem of inferring the causal relation between two variables by comparing the least-squares errors of the predictions in both possible causal 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 method that only requires a regression in both possible causal directions. The performance of this method is compared with different related causal inference methods in various artificial and real-world data sets.

1 Introduction

Causal inference [1, 2] is becoming an increasingly popular topic in machine learning, since the results are often not only of interest to predicting the result of potential interventions, but also to general statistical and machine learning applications [3]. In particular, the identification of the causal relation between only two observed variables is a challenging task [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. As regards the present work, we address this bivariate setting, 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. For instance, is a certain change of the phyiscal state of a patient a symptom or a cause of a certain disease? If we are not able to observe the effect of an intervention on one of the variables, the identification of the correct causal relation generally relies on the exploitation of statistical asymmetries between cause and effect [6, 9, 13, 1, 2]. 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 are compatible with the respective pattern (subject to Markov condition and faithfulness assumption [1, 2]). Whenever there are causal arrows that are common to all DAGs in the class, conditional (in)dependences yield definite statements about causal directions.

In case of bivariate data, however, we rely on those types of 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) [6] guarantees to identify the correct causal direction. For nonlinear relations, the additive noise model (ANM) [9] and its generalization to post-nonlinear models (PNL) [16] identify the causal direction by assuming an independence between cause and noise, where, apart from some exceptions, 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 scenario are based on an *informal* independence assumption stating that the distribution of the cause, denoted by

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

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), [17, 18] 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. The information-geometric approach for causal inference (IGCI) [13] is inspired by the independence assumption and is able 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) [14], 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. According 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 [19]). 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 [20, 21] 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 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 formulation of the theory that justifies the method subject to clear formal assumptions. 3) We perform more 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. An extended version of this work with a more extensive analysis can be found in [22].

The paper is structured as follows: In Section 2, we introduce the used notations and assumptions, which are necessary for main theorem of this work stated in Section 3. An implementation that utilizes this theorem for inferring the causal direction is straightforward and further discussed in Section 4 and evaluated in various artificial and real-world data sets in Section 5.

2 Preliminaries

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

2.1 Notation and problem setting

The goal of this paper is to correctly identify cause and effect variable of given observations X and Y. 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'.

2.2 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

$$\mathbb{E}[(E - \phi(C))^2] \leq \mathbb{E}[(C - \psi(E))^2], \tag{1}$$

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. (1)

can be rewritten in terms of conditional variance as

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

2.3 Assumptions

First, recall that we assume throughout the paper that either X is the cause of Y or vise versa in an unconfounded sense, i.e. there is no common cause. To study the limit of an almost deterministic relation in a mathematically precise way, we consider a family of effect variables E_α by

$$E_\alpha := \phi(C) + \alpha N, \tag{2}$$

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). Here, ϕ is a function that is further specified below. Therefore, (2) does not, a priori, restrict the set of possible causal relations, because for any pair (C, E) one can define the noise N as the residual $N := E - \phi(C)$ and thus obtain $E_1 = E$ for any arbitrary function ϕ.

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

1. **Invertible function:** ϕ is a strictly monotonically increasing twice differentiable function $\phi : [0,1] \to [0,1]$. 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 $E \to 1-E$). We also assume that $\phi^{-1'}$ is bounded.

2. **Compact supports:** The distribution of C has compact support. Without loss of generality, we assume that 0 and 1 are the smallest and the largest values, respectively, 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 each c, $[n_-, n_+]$ is the smallest interval containing the support of $p_{N|c}$. This ensures us to know that $[\alpha n_-, 1+\alpha n_+]$ is the smallest interval containing the support of p_{E_α}. Then the shifted and rescaled variable

   $$\tilde{E}_\alpha := \frac{1}{1 + \alpha n_+ - \alpha n_-}(E_\alpha - \alpha n_-)$$

   attains 0 and 1 as minimum and maximum values and thus is equally scaled as C.

3. **Unbiased noise:** We use the convention $\mathbb{E}[N|c] = 0$ for all values c of C without loss of generality (this can easily be achieved by modifying ϕ). Then ϕ is just the conditional expectation, that is, $\phi(c) = \mathbb{E}[E|c]$.

4. **Unit noise variance:** The expected conditional noise variance is $\mathbb{E}[\mathrm{Var}[N|C]] = 1$, which is also not a proper restriction, because we can scale α accordingly since we are only interested in the limit $\alpha \to 0$.

5. **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

   $$\mathrm{Cov}[\phi', \mathrm{Var}[N|C]p_C] = 0. \tag{3}$$

   More explicitly, (3) reads:

   $$\int_0^1 \phi'(c)\,\mathrm{Var}[N|c]p_C(c)dc - \int_0^1 \phi'(c)dc \int_0^1 \mathrm{Var}[N|c]p_C(c)dc = 0. \tag{4}$$

The justification of (3) 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), (3) reduces to

$$\mathrm{Cov}[\phi', p_C] = 0, \tag{5}$$

which is an independence condition for deterministic relations stated in [23]. Conditions of similar type as (5) have been discussed and justified in [13]. 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 (4), observe first that it *cannot* be justified as stating some kind of 'independence' between p_C and $p_{E|C}$. To see this, note first that (4) states an uncorrelatedness of the two functions $c \mapsto \phi'(c)$ and $c \mapsto \mathrm{Var}[N|c]p_C(c)$. Since the latter function contains the map $c \mapsto \mathrm{Var}[N|c]$, which is a porperty of the conditional $p_{E|C}$ and not of the marginal p_C, it thus contains components from both p_C and $p_{E|C}$. Nevertheless, to justify (4) 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 modelling, that some structural equations may remain unchanged when other parts of a system change (see Chapter 2 in [3] for a literature review).[1]

[1] Note, however, that the assignment $E = \phi(C) + N$ is not a structural equation in a strict sense, because then C and N would need to be statistically independent.

A simple implication of (4) reads

$$\int_0^1 \phi'(c)\,\mathrm{Var}[N|c]p_C(c)dc = 1, \tag{6}$$

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

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

3 Theory

As introduced in Section 2.2, 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. Under the aforementioned assumptions, this can be stated if the noise variance is sufficiently small. 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

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

3.1 Error asymmetry theorem

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

Lemma 1 (Limit of variance ratio) *Let the assumptions 1-4 in Section 2.3 hold. Then the following limit holds:*

$$\lim_{\alpha\to 0} \frac{\mathbb{E}[\mathrm{Var}[C|\tilde{E}_\alpha]]}{\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]]} = \int_0^1 \frac{1}{\phi'(c)^2}\,\mathrm{Var}[N|c]p_C(c)dc \tag{7}$$

The formal proof is a bit technical and can be found in [22], however, the idea is quite simple if we 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:

$$\begin{aligned}\mathrm{Var}[C|\tilde{E}_\alpha = \phi(c)] &\approx \frac{1}{(\phi'(c))^2}\,\mathrm{Var}[\tilde{E}_\alpha|C = c] \\ &= \alpha^2 \frac{1}{(\phi'(c))^2}\,\mathrm{Var}[N|c]. \end{aligned}\tag{8}$$

Taking the expectation of (8) over C and recalling that Assumption 4 implies $\mathbb{E}[\mathrm{Var}[\tilde{E}_\alpha|C]] = \alpha^2\mathbb{E}[\mathrm{Var}[N|C]] = \alpha^2$ already yields (7).

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

Theorem 1 (Error Asymmetry) *Let the assumptions 1-5 in Section 2.3 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 functional relation in Assumption 3 is linear.

To show this, we make particular use of Lemma 1, Assumption 4 and the independence postulate to apply the Cauchy Schwarz inequality to (7), which gives us

$$\begin{aligned}&\int_0^1 \frac{1}{\phi'(c)^2}\,\mathrm{Var}[N|c]p_C(c)dc \\ &\quad\geq \left(\int_0^1 \frac{1}{\phi'(c)}\,\mathrm{Var}[N|c]p_C(c)dc\right)^2 \geq 1.\end{aligned}$$

A more detailed proof can be found in [22].

Note that if 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.

3.2 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 above theoretical insights. Presumably, this depends on features of ϕ, p_C, $p_{N|C}$ in a complicated way. However, the experiments in Section 5 suggest that the asymmetry often appears even for realistic noise levels.

4 Implementation

A causal inference algorithm that exploits Theorem 1 can be formulated in a straightforward way. 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.[2] We call this approach Regression Error based Causal Inference (RECI).

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. This speculative remark is 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 like removing isolated points in low-density regions, might improve the performance.

5 Experiments

For the experiments, we compared our method with three different related causal inference methods in various artificial and real-world data sets. In each evaluation, observations of two variables were given and the goal was to correctly identify cause and effect variables.

5.1 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 are

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

where $\beta \in \mathbb{R}$, $C \perp\!\!\!\perp N$ and N is non-Gaussian. For the experiments, we used a state-of-the-art implementation based on [25].

ANM The ANM approach assumes that

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

where f is nonlinear and $C \perp\!\!\!\perp N$. We used an implementation provided by [15], which uses a Gaussian process regression for the prediction and provides different methods for the evaluation of the causal direction.

IGCI The IGCI approach is able to determine the causal relationship in a deterministic setting

$$E = f(C),$$

under the 'independence assumption' $\mathrm{Cov}[\log f', p_C] = 0$. An implementation was also provided by [15], where we always tested all possible combinations of reference measures and information estimators. Generalizations of IGCI for non-deterministic relations are actually not known and we consider Assumption 5 in Section 2.3 as first step towards possibly more general formulations.

CURE CURE is based on the idea that the distribution p_C does not help for better regressing E on C, while the distribution p_E may help for better regressing C on E. 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 [14], where we used four internal repetition in the artificial data and eight in the real-world data, but only one overall repetition due to the high computation cost.

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 and does not rely on any independence tests.

In the experiments, we always used the same class of regression functions 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. 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 and at most two hidden layers

[2] [24], which appeared after submission, also infers the causal direction via regression, but from a minimum description length perspective.

The logistic and monomial functions cover rather simple regression models, which are probably not able to capture the true function f in most cases. On the other hand, support vector regression and neural networks should be complex enough to capture f. 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 always chose the default parameters, where the parameters of LOG, MON and POLY were fitted by minimizing the least-squared error. In order to have an accurate estimate of the MSE with respect to the function class, we averaged the MSE over all performed runs before comparing them.

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 [14], where a kernel density estimator with a Gaussian kernel is utilized. Note that CURE per default uses this preprocessing step. 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 results 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. Since we did not optimize the choice of functions and estimators, we only summarize the results of the best performing ANM and IGCI methods. Accordingly, we only summarize the results of the best performing MON, POLY and NN setups. Note that the results of IGCI and LiNGAM are constant over all runs since all data points were used.

5.2 Artificial data

For experiments with artificial data, we performed evaluations with simulated cause-effect pairs generated for a benchmark comparison in [15].

5.2.1 Simulated benchmark cause-effect pairs

The work of [15] provides simulated cause-effect pairs with randomly generated distributions and functional relationships under different conditions. As pointed out by [15], the scatter plots of these simulated data look similar to those of real-world data. We took the same data sets as used in [15] and extended the reported results with an evaluation with CURE, LiNGAM, RECI.

The data sets are categorized into four different categories:

- `SIM`: Pairs without confounders. The results are shown in Figure 1(a)
- `SIM-c`: A similar scenario as `SIM`, but with one additional confounder. The results are shown in Figure 1(b)
- `SIM-ln`: Pairs with low noise level without confounder. The results are shown in Figure 1(c)
- `SIM-G`: Pairs where the distributions of C and N are almost Gaussian without confounder. The results are shown in Figure 1(d)

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

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

and with confounder

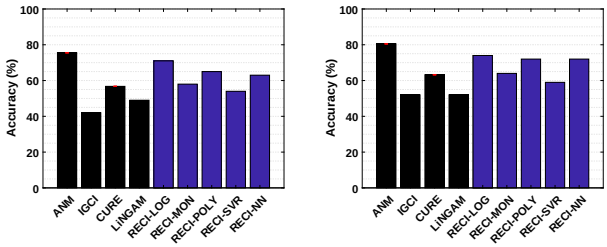
$$\begin{aligned} C' &\sim p_C, N \sim p_N, Z \sim p_Z \\ C'' &= f_C(C', Z) \\ N_C &\sim \mathcal{N}(0,\sigma_C), N_E \sim \mathcal{N}(0,\sigma_E) \\ C &= C'' + N_C, E = f_E(C'', Z, N) + N_E, \end{aligned}$$

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.[3] More details can be found in Appendix C of [15]. Note that adding noise to the cause (as it is done here) can also be considered as a kind of confounding.

In all data sets except `SIM-G`, we normalized the data for RECI. In `SIM-G`, Assumption 2 is violated, since these variables have no compact support due to the nearly Gaussian distribution. Therefore, we standardized the data for RECI and IGCI instead of normalizing them. The good results of RECI show some robustness with respect to a different scaling.

Generally, ANM performs the best in all data sets, while the performance gap, depending on the choice of the class of regression functions, between ANM and RECI is relatively small. On the other hand, in all data sets, RECI always outperforms IGCI, CURE and LiNGAM if a simple logistic or polynomial function is utilized for the regression. Even though Theorem 1 does not exclude cases of a high noise level, it makes a clear statement about low level noise. Therefore,

[3] 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.



(a) SIM (b) SIM-c

(c) SIM-ln (d) SIM-G

Figure 1: The best achieved performances of all methods in the artificial data sets.

as expected, RECI performs better in SIM-ln than in SIM and SIM-c due to the low noise level. In all cases, LiNGAM performs very poorly due to the violations of its core assumptions.

For RECI, similar function classes, such as $\sum_{i=0}^{3} a_i x^i$ and $\sum_{i=0}^{4} a_i x^i$, give similar performances. While LOG performed the best in SIM, SIM-c and SIM-ln, a polynomial function of the form $\sum_{i=0}^{2} a_i x^i$ performed the best in SIM-G.

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, it can be expected that RECI performs significantly better than ANM in cases where the dependency between C and N is very weak or non-existent seeing that RECI explicitly allows an independency. In comparison with IGCI, LiNGAM and CURE, RECI outperforms in almost all data sets. Note that [15] performed more extensive experiments and show more comparisons with ANM and IGCI in these data sets, where additional parameter configurations were tested.

5.3 Real-world data

In real-world data, the true causal relationship generally requires expert knowledge and can still lead to rather philosophical discussions in unclear cases. For our evaluations, we considered the commonly used cause-effect pairs (CEP) benchmark data sets. These benchmark data sets provided, at the time of these evaluations, 106 data sets with given cause and effect variables.[4] However, since we only consider a two variable problem, we omit the multivariate data sets, which leaves 100 data sets for the evaluations. Each data set comes with a corresponding weight. This is because several data sets are too similar to consider them as independent examples, hence they get lower weights. Therefore, the accuracy is a weighted sum over all cause-effect pairs. The experimental 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 experimental conditions as in [14].

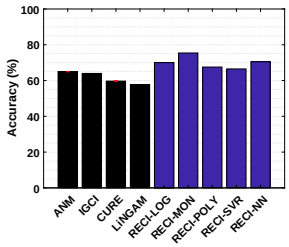


Figure 2: The best achieved performances of all methods in the real-world data sets.

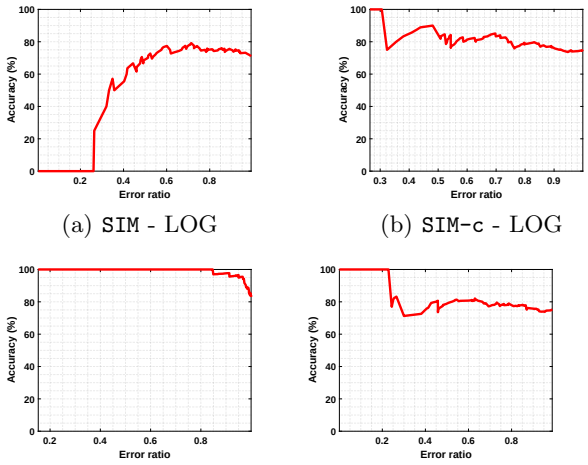
Figure 2 shows the results of the evaluations. In all cases, RECI performs better than all other methods. Surprisingly, the very simple monomial functions $ax^3 + c$ and $ax^2 + c$ perform the best, even though it is very unlikely that these functions are able to capture the true function ϕ. We obtained similar observations in the artificial data sets. The work of [15] provides further evaluations of ANM and IGCI in the original CEP data set with additional parameter configurations. 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 [14].

5.4 Error ratio as rejection criterion

It is not clear how to define a confidence measure for the decision of RECI. However, since Theorem 1 states that the correct causal direction has a smaller error, we evaluated the idea of using the error ratio

$$\frac{\min(\mathbb{E}[\mathrm{Var}[X|Y]], \mathbb{E}[\mathrm{Var}[Y|X]])}{\max(\mathbb{E}[\mathrm{Var}[X|Y]], \mathbb{E}[\mathrm{Var}[Y|X]])} \tag{9}$$

[4] The data set can be found on https://webdav.tuebingen.mpg.de/cause-effect/. More details and discussion about the causal relationship of the first 100 pairs can be found in [15].



(a) SIM - LOG (b) SIM-c - LOG

(c) SIM-G - POLY (d) CEP - MON

Figure 3: The performance of RECI in various data sets with respect to the error ratio (9) as rejection criterion. A small ratio indicates a high error difference.

as a rejection criterion for a decision. 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 $[0, 1]$. We re-evaluated the experimental results by considering only data sets where at most a certain error ratio was obtained by RECI. Figures 3(a)-3(d) show some examples of the performance of RECI if we use the error ratio as rejection criterion. In the figures, an error ratio of 0.2, for instance, can be seen as a surrogate indicator of RECI's accuracy if we only consider data sets where the error ratio was ≤ 0.2 and exclude data sets where the error ratio was > 0.2. In this sense, we can get an idea of how useful the error ratio is as rejection criterion. While Figures 3(b)-3(d) support the intuition that the smaller the error ratio, the higher the decision confidence, Figure 3(a) has a contradictive behavior. However, note that only a few data sets have an error ratio < 0.3 and the majority have an error ratio between 0.3 and 0.9, except for Figure 3(c) where the majority have an error ratio > 0.9.

5.5 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. Using algorithmic information theory, one can, for instance, show that the algorithmic independence of p_C and $p_{E|C}$ implies $K(p_C) + K(p_{E|C}) \leq K(p_E) + K(p_{C|E})$, if K denotes the description length of a distribution in the sense of Kolmogorov complexity, for details see Section 4.1.9 in [3]. 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.

Regarding the computational cost, we want to emphasize that RECI, depending on the implementation details, can have a significantly lower computational cost than ANM and CURE, while providing comparable or even better results. Further, it can be easily implemented and applied. More extensive experiments and comparisons can be found in [22].

6 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. Here, 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 method, 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.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR1666 and JSPS KAKENHI Grant Number JP17K00305, Japan.

References

[1] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[2] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

[3] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017. `http://www.math.ku.dk/~peters/`.

[4] Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, Tokyo, Japan, 2003.

[5] J. W. Comley and D. L. Dowe. General Bayesian networks and asymmetric languages. In *Proceedings of the Second Hawaii International Conference on Statistics and Related fields*, June 2003.

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

[7] X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL, 2006.

[8] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

[9] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696, Red Hook, NY, USA, June 2009. Curran Associates, Inc.

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

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

[12] P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 143–150, Corvallis, OR, USA, July 2010. AUAI Press.

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

[14] E. Sgouritsa, D. Janzing, P. Hennig, and B. Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Artificial Intelligence and Statistics*, pages 847–855, 2015.

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

[16] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 647–655, Arlington, Virginia, United States, June 2009. AUAI Press.

[17] D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[18] J. Lemeire and D. Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249, 7 2012.

[19] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. *Semi-supervised learning in causal and anticausal settings*, chapter 13, pages 129–141. Festschrift in Honor of Vladimir Vapnik. Springer-Verlag, 2013.

[20] P. Blöbaum, T. Washio, and S. Shimizu. Error asymmetry in causal and anticausal regression. *Behaviormetrika*, pages 1–22, 2017.

[21] P. Blöbaum, S. Shimizu, and T. Washio. A novel principle for causal inference in data with small error variance. In *ESANN*, 2017.

[22] P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B.. Schölkopf. Analysis of cause-effect inference by comparing regression errors. eprint `https://arxiv.org/abs/1802.06698`.

[23] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. Semi-supervised learning in causal and anticausal settings. In B. Schölkopf, Z. Luo, and V. Vovk, editors, *Empirical Inference*, Festschrift in Honor of Vladimir Vapnik, pages 129–141. Springer, 2013.

[24] A. Marx and J. Vreeken. Telling cause from effect using mdl-based local and global regression. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 307–316, Nov 2017.

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