

# Distinguishing cause from effect using observational data: methods and benchmarks

**Joris M. Mooij\***

[j.m.mooij@uva.nl](mailto:j.m.mooij@uva.nl)

*Institute for Informatics, University of Amsterdam  
Postbox 94323, 1090 GH Amsterdam, The Netherlands*

**Jonas Peters**

[peters@stat.math.ethz.ch](mailto:peters@stat.math.ethz.ch)

*Seminar for Statistics, ETH Zürich  
Rämistrasse 101, 8092 Zürich, Switzerland*

**Dominik Janzing**

[janzing@tuebingen.mpg.de](mailto:janzing@tuebingen.mpg.de)

*Max Planck Institute for Intelligent Systems  
Spemannstraße 38, 72076 Tübingen, Germany*

**Jakob Zscheischler**

[jzschi@bgc-jena.mpg.de](mailto:jzschi@bgc-jena.mpg.de)

*Max Planck Institute for Biogeochemistry  
Hans-Knöll-Straße 10, 07745 Jena, Germany*

**Bernhard Schölkopf**

[bs@tuebingen.mpg.de](mailto:bs@tuebingen.mpg.de)

*Max Planck Institute for Intelligent Systems  
Spemannstraße 38, 72076 Tübingen, Germany*

**Editor:**

## Abstract

The discovery of causal relationships from purely observational data is a fundamental problem in science. The most elementary form of such a causal discovery problem is to decide whether  $X$  causes  $Y$  or, alternatively,  $Y$  causes  $X$ , given joint observations of two variables  $X, Y$ . This was often considered to be impossible. Nevertheless, several approaches for addressing this bivariate causal discovery problem were proposed recently. In this paper, we present the benchmark data set `CauseEffectPairs` that consists of 88 different “cause-effect pairs” selected from 31 datasets from various domains. We evaluated the performance of several bivariate causal discovery methods on these real-world benchmark data and on artificially simulated data. Our empirical results provide evidence that additive-noise methods are indeed able to distinguish cause from effect using only purely observational data. In addition, we prove consistency of the additive-noise method proposed by [Hoyer et al. \(2009\)](#).

**Keywords:** Causal discovery, additive noise models, information-geometric causal inference

---

\*. Part of this work was done while JMM, JP and JZ were with the MPI Tübingen.

## 1. Introduction

An advantage of having knowledge about causal relationships rather than about statistical associations is that the former enables prediction of the effects of actions that perturb the observed system. While the gold standard for identifying causal relationships is controlled experimentation, in many cases, the required experiments are too expensive, unethical, or technically impossible to perform. The development of methods to identify causal relationships from purely observational data therefore constitutes an important field of research.

An observed statistical dependence between two variables  $X, Y$  can be explained by a causal influence from  $X$  to  $Y$  (“ $X \rightarrow Y$ ”), a causal influence from  $Y$  to  $X$  (“ $Y \rightarrow X$ ”), a (possibly unobserved) common cause that influences both  $X$  and  $Y$  (“confounding”), a (possibly unobserved) common effect that is caused by  $X$  and  $Y$  and is conditioned upon in data acquisition (“selection bias”), or combinations of these (see also Figure 1). Most state-of-the-art causal discovery algorithms that attempt to distinguish these cases based on observational data require that  $X$  and  $Y$  are part of a larger set of observed random variables influencing each other. In that case, under a genericity condition called “faithfulness”, (conditional) (in)dependences between subsets of observed variables allow one to draw partial conclusions regarding their causal relationships (Spirtes et al., 2000; Pearl, 2000; Richardson and Spirtes, 2002).

In this article, we focus on the *bivariate* case, assuming that only two variables, say  $X$  and  $Y$ , have been observed. We simplify the causal discovery problem by assuming no confounding, selection bias and feedback. We study how to distinguish  $X$  causing  $Y$  from  $Y$  causing  $X$  using only purely observational data, i.e., a finite sample of i.i.d. copies drawn from the joint distribution  $\mathbb{P}_{X,Y}$ . Some consider this task to be impossible. For example, Wasserman (2004, Remark 17.16) writes: “We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible with two variables.” Indeed, standard approaches based on (conditional) (in)dependences do not work here, as  $X$  and  $Y$  are typically dependent, and there are no further observed variables to condition on.

Nevertheless, the challenge of distinguishing cause from effect using only observational data has attracted increasing interest recently (Mooij and Janzing, 2010; Guyon et al., 2010, 2014), and knowledge of cause and effect can have implications on the applicability of semi-supervised learning and covariate shift adaptation (Schölkopf et al., 2012). A variety of causal discovery methods have been proposed in recent years (Friedman and Nachman, 2000; Kano and Shimizu, 2003; Shimizu et al., 2006; Sun et al., 2006, 2008; Hoyer et al., 2009; Mooij et al., 2009; Zhang and Hyvärinen, 2009; Janzing et al., 2010; Mooij et al., 2010; Daniusis et al., 2010; Mooij et al., 2011; Shimizu et al., 2011; Janzing et al., 2012; Hyvärinen and Smith, 2013) that were claimed to be able to solve this task. One could argue that all these approaches exploit the *complexity* of the marginal and conditional probability distributions, in one way or the other. On an intuitive level, the idea is that the factorization of the joint density (if it exists)  $p_{C,E}(c,e)$  of cause  $C$  and effect  $E$  into  $p_C(c)p_{E|C}(e|c)$  typically yields models of lower total complexity than the alternative factorization into  $p_E(e)p_{C|E}(c|e)$ . Although the notion of “complexity” is intuitively appealing, it is not obvious how it should be precisely defined. Indeed, each of these methods effectively uses its own measure of complexity.

The main contribution of this work is to provide extensive empirical results on how well two (families of) bivariate causal discovery methods work in practice: *Additive Noise Methods (ANM)* (originally proposed by Hoyer et al., 2009), and *Information Geometric Causal Inference (IGCI)* (originally proposed by Daniušis et al., 2010). Other contributions are a proof of the consistency of the original implementation of ANM (Hoyer et al., 2009) and a detailed description of the `CauseEffectPairs` benchmark data that we collected over the years for the purpose of evaluating bivariate causal discovery methods.

In the next subsection, we give a more formal definition of the causal discovery task we consider in this article. In Section 2 we give a review of ANM, an approach based on the assumed additivity of the noise, and describe various ways of implementing this idea for bivariate causal discovery. In Section 3, we review IGCI, a method that exploits the independence of the distribution of the cause and the functional relationship between cause and effect. This method is designed for the deterministic (noise-free) case, but has been reported to work on noisy data as well. Section 4 gives more details on the experiments that we have performed, the results of which are reported in Section 5. Appendix D describes the `CauseEffectPairs` benchmark data set that we used for assessing the accuracy of various methods. We conclude in Section 6.

### 1.1 Problem setting

Suppose that  $X, Y$  are two real-valued random variables with joint distribution  $\mathbb{P}_{X,Y}$ . This observational distribution corresponds with measurements of  $X$  and  $Y$  in an experiment in which  $X$  and  $Y$  are both (passively) observed. If an external intervention (i.e., from outside the system under consideration) changes some aspect of the system, then in general, this may lead to a change in the joint distribution of  $X$  and  $Y$ . In particular, we will consider a perfect intervention<sup>1</sup> “ $\text{do}(x)$ ” (or more explicitly: “ $\text{do}(X = x)$ ”) that forces the variable  $X$  to have the value  $x$ , and leaves the rest of the system untouched. We denote the resulting interventional distribution of  $Y$  as  $\mathbb{P}_{Y|\text{do}(x)}$ , a notation inspired by Pearl (2000). This interventional distribution corresponds with measurements of  $Y$  in an experiment in which  $X$  has been set to the value  $x$  by the experimenter, after which  $Y$  is observed. Similarly, we may consider a perfect intervention  $\text{do}(y)$  that forces  $Y$  to have the value  $y$ , leading to the interventional distribution  $\mathbb{P}_{X|\text{do}(y)}$  of  $X$ .

In general, the marginal distribution  $\mathbb{P}_X$  can be different from the interventional distribution  $\mathbb{P}_{X|\text{do}(y)}$  for some values of  $y$ , and similarly  $\mathbb{P}_Y$  can be different from  $\mathbb{P}_{Y|\text{do}(x)}$  for some values of  $x$ .

**Definition 1** *We say that  $X$  causes  $Y$  if  $\mathbb{P}_{Y|\text{do}(x)} \neq \mathbb{P}_{Y|\text{do}(x')}$  for some  $x, x'$ .*

Note that here we do not need to distinguish between *direct* and *indirect* causation, as we only consider the two variables  $X$  and  $Y$ . The *causal graph*<sup>2</sup> consists of two nodes, labeled  $X$  and  $Y$ . If  $X$  causes  $Y$ , the graph contains an edge  $X \rightarrow Y$ , and similarly, if

- 
1. In this paper we only consider perfect interventions. Different types of “imperfect” interventions can be considered as well, see e.g., Eberhardt and Scheines (2007); Eaton and Murphy (2007); Mooij and Heskes (2013).
  2. We will not give a precise definition of the causal graph here, as doing so would require distinguishing direct from indirect causation, which makes matters needlessly complicated for our purposes. The reader can consult Pearl (2000) for more details.

$Y$  causes  $X$ , the causal graph contains an edge  $Y \rightarrow X$ . If  $X$  causes  $Y$ , then generically we have that  $\mathbb{P}_{Y|\text{do}(x)} \neq \mathbb{P}_Y$ . Figure 1 illustrates how various causal relationships between  $X$  and  $Y$  generically give rise to different (in)equalities between marginal, conditional, and interventional distributions. When data from all these distributions is available, it becomes straightforward to infer the causal relationship between  $X$  and  $Y$  by checking which (in)equalities hold. Note that the list of possibilities in Figure 1 is not exhaustive, as (i) feedback relationships with a latent variable were not considered; (ii) combinations of the cases shown are possible as well, e.g., (d) can be considered to be the combination of (a) and (b), and both (e) and (f) can be combined with all other cases; (iii) more than one latent variable could be present.

Now suppose that we only have data from the observational distribution  $\mathbb{P}_{X,Y}$  (for example, because doing intervention experiments is too costly). Can we then still infer the causal relationship between  $X$  and  $Y$ ? We will simplify matters by considering only (a) and (b) in Figure 1 as possibilities. In other words, we assume that  $X$  and  $Y$  are dependent (i.e.,  $\mathbb{P}_{X,Y} \neq \mathbb{P}_X \mathbb{P}_Y$ ), there is no confounding (common cause of  $X$  and  $Y$ ), no selection bias (common effect of  $X$  and  $Y$  that is implicitly conditioned on), and no feedback between  $X$  and  $Y$  (a two-way causal relationship between  $X$  and  $Y$ ). Inferring the causal direction between  $X$  and  $Y$ , i.e., deciding which of the two cases (a) and (b) holds, using *only the observational distribution*  $\mathbb{P}_{X,Y}$  is the challenging task that we consider here. If, under certain assumptions, we can decide upon the causal direction, we say that the causal direction is *identifiable* from the observational distribution.

## 2. Additive Noise Models

In this section, we review a class of causal discovery methods that exploits *additivity* of the noise. We only consider the bivariate case here. More details and extensions to the multivariate case can be found in (Hoyer et al., 2009; Peters et al., 2014).

### 2.1 Theory

There is an extensive body of literature on causal modeling and causal discovery that assumes that effects are linear functions of their causes plus independent, Gaussian noise. These models are known as *Structural Equation Models* (SEM) (Wright, 1921; Bollen, 1989) and are popular in econometry, sociology, psychology and other fields. Although the assumptions of linearity and Gaussianity are mathematically convenient, they are not always realistic. More generally, one can define *Functional Models* (also known as *Structural Causal Models* (SCM) or *Non-Parametric Structural Equation Models* (NP-SEM)) (Pearl, 2000) in which effects are modeled as (possibly nonlinear) functions of their causes and latent noise variables. In general, if  $Y \in \mathbb{R}$  is a direct effect of a cause  $X \in \mathbb{R}$  and  $m$  latent causes  $\mathbf{U} = (U_1, \dots, U_m) \in \mathbb{R}^m$ , then it is intuitively reasonable to model this relationship as follows:

$$\begin{cases} Y = f(X, U_1, \dots, U_m), \\ X \perp\!\!\!\perp \mathbf{U}, \quad \mathbf{X} \sim p_X(x), \quad \mathbf{U} \sim p_{\mathbf{U}}(u_1, \dots, u_m) \end{cases} \quad (1)$$

where  $f : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}$  is a (possibly nonlinear) function, and  $p_X(x)$  and  $p_{\mathbf{U}}(u_1, \dots, u_m)$  are the joint densities of the observed cause  $X$  and latent causes  $\mathbf{U}$  (with respect to Lebesgue measure on  $\mathbb{R}$  and  $\mathbb{R}^m$ , respectively). The assumption that  $X$  and  $\mathbf{U}$  are independent is

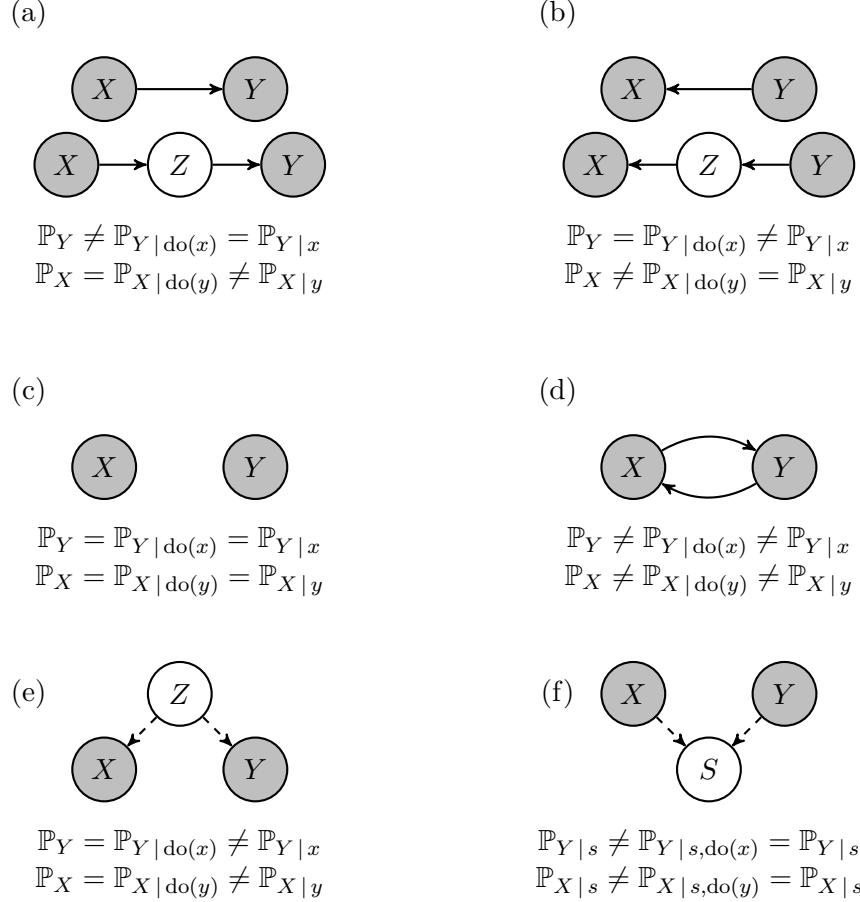


Figure 1: Several possible causal relationships between two observed variables  $X, Y$  and a single latent variable: (a)  $X$  causes  $Y$ ; (b)  $Y$  causes  $X$ ; (c)  $X, Y$  are not causally related; (d) feedback relationship; (e) a hidden confounder  $Z$  explains the observed dependence; (f) conditioning on a hidden selection variable  $S$  explains the observed dependence. All equalities are valid for all  $x, y$ , but the inequalities are only (generically) valid for some  $x, y$ . Note that all inequalities here are only *generic*, i.e., they do not necessarily hold, although they hold *typically*. In all situations except (c),  $X$  and  $Y$  are (generically) dependent, i.e.,  $\mathbb{P}_{X,Y} \neq \mathbb{P}_X \mathbb{P}_Y$ . The basic task we consider in this article is deciding between (a) and (b), using only data from  $\mathbb{P}_{X,Y}$ .

justified by the assumption that there is no confounding (i.e., there are no latent common causes of  $X$  and  $Y$ ), no selection bias (i.e., no common effect of  $X$  and  $Y$  that is conditioned upon), and no feedback between  $X$  and  $Y$ . As the latent causes  $\mathbf{U}$  are unobserved anyway, we can summarize their influence by one “effective” noise variable  $E \in \mathbb{R}$  such that

$$\begin{cases} Y = f_Y(X, E_Y) \\ X \perp\!\!\!\perp E_Y, \quad X \sim p_X(x), \quad E_Y \sim p_{E_Y}(e_Y). \end{cases} \quad (2)$$

One possible way to construct such  $E_Y$  and  $f_Y$  is to define the conditional cumulative density function  $F_{Y|x}(y) := \mathbb{P}(Y \leq y | X = x)$  and its inverse with respect to  $y$  for fixed  $x$ ,

$F_{Y|x}^{-1}$ . Then, one can define  $E_Y$  as the random variable

$$E_Y := F_{Y|X}(Y),$$

and the function  $f_Y$  by

$$f_Y(x, e) := F_{Y|x}^{-1}(e).$$

Assuming that these quantities are well-defined, it is easy to check (e.g., Hyvärinen and Pajunen, 1999, Theorem 1) that (2) holds with  $E_Y$  uniformly distributed on  $[0, 1]$ .

Model (2) does not yield any asymmetry between  $X$  and  $Y$ , as the same construction of an effective noise variable can be performed in the other direction. That gives another model for the joint density  $p_{X,Y}$ , where we could now interpret  $Y$  as the cause and  $X$  as the effect:

$$\begin{cases} X = f_X(Y, E_X) \\ Y \perp\!\!\!\perp E_X, \quad Y \sim p(y), \quad E_X \sim p(e_X). \end{cases} \quad (3)$$

A well-known example is the linear-Gaussian case:

**Example 1** Suppose that

$$\begin{cases} Y = \alpha X + E_X & X \sim \mathcal{N}(\mu_X, \sigma_X^2) \\ E_X \perp\!\!\!\perp X & E_X \sim \mathcal{N}(\mu_{E_X}, \sigma_{E_X}^2). \end{cases}$$

Then:

$$\begin{cases} X = \beta Y + E_Y & Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \\ E_Y \perp\!\!\!\perp Y & E_Y \sim \mathcal{N}(\mu_{E_Y}, \sigma_{E_Y}^2), \end{cases}$$

with

$$\begin{aligned} \beta &= \frac{\alpha \sigma_X^2}{\alpha^2 \sigma_X^2 + \sigma_{E_X}^2}, \\ \mu_Y &= \alpha \mu_X + \mu_{E_X}, \quad \sigma_Y^2 = \alpha^2 \sigma_X^2 + \sigma_{E_X}^2, \\ \mu_{E_Y} &= (1 - \alpha \beta) \mu_X - \beta \mu_{E_X}, \quad \sigma_{E_Y}^2 = (1 - \alpha \beta)^2 \sigma_X^2 + \beta^2 \sigma_{E_X}^2. \end{aligned}$$

Without having access to the interventional distributions, this symmetry apparently prevents us from drawing any conclusions regarding the causal direction.

However, by *restricting* the models (2) and (3) to have lower complexity, asymmetries can be introduced. The work of (Kano and Shimizu, 2003; Shimizu et al., 2006) showed that for *linear* models (i.e., where the functions  $f_X$  and  $f_Y$  are restricted to be linear), *non-Gaussianity* of the input and noise distributions actually allows one to distinguish the directionality of such functional models. Peters and Bühlmann (2014) recently proved that for linear models, Gaussian noise variables with equal variances also lead to identifiability. For high-dimensional variables, the structure of the covariance matrices can be exploited to achieve asymmetries (Janzing et al., 2010; Zscheischler et al., 2011).

More generally, Hoyer et al. (2009) showed that also *nonlinearity* of the functional relationships aids in identifying the causal direction, as long as the influence of the noise is additive. More precisely, they consider the following class of models:

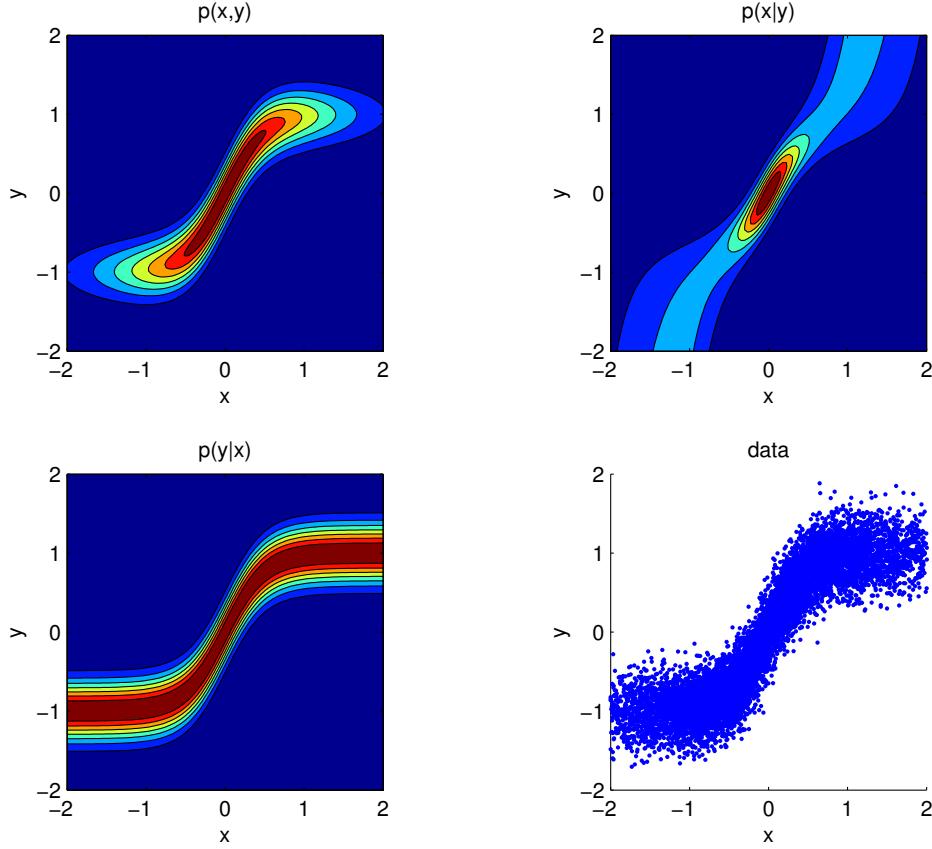


Figure 2: Identifiable ANM with  $Y = \tanh(X) + E$ , where  $X \sim \mathcal{N}(0, 1)$  and  $E \sim \mathcal{N}(0, 0.5^2)$ . Shown are contours of the joint and conditional distributions, and a scatter plot of data sampled from the model distribution. Note that the contour lines of  $p(y|x)$  only shift as  $x$  changes. On the other hand,  $p(x|y)$  differs by more than just its mean for different values of  $y$ .

**Definition 2** A tuple  $(p_X, p_{E_Y}, f_Y)$  consisting of a density  $p_X$ , a density  $p_{E_Y}$  with finite mean, and a measurable function  $f_Y : \mathbb{R} \rightarrow \mathbb{R}$ , defines a **bivariate additive noise model (ANM)**  $X \rightarrow Y$  by:

$$\begin{cases} Y = f_Y(X) + E_Y \\ X \perp\!\!\!\perp E_Y, \quad X \sim p_X, \quad E_Y \sim p_{E_Y}. \end{cases} \quad (4)$$

The induced density  $p(x, y)$  is said to satisfy an additive noise model  $X \rightarrow Y$ .

We are especially interested in cases where the additivity requirement introduces an asymmetry between  $X$  and  $Y$ :

**Definition 3** If the joint density  $p(x, y)$  satisfies an additive noise model  $X \rightarrow Y$ , but does not satisfy any additive noise model  $Y \rightarrow X$ , then we call the ANM  $X \rightarrow Y$  **identifiable**.

Hoyer et al. (2009) proved that additive noise models are generically identifiable. The intuition behind this result is that if  $p(x, y)$  satisfies an additive noise model  $X \rightarrow Y$ ,

then  $p(y|x)$  depends on  $x$  only through its mean, and all other aspects of this conditional distribution do not depend on  $x$ . On the other hand,  $p(x|y)$  will typically depend in a more complicated way on  $y$  (see also Figure 2). The parameters of an ANM have to be carefully “tuned” in order to obtain a non-identifiable ANM. We have already seen an example of such a non-identifiable ANM: the linear-Gaussian case (Example 1). A more exotic example with non-Gaussian distributions was given in (Peters et al., 2014, Example 25). Zhang and Hyvärinen (2009) proved that non-identifiable ANMs necessarily fall into one out of five classes. In particular, their result implies something that we might expect intuitively: if  $f$  is not injective, the ANM is identifiable. The results on identifiability of additive noise models can be extended to the multivariate case (Peters et al., 2014). Further, Mooij et al. (2011) showed that bivariate identifiability even holds generically when feedback is allowed (i.e., if both  $X \rightarrow Y$  and  $Y \rightarrow X$ ), at least when assuming noise and input distributions to be Gaussian. Peters et al. (2011) provide an extension for discrete variables. Zhang and Hyvärinen (2009) give an extension of the identifiability results allowing for an additional bijective transformation of the data, i.e., using a functional model of the form  $Y = \phi(f_Y(X) + E_Y)$ , with  $E_Y \perp\!\!\!\perp X$ , and  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  bijective, which they call the Post-NonLinear (PNL) model.

Following Hoyer et al. (2009), we postulate that:

**Postulate 1** Suppose we are given a joint density  $p(x,y)$  and we know that the causal structure is either that of (a) or (b) in Figure 1. If  $p(x,y)$  satisfies an identifiable additive noise model  $X \rightarrow Y$ , then it is highly likely that we are in case (a), i.e.,  $X$  causes  $Y$ .

This postulate should not be regarded as a rigorous statement, but rather as an empirical assumption: we cannot exactly quantify *how likely* the conclusion that  $X$  causes  $Y$  is, as there is always a possibility that  $Y$  causes  $X$  while  $p_{X,Y}$  happens to satisfy an identifiable additive noise model  $X \rightarrow Y$ . In general, that would require a special tuning of the distribution of  $X$  and the conditional distribution of  $Y$  given  $X$ , which is unlikely.

In this paper, we provide empirical evidence for this postulate. In the next subsection, we will discuss various ways of operationalizing this postulate.

## 2.2 Estimation methods

The following Lemma is helpful to test whether a density satisfies a bivariate additive noise model:

**Lemma 4** Given a joint density  $p(x,y)$  of two random variables  $X, Y$  such that the conditional expectation  $\mathbb{E}(Y | X = x)$  is well-defined for all  $x$ . Then,  $p(x,y)$  satisfies a bivariate additive noise model  $X \rightarrow Y$  if and only if  $E_Y := Y - \mathbb{E}(Y | X)$  has finite mean and is independent of  $X$ .

**Proof** Suppose that  $p(x,y)$  is induced by  $(p_X, p_U, f)$ , say  $Y = f(X) + U$  with  $X \perp\!\!\!\perp U$ ,  $X \sim p_X$ ,  $U \sim p_U$ . Then  $\mathbb{E}(Y | X = x) = f(x) + \nu$ , with  $\nu = \mathbb{E}(U)$ . Therefore,  $E_Y = Y - \mathbb{E}(Y | X) = Y - (f(X) + \nu) = U - \nu$  is independent of  $X$ . Conversely, if  $E_Y$  is independent of  $X$ ,  $p(x,y)$  is induced by the bivariate additive noise model  $(p_X, p_{E_Y}, x \mapsto \mathbb{E}(Y | X = x))$ . ■

In practice, we usually do not have the density  $p(x,y)$ , but rather a finite sample of it. In

that case, we can use the same idea for testing whether this sample comes from a distribution that satisfies an additive noise model: we estimate the conditional expectation  $\mathbb{E}(Y | X)$  by regression, and then test the independence of the residuals  $Y - \mathbb{E}(Y | X)$  and  $X$ .

Suppose we have two data sets, a *training* data set  $\mathcal{D}_N := \{(x_n, y_n)\}_{n=1}^N$  (for estimating the function) and a *test* data set  $\mathcal{D}'_N := \{(x'_n, y'_n)\}_{n=1}^N$  (for testing independence of residuals), both consisting of i.i.d. samples distributed according to  $p(x, y)$ . We will write  $\mathbf{x} = (x_1, \dots, x_N)$ ,  $\mathbf{y} = (y_1, \dots, y_N)$ ,  $\mathbf{x}' = (x'_1, \dots, x'_N)$  and  $\mathbf{y}' = (y'_1, \dots, y'_N)$ . We will consider two scenarios: the “data splitting” scenario where training and test set are independent (typically achieved by splitting a bigger data set into two parts), and the “data recycling” scenario in which the training and test data are identical (where we use the same data twice for different purposes: regression and independence testing).<sup>3</sup>

[Hoyer et al. \(2009\)](#) suggested the following procedure to test whether the data come from a distribution that satisfies an additive noise model.<sup>4</sup> By regressing  $Y$  on  $X$  using the training data  $\mathcal{D}_N$ , an estimate  $\hat{f}_Y$  for the regression function  $x \mapsto \mathbb{E}(Y | X = x)$  is obtained. Then, an independence test is used to estimate whether the predicted residuals are independent of the input, i.e., whether  $(Y - \hat{f}_Y(X)) \perp\!\!\!\perp X$ , using test data  $(\mathbf{x}', \mathbf{y}')$ . If the null hypothesis of independence is not rejected, one concludes that  $p(x, y)$  satisfies an additive noise model  $X \rightarrow Y$ . The regression procedure and the independence test can be freely chosen, but should be non-parametric.

There is a caveat, however: under the null hypothesis that  $p(x, y)$  indeed satisfies an ANM, the error in the estimated residuals may introduce a dependence between the predicted residuals  $\hat{\mathbf{e}}'_Y := \mathbf{y}' - \hat{f}_Y(\mathbf{x}')$  and  $\mathbf{x}'$  even if the true residuals  $\mathbf{y}' - \mathbb{E}(Y | X = \mathbf{x}')$  are independent of  $\mathbf{x}'$ . Therefore, the threshold for the independence test statistic has to be chosen with care: the standard threshold that would ensure consistency of the independence test on its own may be too tight. As far as we know, there are no theoretical results on the choice of that threshold that would lead to a consistent way to test whether  $p(x, y)$  satisfies an ANM  $X \rightarrow Y$ .

We circumvent this problem by assuming *a priori* that  $p(x, y)$  either satisfies an ANM  $X \rightarrow Y$ , or an ANM  $Y \rightarrow X$ , but not both. In that sense, the test statistics of the independence test can be directly compared, and no threshold needs to be chosen. This leads us to Algorithm 1 as a general scheme for identifying the direction of the ANM. In order to decide whether  $p(x, y)$  satisfies an additive noise model  $X \rightarrow Y$ , or an additive noise model  $Y \rightarrow X$ , we simply estimate the regression functions in both directions, estimate the corresponding residuals, measure the dependence of the residuals with respect to the input by some dependence measure  $C$ , and choose the direction that has the lowest dependence.

In principle, any consistent regression method can be used in Algorithm 1. Likewise, in principle any measure of dependence can be used in Algorithm 1 as score function. In the next subsections, we will consider in more detail some possible choices for the score function. Originally, [Hoyer et al. \(2009\)](#) proposed to use the *p*-value of the Hilbert Schmidt Independence Criterion (HSIC), a kernel-based non-parametric independence test. Alternatively, one can also use the HSIC statistic itself as a score, and we will show that this leads

---

3. [Kpotufe et al. \(2014\)](#) refer to these scenarios as “decoupled estimation” and “coupled estimation”, respectively.

4. They only considered the data recycling scenario, but the same idea could be applied to the data splitting scenario.

---

**Algorithm 1** General procedure to decide whether  $p(x, y)$  satisfies an additive noise model  $X \rightarrow Y$  or  $Y \rightarrow X$ .

---

**Input:**

1. an i.i.d. sample  $\mathcal{D}_N := \{(x_i, y_i)\}_{i=1}^N$  of  $X$  and  $Y$  (“training data”);
2. an i.i.d. sample  $\mathcal{D}'_N := \{(x'_i, y'_i)\}_{i=1}^N$  of  $X$  and  $Y$  (“test data”).

**Parameters:**

1. Regression method
2. Score function  $C : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$  for measuring dependence

**Output:** one of  $\{X \rightarrow Y, Y \rightarrow X, ?\}$ .

1. Use the regression method to obtain estimates:
    - (a)  $\hat{f}_Y$  of the regression function  $x \mapsto \mathbb{E}(Y | X = x)$ ,
    - (b)  $\hat{f}_X$  of the regression function  $y \mapsto \mathbb{E}(X | Y = y)$
 using the training data  $\mathcal{D}_N$ ;
  2. Use the estimated regression functions to predict residuals:
    - (a)  $\hat{e}'_Y := \mathbf{y}' - \hat{f}_Y(\mathbf{x}')$
    - (b)  $\hat{e}'_X := \mathbf{x}' - \hat{f}_X(\mathbf{y}')$
 from the test data  $\mathcal{D}'_N$ .
  3. Calculate the scores to measure independence of inputs and estimated residuals on the test data  $\mathcal{D}'_N$ :
    - (a)  $C_{X \rightarrow Y} := C(\mathbf{x}', \hat{e}'_Y)$ ;
    - (b)  $C_{Y \rightarrow X} := C(\mathbf{y}', \hat{e}'_X)$ ;
  4. Output:
 
$$\begin{cases} X \rightarrow Y & \text{if } C_{X \rightarrow Y} < C_{Y \rightarrow X}, \\ Y \rightarrow X & \text{if } C_{X \rightarrow Y} > C_{Y \rightarrow X}, \\ ? & \text{if } C_{X \rightarrow Y} = C_{Y \rightarrow X}. \end{cases}$$
- 

to a consistent procedure. Kpotufe et al. (2014) proposed to use as a score the sum of the estimated differential entropies of inputs and residuals and proved consistency of that procedure. For the Gaussian case, that is equivalent to the score considered in a high-dimensional context and shown to be consistent by Bühlmann et al. (2014). This Gaussian score is also strongly related to a Bayesian score originally proposed by Friedman and Nachman (2000). Finally, we will briefly discuss a Minimum Message Length score that was considered by

Mooij et al. (2010) and another idea (based on minimizing a dependence measure directly) proposed by Mooij et al. (2009).

### 2.2.1 HSIC-BASED SCORES

One possibility, originally proposed by Hoyer et al. (2009), is to use the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) for testing the independence of the estimated residuals with the inputs. See Appendix A.1 for a definition and basic properties of the HSIC independence test.

As proposed by Hoyer et al. (2009), one can use the  $p$ -value of the HSIC statistic under the null hypothesis of independence. This amounts to the following score function:

$$C(\mathbf{u}, \mathbf{v}) := \hat{p}_{\text{HSIC}_{\kappa_{\hat{\ell}(\mathbf{u})}, \kappa_{\hat{\ell}(\mathbf{v})}}}(\mathbf{u}, \mathbf{v}). \quad (5)$$

Here,  $\kappa_\ell$  is a kernel with parameters  $\ell$ , that are estimated from the data. A low HSIC  $p$ -value indicates that we should reject the null hypothesis of independence. Another possibility is to use the HSIC value itself (instead of its  $p$ -value):

$$\widehat{C}(\mathbf{u}, \mathbf{v}) := \widehat{\text{HSIC}}_{\kappa_{\hat{\ell}(\mathbf{u})}, \kappa_{\hat{\ell}(\mathbf{v})}}(\mathbf{u}, \mathbf{v}). \quad (6)$$

An even simpler option is to use a fixed kernel  $k$ :

$$C(\mathbf{u}, \mathbf{v}) := \widehat{\text{HSIC}}_{k,k}(\mathbf{u}, \mathbf{v}). \quad (7)$$

In Appendix A, we prove that under certain technical assumptions (in particular, the kernel  $k$  should be characteristic), Algorithm 1 with score function (7) is a consistent procedure for inferring the direction of the ANM:

**Theorem 5** *Let  $X, Y$  be two real-valued random variables with joint distribution  $\mathbb{P}_{X,Y}$  that either satisfies an additive noise model  $X \rightarrow Y$ , or  $Y \rightarrow X$ , but not both. Suppose we are given sequences of training data sets  $\mathcal{D}_N$  and test data sets  $\mathcal{D}'_N$  (in either the data splitting or the data recycling scenario). Let  $k, l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be two bounded non-negative Lipschitz-continuous characteristic kernels. If the regression procedure used in Algorithm 1 is suitable (c.f. Definition 14) for both  $\mathbb{P}_{X,Y}$  and  $\mathbb{P}_{Y,X}$ , then Algorithm 1 with score (7) is a consistent procedure for estimating the direction of the additive noise model.*

**Proof** See Appendix A. The main technical difficulty consists of the fact that the error in the estimated regression function introduces a dependency between the cause and the estimated residuals. We overcome this difficulty by showing that the dependence is so weak that its influence on the test statistic vanishes asymptotically. ■

In the data splitting case, weakly universally consistent regression methods (Györfi et al., 2002) are suitable. In the data recycling scenario, any regression method that satisfies (31) is suitable.

### 2.2.2 ENTROPY-BASED SCORES

Instead of explicitly testing for independence of residuals and inputs, one can instead use the sum of their differential entropies as a score function (Kpotufe et al., 2014). This can be easily seen using Lemma 1 of Kpotufe et al. (2014), which we reproduce here because it is very instructive:

**Lemma 6** Consider a joint distribution of  $X, Y$  with density  $p(x, y)$ . For arbitrary functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  we have:

$$H(X) + H(Y - f(X)) = H(Y) + H(X - g(Y)) - (I(X - g(Y), Y) - I(Y - f(X), X)).$$

where  $H(\cdot)$  denotes differential Shannon entropy, and  $I(\cdot, \cdot)$  denotes differential mutual information (Cover and Thomas, 2006). ■

The proof is a simple application of the chain rule of differential entropy. If  $p(x, y)$  satisfies an identifiable additive noise model  $X \rightarrow Y$ , then there exists a function  $f$  with  $I(Y - f(X), X) = 0$  (e.g., the regression function  $x \mapsto \mathbb{E}(Y | X = x)$ ), but  $I(X - g(Y), Y) > 0$  for any function  $g$ . Therefore, one can use Algorithm 1 with score function

$$C(\mathbf{u}, \mathbf{v}) := \hat{H}(\mathbf{u}) + \hat{H}(\mathbf{v}) \quad (8)$$

in order to estimate the causal direction, using any estimator  $\hat{H}(\cdot)$  of the differential Shannon entropy.

Kpotufe et al. (2014) note that the advantage of score (8) (based on marginal entropies) over score (6) (based on dependence) is that marginal entropies are cheaper to estimate than dependence (or mutual information). This is certainly the case when considering computation time. However, as we will later, a disadvantage of relying on differential entropy estimators is that these typically are quite sensitive to discretization effects.

### 2.2.3 GAUSSIAN SCORE

The differential entropy of a random variable  $X$  can be upper bounded in terms of its variance (see e.g., Cover and Thomas, 2006, Theorem 8.6.6):

$$H(X) \leq \frac{1}{2} \log(2\pi e) + \frac{1}{2} \log \text{Var}(X), \quad (9)$$

where identity holds in case  $X$  has a Gaussian distribution. Assuming that  $p(x, y)$  satisfies an identifiable Gaussian additive noise model  $X \rightarrow Y$  (with Gaussian input and Gaussian noise distributions), we therefore conclude from Lemma 6:

$$\begin{aligned} \log \text{Var}(X) + \log \text{Var}(Y - \hat{f}(X)) &= 2H(X) + 2H(Y - \hat{f}(X)) - 2\log(2\pi e) \\ &< 2H(Y) + 2H(X - \hat{g}(Y)) - 2\log(2\pi e) \\ &\leq \log \text{Var}Y + \log \text{Var}(X - \hat{g}(Y)) \end{aligned}$$

for any function  $g$ . So in that case, we can use Algorithm 1 with score function

$$C(\mathbf{u}, \mathbf{v}) := \log \widehat{\text{Var}}(\mathbf{u}) + \log \widehat{\text{Var}}(\mathbf{v}). \quad (10)$$

This score was also considered recently by Bühlmann et al. (2014) and shown to lead to a consistent estimation procedure under suitable assumptions.

### 2.2.4 BAYESIAN SCORES

Deciding the direction of the ANM can also be done by applying standard Bayesian model selection. As an example, for the ANM  $X \rightarrow Y$ , one can consider a generative model that models  $X$  as a Gaussian, and  $Y$  as a Gaussian Process conditional on  $X$ . For the ANM  $Y \rightarrow X$ , one considers a similar model with the roles of  $X$  and  $Y$  reversed. Bayesian model selection is performed by calculating the evidences (marginal likelihoods) of these two models, and preferring the model with larger evidence. This is actually a special case (the bivariate case) of an approach proposed by Friedman and Nachman (2000).<sup>5</sup> Considering the negative log marginal likelihoods leads to the following score for the ANM  $X \rightarrow Y$ :

$$C_{X \rightarrow Y} := \min_{\mu, \tau^2, \boldsymbol{\theta}, \sigma^2} (-\log \mathcal{N}(\mathbf{x} | \mu \mathbf{1}, \tau^2 \mathbf{I}) - \log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}}(\mathbf{x}) + \sigma^2 \mathbf{I})) , \quad (11)$$

and a similar expression for  $C_{Y \rightarrow X}$ , the score of the ANM  $Y \rightarrow X$ . Here,  $\mathbf{K}_{\boldsymbol{\theta}}(\mathbf{x})$  is the  $N \times N$  kernel matrix  $K_{ij} = k_{\boldsymbol{\theta}}(x_i, x_j)$  for a kernel with parameters  $\boldsymbol{\theta}$  and  $\mathcal{N}(\cdot | \mu, \Sigma)$  denotes the density of a multivariate normal distribution with mean  $\mu$  and covariance matrix  $\Sigma$ . Typically, one optimizes the hyperparameters  $(\mu, \tau, \boldsymbol{\theta}, \sigma)$  instead of integrating them out for computational reasons. Note that this method skips the explicit regression step, instead it (implicitly) integrates over all possible regression functions. Also, it does not distinguish the data splitting and data recycling scenarios, instead it uses the data directly to calculate the marginal likelihood. Therefore, the structure of the algorithm is slightly different (Algorithm 2). In Appendix B we show that this score is actually closely related to the Gaussian score considered in Section 2.2.3.

### 2.2.5 MINIMUM MESSAGE LENGTH SCORES

In a similar vein as Bayesian marginal likelihoods can be interpreted as measuring likelihood in combination with a complexity penalty, Minimum Message Length (MML) techniques can be used to construct scores that incorporate a trade-off between model fit (likelihood) and model complexity (Grünwald, 2007). Asymptotically, as the number of data points tends to infinity, one would expect the model fit to outweigh the model complexity, and therefore by Lemma 6, simple comparison of MML scores should be enough to identify the direction of an identifiable additive noise model.

A particular MML score was considered by Mooij et al. (2010). This is a special case (referred to in Mooij et al. (2010) as “AN-MML”) of their more general framework that allows for non-additive noise. Like (11), the score is a sum of two terms, one corresponding with the marginal density  $p(x)$  and the other with the conditional density  $p(y | x)$ :

$$C_{X \rightarrow Y} := \mathcal{L}(\mathbf{x}) + \min_{\boldsymbol{\theta}, \sigma^2} (-\log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}}(\mathbf{x}) + \sigma^2 \mathbf{I})) . \quad (12)$$

The second term is an MML score for the conditional density  $p(y | x)$ , and is identical to the the conditional density term in (11). The MML score  $\mathcal{L}(\mathbf{x})$  for the marginal density  $p(x)$  is derived as an asymptotic expansion based on the Minimum Message Length principle for a

---

5. Friedman and Nachman (2000) even hint at using this method for inferring causal relationships, although it seems that they only thought of cases where the functional dependence of the effect on the cause was not injective.

---

**Algorithm 2** Procedure to decide whether  $p(x, y)$  satisfies an additive noise model  $X \rightarrow Y$  or  $Y \rightarrow X$  suitable for Bayesian or MML model selection.

---

**Input:**

1. real-valued random variables  $X, Y$ ;
2. an i.i.d. sample  $\mathcal{D}_N := \{(x_i, y_i)\}_{i=1}^N$  of  $X$  and  $Y$  (“data”);
3. Score function  $C : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$  for measuring model fit and model complexity

**Output:** one of  $\{X \rightarrow Y, Y \rightarrow X\}$ .

1. (a) calculate  $C_{X \rightarrow Y} = C(\mathbf{x}, \mathbf{y})$   
(b) calculate  $C_{Y \rightarrow X} = C(\mathbf{y}, \mathbf{x})$
2. Output

$$\begin{cases} X \rightarrow Y & \text{if } C_{X \rightarrow Y} < C_{Y \rightarrow X}, \\ Y \rightarrow X & \text{if } C_{X \rightarrow Y} > C_{Y \rightarrow X}, \\ ? & \text{if } C_{X \rightarrow Y} = C_{Y \rightarrow X}. \end{cases}$$


---

mixture-of-Gaussians model (Figueiredo and Jain, 2002):

$$\mathcal{L}(\mathbf{x}) = \min_{\boldsymbol{\eta}} \left( \sum_{j=1}^k \log \left( \frac{N\alpha_j}{12} \right) + \frac{k}{2} \log \frac{N}{12} + \frac{3k}{2} - \log p(\mathbf{x} | \boldsymbol{\eta}) \right), \quad (13)$$

where  $p(\mathbf{x} | \boldsymbol{\eta})$  is a Gaussian mixture model:  $p(x_i | \boldsymbol{\eta}) = \sum_{j=1}^k \alpha_j \mathcal{N}(x_i | \mu_j, \sigma_j^2)$ . The optimization problem (13) is solved numerically by means of the algorithm proposed by Figueiredo and Jain (2002), using a small but nonzero value ( $10^{-4}$ ) of the regularization parameter.

Comparing this score with the Bayesian score (11), the main difference is that the former uses a more complicated mixture-of-Gaussians model for the marginal density, whereas (11) uses a simple Gaussian model. We can use (12) in combination with Algorithm 2 in order to estimate the direction of an identifiable additive noise model.

#### 2.2.6 MINIMIZING HSIC DIRECTLY

One can try to apply the idea of combining regression and independence testing into a single procedure (as achieved with the Bayesian score described in Section 2.2.4, for example) more generally. Indeed, a score that measures the dependence between the residuals  $\mathbf{y}' - f_Y(\mathbf{x}')$  and the inputs  $\mathbf{x}'$  can be minimized with respect to the function  $f_Y$ . Mooij et al. (2009) proposed to minimize  $\widehat{\text{HSIC}}(\mathbf{x}, \mathbf{y} - f(\mathbf{x}))$  with respect to the function  $f$ . However, the optimization problem with respect to  $f$  turns out to be a challenging non-convex optimization problem with multiple local minima, and there are no guarantees to find the global minimum. In addition, the performance depends strongly on the selection of suitable kernel bandwidths, for which no automatic procedure is known in this context. Finally,

proving consistency of such a method might be challenging, as the minimization may introduce strong dependences between the residuals. Therefore, we do not consider this method here.

### 3. Information-Geometric Causal Inference

In this section, we review a class of causal discovery methods that exploits independence of the distribution of the cause and the conditional distribution of the effect given the cause. It nicely complements causal inference based on additive noise by employing asymmetries between cause and effect that have nothing to do with noise.

#### 3.1 Theory

Information-geometric causal inference (IGCI) is an approach that builds upon the assumption that for  $X \rightarrow Y$  the marginal distribution  $\mathbb{P}_X$  contains no information about the conditional  $\mathbb{P}_{Y|X}$  and *vice versa*, since they represent independent mechanisms. As [Janzing and Schölkopf \(2010\)](#) illustrated for several toy examples, the conditional and marginal distributions  $\mathbb{P}_Y, \mathbb{P}_{X|Y}$  may then contain information about each other, but it is hard to formalize in what sense this is the case for scenarios that go beyond simple toy models. IGCI is based on the strong assumption that  $X$  and  $Y$  are deterministically related by a bijective function  $f$ , that is,  $Y = f(X)$  and  $X = f^{-1}(Y)$ . Although its practical applicability is limited to causal relations with sufficiently small noise, IGCI provides a setting in which the independence of  $\mathbb{P}_X$  and  $\mathbb{P}_{Y|X}$  provably implies well-defined dependences between  $\mathbb{P}_Y$  and  $\mathbb{P}_{X|Y}$ .

To introduce IGCI, note that the deterministic relation  $Y = f(X)$  implies that the conditional  $\mathbb{P}_{Y|X}$  has no density  $p(y|x)$ , but it can be represented using  $f$  via

$$\mathbb{P}(Y = y|X = x) = \begin{cases} 1 & \text{if } y = f(x) \\ 0 & \text{otherwise.} \end{cases}$$

The fact that  $\mathbb{P}_X$  and  $\mathbb{P}_{Y|X}$  contain no information about each other then translates into the statement that  $\mathbb{P}_X$  and  $f$  contain no information about each other.

Before sketching a more general formulation of IGCI ([Daniušis et al., 2010; Janzing et al., 2012](#)), we begin with the most intuitive case where  $f$  is a strictly monotonously increasing differentiable bijection of  $[0, 1]$ . We then assume: that the following equality is approximately satisfied:

$$\int_0^1 \log f'(x)p(x) dx = \int_0^1 \log f'(x) dx. \quad (14)$$

To see why (14) is an independence between function  $f$  and input density  $p_X$ , we interpret  $x \mapsto \log f'(x)$  and  $x \mapsto p(x)$  as random variables on  $[0, 1]$ . Then the difference between the two sides of (14) is the covariance of these two random variables with respect to the uniform distribution on  $[0, 1]$ . As shown in Section 2 in ([Daniušis et al., 2010](#)),  $p_Y$  is then related to the inverse function  $f^{-1}$  in the sense that

$$\int_0^1 \log f^{-1'}(y)p(y) dy \geq \int_0^1 \log f^{-1'}(y) dy,$$

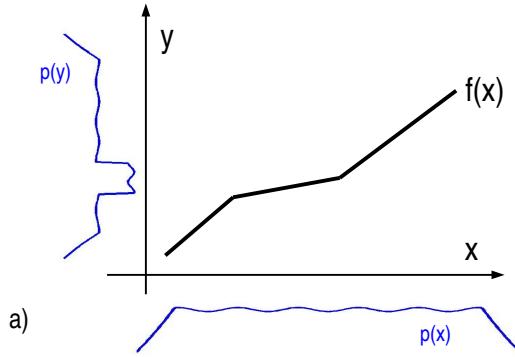


Figure 3: Illustration of the basic intuition behind IGCI. If the density  $p_X$  of the cause  $X$  is not correlated with the slope of  $f$ , then the density  $p_Y$  tends to be high in regions where  $f$  is flat (and  $f^{-1}$  is steep). Source: Janzing et al. (2012)

with equality if and only if  $f'$  is constant. Hence,  $\log f'^{-1}$  and  $p_Y$  are positively correlated. Intuitively, this is because the density  $p_Y$  tends to be high in regions where  $f$  is flat and  $f^{-1}$  is steep (see also Figure 3). Hence, we have shown that  $\mathbb{P}_Y$  contains information about  $f^{-1}$  and hence about  $\mathbb{P}_{X|Y}$  whenever  $\mathbb{P}_X$  does not contain information about  $\mathbb{P}_{Y|X}$  (in the sense that (14) is satisfied), except for the trivial case where  $f$  is linear.

To employ this asymmetry, IGCI introduces the expressions

$$C_{X \rightarrow Y} := \int_0^1 \log f'(x)p(x)dx \quad (15)$$

$$C_{Y \rightarrow X} := \int_0^1 \log f'^{-1}(y)p(y)dy = -C_{X \rightarrow Y}. \quad (16)$$

Since the right hand side of (14) is smaller than zero due to concavity of the logarithm (exactly zero only for constant  $f$ ), IGCI infers  $X \rightarrow Y$  whenever  $C_{X \rightarrow Y}$  is negative. Section 3.5 in (Daniusis et al., 2010) also shows that

$$C_{X \rightarrow Y} = H(Y) - H(X),$$

i.e., the decision rule considers the variable with lower entropy as the effect. The idea is that the function introduces new irregularities to a distribution rather than smoothing the irregularities of the distribution of the cause.

*Generalization to other reference measures:* In the above version of IGCI the uniform distribution on  $[0, 1]$  plays a special role because it is the distribution with respect to which uncorrelatedness between  $p_X$  and  $\log f'$  is defined. The idea can be generalized to other reference distributions. How to choose the right one for a particular inference problem is a difficult question which goes beyond the scope of this article. From a high-level perspective, it is comparable to the question of choosing the right kernel for kernel-based machine learning algorithms; it also is an a-priori structure of the range of  $X$  and  $Y$  without which the inference problem is not feasible.

Let  $u_X$  and  $u_Y$  be densities of  $X$  and  $Y$ , respectively, that we call “reference densities”. One may think of Gaussians as an example for reasonable choices other than the uniform distribution. Let  $u_f$  be the image of  $u_X$  under  $f$  and  $u_{f^{-1}}$  be the image of  $u_Y$  under  $f^{-1}$ . Then we postulate the following generalization of (14):

**Postulate 2** *If  $X$  causes  $Y$  via a deterministic bijective function  $f$  such that the densities  $u_{f^{-1}}$  exist then*

$$\int \log \frac{u_{f^{-1}}(x)}{u(x)} p(x) dx = \int \log \frac{u_{f^{-1}}(x)}{u(x)} u(x) dx. \quad (17)$$

In analogy to the remarks above, this can also be interpreted as uncorrelatedness of the functions  $\log(u_{f^{-1}}/u_X)$  and  $p_X$ . Again, we postulate this because the former expression is a property of the function  $f$  alone (and the reference densities) and should thus be unrelated to  $p_X$ . The special case (14) can be obtained by taking the uniform distribution on  $[0, 1]$  for  $u_X$  and  $u_Y$ .

As generalization of (15,16) we define<sup>6</sup>

$$C_{X \rightarrow Y} := \int \log \frac{u_{f^{-1}}(x)}{u(x)} p(x) dx \quad (18)$$

$$C_{Y \rightarrow X} := \int \log \frac{u_f(y)}{u(y)} p(y) dy = \int \log \frac{u(x)}{u_{f^{-1}}(x)} p(x) dx = -C_{X \rightarrow Y}, \quad (19)$$

where the second equality in (19) follows by substitution of variables. Again, the postulated independence implies  $C_{X \rightarrow Y} \leq 0$  since the right hand side of (17) coincides with  $-D(u_X \| u_{f^{-1}})$  where  $D(\cdot \| \cdot)$  denotes Kullback-Leibler divergence. Hence, we also infer  $X \rightarrow Y$  whenever  $C_{X \rightarrow Y} < 0$ . Daniušis et al. (2010) further show (see eq. (8) therein) that

$$C_{X \rightarrow Y} = D(p_X \| u_X) - D(p_Y \| u_Y).$$

Hence, our decision rule amounts to inferring that the density of the cause is closer to its reference density. This decision rule gets quite simple, for instance, if  $u_X$  and  $u_Y$  are Gaussians with the same mean and variance as  $p_X$  and  $p_Y$ , respectively. Then it again amounts to inferring  $X \rightarrow Y$  whenever  $X$  has larger entropy than  $Y$  after rescaling both  $X$  and  $Y$  to the same variance.

### 3.2 Estimation methods

The specification of the reference measure is essential for IGCI. We describe the implementation for two different choices:

1. *Uniform distribution*: scale and shift  $X$  and  $Y$  such that extrema are mapped onto 0 and 1.
2. *Gaussian distribution*: scale  $X$  and  $Y$  to variance 1.

Given this preprocessing step (see Section 3.5 in (Daniušis et al., 2010)), there are different options for estimating  $C_{X \rightarrow Y}$  and  $C_{Y \rightarrow X}$  from empirical data:

---

6. Note that the formulation in Section 2.3 in (Daniušis et al., 2010) is more general because it uses *manifolds* of reference densities instead of a single density.

---

**Algorithm 3** General procedure to decide whether  $\mathbb{P}_{X,Y}$  is generated by a deterministic function from  $X$  to  $Y$  or from  $Y$  to  $X$ .

---

**Input:** an i.i.d. sample  $\mathcal{D}_N := \{(x_i, y_i)\}_{i=1}^N$  of  $X$  and  $Y$

**Parameters:**

1. normalization procedure, where the scaling factor is either given by the range or the standard deviation of each variable.
2. Score function  $\hat{C}_{X \rightarrow Y}, \hat{C}_{Y \rightarrow X}$
3. Significance threshold  $\alpha > 0$ .

**Output:** one of  $\{X \rightarrow Y, Y \rightarrow X, ?\}$ .

1. Compute  $\hat{C}_{X \rightarrow Y}$  and  $C_{Y \rightarrow X}$  from  $\mathcal{D}_N$ .

2. Infer

$$\begin{cases} X \rightarrow Y & \text{if } \hat{C}_{X \rightarrow Y} < \hat{C}_{Y \rightarrow X} - \alpha, \\ Y \rightarrow X & \text{if } C_{X \rightarrow Y} > C_{Y \rightarrow X} + \alpha, \\ ? & \text{otherwise.} \end{cases}$$


---

1. *Slope-based estimator:*

$$\hat{C}_{X \rightarrow Y} := \frac{1}{N-1} \sum_{j=1}^{N-1} \log \frac{|y_{j+1} - y_j|}{x_{j+1} - x_j}, \quad (20)$$

where we assumed the  $x_j$  to be in increasing order. Since empirical data are noisy, the  $y$ -values need not be in the same order.  $\hat{C}_{Y \rightarrow X}$  is given by exchanging the roles of  $X$  and  $Y$ .

2. *Entropy-based estimator:*

$$\hat{C}_{X \rightarrow Y} := \hat{H}(Y) - \hat{H}(Y), \quad (21)$$

where  $\hat{H}(\cdot)$  denotes some entropy estimator.

The theoretical equivalence between these estimators breaks down on empirical data not only due to finite sample effects but also because of noise. For the slope based estimator, we even have

$$\hat{C}_{X \rightarrow Y} \neq -\hat{C}_{Y \rightarrow X},$$

and thus need to compute both terms separately.

Note that the IGCI implementations discussed here make sense only for continuous variables. This is because the difference quotients are undefined if a value occurs twice. In many empirical data sets, however, the discretization is not fine enough to guarantee this. A very preliminary heuristic removes repeated occurrences by removing data points, but a conceptually cleaner solution would be, for instance, the following procedure: Let  $\tilde{x}_j$  with

$j \leq \tilde{N}$  be the ordered values after removing repetitions and let  $\tilde{y}_j$  denote the corresponding  $y$ -values. Then we replace (20) with

$$\hat{C}_{X \rightarrow Y} := \sum_{j=1}^{\tilde{N}} n_j \log \frac{|\tilde{y}_{j+1} - \tilde{y}_j|}{\tilde{x}_{j+1} - \tilde{x}_j}, \quad (22)$$

where  $n_j$  denotes the number of occurrences of  $\tilde{x}_j$  in the original data set. Here we have ignored the problem of repetitions of  $y$ -values since they are less likely, because they are not ordered if the relation between  $X$  and  $Y$  is noisy (and for bijective deterministic relations, they only occur together with repetitions of  $x$  anyway).

## 4. Experiments

In this section we describe the data that we used for evaluation, implementation details for various methods, and our evaluation criteria. The results of the empirical study will be presented in Section 5.

### 4.1 Implementation details

The complete source code to reproduce our experiments will be made available online under an open source license on the homepage of the first author, <http://www.jorismooij.nl>. We used MatLab on a Linux platform, and made use of external libraries GPML (Rasmussen and Nickisch, 2010) for GP regression and ITE (Szabó, 2014) for entropy estimation. For parallelization, we used the convenient command line tool GNU parallel (Tange, 2011).

#### 4.1.1 REGRESSION

For regression, we used standard Gaussian Process (GP) Regression (Rasmussen and Williams, 2006), using the GPML implementation (Rasmussen and Nickisch, 2010). We used a squared exponential covariance function, constant mean function, and an additive Gaussian noise likelihood. We used the FITC approximation (Quiñonero-Candela and Rasmussen, 2005) as an approximation for exact GP regression in order to reduce computation time. We found that 100 FITC points distributed on a linearly spaced grid greatly reduce computation time (which scales cubically with the number of data points for exact GP regression) without introducing a noticeable approximation error. Therefore, we used this setting as a default for the GP regression.

#### 4.1.2 ENTROPY ESTIMATION

We tried many different empirical entropy estimators, see Table 1. The first method, `1sp`, uses a so-called “1-spacing” estimate (e.g., Kraskov et al., 2004):

$$\hat{H}(\mathbf{x}) := \psi(N) - \psi(1) + \frac{1}{N-1} \sum_{i=1}^{N-1} \log |x_{i+1} - x_i|, \quad (23)$$

where the  $x$ -values should be ordered ascendingly, i.e.,  $x_i \leq x_{i+1}$ , and  $\psi$  is the digamma function (i.e., the logarithmic derivative of the gamma function:  $\psi(x) = d/dx \log \Gamma(x)$ ,

Table 1: Entropy estimation methods. ‘‘ITE’’ refers to the Information Theoretical Estimators Toolbox (Szabó, 2014). The first group of entropy estimators is nonparametric, the second group makes additional parametric assumptions on the distribution of the data.

Name	Implementation	References
1sp	based on (23)	(Kraskov et al., 2004)
3NN	ITE: Shannon_kNN_k	(Kozachenko and Leonenko, 1987)
sp1	ITE: Shannon_spacing_V	(Vasicek, 1976)
sp2	ITE: Shannon_spacing_Vb	(Van Es, 1992)
sp3	ITE: Shannon_spacing_Vpconst	(Ebrahimi et al., 1994)
sp4	ITE: Shannon_spacing_Vplin	(Ebrahimi et al., 1994)
sp5	ITE: Shannon_spacing_Vplin2	(Ebrahimi et al., 1994)
sp6	ITE: Shannon_spacing_VKDE	(Noughabi and Noughabi, 2013)
KDP	ITE: Shannon_KDP	(Stowell and Plumley, 2009)
PSD	ITE: Shannon_PSD_SzegöT	(Ramirez et al., 2009; Gray, 2006) (Grenander and Szegö, 1958)
EdE	ITE: Shannon_Edgeworth	(van Hulle, 2005)
Gau	based on (9)	
ME1	ITE: Shannon_MaxEnt1	(Hyvärinen, 1997)
ME2	ITE: Shannon_MaxEnt2	(Hyvärinen, 1997)

which behaves as  $\log x$  asymptotically for  $x \rightarrow \infty$ ). As this estimator would become  $-\infty$  if a value occurs more than once, we first remove duplicate values from the data before applying (23). There should be better ways of dealing with discretization effects, but we nevertheless include this particular estimator for comparison, as it was also used in previous implementations of the entropy-based IGCI method (Daniušis et al., 2010; Janzing et al., 2012).

We also made use of various entropy estimators implemented in the Information Theoretical Estimators (ITE) Toolbox, release 0.58 (Szabó, 2014). The method 3NN is based on  $k$ -nearest neighbors with  $k = 3$ , all sp\* methods use Vasicek’s spacing method with various corrections, KDP uses k-d partitioning, PSD uses the power spectral density representation and Szegö’s theorem, ME1 and ME2 use the maximum entropy distribution method, EdE uses the Edgeworth expansion. For more details, see the documentation of the ITE toolbox (Szabó, 2014).

#### 4.1.3 INDEPENDENCE TESTING: HSIC

As covariance function for HSIC, we use the popular Gaussian kernel:

$$\kappa_\ell : (x, x') \mapsto \exp\left(-\frac{(x - x')^2}{\ell^2}\right),$$

with bandwidths selected by the median heuristic (Schölkopf and Smola, 2002), i.e., we take

$$\hat{\ell}(\mathbf{u}) := \text{median}\{\|u_i - u_j\| : 1 \leq i < j \leq N, \|u_i - u_j\| \neq 0\},$$

and similarly for  $\hat{\ell}(\mathbf{v})$ . As the product of two Gaussian kernels is characteristic, HSIC with such kernels will pick up any dependence asymptotically (see also Lemma 8 in Appendix A).

The  $p$ -value can either be estimated by using permutation, or can be approximated by a Gamma approximation, as the mean and variance of the HSIC value under the null hypothesis can also be estimated in closed form (Gretton et al., 2008). In this work, we use the Gamma approximation for the HSIC  $p$ -value.

## 4.2 Data sets

We will use both real-world and simulated data in order to evaluate the methods. Here we give short descriptions and refer the reader to Appendix C and Appendix D for details.

### 4.2.1 REAL-WORLD BENCHMARK DATA

The `CauseEffectPairs` (CEP) benchmark data set that we propose in this work consists of different “cause-effect pairs”, each one consisting of samples of a pair of statistically dependent random variables, where one variable is known to cause the other one. It is an extension of the collection of the eight data sets that formed the “CauseEffectPairs” task in the *Causality Challenge #2: Pot-Luck* competition (Mooij and Janzing, 2010) which was performed as part of the NIPS 2008 Workshop on Causality (Guyon et al., 2010). Version 1.0 of the `CauseEffectPairs` collection that we present here consists of 88 pairs, taken from 31 different data sets from different domains. The CEP data is publicly available at (Mooij et al., 2014). Appendix D contains a detailed description of each cause-effect pair and a justification of what we believe to be the ground truths. Scatter plots of all pairs are shown in Figure 4.

### 4.2.2 SIMULATED DATA

As collecting real-world benchmark data is a tedious process (mostly because the ground truths are unknown, and acquiring the necessary understanding of the data-generating process in order to decide about the ground truth is not straightforward), we also studied the performance of methods on simulated data where we can control the data-generating process, and therefore can be certain about the ground truth.

Simulating data can be done in many ways. It is not straightforward to simulate data in a “realistic” way, e.g., in such a way that scatter plots of simulated data look similar to those of the real-world data (see Figure 4). For reproducibility, we describe in Appendix C in detail how the simulations were done. Here, we will just sketch the main ideas.

We sample data from the following structural equation models. If we do not want to model a confounder, we use:

$$\begin{aligned} E_X &\sim p_{E_X}, E_Y \sim p_{E_Y} \\ X &= f_X(E_X) \\ Y &= f_Y(X, E_Y), \end{aligned}$$

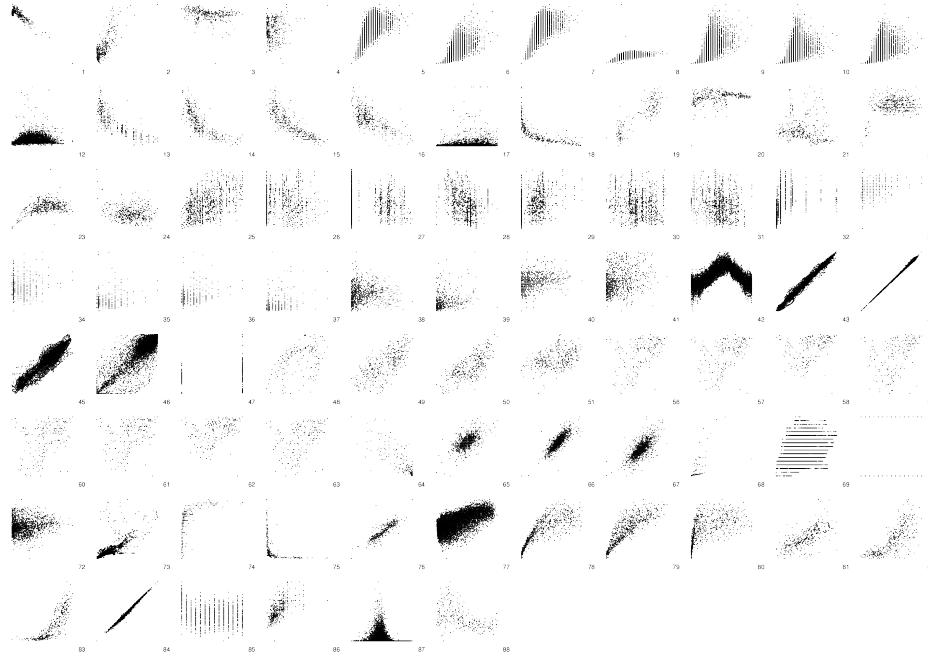


Figure 4: Scatter plots of the 88 cause-effect pairs in the `CauseEffectPairs` benchmark data.

and if we do want to include a confounder  $Z$ , we use:

$$\begin{aligned} E_X &\sim p_{E_X}, E_Y \sim p_{E_Y}, E_Z \sim p_{E_Z} \\ Z &= f_Z(E_Z) \\ X &= f_X(E_X, E_Z) \\ Y &= f_Y(X, E_Y, E_Z). \end{aligned}$$

Here, the noise distributions  $p_{E_X}, p_{E_Y}, p_{E_Z}$  are randomly generated distributions, and the causal mechanisms  $f_Z, f_X, f_Y$  are randomly generated functions. Sampling the random distributions for a noise variable  $E_X$  (and similarly for  $E_Y$  and  $E_Z$ ) is done by mapping a standard-normal distribution through a random function, which we sample from a Gaussian Process. The causal relationship  $f_X$  (and similarly  $f_Y$  and  $f_Z$ ) is drawn from a Gaussian Process as well. After sampling the noise distributions and the functional relationships, we generate data for  $X, Y, Z$ . Finally, Gaussian measurement noise is added to both  $X$  and  $Y$ .

By controlling various hyperparameters, we can control certain aspects of the data generation process. We considered four different scenarios. **SIM** is the default scenario without confounders. **SIM-c** does include a one-dimensional confounder, whose influence on  $X$  and  $Y$  are typically equally strong as the influence of  $X$  on  $Y$ . The setting **SIM-1n** has low noise levels, and we would expect IGCI to work well for this scenario. Finally, **SIM-G** has approximate Gaussian distributions for the cause  $X$  and approximately additive Gaussian noise (on top of a nonlinear relationship between cause and effect); we expect that methods

which make these Gaussianity assumptions will work well in this scenario. Scatter plots of the simulated data are shown in Figures 5–8.

### 4.3 Preprocessing and Perturbations

The following preprocessing was applied to each pair  $(X, Y)$ . Both variables  $X$  and  $Y$  were standardized (i.e., an affine transformation is applied on both variables such that their empirical mean becomes 0, and their empirical standard deviation becomes 1). In order to study the effect of discretization and other small perturbations of the data, one of these four perturbations was applied:

**unperturbed** : No perturbation is applied.

**discretized** : Discretize the variable that has the most unique values such that after discretization, it has as many unique values as the other variable. The discretization procedure repeatedly merges those values for which the sum of the absolute error that would be caused by the merge is minimized.

**undiscretized** : “Undiscretize” both variables  $X$  and  $Y$ . The undiscretization procedure adds noise to each data point  $z$ , drawn uniformly from the interval  $[0, z' - z]$ , where  $z'$  is the smallest value  $z' > z$  that occurs in the data.

**small noise** : Add tiny independent Gaussian noise to both  $X$  and  $Y$  (with mean 0 and standard deviation  $10^{-9}$ ).

Ideally, a causal discovery method should be robust against these and other small perturbations of the data.

### 4.4 Evaluation Measures

As a performance measure, we calculate the weighted accuracy of a method in the following way:

$$\text{accuracy} = \frac{\sum_{m=1}^M w_m \delta_{\hat{d}_m, d_m}}{\sum_{m=1}^M w_m}, \quad (24)$$

where  $d_m$  is the true causal direction for the  $m$ 'th pair (either “ $\leftarrow$ ” or “ $\rightarrow$ ”),  $\hat{d}_m$  is the estimated direction (one of “ $\leftarrow$ ”, “ $\rightarrow$ ”, and “?”), and  $w_m$  is the *weight* of the pair. Note that we are only awarding correct decisions, i.e., if no estimate is given ( $d_m = “?”$ ), this will negatively affect the accuracy. For the simulated pairs, each pair has weight  $w_m = 1$ . For the real-world cause-effect pairs, the weights are specified in Table 4. These weights have been chosen such that the weights of all pairs from the same data set sum to one. The reason is that we cannot consider pairs that come from the same data set as independent. For example, in the case of the **Abalone** data set, the variables “whole weight”, “shucked weight”, “viscera weight”, “shell weight” are strongly correlated. Considering the four pairs (age, whole weight), (age, shucked weight), etc., as independent could introduce a bias. We (conservatively) correct for that bias by downweighting the pairs taken from the **Abalone** data set.

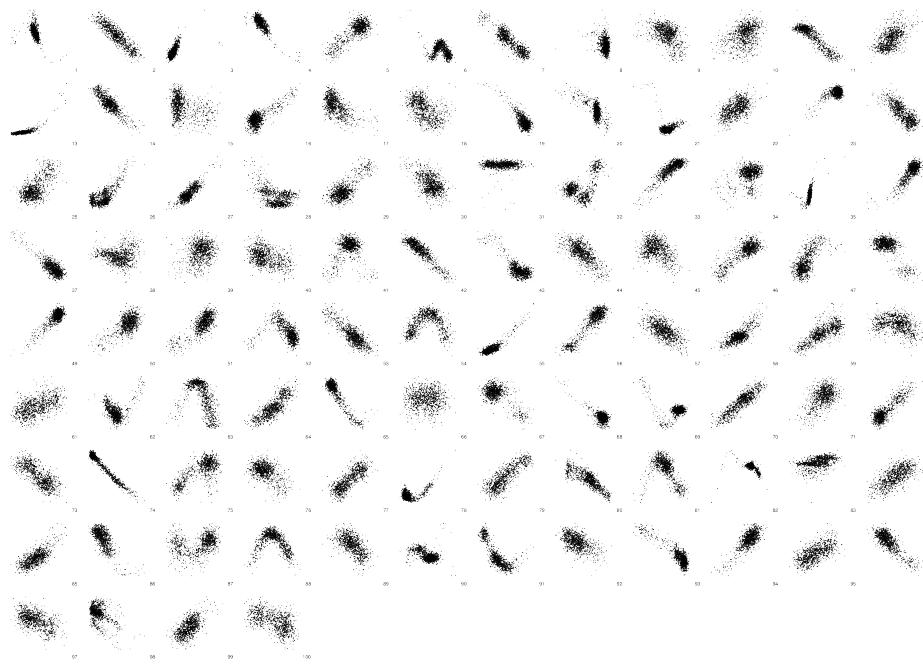


Figure 5: Scatter plots of the cause-effect pairs in simulation scenario **SIM**.

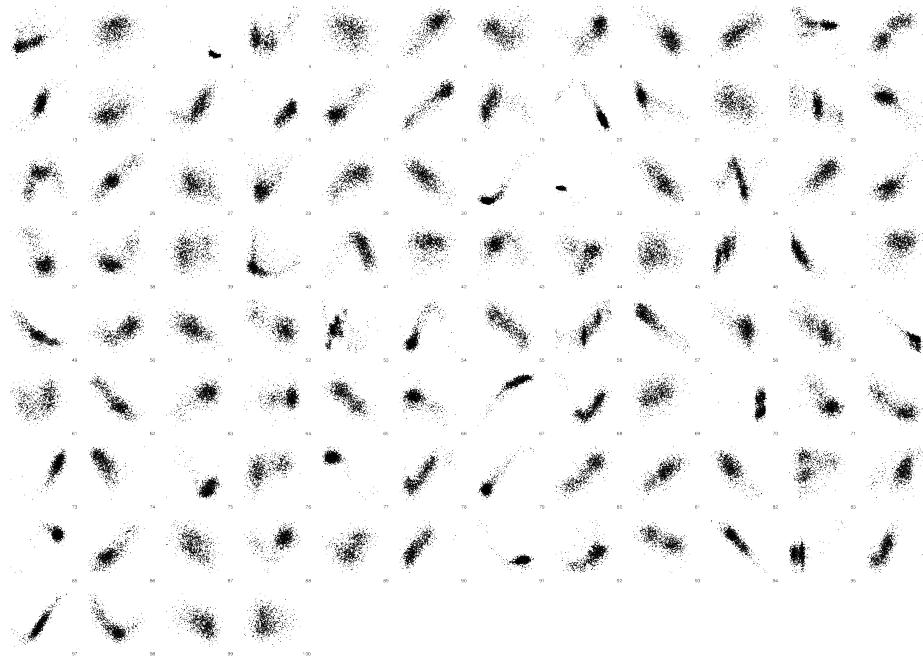


Figure 6: Scatter plots of the cause-effect pairs in simulation scenario **SIM-c**.



Figure 7: Scatter plots of the cause-effect pairs in simulation scenario SIM-1n.

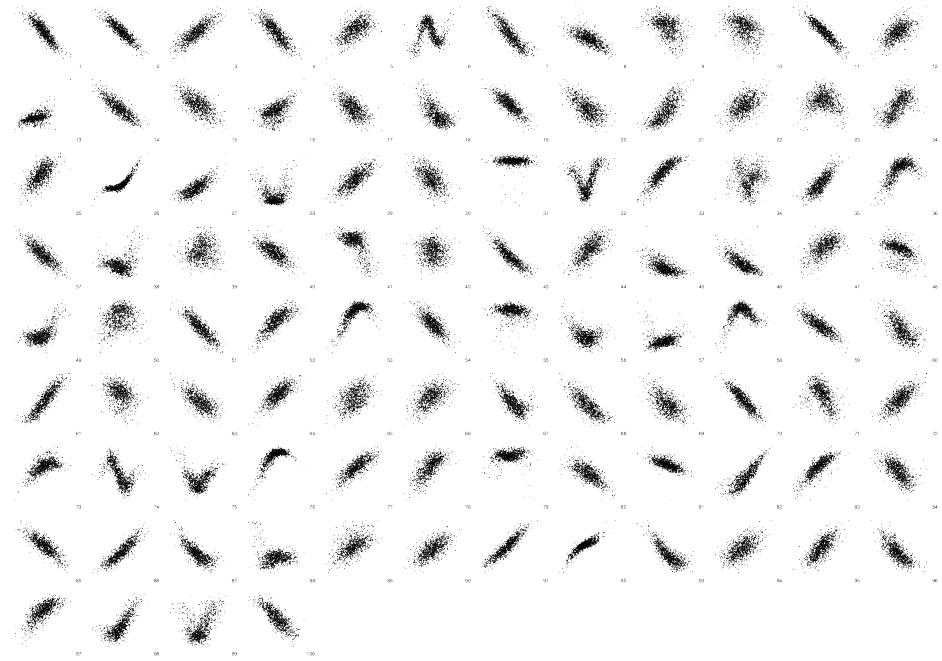


Figure 8: Scatter plots of the cause-effect pairs in simulation scenario SIM-G.

Table 2: The methods that are evaluated in this work.

Name	Algorithm	Score	Details
ANM-pHSIC	1	(5)	Data recycling, adaptive kernel bandwidth
ANM-HSIC	1	(6)	Data recycling, adaptive kernel bandwidth
ANM-HSIC-ds	1	(6)	Data splitting, adaptive kernel bandwidth
ANM-HSIC-fk	1	(6)	Data recycling, fixed kernel
ANM-HSIC-ds-fk	1	(6)	Data splitting, fixed kernel
ANM-ent-...	1	(8)	Data recycling, entropy estimators from Table 1
ANM-Gauss	1	(10)	Data recycling
ANM-FN	2	(11)	
ANM-MML	2	(12)	
IGCI-slope	3	(20)	
IGCI-slope++	3	(22)	
IGCI-ent-...	3	(21)	Entropy estimators from Table 1

In earlier work, we have reported accuracy-decision rate curves, in order to evaluate whether accuracy would increase when only the most confident decisions are taken. Here, we only evaluate the forced-choice scenario (i.e., a decision rate of 100%), because it is too easy to visually over-interpret the significance of such a curve.

## 5. Results

In this section, we report the results of the experiments that we carried out in order to evaluate the performance of various methods. We plot the accuracies as box plots, indicating the estimated (weighted) accuracy (24), the corresponding 68% confidence interval, and the 95% confidence interval, assuming a binomial distribution using the method by Clopper and Pearson (1934).<sup>7</sup> If there were pairs for which no decision was taken, the number of nondecisions is indicated on the corresponding accuracy boxplot. The methods that we evaluated are listed in Table 2.

### 5.1 Additive Noise Models

We start by reporting the results for additive noise methods. Figure 9 shows the accuracies of all ANM methods on different unperturbed data sets, including the CEP benchmark and various simulated data sets. Figure 10 shows the accuracies of the same methods on different perturbations of the CEP benchmark data.

#### 5.1.1 HSIC-BASED SCORES

The ANM methods that use HSIC perform well on all data sets, obtaining accuracies between 63% and 83%. Note that the simulated data (and also the real-world data) deviate in three ways from the assumptions made by the additive noise method: (i) the noise is not

---

7. In earlier work, we ranked decisions according to their confidence, and plotted accuracy versus decision rate. Here we only consider the “forced-choice” scenario, because the accuracy-decision rate curves are easy to overinterpret visually in terms of their significance.

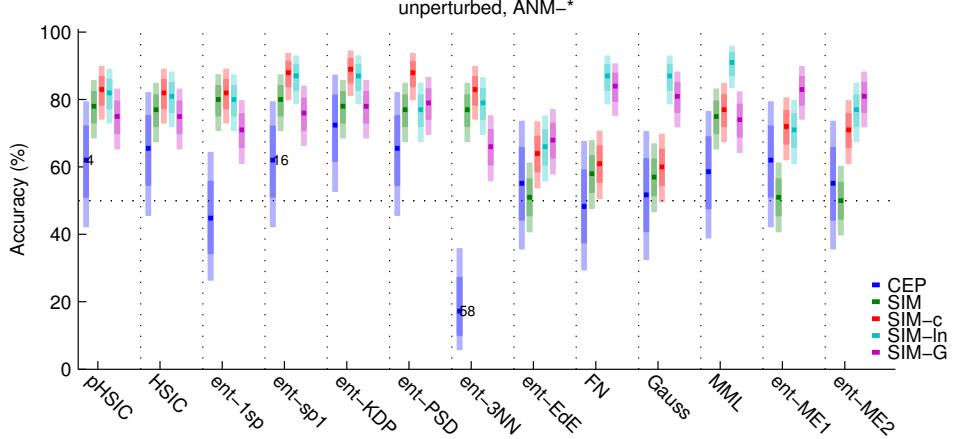


Figure 9: Accuracies of various ANM methods on different (unperturbed) data sets. For the variants of the spacing estimator, only the results for `sp1` are shown, as results for `sp2`, ..., `sp6` were similar.

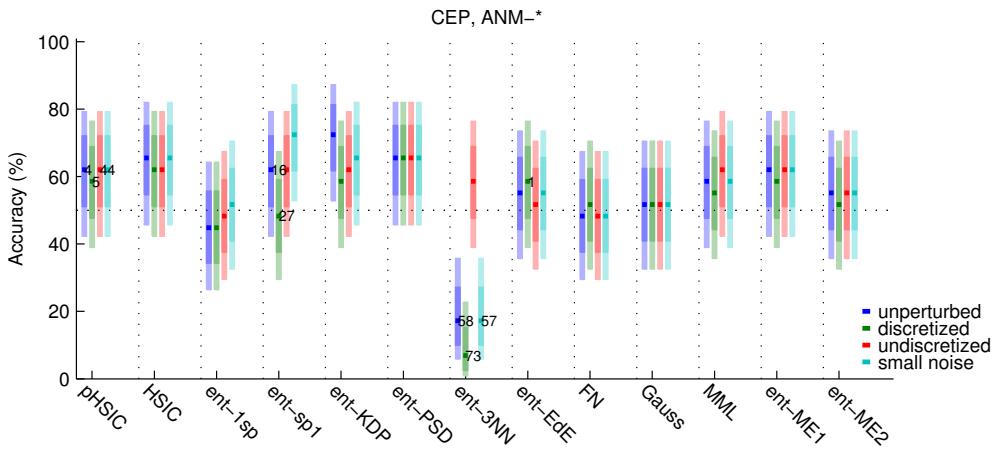


Figure 10: Accuracies of various ANM methods on different perturbations of the CEP benchmark data. For the variants of the spacing estimator, only the results for `sp1` are shown, as results for `sp2`, ..., `sp6` were similar.

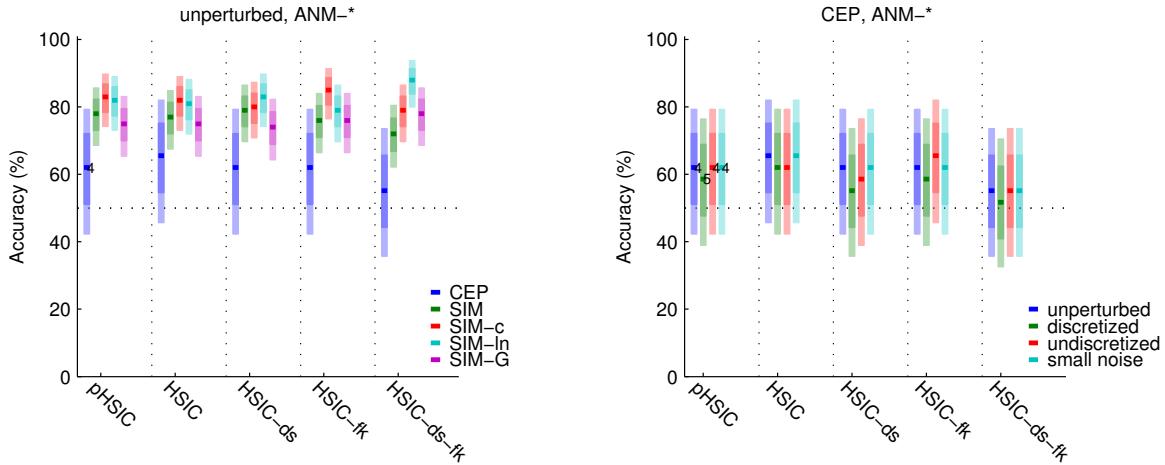


Figure 11: Left: Accuracies of several variants of HSIC-based ANM methods on different (unperturbed) data sets. Right: Accuracies of several variants of HSIC-based ANM methods on different perturbations of the CEP benchmark data.

additive, (ii) a confounder can be present, and (iii) additional “measurement noise” was added to both cause and effect. Moreover, the results turn out to be robust against small perturbations of the data. This shows that the additive noise method can perform well, even in case of model misspecification.

ANM-pHSIC, the method originally used by Hoyer et al. (2009), has an issue with large sample sizes: in those cases, the Gamma-approximated HSIC  $p$ -value can become so small that it underflows (i.e., can no longer be represented as a floating point number). This decreases the number of decisions due to cases where the HSIC  $p$ -value underflows in both directions. ANM-HSIC, which uses the HSIC value itself, does not suffer from this problem. Note that the results of ANM-pHSIC and ANM-HSIC are almost identical, and both are very robust against perturbations of the data, although discretization seems to lower the accuracy slightly.

Figure 11 shows how the performance of the HSIC-based additive noise methods depends on other implementation details: ANM-HSIC-ds uses data splitting, ANM-HSIC-fk uses fixed kernels for HSIC (with bandwidth 0.5), and ANM-HSIC-ds-fk combines both features. Generally, the differences are small, although ANM-HSIC-ds-fk seems to perform a little differently than the other ones. ANM-HSIC-ds-fk is shown to be consistent in Appendix A. If standard GP regression satisfies the property in (31), then ANM-HSIC-fk is also consistent.

### 5.1.2 ENTROPY-BASED SCORES

For the entropy-based score (8), we see in Figure 9 and Figure 10 that the results depend strongly on which entropy estimator is used. The six variants  $\text{sp}1, \dots, \text{sp}6$  of the spacing estimators perform very similarly, so we showed only the results for ANM-ent-sp1.

All (nonparametric) entropy estimators (1sp, 3NN, sp*i*, KDP, PSD) perform well on simulated data, with the exception of ANM-ent-EdE. On the CEP benchmark on the other hand, the performance varies greatly over estimators. This has to do with discretization effects.

Indeed, the differential entropy of a variable that can take only a finite number of values is  $-\infty$ . The way in which differential entropy estimators treat values that occur multiple times differs, and this can have a large influence on the estimated entropy. For example, **ANM-ent-1sp** simply ignores values that occur more than once, which leads to a performance that is around chance level. **ANM-ent-3NN** returns  $-\infty$  (for both  $C_{X \rightarrow Y}$  and  $C_{Y \rightarrow X}$ ) in the majority of the pairs in the CEP benchmark, and **ANM-ent-spi** also return  $-\infty$  in quite a few cases. Also, additional discretization of the data decreases accuracy of these methods as it increases the number of pairs for which no decision can be taken. The only (non-parametric) entropy-based ANM methods that perform well on both the CEP benchmark data and the simulated data are **ANM-ent-KDP** and **ANM-ent-PSD**. Of these two methods, **ANM-ent-PSD** seems more robust under perturbations than **ANM-ent-KDP**, and can compete with the HSIC-based methods.

### 5.1.3 OTHER SCORES

Consider now the results for the “parametric” entropy estimators (**ANM-Gauss**, **ANM-ent-ME1**, **ANM-ent-ME2**), the Bayesian method **ANM-FN**, and the MML method **ANM-MML**.

First, note that **ANM-Gauss** and **ANM-FN** perform very similarly. This means that the difference between these two scores (i.e., the complexity measure of the regression function) does not outweigh the common part (the likelihood) of these two scores. Both these scores do not perform better than chance on the CEP data, probably because the Gaussianity assumption is typically violated on real data. They do obtain high accuracies for the **SIM-1n** and **SIM-G** scenarios. For **SIM-G** this is to be expected, as the assumption that the cause has a Gaussian distribution is satisfied in that scenario. For **SIM-1n** it is not evident why these scores perform so well—it could be that the noise is close to additive and Gaussian in that scenario.

The related score **ANM-MML**, which employs a more sophisticated complexity measure for the distribution of the cause, performs better on the CEP data and the two simulation settings **SIM** and **SIM-c**. This suggests that the Gaussian complexity measure used by **ANM-FN** and **ANM-Gauss** is too simple in practice. Note further that **ANM-ent-MML** performs worse in the **SIM-G** scenario, which is probably due to a higher variance of the MML complexity measure compared with the simple Gaussian entropy measure. The performance of **ANM-MML** on the CEP benchmark is robust against small perturbations of the data.

The parametric entropy estimators **ANM-ent-ME1**, **ANM-ent-ME2** do not perform well on the CEP and **SIM** data, although they do obtain good accuracies on the other simulated data sets. The reasons for this behaviour are not understood; we speculate that the parametric assumptions made by these estimators match the actual distribution of the data in these particular simulation settings quite well.

## 5.2 Information Geometric Causal Inference

Here we report the results of the evaluation of different IGCI variants. Figure 12 shows the accuracies of all the IGCI variants on different (unperturbed) data sets, including the CEP benchmark and different simulation settings. Figure 13 shows the accuracies of the IGCI methods on different perturbations of the CEP benchmark. In both figures, we distinguish two base measures: the uniform and the Gaussian base measure.

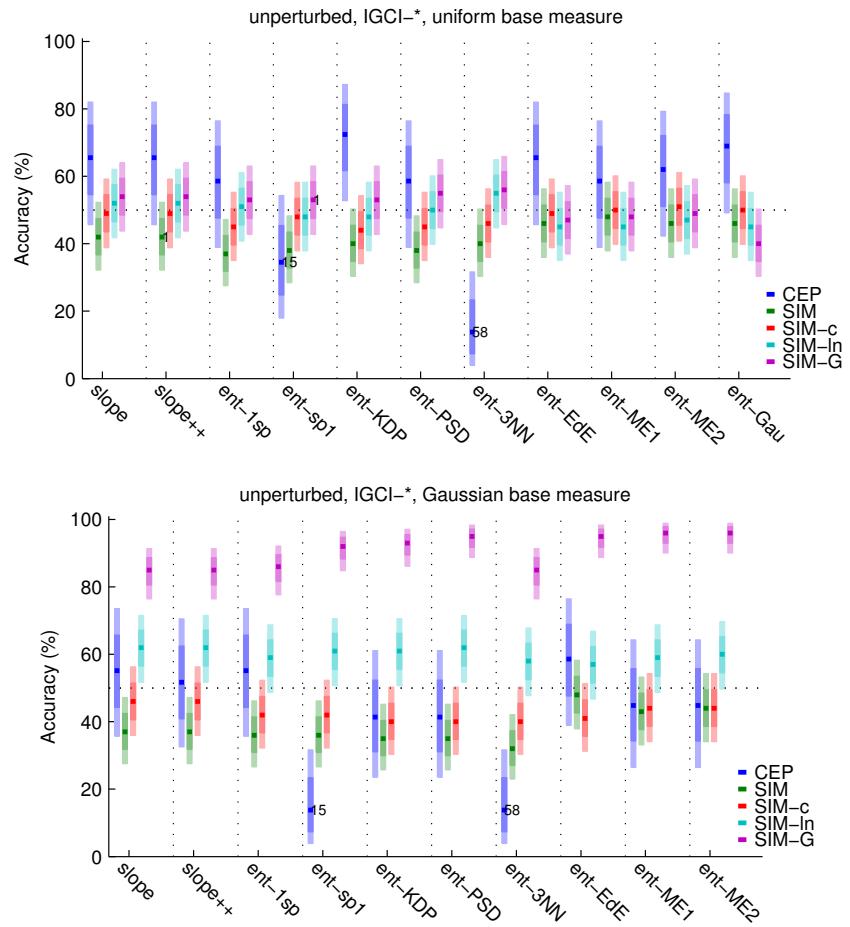


Figure 12: Accuracies for various IGCI methods on different (unperturbed) data sets. Top: uniform base measure. Bottom: Gaussian base measure.

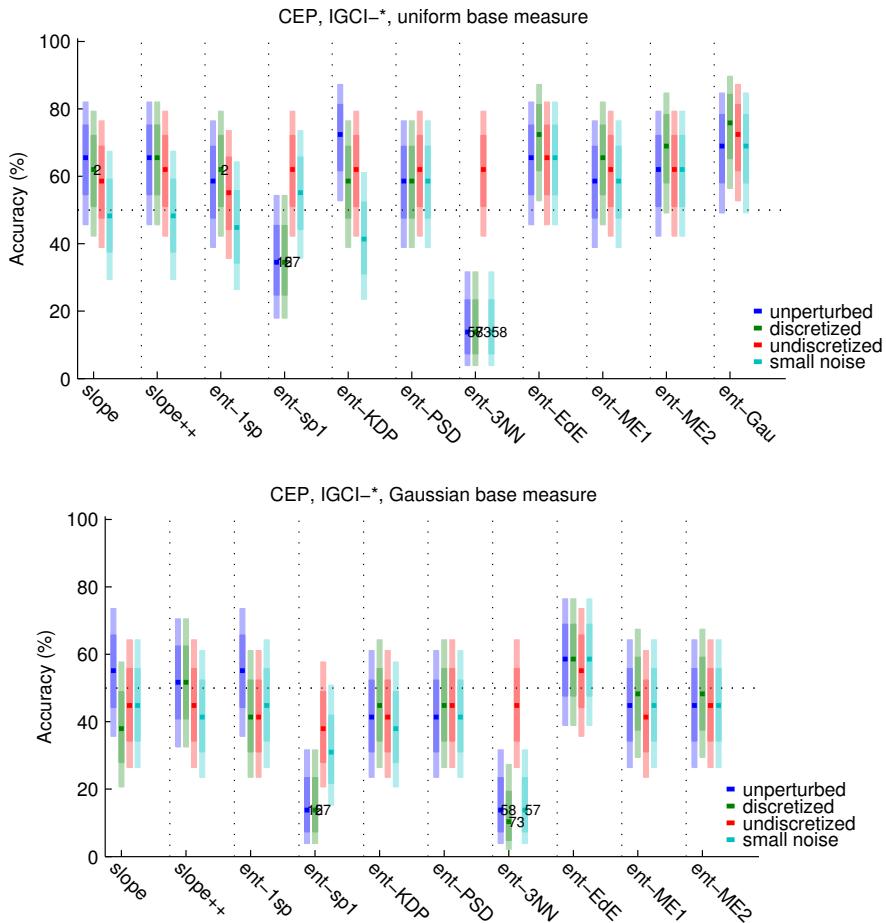


Figure 13: Accuracies for various IGCI methods on different perturbations of the CEP benchmark data. Top: uniform base measure. Bottom: Gaussian base measure.

Note that none of the IGCI methods performs well on our simulated data when using the uniform base measure, as all these accuracies are around chance level. A very different picture emerges when using the Gaussian base measure: here the performance is generally lowest on the **SIM** data (even lower than chance level), somewhat higher for **SIM-c**, somewhat above chance level for **SIM-1n**, and between 80% and 100% for **SIM-G**.

This requires some explanations. Indeed, as IGCI was designed for the bijective deterministic case, one would expect that IGCI would work better on **SIM-1n** (without depending too strongly on the reference measure), because in that scenario the noise is relatively small. This, however, does not turn out to be the case. To understand this phenomenon, we inspect the scatter plots in Figure 7 and observe that the functions in **SIM-1n** are either non-injective or relatively close to being linear. Both can spoil the performance despite having low noise.

For the more noisy settings, earlier experiments showed that **IGCI-slope** and **IGCI-1sp** can perform surprisingly well on simulated data (Janzing et al., 2012). Here, however, we see that the performance of all IGCI variants on noisy data seems to depend strongly on characteristics of the data generation process, and on the implementation details of the IGCI estimator. Often IGCI seems to pick up certain features in the data that turn out to be correlated with the causal direction in some settings, but can be anticorrelated with the causal direction in other settings. In addition, these results suggest that if the distribution of the cause is close to the base measure used in IGCI, then also for noisy data the method may work well (as happens in the **SIM-G** setting). However, for causal relations that are not sufficiently non-linear, performance can drop significantly (even below chance level) in case of a discrepancy between the actual distribution of the cause and the base measure assumed by IGCI.

### 5.2.1 ORIGINAL IMPLEMENTATIONS

Let us now take a closer look at the accuracies of the original methods **IGCI-slope** and **IGCI-ent-1sp** that were proposed in (Daniūsis et al., 2010; Janzing et al., 2012). For both uniform and Gaussian base measures, the performance seems a little better than chance level on the **CEP** benchmark, but not as much as in previous evaluations on earlier versions of the benchmark. The discrepancy with the accuracies of around 80% reported by Janzing et al. (2012) can be explained by the fact that here we evaluate on more cause-effect pairs, and that we chose the weights more appropriately.

It is also interesting to look at the behavior under perturbations of the **CEP** data. When using the uniform base measure, the performance of both **IGCI-slope** and **IGCI-ent-1sp** drops back to chance level if small noise is added. For the Gaussian base measure, the accuracies that are slightly better than chance on the unperturbed **CEP** data become worse than random guessing on the perturbed versions. This observation motivated the introduction of the improved slope-based estimator **IGCI-slope++** that uses (22) instead of (20). This estimator seems to be a bit more robust to additional discretization, but its accuracy is still no better than chance level when adding a small amount of noise.

### 5.2.2 NONPARAMETRIC ENTROPY ESTIMATORS

On the CEP benchmark, the accuracies of the entropy-based IGCI methods depend strongly on the specific entropy estimator. This can again be ascribed to discretization effects. For example, the closely related estimators `IGCI-ent-1sp` and `IGCI-ent-sp1` perform comparably on simulated data, but on the real CEP data, the `IGCI-ent-sp1` estimator obtains a lower score because of nondecisions due to discretization effects. Generally, the behavior seems similar to what we saw for the ANM methods: `3NN` and the spacing estimators  $sp_i$  suffer most from discretization, whereas for `KDP` and `PSD` the discretization does not lead to any nondecisions, but it might still affect accuracy negatively. Indeed, the performance of `KDP` on the CEP data decreases when discretizing more, and is generally not robust to perturbations. `PSD` on the other hand seems to be robust to perturbations. However, while for the uniform base measure the performance of `IGCI-ent-PSD` is above chance level on the CEP data, for the Gaussian base measure it is below chance level. Interestingly, the `EdE` estimator that performed poorly for ANM seems to be the best one on the CEP benchmark (both for the uniform and the Gaussian base measure).

### 5.2.3 PARAMETRIC ENTROPY ESTIMATORS

Let us now consider the accuracies of entropy-based IGCI that use parametric entropy estimators, which make additional assumptions on the distribution. Interestingly, the `Gau` estimator that assumes a Gaussian distribution turns out to work quite well on CEP data (when using the uniform base measure, obviously). This means that the ratio of the size of the support of the distribution and its variance is already quite informative on the causal direction for the CEP data. This might also explain the relatively good performance on this benchmark of `IGCI-ent-ME1` and `IGCI-ent-ME2` when using the uniform base measure.

## 6. Discussion and Conclusion

In this work, we have proposed the `CauseEffectPairs` benchmark data set consisting of 88 real-world cause-effect pairs with known ground truth. We have used this benchmark data in combination with several simulated data sets in order to evaluate various bivariate causal discovery methods based on the Additive Noise Model (ANM) and on Information Geometric Causal Inference (IGCI). Our main conclusions are twofold:

1. The ANM variants `ANM-pHSIC`, `ANM-HSIC`, `ANM-PSD` and `ANM-MML` perform better than chance on both real-world and simulated data, obtaining accuracies between 65% and 80%. Their performance is robust against small perturbations of the data (including discretization).
2. The performance of IGCI-based methods varies greatly depending on implementation details, perturbations of the data and certain characteristics of the data, in ways that we do not fully understand. In many cases, causal relations seem to be too linear for IGCI to work well.

The former conclusion is in line with earlier reports, but the latter conclusion is surprising, considering that good performance of `IGCI-slope` and `IGCI-ent-1sp` has been reported

several times in earlier work (Danušis et al., 2010; Mooij et al., 2010; Janzing et al., 2012; Statnikov et al., 2012).

Whether the performance of a method is significantly better than random guessing is not so clear-cut. For example, the  $p$ -value for ANM-pHSIC under the null hypothesis of random guessing is 0.068 on the CEP benchmark, only slightly larger than the popular  $\alpha = 5\%$  threshold. However, because we tested many different methods, one should correct for multiple testing. Using a conservative Bonferroni correction, the results on a single data set would be far from being significant. In other words, from this study we cannot conclude that the ANM-pHSIC method significantly outperforms random guessing on the CauseEffectPairs benchmark. However, the Bonferroni correction would be overly conservative for our purposes, as many of the methods that we compare are small variations of each other, and their results are clearly dependent. In addition, good performance across data sets increases the significance of the results. Although a proper quantitative evaluation of the significance of our results is a nontrivial exercise, we believe that when combining the results on the CEP benchmark with those on the simulated data, the hypothesis that the good performance of the methods ANM-pHSIC, ANM-HSIC, ANM-PSD and ANM-MML is only due to chance is implausible.

When dealing with real-world data on a computer, variables of a continuous nature are almost always discretized because they have to be represented as floating point numbers. Often, additional rounding is applied, for example because only the most significant digits are recorded. We found that for many methods, especially for those that use differential entropy estimators, coarse discretization of the data causes problems. This suggests that performance of several methods can still be improved, e.g., by using entropy estimators that are more robust to such discretization effects. The HSIC independence measure (and its  $p$ -value) and the PSD entropy estimator were found to be robust against small perturbations of the data, including discretization.

The original ANM method (ANM-pHSIC) proposed by Hoyer et al. (2009) turns out to be one of the best methods. This observation motivated the consistency proof, one of the more theoretical contributions of this work. We expect that extending this consistency result to the multivariate case (see also Peters et al., 2014) should be straightforward.

Concluding, our results provide evidence that distinguishing cause from effect is indeed possible by exploiting certain statistical patterns in the observational data. However, the performance of current state-of-the-art bivariate causal methods still has to be improved further in order to enable practical applications.

## Appendix A. Consistency Proof of ANM-HSIC

In this Appendix, we prove the consistency of Algorithm 1 with score (7), which is closely related to the algorithm originally proposed in Hoyer et al. (2009). The main difference is that the original implementation uses the HSIC  $p$ -value, whereas here, we use the HSIC value itself as a score. Also, we consider the option of splitting the dataset into one part for regression and another part for independence testing. Finally, we fix the HSIC kernel instead of letting its bandwidth be chosen by a heuristic that depends on the data. The reason that we make these small modifications is that they lead to an easier proof of consistency of the method.

We start with recapitulating the definition and basic properties of the Hilbert Schmidt Independence Criterion (HSIC) in Section A.1. Then, we discuss asymptotic properties of non-parametric regression methods in Section A.2. Finally, we combine these ingredients in Section A.3.

### A.1 Consistency of HSIC

We recapitulate the definitions and some asymptotic properties of the Hilbert Schmidt Independence Criterion (HSIC), following mostly the notations and terminology in Gretton et al. (2005). The HSIC estimator that we use here is the original biased estimator proposed in Gretton et al. (2005).

**Definition 7** *Given two random variables  $X \in \mathcal{X}$  and  $Y \in \mathcal{Y}$  with joint distribution  $\mathbb{P}_{X,Y}$ , and bounded kernels  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$  and  $l : \mathcal{Y}^2 \rightarrow \mathbb{R}$ , we define the **population HSIC** of  $X$  and  $Y$  as:*

$$\text{HSIC}_{k,l}(X, Y) := \mathbb{E}_{X,X',Y,Y'}(k(X, X')l(Y, Y')) + \mathbb{E}_{X,X'}(k(X, X'))\mathbb{E}_{Y,Y'}(l(Y, Y')) \\ - 2\mathbb{E}_{X,Y}\left(\mathbb{E}_{X'}(k(X, X'))\mathbb{E}_{Y'}(l(Y, Y'))\right).$$

Here,  $(X, Y)$  and  $(X', Y')$  are two independent random variables distributed according to  $\mathbb{P}_{X,Y}$ .

When  $k$  and  $l$  are clear from the context, we will typically suppress the dependence of the population HSIC on the choice of the kernels  $k$  and  $l$ , simply writing  $\text{HSIC}(X, Y)$  instead. The justification for the name “independence criterion” stems from the following important result (Fukumizu et al., 2008, Theorem 3):

**Lemma 8** *Whenever the product kernel  $k \cdot l$  is characteristic (in the sense of Fukumizu et al. (2008); Sriperumbudur et al. (2010)):  $\text{HSIC}_{k,l}(X, Y) = 0$  if and only if  $X \perp\!\!\!\perp Y$  (i.e.,  $X$  and  $Y$  are independent).  $\blacksquare$*

A special case of this lemma, assuming that  $X$  and  $Y$  have compact domain, was proven originally in Gretton et al. (2005). Intuitively, a characteristic kernel leads to an injective embedding of probability measures into the corresponding Reproducible Kernel Hilbert Space (RKHS). The HSIC is the squared RKHS distance between the embedded joint distribution and the embedded product of the marginals. Given that the embedding is injective, this distance is zero if and only if the variables are independent. Examples of characteristic kernels are Gaussian RBF kernels and Laplace kernels. For more details

on the notion of characteristic kernel, see Sriperumbudur et al. (2010). We will use the following (biased) estimator of the population HSIC Gretton et al. (2005):

**Definition 9** Given two  $N$ -tuples (with  $N \geq 2$ )  $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{X}^N$  and  $\mathbf{y} = (y_1, \dots, y_N) \in \mathcal{Y}^N$ , and bounded kernels  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$  and  $l : \mathcal{Y}^2 \rightarrow \mathbb{R}$ , we define

$$\widehat{\text{HSIC}}_{k,l}(\mathbf{x}, \mathbf{y}) := (N-1)^{-2} \text{tr}(KHLH) = (N-1)^{-2} \sum_{i,j=1}^N \bar{K}_{ij} L_{ij}, \quad (25)$$

where  $K_{ij} = k(x_i, x_j)$ ,  $L_{ij} = l(y_i, y_j)$  are Gram matrices and  $H_{ij} = \delta_{ij} - N^{-1}$  is a centering matrix, and we write  $\bar{K} := KKH$  for the centered Gram matrix  $K$ . Given an i.i.d. sample  $\mathcal{D}_N = \{(x_n, y_n)\}_{n=1}^N$  from  $\mathbb{P}_{X,Y}$ , we define the **empirical HSIC** of  $X$  and  $Y$  estimated from  $\mathcal{D}_N$  as:

$$\widehat{\text{HSIC}}_{k,l}(X, Y; \mathcal{D}_N) := \widehat{\text{HSIC}}_{k,l}(\mathbf{x}, \mathbf{y}).$$

Again, when  $k$  and  $l$  are clear from the context, we will typically suppress the dependence of the empirical HSIC on the choice of the kernels  $k$  and  $l$ . Unbiased estimators of the population HSIC were proposed in later work (Song et al., 2012), but we will not consider those here. A large deviation result for this empirical HSIC estimator is given by (Gretton et al., 2005, Theorem 3):

**Lemma 10** Assume that  $k$  and  $l$  are bounded almost everywhere by 1, and are non-negative. Suppose that the data set  $\mathcal{D}_N$  consists of  $N$  i.i.d. samples from some joint probability distribution  $\mathbb{P}_{X,Y}$ . Then, for  $N \geq 2$  and all  $\delta > 0$ , with probability at least  $1 - \delta$ :

$$|\text{HSIC}_{k,l}(X, Y) - \widehat{\text{HSIC}}_{k,l}(X, Y; \mathcal{D}_N)| \leq \sqrt{\frac{\log(6/\delta)}{\alpha^2 N}} + \frac{c}{N},$$

where  $\alpha^2 > 0.24$  and  $c$  are constants. ■

This directly implies the consistency of the empirical HSIC estimator:<sup>8</sup>

**Corollary 11** Let  $(x_1, y_1), (x_2, y_2), \dots$  be i.i.d. according to  $\mathbb{P}_{X,Y}$ . Defining  $\mathcal{D}_N = \{(x_n, y_n)\}_{n=1}^N$  for  $N = 2, 3, \dots$ , we have for non-negative bounded kernels  $k, l$  that, for  $N \rightarrow \infty$ :

$$\widehat{\text{HSIC}}_{k,l}(X, Y; \mathcal{D}_N) \xrightarrow{P} \text{HSIC}_{k,l}(X, Y). \quad \blacksquare$$

For the special case that  $\mathcal{Y} = \mathbb{R}$ , we will use the following continuity property of the empirical HSIC estimator. It shows that for a Lipschitz-continuous kernel  $l$ , the empirical HSIC is also Lipschitz-continuous in the corresponding argument, but with a Lipschitz constant that scales at least as  $N^{-1/2}$  for  $N \rightarrow \infty$ . This novel technical result will be the key to our consistency proof of Algorithm 1 with score (7).

8. Let  $X_1, X_2, \dots$  be a sequence of random variables and let  $X$  be another random variable. We say that  $X_n$  converges to  $X$  in **probability**, written  $X_n \xrightarrow{P} X$ , if

$$\forall \epsilon > 0 : \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

**Lemma 12** For all  $N \geq 2$ , for all  $\mathbf{x} \in \mathcal{X}^N$ , for all  $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^N$ , for all bounded kernels  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$ , and for all bounded and Lipschitz-continuous kernels  $l : \mathbb{R}^2 \rightarrow \mathbb{R}$ :

$$\left| \widehat{\text{HSIC}}(\mathbf{x}, \mathbf{y}) - \widehat{\text{HSIC}}(\mathbf{x}, \mathbf{y}') \right| \leq \frac{32\lambda C}{\sqrt{N}} \|\mathbf{y} - \mathbf{y}'\|,$$

where  $|k(\xi, \xi')| \leq C$  for all  $\xi, \xi' \in \mathcal{X}$  and  $\lambda$  is the Lipschitz constant of  $l$ .

**Proof** From its definition, (25):

$$\left| \widehat{\text{HSIC}}(\mathbf{x}, \mathbf{y}) - \widehat{\text{HSIC}}(\mathbf{x}, \mathbf{y}') \right| = (N-1)^{-2} \left| \sum_{i,j=1}^N \bar{K}_{ij}(L_{ij} - L'_{ij}) \right|,$$

where  $K_{ij} = k(x_i, x_j)$ ,  $L_{ij} = l(y_i, y_j)$ ,  $L'_{ij} = l(y'_i, y'_j)$  and  $\bar{K} := HKH$  with  $H_{ij} = \delta_{ij} - N^{-1}$ . First, note that  $|K_{ij}| \leq C$  implies that  $|\bar{K}_{ij}| \leq 4C$ :

$$\begin{aligned} |\bar{K}_{ij}| &= \left| K_{ij} - \frac{1}{N} \sum_{i'=1}^N K_{i'j} - \frac{1}{N} \sum_{j'=1}^N K_{ij'} + \frac{1}{N^2} \sum_{i',j'=1}^N K_{i'j'} \right| \\ &\leq |K_{ij}| + \frac{1}{N} \sum_{i'=1}^N |K_{i'j}| + \frac{1}{N} \sum_{j'=1}^N |K_{ij'}| + \frac{1}{N^2} \sum_{i',j'=1}^N |K_{i'j'}| \leq 4C. \end{aligned}$$

Now starting from the definition and using the triangle inequality:

$$\begin{aligned} \left| \sum_{i,j=1}^N \bar{K}_{ij}(L_{ij} - L'_{ij}) \right| &\leq \left| \sum_{i,j=1}^N \bar{K}_{ij}(l(y'_i, y'_j) - l(y'_i, y_j)) \right| \\ &\quad + \left| \sum_{i,j=1}^N \bar{K}_{ij}(l(y_i, y_j) - l(y_i, y_j)) \right|. \end{aligned}$$

For the first term, using Cauchy-Schwartz (in  $\mathbb{R}^{N^2}$ ) and the Lipschitz property of  $l$ :

$$\begin{aligned} \left| \sum_{i,j=1}^N \bar{K}_{ij}(l(y'_i, y'_j) - l(y'_i, y_j)) \right|^2 &\leq \left( \sum_{i,j=1}^N |\bar{K}_{ij}|^2 \right) \left( \sum_{i,j=1}^N |l(y'_i, y'_j) - l(y'_i, y_j)|^2 \right) \\ &\leq 16N^2 C^2 \cdot \lambda^2 N \sum_{j=1}^N |y'_j - y_j|^2 \\ &= 16N^3 C^2 \lambda^2 \|\mathbf{y}' - \mathbf{y}\|^2. \end{aligned}$$

The second term is similar. The result now follows (using that  $\frac{N}{N-1} \leq 2$  for  $N \geq 2$ ).  $\blacksquare$

## A.2 Consistency of nonparametric regression

From now on, we will assume that both  $X$  and  $Y$  take values in  $\mathbb{R}$ . Györfi et al. (2002) provide consistency results for several nonparametric regression methods. Here we briefly discuss the main property (“weak universal consistency”) that is of particular interest in our setting.

Given a distribution  $\mathbb{P}_{X,Y}$ , one defines the **regression function** of  $Y$  on  $X$  as the conditional expectation

$$f(x) := \mathbb{E}(Y | X = x).$$

Given an i.i.d. sample of data points  $\mathcal{D}_N = \{(x_n, y_n)\}_{n=1}^N$  (the “training” data), a regression method provides an estimate of the regression function  $\hat{f}(\cdot; \mathcal{D}_N)$ . The **mean squared error on the training data** (also called “training error”) is defined as:

$$\frac{1}{N} \sum_{n=1}^N \left| f(x_n) - \hat{f}(x_n; \mathcal{D}_N) \right|^2.$$

The **risk** (also called “generalization error”), i.e., the expected  $L_2$  error on an independent test datum, is defined as:

$$\mathbb{E}_X \left| f(X) - \hat{f}(X; \mathcal{D}_N) \right|^2 = \int \left| f(x) - \hat{f}(x; \mathcal{D}_N) \right|^2 d\mathbb{P}_X(x)$$

Note that the risk is a random variable that depends on the training data  $\mathcal{D}_N$ .

If the expected risk converges to zero as the number of training points increases, the regression method is called “weakly consistent”. More precisely, following Györfi et al. (2002):

**Definition 13** A sequence of estimated regression functions  $\hat{f}(\cdot; \mathcal{D}_N)$  is called **weakly consistent for a certain distribution**  $\mathbb{P}_{X,Y}$  if<sup>9</sup>

$$\lim_{N \rightarrow \infty} \mathbb{E}_{\mathcal{D}_N} \left( \mathbb{E}_X \left| f(X) - \hat{f}(X; \mathcal{D}_N) \right|^2 \right) = 0. \quad (26)$$

A regression method is called **weakly universally consistent** if it is weakly consistent for all distributions  $\mathbb{P}_{X,Y}$  with finite second moment of  $Y$ , i.e., with  $\mathbb{E}_Y(Y^2) < \infty$ .

Many popular nonparametric regression methods have been shown to be weakly universally consistent, see e.g., Györfi et al. (2002). One might expect naïvely that if the expected risk goes to zero, then also the expected training error should vanish asymptotically:

$$\lim_{N \rightarrow \infty} \mathbb{E}_{\mathcal{D}_N} \frac{1}{N} \sum_{n=1}^N \left| f(x_n) - \hat{f}(x_n; \mathcal{D}_N) \right|^2 = 0. \quad (27)$$

However, property (27) does not necessarily follow from (26). One would expect that asymptotic results on the training error would actually be easier to obtain than results

---

9. Here,  $\mathbb{E}_{\mathcal{D}_N}$  denotes the expectation value when averaging over data sets  $\mathcal{D}_N$  consisting of  $N$  pairs  $(x_i, y_i)$  that are i.i.d. distributed according to  $\mathbb{P}_{X,Y}$ .

on generalization error, but this question seems to be less well studied in the existing literature. The only result that we know of is (Kpotufe et al., 2014, Lemma 5), which states that a certain box kernel regression method satisfies (27) under certain assumptions on the distribution  $\mathbb{P}_{X,Y}$ . The reason that we bring this up at this point is that property (27) allows to prove consistency even when one uses the same data for both regression and independence testing (see also Lemma 15).

From now on, we will always consider the following setting. Let  $(\xi_1, \eta_1), (\xi_2, \eta_2), \dots$  be i.i.d. according to some joint distribution  $\mathbb{P}_{X,Y}$ . We distinguish two different scenarios:

- “Data splitting”: using half of the data for training, and the other half of the data for testing. In particular, we define  $x_n := \xi_{2n-1}$ ,  $y_n := \eta_{2n-1}$ ,  $x'_n := \xi_{2n}$  and  $y'_n := \eta_{2n}$  for  $n = 1, 2, \dots$ .
- “Data recycling”: using the same data both for regression and for testing. In particular, we define  $x_n := \xi_n$ ,  $y_n := \eta_n$ ,  $x'_n := \xi_n$  and  $y'_n := \eta_n$  for  $n = 1, 2, \dots$ .

In both scenarios, for  $N = 1, 2, \dots$ , we define a training data set  $\mathcal{D}_N := \{(x_n, y_n)\}_{n=1}^N$  (for the regression) and a test data set  $\mathcal{D}'_N := \{(x'_n, y'_n)\}_{n=1}^N$  (for testing independence of residuals). Note that in the data recycling scenario, training and test data are identical, whereas in the data splitting scenario, training and test data are independent.

Define a random variable (the “residual”)

$$E := Y - f(X) = Y - \mathbb{E}(Y | X), \quad (28)$$

and its vector-valued version on the test data:

$$\mathbf{e}'_N := (y'_1 - f(x'_1), \dots, y'_N - f(x'_N)), \quad (29)$$

called the **true residuals**. Using a regression model learned from the training data  $\mathcal{D}_N$ , we obtain an estimate  $\hat{f}(x; \mathcal{D}_N)$  for the regression function  $f(x) = \mathbb{E}(Y | X = x)$ . We then define an estimate of the vector-valued version of  $E$  on the test data:

$$\hat{\mathbf{e}}'_N := (y'_1 - \hat{f}(x'_1; \mathcal{D}_N), \dots, y'_N - \hat{f}(x'_N; \mathcal{D}_N)), \quad (30)$$

called the **predicted residuals**.

**Definition 14** *We call the regression method **suitable** for regressing  $Y$  on  $X$  if the mean squared error between true and predicted residuals vanishes asymptotically in expectation:*

$$\lim_{N \rightarrow \infty} \mathbb{E}_{\mathcal{D}_N, \mathcal{D}'_N} \left( \frac{1}{N} \|\hat{\mathbf{e}}'_N - \mathbf{e}'_N\|^2 \right) = 0. \quad (31)$$

Here, the expectation is taken over both training data  $\mathcal{D}_N$  and test data  $\mathcal{D}'_N$ .

**Lemma 15** *In the data splitting case, any regression method that is weakly consistent for  $\mathbb{P}_{X,Y}$  is suitable. In the data recycling case, any regression method satisfying property (27) is suitable.*

**Proof** Simply rewriting:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathbb{E} \left( \frac{1}{N} \|\hat{\mathbf{e}}'_N - \mathbf{e}'_N\|^2 \right) &= \\ \lim_{N \rightarrow \infty} \mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N \left| (y'_n - \hat{f}(x'_n; \mathcal{D}_N)) - (y'_n - f(x'_n)) \right|^2 \right) &= \\ \lim_{N \rightarrow \infty} \mathbb{E}_{\mathcal{D}_N, \mathcal{D}'_N} \left( \frac{1}{N} \sum_{n=1}^N \left| \hat{f}(x'_n; \mathcal{D}_N) - f(x'_n) \right|^2 \right) \end{aligned}$$

Therefore, (31) reduces to (26) in the data splitting scenario (where each  $x'_n$  is an independent copy of  $X$ ), and reduces to (27) in the data recycling scenario (where  $x'_n = x_n$ ).  $\blacksquare$

In particular, if  $\mathbb{E}(X^2) < \infty$  and  $\mathbb{E}(Y^2) < \infty$ , any weakly universally consistent regression method is suitable both for regressing  $X$  on  $Y$  and  $Y$  on  $X$  in the data splitting scenario.

### A.3 Consistency of ANM-HSIC

We can now prove our main result, stating that the empirical HSIC calculated from the test set inputs and the predicted residuals on the test set (using the regression function estimated from the training set) converges in probability to the population HSIC of the true inputs and the true residuals:

**Theorem 16** *Let  $X, Y \in \mathbb{R}$  be two random variables with joint distribution  $\mathbb{P}_{X,Y}$ . Let  $k, l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be two bounded non-negative kernels and assume that  $l$  is Lipschitz continuous. Suppose we are given sequences of training data sets  $\mathcal{D}_N$  and test data sets  $\mathcal{D}'_N$  (in either the data splitting or the data recycling scenario described above). Suppose we use a suitable regression procedure (c.f. Lemma 15), to obtain a sequence  $\hat{f}(x; \mathcal{D}_N)$  of estimates of the regression function  $\mathbb{E}(Y | X = x)$  from the training data. Defining the true residual  $E$  by (28), and the predicted residuals  $\hat{\mathbf{e}}'_N$  on the test data as in (30), then, for  $N \rightarrow \infty$ :*

$$\widehat{\text{HSIC}}(\mathbf{x}', \hat{\mathbf{e}}'_N) \xrightarrow{P} \text{HSIC}(X, E).$$

**Proof** We start by applying Lemma 12:

$$\left| \widehat{\text{HSIC}}(\mathbf{x}', \hat{\mathbf{e}}'_N) - \widehat{\text{HSIC}}(\mathbf{x}', \mathbf{e}'_N) \right|^2 \leq \left( \frac{32\lambda C}{\sqrt{N}} \right)^2 \|\hat{\mathbf{e}}'_N - \mathbf{e}'_N\|^2$$

where  $\lambda$  and  $C$  are constants. From the suitability of the regression method, (31), it therefore follows that

$$\lim_{N \rightarrow \infty} \mathbb{E}_{\mathcal{D}_N, \mathcal{D}'_N} \left| \widehat{\text{HSIC}}(\mathbf{x}', \hat{\mathbf{e}}'_N) - \widehat{\text{HSIC}}(\mathbf{x}', \mathbf{e}'_N) \right|^2 = 0,$$

i.e.,

$$\widehat{\text{HSIC}}(\mathbf{x}', \hat{\mathbf{e}}'_N) - \widehat{\text{HSIC}}(\mathbf{x}', \mathbf{e}'_N) \xrightarrow{L_2} 0.$$

As convergence in  $L_2$  implies convergence in probability (see, e.g., Wasserman (2004)),

$$\widehat{\text{HSIC}}(\mathbf{x}', \hat{\mathbf{e}}'_N) - \widehat{\text{HSIC}}(\mathbf{x}', \mathbf{e}'_N) \xrightarrow{P} 0.$$

From the consistency of the empirical HSIC, Corollary 11:

$$\widehat{\text{HSIC}}(\mathbf{x}', \mathbf{e}'_N) \xrightarrow{P} \text{HSIC}(X, E).$$

Hence, by taking sums (see e.g., (Wasserman, 2004, Theorem 5.5)), we arrive at the desired statement.  $\blacksquare$

We are now ready to show that Algorithm 1 with score (7) is consistent.

**Corollary 17** *Let  $X, Y$  be two real-valued random variables with joint distribution  $\mathbb{P}_{X,Y}$  that either satisfies an additive noise model  $X \rightarrow Y$ , or  $Y \rightarrow X$ , but not both. Suppose we are given sequences of training data sets  $\mathcal{D}_N$  and test data sets  $\mathcal{D}'_N$  (in either the data splitting or the data recycling scenario). Let  $k, l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be two bounded non-negative Lipschitz-continuous characteristic kernels. If the regression procedure used in Algorithm 1 is suitable for both  $\mathbb{P}_{X,Y}$  and  $\mathbb{P}_{Y,X}$ , then Algorithm 1 with score (7) is a consistent procedure for estimating the direction of the additive noise model.*

**Proof** Define ‘‘population residuals’’  $E_Y := Y - \mathbb{E}(Y | X)$  and  $E_X := X - \mathbb{E}(X | Y)$ . Note that  $\mathbb{P}_{X,Y}$  satisfies a bivariate additive noise model  $X \rightarrow Y$  if and only if  $E_Y \perp\!\!\!\perp X$  (c.f. Lemma 4). Further, by Lemma 8, we have  $\text{HSIC}_{k,l}(X, E_Y) = 0$  if and only if  $X \perp\!\!\!\perp E_Y$ . Similarly,  $\mathbb{P}_{X,Y}$  satisfies a bivariate additive noise model  $Y \rightarrow X$  if and only if  $\text{HSIC}_{l,k}(Y, E_X) = 0$ .

Now, by Theorem 16,

$$C_{X \rightarrow Y} := \widehat{\text{HSIC}}_{k,l}(x', \hat{e}_Y(\mathcal{D}'_N; \mathcal{D}_N)) \xrightarrow{P} \text{HSIC}_{k,l}(X, E_Y),$$

and similarly

$$C_{Y \rightarrow X} := \widehat{\text{HSIC}}_{l,k}(y', \hat{e}_X(\mathcal{D}'_N; \mathcal{D}_N)) \xrightarrow{P} \text{HSIC}_{l,k}(Y, E_X).$$

Because  $\mathbb{P}_{X,Y}$  satisfies an additive noise model only in one of the two directions, this implies that either  $\text{HSIC}_{k,l}(X, E_Y) = 0$  and  $\text{HSIC}_{l,k}(Y, E_X) > 0$  (corresponding with  $X \rightarrow Y$ ), or  $\text{HSIC}_{k,l}(X, E_Y) > 0$  and  $\text{HSIC}_{l,k}(Y, E_X) = 0$  (corresponding with  $Y \rightarrow X$ ). Therefore the test procedure is consistent.  $\blacksquare$

## Appendix B. Relationship between scores (10) and (11)

For the special case of an additive noise model  $X \rightarrow Y$ , the Bayesian score proposed in Friedman and Nachman (2000) was given in (11):

$$C_{X \rightarrow Y} = \min_{\mu, \tau^2, \boldsymbol{\theta}, \sigma^2} \left( -\log \mathcal{N}(\mathbf{x} | \mu \mathbf{1}, \tau^2 \mathbf{I}) - \log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}}(\mathbf{x}) + \sigma^2 \mathbf{I}) \right).$$

It is a sum of the negative log likelihood of a Gaussian model for the inputs:

$$\begin{aligned}
 & \min_{\mu, \tau^2} (-\log \mathcal{N}(\mathbf{x} | \mu \mathbf{1}, \tau^2 \mathbf{I})) \\
 &= \min_{\mu, \tau^2} \left( \frac{N}{2} \log(2\pi\tau^2) + \frac{1}{2\tau^2} \sum_{i=1}^N (x_i - \mu)^2 \right) \\
 &= \frac{N}{2} \log(2\pi e) + \frac{N}{2} \log \left( \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \right)
 \end{aligned} \tag{32}$$

with  $\bar{x} := \frac{1}{N} \sum_{i=1}^N x_i$ , and the negative log marginal likelihood of a GP model for the outputs, given the inputs:

$$\begin{aligned}
 & \min_{\theta, \sigma^2} (-\log \mathcal{N}(\mathbf{y} | 0, \mathbf{K}_\theta(\mathbf{x}) + \sigma^2 \mathbf{I})) \\
 &= \min_{\theta, \sigma^2} \left( \frac{N}{2} \log(2\pi) + \frac{1}{2} \log |\det(\mathbf{K}_\theta(\mathbf{x}) + \sigma^2 \mathbf{I})| + \mathbf{y}^T (\mathbf{K}_\theta(\mathbf{x}) + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \right).
 \end{aligned} \tag{33}$$

Note that (32) is an empirical estimator of the entropy of a Gaussian with variance  $\text{Var}(X)$ , up to a factor  $N$ :

$$H(X) = \frac{1}{2} \log(2\pi e) + \frac{1}{2} \log \text{Var}(X).$$

We will show that (33) is closely related to an empirical estimator of the entropy of the residuals  $Y - \mathbb{E}(Y | X)$ :

$$H(Y - \mathbb{E}(Y | X)) = \frac{1}{2} \log(2\pi e) + \frac{1}{2} \log \text{Var}(Y - \mathbb{E}(Y | X)).$$

This means that the score (11) considered by Friedman and Nachman (2000) is closely related to the Gaussian score (10) for  $X \rightarrow Y$ :

$$C_{X \rightarrow Y} = \log \text{Var}(X) + \log \text{Var}(Y - \hat{f}_Y(X)).$$

The following Lemma shows that standard Gaussian Process regression can be interpreted as a penalized maximum likelihood optimization.

**Lemma 18** *Let  $\mathbf{K}_\theta(\mathbf{x})$  be the kernel matrix (abbreviated as  $\mathbf{K}$ ) and define the negative penalized log-likelihood as:*

$$-\log \mathcal{L}(\mathbf{f}, \sigma^2; \mathbf{y}, \mathbf{K}) := \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - f_i)^2 + \frac{1}{2} \mathbf{f}^T \mathbf{K}^{-1} \mathbf{f} + \frac{1}{2} \log |\det(\mathbf{I} + \sigma^{-2} \mathbf{K})|. \tag{34}$$

Minimizing with respect to  $\mathbf{f}$  yields a minimum at:

$$\hat{\mathbf{f}}_{\sigma, \theta} = \underset{\mathbf{f}}{\operatorname{argmin}} (-\log \mathcal{L}(\mathbf{f}, \sigma^2; \mathbf{y}, \mathbf{K}_\theta)) = \mathbf{K}_\theta (\mathbf{K}_\theta + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \tag{35}$$

and the value at the minimum is given by:

$$\min_{\mathbf{f}} (-\log \mathcal{L}(\mathbf{f}, \sigma^2; \mathbf{y}, \mathbf{K}_\theta)) = -\log \mathcal{L}(\hat{\mathbf{f}}_{\sigma, \theta}, \sigma^2; \mathbf{y}, \mathbf{K}_\theta) = -\log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_\theta + \sigma^2 \mathbf{I}). \tag{36}$$

**Proof** Note that

$$\begin{aligned} & \frac{1}{2} \log |\det \mathbf{K}| - \frac{1}{2} \log |\det (\mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K})| \\ &= \frac{1}{2} \log |\det (\mathbf{I} - (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K})^{-1}| \\ &= \frac{1}{2} \log |\det (\mathbf{I} + \sigma^{-2} \mathbf{K})|, \end{aligned} \quad (37)$$

where we used the Woodbury formula in the last step. Using identities (A.7) and (A.9) in (Rasmussen and Williams, 2006), we obtain:

$$\begin{aligned} & \mathcal{N}(\mathbf{f} | \mathbf{0}, \mathbf{K}) \mathcal{N}(\mathbf{y} - \mathbf{f} | \mathbf{0}, \sigma^2 \mathbf{I}) \\ &= \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}) \mathcal{N}(\mathbf{f} | \hat{\mathbf{f}}, \mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}), \end{aligned} \quad (38)$$

where

$$\hat{\mathbf{f}}_{\sigma, \theta} := \mathbf{K}_\theta (\mathbf{K}_\theta + \sigma^2 \mathbf{I})^{-1} \mathbf{y}.$$

Combining formulas (34), (37) and (38), we can derive the identity

$$\mathcal{L}(\mathbf{f}, \sigma^2; \mathbf{y}, \mathbf{K}) = \mathcal{N}(\mathbf{f} | \mathbf{0}, \mathbf{K}) \mathcal{N}(\mathbf{y} - \mathbf{f} | \mathbf{0}, \sigma^2 \mathbf{I}) |\det (\mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K})|^{1/2} (2\pi)^{N/2}$$

It is now obvious that the penalized likelihood is maximized for  $\mathbf{f} = \hat{\mathbf{f}}_{\sigma, \theta}$  (for fixed hyper-parameters  $\sigma, \theta$ ) and that at the maximum, it has the value

$$\mathcal{L}(\hat{\mathbf{f}}_{\sigma, \theta}, \sigma^2; \mathbf{y}, \mathbf{K}_\theta) = \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_\theta + \sigma^2 \mathbf{I}).$$

■

Note that the estimated function (35) is identical to the mean posterior GP, and the value (36) is identical to the negative logarithm of the marginal likelihood (evidence) of the data according to the GP model.

Making use of Lemma 18, the conditional part (33) in score (11) can be rewritten as:

$$\begin{aligned} & \min_{\sigma^2, \theta} (-\log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_\theta + \sigma^2 \mathbf{I})) \\ &= \min_{\sigma^2, \theta} \left( \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - (\hat{\mathbf{f}}_{\sigma, \theta})_i)^2 + \frac{1}{2} \hat{\mathbf{f}}_{\sigma, \theta}^T \mathbf{K}_\theta^{-1} \hat{\mathbf{f}}_{\sigma, \theta} + \frac{1}{2} \log |\det (\mathbf{I} + \sigma^{-2} \mathbf{K}_\theta)| \right) \\ &= \underbrace{\frac{N}{2} \log(2\pi\hat{\sigma}^2) + \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^N (y_i - (\hat{\mathbf{f}})^i)^2}_{\text{Likelihood term}} + \underbrace{\frac{1}{2} \hat{\mathbf{f}}^T \mathbf{K}_\theta^{-1} \hat{\mathbf{f}} + \frac{1}{2} \log |\det (\mathbf{I} + \hat{\sigma}^{-2} \mathbf{K}_\theta)|}_{\text{Complexity penalty}}, \end{aligned}$$

where  $\hat{f} := \hat{f}_{\hat{\sigma}, \hat{\theta}}$  for the minimizing  $(\hat{\sigma}, \hat{\theta})$ . If the complexity penalty is small compared to the likelihood term around the optimal values  $(\hat{\sigma}, \hat{\theta})$ , we can approximate:

$$\begin{aligned} & \min_{\sigma^2, \theta} (-\log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_\theta + \sigma^2 \mathbf{I})) \\ & \approx \frac{N}{2} \log(2\pi\hat{\sigma}^2) + \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^N (y_i - \hat{f}_i)^2 \\ & \approx \min_{\sigma^2} \left( \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \hat{f}_i)^2 \right) \\ & = \frac{N}{2} \log(2\pi e) + \frac{N}{2} \log \left( \frac{1}{N} \sum_{i=1}^N (y_i - \hat{f}_i)^2 \right). \end{aligned}$$

This shows that there is a close relationship between the two scores (11) and (10).

## Appendix C. Details on the simulated data

### C.1 Sampling from a random density

We first describe how we sample from a random density. First, we sample  $\mathbf{X} \in \mathbb{R}^N$  from a standard-normal distribution:

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{I}_N)$$

and define  $\vec{\mathbf{X}}$  to be the vector that is obtained by sorting  $\mathbf{X}$  in ascending order. Then, we sample a realization  $\mathbf{F}$  of a Gaussian Process with inputs  $\vec{\mathbf{X}}$ , using a kernel with hyperparameters  $\theta$  and white noise with standard deviation  $\sigma$ :

$$\mathbf{F} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_\theta(\vec{\mathbf{X}}) + \sigma^2 \mathbf{I}),$$

where  $\mathbf{K}_\theta(\vec{\mathbf{X}})$  is the Gram matrix for  $\vec{\mathbf{X}}$  using kernel  $k_\theta$ . We use the trapezoidal rule to calculate the cumulative integral of the function  $e^F : \mathbb{R} \rightarrow \mathbb{R}$  that linearly interpolates the points  $(\vec{\mathbf{X}}, \exp(\mathbf{F}))$ . In this way, we obtain a vector  $\mathbf{G} \in \mathbb{R}^N$  where each element  $G_i$  corresponds with  $\int_{\vec{\mathbf{X}}_1}^{\vec{\mathbf{X}}_i} e^F(x) dx$ . As covariance function, we used the Gaussian kernel:

$$k_\theta(\mathbf{x}, \mathbf{x}') = \exp \left( - \sum_{i=1}^D \frac{(x_i - x'_i)^2}{\theta_i^2} \right). \quad (39)$$

We will denote this whole sampling procedure by:

$$\mathbf{G} \sim \mathcal{RD}(\theta, \sigma).$$

### C.2 Sampling cause-effect pairs

We simulate cause-effect pairs as follows. First, we sample three noise variables:

$$\begin{aligned} W_{E_X} &\sim \Gamma(a_{W_{E_X}}, b_{W_{E_X}}) & \mathbf{E}_X &\sim \mathcal{RD}(W_{E_X}, \tau) \\ W_{E_Y} &\sim \Gamma(a_{W_{E_Y}}, b_{W_{E_Y}}) & \mathbf{E}_Y &\sim \mathcal{RD}(W_{E_X}, \tau) \\ W_{E_Z} &\sim \Gamma(a_{W_{E_Z}}, b_{W_{E_Z}}) & \mathbf{E}_Z &\sim \mathcal{RD}(W_{E_Z}, \tau) \end{aligned}$$

where each noise variable has a random characteristic length scale. We then standardize each noise sample  $\mathbf{E}_X$ ,  $\mathbf{E}_Y$  and  $\mathbf{E}_Z$ .

If there is no confounder, we sample  $\mathbf{X}$  from a GP with inputs  $\mathbf{E}_X$ :

$$\begin{aligned} S_{E_X} &\sim \Gamma(a_{S_{E_X}}, b_{S_{E_X}}) \\ \mathbf{X} &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{S_{E_X}}(\mathbf{E}_X) + \tau^2 \mathbf{I}) \end{aligned}$$

and then we standardize  $\mathbf{X}$ . Then, we sample  $\mathbf{Y}$  from a GP with inputs  $(\mathbf{X}, \mathbf{E}_Y) \in \mathbb{R}^{N \times 2}$ :

$$\begin{aligned} S_X &\sim \Gamma(a_{S_X}, b_{S_X}) \\ S_{E_Y} &\sim \Gamma(a_{S_{E_Y}}, b_{S_{E_Y}}) \\ \mathbf{Y} &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{(S_X, S_{E_Y})}((\mathbf{X}, \mathbf{E}_Y)) + \tau^2 \mathbf{I}) \end{aligned}$$

and then we standardize  $\mathbf{Y}$ .

If there is a confounder, we sample  $\mathbf{X}$  from a GP with inputs  $(\mathbf{E}_X, \mathbf{E}_Z) \in \mathbb{R}^{N \times 2}$ :

$$\begin{aligned} S_{E_X} &\sim \Gamma(a_{S_{E_X}}, b_{S_{E_X}}) \\ S_{E_Z} &\sim \Gamma(a_{S_{E_Z}}, b_{S_{E_Z}}) \\ \mathbf{X} &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{(S_{E_X}, S_{E_Z})}((\mathbf{E}_X, \mathbf{E}_Z)) + \tau^2 \mathbf{I}) \end{aligned}$$

and then we standardize  $\mathbf{X}$ . Then, we sample  $\mathbf{Y}$  from a GP with inputs  $(\mathbf{X}, \mathbf{E}_Y, \mathbf{E}_Z) \in \mathbb{R}^{N \times 3}$ :

$$\begin{aligned} S_X &\sim \Gamma(a_{S_X}, b_{S_X}) \\ S_{E_Y} &\sim \Gamma(a_{S_{E_Y}}, b_{S_{E_Y}}) \\ S_{E_Z} &\sim \Gamma(a_{S_{E_Z}}, b_{S_{E_Z}}) \\ \mathbf{Y} &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{(S_X, S_{E_Y}, S_{E_Z})}((\mathbf{X}, \mathbf{E}_Y, \mathbf{E}_Z)) + \tau^2 \mathbf{I}) \end{aligned}$$

and then we standardize  $\mathbf{Y}$ .

Finally, we add measurement noise:

$$\begin{aligned} S_{M_X} &\sim \Gamma(a_{S_{M_X}}, b_{S_{M_X}}) \\ \mathbf{M}_X &\sim \mathcal{N}(\mathbf{0}, S_{M_X}^2 \mathbf{I}) \\ \mathbf{X} &\leftarrow \mathbf{X} + \mathbf{M}_X \\ S_{M_Y} &\sim \Gamma(a_{S_{M_Y}}, b_{S_{M_Y}}) \\ \mathbf{M}_Y &\sim \mathcal{N}(\mathbf{0}, S_{M_Y}^2 \mathbf{I}) \\ \mathbf{Y} &\leftarrow \mathbf{Y} + \mathbf{M}_Y \end{aligned}$$

We considered the four scenarios in Table 3: **SIM**, a scenario without confounders; **SIM-c**, a similar scenario but with one confounder; **SIM-1n**, a scenario with low noise levels (for which we expect IGCI to perform well); **SIM-G**, a scenario with a distribution of  $X$  that is almost Gaussian. We used  $N = 1000$  samples for each pair, and simulated 100 cause-effect pairs for each scenario.

Table 3: Parameter settings used to simulate cause-effect pairs for four scenarios. The common parameters for the four scenarios are:  $\tau = 10^{-4}$ ,  $(a_{WnE_Y}, b_{W_EY}) = (5, 0.1)$ ,  $(a_{W_EZ}, b_{W_EZ}) = (5, 0.1)$ ,  $(a_{S_EZ}, b_{S_EZ}) = (2, 15)$ ,  $(a_{S_X}, b_{S_X}) = (2, 15)$ .

name	# confounders	$(a_{W_{EX}}, b_{W_{EX}})$	$(a_{S_{EX}}, b_{S_{EX}})$	$(a_{S_{EY}}, b_{S_{EY}})$	$(a_{S_{MX}}, b_{S_{MX}})$	$(a_{S_{MY}}, b_{S_{MY}})$
SIM	0	(5, 0.1)	(2, 1.5)	(2, 15)	(2, 0.1)	(2, 0.1)
SIM-c	1	(5, 0.1)	(2, 1.5)	(2, 15)	(2, 0.1)	(2, 0.1)
SIM-ln	0	(5, 0.1)	(2, 1.5)	(2, 300)	(2, 0.01)	(2, 0.01)
SIM-G	0	$(10^6, 10^{-3})$	$(10^6, 10^{-3})$	(2, 15)	(2, 0.1)	(2, 0.1)

## Appendix D. Description of the CauseEffectPairs benchmark set

The `CauseEffectPairs` benchmark set described here is an extension of the collection of the eight data sets that formed the `CauseEffectPairs` task in the *Causality Challenge #2: Pot-Luck* competition (Mooij and Janzing, 2010) that was performed as part of the NIPS 2008 Workshop on Causality (Guyon et al., 2010). Here we describe version 1.0 of the `CauseEffectPairs` benchmark, which consists of 88 “cause-effect pairs”, each one consisting of samples of a pair of statistically dependent random variables, where one variable is known to cause the other one. The task is to identify for each pair which of the two variables is the cause and which one the effect, using the observed samples only. The data are publicly available at (Mooij et al., 2014).

The data sets were selected such that we expect common agreement on the ground truth. For example, the first pair consists of measurements of altitude and mean annual temperature of more than 300 weather stations in Germany. It should be obvious that altitude causes temperature rather than the other way around. Even though part of the statistical dependences may also be due to hidden common causes and selection bias, we expect that there is a significant cause-effect relation between the two variables in each pair, based on our understanding of the data generating process.

The best way to decide upon the ground truth of the causal relationships in the systems that generated the data would be by performing interventions on one of the variables and observing whether the intervention changes the distribution of the other variable. Unfortunately, these interventions cannot be performed in practice for many of the existing pairs because the original data-generating system is no longer available, or because of other practical reasons. Therefore, we have selected data sets in which the causal direction should be clear from the meanings of the variables and the way in which the data were generated. Unfortunately, for many data sets that are publicly available, it is not always clearly documented exactly how the variables are defined and measured.

In selecting the cause-effect pair data sets, we applied the following criteria:

- The minimum number of samples per pair should be a few hundred;
- The variables should have values in  $\mathbb{R}^d$  for some  $d = 1, 2, 3, \dots$ ;
- There should be a significant cause–effect relationship between the two variables;
- The direction of the causal relationship should be known or obvious from the meaning of the variables.

Version 1.0 of the `CauseEffectPairs` collection consists of 88 pairs satisfying these criteria, taken from 31 different data sets from different domains. We refer to these pairs as `pair0001`, ..., `pair0088`. Table 4 gives an overview of the cause-effect pairs. In the following subsections, we describe the cause-effect pairs in detail, and motivate our decisions on the causal relationships present in the pairs. We provide a scatter plot for each pair, where the horizontal axis corresponds with the cause, and the vertical axis with the effect. For completeness, we describe all the pairs in the data set, including those that have been described before in (Mooij and Janzing, 2010).

Table 4: Overview of the pairs in version 1.0 of the `CauseEffectPairs` benchmark.

Pair	Variable 1	Variable 2	Dataset	Ground Truth	Weight
pair0001	Altitude	Temperature	D1	→	1/6
pair0002	Altitude	Precipitation	D1	→	1/6
pair0003	Longitude	Temperature	D1	→	1/6
pair0004	Altitude	Sunshine hours	D1	→	1/6
pair0005	Age	Length	D2	→	1/7
pair0006	Age	Shell weight	D2	→	1/7
pair0007	Age	Diameter	D2	→	1/7
pair0008	Age	Height	D2	→	1/7
pair0009	Age	Whole weight	D2	→	1/7
pair0010	Age	Shucked weight	D2	→	1/7
pair0011	Age	Viscera weight	D2	→	1/7
pair0012	Age	Wage per hour	D3	→	1/2
pair0013	Displacement	Fuel consumption	D4	→	1/4
pair0014	Horse power	Fuel consumption	D4	→	1/4
pair0015	Weight	Fuel consumption	D4	→	1/4
pair0016	Horsepower	Acceleration	D4	→	1/4
pair0017	Age	Dividends from stocks	D3	→	1/2
pair0018	Age	Concentration GAG	D5	→	1
pair0019	Current duration	Ntext interval	D6	→	1
pair0020	Latitude	Temperature	D1	→	1/6
pair0021	Longitude	Precipitation	D1	→	1/6
pair0022	Age	Height	D7	→	1/3
pair0023	Age	Weight	D7	→	1/3
pair0024	Age	Heart rate	D7	→	1/3
pair0025	Cement	Compressive strength	D8	→	1/8
pair0026	Blast furnace slag	Compressive strength	D8	→	1/8
pair0027	Fly ash	Compressive strength	D8	→	1/8
pair0028	Water	Compressive strength	D8	→	1/8
pair0029	Superplasticizer	Compressive strength	D8	→	1/8
pair0030	Coarse aggregate	Compressive strength	D8	→	1/8
pair0031	Fine aggregate	Compressive strength	D8	→	1/8
pair0032	Age	Compressive strength	D8	→	1/8
pair0033	Alcohol consumption	Mean corpuscular volume	D9	→	1/5
pair0034	Alcohol consumption	Alkaline phosphatase	D9	→	1/5
pair0035	Alcohol consumption	Alanine aminotransferase	D9	→	1/5
pair0036	Alcohol consumption	Aspartate aminotransferase	D9	→	1/5
pair0037	Alcohol consumption	Gamma-glutamyl transpeptidase	D9	→	1/5
pair0038	Age	Body mass index	D10	→	1/4
pair0039	Age	Serum insulin	D10	→	1/4
pair0040	Age	Diastolic blood pressure	D10	→	1/4
pair0041	Age	Plasma glucose concentration	D10	→	1/4
pair0042	Day of the year	Temperature	D11	→	1/2
pair0043	Temperature at $t$	Temperature at $t + 1$	D12	→	1/4
pair0044	Surface pressure at $t$	Surface pressure at $t + 1$	D12	→	1/4
pair0045	Sea level pressure at $t$	Sea level pressure at $t + 1$	D12	→	1/4
pair0046	Relative humidity at $t$	Relative humidity at $t + 1$	D12	→	1/4
pair0047	Number of cars	Type of day	D13	↔	1
pair0048	Indoor temperature	Outdoor temperature	D14	↔	1
pair0049	Ozone concentration	Temperature	D15	↔	1/3
pair0050	Ozone concentration	Temperature	D15	↔	1/3
pair0051	Ozone concentration	Temperature	D15	↔	1/3
pair0052	(Temp, Press, SLP, RH)	(Temp, Press, SLP, RH)	D12	↔	0
pair0053	Ozone concentration	(Wind speed, Radiation, Temperature)	D16	↔	0
pair0054	(Displacement, Horsepower, Weight)	(Fuel consumption, Acceleration)	D4	→	0
pair0055	Ozone concentration (16-dim.)	Radiation (16-dim.)	D15	↔	0
pair0056	Female life expectancy, 2000–2005	Latitude of capital	D17	↔	1/12
pair0057	Female life expectancy, 1995–2000	Latitude of capital	D17	↔	1/12
pair0058	Female life expectancy, 1990–1995	Latitude of capital	D17	↔	1/12
pair0059	Female life expectancy, 1985–1990	Latitude of capital	D17	↔	1/12
pair0060	Male life expectancy, 2000–2005	Latitude of capital	D17	↔	1/12
pair0061	Male life expectancy, 1995–2000	Latitude of capital	D17	↔	1/12
pair0062	Male life expectancy, 1990–1995	Latitude of capital	D17	↔	1/12
pair0063	Male life expectancy, 1985–1990	Latitude of capital	D17	↔	1/12
pair0064	Drinking water access	Infant mortality	D17	→	1/12

pair0065	Stock return of Hang Seng Bank	Stock return of HSBC Hldgs	D18	→	1/3
pair0066	Stock return of Hutchison	Stock return of Cheung kong	D18	→	1/3
pair0067	Stock return of Cheung kong	Stock return of Sun Hung Kai Prop.	D18	→	1/3
pair0068	Bytes sent	Open http connections	D19	←	1
pair0069	Inside temperature	Outside temperature	D20	←	1
pair0070	Parameter	Answer	D21	→	1
pair0071	Symptoms (6-dim.)	Classification of disease (2-dim.)	D22	→	0
pair0072	Sunspots	Global mean temperature	D23	→	1
pair0073	CO <sub>2</sub> emissions	Energy use	D17	↔	1/12
pair0074	GNI per capita	Life expectancy	D17	→	1/12
pair0075	Under-5 mortality rate	GNI per capita	D17	↔	1/12
pair0076	Population growth	Food consumption growth	D24	→	1
pair0077	Temperature	Solar radiation	D11	↔	1/2
pair0078	PPFD	Net Ecosystem Productivity	D25	→	1/3
pair0079	Net Ecosystem Productivity	Diffuse PPFD	D25	↔	1/3
pair0080	Net Ecosystem Productivity	Direct PPFD	D25	↔	1/3
pair0081	Temperature	Local CO <sub>2</sub> flux, BE-Bra	D26	→	1/3
pair0082	Temperature	Local CO <sub>2</sub> flux, DE-Har	D26	→	1/3
pair0083	Temperature	Local CO <sub>2</sub> flux, US-PFa	D26	→	1/3
pair0084	Employment	Population	D27	↔	1
pair0085	Time of measurement	Protein content of milk	D28	→	1
pair0086	Size of apartment	Monthly rent	D29	→	1
pair0087	Temperature	Total snow	D30	→	1
pair0088	Age	Relative spinal bone mineral density	D31	→	1

## D1: DWD

The DWD climate data were provided by the Deutscher Wetterdienst (DWD). We downloaded the data from <http://www.dwd.de> and merged several of the original data sets to obtain data for 349 weather stations in Germany, selecting only those weather stations without missing data. After merging the data sets, we selected the following six variables: altitude, latitude, longitude, and annual mean values (over the years 1961–1990) of sunshine duration, temperature and precipitation. We converted the latitude and longitude variables from sexagesimal to decimal notation. Out of these six variables, we selected six different pairs with “obvious” causal relationships: altitude–temperature (**pair0001**), altitude–precipitation (**pair0002**), longitude–temperature (**pair0003**), altitude–sunshine hours (**pair0004**), latitude–temperature (**pair0020**), and longitude–precipitation (**pair0021**).

### pair0001: ALTITUDE → TEMPERATURE

As an elementary fact of meteorology, places with higher altitude tend to be colder than those that are closer to sea level (roughly 1 centigrade per 100 meter). There is no doubt that altitude is the cause and temperature the effect: one could easily think of an intervention where the thermometer is lifted (e.g., by using a balloon) to measure the temperature at a higher point of the same longitude and latitude. On the other hand, heating or cooling a location usually does not change its altitude (except perhaps if the location happens to be the space enclosed by a hot air balloon, but let us assume that the thermometers used to gather this data were fixed to the ground). The altitudes in the DWD data set range from 0 m to 2960 m, which is sufficiently large to detect significant statistical dependences.

One potential confounder is latitude, since all mountains are in the south and far from the sea, which is also an important factor for the local climate. The places with the highest average temperatures are therefore those with low altitude but lying far in the south. Hence this confounder should induce positive correlations between altitude and temperature as opposed to the negative correlation between altitude and temperature that is observed empirically. This suggests that the direct causal relation between altitude and temperature dominates over the confounder.

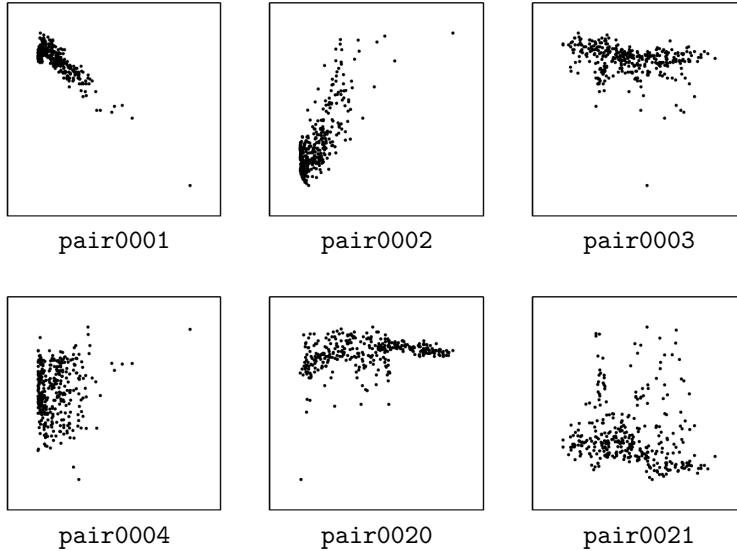


Figure 14: Scatter plots of pairs from D1. **pair0001**: altitude → temperature, **pair0002**: altitude → precipitation, **pair0003**: longitude → temperature, **pair0004**: altitude → sunshine hours, **pair0020**: latitude → temperature, **pair0021**: longitude → precipitation.

#### **pair0002: ALTITUDE → PRECIPITATION**

It is known that altitude is also an important factor for precipitation since rain often occurs when air is forced to rise over a mountain range and the air becomes oversaturated with water due to the lower temperature (orographic rainfall). This effect defines an indirect causal influence of altitude on precipitation via temperature. These causal relations are, however, less simple than the causal influence from altitude to temperature because gradients of the altitude with respect to the main direction of the wind are more relevant than the altitude itself. A hypothetical intervention that would allow us to validate the causal relation could be to build artificial mountains and observe orographic rainfall.

#### **pair0003: LONGITUDE → TEMPERATURE**

To detect the causal relation between longitude and temperature, a hypothetical intervention could be to move a thermometer between West and East. Even if one would adjust for altitude and latitude, it is unlikely that temperature would remain the same since the climate in the West is more oceanic and less continental than in the East of Germany. Therefore, longitude causes temperature.

#### **pair0004: ALTITUDE → SUNSHINE HOURS**

Sunshine duration and altitude are slightly positively correlated. Possible explanations are that higher weather stations are sometimes above low-hanging clouds. Cities in valleys, especially if they are close to rivers or lakes, typically have more misty days. Moving a sunshine sensor above the clouds clearly increases the sunshine duration whereas installing an artificial sun would not change the altitude. The causal influence from altitude to sun-

shine duration can be confounded, for instance, by the fact that there is a simple statistical dependence between altitude and longitude in Germany as explained earlier.

#### pair0020: LATITUDE → TEMPERATURE

Moving a thermometer towards the equator will generally result in an increased mean annual temperature. Changing the temperature, on the other hand, does not necessarily result in a north-south movement of the thermometer. The obvious ground truth of latitude causing temperature might be somewhat “confounded” by longitude, in combination with the selection bias that arises from only including weather stations in Germany.

#### pair0021: LONGITUDE → PRECIPITATION

As the climate in the West is more oceanic and less continental than in the East of Germany, we expect there to be a relationship between longitude and precipitation. Changing longitude by moving in East-West direction may therefore change precipitation, even if one would adjust for altitude and latitude. On the other hand, making it rain locally (e.g., by cloud seeding) will not result in a change in longitude.

## D2: Abalone

The **Abalone** data set (Nash et al., 1994) in the UCI Machine Learning Repository (Bache and Lichman, 2013) contains 4177 measurements of several variables concerning the sea snail *Abalone*. We downloaded the data from <https://archive.ics.uci.edu/ml/datasets/Abalone>. The original data set contains the nine variables sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and number of rings. The number of rings in the shell is directly related to the age of the snail: adding 1.5 to the number of rings gives the age in years. Of these variables, we selected six pairs with obvious cause-effect relationships: age–length (pair0005), age–shell weight (pair0006), age–diameter (pair0007), age–height (pair0008), age–whole weight (pair0009), age–shucked weight (pair0010), age–viscera weight (pair0011).

#### pair0005–pair0011: AGE → {LENGTH, SHELL WEIGHT, DIAMETER, HEIGHT, WHOLE/SHUCKED/VISCERA WEIGHT}

For the variable “age” it is not obvious what a reasonable intervention would be since there is no possibility to change the time. However, waiting and observing how variables change over time can be considered as equivalent to the hypothetical intervention on age (provided that the relevant background conditions do not change too much). Clearly, this “intervention” would change the probability distribution of the length, whereas changing the length of snails (by surgery) would not change the distribution of age (assuming that the surgery does not take years). Regardless of the difficulties of defining interventions, we expect common agreement on the ground truth: age causes all the other variables related to length, diameter height and weight.

There is one subtlety that has to do with how age is measured for these shells: this is done by counting the rings. For the variable “number of rings” however, changing the length of the snail may actually change the number of rings. We here presume that all

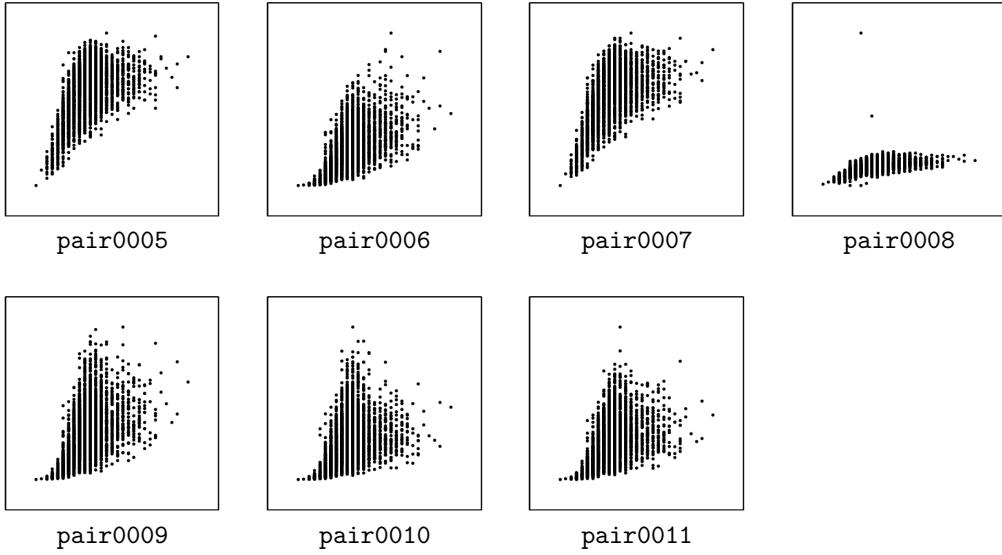


Figure 15: Scatter plots of pairs from D2. **pair0005**: age → length, **pair0006**: age → shell weight, **pair0007**: age → diameter, **pair0008**: age → height, **pair0009**: age → whole weight, **pair0010**: age → shucked weight, **pair0011**: age → viscera weight.

snails have undergone their natural growing process so that the number of rings is a good proxy for the variable age.

### D3: Census Income KDD

The Census Income (KDD) data set (U.S. Department of Commerce, 1994) in the UCI Machine Learning Repository (Bache and Lichman, 2013) has been extracted from the 1984 and 1985 U.S. Census studies. We downloaded the data from [https://archive.ics.uci.edu/ml/datasets/Census-Income+\(KDD\)](https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)). We have selected the following variables: AAGE (age), AHRSPAY (wage per hour) and DIVVAL (dividends from stocks).

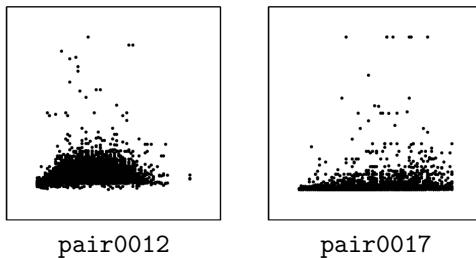


Figure 16: Scatter plots of pairs from D3. **pair0012**: age → wage per hour, **pair0017**: age → dividends from stocks.

**pair0012: AGE → WAGE PER HOUR**

We only used the first 5000 instances for which wage per hour was not equal to zero. The data clearly shows an increase of wage up to about 45 and a decrease for higher age.

As already argued for the **Abalone** data, interventions on the variable “age” are difficult to define. Compared to the discussion in the context of the **Abalone** data set, it seems more problematic to consider waiting as a reasonable “intervention” here, since the relevant (economical) background conditions change rapidly compared to the length of the human life: If someone’s salary is higher than the salary of a 20 year younger colleague *because* of his/her longer job experience, we cannot conclude that the younger colleague will earn the same money 20 years later as the older colleague earns now. Possibly, the factory or even the branch of industry he/she was working in does not exist any more and his/her job experience is no longer appreciated. However, we know that employees sometimes indeed do get a higher income because of their longer job experience. Pretending longer job experience by a fake certificate of employment would be a possible intervention. On the other hand, changing the wage per hour is an intervention that is easy to imagine (though difficult for us to perform) and this would certainly not change the age.

**pair0017: AGE → DIVIDENDS FROM STOCKS**

We only used the first 5000 instances for which dividends from stocks was not equal to zero. Similar considerations apply as for age vs. wage per hour. Doing an intervention on age is not practical, but companies could theoretically intervene on the dividends from stocks, and that would not result in a change of age, obviously. On the other hand, age influences income, and thereby over time, the amount of money that people can invest in stocks, and thereby, the amount of dividends they earn from stocks. This causal relation is a very indirect one, though, and the dependence between age and dividends from stock is less pronounced than that between age and wage per hour.

**D4: Auto-MPG**

The **Auto-MPG** data set in the UCI Machine Learning Repository (Bache and Lichman, 2013) concerns city-cycle fuel consumption in miles per gallon (MPG), i.e., the number of miles a car can drive on one gallon of gasoline, and contains several other attributes, like displacement, horsepower, weight, and acceleration. The original dataset comes from the StatLib library (Meyer and Vlachos, 2014) and was used in the 1983 American Statistical Association Exposition. We downloaded the data from <http://archive.ics.uci.edu/ml/datasets/Auto+MPG> and selected only instances without missing data, thereby obtaining 392 samples.

**pair0013: DISPLACEMENT → FUEL CONSUMPTION**

Displacement is the total volume of air/fuel mixture an engine can draw in during one complete engine cycle. The larger the displacement, the more fuel the engine can consume with every turn. Intervening on displacement (e.g., by increasing the cylinder bore) changes the fuel consumption. Changing the fuel consumption (e.g., by increasing the weight of the

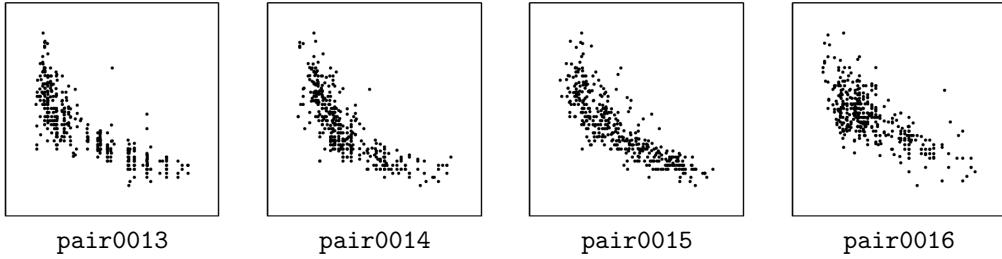


Figure 17: Scatter plots of pairs from D4. **pair0013:** displacement → fuel consumption, **pair0014:** horse power → fuel consumption, **pair0015:** weight → fuel consumption, **pair0016:** horsepower → acceleration, **pair0054:** (displacement,horsepower,weight) → (MPG,acceleration)

car, or changing its air resistance, or by using another gear) will not change the displacement, though.

#### **pair0014: HORSE POWER → FUEL CONSUMPTION**

Horse power measures the amount of power an engine can deliver. There are various ways to define horsepower and different standards to measure horse power of vehicles. In general, though, it should be obvious that fuel consumption depends on various factors, including horse power. Changing horsepower (e.g., by adding more cylinders to an engine, or adding a second engine to the car) would lead to a change in fuel consumption. On the other hand, changing fuel consumption does not necessarily change horse power.

#### **pair0015: WEIGHT → FUEL CONSUMPTION**

There is a strong selection bias here, as car designers use a more powerful motor (with higher fuel consumption) for a heavier car. Nevertheless, the causal relationship between weight and fuel consumption should be obvious: if we intervene on weight, then fuel consumption will change, but not necessarily vice versa.

#### **pair0016: HORSEPOWER → ACCELERATION**

Horsepower is one of the factors that cause acceleration. Other factors are wheel size, the gear used, and air resistance. However, note that when a car is designed, horsepower is chosen with the goal of being able to achieve a certain maximum acceleration, given for example the weight of a car. Indeed, it does not make sense to put a small engine into a big truck. Therefore, there is a strong selection bias on horse power and acceleration.

#### **pair0054: (DISPLACEMENT,HORSEPOWER,WEIGHT) → (MPG,ACCELERATION)**

This pair consists of two multivariate variables that are combinations of the variables we have considered before. The multivariate variable consisting of the three components displacement, horsepower and weight can be considered to cause the multivariate variable comprised of fuel consumption and acceleration.

**D5: GAGurine**

This data concerns the concentration of the chemical compound Glycosaminoglycan (GAG) in the urine of 314 children aged from zero to seventeen years. This is the **GAGurine** data set supplied with the MASS package of the computing language R ([Venables and Ripley, 2002](#)).

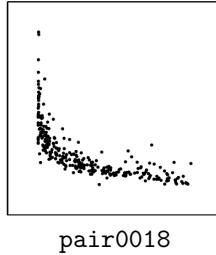


Figure 18: Scatter plots of pairs from D5. pair0018: age → concentration GAG.

**pair0018: AGE → CONCENTRATION GAG**

Obviously, GAG concentration does not cause age, but it could be the other way around, considering the strong dependence between the two variables.

**D6: Old Faithful**

This is the **geyser** data set supplied with the MASS package of the computing language R ([Venables and Ripley, 2002](#)). It is originally described in ([Azzalini and Bowman, 1990](#)) and contains data about the duration of an eruption and the time interval between subsequent eruptions of the Old Faithful geyser in Yellowstone National Park, USA. The data consists of 194 samples and was collected in a single continuous measurement from August 1 to August 15, 1985.

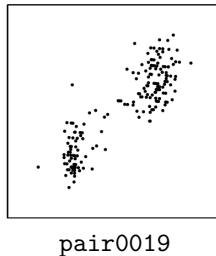


Figure 19: Scatter plots of pairs from D6. pair0019: current duration → next interval.

**pair0019: CURRENT DURATION → NEXT INTERVAL**

The chronological ordering of events implicates that the time interval between the current and the next eruption is an effect of the duration of the current eruption.

## D7: Arrhythmia

The **Arrhythmia** dataset (Guvenir et al., 1997) from the UCI Machine Learning Repository (Bache and Lichman, 2013) concerns cardiac arrhythmia. It consists of 452 patient records and contains many different variables. We downloaded the data from <https://archive.ics.uci.edu/ml/datasets/Arrhythmia> and only used the variables for which the causal relationships should be evident. We removed two instances from the dataset, corresponding with patient lengths of 680 and 780 cm, respectively.

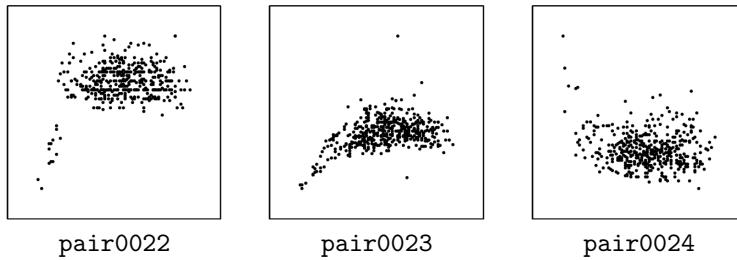


Figure 20: Scatter plots of pairs from D7. pair0022: age → height, pair0023: age → weight, pair0024: age → heart rate.

pair0022–pair0024: AGE → {HEIGHT, WEIGHT, HEART RATE}

As discussed before, “interventions” on age (for example, waiting a few years) may affect height of persons. On the other hand, we know that height does not cause age. The same holds for age and weight and for age and heart rate. It is important to note here that age is simply measured in years since the birth of a person. Indeed, weight, height and also heart rate might influence “biological aging”, the gradual deterioration of function of the human body.

## D8: Concrete Compressive Strength

This data set, available at the UCI Machine Learning Repository (Bache and Lichman, 2013), concerns a systematic study (Yeh, 1998) regarding concrete compressive strength as a function of ingredients and age. Citing (Yeh, 1998): “High-performance concrete (HPC) is a new terminology used in the concrete construction industry. In addition to the three basic ingredients in conventional concrete, i.e., Portland cement, fine and coarse aggregates, and water, the making of HPC needs to incorporate supplementary cementitious materials, such as fly ash and blast furnace slag, and chemical admixture, such as superplasticizer 1 and 2. Several studies independently have shown that concrete strength development is determined not only by the water-to-cement ratio, but that it also is influenced by the content of other concrete ingredients.” Compressive strength is measured in units of MPa, age in days, and the other variables are measured in kilograms per cubic metre of concrete mixture. The dataset was downloaded from <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength> and contains 1030 measurements.

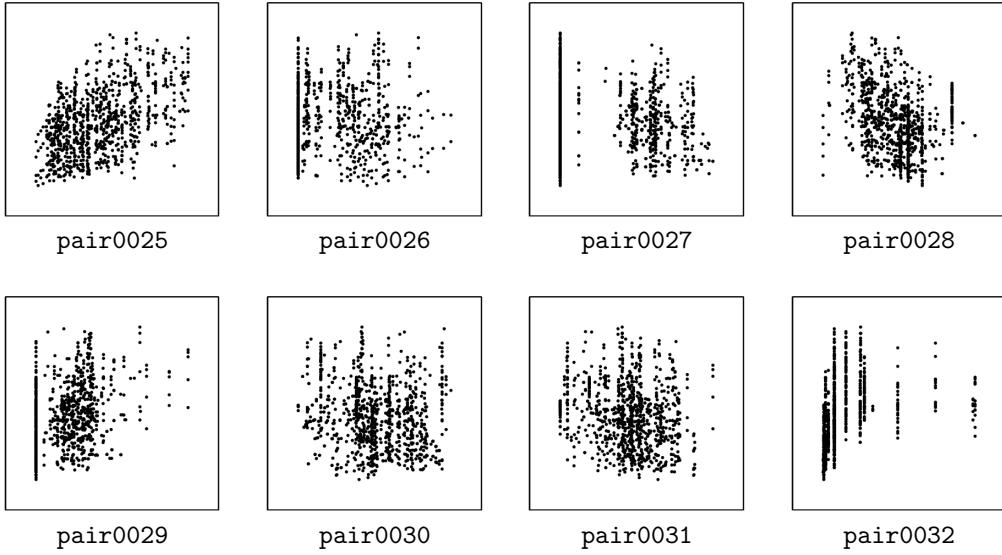


Figure 21: Scatter plots of pairs from D8. pair0025: cement → compressive strength, pair0026: blast furnace slag → compressive strength, pair0027: fly ash → compressive strength, pair0028: water → compressive strength, pair0029: superplasticizer → compressive strength, pair0030: coarse aggregate → compressive strength, pair0031: fine aggregate → compressive strength, pair0032: age → compressive strength.

pair0025–pair0032: {CEMENT, BLAST FURNACE SLAG, FLY ASH, WATER, SUPERPLASTICIZER, COARSE AGGREGATE, FINE AGGREGATE, AGE} → COMPRESSIVE STRENGTH

It should be obvious that compressive strength is the effect, and the other variables are its causes. Note, however, that in practice one cannot easily intervene on the mixture components without simultaneously changing the other mixture components. For example, if one adds more water to the mixture, then as a result, all other components will decrease, as they are measured in kilograms per cubic metre of concrete mixture. Nevertheless, we expect that we can see these interventions as reasonable approximations of “perfect interventions” on a single variable.

## D9: Liver Disorders

This data set, available at the UCI Machine Learning Repository (Bache and Lichman, 2013), was collected by BUPA Medical Research Ltd. It consists of several blood test results, which are all thought to be indicative for liver disorders that may arise from excessive alcohol consumption. Each of the 345 instances constitutes the record of a single male individual. Daily alcohol consumption is measured in number of half-pint equivalents of alcoholic beverages drunk per day. The blood test results are mean corpuscular volume (MCV), alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma-glutamyl transpeptidase (GGT). The data is available at <https://archive.ics.uci.edu/ml/datasets/Liver+Disorders>.

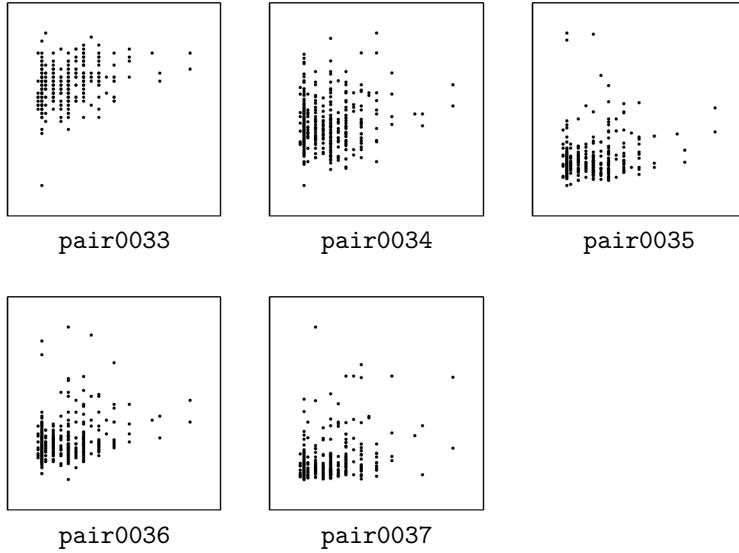


Figure 22: Scatter plots of pairs from D9. **pair0033**: alcohol consumption → mean corpuscular volume, **pair0034**: alcohol consumption → alkaline phosphatase, **pair0035**: alcohol consumption → alanine aminotransferase, **pair0036**: alcohol consumption → aspartate aminotransferase, **pair0037**: alcohol consumption → gamma-glutamyl transpeptidase.

Although one would expect that daily alcohol consumption is the cause, and the blood test results are the effects, this is not necessarily the case. Indeed, citing (Baynes and Dominiczak, 1999): “[...] increased plasma concentrations of acetaldehyde after the ingestion of alcohol [...] causes the individual to experience unpleasant flushing and sweating, which discourages alcohol abuse. Disulfiram, a drug that inhibits ALDH, also leads to these symptoms when alcohol is taken, and may be given to reinforce abstinence from alcohol.” This means that *a priori*, a reverse causation of the chemical whose concentration is measured in one of these blood tests on daily alcohol consumption cannot be excluded with certainty. Nevertheless, we consider this to be unlikely, as the medical literature describes how these particular blood tests can be used to diagnose liver disorders, but we did not find any evidence that these chemicals can be used to *treat* excessive alcohol consumption.

#### **pair0033: ALCOHOL CONSUMPTION → MEAN CORPUSCULAR VOLUME**

The mean corpuscular volume (MCV) is the average volume of a red blood cell. An elevated MCV has been associated with alcoholism (Tønnesen et al., 1986), but there are many other factors also associated with MCV.

#### **pair0034: ALCOHOL CONSUMPTION → ALKALINE PHOSPHOTASE**

Alkaline phosphatase (ALP) is an enzyme that is predominantly abundant in liver cells, but is also present in bone and placental tissue. Elevated ALP levels in blood can be due to many different liver diseases and also bone diseases, but also occur during pregnancy (Braunwald et al., 2001).

**pair0035: ALCOHOL CONSUMPTION → ALANINE AMINOTRANSFERASE**

Alanine Aminotransferase (ALT) is an enzyme that is found primarily in the liver cells. It is released into the blood in greater amounts when there is damage to the liver cells, for example due to a viral hepatitis or bile duct problems. ALT levels are often normal in alcoholic liver disease (Braunwald et al., 2001).

**pair0036: ALCOHOL CONSUMPTION → ASPARTATE AMINOTRANSFERASE**

Aspartate aminotransferase (AST) is an enzyme that is found in the liver, but also in many other bodily tissues, for example the heart and skeletal muscles. Similar to ALT, the AST levels raise in acute liver damage. Elevated AST levels are not specific to the liver, but can also be caused by other diseases, for example by pancreatitis. An AST:ALT ratio of more than 3:1 is highly suggestive of alcoholic liver disease (Braunwald et al., 2001).

**pair0037: ALCOHOL CONSUMPTION → GAMMA-GLUTAMYL TRANSPEPTIDASE**

Gamma-Glutamyl Transpeptidase (GGT) GGT is another enzyme that is primarily found in liver cells. It is rarely elevated in conditions other than liver disease. High GGT levels have been associated with alcohol use (Braunwald et al., 2001).

**D10: Pima Indians Diabetes**

This data set, available at the UCI Machine Learning Repository (Bache and Lichman, 2013), was collected by the National Institute of Diabetes and Digestive and Kidney Diseases in the USA to forecast the onset of diabetes mellitus in a high risk population of Pima Indians near Phoenix, Arizona. Cases in this data set were selected according to several criteria, in particular being female, at least 21 years of age and of Pima Indian heritage. This means that there could be selection bias on age.

We downloaded the data from <https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>. We only selected the instances with nonzero values, as it seems likely that zero values encode missing data. This yielded 768 samples.

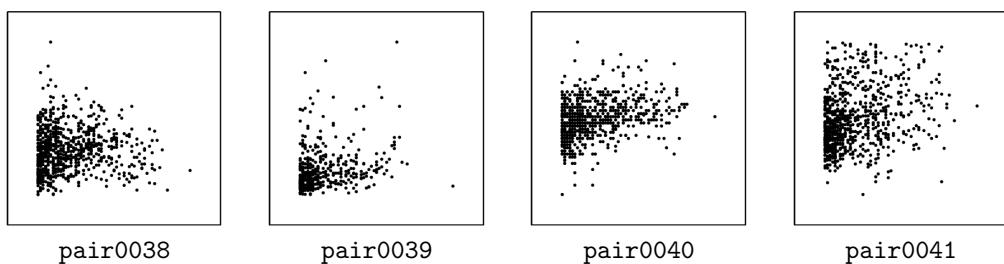


Figure 23: Scatter plots of pairs from D10. pair0038: age → body mass index, pair0039: age → serum insulin, pair0040: age → diastolic blood pressure, pair0041: age → plasma glucose concentration.

**pair0038: AGE → BODY MASS INDEX**

Body mass index (BMI) is defined as the ratio between weight (kg) and the square of height (m). Obviously, age is not caused by body mass index, but as age is a cause of both height and weight, age causes BMI.

**pair0039: AGE → SERUM INSULIN**

2-Hour serum insulin ( $\mu\text{U}/\text{ml}$ ), measured 2 hours after the ingestion of a standard dose of glucose, in an oral glucose tolerance test. We can exclude that serum insulin causes age, and there could be an effect of age on serum insulin. Another explanation for the observed dependence could be the selection bias.

**pair0040: AGE → DIASTOLIC BLOOD PRESSURE**

Diastolic blood pressure (mm Hg). It seems obvious that blood pressure does not cause age. The other causal direction seems plausible, but again, an alternative explanation for the dependence could be selection bias.

**pair0041: AGE → PLASMA GLUCOSE CONCENTRATION**

Plasma glucose concentration, measured 2 hours after the ingestion of a standard dose of glucose, in an oral glucose tolerance test. Similar reasoning as before: we do not believe that plasma glucose concentration causes ages, but it could be the other way around, and there may be selection bias.

**D11: B. Janzing's meteo data**

This data set is from a private weather station, owned by Bernward Janzing, located in Furtwangen (Black Forest), Germany at an altitude of 956 m. The measurements include temperature, precipitation, and snow height (since 1979), as well as solar radiation (since 1986). The data have been archived by Bernward Janzing, statistical evaluations have been published in ([Janzing, 2004](#)), monthly summaries of the weather are published in local newspapers since 1981.



Figure 24: Scatter plots of pairs from D11. pair0042: day of the year → temperature, pair0077: solar radiation → temperature.

**pair0042: DAY OF THE YEAR → TEMPERATURE**

This data set shows the dependence between season and temperature over 25 years plus one month, namely the time range 01/01/1979–01/31/2004. It consists of 9162 measurements.

One variable is the day of the year, represented by an integer from 1 to 365 (or 366 for leap years). The information about the year has been dropped.  $Y$  is the mean temperature of the respective day, calculated according to the following definition:

$$T_{mean} := \frac{T_{morning} + T_{midday} + 2T_{evening}}{4},$$

where morning, midday, and evening are measured at 7:00 am, 14:00 pm, and 21:00 pm (MEZ), respectively (without daylight saving time). Double counting of the evening value is official standard of the German authority “Deutscher Wetterdienst”. It has been defined at a time where no electronic data loggers were available and thermometers had to be read out by humans. Weighting the evening value twice has been considered a useful heuristics to account for the missing values at night.

We consider day of the year as the cause, since it can be seen as expressing the angular position on its orbit around the sun. Although true interventions are infeasible, it is commonly agreed that changing the position of the earth would result in temperature changes at a fixed location due to the different solar incidence angle.

**pair0077: SOLAR RADIATION → TEMPERATURE**

This data set shows the relation between solar radiation and temperature over 23 years, namely the interval 01/01/1986–12/31/2008. It consists of 8401 measurements.

Solar radiation is measured per area in  $\text{W}/\text{m}^2$  averaged over one day on a horizontal surface. Temperature is the averaged daily, as in pair0042. The original data has been processed by us to extract the common time interval. We assume that radiation causes temperature. High solar radiation increases the temperature of the air already at a scale of hours. Interventions are easy to implement: Creating artificial shade on a large enough surface would decrease the air temperature. On longer time scales there might also be an influence from temperature to radiation via the generation of clouds through evaporation in more humid environments. This should, however, not play a role for daily averages.

**D12: NCEP-NCAR Reanalysis**

This data set, available from the NOAA (National Oceanic and Atmospheric Administration) Earth System Research Laboratory website at <http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.surface.html>, is a subset of a reanalysis data set, incorporating observations and numerical weather prediction model output from 1948 to date (Kalnay et al., 1996). The reanalysis data set is a joint product from the National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR). Reanalysis data products aim for a realistic representation of all relevant climatological variables on a spatiotemporal grid. We collected four variables from a global grid of  $144 \times 73$  cells: air temperature (in K, pair0043), surface pressure (in Pascal, pair0044), sea level pressure (in Pascal, pair0045) and relative humidity (in %, pair0045) on two consecutive days, day 50 and day 51 of the year 2000 (i.e., Feb 19th and 20th). Each

data pair consists of  $144 \times 73 - 143 = 10369$  data points, distributed across the globe. 143 data points were subtracted because at the north pole values are repeated across all longitudes.



Figure 25: Scatter plots of pairs from D12. **pair0043:** temperature at  $t \rightarrow$  temperature at  $t+1$ , **pair0044:** surface oressure at  $t \rightarrow$  surface pressure at  $t+1$ , **pair0045:** sea level oressure at  $t \rightarrow$  sea level pressure at  $t+1$ , **pair0046:** relative humidity at  $t \rightarrow$  relative humidity at  $t+1$ , **pair0052:** (temp, press, slp, rh) at  $t \rightarrow$  (temp, press, slp, rh) at  $t+1$ .

Each data point is the daily average over an area that covers  $2.5^\circ \times 2.5^\circ$  (approximately 250 km  $\times$  250 km at the equator). Because causal influence cannot propagate backwards in time, temperature, pressure and humidity in a certain area are partly affected by their value the day before in the same area.

#### **pair0043: TEMPERATURE AT $t \rightarrow$ TEMPERATURE AT $t+1$**

Due to heat storage, mean daily air temperature near surface at any day largely impact daily air temperature at the following day. We assume there is no causation backwards in time, hence the correlation between temperatures at two consecutive days must be driven by confounders (such as large-scale weather patterns) or a causal influence from the first day to the second.

#### **pair0044: SURFACE PRESSURE AT $t \rightarrow$ SURFACE PRESSURE AT $t+1$**

Pressure patterns near the earth's surface are mostly driven by large-scale weather patterns. However, large-scale weather patterns are also driven by local pressure gradients and hence, some of the correlation between surface pressure at two consecutive days stems from a direct causal link between the first and the second day, as we assume there is no causation in time.

#### **pair0045: SEA LEVEL PRESSURE AT $t \rightarrow$ SEA LEVEL PRESSURE AT $t+1$**

Similar reasoning as in **pair0044**.

#### **pair0046: RELATIVE HUMIDITY AT $t \rightarrow$ RELATIVE HUMIDITY AT $t+1$**

Humidity of the air at one day affects the humidity of the following day because if no air movement takes place and no drying or moistening occurs, it will approximately stay the same. Furthermore, as reasoned above, because there is no causation backwards in time, relative humidity at day  $t + 1$  cannot affect humidity at day  $t$ . Note that relative humidity has values between 0 and 100. Values can be saturated in very humid places such as tropical

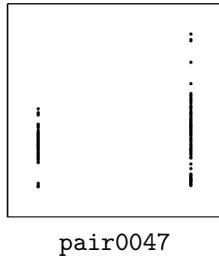
rainforest and approach 0 in deserts. For this reason, the scatter plot looks as if the data were clipped.

**pair0052:** (TEMP, PRESS, SLP, RH) AT  $t \rightarrow$  (TEMP, PRESS, SLP, RH) AT  $t+1$

The pairs pair0043–pair0046 were combined to a 4-dimensional vector. From the reasoning above it follows that the vector of temperature, near surface pressure, sea level pressure and relative humidity at day  $t$  has a causal influence on the vector of the same variables at time  $t + 1$ .

### D13: Traffic

This dataset has been extracted from <http://www.b30-ober schwaben.de/html/tabelle.html>, a website containing various kinds of information about the national highway B30. This is a road in the federal state Baden-Württemberg, Germany, which provides an important connection of the region around Ulm (in the North) with the Lake Constance region (in the South). After extraction, the data set contains 254 samples.



pair0047

Figure 26: Scatter plots of pairs from D13. pair0047: type of day  $\rightarrow$  number of cars.

**pair0047: TYPE OF DAY  $\rightarrow$  NUMBER OF CARS**

One variable is the number of cars per day, the other denotes the type of the respective day, with “1” indicating Sundays and holidays and “2” indicating working days. The type of day causes the number of cars per day. Indeed, introducing an additional holiday by a political decision would certainly change the amount of traffic on that day, while changing the amount of traffic by instructing a large number of drivers to drive or not to drive at a certain day would certainly not change the type of that day.

### D14: Hipel & McLeod

This dataset contains 168 measurements of indoor and outdoor temperatures. It was taken from a book by [Hipel and McLeod \(1994\)](#) and can be downloaded from <http://www.stats.uwo.ca/faculty/mcleod/epubs/mhsets/readme-mhsets.html>.

**pair0048: OUTDOOR TEMPERATURE  $\rightarrow$  INDOOR TEMPERATURE**

Outdoor temperatures can have a strong impact on indoor temperatures, in particular when indoor temperatures are not adjusted by air conditioning or heating. Contrarily, indoor

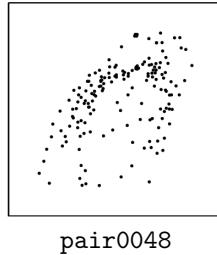


Figure 27: Scatter plots of pairs from D14. **pair0048:** outdoor temperature → indoor temperature.

temperatures will have little or no effect on outdoor temperatures, because the outside environment has a much larger heat capacity.

### D15: Bafu

This data set deals with the relationship between daily ozone concentration in the air and temperature. It was downloaded from [http://www.bafu.admin.ch/luft/luftbelastung/blick\\_zurueck/datenabfrage/index.html](http://www.bafu.admin.ch/luft/luftbelastung/blick_zurueck/datenabfrage/index.html). Lower atmosphere ozone ( $O_3$ ) is a secondary pollutant that is produced by the photochemical oxidation of carbon monoxide (CO), methane ( $CH_4$ ), and non-methane volatile organic compounds (NMVOCs) by OH in the presence of nitrogen oxides ( $NO_x$ ,  $NO + NO_2$ ) (Rasmussen et al., 2012). It is known that ozone concentration strongly correlates with surface temperature (Bloomer et al., 2009). Several explanations are given in the literature (see e.g., Rasmussen et al., 2012). Without going into details of the complex underlying chemical processes, we mention that the crucial chemical reactions are stronger at higher temperatures. For instance, isoprene emissions of plants increase with increasing temperature and isoprene can play a similar role in the generation of  $O_3$  as  $NO_x$  (Rasmussen et al., 2012). Apart from this, air pollution may be influenced indirectly by temperature, e.g., via increasing traffic at ‘good’ weather conditions or an increased occurrence rate of wildfires. All these explanations state a causal path from temperature to ozone. Note that the phenomenon of ozone pollution in the lower atmosphere discussed here should not be confused with the ‘ozone hole’, which is a lack of ozone in the higher atmosphere. Close to the surface, ozone concentration does not have an impact on temperatures. For all three data sets, ozone is measured in  $\mu\text{g}/\text{m}^3$  and temperature in  $^\circ\text{C}$ .

#### **pair0049: TEMPERATURE → OZONE CONCENTRATION**

365 daily mean values of ozone and temperature of year 2009 in Lausanne-César-Roux, Switzerland.

#### **pair0050: TEMPERATURE → OZONE CONCENTRATION**

365 daily mean values of ozone and temperature of year 2009 in Chaumont, Switzerland.

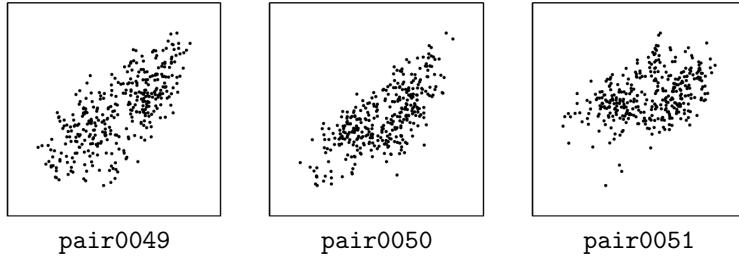


Figure 28: Scatter plots of pairs from D15. pair0049: temperature → ozone concentration, pair0050: temperature → ozone concentration, pair0051: temperature → ozone concentration, pair0055: radiation → ozone concentration.

#### pair0051: TEMPERATURE → OZONE CONCENTRATION

365 daily mean values of ozone and temperature of year 2009 in Davos-See, Switzerland.

#### pair0055: RADIATION → OZONE CONCENTRATION

72 daily mean values of ozone concentrations and radiation in the last 83 days of 2009 at 16 different places in Switzerland (11 days were deleted due to missing data). Solar radiation and surface ozone concentration are correlated (Feister and Balzer, 1991). The deposition of ozone is driven by complex micrometeorological processes including wind direction, air temperature, and global radiation (Stockwell et al., 1997). For instance, solar radiation affects the height of the planetary boundary layer and cloud formation and thus indirectly influences ozone concentrations. In contrast, global radiation is not driven by ozone concentrations close to the surface.

Ozone is given in  $\mu\text{g}/\text{m}^3$ , radiation in  $\text{W}/\text{m}^2$ . The 16 different places are: 1: Bern-Bollwerk, 2: Magadino-Cadenazzo, 3: Lausanne-César-Roux, 4: Payerne, 5: Lugano-Universita, 6: Taenikon, 7: Zuerich-Kaserne, 8: Laegeren, 9: Basel-Binningen, 10: Chauumont, 11: Duebendorf, 12: Rigi-Seebodenalp, 13: Haerkingen, 14: Davos-See, 15: Sion-Aéroport, 16: Jungfraujoch.

#### D16: Environmental

We downloaded ozone concentration, wind speed, radiation and temperature from <http://www.mathe.tu-freiberg.de/Stoyan/umwdat.html>, discussed in Stoyan et al. (1997). The data consist of 989 daily values over the time period from 05/01/1989 to 10/31/1994 observed in Heilbronn, Germany.

#### pair0053: (WIND SPEED, RADIATION, TEMPERATURE) → OZONE CONCENTRATION

As we have argued above in Section D15, wind direction (and speed), air temperature, and global radiation influence local ozone concentrations. Wind can influence ozone concentrations for example in the following way. No wind will keep the the concentration of ozone in a given air parcel constant if no lateral or vertical sources or sinks are prevalent. In contrast, winds can move and disperse and hence mix air with different ozone concentrations.

Ozone concentration is given in  $\mu\text{g}/\text{m}^3$ , wind speed in m/s, global radiation in  $\text{W}/\text{m}^2$  and temperature in  $^\circ\text{C}$ .

### D17: UNdata

The following data were taken from the “UNdata” database of the United Nations Statistics Division at <http://data.un.org>.

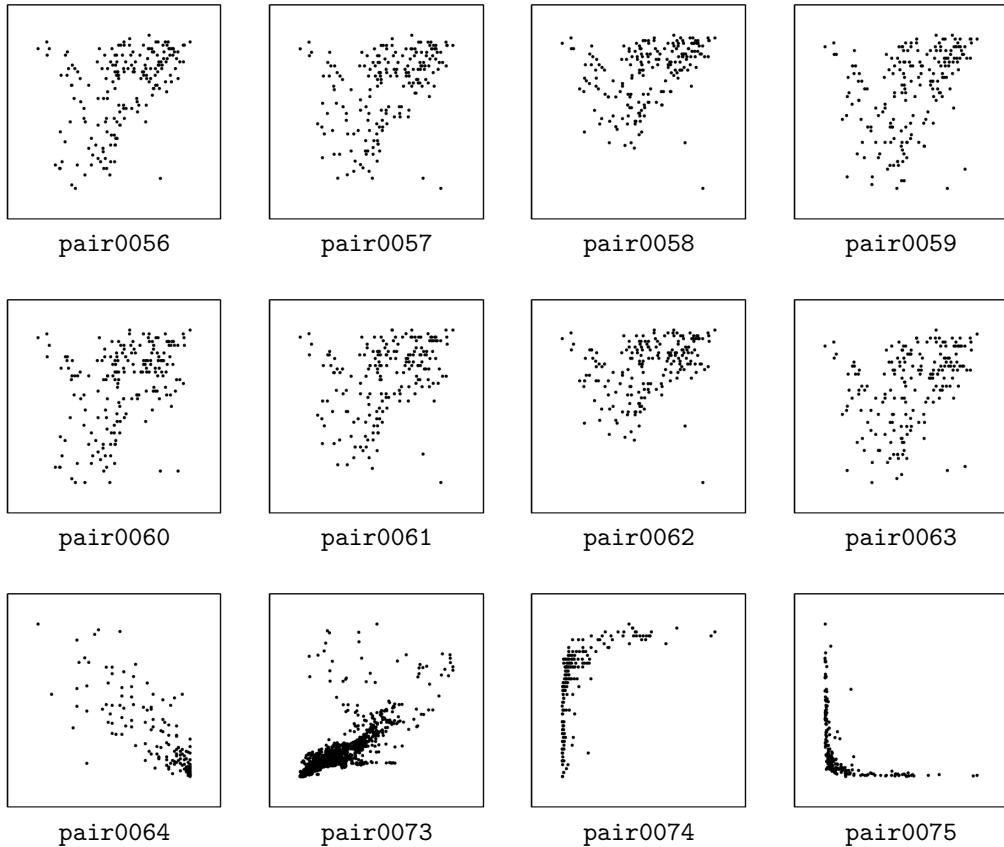


Figure 29: Scatter plots of pairs from D17. pair0056–pair0059: latitude of capital → female life expectancy, pair0060–pair0063: latitude of capital → male life expectancy, pair0064: drinking water access → infant mortality, pair0073: energy use →  $\text{CO}_2$  emissions, pair0074: GNI per capita → life expectancy, pair0075: GNI per capita → under-5 mortality rate.

#### pair0056–pair0059: LATITUDE OF CAPITAL → FEMALE LIFE EXPECTANCY

Pairs pair0056–pair0059 consist of female life expectancy (in years) at birth versus latitude of the country’s capital, for various countries (China, Russia and Canada were removed). The four pairs correspond with measurements over the periods 2000–2005, 1995–2000, 1990–1995, 1985–1990, respectively. The data were downloaded from <http://data.un.org/Data.aspx?d=GenderStat&f=inID%3a37>.

The location of a country (encoded in the latitude of its capital) has an influence on how poor or rich a country is, hence affecting the quality of the health care system and ultimately life expectancy. This influence could stem from abundance of natural resources within the country's borders or the influence neighboring countries have on its economic welfare. Furthermore, the latitude can influence life expectancy via climatic factors. For instance, life expectancy might be smaller if a country frequently experiences climatic extremes. In contrast, it is clear that life expectancy does not have any effect on latitude.

#### pair0060–pair0063: LATITUDE OF CAPITAL → MALE LIFE EXPECTANCY

Pairs pair0060–pair0063 are similar, but concern male life expectancy. The same reasoning as for female life expectancy applies here.

#### pair0064: DRINKING WATER ACCESS → INFANT MORTALITY

Here, one variable describes the percentage of population with sustainable access to improved drinking water sources in 2006, whereas the other variable denotes the infant mortality rate (per 1000 live births) for both sexes. The data were downloaded from <http://data.un.org/Data.aspx?d=WHO&f=inID%3aMBD10> and <http://data.un.org/Data.aspx?d=WHO&f=inID%3aRF03>, respectively, and consist of 163 samples.

Clean drinking water is a primary requirement for health, in particular for infants ([Esrey et al., 1991](#)). Changing the percentage of people with access to clean water will directly change the mortality rate of infants, since infants are particularly susceptible to diseases ([Lee et al., 1997](#)). There may be some feedback, because if infant mortality is high in a poor country, development aid may be directed towards increasing the access to clean drinking water.

#### pair0073: ENERGY USE → CO<sub>2</sub> EMISSIONS

This data set contains energy use (in kg of oil equivalent per capita) and CO<sub>2</sub> emission data from 152 countries between 1960 and 2005, yielding together 5084 samples. Considering the current energy mix across the world, the use of energy clearly results in CO<sub>2</sub> emissions (although in varying amounts across energy sources). Contrarily, a hypothetical change in CO<sub>2</sub> emissions will not affect the energy use of a country on the short term. On the longer term, if CO<sub>2</sub> emissions increase, this may cause energy use to decrease because of fear for climate change.

#### pair0074: GNI PER CAPITA → LIFE EXPECTANCY

We collected the Gross National Income (GNI, in USD) per capita and the life expectancy at birth (in years) for 194 different countries. GNI can be seen as an index of wealth of a country. In general, richer countries have a better health care system than poor countries and thus can take better care of their citizens when they are ill. Reversely, we believe that the life expectancy of humans has a smaller impact on how wealthy a country is than vice versa.

**pair0075: GNI PER CAPITA → UNDER-5 MORTALITY RATE**

Here we collected the Gross National Income (GNI, in USD) per capita and the under-5 mortality rate (deaths per 1000 live births) for 205 different countries. The reasoning is similar as in **pair0074**. GNI as an index of wealth influences the quality of the health care system, which in turn determines whether young children will or will not die from minor diseases. As children typically do not contribute much to GNI per capita, we do not expect the reverse causal relation to be very strong.

**D18: Yahoo database**

These data denote stock return values and were downloaded from <http://finance.yahoo.com>. We collected 1331 samples from the following stocks between January 4th, 2000 and June 17, 2005: Hang Seng Bank (0011.HK), HSBC Hldgs (0005.HK), Hutchison (0013.HK), Cheung kong (0001.HK), and Sun Hung Kai Prop. (0016.HK). Subsequently, the following preprocessing was applied, which is common in financial data processing:

1. Extract the dividend/split adjusted closing price data from the Yahoo Finance data base.
2. For the few days when the price is not available, we use simple linear interpolation to estimate the price.
3. For each stock, denote the closing price on day  $t$  by  $P_t$ , and the corresponding return is calculated as  $X_t = (P_t - P_{t-1})/P_{t-1}$ .

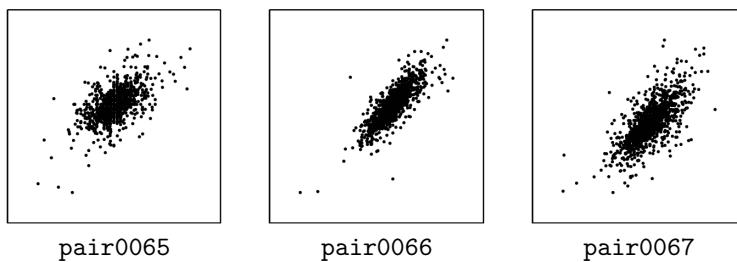


Figure 30: Scatter plots of pairs from D18. **pair0065**: Stock Return of Hang Seng Bank → Stock Return of HSBC Hldgs, **pair0066**: Stock Return of Hutchison → Stock Return of Cheung kong, **pair0067**: Stock Return of Cheung kong → Stock Return of Sun Hung Kai Prop.

**pair0065: STOCK RETURN OF HANG SENG BANK → STOCK RETURN OF HSBC HLDGS**

HSBC owns 60% of Hang Seng Bank. Consequently, if stock returns of Hang Seng Bank change, this should have an influence on stock returns of HSBC Hldgs, whereas causation in the other direction would be expected to be less strong.

**pair0066: STOCK RETURN OF HUTCHISON → STOCK RETURN OF CHEUNG KONG**

Cheung kong owns about 50% of Hutchison. Same reasoning as in **pair0065**.

**pair0067: STOCK RETURN OF CHEUNG KONG → STOCK RETURN OF SUN HUNG KAI PROP.**

Sun Hung Kai Prop. is a typical stock in the Hang Seng Property subindex, and is believed to depend on other major stocks, including Cheung kong.

### D19: Internet traffic data

This dataset has been created from the log-files of a http-server of the Max Planck Institute for Intelligent Systems in Tübingen, Germany. The variable Internet connections counts the number of times an internal website of the institute has been accessed during a time interval of 1 minute (more precisely, it counts the number of URL requests). Requests for non-existing websites are not counted. The variable Byte transferred counts the total number of bytes sent for all those accesses during the same time interval. The values  $(x_1, y_1), \dots, (x_{498}, y_{498})$  refer to 498 time intervals. To avoid too strong dependence between the measurements, the time intervals are not adjacent but have a distance of 20 minutes.

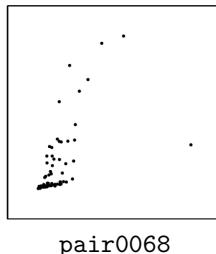


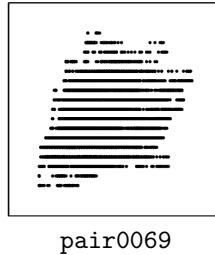
Figure 31: Scatter plots of pairs from D19. pair0068: internet connections → bytes transferred.

**pair0068: INTERNET CONNECTIONS → BYTES TRANSFERRED**

Internet connections causes Bytes transferred because an additional access of the website raises the transfer of data, while transferring more data does not create an additional website access. Note that not every access yields data transfer because the website may still be cached. However, this fact does not spoil the causal relation, it only makes it less deterministic.

### D20: Inside and outside temperature

This bivariate time-series data consists of measurements of inside room temperature ( $^{\circ}\text{C}$ ) and outside temperature ( $^{\circ}\text{C}$ ), where measurements were taken every 5 minutes for a period of about 56 days, yielding a total of 16382 measurements. The outside thermometer was located on a spot that was exposed to direct sunlight, which explains the large fluctuations. The data were collected by Joris M. Mooij.



pair0069

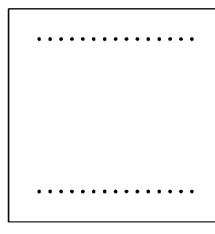
Figure 32: Scatter plots of pairs from D20. pair0069: outside temperature → inside temperature.

#### pair0069: OUTSIDE TEMPERATURE → INSIDE TEMPERATURE

Although there is a causal relationship in both directions, we expect that the strongest effect is from outside temperature on inside temperature, as the heat capacity of the inside of a house is much smaller than that of its surroundings. See also the reasoning for pair0048.

#### D21: Armann & Buelthoff’s data

This dataset is taken from a psychological experiment that artificially generates images of human faces that interpolate between male and female, taking real faces as basis ([Armann and Bülthoff, 2012](#)). The interpolation is done via principal component analysis after representing true face images as vectors in an appropriate high-dimensional space. Human subjects are instructed to label the faces as male or female. The variable “parameter” runs between 0 and 14 and describes the transition from female to male. It is chosen by the experimenter. The binary variable “answer” indicates the answers ‘female’ and ’male’, respectively. The dataset consists of 4499 samples.



pair0070

Figure 33: Scatter plots of pairs from D21. pair0070: parameter → answer.

#### pair0070: PARAMETER → ANSWER

Certainly parameter causes answer. We do not have to talk about *hypothetical* interventions. Instead, we have a true intervention, since “parameter” has been set by the experimenter.

## D22: Acute Inflammations

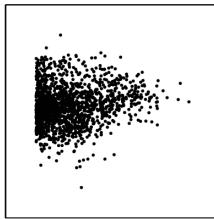
This data set, available at the UCI Machine Learning Repository (Bache and Lichman, 2013), was collected in order to create a computer expert system that decides whether a patient suffers from two different diseases of urinary system (Czerniak and Zarzycki, 2003). We downloaded it from <https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations>. The two possible diseases are acute inflammations of urinary bladder and acute nephritis of renal pelvis origin. As it is also possible to chose none of those, the class variable takes values in  $\{0, 1\}^2$ . The decision is based on six symptoms: temperature of patient (e.g. 35.9), occurrence of nausea (“yes” or “no”), lumbar pain (“yes” or “no”), urine pushing (“yes” or “no”), micturition pains (“yes” or “no”) and burning of urethra, itch, swelling of urethra outlet (“yes” or “no”). These are grouped together in a six-dimensional vector “symptoms”.

**pair0071: SYMPTOMS → CLASSIFICATION OF DISEASE**

One would think that the disease is causing the symptoms but this data set was created artificially. The description on the UCI homepage says: “The data was created by a medical expert as a data set to test the expert system, which will perform the presumptive diagnosis of two diseases of urinary system. (...) Each instance represents an potential patient.” We thus consider the symptoms as the cause for the expert’s decision.

## D23: Sunspot data

The data set consists of 1632 monthly values between May 1874 and April 2010 and therefore contains 1632 data points. The temperature data have been taken from <http://www.cru.uea.ac.uk/cru/data/temperature/> and have been collected by Climatic Research Unit (University of East Anglia) in conjunction with the Hadley Centre (at the UK Met Office) (Morice et al., 2012). The temperature data is expressed in deviations from the 1961–90 mean global temperature of the Earth (i.e., monthly anomalies). The sunspot data (Hathaway, 2010) are taken from the National Aeronautics and Space Administration and were downloaded from <http://solarscience.msfc.nasa.gov/SunspotCycle.shtml>. According to the description on that website, “sunspot number is calculated by first counting the number of sunspot groups and then the number of individual sunspots. The sunspot number is then given by the sum of the number of individual sunspots and ten times the number of groups. Since most sunspot groups have, on average, about ten spots, this formula for counting sunspots gives reliable numbers even when the observing conditions are less than ideal and small spots are hard to see.”



pair0072

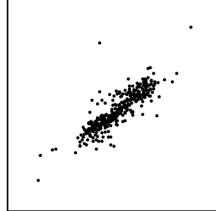
Figure 34: Scatter plots of pairs from D23. pair0072: sunspots → global mean temperature.

**pair0072: SUNSPOTS → GLOBAL MEAN TEMPERATURE**

Sunspots are phenomena that appear temporarily on the sun's surface. Although the causes of sunspots are not entirely understood, there is a significant dependence between the number of sunspots and the global mean temperature anomalies ( $p$ -value for zero correlation is less than  $10^{-4}$ ). There is evidence that the Earth's climate heats and cools as solar activity rises and falls (Haigh, 2007), and the sunspot number can be seen as a proxy for solar activity. Also, we do not believe that the Earth's surface temperature (or changes of the Earth's atmosphere) has an influence on the activity of the sun. We therefore consider number of sunspots causing temperature as the ground truth.

**D24: Food and Agriculture Organization of the UN**

The data set has been collected by Food and Agriculture Organization of the UN (<http://www.fao.org/economic/ess/ess-fs/en/>) and is accessible at <http://www.docstoc.com/docs/102679223/Food-consumption-and-population-growth---FAO>. It covers 174 countries or areas during the period from 1990–92 to 1995–97 and the period from 1995–97 to 2000–02. As one entry is missing, this gives 347 data points. We selected two variables: population growth and food consumption. The first variable indicates the average annual rate of change of population (in %), the second one describes the average annual rate of change of total dietary consumption for total population (kcal/day) (also in %).



pair0076

Figure 35: Scatter plots of pairs from D24. pair0076: population growth → food consumption growth.

**pair0076: POPULATION GROWTH → FOOD CONSUMPTION GROWTH**

We regard population growth to cause food consumption growth, mainly because more people eat more. Both variables are most likely also confounded by the availability of food, driven for instance by advances in agriculture and subsequently increasing yields, but also by national and international conflicts, the global food market and other economic factors. However, for the short time period considered here, confounders which mainly influence the variables on a temporal scale can probably be neglected. Their might also be a causal link from food consumption growth to population growth, for instance one could imagine that if people are well fed, they also reproduce more. However, we assume this link only plays a minor role here.

## D25: Light response data

The filtered version of the light response data was obtained from Moffat (2012). It consists of 721 measurements of Net Ecosystem Productivity (NEP) and three different measures of the Photosynthetic Photon Flux Density (PPFD): the direct, diffuse, and total PPFD. NEP is a measure of the net  $CO_2$  flux between the biosphere and the atmosphere, mainly driven by biotic activity. It is defined as the photosynthetic carbon uptake minus the carbon release by respiration, and depends on the available light. NEP is measured in units of  $\mu\text{mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$ . PPFD measures light intensity in terms of photons that are available for photosynthesis, i.e., with wavelength between 400 nm and 700 nm (visible light). More precisely, PPFD is defined as the number of photons with wavelength of 400–700 nm falling on a certain area per time interval, measured in units of  $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$ . The total PPFD is the sum of PPFD<sub>diffuse</sub>, which measures only diffusive photons, and PPFD<sub>direct</sub>, which measures only direct (solar light) photons. The data was measured over several hectare of a forest in Hainich, Germany (site name DE-Hai, latitude: 51.08°N, longitude: 10.45°E), and is available from <http://fluxnet.ornl.gov>.

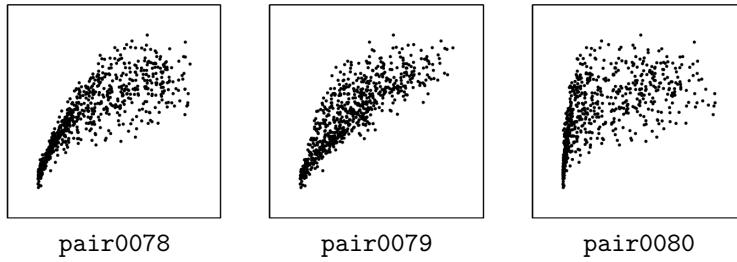


Figure 36: Scatter plots of pairs from D25. pair0078: PPFD → NEP, pair0079: PPFD<sub>diffuse</sub> → NEP, pair0080: PPFD<sub>direct</sub> → NEP.

pair0078–pair0080: {PPFD,PPFD<sub>DIFFUSE</sub>,PPFD<sub>DIRECT</sub>} → NEP

Net Ecosystem Productivity is known to be driven by both the direct and the diffuse Photosynthetic Photon Flux Density, and hence also by their sum, the total PPFD.

## D26: FLUXNET

The data set contains measurements of net  $CO_2$  exchanges between atmosphere and biosphere aggregated over night, and the corresponding temperature. It is taken from the FLUXNET network (Balocchi et al., 2001), available at <http://fluxnet.ornl.gov> (see also Section D25). The data have been collected at a 10 Hz rate and was aggregated to one value per day over one year (365 values) and at three different sites (BE-Bra, DE-Har, US-PFa).  $CO_2$  exchange measurements typically have a footprint of about 1km<sup>2</sup>. The data set contains further information on the quality of the data (“1” means that the value is credible, “NaN” means that the data point has been filled in).

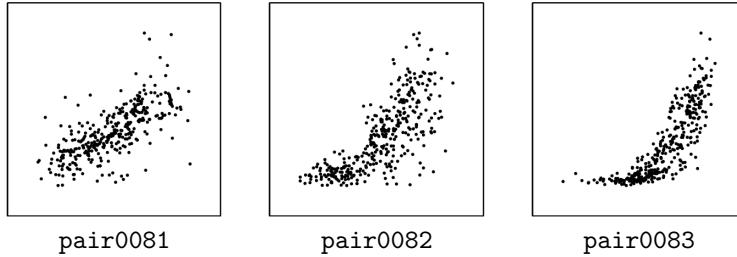


Figure 37: Scatter plots of pairs from D26. **pair0081** (BE-Bra): temperature → local  $CO_2$  flux, **pair0082** (DE-Har): temperature → local  $CO_2$  flux, **pair0083** (US-PFa): temperature → local  $CO_2$  flux.

#### **pair0081–pair0083: TEMPERATURE → LOCAL $CO_2$ FLUX**

Because of lack of sunlight,  $CO_2$  exchange at night approximates ecosystem respiration (carbon release from the biosphere to the atmosphere), which is largely dependent on temperature (e.g. Mahecha et al., 2010). The  $CO_2$  flux is mostly generated by microbial decomposition in soils and maintenance respiration from plants and does not have a direct effect on temperature. We thus consider temperature causing  $CO_2$  flux as the ground truth. The three pairs **pair0081–pair0083** correspond with sites BE-Bra, DE-Har, US-PFa, respectively.

#### **D27: US county-level growth data**

The data set (Wheeler, 2003) contains both employment and population information for 3102 counties in the US in 1980. We downloaded the data from <http://www.spatial-econometrics.com/data/contents.html>. We selected columns eight and nine in the file “countyg.dat”. Column eight contains the natural logarithm of the number of employed people, while column nine contains the natural logarithm of the total number of people living in this county, and is therefore always larger than the number in column eight.

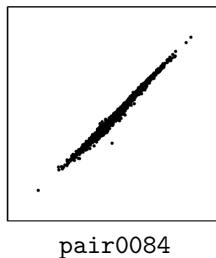


Figure 38: Scatter plots of pairs from D27. **pair0084**: population → employment.

#### **pair0084: POPULATION → EMPLOYMENT**

It seems reasonable that the total population causes the employment and not vice versa. If we increase the number of people living in an area, this has a direct effect on the number of employed people. We believe that the decision to move into an economically strong area is

rather based on the employment rate rather than the absolute number of employed people. There might be an effect that the employment status influences the decision to get children but we regard this effect to be less relevant.

### D28: Milk protein trial

This data set is extracted from that for the milk protein trial used by [Verbyla and Cullis \(1990\)](#). The original data set consists of assayed protein content of milk samples taken weekly from each of 79 cows. The cows were randomly allocated to one of three diets: barley, mixed barley-lupins, and lupins, with 25, 27 and 27 cows in the three groups, respectively. Measurements were taken for up to 19 weeks but there were 38 drop-outs from week 15 onwards, corresponding to cows who stopped producing milk before the end of the experiment. We removed the missing values (drop-outs) in the data set: we did not consider the measurements from week 15 onwards, which contain many drop-outs, and we discarded the cows with drop-outs before week 15. Finally, the data set contains 71 cows and 14 weeks, i.e., 994 samples in total. Furthermore, we re-organized the data set to see the relationship between the milk protein and the time to take the measurement. We selected two variables: the time to take weekly measurements (from 1 to 14), and the protein content of the milk produced by each cow at that time.

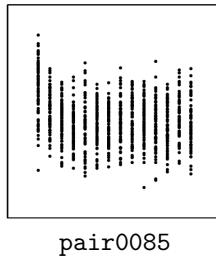


Figure 39: Scatter plots of pairs from D28. pair0085: time of measurement → protein content of milk.

#### pair0085: TIME OF MEASUREMENT → PROTEIN CONTENT OF MILK

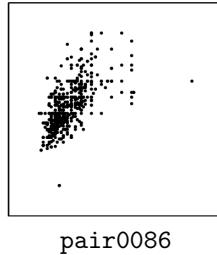
Clearly, the time of the measurement causes the protein content and not vice versa. We do not consider the effect of the diets on the protein content.

### D29: kamernet.nl data

This data was collected by Joris M. Mooij from <http://www.kamernet.nl>, a Dutch website for matching supply and demand of rooms and appartments for students, in 2007. The variables of interest are the size of the apartment or room (in  $\text{m}^2$ ) and the monthly rent in EUR. Two outliers (one with size  $0 \text{ m}^2$ , the other with rent of 1 EUR per month) were removed, after which 666 samples remained.

#### pair0086: SIZE OF APARTMENT → MONTHLY RENT

Obviously, the size causes the rent, and not vice versa.

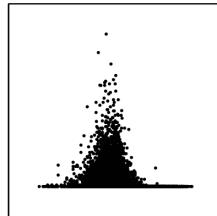


pair0086

Figure 40: Scatter plots of pairs from D29. pair0086: size of apartment → monthly rent.

**D30: Whistler daily snowfall**

The Whistler daily snowfall data is one of the data sets on <http://www.mldata.org>, and was originally obtained from <http://www.climate.weatheroffice.ec.gc.ca/> (Whistler Roundhouse station, identifier 1108906). We downloaded it from <http://www.mldata.org/repository/data/viewslug/whistler-daily-snowfall>. It concerns historical daily snowfall data in Whistler, BC, Canada, over the period July 1, 1972 to December 31, 2009. It was measured at the top of the Whistler Gondola (Latitude:  $50^{\circ}04'04.000''$  N, Longitude:  $122^{\circ}56'50.000''$  W, Elevation: 1835 m). We selected two attributes, mean temperature ( $^{\circ}\text{C}$ ) and total snow (cm). The data consists of 7753 measurements of these two attributes.



pair0087

Figure 41: Scatter plots of pairs from D30. pair0087: temperature → total snow.

**pair0087: TEMPERATURE → TOTAL SNOW**

Common sense tells us that the mean temperature is one of the causes of the total amount of snow, although there may be a small feedback effect of the amount of snow on temperature. Confounders are expected to be present (e.g., whether there are clouds).

**D31: Bone Mineral Density**

This dataset comes from the R package `ElemStatLearn`, and contains measurements of the age and the relative change of the bone mineral density of 261 adolescents. Each value is the difference in the spinal bone mineral density taken on two consecutive visits, divided by the average. The age is the average age over the two visits. We preprocessed the data by taking only the first measurement for each adolescent, as each adolescent has 1–3 measurements.

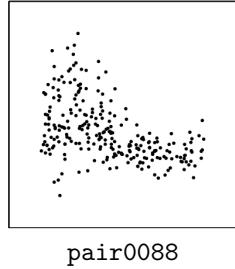


Figure 42: Scatter plots of pairs from D31. pair0088: age → relative bone mineral density.

pair0088: AGE → BONE MINERAL DENSITY

Age must be the cause, bone mineral density the effect.

## Acknowledgements

JMM was supported by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410). JP received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no 326496.

## References

- R. Armann and I. Bülthoff. Male and female faces are only perceived categorically when linked to familiar identities – and when in doubt, he is a male. *Vision Research*, 63:69–80, 2012.
- A. Azzalini and A. W. Bowman. A look at some data on the Old Faithful Geyser. *Applied Statistics*, 39(3):357–365, 1990.
- K. Bache and M. Lichman. UCI Machine Learning Repository, 2013. URL <http://archive.ics.uci.edu/ml>.
- D. Baldocchi, E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. Lee, Y. Malhi, T. Meyers, W. Munger, W. Oechel, K. T. Paw, K. Pelegaard, H. P. Schmid, R. Valentini, S. Verma, T. Vesala, K. Wilson, and S. Wofsy. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, 82(11):2415–2434, 2001.
- J. Baynes and M. H. Dominiczak. *Medical Biochemistry*. Mosby, 1999.
- J. Bloomer, J. W. Stehr, C. A. Piety, R. J. Salawitch, and R. R. Dickerson. Observed relationships of ozone air pollution with temperature and emissions. *Geophysical Letters*, 36(9), 2009.
- K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

- E. Braunwald, A. S. Fauci, D. L. Kasper, S. L. Hauser, D. L. Long, and J. L. Jameson, editors. *Principles of Internal Medicine: Volume 2*. McGraw-Hill, 15th international edition, 2001.
- P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics (to appear), ArXiv e-prints (1207.5136)*, 2014.
- C. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.
- J. Czerniak and H. Zarzycki. Application of rough sets in the presumptive diagnosis of urinary system diseases. In J. Soldek and L. Drobizgiewicz, editors, *Artificial Intelligence and Security in Computing Systems*, pages 41–51. Kluwer Academic Publishers, Norwell, MA, USA, 2003.
- P. Daniušis, D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 143–150, 2010.
- D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 107–114, 2007.
- F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.
- N. Ebrahimi, K. Pflughoeft, and E. S. Soofi. Two measures of sample entropy. *Statistics and Probability Letters*, 20:225–234, 1994.
- S. A. Esrey, J. B. Potash, L. Roberts, and C. Schiff. Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis, and trachoma. *Bulletin of the World Health Organization*, 69(5):609, 1991.
- U. Feister and K. Balzer. Surface ozone and meteorological predictors on a subregional scale. *Atmospheric Environment. Part A. General Topics*, 25(9):1781–1790, 1991.
- M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, March 2002.
- N. Friedman and I. Nachman. Gaussian process networks. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pages 211–219, 2000.
- K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20 (NIPS\*2007)*, pages 489–496, 2008.

- R. M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2:155–239, 2006.
- U. Grenander and G. Szego. *Toeplitz forms and their applications*. University of California Press, 1958.
- A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pages 63–78. Springer-Verlag, 2005.
- A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 585–592, 2008.
- P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.
- H. A. Guvenir, B. Acar, G. Demiroz, and A. Cekin. A supervised machine learning algorithm for arrhythmia analysis. In *Proceedings of the Computers in Cardiology Conference*, Lund, Sweden, 1997.
- I. Guyon, D. Janzing, and B. Schölkopf. Causality: Objectives and assessment. In *JMLR Workshop and Conference Proceedings*, volume 6, pages 1–38, 2010.
- I. Guyon et al. Results and analysis of 2013-2014 ChaLearn Cause-Effect Pair Challenges. In preparation, 2014.
- L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.
- J. D. Haigh. The sun and the earths climate. *Living Reviews in Solar Physics*, 4(2):2298, 2007.
- D. H. Hathaway. The solar cycle. *Living Reviews in Solar Physics*, 7:1, 2010.
- K. W. Hipel and A. I. McLeod. *Time series modelling of water resources and environmental systems*. Elsevier, 1994.
- P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS\*2008)*, pages 689–696, 2009.
- A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems (NIPS\*1996)*, pages 273–279, 1997.
- A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12:429–439, 1999.
- A. Hyvärinen and S. M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14:111–152, 2013.

- B. Janzing. *Sonne, Wind und Schneerekorde: Wetter und Klima in Furtwangen im Schwarzwald, zum 25-jährigen Bestehen der Wetterstation.* self-published, in German, 2004.
- D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.
- D. Janzing, P. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *Proceedings of the International Conference on Machine Learning (ICML 2010)*, pages 479–486, 2010.
- D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182–183:1–31, 2012.
- E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3):437–471, 1996.
- 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.
- L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 23:9–16, 1987.
- S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf. Consistency of causal inference under the additive noise model. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 478–486, 2014.
- A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69:066138, 2004.
- L. Lee, M. R. Rosenzweig, and M. M. Pitt. The effects of improved nutrition, sanitation, and water quality on child health in high-mortality populations. *Journal of Econometrics*, 77(1):209–235, 1997.
- M. D. Mahecha, M. Reichstein, N. Carvalhais, G. Lasslop, H. Lange, S. I. Seneviratne, R. Vargas, C. Ammann, M. A. Arain, A. Cescatti, I. A. Janssens, M. Migliavacca, L. Montagnani, and A. D. Richardson. Global convergence in the temperature sensitivity of respiration at ecosystem level. *Science*, 329(5993):838–840, 2010.
- M. Meyer and P. Vlachos. Statlib: Data, software and news from the statistics community, 2014. URL <http://lib.stat.cmu.edu/>.
- A. M. Moffat. *A new methodology to interpret high resolution measurements of net carbon fluxes between terrestrial ecosystems and the atmosphere.* PhD thesis, Friedrich Schiller University, Jena, 2012.

- J. M. Mooij and T. Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 431–439, 2013.
- J. M. Mooij and D. Janzing. Distinguishing between cause and effect. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 6, pages 147–156, 2010.
- J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, pages 745–52, 2009.
- J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems 23 (NIPS\*2010)*, pages 1687–1695, 2010.
- J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24 (NIPS\*2011)*, pages 639–647, 2011.
- J. M. Mooij, D. Janzing, J. Zscheischler, and B. Schölkopf. CauseEffectPairs repository, 2014. URL <http://webdav.tuebingen.mpg.de/cause-effect/>.
- C. P. Morice, J. J. Kennedy, N. A. Rayner, and P. D. Jones. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The hadcrut4 data set. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 117 (D8), 2012.
- W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait. Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288), 1994.
- H. A. Noughabi and R. A. Noughabi. On the entropy estimators. *Journal of Statistical Computation and Simulation*, 83:784–792, 2013.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228, 2014.
- 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: 2436–2450, 2011.
- J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.
- J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

- D. Ramirez, J. Via, I. Santamaría, and P. Crespo. Entropy and Kullback-Leibler divergence estimation based on Szego’s theorem. In *European Signal Processing Conference (EUSIPCO)*, pages 2470–2474, 2009.
- C. E. Rasmussen and H. Nickisch. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.
- C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- D. J. Rasmussen, A. M. Fiore, V. Naik, L. W. Horowitz, S. J. McGinnis, and M. G. Schlutz. Surface ozone-temperature relationships in the eastern us: A monthly climatology for evaluating chemistry-climate models. *Atmospheric Environment*, 47:142–153, 2012.
- T. Richardson and P. Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):9621030, 2002.
- B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
- B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On Causal and Anticausal Learning. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262, New York, NY, USA, 2012. Omnipress.
- S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear Non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.
- S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.
- L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, May 2012.
- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.
- B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.
- A. Statnikov, M. Henaff, N. I. Lytkin, and C. F. Aliferis. New methods for separating causes from effects in genomics data. *BMC Genomics*, 13:S22, 2012.
- W. R. Stockwell, G. Kramm, H.-E. Scheel, V. A. Mohnen, and W. Seiler. *Ozone formation, destruction and exposure in Europe and the United States*, pages 1–38. Springer, 1997.
- D. Stowell and M. D. Plumley. Fast multidimensional entropy estimation by k-d partitioning. *IEEE Signal Processing Letters*, 16:537–540, 2009.

- D. Stoyan, H. Stoyan, and U. Jansen. *Umwelstatistik: Statistische Verarbeitung und Analyse von Umweltdaten*. Springer, 1997.
- 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, 2006.
- X. Sun, D. Janzing, and B. Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, 71:1248–1256, 2008.
- Z. Szabó. Information theoretical estimators toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014.
- O. Tange. GNU Parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb 2011. URL <http://www.gnu.org/s/parallel>.
- H. Tønnesen, L. Hejberg, S. Frobenius, and J. Andersen. Erythrocyte mean cell volume–correlation to drinking pattern in heavy alcoholics. *Acta medica Scandinavica*, 219:515–518, 1986.
- U.S. Department of Commerce. Website of the U.S. Census Bureau, 1994. URL <http://www.census.gov/>.
- B. van Es. Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, 19:61–72, 1992.
- M. van Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17:1903–1910, 2005.
- O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, 38:54–59, 1976.
- W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 4th edition, 2002.
- A. P. Verbyla and B. R. Cullis. Modelling in repeated measures experiments. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 39(3):341–356, 1990.
- L. Wasserman. *All of Statistics*. Springer, New York, 2004.
- C. H. Wheeler. Evidence on agglomeration economies, diseconomies, and growth. *Journal of Applied Econometrics*, 18(1):79–104, 2003.
- S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.
- I.-C. Yeh. Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, 28:1797–1808, 1998.
- K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655, 2009.

- J. Zscheischler, D. Janzing, and K. Zhang. Testing whether linear equations are causal: A free probability theory approach. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 839–847, 2011.