Information-geometric approach to inferring causal directions

Dominik Janzing[a], Joris Mooij[b], Kun Zhang[a], Jan Lemeire[c,d], Jakob Zscheischler[a], Povilas Daniušis[e], Bastian Steudel[f], Bernhard Schölkopf[a]

[a] *Max Planck Institute for Intelligent Systems, Tübingen, Germany*
[b] *Radboud University, Nijmegen, Netherlands*
[c] *Vrije Universiteit Brussel, Brussels, Belgium*
[d] *Interdisciplinary Institute for Broadband Technology, Ghent, Belgium*
[e] *Vilnius University, Lithuania*
[f] *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*

Abstract

While conventional approaches to causal inference are mainly based on conditional (in)dependences, recent methods also account for the shape of (conditional) distributions. The idea is that the causal hypothesis "X causes Y" imposes that the marginal distribution P_X and the conditional distribution $P_{Y|X}$ represent *independent mechanisms* of nature. Recently it has been postulated that the shortest description of the joint distribution $P_{X,Y}$ should therefore be given by separate descriptions of P_X and $P_{Y|X}$. Since description length in the sense of Kolmogorov complexity is uncomputable, practical implementations rely on other notions of independence. Here we define independence via orthogonality in information space. This way, we can explicitly describe the kind of dependence that occur between P_Y and $P_{X|Y}$ making the causal hypothesis "Y causes X" implausible. Remarkably, this asymmetry between cause and effect becomes particularly simple if X and Y are deterministically related. We present an inference method that works in this case. We also discuss some theoretical results for the non-deterministic case although it is not clear how to employ them for a more general inference method.

Keywords:
deterministic causal relations, pythagorean triple, cause-effect pairs

1. Introduction

The problem of inferring whether X causes Y (write $X \to Y$) or Y causes X from observations $(x_1, y_1), \ldots, (x_m, y_m)$ that are i.i.d. drawn from $P_{X,Y}$ is a particularly challenging task for causal inference [1]. Although this restricted problem ignores other important problems of causal inference (i.e., unobserved common causes or bidirectional influence), it is useful for studying statistical asymmetries between cause and effect. Conventional methods for causal inference [2, 3] focus on conditional independences and thus require observations from at least three variables.

Extending an idea in [4], [5] postulates that $X \to Y$ is only acceptable as causal hypothesis if the shortest description of $P_{X,Y}$ is given by separate descriptions of $P_{Y|X}$ and P_X. Here description length is understood in the sense of algorithmic information ("Kolmogorov complexity") [6, 7, 8]. Note that the postulate is equivalent to saying that $P_{Y|X}$ and P_X are *algorithmically independent* in the sense that knowing P_X does not enable a shorter description of $P_{Y|X}$ and vice versa. To show that this helps in distinguishing between cause and effect for just two observed variables, [5] constructed toy models of causal mechanisms where the causal structure $X \to Y$ yields algorithmic dependences between $P_{X|Y}$ and P_Y. Even though *algorithmic* independence between P_{cause} and $P_{\text{effect}|\text{cause}}$ is an appealing formalization of independence, practical methods must be based on computable criteria.

[9] described a potential asymmetry between cause and effect where independence is meant in terms of *statistical* independence between the cause and the noise term that occurs in the causal mechanism: If Y is a function of X up to an additive noise term that is statistically independent of X, i.e.,

$$Y = f(X) + E \quad \text{with} \quad E \perp\!\!\!\perp X\,, \tag{1}$$

then there is usually (up to some exceptions like the bivariate Gaussian) no such additive noise model from Y to X. In other words, writing X as $X = g(Y) + \tilde{E}$ with some function g will not render the residual term $\tilde{E}$ statistically independent of Y. [10] generalizes the model class to

$$Y = h(f(X) + E) \quad \text{with} \quad E \perp\!\!\!\perp X\,, \tag{2}$$

and show that such a "post-nonlinear (PNL) model" also exists in at most one direction, except for some special cases. If $P_{X,Y}$ is consistent with (1) or (2), respectively, in one direction but not the other, one infers that direction to be the causal one implied by the corresponding model. For the model (1) it has been shown [11] that this kind of reasoning is justified by the above algorithmic independence principle.

Note that these inference methods do not assume that causal relations are always of the above form. They only decide for one of the causal directions *if* one and only one direction admits such a model. The idea is the following: if $X \to Y$ is the correct model, but not of the additive noise form, it is rather unlikely that it generates a joint distribution that admits an additive noise model in the opposite direction. The reason is that this would require rather contrived adjustments between P_X (the marginal distribution of the hypothetical cause) and $P_{Y|X}$ (the conditional distribution of the effect, given the cause) [11]. This article develops an information-geometric principle that does not require the restricted class of additive noise or post-nonlinear models. To this end, we revisit additive noise models in Section 2 and show that entropies can play a key role in describing the kind of dependences between $P_{X|Y}$ and P_Y that can occur if X causes Y. This motivates our information geometric perspective developed in Section 3, which results in an inference method for deterministic causal relations in 4, with an outlook for the non-deterministic case in Appendix Appendix A. The table in Fig. 1 shows how the main results are structured.

Readers who are only interested in our inference method may focus on Section 4, with Subsection 4.3 and 4.4 as its main parts. The other sections provide a general background and describe a large class of asymmetries between cause and effect that could be helpful for developing other information-theoretic methods in the future.

2. Information theoretic view on additive noise models

We consider the additive noise model (1) in the low noise regime (see Fig. 2) and show how the relationship between the input distribution and the conditional one is different for both directions. We use the following notational conventions. $P_{Y|x}$ is the distribution of Y, given a fixed value x while $P_{Y|X}$ denotes the entire conditional distribution. The range of a random variable X will be denoted by D_X. $S(P_{Y|x})$ denotes the (differential) Shannon entropy of $P_{Y|x}$ for fixed x. The function $x \mapsto S(P_{Y|x})$ will also be called the conditional entropy function. Throughout the paper we will assume that all distributions have densities with respect to a fixed reference measure (e.g., the Lebesgue measure for real-valued variables or the counting

Section	**Contents**	**Main Reference**
Section 3	Postulating independence conditions (h_1) - (h_3) for P_{cause} and $P_{\text{effect}\vert\text{cause}}$	Postulate 1 and Definition 1
	Justifying the conditions	Lemmas 1,2
	Rephrasing (h_1)-(h_3) as orthogonality	Theorem 1
Section 4	Implications of (h_3) for deterministic causality	Theorem 2
	Generalizing (h_3) via exponential families	Postulate 2
	Inference method for deterministic case based on the generalized condition (h_3)	Subsections 4.3 and 4.4
Appendix Appendix A	Outlook: employing orthogonality for inferring non-deterministic relations (toy examples, negative results)	Lemma 9 and 10

Figure 1: Structure of the main results

measure for discrete variables). This measure will never appear explicitly and should not be confused with reference probability distributions that occur all over the article. By slightly overloading notation, P_X will stand for both the distribution and the density $x \mapsto P_X(x)$. We will also write $P(x)$ instead of $P_X(x)$ whenever this causes no confusion. For discrete variables X, integrals of the form $\int \cdots P(x)dx$ will be understood as sums by interpreting dx as $d\mu(x)$ where μ denotes the counting measure.

Regarding (1) we observe that $E \perp\!\!\!\perp X$ ensures that the conditional entropy function $S(P_{Y|x})$ is constant in x and coincides with the conditional entropy $S(P_{Y|X})$ (defined by the average $\int S(P_{Y|x})P(x)dx$). In studying how P_Y and $P_{X|Y}$ are then related we first assume that P_X is uniform. Then, $P(y) \approx P_X(f^{-1}(y)) \cdot {f^{-1}}'(y)$ is large for those y-values where $|{f^{-1}}'(y)|$ is large. At the same time, the entropy $S(P_{X|y})$ is large for y-values in regions with large $|{f^{-1}}'(y)|$ (see Fig. 2). Hence, large entropy $S(P_{X|y})$ correlates with high density $P(y)$, assuming that $P(x)$ is constant on the interval under consideration. If P_X is not the uniform distribution, high values of $P(y)$ occur at points where both $|{f^{-1}}'(y)|$ and $P_X(f^{-1}(y))$ are high. We argue later that if the peaks of $P(x)$ do not correlate with the slope of f then the qualitative argument above still holds and $S(P_{X|y})$ again correlates with $P(y)$. This reasoning will be formalized in Section 3.

The first information geometric inference principle that we are going to state in the next section no longer assumes that the entropy $S(P_{Y|x})$ is constant in x if $X \to Y$ is the true causal direction. Instead, it postulates that regions of large $S(P_{Y|x})$ *do not correlate* with regions of large density $P(x)$. The example above shows that dependences between P_Y and $P_{X|Y}$ occurring for the wrong causal direction can appear on the level of correlations between information theoretic expressions (like conditional entropy) computed from the conditional $P_{X|y}$ and the density $P(y)$. We will show that correlations of this type can be phrased as an orthogonality relation in the sense of information geometry.

3. A class of testable independence relations

The intention of this section is to postulate independence conditions between $P_{Y|X}$ and P_X that can be tested empirically. We will describe several options to solve this task.

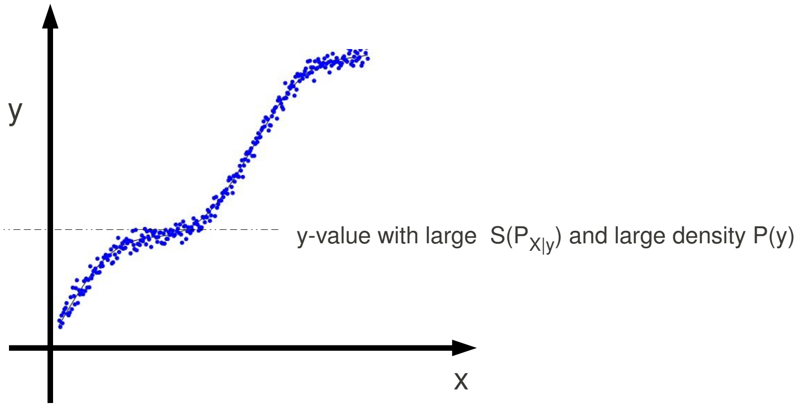

Figure 2: Functional relation with small noise. The conditional entropy function $S(P_{X|y})$ is high at regions with high slope of $f^{-1}(y)$, i.e., small slope of f at this point.

3.1. General structure of our independence relations

The following postulate describes the general structure that all our postulates share:

Postulate 1 (general structure of independence).
Assume that X causes Y. Let $x \mapsto h(x) \in \mathbb{R}$ be any function for which $h(x)$ describes local properties of the conditional $P_{Y|X}$ at the point $X = x$.[1] *Then the "structure function" h and P_X are likely to satisfy*

$$\int h(x)P(x)dx \approx \int h(x)U_X(x)dx\,, \tag{3}$$

where U_X is a reference density for X (not necessarily uniform).

Note that the difference between both sides of (3) can be rephrased as a covariance if we formally consider h and P_X/U_X as functions of a random

[1] Note that we have avoided the more concise formulation "$h(x)$ describes properties of the conditional $P_{Y|x}$" for the following reason: For deterministic relations $Y = f(X)$, the function $h(x) := f'(x)$ expresses a property of $P_{Y|X}$ that is local at $X = x$, but $h(x)$ cannot be derived from $P_{Y|x}$ alone.

variable X with distribution U_X:

$$\begin{aligned}
& \int h(x)P(x)dx - \int h(x)U(x)dx \qquad (4)\\
= \quad & \int h(x)\frac{P(x)}{U(x)}U(x)dx - \int h(x)U(x)dx \int \frac{P(x)}{U(x)}U(x)dx \\
=: \quad & \mathrm{Cov}_{U_X}\left(h, \frac{P_X}{U_X}\right) .
\end{aligned}$$

Therefore (3) formalizes uncorrelatedness between the functions h and P_X/U_X, which is justified by the idea that the way P_X differs from U_X is independent of h.

The postulate remains vague regarding how to choose h and U_X. We will later discuss different reasonable choices, for instance $h(x) := S(P_{Y|x})$ (for non-deterministic relations), $h(x) := f'(x)$ (for deterministic ones) and also $h(x) := \log f'(x)$ (for deterministic monotonically increasing relations). We recommend to choose "non-informative" distributions like uniform ones or Gaussians for U_X. If we assume that "typical" choices of P_X (as long as the choice is independent of h) yield almost the same integral, we also have to assume that changing U_X to some U'_X does not matter too much as long as we have chosen U'_X independently of h. This suggests some robustness under changing the reference measure.

3.2. Probabilistic models as justification

Even after specifying the reference density U_X and the map from the conditional $P_{Y|X}$ to its structure function h, a mathematical justification of (3) can only be given within a probabilistic model about how "nature chooses P_X" or how it "chooses $P_{Y|X}$". To show this, we now consider a random process that generates functions h (which can equivalently be seen as generating random conditionals $P_{Y|X}$):

Lemma 1 (interval-wise random generation of $P_{Y|X}$).
Let X, Y be real-valued. Let $r_j > 0$ with $j \in \mathbb{Z}$ be random numbers iid drawn from a distribution $Q(r)$ with standard deviation σ_r. We then define a piecewise constant function h via $h(x) := r_j$ for $x \in [j, j+1)$ (Fig. 3 shows two options how h may correspond to a conditional $P_{Y|X}$). We then have for every $c > 0$,

$$\left| \int h(x)P(x)dx - \int h(x)U(x)dx \right| \leq c\,\sigma_r \sqrt{\sum_j \left(\int_j^{j+1} P(x) - U(x)\, dx \right)^2} ,$$

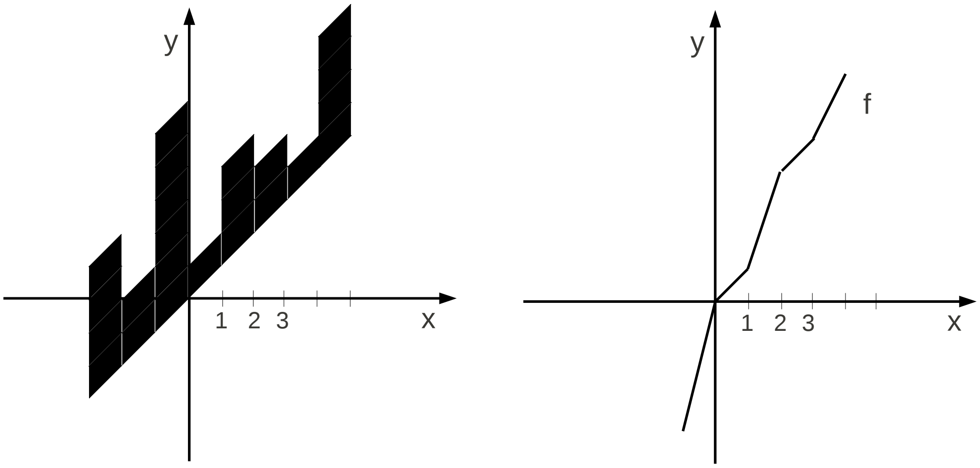

Figure 3: Visualization of two options for interval-wise generation of a conditional $P_{Y|X}$ via dice throws r_j. Left: $P_{Y|X}$ corresponding to $Y = X + E$ where the distribution of $E|_x$ is the uniform distribution on the interval $[0, h(x)]$. Right: the dice determines the slope of f for $Y = f(X)$ via $f'(x) := h(x)$.

with probability $1 - 1/c^2$ *or higher.*

Proof: This is because

$$\int h(x)(P(x) - U(x))dx = \sum_j r_j \left(\int_j^{j+1} P(x) - U(x)\, dx \right)$$

is the sum of independent random variables, each having variance

$$\sigma_r^2 \left(\int_j^{j+1} P(x) - U(x)\, dx \right)^2 .$$

Then the statement follows from Chebyshev's inequality, noting that the expected value of $\int h(x)\big(P(x) - U(x)\big)\, dx$ vanishes. $\Box$.

The example is instructive because it shows that (3) is likely to hold regardless of P_X and U_X provided that the following conditions are satisfied: First, both distributions P_X and U_X have been chosen independently and independently of $P_{Y|X}$. Second, both distributions are sufficiently spread out such that $\beta := \sum_j (\int_j^{j+1} P(x) - U(x)\, dx)^2$ is small. Roughly speaking, if P_X and U_X have width n, then $\beta \in O(1/n)$ and hence (3) holds up to an error in $O(1/\sqrt{n})$. Neglecting one of these conditions, one can easily construct counter examples: First, if one of the distributions P_X or U_X, say U_X, is constructed *after* having seen all the r_j, U_X can be constructed such that $P_X - U_X$ is positive for all intervals $[j, j+1)$ where r_j is large and negative for small r_j. This results in $\int h(x)U(x)dx$ being systematically greater than $\int h(x)U(x)dx$. Second, if one of the distributions, say U_X, is supported by one interval $[j, j+1)$ only, we have $\int h(x)U(x)dx = r_j$, i.e, the right hand side of (3) depends on a single r_j only and can therefore strongly deviate from $\int h(x)P(x)dx$.

One can also construct a probabilistic model where $P_{Y|X}$ and thus h is fixed and instead P_X is generated randomly, for instance by the following procedure. On each interval $[j, j+1)$ multiply $U(x)$ by some random number r_j. Then renormalize the obtained function to obtain $P(x)$. If U_X is spread out over many intervals, (3) holds with high probability. We have skipped the detailed analysis of this example because it becomes too technical.

The following model assumes that P_X is chosen from a prior that is invariant under a group action:

Lemma 2 (symmetric prior).
G be a group acting on the domain of X and $\mathcal{P}(P_X)$ be a probability density on the set of distributions P_X that is G-invariant, i.e.,

$$\mathcal{P}(P_X) = \mathcal{P}(\overline{P}_X)\,,$$

where $\overline{P}_X$ denotes the average of P_X over the action of G.

Then, for any fixed h we have

$$\mathbb{E}_{\mathcal{P}} \int h(x)P_X(x)dx = \mathbb{E}_{\mathcal{P}} \int h(x)\overline{P}_X(x)dx\,,$$

where $\mathbb{E}_{\mathcal{P}}$ denotes the expectation over the prior $\mathcal{P}$.

The result follows immediately from linearity of the expectation. It suggests to choose $\overline{P}_X$ as reference measure whenever one believes that a G-invariant prior is appropriate. The fact that then the expectations of both sides of (3) coincide does not necessarily guarantee that $\int h(x)P_X(x)dx$ and $\int h(x)\overline{P}_X(x)dx$ are close with high probability. However, for "sufficiently large" groups, this follows indeed from concentration-of-measure results (see [12] and [13] for a similar statement with rotations in high-dimensional spaces). To elaborate on this for general groups would go beyond the scope of this paper.

We have seen that the degree to which we can trust (3) heavily relies on the particular probabilistic models for generating P_X and $P_{Y|X}$. Therefore, we cannot provide any confidence levels that would be valid without referring to one of the models above. After deciding, for instance, that the example in Lemma 1 is a good model for the generation of $P_{Y|X}$ we still need to estimate the size of the intervals that correspond to independent random experiments. Then, we only believe in (3) if the interval sizes are sufficiently small compared to the width of P_X and U_X. Example 2 in Section 4 shows, in the context of deterministic relations, that violation of (3) can easily happen for very simple $P_{Y|X}$ and P_X if P_X and U_X differ in large regions.

We also want to mention that Postulate 1 may fail due to "intelligent design" of P_X and $P_{Y|X}$. This is a fundamental limitation not only of our approach, but also of well-known postulates for causal inference like causal faithfulness [2].

3.3. Independence as orthogonality in information space

Our structure functions will be relative-entropy-like expressions because these turned out to be helpful for formalizing asymmetries between cause and effect. We introduce this terminology now. For two densities P, Q for which P is absolutely continuous with respect to Q, the relative entropy (or KL-distance) is defined by

$$D(P\,||\,Q) := \int \log \frac{P(w)}{Q(w)} P(w) dw \geq 0\,.$$

We then define:

Definition 1 (structure functions for the conditional).
Let U_X and U_Y be reference densities for X and Y, respectively and

$$\overrightarrow{P}_Y := \int P(y|x) U(x) dx$$

denote the output distribution obtained by feeding the conditional with the reference input U_X. Similarly, we will later use

$$\overleftarrow{P}(x) := \int P(x|y) U(y) dy\,.$$

Then we define the following "structure functions":

$$\begin{aligned}
h_1(x) &:= \int \log \frac{P(y|x)}{U(y)} P(y|x) dy = D(P_{Y|x} \,||\, U_Y) \\
h_2(x) &:= \int \log \frac{P(y|x)}{\overrightarrow{P}(y)} P(y|x) dy = D(P_{Y|x} \,||\, \overrightarrow{P}_Y) \\
h_3(x) &:= \int \log \frac{\overrightarrow{P}(y)}{U(y)} P(y|x) dy = h_1(x) - h_2(x)\,.
\end{aligned}$$

The reason that we list all three of these functions though the third one can be represented in terms of the other two is because they all yield conditions that have an interpretation in terms of information geometry, relying on the following concept. Three densities (P, R, Q) are said to form a pythagorean triple of distributions if

$$D(P||Q) = D(P||R) + D(R||Q)\,. \tag{5}$$

This terminology is motivated by interpreting relative entropy as a squared distance and the triple thus satisfies the Pythagorean theorem. If condition (5) holds we say that the vector connecting P with R is orthogonal to the one connecting R with Q are orthogonal but keep in mind that this relation is neither symmetric with respect to exchanging the vectors with each other, nor with respect to reversing the arrows.

We will also use the following formulation:

Lemma 3 (orthogonality in information space).
Orthogonality (5) is equivalent to

$$\int \log \frac{R(w)}{Q(w)} P(w) dw = \int \log \frac{R(w)}{Q(w)} R(w) dw\,. \tag{6}$$

The proof is given by straightforward computation. In analogy to our interpretation of (3), we can interpret (6) as "the integral over the log term does not depend on whether it is weighted with P or R". We then find:

Theorem 1 (three orthogonality conditions).
The conditions $\mathrm{Cov}_{U_X}(h_i, P_X/U_X) = 0$ *for* $i = 1, 2, 3$ *are equivalent to*

$$
\begin{aligned}
D(P_{Y,X} \,||\, U_X U_Y) &\overset{h_1}{=} D(P_{Y,X} \,||\, U_X P_{Y|X}) + D(U_X P_{Y|X} \,||\, U_X U_Y) \\
D(P_{Y,X} \,||\, U_X \overrightarrow{P}_Y) &\overset{h_2}{=} D(P_{Y,X} \,||\, U_X P_{Y|X}) + D(U_X P_{Y|X} \,||\, U_X \overrightarrow{P}_Y) \\
D(P_Y \,||\, U_Y) &\overset{h_3}{=} D(P_Y \,||\, \overrightarrow{P}_Y) + D(\overrightarrow{P}_Y \,||U_Y)\,.
\end{aligned}
$$

Proof: Using Lemma 3, the cases h_1 and h_2 are straightforward computations. For case h_3 note that

$$
\int \log \frac{\overrightarrow{P}(y)}{U(y)} P(y|x)P(x)dxdy = \int \log \frac{\overrightarrow{P}(y)}{U(y)} P(y)dy
$$

and

$$
\int \log \frac{\overrightarrow{P}(y)}{U(y)} P(y|x)U(x)dxdy = \int \log \frac{\overrightarrow{P}(y)}{U(y)} \overrightarrow{P}(y)dy\,.
$$

□

To geometrically justify the orthogonality assumption for h_1, we consider the space V of functions of x, y and identify each distribution $Q_{X,Y}$ with the point

$$
((x, y) \mapsto \log Q(x, y)) \in V\,.
$$

Then we observe that the difference vector connecting the points $P_{Y,X}$ and $U_X P_{Y|X}$ only depends on P_X (in the sense that the common term $P_{Y|X}$ cancels when taking the difference between the two points), while the vector pointing from $U_X P_{Y|X}$ to $U_X U_Y$ only depends on $P_{Y|X}$. In high-dimensional spaces it is likely that two vectors are close to orthogonal if they are chosen independently according to a uniform prior. Even though we do not know of any precise statement of this form with respect to information geometric orthogonality, we accept this as another leading intuition on top of the interpretation of "uncorrelatedness" given by Theorem 1. Regarding h_2, we can argue in a similar way. The fact that both joint distributions occurring in the points $U_X P_{Y|X}$ and $U_X \overrightarrow{P}_Y$ do not contain P_X at all, makes it plausible that the vector should be orthogonal to any vector that only depends on P_X. How to geometrically interpret the orthogonality given by h_3 is, however, less clear, but it will be the essential one for Section 4 since it is the only one that is applicable to the deterministic case. Condition (h_1) will be used in the outlook in Appendix Appendix A.

A simple example of a reasonable reference measure is the uniform distribution on an interval $[a, b]$. It is a natural choice whenever the data points are a priori restricted to $[a, b]$. For this example, the conditional relative entropy reduces to a conditional Shannon entropy:

Example 1 (uniform reference measure).
Let the range of X and Y be restricted to the interval $[0, 1]$ and U_X and U_Y be the uniform distributions on $[0, 1]$. Then the orthogonality condition h_1 in Theorem 1 is equivalent to

$$\int S(P_{Y|x})P(x)dx = \int_0^1 S(P_{Y|x})dx\,, \tag{7}$$

and

$$\mathrm{Cov}_{U_X}(S(P_{Y|x}), P(x)) = 0\,. \tag{8}$$

Hence, (7) states that regions with high entropy $S(P_{Y|x})$ do not correlate with regions of high density $P(x)$. If $P_{Y|X}$ and P_X are chosen independently, we assume that this independence assumption will approximately hold. For additive noise models, this is always satisfied because (1) implies that $S(Y|x)$ is constant in x. We have already given an intuitive argument (see also Fig. 2) why (7) is violated in the backward direction (in the low noise regime).

We can define a group of cyclic shifts $(S_t)_{t\in[0,1]}$ with $S_t(x) := (x + t)$ mod 1 having the uniform reference measure as unique invariant measure. Then the covariance in (8) vanishes on the average over all shifted copies of P_X (cf. Lemma 2), although we do not have any result saying that it holds for most shifted copies approximately.

To what extent the above orthogonality relations are approximately satisfied for real-world cause-effect pairs can only be answered by extensive empirical studies. An interesting theoretical question, however, is in which cases the orthogonality in one direction imposes the violation of orthogonality for the converse direction. The simplest model class where this could be confirmed is given by deterministic invertible relations [14]. A remarkable fact is that, for the backward direction, h_3 is always positively correlated with the hypothetical input density (i.e., in fact the output). Appendix Appendix A discusses some cases where the relation between cause and effect is not bijective and only deterministic in one direction. There, we are also able to show violations of orthogonality in backward direction, but sometimes additional independence conditions between P_X and $P_{Y|X}$ other than the orthogonality postulates turn out to be necessary.

4. Deterministic invertible relation

The bijective case where $Y = f(X)$ and $X = f^{-1}(Y)$ seems particularly challenging for causal inference. First, the absence of noise makes additive-noise model based inference impossible [9], and second, methods that use non-invertibility of the functional relation fail [15]. Surprisingly, the "hopeless" noiseless invertible case is one where the theory turns out to be most elegant because violation of one of our orthogonality conditions in backward direction follows easily from the orthogonality in forward direction. Moreover, our simulations suggest that the corresponding inference method is robust with respect to adding some noise; and also the empirical results on noisy real-world data with known ground truth were rather positive. This section largely follows our conference paper [14] but puts the ideas in a broader context and contains more systematic experimental verifications.

4.1. Motivation

We start with a motivating example. For two real-valued variables X and Y, let $Y = f(X)$ with an invertible differentiable function f. Let P_X be chosen independently of f. Then regions of high density P_Y correlate with regions where f has small slope (see Fig. 4). The following Lemma make this phenomenon more explicit:

Lemma 4 (correlations between slope and density).
Let $Y = f(X)$, where f is a differentiable bijection of $[0,1]$ with differentiable inverse f^{-1}. If $\log f'$ and P_X are uncorrelated in the sense that

$$\int \log f'(x) P(x) dx = \int \log f'(x) dx \,, \tag{9}$$

then $\log(f^{-1})'$ and P_Y are positively correlated, i.e.,

$$\int \log(f^{-1})'(x) P(y) dy > \int \log f'(y) dy \,,$$

unless f is the identity.

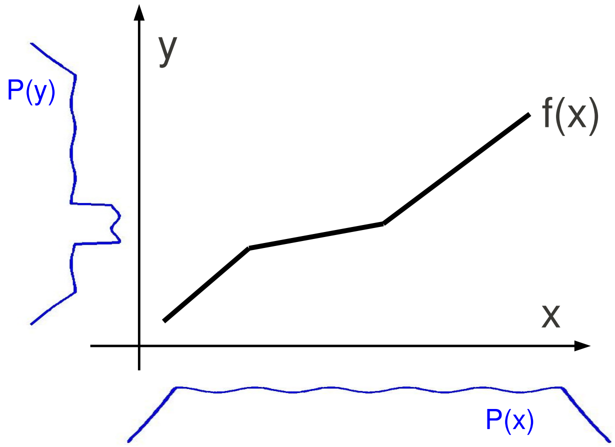

Figure 4: If the structure of the density of P_X is not correlated with the slope of f, then flat regions of f induce peaks of P_Y. The causal hypothesis $Y \rightarrow X$ is thus implausible because the causal mechanism f^{-1} appears to be adjusted to the "input" distribution P_Y.

Note that the terminology "uncorrelated" is justified if we interpret f' and p_X as random variables on the probability space $[0,1]$ with uniform measure (see the interpretation of (3) as uncorrelatedness). The lemma actually follows from more general results shown later, but the proof is so elementary that it is helpful to see:

$$\begin{aligned} & \int_0^1 \log(f^{-1})'(y)P(y)dy - \int_0^1 \log(f^{-1})'(y)dy \\ = & -\int_0^1 \log f'(x)P(x)dx + \int_0^1 \log f'(x)f'(x)dx \\ = & -\int_0^1 \log f'(x)dx + \int_0^1 \log f'(x)f'(x)dx = \int_0^1 (f'(x)-1)\log f'(x)dx \geq 0\,. \end{aligned}$$

The first equality uses standard substitution and exploits the fact that

$$\log(f^{-1})'(f(x)) = -\log f'(x)\,. \tag{10}$$

The second equality uses assumption (9), and the last inequality follows because the integral is non-negative everywhere. Since it can only vanish if Z is constant almost everywhere, the entire statement of Lemma 4 follows.

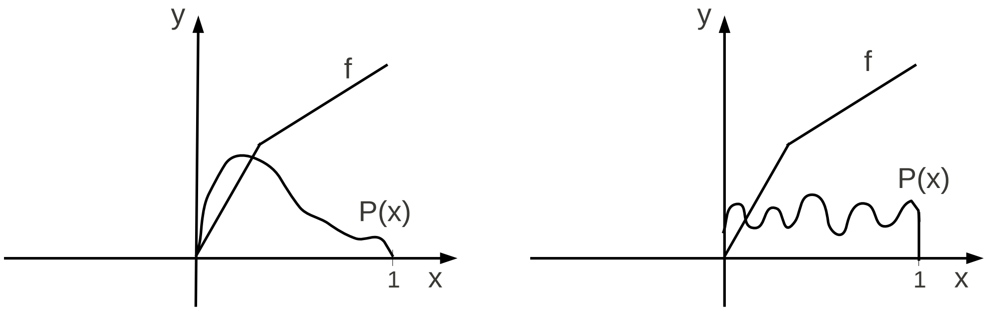

Figure 5: Left: violation of (9) due to a too global deviation of P_X from the uniform measure. Right: P_X oscillating around the constant density ensures uncorrelatedness.

Peaks of P_Y thus correlate with regions of large slope of f^{-1} (and thus small slope of f) if X is the cause. One can show that this observation can easily be generalized to the case where f is a bijection between sets of higher dimension. Assuming that P_X is uncorrelated with the logarithm of the Jacobian determinant $\log |\nabla f|$ implies that P_Y is positively correlated with $\log |\nabla f^{-1}|$.

Before embedding the above insights into our information geometric framework we will show an example where the whole idea fails:

Example 2 (failure of uncorrelatedness).
Let f be piecewise linear with $f'(x) = a$ for all $x < x_0$ and $f'(x) = b$ for all $x \geq x_0$. Then

$$\int_0^1 \log f'(x) P(x) dx - \int_0^1 \log f'(x) dx = (\log a - \log b) \left(P_X([0, x_0]) - x_0\right) .$$

Therefore, uncorrelatedness can fail spectacularly whenever $|P_X([0, x_0]) - x_0|$ is large, meaning that P_X and the uniform measure differ on a larger scale as in figure 5, left. If P_X only oscillates locally around 1, it still holds (figure 5 right).

The fact that the logarithm of the slope turned out to be particularly convenient due to (10), is intimately related to our information geometric framework: We first observe that $\overrightarrow{P}_Y$ and $\overleftarrow{P}_X$ have straightforward generalizations to the deterministic case as the images of U_X and U_Y under f and

$g := f^{-1}$, respectively. If U_X and U_Y are the uniform distributions on $[0, 1]$, they are given by

$$\overrightarrow{P}(y) := g(y) \quad \text{and} \quad \overleftarrow{P}(x) := f'(x)\,. \tag{11}$$

We thus obtain that (9) is equivalent to

$$\int_0^1 \log g'(y) P(y) dy = \int_0^1 \log g'(y) g'(y) dy\,,$$

which can be transformed to

$$\int_0^1 \log \frac{\overrightarrow{P}(y)}{U(y)} P(y) dy = \int_0^1 \log \frac{\overrightarrow{P}(y)}{U(y)} \overrightarrow{P}(y) dy$$

which is equivalent to orthogonality condition (h_3).

One can easily think of mechanisms in nature that violate the model of choosing the function f and the input distribution P_X independently because P_X is the result of intelligent design or a long adaption process, like evolution in biological systems. If the reward of a system can be optimized by controlling the value of y, P_X may over time shifted towards regions with large slope of f and thus (P_X, f) may spectacularly violate (9). Such effects imply a fundamental limitation of our method.

4.2. Identifiability results

Here we rephrase the theory developed in [14] and further elaborate on the asymmetries between cause and effect. Orthogonality (h_3) in Theorem 1 is the only one that is applicable to the deterministic case since it only refers to the image of the uniform input density under the conditional $P_{Y|X}$, which also exists in the deterministic case, while the others refer to the conditional *density* $P(y|x)$ (which does not exist since it would correspond to a delta-"function"). Condition (h_3) can be rephrased in different ways:

Theorem 2 (equivalent formulations of orthogonality (h_3)). *For bijective relations, the following conditions are equivalent:*

(I) Orthogonality (h_3) in Theorem 1:

$$D(P_Y \,||\, U_Y) = D(P_Y \,||\, \overrightarrow{P}_Y) + D(\overrightarrow{P}_Y \,||\, U_Y)\,. \tag{12}$$

(II) Uncorrelatedness between input and transformed density:

$$\mathrm{Cov}_{U_X}\left(\log\frac{\overleftarrow{P}_X}{U_X},\frac{P_X}{U_X}\right)=0\,. \tag{13}$$

(III) Transformed orthogonality:

$$D(P_X\,||\,\overleftarrow{P}_X)=D(P_X\,||\,U_X)+D(U_X\,||\,\overleftarrow{P}_X)\,. \tag{14}$$

(IV) Additivity of irregularities:

$$D(P_Y\,||\,U_Y)=D(P_X\,||\,U_X)+D(\overrightarrow{P}_Y\,||\,U_Y)\,. \tag{15}$$

(V) Additivity of approximation error:

$$D(P_X\,||\,\overleftarrow{P}_X)=D(P_Y\,||\,\overrightarrow{P}_Y)+D(\overrightarrow{P}_Y\,||\,U_Y) \tag{16}$$

Proof: Condition (13) is equivalent to

$$\int\log\frac{\overleftarrow{P}(x)}{U(x)}P(x)dx=\int\log\frac{\overleftarrow{P}(x)}{U(x)}U(x)dx\,,$$

using (4). Due to Lemma 3, this is equivalent to (14). The equivalence between (12) and (14) is immediate by applying f^{-1} to all distributions in (12) because the relative entropy is conserved under bijections. Equivalence between (15) and (12) is obtained by applying f^{-1} only to the first term on the right hand side of (12). By applying f^{-1} to the term on the left and f to the first term on the right hand side, (15) is transformed into (16). $\square$

Later in this section, a generalization of Condition (15) will be our essential postulate. For this reason, we should mention the idea: the distance $D(P_X\,||\,U_X)$ measures the irregularities of the input distribution and $D(\overrightarrow{P}_Y\,||\,U_Y)$ quantifies the irregularities of the function. The amount of irregularities of the output is thus given by the sum of these two terms. This is because the irregularities between input and function are independent, thus they neither "interfere" constructively nor destructively.

Condition (16) also admits an interesting interpretation: assuming that U_X and U_Y are given by smoothing P_X and P_Y, respectively, then $D(P_Y\,||\,\overrightarrow{P}_Y)$

is the error of approximating P_Y by $\overrightarrow{P}_Y$, i.e., the image of the smoothed input. Then (16) implies that the output is less sensitive to smoothing the input than vice versa: imagine the case where some of the peaks of P_Y stem from P_X and some from f. By smoothing the peaks that are caused by f, we generate additional peaks on P_X, while smoothing the ones of P_X just removes those in P_Y that are due to the peaks in the non-smoothed P_X. For all these interpretations it is essential that relative entropy is always non-negative.

Theorem 3 (violations in backward direction).
Let f be non-trivial in the sense that the image of U_X under f does not coincide with U_Y. If one condition (and thus all) in Theorem 2 holds, then the corresponding conditions that exchange the role of X and Y are violated with definite sign:

$$\begin{aligned} D(P_X\,||\,\overleftarrow{P}_X) + D(\overleftarrow{P}_X\,||\,U_X) &> D(P_X\,||\,U_X) && (17) \\ \mathrm{Cov}_{U_Y}\left(\log\frac{\overrightarrow{P}_Y}{U_Y}, \frac{P_Y}{U_Y}\right) &> 0 && (18) \\ D(P_Y\,||\,U_Y) + D(U_Y\,||\,\overrightarrow{P}_Y) &> D(P_Y\,||\,\overrightarrow{P}_Y) && (19) \\ D(P_Y\,||\,U_Y) + D(\overleftarrow{P}_X\,||\,U_X) &> D(P_X\,||\,U_X) && (20) \\ D(P_X\,||\,\overleftarrow{P}_X) + D(\overleftarrow{P}_X\,||\,U_X) &> D(P_Y\,||\,\overrightarrow{P}_Y)\,. && (21) \end{aligned}$$

Proof: reordering (14) yields

$$D(P_X\,||\,U_X) = D(P_X\,||\,\overleftarrow{P}_X) - D(U_X\,||\,\overleftarrow{P}_X) < D(P_X\,||\,\overleftarrow{P}_X) + D(\overleftarrow{P}_X\,||\,U_X)\,,$$

showing ineq. (17). Ineqs. (19)–(21) then follow by applying f^{-1} to some of the terms, but (20) and (21) follow also directly from (15) and (16), respectively. (18) follows because the left hand side is the difference between the right hand and the left hand side of (19). The fact that (12) implies (17) can also be seen in Fig. 6. Moreover, the fact that f^{-1} conserves the shape of the triangle shows that the discrepancy between the two sides of (17) is given by the "symmetrized relative entropy"

$$D(\overleftarrow{P}_X\,||\,U_X) + D(U_X\,||\,\overleftarrow{P}_X)\,. \qquad (22)$$

□

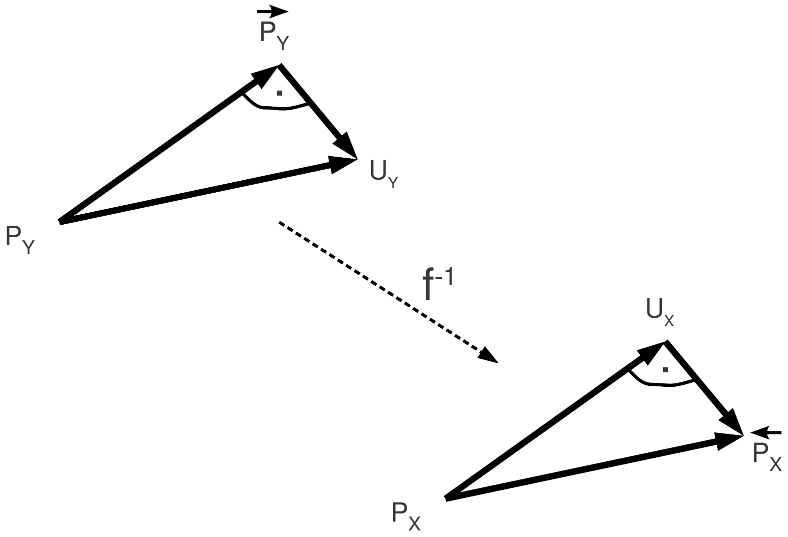

Figure 6: The orthogonality condition (I) is inconsistent with the analog orthogonality for the backward direction: since f^{-1} preserves all distances, a rectangular angle in information space at $\overrightarrow{P}_Y$ implies a rectangular angle at U_X rather than at $\overleftarrow{P}_X$, as it would be required.

Generalization to reference manifolds

The choice of the reference measure is the most delicate part of our method because the structure of a distribution P_X is represented by the vector connecting P_X and U_X.

The uniform distribution on a certain interval may only be a reasonable choice if the range of the respective variable is a priori restricted to this interval. If a real-valued variable has unbounded range and finite variance, the Gaussian with the same mean and variance as P_X is a more natural candidate for U_X (and likewise for Y). However, U_X then depends on P_X via its mean and variance. A better way of expressing the above is then given by introducing families of reference distributions rather than having a *single* reference distribution. We then measure irregularities by the distance of P_X to the exponential family of Gaussians and represent the structure of P_X by the vector that connects P_X to its closest point in the manifold. The family of Gaussians is only one example of a reasonable choice. Even though it will turn out to be a useful one in many cases, the theory below is phrased in terms of general exponential manifolds:

Definition 2 (exponential manifolds).
Let $\Omega \subseteq \mathbb{R}^d$ and assume a finite dimensional vector space V of functions $f : \Omega \to \mathbb{R}$ is given. Then, V defines an exponential manifold $\mathcal{E}$ by the set of probability densities that can be written as[2]

$$P(\omega) \propto e^{v(\omega)} \quad \forall \omega \in \Omega \,.$$

For any density P, $D(P \,||\, \mathcal{E})$ denotes the infimum of $D(P \,||\, Q)$ with $Q \in \mathcal{E}$. If there is a Q with $D(P \,||\, \mathcal{E}) = D(P \,||\, Q)$, it is called projection of P onto $\mathcal{E}$.

Note that the projection is unique whenever it exists [16]. Given appropriate reference manifolds for X and Y (formalizing the set of "smoothest" distributions), our inference method will be based on the following assumption:

[2] It is common to use slightly more general definitions [16] where the exponent also contains a fixed additional function that is not in V. Our formulation ensures that $\mathcal{E}$ contains the constant density whenever Ω has finite measure.

Postulate 2 (orthogonality for reference manifolds).
Let $\mathcal{E}_X$ and $\mathcal{E}_Y$ be "reasonable" reference manifolds for X and Y, respectively. If X causes Y then the conditions of Theorem 2 hold approximately, where U_X and U_Y be the projections of P_X and P_Y onto $\mathcal{E}_X$ and $\mathcal{E}_Y$, respectively.

For reference manifolds (instead of single reference distributions), this postulate requires a slightly different justification. This is explained in Appendix Appendix B.

The choice of the reference manifold is the point where prior knowledge on the respective domain enters into the method in the same way as the choice of the single reference measure did in the theory developed previously. Choosing the family of all Gaussians has the following interesting feature: the distance to the closest Gaussian defines a scale- and location-invariant measure of irregularities of P_X. Choosing a manifold *smaller* than the set of all Gaussians would keep some of the information about location or scale, choosing a *larger* manifold would also remove some of the scale- and location-invariant information about P_X. This is why the Gaussians are a natural choice at least in the one-dimensional case. For multi-dimensional variables X and Y, we will later see that the manifold of *all* Gaussians is often too large because it also removes the information about *relative* scaling of the different components of each variable X and Y. In this case, we will choose a proper submanifold.

4.3. Inference method (general form)

Having derived a long list of asymmetries between cause and effect, we have to chose one that is convenient for inferring the causal direction. To this end, we observe that the additivity of irregularities in (15) obviously implies

$$D(P_X \,||\, U_X) \leq D(P_Y \,||\, U_Y)\,,$$

whenever X causes Y. Generalizing this to reference manifolds (see Postulate 2) implies

$$D(P_X \,||\, \mathcal{E}_X) \leq D(P_Y \,||\, \mathcal{E}_Y)\,, \tag{23}$$

with equality if and only if $D(\overrightarrow{P}_Y \,||\, U_Y) = 0$ (i.e., when the function is so simple that the image of U_X is U_Y). Therefore, our inference method reads:

Information Geometric Causal Inference (IGCI):
Let $\mathcal{E}_X$ and $\mathcal{E}_Y$ be manifolds of "smooth" reference distributions for X and Y,

respectively. Consider the distances of P_X and P_Y to $\mathcal{E}_X$ and $\mathcal{E}_Y$, respectively, as the complexity of the distributions. Define the complexity loss from P_X to P_Y by

$$C_{X\to Y} := D(P_X \,||\, \mathcal{E}_X) - D(P_Y \,||\, \mathcal{E}_Y) \tag{24}$$

Likewise, the loss from P_Y to P_X is given by exchanging the roles of X and Y.
Then, infer that X causes Y if $C_{X\to Y} < 0$, or that Y causes X if $C_{X\to Y} > 0$.

To make this rule applicable, we first derive more explicit forms of $C_{X\to Y}$, which still refer to general reference manifolds. Subsection 4.4 then describes estimators from empirical data that refer to particular reference manifolds.

Lemma 5 ($C_{X\to Y}$ as difference of Shannon entropies).
Let P_X and P_Y be densities on $\mathbb{R}^d$. Assume that U_X and U_Y are the projections of P_X on $\mathcal{E}_X$ and P_Y on $\mathcal{E}_Y$, respectively. Then

$$\begin{aligned} C_{X\to Y} &= (S(U_X) - S(U_Y)) - (S(P_X) - S(P_Y)) && (25)\\ &= (S(U_X) - S(P_X)) - (S(U_Y) - S(P_Y))\,. && (26) \end{aligned}$$

Proof: since U_X is the projection of P_X onto $\mathcal{E}_X$, we have

$$\begin{aligned} D(P_X \,||\, \mathcal{E}_X) &= D(P_X \,||\, U_X) = -S(P_X) - \int P(x)\log U(x)dx \\ &= -S(P_X) + S(U_X)\,. && (27) \end{aligned}$$

To derive the last equation, we first assume that P_X and all densities in $\mathcal{E}_X$ have compact support $\Lambda \subset \mathbb{R}^d$. Then $\mathcal{E}$ contains the uniform distribution $U_X^{(0)}$ since the vector space defining $\mathcal{E}$ clearly contains the constant function $x \mapsto 0$. Because U_X is the projection of P_X onto $\mathcal{E}_X$, $(P_X, U_X, U_X^{(0)})$ form a pythagorean triple [17]. Using Lemma 3, we obtain $-\int P(x)\log U(x)dx = S(U_X)$. For non-compact supports, we consider the restrictions of all densities to an increasing sequence of compact subsets Λ_n. The statement then follows by the limit $n \to \infty$. $\square$

The entropy difference between X and Y can also be rewritten as follows:

Lemma 6 ($C_{X\to Y}$ as mean of log Jacobi determinant).
If f is a diffeomorphism between submanifolds of $\mathbb{R}^d$, then

$$C_{X\to Y} = S(U_X) - S(U_Y) + \int \log|\det \nabla f(x)| P(x)dx\,,$$

where we have used the notations of Lemma 5.

Proof: The entropy of $Y = f(X)$ reads

$$S(P_Y) = S(P_X) + \int P_X(x) \log|\det \nabla f(x)|\; dx\,,$$

thus we have

$$\begin{aligned} C_{X\to Y} &= \big(S(U_X) - S(P_X)\big) - \big(S(U_Y) - S(P_Y)\big) \\ &= S(U_X) - S(U_Y) + \int \log|\det \nabla f(x)|\, P(x)dx\,. \end{aligned} \tag{28}$$

□

Note that $C_{X\to Y}$ is invariant under joint rescaling of P_X and U_X (and likewise for P_Y and U_Y), since $S(U_X)$ changes by the same additive constant as $\det \nabla f$, except for the sign. In the next subsection, we discuss some important cases of domains of X and Y and describe possible choices of reference manifolds and how to empirically estimate $\hat{C}_{X\to Y}$.

4.4. Inference method (explicit form for reference measures on $\mathbb{R}$)

Lemma 5 and 6 reduce the estimation of $C_{X\to Y}$ and $C_{Y\to X}$ to estimating entropies or Jacobians, respectively. In this paper we are mainly concerned with one-dimensional continuous variables. We therefore give the explicit form of the estimators for this case, which will be used in our experiments. For completeness, we also discuss other situations in Subsection 4.5 and propose corresponding reference measures.

Uniform reference measure on intervals

For our motivating example of Subsection 4.1, where X and Y attain values in $[0, 1]$, Lemma 5 and Lemma 6 imply the following two simple versions of IGCI:

1. *Entropy-based IGCI:* infer $X \to Y$ whenever $S(P_X) > S(P_Y)$.
   To implement this in practice, we used the entropy estimator [18]:

   $$\hat{S}(P_X) := \psi(m) - \psi(1) + \frac{1}{m-1} \sum_{i=1}^{m-1} \log |x_{i+1} - x_i| \ , \tag{29}$$

   where the x-values should be ordered ascendingly, i.e., $x_i \leq x_{i+1}$, and ψ is the digamma function.[3] Note that here we set $\log 0 = 0$, i.e., the points with $x_{i+1} = x_i$ don't contribute to the sum. The estimate for $C_{X \to Y}$ based on (29) is then given by:

   $$\hat{C}_{X \to Y} := \hat{S}(P_Y) - \hat{S}(P_X) = -\hat{C}_{Y \to X} \, . \tag{30}$$

2. *Slope-based IGCI:* infer $X \to Y$ whenever

   $$\int_0^1 \log |f'(x)| P(x) dx < \int_0^1 \log |g'(y)| P(y) dx \, .$$

   We introduce the following estimator:

   $$\hat{C}_{X \to Y} := \frac{1}{m-1} \sum_{i=1}^{m-1} \log \left| \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \right| \ , \tag{31}$$

   where the x_i values are ordered, and a similar one for $\hat{C}_{Y \to X}$.

With the assumptions of this section, (30) and (31) coincide exactly, because the ψ-terms cancel when taking the difference between the estimated entropies of X and Y and because ordering the x-values is equivalent to ordering the y-values. In the noisy case, the relation between both methods is not yet understood (see also Section 4.6). (31) then diverges for $m \to \infty$ since the difference of y-values remains finite when the difference of x-values gets closer to zero. Then one has to compensate for this by considering the difference of this estimator and its analog in the reverse direction (obtained by swapping the roles of X and Y).

[3] The digamma function is the logarithmic derivative of the gamma function: $\psi(x) = d/dx \log \Gamma(x)$. It behaves as $\log x$ asymptotically for $x \to \infty$.

Gaussian reference measure on $\mathbb{R}$

Let us discuss the case $d = 1$ first. Lemma 5 and 6 imply that $C_{X\to Y}$ and $C_{Y\to X}$ remain formally the same as for the uniform reference measure after we rescale X and Y such that they have the same variance (note that this ensures $S(U_X) = S(U_Y)$). In contrast, the uniform measure required all data points to lie in $[0, 1]$. The different scaling changes $C_{X\to Y}$ by $\log \sigma_X - \log \sigma_Y$, where σ_X^2 and σ_Y^2 denote the variances of X and Y, respectively, according to the scaling used for uniform measure. Consequently, the methods may infer different directions when σ_X^2 and σ_Y^2 differ significantly, although this did not happen that often in our real world data experiments.

4.5. Inference rule for other variable ranges and reference manifolds

Although our experiments contained only real-valued variables, we sketch how to use IGCI also for variables with other ranges.

Gaussian reference measure on $\mathbb{R}^d$

Suppose now that both X and Y are d-dimensional real random vectors, and that f is a diffeomorphism $\mathbb{R}^d \to \mathbb{R}^d$. Let both $\mathcal{E}_X$ and $\mathcal{E}_Y$ be the manifolds of d-dimensional Gaussian distributions. The projection U_X is the d-variate Gaussian with the same mean vector and covariance matrix as X, denoted by Σ_X. U_Y is derived similarly. The difference of the entropies of U_X and U_Y thus reads $\frac{1}{2}\log(\det \Sigma_X / \det \Sigma_Y)$. Then we can easily compute $C_{X\to Y}$ based on (26). Because the entropy difference $S(U_X) - S(P_X)$ is a measure of non-Gaussianity, the method thus considers the variable that is closer to a Gaussian as the cause.

Isotropic Gaussians as reference on $\mathbb{R}^d$

We will now show that the deterministic case of the method described in [12] and [13] relies on an assumption that implies Postulate 2 for a particular choice of the reference manifold. Let P_X and P_Y be multivariate Gaussians in $\mathbb{R}^d$ with zero mean and X and Y be related by

$$Y = AX\,, \tag{32}$$

where A is an invertible $d \times d$-matrix.[4] For an arbitrary $d \times d$ matrix B let $\tau(B) = \mathrm{tr}(B)/d$ denote the renormalized trace. Then [12] is based on the

[4] Ref. [12] also considers the case $Y = AX + E$, where E is an independent noise term, but we restrict the attention to the deterministic one.

assumption that $X \to Y$ implies approximately

$$\tau(\Sigma_Y) = \tau(\Sigma_X)\,\tau(AA^T)\,, \tag{33}$$

where Σ_X and Σ_Y denote the covariance matrices of X and Y, respectively. In [12] this is further justified by showing that for any given A, choosing Σ_X randomly from a rotation invariant prior ensures that (33) is approximately true with high probability[5]. We now show that this implies Postulate 2 if both $\mathcal{E}_X$ and $\mathcal{E}_Y$ are the manifold of *isotropic* Gaussians, i.e., those whose covariance matrices are multiples of the identity. U_X and U_Y have the same mean as X and Y and their covariance matrices read $\tau(\Sigma_X)\mathbf{I}$ and $\tau(\Sigma_Y)\mathbf{I}$. The relative entropy distance between two Gaussians with equal mean and covariance matrices Σ_1 and Σ_0 is given by

$$D(P_{\Sigma_1}\,||\,P_{\Sigma_0}) = \frac{1}{2}\left(\log\frac{\det\Sigma_0}{\det\Sigma_1} + d\left[\tau(\Sigma_0^{-1}\Sigma_1) - 1\right]\right)\,.$$

The distances to the manifold of isotropic Gaussians thus read [12]

$$\begin{aligned} D(P_X\,||\,\mathcal{E}_X) &= \frac{1}{2}\big(d\log\tau(\Sigma_X) - \log\det(\Sigma_X)\big) && (34)\\ D(P_Y\,||\,\mathcal{E}_Y) &= \frac{1}{2}\big(d\log\tau(\Sigma_Y) - \log\det(\Sigma_Y)\big) && (35)\end{aligned}$$

The covariance matrix of $\overrightarrow{P}_Y$ reads $\tau(\Sigma_X)AA^T$. Hence,

$$D(\overrightarrow{P}_Y\,||\,U_Y) = \frac{1}{2}\left(\log\frac{\tau(\Sigma_Y)^d}{\tau(\Sigma_X)^d\,\det(AA^T)} + d\left[\frac{\tau(\Sigma_X)\,\tau(AA^T)}{\tau(\Sigma_Y)} - 1\right]\right)\,.$$

Due to $\det(\Sigma_Y) = \det(\Sigma_X)\,\det(AA^T)$ we have

$$D(P_Y\,||\,\mathcal{E}_Y) = D(P_X\,||\,\mathcal{E}_X) + D(\overrightarrow{P}_Y\,||\,U_Y) + \frac{d}{2}\left[1 - \frac{\tau(\Sigma_X)\,\tau(AA^T)}{\tau(\Sigma_Y)}\right]\,.$$

Assumption (33) is thus equivalent to condition (V) in Theorem 2. Postulate 2 thus gets an additional justification via a probabilistic scenario where f is fixed and P_X is chosen randomly from a prior that satisfies a certain symmetry condition.

[5] Ref. [13] extends this framwork to the case where the number of dimensions exceeds the number of samples.

For high-dimensional relations that are close to linear, the method above seems more appropriate than the one that uses the set of *all* Gaussians (as opposed to the isotropic ones only) as reference manifold. Allowing for all Gaussians, the method makes use of the non-linearities of f, while it removes the information that is contained in Σ_X. For relations that are close to the linear case, one thus looses the essential information, while taking isotropic Gaussians as reference ensures that only the information that describes the joint (overall) scaling is lost.

Non-uniform reference measure on finite sets

The intuitive explanation of the identifiability of cause and effect used the fact that regions of high density of the effect correlate with regions of high slope of the inverse function. Remarkably, our method is in principle also applicable to bijections between *finite* probability spaces, provided that we ensure that $D(\overrightarrow{P}_Y \,||\, \mathcal{E}_Y) > 0$ (which is not the case if $\mathcal{E}_X$ and $\mathcal{E}_Y$ consist of the uniform distribution only). We omit the details but only give a brief sketch of a special case here.

Assume that both X and Y take values in $\{1, \dots, k\}$ and P_X and P_Y are probability mass functions with $P_Y(y) = P_X\big(g(y)\big)$. (Note that in this discrete case, g is invertible but not monotonic.) Let $\mathcal{E}_X$ and $\mathcal{E}_Y$ be the two-parametric manifold of distributions of "discrete Gaussians" with

$$U(x \,|\, \mu, \sigma) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} ,$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. Then the image of the discrete Gaussians will usually not be a discrete Gaussian and our inference principle becomes non-trivial, yielding preference for one direction. The essential question is, however, under which conditions Postulate 2 is still reasonable. The following explanations provide an idea about this. Assume that k is large and that P_X is a distribution that is close to one of the above discrete Gaussian except for a small number of x-values. Let f be a bijection that preserves most of the points $\{1, \dots, k\}$, while permuting only some of them. It is then likely that this permutation increases the distance to the reference manifold rather than decreasing it. This way of reasoning certainly relies on the assumption that k is large and that the distance of P_X to the reference manifold is not too large. For small k, one can easily construct examples with P_X deviating so strongly from the Gaussians that a significant fraction of permutations decrease the distance to the reference manifold.

4.6. Performance in the noisy regime

The assumption of having a bijective deterministic relation is actually necessary for the IGCI method. Section 4.7, however, will show that the performance on our real data sets was unexpectedly good, even though most of them are obviously noisy. We therefore present some explanations for this fact. Although the noisy case is actually out of scope, the development of future methods could be inspired by understanding the unexpectedly reasonable performance in this regime.

On the one hand, we estimate how small the noise needs to be in order not to spoil the method (Subsubsection 4.6.1). On the other hand we show, that under some conditions noise can even contribute to inferring the correct causal direction (Subsubsection 4.6.2).

First we discuss a case where IGCI necessarily fails. Let Y be generated from X by a linear model with additive noise

$$Y = X + E \quad \text{with} \quad E \perp\!\!\!\perp X\,,$$

hence P_Y is obtained by the convolution $P_Y = P_X * P_E$. For Gaussians as reference manifolds, the projections U_X and U_Y of P_X and P_Y on $\mathcal{E}_X$ and $\mathcal{E}_Y$, respectively, are given by the Gaussians with the same mean and variance. If E is Gaussian, we thus have $U_Y = U_X * P_E$ due to the additivity of means and variances under convolution. We have

$$D(P_X\,||\,\mathcal{E}_X) = D(P_X\,||\,U_X) > D(P_X{*}P_E\,||\,U_X{*}P_E) = D(P_Y\,||\,U_Y) = D(P_Y\,||\,\mathcal{E}_Y)\,,$$

because the convolution with a Gaussian decreases the distance to the set of Gaussians (that it is non-increasing already follows from monotonicity of relative entropy distance under stochastic maps [19]). Hence, (23) is violated and, after renormalizing X and Y to unit variance, the entropy of Y will be greater than the entropy of X. The entropy-based estimator for $C_{X\to Y}$ will thus converge to a positive number, while our theory makes no statement on the slope-based estimator (note that the equivalence of both required deterministic models). Similar arguments hold for $Y = \alpha X + E$, we have restricted the derivation above to $\alpha = 1$ only for technical convenience. Hence, entropy based IGCI with Gaussians as reference manifold fails if the non-linearity of f is small compared to the width of the (Gaussian) noise. The following subsection provides a bound on how relevant small noise can get for the decision made by IGCI.

4.6.1. Robustness of entropy based inference under adding small noise

We restrict the attention again to real-valued X and Y and recall that the entropy generated by adding independent Gaussian noise is related to the Fisher information

$$J(Y) := \mathbb{E}_P \left(\frac{\partial \log P(y)}{\partial y} \right)^2$$

by De Bruijn's identity [19]:

$$\frac{\partial}{\partial t} S(P_{Y+\sqrt{t}Z}) = \frac{1}{2} J(Y + \sqrt{t}Z) \,, \tag{36}$$

where Z is a Gaussian with variance 1 and $Y \perp\!\!\!\perp Z$. The following Lemma provides a lower bound on the non-Gaussianity of the perturbed variable:

Lemma 7 (non-Gaussianity of noisy output).
If $\mathcal{E}_Y$ denotes the manifold of Gaussians and E is Gaussian noise with $E \perp\!\!\!\perp Y$ then the "decrease of non-Gaussianity" is bounded from above by

$$D(P_Y \,||\, \mathcal{E}_Y) - D(P_{Y+E} \,||\, \mathcal{E}_Y) \leq \frac{1}{2} \log \left(1 + [J(Y)\sigma_Y^2 - 1] \frac{\sigma_E^2}{\sigma_Y^2 + \sigma_E^2} \right) ,$$

where σ_Y^2 and σ_E^2 denote the variance of the unperturbed output and the noise, respectively.

Proof: Set $E := \sigma_E Z$ for standard Gaussian Z, then (36) implies

$$\begin{aligned} S(P_{Y+E}) - S(P_Y) &= \int_0^{\sigma_E^2} \frac{\partial}{\partial t} S(P_{Y+\sqrt{t}Z}) dt \\ &= \frac{1}{2} \int_0^{\sigma_E^2} J(Y + \sqrt{t}Z) dt \leq \frac{1}{2} \int_0^{\sigma_E^2} \frac{J(Y) J(\sqrt{t}Z)}{J(Y) + J(\sqrt{t}Z)} dt \,, \end{aligned}$$

where the last inequality is due to the Fisher information inequality [20]

$$\frac{1}{J(Y+W)} \geq \frac{1}{J(Y)} + \frac{1}{J(W)} \,,$$

for arbitrary independent random variables Y and W. Using $J(\sqrt{t}Z) = 1/t$ (which can be checked via straightforward computation), we obtain

$$\begin{aligned} S(P_{Y+E}) - S(P_Y) &\leq \frac{1}{2}\int_0^{\sigma_E^2} \frac{1}{t + 1/J(Y)} dt \\ &= \frac{1}{2}\left[\log\left(\sigma_E^2 + \frac{1}{J(Y)}\right) - \log\left(\frac{1}{J(Y)}\right)\right] \\ &= \frac{1}{2}\log\left(J(Y)\sigma_E^2 + 1\right) . \end{aligned}$$

Recalling from (27) that non-Gaussianity is given by

$$D(P_Y \,||\, \mathcal{E}_Y) = \frac{1}{2}\log(2\pi e\sigma_Y^2) - S(P_Y)\,,$$

because the first term is the entropy of the Gaussian with variance σ_Y^2, the non-Gaussianity changes according to

$$\begin{aligned} D(P_Y \,||\, \mathcal{E}_Y) - D(P_{Y+E} \,||\, \mathcal{E}_Y) &\leq \frac{1}{2}\left(\log\frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_E^2} + \log\left(J(Y)\sigma_E^2 + 1\right)\right) \\ &= \frac{1}{2}\log\left(1 + [J(Y)\sigma_Y^2 - 1]\frac{\sigma_E^2}{\sigma_Y^2 + \sigma_E^2}\right) . \end{aligned}$$

□

Note that Gaussians minimize Fisher information for a given variance and thus $J(Y)\sigma_Y^2 - 1 \geq 0$, with equality for Gaussians. If Y is Gaussian, convolution with a Gaussian cannot decrease non-Gaussianity any further because it is already zero. For non-Gaussian Y, the amount of decrease not only depends on the "sensitivity term" $[J(Y)\sigma_Y^2 - 1]$ but also on the ratio between the variance of the noise and the total variance $\sigma_Y^2 + \sigma_E^2$ of the noisy output.

Lemma 7 assumes Gaussian noise. We expect however that non-Gaussian noise will typically decrease non-Gaussianity even less than Gaussian noise does, except for rare cases of very particularly distributed noise. We therefore propose to use the bound for general noise. To decide whether noise may have reversed the inferred causal arrow, we could proceed as follows. For every hypothetical cause, say, X, we can estimate the density and the function f and thus compute the distribution of the effect Y without noise.

After computing its Fisher information we can estimate the decrease of non-Gaussianity caused by the noise and check whether it is smaller than the difference between $D(P_{Y'} \,||\, \mathcal{E}_Y)$ and $D(P_X \,||\, \mathcal{E}_X)$, where $Y' := Y + E$ denotes the noisy effect.

4.6.2. Performance of slope-based inference in the noisy regime

We will now take a closer look at the estimator (31) in the noisy case. The arguments below are partly heuristic, but simulation studies in Subsection 4.7 support our claims. Assume that the i.i.d sample (x_i, y_i) with $i = 1, \ldots, m$ is generated by an additive noise model (1) with strictly monotonic differentiable f. We assume that the x_i and hence also the $f_i := f(x_i)$ are already ordered ($x_{i+1} \geq x_i$ and $f_{i+1} \geq f_i$). We have for large m

$$\frac{1}{m-1}\sum_{i=1}^{m-1} \log \left|\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\right| = \frac{1}{m-1}\sum_{i=1}^{m-1} \log \left|\frac{(f_{i+1}-f_i)+(e_{i+1}-e_i)}{x_{i+1}-x_i}\right|$$

$$\approx \frac{1}{m-1}\sum_{i=1}^{m-1} \log |e_{i+1}-e_i| - \frac{1}{m-1}\sum_{i=1}^{m-1} \log |x_{i+1}-x_i| \,. \qquad (37)$$

The approximation is based on the observation that the difference $|f_{i+1} - f_i|$ gets negligible compared to $|e_{i+1} - e_i|$ for $m \to \infty$ since the latter term remains finite while the other one converges to zero.

The second term in (37) is actually the entropy estimator (30) up to the term $\psi(m) - \psi(1)$. Without the noise E, the two estimators (30) and (31) coincide with each other, as we have already argued. However, in the noisy regime, the first term tends to dominate as m increases, since it diverges as $m \to \infty$.

Now we write X as $X = \tilde{f}(Y) + \tilde{E}$ with an arbitrary function $\tilde{f}$. To focus on the noise effect, let us assume that X and Y have the same entropy such that the information contained in the nonlinear functions does not help identifying the causal direction, i.e., the estimator (30) would give the same value for $\hat{C}_{X\to Y}$ and $\hat{C}_{Y\to X}$. To investigate the behavior of the estimator (31), denote by $A_{X\to Y}$ the first term of (37), i.e.,

$$A_{X\to Y} := \frac{1}{m-1}\sum_{i=1}^{m-1} \log |e_{i+1}-e_i| \,,$$

and let

$$A_{Y\to X} := \frac{1}{m-1}\sum_{i=1}^{m-1} \log |\tilde{e}_{i+1}-\tilde{e}_i| \,.$$

As the second term of (37) is the same for both directions by assumption, (31) would prefer the direction $X \to Y$ (resp. $Y \to X$) if $A_{X\to Y}$ is smaller (resp. larger) than $A_{Y\to X}$.

The Jacobian matrix associated with the transformation from $(X, E)^T$ to $(X, Y)^T$ is

$$\mathbf{J} = \begin{bmatrix} 1 & 0 \\ f'(X) & 1 \end{bmatrix},$$

and hence $|\mathbf{J}| = 1$, where $|\mathbf{J}|$ denotes the absolute value of the determinant of $\mathbf{J}$. We then have $P_{X,Y} = P_{X,E}/|\mathbf{J}| = P_{X,E}$. As X and E are independent we further have

$$S(X, Y) = S(X) + S(E). \tag{38}$$

On the other hand, we have

$$S(X, Y) = S(Y, \tilde{E}). \tag{39}$$

Except for some special cases (for instance, where f is linear and both X and E are Gaussian) $\tilde{E}$ and Y are dependent [9, 10], i.e., $S(Y)+S(\tilde{E})-S(Y,\tilde{E}) > 0$. Due to (38) and (39), we thus have $S(X) + S(E) < S(Y) + S(\tilde{E})$. As we assumed $S(X) = S(Y)$, finally we have $S(E) < S(\tilde{E})$. Furthermore, as E and $\tilde{E}$ approximately have the same variance,[6] the above inequality implies that $\tilde{E}$ is more Gaussian than E.

Let $\tilde{D}_i := \tilde{E}_{i+1} - \tilde{E}_i$ and $D_i := E_{i+1} - E_i$. Under the condition that E_i are i.i.d., P_{D_i} is the convolution of P_E and P_{-E}. Likewise, $P_{\tilde{D}_i}$ is a convolution of $P_{\tilde{E}}$ with $P_{-\tilde{E}}$. Since $\tilde{E}$ is more Gaussian than E, it is quite likely that $\tilde{D}_i$ is also more Gaussian than D_i. We then consider the following three possible cases.

1. If E is Gaussian (and so are D_i), $\tilde{D}_i$ is also Gaussian (given the above heuristics), and $A_{X\to Y} = A_{Y\to X}$. Hence, the noise does not change the decision.
2. Consider the case where E is super-Gaussian, which, roughly speaking, means that P_E has a sharper peak and longer tails than the Gaussian variables with the same mean and variance. The Laplacian distribution is an example of such distributions. Since $\tilde{E}$ is more Gaussian than E,

[6] Note that E and $\tilde{E}$ have *exactly* the same variance, if both directions are fitted with linear functions (i.e., both f and $\tilde{f}$ are linear) and X and Y have the same variance.

E is more super-Gaussian than $\tilde{E}$. Consequently, D_i take relatively more values that are very close to zero than $\tilde{D}_i$. The function $\log |D_i|$ is concave on $(0, \infty)$ and symmetric w.r.t. the y-axis; we obtain large negative values for D_i that are close to zero. $A_{X \to Y}$ thus gets smaller than $A_{Y \to X}$. That is, *super-Gaussian noise tends to favor the correct direction*, $X \to Y$.

3. Suppose E is sub-Gaussian, which is flatter than the Gaussian variable with the same mean and variance. An example is the uniform distribution. As $\tilde{D}_i$ is more Gaussian (or less sub-Gaussian) than D_i, they take more often values that are very close to zero or very large than D_i. Hence $A_{Y \to X}$ is larger than $A_{X \to Y}$. In other words, *sub-Gaussian noise tends to favor the wrong direction*, $Y \to X$.

Fortunately, super-Gaussian noise occurs quite often in practice. Although we only analyze the noise effect above, one should bear in mind that with the estimator (31), the decision is made based on the joint effect of the properties of the nonlinear function and the noise distribution, which correspond to the second and first terms of (37), respectively.[7]

In the analysis above we assume that the data-generating process in the noisy case can be approximated by the additive noise model. Analyzing the noise effect in more general settings (e.g., in the PNL causal model [10]) is rather complicated, and is not given here. However, in Subsection 4.7 we also give simulation results on the data with a rather complex data generating process and illustrate how the noise influences the performance of IGCI.

4.7. Experiments

In this section we describe some experiments that illustrate the theory above and show that our method can detect the true causal direction in many real-world data sets. Complete source code for the experiments is provided online at `http://webdav.tuebingen.mpg.de/causality/` and `http://parallel.vub.ac.be/igci`. The latter provides an applet showing the data and the results of IGCI.

[7] Rigorously speaking, the noise in the forward direction also changes the best-fitting nonlinear function in the backward direction, which would influence the estimate of $C_{Y \to X}$ as well. As a simple illustration, consider the case where both X and E are uniform. Then the best-fitting function $\tilde{f}$ in the direction $Y \to X$ is no longer linear, and its shape depends on the noise level. However, we skip the details of this aspect.

Simulation studies (I): cause-effect pairs from a larger causal network

We investigate the performance of IGCI in the deterministic and the noisy regime.

To this end, we simulate a causal relation between n variables $X_1, \ldots, X_n$, from which we take different pairs $(Y, X) \equiv (X_i, X_j)$, where X_j is one of the parents of X_i. All causal dependences will be given by structural equations. This ensures that in our pairs not only the effects but also the causes are the outcomes of structural equations – reflecting the fact that causes in the real-world are effects of other variables.

The precise form of the data generating process is as follows. We first generate 20 independent variables $X_1, \ldots, X_{20}$. Their distribution is randomly chosen from two options with equal probability: either the uniform distribution on $[0, 1]$ or a Gaussian mixture distribution GM with the following density:

$$GM(x) = \sum_{i=1}^{g} w_i \phi(x|\mu_i, \sigma_i)\,,$$

where $g \in [1, 5]$, means $\mu_i \in [0, 1]$, standard deviations $\sigma_i \in [0, 1/g]$ and weights $w_i \in [0, 1]$ with $\sum_{i=1}^{g} w_i = 1$. Each parameter is randomly chosen from the interval according to a uniform distribution. Then, 50 variables $X_{21}, \ldots, X_{70}$ are defined according to the following structural equation:

$$X_i = f_i(X_j, \ldots, X_{j+k}) + \lambda_i R_i E_i\,,$$

with j, k defined later, where for each i:

- The function f_i is randomly selected from the following families:

  LIN Linear functions of the form

  $$f(x_j, \ldots, x_{j+k}) = \sum_{j=0}^{k} c_j x_{j+i}\,,$$

  where $c_j \in [-1, 1]$ and k is a natural number randomly according to the probability $1/2^k$.

  POL Polynomials of the form

  $$f(x) = \sum_{i=1}^{n} i c_i x^i\,,$$

with $n \in [1,5]$ and $c_i \in [-1,1]$. The purpose of the factor i is to ensure that the magnitude of each term is similar for $x \in [0,1]$.

MON Monomials of the form $f(x) = x^n$ with $n \in [2,5]$.

ROOT Root functions $f(x) = x^{1/n}$ with $n \in [2,5]$.

MG Cumulative distribution functions of mixtures of Gaussians:

$$f(x) = \sum_{i=1}^{5} \alpha_i \Phi(x|\mu_i, \sigma_i),$$

which is a convex combination of Gaussian cdf's $\Phi(x|\mu_i, \sigma_i)$ with $\alpha_i, \mu_i \in [0,1], \sigma_i \in [0, 0.2]$.

PROD Product functions of the form $f(x_j, x_{j+1}) = x_j x_{j+1}$.

QUOT Quotients $f(x_j, x_{j+1}) = x_j / x_{j+1}$.

- The variables $X_j, \ldots, X_{j+k}$ are chosen randomly from $X_1, \ldots, X_{i-1}$ (the "causally preceding" variables). Note that $k \geq 0$ for the linear function and k=1 for the product and division function. The latter results in non-additive noise, since the study of the relation between one input variable and the output variable is based on marginalizing the data over the second input variable. All other functions have only one independent variable.

- λ_i has the probability of 0.5 to be zero, and is otherwise chosen uniformly between $[0, 0.2]$.

- R_i is the difference of the maximum and the minimum of the function f_i after feeding it with the values of $X_j, \ldots, X_{j+k}$. In this way, the noise is proportional to the range of the function values.

- The noise term E_i is drawn from a Gaussian distribution with mean 0 and variance 1.

Note that a deterministic relation is obtained whenever $k = 0$ (i.e., X_j is the only parent of X_i) and the noise parameter λ vanishes. When a deterministic relation is monotonic and decreasing, we make it an increasing function by replacing each y-value with $1 - y$. We repeat the whole procedure of generating the variables $X_1, \ldots, X_{70}$ 100 times, and each time we generate 200 samples such that each causal decision will be based on $m = 200$ i.i.d. data points.

For each data set generated by this procedure, we apply our inference method to the 50 pairs $(Y, X) \equiv (X_i, X_j)$ for $i = 21, \ldots, 70$ (with randomly chosen j as above). We compared the entropy-based and the slope-based method. We also compare two different families of reference measures: the uniform family (which amounts to preprocessing both components of the data by an affine transformation such that the minimum is 0 and the maximum is 1) and the Gaussian family (where each component of the data is preprocessed by an affine transformation such that it has zero mean and standard deviation 1).

Fig. 7 shows some typical examples of input distributions, relations between input and output, and the corresponding output distribution. Table 1 lists the values $\hat{C}_{X\to Y}$ and $\hat{C}_{Y\to X}$ of the slope-based estimator (31) and the entropy estimator (30), as well as the corresponding decision. Remarkably, the decision was also correct for the linear noisy case (the fourth case). A possible explanation could be the one given in Subsection 4.6.2.

By only taking decisions for $|\hat{C}_{X\to Y} - \hat{C}_{Y\to X}| \geq \delta$ for some threshold δ, one can trade off accuracy (percentage of correct decisions) versus decision rate (percentage of cases in which a decision was taken). Fig. 8 shows the accuracy versus the decision rate for the deterministic relations, and Fig. 9 shows the same for the probabilistic relations. These results show that the method works best for deterministic relations, as expected. For deterministic relations, increasing the threshold increases the accuracy of the method, coming close to an accuracy of 100% for large threshold values. For deterministic relations, the Gaussian reference measure performs somewhat better than the uniform reference measure. For probabilistic relations, however, the picture is rather different. The uniform reference has an increasing accuracy starting at 70% when no threshold is used and reaching 85% for large thresholds. The Gaussian reference on the other hand fails for small thresholds; the accuracy is close to 50% which is the same as random guessing. Only for large thresholds (decision rates smaller than 20%) the accuracy reaches 70%. For both, deterministic and probabilistic relations, the slope-based estimator (31) and the entropy-based one (30) yield similar results.

It is also instructive to check to what extent the above procedure generates joint distributions $P_{Y,X}$ that satisfy our orthogonality assumptions. We

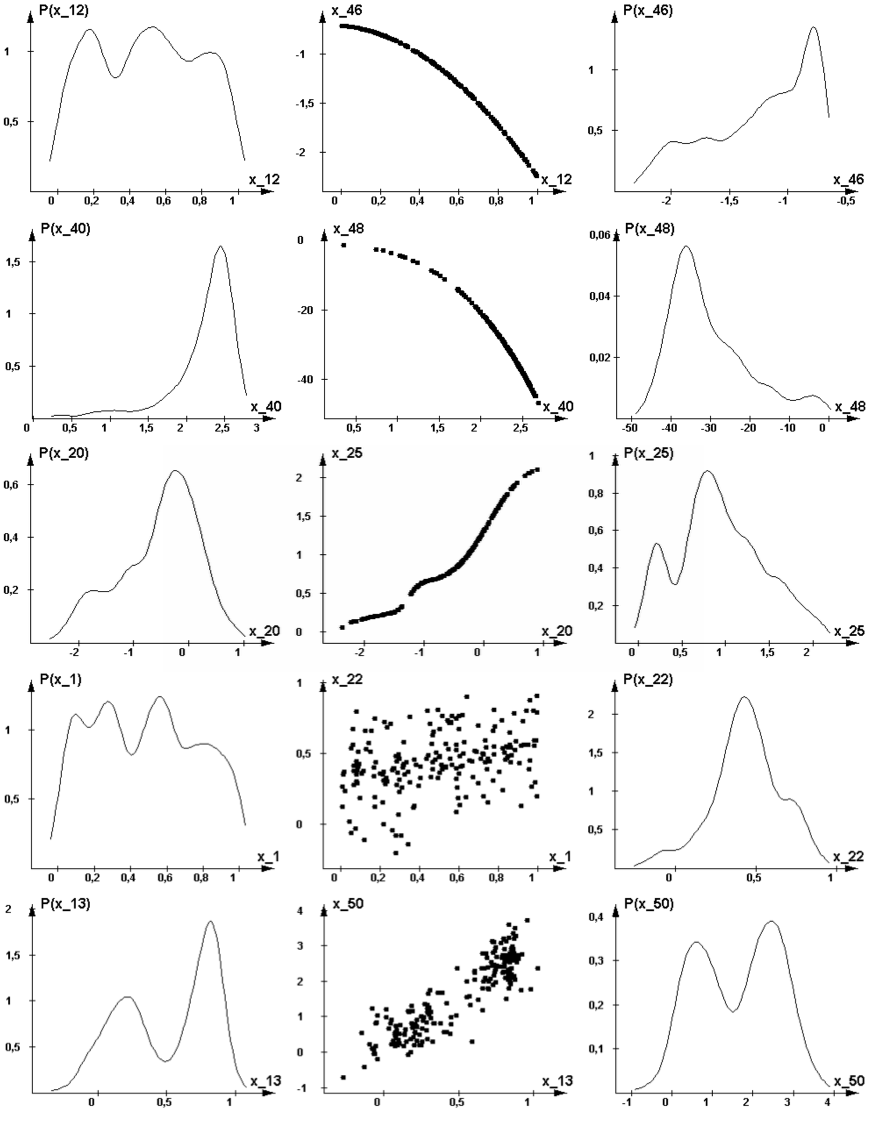

Figure 7: Typical synthetic cause-effect pairs illustrating various different cases: probabilistic versus deterministic relations, correct versus incorrect result and an indecision. The corresponding quantitative description is given in Table 1.

		(31)				(30)			
Function	type	$\hat{C}_{X\to Y}$	$\hat{C}_{Y\to X}$	Dec.	OK	$\hat{S}(P_X)$	$\hat{S}(P_Y)$	Dec.	OK
$X_{46} = f(X_{12})$	POL	-0.29	0.29	+	+	-2.53	-2.83	+	+
$X_{48} = f(X_{40})$	POL	0.73	-0.73	+	-	-3.81	-3.01	+	-
$X_{25} = f(X_{20})$	MG	0.03	-0.03	-	-	-2.84	-2.8	-	-
$X_{22} = f(X_1)$	LIN	5.60	6.44	+	+	-2.64	-2.93	+	+
$X_{50} = f(X_{13})$	MG	5.33	4.71	+	-	-3.14	-2.92	+	-

Table 1: Quantitative description (only for the uniform reference measures) of the typical examples depicted in Figure 7.

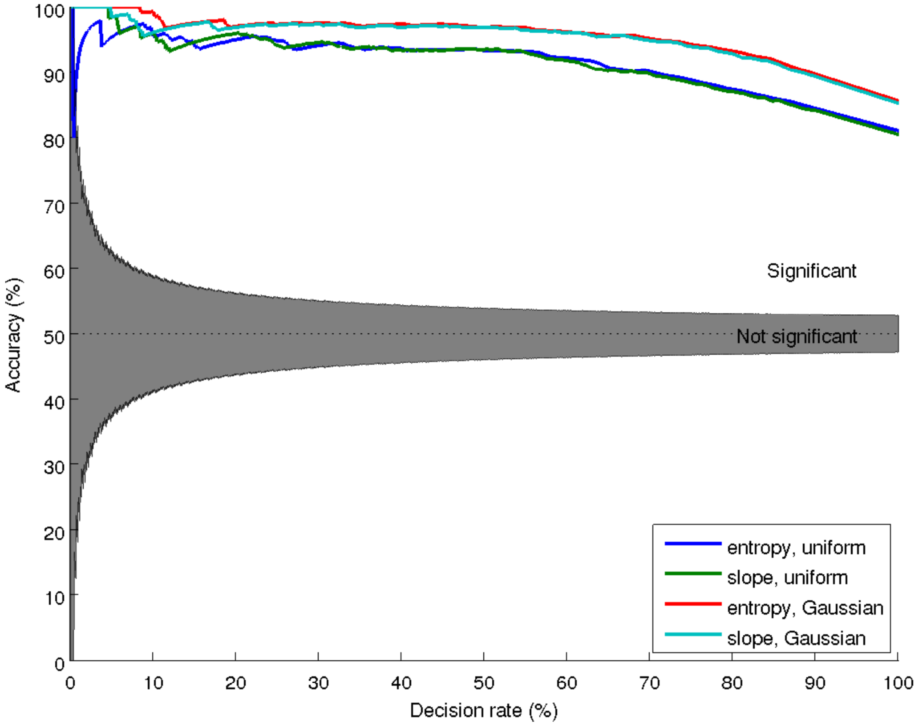

Figure 8: Results of four different implementations of IGCI on simulated deterministic causal relations for about 2000 different (X, Y) pairs.

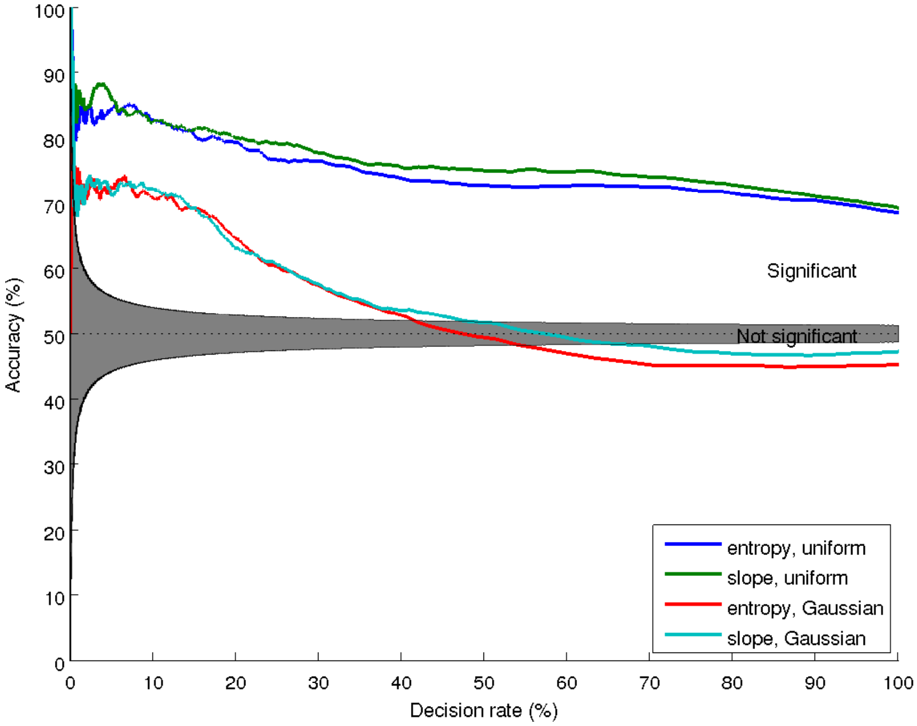

Figure 9: Results of four different implementations of IGCI on simulated probabilistic causal relations for about 3000 different (X, Y) pairs.

therefore compare

$$\mathrm{Cov}_{U_X}(\log f', P_X) \quad \text{to} \quad \mathrm{Cov}_{U_Y}(\log f^{-1'}, P_Y)\,,$$

because the covariance in condition (II) in Theorem 2 and the corresponding expression for the backward direction take this form after we use (11) for the uniform reference measure. The first expression is given by the estimator

$$\widehat{\mathrm{Cov}}(P_X, \log f') := \frac{1}{m-1}\sum_{i=1}^{m-1}\left(1 - \frac{x_{i+1} - x_{i-1}}{2}\right)\log\left|\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right|\,, \qquad (40)$$

and the second by exchanging the roles of X and Y. By the simulation explained above 1575 examples of deterministic strictly monotonic relations were generated. The x-axis of Fig. 10 shows the values (40) and the y-axis the analog one for the backward direction.

The figure confirms our postulate in the sense that the covariance in forward direction is usually closer to zero. Most of the values for the forward direction are in the interval $[-1, 1]$, while many of the backward values even reach values up to 5. It clearly shows that the backward covariance is biased away from zero and that the spread is higher.

4.7.1. Simulation studies (II): performance for different shapes of the noise

We investigate how the performance of slope-based inference (estimator (31)) with uniform reference measure is changed by the shape of the noise. We generate the data according to $Y = f(X) + E$. We use four distributions for P_E, which are the Gaussian, Laplacian (which is super-Gaussian), uniform (which is sub-Gaussian) distributions and a strongly sub-Gaussian distribution (represented by the mixture of Gaussians $0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$). Similarly, four distributions are used for P_X; they are the Gaussian and uniform distributions, a super-Gaussian distribution obtained by passing a Gaussian variable through the power-nonlinearity with exponent 1.5 and keeping the original sign, and a sub-Gaussian one represented by the mixture of Gaussians $0.5\mathcal{N}(-0.5,1) + 0.5\mathcal{N}(0.5,1)$. f has three different forms: $f(X) = X^{1/3}$, $f(X) = X^3$, and $f(X) = X$. Note that in the last case, f is not informative for causal inference at all in the noise-free case.

For each setting, we repeat the simulations 500 times. Figs. 11, 12, and 13 plot the performance as a function of the noise standard deviation in all possible cases of P_E and P_X, with the three forms for f, respectively. One can see that in the second columns of all these figures (corresponding to

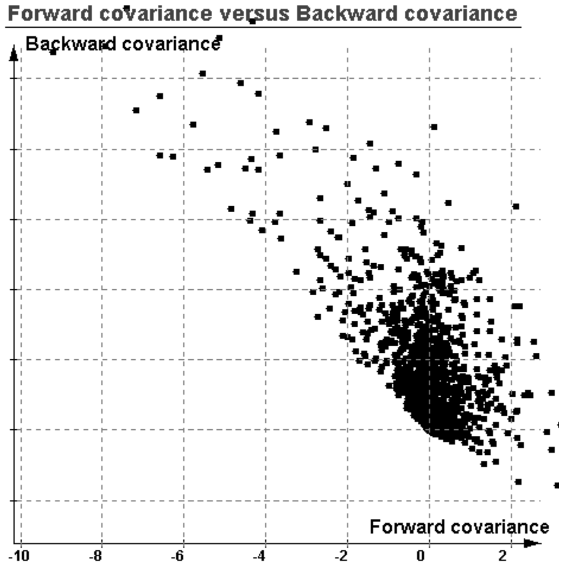

Figure 10: Empirical violations of orthogonality condition (h_3) in forward vs. backward direction (see text).

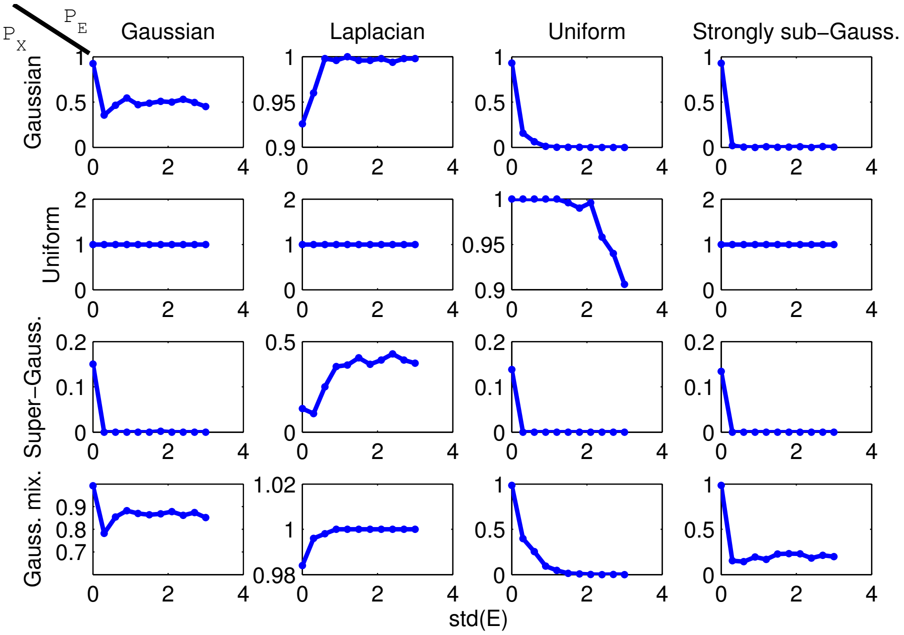

Figure 11: The performance (percentage of correct inferences) at different noise levels under various choices for P_E and P_X with the function $f(X) = X^{1/3}$. Columns from left to right correspond to Gaussian, Laplacian, uniform, and strongly sub-Gaussian (with $P(E) = 0.5\mathcal{N}(-2,1)+0.5\mathcal{N}(2,1)$) noise, respectively. Rows from top to bottom correspond to the Gaussian, uniform, super-Gaussian, and sub-Gaussian distributions for the cause X, respectively.

Laplacian noise) the performance increases along with the noise variance. In the third and fourth columns (corresponding to the uniform and strongly sub-Gaussian noise), the performance tends to become worse as the noise variance increases. As seen from Fig. 12, the function $f(X) = X^3$ is informative for causal inference: the performance is always good, almost regardless of different choices for P_E and P_X. When both P_E and P_X are Gaussian with $f(X) = X^{1/3}$ or $f(X) = X$ (top-left panels in Figs. 11 and 13), the decision is for high noise level like a random guess. Finally, when f is linear and thus not useful for causal inference in the deterministic setting (see Fig. 13), in certain combinations of P_E and P_X, IGCI still infers correctly due to the noise effect.

We then consider a fixed signal-to-noise ratio and change the shape of the noise continuously. To this end, we randomly generate i.i.d. samples for the

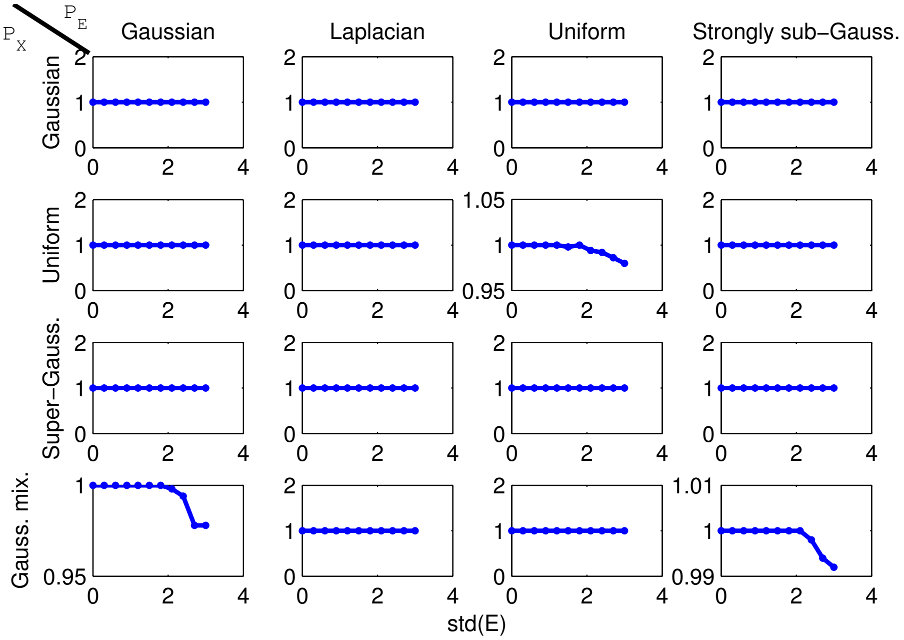

Figure 12: The performance (percentage of correct inferences) at different noise levels with the function $f(X) = X^3$; see the caption of Fig. 11.

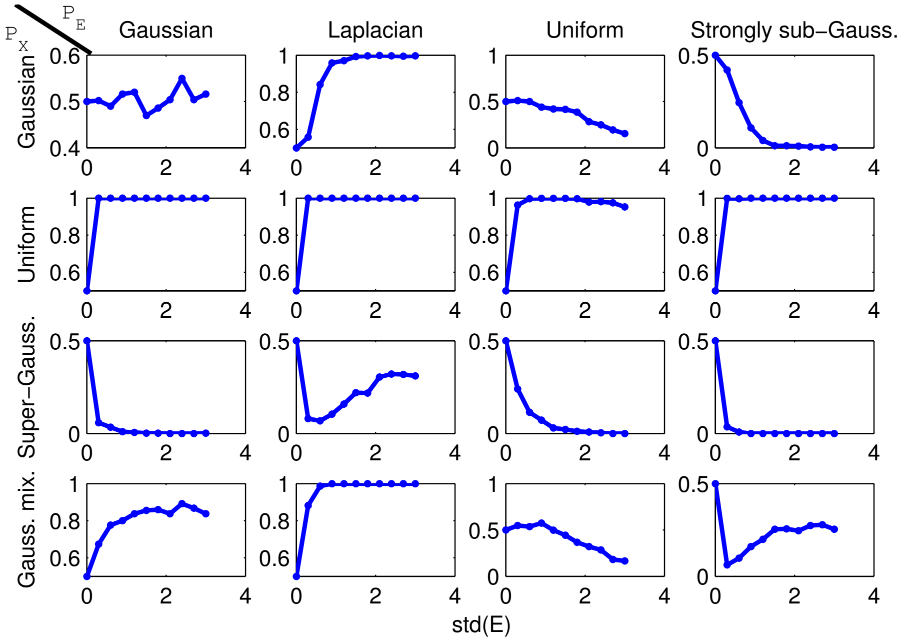

Figure 13: The performance (percentage of correct inferences) at different noise levels with the function $f(X) = X$; see the caption of Fig. 11.

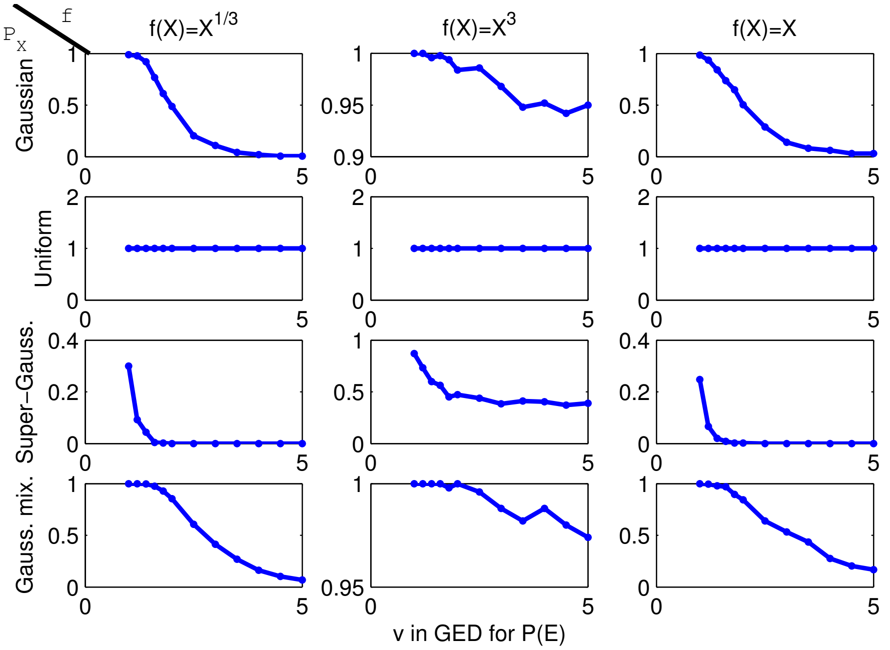

Figure 14: The performance (percentage of correct inferences) with different noise distributions (as indicated by v in (41)) for various P_X and f. The ratio of the noise standard deviation w.r.t. that of $f(X)$ is fixed to 2.

noise term E according to the zero-mean generalized exponential distribution (GED)

$$P(e) = \frac{v}{\sqrt{8}\Gamma(1/v)} \exp\left\{ - \left| \frac{e}{\sqrt{2}\sigma} \right|^v \right\}, \tag{41}$$

where v is the mode, $\Gamma(.)$ the gamma function, and σ the standard deviation. Sub-Gaussian, Gaussian and super-Gaussian noise are obtained for $v > 2$, $v = 2$, and $v < 2$, respectively. In particular, when $v = 1$, we get the Laplacian distribution. The uniform distribution is obtained via the limit $v \to \infty$. We use the ratio-of-uniform method [21] to generate the random numbers.

For various cases of P_X and f we vary v from 1 to 5 in (41), while the ratio of the standard deviation of the noise w.r.t. that of $f(X)$ is fixed to 2. Fig. 14 depicts the performance as a function of v. In all cases of P_X and f under consideration, the performance decreases or remains the same as v increases (i.e., as P_E becomes less super-Gaussian or more sub-Gaussian), which is consistent with the claims in Subsubsection 4.6.2.

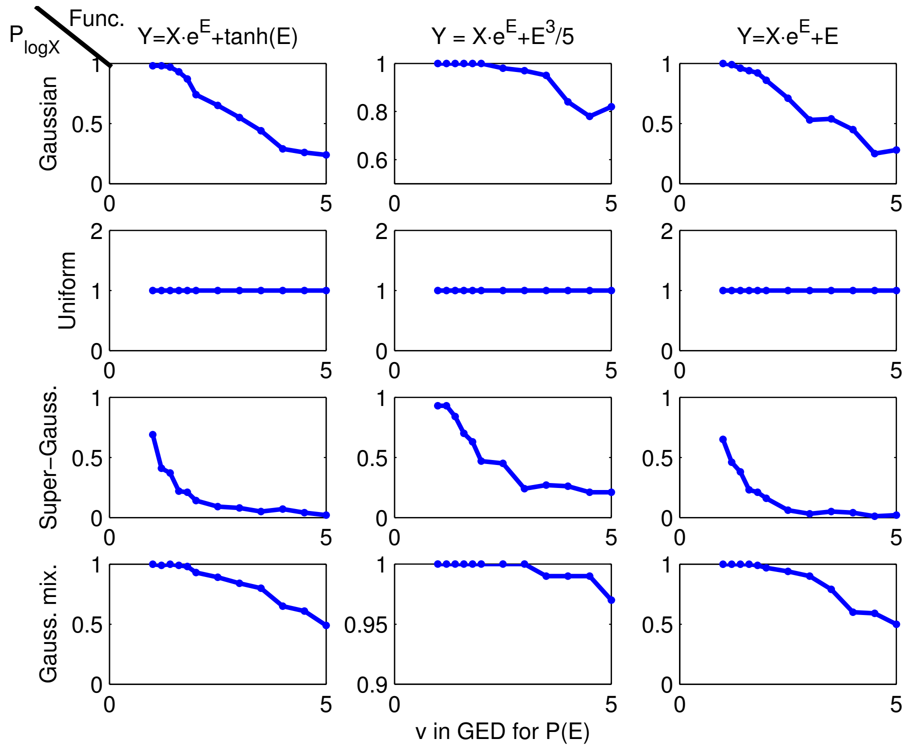

Figure 15: The performance of IGCI on the data generated by rather complex transformations with different noise distributions (as indicated by v in (41)) for various $P_{\log X}$ and transformations. Note that the y-axis labels correspond to the distribution of $\log X$. The variance of the noise E is fixed to 0.45.

As a more general setting, we repeat the above simulations with the data generated by $Y = X \cdot e^E + \tanh(E)$, $Y = X \cdot e^E + \frac{E^3}{5}$, and $Y = X \cdot e^E + E$, respectively; here Y is generated by a multiplicative block together with a nonlinear or linear effect of the noise E. The performance of IGCI as a function of v is given in Fig. 15. Again, as in Fig. 14, one can see that the performance decreases or remains the same as v increases.

Real-world data: Cause-effect pairs

We have also evaluated the IGCI method on real-world data, namely on the extended version of the "Cause-effect pairs" dataset described in [1]. This dataset consists of observations of 70 different pairs of variables from various domains, and the task for each pair is to find which variable is the cause and which variable the effect. For example, one of the pairs consists of 349 measurements of altitude and temperature taken at different weather stations in Germany. Obviously, the altitude is the cause and the temperature is the effect. The complete dataset and a more detailed description of each pair can be found at `http://webdav.tuebingen.mpg.de/cause-effect`. Note that most of the pairs in this data set have high noise levels, so that we do not necessarily expect our method to work well.

In Fig. 16 we show the results for the 70 pairs with the following four variants of IGCI: uniform distribution and Gaussians as reference measures, each case combined with the slope-based and the entropy-based estimator. The absolute value $|\hat{C}_{X\to Y}|$ was used as a heuristic confidence criterion. By taking only those decisions with high absolute value, one can trade off accuracy versus the amount of decisions taken. Fig. 16 shows the accuracy (i.e., the fraction of correct decisions) as a function of the decision rate (i.e., the fraction of decisions taken out of a total of 70 possible decisions, one for each cause-effect pair). If the absolute value of $\hat{C}_{X\to Y}$ was indeed a good measure of the confidence, one would expect that the accuracy is lowest for decision rate 100% (i.e., if all decisions are taken, regardless of the estimated confidence) and increases (more or less) monotonically as the decision rate decreases. A complication here is that the amount of data sets (cause-effect pairs) from which the accuracy can be estimated decreases proportionally to the decision rate. This means that the accuracies reported for low decision rates have higher uncertainty than the accuracies reported for high decision rates. For each decision rate, we have therefore indicated the 95% confidence interval that the accuracy is not significantly different from 50% by a grey area.

The four variants of IGCI yield comparable results. We also conclude that the majority of the decisions does agree with the causal ground truth, and that this agreement is statistically significant for high decision rates. However, the accuracy does not clearly increase with decreasing decision rates. This indicates that the heuristic confidence estimate (the absolute value of the estimated $\hat{C}_{X\to Y}$) is not functioning properly, although it is difficult to draw any final conclusions about this because of the high uncertainty in the accuracy for low decision rates. Nevertheless, considering the amount of noise that is present in many cause-effect pairs, it is surprising that our method works so well: if one always takes a decision, the four IGCI variants have accuracies of $70 \pm 7\%$, $75 \pm 7\%$, $69 \pm 7\%$, and $70 \pm 7\%$, respectively.

Fig. 17 provides comparative results of IGCI (now using only the variant based on (31) with a uniform reference measure) with four other causal

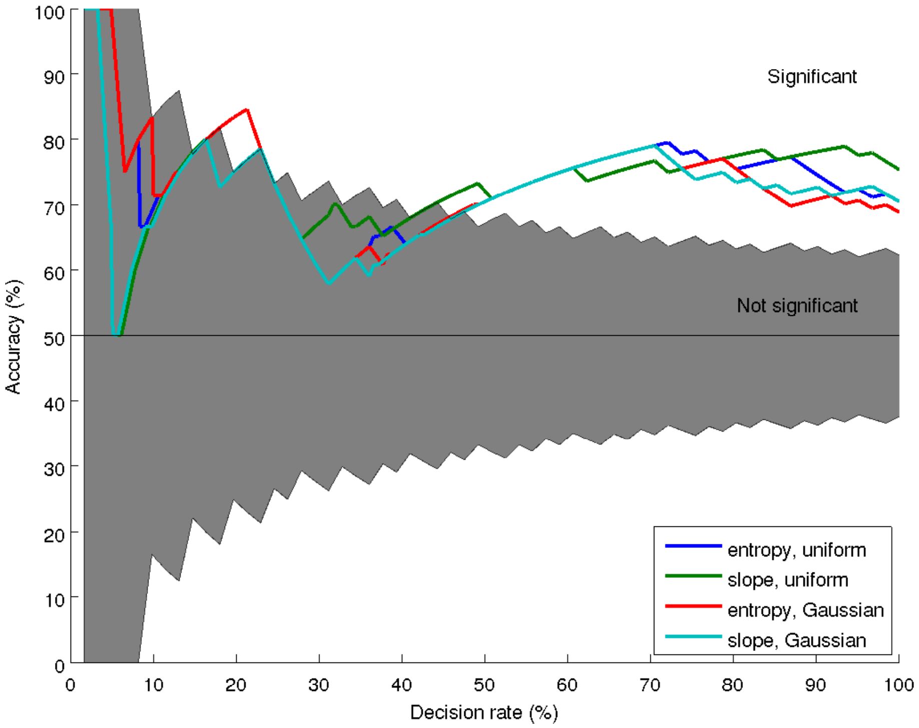

Figure 16: Results of four different implementations of IGCI on 70 cause-effect pairs.

inference methods that are suitable for inferring the causal direction between pairs of variables: LINGAM [22], Additive Noise (AN) [9], the Post-NonLinear (PNL) model [10], and a recent non-parametric method (GPI) [23]. All methods, except for GPI, employ the HSIC independence test [24] for accepting or rejecting the fitted models and use the maximum of the two HSIC p-values (where each p-value corresponds to a possible causal direction) as confidence estimate. The LINGAM method fits functional relationships of the form $Y = \alpha X + E$ to the data, preferring the causal direction for which the noise E is more independent of the hypothetical cause X. The additive noise based method (recall remarks around (1) and [9]) was implemented using the Gaussian Process regression code in the GPML toolbox [25] to find the most likely function f. For post-nonlinear model based inference

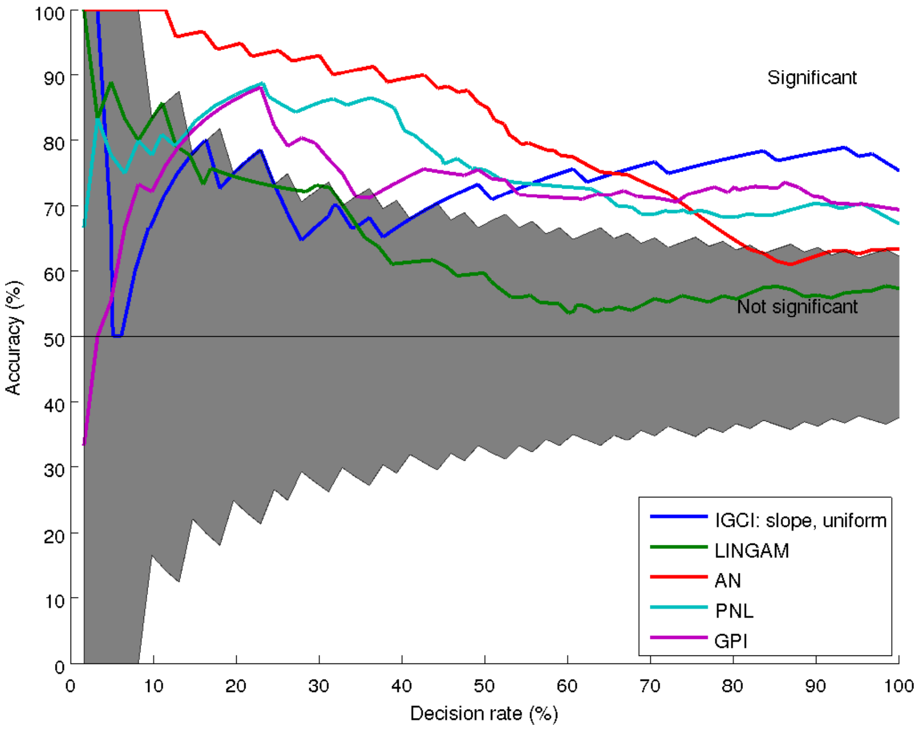

Figure 17: Results of various causal inference methods on 70 cause-effect pairs.

(2), we employed neural networks to model the functions f and h.[8] Finally, the non-parametric GPI method does not assume a particular class of functional relationships, but uses the general model $Y = f(X, E)$ and exploits the smoothness of the function f as one of the criteria for deciding upon the causal direction. For this method, the confidence value is taken to be the approximated Bayes factor between the two models corresponding with the two possible causal directions.

In contrast with the experiments reported in Fig. 16, we used at most 500 data points from each cause-effect pair, because most methods need sig-

[8] The large discrepancy between the results for PNL reported here and those reported in [14] is due to the fact that in [14], we applied a hand-tuned preprocessing method to each pair, whereas here we have treated all pairs equally by using the same preprocessing method for each pair.

nificantly more computation time than IGCI for large sample sizes. Note, however, that the performance of IGCI in this case is comparable with the performance reported in Fig. 16 where we used all data points. This can be explained because for many pairs, the measured values have been discretized, and therefore, the effective number of data points used by IGCI is usually lower than the number of available data points. We have repeated the experiments three times with different subsamples and plotted the average curves in Fig. 17. We observe that for high decision rates, all methods except LINGAM draw causal conclusions that are significantly correlated with the ground truth. IGCI, PNL and GPI yield comparable performances overall. The performance of the additive noise method seems to deviate from these three methods, because its accuracy is somewhat lower if it is forced to always make a decision, but on the other hand, its confidence estimate appears to be more accurate than that of the other methods, because the accuracy increases more quickly (even up to 100%) as the decision rate decreases. Again, it is difficult to draw any definite conclusions about the relative performances of these methods based on only 70 cause-effect pairs.

Real world data: water-levels of the Rhine

The data consists of the water levels of the Rhine[9] measured at 22 different cities in Germany in 15 minute intervals from 1990 to 2008. It is natural to expect that there is a causal relationship between the water levels at the different locations, where "upstream" levels influence "downstream" levels.

We tested our method on all 231 pairs of cities. Since the measurements are actually time series, and the causal influence needs some time to propagate, we performed the experiments with shifted time series, where for each pair of time series, one series was shifted relatively to the other so as to maximize the correlation between both.

Fig. 18 shows for each pair whether the decision is correct or not. It also shows some representative plots of the data. One clearly sees that the noise for two nearby cities is relatively low, but it can be quite large for two distant cities. Nevertheless, our method performed quite well in both situations: the overall accuracy, using the uniform reference measure, is 87% (201 correct decisions). The results for the Gaussian reference measure are similar (202

[9] We are grateful to the German office "Wasser- und Schiffahrtsverwaltung des Bundes", which provides the data upon request.

correct decisions).

4.8. Discussion

The assumption that $P_{\text{effect}|\text{cause}}$ and P_{cause} do not satisfy any “non-generic relation” can be a helpful paradigm for finding novel causal inference rules. Hence, one of the main important challenges consists in describing *what kind of* “non-generic” dependences typically occur in backward direction. A general answer to this question could not be given here, but we have shown that one option for defining dependences in an empirically testable way is given by orthogonality conditions in the sense of information geometry.

We have presented a method that is able to infer deterministic causal relations between variables with various domains. The accuracy of the proposed method was shown to be competitive with existing methods. In terms of computation time, this method is orders of magnitude faster (in particular, it is linear in the number of data points). In addition, it can handle the deterministic case, whereas existing methods only work in the presence of noise.

It would be desirable to have a reliable confidence criterion for our inference method. Moreover, we would like to point out again that in the large noise regime, the present method may completely fail. For a Gaussian reference measure in one dimension, for instance, our entropy-based version necessarily shows the wrong direction when the effect is given by a linear function of the cause plus an independent Gaussian noise. This is because then the effect is more Gaussian than the cause.

A generalization of the information geometric inference method to the case where the relation between cause and effect is not close to a bijective map is not straightforward. In Appendix Appendix A we discuss some toy examples showing that asymmetries between cause and effect can sometimes still be phrased in terms of information geometry.

Acknowledgements

Part of this work has been supported by Deutsche Forschungsgemeinschaft, SPP 1395.

References

[1] J. Mooij and D. Janzing. Distinguishing between cause and effect. *Journal of Machine Learning Research W&CP*, 6:147–156, 2010.

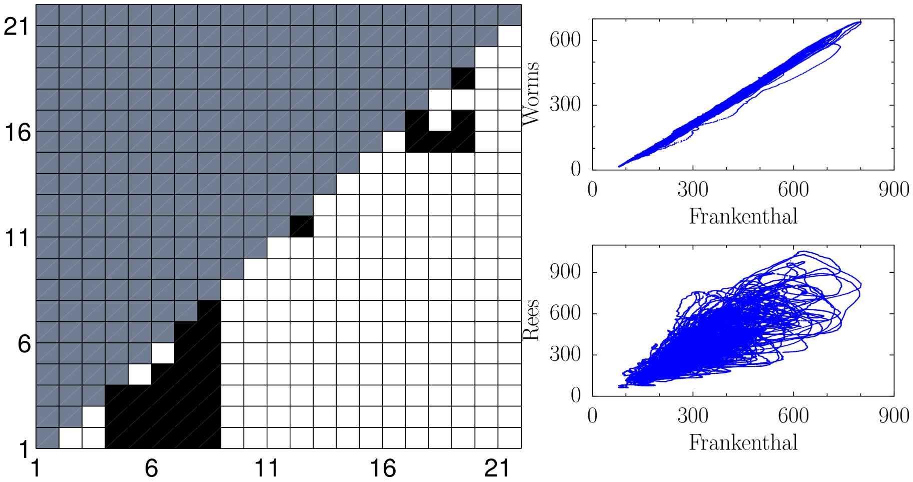

Figure 18: Results for the German Rhine data. All pairs out of in total 22 cities have been tested. White means a correct decision, black is a wrong decision, and the gray part can be ignored. On the right, typical data is illustrated for two measurement stations which are near to each other (top) and for two measurement stations farther apart (bottom), which shows that the noise increases significantly with the distance.

[2] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search (Lecture notes in statistics).* Springer-Verlag, New York, NY, 1993.

[3] J. Pearl. *Causality.* Cambridge University Press, 2000.

[4] J. Lemeire and E. Dirkx. Causal models as minimal descriptions of multivariate systems. `http://parallel.vub.ac.be/∼jan/`, 2006.

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

[6] R. Solomonoff. A formal theory of inductive inference. *Information and Control, Part II*, 7(2):224–254, 1964.

[7] G. Chaitin. A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.*, 22:329–340, 1975.

[8] A. Kolmogorov. Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, 1(1):1–7, 1965.

[9] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. In *Proceedings of the conference Neural Information Processing Systems (NIPS) 2008*, Vancouver, Canada, 2009. MIT Press.

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

[11] D. Janzing and B. Steudel. Justifying additive-noise-based causal discovery via algorithmic information theory. *Open Systems and Information Dynamics*, 17(2):189–212, 2010.

[12] D. Janzing, P. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. *Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel*, 06:479–486, 2010.

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

[14] P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 07:1–8, 2010.

[15] N. Friedman and I. Nachman. Gaussian process networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 211–219, Stanford, CA, USA, 2000. Morgan Kaufmann.

[16] S. Amari and H. Nagaoka. *Methods of Information Geometry.* Oxford University Press, 1993.

[17] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

[18] A. Kraskov, H. Stoegbauer, and P. Grassberger. Estimating Mutual Information. *http://arxiv.org/abs/cond-mat/0305641v1*, 2003.

[19] T. Cover and J. Thomas. *Elements of Information Theory.* Wileys Series in Telecommunications, New York, 1991.

[20] A. Dembo, Cover T., and Thomas J. Information theoretic inequalities. *IEEE Trans. Inform. Theory*, 37:1501–1518, 1991.

[21] A. J. Kinderman and J. F. Monahan. Computer generation of random variables using the ratio of uniform deviates. *ACM Transactions on Mathematical Software*, 3(3):257–260, 1977.

[22] 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.

[23] Joris M. Mooij, Oliver Stegle, Dominik Janzing, Kun Zhang, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems 23 (NIPS*2010)*, 2010. In press.

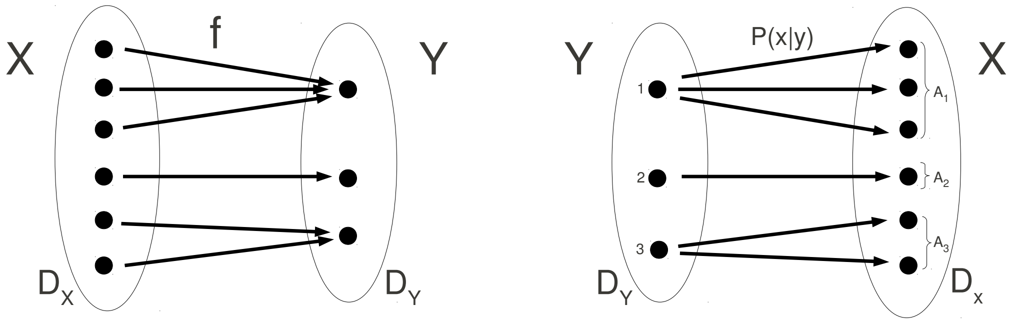

Figure A.19: Left: causal relation given by a deterministic non-injective map from cause to effect. Right: "splitting model", the cause can deterministically be inferred from the effect.

[24] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[25] C. E. Rasmussen and H. Nickisch. Gaussian Processes for Machine Learning (GPML) Toolbox. *Journal of Machine Learning Research*, accepted, 2010.

Appendix A. Outlook: Special cases of non-bijective relations

The following subsections provide a list of toy models, and explore under which conditions a violation of some of the orthogonality conditions can be shown for the backward direction. The models suggest that there is no straightforward extension of our IGCI method to the non-bijective case, although orthogonality could also help in identifying the causal direction.

Appendix A.1. One way deterministic

Let the ranges D_X and D_Y of X and Y, respectively, be finite and $P(X, Y)$ be a distribution for which Y is deterministically determined by X, i.e., $Y = f(X)$ for some surjective, but not necessarily injective function f (without surjectivity the backward model would not be defined), as in Fig. A.19. We show that the orthogonality conditions of Theorem 2 get simple for this case if U_X and U_Y are the uniform distributions on D_X and D_Y, respectively. First consider the orthogonalities that we expect if X causes Y:

Lemma 8 (orthogonalities for surjective functions, $X \to Y$). *Assume that Y is deterministically given by X.* $\mathrm{Cov}_{U_X}(h_i, P_X/U_X) = 0$ *holds trivially for $i = 1$. For $i = 2, 3$ it is equivalent to*

$$\mathrm{Cov}_{U_X}(\log m \circ f, P_X/U_X) = 0\,,$$

where

$$m(y) := |f^{-1}(y)| \tag{A.1}$$

denotes the number of pre-images of y.

Proof: Condition (h_1) is trivial since the function $x \mapsto D(P_{Y|x} \,||\, U_Y)$ is constant (because $P(y|x) = \delta_{y,f(x)}$ and $P_{Y|x}$ thus is a point measure). To rephrase condition (h_2), we first compute

$$\overrightarrow{P}(y) = \frac{m(y)}{|D_X|}\,,$$

and thus obtain

$$\begin{aligned} h_2(x) &= \sum_y \log \frac{P(y|x)}{\overrightarrow{P}(y)} P(y|x) \\ &= \sum_y \log \frac{\delta_{y,f(x)}}{m(y)} \delta_{y,f(x)} + c \\ &= -\log m(f(x)) + c \end{aligned}$$

with $c := \log |D_X|$. The constant term c is clearly irrelevant for the covariance. Since (h_1) is a constant function, uncorrelatedness between $h_3 = h_1 - h_2$ and P_X/U_X is obviously equivalent to uncorrelatedness between h_2 and P_X/U_X. $\Box$

The following Lemma describes the relations that we expect for the same condition $Y = f(X)$ if Y causes X, as in the "splitting model" in Fig. A.19 (right) [10]. The causal relation is now given by a mechanism that splits every y-value into different x-values in the set A_y such that the mapping from x to y is deterministic.

[10] Note that this is the only case in this paper where Y is the cause. The reason is that we want to compare the properties of $P_{X,Y}$ that we expect for the two possible causal directions.

Lemma 9 (orthogonalities for splitting model, $Y \to X$).
Assume again that Y is deterministically given by X (although we now assume Y to be the cause). For the functions $y \mapsto h_i(y)$ the equation $Cov_{U_Y}(h_i, P_Y/U_Y) = 0$ is trivial for $i = 2$ and for $i = 1, 3$ equivalent to

$$\mathrm{Cov}_{U_Y}(S(P_{X|y}), \frac{P(y)}{U(y)}) = 0 \,.$$

By slightly abusing notation, $S(P_{X|y})$ and $P(y)/U(y)$ denote the functions $y \mapsto S(P_{X|y})$ and $y \mapsto P(y)/U(y)$, respectively.

Proof: We first compute

$$P(x|y) = \delta_{y,f(x)} \frac{P_X(x)}{P_Y(f(x))} \,.$$

To rephrase condition (h_2), we compute

$$\overleftarrow{P}(x) = \frac{1}{|D_Y|} \sum_y P(x|y) = \frac{P_X(x)}{|D_Y| P_Y(f(x))} \,.$$

We thus obtain

$$h_2(y) \;=\; \sum_x \log \frac{P(x|y)}{\overleftarrow{P}(x)} P(x|y) = \log |D_Y| \,.$$

Therefore, condition (h_2) becomes trivial.

To reformulate condition (h_1), we observe that $h_1(y) = D(P_{X|y} \,||\, U_X)$ is, up to a sign and and additive constant, given by $S(P_{X|y})$. Since h_2 is a constant, condition (h_3) is equivalent to (h_1). $\Box$

These results show that we obtain reasonable conditions for both directions: if X is the cause, we get uncorrelatedness between input and the logarithm of the number of pre-images. On the other hand, if Y is the cause, we postulate zero correlation between input and conditional entropy. Unfortunately, the orthogonality in one direction does not imply the violation of orthogonality for the other direction. Moreover, the violation of orthogonality in backward direction can have both positive and negative sign. This is shown by the following example.

Example 3 (no definite violation in backward direction).
Let f be such that the number m of pre-images (see (A.1)) is constant. Then all non-trivial conditions of Lemma 8 are satisfied for the hypothesis $X \to Y$ but $y \mapsto S(P_{X|y})$ can be positively or negatively correlated or uncorrelated with P_Y/U_Y. To see this, set $D_Y = \{1,2\}$ and $D_X = \{1,2,3,4\}$ and $P(x|y = 1) = 1/2$ for each $x = 1,2$. On the other hand, $P(x|y = 2) = 1$ for $x = 3$. Then $S(P_{X|y=1}) = \log 2$ and $S(P_{X|y=2}) = 0$. Depending on whether $P(y = 1)$ is greater or smaller than $P(y = 2)$ we can induce positive or negative correlations.

Conversely, assume that $S(P_{X|y})$ is uncorrelated with $P(y)/U(y)$ and all our orthogonality conditions would therefore be consistent with the hypothesis $Y \to X$. To see that then $\log m(f(x))$ can nevertheless be negatively or positively correlated with $P(x)/U(x)$, we consider the following example. Set $D_X = \{1,2,3,4,5\}$ and $D_Y = \{1,2\}$. Let $P_{X|y=1}$ be the uniform distribution on the set $\{1,2\}$ and let $P_{X|y=2}$ be some distribution on $\{3,4,5\}$ that has also the entropy 1 bit. Then $S(P_{X|y})$ is constant in y and thus uncorrelated with $P(y)/U(y)$ and we can design $P(y)$ as we like. To check whether $\log m(f(x))$ is positively or negatively correlated with $P(x)/U(x)$ we observe

$$\begin{aligned}\sum_x \log m(f(x))(P(x) - U(x)) &= \log 2 \left(P_Y(1) - \frac{2}{5}\right) + \log 3 \left(P_Y(2) - \frac{3}{5}\right) \\ &= (\log 2 - \log 3) \left(P_Y(1) - \frac{2}{5}\right) ,\end{aligned}$$

which is positive for $P_Y(1) < 2/5$ and negative for $P_Y(1) > 2/5$.

This result is a bit disappointing at first glance since it questions the information geometric method for the non-bijective case: if violations of orthogonality for the backward direction occur with both possible signs, decision rules get less simple; preferring the direction for which the violation of orthogonality is smaller with respect to its *absolute value* seems less natural than inference rules that work without absolute value. It could therefore be that notions of independence other than our orthogonality conditions are needed. To support this conjecture, we should also mention that in designing P_Y and $P_{X|Y}$ in the above example we have in fact adjusted them to each other, we only did it in a way that is not captured by our orthogonality conditions.

There is, however, the following nice result:

Lemma 10 (number of pre-images and input probability).
For $Y = f(X)$, let $m(y)$ be the number of pre-images of y. If $\log m$ is uncorrelated with P_Y/U_Y then $\log m \circ f$ is negatively correlated with P_X/U_X. On the other hand, if $\log m \circ f$ is uncorrelated with P_X/U_X, then $\log m$ is positively correlated with P_Y/U_Y:

$$\begin{aligned} \mathrm{Cov}_{U_X}\left(\log m \circ f, \frac{P_X}{U_X}\right) &= \mathrm{Cov}_{U_Y}\left(\log m, \frac{P_Y}{U_Y}\right) \\ &- D\left(U_Y \,||\, \frac{m}{|D_X|}\right) - D\left(\frac{m}{|D_X|} \,||\, U_Y\right) . \end{aligned} \tag{A.2}$$

Proof:

$$\begin{aligned} \mathrm{Cov}_{U_X}\left(\log m \circ f, \frac{P_X}{U_X}\right) &= \sum_x \log m(f(x))(P(x) - U(x)) && \text{(A.3)} \\ &= \sum_y \log m(y)\left(P(y) - \frac{m(y)}{|D_X|}\right) \\ &= \sum_y \log m(y)\left(P(y) - U(y) + U(y) - \frac{m(y)}{|D_X|}\right) \\ &= \mathrm{Cov}_{U_Y}\left(\log m, \frac{P_Y}{U_Y}\right) && \text{(A.4)} \\ &- D\left(U_Y \,||\, \frac{m}{|D_X|}\right) - D\left(\frac{m}{|D_X|} \,||\, U_Y\right) . && \text{(A.5)} \end{aligned}$$

□

The term (A.2) measures to what extent m is non-constant. Since $m(y)/|D_X|$ coincides with $\overrightarrow{P}(y)$ this is again our well-known expression (22). Note that the correlations between m and P_Y/U_Y is positive if Y is the effect, while the correlation between $m \circ f$ and P_X/U_X is negative if X is the effect. Here, the different sign of the correlation may seem disturbing. However, in the following special case it turns out to be natural:

Example 4 (all pre-images are equally likely).
For both causal directions $X \to Y$ and $Y \to X$ assume

$$P(x|y) = \frac{\delta_{y,f(x)}}{m(y)} . \tag{A.6}$$

If $Y \to X$ this is not unlikely to occur, because it only means to divide the probability uniformly over all pre-images x of a given y. For $X \to Y$ it can, for instance, occur if P_X is uniform.[11]

We obtain

$$h_3(x) = \sum_y \log \frac{\overrightarrow{P}(y)}{U(y)} P(y|x) = \log \frac{m(f(x))\,|D_Y|}{|D_X|}\,,$$

and

$$h_3(y) = \sum_x \log \frac{\overleftarrow{P}(x)}{U(x)} P(x|y) = \log \frac{|D_X|}{m(y)\,|D_Y|}\,.$$

Therefore, $h_s(x)$ and $h_3(y)$ are, up to irrelevant constants, given by $m(f(x))$ and $-m(y)$, respectively. Hence, Lemma 10 implies that $h_3(y)$ is negatively correlated with $P(y)/U(y)$ if $h_3(x)$ is uncorrelated with $P(x)/U(x)$ and vice versa, which nicely fits into our information geometric framework.

Note, moreover, that $\log m(y)$ coincides with $S(P_{X|y})$ up to a constant (and hence also with $D(P_{X|y}\|U_X)$ up to a sign and a constant). Therefore, uncorrelatedness between $\log m$ and P_Y/U_Y is equivalent to orthogonality condition (h_1) in Theorem 1.

Appendix A.2. Functional relation with small independent noise

In this subsection we revisit the motivating remarks in Section 2 in a more precise way and describe how they fit into our information geometric framework. Consider a so-called additive noise model

$$Y = f(X) + E \text{ with } E \perp\!\!\!\perp X\,.$$

Let f be a bijection of $[0,1]$ and E have compact support $[0,\epsilon]$. Let P_X have support $[0,1]$, the support of Y is thus given by $[0,1+\epsilon]$. By adapting the arguments of Example 1 to the uniform distribution on $[0,1+\epsilon]$ instead of $[0,1]$ one checks easily that orthogonality condition (h_1) is equivalent to uncorelatedness between $S(P_{Y|x})$ and $P(x)$, which holds because $S(P_{Y|x})$ attains the constant value $S(E)$. We now assume that ϵ is so small compared to the curvature of f and the scale of the fluctuations of $P(x)$ that the

[11] If $P(x)$ attains *many* different values, it is, however, unlikely that it always attains the same value within the same A_y.

conditional distribution $P_{X|y}$ is approximately given by the distribution of $f^{-1'}(y)E$, shifted by some y-dependent value. We thus assume

$$S(P_{X|y}) \approx S(E) + \log f^{-1'}(y)\,, \tag{A.7}$$

for all y that are not too close to the boundaries of the interval $[0,1]$. For the backward direction condition, the covariance (h_1) therefore reads

$$\mathrm{Cov}_{U_Y}\left(S(P_{X|y}), \frac{P(y)}{U(y)}\right) = \int_0^{1+\epsilon} S(P_{X|y})(P(y)-U(y))dy \tag{A.8}$$

$$\approx \int_0^1 \log f^{-1'}(y)(P(y)-U(y))dy \tag{A.9}$$

where we have not only used the approximation (A.7) but also neglected the fact that $S(X|y)$ actually has to be integrated over $[0,1+\epsilon]$ rather than $[0,1]$ since the errors are all of order ϵ. Expression (A.9) is positive for small ϵ because in the deterministic limit $\epsilon \to 0$, (A.9) can be transformed into

$$-\int_0^1 \log f'(x)(P(x)-f'(x))dx = D(P_X\,||\,f') + D(f'\,||\,P_X) \geq 0\,, \tag{A.10}$$

where we interpret f' as probability density (due to $f(1)=1$ and $f(0)=0$). Note that in the deterministic invertible case we have $f'(x) = \overleftarrow{P}(x)$ and (A.10) is again a symmetrized relative entropy term. This result shows that additive noise models (in the low noise regime) induce backward models for which the noise depends on the input in a way that leads to violation of orthogonality condition (h_1). It is remarkable that the amount of violation is here described by a term that is similar to the one that occurred in the bijective as well as in the case of Appendix Appendix A.1 even though these cases refer to orthogonality (h_3). This suggests that there is a common principle behind our observations.

Appendix B. Justification of Postulate 2

For single reference densities instead of manifolds we have justified conditions (h_1) to (h_3) by the argument that the structure functions h should not correlate with P_X because they only depend on the conditional $P_{Y|X}$ (i.e., the function f in our case). This justification is not completely convincing if we generalize the setting to manifolds: the functions h_1 and h_3 contain the reference density U_Y, which is defined by projecting the "output" probability P_Y onto $\mathcal{E}_Y$. U_Y thus depends on both P_X and f because P_Y is the image of P_X under f. We therefore justify Postulate 2 in a slightly different way.

Our justification remains quite informal; to provide a more precise version of the below arguments would go beyond the scope of this paper. We start with the following statement:

Observation 1 (projection of random moves).
Let $\mathcal{P}$ be the set of probability distributions over some large (or infinite) probability space. Let $\mathcal{E} \subset \mathcal{P}$ be a low-dimensional exponential manifold and $P \in \mathcal{P}$ be arbitrary. Generate a new point R by moving into some random direction from P, chosen independently of $\mathcal{E}$. Denoting the projections of P and R on $\mathcal{E}$ by $P_{\mathcal{E}}$ and $R_{\mathcal{E}}$, respectively, we obtain for a typical move

$$P_{\mathcal{E}} \approx R_{\mathcal{E}} \,.$$

The approximate equality means that the error made by replacing one point with the other in any relative entropy expression is small compared to $D(P \,||\, R)$ and $D(R \,||\, P)$.

Apart from the approximate equality signs, the statement is also informal by not specifying what a "typical move" means. This would require a probability distribution on the set of possible moves.

Assume now that the pair (P_X, f) is generated as follows. Let $U_X \in \mathcal{E}_X$ be given and obtain P_X by modifying U_X according to some random move. Generate f independently of $\mathcal{E}_X$ and $\mathcal{E}_Y$. We can assume that U_X is the projection of P_X onto $\mathcal{E}_X$ without seriously restricting the random moves because this assumption approximates the typical case. This is seen by applying Observation 1 to the special case $P \in \mathcal{E}$. Let U_Y, as usual, be the projection of P_Y onto $\mathcal{E}_Y$ and W be the projection of $\overrightarrow{P}_Y$ onto $\mathcal{E}_Y$. We now apply Observation 1 and consider P_Y as obtained from $\overrightarrow{P}_Y$ by a random move. This is justified because it is just the map of the move from U_X to P_X under f. Since f and this move have been chosen independently of the manifold $\mathcal{E}_Y$, Observation 1 states

$$W \approx U_Y \,. \tag{B.1}$$

Applying f^{-1} to both sides yields

$$W_f \approx \overleftarrow{P}_X \,, \tag{B.2}$$

where W_f denotes the image of W under f^{-1}. Note that W_f is the point in $f^{-1}(\mathcal{E}_Y)$ that is the closest to U_X because W is the point in $\mathcal{E}_Y$ that is the closest to the image of U_X under f. In the typical case we expect

$$D(P_X \,||\, W_f) \approx D(P_X \,||\, U_X) + D(U_X \,||\, W_f)\,,$$

because the vector connecting U_X with W_f does not depend on P_X, it only depends on f and the manifolds. The vector pointing from U_X to P_X is therefore typically close to orthogonal to the one pointing from U_X to W_f. Together with (B.2) we thus obtain

$$D(P_X \,||\, \overleftarrow{P}_X) \approx D(P_X \,||\, U_X) + D(U_X \,||\, \overleftarrow{P}_X)\,,$$

which is one of the equivalent conditions in Theorem 2.

In Subsection 4.4 we have already mentioned that in the special case of linear relations (32) between high-dimensional Gaussian variables (with isotropic Gaussians as reference manifold), Postulate 2 can be further justified by concentration of measure results. It is instructive to verify that also (B.1) holds for this case. To see this, we recall that U_Y has the covariance matrix

$$\tau(\Sigma_Y)\,\mathbf{I} = \tau(A\Sigma_X A^T)\,\mathbf{I}\,,$$

which is approximately equal to

$$\tau(\Sigma_X)\,\tau(AA^T)\,\mathbf{I}\,,$$

(see Subsection 4.4 and [12]). One checks easily that the latter is the covariance matrix of W, i.e., the isotropic Gaussian that is closest to $\overrightarrow{P}_Y$. Using the notations above, this shows (B.1).