# Generative causality: using Shannon's information theory to detect underlying asymmetry in causal relations

Soumik Purkayastha and Peter X.-K. Song*

*Department of Biostatistics, University of Michigan, Ann Arbor, USA.*

E-mail: *pxsong@umich.edu

**Summary**. Causal investigations in observational studies pose a great challenge in scientific research where randomized trials or intervention-based studies are not feasible. Leveraging Shannon's seminal work on information theory, we consider a framework of *asymmetry* where any causal link between putative cause and effect must be explained through a mechanism governing the cause as well as a generative process yielding an effect of the cause. Under weak assumptions, this framework enables the assessment of whether X is a stronger predictor of Y or vice-versa. Under stronger identifiability assumptions our framework is able to distinguish between cause and effect using observational data. We establish key statistical properties of this framework. Our proposed methodology relies on scalable non-parametric density estimation using fast Fourier transformation. The resulting estimation method is manyfold faster than the classical bandwidth-based density estimation while maintaining comparable mean integrated squared error rates. We investigate key asymptotic properties of our methodology and introduce a data-splitting technique to facilitate inference. The key attraction of our framework is its inference toolkit, which allows researchers to quantify uncertainty in causal discovery findings. We illustrate the performance of our methodology through simulation studies as well as multiple real data examples.

**Keywords:** cross-fitting, entropy, information theory, asymmetry.

## 1. Introduction

Unravelling causal relationships from observational data is a key aspect of scientific inquiry. For the pair of random variables $(X, Y)$, a fundamental question of interest is to determine whether an observed statistical dependence may be explained by a causal influence of either $X$ on $Y$ or vice-versa. Observed statistical dependence can arise from either direct causation between $X$ and $Y$, unobserved common causes ("confounding", see e.g., Pearl et al. (2000)), conditioned common effects ("selection bias", see e.g., Pearl et al. (2000)), or a combination of these. Even under simplifying assumptions of no confounding, no feedback loops, and no

selection bias, directly assessing pairwise causal relationships is a notoriously hard problem (Spirtes and Zhang, 2016).

Although the gold standard for detecting causality involves controlled experimentation, in many cases, such trials are prohibitively expensive, technically challenging, or even unethical (Mooij et al., 2016). Hence, the development of methods that can identify causal relations from purely observational data is of great importance and has attracted increasing interest recently (see e.g., Hoyer et al. (2008); Zhang and Hyvärinen (2009); Hernandez-Lobato et al. (2016); Choi et al. (2020); Tagasovska et al. (2020)). All these approaches exploit the complexity of the marginal and conditional probability distributions, in one way or the other. On an intuitive level, the idea is that the factorization of the joint density $f_{XY}$ of cause $X$ and effect $Y$ into the product of the marginal $f_X$ and the conditional $f_{Y|X}$ yields models of "lower complexity" than the alternative factorization. In other words, these approaches exploit *asymmetries* between $X$ and $Y$ on a distributional level. Asymmetry, hence, may be deemed as an informative reflection of underlying causality (Janzing et al., 2012).

Our understanding of causality is somewhat different from the framework of potential outcomes (Rubin, 2005), which relies on some fundamental assumptions (Pearl, 2010; Robins et al., 2000) that are often unverifiable. In studying asymmetry, we rely on an understanding of causation as a generative process. According to Cox (1990, 1992), the elaboration of an underlying generative process (Engelhardt et al., 2009) is crucial to any claim of causality. In other words, a causal link between $X$ and $Y$ must be explained through (a) a mechanism governing the cause and (b) a generative process yielding an effect of the cause. In an attempt to study such *generative causality*, we consider the following *additive noise model (ANM)*:

$$Y = g(X) + \epsilon, \tag{1}$$

where $X$ is the cause, governed by a density function $f_X$ and $g$ is the generative function that yields the effect $Y$. Further, $\epsilon$ is an additive noise that is statistically independent of $X$. Hoyer et al. (2008) described a potential asymmetry between cause and effect in terms of statistical independence between $X$ and $\epsilon$ that occurs in the causal mechanism described by (1). The authors show that if the above *additive noise model (ANM)* holds, then there is usually (up to some exceptions like the bivariate Gaussian case) no converse ANM from the effect $Y$ to $X$ Janzing et al. (2012). In other words, writing $X = \tilde{g}(Y) + \tilde{\epsilon}$ for some function $\tilde{g}$ will not render the residual term $\tilde{\epsilon}$ statistically independent of $Y$. Zhang and Hyvärinen (2009) expanded the

framework to

$$Y = h(g(X) + \epsilon), \quad \epsilon \perp X, \tag{2}$$

and show that such a *post-nonlinear model (PNM)* also exists in at most one direction, except for some special cases (Janzing et al., 2012). This article develops an information geometric framework for inferring causal direction that does not require the restricted class of ANMs or PNMs.

In Section 2.2, we propose a framework of *weak asymmetry* whereby a comparison of (differential) entropies sheds light on whether predicting $Y$ given knowledge of $X$ involves less uncertainty than the reverse. Next, in Section 2.3, we note how conditional entropies can also detect dependence between $f_{X|Y}$ and $f_Y$ that can occur if $X$ causes $Y$. We formalize this finding in Section 3, where we establish the framework of *strong asymmetry*. In this framework, when $X$ cause $Y$ via a bijective function $g$, i.e., $Y = g(X)$, the difference in entropy of $X$ and $Y$ is always positive, subject to some identifiability assumptions. This yields an asymmetry between cause and effect. Our strong asymmetry framework aligns with the causal discovery framework named *information geometric causal inference (IGCI)* proposed by Janzing et al. (2012). In Section 4 we propose modifications to our framework of asymmetry when accommodating confounders. Next, in Section 5 we leverage a fast Fourier transformation-based density estimation technique to obtain the causal estimands of interest. Furthermore, we introduce a data-splitting and cross-fitting technique to control bias while improving statistical efficiency. Finally, in Sections 6 and 7 we examine the performance of causal estimands through empirical studies and real data applications respectively.

## 2. Studying weak asymmetry through entropy

Arguably, the key problems in pairwise causal discovery lie in inferring (i) whether $X$ and $Y$ are associated and (ii) whether $X \Rightarrow Y$ (i.e., $X$ causes $Y$) from observations $\{(x_i, y_i)\}_{i=1}^n$. We leverage information-theoretic notions of *mutual information (MI)* and *entropy* to answer (i) and (ii) respectively. Through studying association using $MI$, we rule out the case where $X \perp Y$ (i.e., $X$ and $Y$ are independent). Next, within the weak asymmetry framework, we note how a comparison of the entropy of $X$ and $Y$ unearths whether $X$ is a better predictor of $Y$ or if the converse is true.

### 2.1. Review of mutual information (MI) and entropy

Let $X$ and $Y$ be two random variables with joint density function $f_{XY}$. Let $f_X$ and $f_Y$ be the marginal densities of $X$ and $Y$, respectively. Further, let $\mathcal{X}$ and $\mathcal{Y}$ denote the respective support sets of $X$ and $Y$. The mutual information (Shannon, 1948) is:

$$MI(X,Y) = E_{XY}\left\{\log \frac{f_{XY}(X,Y)}{f_X(X)f_Y(Y)}\right\}, \tag{3}$$

where $E_{XY}$ denotes expectation over $f_{XY}$. Some properties that make $MI$ an attractive measure of complex dependence include (i) $MI \geq 0$ with equality if and only if $X$ and $Y$ are independent and (ii) a larger value of $MI$ indicates a stronger dependence between two variables. However, it must be noted that $MI$ is a symmetric measure, and hence unable to quantify any sort of asymmetry between cause and effect in our setting. This drawback is shared by most popular measures of association such as the distance correlation of Székely et al. (2007), the Heller–Heller–Gorfine (HHG) test of Heller et al. (2012), the Hilbert-Schmidt independence criterion (HSIC) test of Gretton et al. (2007). In contrast, a new correlation coefficient proposed by Chatterjee (2020) is asymmetric, but unfortunately does not carry any causal interpretation. In Purkayastha and Song (2023), we develop a fast and consistent estimator of mutual information, called `fastMI`, which provides a powerful test for independence of $X$ and $Y$ that exhibits satisfactory type I error control.

The *joint entropy* of $(X,Y)$ and *marginal entropy* of $X$ are respectively given by

$$\begin{aligned} H(X,Y) &= E_{XY}\left\{-\log f_{XY}(X,Y)\right\}, \\ H(X) &= E_X\left\{-\log f_X(X)\right\}. \end{aligned} \tag{4}$$

Differential entropy measures the randomness of a random variable (Orlitsky, 2003). Further, let $f_{Y|x}$ denote the distribution of $Y \mid X = x$. We define the *conditional entropy function* of $Y \mid X = x$ as:

$$H(Y \mid X = x) = \int_{y\in\mathcal{Y}} -\log\left(f_{Y|x}(y)\right) f_{Y|x}(y)dy \tag{5}$$

Averaging the above function over all values of $x \in \mathcal{X}$, we get the *conditional entropy* $H(Y \mid X)$, which is related to the marginal and joint entropy terms defined in (4) through the following

chain rule:

$$H(X,Y) = H(Y \mid X) + H(X), \tag{6}$$

which may be interpreted as saying that uncertainty about $X$ and $Y$ may be decomposed into uncertainty about $X$ and uncertainty about $Y$ given $X$. Further, it can be shown that $H(Y \mid X) \leq H(Y)$ with equality iff $X \perp Y$. Therefore knowing $X$ never increases entropy of $Y$, and, except when it is irrelevant (i.e., $X$ and $Y$ independent), it always lowers entropy (Orlitsky, 2003). For more details on these information-theoretic terms, please see Cover and Thomas (2005).

### *2.2. Studying weak asymmetry through entropy*

The quantities defined in (4) - (6) and are related to each other through the following decomposition of total entropy:

$$H(X,Y) = H(X \mid Y) + H(Y \mid X) + MI(X,Y). \tag{7}$$

Since both $MI$ as well as $H(X,Y)$ are symmetric measures, we can only rely on comparing the conditional entropy terms to study asymmetry in $(X,Y)$. This approach is supported by our understanding that conditional entropy $H(Y \mid X)$ measures the uncertainty about $Y$ given knowledge of $X$. Observe how $H(X \mid Y) > H(X \mid Y)$ implies there is less uncertainty in $Y$ after conditioning on $X$, rather than the converse.

If $X \Rightarrow Y$, it is reasonable to posit predicting $Y$ given knowledge of $X$ involves less uncertainty than the reverse, thereby giving rise to a potential *weak asymmetry* between cause and effect. Said asymmetry may be studied by comparing the conditional entropies $H(X \mid Y)$ and $H(Y \mid X)$. Futher, from the chain rule given by (6), we note

$$H(X \mid Y) - H(Y \mid X) = H(X) - H(Y), \tag{8}$$

thereby establishing an equivalence between the comparison of marginal entropies and conditional entropies. In the next section, we review ANMs and note how entropies can play a key role in describing asymmetry that can occur if $X \Rightarrow Y$.

### *2.3. Entropy-based understanding of additive noise models (ANMs)*

In the ANM (Hoyer et al., 2008), we note how independence of the error term and cause ensures that the conditional entropy function $H(Y \mid X = x)$ is constant in $x \in \mathcal{X}$ and coincides with the conditional entropy $H(Y \mid X)$. For ease of exposition, let us first assume $f_X$ is uniform and that $g$ in (1) is invertible. Then, for low noise $\epsilon$, we have $f_Y(y) \approx f_X\left(g^{-1}(y)\right) \cdot \left|\nabla g^{-1}(y)\right|$, where the $\nabla$ operator denotes the derivative of a function with respect to its argument. This implies the density of $Y$ is high whenever $\left|\nabla g^{-1}(y)\right|$ is large. At the same time, as seen in

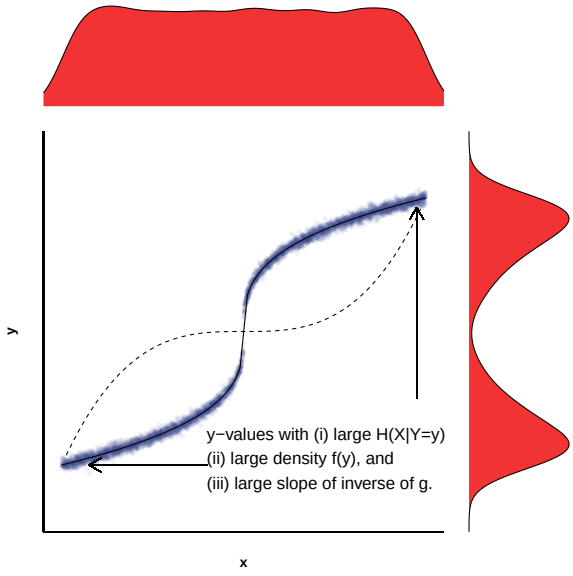


**Fig. 1.** Additive noise model (Hoyer et al., 2008) in low noise regime. We draw observations from $X \sim U(0,1)$ and generate $Y = g(X) + \epsilon$, where $g$ is a bijective function denoted by the solid line and $\epsilon$ is an error term independent of $X$. The uniform density of $X$ is reflected by its kernel density plot (in red, above the scatterplot). Note how peaks of the estimated kernel density plot of $Y$ (in red, to the right of the scatterplot) coincide with regions with a large slope of $g^{-1}(y)$. The conditional entropy function $H(X|y)$ is also found to be high at regions with a large slope of $g^{-1}(y)$ (dashed line), i.e., a small slope of $g$ (solid line). Image motivated by an example from Janzing et al. (2012).

Figure 1, the entropy function $H(X \mid Y = y)$ is also high for regions with large $\left|\nabla g^{-1}(y)\right|$. In other words, the entropy function $H(X \mid Y = y)$ is correlated with high density $f_Y$, assuming a uniform distribution of $X$ in a low-noise regime of model given by (1) where the link function $g$ is invertible. Further, if $f_X$ is not uniform, high values of $f_Y$ occur when both $\left|\nabla g^{-1}(y)\right|$ as well as $f_X\left(g^{-1}(y)\right)$ are simultaneously large. In a situation where the peaks of $f_X$ do not correlate with the slope of $g$, the heuristic argument above still holds, thereby yielding a correlation between $H(X \mid Y = y)$ and $f_Y(y)$. In the next section, we highlight key issues of working with either ANM or PNM and describe an information geometric perspective that results in an inference method for detecting causal relations.

## 3. Studying strong asymmetry through entropy

Since both the ANM as well as PNM approaches rely on assumptions involving the statistical independence of the noise, the special case where $X \Rightarrow Y$ via a deterministic function $g$ poses greater challenges in terms of inferring a direction of causation. First, the absence of noise makes additive noise model-based inference impossible (Hoyer et al., 2008), and second, methods that use non-invertibility of the functional relation fail (Friedman and Nachman, 2000). If $g$ is non-invertible, one may consider the task solvable, so we focus on the harder, invertible case to motivate the approach of *information geometric causal inference* (IGCI) (Janzing et al., 2012). We begin with a motivating example for IGCI and note that it forms the basis of our *strong asymmetry* framework. Later, we extend the strong asymmetry framework to a low-noise regime as well, i.e., when $Y = g(X) + \epsilon$ for "small" noise $\epsilon$.

### *3.1. Introduction*

For a pair of random variables $(X, Y)$, let $X \Rightarrow Y$ with $Y = g(X)$, where $g$ is a bijective function. Without loss of generality, let $g$ be monotonically increasing without loss of generality. Finally, let $X$ be distributed on $[0, 1]$. We will later relax this assumption to cases where $X$ does not have a compact support.

The key assumption of IGCI is that if $X \Rightarrow Y$, the distribution of $X$ and the function $g$ mapping $X$ to $Y$ are "independent" since they correspond to separate mechanisms of nature. Naturally, this postulate is not a rigorous one since there is no formal definition for independence of functions. One possible formalization may be through *algorithmic independence* (Lemeire and Dirkx, 2006; Janzing and Schölkopf, 2010) which invokes the notion of Kolmogorov complexity Kolmogorov (1963). However, since Kolmogorov complexity is not computable (Chaitin et al., 1995), a more practical notion of dependence is needed.

### *3.2. Correlation between slope and density*

In Section 2.3, we observed how, if the peaks of $f_X$ are uncorrelated with the slope of $g$, then the peaks of $f_Y$ are correlated with regions where $g^{-1}$ has high slope. We formalize this observation in the following lemma:

LEMMA 3.1. *Correlations of log-transformed slope and density: Let $X$ and $Y = g(X)$ be random variables defined on $[0, 1]$ where $g$ is a differentiable bijection with a differentiable inverse.*

*If the slope of $g$ and $f_X$ are approximately uncorrelated in the sense that*

$$\int \log\left(\nabla g(x)\right) f_X(x)dx \approx \int \log\left(\nabla g(x)\right) dx,$$

*then the slope of $g^{-1}$ and $f_Y$ are positively correlated, i.e.,*

$$\int \log\left(\nabla g^{-1}(y)\right) f_Y(y)dy \geq \int \log\left(\nabla g^{-1}(y)\right) dy,$$

*with equality if and only if $\nabla g$ is constant.*

Note that the word "uncorrelated" is justified here if we interpret $\nabla g$ and $f_X$ as random variables on the probability space $[0,1]$ with uniform measure:

$$\mathrm{Cov}\left(\log\left(\nabla g(X)\right), f_X(X)\right) = \int \log\left(\nabla g(x)\right) f_X(x)dx - \int \log\left(\nabla g(x)\right) dx \int f_X(x)dx,$$

where the last integral $\int f_X(x)dx = 1$. Hence, Lemma 3.1 asserts that $f_Y$ contains information about $g^{-1}$, and hence, about $f_{X|Y}$ whenever $X \Rightarrow Y$ such that $P_X$ does not contain information $g$, and hence, about $f_{Y|X}$.

### 3.3. Reformulation in terms of information geometry

Before further studying the "asymmetry" of information as described in the previous section, we embed our insights into an information geometric framework. We begin with the following lemma:

LEMMA 3.2. *Orthogonality in information space: Let $p, q$, and $r$ be three probability density functions such that*

$$\int p(x) \log \frac{q(x)}{r(x)} dx = \int q(x) \log \frac{q(x)}{r(x)} dx.$$

*Then we have*

$$KL\left(p \parallel q\right) + KL\left(q \parallel r\right) = KL\left(r \parallel s\right),$$

*where $KL\left(p \parallel q\right)$ denotes the Kullback-Leibler divergence between to probability densities and is given by $KL\left(p \parallel q\right) = \int p(x) \log\left(p(x)/q(x)\right) dx$.*

The three densities $p, q$, and $r$ are said to form a Pythagorean triplet of densities. This terminology is motivated by interpreting relative entropy as a squared distance and the triple thus satisfies the Pythagorean theorem. If the condition of Lemma 3.2 holds, intuitively we

understand that the vector connecting $p$ with $q$ is orthogonal to the one connecting $q$ with $r$ (Amari, 2001). It must be noted, however, that KL divergences are not symmetric. Hence, interchanging the densities or the direction of arrows connecting them will lead to a different scenario. Finally, we note that KL divergences are (i) non-negative, and (ii) invariant under bijective mappings, i.e., $KL(p_g \parallel q_g) = KL(p \parallel q)$, for any bijection $g$, where $p_g$ and $q_g$ are the images of $p$ and $q$ under $g$ respectively under bijective transformation $g$ of their corresponding random variables.

Returning to the orthogonality condition specified by Lemma (3.1) in an information geometric framework, we must first consider a pair of "reference densities". Let $u_X$ and $u_Y$ denote said "reference densities" of $X$ and $Y$, respectively. Note that the reference densities are distinct from the actual probability densities of $X$ and $Y$. In this example, we consider uniform reference distributions. Let $u_g(y)$ be the image of $u_X(x)$ under $y = g(x)$ and $u_{g^{-1}}(x)$ be the image of $u_Y(y)$ under $x = g^{-1}(y)$ given by $u_g(y) := \nabla g^{-1}(y)$ and $u_{g^{-1}}(x) := \nabla g(x)$ respectively. We reformulate Lemma (3.1) in terms of Lemma (3.2) and note that

$$\begin{aligned} KL\left(f_X \parallel u_X\right) + KL\left(u_X \parallel u_{g^{-1}}\right) &= KL\left(f_X \parallel u_{g^{-1}}\right), \\ KL\left(f_X \parallel u_X\right) + KL\left(u_g \parallel u_y\right) &= KL\left(f_y \parallel u_Y\right), \end{aligned} \tag{9}$$

where the second equality follows since KL divergences are preserved under bijective maps. We interpret the second equality of (9) by noting the distance of $f_Y$ from its reference distribution $u_Y$ is equal to the distance of $f_X$ from its reference distribution $u_X$ plus some non-negative term that measures irregularities (or slope) of $g$. *In other words, the function $g$ introduces new irregularities to the density of the effect $Y$ rather than smoothing the irregularities of the density of the cause $X$.*

### *3.4. Generalising IGCI*

In our discussion so far, the uniform distribution on [0, 1] plays a special role because it is the distribution with respect to which uncorrelatedness between $f_X$ and $\log(\nabla g)$ is defined. IGCI can be generalised further to cases where $X$ is no longer restricted to $[0, 1]$; in such cases we will need to specify other appropriate reference distributions. Let $X$ be any real-valued random variable with distribution $f_X$ and let $Y = g(X)$ where $g$ is a bijective function, assumed to be monotonically increasing without loss of generality. Let $u_X$ and $u_Y$ be reference densities (not necessarily uniform) of $X$ and $Y$ respectively. Let $u_g(y)$ be the image of $u_X(x)$ under $y = g(x)$

and $u_{g^{-1}}(x)$ be the image of $u_Y(y)$ under $x = g^{-1}(y)$ given by:

$$u_g(y) := u_X\left(g^{-1}(y)\right) \nabla g^{-1}(y),$$
$$u_{g^{-1}}(x) := u_Y\left(g(x)\right) \nabla g(x).$$

Analogous to Lemma (3.1), correlation between the slope of $g$ and $f_X$ will be reformulated as follows: if $X \Rightarrow Y$ via a deterministic bijective function $g$ such that $u_{g^{-1}}$ has a well-defined density, then we will assume:

$$\int \log\left(\frac{u_{g^{-1}}(x)}{u_X(x)}\right) f_X(x)dx \approx \int \log\left(\frac{u_{g^{-1}}(x)}{u_X(x)}\right) u_X(x)dx. \tag{10}$$

We interpret the above assumption as a statement of uncorrelatedness between the functions $\log\left(u_{g^{-1}}(x)/u_X(x)\right)$ and $f_X/u_X$ with respect to the measure given by the density of $u_X$. We posit that this is a reasonable assumption, since the former quantity is a property of the function $g$ and should thus be unrelated to the latter term involving marginal density $f_X$ (Janzing et al., 2012; Mooij et al., 2016). The special case described by Lemma 3.1 can be recovered by setting $u_X$ and $u_Y$ to $U(0,1)$. Given the assumption in (10), we are able to recover the information theoretic equalities given by (9), as well as the subsequent inequality following from (9),

$$KL(f_X \parallel u_X) \leq KL(f_Y \parallel u_Y). \tag{11}$$

Janzing et al. (2012) study this inequality or asymmetry by means of the following expression:

$$C_{X\Rightarrow Y} := KL(f_Y \parallel u_Y) - KL(f_X \parallel u_X). \tag{12}$$

Hence, IGCI infers $X \Rightarrow Y$ whenever $C_{X\Rightarrow Y} > 0$ is positive. In other words, the decision rule amounts to inferring that the density of the cause is closer to its reference density.

Choosing different reference densities might potentially bias our findings in one particular direction, so we recommend setting similar reference densities for $X$ and $Y$. In the case of compact support, we follow the suggestion of Daniusis et al. (2012) and choose uniform distributions which coincide with the respective ranges of $X$ and $Y$. Since we don't know the actual densities, this translates to a pre-processing step whereby we scale and shift $X$ and $Y$ such that they are individually mapped to $[0,1]$. In the case of distributions with non-compact support, Daniusis et al. (2012) recommend choosing Gaussian distributions with the same mean and variance as

$X$ and $Y$ respectively, implying a pre-processing step where both $X$ and $Y$ are scaled to have unit variance.

Given the pre-processed data, as described in Section 3.4, we follow the estimation procedure from $C_{X\Rightarrow Y}$ from empirical data outlined by Daniusis et al. (2012); Janzing et al. (2012); Mooij et al. (2016). The authors propose the following entropy-based estimator of $C_{X\Rightarrow Y}$:

$$\hat{C}_{X\Rightarrow Y} := \hat{H}(X) - \hat{H}(Y), \tag{13}$$

where $\hat{H}(X)$ and $\hat{H}(Y)$ are estimates of differential entropy of $X$ and $Y$ respectively, as defined earlier in Section 2.1. *Hence, the IGCI decision rule considers the variable with lower differential entropy as the effect.* For the rest of the paper, we will treat $\hat{C}_{X\Rightarrow Y}$ as the sample counterpart of the true coefficient $C_{X\Rightarrow Y} = H(X) - H(Y)$. Hence, if the assumptions of IGCI hold, the sign of $C_{X\Rightarrow Y}$ reveals a *strong asymmetry* between cause and effect.

### 3.5. Strong asymmetry in the low-noise regime

In Section 3.4 we noted that in the *strong asymmetry* framework, we compare the marginal entropies of $X$ and $Y$ and consider the variable with lower entropy as the effect. A surprising finding, which we describe here, is that we can extend this method also works in the low-noise regime. In brief, assuming that $X$ is the cause and that the additive noise model holds, we bound the variance of the noise so that $H(Y)$ remains lower than $H(X)$.

To set up the problem, consider a situation where we set the reference density of $X$ to be the normal distribution with the same mean $\mu_X$ and variance $\sigma_X^2$ as $X$. Then, the deviation of the true density $f_X$ from said reference density is given by $D(p_X) := KL\left(f_X \mid N\left(\mu_x, \sigma_X^2\right)\right)$. The measure $D\left(p_X\right)$ is a measure of *non-Gaussianity*, or departure of $X$ from Gaussian behaviour. Note that the measure may be re-written as $D(p_X) = \log\left(\sqrt{2\pi e\sigma_X^2}\right) - H(X)$.

We further note that the entropy of $X$ is modified by adding white noise $\epsilon \sim N(0,1)$ to it. The modified entropy is related to the underlying Fisher information $I$ by de Bruijn's identity (Cover and Thomas, 2005) as follows:

$$\frac{\delta}{\delta t} H(X + \sqrt{t}\epsilon) = \frac{1}{2} I(X + \sqrt{t}\epsilon),$$

where $X \perp \epsilon$, and $I(X)$ is the Fisher information associated with $X$, defined by

$$I(X) := \mathbb{E}\left(\frac{\delta}{\delta x}\log\left(f_X(X)\right)\right)^2.$$

The following theorem provides an upper-bound on the entropy of $X$ when noise is added to it.

THEOREM 3.3. *Let $X$, having density $f_X$, be contaminated with noise $Z$ such that $X \perp Z$ with $\mathbb{E}(Z) = 0$ and $\mathbb{V}(Z) = 1$. Then, assuming the deviation of $X + \sqrt{\sigma}Z$ from Gaussianity is more than the deviation of $X + \sqrt{\sigma}\epsilon$ from Gaussianity, where $\epsilon \sim N(0,1)$, i.e.,*

$$D\left(f_{X+\sqrt{\sigma}Z}\right) \geq D\left(f_{X+\sqrt{\sigma}\epsilon}\right), \tag{14}$$

*then the following upper-bound on the entropy of $X + \sqrt{\sigma}Z$ holds:*

$$H\left(X + \sqrt{\sigma}Z\right) \leq H(X) + \frac{1}{2}\log\left(\sigma I(X) + 1\right).$$

PROOF. From de Bruijn's identity (Cover and Thomas, 2005), we get

$$H\left(X + \sqrt{\sigma}\epsilon\right) - H(X) = \frac{1}{2}\int_0^\sigma I\left(X + \sqrt{t}\epsilon\right) dt,$$

The integrand on the right may be bounded above by using the Fisher information inequality. For any two random variables $X$ and $Y$ with well-defined Fisher information, the information inequality (Dembo et al., 1991) states $1/I(X+Y) \geq 1/I(X) + 1/I(Y)$. Consequently, we get

$$H\left(X + \sqrt{\sigma}\epsilon\right) - H(X) \leq \frac{1}{2}\int_0^\sigma \frac{I(X)I(\sqrt{t}\epsilon)}{I(X) + I(\sqrt{t}\epsilon)} dt.$$

Note that since $\epsilon \sim N(0,1)$, we have $I(\sqrt{t}\epsilon) = 1/t$. A little algebra reveals the inequality:

$$H\left(X + \sqrt{\sigma}\epsilon\right) - H(X) \leq \frac{1}{2}\log\left(\sigma I(X) + 1\right),$$

Further, assuming the inequality given by (14) holds, we can rewrite the inequality in terms of entropy: $H(X + \sqrt{\sigma}Z) \leq H(X + \sqrt{\sigma}\epsilon)$. This gives rise to the following inequality chain:

$$H\left(X + \sqrt{\sigma}Z\right) \leq H\left(X + \sqrt{\sigma}\epsilon\right) \leq H(X) + \frac{1}{2}\log\left(\sigma I(X) + 1\right),$$

which completes the proof.

Theorem 3.3 presents the basis for showing the robustness of our approach in low-noise regimes. The following lemma (presented without proof) describes this robustness in greater detail.

COROLLARY 3.1. *Let $Y = f(X)$ such that $H(Y) < H(X)$. For any noise $E$ with mean zero and variance $\mathbb{V}(E) = \sigma$ such that*

$$\sigma < \frac{\exp(2H(X) - 2H(Y)) - 1}{I(Y)},$$

*we have $H(Y + \sqrt{\sigma}E) < H(X)$.*

PROOF. Rewriting Theorem 3.3 in terms of $Y$ instead of $X$, we have $H\left(Y + \sqrt{\sigma}Z\right) \leq H(Y) + \frac{1}{2}\log\left(\sigma I(Y) + 1\right)$. Next, utilising the upper-bound on $\sigma$ assumed above, we note $\frac{1}{2}\log\left(\sigma I(Y) + 1\right) < H(X) - H(Y)$, which implies $H\left(Y + \sqrt{\sigma}Z\right) \leq H(X)$. This concludes the proof.

In other words, the *strong asymmetry* framework works in low-noise regimes of $Y = f(X) + E$ and Corollary 3.1 provides a bound on the strength of noise that the method can tolerate while providing a correct decision.

## 4. Adjusting for confounding factors

In Sections 2 and 3 we describe a framework to examine weak and strong asymmetry in bivariate $(X, Y)$ using marginal entropies. In this section, we propose the inclusion of confounding factor(s) $\mathbf{Z}$ within our framework. We will describe a framework to infer asymmetry in $(X, Y)$ from observations $\{(x_i, y_i)\}_{i=1}^n$ given observations $\{\mathbf{z}_i\}_{i=1}^n$ on confounder $\mathbf{Z}$. We will assume that $\mathbf{Z}$ has a density function $f_{\mathbf{Z}}$ with support denoted by $\mathcal{S}_{\mathbf{Z}}$. For now, we will restrict ourselves to adjusting low-dimensional confounder effects.

We explore two related questions: first, we propose a framework for examining asymmetry between $(X, Y)$ given a specific value of observed confounder $\mathbf{Z} = \mathbf{z}$; next, we propose an extension that examines asymmetry between $(X, Y)$ conditioned across all values in the support $\mathcal{S}_{\mathbf{Z}}$. While the former approach will allow for strata-specific comparisons of asymmetry between $(X, Y \mid \mathbf{Z} = \mathbf{z})$ given a fixed value $\mathbf{Z} = \mathbf{z}$, the latter will facilitate population-level comparisons of asymmetry between $(X, Y \mid \mathbf{Z})$ for confounder $\mathbf{Z}$.

### 4.1. Confounding in weak asymmetry framework

Extending the framework of weak asymmetry in Section 2.2, we note that

$$C_{X \Rightarrow Y|\mathbf{Z}=\mathbf{z}} = H(X \mid \mathbf{Z} = \mathbf{z}) - H(Y \mid \mathbf{Z} = \mathbf{z}), \tag{15}$$

serves as a measure to quantify weak asymmetry in $(X, Y)$ for a specific value of observed confounder $\mathbf{Z} = \mathbf{z}$. If weak asymmetry in a specific direction holds for all values $\mathbf{Z} \in \mathcal{S}_Z$, i.e., say, if $C_{X \Rightarrow Y|\mathbf{Z}=\mathbf{z}} > 0$ all $\mathbf{Z} \in \mathcal{S}_Z$, we can further claim that the population-level measure of weak asymmetry in $(X, Y \mid \mathbf{Z})$ is given by

$$\begin{aligned} C_{X \Rightarrow Y|\mathbf{Z}} &= \int f_{\mathbf{Z}}(\mathbf{z}) \left\{ C_{X \Rightarrow Y|\mathbf{Z}=\mathbf{z}} \right\} d\mathbf{z} \\ &= \int f_{\mathbf{Z}}(\mathbf{z}) \left\{ H(X \mid \mathbf{Z} = \mathbf{z}) \right\} d\mathbf{z} - \int f_{\mathbf{Z}}(\mathbf{z}) \left\{ H(Y \mid \mathbf{Z} = \mathbf{z}) \right\} d\mathbf{z} \\ &= H(X \mid \mathbf{Z}) - H(Y \mid \mathbf{Z}). \end{aligned} \tag{16}$$

### 4.2. Confounding in strong asymmetry framework

Extending the framework of strong asymmetry in Section 3.4, let us consider the scenario, for a given value of $\mathbf{z} \in \mathcal{S}_{\mathbf{Z}}$, we have $Y = g(X, \mathbf{Z} = \mathbf{z}) = g_{\mathbf{z}}(X)$, where $g_{\mathbf{z}}$ is a bijective function, assumed to be monotonically increasing without loss of generality. Assuming the conditional density $f_{X|\mathbf{z}}$ and the generative function $g_{\mathbf{z}}$ satisfy the assumption in (10), the measure given by (15) serves as a measure of strong asymmetry in $(X, Y)$ for a specific value of observed confounder $\mathbf{Z} = \mathbf{z}$. Further, if strong asymmetry in a specific direction holds for all values $\mathbf{Z} \in \mathcal{S}_Z$, we can further claim that the population-level measure of strong asymmetry in $(X, Y \mid \mathbf{Z})$ is given by (16).

## 5. Estimation and inference

In either framework of *weak* or *strong asymmetry*, we must estimate and perform inference using $\hat{C}_{X \Rightarrow Y}$. To do so, we must first estimate the underlying marginal densities $f_X$ and $f_Y$, which may be thought of as infinite-dimensional nuisance parameters. In Section 5.1, we describe a density estimation technique that estimates density functions in an accurate and fast manner without incurring the need for tuning parameters. Using the estimated densities, we obtain consistent estimates of $\hat{C}_{X \Rightarrow Y}$. Next, in Section 5.2, we introduce a cross-fitting technique that allows us to provide an inference rule for testing if $X \Rightarrow Y$.

### *5.1. Self-consistent density estimation*

The self-consistent estimator (SCE) was proposed by Bernacchia and Pigolotti (2011); O'Brien et al. (2016) to minimize the mean integrated squared error (MISE) between the estimated density and the true density without incurring any manual parameter tuning. The estimation process relies on fast Fourier transforms (FFT). Utilizing this 'optimal' density estimator in Purkayastha and Song (2023), we propose a plug-in estimator of $MI$, termed as the `fastMI`, which is shown to be consistent and manyfold faster than the original. Extending the usage of the self-consistent density estimator here, we wish to estimate the marginal entropies $\hat{H}(X)$ and $\hat{H}(Y)$, thereby obtaining an estimate of $\hat{C}_{X\Rightarrow Y}$.

Let us consider a random sample denoted by $\mathcal{S} = \{X_1, X_2, \ldots, X_n\}$ from an unknown density $f$ with support $\mathcal{X}$ (without loss of generality, $\mathcal{X} = \mathbb{R}$). We assume $f$ belongs to the Hilbert space of square-integrable functions, given by $\mathcal{L}^2 = \left\{f : \int f^2(x)dx < \infty\right\}$. The SCE is given by $\hat{f} \in \mathcal{L}^2$. First, in order to define $\hat{f}$, we require a kernel function $K$, which belongs to the class of functions given by

$$\mathcal{K} := \left\{K : K(x) \geq 0, K(x) = K(-x)\ \forall\ x; \int K(t)dt = 1\right\}.$$

Specifically, $\hat{f}$ is given by the convolution of a kernel $K$ and the set of delta functions centered on the dataset as follows:

$$\hat{f}(z) \equiv n^{-1}\sum_{j=1}^{n} K(x - X_j) = n^{-1}\sum_{j=1}^{n}\int_{\mathbb{R}} K(s)\delta(x - X_j - s)ds, \quad x \in \mathbb{R} \tag{17}$$

where $\delta(x)$ is the Dirac delta function (Kreyszig, 2020). The optimal $\hat{f}$ is identified by the optimal kernel $\hat{K}$, where "optimality" is intended as minimising the mean integrated square error ($MISE$) between the true density $f$ and the estimator $\hat{f}$:

$$\hat{K} = \underset{K\in\mathcal{K}}{\operatorname{argmin}}\ MISE(\hat{f}, f) = \underset{K\in\mathcal{K}}{\operatorname{argmin}}\ \mathbb{E}\left[\int_{\mathbb{R}}\{\hat{f}(x) - f(x)\}^2 dx\right], \tag{18}$$

where the $\mathbb{E}$ operator denotes taking expectation over the entire support of $f$. The SCE in (17) may be represented equivalently by its inverse Fourier transform pair, $\hat{\phi} \in \mathcal{L}^2$:

$$\hat{\phi}(t) = \mathcal{F}^{-1}\big(\hat{f}(z)\big) = \kappa(t)\mathcal{C}(t), \quad t \in \mathbb{R} \tag{19}$$

where $\mathcal{F}^{-1}$ represents the multidimensional inverse Fourier transformation from space of data

$z \in \mathbb{R}$ to frequency space coordinates $t \in \mathbb{R}$. $\kappa = \mathcal{F}^{-1}(K)$ is the inverse Fourier transform of the kernel $K$ and $\mathcal{C}$ is the empirical characteristic function (ECF) of the data, defined as $\mathcal{C}(t) = n^{-1} \sum_{j=1}^{n} \exp(itZ_j)$. Bernacchia and Pigolotti (2011) derive the optimal transform kernel $\hat{\kappa}$ that minimizes the $MISE$ given by (18), given as follows:

$$\hat{\kappa}(t) = \frac{n}{2(n-1)} \left[ 1 + \sqrt{1 - \frac{4(n-1)}{|n\mathcal{C}(t)|^2}} \right] I_{A_n}(t), \tag{20}$$

where $A_n$ serves as a low-pass filter that yields a stable estimator (see Purkayastha and Song (2023) for a detailed discussion on the filter). We follow the nomenclature of Bernacchia and Pigolotti (2011) and denote $A_n$ as the set of "acceptable frequencies". The optimal transform kernel $\hat{\kappa}$ in (20) may be antitransformed back to the real space to obtain the optimal kernel $\hat{K} \in \mathcal{K}$, which yields the optimal density estimator $\hat{f}$ according to Equation 17. Theorem 5.1 presents the sufficient conditions for the estimate $\hat{f}$ to converge to the true density $f$ for $n \to \infty$.

THEOREM 5.1. *Let the true density $f$ be square integrable and its corresponding Fourier transform $\phi$ be integrable, then the self consistent estimator $\hat{f}$, which is defined by (17) - (20) converges almost surely to the true density as $n \to \infty$, under the additional assumptions $\mathcal{V}(A_n) \to \infty, \mathcal{V}(A_n)/\sqrt{n} \to 0$ and $\mathcal{V}(\bar{A}_n) \to 0$ as $n \to \infty$. Further, assuming $f$ to be continuous on dense support $\mathcal{X}$, we have uniform almost sure convergence of $\hat{f}$ to $f$ as $n \to \infty$.*

Here $\bar{A}_n$ is the complement of $A_n$ and the volume of $A_n$ is given by $\mathcal{V}(A_n)$. The detailed proof of Theorem 5.1 is omitted here. Please see Purkayastha and Song (2023) for the complete proof.

### 5.2. Data splitting and cross-fitting inference

Let $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})\}$ be a random sample drawn from a bivariate distribution $f_{XY}$ with marginal $f_X$ for $X$ and $f_Y$ for $Y$. Since we do not have knowledge of $f_X$ or $f_Y$, we invoke a data-splitting and cross-fitting technique to estimate the underlying density functions as well as the relevant entropy terms: we first split the available data $\mathcal{D}$ into two equal-sized but disjoint sets

$$\mathcal{D}_1 := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}, \quad \mathcal{D}_2 := \{(X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n})\}.$$

Using one data split $\mathcal{D}_1$, we obtain estimates of the marginal density functions $\hat{f}_{X;1}$ and $\hat{f}_{Y;1}$. We provide details on the density estimation method in Section 5.1. The estimated density functions are evaluated for data belonging to the second data split $\mathcal{D}_2$ to obtain the following

estimates:

$$\widehat{H_2}(X) = -\frac{1}{n}\sum_{j=1}^{n} \ln\left(\hat{f}_{X;1}\left(X_{n+j}\right)\right), \quad \widehat{H_2}(Y) = -\frac{1}{n}\sum_{j=1}^{n} \ln\left(\hat{f}_{Y;1}\left(Y_{n+j}\right)\right). \tag{21}$$

Interchanging the roles of data splits $\mathcal{D}_1$ and $\mathcal{D}_2$, we obtain the estimated densities $\hat{f}_{X;2}$ and $\hat{f}_{Y;2}$. The estimated density functions are evaluated for data belonging to data split $\mathcal{D}_1$:

$$\widehat{H_1}(X) = -\frac{1}{n}\sum_{j=1}^{n} \ln\left(\hat{f}_{X;2}\left(X_{j}\right)\right), \quad \widehat{H_1}(Y) = -\frac{1}{n}\sum_{j=1}^{n} \ln\left(\hat{f}_{Y;2}\left(Y_{j}\right)\right). \tag{22}$$

We take averages of the two sets of estimates obtained above to obtain the "cross-fitted" estimates:

$$\hat{H}(X) = \frac{\widehat{H_1}(X) + \widehat{H_2}(X)}{2}, \quad \hat{H}(Y) = \frac{\widehat{H_1}(Y) + \widehat{H_2}(Y)}{2}. \tag{23}$$

We define oracle estimators of $H_X$, and $H_Y$ for each of the data splits and the combined sample:

$$\begin{aligned} H_{0;1}(X) &= -\tfrac{1}{n}\textstyle\sum_{j=1}^{n} \ln\left(f_X\left(X_j\right)\right) \quad H_{0;2}(X) = -\tfrac{1}{n}\textstyle\sum_{j=1}^{n} \ln\left(f_X\left(X_{n+j}\right)\right) \\ H_{0;1}(Y) &= -\tfrac{1}{n}\textstyle\sum_{j=1}^{n} \ln\left(f_Y\left(Y_j\right)\right) \quad H_{0;2}(Y) = -\tfrac{1}{n}\textstyle\sum_{j=1}^{n} \ln\left(f_Y\left(Y_{n+j}\right)\right) \end{aligned} \tag{24}$$

which may be averaged to obtain the "cross-fitted oracle estimates":

$$\begin{aligned} H_0(X) &= \frac{H_{0;1}(X) + H_{0;1}(X)}{2}, \\ H_0(Y) &= \frac{H_{0;1}(Y) + H_{0;1}(Y)}{2}. \end{aligned} \tag{25}$$

The following two theorems establish (1) consistency and (2) asymptotic normality of the cross-fitted estimates.

THEOREM 5.2. *Let the conditions presented in Theorem 5.1 hold when estimating $f_X$ and $f_Y$ using data splits $\mathcal{D}_1$ and $\mathcal{D}_2$. We further assume that the densities are bounded away from zero and infinity on their respective supports. Under these assumptions, we have consistency of the cross-fitted estimates: $\hat{C}_{X\Rightarrow Y} \overset{a.s.}{\to} C_{X\Rightarrow Y}$ as $n \to \infty$.*

PROOF. From Theorem 5.1, for any small $\epsilon$, there exists sufficiently large $n$ such that

$$\begin{aligned} \left|\hat{f}_{X;1}(x) - f_X(x)\right| &< \epsilon, \quad x \in \mathbb{R} \\ \left|\hat{f}_{X;2}(x) - f_X(x)\right| &< \epsilon, \quad x \in \mathbb{R}. \end{aligned}$$

Consider the following Taylor expansions for $j \in 1, \ldots, n$:

$$\ln\left(\hat{f}_{X;1}(X_{n+j})\right) = \ln\left(f_{X;1}(X_{n+j})\right) + \frac{\hat{f}_{X;1}(X_{n+j}) - f_{X;1}(X_{n+j})}{f_{X;1}(X_{n+j})} + o(\epsilon)$$
$$\ln\left(\hat{f}_{X;2}(X_j)\right) = \ln\left(f_{X;2}(X_j)\right) + \frac{\hat{f}_{X;2}(X_j) - f_{X;2}(X_j)}{f_{X;2}(X_j)} + o(\epsilon).$$

Since $f_X$ is bounded below by a constant $B^{-1}$ on its support, we have

$$\left|\hat{H}_1(X) - H_{0;1}(X)\right| \leq \frac{1}{n}\sum_{j=1}^{n}\left|\frac{\hat{f}_{X;2}\left(X_j\right) - f_X\left(X_j\right)}{f_X\left(X_j\right)}\right| \leq \frac{\epsilon}{B},$$
$$\left|\hat{H}_2(X) - H_{0;2}(X)\right| \leq \frac{1}{n}\sum_{j=1}^{n}\left|\frac{\hat{f}_{X;1}\left(X_{n+j}\right) - f_X\left(X_{n+j}\right)}{f_X\left(X_{n+j}\right)}\right| \leq \frac{\epsilon}{B},$$

which together imply $\left|\hat{H}(X) - H_0(X)\right| \overset{a.s.}{\to} 0$, as $n \to \infty$. Further, the strong law of large numbers implies as $n \to \infty$, we have $H_0(X) - H(X) \overset{a.s.}{\to} 0$. Hence, using the inequality obtained above in conjunction with the strong law of large numbers yields $\hat{H}(X) - H(X) \overset{a.s.}{\to} 0$ as $n \to \infty$. Similar results hold for $\hat{H}_Y$ - thus, invoking continuous mapping theorem (Billingsley, 1995), we are able to show $\hat{C}_{X\Rightarrow Y} \overset{a.s.}{\to} C_{X\Rightarrow Y}$ as $n \to \infty$. This concludes the proof.

THEOREM 5.3. *Assuming the conditions presented in Theorem 5.2 hold, we have*

$$\sqrt{n}\begin{pmatrix}\hat{H}(X) - H_0(X)\\ \hat{H}(Y) - H_0(Y)\end{pmatrix} \overset{\mathcal{P}}{\to} 0, \ \textit{as } n \to \infty. \tag{26}$$

PROOF. We consider the following Taylor series expansion:

$$\sqrt{n}\left\{\hat{H}_2(X) - H_{0;2}(X)\right\} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\frac{\hat{f}_{X;1}(X_{n+j}) - f_{X;1}(X_{n+j})}{f_{X;1}(X_{n+j})}\right\} + o_{\mathcal{P}}(n^{-1/2})$$
$$= \frac{S_n(2)}{\sqrt{n}} + o_{\mathcal{P}}(n^{-1/2}).$$

We now show that the leading term $S_n(2)/\sqrt{n} \overset{\mathcal{P}}{\to} 0$ as $n \to 0$, which will establish

$$\sqrt{n}\left\{\hat{H}_2(X) - H_{0;2}(X)\right\} \overset{\mathcal{P}}{\to} 0, \quad \text{as } n \to 0.$$

Using identical arguments, we establish similar results for $\hat{H}_2(Y)$ as well. It is sufficient to show $\mathbb{E}\left[\{S_n(2)/\sqrt{n}\}^2\right] \to 0$ as $n \to \infty$. Note that $\mathbb{E}\left[\{S_n(2)/\sqrt{n}\}^2\right] = \mathbb{E}^2\left[\{S_n(2)/\sqrt{n}\}\right] +$

$\mathbb{V}\left[\{S_n(2)/\sqrt{n}\}\right]$. First, we prove $\mathbb{E}\left[\{S_n(2)/\sqrt{n}\}\right]$:

$$\begin{aligned}\mathbb{E}\left\{\frac{S_n(2)}{\sqrt{n}}\right\} &= \mathbb{E}_{\mathcal{D}_1}\left[\mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\frac{S_n(2)}{\sqrt{n}}\middle|\mathcal{D}_1\right\}\right] \\ &= \mathbb{E}_{\mathcal{D}_1}\left[\mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\frac{\hat{f}_{X;1}(X_{n+j})-f_X(X_{n+j})}{f_X(X_{n+j})}\right)\middle|\mathcal{D}_1\right\}\right] \\ &= \sqrt{n}\,\mathbb{E}_{\mathcal{D}_1}\left[\mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{\hat{f}_{X;1}(X)-f_X(X)}{f_X(X)}\right)\middle|\mathcal{D}_1\right\}\right],\end{aligned}$$

where the inner expectation term is evaluated as follows:

$$\mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{\hat{f}_{X;1}(X)-f_X(X)}{f_X(X)}\right)\middle|\mathcal{D}_1\right\} = \int_{\mathbb{R}}\left(\frac{\hat{f}_{X;1}(x)-f_X(x)}{f_X(x)}\right)f_X(x)dx = 0.$$

The last equality holds since $\hat{\phi}_1$ is the Fourier transform associated with the optimal density function estimator $\hat{f}_{X;1}$, and we know $\hat{\phi}_1(0) = 1$ from Theorem 5.2 and consequently, $\mathbb{E}\left[\{S_n(2)/\sqrt{n}\}\right] = 0$. Next, we consider the term $\mathbb{V}\left[\{S_n(2)/\sqrt{n}\}\right]$:

$$\mathbb{V}\left[\left\{\frac{S_n(2)}{\sqrt{n}}\right\}\right] = \mathbb{E}_{\mathcal{D}_1}\left[\mathbb{V}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{S_n(2)}{\sqrt{n}}\right)\middle|\mathcal{D}_1\right\}\right] + \mathbb{V}_{\mathcal{D}_1}\left[\mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{S_n(2)}{\sqrt{n}}\right)\middle|\mathcal{D}_1\right\}\right],$$

where the second term is zero, as per our calculations above. Conditional on $\mathcal{D}_1$, the terms $\hat{f}_{X;1}(X_{n+j})$ are independent and identically distributed for $1 \leq j \leq n$. We have:

$$\begin{aligned}\mathbb{V}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{S_n(2)}{\sqrt{n}}\right)\middle|\mathcal{D}_1\right\} &= \mathbb{V}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\frac{\hat{f}_{X;1}(X_{n+j})-f_X(X_{n+j})}{f_X(X_{n+j})}\right)\middle|\mathcal{D}_1\right\} \\ &= \mathbb{V}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{\hat{f}_{X;1}(X)-f_X(X)}{f_X(X)}\right)\middle|\mathcal{D}_1\right\} \\ &= \mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{\hat{f}_{X;1}(X)-f_X(X)}{f_X(X)}\right)^2\middle|\mathcal{D}_1\right\},\end{aligned}$$

The last equality follows from $\mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\hat{f}_{X;1}(X)-f_X(X)\right)/f_X(X)\middle|\mathcal{D}_1\right\} = 0$. Moreover,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_2|\mathcal{D}_1}\left\{\left(\frac{\hat{f}_{X;1}(X)-f_X(X)}{f_X(X)}\right)^2\middle|\mathcal{D}_1\right\} &= \int_{\mathbb{R}}\left(\frac{\hat{f}_{X;1}(x)-f_X(x)}{f_X(x)}\right)^2 f_X(x)dx \\ &\leq B\int_{\mathbb{R}}\left(\hat{f}_{X;1}(x)-f_X(x)\right)^2 dx.\end{aligned}$$

where $B$ is a (positive) lower bound for the density $f_X$ over its support. Consequently, we get

$$\mathbb{V}\left[\{S_n(2)/\sqrt{n}\}\right] \leq B \times \mathbb{E}_{\mathcal{D}_1}\left\{\int_{\mathbb{R}}\left(\hat{f}_{X;1}(x) - f_X(x)\right)^2 dx\right\} = B \times MISE(\hat{f}_{X;1}, f_X).$$

Bernacchia and Pigolotti (2011) present an expression of $MISE$ in terms of the optimal kernel and prove that the last expression goes to zero as sample size increases, i.e., $MISE(\hat{f}_{X;1}, f_X) \to 0$ as $n \to \infty$. This allows us to claim $\sqrt{n}\left(\hat{H}_2(X) - H_{0;2}(X)\right) \xrightarrow{\mathcal{P}} 0$ as $n \to \infty$. Note that the arguments presented above are generally valid for any true density function that is bounded away from zero and infinity on its support. Hence, they can also be used to establish similar results involving $\sqrt{n}\left(\hat{H}_2(Y) - H_{0;2}(Y)\right) \xrightarrow{\mathcal{P}} 0$ as $n \to \infty$.

Note that we can interchange the roles of $\mathcal{D}_1$ and $\mathcal{D}_2$ in the proof above to arrive at

$$\sqrt{n}\begin{pmatrix}\hat{H}_1(X) - H_{0;1}(X)\\ \hat{H}_1(Y) - H_{0;1}(Y)\end{pmatrix} \xrightarrow{\mathcal{P}} 0, \text{ as } n \to \infty. \tag{27}$$

Together (26) and (27) can be combined by invoking the continuous mapping theorem to yield:

$$\sqrt{n}\begin{pmatrix}\hat{H}(X) - H_0(X)\\ \hat{H}(Y) - H_0(Y)\end{pmatrix} \xrightarrow{\mathcal{P}} 0, \text{ as } n \to \infty. \tag{28}$$

This concludes the proof.

LEMMA 5.4. *By the multivariate central limit theorem we have* $(H_0(X), H_0(Y))^\intercal$ *jointly converge in distribution to a two-dimensional normal distribution, namely*

$$\sqrt{n}\begin{pmatrix}H_0(X) - H(X)\\ H_0(X) - H(Y)\end{pmatrix} \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \Sigma\right), \tag{29}$$

*as* $n \to \infty$. $\Sigma$ *is a* $2\times 2$ *matrix that describes the variance-covariance terms of* $(H_0(X), H_0(Y))^\intercal$.

Using Theorem 5.3, Lemma 5.4 and the delta method (Casella and Berger, 2021), we get the following result about the asymptotic behaviour of $\sqrt{n}\hat{C}_{X\Rightarrow Y}$.

COROLLARY 5.1. *Assuming the conditions presented in Theorem 5.3 hold, we make use of Lemma 5.4 to note that as* $n \to \infty$, *we have*

$$\sqrt{n}\left(\hat{C}_{X\Rightarrow Y} - C_{X\Rightarrow Y}\right) \xrightarrow{\mathcal{D}} N(0, \sigma_C^2), \tag{30}$$

*where* $\sigma_C^2 := \mathbb{V}\left[\log\left(f_X(X)\right) + \log\left(f_Y(Y)\right)\right]$*, and can be estimated by Monte Carlo methods upon obtaining density estimates* $\hat{f}_X$ *and* $\hat{f}_Y$ *from the data* $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{2n}, Y_{2n})$.

PROOF. Using Theorem 5.3 and Lemma 5.1 in conjunction with Slutsky's theorem (Casella and Berger, 2021), we get

$$\sqrt{n}\begin{pmatrix} \hat{H}(X) - H(X) \\ \hat{H}(Y) - H(Y) \end{pmatrix} \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \Sigma\right).$$

Hence, we note that $\sqrt{n}\left(\hat{C}_{X\Rightarrow Y} - C_{X\Rightarrow Y}\right) \xrightarrow{\mathcal{D}} N\left(0, \sigma_C^2\right)$. This concludes the proof.

### 5.3. *Testing for asymmetry using* $C_{X\Rightarrow Y}$

The sign of $C_{X\Rightarrow Y}$ determines a direction of dependence between $X$ and $Y$. The null hypothesis $H_0 : C_{X\Rightarrow Y} = 0$ signifies an indeterminate case where we are unable to infer a direction of association between $X$ and $Y$. If $C_{X\Rightarrow Y}$ is significantly larger (or smaller) than zero, then we assign $X \Rightarrow Y$ (or $Y \Rightarrow X$). We propose an asymptotic test based on the large sample behaviour of $\hat{C}_{X\Rightarrow Y} = \hat{H}(X) - \hat{H}(Y)$ to test for the null hypothesis $H_0$ described above. Using the result presented in Theorem 5.1, we obtain a 95% asymptotic confidence interval (CI) of $\hat{C}_{X\Rightarrow Y}$ and draw inference on whether $X \Rightarrow Y$.

## 6. Simulation studies

In this section, we report the findings of extensive simulation studies that provide empirical evidence supporting the validity of $\hat{C}_{X\Rightarrow Y}$ to infer weak or strong asymmetry in bivariate $(X, Y)$.

### 6.1. *Coverage probability, bias and standard error of* $\hat{C}_{X\Rightarrow Y}$

We simulate data from $X \sim f_X$ with entropy $H(X)$. Next, we generate $Y = g(X)$ via a bijective function $g$. We consider two different cases: (i) $X \sim$ Lognormal$(5, 1)$ and $g(t) := \log(t)$, which implies $Y \sim$ N$(5, 1)$; and (ii) $X \sim$ Exp(mean $= 1$) and $g(t) := t^{2/3}$, which implies $Y \sim$ Weibull(scale $= 1$, shape $= 3/2$). Note that $f_X$ and $g$ above are chosen in a manner to ensure that the true value of $C_{X\Rightarrow Y}$ has a well-defined and closed-form expression that can be computed easily. This is critical when we evaluate the coverage probability of the $100(1-\alpha)\%$ CI of $\hat{C}_{X\Rightarrow Y}$. Given data of size $n$ from cases (i) or (ii) above, we obtain estimates of $\hat{C}_{X\Rightarrow Y}$. Repeating this process $R = 200$ times, we compute a set of estimates $\{\hat{C}^r_{X\Rightarrow Y}\}_{r=1}^R$ and their associated 95% CIs, as described by Section 5.3. Using this set of estimates, in Table 6.1,

**Table 1.** Examining $\hat{C}_{X\Rightarrow Y}$: coverage probability (bias) [standard error] for different sample sizes $(n)$. We consider two bivariate cases: (i) $X \sim$ Lognormal$(5,1)$ and $Y \sim$ N$(5,1)$ and (ii) $X \sim$ Exp(mean $= 1$) and $Y \sim$ Weibull(scale = $1$, shape = $3/2$).

| Simulation case | $n = 250$ | $n = 500$ | $n = 750$ |
|---|---|---|---|
| Case (i) | 0.91 (0.05) [0.08] | 0.93 (0.03) [0.05] | 0.96 (0.02) [0.04] |
| Case (ii) | 0.94 (0.05) [0.09] | 0.94 (0.04) [0.06] | 0.96 (0.03) [0.05] |

we report the coverage probability, bias, and standard error of $\hat{C}_{X\Rightarrow Y}$ for the simulation cases described above for different sample sizes. We note that the statistic reliably provides an approximate coverage of 95% for sample sizes $n = 500$ or more for both scenarios considered.

### *6.2. Behaviour of $\hat{C}_{X\Rightarrow Y}$ under weak asymmetry*

Using the statistic $\hat{C}_{X\Rightarrow Y}$, we want to investigate the departure from symmetric associations in synthetic bivariate data within our framework of weak asymmetry. We generate a sample of $n = 500$ observations drawn from a bivariate PDF $f_{XY}$ on $\mathbb{R}^2$, through representation involving Sklar's theorem (Sklar, 1959). We specify the underlying copula density function and the two associated marginal densities, given as follows:

(a) **Choice of copula density:** we choose the bivariate Gaussian copula with $\rho = 0.25$. For more details on various copula classes, please see Czado (2019); Joe (2014); Nelsen (2006).

(b) **Choice of marginals:** choice of the marginal densities influences predictive asymmetry in the proposed information-theoretic framework. We consider the Gaussian $N(0, \sigma^2)$, Exp(rate $= \lambda$), and Lognormal(scale $= \gamma$) density functions. We vary the parameters $\sigma$, $\lambda$ and $\gamma$ over a range of values.

Note that increasing the scale parameter $\gamma$ for the lognormal distribution and the dispersion parameter $\sigma$ for the Gaussian distribution yields increased entropy values. In contrast, increasing the rate parameter $\lambda$ for the exponential distribution yields decreased entropy values. For a given association pattern we generate $R = 200$ bivariate $i.i.d.$ samples on $(X, Y)$, each of size $n = 500$, in order to draw summary statistics. For each simulated sample, we study the estimate $\{\hat{C}^r_{X\Rightarrow Y}\}_{r=1}^R$ and compute the mean and quantile-based 95% confidence interval of $\hat{C}_{X\Rightarrow Y}$.

From Figure 2, we note that $\hat{C}_{X\Rightarrow Y}$ captures departure from balanced marginal entropy values in our weak asymmetry framework in bivariate data: our simulation results vary based on how we specify the marginal density functions, which in turn influences the underlying marginal entropy. Subplots (I), (II), (III), and (IV) examine behaviour of $\hat{C}_{X\Rightarrow Y}$ upon changing

the marginal parameters. In (I) and (III), note that increasing the $X$-marginal parameter while keeping the $Y$-marginal fixed causes the entropy of $X$ to rise, relative to $Y$, thereby causing an increase in the value of $\hat{C}_{X\Rightarrow Y}$. In (II) and (IV), note that increasing the $X$-marginal parameter while keeping the $Y$-marginal fixed causes the entropy of $X$ to fall, relative to $Y$, thereby causing a decrease in the value of $\hat{C}_{X\Rightarrow Y}$.

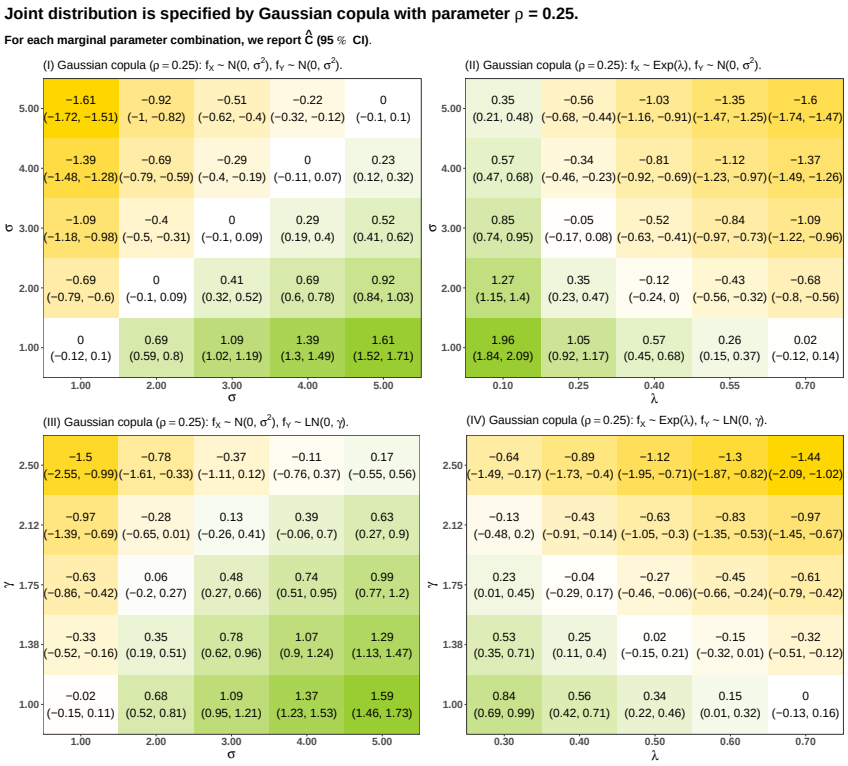


**Fig. 2.** Examining the behaviour of $\hat{C}_{X\Rightarrow Y}$ under weak asymmetry framework.

### *6.3. Behaviour of $\hat{C}_{X\Rightarrow Y}$ under strong asymmetry*

In this section, we show how $\hat{C}_{X\Rightarrow Y}$ can detect the true causal direction in many synthetic datasets when the postulate of Lemma 3.1 is satisfied in the strong asymmetry framework. Further, in some more examples, we will illustrate cases when our method fails. Unless explicitly stated, we normalize the data to $[0, 1]$ (i.e., we choose the uniform reference measure) and use the empirical estimator $\hat{C}_{X\Rightarrow Y} = \hat{H}(X) - \hat{H}(Y)$, as given by (13).

We consider some artificial data sets in various simple deterministic situations. Let $Y = g(X)$, where cause $X \sim f_X$ and $g$ is some bijective function such that regions, where the slope of $g$ is high, do not coincide with peaks of the input density $f_X$. In our experiments, we choose $f_X$ to be the density of $U(0, 1)$ and $N(0, 1)$ distributions. We set $g(x)$ to be one of the following functions: $\mathrm{sgn}(x)|x|^{1/3}, \mathrm{sgn}(x)|x|^{1/2}, \mathrm{sgn}(x)|x|^2$, and $\mathrm{sgn}(x)|x|^3$, where $\mathrm{sgn}(x)$ denotes the sign

**Table 2.** Examining $\hat{C}_{X\Rightarrow Y}(95\%\ \text{CI})$ in the strong asymmetry framework. We consider two choices of $f_X$ and four choices of bijective $g$.

| $f_X$ | $g(x)=\text{sgn}(x)\lvert x\rvert^{1/3}$ | $g(x)=\text{sgn}(x)\lvert x\rvert^{1/2}$ | $g(x)=\text{sgn}(x)\lvert x\rvert^{2}$ | $g(x)=\text{sgn}(x)\lvert x\rvert^{2}$ |
|---|---|---|---|---|
| $X\sim U(0,1)$ | $0.41\ (0.34, 0.48)$ | $0.24\ (0.20, 0.28)$ | $0.26\ (0.20, 0.33)$ | $0.65\ (0.55, 0.77)$ |
| $X\sim N(0,1)$ | $-0.09\ (-0.19, -0.01)$ | $-0.33\ (-0.45, -0.22)$ | $1.02\ (0.87, 1.16)$ | $5.05\ (4.62, 5.61)$ |

of $x$. Note that the first two choices of $g$ violate the conditions of Lemma 3.1 for the input density $X\sim N(0,1)$, so these are cases where we expect our method to yield wrong conclusions.

For each setting, the experiments were repeated $R=200$ times, with $n=1000$ bivariate *i.i.d.* samples on $(X,Y)$ drawn each time. For each simulated sample, we study the estimate $\{\hat{C}^r_{X\Rightarrow Y}\}_{r=1}^R$ and report the mean and quantile-based 95% confidence interval of $\hat{C}_{X\Rightarrow Y}$ in Table 6.3. In all the four cases of $g$ when $X\sim U(0,1)$, $\hat{C}_{X\Rightarrow Y}$ is successfully able to detect the correct causal direction. As expected, our method fails when $X\sim N(0,1)$ and $g$ is given either by $\text{sgn}(x)|x|^{1/3}$ or $\text{sgn}(x)|x|^{1/2}$. A simple graphical investigation reveals a violation of the postulate of Lemma 3.1. However, our method is able to successfully capture the correct causal direction for the other two choices of $g$ (i.e., for $\text{sgn}(x)|x|^2$ or $\text{sgn}(x)|x|^3$) which do not violate the postulate when $X\sim N(0,1)$.

## 7. Real data examples

In this section, we provide real-world evidence of the validity of our method. We choose three illustrative examples from the CauseEffectPairs benchmark dataset presented by Mooij and Janzing (2010). The dataset consists a number of "cause-effect pairs", each one consisting of samples of a pair of statistically dependent random variables, where one variable is known to cause the other one. The task is to identify for each pair which of the two variables is the cause and which one is the effect, using the observed samples only. The data sets were selected such that we expect common agreement on the ground truth. The three data sets we choose are of different sample sizes and are taken from different areas of observational research. Similar to Section 6.3, we normalize the data to $[0,1]$ (i.e., we choose the uniform reference measure) and use the empirical estimator $\hat{C}_{X\Rightarrow Y}=\hat{H}(X)-\hat{H}(Y)$, as given by (13).

### *7.1. Social-economic status ⇒ Language test scores in the Netherlands*

This data was used by Tom et al. (1999) and pertains to a study involving $n=2287$ eighth-grade students, approximately 11 years of age, distributed across 132 classes within 131 schools in the Netherlands. This dataset encompasses two key variables: a language test score denoted

$(X)$ and the socio-economic status of the student's family, represented as $(Y)$.

It is hypothesized that the socio-economic status of the student's family significantly influences the student's language test score. Our method supports this claim, with $\hat{C}_{X \Rightarrow Y}$ (95% CI) given by $-0.312$ $(-0.358, -0.266)$.

### *7.2. Stock Return of Hang Seng Bank ⇒ Stock Return of HSBC Holdings during 2000 - 2005*

These data represent stock return values sourced from Yahoo Finance between January 4, 2000, and June 17, 2005. The data has $n = 1331$ samples from two stocks: Hang Seng Bank and HSBC Holdings. Mooij et al. (2016) further pre-processes the data as follows: (i) when price data is missing for a few days, employ linear interpolation to estimate the missing prices, then (ii) define the closing price on day $t$ for each stock as $P_t$, and compute the corresponding return as $R_t = (P_t - P_{t-1})/P_{t-1}$. We let $X$ denote the stock return data for Hang Seng Bank and $Y$ denote the stock return data for HSBC.

During 2000 - 2005, HSBC owned about 60% of Hang Seng Bank (Mooij et al., 2016). Thus, any fluctuations in Hang Seng Bank's stock returns are expected to exert an influence on the stock returns of HSBC, with a weaker causal influence expected in the opposite direction. This is reflected by our method, with $\hat{C}_{X \Rightarrow Y}$ (95% CI) given by $0.145$ $(0.066, 0.224)$.

### *7.3. Population growth ⇒ Food consumption growth in the UN FAO dataset*

The data set has been collected by Food and Agriculture Organization (FAO) of the United Nations (UN) and is accessible online. This comprehensive dataset encompasses 174 countries or regions over two distinct time intervals, spanning from 1990–92 to 1995–97 and from 1995–97 to 2000–02. In total, there are $n = 347$ data points.

Within this dataset, we focus on two specific variables: population growth $(X)$ and food consumption $(Y)$. The former variable quantifies the average annual rate of population change (expressed in percentage), while the latter variable characterizes the average annual rate of dietary consumption growth for the entire population (measured in kcal/day and also expressed in percentage). Our working hypothesis posits that population growth serves as a primary driver for the growth in food consumption since it is understood that an increased population drives a corresponding increase in overall food consumption. This claim is supported by our method, with $\hat{C}_{X \Rightarrow Y}$ (95% CI) given by $0.113$ $(0.059, 0.167)$.

## 8. Discussion

Asymmetry is an inherent property of bivariate associations and therefore must not be ignored. Most dependence measures mask potential asymmetry by implicitly assuming that variables $X$ and $Y$ are equally dependent on each other, which may be false. We present a new causal discovery framework of asymmetric predictability between two random variables $X$ and $Y$ using $C_{X\Rightarrow Y}$. The measure $C_{X\Rightarrow Y}$ is inspired by Shannon's seminal work on information theory and is a well-justified tool to detect and quantify weak asymmetry, thereby serving as an attractive causal discovery tool. If further assumptions can be made on the functional dependence of the effect on the cause, $C_{X\Rightarrow Y}$ may be used to quantify strong asymmetry, thereby enabling the user to directly examine the direction of causation. While standard causal discovery approaches do not incorporate the effect of confounding, we present an approach to infer directionality even in the presence of (low-dimensional) confounders. Future work will involve extending our method to high dimensional confounders.

A computationally fast and robust Fourier transformation-based method is used to estimate $C_{X\Rightarrow Y}$ instead of conventional kernel-based methods. This approach is data driven and does not rely on user input for bandwidth selection. Moreover, our method consistently performs faster than existing bandwidth-dependent methods - being approximately 4 times faster for bivariate sample sizes of approximately $10^4$ (Purkayastha and Song, 2023). Another contribution of our $C_{X\Rightarrow Y}$ methodology is data-splitting inference, which enjoys key large-sample properties necessary for valid inference. This new approach enables us to establish asymptotic normality for functionals of PDFs, which is widely regarded as a difficult issue to address.

Our simulations and data analysis clearly demonstrate the necessity and universal applicability of the quantification of asymmetric predictability in bivariate associations, thereby establishing an attractive causal discovery framework. Potential applications of our $C_{X\Rightarrow Y}$ framework include mediation analysis and instrumental variable methods, in which implicit assumptions are made about causal directions, often without justification. In the absence of *a priori* knowledge, our framework may serve either as a discovery or confirmatory tool, thereby aiding many applications in current statistical research, particularly in the investigation of causality.

## References

Amari, S.-I. (2001) Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, **47**, 1701–1711.

Bernacchia, A. and Pigolotti, S. (2011) Self-consistent method for density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 407–422. URL: https://doi.org/10.1111/j.1467-9868.2011.00772.x.

Billingsley, P. (1995) *Probability and measure*. A Wiley-Interscience publication. New York [u.a.]: Wiley, 3. ed edn. URL: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+164761632&sourceid=fbw_bibsonomy.

Casella, G. and Berger, R. L. (2021) *Statistical inference*. Cengage Learning.

Chaitin, G. J., Arslanov, A. and Calude, C. (1995) Program-size complexity computes the halting problem. *Tech. rep.*, Department of Computer Science, The University of Auckland, New Zealand.

Chatterjee, S. (2020) A new coefficient of correlation. *Journal of the American Statistical Association*, **116**, 2009–2022. URL: https://doi.org/10.1080/01621459.2020.1758115.

Choi, J., Chapkin, R. and Ni, Y. (2020) Bayesian causal structural learning with zero-inflated poisson bayesian networks. *Advances in neural information processing systems*, **33**, 5887–5897.

Cover, T. M. and Thomas, J. A. (2005) *Elements of Information Theory*. Wiley. URL: https://doi.org/10.1002%2F047174882x.

Cox, D. R. (1990) Role of models in statistical analysis. *Statistical Science*, **5**, 169–174.

— (1992) Causality: some statistical aspects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **155**, 291–301.

Czado, C. (2019) *Analyzing Dependent Data with Vine Copulas: A Practical Guide With R (Lecture Notes in Statistics, 222)*. Springer, paperback edn. URL: https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=3030137848.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K. and Schölkopf, B. (2012) Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*.

Dembo, A., Cover, T. and Thomas, J. (1991) Information theoretic inequalities. *IEEE Transactions on Information Theory*, **37**, 1501–1518.

Engelhardt, H., Kohler, H.-P. and Prskawetz, A. (2009) *Causal analysis in population studies*. Springer.

Friedman, N. and Nachman, I. (2000) Gaussian process networks. *UAI'00: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, **16**.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B. and Smola, A. (2007) A kernel statistical test of independence. *Advances in neural information processing systems*, **20**.

Heller, R., Heller, Y. and Gorfine, M. (2012) A consistent multivariate test of association based on ranks of distances. *Biometrika*, **100**, 503–510. URL: https://doi.org/10.1093/biomet/ass070.

Hernandez-Lobato, D., Morales-Mombiela, P., Lopez-Paz, D. and Suarez, A. (2016) Non-linear causal inference using gaussianity measures. *The Journal of Machine Learning Research*, **17**, 939–977.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J. and Schölkopf, B. (2008) Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, **21**.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniusis, P., Steudel, B. and Scholkopf, B. (2012) Information-geometric approach to inferring causal directions. *Artificial Intelligence*, **182**, 1–31.

Janzing, D. and Schölkopf, B. (2010) Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, **56**, 5168–5194.

Joe, H. (2014) *Dependence Modeling with Copulas (Chapman & Hall/CRC Monographs on Statistics and Applied Probability)*. Chapman and Hall/CRC, hardcover edn. URL: https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1466583223.

Kolmogorov, A. N. (1963) On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, 369–376.

Kreyszig, E. (2020) *Advanced Engineering Mathematics*. Wiley, loose leaf edn. URL: https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1119455928.

Lemeire, J. and Dirkx, E. F. (2006) Causal models as minimal descriptions of multivariate systems. URL: https://api.semanticscholar.org/CorpusID:12004610.

Mooij, J. and Janzing, D. (2010) Distinguishing between cause and effect. In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008* (eds. I. Guyon, D. Janzing

and B. Schölkopf), vol. 6 of *Proceedings of Machine Learning Research*, 147–156. Whistler, Canada: PMLR.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J. and Schölkopf, B. (2016) Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, **17**, 1–102. URL: http://jmlr.org/papers/v17/14-518.html.

Nelsen, R. B. (2006) *An Introduction to Copulas*. Springer New York. URL: https://doi.org/10.1007%2F0-387-28678-0.

O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D. and O'Brien, J. P. (2016) A fast and objective multidimensional kernel density estimation method: fastKDE. *Computational Statistics & Data Analysis*, **101**, 148–160. URL: https://doi.org/10.1016/j.csda.2016.02.014.

Orlitsky, A. (2003) Information theory. In *Encyclopedia of Physical Science and Technology*, 751–769. Elsevier. URL: https://doi.org/10.1016/b0-12-227410-5/00337-9.

Pearl, J. (2010) Brief report: On the consistency rule in causal inference:" axiom, definition, assumption, or theorem?". *Epidemiology*, 872–875.

Pearl, J. et al. (2000) *Models, reasoning and inference*, vol. 19.

Purkayastha, S. and Song, P. X. K. (2023) fastmi: a fast and consistent copula-based estimator of mutual information.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 550–560.

Rubin, D. B. (2005) Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, **100**, 322–331.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423. URL: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sklar, M. J. (1959) Fonctions de repartition a n dimensions et leurs marges.

Spirtes, P. L. and Zhang, K. (2016) Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, **3**.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **35**. URL: https://doi.org/10.1214/009053607000000505.

Tagasovska, N., Chavez-Demoulin, V. and Vatter, T. (2020) Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *Proceedings of the 37th International Conference on Machine Learning* (eds. H. D. III and A. Singh), vol. 119 of *Proceedings of Machine Learning Research*, 9311–9323. PMLR. URL: https://proceedings.mlr.press/v119/tagasovska20a.html.

Tom, A., Bosker, T. A. S. R. J. and Bosker, R. J. (1999) *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage.

Zhang, K. and Hyvärinen, A. (2009) On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, 647–655. Arlington, Virginia, USA: AUAI Press.