We Are Not Your Real Parents: Telling Causal from Confounded using MDL

David Kaltenpoth[°] Jilles Vreeken[•]

Abstract

Given data over variables $(X_1, ..., X_m, Y)$ we consider the problem of finding out whether X jointly causes Y or whether they are all confounded by an unobserved latent variable Z. To do so, we take an information-theoretic approach based on Kolmogorov complexity. In a nutshell, we follow the postulate that first encoding the true cause, and then the effects given that cause, results in a shorter description than any other encoding of the observed variables.

The ideal score is not computable, and hence we have to approximate it. We propose to do so using the Minimum Description Length (MDL) principle. We compare the MDL scores under the models where X causes Y and where there exists a latent variables Z confounding both X and Y and show our scores are consistent. To find potential confounders we propose using latent factor modeling, in particular, probabilistic PCA (PPCA).

Empirical evaluation on both synthetic and real-world data shows that our method, CoCa, performs very well—even when the true generating process of the data is far from the assumptions made by the models we use. Moreover, it is robust as its accuracy goes hand in hand with its confidence.

1 Introduction

Causal inference from observational data, i.e. inferring cause and effect from data that was not collected through randomized controlled trials, is one of the most challenging and important problems in statistics [21]. One of the main assumptions in causal inference is that of *causal sufficiency*. That is, to make sensible statements on the causal relationship between two statistically dependent random variables X and Y, it is assumed that there exists no hidden confounder Z that causes both X and Y. In practice this assumption is often violated—we seldom know all factors that could be relevant, nor do we measure everything—and hence existing methods are prone to spurious inferences.

In this paper, we study the problem of inferring whether X and Y are causally related, or, are more likely jointly caused by an unobserved confounding variable Z. To do so, we build upon the algorithmic Markov condition (AMC) [8]. This recent postulate states that the simplest—measured in terms of Kolmogorov complexity—factorization of the joint distribution coincides with the true causal model. Simply put, this means that if Z causes both X and Y the complexity of the factorization according to this model, $K(P(Z)) + K(P(X|Z)) + K(P(Y|Z))$, will be lower than the complexity corresponding to the model where X causes Y, $K(P(Z)) + K(P(X)) + K(P(Y|X))$. As we obviously do not have access to $P(Z)$, we propose to estimate it using latent factor modelling. Second, as Kolmogorov complexity is not computable, we use the Minimum Description Length (MDL) principle as a well-founded approach to approximate it from above. This is the method that we develop in this paper.

In particular, we consider the setting where given a sample over the joint distribution $P(\mathbf{X}, Y)$ of continuous-valued univariate or multivariate random variable $\mathbf{X} = (X_1, \ldots, X_m)$, and a continuous-valued scalar Y. Although it has received little attention so far, we are not the first to study this problem. Recently, Janzing and Schölkopf [9, 10] showed how to measure the "structural strength of confounding" for linear models using resp. spectral analysis [9] and ICA [10]. Rather than implicitly measuring the significance, we explicitly model the hidden confounder Z via probabilistic PCA. While this means our approach is also linear in nature, it gives us the advantage that we can fairly compare the scores for the models $X \rightarrow Y$ and $X \leftarrow Z \rightarrow Y$, allowing us to define a reliable confidence measure.

Through extensive empirical evaluation on synthetic and real-world data, we show that our method, CoCa, short for Confounded-or-Causal, performs well in practice. This includes settings where the modelling assumptions hold, but also in adversarial settings where they do not. We show that CoCa beats both baselines as well as the recent proposals mentioned above. Importantly, we observe that our confidence score strongly correlates with accuracy. That is, for those cases where we observe a large difference between the scores for causal resp. confounded, we can trust CoCa to provide highly accurate inferences.

The main contributions of this paper are as follows, we

(a) extend the AMC with latent factor models, and propose to instantiate it via probabilistic PCA,
(b) define a consistent and easily computable MDL-score to instantiate the framework in practice,
(c) provide extensive evaluation on synthetic and real data, including comparisons to the state-of-the-art.

This paper is structured as usual. In Sec. 2 we introduce basic concepts of causal inference, and hidden confounders. We formalize our information theoretic approach to inferring

[°] Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany. dkaltenpo@mpi-inf.mpg.de

[•] Helmholtz Center for Information Security and Max Planck Institute for Informatics, Saarbrücken, Germany. jv@cispa-helmholtz.de

causal or confounded in Sec. 3. We discuss related work in Sec. 4, and present the experiments in Sec. 5. Finally, we wrap up with discussion and conclusions in Sec. 6.

2 Causal Inference and Confounding

In this work, we consider the setting where we are given n samples from the joint distribution $P(\mathbf{X}, Y)$ over two statistically dependent continuous-valued random variables $\mathbf{X}$ and Y. We require Y to be a scalar, i.e. univariate, but allow $\mathbf{X} = (X_1, \ldots, X_m)$ to be of arbitrary dimensionality, i.e. univariate or multivariate. Our task is to determine whether it is more likely that $\mathbf{X}$ jointly cause Y, or that there exists an unobserved random variable $\mathbf{Z} = (Z_1, \ldots, Z_k)$ that is the cause of both $\mathbf{X}$ and Y. Before we detail our approach, we introduce some basic notions, and explain why the straightforward solution does not work.

2.1 Basic Setup It is impossible to do causal inference from observational data without making assumptions [21]. That is, we can only reason about what we should observe in the data if we were to change the causal model, if we assume (properties of) a causal model in the first place.

A core assumption in causal inference is that the data was drawn from a probabilistic graphical model, a casual directed acyclic graph (DAG). To have a fighting chance to recover this causal graph G from observational data, we have to make two further assumptions. The first, and in practice most troublesome is that of *causal sufficiency*. This assumption is satisfied if we have measured *all* common causes of *all* measured variables. This is related to Reichenbach's principle of common cause [25], which states that if we find that two random variables X and Y are statistically dependent, denoted as $X \not\perp\!\!\!\perp Y$, there are three possible explanations. Either X causes Y, $X \rightarrow Y$, or, the other way around, Y causes X, $X \leftarrow Y$, or there is a third variable Z that causes both X and Y, $X \leftarrow Z \rightarrow Y$. In order to determine the latter case, we need to have measured Z.

The second additional assumption we have to make is that of *faithfulness*, which is defined as follows.

Definition 1. (Faithfulness) *If a Bayesian network G is faithful to a probability distribution P, then for each pair of nodes X_i and X_j in G, X_i and X_j are adjacent in G iff. $X_i \not\perp\!\!\!\perp X_j \mid \mathbf{Z}$, for each $\mathbf{Z} \subset G$, with $X_i, X_j \notin \mathbf{Z}$.*

In other words, if we measure that X is independent of Y, denoted as $X \perp\!\!\!\perp Y$, there is no direct influence between the two in the underlying causal graph. This is a strong, but generally reasonable assumption; after all, violations of this condition do generally not occur unless the distributions have been specifically chosen to this end.

Under these assumptions, Pearl [21] showed that we can factorize the joint distribution over the measured variables,

$$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(X_i \mid PA_i) \; .$$

That is, we can write it as a product of the marginal distributions of each X_i conditioned on its true causal parents PA_i. This is referred to as the causal Markov condition, and implies that, under all of the above assumptions, we can have a hope of reconstructing the causal graph from a sample from the joint distribution.

2.2 Crude Solutions That Do Not Work Based on the above, many methods have been proposed to infer causal relationships from a dataset. We give a high level overview of the state of the art in Sec. 4. Here, we continue to discuss why it is difficult to determine whether a given pair $\mathbf{X}, Y$ is confounded or not, and in particular, why traditional approaches based on probability theory or (conditional) independence, do not suffice.

To see this, let us first suppose that $\mathbf{X}$ causes Y, and there are is no hidden confounder $\mathbf{Z}$. We then have $P(\mathbf{X}, Y) = P(\mathbf{X})P(Y|\mathbf{X})$, and $\mathbf{X} \not\perp\!\!\!\perp Y$. Now, let us suppose instead that $\mathbf{Z}$ causes $\mathbf{X}$ and Y, i.e. $\mathbf{X} \leftarrow \mathbf{Z} \rightarrow Y$. Then we would have $P(\mathbf{X}, Y, \mathbf{Z}) = P(\mathbf{Z})P(\mathbf{X}|Z)P(Y|\mathbf{Z})$, and, importantly, while $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, we still observe $X \not\perp\!\!\!\perp Y$, and hence cannot determine causal or confounded on that alone.

Moreover, as we are only given a sample over $P(\mathbf{X}, Y)$ for which $X \not\perp\!\!\!\perp Y$ holds, but know nothing about $\mathbf{Z}$ or $P(\mathbf{Z})$, we cannot directly measure $\mathbf{X} \perp\!\!\!\perp Y \mid \mathbf{Z}$. A simple approach would be to see if we can *generate* a $\hat{\mathbf{Z}}$ such that $\mathbf{X} \perp\!\!\!\perp Y \mid \hat{\mathbf{Z}}$; for example through sampling or optimization. However, as we have to assign n values for $\hat{\mathbf{Z}}$, this means we have n degrees of freedom, and it easy to see that under these conditions it is *always* possible to generate a $\hat{\mathbf{Z}}$ that achieves this independence, even when there was no confounding $\mathbf{Z}$. A trivial example is to simply set $\hat{\mathbf{Z}} = \mathbf{X}$.

A similarly flawed idea would be to decide on the likelihoods of the data alone, i.e. to see if we can find a $\hat{\mathbf{Z}}$ for which $P(\hat{\mathbf{Z}})P(\mathbf{X}|\hat{\mathbf{Z}})P(Y|\hat{\mathbf{Z}}) > P(\hat{\mathbf{Z}})P(\mathbf{X})P(Y|\mathbf{X})$. Besides having to choose a prior on $\mathbf{Z}$, as we already achieve equality by initializing $\hat{\mathbf{Z}} = \mathbf{X}$ and have n degrees of freedom, we again virtually always will find a $\hat{\mathbf{Z}}$ for which this holds, regardless of whether there was a true confounder or not.

Essentially, the problem here is that it is too easy to find a $\hat{\mathbf{Z}}$ where these conditions hold, which for a large part is due to the fact that we do not take the complexity of $\hat{\mathbf{Z}}$ into account, and hence face the problem of overfitting. To avoid this, we take an information theoretic approach, such that in principled manner we can take both the complexity of $\hat{\mathbf{Z}}$, as well as its effect on $\mathbf{X}$ and Y into account.

3 Telling Causal from Confounded by Simplicity

We base our approach on the algorithmic Markov condition, which in turn is based on the notion of Kolmogorov complexity. We first give short introductions to both notions, and then develop our approach.

3.1 Kolmogorov Complexity The Kolmogorov complexity of a finite binary string x is the length of the shortest program p^* for a universal Turing machine $\mathcal{U}$ that generates x and then halts [18, 14]. Formally,

$$K(x) = \min \left\{ |p| : p \in \{0,1\}^*, \mathcal{U}(p) = x \right\} \ .$$

That is, program p^* is the most succinct *algorithmic* description of x, or, in other words, the ultimate lossless compressor of that string. For our purpose, we are particularly interested in the Kolmogorov complexity of a distribution P,

$$K(P) = \min \left\{ |p| : p \in \{0,1\}^*, |\mathcal{U}(x,p,q) - P(x)| \leq 1/q \right\} \ ,$$

which is the length of the shortest program p^* for a universal Turing machine $\mathcal{U}$ that approximates P arbitrarily well [18].

By definition, Kolmogorov complexity will make maximal use of any structure in the input that can be used to compress the object. As such it is the theoretically optimal measure for complexity. Due to the halting problem, Kolmogorov complexity is also not computable, nor approximable up to arbitrary precision [18]. The Minimum Description Length (MDL) principle [4], however, provides a statistically well-founded approach to approximate it from above. We will later use MDL to instantiate the framework we define below.

3.2 Algorithmic Markov Condition Recently, Janzing and Schölkopf [8] postulated the *algorithmic* Markov condition (AMC), which states that if X causes Y, the factorization of the joint distribution over X and Y in the true causal direction has a lower Kolmogorov complexity than in the anti-causal direction, i.e.

$$K(P(X)) + K(P(Y|X)) \leq K(P(Y)) + K(P(X|Y))$$

holds up to an additive constant. Moreover, under the assumption of causal sufficiency this allows us to identify the true causal network as the least complex one,

$$K(P(X_1, \ldots, X_m)) = \min_G \sum_{i=1}^{m} K(P(X_i|PA_i)) \ , \qquad (3.1)$$

which again holds up to an additive constant.

3.3 AMC and Confounding Although the algorithmic Markov condition relies on causal sufficiency, it does suggest a powerful inference framework where we do allow variables to be unobserved. For simplicity of notation, as well as generality, let us ignore Y for now, and instead consider the question whether $\mathbf{X}$ is confounded by some factor $\mathbf{Z}$. We can answer this question using the AMC by including a latent variable $\mathbf{Z} = (Z_1, ..., Z_k)$, where we assume the Z_j's to be independent, of which we know the joint distribution corresponding to measured $\mathbf{X}$ and unmeasured $\mathbf{Z}$, $P(\mathbf{X}, \mathbf{Z})$. If this is the case, we can again simply identify the corresponding minimal Kolmogorov complexity network via

$$K(P(\mathbf{X}, \mathbf{Z})) = \min_G \sum_{i=1}^{m} K(P(X_i|PA_i)) + \sum_{j=1}^{k} K(P(Z_j)) \ , \qquad (3.2)$$

where PA_i are now the parents of X_i among $\left\{X_l, Z_j\right\}$ in the extended network. By adding terms $K(P(Z_j))$ we implicitly assume that there is no reverse causality $\mathbf{X} \rightarrow \mathbf{Z}$.

This formulation gives us a principled manner to identify whether a given $P(\mathbf{Z})$ is a (likely) confounder of $\mathbf{X}$. Clearly, with the above we can score the hypothesis $\mathbf{Z} \rightarrow \mathbf{X}$. However, it also allows us to fairly score the hypothesis $\mathbf{Z} \perp\!\!\!\perp \mathbf{X}$, because if we choose $P(\mathbf{Z})$ to be a prior concentrated on a single point, this corresponds to Eq. (3.1) up to an additive constant. By the algorithmic Markov condition, we can now determine the most likely causal model, simply by comparing the two scores and choosing the one with the lower Kolmogorov complexity. This approach does not suffer from the same problems as in Sec. 2.2 as we explicitly take the complexity of $P(\mathbf{Z})$ into account. Moreover, and importantly, this formulation allows us to consider *any* distribution $P(\mathbf{X}, \mathbf{Z})$ with *any* type of latent factor $\mathbf{Z}$.

Two problems, however, do remain with this approach. First, we do not know the true distribution $P(\mathbf{X}, \mathbf{Z})$, nor even distributions $P(\mathbf{X})$ or $P(\mathbf{Z})$. Instead we only have empirical data over $\mathbf{X}$ from which we can approximate $\hat{P}(\mathbf{X})$, but this does not give us explicit information about $\mathbf{Z}$, $P(\mathbf{Z})$ or the joint $P(\mathbf{X}, \mathbf{Z})$. Second, as stated above, Kolmogorov complexity is not computable and the criterion as such therefore not directly applicable. We will deal with the first problem next by making assumptions on the form of $P(\mathbf{X}, \mathbf{Z})$, and then in Sec. 3.5 will instantiate this criterion using the Minimum Description Length (MDL) principle.

3.4 Latent Factor Models Even under the assumption that the Z_j are mutually independent, there are infinitely many possible distributions $P(\mathbf{X}, \mathbf{Z})$, and hence we have to make further choices to make the problem feasible. In our setting, a particularly natural choice is to use latent factor modelling. That is, we say the distribution over $\mathbf{X}, \mathbf{Z}$ should be of the form

$$P(\mathbf{X}, \mathbf{Z}) = P(\mathbf{Z}) \prod_{i=1}^{m} P(X_i|\mathbf{Z})$$

where the distribution of $\mathbf{Z}$ can be arbitrarily complex. Not only does this give us a very clear and interpretable hypothesis, namely that given $P(\mathbf{Z})$, every X_i should be independent of every other member of $\mathbf{X}$, i.e. $X_i \perp\!\!\!\perp X_j \mid \mathbf{Z}$, it also corresponds to the notion that $P(\mathbf{Z})$ should explain away *as much* of the information shared within $\mathbf{X}$ as possible—very much in line with Eq. (3.2). Moreover, from a more practical perspective, it is also a well-studied problem for which advanced techniques exist, such as Factor Analysis [19], GPLVM [16], Deep Generative Models [12, 26, 24], as well as Probabilistic PCA (PPCA) [34].

For the sake of simplicity we will here focus on using PPCA, which has the following linear form

$$\begin{aligned} Z_i &\sim \text{Normal}(0, \sigma_z^2 I) \qquad (3.3)\\ W_i &\sim \text{Normal}(0, \sigma_w^2 I)\\ \mathbf{X}|\mathbf{Z},\mathbf{W} &\sim \text{Normal}(\mathbf{W}^t\mathbf{Z}, \sigma_x^2 I)\ , \end{aligned}$$

and is appropriate if we deal with real-valued variables without any constraints and assume Gaussian noise. If the data does not follow these assumptions one of the other models mentioned above may be a more appropriate choice. An appealing aspect of PPCA is that by marginalizing over $\mathbf{Z}$ we can rewrite it in only terms of the matrix W [34], i.e.

$$\begin{aligned} W_i &\sim \text{Normal}(0, \sigma_w^2 I) \qquad (3.4)\\ \mathbf{X}|\mathbf{W} &\sim \text{Normal}(0, \mathbf{W}\mathbf{W}^t + \sigma_x^2 I)\ , \end{aligned}$$

which both dramatically reduces the computational effort as well as will allow us to make statements about the consistency of our method.

While in the simple form PPCA assumes linear relationships, we can also model non-linear relationships by adding features to conditional distribution $\mathbf{X}|\mathbf{Z},\mathbf{W}$, e.g. using polynomial regression of $\mathbf{X}$ on $\mathbf{Z}$. While this increases the modelling power, it comes with an increase in computational effort as the simplification of Eq. (3.4) no longer holds.

3.5 Minimum Description Length While Kolmogorov complexity is not computable, the Minimum Description Length (MDL) principle [27] provides a statistically well-founded approach to approximate $K(\cdot)$ from above. To achieve this, rather than considering all Turing machines, in MDL we consider a model class $\mathcal{M}$ for which we *know* that every model $M \in \mathcal{M}$ will generate the data and halt, and identify the best model $M^* \in \mathcal{M}$ as the one that describes the data most succinctly without loss. If we instantiate $\mathcal{M}$ with all Turing machines that do so, the MDL-optimal model coincides with Kolmogorov complexity—this is also known as Ideal MDL [4]. In practice, we of course consider smaller model classes that are easier to handle and match our modelling assumptions.

In two-part, or, *crude* MDL, we score models $M \in \mathcal{M}$ by first encoding the model, and then the data given that model,

$$L(X, M) = L(M) + L(X \mid M)\ ,$$

where $L(M)$ and $L(X|M)$ are code length functions for the model, and the data conditional on the model, respectively.

Two-part MDL often works well in practice, but, by encoding the model separately it introduces arbitrary choices. In one part MDL—also known as *refined* MDL—we avoid these choices by encoding the data using the entire model class at once. In order for a code length function to be refined, it has to be asymptotically mini-max optimal. That is, no matter what data X' of the same type and dimensions as X we consider, the refined score for X' is within a constant from the score where we already know its corresponding optimal model M'^*, $L(X' \mid M'^*)$, and this constant is independent of the data. There exist different forms of refined MDL codes [4]. For our setup it is convenient to use the full Bayesian definition,

$$L(X|\mathcal{M}) = -\log \int_{M\in\mathcal{M}} P(X|M)dP(M)$$

where $P(M)$ is a prior on the model class $\mathcal{M}$. In our case, that is, for the PPCA model from 3.3 each pair $\mathbf{Z},\mathbf{W}$ corresponds to one model M, and hence the model class to all possible $\mathbf{Z},\mathbf{W}$ of which the posterior is given by Eq. (3.3), i.e. we have

$$L(\mathbf{X}|\mathcal{M}) = -\log \int p(\mathbf{X}|\mathbf{Z},\mathbf{W})p(\mathbf{Z})p(\mathbf{W})d\mathbf{W}d\mathbf{Z}\ .$$

We can now put all the pieces together, and use the above theory to determine whether a pair $\mathbf{X}, Y$ is more likely causally related or confounded by an unobserved $\mathbf{Z}$.

3.6 Causal or Confounded? Given the above theory, determining which of $\mathbf{X} \to Y$ and $\mathbf{X} \leftarrow \mathbf{Z} \to Y$ is more likely, is fairly straightforward. To do so, we consider two model classes, one for each of these two hypotheses, and determine which of the two leads to the most succinct description of the data sample over $\mathbf{X}$ and Y.

First, we consider the causal model class $\mathcal{M}_{\text{ca}}$ that consists of models where $\mathbf{X}$ causes Y in linear fashion,

$$\begin{aligned} X_i &\sim \text{Normal}(0, \sigma_x^2 I) \qquad (3.5)\\ \mathbf{w} &\sim \text{Normal}(0, \sigma_w^2 I)\\ Y|\mathbf{X},\mathbf{w} &\sim \text{Normal}(\mathbf{w}^t\mathbf{X}, \sigma_y^2)\ . \end{aligned}$$

and writing L_{ca} instead of $L(\cdot|\mathcal{M}_{\text{ca}})$ we encode the data as

$$\begin{aligned} L_{\text{ca}}(\mathbf{X}, Y) &= -\log P(\mathbf{X}) \int P(Y \mid \mathbf{X},\mathbf{w})P(\mathbf{w})d\mathbf{w}\\ &\approx -\log P(\mathbf{X})N^{-1}\sum_{j=1}^{N} P(Y|\mathbf{X},\hat{\mathbf{w}}_j) \end{aligned}$$

where we approximate the integral by sampling N weight vectors $\hat{\mathbf{w}}_i$ from the distribution defined by Eq. (3.5).

Second, we consider the *confounded* model class $\mathcal{M}_{\text{co}}$, where the correlations within $\mathbf{X}$ and Y are entirely explained by a hidden confounder modelled by PPCA, i.e.

$$\begin{aligned} L_{\text{co}}(\mathbf{X}, Y) &= -\log \int p(\mathbf{X}, Y|\mathbf{Z}, \mathbf{W}) p(Z) p(\mathbf{W}) d\mathbf{W} d\mathbf{Z} \\ &\approx -\log N^{-1} \sum_{j=1}^{N} p(\mathbf{X}, Y|\hat{\mathbf{Z}}_j, \hat{\mathbf{W}}_j) \end{aligned}$$

where the N samples for $\hat{\mathbf{Z}}_j, \hat{\mathbf{W}}_j$ are drawn from the model we inferred using PPCA, i.e., according to Eq. (3.3). Like for the causal case, the more samples we consider, the better the approximation, but the higher the computational cost.

By MDL we can now check which hypothesis better explains the data, by simply considering the sign of $L_{\text{co}}(\mathbf{X}, Y) - L_{\text{ca}}(\mathbf{X}, Y)$. If this is less than zero, the confounded model does a better job at describing the data than the causal model and vice versa. We refer to this approach as CoCa.

To make the CoCa scores comparable between different data sets, we further introduce the confidence score

$$C = \frac{L(\mathbf{X}, Y|\mathcal{M}_{\text{co}}) - L(\mathbf{X}, Y|\mathcal{M}_{\text{ca}})}{\max\{L(\mathbf{X}, Y|\mathcal{M}_{\text{co}}), L(\mathbf{X}, Y|\mathcal{M}_{\text{ca}})\}} ,$$

which is simply a normalized version of c that accounts for both the intrinsic complexities of the data as well as the number of samples. If the absolute value of C is small both model classes explain the data approximately equally well, and hence we are not very confident in our result and should perhaps refrain from making a decision.

Last, we consider the question of whether we can say when our method will properly distinguish between the cases we care about? For this, we use a general result for MDL on the consistency of deciding between two model classes when the data is generated by a model contained in either of these classes [4]. That is, if we let $\mathbf{X}^n, Y^n$ be n samples for $\mathbf{X}$ and Y then

$$\lim_{n\to\infty} n^{-1} \left(L_{\text{co}}(\mathbf{X}^n, Y^n) - L(\mathbf{X}^n, Y^n)\right) \begin{cases} \leq 0 & \text{if } M^* \in \mathcal{M}_{\text{co}} \\ \geq 0 & \text{if } M^* \in \mathcal{M}_{\text{ca}} \end{cases}$$

with strict inequalities if M^* is contained in only one of the two classes. This means that in the limit we will infer the correct conclusion if the true model is within the model classes we assume. Moreover, since our refined MDL formulation is also consistent for model selection [4], following Sec. 3.2 we expect that even if M^* is contained in both model classes the shortest description of the model M^* corresponds to the true generative process. Importantly, even when the true model is not in either of our model classes, we can still expect reasonable inferences with regard to these model classes; by the minimax property of refined codes we use, we encode every model as efficiently (up to a constant) as possible, which promises reliable performance and confidence scores even in adversarial cases. As we will see shortly, the experiments confirm this.

4 Related Work

Causal inference is arguably one of the most important problems in statistical inference, and hence has attracted a lot of research attention [28, 21, 32]. The existence of confounders, selection bias and other statistical problems make it impossible to infer causality from observational data alone [21]. When their assumptions hold, constraint-based [32, 33, 36] and score-based [2] causal discovery can, however, reconstruct causal graphs up to Markov equivalence. This means, however, they are not applicable to determine the causal direction between just X and Y.

By making assumptions on the shape of the causal process Additive Noise Models (ANMs) can determine the causal direction between just X and Y. In particular, ANMs assume independence between the cause and the residual (noise), and infer causation if such a model can be found in one direction but not in the other [30, 31, 5, 37]. A more general framework for inferring causation than any of the above is given by the Algorithmic Markov Condition (AMC) [17, 8] which is based on finding the least complex – in terms of Kolmogorov complexity – causal network for the data at hand. Since Kolmogorov complexity is not computable [18], practical instantiations require a computable criterion to judge the complexity of a network, which has been proposed to do using Renyi-entropies [13], information geometry [3, 7, 11], and MDL [1, 20]. All of these methods assume causal sufficiency, however, and are not applicable in the case where there are hidden confounders.

Rather than inferring the causal direction between X and Y, estimating the causal effect of X onto Y is also an active topic of research. To do so in the presence of latent variables, Hoyer et al. [6] solve the overcomplete independent component analysis (ICA) problem, whereas Wang and Blei [35] and Ranganath and Perotte [23] control for plausible confounders using a given factor model.

Most relevant to this paper is the recent work by Janzing and Schölkopf on determining the "structural strength of confounding" for a continuous-valued pair $\mathbf{X}, Y$, which they propose to measure using resp. spectral analysis [9] and ICA [10]. Like us, they also focus on linear relationships, but in contrast to us define a one-sided significance score, rather than a two-sided information theoretic confidence score. In the experiments we will compare to these two methods.

5 Experiments

In this section we empirically evaluate CoCa. In particular, we consider performance in telling causal from confounded for both in-model and adversarial settings on both synthetic and real-world data. We compare to the recent methods by Janzing and Schölkopf [9, 10]. We implemented CoCa in Python using PyMC3 [29] for posterior inference via ADVI [15]. All code is available for research purposes.[1]

Throughout this section we infer one-dimensional factor models $\hat{\mathbf{Z}}$, noting that higher-dimensional $\hat{\mathbf{Z}}$ gave similar results. We use N=500 samples to calculate the MDL scores. All experiments were executed single-threaded on an Intex Xeon E5-2643 v3 machine with 256GB memory running Linux, and each run took on the order of seconds to finish.

5.1 Synthetic Data To see whether CoCa works at all, we start by generating synthetic data with known ground truth close to our assumptions. For the confounded case, we generate samples over $\mathbf{X}, Y$ as follows

$$\begin{aligned} Z_j &\sim p_z , & W_{ij} &\sim p_w \\ \epsilon &\sim \text{Normal}(0,1) & \mathbf{X}, Y &= \mathbf{W}^t \mathbf{Z} + \epsilon \,, \end{aligned}$$

while for the causal case, we generate $\mathbf{X}, Y$ as

$$\begin{aligned} X_i &\sim p_x & w_i &\sim p_w \\ \epsilon &\sim \text{Normal}(0,1) & Y &= \mathbf{w}^t \mathbf{X} + \epsilon \,. \end{aligned}$$

To see how our performance depends on the precise generating process, we consider the following source distributions,

$$\begin{aligned} p_z, p_x, p_w \in \{ &\text{Normal}(0,1), \text{Laplace}(0,1), \\ &\text{LogNormal}(0,1), \text{Uniform}(0,1) \} \,. \end{aligned}$$

We expect best CoCa performance when the generating process uses the Normal or Laplace distributions as these are closest to the assumptions made in Eq. (3.3) and Eq. (3.5).

To see how the accuracy of CoCa depends on the confidence assigned to each inference, we consider decision rate (DR) plots. In these, we consider the accuracy over the top-k pairs sorted descending by absolute confidence, $|C|$. This metric is commonly used in the literature on causal inference as it gives more information about the performance of our classifier than simple accuracy scores.

We consider the case where we fix the dimensionality of $\mathbf{Z}$ to be 3, and vary the dimensionality of $\mathbf{X}$ to be 1, 3, 6, 9 and further restrict p_x, p_z, p_w to be $\mathcal{N}(0,1)$, as these are precisely the model assumptions made by CoCa. We show the resulting DR plot in the left plot of Fig. 1. We see that for all dimensionalities of $\mathbf{X}$ the pairs for which CoCa is most confident are also most likely to be classified correctly. While for $\dim(\mathbf{X}) = 1, 3 \leq \dim(\mathbf{Z})$ there is (too) little information about $\mathbf{Z}$ that can be inferred by the factor model, for $\dim(\mathbf{X}) = 6, 9 > \dim(\mathbf{Z})$ CoCa is both highly confident and accurate over all decisions.

Next, we move away from our model assumptions and aggregate over all the possibilities $p_x = p_z = p_w$ listed above.

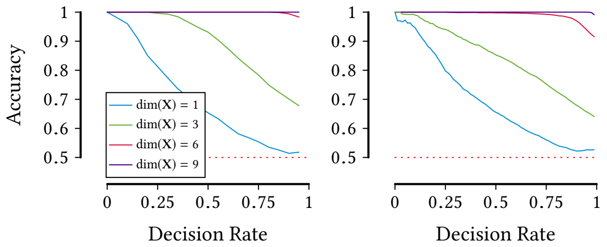

Figure 1: [Higher is better] Accuracy over top-k% pairs sorted by confidence, for different generating models of $\mathbf{X}$, Y, and different dimensionality of $\mathbf{X}$, with $\dim(\mathbf{Z}) = 3$ fixed. The baseline is at 0.5. On the left, $\mathbf{X}$, $\mathbf{Z}$ and W are $\sim$ Normal(0, 1), while on the right, we consider the adversarial case where we use out-of-model source distributions.

We show the results on the right-hand side of Fig. 1. We observe essentially the same pattern, except that all the lines drop off slightly earlier than in the left plot. Experiments where we chose p_x, p_z, p_w independently at random resulted in similar results and are hence not shown for conciseness. This shows us that our method continues to work even when the assumptions we make no longer hold.

Importantly, all results in both experiments are significant with regard to the 95% confidence interval of a fair coin flip—except for $\dim(\mathbf{X}) = 1$ which is significant only for the 75% of tuples where it was most confident. Further, in none of these cases was the method biased towards classifying datasets as causal or confounded.

To see how CoCa fares for a broader variety of combinations of dimensionalities of $\mathbf{X}$ and $\mathbf{Z}$, in Fig. 2 we plot a heatmat of the area under the decision rate curve (AUDR) of CoCa. As expected we see that when $\dim(\mathbf{Z})$ is fixed we become more accurate as $\dim(\mathbf{X})$ increases. Further as $\dim(\mathbf{Z})$ increases for fixed $\dim(\mathbf{X})$ our performance degrades gracefully—this is because we infer a $\hat{\mathbf{Z}}$ of dimensionality one, which deviates further from the true generating process as the dimensionality of the true $\mathbf{Z}$ increases. Note that all CoCa AUDR scores are above 0.75, whereas a random classifier would obtain a score of only 0.5.

Finally, we compare CoCa to the only two competitors we are aware off; the two recent approaches by Janzing and Schölkopf, of which one is based on spectral analysis (SA) [9] and the other on independent component analysis (ICA) [10]. The implementation of both methods require $\mathbf{X}$ to be multidimensional. We hence consider the cases where $\dim(\mathbf{X}) = 3, 6, 9$, while allowing p_x, p_y, p_z to be any of the distributions listed above. We show the results in Fig. 3.

As SA and ICA provide an estimate $\widehat{\beta} \in [0,1]$ measuring the strength of confounding without any two-sided confidence score, we used $|\widehat{\beta} - 1/2|$ as a substitute for such a score. That the corresponding lines are shaped as expected gives

[1] http://eda.mmci.uni-saarland.de/coca/

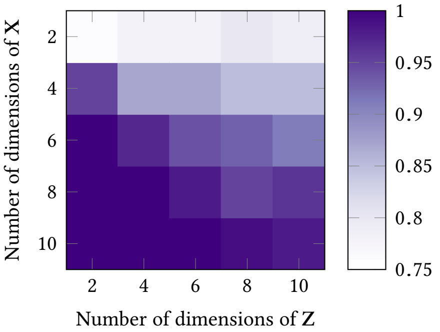

Figure 2: [Darker is better] Area under the Decision Rate Curve heatmap over dimensionality of **X** and **Z**. For fixed dim(**Z**) performance improves as dim(**X**) increases, while for fixed dim(**X**) performance degrades as dim(**Z**) increases. CoCa scores between 0.75 and 1.0, against a baseline of 0.5.

us some assurance that this is a reasonable choice.

We see that for all dimensionalities CoCa outperforms these competitors by a margin where the respective methods are most confident, but also that the overall accuracies are almost indistinguishable. We further note that as the dimensionality of **X** becomes large relative to **Z** the differences in performance between the approaches reduces.

5.2 Simulated Genetic Networks Next, we consider more realistic synthetic data. For this we consider the DREAM 3 data [22] which was originally used to compare different methods for inferring biological networks. We use this data both because the underlying generative network is known, and because the generative dynamics are biologically plausible [22]. That is, the relationships are highly nonlinear, and therefore an interesting case to evaluate how CoCa performs when our assumptions do not hold at all. Out of all networks in the dataset, we consider the ten largest networks, those of 50 and 100 nodes, which are associated with time series of lengths 500 and 1000, respectively. Since CoCa was not designed to work with time series, we treat the data as if it were generated from an i.i.d. source.

For each network we take pairs (X, Y) of univariate X and Y such that either of the following two cases holds

- X has a causal effect on Y and there exists no common parent Z, or
- X, Y have a common parent Z and there are no causal effects between X and Y.

Although in theory we could also consider tuples $(X_1, ..., X_m, Y)$ with $m > 1$, for this dataset there were too few such tuples to have sufficient statistical power. Further, since the original networks are heavily biased towards causality rather than to common parents we take all the confounded tuples and then uniformly sample an equal number of causal tuples from the set of all such tuples.

We show the decision rate plot when applying CoCa to these pairs after aggregating over all the networks in the left-hand side plot of Fig. 4. Like before, we see that CoCa is highly accurate for those tuples where it is most confident. In comparison to the results for $\dim(\mathbf{X}) = 1$ in Fig. 1, we see that performance drops more quickly, which is readily explained by the fact that the simulated dynamics are highly nonlinear. Note however, that our results are nevertheless still statistically significant with regard to a fair coin flip for up the 75% pairs CoCa that is most confident about. To further explain the behavior of CoCa on this dataset, we plot the absolute confidence scores we obtain on the right of Fig. 4. We see that particularly for the first 25% of the decisions the confidences we obtain are much larger than for the remaining pairs. This corresponds very nicely to the plot on the left, as the first 25% of our decisions are also those where we compare most favorably to the baseline.

5.3 Tübingen Benchmark Pairs To consider real-world data suited for causal inference, we now consider the Tübingen benchmark pairs dataset.[2] This dataset consists of (mostly) pairs (X, Y) of univariate variables for which plausible directions of causality can be decided assuming no hidden confounders. For many of these, however, it is either known, or plausible to posit that they are confounded rather than directly causally related. For example, for pairs 65–67 certain stock returns are supposedly causal, but given the nature of the market would likely be better explained by common influences on the returns of the stock options.

We therefore code every pair in the benchmark dataset as either causal (if we think the directly causal part to be stronger), confounded (if we expect the common cause to be the main driver), or unclear (if we are not sure which component is more important) and apply CoCa to the pairs in the first two categories. This leaves 47 pairs, of which we judged 41 to be mostly causal and 6 to be mostly confounded.[3]

In Fig. 5 we show the decision rate plots across the datasets weighed according to the benchmark definition. As in the previous cases, CoCa is most accurate where it is most confident, while declining to the baseline as we try to classify points about which CoCa is less and less certain. We note that for these cases CoCa was biased towards saying that datasets represented truly causal relationships, even when we judged them to be driven by confounding. Despite this, CoCa does better than the naive baseline of "everything is causal" by assigning more confidence to those datasets which according to our judgment were indeed truly causal.

5.4 Optical Data Finally, we consider real world optical data [9]. In these experiments, **X** is a low-resolution (3×3

[2] https://webdav.tuebingen.mpg.de/cause-effect/

[3] The complete list can be found in the online appendix at http://eda.mmci.uni-saarland.de/coca/.

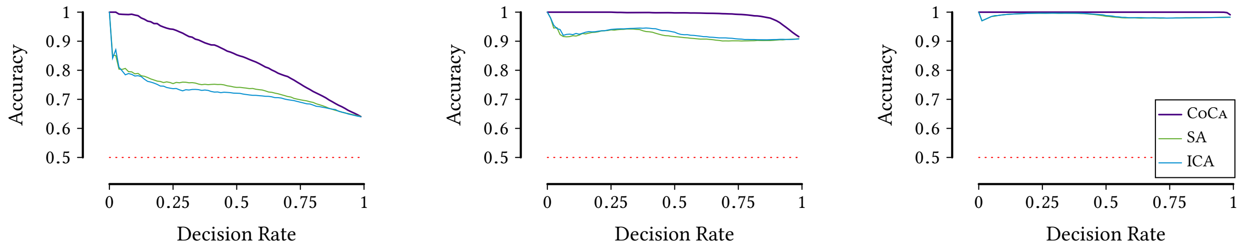

Figure 3: [Higher is better.] Comparing CoCa to the spectral [9] (SA) and ICA-based [10] (ICA) approaches by Janzing and Schölkopf in synthetic data of, from left to right, resp. dim($\mathbf{X}$) = 3, 6, and 9. Baseline accuracy is at 0.5. We see that in all cases, CoCa performs best by a margin, particularly in regions where it is most confident.

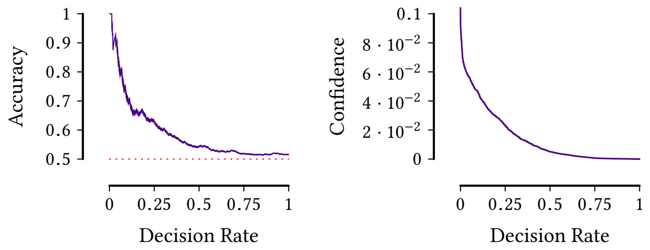

Figure 4: Decision rate and corresponding confidence plots for the genetic networks data. CoCa is accurate when it is confident, even for this adversarial setting.

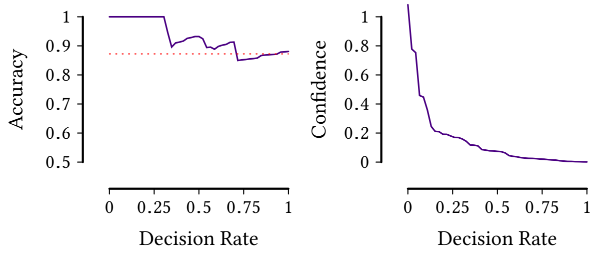

Figure 5: Decision rate plot and its corresponding confidence plot for the Tübingen pairs. The baseline for the decision rate plot is at 0.87. Note the strong correspondence between high confidence and high accuracy.

pixels) image shown on the screen of a laptop and Y is the brightness measured by a photodiode at some distance from the screen. The confounders $\mathbf{Z}$ are an LED in front of the photodiode and another LED in front of the camera, both controlled by random noise, where the strength of confounding is controlled by the brightness of these LEDs.

We evaluate CoCa on each of the provided datasets, and plot the resulting values in Fig. 6. The strength of confounding increases from the left to right, and values larger than zero indicate that CoCa judged the data to be causal, while values smaller than zero indicate confounding.

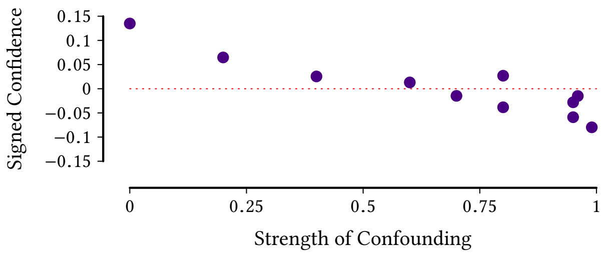

Figure 6: Strength of confounding against the signed confidence of CoCa on the optical data. The confounding strength increases from left to right. Higher positive values indicate a stronger belief in causality of CoCa, while more negative values indicate a stronger belief in confounding.

We see that towards an intermediate confounding strength of 0.5 our method is very uncertain about its classification, while towards the extreme ends of pure causality or pure confounding it is very confident, and correct in being so.

6 Discussion and Conclusions

We considered the problem of distinguishing between the case where the data $(\mathbf{X}, Y)$ has been generated via a genuinely causal model and the case where the apparent cause and effect are in fact confounded by unmeasured variables. We proposed a practical information theoretic way of comparing these cases on the basis of MDL and latent variable models that can be efficiently inferred using variational inference. Through experiments we showed that CoCa works well in practice—including in cases where the data generating process is quite different from our models assumptions. Importantly, we showed that CoCa is particularly accurate when it is also confident, more so than its competitors.

For future work, we will investigate the behavior of CoCa if we use more complex latent variable models, as these allow for modelling more complex relations. These methods, however, also come with a much higher computational cost and without theoretical guarantees of consistency, but may work well in practice. In addition, we would like to be able to infer more complete networks on $(\mathbf{X}, Y)$ while taking into account the presence of confounders. However, this will likely lead to inconsistent inference of edges unless we can find a theoretically well-founded method of telling apart direct and indirect effects. To the best of our knowledge, as of now, no such method is known.

Acknowledgements

David Kaltenpoth is supported by the International Max Planck Research School for Computer Science (IMPRS-CS). Both authors are supported by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government.

References

[1] K. Budhathoki and J. Vreeken. MDL for causal inference on discrete data. In *ICDM*, pages 751–756. IEEE, 2017.

[2] D. M. Chickering. Learning equivalence classes of bayesian-network structures. *JMLR*, 2:445–498, 2002.

[3] P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *UAI*, pages 143–150, 2010.

[4] P. D. Grünwald. *The Minimum Description Length Principle.* MIT press, 2007.

[5] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696, 2009.

[6] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *Int. J. Approx. Reason.*, 49:362–378, 2008.

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

[8] D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE TIT*, 56:5168–5194, 2010.

[9] D. Janzing and B. Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.

[10] D. Janzing and B. Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *ICML*, pages 2245–2253. JMLR, 2018.

[11] D. Janzing, B. Steudel, N. Shajarisales, and B. Schölkopf. Justifying information-geometric causal inference. In *Measures of Complexity*, pages 253–265. Springer, 2015.

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[13] M. Kocaoglu, A. G. Dimakis, S. Vishwanath, and B. Hassibi. Entropic causal inference. In *AAAI*, pages 1156–1162, 2017.

[14] A. Kolmogorov. On tables of random numbers. *Indian J. Stat. Ser. A*, 25(4):369–376, 1963.

[15] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *JMLR*, 18:430–474, 2017.

[16] N. Lawrence. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *JMLR*, 6:1783–1816, 2005.

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

[18] M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications.* Springer Science & Business Media, 2009.

[19] J. C. Loehlin. *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis.* Psychology Press, 1998.

[20] A. Marx and J. Vreeken. Telling cause from effect using MDL-based local and global regression. In *ICDM*, pages 307–316. IEEE, 2017.

[21] J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2009.

[22] R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS One*, 5(2), 2010.

[23] R. Ranganath and A. Perotte. Multiple causal inference with latent confounding. *CoRR*, abs/1805.08273, 2018.

[24] R. Ranganath, L. Tang, L. Charlin, and D. M. Blei. Deep exponential families. In *AISTATS*, 2015.

[25] H. Reichenbach. *The direction of time.* Dover, 1956.

[26] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *CoRR*, abs/1505.05770, 2015.

[27] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(1):465–471, 1978.

[28] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66:688–701, 1974.

[29] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Comp Sci*, 2, 2016.

[30] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7:2003–2030, 2006.

[31] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *JMLR*, 12:1225–1248, 2011.

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

[33] P. Spirtes, C. Meek, and T. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias in computation, causation and discovery. In *UAI.* MIT Press, 1999.

[34] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. R. Statist. Soc. B*, 61:611–622, 1999.

[35] Y. Wang and D. M. Blei. The blessings of multiple causes. *CoRR*, abs/1805.06826, 2018.

[36] J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172:1873–1896, 2008.

[37] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *UAI*, pages 647–655, 2009.

A Appendix

A.1 Coding of the Tübingen Pairs Here we give a full list of which pairs of the Tübingen pairs dataset we considered to be mainly causal, confounded, or which we were uncertain about.

- Causal: 13–16, 25–37, 43–46, 48, 54, 64, 69, 71–73, 76–80, 84, 86–87, 93, 96–98, 100
- Confounded: 65–67, 74–75, 99
- Uncertain: 1–12, 17–24, 38–42, 47, 49–53, 55–63, 68, 70, 81–83, 85, 88–92, 94–95

For example for pairs 5–11 it was unclear to us to what extent the age of an abalone should be considered as a causal factor to its length, height, weight, or other measurements, and to what extent all of these should simply be confounded by the underlying biological processes of development.

As another example, for pair 99 we believed that it is reasonable to suggest that the correlation between language test score of a child and socio-economic status of its family might more plausibly be explained by the intelligence of parents and child — which are strongly correlated themselves.