Leveraging directed causal discovery to detect latent common causes

Ciarán M. Lee,[1, 2] Christopher Hart,[1] Jonathan G. Richens,[1] and Saurabh Johri[1]

[1]*Babylon Health, London, United Kingdom*
[2]*University College London, United Kingdom*

The discovery of causal relationships is a fundamental problem in science and medicine. In recent years, many elegant approaches to discovering causal relationships between two variables from uncontrolled data have been proposed. However, most of these deal only with purely directed causal relationships and cannot detect latent common causes. Here, we devise a general method which takes a purely directed causal discovery algorithm and modifies it so that it can also detect latent common causes. The identifiability of the modified algorithm depends on the identifiability of the original, as well as an assumption that the strength of noise be relatively small. We apply our method to two directed causal discovery algorithms, the Information Geometric Causal Inference of (Daniusis et al., 2010) and the Kernel Conditional Deviance for Causal Inference of (Mitrovic, Sejdinovic, & Teh, 2018), and extensively test on synthetic data—detecting latent common causes in additive, multiplicative and complex noise regimes—and on real data, where we are able to detect known common causes. In addition to detecting latent common causes, our experiments demonstrate that both modified algorithms preserve the performance of the original directed algorithm in distinguishing directed causal relations.

I. INTRODUCTION

Causal knowledge is crucial in medicine, as causal relations—unlike correlations—allow one to reason about the consequences of possible treatments [1–3]. Indeed, determining whether a particular treatment *causes* a reduction in the severity of a disease is an incredibly important problem in healthcare. However, learning from medical records that a treatment is correlated with recovery is not sufficient to conclude that treatment will cause a new patient to recover, due to the presence of latent confounding. More expensive treatment could be given to wealthier patients who are more likely to recover regardless of treatment, due to better lifestyle and diet. Here, the patients wealth acts as a latent confounder, or common cause, between receiving treatment and recovering. In reality, the treatment could be detrimental. Thus, learning the underlying causal structure is vital. This difficult problem is confounded further when no high-grade controlled trial evidence exists—-meaning the standard approach to learning causal structure isn't available. Hence, methods for discovering causal structure from uncontrolled, or observational, data are paramount to making correct treatment decisions.

In recent years, a number of elegant approaches to discovering causal relations between two variables have been proposed [4–13]. However, most of these deal only with purely directed causal relationships and cannot detect latent common causes. That is, given two variables A and B, these algorithms can only distinguish A causes B, graphically depicted in Figure 1(a), from B causes A, depicted in Figure 1(b). Those algorithms that can detect latent common causes either impose stern restrictions on the underlying causal model—such as enforcing linearity [5, 14] or demanding that noise be additive [8]. As the presence of latent common causes is a significant problem that confounds many statistical analyses, surmounting these shortcomings and developing general methods

for detecting latent common causes that do not rely on strong parametric assumptions, such as additive noise or linearity, is an important problem in causal discovery.

In this work, we devise a heuristic method which takes a purely directed causal discovery algorithm and modifies it so that it can also discover latent common causes. That is, our method takes in a causal discovery algorithm which can only distinguish between the causal structures in Figure 1(a) and Figure 1(b), and outputs a causal discovery algorithm which can distinguish between all three structures in Figure 1. The identifiability of the modified algorithm depends on the identifiability of the original. If the original directed algorithm requires a parametric assumption, that noise be additive for example, then so too does the modified algorithm. However, if the directed algorithm does not rely on any parametric assumptions for its identifiability, then neither does the modified algorithm resulting from applying our method. Additionally, identifiability of our method requires a constraint on the strength of latent noise.

We apply our method to two state-of-the-art directed causal discovery algorithms, the Information Geometric Causal Inference (IGCI) algorithm of [15] and the Kernel Conditional Deviance for Causal Inference (KCDC) algorithm of [7], and extensively test on synthetic data—detecting latent common causes in additive, multiplicative and complex noise regimes—as well as on real data, where we are able to detect known common causes. Importantly, and in addition to detecting latent common causes with high accuracy, our experiments demonstrate that both modified algorithms do not sacrifice the performance of the original directed algorithm in distinguishing directed causal relations.

The paper is organised as follows. In Section II we discuss previous and related work. In Section III we introduce and describe our method and discuss identifiability results. Section IV outlines our experiments and Section V contains a discussion of possible future work.

II. RELATED WORK

Methods for discovering the causal structure underlying a data generating process largely fall into two categories. The first, which we term *global* causal discovery, attempts to reconstruct a (partially) undirected version of the DAG. This approach is broadly split into two categories: constraint-based and score-based. The constraint-based approach employs conditional independence tests between the variables in question to determine which should share an edge in the causal structure. Examples include the PC algorithm [16], the IC algorithm [1], as well as algorithms which allow for latent variables [17] and selection bias [18]. The score-based approach introduces a scoring function, such as Minimum Description Length, that evaluates each network with respect to some training data, and searches for the best network according to this function. Hybrid approaches employing both constraint- and score-based techniques to outperform either alone [19].

The main drawback of constraint-based algorithms is that, as they can only recover Markov equivalence classes, they are not always able to orient edges between dependent variables. Given correlated variables A and B, these methods are unable to distinguish the structures depicted in Figure 1. The second category of causal discovery algorithm, which we term *local* or *bivariate* causal discovery, aims to address this by exploiting notions of asymmetry between cause and effect. These methods specify different assumptions that aim to make such asymmetries testable at an observational level.

The first type of assumption proposed in the literature, known as the functional causal model approach, specifies that each effect is a deterministic function of its cause together with some latent, independent noise term [20]. The first such algorithm, termed Linear Non-Gaussian Additive noise Model (LiNGAM) [5], assumes the functions are linear and the latent noise variables are non-Gaussian. Given these assumptions, this method can distinguish between the causal structures in Figure 1(a) and (b); a separate hidden variables extension [21] extends this result to all structures in Figure 1. The other prototypical example, the Additive Noise Model (ANM) [4], allows the effect to be an arbitrary function of the cause, but assumes the effect only depends on the latent noise term additively. The ANM algorithm can distinguish between the two structures in Figure 1(a)-(b) as long as the additivity assumption holds. ANM has been extended to allow the effect to depend on the cause and noise in a non-linear fashion [10]. Finally, the Confounding Additive Noise (CAN) algorithm [8] extends the ANM algorithm to deal with latent common causes, allowing all structures in Figure 1 to be distinguished. We review the CAN model in Section II A.

The second type of assumption stipulates that the cause $P(\text{cause})$ be independent of the mechanism $P(\text{effect}|\text{cause})$. An information geometric approach to measuring such independence has been proposed [6, 15],
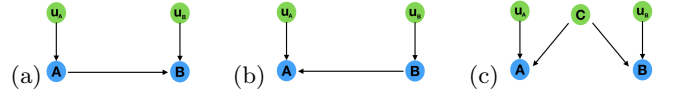

Figure 1: In all cases, blue nodes represent observed variables and green nodes denote latent noise terms. (a) A causes B. (c) B causes A. (d) Latent common cause between $A\&B$.

known as the Information Geometric Causal Inference (IGCI) algorithm. This method can only distinguish the two structures in Figure 1(a)-(b). IGCI does not require specific parametric functional relationships to hold.

The last type of assumption entails that the causal mechanism $P(\text{effect}|\text{cause})$ should be "simpler", given some quantifiable notion of "simple", than the acausal mechanism $P(\text{cause}|\text{effect})$. The Kernel Conditional Deviance Causal discovery (KCDC) algorithm [7] algorithm uses conditional kernel mean embeddings to establish a norm on conditional probabilities and uses the variance of these norms as a simplicity measure.

The IGCI and KCDC algorithms constitute the current state-of-the-art at distinguishing between the two causal structures in Figure 1(a)-(b).

While the ANM algorithm has been extended to deal with latent common causes via the CAN algorithm, its generalisations, such as KCDC and IGCI, have not. The current paper introduces a heuristic method that takes as input any purely directed causal discovery algorithm and outputs a new algorithm that can distinguish all causal structures in Figure 1.

A. Identifying latent common causes using additive noise models

As a prelude to our general method for detecting latent common causes, we review a current state-of-the-art algorithm for distinguishing all structures in Figure 1, the Confounding Additive Noise model (CAN) of [8]. In CAN, one assumes that the functions underlying the causal structure from Figure 1(c) are $A = f(C) + u_A$ and $B = g(C) + u_B$, for arbitrary f, g. The first step of CAN is to fit such a model using samples from A, B. To do this, manifold learning is employed to learn a representation of C from the observed samples of A, B. Next, functions are fitted to map the learned representation of C to A and B using e.g. Gaussian Process regression. The residuals resulting from this regression correspond to representations of the noise terms, u_A and u_B. In fitting this model, one must ensure that C, u_A, u_B are all mutually independent, as implied by the graphical structure of Figure 1(c). To ensure independence, CAN employs a computationally expensive optimisation procedure. See Section IV for further discussion.

With such a model fit using samples of A, B, the next step is to use it to distinguish all structures in Figure 1.

Note that the additive noise assumption for Figure 1(a)[1] means $B = g(A) + u_B$ and $A = u_A$, for arbitrary g. CAN distinguishes Figure 1(a) from Figure 1(c) if the variance of the learned u_A is small, relative to the variance of u_B, and prefers Figure 1(c) if the variances are approximately equal. The case of Figure 1(b) follows by switching A and B. The authors justify this as follows. Suppose A causes B, and, by a slight measurement error, we observe A', which differs from A by a small additive term. Here, we would expect the functions to be as described above for Figure 1(c). But we should not distinguish A from its measurement result A' if both variables almost coincide. In this case, after normalising A, B to unit variance, the variance of u_A would be small compared to u_B, as A is close to A'. In general, the CAN algorithm distinguishes between all structures in Figure 1 using the following decision criteria: i) $A \rightarrow B$ if $\mathrm{var}(u_A)/\mathrm{var}(u_B) \ll 1$, ii) $A \leftarrow B$ if $\mathrm{var}(u_A)/\mathrm{var}(u_B) \gg 1$, and iii) $A \leftarrow C \rightarrow B$ if $\mathrm{var}(u_A)/\mathrm{var}(u_B) \approx 1$. Determining reasonable practical thresholds for the above criteria is a tricky problem, with values being decided on a case by case basis; see [8] for more information.

III. METHODS

We work in the functional causal model[2] framework [20]. Here, a causal structure corresponds to a directed acyclic graph (DAG) between observed and latent variables, with each variable a deterministic function of its parents and some latent, independent noise term. A causal model corresponds to a DAG, the functions, and a specification of the priors over the latent noise terms.

The following two Lemmas and Definition play an important role in our method for distinguishing directed from latent common causes. We define a causal model with no directed arrows between observed variables *purely common cause*. We define two causal models *observationally equivalent* if they are characterised by the same set of distributions over observed variables.

Lemma 1. *Any causal model wherein there is a directed arrow between A & B ($A \rightarrow B$ or $B \rightarrow A$) is observationally equivalent to one that is purely common cause.*

This result has been proved before, in [13] for example, but to see that it holds for the case of Figure 1(a), consider the following. A model for this causal structure corresponds to specifying the functional dependencies between A, B and their parents: $A = f(u_A)$ and $B = g(A, u_B)$. Substituting $f(u_A)$ for A in B, one obtains $B = g(f(u_A), u_B)$. Defining $u'_A := f(u_A)$ results in the relations $A = u'_A$ and $B = g(u'_A, u_B)$. The directed
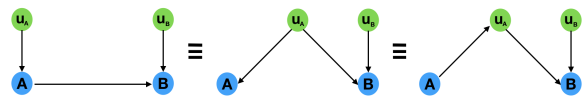

Figure 2: Depiction of three observationally equivalent DAGs discussed in Lemma 2. The symbol $\equiv$ denotes observational equivalence.

arrow from A to B has been replaced by the latent "common cause" u'_A, and Figure 1(a) has been reduced to a purely common cause model. This is outlined graphically in the first two causal structures from Figure 2, where the symbol $\equiv$ denotes observational equivalence.

We say a causal model is in *canonical form* when it is purely common cause. Note that Figure 1(c) is already in canonical form. The following Lemma provides another key observation underlying our approach.

Lemma 2. *Given the causal model in Figure 1(a)[3], the canonical model is consistent with two causal structures: one where u_A is a common cause between A and B, and the other where u_A is the mediator of the causal arrow from A to B. This is depicted in the last two structures of Figure 2.*

Proof. Without loss of generality, the canonical model of Figure 1(a) is given by $A = u_A$ and $B = g(u_A, u_B)$. These functions are consistent with the second causal structure in Figure 2, in which a causal arrow points from u_A to A. But the equality $A = u_A$ is also consistent with the third causal structure in Figure 2, where a causal arrow points from A to u_A. That is, the relationship $A = u_A$ is consistent with both an arrow from u_A to A and an arrow from A to u_A. Note that due to the independence of the noise terms u_A and u_B, this is not true of the other arrows in the causal structure. A causal arrow pointing from B to u_A in Figure 2 would induce correlations between u_A and u_B, in direct contradiction with the independence of noise terms. $\square$

Lemma 2 tells us that in the canonical model of Figure 1(a), the directionality of the causal link between A and u_A is underdetermined, as illustrated in last two structures of Figure 2. All other causal links are fully determined and directed however. If, on the other hand, the original structure had been Figure 1(b), then the causal link between the canonical "common cause", u_B, and B would be underdetermined, with all remaining links determined and directed. Note that if the structure had originally been that of Figure 1(c), then neither arrow from the true common cause, C, to the observed variables is underdetermined.

Here, we focus on directed causal discovery algorithms, D, that decide the causal direction using some quantifiable notion of asymmetry, such as IGCI [15] and KCDC

---

[1] Figure 1(b) follows by interchanging A and B.
[2] Also known as the structural equation model framework.

[3] The case of Figure 1(b) follows by interchanging A and B.

[7], both described in Section II. In general, such algorithms assign a real scalar to each causal direction $v_{A \to B}^D$ and $v_{B \to A}^D$, returning the causal direction with smallest v value. Due to the pervasiveness of statistical noise, such algorithms have a decision threshold δ, such that if $|v_{A \to B}^D - v_{B \to A}^D| < \delta$ the algorithm fails to detect a causal direction; see [7] for further details.

For instance, the KCDC algorithm of [7] uses conditional kernel mean embeddings to establish a norm on conditional probabilities $P(B|A = a)$, $P(A|B = b)$ for each value of $A = a$, $B = b$. If the variance $v_{A \to B}$ of the norms $\{\|P(B|A = a)\|\}_a$ is smaller than the variance $v_{B \to A}$ of $\{\|P(A|B = b)\|\}_b$, the KCDC algorithm returns Figure 1(a), otherwise it returns Figure 1(b).

The following definition formalises the notion of a causal arrow being underdetermined relative to a particular directed causal discovery algorithm.

Definition 1. *A causal link between two variables A and B is* underdetermined *relative to causal discovery algorithm D if $|v_{A \to B}^D - v_{B \to A}^D| < \delta$ for pre-specified decision threshold δ.*

A. Detecting latent common causes

Lemmas 1 & 2 and Definition 1 suggest a method for distinguishing the three causal structures in Figure 1. To employ the method however, the canonical common cause needs to be identified from samples of the obersved variables A, B.

Lemma 3. *Under the assumption that the influence of the latent noise term is small, the canonical common cause can be identified from samples of A, B.*

Proof. Consider the causal model of Figure 1(a)[4], which, following Lemma 1 we can write as $A = u_A$ and $B = g(u_A, u_B)$. Now, a Taylor expansion yields $B = g(u_A, 0) + g'(u_A, 0)u_B + $ (higher order terms in u_B), where g' is the partial derivative of g with respect to u_B. The assumption that the influence of the noise term u_B on B is small corresponds to the fact that we can drop higher order terms in u_B in the above Taylor expansion. Hence, we can write $B = g(u_A, 0) + g'(u_A, 0)u_B$. Without loss of generality, we can assume the noise term is a Gaussian random variable with mean 0. This implies that the expected value of B is $B = g(u_A, 0)$. Thus the expected value of (A, B) samples is $(u_A, g(u_A, 0))$. That is, they are both functions of the canonical common cause u_A alone and can hence be identified from A, B samples. $\square$

The assumption underlying Lemma 3 intuitively corresponds to the requirement that fluctuations around the manifold parameterised by the canonical common cause due to latent noises are small.

---

[4] All other cases follow by a similar argument.

Lemma 3 implies that given the canonical causal model associated with any of the structures from Figure 1, manifold learning can be utilised—as is done in the CAN algorithm [6, Section 3], described in Section II A of the current paper—to determine the approximate parameterisation (up to rescaling) of the canonical "common cause" from samples of the observed A, B.

Under the hypothesis that such an algorithm will learn the most likely latent representation of the data, we determine the most likely causal explanation as follows. In the case of Figure 1(a) and Figure 1(b), the "common cause" corresponds to the variables u_A and u_B respectively. In the case of Figure 1(c) it is the actual common cause, C. Second, given the parameterisation of the "common cause", one can determine the original precanonical causal structure using a purely directed causal discovery algorithm to determine which causal link to the observed variables is underdetermined (if any). If the arrow to A is underdetermined then the causal structure is Figure 1(a); if the arrow to B is underdetermined the structure is Figure 1(b); if no arrow is underdetermined, the structure is Figure 1(c). The causal discovery algorithm described above is summarised in Algorithm 1.

Algorithm 1

Input: A, B samples, manifold learning algorithm M, directed causal discovery algorithm D.

Output: Single causal structure from Figure 1.

1: Run M on $\{A, B\}$ to obtain parameterisation (up to rescaling, etc.) of canonical common cause T that best fits the data.
2: Implement D between $T\&A$ and $T\&B$
3: **if** D outputs $T \to B$ & the causal link between $T\&A$ is underdetermined relative to D **do:**
4:      Output DAG from Figure 1(a)
5: **else if** D outputs $T \to A$ & the causal link between $T\&B$ is underdetermined relative to D **do:**
6:      Output DAG from Figure 1(b)
7: **else if** D outputs $T \to B$ & $T \to A$ **do:**
8:      Conclude T is a common cause of A, B and output DAG from Figure 1(c)
9: **return** DAG output from above

B. Decision criteria

We now introduce a heuristic decision criterion for checking the conditions of Algorithm 1. Consider step 3 of Algorithm 1, where the learned canonical common cause is denoted T. To simultaneously check whether the causal link between A and T is underdetermined, and also whether $T \to B$, are consistent with the data, one could apply the following heuristic. Check if

$$\Delta = \frac{\left| |v_{A \to T} - v_{T \to A}| - |v_{B \to T} - v_{T \to B}| \right|}{\max \left(|v_{A \to T} - v_{T \to A}|, |v_{B \to T} - v_{T \to B}| \right)}$$

lies in the region $(\alpha, 1]$ for some pre-specified decision threshold α close to 1. This follows as the conditions in step 3 imply $|v_{A \to T} - v_{T \to A}| < \delta$ and $|v_{B \to T} - v_{T \to B}| > \delta$, for small δ. That is, the first term in the numerator of Δ is generally close to zero, while the second is not.

Alternatively, if $\Delta \in (0, \alpha]$, then the data is consistent with $T \to B$ & $T \to A$, as in this case $|v_{A \to T} - v_{T \to A}| > \delta$ and $|v_{B \to T} - v_{T \to B}| > \delta$. Hence, both terms in the numerator can be roughly thought of as being closer in magnitude than in the case of steps 3 and 5. This allows us to check whether step 7 of Algorithm1 is true.

Note that the first criterion $\Delta \in (\alpha, 1]$ is also consistent with step 5 of Algorithm 1, as steps 3 and 5 are the same under interchange of A & B. Hence if $\Delta \in (\alpha, 1]$, then either step 3 or step 5 is true, and step 7 is false, as per the above. As the two cases in in step 3 and 5 are purely directed causal structures, they can be distinguished using the original directed causal discovery algorithm.

To assign confidence to the outcome of the above heuristic, one could take a bootstrapped approach and calculate the mean of the $\{\Delta_i\}$'s output by running the algorithm on subsamples of the input data, where Δ_i is the value computed on the ith subsample. We suppress the index i in the below for notational ease. As reasoned above, mean$(\{\Delta\}) \in (\alpha, 1]$ is consistent with a directed causal structure and mean$(\{\Delta\}) \in (0, \alpha]$ with a common cause. Additionally, the variance of the $\{\Delta\}$'s calculated in this manner encodes information of the correct DAG. For instance, small variance var$(\{\Delta\}) \in (0, \gamma]$—where γ is again a pre-specified decision threshold less than 1— is consistent with a directed causal structure and large variance var$(\{\Delta\}) \in (\gamma, 1]$ is consistent with a common cause. This follows because for a directed causal structure only one of the terms in the numerator of each Δ_i must lie in the region $[0, \delta)$, for small δ, and hence cannot vary much. For a common cause, both terms are outside this region and thus can vary more.

These two implication, high mean$(\{\Delta\})$ being indicative of a directed cause and high var$(\{\Delta\})$ of a common cause, can be used in conjunction to determine the causal structure. For example, it could be the case that a particular common cause structure has moderately high mean$(\{\Delta\})$, but such a high var$(\{\Delta\})$ that a directed cause is ruled out, and a common cause is returned by the algorithm. Similar reasoning can be applied to the case of a directed cause whose mean$(\{\Delta\})$ is lower than expected, but has such a low var$(\{\Delta\})$ that a common cause is ruled out.

To summarise, the above decision criteria are as follows. The mean and variance of $\{\Delta\}$'s output by running the algorithm on subsamples of the input data are computed. Recall the possible range of mean$(\{\Delta\})$ is $[0, 1]$. This range is split into a number of regions. We fix region 1 to be $[0, \alpha_1]$, region 2 to be $[\alpha_1, \alpha_2]$, and so on. Each region R is associated with a threshold value for var$(\{\Delta\})$, γ_R. If mean$(\{\Delta\})$ lies in region R, then if var$(\{\Delta\}) \leq \gamma_R$ the algorithm outputs a directed causal structure, whose direction is then determined by

the original directed causal discovery algorithm. But if var$(\{\Delta\}) > \gamma_R$, a common cause structure is output. Note that as the regions get closer to 0, that is, to $[0, \alpha_1]$, the variance threshold γ_R gets increasing small. Hence, $\gamma_R < \gamma_{R+1}$. This encodes our previous reasoning of how mean and variance imply different causal structures. Additionally, there should be a failure mode for very low variance and mean, and very high mean and variance, as these are not indicative of a direct nor a common cause.

As with the decision criteria for the CAN algorithm, described in Section II A, and KCDC [7] setting such thresholds for causal discovery presents a challenge due to the varying impact of noise within experimental data sets. An appropriate method we take in this work is to use data of similar provenance with known causal relations and determine appropriate thresholds for that domain, which are then fixed.

C. Identifiability

Lemmas 1, 2, and 3 establish the identifiability of our method. Given the assumption that the strength of latent noise is relatively small, a representation of the canonical common cause can be extracted from samples of the observed A, B. The discussion in Section III A and Lemmas 1 & 2 show how this representation can be used to distinguish all three causal structures in Figure 1.

The sensitivity of our algorithm to the assumption that the strength of latent noise be relatively small is explored empirically in Section IV C.

Additionally, as our algorithm depends on the original directed causal discovery algorithm, their identifiability is closely related. It may sometimes be difficult to demarcate underdetermination from the unidentifiability of the original algorithm. A directed algorithm could conclude that a causal link is underdetermined when it is not, due to the fact that the model underlying said link makes it unidentifiable. However, this does not seem to happen often in practice, as we see when conducting extensive experiments in many different noise and functional regimes in Section IV.

IV. EXPERIMENTAL RESULTS

We test our method by showing it can turn both the KCDC algorithm of [7] and the IGCI algorithm of [15], which can each only distinguish Figure 1(a) and (b), into algorithms that can distinguish all causal structures in Figure 1. We refer to the modified KCDC and IGCI algorithms output by our method as modKCDC and modIGCI respectively. To benchmark performance of modKCDC and modIGCI, we compare against CAN [8] the only previous causal discovery algorithm explicitly designed to distinguish all structures in Figure 1, which we outlined in Section II A. The CAN algorithm proved to be very computational intensive to run compared to

modKCDC and modIGCI. To reduce computational cost, we only enforced two of the three independence constraints discussed in Section II A: $u_A, u_B \perp C$.

For KCDC, we make use of a radial basis function kernel with hyperparameters fixed in all experiments; there is scope to extend this to multiple kernel majority vote following [7], though performance with a single basic kernel proved effective here. We use the implementation of CAN outlined in [8], using Gaussian Process regression and HSIC to test independence, and employing up to five iterations of the independence optimisation loops, capped at 5000 iterations of the solver each time. For the manifold learning subroutine of modKCDC, modIGCI, and CAN, we employ Isomap [22]. This is the same implementation of manifold learning as originally used in the CAN algorithm [8].

We followed the bootsrapping procedure outlined in Section III B, using 25 bootstraps each randomly sampling 95% of the data. To set the thresholds, we generated synthetic data from directed causal models with additive, multiplicative, and complex noise functions—similar to the models we test on in this section—and determined the threshold values which approximately reproduced the performance of the original directed causal discovery algorithms on these different models. We did this for modKCDC and modIGCI, as well as for CAN, whose decision criteria were discussed in Section II A. This is similar to the way the CAN thresholds were set in [6], and worked well in our experiments despite the fact that they were quite diverse in nature. We report the thresholds set in this manner in the Appendix. These thresholds are used in all experiments.

A. Synthetic data

We first test on synthetic data for both purely directed and purely common causes. The functions used to generate this synthetic data were adapted from the work of [7], who used them to test performance of KCDC and compare its performance to IGCI. As such, we believe using similar synthetic functions and the same range of synthetic noise distributions provides a fair and robust benchmark of KCDC, IGCI, and their modifications.

On the directed synthetic datasets, we compare the modified algorithms to the original to determine whether the modification noticeably reduced their performance in detecting directed causal structure. Our experiments show that modKCDC and modIGCI preserve the high level of performance of the original KCDC and IGCI algorithms on directed data.

1. Directed causal structures

In the below experiments, we sample 100 datasets each of 250 observations and test modKCDC and modIGCI. The CAN algorithm proved to be very computational

intensive to run, taking 300-500 times as long to run on a dataset of size 250 compared to modKCDC and modIGCI. Hence we only tested it on 10 datasets, each of 250 observations. For our three algorithms, we sample $A \in \mathcal{N}(0, 1)$, and test across three different noise regimes: Normal $n_B \in \mathcal{N}(0, 1)$, Uniform $n_B \in \mathcal{U}(0, 1)$, and Exponential $n_B \in \mathcal{E}(1)$. We record the accuracy over all datasets, that is, the percentage of cases where the algorithm output the correct ground truth directed structure. This is the same metric employed by [7] in their experiments. The CAN algorithm was only able to fit a model that satisfied the independence criteria discussed in Section II A and IV in the case of additive noise, as used in experiments (1) and (2) below. This is to be expected as non-additive noise, as in experiments (3)–(6), violates the identifiability requirements of CAN. In these cases we could not report CAN's performance. All results are presented in Table 1.

Additive noise:

$$(1) \; B = \sin(10A) + e^{3A} + n_B$$
$$(2) \; B = A \exp A^2 + n_B$$

Multiplicative noise:

$$(3) \; B = \big(\sin(10A) + e^{3A} \big) e^{n_B}$$
$$(4) \; B = \big(A^2 + A^5 \big) e^{n_B}$$

Complex noise:

$$(5) \; B = A^5 - \sin(A^2 |n_B|)$$
$$(6) \; B = \log(A + 10) + A^{8n_B}$$

In all synthetic direct cause experiments, KCDC and IGCI perform extremely well, achieving 100% accuracy in all situations with the exception of IGCI in experiment (6) with normally distributed noise. In all cases bar one, modKCDC and modIGCI either match this performance, or come within 1-2% of it. It is also interesting to note that the one case in which IGCI performed badly, so did modIGCI. modIGCI failed to match the performance of IGCI in one case however, that of experiment (6) with uniform noise. As IGCI itself performed badly on this experiment with normal noise, it seems to present a challenging dataset for this algorithm, so it is perhaps not overly surprising that modIGCI yielded poor performance in this case.

The CAN algorithm did not perform well on any of these synthetic datasets, failing to fit a model that complied with the independence criteria in all cases beyond additive noise. We only reported accuracy for CAN in cases where it successfully fit a model. This in principle gives CAN an added advantage. In (1), this happened in 5/10, 9/10, 7/10 cases for the three different noise distributions. In (2), this happened in 7/10, 3/10, 5/10 cases. CAN's poor performance shows the difficulty in developing an algorithm that can distinguish purely directed and common causes.

Table 1				
Directed Cause				
Exp.	Algorithm	Normal	Uniform	Expon.
1	modKCDC	99%	100%	98%
	KCDC	100%	100%	100%
	modIGCI	100%	99%	100%
	IGCI	100%	100%	100%
	CAN	40%	33%	57%
2	modKCDC	99%	100%	98%
	KCDC	100%	100%	100%
	modIGCI	100%	100%	100%
	IGCI	100%	100%	100%
	CAN	43%	66%	50%
3	modKCDC	100%	100%	100%
	KCDC	100%	100%	100%
	modIGCI	100%	99%	98%
	IGCI	100%	100%	100%
4	modKCDC	100%	99%	100%
	KCDC	100%	100%	100%
	modIGCI	99%	100%	96%
	IGCI	100%	100%	100%
5	modKCDC	98%	98%	98%
	KCDC	100%	100%	100%
	modIGCI	99%	98%	99%
	IGCI	100%	100%	100%
6	modKCDC	100%	97%	99%
	KCDC	100%	100%	100%
	modIGCI	29%	57%	99%
	IGCI	31%	100%	100%

2. Common cause

The common cause setup closely follows that of Section IV A 1. For these synthetic experiments, the common cause T was sampled from $T \in \mathcal{N}(0,1)$. Here, CAN was only able to fit a model for experiment (2), below. In all other cases we could not report CAN's performance. Results are presented in Table 2.

Additive noise:

$$(1)\ A = \sin(10T) + e^{3T} + n_A$$
$$B = \log(T+10) + T^6 + n_B$$
$$(2)\ A = \log(T+10) + T^6 + n_A$$
$$B = T^2 + T^6 + n_B.$$

Multiplicative noise:

$$(3)\ A = \big(\sin(10T) + e^{3T}\big)e^{n_A}$$
$$B = \big(T^2 + T^6\big)e^{n_B}$$
$$(4)\ A = \big(\sin(10T) + e^{3T}\big)e^{n_A}$$
$$B = \big(\log(T+10) + T^6\big)e^{n_y}$$

Additive and Mulitplicative noise:

$$(5)\ A = \log(T+10) + T^6 + n_A$$
$$B = \big(T^2 + T^6\big)e^{n_B}$$
$$(6)\ A = \sin(10T) + e^{3T} + n_A$$
$$B = \big(T^2 + T^6\big)e^{n_B}$$

Table 2				
Common Cause				
Exp.	Algorithm	Normal	Uniform	Expon.
1	modKCDC	96%	95%	97%
	modIGCI	99%	96%	99%
2	modKCDC	98%	95%	96%
	modIGCI	100%	100%	100%
	CAN	80%	66%	100%
3	modKCDC	94%	99%	95%
	modIGCI	98%	96%	97%
4	modKCDC	95%	96%	96%
	modIGCI	96%	96%	97%
5	modKCDC	97%	100%	95%
	modIGCI	95%	100%	94%
6	modKCDC	96%	95%	96%
	modIGCI	94%	96%	93%

Here, modKCDC and modIGCI achieve impressive accuracy across a range of different synthetic noise functions—including additive and multiplicative noise—and diverse distributions over these noise terms. Despite the fact that neither KCDC nor IGCI could detect latent common causes, our method turned them into algorithms that could detect them. CAN performed well on the only experiment where it managed to fit a model, (2). Again, accuracy was only reported when it fit a model, in principle giving CAN an added advantage. In (2), this happened in 5/10, 3/10, 1/10 cases for the different noise distributions. CAN failed to fit a model in (1), (3)-(6) above, again demonstrating the difficulty in distinguishing common and directed causes.

B. Common cause robustness tests

In the above experiments we tested the robustness of our algorithms in different noise regimes. We now test robustness to more complex functional relationships. First, we test the accuracy of the algorithms in a regime beyond additive and multiplicative noise. We then test on functions drawn from Gaussian Processes. Results are presented in Table 3.

1. Complex noise

(1) $A = T^5 - \sin{(T^2 n_A)}$,   (2) $A = T \sin{(10T|n_A|)}$,

       $B = \left(\log(T^4 + 10)\right)^{2n_B}$ $B = \log(T + 10) + T^{2|n_B|}$

2. Gaussian Process generators

(3) Let $A = f(T)e^{n_A}$ & $B = g(T)e^{n_B}$, with f, g drawn from the same Gaussian Process whose kernel is a sum of polynomial (with only $T^2 + T^6$ terms) and periodic exponential.

(4) f is drawn from a GP with polynomial kernel (only $T^2 + T^6$ terms) and g is drawn from GP with a sum of polynomial (only $T^3 + T^5$ terms) and a periodic exponential kernel.

Table 3

Complex Noise Common Cause		
Experiment	Algorithm	Result
1	modKCDC	93%
	modIGCI	95%
2	modKCDC	89%
	modIGCI	91%
3	modKCDC	90%
	modIGCI	95%
4	modKCDC	91%
	modIGCI	91%

Both modKCDC and modIGCI perform well in these datasets beyond additive and multiplicative noise. Experiments (3) and (4) provide an analysis of robustness to changes in the parameters of the underlying functions.

C. Sensitivity to the assumption of Lemma 3

As stated in Lemma 3, the identifiability of our method is based on the assumption that the influence of latent noise is small. The sensitivity of our method to this assumption will now be empirically tested using modIGCI. We synthetically generate both directed and common cause data from models which violate the assumption. Directed data is generated by sampling $A \in \mathcal{N}(0,1)$ and $u_A \in \mathcal{N}(0,1)$ and inputting to

$$B = \sin{(10A)} + e^{3A} + \lambda e^{e^{n_B}}$$

Common cause data is generated by sampling $T \in \mathcal{N}(0,1)$, $u_A \in \mathcal{N}(0,1)$ & $u_B \in \mathcal{N}(0,1)$ and inserting to

$$A = \sin(3t) + \lambda e^{e^{n_A}}$$
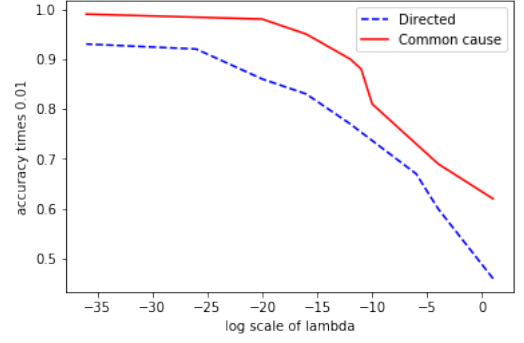$$B = \log(T + 10) + \lambda e^{e^{n_B}}$$

Figure 3: Effect of increasing λ on algorithm accuracy.

Here, the latent noise terms influence the observed variables via a double exponential function, which violates the assumption underlying Lemma 3. In both cases, the smaller λ is, the smaller the influence of the noise terms. We would expect that the smaller λ, the higher the accuracy of our method, and the larger λ, the lower the accuracy of our method. This intuition is borne out in the experiments. Results are plotted in Figure 3.

Despite the fact that the noise influences the observed variables via a double exponential function, the accuracy of modIGCI is quite resilient. For instance, when $\lambda = 1$ the accuracy on directed data is 46% and the accuracy on common cause data is 62%. Full results are in Figure 3.

D. Real data

We now test on real world data. We first consider *NLSchools data*, the 99th pair in the Tübingen cause-effect data[5]. Here 'socioeconomic status' is expected to be a direct cause of 'language score'. However, some researchers have argued that the correlation between these two attributes could be genetic in nature, implying the underlying structure is purely common cause. When applied on this data, modKCDC returns mean($\{\Delta\}$) = 0.933 and var($\{\Delta\}$) = 0.007, with modIGCI returning mean($\{\Delta\}$) = 0.022 and var($\{\Delta\}$) = 5.4×10^{-7}. modKCDC's output—with its high mean and low variance—is clearly indicative of a directed causal structure. While modIGCI has a slightly low mean($\{\Delta\}$), its var($\{\Delta\}$) value is extremely low. Following the discussion at the end of Section III B, we should conclude that modIGCI returns a directed causal structure in this instance. Hence, both modKCDC and modIGCI return a direct cause from 'socioeconomic status' to 'language score', ruling out a purely common cause explanation of this data.

Next, we consider *Breast Cancer Wisconsin (Diagnostic) Data* [23]. Here the diagnosis (malignant or be-

---

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

nign) is the common cause of two attributes, 'perimeter' and 'compactness', as they are conditionally independent given diagnosis. When run on this 'perimeter' and 'compactness' data, modKCDC returns mean($\{\Delta\}$) = 0.491 and var($\{\Delta\}$) = 0.081, and modIGCI returns mean($\{\Delta\}$) = 0.308 and var($\{\Delta\}$) = 0.062. In both cases, mean($\{\Delta\}$) is relatively small, with var($\{\Delta\}$) in both cases relatively large. From the discussion in Section III B, we should conclude that both algorithms are consistent with a common cause. In addition, all values are within the thresholds fixed in Section IV, which are reported in the Appendix, for a common cause. We hence conclude that both algorithms return a common cause between 'perimeter' and 'compactness', agreeing with the ground truth.

Next we consider the *AutoMPG* data set from [23]. Here, attributes 'acceleration' and 'MPG' (miles per gallon) are correlated but have a common cause, the model year. This is checked by performing a conditional independence test between 'acceleration' and 'MPG' given model year. When applied on this data, modKCDC returns mean($\{\Delta\}$) = 0.639 and var($\{\Delta\}$) = 0.086, with modIGCI yielding mean($\{\Delta\}$) = 0.124 and var($\{\Delta\}$) = 0.00062. As per the thresholds reported in the Appendix, we conclude that modKCDC returns a common cause between 'acceleration' and 'MPG'. While modIGCI returned a low mean($\{\Delta\}$) value—which is indicative of a common cause—its value for var($\{\Delta\}$) is also low, which

is not in general consistent with one. Given this relatively large size difference, following the discussion at the end of Section III B one could conclude that modIGCI's output is consistent with a common cause between 'acceleration' and 'MPG'.

V. CONCLUSION

We devised a method to turn a purely directed causal discovery algorithm into one that can also detect latent common causes. In our experiments we took the KCDC [7] and IGCI [15] algorithms, which each could distinguish directed causes but could not detect latent common causes, and showed our method enabled detection of latent common causes while preserving accuracy in distinguishing directed causes.

The setting of the thresholds for the decision criterion from Section III B is a challenging problem. The use of domain heuristic employed here follows both the approach of CAN [8], described in Section II A, and KCDC [7], outlined just before Definition 1. Despite its heuristic nature, it worked quite well in our experiments, despite how varied they were. Ongoing research will investigate the use of statistical tests to analyse the thresholds.

In future work, applications to the burgeoning field of quantum causal models [24–28] will be explored.

[1] J. Pearl, *Causality (2nd edition).* Cambridge University Press, 2009.

[2] J. G. Richens, C. M. Lee, and S. Johri, "Counterfactual diagnosis," *arXiv preprint arXiv:1910.06772*, 2019.

[3] Y. Perov, L. Graham, K. Gourgoulias, J. G. Richens, C. M. Lee, A. Baker, and S. Johri, "Multiverse: Causal reasoning using importance sampling in probabilistic programming," *arXiv preprint arXiv:1910.08091*, 2019.

[4] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Advances in neural information processing systems*, pp. 689–696, 2009.

[5] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, no. Oct, pp. 2003–2030, 2006.

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

[7] J. Mitrovic, D. Sejdinovic, and Y. W. Teh, "Causal inference via kernel deviance measures," in *Advances in Neural Information Processing Systems*, pp. 6986–6994, 2018.

[8] D. Janzing, J. Peters, J. Mooij, and B. Schölkopf, "Identifying confounders using additive noise models," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 249–257, AUAI Press, 2009.

[9] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag, "Learning functional causal models with generative neural networks," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 39–80, Springer, 2018.

[10] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the 25 conference on uncertainty in artificial intelligence*, pp. 647–655, 2009.

[11] J. A. Fonollosa, "Conditional distribution variability measures for causality detection," *arXiv preprint arXiv:1601.06680*, 2016.

[12] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin, "Towards a learning theory of cause-effect inference," *ICML, 1452–1461*, 2015.

[13] C. M. Lee and R. W. Spekkens, "Causal inference via algebraic geometry: feasibility tests for functional causal structures with two binary observed variables," *Journal of Causal Inference*, vol. 5, no. 2, 2017.

[14] D. Kaltenpoth and J. Vreeken, "We are not your real parents: Telling causal from confounded using mdl," in *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 199–207, SIAM, 2019.

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

[16] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search.* Adaptive computation and machine learning, The MIT press, 2nd edition, 2000.

[17] R. Silva, R. Scheine, C. Glymour, and P. Spirtes, "Learning the structure of linear latent variable models," *Journal of Machine Learning Research*, vol. 7, no. Feb, pp. 191–246, 2006.

[18] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 499–506, Morgan Kaufmann Publishers Inc., 1995.

[19] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.

[20] J. Peters, D. Janzing, and B. Schölkopf, "Elements of causal inference-foundations and learning algorithms," 2017.

[21] 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. Reasoning*, vol. 49, pp. 362–378, Oct. 2008.

[22] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[23] M. Lichman, "Uci machine learning repository," *URL http://archive.ics.uci.edu/ml.*, 2013.

[24] J.-M. A. Allen, J. Barrett, D. C. Horsman, C. M. Lee, and R. W. Spekkens, "Quantum common causes and quantum causal models," *Physical Review X*, vol. 7, no. 3, p. 031021, 2017.

[25] C. M. Lee and M. J. Hoban, "Towards device-independent information processing on general quantum networks," *Physical Review Letters*, vol. 120, no. 2, p. 020504, 2018.

[26] C. M. Lee, "Device-independent certification of nonclassical measurements via causal models," *arXiv preprint arXiv:1806.10895*, 2018.

[27] E. Wolfe, R. W. Spekkens, and T. Fritz, "The inflation technique for causal inference with latent variables," *arXiv preprint arXiv:1609.00672*, 2016.

[28] R. Chaves, C. Majenz, and D. Gross, "Information–theoretic implications of quantum causal structures," *Nature communications*, vol. 6, p. 5766, 2015.

Appendix A: Decision thresholds for experimental section of main paper

In Section 4 of the main paper, we use synthetic directed data to set the thresholds for the decision criteria—discussed in Section 3.2 of the main paper—of our modifications of KCDC [7] and IGCI [15], termed modKCDC and modIGCI respectively. To set the thresholds, we generated synthetic data from directed causal models with additive, multiplicative, and complex noise functions—similar to the ones used in the Experiments section—and determined the threshold values which approximately reproduced the performance of the original directed causal discovery algorithms on these three different models. These thresholds are then used in all experiments. While this is somewhat ad-hoc, it is similar to the way CAN's thresholds were set in [6], and worked well in our experiments despite the fact that they were quite diverse in nature. The thresholds were set following the bootstrapping approach detailed in Section 3.2, using 25 bootstraps each randomly sampling 95% of the data.

The thresholds we set for modKCDC are as follows:

1. If $\text{mean}(\{\Delta\}) > 0.9$ and $\text{var}(\{\Delta\}) \leq 0.06$, then we output a directed causal structure oriented using the original causal discovery algorithm.

    - The regime $\text{mean}(\{\Delta\}) > 0.9$ and $\text{var}(\{\Delta\}) > 0.06$ is a failure mode of our algorithm.

2. If $\text{mean}(\{\Delta\}) < 0.25$ and $\text{var}(\{\Delta\}) > 0.03$, a common cause structure is output.

    - The regime $\text{mean}(\{\Delta\}) < 0.25$ and $\text{var}(\{\Delta\}) < 0.03$ is a failure mode of our algorithm.

3. If $0.65 < \text{mean}(\{\Delta\}) \leq 0.9$ and $\text{var}(\{\Delta\}) \leq 0.06$, then a directed causal structure is output.

4. But if $0.65 < \text{mean}(\{\Delta\}) \leq 0.9$ and $\text{var}(\{\Delta\}) > 0.06$, then a common cause structure is output.

5. If $0.25 \leq \text{mean}(\{\Delta\}) \leq 0.65$ and $\text{var}(\{\Delta\}) \leq 0.03$, then a directed causal structure is output.

6. But if $0.25 \leq \text{mean}(\{\Delta\}) \leq 0.65$ and $\text{var}(\{\Delta\}) > 0.03$, then a common cause structure is output.

The thresholds we set for modIGCI are as follows:

1. If $\text{mean}(\{\Delta\}) > 0.9$ and $\text{var}(\{\Delta\}) \leq 0.02$, then we output a directed causal structure oriented using the original causal discovery algorithm.

    - The regime $\text{mean}(\{\Delta\}) > 0.9$ and $\text{var}(\{\Delta\}) > 0.02$ is a failure mode of our algorithm.

2. If $\text{mean}(\{\Delta\}) < 0.25$ and $\text{var}(\{\Delta\}) > 0.01$, a common cause structure is output

    - The regime $\text{mean}(\{\Delta\}) < 0.25$ and $\text{var}(\{\Delta\}) \leq 0.01$ is a failure mode of our algorithm.

3. If $0.45 < \text{mean}(\{\Delta\}) \leq 0.9$ and $\text{var}(\{\Delta\}) \leq 0.02$, then a directed causal structure is output.

4. But if $0.45 < \text{mean}(\{\Delta\}) \leq 0.9$ and $\text{var}(\{\Delta\}) > 0.02$, then a common cause structure is output.

5. If $0.25 \leq \text{mean}(\{\Delta\}) \leq 0.45$ and $\text{var}(\{\Delta\}) \leq 0.01$, then a directed causal structure is output.

6. But if $0.25 \leq \text{mean}(\{\Delta\}) \leq 0.45$ and $\text{var}(\{\Delta\}) > 0.01$, then a common cause structure is output.

The thresholds set for CAN, described in Section 2.1 of the main paper, are as follows. Let $r = \text{var}(u_A)/\text{var}(u_B)$. If $r \leq 0.65$ then a $A \rightarrow B$ is returned. If $r \geq 1.65$, then $A \leftarrow B$ is returned. If $0.65 < r < 1.65$, then a common cause structure is returned.