Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition

David Kaltenpoth[1] **Jilles Vreeken**[1]

[1]CISPA Helmholtz Center for Information Security, Saarbrücken

Abstract

Causal sufficiency is a cornerstone assumption in causal discovery. It is, however, both unlikely to hold in practice as well as unverifiable. When it does not hold, existing methods struggle to return meaningful results. In this paper, we show how to discover the causal network over both observed *and* unobserved variables. Moreover, we show that the causal model is identifiable in the sparse linear Gaussian case. More generally, we extend the algorithmic Markov condition to include latent confounders. We propose a consistent score based on the Minimum Description Length principle to discover the full causal network, including latent confounders. Based on this score, we develop an effective algorithm that finds those sets of nodes for which the addition of a confounding factor Z is most beneficial, then fits a new causal network over both observed as well as inferred latent variables.

1 INTRODUCTION

Discovering causal relationships from observational data is one of the most important open problems in science [Pearl, 2009]. Causal discovery methods aim to identify causal networks from data by reporting only edges that cannot be explained away by any other variables. Most typical causal discovery methods have in common that unless *all* relevant variables have been measured, they return models that include (many) spurious edges and thus lack a causal interpretation. The most commonly used approach is to wish the problem away by assuming *causal sufficiency*, i.e., that all common causes of all observed variables are observed. However, doing this does not make the issue disappear in practice; in many applications, including epidemiology [Kesteloot et al., 2006], economics [Angrist and Pischke, 2009], and bio-medicine Imbens and Rubin [2015], we do not know all the relevant variables, nor would we be able to measure them even if we knew.

In this paper, we present an approach to causal discovery that does not require causal sufficiency over the observed variables. We give conditions under which it is possible to identify joint confounders, and show how to infer these using factor analysis, as well as how to construct a causal network without spurious edges over both the observed X and the discovered confounders Z.

This may seem impossible. After all, given only a sample from $P(X)$, there exist infinitely many joint distributions $P(X, Z)$ consistent with the marginal $P(X)$, and picking out the $P(X, Z)$ that corresponds to the true causal mechanism sounds far-fetched. It is not. First, we can exploit the fact that a causal graph discovered over the observed variables X will contain many spurious edges $X_i \to X_j$ when both are affected by the hidden latent variable Z [Elidan et al., 2000]. That is, by focusing on those subsets of X that are densely connected in the graph, we can determine which variables are likely to share a hidden confounder.

While not every densely connected set of variables necessarily shares a hidden confounder, we can use the *algorithmic Markov condition* (AMC) to find the simplest causal model describing the data [Janzing and Schölkopf, 2010]. In particular, such a model may include latent variables Z affecting the observed X so as to respect the *independence of causal mechanisms* [Parascandolo et al., 2018].

Putting these two ideas together, we have the following natural approach: run a causal discovery algorithm on the observed data, find sets of densely connected variables in the learned graph, learn a latent factor model for each set, and then use the AMC to determine which so-found $P(X, Z)$ are simpler than $P(X)$. If so, add the best newly discovered confounder Z to X and iterate until convergence.

We show that our approach is both theoretically and empirically sound. It provably recovers the true set of confounded nodes under general conditions, while for the sparse linear

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:1016–1026.

Gaussian setting, it is consistent for recovering the entire model. Empirical evaluation shows it to be highly accurate. It improves both over methods that do assume causal sufficiency as well as those which do not. All code, data, results, and proofs can be found online on the authors' website.[1]

2 THEORY

We first describe our problem setting. We then prove identifiability for sparse linear Gaussian (SLG) models and provide a framework for causal discovery under latent confounding. With this, we derive a consistent score for the SLG.

2.1 PROBLEM SETTING

Let $X = (X_1, \ldots, X_m)$ and $Z = (Z_1, \ldots, Z_l)$ be two sets of variables with joint distribution $P(X, Z)$, where X are observed and Z unobserved variables. Our goal is to discover a network over X, Z, that is, a directed acyclic graph (DAG) $G_{X,Z} = (V, E)$ with vertices $X \cup Z$ and edges capturing the causal relationships in $P(X, Z)$. By marginalizing over Z, we obtain a distribution $P(X)$ with a corresponding network G_X. When the variables are clear from the context, we write G for $G_{X,Z}$, respectively G_X.

To permit identifiability of the network $G_{X,Z}$, we have to make some common assumptions [Koller and Friedman, 2009]. First, *Causal Faithfulness*: if U and V are independent given W in $P(X, Z)$, then U and V are d-separated by W in G. Second, the *Causal Markov Condition*: each $Y \in X \cup Z$ is independent of its non-descendants given its parents $\mathrm{Pa}_G(Y)$. Third, we do not assume *Causal Sufficiency* over X, but we *do* assume it over $X \cup Z$. That is, we assume that all common parents of at least two variables $U, V \in (X, Z)$ are included in $X \cup Z$. In other words, all non-causal correlations can be *explained away* by conditioning on the right variables. Last, we assume that all Z_j are jointly independent and that no reverse causation exists, i.e., $\mathrm{Pa}(Z_j) = \emptyset$. Under these assumptions, we have

$$P(X, Z) = \prod_{i=1}^{m} P(X_i \mid \mathrm{Pa}_i) \prod_{j=1}^{l} P(Z_j)\,,$$

where $\mathrm{Pa}_i = \mathrm{Pa}_G(X_i)$. Such factorizations are a cornerstone of causal learning [Pearl, 2009], allowing for the identification of many causal effects. Our goal is the following.

Problem Statement. *Given a sample x^n only from the observed distribution $P(X)$, discover*

- *a (small) set of latent variables Z*
- *a (sparse) network G over X and Z*

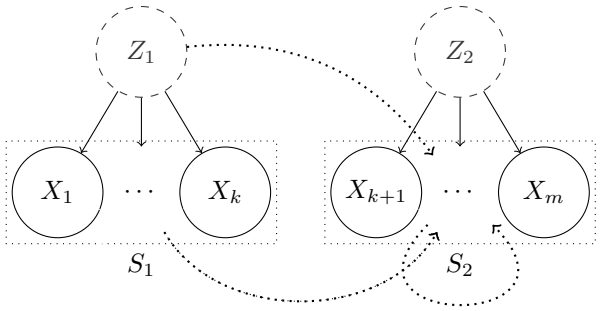


Figure 1: Structural assumptions 1 and 2 of our model. Each Z_i has an edge towards all nodes in S_i (solid), but only few other edges (dotted) are incoming to each S_i.

- *and a (simple) joint distribution $P(X, Z)$ such that*

$$P(X, Z) = \prod_{i=1}^{m} P(X_i \mid \mathrm{Pa}_i) \prod_{j=1}^{l} P(Z_j)\,,$$

factorizes according to the discovered G.

In the following, we note a simple property laying the foundations for discovering all three of these components.

2.2 STRUCTURE OF LATENT CONFOUNDING

To solve our problem, we start with the following observation. Whenever a set of variables $X_S = (X_i)_{i \in S}$, $S \subseteq \{1, \ldots, m\}$, are co-caused by an unmeasured Z, no pair $X_i, X_j \in X_S$ can be made independent by conditioning on any other subset $W \subset X$. Thus, when G captures the independences of P, all pairs in X_S are connected. That is, G contains a *clique* over X_S [Elidan et al., 2000].

Proposition 1 (Confounders and Cliques)**.** *Let $P(X, Z)$ be the joint distribution of X, Z where Z is one-dimensional and let $S = \{i : Z \to X_i\}$. Then any graph G_X capturing the correlations in $P(X)$ contains a clique over X_S.*

When Z is multivariate, each Z_j induces its own clique in G, all of which may overlap. Next, we show that this graphical characterization of confounding is already sufficient to identify the true model in the sparse linear Gaussian case.

2.3 IDENTIFIABILITY FOR THE SPARSE LINEAR GAUSSIAN MODEL

To prove identifiability of the causal model, we assume that $P(X, Z)$ is given by a linear Gaussian structural causal model (SCM, Pearl [2009])

$$X = A^\top X + B^\top Z + \epsilon \tag{1}$$

[1] https://eda.rg.cispa.io/prj/pepsi/

where A encodes the DAG G, $Z \sim N(0, I)$ and $\epsilon \sim N(0, \mathrm{diag}(\sigma_\epsilon^2))$. We further make the following assumptions on the causal model generating our data X, Z.

Assumption 1. *There exists a partition of the variables X into l disjoint sets $S_1, \ldots, S_l$ of sizes $|S_j| \geq 4$ such that for each variable $X_i \in S_j$ the direct causal effect b_{ij} of $Z_j \to X_i$ is non-zero, $b_{ij} \neq 0$.*

This assumption guarantees that each Z_j has an influence on a subset S_j of the variables X that is sufficiently large to recover its parameters b_{ij}. Of course, this would not avail us much if the overlaps between sets are too large, e.g., when two variables $Z_j \neq Z_k$ have exactly the same sets $S_j = S_k$ of downstream effects. To prevent such cases, we introduce our next assumption.

Assumption 2. *There are at most $|S_j| - 4$ edges incoming to vertices in S_j, aside from the edges $Z_j \to S_j$.*

This assumption ensures that the different Z_j, Z_k cannot have too much overlap in their S_j, either through direct connections of $Z_j \to S_k$, or through indirect paths $Z_j \to S_j \to S_k$. That is, the sets S_j are only weakly connected to each other to ensure distinguishability between the effects of different Z_j. Note in particular that since models with $Z_j \to Z_k$ are indistinguishable from models where each node in $X_i \in S_k$ also has an edge $Z_j \to X_i$, this assumption *requires* Z to be jointly independent. Likewise, as we saw in the first example, there also cannot be too many connections between variables *within* S_j, as this makes it impossible to tell which correlations are due to Z_j, and which due to causal effects of variables within S_j.

When all assumptions hold, the causal model is identifiable.

Theorem 2. *Let our distribution $P(X, Z)$ be described by the linear Gaussian SEM given in Eq.* (1)

$$X = A^\top X + B^\top Z + \epsilon\,,$$

for some Z of dimension $l \leq m/4$. Further, let assumptions 1-3 hold. Then the number l of confounders and its parameters B are identifiable up to column permutations and rescaling. Furthermore, if all noise variables ϵ have equal variances, then A is also identifiable.

Unlike this sparse linear Gaussian (SLG) case, causal models with latent variables are generally overparametrized and thus unidentifiable. As we show next, however, it is possible to identify *whether or not* latent confounders are involved.

2.4 ALGORITHMIC MODEL OF CAUSALITY

To determine whether variables X are influenced by latent confounders, we introduce the algorithmic model of causality. We begin by studying $P(X)$ under the assumption that no latent variables Z are involved. In the algorithmic model of causality, the distribution $P(X)$ with graph G corresponds to a set of programs f_i describing how each X_i is generated from Pa_i and mutually independent noise ϵ

$$X_i = f_i(\mathrm{Pa}_i, \epsilon_i)\,, \quad \epsilon_i \perp\!\!\!\perp \mathrm{Pa}_i\ .$$

We can measure the complexity of such a description of $P(X)$ using Kolmogorov complexity [Li and Vitányi, 2009]. Given a universal Turing machine U, it measures the length of the shortest program p approximating a given function f,

$$K(f) \coloneqq \min_p \{|p| \mid \forall u \forall q : |U(u, p, q) - f(u)| \leq 1/q\}\ .$$

Kolmogorov complexity optimally utilizes all available information and therefore measures the length of the best compression of f. If $P(X)$ is the distribution generated by a causal process, the *true* causal model should also compress $P(X)$ *best*. This notion is captured by the algorithmic Markov condition (AMC) [Janzing and Schölkopf, 2010].

Postulate (Algorithmic Markov Condition)**.** *Let G^* be the true causal network for $P(X)$. Then*

$$K(P(X)) \stackrel{+}{=} \sum_{i=1}^{m} K(P(X_i | \,\mathrm{Pa}_{G^*}(X_i)))\,,$$

where $\stackrel{+}{=}$ denotes equality up to an additive constant that depends on the Turing machine U but is independent of P.

In the bivariate case, if X causes Y, it says

$$K(P(X)) + K(P(Y \mid X)) \stackrel{+}{\leq} K(P(Y)) + K(P(X \mid Y)).$$

The AMC states that the true network G^* not only provides the best factorization of P but the best compression of P in general. In particular, $P(X_i \mid \mathrm{Pa}_i)$ and $P(X_j \mid \mathrm{Pa}_j)$ cannot be better compressed jointly so that they do not share any structure. In other words, the causal mechanisms generating X_i and X_j are *algorithmically independent* of each other [Janzing and Schölkopf, 2010].

This independence breaks under latent confounding. Consider the case where $(X_1, \ldots, X_4)$ are given by $X = B^t Z + \epsilon$ with unobserved univariate $Z \sim N(0, 1)$. Then $P(X)$ is fully described by its covariances $\sigma_{ij} = b_i b_j$. However, for 4 variables, there are 6 covariances described by the four parameters $b_1, \ldots, b_4$, rendering them dependent.

Finding models with independent causal mechanisms, therefore, requires us to consider models with latent factors, i.e., distributions $P(X, Z)$ with marginals $P(X)$ and $P(Z) = \prod_j P(Z_j)$, the class of which we refer to as $\mathcal{P}$. The question is whether it is meaningful to compare $P(X, Z)$ with $P(X)$ in terms of Kolmogorov complexity. The following theorem provides a positive answer to this question.

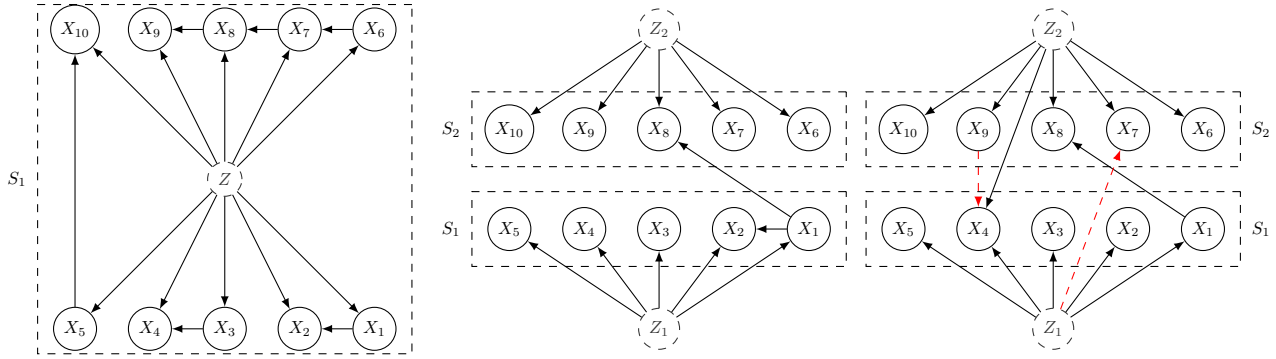


Figure 2: Example graphs illustrating our structural assumptions. (a) All observed variables are confounded by the same factor Z_1, and 6 edges exist between its children S_1. (b) Two different confounders affecting five nodes each, and one additional edge incoming to each of the sets. (c) Z_2 affects one of the nodes in S_1. Furthermore, if we added either of the dashed red edges, too many edges incoming into S_1, respectively, S_2 would render the model unidentifiable.

Theorem 3 (Kolmogorov Does Not Incorrectly Detect Confounders)**.** *For any distribution $P(X)$, we have*

$$\inf_{P(X,Z)\in\mathcal{P}} K(P(X,Z)) \overset{+}{\leq} K(P(X))\,,$$

where the infimum is over all joint distributions $P(X,Z)$ with fixed marginal $P(X)$ and independent Z. Conversely, if a joint distribution $P(X,Z)\in\mathcal{P}$ exists such that

$$K(P(X,Z)) < K(P(X))\,, \tag{2}$$

then the true generating mechanism of X includes latent variables influencing some subset X_S.

First, this result shows that there is no intrinsic bias towards one type of distribution being more complex than the other. Second a confounded model can only offer the best description of the data if the true generating mechanism involves unobserved variables. In this case, we can write

$$K(P(X,Z)) = \sum_{i=1}^{m} K(P(X_i \mid \mathrm{Pa}(X_i))) + \sum_{j=1}^{l} K(P(Z_i))$$

where the parents of X_i are among both X and Z. Note that while we cannot incorrectly infer latent confounders where there are none, we may miss latent variables that do exist.

2.5 INSTANTIATING THE AMC

Using Kolmogorov complexity, we have developed a framework for discovering latent confounders affecting $P(X)$. To use it in practice, however, we need to address two challenges: access only to a sample x and incomputability of Kolmogorov complexity [Li and Vitányi, 2009].

The first issue is resolved in the large sample limit, where measuring the suitability of a causal model using $K(x)$ instead of $K(P(X))$ leads to the same decisions [Marx and Vreeken, 2021]. To solve the second issue, we use the Minimum Description Length (MDL) principle [Grünwald, 2007]. In its simplest form, to encode data x, we choose a model M and encode both M and x given M as

$$L(x,M) = L(x \mid M) + L(M)\,.$$

We then want to find the best model M^* in some model class $\mathcal{M}$. For any $\mathcal{M}$, we have $K(x) \leq \inf_{M\in\mathcal{M}} L(x,M)$, so that MDL provides a well-founded upper bound on K.

In this work, our focus is not on identifying a single best model but on determining whether a subset X_S of the observed variables is confounded or not. To distinguish between model classes, we use refined MDL Grünwald [2007]

$$L(x,\mathcal{M}) \coloneqq -\log \int_{M\in\mathcal{M}} P(x|M)Q(M)dM\,, \tag{3}$$

where Q is a prior on the models $M\in\mathcal{M}$.

For this score to be sound, we next establish a result similar to Eq (2). More precisely, we show that confounded models obtain a better score only if causal sufficiency is violated. To do this formally, we require two things. First, a class $\mathcal{M}_0$ of models without latent factors. Second, a set $\mathfrak{M}$ of model classes $\mathcal{M}$ describing the observed distribution $P(X)$ as marginal of a joint distribution $P(X,Z)$ in some distinct way. For data x sampled from $P(X)$, Eq. (2) then corresponds to asking whether or not

$$\inf_{\mathcal{M}\in\mathfrak{M}} L(x,\mathcal{M}) < L(x,\mathcal{M}_0)\,. \tag{4}$$

The following theorem tells us that, indeed, Eq. (4) holds only if latent variables were involved in generating x.

Theorem 4 (MDL Does Not Incorrectly Detect Confounders)**.** *Let x^n be the observed part of an i.i.d. sample*

from $P \in \mathfrak{M} \cup \{\mathcal{M}_0\}$. Further, assume that Eq. (4) *holds P-almost surely as $n \to \infty$. Then $P \in \mathfrak{M}$.*

Just as in the case of algorithmic causality, i.e., Thm. 3, it is generally impossible to guarantee recovery of the exact true model. However, for the SLG, it is possible to define a consistent score, requiring no further assumptions than those we already made for identifiability. in Sec. 2.3.

Theorem 5 (Consistency of BIC for SLGs)**.** *Let $x = x^n$ be a sample from the SLG of Eq.* (1) *and let assumptions (A1)-(A3) hold. Let $\mathcal{M}$ be the corresponding model class and $\mathcal{M}_0$ the restriction of $\mathcal{M}$ to models with $B = 0$. Let*

$$L(x^n, M) = -\log P(x^n \mid A, B, \sigma_\epsilon^2) + \lambda\|A\|_0 + \lambda\|B\|_0 \tag{5}$$

and $\widehat{A}, \widehat{B}$ its minimizers. Then for $\lambda = \log(n)/2$, the score L is consistent for detecting confounders. That is,

$$\lim_{n\to\infty} P\left(\min_{M\in\mathcal{M}} L(x^n, M) < \min_{M\in\mathcal{M}_0} L(x^n, M)\right) = 1\,.$$

Further, $\widehat{A}$ and $\widehat{B}$ converge to the true A, B in probability,

$$\lim_{n\to\infty} P(\widehat{A} = A, \widehat{B} = B) = 1\,.$$

While it is remarkable that in the linear Gaussian case, the full causal model can be recovered from a sample x from $P(X)$ only, we have to solve two problems before we can put this into practice. First, we do not know B, nor even which of the exponentially many subsets $X_S \subseteq X$ are affected by any one Z_j. Second, even knowing B and Z, optimizing Eq. (5) is NP-hard [Peters and Bühlmann, 2012]. We, therefore, next develop a good heuristic as to which subsets X_S are likely affected by Z and show how standard causal discovery algorithms can be leveraged to find a causal network over both the observed X and the latent Z.

3 THE CDHC ALGORITHM

With the theory we developed, we can now introduce CDHC, our method for Causal Discovery with Hidden Confounders.

3.1 FINDING CONFOUNDED VARIABLES

To find a causal network over both X and its confounders Z, we first need to determine which subsets of X are likely to be confounded. A structure learning algorithm $\mathcal{A}$ can help us determine these sets by discovering connected subsets of variables (Prop. 1). If X_S are densely connected in the discovered graph G, then the Markov boundaries satisfy $X_i \in X_S$ is $\mathrm{MB}(X_i) \approx X_S$. Therefore, we consider the Markov boundary of each node as the seed sets S over which we may infer latent confounders.

3.2 LEARNING LATENT CONFOUNDERS

To evaluate a proposed set of confounded nodes and its associated graph G, we introduce a causal model including latent factors as follows. Based on Theorem 2, given a graph G over (X, Z), we assume that the data is generated from a model in a class $\mathcal{M} = \mathcal{M}(G)$ similar to Probabilistic PCA (PPCA) [Tipping and Bishop, 1999]

$$\begin{aligned} Z_i &\sim N(0,1)\,, & \epsilon &\sim N(0, \sigma_\epsilon^2) \\ A_{ij} &\sim N(0, \sigma_a^2)\,, & B_{ij} &\sim N(0, \sigma_b^2) \\ X &= A^\top X + B^\top Z + \epsilon\,, \end{aligned} \tag{6}$$

where entries of A, B are nonzero only when their corresponding edges are in G. By marginalizing out Z, we obtain

$$X \mid A, B \sim N(0, C(B^\top B + \sigma_\epsilon^2)C^\top)\,,$$

where $C = (I - A^t)^{-1}$. Since the A_{ij} are sampled from a continuous distribution, assumption (A3) holds almost surely so that, unlike PPCA, we do not require $A = 0$.

To evaluate the fit of our model class $\mathcal{M}$ to our data x, we use the score $L(x, \mathcal{M})$ defined in Eq. (3)

$$L(x, \mathcal{M}) = -\log \int P(x \mid A, B)P(A, B)dAdB\,,$$

which we can estimate using standard variational methods [Kucukelbir et al., 2017]. This score is suitable in that it is consistent for causal discovery with latent confounders.

Theorem 6 (Consistency of MDL for SLGs)**.** *Let the assumptions of Thm. 5 hold. Then the minimizer $\widehat{G}$,*

$$\widehat{G} = \mathit{arg\,min}_G L(x^n, \mathcal{M}(G))\,,$$

converges to the ground truth G^* with probability one,

$$\lim_{n\to\infty} P(\widehat{G} = G^*) = 1\,.$$

With this guarantee that our score is sound, we now introduce our method for discovering the entire causal network.

3.3 DISCOVERING THE CAUSAL NETWORK

We can now put all of the above together and present CDHC. We give the pseudo-code as Algorithm 1. We first (line 1) discover a graph G over the observed data x using a score-based structure discovery algorithm $\mathcal{A}$, such as GES [Chickering, 2002], GGSL [Gao et al., 2017] or NOTEARS [Zheng et al., 2018]. We then consider every node X_i and initialize the confounded set X_S with the Markov boundary $\mathrm{MB}(X_i)$ and add a node Z and edges $Z \to X_S$ to G. (l. 4-5). We refine S by greedily adding nodes (l. 6-9), then removing nodes (l. 10-13). After finding the locally optimal set S, we

Algorithm 1: CDHC

input : data x sampled from $P(X)$, algorithm $\mathcal{A}$
output : graph G and distribution $P(X, Z)$

1 $G = (V, E) \leftarrow$ Graph inferred over x using $\mathcal{A}$;
2 **do**
3 **foreach** i $\in \{1, \ldots, m\}$ **do**
4 $X_S \leftarrow$ Markov boundary of X_i in G;
5 $G' \leftarrow (V \cup \{Z\}, E \cup \{Z \rightarrow X_S\})$;
  // Forward phase
6 **do**
7 $j \leftarrow \arg\min_{j \notin S} L(x, G' \cup \{Z \rightarrow X_j\})$;
8 $(S, G') \leftarrow (S \cup \{j\}, G' \cup \{Z \rightarrow X_j\})$;
9 **while** $L(x, G')$ decreases;
  // Backward phase
10 **do**
11 $j \leftarrow \arg\min_{j \in S} L(x, G' \setminus \{Z \rightarrow X_j\})$;
12 $(S, G') \leftarrow (S \setminus \{j\}, G' \setminus \{Z \rightarrow X_j\})$;
13 **while** $L(x, G')$ decreases;
14 $z \leftarrow$ sample from $P(Z \mid X)$;
15 $G[i] \leftarrow$ Graph inferred over (x, z) using $\mathcal{A}$;
  // Use the model with best confounder
16 $G \leftarrow \arg\min_{G[i]} L((x, z), G[i])$;
17 **while** G changes;
18 **return** G and the $P(X, Z)$ associated with G

sample z from $P(Z \mid X)$ and fit a network over (x, z) using $\mathcal{A}$ (l. 14-15). Out of all these networks, we update G to be the best of them (l. 16) and iterate until convergence (l. 17). Finally, we return the discovered network G and distribution $P(X, Z)$ over X and its inferred confounders Z (l. 18).

Note that since our score strictly decreases at every step, our method necessarily converges. Furthermore, we can show that in the large sample limit we are guaranteed to recover the true set of confounded nodes.

Proposition 7 (Consistency of CDHC for Discovering Confounded Nodes)**.** *Let x^n be the an i.i.d. sample from $P \in \mathcal{M}(G^*)$ defined in Eq.* (6)*, let assumptions (A1-3) hold and let S_i^* be the set of nodes affected by Z_i. Assume that $\bigcap_{s \in S_i^*} MB_{G^*}(X_s) \setminus \{Z_i\} \subsetneq S_i^*$. Let $\mathcal{A}$ be a consistent for recovering the Markov equivalence class of the graph G_X for distribution $P(X)$. Let $\widehat{S}_i$ be the set nodes confounded by Z_i discovered by* CDHC*. Then*

$$\lim_{n \to \infty} P(\widehat{S}_i = S_i^*) = 1.$$

While we can recover the correct sets of confounded nodes, to recover the entire graph G^*, we need additional assumptions such as those outlined in Theorem 5.

3.4 COMPLEXITY

Last, we analyze the runtime complexity of CDHC. CDHC employs a loop (l. 2-17) whose inside has complexity $O(C(m, n) + m^2 n) = O(C(m, n))$ – the former, $C(m, n) = \Omega(m^2 n)$, is the runtime for running $\mathcal{A}$ and the latter $m^2 n$ for finding S. Since we can find at most $O(m)$ non-overlapping confounded sets, our worst case runtime is therefore on the order of $O(mC(m, n))$. In general, only few variables are confounded so that in practice, our runtime is roughly $O(C(m, n))$ — the same as that of $\mathcal{A}$ itself.

4 RELATED WORK

Causal inference is one of the most important problems in statistical inference and has attracted a lot of research attention [Pearl, 2009, Spirtes et al., 2000]. Unfortunately, latent confounding makes it impossible to infer causality from observational data without making additional assumptions [Pearl, 2009]. Traditional constraint-based [Spirtes et al., 2000, Zhang, 2008] and score-based [Chickering, 2002, Gao et al., 2017, Zheng et al., 2018] causal discovery methods can reconstruct the true causal network up to Markov equivalence when causal sufficiency holds.

When causal sufficiency does not hold, a number of algorithms such as the FCI family [Spirtes et al., 2000, Colombo et al., 2012, Ogarrio et al., 2016], 3OFF2 [Affeldt et al., 2016] and DCD [Bhattacharya et al., 2021] can find complete partially directed acyclic graphs which can capture correlations due to confounders. However, these networks are generally difficult to interpret and cannot determine which sets of variables share the same latent confounder.

To make the resulting causal networks more interpretable, observational and experimental data can be combined to improve results [Kallus et al., 2018, Kocaoglu et al., 2019]. However, these methods are generally restricted in their ability to rule out or corroborate the existence of latent variables by the scarcity of available experimental data.

Other research controls causal estimates for latent confounders. To do so, Hoyer et al. [2008] solve the overcomplete ICA problem to correct the estimated causal effect of X on Y for confounders, whereas Wang and Blei [2018] and Ranganath and Perotte [2018] use factor models.

Until recent years, only little prior research has tackled the topic of determining which variables share the same latent confounders. Janzing and Schölkopf [2018] considered the case of determining whether the variables X are confounded by finding deviations of the regression vector from theoretical properties in high-dimensional regression. Kaltenpoth and Vreeken [2019] use the AMC [Janzing and Schölkopf, 2010] to infer whether two sets of variables X and Y are causally related or jointly confounded. Silva et al. [2006] proposed a model based on low-rank correlation structures

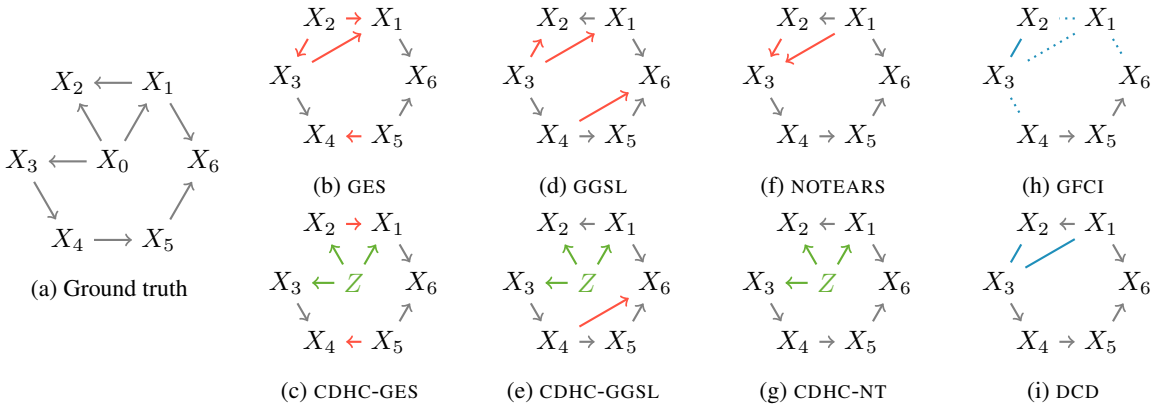


Figure 3: Application of CDHC to synthetic data generated from the network shown in (a). When X_0 is withheld, all base algorithms $\mathcal{A}$ (b,d,f) find a clique of spurious edges on X_1, X_2, X_3 (red). GFCI (h) and DCD (i) indicate that some pairs from X_1, X_2 and X_3 are confounded (blue) but cannot tell that they share the same confounder. In contrast, by applying CDHC (c, e, g), we discover a confounder capturing the effect of X_0 (green) and obtain higher quality networks in all cases.

between observed variables. Elidan et al. [2000] proposed an algorithm for replacing semi-cliques in a discovered causal graph with single nodes based on the idea that such cliques are likely due to latent confounding.

More recently, work based on trek-separation [Kummerfeld and Ramsey, 2016], rank-constraints [Huang et al., 2022], heterogeneous data sources [Zhou et al., 2022], and over-complete ICA-based approaches [Xie et al., 2020, Adams et al., 2021] have been used to obtain the (hierarchical) latent structure of the observed variables. They do not, however, permit edges between the observed variables. This is in line with the recent field of causal representation learning [Schölkopf et al., 2021], where it is assumed that *all* observed correlations are due to causal relations between the unobserved variables. While such assumptions are realistic for data such as images (pixels don't cause each other), they are not reasonable for data gathered from physical, biological, or social systems.

5 EXPERIMENTS

In this section, we evaluate CDHC empirically. We are interested in two things; first, how well it recovers the set of confounded nodes S^*, and second, how well it recovers the entire network. We compare CDHC against NOTEARS [Zheng et al., 2018], 3OFF2 [Affeldt et al., 2016], DCD [Bhattacharya et al., 2021] and GFCI [Ogarrio et al., 2016].[2] We instantiate CDHC with different causal discovery algorithms $\mathcal{A}$ and refer to CDHC using $\mathcal{A}$ as CDHC-$\mathcal{A}$. Specifically, we use GES [Chickering, 2002], GGSL [Gao et al., 2017], and NOTEARS [Zheng et al., 2018]. When clear from the context, we write CDHC for CDHC-GES. We implement CDHC in Python. For comparison with all other methods, we use the implementations provided by the respective authors. All experiments finished within minutes on a commodity laptop. All code and data can be found online, along with additional experiments postponed in the interest of space.[3]

5.1 SYNTHETIC DATA

We evaluate CDHC on synthetic data by generating a random acyclic graph G of size m from the Erdős-Rényi model with parameter $p = 0.3$. We model the causal relationships via a linear SEM, $X = AX + \alpha BZ + \epsilon$ where $A_{ij} \neq 0$ if and only if $(i, j) \in E(G)$. Nonzero values of A, B, Z, ϵ are all $\sim N(0, 3)$. The parameter $\alpha \sim U[1, 8]$ determines the relative strength of confounding. Using this model, we generate 1000 data sets over $m = 50$ variables, of which ten nodes are confounded. Before moving to the general case, we begin by studying how CDHC improves over base algorithms $\mathcal{A}$ on an illustrative example with $\dim(Z) = 1$.

Comparison with Base Algorithms We begin by showing that CDHC produces better results than standard discovery algorithms when not all variables are observed. We consider the network shown in Fig. 3a containing nodes $X_0, ..., X_6$, of which X_1, X_2, X_3 are confounded by X_0. When withholding X_0, none of the base methods find the correct structure over X_1, X_2, X_3. Furthermore, while GFCI and DCD find the variables to be confounded, they cannot tell that all variables share the same confounder. In contrast, by applying CDHC, we consistently find that X_1, X_2, X_3 are confounded while maintaining the quality of the remaining edges.

Confidence and Performance Since we generally do not have access to the ground truth network in practice, we next test how well we can predict the performance of each method from an easily observable quantity. That is, we compare the

[2] GFCI is part of a group of methods, including FCI and RFCI. Preliminary experiments corroborated previous research [Ogarrio et al., 2016] that GFCI performs better than its relatives.

[3] https://eda.rg.cispa.io/prj/pepsi/

Method	Number of confounders 1	2	3	4	5
CDHC	0.43	0.38	0.35	0.23	0.15
DCD	0.35	0.18	0.11	0.07	0.03
3OFF2	0.36	0.2	0.14	0.11	0.04
GFCI	0.22	0.11	0.05	0.02	0.01

Table 1: Comparison of CDHC, DCD, 3OFF2 and GFCI for graphs with varying numbers of latent confounders. While all methods perform well, only CDHC maintains its performance as the number of latent factors increases.

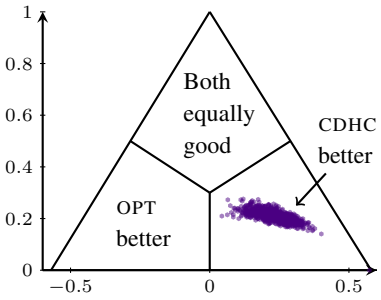


Figure 5: Significance assessment of the improvement of CDHC against its competitors.

confidence—the improvement of the discovered network compared to a baseline—of each method to a range of metrics measuring the quality of our results. We provide details on the computation of the confidences in Appendix A.3.

We evaluate each method based on four criteria: (1) the F_1 score for network recovery including the confounder (F_1^{net}), (2) the F_1 score for the recovery of the set of confounded nodes (F_1^{conf}), (3) the Structural Hamming Distance (SHD) between the discovered and true network, and (4) the Structural Intervention Distance (SID) to measure differences in causal interpretations [Peters and Bühlmann, 2013].

We show the results in the form of decision rate (DR) plots in Fig. 4. First, we sort the result of each method by their confidence. Then we plot the confidence against each metric. On the left of each plot, we include data sets where each method is most confident and increase this number of sets until all are included on the right. We see that CDHC outperforms its competitors by a large margin. Interestingly, the choice of $\mathcal{A}$ has little influence on the performance of CDHC, and the gap between each $\mathcal{A}$ with CDHC-$\mathcal{A}$ is comparable. We include further results for GGSL and GES in Appendix A.4.

Higher-dimensional Z We next consider the effect of including multiple confounders Z_i in our causal model, each influencing non-overlapping sets of 5 variables in a network of $m = 50$ variables. We show the F_1^{conf} scores for one to five confounders in Table 1. We omit NOTEARS, GGSL, and GES since none of them are designed for this task.

We see that for one to three confounders, CDHC performs at a consistent level, but for four and five confounders, its performance decreases due to the difficulty of finding additional sets of confounded nodes. In contrast, while DCD, 3OFF2, and GFCI perform well for single-dimensional confounders, their performance drops immediately upon addition of a second latent confounder. The reason for this is instructive: since they do not model the confounder but instead indicate whether pairs of variables are confounded, they cannot distinguish between different confounders.

Significance To verify that CDHC significantly outperforms its competitors, we use the Bayesian signed rank test [Benavoli et al., 2014]. It explicitly models the probability that one model is significantly better than the other *in practice* by introducing a *region of practical equivalence* (rope) specified by parameter r [Benavoli et al., 2014]. Two methods are considered to perform equally well if the difference in scores for the methods lies in $[-r, r]$. We pick $r = 0.05$ [Benavoli et al., 2014] but the conclusion remains the same for values $r \in (0, 0.15]$. Since the test was designed for *two* competing methods, for each dataset, we compare CDHC with the best-performing competitor, which we refer to as OPT. For each dataset k, we compute the F_1^{net} scores for both CDHC and OPT and compute their differences. We include more detail in Appendix A.5. We show the estimated pos-

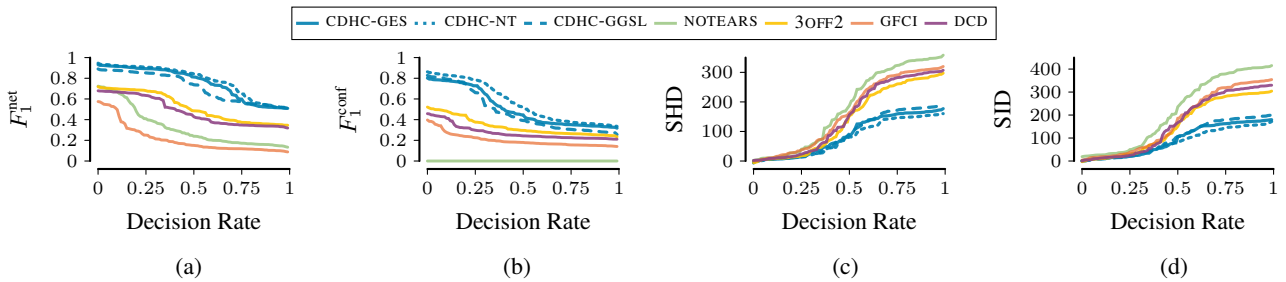
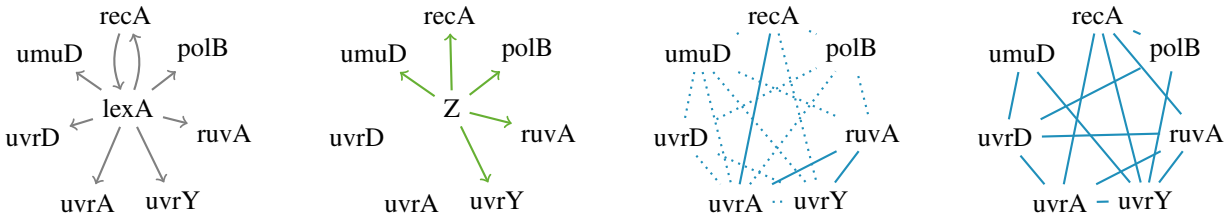


Figure 4: Evaluation on synthetic data. F_1 scores for (a) network recovery and (b) confounded set recovery (higher is better), and (c) Structural Hamming Distance and (d) Structural Intervention Distance (lower is better). Each figure shows the average score over increasing fractions of all datasets, sorted in descending order by the confidence of each method.



(a) Ground Truth (b) CDHC (c) GFCI (d) DCD

Figure 6: Results on SOS DNA repair network in E. coli. CDHC (b) discovers a confounder Z capturing five out of the seven edges (green) in the ground truth network (a). In contrast, GFCI (c) and DCD (d) find many pairs of nodes that are confounded (solid blue), and in the case of GFCI, more pairs yet which *might* be confounded (dotted). However, they do not discover that all nodes share the same confounder, making the resulting networks challenging to interpret.

terior distribution resulting from this comparison in Fig. 5. Here, the top region contains points where both methods are practically equally good, while the bottom left and right corners contain points where OPT, respectively CDHC, perform better. We see that all points lie squarely in the region, indicating CDHC to outperform OPT, suggesting that CDHC performs significantly better than its competitors.

5.2 CASE STUDY: CELLULAR SIGNALING

Finally, we consider real-world data to investigate the interpretability of the results returned by CDHC. In particular, we consider the SOS DNA repair network in E. coli [Ronen et al., 2002]. This data consists of protein levels of eight genes measured every five minutes for five hours, resulting in only 60 samples. Since the governing relationships in gene regulation are highly nonlinear, this tests the applicability of CDHC even when our assumptions do not hold.

Since the ground truth network has been established [Perrin et al., 2003], we can test CDHC by excluding a gene known to have a downstream causal effect on other genes. An excellent candidate is *lexA* as it has a causal influence on *all* of the other genes: it is upstream of six genes and has a bidirectional relationship with the seventh (Fig. 6a). We also applied the other methods on the same data, including here the results of both GFCI and DCD, and postpone the networks discovered by the other methods to Appendix A.7.

We show the results in Fig. 6. For clarity, we focus only on discovering which confounded nodes. In Fig. 6b, we find a striking similarity between the Z discovered by CDHC and the true common parent *lexA*. CDHC correctly identifies five out of seven relationships: four out of six downstream effects, as well as one of the two edges between recA and lexA—which is the most a DAG can do, given that the two edges are mutually exclusive. Next, for GFCI (Fig. 6c) and DCD (Fig. 6d), we indicate definite confounding by solid edges and correlations which could be due to either confounding or causation by dotted edges. GFCI indicates definite confounding for only three out of 16 pairs, while we cannot be certain for the other pairs. The resulting network of DCD indicates definite confounding for many pairs of variables. However, neither method can determine that all variables share the *same* latent confounder. The results of both GFCI and DCD are consistent with many different structures and provide no well-founded way of choosing one over the other. That is, since GFCI and DCD indicate only pairwise confounding without any way to evaluate whether nodes are jointly confounded, interpretation of the resulting networks is difficult. Overall, despite low sample size and violations of our model assumptions, CDHC finds a readily interpretable network close to the ground truth.

6 DISCUSSION AND CONCLUSION

We studied the problem of discovering a causal network over both X and its latent confounders Z. In particular, by exploiting the structure among confounded nodes X_S we proved identifiability in the sparse linear Gaussian model.

We derived a general approach for discovering sets of confounded nodes from the algorithmic model of causality by explicitly modeling latent variables. We showed that including latent variables is only beneficial when the observed X are indeed confounded. We used MDL to determine which sets of variables are confounded and showed that CDHC is consistent for recovering the set of confounded nodes when combined with a consistent causal discovery method.

Evaluation of CDHC on synthetic data showed that it outperformed the state of the art on all considered metrics. Furthermore, on real data it generated results close to the ground truth despite small sample size and large deviations from its model assumptions. We further obtained more readily interpretable results than our competitors.

For the future, we are interested in more general identifiability results as well as developing better algorithms which do not require multiple passes of a causal discovery algorithm.

References

Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34: 22822–22833, 2021.

Séverine Affeldt, Louis Verny, and Hervé Isambert. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. In *BMC bioinformatics*, volume 17, pages 149–165. BioMed Central, 2016.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

A. Benavoli, F. Mangili, G. Corani, M. Zaffalon, and F. Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *ICML*, pages 1026–1034. JMLR, 2014.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2021.

David Maxwell Chickering. Learning equivalence classes of Bayesian-network structures. *JMLR*, 2:445–498, 2002.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals Stat.*, 40(1):294–321, 2012.

Gal Elidan, Noam Lotner, Nir Friedman, and Daphne Koller. Discovering hidden variables: A structure-based approach. *Advances in Neural Information Processing Systems*, 13, 2000.

Tian Gao, Kshitij Fadnis, and Murray Campbell. Local-to-global bayesian network structure learning. In *International Conference on Machine Learning*, pages 1193–1202. PMLR, 2017.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Patrik O. Hoyer, Shohei Shimizu, Antti J. Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *Int. J. Approx. Reason.*, 49:362–378, 2008.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *arXiv preprint arXiv:2210.01798*, 2022.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE TIT*, 56: 5168–5194, 2010.

Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization. In *NIPS*, pages 6920–6931. 2018.

David Kaltenpoth and Jilles Vreeken. We are not your real parents: Telling causal from confounded using mdl. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2019.

Hugo Kesteloot, Susana Sans, and Daan Kromhout. Dynamics of cardiovascular and all-cause mortality in western and eastern europe between 1970 and 2000. *European heart journal*, 27(1):107–113, 2006.

Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *NIPS*, pages 14346–14356. 2019.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *JMLR*, 18:430–474, 2017.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1655–1664, 2016.

Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media, 2009.

Alexander Marx and Jilles Vreeken. Formally justifying mdl-based inference of cause and effect. *arXiv preprint arXiv:2105.01902*, 2021.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Bruno-Edouard Perrin, Liva Ralaivola, Aurélien Mazurie, Samuele Bottani, Jacques Mallet, and Florence d' Alché–Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19:138–148, 2003.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *arXiv:1205.2536*, 2012.

Jonas Peters and Peter Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. *arXiv preprint arXiv:1306.1043*, 2013. URL `https://arxiv.org/abs/1306.1043`.

Rajesh Ranganath and Adler Perotte. Multiple causal inference with latent confounding. *CoRR*, abs/1805.08273, 2018. URL `http://arxiv.org/abs/1805.08273`.

Michal Ronen, Revital Rosenberg, Boris I. Shraiman, and Uri Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *PNAS*, 99, 2002.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2), 2006.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *J. R. Statist. Soc. B*, 61: 611–622, 1999.

Yixin Wang and David M. Blei. The blessings of multiple causes. *JASA*, 114:1574–1596, 2018.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

Jiji 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.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422*, 2018.

Fangting Zhou, Kejun He, and Yang Ni. Causal discovery with heterogeneous observational data. In *Uncertainty in Artificial Intelligence*, pages 2383–2393. PMLR, 2022.

Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition (Supplementary Material)

David Kaltenpoth[1] **Jilles Vreeken**[1]

[1]CISPA Helmholtz Center for Information Security, Saarbrücken

A APPENDIX

A.1 PROOFS

Proof of Proposition 1. For any two $i, j \in S$ we know that, since they are direct descendants of Z, $X_i \not\perp\!\!\!\perp X_j \mid U$ for any $U \subset \{X_1, \dots, X_m\} \setminus \{X_i, X_j\}$. Hence all edges $\{X_i, X_j\}$ are in G so that S is a clique in G. □

Proof of Theorem 2. fix We prove this statement in two steps. First, we show that all b_{ij} are identifiable. Let $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, l\}$. Then, by assumption (A2) there exists a distinct quadruple (X_i, X_u, X_v, X_w) of nodes that are conditionally independent given Z_j. In order to make every quadruple (X_i, X_u, X_v, X_w) be dependent conditional on Z_j, it would have to have either an edge between them or a common predecessor, which would require at least $|S_j| - 3$ incoming edges to S_j from sources that are not Z_j.

Therefore, for any two variables (X_λ, X_μ) in our quadruple we know that $\sigma_{\lambda\mu} = \mathrm{cov}(X_\lambda, X_\mu) = b_\lambda b_\mu$ and in particular

$$\sigma_{iu}\sigma_{vw} = b_i b_u b_v b_w = \sigma_{iv}\sigma_{uw}$$

We can therefore write

$$b_i^2 = \sigma_{iu}\sigma_{iv}/\sigma_{uv}.$$

Furthermore, no quadruple $(X_i, X_{u'}, X_{v'}, X_{w'})$ this is not conditionally independent given Z can satisfy the constraint $\sigma_{iu}\sigma_{vw} = \sigma_{iv}\sigma_{uw}$ by assumption (A3). Hence, all b_{ij}^2 are identifiable, and since we know the sign of at least one b_{ij} for the given j, we therefore know the sign of each b_{ij} for fixed j. However, since j was arbitrary, B is identifiable in its entirety.

Now, knowing the values of B we can determine the distribution $P(X \mid Z)$, which depends only on A and σ_ϵ^2. Since $X \mid Z$ is now a linear Gaussian SEM with equal variances, the identifiability of A and σ_ϵ^2 follows from the work of Peters and Bühlmann [2012] on the identifiability of the equal variance model. □

Proof of Theorem 3. To prove the first statement, let Z be jointly independent and let there be no edges $X \to Z$. Pick P such that $P(Z = 0) = 1$. Then Z contains no information about X so that $K(P(X, Z)) \leq K(P(X)) + K(P(Z)) = K(X) + O(1)$, with constant $K(P(Z)) = O(1)$ independent of $P(X)$.

For the second statement, consider the case where the true generating mechanism for X does not include any latent variables for any subset X_S. Then as noted in the AMC and the discussion preceding it, *all* information needed to compress $P(X)$ is already present in the graph G_X^* giving the optimal factorization of $P(X)$. Hence $K(P(X, Z)) \geq K(P(X)) + K(P(Z|X)) > K(P(X))$. □

Proof of Theorem 4. Assume that Eq. (4) holds in the limit $n \to \infty$. Then there is a model class $\mathcal{M} \in \mathfrak{M}$ such that $L(x \mid \mathcal{M}) < L(x \mid \mathcal{M}_0)$. Further, as we use a refined MDL score that means there is an optimal $P^* \in \mathcal{M}$ such that

$L(x \mid \mathcal{M}) = L(x \mid P^*) + O(n^{-1})$. Due to the consistency [Grünwald, 2007] of refined MDL scores this means that almost surely X has generating distribution $P^*(X)$. As P^* is a joint distribution over (X, Z) this means that x is the observed part of a joint sample (x, z) from $P^*(X, Z)$. □

Proof of Theorem 5. As we've seen in the proof of Theorem 2, for each X_i there exists a distinct quadruple (X_i, X_u, X_v, X_w) conditionally independent given Z_j by (A2). Hence, all correlations between these four variables can be explained by the parameters in B. Furthermore, by Proposition 1, no pair of variables X_μ, X_λ can be d-separated in any DAG over X, so that by setting $b_{ij} = 0$ we would require *at least* four additional entries of A to be non-zero, instead of only one in B.

Hence, since in the limit we have $\widehat{b}_{ij}\widehat{b}_{vj} - \sigma_{iv} \to 0$, the matrix $\widehat{B}$ converges towards B. Furthermore, given a good approximation of $P(X)$ and of B, we obtain a good approximation of A by the results of Van de Geer and Bühlmann [2013]. □

Proof of Theorem 6. This follows directly from Theorem 5 and the fact that for $n \to \infty$, our MDL score is equivalent to BIC [Grünwald, 2007]. □

Proof of Proposition 7. By Proposition 1, we know that S_j^* forms a clique in the graph G^* inferred by a consistent $\mathcal{A}$. This clique is maximal due to no node being in the Markov Blanket of all $s \in S_j^*$. Further, since x^n is a sample from Eq. (6) we know from the MDL principle for selecting nested model classes [Grünwald, 2007] that in the limit no other set can be compressed better by introducing a confounder than S_j^* itself. □

A.2 IMPLEMENTATION DETAILS

We implemented CDHC in Python using PyMC3 for posterior inference using ADVI with default parameters. All code is available for research purposes. Experiments were run single-threaded on a standard commodity laptop and each experiment finished within minutes.

A.3 COMPUTING CONFIDENCES

For CDHC we measure its confidence as the relative gain in compression due to addition of confounders, $C_{\mathcal{A}} = (L_{\mathcal{A}} - L_{\text{CDHC-}\mathcal{A}})/L_{\mathcal{A}} \geq 0$ and for NOTEARS we use the normalized difference between the initial (empty network) and final (discovered network) score obtained from optimization. None of GFCI, 3OFF2 or DCD come with readily computable confidence scores so we treat them in *the way most favorable to them* by assuming that their best performances are also their most confident.

A.4 ADDITIONAL RESULTS FOR GGSL AND GES

We now provide additional details on the results of GGSL and GES in Table 1. We compute all confidences as described in the previous section. For comparison we also include the results of CDHC. Since they are most interpretable, we include only the F_1^{net} score here. For other metrics too, however, both GGSL and GES perform similarly to NOTEARS so that CDHC significantly outperforms them. In particular, neither of them can find any confounders.

	Data evaluated		
Method	1%	50%	100%
CDHC	**0.92**	**0.85**	**0.53**
GGSL	0.68	0.42	0.24
GES	0.64	0.35	0.21

Table 1: Comparison of CDHC, GGSL and GES in terms of F_1^{net} scores. The performance of each method is shown for each of 1%, 50% and 100% of the data evaluated (corresponding to leftmost, median and right-most points in a decision-rate plot). We see that CDHC clearly outperforms both competitors by a large margin and further that both GGSL and GES perform similarly to NOTEARS.

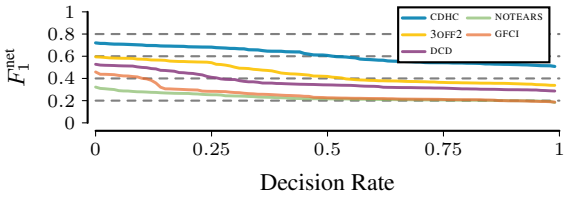


Figure 1: [Higher is better] Decision rate for CDHC and its competitors on the REGED dataset. Overall, CDHC outperforms all other methods both for points where they are confident as well as those where they are not.

A.5 ADDITIONAL DETAILS ON SIGNIFICANCE TESTING

To verify that CDHC significantly outperforms its competitors we use the Bayesian signed rank test [Benavoli et al., 2014]. It explicitly models the probability that one model is significantly better than the other *in practice* by introducing a *region of practical equivalence* (rope) specified by parameter r. Two methods are considered to perform equally well if the difference of scores for the methods lies in $[-r, r]$. We pick $r = 0.05$ [Benavoli et al., 2014] but the conclusion remains the same for values $r \in (0, 0.15]$. Since the test was designed for two competing methods, for each dataset we compare CDHC with the best-performing of its competitors, which we refer to as OPT.

To compare the two methods over all samples, we aggregate the F_1 scores for both CDHC and OPT and take their differences $z_i = F_{1,i}^{\text{OPT}} - F_{1,i}^{\text{CDHC}}, i \in \{1, \dots, q\}$. To include the prior assumption that both methods are equally good, we include a pseudo-observation $z_0 = 0$, i.e. that both methods are precisely equally good. We take weights $w = (w_0, \dots, w_q) \sim \text{Dirichlet}(s, 1, \dots, 1)$ where s corresponds to the number of times we obtained z_0. This is commonly set to be $s = 0.5$, but due to our large number of experiments its influence on the posterior is minor. The posterior probabilities are computed as

$$\theta_{\text{OPT}} = \sum_{i,j=0}^{q} w_i w_j I_{(2r,\infty)}(z_i + z_j)$$

$$\theta_{\text{rope}} = \sum_{i,j=0}^{q} w_i w_j I_{[-2r,2r]}(z_i + z_j)$$

$$\theta_{\text{CDHC}} = \sum_{i,j=0}^{q} w_i w_j I_{(-\infty,-2r)}(z_i + z_j)$$

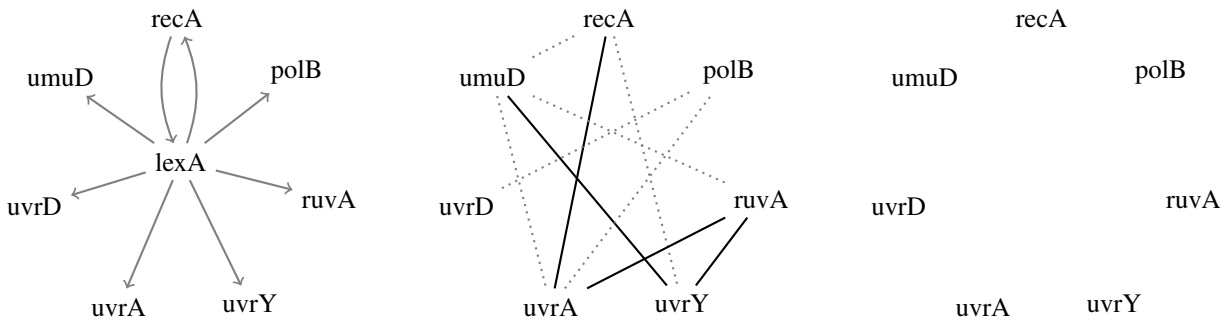
where $\theta_{\text{OPT}}, \theta_{\text{CDHC}}$ are the posterior probabilities that OPT, respectively CDHC are better by at least a margin r, while θ_{rope} is the posterior probability that they perform equally well up to said margin. The distribution of θ is not analytically tractable, but we can evaluate it empirically by sampling values for w. Such a sample is precisely what is depicted in Fig. 5 in barycentric coordinates.

A.6 REALISTIC DATA: REGED

Next, we consider realistic synthetic data from REGED [Guyon et al., 2008], which is based on human lung-cancer microarray gene expression data. Since the available samples are non-i.i.d., the causal relationships are nonlinear, and the ground truth is known from gene intervention studies, it provides a good benchmark for CDHC.

To make CDHC applicable to the REGED dataset we consider the following setup. For each node X_i in the ground truth graph G^* with $k \geq 5$ children, the set of which we denote by $C = C_i$, we select a random subset $R = R_i$ also consisting of k nodes of G^* which are not contained in the Markov boundary of any X_i in G^* and do not have a common parent. We then consider the induced subgraph G_i over the nodes $C_i \cup R_i \cup \{i\}$. However, the data given to each method is only over the variables $X_{C \cup R}$ from which we compute the results in Fig. 1.

We show the results for F_1^{net} for different methods in a DR plot in Fig. 1. Even though the data violates our assumptions, CDHC outperforms its competitors by a large margin. Moreover, even for those sets of variables where CDHC is only moderately confident, it still performs better than its competitors at their *most* confident. This suggests that CDHC works reliably even when the true model deviates from our assumptions.



(a) Ground Truth (b) 3OFF2 (c) NOTEARS

Figure 2: 3OFF2 and NOTEARS on the SOS dataset. As before, only (potentially) confounded pairs are drawn in the figures. We see that 3OFF2, like GFCI and DCD cannot determine all nodes to be jointly confounded. Meanwhile NOTEARS assumes causal sufficiency and therefore finds no indication of confounding.

A.7 OTHER METHODS ON THE SOS NETWORK

In Fig. 2 we show the results of 3OFF2 and NOTEARS on the SOS dataset. Like GFCI, 3OFF2 can only give indications about which pairs might be confounded, and for the majority of pairs it is not confident. Meanwhile, by its very design NOTEARS has no notion that confounders might be involved.

References

A. Benavoli, F. Mangili, G. Corani, M. Zaffalon, and F. Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *ICML*, pages 1026–1034. JMLR, 2014.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov. Datasets of the causation and prediction challenge. Technical report, 2008.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *arXiv:1205.2536*, 2012.

Sara Van de Geer and Peter Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.