Nonlinear Causal Discovery with Latent Confounders

David Kaltenpoth [1] Jilles Vreeken [1]

Abstract

Causal discovery, the task of discovering the causal graph over a set of observed variables $X_1, \ldots, X_m$, is a challenging problem. One of the cornerstone assumptions is that of causal sufficiency: that *all* common causes of *all* measured variables have been observed. When it does not hold, causal discovery algorithms making this assumption return networks with many spurious edges. In this paper, we propose a nonlinear causal model involving hidden confounders. We show that it is identifiable from only the observed data and propose an efficient method for recovering this causal model. At the heart of our approach is a variational autoencoder which parametrizes both the causal interactions between observed variables as well as the influence of the unobserved confounders. Empirically we show that it outperforms other state-of-the-art methods for causal discovery under latent confounding on synthetic and real-world data.

1. Introduction

Causal models are at the heart of science (Pearl, 2009). They are robust as they explain away spurious associations, interpretable as we can see what causes what, and actionable as they allow us to estimate effects. Furthermore, they generalize across different, often novel, environments, making them suitable for many machine learning tasks (Pearl, 2009). However, deriving causal models is challenging. Not only is structure learning NP-hard (Chickering et al., 2004), it is in fact *impossible* to learn a causal model from data without making further assumptions on the generating process (Pearl, 2009). One of the cornerstone assumptions in causal discovery is that of *causal sufficiency*, the assumption that we have measured *all* common causes of *all* variables in our data. Whenever this assumption holds, all correlations between the observed variables can be explained. When causal sufficiency does not hold, however, methods that rely on it will not be able to adequately explain the observed correlations and hence return a causal network with many spurious edges.

Because it is impossible to measure all relevant variables in most applications, detecting and adjusting causal effects for such latent confounders has become an important research theme (Ogarrio et al., 2016; Wang & Blei, 2019; Janzing & Schölkopf, 2018; Kaltenpoth & Vreeken, 2019; Ranganath & Perotte, 2018). Existing work, however, can only do so in restricted settings, e.g., for two high-dimensional variables (Janzing & Schölkopf, 2018) or when additional information is available, such as data from multiple environments, labels, time, or proxies of the latent confounders (Khemakhem et al., 2020).

In contrast, in this paper, we propose a new method for learning a causal directed acyclic graph (DAG) from observational data in the presence of latent confounders. Given data over variables $X_1, \ldots, X_m$, our goal is to determine which X_i share a common cause Z, and return a causal network over both observed variables and hidden confounders.

To do so, we define a nonlinear structural causal model (SCM) with hidden confounders based on the post-nonlinear (PNL) model (Zhang & Hyvärinen, 2010; Zhang & Hyvarinen, 2012). We formally show that our model is identifiable in both the linear and in the strictly nonlinear case when the causal graph over X is sparse.

For discovery, we leverage recent advances in automatic differentiation (Abadi et al., 2016; Baydin et al., 2018) and present a novel method for causal discovery with hidden confounding based on variational autoencoders (Kingma & Welling, 2019). Specifically, the decoder is given by our structural causal model. To ensure that we obtain an acyclic causal network, we use a differentiable penalty based on counting the number of weighted cycles in the graph (Zheng et al., 2018; Yu et al., 2019).

With our approach for Nonlinear Causal Discovery under Latent Confounding, NoCaDiLaC for short, we efficiently obtain a fully directed network over both the observed variables X and their latent confounders Z. We

[1] CISPA Helmholtz Center for Information Security, Germany. Correspondence to: David Kaltenpoth <david.kaltenpoth@cispa.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

show that our method is theoretically sound by proving that it is consistent in a special case. We show through extensive experimental evaluation that our proposed method is not only theoretically sound but also does well in practice. On synthetic data, we significantly outperform a number of other state of the art methods for causal discovery with and without assumptions on causal sufficiency – GFCI (Ogarrio et al., 2016), NoTears (Zheng et al., 2018), DAG-GNN (Yu et al., 2019), 3off2 (Affeldt et al., 2016) and DCD (Bhattacharya et al., 2021). We further illustrate the importance of explicitly modeling latent confounders by comparing NoCaDiLaC with DAG-GNN and DCD on highly nonlinear real-world protein signaling data.

We include full proofs of all results in Appendix B, but sketch the main ideas here. All code and results can be found on the authors' website.[1]

2. Preliminaries

This section introduces notation, defines the problem at hand, and introduces our causal model.

2.1. Notation

We denote observed variables by $X = (X_1, \dots, X_m)^\top$ and unobserved variables by $Z = (Z_1, \dots, Z_l)^\top$, which are governed by a joint distribution $P(X, Z)$. Noise variables are denoted by ϵ, and we write ϵ_x and ϵ_z for noise variables affecting X, respectively Z. We assume that our noise follows a normal distribution $\epsilon \sim N(\mu, \mathrm{diag}(\sigma^2)I)$ where σ^2 contains the variances of the exogenous variables.

We write directed acyclic graphs (DAGs) as $G = (V, E)$ where the nodes V are comprised of the observed variables X and (a subset of) the unobserved Z. The set of all parents of Y in G is given by $\mathrm{Pa}_G(Y)$. We assume that the distribution $P(X, Z)$ corresponds to a graph G satisfying faithfulness, sufficiency, and causal Markov conditions (Koller & Friedman, 2009). We also refer to the parents of Y as the causes of Y, and write $Y_1 \to Y_2$ when Y_1 causes Y_2.

We denote nonlinear functions by τ, ν and assume that they work elementwise, $\tau(y) = (\tau_1(y_1), \dots, \tau_m(y_m))$, and that all nonlinearities τ_i, ν_i are three times differentiable, invertible and consequently strictly monotonic. We write id for the identity function, $\mathrm{id}(y) = y$. We denote matrices by capital letters A, B, C, their transposes by $A^\top, B^\top, C^\top$, and use I to refer to the identity matrix. The entries of the matrix A are a_{ij}. We further call an adjacency matrix A acyclic if the DAG G induced by it is acyclic.

We can now informally state our problem as follows.

Problem Statement (Informal). *Given data only over the observed variables X, recover the causal model over both X and its unobserved confounders Z.*

To formalize this problem, we next describe the structural causal models we are considering.

2.2. Structural Causal Models

To describe our data-generating process, we begin by introducing structural causal models (SCMs, Pearl (2009)) for the observed variables X when no latent confounders Z are involved. If all causal relationships are linear, we can write the model as a standard linear SCM,

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

where each ϵ_i is independent of the parents $\mathrm{Pa}_G(X_i)$ of X_i. Since the non-zero entries of the matrix A encode a DAG, there are no paths of length $m+1$ in G, so that the matrix satisfies $A^{m+1} = 0$. In particular, the matrix $C^\top := (I - A^\top)^{-1} = \sum_{k=0}^{m} A^k$ exists. By rearranging terms, we can therefore equivalently write

$$X = C^\top \epsilon \, .$$

To generalize this, we now include element-wise post-nonlinearities τ in the causal model (Taleb & Jutten, 1999),

$$X = \tau(C^\top \epsilon) \, . \tag{2}$$

As we assume the τ to be invertible, it is straightforward to show this model is equivalent to the well-studied post-nonlinear (PNL) causal model (Zhang & Hyvärinen, 2010), and refer the reader to Appendix A. The PNL model is suitable when sensors or measurements introduce nonlinear distortions in the observed variables. Moreover, this class of nonlinear causal models is known to permit identifiability of the causal DAG under general conditions (Zhang & Hyvarinen, 2012; Peters et al., 2014). These models are therefore appropriate for our setting both due to their ease of computability (Yu et al., 2019) as well as to facilitate theoretical tractability in the following.

Our next step in formalizing the problem is introducing latent factors in the structural causal model.

2.3. Confounders in SCMs

To describe the effects of latent confounders, we replace X by $(X, Z)^\top$ in the model of Eq. (1),

$$\begin{pmatrix} X \\ Z \end{pmatrix} = \begin{pmatrix} A_{XX}^\top & A_{XZ}^\top \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ Z \end{pmatrix} + \begin{pmatrix} \epsilon_x \\ \epsilon_z \end{pmatrix} ,$$

where we assume the matrices $A_{ZX}, A_{ZZ} = 0$ so that all $\mathrm{Pa}(Z_j) = \emptyset$. Note that for Gaussian variables Z, this assumption is a natural one. If $Z_1 \to Z_2$ and Z_2 affects precisely two variables X_1, X_2, then replacing the

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

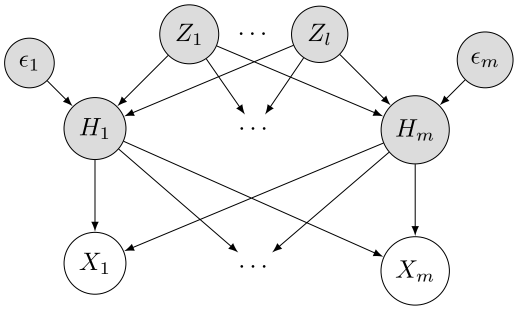


Figure 1: Graphical representation of the generating model in Eq. (4). The variables Z and ϵ generate the correlated noise $H = B^\top Z + \epsilon$, which then generates $X = \tau(C^\top H)$. Shaded nodes are unobserved ϵ, H are unobserved, while unshaded nodes X are observed.

edge $Z_1 \to Z_2$ with direct edges $Z_1 \to X_1, X_2$ and the correct parameters results in an indistinguishable distribution $P(X)$. We will elaborate further on the need for this asumption in Section 3.1. We can therefore focus on the generating mechanism of X, which we write as

$$X = A^\top X + B^\top Z + \epsilon\,, \tag{3}$$

where $A = A_{XX}, B = A_{XZ}$ and $\epsilon = \epsilon_x$. Equivalently, for $C^\top = (I - A^\top)^{-1}$ as above we have

$$X = C^\top (B^\top Z + \epsilon)\,.$$

Adding latent confounders thus replaces the uncorrelated noise ϵ with correlated noise terms $B^\top Z + \epsilon$. It is in this spirit that we generalize our nonlinear model of Eq. (2) as

$$X = \tau(C^\top (B^\top Z + \epsilon))\,, \tag{4}$$

or equivalently $X = \nu(X) + B^\top Z + \epsilon$ where $\nu = \mathrm{id} - \left(\tau \circ C^\top\right)^{-1}$. We graphically depict the data-generating model in Figure 1. Here, $H = B^\top Z + \epsilon$ represents the correlated noise obtained as a mixture of Z and ϵ. Having introduced our causal model with latent confounders, we can state our problem formally.

Problem Statement (Formal). *Given data from*

$$\begin{aligned} X &= \tau(C^\top (B^\top Z + \epsilon_x)) \\ Z &\sim N(0, \mathrm{diag}(\sigma_z^2) I) \\ \epsilon &\sim N(0, \mathrm{diag}(\sigma_x^2) I)\,, \end{aligned}$$

recover matrices A, B capturing the dependencies between the observed variables (up to trivial indeterminacies).

Here, by trivial indeterminacy we mean the following. We can write our problem as an overcomplete independent component analysis (OICA) problem (Comon, 1992),

$$X = \tau\left(\begin{pmatrix} C^\top B^\top & C^\top \end{pmatrix} \begin{pmatrix} Z \\ \epsilon \end{pmatrix}\right)$$

and by writing $Q = \begin{pmatrix} C^\top B^\top & C^\top \end{pmatrix}$ we know that Q can only be identified up to permution and scalar multiplication of its columns. That is, any matrix $Q' = Q \Pi \Lambda$ can induce the same distribution $P(X)$ under rescaling of the noise variables ϵ, where Π is a permutation and Λ a diagonal matrix (Eriksson & Koivunen, 2004). Consequently, when we talk about uniqueness or identifiability of any of the parameters below, we will be talking about uniqueness modulo such trivial indeterminacies.

Since our exogenous noise variables ϵ are all Gaussian, additional *non*-trivial indeterminacies are possible (Eriksson & Koivunen, 2004; Hyvarinen et al., 2019). We therefore study conditions under which our causal model becomes identifiable in the next section.

3. Theory and Methodology

In this section, we first give identifiability guarantees for our causal model. We then describe our method for recovering the causal model over the observed X and unobserved Z and show that it is consistent in the linear setting.

3.1. Identifiability

In general, without further assumptions, our causal model is not identifiable. To see this, consider the model from Eq. (3) with $X = (X_1, \dots, X_4)^\top$ and one-dimensional Z

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

Without further assumptions on A or B, the model has many equivalent parametrizations.

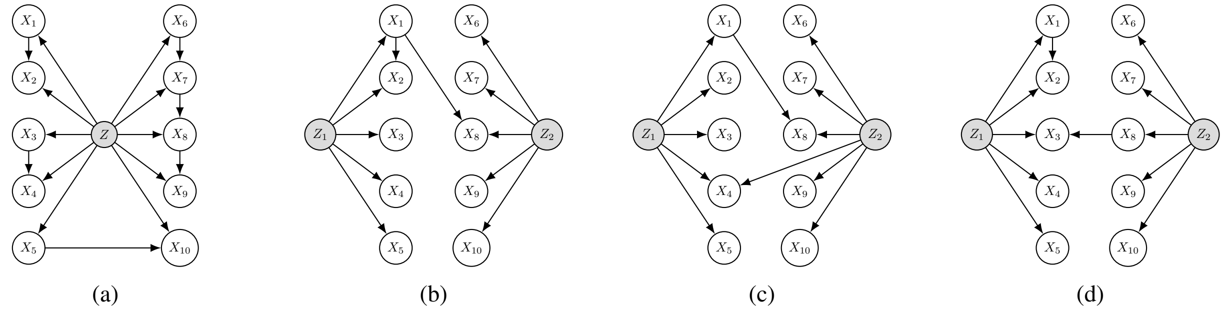
To give the simplest example in which we can prove uniqueness of the parameters, consider the model $X_i = b_i Z + \epsilon_i$. Then no X_i has a causal influence on any other X_j, but all pairs X_i, X_j are correlated with covariances

$$\sigma_{ij} := \mathrm{cov}(X_i, X_j) = b_i b_j\,.$$

If we try to fit a causal graph over only the variables X, we will find all variables to be connected. That is, the causal network over X would form a complete network (Elidan et al., 2000). However, the pairwise correlations satisfy the following constraint for all distinct quadruples i, j, u, v (Silva et al., 2006),

$$\sigma_{ij}\sigma_{uv} = \sigma_{iu}\sigma_{jv}\,,$$

so that the six pairwise correlations lie on a four-dimensional manifold. The inferred causal mechanisms are therefore correlated with each other, violating the principle of independent mechanisms (Janzing & Schölkopf, 2010). To achieve identifiability of our causal model in a more general setting, we make the following assumptions.



(a) (b) (c) (d)

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) As in (b), but this time Z_2 affects one of the nodes in S_1. (d) This graph violates the assumptions because S_1 has *too many* additional edges incoming.

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 the direct causal effect b_{ij} of $Z_j \to X_i$ is non-zero, $b_{ij} \neq 0$.*

As in the example, 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.

Assumption 3. *For all distinct X_i, X_j, X_u, X_v that are not conditionally independent given Z, and their covariances σ_{rs}, we have $\sigma_{ij}\sigma_{uv} \neq \sigma_{iu}\sigma_{jv}$.*

This assumption guarantees that the weights in A are not picked adversarially so as to make it seem as if the variables are purely confounded when they are not. It ensures that when we observe low-dimensional correlation structres that can be more readily explained by a latent confounder as in the example above, these correlations are *in fact* due to such a latent confounder. When the non-zero weights of A are sampled from a continuous probability distribution, this assumption holds with probability 1.

Assumption 4. *Either $\tau = id$ and $\epsilon_x \sim N(0, \sigma_x^2 I)$, or τ_i is strictly nonlinear for all i, $\frac{d^2}{dy^2}\tau_i(y) \neq 0$ for all y.*

While not necessary to determine the effects of the latent confounders Z, this assumption is necessary for the identification of the matrix A determining causal relationships between the observed variables X. The first case corresponds to the well-known identifiability of causal models for linear Gaussian models with equal noise variances (Peters & Bühlmann, 2012). The second case corresponds to the identifiablity of PNL models (Peters et al., 2014).

With these assumptions, we can now state our main result.

Theorem 1 (Identifiability)**.** *Let the distribution $P(X, Z)$ be generated by the model* (4) *and let assumptions 1–4 hold. Then the matrices B, C (and therefore A) are identifiable up to trivial indeterminacies.*

Proof Sketch. Since all nonlinearities τ_i, τ_j are invertible, the mutual information terms $I(X_i, X_j)$ can be correlated similar to the covariances in the example. Further, due to sparsity, for each Z_j, each of its children X_i is part of a quadruple X_i, X_u, X_v, X_w permitting b_{ij} to be inferred. Once the effect of Z on X is known, criteria for PNL models or equal variance Gaussian models can be used, depending on τ, to infer the remaining network over the observed X (Peters et al., 2014; Peters & Bühlmann, 2012). □

Note that we do not need to know the sets S_j to determine B. Instead, the tetrad constraints fully determine the sets S_j up to permutations of its indices. If we assume that Z and τ are normalized and we have some domain knowledge, we can further get rid of the rescaling indeterminacy.

Corollary 2. *Let the assumptions of Theorem 1 hold. Further let all τ_i be increasing and satisfy $\tau_i(1) = 1$, let*

$Z_j \sim N(0,1)$ and for each j let the sign of b_{ij} be known for at least one $X_i \in S_j$. Then B is identifiable up to permutations of its columns.

Next, we show that the model is identifiable with probability tending to one for large numbers of observed variables, even when the causal network is much less sparse than required by assumption (A3).

Theorem 3 (Identifiability for Large Dense Graphs)**.** *Let assumptions 3 and 4 hold and let the true causal graph G over X, Z be sampled from a directed Erdős-Rényi model $ER(m + l, p)$ satisfying assumption 1, with m observed and l unobserved nodes and edge probability $p < 1$. Then*

$$\lim_{m\to\infty} P(A, B \textit{ identifiable}) = 1\,,$$

where the limit is over graphs with fixed topological order.

Proof Sketch. When $p < 1$, for any Z_j with child X_i, for sufficiently large m we are guaranteed to find a suitable quadruple X_i, X_u, X_v, X_w to estimate b_{ij}. Given all b_{ij}, the entries of A corresponding to incoming edges into X_i, are identifiable for the same reason as in Theorem 1. □

Note that we are not interested in the identifiability of the nonlinearities τ_i (Zhang & Hyvärinen, 2010; Zhang & Hyvarinen, 2012). Instead we are interested in discovering the underlying generating DAG G as well as the effects of the latent confounder Z, this is not an issue for our purposes. As long as we can find *any* element-wise nonlinearity ν such that $\nu(X) \sim N(0, \Sigma)$ for some Σ, we have achieved our goal. Next, we develop a method that achieves this task.

3.2. Learning with Autoencoders

In order to learn a nonlinear function ν such that $\nu(X)$ is normally distributed, we make use of variational autoencoders (VAEs) (Kingma & Welling, 2019). The goal of a VAE is to optimize the evidence lower bound (ELBO, Kingma & Welling (2019)) defined as

$$\begin{aligned} L_{\text{ELBO}} &\coloneqq E_{q(H|X)}\left(\log p(X \mid H)\right) \\ &\quad - D\left(q(H \mid X) \mid p(H)\right) \\ &\leq \log p(X) \end{aligned}$$

where D is the Kullback-Leibler divergence and $q(H \mid X)$ is a distribution to be optimized over. That is, we maximize a lower bound on the evidence $\log p(X)$

$$\max_{q(H|X),\, p(X|H)} L_{\text{ELBO}} \leq \log p(X)$$

over some class of distributions $q(H \mid X)$, called encoders, and $p(X \mid H)$, called decoders. Given a set of variables H, we can use Eq. (4) as decoder and write $X = \tau(C^\top H)$ where H takes on the role of $B^\top Z + \epsilon$. We can write the corresponding encoder as

$$H \sim N((I - A^\top)\tau^{-1}(X), \sigma_z^2(X)B^\top B + \sigma_\epsilon^2(X)I)\,.$$

Equivalently, we can split H into the effect of the latent Z and the independent noise ϵ as

$$\begin{aligned} H &= B^\top Z + \epsilon \quad \text{where} \\ Z &\sim N(\mu_z(X), \sigma_z^2(X)I) \\ \epsilon &\sim N(\mu_\epsilon(X), \sigma_\epsilon^2(X)I) \\ \text{s.t.} \quad B^\top \mu_z + \mu_\epsilon &= \left(I - A^\top\right)\tau^{-1}(X)\,. \end{aligned}$$

We show the full model in graphical form in Fig. 3.

In practice, to allow the nonlinearity τ to be learned jointly with its inverse τ^{-1}, it needs to have a closed-form solution for its inverse and allow for ease of gradient computation in the shared parameters θ. Due to its piece-wise linearity, a stack of multiple PReLU functions applied to each variable is therefore a good candidate. We define

$$\begin{aligned} \tau_i(x_i; \theta_i) &= \phi(\lambda_{i2}\phi(\lambda_{i1}x + \beta_{i1}; \alpha_{i1}) + \beta_{i2}; \alpha_{i2}) \\ \text{and} \quad \tau(x; \theta) &= (\tau_1(x_1; \theta_1), \ldots, \tau_m(x_m; \theta_m)) \end{aligned}$$

where $\phi(y; \gamma) = [y]_+ - \gamma[y]_-$ and θ is the vector containing all parameters $\{\lambda_{ij}, \alpha_{ij}, \beta_{ij}\}_{ij}$ (He et al., 2015).

To ensure that our learned model has a causal interpretation, we require the discovered adjacency matrix A to be acyclic. To this end, we use a differentiable penalty $h(A)$ proposed by (Zheng et al., 2018; Yu et al., 2019)

$$\begin{aligned} h(A) &\coloneqq \text{tr}\left((I + A \odot A/m)^m\right) - m \\ &= \text{trace}(I) + \sum_{i=1}^{m} \binom{m}{i} \text{trace}((A \odot A/m)^i) - m \end{aligned}$$

which counts the weighted number of loops of each length i in the graph G with adjacency matrix A. Here $A \odot A$ denotes the Hadamard product, $(A \odot A)_{ij} = A_{ij}^2$. Therefore $h(A) = 0$ if and only if the graph described by the adjacency matrix A is acyclic.

In general, however, we not only care that A is acyclic, we are particularly interested in finding *sparse* matrices A and B encoding simple and interpretable causal networks. Under such a sparsity constraint, our approach is consistent.

Theorem 4 (Consistency under Sparsity)**.** *Let x^n be a sample generated from the model in Eq.* (4) *with $\tau = id$ and let assumptions 1–4 hold. Let L be the score*

$$L(x^n; A, B) \coloneqq -L_{\textit{ELBO}} + \lambda_A \|A\|_0 + \lambda_B \|B\|_0\,,$$

and let $\widehat{A}, \widehat{B}$ be its maximizers subject to acyclicity,

$$\begin{aligned} \widehat{A}, \widehat{B} &= \arg\max_{A,B} L(x^n; A, B) \\ \textit{s.t.}\ \ h(A) &= 0\,. \end{aligned}$$

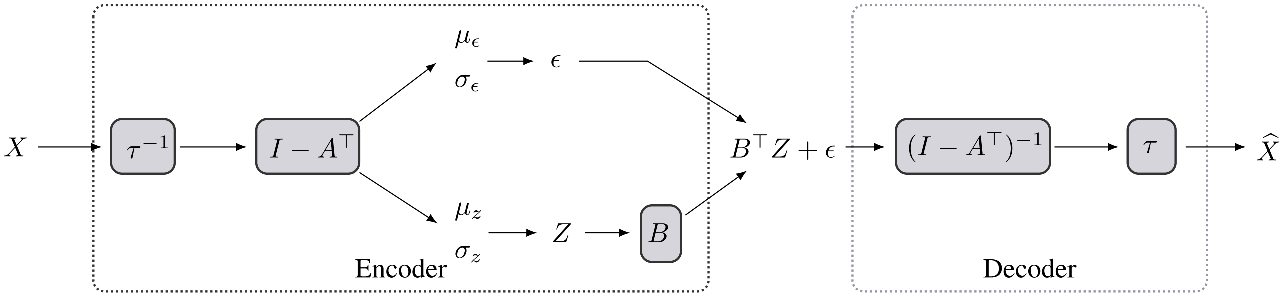


Figure 3: Proposed model architecture. The encoder outputs both noise ϵ and confounder Z. Learnable parameters include the adjacency matrix A of the causal model over the observed variables as well as the matrix B containing the influence of Z on the observed nodes.

Then for small enough $\sigma_\epsilon(X), \sigma_z(X)$ the score L is consistent for recovering the matrices A, B when $\lambda_A = \lambda_B = \log(n)/2$:

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

Proof Sketch. When $\tau = \mathrm{id}$ and $\sigma_z, \sigma_\epsilon = 0$, L_{ELBO} can be written as $\frac{1}{2}\left\|(I - A^\top)X\right\|_F^2$ for where X is the data matrix for x^n. As before, B can be discovered by finding appropriate quadruples, and the network can be approximated by using consistency results for linear Gaussian models with equal variance (Van de Geer & Bühlmann, 2013). □

This result applies only to the linear case, for which the identifiability of overcomplete ICA has been established (Eriksson & Koivunen, 2004). In contrast, for nonlinear ICA, even the *non*-overcomplete case is riddled with difficulties, requiring additional assumptions of specific kinds of sparsity, independence of mechanisms, or auxiliary information (Zheng et al., 2022; Gresele et al., 2021; Khemakhem et al., 2020). To the best of our knowledge, these approaches do not (easily) extend to overcomplete nonlinear ICA, preventing us from providing any better consistency results at this point in time.

In practice, the regularizers $\|\cdot\|_0$ are not differentiable, so we use the common practice of replacing them with L_1 norms $\|\cdot\|_1$. The overall learning problem including the nonlinearities $\tau(\cdot;\theta)$ defined above is therefore

$$\min_{A,B,\theta} f(x^n; A, B, \theta) := -L_{\text{ELBO}} + \lambda_A \|A\|_1 + \lambda_B \|B\|_1 \quad \text{s.t. } h(A) = 0\,.$$

To obtain a fully differentiable optimization target, we use the augmented Lagrangian approach (Bertsekas, 1997)

$$L(A, B, \theta, \lambda) = f(A, B, \theta) + \lambda h(A) + \frac{\rho}{2}\left|h(A)\right|^2\,.$$

which can be solved using dual ascent (Bertsekas, 1997). We include further details on the computation of the terms as well as the optimization in Appendix C.

4. Related Work

Causal inference is arguably one of the most critical research areas in statistical inference and has recently gained much attention among researchers (Pearl, 2009). The existence of hidden confounders and selection bias makes it impossible, however, to infer causality from purely observational data without additional assumptions (Pearl, 2009). When causal sufficiency is assumed, both constraint-based (Spirtes et al., 2000; Zhang, 2008) and score-based (Chickering, 2002; Scanagatta et al., 2015; Ramsey et al., 2017) methods can discover causal networks up to Markov equivalence. These methods are, however, based on discrete optimization and do not benefit from recent advances in automatic differentiation (Abadi et al., 2016; Baydin et al., 2018).

To make use of these advances, Zheng et al. (2018) reformulate network inference as a continuous optimization problem by introducing a differentiable constraint measuring how many cycles a matrix contains. While initially designed for purely linear relationships, it has been generalized to permit nonlinear relationships (Zheng et al., 2020; Yu et al., 2019). However, none of the aforementioned methods can distinguish between networks inside the Markov equivalence class (MEC).

When additional assumptions are made, it becomes possible to distinguish between Markov equivalent networks. Based on the asymmetry between factorizations of the joint distribution in the causal and the anti-causal directions, it is possible to determine causal directions in both the bivariate (Zhang & Hyvarinen, 2012; Daniusis et al., 2012; Janzing et al., 2012) as well as multivariate cases (Peters & Bühlmann, 2012; Bühlmann et al., 2014; Mian et al., 2021).

In the presence of latent confounders, several methods, such as GFCI (Ogarrio et al., 2016; Colombo et al., 2012) or 3OFF2 (Affeldt et al., 2016) and convex optimization-based approaches (Chandrasekaran et al., 2010), can find causal networks in which undirected edges indicate latent confounding. Specifically, Nested Markov Models (NMMs) (Shpitser et al., 2014; 2018; Richardson et al., 2017; Evans & Richardson, 2019) can sometimes provide identifiability of causal models with latent factors by using Verna constraints. The recent approach DCD by Bhattacharya et al. (2021) combines NMMs with the differentiable constraint by Zheng et al. (2018) to discover a partially directed causal network indicating which nodes are likely confounded. However, all of these methods return MECs. In particular, they cannot tell which nodes share a confounder, making their results difficult to interpret.

To relieve this issue, several approaches determine whether sets of variables share a confounder. Janzing & Schölkopf (2018) use geometric properties of high-dimensional regression problems to assess whether variables are confounded, while Kaltenpoth & Vreeken (2019) use the minimum description length principle (Grünwald, 2007) to test whether the confounder improves the compression of the observed data. In contrast, Wang & Blei (2019) and Ranganath & Perotte (2018) explicitly model latent confounders using factor models to adjust causal estimates for their presence. However, none of these approaches infer causal networks in the presence of confounders.

5. Experiments

In this section, we evaluate our method NoCaDiLaC empirically. We are interested in how well it 1) recovers the set of nodes affected by Z and 2) recovers the entire causal network. We compare with other state-of-the-art methods for network discovery, including those that permit latent confounding, using GFCI (Ogarrio et al., 2016), 3off2 (Affeldt et al., 2016), and DCD (Bhattacharya et al., 2021), and those assuming causal sufficiency, NoTears (Zheng et al., 2018) and DAG-GNN (Yu et al., 2019).

We implemented NoCaDiLaC using Tensorflow (Abadi et al., 2016) and perform optimization using Adam (Kingma & Ba, 2014). For our competitors, we use the implementations provided by the authors. We make all code and results available in the supplement.

5.1. Synthetic Data

We first describe our data-generating process, followed by the metrics we use to evaluate our method. We then show the results of each competitor in multiple settings.

Data Generation

We begin by giving an overview of our data generation process. First, we use the Erdős-Rényi model to generate a random DAG with $p = 0.3$. We then generate a corresponding adjacency matrix with $A_{ij} \sim U[-1, 1]$ when $(i, j) \in G$. Our data generating model is given by $x' = A^\top g(x') + \epsilon$ where $g(x') = \mu + \alpha \odot f(x')$ where f is uniformly sampled from linear, quadratic, cubic, exponential, logistic or sinusoidal functions and $(\mu, \alpha) \sim U[-1, 1]^{2m}$ are i.i.d. The noise is $\epsilon \sim N(0, 1)$.

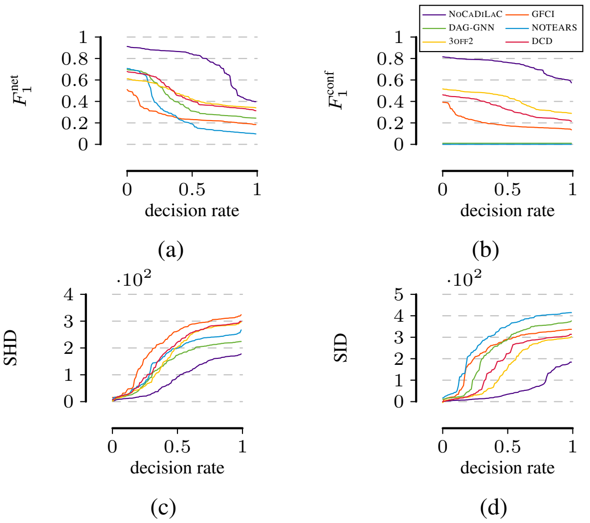


Figure 4: Evaluation on synthetic networks of size 25. We show F_1 scores for (a) network discovery and (b) confounded set recovery (higher is better), as well as (c) structural Hamming distance and (d) structural intervention distance (lower is better). Each figure shows the average score over increasing fractions of all datasets, sorted for each method from most confident to least. We see that NoCaDiLaC outperforms its competitors at all levels. In particular, in confounded set recovery (b), NoCaDiLaC performs better *at its worst* than its competitors *at their best.*

To generate the observed data x, we remove sample x' from the above model and remove l source nodes $z = (x'_{k_1}, \dots, x'_{k_l})$. In the interest of space, we here consider confounders z of dimension $l = 1$ but include analysis for varying l to Appendix E. We use sample size $n = 2500$ and repeat each experiment 500 times.

Evaluation Metrics

To evaluate NoCaDiLaC, we consider the following metrics chosen to capture different aspects of causal network discovery with confounders. To measure how well we recover the overall network, we use the Structural Hamming Distance (SHD) as well as the F_1 score for network discovery, which we denote F_1^{net}. As we are particularly concerned with discovering which nodes are confounded, we further consider the F_1 score for confounded node recovery, denoted F_1^{conf}, which measures how well we recover children of z. Since all of the above metrics measure only the structural similarity between the true and recovered networks, we also use the Structural Intervention Distance

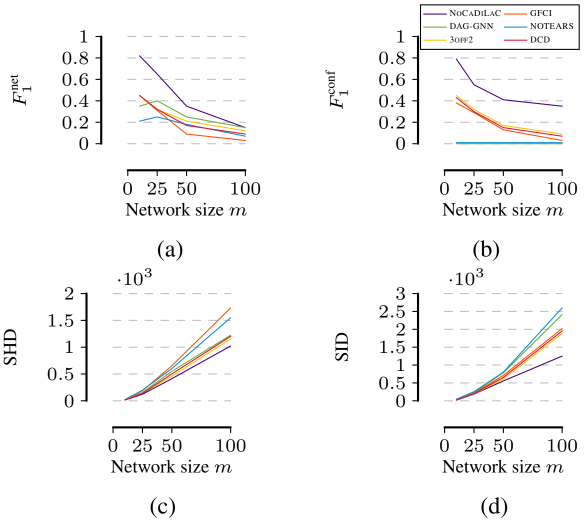


Figure 5: Evaluation on synthetic networks of different sizes. We show F_1 scores for (a) network discovery and (b) confounded set recovery (higher is better), as well as (c) structural Hamming distance and (d) structural intervention distance (lower is better).

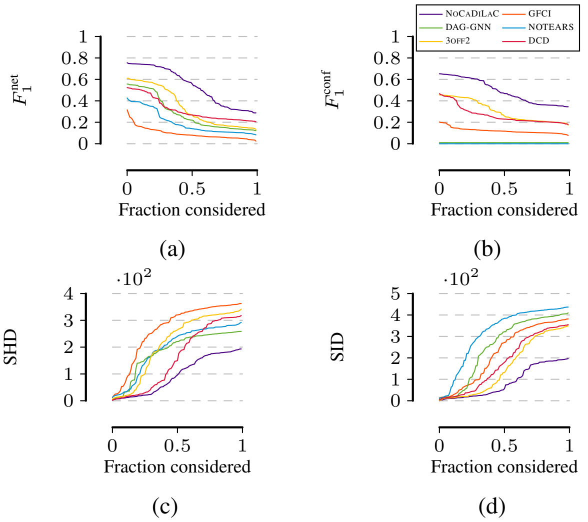


Figure 6: Evaluation on REGED. F_1 scores for (a) network discovery and (b) confounded set recovery (higher is better), as well as (c) structural Hamming distance and (d) structural intervention distance (lower is better).

(SID) (Peters & Bühlmann, 2013), which measures how many interventional distributions differ between the true and the recovered network.

To ensure that our evaluation also has practical significance, we do not only look at the average performance of each method over all generated datasets. Instead, we want a metric that is easy to compute solely from the observed and that we can use to predict the quality of our discovered causal model. To do so, we consider the relationship between the confidence, defined as the improvement of the discovered model over the null model, and the performance measured by the above metrics. Formally, we define the confidence of a method as the relative gain in its score L over the null model

$$C = \frac{L(X;0) - L(X;\theta^*)}{L(X;0)}\,, \tag{5}$$

where $L(X;\theta^*)$ is the score optimized by the method and $L(X;0)$ corresponds to an empty model. We are therefore interested in the relationship between C and the metrics above: a good method should obtain better scores where it is confident than where it is not.

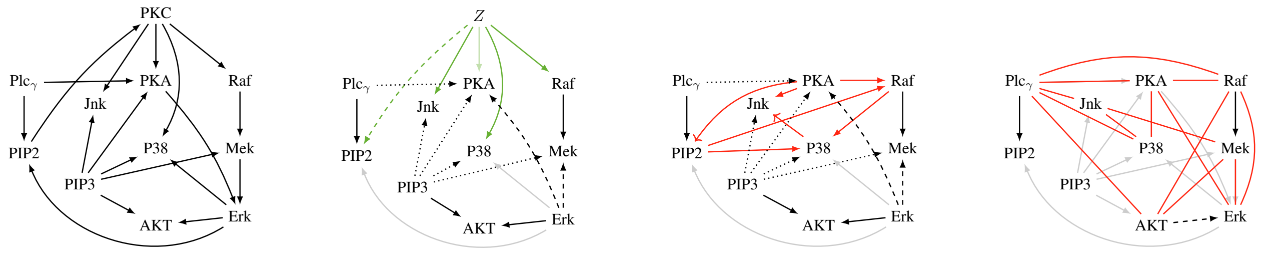
Confidence and Performance

To evaluate the relationship between the confidence C and each used metric, for each competitor we order the generated datasets and recovered networks in descending order by their obtained confidence C. For NoCaDiLaC, DAG-GNN, NoTears, and DCD, we can use the definition of Eq. (5). However, since GFCI and 3off2 do not optimize a score, we order their decisions in the way *most favorable* to them. We include more details in Appendix D.

To show the relationship between confidences and each respective metric, we create *decision rate plots*. They show the average performance over the top $k\%$ datasets, sorted from most to least confident. We show these decision rate plots in Fig. 4. For NoCaDiLaC, all metrics correlate strongly with C, suggesting that it can be used to determine which network discoveries are more reliable than others. For NoTears and DAG-GNN, confidence correlates with F_1^{net}, SID and SHD, but since these methods are not able to find confounders, their F_1^{conf} score is zero throughout. Overall NoCaDiLaC outperforms its competitors by a large margin for all metrics.

Performance for different network sizes

To see how well NoCaDiLaC performs for larger networks, we next test all methods for networks of sizes $m \in \{10, 25, 50, 100\}$. We show the results in Fig. 5. For all metrics, NoCaDiLaC outperforms its competitors by a large margin. While all methods show decreasing performance for increasing network size m, NoCaDiLaC is most robust against increasing network sizes. As the network size increases, the gap becomes smaller for the F_1 scores. This is due primarily to the latent confounder affecting only a smaller fraction of variables as we increase the network size. Meanwhile, especially for SID, the gap becomes more prominent as every incorrect edge between nodes simultaneously affects many other pairs of nodes.



(a) Consensus Network (b) NoCaDiLaC (c) DAG-GNN (d) DCD

Figure 7: Results on the Sachs dataset. NoCaDiLaC discovers a confounder Z capturing the influence of PKC (green edges). In contrast, DAG-GNN finds many edges between nodes influenced by PKC (red) and DCD contains indications of confounding for many pairs of nodes, but neither method can determine that nodes share a confounder. All methods make roughly the same number of errors regarding reversed edges (dashed), including some edges only as indirect paths (dotted), and missing edges (gray).

5.2. REGED Benchmark Data

To evaluate NoCaDiLaC on data that violates our assumptions, we evaluate it on the REGED dataset (Guyon et al., 2008). This dataset contains "re-simulated" created by matching the distribution of real microarray gene expression data. It contains $m = 999$ features and $n = 20\,500$ non-i.i.d. samples. To introduce confounders, we cannot use the whole network to evaluate NoCaDiLaC. Instead, we generate subgraphs of different sizes containing nodes with common parents, whose data we hide from our methods. We include the full details in Appendix F.

We show decision rate plots for all metrics in Fig. 6. While all methods perform worse than on synthetic data, NoCaDiLaC nevertheless performs best by a large margin.

5.3. Case Study: Protein Signaling Network

Last, to see how well NoCaDiLaC works on real-world data, we evaluate it on the widely used Sachs dataset (Sachs et al., 2005) for protein signaling. It contains $n = 7466$ continuous measurements of a total of $m = 11$ phosphorylated proteins and phospholipids in human immune system cells. The consensus network contains 20 edges, which we show in Fig. 7a. Since the graph contains cycles, some mistakes are inevitable. To make the data appropriate for our setting, we remove the node PKC with out-degree four from the network. Note that the edge from PIP2 to PKC violates our assumption that latent confounders have no incoming edges. We show the benefit of explicitly modeling confounders by comparing against DAG-GNN and DCD.

We show the results of this experiment in Fig. 7. In Fig. 7b, we see that NoCaDiLaC automatically discovers a substitute latent factor Z connected to the correct variables (green edges) and thereby takes the place of PKC. For the overall network, only three edges are missing entirely (gray), while two are reversed (dashed), and another three are instead contained as paths of length 2 (dotted). This performance is similar to the result of other state-of-the-art methods on *fully* observed data (Yu et al., 2019). In Fig. 7c, we see the result of DAG-GNN. The absence of PKC from the observed data leads to DAG-GNN inferring a large number of edges (red) between nodes initially connected to PKC while making more mistakes over the remaining variables. We see a similar pattern in the results of DCD in Fig. 7d. Many pairs of children of PKC are considered to be potentially confounded (red), but so are many other pairs of variables that do *not* have PKC as parent. It is also unclear which pairs share the *same* latent parent. Furthermore, DCD misses many more edges than either NoCaDiLaC or DAG-GNN. Overall, NoCaDiLaC produces more accurate and interpretable results than its competitors.

6. Conclusion

In this paper, we proposed a method for discovering a nonlinear causal model in the presence of latent confounders from observational data. We proved the identifiability of our causal model and proposed an effective approach to learn the parameters based on VAEs. We showed the theoretical soundness of our method by proving that the L_{ELBO} with sparsity and acyclicity constraints forms a consistent scoring criterion when the generating model is linear.

Empirical evaluation showed that NoCaDiLaC works well not only on linearly but also on nonlinearly generated data. On real-world data, we saw that compared to its competitors, it produces both quantitatively better results on all metrics as well as qualitatively more interpretable models.

For future work, an interesting avenue will be the combination of recent theoretical insights into nonlinear ICA with our present results to obtain better theoretical guarantees.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A System for Large-Scale Machine Learning. In *OSDI16*, pp. 265–283, 2016.

Affeldt, S., Verny, L., and Isambert, H. 3off2: A Network Reconstruction Algorithm Based on 2-point and 3-point Information Statistics. In *BMC bioinformatics*, volume 17, pp. 149–165. BioMed Central, 2016.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. Automatic Differentiation in Machine Learning: A Survey. *Journal of Marchine Learning Research*, 18:1–43, 2018.

Bertsekas, D. P. *Nonlinear Programming*. Athena Scientific, 1997.

Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. Differentiable Causal Discovery Under Unmeasured Confounding. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, 2021.

Bühlmann, P., Peters, J., Ernest, J., et al. CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression. *The Annals of Statistics*, 42: 2526–2556, 2014.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Latent Variable Graphical Model Selection Via Convex Optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1610–1613. IEEE, 2010.

Chickering, D. M. Learning Equivalence Classes of Bayesian-Network Structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

Chickering, M., Heckerman, D., and Meek, C. Large-Sample Learning of Bayesian Networks Is NP-Hard. *Journal of Machine Learning Research*, 5, 2004.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. Learning High-Dimensional Directed Acyclic Graphs With Latent and Selection Variables. *The Annals of Statistics*, pp. 294–321, 2012.

Comon, P. Independent Component Analysis. 1992.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and Schölkopf, B. Inferring Deterministic Causal Relations. *arXiv preprint arXiv:1203.3475*, 2012.

Elidan, G., Lotner, N., Friedman, N., and Koller, D. Discovering Hidden Variables: A Structure-Based Approach. *Advances in Neural Information Processing Systems*, 13, 2000.

Eriksson, J. and Koivunen, V. Identifiability, Separability, and Uniqueness of Linear ICA Models. *IEEE signal processing letters*, 11(7):601–604, 2004.

Evans, R. J. and Richardson, T. S. Smooth, Identifiable Supermodels of Discrete Dag Models With Latent Variables. *Bernoulli*, 25(2):848–876, 2019.

Gresele, L., Von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. Independent Mechanism Analysis, a New Concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

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

Guyon, I., Aliferis, C. F., Cooper, G. F., Elisseeff, A., Pellet, J., Spirtes, P., and Statnikov, A. R. Design and Analysis of the Causation and Prediction Challenge. In *Causation and Prediction Challenge at WCCI 2008, Hong Kong, June 1-6, 2008*, volume 3 of *JMLR Proceedings*, pp. 1–33. JMLR.org, 2008.

He, K., Zhang, X., Ren, S., and Sun, J. Delving Deep Into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Hyvarinen, A., Sasaki, H., and Turner, R. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Janzing, D. and Schölkopf, B. Causal Inference Using the Algorithmic Markov Condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Janzing, D. and Schölkopf, B. Detecting Non-Causal Artifacts in Multivariate Linear Regression Models. In *International Conference on Machine Learning*, pp. 2245–2253. PMLR, 2018.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-Geometric Approach To Inferring Causal Directions. *Artificial Intelligence*, 182:1–31, 2012.

Kaltenpoth, D. and Vreeken, J. 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.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. An Introduction To Variational Autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

Mian, O. A., Marx, A., and Vreeken, J. Discovering Fully Oriented Causal Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8975–8982, 2021.

Ogarrio, J. M., Spirtes, P., and Ramsey, J. A Hybrid Causal Search Algorithm for Latent Variable Models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pp. 368–379, 2016.

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

Peters, J. and Bühlmann, P. Identifiability of Gaussian Structural Equation Models With Equal Error Variances. *arXiv preprint arXiv:1205.2536*, 2012.

Peters, J. and Bühlmann, P. Structural Intervention Distance (SID) for Evaluating Causal Graphs. *arXiv preprint arXiv:1306.1043*, 2013.

Peters, J., Mooij, J. M., Janzing, D., Schölkopf, B., et al. Causal Discovery With Continuous Additive Noise Models. *Journal of Machine Learning Research*, 15: 2009–2053, 2014.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. A Million Variables and More: The Fast Greedy Equivalence Search Algorithm for Learning High-Dimensional Graphical Causal Models, With an Application To Functional Magnetic Resonance Images. *International journal of data science and analytics*, 3: 121–129, 2017.

Ranganath, R. and Perotte, A. Multiple Causal Inference With Latent Confounding. *arXiv preprint arXiv:1805.08273*, 2018.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. Nested Markov Properties for Acyclic Directed Mixed Graphs. *arXiv preprint arXiv:1701.06686*, 2017.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal Protein-Signaling Networks Derived From Multiparameter Single-Cell Data. *Science*, 308(5721):523–529, 2005.

Scanagatta, M., de Campos, C. P., Corani, G., and Zaffalon, M. Learning Bayesian Networks With Thousands of Variables. In *NIPS*, pp. 1864–1872, 2015.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. Introduction To Nested Markov Models. *Behaviormetrika*, 41(1):3–39, 2014.

Shpitser, I., Evans, R. J., and Richardson, T. S. Acyclic Linear Sems Obey the Nested Markov Property. In *Uncertainty in Artificial Intelligence*, volume 2018. NIH Public Access, 2018.

Silva, R., Scheines, R., Glymour, C., Spirtes, P., and Chickering, D. M. Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research*, 7(2), 2006.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2000.

Taleb, A. and Jutten, C. Source Separation in Post-Nonlinear Mixtures. *IEEE Transactions on signal Processing*, 47(10):2807–2820, 1999.

Van de Geer, S. and Bühlmann, P. ℓ_0-penalized Maximum Likelihood for Sparse Directed Acyclic Graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Wang, Y. and Blei, D. M. The Blessings of Multiple Causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Yu, Y., Chen, J., Gao, T., and Yu, M. DAG-GNN: DAG Structure Learning With Graph Neural Networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019.

Zhang, J. On the Completeness of Orientation Rules for Causal Discovery in the Presence of Latent Confounders and Selection Bias. *Artificial Intelligence*, 172(16-17): 1873–1896, 2008.

Zhang, K. and Hyvärinen, A. Distinguishing Causes From Effects Using Nonlinear Acyclic Causal Models. In *Causality: Objectives and Assessment*, pp. 157–164. PMLR, 2010.

Zhang, K. and Hyvarinen, A. On the Identifiability of the Post-Nonlinear Causal Model. *arXiv preprint arXiv:1205.2599*, 2012.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags With No Tears: Continuous Optimization for Structure Learning. *Advances in neural information processing systems*, 31, 2018.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. P. Learning Sparse Nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Zheng, Y., Ng, I., and Zhang, K. On the Identifiability of Nonlinear ICA: Sparsity and Beyond. *arXiv preprint arXiv:2206.07751*, 2022.

A. Why our Model is a PNL Causal Model

In this section we quickly explain why our proposed model in Eq. 2 can be considered a post-nonlinear (PNL) causal model. Note first that the model as stated is a PNL ICA model (Taleb & Jutten, 1999). We now show that this can in fact be rewritten as PNL causal model (Zhang & Hyvarinen, 2012).

To do so, assume that the topological order is $X_1, \dots, X_m$. Next we show for all $k = 1, \dots, m$ that we can write

$$X_k = \tau_k \left(\sum_{j=1}^{k-1} \alpha_{jk} \tau_j^{-1}(X_j) + \epsilon_k \right) .$$

For $k = 1$ this is trivially true as the sum is empty. For the case $k > 1$, we have

$$\begin{aligned} X_k &= \tau_k \left(\sum_{j=1}^{k-1} c_{jk} \left(\tau_j^{-1}(X_j) - \sum_{l=1}^{j-1} \alpha_{jl} \tau_l^{-1}(X_l) \right) + \epsilon_k \right) \\ &= \tau_k \left(\sum_{j<k} c_{jk} \tau_j^{-1}(X_j) - \sum_{l<k} \left(\sum_{j:l<j<k} c_{jk} \alpha_{lj} \right) \tau_l^{-1}(X_l) + \epsilon_k \right) , \end{aligned}$$

where in the last line, we changed the order in which the dual sum is computed. To simplify this further, we need two things. First, the matrix $C = (I - A)^{-1} = \sum_{i=0}^{m-1} A^i$ has as entries c_{jk} precisely the sum of all path weights for paths $p : j \rightsquigarrow k$ of any length. Second, the inside sum of the second term above, $\sum_{j:l<j<k} c_{jk} \alpha_{lj}$, contains precisely the sum of all paths $p : l \rightsquigarrow k$ of length *at least* two. By swapping the indices j, l in the dual sum, we therefore obtain

$$X_k = \tau_k \left(\sum_{j<k} \left(c_{jk} - \sum_{l:j<l<k} c_{lk} \alpha_{jl} \right) \tau_j^{-1}(X_j) + \epsilon_k \right) ,$$

and the difference between the two weights in the inside parentheses is precisely the weight of the (at most one) path of length one, $X_j \to X_k$, which is precisely α_{jk}. In particular, when all τ_j are strictly nonlinear then the result of Corollary 31(ii) of Peters et al. (2014) guarantees that the causal model (without confounders) is identifiable. We explain how to deal with the inclusion of confounders in the proofs below.

B. Proofs

Proof of Theorem 1. Note first of all that for any variables X_i, X_j we have that the mutual information $I(X_i; X_j) = I(\tau_i^{-1}(X_i); \tau_j^{-1}(X_j))$. Hence, since $X = \tau(C^\top(B^\top Z + \epsilon))$, it suffices to study the fully linear Gaussian case. Note that for two Gaussian variables, we have

$$I(X_i; X_j) = -\frac{1}{2} \log(1 - \rho_{ij}^2)$$

where ρ_{ij} is the pair's correlation coefficient.

It therefore suffices to show that the parameters ρ_{ij} determine the matrices C and B (up to trivial indeterminacies) when assumptions 1-4 hold. To this end, note that the variances of each variable are observable and therefore ρ_{ij} and $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$ contain equal amounts of information.

Now, let Z_j be given. Then by assumption 2, each child X_i of Z_j has a quadruple (X_i, X_u, X_v, X_w) of variables $X_u, X_v, X_w \in S_j$ that is independent given Z_j. Hence, from the example in the text we know that

$$\sigma_{iu}\sigma_{vw} = \sigma_{iv}\sigma_{uw}$$

so that by assumption 4 and the example in the text, we know that

$$b_{ij}^2 = \lambda_j \sigma_{iu} \sigma_{iv} / \sigma_{uv},$$

i.e., that b_{ij} is identifiable up to the scalar multiple λ_j (as well as permutations of Z_j).

Given B, by assumption 4 the matrix C (and therefore A) is also identifiable. When τ_i is strictly nonlinear, this follows from Corollary 31(ii) of (Peters et al., 2014). When $\tau_i = \text{id}$ and all ϵ_x have equal variances, this follows from results on identifiability of linear Gaussian models with equal noise variances (Peters & Bühlmann, 2012). □

Proof of Corollary 2. Under the additional assumptions on normalization of τ_i and the variances of Z_j, the scalar multiple λ_j in the previous proof is limited to be $\lambda_j \in \{-1, 1\}$. Furthermore, if for each Z_j the sign of at least one b_{ij} is known, the direction of $b_{\cdot j}$ is fixed and B thus identifiable up to permutations. □

Proof of Theorem 2. To prove this result, we need to show that for any $p < 1$, for every variable X_i, we can find triples (X_u, X_v, X_w) such that (X_i, X_u, X_v, X_w) are conditionally independent given Z_j as in the previous proof by which to identify the value of b_{ij}.

We do this through a simple counting argument. Let $Y = X_i$ be fixed, and $U = X_u \in X$ be another variable. What is the probability that U is not admissible in a quadruple as above? There are two possibilities:

1. Either, U is ruled out by there existing an edge $X - U$ or an edge $X \leftarrow R \rightarrow U$.

2. Or, U is ruled out by being mutually connected to too many other nodes V, which are not ruled out by step 1.

We begin with the first case. The probability of this occurring is $r := p + p^2 - p^3 = p + (1-p)p^2$ where the last term is subtracted because otherwise, we would be counting the intersection twice. Note that for $p < 1$, we have $r < 1$. Next note that for any $r < 1$ and any $s < 1$ we have

$$r + (1-r)s < 1$$

Hence we need to show that the probability s of the second case above is < 1. This is, however, trivial since U can only be ruled out in this way if there is at least one edge to another variable V, i.e., it cannot have zero edges, such that the probability $s < 1$ for any $p < 1$.

Therefore, the probability of a node U being ruled out is strictly less than 1 so that in the limit $m \to \infty$, we are guaranteed to find at least one valid quadruple. □

Proof of Theorem 3. Let us write $\widehat{X}$ for our reconstruction of X. With $\sigma_\epsilon^2, \sigma_z^2 = 0$, a typical sample loss of our L_{ELBO} score is then given by (Yu et al., 2019)

$$-L_{\text{ELBO}} = \frac{1}{2}\sum \left(X_{ij} - \widehat{X}_{ij}\right)^2 + \frac{1}{2}\sum H_{ij}^2 \,.$$

Furthermore in the case $\tau = \text{id}$ the reconstruction is perfect, $\widehat{X} = X$ and $H = (I - A^\top)X$ so that the L_{ELBO} can be written as

$$-L_{\text{ELBO}} = \frac{1}{2}\left\|(I - A^\top)X\right\|_F^2 \,.$$

Using assumptions 1-4, as in the proof of Theorem 1, for each X_{ij}, there exists a distinct quadruple which is independent given Z_j, which permits us to explain the correlations between them by way of the parameters $b_{\cdot j}$ and since the variables cannot be rendered independent without reference to Z_j, any explanation of the same correlations involving only variables X_j would need six parameters a_{jk} instead of the four parameters from $b_{\cdot j}$, so that due to the use of $\|\cdot\|_0$ penalties the explanation using Z is preferred. We therefore have $\widehat{b}_i\widehat{b}_j - \sigma_{ij} \to 0$ so that $\widehat{B}$ converges to B.

Furthermore, given the influence B of Z on X, we can learn the matrix A from the consistency result of (Van de Geer & Bühlmann, 2013). □

C. Scoring and Optimization

We can estimate the expectation term of the ELBO by Monte Carlo Approximation

$$
\begin{aligned}
&E_{q(Z,\epsilon|X)}\left[\log p(Z,\epsilon \mid X)\right] \\
&\approx \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{m} \frac{\left(X_j - \mu_x(Z_{(k)},\epsilon_{(k)})_j\right)^2}{2\sigma_x(Z_{(k)},\epsilon_{(k)})_j^2} \\
&- 2\log\left(\sigma_x(Z_{(k)},\epsilon_{(k)})_j\right) - c
\end{aligned}
$$

where $Z_{(k)}, \epsilon_{(k)}$ is the k-th sample for the encoding of X. We can compute the KL-divergence as

$$
\begin{aligned}
&D\left(q(Z,\epsilon \mid X) \parallel p(Z,\epsilon)\right) \\
&= \frac{1}{2}\sum_{j}\Big(\sigma_{Z,\epsilon}(X)_j^2 + \mu_{Z,\epsilon}(X)_j^2 - 2\log\left(\sigma_{Z,\epsilon}(X)_j^2\right)\Big)
\end{aligned}
$$

where $\mu_{Z,\epsilon} = (\mu_Z, \mu_\epsilon)$ and similarly for $\sigma_{Z,\epsilon}$.

Next, as noted above, the learning problem

$$
\begin{aligned}
\min_{A,B,\theta} \quad & f(A,B,\theta) \coloneqq -L_{\text{ELBO}} + \lambda_A \|A\|_1 + \lambda_B \|B\|_1 \\
& \text{s.t. } h(A) \coloneqq \text{trace}((I + (A \odot A)/m)^m) - m = 0
\end{aligned}
$$

can be reformulated as

$$
L(A,B,\theta;\lambda,\rho) = f(A,B,\theta) + \lambda h(A) + \frac{\rho}{2}\left|h(A)\right|^2 .
$$

and solved by using a dual ascent approach (Bertsekas, 1997). More specifically, one of the most common updating schemes for the values A, B, θ and parameters λ, ρ is

$$
\begin{aligned}
A^k, B^k, \theta^k &= \text{argmin}_{A,B,\theta} L(A,B,\theta;\lambda^k,\rho^k) \\
\lambda^{k+1} &= \lambda^k + \rho h(A^k) \\
\rho^{k+1} &= \begin{cases} \alpha\rho^k & \text{if } h(A^k) > \gamma h(A^{k-1}) \\ \rho^k & \text{otherwise} \end{cases}
\end{aligned}
$$

where $\alpha > 1, \gamma < 1$ determine how quickly ρ increases.

The first of these equations can be solved using any black box stochastic optimization algorithm readily available in machine learning toolboxes. For the other two equations, we use the commonly used hyperparameters $\alpha = 10, \gamma = 0.25$ (Yu et al., 2019).

D. Evaluating GFCI and 3OFF2

As mentioned in the text, evaluating GFCI and 3OFF2 with our metrics is not straightforward. This is due to both of them using only local scores to discern local structures, and these scores cannot be aggregated straightforwardly. We, therefore, sort the obtained scores of both GFCI and 3OFF2 in the *best possible way* for their evaluation. Even so, we see that NOCADILAC outperforms both of these methods for all metrics.

Further, GFCI does not return a unique directed graph but an equivalence class of graphical models. In particular, latent confounding is possible for many edges in the discovered networks, but for only very few of them, they are confident that they are confounded.

To evaluate the returned equivalence classes, we, therefore, evaluate $F_1^{\text{net}}, F_1^{\text{conf}}$ and SID over all networks in these equivalence classes where feasible – which is infrequently the case – and over a random sample of 10000 networks from these equivalence classes where it is not. We then take the average of these scores to evaluate GFCI as this is the most reasonable evaluation of their results. However, even taking the upper quartile of these scores does not change the fact that NOCADILAC consistently outperforms GFCI.

E. Higher-dimensional Z

We next consider the effect of including multiple confounding factors Z_i in our generating data, each influencing non-overlapping sets of 5 variables in a network of 50 variables total. We show the overall F_1^{conf} scores for one to five confounders in Table 1.

We see that for one to three confounders, NoCaDiLaC performs at a consistent level. However, at four and five confounders, its performance decreases due to the increasing difficulty of fitting a learning a good model of such complexity. In contrast, all of DCD, 3off2, and GFCI perform less well from the start, and their performance drops immediately upon adding a second latent confounder. The reason for this is instructive: since none of them infer a latent confounder but only indicate whether pairs of variables may be confounded, they have no way of distinguishing between sets of nodes confounded by different factors. We, therefore, ran a spectral clustering algorithm on the subgraphs containing only edges due to pairwise confounding to return the correct number of confounded sets. Nevertheless, the returned subgraphs were of low quality, as reflected by the very low F_1^{conf} scores for each of the competitors.

	Number of confounders				
Method	1	2	3	4	5
NoCaDiLaC	0.42	0.38	0.35	0.23	0.15
DCD	0.23	0.14	0.11	0.07	0.03
3off2	0.22	0.12	0.08	0.05	0.03
GFCI	0.21	0.07	0.03	0.02	0.01

Table 1: Comparison of NoCaDiLaC, DCD, 3off2 and GFCI for graphs with varying numbers of latent confounders. While all methods perform well, only NoCaDiLaC maintains its performance as the number of latent factors increases.

F. Details on REGED Setting

As mentioned in the text, we can only use part of the network provided by REGED at once if our goal is to introduce confounders. Instead, we generate a number of subgraphs as follows. For every node with an outdegree of at least 3, we start by taking its children C. We then take the union of the (minimal) Markov blankets of nodes in C as our vertex set $V' = \bigcup_{c \in C} \text{MB}(c)$ and use its induced graph $G' = G[V']$. Note that even though we start with only C as the set of confounded nodes, it is possible that the addition of the Markov blankets adds nodes with common parents that are not themselves in V'. Our confounded set is, therefore, some larger $C' \supseteq C$, which is nevertheless known beforehand.