Causal Reasoning in the Presence of Latent Confounders via Neural ADMG Learning

Matthew Ashman*
University of Cambridge
Cambridge, UK
mca39@cam.ac.uk

Chao Ma*
Microsoft Research
Cambridge, UK
chao.ma@microsoft.com

Agrin Hilmkil
Microsoft Research
Cambridge, UK
agrinhilmkil@microsoft.com

Joel Jennings
Microsoft Research
Cambridge, UK
joeljennings@microsoft.com

Cheng Zhang
Microsoft Research
Cambridge, UK
cheng.zhang@microsoft.com

Abstract

Latent confounding has been a long-standing obstacle for causal reasoning from observational data. One popular approach is to model the data using acyclic directed mixed graphs (ADMGs), which describe ancestral relations between variables using directed and bidirected edges. However, existing methods using ADMGs are based on either linear functional assumptions or a discrete search that is complicated to use and lacks computational tractability for large datasets. In this work, we further extend the existing body of work and develop a novel gradient-based approach to learning an ADMG with non-linear functional relations from observational data. We first show that the presence of latent confounding is identifiable under the assumptions of bow-free ADMGs with non-linear additive noise models. With this insight, we propose a novel neural causal model based on autoregressive flows. This not only enables us to model complex causal relationships behind the data, but also estimate their functional relationships (hence treatment effects) simultaneously. We further validate our approach via experiments on both synthetic and real-world datasets, and demonstrate the competitive performance against relevant baselines.

1 Introduction

Latent confounding has been a long-standing obstacle for causal reasoning from observational data. If not properly accounted for, the presence of latent confounding can lead to incorrect evaluation of causal quantities of interest (Pearl, 2009). Traditional causal discovery methods that account for the presence of latent confoundings, such as the fast causal inference algorithm (FCI) (Spirtes et al., 2000) and its extensions (Colombo et al., 2012; Claassen et al., 2013; Chen et al., 2021), rely on uncovering an equivalence class of acyclic directed mixed graphs (ADMGs) that share the same conditional independencies. Without additional assumptions, however, these methods might return uninformative results as they cannot distinguish between members of the same Markov equivalence class (Bellot & van der Schaar, 2021).

More recently, causal discovery methods based on structural causal models (SCMs) (Pearl, 1998) have been developed for latent confounding (Nowzohour et al., 2017; Wang & Drton, 2020; Maeda & Shimizu, 2020, 2021; Bhattacharya et al., 2021). By assuming that the causal effects follow specific functional forms, they have the advantage of being able to distinguish between members of the same Markov equivalence class (Glymour et al., 2019). Yet, existing approaches either rely on restrictive linear functional assumptions (Bhattacharya et al., 2021; Maeda & Shimizu, 2020; Bellot

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

& van der Schaar, 2021), or discrete search over the space of causal graphs (Maeda & Shimizu, 2021) that are computationally burdensome and unintuitive to use. As a result, modeling non-linear causal relationships between variables in a scalable way remains an outstanding task.

In this work, we seek to utilize recent advances in differentiable causal discovery (Zheng et al., 2018; Bhattacharya et al., 2021) and neural causal models (Lachapelle et al., 2019; Geffner et al., 2022) to overcome these limitations. Our core contribution is to extend the framework of differentiable ADMG discovery for linear models (Bhattacharya et al., 2021) to non-linear cases using neural causal models. This enables us to build efficient and flexible methods capable of discovering non-linear, potentially confounded relationships between variables and perform subsequent causal inference. Specifically, our contributions include:

1. **Sufficient conditions for ADMG identifiability with non-linear SCMs.** We assume: i) the functional relationship follows non-linear additive noise SCM; ii) the effect of observed and latent variables do not modulate each other, and iii) all latent variables confound a pair of non-adjacent observed nodes. Under these assumptions, the underlying ground truth ADMG causal graph is identifiable.

2. **A novel gradient-based framework for learning ADMGs from observational data.** Based on our ADMG identification conditions, we further propose Neural ADMG Learning (N-ADMG), a neural autoregressive-flow-based model capable of learning complex non-linear causal relationships in the presence of latent confounding. N-ADMG utilizes variational inference to approximate posteriors over causal graphs and latent variables, whilst simultaneously learning the model parameters via gradient-based optimization.

3. **Empirical evaluation on synthetic and real-world datasets.** We evaluate N-ADMG on a variety of synthetic and real-world datasets, comparing performance with a number of existing causal discovery and inference algorithms. We find that N-ADMG provides competitive or state-of-the-art results on a range of causal reasoning tasks.

2 Background: acyclic directed mixed graphs (ADMGs) and its magnification

One of the most widely-used graphical representations of causal relationships involving latent confounding is the so-called acyclic directed mixed graph (ADMGs). ADMGs are an extension of DAGs, that contain both directed edges (→) and bidirected edges (↔) between variables. More concretely, the directed edge $x_i \to x_j$ indicates that x_i is an ancestor of x_j, and the bidirected edge $x_i \leftrightarrow x_j$ indicates that x_i and x_j share a common, unobserved ancestor. An ADMG G over a collection of D variables $\mathbf{x} = (x_1, \dots, x_D)$ can be described using two binary adjacency matrices: $G_D \in \mathbb{R}^{D \times D}$, for which an entry of 1 in position (i, j) indicates the presence of the directed edge $x_i \to x_j$, and $G_B \in \mathbb{R}^{D \times D}$, for which an entry of 1 in position (i, j) indicates the presence of the bidirected edge $x_i \leftrightarrow x_j$.

Similar to a DAG, an SCM can be specified to describe the causal relationships implied by an ADMG through the so-called *magnification* process. As formulated in (Peña, 2016), whenever a bidirected edge $x_i \leftrightarrow x_j$ is present according to G_B in an ADMG, we will explicitly add a latent node u_m to represent the confounder of $x_i \leftrightarrow x_j$. Then, the SCM of x_i can be written as $\mathbf{x}$ by $x_i = f_i(\mathbf{x}_{\text{pa}(i;G_D)}, \mathbf{u}_{\text{pa}(i;G_B)}) + \epsilon_i$, where $\mathbf{u}_{\text{pa}(i;G_B)}$ denotes the latent parents of x_i in the set of all latent nodes $\mathbf{u} = (u_1, \dots, u_M)$ added in the magnification process. In the compact form, we can write:

$$[\mathbf{x}, \mathbf{u}] = f_G(\mathbf{x}, \mathbf{u}) + \boldsymbol{\epsilon}. \tag{1}$$

The "magnified SCM" will serve as a practical device for learning ADMGs. Similar to DAGs, magnified SCMs induce a observational distribution on $\mathbf{x}$, denoted by $p_\theta(\mathbf{x}^n; G)$. Note that given an ADMG, the magnified SCM is not unique, since latent variables may be shared. For example, $x_1 \leftrightarrow x_2$, $x_2 \leftrightarrow x_3$, $x_3 \leftrightarrow x_1$ can be magnified into both $\{x_1 \leftarrow u_1 \to x_2, x_2 \leftarrow u_2 \to x_3, x_3 \leftarrow u_3 \to x_1\}$, and $\{u_1 \to x_1, u_1 \to x_2, u_1 \to x_3\}$. Therefore, the of ADMG identifiability does not imply the structural identifiability of the magnified SCM. In this paper, we focus on ADMG identifiability.

3 Establishing ADMGs identifiability under non-linear SCMs

To build flexible methods that are capable of discovering causal relationships under the presence of latent confounding, we first need to establish the identifiability of ADMGs under non-linear SCMs. The concept of structural identifiability of ADMGs is formalized in the following definition:

Definition 1 (ADMG structural identifiability)**.** *For a distribution $p_\theta(\mathbf{x}; G)$, the ADMG $G = \{G_D, G_B\}$ is said to be structurally identifiable from $p_\theta(\mathbf{x}; G)$ if there exists no other distribution $p_{\theta'}(\mathbf{x}; G')$ such that $G \neq G'$ and $p_\theta(\mathbf{x}; G) = p_{\theta'}(\mathbf{x}; G')$.*

Assuming that our model is correctly specified and $p_{\theta^0}(\mathbf{x}; G^0)$ denotes the true data generating distribution, then ADMG structural identifiability guarantees that if we find some $p_\theta(\mathbf{x}; G) = p_{\theta^0}(\mathbf{x}; G^0)$ (by e.g., maximum likelihood learning), we can recover $G = G^0$. In this section, we seek to establish sufficient conditions under which the ADMG identifiability is satisfied. Let $\mathbf{x} = (x_1, \ldots, x_D)$ be a collection of observed random variables, and $\mathbf{u} = (u_1, \ldots, u_M)$ be a collection of unobserved (latent) random variables. Our first assumption is that data generating process can be expressed as a specific non-linear additive noise SCM, in which the effect of the observed and latent variables do not modulate each other :

Assumption 1. *We assume that the data generating process takes the form*

$$[\mathbf{x}, \mathbf{u}]^\top = f_{G_D,\mathbf{x}}(\mathbf{x}; \theta) + f_{G_B,\mathbf{u}}(\mathbf{u}; \theta) + \boldsymbol{\epsilon} \tag{2}$$

where each element of $\boldsymbol{\epsilon}$ is independent of all other variables in the model and θ denotes the parameters of the non-linear functions $f_{G_D,\mathbf{x}}$ and $f_{G_B,\mathbf{u}}$.

This assumption is one of the elements that separate us from previous work of (Bhattacharya et al., 2021; Maeda & Shimizu, 2020) (linearity is assumed) and (Maeda & Shimizu, 2021) (effects are assumed to be fully decoupled, $x_i = \sum_m f_{im}(x_m \in \mathbf{x}_{\text{pa}(i;G_D)}) + \sum_k g_{ik}(u_k \in \mathbf{u}_{\text{pa}(i;G_B)}) + \epsilon_i$). As discussed in Section 2, the discovery of an ADMG between observed variables amounts to the discovery of their ancestral relationships. Therefore, the mapping between ADMGs and their magnified SCMs is not one-to-one, which might cause issues when designing causal discovery methods using SCM-based approaches. To further simplify the underlying latent structure (without losing too much generality), our second assumption assumes that every latent variable is a parentless common cause of a pair of non-adjacent observed variables:

Assumption 2 (Latent variables confound pairs of non-adjacent observed variables)**.** *For each latent variable u_k in the data generating process, there exists a non-adjacent pair x_i and x_j that is unique to u_k, such that $x_i \leftarrow u_k \rightarrow x_j$.*

Arguments for Assumption 2 are made by (Pearl & Verma, 1992), who show that this family of causal graphs is very flexible, and can produce the same conditional independencies amongst the observed variables in any given causal graph. Therefore, it has been argued that without loss of generality, we can assume latent variables to be exogenous, and have exactly two non-adjacent children. This assumption also implies that we can safely specify an magnified SCM to the ADMG as shown in Section 2, which allows us to evaluate and optimize the induced likelihood later on. Given Assumptions 1 and 2, we can prove the ADMG identifiability result:

Proposition 1 (Identifiability of ADMGs under non-linear SCMs)**.** *Assume Assumptions 1 to 2 hold. Then, the ADMG matrices G_D and G_B are identifiable for the data generating process specified in Equation 2.*

Whilst theoretically we could test for these conditions stated in Lemmas 1 to 3 directly, a more efficient approach is to use differentiable maximum likelihood learning. Assuming the model is correctly specified, then ADMG identifiability ensures the maximum likelihood estimate recovers the true graph in the limit of infinite data (Appendix C). Hence, we can design the ADMG identification algorithms via ML learning.

4 Neural ADMG Learning

Whilst we have outlined sufficient conditions under which the underlying ADMG causal structure can be identified, this does not directly provide a framework through which causal discovery and inference can be performed. In this section, we seek to formulate a practical framework for gradient-based ADMG identification. Three challenges that remain are: 1. How can we parameterize the magnified SCM models for ADMG to enable learning flexible causal relationships? 2. How can we optimize the causal graphs in the space of ADMGs as assumed Section 3? 3. How do we learn the ADMG causal structure efficiently, whilst accounting for the missing data ($\mathbf{u}$) and graph uncertainties in the finite data regime? In this section, we present Neural ADMG Learning (N-ADMG), a novel framework that addresses all three challenges.

4.1 Neural auto-regressive flow parameterization

We assume that our model used to learn ADMGs from data is correctly specified. That is, it can be written in the same magnified SCM form as in Equation 2:

$$[\mathbf{x}, \mathbf{u}]^\top = f_{G_D,\mathbf{x}}(\mathbf{x}; \theta) + f_{G_B,\mathbf{u}}(\mathbf{u}; \theta) + \boldsymbol{\epsilon} \tag{3}$$

Following Khemakhem et al. (2020), we factorise the likelihood $p_\theta(\mathbf{x}^n, \mathbf{u}^n|G)$ induced by Equation 3 in an autoregressive manner. We can rearrange Equation 3 as

$$\boldsymbol{\epsilon} = \mathbf{v} - f_{G_D,\mathbf{x}}(\mathbf{x}; \theta) - f_{G_B,\mathbf{u}}(\mathbf{u}; \theta) := g_{\tilde{G}}(\mathbf{v}; \theta) = \mathbf{v} - f_{\tilde{G}}(\mathbf{v}; \theta) \tag{4}$$

where $\mathbf{v} = (\mathbf{x}, \mathbf{u}) \in \mathbb{R}^{D+M}$, and $\tilde{G}$ is the magnified adjacency matrix on $\mathbf{v}$, defined as $\tilde{G}_{i,j} \in \{0, 1\}$ if and only if $v_i \to v_j$. This allows us to express the likelihood as

$$p_\theta(\mathbf{v}^n|G) = p_{\boldsymbol{\epsilon}}(g_{\tilde{G}}(\mathbf{v}^n; \theta)) = \prod_{i=1}^{D+M} p_{\epsilon_i}(g_{\tilde{G}}(\mathbf{v}^n; \theta)_i). \tag{5}$$

Note that we have omitted the Jacobian-determinant term as it is equal to one since $\tilde{G}$ is acyclic (Mooij et al., 2011). Following Geffner et al. (2022), we adopt an efficient, flexible parameterization for the functions f_i taking the form (consistent with Assumption 1)

$$f_i(\mathbf{v}) = \xi_{1,i}\left(\sum_{v_j \in \mathbf{x}}^{D} \tilde{G}_{j,i}\ell_j(v_j)\right) + \xi_{2,i}\left(\sum_{v_j \in \mathbf{u}}^{M} \tilde{G}_{j,i}\ell_j(v_j)\right) \tag{6}$$

where $\xi_{1,i}$, $\xi_{2,i}$ and ℓ_i ($i = 1, ..., D + M$) are MLPs. A naïve implementation would require training $3(D+M)$ neural networks. Instead, we construct these MLPs so that their weights are shared across nodes as $\xi_{1,i}(\cdot) = \xi_{1,i}(\mathbf{e}_i, \cdot)$, $\xi_{2,i}(\cdot) = \xi_{2,i}(\mathbf{e}_i, \cdot)$ and $\ell_i(\cdot) = \ell(\mathbf{e}_i, \cdot)$, with $\mathbf{e}_i \in \mathbb{R}^{D+M}$ a trainable embedding that identifies the output and input nodes respectively.

4.2 ADMG learning via maximizing evidence lower bound

Our ADMG identifiability theory in Section 3 suggests that the ground truth ADMG graph can, in principle, be recovered via maximum likelihood learning of $p_\theta(\mathbf{x}|G)$. However, the aforementioned challenges remain: given a finite number of observations $\mathbf{x}^1, \dots, \mathbf{x}^N$, how do we deal with the corresponding missing data $\mathbf{u}^1, \dots, \mathbf{u}^N$ while learning the ADMG? How do we account for graph uncertainties and ambiguities in the finite data regime? To address these issues, N-ADMG takes a Bayesian approach toward ADMG learning. We jointly model the distribution over the ADMG graph G, the observations $\mathbf{x}^1, \dots, \mathbf{x}^N$, and the corresponding latent variables $\mathbf{u}^1, \dots, \mathbf{u}^N$, as

$$p_\theta(\mathbf{x}^1, \mathbf{u}^1, \dots, \mathbf{x}^N, \mathbf{u}^N, G) = p(G) \prod_{n=1}^{N} p_\theta(\mathbf{x}^n, \mathbf{u}^n|G) \tag{7}$$

where $p_\theta(\mathbf{x}^n, \mathbf{u}^n|G)$ is the neural SCM model specified in Section 4.1, θ denotes the corresponding model parameters, and $p(G)$ is some prior distribution over the graph. Our goal is to learn both the model parameters θ and an approximation to the posterior $q_\phi(\mathbf{u}^1, \dots, \mathbf{u}^N, G) \approx p_\theta(\mathbf{u}^1, \dots, \mathbf{u}^N, G|\mathbf{x}^1, \dots, \mathbf{x}^N)$. This can be achieved jointly using the variational inference framework (Zhang et al., 2018; Kingma & Welling, 2013), in which we maximize the evidence lower bound (ELBO) of $\sum_n \log p_\theta(\mathbf{x}^n)$ given by

$$\begin{aligned}
&\sum_n \log p_\theta(\mathbf{x}^n) \geq \mathcal{L}_{\text{ELBO}}(\theta, \phi) \\
=&\mathbb{E}_{q_\phi(G)}\left[\sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathbf{u}^n|G)}\left[\log p_\theta(\mathbf{x}^n|\mathbf{u}^n, G)\right]\right] - \text{KL}\left[q_\phi(G)||p(G)\right] \\
&- \mathbb{E}_{q_\phi(G)}\left[\sum_{n=1}^{N} \text{KL}\left[q_\phi(\mathbf{u}^n|\mathbf{x}^n, G))||p_\theta(\mathbf{u}^n|G)\right]\right].
\end{aligned} \tag{8}$$

We parameterize approximate posterior as $q_\phi(\mathbf{u}^1, \dots, \mathbf{u}^N, G) = \prod_n q_\phi(\mathbf{u}^n|\mathbf{x}^n) q_\phi(G)$. For $q_\phi(G)$, we use mean-field Bernoulli distributions to parameterize the ADMG adjancency matrices for G_D and G_B. For G_D, edge existence and edge orientation are parameterized separately using the ENCO (Lippe et al., 2021). For the posteriors on $\mathbf{u}^n$, we apply amortized Gaussian VI as in VAE literature (Kingma & Welling, 2013), where $q_\phi(\mathbf{u}^n|\mathbf{x}^n)$ is parameterized via a neural inference net. In Appendix C we demonstrate that maximizing $\mathcal{L}_{\text{ELBO}}(\theta, \phi)$ recovers the true ADMG causal graph in the limit of infinite data.

4.3 Choice of prior over ADMG graphs

As discussed in Section 2, the ADMG G can be parameterized by two binary adjacency matrices, G_D whose entries indicate the presence of a directed edge, and G_B whose edges indicate the presence of a bidirected edge. As discussed in Section 3, a necessary assumption for structural identifiability is that each latent variable is a parent-less confounder of a pair of non-adjacent observed variables. This further implies that the underlying ADMG must be bow-free (both a directed and a bidirected edge cannot exist between the same pair of observed variables). This can be enforced by leveraging the bow-free constrain penalty introduced by Bhattacharya et al. (2021),

$$h(G_D, G_B) = \text{trace}\left(e^{G_D}\right) - D + \text{sum}\left(G_D \circ G_B\right) \tag{9}$$

which is non-negative and zero only if (G_D, G_B) is a bow-free ADMG. As suggested in Geffner et al. (2022), we implement the prior as

$$p(G) \propto \exp\left(-\lambda_{s1}\|G_D\|_F^2 - \lambda_{s2}\|G_B\|_F^2 - \rho h(G_D, G_B)^2 - \alpha h(G_D, G_B)\right) \tag{10}$$

where the coefficients α and ρ are gradually increased when maximizing $\mathcal{L}_{\text{ELBO}}(\theta, \phi)$, following an augmented Lagrangian scheme (Nemirovski, 1999). Prior knowledge about the sparseness of the graphs is introduced by penalizing the norms $\|G_D\|_F^2$ and $\|G_B\|_F^2$ with scaling coefficients λ_{s1} and λ_{s2}.

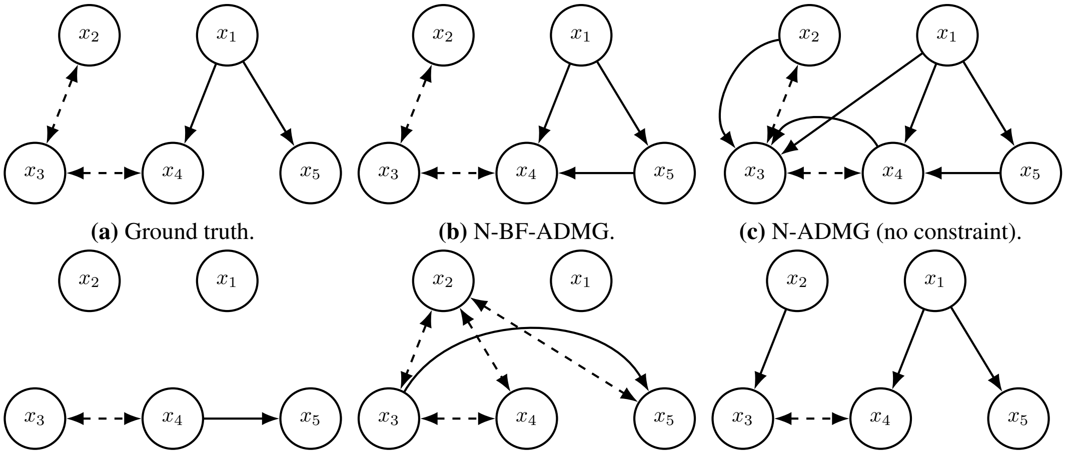
5 Experiments

We evaluate N-ADMG in performing both causal discovery and causal inference on a number of synthetic datasets. Additional result on a real-world dataset can be found in appendix I Note that we run our model *both with and without the bow-free constraint, identified as N-BF-ADMG (our full model) and N-ADMG (for ablation purpose)*, respectively. We compare the performance of our model against five baselines: DECI (Geffner et al., 2022) (which we refer to as N-DAG for consistency), FCI (Spirtes et al., 2000), RCD (Maeda & Shimizu, 2020), CAM-UV (Maeda & Shimizu, 2021), and DCD (Bhattacharya et al., 2021). We evaluate the causal discovery performance using F1 scores for directed and and bidirected adjacency. The expected values of these metrics are reported using the learned graph posterior (which is deterministic for RCD, CAM-UV, and DCD). Causal inference is evaluated using the expected ATE as described in Appendix E. We evaluate the causal inference performance of the causal discovery benchmarks by fixing $q(G)$ to either deterministic or uniform categorical distributions on the learned causal graphs, then learning a non-linear flow-based ANM by optimizing Equation 8 in an identical manner to N-ADMG. A full list of results and details of the experimental set-up are included in Appendix G.

5.1 Synthetic fork-collider dataset

We construct a synthetic fork-collider dataset consisting of five nodes (Figure 1a). The data-generating process is a non-linear ANM with Gaussian noise. Variable pairs (x_2, x_3), and (x_3, x_4) are latent-confounded, whereas (x_4, x_5) share a observed confounder. We evaluate both causal discovery and inference performances. For discovery, we evaluate F-score measure on both G_D and G_B. For causal inference, we choose x_4 as the treatment variable taking and x_2, x_3 and x_5 as the response variables, and evaluate ATE RMSE to benchmark performance.

Figure 1b-1f visualizes the causal discovery results of different methods, and Table 1 summarizes both discovery and inference metrics. our method N-BF-ADMG-G and CAM-UV achieved the best overall performance. N-BF-ADMG-G and N-ADMG-G on average are able to recover all bidirected edges from the data, while CAM-UV can only recover half of the latent variables. Without the bow-free constraint, N-ADMG-G discovers a directed edge from x_4 to x_3, which results in poor ATE RMSE performance. On the other hand, DAG-based method (N-DAG-G) is not able to deal with latent confounders, resulting in the poor f-scores. Linear ANM-based methods (RCD and DCD) perform significantly worse than other methods, resulting in 0 f-scores for directed matrices and largest ATE errors.



(a) Ground truth. **(b)** N-BF-ADMG. **(c)** N-ADMG (no constraint).

(d) RCD. **(e)** DCD. **(f)** CAM-UV.

Figure 1: ADMG identification results on fork-collider dataset.

5.2 Random confounded ER synthetic dataset

Table 1: Causal discovery and inference results for the synthetic fork-collider dataset. The table shows the mean and standard error results across five different random seeds.

METHOD	G_D FSCORE	G_B FSCORE	ATE RMSE
N-BF-ADMG-G	0.64 (0.06)	**0.93 (0.07)**	**0.022 (0.003)**
N-ADMG-G	0.49 (0.02)	**0.99 (0.00)**	0.239 (0.067)
N-DAG-G	0.50 (0.00)	0.00 (0.00)	**0.046 (0.025)**
FCI	0.00 (0.00)	0.75 (0.00)	0.072 (0.015)
RCD	0.00 (0.00)	0.54 (0.00)	0.206 (0.029)
CAM-UV	**0.80 (0.00)**	0.67 (0.00)	**0.017 (0.003)**
DCD	0.00 (0.00)	0.67 (0.00)	0.208 (0.064)

We generate synthetic datasets from ADMG extension of *Erdős-Rényi* (ER) graph model (Lachapelle et al., 2019), see Appendix H.2 for details. We consider the number of nodes, directed edges and latent confounders triplets $(d, e, m) \in \{(4, 6, 2), (8, 20, 6), (12, 50, 10)\}$. The resulting datasets are identified as $\mathbf{ER}(d, e, m)$. Figure 2 compares the performance of N-ADMG with the baselines. All variants of N-ADMG outperform the baselines for most datasets, highlighting its effectiveness relative to other methods (even those that employ similar assumptions). Similar to the fork-collider dataset, we see that methods operating under the assumption of linearity perform poorly when the data-generating process is non-linear. It is worth noting that even when the exogenous noise of N-ADMG is misspecified, the causal discovery performance still exceeds that of other methods in most cases. This robustness is a desirable property as in many settings the form of exogenous noise is unknown.

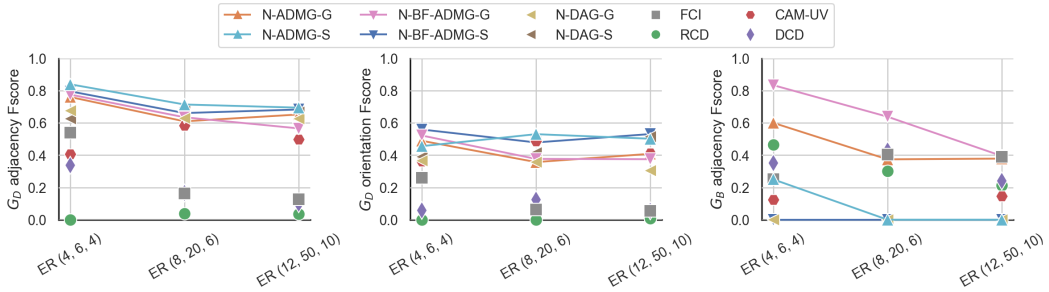


Figure 2: Causal discovery results for the synthetic ER datasets. For readability, the N-ADMG results are connected with lines. The figure shows mean results across five randomly generated datasets.

6 Conclusion

In this work, we proposed Neural ADMG Learning (N-ADMG), a novel framework for gradient-based causal discovery in the presence of latent confounding for nonlinear SCMs. Building upon existing work, we established identifiability theory for ADMGs under latent confounding, and proposed a practical ADMG learning algorithm that is both flexible and efficient. In future work, we will further extend our framework on how to more general cases (e.g., the effect from observed and latent variables can modulate in certain forms; latent variables can confound adjacent variables; etc), and how to extend to more complicated scenarios.

References

Alexis Bellot and Mihaela van der Schaar. Deconfounded score method: Scoring dags with dense unobserved confounding. *arXiv preprint arXiv:2103.15106*, 2021.

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

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 32, 2019.

Wenyu Chen, Mathias Drton, and Ali Shojaie. Causal structural learning via local graphs. *arXiv preprint arXiv:2107.03597*, 2021.

Tom Claassen, Joris M Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 172–181, 2013.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pp. 294–321, 2012.

Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3478–3486. PMLR, 16–18 Apr 2019.

Robin J Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43 (3):625–648, 2016.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Jennifer L Hill. Bayesian nonparametric modelling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. *arXiv preprint arXiv:1906.02226*, 2019.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Takashi Nicholas Maeda and Shohei Shimizu. RCD: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pp. 735–745. PMLR, 2020.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pp. 97–106. PMLR, 2021.

Sara Mohammad-Taheri, Robert Ness, Jeremy Zucker, and Olga Vitek. Do-calculus enables causal reasoning with latent variable models. *arXiv preprint arXiv:2102.06626*, 2021.

Joris M Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise model. *Advances in neural information processing systems 24*, 2011.

A Nemirovski. Optimization II. Numerical methods for nonlinear continuous optimization. 1999.

Christopher Nowzohour, Marloes H Maathuis, Robin J Evans, and Peter Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11(2):5342–5374, 2017.

Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Judea Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

Judea Pearl and Thomas S Verma. A statistical semantics for causation. *Statistics and Computing*, 2(2):91–95, 1992.

Jose M Peña. Learning acyclic directed mixed graphs from observations and interventions. In *Conference on Probabilistic Graphical Models*, pp. 392–402. PMLR, 2016.

Rajesh Ranganath and Adler Perotte. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*, 2018.

Amit Sharma, Vasilis Syrgkanis, Cheng Zhang, and Emre Kıcıman. Dowhy: Addressing challenges in expressing and validating causal assumptions. *arXiv preprint arXiv:2108.13518*, 2021.

Ilya Shpitser, Robin J Evans, Thomas S Richardson, and James M Robins. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, 2014.

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

Dustin Tran and David M Blei. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*, 2017.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270, 1990.

Y Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-Gaussian data. *arXiv preprint arXiv:2007.11131*, 2020.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Jiji Zhang. 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.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

APPENDIX

A Related Work

In this section, we review related works in causal discovery/inference under unobserved confounders.

Causal discovery with latent confounding. In the literature, constraint-based causal discovery methods in the presence of latent confounding have been well-studied (Spirtes et al., 2000; Zhang, 2008; Colombo et al., 2012; Claassen et al., 2013; Chen et al., 2021). In those methods, conditional independence tests are called iteratively to identify the graph skeletons and orientations to the extent that is possible. Without assumptions on the data generating process, these approaches can only identify a Markov equivalence class of causal structures that imply the same conditional independencies. When certain assumptions are made on the data generating process in the form of SCMs (Pearl, 1998), additional constraints can help identify the true causal structure. In the most general case, non-parametric constraints have been identified which can refine causal structures beyond those identified by conditional independency restrictions (Verma & Pearl, 1990; Shpitser et al., 2014; Evans, 2016). Yet, further refinement can be made through the assumption of stricter SCMs. For example, in the linear Gaussian additive noise model (ANM) case, Nowzohour et al. (2017) proposes a score-based approach for finding an equivalent class of bow-free acyclic path diagrams. Both Maeda & Shimizu (2020) and Wang & Drton (2020) develop additional conditional independency tests for linear non-Gaussian ANM case, with Maeda & Shimizu (2021) extending this to more general cases.

Differentiable characterization of causal discovery. All aforementioned approaches employ a search over a discrete space of causal structures, which often requires task-specific search procedures, and imposes a computational burden for large-scale problems. More recently, (Zheng et al., 2018) proposed a differentiable constraint on directed acyclic graphs (DAG), and frames the graph structure learning problem as a differentiable optimization task in the absence of latent confounders. This is further generalized to the latent confounding case (Bhattacharya et al., 2021) through differentiable algebraic constraints that characterize the space of acyclic directed mixed graphs (ADMGs). Given these differentiable constraints, the authors define a score based on the Bayesian information criterion to characterize how well each graph fits the data, enabling the deployment of gradient-based optimization techniques to find the best fitting graph within a discrete space. Nonetheless, this work is limited in that it only considers linear Gaussian ANMs.

Causal inference under latent confounding. Attempting to perform causal inference in the presence of latent confounding can lead to biased estimates (Pearl, 2012). Whilst the observed data distribution may still be identifiable, estimating causal effects are not (Spirtes et al., 2000). A recent string of work has made progress in the case where the effects of multiple interventions are being estimated (Tran & Blei, 2017; Ranganath & Perotte, 2018; Wang & Blei, 2019; D'Amour, 2019). An alternative approach is to simply assume identifiability of the joint distribution over both latent and observed variables given just the observations. Louizos et al. (2017) point out that there are many cases in which this is possible (Khemakhem et al., 2020; Kingma & Welling, 2013). Nevertheless, all these methods assume the underlying causal graph is known. More recently, Mohammad-Taheri et al. (2021) argue that a DAG latent variable model trained on data can be used for down-stream causal inference tasks even if its parameters are non-identifiable, as long as the query can be identified from the observed variables according to the do-calculus.

B Proof of proposition 1

Given Assumptions 1 and 2, to prove proposition 1, we need the following three lemmas which allow us to identify the ADMG matrices, G_D and G_B. These lemmas are extensions of the previous results in Maeda & Shimizu (2021) under our new assumptions above. We adopt the notation $\mathbf{x}_{-i}$ to denote all variables in x excluding x_i.

Definition 2 (Residual faithfulness condition)**.** *We say that nonlinear functions g_i, g_j satisfies the residual faithfulness condition, if: for any two arbitrary subset of $\mathbf{x}$, denote by M and N, when both $(x_i - g_i(M))$ and $(x_i - g_j(M))$ have terms involving the same exogenous noise ϵ_k, then $(x_i - g_i(M))$ and $(x_i - g_j(M))$ are mutually dependent.*

Lemma 1 (Case 1)**.** *Assume the data generation process follows Assumptions 1 and 2. Then, $[G_B]_{i,j} = 1$ and $[G_D]_{i,j} = 0$ if and only if*

$$\forall g_i, g_j,\ [(x_i - g_i(\mathbf{x}_{-i}) \not\perp\!\!\!\perp (x_j - g_j(\mathbf{x}_{-j})], \tag{11}$$

where g_i and g_j denote regression functions satisfying the residual faithfulness condition in Definition 2.

Lemma 2 (Case 2)**.** *Assume the data generation process follows Assumptions 1 and 2. Then, $[G_B]_{i,j} = 0$ and $[G_D]_{i,j} = 0$ if and only if*

$$\exists g_i, g_j,\ [(x_i - g_i(\mathbf{x}_{-(i,j)})) \perp\!\!\!\perp (x_j - g_j(\mathbf{x}_{-(i,j)}))], \tag{12}$$

where g_i and g_j denote regression functions satisfying the residual faithfulness condition in Definition 2.

Lemma 3 (Case 3)**.** *Assume the data generation process follows Assumptions 1 and 2. Then, $[G_B]_{i,j} = 0$ and $[G_D]_{i,j} = 1$ if and only if*

$$\forall g_i, g_j,\ [(x_i - g_i(\mathbf{x}_{-(i,j)})) \not\perp\!\!\!\perp (x_j - g_j(\mathbf{x}_{-j})] \tag{13}$$

$$\exists g_i, g_j,\ [(x_i - g_i(\mathbf{x}_{-i}) \perp\!\!\!\perp (x_j - g_j(\mathbf{x}_{-(i,j)}))], \tag{14}$$

where g_i and g_j denote regression functions satisfying the residual faithfulness condition in Definition 2.

B.1 Proof of Lemmas 1 to 3

Here, we provide the proof for Lemmas 1 to 3 in Section 3. To prove those lemmas, we need the help of the following lemma, which extends Lemma A in (Maeda & Shimizu, 2021), except we don't assume any form for g_i other than non-linearity.

Lemma 4. *Let $s(x_i)$ denote an arbitrary function of x_i. The residual of $s(x_i)$ regressed onto $\mathbf{x}_{-i}$ cannot be independent of ϵ_i:*

$$\forall g_i,\ [s(x_i) - g_i(\mathbf{x}_{-i}) \not\perp\!\!\!\perp \epsilon_i]. \tag{15}$$

Proof. Assume that $[s(x_i) - g_i(\mathbf{x}_{-i}) \perp\!\!\!\perp \epsilon_i]$ holds, then $\mathbf{x}_{-i}$ must contain at least one descendent of x_i as it must have dependence on the noise ϵ_i to cancel effect of ϵ_i in $s(x_i)$. We can express $g_i(\mathbf{x}_{-i})$ as $u_i(\boldsymbol{\epsilon})$. Since g_i operates on variables defined by non-linear transformations of the exogenous noise terms, we cannot express u_i as $a_i(\boldsymbol{\epsilon}_{-i}) + b_i(\epsilon_i)$. $\mathbf{x}_{-i}$ contains a descendent of x_i, so $\boldsymbol{\epsilon}_{-i}$ includes at least one noise term ϵ_k that satisfies $x_i \perp\!\!\!\perp \epsilon_k$ (i.e. is not in x_i). Thus, terms containing ϵ_i cannot be fully removed from $s(x_i) - g_i(\mathbf{x}_{-i})$ and so $[s(x_i) - g_i(\mathbf{x}_{-i}) \perp\!\!\!\perp \epsilon_i]$ does not hold. □

B.2 Proof of Lemma 1

Proof. Define g_i and g_j as $g_i(\mathbf{x}_{-i}) = f_{i,\mathbf{x}}(\mathbf{pa}_{\mathbf{x}}(i)) + g'_i(\mathbf{x}_{-i})$ and $g_j(\mathbf{x}_{-j}) = f_{j,\mathbf{x}}(\mathbf{pa}_{\mathbf{x}}(j)) + g'_i(\mathbf{x}_{-j})$ respectively. Then, Equation 11 becomes equivalent to

$$\forall g'_i, g'_j,\ [(f_{i,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(i)) + \epsilon_i - g'_i(\mathbf{x}_{-i}) \not\perp\!\!\!\perp (f_{j,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(j)) + \epsilon_j - g'_j(\mathbf{x}_{-j})]. \tag{16}$$

Given Lemma 4 and following the same arguments as in Maeda & Shimizu (2021), this is equivalent to

$$(f_{i,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(i)) + \epsilon_i) \not\perp\!\!\!\perp (f_{j,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(j)) + \epsilon_j). \tag{17}$$

Since $\epsilon_i \perp\!\!\!\perp \epsilon_j$, we have

$$(f_{i,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(i)) \not\perp\!\!\!\perp \epsilon_j) \vee (n_i \not\perp\!\!\!\perp f_{j,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(j))) \vee (f_{i,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(i)) \not\perp\!\!\!\perp f_{j,\mathbf{u}}(\mathbf{pa}_{\mathbf{u}}(j))). \tag{18}$$

The first implies the existence of an unobserved mediator between x_j and x_i, the second implies the existence of an unobserved mediator between x_i and x_j, and the third implies the existence of an unobserved confounder. Given the assumption of latent variables being confounders and minimality, this indicates the presence of a latent confounder and no direct cause between x_i and x_j. □

B.3 Proof of Lemma 2

Proof. When Equation 12 holds, Equation 11 does not. Thus, there is no unobserved confounder between x_i and x_j. Assume that x_j is a direct cause of x_i, and that Equation 12 is satisfied for g_i and g_j. x_i contains a nonlinear function of ϵ_j that cannot be removed by $g_i(\mathbf{x}_{-(i,j)})$, thus $(x_i - g_i(\mathbf{x}_{-(i,j)}))\not\perp\!\!\!\perp\epsilon_j$. Similarly, $(x_j - g_j(\mathbf{x}_{-(i,j)}))\not\perp\!\!\!\perp\epsilon_i$. Thus, we have $[(x_i - g_i(\mathbf{x}_{-(i,j)})\not\perp\!\!\!\perp(x_j - g_j(\mathbf{x}_{-(i,j)}))]$ which contradicts our initial assumption. The same arguments apply when x_i is a direct cause of x_j, implying that there can be no causal relationship between x_i and x_j. □

B.4 Proof of Lemma 3

Proof. When Equation 14 holds, Equation 11 does not and so there is no latent confounder between x_i and x_j. When *Equation* 13 holds, Equation 12 does not hold and so there is a direct causal relationship between x_i and x_j.

Assume that x_j is a direct cause of x_i. Define $g_1(\mathbf{x}_{-i}) = f_{i,\mathbf{x}}(\mathbf{pa}_\mathbf{x}(i))$ and $g_2(\mathbf{x}_{-(i,j)}) = f_{j,\mathbf{x}}(\mathbf{pa}_\mathbf{x}(j))$, giving $x_i - g_i(\mathbf{x}_{-i}) = f_{i,\mathbf{u}}(\mathbf{pa}_\mathbf{u}(i)) + \epsilon_i$ and $x_j - g_j(\mathbf{x}_{-(i,j)}) = f_{j,\mathbf{u}}(\mathbf{pa}_\mathbf{u}(j)) + \epsilon_j$. When there is no latent confounder, $f_{i,\mathbf{u}}(\mathbf{pa}_\mathbf{u}(i))\perp\!\!\!\perp f_{j,\mathbf{u}}(\mathbf{pa}_\mathbf{u}(j))$. Thus, $(f_{i,\mathbf{u}}(\mathbf{pa}_\mathbf{u}(i)) + \epsilon_i)\perp\!\!\!\perp(f_{j,\mathbf{u}}(\mathbf{pa}_\mathbf{u}(j)) + \epsilon_j)$ hods and Equation 14 is satisfied.

Now, assume that x_i is a direct cause of x_j. Using Lemma 4, we have

$$\forall g_i,\ [(x_i - g_i(\mathbf{x}_{-i})\not\perp\!\!\!\perp\epsilon_i]. \tag{19}$$

Similarly, since x_j is a function of x_i we also have

$$\forall g_j,\ [(x_j - g_j(\mathbf{x}_{-(i,j)})\not\perp\!\!\!\perp\epsilon_i]. \tag{20}$$

Collectively, this implies

$$\forall g_i, g_j,\ [(x_i - g_i(\mathbf{x}_{-i})\not\perp\!\!\!\perp(x_j - g_j(\mathbf{x}_{-i,j}))] \tag{21}$$

which contradicts Equation 14. Thus, if Equation 14 is satisfied then x_j is a direct cause of x_i. □

B.5 Finishing the proof for proposition 1

Although Lemmas 1 to 3 cannot cover the scenario where $[G_B]_{i,j} = 1$ and $[G_D]_{i,j} = 1$, note this case is already eliminated by Assumption 2 (confounded pairs must be non-adjacent). Since each case leads to mutually exclusive conditional independence/dependence constraints, for any $p_\theta(\mathbf{x}; G)$ specified under the assumptions above, there cannot exist some $p_{\theta'}(\mathbf{x}; G')$ such that $G \neq G'$ and $p_\theta(\mathbf{x}; G) = p_{\theta'}(\mathbf{x}; G')$, thus structural identifiability is satisfied. This completes the proof for our ADMG identifiability result in proposition 1.

C Maximizing $\mathcal{L}_{\text{ELBO}}(\theta, \phi)$ Recovers the Ground Truth ADMG

In Section 3, we have proved the structural identifiability of ADMGs. In this section, we further show that under certain assumptions, maximizing $\mathcal{L}_{\text{ELBO}}(\theta, \phi)$ recovers the true ADMG graph (denoted by G^0) in the infinite data limit. This result is stated in the following proposition:

Proposition 2 (Maximizing $\mathcal{L}_{\text{ELBO}}(\theta, \phi)$ recovers the ground truth ADMG)**.** *Assume that:*

- *Assumptions 1 and 2 (hence the identifiability of ADMGs) holds for the model $p_\theta(\mathbf{x}; G)$.*
- *The model is correctly specified (there exists θ^* such that $p_{\theta^*}(\mathbf{x}; G^0)$ recovers the data-generating process)*
- *Regularity condition: for all θ and G we have $\mathbb{E}_{p(\mathbf{x};G^0)}[|\log p_\theta(\mathbf{x}; G)|] < \infty$.*
- *The variational family of $q_\phi(\mathbf{u}|\mathbf{x}, G)$ is flexible enough, i.e., it contains the true posterior $p_\theta(\mathbf{u}|\mathbf{x}, G)$.*

Then, the solution $(\theta', q'_\phi(G))$ that maximizes $\mathcal{L}_{ELBO}(\theta, \phi)$ satisfies $q'_\phi(G) = \delta(G = G')$, where $G' = G^0$.

The proof of Proposition 2 can be found in Appendix D, which justifies performing causal discovery by maximizing ELBO of the N-ADMG model. Once the model has been trained and the ADMG has been recovered, we could use the trained N-ADMG to perform causal inference tasks, as detailed in appendix E.

D Proof of Proposition 2

To prove Proposition 2, we need the following lemma:

Lemma 5. *Assume a variational distribution $q_\phi(G)$ over a space of graphs $\mathcal{G}_\phi$, where each graph $G \in \mathcal{G}_\phi$ has a non-zero associated weight $w_\phi(G)$. With the soft prior $p(G)$ defined as Equation 10 and bounded $\lambda_1, \lambda_2, \rho, \alpha$, we have*

$$\lim_{N\to\infty} \frac{1}{N}\mathrm{KL}[q_\phi(G)\|p(G)] = 0. \tag{22}$$

Proof. This directly follows from Lemma 1 of (Geffner et al., 2022). □

Now we can proceed to prove Proposition 2:

Proof. For N-ADMG, in the infinite data limit $\mathcal{L}_{\text{ELBO}}$ becomes

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q_\phi(G)q(\mathbf{u}_n|G)}\left[\log p_\theta(\mathbf{x}_n, \mathbf{u}_n|G)\right] - \frac{1}{N}\sum_{n=1}^{N} H[q(\mathbf{u}_n|\mathbf{x}_n, G)] - \underbrace{\frac{1}{N}\mathrm{KL}\left[q(G)||p(G)\right]}_{\to 0} \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N}\sum_{G\in\mathcal{G}_\phi} w_\phi(G)\mathbb{E}_{q(\mathbf{u}|\mathbf{x},G)}\left[\log p_\theta(\mathbf{x}_n|\mathbf{u}_n, G)\right] - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q(G)}\left[\mathrm{KL}\left[q(\mathbf{u}_n|\mathbf{x}_n G)||p(\mathbf{u}_n|\mathbf{x}_n, G)\right]\right].
\end{aligned} \tag{23}$$

where the zeroing of the KL divergence follows from Lemma 5. Given fixed θ, the optimal posterior $q^*(\mathbf{u}_n|\mathbf{x}_n, G)$ satisfies $q^*(\mathbf{u}_n|\mathbf{x}_n, G) = p_\theta(\mathbf{u}_n|\mathbf{x}_n, G)$ due to the flexibility assumption. Thus,

$$\begin{aligned}
&\lim_{N\to\infty} \mathcal{L}_{\text{ELBO}}(\theta, \phi, q^*(\mathbf{u}_n|\mathbf{x}_n, G)) \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N}\sum_{G\in\mathcal{G}_\phi} w_\phi(G)\log p_\theta(\mathbf{x}_n|G) \\
&= \int p(\mathbf{x}; G^0) \sum_{G\in\mathcal{G}_\phi} w_\phi(G)\log p_\theta(\mathbf{x}|G)d\mathbf{x}
\end{aligned} \tag{24}$$

where $p(\mathbf{x}; G^0)$ denotes the data generation distribution with the ground truth ADMG, G^0. Let $(\theta^*, G^*) = \arg\max \int p(\mathbf{x}; G^0)\log p_\theta(\mathbf{x}|G)d\mathbf{x}$ be the MLE solution. Since $\sum_{G\in\mathcal{G}_\phi} w_\phi(G) = 1$, $w_\phi(G) > 0$, we have

$$\sum_{G\in\mathcal{G}_\phi} w_\phi(G)\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_\theta(\mathbf{x}|G)\right] \le \mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta^*}(\mathbf{x}; G^*)\right]$$

with the optimal value of $\sum_{G\in\mathcal{G}_\phi} w_\phi(G)\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_\theta(\mathbf{x}|G)\right]$ is achieved when every graph $G \in \mathcal{G}_\phi$ and associated parameter θ_G satisfies

$$\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta_G}(\mathbf{x}|G)\right] = \mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta^*}(\mathbf{x}|G^*)\right]. \tag{25}$$

Since the model is correctly specified, the MLE solution (θ^*, G^*) satisfies

$$\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta^*}(\mathbf{x}|G^*)\right] = \mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p(\mathbf{x}; G^0)\right]$$

Therefore, condition Equation 25 implies for every graph $G' \in \mathcal{G}_\phi$, $G' = G^0$ under the regularity condition; or equivalently, $\mathcal{G}_\phi = \{G' = G^0\}$. This proves our statement that $q'_\phi(G) = \delta(G = G')$, where $G' = G^0$.

□

E Estimating Treatment Effects

For all experiments we consider, the causal quantity of interest we wish to estimate is the expected average treatment effect (ATE), $\mathbb{E}_{q_\phi(G)}\left[\text{ATE}(\mathbf{a}, \mathbf{b}|G)\right]$, where the expectation is taken with respect to our learned posterior over causal graphs $q_\phi(G)$:

$$\mathbb{E}_{q_\phi(G)}\left[\text{ATE}(\mathbf{a}, \mathbf{b}|G)\right] = \mathbb{E}_{q_\phi(G)}\left[\mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{b}),G)}\left[\mathbf{x}_Y\right] - \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{b}),G)}\left[\mathbf{x}_Y\right]\right]. \tag{26}$$

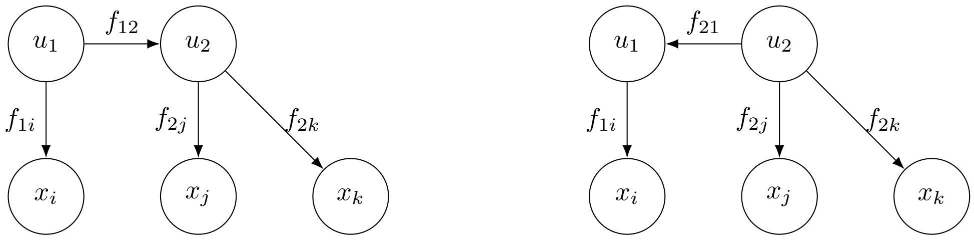
This requires samples from $p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T = \mathbf{b}), G) = p(\mathbf{x}_Y|\mathbf{x}_T = \mathbf{b}, G_{\text{do}(\mathbf{x}_T)})$, where $G_{\text{do}(\mathbf{x}_T)}$ is the 'mutilated' graph obtained by removing incoming edges into $\mathbf{x}_T$. We can achieve this by simulating the learnt SCM on $G_{\text{do}(\mathbf{x}_T)}$ whilst keeping $\mathbf{x}_T = \mathbf{b}$ fixed. Note that $q_\phi(\mathbf{u}|\mathbf{x})$ is not used to estimate the ATE; it suffices to sample $\mathbf{u}$ from the prior distribution, $p(\mathbf{u})$. In our setting, the inference net $q_\phi(\mathbf{u}|\mathbf{x})$ is only used as a means through which the likelihood of the data can be evaluated efficiently, and thus model parameters learned. This is in a similar spirit to deep generative models (Kingma & Welling, 2013).

F Additional Discussions on Structural Identifiability of Latent Variables

In Section 2, we argued that the identifiability of ADMG does not imply the structural identifiability of the magnified SCM. In this section, we will present more discussions on certain identifiability of latent structures (of magnified SCMs). In general, these examples demonstrate that for linear non-Gaussian ANMs the structure of latent variables can be refined beyond the assumption of latent confounders acting between pairs of non-adjacent observed variables, whilst the same techniques cannot achieve the same for non-linear ANMs.

F.1 Determining Causal Structure Amongst Latent Variables

Recently, Cai et al. (2019) demonstrated that it is possible to discover the structure amongst latent variables using their so-called Triad constraints. Their method is limited to the linear non-Gaussian ANM case. In this section, we demonstrate that an analogous constraint isn't available for non-linear ANMs.



(a) **(b)**

Figure 3: The two possible latent variable structures considered by Cai et al. (2019).

Consider the causal graphs shown in Figure 3a and Figure 3b.

Lemma 6 (Linear non-Gaussian identifiability)**.** *In the linear non-Gaussian ANM case, Equation 27 is satisfied only for the causal graph shown in Figure 3b. In the non-linear ANM case, Equation 27 is not satisfied for either Figure 3a or Figure 3b.*

$$\exists g, \quad x_i - g(x_j) \perp\!\!\!\perp x_k \tag{27}$$

Sketch of Proof. For the causal graph in Figure 3a, Equation 27 is equivalent to

$$\exists g \quad f_{1i}(u_1) + \epsilon_i - g(f_{2j}(f_{12}(u_1), u_2), \epsilon_j) \perp\!\!\!\perp f_{2k}(f_{12}(u_1), u_2) + \epsilon_k. \tag{28}$$

On the right, we have some function of the latent variables u_1 and u_2, $f_{2k}(f_{12}(u_1), u_2)$. On the left, we have some function of the same noise terms, $f_{1i}(u_1) - g(f_{2j}(f_{12}(u_1), \epsilon_j)$. To remove u_1 from both sides, we require g to be non-zero and so a term including u_2 is still present. To remove u_2, we again require g to be non-zero and so a term including u_1 is still present. Thus, Equation 27 is not satisfied in either the linear non-Gaussian or non-linear ANM case.

For the causal graph in Figure 3b, Equation 27 is equivalent to

$$\exists g \quad f_{1i}(f_{21}(u_2), u_1) + \epsilon_i - g(f_{2j}(u_2) + \epsilon_j) \perp\!\!\!\perp f_{2k}(u_2) + \epsilon_k. \tag{29}$$

In the linear case, we can construct a linear g to remove u_2 fom the left side so that Equation 27 holds (i.e. $g = \frac{f_{1i} f_{21}}{f_{2j}}$). In the non-linear case, u_1 and u_2 are coupled in the leftmost term and cannot be removed by a term involving u_2 and ϵ_j. Hence, Equation 27 is violated. □

F.2 Determining the Number of Latent Confounders

Here, we consider whether the number of latent confounders can be determined in the non-linear ANM case. Lemma 7 shows that in the linear non-Gaussian case, the number of latent confounders acting between a triplet of confounded observed variables can be determined (by verifying certain constraints on marginal distributions on observed variables), whilst the same approach cannot be used in the non-linear ANM case.

Consider the two causal graphs shown in Figure 4a and Figure 4b.

Lemma 7 (Linear non-Gaussian identifiability)**.** *In the linear non-Gaussian ANM case, Equation 30 is satisfied only for the causal graph shown in Figure 3b. In the non-linear ANM case, Equation 30 is not satisfied for either Figure 3a or Figure 3b.*

$$\exists g, \quad x_i - g(x_j) \perp\!\!\!\perp x_k. \tag{30}$$

Sketch of Proof. For Figure 4b, Equation 30 is equivalent to

$$\exists g \quad f_i(u) + \epsilon_i - g(f_j(u) + \epsilon_j) \perp\!\!\!\perp f_k(u) + \epsilon_k. \tag{31}$$

In the linear non-Gaussian case, it is straightforward to set $g = \frac{f_i}{f_j}$ to remove the common noise term u from the left term and make the two sides independent. In the non-linear case, when $f_i \neq f_j$ the common noise term u cannot be removed from the left as g must be non-linear, and thus produces a term that involves both u and ϵ_j. For Figure 4a, Equation 30 is equivalent to

$$\exists g \quad f_i(u_1, u_2) + \epsilon_i - g(f_j(u_1, u_3) + \epsilon_j) \perp\!\!\!\perp f_k(u_2, u_3) + \epsilon_k. \tag{32}$$

u_2 cannot be removed from the left, and so Equation 30 does not hold in either the linear non-Gaussian or non-linear case. □

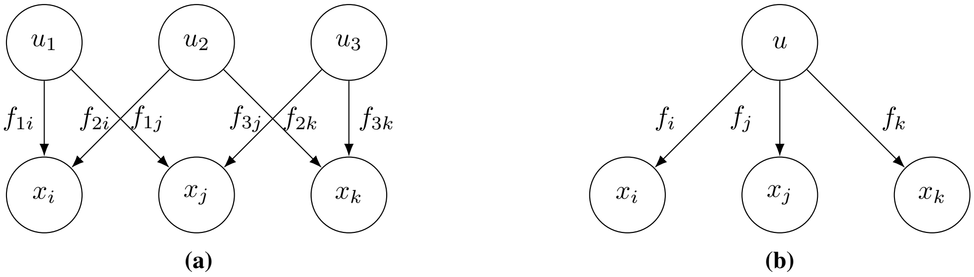


Figure 4: Two possible latent structures that confound each pair of variables x_i, x_j and x_k.

G Optimisation Details

G.1 Optimisation Details for N-ADMG

As discussed in Section 4, we gradually increase the prior hyperparameters ρ and α throughout training. This is done using the augmented Lagrangian procedure for optimisation (). The optimisation process interleaves two steps: 1) optimise the objective for fixed values of ρ and α for a certain number of steps; and ii) update the values of ρ and α. Steps i) and ii) and ran until convergence, or the maximum number of optimisation steps is reached. We describe these two steps in more detail below.

Step i). The objective is optimised for some fixed vales of ρ and α using Adam (). We use a learning rate of 10^{-3} for the model parameters and 5×10^{-3} for the variational parameters. We optimise the objective for a maxmimum of 5000 steps or until convergence (we stop early if the loss does not improve for 1500 optimisation steps, moving to step ii)). During training, we reduce the learning rate by a factor of 10 if the training loss does not improve for 1000 steps a maximum of two times. If we reach the condition a third time, we assume optimisation has converged and move to step ii). We apply annealing to the KL-divergence between the approximate posterior and prior over the latent variables. The annealing contant is fixed for each step i), and increased linearly over the first optimisation loops.

Step ii). We initialise $\rho = 1$ and $\alpha = 0$. At the beginning of step i) we measure the DAG / bow-free penalty $P_1 = \mathbb{E}_{q_\phi(G)}\left[h(G)\right]$. At the beginning of step ii), we measure this penalty again, $P_2 = \mathbb{E}_{q_\phi(G)}\left[h(G)\right]$. If $P_2 < P_1$, we leave ρ unchanged and update $\alpha \leftarrow \alpha + \rho P_2$. Otherwise, if $P_2 \geq 0.65 P_1$, we leave α unchanged and update $\rho \leftarrow 10\rho$. We repeat the steps i) to ii) a maximum of 30 times or until convergence (measured as α or ρ reaching some max value which we set to 10^3 for both).

G.2 Additional Hyperparameters

Prior Hyperparameters. We use the sparsity inducing prior hyperparameters $\lambda_{s,1} = \lambda_{s,2} = 5$.

ELBO approximation. We construct an approximation to the ELBO in Equation 8 using a single sample from the approximate posteriors. For evaluating the gradients of the ELBO we use the Gumbel softmax method with a hard forward pass and soft backward pass with temperature of 0.25.

Neural network architectures. The functions ξ_1, ξ_2 and ℓ used in the likelihood and the inference network used to parameterise $q_\phi(\mathbf{u}|\mathbf{x})$ are all two hidden layer MLPs with 80 hidden units per hidden layer.

Non-Gaussian noise model. For the non-Gaussian noise model

ATE estimation. For ATE estimation we compute expectations by drawing 1000 graphs from the learnt posterior, and for each graph we draw two samples of $\mathbf{x}_Y$ for a total of 2000 samples which we used to form a Monte Carlo estimate.

H Dataset Details

H.1 Synthetic Fork-Collider Dataset

We constructed a 2000 sample synthetic dataset with the causal structure shown in Figure 1 by sampling from the following SEM:

$$\begin{aligned}
[u_1,\ u_2,\ \epsilon_1,\ \epsilon_2,\ \epsilon_3,\ \epsilon_4,\ \epsilon_5]^T &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)\\
x_1 &= \epsilon_1\\
x_2 &= \sqrt{6}\exp(-u_1^2) + 0.1\epsilon_2\\
x_3 &= \sqrt{6}\exp(-u_1^2) + \sqrt{6}\exp(-u_2^2) + 0.2\epsilon_3\\
x_4 &= \sqrt{6}\exp(-u_2^2) + \sqrt{6}\exp(-x_1^2) + 0.1\epsilon_4\\
x_5 &= \sqrt{6}\exp(-x_1^2) + 0.1\epsilon_5.
\end{aligned} \tag{33}$$

Variables u_1 and u_2 are latent confounders acting on variable pairs x_2 and x_3, and x_3 and x_4, respectively.

H.2 Latent confounded ER Dataset

We generate synthetic datasets from ADMG extension of *Erdős-Rényi* (ER) graph model (Lachapelle et al., 2019; Zheng et al., 2020). An **ER**(d, e, m) dataset is generated according the following procedures:

1. Generate a $d \times d$ directed adjacency matrix G_D of a ADMG from *Erdős-Rényi* (ER) graph model, whose expected directed edges equal to e;

2. Simulate $d \times d$ bidirected adjacency matrix G_B via random Bernoulli sampling, whose expected number equals to m;
3. Simulate exogenous noises ϵ_i from a zero mean Gaussian distribution with standard deviation of 0.1;
4. Simulate latent variables $\mathbf{u}$ from a zero mean Gaussian distribution with standard deviation of 0.1;
5. Simulate each observed variable as $x_i = f_i(\mathbf{x}_{\text{pa}(i;G_D)}) + g_i(\mathbf{u}_{\text{pa}(i;G_B)}) + \epsilon_i$, where f_i, g_i are randomly sampled nonlinear functions of the form: $y = \mathbf{w}^T \ e^{-\mathbf{x}^2}$.
6. Remove $\mathbf{u}$ from the sampled dataset.

H.3 IHDP dataset details

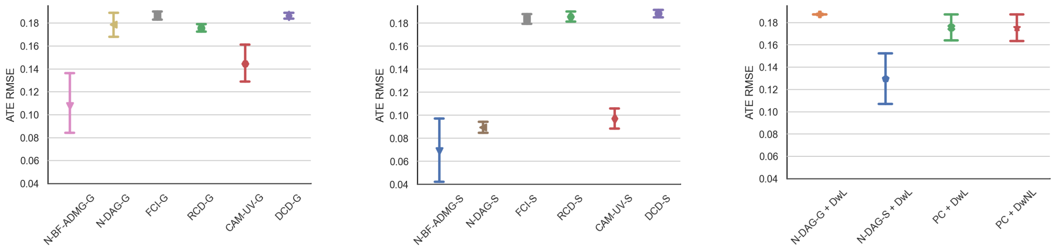
This dataset contains measurements of both infants (birth weight, head circumference, etc.) and their mother (smoked cigarettes, drank alcohol, took drugs, etc) during real-life data collected in a randomised experiment. The main task is to estimate the effect of home visits by specialists on future cognitive test scores of infants. The outcomes of treatments are simulated artificially as in Hill (2011); hence the outcomes of both treatments (home visits or not) on each subject are known. Note that for each subject, our models are only exposed to only one of the treatments; the outcomes of the other potential/counterfactual outcomes are hidden from the mode, and are only used for the purpose of ATE evaluation. To make the task more challenging, additional confoundings are manually introduced by removing a subset (non-white mothers) of the treated children population. In this way we can construct the IHDP dataset of 747 individuals with 6 continuous covariates and 19 binary covariates. We use 10 replicates of different simulations based on setting B (log-linear response surfaces) of Hill (2011), which can downloaded from `https://github.com/AMLab-Amsterdam/CEVAE`. We use a 70%/30% train-test split ratio. Before training our models, all continuous covariates are normalised.

I Additional Results

I.1 Causal inference on Infant Health and Development Program (IHDP) Dataset

For the real-world datasets, we evaluate treatment effect estimation performances on infant health and development program data (IHDP). This dataset contains measurements of both infants and their mother during real-life data collected in a randomized experiment. The main task is to estimate the effect of home visits by specialists on future cognitive test scores of infants, where the ground truth outcomes are simulated as in (Hill, 2011). To make the task more challenging, additional confoundings are manually introduced by removing a subset (non-white mothers) of the treated children population. More details can be found in Appendix H.3. We first perform causal discovery to learn the underlying ADMG of the dataset, and then perform causal inference. Since the true causal graph is unknown, we evaluate the causal inference performance of each method by estimating the ATE RMSE.

Apart from the aforementioned baselines, here we introduce four more methods: PC-DWL (PC algorithm for discovery, DoWhy (Sharma et al., 2021) linear adjustment for inference); PC-DwNL (PC for discovery, DoWhy double machine learning for inference); N-DAG-G-DwL (N-DAG Gaussian for discovery, linear adjustment for inference); and N-DAG-S-DwL (N-DAG Spline for discovery, linear adjustment for inference). Results are summarized in Figure 5a-Figure 5c. Generally, models with non-Gaussian exogenous assumptions tend to have lower ATE estimation errors; while models with linear assumptions (RCD and DCD) have the worst ATE RMSE. Interestingly, the DoWhy-based plug-in estimators tend to worsen the performances of SCM models. However, regardless of the assumptions made on exogenous noise, our method (N-BF-ADMG-G and N-BF-ADMG-S) consistently outperforms all other baselines with the same noise. It is evident that for causal inference in real-world datasets, the ability of N-BF-ADMG to handle latent confoundings and nonlinear causal relationships becomes very effective.



(a) Gaussian exogenous noise.

(b) Spline exogenous noise.

(c) DoWhy baselines.

Figure 5: Causal inference results for the IHDP dataset. The figure shows mean ± standard error results across five random initialisations.