Non-linear Causal Discovery for Additive Noise Model with Multiple Latent Confounders*

Xuanzhi Chen[†], **Wei Chen**[†], **Ruichu Cai**
School of Computer Science
Guangdong University of Technology
Guangzhou, China
xuanzhichen.42@gmail.com

Abstract

Could we teach AI in brain science spectrum to manoeuvre causation via the specific identification entailed by data? How should we further appreciate "causal structures" underneath the data in a complicate learning environment? An environment in which "generic data relations" are prone to be non-linear, and even impacts from the multiple unknown factors are persisting. Existing solutions towards the issue—non-linearity identification with latent confounding—might be either theoretically elusive in formal representation or notoriously difficult in algorithmic computation. Such motivations have driven us to a theory-guided and effective causal discovery algorithm. Concretely, with the existence of multiple latent confounders, we re-analysis the celebrated non-linear Additive Noise Models, and discovered an identification which specifies how the varying degrees of latent confounding are essentially raised by "unobserved parents". Ultimately the identification gives rise to the well-performance algorithm on functional magnetic resonance imaging (fMRI) brain data, bringing our hope that, in practice, the intelligent agents seeking for causation are equipped with vigorous counteraction against the data complexities. Mathematical proof is provided in the supplementary material [link] and Python implementations are open-source in Github [link].

1 Introduction

Causal discovery, namely to recover data generation mechanism represented (vaguely) by causal graphs, is an emerging scientific discipline to spot causal significance from the merely observed data. As for quick starters, we figuratively describe our primary motivation by considering the identification over a non-linear additive-noise-model denoted as "cause-and-effect" $C \to E$. Obviously, $C \to E$ cannot be methodologically identified, if both of their parent are unobserved (e.g. $\overline{pa}_C$ and $\overline{pa}_E$), amounting to an unobserved common cause. **The question is**, whether $C \to E$ keeps identifiable if only **one side** of the parent is unobserved and even triggers the indirected confounding (e.g. $C \leftarrow pa_C \leftarrow \overline{pa}_E \to E$)?

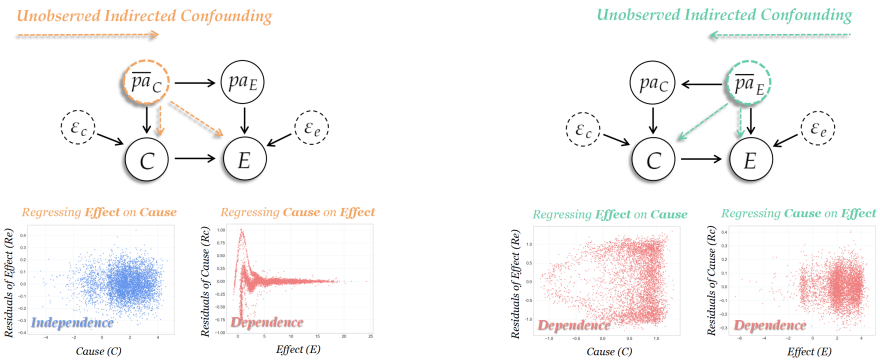

Figure 1: Intuitions of non-linear causal identification under indirected latent confounding.

*This is a complete work that should has been scheduled for submission in 2023, but collaboration between Xuanzhi Chen and other authors came to an early cessation, due to the personal inconvenience of Xuanzhi Chen.

[†]Equal contribution.

Leveraging the popular methodology [P. Hoyer et al. (2008)][Peters, Janzing, and Schölkopf (2017)], namely the regression and the independence test, two illustrated examples in Figure (1) imply the procedure of causal identification after the standard "deconfounding" methodology. The "deconfounding" procedure results in residuals of C and E (denoted as R_C and R_E in followings) by regressing on all the hypothetical observed parents respectively (including C and E). Consequently, the statistical asymmetry, as shown in previous literature, indicates that only if the independence between R_E and C whereas the dependence between R_C and E occurring after "deconfounding", could the direction of "cause-and-effect" $C \to E$ keep identifiable.

"Deconfounding" deserves further discussion. In practice, only a subset of variables relative to a reasonable system could be measured, posing significant challenges on distinguishing causal directions in the circumstance with hidden variables. Specifically, if considering the latent confounding triggered by unobserved parents $\overline{pa}$, the explanation exhibits an intuitive side when we draw comparison with the methodology in linear circumstances: No matter which latent confounder raised by the unobserved parents $\overline{pa}_C$ or $\overline{pa}_E$, the causal information "flow" alongside the indirected confounding path ending up between C and E will be "blocked" by one of their observed parents (e.g. $C \leftarrow pa_C \leftarrow \overline{pa}_E \to E$ is blocked by pa_C; $C \leftarrow \overline{pa}_C \to pa_E \to E$ is blocked by pa_E). Such blocking behaviors entirely "blocks" the information flow with respect to the confounding path, which means being capable of deconfounding by linear regression methodologically. This strategy, unfortunately, could not pay off in terms of non-linear functions, due to the variables' non-linear interaction compromising the effect of regression (Section 3).

Our work in this paper, however, still endeavour to contribute a leeway for appropriately identify causal direction from the data that involves non-linearity and latent confounding. Briefly, **the conclusion** for the above question suggests that the direction of "cause-and-effect" $C \to E$ is able to continuously keep identifiable, only if the confounding is triggered by the unobserved parent $\overline{pa}_C$ rather than $\overline{pa}_E$. Equipping with the theoretical conclusion, we further proposed a "hybrid" causal discovery method that will wisely utilize the regression-independence-test methodology to reach the practical efficiency. Details about the theory and algorithm, along with the experimental results, will be elaborated using the field-related language in the rest of the main part of the paper. **Also see** a relatively light discussion at the end of the article for general ideas from one of the author Xuanzhi Chen.

2 Related Work

Classical causal discovery approaches can be divided into two primary categories: constraint-based and functional-based. In this section, we will briefly review methods with and without the causal sufficiency.

Conventional constraint-based methods, such as the SGS algorithm and the PC algorithm [Spirtes, C. N. Glymour, et al. (2000)], resort to the (conditional) independence test among variables to first identify a skeleton and further determine causal directions by utilizing V-structures and specific rules. Other methods, such as GES [Chickering (2002)], performs the greedy equivalent search and optimizes the BIC score of associating Bayesian networks. These methods, however, inevitably suffer a problematic issue of Markov equivalence class (MEC). Over the last decades, the emerging functional-based methods are on an attempt to assume data generation process based on restricted functional causal models (FCM). Provided the linear additive models with independent non-Gaussian noise (LiNGAM), the ICA-LiNGAM [Shimizu, P. O. Hoyer, et al. (2006)] method optimizes the linear causal transformation based on the independent component analysis(ICA). The Direct-LiNGAM [Shimizu, Inazumi, et al. (2011)] approach, as a variant of ICA-LiNGAM, unravels a causal order by iteratively identifying the most exogenous or the most endogenous variables after regression. Beyond the non-Gaussian noise assumption that entails causation, functional causal models such as ANMs [P. Hoyer et al. (2008)] and CAMs [Bühlmann, Peters, and Ernest (2014)], implying the non-linearity of specific functional classes, can as well exhibit causal asymmetric for identification.

Nevertheless, the above methodology is confined to the causal insufficiency, namely the presence of latent confounders. To alleviate this, constraint-based methods such as the FCI algorithm [Spirtes, C. N. Glymour, et al. (2000)] introduce the notion of partial ancestral graphs (PAG) for causal identification with the sophisticated graphical representation, whereas limitations of the MEC issue are still existing. Functional-based methods attempting to fine-tune the existing basics are similarly facing the challenges for both linear and non-linear cases. In linear cases, the proposed overcomplete ICA [Lewicki and Sejnowski (2000)] approach might suffer local optimum; the pairwise LvLiNGAM [Tashiro et al. (2012)] and Parce-LiNGAM [Tashiro et al. (2014)] approaches demands unacceptable computational time; recent methods such as MLC-LiNGAM [Chen et al. (2021)] and RCD [Maeda and Shimizu (2020)] are rely on the linear setting to advance their strength. In terms of non-linear cases, the application of the ICAN method [Janzing et al. (2012)] is restricted to pairwise variables; the causal identifiability of the CAM-UV approach [Maeda and Shimizu (2021)] might focus more on avoiding false causal inference instead of providing deterministic identification statements.

3 Model, Assumption, and Causal Identification Theory

Let $\mathcal{G}_{\boldsymbol{X}}$ denotes as the directed acyclic graphs (DAG) of variables $\boldsymbol{X} = \{x_1, x_2, \ldots, x_d\}$, with i.i.d. noises $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d\}$. Presuming the additive-noise-models (*ANMs*) [P. Hoyer et al. (2008)], the causal models $< \mathcal{G}_{\boldsymbol{X}}, \boldsymbol{\varepsilon}, \boldsymbol{f}_{\boldsymbol{X}} >$ generated by **Non-linear** functions $\boldsymbol{f}_{\boldsymbol{X}}$ with **M**ultiple **L**atent **C**onfounders, can be formalized in this paper—in form of the *ANMs* generation procedure of following pair $x_j \to x_i$:

$$x_i := \sum_{x_j \in \boldsymbol{pa}_i} f_{ij}(x_j) + \xi_i. \tag{1}$$

Denoting as ***Nonlinear-MLC***, we specify that the "latent confounding" from "unobserved parents" $\overline{\boldsymbol{pa}}$ (observed parents $\boldsymbol{pa}$ analogically) is incorporated to the **extensive noises** $\xi_i := \varepsilon_i \cup \boldsymbol{f}(\overline{\boldsymbol{pa}}_i)$. Additionally mild assumptions required in this article (Section 4.2) are listed as follows:

A-1 Markov Assumption: Independence yielded by $\mathcal{G}_{\boldsymbol{X}}$ is consistent with ones over distributions $P_{\boldsymbol{X}}$.

A-2 Faithfulness Assumption: Distributions $P_{\boldsymbol{X}}$ faithfully encode independence entailed only by $\mathcal{G}_{\boldsymbol{X}}$.

Traditionally, discovering $\mathcal{G}_{\boldsymbol{X}'}$ over a subset $\boldsymbol{X}' \subset \boldsymbol{X}$ might draw on expunging causal functions' effects by linear regression and thus revealing the testable independence noise. But contrasting with linear combinations, the causal model *Nonlinear-MLC* by Equation(1) **cannot** be expanded as:

$$x_i := \sum_{\varepsilon_k \in \boldsymbol{\varepsilon} \setminus \{\varepsilon_i\}} \phi(\varepsilon_k) + \varepsilon_i. \tag{2}$$

Composite non-linear functions $\phi := f(f(...))$ relative to independent noises **cannot** be accessible, in the sense that "composite *ANMs*" do not hold with embedded latent noises $\boldsymbol{\varepsilon}_{\overline{\boldsymbol{pa}}}$, leading to the infeasibility to expunge non-linear effects from within. (e.g. "endogenous dependence $\boldsymbol{\varepsilon}_{\overline{\boldsymbol{pa}}}$" will compromise regression)

Aiming at mitigating this issue, we proposed Lemma 1, arguing as the **latent additive noise models** (***L-ANMs***), to stipulate a novel i.i.d. identifiable condition for the *Nonlinear-MLC* models:

Lemma 1. *Assuming data generation procedures are consistent with Equation (1), the pairwise cause-and-effect $C \to E$ among (**multiple unobserved**) pairs $\boldsymbol{C}^* \to E$ is identifiable if and only if*

$$(\xi_E \perp\!\!\!\perp C) \wedge (\xi_E := \varepsilon_E \cup \boldsymbol{f}(\boldsymbol{C}^*)) \tag{3}$$

is satisfied, where other multiple unobserved causes $\boldsymbol{C}^*$ are denoted as $\boldsymbol{C}^* := \boldsymbol{C} \backslash C = \overline{\boldsymbol{pa}}_E$.

Contributions of the *L-ANMs* Lemma are **three-fold**: **(i)** It is targeted for non-linearity, not necessarily a compulsion for linear cases. **(ii)** A succintly deterministic identification condition, compared to notions in previous work (e.g. "C-ANMs" [Cai et al. (2019)] or "CAM-UV" [Maeda and Shimizu (2021)]). **(iii)** Akin to the well-known identification condition $((\varepsilon_E \perp\!\!\!\perp C) \wedge (\varepsilon_C \perp\!\!\!\perp E \mid C))$ by **Independence Causal Mechanism (ICM)** [Peters, Janzing, and Schölkopf (2017)], Lemma *L-ANMs* permits $(\xi_C \not\perp\!\!\!\perp E \mid C)$, which is essentially how we characterize the latent confounding by unobserved parents in this paper.

Figure 2: Graphical intuitions of Lemma 1 (*Latent-ANMs*) with the symbol "o-o" characterizing the uncertain causal directions "->" or "<-". (a) The structure involving relations between interested cause C and multiple unobserved causes $\boldsymbol{C}^*$ given the effect E. (b) The general *Nonlinear-MLC* model with respect to the structure in (a), where U summarizes the rest of observed variables.

Diving deeper in algorithm design (Section 4), *L-ANM* inspires a seeking for **empirical regressor** $\mathcal{R}_i$ to discover independence between ξ_i and x_j (similar to the binary case ξ_E and C shown in Lemma 1). A scratch of proof with slight algebra in Equation(1) shows:

$$x_i - \underbrace{\sum_{x_h \in \boldsymbol{pa}_i \setminus \{x_j\}} f_{ih}(x_h)}_{\mathcal{R}_i} = f(x_j) + \underbrace{\left(\sum_{x_k \in \overline{\boldsymbol{pa}}_i} f_{ik}(x_k) + \varepsilon_i \right)}_{\xi_i}. \tag{4}$$

To summarize, the essence of *L-ANMs* Lemma reveals a condition that prevents *ANMs* from violations, thus generalizing the *ANMs* identification into an extensive range with presence of latent confounders.

4 A Theory-based Novel Algorithm for Causal Discovery

4.1 Recognize the Graphical Pattern: Maximal Cliques

Existing algorithms [Maeda and Shimizu (2021)][Tashiro et al. (2014)] may stuck in undetermined dependence of variable subsets. The "undetermined dependence" within variable subsets, from the graphical perspective, might indicate the "undirected connection" within maximal cliques $\mathcal{M}$. One might also notice that, in the view of algorithms amidst their computing memory, the undetermined parent-relations amounts to **multiple unobserved parents**. Thus lining up with Lemma 1, fully depending structures in $\mathcal{M}$ (including C and E) limit the anticipant independence between C^* and C, leading to the following corollary drawing on **empirical regressor** $\mathcal{R}$ to counteract that dependence.

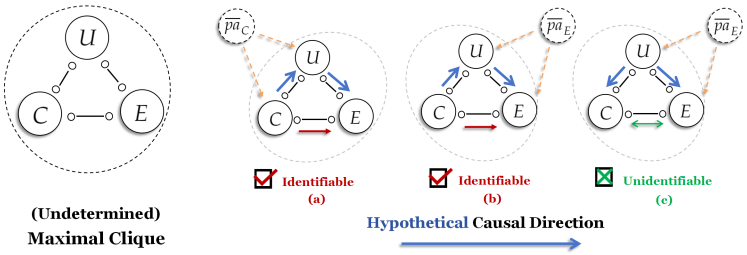

Figure 3: A toy graphical structure, namely $\mathcal{M}_{C,E} = \{C, E, U\}$, illustrates how to determine non-linear identifiability under latent confounding ($\overline{pa}_C$ or $\overline{pa}_E$) by applying Corollary 1 in cases (a), (b), and (c).

Corollary 1. *Assuming data generation procedures are consistent with Equation (1), the pairwise cause-and-effect $C \to E$ over a* ***maximal clique*** *$\mathcal{M}$ is identifiable if and only if*

$$(E - \mathcal{R}_E(\mathcal{M}^*) \perp\!\!\!\perp C) \ \wedge \ (\mathcal{M}^* := \mathcal{M}_{C,E} \backslash \{E\}) \tag{5}$$

is satisfied, where $\mathcal{M}_{C,E}$ represents all observed variables including C and E within a maximal clique.

4.2 Incorporate the Two-Steps Framework of Hybrid Methodology

We further apply the novel identification (implies by Corollary 1) on the maximal cliques $\mathcal{M}$ partitioned over causal skeleton $\mathcal{S}_{\boldsymbol{X}'}$, which stands on the basic ground provided by the PC algorithm [Spirtes, C. N. Glymour, et al. (2000)], along with the algorithm's identification guaranteed by Lemma 2.

Lemma 2. *Suppose that assumptions A1 and A2 hold, every true adjacency pair of variables x_i and x_j in $\mathcal{G}_{\boldsymbol{X}}$ is in accord with the estimated adjacency pair in causal skeleton $\mathcal{S}_{\boldsymbol{X}'}$ of $\mathcal{G}_{\boldsymbol{X}'}$ using PC algorithm.*

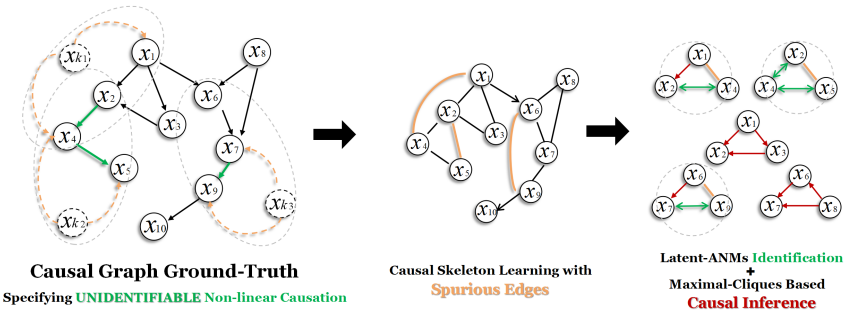

Figure 4: A two-steps method with the spurious edges detecting latent confounders [Chen et al. (2021)].

Algorithm 1 Nonlinear-MLC Algorithm

Require: Data $\boldsymbol{X'} = \{x_1, ..., x_m\}(m < d)$, significant level α
Ensure: Estimated causal graph $\hat{\mathcal{G}_{\boldsymbol{X}'}}$

$\hat{\mathcal{S}_{\boldsymbol{X}'}}, \hat{\mathcal{G}_{\boldsymbol{X}'}} \leftarrow stage1CausalDiscovery(\boldsymbol{X'}, \alpha)$, search $\leftarrow$ True;
while search **do**
  $\hat{\mathcal{G}_{\boldsymbol{X}'}} \leftarrow stage2CausalDiscovery(\boldsymbol{X'}, \alpha, \boldsymbol{\mathcal{M}}(\hat{\mathcal{S}_{\boldsymbol{X}'}}), \hat{\mathcal{G}_{\boldsymbol{X}'}})$, search $\leftarrow$ False;
  if $determinedNewDirections(\hat{\mathcal{G}_{\boldsymbol{X}'}})$ **then**
    search $\leftarrow$ True;
  end if
end while
return($\hat{\mathcal{G}_{\boldsymbol{X}'}}$)

5 Experiments

We use the following causal discovery algorithms as the baseline methods: PC [Spirtes, C. N. Glymour, et al. (2000)], FCI [Spirtes and C. Glymour (1991)], RESIT [Peters, Mooij, et al. (2014)], and CAM-UV [Maeda and Shimizu (2021)]. PC is a constraint-based approach assuming causal sufficiency. Accordingly, FCI servers as an extension of PC algorithm, applying to causal inference with latent confounders. RESIT and CAM-UV are categorized to functional-based approaches. As a variant of DirectLiNGAM [Shimizu, Inazumi, et al. (2011)], RESIT assumes that non-linear additive models hold as for the data generation without presence of latent confounders. CAM-UV [Maeda and Shimizu (2021)], however, further assumes the existence of (general) unobserved variables, tending to avoid the incorrect causal inference. The usages of the baseline methods stated above refer to the python package causal-learn[3].

We use precision, recall, and F1 score as the evaluation indicators for the estimated causal graphs reconstructed by different algorithms. Amidst the experiment, notice that we only extracted directed edges from the adjacency matrix or directly obtained causal pairs for calculating the indicators.

5.1 Performance on Functional Magnetic Resonance Imaging (fMRI) Data

We tested the *Nonlinear-MLC* algorithm for the simple application in related fields of neuroscience, performing causal discovery to established fMRI brain dataset[4] that is on the mathematical basics of (non-linear) dynamic causal models [Friston, Harrison, and Penny (2003)]. We selected the fMRI-dataset (sim3) that entails causal interactions among 15 distinct spatial regions (Regions of Interest, ROI), and characterizes the temporal signals sampled from individuals. We further reconstructed the non-temporal dataset via random sampling by giving a proper width-fixed time window. The goal is to discovery the causal structures (e.g. the mapping networks based on brain functions) over brain regions (denoted as the variable X_i) under the circumstances with omitted variables [5] (shown in Figure 5[6])

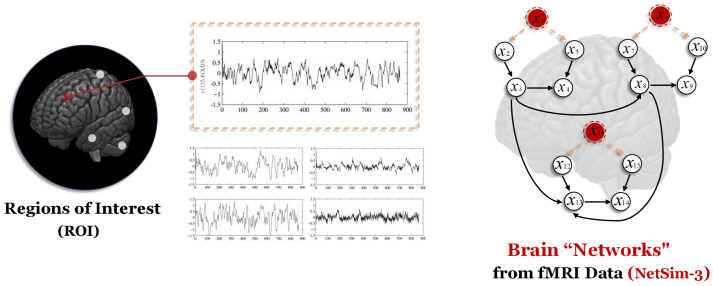

Figure 5: Illustration of the causal structure (with latent confounders) with respect to ROI based on fMRI data. Omitted regions (namely the latent confounder) are marked as red color.

Specifically, we first prioritized an increasing sequence of variables associating with brain regions (e.g. x_0, x_5, and x_{10}) as latent confounders by omitting them from original dataset. Then the dataset was processed by sampling with a size of 1000 from randomly selective 5 individuals, given width-fixed time windows with length of 200. Ultimately, we performed causal discovery approaches to the dataset.

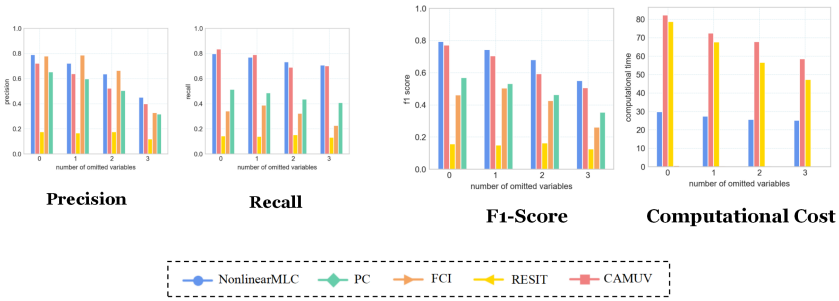

Figure 6: Performance evaluations (precision, recall, f1-score) on fMRI-dataset (sim3).

[3] https://causal-learn.readthedocs.io/en/latest/

[4] https://www.fmrib.ox.ac.uk/datasets/netsim/index.html

[5] Experiment settings virtually mirror the previous work [Maeda and Shimizu (2021)], whereas we primarily consider high dimension causal inference with the variable omitting that renders the latent confounders (instead of the latent intermediates).

[6] Figurative descriptions in Figure 5 are partially referred to the course Introduction to FSL (Andrew Jahn) and the literature [Minati et al. (2015)]

Notice that the relative simplicity of causal structures implied by fMRI data reduces *Nonlinear-MLC*'s advanced maximal-clique-based causal inference (described in Section 4.1) into the directed pairwise causal inference. Meanwhile, the CAM-UV algorithm might infer more hypothetical causal connections (including the redundant connections) without the foundation of the causal skeleton, which rendered its performance with marginally higher recalls but lower precision than the *Nonlinear-MLC* algorithm. Despite of the well-perform precision by the FCI algorithm, it actually determined a small fraction of causal directions that contributes little to the recall (illustrated in Figure 6).

Hence, **we conclude** that performance of the *Nonlinear-MLC* algorithm inclines to a slight advantage in the comprehensive F1 score with practicably lower computational time (exhibited in Fig 6).

5.2 Performance on Simulated Causal Models (Non-linear MLC) Data

We simulated the data generation via first obtained a random DAG in light of the Erdős–Rényi model [Erdős, Rényi, et al. (1960)]. According to a omitted number that specified latent confounders, we then determined the unobserved variables and (evenly) distributed their range of confounding across all observed variables. Provided a topological order converted by the DAG, we consequently simulated each observed variable x_i (in the topological order) by summarizing the effects of both its observed parents $\boldsymbol{pa_i}$ and unobserved parents $\overline{\boldsymbol{pa}}_i$, along with an externally random noise term ε_i, namely

$$x_i = x_i'/std(x_i'), \quad x_i' := \sum_{x_j \in \boldsymbol{pa_i}} \rho_1 \cdot f_{ij}(x_j) + \sum_{x_k \in \overline{\boldsymbol{pa}}_i} \rho_2 \cdot f_{ik}(x_k) + \varepsilon_i, \tag{6}$$

where the symbol $std(\cdot)$ denotes the standard deviation of x_i'. The numerical coefficients ρ_1 and ρ_2 were randomly taken from Uniform(0.3, 0.5) and Uniform(0.6, 0.8) respectively; the types of the non-linear functions f were randomly chose from $\{f(\cdot) \mid sin(\cdot),\ \sqrt{\cdot},\ (\cdot)^3\}$; the external noise ε_i was randomly sampled from Uniform(-10, 10). Since the assumption A-2 (in Section 3) is vulnerably to be violated in practice, **we must highlight** herein that we have heuristically filtered the groups of simulated data that led to low recall of causal skeletons by applying the PC algorithm beforehand (since a low skeleton recall implies weak encodings of (conditional) independence entailed by $\mathcal{G}_{\boldsymbol{X}}$ and further violates A-2).

Fixing the sample size n of 1000 and the dimension d (of observed variables) of 10, we test the average perform of the *Nonlinear-MLC* algorithm by running 50 times of experiments with the variation of omitted variables numbers (from 0 to 3). Plus, we further tested the algorithm's sensitivity in regard of two cases: $n \in [500, 1000, 1500]$ and $d \in [5, 10, 15]$. We set the same ratio of latent confounder as 0.2, and fixed $d = 10$ and $n = 1000$ respectively for the two cases. (Figuratively shown in Figure 7; Table 1)

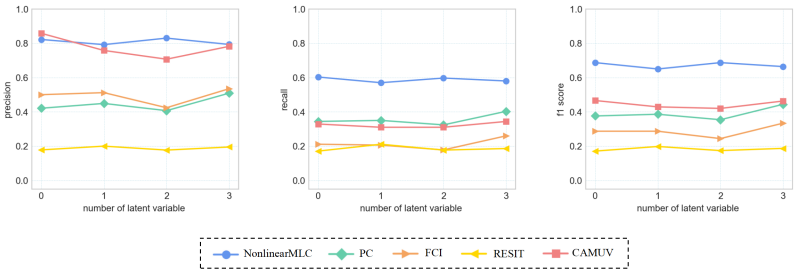

Figure 7: Performance evaluations on simulated data with different numbers of latent confounders.

Table 1: Sensitivity as to samples and dimensions, along with associating computational cost.

Algorithm	F1 Score						Computational Time			
	Sample Size			Dimension			Number of Latent Confounder			
	500	1000	1500	5	10	15	0	1	2	3
PC	0.347	0.385	0.421	0.322	0.385	0.372	0.05	0.07	0.05	0.01
FCI	0.217	0.261	0.473	0.304	0.261	0.197	0.18	0.21	0.17	0.17
RESIT	0.152	0.169	0.174	0.277	0.169	0.124	29.78	29.73	29.82	29.86
CAMUV	0.265	0.374	0.586	0.625	0.374	0.419	14.7	17.59	15.14	15.16
NonlinearMLC	**0.623**	**0.661**	**0.735**	**0.851**	**0.661**	**0.627**	9.42	10.62	11.26	11.67

Figure 5.2 illustrates the average performance of *Nonlinear-MLC* compared with baseline methods. Except for the case of causal sufficiency—precision of our method is slightly lower than the CAM-UV algorithm—*Nonlinear-MLC* outperforms others in presence of latent confounders. Table 5.2 further demonstrates that our method is robust against the changes as to different sample sizes and dimensions.

6 Discussion

Since the schedule of publishing this paper was eventually cancelled, the light discussions in this section were additionally listed by one of the author Xuanzhi Chen. The following question-oriented discussion will specify supplemental perspectives as for this paper, by reviewing some of the equally important ideas (from Xuanzhi's point of view) during his journey of finishing the work.

- **Does the work in this paper truly tackle the issue of "Multiple Latent Confounders"?**

  No quite, I have to admit. Initially I just wanted to "extend" the repertoire of our previous work MLC-LiNGAM [Chen et al. (2021)], an causal discovery algorithm serves in a linear spectrum, by utilizing the conventional (non-linear) additive noise models (ANMs). To this end, I kept that abbreviation "MLC" (Multiple Latent Confounders) for echoing the series of our work.

  However, this might cause a slight exaggeration for the *Nonlinear-MLC* algorithm in this paper because I gradually found that non-linearity in causal inference is tricky than what I have imagined. The "idea of extension" did not fully make sense due to the fact that the (linear) causal discovery strategies, which has paid off in MLC-LiNGAM (e.g. up-down and bottom-up search), cannot just directly fit for *NonlinearMLC*.

  Technically speaking, in presence of multiple latent confounders, a LiNGAM (Linear Non-Gaussian Additive Model) would tend to hold after linear regression, whereas an ANM (non-linear Additive Noise Model) is distortion-prone via "inadequate non-linear regression"—non-linear regression is susceptible to multiple latent confounders. In essence, this therein results in the marjor difference between the MLC-LiNGAM and the *Nonlinear-MLC* algorithm.

  Thus, it is partially the reason why I spent the time than anticipation on this paper. With the help from my advisor Wei Chen, fortunately, we alternatively discovered the *Latent-ANMs* lemma (in Section 3). Despite the lemma primarily functions as a "fine-grained" theory, it might be simper and more distinctive in articulating the non-linear identification, which describes the relations between the cause-variable and the other unobserved patents of the effect-variable.

- **What is the limitation of the *Nonlinear-MLC* algorithm?**

  Though I have featured *Nonlinear-MLC* with emphasis on its theory-guided advantages, such as "maximal clique patterns" and "hybrid methodology", I would like to say the meaning of the algorithm is more about the practicable causal discovery program on its own.

  The strategy of maximal-clique-based causal inference, for instance, do strength the empirical performance of *Nonlinear-MLC*, whereas the algorithm in practice (according to my observation while developing the program) does not necessarily obey this "fine-grained" theoretical strategies all the way (e.g. mostly a maximal clique includes the vertexes that are not more than 3, excluding the necessity for comprehensive analysis given such a simple structure). On top of that, restricted in a established hybrid-based framework, *Nonlinear-MLC* might sometimes become susceptible to the so-called cascading errors—an incorrect estimated causal skeleton (in the first stage of the algorithm) can compromise the subsequent non-linear regression and independence tests.

  Bottomline, I think the ideas of *Nonlinear-MLC*, good or bad, would largely depend on the feedback from users in different fields who want to give the non-linear causal inference a shot. By applying *Nonlinear-MLC*, I wish the users more or less are able to get a rough understanding about the causation with respect to the field-related data they are interested in.

- **Would the algorithm be extended to apply for time series data in the future work?**

  No, and I would not recommend that follow-up work is done to serve as a "time-series version" of *Nonlinear-MLC*, though it might be a good way to quickly grasp the idea and yield another paper for utilitarian purpose. Temporal causal discovery has recently been a popular topic, but I wish we could dig deeper instead of directly launching a parallel extension.

Acknowledgements

Xuanzhi Chen would like to thank Jie Qiao and Zhiyi Huang for the personal discussion about details of the primary identification theory (*L-ANMs*) in this paper. Zhengming Chen helped point out initial mistakes in the paper when Xuanzhi Chen was once making presentation at the DMIR group-meeting.

References

Bühlmann, Peter, Jonas Peters, and Jan Ernest (2014). "CAM: Causal additive models, high-dimensional order search and penalized regression". In: *The Annals of Statistics* 42(6), pp. 2526–2556.

Cai, Ruichu et al. (2019). "Causal discovery with cascade nonlinear additive noise models". In: *arXiv preprint arXiv:1905.09442*.

Chen, Wei et al. (2021). "Causal discovery in linear non-gaussian acyclic model with multiple latent confounders". In: *IEEE Transactions on Neural Networks and Learning Systems*.

Chickering, David Maxwell (2002). "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3(Nov), pp. 507–554.

Erdős, Paul, Alfréd Rényi, et al. (1960). "On the evolution of random graphs". In: *Publ. Math. Inst. Hung. Acad. Sci* 5(1), pp. 17–60.

Friston, Karl J, Lee Harrison, and Will Penny (2003). "Dynamic causal modelling". In: *Neuroimage* 19(4), pp. 1273–1302.

Hoyer, Patrik et al. (2008). "Nonlinear causal discovery with additive noise models". In: *Advances in neural information processing systems* 21.

Janzing, Dominik et al. (2012). "Identifying confounders using additive noise models". In: *arXiv preprint arXiv:1205.2640*.

Lewicki, Michael S and Terrence J Sejnowski (2000). "Learning overcomplete representations". In: *Neural computation* 12(2), pp. 337–365.

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

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

Minati, Ludovico et al. (2015). "Synchronization, non-linear dynamics and low-frequency fluctuations: analogy between spontaneous brain activity and networked single-transistor chaotic oscillators". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25(3).

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Peters, Jonas, Joris M Mooij, et al. (2014). "Causal discovery with continuous additive noise models". In.

Shimizu, Shohei, Patrik O Hoyer, et al. (2006). "A linear non-Gaussian acyclic model for causal discovery." In: *Journal of Machine Learning Research* 7(10).

Shimizu, Shohei, Takanori Inazumi, et al. (2011). "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model". In: *Journal of Machine Learning Research-JMLR* 12(Apr), pp. 1225–1248.

Spirtes, Peter and Clark Glymour (1991). "An algorithm for fast recovery of sparse causal graphs". In: *Social science computer review* 9(1), pp. 62–72.

Spirtes, Peter, Clark N Glymour, et al. (2000). *Causation, prediction, and search*. MIT press.

Tashiro, Tatsuya et al. (2012). "Estimation of causal orders in a linear non-Gaussian acyclic model: a method robust against latent confounders". In: *International conference on artificial neural networks*. Springer, pp. 491–498.

Tashiro, Tatsuya et al. (2014). "ParceLiNGAM: A causal ordering method robust against latent confounders". In: *Neural computation* 26(1), pp. 57–83.