Supplementary Material to: "Non-linear Causal Discovery for Additive Noise Model with Multiple Latent Confounders"

Xuanzhi Chen
School of Computer Science
Guangdong University of Technology
Guangzhou, China
xuanzhichen.42@gmail.com

A Proofs

The appendix provides the proof of Lemma 1 [1] and Corollary 1 (identifiable theory of *Latent-ANMs*).

Data Generation Procedure (*Nonlinear-MLC* Causal Models):

Denote $\mathcal{G}_{\boldsymbol{X}}$ with $\boldsymbol{X} = \{x_1, x_2, \ldots, x_d\}$ as directed acyclic graphs (DAG) and $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d\}$ as i.i.d. latent noises, the generation of pairwise additive-noise-models (ANMs) $x_j \to x_i$ characterized by observed parents (pa) and unobserved parents ($\overline{\boldsymbol{pa}}$) can be formalized as:

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

where $f(\cdot)$ denotes the third order differentiable non-linear functions that are in accord with the ANMs assumption, and an extensively latent noise $\xi_i := \varepsilon_i \cup \boldsymbol{f}(\overline{\boldsymbol{pa}}_i) = \varepsilon_i + \sum_{\ell_k \in \overline{\boldsymbol{pa}}_i} f_{ik}(\ell_k)$ is introduced into Equation(1) for modelling the multiple latent confounding from the multiple unobserved parents $\overline{\boldsymbol{pa}}$.

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*

$$(\xi_E \perp\!\!\!\perp C) \wedge (\xi_E := \varepsilon_E \cup \boldsymbol{f}(\boldsymbol{C}^*)) \tag{2}$$

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

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*

$$(E - \mathcal{R}_E(\mathcal{M}^*) \perp\!\!\!\perp C) \wedge (\mathcal{M}^* := \mathcal{M}_{C,E} \backslash \{E\}) \tag{3}$$

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

A scratch of proofs is shown as the following. Lemma 1 is proved by incorporating multiple latent confounding into structure causal models (SCMs), consisting with the ANMs proof framework [Hoyer et al. (2008)] — restricting non-linear function class over differential equations to exhibit asymmetry.

On the basic of Lemma 1, proofs of Corollary 1 sheds light on their mutual equivalences, given the premises in which leveraging maximal-clique-patterns has properly expunged the confounding effect.

[1] Identifiability guaranteed by Lemma 2 (Section 4.2) has been proven in the literature [Spirtes et al. (2000)].

A.1 Proof of Lemma 1

We proved Lemma 1 by transferring variable descriptions beforehand — from the intuitive pairwise cause-and-effect $C \rightarrow E$ into the standard structure causal models (SCMs) with the variables $\boldsymbol{X} = \{x_1, x_2, \ldots, x_d\}$. The corresponding Lemma relative to SCMs is stated as the following.

Lemma 1. ***(SCMs):*** *Assuming data generation procedures are consistent with Equation (1), the pairwise causal dependence between the effect-variable x_i and one of the associating cause-variables $x_j \in \boldsymbol{pa_i}$ is identifiable if and only if*

$$\{\xi_i \perp\!\!\!\perp x_j\} \wedge \{\xi_i := \varepsilon_i + \sum_{\ell_k \in \overline{\boldsymbol{pa}}_i} f_{ik}(\ell_k)\} \tag{4}$$

is satisfied, where ξ_i is denoted as an extensive noise (e.g. compared to the original noise ε_i that has satisfied $\varepsilon_i \perp\!\!\!\perp x_j$). The extensive noise ξ_i further models the multiple latent confounding from the multiple unobserved parents $\overline{\boldsymbol{pa}}_i$.

Taking the potential latent confounders $\ell_k \in \overline{\boldsymbol{pa}}_i$ into consideration, Lemma 1 (SCMs) provides an independent condition to identify the unambiguous causal directions $\{x_j \rightarrow x_i \mid x_j \in \boldsymbol{pa}_i\}$. Next, suppose we use $x_j \rightarrow x_i$ $(x_j = \boldsymbol{pa}_i)$ to represent any of the identifiable pairs satisfying Lemma 1 (SCMs).

Notice that the proof of Lemma 1 (SCMs) is equal to prove that the *Nonlinear-MLC* causal model only holds in the causal direction $x_j \rightarrow x_i$. According to Equation (1), we further formalize the generation procedure as to a correct causal model $\mathcal{M}_1$ in the following

$$\mathcal{M}_1 : x_i := f_{ij}(x_j) + \mathcal{F}_i(\boldsymbol{\ell}_i) + \varepsilon_i. \tag{5}$$

Where $\mathcal{F}_i(\boldsymbol{\ell}_i) = \sum_{\ell_k \in \overline{\boldsymbol{pa}}_i} f_{ik}(\ell_k)$. Without loss of generality, we slightly distinguish the reversed non-linear function and the latent noise, in the sense that an inversed (incorrect) causal model $\mathcal{M}_2$ satisfies

$$\mathcal{M}_2 : x_j := \tilde{f}_{ji}(x_i) + \tilde{\mathcal{F}}_j(\boldsymbol{\ell}_j) + \tilde{\varepsilon_j}. \tag{6}$$

We factorize the marginal distribution (with multiple unobserved parents) entailed by both models:

$$p(x_i,\ x_j) = \sum_{\boldsymbol{\ell}} p(x_i,\ x_j \mid \ell)\ p(\ell) = \begin{cases} \sum_{\boldsymbol{\ell}_i} \ p(x_i \mid x_j, \ell, \mathcal{M}_1)\ p(x_j \mid \ell, \mathcal{M}_1)\ p(\ell \mid \mathcal{M}_1), \\ \sum_{\boldsymbol{\ell}_j} \ p(x_j \mid x_i, \ell, \mathcal{M}_2)\ p(x_i \mid \ell, \mathcal{M}_2)\ p(\ell \mid \mathcal{M}_2). \end{cases} \tag{7}$$

Notice that the independent noise ε is generalized into (the possibly dependence) ξ, along with the independence $\xi_i \perp\!\!\!\perp x_j$ entailed by the identifiable causal model $\mathcal{M}_1$:

$$\xi = \mathcal{F}(\boldsymbol{\ell}) + \varepsilon = \sum_{\ell_k \in \overline{\boldsymbol{pa}}} f_k(\ell_k) + \varepsilon, \quad \xi_i \perp\!\!\!\perp x_j. \tag{8}$$

Given likelihood functions $\mathcal{L} = \log p(\cdot)$ and injective relations between ξ_i and x_j ($\tilde{\varepsilon_j}$ and x_i), combining Equations (7) and (8) yields

$$\mathcal{L}(\mathcal{M}) = \begin{cases} \mathcal{L}_{\xi_i}(x_i - f_{ij}(x_j)) + \ \mathcal{L}_{x_j}(x_j), & \mathcal{M} = \mathcal{M}_1, \\ \mathcal{L}_{\tilde{\varepsilon_j}}\left(x_j - \tilde{f}_{ji}(x_i) - \tilde{\mathcal{F}}_j(\boldsymbol{\ell})\right) + \ \mathcal{L}_{x_i}(x_i), & \mathcal{M} = \mathcal{M}_2. \end{cases} \tag{9}$$

Additionally, we herein **emphasize** that the strict independence $\xi_i \perp\!\!\!\perp x_j$ ensures the expression of $\mathcal{L}(\mathcal{M} = \mathcal{M}_1)$ in Equation (9). In other words, the conditional independence (between ξ_i and x_j) is insufficient to yield that expression in form of regression-based replacement (e.g. replace $\mathcal{L}_{x_i|x_j,\boldsymbol{\ell_i}}(x_i)$ in eq.(7) by $\mathcal{L}_{\xi_i}(x_i - f_{ij}(x_j))$ in eq.(9)). **The reason** is given by the non-linearity, which implies that the variables' non-linear interaction, compared with linearity, will compromise the effect of regression (**recall** the Introduction and Section 3 in the paper).

Based on the formalism shown in Equation (9), we continue the rest of the proof framework by following the **ANMs identification** [Hoyer et al. (2008)]. Assuming $\tilde{f}$ is third order differentiable we obtain

$$\frac{\partial}{\partial x_j}\left(\frac{\partial^2 \mathcal{L}(\mathcal{M}) / \partial x_j^2}{\partial^2 \mathcal{L}(\mathcal{M}) / \partial x_i \partial x_j}\right) = 0, \quad \mathcal{M} = \mathcal{M}_2. \tag{10}$$

Notice that this is not hold when $\mathcal{M} = \mathcal{M}_1$. To see this, imply

$$\frac{\partial^2 \mathcal{L}(\mathcal{M}_1)}{\partial x_j \, \partial x_i} = -f'_{ij} \mathcal{L}''_{x_i}, \tag{11}$$

and

$$\frac{\partial^2 \mathcal{L}(\mathcal{M}_1)}{\partial x_j^2} = \mathcal{L}''_{\xi_i} (f'_{ij})^2 - \mathcal{L}' f''_{ij} + \mathcal{L}''_{x_j}, \tag{12}$$

then we obtain the analogical differential equation (compared with Equation (10)) constructed by $\mathcal{M}_1$:

$$\frac{\partial}{\partial x_j} \left(\frac{\frac{\partial^2 \mathcal{L}(\mathcal{M}_1)}{\partial x_j^2}}{\frac{\partial^2 \mathcal{L}(\mathcal{M}_1)}{\partial x_i \partial x_j}} \right) = -2 f''_{ij} + \frac{\mathcal{L}'_{\xi_i} f'''_{ij} - \mathcal{L}'''_{x_j}}{\mathcal{L}''_{\xi_i} f'_{ij}} + \frac{\mathcal{L}''_{x_j} f''_{ij} - \mathcal{L}'_{\xi_i} (f''_{ij})^2}{\mathcal{L}''_{\xi_i} (f'_{ij})^2} + \frac{\mathcal{L}'_{\xi_i} \mathcal{L}'''_{\xi_i} f''_{ij} - \mathcal{L}''_{x_j} \mathcal{L}'''_{\xi_i}}{(\mathcal{L}''_{\xi_i})^2}. \tag{13}$$

Notice that here we omit the variable inside the function notation.

In order to vanish Equation (13) (if both of the forward causal model $\mathcal{M}_1$ and backward causal model $\mathcal{M}_2$ hold over the joint probability $p(x_i,\, x_j)$), we are supposed to obtain the following (linear inhomogeneous) differential equation [Hoyer et al. (2008)] for every fix x_i given $\mathcal{L}''_{\xi_i} \cdot f'_{ij} \neq 0$. It is given by

$$\mathcal{L}_{x_j}(x_j)''' = \mathcal{L}_{x_j}(x_j)'' \phi(x_j,\, x_i) \; + \; \eta(x_j,\, x_i) \; , \tag{14}$$

where $\phi(x_j,\, x_i)$ and $\eta(x_j,\, x_i)$ are defined by

$$\phi(x_j,\, x_i) = -\frac{\mathcal{L}'''_{\xi_i} f'_{ij}}{\mathcal{L}''_{\xi_i}} + \frac{f''_{ij}}{f'_{ij}}, \tag{15}$$

and

$$\eta(x_j,\, x_i) = -2\mathcal{L}''_{\xi_i} f''_{ij} f'_{ij} + \mathcal{L}'_{\xi_i} f'''_{ij} + \frac{\mathcal{L}'_{\xi_i} \mathcal{L}'''_{\xi_i} f''_{ij} f'_{ij}}{\mathcal{L}''_{\xi_i}} - \frac{\mathcal{L}'_{\xi_i} (f''_{ij})^2}{f'_{ij}}. \tag{16}$$

Therefore, from Equation (14) - (16) we conclude that the hypothetical $\mathcal{L}_{x_j}$ admitting a backward causal model is limited in a three-dimensional, which contradicts our priority that all possible $\mathcal{L}_{x_j}$ should be infinite-dimensional [Hoyer et al. (2008)]. That is, from the perspective of generic, the *Nonlinear-MLC* causal model only holds in $x_j \rightarrow x_i$ and can not be inverted.

A.2 Proof of Corollary 1

Likewise, Corollary 1 was proven provided the context of standard structure causal models (SCMs). The associating Corollary with respect to SCMs is claimed as the following.

Corollary 1. ***(SCMs):*** *Assuming data generation procedures are consistent with Equation (1), the pairwise causal dependence between the effect-variable x_i and one of the associating cause-variables $x_j \in \boldsymbol{\mathcal{M}}_{ij}$ is identifiable if and only if*

$$\{x_i - \mathcal{R}_i(\boldsymbol{\mathcal{M}}^*_{ij} \; \cup \; \hat{\boldsymbol{pa}}_i)\} \perp\!\!\!\perp \; x_j \tag{17}$$

is satisfied, where $\mathcal{R}(\cdot)$ denotes the non-linear regressor, $\boldsymbol{\mathcal{M}}^*_{ij} := \boldsymbol{\mathcal{M}}_{ij} \setminus \{x_i\}$, and $\hat{\boldsymbol{pa}}_i \subseteq \boldsymbol{pa}_i$. In the view of computing memory in (constraint-based) algorithms, $\hat{\boldsymbol{pa}}$ denotes the determined parent relations, whereas $\boldsymbol{\mathcal{M}}_{ij}$ represent the variables (including x_j and x_i) whose relations remain undetermined within the possible maximal cliques.

Providing identifiable causal directions $\{x_j \to x_i \mid x_j \in \boldsymbol{\mathcal{M}}_{ij}\}$, we assume the causal direction as $x_j \to x_i$ to represent any of the identifiable pairs satisfying Corollary 1 (SCMs). The data generation process of the variable x_i can be formulated as

$$x_i := f_{ij}(x_j) + \sum_{x_t \in \boldsymbol{pa}_i \backslash \{x_j\}} f_{it}(x_t) + \sum_{\ell_k \in \overline{\boldsymbol{pa}}_i} f_{ik}(\ell_k) + \varepsilon_i, \tag{18}$$

According to the causal additive models (CAMs) [Bühlmann, Peters, and Ernest (2014)], the empirical (non-linear) regressor $\mathcal{R}_i$ (for the explaining variable x_i) of general additive models (GAMs) [Maeda and Shimizu (2021)] is defined by

$$\mathcal{R}_i := g_{ij}(x_j) + \sum_{x_t \in \hat{\boldsymbol{pa}}_i} g_{it}(x_t) + \sum_{x_r \in \boldsymbol{\mathcal{M}}^*_{ij}} g_{ir}(x_r), \tag{19}$$

where $g(\cdot)$ denotes the empirical regression function selected from GAMs.

Since $\mathcal{R}_i$ is decomposed into several specific parts to cancel the effect of hypothetical cause-variables, we substitute the regressor $\mathcal{R}(\cdot)$ in Corollary 1 (SCMs) with Equation (18) and (19). We conclude

$$H_i(x) \;\perp\!\!\!\perp\; x_j, \tag{20}$$

where $H_i(x)$ is defined by

$$H_i(x) := \{f_{ij}(x_j) - g_{ij}(x_j)\} + \left\{ \sum_{x_t \in \boldsymbol{pa}_i \backslash \{x_j\}} f_{it}(x_t) - \{ \sum_{x_t \in \hat{\boldsymbol{pa}}_i} g_{it}(x_t) + \sum_{x_r \in \boldsymbol{\mathcal{M}}^*_{ij}} \hat{g_{ir}}(x_r)\} \right\} + \{ \sum_{\ell \in \overline{\boldsymbol{pa}}_i} f_{ik}(\ell_k) + \varepsilon_j\}, \tag{21}$$

We **highlight** that the variable set (including x_i) consisting of a maximal clique $\mathcal{M}_{ij}$ might involve the correct (but undetermined) parent relations in the view of algorithmic memory:

$$\exists\, x_r \in \mathcal{M}^*_{ij},\ x_r \not\perp\!\!\!\perp x_i \;\Rightarrow\; x_r \;\in \boldsymbol{pa}_i. \tag{22}$$

In light of Lemma 1, the anticipant independence (**recall** Section 4.1 in the paper) is defined as

$$x_j \;\perp\!\!\!\perp \boldsymbol{pa}_i \backslash \{\hat{\boldsymbol{pa}}_i \;\cup\; \boldsymbol{\mathcal{M}}^*_{ij}\}. \tag{23}$$

We then consider three of the independence combinations of $H_i(x)$ relative to Equation (20). We have

(1) $f_{ij}(x_j) - g_{ij}(x_j) = 0$, which is ideally required by the GAMs regression.

(2) $Z_i(x) \perp\!\!\!\perp x_j$, where $Z_i(x)$ is defined by

$$Z_i(x) := \sum_{x_t \in \boldsymbol{pa}_i \backslash \{x_j\}} f_{it}(x_t) - \{ \sum_{x_t \in \hat{\boldsymbol{pa}}_i} g_{it}(x_t) + \sum_{x_r \in \boldsymbol{\mathcal{M}}^*_{ij}} \hat{g_{ir}}(x_r)\}. \tag{24}$$

Notice that assuming $\{x_j \;\perp\!\!\!\perp \boldsymbol{pa}_i \backslash \{\hat{\boldsymbol{pa}}_i \cup \boldsymbol{\mathcal{M}}^*_{ij}\}\}$ by Equation (23) enforces Equation (24) to vanish into irrelevant regressing residuals with respect to x_j.

(3) $\xi_i \perp\!\!\!\perp x_j$, where ξ_i is the extensive noise (in the data generation procedure, Equation (1)) defined by

$$\xi_i := \sum_{\ell \in \overline{\boldsymbol{pa}}_i} f_{ik}(\ell_k) + \varepsilon_j. \tag{25}$$

The independence for the identifiable $x_j \to x_i$ has already required by Lemma 1 (SCMs).

Thus, the independence implied by Equation (20) eventually reduces to

$$\{Z_i(x) \;\cup\; \xi_i\} \perp\!\!\!\perp x_j,\ H_i(x) := 0 + Z_i(x) + \xi_i, \tag{26}$$

which represents Corollary 1 (SCMs) and is further satisfied by the sub-conditions (1)-(3).

B Average Performance on Experiments Net-Sim2 and Net-Sim3

Based on the corresponding fMRI dataset (Net-Sim3) and the supplemental dataset (Net-Sim2) with lower variable dimension, the average causal discovery performance of the proposed method (*Nonlinear-MLC* algorithm) and baseline methods is listed as the following:

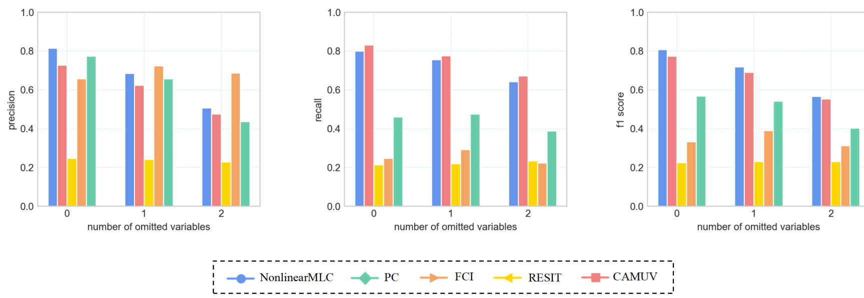

Figure 1: Performance evaluations (precision, recall, f1-score) on fMRI-dataset (sim2).

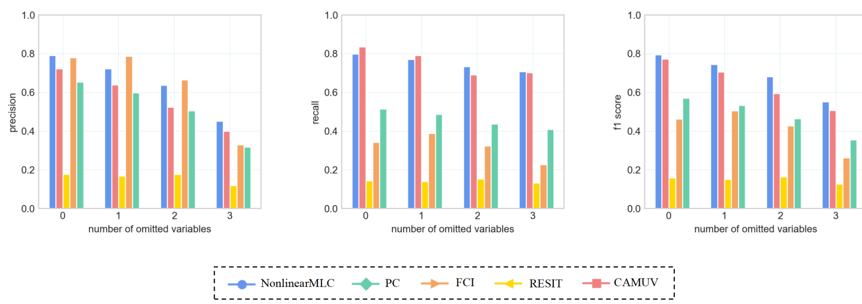

Figure 2: Performance evaluations (precision, recall, f1-score) on fMRI-dataset (sim3).

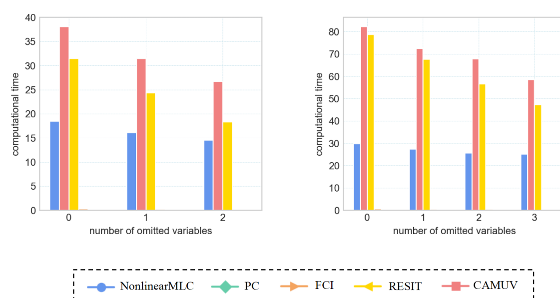

Figure 3: Computational cost of causal discovery on fMRI-dataset (sim2 and sim3).

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.

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

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

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