Identification of Causal Structure in the Presence of Missing Data with Additive Noise Model

Jie Qiao[1], Zhengming Chen[1,3], Jianhua Yu[1], Ruichu Cai[1,2*], Zhifeng Hao[4]

[1]School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China
[2]Peng Cheng Laboratory, Shenzhen 518066, China
[3]Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[4]College of Science, Shantou University, Shantou 515063, China
{qiaojie.chn, chenzhengming1103, philoso521, cairuichu}@gmail.com, haozhifeng@stu.edu.cn

Abstract

Missing data are an unavoidable complication frequently encountered in many causal discovery tasks. When a missing process depends on the missing values themselves (known as *self-masking missingness*), the recovery of the joint distribution becomes unattainable, and detecting the presence of such self-masking missingness remains a perplexing challenge. Consequently, due to the inability to reconstruct the original distribution and to discern the underlying missingness mechanism, simply applying existing causal discovery methods would lead to wrong conclusions. In this work, we found that the recent advances additive noise model has the potential for learning causal structure under the existence of the self-masking missingness. With this observation, we aim to investigate the identification problem of learning causal structure from missing data under an additive noise model with different missingness mechanisms, where the 'no self-masking missingness' assumption can be eliminated appropriately. Specifically, we first elegantly extend the scope of identifiability of causal skeleton to the case with weak self-masking missingness (i.e., no other variable could be the cause of self-masking indicators except itself). We further provide the sufficient and necessary identification conditions of the causal direction under additive noise model and show that the causal structure can be identified up to an IN-equivalent pattern. We finally propose a practical algorithm based on the above theoretical results on learning the causal skeleton and causal direction. Extensive experiments on synthetic and real data demonstrate the efficiency and effectiveness of the proposed algorithms.

Introduction

Missing data are ubiquitous in many fields and causal discovery from missing observational data is challenging due to possible complex *missingness mechanisms*—causal structure among the causal variables and its missingness indicators. 'indicator' is variables that specify whether a value was missing for an observation. With different structures of the missingness mechanisms, following Rubin (1976), the missingness types can be categorized as Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR).

When data are MCAR, the causal variables, and their missingness indicators are independent such that the missingness mechanism is ignorable, and one can perform the listwise deletion that simply drops the samples with missing value (Strobl, Visweswaran, and Spirtes 2018). However, for MNAR, the missingness mechanism is not ignorable, and the simple deletion would introduce bias resulting in incorrect inference (Rubin 2004; Mohan, Pearl, and Tian 2013; Tu et al. 2019). Much effort has been made in studying the recoverability of the MNAR data to correct the spurious correlations from deleted-wised distribution using reweighting by incorporating prior information of the causal graph and missingness mechanisms (Bhattacharya et al. 2020; Mohan and Pearl 2021; Nabi, Bhattacharya, and Shpitser 2020). However, when a missing process depends on the missing values themselves (known as *self-masking* belonging to MNAR), it suffers from a serious identifiability issue such that the origin distribution is not recoverable (Nabi, Bhattacharya, and Shpitser 2020).

Existing works, therefore, mainly focus on causal discovery with no self-masking missingness mechanisms. On one hand, (Gao et al. 2022) addresses the M(C)AR cases using the identifiable additive noise models (ANMs) (Hoyer et al. 2008; Shimizu et al. 2006; Peters and Bühlmann 2014) with a general EM-based framework to perform causal discovery in the presence of missing data. On the other hand, to learn causal structure in MNAR, some constraints of the missingness mechanisms must be made. With the no self-masking assumption, Gain and Shpitser (2018) proposes to extend the constraint-based PC algorithm (Spirtes et al. 2000) by correcting the bias of each conditional independence (CI) test bought by the missing value. While, instead of correcting CI test, Tu et al. (2019) proposes a more efficient post-correction strategy to correct the spurious edges produced by the original PC algorithm.

However, due to the limited recoverability, it is still unclear how to identify causal structure in the presence of the self-masking missingness mechanism, which is believed to be the most commonly encountered in practice (Osborne 2013), e.g., smokers do not answer any questions about their smoking behavior in insurance applications, and people with very high or low income do not disclose their income (Mohan 2018).

In this work, we found that the recent advances additive noise model has the potential for learning causal structure

*Corresponding author.

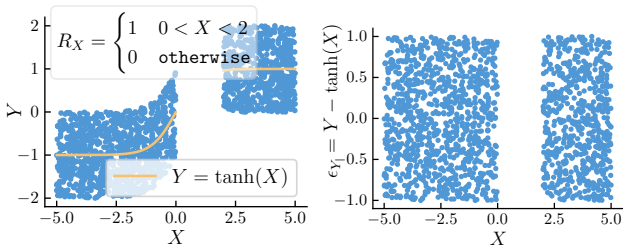


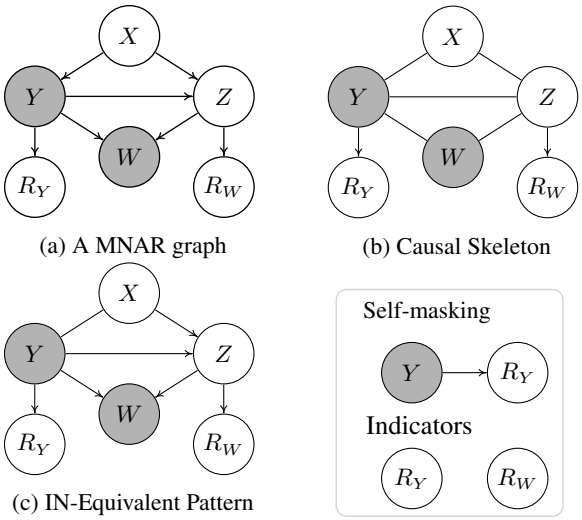
Figure 1: Illustration of self-masking missingness ANM in which X is the self masking missing variable satisfying $R_x \leftarrow X \rightarrow Y$. Panels left show the scatter plot of X and $Y = \tanh(X) + N$ and Panels right show the corresponding noise of ANM.

under the existence of the self-masking missingness mechanism. Take Fig. 1 as an example, suppose a causal pair $R_X \leftarrow X \rightarrow Y$ where X is a self-masking missing variable and R_X is the missingness indicator of X. Although we can only access the distribution of $P(X|R_X = 0)$, the causal direction is still identifiable since the independence noise $X \perp\!\!\!\perp \epsilon_Y$ still holds—considering each separate part follows an identifiable ANM. In contrast, suppose $X \rightarrow Y \rightarrow R_Y$ as shown in Fig. 2(a), unfortunately, ANM is not identifiable and the independence noise will not hold in this case since we can only access the distribution in $R_Y = 0$ such that $X \not\perp\!\!\!\perp \epsilon_Y|R_Y$ (R_Y is the descendent of the collider Y). Nevertheless, if the self-masking of Y is known, we can still identify such causal direction since we know that such dependence noise is brought by R_Y, otherwise, the causal direction can be identified in the first place. With this observation, in this paper, we aim to address the following two questions: 1) How to identify the self-masking missingness mechanism in learning causal skeleton? 2) How to identify the causal direction when ANM is not identifiable due to the self-masking missingness?

In answering the first question, we relax the assumptions in MVPC (Tu et al. 2019) to allow the existence of the weak self-masking missingness and propose a post-correction strategy to correct the spurious edges produced by the self-masking missingness. For the second question, by giving the causal skeleton, we propose an equivalent class for ANM and develop a searching method to identify the causal direction from the equivalent class. Our main contributions are summarized as follows: 1) we propose a practical causal skeleton learning algorithm for learning the skeleton among causal variables and the missingness indicators in the presence of weak self-masking missingness. 2) We provide a theoretical analysis of the identification of the additive noise model under missing data. 3) Combined with the causal skeleton and additive noise model, we show that the causal structure can be identified up to an IN-equivalent pattern. 4) We provide some insightful orientation rules to further orient the IN-equivalent pattern.

Problem Definition

In this paper, we focus on the problem of learning causal structure from missing data based on the additive noise model



(a) A MNAR graph (b) Causal Skeleton

(c) IN-Equivalent Pattern

Figure 2: An example of missingness graph with self-masking missingness. Here, gray nodes are partially observed variables, and white nodes are fully observed variables, R_Y and R_W are the missingness indicators of Y and W, respectively.

(ANM). We begin with the description of the missingness graphs (or *m-graph*) (Mohan and Pearl 2021) and the ANM, used as a representation of complex structures in missing data. Both of them are building blocks for our results.

Missingness Graph. A missingness graph $G(\mathbb{V}, \mathbf{E})$ is a Directed Acyclic Graph (DAG) where $\mathbb{V} = \mathbf{V} \cup \mathbf{V}^* \cup \mathbf{R}$ is the set of nodes and $\mathbf{E}$ is the set of edges. $\mathbf{V}$ is a set of observable variables which can be decomposed into $\mathbf{V} = \mathbf{V}_o \cup \mathbf{V}_m$, where $\mathbf{V}_o$ is the set of fully observed variables, depicted as white nodes in our figure, and $\mathbf{V}_m$ is the set of partially observed variables that contain missing values, depicted as gray nodes in our figure. Each $\mathbf{V}_i^* \in \mathbf{V}^*$ is a proxy variable that is actually observed and $R_{V_i} \in \mathbf{R}$ is the missingness indicator of the partially observed variable such that

$$V_i^* = \begin{cases} V_i & R_{V_i} = 0 \\ ? & R_{V_i} = 1 \end{cases} \tag{1}$$

where $R_{V_i} = 1$ denotes the corresponding entry is missing while $R_{V_i} = 0$ means the entry is observable and V_i^* takes the value of V_i.

Our work is in the framework of causal graphical models. Some concepts used here without explicit definition, such as d-separation, and Markov equivalent class, can be found in standard sources (Spirtes et al. 2000). Besides, we use "$\perp\!\!\!\perp$" to denote the independent relation in dataset and use "$\perp\!\!\!\perp_d$" to denote the d-separation in m-graph. With the m-graph, we say data are MCAR if $\mathbf{V}_m, \mathbf{V}_o \perp\!\!\!\perp_d \mathbf{R}$ holds and MAR if $\mathbf{V}_m \perp\!\!\!\perp_d \mathbf{R}|\mathbf{V}_o$, and MNAR otherwise. We say a path $P = (V_{i_0}, V_{i_1}, \ldots, V_{i_k})$ in G is a *directed path* if it is a sequence of nodes of G where there is a directed edge from V_{i_j} to $V_{i_{(j+1)}}$ for any $0 \leq j \leq k-1$. For simplifying graphical concepts, we use the following symbols $Pa_{V_i} = \{V_j|V_j \rightarrow V_i\}$, $Ch_{V_i} = \{V_j|V_i \rightarrow V_j\}$, $Anc_{V_i} = \{V_j|V_j \rightsquigarrow V_i\}$, $Des_{V_i} = \{V_j|V_i \rightsquigarrow V_j\}$, $Adj_{V_i} = \{V_j|V_j \leftarrow V_i \vee$

$V_j \to V_i\}$ to denote the set of parents, children, ancestors, descendants, and adjacent nodes of V_i, respectively.

Example 1 (Illustrative example for the m-graph)**.** *Consider the m-graph in Fig. 2(a). The full observed variable set is* $\mathbf{V}_o = \{X, Z\}$ *and the partially observed variable set is* $\mathbf{V}_m = \{Y^*, W^*\}$*. Moreover,* R_Y *and* R_W *is the missing indicator of* Y *and* W *respectively, where* R_Y *is a 'self-masking missingness' indicator.*

Additive Noise Model. Additive noise model assumes the observed data has been generated the following way: for each variable $V_i \in \mathbf{V}$ and its causal parents $Pa_{V_i} \subseteq \mathbf{V}$, there is

$$V_i = f_i(Pa_{V_i}) + \varepsilon_i, \quad Pa_{V_i} \perp\!\!\!\perp \varepsilon_i, \tag{2}$$

where ε_i denotes noise variable which is independence of Pa_{X_i} and other noise variables, i.e., $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$, for all $j = 1, \ldots, |\mathbf{V}|$, where $|\mathbf{V}|$ denotes the number of observed variables in graph, and f_i denotes the functional relationship between cause and effect.

ANM is widely used in statistics, multivariate analysis, and causal discovery (Hoyer et al. 2008; Cai et al. 2018; Qiao et al. 2021). In causal discovery, ANM is capable to alleviate the problem of Markov equivalence class and display an identification ability on causal direction, which may be helpful in the missing data setting. In the paper, we only focus on the class of ANMs that are identifiable in complete data, e.g., the linear non-Gaussian model (Shimizu et al. 2006), or the nonlinear additive noise model (Hoyer et al. 2008).

However, without further constraints, it is hard to answer the identification problem in the presence of missing data even in the ANM model. Our theoretical results build upon the basic framework of Tu et al. (2019), following Assumption 1 ~ 3 in Tu et al. (2019), which restrict (i) the missingness indicators are not causes; (ii) conditional independence relations in the observed data also hold in the complete data; (iii) no causal interactions between missingness indicators. Besides, we allow the existence of self-masking missingness (as shown in the following Assumption).

Assumption 1 (Weak self-masking missingness)**.** *Weak self-masking missingness refers to missingness in a variable that is caused only by itself, i.e., for any self-masking missing variable* $V_i \to R_i$*,* $V_i \in \mathbf{V}_m$*, the parent set of* R_i *only contains itself* $Pa_{R_i} = \{V_i\}$*.*

The key difference to existing methods, such as Tu et al. (2019); Gao et al. (2022), is that we relax the 'no self-masking missingness' assumption to a sensible situation (as shown in Assumption 1). Moreover, to ensure the asymptotic correctness of the identification algorithm in the presence of weak self-masking missingness, apart from the assumptions of causal Markov, faithfulness, and causal sufficiency, we further require a mild structural assumption.

Assumption 2 (Structural condition)**.** *For each indicator of a weak self-masking variable* $Z_i \in \mathbf{Z}$*, there exists* $X, Y \in \mathbf{V} \setminus \mathbf{Z}$ *such that* $X \perp\!\!\!\perp_d Y | \mathbf{Z}$ *in the ground truth.*

Note that this assumption generally holds as the real-world structure is generally sparse and such d-separation should often occur. Overall, our **goal** is to identify the causal structure from missing data under more general assumptions that the self-masking missingness is allowed.

Identifiability Results

In this section, we address the problem of structure learning using ANM under the proper assumptions discussed above. To leverage the identification results of ANM (refer to Theorem 3), it is necessary to answer the identification of missing indicators in advance. We will show that the missing indicator can be found in a constraint-based framework that learns the causal skeleton and missing indicators correctly (Theorem 2). Based on this result, we complete the identification of causal direction and show that the causal structure is identified up to an IN-equivalent pattern for ANM (Definition 1) and further provide an insightful orientation rule for further searching the equivalent class. Proofs are given in Qiao et al. (2023).

Identification of Missing Indicators

We start with the identification of missing indicators, used as a building block for the ANM identification. Our results are based on a constraint-based framework in which we represent the learned structure as a causal skeleton that entails a set of d-separation relations. In other words, the causal skeleton characters the Markov equivalence class and ignores the causal direction information (e.g., V-structure). In this section, we mainly focus on the identification of missing indicators, together with the causal skeleton, which are mutually complementary.

To complete the identification of missing indicators, we are required to deal with two challenges: (i) CI test in missing data where weak self-missingness is allowed; (ii) identify the causal relations between variable $\mathbf{V}$ and missing indicator $\mathbf{R}$. The result of the first challenge ensures the correctness of the recovered causal skeleton while the latter finds the structural position of missing indicators in the causal skeleton (e.g., the parent set for each missing indicator).

CI test in missing data in the presence of weak self-masking missingness. Consider the first challenge that the goal is to perform the correct CI tests in missing data and then use them to recover the causal skeleton. Under the 'no self-masking missingness' assumption, a recent approach by Tu et al. (2019) has shown that although there are some special cases, in most cases, the CI relation can be directly tested from data by the Test-wise Deletion based method (TD-based method). Our first finding is extending the CI relations in test-wise deletion data (Tu et al. 2019) to the case of weak self-masking missingness.

Theorem 1. *With Assumption 1 ~ 3 in (Tu et al. 2019) and the assumption of weak self-masking missingness, for any* $X, Y \in \mathbf{V}$*,* $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$*, and their corresponding missingness indicators* $\mathbf{R}_{X,Y,\mathbf{Z}}$*, the CI test between* X, Y *given* $\mathbf{Z}$ *is always consistent with that without the self-masking missingness, i.e.,* $X \perp\!\!\!\perp Y | \{\mathbf{Z}, \mathbf{R}_{X,Y,\mathbf{Z}}\}) \Leftrightarrow X \perp\!\!\!\perp Y | \{\mathbf{Z}, \mathbf{R}_{X,Y,\mathbf{Z}} \setminus \mathbf{R}_S\})$ *and* $X \not\perp\!\!\!\perp Y | \{\mathbf{Z}, \mathbf{R}_{X,Y,\mathbf{Z}}\}) \Leftrightarrow X \not\perp\!\!\!\perp Y | \{\mathbf{Z}, \mathbf{R}_{X,Y,\mathbf{Z}} \setminus \mathbf{R}_S\})$ *where* $\mathbf{R}_S = \{R_i | V_i \to R_i\}$ *is the set of weak self-masking indicators.*

Here, the 'consistent' means that the CI relations that contain the weak self-masking missingness are equal to the CI relations after the elimination of such self-masking missingness. Based on Theorem 1, the existing results for CI relations

in missing data can be naturally extended to the case with the existence of weak self-masking. Thus, we can further extend the result from (Tu et al. (2019), Prop. 2) that the CI relations in the observed MNAR data may be different from those in the complete data. Such phenomenon, as shown in Corollary 1, is due to conditioning on a missingness indicator which is a collider, and thus introduces a spurious dependence.

Corollary 1. *Suppose that X and Y are not adjacent in the true causal graph and that for any variable set $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ such that $X \perp\!\!\!\perp Y | \mathbf{Z}$. Then under Assumption 1 ~ 3 in (Tu et al. 2019) and the assumption of weak self-masking missingness, $X \not\perp\!\!\!\perp Y | \{\mathbf{Z}, R_X, R_Y, \mathbf{R_Z}\}$ if and only if for at least one variable $V \in \{X\} \cup \{Y\} \cup \{\mathbf{Z}\}$, such that X, Y are the direct parents or direct ancestor of R_V.*

Such spurious dependence is supposed to be addressed using methods like Density Ratio Weighted correction (DRW) (Mohan, Pearl, and Tian 2013) to recover the original distribution. However, these methods only hold in the assumption of no self-masking missingness and are not feasible when self-masking missingness exists. Interestingly, we find that the distribution of missing data can still be recovered up to a reasonable conditional distribution that only conditions on the self-masking variables, as given in Prop. 1.

Proposition 1. *With assumptions 1 ~ 3 in (Tu et al. 2019) and assumption of weak self-masking missingness, given a m-graph G, the joint distribution $P(V)$ is recoverable up to $P(V|\mathbf{R}_S)$, where $\mathbf{R}_S = \{R_i | V_i \to R_i\}$ is the collection of the self-masking missingness indicators. Then, we have*

$$P(V|\mathbf{R_S}) = \frac{P(\mathbf{R}_{V\setminus S} = 0, V|\mathbf{R}_S)}{\prod\limits_{i\in\{i|R_i\in\mathbf{R}_{V\setminus S}\}} P\left(R_i = 0|Pa^o_{R_i}, Pa^m_{R_i}, \mathbf{R}_{Pa^m_{R_i}}\right)} \tag{3}$$

where $Pa^o_{R_i} \subseteq V_0$ and $Pa^m_{R_i} \subseteq V_m$ denote the parents of R_i, $\mathbf{R}_{V\setminus S}$ is the non self-masking missingness indicators.

Based on Prop. 1 and Theorem 1, one may correct the CI relations in the recovered distribution and hence ensure the causal skeleton is reconstructed correctly by CI test and correcting method. The implementation will be provided in the algorithm section. By this, the first challenge is solved.

Identifying missing indicators Now, we deal with the second challenge: how to identify the missing indicator variables in the skeleton. Our solution is built upon the constraint-based method that the parent set of missing indicators is identified through CI testing.

One straightforward way is to perform test-wise deletion CI test between each missingness indicator and the variables. As discussed by Tu et al. (2019), if there does not exist self-masking missingness, and the assumption 1 ~ 3 in (Tu et al. 2019) hold, then all missingness indicators can be identified correctly since CI relations in observed data will be consistent with that in complete data, i.e., $R_{V_i} \perp\!\!\!\perp V_j | \mathbf{Z} \Leftrightarrow R_{V_i} \perp\!\!\!\perp V_j^* | \mathbf{Z}$ for any $\mathbf{Z} \subseteq \mathbf{V} \setminus V_j$ where V_j^* is the observed variable in missing data.

However, such a result only holds for the case without self-masking missingness. When a self-masking missingness $V_i \to R_{V_i}$ exist, we have $R_{V_i} \not\perp\!\!\!\perp V_i$ but $R_{V_i} \perp\!\!\!\perp V_i^*$. The reason is that we can only observe the value of V_i when $R_{V_i} = 1$ and missing when $R_{V_i} = 0$ such that R_{V_i} is a constant in testing leading to independence. Such an untestable problem will introduce spurious dependence on weak self-masking missingness indicators. An example is given in the following.

Example 2 (Untestable CI relations in self-masking indicator)**.** *As shown in Fig. 2, since Y being missing, the conditional independence $\{X, W\} \perp\!\!\!\perp R_Y | Y$ is untestable. It will result in the relations between X, W, and R_Y being dependent, and thus X, W is incorrectly identified as the parent of R_Y.*

To tackle this problem, fortunately, we find that such incorrect results could be corrected and the indicators of weak self-masking variables are identifiable by detecting the conflict of CI relations under a mild structure condition (Assumption 2)—there exist two variables that are d-separated by the weak self-masking missingness variable.

Lemma 1 (Identification of self-masking indicator)**.** *Suppose Assumption 1 ~ 3 in (Tu et al. 2019) holds, and further assume weak self-masking missingness and structure condition hold. A variable $Z_i \in \mathbf{Z}$ is a self-masking missingness variable if there exits $X, Y \in \mathbf{V}$ such that $X, Y \not\perp\!\!\!\perp R_{Z_i} | \{R_X, R_Y\}$, and (i) a simply test-wise deletion CI test yields $X \perp\!\!\!\perp Y | \{\mathbf{Z}, R_X, R_Y, \mathbf{R_Z}\}$; or (ii) after the correction by Prop. 1, the CI test yields $X \perp\!\!\!\perp Y | \{\mathbf{Z}, \mathbf{R}_S\}$, where $\mathbf{R}_S$ is test-wise self-masking missingness indicators.*

Roughly speaking, Lemma 1 provides a correcting method for identifying an indicator of self-masking missingness. An illustrative example is provided in the following.

Example 3. *According to Example 2, in Fig. 2(a), X, W will incorrectly be identified as a parent of R_Y. That is, we have $X, W \not\perp\!\!\!\perp R_Y | \{R_X, R_W\}$. Moreover, we can further find that $X \perp\!\!\!\perp W | \{Y, Z, R_Y, R_W\}$ holds which is unreasonable because $X \perp\!\!\!\perp W | Y, R_Y$ is not possible to hold if R_Y is the collider for X and W under faithfulness assumption. Therefore, R_Y must be the weak self-masking missingness indicator and we obtain Lemma 1.*

Based on the result in Lemma 1, the identification of missing indicators is well addressed.

Theorem 2 (Identification of missing indicators)**.** *Suppose Assumption 1 ~ 3 in (Tu et al. 2019) holds, and further assume weak self-masking missingness and structure condition hold. The causal relations of missing indicators are identifiable.*

By this, one may learn the causal skeleton and the causal relations of missing indicators correctly based on the above theoretical results. A practical identification algorithm (*SM-MVPC*) is presented in the 'algorithm' section. Now, we are capable to cope with the next issue—identification of causal direction.

Identification of Causal Direction

In this section, we aim to address the identification of causal directions in the learned causal skeleton by ANM. One of the straightforward ways to identify the causal direction is to recover all relevant distributions that are required for ANM

identifiability. However, the recovery procedure is generally harder and could be inaccurate. Thus, in this work, instead of correcting all distributions, we aim to investigate the identification of causal direction using only the independence noise property in ANM based on the learned causal skeleton.

To do so, we first discuss the identifiable ANM under the complete data and then investigate the identifiability of ANM under the missing data, resulting in a sufficient and necessary condition for that identifiability (Theorem 3). To further leverage the identifiability results from ANM, we formulate the IN-equivalent pattern to characterize the independent noise property (Definition 1), which allows us to orient the causal direction whenever possible by an insightful orientation rule (Theorem 4).

The key to solving the identification of causal direction is the independent noise property of ANM. Below, we will show how the causal direction is identified by ANM in the complete data, which defines the concept of '*identifiable*'.

Remark 1. *Given an ANM $V_i = f_i(Pa_{V_i}) + \varepsilon_i$ in complete data, one can identify the causal direction by testing the independence between the residuals of regression and the hypothesis cause, such that the independence holds only in the correct causal direction, e.g., $Pa_{V_i} \perp\!\!\!\perp V_i - f_i(Pa_{V_i})$.*

However, when data contain missing values, the independence between residual and cause variables may no longer holds. For example, as shown in Fig. 2, the ANM of $X \to Y$ is not identifiable because $Y^* - f_i(X) \not\perp\!\!\!\perp X$ in missing data. Thus, it is necessary to understand in what cases the ANM is identifiable or not in the presence of missing data. In the following theorem, we provide a sufficient and necessary condition for the identifiability of an ANM causal pair under the missing data.

Theorem 3 (Identifiability of ANM in missing data)**.** *Given an m-graph G, an additive noise model*

$$V_i = f_i(Pa_{V_i}) + \varepsilon_i, \quad Pa_{X_i} \perp\!\!\!\perp \varepsilon_i$$

is identifiable in complete data but not identifiable in missing data if and only if there exists a directed path in G that starts from one of the parent $V_j \in Pa_{V_i}$ and ends at missing indicator R_{V_j} or R_{V_i}.

Example 4 (An example for illustrating Theorem 3)**.** *As shown in Fig. 2, one can see that the ANM in the causal pair between Y, Z and W is identifiable even though there exist self-masking variables. However, the causal pair in between X, Y is non-identifiable due to there exists a directed path from X to R_Y.*

Theorem 3 has a graphical implication: in the missing data, the identification of ANM is determined by the structural position of the missing indicator in the causal skeleton. To further characterize the identification of ANM in a causal skeleton, we devise the **I**ndependent **N**oise-equivalent (IN-equivalent) pattern using a partially directed graph:

Definition 1 (IN-equivalent pattern)**.** *An IN-equivalent pattern is a partially directed m-graph, which has the identical adjacency as the original m-graph and which has oriented edge $Pa_{V_i} \to V_i$ if and only if $Pa_{V_i} \perp\!\!\!\perp V_i - f_i(Pa_{V_i}) | \{R_{V_i}, \mathbf{R}_{Pa_{V_i}}\}$ where f_i is a regression of V_i on Pa_{V_i}.*

In other words, the IN-equivalent pattern encodes the set of independent noise of ANM, where the direct edge represents the identifiable ANM in this edge while the bi-directed edge represents that there is no distinguishable independent noise of ANM for identifying such causal direction.

Similar to Meek's rule (Meek 1995) in the Markov equivalence pattern, given the IN-equivalent pattern, we are able to further orient the undetermined edges by making no identifiable ANM structure and no cycle in the pattern. Our idea is derived from the fact that if (i) an undirected edge $X - Y$ in the IN-equivalent pattern is not identified as $X \to Y$ (causal direction) due to Theorem 3, and (ii) ANM reject the reverse direction $X \leftarrow Y$ by the violation of independent noise, then we can infer $X \to Y$ uniquely because unidentifiable ANM for $X \to Y$ is only caused by (i). To characterize such fact and further consider the no cycle constraint, we are able to graphically characterize the identifiability of ANM in missing data using the following potential non-identifiable paths:

Definition 2 (Potential Non-identifiable Paths)**.** *For a node $V_i \in \mathbf{V}$, the potential non-identifiable paths $\mathbf{P}_i$ is a set of paths w.r.t. V_i on the missingness graph $G(\mathbb{V}, \mathbf{E})$ such that for each path in $\mathbf{P}_i$ satisfies: 1) each edge in the path is either $\to$ or $-$; 2) the path starts from $V_j \in \{V_j | V_j - V_i \in \mathbf{E}\}$ which is one of the undirected neighbors of V_i, and the corresponding end point is R_{V_i} or R_{V_j}; 3) for every undirected edge $V_k - V_l$ in the path it can not create cycle in the graph if we orient $V_k \to V_l$, i.e., there does not exist a directed path from V_l to V_k.*

Based on the potential non-identifiable paths, we conclude the following orientation rule:

Theorem 4 (Orientation Rule)**.** *Given an IN-equivalent pattern represented by a partially directed m-graph $G(\mathbb{V}, \mathbf{E})$, where edges $\mathbf{E}$ may contain directed edges and undirected edges. If the potential non-identifiable paths of V_i is empty, then we orient every undirected neighbor $V_j \in \{V_j | V_j - V_i\}$ as $V_i \to V_j$.*

Example 5 (Illustrating for orientation rule)**.** *Take Fig. 1(c) as an example, for node Y, there exists one potential non-identifiable path $X - Y \to R_X$. Meanwhile, for X there does not exist such type of path. We infer $X \to Y$ by Theorem 4.*

Intuitively, since all IN-equivalent graphs share the same independence noise, if there exists an orientation that introduces a new independence noise, then such an orientation is invalid as it will break the equivalent pattern. Thus, if we can find a unique structure that makes no identifiable ANM structure in missing data, then we can orient the edge whenever possible.

The Algorithms for Learning Causal Structure from Missing Data

In this section, we provide the implementation of identifiability results, consisting of two practical algorithms for learning the causal skeleton and causal direction, respectively.

As shown in Algorithm 1 (named as Self-Masking MVPC, abbreviated as SM-MVPC), it is an extension of MVPC (Tu et al. 2019) for learning causal skeleton allowing the presence of self-masking missingness. Here, due to space limitation,

Algorithm 1: Self-Masking MVPC (Simplified)

// Learning skeleton by PC algorithm in deleted data
$G \leftarrow$ Test-wise deletion PC (Spirtes et al. 2000);
// Learning missing indicators in skeleton
Detect direct causes of missingness indicators by testing CI relations between each missingness indicator and the variables in $\mathbf{V}$;
Find the weak self-masking indicators by Lemma 1;
// Reconstructing the causal skeleton by a correcting strategy
Detecting potential extraneous edges and recovering the true causal skeleton in an iteration way;

Algorithm 2: Learning Causal Structure from Missing Data (LCS-MD)

Input: The causal skeleton G (including all missing indicators) learned by *SM-MVPC* and the dataset $\mathcal{D}$ with the observed variable set $\mathbf{V}$
Output: A partially directed acyclic graph among the observed variables

// Estimating IN-Equivalent pattern
for $V_i \in \mathbf{V}$ **do**
  Find the maximum parents set from adjacent with $\mathrm{Pa}_{V_i} \subseteq \mathrm{Adj}_{V_i}$;
  if $V_i - f_i(\mathrm{Pa}_{V_i}) \perp\!\!\!\perp \mathrm{Pa}_{V_i}$ **then**
    Orient $\mathrm{Pa}_{V_i} \to V_i$ in G;
// Perform orientation rule on IN-equivalent pattern
while *no undirected edges in G can be oriented;* **do**
  Orient the undirected edges based on Theorem 4.
Return partially directed acyclic graph G

we only present a simplified version of the full algorithm, and the complete algorithm is given in Appendix. In SM-MVPC, we first apply the classic PC algorithm in deleted data to obtain an initial skeleton among $\mathbf{V}$ and further to find the missingness indicators based on Theorem 2. Finally, we iteratively remove potential redundant edges in the recovered distribution (Prop. 1) to obtain the true causal skeleton.

Next, we provide an algorithm for learning causal direction based on the output of SM-MVPC, as shown in Algorithm 2 (*LCS-MD*). We first learn the causal direction by enumerating all candidate parent sets of each node and testing the independence between residuals of regression and hypothesis causal variables. The above procedure outputs an IN-equivalent pattern. Next, we orient the undirected edges according to the orientation rules (Theorem 4). Our algorithm outputs a partially directed acyclic graph with maximum direction information.

Theorem 5 (Soundness of LCS-MD)**.** *Suppose that the data over variables $\mathbf{V}$ was generated by ANM and assumptions 1 ~ 3 in (Tu et al. 2019), assumptions of weak self-masking missingness and structural condition hold. Let G denote the output of LCS-MD, all directed edge in G is correctly oriented.*

Experiments

In this section, we verify the proposed method with baselines through simulation studies and real-world dataset studies. Our baseline methods include MVPC (Tu et al. 2019) and MissDAG (Gao et al. 2022). We evaluate the methods in estimating the skeleton and inferring causal directions, respectively.

Synthetic Experiments

In synthetic experiments, we conducted two different control experiments: (1) the sensitivity of sample size, and (2) the sensitivity of the number of weak self-masking missingness for testing the performance of causal skeleton learning (including the missingness indicators) and causal structure learning, respectively. Each experiment is conducted on MNAR with the existence of weak-self masking missingness. The sensitivity experiments are controlled by traversing the controlled parameter while keeping other setting fixed as default. The ranging of sample size, and the number of weak self-masking missingness: $\{3000, 5000, \mathbf{7000}, 9000, 15000\}$, $\{1, 2, \mathbf{3}, 4\}$, respectively. The default setting is marked as bold. Each experiment is conducted 100 times with different seeds. The causal graph has 10 variables with at least 3 missing variables.

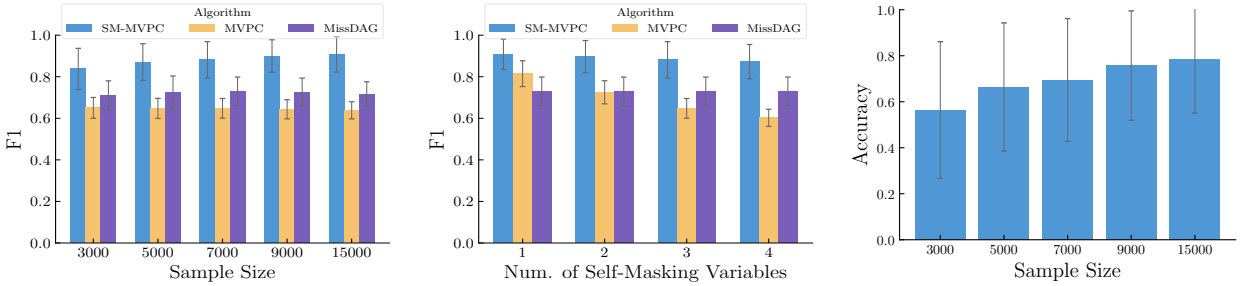
Our data generation process follows the constraint of the assumptions in this work and the missing data are generated according to the missingness indicators following Liu and Constantinou (2021). In detail, for each causal pair, f_i is constructed from a two-layer random MLP with 50 hidden layers $V_i = f(Pa_{V_i}) + \varepsilon_{V_i}$, where the noise $\varepsilon_{V_i} \sim \mathcal{U}(-1, 1)$.

In all experiments, Structural Hamming Distance (SHD), Precision, Recall, and F1 are used as the evaluation metrics. For the independence test, the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005; Zhang et al. 2018) test is used, and we set the significance level as $\alpha = 0.01$. Due to the space limitation, only the F1 metric is provided, other metrics are provided in the supplementary material.

Results in Causal Skeleton Identification. Results are tested with the skeleton of the ground truth. As shown in Fig. 3, one can see that SM-MVPC achieves the best performance in all cases since existing methods are not designed for the MNAR with self-masking cases, which shows the effectiveness of our proposal in learning causal skeletons.

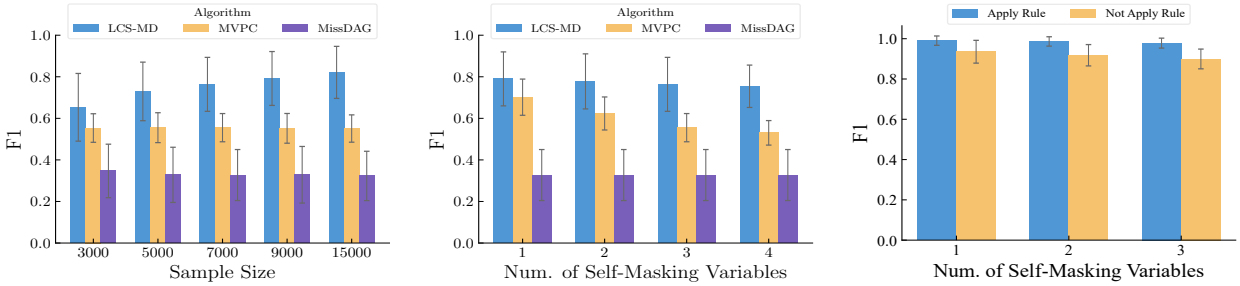
For the sensitivity of the number of weak self-masking variables given in Fig. 3(b), our method is not sensitive while the performance of MVPC decreases as the number of weak self-masking variables increases. It suggests the correctness of our method and the existence of redundant edges in MVPC due to the self-masking variables. For other methods, based on Theorem 1 it is reasonable to be not sensitive. In addition, in Fig. 3, the average accuracy of the identification of self-masking increases as the sample size grows which also verifies the effectiveness in identifying the missing indicator.

Results in Causal Direction Identification. For learning causal structure, as shown in Fig. 4, our method outperforms all the baseline methods. Moreover, compared with the experiments in the skeleton, the performance gap with baseline



(a) Sensitivity of sample size (b) Sensitivity of the weak self-masking (c) Accuracy of self-masking indicator

Figure 3: Experiments for Skeleton Learning



(a) Sensitivity of sample size (b) Sensitivity of the weak self-masking (c) Applying rule with correct skeleton

Figure 4: Experiments for Structure Learning

methods is even larger, which is thanks to the identifiability of ANM. As for the sensitivity, all methods are sensitive to the sample size. While our method is not sensitive to the weak self-masking variables but MVPC is sensitive. The above results are consistent with the experiments on the skeleton. Moreover, to verify the effectiveness of the orientation rule, we apply the rule on a correct IN-equivalent pattern and the results in Fig. 4 show promising as it successfully identified most of the directions that are not identifiable by ANM.

Real World Experiments

In this section, we applied our method to a real-world dataset, Cognition and Aging in the Chronic Fatigue Syndrome (CFS), provided by (Heins et al. 2013). The dataset contains six variables with 183 records: fatigue, focusing on symptoms (focusing), sense of control over fatigue (control), objective activity (oActivity), physical activity (pActivity), and physical functioning (functioning). It is designed to investigate the cause of fatigue. Moreover, this dataset contains a few missing values. The result is given in Fig. 5, and LCS-MD finds that pAactivity, focusing, control and functioning are the direct causes of fatigue which is consistent with the conclusion drawn by (Heins et al. 2013; Rahmadi et al. 2017). In addition, objective activity also has an indirect effect on fatigue by the pActivity variable. This is reasonable because physical activity (pActivity) also belongs to one type of objective activity. Overall, the result verifies the effectiveness of our method.

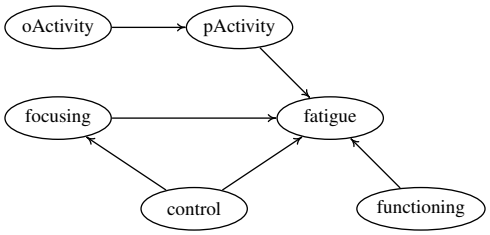


Figure 5: Result from LCS-MD in the Chronic Fatigue Syndrom (CFS) study.

Conclusion

In this work, we studied the causal discovery problem on the observational data with missing values. By taking the advantage of additive noise model, we greatly extend the identifiability results of the causal discovery methods, including that the causal skeleton is identifiable up to the case with weak self-masking missingness and the causal direction is identifiable up to an IN-equivalent pattern. Based on the theoretical results, we propose two algorithms, SM-MVPC and LCS-MD, to discover the causal structure on the data with missing values. The proposed theorems and the algorithms take a meaningful step in understanding the missingness mechanism. How to relax the assumptions and how to improve efficiency would be interesting future directions.

Acknowledgements

This research was supported in part by National Key R&D Program of China (2021ZD0111501), National Science Fund for Excellent Young Scholars (62122022), Natural Science Foundation of China (61876043, 61976052), the major key project of PCL (PCL2021A12). ZM's research was supported by the China Scholarship Council (CSC).

References

Bhattacharya, R.; Nabi, R.; Shpitser, I.; and Robins, J. M. 2020. Identification in missing data models represented by directed acyclic graphs. In *Uncertainty in Artificial Intelligence*, 1149–1158. PMLR.

Cai, R.; Qiao, J.; Zhang, Z.; and Hao, Z. 2018. Self: structural equational likelihood framework for causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Gain, A.; and Shpitser, I. 2018. Structure learning under missing data. In *International Conference on Probabilistic Graphical Models*, 121–132. PMLR.

Gao, E.; Ng, I.; Gong, M.; Shen, L.; Huang, W.; Liu, T.; Zhang, K.; and Bondell, H. 2022. MissDAG: Causal Discovery in the Presence of Missing Data with Continuous Additive Noise Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 5024–5038. Curran Associates, Inc.

Gretton, A.; Herbrich, R.; Smola, A. J.; Bousquet, O.; and Schölkopf, B. 2005. Kernel Methods for Measuring Independence. *Journal of Machine Learning Research*, 6: 2075–2129.

Heins, M. J.; Knoop, H.; Burk, W. J.; and Bleijenberg, G. 2013. The process of cognitive behaviour therapy for chronic fatigue syndrome: which changes in perpetuating cognitions and behaviour are related to a reduction in fatigue? *Journal of psychosomatic research*, 75(3): 235–241.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

Liu, Y.; and Constantinou, A. C. 2021. Greedy structure learning from data that contains systematic missing values. *arXiv preprint arXiv:2107.04184*.

Meek, C. 1995. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 403–410.

Mohan, K. 2018. On Handling Self-Masking and Other Hard Missing Data Problems. In *AAAI Symposium 2018*.

Mohan, K.; and Pearl, J. 2021. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534): 1023–1037.

Mohan, K.; Pearl, J.; and Tian, J. 2013. Graphical models for inference with missing data. *Advances in neural information processing systems*, 26.

Nabi, R.; Bhattacharya, R.; and Shpitser, I. 2020. Full law identification in graphical models of missing data: Completeness results. In *International Conference on Machine Learning*, 7153–7163. PMLR.

Osborne, J. W. 2013. *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

Peters, J.; and Bühlmann, P. 2014. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1): 219–228.

Qiao, J.; Cai, R.; Zhang, K.; Zhang, Z.; and Hao, Z. 2021. Causal discovery with confounding cascade nonlinear additive noise models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(6): 1–28.

Qiao, J.; Chen, Z.; Yu, J.; Cai, R.; and Hao, Z. 2023. Identification of Causal Structure in the Presence of Missing Data with Additive Noise Model. arXiv:2312.12206.

Rahmadi, R.; Groot, P.; Heins, M.; Knoop, H.; Heskes, T.; et al. 2017. Causality on cross-sectional data: stable specification search in constrained structural equation modeling. *Applied Soft Computing*, 52: 687–698.

Rubin, D. B. 1976. Inference and missing data. *Biometrika*, 63(3): 581–592.

Rubin, D. B. 2004. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; Kerminen, A.; and Jordan, M. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Strobl, E. V.; Visweswaran, S.; and Spirtes, P. L. 2018. Fast causal inference with non-random missingness by test-wise deletion. *International journal of data science and analytics*, 6(1): 47–62.

Tu, R.; Zhang, C.; Ackermann, P.; Mohan, K.; Kjellström, H.; and Zhang, K. 2019. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1762–1770. PMLR.

Zhang, Q.; Filippi, S.; Gretton, A.; and Sejdinovic, D. 2018. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28: 113–130.