Causal Discovery with Linear Non-Gaussian Models under Measurement Error: Structural Identifiability Results

Kun Zhang[†], Mingming Gong[⋆†], Joseph Ramsey[†], Kayhan Batmanghelich[⋆], Peter Spirtes[†], Clark Glymour[†]
[†]Department of philosophy, Carnegie Mellon University
[⋆]Department of Biomedical Informatics, University of Pittsburgh

Abstract

Causal discovery methods aim to recover the causal process that generated purely observational data. Despite its successes on a number of real problems, the presence of measurement error in the observed data can produce serious mistakes in the output of various causal discovery methods. Given the ubiquity of measurement error caused by instruments or proxies used in the measuring process, this problem is one of the main obstacles to reliable causal discovery. It is still unknown to what extent the causal structure of relevant variables can be identified in principle. This study aims to take a step towards filling that void. We assume that the underlining process or the measurement-error free variables follows a linear, non-Guassian causal model, and show that the so-called ordered group decomposition of the causal model, which contains major causal information, is identifiable. The causal structure identifiability is further improved with different types of sparsity constraints on the causal structure. Finally, we give rather mild conditions under which the whole causal structure is fully identifiable.

1 INTRODUCTION

Understanding and using causal relations among variables of interest has been a fundamental problem in various fields, including biology, neuroscience, and social sciences. Since interventions or controlled randomized experiments are usually expensive or even impossible to conduct, discovering causal information from observational data, known as causal discovery (Spirtes et al., 2001; Pearl, 2000), has been an important task and received much attention in computer science, statistics, and philosophy. Roughly speaking, methods for causal discovery are categorized into constraint-based ones, such as the PC algorithm (Spirtes et al., 2001), and score-based ones, such as Greedy Equivalence Search (GES) (Chickering, 2002).

Almost all current causal discovery methods assume that the recorded values are realizations of the variables of interest. Typically, however, the measured values are not identical to the values of the variables that they are intended to measure. The measuring process may involve nonlinear distortion, as already address by the post-nonlinear causal model (Zhang & Hyvärinen, 2009; Zhang & Chan, 2006), and may introduce a lot of error. For instance, in neuroscience the measured brain signals obtained by functional magnetic resonance (fMRI) usually contain error introduced by instruments. In this paper, we consider the so-called random measurement error model, as defined by Scheines & Ramsey (2017), in which observed variables X_i, $i = 1, ..., n$, are generated from the underlying measurement-error-free variables $\tilde{X}_i$ with additive measurement errors E_i:

$$X_i = \tilde{X}_i + E_i. \quad (1)$$

We further assume that the errors E_i are mutually independent and independent from $\tilde{X}_i$. Putting the causal model for $\tilde{X}_i$ and the random measurement error model together, we have the whole process that generates the measured data. We call this process the CAusal Model with Measurement Error (CAMME).

Generally speaking, because of the presence of measurement errors, the d-separation patterns among X_i are different from those among the underlying variables $\tilde{X}_i$. This generating process has been called the random measurement error model in (Scheines & Ramsey, 2017). According to the causal Markov condition (Spirtes et al., 2001; Pearl, 2000), observed variables X_i and the underlying variables $\tilde{X}_i$ may have different conditional independence/dependence relations and, as a consequence, the output of approaches to causal discovery that exploit conditional independence and dependence relations are unreliable in the

presence of such errors, as demonstrated in (Scheines & Ramsey, 2017). In Section 2 we will give an example to show how conditional independence/dependence between the variables is changed by measurement error, and discuss its implication in applications of causal discovery to real problems. Furthermore, because of the measurement error, the structural equation models according to which the measurement-error-free variables $\tilde{X}_i$ are generated usually do not hold for the observed variables X_i. (In fact, X_i follow error-in-variables models, for which the identifiability of the underlying causal relation is not clear.) Hence, approaches based on structural equation models, such as the linear, non-Gaussian, acyclic model (LiNGAM (Shimizu et al., 2006)), will generally fail to find the correct causal direction.

In this paper, we aim to estimate the causal model underlying the measurement-error-free variables $\tilde{X}_i$ from their observed values X_i contaminated by random measurement error. We assume linearity of the causal model and causal sufficiency relative to $\{\tilde{X}_i\}_{i=1}^n$. We particularly focus on the case where the causal structure for $\tilde{X}_i$ is represented by a Directed Acyclic Graph (DAG), although this condition can be weakened. In order to develop principled causal discovery methods to recover the causal model for $\{\tilde{X}_i\}_{i=1}^n$ from observed values of $\{X_i\}_{i=1}^n$, we have to address theoretical issues include 1) whether the causal model of interest is completely or partially identifiable from the contaminated observations and 2) what are the precise identifiability conditions.

There exist causal discovery methods, such as the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2001), to deal with confounders, i.e., hidden direct common causes. However, they cannot estimate the causal relations among the "latent" variables, which is what we aim to recover in this paper. Silva et al. (2006) and Kummerfeld et al. (2014) have provided algorithms for recovering latent variables and their causal relations when each latent variable has multiple measured effects; Shimizu et al. (2011a) further applied LiNGAM to the recovered latent variables to improve the estimated causal relations between them. Their problem is different from the measurement error setting we consider, where clustering for latent common causes is not required and each measured variable is the direct effect of a single "true" variable. As discussed in Section 3, their models can be seen as special cases of our setting.

2 EFFECT OF MEASUREMENT ERROR

Suppose we observe variables X_1, X_2, and X_3, which are generated from measurement-error-free variables $\tilde{X}_i$ according to the structure given in Figure 1. By the Markov condition and Faithfulness assumption, all three of the $\tilde{X}_i$ variables are dependent on one another, while $\tilde{X}_1$ and $\tilde{X}_3$ are conditionally independent given $\tilde{X}_2$. That conditional independence does not hold for X_i, the variables actually observable. The measurement error E_2 produces the trouble. We will treat the distributions as Gaussian purely for illustration; again, the point is general.

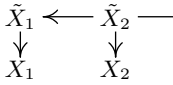

Figure 1: A linear CAMME to demonstrate the effect of measurement error on conditional independence and dependence relationships. For simplicity, we consider the special case where there is measurement error only in X_2, i.e., $X_2 = \tilde{X}_2 + E_2$, but $X_1 = \tilde{X}_1$ and $X_3 = \tilde{X}_3$.

Let $\tilde{\rho}_{12}$ be the correlation coefficient between $\tilde{X}_1$ and $\tilde{X}_2$ and $\tilde{\rho}_{13,2}$ be the partial correlation coefficient between $\tilde{X}_1$ and $\tilde{X}_3$ given $\tilde{X}_2$, which is zero. Let ρ_{12} and $\rho_{13,2}$ be the corresponding correlation coefficient and partial correlation coefficient in the presence of measurement error. We let $\tilde{\rho}_{12} = \tilde{\rho}_{23} = \tilde{\rho}$ to make the argument simpler, but the point is quite general. So we have $\rho_{13} = \tilde{\rho}_{13} = \tilde{\rho}_{12}\tilde{\rho}_{23} = \tilde{\rho}^2$. Let $\gamma = \frac{\mathbb{S}\mathrm{td}(E_2)}{\mathbb{S}\mathrm{td}(\tilde{X}_2)}$. For the data with measurement error, we have

$$\begin{aligned}
\rho_{12} &= \frac{\mathbb{C}\mathrm{ov}(X_1, X_2)}{\mathbb{V}\mathrm{ar}^{1/2}(X_1)\mathbb{V}\mathrm{ar}^{1/2}(X_2)} \\
&= \frac{\mathbb{C}\mathrm{ov}(\tilde{X}_1, \tilde{X}_2)}{\mathbb{V}\mathrm{ar}^{1/2}(\tilde{X}_1)(\mathbb{V}\mathrm{ar}(\tilde{X}_2) + \mathbb{V}\mathrm{ar}(E_2))^{1/2}} \\
&= \frac{\tilde{\rho}}{(1+\gamma^2)^{1/2}}; \\
\rho_{13,2} &= \frac{\rho_{13} - \rho_{12}\rho_{23}}{(1-\rho_{12}^2)^{1/2}(1-\rho_{23}^2)^{1/2}} \\
&= \frac{\tilde{\rho}_{13} - \frac{\tilde{\rho}_{12}\tilde{\rho}_{23}}{1+\gamma^2}}{\big(1 - \frac{\tilde{\rho}^2}{(1+\gamma^2)}\big)^{1/2}\big(1 - \frac{\tilde{\rho}^2}{(1+\gamma^2)}\big)^{1/2}} \\
&= \frac{r^2\tilde{\rho}^2}{1+\gamma^2-\tilde{\rho}^2}.
\end{aligned}$$

As the variance of the measurement error in X_2 increases, γ become larger, and ρ_{12} decreases and finally goes to zero; in contrast, $\rho_{13,2}$, which is zero for the measurement-error-free variables, is increasing and finally converges to $\tilde{\rho}^2$. See Figure 2 for an illustration. In other words, in this example as the variance of the measurement error in X_2 increases, X_1 and X_2 become more and more independent, while X_1 and X_3 are conditionally more and more dependent given X_2. However, for the measurement-error-free variables, $\tilde{X}_1$ and $\tilde{X}_2$ are dependent and $\tilde{X}_1$ and $\tilde{X}_3$ and conditionally independent given $\tilde{X}_2$. The PC algorithm and other methods that explicitly or implicitly exploit con-

ditional independence and dependence relations will find an edge between X_1 and X_3 that does not exist between X_1 and X_3 . Multiple regression of X_3 on X_1 and X_2, or X_1 on X_3 and X_2, will make the same error.

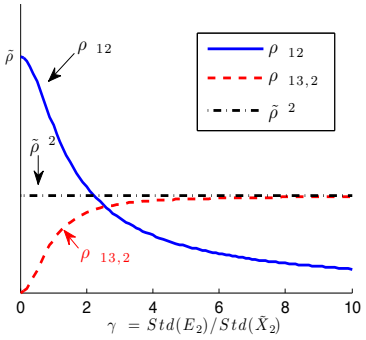

Figure 2: The correlation coefficient ρ_{12} between X_1 and X_2 and partial correlation coefficient $\rho_{13,2}$ between X_1 and X_3 given X_2 as functions of γ, the ratio of the standard deviation of measurement error to the that of $\tilde{X}_2$. We have assumed that the correlation coefficient between $\tilde{X}_1$ and $\tilde{X}_2$ and that between $\tilde{X}_2$ and $\tilde{X}_3$ are the same (denoted by $\tilde{\rho}$), and that there is measurement error only in X_2.

Roughly speaking, originally conditionally independent (or dependent) variables will become less independent (or dependent), due to the effect of measurement error. In order to correctly detect conditional independence relations between measurement-error-free variables from the observed noisy values, one may use a very small significance level (or type I error level, α) when performing conditional independence tests–the smaller the significance level, the less often the independence null hypothesis is rejected, and more pairs of variables are likely to be considered as conditionally independent. This, inevitably, risks high type II errors (i.e., conditionally dependent variable pairs are likely to be considered as independent), especially when the sample size is relatively small. Therefore it is desirable to develop principled causal discovery methods to deal with measurement error.

One might apply other types of methods instead of the constraint-based ones for causal discovery from data with measurement error. In fact, as the measurement-error-free variables are not observable, $\tilde{X}_2$ in Figure 1 is actually a confounder for observed variables. As a consequence, generally speaking, due to the effect of the confounders, the independence noise assumption underlying functional causal model-based approaches, such as the method based on the linear, non-Gaussian, acyclic model (Shimizu et al., 2006), will not hold for the observed variables any more. Figure 3 gives an

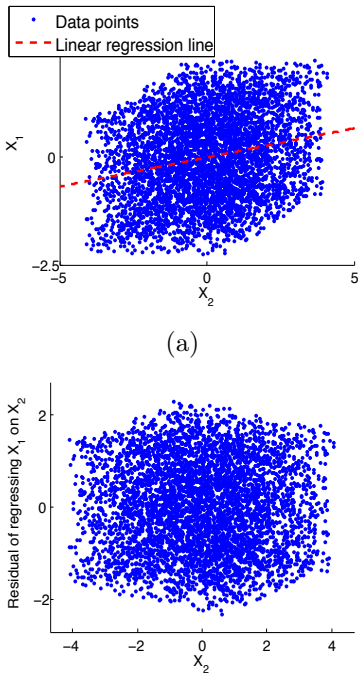

Figure 3: Illustration on how measurement error leads to dependence between regression residual and contaminated cause. (a) Scatter plot of X_2 and X_1 with measurement error in X_2 together with the regression line. (b) Scatter plot of the regression residual and X_2. Note that if we regress $\tilde{X}_1$ on $\tilde{X}_2$, the residual is independent from $\tilde{X}_2$.

illustration on this. Figure 3(a) shows the scatter plot of X_1 vs. X_2 and the regression line from X_2 to X_1, where $\tilde{X}_2$, the noise in $\tilde{X}_1$, and the measurement error E_2, are all uniformly distributed ($\rho = 0.4$, and $\gamma = 1.4$). As seen from Figure 3(b), the residual of regressing X_1 on X_2 is not independent from X_2, although the residual of regressing $\tilde{X}_1$ on $\tilde{X}_2$ is independent from $\tilde{X}_2$. As a result, the functional causal model-based approaches to causal discovery may also fail to find the causal structure of the measurement-error-free variables from their contaminated observations. The effect of measurement error on causal direction identification in the two-variable case was also studied by Wiedermann et al. (2018) under some further assumptions.

3 MODEL CANONICAL REPRESENTATION

Let $\tilde{G}$ be the acyclic causal model over $\tilde{X}_i$. Here we call it *measurement-error-free causal model*. Let $\mathbf{B}$ be the corresponding causal adjacency matrix for $\tilde{X}_i$, in which B_{ij} is the coefficient of the direct causal influence

from $\tilde{X}_j$ to $\tilde{X}_i$ and $B_{ii} = 0$. We have,

$$\tilde{\mathbf{X}} = \mathbf{B}\tilde{\mathbf{X}} + \tilde{\mathbf{E}}, \tag{2}$$

where the components of $\tilde{\mathbf{E}}$, $\tilde{E}_i$, have non-zero, finite variances. Then $\tilde{\mathbf{X}}$ is actually a linear transformation of the error terms in $\tilde{\mathbf{E}}$ because (2) implies

$$\tilde{\mathbf{X}} = \underbrace{(\mathbf{I} - \mathbf{B})^{-1}}_{\triangleq \mathbf{A}} \tilde{\mathbf{E}}. \tag{3}$$

Now let us consider two types of nodes of $\tilde{G}$, namely, leaf nodes (i.e., those that do not influence any other node) and non-leaf nodes. Accordingly, the noise term in their structural equation models also has distinct behaviors: If $\tilde{X}_i$ is a leaf node, then $\tilde{E}_i$ influences only $\tilde{X}_i$, not any other; otherwise $\tilde{E}_i$ influences $\tilde{X}_i$ and at least one other variable, $\tilde{X}_j$, $j \neq i$. Consequently, we can decompose the noise vector into two groups: $\tilde{\mathbf{E}}^L$ consists of the l noise terms that influence only leaf nodes, and $\tilde{\mathbf{E}}^{\mathrm{NL}}$ contains the remaining noise terms. Equation (3) can be rewritten as

$$\tilde{\mathbf{X}} = \mathbf{A}^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}} + \mathbf{A}^L\tilde{\mathbf{E}}^L = \tilde{\mathbf{X}}^* + \mathbf{A}^L\tilde{\mathbf{E}}^L, \tag{4}$$

where $\tilde{\mathbf{X}}^* \triangleq \mathbf{A}^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}}$, $\mathbf{A}^{\mathrm{NL}}$ and $\mathbf{A}^L$ are $n \times (n-l)$ and $n \times l$ matrices, respectively. Here both A^L and $\mathbf{A}^{\mathrm{NL}}$ have specific structures. All entries of $\mathbf{A}^L$ are 0 or 1; for each column of $\mathbf{A}^L$, there is only one non-zero entry. In contrast, each column of $\mathbf{A}^{\mathrm{NL}}$ has at least two non-zero entries, representing the influences from the corresponding non-leaf noise term.

We give a more formal way to derive the above result and make it clear how $\mathbf{A}^{\mathrm{NL}}$ and $\mathbf{A}^L$ depend on $\mathbf{B}$. For any graph $\tilde{G}$ there always exists a suitable permutation matrix, denoted by $\mathbf{\Omega}$, such that the last l elements of the permuted variables $\mathbf{\Omega}\tilde{\mathbf{X}}$ are all leaf nodes. Hence, $\mathbf{\Omega}\tilde{\mathbf{E}} = \begin{bmatrix} \tilde{\mathbf{E}}^{\mathrm{NL}} \\ \tilde{\mathbf{E}}^L \end{bmatrix}$. Accordingly, (2) implies that

$$\mathbf{\Omega}\tilde{\mathbf{X}} = \mathbf{B}_\Omega \cdot \mathbf{\Omega}\tilde{\mathbf{X}} + \mathbf{\Omega}\tilde{\mathbf{E}}, \tag{5}$$

where $\mathbf{B}_\Omega = \mathbf{\Omega}\mathbf{B}\mathbf{\Omega}^\intercal$. Since the last l variables in $\mathbf{\Omega}\tilde{\mathbf{X}}$ are leaf nodes, the last l columns of $\mathbf{B}_\Omega$ are zero. Let $\mathbf{B}_\Omega^{\mathrm{NL}}$ be the causal influence matrix for the non-leaf nodes and $\mathbf{B}_\Omega^L$ denote the causal influence from non-leaf nodes to leaf nodes. We have $\mathbf{B}_\Omega = \begin{bmatrix} \mathbf{B}_\Omega^{\mathrm{NL}} & \mathbf{0} \\ \mathbf{B}_\Omega^L & \mathbf{0} \end{bmatrix}$. Consequently,

$$(\mathbf{I} - \mathbf{B}_\Omega)^{-1} = \begin{bmatrix} (\mathbf{I} - \mathbf{B}^{\mathrm{NL}})^{-1} & \mathbf{0} \\ \mathbf{B}^L(\mathbf{I} - \mathbf{B}^N)^{-1} & \mathbf{I} \end{bmatrix}. \tag{6}$$

Combining (5) and (6) gives

$$\begin{aligned} \tilde{\mathbf{X}} &= \mathbf{\Omega}^\intercal(\mathbf{I} - \mathbf{B}_\Omega)^{-1}\mathbf{\Omega}\tilde{\mathbf{E}} \\ &= \underbrace{\mathbf{\Omega}^\intercal \cdot \begin{bmatrix} \mathbf{I} \\ \mathbf{B}^L \end{bmatrix} \cdot (\mathbf{I} - \mathbf{B}^{\mathrm{NL}})^{-1}}_{\mathbf{A}^{\mathrm{NL}}} \tilde{\mathbf{E}}^{\mathrm{NL}} + \underbrace{\mathbf{\Omega}^\intercal \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{E}}^L \end{bmatrix}}_{\mathbf{A}^L\tilde{\mathbf{E}}^L}. \end{aligned}$$

Further consider the generating process of observed variables X_i. Combining (1) and (4) gives

$$\begin{aligned} \mathbf{X} &= \tilde{\mathbf{X}}^* + \mathbf{A}^L\tilde{\mathbf{E}}^L + \mathbf{E} = \mathbf{A}^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}} + (\mathbf{A}^L\tilde{\mathbf{E}}^L + \mathbf{E}) \\ &= \mathbf{A}^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}} + \mathbf{E}^* \qquad (7) \\ &= \begin{bmatrix} \mathbf{A}^{\mathrm{NL}} & \mathbf{I} \end{bmatrix} \cdot \begin{bmatrix} \tilde{\mathbf{E}}^{\mathrm{NL}} \\ \mathbf{E}^* \end{bmatrix}, \qquad (8) \end{aligned}$$

where $\mathbf{E}^* = \mathbf{A}^L\tilde{\mathbf{E}}^L + \mathbf{E}$ and $\mathbf{I}$ denotes the identity matrix. To make it more explicit, we give how $\tilde{X}_i^*$ and E_i^* are related to the original CAMME process:

$$\tilde{X}_i^* = \begin{cases} \tilde{X}_i, & \text{if } \tilde{X}_i \text{ is not a leaf node in } \tilde{G}; \\ \tilde{X}_i - \tilde{E}_i, & \text{otherwise;} \end{cases} \text{, and} \tag{9}$$

$$E_i^* = \begin{cases} E_i, & \text{if } \tilde{X}_i \text{ is not a leaf node in } \tilde{G}; \\ E_i + \tilde{E}_i, & \text{otherwise.} \end{cases}$$

Clearly E_i^*s are independent across i, and as we shall see in Section 4, the information shared by difference X_i is still captured by $\tilde{\mathbf{X}}^*$. For each CAMME specified by (2) and (1), there always exists an observationally equivalent representation in the form of (7). We call the representation (7) the canonical representation of the CAMME (CR-CAMME).

Example Set 1 Consider the following example with three observed variables X_i, $i = 1, 2, 3$, for which $\tilde{X}_1 \to \tilde{X}_2 \leftarrow \tilde{X}_3$, with causal relations $\tilde{X}_2 = a\tilde{X}_1 + b\tilde{X}_3 + \tilde{E}_2$. That is,

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ a & 0 & b \\ 0 & 0 & 0 \end{bmatrix}, \text{ and } \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ a & 1 & b \\ 0 & 0 & 1 \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \mathbf{X} &= \tilde{\mathbf{X}} + \mathbf{E} = \tilde{\mathbf{X}}^* + \mathbf{E}^* \\ &= \begin{bmatrix} 1 & 0 \\ a & b \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \tilde{E}_1 \\ \tilde{E}_3 \end{bmatrix} + \begin{bmatrix} E_1 \\ \tilde{E}_2 + E_2 \\ E_3 \end{bmatrix} \\ &= \left[\begin{array}{cc|ccc} 1 & 0 & 1 & 0 & 0 \\ a & b & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array}\right] \cdot \begin{bmatrix} \tilde{E}_1 \\ \tilde{E}_3 \\ E_1 \\ \tilde{E}_2 + E_2 \\ E_3 \end{bmatrix}. \end{aligned}$$

In causal discovery from observations in the presence of measurement error, we aim to recover information of the measurement-error-free causal model $\tilde{G}$. Let us define a new graphical model, $\tilde{G}^*$. It is obtained by replacing variables $\tilde{X}_i$ in $\tilde{G}$ with variables $\tilde{X}_i^*$. In other words, it has the same causal structure and causal parameters (given by the $\mathbf{B}$ matrix) as $\tilde{G}$, but with

variables $\tilde{X}_i^*$ as its nodes. If we manage to estimate the structure of and the involved causal parameters in $\tilde{G}^*$, then the causal model of interest, $\tilde{G}$, is recovered. We defined the graphical model $\tilde{G}^*$ because we cannot fully estimate the distribution of measurement-error-free variables $\tilde{\mathbf{X}}$, but might be able to estimate that of $\tilde{\mathbf{X}}^*$ under proper assumptions, as shown in Section 4.

Compared to $\tilde{G}$, $\tilde{G}^*$ involves some deterministic causal relations because each leaf node is a deterministic function of its parents (the noise in leaf nodes has been removed; see (9)). For instance, suppose in $\tilde{G}^*$, $\mathrm{PA}(\tilde{X}_3^*) = \{\tilde{X}_1^*, \tilde{X}_2^*\}$, where $\mathrm{PA}(\tilde{X}_3^*)$ denotes the set of parents of $\tilde{X}_3^*$ in $\tilde{G}^*$, and that $\tilde{X}_3$ is a leaf node. Then each of $\tilde{X}_1$, $\tilde{X}_2$, and $\tilde{X}_3$ is a deterministic function of the remaining two. More generally, let $\tilde{X}_l^*$ be a leaf node in the causal graph $\tilde{G}^*$; then each of the variables in $\{\tilde{X}_l^*\} \cup \mathrm{PA}(\tilde{X}_l^*)$, denoted by $\tilde{X}_k^*$, is a deterministic function of the remaining variables.

To make it possible to identify the structure of $\tilde{G}$ from the distribution of $\mathbf{X}$, in what follows we assume the distribution of $\tilde{\mathbf{X}}^*$ satisfies the following assumption.

A0. The causal Markov condition holds for $\tilde{G}$ and the distribution of $\tilde{X}_i$ is faithful w.r.t. $\tilde{G}$. Furthermore, the distribution of $\tilde{X}_i^*$ is *non-deterministically faithful* w.r.t. $\tilde{G}^*$, in the sense that if there exists $\mathbf{S}$, a subset of $\{\tilde{X}_k^* : k \neq i, k \neq j\}$, such that neither of $\tilde{X}_i^*$ and $\tilde{X}_j^*$ is a deterministic function of $\mathbf{S}$ and $\tilde{X}_i^* \perp\!\!\!\perp \tilde{X}_j^* \,|\, \mathbf{S}$ holds, then $\tilde{X}_i^*$ and $\tilde{X}_j^*$ (or $\tilde{X}_i$ and $\tilde{X}_j$) are d-separated by $\mathbf{S}$ in $\tilde{G}^*$.

This non-deterministically faithfulness assumption excludes a particular type of parameter coupling in the causal model for $\tilde{X}_i$. in Figure 4 we give a causal model in which the causal coefficients are carefully chosen so that this assumption is violated: because $\tilde{X}_3^* = a\tilde{X}_1^* + b\tilde{X}_2^*$ and $\tilde{X}_4^* = 2a\tilde{X}_1^* + 2b\tilde{X}_2^* + E_4^*$, we have $\tilde{X}_4^* = 2\tilde{X}_3^* + E_4^*$, implying $\tilde{X}_4^* \perp\!\!\!\perp \tilde{X}_1^* \,|\, \tilde{X}_3^*$ and $\tilde{X}_4^* \perp\!\!\!\perp \tilde{X}_2^* \,|\, \tilde{X}_3^*$, which are not given by the causal Markov condition on $\tilde{G}$. We note that this non-deterministic faithfulness is defined for the distribution of the constructed variables $\tilde{X}_i^*$, not the measurement-error-free variables $\tilde{X}_i$. (Bear in mind their relationship given in (9).) This assumption is generally stronger than the faithfulness assumption for the distribution of $\tilde{X}_i$. In particular, in the causal model given in Figure 4, the distribution of $\tilde{X}_i$ is still faithful w.r.t. $\tilde{G}$. Below we call the conditional independence relationship between $\tilde{X}_i^*$ and $\tilde{X}_j^*$ given $\mathbf{S}$ where neither of $\tilde{X}_i^*$ and $\tilde{X}_j^*$ is a deterministic function of $\mathbf{S}$ *non-deterministic conditional independence*.

Now we have two concerns. One is whether essential information of the CR-CAMME is identifiable from observed values of $\mathbf{X}$. The other is what information of the original CAMME, in particular, the causal model over $\tilde{X}_i$, can be estimated from the above identifiable information of the CR-CAMME. Although the transformation from the original CAMME to a CR-CAMME is straightforward, without further knowledge there does not necessarily exist a unique CAMME corresponding to a given CR-CAMME: first, the CR-CAMME does not tell us which nodes $\tilde{X}_i$ are leaf nodes in $\tilde{G}$; second, even if $\tilde{X}_i$ is known to be a leaf node, it is impossible to separate the measurement error E_i from the noise $\tilde{E}_i$ in E_i^*. Fortunately, we are not interested in everything of the original CAMME, but only the causal graph $\tilde{G}$ and the corresponding causal influences $\mathbf{B}$. Accordingly, in the next section we will explore what information of the CR-CAMME is identifiable from the observations of $\mathbf{X}$ and how to further reconstruct necessary information of the original CAMME.

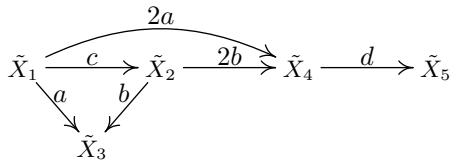

Figure 4: A specification of the causal model $\tilde{G}$ in which $\tilde{X}_i^*$ are not *non-deterministically faithful* w.r.t. $\tilde{G}$ because of parameter coupling.

In the measurement error model (1) we assumed that each observed variable X_i is generated from its own latent variable $\tilde{X}_i$. We note that in case multiple observed variables are generated from a single latent variable or a single observed variable is generated by multiple latent variables (see, e.g., Silva et al. (2006)), we can still use the CR-CAMME to represent the process. In the former case, certain rows of $\mathbf{A}^{\mathrm{NL}}$ are identical. For instance, if X_1 and X_2 are generated as noisy observations of the same latent variable, then in (7) the first two rows of $\mathbf{A}_{\mathrm{NL}}$ are identical. (More generally, if one allows different coefficients to generate them from the latent variable, the two rows are proportional to each other.) Let us then consider an example in the latter case. Suppose X_3 is generated by latent variables $\tilde{X}_1$ and $\tilde{X}_2$, for each of which there is also an observable counterpart. Write the causal model as $X_3 = f(\tilde{X}_1, \tilde{X}_2) + E_3$ and introduce the latent variable $\tilde{X}_3 = f(\tilde{X}_1, \tilde{X}_2)$, and then we have $X_3 = \tilde{X}_3 + E_3$. The CR-CAMME formulation then follows.

4 IDENTIFIABILITY IN THE LINEAR, NON-GAUSSIAN CASE

The CR-CAMME (7) has a form of the factor analysis model (FA) (Everitt, 1984), which has been a funda-

mental tool in data analysis. Accordingly, one can study the identifiability for CAMME by making use of the identifiability of FA, as reported by Zhang et al. (2017). The identifiability of FA, however, replies heavily on the assumption that there are a relatively large number of leaf variables in the causal graph $\tilde{G}$ (Bekker & ten Berge, 1997), which seems rather strong. Moreover, it has been shown that second-order statistics usually is not informative enough to recover a unique causal model (Spirtes et al., 2001). Interestingly, we show that the identifiability results can greatly benefit from the non-Gaussianity assumption on the data. In this paper we make the following assumption on the distribution of $\tilde{E}_i$:

A1. All $\tilde{E}_i$ are non-Gaussian.

We note that under the above assumption, $\mathbf{A}^{\mathrm{NL}}$ in (8) can be estimated up to the permutation and scaling indeterminacies (including the sign indeterminacy) of the columns, as given in the following lemma. This can be achieved by using overcomplete Independent Component Analysis (ICA) (Hyvärinen et al., 2001).

Lemma 1. *Suppose assumption A1 holds. Given* $\mathbf{X}$ *which is generated according to (8),* $\mathbf{A}^{\mathrm{NL}}$ *is identifiable up to permutation and scaling of columns as the sample size* $N \to \infty$.

Proof. This lemma is implied by Theorem 10.3.1 in (Kagan et al., 1973) or Theorem 1 in (Eriksson & Koivunen, 2004). □

What information of the causal structure $\tilde{G}$ can we recover? Can we apply existing methods for causal discovery based on LiNGAM, such as ICA-LiNGAM (Shimizu et al., 2006) and Direct-LiNGAM (Shimizu et al., 2011b), to recover it? LiNGAM assumes that the system is non-deterministic: each variable is generated as a linear combination of its direct causes plus a non-degenerate noise term. As a consequence, the linear transformation from the vector of observed variables to the vector of independent noise terms is a square matrix; ICA-LiNGAM applies certain operations to this matrix to find the causal model, and Direct-LiNGAM estimates the causal ordering by enforcing the property that the residual of regressing the effect on the root cause is always independent from the root cause.

In our case, $\mathbf{A}^{\mathrm{NL}}$, the essential part of the mixing matrix in (8), is $n \times r$, where $r < n$. In other words, for some of the variables $\tilde{X}_i^*$, the causal relations are deterministic. (In fact, if $\tilde{X}_k$ is a leaf node in $\tilde{G}$, $\tilde{X}_k^*$ is a deterministic function of $\tilde{X}_k$'s direct causes.) As a consequence, unfortunately, the above causal analysis methods based on LiNGAM, including ICA-LiNGAM and Direct-LiNGAM, do not apply. We will see how to recover information of $\tilde{G}$ by analyzing the estimated $\mathbf{A}^{\mathrm{NL}}$.

We will show that some group structure and the group-wise causal ordering in $\tilde{G}$ can always be recovered. Before presenting the results, let us define the following ordered group decomposition according to causal structure $\tilde{G}$.

Definition 2 (ordered group decomposition). *Consider the causal model* $\tilde{G}^*$. *Decompose all involved nodes into disjoint groups in the following way. First put all leaf nodes which share the same direct-and-only-direct cause in the same group; further incorporate the corresponding direct-and-only-direct cause in the same group. Here we say a node* $\tilde{X}_i^*$ *is the "direct-and-only-direct" cause of* $\tilde{X}_j^*$ *if and only if* $\tilde{X}_i^*$ *is a direct cause of* $\tilde{X}_j^*$ *and there is no other directed path from* $\tilde{X}_i^*$ *to* $\tilde{X}_j^*$. *After forming all groups each of which involves at least one leaf node, each of the remaining nodes forms a separate group.* **Each node is guaranteed to be in one and only one group**. *We call the set of all such groups* **ordered according to the causal ordering of the non-leaf nodes** *in DAG* $\tilde{G}^*$ *an ordered group decomposition of* $\tilde{G}^*$, *denoted by* $\mathcal{G}_{\tilde{G}^*}$.

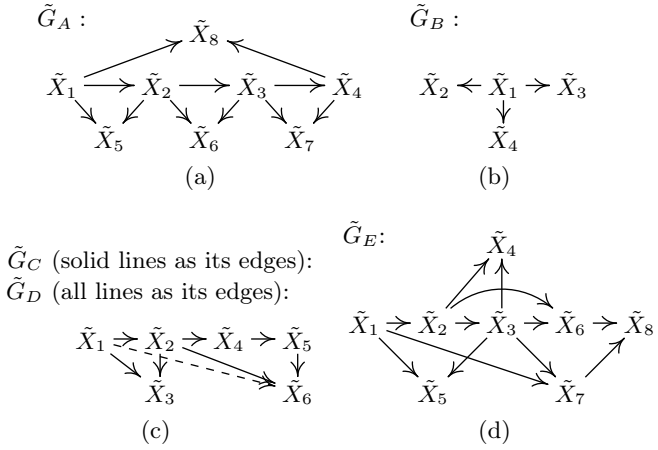

Figure 5: A set of causal DAGs $\tilde{G}$ as illustrative examples. (a) DAG $\tilde{G}_A$. (b) $\tilde{G}_B$. (c) Two DAGs $\tilde{G}_C$ and $\tilde{G}_D$. (d) $\tilde{G}_E$.

Example Set 2 As seen from the process of ordered group decomposition, each non-leaf node is in one and only one ordered group, and it is possible for multiple leaf nodes to be in the same group. Therefore, in total there are $(n - l)$ ordered groups. For example, for $\tilde{G}_A$ given in Figure 5(a), a corresponding group structure for the corresponding $\tilde{G}^*$ is $\mathcal{G}_{\tilde{G}_A^*} = (\{\tilde{X}_1^*\} \to \{\tilde{X}_2^*, \tilde{X}_5^*\} \to \{\tilde{X}_3^*, \tilde{X}_6^*\} \to \{\tilde{X}_4^*, \tilde{X}_7^*, \tilde{X}_8^*\})$, and for $\tilde{G}_B$ in Figure 5(b), there is only one group:

$\mathcal{G}_{\tilde{G}_B^*} = (\{\tilde{X}_1^*, \tilde{X}_2^*, \tilde{X}_3^*, \tilde{X}_4^*\})$. For both $\tilde{G}_C$ and $\tilde{G}_D$, given in Figure 5(c), an ordered group decomposition is $(\{\tilde{X}_1^*\} \to \{\tilde{X}_2^*, \tilde{X}_3^*\} \to \{\tilde{X}_4^*\} \to \{\tilde{X}_5^*, \tilde{X}_6^*\})$.

Note that the causal ordering and the ordered group decomposition of given variables according to the graphical model $\tilde{G}^*$ may not be unique (this will actually give rise to the possibility of distinguishing between the non-leaf and leaf node in the group, as shown next). For instance, if $\tilde{G}^*$ has only two variables $\tilde{X}_1^*$ and $\tilde{X}_2^*$ which are not adjacent, both decompositions $(\{\tilde{X}_1^*\} \to \{\tilde{X}_2^*\})$ and $(\{\tilde{X}_2^*\} \to \{\tilde{X}_1^*\})$ are correct. Consider $\tilde{G}^*$ over three variables, $\tilde{X}_1^*, \tilde{X}_2^*, \tilde{X}_3^*$, where $\tilde{X}_1^*$ and $\tilde{X}_2^*$ are not adjacent and are both causes of $\tilde{X}_3^*$; then both $(\{\tilde{X}_1^*\} \to \{\tilde{X}_2^*, \tilde{X}_3^*\})$ and $(\{\tilde{X}_2^*\} \to \{\tilde{X}_1^*, \tilde{X}_3^*\})$ are valid ordered group decompositions.

We first present a procedure to construct the ordered group decomposition and the causal ordering among the groups from the estimated $\mathbf{A}^{\mathrm{NL}}$. We will further show that the recovered ordered group decomposition is always asymptotically correct under assumption A1.

4.1 Construction and Identifiability of ordered Group Decomposition

First of all, Lemma 1 tells us that $\hat{\mathbf{A}}^{\mathrm{NL}}$ in (8) is identifiable up to permutation and scaling columns. Let us start with the asymptotic case, where the columns of the estimated $\mathbf{A}^{\mathrm{NL}}$ from values of X_i are a permuted and rescaled version of the columns of $\mathbf{A}^{\mathrm{NL}}$. In what follows the permutation and rescaling of the columns of $\mathbf{A}^{\mathrm{NL}}$ does not change the result, so below we just work with the true $\mathbf{A}^{\mathrm{NL}}$, instead of its estimate.

$\tilde{X}_i^*$ and $\tilde{X}_i$ follow the same causal DAG, $\tilde{G}$, and $\tilde{X}_i^*$ are causally sufficient, although some variables among them (corresponding to leaf nodes in $\tilde{G}^*$) are determined by their direct causes. Let us find the causal ordering of $\tilde{X}_i^*$. If there are no deterministic relations and the values of $\tilde{X}_i^*$ are given, the causal ordering can be estimated by recursively performing regression and checking independence between the regression residual and the predictor (Shimizu et al., 2011b). Specifically, if one regresses all the remaining variables on the root cause, the residuals are always independent from the predictor (the root cause). After detecting a root cause, the residuals of regressing all the other variables on the discovered root cause are still causally sufficient and follow a DAG. One can repeat the above procedure to find a new root cause over such regression residuals, until no variable is left.

However, in our case we have access to $\mathbf{A}^{\mathrm{NL}}$ but not the values of $\tilde{X}_i^*$. Fortunately, the independence between regression residuals and the predictor can still be checked by analyzing $\mathbf{A}^{\mathrm{NL}}$. Recall that $\tilde{\mathbf{X}}^* = \mathbf{A}^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}}$, where the components of $\tilde{\mathbf{E}}^{\mathrm{NL}}$ are independent. Without loss of generality, here we assume that all components of $\tilde{\mathbf{E}}^{\mathrm{NL}}$ are standardized, i.e., they have a zero mean and unit variance. Denote by $\mathbf{A}_{i\cdot}^{\mathrm{NL}}$ the ith row of $\mathbf{A}^{\mathrm{NL}}$. We have $\mathbb{E}[\tilde{X}_j^*\tilde{X}_i^*] = \mathbf{A}_{j\cdot}^{\mathrm{NL}}\mathbf{A}_{i\cdot}^{NL\intercal}$ and $\mathbb{E}[\tilde{X}_i^{*2}] = \mathbf{A}_{j\cdot}^{\mathrm{NL}}\mathbf{A}_{i\cdot}^{NL\intercal} = ||\mathbf{A}_{i\cdot}^{\mathrm{NL}}||^2$. The regression model for $\tilde{X}_j^*$ on $\tilde{X}_i^*$ is

$$\tilde{X}_j^* = \frac{\mathbb{E}[\tilde{X}_j^*\tilde{X}_i^*]}{\mathbb{E}[\tilde{X}_i^{*2}]}\tilde{X}_i^* + R_{j\leftarrow i} = \frac{\mathbf{A}_{j\cdot}^{\mathrm{NL}}\mathbf{A}_{i\cdot}^{\mathrm{NL}\intercal}}{||\mathbf{A}_{i\cdot}^{\mathrm{NL}}||^2}\tilde{X}_i^* + R_{j\leftarrow i}.$$

Here the residual can be written as

$$\begin{aligned} R_{j\leftarrow i} &= \tilde{X}_j^* - \frac{\mathbf{A}_{j\cdot}^{\mathrm{NL}}\mathbf{A}_{i\cdot}^{\mathrm{NL}\intercal}}{||\mathbf{A}_{i\cdot}^{\mathrm{NL}}||^2}\tilde{X}_i^* \\ &= \underbrace{\left(\mathbf{A}_{j\cdot}^{\mathrm{NL}} - \frac{\mathbf{A}_{j\cdot}^{\mathrm{NL}}\mathbf{A}_{i\cdot}^{\mathrm{NL}\intercal}\mathbf{A}_{i\cdot}^{\mathrm{NL}}}{||\mathbf{A}_{i\cdot}^{\mathrm{NL}}||^2}\right)}_{\triangleq\alpha_{j\leftarrow i}}\tilde{\mathbf{E}}^{\mathrm{NL}}. \end{aligned} \tag{10}$$

If for all j, $R_{j\leftarrow i}$ is either zero or independent from $\tilde{X}_i^*$, we consider $\tilde{X}_i^*$ as the current root cause and put it and all the other variables which are deterministically related to it in the first group, which is a *root cause group*. Now the problem is whether we can check for independence between nonzero residuals $R_{j\leftarrow i}$ and the predictor $\tilde{X}_i^*$. Interestingly, the answer is yes, as stated in the following proposition.

Proposition 3. *Suppose assumption A1 holds. For variables $\tilde{\mathbf{X}}^*$ generated by (7), regression residual $R_{j\leftarrow i}$ given in (10) is independent from variable $\tilde{X}_i^*$ if and only if*

$$\left|\left|\alpha_{j\leftarrow i} \circ \mathbf{A}_{i\cdot}^{\mathrm{NL}}\right|\right|_2 = 0, \tag{11}$$

where $\circ$ denotes entrywise product.

So we can check for independence between the predictor and regression residual as if the values of $\tilde{\mathbf{X}}^*$ were given. Consequently, we can find the root cause group.

We then consider the residuals of regressing all the remaining variables $\tilde{X}_k^*$ on the discovered root cause as a new set of variables. Note that like the variables $\tilde{X}_j^*$, these variables are again linear mixtures of $\tilde{E}_i$. Repeating the above procedure on this new set of variables will give the second root cause and its ordered group. Applying this procedure repeatedly until no variable is left finally discovers all ordered groups following the causal ordering. The constructed ordered group decomposition is asymptotically correct, as stated in the following proposition. We denote by `OICA+Reg` the above two-stage procedure: we first apply overcomplete ICA to find an estimate of $\mathbf{A}^{\mathrm{NL}}$, and then do regression and check for independence between the residuals and the current candidate root cause by analyzing $\mathbf{A}^{\mathrm{NL}}$.

Proposition 4. (Identifiable ordered group decomposition) *Let X_i be generated by the CAMME with the corresponding measurement-error-free variables generated by the causal DAG $\tilde{G}$ and suppose assumptions A0 and A1 hold. The ordered group decomposition constructed by the above procedure is asymptotically correct, in the sense that as the sample size $N \to \infty$, if non-leaf node $\tilde{X}_i$ is a cause of non-leaf node $\tilde{X}_j$, then the ordered group which $\tilde{X}_i$ is in precedes the group which $\tilde{X}_j$ belongs to. However, the causal ordering among the nodes within the same ordered group may not be identifiable.*

The result of Proposition 4 applies to any DAG structure $\tilde{G}$. Clearly, the identifiability can be naturally improved if additional assumptions on the causal structure $\tilde{G}$ hold. In particular, to recover information of $\tilde{G}$, it is essential to answer the following questions.

- Can we determine which nodes in an ordered group are leaf nodes?
- Can we find the causal edges into a particular node?

Below we will show that under rather mild assumptions, the answers to both questions are yes.

4.2 Identifying Leaf Nodes and Individual Causal Edges

If for each ordered group we can determine which variable is the non-leaf node, the causal ordering among the variables $\tilde{X}_i^*$ is then fully known. The causal structure in $\tilde{G}^*$ as well as the causal model can then be readily estimated by regression: for a leaf node, its direct causes are those non-leaf nodes that determine it; for a non-leaf node, we can regress it on all non-leaf nodes that precede it according to the causal ordering, and those predictors with non-zero linear coefficients are its parents. This way the structure can be estimated uniquely under Assumption A0, although whether the causal parameters in the causal model are uniquely identifiable is another issue for investigation.

Now the goal is to see whether it is possible to find out which variables in a given ordered group are leaf nodes; if all leaf nodes are found, then the remaining one is the (only) non-leaf node in the considered ordered group. Below we will show that it is possible to find leaf nodes by "looking backward" or "looking forward"; the former makes use of the parents of the variables in the considered group, and the latter exploits the fact leaf nodes do not have any child.

Proposition 5. (Leaf node determination by "looking backward") *Suppose the observed data were generated by the CAMME where Assumptions A0 and A1 hold.*[1] *Let the sample size $N \to \infty$. Then if assumption A2 holds, leaf node O is correctly identified from observations of $\mathbf{X}$ (more specifically, from the estimated $\mathbf{A}^{\mathrm{NL}}$ or the distribution of $\tilde{\mathbf{X}}^*$).*

A2. According to $\tilde{G}^*$, for leaf node O in the considered ordered group $g^{(k)}$, at least one of its parents is not a parent of the non-leaf node in $g^{(k)}$ or some other leaf node in $g^{(k)}$.

Example Set 3 Suppose Assumptions A0 and A1 hold.

- For $\tilde{G}_A$ in Figure 5(a), assumption A2 holds for $\tilde{X}_7^*$ and $\tilde{X}_8^*$ in the ordered group $\{\tilde{X}_4^*, \tilde{X}_7^*, \tilde{X}_8^*\}$: each of them has a parent which is not a parent of the other; so both of them are identified to be leaf nodes from the estimated $\mathbf{A}^{\mathrm{NL}}$ or the distribution of $\tilde{\mathbf{X}}^*$, and $\tilde{X}_4^*$ can then be determined as a non-leaf node.
- For $\tilde{G}_B$, we cannot detect which node is a leaf node or a non-leaf node.
- For both $\tilde{G}_C$ and $\tilde{G}_D$ in Figure 5(c), $\tilde{X}_6^*$, in the ordered group $\{\tilde{X}_5^*, \tilde{X}_6^*\}$, follows assumption A2 and can be found to be a leaf node from the matrix $\mathbf{A}^{\mathrm{NL}}$; accordingly, $\tilde{X}_5^*$ has to be a non-leaf node.
- For $\tilde{G}_E$ in Figure 5(d), assumption A2 holds for all leaf nodes, $\tilde{X}_4^*$, $\tilde{X}_5^*$, and $\tilde{X}_8^*$, which can then be found to be leaf nodes.

We can also determine leaf nodes by looking at the relationships between the considered variables and the variables causally following them, as stated in the following proposition.

Proposition 6. (Leaf node determination by "looking forward") *Suppose the observed data were generated by the CAMME where Assumptions A0 and A1 hold. Then as the sample size $N \to \infty$, we can correctly identify the leaf node U in the considered ordered group $g^{(k)}$ from values of $\mathbf{X}$ if assumption A3 holds for it:*

A3. For leaf node U in $g^{(k)}$, there exists at least one node causally following $g^{(k)}$ that 1) is d-separated from U by a subset of variables in $g^{(1)} \cup g^{(2)} ... \cup g^{(k)} \setminus \{U\}$ which does not include all parents of U and 2) is a child of the non-leaf node in $g^{(k)}$.

[1] In this non-Gaussian case (implied by assumption A1), the result reported in this proposition may still hold if one avoids the non-deterministic faithfulness assumption and assumes a weaker condition; however, for simplicity of the proof we currently still assume non-deterministic faithfulness.

Example Set 4 Let Assumptions A0 and A1 hold.

- For data generated by $\tilde{G}_A$ in Figure 5(a), we already found $\tilde{X}_4^*$ in ordered group $\{\tilde{X}_4^*, \tilde{X}_7^*, \tilde{X}_8^*\}$ to be a non-leaf node because of Proposition 5. Proposition 6 further indicates that $\tilde{X}_2^*$ (in group $\{\tilde{X}_2^*, \tilde{X}_5^*\}$) and $\tilde{X}_3^*$ (in group $\{\tilde{X}_3^*, \tilde{X}_6^*\}$) are non-leaf nodes, and all leaf nodes are identified.
- For $\tilde{G}_B$ in Figure 5(b), there is only one ordered group, and it does not provide further information by looking "backward" or "forward", and it is impossible to find the non-leaf node with Proposition 5 or 6.
- For both $\tilde{G}_C$ and $\tilde{G}_D$ in Figure 5(c), $\tilde{X}_6^*$ was found to be a leaf node due to Proposition 5; thanks to Proposition 6, the other leaf node, $\tilde{X}_3^*$, was also detected. In particular, in $\tilde{G}_C$, for leaf node $\tilde{X}_3^*$ both $\tilde{X}_4^*$ and $\tilde{X}_6^*$ satisfy the two conditions in Assumption A3; however, in $\tilde{G}_D$, for leaf node $\tilde{X}_3^*$ only $\tilde{X}_4^*$ satisfies them. All leaf nodes were successfully found.
- For $\tilde{G}_E$ in Figure 5(d), Proposition 5 already allows us to identify all leaf nodes, $\tilde{X}_4^*$, $\tilde{X}_5^*$, and $\tilde{X}_8^*$. The assumptions in Propositions 5 and 6 are not exclusive: Assumption A3 also holds for $\tilde{X}_4^*$ (for it $\tilde{X}_7^*$ satisfies the two conditions), we can alternatively identify this leaf node by making use of Proposition 6.

For contaminated data generated by any of $\tilde{G}_A$, $\tilde{G}_C$, $\tilde{G}_D$, and $\tilde{G}_E$, now we can find all leaf nodes in the measurement-error-free causal model. One can then immediately estimate the whole structure of the measurement-error-free model.

The above two propositions are about the identifiably of leaf nodes in the measurement-error-free causal model. By applying them to all leaf nodes, we have (sufficient) conditions under which the causal graph of $\tilde{G}$ is fully identifiable.

Proposition 7. (Full identifiability) *Suppose the observed data were generated by the CAMME where Assumptions A0 and A1 hold. Assume that for each leaf node in $\tilde{G}^*$, at least one of the the two assumptions, A2 and A3, holds. Then as the sample size $N \to \infty$, the causal structure $\tilde{G}$ is fully identifiable from the observations with random measurement error.*

In the general case, the causal structure $\tilde{G}$ might not be fully identifiable, and the above propositions may allow partial identifiability of the underlying causal structure. Roughly speaking, the ordered group decomposition is identifiable in the non-Gaussian case; with Propositions 5 and 6 one can further identify some leaf nodes as well as their parents.

5 CONCLUSION AND DISCUSSIONS

The measured values of variables of interest in various fields, including the social sciences, neuroscience, and biology, are often contaminated by measurement error. Unfortunately, the output of existing causal discovery methods is sensitive to the existence of measurement error, and it is desirable to develop causal discovery methods that can estimate the causal model for the measurement-error-free variables without using much prior knowledge about the measurement error. To this end, this paper investigates identifiability conditions for the underlying measurement-error-free causal structure given contaminated observations. We have shown that under appropriate conditions, the causal structure of interest is partially or even fully identifiable.

We formulated four assumptions. Assumption A0 is about the Markov condition and non-deterministic faithfulness assumption for causal model $\tilde{G}^*$. Assumption A1 is about the distribution of the underlying noise terms in the causal process. The remaining two are about particular types of "sparsity" of the underlying causal graph. We note that in principle, all assumptions except A0 are testable from the observed data. This suggests that it is possible to develop practical causal discovery methods to deal with measurement error that are able to produce reliable information at least in the asymptotic case. In addition, it is worth noting that some involved assumptions may be weakened. For instance, faithfulness is not required to find the correct ordered group decomposition, but just needed for detecting leaf nodes in the ordered groups. Suppose Assumptions A0 and A1 hold; we conjecture that the necessary and sufficient condition for the non-leaf node to be identifiable is that at least one of the two assumptions, A2 and A3, holds. To falsify or prove this conjecture is part of our future work.

It is worth noting that various kinds of background knowledge of the causal model may further help improve the identifiability of the measurement-error-free causal model. For instance, if one knows that all causal coefficients are smaller than one in absolute value, then the measurement-error-free causal model in Figure 5(b) is immediately identifiable from contaminated data. Our future research further includes 1) establishing identifiability conditions that allow cycles in the measurement-error-free causal model in light of ubiquity of cycles in causal models, 2) developing computationally efficient algorithms for causal discovery under measurement error based on the established theory, and 3) proposing efficient methods for particular cases where each measurement-error-free variable has multiple measured effects or multiplied measurement-error-free variables generate a single measured effect.

References

Bekker, P. A. and ten Berge, J. M. F. Generic global indentification in factor analysis. *Linear Algebra and its Applications*, 264:255–263, 1997.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Eriksson, J. and Koivunen, V. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

Everitt, B. S. *An introduction to latent variable models*. London: Chapman and Hall, 1984.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

Kagan, A. M., Linnik, Y. V., and Rao, C. R. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

Kummerfeld, E., Ramsey, J., Yang, R., Spirtes, P., and Scheines, R. Causal clustering for 2-factor measurement models. In Calders, T., Esposito, F., Hüllermeier, R., and Meo, R. (eds.), *Proc. ECML PKDD*, pp. 34–49, 2014.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

Scheines, R. and Ramsey, J. Measurement error and causal discovery. In *Proc. CEUR Workshop 2016*, pp. 1–7, 2017.

Shimizu, S., Hoyer, P.O., Hyvärinen, A., and Kerminen, A.J. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7: 2003–2030, 2006.

Shimizu, S., Hoyer, P. O., and Hyvärinan, A. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72:2024–2027, 2011a.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. Directlingam: Adirect method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, pp. 1225–1248, 2011b.

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2001.

Wiedermann, W., Merkle, E. C., and von Eye, A. Direction of dependence in measurement error models. *British Journal of Mathematical and Statistical Psychology*, 71:117–145, 2018.

Zhang, K. and Chan, L. Extensions of ICA for causality discovery in the hong kong stock market. In *Proc. 13th International Conference on Neural Information Processing (ICONIP 2006)*, 2006.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., and Glymour, C. Causal discovery in the presence of measurement error: Identifiability conditions. In *UAI 2017 Workshop on Causality: Learning, Inference, and Decision-Making*, 2017.