Causal Discovery in Linear Models with Unobserved Variables and Measurement Error

Yuqin Yang[*]
Georgia Institute of Technology

Mohamed Nafea
Missouri University of Science & Technology

Negar Kiyavash
École Polytechnique Fédérale de Lausanne (EPFL)

Kun Zhang
Carnegie Mellon University & MBZUAI

AmirEmad Ghassami
Boston University

Abstract

The presence of unobserved common causes and the presence of measurement error are two of the most limiting challenges in the task of causal structure learning. Ignoring either of the two challenges can lead to detecting spurious causal links among variables of interest. In this paper, we study the problem of causal discovery in systems where these two challenges can be present simultaneously. We consider linear models which include four types of variables: variables that are directly observed, variables that are not directly observed but are measured with error, the corresponding measurements, and variables that are neither observed nor measured. We characterize the extent of identifiability of such model under separability condition (i.e., the matrix indicating the independent exogenous noise terms pertaining to the observed variables is identifiable) together with two versions of faithfulness assumptions and propose a notion of observational equivalence. We provide graphical characterization of the models that are equivalent and present a recovery algorithm that could return models equivalent to the ground truth.

1 Introduction

Causal structure learning, also known as causal discovery, from observational data has been studied extensively in the literature. The majority of work assume that there are no unobserved variables in the system that can affect more than one other variable, i.e., no unobserved common causes, and that variables are measured without error. This leads to identification of the underlying structure up to Markov equivalence class in general [15, 3], and complete identification of the structure under further model assumptions such as assuming a linear non-Gaussian model (LiNGAM, [12, 13]). However, in majority of real-world settings, the researcher will not be able to observe all the relevant variables and hence cannot rule out the presence of unobserved common causes, and further many of the variables may have been measured with error. This necessitates approaches for causal discovery capable of dealing with these two important challenges.

In this work, we study the problem of causal discovery from observational data in the presence of both aforementioned challenges, i.e., in settings with both unobserved common causes and measurement error. We consider a special type of linear structural equation model (SEM) as the underlying data generating process which includes four types of variables: variables that are directly observed (called *observed variables*), variables that are not directly observed but are measured with error (called *measured variables*), the corresponding *measurement variables*, and variables that are neither

[*]Correspondence to: yuqinyang767@gmail.com.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

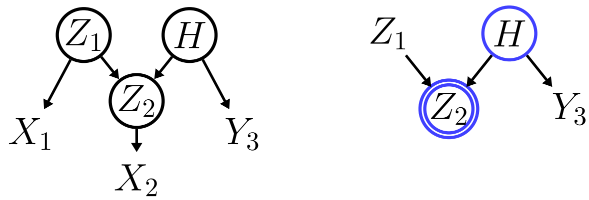

Figure 1: **Left**: Diagram of an LV-SEM-ME (see Appendix B.2 for more details). We observe Y_3 and (X_1, X_2), which are noisy measurements of (Z_1, Z_2). **Right**: Diagram of the corresponding canonical model: Z_2 is an mleaf variable; H is an unobserved variable; Z_1, Y_3 are cogent variables.

directly observed nor measured (called *unobserved variables*). We refer to this model as *linear latent variable SEM with measurement error* (linear LV-SEM-ME). We study the identifiability of linear LV-SEM-MEs in a setup where the independent exogenous noise terms that causally (directly or indirectly) affect each observed variable can be distinguished from each other. That is, the mixing matrix of the linear system that transforms exogenous noise terms to observed variables is recoverable up to permutation and scaling of the columns. This can be satisfied, for example, if all independent exogenous noise terms are non-Gaussian. We note that measurement error challenge is essentially a special case of unobserved variable challenge, in which we observe a measurement of an underlying unobserved variable of interest. Yet, the observed measurement variable usually have special properties (such as not being affected by other variables) that can be leveraged to improve the identification power. Hence, our point of view in this work is to allow for coexistence of the challenges of unobserved common causes and measurement error, while leveraging the properties of the measurement variables to improve identification.

We study the identifiability of linear LV-SEM-ME under two faithfulness assumptions. The first assumption prevents existence of zero total causal effects of a variable on its descendants, which we refer to as *conventional faithfulness assumption* as it is widely assumed in the literature. The second assumption prevents additional parameter cancellation or proportionality among specific edges, which we refer to as *LV-SEM-ME faithfulness assumption*. Both assumptions are mild in that their violations are zero-probability events. We show that under conventional faithfulness assumption, the model can be identified up to an equivalence class characterized by an ordered grouping of the variables which we call *ancestral ordered grouping* (AOG). Further, under LV-SEM-ME faithfulness assumption, the model can be learned up to an equivalence class characterized by a more refined ordered grouping which we call *direct ordered grouping* (DOG). We provide a graphical characterization of the elements of the equivalence class in which the induced graph on each ordered group includes a star structure where any member (variable) of the group is a potential center of the star. Specifically, every element in the equivalence class corresponds to distinct assignments of the centers of the star graphs, yet it possesses the same ordered grouping of variables and the same unlabelled structure on each group as the rest of the elements in the equivalence class. Models in the same AOG equivalence class are consistent with the same set of causal orders among groups, and models in the same DOG equivalence class share the same unlabeled graph structure, i.e., the causal diagrams are isomorphic. Lastly, we provide a recovery algorithm that could return all models in the AOG and DOG equivalence classes. Previous work presented in [19] only considered confounders and measurement error separately. Moreover, a different definition for AOG and DOG was used.

2 Model description

We start with a formal definition of the model that we consider in this work.

Definition 1 (General linear LV-SEM-ME) *A general linear LV-SEM-ME consists of two sets of variables $\mathcal{V}$ and $\mathcal{X}$. Variables in $\mathcal{V}$ can be arranged in a causal order, and each variable $V_i \in \mathcal{V}$ is generated as a linear combination of a subset $Pa(V_i) \subset \mathcal{V}$ (called its direct parents), plus an exogenous noise term N_{V_i}, where $\{N_{V_i}\}_{V_i \in \mathcal{V}}$ are jointly independent. Further, $\mathcal{V}$ can be partitioned into three sets $\mathcal{Y}$, $\mathcal{Z}$ and $\mathcal{H}$. Variables in $\mathcal{Y}$ are observed (without error). Variables in $\mathcal{Z}$ are measured with error, where each variable $X_i \in \mathcal{X}$ is a noisy measurement of one corresponding variable $Z_i \in \mathcal{Z}$ plus an exogenous noise term N_{X_i} (which we call measurement error of Z_i). Variables* *in $\mathcal{H}$ are neither observed nor measured with error. We refer to variables in $\mathcal{H}$, $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{X}$ as unobserved variables, observed variables, measured variables, and measurements, respectively.*

We define a measured leaf variable (mleaf variable) as a measured variable in $\mathcal{Z}$ that has no other children besides its noisy measurement. We define a *cogent variable* as a variable in $\mathcal{Z} \cup \mathcal{Y}$ that is not an mleaf. As mentioned earlier, we study the problem of recovering the linear LV-SEM-ME from observations of $(\mathcal{Y}, \mathcal{X})$. We focus on recovering the subset of linear LV-SEM-ME, called *canonical LV-SEM-ME*, defined as follows. We add two restrictions on the model for the sake of model identifiability based on observational data, where both have been considered in literature [6, 20, 19]. See Appendix B.1 for detailed discussion about the restrictions on the canonical form.

Definition 2 (Canonical linear LV-SEM-ME) *A canonical LV-SEM-ME is an LV-SEM-ME where variables in $\mathcal{H}$ are roots and confounders, and mleaf variables do not have exogenous noise terms.*

The matrix form of the canonical LV-SEM-ME can be written as

$$H = N_H, \tag{1a}$$

$$\begin{bmatrix} Z^L \\ Z^C \\ Y \end{bmatrix} = \mathbf{B}H + \begin{bmatrix} \mathbf{D} \\ \mathbf{C}_Z \\ \mathbf{C}_Y \end{bmatrix} \begin{bmatrix} Z^C \\ Y \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ N_{Z^C} \\ N_Y \end{bmatrix}, \tag{1b}$$

$$X = \begin{bmatrix} Z^L \\ Z^C \end{bmatrix} + \begin{bmatrix} N_{X^L} \\ N_{X^C} \end{bmatrix}, \tag{1c}$$

where H, Y, X represent the vector of unobserved variables, observed variables, measurements, respectively. Z^L represent the vector of mleaf variables, and Z^C represent the vector of measured variables that are not mleaf variables. N_H, N_Y, N_{Z^C} are the noise vectors corresponding to H, Y, Z^C, respectively. N_{X^L} (resp. N_{X^C}) represent the measurement error of the variables in Z^L (resp. Z^C). $\mathbf{B}$ represent the causal connections from unobserved variables H to $[Z^L; Z^C; Y]$, $\mathbf{C}$ represent the causal connections among cogent variables, which is partitioned into $[\mathbf{C}_Z; \mathbf{C}_Y]$ according to $[Z^C; Y]$. $\mathbf{D}$ represent the causal connections from cogent variables to mleaf variables. Note that the right hand side of Equation (1b) is not a function of Z^L since variables on the left hand side do not have mleaf variables as their parents.

We define the causal diagram of a linear LV-SEM-ME as a directed graph, where the nodes are all variables in $\mathcal{V} \cup \mathcal{X}$. For any two variables $W_1, W_2 \in \mathcal{V} \cup \mathcal{X}$, there is a directed edge from W_1 to W_2 if and only if $W_1 \in Pa(W_2)$. Due to the causal order of $\mathcal{V} \cup \mathcal{X}$, the causal diagram is acyclic.

Problem description We consider a setting with known *observability indicators*, that is, we know whether each variable is observed without error (i.e., belongs to Y) or measured with error (i.e., belongs to X). Suppose we have n i.i.d. observations of the variables $\{X, Y\}$. The task is to recover all linear LV-SEM-MEs which have the same observational distribution up to the noise distributions.

3 Identification analysis

In this section we study identification for our model of interest. We first consider the problem for two special submodels in Section 3.2: 1) If $\mathcal{H} = \emptyset$, i.e., all unobserved variables have noisy measurements, then the model is linear SEM-ME. 2) If $\mathcal{Z} = \emptyset$, i.e., all unobserved variables are roots and confounders, then the model is linear LV-SEM. Identification analysis for these two special cases have been studied in [19], yet different technique was used in that work for the linear LV-SEM. We then study the general form in Section 3.3 where both challenges can be present in the system simultaneously, which is our main identification result. In both subsections, we study identification under two faithfulness assumptions where, as will be discussed shortly, the first one, referred to as the conventional faithfulness, is a weaker assumption.

3.1 Separability and faithfulness assumptions

In the following we present two assumptions for the identifiability of SEM-ME and LV-SEM, namely *separability assumption* and *faithfulness assumption*. Note that for the identifiability of LV-SEM-ME, we need one extra assumption which we will present in Section 3.3.

Separability assumption. We first deduce the mixing matrix that transforms independent exogenous noise terms to the observed variables $[X;\ Y]$. Denote $V^C = [Z^C; Y]$ as the vector of cogent variables. We can write variables in $[Z^L; V^C]$ as linear combinations of the exogenous noise terms:

$$\begin{bmatrix} Z^L \\ V^C \end{bmatrix} = \mathbf{W}^* \begin{bmatrix} N_H \\ N_{V^C} \end{bmatrix}, \qquad \text{where} \quad \mathbf{W}^* = \begin{bmatrix} \mathbf{B}^L + \mathbf{D}(\mathbf{I}-\mathbf{C})^{-1}\mathbf{B}^C & \mathbf{D}(\mathbf{I}-\mathbf{C})^{-1} \\ (\mathbf{I}-\mathbf{C})^{-1}\mathbf{B}^C & (\mathbf{I}-\mathbf{C})^{-1} \end{bmatrix}. \tag{2}$$

$\mathbf{B}$ is partitioned into $[\mathbf{B}^L; \mathbf{B}^C]$ according to $[Z^L; V^C]$, $N_{V^C} = [N_{Z^C}; N_Y]$, and $\mathbf{I}$ represents the identity matrix. Combined with Equation (1c), the overall mixing matrix $\mathbf{W}$ can be written as

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{W}^* & \begin{matrix} \mathbf{I} \\ \mathbf{0} \end{matrix} \end{bmatrix}}_{\mathbf{W}} \begin{bmatrix} N_H \\ N_{V^C} \\ N_{X^L} \\ N_{X^C} \end{bmatrix}. \tag{3}$$

We note that because mleaf variables have no exogenous noise, each column in $\mathbf{W}^*$ either has at least two non-zero entries, or has one non-zero entry where the non-zero entry is not in X.

Assumption 1 (Separability) *The mixing matrix $\mathbf{W}$ in Equation (3) can be recovered from observations of $[X;\ Y]$ up to permutation and scaling of its columns.*

Separability assumption states that the independent exogenous noise terms pertaining to this mixture can be separated. An example of a setting where this assumption holds is when all exogenous noises are non-Gaussian. Further, given the recovered matrix $\mathbf{W}$, the matrix $\mathbf{W}^*$ can be recovered by removing the one-hot column vectors in $\mathbf{W}$ where the non-zero entry is in a row corresponding to a variable in $\mathcal{X}$. Please refer to Appendix C.1 for more details.

Faithfulness assumption. For each variable V_i, define possible parent set of V_i, $PP(V_i)$, as the union of $An(V_i) \setminus \mathcal{H}$ and the set of mleaf variables whose parent sets are subsets of $An(V_i)$. We provide two versions of faithfulness assumption that will be used in our identification results.

Assumption 2 (Conventional faithfulness) *The total causal effect of any variable $V_i \in \mathcal{V}$ on its descendant $V_j \in \mathcal{V}$ is not zero.*

Assumption 3 (LV-SEM-ME faithfulness) *For each variable $V_i \in \mathcal{Z} \cup \mathcal{Y}$ and any pair of subsets $(J, K) \subseteq (\mathcal{H} \cup \mathcal{V}^C) \times \mathcal{Z} \cup \mathcal{Y}$ that satisfies at least one of the conditions below, the rank of the submatrix $\mathbf{W}_{K \cup \{V_i\}, J}$ is equal to the size of the smallest set that blocks all directed paths from J to $K \cup \{V_i\}$:*

- *(a) $J \subseteq An(V_i)$, $K \subseteq PP(V_i)$;*
- *(b) $J \subseteq An(V_i) \setminus \{V_j\}$, $K \subseteq PP(V_j)$, when V_i is a mleaf variable and V_j is a parent of V_i.*

Assumption 2 is widely assumed in the literature, and hence we refer to it as the conventional faithfulness. It requires that when multiple causal paths exist from any (observed or unobserved) variable to its descendants, their combined effect (i.e., sum of products of path coefficients) is not equal to zero. Assumption 3 provides a stronger notion of faithfulness. The intuition of Assumption 3 is as follows. The structure of the causal diagram in the data generating process implies proportionality in the corresponding entries in the mixing matrix. For example, in the structure $V_1 \to V_2 \to V_3$, the mixing matrix with rows corresponding to $\{V_2, V_3\}$ and columns corresponding to $\{V_1, V_2\}$ is of rank 1. However, there may exists extra proportionality among the entries in the mixing matrix that are not enforced by the graph. This extra proportionality may result in the data distribution corresponding to an alternative model that does not always happen. Note that both assumptions are violated with probability zero. Please refer to Appendix C.2 for more discussion about both faithfulness assumptions.

3.2 Identifiability of SEM-ME and LV-SEM

In this subsection, we summarize the identification results of SEM-ME and LV-SEM under faithfulness and separability assumptions, where the extent of identifiability can be described by two graphical characterizations of equivalence, namely Ancestral ordered grouping (AOG) equivalence and Direct ordered grouping (DOG) equivalence. See Appendix D for more detailed analysis.

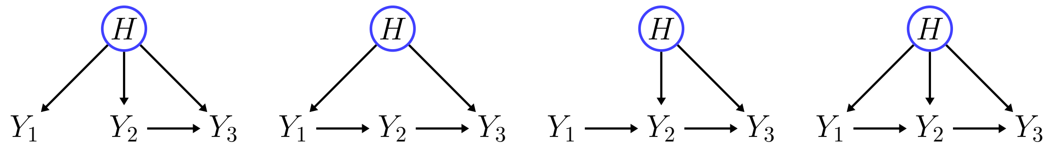

Figure 2: Considered models in simulation: NCO, Front-Door, IV, and Union Model.

Definition 3 (Ancestral/direct ordered grouping) *The AOG/DOG of a linear SEM-ME (resp. LV-SEM) is a partition of $\mathcal{Z}^L \cup \mathcal{V}^C$ (resp. $\mathcal{H} \cup \mathcal{V}^C$) into distinct sets, described as follows:*

(1) Assign each cogent variable in $\mathcal{V}^C$ to a distinct group.

*(2) (i) **SEM-ME**: For each mleaf variable $Z_j \in \mathcal{Z}^L$, assign Z_j to the same ordered group as Z_i if certain graphical condition is satisfied. Otherwise, assign Z_j to a separate group.*

*(ii) **LV-SEM**: For each unobserved variable $H_j \in \mathcal{H}$, assign H_j to the same ordered group as V_i if certain graphical condition is satisfied. Otherwise, assign H_j to a separate group.*

Definition 4 (AOG/DOG equivalence class) *The AOG/DOG equivalence class of a linear SEM-ME (resp. LV-SEM) is a set of models where the elements of this set all have the same mixing matrix (up to permutation and scaling) and same ancestral/direct ordered groups.*

The definitions of AOG and DOG only differs in the graphical condition. It was shown in [19] that all models in the AOG equivalence class are consistent with the same set of causal orders among the groups (i.e., if a causal order on the groups is consistent with one model in the class, it is consistent with all the models in the class), but not necessarily all the same edges across the groups.[2] That is, based on AOG, the set of causal orders among groups are identifiable but not the edges across groups.

Proposition 1 *(a) Models in the same DOG equivalence class of a SEM-ME have the same unlabeled graph structure, i.e., the causal diagrams of these models are isomorphic.*

(b) Models in the same DOG equivalence class of an LV-SEM have the same graph structure.

For a SEM-ME, according to Definition 3, there is at most one cogent variable in each ancestral ordered group. Furthermore, each observed cogent variable belongs to a separate group. Each mleaf node Z_j is assigned to the ancestral ordered group of at most one of its parents. Hence, if a group has more than one variable, then there must be exactly one measured cogent variable, and the rest of the nodes are mleaf nodes which are children of this node. This concludes that the induced structure on each ancestral ordered group is a star graph. Similar property holds for LV-SEM. Define the center of the ancestral ordered group as the cogent variable, or the mleaf variable (resp. unobserved variable) if the group does not include a cogent variable. [19] showed that fixing the center of the ancestral ordered groups for SEM-ME, and fixing the exogenous noise term of the center of the ancestral ordered groups as well as the choice of scaling and permutation of the columns of $\mathbf{B}$ for LV-SEM, leads to unique identification of the models. Therefore, by considering all the candidates for the center, models in the same AOG equivalence class of a SEM-ME can be enumerated by switching the center of each group with other nodes that are in the same group. Models in the same AOG equivalence class of an LV-SEM can be enumerated by switching the exogenous noise of the center of each group with the noise of other nodes in the same group.

As for DOG, we can see from graphical condition in Definition 8 that DOG is a more refined partition compared with AOG. Therefore, similar to AOG equivalence class, models in the DOG equivalence class are also consistent with the same set of causal orders among the groups, and the induced structure on each direct ordered group is also a star graph. Further, models in the same DOG equivalence class of a SEM-ME and an LV-SEM can also be enumerated by switching the center of each group, and switching the exogenous noise of the center of each group with the noise of other nodes in the same group, respectively. For the properties that only hold for DOG, it was shown in [19] that models in the same DOG equivalence class have the same edges across the groups. Combined with the star structure within each group, the following proposition provides a graphical characterization of the DOG equivalence class for SEM-ME and LV-SEM.

[2]The identification results in Theorem 1 implies that AOG is the finest partitioning that satisfy this property under Assumption 2, and DOG is the finest partitioning under Assumption 3.

Proposition 2 *(a) Models in the same DOG equivalence class of a SEM-ME have the same unlabeled graph structure, i.e., the causal diagrams of these models are isomorphic.*

(b) Models in the same DOG equivalence class of an LV-SEM have the same graph structure.

We provide the identification results for SEM-ME and LV-SEM below. Please refer to Appendix D for detailed discussion about these results.

Theorem 1 *We have the following results regarding the identification of SEM-ME and LV-SEM:*

(a) Under Assumptions 1 and 2, the linear SEM-ME (resp. LV-SEM) can be identified up to its AOG equivalence class.

(b) Under Assumptions 1 and 3, the linear SEM-ME (resp. LV-SEM) can be identified up to its DOG equivalence class with probability one.

3.3 Identifiability of LV-SEM-ME

In this subsection, we provide identification results for a system in which both unobserved confounders and measurement errors can co-exist. For this case, we need an extra *minimality assumption*.

Assumption 4 (Minimality) *We assume the linear LV-SEM-ME M is* minimal*, that is, there does not exist any other linear LV-SEM-ME M' such that M' has strictly fewer unobserved variables than M, the same observability indicators of the variables, and the same mixing matrix as M up to permutation and scaling of the columns.*

Minimality assumption asserts that the ground-truth model has fewer (or equal) unobserved variables than any other models that has the same mixing matrix and observability indicator. This assumption is required since we cannot infer the number of unobserved variables without prior knowledge of the system. See Appendix C.3 for more discussion and an equivalent graphical characterization. Equipped with minimality assumption, we are ready to present our main identification result.

Definition 5 (AOG and DOG of LV-SEM-ME) *The DOG (resp. AOG) of an LV-SEM-ME consists of a partition of the variables in $\mathcal{H} \cup \mathcal{Z}^L \cup \mathcal{V}^C$ described as follows:*

(1) Assign each cogent variable $V_i \in \mathcal{V}^C$ to a distinct group.

(2) Assign the mleaf variables in $\mathcal{Z}^L$ to the groups of measured cogent variables or a separate group following the same condition for DOG (resp. AOG) in SEM-ME (i.e., (2)(i) in Def. 3).

(3) Assign the unobserved variables in $\mathcal{H}$ to the groups of cogent variables or a separate group following the same condition for DOG (resp. AOG) in LV-SEM (i.e., (2)(ii) in Def. 3).

Using Definition 5, the AOG and DOG equivalence classes of an LV-SEM-ME are defined as in Definition 4. Similar to Section 3.2, models in the same AOG and DOG equivalence classes are consistent with the different sets of causal orders among the groups, and models in the same DOG equivalence class have isomorphic graph structure, summarized below.

Proposition 3 *Models in the same DOG equivalence class of an LV-SEM-ME have the same unlabeled graph structure, i.e., the causal diagrams of these models are isomorphic.*

However, unlike the results in Section 3.2, the induced structure on each group may not be a star graph if there is an edge from the unobserved variables to the mleaf variables in the same group. We extend our approach by defining the center of a group as the cogent variable in that group (if it exists), or the only mleaf or unobserved variable in the group. In this case, all members in the equivalence classes can be enumerated by either switching the center with any mleaf variables in the group, and/or switching the exogenous noises of the center with the noises of any unobserved variables in the group. For example, a group with one measured cogent variable, one mleaf variable and one unobserved variable has three equivalents (switching the cogent with the mleaf, switching the noise of the cogent with the noise of the unobserved confounder, and both). Below, we show that the introduced notion of equivalnece, is indeed the extent of identification in the generalized LV-SEM-ME.

Theorem 2 *We have the following results regarding the identification in LV-SEM-ME:*

Table 1: Performance of the algorithms on estimating the causal effect in four different models. Mean, 20%, 80% stand for the average, 20% and 80% percentiles of the estimation error, respectively.

	DOGEC (ours)			**GRICA** [16]			**lvLiNGAM**[11]			**Cross-Moment**[7]		
	Mean	20%	80%	Mean	20%	80%	Mean	20%	80%	Mean	20%	80%
NCO	**0.08**	**0.06**	**0.11**	0.38	0.30	0.43	1.15	0.88	1.49	**0.08**	**0.04**	**0.10**
Front	**0.02**	**0.01**	**0.02**	0.46	0.35	0.57	0.84	0.60	0.93	0.23	0.03	0.23
IV	**0.19**	**0.02**	**0.36**	0.50	0.37	0.59	0.84	0.69	0.99	1.33	0.49	1.21
Union	**0.11**	**0.08**	**0.14**	0.39	0.30	0.42	0.73	0.28	0.89	0.30	0.23	0.38

(a) *Under Assumptions 1, 2 and 4, the linear LV-SEM-ME can be identified up to its AOG equivalence class.*

(b) *Under Assumptions 1, 3 and 4, the linear LV-SEM-ME can be identified up to its DOG equivalence class with probability one.*

4 Recovery algorithm and numerical experiments

We first describe the recovery algorithm for the introduced LV-SEM-ME model. The algorithm includes two main parts. The first part, namely AOG recovery algorithm (Algorithm 1), returns the ancestral ordered grouping of the underlying model given mixing matrix and observability indicator. The second part is to return all members in the AOG and DOG equivalence class (Algorithm 2). As described above, given the recovered AOG, members in the AOG equivalence class can be enumerated by switching the center with any mleaf variables in the group and/or switching the exogenous noises of the center with the noises of any unobserved variables in the group. For recovering the members in the DOG equivalence class, we show that members in the AOG equivalence class of the ground-truth but not DOG equivalence class has strictly more edges than the ground truth (Proposition 7). Therefore, members in the DOG equivalence class can be recovered by finding the models in the AOG equivalence class that has the fewest number of edges. See Appendix E for more details.

We evaluated the performance of our algorithm on synthetic data. We consider the following four models, all with three observed variables and one unobserved variable: (1) Negative Control Outcome (NCO) model; (2) Front-door model; (3) Instrumental Variable (IV) model; (4) A union model that can be considered as a generalization of all three models. See Figure 2 for causal diagrams of four models. Our theoretical results imply that all four models are uniquely identifiable, i.e., the DOG equivalence class only includes the ground truth. The edge weights are randomly generated from $[0.5, 0.9]$, and the noises are sampled from uniform distribution. The task is to estimate the edge weight from Y_2 to Y_3. We use Reconstruction ICA (RICA) [9] to recover the mixing matrix.

We compare our recovery algorithm with lvLiNGAM algorithm [11], Graphical RICA (GRICA) algorithm [16] and cross-moment approach [7] on 10 randomly generated models. We note that both GRICA and cross-moment assume that the causal graph is known. However, this is not given as input in our algorithm. We calculate the relative error between the true value and the estmated value, and report the average error, as well as the 20% and 80% percentile of the errors in Table 1. The results show that our method significantly outperforms the baselines even without knowing the true causal graph, which aligns with our theoretical results.

5 Conclusion

We studied causal discovery in the presence of unobserved variables and measurement error by defining the linear LV-SEM-ME model. We characterized its identifiability under separability and two faithfulness assumptions, and showed that it can be identified up to AOG and DOG equivalence classes respectively. We provided graphical characterization of the models that are equivalent and presented a recovery algorithm that could return models equivalent to the ground truth. A remaining challenge in our proposed methodology is its reliance on the accuracy of the provided estimate of the mixing matrix of the underlying model. Hence, devising accurate approaches for estimating the mixing matrix is an important direction of future research.

References

[1] Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Merle Behr, Chris Holmes, Axel Munk, et al. Multiscale blind source separation. *The Annals of Statistics*, 46(2):711–744, 2018.

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

[4] Jan Eriksson and Visa Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE signal processing letters*, 11(7):601–604, 2004.

[5] Yoni Halpern, Steven Horng, and David Sontag. Anchored discrete factor analysis. *arXiv preprint arXiv:1511.03299*, 2015.

[6] Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[7] Yaroslav Kivva, Saber Salehkaleybar, and Negar Kiyavash. A cross-moment approach for causal effect estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1655–1664, 2016.

[9] Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in neural information processing systems*, pages 1017–1025, 2011.

[10] Basil Saeed, Anastasiya Belyaeva, Yuhao Wang, and Caroline Uhler. Anchored causal inference in the presence of measurement error. In *Conference on uncertainty in artificial intelligence*, pages 619–628. PMLR, 2020.

[11] Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020.

[12] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[13] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

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

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

[16] Daniele Tramontano, Yaroslav Kivva, Saber Salehkaleybar, Mathias Drton, and Negar Kiyavash. Causal effect identification in lingam models with latent confounders. In *Forty-first International Conference on Machine Learning*, 2024.

[17] F. Xie, R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang. Generalized independent noise condition for estimating linear non-gaussian latent variable causal graphs. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[18] Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-Gaussian latent hierarchical structure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24370–24387. PMLR, 17–23 Jul 2022.

[19] Yuqin Yang, AmirEmad Ghassami, Mohamed Nafea, Negar Kiyavash, Kun Zhang, and Ilya Shpitser. Causal discovery in linear latent variable models subject to measurement error. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 874–886. Curran Associates, Inc., 2022.

[20] Kun Zhang, Mingming Gong, Joseph D Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. Causal discovery with linear non-gaussian models under measurement error: Structural identifiability results. In *UAI*, pages 1063–1072, 2018.

A Related work

The two aforementioned challenges have been addressed in several work in the literature. Regarding the challenge of unobserved common causes, constraint-based algorithms such as the Fast Causal Inference (FCI) algorithm has been widely used [15]. However, these approaches are often unable to identify the direction of the majority of causal connections in the recovered graph. In the presence of the unobserved common causes, methods based on linearity and non-Gausianity also in general are not able to learn the structure uniquely. Proposed methods such as latent variable LiNGAM [6, 11] and partially observed LiNGAM [1] provide graphical conditions for unique identifiability of the model. These conditions are usually non-trivial, which demonstrates the difficulty of the task of causal discovery in the presence of unobserved common causes.

There are fewer work focusing on the challenge of causal discovery in the presence of measurement error. Majority of the work [14, 8, 17, 18] assume that each unobserved (but measured) variable has at least two measurements. This assumption facilitates the task of discovery and usually leads to unique identifiability of the model, yet may not hold in many settings.[3] Without this assumption, [5] consider the model where the measured variables are binary, and [10] consider the case where the relations among measured variables are nonlinear, and the measurement errors are Gaussian. In linear models, [20] provide conditions for unique identifiability of linear Gaussian and non-Gaussian models. None of the aforementioned work on measurement error considered the case when the model is not uniquely identifiable (i.e., there are other models that are observationally equivalent to the true model). Hence, characterization of observational equivalence for measurement error models is missing in the literature.

B Detailed description of LV-SEM-ME model

B.1 Additional restriction on canonical LV-SEM-ME model

We add the following two restrictions on the model for the sake of model identifiability based on observational data.

- Firstly, as discussed in [20, 19], for any mleaf variable Z_i, the exogenous noise term N_{Z_i} is not distinguishable from its measurement error N_{X_i}. Specifically, for any two models that only differ in Z_i and X_i for some mleaf variable Z_i but have the same sum $N_{Z_i} + N_{X_i}$, they have the same observational distribution. This follows because Z_i is not observed, and N_{Z_i} only influences its noisy measurement X_i and no other observed variables in $\mathcal{Y}$. Therefore, we consider the model where $N_{Z_i} = 0$ for all mleaf variables for model identifiability, i.e., we assume that mleaf variables are deterministically generated from its direct parents.
- Secondly, we assume that variables in $\mathcal{H}$ are all root variables (i.e., have no direct parents) and confounders (i.e., have at least two children). This is because for any linear latent variable model with a non-root latent variable, there exists an equivalent latent variable

[3] As seen in Section 2, we avoid positing this assumption in our work.

model where the latent variables are all roots, such that both models have the same joint distribution across observed variables and total causal effect between any pair of observed variables [6].

Due to the aforementioned restrictions, we focus on recovering the subset of linear LV-SEM-ME, called *canonical LV-SEM-ME*, defined as in Definition 2.

B.2 Example of a linear LV-SEM-ME

Example 1 *Figure 1 shows an example of a causal diagram including unobserved variable H, observed variable $\{Y_3\}$, measured variables $\{Z_1, Z_2\}$ and their corresponding measurements $\{X_1, X_2\}$. The generating model is as follows.*

$$\begin{aligned} H &= N_H, \\ Z_1 &= N_{Z_1}, \\ Z_2 &= b_2 H + a_{21} Z_1 + N_{Z_2}, \\ Y_3 &= b_3 H + N_{Z_3}, \\ X_1 &= Z_1 + N_{X_1}, \\ X_2 &= Z_2 + N_{X_2}. \end{aligned}$$

We note that Z_2 is an mleaf variable; it has no other children except for X_2. $\{Z_1, Y_3\}$ are cogent variables. Therefore, in the canonical form, the exogenous noise of Z_2 is 0, and the measurement error of Z_2 is $N_{Z_2} + N_{X_2}$. In the matrix form of Equation (1b), we have $Z^L = [Z_2]$, $Z^C = [Z_1]$, $Y = [Y_3]$, $\mathbf{B} = [b_2, 0, b_3]^\top$, $\mathbf{D} = [a_{21}, 0]$, $\mathbf{C}_Z = \mathbf{C}_Y = [0, 0]$.

The mixing matrix $\mathbf{W}$ in this example can be written as

$$\begin{bmatrix} X_2 \\ X_1 \\ Y_3 \end{bmatrix} = \begin{bmatrix} b_2 & a_{21} & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ b_3 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} N_H \\ N_{Z_1} \\ N_{Y_3} \\ N_{X_2} \\ N_{X_1} \end{bmatrix}.$$

The leftmost three columns correspond to $\mathbf{W}^*$, and $\mathbf{I}$ is of dimension 2×2.

C Detailed discussion on the identification assumptions

C.1 Separability assumption

Separability assumption states that the independent exogenous noise terms pertaining to the mixture in Equation (3) can be separated, i.e., the mixing matrix can be recovered up to permutation and scaling of its columns. An example of a setting where this assumption holds is when all exogenous noises are non-Gaussian. In this case, if the model satisfies the requirement in [4, Theorem 1], then overcomplete Independent Component Analysis (ICA) can be used to recover the mixing matrix up to permutation and scaling of its columns. Another example where separability assumption is satisfied is the setup in which the noise terms are piecewise constant functionals satisfying a set of mild conditions [2]. On the other hand, an example where this assumption is violated is when all exogenous noise terms have Gaussian distributions. In this case, the mixing matrix can only be recovered up to an orthogonal transformation.

We note that under separability assumption, the matrix $\mathbf{W}^*$ can also be recovered up to permutation and scaling of its columns. Suppose the recovered mixing matrix is $\hat{\mathbf{W}}$, and we also have the information whether each variable is observed (i.e., belongs to $\mathcal{Y}$) or measured with error (i.e., belongs to $\mathcal{X}$). Then the matrix $\mathbf{W}^*$ can be recovered by removing the one-hot column vectors in $\hat{\mathbf{W}}$ where the non-zero entry is in a row corresponding to a variable in $\mathcal{X}$.

The justification of the above approach is as follows: If a variable X_i is measured with error, then there must exist one column in $\hat{\mathbf{W}}$ that corresponds to the measurement error N_{X_i} (or $N_{X_i} + N_{Z_i}$ for mleaf variables in the original form). This column has only one non-zero entry in the row corresponding to the measurement X_i. The columns that we remove in the procedure above correspond to such measurement errors. We note that by removing these columns from $\hat{\mathbf{W}}$ we are not losing any information. This is due to the fact that we can simply recover a matrix (permutationally) equivalent to $\hat{\mathbf{W}}$ from $\mathbf{W}^*$: Since we know which variables are measured with error, we can simply add the corresponding one-hot column vectors back to the matrix. The order of adding these is irrelevant (no information) since it is arbitrary in the original $\hat{\mathbf{W}}$ as well (recall that $\mathbf{W}$ is identifiable up to permutation and scaling of the columns).

C.2 Faithfulness assumption

For each variable $V_i \in \mathcal{V}$, define the ancestor set, $An(V_i)$ as the set of variables that has a direct path to V_i (excluding V_i). Define the cogent ancestor set $An_V(V_i)$ as $An(V_i) \setminus \mathcal{H}$. Define the possible parent set of V_i, $PP(V_i)$, as the union of $An_V(V_i)$ and the set of mleaf variables whose parent sets are subsets of $An(V_i)$ (excluding V_i itself if V_i is an mleaf). For two sets of variables $J \subseteq \mathcal{H} \cup \mathcal{V}^C$ and $K \subseteq \mathcal{Z} \cup \mathcal{Y}$, define the matrix $\mathbf{W}_K^J$ as the submatrix of $\mathbf{W}^*$ where the rows correspond to the variables in K, and the columns correspond to the exogenous noise terms of variables in $\mathcal{J}$. A set of variables $\mathcal{B}$ is said to be a bottleneck from J to K if for every directed path from a variable in J to a variable in K, at least one variable on this path (including the start and end nodes) is in $\mathcal{B}$. $\mathcal{B}$ is said to be a minimal bottleneck from J to K if no other bottlenecks from J to K have fewer number of variables.

Assumption 2 is widely assumed in the literature, and hence we refer to it as the conventional faithfulness. It requires that when multiple causal paths exist from any (observed or unobserved) variable to its descendants, their combined effect (i.e., sum of products of path coefficients) is not equal to zero. Note that Assumption 2 is a special case of Assumption 3(a) with $K = \emptyset$ and J being singleton set of any ancestor of V_i. The intuition of Assumption 3 is as follows. The structure of the causal diagram in the data generating process implies proportionality in the corresponding entries in the mixing matrix. For example, in the structure $V_1 \to V_2 \to V_3$, the mixing matrix with rows corresponding to $\{V_2, V_3\}$ and columns corresponding to $\{V_1, V_2\}$ is of rank 1. However, there may exists extra proportionality among the entries in the mixing matrix that are not enforced by the graph. This extra proportionality may result in the data distribution corresponding to an alternative model that does not always happen. Faithfulness assumption prevents such extra proportionality in the generating model. A similar bottleneck faithfulness has been proposed in [1], where they consider any pair of subset (J, K). Assumption 2 is strictly weaker than that assumption.

Remark 1 *Both Assumptions 2 and 3 are violated with probability zero if all model coefficients are drawn randomly and independently from continuous distributions. However, Assumption 2 is only focused on marginal independencies and requires that an ancestor of a variable should not be independent of it due to parameter cancellation. In practice, due to sample size limitations, an approximate cancellation may be perceived as an actual cancellation. Therefore, although Assumptions 2 and 3 both rule out measure-zero subsets of the model, and may be perceived equally weak assumptions mathematically, in practice, Assumption 2 may be preferred. Due to this reason, in the following, we provide results under Assumption 2 and Assumption 3 separately.*

C.3 Minimality assumption

Minimality assumption asserts that the ground-truth model has fewer (or equal) unobserved variables than any other models that has the same mixing matrix and observability indicator. This assumption is required since we cannot infer the number of unobserved variables without prior knowledge of the system. Recall from Equation (2) that the number of columns of $\mathbf{W}^*$ is the sum of the number of cogent variables and unobserved variables. However, the number of each type of variable is not known apriori under only separability assumption.

Minimality assumption is always required when unobserved variables are present in the system. Specifically, it is often assumed that the ground-truth model either has the fewest number of edges [1], or the fewest number of unobserved variables [11]. Our minimality condition falls into the latter case, and in Proposition 4 below, we show that the minimality assumption has a equivalent graphical characterization.

Proposition 4 (Minimality) *Under Assumption 2, a linear LV-SEM-ME is not minimal if and only if there exists an unobserved variable H_i and a mleaf child Z_j of H_i, such that for any other child*

V_k of H_i, $An(Z_j) \subseteq An(V_k)$[4]*. This is equivalent to the following. Any (observed or unobserved) parent of Z_j is also an ancestor of V_k.*

We note that there is a similarity between the condition in Proposition 4 and the condition in the definition of AOG in Definition 6. The reason for this similarity is that both intend to characterize the situation in which the location of an exogenous source is unknown. The former condition considers the case where this source belongs to a unobserved variable, while the latter considers the case where this source belongs to a measured cogent variable.

D Identifiability results of SEM-ME and LV-SEM

D.1 AOG for SEM-ME and LV-SEM

We present the formal definition of Ancestral ordered grouping (AOG) and AOG equivalence class below.

Definition 6 (Ancestral ordered grouping (AOG)) *The AOG of a SEM-ME (resp. LV-SEM) is a partition of $\mathcal{Z}^L \cup \mathcal{V}^C$ (resp. $\mathcal{H} \cup \mathcal{V}^C$) into distinct sets. This partition is described as follows:*

(1) Assign each cogent variable in $\mathcal{V}^C$ to a distinct group.

*(2) (i) **SEM-ME**: For each mleaf variable $Z_j \in \mathcal{Z}^L$, if it has one measured parent $Z_i \in \mathcal{Z} \cap \mathcal{V}^C$ such that Z_j has no other parents, or all other parents of Z_j are also ancestors of Z_i, assign Z_j to the same group as Z_i. Otherwise, assign Z_j to a separate group (with no cogent variable).*

*(ii) **LV-SEM**: For each unobserved variable $H_j \in \mathcal{H}$, if it has one cogent child $V_i \in \mathcal{V}^C$ such that all other children of H_j are also descendants of V_i, assign H_j to the same group as V_i. Otherwise, assign H_j to a separate group (with no cogent variable).*

Definition 7 (AOG equivalence class) *The AOG equivalence class of a linear SEM-ME (resp. LV-SEM) is a set of models where the elements of this set all have the same mixing matrix (up to permutation and scaling) and same ancestral ordered groups.*

D.2 Identification Under LV-SEM-ME Faithfulness

We first present a graphical condition that is used to characterize DOG equivalence class, and then formally show that this notion of equivalence is the extent of identifiability as in Theorem 1(b). Lastly, we add some discussion about the difference between the different extents of identifiable, and provide an example to show the difference.

Condition 1 (SEM-ME edge identifiability) *For a given edge from a measured cogent variable Z_i to an mleaf variable Z_l, at least one of the following two conditions is satisfied: (a) $Pa(Z_l) \setminus \{Z_i\}$ is not a subset of $Pa(Z_i)$. That is, there exists another parent V_j of Z_l, which is not a parent of Z_i. (b) $Pa(Z_l)$ is not a subset of $\cap_{V_k \in Ch(Z_i) \setminus \{Z_l\}} Pa(V_k)$. That is, there exists a child V_k of Z_i and a parent V_j of Z_l such that V_j is not a parent of V_k.*

Condition 2 (LV-SEM edge identifiability) *For a given edge from an unobserved variable H_l to a cogent variable V_i, there exists another cogent child V_j of H_l, such that at least one of the following two conditions is satisfied: (a) V_i is not a direct parent of V_j. (b) $Pa(V_i)$ is not a subset of $Pa(V_j)$. That is, there exists an observed (or unobserved) parent V_k (or H_k) of V_i that is not a parent of V_j.*

Definition 8 (Direct ordered grouping (DOG)) *The DOG of a linear SEM-ME (resp. LV-SEM) is a partition of $\mathcal{Z}^L \cup \mathcal{V}^C$ (resp. $\mathcal{H} \cup \mathcal{V}^C$) into distinct sets. This partition is described as follows:*

(1) Assign each cogent variable in $\mathcal{V}^C$ to a distinct group.

*(2) (i) **SEM-ME**: For each mleaf variable $Z_j \in \mathcal{Z}^L$, if it has one measured parent $Z_i \in \mathcal{Z} \cap \mathcal{V}^C$ such that the edge from Z_i to Z_j violates Condition 1, assign Z_j to the same ordered group as Z_i. Otherwise, assign Z_j to a separate ordered group (with no cogent variable).*

[4] Note that we do not include Z itself in $An(Z)$.

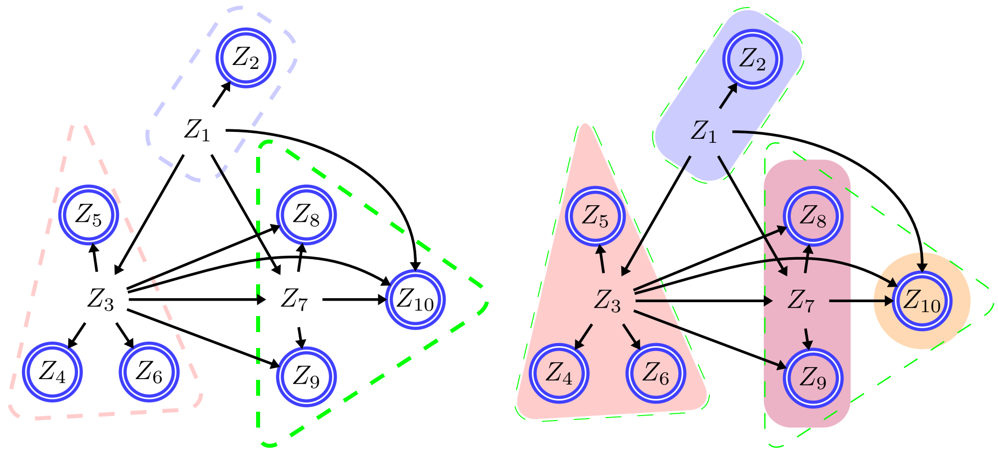

Figure 3: **Left**: Diagram of the SEM-ME and the corresponding AOG considered in Example 2. **Right**: DOG of the SEM-ME.

(ii) ***LV-SEM****: For each unobserved variable $H_j \in \mathcal{H}$, if it has one cogent child $V_i \in \mathcal{V}^C$ such that the edge from H_j to V_i violates Condition 2, assign H_j to the same ordered group as V_i. Otherwise, assign H_j to a separate ordered group (with no cogent variable).*

Definition 9 (DOG equivalence class) *The DOG equivalence class of a linear SEM-ME (resp. LV-SEM) is a set of models where the elements of this set all have the same mixing matrix (up to permutation and scaling) and same direct ordered groups.*

Theorem 3 *Under Assumptions 1 and 3, the linear SEM-ME (resp. LV-SEM) can be identified up to its DOG equivalence class with probability one.*

DOG equivalence class provides a much more distinctive characterization of the causal relations among the variables compared to the characterization given for AOG equivalence class. This is the benefit achieved by extending the faithfulness assumption from Assumption 2 to Assumption 3. We re-emphasize that, as explained in Remark 1, both these two assumptions are violated with probability zero, but Assumption 2 may be easier to be incorporated in practice. Therefore, in light of Theorem 1(b), we see the trade-off between the Assumptions 2 and 3.

Lastly, as shown in Theorem 1(b), for an LV-SEM, the only undetermined part in the DOG equivalence class pertains to the assignment of the exogenous noises and coefficients, but the structure is the same. Consequently, if only the identification of the structure without weights is of interest, Assumptions 1 and 3 are sufficient.

Remark 2 *The definition for AOG and DOG of a SEM-ME in this work is different from previous work [19]. In that work, mleaf variables are either assigned to the groups of their (observed or measured) parents or a separate group. In contrast, in this work, mleaf variables cannot be assigned to the groups of their observed parents. This change is based on using the information about which cogent variables are measured and which ones are directly observed. Specifically, models in the same equivalence class defined in [19] may have different labeling of $\mathcal{Y}$ and $\mathcal{Z}$ among variables, while in this work, models in the same equivalence class have the same mixing matrix, AOG/DOG and the same labeling of $\mathcal{Y}$ and $\mathcal{Z}$. Therefore, this change leads to smaller equivalence classes and hence more identification power.*

Example 2 *Figure 3 shows an example of a causal diagram of an LVSEM-ME with 10 measured variables. (Z_1, Z_3, Z_7) are cogent variables, and the remaining are mleaf variables.*

The AOG of the model is shown on the left, and the DOG of the model is shown on the right. We note that Z_{10} belongs to the same ancestral ordered group as Z_7 since all other parents of Z_{10} are also parents of Z_7. However, Z_{10} does not belong to the same direct ordered group as Z_7. This is because Z_8 is a child of Z_7, Z_1 is a parent of Z_{10}, but Z_1 is not a parent of Z_8. Therefore the edge $Z_7 \to Z_{10}$ violates the Condition 1(b).

E Detailed description of the algorithm

In this section, we present recovery algorithms for the introduced LV-SEM-ME model. We first present AOG recovery algorithm (Algorithm 1) in Section E.1. The algorithm returns the AOG of the underlying model, and it is used in both AOG equivalence class and DOG equivalence class recovery algorithms. We then show how to recover all models in the AOG and DOG equivalence classes in Section E.2 based on the recovered AOG (Algorithm 2).

Both Algorithm 1 and 2 are based on the mixing matrix $\mathbf{W}^*$ defined in Equation (2). Therefore, $\mathbf{W}^*$ will be the input to both algorithms. We note that $\mathbf{W}^*$ can be recovered from the observational data by first recovering the overall mixing matrix $\mathbf{W}$ (cf. (3)) using methods such as overcomplete ICA [4] when the exogenous noises are assumed to be non-Gaussian. Then, $\mathbf{W}^*$ can be deduced from $\mathbf{W}$ by removing certain columns as described in Section 3.1.

E.1 AOG Recovery

The following property can be implied from the definition of AOG, which shows that under conventional faithfulness assumption, we can identify the AOG of the model only based on the support of the mixing matrix $\mathbf{W}^*$.

Proposition 5 *Under Assumptions 2 and 4,*

(a) One mleaf variable and one measured cogent variable belong to the same ancestral ordered group if and only if the two rows in $\mathbf{W}^*$ corresponding to these variables have the same support. Further, for any cogent variable V_i and its descendant V_j, the row support of V_i must be a subset of the row support of V_j.

(b) One unobserved variable and one cogent variable belong to the same ancestral ordered group if and only if the two columns in $\mathbf{W}^*$ corresponding to the exogenous noise terms of these variables have the same support. Further, for any cogent variable V_i and its ancestor V_j, the column support of N_{V_i} must be a subset of the column support of N_{V_j}.

The proof of Proposition 5 directly follow from the definition of AOG, and hence are omitted.

Equipped with Proposition 5, we propose an iterative algorithm for recovering the AOG in Definition 5 from $\mathbf{W}^*$. The pseudo-code of the proposed method is presented in Algorithm 1. In the first iteration, the Algorithm randomly chooses a row in $\mathbf{W}^*$ with the fewest number of non-zero entries and finds all other rows with the same support. Denote the selected rows as $\mathcal{Z}_{\mathcal{J}}$, and columns corresponding to these non-zero entries as $\mathcal{N}_I$. Each of the selected row may either correspond to a cogent variable or an mleaf variable, and each of the column in $\mathcal{N}_I$ may either correspond to the exogenous noise of an cogent variable or an unobserved confounder. The task is to decide whether there exists a cogent variable (and its associated exogenous noise). We first select the columns in $\mathcal{N}_I$ with the fewest number of non-zero entries in $\mathbf{W}^*$. Denote this subset as $\mathcal{N}_J$. Then noises in $\mathcal{N}_I \setminus \mathcal{N}_J$ must correspond to unobserved variables and are assigned to separate groups in $\mathcal{C}_{unobserved}$.

We check whether any of the rows in $\mathcal{Z}_{\mathcal{J}}$ can be a cogent variable. If there is one observed variable in $\mathcal{Z}_J$ then it must correspond to the cogent variable. If all variables are unobserved, then we look at the submatrix $\mathbf{W}_0$ of $\mathbf{W}^*$ with the rows corresponding to the (column) support of any variable in $\mathcal{N}_{\mathcal{J}}$, and columns corresponding to the (row) support of $\mathcal{Z}_{\mathcal{J}}$. If $\mathcal{Z}_{\mathcal{J}}$ includes a cogent variable, then its corresponding exogenous noise must be in $\mathcal{N}_{\mathcal{J}}$, and the remaining noises are unobserved confounders. Since all noises in $\mathcal{N}_{\mathcal{J}}$ have the same number of non-zero entries, they must have the same support. Further, any row that includes this exogenous noise must be a descendant of the cogent variable, and must include all the non-zero columns of the cogent variable under Assumption 2. This implies that $\mathbf{W}_0$ does not include any non zero entry. Therefore, if $\mathbf{W}_0$ includes any zero entry, then none of the rows in $\mathcal{Z}_{\mathcal{J}}$ would correspond to a cogent variable. The rows in $\mathcal{Z}_{\mathcal{J}}$ and the columns in $\mathcal{N}_{\mathcal{J}}$ all belong to separate groups in $\mathcal{C}_{mleaf}$ and $\mathcal{C}_{unobserved}$ as they correspond to mleaf variables and unobserved variables. If $\mathbf{W}_0$ does not include any zero entry, then under minimality assumption, one of the rows in $\mathcal{Z}_{\mathcal{J}}$ correspond to a cogent variable. Therefore all noises in $\mathcal{N}_J$ and all rows in $\mathcal{Z}_{\mathcal{J}}$ belong to a single ancestral ordered group in $\mathcal{C}_{cogent}$. The algorithm then removes the rows in $\mathcal{N}_J$ and columns in $\mathcal{N}_I$. Denote the remaining matrix as $\tilde{\mathbf{W}}$.

In the second iteration, the Algorithm again chooses the rows $\mathbf{w}$ with the fewest number of non-zero entry in $\tilde{\mathbf{W}}$. However, we note that the row with the fewest number of non-zero entry in $\mathbf{W}^*$ is not a cogent variable if some non-zero entries are removed in the first iteration. Therefore, among these columns, we select one column with the fewest number of non-zero entries in the large matrix $\mathbf{W}^*$. Then, it selects all other rows in $\tilde{\mathbf{W}}$ with the same support as $\mathbf{w}$. Denote the selected rows as $\mathcal{Z}_{\mathcal{I}}$. Rows in $\mathcal{Z}_{\mathcal{I}}$ that has more non-zero entries than $\mathbf{w}$ in the large matrix $\mathbf{W}^*$ must be mleaf variables (otherwise Assumption 2) is violated) and are assigned to separate groups in $\mathcal{C}_{mleaf}$. Rows that has the same number of non-zero entries as $\mathbf{w}$ may either correspond to a cogent variable or an mleaf variable, and we can use the same procedure as in the first iteration to distinguish between them. Finally, this procedure is repeated until all variables and noises are assigned to ordered groups.

The explanation above implies the identifiability result of Algorithm 1, summarized as follows.

Proposition 6 *Under Assumptions 2 and 4, the concatenation of the outputs $(\mathcal{C}_{unobserved}, \mathcal{C}_{cogent}, \mathcal{C}_{mleaf})$ of Algorithm 1 is the AOG of the LV-SEM-ME corresponding to the input $\mathbf{W}^*$ in a causal order consistent with the AOG equivalence class of the LV-SEM-ME.*

Computational complexity Algorithm 1 includes p_c steps of iteration, where p_c is the number of cogent variables. Each iteration requires $O(mn)$ calculation and needs $O(mn)$ space, where m, n are the dimensions of the mixing matrix $\mathbf{W}^*$. Recall that $m = p_c + p_{ml}$, and $n = p_c + p_H$, where p_{ml} and p_H stand for the number of mleaf and unobserved variables. Therefore, the total time complexity of the algorithm is $O(p_c mn)$, and the space complexity is $O(mn)$.

E.2 Model Recovery in the AOG and DOG Equivalence Class

In this section, we present our algorithm for recovering the models in the equivalence class using $\mathbf{W}^*$. The psuedo-code of the algorithm is presented in Algorithm 2.

AOG Equivalence class. Recall from Section 3.3 that all members in the AOG equivalence class can be enumerated by switching the center with any mleaf variables in the group and/or switching the exogenous noises of the center with the noises of any unobserved variables in the group. Therefore, given the mixing matrix $\mathbf{W}^*$, Algorithm 2 first recovers the AOG of the true model using Algorithm 1. Then, it enumerates all possible choices of centers and the noises for each group. Note that groups with only mleaf variables or noises of unonserbed variables (i.e., in $\mathcal{C}_{mleaf}$ or $\mathcal{C}_{unobserved}$) only have one choice. Therefore, we only need to consider all the groups that include cogent variables (i.e., in $\mathcal{C}_{cogent}$). Denote each single selection of the centers in these groups as row, and the noises as col. The next step is to recover the model parameters $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ based on $\mathbf{W}^*$ and the selected row and col following Equation (2). Denote the variables not in row as row^C, and the noises not in col as col^C. The selected centers in row correspond to V^C, and the selected noises in col correspond to N_{V^C} in (2). Similarly, variables in row^C and noises in col^C correspond to Z^L and N_H, respectively. Therefore $\mathbf{C}$, $\mathbf{B}^C$, $\mathbf{D}$, $\mathbf{B}^L$ can be calculated following lines 6-9 in Algorithm 2. Finding model parameters $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ for all possible choices of row and col gives us all models in the AOG equivalence class.

DOG Equivalence class. Proposition 7 can be leveraged to recover models in the DOG equivalence class of the ground-truth given AOG, which states that the ground-truth model has strictly fewer edges than any model in the AOG equivalence class but does not belong to the DOG equivalence class.

Proposition 7 *Suppose an LV-SEM-ME satisfies Assumptions 1 and 3. Any model that belongs to the same AOG equivalence class but does not belong to the same DOG equivalence class has strictly more edges than any member in the DOG equivalence class.*

Recall from Section 3.3 that models in the same DOG equivalence class all have the same unlabeled graph structure, hence the same number of edges. Therefore, using Proposition 7 given the members of the AOG equivalence class of the true model, members in the DOG equivalence class can be found by finding all models in the AOG equivalence class that have the fewest number of edges in the recovery output.

To conclude, the identifiability of Algorithm 2 can be summarized by the following proposition.

Algorithm 1: AOG Recovery algorithm for linear LV-SEM-ME.

Input: Recovered mixing matrix $\mathbf{W}^*$, observability indicators of the variables (whether each variable is observed or measured).

```
Define n as the vector with the i-th entry as the number of non-zero entries in row i of W*.
Define m as the vector with the j-th entry as the number of non-zero entries in column j of W*.
Initialize W̃ ← W*, C_cogent ← [ ], C_mleaf ← [ ], C_unobserved ← [ ].
while W̃ is not empty do
    Find the rows in W̃ that contain the fewest number of non-zero entries. Among them, choose
      one with the lowest corresponding value in n (break ties at random). Denote the selected
      row by w, the variable corresponding to this row by Z_w, and its corresponding value in n
      by n_0.
    Consider the rows in W̃ with the same support (non-zero entries) as w, including w itself.
      Denote the set of variables that correspond to these rows as Z_I, and the the noise terms
      corresponding to the support of w as N_I.
    Denote the set of variables in Z_I with corresponding value n_0 in n as Z_J, and the set of
      noise terms in N_I with the smallest corresponding value in m as N_J.
    Assign each variable in Z_I \ Z_J to a separate group in C_mleaf. Assign each exogenous noise
      term in N_I \ N_J to a separate group in C_unobserved.
    Randomly select one noise term N_m in N_J. Consider the submatrix W_0 of W* where the
      rows correspond to the (column) support of N_m, and the columns correspond to the (row)
      support of Z_w.
    if W_0 includes any zero entry then
        Assign each variable in Z_J to a separate group in C_mleaf. Assign each exogenous noise
          term in N_J to a separate group in C_unobserved.
    else if Z_J includes any observed variable then
        Assign the observed variable in Z_J and all exogenous noise terms in N_J to a single
          group in C_cogent. Assign each remaining variable in Z_J to a separate group in C_mleaf.
    else
        Assign all variables in Z_J and all exogenous noise terms in N_J to a single group in
          C_cogent.
    Remove from W̃ the rows corresponding to the variables in Z_I, and the columns
      corresponding to the noise terms in N_I. Remove the corresponding entries in m and n.
```

Output: $\mathcal{C}_{unobserved}$, $\mathcal{C}_{cogent}$, $\mathcal{C}_{mleaf}$.

Proposition 8 *(a) Under Assumptions 2 and 4, Algorithm 2 outputs $\mathcal{M}_{AOG}$ which is the AOG equivalence class of the LV-SEM-ME corresponding to the input $\mathbf{W}^*$.*

(b) Under Assumptions 3 and 4, Algorithm 2 outputs both $\mathcal{M}_{AOG}$ and $\mathcal{M}_{DOG}$, where $\mathcal{M}_{DOG}$ is the DOG equivalence class of the LV-SEM-ME corresponding to the input $\mathbf{W}^*$.

Computational complexity Recovering AOG of the model requires $O(p_c mn)$ time complexity and $O(mn)$ space complexity as discussed above. Recovering the underlying model given each choice of row and col requires $O(p_c mn)$ calculation and $O(mn)$ space. Calculating the number of edges in each single model requires $O(mn)$ calculation and $O(1)$ space. Denote the total number of choices of the centers and noises (i.e., size of Row and Col) by M_{Row} and M_{Col}. Note that M_{Row} and M_{Col} are bounded by $|D|^{p_c}$, where $|D|$ is the maximum group size. Therefore, total time complexity for recovering $\mathcal{M}_{AOG}$ and $\mathcal{M}_{DOG}$ is $O(M_{Row}M_{Col}p_c mn)$, and the space complexity is $O(mn)$.

F Proofs

F.1 Proof of Proposition 4

The proof includes two parts. We first show the sufficiency: For an unobserved variable H_i in the ground truth model M, if it has an mleaf child that satisfies the condition described in Proposition

Algorithm 2: Recovering all models in the AOG / DOG Equivalence Class

Input: Recovered mixing matrix $\mathbf{W}^*$, observability indicators of the variables.

Obtain the AOG of the true model using Algorithm 1.

Denote row as a selection of centers in the groups in $\mathcal{C}_{cogent}$. Define Row as the set of all rows.

Denote col as a selection of noises in the groups in $\mathcal{C}_{cogent}$. Define Col as the set of all cols.

Initialize $\mathcal{M}_{AOG} = \emptyset$.

for *all* $row \in Row$ *and* $col \in Col$ **do**

  $\mathbf{C} = \mathbf{I} - (\mathbf{W}^*[row, col])^{-1}$.

  $\mathbf{B}^C = (\mathbf{I} - \mathbf{C})\mathbf{W}^*[row, col^C]$.

  $\mathbf{D} = \mathbf{W}^*[row^C, col](\mathbf{I} - \mathbf{C})$.

  $\mathbf{B}^L = \mathbf{W}^*[row^C, col^C] - \mathbf{D}(\mathbf{I} - \mathbf{C})^{-1}\mathbf{B}^C$.

  $M = \{\mathbf{B} = [\mathbf{B}^L; \mathbf{B}^C], \mathbf{C}, \mathbf{D}\}$. Add M to $\mathcal{M}_{AOG}$.

Define $\mathcal{M}_{DOG}$ as the set of models in $\mathcal{M}_{AOG}$ that have the fewest total number of non-zero entries in $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$.

Output: $\mathcal{M}_{AOG}$, $\mathcal{M}_{DOG}$.

4, then M is not minimal, i.e., there exists an alternative model M' without H_i that has the same mixing matrix and satisfies Assumption 1. Next, we show the necessity: If M is not minimal, then there must exist an unobserved variable H_i and one of its mleaf child Z_j such that the described condition is satisfied.

F.1.1 Proof of sufficiency

Suppose there exists latent variable H_i and a mleaf child Z_j of H_i in M such that the condition described in Proposition 4 holds. In the following we construct the alternative model M' that includes all variables in M except for H_i. The idea is to consider N_{H_i} as the exogenous noise term of Z_j. Further, for any other child V_k of H_i, replace the edge $H_i \to V_k$ in M by edges from Z_j (and parents of Z_j) to V_k in M'.

The structural equation of Z_j in M can be written as

$$Z_j = \sum_{l:V_l \in Pa_M(Z_j) \cap \mathcal{V}^C} a_{jl} V_l + \sum_{l_j : H_{l_j} \in Pa_M(Z_j) \cap \mathcal{H}} b_{jl_j} H_{l_j}. \tag{4}$$

Consider N_{H_i} as the exogenous noise term of Z_j in M'. The structural equation of Z_j in M' is

$$Z_j = \sum_{l:Z_l \in Pa_M(Z_j) \cap \mathcal{V}^C} a_{jl} Z_l + \sum_{l_j : H_{l_j} \in Pa_M(Z_j) \cap \mathcal{H} \setminus \{H_i\}} b_{jl_j} H_{l_j} + b_{ji} N_{H_i}. \tag{5}$$

For any other children V_k of H_i in M, the structural equation of V_k in M can be written as

$$V_k = \sum_{l:Z_l \in Pa_M(V_k) \cap \mathcal{V}^C} a_{kl} Z_l + \sum_{l_k : H_{l_k} \in Pa_M(V_k) \cap \mathcal{H}} b_{kl_k} H_{l_k} + N_{V_k}. \tag{6}$$

By considering N_{H_i} in Equation (6) to be the exogenous noise of Z_j, the structural equation of Z_m in M' can be written as

$$\begin{aligned} V_k = & \sum_{l:Z_l \in Pa_M(V_k) \cap \mathcal{V}^C} a_{kl} Z_l + \sum_{l_k : H_{l_k} \in Pa_M(V_k) \cap \mathcal{H} \setminus \{H_i\}} b_{kl_k} H_{l_k} + N_{Z_m} + \\ & b_{ki} b_{ji}^{-1} \left(Z_j - \sum_{l:Z_l \in Pa_M(Z_j) \cap \mathcal{V}^C} a_{jl} Z_l - \sum_{l_j : H_{l_j} \in Pa_M(Z_j) \cap \mathcal{H} \setminus \{H_i\}} b_{jl_j} H_{l_j} \right). \end{aligned} \tag{7}$$

Since H_i and Z_j satisfy the condition in Proposition 4, V_k cannot be an ancestor of Z_j or variables in $Pa_M(Z_j)$ in M, otherwise we have $V_k \in \cup_{Z \in Pa(Z_j)} An_M(Z = An_M(Z_j) \subseteq An_M(V_k)$. This implies that M' is still acyclic. Further, since $An_M(Z_j) \subseteq An_M(V_k)$, there are no additional ancestors introduced to V_k in M' compared with M. Lastly, we note that there might be edge cancellations in (4). In particular, the coefficient of the direct edge from a variable in $Z \in Pa_M(V_k) \cap \mathcal{V}^C$ to V_k may change in M' if $Z \in Pa_M(Z_j)$ and hence may be cancelled out. However, Z is still an ancestor of V_k in M', as there is the path $Z \to Z_j \to V_k$. As M and M' has the same mixing matrix, the conventional faithfulness is still satisfied.

In conclusion, if such H_i and Z_j exists in $\mathcal{M}$, then there exists an alternative model $\mathcal{M}'$ such that we cannot distinguish $\mathcal{M}'$ from $\mathcal{M}$ under Assumption 1 while having one less latent variable. Hence the sufficiency is proved.

F.1.2 Proof of necessity

Suppose M is not minimal. Then there exists an alternative model M' that has the same mixing matrix as M and also satisfies conventional faithfulness assumption, while having fewer unobserved variables. Without loss of generality, suppose M' is minimal. Note that since both models correspond to the same mixing matrix, this implies that the number of measured cogent variables in M is strictly less than M', which equals to the number of columns of in the mixing matrix subtract the number of unobserved variables.

Now, we partition the cogent and mleaf variables in M and M' as follows, where we put measured variables with the same row support in the mixing matrix in the same group. We note that in M', this is the same partition as the ancestral ordered grouping among these variables.

Consider the set of measured cogent variables in M. According to the definition, they must have different row support hence each of them must belong to a separate group. Therefore, since M has fewer measured cogent variables than M', there exists at least one group $\mathcal{G}$, where variables $\mathcal{G}$ are all mleaf variables in M and one of the variables in $\mathcal{G}$ is a measured cogent variable in M'. Denote this measured cogent variable in M' as Z_j. Consider the column corresponding to the exogenous noise term of $\mathcal{G}$ in the mixing matrix.

We first show that this column corresponds to a latent confounder in M by contradiction. Suppose this column corresponds to the exogenous noise of a measured cogent variable Z_i in M. Therefore Z_i must be an ancestor of Z_j in M, and the entry with row corresponding to Z_i and column corresponding to this noise is not zero. Denote $Supp(Z_i)$, $Supp(Z_j)$ as the support of the row corresponding to Z_i and Z_j, respectively. Since M satisfies conventional faithfulness, the total causal effects from all ancestors of Z_i on Z_j in M is not zero. Hence $Supp(Z_i) \subseteq Supp(Z_j)$. Similarly, consider Z_i in M'. Since Z_j is the a measured cogent variable, Z_j is an ancestor of Z_i, and $Supp(Z_j) \subseteq Supp(Z_i)$. Therefore we have $Supp(Z_j) = Supp(Z_i)$, and hence both belong to the same group $\mathcal{G}$. However, no variables in $\mathcal{G}$ are cogent variables in M, which leads to a contradiction. Therefore, this column must correspond to a latent confounder in M denote this latent confounder as H_i in M.

Next, we consider any other child V_k of H_i in M. V_k must be an descendant of Z_j, hence $Supp(Z_j) \subseteq Supp(V_k)$. As M satisfies conventional faithfulness, this implies that $An(Z_j) \subseteq An(V_k)$. Therefore the condition in the proposition is satisfied on M.

F.2 Enumerating all models in the AOG and DOG equivalence classes by different choice of centers

We first show that an LV-SEM-ME can be uniquely deduced given the mixing matrix $\mathbf{W}^*$ of the LV-SEM-ME, and a choice of the centers (and their corresponding exogenous noises) in each group. This has been described in Algorithm 2, where the matrices $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ can be found through matrix calculation.

In this following, given the ground-truth model M, we will show how to deduce the structural equations of the variables in the alternative model M', where M and M' has the same mixing matrix and the same (ancestral ordered or directed ordered) grouping of the variables. Specifically, we consider the case when the centers of the groups are the same between M and M' except for one group. We denote the center of this only group in M as V_i with exogenous noise N_{V_i}, while the center and the corresponding exogenous noise in M' are Z_j and N_{H_l}, where Z_j and H_k belong to the same group as V_i in M. We will show the structural equations of all variables that are affected by this difference. This will be used in the proof of AOG and DOG results in F.3 and F.4, respectively.

The structural equation of V_i in M can be written as:

$$V_i = \sum_{m_i : H_{m_i} \in Pa_{\mathcal{H}}(V_i)} b_{im_i} H_{m_i} + \sum_{n_i : V_{n_i} \in Pa_{\mathcal{V}^C}(V_i)} a_{in_i} V_{n_i} + N_{V_i} \tag{8}$$

Note that $H_l \in Pa_{\mathcal{H}}(V_i)$. For Z_j, since it is an mleaf child of V_i, we have:

$$Z_j = a_{ji} V_i + \sum_{m_j : H_{m_j} \in Pa_{\mathcal{H}}(Z_j)} b_{jm_j} H_{m_j} + \sum_{n_j : V_{n_j} \in Pa_{\mathcal{V}^C}(Z_j) \backslash \{V_i\}} a_{jn_j} V_{n_j}. \tag{9}$$

We can also write down the equations of any other children V_{k_i} of V_i, and any other children V_{k_l} of H_l:

$$\begin{aligned}
V_{k_i} &= a_{k_i i} V_i + \sum_{m_{k_i} : H_{m_{k_i}} \in Pa_{\mathcal{H}}(V_{k_i})} b_{k_i m_{k_i}} H_{m_{k_i}} + \sum_{n_{k_i} : V_{n_{k_i}} \in Pa_{\mathcal{V}^C}(V_{k_i}) \backslash \{V_i\}} a_{k_i n_{k_i}} V_{n_{k_i}} + N_{V_{k_i}}, \\
V_{k_l} &= b_{k_l l} H_l + \sum_{m_{k_l} : H_{m_{k_l}} \in Pa_{\mathcal{H}}(V_{k_l}) \backslash \{H_l\}} b_{k_l m_{k_l}} H_{m_{k_l}} + \sum_{n_{k_l} : V_{n_{k_l}} \in Pa_{\mathcal{V}^C}(V_{k_l})} a_{k_l n_{k_l}} V_{n_{k_l}} + N_{V_{k_l}}.
\end{aligned} \tag{10}$$

(11)

Now, consider M'. We can first write down the equation for V_i, which is now a mleaf:

$$V_i = a_{ji}^{-1} \left(Z_j - \sum_{m_j : H_{m_j} \in Pa_{\mathcal{H}}(Z_j)} b_{jm_j} H_{m_j} - \sum_{n_j : V_{n_j} \in Pa_{\mathcal{V}^C}(Z_j) \backslash \{V_i\}} a_{jn_j} V_{n_j} \right). \tag{12}$$

For Z_j, by plugging in V_i from (8) to (9), we have:

$$\begin{aligned}
Z_j = a_{ji} &\left(b_{il} H_l + \sum_{m_i : H_{m_i} \in Pa_{\mathcal{H}}(V_i) \backslash \{H_l\}} b_{im_i} H_{m_i} + \sum_{n_i : V_{n_i} \in Pa_{\mathcal{V}^C}(V_i)} a_{in_i} V_{n_i} + N_{V_i} \right) \\
&+ \sum_{m_j : H_{m_j} \in Pa_{\mathcal{H}}(Z_j)} b_{jm_j} H_{m_j} + \sum_{n_j : V_{n_j} \in Pa_{\mathcal{V}^C}(Z_j) \backslash \{V_i\}} a_{jn_j} V_{n_j}.
\end{aligned}$$

Next, since N_{H_l} is the exogenous noise term of Z_j, and N_{V_i} is the exogenous noise of H_l, we can rewrite it as

$$\begin{aligned}
Z_j = a_{ji} &\left(H_l + \sum_{m_i : H_{m_i} \in Pa_{\mathcal{H}}(V_i) \backslash \{H_l\}} b_{im_i} H_{m_i} + \sum_{n_i : V_{n_i} \in Pa_{\mathcal{V}^C}(V_i)} a_{in_i} V_{n_i} \right) \\
&+ \sum_{m_j : H_{m_j} \in Pa_{\mathcal{H}}(Z_j)} b_{jm_j} H_{m_j} + \sum_{n_j : V_{n_j} \in Pa_{\mathcal{V}^C}(Z_j) \backslash \{V_i\}} a_{jn_j} V_{n_j} + a_{ji} b_{il} N_{H_l}.
\end{aligned} \tag{13}$$

For V_{k_i}, by plugging in V_i from (12) to (10), we have

$$\begin{aligned}
V_{k_i} = \frac{a_{k_i i}}{a_{ji}} &\left(Z_j - \sum_{m_j : H_{m_j} \in Pa_{\mathcal{H}}(Z_j)} b_{jm_j} H_{m_j} - \sum_{n_j : V_{n_j} \in Pa_{\mathcal{V}^C}(Z_j) \backslash \{V_i\}} a_{jn_j} V_{n_j} \right) \\
&+ \sum_{m_{k_i} : H_{m_{k_i}} \in Pa_{\mathcal{H}}(V_{k_i})} b_{k_i m_{k_i}} H_{m_{k_i}} + \sum_{n_{k_i} : V_{n_{k_i}} \in Pa_{\mathcal{V}^C}(V_{k_i}) \backslash \{V_i\}} a_{k_i n_{k_i}} V_{n_{k_i}} + N_{V_{k_i}}.
\end{aligned} \tag{14}$$

Lastly, for V_{k_l}, we substitute H_l in (11) by N_{H_l} in (13) and have

$$\begin{aligned}
V_{k_l} = \frac{b_{k_l l}}{a_{ji} b_{il}} Z_j - \frac{b_{k_l l}}{b_{il}} \left(H_l + \sum_{m_i : H_{m_i} \in Pa_{\mathcal{H}}(V_i) \setminus \{H_l\}} b_{i m_i} H_{m_i} + \sum_{n_i : V_{n_i} \in Pa_{\mathcal{V}C}(V_i)} a_{i n_i} V_{n_i} \right) \\
- \frac{b_{k_l l}}{a_{ji} b_{il}} \left(\sum_{m_j : H_{m_j} \in Pa_{\mathcal{H}}(Z_j)} b_{j m_j} H_{m_j} + \sum_{n_j : V_{n_j} \in Pa_{\mathcal{V}C}(Z_j) \setminus \{V_i\}} a_{j n_j} V_{n_j} \right) \\
+ \sum_{m_{k_l} : H_{m_{k_l}} \in Pa_{\mathcal{H}}(V_{k_l}) \setminus \{H_l\}} b_{k_l m_{k_l}} H_{m_{k_l}} + \sum_{n_{k_l} : V_{n_{k_l}} \in Pa_{\mathcal{V}C}(V_{k_l})} a_{k_l n_{k_l}} V_{n_{k_l}} + N_{V_{k_l}}.
\end{aligned} \tag{15}$$

We note that if $V_i \in Pa(V_{k_l})$, then we have to replace $a_{k_l i} V_i$ by the right hand side of Equation (12).

To summarize, compared with M, the changes in the parent-child relationships among variables in M' can be summarized as follows:

(i) For V_i, since it is an mleaf node in M', its parents in M' are the parents of Z_j in M (excluding V_i itself), plus Z_j.

(ii) For Z_j, since it is the new center, its parents in M' are the parents of Z_j itself in M (excluding V_i), plus the parents of V_i in M.

(iii) For any child V_{k_i} of V_i in M (other than Z_j), compared with its parent set in M, the new parent set in M replaces V_i by Z_j, and includes additional variables that are the parents of Z_j in M.

(iv) For any child V_{k_l} of H_l in M (other than V_i), compared with its parent set in M, the new parent set in M' additionally includes the new center Z_j and all parents of Z_j in M' (i.e., parents of V_i and Z_j in M).

Note that there may be changes of model coefficients if the "additional variables" are already in the parent set of V_{k_i} and V_{k_l}. In particular, this change may lead to removal of variables from the parent set if the coefficients cancel out each other.

F.3 Proof of the identification result under conventional faithfulness

In this subsection, we provide the proof of the result regarding the AOG equivalence class in LV-SEM-ME, i.e., Theorem 2(a). Note that the proof of the result for SEM-ME and LV-SEM, i.e., Theorem 1(a), can be deduced from it.

To show that the extent of identifiablity of an LV-SEM-ME under conventional faithfulness is the AOG equivalence class, we need to show that, for the ground-truth model M, any other model M' in the AOG equivalence class of M, and any model M'' that has the same mixing matrix but does not belong to the AOG equivalence class of M:

(1.a) M' satisfies conventional faithfulness.

(1.b) M' is consistent with any causal order among the ancestral ordered groups that is consistent with M.

(1.c) M'' violates conventional faithfulness.

Recall that for each cogent variable V_i, the ancestral ordered group of V_i includes its mleaf child Z_j if V_i is measured and all other parents of Z_j are also ancestors of V_i, and unobserved parent H_l if all other children of H_l are also descendants of V_i.

Proof of (1.a). We note that it suffices to show (a) when the choices of centers (and/or the corresponding exogenous noise) of M' only differs from the choices of M in one group. This is because if there are p differences in the choices of centers, then we can always find a finite sequence of models $M_0 \to M_1 \to \cdots \to M_p$, where $M_0 = M$, $M_p = M'$, and for each M_{k_p}, $k_p \in [p]$, the choices of centers only differs from the choices of centers of $M_{k_p - 1}$ in one group. If (a) hold for models with one difference of the choices, then following the sequence of the models, (a) must also hold for M_p.

We prove by contradiction. Suppose V is an ancestor of V' in M' and the total causal effect from V to V' is zero. Note that the total causal effect from V to V' is equal to the sum of path products from the exogenous noise of V to V'. Suppose the exogenous noise of V in M' is the exogenous noise of V_0 in M. Since M satisfies Assumption 2, V_0 is not an ancestor of V' in M. This means that the added edges in M' introduces additional ancestors to V'.

Note that if $V' = Z_j$, then we compare the addition of ancestors to Z_j in M' with the ancestors of V_j in M as both are the center variable in the corresponding model. Similarly, we compare the addition of ancestors to V_j in M' with the ancestors of V_i in M.

However, as we described in Appendix F.2, all added edges in M' can be categorized as follows:

- If $V' = Z_j$. According to (ii), all added edges in M' are from parents of Z_j in M to Z_j. However, parents of Z_j are all ancestors of V_i in M. Therefore no additional ancestors are introduced.
- If $V' = V_{k_i}$. According to (iii), all added edges in M' are also from parents of Z_j in M to Z_j. Since there is V_i is a parent of V_{k_i} in M, parents of Z_j are all ancestors of V_{k_i} in M.
- If $V' = V_{k_l}$. According to (iv), all added edges in M' are also from parents of V_i or Z_j in M to Z_j. Since V_i is an ancestor of V_{k_l} in M, parents of V_i or Z_j are all ancestors of V_{k_l}.

In conclusion, the added edges in M' does not introduce any additional ancestors to V', which leads to a contradiction. Therefore M' satisfies Assumption 2.

Proof of (1.b). Since M and M' both satisfies conventional faithfulness, then Proposition 5 holds. Further, for any group g, define $Supp(g)$ as the row support of variables in g in $\mathbf{W}^*$ if g includes any observed or measured variables, and if $g = \{H_i\} \subseteq \mathcal{H}$, define $Supp(g)$ as $\{N_{H_i}\}$. Similarly, define $Supp(N_g)$ as the column support of the exogenous noises of the variables in g in $\mathbf{W}^*$ if g includes any cogent or unobserved variables, and if $g = \{Z_L\} \subseteq \mathcal{Z}^L$, define $Supp(g)$ as $\{Z_L\}$. We have the following property.

Proposition 9 *For two different ancestral ordered groups g_i and g_j, the following three conditions are equivalent:*

(i) There exists a causal path from one variable in the group of g_i to one variable in the group of g_j.

(ii) $Supp(g_i) \subset Supp(g_j)$.

(iii) $Supp(N_{g_j}) \subset Supp(N_{g_i})$.

Therefore, given M, a causal order among the groups is consistent with the causal order among the variables if and only if the set relations in the mixing matrix hold. Since M and M' hae the same mixing matrix and satisfy Assumption 2, they are consistent with the same set of causal order among the groups.

Proof of (1.c). Suppose M'' does not belong to the AOG equivalence class of M. First, consider the number of cogent variables in M'', which is equal to the number of columns of $\mathbf{W}^*$ minus the number of unobserved variables. Given that M is minimal, M'' cannot have more cogent variables than M. Similarly, if M'' has fewer cogent variables than M, then M'' is not minimal. Therefore, M and M'' must have equal number of cogent variables.

Next, consider the cogent variables in M'', and their position in the AOG of M. Denote these cogent variables as $\mathcal{P}$. Firstly, note that no two variables in $\mathcal{P}$ belongs to the same ancestral ordered group of M. This is because the mixing matrix corresponding to $\mathcal{P}$ must be lower-triangular following the causal order, which is impossible when two variables in $\mathcal{P}$ have the same row support. Therefore, $\mathcal{P}$ includes at most one variable in each group. Similarly, the exogenous noises of variables in $\mathcal{P}$ in M'', denoted by $N_{\mathcal{P}}$, includes the exogenous noise of at most one variable in each group. Suppose g is the a group with cogent variables where either (I) g does not include any variable in $\mathcal{P}$, or (II) g does not include any variable whose corresponding exogenous noise is in $N_{\mathcal{P}}$. Note that such a group must exist, otherwise M'' belongs to the same AOG equivalence class. Denote this cogent variable in M as V_i, and its exogenous noise N_{V_i}.

(I): Suppose g does not include any variable in $\mathcal{P}$. Then V_i is an mleaf in M''. Consider N_{V_i} in M''. If there exists one parent of V_i in M'' that also includes N_{V_i} (i.e., is affected by N_{V_i} directly or indirectly), denote this variable as V_j, which must be in $\mathcal{P}$ and not in g. Since V_j includes N_{V_i}, it must be a descendant of V_i in M. However, since M satisfies conventional faithfulness and $V_j \notin g$, the row support of V_j must be a superset of V_i. This implies that conventional faithfulness is violated on M''.

Therefore, none of the parents of V_i in M'' is affected by N_{V_i}, and there must be a directed edge from N_{V_i} to V_i. Since V_i is an mleaf, N_{V_i} corresponds to a latent confounder in M''. Further, for any other children V of N_{V_i} in M'', it must be a descendant of V_i in M, and hence $Supp(V_i) \subseteq Supp(V)$. According to Proposition 4, this implies that M'' is not minimal.

(II): Suppose g does not include any variable whose exogenous noise is in $N_{\mathcal{P}}$. Then N_{V_i} corresponds to an unobserved confounder in M''. Similarly, consider V_i in M''. If there is one cogent variable in M'' that is a child of N_{V_i} and affects V_i (or V_i is this cogent variable), denote the exogenous noise of this cogent variable as N_V, which is in $N_{\mathcal{P}}$. Since N_V is an ancestor of V_i in M, $Supp(N_{V_i}) \subseteq Supp(N_V)$. Further, as N_V is not the exogenous noise of any (observed or latent) variable in g, $Supp(N_{V_i})$ must be a strict subset of $Supp(N_V)$. As any descendants of N_V must also be a descendant of N_{V_i} in M'', this implies that conventional faithfulness is violated on M''.

Therefore, none of the parents of V_i in M'' is affected by N_{V_i}, and V_i cannot be a cogent variable. Hence V_i is an mleaf variable in M'' and is directly affected by N_{V_i}. Similarly, for any other children V of N_{V_i} in M'', it must be a descendant of V_i in M, and hence $Supp(V_i) \subseteq Supp(V)$. According to Proposition 4, this implies that M'' is not minimal.

F.4 Proof of the identification result under LV-SEM-ME faithfulness

In this subsection, we provide the proofs of all results regarding the DOG equivalence class in LV-SEM-ME, i.e., Proposition 3, and 7, and Theorem 2(b). These proof of the results for for SEM-ME and LV-SEM, i.e., Proposition 2 and Theorem 1(b) can be deduced from it.

To show that the extent of identifiablity of an LV-SEM-ME under LV-SEM-ME faithfulness is the DOG equivalence class, the proof includes two parts.

First, we show that, for the ground-truth model M, any other model M' in the DOG equivalence class:

- (2.a) M' does not add extra edge compared with M.
- (2.b) M' does not remove any edge compared with M.
- (2.c) M' satisfies LV-SEM-ME faithfulness.

Therefore we cannot distinguish M' from M. Next, any other model M'' that is in the AOG equivalence class of M but not the DOG equivalence class:

- (2.d) M'' adds at least one extra edge compared with M.
- (2.e) M'' does not remove any edge compared with M, and there is at least one added edge that is not removed.
- (2.f) M'' violates LV-SEM-ME faithfulness.

Therefore we can disting huish M' from M.

F.4.1 Model within the same DOG equivalence class

Similar to (1.a) in the AOG proof, we only need to show the result if M' only differs from M in the choices of centers (and/or the corresponding exogenous noise) in one group. If there are p differences in the choices of centers between M and M', then we can always find a finite sequence of models $M_0 \to M_1 \to \cdots \to M_p$, where $M_0 = M$, $M_p = M'$, and for each M_{k_p}, $k_p \in [p]$, the choices of centers only differs from the choices of centers of $M_{k_p - 1}$ in one group.

In the following, we show (2.a) - (2.c) together for the one difference in the choices of centers. Specifically, for (2.b), we show that if one edge is cancelled out, then M violates LV-SEM-ME faithfulness. Further, for (2.c), we show that a single change still preserves LV-SEM-ME faithfulness. Following the sequence of the models, all three properties hold for M'.

Proof of (2.a). Following the description in Appendix F.2, after replacing the center V_i in M with Z_j, and replacing the exogenous noise N_{V_i} in M with N_{H_l}, where Z_j and H_l belong to the same direct ordered group as V_i in M, all added edges in M' can be categorized as follows:

- For V_i: No edges are added.
- For Z_j: According to (ii), all added edges in M' are from parents of Z_j (excluding V_i) in M to Z_j. However, according to Condition 1(a), since Z_j belongs to the same direct ordered group as V_i, $Pa(Z_j) \setminus \{V_i\}$ is a subset of $Pa(V_i)$. Therefore no edges are added.
- For V_{k_i}: According to (iii), all added edges in M' are from parents of Z_j in M (excluding V_i) to V_{k_i}. Still, according to Condition 1(b), since Z_j belongs to the same direct ordered group as V_i, $Pa(Z_j)$ is a subset of $Pa(V_{k_i})$. Therefore no edges are added.
- For V_{k_l}: According to (iv), the added edges are from Z_j and parents of V_i and V_j in M to V_{k_l}. Since H_l belongs to the same direct ordered group as V_i in M, $\{V_i\} \cap Pa(V_i) \subseteq Pa(V_{k_l})$ in M. Since the center is replaced by Z_j and $Pa(Z_j) \setminus \{V_i\}$ is a subset of $Pa(V_i)$, no edges are added.

Therefore no edges are added to M' compared with M.

Proof of (2.b). We show that if any edge is cancelled out in M' described in Appendix F.2, then M violates LV-SEM-ME faithfulness. Similarly, all removed edges in M' can be categorized as follows:

- For V_i: No edges are removed.
- For Z_j: According to (ii), the edge from V to V_i in M may be cancelled out in M' if V is also a parent of Z_j. Suppose the edge from V to V_i is cancelled out in M' because of this. We show that if this happens, then M violates LV-SEM-ME faithfulness. Note that V can be either cogent or unobserved.

  If V is cogent: Consider variable Z_j, the set of cogent ancestors $J = An_{M'}(Z_j) \cap \mathcal{V}^C$ in M', and the set of cogent parents $K = Pa_{M'}(Z_j) \cap \mathcal{V}^C$ in M'. We have $An_{M'}(Z_j) \cap \mathcal{V}^C = An_V(V_j) \setminus \{V_i\}$ and $K \subset Pa_V(V_i)$, because $V \notin K$. Next, following the structural equation of Z_j in M', we have: $Rank(\mathbf{W}^J_{K \cup \{Z_j\}}) = |K|$. However, the minimal bottleneck from J to $K \cup \{Z_j\}$ in M is at least $|K| + 1$, because $K \subset J$, and there is one extra path in M ($V \to V_i \to Z_j$) that is not blocked by K. Therefore, Z_j violates Assumption 3(b).

  If V is unobserved: Consider the set of cogent ancestors $J = An_{M'}(Z_j) \cap \mathcal{V}^C$ plus $\{V\}$ in M', and the set of cogent parents $K = Pa_{M'}(Z_j) \cap \mathcal{V}^C$ in M'. In this case, K may equal to $Pa_V(V_i)$. However, we can still show that $Rank(\mathbf{W}^J_{K \cup \{Z_j\}}) = |K|$ (since there is no direct connection from V to Z_j in M'), but the minimal bottleneck from J to $K \cup \{Z_j\}$ in M is at least $|K| + 1$, because the path $V \to V_i \to Z_j$ is not blocked. Therefore Z_j still violates Assumption 3(b).
- For V_{k_i}: According to (iii), the edge from V to V_{k_i} in M may be cancelled out in M' if V is also a parent of Z_j. Suppose the edge from V to V_{k_i} is cancelled out in M' because of this. Similarly, V can be either cogent or observed.

  If V is cogent: Consider variable V_{k_i}, $J = An_V(V_{k_i}) \setminus \{V_i\} \cup \{H_l\}$, and the set of cogent parents $K = Pa_{M'}(V_{k_i}) \cap \mathcal{V}^C$ in M'. We have $K \subset PP(V_{k_i})$ (note that Z_j is an mleaf in M), and $|K| < Pa_V(V_{k_i})$. Further, J includes all variables with the exogenous noise corresponding to the cogent ancestors of Z_j in M'. Therefore, we have $Rank(\mathbf{W}^J_{K \cup \{V_{k_i}\}}) = |K|$. However, the minimal bottleneck from J to $K \cup \{V_{k_i}\}$ in M is at least $|K| + 1$. Firstly, $K \setminus \{Z_j\}$ is a subset of J so they are included in any bottleneck. Additionally, there are two distinct paths from J to $K \cup \{V_{k_i}\}$ that cannot be blocked by $K \setminus \{Z_j\}$: $V \to V_{k_i}$ and $H_l \to V_i \to Z_j$. Therefore V_{k_i} violates Assumption 3(a).

  If V is unobserved: Similarly, consider $J = An_V(V_{k_i}) \setminus \{V_i\} \cup \{V, H_l\}$ and $K = Pa_{M'}(V_{k_i}) \cap \mathcal{V}^C$. Note that $V \neq H_l$ as V is a parent of V_{k_i} in M. The results above still hold, as $Rank(\mathbf{W}^J_{K \cup \{V_{k_i}\}}) = |K|$, and the same two paths, $V \to V_{k_i}$ and $H_l \to V_i \to Z_j$, cannot be blocked by $K \setminus \{Z_j\}$. Therefore V_{k_i} violates Assumption 3(a).
- For V_{k_l}: According to (iv), the edge from V to V_{k_l} in M may be cancelled out in M' if $V = Z_j$, or $V \in Pa_{M'}(Z_j) \cap V^C$. Suppose the edge from V to V_{k_l} is cancelled out in M' because of this.

If V is cogent and $V \in Pa_V(V_i)$: Consider variable V_{k_l}, the set of cogent ancestors $J = An_V(V_{k_l}) \setminus \{V_i\} \cup \{H_l\}$ in M, and the set of cogent parents $K = Pa_{M'}(V_{k_l}) \cap \mathcal{V}^C$ in M'. Similarly, we have $K \subset PP(V_{k_l})$, and $|K| < Pa_V(V_{k_l})$. We have $Rank(\mathbf{W}^J_{K \cup \{V_{k_l}\}}) = |K|$. Further, the minimal bottleneck from J to $K \cup \{V_{k_l}\}$ in M is at least $|K| + 1$. Firstly, $K \setminus \{Z_j\}$ is a subset of J so they are included in any bottleneck. Additionally, there are two distinct paths from J to $K \cup \{V_{k_l}\}$ that cannot be blocked by $K \setminus \{Z_j\}$: $V \to V_{k_l}$ and $H_l \to V_i \to Z_j$. Therefore V_{k_l} violates Assumption 3(a).

If $V = Z_j$: Consider the same J and K as above. The main difference here is that $Z_j \notin K$ and hence $K \subset J$. Therefore, $Rank(\mathbf{W}^J_{K \cup \{V_{k_l}\}}) = |K|$, and the minimal bottleneck at least includes all variables in K, and one extra variable on the edge $H_l \to V_{k_l}$. Therefore V_{k_l} violates Assumption 3(a).

If V is unobserved: Consider $J = An_V(V_{k_l}) \setminus \{V_i\} \cup \{H_l, V\}$ in M, and $K = Pa_{M'}(V_{k_l}) \cap \mathcal{V}^C$ in M'. We have $Rank(\mathbf{W}^J_{K \cup \{V_{k_l}\}}) = |K|$, and the minimal bottleneck includes all variables in $K \setminus \{Z_j\}$, as well as two variables on the paths $V \to V_{k_l}$ and $H_l \to V_i \to Z_j$. Therefore V_{k_l} violates Assumption 3(a).

In conclusion, if an edge is cancelled in M', then M violates Assumption 3.

Proof of (2.c). From (2.a) and (2.b), we show that M' has the same unlabeled graph structure as M. We can construct a mapping σ_M on variables in $\mathcal{Z} \cup \mathcal{Y}$, where

$$\sigma(V_i) = Z_j; \quad \sigma(Z_j) = V_i; \quad \sigma(V) = V, \quad V \neq V_i, Z_j.$$

Similarly, define a mapping σ_N on variables in $\mathcal{V}^C \cup \mathcal{H}$, where

$$\sigma_N(V_i) = H_l; \quad \sigma(H_l) = V_i; \quad \sigma(V) = V, \quad V \neq V_i, H_l.$$

We note that for all variable V except V_i and Z_j, $PP_M(V) = PP_{M'}(V)$, $An_{M'}(V) = An_M(V)$. This is because $PP(V)$ either includes both V_i and Z_j, or include neither of them, as all parents of Z_j in M are ancestors of V_i. Similarly, $An_M(V)$ either includes both H_l and V_i, or neither of them, because H_l is a parent of V_i, and all other children of H_l are descendants of V_i.

In the following, we show that if M satisfies Assumption 3, then M' also satisfies Assumption 3 with probability one. In particular, following the rewriting procedure described in Appendix F.2, we can construct an invertible mapping between the model parameters in M and in M'. Since Assumption 3 is violated with probability zero on the model parameters in M, the same results hold on the model parameters in M'.

We note that the differences in the model parameters that are different between M and M' can be summarized as follows:

(i) The edge weight from Z_j to V_i is the inverse of the edge weight from V_i to Z_j, a_{ji};

(ii) The edge weight from other parents of V_i to V_i in M' can be written as the edge weight from other parents of Z_j to Z_j in M multiplied by $-a_{ji}^{-1}$;

(iii) The edge weight from $Pa(Z_j)$ to Z_j in M' can be written as a function of a_{ji}, the edge weight from $Pa(V_i)$ to V_i in M and the edge weight from $Pa(Z_j)$ to Z_j;

(iv) The edge weight from Z_j to V_{k_i} is edge weight from V_i to V_{k_i}, $a_{k_i i}$ divided by a_{ji};

(v) The edge weight from other parents of V_{k_i} to V_{k_i} in M' can be written as a function of $a_{k_i i}$, a_{ji} the edge weight from $Pa(V_{k_i})$ to V_{k_i} and the edge weight from $Pa(Z_j)$ to Z_j in M;

(vi) The edge weight from H_l to V_{k_l} is edge weight from H_l to V_{k_l}, $b_{k_l l}$, divided by b_{il} (note that b_{il} is included in (iii));

(vii) The edge weight from Z_j to V_{k_l} is a function of the edge weight from V_i to V_{k_l} ($a_{k_l i}$), a_{ji}, b_{il}, $b_{k_l l}$;

(viii) The edge weight from $Pa(V_{k_l})$ to V_{k_l} in M' can be written as a function of the edge weight from $Pa(V_{k_l})$ to V_{k_l}, and the edge weight from $Pa(Z_j)$ to Z_j, and $Pa(V_i)$ to V_i in M.

Therefore, if we arrange the model parameters in M and in M' following (i)-(viii) above, we can clearly see that the mapping that translates model parameters in M to model parameters in M' is invertible. This is consistent with the fact that since M and M' belong to the same DOG, we can equivalently write the model parameters in M using model parameters in M'. Therefore, since the model parameters in M satisfy Assumption 3 with probability one, the model parameters in M' also satisfy Assumption 3 with probability one.

Lastly, we note that, since the number of models in the DOG equivalence class is finite, if the ground-truth model satisfies Assumption 3, then the probability that models in the DOG equivalence class all satisfy Assumption 3 is also one.

F.4.2 Model outside the DOG equivalence class

Without loss of generality, we only need to show the result if all the differences in the choices of centers in M'' are outside the corresponding direct ordered group of the cogent variable in M, but in the same ancestral ordered group. That is, we consider M to be the model that is closest to M'' in terms of the differences of the centers. We have shown above that this closest model also satisfies LV-SEM-ME faithfulness and share the same unlabeled graph structure.

Proof of (2.d). We first show that if M'' differs from M only in one choice of the centers, then there will be at least one edge added to M''. The proof also follows the procedure described in Appendix F.2. That is, if Z_j belong to the same ancestral ordered group but not the same direct ordered group as V_i, then Condition 1 is satisfied. In this case, there will be one additional edge either from parents of Z_j in M to Z_j, or to V_{k_i}. Similarly, if H_l belong to the same ancestral ordered group but not the same direct ordered group as V_i, then Condition 2 is satisfied. In this case, there will be one additional edge from Z_j, or the parents of Z_j or V_i to V_{k_l}. In conclusion, there will be at least one added edge in M'' compared with M.

We note that if there are multiple differences in the centers between M'' and M, then we may not able to use the above analysis. In particular, since one added edge may change the parent-child relation among variable, it might be the case that no edges are added to M_2 compared with M_1. Nevertheless, we can still show that M_2 has at least one more edge compared with M_0. Repeating this procedure following the chain $M_0 \to M_1 \to \cdots \to M_p$ completes the proof.

Proof of (2.e). We note that we can use the same method as in (2.b) to show that if any edge is removed in M, then M violates Assumption 3. Specifically, in (2.b), we only used the fact that Z_j is a child of V_i, and H_l is a parent of V_i. Both still hold true for AOG. In the following, we will only show the latter part, that is, there is at least one added edge that is not removed.

We prove by contradiction. Suppose all added edges are removed in M'', and suppose the edge from V_0 to V is added but finally removed. We further assume that among all the added edges $V_1 \to V_2$, V_0 has the smallest index (following the causal order in M), and for any $V' \in De(V_0) \cap An(V)$, no edges from V_0 to V' is added. That is, for any causal path from V_0 to V, no edges are added from V_0 to any other variable V' on this path. Note that V_0 and V refers to their position (center or non-center) in M, i.e., they may represent different variables in M'' and M_q.

Suppose M_s is the first model following the sequence where this edge is added, and M_{q-1} is the last model following the sequence where this edge is present. Denote the center of the ancestral ordered group that changes between M_{s-1} and M_s as V_{i_s}, and between M_{q-1} and M_q as V_{i_q}. Since the edge from V_0 to V is added in M_s, it must belong to one of the following three cases: There is a causal path $V_0 \to V_{i_s} \to V$, or there exists a latent confounder with $V_0 \to V_{i_s} \leftarrow H_{l_s} \to V$ (note that V_0 may be V_{i_s}) or $V_0 \to Z_{j_s} \leftarrow V_{i_s} \leftarrow H_{l_s} \to V$ in M_{q-1}. Note that in the latter two cases, since Z_{j_s}, V_{i_s}, H_{l_s} all belong to the same ancestral ordered group, V_0 is an ancestor of V_{i_s}, and V_{i_s} is an ancestor of V.

We consider the following cases:

- $V = Z_{j_q}$. This step does not include edge removals.
- $V = V_{i_q}$. Since the edge is removed in M_q, $V_0 \in Pa_{M_{q-1}}(Z_{j_q})$ in M_{q-1}, where V_0 is not a parent of V_{i_q} in M, and is not a parent of Z_{j_q} in M''. We note that the edge from V_0 to Z_{j_q} cannot be an added edge. Because this edge is not removed in M_q, and will never be removed for all $M_{q+1}, \cdots, M_p = M''$. Therefore this edge is in M.
  Consider Z_{j_q}, and consider J as the set of (unobserved or cogent) variable V' in M where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_{M''}(Z_{j_q}) \cap \mathcal{V}^C)$. In other

words, $J = An_{M''}(Z_{j_q}) \cap \mathcal{V}^C$ if V_0 is a cogent variable, and J additionally includes V_0 if it is unobserved. Further, consider $K = Pa_{M''}(Z_{j_q}) \cap \mathcal{V}^C$. It follows that $K \subseteq PP_M(V_{i_q})$, and $J \subseteq An_M(Z_{j_q}) \setminus \{V_{i_q}\}$.

We have $Rank(\mathbf{W}^J_{K \cup \{Z_{j_q}\}}) = |K|$. However, note that we can find K distinct paths from variables in J to variables in K. Specifically, for each variable V' in K, define $f(V')$ as the variable in J whose exogenous noise in M is the exogenous noise of V' in M''. Note that $f(V') = V'$ if there is no change on V' and $N_{V'}$ on both models. Further, $f(V')$ and V' must belong to the same ancestral ordered group and hence there is a causal path $f(V') \rightsquigarrow V'$ in M. Therefore, a minimal bottleneck from J to $K \cup \{Z_{j_q}\}$ must at least include K variables. However, there is still one path $V_0 \to Z_{j_q}$ that is not blocked, as V_0 is not a parent of Z_{j_q} in M'' (meaning $V_0 \notin K$). Hence M violates Assumption 3.

- $V = V_{k_i}$ for some V_{k_i}. Note that this variable V_{k_i} may have different positions (center or non-center) in M_q and M''. For simplicity of notation, denote this variable as V_{k_q}. Then there exist paths $V_0 \to V_{i_q} \to V_{k_q}$, $V_0 \to V_{k_q}$, and $V_0 \to Z_{j_q}$ in M_{q-1}. Note that the edge $V_{i_q} \to V_{k_q}$ and $V_0 \to Z_{j_q}$ are in M.

  If V_{k_q} does not change the position. Consider J as the set of (unobserved or cogent) variable V' in M where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_{M''}(V_{k_q}) \cap \mathcal{V}^C)$, and $K = Pa_{M''}(V_{k_q}) \cap \mathcal{V}^C$. It follows that $K \subseteq PP_M(V_{k_q})$, and $J \subseteq An_M(V_{k_q})$. Further, $Rank(\mathbf{W}^J_{K \cup \{V_{k_q}\}}) = |K|$. However, a minimal bottleneck from J to $K \cup \{V_{k_q}\}$ must at least include: One variable between V' and $f(V')$ for each $V' \in K$, except for Z_{j_q}; One variable on the edge $V_{i_q} \to V_{k_q}$; One variable on the edge $V_0 \to Z_{j_q}$. Therefore the size of the minimal bottleneck is at least $|K| + 1$. Hence M violates Assumption 3.

  If V_{k_q} changes the position from non-center to center. That is, it can be denoted by Z_{j_r} for some $q < r \leq p$. Note that the reason we consider M'' in the above analysis is because there may still be edge additions or removals on V_{k_q} after M_q. However, if $V = Z_{j_r}$, then we know that there will be no edge additions/removals on V after M_r. Therefore, consider J as the set of variable V' in M_r where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_{M_r}(Z_{j_r}) \cap \mathcal{V}^C)$, and $K = Pa_{M_r}(Z_{j_r}) \cap \mathcal{V}^C$. This implies that $J \subseteq An(Z_{j_r}) \setminus \{V_{i_r}\}$, and $K \subseteq PP(V_{i_r})$. We have $Rank(\mathbf{W}^J_{K \cup \{Z_{j_r}\}}) = |K|$. However, a minimal bottleneck from J to $K \cup \{V_{k_q}\}$ must at least include: One variable between V' and $f(V')$ for each $V' \in K$, except for Z_{j_q}; One variable on the edge $V_{i_q} \to V_{k_q}$; One variable on the edge $V_0 \to Z_{j_q}$. Therefore the size of the minimal bottleneck is at least $|K| + 1$. Hence M violates Assumption 3.

  If V_{k_q} changes the position from center to non-center. That is, it can be denoted by V_{i_r} for some $q < r \leq p$. Similarly, there are no edge additions or removals involving V_{i_r} after M_r. In this case, we consider the variable V_{i_r} in M_{r-1}, i.e., J as the set of variable V' in M_r where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_{M_{r-1}}(V_{i_r}) \cap \mathcal{V}^C)$, and $K = Pa_{M_{r-1}}(V_{i_r}) \cap \mathcal{V}^C$. The following is the same as when V_{k_q} does not change the position, and we can conclude that M violates Assumption 3.

- $V = V_{k_l}$ for some V_{k_l}. Similarly, for notation simplicity, denote this variable as V_{l_q}. Note that this is different from H_{l_q}, which refers to the unobserved variable that is in the same group as V_{i_q}. Then there exist paths $V_0 \to V_{i_q} \leftarrow H_{l_q} \to V_{l_q}$ or $V_0 \to Z_{j_q} \leftarrow V_{i_q} \leftarrow H_{l_q} \to V_{l_q}$ in M_{q-1}.

  If V_{l_q} does not change the position. Consider J as the set of variable V' in M where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_{M''}(V_{l_q}) \cap \mathcal{V}^C)$, and $K = Pa_{M''}(V_{l_q}) \cap \mathcal{V}^C$. It follows that $K \subseteq PP_M(V_{l_q})$, and $J \subseteq An_M(V_{l_q})$, and $Rank(\mathbf{W}^J_{K \cup \{V_{l_q}\}}) = |K|$.

  - If V_{i_q} (Z_{j_q} if the center is replaced) is not a parent of V_{l_q} in M''. Note that this includes the case when $V_0 = V_{i_q}$. Similar to the above analysis for V_{k_i}, for each variable V' in K, define $f(V')$ as the variable in J whose exogenous noise in M is the exogenous noise of V' in M''. Then any bottleneck from J to K must include at least one variable between V' and $f(V')$ for each $V' \in K$. Further, it must also include one variable on the edge $H_{l_q} \to V_{l_q}$. Therefore the size of the minimal bottleneck must be at least $|K| + 1$.

   - If V_{i_q} (Z_{j_q} if the center is replaced) is a parent of V_{l_q} in M. Then any bottleneck from J to K must include at least one variable between V' and $f(V')$ for each $V' \in K$ except for V_{i_q}. Further, it must also include one variable on the edge $H_{l_q} \to V_{l_q}$, and one variable on the edge $V_0 \to V_{i_q}$ (or $V_0 \to Z_{j_q}$ if Z_j is the center node in M''). Therefore the size of the minimal bottleneck must also be at least $|K| + 1$.

Therefore, in both cases, M violates Assumption 3.

If V_{l_q} changes its position. Similar to the analysis described in V_{k_i}, V_{l_q} can be denoted by Z_{j_r} or V_{i_r} for some $q < r \leq p$. Since there are no edge additions or removals involving V_{i_r} after M_r, we can consider the sets J and K defined on M_{r-1} if $V = V_{i_r}$, and the sets J and K defined on M_r if $V = Z_{j_r}$. Following the same analysis as above, we can conclude that M violates Assumption 3.

In conclusion, we show that if all added edges in $M_1, \cdots, M_p$ are removed, then M violates Assumption 3. Therefore at least one of the added edge to M'' is not removed.

Proof of (2.f). Lastly, we show that because of this added edge, M'' violates Assumption 3. Suppose the edge from V_0 to V is added in M''. Note that this edge does not introduce any additional ancestral relations as M'' and M belong to the same AOG equivalence class. Therefore for each center or non-center variable V, the ancestor set and possible parent set of V remain the same. The proof of this claim is actually the same as (2.b), but in a reversed manner.

- $V = Z_j$. Then V is a center node in M'', and $V_0 \in Pa_M(Z_j) \setminus Pa_M(V_i)$. Consider V_i in M'', which is an mleaf. Consider J as the set of (unobserved or cogent) variable V' in M'' where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_M(V_i) \cap \mathcal{V}^C)$, and consider $K = Pa_M(X_i) \cap \mathcal{V}^C$. We have $J \subseteq An_{M''}(V_i \setminus \{Z_j\})$, and $K \subseteq PP_{M''}(V_j)$. Further, according to the structural equation of V_i in M, $Rank(\mathbf{W}^J_{K \cup \{V_i\}}) = |K|$.

  For each variable V' in K, define $f(V')$ as the variable in J whose exogenous noise in M'' is the exogenous noise of V' in M. Then the minimal bottleneck must include at least one variable between V' and $f(V')$ for each $V' \in K$, and one variable on the path $V_0 \to Z_j \to V_i$ on M''. Therefore the size of the minimal bottleneck must be at least $|K| + 1$.

- $V = V_{k_i}$. Then $V_0 \in Pa_M(Z_j) \setminus Pa_M(V_{k_i})$. Without loss of generality, suppose V_{k_i} does not change its position between M and M''. If it changes then we can use the same procedure as described in (2.e).

  Consider J as the set of variable V' in M'' where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_M(V_{k_i}) \cap \mathcal{V}^C)$, and consider $K = Pa_M(V_{k_i}) \cap \mathcal{V}^C$. We have $Rank(\mathbf{W}^J_{K \cup \{V_{k_i}\}}) = |K|$. However, any bottleneck from J to K includes one variable between V' and $f(V')$, for all $V' \in K$. Additionally, it needs to cover the edge $V_0 \to V$. Therefore the size of the minimal bottleneck must be at least $|K| + 1$.

- $V = V_{k_l}$. Similarly, we assume that V_{k_i} does not change its position between M and M''. Consider J as the set of variable V' in M'' where $N_{V'}$ corresponds to the exogenous noise of a variable in $\{V_0\} \cup (An_M(V_{k_l}) \cap \mathcal{V}^C)$, and consider $K = Pa_M(V_{k_l}) \cap \mathcal{V}^C$. We have $Rank(\mathbf{W}^J_{K \cup \{V_{k_l}\}}) = |K|$. However, any bottleneck from J to K includes one variable between V' and $f(V')$, for all $V' \in K$. Additionally, it needs to cover the edge $V_0 \to V$. Therefore the size of the minimal bottleneck must be at least $|K| + 1$.