Causal Discovery in Linear Latent Variable Models Subject to Measurement Error

Anonymous Author(s)
Affiliation
Address
email

Abstract

We focus on causal discovery in the presence of measurement error in linear systems where the mixing matrix, i.e., the matrix indicating the independent exogenous noise terms pertaining to the observed variables, is identified up to permutation and scaling of the columns. We demonstrate a somewhat surprising connection between this problem and causal discovery in the presence of unobserved parentless causes, in the sense that there is a mapping, given by the mixing matrix, between the underlying models to be inferred in these problems. Consequently, any identifiability result based on the mixing matrix for one model translates to an identifiability result for the other model. We characterize to what extent the causal models can be identified under a two-part faithfulness assumption. Under only the first part of the assumption (corresponding to the conventional definition of faithfulness), the structure can be learned up to the causal ordering among an ordered grouping of the variables but not all the edges across the groups can be identified. We further show that if both parts of the faithfulness assumption are imposed, the structure can be learned up to a more refined ordered grouping. As a result of this refinement, for the latent variable model with unobserved parentless causes, the structure can be identified. Based on our theoretical results, we propose causal structure learning methods for both models, and evaluate their performance on synthetic data.

1 Introduction

Learning causal structure among the variables of a system from observational data has received considerable attention in the literature, since subject matter knowledge on causal relationships is often incomplete or impossible to obtain in many applications [20, 14]. In many real-life systems, not all variables of interest or all direct common causes of those variables can be observed. This necessitates approaches for structure learning capable of dealing with latent variables. For the case that no a-priori restrictions on the functional form of causal mechanisms are imposed, constraint-based algorithms such as the Fast Causal Inference (FCI) algorithm have been proposed [20]. Such approaches are often unable to identify the direction of the majority of causal connections in the system, which motivates placing additional, often parametric, restrictions on the generating model. One of the most commonly used restrictions is to assume that the model is linear with non-Gaussian exogenous noise terms [18, 19, 10, 8]. This generating model leads to identification of the existence and orientation of all edges in a causal model, assuming that no latent variables that are common causes of two or more variables exist [18, 19]. Moreover, in such models even when latent variable common causes exist, it is still possible to orient additional edges compared to algorithms such as FCI [10, 17, 1].

In certain application, although some of the variables of interest (called underlying variables) are latent, we may have access to noisy measurements of them [15, 11, 22, 16, 9]. In this case, conditional independence patterns among observed variables (that are noisy measurements of underlying variables) are different from those among the underlying (measurement-error-free) variables. In general, constraint-based approaches to causal discovery are not able to correct for the difference in

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

observed independence due to the presence of measurement error, and thus will produce erroneous edge adjacencies and orientations. Similarly, naive application of methods based on non-Gaussian exogenous noise assumption may also fail in recovering the correct causal relations in the presence of measurement error. This is due to the fact that the extra measurement noise terms break the asymmetry utilized by those methods.

In this paper, we bring additional insight into the problem of causal discovery under measurement error by showing a surprising connection between linear structural equation models (SEMs) where variables are measured with error, and linear SEMs with hidden variables. In particular, we consider two linear SEMs. In the first model, the unobserved variables are in fact of interest and can appear in any part of the causal structure of the system, but we assume that noisy measurements of these unobserved variables are available. We refer to this model as the linear SEM with measurement error (SEM-ME). The second model does not require measurements of the latent variables, as they are not of primary interest, yet we assume that all latent variables are root variables, meaning that they are not causally affected by any other latent or observed variables in the system. We refer to this model as the linear SEM with unobserved roots (SEM-UR). It is noteworthy that assuming latent variables are limited to be root variables does not affect the estimated total causal effects among the observed variables (see Remark 1 in Section 2.2), and provides a representation of a latent variable model that is consistent with these casual effects. For this reason, this restriction is common in the literature of causal inference [6, 7, 10, 3].

We study the identification of linear SEM-MEs and SEM-URs in a setup where the independent exogenous noise terms that causally (directly or indirectly) affect each observed variable can be identified. That is, the mixing matrix of the linear system that transforms exogenous noise terms to observed variables is identified up to permutation and scaling of the columns. This can be satisfied, for example, if all independent exogenous noise terms are non-Gaussian. Our first main contribution is presenting a mapping between the linear SEM-MEs and SEM-URs, which demonstrates a correspondence between a weighted causal diagram[1] generated by SEM-UR and a set of weighted causal diagrams from the SEM-ME (Theorem 1). The models in this correspondence all possess the same mixing matrix. Consequently, any identifiability result based on the mixing matrix for one model is applicable to the other model. This allows us to study the problem of causal discovery in these two models together.

Additionally, we study the identifiability of linear SEM-MEs and SEM-URs, and show how it benefits from a two-part faithfulness assumption. The first part prevents zero total causal effect of a variable on its descendants and the second part prevents cancellation or proportionality among specific edges. Together, these two prevent parameter couplings in the model. Our second main contribution is to characterize the extent of identifiability of the causal model under our faithfulness assumption. We demonstrate that if only the first part of the faithfulness assumption is imposed, the model can be learned up to an equivalence class characterized by an ordered grouping of the variables which we call the *ancestral ordered grouping* (AOG), but not all the edges across the groups, or the group centers (defined in Section 4.3), are identifiable (Theorem 2). This characterization is a refined version of the original ordered grouping proposed in [22]. We further show that if both parts of the faithfulness assumption are imposed, the model can be learned up to an equivalence class characterized by a more refined ordered grouping which we call *direct ordered grouping* (DOG) (Theorem 3). This characterization only leaves the center of each group in the SEM-ME and the noise term corresponding to the center of each group in the SEM-UR unidentified. Importantly, for SEM-UR, the difference in the equivalent models is only in the location of the exogenous noise terms. This means that the ground-truth structure is uniquely identifiable (Corollary 3). Based on our identification results, we propose causal structure learning methods for both SEM-ME and SEM-UR, and evaluate their performance on synthetic data.

2 Model description

In this section, we first introduce the two models considered in this paper, namely linear SEM-ME and SEM-UR. We then discuss a model assumption needed for our identifiability results.

2.1 Linear structural equation model with measurement error (SEM-ME)

Definition 1 (Linear SEM-ME) *A linear SEM-ME consists of a set of "underlying" variables which can be partitioned into unobserved underlying variables* $\mathcal{Z} = \{Z_1, ..., Z_{p_z}\}$ *and observed*

[1] A weighted causal diagram is the directed graph corresponding to the linear model which associates the coefficient of variable X_1 in the structural equation for X_2 to the edge from X_1 to X_2.

underlying variables $\mathcal{Y} = \{Y_1, ..., Y_{p-p_z}\}$. In addition, we have another set of observed variables $\mathcal{U} = \{U_1, \cdots, U_{p_z}\}$ corresponding to the noisy measurements of $\mathcal{Z}$. The underlying variables in $\mathcal{Z} \cup \mathcal{Y}$ can be arranged in a total (causal) order (such that no latter variable in the order can cause any earlier variable), and variables $V_i \in \mathcal{Z} \cup \mathcal{Y}$ and $U_i \in \mathcal{U}$ are generated as follows:

$$V_i = \sum_{j:\, V_j \in Pa(V_i)} c_{ij} V_j + N_{V_i} \quad i \in [p]; \qquad U_i = Z_i + N_{U_i}, \quad i \in [p_z], \tag{1}$$

where $[n] := \{1, 2, \cdots, n\}$, $Pa(V_i) \subseteq \mathcal{Z} \cup \mathcal{Y}$ denotes direct causes of V_i, and N_{V_i}, N_{U_i} are the exogenous noise terms corresponding to V_i and U_i, respectively.

We define the weighted causal diagram of the linear SEM-ME as a weighted directed graph where the nodes are the variables in $\mathcal{Z} \cup \mathcal{Y}$. There is a directed edge (causal connection) from V_i to V_j with weight c_{ji} if and only if $c_{ji} \neq 0$, for $V_i, V_j \in \mathcal{Z} \cup \mathcal{Y}$. Because of our causal order assumption, the causal diagram will be acyclic. We use the terms node and variable interchangeably. The linear SEM-ME in this work is a generalization of the linear causal model with measurement error proposed in [22], in the sense that some of the underlying variables can be observed.

We define an *unobserved leaf node* (`u-leaf` node) as an unobserved underlying variable in $\mathcal{Z}$ that does not have any children in $\mathcal{Z} \cup \mathcal{Y}$. The rest of the underlying variables are referred to as *non u-leaf nodes* (`nu-leaf` nodes): These include unobserved non-leaf underlying variables, which have children in $\mathcal{Z} \cup \mathcal{Y}$, and observed underlying variables in $\mathcal{Y}$. Similar to the argument in [22], given observed data generated by the linear SEM-ME in Equation (1), for a `u-leaf` node Z_i, its exogenous noise term N_{Z_i} is not distinguishable from its measurement error N_{U_i}. This follows because Z_i is not observed, and N_{Z_i} only influences one observed variable which is the noisy measurement U_i. Therefore, we restrict our focus to the following canonical form.

Definition 2 (Canonical form of a linear SEM-ME) *The canonical form of a linear SEM-ME is the one in which all* `u-leaf` *nodes have no exogenous noise terms. Specifically,*

$$\begin{bmatrix} Z^L \\ Z^{NL} \\ Y \end{bmatrix} = \begin{bmatrix} \mathbf{D} \\ \mathbf{C}_Z \\ \mathbf{C}_Y \end{bmatrix} Z^{NU} + \begin{bmatrix} 0 \\ N_{Z^{NL}} \\ N_Y \end{bmatrix}, \qquad U = \begin{bmatrix} Z^L \\ Z^{NL} \end{bmatrix} + \begin{bmatrix} N_{Z^L} + N_{U^L} \\ N_{U^{NL}} \end{bmatrix}, \tag{2}$$

where Z^{NL}, Z^L, Y are the vectors of unobserved underlying non-leaf variables, `u-leaf` *nodes, and observed underlying variables, respectively. $N_{Z^{NL}}$, N_{Z^L}, and N_Y are the corresponding noise vectors. $Z^{NU} = [Z^{NL};\ Y]$ is the vector of* `nu-leaf` *nodes. U is the vector of noisy measurements and $N_{U^{NL}}$ and N_{U^L} denote the noise vectors corresponding to the measurements of Z^{NL} and Z^L, respectively. $\mathbf{C}$ represents the causal connections among* `nu-leaf` *nodes, and can be partitioned into $[\mathbf{C}_Z;\ \mathbf{C}_Y]$ corresponding to $[Z^{NL};\ Y]$, respectively. $\mathbf{D}$ represents the causal connections from* `nu-leaf` *nodes to* `u-leaf` *nodes. Note that $\mathbf{C}$ can be arranged into a strictly lower triangular matrix via simultaneous column and row permutations (due to the acyclicity assumption).*

2.2 Linear structural equation model with unobserved roots (SEM-UR)

Definition 3 (Linear SEM-UR) *A linear SEM-UR consists of a set of observed variables $\mathcal{X} = \{X_1, \cdots, X_q\}$ and another set of latent variables $\mathcal{H} = \{H_1, \cdots, H_l\}$. All latent variables are assumed to be root variables (have no direct causes). The observed variables are arranged in a causal order, and each observed variable is directly influenced by a linear combination of other observed and latent variables, plus an independent exogenous noise term.*

$$H_i = N_{H_i}, \quad i \in [l]; \qquad X_j = \sum\nolimits_{i \in [l]} b_{ji} H_i + \sum\nolimits_{k<j} a_{jk} X_k + N_{X_j}, \qquad j \in [q]. \tag{3}$$

We define the weighted causal diagram of the linear SEM-UR as a directed graph where the nodes are the observed and latent variables in $\mathcal{X} \cup \mathcal{H}$. There is a causal connection from X_i (resp. H_i) to X_j with weight a_{ji} (resp. b_{ji}) if and only if a_{ji} (resp. b_{ji}) $\neq 0$. Because of our causal order assumption, the causal diagram will be acyclic. Note that unlike in [10] and [17], we consider a broader model which allows latent variables to have a single observed child. However, unlike the work [1], we restrict all latent variables to roots; see Remark 1.

A linear SEM-UR can be written in the following matrix from:

$$H = N_H; \qquad X = \mathbf{B}H + \mathbf{A}X + N_X, \tag{4}$$

where $X = [X_1\ \cdots\ X_q]^\top$ and $H = [H_1\ \cdots\ H_l]^\top$ are the vectors of observed and latent variables, respectively. $N_X = [N_{X_1}\ \cdots\ N_{X_q}]^\top$, $N_H = [N_{H_1}\ \cdots\ N_{H_l}]^\top$ represent the vectors of independent

noise terms associated with X, H, respectively. $\mathbf{B}$ represents the causal connections from latent to observed variables, and $\mathbf{A}$ represents the causal connections among observed variables, which can be assumed to be a strictly lower-triangular matrix (due to the acyclicity assumption).

Remark 1 *In our definition of linear SEM-UR, we require that all the latent variables are parentless. This assumption is common in the literature of causal inference, as it does not restrict the feasible total causal effects (i.e., sum of products of path coefficients) among the observed variables [6, 7, 3]. To explain this, suppose we start from a general linear latent variable model in which latent variables are allowed to have parents. Hoyer et al. [10] proposed an algorithm which maps such general model to one in which latent variables are parentless. They showed that the resulting model is observationally and causally equivalent to the original model, i.e., the joint distributions of observed variables are identical, and all causal effects of observed variables on other observed variables are identical in the models before and after the mapping.*

2.3 Separability assumption

For both models in Sections 2.1 and 2.2, we assume that the noise components can be separated from observations (i.e., the mixing matrices are recoverable). We first describe the mixing process for each of the two models, and then state our separability assumption.

From Equation (2), we can write all underlying variables in the linear SEM-ME in terms of the independent noise terms, as follows.

$$\begin{bmatrix} Z^L \\ Z^{NL} \\ Y \end{bmatrix} = \mathbf{W}^{ME} \begin{bmatrix} N_{Z^{NL}} \\ N_Y \end{bmatrix}, \qquad \text{where} \quad \mathbf{W}^{ME} = \begin{bmatrix} \mathbf{D}(\mathbf{I}-\mathbf{C})^{-1} \\ (\mathbf{I}-\mathbf{C})^{-1}_Z \\ (\mathbf{I}-\mathbf{C})^{-1}_Y \end{bmatrix}. \tag{5}$$

$\mathbf{I}$ is the identity matrix, and $(\mathbf{I}-\mathbf{C})^{-1}_Z$ and $(\mathbf{I}-\mathbf{C})^{-1}_Y$ represent the rows of $(\mathbf{I}-\mathbf{C})^{-1}$ corresponding to Z^{NL} and Y, respectively. Using the relation between U and Z in Equation (2), the two types of observed variables, i.e., measurements U and observed underlying variables Y, can be written as

$$\begin{bmatrix} U \\ Y \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{W}^{ME} & \mathbf{I} \\ & \mathbf{0} \end{bmatrix}}_{\mathbf{W}} \begin{bmatrix} N_{Z^{NL}} \\ N_Y \\ N_{Z^L} + N_{U^L} \\ N_{U^{NL}} \end{bmatrix}. \tag{6}$$

We refer to $\mathbf{W}$ as the overall mixing matrix of the system. Given any column permuted and rescaled version of $\mathbf{W}$, we can recover matrix $\mathbf{W}^{ME}$ by removing the submatrix $[\mathbf{I} \quad \mathbf{0}]^\top$ of size $p \times p_z$ corresponding to U. Recall that p_z is the cardinality of Z (which is the same as the cardinality of U).

Similarly, according to Equation (4), observed variables in the linear SEM-UR can be written in terms of the independent noise terms as

$$X = \mathbf{W}^{UR} \begin{bmatrix} N_H \\ N_X \end{bmatrix}, \qquad \text{where} \quad \mathbf{W}^{UR} = \left[(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} \quad (\mathbf{I}-\mathbf{A})^{-1}\right]. \tag{7}$$

Assumption 1 (Separability) *The linear SEM-ME (resp. linear SEM-UR) is separable, that is, the mixing matrix $\mathbf{W}^{ME}$ in Equation* (5) *(resp. $\mathbf{W}^{UR}$ in Equation* (7)*) can be recovered from observations of $[U;\ Y]$ (resp. X) up to permutation and scaling of its columns.*

Separability assumption states that for every observed mixture, the independent exogenous noise terms pertaining to this mixture 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 [5, Theorem 1] (SEM-ME always does), 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.

3 Mapping between the weighted graphs of the models for identifiability

A key observation in this work is that there exists a mapping between the weighted causal diagrams of the introduced SEM-ME and SEM-UR, which leads to the corresponding mixing matrices $\mathbf{W}^{ME}$ and $\mathbf{W}^{UR}$ being transposes of one another.

Define the mapping φ from the set of weighted causal diagrams of linear SEM-URs, denoted by $\mathcal{M}^{UR}$, to the set of weighted causal diagrams of canonical linear SEM-MEs, denoted by $\mathcal{M}^{ME}$, where for each diagram $M \in \mathcal{M}^{UR}$, $\varphi(M) \subset \mathcal{M}^{ME}$ is constructed as follows: (a) Replace each latent variable in M with a Z variable. (b) Replace each observed root variable in M with a Y variable. (c) Replace the rest of the observed variables in M with either Z or Y variables. (d) Reverse all the edges in M. Note that $\varphi(M)$ encompasses a set of models, corresponding to the possible choices in step (c).

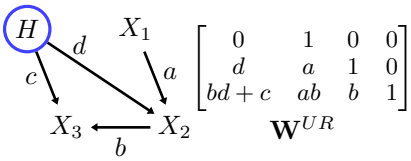
Similarly define the mapping ϕ from the set $\mathcal{M}^{ME}$, to the set $\mathcal{M}^{UR}$, where for each diagram $M' \in \mathcal{M}^{ME}$, $\phi(M') \in \mathcal{M}^{UR}$ is constructed as follows. (a) Replace each `u-leaf` variable in M' with a latent variable. (b) Replace the rest of the variables in M' with observed variables. (c) Reverse all the edges in M'. Note that from the definition of the two mappings, for any given $M \in \mathcal{M}^{UR}$, $M = \phi(\varphi(M))$. For any given $M' \in \mathcal{M}^{ME}$, $M' \in \varphi(\phi(M'))$.

Theorem 1 *Let $M \in \mathcal{M}^{UR}$ with mixing matrix $\mathbf{W}^{UR}(M)$ and M' be a corresponding element in $\mathcal{M}^{ME}$ with mixing matrix $\mathbf{W}^{ME}(M')$, where $M' \in \varphi(M)$ and $M = \phi(M')$. Then there exist a permutation of the columns of $\mathbf{W}^{UR}(M)$, denoted by $\tilde{\mathbf{W}}^{UR}(M)$, and a permutation of the columns of $\mathbf{W}^{ME}(M')$, denoted by $\tilde{\mathbf{W}}^{ME}(M')$, which satisfy $\tilde{\mathbf{W}}^{ME}(M') = (\tilde{\mathbf{W}}^{UR}(M))^\top$.*

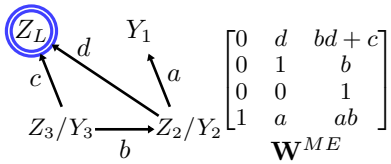
The proofs of all the results are provided in the Appendix. Note that as it can be seen from the proof of Theorem 1, whether the `nu-leaf` variables, that are non-leaf in the SEM-ME, are observed or not does not change the corresponding mixing matrix.

Remark 2 *The mapping introduced here is between a weighted diagram in SEM-UR and a set of weighted diagrams in SEM-ME. The elements of the set in fact correspond to different settings of the non-leaf underlying variables being observed or not. Note that in most applications, it is clear to the researcher whether an underlying variable is observed or measured with error. Therefore, only one element of the set is relevant to the application and the set reduces to a single element. Consequently, in our identifiability results, without loss of generality, we assume the set is a singleton.*

Theorem 1 implies that under separability assumption, any identifiability result based on the mixing matrix for one model are applicable to the other model, by reversing all edges. This allows us to study the problem of causal discovery of these two models simultaneously.



(a) Linear SEM-UR



(b) Linear SEM-ME

Figure 1: Example explaining the mapping.

Example 1 *Consider the linear SEM-UR in Figure 1(a) comprised of three observed and one latent variable (denoted by a single circle). The corresponding linear SEM-ME under the mapping φ is shown in Figure 1(b). Specifically, (i) the latent variable H is mapped to the* `u-leaf` *variable Z_L (denoted by a double circle); (ii) the observed root X_1 is mapped to the observed underlying variable Y_1; and (iii) observed variables X_2 and X_3 could be mapped to either observed or unobserved underlying variables; hence the UR graph will be mapped to a set of 4 graphs in the ME model. The mixing matrices for both models are shown in Figure 1. It can be readily seen that there exists a permutation of the columns of $\mathbf{W}^{UR}$, which is equal to the transpose of $\mathbf{W}^{ME}$.*

4 Identifiability of models

In this section, we study the extent of identifiability of linear SEM-ME and SEM-UR under separability assumption. We consider both a weaker notion of structure identification, where only the graph structure of the model is obtained, and a stronger notion of model identification where, in addition to the graph structure, model parameters are recovered too. Recall from Theorem 1 that identifying one model is equivalent to identifying the other. Therefore, we will mainly explain the results on the recovery of linear SEM-ME. We first present SEM-ME and SEM-UR faithfulness assumptions, which are extensions of the conventional faithfulness in linear causal models. We discuss model identifiability under conventional faithfulness assumptions, which can be characterized by *ancestral ordered grouping* (AOG), explained in Section 4.2, and show the limitation of these results. Under our proposed faithfulness assumptions, the extent of identifiability can be characterized by *direct*

ordered grouping (DOG), explained in Section 4.3. This leads to a subsequent result on structure identifiability.

4.1 Faithfulness assumption

Assumption 2 (SEM-ME faithfulness) *(a) The total causal effect of an underlying variable $V_i \in \mathcal{Z} \cup \mathcal{Y}$ on its descendant $V_j \in \mathcal{Z} \cup \mathcal{Y}$ is not equal to zero. (b) Denote $\mathbf{W}_0$ as any $m \times m$ submatrix of $\mathbf{W}^{ME}$ involving underlying variables $V_{i_1}, \cdots, V_{i_m}$ and noise terms $N_{V_{j_1}}, \cdots, N_{V_{j_m}}$, such that $V_{j_{[m]}} = \{V_{j_1} \cdots, V_{j_m}\}$ are ancestors of V_{i_m}, and the support of the rows in $\mathbf{W}^{ME}$ corresponding to $V_{i_{[m-1]}} = \{V_{i_1}, \cdots, V_{i_{m-1}}\}$ are subsets of that of V_{i_m}; $m \geq 2$. Then $\mathbf{W}_0$ is not full-rank if and only if $V_{i_{[m-1]}}$ are ancestors of V_{i_m}, and every causal path from $V_{j_{[m]}}$ to V_{i_m} passes through $V_{i_{[m-1]}}$.*

Assumption 3 (SEM-UR faithfulness) *(a) The total causal effect of an observed (latent) variable $X_i \in \mathcal{X}$ ($H_i \in \mathcal{H}$) on its descendant $X_j \in \mathcal{X}$ is not equal to zero. (b) Denote $\mathbf{W}_0$ as any $m \times m$ submatrix of $\mathbf{W}^{UR}$ involving observed variables $X_{i_1}, \cdots, X_{i_m}$, and noise terms $N_{j_{[m]}} = \{N_{j_1}, \cdots, N_{j_m}\}$, where variables in $X_{i_{[m-1]}} = \{X_{i_1}, \cdots, X_{i_{m-1}}\}$ and variables corresponding to $N_{j_{[m]}}$ (observed or latent) are all ancestors of X_{i_m}; $m \geq 2$. Then $\mathbf{W}_0$ is not full-rank if and only if every causal path from variables corresponding to $N_{j_{[m]}}$ to X_{i_m} passes through $X_{i_{[m-1]}}$.*

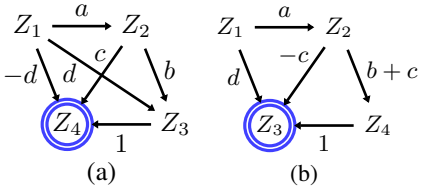
Assumption 2*(a)* and 3*(a)* are the conventional faithfulness assumption in linear causal models. They require that when multiple causal paths exist from any (observed or unobserved) variable to its descendants, their combined effect (i.e., sum of path products) is not equal to zero, see [20]. These assumptions are equivalent to requiring that all variables to be marginally dependent on their descendants. Assumption 2*(b)* and 3*(b)* prevent certain parameter cancellation or proportionality. We note that Assumption 2*(b)* holds if certain conditional dependency conditions are satisfied (see Appendix for more details). This implies that Assumption 2*(b)* is weaker than the non-deterministic faithfulness in [22]. Assumptions 2 and 3 are satisfied with probability one if all model coefficients are drawn randomly and independently from continuous distributions [13]. As an illustration of violation of Assumption 2*(b)*, consider the linear SEM-ME generated from the graph in Figure 2(a). The mixing matrix is given in the figure. This model satisfies Assumption 2*(a)*. However, the 2×2 submatrix $[a \;\; 1; a(b+c) \;\; b+c]$ is of rank 1, but not all causal paths from Z_1 to Z_4 passes through Z_2. This is caused by the parameter cancellation on the triangle (Z_1, Z_3, Z_4). Therefore, the model violates Assumption 2*(b)*. Furthermore, this path cancellation introduces additional conditional independencies in the model: We can write Z_4 as $(b+c)Z_2 + N_{Z_3}$, implying $Z_1 \perp\!\!\!\perp Z_4 | Z_2$.

Remark 3 *Assumption 3 (SEM-ME faithfulness) is equivalent to bottleneck faithfulness in [1] and the counterpart for the SEM-ME (Assumption 2) is an extension of bottleneck faithfulness, where we consider the support corresponding to $V_{i_1}, \cdots, V_{i_{m-1}}$ in $\mathbf{W}^{ME}$ being a subset of that corresponding to V_{i_m}. In SEM-URs, this is equivalent to $V_{i_1}, \cdots, V_{i_{m-1}}$ being ancestors of V_{i_m}. However, in SEM-MEs this is not necessarily true [21]. See Appendix for a detailed comparison with [1].*

4.2 Ancestral ordered grouping

We first study the extent of identifiability of the models only under conventional faithfulness, i.e., Assumptions 2*(a)* and 3*(a)*.

Definition 4 (Ancestral ordered grouping (AOG)) *The AOG of a SEM-ME (resp. SEM-UR) is a partition of $\mathcal{Z} \cup \mathcal{Y}$ (resp. $\mathcal{H} \cup \mathcal{X}$) into distinct sets. This partition is described as follows: (1) Assign each* `nu-leaf` *node (resp. observed node) to a distinct group. (2) SEM-ME: For each* `u-leaf` *node $Z_j \in \mathcal{Z}$ and each of its parents V_i, if Z_j has no other parents, or all other parents of Z_j are also ancestors of V_i, assign Z_j to the same group as V_i. Otherwise, assign Z_j to a separate group (with no* `nu-leaf` *node). (2) SEM-UR: For each latent variable H_j and each of its children X_i, if H_j has no other children, or all other children of H_j are also descendants of X_i, assign H_j to the same group as X_i. Otherwise, assign H_j to a separate group (with no observed variable).*



$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ ab+d & b & 1 \\ a(b+c) & b+c & 1 \end{bmatrix} \begin{bmatrix} N_{Z_1} \\ N_{Z_2} \\ N_{Z_3} \end{bmatrix}$$

Figure 2: Example explaining AOG and SEM-ME faithfulness assumption.

The AOG characterization is a refined version of the original ordered grouping proposed in [22], where no variables in later groups causes variables in earlier groups. Yet, the grouping introduced in that work is order dependent and hence in general not unique. Moreover, it does not always

characterize the extent of identifiability. Our AOG characterization also uses a similar idea as the learning approach in [17]; however, that work allows for latent variables which have parents and in general does not learn the full structure. Note that the AOG can be recovered based on the support of the mixing matrix under Assumption 2*(a)* (resp. 3*(a)*). Please refer to Appendix for more details.

According to Definition 4, there is at most one `nu-leaf` node in each ancestral ordered group of a SEM-ME. Furthermore, each `u-leaf` node Z_j is assigned to the ancestral ordered group of at most one of its parents: Following the true causal order among `nu-leaf` nodes, only the last parent of Z_j may have the same ancestor set. Hence, if a group has more than one node, then there must be exactly one `nu-leaf` node, and the rest of the nodes are `u-leaf` nodes which are children of this node. This concludes that the structure induced on each ancestral ordered group is a star graph. Similar property holds for SEM-UR. Define the center of the ancestral ordered group as the `nu-leaf` node (resp. the observed node) or the `u-leaf` node (resp. latent node) if the group is of size one. The identifiability of a SEM-ME (resp. SEM-UR) under Assumptions 2*(a)* (resp. 3*(a)*) can be descried by Theorem 2.

Theorem 2 *Under Assumptions 1 and 2(a) (resp. 3(a)), the linear SEM-ME (resp. SEM-UR) can be identified up to a set of models, where the elements of this set all have the same ancestral ordered groups and causal order among the groups, but not necessarily all the edges across the groups. Further, elements in the set can be characterized by the choice of the center of the groups (resp. exogenous noise terms associated to the center of the groups). All the coefficients are identified if and only if the centers (resp. exogenous noise assignments) are identified.*

We refer to the class of models characterized in Theorem 2 as the AOG equivalence class.

Corollary 1 *(i) Denote the AOG of a linear SEM-ME (resp. SEM-UR) as $\{\mathcal{G}^{(g)}\}_{g\in[g_k]}$, where g_k is the number of ancestral ordered groups. The size of the AOG equivalence class, described in Theorem 2, is equal to $\prod_{g\in[g_l]}|\mathcal{G}^{(g)}|$. (ii) Under Assumptions 1 and 2(a) (resp. 3(a)), a linear SEM-ME (resp. SEM-UR) can be uniquely identified if and only if no* `u-leaf` *node (resp. latent variable) has precisely the same ancestors (resp. descendants) as any* `nu-leaf` *node (resp. observed variable).*

As stated in Theorem 2, in general, the edges across the groups cannot all be discovered and the center of the groups (or the exogenous noise term corresponding to it) are not identifiable. This is due to the fact that when only Assumption 2*(a)* holds, certain parameter cancellation or proportionality in the model can occur. This is illustrated in the following example.

Example 2 *Consider the causal graph shown in Figure 2(a). The AOG is $\{\{Z_1\},\{Z_2\},\{Z_3,Z_4\}\}$. Depending on whether Z_3 or Z_4 is the center of the last group, the AOG equivalence class includes the ground-truth and the graph in Figure 2(b). Since both models generate the same mixing matrix, an identification algorithm merely based on the mixing matrix cannot distinguish the two graphs. Note that not only can we not learn the direction of the edge in group $\{Z_3,Z_4\}$, but also the existence of some of the edge across groups such as $Z_1\to Z_4$ cannot be established in the ground-truth because Assumption 2(b) is violated.*

4.3 Identifiability of the model up to equivalence classes

Under Assumptions 1 and 2 (resp. 3), the extent of identifiability of a SEM-ME (resp. SEM-UR) can be characterized by the direct ordered grouping defined in Definition 5 below. We first present two conditions that are used to implement the DOG. We then give the formal definition of DOG. In the following, we use $Ch(V)$ to denote the set of children of variable V in the causal diagram.

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

Similarly, we can translate these conditions to linear SEM-UR.

Condition 2 (SEM-UR edge identifiability) *For a given edge from a latent variable H_l to an observed variable X_i, at least one of the following two conditions is satisfied: (a) $Ch(H_l)\setminus\{X_i\}$ is not a subset of $Ch(X_i)$. That is, there exists another observed child X_j of H_l, which is not a child of X_i. (b) $Ch(H_l)$ is not a subset of $\cap_{X_k\in Pa(X_i)\setminus\{H_l\}}Ch(X_k)$. That is, there exist an (observed or latent) parent X_k of X_i and a child X_j of X_l such that X_j is not a child of X_k.*

Figure 3(a) demonstrates an example of a graph structure which satisfies Assumption 2 while containing an edge which violates Condition 1.

Condition 1 is an equivalent formulation of the conditions for unique identifiability derived in [22]. In that work, Zhang et al. proved that a linear SEM-ME can be uniquely recovered if Condition 1 is satisfied for every edge from an `nu-leaf` node to a `u-leaf` node.[2] They also conjectured that this condition is necessary for identifiability. We prove their conjecture here, although our faithfulness assumption is weaker than theirs. Furthermore, we demonstrate that the conditions can be used to characterize an equivalence class of the models which characterizes the extent of identifiability; the same assertion holds for linear SEM-UR. This characterization is done based on the notion of direct ordered grouping, defined as follows.

Definition 5 (Direct ordered grouping (DOG)) *The DOG of a linear SEM-ME (resp. SEM-UR) is a partition of $\mathcal{Z} \cup \mathcal{Y}$ (resp. $\mathcal{H} \cup \mathcal{X}$) into distinct sets. This partition is described as follows:*
(1) Assign each `nu-leaf` *node (resp. observed node) to a distinct ordered group.*
(2) SEM-ME: For each `u-leaf` *node $Z_j \in \mathcal{Z}$ and each of its parent V_i, if the edge from V_i to Z_j violates Condition 1, assign Z_j to the same ordered group as V_i. Otherwise, assign Z_j to a separate ordered group (with no* `nu-leaf` *node).*
(2) SEM-UR: For each latent variable H_j and each of its child X_i, if the edge from H_j to X_i violates Condition 2, assign H_j to the same ordered group as X_i. Otherwise, assign H_j to a separate ordered group (with no observed variable).

Theorem 3 *Under Assumptions 1 and 2 (resp. 3), the linear SEM-ME (resp. SEM-UR) can be identified up to a set of models where the elements of this set all have the same directed order groups. Further, elements in the set can be characterized by the choice of the center of the groups (resp. exogenous noise terms associated to the center of the groups). All the coefficients are identified if and only if the centers (resp. exogenous noise assignments) are identified.*

We refer to the class of models characterized in Theorem 3 as the DOG equivalence class. Theorem 3 states that we can learn the structure among the groups, but the center of the group (or the noise term corresponding to it) will remain unidentified. Unlike the case of the AOG equivalence class, all the edges across the groups will be identified. That is, the choice of center does not change the edges across groups in DOG equivalence class, while it may change the edge in AOG equivalence class. Furthermore, the size of the direct ordered groups are smaller than the size of the ancestral ordered groups. As an example, for ground-truth in Figure 2(b), all direct ordered group are of size one.

Corollary 2 *(i) The counterpart of Corollary 1(i) is true for the case of DOG as well and we do not repeat it here. (ii) Under Assumptions 1 and 2 (resp. 3), a linear SEM-ME (resp. SEM-UR) can be uniquely identified if and only if for every edge from an* `nu-leaf` *node (resp. latent variable) to a* `u-leaf` *node (resp. observed variable), Condition 1 (resp. 2) is satisfied.*

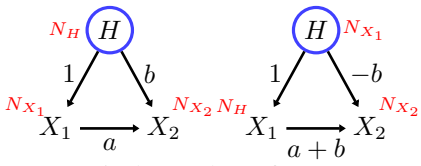
Identifiability of the structure of the model. As seen in Theorem 3, the only undetermined part in the DOG equivalence class pertains to the assignment of the exogenous noises and coefficients, but the structure itself is the same. Consequently, if only the identification of the structure without weights is of interest, Assumptions 1 and 3 are sufficient. See Figure 3(a) for an example of two DOG equivalent SEM-UR models. This identifiability result is summarized in the following corollary.

Corollary 3 *Under Assumptions 1 and 3, the structure of a SEM-UR can be uniquely recovered.*

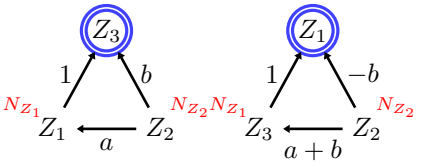
4.4 Recovery algorithm

We note that our faithfulness assumption implies the ground-truth generating model to have strictly fewer edges than any model which does not belong to its DOG equivalence class but has the same mixing matrix. This property can be leveraged to recover the ground-truth structure or a member of its equivalence class. We state this formally in the following.

Proposition 1 *Suppose a SEM-ME (resp. linear SEM-UR) satisfies Assumptions 1 and 2 (resp. 3). Any model that has the same mixing matrix $\mathbf{W}^{ME}$ (resp. $\mathbf{W}^{UR}$) but does not belong to the same DOG equivalence class violates Assumption 2 (resp. 3), and has strictly more edges than any member in the equivalence class.*



(a) Equivalence class of a SEM-UR



(b) Equivalence class of a SEM-ME

Figure 3: DOG Equivalence classes.

[2]Note that in their model, Zhang et al. assumed that all the underlying variables are unobserved.

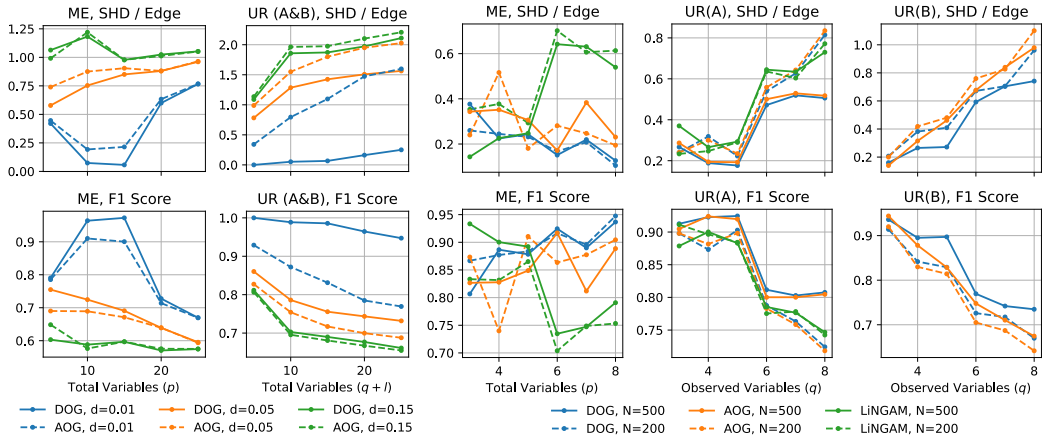


Figure 4: Simulation results of Cases (1) and (2). The x-axis corresponds to the parameters in Equation (1) for SEM-ME and Equation (3) for SEM-UR. In SEM-UR we compare the recovery of both $\mathbf{A}$ and $\mathbf{B}$ in Equation (4). In Case (2), the sample size is `N` times the value on the x-axis.

Equipped with Proposition 1, the steps of our recovery algorithm are as follows: (1) Recover the mixing matrix $\mathbf{W}^{ME}$ (resp. $\mathbf{W}^{UR}$) from observational data. (2) Return the AOG of the true model by comparing the support across different rows/columns in $\mathbf{W}^{ME}$ (resp. $\mathbf{W}^{UR}$). Please refer to the Appendix for details and pseudo-code for recovering AOG. (3) For all possible choices of the center (resp. noise term associated to the center) of the ancestral ordered groups, find a choice that leads to the graph with fewest number of edges in the recovery output.

5 Simulations

We evaluated the performance of our recovery algorithm on randomly generated linear SEM-MEs and SEM-URs with different number of variables. We considered two cases: (1) when a noisy version of the mixing matrix is given, where the noise is Gaussian with different choices of variance denoted by d^2, and (2) when synthetic data comes from a linear generating model with non-Gaussian noises with different sample sizes, and the mixing matrix needs to be estimated. We used Reconstruction ICA algorithm [12] as our overcomplete ICA method. For Case (1), we compared our recovery method (which we refer to as DOG) with an approach based on AOG equivalence class (similar to the learning method in [17]). In addition to the AOG method, for Case (2), we compared our method with ICA-LiNGAM [18]. Note that the ICA-LiNGAM method is not designed for systems with latent variables, but benefits from strong performance for recovering the mixing matrix in causally sufficient settings. The goal of this comparison was to demonstrate the necessity of using methods designed specifically to handle latent variables.

We compared the recovery of the adjacency matrix to the ground-truth, where we used normalized structural Hamming distance (SHD) and F1 score as our performance metrics. Our results for Cases (1) and (2) are shown in the left two and right three columns in Figure 4, respectively. As seen in these figures, our recovery algorithm outperforms recovery based on AOG. In particular, our method recovers the structure in SEM-UR model, and can recover part of the structure in SEM-ME that is shared among DOG equivalence classes. Moreover, both methods outperform ICA-LiNGAM. Please refer to Appendix for additional results and analysis.

6 Conclusion

We considered the problem of causal discovery in linear latent variable models under two settings; (1) in the presence of measurement error, and (2) with unobserved parentless causes. We demonstrated a mapping between these two models that preserves their mixing matrix. Based on this mapping, we derived identifiability results for both models under different faithfulness assumptions, and proposed structure learning algorithms. Similar to other causal discovery methods which are based on the mixing matrix, the performance of our proposed recovery method relies on the accuracy of the utilized mixing matrix estimation approach. Hence, the recovered structure should be interpreted with caution if the mixing matrix estimation approach is unreliable. Devising more accurate approaches for estimating the mixing matrix, as well as extending the method proposed in this work to non-linear models are important directions 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] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *arXiv preprint arXiv:2109.13471*, 2021.

[4] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

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

[6] Robin J Evans. Graphs for margins of bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

[7] Robin J Evans. Margins of discrete bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

[8] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

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

[10] 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.

[11] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

[12] 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.

[13] Chris Meek. Strong-completeness and faithfulness in belief networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Citeseer, 1995.

[14] Judea Pearl. *Causality*. Cambridge university press, 2009.

[15] Judea Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] Yuqin Yang, Mohamed Nafea, AmirEmad Ghassami, and Negar Kiyavash. Causal discovery in linear structural causal models with deterministic relations. In *First Conference on Causal Learning and Reasoning*, 2022.

[22] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please refer to Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We believe that causal structure learning from observational data avoids potential negative societal impacts in causal discovery that might be associated with conducting randomized experiments.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions 1 and 2 (resp. 3) are our identifiability assumptions for linear SEM-ME (resp. SEM-UR).
   (b) Did you include complete proofs of all theoretical results? [Yes] Please refer to the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please refer to the Appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] This is explained in detail in the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Please refer to the Appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] Simulations are conducted using the CPU of a MacBook laptop.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cited the original paper of the baseline algorithm.
   (b) Did you mention the license of the assets? [N/A] Our simulations are conducted on synthetic data only.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]