Causal Representation Learning Made Identifiable by Grouping of Observational Variables

Hiroshi Morioka [1] **Aapo Hyvärinen** [2]

Abstract

A topic of great current interest is Causal Representation Learning (CRL), whose goal is to learn a causal model for hidden features in a data-driven manner. Unfortunately, CRL is severely ill-posed since it is a combination of the two notoriously ill-posed problems of representation learning and causal discovery. Yet, finding practical identifiability conditions that guarantee a unique solution is crucial for its practical applicability. Most approaches so far have been based on assumptions on the latent causal mechanisms, such as temporal causality, or existence of supervision or interventions; these can be too restrictive in actual applications. Here, we show identifiability based on novel, weak constraints, which requires no temporal structure, intervention, nor weak supervision. The approach is based on assuming the observational mixing exhibits a suitable grouping of the observational variables. We also propose a novel self-supervised estimation framework consistent with the model, prove its statistical consistency, and experimentally show its superior CRL performances compared to the state-of-the-art baselines. We further demonstrate its robustness against latent confounders and causal cycles.

1. Introduction

Causal discovery aims to learn causal interactions among observed variables in a data-driven manner (Pearl, 2000). The goal is to estimate a causal graph, also called an adjacency matrix, from passively observed data, with minimal assumptions. It plays an important role in a wide variety of fields, enabling fundamental insight into causal mechanisms latent in the data; importantly, this is possible without conducting expensive and time-consuming interventional experiments. However, the problem is ill-posed in general, and thus the main focus of causal discovery research is to find conditions where the causal graph can be uniquely determined (Andersson et al., 1997; Spirtes et al., 2001). A large number of studies have been conducted so far; they have basically found that imposing some asymmetricity into the model, such as nonlinearity or non-Gaussianity, enables its unique identification (Hoyer et al., 2008a; Peters et al., 2014; Shimizu et al., 2006; 2011; Zhang & Hyvärinen, 2009).

A crucial and implicit assumption of most causal discovery research is that we know exactly *what* constitutes the *causal variables*; in most cases, we implicitly assume that each observational variable corresponds to a single causal variable, i.e. a node in the causal graph. However, this is not necessarily true, for example when what is actually observed is raw, high-dimensional sensory data. Consider natural images: we do not really know in advance what kinds of objects are present, while the causal interactions should probably be modeled on the level of the objects. Therefore, in order to understand what kind of causal mechanism is generating such low-level sensory data, we also need to extract the "high-level" causal variables constructing the causal graph by performing *representation learning* (Bengio et al., 2013) simultaneously.

Nonlinear representation learning has its own problems of identifiability. Recent work has solved the identifiability problem in the context of Nonlinear Independent Component Analysis (NICA) by assuming temporal structure or an additional (conditioning, possibly unobservable) auxiliary variable (Hyvärinen & Morioka, 2016; Hälvä et al., 2021; Sprekeler et al., 2014). However, if the components are mutually independent, it seems impossible to model causal connections between them, and thus such theory is not directly applicable to this case. Thus, we need to go beyond independent components (Zhang & Hyvärinen, 2010; Khemakhem et al., 2020b) and build an explicit model of the dependencies resulting from their causal interactions.

Such simultaneous learning of the causal variables (representation learning) and their causal graph (causal discovery) has been a focus of intense attention recently, resulting is

[1]RIKEN Center for Advanced Intelligence Project, Kyoto, Japan [2]Department of Computer Science, University of Helsinki, Helsinki, Finland. Correspondence to: Hiroshi Morioka <hiroshi.morioka@riken.jp>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

what is called as Causal Representation Learning (CRL) (Schölkopf et al., 2021). Since both of the two separate problems combined here are known to be ill-posed, CRL seems to be even more severely ill-posed, and much less is known on identifiability of CRL. Yet, finding identifiability conditions is crucial for its interpretability, applicability, and reproducibility. So far very limited frameworks were proposed, and many of them are based on heavy assumptions on the causal mechanisms, such as supervision or intervention on latent variables or causal graphs (Brehmer et al., 2022; Shen et al., 2022; Yang et al., 2021), or temporal causality (dynamics) (Lachapelle et al., 2022; Lippe et al., 2022).

Here, we propose a new model for CRL based on a novel approach assuming that the observed variables follow a certain *grouping* structure known a priori, as illustrated in Fig. 1c. Such grouping is common and naturally appears in many practical situations. For example, the variables could be grouped based on which measurement sensor they come from in multimodal data; or which time point in causal dynamics, or geographical location in sensor networks they are measured at. We further assume that the causal interactions are *pairwise* as in some Markov network models. Under these assumptions, we prove identifiability with much weaker, and very different, constraints than previous work. The model in particular is able to consider instantaneous causal relations rather than temporal (Granger) causality, while autoregressive (AR) dynamics are further contained as a special case. Nor does our model assume any supervision or interventions. Our experiments on synthetic data as well as a realistic high-dimensional image and gene regulatory network datasets show that our framework can indeed extract latent causal variables and their causal structure, with better performance than the state-of-the-art baselines.

2. Related Works

The general form of the CRL problem can be defined as an estimation of a set of causally-related latent variables $\mathbf{s} = [s_1, \dots, s_{D_S}]^\mathsf{T} \in \mathbb{R}^{D_S}$, or causal variables for short, together with their causal structure. We typically assume that the observed data $\mathbf{x} \in \mathbb{R}^{D_\mathcal{X}}$ with $D_\mathcal{X} \geq D_\mathcal{S}$ are obtained via an unknown observational mixing $\mathbf{f} : \mathbb{R}^{D_S} \to \mathbb{R}^{D_\mathcal{X}}$ as

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \tag{1}$$

where the latent causal variables $s_1, \dots, s_{D_S}$ are *not* mutually independent but follow a causal model to be specified. Typically, the causal model would be a Structural Equation Model (SEM) (Shen et al., 2022; Yang et al., 2021) which has a well-defined causal semantics, or something simpler such as a Bayesian network (BN) as a kind of proxy.

In this work we only consider the case of independent and identically distributed (i.i.d.) sampling, which means the different observations of $\mathbf{x}$ are independent of each other and there is no time structure. This obviously implies the causal relations must also be instantaneous. Note that estimating instantaneous causality is more challenging compared to temporal causality (Lachapelle et al., 2022; Li et al., 2020; Lippe et al., 2022; 2023; Yao et al., 2022a;b), since we do not have prior knowledge about the causal direction or causal ordering between variables (from past to future; Fig. 1b) in instantaneous causality.

CRL can be seen as a generalization of NICA and causal discovery, both of which are known to be ill-posed without any specific assumptions. NICA can actually be seen as a special case of CRL, where the latent variables follow the degenerate causal graph in the sense of not having any causal relations. Recent studies have shown that NICA can be given identifiability by assuming temporal structure (Hyvarinen et al., 2019; Hälvä et al., 2021; Klindt et al., 2021; Sprekeler et al., 2014) instead of i.i.d. sampling. On the other hand, causal discovery is also a special case of CRL, where the causal variables are observed directly. Many studies have shown that some kind of asymmetricity of the statistical causal model enables the identifiability (Hoyer et al., 2008a; Park & Raskutti, 2015; Peters et al., 2014; Shimizu et al., 2006; 2011; Zhang & Hyvärinen, 2009).

The CRL model thus violates the important assumptions of the both problems (mutual independence, non-i.i.d., and direct observability). The research goal of CRL is thus to find the practical conditions for the identifiability of the model. Some studies have shown the identifiability in the instantaneous causality case, but they require supervision or intervention on the causal variables (Ahuja et al., 2022a;b; Brehmer et al., 2022; Shen et al., 2022; Yang et al., 2021) (Fig. 1a), or access to some latent information such as mixture oracles (Kivva et al., 2021), which might be too restrictive in actual applications. Recently some CRL studies proposed to use a grouping of variables instead of supervisions (Daunhawer et al., 2023; Lyu et al., 2022; Morioka & Hyvarinen, 2023; Sturma et al., 2023; Yao et al., 2023), similarly to this study. Most of them, except for Morioka & Hyvarinen (2023), especially focused on the intersections between groups. Sturma et al. (2023) showed identifiability of the intersection of the latent variables across all groups, based on linearity of the causal and observational models. Daunhawer et al. (2023); Lyu et al. (2022); Yao et al. (2023) considered more general causal and observational models, though the identifiability is limited to up to intersection-wise transformations. Morioka & Hyvarinen (2023) and our study instead focus on the group-specific causal variables not shared across groups. Morioka & Hyvarinen (2023) assumed a component-wise dependency as in NICA, and can be seen as a very special case of this study (see Section 8). A more detailed discussion about the related works are given in Supplementary Material I.

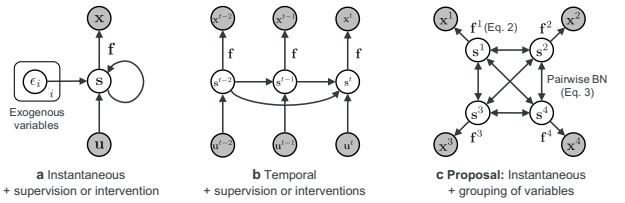

Figure 1. Comparison of the graphical models of major CRL frameworks whose goal is to estimate latent causal variables $\mathbf{s}$ from the low-level observations $\mathbf{x}$, usually with supervision or intervention $\mathbf{u}$. Our proposal in (c) is based on the grouping of variables (Eq. 2; $M = 4$ groups here) and the causal model based on a pairwise BN (Eq. 3), and does not require any supervision or intervention, greatly generalizing the existing models.

3. Model Definition

Our basic idea is to impose some constraints on the observational model $\mathbf{f}$, based on *grouping of the observed variables*, together with some Markov-like (pairwise) constraints on the causal interactions between the groups (Fig. 1c). Next we first define our observational model and then the causal model.

Observation Model As the original approach in our model, we assume that the observational mixing can be separated into $M > 1$ non-overlapping groups, as found in many practical cases. After appropriate permutations of the elements of $\mathbf{s}$ and $\mathbf{x}$ without loss of generality, we assume that the observation model Eq. 1 can be expressed as

$$\begin{aligned} \mathbf{x} &= \left[\mathbf{x}^1, \quad \ldots, \quad \mathbf{x}^M\right] = \mathbf{f}(\left[\mathbf{s}^1, \quad \ldots, \quad \mathbf{s}^M\right]) \\ &= \left[\mathbf{f}^1(\mathbf{s}^1), \quad \ldots, \quad \mathbf{f}^M(\mathbf{s}^M)\right], \end{aligned} \tag{2}$$

where $\mathbf{x}^m = [x^m_1, \ldots, x^m_{d^m_\mathcal{X}}]^\mathsf{T} \in \mathbb{R}^{d^m_\mathcal{X}}$ and $\mathbf{s}^m = [s^m_1, \ldots, s^m_{d^m_\mathcal{S}}]^\mathsf{T} \in \mathbb{R}^{d^m_\mathcal{S}}$ are the m-th group of the observational and latent variables respectively. Each element of the latent and the observational variables belongs to only one of the groups with index in $\mathcal{M} = \{1, \ldots, M\}$, which means that the m-th observational group $\mathbf{x}^m$ is generated only as a function of $\mathbf{s}^m$, without any observational contaminations from the other groups: i.e., $\mathbf{x}^m = \mathbf{f}^m(\mathbf{s}^m)$, $D_\mathcal{S} = \sum_{m=1}^M d^m_\mathcal{S}$, and $D_\mathcal{X} = \sum_{m=1}^M d^m_\mathcal{X}$. The number of variables in a group can be different across groups. We usually denote the group index as a superscript, which should not be confused with an exponent; the element index as denoted by a subscript. Note that when $M = 1$ this model corresponds to the general CRL (Eq. 1), and when $M = D_\mathcal{S}$ this model simply leads to the ordinary causal discovery problem without observational mixing.

Such grouping structure is not anything unrealistic, and can be seen in many practical situation, as described in the following illustrative examples.

Illustrative Example 1: Causally Related Sensor Measurements The most intuitive example would be where m is a sensor index. Data is then obtained from a set of M sensors, each measuring different but causally-related multidimensional physical quantities $\mathbf{x}^{m(n)} \in \mathbb{R}^{d^m_\mathcal{X}}$ for each sample n. For example, in single-cell multiomics data (Burkhardt et al., 2022), each cell (n) could be measured to give chromatin accessibility (DNA) as $\mathbf{x}^{1(n)}$, gene expressions (RNA) as $\mathbf{x}^{2(n)}$, and protein levels as $\mathbf{x}^{3(n)}$. These are all multi-dimensional quantities representing causally-interacting latent high-level features $\{\mathbf{s}^m\}_m$. There exist many other possible applications consistent with this observational model; e.g., multimodal biomedical data (Acosta et al., 2022), simultaneous measurements of brain and behavior (Hebart et al., 2023), climate monitoring sensor networks (Longman et al., 2018), and so on, where m corresponds to sensor modalities or locations (see Section 8 for more details).

Illustrative Example 2: Causal Dynamics Although we focus on independent data samples rather than dynamics, our model can also implement dynamics with dependency across *time* by simply defining the group-index m as time-index t (see Figs. 1b and c). We then obtain low-level observations $\mathbf{x}^t$ (such as images) from high-level latent causal process composed of multidimensional variable $\mathbf{s}^t$ through time-dependent mixing model $\mathbf{x}^t = \mathbf{f}^t(\mathbf{s}^t)$ for each time point t. In this case, our model gives a generic form of a time series model, which is actually more general compared to some existing studies of CRL based on dynamics (Lachapelle et al., 2022; Lippe et al., 2022; Yao et al., 2022a;b) in the sense that the mixing function $\mathbf{f}^t$ changes as a function of time t, which can happen in many practical situations (such as changes of the camera angle capturing the images). In this case, we assume we observe the same time-series many times, i.e. we have $\mathbf{x}^{t(n)}$ where n is the index of the time series realization (e.g., capturing images with multiple sequences n with the same transition of camera angles across t every time).

Causally Structured Latent Variable Next we model the causal structure of the latent variables based on a BN, which is in particular *pairwise*, in the spirit of pairwise Markov random fields. Denote by $\phi(\cdot, \cdot)$ a potential function representing causal relations between two variables, which is the same for all variable pairs. Further, denote by $\bar{\phi}^m$ group-wise potential functions representing causal relations *inside* a group, i.e. between the elements of $\mathbf{s}^m$, which are not restricted in any way. We assume that the joint

distribution is factorized as

$$p(\mathbf{s}) \propto \left[\prod_{m\in\mathcal{M}} \exp\left(\bar{\phi}^m(\mathbf{s}^m)\right)\right] \times \prod_{m\neq m'} \prod_{(a,b)\in\mathcal{V}_S^m\times\mathcal{V}_S^{m'}} \exp\left(\lambda_{ab}^{mm'}\phi(s_a^m, s_b^{m'})\right), \tag{3}$$

where we denote the set of indices of the latent variables belonging to the m-th group by $\mathcal{V}_S^m$ ($|\mathcal{V}_S^m| = d_S^m$). The idea is to have a model of dependencies between variables which is so general that the estimation of the representation is not biased towards independent components. The variables in one group $\mathbf{s}^m$ can causally affect all variables on the other groups $m' \neq m$, and also in the same group m, in a complex manner, thus breaking any independence of variables. The coefficient $\lambda_{ab}^{mm'} \in \mathbb{R}$ modulates the strength of the causal relation $s_a^m \to s_b^{m'}$, which is constantly zero if s_a^m is not a direct causal parent of $s_b^{m'}$. The sets of the coefficients $\mathbf{L} = \{\mathbf{L}^{mm'}\}_{(m,m')}, \mathbf{L}^{mm'} = (\lambda_{ab}^{mm'})_{a\in\mathcal{V}_S^m, b\in\mathcal{V}_S^{m'}}$ can be interpreted as (inter-group) weighted adjacency matrices.

This model is rather general, and includes a form of exponential family BNs (see Supplementary Material B). This allows for assigning causal semantics to the model by considering its equivalence to some SEMs, such as causal additive models (CAMs; Bühlmann et al. (2014)), which are more general than linear SEMs on which some CRL frameworks are based (Shen et al., 2022; Sturma et al., 2023; Yang et al., 2021). Our model is not even restricted to Gaussian BNs as in CAMs. Although Morioka & Hyvarinen (2023) used a similar causal model, their model can be seen as a very special case of ours (see Section 8); especially, Eq. 3 does not require any mutual independence across variables.

Note that Eq. 3 just represents the factorization of the joint distribution and does not incorporate any causal directional assumptions between variables. We thus need some additional assumptions for the identifiability of this factorization model as a *causal* model as shown in Theorems below, similarly to causal discovery based on BN.

Illustrative Example 1: Causally Related Sensor Measurements In the single-cell multiomics example, the model says there are causal relations between (and within) the groups, which can be consistent with what is known as the *central dogma* in molecular biology; DNA ($\mathbf{x}^1$) $\to$ RNA ($\mathbf{x}^2$) $\to$ Protein ($\mathbf{x}^3$). Our model considers that they are interacting on some high-level latent space $\{\mathbf{s}^m\}_m$, such as something related to transcription factors.

Illustrative Example 2: Causal Dynamics In the temporal dynamics example above, it is natural to have causal relations across the time-index t (which is the same as the group-index m). Our model extends the previous models in the sense that any pairs of latent variables can be causally related across time (no sparseness is required unlike Lachapelle et al. (2022)). It is also worth mentioning that temporal causality from the past to the present is a special case in our model since our model (Eq. 3) does not restrict the causal directions between the groups.

4. Identifiability of Representation Learning

Based on the grouping assumption of the observational model (Eq. 2), together with the latent variable model given above (Eq. 3), we can prove new identifiability results of the CRL model. In this section, we first consider identifiability of the latent variables. We assume that each mixing function $\mathbf{f}^m$ is invertible and a C^2 diffeomorphism (thus $d_S^m = d_X^m$; we later discuss the case $d_S^m < d_X^m$). Apart from that, we do not assume any parametric form for each $\mathbf{f}^m$. We consider the situation where the support of the distribution of each variable is connected (i.e. an interval), and without loss of generality, the same across all variables, denoted as $\bar{S}$. We denote $\phi^{112}(x,y) = \frac{\partial^3}{\partial x^2\partial y}\phi(x,y)$, $\phi^{122}(x,y) = \frac{\partial^3}{\partial x\partial y^2}\phi(x,y)$, and $\phi^{12}(x,y) = \frac{\partial^2}{\partial x\partial y}\phi(x,y)$. Those functions are said to be *uniformly dependent* (Definition 2 in Supplementary Material B) if the set of zeros of the function does not contain any open subset in the support of the input distribution. (**Neighbor**) For a variable s_a^m, we call $s_b^{m'}$ in some *other* group a *neighbor* if either or both of the adjacency coefficients $\lambda_{ab}^{mm'}$ and $\lambda_{ba}^{m'm}$ are non-zero. The identifiability condition is then given in the following Theorem, proven in Supplementary Material C;

Theorem 1. *Assume the generative model given by Eqs. 2 and 3, and also the following:*

A1 (Nondegeneracy of the graph) For any group m in $\mathcal{M}$ (or in a subset of $\mathcal{M}$, that we call "the groups of interest"), each variable has a (at least one) neighbor in some other group, and the collection of inter-group adjacency matrices $\bar{\mathbf{L}}^m$ given below has full row-rank after removing all-zero rows:

$$\bar{\mathbf{L}}^m = \begin{bmatrix} \mathbf{L}^{m1}, & \dots, & \mathbf{L}^{mM} \\ (\mathbf{L}^{1m})^\mathsf{T}, & \dots, & (\mathbf{L}^{Mm})^\mathsf{T} \end{bmatrix}, \tag{4}$$

A2 (Causal function) ϕ^{12}, ϕ^{112}, and ϕ^{122} have uniform dependency, and for any open subset B of $\bar{S}$, there exist some $z_1 \neq z_2 \in \bar{S}$ such that any of the following conditions does not hold for ϕ^{12}: $\phi^{12}(s,z_1) = c_1\phi^{12}(s,z_2)$, $\phi^{12}(z_1,s) = c_2\phi^{12}(z_2,s)$, and $\phi^{12}(s,z_1) = c_3\phi^{12}(z_1,s)$ for all $s \in B$ with some constants $c_1, c_2, c_3 \in \mathbb{R}$.

Then, for all groups m in $\mathcal{M}$ (or in the groups of interests), $\mathbf{s}^m$ can be recovered up to permutation and variable-wise invertible transformations from the distribution of $\mathbf{x}$.

The Assumption A1 is rather intuitive, and requires the variables (rows) in each group m to have distinctive sets of

(or causal strengths to) neighbors (columns), for both causal directions (upper and lower halves), as expressed by the full row-rank condition. This indicates nondegeneracy of the graph, which is known to be crucial in CRL (Kivva et al., 2021; Morioka & Hyvarinen, 2023). Note that A1 does not require the causal graph to be *directed* as in Theorem 3 given below, though requires it to be *asymmetric*. Note also that the variables need to have neighbors only on some of the other groups but not on all of them; e.g. in practice, groups somehow "near-by" in space or time. This condition can be evaluated separately for each group m; we cannot identify the latent variables of the groups not satisfying A1, while they do not affect the identifiability of the other groups.

The Assumption A2 requires the (cross-derivative of) ϕ to be non-factorizable (the first two equations) and asymmetric (the last equation). The asymmetricity is a well-known requirement for causal discovery (Hoyer et al., 2008a; Peters et al., 2014), which excludes linear Gaussian SEMs, and thus reasonable for CRL as well. The non-factorizability cannot be satisfied when the cross-derivative of ϕ is factorized into variable-wise scalar functions, which in turn requires sufficiently complex dependency of variables. In the exponential family BN characterization of Eq. 3, the models with model order more than 1 can satisfy this (Supplementary Material B). We can give an alternative condition not requiring this non-factorizability, by some additional constraints on the causal graph (Proposition 1 in Supplementary Material D), which then allows exponential family BNs with order 1, including CAMs. This resembles the requirement for some representation learning (e.g., non-quasi-Gaussianity in Hyvarinen & Morioka (2017)), and is thus reasonable for CRL as well.

5. Representation Learning Algorithm

We now propose a self-supervised estimation framework called Grouped Causal Representation Learning (G-CaRL). Again, we start by learning the representation, i.e. learning to invert the mixing functions $\{\mathbf{f}^m\}$. To this end, we propose a new contrastive learning method where the pretext task is to discriminate (classify) the following two datasets obtained from the same observations:

$$\begin{aligned} &\mathbf{x}^{(n)} = \left[\mathbf{x}^{1(n)}, \quad \dots, \quad \mathbf{x}^{M(n)}\right] \\ &\text{vs.} \quad \mathbf{x}^{(n_*)} = \left[\mathbf{x}^{1(n_*^1)}, \quad \dots, \quad \mathbf{x}^{M(n_*^M)}\right] \end{aligned} \tag{5}$$

where n indicates the sample index, while n_*^m is a shuffled index, generated in practice by randomly selecting a sample index separately for each group m (note that different groups have different sample indices in $\mathbf{x}^{(n_*)}$). We then learn a nonlinear logistic regression (LR) system which discriminates the two classes, using a cross-entropy loss with a specific form of the regression function with

$$\begin{aligned} r(\mathbf{x}) = c + \sum_{m\in\mathcal{M}} \bar{\psi}^m(\mathbf{h}^m(\mathbf{x}^m)) \\ + \sum_{m\neq m'} \sum_{(a,b)\in\mathcal{V}_{\mathcal{S}}^m\times\mathcal{V}_{\mathcal{S}}^{m'}} w_{ab}^{mm'}\psi(h_a^m(\mathbf{x}^m), h_b^{m'}(\mathbf{x}^{m'})), \end{aligned} \tag{6}$$

where $\mathbf{h}^m : \mathbb{R}^{d_{\mathcal{X}}^m} \to \mathbb{R}^{d_{\mathcal{S}}^m}$ is a group-wise (nonlinear) feature extractor, h_a^m is the a-th element of $\mathbf{h}^m$, $\bar{\psi}^m : \mathbb{R}^{d_{\mathcal{S}}^m} \to \mathbb{R}$ and $\psi : \mathbb{R}^2 \to \mathbb{R}$ are scalar-valued nonlinear functions, and $w_{ab}^{mm'}$ and $c \in \mathbb{R}$ are weight and bias parameters. Importantly, this regression function is designed to be consistent with the causal model defined in Section 3 (Eq. 3): the feature extractors $\mathbf{h}^m$ and the weight parameters $w_{ab}^{mm'}$ correspond to the causal variables $\mathbf{s}^m$ and the adjacency coefficients $\lambda_{ab}^{mm'}$ in the causal model (Eq. 3), respectively. This indicates that learning those parameters in a data-driven manner should lead to CRL automatically, as justified in the following Theorems. The observational grouping indices are assumed to be given in advance, while we only need the information of the size of groups $d_{\mathcal{S}}^m$ for the latent variables. The nonlinear functions are typically learned as neural networks with universal approximation capacity (Hornik et al., 1989). Any optimization method can be used to minimize the loss (see Supplementary Algorithm 1 for example), though the theorem below assumes it gives the optimal solution without getting stuck in a local optimum. After the convergence, we obtain the following consistency Theorem, proven in Supplementary Material E;

Theorem 2. *Assume the same as those in Theorem 1, and:*

- *B1 (Learning) We train a nonlinear LR system (Eq. 6) with universal approximation capability to discriminate two datasets $\mathbf{x}^{(n)}$ and $\mathbf{x}^{(n_*)}$ (Eq. 5).*
- *B2 ($\mathbf{h}$) The functions $\mathbf{h}^m$ are C^2 diffeomorphisms.*

Then, for all groups m in $\mathcal{M}$ (or in the groups of interests in A1), in the limit of infinite samples n, the function $\mathbf{h}^m(\mathbf{x}^m)$ gives the latent variables on the m-th group $\mathbf{s}^m$, up to permutation and variable-wise invertible transformations.

Interestingly, in spite of the lack of clear relevance of this pretext task to CRL at first sight, this theorem actually shows that learning the correct representation is achieved by learning the functions $\mathbf{h}^m(\cdot)$ through the optimization of the regression function Eq. 6. This Theorem is basically based on the well-known properties of the logistic regression (Gutmann & Hyvärinen, 2012). Intuitively, the group-wise shuffling applied to $\mathbf{x}^{(n_*)}$ (Eq. 5) breaks the causal relations between groups, which means for the LR to discriminate the two datasets in Eq. 5 properly, it needs to capture the inter-group causal relations in the latent space, by disentangling the observational mixing. Thanks to the compatibility of the factorization assumptions in the generative (Eq. 3)

and in the regression model (Eq. 6) as mentioned above, the optimal model is achievable only when $\mathbf{h}^m$ essentially gives the inverse model of $\mathbf{f}^m$, which automatically leads to CRL.

6. Identifiability of Causal Discovery

Next, we consider how to learn the causal graph $\mathbf{L}$. In our model, this can be achieved by estimating the (weighted) adjacency coefficients $\lambda_{ab}^{mm'}$ in Eq. 3 from the observations in a data-driven manner, similarly to many causal discovery frameworks based on BN (Choi et al., 2020; Park & Park, 2019b). We can actually achieve this simultaneously with the representation learning by using the self-supervised algorithm in Section 5, where the adjacency coefficients $\{\lambda_{ab}^{mm'}\}$ are learned as the weight parameters $\{w_{ab}^{mm'}\}$ jointly with the inverse models $\mathbf{h}^m$ in the regression function (Eq. 6).

The identifiability of $\mathbf{L}$ requires additional assumptions on $\mathbf{L}$ and ϕ, given in the following Theorem, proven in Supplementary Material F. We first give the definitions of some terms used in the Theorem. (**Directed**) We call a causal relation between two variables s_a^m and $s_b^{m'}$ *directed* if only one of $\lambda_{ab}^{mm'}$ and $\lambda_{ba}^{m'm}$ has a non-zero value. (**Co-parent and co-child**) For a variable s_a^m, we call $s_{a'}^m$ in the *same* group a *co-parent* (respectively *co-child*) of s_a^m if it shares at least one child (respectively parent) $s_b^{m_*}$ on some *other* group ($m_* \in \mathcal{M} \setminus m$) with s_a^m. The $s_b^{m_*}$ can be arbitrarily selected for each co-parent and co-child.

Theorem 3. *Assume the same as those in Theorem 1, and:*

C1 (Causal graph) The inter-group causal relations of variables are all directed, and for every group-pair (m, m') in the groups of interest, all variables in a group m (and m') have both a co-parent and a co-child in the same group. In addition, any variables in the group m (and m') can be reached from any other variables in the same group by moving from a variable to one of its co-parents, possibly by multiple hops (and similarly for co-children).

C2 (Asymmetricity) There is no open subset B of $\bar{\mathcal{S}}$ such that for all $x \neq y \in B$, it holds

$$\phi^{12}(x, y) = c\phi^{12}(y, x) \tag{7}$$

with some constant $c \in \mathbb{R}$.

Then, for all group-pairs (m, m') satisfying A1 and C1, $(\mathbf{L}^{mm'}, \mathbf{L}^{m'm})$ are identifiable up to permutation of variables, linear scaling, and matrix transpose.

This theorem shows that we can identify the causal graphs $\mathbf{L}$ from the data distribution up to linear scaling and matrix transpose. This also indicates that the graph can be estimated indeed consistently by the learning algorithm in Section 5 as claimed above (we omit the proof of the consistency since it can be easily shown by following those of Theorems 2 and 3). The indeterminacy of matrix transpose comes from the lack of specification of the functional form of ϕ; a transposition of the adjacency matrix could be compensated by switching the order of the two arguments of ϕ, which can be resolved by some prior information on ϕ.

The function ϕ is required to be asymmetric (C2), which is a well-known requirement for causal discovery as mentioned in Section 4, though C2 is a bit stronger than that in A2.

The causal graph assumption (C1) is a special condition related to the grouping of variables (we can also consider alternative condition which requires only either co-parent or co-child for each variable; Supplementary Material G). This would be fulfilled as long as the connections between groups are not too sparse (See Supplementary Material H for some illustrative discussion). One typical example would be fully-connected causal effects from one group to another; e.g., fully-connected causality DNA $\rightarrow$ RNA and RNA $\rightarrow$ Protein on their latent space in Illustrative Example 1, and fully-connected temporal causality to some (or all) subsequent time-point(s) in Illustrative Example 2, since in those cases all other variables in the same group are either (or both) co-parents or co-children. Of course they do not need to be *fully*-connected in practice, and denseness is not a necessary condition in theory. The graph is required to be directed though not necessarily *acyclic*.

7. Experiments

To validate the effectiveness of our framework, we compare it to several baselines in two simulation settings and two more realistic scenarios (see Supplementary Material J for the details; the implementation of G-CaRL is available at `https://github.com/hmorioka/GCaRL`). The baselines only include *unsupervised* frameworks with *instantaneous* (causal) interactions, since our experimental setting does not include supervision, intervention, nor temporal causality. Specifically, we compared with three CRL frameworks MVCRL (Yao et al., 2023), CausalVAE (Yang et al. (2021) in unsupervised setting), and CCL (Morioka & Hyvarinen, 2023), and three representation learning (RL) frameworks MFCVAE (Falck et al. (2021); Kivva et al. (2022)), VaDE (Jiang et al. (2017); Kivva et al. (2022); Willetts & Paige (2022)), and β-VAE (Higgins et al., 2017). See Supplementary Material K for the details. Daunhawer et al. (2023); Lyu et al. (2022) are not applicable since they are limited to two-group settings. We also applied Kivva et al. (2021) but it failed due to the difficulty of the estimation of the mixture model in our data. For a fair comparison, we used group-wise structures for the encoders of those baselines similarly to G-CaRL. For baselines which do not estimate the (part of) causal graph, we additionally applied a

causal discovery framework to the estimated latent variables as post-processing, from a wide variety of choices so as to maximize the performances (see Supplementary Fig. 4). We only evaluated the *inter-group* causal graphs, since only those are identifiable in our model (Theorem 3).

7.1. Simulation 1: DAG

We first examined the performance with latent DAG models (Supplementary Fig. 7a shows some examples). The number of groups (M) was fixed to 3, the number of variables was 10 for each group ($d_{\mathcal{S}}^m = 10, D_{\mathcal{S}} = 30$). The latent variables are observed through nonlinear mixings randomly generated as a multilayer-perceptron (MLP) for each group.

The latent variables and the causal graphs were reconstructed reasonably well by G-CaRL, with much higher performances than the baselines (Fig. 2a). The baselines did not work well basically because of the lack of representation capability (CCL), lack of supervision in our setting (CausalVAE), or lack of explicit considerations of the latent causality (others). The worst performance of CCL indicates the inadequacy of assuming mutual independence among some variables in this dataset.

Supplementary Fig. 5a shows how the complexity of the mixing model (L), the number of variables $D_{\mathcal{S}}$, groups M, and sample size n affect the performances; a higher L, $D_{\mathcal{S}}$, and M make learning more difficult, while a larger n makes it possible to achieve higher performances, as expected.

7.2. Simulation 2: Cyclic Graphs with Latent Confounders

To show the robustness of G-CaRL on more complex causal models, we next examined the performances with directed cycles and latent confounders with the same number as the observable variables (Fig. 2b; see Supplementary Fig. 7b for the difficulty of this setting). G-CaRL showed reasonably good performance even in this difficult condition, though it requires larger number of samples than Simulation 1 (Supplementary Fig. 5b). Note that this causal model is difficult even for the conventional causal discovery frameworks directly applied to the true latent variables (Supplementary Fig. 4). This result shows the effectiveness of the causal model of G-CaRL against the existence of causal cycles and latent confounders. Supplementary Fig. 5b shows the effects of the model complexity to the estimation performances, and they show a trend similar to Simulation 1.

7.3. Recovery of Gene Regulatory Network

We also evaluated G-CaRL on a more realistic causal data for showing the robustness against misspecification of the causal model. We used synthetic single-cell gene expression data generated by SERGIO (Dibaeinia & Sinha, 2020), where each gene expression is governed by a stochastic differential equation (SDE) derived from a chemical Langevin equation. Due to such generative process, the true causal relations cannot really be represented by our causal model (Eq. 3). The causal graph was designed to be a DAG (as required of SERGIO) similarly to Simulation 1, but with latent confounders similarly to Simulation 2 (Supplementary Fig. 7c shows examples). We used the same setting for the observational mixings as those in the simulations above. G-CaRL showed the best performances among the baselines (Fig. 2c), which suggests the robustness of G-CaRL against misspecification of the causal model, and the good applicability of G-CaRL to real datasets.

7.4. High-Dimensional Image Observations

We also evaluated G-CaRL on a more realistic observational model, by using a high-dimensional image dataset (3DIdent; Zimmermann et al. (2021)). Images of a tea cup were generated based on ten latent factors for each group, which were causally interacting within/across groups with directed cycles under influence of latent confounders as in Simulation 2 (see Supplementary Fig. 8 for the difficulty of this setting). We compared the performance with two baselines MVCRL and CCL, which do not require learning a decoder, since learning both encoder and decoder should be unstable in this high-dimensional setting.

The estimation performances were reasonably good even in such high-dimensional observations with a complex causal model (Fig. 2d). This suggests the good applicability of G-CaRL to high-dimensional real data.

8. Discussion

Our proposal extends the existing CRL models in many aspects; 1) the framework is unsupervised (only requires *grouping of variables* independent of the samples) rather than supervised (Brehmer et al., 2022; Shen et al., 2022; Yang et al., 2021), 2) we consider instantaneous causality rather than temporal (Lachapelle et al., 2022; Lippe et al., 2022; Yao et al., 2022a;b), while the temporal causality is also contained as a special case, 3) the observational mixing can be group-(or time-)dependent rather than invariant (Lachapelle et al., 2022; Lippe et al., 2022; Morioka & Hyvarinen, 2023; Yao et al., 2022a;b), 4) the latent variables can be nonlinearly causally related rather than linearly (Shen et al., 2022; Yang et al., 2021), nor is sparseness (Lachapelle et al., 2022) necessary, and 5) the causal graph can be cyclic, which is even more general than commonly used models for simple causal discovery.

Meanwhile, our framework also has some connections to existing CRL; it can be understood that our theorems are virtually using the *variables on the other groups* as auxil-

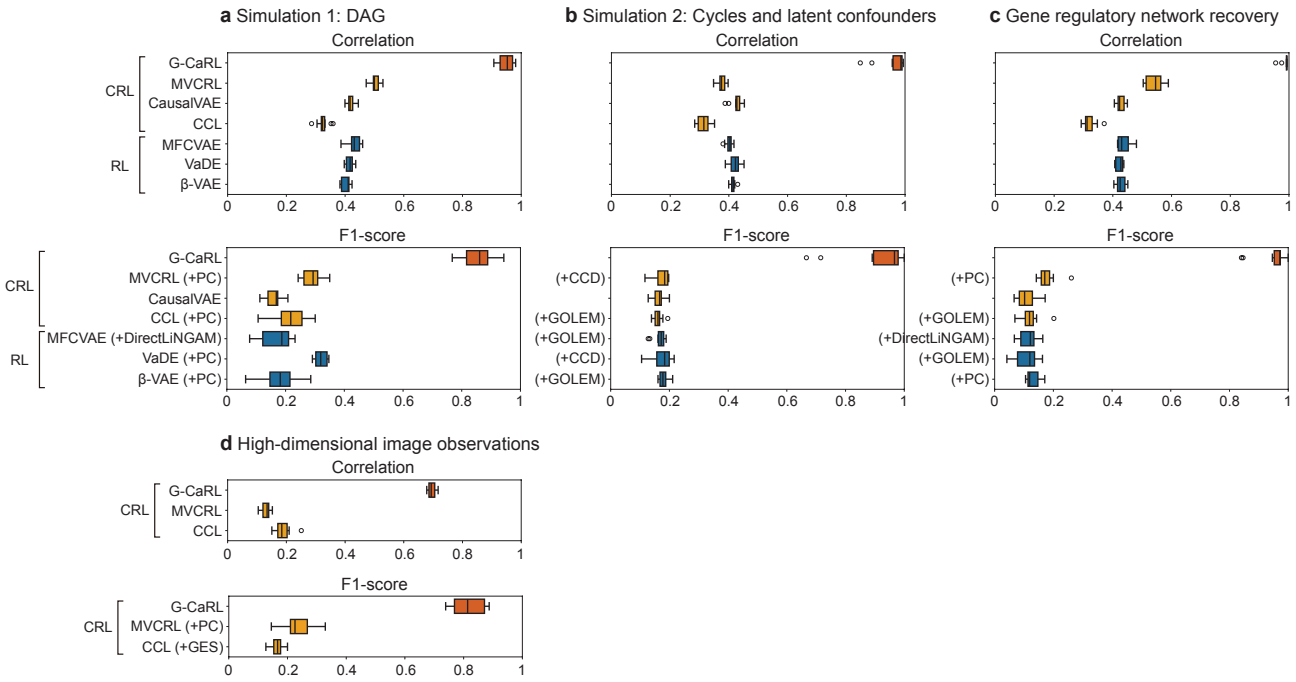

Figure 2. Comparison of CRL performance by the proposed G-CaRL and the baselines. The performances are measured by correlation for the latent variables, and by F1-score for the causal graphs, excluding the intra-group sub-graphs. The parentheses after the names of some (C)RL frameworks indicate the causal discovery frameworks additionally applied as post-processing.

iary variables in the context of (weakly-)supervised CRL. Importantly, instead of requiring *additional* auxiliary variables (supervision) for each sample as in those existing CRL, this study nicely utilized the grouping structure to obtain them automatically. Our estimation framework based on the group-wise shuffling somewhat resembles some multimodal CRL based on contrastive learning (e.g., Daunhawer et al. (2023)). This study nicely extends them to more than two-group settings, and gave theoretically guarantee of the identifiability of the (group-specific) latent variables up to variable-wise transformations, which is quite new.

This study greatly generalizes the recently proposed CRL framework CCL (Morioka & Hyvarinen, 2023) in many aspects; CCL assumes that 1) the mixing functions $\mathbf{f}^m$ are the same for all groups (called *nodes* in CCL) m rather than group-specific as in ours (Eq. 2), 2) the causal relations are component-wise, similarly to NICA (in other words, the adjacency coefficients $\lambda_{ab}^{mm'}$ in Eq. 3 can have non-zero values only when $a = b$, rather than all pairs of (a, b) as in ours), and 3) the causal graph is a forest, while ours can be more general and even cyclic. These indicate much higher generality and applicability of our model. Such generality requires our new estimation framework G-CaRL, which is very different from CCL, though both of them can be categorized as self-supervised (contrastive) learning; CCL used group-paired data for taking contrast, while our G-CaRL used group-wise-shuffled data (Eq. 5).

Our model can be seen as a generalization of NICA, in the sense that 1) it does not require mutual independence of variables, and further, 2) the mixing function is time/group-dependent, which is completely new in NICA. Independently Modulated Component Analysis (Khemakhem et al., 2020b) was recently proposed as an extension of NICA to allow dependency across variables, but it requires an auxiliary variable unlike our framework.

There exist many possible applications where our grouping assumptions are applicable: for example, 1) multimodal measurements, for example consisting of brain activities (group 1), external stimuli (group 2), and behaviors (group 3) of animals or humans (Hebart et al., 2023). Each group has multidimensional observations (multiple brain regions, stimuli, and behaviors) as time series (samples). Our framework would give new insight into how the brain is organizing behaviors based on external factors at an abstract level. 2) Sensor network for example in climate monitoring (Longman et al., 2018), where each group is a single sensor location, and each sensor is measuring such as temperature, humidity, rainfall, pollutants, etc. (*variables*) in the location. Our method would extract some hidden causal relations between sensor locations on high-level feature space. 3) Medical multimodal data (Acosta et al., 2022), where *groups* are dif-

ferent types of medical record obtained from subjects (samples); e.g., genetic factors (group 1), self-report lifestyle (group 2), and clinical diagnosis by doctors (group 3). Our framework would presumably find high-level connections between those groups; e.g., what kinds of (combinations of) genetic factors and/or (combinations of) lifestyles are causing some type of diseases, and so on. 4) Our model can also consider very general types of temporal causality (see Illustrative Example 2).

Although we considered the case $d_{\mathcal{X}}^m = d_{\mathcal{S}}^m$ for theoretical convenience, we can extend our framework to $d_{\mathcal{X}}^m \geq d_{\mathcal{S}}^m$, as empirically validated in Section 7.4. One approach is to assume that $\mathbf{f}^m$ is injective and a C^2 diffeomorphism between $\mathcal{S}^m$ and $\mathcal{X}^m \in \mathbb{R}^{d_{\mathcal{X}}^m}$ which is a C^2 differentiable manifold, as in Hälvä et al. (2021). Our proof can be adapted for that case, following that study. Another approach is to assume that $d_{\mathcal{X}}^m - d_{\mathcal{S}}^m$ variables are not causally related to any other groups (but possibly within each group, implicitly included in $\bar{\phi}^m$ in Eq. 3), in which case our estimation framework would automatically ignore them, which does not affect the identifiability of the $d_{\mathcal{S}}^m$ variables. We can also consider noisy mixing models via a standard deconvolution argument as in Khemakhem et al. (2020a).

Our Theorems assume grouping of the observational mixing without any observational contamination across groups (i.e., there are only group-specific variables), in contrast to some other group-based CRL frameworks focusing rather on intersections of groups (Daunhawer et al., 2023; Lyu et al., 2022; Yao et al., 2023). Our assumption should be satisfied in many practical applications, as described in Illustrative Examples above. In addition, it makes our causal assumptions of the (group-specific) latent variables much weaker than those studies; our group-specific variables can be causally related across groups in our model (Eq. 3) rather than mutually independent (e.g., Corollary 3.10 and 3.11 of Yao et al. (2023)). Furthermore, our assumption makes the latent variables identifiable up to variable-wise transformations (Theorem 1) rather than block-wise transformations.

The identifiability of the causal graphs only applies for connections *across* groups, and those *within* each group are left unknown. Nevertheless, since the latent variables are guaranteed to be identifiable, we can apply existing causal discovery framework for estimating them as a post-processing. Although the ϕ is assumed to be the same across all group-pairs for simplicity, it could be different across them, but that might require some additional assumptions.

Future research might consider relaxing some of the assumptions in our work. Although the grouping of variables without overlaps between groups can be satisfied in plenty of situations in practice (see above), allowing such overlaps should make our framework even more applicable to a wider variety of situations. Multivariate (vector) causal variables might be more favorable for representing complex and high-level phenomena behind data, compared to univariate causal variables as considered in this study. In such case, vector-wise identifiability as in (Daunhawer et al., 2023; Yao et al., 2023) rather than component-wise one (Theorem 1) might be enough, which is weaker and thus might relax some of the assumptions in this study. Applying our framework to more realistic datasets, such as iTHOR (Kolve et al., 2022), would be an important future direction for assessing its applicability and robustness.

9. Conclusion

This study proposed a new identifiable model for CRL, together with its self-supervised estimation framework G-CaRL. The new approach is the assumption of the grouping of the observational variables, which appears naturally in many practical applications such as multi-sensor measurements or time series. Such an assumption allowed us to significantly weaken any other assumptions required on the latent causal mechanisms in existing frameworks. In contrast to existing CRL models, our model does not require temporal structure (although it can use it as a special case), nor does it assume any supervision or interventions. Although our model restricts the inter-group causal relations of variables to some extent, it allows nonlinearity and even cycles, which is more general than most of the causal discovery models. Numerical experiments showed better performances compared to the state-of-the-art baselines, thus making G-CaRL a promising candidate for real-world CRL in a wide variety of fields.

Acknowledgements

This research was supported in part by JST PRESTO JPMJPR2028, JSPS KAKENHI 24H02177, 22H05666, and 22K17956. A.H. was funded by a Fellow Position from CIFAR, and the Academy of Finland (project #330482). We also would like to thank the anonymous reviewers for very useful comments that helped us improve the manuscript.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.

Ahuja, K., Hartford, J. S., and Bengio, Y. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15516–15528, 2022a.

Ahuja, K., Wang, Y., Mahajan, D., and Bengio, Y. Interventional causal representation learning. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022b.

Andersson, S. A., Madigan, D., and Perlman, M. D. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013.

Bollen, K. A. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

Brehmer, J., de Haan, P., Lippe, P., and Cohen, T. S. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38319–38331, 2022.

Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Schölkopf, B., and Ravikumar, P. Learning linear causal representations from interventions under general nonlinear mixing. arXiv, 2023.

Bühlmann, P., Peters, J., and Ernest, J. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014.

Burkhardt, D., Luecken, M., Benz, A., Holderrieth, P., Bloom, J., Lance, C., Chow, A., and Holbrook, R. Open problems - multimodal single-cell integration. Kaggle, 2022.

Cai, R., Xie, F., Glymour, C., Hao, Z., and Zhang, K. Triad constraints for learning causal structure of latent variables. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chen, R. T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Chickering, D. M. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2003.

Choi, J., Chapkin, R., and Ni, Y. Bayesian causal structural learning with zero-inflated Poisson bayesian networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5887–5897, 2020.

Comon, P. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., and Vogt, J. E. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Dibaeinia, P. and Sinha, S. SERGIO: A single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11(3):252–271.e11, 2020.

Falck, F., Zhang, H., Willetts, M., Nicholson, G., Yau, C., and Holmes, C. C. Multi-facet clustering variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8676–8690, 2021.

Geiger, D. and Heckerman, D. Learning Gaussian networks. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pp. 235–243, 1994.

Gong, M., Zhang, K., Schoelkopf, B., Tao, D., and Geiger, P. Discovering temporal causal relations from subsampled data. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1898–1906, 2015.

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 217–227, 2020.

Gresele, L., Kügelgen, J. V., Stimper, V., Schölkopf, B., and Besserve, M. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012.

Hälvä, H. and Hyvärinen, A. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pp. 939–948, 2020.

Hälvä, H., Le Corff, S., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvarinen, A. Disentangling identifiable features from noisy data with structured nonlinear ica. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1624–1633, 2021.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer, New York, NY, 2001.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C. I. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Horan, D., Richardson, E., and Weiss, Y. When is unsupervised disentanglement possible? In *Advances in Neural Information Processing Systems*, volume 34, pp. 5150–5161, 2021.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pp. 689–696, 2008a.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008b.

Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NIPS) 29*, pp. 3765–3773. 2016.

Hyvarinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In *AISTATS*, pp. 460–469, 2017.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.*, 12(3):429 – 439, 1999.

Hyvärinen, A. and Smith, S. M. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14(1): 111–152, 2013.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010.

Hyvarinen, A., Sasaki, H., and Turner, R. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *AISTATS*, pp. 859–868, 2019.

Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, number 8, pp. 1965–1972, 2017.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, 2020a.

Khemakhem, I., Monti, R., Kingma, D., and Hyvarinen, A. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12768–12778, 2020b.

Kim, H. and Mnih, A. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2649–2658, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Kivva, B., Rajendran, G., Ravikumar, P., and Aragam, B. Learning latent causal graphs via mixture oracles. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18087–18101, 2021.

Kivva, B., Rajendran, G., Ravikumar, P., and Aragam, B. Identifiability of deep generative models without auxiliary information. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15687–15701, 2022.

Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021.

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., Kembhavi, A., Gupta, A., and Farhadi, A. Ai2-thor: An interactive 3d environment for visual ai. *arXiv*, 2022.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 366–374, 2008.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., PRIOL, R. L., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177, pp. 428–484, 2022.

Leeb, F., Lanzillotta, G., Annadani, Y., Besserve, M., Bauer, S., and Schölkopf, B. Structure by architecture: Disentangled representations without regularization. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

Li, Y., Torralba, A., Anandkumar, A., Fox, D., and Garg, A. Causal discovery in physical systems from videos. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9180–9192, 2020.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. CITRIS: Causal identifiability from temporal intervened sequences. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023.

Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., van den Hengel, A., Zhang, K., and Shi, J. Q. Identifying weight-variant latent causal models. arXiv, 2023.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4114–4124, 2019.

Locatello, F., Poole, B., Raetsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6348–6359, 2020.

Longman, R. J., Giambelluca, T. W., Nullet, M. A., Frazier, A. G., Kodama, K., Crausbay, S. D., Krushelnycky, P. D., Cordell, S., Clark, M. P., Newman, A. J., and Arnold, J. R. Compilation of climate data from heterogeneous networks across the Hawaiian islands. *Scientific Data*, 5 (1):180012, 2018.

Lyu, Q., Fu, X., Wang, W., and Lu, S. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022.

Maeda, T. N. and Shimizu, S. RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 735–745, 2020.

Monti, R. P., Zhang, K., and Hyvärinen, A. Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 186–195, 2020.

Moran, G. E., Sridhar, D., Wang, Y., and Blei, D. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2021.

Morioka, H. and Hyvarinen, A. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 3399–3426, 2023.

Morioka, H., Calhoun, V., and Hyvärinen, A. Nonlinear ica of fmri reveals primitive temporal structures linked to rest, task, and behavioral traits. In 218 (ed.), *NeuroImage*, pp. 116989, 2020.

Morioka, H., Hälvä, H., and Hyvarinen, A. Independent innovation analysis for nonlinear vector autoregressive process. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1549–1557, 2021.

Munkres, J. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Ng, I., Ghassami, A., and Zhang, K. On the role of sparsity and dag constraints for learning linear dags. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17943–17954, 2020.

Park, G. and Park, H. Identifiability of generalized hypergeometric distribution (GHD) directed acyclic graphical models. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 158–166, 2019a.

Park, G. and Park, S. High-dimensional Poisson structural equation model learning via ℓ_1-regularized regression. *Journal of Machine Learning Research*, 20(95): 1–41, 2019b.

Park, G. and Raskutti, G. Learning large-scale Poisson DAG models based on overdispersion scoring. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

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

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated dag! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27772–27784, 2021.

Roeder, G., Metz, L., and Kingma, D. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9030–9039, 2021.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Towards causal representation learning. arXiv, 2021.

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

Shimizu, S. and Bollen, K. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions. *Journal of Machine Learning Research*, 15(76):2629–2652, 2014.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011.

Sorrenson, P., Rother, C., and Köthe, U. Disentanglement by nonlinear ica with general incompressible-flow networks (GIN). In *International Conference on Learning Representations*, 2020.

Spirtes, P. and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, 2001.

Sprekeler, H., Zito, T., and Wiskott, L. An extension of slow feature analysis for nonlinear blind source separation. *Journal of Machine Learning Research*, 15(26):921–947, 2014.

Squires, C., Seigal, A., Bhate, S., and Uhler, C. Linear causal disentanglement via interventions. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

Sturma, N., Squires, C., Drton, M., and Uhler, C. Unpaired multi-domain causal representation learning. arXiv, 2023.

Varici, B., Acarturk, E., Shanmugam, K., Kumar, A., and Tajer, A. Score-based causal representation learning with interventions. arXiv, 2023.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467, 2021.

von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions. arXiv, 2023.

Wang, Y., Blei, D., and Cunningham, J. P. Posterior collapse and latent variable non-identifiability. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5443–5455, 2021.

Willetts, M. and Paige, B. I don't need u: Identifiable non-linear ica without side information. arXiv, 2022.

Wu, P. and Fukumizu, K. Causal mosaic: Cause-effect inference via nonlinear ICA and ensemble method. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1157–1167, 2020.

Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., and Zhang, K. Generalized independent noise condition for estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14891–14902, 2020.

Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., and Zhang, K. Identification of linear non-Gaussian latent hierarchical structure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 24370–24387, 2022.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9593–9602, 2021.

Yang, X., Wang, Y., Sun, J., Zhang, X., Zhang, S., Li, Z., and Yan, J. Nonlinear ICA using volume-preserving transformations. In *International Conference on Learning Representations*, 2022.

Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., von Kügelgen, J., and Locatello, F. Multi-view causal representation learning with partial observability. arXiv, 2023.

Yao, W., Chen, G., and Zhang, K. Learning latent causal dynamics. arXiv, 2022a.

Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 647–655, 2009.

Zhang, K. and Hyvärinen, A. Source separation and higher-order causal analysis of MEG and EEG. In *Proc. 26th Conference on Uncertainty in Artificial Intelligence (UAI2010)*, Catalina Island, California, 2010.

Zheng, X., Aragam, B., K, P. R., and Xing, E. P. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 3414–3425, 2020.

Zheng, Y. and Zhang, K. Generalizing nonlinear ICA beyond structural sparsity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zheng, Y., Ng, I., and Zhang, K. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12979–12990, 2021.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 10(476):1418–1429, 2006.

A. Supplementary Materials for *Causal Representation Learning Made Identifiable by Grouping of Observational Variables*

B. Exponential Family Bayesian Network Representation of the Causal Model

We consider a special case where the potential function ϕ is factorized as

$$\phi(x, y) = \boldsymbol{\eta}(x)^{\mathsf{T}} \mathbf{T}(y), \tag{8}$$

where the factors $\boldsymbol{\eta}(x) = [\eta_1(x), \ldots, \eta_N(x)]^{\mathsf{T}}$ and $\mathbf{T}(y) = [T_1(y), \ldots, T_N(y)]^{\mathsf{T}}$ are some N-dimensional vector functions of a scalar input, which are differentiable. We assume that the factors $\boldsymbol{\eta}$ and $\mathbf{T}$ are minimal without loss of generality, whose definition is given below (Definition 1). In this parameterization, if the whole causal graph is acyclic and the intra-group causal relations are given in the pairwise form as in those of inter-groups, the conditional distribution of a variable s_a^m is given, from Eqs. 3, by the following (conditional) exponential family of order N,

$$\begin{aligned} p(s_a^m|\mathrm{pa}(s_a^m)) &= \frac{1}{Z_a^m(\mathrm{pa}(s_a^m))} h_a^m(s_a^m) \exp\left(\sum_{s_b^{m'} \in \mathrm{pa}(s_a^m)} \lambda_{ba}^{m'm} \boldsymbol{\eta}(s_b^{m'})^{\mathsf{T}} \mathbf{T}(s_a^m) \right), \\ &= \frac{1}{Z_a^m(\mathrm{pa}(s_a^m))} h_a^m(s_a^m) \exp\left(\tilde{\boldsymbol{\eta}}_a^m(\mathrm{pa}(s_a^m))^{\mathsf{T}} \mathbf{T}(s_a^m) \right) \end{aligned} \tag{9}$$

where $\mathrm{pa}(s_a^m)$ is the set of parents of the variable s_a^m, $\mathbf{T}(s_a^m)$ represents the sufficient statistic of the conditional distribution of s_a^m. The overall natural parameter $\tilde{\boldsymbol{\eta}}_a^m(\mathrm{pa}(s_a^m)) = \sum_{s_b^{m'} \in \mathrm{pa}(s_a^m)} \lambda_{ba}^{m'm} \boldsymbol{\eta}(s_b^{m'})$ is simply given as a summation of the causal effects from the all parents, depending on the causal strengths $\lambda_{ba}^{m'm}$ and the function $\boldsymbol{\eta}$. The base measure h_a^m and the partition function Z_a^m depend on the type of the factors and the graph structure. The crucial point is the additivity of the (nonlinear) causal effects from the parents, which determines the natural parameters of the target variable distribution. Apart from that, this parameterization is not very restrictive, since exponential families have universal approximation capabilities (Sriperumbudur et al., 2017).

This parameterization also simplifies some of the assumptions of Theorems. The non-factorizability of ϕ (Assumption A2 of Theorem 1) can be satisfied if the model order N is more than one, as shown in Lemma 1 given below. Although this exponential family parameterization with order $N = 1$ cannot satisfy the non-factorizability, they can be supported by the other variant of the identifiability condition (Proposition 1), which does not require such non-factorizability. The asymmetricity assumption of ϕ (A2 and C2) indicates that the sets of the elements (functions) need to be sufficiently different between $\boldsymbol{\eta}$ and $\mathbf{T}$.

Some SEMs can be represented by Eq. 9 We can also show that some state-equation models (SEMs) can be represented by this parameterization as a special case. One such example is causal additive models (CAMs; Bühlmann et al. (2014)), given by

$$\mathbf{s} = \mathbf{L}\boldsymbol{\beta}(\mathbf{s}) + \boldsymbol{\epsilon}, \tag{10}$$

where $\mathbf{L} \in \mathbb{R}^{D_{\mathcal{S}} \times D_{\mathcal{S}}}$ is an adjacency matrix, $\boldsymbol{\beta}(\mathbf{s}) = [\beta(s_1), \ldots, \beta(s_{D_{\mathcal{S}}})]^{\mathsf{T}}$ is an element-wise (nonlinear) function of $\mathbf{s}$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma\mathbf{I})$ is $D_{\mathcal{S}}$-dimensional additive Gaussian noise with diagonal covariance matrix. In this SEM, the conditional distribution of a variable s_a^m is given by

$$p(s_a^m|\mathrm{pa}(s_a^m)) = \frac{1}{Z} \exp\left\{ -\frac{1}{2\sigma^2} \left(s_a^m - \sum_{s_b^{m'} \in \mathrm{pa}(s_a^m)} \lambda_{ba}^{m'm} \beta(s_b^{m'}) \right)^2 \right\} \tag{11}$$

$$= \left[\frac{1}{Z} \exp\left(-\frac{\left(\sum_{s_b^{m'} \in \mathrm{pa}(s_a^m)} \lambda_{ba}^{m'm} \beta(s_b^{m'}) \right)^2}{2\sigma^2} \right) \right] \left[\exp\left(-\frac{(s_a^m)^2}{2\sigma^2} \right) \right]$$

$$\times \left[\exp\left(\sum_{s_b^{m'} \in \mathrm{pa}(s_a^m)} \frac{\lambda_{ba}^{m'm}}{\sigma^2} s_a^m \beta(s_b^{m'}) \right) \right], \tag{12}$$

where Z denotes a normalizing constant. We can clearly see that the first, the second, and the third factors correspond to those of Eq. 9, respectively, and the causal function ϕ of this model is given by $\phi(x, y) = y\beta(x)$. Therefore, the CAMs given by Eq. 10 can be represented by our causal model (Eqs. 3), especially by the exponential family parameterization (Eqs. 8 and 9) of order $N = 1$, linear sufficient statistics $\mathbf{T}(y) = y$, and (nonlinear) natural parameter $\boldsymbol{\eta}(x) = \beta(x)$. As mentioned above, such model with order $N = 1$ cannot satisfy the non-factorizability (A2), and thus needs to consider Proposition 1 instead of Theorem 1.

This model includes linear Gaussian SEMs as a special case, where β is a linear function. However, they do not satisfy the conditions of both Theorem 1 (uniform-dependency, non-factorizability, and asymmetricity) and Proposition 1 (uniform-dependency and asymmetricity), where the definition of uniform-dependency is given below (Definition 2). Thus linear Gaussian SEMs cannot have identifiability in our model in any case, which is consistent with the well-known result of causal discovery (Hoyer et al., 2008a; Peters et al., 2014).

Note that the exponential family BNs (Eq. 9) can represent many other causal models in addition to CAMs; crucially, the distributional type of Eq. 9 is not restricted to Gaussian (more specifically, the sufficient statistic $\mathbf{T}$ can be nonlinear and multidimensional), unlike many conventional causal models based on additive Gaussian error terms (e.g., Eq. 10).

Definition 1. *(Minimality) We say that a function $\boldsymbol{\alpha} : \mathbb{R} \to \mathbb{R}^N$ is minimal if for any open subset $\mathcal{X}$ of $\mathbb{R}$ the following is true:*

$$\left(\exists \boldsymbol{\theta} \in \mathbb{R}^N \mid \forall x \in \mathcal{X}, \boldsymbol{\theta}^\mathsf{T} \boldsymbol{\alpha}(x) = \textit{const.}\right) \implies \boldsymbol{\theta} = \mathbf{0}. \tag{13}$$

The minimality is similar to the linear independence of the elements, but stronger; minimality also forbids the existence of elements which only have differences of scaling and biases. Note that a non-minimal model can always be reduced to minimal one via a suitable transformation and reparameterization.

Definition 2. *(Uniform-dependency) We call a function $q(x, y) : \bar{\mathcal{S}} \times \bar{\mathcal{S}} \to \mathbb{R}$ is uniform-dependent if the set of zeros of $q(x, y)$ is a meagre subset of $\bar{\mathcal{S}} \times \bar{\mathcal{S}}$, i.e., it contains no open subset.*

Lemma 1. *In the exponential family characterization of the causal model (Eq. 8), the non-factorizability conditions in Assumption A2 can be satisfied if the model order $N \geq 2$.*

Proof. (Non-factorizability) We firstly show the non-factorizability condition of the first equation $\phi^{12}(s, z_1) = c_1\phi^{12}(s, z_2)$ in Assumption A2; the second equation can be proven in the same manner. We give a proof by contradiction. We suppose the negation; there exist some open subset $B \subset \bar{\mathcal{S}}$ such that the equation $\phi^{12}(s, z_1) = c_1\phi^{12}(s, z_2)$ hold for all $s \in B$ for any $z_1 \neq z_2$. We consider one of such open subset B here. By substituting Eq. 8 into the equation, we have

$$\boldsymbol{\eta}'(s)^\mathsf{T} \left(\mathbf{T}'(z_1) - c_1\mathbf{T}'(z_2)\right) = 0. \tag{14}$$

From Lemma 3 of Khemakhem et al. (2020a), there exist N distinctive values s_1 to s_N such that $(\boldsymbol{\eta}'(s_1), \ldots, \boldsymbol{\eta}'(s_N))$ are linearly independent. By substituting those values into Eq. 14 with concatenating vertically, we obtain

$$\begin{bmatrix} \boldsymbol{\eta}'(s_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\eta}'(s_N)^\mathsf{T} \end{bmatrix} \left(\mathbf{T}'(z_1) - c_1\mathbf{T}'(z_2)\right) = \mathbf{0}. \tag{15}$$

Since the first factor ($N \times N$ matrix) has full-rank, we have

$$\mathbf{T}'(z_1) - c_1\mathbf{T}'(z_2) = \mathbf{0}. \tag{16}$$

However, this contradicts the fact that there should exist at leatst two distinctive values z_1 and z_2 such that $(\mathbf{T}'(z_1), \mathbf{T}'(z_2))$ are linearly independent, again from Lemma 3 of Khemakhem et al. (2020a). From this contradiction, we conclude that we can make the equation not hold by properly choosing some z_1 and z_2, which indicates the non-factorizability.

□

C. Proof of Theorem 1

Proof. We denote by $\mathcal{S} = \mathcal{S}^1 \times \ldots \times \mathcal{S}^M$ the support of the distribution of $\mathbf{s}$, where $\mathcal{S}^m = \mathcal{S}^m_1 \times \ldots \times \mathcal{S}^m_{d^m_\mathcal{S}}$ is that of the distribution of each group $\mathbf{s}^m$, and $\mathcal{S}^m_a \subset \mathbb{R}$ is that of the a-th element. We consider the situation where each $\mathcal{S}^m_a$ is connected (i.e. an interval), and additionally, without loss of generality, $\mathcal{S}^m_a$ are the same across all variables, denoted as $\bar{\mathcal{S}}$.

Writing the joint log-density of the random vector $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^M)$ for the two parameterizations, yields

$$\begin{aligned}
&\log p\left(\mathbf{g}^1(\mathbf{x}^1), \ldots, \mathbf{g}^M(\mathbf{x}^M)\right) + \sum_{m\in\mathcal{M}} \log|\det \mathbf{J}_{\mathbf{g}^m}(\mathbf{x}^m)| \\
&= \log \tilde{p}\left(\tilde{\mathbf{g}}^1(\mathbf{x}^1), \ldots, \tilde{\mathbf{g}}^M(\mathbf{x}^M)\right) + \sum_{m\in\mathcal{M}} \log|\det \mathbf{J}_{\tilde{\mathbf{g}}^m}(\mathbf{x}^m)|,
\end{aligned} \tag{17}$$

where we denote the (true) demixing models as $\mathbf{g}^m = (\mathbf{f}^m)^{-1}$, and their other parameterizations as $\tilde{\mathbf{g}}^m$, and $\mathbf{J}$ is the Jacobian for the change of variables. Let a compound demixing-mixing function $\mathbf{v}^m(\mathbf{s}^m) = \tilde{\mathbf{g}}^m \circ \mathbf{f}^m(\mathbf{s}^m)$, we then have

$$\begin{aligned}
&\log p\left(\mathbf{s}^1, \ldots, \mathbf{s}^M\right) + \sum_{m\in\mathcal{M}} \log|\det \mathbf{J}_{\mathbf{g}^m}(\mathbf{f}^m(\mathbf{s}^m))| \\
&= \log \tilde{p}\left(\mathbf{v}^1(\mathbf{s}^1), \ldots, \mathbf{v}^M(\mathbf{s}^M)\right) + \sum_{m\in\mathcal{M}} \log|\det \mathbf{J}_{\tilde{\mathbf{g}}^m}(\mathbf{f}^m(\mathbf{s}^m))|.
\end{aligned} \tag{18}$$

We substitute the factorization model Eq. 3 into this, and differentiate the both sides with respect to s^m_a and $s^{m'}_b$, where $a \in \mathcal{V}^m_\mathcal{S}$, $b \in \mathcal{V}^{m'}_\mathcal{S}$, and obtain

$$\begin{aligned}
&\frac{\partial^2}{\partial s^m_a \partial s^{m'}_b}\left(\lambda^{mm'}_{ab}\phi(s^m_a, s^{m'}_b) + \lambda^{m'm}_{ba}\phi(s^{m'}_b, s^m_a)\right) \\
&= \frac{\partial^2}{\partial s^m_a \partial s^{m'}_b}\sum_{(i,j)}\left(\tilde{\lambda}^{mm'}_{ij}\tilde{\phi}\left(v^m_i(\mathbf{s}^m), v^{m'}_j(\mathbf{s}^{m'})\right) + \tilde{\lambda}^{m'm}_{ji}\tilde{\phi}\left(v^{m'}_j(\mathbf{s}^{m'}), v^m_i(\mathbf{s}^m)\right)\right).
\end{aligned} \tag{19}$$

The Jacobians disappeared here due to the grouped-observational assumption and the cross-derivatives.

By collecting the cross-derivatives for all $a \in \mathcal{V}^m_\mathcal{S}$ and $b \in \mathcal{V}^{m'}_\mathcal{S}$, with a giving row index and b the column index, we have a matrix equation of the size $d^m_\mathcal{S} \times d^{m'}_\mathcal{S}$,

$$\begin{aligned}
&\left(\frac{\partial^2}{\partial s^m_a \partial s^{m'}_b}\left(\lambda^{mm'}_{ab}\phi(s^m_a, s^{m'}_b) + \lambda^{m'm}_{ba}\phi(s^{m'}_b, s^m_a)\right)\right)_{a\in\mathcal{V}^m_\mathcal{S}, b\in\mathcal{V}^{m'}_\mathcal{S}} \\
&= \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \\
&\cdot \left(\frac{\partial^2}{\partial v^m_a \partial v^{m'}_b}\left(\tilde{\lambda}^{mm'}_{ab}\tilde{\phi}\left(v^m_a(\mathbf{s}^m), v^{m'}_b(\mathbf{s}^{m'})\right) + \tilde{\lambda}^{m'm}_{ba}\tilde{\phi}\left(v^{m'}_b(\mathbf{s}^{m'}), v^m_a(\mathbf{s}^m)\right)\right)\right)_{a\in\mathcal{V}^m_\mathcal{S}, b\in\mathcal{V}^{m'}_\mathcal{S}} \\
&\cdot \mathbf{J}_{\mathbf{v}^{m'}}(\mathbf{s}^{m'}).
\end{aligned} \tag{20}$$

We then focus on the a-th row of Eq. 20, and differentiate each element of the both sides with respect to $s^m_{a'}$, $a' \neq a$. Concatenating it horizontally with substituting some K^m vectors $\{\mathbf{z}^{m_k}_k\}_{k=1}^{K^m}$ into $\mathbf{s}^{m'}$, each of which is on some group $m_k \neq m$ with allowing repetitions, we have a vector equation of the size $1 \times \sum_{k=1}^{K^m} d^{m_k}_\mathcal{S}$

$$\begin{aligned}
\mathbf{0}^\mathsf{T} &= \left[(\mathbf{v}^m)^{a\times a'}(\mathbf{s}^m)^\mathsf{T}, \quad (\mathbf{v}^m)^{aa'}(\mathbf{s}^m)^\mathsf{T}\right] \\
&\cdot \left[\tilde{\mathbf{\Phi}}^{mm_1}(\mathbf{v}^m(\mathbf{s}^m), \mathbf{v}^{m_1}(\mathbf{z}^{m_1}_1)), \ldots, \tilde{\mathbf{\Phi}}^{mm_{K^m}}(\mathbf{v}^m(\mathbf{s}^m), \mathbf{v}^{m_{K^m}}(\mathbf{z}^{m_{K^m}}_{K^m})\right] \\
&\cdot \begin{bmatrix} \mathbf{J}_{\mathbf{v}^{m_1}}(\mathbf{z}^{m_1}_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{\mathbf{v}^{m_{K^m}}}(\mathbf{z}^{m_{K^m}}_{K^m}) \end{bmatrix},
\end{aligned} \tag{21}$$

where $(\mathbf{v}^m)^{a\times a'}(\mathbf{s}^m) = \left[\frac{\partial}{\partial s_a^m} v_1^m(\mathbf{s}^m)\frac{\partial}{\partial s_{a'}^m} v_1^m(\mathbf{s}^m), \quad \ldots, \quad \frac{\partial}{\partial s_a^m} v_{d_\mathcal{S}^m}^m(\mathbf{s}^m)\frac{\partial}{\partial s_{a'}^m} v_{d_\mathcal{S}^m}^m(\mathbf{s}^m)\right]^\mathsf{T}$ and $(\mathbf{v}^m)^{aa'}(\mathbf{s}^m) = \left[\frac{\partial^2}{\partial s_a^m \partial s_{a'}^m} v_1^m(\mathbf{s}^m), \quad \ldots, \quad \frac{\partial^2}{\partial s_a^m \partial s_{a'}^m} v_{d_\mathcal{S}^m}^m(\mathbf{s}^m)\right]^\mathsf{T}$ are $d_\mathcal{S}^m$ dimensional vectors, and $\tilde{\mathbf{\Phi}}^{mm_k}(\mathbf{y}^m, \mathbf{y}^{m_k})$ is a $2d_\mathcal{S}^m \times d_\mathcal{S}^{m_k}$ matrix given as a collection of cross-derivatives of $\tilde{\phi}$,

$$\tilde{\mathbf{\Phi}}^{mm_k}(\mathbf{y}^m, \mathbf{y}^{m_k}) = \begin{bmatrix} \left(\frac{\partial^3}{\partial y_a^{m2}\partial y_b^{m_k}}\left(\tilde{\lambda}_{ab}^{mm_k}\tilde{\phi}(y_a^m, y_b^{m_k}) + \tilde{\lambda}_{ba}^{m_k m}\tilde{\phi}(y_b^{m_k}, y_a^m)\right)\right)_{a\in\mathcal{V}_\mathcal{S}^m, b\in\mathcal{V}_\mathcal{S}^{m_k}} \\ \left(\frac{\partial^2}{\partial y_a^m \partial y_b^{m_k}}\left(\tilde{\lambda}_{ab}^{mm_k}\tilde{\phi}(y_a^m, y_b^{m_k}) + \tilde{\lambda}_{ba}^{m_k m}\tilde{\phi}(y_b^{m_k}, y_a^m)\right)\right)_{a\in\mathcal{V}_\mathcal{S}^m, b\in\mathcal{V}_\mathcal{S}^{m_k}} \end{bmatrix}. \tag{22}$$

The left-hand-side is now a zero-vector due to the pair-wise factorization assumption of the joint distribution (Eq. 3).

We show here the second factor on the right-hand side of Eq. 21 has full row-rank for all $\mathbf{s}^m \in A$ in any open subset A of $\mathcal{S}^m$, by properly choosing the K^m vectors $\{\mathbf{z}_k^{m_k}\}_{k=1}^{K^m}$, due to the assumptions. We differentiate each of a-th row of the both sides of Eq. 20 with respect to s_a^m again, and get

$$\begin{aligned} &\left(\frac{\partial^3}{\partial s_a^{m2}\partial s_b^{m'}}\left(\lambda_{ab}^{mm'}\phi(s_a^m, s_b^{m'}) + \lambda_{ba}^{m'm}\phi(s_b^{m'}, s_a^m)\right)\right)_{a\in\mathcal{V}_\mathcal{S}^m, b\in\mathcal{V}_\mathcal{S}^{m'}} \\ &= \left[\mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \circ \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T}, \quad \mathbf{J}^*_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T}\right] \tilde{\mathbf{\Phi}}^{mm'}\left(\mathbf{v}^m(\mathbf{s}^m), \mathbf{v}^{m'}(\mathbf{s}^{m'})\right)\mathbf{J}_{\mathbf{v}^{m'}}(\mathbf{s}^{m'}), \end{aligned} \tag{23}$$

where $\circ$ is Hadamard product, $\mathbf{J}^*_{\mathbf{v}^m}(\mathbf{s}^m) = (\frac{\partial^2}{\partial s_j^2} v_i^m(\mathbf{s}^m))_{i,j}$ is the row-wise derivatives of the Jacobian $\mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m) = (\frac{\partial}{\partial s_j} v_i^m(\mathbf{s}^m))_{i,j}$, and $\tilde{\mathbf{\Phi}}^{mm'}$ is given by Eq. 22. Since Eqs. 20 and 23 have some common factors, we can concatenate Eqs. 20 and 23 vertically, and represent them as a single matrix equation of the size $2d_\mathcal{S}^m \times d_\mathcal{S}^{m'}$,

$$\begin{aligned} &\mathbf{\Phi}^{mm'}\left(\mathbf{s}^m, \mathbf{s}^{m'}\right) \\ &= \begin{bmatrix} \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \circ \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} & \mathbf{J}^*_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \\ \mathbf{0} & \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \end{bmatrix} \tilde{\mathbf{\Phi}}^{mm'}\left(\mathbf{v}^m(\mathbf{s}^m), \mathbf{v}^{m'}(\mathbf{s}^{m'})\right)\mathbf{J}_{\mathbf{v}^{m'}}(\mathbf{s}^{m'}), \end{aligned} \tag{24}$$

where $\mathbf{\Phi}^{mm'}\left(\mathbf{s}^m, \mathbf{s}^{m'}\right)$ is a $2d_\mathcal{S}^m \times d_\mathcal{S}^{m'}$ matrix, which has the same form as Eq. 22 and is given by

$$\mathbf{\Phi}^{mm'}(\mathbf{y}^m, \mathbf{y}^{m'}) = \begin{bmatrix} \left(\frac{\partial^3}{\partial y_a^{m2}\partial y_b^{m'}}\left(\lambda_{ab}^{mm'}\phi(y_a^m, y_b^{m'}) + \lambda_{ba}^{m'm}\phi(y_b^{m'}, y_a^m)\right)\right)_{a\in\mathcal{V}_\mathcal{S}^m, b\in\mathcal{V}_\mathcal{S}^{m'}} \\ \left(\frac{\partial^2}{\partial y_a^m \partial y_b^{m'}}\left(\lambda_{ab}^{mm'}\phi(y_a^m, y_b^{m'}) + \lambda_{ba}^{m'm}\phi(y_b^{m'}, y_a^m)\right)\right)_{a\in\mathcal{V}_\mathcal{S}^m, b\in\mathcal{V}_\mathcal{S}^{m'}} \end{bmatrix}. \tag{25}$$

We concatenate Eq. 24 horizontally with substituting the same vectors $\{\mathbf{z}_k^{m_k}\}_{k=1}^{K^m}$ used above into $\mathbf{s}^{m'}$, then get a matrix equation of the size $2d_\mathcal{S}^m \times \sum_{k=1}^{K^m} d_\mathcal{S}^{m_k}$

$$\begin{aligned} &\left[\mathbf{\Phi}^{mm_1}(\mathbf{s}^m, \mathbf{z}_1^{m_1}), \ldots, \mathbf{\Phi}^{mm_{K^m}}(\mathbf{s}^m, \mathbf{z}_{K^m}^{m_{K^m}})\right] \\ &= \begin{bmatrix} \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \circ \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} & \mathbf{J}^*_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \\ \mathbf{0} & \mathbf{J}_{\mathbf{v}^m}(\mathbf{s}^m)^\mathsf{T} \end{bmatrix} \\ &\cdot \left[\tilde{\mathbf{\Phi}}^{mm_1}(\mathbf{v}^m(\mathbf{s}^m), \mathbf{v}^{m_1}(\mathbf{z}_1^{m_1})), \ldots, \tilde{\mathbf{\Phi}}^{mm_{K^m}}(\mathbf{v}^m(\mathbf{s}^m), \mathbf{v}^{m_{K^m}}(\mathbf{z}_{K^m}^{m_{K^m}})\right] \\ &\cdot \begin{bmatrix} \mathbf{J}_{\mathbf{v}^{m_1}}(\mathbf{z}_1^{m_1}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{\mathbf{v}^{m_{K^m}}}(\mathbf{z}_{K^m}^{m_{K^m}}) \end{bmatrix}. \end{aligned} \tag{26}$$

From Lemma 2 given below, we can choose the vectors $\{\mathbf{z}_k^{m_k}\}_{k=1}^{K^m}$ so as to make the left-hand side has full row-rank $2d_\mathcal{S}^m$ for all $\mathbf{s}^m$ in any open subset of $\mathcal{S}^m$ based on the assumptions, which implies that the second factor (the concatenation of $\tilde{\mathbf{\Phi}}^{mm_k}$) in the right-hand side has full row-rank $2d_\mathcal{S}^m$ as well. Therefore, the second factor on the right-hand side of Eq. 21 has full row-rank.

Since the last term of Eq. 21 (collection of Jacobians) has full rank because all $\mathbf{v}^m$ are invertible, we can multiply the both sides by its inverse. In addition, since the second factor of Eq. 21 has full row-rank due to the discussion above, we can multiply the both sides of Eq. 21 with its pseudo-inverse from the right side, and finally get

$$\left[(\mathbf{v}^m)^{a\times a'}(\mathbf{s}^m)^\mathsf{T}, \quad (\mathbf{v}^m)^{aa'}(\mathbf{s}^m)^\mathsf{T}\right] = \mathbf{0}^\mathsf{T}, \tag{27}$$

which is true for the all combinations of $a \neq a' \in \mathcal{V}_\mathcal{S}^m$. This particularly indicates that, $\frac{\partial}{\partial s_a^m} v_j^m(\mathbf{s}^m) \cdot \frac{\partial}{\partial s_{a'}^m} v_j^m(\mathbf{s}^m) = 0$ for all $1 \leq j \leq d_\mathcal{S}^m$, $a \neq a'$. This means that the Jacobian of $\mathbf{v}^m$ has at most one non-zero entry in each row. Now, by invertibility and continuity of $\mathbf{J}_{\mathbf{v}^m}$, we deduce that the location of the non-zero entries are fixed and do not change as a function of $\mathbf{s}^m$. This proves that $v_a^m(\mathbf{s}^m)$ is represented by only one variable $s_{\sigma^m(a)}^m$ up to a scalar (variable-specific) invertible transformation for each $a \in \mathcal{V}_\mathcal{S}^m$, where $\sigma^m(a) : \mathcal{V}_\mathcal{S}^m \to \mathcal{V}_\mathcal{S}^m$ represents a permutation of variables, which is indeterminate, and the Theorem is proven. □

Lemma 2. *With assumptions of Theorem 1, we have the following for all group m satisfying A1: For any open subset A of $(\bar{\mathcal{S}})^{d_\mathcal{S}^m}$, there exists a set of $K^m \geq 1$ vectors $\{\mathbf{z}_k \in (\bar{\mathcal{S}})^{d_\mathcal{S}^{m_k}}\}_{k=1}^{K^m}$, each of which belongs to some other group $m_k \neq m$ with allowing repetitions, such that the concatenated matrix*

$$\left[\mathbf{\Phi}^{mm_1}(\mathbf{s}^m, \mathbf{z}_1), \quad \ldots, \quad \mathbf{\Phi}^{mm_{K^m}}(\mathbf{s}^m, \mathbf{z}_{K^m})\right] \tag{28}$$

with the size $2d_\mathcal{S}^m \times \sum_{k=1}^{K^m} d_\mathcal{S}^{m_k}$ has full row-rank $2d_\mathcal{S}^m$ for all $\mathbf{s}^m$ in A.

Proof. To show that there indeed exists such a set of vectors, we especially select here the groups $\{m_k\}$ as $\mathcal{M} \setminus m$ repeating twice, with some specific values of $\mathbf{z}_k$ for each; more specifically, $\left[m_1, \quad \ldots, \quad m_{K^m}\right] = \left[1, \ldots, \quad m-1, \quad m+1, \ldots, \quad M, \quad 1, \ldots, \quad m-1, \quad m+1, \ldots, \quad M\right]$, $K^m = 2(M-1)$, and $\mathbf{z}_k = \mathbf{z}_1 = [z_1, \ldots, z_1]^\mathsf{T}$ for the first half $k = 1, \ldots, M-1$, and $\mathbf{z}_k = \mathbf{z}_2 = [z_2, \ldots, z_2]^\mathsf{T}$ for the second half $k = M, \ldots, 2(M-1)$ with some z_1 and $z_2 \in \bar{\mathcal{S}}$ (note that the size of those vectors can be different across k). We denote a collection of the all inter-group adjacency coefficients related to the group m as

$$\bar{\mathbf{L}}^m = \begin{bmatrix}\mathbf{L}^{m:} \\ \mathbf{L}^{:m}\end{bmatrix} = \begin{bmatrix}\mathbf{L}^{m1}, & \ldots, & \mathbf{L}^{mM} \\ (\mathbf{L}^{1m})^\mathsf{T}, & \ldots, & (\mathbf{L}^{Mm})^\mathsf{T}\end{bmatrix}, \tag{29}$$

which is a $2d_\mathcal{S}^m \times \sum_{m' \in \mathcal{M}\setminus m} d_\mathcal{S}^{m'}$ matrix given in Assumption A1, where $\mathbf{L}^{m:}$ and $\mathbf{L}^{:m}$ denote upper and lower-half matrices of $\bar{\mathbf{L}}^m$, corresponding to the adjacency coefficients from group m to the other groups, and those from the other groups to the group m, respectively. Substituting those values into Eq. 28, we obtain a $2d_\mathcal{S}^m \times 2\sum_{m' \in \mathcal{M}\setminus m} d_\mathcal{S}^{m'}$ matrix with a factorized form

$$\begin{aligned}&\left[\mathbf{\Phi}^{m1}(\mathbf{s}^m, \mathbf{z}_1), \quad \ldots, \quad \mathbf{\Phi}^{mM}(\mathbf{s}^m, \mathbf{z}_1), \quad \mathbf{\Phi}^{m1}(\mathbf{s}^m, \mathbf{z}_2), \quad \ldots, \quad \mathbf{\Phi}^{mM}(\mathbf{s}^m, \mathbf{z}_2)\right] \\ &= \begin{bmatrix}\mathbf{B}^{112}(\mathbf{s}^m, z_1), & \mathbf{B}^{122}(z_1, \mathbf{s}^m), & \mathbf{B}^{112}(\mathbf{s}^m, z_2), & \mathbf{B}^{122}(z_2, \mathbf{s}^m) \\ \mathbf{B}^{12}(\mathbf{s}^m, z_1), & \mathbf{B}^{12}(z_1, \mathbf{s}^m), & \mathbf{B}^{12}(\mathbf{s}^m, z_2), & \mathbf{B}^{12}(z_2, \mathbf{s}^m)\end{bmatrix} \begin{bmatrix}\mathbf{L}^{m:}, & \mathbf{0} \\ \mathbf{L}^{:m}, & \mathbf{0} \\ \mathbf{0}, & \mathbf{L}^{m:} \\ \mathbf{0}, & \mathbf{L}^{:m}\end{bmatrix},\end{aligned} \tag{30}$$

where $\mathbf{B}^{112}(\mathbf{s}^m, z) = \operatorname{diag}\left(\phi^{112}(s_a^m, z)\right)_{a\in\mathcal{V}_\mathcal{S}^m}$, $\mathbf{B}^{122}(z, \mathbf{s}^m) = \operatorname{diag}\left(\phi^{122}(z, s_a^m)\right)_{a\in\mathcal{V}_\mathcal{S}^m}$, $\mathbf{B}^{12}(\mathbf{s}^m, z) = \operatorname{diag}\left(\phi^{12}(s_a^m, z)\right)_{a\in\mathcal{V}_\mathcal{S}^m}$, and $\mathbf{B}^{12}(z, \mathbf{s}^m) = \operatorname{diag}\left(\phi^{12}(z, s_a^m)\right)_{a\in\mathcal{V}_\mathcal{S}^m}$. Note that those cross-derivatives of the function ϕ have uniform-dependency from Assumption A2 (Definition 2).

Considering that the adjacency matrix $\bar{\mathbf{L}}^m$ possibly has some rows with all-zeros, depending on the graph structure, we explicitly divide the set of latent variable indices $\mathcal{V}_\mathcal{S}^m$ into three groups $[\mathcal{V}_\mathrm{b}, \mathcal{V}_\mathrm{p}, \mathcal{V}_\mathrm{c}]$ (we omit the group index m for simplicity here); the variables with indices $\mathcal{V}_\mathrm{b}$ are both parents and children of some variables in some other group, the variables with $\mathcal{V}_\mathrm{p}$ are parents (but not children) of some variable in some other group, and the variables with $\mathcal{V}_\mathrm{c}$ are children (but not parents) of some variable in some other group. We assume without loss of generality that the variable indices $\mathcal{V}_\mathcal{S}^m$ are sorted

in the order $[\mathcal{V}_\mathrm{b}, \mathcal{V}_\mathrm{p}, \mathcal{V}_\mathrm{c}]$. Eq. 30 can then be re-written as

$$
\begin{aligned}
&\left[\boldsymbol{\Phi}^{m1}(\mathbf{s}^m, \mathbf{z}_1), \quad \ldots, \quad \boldsymbol{\Phi}^{mM}(\mathbf{s}^m, \mathbf{z}_1), \quad \boldsymbol{\Phi}^{m1}(\mathbf{s}^m, \mathbf{z}_2), \quad \ldots, \quad \boldsymbol{\Phi}^{mM}(\mathbf{s}^m, \mathbf{z}_2)\right] \\
&= \begin{bmatrix}
\mathbf{B}^{112}(\mathbf{s}^m_{\mathcal{V}_\mathrm{b}}, z_1), & \mathbf{0}, & \mathbf{B}^{122}(z_1, \mathbf{s}^m_{\mathcal{V}_\mathrm{b}}), & \mathbf{0}, \\
\mathbf{0}, & \mathbf{B}^{112}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_1), & \mathbf{0}, & \mathbf{0}, \\
\mathbf{0}, & \mathbf{0}, & \mathbf{0}, & \mathbf{B}^{122}(z_1, \mathbf{s}^m_{\mathcal{V}_\mathrm{c}}), \\
\mathbf{B}^{12}(\mathbf{s}^m_{\mathcal{V}_\mathrm{b}}, z_1), & \mathbf{0}, & \mathbf{B}^{12}(z_1, \mathbf{s}^m_{\mathcal{V}_\mathrm{b}}), & \mathbf{0}, \\
\mathbf{0}, & \mathbf{B}^{12}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_1), & \mathbf{0}, & \mathbf{0}, \\
\mathbf{0}, & \mathbf{0}, & \mathbf{0}, & \mathbf{B}^{12}(z_1, \mathbf{s}^m_{\mathcal{V}_\mathrm{c}}),
\end{bmatrix} \\
&\qquad \begin{bmatrix}
\mathbf{B}^{112}(\mathbf{s}^m_{\mathcal{V}_\mathrm{b}}, z_2), & \mathbf{0}, & \mathbf{B}^{122}(z_2, \mathbf{s}^m_{\mathcal{V}_\mathrm{b}}), & \mathbf{0} \\
\mathbf{0}, & \mathbf{B}^{112}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_2), & \mathbf{0}, & \mathbf{0} \\
\mathbf{0}, & \mathbf{0}, & \mathbf{0}, & \mathbf{B}^{122}(z_2, \mathbf{s}^m_{\mathcal{V}_\mathrm{c}}) \\
\mathbf{B}^{12}(\mathbf{s}^m_{\mathcal{V}_\mathrm{b}}, z_2), & \mathbf{0}, & \mathbf{B}^{12}(z_2, \mathbf{s}^m_{\mathcal{V}_\mathrm{b}}), & \mathbf{0} \\
\mathbf{0}, & \mathbf{B}^{12}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_2), & \mathbf{0}, & \mathbf{0} \\
\mathbf{0}, & \mathbf{0}, & \mathbf{0}, & \mathbf{B}^{12}(z_2, \mathbf{s}^m_{\mathcal{V}_\mathrm{c}})
\end{bmatrix}
\begin{bmatrix}
\mathbf{L}^{m:}_{\mathcal{V}_\mathrm{b}}, & \mathbf{0} \\
\mathbf{L}^{m:}_{\mathcal{V}_\mathrm{p}}, & \mathbf{0} \\
\mathbf{L}^{:m}_{\mathcal{V}_\mathrm{b}}, & \mathbf{0} \\
\mathbf{L}^{:m}_{\mathcal{V}_\mathrm{c}}, & \mathbf{0} \\
\mathbf{0}, & \mathbf{L}^{m:}_{\mathcal{V}_\mathrm{b}} \\
\mathbf{0}, & \mathbf{L}^{m:}_{\mathcal{V}_\mathrm{p}} \\
\mathbf{0}, & \mathbf{L}^{:m}_{\mathcal{V}_\mathrm{b}} \\
\mathbf{0}, & \mathbf{L}^{:m}_{\mathcal{V}_\mathrm{c}}
\end{bmatrix},
\end{aligned}
\tag{31}
$$

where $\mathbf{L}^{m:}_{\mathcal{V}_\mathrm{b}}$ denotes a submatrix of $\mathbf{L}^{m:}$ corresponding to the indices (rows) $\mathcal{V}_\mathrm{b}$, and similarly for $\mathbf{L}^{m:}_{\mathcal{V}_\mathrm{p}}$, $\mathbf{L}^{:m}_{\mathcal{V}_\mathrm{b}}$, and $\mathbf{L}^{:m}_{\mathcal{V}_\mathrm{c}}$. Now the second factor of the right-hand side has the size $2(2|\mathcal{V}_\mathrm{b}|+|\mathcal{V}_\mathrm{p}|+|\mathcal{V}_\mathrm{c}|)\times 2\sum_{m'\in\mathcal{M}\setminus m} d^{m'}_{\mathcal{S}}$, and has full row-rank from the assumption (note that the number of rows of this factor is lower than that in Eq. 30 since we removed all-zero rows). Since the number of rows of the first factor ($2d^m_{\mathcal{S}}$) is always smaller or equal to that of the second factor, $2(2|\mathcal{V}_\mathrm{b}|+|\mathcal{V}_\mathrm{p}|+|\mathcal{V}_\mathrm{c}|) \geq 2d^m_{\mathcal{S}}$, what we need to show for this Lemma is the full row-rankness of the first factor. From its structure, we can show this separately for each subset of its rows corresponding to $\mathcal{V}_\mathrm{b}$, $\mathcal{V}_\mathrm{p}$, and $\mathcal{V}_\mathrm{c}$.

The Rows Corresponding to $\mathcal{V}_\mathbf{b}$ We start from the submatrix (rows) corresponding to $\mathcal{V}_\mathrm{b}$. We especially consider the $2|\mathcal{V}_\mathrm{b}| \times 2|\mathcal{V}_\mathrm{b}|$ submatrix given by

$$
\begin{bmatrix}
\mathbf{B}^{112}(\mathbf{s}^m_{\mathcal{V}_\mathrm{b}}, z_1), & \mathbf{B}^{122}(z_1, \mathbf{s}^m_{\mathcal{V}_\mathrm{b}}) \\
\mathbf{B}^{12}(\mathbf{s}^m_{\mathcal{V}_\mathrm{b}}, z_1), & \mathbf{B}^{12}(z_1, \mathbf{s}^m_{\mathcal{V}_\mathrm{b}})
\end{bmatrix}.
\tag{32}
$$

If this submatrix has full-rank, the submatrix (rows) of the first factor of Eq. 31 corresponding to $\mathcal{V}_\mathrm{b}$ also has full row-rank. Considering that the matrices $\mathbf{B}^{112}$, $\mathbf{B}^{122}$, and $\mathbf{B}^{12}$ are all diagonal, we focus on a 2×2 submatrix corresponding to a variable index $a \in \mathcal{V}_\mathrm{b}$, given by

$$
\begin{bmatrix}
\phi^{112}(s^m_a, z_1) & \phi^{122}(z_1, s^m_a) \\
\phi^{12}(s^m_a, z_1) & \phi^{12}(z_1, s^m_a)
\end{bmatrix}.
\tag{33}
$$

Calculating the determinant, with the uniform dependency assumption of the all elements (Assumption A2), this submatrix has full-rank (non-zero determinant) if the following condition does not hold:

$$
\begin{aligned}
&\frac{\phi^{112}(s^m_a, z_1)}{\phi^{12}(s^m_a, z_1)} = \frac{\phi^{122}(z_1, s^m_a)}{\phi^{12}(z_1, s^m_a)} \\
\implies &\frac{\partial}{\partial s^m_a} \log|\phi^{12}(s^m_a, z_1)| = \frac{\partial}{\partial s^m_a} \log|\phi^{12}(z_1, s^m_a)| \\
\implies &\phi^{12}(s^m_a, z_1) = c(z_1)\phi^{12}(z_1, s^m_a),
\end{aligned}
\tag{34}
$$

for all s^m_a with some constant $c(z_1)$ not dependent on s^m_a. This is exactly the condition assumed in Assumption A2. Since this is true for each 2×2 submatrices corresponding to all $a \in \mathcal{V}_\mathrm{b}$, we conclude that the matrix Eq. 32 has full-rank, and thus the submatrix (rows) corresponding to $\mathcal{V}_\mathrm{b}$ has full row-rank $2|\mathcal{V}_\mathrm{b}|$.

The Rows Corresponding to $\mathcal{V}_\mathbf{p}$ We next show the full row-rankness of the submatrix (rows) corresponding to $\mathcal{V}_\mathrm{p}$. We especially consider the $2|\mathcal{V}_\mathrm{p}| \times 2|\mathcal{V}_\mathrm{p}|$ submatrix given by

$$
\begin{bmatrix}
\mathbf{B}^{112}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_1), & \mathbf{B}^{112}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_2) \\
\mathbf{B}^{12}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_1), & \mathbf{B}^{12}(\mathbf{s}^m_{\mathcal{V}_\mathrm{p}}, z_2)
\end{bmatrix}.
\tag{35}
$$

If this submatrix has full-rank, the submatrix (rows) of the first factor of Eq. 31 corresponding to $\mathcal{V}_{\mathrm{p}}$ also has full row-rank. We focus on a 2×2 submatrix corresponding to a variable index $a \in \mathcal{V}_{\mathrm{p}}$, given by

$$\begin{bmatrix} \phi^{112}(s_a^m, z_1) & \phi^{112}(s_a^m, z_2) \\ \phi^{12}(s_a^m, z_1) & \phi^{12}(s_a^m, z_2) \end{bmatrix}. \tag{36}$$

Calculating the determinant, with the uniform dependency assumption of the all elements (Assumption A2), this submatrix has full-rank (non-zero determinant) if the following condition does not hold:

$$\begin{aligned} &\frac{\phi^{112}(s_a^m, z_1)}{\phi^{12}(s_a^m, z_1)} = \frac{\phi^{112}(s_a^m, z_2)}{\phi^{12}(s_a^m, z_2)} \\ \Longrightarrow\ &\frac{\partial}{\partial s_a^m} \log|\phi^{12}(s_a^m, z_1)| = \frac{\partial}{\partial s_a^m} \log|\phi^{12}(s_a^m, z_2)| \\ \Longrightarrow\ &\phi^{12}(s_a^m, z_1) = c(z_1, z_2)\phi^{12}(s_a^m, z_2), \end{aligned} \tag{37}$$

for all s_a^m with some constant $c(z_1, z_2)$ not dependent on s_a^m. This is exactly the condition assumed in Assumption A2. Since this is true for each 2×2 submatrices corresponding to the all $a \in \mathcal{V}_{\mathrm{p}}$, we conclude that the matrix Eq. 35 has full-rank, and thus the submatrix (rows) corresponding to $\mathcal{V}_{\mathrm{p}}$ has full row-rank $2|\mathcal{V}_{\mathrm{p}}|$.

The Rows Corresponding to $\mathcal{V}_{\mathbf{c}}$ We lastly show the full row-rankness of the submatrix (rows) corresponding to $\mathcal{V}_{\mathrm{c}}$. With the similar discussion to that for the rows $\mathcal{V}_{\mathrm{p}}$ given above, this submatrix has full row-rank $2|\mathcal{V}_{\mathrm{c}}|$ if the following condition does not hold:

$$\phi^{12}(z_1, s_a^m) = c(z_1, z_2)\phi^{12}(z_2, s_a^m), \tag{38}$$

for all s_a^m with some constant $c(z_1, z_2)$ not dependent on s_a^m. This is exactly the condition assumed in Assumption A2.

Combining the all results above, we finally conclude that the first factor in Eq. 31 has full row-rank $2d_{\mathcal{S}}^m$ $(= 2|\mathcal{V}_{\mathrm{b}}| + 2|\mathcal{V}_{\mathrm{p}}| + 2|\mathcal{V}_{\mathrm{c}}|)$. This indicates that the right-hand of Eq. 31 has full row-rank, and so does the left-hand side. Then the Lemma is proven.

□

D. Alternative Identifiability Condition of Theorem 1

In Theorem 1, we can weaken the constraints on the causal function ϕ (Assumption A2) by strengthening that on the causal graph (A1), especially by assuming that all variables have both parent and child in some other group. The alternative conditions of Theorem 1 is given in the following Proposition:

Proposition 1. *Assume the generative model given by Eqs. 2 and 3, and also the following:*

A'1 (Nondegeneracy of the graph) For any group m in $\mathcal{M}$ (or in a subset of $\mathcal{M}$, that we call "the groups of interest"), each variable has both a (at least one) parent and child in some other group, and the collection of inter-group adjacency matrices $\bar{\mathbf{L}}^m$ given below has full row-rank:

$$\bar{\mathbf{L}}^m = \begin{bmatrix} \mathbf{L}^{m1}, & \dots, & \mathbf{L}^{mM} \\ (\mathbf{L}^{1m})^{\mathsf{T}}, & \dots, & (\mathbf{L}^{Mm})^{\mathsf{T}} \end{bmatrix}, \tag{39}$$

A'2 (Causal function) ϕ^{12} has uniform dependency, and either of the following conditions is satisfied:

- *(a) Both ϕ^{112} and ϕ^{122} have uniform dependency, and for any open subset B of $\bar{\mathcal{S}}$, there exist some $z \in \bar{\mathcal{S}}$ such that the following condition does not hold for ϕ^{12}: $\phi^{12}(s, z) = c\phi^{12}(z, s)$ with some constants $c \in \mathbb{R}$ for all $s \in B$.*
- *(b) Either one of ϕ^{112} and ϕ^{122} has uniform dependency, and the other one is constantly zero.*

Then, for all groups m in $\mathcal{M}$ (or in the groups of interests), $\mathbf{s}^m$ can be recovered up to permutation and variable-wise invertible transformations from the distribution of the observations $\mathbf{x}$.

Assumption A'1 is similar to A1, while it requires all variables to have both parent and child. Note that in this case the matrix $\bar{\mathbf{L}}^m$ has at least one non-zero element for each row, and thus does not have all-zero rows. Although A'1 imply that the whole causal graph cannot be acyclic, this does not mean we cannot identify any of the latent variables in acyclic graphs; since A'1 is *group-wise*, we can still identify the variables on the groups (groups of interest) satisfying them. For example, in the illustrative example of the causal dynamics (Illustrative Example 2), we cannot identify the latent variables on the first (no-parents) and the last (no-children) time points (groups), while we would be able to identify the other time-points (groups) t since they usually obtain significant causal effects from nearby preceding time-points (groups) $< t$, then cause the other nearby subsequent time-points (groups) $> t$.

The assumptions on the function ϕ (A'2) is weaker than that of Theorem 1 (A2) since they do not require the factorizability of ϕ and also uniform-dependency of either ϕ^{112} or ϕ^{122} in A'2b. For example, Gaussian CAMs (Eq. 10; $N = 1$ and linear $\mathbf{T}$ in Eq. 9) are allowed in A'2b, while they are not accepted in Theorem 1.

Proof. Proof is basically the same as that of Theorem 1 (Supplementary Material C), while we only need to consider the rows $\mathcal{V}_{\mathrm{b}}$ in Lemma 2 since all of the variables have both parent and child in some other group here. We thus only need to show the full row-rankness of Eq. 33 in Lemma 2, corresponding to the rows $\mathcal{V}_{\mathrm{b}}$, under the condition A'2a or A'2b. Condition A'2a represents the asymmetricity, which is actually the same as that assumed in Theorem 1, and thus can be proven by the same discussion given in the proof of Lemma 2. We can also easily see the full row-rankness of Eq. 33 under Condition A'2b since the determinant of the matrix (Eq. 33) is non-zero from the uniform dependency assumptions.

Then the Proposition is proven. □

E. Proof of Theorem 2

By well-known theory (Gutmann & Hyvärinen, 2012; Hastie et al., 2001), after convergence of logistic regression, with infinite data and a function approximator with universal approximation capability, the regression function (Eq. 6) will equal the difference of the log-pdfs in the two classes $\mathbf{x}^{(n)}$ and $\mathbf{x}^{(n_*)}$ in Eq. 5:

$$\begin{aligned}
&\sum_{m\in\mathcal{M}} \bar{\psi}^m(\mathbf{h}^m(\mathbf{x}^m)) + \sum_{m\neq m'} \sum_{(a,b)\in\mathcal{V}_S^m\times\mathcal{V}_S^{m'}} w_{ab}^{mm'}\psi(h_a^m(\mathbf{x}^m), h_b^{m'}(\mathbf{x}^{m'})) + c \\
&= \log p_{\mathbf{x}}(\mathbf{x}^1,\dots,\mathbf{x}^M) - \log p_{\mathbf{x}^*}(\mathbf{x}^1,\dots,\mathbf{x}^M) \\
&= \log p(\mathbf{g}^1(\mathbf{x}^1),\dots,\mathbf{g}^M(\mathbf{x}^M)) - \log p_{\mathbf{s}^*}(\mathbf{g}^1(\mathbf{x}^1),\dots,\mathbf{g}^M(\mathbf{x}^M)) \\
&= \log p(\mathbf{g}^1(\mathbf{x}^1),\dots,\mathbf{g}^M(\mathbf{x}^M)) - \sum_{m\in\mathcal{M}} \log p^m(\mathbf{g}^m(\mathbf{x}^m)),
\end{aligned} \tag{40}$$

where $p_{\mathbf{x}}$, $p_{\mathbf{x}^*}$, and $p_{\mathbf{s}^*}$ are the joint densities of the observational vector $\mathbf{x}^{(n)}$ (the first dataset in Eq. 5), observational vector with randomized samples for each group $\mathbf{x}^{(n_*)}$ (the second dataset in Eq. 5), and that on the latent space $\mathbf{s}^{(n_*)}$, respectively, and p^m is the marginal distribution of the m-th latent variable group, $\mathbf{g}^m = (\mathbf{f}^m)^{-1}$ are the (true) demixing models. The second equation comes from the well-known theory that the changes of variables do not change the density-ratio (subtraction of log-densities; the Jacobians for the changes of variables cancel out), and the third equation comes from the fact that there is no causal relations across groups on the shuffled dataset because the samples are obtained randomly and independently for each group (while causal relations can still exist within each group, implicitly involved in p^m).

Let a compound demixing-mixing function $\mathbf{v}^m(\mathbf{s}^m) = \mathbf{h}^m \circ \mathbf{f}^m(\mathbf{s}^m)$, we then have

$$\begin{aligned}
&\log p(\mathbf{s}^1,\dots,\mathbf{s}^M) - \sum_{m\in\mathcal{M}} \log p^m(\mathbf{s}^m) \\
&= \sum_{m\in\mathcal{M}} \bar{\psi}^m(\mathbf{v}^m(\mathbf{s}^m)) + \sum_{m\neq m'} \sum_{(a,b)\in\mathcal{V}_S^m\times\mathcal{V}_S^{m'}} w_{ab}^{mm'}\psi(v_a^m(\mathbf{s}^m), v_b^{m'}(\mathbf{s}^{m'})) + c.
\end{aligned} \tag{41}$$

We substitute the factorization model Eq. 3 into this, and differentiate the both sides with respect to s_a^m and $s_b^{m'}$, where

$a \in \mathcal{V}_S^m$, $b \in \mathcal{V}_S^{m'}$, $m \neq m'$, and then obtain,

$$\frac{\partial^2}{\partial s_a^m \partial s_b^{m'}} \left(\lambda_{ab}^{mm'} \phi(s_a^m, s_b^{m'}) + \lambda_{ba}^{m'm} \phi(s_b^{m'}, s_a^m) \right) = \frac{\partial^2}{\partial s_a^m \partial s_b^{m'}} \sum_{(i,j)} \left(w_{ij}^{mm'} \psi \left(v_i^m(\mathbf{s}^m), v_j^{m'}(\mathbf{s}^{m'}) \right) + w_{ji}^{m'm} \psi \left(v_j^{m'}(\mathbf{s}^{m'}), v_i^m(\mathbf{s}^m) \right) \right). \tag{42}$$

Now compare this equation to Eq. 19 of the proof of Theorem 1 in Supplementary Material C. The functions ψ and $\tilde{\phi}$, and the coefficients $\tilde{\lambda}_{ij}^{mm'}$ and $w_{ij}^{mm'}$ denote the same things in the two proofs. Now, we can proceed with the proof of Theorem 1, and the consistency of the estimation framework is thus proven.

F. Proof of Theorem 3

Proof. From the result of Theorem 1 with the required assumptions, for each $m \in \mathcal{M}$, we so far have the identifiability of the latent variables $(\mathbf{s}^m)$ up to variable-wise nonlinear scalings and a permutation; i.e., the compound function $\mathbf{v}^m$ in the proof of Theorem 1 (Supplementary Material C) is given by, for each element,

$$v_a^m(\mathbf{s}^m) = k_{\sigma^m(a)}^m(s_{\sigma^m(a)}^m), \tag{43}$$

where $k_{\sigma^m(a)}^m : \mathbb{R} \to \mathbb{R}$ is a scalar invertible functions, and $\sigma^m(a) : \mathcal{V}_S^m \to \mathcal{V}_S^m$ represents the permutation of variables, which are indeterminate according to Theorem 1. Without loss of generality, we assume that the variables were sorted properly ($\sigma^m(a) = a$), and the nonlinear functions were scaled properly so that the image is embedded on the same space (interval) to that of the input (i.e. $s_a^m \in \mathcal{S}_a^m \to k_a^m(s_a^m) \in \mathcal{S}_a^m$).

We now discuss identifiability of the model (Eq. 3) by considering two sets of parameters $\theta = \{\mathbf{L}, \phi\}$ (true) and $\tilde{\theta} = \{\tilde{\mathbf{L}}, \tilde{\phi}\}$ (another parameterization or estimate) satisfying the assumptions of the Theorem, such that they both give the same data distributions $p(\mathbf{x}; \theta) = p(\mathbf{x}; \tilde{\theta})$.

Resolving the Element-wise Nonlinear Scaling: We first show that the element-wise (nonlinear) scaling k_a^m can be resolved to some extent by some additional assumptions given in Theorem 3; more specifically, the scaling k_a^m can be given by the same function k^m for each group m, rather than the variable-wise manner (Eq. 43). We focus on co-parents s_a^m and $s_{a'}^m \in \mathcal{V}_S^m$ and their child $s_b^{m_*} \in \mathcal{V}_S^{m_*}$, assumed in Assumption C1. We then have the (a, b)-th element of Eq. 20 (causal relation between s_a^m and $s_b^{m_*}$) with substituting Eq. 43,

$$\frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}} \left(\lambda_{ab}^{mm_*} \phi(s_a^m, s_b^{m_*}) + \lambda_{ba}^{m_*m} \phi(s_b^{m_*}, s_a^m) \right) = \frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}} \left(\tilde{\lambda}_{ab}^{mm_*} \tilde{\phi}(k_a^m(s_a^m), k_b^{m_*}(s_b^{m_*})) + \tilde{\lambda}_{ba}^{m_*m} \tilde{\phi}(k_b^{m_*}(s_b^{m_*}), k_a^m(s_a^m)) \right), \tag{44}$$

and likewise the (a', b)-th element (causal relation between variables $s_{a'}^m$ and $s_b^{m_*}$).

From Assumption C1, $\lambda_{ab}^{mm_*} \neq 0$ and $\lambda_{a'b}^{mm_*} \neq 0$ on the left-hand side (true parameter θ; the opposite directions are zeros $\lambda_{ba}^{m_*m} = \lambda_{ba'}^{m_*m} = 0$ since the graph is directed; Assumption C1). By taking a division of Eq. 44 corresponding to those two variable-pairs, which is possible thanks to the uniform-dependency of the cross-derivatives of the functions (Assumption A2), we obtain four possible equations, depending on which combination between $(\tilde{\lambda}_{ab}^{mm_*}, \tilde{\lambda}_{ba}^{m_*m})$ and $(\tilde{\lambda}_{a'b}^{mm_*}, \tilde{\lambda}_{ba'}^{m_*m})$ has

non-zero values on the right-hand side;

$$\frac{\lambda_{ab}^{mm_*}\frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}}\phi(s_a^m, s_b^{m_*})}{\lambda_{a'b}^{mm_*}\frac{\partial^2}{\partial s_{a'}^m \partial s_b^{m_*}}\phi(s_{a'}^m, s_b^{m_*})} = \begin{cases} \frac{\tilde{\lambda}_{ab}^{mm_*}\frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}}\tilde{\phi}(k_a^m(s_a^m), k_b^{m_*}(s_b^{m_*}))}{\tilde{\lambda}_{a'b}^{mm_*}\frac{\partial^2}{\partial s_{a'}^m \partial s_b^{m_*}}\tilde{\phi}(k_{a'}^m(s_{a'}^m), k_b^{m_*}(s_b^{m_*}))} & (\tilde{\lambda}_{ab}^{mm_*}\tilde{\lambda}_{a'b}^{mm_*} \neq 0), \\ \frac{\tilde{\lambda}_{ba}^{m_*m}\frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}}\tilde{\phi}(k_b^{m_*}(s_b^{m_*}), k_a^m(s_a^m))}{\tilde{\lambda}_{ba'}^{m_*m}\frac{\partial^2}{\partial s_{a'}^m \partial s_b^{m_*}}\tilde{\phi}(k_b^{m_*}(s_b^{m_*}), k_{a'}^m(s_{a'}^m))} & (\tilde{\lambda}_{ba}^{m_*m}\tilde{\lambda}_{ba'}^{m_*m} \neq 0), \\ \frac{\tilde{\lambda}_{ab}^{mm_*}\frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}}\tilde{\phi}(k_a^m(s_a^m), k_b^{m_*}(s_b^{m_*}))}{\tilde{\lambda}_{ba'}^{m_*m}\frac{\partial^2}{\partial s_{a'}^m \partial s_b^{m_*}}\tilde{\phi}(k_b^{m_*}(s_b^{m_*}), k_{a'}^m(s_{a'}^m))} & (\tilde{\lambda}_{ab}^{mm_*}\tilde{\lambda}_{ba'}^{m_*m} \neq 0), \\ \frac{\tilde{\lambda}_{ba}^{m_*m}\frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}}\tilde{\phi}(k_b^{m_*}(s_b^{m_*}), k_a^m(s_a^m))}{\tilde{\lambda}_{a'b}^{mm_*}\frac{\partial^2}{\partial s_{a'}^m \partial s_b^{m_*}}\tilde{\phi}(k_{a'}^m(s_{a'}^m), k_b^{m_*}(s_b^{m_*}))} & (\tilde{\lambda}_{ba}^{m_*m}\tilde{\lambda}_{a'b}^{mm_*} \neq 0). \end{cases} \tag{45}$$

On the right-hand side (estimate $\tilde{\theta}$), only one of them is possible due to the directed causal graph assumption (Assumption C1). The first two cases are when the causal directions are the same between the two variable-pairs on the parameterization $\tilde{\theta}$, similarly to θ (but possibly both flipped from θ), while they are opposite each other in the latter two cases.

We first show that the latter two cases of Eq. 45 (opposite causal directions between the two pairs (a, b) and (a', b)) contradict the assumptions, as we expected. We replace s_a^m and $s_{a'}^m$ by a common variable $y_1 \in \bar{\mathcal{S}}$ (this is possible because we consider the case where the supports of the all latent variables are the same, denoted as $\bar{\mathcal{S}}$), and $s_b^{m_*}$ by $y_2 \in \bar{\mathcal{S}}$, then obtain

$$\frac{\lambda_{ab}^{mm_*}}{\lambda_{a'b}^{mm_*}} = \begin{cases} \frac{\tilde{\lambda}_{ab}^{mm_*}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_a^m(y_1), k_b^{m_*}(y_2))}{\tilde{\lambda}_{ba'}^{m_*m}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_b^{m_*}(y_2), k_{a'}^m(y_1))} & (\tilde{\lambda}_{ab}^{mm_*}\tilde{\lambda}_{ba'}^{m_*m} \neq 0), \\ \frac{\tilde{\lambda}_{ba}^{m_*m}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_b^{m_*}(y_2), k_a^m(y_1))}{\tilde{\lambda}_{a'b}^{mm_*}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_{a'}^m(y_1), k_b^{m_*}(y_2))} & (\tilde{\lambda}_{ba}^{m_*m}\tilde{\lambda}_{a'b}^{mm_*} \neq 0), \end{cases} \tag{46}$$

where the left-hand-side is constant. However, Lemma 3 given below indicates that these contradict the assumptions, and thus the latter two cases of Eq 45 are indeed excluded.

On the other hand, in the first two cases of Eq 45, we again replace the variables by y_1 and y_2, then obtain

$$\frac{\lambda_{ab}^{mm_*}}{\lambda_{a'b}^{mm_*}} = \begin{cases} \frac{\tilde{\lambda}_{ab}^{mm_*}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_a^m(y_1), k_b^{m_*}(y_2))}{\tilde{\lambda}_{a'b}^{mm_*}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_{a'}^m(y_1), k_b^{m_*}(y_2))} & (\tilde{\lambda}_{ab}^{mm_*}\tilde{\lambda}_{a'b}^{mm_*} \neq 0), \\ \frac{\tilde{\lambda}_{ba}^{m_*m}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_b^{m_*}(y_2), k_a^m(y_1))}{\tilde{\lambda}_{ba'}^{m_*m}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}(k_b^{m_*}(y_2), k_{a'}^m(y_1))} & (\tilde{\lambda}_{ba}^{m_*m}\tilde{\lambda}_{ba'}^{m_*m} \neq 0). \end{cases} \tag{47}$$

We now show that those equations are possible only when $k_a^m = k_{a'}^m$ due to the assumptions. From Assumption C1, there exists a path from a variable to any other variable by following the co-parents on group m, and for each co-parents we have either one of the cases in Eq. 47. However, once whether the former or the latter case of Eq. 47 is determined for some co-parent, all other co-parents also need to have the same side of the equation, since the existence of inconsistent causal directions is not allowed due to Lemma 3. This indicates that we have a relation of either

$$\begin{aligned} \frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_a^m(y_1), k_b^{m_*}(y_2)\right) &= \alpha_{1aa'}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_{a'}^m(y_1), k_b^{m_*}(y_2)\right), \\ \frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_b^{m_*}(y_2), k_a^m(y_1)\right) &= \beta_{1aa'}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_b^{m_*}(y_2), k_{a'}^m(y_1)\right), \end{aligned} \tag{48}$$

consistently for all co-parents $(a, a') \in \mathcal{V}_{\mathcal{S}}^m \times \mathcal{V}_{\mathcal{S}}^m$ assumed in C1, where $\alpha_{1aa'}$ and $\beta_{1aa'}$ are some scalar constants depending on the co-parents.

We next consider the co-children s_a^m and $s_{a''}^m$ and their parent $s_{b'}^{m_\dagger}$, assumed in Assumption C1 ($m_\dagger$ does not need to be same as m_*). Based on the same discussions for the co-parents given above, we have a relation of either

$$\begin{aligned} \frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_{b'}^{m_\dagger}(y_2), k_a^m(y_1)\right) &= \alpha_{2aa''}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_{b'}^{m_*}(y_2), k_{a''}^m(y_1)\right), \\ \frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_a^m(y_1), k_{b'}^{m_\dagger}(y_2)\right) &= \beta_{2aa''}\frac{\partial^2}{\partial y_1 \partial y_2}\tilde{\phi}\left(k_{a''}^m(y_1), k_{b'}^{m_\dagger}(y_2)\right), \end{aligned} \tag{49}$$

consistently for all co-children $(a, a'') \in \mathcal{V}_S^m \times \mathcal{V}_S^m$ assumed in C1, where $\alpha_{2aa'}$ and $\beta_{2aa'}$ are some scalar constants depending on the co-children. The former and the latter cases in Eqs. 48 and 49 correspond to each other; if we have the first case of Eq. 48, we also have the first case of Eq. 49, excluding the second cases, and vice versa. We show this by contradiction; we suppose there exist three variables with causal relation $s_{b'}^{m_\dagger} \to s_a^m \to s_b^{m_*}$ as assumed in C1 on the true model θ, but they are wrongly learned as $s_{b'}^{m_\dagger} \leftarrow s_a^m \to s_b^{m_*}$ on the estimate $\tilde{\theta}$ (the following discussions are the same when they are learned as $s_{b'}^{m_\dagger} \to s_a^m \leftarrow s_b^{m_*}$ as well). This gives us two equations from Eq. 44,

$$\begin{aligned}
\lambda_{b'a}^{m_\dagger m} \frac{\partial^2}{\partial s_a^m \partial s_{b'}^{m_\dagger}} \phi(s_{b'}^{m_\dagger}, s_a^m) &= \tilde{\lambda}_{ab'}^{m m_\dagger} \frac{\partial^2}{\partial s_a^m \partial s_{b'}^{m_\dagger}} \tilde{\phi}(k_a^m(s_a^m), k_{b'}^{m_\dagger}(s_{b'}^{m_\dagger})), \\
\lambda_{ab}^{m m_*} \frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}} \phi(s_a^m, s_b^{m_*}) &= \tilde{\lambda}_{ab}^{m m_*} \frac{\partial^2}{\partial s_a^m \partial s_b^{m_*}} \tilde{\phi}(k_a^m(s_a^m), k_b^{m_*}(s_b^{m_*})).
\end{aligned} \tag{50}$$

We substitute $s_{b'}^{m_\dagger}$ with $(k_{b'}^{m_\dagger})^{-1} \circ k_b^{m_*}(s_b^{m_*})$, which makes the right-hand side the same up to scaling, after applying the change of variables and canceling out the derivatives of the functions k on the both sides. However, this contradicts Lemma 3 given below. This indicates that the relations of *parents* and *children* from a variable are preserved between θ and $\tilde{\theta}$, though could be flipped, and thus the cases of Eqs. 48 and 49 are consistent.

Now we focus on the first cases of Eqs. 48 and 49. Since those equations are true for any variable-pairs, we especially consider a pair (a, a') for both of them, and divide the both-sides (after replacing y_2 by y_3 for the second equation), which is possible thanks to the uniform dependency (Assumption A2), and then obtain

$$\begin{aligned}
&\frac{\frac{\partial^2}{\partial y_1 \partial y_2} \tilde{\phi}\left(k_a^m(y_1), k_b^{m_*}(y_2)\right)}{\frac{\partial^2}{\partial y_1 \partial y_3} \tilde{\phi}\left(k_{b'}^{m_\dagger}(y_3), k_a^m(y_1)\right)} = \frac{\alpha_{1aa'} \frac{\partial^2}{\partial y_1 \partial y_2} \tilde{\phi}\left(k_{a'}^m(y_1), k_b^{m_*}(y_2)\right)}{\alpha_{2aa'} \frac{\partial^2}{\partial y_1 \partial y_3} \tilde{\phi}\left(k_{b'}^{m_\dagger}(y_3), k_{a'}^m(y_1)\right)}, \\
\implies &\frac{\tilde{\phi}^{12}\left(k_a^m(y_1), k_b^{m_*}(y_2)\right)}{\tilde{\phi}^{12}\left(k_{b'}^{m_\dagger}(y_3), k_a^m(y_1)\right)} = \frac{\alpha_{1aa'}}{\alpha_{2aa'}} \cdot \frac{\tilde{\phi}^{12}\left(k_{a'}^m(y_1), k_b^{m_*}(y_2)\right)}{\tilde{\phi}^{12}\left(k_{b'}^{m_\dagger}(y_3), k_{a'}^m(y_1)\right)},
\end{aligned} \tag{51}$$

where $\phi^{12}(x, y) = \frac{\partial^2}{\partial x \partial y} \tilde{\phi}(x, y)$, and the derivatives of the scalar functions k_a^m, $k_{a'}^m$, $k_b^{m_*}$, and $k_{b'}^{m_\dagger}$ canceled out between the left- and the right-hand sides or between the numerator and the denominator of the equation (note that they are non-zeros almost everywhere due to the invertibility). Since this equation is true for any choices of y_2 and y_3, we set $y_2 = (k_b^{m_*})^{-1} \circ k_{a'}^m(y_1)$ and $y_3 = (k_{b'}^{m_\dagger})^{-1} \circ k_a^m(y_1)$, and we get

$$\begin{aligned}
&\frac{\tilde{\phi}^{12}\left(k_a^m(y_1), k_{a'}^m(y_1)\right)}{\tilde{\phi}^{12}\left(k_a^m(y_1), k_a^m(y_1)\right)} = \frac{\alpha_{1aa'}}{\alpha_{2aa'}} \cdot \frac{\tilde{\phi}^{12}\left(k_{a'}^m(y_1), k_{a'}^m(y_1)\right)}{\tilde{\phi}^{12}\left(k_a^m(y_1), k_{a'}^m(y_1)\right)}, \\
\implies &\tilde{\phi}^{12}\left(k_a^m(y_1), k_{a'}^m(y_1)\right)^2 = \frac{\alpha_{1aa'}}{\alpha_{2aa'}} \cdot \tilde{\phi}^{12}\left(k_a^m(y_1), k_a^m(y_1)\right) \tilde{\phi}^{12}\left(k_{a'}^m(y_1), k_{a'}^m(y_1)\right).
\end{aligned} \tag{52}$$

Since this equation indicates symmetricity of $\tilde{\phi}^{12}$ (flipping $k_a^m(y_1)$ and $k_{a'}^m(y_1)$ on the left-hand side gives the same value on the right-hand side), which is prohibited by Assumption C2, we need to have $k_a^m = k_{a'}^m$. Therefore we conclude that $k_a^m = k_{a'}^m$. Since this is true for each index-pair (a, a') in the group m considered in Assumption C1, k_a^m can be given as a single function k^m for all variable index $a \in \mathcal{V}_S^m$. This is also true when we focus on the latter cases Eqs. 48 and 49.

Identifiability of the Causal Graph: With the same discussions above, the functions $\{k_b^{m'}\}_b$ on a group m' can be also simply denoted as $k^{m'}$ for all b, based on the relations of the variables $\{s_b^{m'}\}_b$ on the group m' to the variables on the other groups. Using this to Eq. 44 with a group-pair (m, m'), and by gathering this equation for all variable-index-pairs $(a, b) \in \mathcal{V}_S^m \times \mathcal{V}_S^{m'}$ on the group-pair (m, m') in a matrix form (a giving rows, and b columns), and also by replacing all $\{s_a^m\}_a$ by a common variable $y_1 \in \bar{\mathcal{S}}$ and similarly all $\{s_b^{m'}\}_b$ by $y_2 \in \bar{\mathcal{S}}$, we get a matrix equation of the size $d_S^m \times d_S^{m'}$,

$$\begin{aligned}
&\mathbf{L}^{mm'} \frac{\partial^2}{\partial y_1 \partial y_2} \left(\phi(y_1, y_2)\right) + (\mathbf{L}^{m'm})^\mathsf{T} \frac{\partial^2}{\partial y_1 \partial y_2} \left(\phi(y_2, y_1)\right) \\
&= \tilde{\mathbf{L}}^{mm'} \frac{\partial^2}{\partial y_1 \partial y_2} \left(\tilde{\phi}(k^m(y_1), k^{m'}(y_2))\right) \\
&+ (\tilde{\mathbf{L}}^{m'm})^\mathsf{T} \frac{\partial^2}{\partial y_1 \partial y_2} \left(\tilde{\phi}(k^{m'}(y_2), k^m(y_1))\right).
\end{aligned} \tag{53}$$

The factors other than the adjacency matrices $\mathbf{L}^{mm'}$, $\mathbf{L}^{m'm}$, $\tilde{\mathbf{L}}^{mm'}$, and $\tilde{\mathbf{L}}^{m'm}$ are now scalar values and do not change across rows and columns.

Firstly, for the elements of Eq. 53 where $\lambda_{ab}^{mm'} = \lambda_{ba}^{m'm} = 0$ on the left-hand side, the corresponding coefficients $\tilde{\lambda}_{ab}^{mm'}$ and $\tilde{\lambda}_{ba}^{m'm}$ on the right-hand side should be also zeros, due to the directed causal graph assumption (Assumption C1) and uniform dependency of the cross-derivative of $\tilde{\phi}$ (A2).

We next focus on the elements of Eq. 53 where $\mathbf{L}^{mm'}$ are non-zeros (the corresponding elements of $(\mathbf{L}^{m'm})^\mathsf{T}$ are constantly zeros due to the directed causal graph assumption C1). In this case, we can say that only either of $\tilde{\mathbf{L}}^{mm'}$ or $(\tilde{\mathbf{L}}^{m'm})^\mathsf{T}$ are non-zeros consistently for all of the corresponding elements. This is because, if both of $\tilde{\mathbf{L}}^{mm'}$ and $(\tilde{\mathbf{L}}^{m'm})^\mathsf{T}$ have non-zero values on some different elements, it is easy to show that it contradicts the asymmetricity of the function $\tilde{\phi}$ (Assumption C2). This is the same for the case when we focus on the elements where $(\mathbf{L}^{m'm})^\mathsf{T}$ are non-zeros.

Similarly, this is also true when we focus on the right-hand side; if some elements of $\tilde{\mathbf{L}}^{mm'}$ are non-zeros, only the corresponding elements of either $\mathbf{L}^{mm'}$ or $(\mathbf{L}^{m'm})^\mathsf{T}$ are consistently non-zeros, due to the asymmetricity of ϕ (Assumption C2).

These indicate that we can identify the causal graph $\mathbf{L}^{mm'}$ and $\mathbf{L}^{m'm}$ up to scaling and matrix-transpose (flipping of $\tilde{\mathbf{L}}^{mm'}$ and $(\tilde{\mathbf{L}}^{m'm})^\mathsf{T}$). Additionally considering the permutation of variables for each group, which are indeterminate according to Theorem 1, we eventually have: if we have two sets of causal graphs $\mathbf{L}$ and $\tilde{\mathbf{L}}$ giving the same data distributions, we have $(\tilde{\mathbf{L}}^{mm'}, \tilde{\mathbf{L}}^{m'm}) = c^{mm'}(\mathbf{L}^{mm'}_{\boldsymbol{\sigma}^m \boldsymbol{\sigma}^{m'}}, \mathbf{L}^{m'm}_{\boldsymbol{\sigma}^{m'} \boldsymbol{\sigma}^m})$ or $c^{mm'}((\mathbf{L}^{m'm}_{\boldsymbol{\sigma}^{m'} \boldsymbol{\sigma}^m})^\mathsf{T}, (\mathbf{L}^{mm'}_{\boldsymbol{\sigma}^m \boldsymbol{\sigma}^{m'}})^\mathsf{T})$ with a scalar constant $c^{mm'}$, and permutations of variables (rows and columns) represented by $\boldsymbol{\sigma}^m$ and $\boldsymbol{\sigma}^{m'}$ on groups m and m', respectively. Theorem is then proven.

□

Lemma 3. *Assume $k_{\{1,2,3\}} : \mathbb{R} \to \mathbb{R}$ are C^1 scalar invertible functions, and $\phi(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$ is a function whose cross-derivative satisfies the asymmetricity (Assumption C2). Then for any open subset B of $\bar{\mathcal{S}}$, the following relation cannot hold;*

$$\frac{\partial^2}{\partial x \partial y} \phi(k_1(x), k_2(y)) = \gamma \frac{\partial^2}{\partial x \partial y} \phi(k_2(y), k_3(x)) \tag{54}$$

with some scalar constant $\gamma \neq 0 \in \mathbb{R}$, for all x and $y \in B$.

Proof. We give a proof by contradiction. We suppose the negation; there exist an open subset $B \subset \bar{\mathcal{S}}$ such that the equation hold for all x and $y \in B$. By the chain rule of the derivatives,

$$\phi^{12}(k_1(x), k_2(y)) \frac{\partial}{\partial x} k_1(x) = \gamma \phi^{12}(k_2(y), k_3(x)) \frac{\partial}{\partial x} k_3(x), \tag{55}$$

where $\phi^{12}(x, y) = \frac{\partial^2}{\partial x \partial y} \phi(x, y)$, and the derivatives of $k_{\{1,2,3\}}$ are non-zeros almost everywhere from their invertibility. The derivatives of $k_2(y)$ on both sides canceled out.

By a simple calculation from Eq. 55 and the uniform dependency (Assumption C2 with $c = 0$), we can say that a function $Q(x, y) = \frac{\phi^{12}(k_1(x), k_2(y))}{\phi^{12}(k_2(y), k_3(x))}$ does not depend on y. Due to this, we can consider $Q(x, y)$ with two different values of y, especially $y = (k_2)^{-1} \circ k_1(x)$ and $(k_2)^{-1} \circ k_3(x)$, and obtain an equation

$$\begin{aligned} &\frac{\phi^{12}(k_1(x), k_1(x))}{\phi^{12}(k_1(x), k_3(x))} = \frac{\phi^{12}(k_1(x), k_3(x))}{\phi^{12}(k_3(x), k_3(x))}, \\ \implies &\left(\phi^{12}(k_1(x), k_3(x))\right)^2 = \phi^{12}(k_1(x), k_1(x)) \phi^{12}(k_3(x), k_3(x)). \end{aligned} \tag{56}$$

Since this equation indicates symmetricity of ϕ^{12} (flipping $k_1(x)$ and $k_3(x)$ on the left-hand side gives the same value on the right-hand side), which is prohibited by Assumption C2, we need to have $k_1 = k_3$. However, substituting this result to Eq. 55 indicates that this is contradictory to the asymmetricity of ϕ^{12} (Assumption C2). From this contradiction, we conclude that Eq. 54 cannot hold with those assumptions, and thus the Lemma is proven.

□

G. Alternative Identifiability Condition of Theorem 3

By adding an additional constraint on the causal function ϕ, we can weaken the assumption on the causal graph $\mathbf{L}$ in Theorem 3. The alternative condition of Theorem 3 is given below:

Proposition 2. *Assume the same as those in Theorem 1, and also:*

C'1 (Causal graph) The inter-group causal relations of variables are all directed, and for every group-pair (m, m') in the groups of interest, all variables in a group m (and m') have either co-parent or co-child in the same group. In addition, any variables in the group m (and m') can be reached from any other variables in the same group by moving from a variable to one of its co-parents or co-children, possibly by multiple hops.

C'2 (Asymmetricity) There is no open subset B of $\bar{\mathcal{S}}$ such that for all $x \neq y \in B$, it holds

$$\phi^{12}(x, y) = c\phi^{12}(y, x) \tag{57}$$

with some constant $c \in \mathbb{R}$.

C'3 (Non-factorizability) There is no open subset B of $\bar{\mathcal{S}}$ such that for all $x \neq y \in B$, it holds

$$\phi^{12}(x, y) = \alpha(x)\beta(y) \exp\left(\gamma(x, y) - \gamma(y, x)\right) \tag{58}$$

with some scalar functions α, β, and γ.

Then, for all group-pairs (m, m') satisfying A1 and C'1, $(\mathbf{L}^{mm'}, \mathbf{L}^{m'm})$ are identifiable up to permutation of variables, linear scaling, and matrix transpose.

This Proposition requires additional constraint on the function ϕ (C'3) compared to Theorem 3. It restricts some factorization form of the cross-derivative of ϕ. This is similar to the non-factorizability in A2 of Theorem 1, but a bit stronger. Such restriction on the factorization of the (cross-derivative of) ϕ is reasonable because the factorization of ϕ^{12} into some input-variable-wise factors α and β would not be informative enough to fully determine the causal direction, which is also the case for the factorization into an anti-symmetric function with some factor γ.

Thanks to such stronger constraint on ϕ, the assumption on the causal graph (C'1) is weaker than C1 (Theorem 3); it requires only either co-parent or co-children for each variable, rather than both of them as in C1. In addition, we can consider both co-parents and co-children pairs to reach from one variable to another (see Supplementary Material H for some illustrative examples).

Proof. Proof is basically the same as that of Theorem 3 (Supplementary Material F). For showing $k_a^m = k_{a'}^m$ from Eq. 47, we use Lemma 4 given below, which only requires either co-parent or co-child for each variable in contrast to both of them as in Theorem 3 (from Eq. 48 to Eq. 52), thanks to the additional assumption of the non-factorizability of the function ϕ (C'3). The remaining proof is the same as that of Theorem 3 (Supplementary Material F), and thus omitted. □

Lemma 4. *Assume we have*

$$\frac{\partial^2}{\partial x \partial y}\phi(k_1(x), k_2(y)) = \gamma \frac{\partial^2}{\partial x \partial y}\phi(k_3(x), k_2(y)) \tag{59}$$

for all x and y in an open subset B of $\bar{\mathcal{S}}$, where $k_{\{1,2,3\}} : \mathbb{R} \to \mathbb{R}$ are C^1 scalar invertible functions, $\gamma \neq 0 \in \mathbb{R}$ is a constant scalar value, and $\phi(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$ is a function whose cross-derivative satisfies the uniform-dependency (Definition 2) and the condition C'3. Then we have $k_1(x) = k_3(x)$, and $\gamma = 1$. In addition, this is also the case even if the two arguments of ϕ are switched on the both-sides.

Proof. By the chain rule of the derivatives,

$$\phi^{12}(k_1(x), k_2(y)) \frac{\partial}{\partial x} k_1(x) = \gamma \phi^{12}(k_3(x), k_2(y)) \frac{\partial}{\partial x} k_3(x), \tag{60}$$

where $\phi^{12}(x,y) = \frac{\partial^2}{\partial x \partial y}\phi(x,y)$, and the derivatives of $k_{\{1,2,3\}}$ are non-zeros almost everywhere from their invertibility. The derivatives of $k_2(y)$ on both sides canceled out.

By a simple calculation from Eq. 60 and the uniform dependency, we can see that a function $Q(x,y) = \frac{\phi^{12}(k_1(x),k_2(y))}{\phi^{12}(k_3(x),k_2(y))}$ does not depend on y. Due to this, we can consider $Q(x,y)$ with two different values of y, especially $y = (k_2)^{-1} \circ k_1(x)$ and $(k_2)^{-1} \circ k_3(x)$, and obtain an equation

$$\begin{aligned} &\frac{\phi^{12}(k_1(x),k_1(x))}{\phi^{12}(k_3(x),k_1(x))} = \frac{\phi^{12}(k_1(x),k_3(x))}{\phi^{12}(k_3(x),k_3(x))}, \\ \implies &\phi^{12}(k_1(x),k_3(x))\phi^{12}(k_3(x),k_1(x)) = \phi^{12}(k_1(x),k_1(x))\phi^{12}(k_3(x),k_3(x)). \end{aligned} \tag{61}$$

However, this equation indicates that the function ϕ^{12} can be factorized as Eq. 58 from Lemma 5 given below, which is prohibited by Assumption C'3 unless $k_1 = k_3$. Therefore we have $k_1 = k_3$ by contradiction. Putting this into Eq. 59 also indicates that $\gamma = 1$. The same result can be obtained in the same manner even if the two arguments of ϕ are switched on the both-sides of Eq. 59. Then Lemma is proven.

□

Lemma 5. *The equation of a two variable function q with uniform-dependency given as*

$$q(x,y)q(y,x) = q(x,x)q(y,y) \tag{62}$$

holds if and only if the function q can be factorized as

$$q(x,y) = \alpha(x)\beta(y)\exp\left(\gamma(x,y) - \gamma(y,x)\right) \tag{63}$$

for some scalar functions α, β, and γ.

Proof. We take absolute values and logarithms on both sides of Eq. 62, and then take derivatives with respect to x and y, and obtain

$$\frac{\partial^2}{\partial x \partial y}\left(\log|q(x,y)|\right) + \frac{\partial^2}{\partial x \partial y}\left(\log|q(y,x)|\right) = 0. \tag{64}$$

This skew-symmetric equation holds if and only if the function can be represented by

$$\frac{\partial^2}{\partial x \partial y}\left(\log|q(x,y)|\right) = \bar{\gamma}(x,y) - \bar{\gamma}(y,x), \tag{65}$$

with some scalar function $\bar{\gamma}$. Taking integrals and exponential both sides (also considering the possible flipping of signs), we obtain Eq. 63, where γ corresponds to the integral function of $\bar{\gamma}$. Note that this factorization form of q indeed gives the Eq. 62. Then the Lemma is proven. □

H. Illustrative Discussion about Assumptions C1 and C'1

We give here additional discussion about Assumption C1 (Theorem 3) and C'1 (Proposition 2). Firstly, Fig. 3a illustrates the definitions of the *co-parent* and *co-child*. From the perspective of variable a in group m (s_a^m), variable b (s_b^m) is a *co-child* since it has the same parent in some other group m^* ($s_{a^*}^{m^*}$). Similarly, from the perspective of variable c in group m (s_c^m), variable d (s_d^m) is a *co-parent* since it has the same child in some other group m^{**} ($s_{d^{**}}^{m^{**}}$). The parent $s_{a^*}^{m^*}$ and the child $s_{d^{**}}^{m^{**}}$ to be considered can be arbitrary selected from any other group and variable-index, and m^* can be even the same to m^{**} (but not m). In this example, group m does not satisfy both C1 and C'1 since variables do not have either co-parent or co-child (C1), and thus some variables cannot be reached from some other variables through co-parent-and-child-paths (C1 and C'1).

For illustrative purpose, we start from a rather sparse causal graph as shown in Fig. 3b, where group m can satisfy Assumption C1 (and also C'1; note that we here focus only on a single group m for simplicity, while C1 and C'1 actually require both of the two target groups (m, m') to satisfy the same condition to identify the causal graph between them). In this example, every variable in the group m has both (at least one) co-parent and co-child in the same group, and

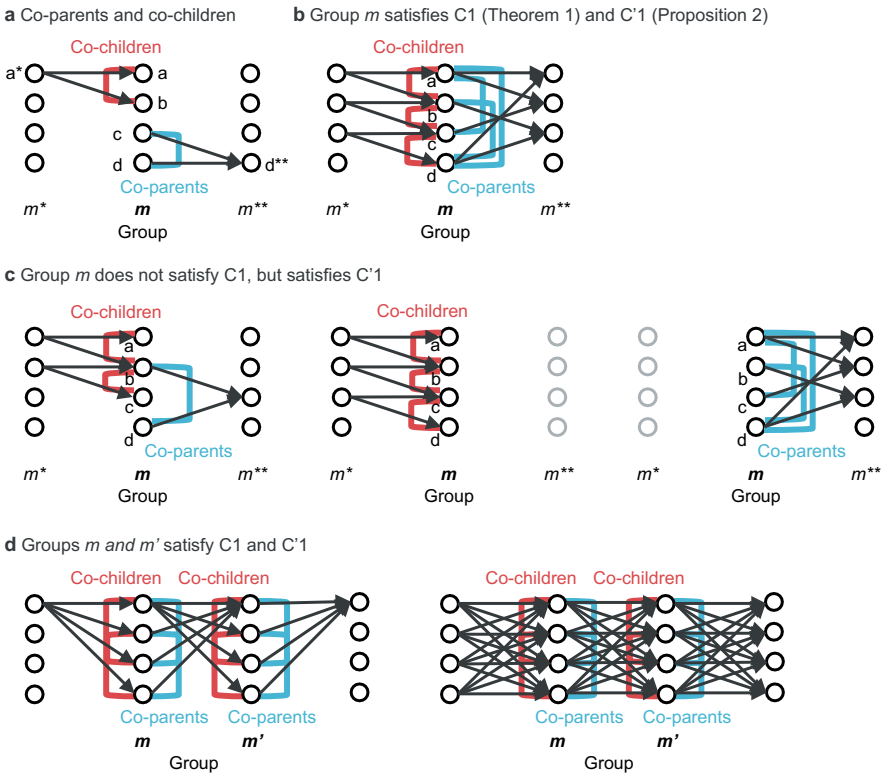

Figure 3. (**a**) Illustrative description of the *co-parents* and *co-children*. (**b**–**d**) Illustrative examples of some causal graphs, which (do not) satisfy Assumption C1 (Theorem 3) and/or C'1 (Proposition 2).

can reach any other variables in the same group by moving to one of its co-children (red paths; similarly for co-parents, blue paths), possibly by multiple hops. For example, we can reach from s_a^m to s_d^m on the co-children side via the path $s_a^m \to s_b^m \to s_c^m \to s_d^m$ in three hops (red paths), and via the path $s_a^m \to s_d^m$ in a single hop on the co-parents side (blue paths).

On the other hand, if we remove some edges from Fig. 3b, C1 cannot be satisfied anymore; in the all three examples shown in Fig. 3c, some variables do not have either co-parent or co-child, and thus cannot reach some other variables on either or both of the co-parent-side and co-child-side. On the other hand, all of them can satisfy C'1 (Proposition 2) since every variable has *either* co-parent or co-child in the same group, and can reach any other variables in the same group by moving through *either* co-parent-path or co-child-path for each hop. Those examples indicate that C'1 allows much sparser graphs than C1, though it requires an additional assumption on the function ϕ (C'3).

Assumption C1 would be fulfilled as long as the connections between groups are not too sparse. Fig. 3d show some such examples, where both of the two target groups (m, m') satisfy C1, and thus the causal graph between them can be identified. Especially the right side of Fig. 3d corresponds to the example given in Section 6; fully-connected autoregressive temporal causality to the subsequent time-point (group), where the groups m and m' correspond to time-points $t-1$ and t, respectively. As we can see, for every variable in a group (time-point in time-series cases), all other variables in the same group are both co-parents and co-children in this case. Note that due to Assumption A1 of Theorem 1, causal strengths need to be sufficiently different across edges in this case.

I. Related Works

Disentangled Representation Learning and Nonlinear ICA Revealing fundamental representation (latent variables $\mathbf{s}$) generating the observational data $\mathbf{x}$ in a data-driven manner is called representation learning (Bengio et al., 2013). This is supposed to be achieved by assuming some underlying observational process (e.g., mixing function $\mathbf{f}$ in Eq. 1; $\mathbf{x} = \mathbf{f}(\mathbf{s})$), and then by disentangling it (estimating the inverse model $\mathbf{g} = \mathbf{f}^{-1}$) in a data-driven manner from the observations. However,

such inverse problem is in general ill-posed, and there could exist many different combinations of components $\mathbf{s}$ and mixing function $\mathbf{f}$ which can explain the same observational data. The main focus of representation learning is thus to find the conditions where the components can be determined uniquely, for its interpretability, applicability, and reproducibility. Such model is called *identifiable*, and satisfies the condition $\forall(\theta, \theta'), p(\mathbf{x}; \theta) = p(\mathbf{x}; \theta') \Rightarrow \theta = \theta'$, which indicates that we can uniquely identify the parameters θ of the model (the observation model $\mathbf{f}$ or equivalently the demixing model $\mathbf{g} = \mathbf{f}^{-1}$) from the data distribution alone.

Independent component analysis (ICA) is one of such representation learning frameworks made to identifiable based on some assumptions on the latent variables. As the name suggests, ICA assumes that the latent components are *mutually independent*, $p(\mathbf{s}) = \prod_i p_i(s_i)$, as in many other representation learning frameworks including variants of variational autoencoders (VAEs) (Chen et al., 2018; Higgins et al., 2017; Kim & Mnih, 2018; Kingma & Welling, 2014), and so on. However, it is well-known that such independence assumption alone is not sufficient for the identifiability (Comon, 1994). The key idea of ICA is thus to give some additional assumptions on the components so as to give the identifiability. For example, it is well-known that when the mixing $\mathbf{f}$ is linear and samples are independent and identically distributed (i.i.d.), an additional assumption of non-Gaussianity (up to one Gaussian component) can make the model identifiable (Comon, 1994).

When the observational mixing $\mathbf{f}$ is nonlinear, the identifiability condition becomes much severer; it is known that the i.i.d. with non-Gaussianity assumption successfully used in the linear case above is not acceptable anymore in the nonlinear models (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). Nonlinear ICA (NICA) was firstly shown to be identifiable by assuming that the components are not i.i.d., but rather modulated depending on some additional (possibly latent) information associated with each sample (Khemakhem et al., 2020a; Hyvärinen & Morioka, 2016; Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Klindt et al., 2021; Sorrenson et al., 2020; Sprekeler et al., 2014). More specifically, by assuming that the components are *conditionally* mutually independent given an (possibly unobservable) additional auxiliary variable $\mathbf{u}$, i.e., $p(\mathbf{s}|\mathbf{u}) = \prod_i p_i(s_i|\mathbf{u})$, and the p_i is sufficiently modulated by $\mathbf{u}$, the model was shown to be identifiable (Hyvarinen et al., 2019). There exist many ways to consider $\mathbf{u}$; temporal- or spatial-segment-index (Hyvärinen & Morioka, 2016; Morioka et al., 2020), components on the other samples (Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Hälvä et al., 2021; Klindt et al., 2021), state-index of hidden Markov process (Hälvä & Hyvärinen, 2020), positive-negative-samples (Zimmermann et al., 2021), observation-target-samples (Roeder et al., 2021), and so on. Efficient estimation framework is also a crucial problem of NICA. Maximum-likelihood estimation (MLE) (Hälvä & Hyvärinen, 2020) or variational estimation (Hälvä et al., 2021; Khemakhem et al., 2020a) were shown to give reasonable inference. On the other hand, if $\mathbf{u}$ is directly observable, a self-supervised (or weakly-supervised) learning framework might be also proposed by using it as a target label or weak supervision (Hyvärinen & Morioka, 2016; Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Morioka et al., 2020; 2021; Zimmermann et al., 2021), which are empirically known to achieve better performances (Morioka et al., 2021).

The other direction for the identifiable NICA is to give some constraints on the mixing models $\mathbf{f}$, with weaker assumptions on the latent components; such as local isometry (Horan et al., 2021), volume preservation (Yang et al., 2022), independent mechanism (Gresele et al., 2021), sparseness (Moran et al., 2021), Brenier map (Wang et al., 2021), and a piecewise affine function (Kivva et al., 2022). It was found recently that local isometry gives identifiability up to linear transformation (Horan et al., 2021). Yang et al. (2022) showed identifiability by assuming that $\mathbf{f}$ is volume-preserving, the latent components are factorial multivariate Gaussian, and there exist two distinctive observations of auxiliary variable. Structural sparsity of the observational model is shown to give the identifiability (Zheng et al., 2022; Zheng & Zhang, 2023). Assuming multiple views from latent variables was also studied (Gresele et al., 2020; Locatello et al., 2020). Independent mechanism analysis (IMA) was shown to solve some well-known type of indeterminacy of NICA (Gresele et al., 2021), though its full-identifiability has not been resolved. Moran et al. (2021) proposed sparse VAE and showed the identifiability with additional anchor feature assumption. Wang et al. (2021) showed VAE with Brenier map enables identifiability without auxiliary information. Willetts & Paige (2022) empirically showed the possibility of identifiable NICA with clustering structure of the latent variables based on a Gaussian mixture model (VaDE; Jiang et al. (2017)), and Kivva et al. (2022) actually showed the identifiability of such model by additionally assuming that $\mathbf{f}$ is a piecewise affine function. They also showed that those assumptions make the latent components identifiable up to an affine transform even with (conditional) dependency between them, though (conditional) mutual independence is required if we need stronger identifiability.

Causal Discovery Causal discovery aims to estimate causal relations among variables from their observations in a data-driven manner. One of the major approaches is called a model-based approach, which models causal relations of variables based on some parametric models, such as Bayesian networks (BNs) or state-equation models (SEMs), and then estimate

the causal graph (or adjacency matrix) from the observations in a data-driven manner. One of the most general models is BNs (Pearl, 2000), which represent a causal graph among variables by a factorization of their joint distribution into some conditional distributions representing the conditional independence of the variables; i.e., $p(\mathbf{x}) = \prod_{a \in \mathcal{V}} p_a(x_a | \text{ pa}(x_a))$. Although BNs are flexible, recovering the graph from the joint distribution alone is not generally possible because many different graphs can have exactly the same joint distribution (Andersson et al., 1997; Spirtes et al., 2001). Some studies showed that suitable assumptions on the type of the conditional distributions enable identifiability of the causal structure, such as Poisson distribution (Park & Raskutti, 2015; Park & Park, 2019b), generalized hypergeometric distribution (Park & Park, 2019a), and zero-inflated Poisson model (Choi et al., 2020). A very closely related framework is given by SEMs (Bollen, 1989). Since SEMs are not generally identifiable (Bollen, 1989; Geiger & Heckerman, 1994; Pearl, 2000), similarly to BNs, some further assumptions were proposed to guarantee the identifiability: linear acyclic models with non-Gaussian noise (Shimizu et al., 2006; 2011), additive noise models excluding linear functions (Hoyer et al., 2008a; Hyvärinen & Smith, 2013; Peters et al., 2014), post-nonlinear models (Zhang & Hyvärinen, 2009), and so on. The SEMs can be also extended to time series (Gong et al., 2015; Hyvärinen et al., 2010), and models with latent confounding factors (Hoyer et al., 2008b; Maeda & Shimizu, 2020; Shimizu & Bollen, 2014), and so on. More recently, general nonlinear SEMs with non-additive noise have been proven to be identifiable by assuming nonstationarity of the noise (Monti et al., 2020; Wu & Fukumizu, 2020), though limited to bivariate settings.

Causal Representation Learning Causal representation learning (CRL) (Schölkopf et al., 2021) assumes the same observational models as NICA (Eq. 1), while the latent variables are not mutually independent but causally dependent each other; the causal mechanism $p(\mathbf{s})$ is given, for example, by a BN or SEM as in causal discovery studies (see above). The focus of CRL is to estimate both of the (high-level) latent causal variables and the causal graph from the (low-level) observations simultaneously, which would be achievable by jointly performing representation learning and causal discovery. However, CRL is supposed to be highly ill-posed without any assumptions, as a combination of two notoriously ill-posed problems of NICA and causal discovery (see above), and the degree of the indeterminacy should be even worse compared to those two individual problems; causal discovery can be seen as a special case of CRL, where the latent variables are directly observable ($\mathbf{f}$ is the identity mapping), and NICA can be seen as a special case of CRL as well, where all variables are mutually independent and thus $p(\mathbf{s})$ is simply given by a product of variable-wise distributions. Unfortunately, simply assuming causal structure on the latent space (Leeb et al., 2022) would not be enough for giving identifiability. Some studies recently succeeded to guarantee the identifiability by assuming some constraints on the latent variables, such as linear SEM with supervision or intervention (Buchholz et al., 2023; Liu et al., 2023; Shen et al., 2022; Squires et al., 2023; Varici et al., 2023; Yang et al., 2021), or more general causal relations but with do-interventions (Ahuja et al., 2022b), perfect intervention with unknown targets (von Kügelgen et al., 2023), or linear mixing (Varici et al., 2023), discrete latent variables with *mixture oracles* (Kivva et al., 2021), purity of the children of (subsets of) variables (Cai et al., 2019; Xie et al., 2020; 2022), or access to paired counterfactual data (Ahuja et al., 2022a; Brehmer et al., 2022). Independently Modulated Component Analysis (IMCA) (Khemakhem et al., 2020b) was proposed as an extension of NICA to allow some dependency across variables, though it requires weak-supervision (observable auxiliary variables), and only considers one-to-one relations between the latent variables and the auxiliary variables. Crucially, all of those frameworks require some level of supervision or intervention for each sample for the identifiability. Kivva et al. (2021) requires information of the number of components $k(S)$ of the mixture model (so-called *mixture oracle*) for every subsets S of observational variables. Estimating the number of componetns k itself is a long-standing challenge of finite mixture models, and infeasible or unstable in high-dimensional cases in practice, though they proposed a heuristic method for low dimensional case. Other studies (Lachapelle et al., 2022; Lippe et al., 2022; Yao et al., 2022a;b) have used temporal causal relations and are thus only applicable to time-series data (many of them also require weak-supervisions or interventions). Although Lippe et al. (2023) extended the temporal causality to also include instantaneous one, the identifiability condition is still highly dependent on the temporal structure.

Recently the concept of grouping of variables for CRL, similarly to our work, was proposed by Daunhawer et al. (2023); Lyu et al. (2022); Morioka & Hyvarinen (2023); Sturma et al. (2023); Yao et al. (2023). Most of them, except for Morioka & Hyvarinen (2023), especially focused on the intersections between groups. Sturma et al. (2023) showed that by considering multiple domains (corresponding to *groups* in this study) sharing some latent variables, the latent variables shared across all domains can be identified. Their causal and observational models are limited to linear models since their identifiability theorem and estimation algorithm are in principle based on linear ICA. Daunhawer et al. (2023); Lyu et al. (2022); Yao et al. (2023) considered more general causal and observational models; a nonlinear observational mixing for each view/modality (corresponding to *group* in this study) without assuming latent causal models explicitly. Those studies showed the identifiability of the latent variables corresponding to the intersection of the (subset of) groups. However, due to their

less-restrictive models compared to Sturma et al. (2023), the identifiability is limited to only up to block(intersection)-wise transformations. Lyu et al. (2022); Yao et al. (2023) also showed the identifiability of the group-specific (private) variables by assuming that they are independent on the intersections. Daunhawer et al. (2023); Lyu et al. (2022) are limited to two group settings, while Yao et al. (2023) extended them to more than two groups. Morioka & Hyvarinen (2023) proposed a CRL framework called connectivity-contrastive learning (CCL) designed for (homogeneous-)sensor-network-type architectures. In contrast to the group-based CRL frameworks mentioned above, the goal of this framework is to estimate the (causally-related) group-specific (private) variables. Their model can be seen as a very special case of ours; 1) CCL assumes the mixing functions $\mathbf{f}^m$ are the same for all groups m rather than group-specific as in ours (Eq. 2; the dimensions $d_{\mathcal{S}}^m$ can be also different across groups m in ours), 2) only considers component-wise relations between groups similarly to NICA (in other words, the adjacency coefficients $\lambda_{ab}^{mm'}$ in Eq. 3 can have non-zero values only when $a = b$, rather than all pairs of (a, b) as in ours)[1], and 3) the graph is a forest, while ours can be more general and even cyclic. These indicate much higher generality and applicability of our model (see Illustrative Examples) compared to CCL (homogeneous-sensor-network-type architectures). We also emphasize that the estimation frameworks are very different too, though both of them can be categorized as self-supervised (contrastive) learning; CCL uses group(node)-paired data for taking contrast, while our G-CaRL uses group-wise-shuffled data (Eq. 5).

J. Implementation Detail for Experiments

We give here more detail on the data generation, training, and evaluation in Experiments (Section 7). The codes are available at https://github.com/hmorioka/GCaRL.

J.1. Simulation 1: DAG

Data Generation We generated artificial data based on the generative model described in Section 3. Basically, the latent variables $\mathbf{s}^{(n)}$ were generated probabilistically for each sample n based on the pairwise BN causal model parameterized by an adjacency matrix $\mathbf{L}$ (Eq. 3), and then observed through nonlinear observational mixings $\mathbf{f}^m$ for each group m, after being divided into M groups (Eq. 2).

The whole causal graph $\mathbf{L}$ was designed to be a DAG (see Supplementary Fig. 7a for some examples). More specifically; the variables are causally ordered from 1 to $D_{\mathcal{S}}$, such that no later variable b causes any earlier variable $a < b$, and divided into non-overlapping M groups in order (i.e., the first $d_{\mathcal{S}}^1$ variables are group-1, and the next $d_{\mathcal{S}}^2$ variables are group-2, and so on). The whole causal graph $\mathbf{L}$ was generated separately for each sub-graphs; intra-group sub-graphs denoted as $\{\mathbf{L}^{mm}\}_{m\in\mathcal{M}}$, and inter-group sub-graphs $\{\mathbf{L}^{mm'}\}_{(m,m')}$. The intra-group sub-graph $\mathbf{L}^{mm} \in \mathbb{R}^{d_{\mathcal{S}}^m \times d_{\mathcal{S}}^m}$ was generated as a DAG for each group m, where each variable s_a^m was give one other randomly selected variable $s_{a'}^m$, $a' < a$, in the same group m as a parent (except for the first variable in the group). We used a special structure for the first group $\mathbf{L}^{11}$, where the number of parents (if they have) were fixed to 2 to avoid strong correlations between variables within the group, which can happen when the variables have only one parent. The inter-group sub-graphs $\mathbf{L}^{mm'} \in \mathbb{R}^{d_{\mathcal{S}}^m \times d_{\mathcal{S}}^{m'}}$ were generated randomly for each group-pair (m, m'), $m < m'$, so that each variable s_a^m on a group m has (almost) two children on other groups m', and conversely, each variable $s_b^{m'}$ on a group m' has (almost) two parents on the group m. The inter-group sub-graph on the opposite direction $\mathbf{L}^{m'm}$ is empty (zero-matrix) due to the causal ordering. In total, each variable on the m-th group ($m \geq 2$) has almost $2m - 1$ causal parents on average. The non-zero values of $\mathbf{L}$ were randomly drawn from $[0.9, 1]$, and then divided by the number of parents for each variable (column) so that the standard deviations of the variables were approximately the same regardless of the number of parents (also see below).

The latent variables were then sampled based on the following conditional distribution for each variable s_a^m at n-th sample:

$$s_a^{m(n)} \sim \exp \sum_{m' \neq m} \sum_{b \in \mathcal{V}_{\mathcal{S}}^{m'}} -\lambda_{ba}^{m'm} \left| s_a^{m(n)} - \alpha \tanh(\beta s_b^{m'(n)}) \right|, \tag{66}$$

where $\alpha = 3$ and $\beta = 0.8$ are scalar coefficients. This indicates that the sample $s_a^{m(n)}$ is randomly generated through a (piecewise) Laplace distribution with a standard deviation modulated by the inverse of the (summation of) $\lambda_{ba}^{m'm}$, and its

[1] Note the difference of the notation of (a, b) in our model Eq. 3 and that in Eq. 2 of Morioka & Hyvarinen (2023). The indices (a, b) in our model indicate the *variables*, while those in CCL indicate the *groups* (called nodes in CCL). The causal graphs are considered separately for each component in CCL (j in Eq. 2 of Morioka & Hyvarinen (2023)), while our model can even consider causal relations between every components (variables) without assuming independence.

average is biased by the activities of its parents, after nonlinearly transformed by $\tanh(\cdot)$. The non-parental variables do not directly influence s_a^m because the corresponding coefficients $\lambda_{ba}^{m'm}$ are set to zeros as mentioned above. This sampling distribution indicates that the function ϕ in Eq. 3 is given by,

$$\phi(x, y) = |y - \alpha \tanh(\beta x)| . \tag{67}$$

This function ϕ slightly violates the Assumption A2 of Theorem 1, so we can investigate the robustness of our theory at the same time. A typical smooth approximation of the Laplace density, such as $\phi(x, y) = -\sqrt{(y - \alpha \tanh(\beta x))^2 + \epsilon}$ with some small ϵ, would satisfy the assumption.

For the observation model $\mathbf{f}^m : \mathbb{R}^{d_{\mathcal{S}}^m} \to \mathbb{R}^{d_{\mathcal{X}}^m}$, we used a multilayer perceptron (MLP) with L layers (excluding the input layer) with random parameters, which takes a $d_{\mathcal{S}}^m$-dimensional latent variable $\mathbf{s}^{m(n)}$ and then outputs a $d_{\mathcal{X}}^m$-dimensional observation $\mathbf{x}^{m(n)}$ for each group $m \in \mathcal{M}$ and sample n. To guarantee the invertibility, we fixed $d_{\mathcal{S}}^m = d_{\mathcal{X}}^m = d^m$ and the number of units of each layer to d^m, and used leaky ReLU units for the nonlinearity except for the last layer which has no-nonlinearity.

The number of groups (M) was 3, the number of variables on each group ($d_{\mathcal{S}}^m = d_{\mathcal{X}}^m = d^m$) was 10 for all groups (i.e., the number of variables $D_{\mathcal{S}}$ was 30 in total), the number of data points n was $2^{16} = 65,536$, and the complexity (the number of layers) of the observational mixing model L was 3. We also evaluated the performances with changing those parameters to see how they affect the estimation performances (Supplementary Fig. 5a).

Training (G-CaRL) We trained the nonlinear regression function in Eq. 6 with the observed data by G-CaRL. We adopted MLP for each $\mathbf{h}^m : \mathbb{R}^{d_{\mathcal{X}}^m} \to \mathbb{R}^{d_{\mathcal{S}}^m}$ (h_{MLP}), whose outputs are supposed to represent the latent variables after the training (Theorem 2). The number of layers was selected to be the same as that of the observation model (L), and the number of units in each layer was $2d^m$ except for the output (d^m), so as to make it have enough number of parameters as the demixing model. A *maxout* unit was used as the activation function in the hidden layers, which was constructed by taking the maximum across two affine fully connected weight groups, while no-nonlinearity was applied at the output (last layer).

The function $\psi : \mathbb{R}^2 \to \mathbb{R}$ in the regression function (Eq. 6) was parameterized by

$$w_{ab}^{mm'} \psi(x, y) = w_{1ab}^{mm'} \left| w_{2ab}^{mm'} y + |w_{2ab}^{mm'}| \text{MLP}(x) \right| , \tag{68}$$

where $w_{1ab}^{mm'}$ and $w_{2ab}^{mm'}$ are weight parameters, which are supposed to give the estimation of the causal structure $(\tilde{\lambda}_{ab}^{mm'})$ by $(w_{1ab}^{mm'} |w_{2ab}^{mm'}|)$ after training (see Supplementary Fig. 7a for some examples). The nonlinear function $\text{MLP}(\cdot) : \mathbb{R} \to \mathbb{R}$ was parameterized by a learnable two-layer MLP with hyperbolic tangent (tanh) activation units. This function has enough degree of freedom to represent the true ϕ (Eq. 67). The intra-group functions $\bar{\psi}^m$ in Eq. 6 were also parameterized by MLP.

Those nonlinear functions were then trained by back-propagation with a momentum term (SGD) so as to optimize the cross-entropy loss of LR with a regression function Eq. 6 (Supplementary Algorithm 1). The initial parameters were randomly drawn from a uniform distribution or some non-informative constant values. The number of iterations for the optimization depends on the complexity of the model; e.g., convergence of a three-layer model by G-CaRL took about 3 hours (Intel Xeon 3.6 GHz 16 core CPUs, 384 GB Memory, NVIDIA Tesla A100 GPU), though we continued the training longer for safety.

Evaluation We evaluated the estimation performances of the latent variables and the causal structures by comparing the estimations with the true values. The learning was performed for 10 runs with changing the parameters of the observation model and the causal structures.

The estimated latent variables $\mathbf{h}^m(\cdot)$ were evaluated by their Pearson correlation to the true values $\mathbf{s}^m$ across samples. Since the order of the variable index is undetermined for each group (Theorems 1 and 2), we performed an optimal assignment of the variable indices $(\boldsymbol{\sigma}^m(\cdot) : \mathcal{V}_{\mathcal{S}}^m \to \mathcal{V}_{\mathcal{S}}^m)$ between the estimations and the true ones by the Munkres assignment algorithm (Munkres, 1957), maximizing the mean absolute-correlation coefficients, for each group m. The variable-wise accuracies (correlations) were then averaged over all variables.

For evaluations of the estimated causal structures $\tilde{\mathbf{L}}^{mm'} = (\tilde{\lambda}_{ab}^{mm'}) = (w_{1ab}^{mm'} |w_{2ab}^{mm'}|)$ (see Supplementary Fig. 7a for some examples), we at first converted them into binary directed (not necessarily DAG) adjacency matrices by the following procedure: we determined the causal direction on every pairs $(a, b) \in \mathcal{V}_{\mathcal{S}}^m \times \mathcal{V}_{\mathcal{S}}^{m'}$ by comparing the absolute values of

$\tilde{\lambda}_{ab}^{mm'}$ and $\tilde{\lambda}_{ba}^{m'm}$; direction is $s_a^m \to s_b^{m'}$ if $|\tilde{\lambda}_{ab}^{mm'}| > |\tilde{\lambda}_{ba}^{m'm}|$, and vice versa. We then removed edges whose absolute weights were less than a specific ratio (35% for Simulation 1) of the maximum absolute values of both $\tilde{\mathbf{L}}^{mm'}$ and $\tilde{\mathbf{L}}^{m'm}$ for each group-pair (m, m'). If both $|\tilde{\lambda}_{ab}^{mm'}|$ and $|\tilde{\lambda}_{ba}^{m'm}|$ are under the threshold, s_a^m and $s_b^{m'}$ are considered to have no direct causal relation. The obtained adjacency matrices were then compared with the (binarized) true causal structure $(\lambda_{ab}^{mm'})$, and evaluated by F1-score ($= 2 \cdot$ precision $\cdot$ recall / (precision + recall)). This kind of hard thresholding is known to be effective to reduce the number of false discoveries (Zheng et al., 2018), and seems to be especially important for methods like G-CaRL which do not explicitly impose sparseness or DAG structure constraints for the estimation. The threshold-ratio was determined separately for each experiment (simulation 1 and 2, and gene regulatory network recovery), but it was not changed across the parameter settings or runs within each experiment. Our preliminary analyses showed that the G-CaRL framework was not so sensitive to the selection of the threshold values, which can be seen from the ROC curves with varying threshold (Supplementary Fig. 6). Although G-CaRL is supposed to have indeterminacy of the causal graphs with its group-pair-wise matrix-transposes (Theorem 3), we solved that indeterminacy by giving some level of constraints to the function ψ (Eq. 68), where the function cannot be fitted into the opposite direction by its design. Therefore we directly used $\tilde{\mathbf{L}}^{mm'}$ as the final guess of $\mathbf{L}^{mm'}$ for each group-pair (m, m') without considering the possible matrix transpose. For solving the possible permutation of variables, we used the same permutations $\boldsymbol{\sigma}^m$ and $\boldsymbol{\sigma}^{m'}$ estimated based on the latent variables above. We only evaluated the *inter-group* causal connections, since only those are identifiable in our model (Theorem 3).

Baselines The baselines only include *unsupervised* frameworks with *instantaneous* (causal) dependency, since our experimental setting does not include supervision, intervention, nor temporal causality. Specifically, we showed the comparisons to three CRL frameworks MVCRL (Yao et al., 2023), CausalVAE (Yang et al. (2021) in unsupervised setting), and CCL (Morioka & Hyvarinen, 2023), and three representation learning frameworks MFCVAE (Falck et al. (2021); Kivva et al. (2022)), VaDE (Jiang et al., 2017; Kivva et al., 2022; Willetts & Paige, 2022) and β-VAE (Higgins et al., 2017) (see Supplementary Material K for details). We used publicly available implementations of them. We also applied a CRL framework Kivva et al. (2021) as well, though it failed due to the difficulty of the estimation of the mixture model in our data. Note that we cannot apply many of the existing CRL frameworks designed for instantaneous causal models (e.g., Shen et al. (2022)) since they usually require supervision or interventions, which are not available in this experiment. We cannot apply the multi-domain unsupervised CRL framework proposed by Sturma et al. (2023) since the groups have no overlap in our setting. Daunhawer et al. (2023); Lyu et al. (2022) are not applicable since they are limited to two group settings. We do not consider frameworks based on temporal causality, such as Lachapelle et al. (2022); Lippe et al. (2022); Yao et al. (2022a;b), since the samples are generated independently, with only instantaneous causality in this experiment. For a fair comparison, we used the same architecture (group-wise disentanglement) for the encoders of those baselines as in the feature extractors $\mathbf{h}^m$ of G-CaRL.

For baselines which do not estimate the (whole or part of) causal graph by themselves (MVCRL, CCL, MFCVAE, VaDE, and β-VAE), we additionally applied a causal discovery framework to the estimated latent variables as a post-processing, from a wide variety of selections so as to maximize the performances (Supplementary Fig. 4); DirectLiNGAM (Shimizu et al., 2011), NOTEARS (Zheng et al., 2018), NOTEARS-MLP (Zheng et al., 2020), GOLEM (Ng et al., 2020), PC (Spirtes & Glymour, 1991), CAM (Bühlmann et al., 2014), and CCD (Lacerda et al., 2008). Briefly, DirectLiNGAM, NOTEARS, and GOLEM are specialized at linear DAGs, CAM and NOTEARS-MLP are for nonlinear DAGs, and CCD assumes existence of directed cycles (See Supplementary Material K).

We used the same evaluation criteria for those baselines to that of G-CaRL (causal direction determination, thresholding, and variable assignments). The threshold was determined separately for each method, so as to maximize the F1-score (see Supplementary Fig. 6 for the effect of the varying threshold). Although some causal discovery frameworks listed above have a function to adjust the threshold so as to make the estimated graphs DAG, we instead applied the same thresholding method to ours, without constraining the acyclicity on the final graph. For some causal discovery frameworks which output a binarized adjacency matrix, we directly compared them with the binarized true adjacency matrices, after variable assignments. Since some of them output graphs possibly with some bi-directional (or undetermined) edges, we gave the *true* directions to them favorably.

To see the difficulty of our latent causal model for the conventional causal discovery frameworks, we also applied the causal discovery frameworks listed above directly to the latent variables (Supplementary Fig. 4). In this case, we applied them after standardizing the latent variables (zero-mean and unit-variance for each variable), as suggested by (Reisach et al., 2021).

J.2. Simulation 2: Cyclic Graphs with Latent Confounders

We give here more detail on the data generation and training in Simulation 2 (Section 7.2). Evaluation methods are the same to those in Simulation 1 (see Supplementary Material J.1),

Data Generation We generated artificial data in a similar manner to Simulation 1, though the causal graphs were designed to be much more complex, due to the presence of directed cycles and latent confounders. Basically, we firstly generated latent variables with twice of the target size of variables with possible cycles, and then simply masked half of them as unobservable variables (latent confounders) alternately for each group; there are 10 observable (non-confounder) variables and 10 latent confounders for each group m ($d_{\mathcal{S}}^m = 10, D = 30$ for observable variables).

The whole causal graph (including latent confounders) was generated separately for each group-pair as in Simulation 1, but here without considering the causal order. The intra-group sub-graphs $\mathbf{L}^{mm}$ was generated randomly so that each variable s_a^m has one other randomly selected variable $s_{a'}^m$, $a' \neq a$, in the same group as a parent. The inter-group sub-graphs $\mathbf{L}^{mm'}$ were generated randomly for each group-pair (m, m') so that each variable s_a^m on a group m has two children on other groups m', and conversely, each variable $s_b^{m'}$ on a group m' has (almost) two parents on the group m. And similarly for the opposite direction $\mathbf{L}^{m'm}$. At this point, each variable is supposed to have $2M - 1$ causal parents (including the latent confounders). The non-zero values of $\mathbf{L}$ were randomly drawn from $[0.9, 1]$.

The latent variables were then sampled based on the following conditional distribution for each variable s_a^m at n-th sample:

$$s_a^{m(n)} \sim \exp\left(\sum_{m' \neq m} \sum_{b \in \mathcal{V}_{\mathcal{S}}^{m'}} -\frac{\lambda_{ba}^{m'm}}{|\text{pa}(s_a^m)|} \left(s_b^{m'(n)} + |\text{pa}(s_a^m)| \text{Relu}(s_a^{m(n)}) \right)^2 \right), \tag{69}$$

where $\text{pa}(s_a^m)$ is the set of parents (including latent confounders) of variable s_a^m, deduced from the adjacency matrix, $|\text{pa}(s_a^m)|$ is the number of parents, and $\text{Relu}(x) = \max(0, x)$ is a rectified linear unit. This indicates that the activity $s_a^{m(n)}$ is randomly generated through a Gaussian distribution with a standard deviation modulated by the inverse of root of summation of $\lambda_{ba}^{m'm}$, and its average is negatively biased by the positive-, but not by negative-, activities of its parents (nonlinear inhibitory connection). The non-parental variables do not directly influence s_a^m because the corresponding coefficients $\lambda_{ba}^{m'm}$ are zeros as mentioned above. The inverse-scaling of $\lambda_{ba}^{m'm}$ by $|\text{pa}(s_a^m)|$ was used so that the (conditional) standard deviations of variables were approximately the same regardless of the number of parents. This sampling distribution indicates that the function ϕ in Eq. 8 is given by, with a simple calculation,

$$\phi(x, y) = y\text{Relu}(x). \tag{70}$$

Since this causal graph can have a directed cycle, we generated the data realizations based on Gibbs sampling.

After generating the latent variables, we masked half of the variables as latent confounders alternately. Since each variable needs to be causally related to at least one of the variables on some other group (Assumption A1), we generated the causal graph under a constraint that each variable s_a^m has one observable child and one observable parent on all of the other groups $m' \neq m$ after the masking, in the graph generation above.

We used MLPs for the observation models $\mathbf{f}^m : \mathbb{R}^{d_{\mathcal{S}}^m} \to \mathbb{R}^{d_{\mathcal{X}}^m}$, as in Simulation 1.

The number of groups (M) was 3, the number of the observable variables was 10 for all groups (i.e., $d_{\mathcal{S}}^m = d_{\mathcal{X}}^m = d^m = 10$; the number of variables $D_{\mathcal{S}}$ was 30 in total, and the number of latent confounds was also 30), The number of data points n was 2^{20}, and the complexity (the number of layers) of the observational mixing model L was 3. We also evaluated the performances with changing those parameters to see how they affect the estimation performances (Supplementary Fig. 5b).

Training (G-CaRL) We train the nonlinear regression function in Eq. 6 with the observed data by G-CaRL. The model is basically the same as that used in Simulation 1, except for the regression function. The function $\psi : \mathbb{R}^2 \to \mathbb{R}$ in the regression function (Eq. 6) was parameterized by

$$\psi(x, y) = y \max(a_1^{mm'}(x - b_1^{mm'}), a_2^{mm'}(x - b_2^{mm'})) \tag{71}$$

with some scalar parameters $a_1^{mm'}, b_1^{mm'}, a_2^{mm'}$ and $b_2^{mm'}$. This is based on the idea of maxout unit, and has enough degree of freedom to represent the causal effect function ϕ (Eq. 70). The intra-group functions $\bar{\psi}^m$ were parameterized by MLP. The weight parameters $w_{ab}^{mm'}$ in the regression function (Eq. 6) are supposed to give the estimation of the causal structure $(\tilde{\lambda}_{ab}^{mm'})$ after training (Theorem 3; see Supplementary Fig. 7b for some examples).

J.3. Recovery of Gene Regulatory Network

We used synthetic single-cell gene expression data generated by SERGIO (Dibaeinia & Sinha, 2020), where each gene expression is governed by a stochastic differential equation (SDE) derived from a chemical Langevin equation, with activating or repressing causal interactions with the other genes. The gene expression data generated by SERGIO were shown to be statistically comparable to real experimental data (Dibaeinia & Sinha, 2020). We used the same parameters for the differential equations as in (Dibaeinia & Sinha, 2020), but changed the hill coefficient from 2 to 6 to make the causal relations more nonlinear.

The causal graph was designed to be a DAG (as required of SERGIO) similarly to Simulation 1, but with latent confounders similarly to Simulation 2 (Supplementary Fig. 7c shows examples). More specifically, the intra-group sub-graphs $\mathbf{L}^{mm}$ was generated randomly in same way used in Simulation 1, while the inter-group sub-graphs $\mathbf{L}^{mm'}$ were generated randomly for each group-pair (m, m') in the same way used in Simulation 2, but only for the group pairs $m < m'$. To make almost all genes have children on some other group, we designated the last gene of each group $m < M$ as the leaves (have no children), and connected genes on the last group (group-M) to those genes as parents. The number of variables (genes) including the latent confounders was fixed to 60, we then masked half of them as the latent confounders, and divided them into 3 groups (i.e., $M = 3$, and $d_S^m = 10, D = 30$ for non-latent-confounder variables). The maximum contributions (weights of edges) from parental genes to target genes were set to 0.25 for all edges. We set half of the parents as activating, and the others as repressing for each gene. See Supplementary Fig. 7c for some example. The genes (variables) which do not have any parents were assigned as master regulators (MRs), and controlled by basal production rates, randomly selected from $[0.25, 0.75]$. We fixed the number of samples to 2^{18}.

For the observation model $\mathbf{f} : \mathbb{R}^{d_S^m} \to \mathbb{R}^{d_X^m}$, we used a multilayer perceptron (MLP) with L layers (excluding the input layer) similarly to Simulations 1 and 2 because there is no known realistic settings of the observational mixings in this kind of gene expression data, to the best of our knowledge. The complexity (the number of layers) of the observational mixing model L was fixed 3, similarly to Simulations 1 and 2.

The function $\psi : \mathbb{R}^2 \to \mathbb{R}$ in the regression function (Eq. 6) was parameterized as

$$\psi(x, y) = y \sum_{k=1}^{K} a_k \tanh(b_k x + c_k), \tag{72}$$

where $K = 5$ is a model order, a_k, b_k, and c_k are trainable scalar parameters. The weight parameters $w_{ab}^{mm'}$ in the regression function (Eq. 6) are supposed to give the estimation of the causal structure ($\tilde{\lambda}_{ab}^{mm'}$) after training (Theorem 3; see Supplementary Fig. 7c for some examples).

J.4. High-Dimensional Image Observation

Data Generation We additionally evaluated our framework by using high-dimensional image-generation process as the observational model. We especially used 3DIdent image dataset (Zimmermann et al., 2021). The dataset is compose of images of an object (tea pot) rendered with image-size 224×224, conditioned on ten continuous factors (called hereafter *image-factors*; see Supplementary Fig. 8); three dimensional positions of the object (PosX_{obj}, PosY_{obj}, and PosZ_{obj}), three dimensional rotation of the object in Euler angles (RotX_{obj}, RotY_{obj}, and RotZ_{obj}), the color of the object and the ground of the scene ($\text{Color}_{\text{obj}}$ and Color_{bg}), and the position and color of the spotlight ($\text{Pos}_{\text{light}}$ and $\text{Color}_{\text{light}}$).

We firstly generated *candidates* of the latent causal variables based on the same way used in Simulation 2 (cyclic graphs with latent confounders), especially with fixing the number of latent variables to 10 (there also exist 10 additional latent confounders) for all group (see Supplementary Fig. 8 for example). We then picked a set of ten image-factors closest to the generated ten dimensional variables (after scaling-normalization) from the dataset, and then adopted it as the actual latent causal variables ($\mathbf{s}^m$) and the corresponding image as the observational image ($\mathbf{x}^m$), for each group m. Note that we cannot find the image-factors which exactly matches the generated ten dimensional variables since the number of images are limited (250,000) in this dataset. This indicates possible misspecification of the causal model, which enables the evaluation of the robustness of G-CaRL. We fixed the number of groups M to 3, and the number of data points n to 2^{20}.

Training (G-CaRL) Since the observations are high-dimensional images, we used here convolutional neural networks as the feature extractors $\{\mathbf{h}^m\}$. More specifically, we used ResNet-18 (He et al., 2016) additionally with hyperbolic tangent units and a fully-connected layer on top of it, which encodes an input image into 10-dimensional features. Those 10 features

are supposed to represent the original image-factors after the training. We trained a single feature extractor shared across all groups since the observational models are supposed to be the same across groups in this experiment. The learning was performed for 10 runs with changing the parameters of the causal structures.

Evaluation Due to the much higher nonlinearity of the observational model compared to the former simulations, we slightly changed the evaluation criteria, which is the same for the baselines too: 1) for evaluating the estimation of the latent variables, we used Spearman's rank correlation instead of Pearson correlation, and 2) for evaluating the causal graph, we optimally chose either the original estimate or its matrix-transpose so as to maximize the F1-score, since the estimated causal graph could be flipped (matrix-transposed) from the true one sometimes as suggested by Theorem 3, which did not happen in the previous simulations. We found that only considering a matrix-transpose of the whole causal graph was enough, rather than group-pair-specific matrix transposes as suggested in Theorem 3. The other evaluation methods are the same to those in Simulation 1 and 2.

K. Details of Baselines

CCL Connectivity-contrastive learning (CCL; Morioka & Hyvarinen (2023)) is a CRL framework based on self-supervised learning, whose generative model can be seen as a special case of ours. CCL assumes a sensor-network-type generative model with homogeneous observations, where multidimensional observations are obtained for each group (called *node* in Morioka & Hyvarinen (2023)) from latent components (corresponding to *latent variables* in this study), which are causally-related across groups while mutually-independent across components. More specifically, the observational model is given by

$$\begin{aligned}\mathbf{x} &= \left[\mathbf{x}^1, \quad \ldots, \quad \mathbf{x}^M\right] \\ &= \left[\mathbf{f}(\mathbf{s}^1), \quad \ldots, \quad \mathbf{f}(\mathbf{s}^M)\right],\end{aligned} \tag{73}$$

with a group-common (node-homogeneous) observational mixing $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, and the joint distribution of the latent causal variables $\mathbf{s}$ are assumed to be factorized as

$$p(\mathbf{s}) \propto \prod_{a \in \mathcal{V}_S} \left[\prod_{m \in \mathcal{M}} \exp\left(\bar{\phi}_a^m(s_a^m)\right) \prod_{m \neq m'} \exp\left(\lambda_a^{mm'} \phi_a(s_a^m, s_a^{m'})\right) \right], \tag{74}$$

with some (component-specific) potential functions $\bar{\phi}_a^m : \mathbb{R} \to \mathbb{R}$ and $\phi_a(s_a^m, s_a^{m'}) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and (component-wise) adjacency coefficients between groups $\{\lambda_a^{mm'}\}_{(m,m')}$. They then showed the identifiability of this model as a causal model with some causal structural assumptions, such as asymmetricity of the causal graph $\{\lambda_a^{mm'}\}_{(m,m')}$ and the potential functions ϕ_a. Although their models somewhat resembles to ours, they can be seen as a very special case of ours (see Section 8). They also proposed a self-supervised learning framework called CCL, which estimates the latent components and the causal graph simultaneously, based on a multinomial logistic regression (MLR), whose pretext task is to well predict the group-pair label of group-paired observations, for every group-pairs and samples. Note that the estimation framework is much different from ours, though both of them are categorized as self-supervised learning.

In our experiment, since CCL can consider only the diagonal part of the adjacency matrix $\mathbf{L}^{mm'}$ for each group-pair (m, m') (compare Eq. 74 to our model Eq. 3; also see Supplementary Material I), we applied a causal discovery algorithm to the estimated latent components as a post-processing for reconstructing the whole causal graph.

MVCRL Multi-view CRL (MVCRL; Yao et al. (2023)) is a CRL framework targeting multi-view data, based on alignments of the (embeddings of) intersections of the latent variables (called *content*) between views (for the sake of consistency, we hereafter call *views* as *groups*). Some multimodal representation learning frameworks can be seen as a special case of their work (Ahuja et al., 2022a; Daunhawer et al., 2023; Gresele et al., 2020; von Kügelgen et al., 2021; Locatello et al., 2020). It assumes group-wise observational model

$$\mathbf{x}^m = \mathbf{f}^m(\mathbf{s}^m), \tag{75}$$

for each of the groups $m \in \mathcal{M}$, similarly to ours (Eq. 2), while it considers overlaps of the latent variables (and possibly those of the observations) for (some) subsets of groups $\mathcal{M}$. The authors then showed that sets of latent variables, each of

which is the intersection of some subset of groups (called *content*), can be block-identified. By additionally assuming that the non-shared (view-specific, called *style*) variables are independent on the shared variables (content), the non-shared variables are also block-identifiable. MVCRL seeks to learn the embeddings of the latent variables by optimizing the alignments of the contents shared between views.

In our experiments, although our data do not have *contents variables* shared across groups, we considered that all group-specific variables $\mathbf{s}^m$ are the *contents* to be aligned for all subsets of groups in the estimation by MVCRL. This should be practically better than explicitly considering them as group-specific (independent) variables, since they should have some similarity (though not *equivalence* like content variables) between groups due to their inter-group causal relations in our settings.

CausalVAE CausalVAE (Yang et al., 2021) is a CRL framework based on VAE with some causal structural assumptions on the latent embedding. The authors showed that some causal assumptions, such as linear directed acyclic causal graphs, and (weak) supervision on the latent variables give the identifiability of the model up to some indeterminacy. We especially used its unsupervised setting (CausalVAE-unsup, (Yang et al., 2021)) as a baseline, which is composed of an encoder and a decoder, as in vanilla-VAE, and also a Causal Layer to represent the causal relations of the latent variables. More specifically, the input signals (observations $\mathbf{x}$) passes through an encoder to obtain independent exogenous factors $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I})$, which are passed through a Causal Layer to generate causal variables $\mathbf{s}$ via a linear SEM as

$$\mathbf{s} = \mathbf{A}^\mathsf{T}\mathbf{s} + \boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{A}^\mathsf{T})^{-1}\boldsymbol{\epsilon}, \tag{76}$$

where $\mathbf{A}$ is a matrix parameter representing an adjacency matrix, which are then taken by the decoder to reconstruct the original observation $\mathbf{x}$. CausalVAE estimates the model by optimizing the evidence lower bound (ELBO), similarly to vanilla-VAE, but additionally with regularization on DAG-ness of $\mathbf{A}$.

β-VAE β-VAE (Higgins et al., 2017) is a representation learning framework based on the vanilla-VAE (Kingma & Welling, 2014), but with an adjustable regularization parameter β on the distribution of the embeddings. The generative model is the same as the vanilla-VAE; the observations $\mathbf{x}$ are obtained from the latent variables $\mathbf{s}$ through an unknown observational mixing $\mathbf{f}$ as

$$\mathbf{x} = \mathbf{f}(\mathbf{s}) + \boldsymbol{\epsilon} \tag{77}$$

with observational noise $\boldsymbol{\epsilon}$, and the joint distribution of the latent variables $\mathbf{s}$ are assumed to be the standard multivariate Gaussian distribution

$$p(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \mathbf{0}, \mathbf{I}), \tag{78}$$

where $\mathcal{N}(\cdot; \mathbf{0}, \mathbf{I})$ is the Gaussian density with zero-mean and unit diagonal covariance, which implies that the latent variables are assumed to be mutually orthogonal. β-VAE additionally has a regularization parameter β for adjusting the strength of the KL-divergence regularization controlling the discrepancy between the encoded latent variable distributions and their priors. When $\beta = 1$, β-VAE simply leads to the vanilla-VAE.

In our experiments, we used the setting of $\beta = 1$, which actually corresponds to the original-VAE (Kingma & Welling, 2014), since it gives the best estimation performance of the latent variables.

VaDE VaDE (Jiang et al., 2017) is a representation learning framework based on VAE with Gaussian mixture priors. The observational model is the same as the vanilla-VAE (Eq. 77), while the joint distribution of the latent variables is assumed to be given by a Gaussian mixture model as

$$p(\mathbf{s}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k > 0, \tag{79}$$

where $N(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ of the k-th mixture component. Willetts & Paige (2022) later empirically showed that the estimation of this model gives rather consistent results across estimations, Kivva et al. (2022) then showed that the identifiability of such model can be given by additionally assuming that $\mathbf{f}$ is a piecewise affine function. VaDE model can be seen as an unsupervised version of identifiable-VAE

(iVAE; Khemakhem et al. (2020a)), where the indices of the mixture components are unknown rather than being given as an observable auxiliary variable for all samples as in Khemakhem et al. (2020a).

In our experiments, we used the implementation by Kivva et al. (2022), and fixed the number of mixture components to be 5.

MFCVAE Multi-facet clustering variational autoencoder (MFCVAE; Falck et al. (2021)) extends the idea of VaDE (Jiang et al., 2017) so as to jointly consider multiple aspects (facets) of clustering features hidden in the data, such as *color* and *shape* in images. MFCVAE assumes that the observations are obtained from J multidimensional latent facets $\{\mathbf{z}_1, \ldots, \mathbf{z}_J\}$ as

$$\mathbf{x} = \mathbf{f}(\mathbf{z}_1, \ldots, \mathbf{z}_J) + \boldsymbol{\epsilon} \tag{80}$$

with observational noise $\boldsymbol{\epsilon}$, and each facet j has its own unique clustering structure represented by a multivariate Gaussian mixture model with K_j mixture components

$$p(\mathbf{z}_j) = \sum_{k_j=1}^{K_j} p(k_j)\mathcal{N}(\mathbf{z}_j; \boldsymbol{\mu}_{k_j}, \boldsymbol{\Sigma}_{k_j}), \quad p(k_j) = \mathrm{Cat}(\boldsymbol{\pi}_j), \tag{81}$$

where $\mathrm{Cat}(\cdot)$ denotes a categorical distribution, $\boldsymbol{\pi}_j$ is the K_j-dimensional vector of mixing weights, and $\boldsymbol{\mu}_{k_j}$ and $\boldsymbol{\Sigma}_{k_j}$ are the mean vector and (diagonal or full) covariance matrix of the k_j-th mixture component in facet j. MFCVAE then learns clustering on multiple facets jointly, from shared (or progressively-trained ladder-architectured) latent variables in a fully unsupervised and end-to-end manner. Kivva et al. (2022) later showed that the model is identifiable by additionally assuming that $\mathbf{f}$ is a piecewise affine function.

In our experiments, we fixed the number of facets to 3. The dimension of the latent space was set to be 5 for each facet and number of mixture components to be 25, as recommended in Falck et al. (2021).

DirectLiNGAM DirectLiNGAM (Shimizu et al., 2011) assumes SEMs with linear DAG and non-Gaussian errors. In the first step, DirectLiNGAM finds the causal order of variables by iteratively finding a root variable by performing regression and independence testing for each pair of variables, extracting one which is exogenous to the others, and then removing the effect of the root variable from the other ones. DirectLiNGAM then eliminates unnecessary edges using AdaptiveLasso (Zou, 2006), and outputs a weighted adjacency matrix.

NOTEARS NOTEARS (Zheng et al., 2018) assumes linear SEMs of DAG. It estimates a weighted adjacency matrix by minimizing a least-squares loss in scoring DAGs with regularization terms imposing sparseness and DAG-ness of the adjacency matrix. Since NOTEARS formulates the structure learning problem as a continuous optimization problem over real matrices, it can effectively avoid the traditional combinatorial optimization problem (NP-hard) of learning DAGs. We used the default parameters.

NOTEARS-MLP NOTEARS-MLP (Zheng et al., 2020) is an extension of NOTEARS (Zheng et al., 2018) to general nonparametric DAG models. NOTEARS-MLP models variable-wise nonlinear causal functions by MLPs, which are learned based on continuous optimization problem with regularizations for the sparseness of the MLP parameters and for DAG-ness of the causal functions. We used the default parameters.

GOLEM GOLEM (Ng et al., 2020) is an efficient version of NOTEARS (Zheng et al., 2018), which can reduce number of optimization iterations. GOLEM assumes linear DAGs, and performs multivariate Gaussian maximum likelihood estimation (MLE) with a soft version of the differentiable acyclicity constraint proposed in (Zheng et al., 2018). There are two proposed models; equal(EV) or unequal (NV) noise variances. We used EV here since the estimation performances were better than NV. We used the hyper-parameters used in (Ng et al., 2020).

PC PC algorithm (Spirtes & Glymour, 1991) is a constraint-based method. PC algorithm firstly constructs an undirected graph by removing edges from a fully connected graph based on independence and conditional independence tests. It then constructs a DAG by directing the edges based on the information of separation sets and with some additional assumptions (no new v-structures and directed cycles).

Algorithm 1 Pseudo-code of Grouped Causal Representation Learning (G-CaRL) based on stochastic gradient descent (SGD) optimization

Input: A set of observational data $\{\mathbf{x}^{(n)}\}_n$, observational grouping indices $\{\mathcal{V}_{\mathcal{X}}^m\}_m$, and hyper-parameters for SGD optimization.

1: Initialization: Initialize the parameters of the regression function (Eq. 6) with random values.
2: **repeat**
3: Randomly pick some samples n (mini-batch) from the observations $\{\mathbf{x}^{(n)}\}$ (label 1) and $\{\mathbf{x}^{(n_*)}\}$ (label 0) (Eq. 5), where $\mathbf{x}^{(n_*)}$ are generated artificially from the original $\{\mathbf{x}^{(n)}\}_n$ by shuffling the index n over all samples separately for each group m, indicated by $\mathcal{V}_{\mathcal{X}}^m$.
4: Update the parameters of the regression function (Eq. 6) so as to minimize the objective function (cross-entropy) of the logistic regression, discriminating the labels 1 and 0.
5: **until** the objective function converges.
6: **return** the trained (group-wise) nonlinear feature extractors $\{\mathbf{h}^m(\cdot)\}$ representing the latent causal variables (Theorem 2), and the weight parameters $\{w_{ab}^{mm'}\}$ representing the causal structures (Theorem 3).

GES Greedy equivalent search (GES) (Chickering, 2003) algorithm is a score-based method. GES starts with an empty graph and iteratively adds directed edges such that the improvement of Bayesian score (BIC score) is maximized, until no single edge addition increases the score (forward phase). GES then iteratively removes edges until no more improvements in the score can be made by single-edge deletions (backward phase).

CAM Causal additive model (CAM) (Bühlmann et al., 2014) assumes SEMs specified by DAG and additive Gaussian errors, which is an extension of linear SEMs by allowing for variable-wise scalar nonlinear functions (Collorary 31 in (Peters et al., 2014)). CAM at first estimates the causal order of variables based a greedy search algorithm so as to maximize the likelihood, then non-relevant edges were removed (pruning) by a sparse regression technique implemented based on significance testing of covariates.

CCD Cyclic causal discovery algorithm (CCD) (Lacerda et al., 2008) assumes that the data are causally sufficient (no latent variables), while possibly includes directed cycles. CCD extracts cyclic models using conditional independence tests, as with PC (Spirtes & Glymour, 1991). The output of CCD algorithm is a cyclic partial ancestral graph (PAG), which is a graphical object that represents a set of causal Bayesian networks that cannot be distinguished by the algorithm.

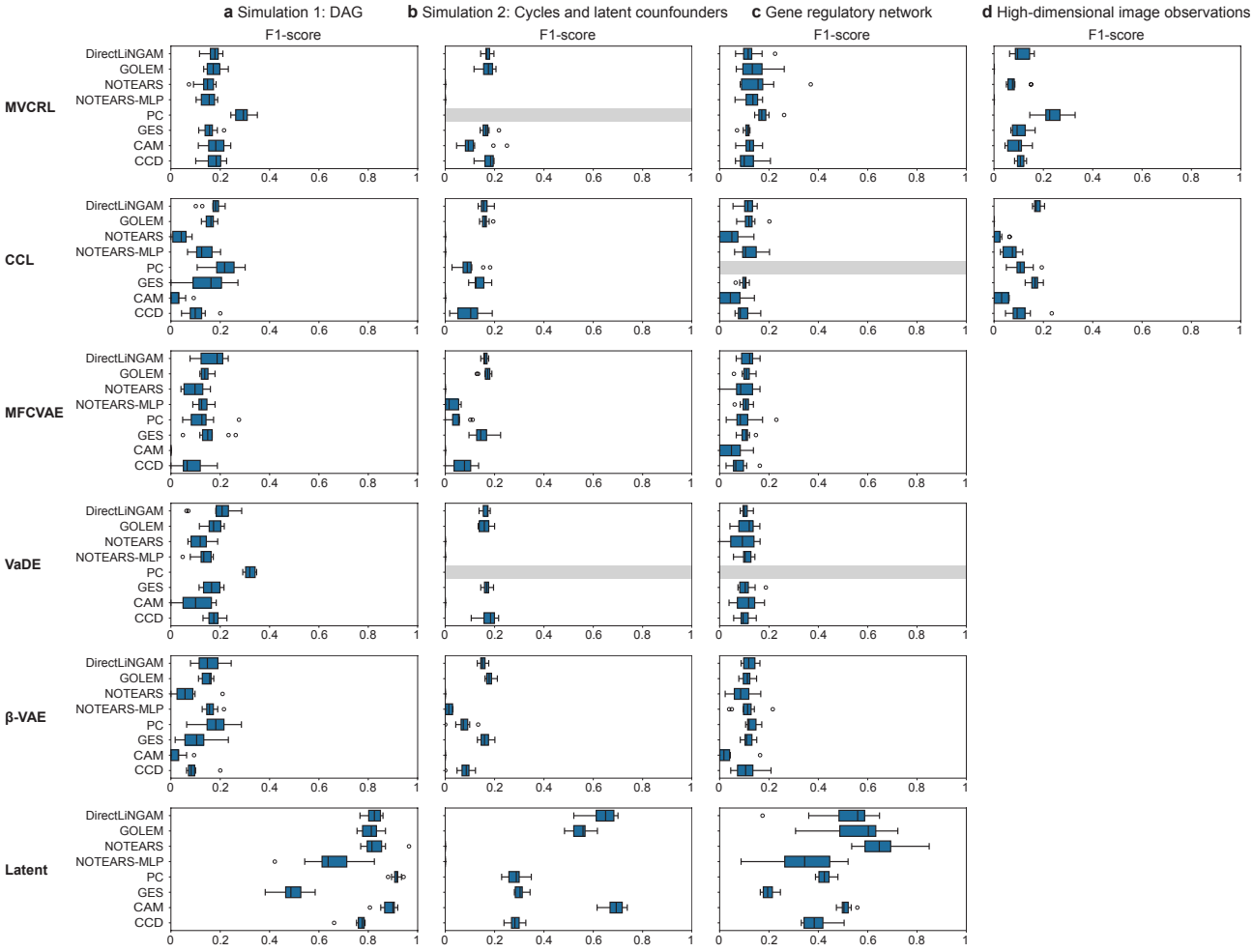

Figure 4. Comparison of a set of causal discovery frameworks (rows in each panel; measured by F1-score) applied to the baseline representation learning frameworks (rows), or directly to the latent variables (the last row: *Latent*; omitted in **d** since it is the same as **b**). We discarded some causal discovery frameworks (shaded by grey) on some panels since they did not converge within practical calculation time. (**a**) Simulation 1, (**b**) Simulation 2, (**c**) gene regulatory network recovery, and (**d**) high-dimensional image observations. Only the best performance for each panel was reported in Fig. 2.

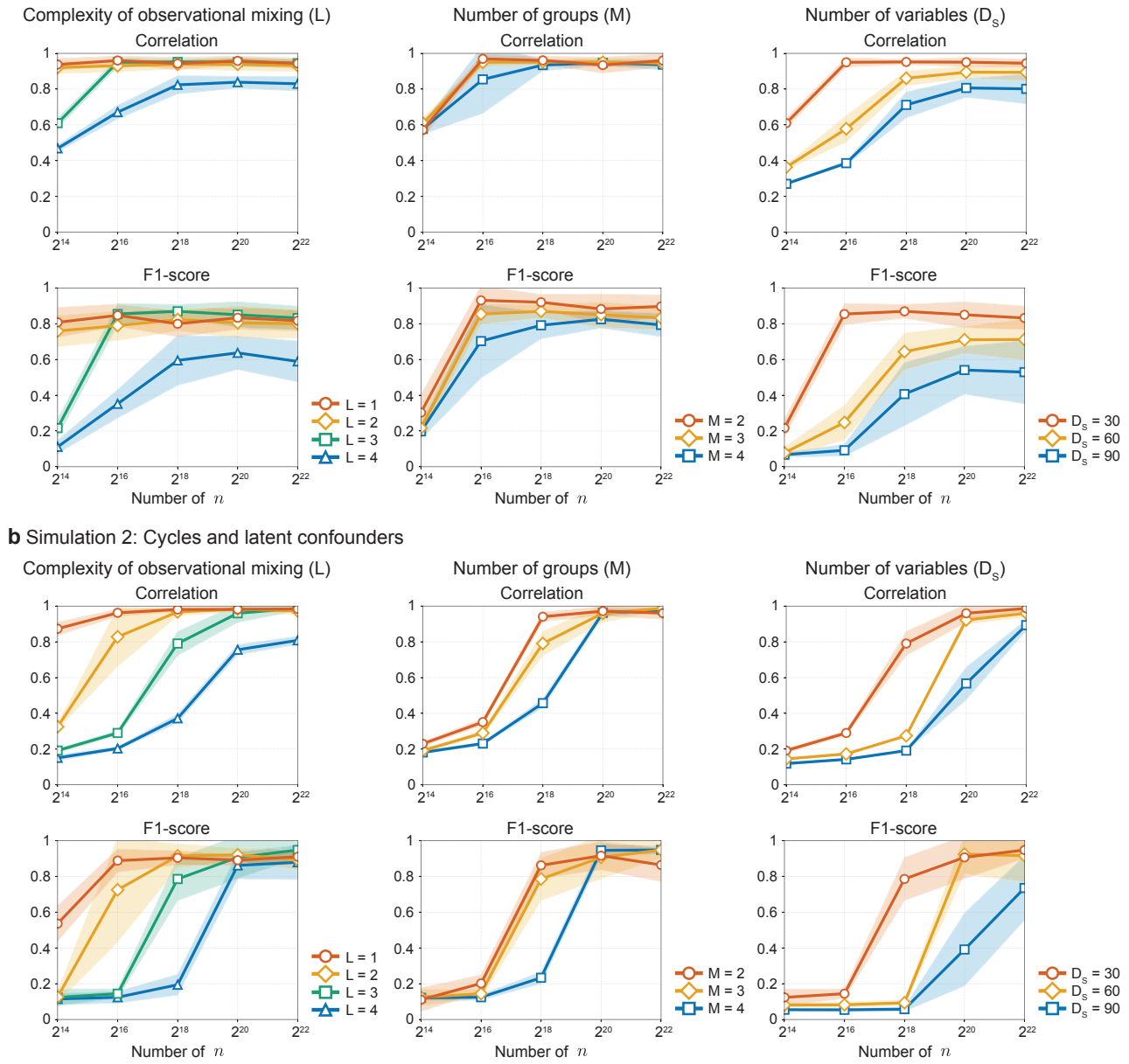

Figure 5. Estimation performances of the latent variables (Pearson correlation) and the causal structures (F1-score) by the proposed framework G-CaRL, but different settings of (Left) the complexity of the observation models (the number of MLP-layers L of the observation function $\mathbf{f}$), (Middle) the number of groups (M), and (Right) the number of variables ($D_{\mathcal{S}}$), with changing the number of samples n. Simulation 1 (basic DAG). (**b**) Simulation 2 (cycles and latent confounders). The values are the averages of 10 runs for each setting, and the shaded regions show the standard deviations. Fig. 2a corresponds to the case $L = 3$, $M = 3$, $D_{\mathcal{S}} = 30$, and $n = 2^{16}$ in **a**, and Fig. 2b corresponds to the case $L = 3$, $M = 3$, $D_{\mathcal{S}} = 30$, and $n = 2^{20}$ in **b**.

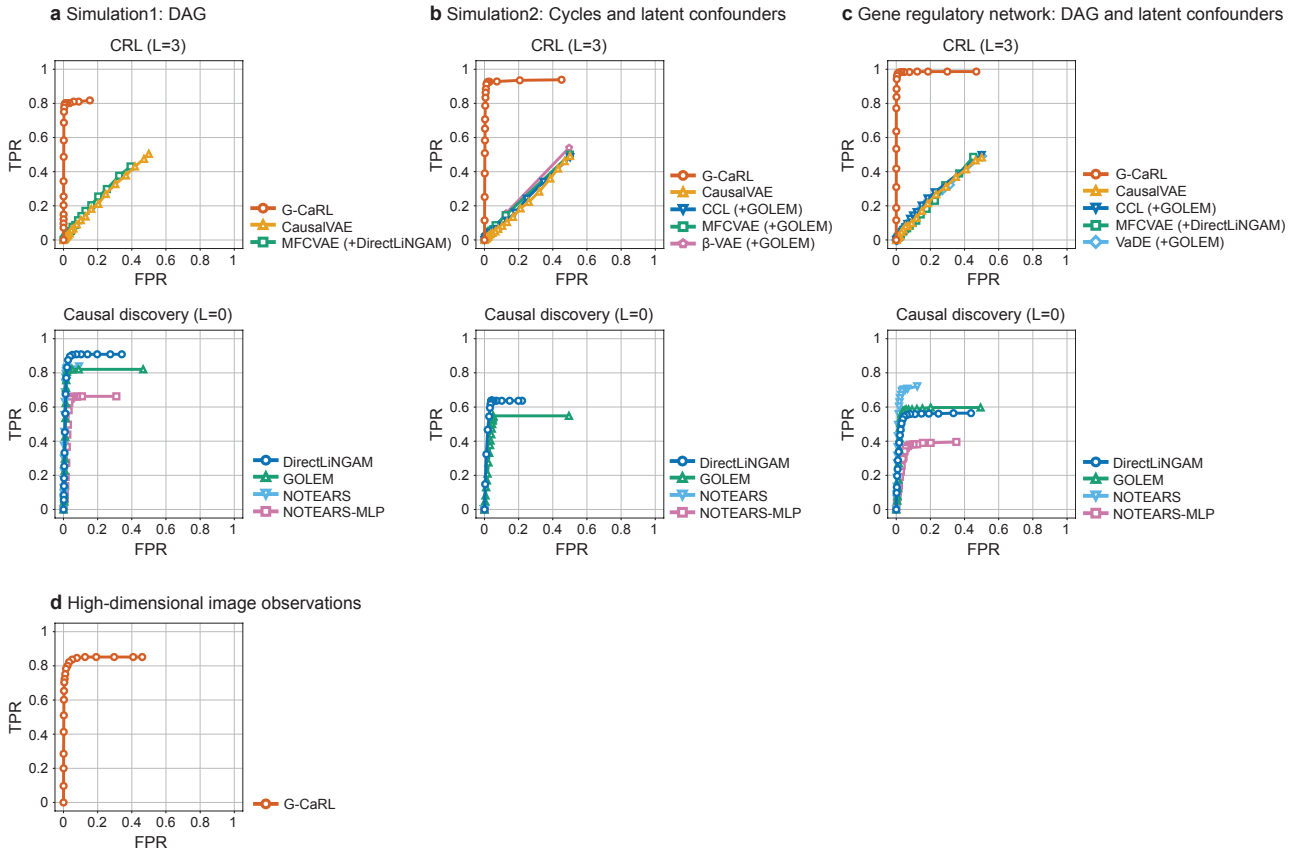

Figure 6. Illustration of the effect of the threshold for each method on (**a**) Simulation 1, (**b**) Simulation 2, (**c**) the gene regulatory network recovery task, and (**d**) the high-dimensional image observations. The upper panels show the results with unknown observational mixings (CRL), and lower panels show the results when we applied the causal discovery frameworks directly to the latent variables (causal discovery; omitted in **d** since it is the same as **b**). For each panel, ROC curve shows false positive rate (FPR) and true positive rate (TPR) with varying level of threshold, from 0% to 100% with interval of 5%, for each method. The values are the averages of 10 runs for each threshold. In upper panels, we only showed the curves for the (combinations of) frameworks in Fig. 2 which give *wighted* adjacency matrices and thus require thresholding. This result shows that G-CaRL was not so sensitive to the selection of the threshold values.

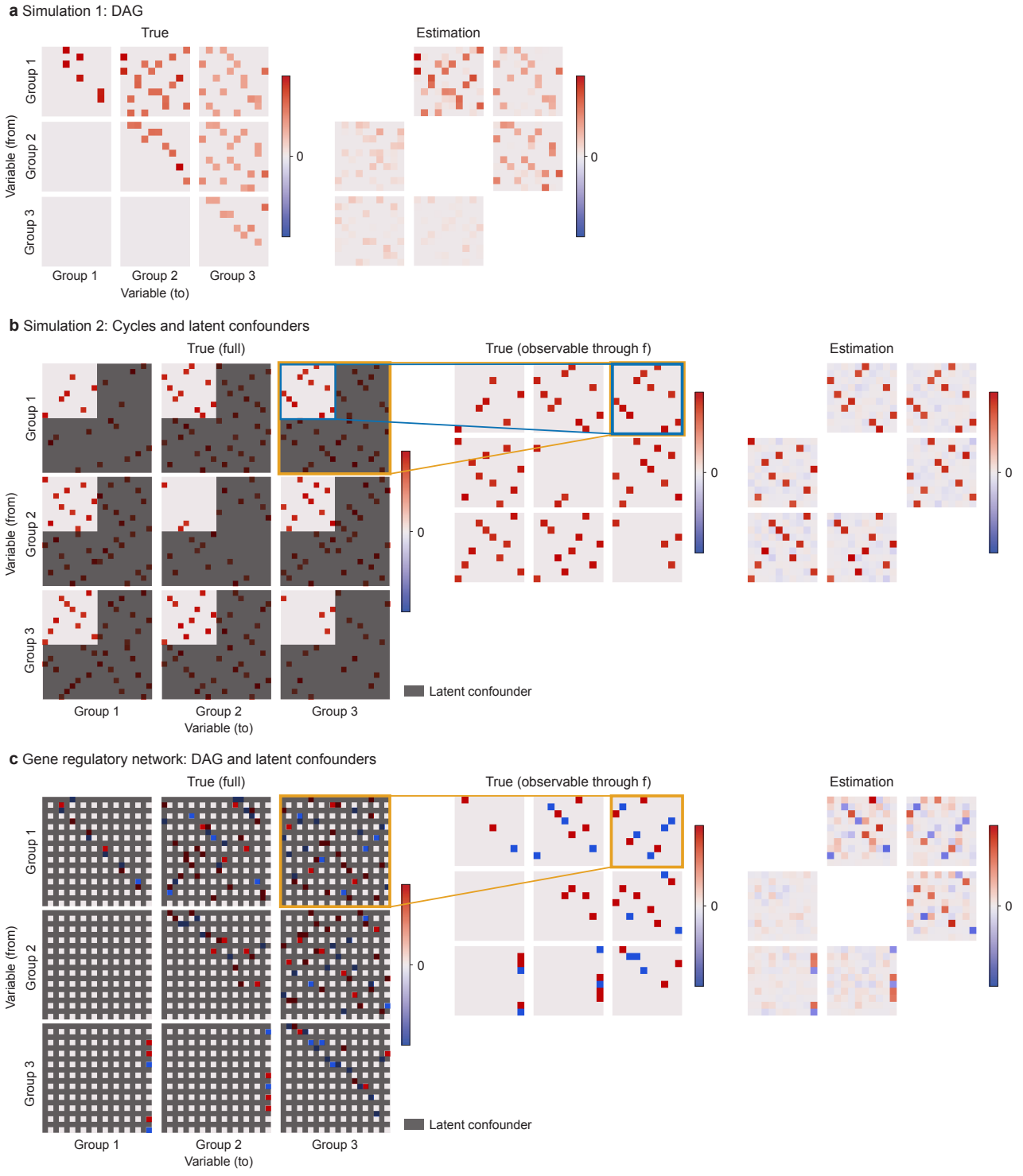

Figure 7. Example of the true causal structures (weighted adjacency matrices) and the estimations by G-CaRL (before causal direction determination and thresholding) in Simulation 1 (**a**), Simulation 2 (**b**), and the gene regulatory network recovery task (**c**). G-CaRL only identifies the inter-group-parts of the adjacency matrix, and thus the block-diagonal-parts are left unknown.

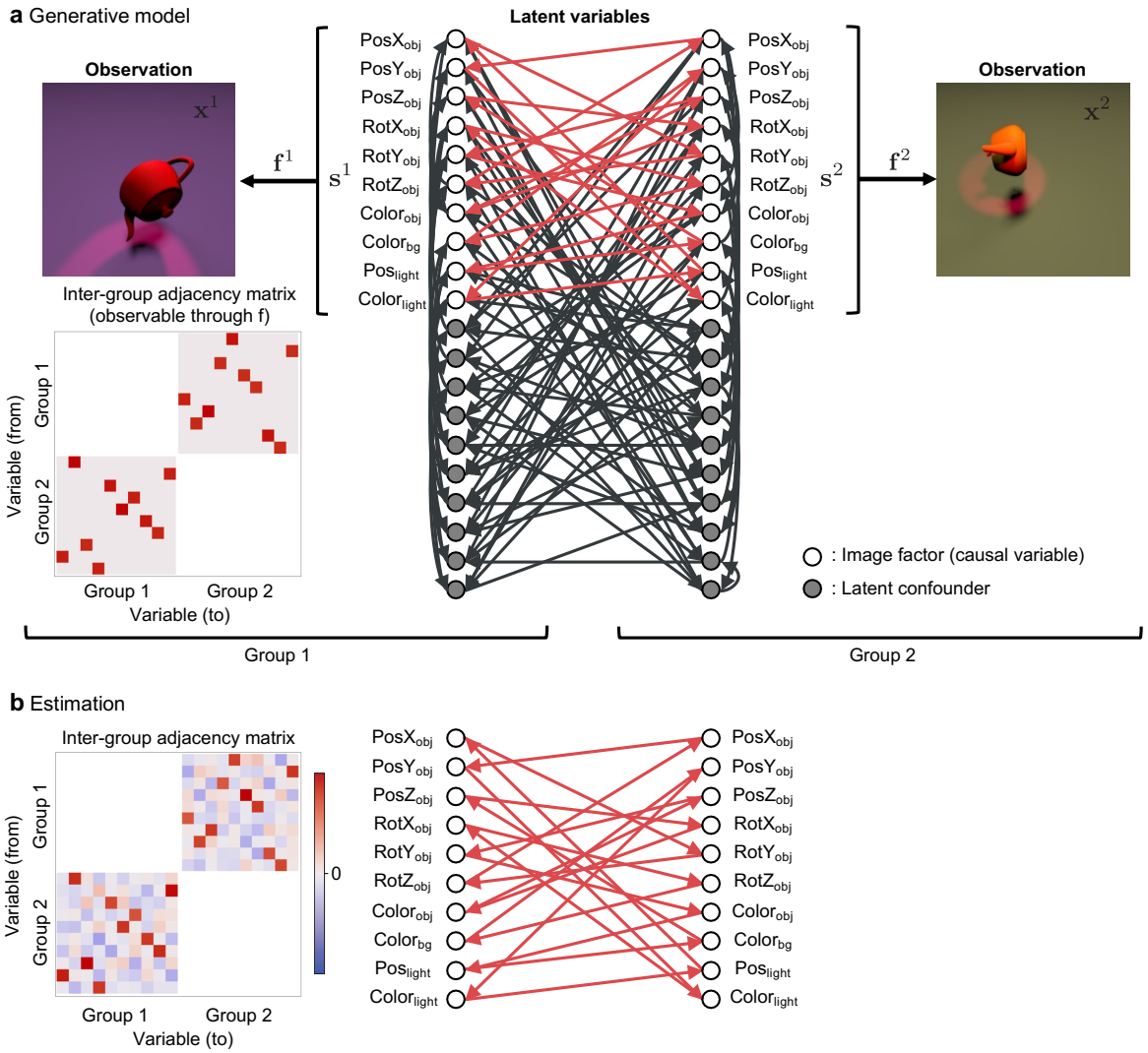

Figure 8. Evaluation of G-CaRL on high-dimensional image observations. We fixed the number of groups to $M = 2$ in this figure for brevity. (**a**) Example of the true causal structures (weighted adjacency matrices) with high-dimensional image observations, and (**b**) the estimation by G-CaRL (the directed graph visualization is after applying thresholding). Each group has ten image-factors (latent causal variables; white circles) conditioning the observation images, and additional ten latent confounders (gray circles) affecting the other variables. The ten image-factors are composed of XYZ positions of the object (PosX_{obj}, PosY_{obj}, and PosZ_{obj}), three dimensions describe the rotation of the object in Euler angles (RotX_{obj}, RotY_{obj}, and RotZ_{obj}), the color of the object and the ground of the scene ($\text{Color}_{\text{obj}}$ and Color_{bg}), and the position and color of the spotlight ($\text{Pos}_{\text{light}}$ and $\text{Color}_{\text{light}}$). The latent confounders do not have such physical interpretation, but still affect the observation images indirectly. What we aim to estimate are the causal edges colored by red in **a**, connecting the image-factors between groups, which are observable as high-dimensional images. G-CaRL only identifies the inter-group-parts of the adjacency matrix, and thus the intra-group graphs are left unknown. The causal graphs related to the latent confounders are also left unknown since the latent confounders are not observable.