Connectivity-Contrastive Learning: Combining Causal Discovery and Representation Learning for Multimodal Data

Hiroshi Morioka
RIKEN AIP

Aapo Hyvärinen
University of Helsinki
Université Paris-Saclay, Inria

Abstract

Causal discovery methods typically extract causal relations between multiple nodes (variables) based on univariate observations of each node. However, one frequently encounters situations where each node is multivariate, i.e. has multiple observational modalities. Furthermore, the observed modalities may be generated through an unknown mixing process, so that some original latent variables are entangled inside the nodes. In such a multimodal case, the existing frameworks cannot be applied. To analyze such data, we propose a new causal representation learning framework called connectivity-contrastive learning (CCL). CCL disentangles the observational mixing and extracts a set of mutually independent latent components, each having a separate causal structure between the nodes. The actual learning proceeds by a novel self-supervised learning method in which the pretext task is to predict the label of a pair of nodes from the observations of the node pairs. We present theorems which show that CCL can indeed identify both the latent components and the multimodal causal structure under weak technical assumptions, up to some indeterminacy. Finally, we experimentally show its superior causal discovery performance compared to state-of-the-art baselines, in particular demonstrating robustness against latent confounders.

1 INTRODUCTION

Estimating causal relations among a set of random variables based on purely observational data is called causal discovery, and has a great importance in a wide variety of fields, such as medicine, biology, finance, social sciences, and neurosciences. Causal discovery is generally formulated as estimation of a directed-acyclic graph (DAG) representing the causal relations of variables. However, learning a DAG from data is known to be highly challenging because the number of possible DAGs grows super-exponentially as a function of the number of variables (Robinson, 1977), and many different graphs can represent exactly the same data distribution (Andersson et al., 1997; Spirtes et al., 2001). There are mainly three approaches to tackle this problem (Glymour et al., 2019): the constraint-based, score-based, and asymmetry-based approaches. The score-based approach posits a scoring criterion for DAGs and then searches for the model with the highest score given the observations, while the constraint-based approach is based on conditional independence tests among variables. The asymmetry-based approach, which we use in this paper, formulates the problem as a continuous optimization of a statistical causal model, such as Bayesian network (BN) or structural equation model (SEM), with assumptions that make the model suitably asymmetric, such as nonlinearity or non-Gaussianity.

While most frameworks focus on causal discovery given a univariate variable for each node, in many applications we have access to multimodal observations (variables) for each node, for example with the same set of physical quantities across nodes. Examples include any kinds of sensor networks, such as sensor-arrays for climate monitoring (where a *node* is a single sensor location, and the *modalities* include temperature, humidity, rainfall (Longman et al., 2018), pollutants (Reani et al., 2022), and so on), or simultaneous measurements of multiple brain-imaging modalities (Huster et al., 2013; Shin et al., 2018) (where a *node* is a brain region), and so on.

If all variables are allowed to have causal relationships, the causal structure in such data can be simply represented by an adjacency matrix with the size of (modality $\times$ node) $\times$ (modality $\times$ node), though this leads to a very large matrix. On the other hand, if the measurement modalities are statistically mutually independent, as assumed in (Chen et al., 2021; Wang et al., 2020), the causal structure can be fully described by a three-way tensor (modality $\times$ node $\times$

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

node) representing separate modality-wise causal structures with no connections between modalities. Unfortunately in many cases, clearly including the examples just given, both of these simple frameworks fail because of dependencies *across* modalities, not due to causal relations but a mixing effect caused by (unknown) measurement process. An observational mixing clearly violates the assumption of causal relations between variables, which makes the conventional frameworks completely inapplicable.

Fortunately, there exists practical evidence, especially in the context of the recently developed nonlinear independent component analysis (NICA) (Hyvärinen and Morioka, 2016; Hyvarinen and Morioka, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020), that observations can often be disentangled into mutually independent latent components, which is a form of representation learning. If we can disentangle the observational mixing, obtaining *latent* modalities (components) which are mutually independent, the causal structure across node variables can often be properly described using causal connections *inside* those latent components. We can thus describe the causal structure hidden in the data by a three-way tensor (*component* $\times$ node $\times$ node). This is never possible by the existing NICA or causal discovery frameworks alone.

In this study, we propose a completely new framework, connectivity-contrastive learning (CCL), for causal discovery from multimodal node observations with unknown observational mixing (Fig. 1b). CCL jointly performs representation learning and causal discovery so that while the representation is learned for decomposing the observational modalities into latent modalities (components) *inside* each node (e.g. sensor location, brain region), the component-wise causal relations *over* node-variables are learned simultaneously. CCL assumes a generative model where the observational modalities are obtained as node-wise nonlinear mixtures of mutually independent latent components (Fig. 1a). Each component includes multiple node variables, which are causally generated based on a pairwise BN characterized by a component-specific adjacency matrix together with asymmetric (nonlinear) causal effect functions. CCL then learns the latent components together with their causal structures from the observations in a data-driven manner, by a new self-supervised learning framework: The pretext task is to predict the pair index, or location in the graph, given multimodal observations from every pairs of nodes. Although this self-supervised framework may not seem relevant to representation learning or causal discovery at first glance, we prove the identifiability of the latent components as well as of the multimodal causal structures (up to some inevitable indeterminacies) by proving the consistency of the model estimation by CCL. Experiments show that CCL can estimate multimodal causal structures, and beats the existing state-of-the-art methods, both on simple simulated data and a synthetic gene regulatory network recovery task, and this both with and without nonlinear observational mixing or latent confounders.

2 MODEL DEFINITION

We first present a short overview of the model definition (Fig. 1a). We obtain a two-dimensional matrix observation for each sample n; $\mathbf{X}^{(n)} = (x_{ai})^{(n)} \in \mathbb{R}^{p\times d}$, whose elements are obtained from multiple nodes $a \in \mathcal{V}$ ($|\mathcal{V}| = p$) with multiple observational modalities $i \in \{1, \dots, d\}$.[1] This is a general situation to consider, for example, a sensor array measuring multiple physical quantities (i) having unknown relations across some geographical locations (a) at different time points (n). We then assume that the observations are actually generated from a causally-structured latent matrix $\mathbf{S}^{(n)} = (s_{aj})^{(n)} \in \mathbb{R}^{p\times d}$, where a is the node from the same set $\mathcal{V}$ while $j \in \{1, \dots, d\}$ is a component index. The elements of $\mathbf{S}^{(n)}$ are assumed mutually independent across components j, while causally-structured across nodes $a \in \mathcal{V}$ for each j, as in Chen et al. (2021); Wang et al. (2020). If the latent matrix is directly given as the observation (i.e., $\mathbf{X}^{(n)} = \mathbf{S}^{(n)}$), the problem resembles that of Chen et al. (2021); Wang et al. (2020), and thus those multitask causal discovery frameworks would be applicable. However, we additionally assume that the observational matrix is obtained through an unknown observational mixing for each node a, which happens frequently in many practical situations, including the example of the sensor arrays above. This study considers the case where the all nodes has the same measurement process. Unfortunately, such observational mixings generally break the causal relations between elements of $\mathbf{X}$, and thus make the existing causal discovery frameworks on the observational space impossible. Our goal is to jointly estimate the latent matrix and its causal structures from the observations in a data-driven manner, which was never possible before.

Causally Structured Latent Components The latent matrix $\mathbf{S}^{(n)} = (s_{aj})^{(n)} \in \mathbb{R}^{p\times d}$ is generated probabilistically for each sample n, mutually independently across components j, while with causal structures across nodes $a \in \mathcal{V}$ within each j (Fig. 1a), based on a pairwise BN model, as described in detail below. Note that, based on the two-dimensional structure of $\mathbf{S}^{(n)}$ for given n, the elements of each *component* j corresponds to a vector $(s_{aj})_{a\in\mathcal{V}}$, i.e. the j-th column of $\mathbf{S}$, representing its realizations over the nodes $a \in \mathcal{V}$. This is a general setting to represent causal relations of multiple nodes having multiple (mutually independent) modalities j, similarly to Chen et al. (2021); Wang et al. (2020), and thus would be useful for many practical situations. We assume that the samples are generated independently, thus n does not need to be *time*, as in stud-

[1] We sometimes omit the sample index n below when it represents a random variable instead of a sample, or it is apparent from the context.

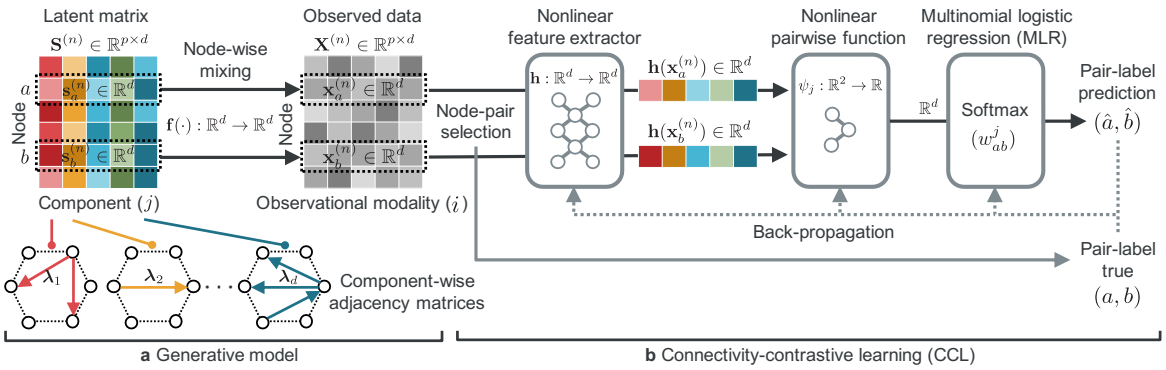

Figure 1: (**a**) Generative model combining causal structure and observational mixing. We obtain a two-dimensional matrix observation $\mathbf{X}^{(n)}$ for each sample n, whose elements are obtained from p nodes (rows) with d observational modalities (columns). The observations are obtained from a latent matrix $\mathbf{S}^{(n)}$ with the same set of nodes through unknown node-wise nonlinear mixings, which elements are mutually independent across components (columns) while causally-structured between nodes (rows) based on component-specific pairwise bayesian networks. This is a general model to consider, for example, a time series (n) from sensor array on multiple locations (row) with multiple physical quantities (column). (**b**) CCL jointly performs representation learning and causal discovery. We train a feature extractor by self-supervised multinomial logistic regression (MLR) which attempts to predict the pair-label (a, b) from paired observations $(\mathbf{x}_a, \mathbf{x}_b)$ for every pairs and samples, through ordinary back-propagation training. After training, the feature extractor $\mathbf{h}(\cdot)$ represents the latent components, and the weight parameters (w^j_{ab}) of MLR represent the adjacency matrices, up to some indeterminacy.

ies based on temporal dependency (e.g., Granger (1980); Lippe et al. (2022)). Our method is thus more general in the sense that no assumptions on possible time dependencies are made.

We model the causality of the node variables $(s_{aj})_{a\in\mathcal{V}}$ for each component j by a BN with some assumptions on it. BNs generally represent a causal graph by a factorization of the joint distribution by a product of conditional distributions (e.g., Choi et al. (2020); Park and Raskutti (2015));

$$p_j((s_{aj})_{a\in\mathcal{V}}) = \prod_{a\in\mathcal{V}} p_{aj}(s_{aj}|\,\mathrm{pa}(s_{aj})) \tag{1}$$

where $\mathrm{pa}(s_{aj})$ is the set of parents of variable s_{aj} on the causal graph of component j. Since this model is too general and is not identifiable in general (Andersson et al., 1997; Spirtes et al., 2001), we assume that Eq. 1 is further factorized and parameterized by the following *pairwise* form;

$$p_j((s_{aj})_{a\in\mathcal{V}}) \propto \prod_{a\in\mathcal{V}} \bar{q}_{aj}(s_{aj}) \prod_{(a,b)\in\mathcal{E}} \exp\left(\lambda^j_{ab}\phi_j(s_{aj}, s_{bj})\right), \tag{2}$$

where $\mathcal{E} \subset \mathcal{V}^2$ is the all directed pairs of nodes excluding the self-pairs ($\forall a$, $(a,a) \notin \mathcal{E}$), $\bar{q}_{aj}(\cdot) : \mathbb{R} \to (0,\infty)$ is a node potential function, $\phi_j(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is a nonlinear (asymmetric) pairwise causal effect function, and $\boldsymbol{\lambda}_j = (\lambda^j_{ab})_{(a,b)\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$ is coefficients modulating the causal effects across node-pairs. This factorization is possible (Eq. 1 can be given by Eq. 2), for example, when the causal graph is a (union of) directed rooted tree(s), where each node has only (up to) one parent, and each (cross-term of) conditional distribution $p_{bj}(s_{bj}|s_{aj})$ is given by an exponential family parameterized by λ^j_{ab} and ϕ_j. The function ϕ_j is specific to each component j and does not vary across pairs, while the coefficients $\boldsymbol{\lambda}_j$ controls the strengths of the relations across node-pairs. The coefficients (λ^j_{ab}) are supposed to be non-zero only when a is the direct cause of b (or the opposite, depending on the functional form of ϕ_j) on the causal graph of component j (note that the exponential function is constant when $\lambda^j_{ab} = 0$). Hence, although the nodes $\mathcal{V}$ and the pairs $\mathcal{E}$ are the same across components, the actual causal structures and ordering can be different across j due to the modulations by $\boldsymbol{\lambda}_j$. From those properties, $\boldsymbol{\lambda}_j$ can be interpreted as the *causal graph* (vectorized *adjacency matrix*), and estimating $\boldsymbol{\lambda}_j$ is equivalent to the *causal discovery* in this model. In some special cases, this pairwise BN causal model can be also represented by a nonlinear SEM (see Supplementary Material E), though such translation is not always possible. We will find below assumptions of ϕ_j and $\boldsymbol{\lambda}_j$ necessary for the identifiability of the model.

Observation Model The observed data consists of matrices $\mathbf{X}^{(n)} = (x_{ai})^{(n)} \subset \mathbb{R}^{p\times d}$, where $i = 1, \ldots, d$ is the observation modality index, while $a \in \mathcal{V}$ is the same node index as in the latent matrix $(s_{aj})^{(n)}$, for each sample n (Fig. 1a). The observations are obtained from the latent variables through an unknown observational mixing across components j, for each node a and sample n;

$$\mathbf{x}^{(n)}_a = \mathbf{f}(\mathbf{s}^{(n)}_a), \tag{3}$$

where $\mathbf{s}_a^{(n)} = (s_{aj})_j^{(n)} \in \mathbb{R}^d$ is the a-th row of $\mathbf{S}^{(n)}$, representing the set of d components at node $a \in \mathcal{V}$, $\mathbf{x}_a^{(n)} = (x_{ai})_i^{(n)} \in \mathbb{R}^d$ is similarly the a-th row of $\mathbf{X}^{(n)}$, i.e. the set of the d observational modalities at the same node, and $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$ is the observational mixing function, which is common across nodes. (See Section 6 for the possibility of different dimensions for the observational modalities and the latent components.) Importantly, the fundamental difference between the latent and the observational matrices is the *lack* or *presence*, respectively, of statistical dependence over the second index (components j or modalities i), induced by the observational mixing in the latter case. Although such dependency is inherent in many applications including the examples above, it breaks the causal relations between variables, and thus make the existing causal discovery frameworks inapplicable.

This observational model resembles NICA (Hyvarinen et al., 2019), except that the elements of each component here is a *vector* representing multiple nodes $a \in \mathcal{V}$ with causal relationships (Fig. 1a). Such combination of causality and observational mixing makes existing NICA frameworks inapplicable and thus requires a new framework which jointly performs the demixing and the causal discovery. Treating multimodal causal discovery as a novel type of nonlinear ICA, with this definition of a causal structure inside each component, and in particular in a directed, causal way, is the originality of our model.

Illustrative Example Consider a battery of environmental sensors which measure d physical quantities at p different geographical locations, and at different time points. Thus, the observed data is a tensor with the three indices: location (a), physical quantity (i), and time (n). The measurements are not independent across locations; the measured phenomena at one location are causally affected by those at other (nearby) locations but in a very complex way. Then, the multidimensional signal at geographical location a and time point n can be used in our model as $\mathbf{x}_a^{(n)}$ as in Eq. 3. The observed matrix $\mathbf{X}^{(n)}$ contains a temporal snapshot of the multidimensional measurements $\mathbf{x}_a^{(n)}$ at all locations $a \in \mathcal{V}$, and is observed for many time points n. This measured signal is assumed to be a mixture of underlying components $\mathbf{s}_a^{(n)}$ that we want to recover, since the sensors do not directly give the essential disentangled quantities. The component at a given location $s_{aj}^{(n)}$ has component-specific causal relations with those on the other locations $(s_{bj}^{(n)})_{b \neq a}$ based on unknown causal graphs $\boldsymbol{\lambda}_j$ and relations ϕ_j.

3 CONNECTIVITY-CONTRASTIVE LEARNING

We propose a novel self-supervised method called CCL for estimating the model just defined (Fig. 1b). Estimation is based on a multinomial logistic regression (MLR) problem, whose goal is to well predict the node-pair label $(a, b) \in \mathcal{E}$ of paired node observations $(\mathbf{x}_a^{(n)}, \mathbf{x}_b^{(n)})$, for every node-pair $\mathcal{E}$ and sample n. The learning procedure is based on the minimization of a softmax loss, as generally done in supervised classification problems, but here with a specific form of the softmax function, and *node-paired* inputs;

$$L = -\sum_n \sum_{(a,b)\in\mathcal{E}} \frac{\exp(\sum_{j=1}^d z_{ab}^j(\mathbf{x}_a{}^{(n)}, \mathbf{x}_b{}^{(n)}))}{\sum_{(l,m)\in\mathcal{E}} \exp(\sum_{j=1}^d z_{lm}^j(\mathbf{x}_a{}^{(n)}, \mathbf{x}_b{}^{(n)}))}, \tag{4}$$

where $z_{ab}^j(\mathbf{y}_1, \mathbf{y}_2) = w_{ab}^j \psi_j(h_j(\mathbf{y}_1), h_j(\mathbf{y}_2)) + w_{ba}^j \psi_j(h_j(\mathbf{y}_2), h_j(\mathbf{y}_1)) + \bar{\psi}_{ab}^j(h_j(\mathbf{y}_1)) + \bar{\psi}_{ba}^j(h_j(\mathbf{y}_2)) + b_{ab}$, (w_{ab}^j) and (b_{ab}) are the weight and bias parameters, h_j is the jth element of a (nonlinear) feature extractor $\mathbf{h}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, and $\psi_j(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ and $\bar{\psi}_{ab}^j(\cdot) : \mathbb{R} \to \mathbb{R}$ are scalar-valued nonlinear functions. All of those parameters are learned from the observational data $\{\mathbf{X}^{(n)}\}_n$ in a data-driven manner so as to optimize the loss function. The nonlinear functions are assumed to have universal approximation capacity (Hornik et al., 1989), and would typically be learned as neural networks. We can use any optimization method; a pseudo-code based on a basic SGD is given in Supplementary Algorithm 1.

CCL can be seen as a new type of contrastive-learning, which performs representation learning by taking contrast of pairwise marginal distributions (numerator of Eq. 4) with that of the all other node-pairs (denominator). This is the reason why this framework is called *connectivity*-contrastive learning. Interestingly, although this learning framework may not seem to be related to representation learning or causal discovery at first sight, our theorems below show that it actually achieves both of them simultaneously, by learning the function $\mathbf{h}(\cdot)$ and the weights (w_{ab}^j) of the softmax function (Eq. 4). The crucial point is the functional form of Eq. 4 specially designed to be consistent with the generative model shown above (such as mutual independence across j represented by the summation $\sum_j$, and the pairwise BNs parameterized by ψ_j and w_{ab}^j). With some additional assumptions on the generative model, we can guarantee that the best loss-score can be only achieved by the model satisfying those assumptions (up to some extent), which automatically leads to the representation learning and causal discovery.

The identification of the latent components is given by the following Theorem, proven in Supplementary Material A:

Theorem 1. *Assume the following:*

1. *($\mathbf{X}$) We obtain a dataset of two-dimensional observations $\{\mathbf{X}^{(n)}\}_n$ generated as node-wise nonlinear mixtures of latent components (Eq. 3), where the unknown mixing $\mathbf{f}$ is invertible and sufficiently smooth.*

2. *($\mathbf{S}$) The elements of latent matrices $\{\mathbf{S}^{(n)}\}_n$ are generated mutually independently across components j, while with causal relations across nodes $a \in \mathcal{V}$ represented by a pairwise BN (Eq. 2) within each j, for each sample n.*

3. *(ϕ) The causal effect function ϕ_j in Eq. 2 is asymmetric, in the sense that there is at least one point $(z_{1j}, z_{2j}) \in \mathbb{R}^2$ where $\phi_j^{12}(z_{1j}, z_{2j}) \neq \phi_j^{12}(z_{2j}, z_{1j})$ for each j, where $\phi_j^{12}(\eta_1, \eta_2) = \frac{\partial^2}{\partial \eta_1 \partial \eta_2} \phi_j(\eta_1, \eta_2)$.*

4. *($\boldsymbol{\lambda}$) The underlying undirected structure of $\boldsymbol{\lambda}_j$ is acyclic for all j, and each $\boldsymbol{\lambda}_j$ is sufficiently distinct across components j, in the sense that the concatenated matrix $[\bar{\mathbf{L}}, \bar{\mathbf{L}}']$ has full column rank $2d$, where $\bar{\mathbf{L}} = (\lambda_{ab}^j - \lambda_{a^*b^*}^j)_{(a,b),j}$ is a matrix of modulation coefficients from which some pivot pair (a^*, b^*) is subtracted, with the all pairs $(a, b) \in \mathcal{E}$ giving row index and the component j the column index, and similarly with $\bar{\mathbf{L}}' = (\lambda_{ba}^j - \lambda_{b^*a^*}^j)_{(a,b),j}$.*

5. *(CCL) We train MLR given by the loss function Eq. 4 with universal approximation capability.*

6. *($\mathbf{h}$) The function $\mathbf{h}$ in Eq. 4 is invertible.*

7. *(ψ) Each ψ_j in Eq. 4 is asymmetric, in the sense that $\frac{\partial}{\partial z_1} \log \psi_j^{12}(z_1, z_2) \neq \frac{\partial}{\partial z_1} \log \psi_j^{12}(z_2, z_1)$, for all j and for almost all $(z_1, z_2) \in \mathbb{R}^2$, where $\psi_j^{12}(\eta_1, \eta_2) = \frac{\partial^2}{\partial \eta_1 \partial \eta_2} \psi_j(\eta_1, \eta_2)$.*

Then, in the limit of infinite samples, $\mathbf{h}$ in the regression function provides a consistent estimator of the latent components: The output of the feature extractor $\mathbf{h}(\mathbf{x}_a^{(n)}) = (h_1(\mathbf{x}_a^{(n)}), \ldots, h_d(\mathbf{x}_a^{(n)}))^T$ gives the latent components $\mathbf{s}_a^{(n)} = (s_{aj})_j^{(n)}$, up to permutation and scalar (component-wise) invertible transformations for all $a \in \mathcal{V}$ and n.

This theorem guarantees the (local) convergence (i.e. statistical consistency) of the learning algorithm, which immediately implies the identifiability of the components, up to some indeterminacy.

Assumption 3 implies that ϕ_j needs to have sufficient asymmetricity, and excludes for example a linear autoregressive model with Gaussian innovations; $p_{bj}(s_{bj}|s_{aj}) \propto \exp(s_{bj} - \lambda_{ab}^j s_{aj})^2$, where ϕ_j is obtained as a symmetric form $\phi_j(s_{aj}, s_{bj}) \propto s_{aj} s_{bj}$, which is consistent with the well-known result in the causal discovery studies (Shimizu et al., 2006).

The acyclicity of the *underlying* undirected causal structure of $\boldsymbol{\lambda}_j$[2] (Assumption 4) is required to ensure that the cross-term of the pairwise marginal distribution $p_{ab}^j(s_{aj}, s_{bj})$ is given by a specific form parameterized only by λ_{ab}^j, λ_{ba}^j, and $\phi_j(\cdot, \cdot)$ for every node-pair $(a, b) \in \mathcal{E}$ and component j (see Supplementary Material A), and it includes forests with asymmetric weights, directed forests, and so on. Although this assumption excludes general DAGs, which can have cycles in the underlying undirected structures in general, we will experimentally show that CCL can still work on more general types of graphs (see Experiments 5). Note that for Theorem 1 (extraction of latent components), the graph structures do not need to be strictly directed, but just need to be asymmetric (i.e. both λ_{ab}^j and λ_{ba}^j can have non-zero values), in contrast to the Theorem 2 below for the causal discovery on directed graphs.

The assumption also requires the graph structures to be different across components j to some extent, so as to have a sufficiently strong and diverse difference of the pairwise distributions of the variables across node-pairs $(a, b) \in \mathcal{E}$, to be discriminated by MLR. This assumption requires at least $|\mathcal{E}| = p^2 - p \geq 2d$, which would not be difficult since we usually have much larger number of node-pairs than the components; i.e., $|\mathcal{E}| = p^2 - p \gg 2d$.

The assumptions of the nonlinear functions to be trained in Eq. 4 (Assumptions 6 and 7) apply *after learning*. Although they are not trivial, we assume they are only necessary to have a rigorous theory, and immaterial in practice.

The identifiability of the adjacency matrices $\boldsymbol{\lambda}_j$ requires additional assumptions on $\boldsymbol{\lambda}_j$ and (w_{ab}^j), and is given by the following theorem, proven in Supplementary Material B:

Theorem 2. *In addition to the assumptions of Theorem 1, assume the following:*

8. *($\boldsymbol{\lambda}$) The causal structures given by $\boldsymbol{\lambda}_j$ are never bidirectional (meaning at least one of λ_{ab}^j and λ_{ba}^j is 0 for each $(a, b) \in \mathcal{E}$) for all components j, and there is at least one pair $(a^*, b^*) \in \mathcal{E}$ where $\lambda_{a^*b^*}^j = \lambda_{b^*a^*}^j = 0$ for all j.*

9. *(w) $(w_{ab}^j - w_{a^*b^*}^j)_{(a,b)\in\mathcal{E}_j'} \in \mathbb{R}^{|\mathcal{E}_j'|}$ and $(w_{ba}^j - w_{b^*a^*}^j)_{(a,b)\in\mathcal{E}_j'} \in \mathbb{R}^{|\mathcal{E}_j'|}$ are linearly independent, or one of them is a zero vector, where $\mathcal{E}_j'$ is the set of pairs $(a, b)_{a<b}$ in the causal graph of j whose nodes are arranged in a causal order, such that no later node b causes any earlier node $a < b$, for each j.*

Then, $\mathbf{w}_{\sigma(j)} = (w_{ab}^{\sigma(j)})_{(a,b)\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$ in the regression function gives either $\boldsymbol{\lambda}_j = (\lambda_{ab}^j)_{(a,b)\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$ or its matrix-transpose $\boldsymbol{\lambda}_j' = (\lambda_{ba}^j)_{(a,b)\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$, up to a linear scaling and a bias, where $\sigma(j)$ represents the permutation of components (that is indeterminate according to Theorem 1).

This theorem shows that the weight parameters (w_{ab}^j) in the regression function (Eq. 4) represent the hidden causal structures (λ_{ab}^j) up to some indeterminacy after the training.

CCL cannot perfectly determine the causal graphs in the

[2]Note that $\boldsymbol{\lambda}_j$ themselves are not undirected.

sense whether the estimation $\mathbf{w}_{\sigma(j)}$ gives $\boldsymbol{\lambda}_j$ or its matrix-transpose $\boldsymbol{\lambda}'_j$, and these transposes can be different across j. This is caused by the indeterminacy of the functional causal direction of ϕ_j in Eq. 2, and thus inevitable; if the functional direction of ϕ_j is flipped and $\boldsymbol{\lambda}_j$ is transposed, they simply lead to the same model, and thus cannot be distinguished only from the observations without further assumptions on ϕ_j. In practice, the directions can be determined by a prior knowledge about the causal direction of at least one edge for each j.

Assumption 8 together with 4 requires the graph structures to be directed forests. We emphasize again that although this assumption is stronger than general DAGs assumed in ordinary causal discovery, our experiments show that CCL can still work well on more general DAGs. The existence of a node-pair without any connection (end of Assumption 8) should be easily satisfied in practice.

Note that the assumptions require the graph structures to be only *partially* distinctive across components, rather than completely different; at least one edge needs to be different from that on the other components, but the other parts of the graphs can have the same structures.

The assumption on (w_{ab}^j) (Assumption 9) is not trivial, but it can be easily verified after learning, and could be achieved by extra constraints during training.

4 ALTERNATIVE THEORY

We can also construct an identifiability theory of the latent components based on rather different assumptions. To this end, we adapt the theory of Hyvarinen and Morioka (2017); Hälvä et al. (2021). While those papers considered stationary time series, they did it in a way that has some resemblance to a multimodal modelling: They transformed time series into a multimodal data where the modalities are obtained by taking time windows, in particular $(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)})$ in the basic case. On the other hand, they make no reference to causality, but the asymmetric models we use for the causal discovery can be considered as simply one possible model of dependencies that still fit some of the assumptions of those studies.

The alternative theorem guarantees the identifiability of the latent components (but not the causal structures), only making assumptions on a *single* specific node-pair, rather than considering the whole set of pairs as in CCL. Due to lack of space, please refer to Supplementary Material C for the detailed description of the theorem and the proof. Basically, we assume that there exists a single node-pair $(a^*, b^*) \in \mathcal{E}$ where the joint distributions of the paired latent variables (s_{a^*j}, s_{b^*j}) are not *locally quasi-Gaussian* (Hälvä et al., 2021) for all j. We can further propose an estimation framework consistent with that theory, which we call *CCLalt*, and estimate the latent components only from the observations obtained from that single node-pair. We prove the identifiability of the latent components by proving the consistency of the model estimation by CCLalt (up to the same level of indeterminacies given by CCL).

This alternative theorem has an advantage that the constraint on the whole causal graph is weaker than that in CCL (Assumption 4) because it is only based on the joint distributions on a single node-pair, though it still would imply some level of constraints on the whole causal graphs.

5 EXPERIMENTS

5.1 Simulation 1: Multimodal DAG Causal Discovery

Data Generation We generated artificial data based on the generative model described in Section 2. The number of nodes (p) was fixed to 30, the number of modalities and components (d) was 10, and the number of data points n was $4,096$. The causal graphs of the latent components were designed to be DAGs (no directed cycles), though their underlying undirected graphs can have cycles and thus do not really satisfy Assumption 4. We obtained the observations directly from the latent matrix ($\mathbf{X}^{(n)} = \mathbf{S}^{(n)}$; denoted as $L = 0$), or with nonlinear observational mixings ($L = 3$). See Supplementary Material E for more details of the experimental settings.[3]

Training and Evaluation For CCL, we trained the MLR (Eq. 4) from the observed data, with using multilayer perceptrons (MLPs) for the nonlinear functions. CCLalt is also applied to the same data. Note that CCLalt can perform only the latent components estimation, but not causal discovery.

The estimated latent components and the causal structures were then evaluated by comparing them to the true ones. The latent components were evaluated by Pearson correlation, while the adjacency matrices are evaluated by precision, recall, and F1-score, after binarizing both of the estimations and the true ones. All of them were averaged over 10 runs.

For comparison, we also applied PC (Spirtes and Glymour, 1991), FGS (Ramsey et al., 2017), GFCI (Ogarrio et al., 2016), CAM (Bühlmann et al., 2014), DirectLiNGAM (Shimizu et al., 2011), NOTEARS-linear (Zheng et al., 2018), NOTEARS-MLP (Zheng et al., 2020), and MultiDAG (Chen et al., 2021) to the same data. See Supplementary Material G for their detail. Basically, all baselines estimate all of the d-component causal graphs separately or jointly, and then are evaluated with the same way as in CCL, so as to make the comparisons as fair as possible.

Results Firstly and notably, CCL showed the best causal discovery F1-score compared to the baselines on the directly-

[3]The codes are available at `https://github.com/hmorioka/CCL`.

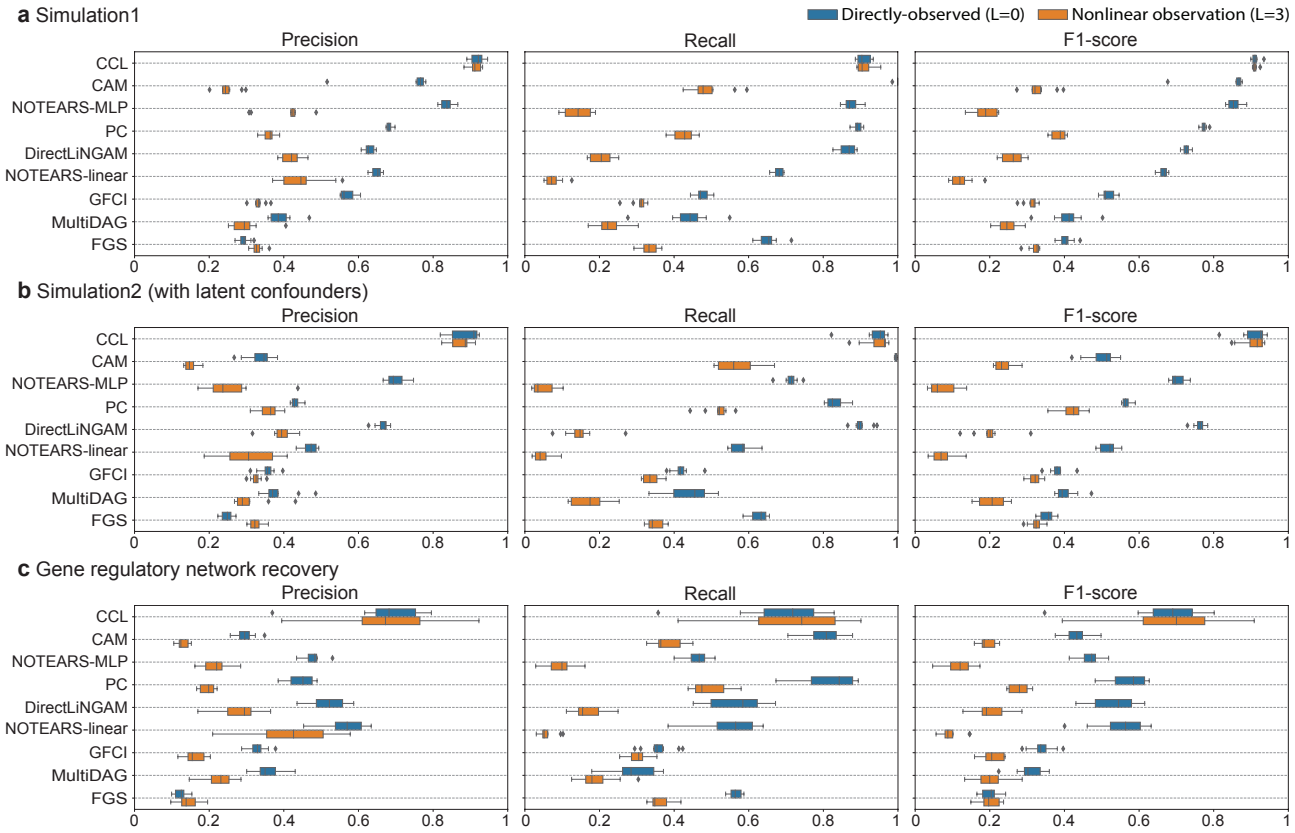

Figure 2: Comparisons of the causal discovery performances of multimodal DAGs by the proposed framework CCL and the baselines. (**a**) Simulation 1: The estimation performances measured by precision, recall, and F1-score. The performances were evaluated for each of directly-observed case ($L = 0$) and the case with unknown nonlinear observational mixings across modalities ($L = 3$). CCL shows the best F1-scores on both settings. (**b**) Simulation 2: Evaluation of robustness against presence of latent confounders. The similar performances of CCL to those in **a** show the robustness of CCL. (**c**) Gene regulatory network recovery task: CCL showed the best performances.

observed cases (Fig. 2a; $L = 0$). This is presumably due to the estimation principle of CCL, which jointly estimates the multimodal causal structures by combining the whole information from the all modalities, rather than estimating them separately as in many of the baselines. Importantly, although the true causal graphs have cycles in their underlying undirected graphs and thus do not satisfy Assumption 4 (see Supplementary Fig. 6 for some examples), the results show that CCL can still work on this more general type of DAGs. Although MultiDAG also considers multimodal structures, the performances were worse than CCL presumably because of the *nonlinear* causal model here. Especially nonlinear methods (CCL, CAM, and NOTEARS-MLP) worked well here because they can consider nonlinear causal effects.

Crucially, CCL could reconstruct the causal structures even with unknown nonlinear observational mixings ($L = 3$) with comparable performances to those without the mixings ($L = 0$; Fig. 2a, and Supplementary Fig. 6 for some examples). Supplementary Figs. 3 and 4 show how the complexity L, the number of nodes p, modalities d, and n affect the performances; a higher L, p, and d make learning more difficult, and a larger amount of n make it possible to achieve higher performances, as expected. The performances of the baselines were hugely deteriorated by the observational mixings because they cannot perform the demixing (representation learning) by themselves. We also applied the baselines on the latent components estimated by CCL (the output of $\mathbf{h}$; Supplementary Fig. 5). The results show improved estimation performances for the baselines compared to Fig. 2a, but they are still worse than CCL. This result indicates the importance of performing both representation learning and causal discovery simultaneously, as in CCL.

CCL could also reconstruct the latent components reasonably well even with unknown nonlinear observational mixings across modalities (Supplementary Fig. 3a Correlation). CCLalt also succeeded to reconstruct the latent components on the same dataset (Supplementary Fig. 8). However, the performances were slightly worse than CCL, presumably

because it only uses the information on a specific node-pair, unlike CCL which utilizes the whole node-pair information.

5.2 Simulation 2: Latent Confounders

To show robustness of CCL against presence of unobservable latent confounders, we evaluated CCL by artificial data similarly to Simulation 1, but now with unobservable (masked) nodes. More specifically, we firstly generated latent components with 60 nodes with the same manner to Simulation 1, and then simply masked half of the nodes as unobservable nodes (latent confounders) alternately; there are $p = 30$ observable nodes and 30 latent confounders. We used the same regression model, training, and evaluation methods. CCL showed slightly lower but comparable performances to those in Simulation 1 even with the latent confounders (Fig. 2b). On the other hand, the performances of the baselines were severely deteriorated by the presence of the confounders. This result shows the higher robustness of CCL against latent confounders.

The latent components were also well-reconstructed by CCL even with latent confounders, though it required larger number of samples compared to the case without latent confounders (Supplementary Fig. 3b Correlation).

5.3 Recovery of Gene Regulatory Network

Methods We also evaluated CCL on a more realistic causal model. Since real (especially multimodal) data generally do not have information of the true causal relations behind, we used synthetic single-cell gene expression data generated by SERGIO (Dibaeinia and Sinha, 2020), similarly to (Chen et al., 2021). We emulated a situation where we have latent matrices $\{(s_{aj})^{(n)}\}_n$ representing steady-state expression levels of causally-interacting multiple genes (nodes a), measured individually from many cells (samples n) under multiple conditions (components j) each, such as different developmental stages of the cells (Kojima et al., 2017) or perturbations (Sachs et al., 2005), which are known to lead to distinctive causal interactions of genes. The observations $\{(x_{ai})^{(n)}\}_n$ are then obtained as gene-wise (node-wise) nonlinear mixtures of the expression levels across the multiple conditions (components). The goal of this experiment is to estimate the condition-specific gene regulatory networks (GRNs) from the nonlinearly mixed observations.

We used similar causal DAGs as in the previous sections, but with more difficult settings; We generated the expression level data of 100 genes (nodes a) in 8,192 cells (samples n) under 10 different measurement conditions of the cells (components j) having distinctive GRNs. Each GRN is designed so that each gene has (approximately) two activator and two repressor parents. We then selected half of the genes as observable nodes alternately ($p = 50$), and left the others as latent confounders. See Supplementary Material F for some additional information.

Results CCL showed the best performances among the baselines (Fig. 2c), which indicates the good applicability of CCL to real datasets for causal discovery. Although MultiDAG was originally applied on synthetic data generated by SERGIO in Chen et al. (2021) and showed reasonable performances, its performance was worse here. This would be because of the different settings of the gene dynamics, which are more close to Dibaeinia and Sinha (2020) here, and the denser connections of the GRNs.

6 DISCUSSION

This study has an important novelty from the both aspects of representation learning and causal discovery. Although the multimodal (multi-task) causal structure considered here is similar to those in Chen et al. (2021); Wang et al. (2020) and itself would have great practical importance, there are further advances in this study. Firstly, CCL assumes *nonlinear* causal relations between variables, rather than linear ones. Secondly, CCL utilizes the *differences* of the causal structures across modalities for the estimation, rather than the *consistency* (same causal orderings) as in Chen et al. (2021); Wang et al. (2020), which might be too restrictive in some applications. Thirdly and the most importantly, CCL also performs disentanglements of observational mixings (representation learning) *jointly* with the causal discovery on the latent space, which was never considered before. In the aspect of the representation learning, CCL can be seen as a novel NICA framework, but it considers two-dimensional structures of the variables (multimodal observations from multiple nodes) and also estimates hidden causal structures, which are both new compared to existing NICA (Hälvä and Hyvärinen, 2020; Hyvarinen et al., 2019; Khemakhem et al., 2020; Morioka et al., 2021). See Supplementary Material D for more details of the related works.

CCL is not a simple aggregation of representation learning and causal discovery; those two are not mutually orthogonal, but closely related each other. Firstly, without the representation learning, we cannot perform the causal discovery because the causal relations between variables are broken by unknown observational mixings (Eq. 3). Secondly, without assuming the causal structures, representation learning (demixing) is infeasible because of the well-known indeterminacy of NICA (Hyvärinen and Pajunen, 1999). CCL smartly utilizes such hidden causal structures to perform the two tasks jointly as a novel contrastive learning, achieving better causal discovery performance than any baseline.

CCL is based on the idea of contrastive-learning (Gutmann and Hyvärinen, 2012; Sugiyama et al., 2012) or self-supervised learning Jaiswal et al. (2021), which have been receiving increasing attention recently. However, this is the first study which uses the *node-pair-indices* of node-paired inputs as the auxiliary labels for taking contrasts, and then jointly achieves causal discovery, which is also

quite new. Although the definition of contrast in CCL seems to be conceptually different from other major contrastive-learning frameworks measuring dissimilarity of some data samples (positive samples) to the other set of samples (negative samples), they are somewhat connected. The softmax function in Eq. 4 represents the contrast of how likely the paired-observation $(\mathbf{x}_a, \mathbf{x}_b)$ is obtained from the true node-pair (a, b) (numerator) compared to the all other pairs $\mathcal{E}$ (denominator). This indicates that minimizing Eq. 4 implicitly enforces the feature values from the positive samples $(\mathbf{x}_a, \mathbf{x}_b)$ to be somehow distinctive from those from the all other negative samples $(\mathbf{x}_l, \mathbf{x}_m)_{(l,m)\neq(a,b)}$, similarly to the other contrastive-learning frameworks.

The identifiability is basically guaranteed by *nonstationarity* of the distributions of paired observations $(\mathbf{x}_a, \mathbf{x}_b)$ across node-pairs $\mathcal{E}$, which is satisfied by the distinctiveness of the causal structures across components (Assumption 4). CCL thus can be seen as an extension of NICA based on temporal nonstationarity (Hyvärinen and Morioka, 2016) to causal graphs and multimodal data.

We can consider some possible generalizations of CCL. Firstly, the exponential function in Eq. 2 representing the (cross-term of) conditional distribution can be higher order: $\exp\left(\sum_k \lambda_{ab}^{jk} \phi_{jk}(s_{aj}, s_{bj})\right)$ with $k > 1$. Such a model would have higher representational power, and should be useful to model the conditional distributions based on, for example, a Gaussian distribution with scaling and bias, Beta distribution, and so on. Our proofs can be easily extended to this model. Secondly, although we have focused exclusively on pairwise interaction models (Eq. 2), it should be possible to develop CCL for higher-order interactions, where the exponential function in Eq. 2 is represented by a combination of more than two latent node variables, where the causal graph $\boldsymbol{\lambda}_j$ is represented by a high-order tensor.

We can also generalize CCLalt to consider multiple node-pairs simultaneously, rather than focusing only on a single node-pair. This would weaken the assumption of the non-quasi-Gaussianities of the "*all* components on a specific node-pair," to those of "*some* of the components for each node-pair." This should also improve the performance compared to the single node-pair case used above.

Although we assume that the number of the latent components and the observational modalities are the same (d), it would be possible to consider, especially, the case where the number of the latent components is smaller than that of the observation modalities, with the similar claim used in Hyvärinen and Morioka (2016); it is enough to assume that while we formally have the same dimensions, there exist some latent components which do not have any causal relations across nodes, and those would be automatically ignored in CCL learning.

Since CCL needs to solve a $|\mathcal{E}| = p^2 - p$ class classification problem, it would be computationally inefficient in large network data (say p is more than several hundreds). Using special classifiers designed for a high number of classes (e.g., Babbar and Schölkopf (2017)) should be able to solve such issue. Nevertheless, CCL shows decent scalability up to $p = 64$ (or possibly up to 128; around 10 thousands of pair-labels) in our experimental setting (Supplementary Fig. 4b), which is comparable to the state-of-the-art baselines (e.g., see Fig. 3 of Zheng et al. (2018)), even with nonlinear observational mixings in CCL.

7 CONCLUSION

This study proposed a novel framework called CCL for jointly performing representation learning and causal discovery. In contrast to the conventional causal discovery frameworks which assume univariate node-variables, CCL assumes that each node has multiple observational modalities (variables) with mutual dependency due to an unknown observational mixing process. CCL then estimates mutually independent latent components together with a specific causal structure over node-variables for each of them, using a novel self-supervised learning method in a data-driven manner. Our theorems showed identifiability of the model and the consistency of the estimation method. A crucial assumption is the pairwise BN causal model with distinctive graph structures across components. Experiments using synthetic data and more-realistic gene expression data showed that CCL works better than the state-of-the-art causal discovery baselines, both when nonlinear observational mixing is present or not, and even in the case with latent confounders. Since such multimodal causal structures together with observational mixing and latent confounders are inherent in many data, CCL has a great potential for application.

Acknowledgements

This research was supported in part by JST PRESTO JPMJPR2028, JSPS KAKENHI 22H05666, 22K17956, and 19K20355. 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.

References

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

R. Babbar and B. Schölkopf. DiSMEC: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 721–729, 2017.

H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

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

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

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1597–1607, 2020.

X. Chen, H. Sun, C. Ellington, E. Xing, and L. Song. Multi-task learning of order-consistent causal graphs. In *Advances in Neural Information Processing Systems*, volume 34, pages 11083–11095, 2021.

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

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

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

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

C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

J. Fan. Multi-mode deep matrix and tensor factorization. In *International Conference on Learning Representations*, 2022.

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

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

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

C. W. J. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2: 329–352, 1980.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

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

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

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

R. A. Harshman and M. E. Lundy. PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72, 1994.

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

H. Hosoya. CIGMO: Learning categorical invariant deep generative models from grouped data. In *Workshop on Weakly Supervised Learning (ICLR 2021)*, 2021.

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

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

R. J. Huster, S. Debener, T. Eichele, and C. S. Herrmann. Methods for simultaneous EEG-fMRI: An introductory

review. *Journal of Neuroscience*, 32(18):6053–6060, 2013.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, 10(3):626–634, 1999.

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

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

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

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

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

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

A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2021.

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

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

Y. Kojima, K. Sasaki, S. Yokobayashi, Y. Sakai, T. Nakamura, Y. Yabuta, F. Nakaki, S. Nagaoka, K. Woltjen, A. Hotta, T. Yamamoto, and M. Saitou. Evolutionarily distinctive transcriptional and signaling programs drive human germ cell lineage specification from pluripotent stem cells. *Cell Stem Cell*, 21(4):517–532.e5, 2017.

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

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

T. N. Maeda and S. Shimizu. 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*, pages 735–745, 2020.

F. Miwakeichi, E. Martinez-Montes, P. A. Valdés-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi. Decomposing EEG data into space–time–frequency components using parallel factor analysis. *NeuroImage*, 22(3):1035–1045, 2004.

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

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

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

N. Noorshams and M. J. Wainwright. Belief propagation for continuous state spaces: Stochastic message-passing with quantitative guarantees. *Journal of Machine Learning Research*, 14(49):2799–2835, 2013.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, volume 52, pages 368–379, 2016.

Z. Pan, Z. Wang, and S. Zhe. Streaming nonlinear bayesian tensor decomposition. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 490–499, 2020.

G. Park and H. Park. 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, pages 158–166, 2019a.

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

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

J. Pearl, editor. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

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

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

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.

M. Reani, D. Lowe, A. Gledson, D. Topping, and C. Jay. UK daily meteorology, air quality, and pollen measurements for 2016–2019, with estimates for missing data. *Scientific Data*, 9(1):43, 2022.

R. W. Robinson. Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics V*, pages 28–43, 1977.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

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

S. Shimizu and K. Bollen. 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.

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

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

J. Shin, A. von Lühmann, D.-W. Kim, J. Mehnert, H.-J. Hwang, and K.-R. Müller. Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. *Scientific Data*, 5(1):180003, 2018.

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

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 499–506, 1995.

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

E. B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willsky. Nonparametric belief propagation. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, 2003.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Y. Wang, S. Segarra, and C. Uhler. High-dimensional joint estimation of multiple directed Gaussian graphical models. *Electronic Journal of Statistics*, 14(1):2439 – 2483, 2020.

P. Wu and K. Fukumizu. 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, pages 1157–1167, 2020.

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

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

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

S. Zhe, K. Zhang, P. Wang, K.-c. Lee, Z. Xu, Y. Qi, and Z. Ghahramani. Distributed flexible nonlinear tensor factorization. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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

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

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

A PROOF OF THEOREM 1

Firstly, we reformulate the parameterized pairwise BN causal model (Eq. 2) by the following pairwise factor graph;

$$p_j((s_{aj})_{a\in\mathcal{V}}) \propto \prod_{a\in\mathcal{V}} \bar{q}_{aj}(s_{aj}) \prod_{(a,b)\in\bar{\mathcal{E}}} q^j_{ab}(s_{aj}, s_{bj}), \tag{5}$$

where $\bar{\mathcal{E}} \subset \mathcal{E}$ is the set of the all *undirected* pairs (i.e., combinations) of nodes, and q^j_{ab} is the pairwise clique potential on each node-combination, given by

$$q^j_{ab}(s_{aj}, s_{bj}) \propto \exp\left(\lambda^j_{ab}\phi_j(s_{aj}, s_{bj}) + \lambda^j_{ba}\phi_j(s_{bj}, s_{aj})\right), \tag{6}$$

which is simply a product of the two (directed) exponential factors in Eq. 2 on that combination. Note that $q^j_{ab}(s_{aj}, s_{bj}) = q^j_{ba}(s_{bj}, s_{aj})$ due to its symmetric functional form, and the clique potential is constant when s_{aj} and s_{bj} are not causally related ($\lambda^j_{ab} = \lambda^j_{ba} = 0$). In the pairwise factor graph represented by Eq. 5, the marginal distribution of a paired latent variables (s_{aj}, s_{bj}) for each $(a, b) \in \bar{\mathcal{E}}$ is given (or approximated) by brief propagation (BP) (Noorshams and Wainwright, 2013; Pearl, 1988; Sudderth et al., 2003), after the convergence, as

$$p^j_{ab}(s_{aj}, s_{bj}) \propto \bar{q}_{aj}(s_{aj})\bar{q}_{bj}(s_{bj})q^j_{ab}(s_{aj}, s_{bj}) \left[\prod_{k\in\mathcal{N}_j(a)\setminus b} m^j_{ka}(s_{aj})\right] \left[\prod_{k\in\mathcal{N}_j(b)\setminus a} m^j_{kb}(s_{bj})\right], \tag{7}$$

where $\mathcal{N}_j(a) \setminus b$ indicates the neighbors of node a except for node b in the graph of component j, $m^j_{ka}(s_{aj})$ is a message sending information from node k to a on component j, as a function of the state of node a (i.e., s_{aj}), via the recursive form

$$m^j_{ka}(s_{aj}) \propto \int \bar{q}_{kj}(s_{kj})q^j_{ka}(s_{kj}, s_{aj}) \prod_{l\in\mathcal{N}_j(k)\setminus a} m^j_{lk}(s_{kj})ds_{kj}. \tag{8}$$

The important implication of Eq. 7 is that the marginal distribution $p^j_{ab}(s_{aj}, s_{bj})$ is given as a combination of the pairwise clique potential $q^j_{ab}(s_{aj}, s_{bj})$ (Eq. 6) and the other functions only dependent on either s_{aj} or s_{bj}. For acyclic factor graphs (Assumptions 2 and 4), BP is guaranteed to converge after a finite number of iterations and Eq. 7 yields the exact marginal distributions, whereas for cyclic graphs it yields only the approximations, though still known to be a good one. Note that we do not need to perform BP for CCL, but just need to guarantee that the pairwise marginal distribution (Eq. 7) is given as a combination of the pairwise clique potential (Eq. 6) and univariate functions of s_{aj} and s_{bj}, which is satisfied by the existence of the exact solution by BP on the graph structures with Assumptions 2 and 4.

From the generative model, the pairwise marginal distribution of paired observations $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = (\mathbf{f}(\boldsymbol{\eta}_1), \mathbf{f}(\boldsymbol{\eta}_2)) \in \mathbb{R}^{d\times 2}$ on each node-pair $(a, b) \in \mathcal{E}$ is given by, using the probability transformation formula and that of the latent components Eq. 7,

$$\begin{aligned}
&\log p_{ab}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \\
&= \log p_{ab}(\mathbf{g}(\boldsymbol{\xi}_1), \mathbf{g}(\boldsymbol{\xi}_2)) + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)| \\
&= \sum_{j=1}^{d} \log p^j_{ab}(g_j(\boldsymbol{\xi}_1), g_j(\boldsymbol{\xi}_2)) + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)| \\
&= \sum_{j=1}^{d} \lambda^j_{ab}\phi_j(g_j(\boldsymbol{\xi}_1), g_j(\boldsymbol{\xi}_2)) + \lambda^j_{ba}\phi_j(g_j(\boldsymbol{\xi}_2), g_j(\boldsymbol{\xi}_1)) + \log\mu^j_{ab}(g_j(\boldsymbol{\xi}_1)) + \log\mu^j_{ba}(g_j(\boldsymbol{\xi}_2)) \\
&- \log Z^j_{ab} + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)|,
\end{aligned} \tag{9}$$

where $\mathbf{J}$ denotes the Jacobian, $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ is the (true) inverse function of the observational mixing $\mathbf{f}$ (Eq. 3), and thus $\eta_{1j} = g_j(\boldsymbol{\xi}_1)$ gives a single latent component for each $j = 1, \ldots, d$ by definition. The second equation comes from the mutual independence of the latent components, and third equation comes from the pairwise marginal distribution of the latent node-variables given by BP (Eq. 7) and the pairwise clique potential model (Eq. 6). $\mu^j_{ab}(\eta_{1j})$ represents functions depending only on a single variable η_{1j} in Eq. 7 (the node potential function and the messages), and similarly for $\mu^j_{ba}(\eta_{2j})$. Z^j_{ab} denotes the partition functions (normalization constants) of the marginal distributions.

On the other hand, on the optimal discrimination relation of MLR given by Eq. 4, the softmax function in Eq. 4 is supposed to represent the posterior distribution of the node-pair index $p((a,b)|\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ given a paired observation $(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ after the training. By applying Bayes rule on this, after dividing all the exponential terms by the one corresponding to the pair (a^*,b^*) assumed in Assumption 4 to avoid the well-known indeterminacy of the softmax function,

$$\begin{aligned}
&\log p_{ab}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\\
&=\sum_{j=1}^{d}(w_{ab}^j-w_{a^*b^*}^j)\psi_j(h_j(\boldsymbol{\xi}_1),h_j(\boldsymbol{\xi}_2))+(w_{ba}^j-w_{b^*a^*}^j)\psi_j(h_j(\boldsymbol{\xi}_2),h_j(\boldsymbol{\xi}_1))\\
&+\bar{\psi}_{ab}^j(h_j(\boldsymbol{\xi}_1))-\bar{\psi}_{a^*b^*}^j(h_j(\boldsymbol{\xi}_1))+\bar{\psi}_{ba}^j(h_j(\boldsymbol{\xi}_2))-\bar{\psi}_{b^*a^*}^j(h_j(\boldsymbol{\xi}_2))\\
&+\log p_{a^*b^*}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)+\alpha_{ab},
\end{aligned}\tag{10}$$

where $\alpha_{ab}=b_{ab}-b_{a^*b^*}-\log p(a,b)+\log p(a=a^*,b=b^*)$, which is a collection of the terms not dependent on $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$.

Setting Eq. 9 and Eq. 10 to be equal for arbitrary $(a,b)\in\mathcal{E}$, we have:

$$\begin{aligned}
&\sum_{j=1}^{d}(w_{ab}^j-w_{a^*b^*}^j)\psi_j(h_j(\boldsymbol{\xi}_1),h_j(\boldsymbol{\xi}_2))+(w_{ba}^j-w_{b^*a^*}^j)\psi_j(h_j(\boldsymbol{\xi}_2),h_j(\boldsymbol{\xi}_1))\\
&+\bar{\psi}_{ab}^j(h_j(\boldsymbol{\xi}_1))-\bar{\psi}_{a^*b^*}^j(h_j(\boldsymbol{\xi}_1))+\bar{\psi}_{ba}^j(h_j(\boldsymbol{\xi}_2))-\bar{\psi}_{b^*a^*}^j(h_j(\boldsymbol{\xi}_2))+\alpha_{ab}\\
&=\sum_{j=1}^{d}(\lambda_{ab}^j-\lambda_{a^*b^*}^j)\phi_j(g_j(\boldsymbol{\xi}_1),g_j(\boldsymbol{\xi}_2))+(\lambda_{ba}^j-\lambda_{b^*a^*}^j)\phi_j(g_j(\boldsymbol{\xi}_2),g_j(\boldsymbol{\xi}_1))\\
&+\log\mu_{ab}^j(g_j(\boldsymbol{\xi}_1))-\log\mu_{a^*b^*}^j(g_j(\boldsymbol{\xi}_1))+\log\mu_{ba}^j(g_j(\boldsymbol{\xi}_2))-\log\mu_{b^*a^*}^j(g_j(\boldsymbol{\xi}_2))+z_{ab}^j
\end{aligned}\tag{11}$$

where $z_{ab}^j=-\log Z_{ab}^j+\log Z_{a^*b^*}^j$, and we substituted Eq. 9 with $(a,b)=(a^*,b^*)$ into Eq. 10. Importantly, the Jacobians do not appear in Eq. 11 because of the subtraction by the pivot condition $(a,b)=(a^*,b^*)$.

By collecting Eq. 11 for all the candidates of $(a,b)\in\mathcal{E}$ into rows;

$$\begin{aligned}
&\bar{\mathbf{W}}\boldsymbol{\psi}(\mathbf{h}(\boldsymbol{\xi}_1),\mathbf{h}(\boldsymbol{\xi}_2))+\bar{\mathbf{W}}'\boldsymbol{\psi}(\mathbf{h}(\boldsymbol{\xi}_2),\mathbf{h}(\boldsymbol{\xi}_1))+\bar{\boldsymbol{\psi}}(\mathbf{h}(\boldsymbol{\xi}_1))+\bar{\boldsymbol{\psi}}'(\mathbf{h}(\boldsymbol{\xi}_2))+\boldsymbol{\alpha}\\
&=\bar{\mathbf{L}}\boldsymbol{\phi}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)+\bar{\mathbf{L}}'\boldsymbol{\phi}(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1)+\bar{\boldsymbol{\mu}}(\boldsymbol{\eta}_1)+\bar{\boldsymbol{\mu}}'(\boldsymbol{\eta}_2)+\mathbf{z},
\end{aligned}\tag{12}$$

where $\bar{\mathbf{W}}$ and $\bar{\mathbf{W}}'\in\mathbb{R}^{|\mathcal{E}|\times d}$ are matrices of $(w_{ab}^j-w_{a^*b^*}^j)$ and $(w_{ba}^j-w_{b^*a^*}^j)$ respectively, with the pairs $(a,b)\in\mathcal{E}$ giving row index and the components j column index, $\boldsymbol{\psi}(\mathbf{h}(\boldsymbol{\xi}_1),\mathbf{h}(\boldsymbol{\xi}_2))=(\psi_1(h_1(\boldsymbol{\xi}_1),h_1(\boldsymbol{\xi}_2)),\ldots,\psi_d(h_d(\boldsymbol{\xi}_1),h_d(\boldsymbol{\xi}_2)))^T$, $\boldsymbol{\phi}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)=(\phi_1(\eta_{11},\eta_{21}),\ldots,\phi_d(\eta_{1d},\eta_{2d}))^T$, $\bar{\boldsymbol{\psi}}(\mathbf{h}(\boldsymbol{\xi}_1))=\left(\sum_{j=1}^d\bar{\psi}_{ab}^j(h_j(\boldsymbol{\xi}_1))-\bar{\psi}_{a^*b^*}^j(h_j(\boldsymbol{\xi}_1))\right)_{(a,b)\in\mathcal{E}}$, $\bar{\boldsymbol{\psi}}'(\mathbf{h}(\boldsymbol{\xi}_2))=\left(\sum_{j=1}^d\bar{\psi}_{ba}^j(h_j(\boldsymbol{\xi}_2))-\bar{\psi}_{b^*a^*}^j(h_j(\boldsymbol{\xi}_1))\right)_{(a,b)\in\mathcal{E}}$, $\bar{\boldsymbol{\mu}}(\boldsymbol{\eta}_1)=\left(\sum_{j=1}^d\log\mu_{ab}^j(\eta_{1j})-\log\mu_{a^*b^*}^j(\eta_{1j})\right)_{(a,b)\in\mathcal{E}}$, $\bar{\boldsymbol{\mu}}'(\boldsymbol{\eta}_2)=\left(\sum_{j=1}^d\log\mu_{ba}^j(\eta_{2j})-\log\mu_{b^*a^*}^j(\eta_{2j})\right)_{(a,b)\in\mathcal{E}}$, $\boldsymbol{\alpha}=(\alpha_{ab})_{(a,b)\in\mathcal{E}}$, and $\mathbf{z}=\left(\sum_{j=1}^d z_{ab}^j\right)_{(a,b)\in\mathcal{E}}$. Let a compound demixing-mixing function $\mathbf{v}(\boldsymbol{\eta}_1)=\mathbf{h}\circ\mathbf{f}(\boldsymbol{\eta}_1)$, we then have

$$\begin{aligned}
&\bar{\mathbf{W}}\boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_1),\mathbf{v}(\boldsymbol{\eta}_2))+\bar{\mathbf{W}}'\boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_2),\mathbf{v}(\boldsymbol{\eta}_1))+\bar{\boldsymbol{\psi}}(\mathbf{v}(\boldsymbol{\eta}_1))+\bar{\boldsymbol{\psi}}'(\mathbf{v}(\boldsymbol{\eta}_2))+\boldsymbol{\alpha}\\
&=\bar{\mathbf{L}}\boldsymbol{\phi}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)+\bar{\mathbf{L}}'\boldsymbol{\phi}(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1)+\bar{\boldsymbol{\mu}}(\boldsymbol{\eta}_1)+\bar{\boldsymbol{\mu}}'(\boldsymbol{\eta}_2)+\mathbf{z}.
\end{aligned}\tag{13}$$

We firstly show that the concatenated matrix $[\bar{\mathbf{W}},\bar{\mathbf{W}}']\in\mathbb{R}^{|\mathcal{E}|\times 2d}$ has full column rank $2d$. We differentiate Eq. 13 with respect to $\eta_{1k},\eta_{2k},1\le k\le d$, and obtain

$$\left[\bar{\mathbf{W}},\bar{\mathbf{W}}'\right]\frac{\partial^2}{\partial\eta_{1k}\partial\eta_{2k}}\begin{bmatrix}\boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_1),\mathbf{v}(\boldsymbol{\eta}_2))\\ \boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_2),\mathbf{v}(\boldsymbol{\eta}_1))\end{bmatrix}=\left[\bar{\mathbf{L}},\bar{\mathbf{L}}'\right]\begin{bmatrix}\boldsymbol{\phi}_k^{12}(\eta_{1k},\eta_{2k})\\ \boldsymbol{\phi}_k^{12}(\eta_{2k},\eta_{1k})\end{bmatrix},\tag{14}$$

where $\boldsymbol{\phi}_k^{12}(\eta_{1k},\eta_{2k})=(0,\ldots,0,\frac{\partial^2}{\partial\eta_{1k}\partial\eta_{2k}}\phi_k(\eta_{1k},\eta_{2k}),0,\ldots,0)^T\in\mathbb{R}^d$ such that the non-zero entry is at index k. Now we concatenate Eq. 14 into columns with changing k, with also flipping the top and the bottom half of the vectors, we get

$$\left[\bar{\mathbf{W}},\bar{\mathbf{W}}'\right]\tilde{\mathbf{Q}}=\left[\bar{\mathbf{L}},\bar{\mathbf{L}}'\right]\begin{bmatrix}\boldsymbol{\Phi}^{12}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) & \boldsymbol{\Phi}^{12}(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1)\\ \boldsymbol{\Phi}^{12}(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1) & \boldsymbol{\Phi}^{12}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\end{bmatrix},\tag{15}$$

where $\mathbf{\Phi}^{12}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \text{diag}\left(\frac{\partial^2}{\partial\eta_{11}\partial\eta_{21}}\phi_1(\eta_{11}, \eta_{21}), \ldots, \frac{\partial^2}{\partial\eta_{1d}\partial\eta_{2d}}\phi_d(\eta_{1d}, \eta_{2d})\right)$, and $\tilde{\mathbf{Q}}$ is a collection of partial derivatives of ψ with respect to the same variables. Now, from Assumption 3, we can find d sets of points $(z_{11}, z_{21}), \ldots, (z_{1d}, z_{2d})$ which make the collection of the partial derivatives of ϕ (the second factor in the right-hand side) full-rank $(2d)$. Since $[\bar{\mathbf{L}}, \bar{\mathbf{L}}']$ is full column rank (Assumption 4), the right-hand side of Eq. 15 has full column rank $(2d)$, and so does the left-hand side. This implies that $[\bar{\mathbf{W}}, \bar{\mathbf{W}}']$ has full column rank $2d$.

For the main result of the Theorem, what we need to prove is that $\mathbf{v}$ is an invertible element-wise function, in the sense that $v_j(\boldsymbol{\eta}_1)$ is a function of only one $\eta_{1\sigma(j)}$ for all j, where $\sigma(j)$ represents the permutation of components. Since $\mathbf{v}$ is invertible because both $\mathbf{h}$ and $\mathbf{f}$ are invertible, the proof can be done by showing that the product of any two distinct partial derivatives of any component is always zero. Along with invertibility, this means that each component depends exactly on one variable. We differentiate Eq. 13 with respect to $\boldsymbol{\eta}_2$, $\eta_{1k}, 1 \leq k \leq d$, and $\eta_{1l}, k < l \leq d$, and get

$$\left[\bar{\mathbf{W}}, \bar{\mathbf{W}}'\right] \frac{\partial^3}{\partial\eta_{1k}\partial\eta_{1l}\partial\boldsymbol{\eta}_2} \begin{bmatrix} \boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_1), \mathbf{v}(\boldsymbol{\eta}_2)) \\ \boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_2), \mathbf{v}(\boldsymbol{\eta}_1)) \end{bmatrix} = \mathbf{0}. \tag{16}$$

From the full column rank of $\left[\bar{\mathbf{W}}, \bar{\mathbf{W}}'\right]$ and the calculation of differentials, we get

$$\frac{\partial^3}{\partial\eta_{1k}\partial\eta_{1l}\partial\boldsymbol{\eta}_2} \begin{bmatrix} \boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_1), \mathbf{v}(\boldsymbol{\eta}_2)) \\ \boldsymbol{\psi}(\mathbf{v}(\boldsymbol{\eta}_2), \mathbf{v}(\boldsymbol{\eta}_1)) \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi}^{112}(\mathbf{v}(\boldsymbol{\eta}_1), \mathbf{v}(\boldsymbol{\eta}_2)) & \mathbf{\Psi}^{12}(\mathbf{v}(\boldsymbol{\eta}_1), \mathbf{v}(\boldsymbol{\eta}_2)) \\ \mathbf{\Psi}^{221}(\mathbf{v}(\boldsymbol{\eta}_2), \mathbf{v}(\boldsymbol{\eta}_1)) & \mathbf{\Psi}^{21}(\mathbf{v}(\boldsymbol{\eta}_2), \mathbf{v}(\boldsymbol{\eta}_1)) \end{bmatrix} \begin{bmatrix} \mathbf{N}^{k\times l}(\boldsymbol{\eta}_1) \\ \mathbf{N}^{kl}(\boldsymbol{\eta}_1) \end{bmatrix} \mathbf{J}_{\mathbf{v}}(\boldsymbol{\eta}_2) = \mathbf{0}, \tag{17}$$

where $\mathbf{\Psi}^{iij}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \text{diag}\left(\frac{\partial^3}{\partial\eta_{i1}\partial\eta_{i1}\partial\eta_{j1}}\psi_1(\eta_{11}, \eta_{21}), \ldots, \frac{\partial^3}{\partial\eta_{id}\partial\eta_{id}\partial\eta_{jd}}\psi_d(\eta_{1d}, \eta_{2d})\right)$, $\mathbf{\Psi}^{ij}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \text{diag}\left(\frac{\partial^2}{\partial\eta_{i1}\partial\eta_{j1}}\psi_1(\eta_{11}, \eta_{21}), \ldots, \frac{\partial^2}{\partial\eta_{id}\partial\eta_{jd}}\psi_d(\eta_{1d}, \eta_{2d})\right)$, $i, j \in \{1, 2\}$, $\mathbf{N}^{k\times l}(\boldsymbol{\eta}_1) = \text{diag}\left(v_1^k(\boldsymbol{\eta}_1)v_1^l(\boldsymbol{\eta}_1), \ldots, v_d^k(\boldsymbol{\eta}_1)v_d^l(\boldsymbol{\eta}_1)\right)$, $\mathbf{N}^{kl}(\boldsymbol{\eta}_1) = \text{diag}\left(v_1^{kl}(\boldsymbol{\eta}_1), \ldots, v_d^{kl}(\boldsymbol{\eta}_1)\right)$, $v_i^k(\boldsymbol{\eta}_1) = \frac{\partial}{\partial\eta_{1k}}v_i(\boldsymbol{\eta}_1)$, $v_i^{kl}(\boldsymbol{\eta}_1) = \frac{\partial^2}{\partial\eta_{1k}\partial\eta_{1l}}v_i(\boldsymbol{\eta}_1)$, and $\mathbf{J}_{\mathbf{v}}$ is the Jacobian of $\mathbf{v}$. From Assumption 7, the matrix of the collection of the derivatives of ψ_j,

$$\begin{bmatrix} \frac{\partial^3}{\partial\eta_{1j}\partial\eta_{1j}\partial\eta_{2j}}\psi_j(\eta_{1j}, \eta_{2j}) & \frac{\partial^2}{\partial\eta_{1j}\partial\eta_{2j}}\psi_j(\eta_{1j}, \eta_{2j}) \\ \frac{\partial^3}{\partial\eta_{1j}\partial\eta_{1j}\partial\eta_{2j}}\psi_j(\eta_{2j}, \eta_{1j}) & \frac{\partial^2}{\partial\eta_{1j}\partial\eta_{2j}}\psi_j(\eta_{2j}, \eta_{1j}) \end{bmatrix} \tag{18}$$

has full rank 2 for almost all of η_{1j} and η_{2j}, for all $1 \leq j \leq d$, which implies that the first factor of Eq. 17 has full-rank $2d$ for almost always. The Jacobian $\mathbf{J}_{\mathbf{v}}$ has full rank d from the invertibility of $\mathbf{v}$. By multiplying the inverses of them to both sides, we get

$$\begin{bmatrix} \mathbf{N}^{k\times l}(\boldsymbol{\eta}_1) \\ \mathbf{N}^{kl}(\boldsymbol{\eta}_1) \end{bmatrix} = \mathbf{0}, \tag{19}$$

In particular, $v_j^k(\boldsymbol{\eta}_1)v_j^l(\boldsymbol{\eta}_1) = 0$ for all $1 \leq j \leq d$, $1 \leq k \leq d$, and $k < l \leq d$. This means that the Jacobian of $\mathbf{v}$ has at most one non-zero entry in each row. Now, by invertibility and continuity of $\mathbf{J}_{\mathbf{v}}$, we deduce that the location of the non-zero entries are fixed and do not change as a function of $\boldsymbol{\eta}_1$. This proves that $v_j(\boldsymbol{\eta}_1)$ is represented by only one $\eta_{1\sigma(j)}$ up to a scalar (component-specific) invertible transformation, and the Theorem is proven.

B PROOF OF THEOREM 2

From the result of Theorem 1 with the required assumptions, the j-th element of $\mathbf{v}(\boldsymbol{\eta}_1)$ represents an invertible transformation of a single component $\eta_{1\sigma(j)}$; we have $v_j(\boldsymbol{\eta}_1) = k_{\sigma(j)}(\eta_{1\sigma(j)})$, where $k_{\sigma(j)}$ is a scalar invertible function, and $\sigma(j)$ is the permutation of components, which is indeterminate according to Theorem 1. Without loss of generality, we assume that the estimated components were sorted properly ($\sigma(j) = j$). Using this result to Eq. 15, with only focusing on the elements related to component j, we have

$$\left[\bar{\mathbf{w}}_j, \bar{\mathbf{w}}_j'\right] \begin{bmatrix} \psi_j^{12}(k_j(\eta_{1j}), k_j(\eta_{2j})) & \psi_j^{12}(k_j(\eta_{2j}), k_j(\eta_{1j})) \\ \psi_j^{12}(k_j(\eta_{2j}), k_j(\eta_{1j})) & \psi_j^{12}(k_j(\eta_{1j}), k_j(\eta_{2j})) \end{bmatrix} = \left[\bar{\boldsymbol{\lambda}}_j, \bar{\boldsymbol{\lambda}}_j'\right] \begin{bmatrix} \phi_j^{12}(\eta_{1j}, \eta_{2j}) & \phi_j^{12}(\eta_{2j}, \eta_{1j}) \\ \phi_j^{12}(\eta_{2j}, \eta_{1j}) & \phi_j^{12}(\eta_{1j}, \eta_{2j}) \end{bmatrix}, \tag{20}$$

where $\bar{\mathbf{w}}_j$, $\bar{\mathbf{w}}_j'$, $\bar{\boldsymbol{\lambda}}_j$ and , $\bar{\boldsymbol{\lambda}}_j'$ are the jth column of $\bar{\mathbf{W}}$, $\bar{\mathbf{W}}'$, $\bar{\mathbf{L}}$, and $\bar{\mathbf{L}}'$, respectively. We then assume that, without loss of generality, the nodes were arranged in a causal order on the j-th component, such that no later node b causes any earlier node $a < b$, which is possible due to the directed acyclic causal structure assumption (Assumptions 4 and 8). We denote the set of

node-pairs $(a,b)_{a<b}$ on the causally-ordered graph of component j as $\mathcal{E}'_j$. This means that only the half of the elements of $\boldsymbol{\lambda}_j$ corresponding to the node-pairs $\mathcal{E}'_j$ can have non-zero values, while the elements of $\boldsymbol{\lambda}'_j$ on the same pairs are zeros because they represent the edges on the opposite directions. We now substitute the point (a^*, b^*) in Assumptions 8 and also (z_{1j}, z_{2j}) in Assumptions 3 into Eq. 20. We then have, by only considering the rows corresponding to the node-pairs $(a,b) \in \mathcal{E}'_j$,

$$\left[\bar{\mathbf{w}}_{j\star}, \bar{\mathbf{w}}'_{j\star}\right] \begin{bmatrix} \psi_j^{12}(k_j(z_{1j}), k_j(z_{2j})) & \psi_j^{12}(k_j(z_{2j}), k_j(z_{1j})) \\ \psi_j^{12}(k_j(z_{2j}), k_j(z_{1j})) & \psi_j^{12}(k_j(z_{1j}), k_j(z_{2j})) \end{bmatrix} = [\boldsymbol{\lambda}_{j\star}, \mathbf{0}] \begin{bmatrix} \phi_j^{12}(z_{1j}, z_{2j}) & \phi_j^{12}(z_{2j}, z_{1j}) \\ \phi_j^{12}(z_{2j}, z_{1j}) & \phi_j^{12}(z_{1j}, z_{2j}) \end{bmatrix}, \tag{21}$$

where $\bar{\mathbf{w}}_{j\star}$, $\bar{\mathbf{w}}'_{j\star}$, and $\boldsymbol{\lambda}_{j\star} \in \mathbb{R}^{|\mathcal{E}'_j|}$ are the elements of $\bar{\mathbf{w}}_j$, $\bar{\mathbf{w}}'_j$, and $\boldsymbol{\lambda}_j \in \mathbb{R}^{|\mathcal{E}_j|}$ on the node-pairs $(a,b) \in \mathcal{E}'_j$, respectively. We used here $\bar{\boldsymbol{\lambda}}_{j\star} = (\lambda^j_{ab} - \lambda^j_{a^*b^*})_{(a,b)\in\mathcal{E}'_j} = \boldsymbol{\lambda}_{j\star}$ because $\lambda^j_{a^*b^*} = 0$ (Assumptions 8), and $\bar{\boldsymbol{\lambda}}'_{j\star} = (\lambda^j_{ba} - \lambda^j_{b^*a^*})_{(a,b)\in\mathcal{E}'_j} = \mathbf{0}$ because of the causal ordering as mentioned above.

We firstly show that the second factor (2×2 matrix) of the left-hand side has full rank. If the matrix does not have full rank, it means that $\psi_j^{12}(k_j(z_{1j}), k_j(z_{2j})) = \psi_j^{12}(k_j(z_{2j}), k_j(z_{1j}))$. However, this implies that $\phi_j^{12}(z_{1j}, z_{2j}) = \phi_j^{12}(z_{2j}, z_{1j})$ from the relations of the elements of both sides of Eq. 21, which is contradictory to Assumption 3, and thus the matrix should have full rank.

Now, since the second factors (2×2 matrices) on both sides have full rank and the first factor of the right-hand side has rank-1, so that of the left-hand side does. This excludes the case where $\bar{\mathbf{w}}_{j\star}$ and $\bar{\mathbf{w}}'_{j\star}$ are linearly independent in Assumption 9, and thus either $\bar{\mathbf{w}}_{j\star}$ or $\bar{\mathbf{w}}'_{j\star}$ is a zero vector. This indicates that $\boldsymbol{\lambda}_{j\star}$ is given by either $\mathbf{w}_{j\star}$ or $\mathbf{w}'_{j\star}$ up to a linear scaling and a bias (note that $\bar{\mathbf{w}}_{j\star} = (w^j_{ab} - w^j_{a^*b^*})_{(a,b)\in\mathcal{E}'_j}$ is biased by $w^j_{a^*b^*}$), and the other one (either $\mathbf{w}_{j\star}$ or $\mathbf{w}'_{j\star}$) is a vector with a constant value.

C ALTERNATIVE IDENTIFIABILITY THEORY

A key assumption here is non-quasi-gaussianity, introduced in Hyvarinen and Morioka (2017), and refined in Hälvä et al. (2021). The relevant definition is

Definition 1. *A two-dimensional random vector (x, y) is called **locally quasi-Gaussian** if there is an open subset $A \in \mathbb{R}^n$, a function $\alpha : \mathbb{R} \to \mathbb{R}$, and a constant $c \in \mathbb{R}$ such that*

$$\frac{\partial^2 \log p_{xy}(x,y)}{\partial x \partial y} = c\,\alpha(x)\alpha(y) \quad \forall (x,y) \in A. \tag{22}$$

This generalizes the definition of Gaussianity since for α equal to a constant, we get the case of the Gaussian bivariate distribution.

In the alternative theorem, we focus only on a single specific node-pair $(a^*, b^*) \in \mathcal{E}$ which satisfies the assumptions shown below, rather than the all pairs $\mathcal{E}$ simultaneously as in the previous theorems. We then train a (nonlinear) feature extractor $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ based on a logistic regression (LR) which discriminates two datasets obtained from the node-pair; $\{(\mathbf{x}^{(n)}_{a^*}, \mathbf{x}^{(n)}_{b^*})\}_n$ and $\{(\mathbf{x}^{(n)}_{a^*}, \mathbf{x}^{*}_{b^*})\}_n$, where $\mathbf{x}^{(n)}_{a^*} \in \mathbb{R}^d$ and $\mathbf{x}^{(n)}_{b^*} \in \mathbb{R}^d$ are the a^*-th and b^*-th rows of $\mathbf{X}^{(n)}$, and $\mathbf{x}^{*}_{b^*} \in \mathbb{R}^d$ is obtained randomly from the distribution of $\mathbf{x}_{b^*}$, in practice by randomly permuting (shuffling) the sample index n of $\mathbf{x}^{(n)}_{b^*}$. We use a regression function of the form

$$r(\mathbf{x}^{(n)}_{a^*}, \mathbf{x}^{(n)}_{b^*}) = \sum_{j=1}^{d} \psi_j(h_j(\mathbf{x}^{(n)}_{a^*}), h_j(\mathbf{x}^{(n)}_{b^*})) + \bar{\psi}_{a^*j}(h_j(\mathbf{x}^{(n)}_{a^*})) + \bar{\psi}_{b^*j}(h_j(\mathbf{x}^{(n)}_{b^*})) + b \tag{23}$$

for the LR, where h_j is the jth element of the feature extractor $\mathbf{h}(\cdot)$, $\psi_j(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$, $\bar{\psi}_{a^*j}(\cdot) : \mathbb{R} \to \mathbb{R}$, and $\bar{\psi}_{b^*j}(\cdot) : \mathbb{R} \to \mathbb{R}$ are some additional nonlinear functions to be learned from the data with universal approximation capacity, and b is a bias parameter.

Using the definition above and adapting the theory of the aforementioned papers, we obtain the following identifiability theorem, where instead of requiring the causal structures to be acyclic and different, we require them to be non-quasi-Gaussian on the specific node-pair:

Theorem 3. *Consider a data model where Assumptions 1, 2, and 6 of Theorem 1 hold in addition to*

10. *There exists and we have a specific node-pair $(a^*, b^*) \in \mathcal{E}$ satisfying the following; for all j, the marginal distributions of two variables (s_{a^*j}, s_{b^*j}) are given as,*

$$p^j_{a^*b^*}(s_{a^*j}, s_{b^*j}) \propto \bar{q}_{a^*j}(s_{a^*j})\bar{q}_{b^*j}(s_{b^*j}) \exp\left(\phi_j(s_{a^*j}, s_{b^*j})\right), \tag{24}$$

where $\phi_j(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$, $\bar{q}_{a^*j}(\cdot)$ and $\bar{q}_{b^*j}(\cdot) : \mathbb{R} \to \mathbb{R}$ are some nonlinear functions, and are not locally quasi-Gaussian (Definition 1).

11. *(LR) We train LR with the regression function Eq. 23 with universal approximation capability, to discriminate two datasets $\{(\mathbf{x}^{(n)}_{a^*}, \mathbf{x}^{(n)}_{b^*})\}_n$ and $\{(\mathbf{x}^{(n)}_{a^*}, \mathbf{x}^{*}_{b^*})\}_n$.*

Then, the latent components are identifiable, in the sense that the feature extractor $\mathbf{h}(\mathbf{x}^{(n)}_a) = (h_1(\mathbf{x}^{(n)}_a), \dots, h_d(\mathbf{x}^{(n)}_a))^T$ gives the latent components $\mathbf{s}^{(n)}_a = (s_{aj})^{(n)}_j$, up to permutation and scalar (component-wise) invertible transformations for all $a \in \mathcal{V}$ and n.

Proof. From the generative model with Assumption 10, the pairwise marginal distribution of a pair of observations $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = (\mathbf{f}(\boldsymbol{\eta}_1), \mathbf{f}(\boldsymbol{\eta}_2)) \in \mathbb{R}^{d\times 2}$ on the node-pair $(a^*, b^*) \in \mathcal{E}$ is given by, using the probability transformation formula,

$$\begin{aligned}
&\log p_{a^*b^*}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \\
&= \log p_{a^*b^*}(\mathbf{g}(\boldsymbol{\xi}_1), \mathbf{g}(\boldsymbol{\xi}_2)) + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)| \\
&= \sum_{j=1}^{d} \log p^j_{a^*b^*}(g_j(\boldsymbol{\xi}_1), g_j(\boldsymbol{\xi}_2)) + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)| \\
&= \sum_{j=1}^{d} \phi_j(g_j(\boldsymbol{\xi}_1), g_j(\boldsymbol{\xi}_2)) + \log \bar{q}_{a^*j}(g_j(\boldsymbol{\xi}_1)) + \log \bar{q}_{b^*j}(g_j(\boldsymbol{\xi}_2)) \\
&- \log Z^j_{a^*b^*} + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)|,
\end{aligned} \tag{25}$$

where $\mathbf{J}$ denotes the Jacobian, $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ is the (true) inverse function of the observational mixing $\mathbf{f}$ (Eq. 3), and thus $\eta_{1j} = g_j(\boldsymbol{\xi}_1)$ gives a single component for each $j = 1, \dots, d$ by definition. The second equation comes from the mutual independence of the latent components, and third equation comes from the pairwise marginal distribution of the latent components given by Assumption 10. $Z^j_{a^*b^*}$ denotes the partition functions (normalization constants) of the marginal distributions.

On the other hand, the joint distribution of the same observations $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ on the shuffled data class is given by the form,

$$\begin{aligned}
&\log p^*_{a^*b^*}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \\
&= \log p^*_{a^*b^*}(\mathbf{g}(\boldsymbol{\xi}_1), \mathbf{g}(\boldsymbol{\xi}_2)) + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)| \\
&= \sum_{j=1}^{d} \log p^{j*}_{a^*b^*}(g_j(\boldsymbol{\xi}_1), g_j(\boldsymbol{\xi}_2)) + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)| \\
&= \sum_{j=1}^{d} \log p^*_{a^*j}(g_j(\boldsymbol{\xi}_1)) + \log p^*_{b^*j}(g_j(\boldsymbol{\xi}_2)) + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_1)| + \log|\det \mathbf{Jg}(\boldsymbol{\xi}_2)|,
\end{aligned} \tag{26}$$

where $p^*_{a^*j}$ and $p^*_{b^*j}$ are the marginal distributions of the latent variables on nodes a^* and b^* on component j, respectively. This factorization comes from the fact that the two variables are independent here because of the shuffling across samples.

According to well-known theory of LR (Gutmann and Hyvärinen, 2012), when training LR which discriminates the data with the original combination and the shuffled data, with the regression function Eq. 23, we will asymptotically have

$$r(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \log p_{a^*b^*}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) - \log p^*_{a^*b^*}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2), \tag{27}$$

i.e. the regression function will asymptotically give the difference of the log-probabilities in the two classes. By substituting

the regression function model Eq. 23 and the distributions Eq. 25 and Eq. 26 into the equation above, we have

$$\sum_{j=1}^{d} \psi_j(h_j(\boldsymbol{\xi}_1), h_j(\boldsymbol{\xi}_2)) + \bar{\psi}_{a^* j}(h_j(\boldsymbol{\xi}_1)) + \bar{\psi}_{b^* j}(h_j(\boldsymbol{\xi}_2)) + b$$
$$= \sum_{j=1}^{d} \phi_j(g_j(\boldsymbol{\xi}_1), g_j(\boldsymbol{\xi}_2)) + \log \bar{q}_{a^* j}(g_j(\boldsymbol{\xi}_1)) - \log p^*_{a^* j}(g_j(\boldsymbol{\xi}_1)) + \log \bar{q}_{b^* j}(g_j(\boldsymbol{\xi}_2)) - \log p^*_{b^* j}(g_j(\boldsymbol{\xi}_2)) - \log Z^j_{a^* b^*}. \tag{28}$$

Consider Eq. 28, and compare it with Eq. (26) of Hyvarinen and Morioka (2017). The functions g and h denote the same things in the two proofs. The two nodes a^*, b^* in Eq. 28 formally correspond to the two time points $t, t-1$ in Hyvarinen and Morioka (2017). Thus, we have the equivalent set of terms (B_j, Q_j) of Hyvarinen and Morioka (2017) and (ψ_j, ϕ_j) here. The other terms are immaterial since they depend only on either $\boldsymbol{\xi}_1$ or $\boldsymbol{\xi}_2$, and thus they will disappear later in the proof anyway. Now, we can proceed with the proof of Hyvarinen and Morioka (2017), taking into account the small correction pointed out in Hälvä et al. (2021) in their last paragraph of Section 4, and the identifiability is thus proven.

□

D RELATED WORKS

Causal Discovery by Asymmetricity Bayesian networks (BNs) (Pearl, 2000) represent a causal graph among variables by a factorization of their joint distribution into some conditional distributions representing the conditional independence of the variables. 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, such as the Poisson distribution (Park and Raskutti, 2015; Park and Park, 2019b), the generalized hypergeometric distribution (Park and Park, 2019a), and the zero-inflated Poisson model (Choi et al., 2020), enable identifiability of the causal structure. A very closely related framework is given by structural equation models (SEMs) (Bollen, 1989). Since SEMs are not generally identifiable (Bollen, 1989; Geiger and Heckerman, 1994; Pearl, 2000), similarly to BNs, further assumptions have been proposed to guarantee the identifiability: linear acyclic models with non-Gaussian noise (influence) (Shimizu et al., 2006, 2011), additive noise models excluding linear functions (Hoyer et al., 2008a; Hyvärinen and Smith, 2013; Peters et al., 2014), post-nonlinear models (Zhang and 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 and Shimizu, 2020; Shimizu and Bollen, 2014), for example. 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 and Fukumizu, 2020), though limited to bivariate settings. Our CCL is similar to those frameworks in that it represents the causal structure by a probabilistic graphical model with some assumptions on the graph structures and the asymmetricity of the causal relations. However, we consider multimodal (multidimensional) node observations, with data-driven disentanglements of unknown observational mixing across the modalities, which are quite new.

As a first approach to dealing with multimodal node observations, one can consider simply stacking multiple DAGs with the same set of nodes across modalities, and then representing the causal structure as a three-way adjacency tensor (modality $\times$ node $\times$ node), as in Chen et al. (2021); Wang et al. (2020). Such an approach assumes the modalities are mutually independent (there are no connections between the different modalities), thus modelling each adjacency matrix independently. Some recent studies showed that some assumptions on the similarity (or *consistency*) across modalities, in particular concerning the causal order, enable identifiability and consistency of the joint estimator together with better estimation performance compared to individual estimators (Chen et al., 2021; Wang et al., 2020). In contrast, CCL utilizes the *differences* of the causal structures across modalities for the estimation, rather than their similarity (such as the same causal ordering) which might be too restrictive in some applications. In addition, CCL has a fundamental advantage compared to such previous works, as performing disentanglements of observational mixings jointly with the causal discovery on the disentangled latent space.

Disentangled Representation Learning In many applications, there is an unknown process generating the observed data as a mixing of some underlying latent components. Thus, disentangling (demixing) the observations into the latent components in an unsupervised data-driven manner, would have great utility for the generalizability, robustness, interpretability, and explainability of a model. Recently, new frameworks have been proposed based on NICA (Hälvä and Hyvärinen, 2020;

Hyvärinen and Morioka, 2016; Hyvarinen and Morioka, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020; Morioka et al., 2021), which have shown that some assumptions on the latent components, such as mutual independence and temporal nonstationarity, enable their identifiability. Our framework CCL can be also seen as a novel NICA framework, but it considers two-dimensional variables (multimodal observations from multiple nodes) and also gives hidden causal structures, which are both new compared to NICA.

Causal representation learning, which is a combination of representation learning and causal discovery, has been receiving increasing attention recently (Schölkopf et al., 2021). Methods were proposed to extract a set of latent variables having causal relations from observational variables, for example by assuming linear causal models (Yang et al., 2021) or discrete latent variables (Kivva et al., 2021). On the other hand, CCL assumes a two dimensional observation (node $\times$ modality) for each sample, then finds a latent matrix (node $\times$ component) by disentangling the observational modalities into latent components (2nd axis), and jointly estimates component-wise causal structures between node variables (1st axis) with nonlinear causal relations. Other studies (Lippe et al., 2022; Yao et al., 2022) have used temporal causal relations and are thus only applicable to time-series data.

Considering that the samples $(x_{ai})^{(n)}$ has three indices, node a, modality i, and sample n, this model can be seen as a type of tensorial data analysis. In that domain, related classical methods include frameworks such as PARAFAC/CANDECOMP or the Tucker decomposition (Harshman and Lundy, 1994; Miwakeichi et al., 2004). Some recent work in deep learning has started extending such (multi-)linear methods to the nonlinear case, e.g. (Fan, 2022; Hosoya, 2021; Pan et al., 2020; Zhe et al., 2016), but they generally do not model causal structures. Identifiability of PARAFAC typically requires that the modalities are sufficiently different from each other, which is something we use in this study as well.

Self-supervised Learning and Contrastive Learning Self-supervised leaning has been receiving attention as a new way to learn hidden representation of data in a data-driven manner. In contrast to ordinary supervised learning, which explicitly uses the labels assigned to each data point in advance (such as "dog" or "cat" in the image-classification), self-supervised leaning instead uses data structures *inherent* in the data as the target labels. The structure to be used depends on the data type, and also on what kind of representation we want to extract from the data. Examples include spatial neighboring structures of images (Doersch et al., 2015), spatial transformation invariance of images (Chen et al., 2020), sequential structure of natural languages (Devlin et al., 2018), temporal proximity of time-series data (Banville et al., 2021; Hyvärinen and Morioka, 2016; Hyvarinen and Morioka, 2017), and so on. Some studies showed that it can match or even exceeds the performances achieved by supervised learning frameworks (Chen et al., 2020; Grill et al., 2020).

Self-supervised learning is closely related to the idea of contrastive-learning which takes a contrast of data distributions on some different conditions for model estimation. Contrastive-learning has theoretical justification as a density-ratio estimation (Gutmann and Hyvärinen, 2012; Sugiyama et al., 2012). NICA is also based on this idea, and gives identifiability of the latent components by assuming some temporal (or spatial) structures of the data, and then taking a contrast on some different observational conditions (Hyvärinen and Morioka, 2016; Hyvarinen and Morioka, 2017; Hyvarinen et al., 2019). Our new framework CCL learns representation of data (latent components and causal structures) with a novel contrastive-learning framework, which uses a *node-pair index* as a target label for every node-paired observations. The name *connectivity*-contrastive learning comes from such idea of taking contrast of data distributions across all node-pairs for its learning.

E IMPLEMENTATION DETAIL FOR SIMULATION 1

We give here more detail on the data generation, training, and evaluation in Simulation 1 (Section 5.1). Also see the Supplementary Code for the implementation.

Data Generation We generated artificial data based on the generative model described in Section 2. The number of nodes (p) was fixed to 30, the number of modalities and components (d) was 10, and the number of data points n was fixed to $2^{12} = 4,096$. First, the latent node variables $(s_{aj})_{a\in\mathcal{V}}^{(n)}$ (j-th column of $\mathbf{S}^{(n)}$) were generated probabilistically for each sample n and component j, based on the pairwise BN (Eq. 2). The weighted adjacency matrices $\boldsymbol{\lambda}_j$ were designed so that they are DAGs and have distinctive structures across components j (see some examples of the generated causal structures in Supplementary Fig. 6). More specifically, for each component j, each node b was given three parents (unless there is no enough number of candidates) randomly selected from nodes $a < b$ within some ranges of indices from the target node. To avoid the graphs to have small underlying undirected cycles, we selected the parents so that they are not the parents of the other parents. Although these graphs can have cycles in their underlying undirected graphs and thus do not really satisfy

Assumption 4, the results show that CCL can still work on this more general type of DAGs. The latent variables were then sampled based on the following distribution for each component j, node b, and sample n:

$$s_{bj}^{(n)} \sim \exp\left(\sum_{a \in \mathrm{pa}_j(b)} -\frac{\lambda_{ab}^j}{|\mathrm{pa}_j(b)|} \left(s_{bj}^{(n)} + |\mathrm{pa}_j(b)| \mathrm{Relu}(s_{aj}^{(n)}) \right)^2 \right), \tag{29}$$

where $\mathrm{pa}_j(b)$ is the set of parents of node b on component j, deduced from the adjacency matrix $\boldsymbol{\lambda}_j$, $|\mathrm{pa}_j(b)|$ the number of parents, and $\mathrm{Relu}(x) = \max(0, x)$ is a rectified linear unit. This model indicates that the activity $s_{bj}^{(n)}$ is randomly generated through a Gaussian distribution with a standard deviation modulated by the inverse of root of summation of $\lambda_{\mathrm{pa}(b)b}^j$, and its average is negatively biased by positive-, but not by negative-, activities of its parents (nonlinear inhibitory connection). The non-zero values of $\boldsymbol{\lambda}_j$ were randomly generated so that the standard deviation parameters $(1/\sqrt{2\boldsymbol{\lambda}_j})$ distributed uniformly on $[0.7, 1]$. The inverse-scaling of λ_{ab}^j by $|\mathrm{pa}_j(b)|$ was used so that the (conditional) standard deviations of nodes were approximately the same regardless of the number of parents. This sampling distribution indicates that the function ϕ_j in Eq. 2 is given by, with a simple calculation,

$$\phi_j(s_{aj}, s_{bj}) = s_{bj} \mathrm{Relu}(s_{aj}) \tag{30}$$

which is designed to satisfy Assumption 3 due to its asymmetricity and nonlinearity. This model can be also represented by a (probabilistic) nonlinear SEM;

$$(s_{aj})_{a \in \mathcal{V}} = \mathrm{vec}^{-1}(\boldsymbol{\lambda}_j) \mathrm{Relu}((s_{aj})_{a \in \mathcal{V}}) + \boldsymbol{\epsilon}(\boldsymbol{\lambda}_j), \tag{31}$$

where $\mathrm{vec}^{-1}(\cdot)$ is the inverse of the vectorization of the vectorized adjacency matrix $\boldsymbol{\lambda}_j$, with row-wise weighting based on the number of non-zero elements, $\mathrm{Relu}(\cdot)$ is an element-wise rectified linear unit, and $\boldsymbol{\epsilon}$ is a p-dimensional orthogonal Gaussian noise with zero-mean and element-specific variances depending on the graph structure. Note that such translation of pairwise BN into SEM is not always possible.

For the observation model $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, we used a multilayer perceptron (MLP) with L layers (excluding the input layer) with random parameters, which takes a d-dimensional latent component $\mathbf{s}_a^{(n)} = (s_{aj})_j^{(n)}$ (a-th row of $\mathbf{S}^{(n)}$) and then outputs a d-dimensional observation $\mathbf{x}_a^{(n)} = (x_{ai})_i^{(n)}$ (a-th row of $\mathbf{X}^{(n)}$) for each node $a \in \mathcal{V}$ and sample n. To guarantee the invertibility, we fixed the number of units of each layer to d, and used leaky ReLU units for the nonlinearity except for the last layer which has no-nonlinearity. $L = 0$ means that the latent components are directly given as the observations, without any observational mixings across modalities (i.e., $\mathbf{X}^{(n)} = \mathbf{S}^{(n)}$).

Training (CCL) We train the nonlinear regression function in Eq. 4 with the observed data by CCL. We adopted MLP for $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ (*h-MLP*), whose outputs are supposed to represent the latent components after the training (Theorem 1). 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 $2d$ except for the output (d), 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). In directly-observed cases ($L = 0$), we still used one-layer h-MLP (linear transformation) for CCL, whose parameters need to be learned from the data, though it is redundant and not necessary in practice. The function $\psi_j : \mathbb{R}^2 \to \mathbb{R}$ was modeled by

$$\psi_j(x, y) = \max(c_x^{j(1)}(x - d_x^j), c_x^{j(2)}(x - d_x^j)) \times \max(c_y^{j(1)}(y - d_y^j), c_y^{j(2)}(y - d_y^j)), \tag{32}$$

where $c_x^{j(1)}$, $c_x^{j(2)}$, d_x^j, $c_y^{j(1)}$, $c_y^{j(2)}$, and d_y^j are trainable scalar parameters. This function is based on the idea of maxout unit, and has enough degree of freedom to represent ϕ_j (Eq. 30). For the functions $\bar{\psi}_{ab}^j$, we used a functional form $\bar{\psi}_{ab}^j(h_j(\cdot)) = \max(c_{ab}^{j(1)} h_j(\cdot) + d_{ab}^{j(1)}, c_{ab}^{j(2)} h_j(\cdot) + d_{ab}^{j(2)})^2$, where $c_{ab}^{j(1)}$, $c_{ab}^{j(2)}$, $d_{ab}^{j(1)}$, and $d_{ab}^{j(2)}$ are trainable parameters.

Those nonlinear functions were then trained by back-propagation with a momentum term so as to optimize the loss function of MLR (Eq. 4; Fig. 1), whose feature extractor $\mathbf{h}(\cdot)$ is supposed to represent the latent components (Theorem 1), and the weight parameters (w_{ab}^j) are supposed to represent the causal structure after the training (Theorem 2; see Supplementary Fig. 6 for some examples). The initial parameters were randomly drawn from a uniform distribution. The training of a three-layer model by CCL took about 1.5 hours (Intel Xeon 3.6 GHz 16 core CPUs, 384 GB Memory, NVIDIA Tesla A100 GPU).

The pseudo-code based on a basic stochastic gradient descent (SGD) is given in Supplementary Algorithm. 1, which can be easily implemented based on ordinary neural network training.

Training (CCLalt) We train the nonlinear regression function in Eq. 23 with the observed data by CCLalt. We used an MLP for the feature extractor $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$, similarly to CCL. We modeled ψ_j and $\bar{\psi}_{a^* j}$ by the same models as ψ_j and $\bar{\psi}^j_{ab}$ in CCL, respectively. We gave the node-pair $(1, 2) \in \mathcal{E}$ as (a^*, b^*) in Assumption 10 because we know that that node-pair has causal relations on all of the components, from the data generation process used here (see Supplementary Fig. 6).

Those nonlinear functions were then trained by back-propagation with a momentum term so as to predict whether each input came from the original paired observations $\{(\mathbf{x}^{(n)}_{a^*}, \mathbf{x}^{(n)}_{b^*})\}_n$ or the shuffled ones $\{(\mathbf{x}^{(n)}_{a^*}, \mathbf{x}^{*}_{b^*})\}_n$ by LR. The initial parameters were randomly drawn from a uniform distribution. The training of a three-layer model by CCLalt took about 1 hour (Intel Xeon 3.6 GHz 16 core CPUs, 384 GB Memory, NVIDIA Tesla A100 GPU).

Evaluation We evaluated the estimation performances of the latent components and the causal structures by comparing the estimations with the true values.

The estimated latent components $\mathbf{h}(\cdot)$ were evaluated by their Pearson correlation to the true values. Since the order of the components j is undetermined (Theorem 1), we performed an optimal assignment of components between the estimations and the true ones by the Munkres assignment algorithm (Munkres, 1957), maximizing the mean absolute-correlation coefficients.

For evaluations of the estimated causal structures (w^j_{ab}), 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 $(a, b) \in \mathcal{E}$ on component j by comparing the absolute values of w^j_{ab} and w^j_{ba}; direction is $a \to b$ on j if $|w^j_{ab}| > |w^j_{ba}|$, and vice versa. We then removed edges whose absolute weights were less than a specific ratio (35% for Simulation 1) of the maximum values of $|w^j_{ab}|$ across edges for each component. If both $|w^j_{ab}|$ and $|w^j_{ba}|$ are under the threshold, a and b are considered to have no direct causal relation. The obtained adjacency matrices were then compared with the (binarized) true causal structure (λ^j_{ab}), and evaluated by precision, recall, and 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 CCL which do not explicitly impose sparseness or DAG structure constraints for the estimation. The threshold 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 CCL 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. 7). Since CCL has indeterminacy of the estimation of the causal structure with its component-wise matrix-transpose (Theorem 2), we optimally chose either $(w^j_{ab})_{(a,b)\in\mathcal{E}}$ or its matrix-transpose $(w^j_{ba})_{(a,b)\in\mathcal{E}}$ as its final guess of $(\lambda^j_{ab})_{(a,b)\in\mathcal{E}}$ for each component j. Again, we performed an optimal assignment of components j between the estimations and the true ones by the Munkres assignment algorithm (Munkres, 1957) so as to maximize the F1-score because of the indeterminacy of the order of the components. The learning was performed for 10 runs with changing the parameters of the observation model and the causal structures, for both the directly-observed case ($L = 0$) and the nonlinear-observational-mixture case ($L = 3$).

For comparison, we also applied PC (Spirtes and Glymour, 1991), FGS (Ramsey et al., 2017), GFCI (Ogarrio et al., 2016), CAM (Bühlmann et al., 2014), DirectLiNGAM (Shimizu et al., 2011), NOTEARS-linear (Zheng et al., 2018), NOTEARS-MLP (Zheng et al., 2020), and MultiDAG (Chen et al., 2021) to the same data. See Supplementary Material G for the details of the baselines. Briefly, MultiDAG assumes multimodal causal structures explicitly, while the others focus on single modalities. FGS, GFCI, DirectLiNGAM, NOTEARS-linear, and MultiDAG are specialized at linear models, CAM and NOTEARS-MLP are for nonlinear models, and GFCI assumes presence of latent confounder. We used publicly available implementations of them. Basically, all baselines estimate all of the d-component graphs, and are evaluated based on the average of the graph-wise evaluations, in the same way as CCL. For the directly-observed case ($L = 0$), we applied them separately to each modality i (or jointly in MultiDAG) of the observations. For the nonlinear-observation case ($L = 3$), we applied them after linear-ICA (Hyvärinen, 1999) across modalities because they cannot perform representation learning by themselves. Note that we cannot apply NICA such as Hyvärinen and Morioka (2016); Hyvarinen and Morioka (2017) for this representation learning because there is no existing consistent NICA frameworks for the generative model of this study. For the frameworks which output weighted adjacency matrices (DirectLiNGAM, NOTEARS-linear, NOTEARS-MLP, and MultiDAG), we used the same evaluation criteria to that of CCL (causal direction determination, thresholding, and component assignments). The threshold was determined separately for each method, so as to maximize the F1-score (see Supplementary Fig. 7 for the effect of the varying threshold). The other frameworks which output a binarized adjacency matrix, we directly compared them with the binarized true adjacency matrices. Since some of them output graphs possibly with some bi-directional (or undetermined) edges, we gave the *true* directions to them favorably. For a fair comparison, we optimally chose either the originally estimated adjacency matrix or its transpose as the final guess for each modality.

Although we also applied RCD (Maeda and Shimizu, 2020), which assumes a linear SEM model with latent confounders, we do not report the results here because it eventually estimated that many of the edges were affected by confounders, and did not give proper estimations of the causal directions.

F ADDITIONAL INFORMATION FOR GENE REGULATORY NETWORK RECOVERY

We used synthetic single-cell gene expression data generated by SERGIO (Dibaeinia and 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 experimental data (Dibaeinia and Sinha, 2020). We used the same parameters for the differential equations as in (Dibaeinia and Sinha, 2020), but changed the hill coefficient from 2 to 6 to make the causal relations more nonlinear. The first (smallest-indexed) 15 genes were assigned as master regulators (MRs), having no parents, and controlled by basal production rates, randomly selected from $[0.25, 0.75]$. The maximum contributions (weights of edges) from parental genes to target genes were set to 0.25 for all edges. For the observation model $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, 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 function $\psi_j : \mathbb{R}^2 \to \mathbb{R}$ was modeled by

$$\psi_j(x, y) = y \times \sum_{k=1}^{K} a_{jk} \tanh(b_{jk}x + c_{jk}), \tag{33}$$

where $K = 5$ is a model order, a_{jk}, b_{jk}, and c_{jk} are trainable scalar parameters.

We omitted some runs of PC for $L = 3$ because they did not converge within reasonable calculation time.

G DETAILS OF BASELINE METHODS

PC PC algorithm (Spirtes and 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).

FGS Fast greedy search (FGS) (Ramsey et al., 2017) is a score-based method and assumes DAG, which is an extension of greedy equivalent search (GES) algorithm (Chickering, 2003). 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). FGS improved performances and speeds of GES with some techniques, such as parallelization.

GFCI GFCI (Ogarrio et al., 2016) assumes DAG and presences of latent confounders, which is an extension of constraint-based method FCI (Spirtes et al., 1995), by using a score-based greedy algorithm FGS (Ramsey et al., 2017). GFCI produces a partial ancestral graph (PAG), which is a representation of a set of causal networks that may include hidden confounders, and consists of four types of edges; 1) directed, 2) no causal relation but there is an unmeasured variable that is a cause of both variables, 3) directed or there is an unmeasured variable that is a cause of both variables, or both, 4) directed but the direction is undtermined or there is an unmeasured variable that is a cause of both variables, or both. For comparisons to the other methods, we treated 1 and 3 as directed edges and 4 as bi-directed edges. We set the upper bound on the maximum degree of graphs to 10 to make the computation feasible.

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. 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.

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

Algorithm 1 Pseudo-code of connectivity-contrastive learning (CCL) based on stochastic gradient descent (SGD)

Input: A dataset of observational matrices $\{\mathbf{X}^{(n)}\}_n$, hyper-parameters for the optimization by stochastic gradient descend (SGD).

1: Initialization: Initialize the parameters of the softmax function in Eq. 4 with random values.
2: **repeat**
3: Randomly select some node-pairs $(a, b) \in \mathcal{E}$. Note that those pairs do not necessarily need to have causal relations behind, as far as the total graphs satisfy Assumption 4.
4: Randomly pick some samples (mini-batch) of paired observations $(\mathbf{x}_a^{(n)}, \mathbf{x}_b^{(n)}) \in \mathbb{R}^{d\times 2}$, which are the a-th and b-th rows of $\mathbf{X}^{(n)}$, for each selected node-pair.
5: Update the parameters of the softmax function in Eq. 4 so as to optimize the objective function with back-propagation (SGD). This equivalents to make the model better predict the true node-pair label, assigned at step 3, for each input.
6: **until** the objective function Eq. 4 converges.
7: **return** The trained nonlinear feature extractor $\mathbf{h}(\cdot)$ and the weight parameters (w_{ab}^j) in Eq. 4.

and independence testing for each pair of nodes, 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-linear NOTEARS-linear (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-linear 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-linear (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.

MultiDAG MultiDAG (Chen et al., 2021) jointly estimates multiple causal structures corresponding to multiple task conditions. MultiDAG assumes that the causal orders are consistent across tasks, and then estimates multi-task (linear) adjacency matrices by jointly minimizing reconstruction error of structural equation models, with using DAG-ness constraints proposed by (Zheng et al., 2018). We used the default parameters.

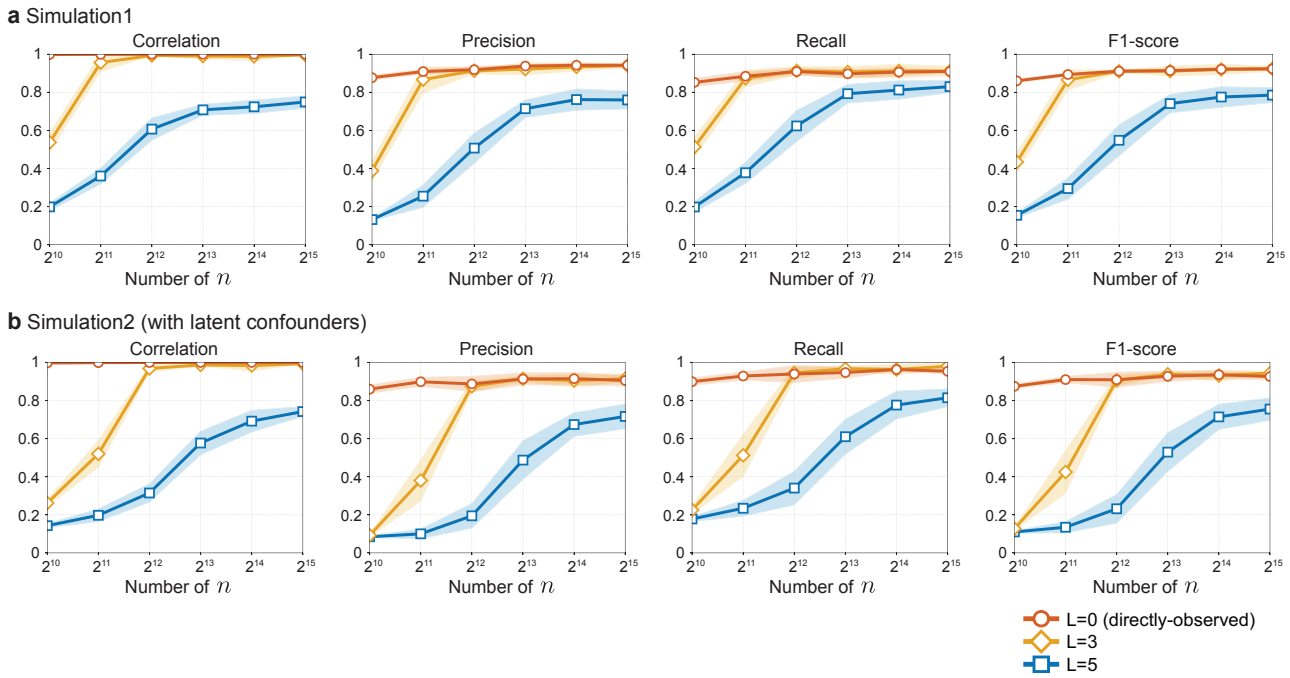

Figure 3: Estimation performances of the latent components (Pearson correlation) and the causal structures (precision, recall, and F1-score) by the proposed framework CCL, with different settings of the complexity of the observation models (the number of MLP-layers L of the observation function $\mathbf{f}$) and the number of n, on Simulation 1 (**a**) and Simulation 2 (**b**; with latent confounders). $L = 0$ indicates that the latent components were directly obtained as the observations, while the observations were (unknown) nonlinear mixtures of the latent components in $L > 0$. The values are the averages of 10 runs for each setting, and the shaded regions show the standard deviations.

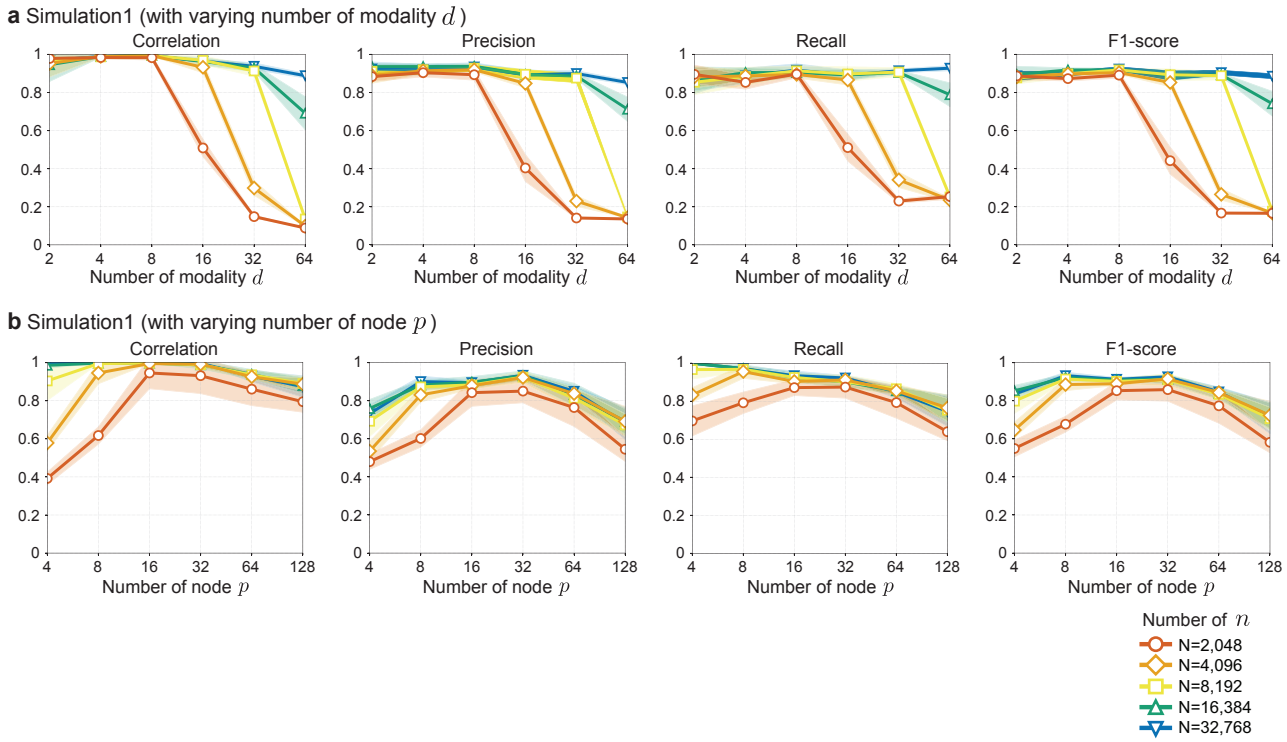

Figure 4: (**a**) Estimation performances of the latent components (Pearson correlation) and the causal structures (precision, recall, and F1-score) by CCL, with different settings of the number of modalities and components d, and the number of n, with fixing the number of nodes $p = 30$ and $L = 3$. The values are the averages of 10 runs for each setting, and the shaded regions show the standard deviations. (**b**) Same as **a**, but with changing the number of nodes p, with fixing $d = 10$. The large number of nodes causes difficulty of estimations, as expected. In addition, small number of nodes also leads to worse estimation performances presumably because the modulation of the causal structures across modalities are not enough for the estimations by CCL (Assumption 4) in our data generation method (Supplementary Material E).

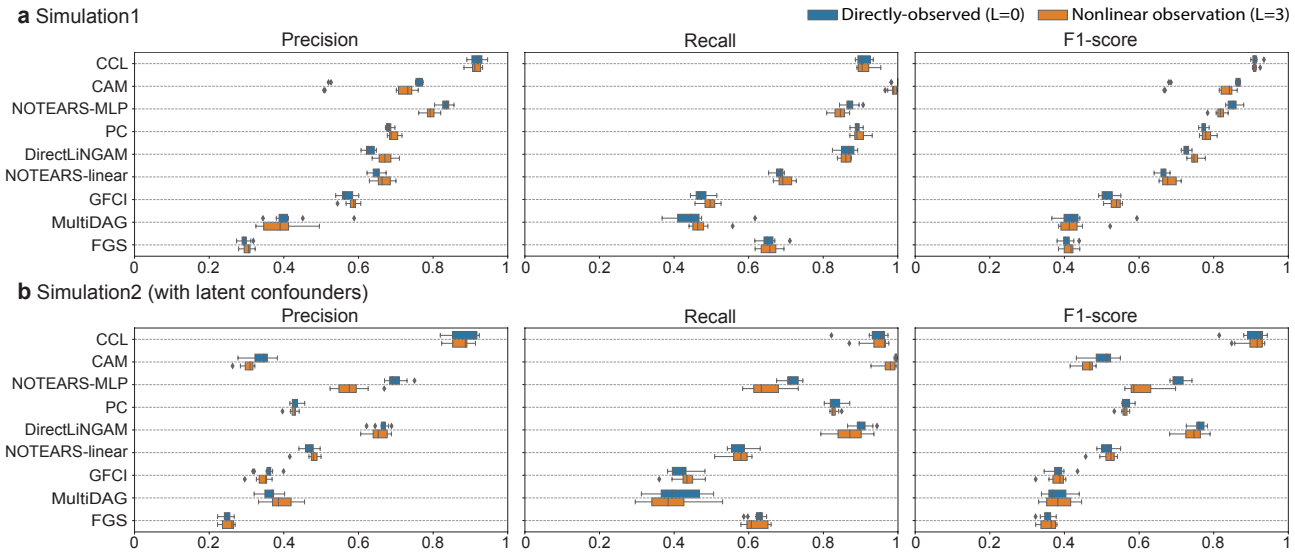

Figure 5: Causal discovery performances by the baseline frameworks applied to the latent components estimated by CCL (the output of the feature extractor $\mathbf{h}$). In the nonlinear observation cases ($L = 3$), the performances of the baselines are better than those directly applied on the observations (after linear-ICA, Fig. 2) due to the nonlinear demixing by CCL, but still lower than that of CCL. This result indicates the importance of performing both representation learning and causal discovery simultaneously, as in CCL.

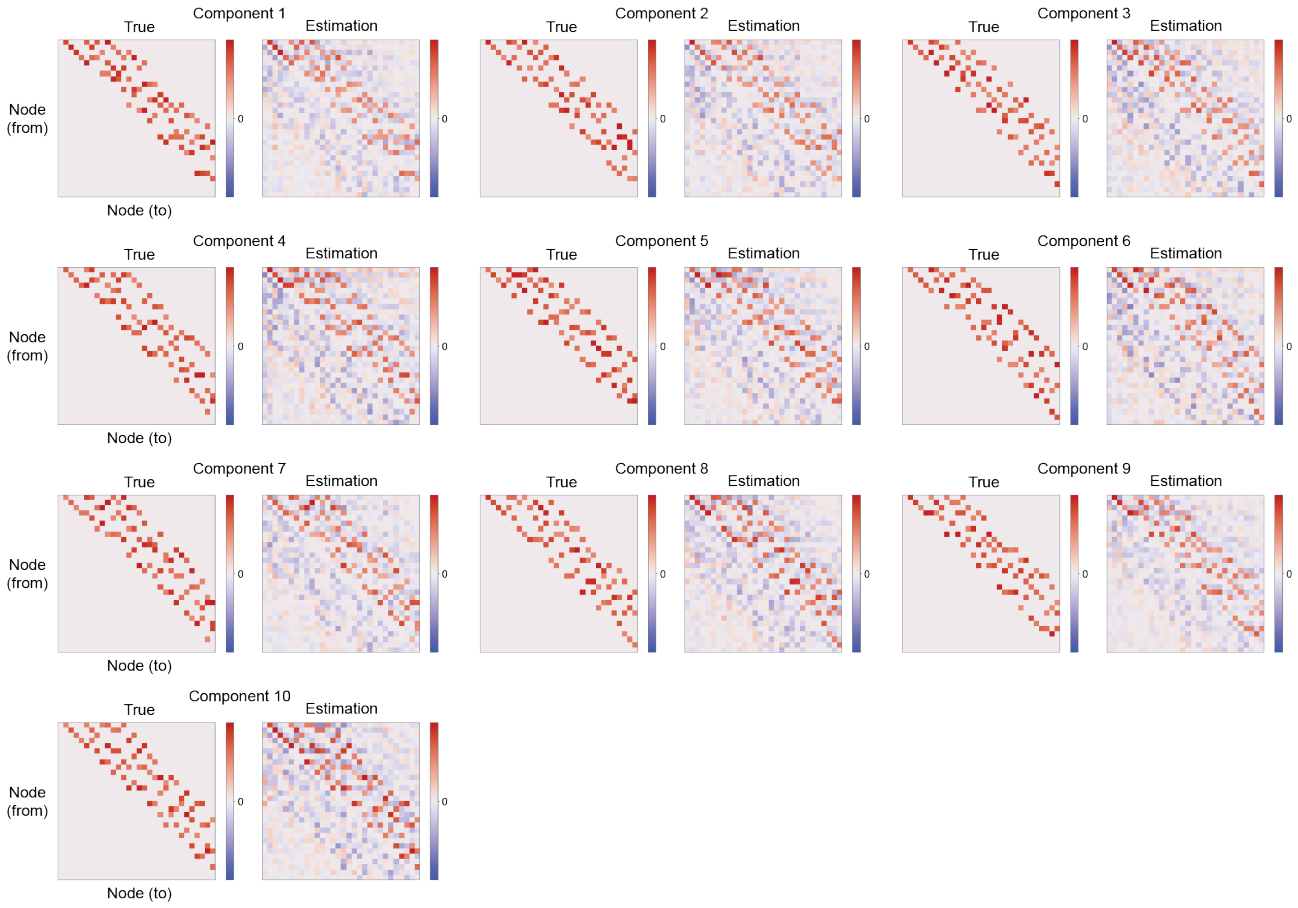

Figure 6: Example of the true causal structures and the estimations by CCL in Simulation 1 (nonlinear-mixture case $L = 3$). We showed the weight parameters of MLR (raw values, before threshold) as the estimated causal structures. Note that some of the estimations were matrix-transposed from the original values so as to match the true one, due to the indeterminacy of the causal discovery by CCL (see Theorem 2).

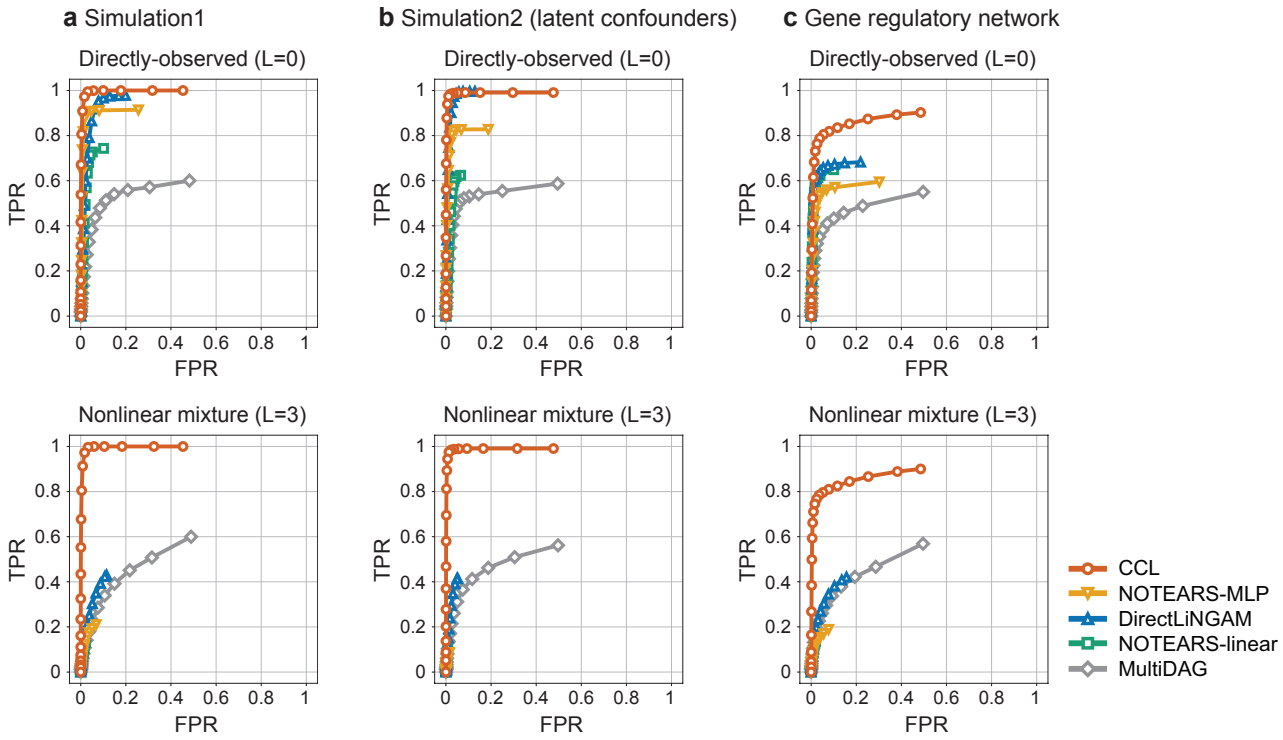

Figure 7: Illustration of the effect of the threshold for CCL, DirectLiNGAM, NOTEARS-linear, NOTEARS-MLP, and MultiDAG. For each panel, ROC curve shows false positive rate (FPR) and true positive rate (TPR) with varying level of threshold, from 0% to 100% with interval of 5%, for each method. The values are the averages of 10 runs for each threshold. This result shows that CCL was not so sensitive to the selection of the threshold values.

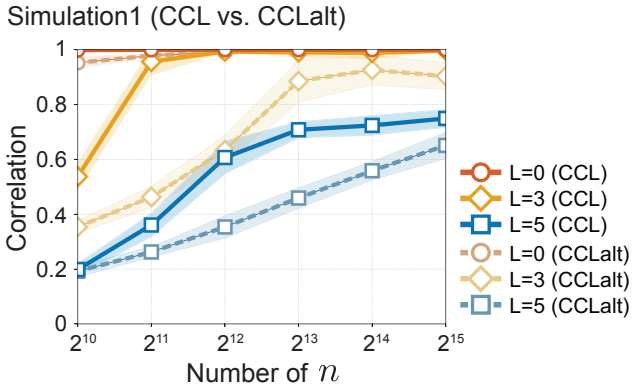

Figure 8: Comparison of the estimation performances of the latent components (Pearson correlation) by CCL and CCLalt, with different settings of the complexity of the observation models (the number of MLP-layers L of the observation function $\mathbf{f}$) and the number of n, on Simulation 1. $L = 0$ indicates that the latent components were directly obtained as the observations, while the observations were (unknown) nonlinear mixtures of the latent components in $L > 0$. The values are the averages of 10 runs for each setting, and the shaded regions show the standard deviations.