Identifying Weight-Variant Latent Causal Models

Yuhang Liu[1,2]   YUHANG.LIU01@ADELAIDE.EDU.AU
Zhen Zhang[1,2]   ZHEN.ZHANG02@ADELAIDE.EDU.AU
Dong Gong[3]   EDGONG01@GMAIL.COM
Mingming Gong[4]   MINGMING.GONG@UNIMELB.EDU.AU
Biwei Huang[5]   BIH007@UCSD.EDU
Anton van den Hengel[1,2]   ANTON.VANDENHENGEL@ADELAIDE.EDU.AU
Kun Zhang[6]   KUNZ1@CMU.EDU
Javen Qinfeng Shi[1,2]   JAVEN.SHI@ADELAIDE.EDU.AU

[1] *Responsible AI Research Centre, Australia*

[2] *Australian Institute for Machine Learning, Adelaide University, Australia*

[3] *School of Computer Science and Engineering, The University of New South Wales, Australia*

[4] *School of Mathematics and Statistics, The University of Melbourne, Australia*

[5] *Halicioğlu Data Science Institute, University of California San Diego, USA*

[6] *Department of Philosophy, Carnegie Mellon University, USA*

Abstract

The task of causal representation learning aims to uncover latent higher-level causal variables that affect lower-level observations. Identifying the true latent causal variables from observed data, while allowing instantaneous causal relations among latent variables, remains a challenge, however. To this end, we start with the analysis of three intrinsic indeterminacies in identifying latent variables from observations: transitivity, permutation indeterminacy, and scaling indeterminacy. We find that transitivity acts as a key role in impeding the identifiability of latent causal variables. To address the unidentifiable issue due to transitivity, we introduce a novel identifiability condition where the underlying latent causal model satisfies a linear-Gaussian model, in which the causal coefficients and the distribution of Gaussian noise are modulated by an additional observed variable. Under certain assumptions, including the existence of a reference condition under which latent causal influences vanish, we can show that the latent causal variables can be identified up to trivial permutation and scaling, and that partial identifiability results can still be obtained when this reference condition is violated for a subset of latent variables. Furthermore, based on these theoretical results, we propose a novel method, termed Structural caUsAl Variational autoEncoder (SuaVE), which directly learns causal representations and causal relationships among them, together with the mapping from the latent causal variables to the observed ones. Experimental results on synthetic and real data demonstrate the identifiability and consistency results and the efficacy of SuaVE in learning causal representations.

Project Page: https://sites.google.com/view/yuhangliu/projects/suave

Keywords: Causal Representation Learning, Latent Causal Models, Nonlinear ICA

1. Introduction

While there is no universal formal definition, one widely accepted feature of disentangled representations is that a change in one dimension corresponds to a change in one factor of variation in the underlying model of the data, while having little effect on others (Bengio et al., 2013). The underlying model is rarely available for interrogation, however, which makes learning disentangled representations challenging. Several excellent works for disentangled representation have been proposed that focus on enforcing independence over the latent variables that control the factors of variation (Higgins et al., 2017; Chen et al., 2018; Locatello et al., 2019; Kim and Mnih, 2018; Locatello et al., 2020). In many applications, however, the latent variables are not statistically independent, which is at odds with the notion of disentanglement, *i.e.*, foot length and body height exhibit strong correlation in the observed data (Träuble et al., 2021).

Causal representation learning avoids the aforementioned limitation, as it aims to learn a representation that exposes the unknown high-level latent causal variables, and the relationships between them, from a set of low-level observations (Schölkopf et al., 2021). Unlike disentangled representation learning, it allows the possible causal relations between latent variables. In fact, disentangled representation learning can be viewed as a special case of causal representation learning where the latent variables have no causal influences (Schölkopf et al., 2021). One of the most prominent additional capabilities of causal representations is the ability to represent interventions and to make predictions regarding such interventions (Pearl, 2000), which enables the generation of new samples that do not lie within the distribution of the observed data. This can be particularly useful to improve the generalization of the resulting model. Causal representations also enable answering counterfactual questions, *e.g.*, would a given patient have suffered heart failure if they had started exercising a year earlier?

Despite its advantages, causal representation learning is a notoriously hard problem. Without certain assumptions, identifying the true latent causal model from observed data is generally not possible. There are four primary approaches to achieve identifiability: 1) adapting (weakly) supervised methods with given latent causal graphs or/and labels (Kocaoglu et al., 2018; Yang et al., 2021), 2) imposing special graphical conditions with bottleneck graphical conditions (Adams et al., 2021; Xie et al., 2020a; Lachapelle et al., 2021), 3) using temporal information (Yao et al., 2021; Lippe et al., 2022), 4) using hard interventional data (Ahuja et al., 2023; Seigal et al., 2022). A brief review is provided in Section 2. In supervised approaches, when either the labels for latent variables or the true latent causal graph are available, the challenging identifiability problem in the latent space can be alleviated, or even bypassed by leveraging existing identifiability results in observed sapce (Zhang and Hyvarinen, 2012; Peters et al., 2014). However, such methods tend to rely heavily on domain-specific prior knowledge and are often associated with high annotation costs. For the second approach, many true latent causal graphs do not satisfy the assumed special graph structure. The temporal approach is only applicable when temporal information or temporal intervened information between latent factors is available. For using intervention data, most studies focus on hard interventions, particularly single-node and paired hard interventions. In contrast, our work considers the changes of causal influences, which is more aligned with soft interventions.

In this work, we explore a new direction in the identifiability of latent causal variables by examining the changes in causal influences between latent causal variables. Our approach is motivated by recent advances in nonlinear ICA, which, broadly speaking, demonstrate that latent independent variables can be identified when an additional observed variable **u** is introduced to modulate their distributions (Hyvarinen et al., 2019; Khemakhem et al., 2020). This raises a natural question:

Beyond the independent latent variables in nonlinear ICA, with causal relationships between latent variables, what additional assumptions are required for identifiability of latent causal variables?

- To answer the question above, we begin by analyzing three intrinsic indeterminacies in the latent space (see Section 3): transitivity, permutation indeterminacy, and scaling indeterminacy. This analysis leads to the following key insights. 1) Transitivity is the scourge of identifiability of latent causal models. 2) Permutation indeterminacy nature enables us to pre-define a causal order, avoiding troublesome directed acyclic graph (DAG) constraints. 3) Scaling and permutation indeterminacy only allow recovering latent causal variables up to permutation and scaling, not the exact values.

- To overcome the challenge of transitivity, we model the underlying causal generative process with weight-variate linear Gaussian models (see Section 4). In this model, both the weights (*i.e.*, causal influences) and the mean and variance of the Gaussian noise are modulated by an additional observed variable **u**.

- By exploring the changes in causal influences, i.e., the existence of a reference environment under which certain causal influences between latent variables vanish, we demonstrate that the latent causal variables can be recovered up to a trivial permutation and scaling. Combined with the independent causal mechanism, a common assumption in causality, we show that the latent causal structure is also identifiable. Further, realizing that requiring all weights to change may be challenging in real applications, we show partial identifiability results if only part of the weights change in Section 5.

- Based on our analysis, in Section 6, we additionally propose a novel method, Structural caUsAl Variational autoEncoder (SuaVE), for learning latent causal variables and latent causal graph among them. Section 7 verifies the efficacy of the proposed method on both synthetic and real fMRI data.

The key to the identifiability results in this work lies in leveraging changes in the causal influences (*e.g.*, weights in linear models) between latent causal variables. These changes correspond to distribution shifts, which can be interpreted as interventions acting on the latent variables. Such distribution shifts commonly occur across various domains, including medical imaging (Chandrasekaran et al., 2021), biogeography (Pinsky et al., 2020), and finance (Gibbs and Candes, 2021). Analogous to studies of distribution shifts in observed data (Ghassami et al., 2018; Huang et al., 2020), shifts in the latent space enable comparative analyses that reveal the underlying causal mechanisms governing relationships between latent variables. Thus, investigating distribution shifts offers a promising avenue for achieving identifiability in causal representation learning.

2. Related work

Due to the inherent challenges of identifiability in causal representation learning, most existing methods address this issue by imposing specific assumptions. We therefore provide a brief review of related work grounded on this basis.

(Weakly) Supervised Causal Representation Learning   Approaches in this category assume either known latent causal graphs or access to labels. For example, CausalGAN (Kocaoglu et al., 2018) requires prior knowledge of the latent causal graph structure, which limits its practical applicability. CausalVAE (Yang et al., 2021) relies on additional labels to supervise latent variable learning. However, such labels are often unavailable, and manual annotation can be costly and error-prone. Von Kügelgen et al. (2021) explore self-supervised causal representation learning using a known but complex causal graph between content and style factors. Meanwhile, Brehmer et al. (2022) propose a weakly supervised approach that assumes access to paired data before and after unknown interventions on the system.

Special Graphical Structure   Most recent advances in identifiability focus on imposing special graphical structure constraints (Silva et al., 2006; Shimizu et al., 2009; Anandkumar et al., 2013; Frot et al., 2019; Cai et al., 2019; Xie et al., 2020a, 2022a; Kivva et al., 2021). Adams et al. (2021) provided a unifying perspective, advocating for sparser models that fit the observations better. However, their analysis mainly considers linear causal relationships both between latent causal variables and between latent and observed variables, which often violated in real-world complex scenarios. Lachapelle et al. (2021) addressed nonlinear causal relations by assuming special sparse graphical structures. Yet, latent causal graphs encountered in practice are more complex and may not conform to strict sparsity assumptions. In contrast, our approach defines a function class for latent variables without imposing restrictive constraints on the underlying graph structure. Additionally, Kivva et al. (2021) focus on discrete latent causal variables within relatively limited graph structures, while our work considers continuous latent causal variables with flexible general graph configurations.

Temporal Information   The temporal constraint that the effect cannot precede the cause has been used repeatedly in latent causal representation learning (Yao et al., 2021; Lippe et al., 2022; Yao et al., 2022). For example, Yao et al. (2021) recover latent causal variables and the relations between them using Variational AutoEncoders and enforcing constraints in causal process prior. Lippe et al. (2022) learn causal representations from temporal sequences, which requires the underlying causal factors to be intervened. All of these works can be regarded as special cases of exploring the change of causal influences among latent variables in time series data. The proposed approach in this work tends to be more general, since the auxiliary variable **u** can represent not only time indices but also domain indices or virtually any additional side information.

Data from Hard Interventions   Very recently, there have been some works exploring interventional data (Brehmer et al., 2022; Ahuja et al., 2023; Seigal et al., 2022; Buchholz et al., 2023; Varici et al., 2023). Most of them consider hard intervention, and more restricted single-node and paired hard interventions (Ahuja et al., 2023; Seigal et al., 2022; Buchholz et al., 2023; Varici et al., 2023, 2024; Buchholz et al., 2024; von Kügelgen et al., 2024), which can only capture some special changes of causal influences. By contrast, this work studies

unpaired data, and employs soft interventions to model a broader range of possible changes. This is especially important considering that distribution shifts in real applications may be arbitrary. From this perspective, soft interventions could be easier to achieve for latent variables than hard interventions. To clarify this distinction, we introduce hard and soft interventions in the following.

Preliminaries: Hard Intervention and Soft Intervention Generally speaking, hard intervention (sometimes called perfect intervention) fixes a causal variable to a specific value, thereby breaking its dependence on all of its parent variables. In contrast, soft intervention modifies the distribution of a causal variable conditional on its parent nodes, such as by altering its parameters or noise distribution, without completely severing the causal influence from its parents. This added flexibility allows soft interventions to capture a broader and more realistic range of changes in causal influences, enabling the model to account for a broader array of changes in causal influences. This is particularly valuable for capturing changes in causal influences driven by uncontrollable behaviors, such as environmental fluctuations, within a causal system. Moreover, soft interventions generalize hard interventions, since a hard intervention can be viewed as a special case of a soft intervention where the distribution of the targeted causal variable collapses to a Dirac delta distribution.

3. Indeterminacies in Latent Causal Models

In this section, we first build a connection between nonlinear ICA and causal representation learning, by exploiting the correspondence between the independence of latent variables in nonlinear ICA and the independence of latent noise variables in causal systems. We then consider three indeterminacies in latent space: transitivity, permutation indeterminacy, and scaling indeterminacy, and analyze their impact on identifiability.

3.1 Relating Causal Representation Learning with Identifiable Nonlinear ICA
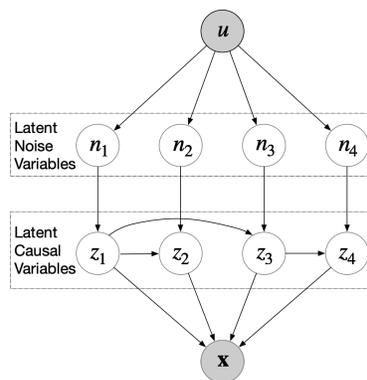

Figure 1: Latent Causal Models with $\mathbf{u}$.

Causal representation learning aims to uncover latent higher-level causal variable that can explain lower-level raw observations. Specifically, we assume that the observed variables $\mathbf{x}$ are influenced by latent causal variables z_i, and the causal structure among z_i can be any directed

acyclic graph (which is unknown). For each latent causal variable z_i, there is a corresponding latent noise variable n_i, as shown in Figure 1, which represents some unmeasured factors that influence z_i. The latent noise variables n_i are assumed to be independent with each other, *conditional on the observed variable* $\mathbf{u}$[1], in a causal system (Peters et al., 2017), so it is natural to leverage recent progress in nonlinear ICA (Hyvarinen et al., 2019; Khemakhem et al., 2020), which has shown that the independent latent noise variables n_i are identifiable under relatively mild assumptions, *i.e.*, one of the main assumptions is that $\{n_i\}$ are Gaussian distributed with their mean and variance modulated by an observed variable $\mathbf{u}$. Taking one step further, our goal is to recover the latent causal variables z_i. However, with the assumptions for the identifiability of latent noise variables n_i from nonlinear ICA, it is still insufficient to identify the latent causal variables z_i. The reason of such non-identifiability will be given in Section 3.2. Given this fact, a further question is what additional conditions are needed to recover the latent causal variables z_i. The corresponding identifiability conditions will be given in Section 4.

Understanding u   Essentially, the introduced $\mathbf{u}$ as a surrogate variable is to characterize the changes of independent noise variables n_i conditional on $\mathbf{u}$. In some cases, such a surrogate variable may be unavailable or hard to obtain in practice (Liu et al., 2021; Creager et al., 2021). Fortunately, it has been shown that learning or inferring such a surrogate variable is possible with additional auxiliary information (Lin et al., 2022). These information is often cheaply available for every input in practice (Xie et al., 2020b; Wang et al., 2020). Examples include time index of the data in time series forecasting tasks (Mudelsee, 2019), locations (longitude and latitude) of collected satellite data in remote sensing (Rußwurm et al., 2020), modality index in multi-modality dataset, and label in natural images.

3.2 Transitivity: the Challenge of Identifying Causal Representations

Even with the identifiable n_i, it is impossible to identify the latent causal variables z_i without additional assumptions. To interpret this point, for simplicity, let us only consider the influences of z_1 and z_2 on $\mathbf{x}$ in Figure 1. According to the graph structure in Figure 1, assume that $z_1 := n_1$, $z_2 := z_1 + n_2$ and $\mathbf{x} := \mathbf{f}(z_1, z_2)$ (case 1). We then consider the graph structure shown in the left column of Figure 2, where the edge $z_1 \rightarrow z_2$ has been removed, and assume that $z_1 := n_1$, $z_2 := n_2$ and $\mathbf{x} := \mathbf{f} \circ \mathbf{g}(z_1, z_2)$ where $\mathbf{g}(z_1, z_2) = [z_1, z_1 + z_2]$ (case 2). Interestingly, we find that the causal models in case 1 and case 2 generate the same observed data $\mathbf{x}$, which implies that there are two different causal models to interpret the same observed data. Clearly, z_2 in both the two equivalent causal structures are different, and thus z_2 is unidentifiable.

Similarly, we can cut the edge $z_1 \rightarrow z_3$ in Figure 1, obtaining another equivalent causal graph as shown in the right of Figure 2. That is, we can have two different z_3 to interpret the same observed data, and thus z_3 is unidentifiable. Since there exist many different equivalent causal structures, the latent causal variables in Figure 1 are unidentifiable. Such a result is because the effect of z_1 on z_2 (or z_3) in latent space can be 'absorbed' by the nonlinear function from $\mathbf{z}$ to $\mathbf{x}$. We term this indeterminacy transitivity in this work. We

---

1. For convenience in the later parts, with a slight abuse of definition, independent n_i means that n_i are mutually independent conditional on the observed variable $\mathbf{u}$.
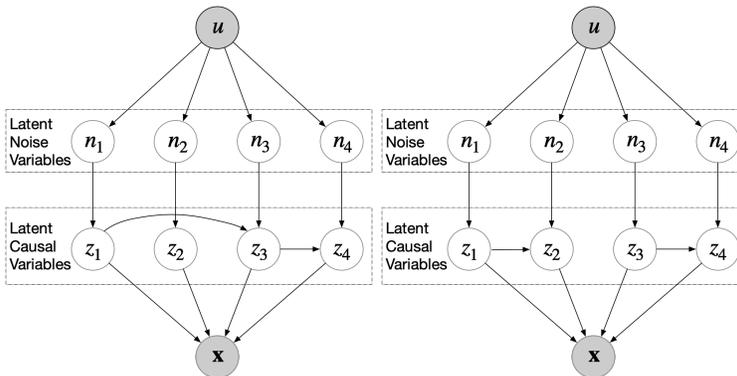
Figure 2: Two equivalent structures.

will show how to handle this challenge in Section 4. Before that, we introduce the other two indeterminacies in latent space, which assists in understanding our identifiability result.

3.3 Scaling Indeterminacy in Causal Representations

The scaling indeterminacy of latent causal variables is also an intrinsic indeterminacy in latent space. Again, for simplicity, we only consider the influence of z_1 and z_2 on $\mathbf{x}$ in Figure 1, and assume that $z_1 := n_1$, $z_2 := z_1 + n_2$ and $\mathbf{x} := \mathbf{f}(z_1, z_2)$. Under this setting, if the value of z_1 is scaled by s, $i.e.$, $s \times z_1$, we can easily obtain the same observed data $\mathbf{x}$ by: 1) letting $z_2 := \frac{1}{s} \times z_1 + n_2$ and 2) $\mathbf{x} := \mathbf{f} \circ \mathbf{g}(z_1, z_2)$ where $\mathbf{g}(z_1, z_2) = [s \times z_1, z_1]$. This indeterminacy is because the scaling of the latent variables z_i can be 'absorbed' by the nonlinear function from $\mathbf{z}$ to $\mathbf{x}$ and the causal functions among the latent causal variables. Therefore, without additional information to determine the values of the latent causal variables z_i, it is only possible to identify the latent causal variable up to scaling, not exactly recovering the values. In general, this scaling does not affect identifying the causal structure among the latent causal variables. We will further discuss this point in Section 4.

3.4 Permutation Indeterminacy in Causal Representations

Due to the nature of ill-posedness, latent causal representation learning suffers from permutation indeterminacy, where the recovered latent causal representations have an arbitrary permutation. For example, assume that the underlying (synthetic) latent causal representations are corresponding to the size, color, shape, and location of an object, and we obtain the recovered latent causal representations z_1, z_2, z_3, z_4. Permutation indeterminacy means that we can not ensure that the recovered z_1 represents which specified semantic information, $e.g.$, the size or the color. Therefore, without additional information, it is only possible to identify the latent causal representations up to permutation.

However, the flexibility and ambiguity above in latent space also offer an advantage, i.e., by pre-defining a causal order, we can avoid constraints like those proposed in Zheng et al. (2018), facilitating the learning of latent causal representations and causal structure. For example, For example, assume that the true latent causal relations for generating data are: size $\rightarrow$ color $\rightarrow$ shape $\rightarrow$ location. During the learning process, we can pre-define a
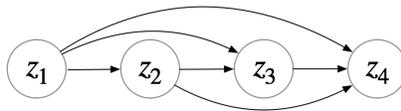
Figure 3: A Causal Fully-connected Graph, based on pre-defined causal order z_1, z_2, z_3, z_4, thanks to permutation indeterminacy in latent space.

causal order z_1, z_2, z_3, z_4, without specifying semantic information for the nodes. Since z_1 is the first node in the predefined causal order, z_1 is enforced to the semantic information of the first node in the true underlying causal order, *i.e.*, the size. Similarly, z_2 is enforced to the semantic information of the second node in the true underlying causal order, *i.e.*, the color. Furthermore, the predefined causal order allows us to naturally establish a causal fully-connected graph as depicted in Figure 3, ensuring the estimation of a directed acyclic graph (DAG) in learning causal representation and avoiding DAG constraints, such as those proposed by Zheng et al. (2018). We will further demonstrate how to implement this in the proposed method in Section 6.

4. Identifiable Causal Representations with Weight-variant Linear Gaussian Models

As discussed in Section 3.2, the key factor that impedes identifiable causal representations is the transitivity in latent space. Note that the transitivity is because the causal influences between the latent causal variables may be 'absorbed' by the nonlinear mapping from latent variables $\mathbf{z}$ to the observed variable $\mathbf{x}$. To address this issue, motivated by identifying causal structures with the change of causal influences in observed space (Ghassami et al., 2018; Huang et al., 2019, 2020), we allow causal influences between latent causal variables to be modulated by an additionally observed variable $\mathbf{u}$, represented by the red edge in Figure 4. Intuitively, variant causal influences between latent causal variables cannot be 'absorbed' by an invariant nonlinear mapping from $\mathbf{z}$ to $\mathbf{x}$, breaking the transitivity, resulting in identifiable causal representations. Specifically, we explore latent causal generative models where the observed data $\mathbf{x}$ is generated by the latent causal variables $\mathbf{z}$, allowing for any potential graph structures among $\mathbf{z}$. In addition, there exist latent noise variables $\mathbf{n}$, known as exogenous variables in causal systems, corresponding to latent causal variables. We introduce a surrogate variable $\mathbf{u}$ characterizing the changes in the distribution of $\mathbf{n}$, as well as the causal influences among latent causal variables $\mathbf{z}$. Here $\mathbf{u}$ could be environment, domain, or time index. More specifically, we parameterize the latent causal generative models as follows:

$$n_i :\sim \mathcal{N}(\beta_{i,1}(\mathbf{u}), \beta_{i,2}(\mathbf{u})), \tag{1}$$

$$z_i := \boldsymbol{\lambda}_i^T(\mathbf{u})(\mathbf{z}) + n_i, \tag{2}$$

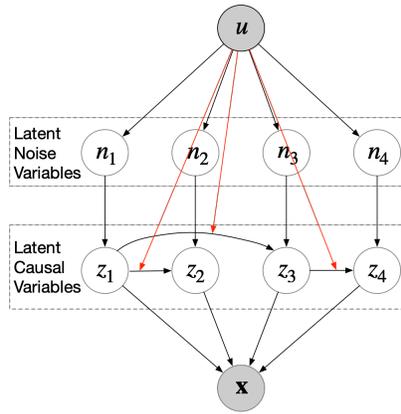$$\mathbf{x} := \mathbf{f}(\mathbf{z}, \boldsymbol{\varepsilon}). \tag{3}$$

where

Figure 4: The proposed Latent Causal Models. Here we use the 'red' lines to indicate the changes of weights across $\mathbf{u}$, which is different from the definition of edges (*i.e.*, causal direction) in standard causal model graphs.

- each noise term n_i is Gaussian distributed with mean $\beta_{i,1}(\mathbf{u})$ and variance $\beta_{i,2}(\mathbf{u})$, $\mathbf{n} \in \mathbb{R}^n$; both $\beta_{i,1}$ and $\beta_{i,2}$ can be nonlinear mappings. Moreover, the distribution of n_i is modulated by the observed variable $\mathbf{u}$.

- In Eq. (2), $\boldsymbol{\lambda_i}(\mathbf{u})$ denote the vector corresponding to the causal weights from $\mathbf{z}$, $\mathbf{z} \in \mathbb{R}^n$, to z_i. Specifically, $\boldsymbol{\lambda_i}(\mathbf{u}) = [\lambda_{1,i}(\mathbf{u}), ..., \lambda_{n,i}(\mathbf{u})]$, where each $\lambda_{j,i}$ could be a nonlinear mapping.

- In Eq. (3), $\mathbf{f}$ denote a nonlinear mapping from $\mathbf{z}$ to $\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^d$ and $\boldsymbol{\varepsilon}$ is independent noise with probability density function $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$, $\boldsymbol{\varepsilon} \in \mathbb{R}^{d-n}$.

In Eq. (1), we assume the latent noise variables to be Gaussian, thus the joint distribution can be expressed by the following exponential family:

$$p_{(\mathbf{T_n}, \boldsymbol{\beta})}(\mathbf{n}|\mathbf{u}) = \frac{1}{Z_{\mathbf{n}}(\boldsymbol{\beta}, \mathbf{u})} \exp\left(\mathbf{T_n}^T(\mathbf{n})\boldsymbol{\eta_n}(\mathbf{u})\right), \qquad (4)$$

where $Z_{\mathbf{n}}(\boldsymbol{\beta}, \mathbf{u})$ denotes the normalizing constant, and $\mathbf{T_n}(\mathbf{n})$ denotes the sufficient statistic for $\mathbf{n}$, whose the natural parameter $\boldsymbol{\eta_n}(\mathbf{u})$ depends on $\beta_{i,1}$ and $\beta_{i,2}$.

In Eq. (2), we assume the latent causal model of each z_i satisfies a linear causal model with the causal weights being modulated by $\mathbf{u}$, and we call it *weight-variant linear Gaussian models*. Therefore, $p(\mathbf{z}|\mathbf{u})$ satisfies the following multivariate Gaussian distribution:

$$p_{(\boldsymbol{\lambda}, \boldsymbol{\beta})}(\mathbf{z}|\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma), \qquad (5)$$

9

with the mean $\boldsymbol{\mu}$ and the covariance matrix Σ computed by the following recursion relations (Bishop and Nasrabadi, 2006; Koller and Friedman, 2009):

$$
\begin{aligned}
\mu_i &= \sum_{j \in \mathrm{pa}_i} \lambda_{j,i}(\mathbf{u})\mu_j + \beta_{i,1}(\mathbf{u}), \\
\Sigma_{i,i} &= \sum_{j \in \mathrm{pa}_i} \lambda_{j,i}^2(\mathbf{u})\Sigma_{j,j} + \beta_{i,2}(\mathbf{u}), \\
\Sigma_{i,j} &= \sum_{k \in \mathrm{pa}_j} \lambda_{k,j}(\mathbf{u})\Sigma_{i,k}, \quad for \quad i \neq j,
\end{aligned}
\tag{6}
$$

where pa_i denotes the parent nodes of z_i. Furthermore, this multi-variate Gaussian can be re-formulated with the following exponential family:

$$
p_{(\mathbf{T_z}, \boldsymbol{\lambda}, \boldsymbol{\beta})}(\mathbf{z}|\mathbf{u}) = \frac{1}{Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u})} \exp\left(\mathbf{T}_{\mathbf{z}}^T(\mathbf{z})\boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}) \right),
\tag{7}
$$

where the parameter $\boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}) = [\Sigma^{-1}\boldsymbol{\mu}; -\frac{1}{2}\mathrm{vec}(\Sigma^{-1})]$, $Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u})$ denotes the normalizing constant, and the sufficient statistic $\mathbf{T}_{\mathbf{z}}(\mathbf{z}) = [\mathbf{z}; \mathrm{vec}(\mathbf{z}\mathbf{z}^T)]$ ('vec' denotes the vectorization of a matrix). We further denote by $\mathbf{T}_{\min}$ the minimal sufficient statistic and by k its dimension, with $2n \leq k \leq n + (n(n+1))/2$. In particular, $k = 2n$ corresponds to the case when $\mathbf{T}_{\min}(\mathbf{z}) = [z_1, z_2, ..., z_n, z_1^2, z_2^2, ..., z_n^2]$, or in other words, there is no edges among the latent variables $\mathbf{z}$, while $k = n + (n(n+1))/2$ corresponds to a full-connected causal graph over $\mathbf{z}$. So with different causal structures over $\mathbf{z}$, the dimension k may vary. For the graph structure in Figure 1, $\mathbf{T}_{\min}(\mathbf{z}) = [z_1, z_2, z_3, z_4, z_1^2, z_2^2, z_3^2, z_4^2, z_1 z_2, z_1 z_3, z_3 z_4]$ and $k = 11$.

Given the above, the following theorem shows that under certain assumptions on the nonlinear mapping $\mathbf{f}$ and the variability of $\mathbf{u}$, the latent variables $\mathbf{z}$ can be identifiable up to trivial permutation and scaling.

Theorem 1 *Suppose latent causal variables $\mathbf{z}$ and the observed variable $\mathbf{x}$ follow the generative models defined in Eq. (1)- Eq. (3), with parameters $(\mathbf{f}, \boldsymbol{\lambda}, \boldsymbol{\beta})$. Assume the following holds:*

(i) $\boldsymbol{\varepsilon}$ is independent noise with probability density function $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$, which is always finite.

(ii) The function $\mathbf{f}$ in Eq. (3) is invertible and smooth.

(iii) There exist $2n + 1$ distinct points $\mathbf{u}_{\mathbf{n},0}, \mathbf{u}_{\mathbf{n},1}, ..., \mathbf{u}_{\mathbf{n},2n}$ such that the matrix

$$
\mathbf{L_n} = (\boldsymbol{\eta}_{\mathbf{n}}(\mathbf{u}_{\mathbf{n},1}) - \boldsymbol{\eta}_{\mathbf{n}}(\mathbf{u}_{\mathbf{n},0}), ..., \boldsymbol{\eta}_{\mathbf{n}}(\mathbf{u}_{\mathbf{n},2n}) - \boldsymbol{\eta}_{\mathbf{n}}(\mathbf{u}_{\mathbf{n},0}))
\tag{8}
$$

of size $2n \times 2n$ is invertible.

(iv) There exist $k + 1$ distinct points $\mathbf{u}_{\mathbf{z},0}, \mathbf{u}_{\mathbf{z},1}, ..., \mathbf{u}_{\mathbf{z},k}$ such that the matrix

$$
\mathbf{L_z} = (\boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},1}) - \boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},0}), ..., \boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},k}) - \boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},0}))
\tag{9}
$$

of size $k \times k$ is invertible.

(v) *The function class of $\lambda_{i,j}$ can be expressed by a Taylor series: for each $\lambda_{i,j}$, $\lambda_{i,j}(\mathbf{u} = \mathbf{0}) = 0$,*

then the true latent causal variables $\mathbf{z}$, which are learned by matching the true marginal data distribution $p(\mathbf{x}|\mathbf{u})$, are related to the estimated latent causal variables $\hat{\mathbf{z}}$ by the following relationship: $\mathbf{z} = \mathbf{P}\hat{\mathbf{z}} + \mathbf{c}$, where $\mathbf{P}$ denotes the permutation matrix with scaling, $\mathbf{c}$ denotes a constant vector.

Proof sketch The proof can be done according to the following intuition. With the support of assumptions (i)-(iii), one can utilize the identifiability result from nonlinear ICA to identify the latent noise variables $\mathbf{n}$ up to permutation and scaling, *i.e.*, $\mathbf{n} = \mathbf{P}\hat{\mathbf{n}} + \mathbf{c_n}$ where $\hat{\mathbf{n}}$ denotes the recovered latent noise variables obtained by matching the true marginal data distribution, and $\mathbf{c_n}$ notes a constant vector. Next, using a similar technique, we can utilize assumption (iv) to establish a linear transformation between the true latent causal variables $\mathbf{z}$ and the recovered latent causal variables $\hat{\mathbf{z}}$, *i.e.*, $\mathbf{z} = \mathbf{A}(\hat{\mathbf{z}}) + \mathbf{c_z}$ where $\mathbf{c_z}$ denotes a constant vector, and $\mathbf{A}$ denotes a liner transformation. Finally, combining the identifiability result for the latent noise variables $\mathbf{n}$ with assumption (v), the linear transformation can be reduced to permutation and scaling, *i.e.*, $\mathbf{z} = \mathbf{P}(\hat{\mathbf{z}}) + \mathbf{c_z}$. See Appendix A.1 for details.

Assumptions (i)-(iv) Assumptions (i)-(iii) are originally developed by nonlinear ICA (Hyvarinen and Morioka, 2016; Hyvarinen et al., 2019; Khemakhem et al., 2020; Sorrenson et al., 2020). We here consider unitizes these assumptions considering the following two main reasons. 1) These assumptions have been verified to be practicable in diverse real-world application scenarios Kong et al. (2022); Xie et al. (2022b). 2) Our result eliminates the need to know the exact dimensionality of latent causal or noise variables, which is in contrast to existing methods that require prior knowledge of the dimensionality, due to imposing the two-parameter Gaussian noise variables (Sorrenson et al., 2020). Assumptions (iv)-(v) originally introduced by this work. Intuitively, assumption (iv) is similar to assumption (iii) in the hope that the auxiliary variable $\mathbf{u}$ must have a sufficiently strong and diverse effect on the distributions of latent causal variables $\mathbf{z}$, similar to latent noise variables $\mathbf{n}$.

Assumptions (v) and sufficient change Among the assumptions above, the only one that cannot be considered weak or natural is assumption (v). Essentially, it is to ensure *sufficient changes* in causal influence. For simplicity, consider two variables, z_1 and z_2, and $z_2 := \lambda(\mathbf{u})z_1 + n_2$, where $\lambda(\mathbf{u}) = \hat{\lambda}(\mathbf{u}) + b$, with λ and $\hat{\lambda}$ belonging to the same function class. As a consequence, while the causal influence $\lambda(\mathbf{u})$ changes as a whole across $\mathbf{u}$, there always exists a part bz_1 that remains unchanged across $\mathbf{u}$. This unchanged term bz_1 across $\mathbf{u}$ can be absorbed into $\mathbf{f}$ the mapping from $\mathbf{z}$ to $\mathbf{x}$, resulting in a possible solution $\hat{z}_2 := \hat{\lambda}(\mathbf{u})z_1 + n_2$, instead of the groundtruth $(\hat{\lambda}(\mathbf{u}) + b)z_1 + n_2$, which leads to an unidentifiable outcome. Essentially, this unidentifiable result occurs because a part of a weight remains unchanged across $\mathbf{u}$, a situation we refer to as *insufficient change*. The objective of assumption (v) is to prevent such insufficient changes. Specifically, by constraining the function class of λ such that $\lambda(\mathbf{u} = \mathbf{0}) = 0$ and thus $\hat{\lambda}(\mathbf{u} = \mathbf{0}) = 0$, $\lambda(\mathbf{u})$ can not be expressed by $\hat{\lambda}(\mathbf{u}) + b$ with $b \neq 0$. Intuitively, assumption (v) suggests that there is a scenario where an existing edge can be removed at a certain $\mathbf{u}$, *i.e.*, $\lambda(\mathbf{u} = \mathbf{0}) = 0$. By removing the edge, we can ensure that no part of the weight corresponding to the edge remains unchanged. It is important to emphasize that assumption (v) is to limit the function class, so that λ do not include

a constant term, such as b, by assuming $\lambda(\mathbf{u} = \mathbf{0}) = 0$. Once samples are drawn from this limited function class, the assumption is satisfied, making the observed data generated by these samples effective for inferring latent causal variables. Therefore, we do not need the samples to necessarily include the specific point $\mathbf{u} = \mathbf{0}$, to generate the corresponding observed data, which is then used for inferring the latent causal variables.

Assumption (v) as a Bridge Between Soft and Hard Interventions   Assumption (v) implies that, although a causal edge may exist between two latent variables, there can be a reference condition under which this influence effectively vanishes, i.e., $\lambda(\mathbf{u} = \mathbf{0}) = 0$. This bears a close connection to hard interventions, which remove causal dependencies by construction. Unlike existing approaches that explicitly rely on hard interventions (Ahuja et al., 2023; Seigal et al., 2022; Buchholz et al., 2023; Varici et al., 2023), our formulation characterizes a restricted function class for $\lambda(\mathbf{u})$ that enables identifiability without requiring direct access to such interventions. We view this as a conceptual link between soft and hard interventions, while a systematic unification is beyond the scope of this work.

Potential Application   Assumption (v) can arise naturally in biological imaging settings, such as cell imaging experiments where samples are exposed to different small-molecule compounds. In this context, latent variables may represent protein groups, and causal edges correspond to protein–protein interactions (PPIs). Certain compounds are known to disrupt or inhibit specific PPIs, effectively suppressing the corresponding causal influences (Arkin and Wells, 2004; Lu et al., 2020). Such scenarios provide a plausible instance where $\lambda_{i,j}(\mathbf{u} = \mathbf{0}) = 0$ approximately holds, making the assumption empirically relevant.

Insights   Intuitively, allowing causal influences among latent causal variables to change leads to varying causal influences. Such variations cannot be 'absorbed' by an invariant nonlinear mapping from $\mathbf{z}$ to $\mathbf{x}$, which breaks the transitivity mentioned in Section 3.2, resulting in identifiability. For example, assume that $z_1 := n_1$, $z_2 := \lambda(\mathbf{u})z_1 + n_2$ and $\mathbf{x} := \mathbf{f}(z_1, z_2)$. If we allow the causal influence $\lambda(\mathbf{u})z_1$ to be 'absorbed' by the mapping $\mathbf{f}$, we would have $z_1 := n_1$, $z_2 := n_2$, and the resulting mapping, $\mathbf{f}' = \mathbf{f} \circ \mathbf{g}$ where $\mathbf{g}(z_1, z_2) = [z_1, \lambda(\mathbf{u})z_1 + z_2]$, would need to include the original $\mathbf{f}$ as well as the term $\lambda(\mathbf{u})$ in $\mathbf{g}$, to obtain the same obervation $\mathbf{x}$. This means that the resulting mapping $\mathbf{f}'$ would change across $\mathbf{u}$ due to the term $\lambda(\mathbf{u})$, which violates the fact that the mapping from $\mathbf{z}$ to $\mathbf{x}$ remains unchanged across $\mathbf{u}$ as defined in Eq. (3).

Identifiability of Causal Structures among Latent Variables   Theorem 1 has shown that latent causal variables can be identified up to trivial permutation and linear scaling. With this result, the identifiability of causal structure in the *latent space* reduces to the identifiability of the causal structure in the *observational space*. Denote by $G_{\mathbf{z}}^u$ the causal graph over latent variables $\mathbf{z}$ when $\mathbf{u} = u$, and let $G_{\mathbf{z}}^{\mathrm{union}} = \cup_u G_{\mathbf{z}}^u$ be the union of causal graphs across different values of $\mathbf{u}$. Similarly, we define $G_{\mathbf{z} \cup \mathbf{u}}^{\mathrm{union}}$ to be the union of causal graphs over $\mathbf{z} \cup \mathbf{u}$. With the help of recent progress in causal discovery from heterogeneous data (Huang et al., 2020), the following corollary shows that the causal structure among latent variables $\mathbf{z}$ is identifiable up to the Markov equivalence class of $G_{\mathbf{z}}^{\mathrm{union}}$. Moreover, although there exists scaling indeterminacy for the recovered latent variables as stated in Theorem 1, it does not affect the identifiability of the causal structure.

Corollary 2 *Suppose latent causal variables* $\mathbf{z}$ *and the observation* $\mathbf{x}$ *follow the generative models defined in Eq. (1)- Eq. (3), and that the conditions in Theorem 1 hold. Then, under the assumption that* $G_{\mathbf{z}}^{union}$ *is a DAG and that the joint distribution over* $\mathbf{z} \cup \mathbf{u}$ *is Markov and faithful to* $G_{\mathbf{z} \cup \mathbf{u}}^{union}$, *the causal structure among variables* $\mathbf{z}$ *can be identified up to the Markov equivalence class of* $G_{\mathbf{z}}^{union}$. *That is,* $MEC(G_{\mathbf{z}}^{union})$ *is identifiable, where* $MEC(\cdot)$ *denotes the Markov equivalence class.*

Proof Sketch The proof can be done by first establishing that Theorem 1 identifies the latent causal variables up to trivial permutation and linear scaling, which reduces the problem to identifying the causal structure in the observed space. Then, by leveraging the results from Huang et al. (2020), which ensure identifiability of the causal structure in the observed space under the Markov condition and faithfulness assumption, we can identify the causal structure up to the Markov equivalence class. Next, we show that the Markov equivalence class over $\mathbf{z}$ remains unchanged after removing the domain variable $\mathbf{u}$ and its edges, preserving the graph's skeleton and directions. Finally, since linear scaling does not affect conditional independence relationships, the identifiability of the causal structure in the latent space is preserved. As a result, the causal structure among the latent variables $\mathbf{z}$ can be identified up to the Markov equivalence class. Please refer to Appendix A.3 for more details.

In addition, with the help of the independent causal mechanisms (ICM) principle (Ghassami et al., 2018; Huang et al., 2020; Schölkopf et al., 2021), the following corollary shows that the causal structure among latent variables $\mathbf{z}$ is fully identifiable.

Corollary 3 *Suppose the conditions in Corollary 2 hold. Denote by* $\boldsymbol{\theta}_i$ *the involved parameters in the causal mechanism of* z_i, *with* $\boldsymbol{\theta}_i = (\beta_{i,1}; \beta_{i,2}; \boldsymbol{\lambda}_i)$, *and denote by* $z_{pa(i)}$ *the parents of* z_i *in* $G_{\mathbf{z}}^u$. *If* $\boldsymbol{\theta}_i$ *and* $\boldsymbol{\theta}_{pa(i)}$ *change independently across different values of* $\mathbf{u}$ *for any* z_i, *then the acyclic causal structure among latent variables* $\mathbf{z}$ *can be fully identified. That is,* $G_{\mathbf{z}}^{union}$ *is fully identifiable.*

Proof Sketch To prove this corollary, we first introduce the ICM principle, which states that in a causally sufficient system, the causal modules and their parameters change independently across different domains. This principle holds in the causal direction, meaning that changes in one module do not affect others. However, this independence generally breaks down in the anti-causal direction. Clearly, The scaling indeterminacy for the recovered latent variables does not affect this principle. Please refer to Appendix A.4 for more details.

5. Change of Part of Weights and Partial Identifiability Result

The aforementioned theoretical result necessitates that all weights undergo changes across $\mathbf{u}$, as constrained by the assumption (v) in Theorem 1. However, in practical applications, this assumption may not hold true. Consequently, two fundamental questions naturally arise: Is this assumption necessary for identifiability in the absence of any supplementary assumptions? Alternatively, can we obtain partial identifiability results if only some of the weights change across $\mathbf{u}$? In fact, when part of weights change, we can still provide partial identifiability results, as outlined below.

Theorem 4 *Suppose latent causal variables* $\mathbf{z}$ *and the observed variable* $\mathbf{x}$ *follow the generative models defined in Eq.* (1)*-Eq.* (3)*. Under the condition that the assumptions* (i)*-*(iv) *in Theorem 1 are satisfied, for each* z_i,

    (a) *if* z_i *is a root node or all weights to* z_i, $\boldsymbol{\lambda}_i$ *in Eq.* (2)*, meet assumption* (v) *in Theorem 1, then the true* z_i *is related to the recovered one* $\hat{z}_j$, *obtained by matching the true marginal data distribution* $p(\mathbf{x}|\mathbf{u})$, *by the following relationship:* $z_i = s\hat{z}_j + c$, *where* s *denotes scaling, c denotes a constant,*

    (b) *if there exists an unchanged weight* $\lambda_{j,i}$ *from the parent node* z_j *to* z_i, *in Eq.* (2)*, then* z_i *is unidentifiable,*

    (c) *if there exists an weight that can be expressed as* $\lambda_{j,i}(\mathbf{u}) + b$, *where* b *is a non-zero constant, from the parent node* z_j *to* z_i, *then* z_i *is unidentifiable.*

Proof sketch This can be proved by the fact that regardless of the assumption (v) in Theorem 1 , two results hold in Theorem 1, i.e., $\mathbf{z} = \mathbf{A}(\hat{\mathbf{z}}) + \mathbf{c_z}$, and $\mathbf{n} = \mathbf{P}\hat{\mathbf{n}} + \mathbf{c_n}$. Then using the change of all weights in $\boldsymbol{\lambda}_i$, we can prove (a). To prove (b), we demonstrate that it is always possible to construct an alternative solution $\hat{z}_i$ by removing the term $\lambda_{j,i} z_j$, corresponding to unchanged weight, in the true z_i, which is capable of generating the same observations $\mathbf{x}$. As a result, we can show non-identifiable result. To prove (c), similar to (b), we can construct an alternative solution $\hat{z}_i$ by removing the term bz_j, corresponding to the constant b, in true z_i.

Insights 1) The result (b) of Theorem 4 implies the necessity of changing weights, without additional assumptions. This can be proven by the contrapositive of (b), i.e., if z_i is identifiable, then there do not exist unchanged weights from the parent nodes to z_i. 2) The result (c) of Theorem 4 implies the necessity of eliminating the corner case where $\lambda(\mathbf{u}) = \lambda'(\mathbf{u}) + b$, without additional assumptions. Similarly, this can be proven by the contrapositive of (c), i.e., if z_i is identifiable, then there do not exist a weight that can be expressed as $\lambda(\mathbf{u}) + b$. Essentially, this implies that sufficient changes in weights are required for identifiability. Informally, if $z_i = \boldsymbol{\lambda}_i^T \mathbf{z} + n_i$, then sufficient changes mean that $\boldsymbol{\lambda}_i^T$ do not include non-zero constant across $\mathbf{u}$. 3) Together with result (a) of Theorem 4, Theorem 4 implies that the entire latent space can theoretically be partitioned into two distinct subspaces: one subspace pertains to invariant latent variables, while the other encompasses variant variables. This partitioning may be valuable for applications that focus on learning invariant latent variables to adapt to varying environments, such as domain adaptation or generalization.

Discussion 1: Partial Causal Structure Partial identifiability of latent causal variables in Theorem 4 does not necessarily guarantee the unique recovery of the corresponding partial latent causal graph structure. However, a probable result can still be achieved. Specifically, if there are no interactions (edges) between the two latent subspaces in the ground truth graph structure, it becomes possible to recover the latent causal structure within the latent variant space. When interactions do exist, examining how they affect the recovery of the latent causal graph structure is an intriguing area for further exploration. Additionally, it is valuable to investigate how such partial results influence the outcomes of interventions and counterfactual inference.

Discussion 2: Parent nodes do not impact children Perhaps counterintuitively, the partial identifiability result in Theorem 4 suggests that the identifiability of z_i remains possible even when its parent nodes are unidentifiable. This is primarily due to the identifiability of all latent noise variables $\mathbf{n}$, regardless of changes in the weights. In this context, for latent linear causal models, all necessary information to recover latent causal variables is encapsulated within these identifiable latent noise variables. As a result, it is possible to recover z_i, irrespective of the identifiability of its parent nodes.

Discussion 3: Potential Implications While partial identifiability may be more realistic in practice, since only some causal influences vary across environments, it still poses challenges in determining whether the latent causal variables have been successfully identified or are practically useful, especially in the absence of ground-truth latent variables. Nevertheless, such partial identifiability results may be valuable in downstream applications, particularly in domain shift scenarios such as domain adaptation, domain generalization, and transfer learning (Kong et al., 2022; Li et al., 2023; Liu et al., 2025). In these settings, partially identified latent causal variables that exhibit stable causal relationships with the target label across environments may capture invariant mechanisms useful for robust prediction. Their relevance can be empirically assessed through predictive performance, stability under distribution shifts.

6. Learning Causal Representations with Weight-Variant Linear Models

Based on the identifiable results above, in this section, we propose a structural causal variational autoencoder (SuaVE) to learn latent causal representations. We first propose a structural causal model prior relating to the proposed weight-variant linear Gaussian model. We then show how to incorporate the proposed prior into the traditional VAE framework (Kingma and Welling, 2013a), together with a variational posterior.

DAG constraints To address the complexity of combinatorial search, recent methods have proposed exact characterizations of DAGs that enable tackling the problem through continuous optimization techniques (Zheng et al., 2018; Yu et al., 2019; Zheng et al., 2020; He et al., 2021). However, this approach does not guarantee the absence of cycles at any stage of training, and solutions often require post-processing (Zantedeschi et al., 2023). By contrast, as mentioned in Section 3.4, the permutation indeterminacy in latent space allows us to predefine a causal order without specifying semantic information. With the guarantee of identifiability, the nodes in the predefined order are enforced to learn the corresponding semantic information in the true causal order for generating data, ensuring each node learns its designated role. Given this, we can incorporate a predefined causal order in the following prior and posterior models.

Prior Model According to the causal model for the latent causal variables, *i.e.*, Eq. (1) and Eq. (2), we construct a linear Gaussian prior with weights and noise depending on $\mathbf{u}$. Since we can pre-define the causal order as mentioned above, let us assume the causal order is $z_1, z_2, ..z_n$. Given this, we can create a fully connected causal graph based on the predefined causal order, where each z_i has parent nodes $z_{i'}$ for $i' < i$. Summarizing the above,

we propose the following prior model:

$$p(\mathbf{z}|\mathbf{u}) = p(z_1|\mathbf{u}) \prod_{i=2}^{n} p(z_i|\mathbf{z}_{<i}, \mathbf{u}) = \prod_{i=1}^{n} \mathcal{N}(\mu_{z_i}, \delta_{z_i}^2), \tag{10}$$

where $\mu_{z_i} = \sum_{i'<i} \lambda_{i',i}(\mathbf{u})z_{i'} + \beta_{i,1}(\mathbf{u}), \delta_{z_i}^2 = \beta_{i,2}(\mathbf{u})$, and the corresponding weight matrix $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_1(\mathbf{u}), ..., \boldsymbol{\lambda}_n(\mathbf{u})]$ can be constrained as a upper triangular matrix with zero-value diagonal elements. Again, the proposed prior naturally ensures a directed acyclic graph estimation, avoiding traditional DAG constraints.

Variational Posterior and Objective The nature of the proposed prior in Eq. (10) gives rise to the following variational posterior:

$$q(\mathbf{z}|\mathbf{x}, \mathbf{u}) = q(z_1|\mathbf{x}, \mathbf{u}) \prod_{i=2}^{n} q(z_i|\mathbf{z}_{<i}, \mathbf{x}, \mathbf{u}) = \prod_{i=1}^{n} \mathcal{N}(\mu_{z_i}', \delta_{z_i}'^2), \tag{11}$$

where $\mu_{z_i}' = \sum_{i'<i} \lambda_{i',i}'(\mathbf{u})z_{i'} + \beta_{i,1}'(\mathbf{x}, \mathbf{u}), \delta_{z_i}'^2 = \beta_{i,2}'(\mathbf{x}, \mathbf{u})$. Therefore, we can arrive at the following evidence lower bound (ELBO):

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u})}(\log p(\mathbf{x}|\mathbf{z}, \mathbf{u})) - D_{KL}(q(\mathbf{z}|\mathbf{x}, \mathbf{u})||p(\mathbf{z}|\mathbf{u})), \tag{12}$$

where D_{KL} denotes Kullback–Leibler divergence, a measure of how one probability distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{u})$ is different from a second, reference probability distribution $p(\mathbf{z}|\mathbf{u})$. We implement assumption (v) and incorporate it into the evidence lower bound by the following objective:

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u})}(\log p(\mathbf{x}|\mathbf{z}, \mathbf{u})) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \mathbf{u})||p(\mathbf{z}|\mathbf{u})) - \gamma_1 \sum_{i,j} ||\lambda_{i,j}(\mathbf{u} = \mathbf{0})||_1 - \gamma_2 \sum_{i,j} ||\lambda_{i,j}'(\mathbf{u} = \mathbf{0})||_1. \tag{13}$$

In our implementation, the observed variable $\mathbf{u}$ is represented using one-hot encoding, so that in the experiments $\mathbf{u}$ takes values in a finite discrete set. In this discrete setting, Assumption (v) can in principle be satisfied exactly by fixing $\lambda_{i,j}(\mathbf{u} = 0) = 0$ by construction and learning $\lambda_{i,j}(\mathbf{u})$ only for $\mathbf{u} \neq 0$. In our implementation, we instead adopt a simple penalty that encourages $\lambda_{i,j}(\mathbf{u} = 0)$ to be close to zero, which serves as a convenient implementation choice. This penalty is included in the last two terms of Eq. (13), where γ_1 and γ_2 are hyperparameters that balance the penalty with the ELBO. Figure 5 illustrates the overall implementation of the proposed method.

Discussion: Relating to hierarchical VAE Beyond standard VAEs, recent advanced works aim to explore hierarchical VAEs with the hope that they could improve the evidence lower bound in standard VAE and decrease reconstruction error compared to standard VAEs. More importantly, the stack of latent variables in hierarchical VAEs might learn a feature hierarchy, where different layers capture information at varying levels of detail and abstraction (Kingma et al., 2016; Sønderby et al., 2016; Maaløe et al., 2019; Vahdat and Kautz, 2020). To this end, hierarchical VAE models often employ the following generative model: $z_1 \to z_2 \to ... \to z_n \to \mathbf{x}$. Interestingly, hierarchical VAE models can be regarded
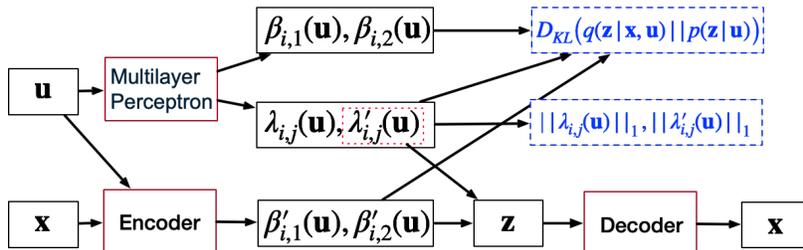
Figure 5: Implementation framework to learn linear Gaussian causal representations.

as a special case of the proposed prior model in Eq. (10), where the fully-connected graph is reduced to the simpler connections used in the generative model of hierarchical VAEs. Essentially, the proposed SuaVE also aims to learn hierarchical latent variables. Due to permutation indeterminacy, we can model the prior or posterior as hierarchical structures. The difference is that the edges among latent variables of the proposed SuaVE are learned adaptively from data, unlike the simple connections in hierarchical VAEs. Importantly, the proposed SuaVE has rigorous theoretical justification for learning hierarchical latent variables and the structures among them, whereas hierarchical VAEs lack such support without further assumptions. Further exploring the relationship between hierarchical VAEs and latent causal models is both interesting and worthwhile.

7. Experiments

7.1 Synthetic Data

We first conduct experiments on synthetic data, generated by the following process: we divide the latent noise variables into M segments, where each segment corresponds to one conditional variable $\mathbf{u}$ as the segment label. Within each segment, we first sample the mean $\beta_{i,1}$ and variance $\beta_{i,2}$ from uniform distributions $[-2, 2]$ and $[0.01, 3]$, respectively. Subsequently, we sample the weights $\lambda_{i,j}$ from a uniform distribution over $[0.1, 2]$. Then for each segment, we generate the latent causal samples according to the generative model in Eq. (2). Finally, we obtain the observed data samples $\mathbf{x}$ by an invertible nonlinear mapping. More details can be found in Appendix.

Comparison We compare the proposed method with identifiable VAE (iVAE) (Khemakhem et al., 2020), β-VAE (Higgins et al., 2017), CausalVAE (Yang et al., 2021), and vanilla VAE (Kingma and Welling, 2013b). Among them, iVAE has been proven to be identifiable so that it is able to learn the true independent noise variables with certain assumptions. While β-VAE has no theoretical support, it has been widely used in various disentanglement tasks. Note that both methods assume that the latent variables are independent, and thus they cannot model the relationships among latent variables. To make a fair comparison in the unsupervised setting, we implement an unsupervised version of CausalVAE, which is not identifiable.

Performance metric Since the proposed method can recover the latent causal variables up to trivial permutation and linear scaling, we compute the mean of the Pearson correlation
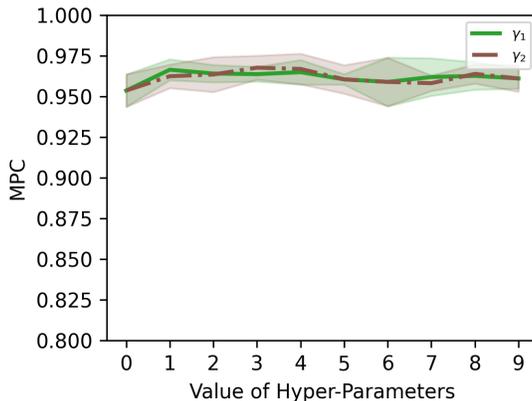
Figure 6: Performances of the proposed SuaVE with different values of hyper-parameters γ_1 and γ_2. Implementation details can be found in Appendix A.6.2.

coefficient (MPC) to evaluate the performance of our proposed method. The Pearson correlation coefficient is a measure of linear correlation between the true latent causal variables and the recovered latent causal variables. Note that the Pearson coefficient is suitable for iVAE, since it has been shown that iVAE can also recover the latent noise variables up to linear scaling under the setting where the mean and variance of the latent noise variables are changed by $\mathbf{u}$ (Sorrenson et al., 2020). To remove the permutation effect, following Khemakhem et al. (2020), we first calculate all pairs of correlation and then solve a linear sum assignment problem to obtain the final results. A high correlation coefficient means that we successfully identified the true parameters and recovered the true variables, up to component-wise linear transformations.

Ablation Study We first conduct experiments on the proposed method, varying the hyper-parameters γ_1 and γ_2. Figure 6 shows the results of performance of the proposed SuaVE with different values of hyper-parameters. The performance remains relatively stable on synthetic data. This suggests that on randomly generated synthetic data, the variation in weights generally satisfies assumption (v), which requires avoiding insufficient change—specifically, the scenario where a portion of the weights remains unchanged across $\mathbf{u}$. This stability occurs because the weights are independently and randomly generated without any inherent patterns, making it unlikely that they would fit any specific nonlinear/linear function with a constant term. Consequently, the proposed method naturally avoids the special case where some weights remain unchanged. Therefore, in the subsequent experiments, we set the values of the hyper-parameters γ_1 and γ_2 to zero.

Analysis of Comparison Results We compared the performance of the proposed SuaVE to some variants of VAE mentioned above. We used the same network architecture for encoder and decoder parts in all these models. In particular, we add a sub-network to model the conditional prior in both iVAE and the proposed SuaVE. We further assign a linear SCM sub-network to model the relations among latent causal variables in the proposed SuaVE. We trained these 5 models on the dataset described above, with different numbers
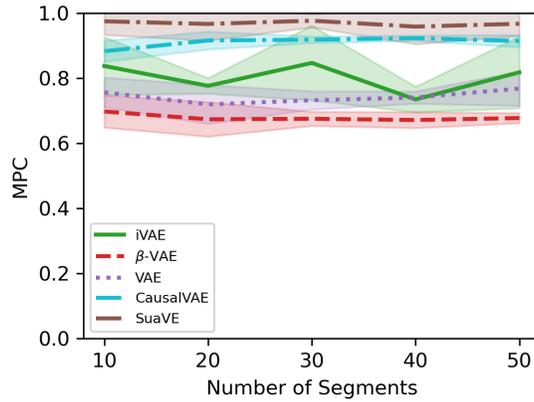
Figure 7: Performances of the proposed SuaVE in comparison to iVAE, β-VAE, VAE and CausalVAE in recovering the latent causal variables on synthetic data with different numbers of segments. Implementation details can be found in Appendix A.6.1.
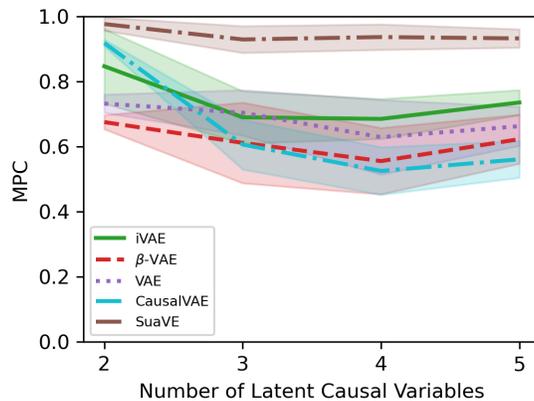

Figure 8: Performances of the proposed SuaVE in comparison to iVAE, β-VAE, VAE and CausalVAE in recovering the latent causal variables on synthetic data with different numbers of the latent causal variables. Details can be found in Appendix A.6.

of segments and different numbers of latent causal variables. For each method, we use 5 different random seeds for data sampling. Figure 7 shows the performance on two latent causal variables with different numbers of segment. The proposed SuaVE obtains the score 0.96 approximately for all the different numbers of segment. In contrast, β-VAE, VAE and CausalVAE fail to achieve a good estimation of the true latent variables, since they are not identifiable. iVAE obtains unsatisfactory results, since its theory only holds for i.i.d. latent variables. Figure 8 shows the performance in recovering latent causal variables on synthetic data with different numbers of the latent causal variables. Figure 9 shows the structural Hamming distance (SHD), a standard metric for comparing graphs via their
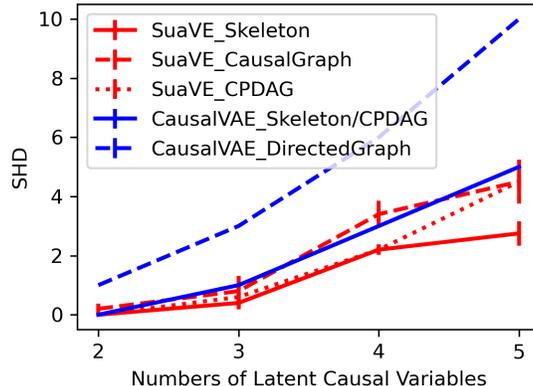
19

Figure 9: The SHD obtained by the proposed SuaVE and CausalVAE. Since CausalVAE is not identifiable, it obtains fully connected graphs, the SHD of the recovered CPDAG is the same as one of the recovered skeletons. Implementation details can be found in Appendix A.6.

adjacency matrices, for the recovered skeletons (the undirected version of the graph with all directed edges replaced by undirected edges), causal graphs, and completed partial directed acyclic graphs (CPDAGs, which capture the set of all DAGs that are Markov equivalent), by SuaVE and CausalVAE. Since CausalVAE is not identifiable, it obtains fully connected graphs, the SHD of the recovered CPDAG is the same as one of the recovered skeletons.

7.1.1 Experiments on Changes of Part of Weights

In this part, we conduct experiments on changes of part of weighs, to verify the results in Theorem 4. To this end, we consider two cases on the following causal graph: $z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z4$. The data details are similar to those of the 4-dimensional case mentioned in Appendix A.6.2, except that z_3 is a parent node of z_4. *Case 1:* the weight on the edge of $z_1 \rightarrow z_2$ (and accordingly, $z_2 \rightarrow z_3$, $z_3 \rightarrow z_4$) remains unchanged across $\mathbf{u}$, while the weights on the remaining edges change across $\mathbf{u}$, *Case 2:* the weight on the edge of $z_1 \rightarrow z_2$ (and accordingly, $z_2 \rightarrow z_3$, $z_3 \rightarrow z_4$) includes a constant part (*i.e.*, $\lambda_{1,2}(\mathbf{u}) + b$, b is a non-zero constant.) across $\mathbf{u}$, while the weights on the remaining edges change across $\mathbf{u}$. Again, the network architecture, optimization, and hyper-parameters are the same as used for experiments of Figure 7 (a). Figures 10 and 11 show the performance of the proposed method in the settings corresponding to *Case 1* and *Case 2* mentioned above. According to MPC, we observe that unchanged weights in *Case 1* or weights involving a constant term in *Case 2* lead to non-identifiability. Conversely, changing weights contribute to the identifiability of the corresponding nodes. These empirical results align with the partial identifiability results in Theorem 4.
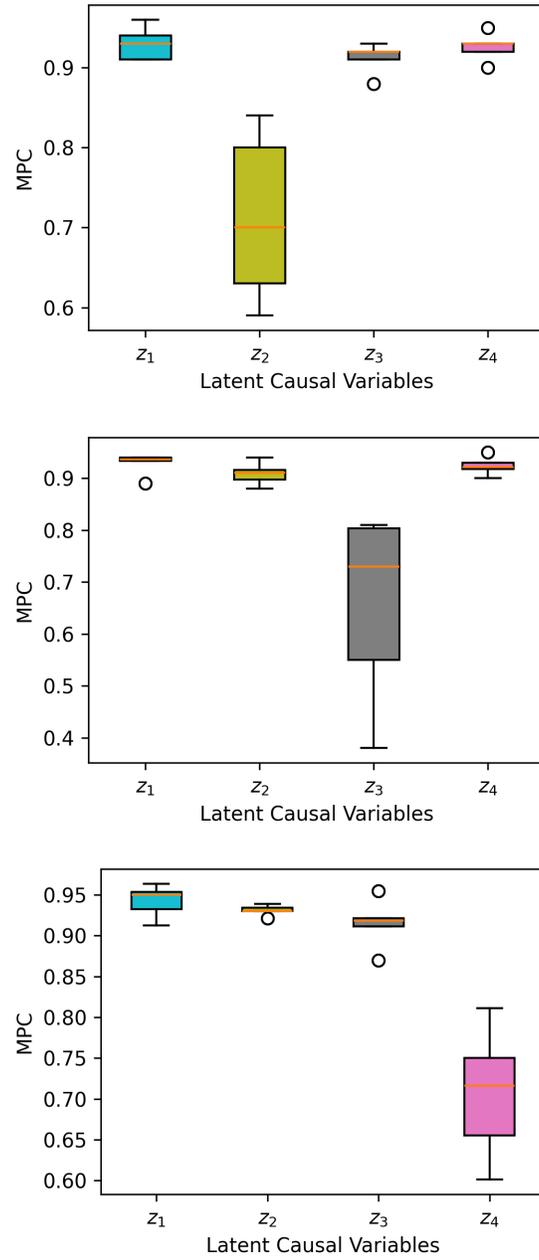
20

Figure 10: *Case 1:* Performances of the proposed method with the change of part of weights. The ground truth of the causal graph is $z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z_4$. From top to bottom: keeping weight on $z_1 \rightarrow z_2$ (and accordingly, $z_2 \rightarrow z_3$, and $z_3 \rightarrow z_4$) unchanged. Those results are consistent with the analysis of partial identifiability results in Theorem 4.
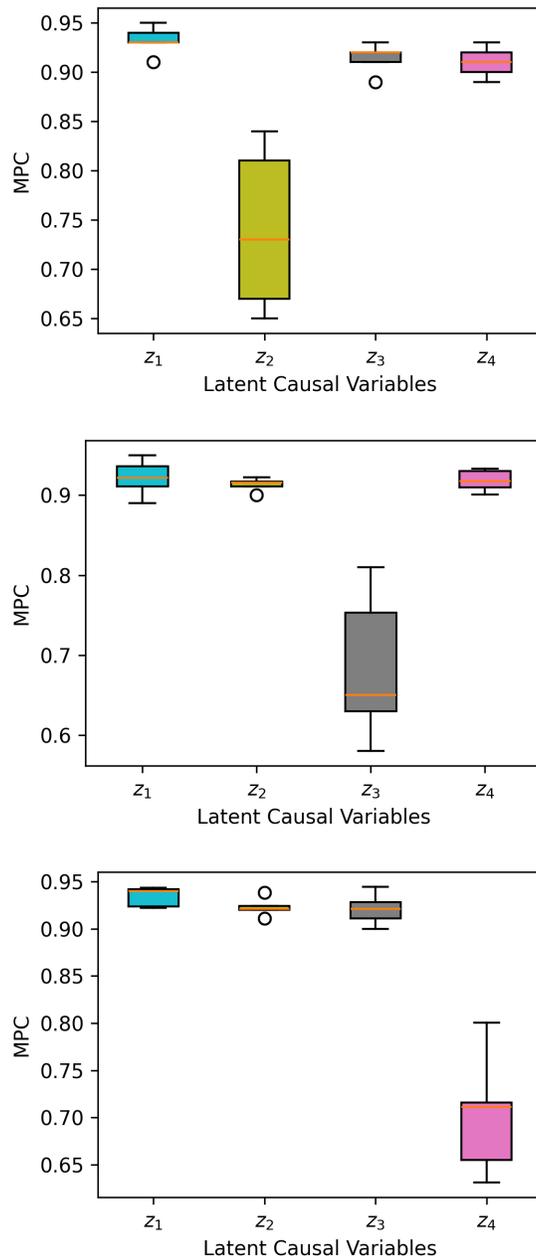
Figure 11: *Case 2:* Performances of the proposed method with the change of part of weights. The ground truth of the causal graph is $z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z_4$. From top to bottom: weight on $z_1 \rightarrow z_2$ (and accordingly, $z_2 \rightarrow z_3$, and $z_3 \rightarrow z_4$) includes a non-zero constant part and changing part, *i.e.*, $\lambda(\mathbf{u}) + b$. Those results are consistent with the analysis of partial identifiability results in Theorem 4.

7.1.2 EXPERIMENTS ON I.I.D. z_i

Our identifiability result includes nonlinear ICA as a special case, in which the latent causal variables do not exhibit any causal relationships among themselves, effectively collapsing the causal structure into a fully disconnected graph. To verify this point, we conduct experiments on the case where any two latent causal variables z_i are independent given $\mathbf{u}$. Data details are the same as mentioned in section A.6.2, but we here enforce $\lambda_{i,j}(\mathbf{u}) = 0$ for each i, j. Again, the network architecture, optimization and hyper-parameters are the same as section A.6.2, except for learning rate (here we set $1r = 1e - 2$). Figure 12 shows the performances of the proposed SuaVE and iVAE. We can see that iVAE is slightly better than the proposed SuaVE, both obtain satisfying results in terms of MPC.
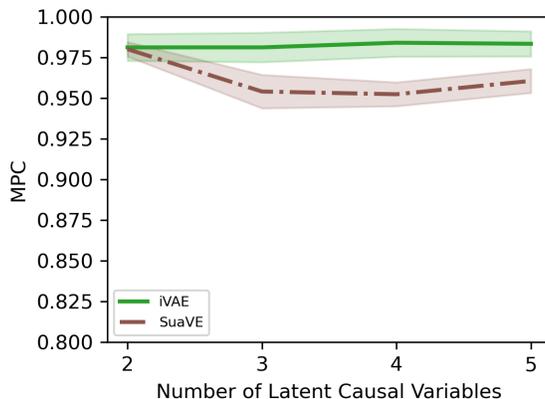

Figure 12: Performances of the proposed SuaVE and iVAE in recovering the independent latent causal variables conditional $\mathbf{u}$.

7.1.3 PERFORMANCE FOR VIOLATED ASSUMPTION ON DISTRIBUTION

To further understand the assumptions of our identifiability result, we conduct experiments in a setting where the assumption on distribution is violated. Table 1 shows the performance when the distribution assumption is violated. We observe that violating the distribution assumption leads to unsatisfactory performance in terms of MPC. The possible reasons for the results with Uniform (Laplace, Gamma) Noise may be: 1) the proposed method enforces Gaussian prior and posterior distributions to approximate non-Gaussian noise, and 2) the settings with Uniform (Laplace, Gamma) noise may be inherently non-identifiable.

7.2 Image Data

We further verify our identifiability results and the proposed method on high-dimensional image data from the chemistry dataset proposed in Ke et al. (2021), which is corresponding to simple chemical reactions where the state of an element can cause changes to another variable's state. The environment consists of a number of objects whose positions are kept fixed and thus, uniquely identifiable, while the colors (states) of the objects affected by

Table 1: Performance for violated assumptions. Uniform (Laplace, Gamma) Noise: the distribution of latent noise variables. Matching Assumptions: a setting that matches our theoretical assumptions.

Generative process	MPC(Mean $\pm$ Std)
Uniform Noise	0.83 ± 0.09
Laplace Noise	0.83 ± 0.03
Gamma Noise	0.82 ± 0.05
Matching Assumptions	0.95 ± 0.01

this variable (according to the causal graph) can change. We use a weight-variant linear Gaussian model as mentioned in section A.6.2 for a 3-dimensional case, to generate the latent variables, and obtain the corresponding observational images. A visualization of the environment can be found in Figure 13.
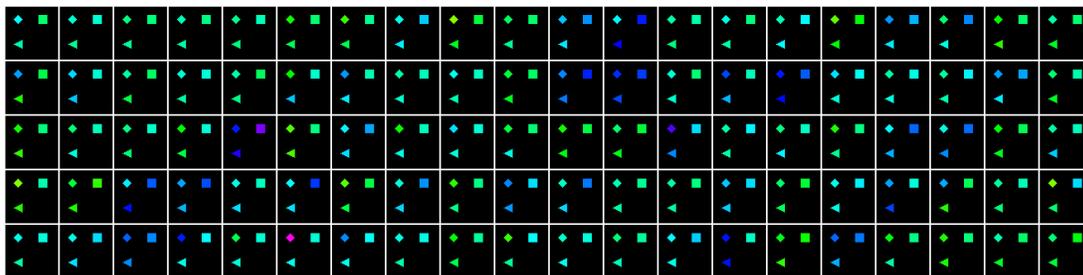

Figure 13: Demonstration of the chemistry environment. The colors (states) of the objects change according to the causal graph: the 'diamond' causes the 'triangle', and the 'triangle' causes the 'square'.

Figure 14 shows the MPC obtained by different methods. The proposed method performs better than β-VAE (Higgins et al., 2017) and CausalVAE (Yang et al., 2021). In term of MPC, the proposed method successfully learns causal representations, while the others fails to learn latent causal variables. To further verify whether the learned causal structure by the proposed method is consistent with the true underlying latent structure, we investigate the results of intervening on each learned variable. Figure 15 shows the intervention results on each learned latent variable by the proposed SuaVE. We can see from Figure 15 that:

- Intervention on z_1 (the 'diamond') cause the change of both z_2 and z_3 (the colors of 'triangle' and 'square'), as depicted in the left in Figure 15.

- Intervention on z_2 (the 'triangle') only cause the change of z_3 (the color of 'square'), in the medium in Figure 15.

- Intervention on the changes of z_3 (the 'square') *do not* cause the change of z_1 and z_2 (the 'diamond' and the 'triangle'), in the right in Figure 15

24

	z_1	z_2	z_3
$\hat{z}_1$	0.992	0.751	0.614
$\hat{z}_2$	0.741	0.990	0.840
$\hat{z}_3$	0.385	0.557	0.917

Ours

	z_1	z_2	z_3
$\hat{z}_1$	0.067	0.552	0.628
$\hat{z}_2$	0.508	0.265	0.346
$\hat{z}_3$	0.117	0.429	0.635

β-VAE

	z_1	z_2	z_3
$\hat{z}_1$	0.253	0.512	0.228
$\hat{z}_2$	0.598	0.265	0.046
$\hat{z}_3$	0.117	0.429	0.515

CausalVAE

Figure 14: MPC obtained by different methods on the image dataset. Supported by our identifiability analysis, the proposed method achieves satisfactory MPC, indicating its success in recovering latent causal representations.

All these observations are consistent with the ground truth latent causal structure, *i.e.*, $z_1 \to z_2 \to z_3$, together with results in Figure 14, demonstrating that the proposed method successfully learns latent causal structures. Similar, we inversigate the change in reconstructions while traversing each learned latent variable obtained by β-VAE (Higgins et al., 2017) and CausalVAE (Yang et al., 2021). Figure 16 and 17 show traversal results on the learned representations by β-VAE (Higgins et al., 2017) and CausalVAE (Yang et al., 2021), respectively. we observed that changing each learned variable results in changes to the colors of all objects. Together with the results in Figure 14, these suggests that β-VAE (Higgins et al., 2017) and CausalVAE (Yang et al., 2021) tend to learn representations that entangle the true underlying latent causal variables and fail to recover the true underlying latent causal structure.

7.3 fMRI Data

Following Ghassami et al. (2018), we applied the proposed method to fMRI hippocampus dataset (Laumann and Poldrack, 2015), which contains signals from six separate brain regions: perirhinal cortex (PRC), parahippocampal cortex (PHC), entorhinal cortex (ERC), subiculum (Sub), CA1, and CA3/Dentate Gyrus (DG) in the resting states on the same person in 84 successive days. Each day is considered as different $\mathbf{u}$, thus $\mathbf{u}$ is a 84-dimensional vector. Since we are interested in recovering latent causal variables, we treat the six signals as latent causal variables by applying a random nonlinear mapping on them to obtain observed data. We then apply various methods on the observed data to recover the six signals. It is worth noting that, with respect to Assumption (v), one of the key conditions in our theory, the context of fMRI might be interpreted as providing an approximate rather than a literal reference condition. For example, different cognitive tasks or resting states are known to substantially suppress or activate specific effective connections between brain regions. While such modulations may not constitute perfect hard interventions, they may weaken certain causal influences, serving as empirical analogues of the reference condition assumed in our theory. Our empirical results suggest that such approximate satisfaction of the assumption might already be sufficient in practice.

Figure 19 (a) shows the performance of the proposed SuaVE in comparison to iVAE, β-VAE, VAE and CausalVAE in recovering the latent six signals. β-VAE aims to recover
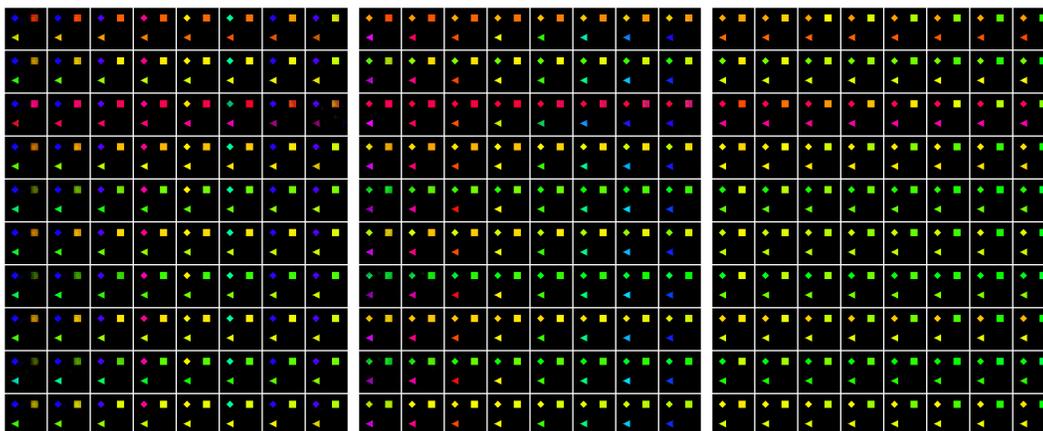
Figure 15: Intervention results on the learned latent variables by the proposed method SuaVE. From left to right: intervention on the learned z_1, z_2, z_3, respectively. The vertical axis denotes different samples, The horizontal axis denotes enforcing different values on the learned causal representation. We can see that: 1) (from the left) the changes of z_1 (the 'diamond') cause the change of both z_2 and z_3 (the 'triangle' and the 'square'). 2) (from the middle) the changes of z_2 (the 'triangle') only cause the change of z_3 (the 'square'). 3) (from the right) the changes of z_3 (the 'square') do not cause the change of z_1 and z_2 (the 'diamond' and the 'triangle').
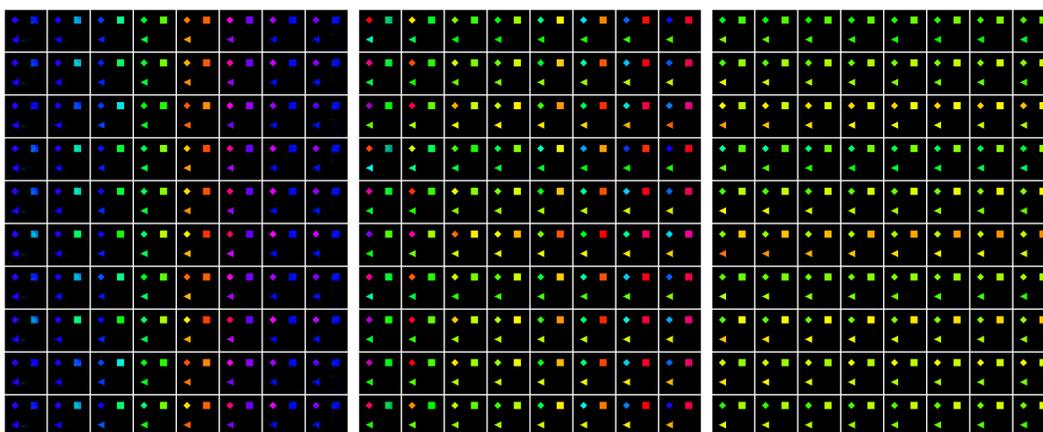
Figure 16: Traversal results on the learned latent variables by β-VAE (Higgins et al., 2017). From left to right: Traversal on the learned z_1, z_2, z_3, respectively. The vertical axis denotes different samples, The horizontal axis denotes enforcing different values on the learned causal representation. We can see that from the left to the right with changes of anyone z_i, the remaining two variables always changes, which means that each learned variable by β-VAE (Higgins et al., 2017) tend to be mixtures of the true underlying latent causal variables.
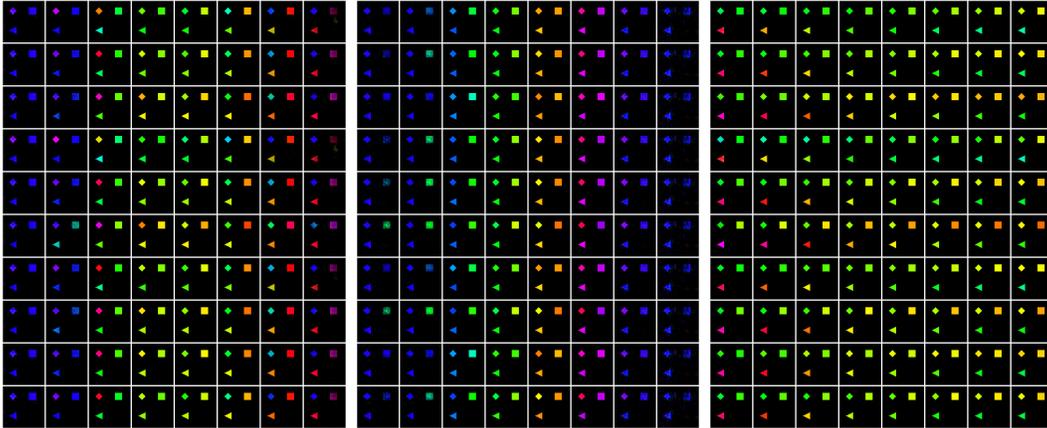
26

Figure 17: Traversal results on the learned latent variables by CausalVAE (Yang et al., 2021). From left to right: Traversal on the learned z_1, z_2, z_3, respectively. The vertical axis denotes different samples, The horizontal axis denotes enforcing different values on the learned causal representation. Again, we can see that from the left to the right with changes of anyone z_i, the remaining two latent variables also change. These results mean that each learned variable tend to be mixtures of the true underlying latent causal variables, and CausalVAE (Yang et al., 2021) fails to recovering latent causal structure.

independent latent variables, and it obtains an interesting result: enforcing independence (*e.g.*, $\beta = 25, 50$) leads to worse MPC, and relaxing it (though contracting to its own independence assumption) improves the result (*e.g.*, $\beta = 4$). This is because the latent variables given the time index are not independent in this dataset.

We further verify the ability of the proposed method to recover causal skeletons. We use the anatomical connections (Bird and Burgess, 2008) as a reference as shown in Figure 19 (b), since causal connections among the six regions should not exist if there is no anatomical connection. To obtain the final estimated graph, note that in the computing MPC, we have solved a linear sum assignment problem, which matches the semantic information of the estimated variables with the one of the true underlying latent variables. After matching all semantic information, we can use a threshold (0.1) to removed a edge separately across **u**, and once an edge appears in more than 70% of all **u**, we keep this edge in the final graph. Figure 19 (c) shows an example of the estimated causal skeleton by the proposed SuaVE. Averaging over 5 different random seeds for the used nonlinear mapping from latent space to observed space, structural Hamming distance (SHD) obtained by the proposed SuaVE is $\mathbf{5.0 \pm 0.28}$. In contrast, iVAE, β-VAE and VAE assume latent variables to be independent, and thus can not obtain the causal skeleton. We further analyse the result of β-VAE with $\beta = 4$ by using the PC algorithm (Spirtes et al., 2001) to discovery skeleton on the recovered latent variables, and obtain $\mathbf{SHD = 5.8 \pm 0.91}$. This means that: 1) β-VAE with $\beta = 4$ does not ensure the strong independence among the recover latent variables as it is expected, 2) the proposed SuaVE is an effective one-step method to discovery the

skeleton. The unsupervised version of CausalVAE is not identifiable, and can also not ensure the relations among the recover variables to be causal skeleton in principle. In experiments we found that CausalVAE always obtains fully-connected graphs for the 5 different random seeds, for which SHD is $\mathbf{9.0 \pm 0}$.
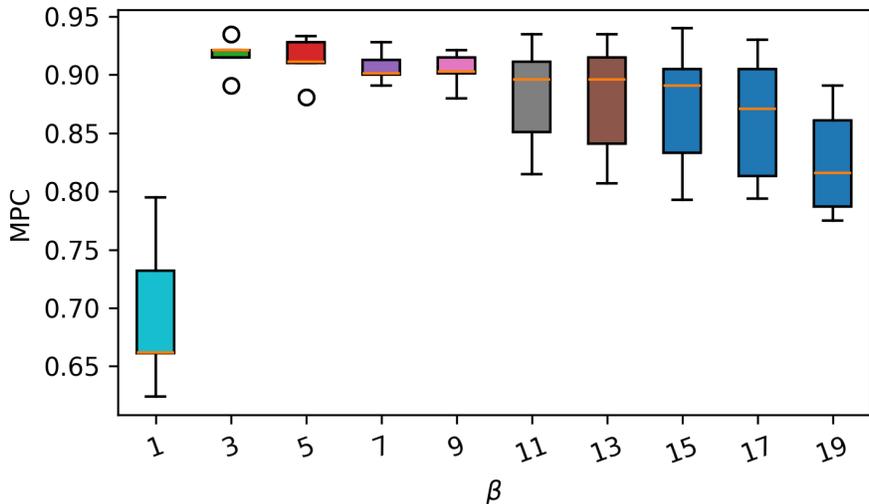

Figure 18: Performances of β-VAE with different β in recovering fMRI data.

We further investigate the performance of β-VAE across different values of β. Figure 18 illustrates the performance trend as β varies. β-VAE suffers from two critical limitations: In practical applications, determining an appropriate value for β is challenging due to the inherent uncertainty in the underlying latent variables. This lack of guidance can lead to suboptimal model performance if β is not carefully tuned. For $\beta \geq 1$, as the value of β increases, there is an inevitable trade-off between disentanglement and reconstruction quality. Specifically, improving disentanglement typically comes at the cost of degraded reconstruction, as compared to a standard VAE. This trade-off is particularly problematic in scenarios where preserving reconstruction quality is crucial, making it difficult to balance these competing objectives in real-world tasks. These challenges highlight the practical difficulties in applying β-VAE effectively, especially when considering applications where the precise nature of the latent structure is unknown and both disentanglement and reconstruction are essential.

Figure 20 (a) show the edges in the anatomical ground truth reported in Ghassami et al. (2018). Figure 20 (b) show the latent recovered edges by the proposed SuaVE, in the setting where we treat fMRI data as latent causal variables. For comparison, Figure 20 (c) shows the recovered causal graph reported in (Ghassami et al., 2018), in the setting where fMRI data are treated as observed variables. That is, the setting in (b) is more challenging than the setting in (c).
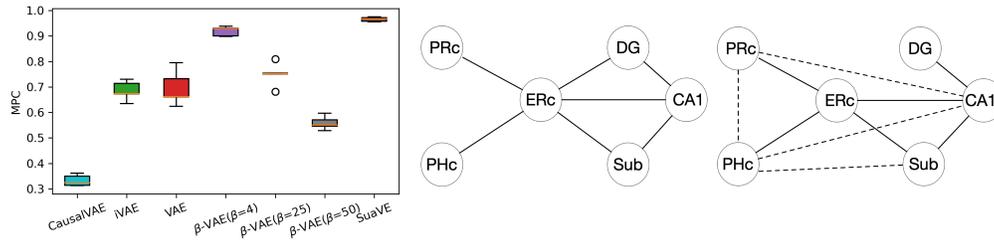
Figure 19: Performance on fMRI Data. (a) the performance of the proposed SuaVE in comparison to iVAE, β-VAE, VAE and CausalVAE in recovering the latent six signals. (b) The skeleton of the anatomical connections given in Bird and Burgess (2008). (c) The recovered skeleton by the proposed SuaVE, where the dashed lines indicate errors.
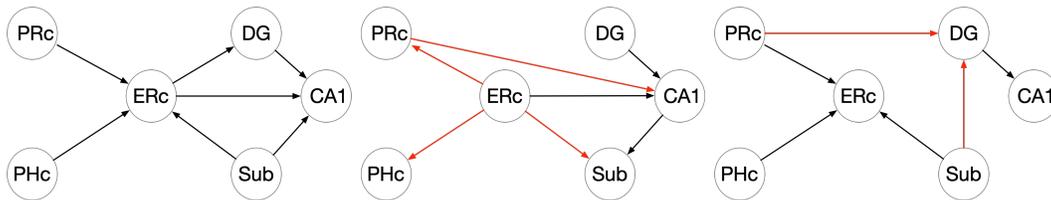

Figure 20: Recovered graphs on fMRI Data. (a) the anatomical connections given in Bird and Burgess (2008). (b) The recovered causal graph by the proposed SuaVE. (c) The recovered causal graph reported in (Ghassami et al., 2018). Note that the setting in (b) treats fMRI data as latent variables, while the setting in (c) treats fMRI data as observed variables.

8. Conclusion

Identifying latent causal variables from observed data is generally impossible without additional assumptions. Motivated by recent advances in the identifiability of nonlinear ICA, we investigate three fundamental sources of indeterminacy in the latent space, *i.e.*, transitivity, permutation, and scaling, which offer key insights into the challenges and opportunities in identifying latent causal variables. Among these, we show transitivity as the primary obstacle to identifiability and propose a new research direction that leverages changes in causal influences between latent variables to overcome it. Under weight-variant linear Gaussian models and mild conditions inspired by nonlinear ICA, we prove that latent causal variables can be identified up to permutation and scaling. Importantly, we show that partial identifiability remains possible even when only a subset of the causal influences varies, making our results more applicable in practice. Building on these theoretical insights, we introduce Structural caUsAl Variational autoEncoder (SuaVE), a novel method for jointly learning latent causal variables and their underlying causal graph. Empirical results on both synthetic and fMRI data validate our identifiability theory and demonstrate the practical effectiveness of the proposed method.

Acknowledgments

Appendix A.

A.1 The proof of Theorem 1

The proof of Theorem 1 is done in three steps. Step I is to show the identifiability result of nonlinear ICA holds in our setting where there is an additional mapping from $\mathbf{n}$ to $\mathbf{z}$, and the mapping changes across $\mathbf{u}$. Step II consists of building a linear transformation between the true latent variables $\mathbf{z}$ and the estimated ones $\hat{\mathbf{z}}$. Step III shows that the linear transformation in Step II can be reduced to permutation transformation by using the result of Step I and the assumption (v).

For convenience, let us discuss the property of the mapping from $\mathbf{n}$ to $\mathbf{z}$ first, which plays a key role in the proof. Let function $\mathbf{g_u}$ denote the mapping from $\mathbf{n}$ to $\mathbf{z}$. That is:

$$\mathbf{z} = \mathbf{g_u}(\mathbf{n}) = (\mathbf{I} - \boldsymbol{\lambda}^T(\mathbf{u}))^{-1}\mathbf{n}, \tag{14}$$

where $\mathbf{I}$ denotes the identity matrix. Due to DAG constraints in a causal system and the assumption of linear models on $\mathbf{z}$, we have 1) the determinant of the Jacobian matrix is equal to 1, *i.e.*, $|\det \mathbf{J}_{\mathbf{g_u}^{-1}}(\mathbf{z})| = |\det \mathbf{J}_{\mathbf{g_u}}(\mathbf{u})| = 1$, 2) the mapping $\mathbf{g_n}$ is invertible.

Step I: Suppose we have two sets of parameters $\theta = (\mathbf{f}, \mathbf{T_n}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\varepsilon})$ and $\hat{\theta} = (\hat{\mathbf{f}}, \hat{\mathbf{T}}_{\mathbf{n}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}})$ such that $p_{(\mathbf{f}, \mathbf{T_n}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\varepsilon})}(\mathbf{x}|\mathbf{u}) = p_{(\hat{\mathbf{f}}, \hat{\mathbf{T}}_{\mathbf{n}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}})}(\mathbf{x}|\mathbf{u})$ for all pairs $(\mathbf{x}, \mathbf{u})$. Due to the assumptions (i) and (ii), by expanding these expressions via the change of variables formula and taking the logarithm we find:

$$\log|\mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) + \log|\mathbf{J}_{\mathbf{g_u}^{-1}}(\mathbf{z})| + \log p_{(\mathbf{T_n}, \boldsymbol{\beta})}(\mathbf{n}|\mathbf{u})$$
$$= \log|\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \log|\mathbf{J}_{(\hat{\mathbf{g}}_{\mathbf{u}})^{-1}}(\hat{\mathbf{z}})| + \log p_{(\hat{\mathbf{T}}_{\mathbf{n}}, \hat{\boldsymbol{\beta}})}(\hat{\mathbf{n}}|\mathbf{u}), \tag{15}$$

Using the exponential family Eq. (4) to replace $p_{(\mathbf{T_n}, \boldsymbol{\beta})}(\mathbf{n}|\mathbf{u})$, we have:

$$\log|\mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) + \log|\mathbf{J}_{\mathbf{g_u}^{-1}}(\mathbf{z})| + \mathbf{T}_{\mathbf{n}}^T(\mathbf{n})\boldsymbol{\eta}_{\mathbf{n}}(\mathbf{u}) - \log Z_{\mathbf{n}}(\boldsymbol{\beta}, \mathbf{u}) = \tag{16}$$

$$\log|\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \log|\mathbf{J}_{(\hat{\mathbf{g}}_{\mathbf{u}})^{-1}}(\hat{\mathbf{z}})| + \hat{\mathbf{T}}_{\mathbf{n}}^T(\hat{\mathbf{n}})\hat{\boldsymbol{\eta}}_{\mathbf{n}}(\mathbf{u}) - \log \hat{Z}_{\mathbf{n}}(\hat{\boldsymbol{\beta}}, \mathbf{u}), \tag{17}$$

then by using

$$|\det \mathbf{J}_{\mathbf{g_u}^{-1}}(\mathbf{z})| = 1, |\det \mathbf{J}_{\hat{\mathbf{g}}_{\mathbf{u}}^{-1}}(\mathbf{z})| = 1. \tag{18}$$

here the latter holds since $\hat{g}_u$ must be the same function class as g_u, Eqs. 16 and 17 can be rewriteed as:

$$\log|\mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) + \mathbf{T}_{\mathbf{n}}^T(\mathbf{n})\boldsymbol{\eta}_{\mathbf{n}}(\mathbf{u}) - \log Z_{\mathbf{n}}(\boldsymbol{\beta}, \mathbf{u}) = \tag{19}$$

$$\log|\mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \hat{\mathbf{T}}_{\mathbf{n}}^T(\hat{\mathbf{n}})\hat{\boldsymbol{\eta}}_{\mathbf{n}}(\mathbf{u}) - \log \hat{Z}_{\mathbf{n}}(\hat{\boldsymbol{\beta}}, \mathbf{u}). \tag{20}$$

The following proof is similar to the proof of Sorrenson et al. (2020). For completeness, we present a slightly simplified proof. With different points in assumption (iii), we can subtract

this expression with $\mathbf{u_{n,0}}$ by the expression with $l_\mathbf{n}$, $l_\mathbf{n} = 1, ..., 2n$. Since the Jacobian and noise terms do not depend on $\mathbf{u}$, the Jacobian and noise terms will be canceled out:

$$\log \frac{Z_\mathbf{n}(\boldsymbol{\beta}, \mathbf{u_{n,0}})}{Z_\mathbf{n}(\boldsymbol{\beta}, \mathbf{u_{n,l_n}})} + \mathbf{T}_\mathbf{n}^T(\mathbf{n})\big(\boldsymbol{\eta}_\mathbf{n}(\mathbf{u_{n,l_n}}) - \boldsymbol{\eta}_\mathbf{n}(\mathbf{u_{n,0}})\big) = \tag{21}$$

$$\log \frac{\hat{Z}_\mathbf{n}(\hat{\boldsymbol{\beta}}, \mathbf{u_{n,0}})}{\hat{Z}_\mathbf{n}(\hat{\boldsymbol{\beta}}, \mathbf{u_{n,l_n}})} + \hat{\mathbf{T}}_\mathbf{n}^T(\hat{\mathbf{n}})\big(\hat{\boldsymbol{\eta}}_\mathbf{n}(\mathbf{u_{n,l_n}}) - \hat{\boldsymbol{\eta}}_\mathbf{n}(\mathbf{u_{n,0}})\big). \tag{22}$$

Combining the $2n$ expressions into a single matrix equation we can write it in terms of $\mathbf{L_n}$ from assumption (iii):

$$\mathbf{L}_\mathbf{n}^T \mathbf{T}_\mathbf{n}(\mathbf{n}) = \hat{\mathbf{L}}_\mathbf{n}^T \hat{\mathbf{T}}_\mathbf{n}(\hat{\mathbf{n}}) + \mathbf{b_n}, \tag{23}$$

where $\mathbf{L_n}$ is defined as the assumption (iii), $\mathbf{b}_{\mathbf{n},l_n} = \log \frac{\hat{Z}_\mathbf{n}(\hat{\boldsymbol{\beta}}, \mathbf{u_{n,0}}) Z_\mathbf{n}(\boldsymbol{\beta}, \mathbf{u_{n,l_n}})}{\hat{Z}_\mathbf{n}(\hat{\boldsymbol{\beta}}, \mathbf{u_{n,l_n}}) Z_\mathbf{n}(\boldsymbol{\beta}, \mathbf{u_{n,0}})}$. Since $\mathbf{L}_\mathbf{n}^T$ is invertible by the assumption (iii), we can multiply this expression by its inverse:

$$\mathbf{T}_\mathbf{n}(\mathbf{n}) = \mathbf{A}_\mathbf{n} \hat{\mathbf{T}}_\mathbf{n}(\hat{\mathbf{n}}) + \mathbf{c_n}, \tag{24}$$

According to lemma 3 in Khemakhem et al. (2020) and Proof for Theorem 1 in Sorrenson et al. (2020) we can show that $\mathbf{A_n}$ has full rank. Since we assume the noise to be Gaussian, Eq. (24) can be re-expressed as:

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{n}^2 \end{pmatrix} = \mathbf{A}_\mathbf{n} \begin{pmatrix} \hat{\mathbf{n}} \\ \hat{\mathbf{n}}^2 \end{pmatrix} + \mathbf{c_n}, \tag{25}$$

Then, by the contradiction between 1) for every n_i, there exists a polynomial with a degree at most 2, and 2) for every n_i^2, there also exists a polynomial with a degree at most 2 (Sorrenson et al., 2020), we can obtain: $n_i = A_{i,j} \hat{n}_j + c_i$. For simplicity, in the following we neglect the noise term $\boldsymbol{\varepsilon}$ in Eq (3). As a result, we can express the result in vector form as:

$$\mathbf{n} = \mathbf{P}\hat{\mathbf{n}} + \mathbf{c_n}, \tag{26}$$

where $\mathbf{P}$ denote the permutation matrix with scaling. More details can be found in Sorrenson et al. (2020). Note that this simplification is for convenience only, the identifiability result still holds even when the noise term is included.

Step II: Again, suppose we have two sets of parameters $\theta = (\mathbf{f}, \mathbf{T_z}, \boldsymbol{\lambda}, \boldsymbol{\beta})$ and $\hat{\theta} = (\hat{\mathbf{f}}, \hat{\mathbf{T}}_\mathbf{z}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}})$ such that $p_{(\mathbf{f}, \mathbf{T_z}, \boldsymbol{\lambda}, \boldsymbol{\beta})}(\mathbf{x}|\mathbf{u}) = p_{(\hat{\mathbf{f}}, \hat{\mathbf{T}}_\mathbf{z}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}})}(\mathbf{x}|\mathbf{u})$ for all pairs $(\mathbf{x}, \mathbf{u})$. Due to the assumptions (i) and (ii), by expanding these expressions via the change of variables formula and taking the logarithm we have:

$$\log |\det \mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \log p_{(\mathbf{T_z}, \boldsymbol{\lambda}, \boldsymbol{\beta})}(\mathbf{z}|\mathbf{u}) + \log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) = \log |\det \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \log p_{(\hat{\mathbf{T}}_\mathbf{z}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}})}(\hat{\mathbf{z}}|\mathbf{u}), \tag{27}$$

Using the exponential family Eq. (7) to replace $p_{(\mathbf{T_z}, \boldsymbol{\lambda}, \boldsymbol{\beta})}(\mathbf{z}|\mathbf{u})$:

$$\log |\det \mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \mathbf{T}_\mathbf{z}^T(\mathbf{z}) \boldsymbol{\eta}_\mathbf{z}(\mathbf{u}) - \log Z_\mathbf{z}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u}) + \log p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) = \tag{28}$$

$$\log |\det \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \hat{\mathbf{T}}_\mathbf{z}^T(\hat{\mathbf{z}}) \hat{\boldsymbol{\eta}}_\mathbf{z}(\mathbf{u}) - \log \hat{Z}_\mathbf{z}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \mathbf{u}). \tag{29}$$

Let $\mathbf{u}_{\mathbf{z},0}, \mathbf{u}_{\mathbf{z},1}, ..., \mathbf{u}_{\mathbf{z},k}$ be the points provided by the assumption (iv). We define $\bar{\boldsymbol{\eta}}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}) = \boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}) - \boldsymbol{\eta}_{\mathbf{z},}(\mathbf{u}_{\mathbf{z},0})$. We plug each of those $\mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}, l_{\mathbf{z}} = 1, ..., k$ in Eq. (29) to obtain $k + 1$ equations. We subtract the first equation for $\mathbf{u}_{\mathbf{z},0}$ from the remaining k equations. For example, for $\mathbf{u}_{\mathbf{z},0}$ and $\mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}$ we have two equations:

$$\log|\det \mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \mathbf{T}_{\mathbf{z}}^{T}(\mathbf{z})\boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},0}) - \log Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u}_{\mathbf{z},0}) + \log p_{\varepsilon}(\boldsymbol{\varepsilon}) =$$
$$\log|\det \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \hat{\mathbf{T}}_{\mathbf{z}}^{T}(\hat{\mathbf{z}})\hat{\boldsymbol{\eta}}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},0}) - \log \hat{Z}_{\mathbf{z}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \mathbf{u}_{\mathbf{z},0}), \tag{30a}$$

$$\log|\det \mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \mathbf{T}_{\mathbf{z}}^{T}(\mathbf{z})\boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}) - \log Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}) + \log p_{\varepsilon}(\boldsymbol{\varepsilon}) =$$
$$\log|\det \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \hat{\mathbf{T}}_{\mathbf{z}}^{T}(\hat{\mathbf{z}})\hat{\boldsymbol{\eta}}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}) - \log \hat{Z}_{\mathbf{z}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \mathbf{u}_{\mathbf{z},l_{\mathbf{z}}}), \tag{30b}$$

Using Eq. (30b) subtracts Eq. (30a), canceling the terms that do not include $\mathbf{u}$, we have:

$$\log \frac{Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u}_{\mathbf{z},0})}{Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u}_{\mathbf{z},l_{\mathbf{z}}})} + \mathbf{T}_{\mathbf{z}}(\mathbf{z})\big(\boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},l_{\mathbf{n}}}) - \boldsymbol{\eta}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},0})\big)$$
$$= \log \frac{\hat{Z}_{\mathbf{z}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \mathbf{u}_{\mathbf{z},0})}{\hat{Z}_{\mathbf{z}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \mathbf{u}_{\mathbf{z},l_{\mathbf{z}}})} + \hat{\mathbf{T}}_{\mathbf{z}}(\hat{\mathbf{z}})\big(\hat{\boldsymbol{\eta}}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},l_{\mathbf{n}}}) - \hat{\boldsymbol{\eta}}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z},0})\big). \tag{31}$$

Let $\mathbf{L}_{\mathbf{z}}$ be the matrix defined in Eq. (9) in assumption (iv), and $\hat{\mathbf{L}}_{\hat{\mathbf{z}}}$ similarly ($\hat{\mathbf{L}}_{\hat{\mathbf{z}}}$ is not necessarily invertible). Define $b_{\mathbf{z},l_{\mathbf{z}}} = \log \frac{\hat{Z}_{\mathbf{z}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \mathbf{u}_{\mathbf{z},0})Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u}_{\mathbf{z},l_{\mathbf{z}}})}{Z_{\mathbf{z}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{u}_{\mathbf{z},0})\hat{Z}_{\mathbf{z}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \mathbf{u}_{\mathbf{z},l_{\mathbf{z}}})}$. Expressing Eq. (31) for all points in matrix form, we get:

$$\mathbf{L}_{\mathbf{z}}^{T}\mathbf{T}_{\mathbf{z}}(\mathbf{z}) = \hat{\mathbf{L}}_{\mathbf{z}}^{T}\hat{\mathbf{T}}_{\mathbf{z}}(\hat{\mathbf{z}}) + \mathbf{b}_{\mathbf{z}}. \tag{32}$$

We multiply both sides of Eq. (32) by the inverse of $\mathbf{L}_{\mathbf{z}}^{T}$ (by assumption (iv)) to find:

$$\mathbf{T}_{\mathbf{z}}(\mathbf{z}) = \mathbf{A}_{\mathbf{z}}\hat{\mathbf{T}}_{\mathbf{z}}(\hat{\mathbf{z}}) + \mathbf{c}_{\mathbf{z}}, \tag{33}$$

where $\mathbf{A}_{\mathbf{z}} = (\mathbf{L}_{\mathbf{z}}^{T})^{-1}\hat{\mathbf{L}}_{\hat{\mathbf{z}}}^{T}$ and $\mathbf{c}_{\mathbf{z}} = (\mathbf{L}_{\mathbf{z}}^{T})^{-1}\mathbf{b}_{\mathbf{z}}$. As mentioned in Eq. (7), the sufficient statistic $\mathbf{T}_{\mathbf{z}}(\mathbf{z}) = [\mathbf{z}; \text{vec}(\mathbf{z}\mathbf{z}^{T})]$. In this case, the relationship Eq. (33) becomes:

$$\begin{pmatrix} \mathbf{z} \\ \mathbf{z}^{2} \\ \mathbf{z}_{i \neq j} \end{pmatrix} = \mathbf{A}_{\mathbf{z}} \begin{pmatrix} \hat{\mathbf{z}} \\ \hat{\mathbf{z}}^{2} \\ \hat{\mathbf{z}}_{i \neq j} \end{pmatrix} + \mathbf{c}_{\mathbf{z}}, \tag{34}$$

where $\mathbf{z}$ denotes $[z_{1}, ..., z_{i}]$, $\mathbf{z}^{2}$ denotes $[z_{1}^{2}, ..., z_{i}^{2}]$, $\mathbf{z}_{i \neq j}$ denotes the vector whose elements are $z_{i}z_{j}$ for all $i \neq j$, $\mathbf{A}_{\mathbf{z}}$ in block matrix can be rewritten as:

$$\mathbf{A}_{\mathbf{z}} = \begin{pmatrix} \mathbf{A}^{(1)} & \mathbf{A}^{(2)} & \mathbf{A}^{(3)} \\ \mathbf{A}^{(4)} & \mathbf{A}^{(5)} & \mathbf{A}^{(6)} \\ \mathbf{A}^{(7)} & \mathbf{A}^{(8)} & \mathbf{A}^{(9)} \end{pmatrix} \tag{35}$$

and $\mathbf{c}_{\mathbf{z}}$ as:

$$\mathbf{c}_{\mathbf{z}} = \begin{pmatrix} \mathbf{c}^{(1)} \\ \mathbf{c}^{(2)} \\ \mathbf{c}^{(3)} \end{pmatrix}. \tag{36}$$

Then, we have:

$$\mathbf{z} = \mathbf{A}^{(1)}\hat{\mathbf{z}} + \mathbf{A}^{(2)}\hat{\mathbf{z}}^2 + \mathbf{A}^{(3)}\hat{\mathbf{z}}_{i \neq j} + \mathbf{c}^{(1)}, \tag{37}$$

$$\mathbf{z}^2 = \mathbf{A}^{(4)}\hat{\mathbf{z}} + \mathbf{A}^{(5)}\hat{\mathbf{z}}^2 + \mathbf{A}^{(6)}\hat{\mathbf{z}}_{i \neq j} + \mathbf{c}^{(2)}. \tag{38}$$

So we can write for each z_i:

$$z_i = \sum_j (A_{i,j}^{(1)}\hat{z}_j) + \sum_j (A_{i,j}^{(2)}\hat{z}_j^2) + (\mathbf{A}_{i,:}^{(3)}\hat{\mathbf{z}}_{i \neq j}) + c_i^{(1)}, \tag{39}$$

$$z_i^2 = \sum_j (A_{i,j}^{(4)}\hat{z}_j) + \sum_j (A_{i,j}^{(5)}\hat{z}_j^2) + (\mathbf{A}_{i,:}^{(6)}\hat{\mathbf{z}}_{i \neq j}) + c_i^{(2)}. \tag{40}$$

Squaring Eq. (39), we have:

$$z_i^2 = \underbrace{(\sum_j (A_{i,j}^{(2)}\hat{z}_j^2))^2}_{(\mathbf{a})} + \underbrace{(\sum_j (A_{i,j}^{(1)}\hat{z}_j))^2}_{(\mathbf{b})} + \underbrace{(\mathbf{A}_{i,:}^{(3)}\hat{\mathbf{z}}_{i \neq j})^2}_{(c)} + (c_i^{(1)})^2 + \ldots.. \tag{41}$$

It is notable that Eq. (40) and Eq. (41) are derived from Eq. (33) which holds for arbitrary $\mathbf{x}$, then by the assumption (ii) in Theorem 1 that $\mathbf{f}$ is a bijective mapping from $\mathbf{z}$ to $\mathbf{x}$, and thus Eq. (40) and Eq. (41) must holds for $\mathbf{z}$ everywhere. Then by the fact that the right sides of the Eq. (40) and Eq. (41) are both polynomials with finite degree, we have each coefficients of the two polynomials must be equal. In more detail, for the term (a) in Eq. (41):

$$(\sum_j (A_{i,j}^{(2)}\hat{z}_j^2))^2 = \sum_j (A_{i,j}^{(2)})^2 \hat{z}_j^4 + \sum_{j \neq j'} (2A_{i,j}^{(2)} A_{i,j'}^{(2)} \hat{z}_j^2 \hat{z}_{j'}^2). \tag{42}$$

Compared with Eq. (40), since there is no term $\hat{z}_j^4$ in Eq. (40), we must have that:

$$\mathbf{A}^{(2)} = 0. \tag{43}$$

For the term (c) in Eq. (41):

$$(\mathbf{A}_{i,:}^{(3)}\hat{\mathbf{z}}_{i \neq j})^2 = \sum_{j', i \neq j} (A_{i,j'}^{(3)})^2 (\hat{z}_i \hat{z}_j)^2 + \ldots \tag{44}$$

Compared with Eq. (40), since there is no term $(\hat{z}_i \hat{z}_j)^2$ in Eq. (40), we must have that:

$$\mathbf{A}^{(3)} = 0 \tag{45}$$

As a result, Eq. (37) becomes:

$$\mathbf{z} = \mathbf{A}^{(1)}\hat{\mathbf{z}} + \mathbf{c}^{(1)}. \tag{46}$$

The above equation indicates the latent causal variables can be recovered up to linear transformation.

Step III We then show that the linear transformation matrix $\mathbf{A}^{(1)}$ in Eq. (46) must be a permutation matrix. As we mentioned in step I, for the Gaussian noise variables $\mathbf{n}$, we have:

$$\mathbf{n} = \mathbf{P}\hat{\mathbf{n}} + \mathbf{c}_n, \tag{47}$$

where $\mathbf{P}$ denote the permutation matrix with scaling. With Eq. (47) and Eq. (14), the Eq. (46) can be rewritten as follows:

$$\mathbf{B}(\mathbf{P}\hat{\mathbf{n}} + \mathbf{c}_n) = \mathbf{A}^{(1)}(\hat{\mathbf{B}}\hat{\mathbf{n}}) + \mathbf{c}^{(1)}. \tag{48}$$

$$\Rightarrow (\mathbf{B}\mathbf{P} - \mathbf{A}^{(1)}\hat{\mathbf{B}})\hat{\mathbf{n}} = \mathbf{c}^{(1)} - \mathbf{B}\mathbf{c}_n, \tag{49}$$

where $\mathbf{B} = (\mathbf{I} - \boldsymbol{\lambda}^T(\mathbf{u}))^{-1}$ and $\hat{\mathbf{B}} = (\mathbf{I} - \hat{\boldsymbol{\lambda}}^T(\mathbf{u}))^{-1}$. By differentiating Eq. (49) with respect to $\hat{\mathbf{n}}$, we have:

$$(\mathbf{B}\mathbf{P} - \mathbf{A}^{(1)}\hat{\mathbf{B}})\mathbf{I} = \mathbf{0}, \tag{50}$$

where $\mathbf{I}$ denote the identity matrix, which implies that:

$$\mathbf{B}\mathbf{P} = \mathbf{A}^{(1)}\hat{\mathbf{B}}, \tag{51}$$

that is,

$$(\mathbf{I} - \boldsymbol{\lambda}^T(\mathbf{u}))^{-1}\mathbf{P} = \mathbf{A}(\mathbf{I} - \hat{\boldsymbol{\lambda}}^T(\mathbf{u}))^{-1}, \tag{52}$$

for convenience, we use $\mathbf{A}$ to replace $\mathbf{A}^{(1)}$ here and in the following. By inverting both sides of the equation, we have

$$\mathbf{P}^{-1}(\mathbf{I} - \boldsymbol{\lambda}^T(\mathbf{u})) = (\mathbf{I} - \hat{\boldsymbol{\lambda}}^T(\mathbf{u}))\mathbf{A}^{-1}. \tag{53}$$

Without loss of generality, suppose that $\boldsymbol{\lambda}^T$ and $\hat{\boldsymbol{\lambda}}^T$ denote lower triangular matrices, respectively.

Elements above the diagonal We can see from Eq. (53) that since the left is a lower triangular matrix, and $(\mathbf{I} - \hat{\boldsymbol{\lambda}}^T(\mathbf{u}))$ in the right is also a lower triangular matrix, $\mathbf{A}^{-1}$ must be a lower triangular matrix. That is, elements above the diagonal in matrix $\mathbf{A}^{-1}$ are equal to 0.

The diagonal elements Assume $\mathbf{P}^{-1}$ to be diagonal with elements $s_{1,1}, s_{2,2}, s_{3,3}, \ldots$. By expanding the both sides of Eq. (53), we have:

$$\begin{pmatrix} s_{1,1} & 0 & 0 & \ldots \\ -s_{2,2}\lambda_{1,2}(\mathbf{u}) & s_{2,2} & 0 & \ldots \\ -s_{3,3}\lambda_{1,3}(\mathbf{u}) & -s_{3,3}\lambda_{2,3}(\mathbf{u}) & s_{3,3} & \ldots \\ . & . & . & \ldots \end{pmatrix}$$
$$= \begin{pmatrix} A_{1,1} & 0 & 0 & \ldots \\ A_{2,1} - A_{1,1}\hat{\lambda}_{1,2}(\mathbf{u}) & A_{2,2} & 0 & \ldots \\ A_{3,1} - \sum_{i=1}^{2} A_{i,1}\hat{\lambda}_{i,3}(\mathbf{u}) & A_{3,2} - A_{2,2}\hat{\lambda}_{2,3}(\mathbf{u}) & A_{3,3} & \ldots \\ . & . & . & \ldots \end{pmatrix}. \tag{54}$$

Comparing the left and the right of Eq. (54), we can see that the diagonal elements of $\mathbf{A}^{-1}$ must be $s_{1,1}, s_{2,2}, s_{3,3}, \ldots$.

Elements below the diagonal By comparison between the left and right of (54), for each element $A_{i,j}$ where $i > j$ in the right of Eq. (54), we have: $A_{i,j} - \sum_{i'<i} \mathbf{A}_{i',j}\hat{\lambda}_{i',j}(\mathbf{u}) = s_{j,j}\lambda_{j,i}(\mathbf{u})$. Then by assumption (v), we have $A_{i,j} = 0$.

As a result, the matrix $\mathbf{A}^{-1}$ is the same as the matrix $\mathbf{P}^{-1}$. That is,

$$\mathbf{z} = \mathbf{P}\hat{\mathbf{z}} + \mathbf{c_z}. \tag{55}$$

A.2 Understanding Assumptions in Theorem 4.1

Gaussian assumption on the latent noise variables One of our model assumption is enforcing Gaussian distribution on the latent noise variables. Note that the assumptions of nonlinear ICA (Khemakhem et al., 2020) on the noise could be broad exponential family distribution, *e.g.*, Gaussian distributions, Laplace, Gamma distribution, and so on. This work considers Gaussian distribution, mainly because it can be straightforwardly implemented for the re-parameterization trick in VAE. 2) It is convenient for simplifying the proof process. It is worthwhile and promising to extend the Gaussian assumption to exponential family for future work.

Linear model assumption for the latent causal variables As the first work to discuss the relation between the change of causal influences and identifiability of latent causal model, this work simply considers linear models for the latent causal variables, because it can be directly parameterized, which helps us to analyze the challenge of identifiability and how to handle it. In addition, together with Gaussian noise, we arrive at the same probabilistic model as linear Gaussian Bayesian networks, which is a well-studied model as mentioned in Eq. (6). We expect that the proposed weights-variant linear models can motivate more general functional classes in future work, *e.g.*, nonlinear additive noise models.

Changes of weights We allow changes of weights (*i.e.*, causal influences) among latent causal variables across $\mathbf{u}$ to handle the transitivity problem. We argue that this may be not a necessary condition for identifiability (Existing methods, including sparse graph structure and temporal information, could be regarded as feasible ways to handle the transitivity.), but it is a sufficient condition to handle the transitivity, and provide a new research line for causal representation learning. In fact, exploring the change of causal influences is not new in observed space (Ghassami et al., 2018; Huang et al., 2020). In addition, identifying latent causal variables by randomly chosen unknown hard intervention can also be regarded as a special change of causal influences (Brehmer et al., 2022). From this viewpoint, we argue that exploring the change of causal influences for identifying latent causal representation may be a promising way and not too restricted in reality.

We emphasize that changes of weights among latent variables across different values of $\mathbf{u}$ are essential for achieving full identifiability. However, this does not mean that all weights must vary continuously or at every point in $\mathbf{u}$. Some weights may remain unchanged at certain values of $\mathbf{u}$, as long as they exhibit variation at other values of $\mathbf{u}$. This flexibility is both theoretically sufficient and practically relevant. See Section 7.1.1 for experimental evidence and further discussion on this point.

A.3 The Proof of Corollary 2

Theorem 1 has shown that the latent causal variables $\mathbf{z}$ can be identified up to trivial permutation and linear scaling. Hence, the identifiability of the causal structure in latent space can be reduced to the identifiability of the causal structure in observed space. Moreover, we need to show that the linear scaling does not affect theoretical identifiability of the causal structure.

For the identifiability of the causal structure in observed space from heterogeneous data, fortunately, we can leverage the results from Huang et al. (2020). Corollary 4.2 relies on the Markov condition and faithfulness assumption and the assumption that the latent change factor (i.e., causal strength in the linear case) can be represented as a function of the domain index $\mathbf{u}$. Hence, it relies on the same assumptions as that in Huang et al. (2020).

It has been shown in Huang et al. (2020) that if the joint distribution over $\mathbf{z}$ and $\mathbf{u}$ ($\mathbf{z}$ and $\mathbf{u}$ are observed variables here) are Markov and faithful to the augmented graph, then the causal structure over $\mathbf{z} \cup \mathbf{u}$ can be identified up to the Markov equivalence class, by making use of the conditional independence relationships.

Next, we show that the Markov equivalence class over $\mathbf{z}$ is also identifiable. Denote by M the Markov equivalence class over $\mathbf{z} \cup \mathbf{u}$, and by M_z the Markov equivalence class over $\mathbf{z}$. Then after removing variable $\mathbf{u}$ in M and its edges, the resulting graph (denoted by M_z') is the same as M_z. This is because of the following reasons. First, it is obvious that M_z' and M_z have the same skeleton. Second, in this paper, $\mathbf{u}$ has an edge over every z_i when considering $\mathbf{n}$ as latent noise variables, because all causal strength and noise distributions change with $\mathbf{u}$. Hence, there is no v-structure over $\mathbf{u}$, z_i, and z_j, so it is not possible to have more oriented edges in M_z'. Therefore, M_z' and M_z have the same skeleton and the same directions.

Moreover, the conditional independence relationships will not be affected by the linear scaling of the variables, so the conditional independence relationships still hold in the identified latent variables. Furthermore, for linear-Gaussian models, independence equivalence is the same as distributional equivalence. This is because of the following reasons. First, since we are concerned with linear Gaussian models over the latent variables, the identifiability up to equivalence class also holds for score-based methods that use the likelihood of data as objective functions. This is because independence equivalence, i.e., two DAGs have the same conditional independence relations, is the same as distributional equivalence for linear-Gaussian models, i.e., two Bayesian networks corresponding to the two DAGs can define the same probability distribution. That is to say, score-based methods also find the structure based on independence relations implicitly. Because the scaling does not change the independence relations, it will also not affect the identifiability of the graph structure.

Therefore, the causal structure among latent variables $\mathbf{z}$ can be identified up to the Markov equivalence class.

A.4 The Proof of Corollary 3

To prove this corollary, let us introduce ICM first.

Definition 5 *(Independent Causal Mechanisms). In a causally sufficient system, the causal modules, as well as their included parameters, change independently across domains.*

Here we follow Ghassami et al. (2018) to define ICM principle. Another slightly different definition appears in Peters et al. (2017); Schölkopf et al. (2021). Both two definitions imply the same principle that the causal modules change independently across $\mathbf{u}$. According to ICM principle, in the causal direction, the causal modules, as well as their included parameters, change independently across $\mathbf{u}$. However, such independence generally does not hold in the anti-causal direction (Ghassami et al., 2018; Huang et al., 2020). For example, consider z_1 and z_2 in Figure 1. According to model definition in Eq. (1)- Eq. (3), for causal direction $z_1 \rightarrow z_2$, we have:

$$n_i :\sim \mathcal{N}(\beta_{i,1}(\mathbf{u}), \beta_{i,2}(\mathbf{u})), \qquad z_1 := n_1, \qquad z_2 := \lambda_{1,2}(\mathbf{u})z_1 + n_2. \tag{56}$$

For the reverse direction, we have

$$
\begin{aligned}
z_1 &= \lambda_1'(\mathbf{u})z_2 + n_1', & z_2 &= n_2' \\
\lambda_1'(\mathbf{u}) &= \frac{1}{\lambda_{1,2}(\mathbf{u})}, \quad n_1' = -\frac{1}{\lambda_{1,2}(\mathbf{u})}n_2 & n_2' &= \lambda_{1,2}(\mathbf{u})z_1 + n_2
\end{aligned}
\tag{57}
$$

In this case, the module $p(z_2)$ is dependent of the module $p(z_1|z_2)$, $i.e.$, both modules depend on the same parameter $\lambda_{1,2}(\mathbf{u})$, which violates ICM principle. Clearly, the scaling indeterminacy for the recovered latent variables does not affect such dependence. As a result, regardless of the scaling, with the help of ICM, we can fully identify the acyclic causal structure among latent variables $\mathbf{z}$.

A.5 The proof of Theorem 4

Throughout the proof process in Section A.1, Eq. (54) holds without the need for assumption (v). Note that this also implies that $\mathbf{A}^{-1}$ is a lower triangular matrix. Then consider the following two cases.

- For the case where z_i is a root node or all weights on all paths from parent nodes of z_i to z_i meet assumption (v), by using assumption (v) and by comparison both sides of Eq. (54), we have: for all $j < i$, $A_{i,j} = 0$, which implies that we can obtain that $A_{i,i}^{-1}z_i = \hat{z}_i + c_i'$.

- If there exists an unchanged weight on all paths from parent nodes of z_i to z_i across $\mathbf{u}$, we demonstrate that it is always possible to construct an alternative solution, which is different from the true $\mathbf{z}$, but capable of generating the same observations $\mathbf{x}$. Suppose that for z_i, there is an unchanged weight $\lambda_{j,i}$ across $\mathbf{u}$, related to the parent node z_j. Then, we can always construct new latent variables $\mathbf{z}'$ as: for all $k \neq i$, $z_k' = z_k$, and $z_i' = z_i - \lambda_{j,i}z_j$. Given this, we can construct a matrix $\mathbf{M}$, so that $\mathbf{Mz}' = \mathbf{z}$. It is clear the matrix $\mathbf{M}$ is invertible. In addition, all the weights of the matrix $\mathbf{M}$ are constant and thus do not depend on $\mathbf{u}$. As a result, we can construct a mapping from $\mathbf{z}'$ to $\mathbf{x}$, $e.g.$, $\mathbf{f}(\mathbf{Mz}')$, which is invertible and do not depend on $\mathbf{u}$, and can create the same data $\mathbf{x}$ generated by $\mathbf{f}(\mathbf{z})$. Therefore, the alternative solution $\mathbf{z}'$ can lead to a non-identifiability result.

- For case that there exists an weight $\lambda_{j,i}(\mathbf{u}) + b$, we can always construct new latent variables $\mathbf{z}'$ as: for all $k \neq i$, $z_k' = z_k$, and $z_i' = z_i - bz_j$. Given this, we can construct

a matrix $\mathbf{M}$, so that $\mathbf{M}\mathbf{z}' = \mathbf{z}$. It is clear the matrix $\mathbf{M}$ is invertible. In addition, all the weights of the matrix $\mathbf{M}$ are constant and thus do not depend on $\mathbf{u}$. Finally, we can construct a mapping from $\mathbf{z}'$ to $\mathbf{x}$, $e.g.$, $\mathbf{f}(\mathbf{M}\mathbf{z}')$, which do not depend on $\mathbf{u}$, and can create the same data $\mathbf{x}$.

A.6 Implementation Details

A.6.1 Experiments for Figure 7

Data For experimental results of Figure 7 the number of segments ($i.e.$, $\mathbf{u}$) M is 10, 20, 30, 40 and 50 respectively. For each segment, the number of latent causal variables is 2 and the sample size is 1000. We consider the following structural causal model

$$n_1 := \mathcal{N}(\beta_{1,1}(\mathbf{u}), \beta_{1,2}(\mathbf{u})), \qquad n_2 := \mathcal{N}(\beta_{2,1}(\mathbf{u}), \beta_{2,2}(\mathbf{u})), \tag{58}$$
$$z_1 := n_1, \qquad z_2 := \lambda_{1,2}(\mathbf{u})z_1 + n_2, \tag{59}$$

where we sample the mean $\beta_{i,1}(\mathbf{u})$ and variance $\beta_{i,2}(\mathbf{u})$ from uniform distributions $[-2, 2]$ and $[0.01, 3]$, respectively. We sample the weights $\lambda_{1,2}(\mathbf{u})$ from uniform distributions $[0.1, 2]$. We sample latent variable z according to Eqs. 58 and 59, and then mix them using a 2-layer multi-layer perceptron (MLP) to generate observed $\mathbf{x}$.

Network and Optimization For all methods, we used a encoder, $i.e.$, 3-layer fully connected network with 30 hidden nodes and Leaky-ReLU activation functions, and decoder, $i.e.$, 3-layer fully connected network with 30 hidden nodes and Leaky-ReLU activation functions. We also use 3-layer fully connected network with 30 hidden nodes and Leaky-ReLU activation functions. For optimization, we use Adam optimizer with learning rate $1e - 3$. For hyper-parameters, we set $\beta = 4$ for the β-VAE. For CausalVAE, we set the hyper-parameters as suggested in Yang et al. (2021).

A.6.2 Experiments for Figures 6 and 8

Data For experimental results of Figures 6 and 8, the number of segments is 30 and sample size is 1000, while the number of latent causal variables is 2,3,4,5 respectively. For 2-dimensional latent causal variables, the setting is the same as experiments for Figure 7 (a). For 3-dimensional case, we consider the following structural causal model:

$$n_1 := \mathcal{N}(\beta_{1,1}(\mathbf{u}), \beta_{1,2}(\mathbf{u})), \quad n_2 := \mathcal{N}(\beta_{2,1}(\mathbf{u}), \beta_{2,2}(\mathbf{u})), \quad n_3 := \mathcal{N}(\beta_{3,1}(\mathbf{u}), \beta_{3,2}(\mathbf{u})), \tag{60}$$
$$z_1 := n_1, \qquad z_2 := \lambda_{1,2}(\mathbf{u})z_1 + n_2, \qquad z_3 := \lambda_{2,3}(\mathbf{u})z_2 + n_3. \tag{61}$$

For 4-dimensional case, we consider the following structural causal model:

$$n_1 := \mathcal{N}(\beta_{1,1}(\mathbf{u}), \beta_{1,2}(\mathbf{u})), \qquad n_2 := \mathcal{N}(\beta_{2,1}(\mathbf{u}), \beta_{2,2}(\mathbf{u})), \tag{62}$$
$$n_3 := \mathcal{N}(\beta_{3,1}(\mathbf{u}), \beta_{3,2}(\mathbf{u})), \qquad n_4 := \mathcal{N}(\beta_{4,1}(\mathbf{u}), \beta_{4,2}(\mathbf{u})), \tag{63}$$
$$z_1 := n_1, \qquad z_2 := \lambda_{1,2}(\mathbf{u})z_1 + n_2, \tag{64}$$
$$z_3 := \lambda_{2,3}(\mathbf{u})z_2 + n_3, \qquad z_4 := \lambda_{2,4}(\mathbf{u})z_2 + n_4. \tag{65}$$

For 5-dimensional case, we consider the following structural causal model:

$$n_1 := \mathcal{N}(\beta_{1,1}(\mathbf{u}), \beta_{1,2}(\mathbf{u})), \quad n_2 := \mathcal{N}(\beta_{2,1}(\mathbf{u}), \beta_{2,2}(\mathbf{u})), \tag{66}$$

$$n_3 := \mathcal{N}(\beta_{3,1}(\mathbf{u}), \beta_{3,2}(\mathbf{u})), \quad n_4 := \mathcal{N}(\beta_{4,1}(\mathbf{u}), \beta_{4,2}(\mathbf{u})), \quad n_5 := \mathcal{N}(\beta_{5,1}(\mathbf{u}), \beta_{5,2}(\mathbf{u})), \tag{67}$$

$$z_1 := n_1, \qquad\qquad\qquad z_2 := \lambda_{1,2}(\mathbf{u})z_1 + n_2, \tag{68}$$

$$z_3 := \lambda_{2,3}(\mathbf{u})z_2 + n_3, \qquad z_4 := \lambda_{2,4}(\mathbf{u})z_2 + n_4, \qquad z_5 := \lambda_{4,5}(\mathbf{u})z_4 + n_5. \tag{69}$$

Again, for all these Eqs 60-69, we sample the mean $\beta(\mathbf{u})_{i,1}$ and variance $\beta(\mathbf{u})_{i,2}$ from uniform distributions $[-2, 2]$ and $[0.01, 3]$, respectively. We sample the weights $\lambda_{i,j}(\mathbf{u})$ from uniform distributions $[0.1, 2]$. We sample latent variable $\mathbf{z}$ according to these equations, and then mix them using a 2-layer MLP to generate observed $\mathbf{x}$.

Network and Optimization   The network architecture, optimization and hyper-parameters are the same as used for experiments of Figure 7 (a), except for the 5-dimensional case where we use 40 hidden nodes for each linear layer.

A.6.3 Experiments for Table 1

For experimental results of Table 1, the number of segments is 30 and the sample size is 1000, and the number of latent causal variables is 3. For the case of the matching assumptions, we consider the structural causal model as mentioned in Eqs. (60)-(61). For the cases of uniform, Laplace and Gamma noise, we replace the original Gaussian noise by using uniform noise, Laplace and Gamma noise, respectively, while the remaining parts are the same as Eqs. (60)-(61). The network architecture, optimization, and hyper-parameters are the same as used for experiments of Figure 7 (a).

A.6.4 Implementation of Enforcing Assumption (v)

To enforce assumption (v), which requires $\lambda_{i,j}(\mathbf{u} = \mathbf{0}) = 0$ in the prior model and $\lambda'_{i,j}(\mathbf{u} = \mathbf{0}) = 0$ in the posterior model, we augment the training process with *synthetic pairs* that regularize model behavior at $\mathbf{u} = \mathbf{0}$, in the absence of observed data at this point. We construct synthetic pairs of the form: (input: $\mathbf{u} = \mathbf{0}$,   target: 0), where the target value "0" is not associated with the observed variable $\mathbf{x}$, but instead reflects the expected output of a specific model component evaluated at $\mathbf{u} = \mathbf{0}$. Specifically, for the prior model Eq. (10), this corresponds to enforcing $\lambda_{i,j}(\mathbf{0}) = 0$. For the inference model Eq. (11), it enforces $\lambda'_{i,j}(\mathbf{0}) = 0$. These synthetic pairs are therefore not part of the data distribution $(\mathbf{u}, \mathbf{x})$, but rather encode theoretical constraints derived from assumption (v). We integrate the synthetic pairs into the training objective alongside observed data $(\mathbf{u}, \mathbf{x})$. For each pair, we apply an L_1 penalty on the model output at $\mathbf{u} = \mathbf{0}$: $\|\lambda_{i,j}(\mathbf{0})\|_1$ for the prior model, $\|\lambda'_{i,j}(\mathbf{0})\|_1$ for the inference model. These terms are added to the ELBO loss as shown in Eq. (13), enabling the model to conform to assumption (v).

A.7 Detailed Results on fMRI

Figures 21-25 show the recovered latent six signals (Blue) and the true ones (Red) within one day by the proposed SuaVE, iVAE, VAE, β-VAE, and CausalVAE, respectively. Regarding to implementation details, for fMRI data, again, we used the same network architecture
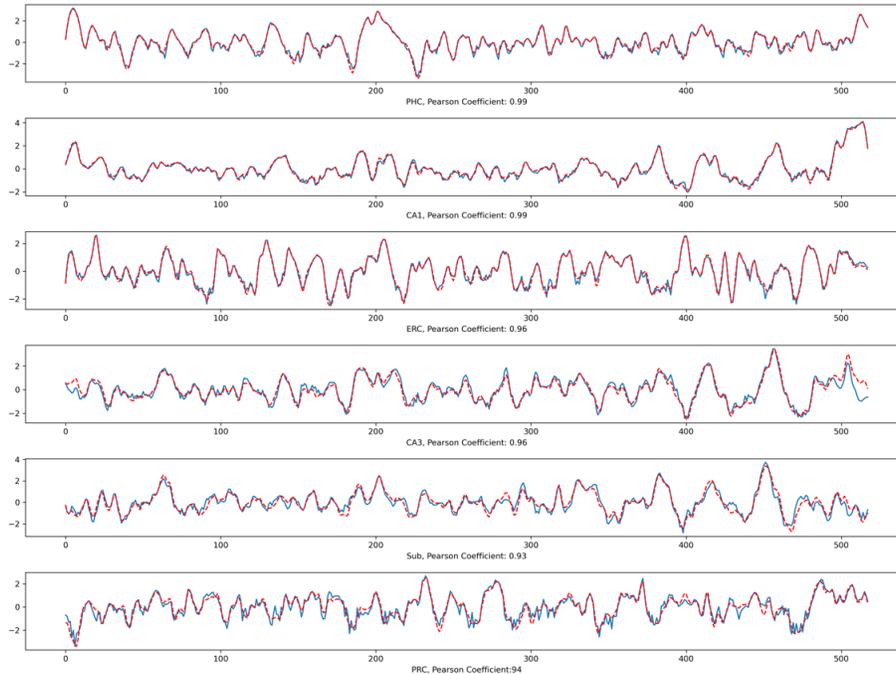
Figure 21: Recovered six signals (Blue) and the true ones (Red) within one day by SuaVE.

for encoder (*i.e.*, 3-layer fully connected network with 30 hidden nodes for each layer) and decoder (*i.e.*, 3-layer fully connected network with 30 hidden nodes for each layer) parts in all these models. For prior model in the proposed SuaVE and iVAE, we use 3-layer fully connected network with 30 hidden nodes for each layer. We assign an 3-layer fully connected network with 30 nodes to generate the weights to model the relations among latent causal variables in the proposed SuaVE. For hyper-parameters, we set $\beta = 4, 25, 50$ for β-VAE. For CausalVAE, we use the hyper-parameters setting as recommended in Yang et al. (2021), for generating observed $\mathbf{x}$, we use an invertible 2-layer multi-layer perceptron on the fMRI with different 5 random seeds.
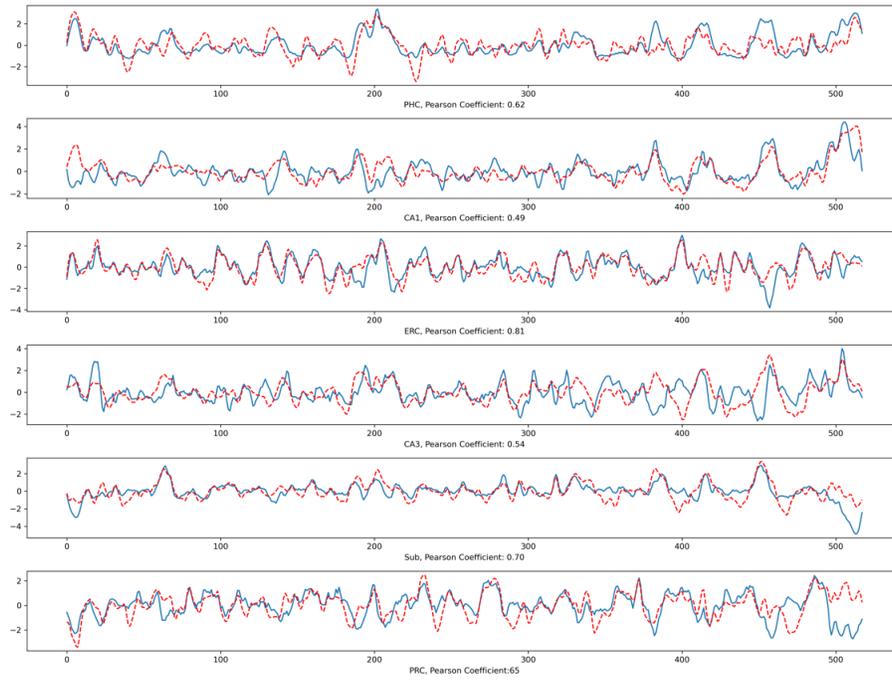
Figure 22: Recovered six signals (Blue) and the true ones (Red) within one day by iVAE.
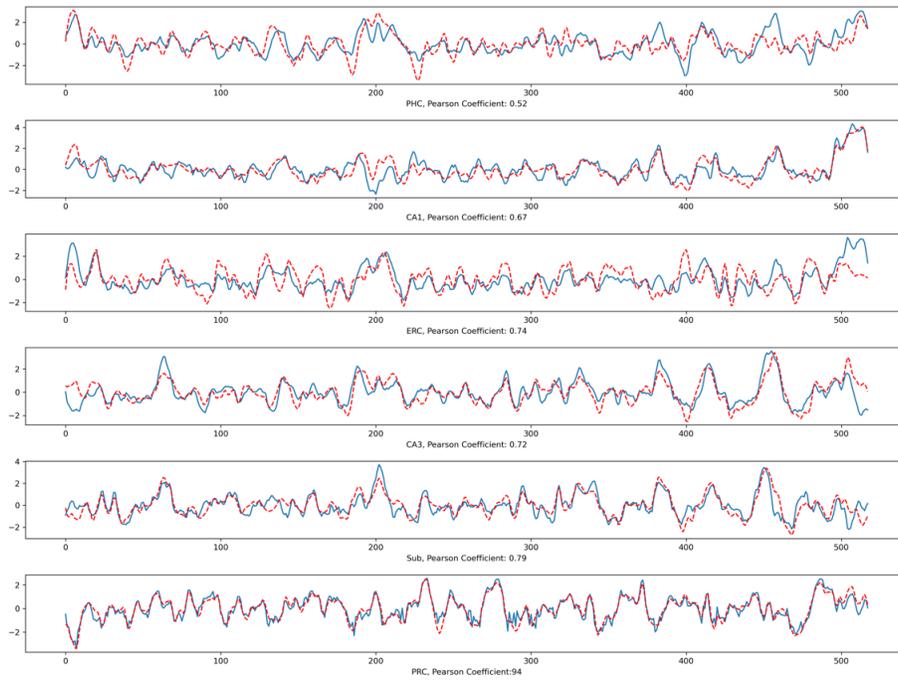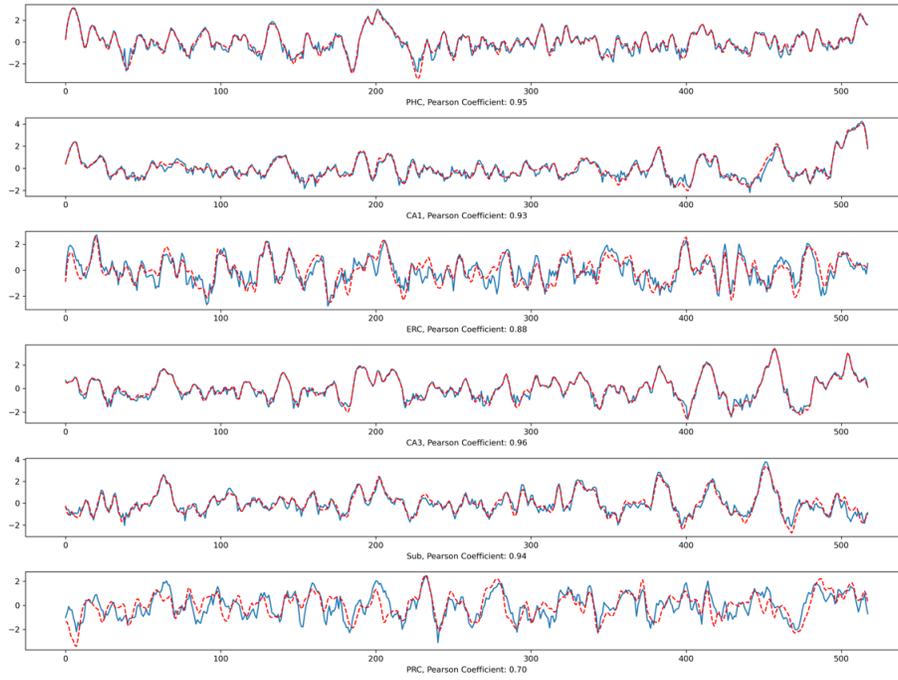

Figure 23: Recovered six signals (Blue) and the true ones (Red) within one day by VAE.

Figure 24: Recovered six signals (Blue) and the true ones (Red) within one day by β-VAE.
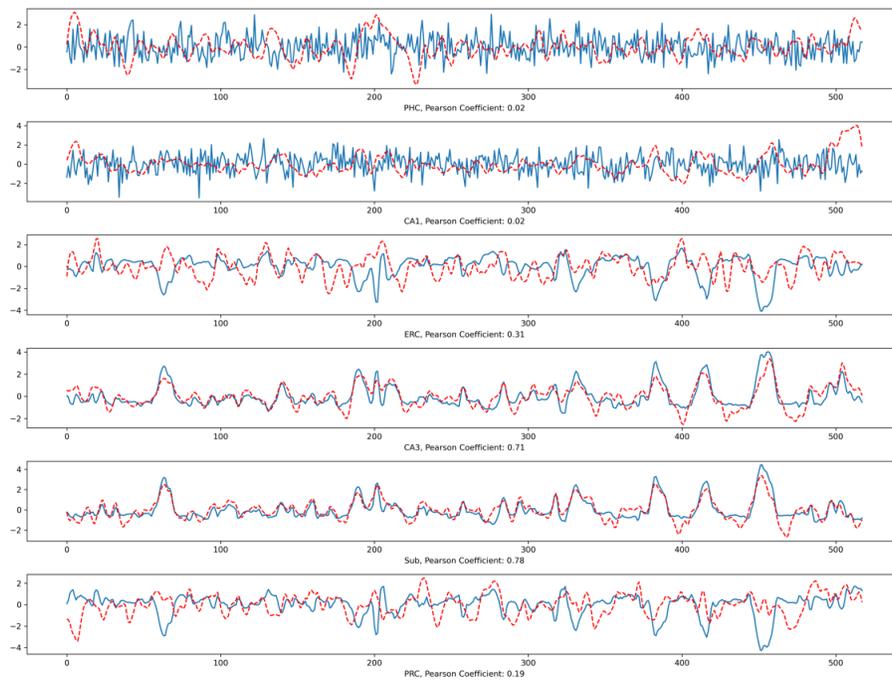

Figure 25: Recovered six signals (Blue) and the true ones (Red) by CausalVAE.

References

Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. In *NeurIPS*, 2021.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023.

Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham Kakade. Learning linear bayesian networks with latent variables. In *ICML*, pages 249–257, 2013.

Michelle R Arkin and James A Wells. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature reviews Drug discovery*, 3(4):301–317, 2004.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

Chris M Bird and Neil Burgess. The hippocampus and memory: insights from spatial processing. *Nature Reviews Neuroscience*, 9(3):182–194, 2008.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *NeurIPS*, 2019.

Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, 2021.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In *NeurIPS*, 2018.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

Benjamin Frot, Preetam Nandy, and Marloes H Maathuis. Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):459–487, 2019.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, 31, 2018.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 596–605, 2021.

I. Higgins, Loïc Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

B. Huang, K. Zhang, M. Gong, and C. Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. *ICML*, 2019.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *JMLR*, 21(89), 2020.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTAS*, pages 2207–2217. PMLR, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018.

D. P. Kingma and M Welling. Auto-encoding variational bayes. *NeurIPS*, 2013a.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013b.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, volume 29, pages 4743–4751, 2016.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. *NeurIPS*, 2021.

Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *ICLR*, 2018.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International conference on machine learning*, pages 11455–11472. PMLR, 2022.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. *arXiv preprint arXiv:2107.10098*, 2021.

Timothy O. Laumann and Russell A. Poldrack, 2015. URL `https://openfmri.org/dataset/ds000031/`.

Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. Subspace identification for multi-source domain adaptation. *Advances in Neural Information Processing Systems*, 36:34504–34518, 2023.

Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pages 6804–6814. PMLR, 2021.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Latent covariate shift: Unlocking partial identifiability for multi-source domain adaptation. *Transactions on Machine Learning Research*, 2025.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *ICML*, 2020.

Haiying Lu, Qiaodan Zhou, Jun He, Zhongliang Jiang, Cheng Peng, Rongsheng Tong, and Jianyou Shi. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal transduction and targeted therapy*, 5 (1):213, 2020.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. In *NeurIPS*, 2019.

Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

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

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15(58):2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press, 2017.

Malin L Pinsky, Rebecca L Selden, and Zoë J Kitchel. Climate-driven shifts in marine species ranges: Scaling from organisms to communities. *Annual review of marine science*, 12:153–179, 2020.

Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 200–201, 2020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Anna Seigal, Chandler Squires, and Caroline Uhler. Linear causal disentanglement via interventions. *arXiv preprint arXiv:2211.16467*, 2022.

Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *JMLR*, 7(2), 2006.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *NeurIPS*, 2016.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

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

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *ICML*, pages 10401–10412. PMLR, 2021.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in neural information processing systems*, 2021.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12 (2):207, 2020.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *NeurIPS*, 2020a.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pages 24370–24387. PMLR, 2022a.

Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. *arXiv preprint arXiv:2012.04550*, 2020b.

Shaoan Xie, Lingjing Kong, Mingming Gong, and Kun Zhang. Multi-domain image generation and translation with identifiability guarantees. In *The Eleventh International Conference on Learning Representations*, 2022b.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. In *CVPR*, 2021.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

Weiran Yao, Guangyi Chen, and Kun Zhang. Learning latent causal dynamics. *arXiv preprint arXiv:2202.04828*, 2022.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *ICML*, 2019.

Valentina Zantedeschi, Luca Franceschi, Jean Kaddour, Matt J Kusner, and Vlad Niculae. Dag learning on the permutahedron. *arXiv preprint arXiv:2301.11898*, 2023.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *NeurIPS*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. Pmlr, 2020.