Causal Disentanglement for Time Series

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, in the field of nonlinear Independent Component Analysis (ICA), strong identifiability results for disentanglement have been established by using certain side information, such as class labels or history information for time series, in addition to independence. However, most existing work is constrained by functional form assumptions such as stationary independent sources or further with linear transitions and distribution assumptions such as exponential family distribution. It is unknown whether the underlying latent processes and their causal relations are identifiable if they have arbitrary, nonparametric causal influences in between. We propose a principled framework called *LCD-NM* to recover time-delayed latent causal variables and identify their relations from measured temporal data under stationary environments and under different distribution shifts. Specifically, the framework factorizes unknown distribution shifts into transition distribution changes caused by fixed dynamics, time-varying latent causal relations, and global changes in observation. We establish the identifiability theories of nonparametric latent causal processes from their nonlinear mixtures under fixed dynamics and analyze how distribution changes can further benefit the identifiability. Through experiments, we show that time-delayed latent causal influences are reliably identified and that our approach considerably outperforms existing baselines that do not correctly exploit this modular representation of changes. Our results demonstrate that disentanglement in time-series settings seems promising both in stationary environments and general nonstationary environments, in which the latent processes have nonparametric causal influences in between.

1 Introduction

Causal discovery from time-series data is a crucial task in numerous fields [1, 2, 3]. Most existing work focuses on estimating the temporal causal relations among observed variables. However, in many real-world scenarios, the observed signals (e.g., image pixels in videos) do not have direct causal edges, but are generated by latent temporal processes or confounders that are causally related. Inspired by these scenarios, this work aims to uncover causally-related latent processes and their relations from observed temporal variables. Estimating latent causal structure from observations, which we assume are unknown (but invertible) nonlinear mixtures of the latent processes, is very challenging. It has been found in [4, 5] that without exploiting an appropriate class of assumptions in estimation, the latent variables are not identifiable or disentangled in the most general case. As a result, one cannot make causal claims on the recovered relations in the latent space.

Recently, in the field of nonlinear Independent Component Analysis (ICA), strong identifiability results of the latent variables have been established [6, 7, 8, 9, 10] by using certain side information, such as class labels, in addition to independence. For time-series data, history information is widely used as the side information for the disentanglement of latent processes. To establish identifiability, the existing approaches enforce different sets of functional and distributional form assumptions as

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

Table 1: Attributes of nonlinear ICA theories for time-series. A check denotes that a method has an attribute or can be applied to a setting, whereas a cross denotes the opposite. † indicates our approach.

Theory	Time-varying Relation	Causally-related Process	Partitioned Subspace	Nonparametric Transition	Applicable to Stationary Environment
PCL	✗	✗	✗	✓	✓
GCL	✓	✗	✗	✓	✓
HM-NLICA	✗	✗	✗	✗	✗
SlowVAE	✗	✗	✗	✗	✓
SNICA	✓	✓	✗	✗	✗
i-VAE	✓	✗	✗	✗	✗
LEAP	✗	✓	✗	✓	✗
LCD-NM †	✓	✓	✓	✓	✓

constraints in estimation; for example, **(1)** PCL [7], GCL [8], HM-NLICA [11] and SlowVAE [12] assume mutually-independent sources in the data generating process. However, this assumption may severely distort the identifiability if the latent variables have time-delayed causal relations in between (i.e., causally-related process); **(2)** SlowVAE [12] and SNICA [13] assume linear relations, which may distort the identifiability results if the underlying transitions are nonlinear, and **(3)** SlowVAE [12] assumes that the process noise is drawn from Laplacian distribution; i-VAE [9] assumes that the conditional transition distribution is part of the exponential family. However, in real-world scenarios, one cannot choose a proper set of functional and distributional form assumptions without knowing in advance the parametric forms of the latent temporal processes. Our first step is hence to understand under what conditions the latent causally processes are identifiable if they have nonparametric transitions in between. With the proposed condition, our approach allows recovery of latent temporal causally-related processes in stationary environments without knowing their parametric forms in advance. On the other hand, nonstationarity has greatly improved the identifiability results for learning the latent causal structure [14, 15, 16]. For instance, LEAP [14] established the identifiability of latent temporal processes, but in limited nonstationary cases, under the condition that the distribution of the noise terms of the latent processes varies across segments. Our second step is to analyze how distribution shifts benefit our stationary condition and to extend our condition to a general nonstationary case. Accordingly, our approach enables the recovery of latent temporal causal processes in a general nonstationary environment with time-varying relations such as changes in the influencing strength or switching some edges off [17] over time or domains.

Given the identifiability results, we propose a learning framework, called *LCD-NM*, to recover nonparametric time-delayed latent causal variables and identify their relations from measured temporal data under stationary environments and under nonstationary environments in which it is unknown in advance how the joint distribution changes across domains (we define it as "unknown distribution shifts"). For instance, Fig. 1 shows an example of multiple video domains of a physical system under different mass, gravity, and environment rendering settings[1]. With *LCD-NM*, the differences across segments are characterized by the learned change factors $\hat{\boldsymbol{\theta}}_r^{\text{dyn}}$ of domain r (note that domain index is given to the model) that encode changes in transition dynamics, and changes in observation or styles modeled by $\hat{\boldsymbol{\theta}}_r^{\text{obs}}$ (we use "causal dynamics" and "latent causal relations/influences" inter-

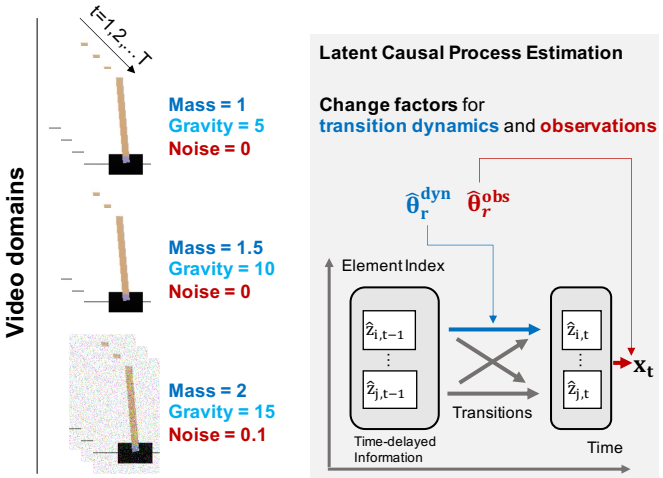

Figure 1: **LCD-NM**: learning Latent Causal Dynamics from Nonliner Mixtures. We exploit fixed causal dynamics and distribution changes from changing causal influences and observation changes to identify the underlying causal processes. $\hat{z}_{i,t}$ denotes the estimated latent process. $\hat{\theta}_r^{\text{dyn}}$ is the change factor for transition dynamics, i.e., representing mass and gravity in this example. $\hat{\theta}_r^{\text{obs}}$ is the change factor for observation, i.e., noise scale.

[1]The variables and functions with "hat" are estimated by the model; the ones without "hat" are ground truth.

changeably). We then present a generalized time-series data generative model that takes these change factors as arguments for modeling the distribution changes. Specifically, we factorize unknown distribution shifts into transition distribution changes in stationary processes, time-varying latent causal relations, and global changes in observation by constructing partitioned latent subspaces, and propose provable conditions under which nonparametric latent causal processes can be identified from their nonlinear invertible mixtures. We demonstrate through a number of real-world datasets, including video and motion capture data that time-delayed latent causal influences are reliably identified from observed variables under stationary environments and unknown distribution shifts. Through experiments, we show that our approach considerably outperforms existing baselines that do not correctly leverage this factorized representation of changes. Our results show that disentanglement in time-series settings seems promising both in stationary and general nonstationary environments, in which the latent processes have arbitrary, nonparametric causal influences in between.

2 Related Work

Causal Discovery from Time Series Inferring the causal structure from time-series data is critical to many fields including machine learning [1], econometrics [2], and neuroscience [3]. Most existing work focuses on estimating the temporal causal relations between observed variables. For this task, constraint-based methods [18] apply the conditional independence tests to recover the causal structures, while score-based methods [19, 20] define score functions to guide a search process. Furthermore, [21, 22] propose to fuse both conditional independence tests and score-based methods. The Granger causality [23] and its nonlinear variations [24, 25] are also widely used.

Nonlinear ICA for Time Series Temporal structure and nonstationarities were recently used to achieve identifiability in nonlinear ICA. Time-contrastive learning (TCL [6]) used the independent sources assumption and leveraged sufficient variability in variance terms of different data segments. Permutation-based contrastive (PCL [7]) proposed a learning framework which discriminates between true independent sources and permuted ones, and identifiable under the uniformly dependent assumption. HM-NLICA [11] combined nonlinear ICA with a Hidden Markov Model (HMM) to automatically model nonstationarity without manual data segmentation. i-VAE [9] introduced VAEs to approximate the true joint distribution over observed and auxiliary nonstationary regimes. Their work assumes that the conditional distribution is within exponential families to achieve the identifiability of the latent space. The most recent literature on nonlinear ICA for time-series includes LEAP [14] and (i-)CITRIS [26, 27]. LEAP proposed a nonparametric condition leveraging the nonstationary noise terms. However, all latent processes are changed across contexts and the distribution changes need to be modeled by nonstationary noise and it does not exploit the stationary nonparametric components for identifiability. Alternatively, CITRIS proposed to use intervention target information for identification of scalar and multidimensional latent causal factors. This approach does not suffer from functional or distributional form constraints, but needs access to active intervention.

3 Problem Formulation

3.1 Time Series Generative Model

As a **fundamental case**, we first present a regular, stationary time-series generative process where the observations $\mathbf{x}_t$ comes from a nonlinear (but invertible) mixing function $\mathbf{g}$ that maps the time-delayed causally-related latent variables $\mathbf{z}_t$ to $\mathbf{x}_t$. The latent variables or processes $\mathbf{z}_t$ have stationary, nonparametric time-delayed causal relations. Let τ be the time lag:

$$\underbrace{\mathbf{x}_t = g(\mathbf{z}_t)}_{\text{Nonlinear mixing}}, \quad \underbrace{z_{it} = f_i\left(\{z_{j,t-\tau} | z_{j,t-\tau} \in \mathbf{Pa}(z_{it})\}, \epsilon_{it}\right)}_{\text{Stationary nonparametric transition}} \; with \; \underbrace{\epsilon_{it} \sim p_{\epsilon_i}}_{\text{Stationary noise}} \; .$$

Note that with nonparametric causal transitions, the noise term $\epsilon_{it} \sim p_{\epsilon_i}$ (where p_{ϵ_i} denotes the distribution of ϵ_{it}) and the time-delayed parents $\mathbf{Pa}(z_{it})$ of z_{it} (i.e., the set of latent factors that directly cause z_{it}) are interacted and transformed in an arbitrarily nonlinear way to generate z_{it}. Under stationarity assumptions, the mixing function $\mathbf{g}$, the transition functions f_i and the noise distributions p_{ϵ_i} are invariant. Finally, we assume that the noise terms are mutually-independent (i.e., spatially and temporally independent), which implies that instantaneous causal influence between latent causal processes is not allowed by the formulation. The stationary time-series model in the fundamental case is used to establish the identifiability results under fixed causal dynamics in Section 4.1.

We further consider two violations of the stationarity assumptions in the fundamental case, which lead to two nonstationary time series models. Let $\mathbf{u}$ denote the domain or regime index. Suppose there exist m regimes of data, i.e., u_r with $r = 1, 2, ..., m$, with unknown distribution shifts. In practice, the changing parameters of the joint distribution across domains often lie in a low-dimensional manifold [28]. Moreover, if the distribution is causally factorized, the distributions are often changed in a minimal and sparse way [29]. Based on these assumptions, we introduce the low-dimensional minimal change factor $(\boldsymbol{\theta}_r^{\text{dyn}}, \boldsymbol{\theta}_r^{\text{obs}})$, which was proposed in [30], to respectively capture distribution shifts in transition functions and observation. The vector $\boldsymbol{\theta}_r = (\boldsymbol{\theta}_r^{\text{dyn}}, \boldsymbol{\theta}_r^{\text{obs}})$ has a constant value in each domain but varies across domains. The formulation of the nonstationary time-series model is in line with [30]. The nonstationary model is used to establish the identifiability results under nonstationary cases in Section 4.2, where we show that the violation of stationarity in both ways can even further improve the identifiability results. We first present the two nonstationary cases. **(1) Changing Causal Dynamics**. The causal influences between the latent temporal processes are changed across domains in this setting. We model it by adding the transition change factors $\boldsymbol{\theta}_r^{\text{dyn}}$ as inputs to the transition function: $z_{it} = f_i\left(\{z_{j,t-\tau} | z_{j,t-\tau} \in \mathbf{Pa}(z_{it})\}, \boldsymbol{\theta}_r^{\text{dyn}}, \epsilon_{it}\right)$. **(2) Observation Changes.** The global properties of the time series (e.g., video styles) are changed across domains in this setting. Our model captures them using latent variables that represent global styles; these latent variables are generated by a bijection f_i that transforms the noise terms $\epsilon_{i,t}$ into the latent with change factor $\boldsymbol{\theta}_r^{\text{obs}}$: $z_{i,t} = f_i\left(\boldsymbol{\theta}_r^{\text{obs}}, \epsilon_{i,t}\right)$. Finally, we can deal with a more general nonstationary case by combining the three types of latent processes in the latent space in a modular way. **Modular Distribution Shifts.** The latent space has three blocks $\mathbf{z}_t = (\mathbf{z}_t^{\text{fix}}, \mathbf{z}_t^{\text{chg}}, \mathbf{z}_t^{\text{obs}})$ where $z_{s,t}^{\text{fix}}$ denotes the sth component of the fixed dynamics parts, $z_{c,t}^{\text{chg}}$ denotes the cth component of the changing dynamics parts, and $\mathbf{z}_{o,t}^{\text{obs}}$ is the oth component of the global changes. The functions $[f_s, f_c, f_o]$ capture fixed and changing transitions and observation changes for each dimension of $\mathbf{z}_t$ in Eq. 1.

$$\begin{cases} z_{s,t}^{\text{fix}} &= f_s\left(\{z_{i,t-\tau} | z_{i,t-\tau} \in \mathbf{Pa}(z_{s,t}^{\text{fix}})\}, \epsilon_{s,t}\right), \\ z_{c,t}^{\text{chg}} &= f_c\left(\{z_{i,t-\tau} | z_{i,t-\tau} \in \mathbf{Pa}(z_{c,t}^{\text{chg}})\}, \boldsymbol{\theta}_r^{\text{dyn}}, \epsilon_{c,t}\right), \\ z_{o,t}^{\text{obs}} &= f_o\left(\boldsymbol{\theta}_r^{\text{obs}}, \epsilon_{o,t}\right), \\ \mathbf{x}_t &= \mathbf{g}(\mathbf{z}_t). \end{cases} \tag{1}$$

3.2 Identifiability of Latent Causal Processes and Time-Delayed Latent Causal Relations

We define the identifiability of time-delayed latent causal processes in the representation function space in **Definition 1**. Furthermore, if the estimated latent processes can be identified at least up to permutation and component-wise invertible nonlinearities, the latent causal relations are also immediately identifiable because conditional independence relations fully characterize time-delayed causal relations in a time-delayed causally sufficient system, in which there are no latent causal confounders in the (latent) causal processes. Note that invertible component-wise transformations on latent causal processes do not change their conditional independence relations.

Definition 1 (Identifiable Latent Causal Processes)**.** *Formally let $\{\mathbf{x}_t\}_{t=1}^T$ be a sequence of observed variables generated by the true temporally causal latent processes specified by $(f_i, \boldsymbol{\theta}_r, p(\epsilon_i), \mathbf{g})$ given in Eq. 1. A learned generative model $(\hat{f}_i, \hat{\boldsymbol{\theta}}_r, \hat{p}(\epsilon_i), \hat{\mathbf{g}})$ is observationally equivalent to $(f_i, \boldsymbol{\theta}_r, p(\epsilon_i), \mathbf{g})$ if the model distribution $p_{\hat{f}, \hat{\boldsymbol{\theta}}_r, \hat{p}_\epsilon, \hat{\mathbf{g}}}(\{\mathbf{x}_t\}_{t=1}^T)$ matches the data distribution $p_{f, \boldsymbol{\theta}_r, p_\epsilon, \mathbf{g}}(\{\mathbf{x}_t\}_{t=1}^T)$ everywhere. We say latent causal processes are identifiable if observational equivalence can lead to identifiability of the latent variables up to permutation π and component-wise invertible transformation T:*

$$p_{\hat{f}_i, \hat{\boldsymbol{\theta}}_r, \hat{p}_{\epsilon_i}, \hat{\mathbf{g}}}(\{\mathbf{x}_t\}_{t=1}^T) = p_{f_i, \boldsymbol{\theta}_r, p_{\epsilon_i}, \mathbf{g}}(\{\mathbf{x}_t\}_{t=1}^T) \Rightarrow \hat{\mathbf{g}}(\mathbf{x}_t) = \mathbf{g} \circ \pi \circ T, \quad \forall \mathbf{x}_t \in \mathcal{X}, \tag{2}$$

where $\mathcal{X}$ is the observation space.

4 Identifiability Theory

We establish the identifiability theory of nonparametric time-delayed latent causal processes under three different types of distribution shifts. W.l.o.g., we consider the latent processes with maximum time lag $L = 1$. Extentions to arbitrary time lags have been discussed in Appendix S1.5. Let k be the element index of the latent space $\mathbf{z}_t$ and the latent size is n. In particular, **(1)** under fixed

causal dynamics, we leverage the distribution changes $p(z_{k,t}|\mathbf{z}_{t-1})$ for different values of $\mathbf{z}_{t-1}$; **(2)** under changing causal dynamics, we exploit the changing causal influences on $p(z_{k,t}|\mathbf{z}_{t-1}, u_r)$ under different domain u_r, and **(3)** under global observation changes, the nonstationarity $p(z_{k,t}|u_r)$ under different values of u_r is exploited. The proofs are provided in Appendix S1. The comparisons between existing theories are in Appendix S1.3.

4.1 Identifiability under Fixed Causal Dynamics

Let $\eta_{kt} \triangleq \log p(z_{k,t}|\mathbf{z}_{t-1})$. Assume that η_{kt} is twice differentiable in $z_{k,t}$ and is differentiable in $z_{l,t-1}$, $l = 1, 2, ..., n$. Note that the parents of $z_{k,t}$ may be only a subset of $\mathbf{z}_{t-1}$; if $z_{l,t-1}$ is not a parent of $z_{k,t}$, then $\frac{\partial \eta_{kt}}{\partial z_{l,t-1}} = 0$. Below we provide a *sufficient condition* for the identifiability of $\mathbf{z}_t$, followed by a discussion of specific unidentifiable and identifiable cases to illustrate how general it is.

Theorem 1 (Identifiablity under a Fixed Temporal Causal Model)**.** *Suppose there exists invertible function $\hat{\mathbf{g}}$ that maps $\mathbf{x}_t$ to $\hat{\mathbf{z}}_t$, i.e.,*

$$\hat{\mathbf{z}}_t = \hat{\mathbf{g}}(\mathbf{x}_t) \tag{3}$$

such that the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$. Let

$$\mathbf{v}_{k,t} \triangleq \left(\frac{\partial^2 \eta_{kt}}{\partial z_{k,t} \partial z_{1,t-1}}, \frac{\partial^2 \eta_{kt}}{\partial z_{k,t} \partial z_{2,t-1}}, ..., \frac{\partial^2 \eta_{kt}}{\partial z_{k,t} \partial z_{n,t-1}}\right)^\intercal, \ \mathring{\mathbf{v}}_{k,t} \triangleq \left(\frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{1,t-1}}, \frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{2,t-1}}, ..., \frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{n,t-1}}\right)^\intercal. \tag{4}$$

If for each value of $\mathbf{z}_t$, $\mathbf{v}_{1,t}, \mathring{\mathbf{v}}_{1,t}, \mathbf{v}_{2,t}, \mathring{\mathbf{v}}_{2,t}, ..., \mathbf{v}_{n,t}, \mathring{\mathbf{v}}_{n,t}$, as $2n$ vector functions in $z_{1,t-1}$, $z_{2,t-1}$, ..., $z_{n,t-1}$, are linearly independent, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.

The linear independence condition in Theorem 1 is the core condition to guarantee the identifiability of $\mathbf{z}_t$ from the observed $\mathbf{x}_t$. To make this condition more intuitive, below we consider specific unidentifiable cases, in which there is no temporal dependence in $\mathbf{z}_t$ or the noise terms in $\mathbf{z}_t$ are additive Gaussian, and two identifiable cases, in which $\mathbf{z}_t$ has additive, heterogeneous noise or follows some linear, non-Gaussian temporal process.

Let us start with two unidentifiable cases. In case N1, $\mathbf{t}_t$ is an independent and identically distributed (i.i.d.) process, i.e., there is no causal influence from any component of $\mathbf{z}_{t-1}$ to any $z_{k,t}$. In this case, $\mathbf{v}_{k,t}$ and $\mathring{\mathbf{v}}_{k,t}$ (defined in Eq. 4) are always $\mathbf{0}$ for $k = 1, 2, ..., n$, since $p(z_{k,t} \mid \mathbf{z}_{t-1})$ does not involve $\mathbf{z}_{t-1}$. So the linear independence condition is violated. In fact, this is the regular nonlinear ICA problem with i.i.d. data, and it is well-known that the underlying independent variables are not identifiable [5]. In case $\mathtt{N}_2$, all $z_{k,t}$ follow an additive noise model with Gaussian noise terms, i.e.,

$$\mathbf{z}_t = \mathbf{q}(\mathbf{z}_{t-1}) + \epsilon_t, \tag{5}$$

where $\mathbf{q}$ is a transformation and the components of the Gaussian vector ϵ_t are independent and also independent from $\mathbf{z}_{t-1}$. Then $\frac{\partial^2 \eta_{kt}}{\partial z_{k,t}^2}$ is constant, and $\frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{l,t-1}} \equiv 0$, violating the linear independence condition. In the following proposition we give some alternative solutions and verify the unidentifiability in this case.

Proposition 1 (Unidentifiability under Gaussian Noise)**.** *Suppose $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$ was generated by Eq. 5, where the components of ϵ_t are mutually independent Gaussian and also independent from $\mathbf{z}_{t-1}$. Then any $\hat{\mathbf{z}}_t = \mathbf{D}_1 \mathbf{U} \mathbf{D}_2 \cdot \mathbf{z}_t$, where $\mathbf{D}_1$ is an arbitrary non-singular diagonal matrix, $\mathbf{U}$ is an arbitrary orthogonal matrix, and $\mathbf{D}_2$ is a diagonal matrix with $\mathbb{V}ar^{-1/2}(\epsilon_{k,t})$ as its kth diagonal entry, is a valid solution to satisfy the condition that the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$.*

Roughly speaking, for a randomly chosen conditional density function $p(z_{k,t} \mid \mathbf{z}_{t-1})$ in which $z_{k,t}$ is not independent from $\mathbf{z}_{t-1}$ (i.e., there is temporal dependence in the latent processes) and which does not follow an additive noise model with Gaussian noise, the chance for its specific second- and third-order partial derivatives to be linearly dependent is slim. Now let us consider two cases in which the latent temporally processes $\mathbf{z}_t$ are naturally identifiable. First, consider case Y1, where $z_{k,t}$ follows a heterogeneous noise process, in which the noise variance depends on its parents:

$$z_{k,t} = q_k(\mathbf{z}_{t-1}) + \frac{1}{b_k(\mathbf{z}_{t-1})} \epsilon_{k,t}. \tag{6}$$

Here we assume $\epsilon_{k,t}$ is standard Gaussian and $\epsilon_{1,t}, \epsilon_{2,t}, .., \epsilon_{n,t}$ are mutually independent and independent from $\mathbf{z}_{t-1}$. $\frac{1}{b_k}$, which depends on $\mathbf{z}_{t-1}$, is the standard deviation of the noise in $z_{k,t}$. (For conciseness, we drop the argument of b_k and q_k when there is no confusion.) Note that in this model, if q_k is 0 for all $k = 1, 2, ..., n$, it reduces to a multiplicative noise model. The identifiability result of $\mathbf{z}_t$ is established in the following corollary.

Corollary 1 (Identifiability under Heterogeneous Noise)**.** *Suppose $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$ was generated according to Eq. 6, and Eq. 3 holds true. If $b_k \cdot \frac{\partial b_k}{\partial \mathbf{z}_{t-1}}$ and $b_k \cdot \frac{\partial b_k}{\partial \mathbf{z}_{t-1}}(z_{k,t} - q_k) - b_k^2 \cdot \frac{\partial q_k}{\partial \mathbf{z}_{t-1}}$, with $k = 1, 2, ..., n$, which are in total $2n$ function vectors in $\mathbf{z}_{t-1}$, are linearly independent, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.*

Let us then consider another special case, denoted by Y2, with a linear, non-Gaussian temporal model for $\mathbf{z}_t$: the latent processes follow Eq. 5, with $\mathbf{q}$ being a linear transformation and $\epsilon_{k,t}$ following a particular class of non-Gaussian distributions. The following corollary shows that $\mathbf{z}_t$ is identifiable as long as each $z_{k,t}$ receives causal influences from some components of $\mathbf{z}_{t-1}$.

Corollary 2 (Identifiability under a Specific Linear, Non-Gaussian Model for Latent Processes)**.** *Suppose $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$ was generated according to Eq. 5, in which $\mathbf{q}$ is a linear transformation and for each $z_{k,t}$, there exists at least one k' such that $c_{k,k'} \triangleq \frac{\partial z_{k,t}}{\partial z_{k',t-1}} \neq 0$. Assume the noise term $\epsilon_{k,t}$ follows a zero-mean generalized normal distribution:*

$$p(\epsilon_{k,t}) \propto e^{-\lambda|\epsilon_{k,t}|^\beta}, \text{ with positive } \lambda \text{ and } \beta > 2 \text{ and } \beta \neq 3. \tag{7}$$

If Eq. 3 holds, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of permuted $\hat{\mathbf{z}}_t$.

4.2 Further Benefits from Changing Causal Influences

Let $\eta_{kt}(u_r) \triangleq \log p(z_{k,t}|\mathbf{z}_{t-1}, u_r)$ where $r = 1, ..., m$. [14] established the identifiability of the latent temporal causal processes $\mathbf{z}_t$ in certain nonstationary cases, under the condition that the noise term in each $z_{k,t}$, relative to its parents in $\mathbf{z}_{t-1}$, changes across m contexts corresponding to $\mathbf{u} = u_1, u_2, ..., u_m$. Here we show that the identifiability result shown in the previous section can further benefit from nonstationarity of the causal model, and that our identifiability condition is generally much weaker than that in [14]: we allow changes in the noise term or causal influence on $z_{k,t}$ from its parents in $\mathbf{z}_{t-1}$, and our "sufficient variability" condition is just a necessary condition for that in [14] because of the additional information that one can leverage. Let $\mathbf{v}_{k,t}(u_r)$ be $\mathbf{v}_{k,t}$, which is defined in Eq. 4, in the u_r context. Similarly, Let $\mathring{\mathbf{v}}_{k,t}(u_r)$ be $\mathring{\mathbf{v}}_{k,t}$ in the u_r context. Let

$$\mathbf{s}_{k,t} \triangleq \left(\mathbf{v}_{k,t}(u_1)^\intercal, ..., \mathbf{v}_{k,t}(u_m)^\intercal, \Delta_2^2, ..., \Delta_m^2\right)^\intercal, \ \mathring{\mathbf{s}}_{k,t} \triangleq \left(\mathring{\mathbf{v}}_{k,t}(u_1)^\intercal, ..., \mathring{\mathbf{v}}_{k,t}(u_m)^\intercal, \Delta_2, ..., \Delta_m\right)^\intercal, \tag{8}$$

where $\Delta_r^2 = \frac{\partial^2 \eta_{kt}(u_r)}{\partial z_{k,t}^2} - \frac{\partial^2 \eta_{kt}(u_{r-1})}{\partial z_{k,t}^2}$ and $\Delta_r = \frac{\partial \eta_{kt}(u_r)}{\partial z_{k,t}} - \frac{\partial \eta_{kt}(u_{r-1})}{\partial z_{k,t}}$. As provided below, in our case, the identifiablity of $\mathbf{z}_t$ is guaranteed by the linear independence of the whole function vectors $\mathbf{s}_{k,t}$ and $\mathring{\mathbf{s}}_{k,t}$, with $k = 1, 2, ..., n$. However, the identifiability result in [14] relies on the linear independence of only the last $m - 1$ components of $\mathbf{s}_{k,t}$ and $\mathring{\mathbf{s}}_{k,t}$ with $k = 1, 2, ..., n$; this linear independence is generally a much stronger and restricted condition.

Theorem 2 (Identifiability under Changing Causal Dynamics)**.** *Suppose $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$ and that the conditional distribution $p(z_{k,t} \mid \mathbf{z}_{t-1})$ may change across m values of the context variable $\mathbf{u}$, denoted by u_1, u_2, ..., u_m. Suppose the components of $\mathbf{z}_t$ are mutually independent conditional on $\mathbf{z}_{t-1}$ in each context. Assume that the components of $\hat{\mathbf{z}}_t$ produced by Eq. 3 are also mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$. If the $2n$ function vectors $\mathbf{s}_{k,t}$ and $\mathring{\mathbf{s}}_{k,t}$, with $k = 1, 2, ..., n$, are linearly independent, then $\hat{\mathbf{z}}_t$ is a permuted invertible component-wise transformation of $\mathbf{z}_t$.*

Theorem 3 (Identifiability under Observation Changes)**.** *Suppose $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$ and that the conditional distribution $p(z_{k,t} \mid \mathbf{u})$ may change across m values of the context variable $\mathbf{u}$, denoted by u_1, u_2, ..., u_m. Suppose the components of $\mathbf{z}_t$ are mutually independent conditional on $\mathbf{u}$ in each context. Assume that the components of $\hat{\mathbf{z}}_t$ produced by Eq. 3 are also mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$. If the $2n$ function vectors $\mathbf{s}_{k,t}$ and $\mathring{\mathbf{s}}_{k,t}$, with $k = 1, 2, ..., n$, are linearly independent, then $\hat{\mathbf{z}}_t$ is a permuted invertible component-wise transformation of $\mathbf{z}_t$.*

Corollary 3 (Identifiability under Modular Distribution Shifts)**.** *Assume the data generating process in Eq. 1. If the three partitioned latent components $\mathbf{z}_t = (\mathbf{z}_t^{fix}, \mathbf{z}_t^{chg}, \mathbf{z}_t^{obs})$ respectively satisfy the conditions in **Theorem** 1, **Theorem** 2, and **Theorem** 3, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.*

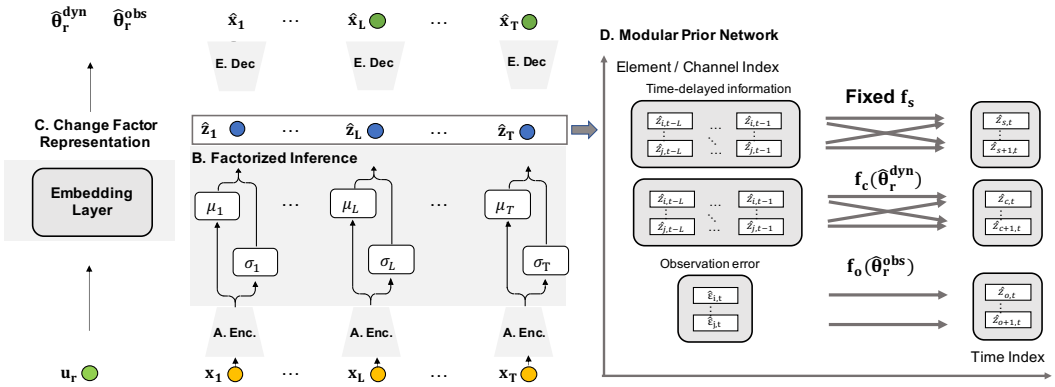

Figure 2: **Framework overview:** Encoder (A) and Decoder (E) using MLP or CNN for specific data types; (B) Factorized inference network that approximates the posteriors of latent variables $\hat{\mathbf{z}}_t$, (C) Change factor representation learning with embedding layer, and (D) Modular prior network that factorizes distribution changes, i.e., fixed causal dynamics, changing causal dynamics, and global changes, into three partitioned latent causal processes constraint (Thm 1,2 and 3).

5 Our Approach

5.1 LCD-NM: Learning Latent Causal Dynamics from Nonlinear Mixtures

Given our identifiability results, we propose LCD-NM framework to estimate the latent causal dynamics under modular distribution shifts, by extending Sequential Variational Auto-Encoders [31] with tailored modules to model different distribution shifts, and enforcing the conditions in Sec. 4 as constraints. We give the estimation procedure of the latent causal dynamics model in Eq. 1. The model architecture is showcased in Fig. 2. The framework has the following three major components. The implementation details are in Appendix S3.3. Specifically, we leverage the partitioned estimated latent subspaces $\hat{\mathbf{z}}_t = (\hat{\mathbf{z}}_t^{\text{fix}}, \hat{\mathbf{z}}_t^{\text{chg}}, \hat{\mathbf{z}}_t^{\text{obs}})$ and model their distribution changes in conditional transition priors. We use $[f_s, f_c, f_o]$ to capture causal relations as in Eq. 1, where f_s captures fixed causal influences, f_c for changing causal influences, and f_o for observation changes. Accordingly, we learn $[\hat{f}_s^{-1}, \hat{f}_c^{-1}, \hat{f}_o^{-1}]$ to output random process noise from the estimated direct cause (lagged states $\hat{\mathbf{z}}_{\text{Hx}}$) and effect (current states) variables $\hat{\mathbf{z}}_t$.

Change Factor Representation We learn to embed domain index $\mathbf{u}_r$ into low-dimensional change factors $(\hat{\boldsymbol{\theta}}_r^{\text{dyn}}, \hat{\boldsymbol{\theta}}_r^{\text{obs}})$ in Fig. 2C and insert them as external inputs to the (inverse) dynamics function $\hat{f}_c^{-1}(\hat{\boldsymbol{\theta}}_r^{\text{dyn}})$, or the observation bijector $\hat{f}_o^{-1}(\hat{\boldsymbol{\theta}}_r^{\text{obs}})$, respectively. And hence the distribution shifts are captured and utilized in the implementation.

Modular Prior Network We follow standard conditional normalizing flow formulation [32, 14]. Let $\mathbf{z}_{\text{Hx}}$ denote the lagged latent variables up to maximum time lag L. In particular, for **1) fixed causal dynamics processes** $\hat{\mathbf{z}}_t^{\text{fix}}$, their transition priors are obtained by first learning inverse transition functions f_s^{-1} that take the estimated latent variables and output random noise terms, and applying the change of variables formula to the transformation: $p(\hat{z}_{s,t}^{\text{fix}}|\hat{\mathbf{z}}_{\text{Hx}}) = p_{\epsilon_s}\left(\hat{f}_s^{-1}(\hat{z}_{s,t}^{\text{fix}}, \hat{\mathbf{z}}_{\text{Hx}})\right)\left|\frac{\partial \hat{f}_s^{-1}}{\partial \hat{z}_{s,t}^{\text{fix}}}\right|$; for **2) changing causal dynamics**, we evaluate $p(\hat{z}_{c,t}^{\text{chg}}|\hat{\mathbf{z}}_{\text{Hx}}, \mathbf{u}_r) = p_{\epsilon_c}\left(\hat{f}_c^{-1}(\hat{z}_{c,t}^{\text{chg}}, \hat{\mathbf{z}}_{\text{Hx}}, \hat{\boldsymbol{\theta}}_r^{\text{dyn}})\right)\left|\frac{\partial \hat{f}_c^{-1}}{\partial \hat{z}_{c,t}^{\text{chg}}}\right|$ by learning a holistic inverse dynamics $\hat{f}_c^{-1}$ that takes the estimated change factors for dynamics $\hat{\boldsymbol{\theta}}_r^{\text{dyn}}$ as inputs, and similarly for **3) observation changes** $\hat{\mathbf{z}}_t^{\text{obs}}$, we learn to project them to invariant noise terms by $\hat{f}_o^{-1}$ which takes the change factors $\boldsymbol{\theta}_r^{\text{obs}}$ as arguments, and obtains $p(\hat{z}_{o,t}^{\text{obs}}|\mathbf{u}_r) = p_{\epsilon_o}\left(\hat{f}_o^{-1}(\hat{z}_{o,t}^{\text{obs}}, \hat{\boldsymbol{\theta}}_r^{\text{obs}})\right)\left|\frac{\partial \hat{f}_o^{-1}}{\partial \hat{z}_{o,t}^{\text{obs}}}\right|$ as the prior. **Conditional independence of the estimated latent variables** $p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{\text{Hx}})$ is enforced by summing up all estimated component densities when obtaining the joint $p(\mathbf{z}_t|\mathbf{z}_{\text{Hx}}, \mathbf{u})$ in Eq. 9. Given that the Jacobian is lower-triangular, we can compute its determinant as the product of diagonal terms. The detailed derivations are given in Appendix S3.1.

$$\log p\left(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{\text{Hx}}, \mathbf{u}_r\right) = \underbrace{\sum_{i=1}^{n} \log p(\hat{\epsilon}_i|\mathbf{u}_r)}_{\text{Conditional indepdence}} + \underbrace{\sum_{i=1}^{n} \log \left|\frac{\partial \hat{f}_i^{-1}}{\partial \hat{z}_{it}}\right|}_{\text{Low-triangular Jacobian matrix}} \tag{9}$$

Factorized Inference We infer the posteriors of each time step $q(\hat{\mathbf{z}}_t|\mathbf{x}_t)$ using only the observation at that time step, because in Eq. 1, $\mathbf{x}_t$ preserves all the information of the current system states so the joint probability $q(\hat{\mathbf{z}}_{1:T}|\mathbf{x}_{1:T})$ can be factorized into product of these terms. We approximate the posterior $q(\hat{\mathbf{z}}_t|\mathbf{x}_t)$ with an isotropic Gaussian with mean and variance from the inference network.

Optimization We train LCD-NM using the ELBO objective $\mathcal{L}_{\text{ELBO}} = \frac{1}{N}\sum_{i\in N} \mathcal{L}_{\text{Recon}} - \beta\mathcal{L}_{\text{KLD}}$, in which we use mean-squared error (MSE) for the reconstruction likelihood $\mathcal{L}_{\text{Recon}}$. The KL divergence $\mathcal{L}_{\text{KLD}}$ is estimated via a sampling approach since with a learned nonparametric modular transition prior, the distribution does not have an explicit form. Specifically, we obtain the log-likelihood of the posterior, evaluate the prior – $\log p\left(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{\text{Hx}}, \mathbf{u}_r\right)$ in Eq. 9, and compute their mean difference in a batch $\mathcal{B}$ as the KL loss: $\mathcal{L}_{\text{KLD}} = \sum_{i\in\mathcal{B}} \mathbb{E}_{\hat{\mathbf{z}}_t\sim q\left(\hat{\mathbf{z}}_t^{(i)}|\mathbf{x}_t^{(i)}\right)} \log q(\hat{\mathbf{z}}_t|\mathbf{x}_t) - \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{\text{Hx}}, \mathbf{u}_r)$.

5.2 Causal Visualization

For visualization purposes, when the underlying latent processes have sparse causal relations, we fit LassoNet [33] on the latent processes recovered by LCD-NM to interpret the causal relations. Specifically, we fit LassoNet to predict $\hat{\mathbf{z}}_t$ using the estimated history information $\hat{\mathbf{z}}_{\text{Hx}} = \{\hat{\mathbf{z}}_{t-\tau}\}_{\tau=1}^{L}$ up to maximum time lag L. Note that this postprocessing step is optional – the latent causal relations have already been captured in the learned transition functions in LCD-NM. Also, our identifiability conditions do not rely on the sparsity of causal relations in the latent processes.

6 Experimental Results

We evaluate the identifiability results of *LCD-NM* on a number of simulated and real-world time-series datasets. We first introduce the evaluation metrics and baselines. **(1) Evaluation Metrics.** To evaluate the identifiability of the latent variables, we compute Mean Correlation Coefficient (MCC) on the test dataset. MCC is a standard metric in the ICA literature for continuous variables which measure the identifiability of the learned latent causal processes. MCC is close to 1 when latent variables are identifiable up to permutation and component-wise invertible transformation in the noiseless case. **(2) Baselines.** Nonlinear ICA methods are used: **(1)** BetaVAE [34] which ignores both history and nonstationarity information; **(2)** iVAE [9] and TCL [6] which leverage nonstationarity to establish identifiability but assumes independent factors, and **(3)** SlowVAE [12] and PCL [7] which exploit temporal constraints but assume independent sources and stationary processes, and **(4)** LEAP [14] which assumes nonstationary, causal processes but only models nonstationary noise. Two other disentangled deep state-space models with nonlinear dynamics models: Kalman VAE (KVAE [35]) and Deep Variational Bayes Filters (DVBF [36]), are also used for comparisons.

Table 2: MCC scores and their standard deviations for the three simulation settings over 3 random seeds. Note: The symbol "–" represents that this method is not applicable to this dataset.

Experiment Settings	Method								
	LCD-NM	LEAP	SlowVAE	PCL	i-VAE	TCL	betaVAE	KVAE	DVBF
Fixed	**0.954 ±0.009**	–	0.411 ±0.022	0.516 ±0.043	–	–	0.353 ±0.001	0.832 ±0.038	0.778 ±0.045
Changing	**0.958 ±0.017**	0.726 ±0.187	0.511 ±0.062	0.599 ±0.041	0.581 ±0.083	0.399 ±0.021	0.523 ±0.009	0.711 ±0.062	0.648 ±0.071
Modular	**0.993 ±0.001**	0.657 ±0.108	0.406 ±0.045	0.564 ±0.049	0.557 ±0.005	0.297 ±0.078	0.433 ±0.045	0.632 ±0.048	0.678 ±0.074

6.1 Simulated Experiments

We generate synthetic datasets that satisfy our identifiability conditions in the theorems following the procedures in Appendix S2.1.1. As in Table 2, our framework can recover the latent processes under fixed dynamics (heterogeneous noise model), under changing causal dynamics, and under modular distribution shifts with high MCCs (>0.95). The baselines that do not exploit history (i.e., βVAE, i-VAE, TCL), with independent source assumptions (SlowVAE, PCL), considers limited nonstationary cases (LEAP) distort the identifiability results. KVAE and DVBF achieve MCCs (0.8) under fixed dynamics but distorts the identifiability under changing dynamics and modular shift setting because they don't model changing causal relations and observation changes.

6.2 Real-world Applications

Video Data – Modified Cartpole Environment We evaluate LCD-NM on the modified cartpole [30] video dataset and compare the performances with the baselines. Modified Cartpole is a nonlinear dynamical system with cart positions x_t and pole angles θ_t as the true state variables. The dataset descriptions are in Appendix S2.1.2. We use 6 source domains with different gravity values $g = \{5, 10, 15, 20, 25, 30\}$. Together with the 2 discrete actions (i.e., left and right), we have 12 segments of data with changing causal dynamics. We fit LCD-NM with two-dimensional change factors $\boldsymbol{\theta}_r^{\text{dyn}}$. We set the latent size $n = 8$ and the lag number $L = 2$. In Fig. 3, the latent causal processes are recovered, as seen from (a) high MCC for the latent causal processes; (b) the latent factors are estimated up to component-wise transformation; (c) LCD-NM outperforms the baselines and (d) the latent traversals confirm the two recovered latent variables correspond to the position and pole angle.

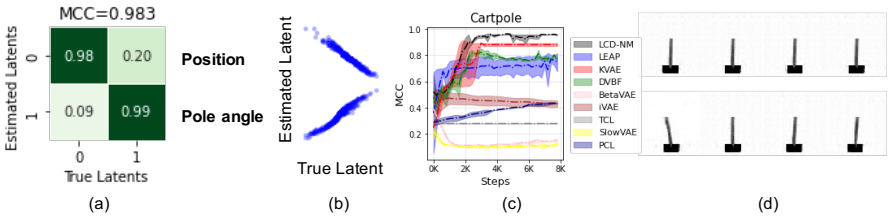

Figure 3: Modified Cartpole results: (a) MCC for causally-related factors; (b) scatterplots between estimated and true factors; (c) baseline comparisons, and (d) latent traversal on a fixed video frame;.

Motion Capture Data – CMU-Mocap We experimented with another real-world motion capture dataset (CMU-Mocap). The dataset descriptions are in Appendix S2.1.2. We fit LCD-NM with 11 trials of motion capture data for subject #8. The 11 trials contain walk cycles with very different styles (e.g., slow walk, stride). We set latent size $n = 3$ and lag number $L = 2$. The differences between trials are captured through learning the 2-dimensional change factors for each trial. In Fig. 4(a), the learned change factors group similar walk styles into clusters; in Panel (c), three latent variables (which seem to be pitch, yaw, roll rotations) are found to explain most of the variances of human walk cycles. The learned latent coordinates (Panel b) show smooth cyclic patterns with differences among different walking styles. For the discovered skeleton (Panel d), roll and pitch of walking are found to be causally-related while yaw has independent dynamics.

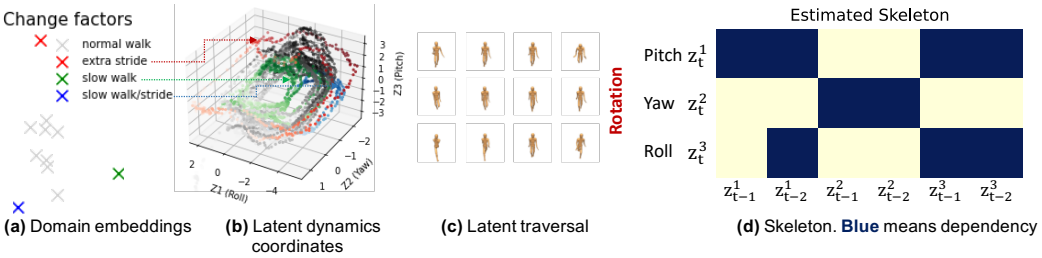

Figure 4: CMU-Mocap results (Subject #8): (a) learned change factors; (b) latent coordinates dynamics for 11 trials; (c)) latent traversal by rendering the reconstructed point clouds into the video frame; (d) estimated causal skeleton (blue means dependency).

7 Conclusion

In this paper, without relying on parametric or distribution assumptions, we established identifiability theories for nonparametric latent causal processes from their observed nonlinear mixtures in stationary environments and under unknown distribution shifts. The basic limitation of this work is that the underlying latent processes are assumed to have no instantaneous causal relations but only time-delayed influences. If the time resolution of the observed time series is the same as the causal frequency (or even higher), this assumption naturally holds. However, if the resolution of the time series is much lower, then it is usually violated and one has to find a way to deal with instantaneous causal relations. On the other hand, it is worth mentioning that the assumptions are generally testable; if the assumptions are actually violated, one can see that the results produced by our method may not be reliable. Extending our theories and framework to address the issue of instantaneous dependency or instantaneous causal relations will be one line of our future work.

References

[1] Carlo Berzuini, Philip Dawid, and Luisa Bernardinell. *Causality: Statistical perspectives and applications*. John Wiley & Sons, 2012.

[2] Eric Ghysels, Jonathan B Hill, and Kaiji Motegi. Testing for granger causality with mixed frequency data. *Journal of Econometrics*, 192(1):207–230, 2016.

[3] Karl Friston. Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS biology*, 7(2):e1000033, 2009.

[4] 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 *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[5] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[6] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016.

[7] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[8] 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.

[9] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

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

[11] Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

[12] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

[13] Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *arXiv preprint arXiv:2106.09620*, 2021.

[14] Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *International Conference on Learning Representations*, 2022.

[15] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[16] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[17] Yunzhu Li, Antonio Torralba, Animashree Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *arXiv preprint arXiv:2007.00631*, 2020.

[18] Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128, 2010.

[19] Kevin P Murphy et al. Dynamic bayesian networks. *Probabilistic Graphical Models, M. Jordan*, 7:431, 2002.

[20] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

[21] Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pages 23–47. PMLR, 2018.

[22] Daniel Malinsky and Peter Spirtes. Learning the structure of a nonstationary vector autoregression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2986–2994. PMLR, 2019.

[23] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[24] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality. *arXiv preprint arXiv:1802.05842*, 2018.

[25] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. *arXiv preprint arXiv:2006.10833*, 2020.

[26] 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.

[27] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. icitris: Causal representation learning for instantaneous temporal effects. *arXiv preprint arXiv:2206.06169*, 2022.

[28] Petar Stojanov, Mingming Gong, Jaime Carbonell, and Kun Zhang. Data-driven approach to multiple-source domain adaptation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3487–3496. PMLR, 2019.

[29] 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.

[30] Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. Adarl: What, where, and how to adapt in transfer reinforcement learning. *arXiv preprint arXiv:2107.02729*, 2021.

[31] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *arXiv preprint arXiv:1803.02991*, 2018.

[32] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[33] Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

[34] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[35] Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. *Advances in neural information processing systems*, 30, 2017.

[36] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick Van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv preprint arXiv:1605.06432*, 2016.

[37] Alessio Spantini, Daniele Bigoni, and Youssef M. Marzouk. Inference via low-dimensional couplings. *J. Mach. Learn. Res.*, 19:66:1–66:71, 2018.

[38] Kiyotoshi Matsuoka, Masahiro Ohoya, and Mitsuru Kawamoto. A neural net for blind separation of nonstationary signals. *Neural networks*, 8(3):411–419, 1995.

[39] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

[40] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.

[41] Clive WJ Granger. Implications of aggregation with common factors. *Econometric Theory*, 3(02):208–222, 1987.

[42] M. Gong*, K. Zhang*, D. Tao, P. Geiger, and B. Schölkopf. Discovering temporal causal relations from subsampled data. In *Proc. 32th International Conference on Machine Learning (ICML 2015)*, 2015.

[43] D. Danks and S. Plis. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, pages 1–10, 2013.

[44] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao. Causal discovery from temporally aggregated time series. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI'17)*, 2017.

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

[46] J. Adams, N. R. Hansen, and K. Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]
(b) Did you mention the license of the assets? [Yes]
(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplement to

"Causal Disentanglement for Time Series"

We divide the supplement into four sections. In S1, we give the proofs of the theorems and compare our theory with the existing nonlinear ICA work in terms of mathematical formulation and assumptions. In S2, we provide the descriptions of the datasets used in this paper and evaluation metrics. In S3, we first derive the computation of the modular prior likelihood. The model architecture, hyperparameter selection and training are also explained in detail.

S1 Identifiability Theory **14**

S1.1 Proof for Theorem 1 14

S1.2 Proof for Theorem 2 and 3 18

S1.3 Comparisons with Existing Nonlinear ICA Theories 19

S1.4 Discussion of the Assumptions 20

S1.4.1 Linear Independence Condition 20

S1.4.2 Independent Noise Condition 21

S1.4.3 Causal Influences between Observed Variables 22

S1.5 Extension to Multiple Time Lags 22

S2 Experiment Settings **22**

S2.1 Datasets 22

S2.1.1 Synthetic Dataset Generation 22

S2.1.2 Real-world Dataset 23

S2.2 Mean Correlation Coefficient 24

S3 Implementation Details **24**

S3.1 Modular Prior Likelihood Derivation 24

S3.2 Comparisons with AdaRL [30] 25

S3.3 Network Architecture 25

S3.4 Hyperparameter and Training 25

S4 Additional Experiment Results **26**

S1 Identifiability Theory

The observed variables were generated according to :

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t), \tag{10}$$

in which $\mathbf{g}$ is invertible, and z_{it}, as the ith component of $\mathbf{z}_t$, is generated by (some) components of $\mathbf{z}_{t-1}$ and noise E_{it}. $E_{1t}, E_{2t}, ..., E_{nt}$ are mutually independent. In other words, the components of $\mathbf{z}_t$ are mutually independent conditional on $\mathbf{z}_{t-1}$. Let $\eta_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_{t-1})$. Assume that $\eta_k(t)$ is twice differentiable in z_{kt} and is differentiable in $z_{l,t-1}$, $l = 1, 2, ..., n$. Note that the parents of z_{kt} may be only a subset of $\mathbf{z}_{t-1}$; if $z_{l,t-1}$ is not a parent of z_{kt}, then $\frac{\partial \eta_k}{\partial z_{l,t-1}} = 0$.

S1.1 Proof for Theorem 1

Theorem S1 (Identifiablity under a Fixed Temporal Causal Model)**.** *Suppose there exists invertible function* $\mathbf{f}$*, which is the estimated mixing function (i.e., we use* $\mathbf{f}$ *and* $\hat{\mathbf{g}}$ *interchangeably in Appendix)*

that maps $\mathbf{x}_t$ to $\hat{\mathbf{z}}_t$, i.e.,

$$\hat{\mathbf{z}}_t = \mathbf{f}(\mathbf{x}_t) \tag{11}$$

such that the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$. Let

$$\begin{aligned} \mathbf{v}_{k,t} &\triangleq \Big(\frac{\partial^2 \eta_{kt}}{\partial z_{k,t} \partial z_{1,t-1}}, \frac{\partial^2 \eta_{kt}}{\partial z_{k,t} \partial z_{2,t-1}}, ..., \frac{\partial^2 \eta_{kt}}{\partial z_{k,t} \partial z_{n,t-1}}\Big)^\intercal \\ \mathring{\mathbf{v}}_{k,t} &\triangleq \Big(\frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{1,t-1}}, \frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{2,t-1}}, ..., \frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{n,t-1}}\Big)^\intercal. \end{aligned} \tag{12}$$

If for each value of $\mathbf{z}_t$, $\mathbf{v}_{1t}, \mathring{\mathbf{v}}_{1t}, \mathbf{v}_{2t}, \mathring{\mathbf{v}}_{2t}, ..., \mathbf{v}_{nt}, \mathring{\mathbf{v}}_{nt}$, as $2n$ vector functions in $z_{1,t-1}$, $z_{2,t-1}$, ..., $z_{n,t-1}$, are linearly independent, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.

Proof. Combining Eq. 10 and Eq. 11 gives $\mathbf{z}_t = \mathbf{g}^{-1}(\mathbf{f}^{-1}(\hat{\mathbf{z}}_t)) = \mathbf{h}(\hat{\mathbf{z}}_t)$, where $\mathbf{h} \triangleq \mathbf{g}^{-1} \circ \mathbf{f}^{-1}$. Since both $\mathbf{f}$ and $\mathbf{g}$ are invertible, $\mathbf{h}$ is invertible. Let $\mathbf{H}_t$ be the Jacobian matrix of the transformation $h(\hat{\mathbf{z}}_t)$, and denote by $\mathbf{H}_{kit}$ its (k, i)th entry.

First, it is straightforward to see that if the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$, then for any $i \neq j$, $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $\hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\})$. Mutual independence of the components of $\hat{\mathbf{z}}_t$ conditional on $\hat{\mathbf{z}}_{t-1}$ implies that $\hat{z}_{it}$ is independent from $\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}$ conditional on $\hat{\mathbf{z}}_{t-1}$, i.e.,

$$p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1}) = p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\})).$$

At the same time, it also implies $\hat{z}_{it}$ is independent from $\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\}$ conditional on $\hat{\mathbf{z}}_{t-1}$, i.e.,

$$p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1}) = p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\})).$$

Combining the above two equations gives $p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\})) = p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}))$, i.e., for $i \neq j$, $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $\hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\})$.

We then make use of the fact that if $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $\hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\})$, then

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0,$$

assuming the cross second-order derivative exists [37]. Since $p(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1}) = p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}) p(\hat{\mathbf{z}}_{t-1})$ while $p(\hat{\mathbf{z}}_{t-1})$ does not involve $\hat{z}_{it}$ or $\hat{z}_{jt}$, the above equality is equivalent to

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0. \tag{13}$$

The Jacobian matrix of the mapping from $(\mathbf{x}_{t-1}, \hat{\mathbf{z}}_t)$ to $(\mathbf{x}_{t-1}, \mathbf{z}_t)$ is $\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ * & \mathbf{H}_t \end{bmatrix}$, where $*$ stands for a matrix, and the (absolute value of the) determinant of this Jacobian matrix is $|\mathbf{H}_t|$. Therefore $p(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1}) = p(\mathbf{z}_t, \mathbf{x}_{t-1}) \cdot |\mathbf{H}_t|$. Dividing both sides of this equation by $p(\mathbf{x}_{t-1})$ gives

$$p(\hat{\mathbf{z}}_t \,|\, \mathbf{x}_{t-1}) = p(\mathbf{z}_t \,|\, \mathbf{x}_{t-1}) \cdot |\mathbf{H}_t|. \tag{14}$$

Since $p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}) = p(\mathbf{z}_t \,|\, \mathbf{g}(\mathbf{z}_{t-1})) = p(\mathbf{z}_t \,|\, \mathbf{x}_{t-1})$ and similarly $p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}) = p(\hat{\mathbf{z}}_t \,|\, \mathbf{x}_{t-1})$, Eq. 14 tells us

$$\log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}) = \log p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}) + \log |\mathbf{H}_t| = \sum_{k=1}^{n} \eta_{kt} + \log |\mathbf{H}_t|. \tag{15}$$

Its partial derivative w.r.t. $\hat{z}_{it}$ is

$$\begin{aligned} \frac{\partial \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it}} &= \sum_{k=1}^{n} \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \frac{\partial z_{kt}}{\partial \hat{z}_{it}} - \frac{\partial \log |\mathbf{H}_t|}{\partial \hat{z}_{it}} \\ &= \sum_{k=1}^{n} \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \mathbf{H}_{kit} - \frac{\partial \log |\mathbf{H}_t|}{\partial \hat{z}_{it}}. \end{aligned}$$

Its second-order cross derivative is

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{n} \Big(\frac{\partial^2 \eta_{kt}}{\partial z_{kt}^2} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) - \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}}. \tag{16}$$

The above quantity is always 0 according to Eq. 13. Therefore, for each $l = 1, 2, ..., n$ and each value $z_{l,t-1}$, its partial derivative w.r.t. $z_{l,t-1}$ is always 0. That is,

$$\frac{\partial^3 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt} \partial z_{l,t-1}} = \sum_{k=1}^{n} \Big(\frac{\partial^3 \eta_{kt}}{\partial z_{kt}^2 \partial z_{l,t-1}} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial^2 \eta_{kt}}{\partial z_{kt} \partial z_{l,t-1}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) \equiv 0, \tag{17}$$

where we have made use of the fact that entries of $\mathbf{H}_t$ do not depend on $z_{l,t-1}$.

If for any value of $\mathbf{z}_t$, $\mathbf{v}_{1t}, \mathring{\mathbf{v}}_{1t}, \mathbf{v}_{2t}, \mathring{\mathbf{v}}_{2t}, ..., \mathbf{v}_{nt}, \mathring{\mathbf{v}}_{nt}$ are linearly independent, to make the above equation hold true, one has to set $\mathbf{H}_{kit}\mathbf{H}_{kjt} = 0$ or $i \neq j$. That is, in each row of $\mathbf{H}_t$ there is only one non-zero entry. Since h is invertible, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$. □

The linear independence condition in Theorem S1 is the core condition to guarantee the identifiability of $\mathbf{z}_t$ from the observed $\mathbf{x}_t$. Roughly speaking, for a randomly chosen conditional density function $p(z_{kt} \,|\, \mathbf{z}_{t-1})$, the chance for this constraint to hold on its second- and third-order partial derivatives is slim. For illustrative purposes, below we make this claim more precise, by considering a specific unidentifiable case, in which the noise terms in $\mathbf{z}_t$ are additive Gaussian, and two identifiable cases, in which $\mathbf{z}_t$ has additive, heterogeneous noise or follows some linear, non-Gaussian temporal process.

Let us start with an unidentifiable case. If all z_{kt} follow the additive noise model with Gaussian noise terms, i.e.,

$$\mathbf{z}_t = \mathbf{q}(\mathbf{z}_{t-1}) + \mathbf{E}_t, \tag{18}$$

where $\mathbf{q}$ is a transformation and the components of the Gaussian vector $\mathbf{E}_t$ are independent and also independent from $\mathbf{z}_{t-1}$. Then $\frac{\partial^2 \eta_{kt}}{\partial z_{kt}^2}$ is constant, and $\frac{\partial^3 \eta_{kt}}{\partial z_{kt}^2 \partial z_{l,t-1}} \equiv 0$, violating the linear independence condition. In the following proposition we give some alternative solutions and verify the unidentifiability in this case.

Proposition S1 (Unidentifiability under Gaussian noise)**.** *Suppose $\mathbf{x}_t$ was generated according to Eq. 10 and Eq. 18, where the components of $\mathbf{E}_t$ are mutually independent Gaussian and also independent from $\mathbf{z}_{t-1}$. Then any $\hat{\mathbf{z}}_t = \mathbf{D}_1 \mathbf{U} \mathbf{D}_2 \cdot \mathbf{z}_t$, where $\mathbf{D}_1$ is an arbitrary non-singular diagonal matrix, $\mathbf{U}$ is an arbitrary orthogonal matrix, and $\mathbf{D}_2$ is a diagonal matrix with $\mathbb{V}ar^{-1/2}(E_{kt})$ as its kth diagonal entry, is a valid solution to satisfy the condition that the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$.*

Proof. In this case we have

$$\hat{\mathbf{z}}_t = \mathbf{D}_1 \mathbf{U} \mathbf{D}_2 \cdot \mathbf{q}(\mathbf{z}_{t-1}) + \mathbf{D}_1 \mathbf{U} \mathbf{D}_2 \cdot \mathbf{E}_t.$$

It is easy to verify that the components of $\mathbf{D}_1 \mathbf{U} \mathbf{D}_2 \cdot \mathbf{E}_t$ are mutually independent and are independent from $\mathbf{D}_1 \mathbf{U} \mathbf{D}_2 \cdot \mathbf{q}(\mathbf{z}_{t-1})$. As a consequence, $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$. □

Now let us consider some cases in which the latent temporally processes $\mathbf{z}_t$ are naturally identifiable under some technical conditions. Let us first consider the case where z_{kt} follow a heterogeneous noise process, in which the noise variance depends on its parents:

$$a_{kt} = q_k(\mathbf{z}_{t-1}) + \frac{1}{b_k(\mathbf{z}_{t-1})} E_{kt}. \tag{19}$$

Here we assume E_{kt} is standard Gaussian and $E_{1t}, E_{2t}, .., E_{nt}$ are mutually independent and independent from $\mathbf{z}_{t-1}$. $\frac{1}{b_k}$, which depends on $\mathbf{z}_{t-1}$, is the standard deviation of the noise in z_{kt}. (For conciseness, we drop the argument of b_k and q_k when there is no confusion.) Note that in this model, if q_k is 0 for all $k = 1, 2, ..., n$, it reduces to a multiplicative noise model. The identifiability result of $\mathbf{z}_t$ is established in the following proposition.

Corollary S1 (Identifiablity under Heterogeneous Noise)**.** *Suppose $\mathbf{x}_t$ was generated according to Eq. 10 and Eq. 19. Suppose Eq. 11 holds true. If $b_k \cdot \frac{\partial b_k}{\partial \mathbf{z}_{t-1}}$ and $b_k \cdot \frac{\partial b_k}{\partial \mathbf{z}_{t-1}}(z_{kt} - q_k) - b_k^2 \cdot \frac{\partial q_k}{\partial \mathbf{z}_{t-1}}$, with $k = 1, 2, ..., n$, which are in total $2n$ function vectors in $\mathbf{z}_{t-1}$, are linearly independent, then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.*

Proof. Under the assumptions, one can see that

$$\eta_{kt} = \log p(z_{kt} \,|\, \mathbf{z}_{t-1}) = -\frac{1}{2}\log(2\pi) + \log b_k - \frac{b_k^2}{2}(z_{kt} - q_k)^2.$$

Consequently, one can find

$$\begin{aligned}
\frac{\partial^3 \eta_{kt}}{\partial z_{kt}^2 \partial z_{l,t-1}} &= -b_k \cdot \frac{\partial b_k}{\partial z_{l,t-1}}, \\
\frac{\partial^2 \eta_{kt}}{\partial z_{kt} \partial z_{l,t-1}} &= -b_k \cdot \frac{\partial b_k}{\partial z_{l,t-1}}(z_{kt} - q_k) + b_k^2 \cdot \frac{\partial q_k}{\partial z_{l,t-1}}.
\end{aligned}$$

Then the linear independence of $\mathbf{v}_{kt}$ and $\mathring{\mathbf{v}}_{kt}$ (defined in Eq. 12), with $k = 1., 2, ..., n$, reduces to the linear independence condition in this proposition. Theorem S1 then implies that $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$. □

Let us then consider another special case, with linear, non-Gaussian temporal model for $\mathbf{z}_t$: the latent processes follow Eq. 18, with $\mathbf{q}$ being a linear transformation and E_{kt} following a particular class of non-Gaussian distributions. The following corollary shows that $\mathbf{z}_t$ is identifiable as long as each z_{kt} receives causal influences from some components of $\mathbf{z}_{t-1}$.

Corollary S2 (Identifiablity under a Specific Linear, Non-Gaussian Model for Latent Processes)**.** *Suppose $\mathbf{x}_t$ was generated according to Eq. 10 and Eq. 18, in which $\mathbf{q}$ is a linear transformation and for each z_{kt}, there exists at least one k' such that $c_{kk'} \triangleq \frac{\partial z_{kt}}{\partial z_{k',t-1}} \neq 0$. Assume the noise term E_{kt} follows a zero-mean generalized normal distribution:*

$$p(E_{kt}) \propto e^{-\lambda |e_{kt}|^\beta}, \text{ with positive } \lambda \text{ and } \beta > 2 \text{ and } \beta \neq 3. \tag{20}$$

Suppose Eq. 11 holds true. Then $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.

Proof. In this case, we have

$$\begin{aligned}
\frac{\partial^3 \eta_{kt}}{\partial z_{kt}^2 \partial z_{k',t-1}} &= -\lambda \cdot \text{sgn}(e_{kt}) \cdot \alpha(\beta - 1)(\beta - 2)|e_{kt}|^{\beta-3} c_{kk'}, \\
\frac{\partial^2 \eta_{kt}}{\partial z_{kt} \partial z_{k',t-1}} &= -\lambda\beta(\beta - 1)|e_{kt}|^{\beta-2} c_{kk'}.
\end{aligned} \tag{21}$$

We know that $|e_{lt}|^{\beta-2}$ and $|e_{lt}|^{\beta-3}$ are linearly independent (because their ratio, $|e_{lt}|$, is not constant). Furthermore, $|e_{lt}|^{\beta-2}$ and $|e_{lt}|^{\beta-3}$, with $l = 1, 2, ..., n$, are $2n$ linearly independent functions (because of the different arguments involved).

Suppose there exist α_{l1} and α_{l2}, with $l = 1, 2, ..., n$, such that

$$\sum_{l=1}^{n} \left(\alpha_{l1}\mathbf{v}_{lt} + \alpha_{l2}\mathring{\mathbf{v}}_{lt}\right) = 0. \tag{22}$$

It is assumed that for each $k = 1, 2, ..., n$, there exists at least one k' such that $c_{kk'} \neq 0$. Eq. 22 then implies that for any k we have

$$\alpha_{k1} c_{kk'} |e_{kt}|^{\beta-2} + \alpha_{k2} c_{kk'} |e_{kt}|^{\beta-3} + \sum_{l \neq k} \left(\alpha_{l1} c_{lk'} |e_{lt}|^{\beta-2} + \alpha_{l2} c_{lk'} |e_{lt}|^{\beta-3}\right) = 0. \tag{23}$$

Since $|e_{lt}|^{\beta-2}$ and $|e_{lt}|^{\beta-3}$, with $l = 1, 2, ..., n$, are linearly independent and $c_{kk'} \neq 0$, to make the above equation hold, one has to set $\alpha_{k1} = \alpha_{k2} = 0$. As this applies to any k, we know that for Eq. 22 to be satisfied, α_{l1} and α_{l2} must be 0, for all $l = 1, 2, ..., n$. That is, $\mathbf{v}_{1t}, \mathring{\mathbf{v}}_{1t}, \mathbf{v}_{2t}, \mathring{\mathbf{v}}_{2t}, ..., \mathbf{v}_{nt}, \mathring{\mathbf{v}}_{nt}$ are linearly independent. The linear independence condition in Theorem S1 is satisfied. Therefore $\mathbf{z}_t$ must be an invertible, component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$. □

S1.2 Proof for Theorem 2 and 3

Let $\mathbf{v}_{kt}(u_r)$ be $\mathbf{v}_{kt}$, which is defined in Eq. 12, in the u_r context. Similarly, Let $\mathring{\mathbf{v}}_{kt}(u_r)$ be $\mathring{\mathbf{v}}_{kt}$ in the u_r context. Let

$$\mathbf{s}_{kt} \triangleq \Big(\mathbf{v}_{kt}(u_1)^\intercal, ..., \mathbf{v}_{kt}(u_m)^\intercal, \frac{\partial^2 \eta_{kt}(u_2)}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(u_1)}{\partial z_{kt}^2}, ..., \frac{\partial^2 \eta_{kt}(u_m)}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(u_{m-1})}{\partial z_{kt}^2}\Big)^\intercal,$$

$$\mathring{\mathbf{s}}_{kt} \triangleq \Big(\mathring{\mathbf{v}}_{kt}(u_1)^\intercal, ..., \mathring{\mathbf{v}}_{kt}(u_m)^\intercal, \frac{\partial \eta_{kt}(u_2)}{\partial z_{kt}} - \frac{\partial \eta_{kt}(u_1)}{\partial z_{kt}}, ..., \frac{\partial \eta_{kt}(u_m)}{\partial z_{kt}} - \frac{\partial \eta_{kt}(u_{m-1})}{\partial z_{kt}}\Big)^\intercal.$$

As provided below, in our case, the identifiablity of $\mathbf{z}_t$ is guaranteed by the linear independence of the whole function vectors $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$, with $k = 1, 2, ..., n$. However, the identifiability result in Yao et al. (2021) relies on the linear independence of only the last $m - 1$ components of $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$ with $k = 1, 2, ..., n$; this linear independence is generally a much stronger condition.

Theorem S2 (Identifiability under Changing Causal Dynamics)**.** *Suppose the observed processes $\mathbf{x}_t$ was generated by Eq. 10 and that the conditional distribution $p(z_{kt} \mid \mathbf{z}_{t-1})$ may change across m values of the context variable $\mathbf{u}$, denoted by u_1, u_2, ..., u_m. Suppose the components of $\mathbf{z}_t$ are mutually independent conditional on $\mathbf{z}_{-1}$ in each context. Assume that the components of $\hat{\mathbf{z}}_t$ produced by Eq. 11 are also mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$. If the $2n$ function vectors $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$, with $k = 1, 2, ..., n$, are linearly independent, then $\hat{\mathbf{z}}_t$ is a permuted invertible component-wise transformation of $\mathbf{z}_t$.*

Proof. As in the proof of Theorem S1, because the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$, we know that for $i \neq j$,

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; \mathbf{u})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{n} \Big(\frac{\partial^2 \eta_{kt}(\mathbf{u})}{\partial z_{kt}^2} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}(\mathbf{u})}{\partial z_{kt}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) - \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} \equiv 0. \quad (24)$$

Compared to Eq. 16, here we allow $p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1})$ to depend on $\mathbf{u}$. Since the above equation is always 0, taking its partial derivative w.r.t. $z_{l,t-1}$ gives

$$\frac{\partial^3 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; \mathbf{u})}{\partial \hat{z}_{it} \partial \hat{z}_{jt} \partial z_{l,t-1}} = \sum_{k=1}^{n} \Big(\frac{\partial^3 \eta_{kt}(\mathbf{u})}{\partial z_{kt}^2 \partial z_{l,t-1}} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial^2 \eta_{kt}(\mathbf{u})}{\partial z_{kt} \partial z_{l,t-1}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) \equiv 0. \quad (25)$$

Similarly, Using different values for $\mathbf{u}$ in Eq. 24 take the difference of this equation across them gives

$$\begin{aligned} &\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; u_{r+1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} - \frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; u_{r+1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} \\ &= \sum_{k=1}^{n} \Big[\Big(\frac{\partial^2 \eta_{kt}(u_{r+1})}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(u_r)}{\partial z_{kt}^2} \Big) \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \Big(\frac{\partial \eta_{kt}(u_{r+1})}{\partial z_{kt}} - \frac{\partial \eta_{kt}(u_r)}{\partial z_{kt}} \Big) \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big] \equiv 0. \end{aligned} \quad (26)$$

Therefore, if $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$, for $k = 1, 2, ..., n$, are linearly independent, $\mathbf{H}_{kit} \mathbf{H}_{kjt}$ has to be zero for all k and $i \neq j$. Then as shown in the proof of Theorem S1, $\hat{\mathbf{z}}_t$ must be a permuted component-wise invertible transformation of $\mathbf{z}_t$. □

Theorem S3 (Identifiability under Observation Changes)**.** *Suppose $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$ and that the conditional distribution $p(z_{k,t} \mid \mathbf{u})$ may change across m values of the context variable $\mathbf{u}$, denoted by u_1, u_2, ..., u_m. Suppose the components of $\mathbf{z}_t$ are mutually independent conditional on $\mathbf{u}$ in each context. Assume that the components of $\hat{\mathbf{z}}_t$ produced by Eq. 3 are also mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$. If the $2n$ function vectors $\mathbf{s}_{k,t}$ and $\mathring{\mathbf{s}}_{k,t}$, with $k = 1, 2, ..., n$, are linearly independent, then $\hat{\mathbf{z}}_t$ is a permuted invertible component-wise transformation of $\mathbf{z}_t$.*

Proof. As in the proof of Theorem S2, because $\mathbf{z}_t$ is not dependent on the history $\mathbf{z}_{t-1}$ so are the components of $\hat{\mathbf{z}}_t$, the conditioning on $\hat{\mathbf{z}}_t$ in Eq. 24 and the following equations can be removed because of the independence. This directly leads to the same conclusion as in Theorem S2.

□

Corollary S3 (Identifiability under Modular Distribution Shifts)**.** *Assume the data generating process in Eq. 1. If the three partitioned latent components* $\mathbf{z}_t = (\mathbf{z}_t^{fix}, \mathbf{z}_t^{chg}, \mathbf{z}_t^{obs})$ *respectively satisfy the conditions in* ***Theorem*** *1,* ***Theorem*** *2, and* ***Theorem*** *3, then* $\mathbf{z}_t$ *must be an invertible, component-wise transformation of a permuted version of* $\hat{\mathbf{z}}_t$*.*

Proof. Because the three partitioned subspaces $(\mathbf{z}_t^{\text{fix}}, \mathbf{z}_t^{\text{chg}}, \mathbf{z}_t^{\text{obs}})$ are conditional independent given the history and domain index, it is straightforward to factorize the joint conditional log density into three components. By using the proof in Theorem 1, 2, and 3, we can directly derive the same quantity as in Eq. 15 or Eq. 24. Therefore, if $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$, for $k = 1, 2, ..., n$, are linearly independent, $\mathbf{H}_{kit}\mathbf{H}_{kjt}$ has to be zero for all k and $i \neq j$. Then as shown in the proof of Theorem S1, $\hat{\mathbf{z}}_t$ must be a permuted component-wise invertible transformation of $\mathbf{z}_t$. □

S1.3 Comparisons with Existing Nonlinear ICA Theories

We compare our established theory with (1) PCL [7], (2) SlowVAE [12], (3) i-VAE [9], (4) GCL [8] and (5) LEAP [14] in terms of their mathematical formulation and assumptions.

PCL [7] The formulation of the underlying processes in PCL is in Eq. 27:

$$\log p(z_{i,t}|z_{i,t-1}) = G(z_{i,t} - \rho z_{i,t-1}) \quad \text{or} \quad \log p(z_{i,t}|z_{i,t-1}) = -\lambda\left(z_{i,t} - r(z_{i,t-1})\right)^2 + \text{const}, \tag{27}$$

where G is a non-quadratic function corresponding to the log-pdf of innovations, $\rho < 1$ is the regression coefficient, r is some nonlinear, strictly monotonic regression, and λ is a positive precision parameter.

PCL is applicable to **stationary environments** only when the sources z_{it} are mutually independent (see Assumption 1 of Theorem 1 in PCL) and follow functional and distribution assumptions in Eq. 27. Our formulation allows latent variables to have arbitrary, nonparametric time-delayed causal relations in between without functional form or distribution assumptions.

SlowVAE [12] The formulation of the underlying sources in SlowVAE is in Eq. 28:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \prod_{i=1}^{d} \frac{\alpha\lambda}{2\Gamma(1/\alpha)} \exp -(\lambda|z_{i,t} - z_{i,t-1}|^{\alpha}) \quad with \quad \alpha < 2, \tag{28}$$

where the latent processes have independent, identity transitions with generalized Laplacian noises.

SlowVAE established identifiability under stationary, mutually independent processes (similar to PCL [7]), in which time-delayed causal influences are not allowed. Furthermore, it assumes that the transition function of each independent process is an identity function and the process noise has generalized Laplacian distribution. Our **Theorem 1** includes both SlowVAE and PCL as special cases, in the sense that (1) we remove the functional and distributional assumptions to allow the latent processes to have nonparametric causal influences in between, and (2) our **Corollay 2**, which is an illustrative example of **Theorem 1**, further completes to Eq. 28 by allowing **linear time-delayed transitions** in the latent process with **non-Gaussian noises**.

i-VAE [9] Similar to TCL [6] and GIN [10], i-VAE exploits the nonstationarity brought by class labels on the distribution of latent variables. As one can see from Eq. 29, the latent variables are conditionally independent, without causal relations in between while all of our theorems consider (time-delayed) causal relations between latent variables. In addition, iVAE assumes the modulation of class labels on latent distributions is limited within the exponential family distribution. On the contrary, our nonparametric conditions (Theorems 1,2,3) allow any kind of modulation caused by fixed, changing transition dynamics or observation changes without those strong assumptions on the distribution of latent variables or noise distribution (i.e., SlowVAE [12]).

$$p_{T,\lambda}(\mathbf{z}|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{z_i(\mathbf{u})} \exp\left[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(\mathbf{u})\right] \tag{29}$$

GCL [8] The formulation of the underlying sources in GCL is in Eq 30, which is also described by Eqs. 4,15 in the original paper [8]:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \prod_{i=1}^{d} p_i(z_{i,t}|z_{i,t-1}), \tag{30}$$

where the latent processes are free from the exponential family distribution assumptions but still constrained within mutually-independent processes. On the contrary, our work considers **causally-related** latent space in which cross causal relations between latent variables can be recovered. Additionally, we want to mention that we provide an antenna tube in the schema of the proof in **Theorems** 1-2-3, which is a more direct way of using sufficient variability conditions than [8].

LEAP [14] LEAP in Eq. 31 considers one special case of nonstationarity caused by changes in noise distributions, while our work can allow changing causal relations over context. Furthermore, because LEAP assumes all latent processes are changed across contexts, it doesn't use or benefit from the fixed time-delayed causal relations for identifiability. On the contrary, our work exploits the modular distribution changes from the fixed causal dynamics, changing dynamics, and observation changes, and hence our identifiability conditions are generally weaker than [14].

$$\underbrace{\mathbf{x}_t = g(\mathbf{z}_t)}_{\text{Nonlinear mixing}}, \quad \underbrace{z_{it} = f_i\left(\{z_{j,t-\tau}|z_{j,t-\tau} \in \mathbf{Pa}(z_{it})\}, \epsilon_{it}\right)}_{\text{Nonparametric transition}} \; with \; \underbrace{\epsilon_{it} \sim p_{\epsilon_i|\mathbf{u}}}_{\text{Nonstationary noise}} . \tag{31}$$

S1.4 Discussion of the Assumptions

We first explain and justify each critical assumption in the proposed conditions. We then discuss how restrictive or mild the conditions are in real applications.

S1.4.1 Linear Independence Condition

Our proposed linear independence condition is a combination of stationary identifiability conditions (Eq. 4) in each context u_r, plus the identifiability conditions for nonrecurrent influences (Eq. 8). The condition is essential to make each row of the Jacobian matrix $\mathbf{H}_t$ of the indeterminacy function of the learned latent space in Eq. 17 to have only one non-zero entry, thus making the learned latent variables identifiable up to permutation and component-wise invertible transformations.

In **stationary environments**, this condition essentially states that, during the generation of $\mathbf{z}_t$, if either of the two conditions is satisfied, then the linear independence condition holds in general.

(1) If the history information $\mathbf{z}_{\text{Hx}} = \{\mathbf{z}_{t-\tau}\}_{\tau=1}^{L}$ up to maximum time lag L, and the process noise ϵ_t are coupled in a nontrivial way for generating $\mathbf{z}_t$ (e.g., heterogeneous noise process in Eq. 32), such that $\mathbf{z}_{\text{Hx}}$ can modulate the variance or higher-order statistics of the conditional distribution $p(\mathbf{z}_t|\mathbf{z}_{\text{Hx}})$, then the linear independence condition generally holds;

$$z_{k,t} = q_k(\mathbf{z}_{t-1}) + \frac{1}{b_k(\mathbf{z}_{t-1})}\epsilon_{k,t}. \tag{32}$$

(2) If the latent transition is an additive noise model (then (1) is violated) but the process noise is non-Gaussian, it will be extremely hard for the linear independence condition to be violated. Roughly speaking, for a randomly chosen conditional density function $p(z_{k,t} \,|\, \mathbf{z}_{t-1})$ in which $z_{k,t}$ is not independent from $\mathbf{z}_{t-1}$ (i.e., there is temporal dependence in the latent processes) and which does not follow an additive noise model with Gaussian noise, the chance for its specific second- and third-order partial derivatives to be linearly dependent is slim.

In **nonstationary environments**, this condition was introduced in GCL [8], namely, "sufficient variability", to extend the modulated exponential families [6] to general modulated distributions. Essentially, the condition says that the nonstationary regimes $\mathbf{u}$ must have a sufficiently complex and diverse effect on the transition distributions. In other words, if the underlying distributions are composed of relatively many domains of data, the condition generally holds true. For instance, in the linear Auto-Regressive (AR) model with Gaussian innovations where only the noise variance

changes, the condition reduces to the statement in [38] that the variance of each noise term fluctuates somewhat independently of each other in different nonstationary regimes. Then the condition is easily attained if the variance vector of noise terms in any regime is not a linear combination of variance vectors of noise terms in other regimes.

We further illustrate the condition using the example of modulated conditional exponential families in [8]. Let the log-pdf $q(\mathbf{z}_t|\{\mathbf{z}_{t-\tau}\}, \mathbf{u})$ be a conditional exponential family distribution of order k given nonstationary regime $\mathbf{u}$ and history $\mathbf{z}_{\text{Hx}} = \{\mathbf{z}_{t-\tau}\}$:

$$q(z_{it}|\mathbf{z}_{\text{Hx}}, \mathbf{u}) = q_i(z_{it}) + \sum_{j=1}^{k} q_{ij}(z_{it})\lambda_{ij}(\mathbf{z}_{\text{Hx}}, \mathbf{u}) - \log Z(\mathbf{z}_{\text{Hx}}, \mathbf{u}), \tag{33}$$

where q_i is the base measure, q_{ij} is the function of the sufficient statistic, λ_{ij} is the natural parameter, and $\log Z$ is the log-partition. Loosely speaking, the sufficient variability holds if the modulation of by $\mathbf{u}$ on the conditional distribution $q(z_{it}|\mathbf{z}_{\text{Hx}}, \mathbf{u})$ is not too simple in the following sense:

1. Higher order of k ($k > 1$) is required. If $k = 1$, the sufficient variability cannot hold;
2. The modulation impacts λ_{ij} by $\mathbf{u}$ must be linearly independent across regimes $\mathbf{u}$. The sufficient statistics functions q_{ij} cannot be all linear, i.e., we require higher-order statistics.

Further details of this example can be found in Appendix B of [8]. In summary, we need the modulation by $\mathbf{u}$ to have diverse (i.e., distinct influences) and complex impacts on the underlying data generation process.

Applicability By combining the stationary and nonstationary conditions, our proposed identifiability condition is generally mild, in the sense that if there is at least one regime r out of the m contexts which satisfies the stationary identifiability conditions, OR, if the overall nonstationary influences are diverse and complex, thus satisfying the nonstationary identifiability conditions, the latent temporal causal processes are identifiable. For stationary conditions, the only situation where we find the latent processes unidentifiable is when the latent temporal transition is described by a Gaussian additive noise model, which violates both (1) and (2). However, for real-world data, it is very unlikely for the process noise to be perfectly Gaussian. Nonstationarity seems to be prominent in many kinds of temporal data. For example, nonstationary variances are seen in EEG/MEG, and natural video, and are closely related to changes in volatility in financial time series [6]. The data that most likely satisfy the nonstationary condition is a collection of multiple trials/segments of data with different temporal dynamics in between.

S1.4.2 Independent Noise Condition

The IN condition was introduced in the Structural Equation Model (SEM), which represents effect Y as a function of direct causes X and noise E:

$$Y = f(X, E) \quad with \quad \underbrace{X \perp\!\!\!\perp E}_{\text{IN condition}} . \tag{34}$$

If X and Y do not have a common cause, as seen from the causal sufficiency assumption of structural equation models in Chapter 1.4.1 of Pearl's book [39], the IN condition states that the unexplained noise variable E is statistically independent of cause X. IN is a direct result of assuming causal sufficiency in SEM. The main idea for the proof is that if IN is violated, then by the common cause principle [40], there exist hidden confounders that cause their dependence, thus violating the causal sufficiency assumption. Furthermore, the noise terms in different variables are mutually independent for a causally sufficient system with acyclic causal relations. The main idea is that when the noise terms are dependent, it is customary to encode such dependencies by augmenting the graph with hidden confounder variables [39], which means that the system is not causally sufficient.

This paper assumes that the underlying latent processes form a casually-sufficient system without latent causal confounders. Then, the process noise terms ϵ_{it} are mutually independent, and moreover, the process noise terms ϵ_{it} are independent of direct cause/parent nodes $\mathbf{Pa}(z_{it})$ because of time information (the causal graph is acyclic because of the temporal precedence constraint).

Applicability Loosely speaking, if there are no latent causal confounders in the (latent) causal processes and the sampling frequency is high enough to observe the underlying dynamics, then the IN condition assumed in this paper is satisfied in a causally-sufficient system and, moreover, there is no instantaneous causal influence (because of the high enough resolution). At the same time, we acknowledge that there exist situations where the resolution is low and there appears to be instantaneous dependence. However, several pieces of work deal with causal discovery from measured time series in such situations; see. e.g., [41, 42, 43, 44]. In case there are instantaneous causal relations among latent causal processes, one would need additional sparsity or minimality conditions to recover the latent processes and their relations, as demonstrated in [45, 46]. How to address the issue of instantaneous dependency or instantaneous causal relations in the latent processes will be one line of our future work.

S1.4.3 Causal Influences between Observed Variables

Although causal discovery between observed variables is not the main focus of our work, our model can discover causal relations between observed variables as a special case, thanks to the nonlinear mixing function assumed in this paper. In our formulation in Eq. 1, we assume the observations x_t are nonlinear, invertible mixtures of latent processes z_t. However, if in the data generating process, one observed variable x_i has direct causal edges with the other observed variable x_j, the mixing function that generates x_i and x_j will just be reduced to identity mappings of z_i and z_j, which is a special case of the nonlinear invertible mixing function in Eq. 1.

S1.5 Extension to Multiple Time Lags

For the sake of simplicity, we consider one lag for the latent processes in Section 3 (and only in Section 3). Our identifiability proof can actually be applied for arbitrary lags directly. For instance, in the stationary case in Eq. 4, one can simply re-define $\eta_{kt} \triangleq \log p(z_{k,t}|\mathbf{z}_{Hx})$, where $\mathbf{z}_{Hx}$ denotes the lagged latent variables up to maximum time lag L. We plug it into Eq. 4, and take derivatives with regard to $z_{1,t-\tau}, \ldots, z_{n,t-\tau}$, which can be any latent temporal variables at lag τ, instead of $z_{1,t-1}, \ldots, z_{n,t-1}$. If there exists one τ (out of the L lags) that satisfies the condition, then the stationary latent processes are identifiable. Similarly, for Eq. 8, one can simply re-define $\eta_kt(u_r) \triangleq \log p(z_{k,t}|\mathbf{z}_{Hx}, u_r)$ and plug it into Eq. 8. No extra changes are needed.

S2 Experiment Settings

S2.1 Datasets

S2.1.1 Synthetic Dataset Generation

We consider three representative simulation settings to validate the identifiability results under fixed causal dynamics, changing causal dynamics, and modular distribution shift which contains fixed dynamics, changing dynamics, and global changes together in the latent processes. For synthetic datasets with fixed and changing causal dynamics, we set latent size $n = 8$. For the modular shift dataset, we add one dimension for global observation changes. The lag number of the process is set to $L = 2$. The mixing function g is a random three-layer MLP with LeakyReLU units.

Fixed Causal Dynamics For the fixed causal dynamics. We generate 100,000 data points according to Eq. (6), where the latent size is $n = 8$, lag number of the process is $L = 2$. We apply a 2-layer MLP with LeakyReLU as the state transition function. The process noise are sampled from i.i.d. Gaussian distribution ($\sigma = 0.1$). The process noise terms are coupled with the history information through multiplication with the average value of all the time-lagged latent variables.

Changing Causal Dynamics We use a Gaussian additive noise model with changes in the influencing strength as the latent processes. To add changes, we vary the values of the first layer of the MLP across the 20 segments and generate 7,500 samples for each segment. The entries of the kernel matrix of the first layer are uniformly distributed between $[-1, 1]$ in each domain.

Modular Distribution Shifts The latent space of this dataset is partitioned into 6 fixed dynamics components under the heterogeneous noise model, 2 changing components with changing causal

dynamics and 1 component modulated by domain index only. The fixed and changing dynamics components follow the same generating procedures above. The global change component is sampled from i.i.d Gaussian distribution whose mean and variance are modulated by domain index. In particular, distribution mean terms are uniformly sampled between $[-1, 1]$ and variance terms are uniformly sampled between $[0.01, 1]$.

S2.1.2 Real-world Dataset

Modified Cartpole The Cartpole problem [30] "consists of a cart and a vertical pendulum attached to the cart using a passive pivot joint. The cart can move left or right. The task is to prevent the vertical pendulum from falling by putting a force on the cart to move it left or right. The action space consists of two actions: moving left or right."

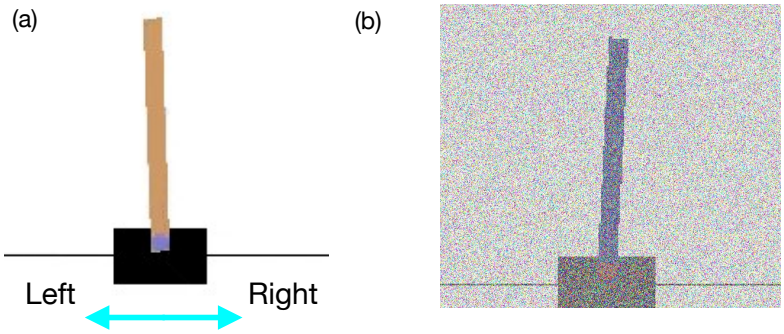

Figure S1: "Visual examples of Cartpole game and change factors. (a) Cartpole game; (b) Modified Cartpole game with Gaussian noise on the image. The light blue arrows are added to show the direction in which the agent can move." Figure source: [30].

The original dataset [30] introduces "two change factors respectively for the state transition dynamics θ_k^{dyn}: varying gravity and varying mass of the cart, and a change factor in the observation function θ_k^{obs} that is the image noise level. Fig. S1 gives a visual example of Cartpole game, and the image with Gaussian noise. The images of the varying gravity and mass look exactly like the original image. Specifically, in the gravity case, we consider source domains with gravity $g = \{5, 10, 20, 30, 40\}$. We take into account both interpolation (where the gravity in the target domain is in the support of that in source domains) with $g = \{15\}$, and extrapolation (where it is out of the support w.r.t. the source domains) with $g = \{55\}$. Similarly, we consider source domains where the mass of the cart is $m = \{0.5, 1.5, 2.5, 3.5, 4.5\}$, while in target domains it is $m = \{1.0, 5.5\}$. In terms of changes on the observation function, we add Gaussian noise on the images with variance $\sigma = \{0.25, 0.75, 1.25, 1.75, 2.25\}$ in source domains, and $\sigma = \{0.5, 2.75\}$ in target domains. The detailed settings in both source and target domains are in Table S1."

	Gravity	Mass	Noise
Source domains	$\{5, 10, 20, 30, 40\}$	$\{0.5, 1.5, 2.5, 3.5, 4.5\}$	$\{0.25, 0.75, 1.25, 1.75, 2.25\}$
Interpolation set	$\{15\}$	$\{1.0\}$	$\{0.5\}$
Extrapolation set	$\{55\}$	$\{5.5\}$	$\{2.75\}$

Table S1: "The settings of source and target domains for modified Cartpole experiments" [30].

CMU-Mocap CMU MoCap (`http://mocap.cs.cmu.edu/`) is an open-source human motion capture dataset with various motion capture recordings (e.g., walk, jump, basketball, etc.) performed by over 140 subjects. In this work, we fit our model on 11 trials of "walk" recordings (Subject #8). Skeleton-based measurements have 62 observed variables corresponding to the locations of joints (e.g., head, foot, shoulder, wrist, throat, etc.) of the human body at each time step.

S2.2 Mean Correlation Coefficient

MCC is a standard metric for evaluating the recovery of latent factors in ICA literature. MCC first calculates the absolute values of the correlation coefficient between every ground-truth factor against every estimated latent variable. Pearson correlation coefficients or Spearman's rank correlation coefficients can be used depending on whether componentwise invertible nonlinearities exist in the recovered factors. The possible permutation is adjusted by solving a linear sum assignment problem in polynomial time on the computed correlation matrix.

S3 Implementation Details

S3.1 Modular Prior Likelihood Derivation

Let us start with an illustrative example of stationary latent causal processes consisting of two time-delayed latent variables, i.e., $\mathbf{z}_t = [z_{1,t}, z_{2,t}]$ with maximum time lag $L = 1$, i.e., $z_{i,t} = f_i(\mathbf{z}_{t-1}, \epsilon_{i,t})$ with mutually independent noises. Let us write this latent process as a transformation map $\mathbf{f}$ (note that we overload the notation f for transition functions and for the transformation map):

$$\begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \\ z_{1,t} \\ z_{2,t} \end{bmatrix} = \mathbf{f}\left(\begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \\ \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}\right). \tag{35}$$

By applying the change of variables formula to the map $\mathbf{f}$, we can evaluate the joint distribution of the latent variables $p(z_{1,t-1}, z_{2,t-1}, z_{1,t}, z_{2,t})$ as:

$$p(z_{1,t-1}, z_{2,t-1}, z_{1,t}, z_{2,t}) = p(z_{1,t-1}, z_{2,t-1}, \epsilon_{1,t}, \epsilon_{2,t}) / \left|\det \mathbf{J}_{\mathbf{f}}\right|, \tag{36}$$

where $\mathbf{J}_{\mathbf{f}}$ is the Jacobian matrix of the map $\mathbf{f}$, which is naturally a low-triangular matrix:

$$\mathbf{J}_{\mathbf{f}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{\partial z_{1,t}}{\partial z_{1,t-1}} & \frac{\partial z_{1,t}}{\partial z_{2,t-1}} & \frac{\partial z_{1,t}}{\partial \epsilon_{1,t}} & 0 \\ \frac{\partial z_{2,t}}{\partial z_{1,t-1}} & \frac{\partial z_{2,t}}{\partial z_{2,t-1}} & 0 & \frac{\partial z_{2,t}}{\partial \epsilon_{2,t}} \end{bmatrix}.$$

Given that this Jacobian is triangular, we can efficiently compute its determinant as $\prod_i \frac{\partial z_{i,t}}{\partial \epsilon_{i,t}}$. Furthermore, because the noise terms are mutually independent, and hence $\epsilon_{i,t} \perp \epsilon_{j,t}$ for $j \neq i$ and $\epsilon_t \perp \mathbf{z}_{t-1}$, we can write the RHS of Eq. 36 as:

$$\begin{aligned} p(z_{1,t-1}, z_{2,t-1}, z_{1,t}, z_{2,t}) &= p(z_{1,t-1}, z_{2,t-1}) \times p(\epsilon_{1,t}, \epsilon_{2,t}) / \left|\det \mathbf{J}_{\mathbf{f}}\right| \quad (\text{because } \epsilon_t \perp \mathbf{z}_{t-1}) \\ &= p(z_{1,t-1}, z_{2,t-1}) \times \prod_i p(\epsilon_{i,t}) / \left|\det \mathbf{J}_{\mathbf{f}}\right| \quad (\text{because } \epsilon_{1,t} \perp \epsilon_{2,t}) \end{aligned} \tag{37}$$

Finally, by canceling out the marginals of the lagged latent variables $p(z_{1,t-1}, z_{2,t-1})$ on both sides, we can evaluate the transition prior likelihood as:

$$p(z_{1,t}, z_{2,t} | z_{1,t-1}, z_{2,t-1}) = \prod_i p(\epsilon_{i,t}) / \left|\det \mathbf{J}_{\mathbf{f}}\right| = \prod_i p(\epsilon_{i,t}) \times \left|\det \mathbf{J}_{\mathbf{f}}^{-1}\right|. \tag{38}$$

Now we generalize this example and derive the modular prior likelihood below.

Fixed Causal Dynamics Let $\{f_s^{-1}\}_{s=1,2,3\ldots}$ be a set of learned inverse fixed dynamics transition functions that take the estimated latent causal variables in the fixed dynamics subspace and lagged latent variables, and output the noise terms, i.e., $\hat{\epsilon}_{s,t} = f_s^{-1}\left(\hat{z}_{s,t}^{\text{fix}}, \{\hat{\mathbf{z}}_{t-\tau}\}\right)$.

Design transformation $\mathbf{A} \to \mathbf{B}$ with low-triangular Jacobian as follows:

$$\underbrace{\left[\hat{\mathbf{z}}_{t-L},\ldots,\hat{\mathbf{z}}_{t-1},\hat{\mathbf{z}}_t^{\text{fix}}\right]^\top}_{\mathbf{A}} \text{ mapped to } \underbrace{\left[\hat{\mathbf{z}}_{t-L},\ldots,\hat{\mathbf{z}}_{t-1},\hat{\epsilon}_{s,t}\right]^\top}_{\mathbf{B}},\ with\ \mathbf{J}_{\mathbf{A}\to\mathbf{B}} = \begin{pmatrix} \mathbb{I}_{nL} & 0 \\ * & \text{diag}\left(\frac{\partial f_{s,i}^{-1}}{\partial \hat{z}_{it}^{\text{fix}}}\right) \end{pmatrix}. \tag{39}$$

Similar to Eq. 38, we can obtain the joint distribution of the estimated fixed dynamics subspace as:

$$\log p(\mathbf{A}) = \underbrace{\log p\left(\hat{\mathbf{z}}_{t-L},\ldots,\hat{\mathbf{z}}_{t-1}\right) + \sum_{i=1}^{n} \log p(\hat{\epsilon}_{s,t})}_{\text{Because of mutually independent noise assumption}} + \log\left(\left|\det\left(\mathbf{J}_{\mathbf{A}\to\mathbf{B}}\right)\right|\right). \tag{40}$$

$$\log p\left(\hat{\mathbf{z}}_t^{\text{fix}} | \{\hat{\mathbf{z}}_{t-\tau}\}_{\tau=1}^{L}\right) = \sum_{i=1}^{n} \log p(\hat{\epsilon}_{s,t}) + \sum_{i=1}^{n} \log \left|\frac{\partial f_s^{-1}}{\partial \hat{z}_{s,t}}\right| \tag{41}$$

Changing Causal Dynamics The differences from fixed dynamics are that the learned inverse changing dynamics transition functions take additional learned change factors of the context as input arguments to out the noise terms, i.e, $\hat{\epsilon}_{c,t} = f_c^{-1}\left(\hat{z}_{c,t}^{\text{chg}}, \{\hat{\mathbf{z}}_{t-\tau}\}, \mathbf{u}_k\right) = f_c^{-1}\left(\hat{z}_{c,t}^{\text{chg}}, \{\hat{\mathbf{z}}_{t-\tau}\}, \boldsymbol{\theta}_k^{\text{dyn}}\right)$.

$$\log p\left(\hat{\mathbf{z}}_t^{\text{chg}} | \{\hat{\mathbf{z}}_{t-\tau}\}_{\tau=1}^{L}, \mathbf{u}_k\right) = \sum_{i=1}^{n} \log p(\hat{\epsilon}_{c,t}|\mathbf{u}_k) + \sum_{i=1}^{n} \log \left|\frac{\partial f_c^{-1}}{\partial \hat{z}_{c,t}}\right| \tag{42}$$

Observation Changes The global observation changes are captured by the learned inverse f_o^{-1},which takes the estimated latent subspace and the learned change factors for global observation $\boldsymbol{\theta}_k^{\text{obs}}$ of context k, and output random noise, i.e, $\hat{\epsilon}_{o,t} = f_o^{-1}\left(\hat{z}_{o,t}^{\text{obs}}, \mathbf{u}_k\right) = f_c^{-1}\left(\hat{z}_{c,t}^{\text{chg}}, \boldsymbol{\theta}_k^{\text{obs}}\right)$.

$$\log p\left(\hat{\mathbf{z}}_t^{\text{obs}} | \mathbf{u}_k\right) = \sum_{i=1}^{n} \log p(\hat{\epsilon}_{o,t}|\mathbf{u}_k) + \sum_{i=1}^{n} \log \left|\frac{\partial f_o^{-1}}{\partial \hat{z}_{o,t}}\right|. \tag{43}$$

S3.2 Comparisons with AdaRL [30]

In terms of the implementation, in our work, we enforce the independent noise or conditional independence condition explicitly in Eq. 9 (derived in Appendix S3.1) for the identifiability of the latent processes. But disentanglement is not the main goal of AdaRL so it used Mixture Density Network (MDN) to approximate the transition prior. Our framework is simpler than AdaRL in the inference module (we use only $\mathbf{x}_t$ to infer $\mathbf{z}_t$) and the loss function (we don't have the prediction branch or the sparsity loss), thanks to the nonlinear ICA formulation in Eq. 1. Our identifiability conditions do not rely on the sparsity constraints in the underlying data generating process.

S3.3 Network Architecture

We summarize our network architecture below and describe it in detail in Table S2 and Table S3.

S3.4 Hyperparameter and Training

Hyperparameter Selection The hyperparameters include β, which is the weight of KLD terms, as well as the latent size n and maximum time lag L. We use the ELBO loss to select the best pair of β because low ELBO loss always leads to high MCC. We always set a larger latent size than the true latent size. This is critical in real-world datasets because restricting the latent size will hurt the reconstruction performances and over-parameterization makes the framework robust to assumption violations. For the maximum time lag L, we set it by the rule of thumb. For instance, we use $L = 2$ for temporal datasets with a latent physics process (e.g, cartpole, cmu-mocap).

Training Details The models were implemented in `PyTorch` 1.8.1. The VAE network is trained using AdamW optimizer for a maximum of 50 epochs and early stops if the validation ELBO loss does not decrease for five epochs. A learning rate of 0.002 and a mini-batch size of 64 are used. We have used three random seeds in each experiment and reported the mean performance with standard deviation averaged across random seeds.

Table S2: Architecture details. BS: batch size, T: length of time series, i_dim: input dimension, z_dim: latent dimension, LeakyReLU: Leaky Rectified Linear Unit.

Configuration	Description	Output
1. MLP-Encoder	Encoder for Synthetic Data	
Input: $\mathbf{x}_{1:T}$	Observed time series	BS $\times$ T $\times$ i_dim
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	Temporal embeddings	BS $\times$ T $\times$ z_dim
2. MLP-Decoder	Decoder for Synthetic Data	
Input: $\hat{\mathbf{z}}_{1:T}$	Sampled latent variables	BS $\times$ T $\times$ z_dim
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	128 neurons, LeakyReLU	BS $\times$ T $\times$ 128
Dense	i_dim neurons, reconstructed $\hat{\mathbf{x}}_{1:T}$	BS $\times$ T $\times$ i_dim
5. Factorized Inference Network	Bidirectional Inference Network	
Input	Sequential embeddings	BS $\times$ T $\times$ z_dim
Bottleneck	Compute mean and variance of posterior	$\mu_{1:T}, \sigma_{1:T}$
Reparameterization	Sequential sampling	$\hat{\mathbf{z}}_{1:T}$
6. Modular Prior	Nonlinear Transition Prior Network	
Input	Sampled latent variable sequence $\hat{\mathbf{z}}_{1:T}$	BS $\times$ T $\times$ z_dim
InverseTransition	Compute estimated residuals $\hat{\epsilon}_{it}$	BS $\times$ T $\times$ z_dim
JacobianCompute	Compute $\log\left(\lvert\det\left(\mathbf{J}\right)\rvert\right)$	BS

Training Stability We have used several tricks to improve training stability: (1) we use a slightly larger latent size than the true latent size for real-world datasets in order to make sure the meaning latent variables are among the recovered latent; (2) we use AdamW optimizer as a regularizer to prevent training from being interrupted by overflow or underflow of variance terms of VAE; (3) we use a large learning rate (0.002) during training.

Computing Hardware We used a machine with the following CPU specifications: Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz; 8 CPUs, four physical cores per CPU, a total of 32 logical CPU units. The machine has two GeForce GTX 1080 Ti GPUs with 11GB GPU memory.

Reproducibility We've included the code for the framework and all experiments in the supplement. We plan to release our code under the MIT License after the paper review period.

S4 Additional Experiment Results

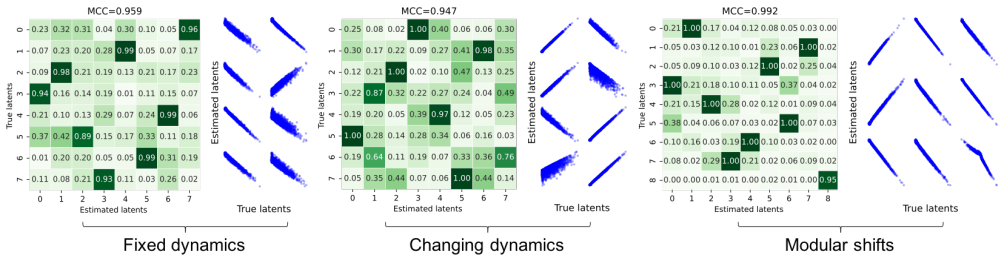

Figure S2: Results for three synthetic datasets: in each block, the left shows the MCC for causally-related and the left are scatterplots between estimated and true factors.

Table S3: Architecture details on CNN encoder and decoder. BS: batch size, T: length of time series, h_dim: hidden dimension, z_dim: latent dimension, F: number of filters, (Leaky)ReLU: (Leaky) Rectified Linear Unit.

Configuration	**Description**	**Output**
3.1.1 CNN-Encoder	Feature Extractor	
Input: $\mathbf{x}_{1:T}$	RGB video frames	BS $\times$ T $\times$ 3 $\times$ 64 $\times$ 64
Conv2D	F: 32, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 32 $\times$ 64 $\times$ 64
Conv2D	F: 32, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 32 $\times$ 32 $\times$ 32
Conv2D	F: 32, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 32 $\times$ 16 $\times$ 16
Conv2D	F: 64, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 64 $\times$ 8 $\times$ 8
Conv2D	F: 64, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 64 $\times$ 4 $\times$ 4
Conv2D	F: 128, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 128 $\times$ 1 $\times$ 1
Dense	F: 2 * z_dim = dimension of hidden embedding	BS $\times$ T $\times$ 2 * z_dim
4.1 CNN-Decoder	Video Reconstruction	
Input: $\mathbf{z}_{1:T}$	Sampled latent variable sequence	BS $\times$ T $\times$ z_dim
Dense	F: 128 , LeakyReLU	BS $\times$ T $\times$ 128 $\times$ 1 $\times$ 1
ConvTranspose2D	F: 64, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 64 $\times$ 4 $\times$ 4
ConvTranspose2D	F: 64, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 64 $\times$ 8 $\times$ 8
ConvTranspose2D	F: 32, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 32 $\times$ 16 $\times$ 16
ConvTranspose2D	F: 32, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 32 $\times$ 32 $\times$ 32
ConvTranspose2D	F: 32, BatchNorm2D, LeakyReLU	BS $\times$ T $\times$ 32 $\times$ 64 $\times$ 64
ConvTranspose2D	F: 3, estimated scene $\hat{\mathbf{x}}_{1:T}$	BS $\times$ T $\times$ 3 $\times$ 64 $\times$ 64