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 *LiLY* 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 and time-varying latent causal relations, and by 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 properly 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 and Related Work

Causal discovery from time-series data [1, 2] is a crucial task in numerous fields, including machine learning [3], econometrics [4], and neuroscience [5]. Most existing work focuses on estimating the temporal causal relations between observed variables. For this task, constraint-based methods [6] apply the conditional independence tests to recover the causal structures, while score-based methods [7, 8] define score functions to guide a search process. Furthermore, the hybrid methods [9, 10] propose to fuse both conditional independence tests and score-based methods. The Granger causality [11] and its nonlinear variations [12, 13] are commonly used. Recently, causal discovery with latent variables has attracted much attention. Estimating latent causal structure from observations is challenging because the latent variables, even with independent factors of variation [14], are not identifiable in the most general case [15]. Several pieces of work aim to uncover causally related latent variables under linear relations or minimality assumptions. As early as [16], vanishing Tetrad conditions were exploited to identify latent variables in linear-Gaussian models [17]. Generalized Independent Noise (GIN) condition estimates linear, non-Gaussian latent variable causal graph [18]. In the field of nonlinear Independent Component Analysis (ICA), identifiability results have been established [19, 20, 21, 22, 23] by using certain side information, such as class labels, in addition to

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

independence. For time-series data, history information is widely used as side information for nonlinear ICA. However, most existing work that establishes identifiability results considers either stationary independent sources such as PCL [20], SlowVAE [24] or under linear transition assumptions such as SlowVAE [24] and SNICA [25], or with certain structure such as Markov properties in HM-NLICA [26]. Furthermore, most existing work requires special assumptions on the distribution of latent variables or process noise distribution, such as exponential family distribution [22] or Laplacian noise [24]. Alternatively, distribution changes or nonstationarity have proved to greatly improve the identifiability results for learning the latent causal structure [27, 28, 29]. For example, LEAP [27] established the identifiability of the latent temporal processes in limited nonstationary cases, under the condition that the distribution of the noise terms of the latent processes vary across segments.

Table 1: Attributes of nonlinear ICA theories for time-series. A check denotes that a method has an attribute, 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	✗	✗	✗	✓	✓
HM-NLICA	✗	✗	✗	✗	✗
SlowVAE	✗	✗	✗	✗	✓
SNICA	✓	✓	✗	✗	✗
LEAP	✗	✓	✗	✓	✗
LiLY †	✓	✓	✓	✓	✓

Questions arise that whether the latent causal processes are identifiable if they have arbitrary, nonparametric causal influences in between. Specifically, this paper provides answers to: **1)** *in stationary environments, can the special constraints on the latent distribution, such as exponential familiy [22], Laplacian noise [24, 27] and independent processes be relaxed*, and **2)** *in nonstationary environments, can the latent causal processes be identified in a general nonstationary case which includes changing causal relations such as changes in the influencing strength or switching some edges off [30], nonstationary noise distributions across domains [27], and domain shifts in observation [31]*.

In this work, we propose a principled framework, called *LiLY*, to recover nonparametric time-delayed latent causal variables and identify their relations from measured temporal data under stationary environments, while also further benefiting from unknown distribution shifts. For instance, Fig. 1 shows an example of multiple video segments of a physical system (i.e., cart-pole) under different mass, gravity and environment rendering settings. With *LiLY*, the differences across segments are characterized by change factors $\boldsymbol{\theta}_k^{\text{dyn}}$ of segment k that encode changes in transition dynamics, and changes in observation or styles modeled by $\boldsymbol{\theta}_k^{\text{obs}}$. We then present a generalized time-series data generative model that takes these change factors as arguments for modeling the distribution changes. Specifically, we factorizes 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 mixtures. We

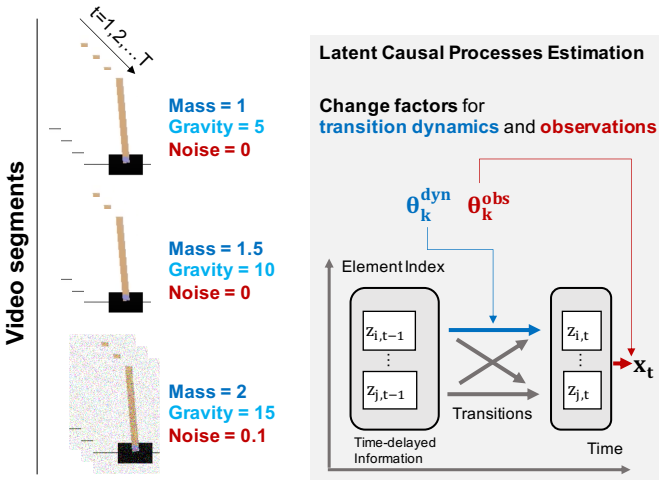

Figure 1: **LiLY**: Learning Latent causal dYnamics under modular distribution shift. We exploit fixed causal dynamics and distribution changes from changing causal influences and global observation changes to identify the underlying causal dynamics. Note $z_{i,t}$ denotes the latent temporal process. θ_k^{dyn} is the low-dimensional change factor for transition dynamics, i.e., mass and gravity in this example. θ_k^{obs} is the change factor for observation, i.e., image noise in this example.

demonstrate through a number of synthetic and 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 under different unknown distribution shifts. Through experiments, we show that our approach considerably outperforms existing baselines that do not properly leverage this factorized representation of changes. Our results show that disentanglement in time-series settings seems promising both in stationary environments and general nonstationary environments, in which the latent processes have arbitrary, nonparametric causal influences in between.

Our framework can be differentiated from the prior work in five aspects as summarized in Table 1:

1. In stationary setting, our framework does not rely on distribution assumptions such as exponential family and Laplacian noise, or functional form assumptions such as independent processes and linear assumptions. Hence the identifiability results are generally applicable for causal process learning in **nonparametric** time-series settings even in **stationary** environments;
2. In nonstationary environements, we consider a **very general case** which includes changing causal dynamics such as changes in the influencing strength or switching some edges off, nonstationary noise distributions across domains , and domain shifts in observation, etc., as special cases;
3. We allow both stationary and nonstationary latent processes to co-exist by using **partitioned latent subspaces** in the data generating processes. The identifiability results under this setting are appealing to real-world data in which whether and how the distribution changes are unknown.

2 Problem Formulation

We present a modular time-series data generative model that factorizes distribution changes into fixed dynamics, time-varying transitions and observation by partitioned latent subspaces. We define the properties of the generative model which suffice the identifiability of time-delayed causal dynamics.

2.1 A Modular Representation of Changes

Following [32, 33], we can represent the changes in a compact way with a low-dimensional vector $\boldsymbol{\theta}_k$. Specifically, we partition the latent space to reflect distribution changes under (1) fixed dynamics, (2) changing dynamics and (3) observation changes. Let $\mathbf{u}$ denotes domain or segment index. Suppose there exist K domains or segments, i.e., u_k with $k = 1, 2, ..., K$. In each domain, we observe time series $\{\mathbf{x}_t\}_{t=1}^T$ generated by an arbitrary invertible nonlinear mixture of the underlying temporal processes $\mathbf{z}_t$. We partition the latent space into 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 (e.g., styles). Because of Independent Causal Mechanism (ICM [34, 35]), a compact representation of changes should disentangle changes in transition dynamics and appearance. We hence assume the data generating process in each segment $\mathbf{u}_k$ can be described by the transition function $[f_s, f_c, f_o]$ that respectively reflects fixed and changing transitions and observation changes for each dimension of $\mathbf{z}_t$, which takes its parent nodes change factors, $(\boldsymbol{\theta}_k^{\text{dyn}}, \boldsymbol{\theta}_k^{\text{obs}})$ and noise terms as inputs, and the mixing function $\mathbf{g}$ as in Eq. 1, where $\mathbf{Pa}(z_{i,t})$ denote the set of (time-delayed) parent nodes of $z_{i,t}$, and $\epsilon_{s,t}, \epsilon_{c,t}, \epsilon_{o,t}$ terms are mutually independent (i.e., spatially and temporally independent) random noises in each segment. The vector $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_k^{\text{dyn}}, \boldsymbol{\theta}_k^{\text{obs}})$ embeds the information of domain u_k and has a constant value in each segment, but varies across domains respectively in the transition dynamics and observation.

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

Remark 1. *The mutual independence of noise terms over space and time imply that: (1) $\epsilon_{i,t}$ are independent from the parent nodes $\mathbf{Pa}(z_{i,t})$, and (2) $z_{i,t}$ are conditional independent given the history information $\mathbf{z}_{Hx} = \{\mathbf{z}_{t-\tau}\}_{\tau=1}^L$ up to maximum time lag L and domain embeddings or index.*

This Remark suggests that the Independent Noise (IN) condition [36] and Conditional Independence (CI) condition [21], both of which are common assumptions for causal discovery, are naturally satisfied in time-series data under random noise.

2.2 Identifiability of Time-Delayed Causal Dynamics

We define the identifiability of time-delayed latent causal dynamics in the representation function space. If the latent variables can be identified at least up to permutation and component-wise invertible nonlinearities, we say that latent causal dynamics are also identifiable because conditional independence relations fully characterize time-delayed causal relations in a time-delayed causally sufficient system.

Definition 1 (Identifiable Latent Causal Dynamics)**.** *Formally let $\{\mathbf{x}_t\}_{t=1}^T$ be a sequence of observed variables generated by the true temporally causal latent processes specified by $(f_i, \boldsymbol{\theta}_k, p(\epsilon_i), \mathbf{g})$ given in Eq. 1. A learned generative model $(\hat{f}_i, \hat{\boldsymbol{\theta}}_k, \hat{p}(\epsilon_i), \hat{\mathbf{g}})$ is observationally equivalent to $(f_i, \boldsymbol{\theta}_k, p(\epsilon_i), \mathbf{g})$ if the model distribution $p_{\hat{f},\hat{\boldsymbol{\theta}}_k,\hat{p}_\epsilon,\hat{\mathbf{g}}}(\{\mathbf{x}_t\}_{t=1}^T)$ matches the data distribution $p_{f,\boldsymbol{\theta}_k,p_\epsilon,\mathbf{g}}(\{\mathbf{x}_t\}_{t=1}^T)$ 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:*

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

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

3 Identifiability Theory

In this section, 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 $L = 1$. 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 $p(z_{k,t}|\mathbf{z}_{t-1}, u_k)$ under different domain u_k, and **(3)** under global observation changes, the nonstationarity $p(z_{k,t}|u_k)$ under different values of u_k is exploited. We illustrate through examples that, even without explicit changes in causal influences, the latent causal processes are generally identifiable if the process noises are not perfectly Gaussian. We show that the identifiability results can further benefit from nonstationarity with weaker identifiability conditions. This identifiability theory extends the prior work [27, 24, 20] by allowing causal relations between latent factors, instead of independent sources, and partitioned subspaces with both fixed and/or changing causal influences. The proofs are provided in Appendix S1. The comparisons between existing theories are in Appendix S1.3.

3.1 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

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

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. 11) 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 th 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 [15].

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 under some technical conditions. 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 proposition.

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 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 $\hat{\mathbf{z}}_t$.

3.2 Further Benefits from Changing Causal Influences

[27] 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 our identifiability condition is generally much weaker than that in [27]: 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 [27] 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. 11, 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

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

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

where $\Delta_i^2 = \frac{\partial^2 \eta_{kt}(u_i)}{\partial z_{k,t}^2} - \frac{\partial^2 \eta_{kt}(u_{i-1})}{\partial z_{k,t}^2}$ and $\Delta_i = \frac{\partial \eta_{kt}(u_i)}{\partial z_{k,t}} - \frac{\partial \eta_{kt}(u_{i-1})}{\partial z_{k,t}}$. 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 [27] 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 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 $\mathbf{z}_{-1}$ 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 $\hat{\mathbf{z}}_t$.*

4 LiLY: Learning Latent Causal Dynamics

Given our identifiability results, we propose Learning Latent causal dYnamics (LiLY) framework to estimate the latent causal dynamics under modular distribution shifts, by extending Sequential Variational Auto-Encoders [37] with tailored modules to model different distribution shifts, and enforcing the conditions in Sec. 3 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 S3.2.

Modular Prior Network The core component of the estimation is to learn the three types of distribution changes. 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 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 $[f_s^{-1}, f_c^{-1}, f_o^{-1}]$ to output random process noise from direct cause (lagged states $\mathbf{z}_{\text{Hx}}$) and effect (current states) variables. Modeling of distribution shifts is achieved by adding change factors $[\hat{\theta}_k^{\text{dyn}}, \hat{\theta}_k^{\text{obs}}]$ as external inputs to $[f_s^{-1}, f_c^{-1}, f_o^{-1}]$. Change factors are learnable embeddings of segment index $\mathbf{u}_k$ in Fig. 2C. And hence the distribution shifts are captured and utilized in the implementation.

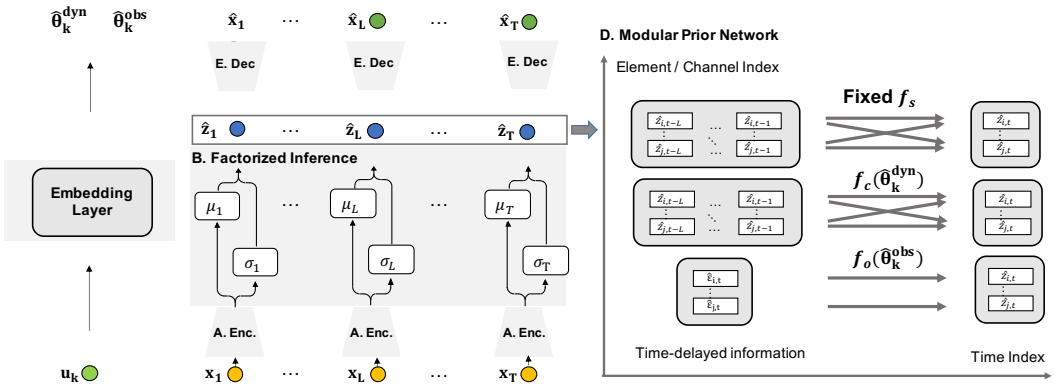

Figure 2: **Framework overview:** Encoder (A) and Decoder (E) using MLP or CNN for specific data types; (B) Factorized inference network that approximates the posteriors of latent variables $\hat{\mathbf{z}}_t$, (C) Change factor representation learning with embedding layer, and (D) Modular prior network 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).

We follow standard conditional normalizing flow formulation [38, 27]. 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: $p(\hat{z}_{s,t}^{\text{fix}}|\hat{\mathbf{z}}_{\text{Hx}}) = p_{\epsilon_s}\left(f_s^{-1}(\hat{z}_{s,t}^{\text{fix}}, \hat{\mathbf{z}}_{\text{Hx}})\right)\left|\frac{\partial f_s^{-1}}{\partial \hat{z}_{s,t}^{\text{fix}}}\right|$; for **2) changing causal dynamics**, we evaluate $p(\hat{z}_{s,t}^{\text{chg}}|\hat{\mathbf{z}}_{\text{Hx}}, \mathbf{u}_k) = p_{\epsilon_c}\left(f_c^{-1}(\hat{z}_{c,t}^{\text{chg}}, \hat{\mathbf{z}}_{\text{Hx}}, \hat{\boldsymbol{\theta}}_k^{\text{dyn}})\right)\left|\frac{\partial f_c^{-1}}{\partial \hat{z}_{c,t}^{\text{chg}}}\right|$ by learning a holistic inverse dynamics f_c^{-1} that takes the estimated change factors for dynamics $\hat{\boldsymbol{\theta}}_k^{\text{dyn}}$) as inputs, and similarly for **3) observation changes** $\hat{\mathbf{z}}_t^{\text{obs}}$, we learn to project them to invariant noise terms by f_o^{-1} which takes the change factors $\boldsymbol{\theta}_k^{\text{obs}}$ as arguments, and obtains $p(\hat{z}_{o,t}^{\text{obs}}|\mathbf{u}_k) = p_{\epsilon_o}\left(f_o^{-1}(\hat{z}_{o,t}^{\text{obs}}, \hat{\boldsymbol{\theta}}_k^{\text{obs}})\right)\left|\frac{\partial f_o^{-1}}{\partial \hat{z}_{o,t}^{\text{obs}}}\right|$ 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. 8. 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.

$$\log p\left(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{\text{Hx}}, \mathbf{u}_k\right) = \underbrace{\sum_{i=1}^{n} \log p(\hat{\epsilon}_i|\mathbf{u}_k)}_{\text{Conditional indepdence}} + \underbrace{\sum_{i=1}^{n} \log\left|\frac{\partial f_i^{-1}}{\partial \hat{z}_{it}}\right|}_{\text{Low-triangular Jacobian matrix}} \tag{8}$$

Factorized Inference We infer the posteriors of each time step $q(\hat{\mathbf{z}}_t|\mathbf{x}_t, \hat{\boldsymbol{\theta}}_k^{\text{obs}})$ using only the observation at that time step and with the change factors $\hat{\boldsymbol{\theta}}_k^{\text{obs}}$ allocated for global changes, 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 into product of these terms.

Change Factor Representation To model distribution changes, instead of learning separate transformation or dynamics in each domain [27], we learn to embed domain index $\mathbf{u}_k$ into low-dimensional change factors $(\hat{\boldsymbol{\theta}}_k^{\text{dyn}}, \hat{\boldsymbol{\theta}}_k^{\text{obs}})$ and insert them as inputs to the inverse dynamics function, or the encoder/decoder function, respectively.

Optimization We train LiLY 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 $\mathcal{L}_{\text{Recon}}$, and use sampling approach to estimate the $\mathcal{L}_{\text{KLD}}$, since with a learned nonparametric modular transition prior, KLD does not have an explicit form.

Connections to Theory To ensure identifiability, what we assume in the theoretical analysis for the estimation framework are the distribution shift modeling through change factors and mutual independence of the estimated noise. The first was achieved by modular prior and the second was enforced by Eq. 8 during the evaluation of transition priors. The other conditions include assumptions over data properties and do not need to be enforced as constraints in the implementation.

5 Experimental Results

We evaluate the identifiability results of *LiLY* on a number of simulated and real-world time-series datasets. We first introduce the evaluation metrics and baselines below.

Evaluation Metrics To evaluate the identifiability of the latent variables, We compute Mean Correlation Coefficient (MCC) on the validation 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 componentwise invertible transformation in the noiseless case.

Baselines The first group of baselines are nonlinear ICA methods: **(1)** BetaVAE [39] which ignores both history and nonstationarity information; **(2)** iVAE [22] and TCL [19] which leverage nonstationarity to establish identifiability but assumes independent factors, and **(3)** SlowVAE [24] and PCL [20] which exploits temporal constraints but assumes independent sources and stationary processes, and **(4)** LEAP [27] which assumes nonstationary, causal processes but models all types of distribution changes in nonstationary noise. Two other disentangled deep state-space models with nonlinear dynamics model, i.e., Kalman VAE (KVAE [40]) and Deep Variational Bayes Filters (DVBF [41]) are also used for comparisons.

5.1 Simulated Experiments

Experiment Setting We generate synthetic datasets that satisfy our identifiability conditions in the theorems following the procedures described in Appendix S2.1.1. In particular, we consider three representative simulation settings respectively 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.

Results 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) cannot recover the latent processes. SlowVAE and PCL distort the results due to independent source assumptions. LEAP, although exploits the history and nonstationarity information, considers only limited nonstationary cases in which the noise distributions are changing, thus cannot recover the causal latent processes under changing causal dynamics and modular distribution shifts. 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.

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								
	LiLY	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

5.2 Real-world Applications

Video Data – Modified Cartpole Environment We evaluate LiLY on the modified cartpole [33] 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 LiLY with two-dimensional change factors $\boldsymbol{\theta}_k^{\text{dyn}}$. We set the latent size $n = 8$ and the lag number $L = 2$. In Fig. 4, the latent causal processes are recovered,

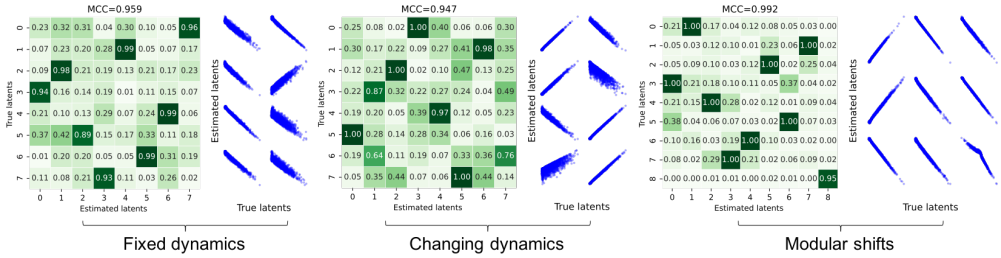

Figure 3: Results for three synthetic datasets: in each block, the left shows the MCC for causally-related and the left are scatterplots between estimated and true factors.

as seen from (a) high MCC for the latent causal processes; (b) the latent factors are estimated up to componentwise transformation; (c) LiLY outperforms the baselines that do not model nonstationarity properly and (d) the latent traversals confirm the two recovered latent variables correspond to the position and pole angle.

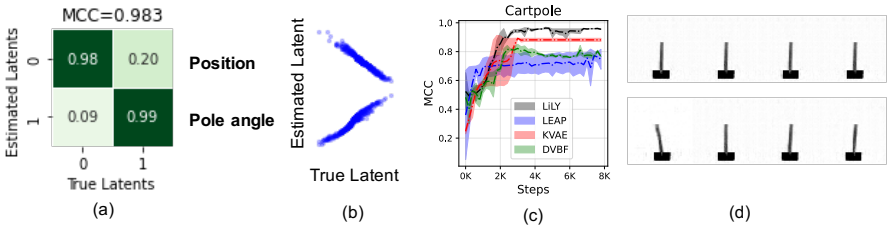

Figure 4: Modified Cartpole results: (a) MCC for causally-related factors; (b) scatterplots between estimated and true factors; (c) baseline comparisons, and (d) latent traversal on a fixed video frame;.

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 LiLY 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. 5(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.

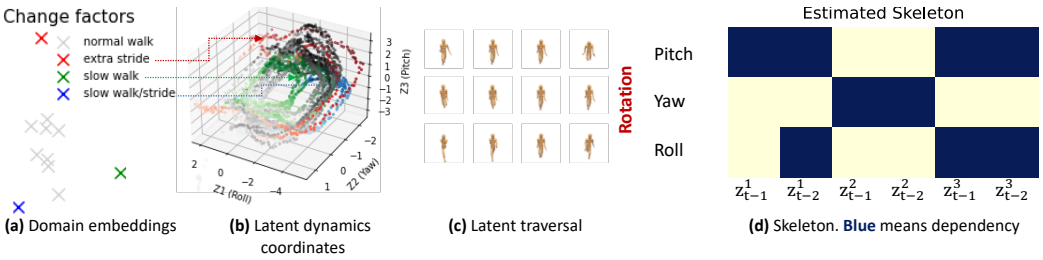

Figure 5: CMU-Mocap results (Subject #8): (a) learned change factors; (b) latent coordinates dynamics for 11 trials; (c)) latent traversal by rendering the reconstructed point clouds into the video frame; (d) estimated causal skeleton (blue means dependency).

6 Conclusion

In this paper, without relying on paricular distribution assumptions such as exponential family distributions or Laplacian noise, we established identifiability theories for nonparametric latent causal processes from their observed nonlinear mixtures under stationary environments and under unknown distribution shifts. The theories have been validated on a number of real-world datasets including videos and motion capture data. 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] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Spring-Verlag Lectures in Statistics, 1993.

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

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

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

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

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

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

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

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

[16] Charles Spearman. Pearson's contribution to the theory of two factors. *British Journal of Psychology*, 19:95–101, 1928.

[17] 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.

[18] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *arXiv preprint arXiv:2010.04917*, 2020.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In *Advances in Neural Information Processing Systems*, 2021.

[32] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020.

[33] 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.

[34] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012.

[35] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

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

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

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

[39] 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.

[40] 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.

[41] 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.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [YES]
   (b) Did you describe the limitations of your work? [YES]
   (c) Did you discuss any potential negative societal impacts of your work? [YES]
   (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? [YES]
   (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? [YES]
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? [YES]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [YES]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [YES]