Supplement to

"Temporally Disentangled Representation Learning under Unknown Nonstationarity"

Appendix organization:

A Identifiability

Assume we observe n-dimensional time-series data at discrete time steps, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, where each $\mathbf{x}_t$ is generated from time-delayed causally related hidden components $\mathbf{z}_t \in \mathbb{R}^n$ by the invertible mixing function:

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t). \tag{1}$$

In addition to latent components $\mathbf{z}_t$, there is an extra hidden component c_t which is a discrete variable with cardinality $\mid c_t \mid = C$, it follows first-order Markov process controlled by a $C \times C$ transition matrix $\mathbf{A}$, in which the i, j-th entry $A_{i,j}$ is the probability to transit from state i to j.

$$c_1, c_2, \ldots, c_t \sim \text{Markov Chain}(\mathbf{A}) \tag{2}$$

For $i \in \{1, \ldots, n\}$, z_{it}, as the i-th component of $\mathbf{z}_t$, is generated by (some) components of history information $\mathbf{z}_{t-1}$, discrete nonstationary indicator c_t, and noise ϵ_{it}.

$$z_{it} = f_i(\{z_{j,t-1} \mid z_{j,t-\tau} \in \mathbf{Pa}(z_{it})\}, c_t, \epsilon_{it}) \quad with \quad \epsilon_{it} \sim p_{\epsilon_i|c_t} \tag{3}$$

where $\mathbf{Pa}(z_{it})$ is the set of latent factors that directly cause z_{it}, which can be any subset of $\mathbf{z}_{\text{Hx}} = \{\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots, \mathbf{z}_{t-L}\}$ up to history information maximum lag L. The components of $\mathbf{z}_t$ are mutually independent conditional on $\mathbf{z}_{\text{Hx}}$ and c_t.

A.1 Identifiability of Nonstationary Hidden States

Theorem 1. *(identifiability of the nonstationarity with Markov Assumptions) Suppose the observed data is generated following the nonlinear ICA framework as defined in Eqs. (1), (2) and (3). And Suppose the following assumptions (Markov Assumptions) hold:*

i For the Markov process, the number of latent states, C, is known.

ii The transition matrix $\mathbf{A}$ is full rank.

Use $\mu_1, \ldots, \mu_C \in \mathbb{R}^n$ to denote nonparametric probability distributions of the C emission distributions $\mu_c = p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, c)$. Then the parameters $\mathbf{A}$ and $M = (\mu_1, \ldots, \mu_C)$ are identifiable given the distribution, $\mathbb{P}^{(3)}_{\mathbf{A},M}$, of at least 4 consecutive observations $\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \mathbf{x}_{t+3}$, up to label swapping of the hidden states, that is:

If $\widetilde{\mathbf{A}}$ is a $C \times C$ transition matrix, if $\widetilde{\pi}(c)$ is a stationary distribution of $\widetilde{\mathbf{A}}$ with $\widetilde{\pi}(c) > 0 \; \forall c \in \{1, \ldots, C\}$, and if $\tilde{M} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_C)$ are C probability distributions on $\mathbb{R}^n$ that verify the equality of the distribution functions $\mathbb{P}^{(3)}_{\widetilde{\mathbf{A}},\tilde{M}} = \mathbb{P}^{(3)}_{\mathbf{A},M}$, then there exists a permutation σ of the set $\{1, \ldots, C\}$ such that for all $k, l = 1, \ldots, C$ we have $\tilde{A}_{k,l} = A_{\sigma(k),\sigma(l)}$ and $\tilde{\mu}_k = \mu_{\sigma(k)}$.

Proof. Suppose we have:

$$\tilde{p}(\mathbf{x}_1, \ldots, \mathbf{x}_T) = p(\mathbf{x}_1, \ldots, \mathbf{x}_T) \tag{4}$$

where $p(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ has transition matrix $\mathbf{A}$ and emission distributions $(\mu_1, \ldots, \mu_C)$, similarly for $\tilde{p}(\mathbf{x}_1, \ldots, \mathbf{x}_T)$.

We consider four consecutive observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and corresponding four discrete elements c_0, c_1, c_2, c_3.

$$\begin{aligned}
p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \mid \mathbf{x}_0) &= \sum_{c_1,c_2,c_3} p(c_1)p(\mathbf{x}_1 \mid \mathbf{x}_0, c_1) \cdot A_{c_1,c_2} p(\mathbf{x}_2 \mid \mathbf{x}_1, c_2) \cdot A_{c_2,c_3} p(\mathbf{x}_3 \mid \mathbf{x}_2, c_3) \\
&= \sum_{c_1,c_2} p(c_1) A_{c_1,c_2} p(\mathbf{x}_1 \mid \mathbf{x}_0, c_1) \cdot p(\mathbf{x}_2 \mid \mathbf{x}_1, c_2) \cdot \left(\sum_{c_3} A_{c_2,c_3} p(\mathbf{x}_3 \mid \mathbf{x}_2, c_3) \right) \\
&= \sum_{c_2} \left(\sum_{c_1} p(c_1) A_{c_1,c_2} p(\mathbf{x}_1 \mid \mathbf{x}_0, c_1) \right) \cdot p(\mathbf{x}_2 \mid \mathbf{x}_1, c_2) \cdot \left(\sum_{c_3} A_{c_2,c_3} p(\mathbf{x}_3 \mid \mathbf{x}_2, c_3) \right) \\
&= \sum_{c_2} \pi_{c_2} \underbrace{\left(\sum_{c_1} \frac{\pi_{c_1} A_{c_1,c_2}}{\pi_{c_2}} \mu_{c_1} \right)}_{\bar{\mu}_{c_2}} \cdot \mu_{c_2} \cdot \underbrace{\left(\sum_{c_3} A_{c_2,c_3} \mu_{c_3} \right)}_{\dot{\mu}_{c_2}}
\end{aligned} \tag{5}$$

where $\pi_{c_i} = p(c_i)$. Since $\mathbf{A}$ has full rank and the probability measures $\mu_1, \ldots, \mu_C$ are linearly independent, the probability measures $\{\bar{\mu}_{c_2} = \sum_{c_1} \frac{\pi_{c_1} A_{c_1,c_2}}{\pi_{c_2}} \mu_{c_1} \,|\, c_2 = 1, \ldots, C\}$ are linearly independent, and the probability measures $\{\dot{\mu}_{c_2} = \sum_{c_3} A_{c_2,c_3} \mu_{c_3} \,|\, c_2 = 1, \ldots, C\}$ are also linearly independent. Thus, applying Theorem 9 of [46], there exists a permutation σ of $\{1, \ldots, C\}$ such that, $\forall i \in \{1, \ldots, C\}$:

$$\tilde{\mu}_i = \mu_{\sigma(i)}$$
$$\sum_j \tilde{A}_{i,j} \tilde{\mu}_j = \sum_j A_{\sigma(i),j} \mu_j$$

This gives easily $\forall i \in \{1, \ldots, C\}$:

$$\sum_j \tilde{A}_{i,j} \mu_{\sigma(j)} = \sum_j A_{\sigma(i),\sigma(j)} \mu_{\sigma(j)}.$$

Since the conditional distributions μ_i are linearly independent, we can establish the equivalence between $\tilde{\mathbf{A}}$ and $\mathbf{A}$ via permutation σ,

$$\tilde{A}_{j,i} = A_{\sigma(j),\sigma(i)}, \tag{6}$$

then the theorem is proved. □

For notational simplicity, and without loss of generality, we assume the components are ordered such that $c = \sigma(c)$. That leads us to the identifiability of the nonstationarity in the system i.e. up to label swapping of the hidden states, the conditional emission distributions $p(\mathbf{x}_t|\mathbf{x}_{t-1}, c_t)$ and transition matrix $\mathbf{A}$ are identifiable, hence providing us a bridge to further leverage the temporal independence condition in the latent space to establish the identifiability result for demixing function or in other words the latent variables $\mathbf{z}_t$.

A.2 Identifiability of Latent Causal Processes

To incorporate nonlinear ICA into the Markov Assumption we define the emission distribution $p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, c)$ as a deep latent variable model. First, the latent independent component variables $\mathbf{z}_t \in \mathbb{R}^n$ are generated from a factorial prior, given the hidden state c_t and previous $\mathbf{z}_{t-1}$, as

$$p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, c_t) = \prod_{k=1}^{n} p(z_{kt} \,|\, \mathbf{z}_{t-1}, c_t). \tag{7}$$

Second, the observed data $\mathbf{x}_t \in \mathbb{R}^n$ is generated by a nonlinear mixing function as in Eq. (1) which is assumed to be bijective with inverse given by $\mathbf{z}_t = \mathbf{g}^{-1}(\mathbf{x}_t)$. Let $\eta_{kt}(c_t) \triangleq \log p(z_{kt}|\mathbf{z}_{t-1}, c_t)$, and assume that $\eta_{kt}(c_t)$ 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 c_t and a subset of $\mathbf{z}_{t-1}$; if $z_{l,t-1}$ is not a parent of z_{kt}, then $\frac{\partial \eta_k}{\partial z_{l,t-1}} = 0$.

Theorem 2. *Suppose there exists an invertible function $\hat{\mathbf{g}}^{-1}$, which is the estimated demixing function that maps $\mathbf{x}_t$ to $\hat{\mathbf{z}}_t$, i.e.,*

$$\hat{\mathbf{z}}_t = \hat{\mathbf{g}}^{-1}(\mathbf{x}_t) \tag{8}$$

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

$$\begin{aligned} \mathbf{v}_{k,t}(c) &\triangleq \Big(\frac{\partial^2 \eta_{kt}(c)}{\partial z_{k,t} \partial z_{1,t-1}}, \frac{\partial^2 \eta_{kt}(c)}{\partial z_{k,t} \partial z_{2,t-1}}, ..., \frac{\partial^2 \eta_{kt}(c)}{\partial z_{k,t} \partial z_{n,t-1}} \Big)^\intercal, \\ \mathring{\mathbf{v}}_{k,t}(c) &\triangleq \Big(\frac{\partial^3 \eta_{kt}(c)}{\partial z_{k,t}^2 \partial z_{1,t-1}}, \frac{\partial^3 \eta_{kt}(c)}{\partial z_{k,t}^2 \partial z_{2,t-1}}, ..., \frac{\partial^3 \eta_{kt}(c)}{\partial z_{k,t}^2 \partial z_{n,t-1}} \Big)^\intercal. \end{aligned} \tag{9}$$

And

$$\begin{aligned} \mathbf{s}_{kt} &\triangleq \Big(\mathbf{v}_{kt}(1)^\intercal, ..., \mathbf{v}_{kt}(C)^\intercal, \frac{\partial^2 \eta_{kt}(2)}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(1)}{\partial z_{kt}^2}, ..., \frac{\partial^2 \eta_{kt}(C)}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(C-1)}{\partial z_{kt}^2} \Big)^\intercal, \\ \mathring{\mathbf{s}}_{kt} &\triangleq \Big(\mathring{\mathbf{v}}_{kt}(1)^\intercal, ..., \mathring{\mathbf{v}}_{kt}(C)^\intercal, \frac{\partial \eta_{kt}(2)}{\partial z_{kt}} - \frac{\partial \eta_{kt}(1)}{\partial z_{kt}}, ..., \frac{\partial \eta_{kt}(C)}{\partial z_{kt}} - \frac{\partial \eta_{kt}(C-1)}{\partial z_{kt}} \Big)^\intercal. \end{aligned} \tag{10}$$

If for each value of $\mathbf{z}_t$, $\mathbf{s}_{1t}, \mathring{\mathbf{s}}_{1t}, \mathbf{v}_{2t}, \mathring{\mathbf{s}}_{2t}, ..., \mathbf{s}_{nt}, \mathring{\mathbf{s}}_{nt}$, as $2n$ function vectors $\mathbf{s}_{k,t}$ and $\mathring{\mathbf{s}}_{k,t}$, with $k = 1, 2, ..., n$, are linearly independent, then $\hat{\mathbf{z}}_t$ must be an invertible, component-wise transformation of a permuted version of $\mathbf{z}_t$.

Proof. Combining (1) and (6) gives $\mathbf{z}_t = (\mathbf{g}^{-1} \circ \hat{\mathbf{g}})(\hat{\mathbf{z}}_t) = \mathbf{h}(\hat{\mathbf{z}}_t)$, where $\mathbf{h} \triangleq \mathbf{g}^{-1} \circ \hat{\mathbf{g}}$. Since both $\mathbf{g}$ and $\hat{\mathbf{g}}$ are invertible, $\mathbf{h}$ is invertible. Let $\mathbf{H}_t$ be the Jacobian matrix of the transformation $\mathbf{h}(\hat{\mathbf{z}}_t)$, 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 previous $\hat{\mathbf{z}}_{t-1}$ and current c_t, then for any $i \neq j$, $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $\hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \{c_t\}$. 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 $\hat{\mathbf{z}}_{t-1}$ and c_t, i.e.,

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

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

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

Combining the above two equations gives

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

i.e., for $i \neq j$, $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $\hat{\mathbf{z}}_{t-1} \cup (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \{c_t\}$.

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

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

assuming the cross second-order derivative exists [47]. Since $p(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1}, c_t) = p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}, c_t) p(\hat{\mathbf{z}}_{t-1}, c_t)$ while $p(\hat{\mathbf{z}}_{t-1}, c_t)$ does not involve $\hat{z}_{it}$ or $\hat{z}_{jt}$, the above equality is equivalent to

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}, c_t)}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0. \tag{11}$$

Then for any c_t, 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 $p(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1} | c_t) = p(\mathbf{z}_t, \mathbf{x}_{t-1} | c_t) \cdot |\mathbf{H}_t|$. Dividing both sides of this equation by $p(\mathbf{x}_{t-1} | c_t)$ gives

$$p(\hat{\mathbf{z}}_t \,|\, \mathbf{x}_{t-1}, c_t) = p(\mathbf{z}_t \,|\, \mathbf{x}_{t-1}, c_t) \cdot |\mathbf{H}_t|. \tag{12}$$

Since $p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, c_t) = p(\mathbf{z}_t \,|\, \mathbf{g}(\mathbf{z}_{t-1}), c_t) = p(\mathbf{z}_t \,|\, \mathbf{x}_{t-1}, c_t)$ and similarly $p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}, c_t) = p(\hat{\mathbf{z}}_t \,|\, \mathbf{x}_{t-1}, c_t)$, Eq. 12 tells us

$$\log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}, c_t) = \log p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, c_t) + \log |\mathbf{H}_t| = \sum_{k=1}^{n} \eta_{kt}(c_t) + \log |\mathbf{H}_t|. \tag{13}$$

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

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

Its second-order cross-derivative is

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}, c_t)}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{n} \Big(\frac{\partial^2 \eta_{kt}(c_t)}{\partial z_{kt}^2} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}(c_t)}{\partial z_{kt}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) - \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}}. \tag{14}$$

The above quantity is always 0 according to Eq. (11). 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,

$$\frac{\partial^3 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}, c_t)}{\partial \hat{z}_{it} \partial \hat{z}_{jt} \partial z_{l,t-1}} = \sum_{k=1}^{n} \Big(\frac{\partial^3 \eta_{kt}(c_t)}{\partial z_{kt}^2 \partial z_{l,t-1}} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial^2 \eta_{kt}(c_t)}{\partial z_{kt} \partial z_{l,t-1}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) \equiv 0, \tag{15}$$

where we have made use of the fact that entries of $\mathbf{H}_t$ do not depend on $z_{l,t-1}$. Using different values r for c_t in Eq. (14) take the difference of this equation across them gives

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; r+1)}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} - \frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; r)}{\partial \hat{z}_{it} \partial \hat{z}_{jt}}$$
$$= \sum_{k=1}^{n} \Big[\Big(\frac{\partial^2 \eta_{kt}(r+1)}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(r)}{\partial z_{kt}^2} \Big) \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \Big(\frac{\partial \eta_{kt}(r+1)}{\partial z_{kt}} - \frac{\partial \eta_{kt}(r)}{\partial z_{kt}} \Big) \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big] \equiv 0. \tag{16}$$

If for any value of $\mathbf{z}_t$, $\mathbf{s}_{1t}, \mathring{\mathbf{s}}_{1t}, \mathbf{s}_{2t}, \mathring{\mathbf{s}}_{2t}, ..., \mathbf{s}_{nt}, \mathring{\mathbf{s}}_{nt}$ 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 $\hat{\mathbf{z}}_t$. □

So far, the identifiability result has been established without observing the nonstationarity indicators such as domain indices.

A.3 Discussion on Assumptions in Theorem 2

This condition was initially introduced in GCL [11], namely, "sufficient variability", to extend the modulated exponential families [9] to general modulated distributions. Essentially, the condition says that the nonstationary domains c 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. Loosely speaking, the sufficient variability holds if the modulation of by c on the conditional distribution $q(z_{it}|\mathbf{z}_{\text{Hx}}, c)$ 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 domains c. 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 [11] and Appendix S1.4.1 of [18]. In summary, we need the domains denoted by c to have diverse (i.e., distinct influences) and complex impacts on the underlying data generation process.

B Implementation Details

B.1 Reproducibility

All experiments are done in a GPU workstation with CPU: Intel i7-13700K, GPU: NVIDIA RTX 4090, Memory: 128 GB. The code can be found via `https://github.com/xiangchensong/nctrl`.

B.2 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):

$$\begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \\ z_{1,t} \\ z_{2,t} \end{bmatrix} = \mathbf{f} \left(\begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \\ \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \right). \tag{17}$$

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:

$$p(z_{1,t-1}, z_{2,t-1}, z_{1,t}, z_{2,t}) = p(z_{1,t-1}, z_{2,t-1}, \epsilon_{1,t}, \epsilon_{2,t}) / \left| \det \mathbf{J}_{\mathbf{f}} \right|, \tag{18}$$

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

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

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. 18 as:

$$\begin{aligned} p(z_{1,t-1}, z_{2,t-1}, z_{1,t}, z_{2,t}) &= p(z_{1,t-1}, z_{2,t-1}) \times p(\epsilon_{1,t}, \epsilon_{2,t}) / \left|\det \mathbf{J_f}\right| \quad \text{(because } \epsilon_t \perp \mathbf{z}_{t-1}\text{)} \\ &= p(z_{1,t-1}, z_{2,t-1}) \times \prod_i p(\epsilon_{i,t}) / \left|\det \mathbf{J_f}\right| \quad \text{(because } \epsilon_{1,t} \perp \epsilon_{2,t}\text{)} \end{aligned} \tag{19}$$

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:

$$p(z_{1,t}, z_{2,t} | z_{1,t-1}, z_{2,t-1}) = \prod_i p(\epsilon_{i,t}) / \left|\det \mathbf{J_f}\right| = \prod_i p(\epsilon_{i,t}) \times \left|\det \mathbf{J_f^{-1}}\right|. \tag{20}$$

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

Let $\{f_i^{-1}\}_{i=1,2,3...}$ be a set of learned inverse transition functions that take the estimated latent causal variables, and output the noise terms, i.e., $\hat{\epsilon}_{i,t} = f_i^{-1}\left(\hat{z}_{i,t}, \{\hat{\mathbf{z}}_{t-\tau}, c_t\}\right)$.

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

$$\underbrace{\left[\hat{\mathbf{z}}_{t-L}, \ldots, \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t\right]^\top}_{\mathbf{A}} \text{ mapped to } \underbrace{\left[\hat{\mathbf{z}}_{t-L}, \ldots, \hat{\mathbf{z}}_{t-1}, \hat{\epsilon}_{i,t}\right]^\top}_{\mathbf{B}}, \; with \; \mathbf{J}_{\mathbf{A}\to\mathbf{B}} = \begin{pmatrix} \mathbb{I}_{nL} & 0 \\ * & \text{diag}\left(\frac{\partial f_{i,j}^{-1}}{\partial \hat{z}_{jt}}\right) \end{pmatrix}. \tag{21}$$

Similar to Eq. 20, we can obtain the joint distribution of the estimated dynamics subspace as:

$$\log p(\mathbf{A}) = \underbrace{\log p\left(\hat{\mathbf{z}}_{t-L}, \ldots, \hat{\mathbf{z}}_{t-1}\right) + \sum_{j=1}^{n} \log p(\hat{\epsilon}_{i,t})}_{\text{Because of mutually independent noise assumption}} + \log\left(\left|\det\left(\mathbf{J}_{\mathbf{A}\to\mathbf{B}}\right)\right|\right). \tag{22}$$

$$\log p\left(\hat{\mathbf{z}}_t | \{\hat{\mathbf{z}}_{t-\tau}\}_{\tau=1}^{L}, c_t\right) = \sum_{j=1}^{n} \log p(\hat{\epsilon}_{i,t} | c_t) + \sum_{i=j}^{n} \log \left|\frac{\partial f_i^{-1}}{\partial \hat{z}_{i,t}}\right| \tag{23}$$

B.3 Derivation of ELBO

Then the second part is to maximize the Evidence Lower BOund (ELBO) for the VAE framework, which can be written as:

$$
\begin{aligned}
\text{ELBO} \triangleq & \log p_{\text{data}}(\mathbf{X}) - D_{KL}(q_\phi(\mathbf{Z}|\mathbf{X})||p_{\text{data}}(\mathbf{Z}|\mathbf{X})) \\
= & \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{X})} \log p_{\text{data}}(\mathbf{X}|\mathbf{Z}) - D_{KL}(q_\phi(\mathbf{Z}|\mathbf{X})||p_{\text{data}}(\mathbf{Z}|\mathbf{X})) \\
= & \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{X})} \log p_{\text{data}}(\mathbf{X}|\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{X})} \left[\log q_\phi(\mathbf{Z}|\mathbf{X}) - \log p_{\text{data}}(\mathbf{Z})\right] \\
= & \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{X})} \left[\log p_{\text{data}}(\mathbf{X}|\mathbf{Z}) + \underbrace{\log p_{\text{data}}(\mathbf{Z})}_{\mathbb{E}_{\mathbf{c}}\left[\sum_{t=1}^{T} \log p(\mathbf{z}_t|\mathbf{z}_{t-1}, c_t)\right]} - \log q_\phi(\mathbf{Z}|\mathbf{X}) \right] \\
= & \mathbb{E}_{\mathbf{z}_t} \left[\underbrace{\sum_{t=1}^{T} \log p_{\text{data}}(\mathbf{x}_t|\mathbf{z}_t)}_{-\mathcal{L}_{\text{Recon}}} + \underbrace{\mathbb{E}_{\mathbf{c}} \left[\sum_{t=1}^{T} \log p_{\text{data}}(\mathbf{z}_t|\mathbf{z}_{\text{Hx}}, c_t) \right] - \sum_{t=1}^{T} \log q_\phi(\mathbf{z}_t|\mathbf{x}_t)}_{-\mathcal{L}_{\text{KLD}}} \right]
\end{aligned}
\tag{24}
$$

B.4 Synthetic Dataset Generation

We generated two synthetic datasets (A and B) with different nonlinear mixing functions. In this section we will introduce the detailed implementation of the generation. The generation can be split into steps (1) sample c_t from a Markov chain, (2) generate $\mathbf{z}_t$ with different transition functions f_{c_t} with respect to c_t, and (3) generate observation $\mathbf{x}_t$ via mixing function $\mathbf{g}$.

B.4.1 Sample c_t from Markov chain

We first randomly initialized a Markov chain with transition matrix $\mathbf{A}$ and sample 20,000 steps.

B.4.2 Generation of latent variables $\mathbf{z}_t$

We first randomly initialized $|C| = 5$ different transition functions $\{f_1, f_2, \ldots, f_{|C|}\}$ with different MLPs, and generate $\mathbf{z}_t = f_{c_t}(\mathbf{z}_{\text{Hx}})$. The dimensions are set to 8 for fair comparison.

B.4.3 Generation of observations $\mathbf{x}_t$

The difference between datasets A and B is the mixing function. We use a two-layer randomly initialized MLP for dataset A and a three-layer MLP for dataset B. For each linear layer in the MLP, we use condition number of the weight matrix to filter out ones that are not "invertible".

B.5 Modified CartPole Dataset Generation

Similar to the synthetic datasets, we also sample from a Markov chain and get c_t. For the modified CartPole, we initialized 5 different environments which have different combinations of hyperparameters such as gravity, pole mass, etc. A detailed comparison is listed in Table 1.

Table 1: Different configs for different Modified CartPole environments.

Environment ID	Gravity	Pole Mass	Noise Scale
0	9.8	0.2	0.01
1	24.79	0.5	0.01
2	3.7	1.0	0.01
3	11.15	1.5	0.01
4	0.62	2.0	0.01

At each time step t the environment will load the corresponding hyperparameters for given c_t and update the states $\mathbf{z}_t$ according to the configuration given c_t. The nonlinear mixing function from states to observations $\mathbf{x}_t$ is fixed by a rendering method in the gym package.

B.6 MoSeq Dataset

In the MoSeq dataset, the observations $\mathbf{x}_t$ are taken to be the first 10 principal components of depth camera video data of mice exploring an open field. The dataset consists of 20-minute depth camera recordings of 24 mice. In preprocessing, the videos are cropped and centered around the mouse centroid and then filtered to remove recording artifacts. Finally, the preprocessed video is projected onto the top principal components to obtain a 10-dimensional time series.

B.7 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.

B.8 Network Architecture

We summarize our network architecture below and describe it in detail in Table 2 and Table 3.

Table 2: 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**
ARHMM	Autoregressive HMM for Synthetic Data	
Input: $\mathbf{x}_{1:T}$	Observed time series	BS $\times$ T $\times$ i_dim
Emission Module	Compute $\mu_{\mathbf{z}_{t+1}}, \sigma_{\mathbf{z}_{t+1}}$	BS $\times$ T $\times$ 2 $\times$ z_dim
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
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
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	$\hat{\mathbf{z}}_{1:T}$
Prior Network	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

Table 3: 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**
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
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