LEARNING TEMPORAL CAUSAL REPRESENTATION UNDER NON-INVERTIBLE GENERATION PROCESS SUPPLEMENTARY MATERIALS

Anonymous authors
Paper under double-blind review

A1 IDENTIFIABILITY THEORY

A1.1 PROOF FOR THEOREM 1

Let us first shed light on the identifiability theory on the special case with $\tau = 1$, i.e.,

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t), \quad z_{it} = f_i\left(\mathbf{z}_{t-1}, \epsilon_{it}\right), \quad \mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu}). \tag{1}$$

Theorem A1 (Identifiability under Non-invertible Generative Process)**.** *For a series of observations $\mathbf{x}_t$ and estimated latent variables $\hat{\mathbf{z}}_t$, suppose there exists function $\hat{\mathbf{g}}, \hat{\mathbf{m}}$ which subject to*

$$\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_t), \quad \hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{x}_{t:t-\mu}). \tag{2}$$

If assumptions

- *(conditional independence) the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$,*
- *(sufficiency) let $\eta_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_{t-1})$, and*

$$\begin{aligned} \mathbf{v}_{k,t} &\triangleq \Big(\frac{\partial^2 \eta_{kt}}{\partial z_{k,t}\partial z_{1,t-1}}, \frac{\partial^2 \eta_{kt}}{\partial z_{k,t}\partial z_{2,t-1}}, ..., \frac{\partial^2 \eta_{kt}}{\partial z_{k,t}\partial z_{n,t-1}}, \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}\Big)^\intercal \\ \mathring{\mathbf{v}}_{k,t} &\triangleq \Big(\mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}, \frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{1,t-1}}, \frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{2,t-1}}, ..., \frac{\partial^3 \eta_{kt}}{\partial z_{k,t}^2 \partial z_{n,t-1}}\Big)^\intercal, \end{aligned} \tag{3}$$

  for each value of $\mathbf{z}_t$, $\mathbf{v}_{1t}, \mathring{\mathbf{v}}_{1t}, \mathbf{v}_{2t}, \mathring{\mathbf{v}}_{2t}, ..., \mathbf{v}_{nt}, \mathring{\mathbf{v}}_{nt} \in \mathbf{R}^{2n}$, as $2n$ vector functions in $z_{1,t-1}$, $z_{2,t-1}$, ..., $z_{n,t-1}$, are linearly independent,

- *(continuity) $\hat{\mathbf{z}}$ is defined on a continuous manifold, and $\mathbf{m}, \hat{\mathbf{m}}, \mathbf{g}, \hat{\mathbf{g}}$ are secondary differentiable, i.e., $\frac{\partial^2 z_i}{\partial a \partial b}$ for $a, b \in \{\hat{z}_{i,t} \mid \forall i\} \cup \{x_{i,j} \mid \forall i, \forall j = t, t-1, \cdots, t-\mu\}$ exists,*

are satisfied, then $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$ with regard to context $\{\mathbf{x}_j \mid \forall j = t, t-1, \cdots, t-\mu\}$.

Proof. For any t, combining Eq 1 and Eq 2 gives

$$\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu}) = \mathbf{m}(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t), \mathbf{x}_{t-1:t-\mu}). \tag{4}$$

as well as $\hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{g}(\mathbf{z}_t), \mathbf{x}_{t-1:t-\mu})$ similarly. Upon Eq 4, we have an unified partially invertible function $\mathbf{z}_t = \mathbf{h}(\hat{\mathbf{z}}_t|\mathbf{x}_{t-1:t-\mu})$ where $\mathbf{h} = \mathbf{m} \circ \hat{\mathbf{g}}$ with Jacobian $\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} = \mathbf{H}_t(\hat{\mathbf{z}}_t; \mathbf{x}_{t-1:t-\mu})$. By *partially invertible* it means that $\mathbf{z}$ and $\hat{\mathbf{z}}$ are in one-to-one correspondence for any context observations $\mathbf{x}_{t-1:t-\mu}$ that are fixed. Let us consider the mapping from joint distribution $(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-1})$ to $(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-1})$, i.e.,

$$P(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-1}) = P(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-1}) \,/\, |\mathbf{J}_t|, \tag{5}$$

where

$$\mathbf{J}_t = \begin{bmatrix} \frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} & \mathbf{0} \\ * & \mathbf{I} \end{bmatrix}. \tag{6}$$

which is a lower triangle matrix, where $\mathbf{I}$ infers eye matrix and $*$ infers any possible matrix. Thus, we have determinant $|\mathbf{J}_t| = |\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t}| = |\mathbf{H}_t|$. Dividing both sides of Eq 5 by $P(\mathbf{x}_{t-1:t-\mu-1})$ gives

$$\textbf{LHS} = P(\mathbf{z}_t|\mathbf{x}_{t-1:t-\mu-1}) = P(\mathbf{z}_t|\mathbf{z}_{t-1}), \tag{7}$$

since $\mathbf{z}_t$ and $\mathbf{x}_{t-1:t-\mu-1}$ are independent conditioned on $\mathbf{z}_{t-1}$. Similarly we can derive the estimated distribution $\textbf{RHS} = P(\hat{\mathbf{z}}_t|\mathbf{x}_{t-1:t-\mu-1}) = P(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1})$ as well, which yields to

$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) = P(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}) \,/\, |\mathbf{H}_t|. \tag{8}$$

From a direct observation, if the components of $\hat{\mathbf{z}}_t$ are mutually independent given $\hat{\mathbf{z}}_{t-1}$,, then for any distinct $i \neq j$, $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $(\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1}$. This mutual independence of the components of $\hat{\mathbf{z}}_t$ based on $\hat{\mathbf{z}}_{t-1}$ implies two things:

- $\hat{z}_{it}$ is independent from $\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}$ conditional on $\hat{\mathbf{z}}_{t-1}$. Formally,
$$p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1}) = p(\hat{z}_{it} \,|\, (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1}).$$
- $\hat{z}_{it}$ is independent from $\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\}$ conditional on $\hat{\mathbf{z}}_{t-1}$. Represented as:
$$p(\hat{z}_{it} \,|\, \hat{\mathbf{z}}_{t-1}) = p(\hat{z}_{it} \,|\, (\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}\}) \cup \hat{\mathbf{z}}_{t-1}).$$

From these two equations, we can derive:

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

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

Leveraging an inherent fact, i.e., if $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}) \cup \hat{\mathbf{z}}_{t-1}$, the subsequent equation arises:

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

assuming the cross second-order derivative exists. Given that $p(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1}) = p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}) p(\hat{\mathbf{z}}_{t-1})$ and $p(\hat{\mathbf{z}}_{t-1})$ remains independent of $\hat{z}_{it}$ or $\hat{z}_{jt}$, the above equality is equivalent to

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0. \tag{9}$$

Referencing Eq 7, it gets expressed as:

$$\log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1}) = \log p(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}) + \log |\mathbf{H}_t| = \sum_{k=1}^{n} \eta_{kt} + \log |\mathbf{H}_t|. \tag{10}$$

The partial derivative w.r.t. $\hat{z}_{it}$ is presented below:

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

The second-order cross derivative can be depicted as:

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \,|\, \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{n} \Big(\frac{\partial^2 \eta_{kt}}{\partial z_{kt}^2} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) + \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}}. \tag{11}$$

According to Eq 9, the right-hand side of the presented equation consistently equals 0. Therefore, for each index l ranging from 1 to n, and every associated value of $z_{l,t-1}$, its partial derivative with respect to $z_{l,t-1}$ remains 0. That is,

$$\sum_{k=1}^{n} \Big(\frac{\partial^3 \eta_{kt}}{\partial z_{kt}^2 \partial z_{l,t-1}} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial^2 \eta_{kt}}{\partial z_{kt} \partial z_{l,t-1}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) \equiv 0, \tag{12}$$

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

Considering any given 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 $i \neq j$. In other words, each row of $\mathbf{H}t$ consists of at most a single non-zero entry. Given that $\mathbf{h}$ has continuous domain and is partially invertible with regard to context $\{\mathbf{x}_j \mid \forall j \neq t\}$, according to Lemma A2, $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$ with regard to context.

□

A1.2 Extension to Multiple Transition Time Lag τ

For the sake of simplicity, we consider only one special case with $\tau = 1$ in Theorem A1. Our identifiability theorem can be actually extended to arbitrary lags directly. For any given τ, according to modularity we have different conclusion at Eq 7 as $P(\mathbf{z}_t|\mathbf{x}_{t-1:t-\mu-\tau}) = P(\mathbf{z}_t|\mathbf{z}_{t-1:t-\tau})$. Similarity $\mathbf{RHS} = P(\hat{\mathbf{z}}_t|\mathbf{x}_{t-1:t-\mu-\tau}) = P(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1:t-\tau})$ holds true as well. In addition, some modifications are needed in sufficiency assumption, i.e., re-define $\eta_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_{t-1:t-\tau})$ and there should be at least $2n$ linear independent vectors for $\mathbf{v}, \mathring{\mathbf{v}}$ with regard to $z_{l,\eta}$ where $l = 1, 2, \cdots, n$ and $t - \tau \leq \mu \leq t - 1$. No extra changes are needed.

A1.3 Identifiability under Time-Delayed Mixing Process

As a more general case, the non-invertibility can be introduced by the neighboring latent variables. When the effect from other latent variables diminishes, this setting will be reduced to the basic scenario as described in Theorem A1. To formalize this problem, consider a time-delayed mixing generative process with a transition lag of $\tau = 2$ and a mixing lag of $r = 1$:

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t; \mathbf{z}_{t-1}), \quad z_{it} = f_i\left(\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \epsilon_{it}\right), \tag{13}$$

where $\mathbf{z}_t$ can be recovered by current observation as well as μ previous observations $\mathbf{x}_{t:t-\mu}$, i.e.,

$$\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu}). \tag{14}$$

Corollary A1 (Identifiability under Time-Delayed Mixing Process)**.** *For a series of observations* $\mathbf{x}$ *and estimated latent variables* $\hat{\mathbf{z}}_t$*, suppose there exists function* $\hat{\mathbf{g}}$ *and* $\hat{\mathbf{m}}$ *which satisfies*

$$\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_t; \hat{\mathbf{z}}_{t-1}), \quad \hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{x}_{t:t-\mu}). \tag{15}$$

If assumptions

- *(conditional independence) the components of* $\hat{\mathbf{z}}_t$ *are mutually independent conditional on* $\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_{t-2}$*,*
- *(sufficiency) let* $\phi_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_{t-1}, \mathbf{z}_{t-2})$*, and*

$$\begin{aligned} \mathbf{v}_{k,t} &\triangleq \Big(\frac{\partial^2 \phi_{kt}}{\partial z_{k,t} \partial z_{1,t-2}}, \frac{\partial^2 \phi_{kt}}{\partial z_{k,t} \partial z_{2,t-2}}, ..., \frac{\partial^2 \phi_{kt}}{\partial z_{k,t} \partial z_{n,t-2}}, \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0} \Big)^\intercal \\ \mathring{\mathbf{v}}_{k,t} &\triangleq \Big(\mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}, \frac{\partial^3 \phi_{kt}}{\partial z_{k,t}^2 \partial z_{1,t-2}}, \frac{\partial^3 \phi_{kt}}{\partial z_{k,t}^2 \partial z_{2,t-2}}, ..., \frac{\partial^3 \phi_{kt}}{\partial z_{k,t}^2 \partial z_{n,t-2}} \Big)^\intercal, \end{aligned} \tag{16}$$

  for each value of $\mathbf{z}_t, \mathbf{v}_{1t}, \mathring{\mathbf{v}}_{1t}, \mathbf{v}_{2t}, \mathring{\mathbf{v}}_{2t}, ..., \mathbf{v}_{nt}, \mathring{\mathbf{v}}_{nt} \in \mathbf{R}^{2n}$*, as* $2n$ *vector functions in* $z_{1,t-2}$*,* $z_{2,t-2}$*, ...,* $z_{n,t-2}$*, are linearly independent,*

- *(continuity)* $\hat{\mathbf{z}}$ *is defined on a continuous manifold, and* $\mathbf{m}, \hat{\mathbf{m}}, \mathbf{g}, \hat{\mathbf{g}}$ *are secondary differentiable, i.e.,* $\frac{\partial^2 z_i}{\partial a \partial b}$ *for* $a, b \in \{\hat{z}_{i,t} \mid \forall i\} \cup \{x_{i,j} \mid \forall i, \forall j = t, t-1, \cdots, t-\mu-1\}$ *exists,*

are satisfied, then $\mathbf{z}_t$ *must be a component-wise transformation of a permuted version of* $\hat{\mathbf{z}}_t$ *with regard to context* $\{\mathbf{x}_j \mid \forall j = t, t-1, \cdots, t-\mu-1\}$.

Proof. For any t, combining Eq 13 and Eq 14 gives

$$\begin{aligned}\mathbf{z}_t &= \mathbf{m}(\mathbf{x}_{t:t-\mu}) \\ &= \mathbf{m}(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1}), \mathbf{x}_{t-1:t-\mu}) \\ &= \mathbf{m}(\hat{\mathbf{g}}(\hat{\mathbf{z}}_t, \hat{\mathbf{m}}(\mathbf{x}_{t-1:t-\mu-1})), \mathbf{x}_{t-1:t-\mu}),\end{aligned} \tag{17}$$

as well as $\hat{\mathbf{z}}_t = \hat{\mathbf{m}}(\mathbf{g}(\mathbf{z}_t, \mathbf{m}(\mathbf{x}_{t-1:t-\mu-1})), \mathbf{x}_{t-1:t-\mu})$ similarly. Upon Eq 17, we have an unified partially invertible function $\mathbf{z}_t = \mathbf{h}(\hat{\mathbf{z}}_t | \mathbf{x}_{t-1:t-\mu-1})$ where $\mathbf{h} = \mathbf{m} \circ \hat{\mathbf{g}}$ with Jacobian $\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} = \mathbf{H}_t(\hat{\mathbf{z}}_t; \mathbf{x}_{t-1:t-\mu-1})$. By *partially invertible* it means that $\mathbf{z}$ and $\hat{\mathbf{z}}$ are in one-to-one correspondence for any context observations $\mathbf{x}_{t-1:t-\mu-1}$ that are fixed. Let us consider the mapping from joint distribution $(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-2})$ to $(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-2})$, i.e.,

$$P(\mathbf{z}_t, \mathbf{x}_{t-1:t-\mu-2}) = P(\hat{\mathbf{z}}_t, \mathbf{x}_{t-1:t-\mu-2}) \,/\, |\mathbf{J}_t|, \tag{18}$$

where

$$\mathbf{J}_t = \begin{bmatrix} \frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t} & \mathbf{0} \\ * & \mathbf{I} \end{bmatrix}, \tag{19}$$

which is a lower triangle matrix, where $\mathbf{I}$ infers eye matrix and $*$ infers any possible matrix. Thus, we have determinant $|\mathbf{J}_t| = |\frac{\partial \mathbf{z}_t}{\partial \hat{\mathbf{z}}_t}| = |\mathbf{H}_t|$. Dividing both sides of Eq 18 by $P(\mathbf{x}_{t-1:t-\mu-2})$ gives

$$\textbf{LHS} = P(\mathbf{z}_t | \mathbf{x}_{t-1:t-\mu-2}) = P(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}), \tag{20}$$

since $\mathbf{z}_t$ and $\mathbf{x}_{t-1:t-\mu-2}$ are independent conditioned on $\mathbf{z}_{t-1}, \mathbf{z}_{t-2}$. Similarly, $\textbf{RHS} = P(\hat{\mathbf{z}}_t | \mathbf{x}_{t-1:t-\mu-2}) = P(\hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_{t-2})$ holds true as well, which yields to

$$P(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}) = P(\hat{\mathbf{z}}_t | \hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_{t-2}) \,/\, |\mathbf{H}_t|. \tag{21}$$

The rest part of proof is very similar to its counterpart in Theorem A1 with transition lag of $\tau = 2$, so we simply omit the same part. The only difference to be notified is that since the partially invertible function $\mathbf{h}$ is defined on $\mathbf{x}_{t-1:t-\mu-1}$, which leads to an effect that the Jacobian matrix $\mathbf{H}_t$ is a function of $\mathbf{z}_t$, according to $\mathbf{z}_{t-1} = \mathbf{m}(\mathbf{x}_{t-1:t-\mu-1})$.

In this case, when it comes to the secondary derivative equation

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{n} \Big(\frac{\partial^2 \eta_{kt}}{\partial z_{kt}^2} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}}{\partial z_{kt}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) + \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}}, \tag{22}$$

where the Jacobian item $\frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}}$ cannot be eliminated by derive it with respect to $z_{l,t-1}$. A further preceding time step is needed to enforce the Jacobian item to be zero as mentioned in the sufficiency assumption, i.e.,

$$\frac{\partial^3 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt} \partial z_{l,t-2}} = 0. \tag{23}$$

□

Note that in the proof of Corollary A1, we require the transition lag $\tau = 2$ to be larger than the mixing lag $r = 1$. As long as this inequality $\tau > r$ is satisfied, the parameters τ and μ can be extended to arbitrary numbers following a similar modification in Appendix A1.2.

A1.4 NECESSITY OF CONTINUITY

Let us first give an extreme example to illustrate the importance of extra constraints for identifiability. Consider 4 independent random variables u, v, x, y subjects to standard normal distribution respectively. Suppose that there exist an invertible function $(x, y) = \mathbf{h}(u, v)$ satisfies

$$\begin{cases} x = \mathbb{I}(x + y > 0) \cdot u + \mathbb{I}(x + y \leq 0) \cdot v \\ y = \mathbb{I}(x + y > 0) \cdot v + \mathbb{I}(x + y \leq 0) \cdot u. \end{cases} \tag{24}$$

Notice that the Jacobian from (u, v) to (x, y) contains at most one non-zero entry for each column or row. However, the result (x, y) is still entangled, and the identifiability of (u, v) is not achieved. What if now we notate latent variable as $\hat{\mathbf{z}} = (u, v)$, estimated latent variable as $\mathbf{z} = (x, y)$ and the transition process with two mixing functions as $\mathbf{h} = \mathbf{g}^{-1} \circ \hat{\mathbf{g}}$?

In the literature of nonlinear ICA, the gap between $\mathbf{H}_{ij} \cdot \mathbf{H}_{ik} = 0$ when $j \neq k$ and identifiability is ill-discussed. In linear ICA, since the Jacobian is a constant matrix, these two statements are equivalent. Nevertheless, in nonlinear ICA, $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$ is not a constant, but a function of $\hat{\mathbf{z}}$, which may leads to the failure of identifiability as shown in Eq 24.

The counterexamples can still be easily constructed even if function $\mathbf{h}$ is continuous. For brevity, let us denote a segment-wise linear indicator function as $f(u, v) = \min(\max(0, u + v + 0.5), 1)$, and we have $\mathbf{h}$ as

$$\begin{cases} x = f(u, v) \cdot u + (1 - f(u, v)) \cdot v \\ y = f(u, v) \cdot v + (1 - f(u, v)) \cdot u. \end{cases} \tag{25}$$

When u, v, x, y are independent uniform distributions on $[-2, -1] \cup [1, 2]$, all conditions are still satisfied while the identifiability cannot be achieved.

To fill this gap, two more assumptions are needed. The domain $\hat{\mathcal{Z}}$ of $\hat{\mathbf{z}}$ should be continuous, i.e., for any $\hat{\mathbf{z}}^{(1)}, \hat{\mathbf{z}}^{(2)} \in \hat{\mathcal{Z}}$, there exists a continuous path connecting $\hat{\mathbf{z}}^{(1)}$ and $\hat{\mathbf{z}}^{(2)}$ with all points of the path are in $\hat{\mathcal{Z}}$. In addition, function $\mathbf{h}$ should be second-order differentiable.

Lemma A1 (Disentanglement with Continuity)**.** *For $\hat{\mathbf{z}}$ defined on continuous domain $\hat{\mathcal{Z}} \subset \mathbb{R}^n$ and second order differentiable invertible function $\mathbf{h}$ which satisfies $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}})$, if there exists at most one non-zero entry in each row of the Jacobian matrix $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$, $\hat{\mathbf{z}}$ is a disentangled version of $\mathbf{z}$ up to a permutation and a element-wise nonlinear operation.*

Proof. According to Inverse function theorem, since the inverse function exists at a point $\hat{\mathbf{z}}$, the derivative $\mathbf{h}'(\hat{\mathbf{z}})$ is invertible at $\hat{\mathbf{z}}$ and the determinant of the Jacobian matrix $\mathbf{H}$ at $\hat{\mathbf{z}}$ is of full rank. In addition, since $\hat{\mathbf{z}}$ is defined on continuous domain and $\mathbf{h}$ is second order differentiable, the range of $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}$ for any i is continuous. That is, the range for $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}$ is defined on the n-dimensional axis except $\mathbf{0}$, which leads to 2^n separated blocks.

If there exist two $\hat{\mathbf{z}}^{(1)}$ and $\hat{\mathbf{z}}^{(2)}$ with different entries $j \neq k$ subjects to

$$\begin{cases} \frac{\partial z_i^{(1)}}{\partial \hat{z}_j^{(1)}} = 0, \quad \frac{\partial z_i^{(1)}}{\partial \hat{z}_k^{(1)}} = a \neq 0, \\ \frac{\partial z_i^{(2)}}{\partial \hat{z}_j^{(2)}} = b \neq 0, \quad \frac{\partial z_i^{(1)}}{\partial \hat{z}_k^{(1)}} = 0, \end{cases} \tag{26}$$

who belongs to 2 different blocks, the assumption will be violated that there exists at least one path connecting $\hat{\mathbf{z}}^{(1)}$ and $\hat{\mathbf{z}}^{(2)}$ without stepping into $\{\mathbf{0}\} \cup \{\mathbf{x} | \exists j \neq k, x_j \cdot x_k \neq 0\}$. Thus, such a case is not allowed, and the identifiability is assured. □

When it comes to partially invertible function with regard to side information $\mathbf{c}$, the proof is the same with only a modification on conditions. That is, the continuous domain assumption is applied to $(\mathbf{z}, \mathbf{c})$, and the second order differentiable is extended to both $\mathbf{z}$ and $\mathbf{c}$, i.e., $\frac{\partial^2 z_i}{\partial a \partial b}$ for $a, b \in \{z | \mathbf{z}_i\} \cup \{c | \mathbf{c}_i\}$ when $a \neq b$ exists.

Lemma A2 (Disentanglement with Continuously for Partially Invertible Function)**.** *For $\hat{\mathbf{z}}, \mathbf{c}$ defined on a continuous domain $\hat{\mathcal{Z}} \subset \mathbb{R}^n, \mathcal{C} \subset \mathbb{R}^m$ respectively and second order differentiable partially*

invertible function $\mathbf{h}$ *which satisfies* $\mathbf{z} = \mathbf{h}(\hat{\mathbf{z}}, \mathbf{c})$*, i.e.,* $\frac{\partial^2 z_i}{\partial a \partial b}$ *for* $a, b \in \{z|\mathbf{z}_i\} \cup \{c|\mathbf{c}_i\}$ *when* $a \neq b$ *exists, if there exists at most one non-zero entry in each row of the Jacobian matrix* $\mathbf{H} = \frac{\partial \mathbf{z}}{\partial \hat{\mathbf{z}}}$*,* $\hat{\mathbf{z}}$ *is a disentangled version of* $\mathbf{z}$ *up to a permutation and an element-wise nonlinear operation.*

Proof. Similar to Lemma A1, the range for $\frac{\partial z_i}{\partial \hat{\mathbf{z}}}$ is defined on the n-dimensional axis except $\mathbf{0}$, which leads to 2^n separated blocks. If there exist two pairs of $(\hat{\mathbf{z}}^{(1)}, \mathbf{c}^{(1)})$ and $(\hat{\mathbf{z}}^{(2)}, \mathbf{c}^{(2)})$ with different entries $j \neq k$ subjects to

$$\begin{cases} \frac{\partial z_i^{(1)}}{\partial \hat{z}_j^{(1)}} = 0, \quad \frac{\partial z_i^{(1)}}{\partial \hat{z}_k^{(1)}} = a \neq 0, \\ \frac{\partial z_i^{(2)}}{\partial \hat{z}_j^{(2)}} = b \neq 0, \quad \frac{\partial z_i^{(1)}}{\partial \hat{z}_k^{(1)}} = 0, \end{cases} \tag{27}$$

who belongs to 2 different blocks, the assumption will be violated as well. Thus, such a case is not allowed, and the identifiability is assured. □

A1.5 Identifiability Benefits from Non-Stationarity

We can further leverage the advantage of non-stationary data for identifiability. Let $\mathbf{v}_{kt}(u_r)$ be $\mathbf{v}_{kt}$, which is defined in Eq 3, 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 identifiability of $\mathbf{z}_t$ is guaranteed by the linear independence of the whole function vectors $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$, with $k = 1, 2, ..., n$. This linear independence is generally a much stronger condition.

Corollary A2 (Identifiability under Non-Stationary Process)**.** *Suppose* $\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$*,* $\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu})$ *and that the conditional distribution* $p(z_{k,t} \mid \mathbf{z}_{t-1}, \mathbf{u})$ *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* $\mathbf{z}_{t-1}$ *in each context. Assume that the components of* $\hat{\mathbf{z}}_t$ *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. Drawing upon the arguments in the proof of Theorem A1, given that the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$, we know that for $i \neq j$,

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; \mathbf{u})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{n} \Big(\frac{\partial^2 \eta_{kt}(\mathbf{u})}{\partial z_{kt}^2} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}(\mathbf{u})}{\partial z_{kt}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) - \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} \equiv 0. \tag{28}$$

In contrast to Eq 11, we now allow $p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1})$ to depend on $\mathbf{u}$. Given that the aforementioned equation is always 0, its partial derivative w.r.t. $z_{l,t-1}$ yields

$$\frac{\partial^3 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; \mathbf{u})}{\partial \hat{z}_{it} \partial \hat{z}_{jt} \partial z_{l,t-1}} = \sum_{k=1}^{n} \Big(\frac{\partial^3 \eta_{kt}(\mathbf{u})}{\partial z_{kt}^2 \partial z_{l,t-1}} \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial^2 \eta_{kt}(\mathbf{u})}{\partial z_{kt} \partial z_{l,t-1}} \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big) \equiv 0. \tag{29}$$

Similarly, when using varied values for $\mathbf{u}$ in Eq 28, computing the difference between these instances yields

$$\begin{aligned} &\frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; u_{r+1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} - \frac{\partial^2 \log p(\hat{\mathbf{z}}_t \mid \hat{\mathbf{z}}_{t-1}; u_{r+1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} \\ =& \sum_{k=1}^{n} \Big[\Big(\frac{\partial^2 \eta_{kt}(u_{r+1})}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(u_r)}{\partial z_{kt}^2} \Big) \cdot \mathbf{H}_{kit} \mathbf{H}_{kjt} + \Big(\frac{\partial \eta_{kt}(u_{r+1})}{\partial z_{kt}} - \frac{\partial \eta_{kt}(u_r)}{\partial z_{kt}} \Big) \cdot \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \Big] \equiv 0. \end{aligned} \tag{30}$$

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 $i \neq j$. Building on the insights from the proof of Theorem A1, $\hat{\mathbf{z}}_t$ is compelled to be a permutation of a component-wise invertible transformation of $\mathbf{z}_t$. □

setting	$\tau = 1, r = 2$	$\tau = 2, r = 1$
CaRiNG	0.9436	0.9131
CaRiNG (lagged decoder)	0.9250	0.9220
TDRL	0.8947	0.7519

Table A1: Ablation study on different settings for **UG-TDMP**. (a) The second column is a more difficult scenario compared to the first, where the performance of **CaRiNG** remains good while that of baseline decreases significantly. (b) Omit the time-lagged latent variables in the decoder will not damage the performance much, but one can enjoy the benefits from a much simpler model.

A2 Experiment Settings

A2.1 Reproducibility

All experiments are done in a GPU workstation with CPU: Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz, GPU: Tesla V100. The source code and the generated data for the simulation experiments are attached in the supplementary materials.

A2.2 Synthetic Dataset Generation

In this section, we give 2 representative simulation settings for **NG** and **NG-TDMP** respectively to reveal the identifiability results. For each synthetic dataset, we set latent space to be 3, i.e., $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^3$.

Non-invertible Generation For **NG**, we set the transition lag as $\tau = 1$. We first generate $10,000$ data points from uniform distribution as the initial state $\mathbf{z}_0 \sim U(0, 1)$. For $t = 1, \cdots, 9$, each latent variable $\mathbf{z}_t$ will be generated from the proceeding latent variable $\mathbf{z}_{t-1}$ through a nonlinear function $\mathbf{f}$ with a non-additive zero-biased Gaussian noise ϵ_t ($\sigma = 0.1$), i.e., $\mathbf{z}_t = \mathbf{f}(\mathbf{z}_t, \epsilon_t)$. To introduce the non-invertibility, the mixing function $\mathbf{g}$ leverages only the first two entries of the latent variables to generate the 2-d observation $\mathbf{z}_t = \mathbf{g}(x_{1,t}, x_{2,t}) \in \mathcal{Z} \subseteq \mathbb{R}^2$.

Time-Delayed Mixing Process For **UG-TDMP**, we set the transition lag as $\tau = 1$ and mixing lag $r = 2$. Similar to the Non-invertible Generation scenario, we generate the initial states from uniform distribution and the subsequent latent variables following a nonlinear transition function. The noise is also introduced in a nonlinear Gaussian ($\sigma = 0.1$) way. The mixing process is a nonlinear function with regard to $\mathbf{z}_t$ plus a side information from previous steps $\mathbf{z}_{t-1:t-2}$, i.e.,

$$\mathbf{x}_t = A_{3\times3} \cdot \sigma\big(B_{3\times3} \cdot \sigma(C_{3\times3} \cdot \mathbf{z}_t)\big) + \begin{bmatrix} 0 \\ 0 \\ D_{3\times1}\mathbf{z}_{t-1} + E_{3\times1}\mathbf{z}_{t-2} \end{bmatrix}, \tag{31}$$

where σ refers to the ReLU function and the capital characters refer to matrices. Note that we make two modifications to show the advantage of **CaRiNG** . The reason we consider larger mixing lag is that it is a much more difficult scenario to handle, with more distribution from the mixing process and less dynamic information from transition. We run experiments in both scenarios with different transition and mixing lag. Besides, we also find out that even without time-lagged latent variables in the decoder, it leads to a smaller model that is more stable and easy to train. Refer to Table A1 for a detailed ablation study.

Post-processing Precedure During the generating process, we did not explicitly enforce the data to meet the constraint $\mathbf{z}_t = \mathbf{m}(\mathbf{x}_{t:t-\mu})$. On the contrary, we implement a checker to filter the data that is qualified. To be more precise, we do linear regression from $\mathbf{x}_{t:t-\mu}$ to $\mathbf{z}_t$ to figure out how much information of latent variables can be recovered from observation series in the best case. We choose the smallest μ when the amount of information that can be recovered is acceptable. We set $\mu = 2$ for **UG** and $\mu = 4$ for **UG-TDMP**.

A2.3 IMPLEMENTATION DETAILS

A2.3.1 SYNTHETIC DATA

Network Architecture To implement the Sequence-to-Step encoder, we leverage the *torch.unfold* to generate the nesting observations. Let us denote $\mathbf{x}_t^{(\mu)} = [\mathbf{x}_t, \cdots, \mathbf{x}_{t-\mu}]$ as inputs. For the time steps that do not exist, we simply pad them with zero. Refer to Table A2 for detailed network architecture.

Training Details The models were implemented in PyTorch 1.11.0. An AdamW optimizer is used for training this network. We set the learning rate as 0.001 and the mini-batch size as 64. We train each model under four random seeds $(770, 771, 772, 773)$ and report the overall performance with mean and standard deviation across different random seeds.

Table A2: Architecture details. BS: batch size, T: length of time series, i_dim: input dimension, o_dim: output dimension, z_dim: latent dimension, LeakyReLU: Leaky Rectified Linear Unit.

Configuration	**Description**	**Output**
1. Sequence-to-Step Encoder	Encoder for Synthetic Data	
Input: $\mathbf{x}_{1:T}^{(\mu)}$	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. Step-to-Step 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 $\times$ o_dim
3. 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	$\hat{\mathbf{z}}_{1:T}$
4. 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

A2.3.2 REAL-WORLD DATASET

Network Architecture We choose HCRN (Le et al., 2020) (without classification head) as the encoder backbone of **CaRiNG** on the real-world dataset: SUTD-TrafficQA. Given that HCRN is an encoder that calculates the cross attention between visual input and text input sequentially, we apply a decoder, which shares the same structure as the Step-to-Step Decoder shown in Table A2 to reconstruct the visual feature embedded with the temporal information. As it goes to transition prior, we use the Modular Prior shown in Table A2. This encoder-decoder structure can guide the model to learn the hidden representation with identifiable guarantees under the non-invertible generation process.

A3 MORE VISUALIZATION RESULTS ON REAL-WORLD DATA

As shown Table A1, we provide some positive examples and also fail cases to analyze our model. From the top two examples, we can find that our method can solve the occlusions well. From the

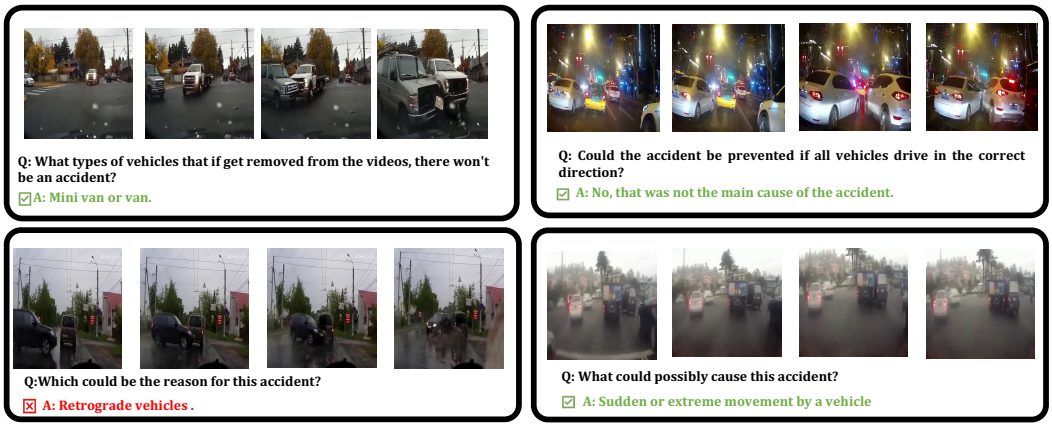

Figure A1: **Qualitative resutls on SUTD-TrafficQA dataset.** We provide some positive examples and also fail cases to analyze our model.

bottom right one, we find that our model can solve the blurred situation. However, when the alignment between visual and textual domains is difficult. The model may fail.

A4 Related Work

A4.1 Causal Discovery with Latent Variables

Some studies have aimed to discover causally related latent variables, such as Silva et al. (2006); Kummerfeld & Ramsey (2016); Huang et al. (2022) leverage the vanishing Tetrad conditions Spearman (1928) or rank constraints to identify latent variables in linear-Gaussian models, and Shimizu et al. (2009); Cai & Xie (2019); Xie et al. (2020; 2022) draw upon non-Gaussianity in their analysis for linear, non-Gaussian scenarios. Furthermore, some methods aim to find the structure beyond the latent variables, resulting in the hierarchical structure. Some hierarchical model-based approaches assume tree-like configurations, such as Pearl (1988); Zhang (2004); Choi et al. (2011); Drton et al. (2017), while the other methods assume a broader hierarchical structure Xie et al. (2022); Huang et al. (2022). However, these methods remain confined to linear frameworks and face escalating challenges with intricate datasets, such as videos.

A4.2 Nonlinear ICA for Time Series Data

Nonlinear ICA represents an alternative methodology to identify latent causal variables within time series data. Such methods leverage auxiliary data—like class labels and domain indices—and impose independence constraints to facilitate the identifiability of latent variables. To illustrate: Time-contrastive learning (TCL (Hyvarinen & Morioka, 2016)) adopts the independent sources premise and capitalizes on the variability in variance across different data segments. Furthermore, Permutation-based contrastive (PCL (Hyvarinen & Morioka, 2017)) puts forth a learning paradigm that distinguishes genuine independent sources from their permuted counterparts. Furthermore, i-VAE (Khemakhem et al., 2020) utilizes deep neural networks, VAEs, to closely approximate the joint distribution encompassing observed and auxiliary non-stationary regimes. Recent work, exemplified by LEAP (Yao et al., 2022b), has tackled both stationary and non-stationary scenarios in tandem. In the stationary context, LEAP postulates a linear non-Gaussian generative process. For the non-stationary context, it assumes a nonlinear generative process, gaining leverage from auxiliary variables. Advancing beyond LEAP, TDRL (Yao et al., 2022a) initially extends the linear non-Gaussian generative assumption to a nonlinear formulation for stationary scenarios. Subsequently, it broadens the non-stationary framework to accommodate structural shifts, global alterations, and combinations thereof. Additionally, CITRIS (Lippe et al., 2022b;a) champions the use of intervention target data to precisely identify scalar and multi-dimensional latent causal factors. However, a

common thread across these methodologies is the presumption of an invertible generative process, a stance that often deviates from the realities of actual data.

A5 Broader Impacts, Limitation, and Future Work

This study introduces both a theoretical framework and a practical approach for extracting causal representations from time-series data. Such advancements enable the development of more transparent and interpretative models, enhancing our grasp of causal dynamics in real-world settings. This approach may benefit many real-world applications, including healthcare, auto-driving, and finance, but it could also be used illegally. For example, within the financial sphere, it can be harnessed to decipher ever-evolving market trends, optimizing predictions and thereby influencing investment and risk management decisions. However, it's imperative to note that any misjudgment of causal relationships could lead to detrimental consequences in these domains. Thus, establishing causal links must be executed with precision to prevent skewed or biased inferences.

Theoretically, though allowing for the non-invertible generation process, our theoretical assumptions still fall short of fully capturing the intricacies of real-world scenarios. For example, identifiability requires the absence of instantaneous causal relations, i.e., relying solely on time-delayed influences within the latent causal dynamics. Furthermore, we operate under the presumption that the number of variables remains consistent across different time steps, signifying that no agents enter or exit the environment. Moving forward, we aim to broaden our framework to ensure identifiability in more general settings, embracing instantaneous causal dynamics and the flexibility for variables to either enter or exit.

In our experiments, we evaluate our approach with both simulated and real-world datasets. However, our simulation relies predominantly on data points, creating a gap from real-world data. Concurrently, the real datasets lack the presence of ground truth latent variables. In the future, we plan to develop a benchmark specifically tailored for the causal representation learning task. This benchmark will harness the capabilities of game engines and renderers to produce videos embedded with ground-truth latent variables.

References

Ruichu Cai and Feng Xie. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 2019.

Myung Jin Choi, Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

Mathias Drton, Shaowei Lin, Luca Weihs, and Piotr Zwiernik. Marginal likelihood and model selection for gaussian latent tree and forest models. 2017.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *arXiv preprint arXiv:2210.01798*, 2022.

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.

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

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*, pp. 2207–2217. PMLR, 2020.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1655–1664, 2016.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9972–9981, 2020.

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*, 2022a.

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*, pp. 13557–13603. PMLR, 2022b.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

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

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

Charles Spearman. Pearson's contribution to the theory of two factors. *British Journal of Psychology*, 19(1):95, 1928.

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.

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

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=Vi-sZWNA_Ue.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Nevin L Zhang. Hierarchical latent class models for cluster analysis. *The Journal of Machine Learning Research*, 5:697–723, 2004.