Latent Processes Identification From Multi-View Time Series

Zenan Huang[1,4] , **Haobo Wang**[4] , **Junbo Zhao**[4] and **Nenggan Zheng**[*1,2,3,4]
[1]Qiushi Academy for Advanced Studies (QAAS), Zhejiang University
[2]The State Key Lab of Brain-Machine Intelligence, Zhejiang University
[3]CCAI by MOE and Zhejiang Provincial Government (ZJU)
[4]College of Computer Science and Technology, Zhejiang University
{lccurious, wanghaobo, j.zhao, zng}@zju.edu.cn,

Abstract

Understanding the dynamics of time series data typically requires identifying the unique latent factors for data generation, *a.k.a.*, latent processes identification. Driven by the independent assumption, existing works have made great progress in handling single-view data. However, it is a non-trivial problem that extends them to multi-view time series data because of two main challenges: (i) the complex data structure, such as temporal dependency, can result in violation of the independent assumption; (ii) the factors from different views are generally overlapped and are hard to be aggregated to a complete set. In this work, we propose a novel framework MuLTI that employs the contrastive learning technique to invert the data generative process for enhanced identifiability. Additionally, MuLTI integrates a permutation mechanism that merges corresponding overlapped variables by the establishment of an optimal transport formula. Extensive experimental results on synthetic and real-world datasets demonstrate the superiority of our method in recovering identifiable latent variables on multi-view time series. The code is available on https://github.com/lccurious/MuLTI.

1 Introduction

Detecting causal relationships from time series based on observations is a challenging problem in many fields of science and engineering [Spirtes *et al.*, 2000]. A thorough grasp of causal relationships, interaction pathways, and time lags is valuable for interpreting and modeling temporal processes [Pearl, 2000]. Existing works [Chickering, 2002; Tsamardinos *et al.*, 2006; Zhang, 2008; Hoyer *et al.*, 2008; Zheng *et al.*, 2018] typically rely on predefined variables. However, such a strategy is not directly applicable to real-world scenarios, where data are intertwined with unknown generation processes, and causal variables are not readily available. Therefore, identifying the latent sources is crucial for interpreting underlying causal relations and elucidating the genuine dynamics inherent in the temporal data.

*Corresponding author.

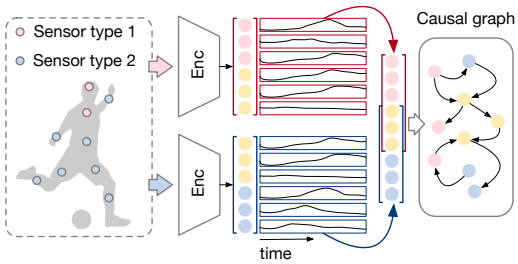

Figure 1: The diagram depicts a pipeline analyzing multi-view physiological time-series data. This pipeline learns temporal embeddings from both views, aligns variables considering their dependencies, and effectively reveals underlying variables and relationships.

To cope with this problem, non-linear independent component analysis (nICA) [Hyvärinen and Pajunen, 1999] has shown promising results in terms of identifiability, by effectively exploiting the underlying structure of the data. On time series data, most nICA techniques attempt to utilize the temporal structure among sequential observations, such as temporal contrastive learning [Hyvärinen and Morioka, 2016], permutation contrastive learning [Hyvärinen and Morioka, 2017], and generalized contrastive learning [Hyvärinen *et al.*, 2019]. However, these methods inherently assume the latent components are independent, which may not hold in practice. Moreover, in broader real-world scenarios, causal discovery tasks mostly require identifying dependent relations from extensive data collections, which may further restrict the identifiability of nICA techniques. Take Figure 1 as an example, understanding the physiological processes of human actions requires jointly analyzing the data across multiple kinds of sensors and a physiological signal may influence multiple downstream regions with different time lags. Similar cases also exist in a wide variety of real-world scenarios, such as multi-view sensors information fusion [He *et al.*, 2021], multi-market stock index analysis [Wang *et al.*, 2020], *etc*. Hitherto, few efforts have been made to address such challenging yet realistic problems.

The example above, which we call the multi-view latent process identification (MVLPI) problem, poses two combinatorial challenges. First, the diverse time delays in interactions among variables in time series data complicate direct estimation of latent variables. Second, different views typically

correlate with subsets of latent factors. Due to identifiability issues, merely inverting these views does not guarantee alignment of the recovered factors. Hence, aggregating them to the final complete latent variables is non-trivial.

In this paper, we propose the **Mu**lti-view **L**aten**T** Processes **I**dentification (dubbed **MuLTI**) framework that learns identifiable causal related variables from the multi-view data. MuLTI identifies latent variables from multi-view observations using three strategies: 1) It reformulates causally related process distributions using conditional dependence, replacing original latent variables with independent causal process noises; 2) It applies contrastive learning to maximize mutual information between prior and posterior conditional distributions; 3) It aligns partially overlapping latent variables using learnable permutation matrices, optimized with the Sinkhorn method. The first two strategies ensure the learning of identifiable latent variables, while the last one bridges the causal relations across multiple views. Given that only a subset of each view's source components correspond, matching and merging shared latent variables resemble the sorting and alignment of top-k shared components among estimated variables.

We evaluate our method on both synthetic and real-world data, spanning multivariable time series and visual tasks. Experiment results show that the latent variables are reliably identified from observational multi-view data. To the best of our knowledge, the learning of causal dependent latent variables from multi-view data has no prior solution. The proposed framework may also serve as a factor analyzing tool for extracting reliable features for downstream tasks of multi-view learning. More theoretical and empirical results can be found in Appendix[1].

2 Methodology

2.1 Problem Formulation

In contrast to the single-view setup, the MVLPI problem presents a unique challenge: only a portion of the latent factors can be recovered from an individual view. Thus, it is necessary to revisit the data-generating procedure of the MVLPI task. As shown in Figure 2, we consider the MVLPI task on the time series data, with the goal of recovering the latent factor $\boldsymbol{z}_t \in \mathbb{R}^d$ at each time step t, which uniquely generates the observed views. To achieve this, we make a common assumption that the current state is spatiotemporally dependent on the historical states,

$$\boldsymbol{z}_{i,t} = f_i(\mathbf{Pa}(z_{i,t}), \epsilon_{i,t}), \tag{1}$$

where $\mathbf{Pa}(z_{i,t})$ denotes causal parents of i-th factor at current step t, $\epsilon_{i,t}$ denotes a noise component of causal transition. While $\boldsymbol{z}_t$ contains a complete set of latent factors that we are truly interested in, we note that each observed view may be dependent on only a subset of it. Formally, we have the following data-generating process,

$$\boldsymbol{x}_t^1 = g^1(\boldsymbol{z}_t^1), \quad \boldsymbol{x}_t^2 = g^2(\boldsymbol{z}_t^2), \quad \text{where } \boldsymbol{z}_t = \boldsymbol{z}_t^1 \cup \boldsymbol{z}_t^2, \tag{2}$$

where $\boldsymbol{x}_t^v$ is the v-th observed view at time step t, and $g^v(\cdot)$ is the corresponding map function. Note that here we consider a 2-view setup for the sake of brevity, while our framework can be easily generalized to handle more views.

[1]https://arxiv.org/abs/2305.08164

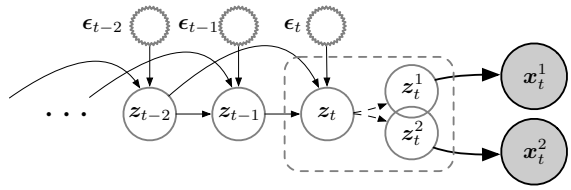

Figure 2: Graphical model of data generation. Each view may be generated by part of latent variables individually.

Following the above definition, we can identify the factors from the observed views, and aggregate them into the complete version $\boldsymbol{z}_t$. However, as each view performs a distinctive nonlinear mixing of the source factors, it is challenging to identify the corresponding latent variables entangled in views or to ensure the factors from different views are in the same subspace. Second, though we slightly abuse the notation '$\cup$' to show the *aggregation* relation of view-specific and the complete factors, it is non-trivial to reconstruct the complete $\boldsymbol{z}_t$ from a set of indeterminate estimations in practice. To remedy this problem, we introduce our novel **MuLTI** framework which comprises a contrastive learning module for enhanced identifiability alongside a novel merge operator that aggregates the view-specific latent variables to a complete one. The whole procedure is illustrated in Figure 3.

2.2 Contrastive Learning Module

Our goal is to learn an inverse function $r(\cdot)$ from the observed data. This is achieved by: (1) learning individual inverse functions $r^v : \mathcal{X}^v \mapsto \mathcal{Z}^v$; (2) introducing a merge function $m : \{\mathcal{Z}^v\} \mapsto \mathcal{Z}$, such that $r(\cdot) = m(\{r^v(\boldsymbol{x}^v)\})$. Note that each $r^v(\cdot)$ corresponds to an inverse of $g^v(\cdot)$. We will introduce the merge operator in Sec 2.4, here we omit the superscript v to represent the merged functions and source variables. Formally, the ultimate objective is to optimize $r(\cdot)$ to create a consistent mapping, $h = r \circ g$, between the estimated $\hat{\boldsymbol{z}}_t = h(\boldsymbol{z}_t)$ and the true $\boldsymbol{z}_t$, *e.g.*, an isometry transformation. We exploit the spatiotemporal dependencies of $\boldsymbol{z}_t$ to infer the values of latent variables from preceding states, as shown in Eq. (1). To achieve this, we introduce the causal transition function $f(\cdot)$ to approximate the inherient conditional distribution $p(\boldsymbol{z}_t|\boldsymbol{z}_{H_t})$, where $\boldsymbol{z}_{H_t}$ denotes the set of preceding states of $\boldsymbol{z}_t$. Empirically, we estimate an alternative value of the complete latent factor by $\tilde{\boldsymbol{z}}_t = f(\boldsymbol{z}_{H_t})$.

Motivated by the results of nICA [Hyvärinen *et al.*, 2019; Zimmermann *et al.*, 2021], we can identify the underlying factors of data with a certain low indeterminacy by carefully exploiting the inherent structure of data via a variety of conditional distributions. To this end, we can achieve model identifiability – that is, the ability to exactly recover the variables $\boldsymbol{z}_t$ and their associated causal relations – by learning the function r, in such a way as to ensure h is an isometry. Specifically, $h : \mathcal{Z} \mapsto \mathcal{Z}$ must statisfy the condition $\delta(\boldsymbol{z}_t, \tilde{\boldsymbol{z}}_t) = \delta(h(\boldsymbol{z}_t), h(\tilde{\boldsymbol{z}}_t))$ everywhere, where $\delta(\cdot,\cdot)$ is a metric. Using the conditional dependency in sequential latent states $\{\boldsymbol{z}_t\}$, a key insight is to estimate the function f for causal relations modeling and estimate reverse function r:

$$\max_{r,f} \mathcal{I}(\tilde{\boldsymbol{z}}_t; \boldsymbol{z}_t) = \mathcal{I}(f(\boldsymbol{z}_{H_t}); \boldsymbol{z}_t), \tag{3}$$

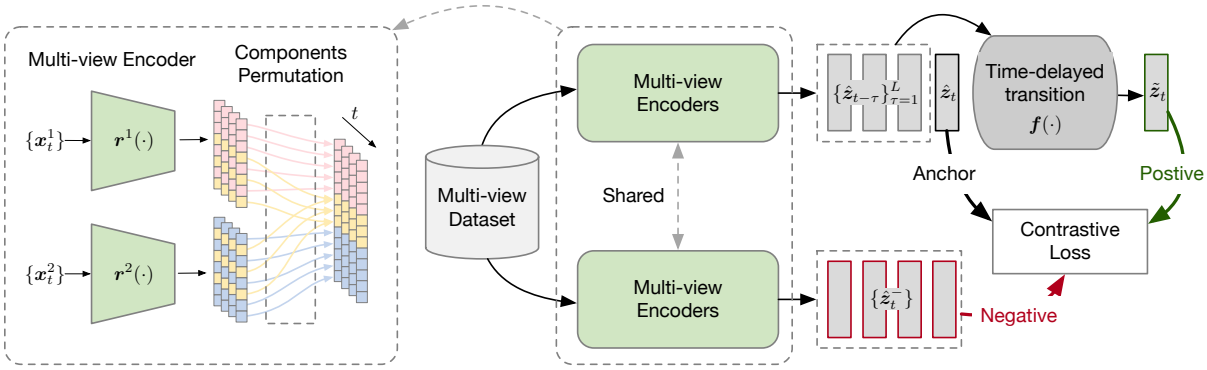

Figure 3: Illustration of our MuLTI framework. On one hand, we recover the view-specific latent factors $\hat{\boldsymbol{z}}_t^v$ from individual views, which are then merged to obtain $\hat{\boldsymbol{z}}_t$. On the other hand, we exploit the temporal dependency to obtain a causal transited latent factor $\tilde{\boldsymbol{z}}_t$ from previously estimated $\hat{\boldsymbol{z}}_t$. Thereafter, we regard them as positive pairs to optimize the contrastive loss, which serves as a surrogate of mutual information maximization to achieve identifiability.

where $\tilde{\boldsymbol{z}}_t = f(\boldsymbol{z}_{H_t})$ is the estimated latent variable based on the previous latent variables $\boldsymbol{z}_{H_t}$, $\mathcal{I}(\cdot;\cdot)$ indicates the mutual information. We denote the prior distribution as $p(\tilde{\boldsymbol{z}}_t|\boldsymbol{z}_{H_t})$ and the modeled distribution as $q_{h,f}(\tilde{\boldsymbol{z}}_t|\boldsymbol{z}_{H_t})$. The optimization problem defined in Eq. (3) can also be viewed as a minimization problem for the cross-entropy $H(\cdot\|\cdot)$,

$$\min_{f,r} H\left(p(\tilde{\boldsymbol{z}}_t|\boldsymbol{z}_{H_t})\|q_{f,r}(\tilde{\boldsymbol{z}}_t|\boldsymbol{z}_{H_t})\right). \tag{4}$$

Nevertheless, directly optimizing the Eq. (4) is typically difficult. To this end, we employ contrastive learning as a surrogate, given that it has been proven to be a variational bound of mutual information maximization [van den Oord *et al.*, 2018]. Formally, the contrastive loss is defined as follows,

$$\mathcal{L}_{\text{contr}}(r, f; \mu, M) := \mathop{\mathbb{E}}_{\substack{(\boldsymbol{x},\tilde{\boldsymbol{x}})\sim p_{\text{pos}} \\ \{\boldsymbol{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{-\delta(\boldsymbol{z},\tilde{\boldsymbol{z}})/\mu}}{e^{-\delta(\boldsymbol{z},\tilde{\boldsymbol{z}})/\mu} + \sum_{i=1}^M e^{-\delta(\boldsymbol{z},\boldsymbol{z}^-)/\mu}}\right], \tag{5}$$

$M \in \mathbb{Z}_+$ represents the fixed number of negative samples, p_{data} denotes the distribution of all observations, and p_{pos} signifies the distribution of positive pairs, we choose the ℓ_1 as the metric, represented by $\delta(\cdot,\cdot)$, $\mu \geq 0$ represents the temperature. Here, we omit the subscript t for simplicity.

As mentioned earlier, we can infer the latent variable $\boldsymbol{z}_t$ in two ways: directly, with $\hat{\boldsymbol{z}}_t = r(\{\boldsymbol{x}_t^v\})$, and indirectly, with $\tilde{\boldsymbol{z}}_t = f(\boldsymbol{z}_{H_t})$. Given the expectation that these two inferred variables should be similar, we define the pair $(\hat{\boldsymbol{z}}_t, \tilde{\boldsymbol{z}}_t)$ as a positive pair. Subsequently, we sample data from the marginal distribution of observations, estimate latent variables $\boldsymbol{z}_t^-$, and form negative pairs with $\hat{\boldsymbol{z}}_t$. At this point, the contrastive learning minimizes the cross-entropy between the ground-truth latent conditional distribution $p(\boldsymbol{z}_t|\boldsymbol{z}_{H_t})$ and a specific model distribution $q_{r,h}(\boldsymbol{z}_t|\boldsymbol{z}_{H_t})$. Thus, we have $\delta(h(\boldsymbol{z}_t), h(f(\boldsymbol{z}_{H_t}))) \propto \delta(\boldsymbol{z}_t, f(\boldsymbol{z}_{H_t}))$ for all $\boldsymbol{z}_t$ and $f(\boldsymbol{z}_{H_t})$. This implies that the ground-truth $\boldsymbol{z}_t$ lies in the isometric space of the estimated $\hat{\boldsymbol{z}}_t$, and can be further recovered through simple transformations.

2.3 Parametric Causal Transition

To represent the causal transition process of $f(\cdot)$, we consider a widely-used parametric formulation that aligns well with the Granger causality [Ding *et al.*, 2006]. To parameterize the dependence of $z_{i,t}$ on its causal parents $\mathbf{Pa}(z_{i,t})$ in Eq. (1), we model the causal transition as following vector autoregressive (VAR) process. Let $\boldsymbol{A}_\tau \in \mathbb{R}^{d\times d}$ be the full rank state transition matrix at lag τ. Assuming the true $\boldsymbol{z}_t$ is known, the causal transition can be represented as follow,

$$\boldsymbol{z}_t = f(\boldsymbol{z}_{H_t}, \boldsymbol{\epsilon}_t) = \sum\nolimits_{\tau=1}^{L} \boldsymbol{A}_\tau \boldsymbol{z}_{t-\tau} + \boldsymbol{\epsilon}_t, \tag{6}$$

here, L is max time lag, $\boldsymbol{\epsilon}_t$ is an additive noise variable. Based on this formulation, we can reformulate the conditional distribution $p(\boldsymbol{z}_t|\boldsymbol{z}_{H_t})$ by changing the variables:

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{H_t}) = p(\boldsymbol{z}_t - \sum\nolimits_{\tau=1}^{L} \boldsymbol{A}_\tau \boldsymbol{z}_{t-\tau}) = p(\boldsymbol{\epsilon}_t). \tag{7}$$

As the ground-truth $\boldsymbol{z}_t$ is not readily available, we reuse the estimated $\hat{\boldsymbol{z}}_t$ to calculate Eq. (6). Through this reformulation, we can represent the causal transition using a mutually independent noise distribution.

Through the above proper definitions of the contrastive pairs and the causal transition f, minimizers of contrastive loss determine the h up to an isometry:

Theorem 1 (Minimizers of contrastive objective recover the latent variables and their relations)**.** *Let latent space $\mathcal{Z}$ be a convex body in $\mathbb{R}^d$, $h = r \circ g : \mathcal{Z} \mapsto \mathcal{Z}$, f be a causal transition function, and δ be a metric, induced by a norm. If g is differentiable and injective, r, f are expressive enough to be minimizers of contrastive objective $\mathcal{L}_{\text{contr}}$ in Eq. (5) for $M \to +\infty$, then, we have $h = r \circ g$ is invertible and affine mapping and f captures the ground-truth causal relations.*

Theorem 1 guarantees the identifiability of the encoder r and causal transition function f (please refer to Appendix C for the proof). Specifically, the contrastive loss enforces $\delta(h(\tilde{\boldsymbol{z}}_t), h(\boldsymbol{z}_t)) \propto \delta(\tilde{\boldsymbol{z}}_t, \boldsymbol{z}_t)$ almost everywhere, which leads h to be an isometry, *i.e.*, there exists an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ such that $\hat{\boldsymbol{z}}_t = \boldsymbol{U}\boldsymbol{z}_t$ and $\hat{\boldsymbol{A}}_\tau = \boldsymbol{A}_\tau \boldsymbol{U}^\top$. Further, when $\boldsymbol{\epsilon}_t$ is non-Gaussian (*e.g.*, Laplacian), and metric

$\delta(\cdot,\cdot)$ is non-isotropic (*e.g.*, ℓ_1 norm), then there is only a channel permutation π between estimations and ground-truth such that $\hat{z}_{i,t} = s_i z_{\pi(i),t}$, where s_i is a scaling constant.

2.4 Merge Operation

As mentioned earlier, a significant challenge in learning a spatiotemporal dependent source from multi-view data lies in aggregating the view-specific source, $\boldsymbol{z}_t^v$, into the complete latent source, $\boldsymbol{z}_t$. This is because each view only holds a portion of the total information. Given that sources estimated from each view exhibit unique uncertainties, ensuring their alignment within the same subspace to yield the desired complete source presents a non-trivial task. Yet, given that a portion of the information is shared among different sources, it is feasible to establish transformation between them by identifying a common source. We subsequently cast the problem of identifying a common source as one of the optimal transport objectives, which involves searching plans to move factors from view-specific sources to the common source.

Permutation learning. Assume there is a common source $\boldsymbol{c}_t \in \mathbb{R}^{d_c}$ shared across all view-specific sources $\boldsymbol{z}_t^v$. While directly searching for the correspondence of common components can be infeasible, we propose to enforce the $\boldsymbol{c}_t^v$ be in the top-d_c channels of every $\boldsymbol{z}_t^v$ via permutation learning. Formally, we can achieve this by multiplying the $\boldsymbol{z}_t^v$ with a doubly stochastic matrix $\boldsymbol{B}^v \in \mathfrak{B}_{d_v}$ as follows,

$$\bar{\boldsymbol{z}}_t^v = \boldsymbol{B}^v \hat{\boldsymbol{z}}_t^v, \quad \boldsymbol{c}_t^v = \bar{\boldsymbol{z}}_{1:d_c,t}^v. \tag{8}$$

After that, we can impose the permutations that generate embeddings whose top d_c entries are most correlated:

$$\max_{\{\boldsymbol{B}^v\}} \mathbb{E}_{v' \neq v} \mathrm{Tr}\left((\boldsymbol{B}^v \hat{\boldsymbol{z}}_t^v)_{1:d_c} (\boldsymbol{B}^{v'} \hat{\boldsymbol{z}}_t^{v'})_{1:d_c}^\top \right). \tag{9}$$

In essence, our goal is to ensure that all extracted instances of $\{\boldsymbol{c}_t^v\}$ have entries closely resembling those of the identified common components. Notably, the procedure outlined above equates to regularizing these $\{\boldsymbol{c}_t^v\}$ so that they are concentrated around their mean. To achieve this, we first estimate a mean center from the currently extracted common sources:

$$\boldsymbol{c}_t^* = \arg\min_{\boldsymbol{c}_t \in \mathbb{R}^{d_c}} \sum_v \|\boldsymbol{c}_t^v - \boldsymbol{c}_t\|^2. \tag{10}$$

In practice, we achieve the objective of Eq. (9) by updating the transport plans $\{\boldsymbol{B}^v\}$ and minimizing transport cost from top-d_c view specific components to $\boldsymbol{c}_t^*$ in an alternating manner:

$$\mathcal{L}_m = \sum_{v=1}^{2} \mathbb{E}\left[\|(\boldsymbol{B}^v \boldsymbol{z}_t^v)_{1:d_c} - \boldsymbol{c}_t^*\|^2 \right] + \eta R(\boldsymbol{B}^v), \tag{11}$$

where $R(\boldsymbol{B}^v) = -\sum_{i,j} B_{i,j}^v \log(B_{i,j}^v)$, $\eta > 0$ is a constant. This criterion effectively results in learning that $\boldsymbol{c}_t^1 = \boldsymbol{c}_t^2$, which represents the ultimate solution for the merging of latent variables.

After this reformulation, with the center $\boldsymbol{c}_t^*$ fixed, we can search for the locally optimal $\{\boldsymbol{B}_v\}$ separately. Crucially, each sub-problem now becomes a standard combinatorial assignment problem [Peyré and Cuturi, 2019], which involves the search for optimal transport plans that assign common components from view-specific sources to the common source $\boldsymbol{c}_t^*$. Ultimately, by iteratively applying the center mean estimation procedure from Eq. (10) and utilizing the Sinkhorn algorithm, we can achieve the minimizer of our final objective, as expressed in Eq. (11) (see Appendix B for additional details). We can consider the application of the transformation $\boldsymbol{B}^v$ as a process of smooth component sorting. When we can identify estimated latent sources up to a permutation transformation, we can reduce the doubly stochastic matrices $\boldsymbol{B}^v$ to permutation matrices, *i.e.*, $\{\boldsymbol{P}^v | \boldsymbol{P}^v \in \{0,1\}^{d_v \times d_v}\}$. This resolution addresses the issue of non-corresponding components.

For merge operator m, we concanate the $\boldsymbol{c}_t^*$ with all remaining private components:

$$\hat{\boldsymbol{z}}_t = \begin{bmatrix} \boldsymbol{c}_t^* \\ \bar{\boldsymbol{z}}_{d_c+1:d_1,t}^1 \\ \bar{\boldsymbol{z}}_{d_c+1:d_2,t}^2 \end{bmatrix} = m\left(\begin{bmatrix} \hat{\boldsymbol{z}}_t^1 \\ \hat{\boldsymbol{z}}_t^2 \end{bmatrix} \right). \tag{12}$$

Consequently, we can input the merged latent source into the contrastive learning and the inference of causal relations, as outlined in Sec 2.2, thereby completing the entire objective, which is stated in the following theorem (please refer to Appendix C.3 for the proof):

Theorem 2 (Minimizers of the multi-view objective maintains the channel corespondency)**.** *Let $\mathcal{Z}^v \subseteq \mathbb{R}^{d_v}$, $\mathcal{Z} = \cup \mathcal{Z}^v$ and $\cap \mathcal{Z}^v \neq \emptyset$. If $\boldsymbol{B}^v \in \mathfrak{B}_{d_v}$ be a doubly stochastic matrix, $r, f, \{\boldsymbol{B}^v\}$ are the minimizers of contrastive objective for $M \to +\infty$, then, we have $h = r \circ g$ is an isometry, and $\{\boldsymbol{B}^v\}$ can rearrange the components of each view source such that common components from each view source are aligned.*

3 Optimization

We use deep neural networks to implement $f(\cdot)$ and $r^v(\cdot)$. The entire optimization process primarily involves two steps: implementing contrastive learning on the estimated sources and learning permutations for the view-specific sources. We jointly train view-specific inverse function networks as well as a causal transition network. Moreover, the permutation matrices $\{\boldsymbol{B}^v\}$ are optimized alternatively with the main networks during this process.

Enhance noises distribution learning. Contrastive learning is designed to minimize the distance between positive pairs, in other words, it brings $\hat{\boldsymbol{z}}_t$ and $f(\hat{\boldsymbol{z}}_{H_t})$ closer together. To emphasize this property, we further employ an objective to minimize the residual as:

$$\mathcal{L}_\epsilon = \mathbb{E}[\delta(\hat{\boldsymbol{z}}_t, f(\hat{\boldsymbol{z}}_{H_t}))]. \tag{13}$$

In parallel, to enhance the mutual independence property of noise $\epsilon_{i,t}$, we follow the approach of [Yao *et al.*, 2021] and create a discriminator network $\mathcal{D}(\cdot)$ implemented by MLPs. This network is used for discriminating the $\{\hat{\epsilon}_{i,t}\}$ and randomly permutated versions $\{\hat{\epsilon}_{i,t}\}^{\text{perm}}$,

$$\mathcal{L}_\mathcal{D} = \mathbb{E}[\log \mathcal{D}(\{\hat{\epsilon}_{i,t}\}) + \log(1 - \mathcal{D}(\{\hat{\epsilon}_{i,t}\}^{\text{perm}}))]. \tag{14}$$

Combining all these elements with weights, the objective is:

$$\mathcal{L} = \mathcal{L}_{\text{contr}} + \beta_1 \mathcal{L}_m + \beta_2 \mathcal{L}_\epsilon + \beta_3 \mathcal{L}_\mathcal{D}. \tag{15}$$

For stable training, we perform the optimization of $\mathcal{L}_{\text{contr}}$ and $\mathcal{L}_m$ alternately.

4 Experiments

In this section, we present our experimental results on three multi-view scenarios to validate the superiority of MuLTI. More empirical results can be found in Appendix D.

4.1 Datasets

We use synthetic and real-world datasets for evaluating the latent process identification task. Different view-specific encoders are used in different datasets for extracting latent variables. In what follows, we briefly introduce the datasets.

Multi-view VAR is a synthetic dataset modified from [Yao *et al.*, 2021]. To generate the latent process, we first sample transition matrices $\{\boldsymbol{A}_\tau\}_{\tau=1}^{L}$ from a uniform distribution $U[-0.5, 0.5]$, and then sample initial states $\{\boldsymbol{z}_t\}_{t=1}^{L}$ from a normal distribution to generate sequences $\{\boldsymbol{z}_t\}_{t=1}^{T}$. To generate multi-view time series, we randomly select d_1 and d_2 dimensions of latent variables for each view. The nonlinear observations of each view are created by mixing latent variables with randomly parameterized MLPs $g^v(\cdot)$ following the previous work [Hyvärinen and Morioka, 2017].

Mass-spring system is a video dataset adopted from [Li *et al.*, 2020] specifically designed for a multi-view setting. A mass-spring system consists of 8 movable balls; the balls are either connected by springs or are not connected at all. As a result of random external forces, the system exhibits different dynamic characteristics depending on its current state and internal constraints. We create two views for this video dataset, as shown in Figure 4, in each of which only 5 of the 8 balls are observable. Consequently, each pair of views includes two videos, each with a duration of 80 frames. This dataset contains 5,000 pairs of video clips in total.

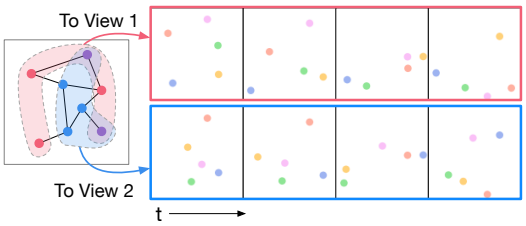

Figure 4: Illustration of multi-view Mass-spring system.

Multi-view UCI Daily and Sports Activities [Altun *et al.*, 2010] is a multivariate time series dataset comprising 9,120 sequences, capturing sensor data for 19 different human actions performed by 8 subjects. Each sample activity contains 125 time steps, each captured by nine sensors, yielding a total of 45 dimensions. We construct the multi-view setting following the [Li *et al.*, 2016]. The 27 dimensions located on the torso, right arm, and left arm are collectively regarded as view 1, while the remaining 18 dimensions on the left and right legs constitute view 2.

4.2 Evaluation Setup

To measure the identifiability of latent causal variables, we compute Mean Correlation Coefficient (MCC) on the validation datasets for revealing the indeterminacy up to permutation transformations, and R^2 for revealing the indeterminacy up to linear transformations. We assess the accuracy of our causal relations estimations by comparing them with the actual data structure, quantified via the Structural Hamming Distance (SHD) on the validation datasets.

Baselines. We compare our method with following baselines: (1) BetaVAE [Higgins *et al.*, 2017] neither accounts for the temporal structure nor provides an identifiability guarantee; (2) SlowVAE [Klindt *et al.*, 2021], PCL [Hyvärinen and Morioka, 2017], and GCL [Hyvärinen *et al.*, 2019] identify independent sources from observations; (3) ShICA [Richard *et al.*, 2021] identifies the shared sources from multi-view observations; (4) CL-ICA [Zimmermann *et al.*, 2021] recovers the latent variables via contrastive learning; (5) LEAP [Yao *et al.*, 2021] utilizes temporally engaged sources and causal mechanism but only suitable single view data.

For multi-view time series adaptation of baseline methods – β-VAE, SlowVAE, PCL, CL-ICA, and LEAP – we concatenate the series into a single-view format, aligning them by the feature dimension. In the GCL setting, we designate as positive pairs the views originating from the same source, and negative pairs those stemming from distinct sources.

Implementation details. For the Multi-view VAR dataset, we use architectures composed of MLPs and Leaky ReLU units. Ground-truth latent variable dimension d is set to 10, the noise distribution is set to Laplacian$(0, 0.05)$. We set the batch size to 2400, employ the Adam optimizer with a learning rate of 0.001, and utilize $\beta_1 = 0.01, \beta_2 = 0.01, \beta_3 = 1e-5$. To verify the influence of the overlapping ratio of views, we select $d_c \in \{10, 4, 2, 0\}$ for evaluation. Here, $d_c = 10$ and $d_c = 0$ indicate situations where the latent variables for generating views are fully overlapped and have no overlaps, respectively. To ensure a fair comparison, all baseline methods utilize similar encoders, and the time-lag L of the causal transition module is set to equal the ground truth.

For the Mass-spring system, we create view-specific encoders following [Li *et al.*, 2020], which has similar architectures to the unsupervised keypoints discovery perception module. We set the time lag $L = 2$ for the causal transition module as the approximated process corresponds to a mass-spring system, which is a dynamic system of second order. The dimension of estimated latent variables is set to be the same as ground-truth, *i.e.*, 8 coordinates (x, y) account for $d = 16$. In practice, we first pre-train two pairs of keypoint encoder-decoder in each view. Then, the pre-trained encoders are taken as $r^v(\cdot)$ corresponding to each view.

For the Multi-view UCI dataset, we construct two separate encoders with MLPs for two views. The latent dimensions of view 1 and view 2 are set to $d_1 = 12$ and $d_2 = 9$ respectively, the shared latent dimension is set to $d_c = 3$, and the complete latent dimension accounts for $d = 18$. The max time lag of the causal transition module is set to $L = 1$. We initially train the MuLTI on the complete dataset, following which we extract the latent variables for downstream tasks.

4.3 Main Results

MuLTI achieves the best identifiability. In Table 1, we report both R^2 and MCC for revealing the identifiability of each method. The results highlight MuLTI's superior performance in terms of identifiability across most settings, particularly

Methods	R^2 (%)				MCC			
	$d_c = d$	$d_c = 4$	$d_c = 2$	$d_c = 0$	$d_c = d$	$d_c = 4$	$d_c = 2$	$d_c = 0$
β-VAE	40.03$\pm$5.31	41.43$\pm$4.22	40.67$\pm$3.44	38.36$\pm$2.71	29.37$\pm$8.20	38.31$\pm$6.31	31.92$\pm$5.61	41.88$\pm$4.62
SlowVAE	60.32$\pm$5.56	63.21$\pm$5.51	62.12$\pm$6.13	61.23$\pm$3.17	50.32$\pm$3.49	51.24$\pm$4.71	52.55$\pm$2.39	54.21$\pm$3.07
PCL	69.78$\pm$3.20	80.24$\pm$3.32	75.71$\pm$2.71	74.56$\pm$2.90	52.64$\pm$2.33	54.63$\pm$2.12	53.42$\pm$1.91	55.61$\pm$2.62
GCL	73.45$\pm$4.34	82.41$\pm$3.12	77.65$\pm$2.52	73.21$\pm$3.31	53.55$\pm$1.71	57.32$\pm$2.41	52.42$\pm$1.70	55.23$\pm$2.31
ShICA	41.08$\pm$5.76	39.65$\pm$4.51	37.61$\pm$4.32	38.43$\pm$3.12	28.71$\pm$5.21	37.21$\pm$5.13	29.31$\pm$4.97	31.47$\pm$3.61
CL-ICA	21.48$\pm$6.71	29.37$\pm$4.32	24.27$\pm$3.33	23.89$\pm$2.71	26.79$\pm$5.51	33.35$\pm$4.51	29.39$\pm$5.11	23.68$\pm$4.90
LEAP	**99.59**$\pm$0.05	99.67$\pm$0.09	99.48$\pm$0.07	**99.74**$\pm$0.06	57.56$\pm$2.44	62.48$\pm$2.12	65.44$\pm$1.61	67.75$\pm$1.32
MuLTI	99.46$\pm$0.09	**99.72**$\pm$0.05	**99.68**$\pm$0.07	99.45$\pm$0.08	**99.80**$\pm$0.12	**99.27**$\pm$0.43	**99.39**$\pm$0.09	**99.43**$\pm$0.04

Table 1: Identifiability results on the VAR process ($d = 10, L = 2$). Mean+standard deviation over 5 random seeds.

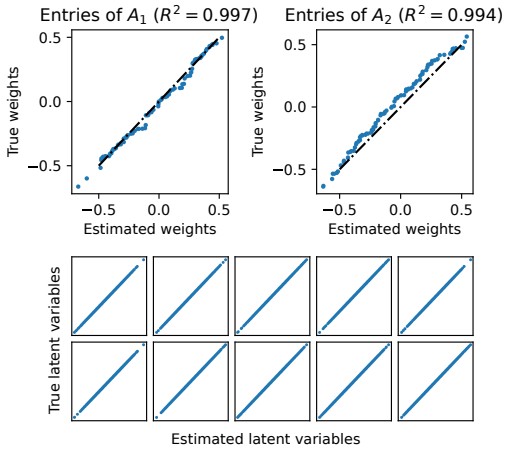

Figure 5: Results for latent VAR process ($d = 10, d_c = 2$) identification. Scatters are paired estimations and ground-truth.

in the MCC metrics, where it significantly outperforms all other methods. The results indicate that methods like β-VAE, SlowVAE, ShICA, CL-ICA, PCL, and GCL, which primarily depend on the independence of sources and do not model the causal transition, fail to recover the latent variables, even up to linear transformations. LEAP successfully identifies the latent process up to linear transformations using causal transition constraints. Yet, it encounters difficulties when trying to recover the process up to a permutation transformation, especially when observations come from distinct views, despite the conditions being conducive to permutation identifiability.

MuLTI successfully identifies the latent process. We show that MuLTI successfully identifies the latent process on both the Multi-view VAR dataset and the Mass-spring system. Figure 5 demonstrates the performance of our MuLTI in estimating the latent process where $d = 10, d_c = 2$. The results suggest that our model nearly perfects the learning of the VAR transition matrices, with the identified latent variables closely resembling the true variables, thus implying a comprehensive identification of the latent process.

The objective of the Mass-spring system task is to estimate latent variables, representing the coordinates of each ball, and transition matrices that elucidate their connections. As displayed in the left panel of Figure 6, the estimated latent

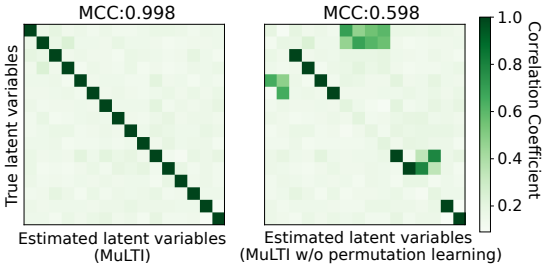

Figure 6: Correlation matrices of recovered latent variables vs. ground-truth latent variables of Mass-spring system.

variables precisely match the ground-truth, and the recovered transition matrices correspond to the ground-truth ball connections, as indicated by an SHD $= 0$.

MuLTI improves downstream task. Given that causal variables are often not readily available, directly assessing their identification and relationships in real-world data is challenging. However, under most conditions, identifying the latent process from observations can simplify downstream tasks (*e.g.*, classification and clustering) or directly aid in the extraction of meaningful features. Therefore, we employ a real-world multi-view time series dataset to demonstrate the effectiveness of latent process identification. We employ multivariate time series classifiers - specifically, reservoir computing [Bianchi *et al.*, 2021]. This approach encodes the multivariate time series data into a vectorial representation in an unsupervised manner, allowing us to evaluate learned latent variables $\{\hat{z}_t\}_{t=1}^T$ in the downstream task.

To demonstrate the efficacy of the learned latent representations, we independently train both the baseline methods and MuLTI on the entire dataset, subsequently extracting 18-dimensional features to form a new dataset. Then, we train and evaluate the RC classifier on the new datasets extracted by each method. The results are shown in Table 2. We report classification accuracy with varying training set sizes, *i.e.*, we randomly select $N_{tr} \in \{10, 20, 30, 40, 50\}$ samples per class for training, using the remaining samples for testing. Compared to the baseline methods, the pre-trained features from MuLTI enhance the downstream performance of the RC classifier, particularly for small training set classification tasks.

N_{tr}	raw data	+β-VAE	+LEAP	+ShICA	+MuLTI
10	74.21$\pm$0.26	72.53$\pm$1.13	73.57$\pm$1.90	73.17$\pm$0.77	**85.59**$\pm$0.05
20	90.00$\pm$0.38	85.82$\pm$0.53	86.10$\pm$0.86	86.59$\pm$0.51	**92.14**$\pm$0.38
30	92.92$\pm$0.74	89.37$\pm$1.54	88.90$\pm$0.69	91.29$\pm$0.31	**94.18**$\pm$0.14
40	94.95$\pm$0.16	91.16$\pm$0.49	91.79$\pm$0.07	94.01$\pm$0.55	**96.15**$\pm$0.81
50	94.92$\pm$0.28	92.53$\pm$0.04	93.00$\pm$0.03	94.78$\pm$0.04	**96.82**$\pm$0.30

Table 2: Classification accuracy on Multi-view UCI dataset. N_{tr} is the number of training samples randomly chosen from each class.

MuLTI	R^2 (%)	MCC
w/o Causal transition	44.67$\pm$2.31	33.47$\pm$7.60
w/o Permutation learning	99.32$\pm$0.07	77.45$\pm$0.47
w/o Residuals minimization	99.48$\pm$0.04	82.32$\pm$2.63
w/o Noise discriminator	99.23$\pm$0.05	98.32$\pm$0.27

Table 3: Ablation results on VAR dataset ($d = 10, d_c = 4, L = 2$).

4.4 Ablation Studies

Module ablations. To verify the effectiveness of each module within MuLTI, we conduct ablation studies by individually removing each module. The results tested on the Multi-view VAR dataset are shown in Table 3. When the causal transition module is removed, positive pairs are formed from temporally adjacent states, *i.e.*, $(\hat{\boldsymbol{z}}_t, \hat{\boldsymbol{z}}_{t+1})$. In the setting of removing permutation learning, the view-specific variables are simply concatenated along the feature dimension and mapped with a linear layer to match the dimension of the transition function. Results indicate the crucial role of the causal transition function in identifying the spatial-temporal dependent latent process. As shown, permutation learning also greatly contributes to the identifiability under this multi-view setting. Furthermore, Figure 6 from the Mass-spring system experiment reveals that, without permutation learning, only partial identification of the latent variables is possible.

Effect of noise and distance metrics. To evaluate the effect of noise distribution type and distance metric $\delta(\cdot, \cdot)$, we conduct experiments under different combinations. The results shown in Table 4 indicate that the Laplacian noise condition accompanied with ℓ_1 norm is key to identifying the ground-truth latent processes up to permutation equivalence. Nevertheless, in most conditions, we can still identify the latent process up to linear equivalence.

Noise type	$\delta(\cdot, \cdot)$	R^2 (%)	MCC
Identity	/	66.12$\pm$2.77	43.91$\pm$2.56
Supervised	/	99.78$\pm$0.07	99.92$\pm$0.02
Normal ($\sigma = 0.1$)	ℓ_2	99.72$\pm$0.03	67.21$\pm$1.41
Laplace ($\lambda = 0.1$)	ℓ_2	99.71$\pm$0.02	66.74$\pm$0.21
Normal ($\sigma = 0.1$)	ℓ_1	99.69$\pm$0.04	70.52$\pm$0.27
Laplace ($\lambda = 0.1$)	ℓ_1	99.74$\pm$0.02	99.31$\pm$0.07
Laplace ($\lambda = 0.05$)	ℓ_1	99.77$\pm$0.03	99.81$\pm$0.02

Table 4: Identifiability with different noise prior & metric. Note that the *Identity* indicates the untrained model, and *Supervised* indicates the model trained with MSE between ground-truth and estimated latent variables. Mean+standard deviation over 5 random seeds.

5 Related Works

Multi-view representation learning. Many realistic conditions involve data from multiple channels and domains; these data must be analyzed together in order to obtain a reliable estimation of latent relations, *e.g.* neurophysiology and behavior time series data [Urai *et al.*, 2022]. Typical methods of learning share representations from multi-view data include Canonical Correlation Analysis (CCA) [Hotelling, 1992] and its variants [Akaho, 2006; Andrew *et al.*, 2013; Lai and Fyfe, 2000; Wang *et al.*, 2015; Zhao *et al.*, 2017], but most of these methods only consider learning the common source of each view. More recently, some improvements are done in dividing the sources into common and private components [Lyu *et al.*, 2021; Huang *et al.*, 2018; Hwang *et al.*, 2020; Luo *et al.*, 2022]. However, it is not the case when there are relations (causal influences) between source components, due to most of them relying on independent source assumptions [Gresele *et al.*, 2019]. In contrast, our work seeks to recover latent variables when there are causal relations between them.

Identifiability. Identifiability means the estimated latent variables are exactly equivalent to the underlying ones or at least up to simple transformation. Traditional independent component analysis (ICA) [Belouchrani *et al.*, 1997] assumes sources are linear mixed, but such linearity relation may not hold in many applications. To overcome this limitation, the non-linear ICA (nICA) [Hyvärinen and Pajunen, 1999] has attracted great attention. A variety of practical methods [Hyvärinen and Morioka, 2017; Hyvärinen *et al.*, 2019; Khemakhem *et al.*, 2020; Sorrenson *et al.*, 2020] have been proposed to achieve identifiability through discriminating data structures, such as manually constructing positive and negative pairs and training models to produce features to distinguish whether input data are distorted, such as temporal structure [Hyvärinen and Morioka, 2016] and randomized distorted data pairs [Hyvärinen *et al.*, 2019]. More recently, Gresele *et al.* adopt similar practices into multi-view data, but still limited in source independent assumptions. Yao *et al.* make attempts on dependent sources but their method can only handle single-view data. In this work, we explore a general case that identifies dependent processes from multi-view temporal data.

6 Conclusion

In this work, we make the first attempt toward a general setting of the multi-view latent process identification problem and propose a novel MuLTI framework based on contrastive learning. The key idea is to identify the latent variables and dependent structure through the construction of conditional distribution and aggregating variables across views. Empirical results suggest that our methods can successfully identify the spatial-temporal dependent latent process from multi-view time series observations. We hope our work can raise more attention to the importance of identifying and aggregating latent factors in understanding multi-view data.

Acknowledgments

This work is supported by the National Key R&D Program of China (2020YFB1313501), National Natural Science Foundation of China (T2293723, 61972347), Zhejiang Provincial Natural Science Foundation (LR19F020005), the Key R&D program of Zhejiang Province (2021C03003, 2022C01119, 2022C01022), the Fundamental Research Funds for the Central Universities (No. 226-2022-00051). JZ also wants to thank the support by the NSFC Grants (No. 62206247) and the Fundamental Research Funds for the Central Universities.

References

[Akaho, 2006] Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

[Altun *et al.*, 2010] Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 2010.

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255. PMLR, 2013.

[Belouchrani *et al.*, 1997] Adel Belouchrani, Karim Abed-Meraim, J-F Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, 1997.

[Bianchi *et al.*, 2021] Filippo Maria Bianchi, Simone Scardapane, Sigurd Løkse, and Robert Jenssen. Reservoir Computing Approaches for Representation and Classification of Multivariate Time Series. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Chickering, 2002] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 2002.

[Ding *et al.*, 2006] Mingzhou Ding, Yonghong Chen, and Steven L Bressler. Granger causality: Basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*, 2006.

[Gresele *et al.*, 2019] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Proc. of UAI*, 2019.

[He *et al.*, 2021] Guoliang He, Han Wang, Shenxiang Liu, and Bo Zhang. CSMVC: A Multiview Method for Multivariate Time-Series Clustering. *IEEE Transactions on Cybernetics*, 2021.

[Higgins *et al.*, 2017] Irina Higgins, Loïc 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. In *Proc. of ICLR*, 2017.

[Hotelling, 1992] Harold Hotelling. Relations Between Two Sets of Variates. In *Breakthroughs in Statistics*. Springer New York, 1992.

[Hoyer *et al.*, 2008] Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Proc. of NeurIPS*, 2008.

[Huang *et al.*, 2018] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. of ECCV*, 2018.

[Hwang *et al.*, 2020] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. In *Proc. of NeurIPS*, 2020.

[Hyvärinen and Morioka, 2016] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Proc. of NeurIPS*, 2016.

[Hyvärinen and Morioka, 2017] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proc. of AISTATS*, 2017.

[Hyvärinen and Pajunen, 1999] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 1999.

[Hyvärinen *et al.*, 2019] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proc. of AISTATS*, 2019.

[Khemakhem *et al.*, 2020] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proc. of AISTATS*, 2020.

[Klindt *et al.*, 2021] David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *Proc. of ICLR*, 2021.

[Lai and Fyfe, 2000] P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 2000.

[Li *et al.*, 2016] Sheng Li, Yaliang Li, and Yun Fu. Multi-view time series classification: A discriminative bilinear projection approach. In *Proc. of CIKM*, 2016.

[Li *et al.*, 2020] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. In *Proc. of NeurIPS*, 2020.

[Luo *et al.*, 2022] Dixin Luo, Hongteng Xu, and Lawrence Carin. Differentiable Hierarchical Optimal Transport for Robust Multi-View Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Lyu *et al.*, 2021] Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *Proc. of ICLR*, 2021.

[Pearl, 2000] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2000.

[Peyré and Cuturi, 2019] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019.

[Richard *et al.*, 2021] Hugo Richard, Pierre Ablin, Bertrand Thirion, Alexandre Gramfort, and Aapo Hyvarinen. Shared Independent Component Analysis for Multi-Subject Neuroimaging. In *Proc. of NeurIPS*, 2021.

[Sorrenson *et al.*, 2020] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). In *Proc. of ICLR*, 2020.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

[Tsamardinos *et al.*, 2006] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 2006.

[Urai *et al.*, 2022] Anne E. Urai, Brent Doiron, Andrew M. Leifer, and Anne K. Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 2022.

[van den Oord *et al.*, 2018] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[Wang *et al.*, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. On deep multi-view representation learning. In *Proc. of ICML*, 2015.

[Wang *et al.*, 2020] Heyuan Wang, Tengjiao Wang, and Yi Li. Incorporating expert-based investment opinion signals in stock prediction: A deep learning framework. In *Proc. of AAAI*, 2020.

[Yao *et al.*, 2021] Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning Temporally Causal Latent Processes from General Temporal Data. In *Proc. of ICLR*, 2021.

[Zhang, 2008] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 2008.

[Zhao *et al.*, 2017] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Proc. of AAAI*, 2017.

[Zheng *et al.*, 2018] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with NO TEARS: continuous optimization for structure learning. In *Proc. of NeurIPS*, 2018.

[Zimmermann *et al.*, 2021] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proc. of ICML*, 2021.