Causal Discovery-Driven Change Point Detection in Time Series

Shanyun Gao Purdue University
Raghavendra Addanki Adobe Research
Tong Yu Adobe Research
Ryan A. Rossi Adobe Research
Murat Kocaoglu Purdue University

Abstract

Change point detection in time series aims to identify moments when the probability distribution of time series changes. It is widely applied in many areas, such as human activity sensing and medical science. In the context of multivariate time series, this typically involves examining the joint distribution of multiple variables: If the distribution of any one variable changes, the entire time series undergoes a distribution shift. However, in practical applications, we may be interested only in certain components of the time series, exploring abrupt changes in their distributions while accounting for the presence of other components. Here, assuming an underlying structural causal model that governs the time-series data generation, we address this task by proposing a two-stage non-parametric algorithm that first learns parts of the causal structure through constraint-based discovery methods, and then employs conditional relative Pearson divergence estimation to identify the change points. The conditional relative Pearson divergence quantifies the distribution difference between consecutive segments in the time series, while the causal discovery method allows a focus on the causal mechanism, facilitating access to independent and identically distributed (IID) samples. Theoretically, the typical assumption of samples being IID in conventional change point detection methods can be relaxed based on the Causal Markov Condition. Through experiments on both synthetic and real-world datasets, we validate the correctness and utility of our approach.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

1 INTRODUCTION

Change point analysis aims to detect distribution shifts in observational time series data. This topic has been explored extensively and is popular in many areas, such as human activity analysis [Brahim-Belhouari and Bermak, 2004, Cleland et al., 2014], image analysis [Radke et al., 2005] and financial markets [Talih and Hengartner, 2005].

Traditionally, change point methods focus on detecting shifts in the **joint distribution** of all variables in the time series. There is an implicit assumption that the entire joint distribution of all the variables significantly shifts whenever a change occurs in any of the variables in the multivariate time series. This assumption works well if we only care about broad, overarching changes, but it can be restrictive and overlook an important real-world consideration: in many scenarios, we are more concerned with local changes rather than global shifts. Motivated by the causal invariance principle and invoking a causal modeling perspective, we are interested in detecting changes that specifically affect the **causal mechanisms** governing the variables in the time series.

In the field of human health and medicine, researchers aim at identifying subtle changes in specific patient conditions before the overall health deteriorates significantly, such as capturing antecedent signs and symptoms of sepsis in [Shashikumar et al., 2017] and [Goh et al., 2021]. Similarly, detecting signal changes within vast datasets to predict anomalies before the onset of overall financial distress is a critical focus in finance, as demonstrated in [Koyuncugil and Ozgulbas, 2012] and [Kliestik et al., 2018]. Therefore, it is essential to shift our focus from global joint distribution changes to causal mechanism changes in practical applications.

Additionally, from a theoretical perspective, many existing methods, including [Aminikhanghahi and Cook, 2017, Harchaoui et al., 2008, Liu et al., 2013, Saggioro et al., 2020], require independent and identically distributed (IID) samples. Although many of these methods are robust to non-IID samples in simulated data such as [Liu et al., 2013], they lack theoretical

guarantees, which are needed for trustworthy deployment of these algorithms in the real world, especially in safety-critical applications as in [Liu et al., 2018].

Driven by both practical needs and theoretical importance, we propose a novel change point detection algorithm that integrates change point detection with causal discovery. First, causal structure helps us obtain a more fine-grained look at the joint distribution by disentangling the causal mechanism of each component in the time series. Second, relying on the Causal Markov Condition assumption in the Structural Causal Model (SCM) framework of [Pearl, 2009], the correlated samples become independent and identically distributed (IID) when conditioned on their causal parents. In this context, our change point detection method aims to identify shifts in the **causal mechanisms**, specifically, the conditional distribution of the target variable given its parents.

In this paper, we summarize our main contributions as follows:

- We propose a novel non-parametric algorithm, called Causal-RuLSIF, for detecting change points in causal mechanisms within discrete-valued time series data. The algorithm assumes an underlying *Mechanism-Shift* SCM but does not impose constraints on the causal mechanisms or the data distributions. Our algorithm introduces a novel dynamic RuLSIF estimator for detecting change points, on the basis of [Yamada et al., 2013] and [Liu et al., 2013], and integrates the PCMCI algorithm for causal discovery from [Runge et al., 2019].
- We validate our method with synthetic simulations on both *soft mechanism change* and *hard mechanism change* time series, demonstrating that it reliably detects change points with high probability. We also employ our method in an air pollution application.

2 RELATED WORK

There are various change point detection methods based on the compatibility between the statistical feature shifts they detect and the characteristics of the data. Likelihood ratio methods compare the probability density of two consecutive intervals of the time series data and the significant difference in the probability density implies the change point. One representative estimator is the relative unconstrained least-squares importance fitting (RuLSIF) in [Yamada et al., 2013] and [Liu et al., 2013]. Kernel-based methods in [Harchaoui et al., 2008] and [Harchaoui et al., 2009] utilize kernel-based test statistics to test the homogeneity of sliding windows in time series data. Other methods include Probabilistic methods, Graph-based methods, clustering methods and Subspace modeling. See [Chib, 1998], [Friedman and Rafsky, 1979], [Keogh et al., 2001] and [Itoh and Kurths, 2010], respectively. See [Aminikhanghahi and Cook, 2017] for a survey. As discussed in the instruction section, such existing methods concentrate on joint features associated with the whole time series, with many assuming IID samples.

In addition to these methods, some change point detection techniques specifically target different challenges or leverage additional information embedded in the dataset. [Cho and Fryzlewicz, 2015], [Barigozzi et al., 2018], [Kovács et al., 2023] introduced methods for detecting multiple change points in high-dimensional time series. [Killick et al., 2012], [Maidstone et al., 2017] focus on increasing the computational efficiency. [Safikhani and Shojaie, 2022] proposed a method for estimating structural change points and parameters in high-dimensional piecewise VAR models. [Qiu et al., 2012] applied Granger causality for anomaly detection in time series data. [Bardet et al., 2010], [Diop and Kengne, 2022] detect change points in a large class of causal time series models including AR(∞), ARCH(∞) and TARCH(∞) models. More recently, [Huang et al., 2024] introduced a change point detection method within the context of a response variable Y and covariates X, assuming a linear SCM that generates Y from X.

To the best of our knowledge, no other non-parametric change point detection models are capable of identifying shifts in causal mechanisms without imposing constraints on the form of the mechanism.

Additionally, there is a noteworthy "bonus" contribution to the field of causal discovery. Our change point detection algorithm, when integrated with a postprocessing causal discovery step, is, to our knowledge, the first non-parametric method capable of learning sudden-shift causal structures from non-stationary time series data without assuming strict periodicity as in [Gao et al., 2023]. More related work on this contribution can be found in Appendix B.

3 CAUSAL-RuLSIF: DETECTING CHANGE POINT IN A CAUSAL TIME SERIES

In this section, we present the framework and problem formulation for change point detection in causal time series, including the key assumptions.

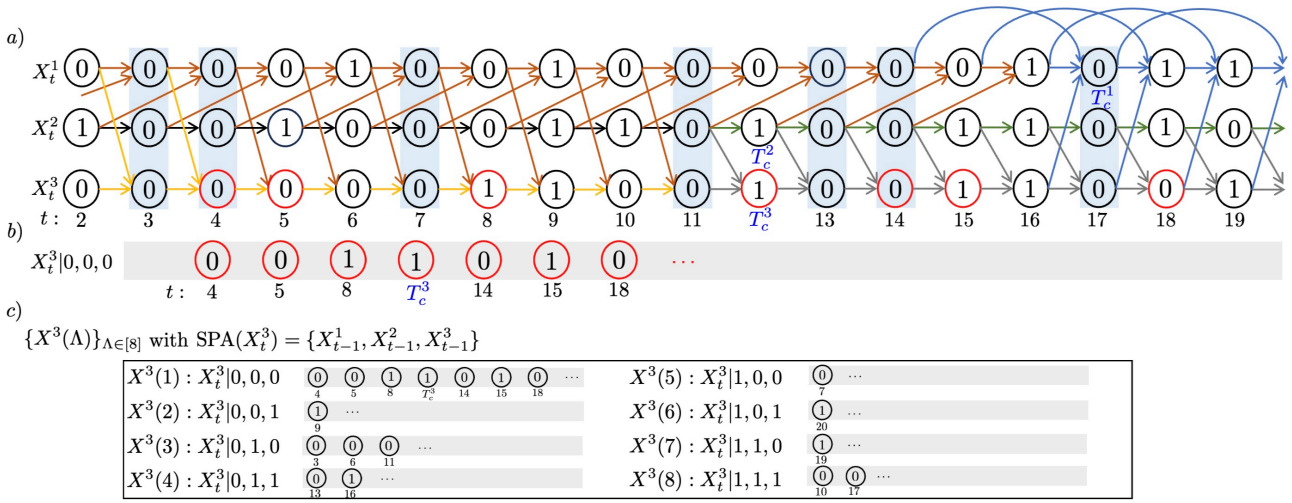


Figure 1: An illustration of collecting time series segments (Def.3.3) in Causal-RuLSIF. a). The Causal Graph of a 3-variate binary time series $V = \{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3\}$ with a *Mechanism-Shift* SCM (Def.3.1), where $c^1 = c^2 = c^3 = 1$, indicating one change point in each univariate time series. Note that the location of change points can differ across univariate time series. E.g., the change point for X^j, where $j \in [3]$, is denoted as T_c^j, and in this case $T_c^1 \neq T_c^2 = T_c^3$. The edges of different colors represent distinct causal mechanisms. E.g., for $\mathbf{X}^1$, the causal mechanism indicated by the brown edges shifts to a different causal mechanism represented by the blue edges at the time point T_c^1. There are no restrictions on the causal mechanisms. E.g., for $\mathbf{X}^1$, the two causal mechanisms before and after T_c^1 differ in both the parent sets and the deterministic function $f_{1,t}(\cdot)$; we refer to this scenario as *hard mechanism change*. In contrast, for $\mathbf{X}^2$, the two causal mechanisms differ only in $f_{2,t}(\cdot)$ while the parent set remains invariant over time; we refer to this situation as *soft mechanism change*. b). Construction of a time series segment for $\mathbf{X}^3$. With $\mathrm{PA}(X^3_{t<T_c^3}) = \{X_{t-1}^1, X_{t-1}^3\}$, $\mathrm{PA}(X^3_{t\geq T_c^3}) = \{X_{t-1}^2, X_{t-1}^3\}$, leading to $\mathrm{SPA}(X_t^3) = \mathrm{PA}(X^3_{t<T_c^3}) \cup \mathrm{PA}(X^3_{t\geq T_c^3})$ for all t (Def.3.2). Given that $|\mathrm{SPA}(X_t^3)| = 3$ and the domain size of binary variables is 2, one specific configuration of $\mathrm{SPA}(X_t^3)$ is $\{0,0,0\}$. To form the corresponding time series segment, we need to collect all variables $X_t^3 \in \mathbf{X}^3$ for which t satisfies $\mathrm{spa}(X_t^3) = \{X_{t-1}^1 = 0, X_{t-1}^2 = 0, X_{t-1}^3 = 0\}$. c). Collection of 8 time series segments for $\mathbf{X}^3$ as there are total 8 parents' configurations, starting from $\{0,0,0\}$, $\{0,0,1\}$, $\{0,1,0\}$, and continuing to $\{1,1,1\}$.

3.1 Preliminaries

Let $\mathcal{G}(V, E)$ denote the underlying causal graph. The set of all incoming neighbors for each variable $X \in V$ is defined as the parent set, denoted by $\mathrm{PA}(X)$. For any $X, Y \in V$ and $S \subset V$, we denote the conditional independence: X is independent of Y conditioned on S, by $X \perp\!\!\!\perp Y \mid S$.

For simplicity, let's define sets: $[b] := \{1, 2, ..., b\}$ and $[a, b] := \{a, a+1, ..., b\}$, where $a, b \in \mathbb{N}$. Let $X_t^j \in \mathbb{R}$ represent the variable of jth *component* univariate time series at time t; $\mathbf{X}^j = \{X_t^j\}_{t\in[T]} \in \mathbb{R}^T$ denote a *component* univariate time series which is a component in multivariate time series V and $\mathbf{X}_t = \{X_t^j\}_{j\in[n]} \in \mathbb{R}^n$ denote a slice of all variables at time point t. Note that $V = \{\mathbf{X}^j\}_{j\in[n]} = \{\mathbf{X}_t\}_{t\in[T]} \in \mathbb{R}^{n\times T}$ represents a n-variate time series, where it can be interpreted either as a collection of n univariate time series (viewed horizontally) or as a sequence of time slices across all variables (viewed vertically). By default, we assume $n > 1$ and hence $\mathbf{X}^j \subsetneq V$, and $p(V) \neq 0$, where $p(.)$ denotes the probability. For discrete-valued time series V, we assume that the domain of each component $\mathbf{X}^j \subsetneq V$, denoted by $D = \{d_1, \cdots, d_s\}$, is the same, that is, $X_t^j \in D$ for all $j \in [n]$ and $t \in [T]$.

In this paper, the subscript notation over $\{\cdot\}$ indicates that the elements within $\{\cdot\}$ are aggregated across all values of the subscripts.

As $\mathrm{PA}(X_t^j)$ represents a set containing random variables, $\mathrm{pa}(X_t^j)$ refers to one specific configuration. Consider $\mathrm{PA}(X_t^j) = \{X_{t_1}^{i_1}, X_{t_2}^{i_2}, \cdots\}$ where $1 \leq i_1 \leq i_2 \leq \cdots \leq n$ and $T \geq t_1 \geq t_2 \geq \cdots \geq 1$. E.g., $\mathrm{pa}(X) = \{1, 0, 1, 2, \cdots\}$ corresponds to a particular configuration (or instantiation). We assume that the configuration set is ordered by the parent set $\{X_{t_1}^{i_1}, X_{t_2}^{i_2}, \cdots\}$, prioritizing ascending variable indices over descending time indices.

We start by defining a type of Structural Causal Model (SCM) that captures our setting.

Definition 3.1 (*Mechanism-Shift* SCM)**.** *A Mechanism-Shift SCM is a tuple* $\mathcal{M} = \langle V, \mathcal{F}, \mathcal{E}, \mathbb{P} \rangle$ *where there exists a* $\tau_{\max} \in \mathbb{N}^+$*, defined as:* $\tau_{\max} \coloneqq \max\{\tau : X^i_{t-\tau} \in \mathrm{PA}(X^j_t), i, j \in [n], t \in [T]\}$*, such that each variable* $X^j_{t>\tau_{\max}} \in V$ *is a deterministic function of its parent set* $\mathrm{PA}(X^j_{t>\tau_{\max}}) \in V$ *and an unobserved (exogenous) variable* $\epsilon^j_{t>\tau_{\max}} \in \mathcal{E}$*:*

$$X^j_t = f_{j,t}(\mathrm{PA}(X^j_t), \epsilon^j_t),\ j \in [n], t \in [\tau_{\max}+1, T], \quad (1)$$

and for each $j \in [n]$*, there exists an ascending sequence of time points* $\{T^j_1, T^j_2, \cdots, T^j_{c^j}\}$ *with* $c^j \in \mathbb{N}^+, \tau_{\max} < T^j_1 < \cdots < T^j_{c^j} < T$ *such that:*

$$a)\ f_{j,t_1} = f_{j,t_2},\ \text{if } \forall c \in [c^j] \text{ s.t. } T^j_c \notin [t_1, t_2]; \quad (2)$$

$$b)\ f_{j,t_1} \neq f_{j,t_2},\ \text{if } \exists c \in [c^j] \text{ s.t. } T^j_{c-1} \leq t_1 < T^j_c \leq t_2 < T^j_{c+1}; \quad (3)$$

$$c)\ \mathrm{PA}(X^j_{t_1}) = \{X^i_{t_1-s} : X^i_{t_2-s} \in \mathrm{PA}(X^j_{t_2}), i \in [n]\},\ \text{if } \forall c \in [c^j] \text{ s.t. } T^j_c \notin [t_1, t_2]; \quad (4)$$

$$d)\ \epsilon^j_t \text{ are i.i.d. } \forall t \in [T]. \quad (5)$$

are satisfied for all $t_1, t_2 \in [\tau_{\max}+1, T]$*, where* $f_{j,t}, f_{j,t_1}, f_{j,t_2} \in \mathcal{F}$ *and* $\{\epsilon^j_t\}_{t\in[T]}$ *are jointly independent with probability measure* $\mathbb{P}$*.* $\tau_{\max}$ *is the finite maximal lag in the causal graph* $\mathcal{G}$*. Define* $T^j_0 \coloneqq \tau_{\max}$ *and* $T^j_{c^j+1} = T$*.*

This indicates that within the univariate time series $\mathbf{X}^j$ in V, there is a finite number of change points, denoted by c^j. The variables in $\mathbf{X}^j$ before and after each change point should exhibit distinct causal mechanisms, as illustrated in b), without overlapping with other change points. Two variables in $\mathbf{X}^j$ that do not span any change points should share the same function and time-shift invariant parents, as depicted in a) and c). An instance of this model is shown in Fig. 1a).

Definition 3.2 (*Illusory Parent Sets*)**.** *For a univariate time series* $\mathbf{X}^j \in V$ *with Mechanism-Shift SCM having change points* $\{T^j_1, T^j_2, \cdots, T^j_{c^j}\}$ *with* $c^j \in \mathbb{N}^+, \tau_{\max} < T^j_1 < \cdots < T^j_{c^j} < T$*, parent set index* $\mathrm{pInd}^j_{k\in[c^j+1]}$ *is defined as:*

$$\mathrm{pInd}^j_k \coloneqq \{(\tau_i, y_i)\}_{i\in[m]},\ \text{given}\ \mathrm{PA}(X^j_t) = \{X^{y_1}_{t-\tau_1}, X^{y_2}_{t-\tau_2}, ..., X^{y_m}_{t-\tau_m}\}, \quad (6)$$

for all $t \in [T^j_{k-1}, T^j_k]$*, where* $m = |\mathrm{PA}(X^j_t))|$*,* τ_i *is the time lag and* y_i *is the variable index; set* $T^j_0 = \tau_{max}$ *and* $T^j_{c^j+1} = T$*. Given* pInd^j_k*, Illusory Parent Sets are defined as:*

$$\mathrm{PA}_k(X^j_t) = \{X^{y_i}_{t-\tau_i} : (\tau_i, y_i) \in \mathrm{pInd}^j_k\},\ \forall k \in \{k : t \notin [T^j_{k-1}, T^j_k]\} \quad (7)$$

Essentially, the illusory parent sets of X^j_t are time-shifted versions of the parent sets of other variables in $\mathbf{X}^j$ that exhibit different causal mechanisms than X^j_t across change points. These sets generalize a concept described in [Gao et al. 2023]. There should be c^j illusory parent sets for each variable X^j_t.

For simplicity, we extend *Illusory Parent Sets* to also include the true parent set of X^j_t, resulting in each variable X^j_t having $c^j + 1$ illusory parent sets, one of which is $\mathrm{PA}(X^j_t)$. The term *illusory* now indicates that these parent sets *may* not exist for t but must be valid for some $t' \in [T]$.

Further, we define the union parent set as:

$$\mathrm{SPA}(X^j_t) \coloneqq \cup_{k\in[c^j+1]}\mathrm{PA}_k(X^j_t), t \in [\tau_{\max}+1, T] \quad (8)$$

E.g., in Fig. 1a), $\mathrm{PA}(X^1_t) = \{X^1_{t-1}, X^2_{t-2}\}$ with $t < T^1_{cp}$ and $\mathrm{PA}(X^1_t) = \{X^1_{t-1}, X^1_{t-3}, X^3_{t-1}\}$ with $t \geq T^1_{cp}$. As $c^1 = 1$, we have two illusory parent sets with index $k \in [c^1 + 1]$. Specifically, the two illusory parent sets of $\mathbf{X}^1$ are $\mathrm{PA}_1(X^1_t) = \{X^1_{t-1}, X^2_{t-2}\}$ and $\mathrm{PA}_2(X^1_t) = \{X^1_{t-1}, X^1_{t-3}, X^3_{t-1}\}$, leading to $\mathrm{SPA}(X^1_t) = \{X^1_{t-1}, X^1_{t-3}, X^2_{t-2}, X^3_{t-1}\}$.

Definition 3.3 (*time series segments*)**.** *For a univariate discrete-valued time series* $\mathbf{X}^j \in V$ *with Mechanism-Shift SCM having change points set* $\{T^j_1, T^j_2, \cdots, T^j_{c^j}\}$ *with* $c^j \in \mathbb{N}^+, \tau_{\max} < T^j_1 < \cdots < T^j_{c^j} < T$*, and finite domain set* $D = \{d_1, d_2, \cdots, d_s\}$*, the time series segments are a collection of non-overlapping non-empty subsets* $\{X^j(\Lambda)\}_{\Lambda\in[s^{|\mathrm{SPA}(X^j_t)|}]}$ *such that:*

$$X^j(\Lambda) \coloneqq \{X^j_t : t \in [\tau_{max}+1, T], \mathrm{pa}(X^j_t) = \Sigma^j_\Lambda\},\ \Lambda \in [s^{|\mathrm{SPA}(X^j_t)|}]. \quad (9)$$

Here, Σ *represents the configuration matrix of* $\mathrm{SPA}(X^j_t)$*, with each row representing a unique configuration of* $\mathrm{SPA}(X^j_t)$*.* Λ *refers to the row index of one specific configuration in* Σ*. Since the domain size is* $|D| = s$*,* Σ *has* $s^{|\mathrm{SPA}(X^j_t)|}$ *rows, representing all possible configurations of* $\mathrm{SPA}(X^j_t)$*. Based on the above definition, we can observe that:*

$$\bigcup_{\Lambda\in[s^{|\mathrm{SPA}(X^j_t)|}]} \{X^j(\Lambda)\} = \mathbf{X}^j, \quad (10)$$

$$\bigcap_{\Lambda_1\neq\Lambda_2\in[s^{|\mathrm{SPA}(X^j_t)|}]]} \{X^j(\Lambda_1), X^j(\Lambda_2)\} = \emptyset. \quad (11)$$

In summary, the time series segments $\{\mathbf{X}^j(\Lambda)\}$ partition $\mathbf{X}^j$ into multiple non-overlapping, non-empty sub-time series, conditioned on the configurations of the union parent set $\mathrm{SPA}(X^j_t)$. Variables X^j_t within the same $\mathbf{X}^j(\Lambda)$ share identical configurations of $\mathrm{SPA}(X^j_t)$. In Fig. 1b) and 1c), $\mathbf{X}^3$ is divided into 8 time series

segments, with each segment having the same configuration spa(X_t^3). Refer to Appendix A.1 for a detailed example.

RuLSIF: Robust Distributional Distance Estimation. Given two distributions $p(\cdot)$ and $p'(\cdot)$ defined over the same support, there exist many metrics measuring the distance between them. In Yamada et al., 2013, the authors proposed a method named RuLSIF, with an α-relative divergence estimation $r_\alpha(x)$ and a corresponding metric, α-relative Pearson Divergence PE_α defined as following:

$$r_\alpha(x) \coloneqq \frac{p(x)}{(1-\alpha)p(x)+\alpha p'(x)} \coloneqq \frac{p(x)}{q_\alpha(x)} \tag{12}$$

$$PE_\alpha \coloneqq \frac{1}{2}\mathbb{E}_{x\sim q_\alpha}[(r_\alpha(x)-1)^2] \tag{13}$$

where α is a parameter used to bound $r_\alpha(x)$.

In our work, we innovatively introduce a dynamic parameter, β_i, within the shifting window framework, rather than relying solely on a fixed hyperparameter α in PE_α. The parameter β_i represents the *concentration* of the distribution $p'(x)$ within a mixture distribution, where i denotes the window index.

As the shifting window moves from $t=0$ to $t=T$ with a window size of $2n_{\text{w}}$ and a stride size of n_{st}, the window index i increases. Notably, β_i transitions from 0 to 1 when the shifting window crosses a change point. Specifically, if a change point occurs in the second half of the window, the samples in that half are drawn from a mixture distribution represented by $(1-\beta_i)p(x)+\beta_i p'(x)$. Here, β_i indicates the proportion of samples in the second half derived from $p'(x)$, reflecting the change in the underlying distribution of data as the window shifts.

Based on this, the dynamic density ratio and the corresponding Pearson Divergence (PE) score are defined as follows:

$$r_{\alpha\beta_i}(x) \coloneqq \frac{p(x)}{(1-\alpha\beta_i)p(x)+\alpha\beta_i p'(x)} \coloneqq \frac{p(x)}{q_{\alpha\beta_i}(x)} \tag{14}$$

$$PE_{\alpha\beta_i} \coloneqq \frac{1}{2}\mathbb{E}_{x\sim q_{\alpha\beta_i}}[(r_{\alpha\beta_i}(x)-1)^2] \tag{15}$$

Further details on dynamic RuLSIF, including the rationale for its selection as the metric and the official definition of β_i are provided in Appendix A.2.

Assumptions. We highlight three key assumptions here: (1) For any univariate time series, there is at most one change point, or if multiple change points exist, the temporal distance between any two must be at least Δ_c. (2) There is a lower bound on the dynamic PE divergence. (3) The number of change points is known. The first assumption ensures that multiple change points do not have a cancellation effect making it difficult to identify the change point. The second assumption is commonly applied in distributional distance estimation, particularly in finite sample and discrete settings. The third assumption offers theoretical guarantees but can be relaxed in practice, as demonstrated by the experimental results. The formalization of these three assumptions is provided in Appendix A.3 , which also includes other assumptions established in prior works.

4 CAUSAL-RuLSIF ALGORITHM

In this section, we first introduce an algorithm named Causal-RuLSIF. We then demonstrate the correctness of Causal-RuLSIF in accurately estimating the change point within a specified confidence interval. Finally, we provide a computational complexity analysis, highlighting both the contribution of the algorithm and its limitations.

Overview of Algorithm 1 Causal-RuLSIF: Please note that in this section, we assume a maximum of one change point per univariate time series; however, the framework can be easily extended to accommodate multiple change points. The parameters n_{w}, n_{st}, and α are set at the beginning. The value of τ_{ub} establishes the upper bound for the search space of $\tau_{\max}$ in PCMCI. Each component $\mathbf{X}^j$, where $j \in [n]$, is analyzed sequentially.

Using PCMCI [Runge et al., 2019], we obtain $\widehat{\text{SPA}}(X_t^j)$ for each variable $X_t^j \in \mathbf{X}^j$ (line 2). To ensure balanced samples as required in Theorem A.1, PCMCI is applied to non-overlapping consecutive intervals, and the edges obtained are collected. Based on the parent configurations of $\widehat{\text{SPA}}(X_t^j)$, we construct time series segments denoted as $\{X^j(\Lambda)\}_\Lambda$.

Next, we apply dynamic RuLSIF on sliding windows over $\{X^j(\Lambda)\}_\Lambda$, resulting in the divergence series represented by $\{\widehat{PE}^{j,\Lambda}_{\alpha\beta_i}\}_{i,\Lambda}$ (lines 6-8). Note that the number of divergence series corresponds to the number of time series segments. The change point estimator, denoted as $\widehat{T}_c^j$, is the window index i that maximizes $\{\widehat{PE}^{j,\Lambda}_{\alpha\beta_i}\}_{i,\Lambda}$ (lines 9-10).

Since $\widehat{T}_c^j$ represents the window index within the time series segments rather than the original time index in $\mathbf{X}^j$, it must be projected back to $\mathbf{X}^j$. The final change point estimator is then defined as $\widetilde{T}_c^j$ (line 11). After obtaining an accurate estimator $\widehat{T}_c^j$, as established in Theorem 4.2, one may optionally run PCMCI on the samples X_t^j before and after the change point $\widetilde{T}_c^j$ to fully learn the underlying causal graph.

Algorithm 1 Causal-RuLSIF

Input: A n-variate time series $V = (\mathbf{X}^1, \cdots, \mathbf{X}^n)$ with domain set $D = \{d_1, ...d_s\}$. Set appropriate τ_{ub}, n_{w}, n_{st} and α.
A superset of the parent set is obtained using PCMCI with τ_{ub} and denote it by $\widehat{\text{SPA}}(X_t^j)$ $\forall j, t$.
for $\mathbf{X}^j$ where $j \in [n]$ **do**
  $\widehat{\text{PA}}(X_t^j) \leftarrow \widehat{\text{SPA}}(X_t^j)$
  Construct $\{X^j(\Lambda)\}$ based on configurations of $\widehat{\text{SPA}}(X_t^j)$.
  for $X^j(\Lambda)$ where $\Lambda \in [s^{|\widehat{\text{SPA}}(X_t^j)|}]$ **do**
    Store divergence score series $\{\widehat{PE}_{\alpha\beta_i}^{j,\Lambda}\}$ with RuLSIF on sliding windows shifting over $X^j(\Lambda)$ where $i \in [\lfloor \frac{T_{\text{sub}} - 2n_{\text{w}}}{n_{\text{st}}} \rfloor + 1]$ and $T_{\text{sub}} := |X^j(\Lambda)|$.
  end for
  $\widehat{\Lambda} \leftarrow \arg_\Lambda \max\{\widehat{PE}_{\alpha\beta_i}^{j,\Lambda}\}_{i,\Lambda}$ ▷ Pick the Λth time series segments whose maximum value over all segments is maximum.
  $\widehat{T}_c^j \leftarrow \arg_i \max\{\widehat{PE}_{\alpha\beta_i}^{j,\widehat{\Lambda}}\}_i$ ▷ Pick the ith window index with maximum PE score on $X^j(\widehat{\Lambda})$.
  Project $\widehat{T}_c^j$, where $j \in [n]$, back to the original time series $\mathbf{X}^j$ with $\widetilde{T}_c^j = \frac{1}{2}(t_{\widehat{T}_c^j} + t_{\widehat{T}_c^j + 1})$, where $t_{\widehat{T}_c^j}$ is the time index in $\mathbf{X}$ corresponding to the midpoint of the window indexed by $\widehat{T}_c^j$ in $X^j(\widehat{\Lambda})$.
  Consider $X_{t-\tau}^j \in \widehat{\text{PA}}(X_t^j)$. Remove $X_{t-\tau}^j$ from $\widehat{\text{PA}}(X_t^j)$ if $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \left(\widehat{\text{SPA}}(X_t^j) \cup \widehat{\text{SPA}}(X_{t-\tau}^i)\right) \setminus X_{t-\tau}^i$ on samples $\{X_t^j\}_{t < \widetilde{T}_c^j}$ and $\{X_t^j\}_{t \geq \widetilde{T}_c^j}$ respectively.
end for
return $\widetilde{T}_c^j$ and $\widehat{\text{PA}}(X_t^j)$ $\forall j, t$.

4.1 Theoretical Guarantees

Theorem 4.1 establishes that if the window W_i only contains samples from a single causal mechanism, meaning there is no change point included in this window, the estimated relative Pearson Divergence $\widehat{PE}_{\alpha\beta_i}$ is close to zero with high probability. Theorem 4.2 provides a confidence interval for the change point estimator $\widehat{T}_c^j$. Proofs can be found in Appendix A.4.

Theorem 4.1. *Let $\{\widehat{PE}_{\alpha\beta_i}\}_i$ be the estimated PE series for one time series segment $X^j(\Lambda) \subsetneq \mathbf{X}^j \subsetneq V$ and T_c^j denote the true change point in this time series segments. Under certain assumptions, we have that $\forall i \in \{i : in_{\text{st}} + 2n_{\text{w}} - 1 < T_c^j\}$*

$$\Pr\big(\max\{\widehat{PE}_{\alpha\beta_i}\}_i < o(1)\big) > 1 - \frac{a_{\text{w}} - 2}{b_{\text{st}} \log T_{\text{sub}}} - \frac{a_{\text{w}}}{T_{\text{sub}}},$$

where $b_{\text{st}} = \lfloor \frac{\log T_{\text{sub}}}{n_{\text{st}}} \rfloor$, $a_{\text{w}} = \lceil \frac{T_{\text{sub}}}{n_{\text{w}}} \rceil$ and $T_{\text{sub}} := |X^j(\Lambda)|$.

The window index i satisfying $in_{\text{st}} + 2n_{\text{w}} - 1 < T_c^j$ guarantees that all the samples in W_i are collected from the same distribution. Theorem 4.1 states that the maximum estimated PE divergence series obtained from such windows are bounded by any positive constant with probability $1 - \frac{a_{\text{w}} - 2}{b_{\text{st}} \log T_{\text{sub}}} - \frac{a_{\text{w}}}{T_{\text{sub}}}$ given large enough n_{w}. Note that a_{w} and b_{st} are constants whose specific values are determined by the chosen window $2n_{\text{w}}$ and stride size n_{st}.

Theorem 4.2. *Let $\{\widehat{PE}_{\alpha\beta_i}\}_i$ be the estimated PE series for one time series segment $X^j(\Lambda) \subsetneq \mathbf{X}^j \subsetneq V$ and T_c^j denote the true change point in this time series segments. $\widehat{T}_c^j$ denotes the estimator of T_c^j obtained by:*

$$\widehat{T}_c^j = \arg_i \max\{\widehat{PE}_{\alpha\beta_i}\}_i \tag{16}$$

Under certain assumptions, we have that given large enough n_{w}, $\forall i \in [\tau_{max} + 1, T]$

$$\Pr\big(|\widehat{T}_c^j - T_c^j| < 2n_{\text{w}}\big) > \left(1 - \frac{a_{\text{w}} - 2}{b_{\text{st}} \log T_{\text{sub}}} - \frac{a_{\text{w}}}{T_{\text{sub}}}\right)\left(1 - \frac{1}{n_{\text{w}}}\right). \tag{17}$$

where $b_{\text{st}} = \lfloor \frac{\log T_{\text{sub}}}{n_{\text{st}}} \rfloor$, $a_{\text{w}} = \lceil \frac{T_{\text{sub}}}{n_{\text{w}}} \rceil$, and $T_{\text{sub}} := |X^j(\Lambda)|$.

The change point estimator has an estimation error smaller than the total window size with a known probability.

4.2 Computational Complexity Analysis

There is a trade-off between computational complexity and result accuracy in the proposed algorithm, with a preference for the latter when significant causal relationships exist. The proposed algorithm functions in two distinct phases as follows:

Phase One (causal discovery) focuses on uncovering the underlying causal structure of the entire non-stationary time series. Phase Two (change point detection) centers on change point detection within each time series segment, leveraging the causal structures estimated in Phase One.

In Phase One, the worst-case computational complexity in PCMCI Runge et al. 2019 is given by $n^3\tau_{\max}^2 + n^2\tau_{\max}$ total conditional independence tests. The running time of PCMCI is then scaled by the time series length T and the size of the conditioning set in conditional independence tests.

In Phase Two, the leading term of the asymptotic convergence rates of $\widehat{PE}$ is $n_{\text{w}}^{-1/2}$, as discussed in Yamada et al., 2013. In our algorithm, each time series segment contains a total of $\left\lfloor \frac{T_{\text{sub}} - 2n_{\text{w}}}{n_{\text{st}}} \right\rfloor + 1$ sliding windows, which implies that the number of $\widehat{PE}$ estimates

is also $\left\lfloor \frac{T_{\text{sub}}-2n_{\text{w}}}{n_{\text{st}}} \right\rfloor + 1$. Consequently, the total number of estimators $\{\widehat{PE}^{j,\Lambda}_{\alpha\beta_i}\}_{i,\Lambda}$ for $\mathbf{X}^j$ with $j \in [n]$ can be expressed as:

$$|\{\widehat{PE}^{j,\Lambda}_{\alpha\beta_i}\}_{i,\Lambda}| \approx \left(\left\lfloor \frac{\frac{T}{|s^{|\widehat{\text{SPA}}(X_t^j)|}|} - 2n_{\text{w}}}{n_{\text{st}}} \right\rfloor + 1 \right) |s^{|\widehat{\text{SPA}}(X_t^j)|}| \tag{18}$$

The RuLSIF method for estimating relative PE divergence is computationally efficient as the optimization process utilizes a kernel-based approach with a square loss.

The complexity of Phase One increases cubically with increasing dimensions of n. Compared to existing high-dimensional methods in Killick et al., 2012 and Kovács et al., 2023, the computational complexity of Phase Two depends on the sparsity of the underlying causal structure among the n-variate time series, rather than directly on n. If the causal structure is sparse, the complexity of Phase Two remains unaffected. Specifically, for a target time series $\mathbf{X}^j$, adding one more time series $\mathbf{X}^i$ does not increase the complexity of detecting change points in $\mathbf{X}^j$, assuming there are no time-lagged causal effects from $\mathbf{X}^i$ to $\mathbf{X}^j$. This advantage distinguishes our approach from existing methods.

However, a limitation arises when the parent size of each variable increases linearly with n, as the number of time series segments for $\mathbf{X}^j$ in Phase Two will increase exponentially with the size of the superset parent set $\widehat{\text{SPA}}(X_t^j)$.

As shown in Fig. 7(a) (Runtime for algorithms), the additional computational time required by CausalRuLSIF, compared to RuLSIF, arises from Phase One and the need to estimate the PE series across multiple univariate time series segments (Phase Two), rather than processing the entire time series like other baselines. This introduces a trade-off between analyzing multiple univariate time series segments and handling one high-dimensional data. The computational burden is balanced with the advantages of segment-wise analysis.

The complexity of Phase One scales with T, depending on the specific choice of the conditional independence (CI) test. As the proposed algorithm applied sliding window techniques, the complexity of Phase Two increases linearly with T, consistent with other methods in Killick et al., 2012 and Kovács et al., 2023.

5 EXPERIMENTS

In this section, we present an empirical evaluation of our approach compared to existing methods, using both synthetic and real-world datasets. Section 5.1 analyzes simulation results on binary multivariate time series, while Section 5.2 offers a case study. The Python code is provided at https://github.com/CausalML-Lab/CausalCPD.

Sample efficiency presents a significant challenge for our algorithm, given that $T_{\text{sub}} \approx T/|s^{|\widehat{\text{SPA}}(X_t^j)|}|$. To address this issue, we either implement the k-PC algorithm from Kocaoglu, 2024 or directly constrain the candidate parent set using the top-K causal strengths derived from the PCMCI algorithm. These enhancements enable our algorithm to effectively handle non-binary discrete-valued multivariate time series with larger domain sizes ($D > 2$) and a greater number of component time series ($n > 3$). Comprehensive experimental results are provided in Appendix E.

Additionally, we extend our approach to detect multiple change points with prior knowledge of their quantity (Assumption A8). The experimental results and corresponding theoretical guarantees are also available in Appendix F.

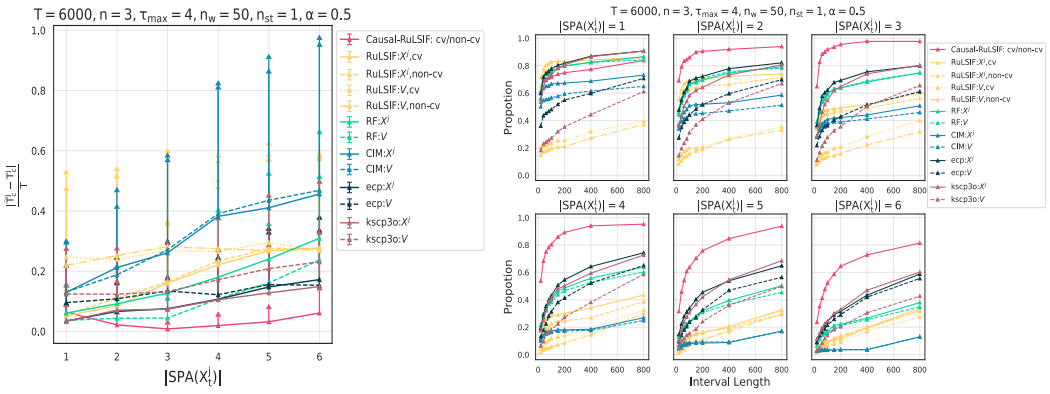
5.1 Simulations on binary multivariate time series

In this section, we have five baseline algorithms, including RuLSIF algorithm in Liu et al., 2013, traditional method change in mean (CIM) in Vostrikova, 1981, changeforest algorithm (RF) in Londschien et al., 2023, ecp algorithm in James and Matteson, 2013 and kscp3o algorithm in Zhang et al., 2017. All experimental code will be available online.

The details of synthetic binary time series generation can be found in Appendix C.

We have two methods to evaluate the performance of the algorithms. The first method is to calculate the estimation error using $\frac{|\widetilde{T}_c^j - T_c^j|}{T}$. The second method is to construct an ROC curve by setting an interval length, denoted as $2Q$. With the change point estimator $\widetilde{T}_c$ and interval length Q, we increment a counter by 1 if the true change point T_c^j falls within the interval $[\widetilde{T}_c^j - Q, \widetilde{T}_c^j + Q]$. This count is then averaged over the total number of univariate time series in the 100 random trials. This is a common metric for measuring the performance of the change point detection algorithm, as described in Liu et al., 2013 and Harchaoui et al., 2008.

Please note that in the RuLSIF method, the kernel width σ and the regularization parameter in the kernel



(a) Influence of $|\mathrm{SPA}(X_t^j)|$ on estimation error (b) ROC curves for different $|\mathrm{SPA}(X_t^j)|$

Figure 2: Causal-RuLSIF is tested on 3-multivariate time series with $T = 6000, \tau_{\max} = 4, n_{\mathrm{w}} = 50, n_{\mathrm{st}} = 1$ with *soft mechanism change*. Every line with a different color corresponds to a different algorithm and different linestyle corresponds to a different setting. X^j in the legend means the algorithm is applied to each univariate time series while V means the algorithm is used for the whole n-variate time series V. Every marker corresponds to the average error or average accuracy rate over 100 random trials. The error bar represents the standard error for the averaged statistics. a) Influence of $|\mathrm{SPA}(X_t^j)|$ on estimation error $\frac{|\widetilde{T}_c^j - T_c^j|}{T}$. b) ROC curves for different $|\mathrm{SPA}(X_t^j)|$.

function are typically chosen using cross-validation, as outlined in Liu et al. 2013 and Harchaoui et al., 2008. This approach is justified, as high-dimensional data can render these parameters more sensitive. However, for binary time series, cross-validation may not be necessary when applying our method, Causal-RuLSIF. One possible reason is that our algorithm avoids the complexities associated with high-dimensional data V by focusing on the analysis of one-dimensional time series segments $X^j(\Lambda) \subsetneq V$. As shown in Fig 2(a) and Fig 2(b), the red line represents our algorithm. With or without the cross-validation technique does not influence its performance. For RuLSIF, the cross-validation (cv) and no cross-validation (non-cv) do not overlap.

For all the baselines, we either apply them to the entire n-variate time series or to each component $\mathbf{X}^j$. In the former scenario, multiple change point estimations are expected, making it difficult to determine which corresponds to which univariate time series. For the RuLSIF method, we select the estimations with the top n change scores and randomly assign them to each univariate time series. Regarding RF, CIM, ecp, and kscp3o, even if the optimal estimations are selected when T_c^j is known, their performance does not surpass ours. In the case of RF and CIM, the change point detection relies on the significance of certain statistics. If these methods fail to detect any change point, we set the estimated change point to T.

From Fig 2(a) and Fig 2(b), Causal-RuLSIF is not optimal when $|\mathrm{SPA}(X_t^j)| = 1$. In this scenario, each variable X_t^j only receives an incoming edge from its only parent X_{t-1}^j, indicating no correlation among different time series. This special structure makes certain baselines more advantageous, particularly considering the potential false positive edges identified by the proposed algorithm, which lacks prior knowledge of the absence of correlations and limited sample size. Therefore, it is reasonable for some baselines to exhibit better performance when applied individually to each time series.

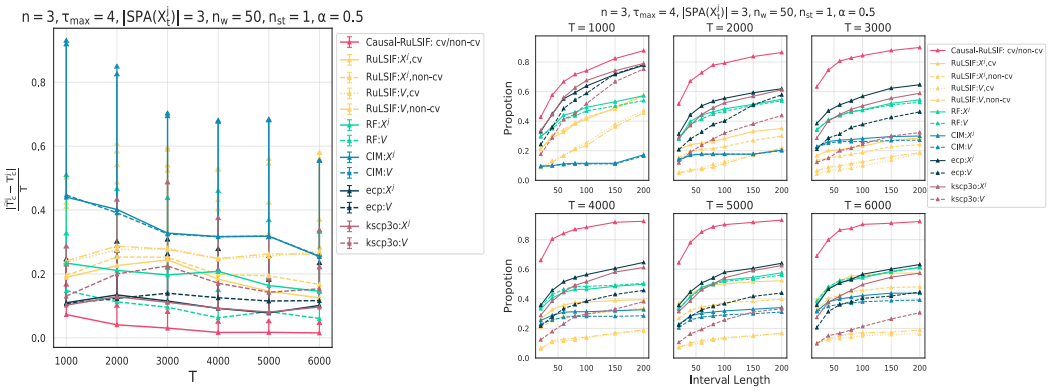
However, when $|\mathrm{SPA}(X_t^j)| > 1$, our algorithm outperforms others, as it focuses on shifts in causal mechanisms given other time series. The performance of Causal-RuLSIF decreases as $|\mathrm{SPA}(X_t^j)|$ increases because the effective sample size decreases with the number of parent configurations, which is $2^{|\mathrm{SPA}(X_t^j)|}$ for binary time series.

Additional experiments on *hard mechanism change*, the impact of n_{w}, the relative location of T_c^1 and T_c^2, and runtime analysis are provided in Appendix D.

From Fig 3(a) and Fig 3(b), it is evident that increasing the length T of the time series V enhances the performance of Causal-RuLSIF, thus practically validating the consistency of the algorithm.

5.2 Case Study

Here, we construct an experiment with a real-world air pollution dataset. This dataset monitors the amount



(a) Influence of T on estimation error

(b) ROC curves for different T

Figure 3: Causal-RuLSIF is tested on 3-multivariate time series with $|\text{SPA}(X_t^j)| = 3, \tau_{max} = 4, n_{\text{w}} = 50, n_{\text{st}} = 1$ with *soft mechanism change*. Every line with a different color corresponds to a different algorithm and different linestyle corresponds to a different setting. X^j in the legend means the algorithm is applied to each univariate time series while V means the algorithm is used for the whole n-variate time series V. Every marker corresponds to the average error or average accuracy rate over 100 random trials. The error bar represents the standard error for the averaged statistics. a) Influence of T on estimation error $\frac{|\widetilde{T}_c^j - T_c^j|}{T}$. b) ROC curves for different T.

Table 1: Causal-RuLSIF in PM_{10} dataset

X	$\widetilde{T}_c^j$	$\widehat{\text{PA}}(X^j_{t<\widetilde{T}_c^j})$; $\widehat{\text{PA}}(X^j_{t\geq\widetilde{T}_c^j})$
$\mathbf{X}^{\text{Fr}}$	05/08/23	$\{X^{\text{Fr}}_{t-1}\}$; $\{X^{\text{Fr}}_{t-1,t-2,t-3}\}$
$\mathbf{X}^{\text{Mono}}$	02/01/23	$\{X^{\text{Mono}}_{t-1}\}$; $\{X^{\text{Mono}}_{t-1,t-2}, X^{\text{Mont}}_{t-3}\}$
$\mathbf{X}^{\text{Mont}}$	04/04/23	$\{X^{\text{Mont}}_{t-1}\}$;$\{X^{\text{Mont}}_{t-1}\}$

of PM_{10} (coarse particles with a diameter between 2.5 and 10 micrometers) in the air. The 3-variate time series data records the hourly concentration of PM_{10} across three counties in California—Fresno, Mono and Monterey—from Jan to June 2023. There are a total of 4305 samples. Let $\mathbf{X}^{\text{Fr}}$, $\mathbf{X}^{\text{Mono}}$ and $\mathbf{X}^{\text{Mont}}$ denote the indicators of PM10 exceeding 10 across the three counties.

Using Causal-RuLSIF, the change point estimators $\widetilde{T}_c^j$ for each $\mathbf{X}^j$ where $j \in [3]$ are shown in the Table 1, along with the parent sets before and after the estimated change point.

Based on the results, assuming there is one change point in PM10 concentration in Fresno, Mono, and Monterey during the first half of 2023, the causal mechanism of PM_{10} in Fresno is likely to shift on May 8, 2023. Additionally, while the PM_{10} levels in Fresno and Monterey are not influenced by other counties, a causal link from Monterey to Mono has emerged after February 2, 2023.

Without the ground truth and relevant knowledge about air pollution and other climate-related information, it is difficult to determine the significance of the case study results. We hope this real data application can offer insights for experts in other fields on detecting change points in causal mechanisms in practice. Additional case studies can be found in Appendix G.

6 CONCLUSION

In this paper, we introduced a novel change point detection algorithm, Causal-RuLSIF, to identify significant changes in causal mechanisms for discrete-valued time series data. By integrating a post-processing causal discovery stage with a novel dynamic divergence estimation, our algorithm accurately detects when causal mechanism shifts occur without imposing constraints on the form of the shift. We provide a theoretical uncertainty analysis of the change point estimator. Our empirical evaluation demonstrates the consistency and robustness of the proposed algorithm. The limitations of the algorithm are discussed in Appendix H.

7 ACKNOWLEDGEMENTS

This research has been supported in part by NSF CAREER 2239375, IIS 2348717, Amazon Research Award and Adobe Research. We sincerely thank the anonymous reviewers for their insightful and constructive feedback, which greatly improved the quality of this manuscript.

References

Sofiane Brahim-Belhouari and Amine Bermak. Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis*, 47(4):705–712, 2004.

Ian Cleland, Manhyung Han, Chris Nugent, Hosung Lee, Sally McClean, Shuai Zhang, and Sungyoung Lee. Evaluation of prompted annotation of activity data recorded from a smart phone. *Sensors*, 14(9): 15861–15879, 2014.

Richard J Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: a systematic survey. *IEEE transactions on image processing*, 14(3):294–307, 2005.

Makram Talih and Nicolas Hengartner. Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3):321–341, 2005.

Supreeth P Shashikumar, Matthew D Stanley, Ismail Sadiq, Qiao Li, Andre Holder, Gari D Clifford, and Shamim Nemati. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *Journal of electrocardiology*, 50 (6):739–743, 2017.

Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1): 711, 2021.

Ali Serhan Koyuncugil and Nermin Ozgulbas. Financial early warning system model and data mining application for risk detection. *Expert systems with Applications*, 39(6):6238–6253, 2012.

Tomas Kliestik, Maria Misankova, Katarina Valaskova, and Lucia Svabova. Bankruptcy prevention: new effort to reflect on legal and social changes. *Science and Engineering Ethics*, 24:791–803, 2018.

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

Zaid Harchaoui, Eric Moulines, and Francis Bach. Kernel change-point analysis. *Advances in neural information processing systems*, 21, 2008.

Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11), 2020.

Siqi Liu, Adam Wright, and Milos Hauskrecht. Change-point detection method for clinical decision support system rule monitoring. *Artificial intelligence in medicine*, 91:49–56, 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5):1324–1370, 2013.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

Zaid Harchaoui, Félicien Vallet, Alexandre Lung-Yut-Fong, and Olivier Cappé. A regularized kernel-based approach to unsupervised audio segmentation. In *2009 IEEE international conference on acoustics, speech and signal processing*, pages 1665–1668. IEEE, 2009.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of econometrics*, 86(2):221–241, 1998.

Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. An online algorithm for segmenting time series. In *Proceedings 2001 IEEE international conference on data mining*, pages 289–296. IEEE, 2001.

Naoki Itoh and Jürgen Kurths. Change-point detection of climate time series by nonparametric method. In *Proceedings of the world congress on engineering and computer science*, volume 1, pages 445–448, 2010.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2):475–507, 2015.

Matteo Barigozzi, Haeran Cho, and Piotr Fryzlewicz. Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 206(1):187–225, 2018.

Solt Kovács, Peter Bühlmann, Housen Li, and Axel Munk. Seeded binary segmentation: a general

methodology for fast and optimal changepoint detection. *Biometrika*, 110(1):249–256, 2023.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and computing*, 27:519–533, 2017.

Abolfazl Safikhani and Ali Shojaie. Joint structural break detection and parameter estimation in high-dimensional nonstationary var models. *Journal of the American Statistical Association*, 117(537):251–264, 2022.

Huida Qiu, Yan Liu, Niranjan A Subrahmanya, and Weichang Li. Granger causality for time-series anomaly detection. In *2012 IEEE 12th international conference on data mining*, pages 1074–1079. IEEE, 2012.

Jean-Marc Bardet, William Chakry Kengne, and Olivier Wintenberger. Detecting multiple change-points in general causal time series using penalized quasi-likelihood. *arXiv preprint arXiv:1008.0054*, 2010.

Mamadou Lamine Diop and William Kengne. Epidemic change-point detection in general causal time series. *Statistics & Probability Letters*, 184:109416, 2022.

Shimeng Huang, Jonas Peters, and Niklas Pfister. Causal change point detection and localization. *arXiv preprint arXiv:2403.12677*, 2024.

Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan A. Rossi, and Murat Kocaoglu. Causal discovery in semi-stationary time series. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dYeUvLUxBQ.

Murat Kocaoglu. Characterization and learning of causal graphs with small conditioning sets. *Advances in Neural Information Processing Systems*, 36, 2024.

Lyudmila Yur'evna Vostrikova. Detecting "disorder" in multidimensional random processes. In *Doklady akademii nauk*, volume 259, pages 270–274. Russian Academy of Sciences, 1981.

Malte Londschien, Peter Bühlmann, and Solt Kovács. Random forests for change point detection. *Journal of Machine Learning Research*, 24(216):1–45, 2023.

Nicholas A James and David S Matteson. ecp: An r package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*, 2013.

Wenyu Zhang, Nicholas A James, and David S Matteson. Pruning and nonparametric multiple change point detection. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 288–295. IEEE, 2017.

Judea Pearl. Causality: models, reasoning, and inference, 1980.

David E Allen, Michael McAleer, Robert J Powell, and Abhay K Singh. Non-parametric multiple change point analysis of the global financial crisis. *Annals of Financial Economics*, 13(02):1850008, 2018.

Hao Chen and Lynna Chu. Graph-based change-point analysis. *Annual Review of Statistics and Its Application*, 10:475–499, 2023.

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

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

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

Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*, pages 1898–1906. PMLR, 2015.

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

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *International conference on machine learning*, pages 2901–2910. PMLR, 2019.

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

Daigo Fujiwara, Kazuki Koyama, Keisuke Kiritoshi, Tomomi Okawachi, Tomonori Izumitani, and Shohei

Shimizu. Causal discovery for non-stationary non-linear time series data using just-in-time modeling. In *2nd Conference on Causal Learning and Reasoning*, 2023.

Arwa Alanqary, Abdullah Alomar, and Devavrat Shah. Change point detection via multivariate singular spectrum analysis. *Advances in Neural Information Processing Systems*, 34:23218–23230, 2021.

CHECKLIST

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable]
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable]

Causal Discovery-Driven Change Point Detection in Time Series: Supplementary Materials

Appendix Outline

In Section A, Section A.1 provides a toy example of time series segments. Section A.2 introduces the proposed dynamic relative PE divergence and explains our rationale for modifying relative PE divergence in our algorithm. In Section A.3, assumptions are stated. Section A.4 contains detailed proofs of theorems. Section B discusses related work on causal discovery methods.

Details of the simulated time series generation process can be found in Section C. Additional experimental results, continuing from the experiment section in the main paper, are presented in Section D. Section E addresses sample efficiency issues in practice, and Section F extends our algorithm to handle multiple change point cases. A cautionary case study demonstrating the necessary requirements of our algorithm is provided in Section G. Finally, the limitations are discussed in Section H.

A Preliminaries

A.1 Time Series Segments

In Fig. 1a), for $\mathbf{X}^1$, $\text{SPA}(X_t^1) = \{X_{t-1}^1, X_{t-3}^1, X_{t-2}^2, X_{t-1}^3\}$. Assuming V is binary time series with $D = [0, 1]$, we have:

$$\Sigma^1 \coloneqq \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \cdots \\ s^{|\text{SPA}(X_t^j)|} \end{array} \begin{array}{c} \begin{array}{cccc} X_{t-1}^1 & X_{t-3}^1 & X_{t-2}^2 & X_{t-1}^3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 \end{bmatrix} \\ s^{|\text{SPA}(X_t^j)|} \times |\text{SPA}(X_t^j)| \end{array} \tag{19}$$

With $s = |D| = 2$ and $|\text{SPA}(X_t^j)| = 4$, there are $16 = 2^4$ configurations of $\text{SPA}(X_t^j)$. Hence there are 16 time series segments of $\mathbf{X}^1$. More specifically, $X^1(1) \coloneqq \{X_t^j : t \in [\tau_{\max} + 1, T], \text{pa}(X_t^j) = [0, 0, 0, 0]\}$, $X^1(2) \coloneqq \{X_t^j : t \in [\tau_{\max} + 1, T], \text{pa}(X_t^j) = [0, 1, 0, 0]\}$ and so on. For $\mathbf{X}^3$, $\text{SPA}(X_t^3) = \{X_{t-1}^1, X_{t-1}^2, X_{t-1}^3\}$ and hence there are total $2^3 = 8$ configurations. Fig.1b) shows the construction of $X^3(1)$ and Fig.1c) illustrates 8 total time series segments of $\mathbf{X}^3$.

A.2 RuLSIF: Robust Distribution Comparison

Given two distributions $p(\cdot)$ and $p'(\cdot)$ defined over the same support, there exist many metrics measuring the distance between them. In Yamada et al. 2013, proposed a method named RuLSIF, with an α-relative divergence estimation $r_\alpha(x)$ and a corresponding metric, α-relative Pearson Divergence PE_α defined as following:

$$r_\alpha(x) \coloneqq \frac{p(x)}{(1-\alpha)p(x) + \alpha p'(x)} \coloneqq \frac{p(x)}{q_\alpha(x)} \tag{20}$$

$$PE_\alpha \coloneqq \frac{1}{2}\mathbb{E}_{x \sim q_\alpha}[(r_\alpha(x) - 1)^2] \tag{21}$$

where α is a parameter used to bound the value of $r_\alpha(x)$. With IID samples, we can obtain a direct density-ratio estimator $\hat{r}_\alpha(x)$ using a kernel function, by minimizing a squared loss function. The estimator has been proven to have a non-parametric convergence speed. In Liu et al., 2013, the method was applied to tackle the change point

detection problem in time series using sliding windows. By dividing the time series into retrospective segments, a sequence of PE divergence estimated by the RuLSIF method is obtained by assessing the distribution divergence between samples from two consecutive segments. Peaks in the divergence score PE_α can indicate change points within the joint distribution. The retrospective segments in Liu et al., 2013 are high-dimensional, even for univariate time series V.

In our proposed method, we also utilize sliding windows. However, since the focus shifts from the joint distribution to the causal mechanism, we do not construct high-dimensional retrospective segments from V. Instead, we create one-dimensional consecutive segments on the time series segments for each $\mathbf{X}^j \subsetneq V$. This approach avoids the Curse of dimensionality and enhances the method's robustness to the hyperparameters in the kernel functions, as will be verified in the experiment section.

Since we construct time series segments in our framework, the samples are IID. The distributions that need to be compared are formalized as follows:

$$p(X^j_{t_1}|\text{spa}(X^j_{t_1}) = \Sigma_\Lambda) \text{ vs } p(X^j_{t_2}|\text{spa}(X^j_{t_2}) = \Sigma_\Lambda) \tag{22}$$

where $\Lambda \in [s^{|\text{SPA}(X^j_t)|}]$, $t_1 < T_c$ and $t_2 \geq T_c$.

To simplify matters, we use $p(x)$ to denote $p(X^j_{t_1}|\text{spa}(X^j_{t_1}) = \Sigma_\Lambda)$ and $p'(x)$ represent $p(X^j_{t_2}|\text{spa}(X^j_{t_2}) = \Sigma_\Lambda)$ in the rest of the paper.

Fig.4(a) provides a toy example illustrating how the sliding window operates on time series segments. A single PE divergence score is generated using two sets of samples, one from each half of the window. As the sliding window shifts from the start to the end of a time series segment, a PE divergence series is obtained. This series is represented in each subplot in Fig. 4(b) and Fig. 4(c). Let W_i denote the ith window, where W_i^1 represents the first half and W_i^2 represents the second half.

In the ith sliding window containing one change point, without loss of generality, assume that the change point is within the second half, W_i^2, of the window. Thus, while the samples in W_i^1 come from $p(x)$, the samples in W_i^2 come from a mixture distribution $(1-\beta_i)p(x) + \beta_i p'(x)$, where β_i is the proportion of samples from $p'(x)$ within W_i^2. Note that β_i is an unknown parameter, while i is the known window index. Therefore, it is more accurate to denote the divergence score for the ith window as $PE_{\alpha\beta_i}$. To clarify, we will use $PE_{\alpha\beta_i}$ instead. By replacing $p'(x)$ with $(1-\beta_i)p(x) + \beta_i p'(x)$ in Eq.20, we obtain the $\alpha\beta_i$-relative divergence estimation $r_{\alpha\beta_i}(x)$ and the $\alpha\beta_i$-relative Pearson Divergence for the ith window:

$$r_{\alpha\beta_i}(x) \coloneqq \frac{p(x)}{(1-\alpha\beta_i)p(x) + \alpha\beta_i p'(x)} \coloneqq \frac{p(x)}{q_{\alpha\beta_i}(x)} \tag{23}$$

$$PE_{\alpha\beta_i} \coloneqq \frac{1}{2}\mathbb{E}_{x\sim q_{\alpha\beta_i}}[(r_{\alpha\beta_i}(x) - 1)^2] \tag{24}$$

and

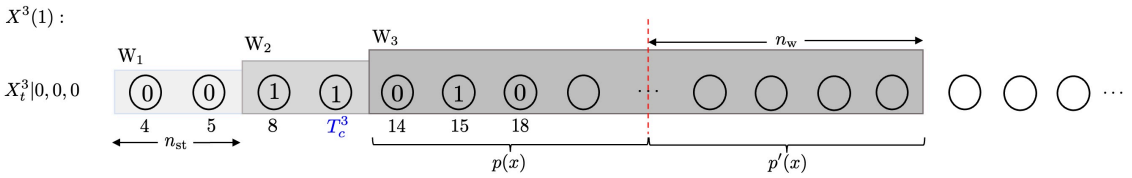
$$\beta_i \coloneqq (1 - \frac{T_c - (in_{\text{st}} + n_{\text{w}})}{n_{\text{w}}})\mathbf{1}(T_c - in_{\text{st}} \geq n_{\text{w}} \text{ and } T_c - in_{\text{st}} < 2n_{\text{w}}) \tag{25}$$

where T_{sub} is the target time series segment, n_{w} is the half window size and n_{st} is the stride size.

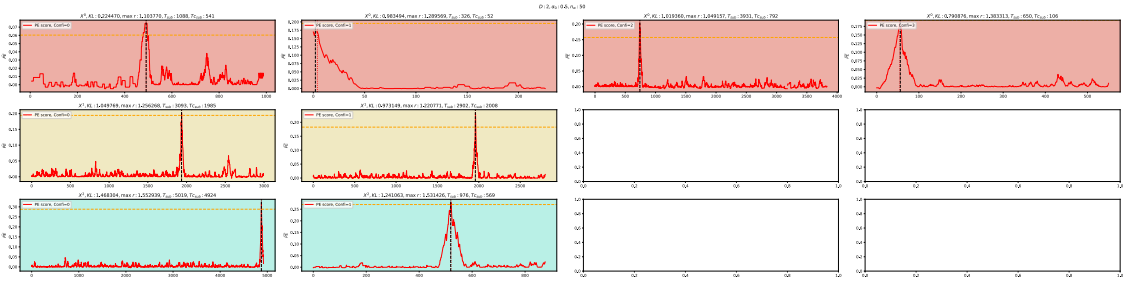
Fig. 4(b) and Fig. 4(c) display the PE divergence series for a 3-variate time series with domain sizes 2 and 3, respectively. These series were generated by applying a shifting window to each segment of the univariate time series $\mathbf{X}^j$ for $j \in [3]$. In each figure, each subplot represents the PE divergence series for a single time series segment, while the rows of subplots collectively show multiple PE divergence series for the corresponding univariate time series $\mathbf{X}^j$ for $j \in [3]$. Each PE divergence value is calculated from a window, forming the PE divergence series by sliding the window from $t = 0$ to $t = T_{\text{sub}} - 2n_{\text{w}}$.

The reasons we chose the α-relative density ratio PE as the fundamental divergence measure in our framework are:

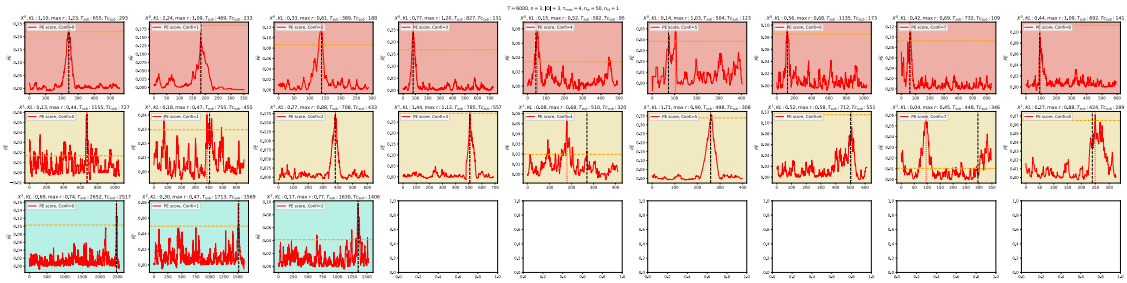
- One motivation for choosing relative PE divergence is that its asymptotic properties have been explored in Yamada et al., 2013, allowing us to directly use these properties in our theorem. In Yamada et al., 2013, the authors theoretically demonstrate that the α-relative PE divergence estimator, based on α-relative



(a) The first time series segment $X^3(1)$ from Fig.1b) for $\mathbf{X}^3$



(b) α-relative Pearson (PE) divergence series for an one arbitrary 3-variate time series with domain size 2



(c) α-relative Pearson (PE) divergence series for an one arbitrary 3-variate time series with domain size 3

Figure 4: a) The first time series segment $X^3(1)$ from Fig.1b) for $\mathbf{X}^3$. This toy example illustrates the sliding windows with n_w, n_{st} for $\{X^3(\Lambda)\}_{\Lambda\in[8]}$. b) and c) illustrate the PE divergence series under different parent configurations for an arbitrary 3-variate time series. The y-axis represents the PE score, and the x-axis represents the index i of the shifting window. The plots are organized into three rows, each corresponding to a univariate time series X^j, where $j \in [3]$. Each column represents a particular parent configuration. For instance, there are four configurations for $\mathbf{X}^1$, while $\mathbf{X}^2$ and $\mathbf{X}^3$ each have only two configurations in (b). The black vertical line indicates the true $\widehat{T}_c^j$ within that time series segment, and the red vertical line indicates the location of the maximum estimated PE score through Causal RuLSIF. The yellow horizontal line represents the true PE score.

density-ratio approximation, offers a more favorable non-parametric convergence speed than the standard density-ratio approach. Additionally, with a correctly specified parametric setup, the asymptotic variance of the proposed α-relative PE divergence estimator remains independent of model complexity. This implies that the proposed estimator resists overfitting, even when applied to complex models.

- As a squared-loss variant of KL divergence, PE divergence estimator is computationally cheaper than KL divergence because it does not involve the log term. Compared to the unbounded density ratio $r(x) = \frac{p(x)}{p'(x)}$ in PE divergence, the α-relative density ratio $r_\alpha(x) = \frac{p(x)}{\alpha p'(x)+(1-\alpha)p(x)}$ in α-relative PE divergence is always bounded above by $\frac{1}{\alpha}$. Relative PE divergence converges faster than PE divergence.

- The hyperparameter α in the α-relative PE divergence can be modified as a dynamic index in the shifting window framework, representing the 'mixture' of two distributions. The denominator in $r_\alpha(x) = \frac{p(x)}{\alpha p'(x)+(1-\alpha)p(x)}$ can be interpreted as a mixture distribution. A varying α_t can then act as a dynamic index for the "concentration" of $p'(x)$ in this mixture as the shifting window moves from $t = 0$ to $t = T$. This change allows for a transition from a static definition of α-relative PE divergence to a dynamic conditional Relative PE divergence by incorporating β_i, as defined in Eq. 25, making it more suitable for time series data.

A.3 Assumptions

A1. Sufficiency: There are no unobserved confounders.

A2. Causal Markov Condition: Each variable X is independent of all its non-descendants, given its parents $\mathrm{PA}(X)$ in $\mathcal{G}$.

A3. Faithfulness Condition [Pearl, 1980]: Let P be a probability distribution generated by $\mathcal{G}$. $\langle \mathcal{G}, P \rangle$ satisfies the Faithfulness Condition if and only if every conditional independence relation true in P is entailed by the Causal Markov Condition applied to $\mathcal{G}$.

A4. No Contemporaneous Causal Effects: Edges between variables at the same time are not allowed.

A5. Temporal Priority: Causal relations always point from the past to the future.

A6. Boundary Separation Assumption: There must be a minimum buffer period at both the beginning and the end of the time series where change points cannot be detected. More specifically, the change point for each time series $\mathbf{X}^j \in V$ cannot occur within the specified window size n_{w}.

A7. One change point per component time series/if multiple change points exist, the temporal distance between any two must be at least Δ_c: $\forall j$, $c^j = 1$, that is, each time series $\mathbf{X}^j \in V$ has only one change point/$\forall j$, if $c^j > 1$, then $\forall c \in [c^j], T_c^j - T_{c-1}^j > \Delta_c$.

A8. The number of change points is known.

A9. Minimum Pearson Divergence: For each $\mathbf{X}^j \subsetneq V$, there should exist a window W^i satisfying $PE^2_{\alpha\beta_i} > \|r_{\alpha\beta_i}\|^2_\infty + \frac{(1-\alpha\beta_i)^2\|r_{\alpha\beta_i}\|^4_\infty}{4} + \frac{\alpha^2\beta_i^2\|r_{\alpha\beta_i}\|^4_\infty}{4}$ in at least one time series segments $X^j(\Lambda)$.

Assumptions **A1-A5** are conventional and commonly employed in causal discovery methods for time series data. Our approach requires specific Assumptions **A6-A9** to be in place. To clarify, assumption **A6** is essential because our algorithm utilizes a series of sliding windows to obtain the divergence score by measuring the first half and the second half of the samples within each window. If the change point is too close to the beginning or end of the time series, the divergence score will not be significant enough to be detected. Assumption **A7** is required since the sliding windows are not directly applied on the original time series. Instead, multiple sub-time series are created and hence it is hard to impose the constraint on the minimum distance among multiple change points, such as in [Harchaoui et al., 2009], [Allen et al., 2018] and [Chen and Chu, 2023]. Assumption **A8** provides a theoretical guarantee. Assumption **A9** is necessary for the proof, as it guarantees a significant difference between two distinct causal mechanisms. This assumption is crucial because detecting the change point successfully becomes highly unlikely if the two causal mechanisms are extremely similar.

A.4 Theoretical Guarantees

Theorem A.1 ensures that in the initial step of our algorithm, a superset of the true union parent set $\mathrm{SPA}(X_t^j)$ can be obtained for all $j \in [n]$. This guarantees the correct construction of time series segments $X^j(\Lambda)$, with IID samples. Theorem A.2 establishes that if the window W_i contains samples from a single conditional distribution, the estimated relative Pearson Divergence $\widehat{PE}_{\alpha\beta_i}$ is close to zero with high probability. The lemma A.3 states that, as n_{w} increases, the estimated relative Pearson Divergence will be close to the true relative Pearson Divergence up to some constant with high probability. The lemma A.4 shows that the relative Pearson Divergence will achieve maximum if all the samples in the first half window are from one distribution denoted by p and the samples in the second half window are all from another distribution denoted by p'. Theorem A.5 establishes a confidence interval for the change point estimator $\widehat{T}_c^j$.

Theorem A.1. *Let $\mathrm{SPA}(X_t^j)$ denote the union parent set of X_t^j and $\widehat{\mathrm{SPA}}(X_t^j)$ denote the estimated union parent set obtained from PCMCI algorithm on time series $\mathbf{X}^j$ with a Mechanism-Shift SCM, and the change point T_c^j satisfies $T_c^j = \frac{T}{2}$. Under assumptions **A1-A5,A7** and with an oracle (infinite sample size limit), we have that:*

$$\mathrm{SPA}(X_t^j) \subseteq \widehat{\mathrm{SPA}}(X_t^j) \tag{26}$$

Theorem A.1 asserts that when samples are balanced, with half originating from one causal mechanism and the remaining half from another, the estimated union parent set encompasses the true union parent set.

Proof. The proof follows the same logic as the Lemma 3.2 and 3.3 in Gao et al., 2023. In the semi-stationary SCM, the samples are from multiple causal mechanisms and due to periodicity, heterogeneous samples from different causal mechanisms are balanced. However, in the mechanism-shift SCM, as per Assumption **A7**, there are only two causal mechanisms, and the samples are inherently unbalanced without specific clarification. With additional assumption $T_c^j = \frac{T}{2}$, we can draw the same conclusion as in Gao et al., 2023. □

Note that $\widehat{PE}^{j,\Lambda}_{\alpha\beta_i}$ series with parameter α and window index i is a function of n_{w}, n_{st}, time series index j and time series segments index Λ, shown as Eq. 23-25. For simplicity, we use $\widehat{PE}_{\alpha\beta_i}$ instead of $\widehat{PE}^{j,\Lambda}_{\alpha\beta_i}(n_{\text{w}}, n_{\text{st}})$ in the next section. Furthermore, we need to emphasize that T_{sub} is not the length of V but the length of the corresponding specific time series segment $X^j(\Lambda)$.

Theorem A.2. *Let $\{\widehat{PE}_{\alpha\beta_i}\}_i$ be the estimated PE series for one time series segments $X^j(\Lambda) \subsetneq \mathbf{X}^j \subsetneq V$ and T_c^j denote the true change point in this time series segments. Under assumption **A6-A7**, we have that $\forall i \in \{i : in_{\text{st}} + 2n_{\text{w}} - 1 < T_c^j\}$*

$$P\big(\max_i\{\widehat{PE}_{\alpha\beta_i}\}_i < o_p(1)\big) > 1 - \frac{a_{\text{w}} - 2}{b_{\text{st}} \log T_{\text{sub}}} - \frac{a_{\text{w}}}{T_{\text{sub}}} \tag{27}$$

where $b_{\text{st}} = \lfloor \frac{\log T_{\text{sub}}}{n_{\text{st}}} \rfloor$ and $a_{\text{w}} = \lceil \frac{T_{\text{sub}}}{n_{\text{w}}} \rceil$.

The window index i satisfying $in_{\text{st}} + 2n_{\text{w}} - 1 < T_c$ guarantees that all the samples in W_i are collected from the same distribution. Theorem A.2 states that the maximum estimated PE score obtained from such windows are bounded by any positive constant with probability $1 - \frac{a_{\text{w}}-2}{b_{\text{st}} \log T_{\text{sub}}} - \frac{a_{\text{w}}}{T_{\text{sub}}}$ if n_{w} are larger than some threshold. In other words, $\forall k > 0, \exists N$ such that $\forall n_{\text{w}} \geq N$:

$$P(\max_i\{\widehat{PE}_{\alpha\beta_i}\}_i < k) > 1 - \frac{a_{\text{w}} - 2}{b_{\text{st}} \log T_{\text{sub}}} - \frac{a_{\text{w}}}{T_{\text{sub}}} \tag{28}$$

Proof. $\forall i \in \{i : in_{\text{st}} + 2n_{\text{w}} - 1 < T_c^j\}$, the asymptotic expectation and variance of $\widehat{PE}_{\alpha\beta_i}$ is given by:

$$\mathbb{E}(\widehat{PE}_{\alpha\beta_i}) = PE_{\alpha\beta_i} + o_p(\frac{1}{\sqrt{n_{\text{w}}}}) \tag{29}$$

$$= o_p(\frac{1}{\sqrt{n_{\text{w}}}}) \tag{30}$$

$$\mathbb{V}(\widehat{PE}_{\alpha\beta_i}) = o_p(\frac{1}{n_{\text{w}}}) \tag{31}$$

The proof of the asymptotic expectation and variance can be found in Section B of Yamada et al., 2013. By Chebyshev's inequality, we have:

$$p\big(|\widehat{PE}_{\alpha\beta_i}| \geq k o_p(\frac{1}{\sqrt{n_{\text{w}}}})\big) \leq \frac{1}{k^2} \tag{32}$$

Denote A_i as the event that $|\widehat{PE}_{\alpha\beta_i}| \geq k o_p(\frac{1}{\sqrt{n_{\text{w}}}})$. By Boole's inequality, the union bound of the series for $i \in \{i : in_{\text{st}} + 2n_{\text{w}} - 1 < T_c^j\}$ is given by:

$$P\big(\cup_{i=0}^{\lfloor \frac{T_c^j - 2n_{\text{w}} + 1}{n_{\text{st}}} \rfloor} A_i\big) \leq \sum_{i=0}^{\lfloor \frac{T_c^j - 2n_{\text{w}} + 1}{n_{\text{st}}} \rfloor} p\big(A_i\big) \leq \frac{\lfloor \frac{T_c^j - 2n_{\text{w}} + 1}{n_{\text{st}}} \rfloor + 1}{k^2} \tag{33}$$

$$P\big(\cap_{i=0}^{\lfloor \frac{T_c^j - 2n_{\text{w}} + 1}{n_{\text{st}}} \rfloor} A_i^c\big) \geq 1 - \frac{\lfloor \frac{T_c^j - 2n_{\text{w}} + 1}{n_{\text{st}}} \rfloor + 1}{k^2} \tag{34}$$

That is, the probability that the maximum value of $\{\widehat{PE}_{\alpha\beta_i}\}_{i=0}^{in_{\text{st}}+2n_{\text{w}}-1<T_c^j}$ is less than $ko_p(\frac{1}{\sqrt{n_{\text{w}}}})$ is larger than $1-\frac{\lfloor\frac{T_c^j-2n_{\text{w}}+1}{n_{\text{st}}}\rfloor+1}{k^2}$. The same results hold for $\forall i\in\{i: in_{\text{st}}\geq T_c^j\}$.

Let $n_{\text{st}}=b_{\text{st}}\log T_{\text{sub}}$, $k^2=n_{\text{w}}$ and make n_{w} a proportion of T_{sub}, which can be denoted by $\frac{T_{\text{sub}}}{a_{\text{w}}}$ where b_{st} and a_s are both constants.

We have:

$$P\big(\cap_{i=0}^{\lfloor\frac{T_c^j-\frac{2T_{\text{sub}}}{a_{\text{w}}}+1}{b_{\text{st}}\log T_{\text{sub}}}\rfloor}A_i^c\big)>1-\frac{a_{\text{w}}-2}{b_{\text{st}}\log T_{\text{sub}}}-\frac{a_{\text{w}}}{T_{\text{sub}}} \tag{35}$$

In other words, the likelihood that the maximum $PE_{\alpha\beta_i}$, encompassing all samples before T_c^j or after T_c^j, is less than $o(1)$ is greater than $1-\frac{a_{\text{w}}-2}{b_{\text{st}}\log T_{\text{sub}}}-\frac{a_{\text{w}}}{T_{\text{sub}}}$. More specifically:

$$P\big(\max_i\{\widehat{PE}_{\alpha\beta_i}\}_i<o_p(1)\big)>1-\frac{a_{\text{w}}-2}{b_{\text{st}}\log T_{\text{sub}}}-\frac{a_{\text{w}}}{T_{\text{sub}}} \tag{36}$$

□

Lemma A.3. *Let $\{\widehat{PE}_{\alpha\beta_i}\}_i$ be the estimated PE series for one time series segments $X^j(\Lambda)\subsetneq\mathbf{X}^j\subsetneq V$ and T_c^j denote the true change point in this time series segments. Under assumption **A6-A7**, we have that $\forall i\in\{i: T_c^j<in_{\text{st}}+2n_{\text{w}}-1<T_c^j+n_{\text{w}}\}$,*

$$p\big(PE_{\alpha\beta_i}-c_i-(\sqrt{n_{\text{w}}}-1)o_p(\frac{1}{\sqrt{n_{\text{w}}}})\leq\widehat{PE}_{\alpha\beta_i} \tag{37}$$

$$\leq PE_{\alpha\beta_i}+c_i+(\sqrt{n_{\text{w}}}+1)o_p(\frac{1}{\sqrt{n_{\text{w}}}})\big)>1-\frac{1}{n_{\text{w}}} \tag{38}$$

where $c_i^2:=\|r_{\alpha\beta_i}\|_\infty^2+\frac{(1-\alpha\beta_i)^2\|r_{\alpha\beta_i}\|_\infty^4}{4}+\frac{\alpha^2\beta_i^2\|r_{\alpha\beta_i}\|_\infty^4}{4}$.

The window index i satisfying $T_c^j<in_{\text{st}}+2n_{\text{w}}-1<T_c^j+n_{\text{w}}$ guarantees that all the samples in W_i^1 are collected from $p(x)$ and samples in W_2^1 are from a mixture distribution $(1-\beta_i)p(x)+\beta_i p'(x)$. Lemma A.3 states that the estimated error between the estimated PE series and the true PE value is smaller than a constant with probability $1-\frac{1}{n_{\text{w}}}$ if n_{w} are larger than some threshold. In other words, $\forall k>0,\exists N$ such that $\forall n_{\text{w}}\geq N$:

$$P(|\widehat{PE}_{\alpha\beta_i}-PE_{\alpha\beta_i}|<k+c_i)>1-\frac{1}{n_{\text{w}}} \tag{39}$$

Proof. $\forall i\in\{i: T_c^j<in_s+2n_{\text{w}}-1<T_c^j+n_{\text{w}}\}$, the asymptotic expectation and variance of $\widehat{PE}_{\alpha\beta_i}$ is given by:

$$\mathbb{E}(\widehat{PE}_{\alpha\beta_i})=PE_{\alpha\beta_i}+o_p(\frac{1}{\sqrt{n_{\text{w}}}}) \tag{40}$$

$$\mathbb{V}(\widehat{PE}_{\alpha\beta_i})\leq\frac{\|r_{\alpha\beta_i}\|_\infty^2}{n_{\text{w}}}+\frac{(1-\alpha\beta_i)^2\|r_{\alpha\beta_i}\|_\infty^4}{4n_{\text{w}}}+\frac{\alpha^2\beta_i^2\|r_{\alpha\beta_i}\|_\infty^4}{4n_{\text{w}}}+o_p(\frac{1}{n_{\text{w}}}). \tag{41}$$

For simplicity, let $\sigma_i^2:=\frac{\|r_{\alpha\beta_i}\|_\infty^2}{n_{\text{w}}}+\frac{(1-\alpha\beta_i)^2\|r_{\alpha\beta_i}\|_\infty^4}{4n_{\text{w}}}+\frac{\alpha^2\beta_i^2\|r_{\alpha\beta_i}\|_\infty^4}{4n_{\text{w}}}$.

By Chebyshev's inequality, we have:

$$p\big(|\widehat{PE}_{\alpha\beta_i}-PE_{\alpha\beta_i}-o_p(\frac{1}{\sqrt{n_{\text{w}}}})|\geq k(\sigma_i+o_p(\frac{1}{\sqrt{n_{\text{w}}}}))\big)\leq\frac{1}{k^2} \tag{42}$$

$$p\big(|\widehat{PE}_{\alpha\beta_i}-PE_{\alpha\beta_i}-o_p(\frac{1}{\sqrt{n_{\text{w}}}})|<k(\sigma_i+o_p(\frac{1}{\sqrt{n_{\text{w}}}}))\big)\geq1-\frac{1}{k^2} \tag{43}$$

Hence we have:

$$p\big(PE_{\alpha\beta_i}-k\sigma_i-(k-1)o_p(\frac{1}{\sqrt{n_{\text{w}}}})\leq\widehat{PE}_{\alpha\beta_i}\leq PE_{\alpha\beta_i}+k\sigma_i+(k+1)o_p(\frac{1}{\sqrt{n_{\text{w}}}})\big)>1-\frac{1}{k^2} \tag{44}$$

Let $k = \sqrt{n_{\mathrm{w}}}$ and make n_{w} a proportion of T, which can be denoted by $\frac{T}{a_w}$ where a_s is a constant.

We have:

$$p\big(PE_{\alpha\beta_i} - c_i - (\sqrt{n_{\mathrm{w}}} - 1)o_p(\frac{1}{\sqrt{n_{\mathrm{w}}}}) \leq \widehat{PE}_{\alpha\beta_i} \leq PE_{\alpha\beta_i} + c_i + (\sqrt{n_{\mathrm{w}}} + 1)o_p(\frac{1}{\sqrt{n_{\mathrm{w}}}})\big) \tag{45}$$

$$> 1 - \frac{1}{n_{\mathrm{w}}} \tag{46}$$

where $c_i^2 \coloneqq \|r_{\alpha\beta_i}\|_\infty^2 + \frac{(1-\alpha\beta_i)^2\|r_{\alpha\beta_i}\|_\infty^4}{4} + \frac{\alpha^2\beta_i^2\|r_{\alpha\beta_i}\|_\infty^4}{4}$.

□

Since we only discuss the discrete distribution, therefore we can assume that there are k realizations of any variable X_t^j from 1 to k. For simplicity, we denote $p(x = k)$ as p_h and $p'(x = k)$ as p'_h.

Lemma A.4. *Let $\{PE_{\alpha\beta_i}\}_i$ be the estimated PE series for one time series segments $X^j(\Lambda) \subsetneq \mathbf{X}^j \subsetneq V$ and T_c^j denote the true change point in this time series segments. Under assumption **A6-A7**, we have that $\forall i \in \{i : T_c^j < in_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c^j + n_{\mathrm{w}}\}$,*

- $PE_{\alpha\beta_i} > 0$
- *$PE_{\alpha\beta_i}$ is a monotonically increasing function regarding i (or β_i).*
- $\max PE_{\alpha\beta_i} = \frac{1}{2}\sum_{h=1}^{k} \frac{p_h^2}{(1-\alpha\beta^*)p_h + \alpha\beta^* p'_h} - \frac{1}{2}$ *where β^* is the largest proportion of samples from $p'(x)$ over those second windows $\{W_i^2\}_{i:T_c^j < in_{\mathrm{st}}+2n_{\mathrm{w}}-1<T_c^j+n_{\mathrm{w}}}$. For appropriate n_{w} and n_{st}, $\max PE_{\alpha\beta_i}$ can achieve the maximum when all samples of W_i^1 are from $p(x)$ and all samples of W_i^2 are from $p'(x)$, that is, $\max_i \beta_i = 1$.*

Lemma A.4 suggests that the true PE series derived from the sequence of sliding windows is a monotonically increasing function with respect to β_i if the samples in the second half of the sliding windows are from a mixture distribution, indicating that the change point is located within the second half of the sliding windows. Similarly, employing the same reasoning, the true PE series becomes a monotonically decreasing function with respect to β_i if the samples in the first half of the windows are from a mixture distribution, that is, $i \in \{i : T_c^j + n_{\mathrm{w}} < in_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c^j + 2n_{\mathrm{w}}\}$. In that case, β_i represents the proportion of samples from $p(x)$ over the first half windows.

Proof. Treat $PE_{\alpha\beta_i}$ as a function of $\alpha \coloneqq \alpha\beta_i$.

$$PE_{\alpha\beta_i} = \frac{1}{2}\mathbb{E}_{p(x)}[r_{\alpha\beta_i}(x)] - \frac{1}{2} \tag{47}$$

$$= \frac{1}{2}\sum_{h=1}^{k} \frac{p_h}{(1-\alpha)p_h + \alpha p'_h} p_h - \frac{1}{2} \tag{48}$$

$$= \frac{1}{2}\sum_{h=1}^{k} \frac{p_h}{1 + (\frac{p'_h}{p_h} - 1)\alpha} - \frac{1}{2} \tag{49}$$

We have:

$$f(\alpha) \coloneqq \sum_{h=1}^{k} \frac{p_h}{1 + (\frac{p'_h}{p_h} - 1)\alpha} \tag{50}$$

$$f'(\alpha) = \sum_{h=1}^{k} \frac{-p_h(\frac{p'_h}{p_h} - 1)}{(1 + (\frac{p'_h}{p_h} - 1)\alpha)^2} \tag{51}$$

$$f''(\alpha) = \sum_{h=1}^{k} \frac{2p_h(\frac{p'_h}{p_h} - 1)^2}{(1 + (\frac{p'_h}{p_h} - 1)\alpha)^3} \tag{52}$$

Since $\frac{p_h'}{p_h} > 0$ and $0 \leq \alpha \leq 1$, we have $1 + (\frac{p_h'}{p_h} - 1)\alpha > 0$ for any $j \in [k]$, hence $f''(\alpha) > 0$, leading to an increasing function $f'(\alpha)$.

$$f'(\alpha) \geq f'(0) = \sum_{h=1}^{k} (p_h - p_h') = 0 \tag{53}$$

Hence $f(\alpha)$ is an increasing function in terms of α and

$$f(\alpha) \geq f(0) = \sum_{h=1}^{k} p_h = 1, \text{ with } \alpha \in [0, 1] \tag{54}$$

□

Theorem A.5. *Let $\{\widehat{PE}_{\alpha\beta_i}\}_i$ be the estimated PE series for one time series segments $X^j(\Lambda) \subsetneq \mathbf{X}^j \subsetneq V$ and T_c^j denote the true change point in this time series segments. $\widehat{T}_c^j$ denotes the estimator of T_c^j obtained by:*

$$\widehat{T}_c^j = \arg_i \max\{\widehat{PE}_{\alpha\beta_i}\}_i \tag{55}$$

*Under assumption **A6-A9**, we have that given large enough n_{w}, $\forall i \in [\tau_{max} + 1, T]$*

$$P\big(|\widehat{T}_c^j - T_c^j| < 2n_{\mathrm{w}}\big) > (1 - \frac{a_{\mathrm{w}} - 2}{b_{\mathrm{st}} \log T_{\mathrm{sub}}} - \frac{a_{\mathrm{w}}}{T_{\mathrm{sub}}})(1 - \frac{1}{n_{\mathrm{w}}}) \tag{56}$$

where $b_{\mathrm{st}} = \lfloor \frac{\log T_{\mathrm{sub}}}{n_{\mathrm{st}}} \rfloor$ and $a_{\mathrm{w}} = \lceil \frac{T_{\mathrm{sub}}}{n_{\mathrm{w}}} \rceil$.

Proof. From Theorem A.2, $\forall k_1 > 0, \exists N_1$ such that $\forall n_{\mathrm{w}} \geq N_1$, we have $\forall i_1 \in \{i_1 : i_1 n_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c\}$:

$$P\big(\max_{i_1}\{\widehat{PE}_{\alpha\beta_{i_1}}\}_{i_1} < k_1\big) > 1 - \frac{a_{\mathrm{w}} - 2}{b_{\mathrm{st}} \log T_{\mathrm{sub}}} - \frac{a_{\mathrm{w}}}{T_{\mathrm{sub}}} \tag{57}$$

From Lemma A.3, $\forall k_2 > 0, \exists N_2$ such that $\forall n_{\mathrm{w}} \geq N_2$, we have $\forall i_2 \in \{i : T_c^j < i_2 n_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c^j + n_{\mathrm{w}}\}$:

$$P(|\widehat{PE}_{\alpha\beta_{i_2}} - PE_{\alpha\beta_{i_2}}| < k_2 + c_{i_2}) > 1 - \frac{1}{n_{\mathrm{w}}} \tag{58}$$

With Assumption **A9**, $\exists i_2$ such that $PE_{\alpha\beta_{i_2}} > c_{i_2}$, hence $\exists k_1, k_2 > 0$:

$$PE_{\alpha\beta_{i_2}} \geq c_{i_2} + k_2 + k_1 \tag{59}$$

Hence we can replace k_1 with $PE_{\alpha\beta_{i_2}} - c_{i_2} - k_2$ in Eq.57 and then we have:

$$P\big(\max_{i_1}\{\widehat{PE}_{\alpha\beta_{i_1}}\}_{i_1} < PE_{\alpha\beta_{i_2}} - c_{i_2} - k_2\big) > 1 - \frac{a_{\mathrm{w}} - 2}{b_{\mathrm{st}} \log T_{\mathrm{sub}}} - \frac{a_{\mathrm{w}}}{T_{\mathrm{sub}}} \tag{60}$$

From Lemma A.3, we have:

$$PE_{\alpha\beta_{i_2}} - k_2 - c_{i_2} < \widehat{PE}_{\alpha\beta_{i_2}} < PE_{\alpha\beta_{i_2}} + k_2 + c_{i_2} \tag{61}$$

holds with probability $1 - \frac{1}{n_{\mathrm{w}}}$.

As samples in W_{i_1} and W_{i_2} are IID, the events in Eq.60 and Eq.61 are independent, resulting in:

$$P(\max_{i_1}\{\widehat{PE}_{\alpha\beta_{i_1}}\}_{i_1} < \widehat{PE}_{\alpha\beta_{i_2}}) > (1 - \frac{a_{\mathrm{w}} - 2}{b_{\mathrm{st}} \log T_{\mathrm{sub}}} - \frac{a_{\mathrm{w}}}{T_{\mathrm{sub}}})(1 - \frac{1}{n_{\mathrm{w}}}) \tag{62}$$

for $\forall i_1 \in \{i_1 : i_1 n_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c\}$ and $\exists i_2 \in \{i : T_c^j < i_2 n_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c^j + n_{\mathrm{w}}\}$.

From Lemma A.4, we know that $\exists \beta^*$ such that $PE_{\alpha\beta^*} = \max PE_{\alpha\beta_{i_2}}$. Denote $i_2^* = \arg_{i_2} \max PE_{\alpha\beta_{i_2}}$.

Since $PE_{\alpha\beta_{i_2}}$ is a monotonically increasing function regarding β_{i_2}, then the true change point T_c^j must satisfying:

$$PE_{\alpha\beta_{\frac{T_c^j - n_{\mathrm{w}}}{n_{\mathrm{st}}}}} \geq PE_{\alpha\beta^*} \geq \widehat{PE}_{\alpha\beta_{i_1}}, \tag{63}$$

$\forall i_1 \in \{i : i_1 n_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c^j\}$ as n_{st}, n_{w} and T has jointly determined β_i. The proof for situations that $\{i_1 : i_1 n_{\mathrm{st}} \geq T_c^j\}$ and $i_2 \in \{i : T_c^j + n_{\mathrm{w}} < i n_{\mathrm{st}} + 2n_{\mathrm{w}} - 1 < T_c^j + 2n_{\mathrm{w}}\}$ follows the same logic.

With Eq.63, we finally have:

$$P\big(|\widehat{T}_c^j - T_c^j| < 2n_{\mathrm{w}}\big) > (1 - \frac{a_{\mathrm{w}} - 2}{b_{\mathrm{st}} \log T_{\mathrm{sub}}} - \frac{a_{\mathrm{w}}}{T_{\mathrm{sub}}})(1 - \frac{1}{n_{\mathrm{w}}}) \tag{64}$$

□

B Related work

Identifying causal relationships in stationary time series is more challenging than in IID samples, and finding these relationships in non-stationary time series is even harder. In our scenario, when change points occur, the *Mechanism-Shift* SCM leads to a non-stationary time series. Each mechanism in this model corresponds to a stationary time series. Once we accurately detect the change points, we can divide the non-stationary time series into multiple stationary components. This simplifies the task of finding causal relationships in non-stationary time series to that of discovering them in stationary time series. Therefore, our algorithm can be viewed as a method for causal discovery in non-stationary time series, driven by change point detection.

Numerous efforts have been made recently to adapt causal discovery algorithms originally designed for IID data to work with stationary time series data, such as [Runge et al., 2019], [Entner and Hoyer, 2010], [Hyvärinen et al., 2010], [Peters et al., 2013] and [Pamfil et al., 2020].

Most causal discovery methods for non-stationary time series rely on parametric models. Examples include the vector autoregressive model used in [Gong et al., 2015] and [Malinsky and Spirtes, 2019], as well as linear and non-linear state-space models in [Huang et al., 2019]. Other approaches focus on linear causal relationships, as seen in [Saggioro et al., 2020]. A non-parametric method called CD-NOD, described in [Huang et al., 2020], can identify time-lagged causal relationships in non-stationary time series. Additionally, [Fujiwara et al., 2023] introduced the JIT-LiNGAM algorithm, which combines the LiNGAM approach with a JIT framework to create a local approximated linear causal model for non-linear and non-stationary data. In [Gao et al., 2023], it is assumed that the underlying causal mechanisms of each time series change sequentially and periodically.

We would like to thoroughly clarify the differences between the setting in [Gao et al., 2023] and our paper to prevent any potential confusion.

Our work Causal-RuLSIF and the previous work of PCMCI$_\Omega$ in [Gao et al., 2023] are designed for distinct settings. We highlight two significant differences below.

Periodicity In PCMCI$_\Omega$, the Structural Causal Model (SCM) underlying the non-stationary time series is defined for a specific type of non-stationarity called semi-stationary time series. These are characterized by a finite number of different causal mechanisms occurring periodically over time. In contrast, our work does not assume such periodicity but rather a change point in the causal mechanism.

For example, consider two different causal mechanisms, denoted as A and B. The shifting causal mechanisms over time can be expressed as $[A, B, A, B, A, B, \cdots]$. PCMCI$_\Omega$ can detect the periodicity of such changes. On the other hand, Causal-RuLSIF is designed for scenarios where a time series $\mathbf{X}^j \in V$ transitions from one causal mechanism to another without periodicity, such as $[A, A, A, \cdots, A, B, B, \ldots, B]$. Causal-RuLSIF identifies the change point where the mechanism shifts from A to B. While Causal-RuLSIF can not be applied to settings like $[A, B, A, B, A, B, \cdots]$, PCMCI$_\Omega$ also fails to handle settings like $[A, A, A, \cdots, A, B, B, \cdots, B]$.

Mechanism Change Additionally, PCMCI$_\Omega$ relies on the Hard Mechanism Change Assumption, which requires that the incoming edges of causal mechanisms A and B do not share time-invariant parent sets, reflecting a time-variant structure. In contrast, Causal-RuLSIF does not impose constraints on the causal graph structure; it directly assesses changes in the conditional distribution using a distribution divergence score. As a result, Causal-RuLSIF can handle both hard and soft mechanism changes.

C Generate Binary-valued Time Series

The generation process has three steps.

1. Determine time series length T, number of time series n of the multivariate time series, data domain D, maximum time lag $\tau_{\max}$. Randomly generate T_c^j, $j \in [n]$, satisfying Definition 3.1 and Assumption **A6-A7**. Additionally, in order to guarantee enough samples in each time series segment, we also control the size of the union parent set, that is, $|\text{SPA}X_t^j|$.

   With n, $\tau_{\max}$ and the number of change points for each component time series c^j, one binary edge array with dimension $[n, c^j + 1, n, \tau_{\max} + 1]$ is randomly generated, where the first n denotes the index of parent variable, c^j denotes the number of change point, the second n represents the index of target variable (child) and $\tau_m ax + 1$ denotes the time lag. With Assumption **A7**, $c^j = 1$ for all $j \in [n]$. 1 in this binary edge array means that there is an edge from the parent variable to the child variable with the corresponding time lag; 0 means that there is no cause-effect between two corresponding variables.

   For a soft mechanism change, the parent set remains the same before and after the change point, without intersecting other change points, so the binary edge arrays representing the different causal mechanisms should be identical. In contrast, for a hard mechanism change, at least two of these binary edge arrays must differ.

2. With the randomly generated binary edge array controlled by $|\text{SPA}X_t^j|$, a full causal graph for this n-variate time series is obtained. Given this edge array and data domain D, parent configuration matrices $\{\Sigma^j\}_{j\in[n]}$ are obtained. Corresponding to $\{\Sigma^j\}_{j\in[n]}$, conditional probability tables (CPTs) are randomly generated.

3. Fill the starting points $\{X^j_{t\le\tau_{\max}}\}_{j\in[n]}$ with randomly generated binary data. Starting from time point $t > \tau_{\max}$, generate vector $\mathbf{X}_t$ over time according to the CPTs, until t achieves T.

Please note that since our algorithm can be extended to handle multiple change points, the generation process can also be easily extended to accommodate such settings.

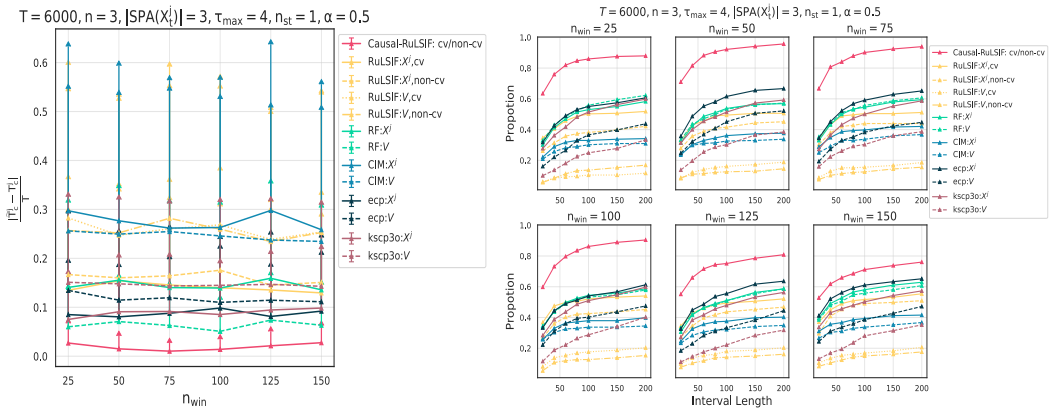
D Additional Experiments

From Fig 5(a) and Fig 5(b), it's apparent that a large n_{w} may affect the performance of Causal-RuLSIF. This issue still stems from sample efficiency. The total number of windows for a specific time series segment $X^j(\Lambda)$ is given by $\lfloor \frac{|T_{\text{sub}}|-2n_{\text{w}}}{n_{\text{st}}} \rfloor + 1$, where $T_{\text{sub}} = |X^j(\Lambda)|$. Increasing n_{w} decreases the total number of windows, resulting in less information collected, as smaller n_{w} is sufficient to estimate PE scores with kernel functions.

As demonstrated in Fig 6(a) and Fig 6(b), the performance of Causal-RuLSIF remains consistent regardless of the distance between the change points T_c^1 of $\mathbf{X}^1$ and T_c^2 of $\mathbf{X}^2$.

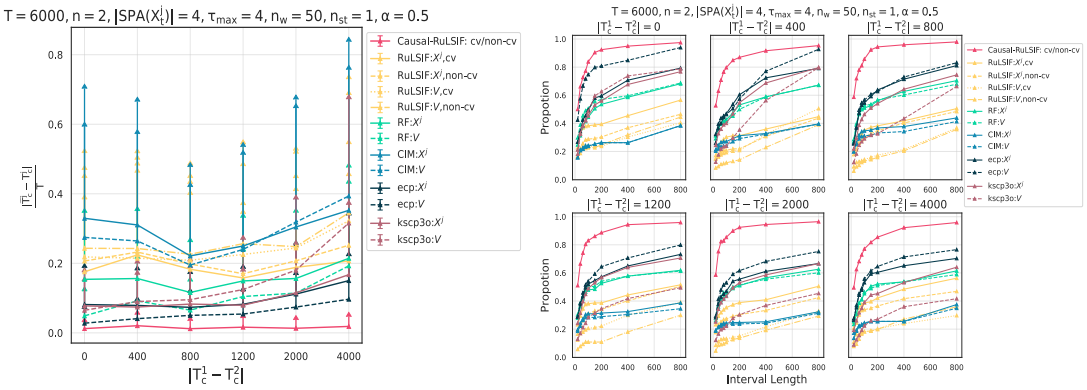
Fig. 7(a) shows the running time needed for each algorithm in seconds. Cross-validation techniques is time-consuming. It is beneficial that cross-validation is not needed for binary time series data for Causal-ReLSIF. Additionally, in order to have a better illustration of implementing dynamic divergence measurement on time series segments, we have conducted experiments on randomly generated binary time series and plot the PE divergence series obtained from the time series segments.

The estimated PE divergence series shown in Fig. 7(b) for one time series segment verifies Theorem 4.1, Lemma A.4 and Theorem 4.2 empirically. For window index i smaller than around 500 or larger than around 900, $\hat{PE}_i$ score is very close to zero. For window index i between 500 and 900, $\widehat{PE}_i$ first monotonically increases and then monotonically decreases over the window index i and the maximum value is achieved at $\widehat{T}_c = 864$, which equals to the true change point in this time series segment. Furthermore, the maximum of the PE score line is equal to the true PE value, which is denoted by the dashed orange horizontal line. In the title of the subgraphs, KL represents the Kullback-Leibler (KL) divergence between the two conditional distributions of X^0 before and after the change point, while $\max r$ indicates the maximum density ratio $\frac{p(x)}{(1-\alpha\beta_i)p(x)+\alpha\beta_i p'(x)}$ over all x in its domain.



(a) Influence of n_{w} on estimation error (b) ROC curves for different n_{w}

Figure 5: Causal-RuLSIF is tested on 3-multivariate binary time series with $T = 6000, \tau_{\max} = 4, |\mathrm{SPA}(X_t^j)| = 3, n_{\mathrm{st}} = 1$ with *hard mechanism change*. Every line with a different color corresponds to a different algorithm and different linestyle corresponds to a different setting. X^j in the legend means the algorithm is applied to each component time series while V means the algorithm is used for the whole n-variate time series V. Every marker corresponds to the average error rate or average accuracy rate over 100 random trials. The error bar represents the standard error for the averaged statistics. a) Influence of n_{w} on estimation error $\frac{|\widetilde{T}_c^j - T_c^j|}{T}$. b) ROC curves for different n_{w}.

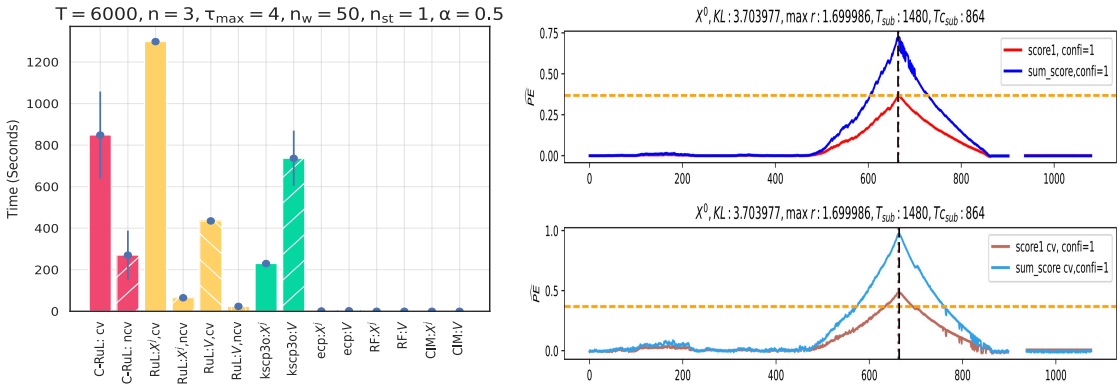


(a) Influence of $|T_c^1 - T_c^2|$ on estimation error (b) ROC curves for different $|T_c^1 - T_c^2|$

Figure 6: Causal-RuLSIF is tested on 3-multivariate binary time series with $T = 6000, \tau_{\max} = 4, |\mathrm{SPA}(X_t^j)| = 3, n_{\mathrm{w}} = 50, n_{\mathrm{st}} = 1$ with *soft mechanism change*. Every line with a different color corresponds to a different algorithm and different linestyle corresponds to a different setting. X^j in the legend means the algorithm is applied to each component time series while V means the algorithm is used for the whole n-variate time series V. Every marker corresponds to the average error rate or average accuracy rate over 100 random trials. The error bar represents the standard error for the averaged statistics. a) Influence of $|T_c^1 - T_c^2|$ on estimation error $\frac{|\widetilde{T}_c^j - T_c^j|}{T}$. b) ROC curves for different $|T_c^1 - T_c^2|$.

E Sample Efficiency: k-PC and Top-K parents

As noted in the main paper, sample efficiency is a fundamental issue in our algorithm. The effective sample size of each time series segment, $|T_{\mathrm{sub}}| \approx T/|s^{|\widehat{\mathrm{SPA}}(X_t^j)|}|$, decreases exponentially with domain size s and parent size $\widehat{\mathrm{SPA}}(X_t^j)$, whereas the baseline method has a sample size of T. This is a necessary trade-off if we focus on the **causal mechanims** shift rather than the shift in **joint distributions**.



(a) Runtime (in sec.) for algorithms

(b) One estimated PE divergence series obtained on one time series segment

Figure 7: Runtime and one illustration of time series segment.

In practice, we can address this issue through two approaches: the k-PC algorithm and the top-K causal strength selection.

Specifically, we can improve the power of the CI tests by applying the k-PC algorithm from Kocaoglu, 2024 during the PC stage of PCMCI, which restricts the size of the conditioning set to k. Furthermore, when generating time series segments with $\mathrm{SPA}(X_t^j)$, we can directly limit the size of $\mathrm{SPA}(X_t^j)$ by selecting only the parents with the top K causal strengths based on the statistics from the MCI tests in PCMCI. As noted in Runge et al., 2019, the MCI test statistic can be interpreted as a measure of causal strength, enabling the meaningful ranking of causal links in large-scale studies.

Given a large multivariate time series dataset with $n = 4, 6, 8, 10$ univariate time series, Fig. 8(a) and 8(b) demonstrate that more accurate time series segments result in more accurate estimates of $\widehat{T_c^j}$, highlighting the need for causal-driven change point detection under the *Mechanism-shift* SCM. Our proposed algorithm, Causal-RuLSIF with the Top 1 parent, may perform worse than some baselines due to sample efficiency limitations. Additionally, in this experiment, when no significant parent is identified during causal discovery, we select the parent with the highest causal influence, even if it lacks significance, which can further decrease the performance of the Top 1 parent method. However, Causal-RuLSIF with the Top 3 parents outperforms other baselines.

F Multiple Change Points

Our algorithm can be readily extended to handle multiple change point problems, subject to **Assumption A7**.

If the temporal distance between two consecutive change points in time series segments exceeds the window size, i.e., $\Delta_c > 2n_{\mathrm{w}}$, the theoretical guarantee for detecting each individual change point, as discussed in the main paper, remains valid.

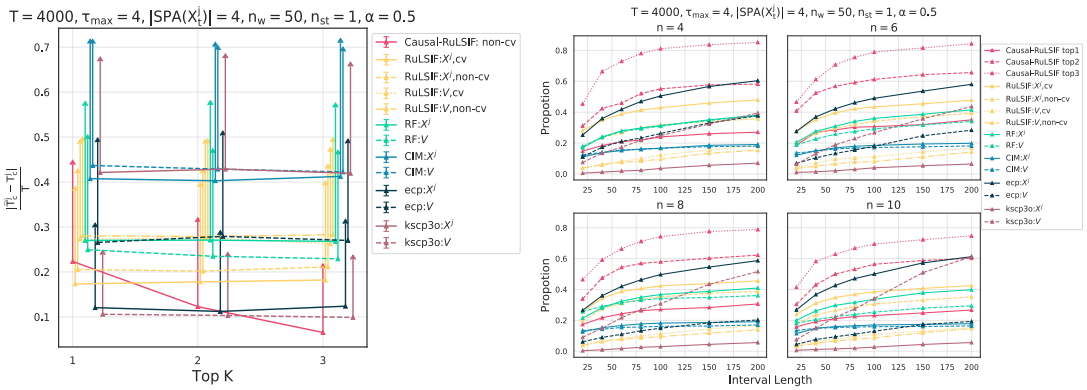
Fig. 9(a) and Fig. 9(b) illustrate the performance of algorithms on time series where each univariate time series contains at most 4 change points, specifically, $c^j \leq 4$ for $j \in [3]$. Our algorithm outperforms the other baselines.

G Case Study: A Cautionary Counterexample

When applying our causal-driven algorithm to a real dataset, it is crucial to ensure that there are causal relationships among the n-variate time series. If no causal relationships exist within the n-variate time series, we do not recommend using our algorithm. This is because the identified causal relationships may not be meaningful, and the efficiency of the sample will diminish due to the presence of misleading or meaningless causal relationships.

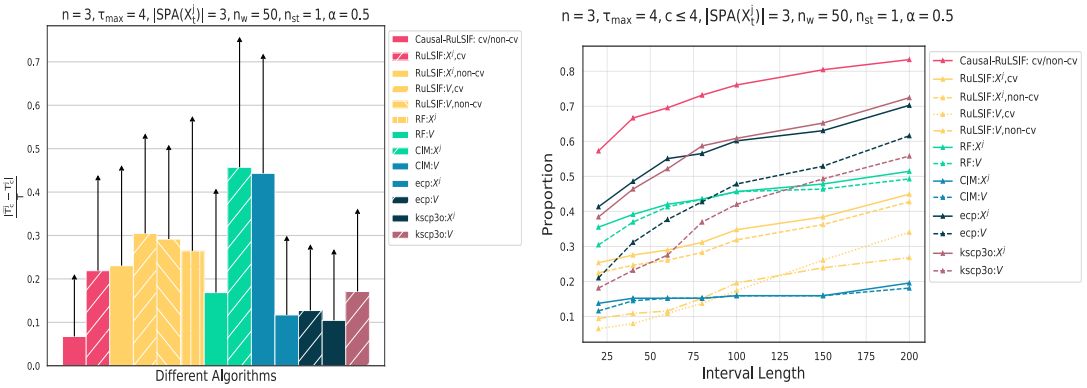
Additionally, please carefully check whether all the assumptions in section A.3 are met.

To illustrate this warning, we conduct a case study on a human activity dataset, as used in Alanqary et al., 2021.



(a) Influence of selecting top-K parents on estimation error (b) ROC curves for different n-variate time series

Figure 8: Causal-RuLSIF is tested on n-variate time series where $T = 6000, \tau_{\max} = 4, |\text{SPA}(X_t^j)| = 4, n_{\text{st}} = 1, n = [4, 6, 8, 10]$ with *soft mechanism change*. The domain size $s = 3$. Every line with a different color corresponds to a different algorithm and different linestyle corresponds to a different setting. X^j in the legend means the algorithm is applied to each component time series while V means the algorithm is used for the whole n-variate time series V. Every marker corresponds to the average error rate or accuracy rate over 50 random trials. The error bar represents the standard error for the averaged statistics. a) Influence of Top-K parents on estimation error $\frac{|\widetilde{T}_c^j - T_c^j|}{T}$. b) ROC curves for different n-variate time series.



(a) Estimation error (b) ROC curves

Figure 9: Causal-RuLSIF is tested on 3-multivariate binary time series with $T = 6000, \tau_{\max} = 4, |\text{SPA}(X_t^j)| = 3, n_{\text{w}} = 50, n_{\text{st}} = 1$ with *soft mechanism change*. Every line with a different color corresponds to a different algorithm and different linestyle corresponds to a different setting. X^j in the legend means the algorithm is applied to each component time series while V means the algorithm is used for the whole n-variate time series V. Every marker corresponds to the average error rate or average accuracy rate over 50 random trials. The error bar represents the standard error for the averaged statistics. a) Estimation error with different algorithms. b) ROC curves for different algorithms.

This dataset was collected using a portable three-axis accelerometer on human beings. The three dimensions of the time series reflect the acceleration recorded by the device along the x, y, and z axes. The change points indicate transitions between six activities: standing, walking, jogging, skipping, climbing up stairs, and climbing down stairs. We select a sequence of length 1000 that contains only one change point.

First, common sense suggests that there are no clear causal relationships among this 3-variate time series. Second, **Assumption A1** is violated since the three axes are affected by human movement, which implies the existence of an unobserved confounder.

We continue to apply causal-RuLSIF to this dataset to highlight its inappropriateness. As illustrated in Fig. 10, while the estimated change point $\widetilde{T}_c^j$ on the y-axis and z-axis is close to the true change point T_c^j, this does not confirm the accuracy of the estimates; instead, it demonstrates the robustness of the causal-driven algorithm. Additionally, the deviation of $\widetilde{T}_c^j$ on the x-axis is consistent with our previous observations. In this type of dataset, we should concentrate on the shift in the **joint distribution** of the time series rather than the **causal mechanism** shift.

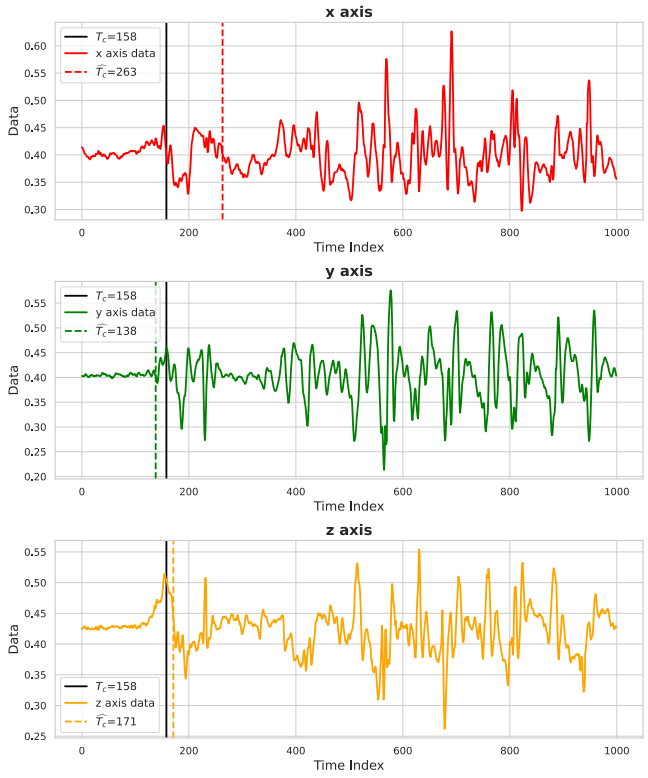


Figure 10: Causal-RuLSIF on human activity dataset. There are three subplots corresponding to the x-axis, y-axis, and z-axis, respectively. The black vertical line represents the true change point T_c^j, while each colored vertical line indicates the estimated change point $\widetilde{T}_c^j$.

H Limitations

Here we discuss the limitations of our proposed algorithm:

Discrete-valued time series with infinite domain size and Continuous-valued time series: If the domain size for the discrete data is large, a longer time series is required. Furthermore, the number of samples in each time series segment decreases exponentially as the domain size increases. This is a necessary trade-off if we want to avoid imposing model constraints on the distribution, such as linearity, and if we want to establish theoretical guarantees for non-IID time series data. The inherent limitation in sample efficiency arises because the conditional distribution is determined by the specific realization of a particular event.

Continuous-valued time series involves working with conditional probability density functions rather than conditional probability tables (CPTs). We have tried to extend this algorithm to probability density, but it is very challenging. For continuous-valued time series, the number of parent configurations is infinite, making it difficult to create a finite number of time series segments unless additional processing, such as categorizing the data or generating discrete analogs of a continuous distribution, is applied.

Sample Efficiency: This limitation has been addressed to some extent, as discussed in Appendix E, but it remains a fundamental issue that we aim to explore in future work. In the current framework, without any special operation like k-PC and Top-K parents, the effective sample size of each time series segment,

$|T_{\text{sub}}| \approx T/|s^{|\widehat{\text{SPA}}(X_t^j)|}|$, diminishes exponentially with domain size s and parent size $\widehat{\text{SPA}}(X_t^j)$, in contrast to the baseline method, which maintains a sample size of T. The challenge lies in how to aggregate the information from each time series segment more efficiently.

Smooth Causal Mechanism Change: With the *Mechanism-shift* SCM, there is a constraint that two different mechanisms switch instantaneously. However, in some real-world scenarios, the causal mechanism shift may occur gradually over time.

Causal Relations and Assumptions: As mentioned in Section G, before implementing our causal-driven algorithm, it is crucial to apply common sense or expert judgment to assess whether potential causal relationships exist. All the assumptions in Section A.3 should be met before using the algorithm.

H.1 Experiments Under Assumption Violations

In order to comprehensively evaluate the proposed algorithm under these limitations, we conducted experiments that violate Assumption **A6** Boundary Separation Assumption by placing change points near $t = 1$ and $t = T$, specifically within 50 time points (Case B). We compared this with a standard setting (Case A), where change points are located within the interval $[50, T - 50]$, satisfying Assumption **A6**.

The algorithm's performance on continuous data after discretization varies depending on how the discretization process affects the underlying causal structure of the original data. To assess this, we also evaluated the algorithm on continuous data discretized using the 5-number summary (Case C).

The results are presented below. The table reports the average estimated error $\frac{\widetilde{T}_c^j - T_c^j}{T_c^j}$ and its standard deviation over 50 random trials. In all experiments, we set $T = 1500$.

Case A (normal)		Case B (without A6)		Case C (discretization)	
Average	Std	Average	Std	Average	Std
0.04	0.10	0.51	0.31	0.23	0.21