Appendix

A PCMCI Algorithm

The PCMCI algorithm is proposed by Runge et al. [2019], aiming to detect time-lagged causal relations in a window causal graph. There are two stages of PCMCI: the condition-selection stage and the causal discovery stage. In the first stage, unnecessary edges are removed based on the conditional independencies from an initialized partially connected graph where Assumption **A4-A5** should be satisfied. In the second stage, Momentary Conditional Independence tests (MCI) are used to further remove the false positive edges caused by autocorrelations in time series data. More specifically, these two steps can be briefly formalized as follows:

- PC_1 in Algorithm A1: Condition selection stage. PC_1 is a variant of the skeleton-discovery part of the PC algorithm in a more robust version named stable-PC Le et al. [2016]. The goal in this stage is to obtain a superset of the parents $\widehat{\mathrm{Pa}}(X_t^j)$ for all variables $X_{t\in[\tau_{\max}+1,T]}^{j\in[n]} \in \mathbf{V}$. Initialize $\widehat{\mathrm{Pa}}(X_t^j) = \{X_{t-\tau}^i\}_{i\in[n],\tau\in[\tau_{\max}]}$. $\widehat{\mathrm{Pa}}(X_t^j)$ will remove $X_{t-\tau}^i$ if

$$X_{t-\tau}^i \perp\!\!\!\perp X_t^j \,\Big|\, \widehat{\mathrm{Pa}}(X_t^j)\backslash\{X_{t-\tau}^i\} \tag{1}$$

- MCI in Algorithm A2: Causal discovery stage. In this stage, do MCI tests for all variable pairs $(X_{t-\tau}^i, X_t^j)$ with $i, j \in [n]$ and time delays $\tau \in [\tau_{\max}]$:

$$MCI(X_{t-\tau}^i, X_t^j | \widehat{\mathrm{Pa}}(X_t^j)\backslash\{X_{t-\tau}^i\}, \widehat{\mathrm{Pa}}(X_{t-\tau}^i)) \tag{2}$$

  where $\widehat{\mathrm{Pa}}(X_t^j)$ and $\widehat{\mathrm{Pa}}(X_{t-\tau}^i)$ are estimated from the PC_1 stage.

Note that $\tau_{\max}$ in this section is the same as τ_{ub} in the main paper, serving as the upper bound for the time lag that exhibits causal effects. On the other hand, $\tau_{\max}$ in the main paper denotes the maximum time lag observed within the multivariate time series. Essentially, in the main paper, τ_{ub} is a parameter that must be fed into the algorithm, and $\tau_{\max}$ is observed from the true causal graph. As a default, we assume τ_{ub} is configured with a value greater than $\tau_{\max}$, ensuring that the algorithm uncovers the correct causal relations. See Fig.1 for more detail.

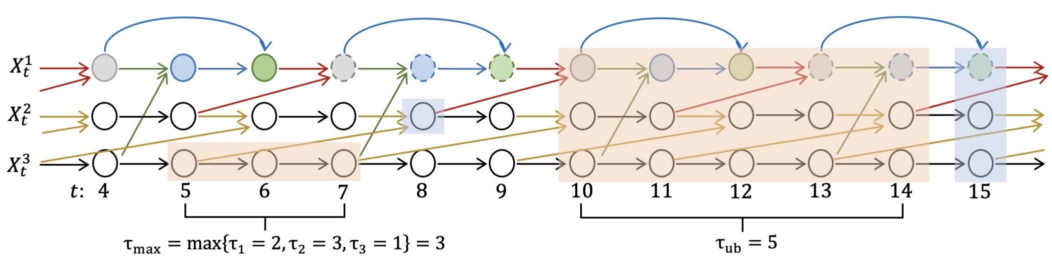


Figure 1: Set τ_{ub} to be 5, then all parent candidates of variables at $t = 15$ are included in the large orange box, ranging from $t = 10$ to $t = 14$. Consequently, the algorithm will only examine causal effects with a time lag not exceeding 5. In the causal graph, $\tau_{\max}$ is 3, representing the maximum time lag observed among the 3-variate time series. Specifically, the maximum time lag for each component time series is $\tau_1 = 2, \tau_2 = 3, \tau_3 = 1$, respectively, and $\tau_{\max}$ represents the largest value among these three maximum lags.

Algorithm A1 $PC_{q_{\max}}$

Input: A n-variate time series $V = (\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, ..., \mathbf{X}^n)$, target time series $\mathbf{X}^j$, maximum time lag $\tau_{\max}$, significance threshold α_{PC}, maximum condition dimension $p_{\max}$ (default $p_{\max} = n\tau_{\max}$), maximum number of combinations $q_{\max}$ (default $q_{\max}$ = 1), conditional independence test function CI

function $CI(X, Y, \mathbf{Z})$
    Test $X \perp\!\!\!\perp Y | \mathbf{Z}$ using test statistic measure I
    return p-value, test statistic value I
Initialize preliminary set of parents $\widehat{Pa}(X_t^j) = \{X_{t-\tau}^i : i \in \{1, ..., n\}, \tau \in \{1, ..., \tau_{\max}\}\}$
Initialize dictionary of test statistic values $I^{\min}(X_{t-\tau}^i \to X_t^j) = \infty \; \forall X_{t-\tau}^i \in \widehat{Pa}(X_t^j)$
for $p = 0, 1, 2, ..., p_{\max}$ **do**
    if $|\widehat{Pa}(X_t^j)| - 1 < p$ **then**
        Break for-loop
    end if
    for all $X_{t-\tau}^i$ in $\widehat{Pa}(X_t^j)$ **do**
        $q = -1$
        for all lexicographically chosen subsets $\mathcal{S} \subseteq \widehat{Pa}(X_t^j) \setminus \{X_{t-\tau}^i\}$ with $|\mathcal{S}| = p$ **do**
            $q = q + 1$
            if $q \geq q_{\max}$ **then**
                Break from inner for-loop
            end if
            Run CI test to obtain (p-value, I) $\leftarrow CI(X_{t-\tau}^i, X_t^j, \mathcal{S})$
            if $|I| < I^{\min}(X_{t-\tau}^i \to X_t^j)$ **then** ▷Store min. I of parent among all tests
                $I^{\min}(X_{t-\tau}^i \to X_t^j) = |I|$
            end if
            if p-value $> \alpha_{PC}$ **then** ▷Removed only after all $X_{t-\tau}^i$ have been tested
                Mark $X_{t-\tau}^i$ for removal from $\widehat{Pa}(X_t^j)$
                Break from inner for-loop
            end if
        end for
    end for
    Remove non-significant parents from $\widehat{Pa}(X_t^j)$
    Sort parents in $\widehat{Pa}(X_t^j)$ by $I^{\min}(X_{t-\tau}^i \to X_t^j)$ from largest to smallest
end for
return $\widehat{Pa}(X_t^j)$

Algorithm A2 MCI

Input: A n-variate time series $V = (\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, ..., \mathbf{X}^n)$, sorted parents $\widehat{Pa}(X_t^j)$ for all variables X^j estimated with Algorithm A1, maximum time lag $\tau_{\max}$, maximum number p_X of parents of variable X^i, and conditional independence test function CI
for all $(X_{t-\tau}^i, X_t^j)$ with $i, j \in \{1, ..., n\}, \tau \in \{0, ..., \tau_{\max}\}$, excluding (X_t^j, X_t^j) **do**
    Remove $X_{t-\tau}^i$ from $\widehat{Pa}(X_t^j)$ if necessary
    Define $\widehat{Pa}_{p_X}(X_{t-\tau}^i)$ as the first p_X parents from $\widehat{Pa}(X_t^i)$, shifted by τ
    Run MCI test to obtain (p-value, I) $\leftarrow CI(X_{t-\tau}^i, X_t^j, \mathbf{Z} = \{\widehat{Pa}(X_t^j), \widehat{Pa}_{p_X}(X_{t-\tau}^i)\})$
end for
Optionally adjust p-value of all links by False Discovery Rate-approach (FDR)
return p-value and MCI test statistic values

B PCMCI$_\Omega$

For simplicity's sake, define sets: $[b] := \{1, 2, ..., b\}$ and $[a, b] := \{a, a+1, ..., b\}$, where $a, b \in \mathbb{N}$.

Algorithm B1 $PCMCI_\Omega$

Input: A n-variate time series $V = (\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, ..., \mathbf{X}^n)$, periodicity upper bound ω_{ub}, time lag upper bound τ_{ub}. By default, we assume τ_{ub} and ω_{ub} are larger than their true value.

A superset of parent set is obtained using PCMCI with τ_{ub} and denote it by $\widehat{S\text{Pa}}(X_t^j)\ \forall j, t$.

for $\mathbf{X}^j$ where $j \in [n]$ **do**

  for a guess $\omega \in [\omega_{ub}]$ of ω_j **do**

    Let $\widehat{\Pi}^j := \{\widehat{\Pi}_k^j | k \in [\omega]\}$ where $\widehat{\Pi}_k^j = \{2\tau_{\text{ub}} + k, 2\tau_{\text{ub}} + \omega + k, 2\tau_{\text{ub}} + 2\omega + k, \cdots\}$.

    for $k \in [\omega]$ **do**

      Initialize the parent set for $X_t^j, t \in \{t : t \geq 2\tau_{\text{ub}}, t \in \widehat{\Pi}_k^j\}$ (with guess ω) denoted by $\widehat{\text{Pa}}_\omega(X_t^j) \leftarrow \widehat{S\text{Pa}}(X_t^j)$.

      Consider $X_{t-\tau}^i \in \widehat{\text{Pa}}_\omega(X_t^j)$. Remove $X_{t-\tau}^i$ from $\widehat{\text{Pa}}_\omega(X_t^j)$ if $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \left(\widehat{S\text{Pa}}(X_t^j) \cup \widehat{S\text{Pa}}(X_{t-\tau}^i)\right) \setminus X_{t-\tau}^i$ using a CI Test with samples $t \in \{t : t \geq 2\tau_{\text{ub}}, t \in \widehat{\Pi}_k^j\}$.

      Store $\widehat{\text{Pa}}_\omega(X_t^j)$ for $X_t^j, t \in \{t : t \geq 2\tau_{\text{ub}}, t \in \widehat{\Pi}_k^j\}$.

    end for

  end for

  ....

  if there exists *turning points* $S_j,\ S_j \in [\omega_{\text{ub}}]$ **then**

    $\widehat{\omega}_j \leftarrow \min S_j$

  else

  ....

    $\widehat{\omega}_j \leftarrow \arg\min_{\omega \in [\omega_{\text{ub}}]} \max_{k \in [\omega]} |\widehat{\text{Pa}}_\omega(X^j_{t \in \widehat{\Pi}_k^j})|$.

  end if

  Set $\widehat{\text{Pa}}(X_t^j) \leftarrow \widehat{\text{Pa}}_{\widehat{\omega}_j}(X_t^j)$ for $X_t^j, t \in \{t : t \geq 2\tau_{\text{ub}}\}$.

end for

return $\hat{\omega}_j$ and $\widehat{\text{Pa}}(X_t^j)\ \forall j \in [n], t \geq 2\tau_{\text{ub}}$.

C Soundness of PCMCI$_\Omega$

C.1 Stationary Markov Chain

Claim: Any discrete-valued time series V with *Semi-Stationary* Structural Causal Model (SCM) satisfying assumption **A1, A2, A4, A5** can be written as a Markov chain $\{Z_n\}$ as long as this Markov chain satisfies $\text{Pa}(Z_n) \subset Z_n \cup Z_{n-1}$ for all n, where Z_n is a set of variables in V. This Markov chain has a finite number of states if all time series in V are discrete-valued time series.

Note that when the notation n is related to a Markov chain Z_n, it means the running index. In the context of $X_t^{j \in [n]}$, n represents the index of component time series within the n-variate time series.

To simplify, assume that one associated Markov chain of $V = \{\mathbf{X}, \mathbf{Y}\}$ has $Z_n = \{X_t, Y_t, X_{t-1}, Y_{t-1}\}$ with $t \in \{t \in \mathbb{N}^+ :, t \leq T\}$ satisfying $\text{Pa}(Z_n) \subset Z_n \cup Z_{n-1}$. Here, the notation for the time points of variables is simplified as t and $t-1$, even though it should be a function of n, the running index of the Markov chain. Note that $Z_{n-1} = \{X_{t-2}, Y_{t-2}, X_{t-3}, Y_{t-3}\}$ rather than $\{X_{t-1}, Y_{t-1}, X_{t-2}, Y_{t-2}\}$, as the simplified notation could erroneously suggest the latter sequence. A simple proof is shown below through Markov assumption (**A2**).

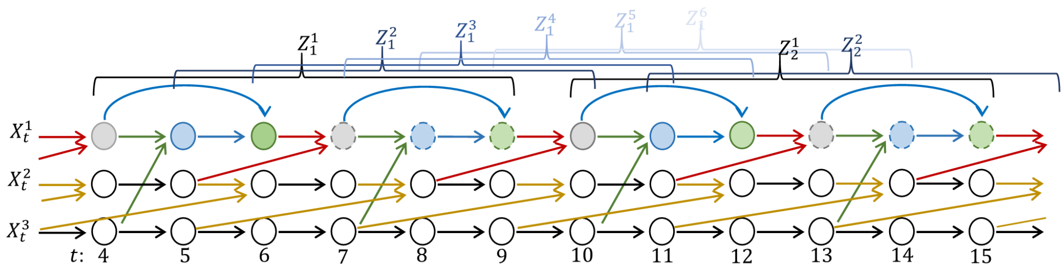


Figure 2: Partial causal graph for 3-variate time series $V = \{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3\}$ with a Semi-Stationary SCM where $\tau_{\max} = 3$, $\omega_1 = 3$, $\omega_2 = 2$, $\omega_3 = 1$, $\Omega = 6$ and $\delta = 6$. The first 3(=$\tau_{\max}$) time slices $\{\mathbf{X}_t\}_{1 \leq t \leq 3}$ are the starting points. The same color edges denote the same causal mechanism. E.g. for $\mathbf{X}^1$: there are 3 ($= \omega_j$) time partition subsets $\{\Pi_k^1\}_{1 \leq k \leq 3}$. The time points t of nodes X_t^1 sharing the same filling color are in the same time partition subsets. The time points t of nodes X_t^1 sharing both the same filling color and the same outline shape are in the same homogenous time partition subsets. There are 6 ($= \delta$) different Markov chains in this multivariate time series V, and the first element of these 6 Markov chains is shown as $\{Z_1^q\}_{1 \leq q \leq 6}$ and are tinted with a gradient of blue hues. Z_1^1 and Z_2^1 denote the first two elements of the first Markov chain while Z_1^2 and Z_2^2 denote the first two elements of the second Markov chain.

Proof.

$$
\begin{aligned}
& p(Z_n | Z_{n-1}, Z_{n-2}, ...) && (3)\\
&= p(X_t, Y_t, X_{t-1}, Y_{t-1} | Z_{n-1}, Z_{n-2}, ...) && (4)\\
&= p(X_t | Z_n \cup Z_{n-1} \setminus X_t, Z_{n-2}, \cdots) p\Big(Y_t | Z_n \cup Z_{n-1} \setminus (X_t \cup Y_t), Z_{n-2}, \cdots\Big) \cdots && (5)\\
&= p\Big(X_t | \mathrm{Pa}(X_t), Z_n \cup Z_{n-1} \setminus (X_t \cup \mathrm{Pa}(X_t))\Big) && (6)\\
&\quad \times p\Big(Y_t | \mathrm{Pa}(Y_t), Z_n \cup Z_{n-1} \setminus (X_t \cup Y_t \cup \mathrm{Pa}(Y_t))\Big) \\
&\quad \times p\Big(X_{t-1} | \mathrm{Pa}(X_{t-1}), Z_n \cup Z_{n-1} \setminus (X_t \cup Y_t \cup X_{t-1} \cup \mathrm{Pa}(X_{t-1}))\Big) \cdots \\
&= p(X_t | Z_n \cup Z_{n-1} \setminus X_t) p\Big(Y_t | Z_n \cup Z_{n-1} \setminus (X_t \cup Y_t)\Big) \cdots && (7)\\
&= p(X_t, Y_t, X_{t-1}, Y_{t-1} | Z_{n-1}) && (8)\\
&= p(Z_n | Z_{n-1}) && (9)
\end{aligned}
$$

□

Assume that the space of both X_t and Y_t with $t < T$ are $\{1, 2\}$. There are total $2^4 = 16$ states of Markov Chain $\{Z_n\} = \{\{X_t, Y_t, X_{t-1}, Y_{t-1}\}\}$. The transition probability $\mathbf{P}$ for this Markov Chain is illustrated as a 16×16 matrix:

$$
\mathbf{P} = \begin{array}{c} \\ (1,1,1,1) \\ (2,1,1,1) \\ \cdots \end{array}
\begin{array}{c}
\begin{array}{ccc} (1,1,1,1) & (2,1,1,1) & \cdots \end{array} \\
\left[\begin{array}{ccc} p_{1,1} & p_{1,2} & \cdots \\ p_{2,1} & p_{2,2} & \cdots \\ \cdots & \cdots & \cdots \end{array}\right]_{16 \times 16}
\end{array}
$$

where $(1, 1, 1, 1)$ means $X_t = 1, Y_t = 1, X_{t-1} = 1, Y_{t-1} = 1$. Each row in this transition probability matrix is a conditional distribution of Z_n given one realization of Z_{n-1}. Each entry is a probability of having one specific realization of Z_n given one realization of Z_{n-1}. This probability can be decomposed by conditional distributions based on Markov assumption (**A2**). Take $p_{1,1}$ as an example:

$$
\begin{aligned}
p_{1,1} &= p(X_t = 1, Y_t = 1, X_{t-1} = 1, Y_{t-1} = 1 | X_{t-2} = 1, Y_{t-2} = 1, X_{t-3} = 1, Y_{t-3} = 1) && (10)\\
&= p(X_t = 1 | \mathrm{Pa}(X_t)) p(Y_t = 1 | \mathrm{Pa}(Y_t)) p(X_{t-1} = 1 | \mathrm{Pa}(X_{t-1})) p(Y_{t-1} = 1 | \mathrm{Pa}(Y_{t-1})) && (11)
\end{aligned}
$$

where $\mathrm{Pa}(.)$ here are realizations, not random variables.

For time series V with *Semi-Stationary* SCM, there are (potentially) δ different Markov chains $\{Z_n^q\}, q \in [\delta]$:

$$Z_n^q = \{\mathbf{X}_{\tau_{\max}+q+(n-1)\delta}, \mathbf{X}_{\tau_{\max}+q+1+(n-1)\delta}, ..., \mathbf{X}_{\tau_{\max}+q-1+n\delta}\},$$

where $n \in \{n : n \in \mathbb{N}^+, \tau_{\max} + q - 1 + n\delta \leq T\}$, $\delta = \lceil \frac{\tau_{\max}+1}{\Omega} \rceil \Omega$. As proved in the claim, such a Markov chain exists as long as $\mathrm{Pa}(Z_n^q) \subset Z_n^q \cup Z_{n-1}^q$ for all n. The value of δ can guarantee the existence of such Markov chain because δ is larger than $\tau_{\max} + 1$ and is a multiple of Ω, that is, a multiple of all $\{\omega_j\}_{j\in[n]}$. By doing so, $\mathrm{Pa}(Z_n^q) \subset Z_n^q \cup Z_{n-1}^q$ is satisfied; for any variable X_t^j, there exists $q \in [\delta]$ and $n \in \mathbb{N}^+$ such that variable X_t^j and its parent set $\mathrm{Pa}(X_t^j)$ can be included in Z_n^q; and the causal mechanism generating Z_n^q is invariant for different n. The state space of $\{Z_n^q\}$ is the set containing all possible realizations of $\{\mathbf{X}_{\tau_{\max}+q+(i-1)+(n-1)\delta}\}_{i\in[\delta],n\in\mathbb{N}}$. The transition probabilities between the states are the product of associated causal mechanisms based on Markov assumption (**A2**).

Determined by the starting slice $\mathbf{X}_t$ where $\tau_{\max} < t \leq \tau_{\max} + \delta$, there should be δ potentially different Markov chains $\{Z_n^q\}$ where $1 \leq q \leq \delta$. To be more specific, those Markov chains are:

$$\begin{aligned}
\text{Markov Chain 1: } Z_n^1 &= \{\mathbf{X}_{\tau_{\max}+1+(n-1)\delta}, \mathbf{X}_{\tau_{\max}+2+(n-1)\delta}, ..., \mathbf{X}_{\tau_{\max}+n\delta}\}, && (12)\\
&\text{where } n \in \{n : n \in \mathbb{N}^+, \tau_{\max} + n\delta \leq T\}.\\
\text{Markov Chain 2: } Z_n^2 &= \{\mathbf{X}_{\tau_{\max}+2+(n-1)\delta}, \mathbf{X}_{\tau_{\max}+3+(n-1)\delta}, ..., \mathbf{X}_{\tau_{\max}+1+n\delta}\}, && (13)\\
&\text{where } n \in \{n : n \in \mathbb{N}^+, \tau_{\max} + 1 + n\delta \leq T\}.\\
&\vdots\\
\text{Markov Chain } \delta\text{: } Z_n^\delta &= \{\mathbf{X}_{\tau_{\max}+n\delta}, \mathbf{X}_{\tau_{\max}+1+n\delta}, ..., \mathbf{X}_{\tau_{\max}-1+(n+1)\delta}\}, && (14)\\
&\text{where } n \in \{n : n \in \mathbb{N}^+, \tau_{\max} - 1 + (n+1)\delta \leq T\}.
\end{aligned}$$

Given Irreducible and Aperiodic Markov Chain assumption (**A7**), discrete-time Markov chain $\{Z_n^q\}_{0<n},\ q \in [\delta]$ with finite states should be a stationary and ergodic Markov chain, and there is a unique stationary distribution π_q (Bertsekas and Tsitsiklis [2008], Karlin [2014]). Additionally, the large power of the associate transition matrix $\mathbf{P}_q$ will eventually converge to a matrix in which each row is the stationary distribution π_q. Equivalently,

$$\lim_{n\to\infty} p(Z_n^q = a | Z_1^q = b) = p(Z_n^q = a), \forall a, b \in S. \tag{15}$$

where S is the state space of Z_n^q.

In other words, after a sufficiently long time, equivalently, n is large enough, the distribution of $\{Z_n^q\}$ does not change with increasing n. That is, for large enough n:

$$p(Z_{n_1}^q) = p(Z_{n_2}^q), \forall n_1, n_2 > n. \tag{16}$$

Returning from the stationary and ergodic Markov chains $\{Z_n^q\},\ q \in [\delta]$ back to the original data V through Eq.(12) to Eq.(14), the distribution of the original data V must adhere to the following condition:

$$\begin{aligned}
&p(\mathbf{X}_{\tau_{\max}+q+n_1\delta}, \mathbf{X}_{\tau_{\max}+q+1+n_1\delta}, ..., \mathbf{X}_{\tau_{\max}+q+\delta-1+n_1\delta})\\
&= p(\mathbf{X}_{\tau_{\max}+q+n_2\delta}, \mathbf{X}_{\tau_{\max}+q+1+n_2\delta}, ..., \mathbf{X}_{\tau_{\max}+q+\delta-1+n_2\delta})
\end{aligned} \tag{17}$$

for any $q \in [\delta]$ and $n_1, n_2 > n$.

Given these clarifications, we can naturally introduce a more refined time partition that is based on, yet finer than, the time partition defined in Definition 2.3 in the main paper.

Definition C.1 (*Homogenous Time Partition*)**.** For a univariate time series $\mathbf{X}^j$ in a Semi-Stationary SCM with periodicity ω_j, the time partition Π_k^j of $\mathbf{X}^j$ can be further divided into a series of non-overlapping and non-empty subsets $\{\pi_{(k,s)}^j\}_{1\leq s\leq \frac{\delta}{\omega_j}}$. For each $t \in [\tau_{\max} + 1, T]$, there exists $k \in [\omega_j]$ so that $t \in \Pi_k^j$ and further there exists $s \in [\frac{\delta}{\omega_j}]$ so that $t \in \pi_{(k,s)}^j$. $\pi_{(k,s)}^j$ can be written as:

$$\pi_{(k,s)}^j := \{t : \tau_{\max} + 1 \leq t \leq T, (t \bmod \omega_j) + 1 = k, (t \bmod \frac{\delta}{\omega_j}) + 1 = s\}. \tag{18}$$

With this definition, we have $\cup_{s=1}^{\frac{\delta}{\omega_j}} \pi^j_{(k,s)} = \Pi^j_k$. While time partition Π^j_k guarantees that all variables in $\{X^j_t\}_{t\in\Pi^j_k}$ share the same causal mechanism, homogenous time partition $\pi^j_{(k,s)}$ guarantees that all variables in $\{X^j_t\}_{t>t', t\in\pi^j_{(k,s)}}$ share the same distribution where t' represent the steps needed by the associated Markov chain to achieve equilibrium.

Fig.2 shows a partial causal graph for a 3-variate time series with Semi-Stationary SCM. $\tau_{\max} = 3$ means that the causal mechanisms start from $t = 4$, and the random variables with $t \in \{1, 2, 3\}$ are random noises. For the first time series $\mathbf{X}^1$, the periodicity ω_1 is 3. And the periodicity of the time series $\mathbf{X}^2$ and $\mathbf{X}^3$ is 2 and 1, respectively. The periodicity of the whole time series V is obtained by LCM$(3, 2, 1) = 6$. $\delta = \lceil \frac{\tau_{\max}+1}{\Omega} \rceil \Omega = \lceil \frac{3+1}{6} \rceil \times 6 = 1 \times 6 = 6$.

In Fig.2, periodicity $\omega_1 = 3$ means that the causal mechanisms repeat every three time points and hence there are three time partition subsets $\Pi^1_k, k \in [3]$. More specifically, $\Pi^1_1 = \{4, 7, 10, 13, ..., 4 + 3N, ...\}, \Pi^1_2 = \{5, 8, 11, 14, ..., 5 + 3N, ...\}, \Pi^1_3 = \{6, 9, 12, 15, ..., 6 + 3N, ...\}$ where $N \in \mathbb{N}^+$. Random variables $\{X^1_t\}$ with t in the same time partition subset share the same causal mechanism. However, they may not share the same marginal distribution.

Still in Fig.2, based on the definition of homogenous time partition, time partition subset Π^1_1 for $\mathbf{X}^1$ can be further decomposed as $\pi^1_{(k=1,s=1)} = \{4, 10, ..., 4 + \delta N, ...\}, \pi^1_{(k=1,s=2)} = \{7, 13, ..., 7 + \delta N, ...\}$. where $s \in [\frac{\delta}{\omega_1}]$. After a long run n, Z^1_n and Z^1_{n+1} will eventually share the same distribution, that is, all the variables inside Z^q_n will share the same joint or marginal distribution as the corresponding variables inside Z^q_{n+1}. To illustrate this, we assume that this Markov chain has already achieved its equilibrium at time point $t = 4$. Based on Eq.(12) and Eq.(17), we have:

$$p(\mathbf{X}_4, \mathbf{X}_5, ..., \mathbf{X}_9) = p(\mathbf{X}_{10}, \mathbf{X}_{11}, ..., \mathbf{X}_{15}) = p(\mathbf{X}_{16}, \mathbf{X}_{17}, ..., \mathbf{X}_{21}) = \cdots \tag{19}$$

From the identical joint distribution, we can further have:

$$p(X^1_4) = p(X^1_{10}) = p(X^1_{16}) = \cdots \tag{20}$$

as $X^1_4 \in \mathbf{X}_4$, $X^1_{10} \in \mathbf{X}_{10}$ and $X^1_{16} \in \mathbf{X}_{16}$.

Therefore, for sufficiently large values of t ensuring that Z^1_n has reached its stationary distribution, all variables within $\{X^j_t\}_{t\in\pi^j(k,s)}$ will share the same distribution.

In Fig.2, there are $6(= \delta)$ potentially different Markov chains $\{Z^q_n\}, q \in [\delta]$ in V. For any time window with length δ, $\{\mathbf{X}_t, ..., \mathbf{X}_{t+\delta-1}\}$, there exists $q \in [\delta], n \in \mathbb{N}^+$, so that this time window can be completely included in Z^q_n. For instance, set $\{\mathbf{X}_5, \mathbf{X}_6, ..., \mathbf{X}_{10}\}$ is in Z^2_1, which is the first element of Markov chain $\{Z^2_n\}$.

Constructing Markov chains and applying the Irreducible and Aperiodic Markov Chain assumption (**A7**) enable us to obtain a consistent estimator for the conditional and joint distributions of interest.

C.2 Consistent Estimator

The conditional distributions for variables in $\{X^j_t\}_{t\in\Pi^j_k}$ are the same, that is, $p(x^j_{t_1}|\mathrm{Pa}(x^j_{t_1})) = p(x^j_{t_2}|\mathrm{Pa}(x^j_{t_2})),\ \forall t_1, t_2 \in \Pi^j_k$. For simplicity, denote

$$p_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t)) := p(x^j_t|\mathrm{Pa}(x^j_t)),\ \forall t \in \Pi^j_k \tag{21}$$

Consider an indicator function such that $\mathbb{1}(x^j_t, \mathrm{Pa}(x^j_t)) = 1$ if configuration $(x^j_t, \mathrm{Pa}(x^j_t))$ has realized, otherwise $\mathbb{1}(x^j_t, \mathrm{Pa}(x^j_t)) = 0$.

Since every $t \in \pi^j_{(k,s)}$ is apart from each other with $N\delta$ steps where $N \in \mathbb{N}^+$, and there must exist $q \in [\omega_j]$ and $n_1 \in \mathbb{N}^+$ so that $\{x^j_t, \mathrm{Pa}(x^j_t)\} \in Z^q_{n_1}$, then for the same q, there must exist another n_2 so that $\{x^j_{t+N\delta}, \mathrm{Pa}(x^j_{t+N\delta})\} \in Z^q_{n_2}$. Hence, we have $\{\mathbb{1}(x^j_t, \mathrm{Pa}(x^j_t))\}_{t\in\pi^j_{(k,s)}} = \{f(Z^{q(t)}_{n_1(t)})\}_{t\in\pi^j_{(k,s)}}$ with some function $f : \mathbb{R}^{n\times\delta} \to \mathbb{R}^1$ satisfying $E|f(Z^{q(t)}_{n_1(t)})| < \infty$. Since the value of t determines q and n_1, we use $q(t)$ and $n_1(t)$ to emphasize their relations. For large enough $t > t'$,

$\{\mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j))\}_{t>t', t\in\pi^j_{(k,s)}}$ are identical samples where t' is the time point needed by the associate Markov chain to achieve its equilibrium after $n_1(t')$ steps.

Without loss of generality, we assume T is a multiple of δ all the time.

We can construct an estimator of $p(x_t^j, \mathrm{Pa}(x_t^j))$ with large enough t as:

$$\hat{p}(x_t^j, \mathrm{Pa}(x_t^j)) = \frac{\delta}{T} \sum_{t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) \tag{22}$$

where k, s is determined by t and there must exist one and only one k, s satisfying $t \in \pi^j_{(k,s)}$. Now, we are going to show this estimator is consistent.

We first decompose the estimator into two parts: time point $t \le t'$ and $t > t'$, where t' represents the time point when the equilibrium of the associated Markov chain is achieved.

$$\begin{aligned}
&\frac{\delta}{T} \sum_{t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) && (23)\\
&= \frac{\delta}{T}\Bigg(\sum_{t\le t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) + \sum_{t>t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j))\Bigg) && (24)\\
&= \frac{\delta}{T} \sum_{t\le t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) + \frac{\delta}{T} \sum_{t>t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) && (25)\\
&= \frac{\delta}{T} \sum_{t\le t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) + \frac{\delta}{T-t'}\frac{T-t'}{\delta}\frac{\delta}{T} \sum_{t>t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) && (26)\\
&= \frac{\delta}{T} \sum_{t\le t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) + \frac{\delta}{T-t'}\frac{T-t'}{\delta}\frac{\delta}{T} \sum_{t>t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) && (27)\\
&= \frac{\delta}{T} \sum_{t\le t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) + \frac{T-t'}{T}\Bigg(\frac{\delta}{T-t'} \sum_{t>t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j))\Bigg) && (28)\\
& && (29)
\end{aligned}$$

Take a limit of Eq.(23), we have:

$$\begin{aligned}
&\lim_{T\to\infty} \frac{\delta}{T} \sum_{t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) && (30)\\
&= \lim_{T\to\infty} \frac{\delta}{T} \sum_{t\le t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j)) + \lim_{T\to\infty} \frac{T-t'}{T}\Bigg(\frac{\delta}{T-t'} \sum_{t>t', t\in\pi^j_{(k,s)}} \mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j))\Bigg) && (31)\\
&= 0 + \lim_{T\to\infty} \frac{T-t'}{T}\Bigg(\frac{1}{n_1(T)-n_1(t')} \sum_{n_1(t)>n_1(t')}^{n_1(T)} f(Z^{q(t)}_{n_1(t)})\Bigg), \text{where } t > t', t \in \pi^j_{(k,s)} && (32)\\
&\overset{\text{Birkhoff's Ergodic Theorem}}{=\!=\!=\!=\!=\!=} 0 + E\Bigg(f(Z^{q(t)}_{n_1(t)})\Bigg) && (33)\\
&= E\Bigg(\mathbb{1}(x_t^j, \mathrm{Pa}(x_t^j))\Bigg), \text{where } t > t', t \in \pi^j_{(k,s)} && (34)\\
&= p(x_t^j, \mathrm{Pa}(x_t^j)), \text{where } t > t', t \in \pi^j_{(k,s)} && (35)
\end{aligned}$$

Denote

$$p_{t\in\pi^j_{(k,s)}}(x_t^j, \mathrm{Pa}(x_t^j)) \coloneqq p(x_t^j, \mathrm{Pa}(x_t^j)), \text{where } t > t', t \in \pi^j_{(k,s)} \tag{36}$$

Based on the definition of homogenous time partition and time partition, $p_{t\in\pi^j_{(k,s)}}(x^j_t|\mathrm{Pa}(x^j_t)) = p_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t)),\ \forall s \in [\frac{\delta}{\omega_j}]$.

Similar to Eq.(22), one estimator of $p_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t)),\ \forall k = [\omega_j]$ is

$$\hat{p}_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t)) = \frac{\sum_{t\in\Pi^j_k} \mathbb{1}(x^j_t, \mathrm{Pa}(x^j_t))}{\sum_{t\in\Pi^j_k} \mathbb{1}(\mathrm{Pa}(x^j_t))} \tag{37}$$

$$= \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} \sum_{t\in\pi^j_{(k,s)}} \mathbb{1}(x^j_t, \mathrm{Pa}(x^j_t))}{\sum_{s=1}^{\frac{\delta}{\omega_j}} \sum_{t\in\pi^j_{(k,s)}} \mathbb{1}(\mathrm{Pa}(x^j_t))} \tag{38}$$

$$= \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} \frac{T}{\delta} \sum_{t\in\pi^j_{(k,s)}} \mathbb{1}(x^j_t, \mathrm{Pa}(x^j_t))}{\sum_{s=1}^{\frac{\delta}{\omega_j}} \frac{T}{\delta} \sum_{t\in\pi^j_{(k,s)}} \mathbb{1}(\mathrm{Pa}(x^j_t))} \tag{39}$$

Take a limit of Eq.(37), we have:

$$\lim_{T\to\infty} \hat{p}_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t)) \tag{40}$$

$$\overset{\text{Eq.(35)}}{=\!=\!=} \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi^j_{(k,s)}}(x^j_t, \mathrm{Pa}(x^j_t))}{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi^j_{(k,s)}}(\mathrm{Pa}(x^j_t))} \tag{41}$$

$$= \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi^j_{(k,s)}}(x^j_t|\mathrm{Pa}(x^j_t)) p_{t\in\pi^j_{(k,s)}}(\mathrm{Pa}(x^j_t))}{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi^j_{(k,s)}}(\mathrm{Pa}(x^j_t))} \tag{42}$$

$$\xlongequal{p_{t\in\pi^j_{(k,s)}}(x^j_t|\mathrm{Pa}(x^j_t)) \text{ are same for all } s} \frac{p_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t)) \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi^j_{(k,s)}}(\mathrm{Pa}(x^j_t))}{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi^j_{(k,s)}}(\mathrm{Pa}(x^j_t))} \tag{43}$$

$$= p_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t)) \tag{44}$$

Hence, $\hat{p}_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t))$ is a consistent estimator of $p_{t\in\Pi^j_k}(x^j_t|\mathrm{Pa}(x^j_t))$.

Similarly, we construct an estimator of $p(x^j_t| \cup_h \mathrm{Pa}_h(x^j_t))$ where $t \in [T]$:

$$\hat{p}(x^j_t| \cup_h \mathrm{Pa}_h(x^j_t)) = \sum_t \mathbb{1}(x^j_t| \cup_h \mathrm{Pa}_h(x^j_t)) \tag{45}$$

$$= \frac{\sum_t \mathbb{1}(x^j_t, \cup_h \mathrm{Pa}_h(x^j_t))}{\sum_t \mathbb{1}(\cup_h \mathrm{Pa}_h(x^j_t))}. \tag{46}$$

We will prove that this estimator is converged as T goes to infinity in Lemma D.2. Hence, it is a consistent estimator.

In this section, we have proved that $\hat{p}(x^j_t, \mathrm{Pa}(x^j_t))$ in Eq.(22) is a consistent estimator of $p(x^j_t, \mathrm{Pa}(x^j_t))$ using samples with t in the same homogenous time partition subset and $\hat{p}(x^j_t|\mathrm{Pa}(x^j_t))$ in Eq.(37) is a consistent estimator of $p(x^j_t|\mathrm{Pa}(x^j_t))$ using samples with t in the same time partition subset.

D Theorem

Theorem D.1. *Let $\widehat{\mathcal{G}}$ be the estimated graph using the Algorithm PCMCI$_\Omega$. Under assumptions **A1-A7** and with an oracle (infinite sample size limit), we have that:*

$$\widehat{\mathcal{G}} = \mathcal{G} \tag{47}$$

almost surely.

Lemma D.2. *Denote that* $\{\mathrm{Pa}_k(X_t^j)\}_{k\in[\omega_j]}$ *contain the true and illusory parent sets, where* ω_j *is the true periodicity of* $\mathbf{X}^j$. *For any random variable* X_t^j *with large enough* t, *under assumptions **A1-A7** and with an oracle (infinite sample size limit), we have:*

$$p\bigg(p(X_t^j|\cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X_t^j)) \neq p(X_t^j|\cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X_t^j)\setminus y)\bigg) = 1,\ \forall y \in \cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X_t^j) \tag{48}$$

Here, $p(X_t^j|\cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X_t^j)) = \lim_{T\to\infty}\hat{p}(X_t^j|\cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X_t^j))$.

Proof. We first prove that there exist a sequence of coefficients $\{\alpha_k\}_{k\in[\omega_j]}$ satisfying $\sum_{k=1}^{\omega_j}\alpha_k = 1$ so that:
$\forall$ configuration $\cup_h \mathrm{Pa}_h(x_t^j)$,

$$\hat{p}(x_t^j|\cup_h \mathrm{Pa}_h(x_t^j)) = \sum_{k=1}^{\omega_j}\alpha_k\hat{p}_k(x_t^j|\mathrm{Pa}(x_t^j)) \tag{49}$$

If this is correct, then $\hat{p}(x_t^j|\cup_h \mathrm{Pa}_h(x_t^j))$ would be a consistent estimator of $p(x_t^j|\cup_h \mathrm{Pa}_h(x_t^j))$. Based on Eq.(46), we have:

$$\hat{p}(x_t^j|\cup_h \mathrm{Pa}_h(x_t^j)) \tag{50}$$

$$= \frac{\sum_t \mathbb{1}(x_t^j, \cup_h\mathrm{Pa}_h(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))} \tag{51}$$

$$= \frac{\sum_t\sum_k \mathbb{1}(x_t^j, \cup_h\mathrm{Pa}_h(x_t^j))\mathbb{1}(t\in\Pi_k^j)}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))} \tag{52}$$

$$= \sum_k \frac{\sum_t \mathbb{1}(x_t^j, \cup_h\mathrm{Pa}_h(x_t^j))\mathbb{1}(t\in\Pi_k^j)}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))} \tag{53}$$

$$= \sum_k \frac{\sum_{t\in\Pi_k^j} \mathbb{1}(x_t^j, \cup_h\mathrm{Pa}_h(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))} \tag{54}$$

$$= \sum_k \Big(\frac{\sum_{t\in\Pi_k^j} \mathbb{1}(x_t^j, \cup_h\mathrm{Pa}_h(x_t^j))}{\cancel{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}}\frac{\cancel{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}}{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}\frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}\Big) \tag{55}$$

$$= \sum_k \Big(\frac{\sum_{t\in\Pi_k^j} \mathbb{1}(x_t^j, \cup_h\mathrm{Pa}_h(x_t^j))}{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}\frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_k\mathrm{Pa}_{t\in\Pi_k^j}(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}\Big) \tag{56}$$

$$= \sum_k \big(\hat{p}_{t\in\Pi_k^j}(x_t^j|\cup_h \mathrm{Pa}_h(x_t^j))\frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}\big) \tag{57}$$

$$= \sum_k \big(\hat{p}_{t\in\Pi_k^j}(x_t^j|\mathrm{Pa}(x_t^j), \cup_h\mathrm{Pa}_h(x_t^j)\setminus\mathrm{Pa}(x_t^j))\frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}\big) \tag{58}$$

$$= \sum_k \big(\hat{p}_{t\in\Pi_k^j}(x_t^j|\mathrm{Pa}(x_t^j))\frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}\big) \tag{59}$$

$$= \sum_k \alpha_k(T)\hat{p}_{t\in\Pi_k^j}(x_t^j|\mathrm{Pa}(x_t^j)), \tag{60}$$

$$\text{where } \alpha_k(T) = \frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}{\sum_t \mathbb{1}(\cup_h\mathrm{Pa}_h(x_t^j))}. \tag{61}$$

Using the same logic in Eq.(30)-(35), we can decompose the numerator and denominator of α_k with homogenous time partition until each component converges to a stationary distribution.

$$\alpha_k(T) = \frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j))}{\sum_k \sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j))} \tag{62}$$

$$= \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} \frac{T}{\delta} \sum_{t\in\pi_{(k,s)}^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j))}{\sum_k \sum_{s=1}^{\frac{\delta}{\omega_j}} \frac{T}{\delta} \sum_{t\in\pi_{(k,s)}^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j))} \tag{63}$$

$$\lim_{T\to\infty} \alpha_k(T) = \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j))}{\sum_k \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j))} \tag{64}$$

Without loss of generality, assume $y \in \mathrm{Pa}(x_t^j)$, where $t \in \Pi_j^1$ and $y \notin \mathrm{Pa}(x_t^j)$, where $t \notin \Pi_j^1$. Then we have

$$\hat{p}(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j) \setminus y) = \frac{\sum_t \mathbb{1}(x_t^j, \cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_t \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} \tag{65}$$

$$= \sum_{k=2}^{\omega_j} \big(\hat{p}_{t\in\Pi_k^j}(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j) \setminus y) \frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_t \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}\big) \tag{66}$$
$$+ \frac{\sum_{t\in\Pi_1^j} \mathbb{1}(x_t^j, \cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_{t\in\Pi_1^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} \frac{\sum_{t\in\Pi_1^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_t \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}$$

$$= \sum_{k=2}^{\omega_j} \beta_k(T) \hat{p}_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) + \beta_1(T) \hat{p}_{t\in\Pi_1^j}(x_t^j | \mathrm{Pa}(x_t^j) \setminus y) \tag{67}$$

$$\text{where } \beta_k(T) = \frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_t \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} \tag{68}$$

Similarly, we have:

$$\beta_k(T) = \frac{\sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_k \sum_{t\in\Pi_k^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} \tag{69}$$

$$= \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} \frac{T}{\delta} \sum_{t\in\pi_{(k,s)}^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_k \sum_{s=1}^{\frac{\delta}{\omega_j}} \frac{T}{\delta} \sum_{t\in\pi_{(k,s)}^j} \mathbb{1}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} \tag{70}$$

$$\lim_{T\to\infty} \beta_k(T) = \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_k \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} \tag{71}$$

Proving $p(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j)) \neq p(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j) \setminus y)$ is equal to proving:

$$p(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j)) - p(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j) \setminus y) \neq 0 \tag{72}$$

Substitutes Eq.(60) and Eq.(67) in Eq.(72), we have the following derivation:

$$
\begin{aligned}
& p(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j)) - p(x_t^j | \cup_h \mathrm{Pa}_h(x_t^j) \setminus y) && (73)\\
&= \lim_{T\to\infty} \left(\sum_{k=1}^{\omega_j} \alpha_k(T) \hat{p}_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) \right) - && (74)\\
& \lim_{T\to\infty} \left(\sum_{k=2}^{\omega_j} \beta_k(T) \hat{p}_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) + \beta_1(T) \hat{p}_{t\in\Pi_1^j}(x_t^j | \mathrm{Pa}(x_t^j) \setminus y) \right) \\
&= \sum_{k=1}^{\omega_j} \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j))}{\sum_k \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j))} p_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) && (75)\\
& - \Bigg(\sum_{k=2}^{\omega_j} \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_k \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} p_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) \\
& + \frac{\sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\Pi_{(1,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)}{\sum_k \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y)} p_{t\in\Pi_1^j}(x_t^j | \mathrm{Pa}(x_t^j) \setminus y) \Bigg)
\end{aligned}
$$

After equating the denominators, the numerator is:

$$
\begin{aligned}
& \left(\sum_{k=1}^{\omega_j} \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j)) p_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) \right) \left(\sum_{k=1}^{\omega_j} \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y) \right) \\
& - \Bigg(\sum_{k=2}^{\omega_j} \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y) p_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) \\
& + \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\Pi_{(1,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y) p_{t\in\Pi_1^j}(x_t^j | \mathrm{Pa}(x_t^j) \setminus y) \Bigg) \\
& \times \left(\sum_{k=1}^{\omega_j} \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j)) \right) && (76)
\end{aligned}
$$

For the sake of simplicity, denote

$$
\begin{aligned}
a_k &:= \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j)) && (77)\\
b_k &:= \sum_{s=1}^{\frac{\delta}{\omega_j}} p_{t\in\pi_{(k,s)}^j}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y) && (78)\\
c_k &:= p_{t\in\Pi_k^j}(x_t^j | \mathrm{Pa}(x_t^j)) && (79)\\
c_1' &:= p_{t\in\Pi_1^j}(x_t^j | \mathrm{Pa}(x_t^j) \setminus y) && (80)
\end{aligned}
$$

After substituting the simple notations in Eq.(76):

$$(\sum_{k=1}^{\omega_j} a_k c_k)(\sum_{k=1}^{\omega_j} b_k) - (\sum_{k=2}^{\omega_j} b_k c_k + b_1 c_1')(\sum_{k=1}^{\omega_j} a_k) \tag{81}$$

$$= \sum_{k=1}^{\omega_j} (c_k - c_{1'}) a_k b_1 + \sum_{k=1}^{\omega_j} \sum_{i>1, i\neq k}^{\omega_j} (c_k - c_i) a_k b_i \tag{82}$$

$$= b_1 \sum_{k=1}^{\omega_j} c_k a_k - c_{1'} b_1 \sum_{k=1}^{\omega_j} a_k + \sum_{k=1}^{\omega_j} c_k a_k \sum_{i>1, i\neq k}^{\omega_j} b_i - \sum_{k=1}^{\omega_j} a_k \sum_{i>1, i\neq k}^{\omega_j} c_i b_i \tag{83}$$

Define

$$V_t = \{\mathbf{X}_{t'} | 0 < t' < t\} \tag{84}$$

That is, V_t contains all the nodes before time point t.

Denote $\{b_{t_i}\}_{i\in[n]} = \cup_h \mathrm{Pa}_h(x_t^j)$ and assume $\{b_{t_i}\}_{1\leq i\leq n_1<n} = \mathrm{Pa}_1(x_t^j)$, where $t \in \Pi^j_{(k,s)}$

We express $p_{t\in\Pi^j_{(k,s)}}(\cup_h \mathrm{Pa}_h(x_t^j))$ by marginalizing all other random variables occurring before the latest variables in $\cup_h \mathrm{Pa}_h(x_t^j)$ and utilizing the Causal Markov assumption (**A2**):

$$p_{t\in\Pi^j_{(k,s)}}(\cup_h \mathrm{Pa}_h(x_t^j)) \tag{85}$$

$$= p(\cup_h \mathrm{Pa}_h(x_t^j) | t \in \Pi^j_{(k,s)}) \tag{86}$$

$$= \sum_{V_h \setminus \{b_{t_i}\}_{i\in[n]}} p(b_{t_1}, b_{t_2}, ...b_{t_n}, V_h \setminus \{b_{t_i}\}_{i\in[n]} | h = \max\{t_i, 1 \leq i \leq n\}, t \in \Pi^j_{(k,s)}) \tag{87}$$

$$= \sum_{\{\mathrm{Pa}(b_{t_i})\}_{i\in[n]}} p(b_{t_i}|\mathrm{Pa}(b_{t_i})) \sum_{V_{\tau_{\max}}} \sum_{\tau_{\max}<t'\leq h} \sum_{j\in[n]} \sum_{x^j_{t'}, \mathrm{Pa}(x^j_{t'})} p\big(x^j_{t'}|\mathrm{Pa}(x^j_{t'})\big) p(V_{\tau_{\max}}) \tag{88}$$

Note that $x^j_{t'} \in V_h \setminus \{b_{t_i}\}_{i\in[n]}$.

This joint distribution is now represented by conditional distributions of one related variable given its parents.

Similarly, assume $y = b_{t_1}$, we have

$$p_{t\in\Pi^j_{(k,s)}}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus y) \tag{89}$$

$$= p_{t\in\Pi^j_{(k,s)}}(\cup_h \mathrm{Pa}_h(x_t^j) \setminus b_{t_1}) \tag{90}$$

$$= p(\cup_h \mathrm{Pa}_h(x_t^j \setminus y) | t \in \Pi^j_{(k,s)}) \tag{91}$$

$$= \sum_{V_{t_n} \setminus \{b_{t_i}\}_{i\neq 1}} p(b_{t_2}, ...b_{t_{n-1}}, b_{t_n}, V_{t_n} \setminus \{b_{t_i}\}_{i\neq 1} | h = \max\{t_i, 2 \leq i \leq n\}) \tag{92}$$

$$= \sum_{\{\mathrm{Pa}(b_{t_i})\}_{i\neq 1}} p(b_{t_i}|\mathrm{Pa}(b_{t_i})) \sum_{V_{\tau_{\max}}} \sum_{\tau_{\max}<t'\leq h} \sum_{j\in[n]} \sum_{x^j_{t'}, \mathrm{Pa}(x^j_{t'})} p\big(x^j_{t'}|\mathrm{Pa}(x^j_{t'})\big) p(V_{\tau_{\max}}) \tag{93}$$

Note that $x^j_{t'} \in V_{t_n} \setminus \{b_{t_i}\}_{i\neq 1}$.

$p_{t\in\Pi_1^j}(x_t^j|\mathrm{Pa}(x_t^j)\setminus y)$ can also be represented by those conditional distributions based on Bayes rule.

$$p_{t\in\Pi_1^j}(x_t^j|\mathrm{Pa}(x_t^j)\setminus b_{t_1}) \tag{94}$$

$$=\frac{p_{t\in\Pi_1^j}(x_t^j, b_{t_2},..,b_{t_{n_1}})}{p_{t\in\Pi_1^j}(b_{t_2},...,b_{t_{n_1}})} \tag{95}$$

$$=\frac{\sum_{b_{t_1}} p_{t\in\Pi_1^j}(x_t^j, b_{t_1},...,b_{t_{n_1}})}{\sum_{b_{t_1}} p_{t\in\Pi_1^j}(b_{t_1},...,b_{t_{n_1}})} \tag{96}$$

$$=\frac{\sum_{b_{t_1}} p_{t\in\Pi_1^j}(x_t^j, \mathrm{Pa}_1(x_t^j))}{\sum_{b_{t_1}} p_{t\in\Pi_1^j}(\mathrm{Pa}_1(x_t^j))} \tag{97}$$

$$=\frac{\sum_{b_{t_1}} p(x_t^j|\mathrm{Pa}_1(x_t^j))p_{t\in\Pi_1^j}(b_{t_1},...,b_{t_{n_1}})}{\sum_{b_{t_1}} p_{t\in\Pi_1^j}(b_{t_1},...,b_{t_{n_1}})} \tag{98}$$

$$=\frac{\sum_{b_{t_1}} p(x_t^j|\mathrm{Pa}_1(x_t^j))\sum_{V_h\setminus\{b_{t_{i\in[n_1]}}\}} p_{t\in\Pi_1^j}(b_{t_1},...b_{t_{n_1}}, V_h\setminus\{b_{t_{i\in[n_1]}}\}|h=\max\{t_{i\in[n_1]}\})}{\sum_{b_{t_1}}\sum_{V_h\setminus\{b_{t_{i\in[n_1]}}\}} p_{t\in\Pi_1^j}(b_{t_1},...b_{t_{n_1}}, V_h\setminus\{b_{t_{i\in[n_1]}}\}|h=\max\{t_{i\in[n_1]}\})} \tag{99}$$

$$=\frac{\sum_{b_{t_1}} AB}{\sum_{b_{t_1}} CD} \tag{100}$$

where

$$A = p(x_t^j|\mathrm{Pa}_1(x_t^j))\sum_{\{\mathrm{Pa}(b_{t_i})\}_{i\in[n_1]}} p(b_{t_i}|\mathrm{Pa}(b_{t_i})) \tag{101}$$

$$B = \sum_{V_{\tau_{\max}}}\sum_{\tau_{\max}<t'\leq h}\sum_{j\in[n]}\sum_{x_{t'}^j,\mathrm{Pa}(x_{t'}^j)} p\big(x_{t'}^j|\mathrm{Pa}(x_{t'}^j)\big)p(V_{\tau_{\max}}) \tag{102}$$

$$C = \sum_{\{\mathrm{Pa}(b_{t_i})\}_{i\in[n_1]}} p(b_{t_i}|\mathrm{Pa}(b_{t_i})) \tag{103}$$

$$D = \sum_{V_{\tau_{\max}}}\sum_{\tau_{\max}<t'\leq h}\sum_{j\in[n]}\sum_{x_{t'}^j,\mathrm{Pa}(x_{t'}^j)} p\big(x_{t'}^j|\mathrm{Pa}(x_{t'}^j)\big)p(V_{\tau_{\max}}) \tag{104}$$

Note that $t\in\Pi_1^j$ for distributions in above section from Eq.(94) to Eq.(104) and that $x_{t'}^j\in V_h\setminus\{b_{t_i}\}_{i\in[n_1]}$.

Hence, every term in Eq.(83) can be expressed as a function of those conditional distributions. Substituting Eq.(88), Eq.(93) and Eq.(100) in Eq.(83), we have a polynomial equation only composed of conditional distributions $\{p\big(x_{t'}^j|\mathrm{Pa}(x_{t'}^j)\big)\}_{j\in[n],t'\leq t}$ except the joint distribution of the starting points $p(V_{\tau_{\max}})$. Note that the conditional distributions of variables in $\{X_t^j\}_{t\in\Pi_k^j}, j\in[n], k\in[\omega_j]$ are the same. Since sets do not allow duplicate values, set $\{p\big(x_{t'}^j|\mathrm{Pa}(x_{t'}^j)\big)\}_{j\in[n],t'\leq t}$ contains only different conditional distributions. There should be potentially total $\sum_{j=1}^n\omega_j$ different causal mechanisms. The total number of conditional probabilities should be jointly determined by the number of causal mechanisms and also the number of realizations that variables can take. After adjusting those conditional distributions by the linear restriction $\sum_y p(x|y)=1$, all components in the set $\{p\big(x_{t'}^j|\mathrm{Pa}(x_{t'}^j)\big)\}_{j\in[n],t'\leq t}$ are mutually independent, and $p(V_{\tau_{\max}})$ is also independent of all the causal mechanisms because the first starting points are random noises. That is, upon adjustments, all the terms in Eq.(83) should be rendered independent of each other, without any imposed constraints across them.

After expanding all the summations in Eq.(83), the coefficients of this polynomial equation are either 1 or -1. Each coefficient is accompanied by one unique monomial as index (k,s) in the joint distribution $p_{t\in\pi_{k,s}^j}$ determined a unique product of conditional distributions, i.e., with a different pair of (k,s), the product should be different. Considering all random and independent conditional distributions in $\{p\big(x^j_{t'}|\mathrm{Pa}(x^j_{t'})\big)\}_{j\in[n],t'\leq t}$, the polynomial is not identically zero, and the probability of choosing a root of this polynomial is zero.

Denote the polynomial equation in Eq.(83) as A, we have:

$$p(A=0)=0 \tag{105}$$

Back to the original Eq.(73), we finally have $p\Big(p(x^j_t|\cup_h \mathrm{Pa}_h(x^j_t)) \neq p(x^j_t|\cup_h \mathrm{Pa}_h(x^j_t)\setminus y)\Big) = 1,\ \forall y \in \cup_h \mathrm{Pa}_h(x^j_t)$.

□

Lemma D.3. *Let $\widehat{S\mathrm{Pa}}(\mathbf{X}^j_t)$ denote the estimated superset of parent set for $\mathbf{X}^j \in V$ obtained from the Algorithm B1 (line 2). $\{\mathrm{Pa}_k(X^j_t)\}_{k\in[\omega_j]}$ contain the true and illusory parent sets, where ω_j is the true periodicity of $\mathbf{X}^j$. Under assumptions **A1-A7** and with an oracle (infinite sample size limit), we have:*

$$\cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X^j_t) \subseteq \widehat{S\mathrm{Pa}}(X^j_t),\ \forall t\in[\tau_{\max}+1,T]$$

almost surely.

Proof. Assume the contrary, i.e., there exists $s \in \cup_k \mathrm{Pa}_k(X^j_t)\setminus \widehat{S\mathrm{Pa}}(X^j_t)$. From Lemma D.2, we have $X^j_t \not\perp\!\!\!\perp s \,\Big|\, \cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X^j_t)\setminus s$. By the Definition 2.4, we have $\mathrm{Pa}(X^j_t) \subset \cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X^j_t)$. If $s \notin \mathrm{Pa}(X^j_t)$, by the causal Markov property (**A2**), the dependence relation can not be true, because s is a non-descendant of X^j_t. If $s\in \mathrm{Pa}(X^j_t)$, our Algorithm would have concluded that $X^j_t \not\perp\!\!\!\perp s \,\Big|\, \widehat{S\mathrm{Pa}}(X^j_t)$ (line 2) with a consistent CI test, evident from the causal Markov property, contradicting our assumption. Hence, the lemma. □

Lemma D.4. *Let $\widehat{Pa}(X^j_t)$ denote the estimated parent set for $\mathbf{X}^j \in V$ obtained from the Algorithm B1 (line 19) assuming that true ω_j has obtained (line 17). $\{\mathrm{Pa}_k(X^j_t)\}_{k\in[\omega_j]}$ contain the true and illusory parent sets. Under assumptions **A1-A7** and with an oracle (infinite sample size limit), we have:*

$$\widehat{Pa}(X^j_t) = Pa(X^j_t),\ \forall t\in[\tau_{\max}+1,T] \tag{106}$$

almost surely.

Proof. From Lemma D.3,

$$Pa(X^j_t) \subset \cup_{k=1}^{\omega_j}\mathrm{Pa}_k(X^j_t) \subseteq \widehat{S\mathrm{Pa}}(X^j_t),\ \forall t\in[\tau_{\max}+1,T], j\in[n] \tag{107}$$

In Runge et al. [2019], the author proved $\widehat{Pa}(X^j_t) = Pa(X^j_t)$ if we run PCMCI on stationary time series. Using the same logic, we have the following proof.

Suppose $X^i_{t-\tau} \notin \widehat{Pa}(X^j_t)$ but $X^i_{t-\tau}\in Pa(X^j_t)$. With a consistent conditional independence test and correct time partition, the MCI test (line 8 in Algorithm B1) will remove $X^i_{t-\tau}$ from $\widehat{Pa}_{\omega_j}(X^j_t)$ if and only if:

$$X^i_{t-\tau} \perp\!\!\!\perp X^j_t \,\Big|\, \widehat{SPa}(X^j_t)\setminus\{X^i_{t-\tau}\}, \widehat{SPa}(X^i_{t-\tau}) \tag{108}$$

Based on Eq.(107), the rule is equivalent to removing $X^i_{t-\tau}$ from $\widehat{Pa}_{\omega_j}(X^j_t)$ if and only if:

$$\begin{aligned} X^i_{t-\tau} \perp\!\!\!\perp X^j_t \,\Bigg|\, \Bigg\{ & Pa(X^j_t)\setminus\{X^i_{t-\tau}\}, Pa(X^i_{t-\tau}) \,, \\ & \widehat{SPa}(X^j_t)\setminus(Pa(X^j_t)\cup\{X^i_{t-\tau}\}), \widehat{SPa}(X^i_{t-\tau})\setminus Pa(X^i_{t-\tau})\Bigg\} \end{aligned} \tag{109}$$

$$\Rightarrow X^i_{t-\tau} \perp\!\!\!\perp X^j_t \,\Big|\, Pa(X^j_t)\setminus\{X^i_{t-\tau}\}, Pa(X^i_{t-\tau}) \tag{110}$$

Based on Causal Markov Condition assumption (**A2**) and Faithfulness Condition (**A3**), from Eq.(110) we have $X^i_{t-\tau} \notin Pa(X^j_t)$. In other words, if $X^i_{t-\tau} \notin \widehat{Pa}(X^j_t)$ then $X^i_{t-\tau} \notin Pa(X^j_t)$. That is, $Pa(X^j_t) \subseteq \widehat{Pa}(X^j_t)$

Suppose $X^i_{t-\tau} \in \widehat{Pa}(X^j_t)$ but $X^i_{t-\tau} \notin Pa(X^j_t)$. By the contraposition of Faithfulness (**A1**), we know that $X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \Big| \widehat{Pa}(X^j_t) \setminus \{X^i_{t-\tau}\}, \widehat{Pa}(X^i_{t-\tau})$. Denote $W = \left\{\widehat{Pa}(X^j_t) \setminus \{Pa(X^j_t), X^i_{t-\tau}\}\right\} \cup \left\{\widehat{Pa}(X^i_{t-\tau}) \setminus Pa(X^i_{t-\tau})\right\}$. Since $X^i_{t-\tau} \notin Pa(X^j_t)$, based on Causal Markov Condition assumption (**A2**),

$$
\begin{aligned}
& W \cup X^i_{t-\tau} \perp\!\!\!\perp X^j_t \Big| Pa(X^j_t) \\
& \Rightarrow W \cup X^i_{t-\tau} \perp\!\!\!\perp X^j_t \Big| Pa(X^j_t), Pa(X^i_{t-\tau}) \\
& \xRightarrow{\text{Weak Union}} X^i_{t-\tau} \perp\!\!\!\perp X^j_t \Big| \{Pa(X^j_t), Pa(X^i_{t-\tau})\} \cup W \\
& \Rightarrow X^i_{t-\tau} \perp\!\!\!\perp X^j_t \Big| \widehat{Pa}(X^j_t) \setminus \{X^i_{t-\tau}\}, \widehat{Pa}(X^i_{t-\tau})
\end{aligned}
$$

This is contrary to the assumption so that there is no such $X^i_{t-\tau}$ satisfying $X^i_{t-\tau} \in \widehat{Pa}(X^j_t)$ but $X^i_{t-\tau} \notin Pa(X^j_t)$. In other words, if $X^i_{t-\tau} \in \widehat{Pa}(X^j_t)$, then $X^i_{t-\tau} \in Pa(X^j_t)$. That is, $\widehat{Pa}(X^j_t) \subseteq Pa(X^j_t)$. Combined with the previous conclusion that $Pa(X^j_t) \subseteq \widehat{Pa}(X^j_t)$, we have $\widehat{Pa}(X^j_t) = Pa(X^j_t)$.

□

Based on Lemma D.2, Lemma D.3 and Lemma D.4, we can identify the true ω_j for $\mathbf{X}^j$ through Lemma D.5.

Lemma D.5. *Let ω_j denote the true periodicity for $\mathbf{X}^j \in V$ and $\widehat{\mathrm{Pa}}(X^j_{t \in \Pi^j_k})$ denote the estimated parent set for X^j_t obtained from Algorithm B1 where $t \in \Pi^j_k$. Define:*

$$\widehat{\omega}_j = \arg \min_{\omega \in [\omega_{ub}]} \max_{k \in [\omega]} |\widehat{\mathrm{Pa}}(X^j_{t \in \Pi^j_k})| \tag{111}$$

*Under assumptions **A1-A7** and with an oracle (infinite sample limit), we have that $\hat{\omega}_j = \omega_j,\ \forall j \in [n]$ almost surely.*

Proof. Assume the contrary that $\hat{\omega}_j \neq \omega_j$, then in the Algorithm B1, we have an incorrect time partition $\widehat{\Pi}^j$. Hence, CI tests that are performed use samples with different causal mechanisms. $\hat{p}(X^j_t | \cup^{\omega_j}_{k=1} \mathrm{Pa}_k(X^j_t))$ in Eq.(50) is estimated from a mixture of two or more time partition subsets, say Π^j_1 and Π^j_2. We can apply Lemma D.2 where $\cup^{\omega_j}_{k=1}\mathrm{Pa}_k(X^j_t)$ is replaced by $\cup^2_{k=1}\mathrm{Pa}_k(X^j_t)$ and then in Lemma D.3, $\widehat{S\mathrm{Pa}}(X^j_t)$ is replaced by $\widehat{\mathrm{Pa}}_{\hat{\omega}_j}(X^j_t)$ and hence $\cup^2_{k=1}\mathrm{Pa}_k(X^j_t) \subseteq \widehat{\mathrm{Pa}}_{\hat{\omega}_j}(X^j_t)$ where $\widehat{\mathrm{Pa}}_{\hat{\omega}_j}(X^j_t)$ is obtained from samples with t from the mixture of two different partition subsets (line 8). Hence, with $\hat{\omega}_j$, $|\widehat{\mathrm{Pa}}_{\hat{\omega}_j}(X^j_t)| \geq |\cup^2_{k=1} \mathrm{Pa}_k(X^j_t)|$ using mixture samples $t \in \cup^2_{k=1}\Pi^j_k$. However, with true ω_j, we have $|\widehat{\mathrm{Pa}}_{\omega_j}(X^j_t)| = |\mathrm{Pa}(X^j_t)|$ based on Lemma D.4. With Assumption **A6** the Hard Mechanism Change, $|\cup^2_{k=1} \mathrm{Pa}_k(X^j_t)| > |\mathrm{Pa}(X^j_t)|$ so that ω_j always leads to a smaller size of estimated parent sets than $\hat{\omega}_j$, contrary to the definition of $\hat{\omega}_j$. Hence, $\hat{\omega}_j = \omega_j$. □

With those lemmas, we can prove Theorem 1.

Proof. Assuming that a correct ω_j has already been obtained, from Lemma D.4 we have

$$\widehat{Pa}(X^j_t) = Pa(X^j_t),\ \forall t \geq 2\tau_{\text{ub}}, j \in [n]$$

From Lemma D.5, we know that a correct ω_j must be obtained with consistent CI tests, that is, $\hat{\omega}_j = \omega_j, \forall j \in [n]$. Therefore from Algorithm B1, we have

$$\widehat{Pa}(X^j_t) = Pa(X^j_t),\ \forall t \geq 2\tau_{\text{ub}}, j \in [n]$$

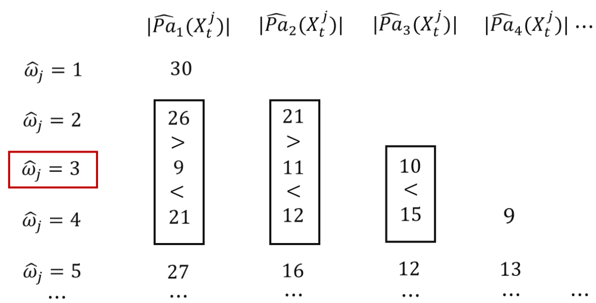


Figure 3: In the above illustration of the "turning point," the sizes of parent sets for different estimates $\hat{\omega}_j$ are depicted as $|\widehat{\mathrm{Pa}}_k(X_t^j)|, k \in [\hat{\omega}_j]$. It is worth noting that $\widehat{\mathrm{Pa}}_k(X_t^j)$ represents either the true parent set or the illusory parent set of X_t^j. In this context, we are interested in the sizes of these parent sets. The first occurrence of the "turning point" happens at $\hat{\omega}_j = 3$ since the sizes of parent sets obtained when $\hat{\omega}_j = 2$ and $\hat{\omega}_j = 4$ are larger than the corresponding size when $\hat{\omega}_j = 3$, respectively. The term "turning point" denotes that as $\hat{\omega}_j$ increases, the size of the parent set initially decreases and then starts increasing once the local minimum is reached. The corresponding relations exist because as long as $\hat{\omega}_j$ is not a multiple of the true ω_j, the estimated time partition subsets with $\hat{\omega}_j$ must be a mixture of some correct time partition subsets with ω_j. Therefore, it is reasonable to use this trick rather than looking at the maximum size of the parent sets $\widehat{\mathrm{Pa}}_k(X_t^j), k \in [\hat{\omega}_j]$ (line 17 in Algorithm B1).

If the causal mechanism is fixed across time, i.e., $\omega_j = 1, j \in [n]$, the proof of PCMCI Runge et al. [2019] showed that for all $\mathbf{X}^j \in V$,

$$X_{t-\tau}^i \to X_t^j \notin \mathcal{G} \Rightarrow X_{t-\tau}^i \to X_t^j \notin \widehat{\mathcal{G}}$$

$$X_{t-\tau}^i \to X_t^j \in \mathcal{G} \Rightarrow X_{t-\tau}^i \to X_t^j \in \widehat{\mathcal{G}}$$

Therefore $\widehat{\mathcal{G}} = \mathcal{G}$.

If $\exists \omega_j > 1$, we can simply separate the whole graph $\mathcal{G}$ into sub graphs $\{\mathcal{G}_k^{\omega_j}\}_{k \in [\omega_j]}$ consisting of only target variable X_t^j with corresponding $t \in \{\Pi_k^j\}_{k \in [\omega_j]}$ and parent variables $X_{t'}^i \in \mathrm{Pa}(X_t^j)$. Focusing only on one time partition subset $\Pi_k^j, k \in [\omega_j]$, we have

$$\widehat{\mathcal{G}}_k^{\omega_j} = \mathcal{G}_k^{\omega_j} \tag{112}$$

for any $k \in [\omega_j]$ and $j \in [n]$ based on the proof of Proposition 1 in the supplementary materials of Runge et al. [2019].

Each sub-graph $\mathcal{G}_k^{\omega_j}$ includes only variable X_t^j, the edges entering X_t^j for time points $t \in \Pi_k^j$ and the corresponding parent variables $X_{t'}^i \in Pa(X_t^j)$. Given $\Pi^j = \bigcup\limits_{k \in [\omega_j]} \Pi_k^j$ and $V = \bigcup\limits_{j \in [n]} \mathbf{X}^j$, we have:

$$\widehat{\mathcal{G}} = \bigcup_{j \in [n],\ k \in [\omega_j]} \widehat{\mathcal{G}}_k^{\omega_j} \tag{113}$$

$$\mathcal{G} = \bigcup_{j \in [n],\ k \in [\omega_j]} \mathcal{G}_k^{\omega_j} \tag{114}$$

On the basis of Eq.(112), we finally have:

$$\widehat{\mathcal{G}} = \mathcal{G}$$

□

E Turning Points

Given infinite samples, our estimate $\hat{\omega}_j$ (line 17 in Algorithm B1) is the exact value ω_j (see Lemma D.5). However, for finite samples, estimating ω_j by the equation in line 17 in Algorithm B1 does not yield good performance when T is small. While searching, larger guesses ω lead to finer time partitions in Π_j, resulting in smaller sizes for Π_j^k (see Line 5 in Algorithm B1). Due to the power limit of CI tests on a smaller sample given by Π_j^k, the number of false negative edges increases. In order to solve this issue, we introduce *turning points*. A *turning point* is a guess $\hat{\omega}$ satisfying:

$$\max_t |\widehat{\mathrm{Pa}}_{\hat{\omega}}(X_t^j)| < \min\{\max_t |\widehat{\mathrm{Pa}}_{\hat{\omega}-1}(X_t^j)|, \max_t |\widehat{\mathrm{Pa}}_{\hat{\omega}+1}(X_t^j)|\}$$

where $|\widehat{\mathrm{Pa}}_{\hat{\omega}}(X_t^j)|$ is the estimated parent set for X_t^j with periodicity guess $\hat{\omega}$. See line 19 in Algorithm B1.

We illustrate it with a special example in Fig.3. If there are several turning points, then $\hat{\omega}_j$ is the first turning point. If there is no turning point, then we obtain $\hat{\omega}_j$ using Line 17 of Algorithm B1.

The concept of the turning point is not based on any formal theorem but rather on experimental observations. In experiments, the turning point often corresponds to a multiple of the true periodicity when T is not large. This occurs due to the limitations of CI tests on finite samples. In such cases, the causal graph can still be correct because the estimated time partition remains accurate. In these experiments, the accuracy rate is calculated by considering $\{N\omega_j\}_{N\in\lfloor\frac{\omega_{\text{ub}}}{\omega_j}\rfloor}$ as correct estimations.

F Computational Complexity

Executing the PCMCI algorithm on the entire time series constitutes the initial phase of the proposed approach (Algorithm B1 line 2). The algorithm's worst-case overall computational complexity is $O(n^3\tau_{\text{ub}}^2) + O(n^2\tau_{\text{ub}})$, discussed in Runge et al. [2019]. Here, the symbol n denotes n-variate time series and τ_{ub} represents the upper boundary for time lags.

The subsequent computational load stemming from the remaining components of our algorithm follows a complexity of $O(\omega_{\text{ub}}^2 n^2\tau_{\text{ub}})$. This encompasses the $O(n^2\tau_{\text{ub}})$ complexity associated with conducting Momentary Conditional Independence (MCI) tests on all n univariate time series. The parameter ω_{ub}^2 here arises due to the search procedure involving ω, iterating through values from 1 to ω_{ub} for all n univariate time series.

The runtime of the computation is further influenced by the scaling behavior of the CI test concerning the dimensionality of the conditioning set and the temporal series length T. For further details, see section 5.1 in Runge et al. [2019].

G Experiments

All experiments, including those detailed in the main paper, are conducted on a single node with one core, utilizing 512 GB of memory in the Gilbreth cluster at Purdue University.

Here, we describe how to calculate the metrics (F_1 score, Adjacency Precision, and Adjacency Recall) in our setting. In stationary time series, the output of the causal discovery algorithm is typically an adjacency matrix with dimensions $[n, n, \tau_{\max}+1]$. Within the three-dimensional binary array, the value 1 signifies an edge pointing from one variable to another with a specific time lag, while 0 indicates the absence of an edge. For instance, if element $[i, j, k]$ in the matrix is 1, then there is an edge pointing from X_{t-k}^i to X_t^j. In semi-stationary time series, due to the presence of multiple causal mechanisms, the binary edge matrix is a four-dimensional array with dimensions $[n, \Omega, n, \tau_{\max}+1]$, where Ω is defined as Eq.(7) in the main paper. This expanded binary matrix is constructed based on the edge matrix of each variable $X_t^j, j \in [n]$, through repetition. For instance, if $\Omega = 2\omega j$, setting the third dimension of the large binary matrix to j should yield ω_j potentially different parent sets (including illusory and true parent sets), each appearing twice.

We should have two such binary arrays, one representing the ground truth with dimensions $[n, \Omega, n, \tau_{\max}+1]$ and one obtained from the algorithm with dimensions $[n, \widehat{\Omega}, n, \tau_{\max}+1]$. If the estimator $\widehat{\Omega}$ is incorrect, those two binary arrays will have different sizes, so we can not directly compare them. To solve this problem, we do the same operation and calculate the least common multiple of Ω and $\widehat{\Omega}$. Denoting this least common multiple as LCM(Ω,$\widehat{\Omega}$), we create two four-dimensional binary arrays with dimensions of $[n, \text{LCM}(\Omega, \widehat{\Omega}), n, \tau_{\max}+1]$ based on the true edge array and the estimated edge array, respectively, through repetition. The metrics are then computed by comparing the values in these two arrays.

G.1 More Discussion regarding the Case Study

As stated in the main paper, we express our inability to comment on the significance of the case study results. We open a door for the related experts; if assumptions **A1-A7** are satisfied, the stationary assumption may not hold in this real-world dataset, and such periodicity exists. However, if the finding is not correct from an expert's viewpoint, the following assumptions may be violated:

- Assumption **A4** No Contemporaneous Causal Effects: There is a possibility of potential causal effects from X_t^{ta} to X_t^{cp} that the algorithm is unable to capture.
- Assumption **A6** Hard Mechanism Change combined with limited power of CI tests: If there is a soft mechanism change in the variables, the reliability of the CI test of two variables given their parents will be influenced by the skewed distribution of the parent variables. This effect will be exacerbated by the fact that the sample size will be shrunk by $\hat{\omega}$.

We provide a sound and robust algorithm for experts in various fields who are interested in validating the presence of periodicity within the causal mechanisms specific to their domain.

G.2 Experiments on Continuous-valued Time Series with Exponential Noise

Considering that VARLiNGAM is a temporal extension of LiNGAM and LiNGAM is an algorithm designed for non-Gaussian data, following the work in Pamfil et al. [2020], we also construct experiments on continuous-valued time series data with Exponential noise. Shown as Fig.4(a), the performance of PCMCI$_\Omega$, PCMCI and VARLiNGAM, are quite similar with their performance on Gaussian noise. The recall rate of DYNOTEARS, however, gets worse with Exponential noise.

G.3 Experiments on Binary Time Series

Similar to the process of generating continuous-valued time series, the generation of binary time series also involves three steps. However, the main difference lies in the last two steps. In the third step, we simulate the conditional distributions of each child variable based on all possible combinations of parent variable values. Subsequently, we randomly generate the value of the child variable by considering the corresponding conditional distribution given its parent sets.

For discrete-valued time series, a longer time length is required. To evaluate performance, we conduct a series of experiments following the same methodology as described in section 4.1. Fig.4(b) illustrates the variation in comprehensive performance with respect to $\omega_{\max}$. PCMCI$_\Omega$ demonstrates a similar performance to PCMCI in terms of the $F1$ score, indicating a well-balanced trade-off between precision and recall. This outcome is expected since discrete-valued time series demand larger sample sizes, and the increases in $\omega_{\max}$ negatively impact the power of MCI tests. This observation is further supported by Fig.5(a), where an increase in time length T from 4000 to 12000 does not lead to a significant improvement in the accuracy rate of $\hat{\omega}$, while the accuracy decreases rapidly with higher values of $\omega_{\max}$.

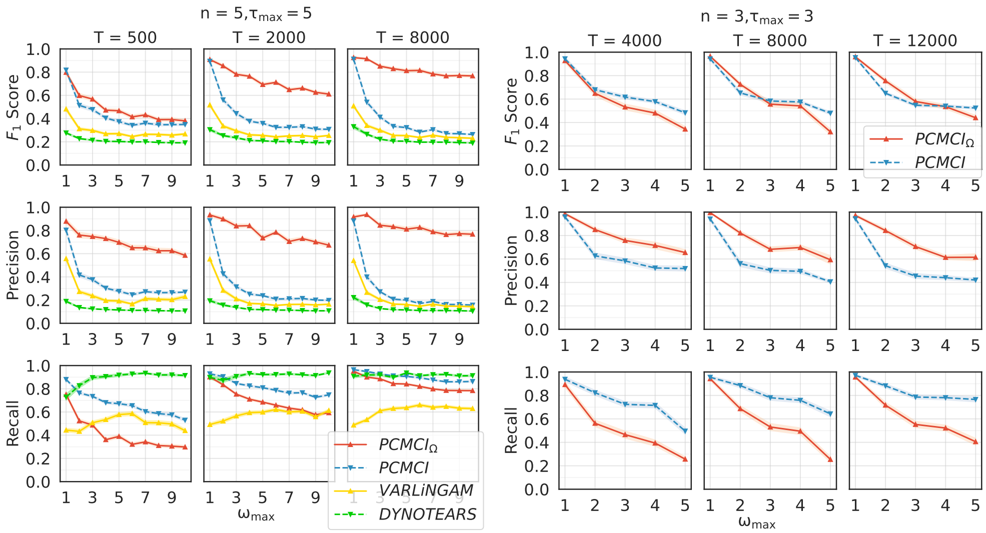
Comparing these results to the experiments conducted on continuous-valued time series, it becomes evident that the demand for efficient samples is even more substantial for binary time series, and the influence of increasing $\omega_{\max}$ on performance becomes more pronounced.

Fig.5(b) shows how the performance of the algorithm varies across $\tau_{\max}$ and the same trade-off between recall and precision has been shown.

G.4 More experiments on Continuous-valued time series

In this section, we conduct more experiments with continuous-valued time series with Gaussian noises.

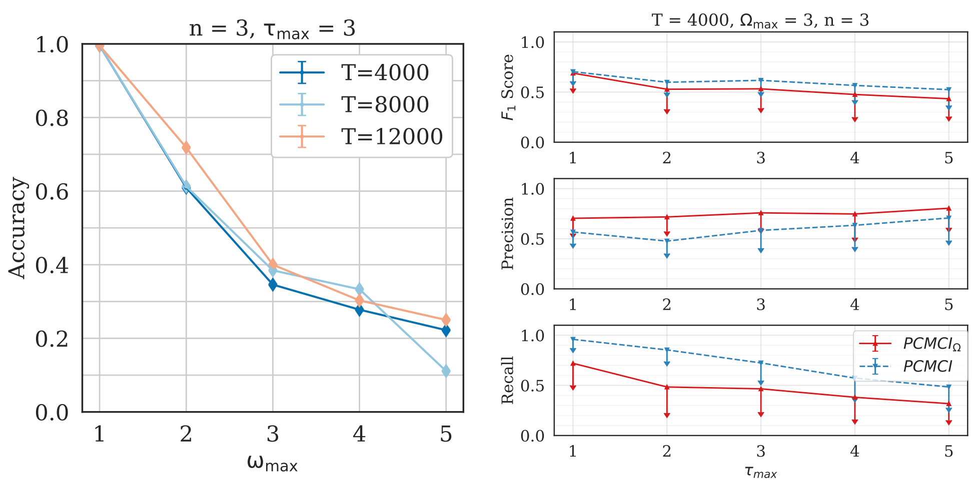
In Fig.6(a), we test our algorithm with and without utilizing the *turning point* rule. See lines 13-14 in Algorithm B1 and section E for more information about the *turning point* rule. Let PCMCI$_\Omega$ TP



(a) Exponential noise

(b) Binary time series

Figure 4: a)F_1 Score, Adjacency Precision, and Adjacency Recall when $\omega_{\max}$ varies for experiments on continuous-valued time series with Exponential noise, length $T = \{500, 2000, 8000\}$, $\tau_{\max} = 5$ and $n = 5$. b) F_1 Score, Adjacency Precision, and Adjacency Recall when $\omega_{\max}$ varies for experiments on binary time series with length $T = \{4000, 8000, 12000\}$, $\tau_{\max} = 3$ and $n = 3$.



(a) Accuracy of $\hat{\omega}$

(b) Performance when $\tau_{\max}$ varies

Figure 5: PCMCI$_\Omega$ is tested on 3-variate binary time series. Every marker corresponds to the average accuracy rate or average running time over 100 trials. a) The accuracy rate of $\hat{\omega}$ for different time series lengths and different $\omega_{\max}$. b) F_1 Score, Adjacency Precision, and Adjacency Recall when $\tau_{\max}$ varies for experiments with time series length $T = 4000$, $\omega_{\max} = 3$ and $n = 3$.

denote the version of PCMCI$_\Omega$ that the *turning point* rule is utilized in choosing ω. PCMCI$_\Omega$ non-TP means that the *turning point* rule is not applied and ω is chosen directly according to Lemma D.5.

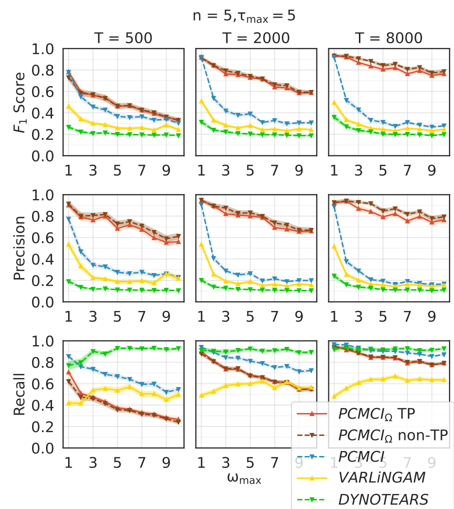
Fig.6(a) shows that the algorithm PCMCI$_\Omega$ non-TP and PCMCI$_\Omega$ TP have similar performance with various T and $\omega_{\max}$. With $T = 500$, PCMCI$_\Omega$ non-TP yields slightly larger standard errors for those metrics, compared to PCMCI$_\Omega$ TP. As time length T increases, the performance of the algorithm PCMCI$_\Omega$ non-TP has consistently increased and is even slightly better than PCMCI$_\Omega$ TP.

The consistent performance of PCMCI_Ω under different chosen rules of ω supports our theoretical result; that is, the correct periodicity leads to the most sparse causal graph.

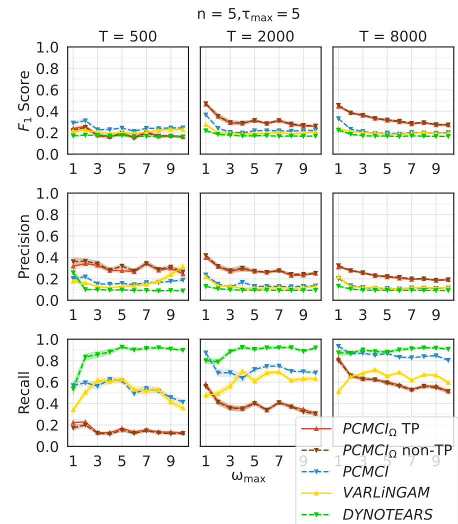
In Fig.6(b), non-stationary time series are produced instead of semi-stationary ones. Consequently, the causal mechanisms for each univariate time series no longer appear sequentially and periodically. The proposed method performs slightly better in terms of F1 score and precision. However, the recall rate is the worst compared to other baselines.

In Fig.6(c), we conduct experiments in the nonlinear setting. The proposed algorithms PCMCI_Ω TP and PCMCI_Ω non-TP perform the best.

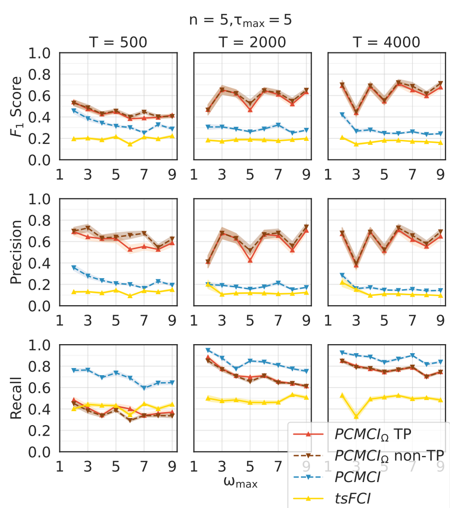
In Fig.6(d), with $\omega_{\text{ub}} < \omega_{\max}$, the performance of the proposed algorithm is significantly worse compared to the scenario where $\omega_{\text{ub}} > \omega_{\max}$. However, with $\omega_{\text{ub}} < \omega_{\max}$, the proposed algorithm can still detect a less dense graph in comparison to other baselines. Based on these outcomes, it is essential to maintain a slightly higher ω_{ub} without significantly impacting the number of efficient samples utilized in each CI test.



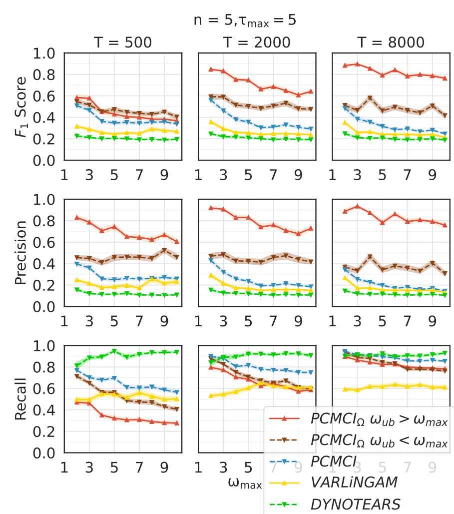
(a) Performance with and without turning point



(b) Performance in non-stationary setting without periodicity



(c) Performance in non-linear SCM



(d) Performance when $\omega_{\text{ub}} > \omega_{max}$ and $\omega_{\text{ub}} < \omega_{max}$

Figure 6: Multiple algorithms are tested on 5-variate time series with different time lengths T. Every line corresponds to a different algorithm. Every marker corresponds to the average performance over 50 trials. In (a), the consistent performance of PCMCI under different chosen rules of ω supports our theoretical result; that is, the correct periodicity ω leads to the most sparse causal graph. In (b), data sets are in a non-stationary setting without periodicity. In (c), the structural causal model (SCM) is non-linear. In (d), algorithm PCMCI$_\Omega$ are tested under conditions that $\omega_{\text{ub}} > \omega_{\text{max}}$ and $\omega_{\text{ub}} < \omega_{\text{max}}$ respectively.

References

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.

Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1483–1495, 2016.

Dimitri Bertsekas and John N Tsitsiklis. *Introduction to probability*, volume 1. Athena Scientific, 2008.

Samuel Karlin. *A first course in stochastic processes*. Academic press, 2014.

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.