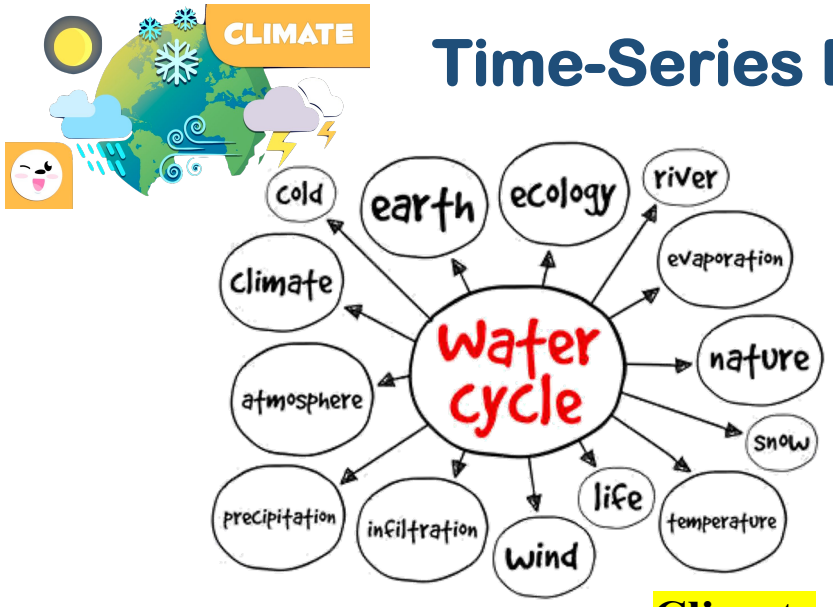
ICML | 2024
Forty-first International
Conference on Machine Learning

Learning Causal Relations from Subsampled Time Series with Two Time-Slices

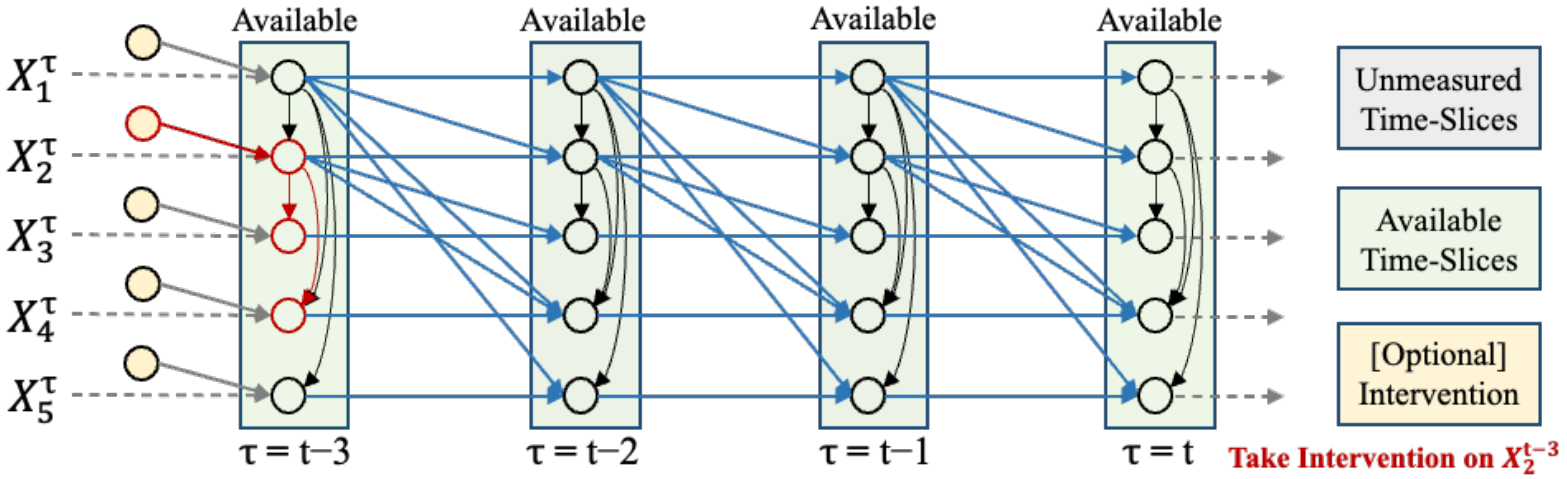
Anpeng Wu[1], Haoxuan Li[2], Kun Kuang[1*], Keli Zhang[3], Fei Wu[145*]

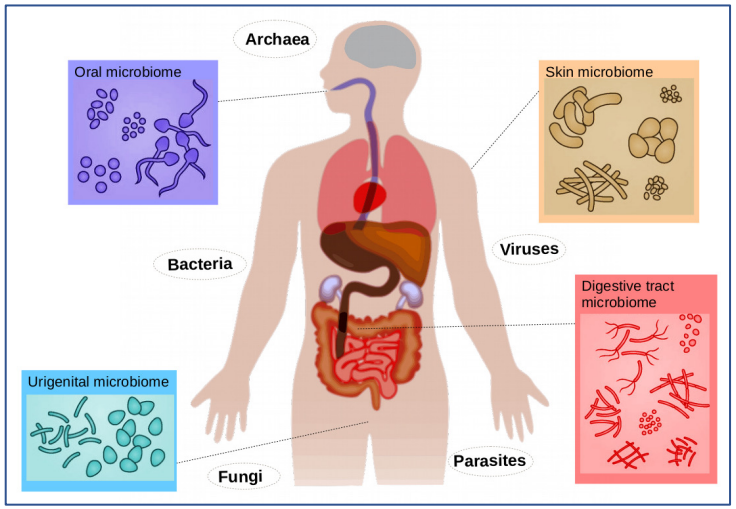
1 Department of Computer Science and Technology, Zhejiang University, Hangzhou, China
2 Center for Data Science, Peking University, Peking, China
3 Huawei Noah's Ark Lab, Huanwei, Shenzhen, China
4 Shanghai Institute for Advanced Study, Zhejiang University, Shanghai, China
5 Shanghai AI Laboratory, Shanghai, China
anpwu@zju.edu.cn
*Corresponding authors

Time-Series Data is ubiquitous in real-applications.

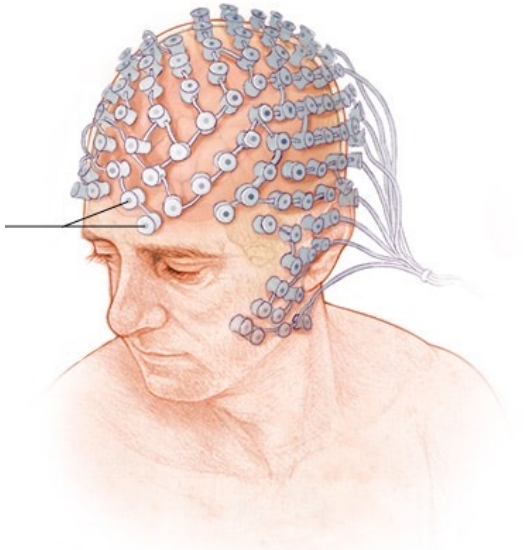


Climate

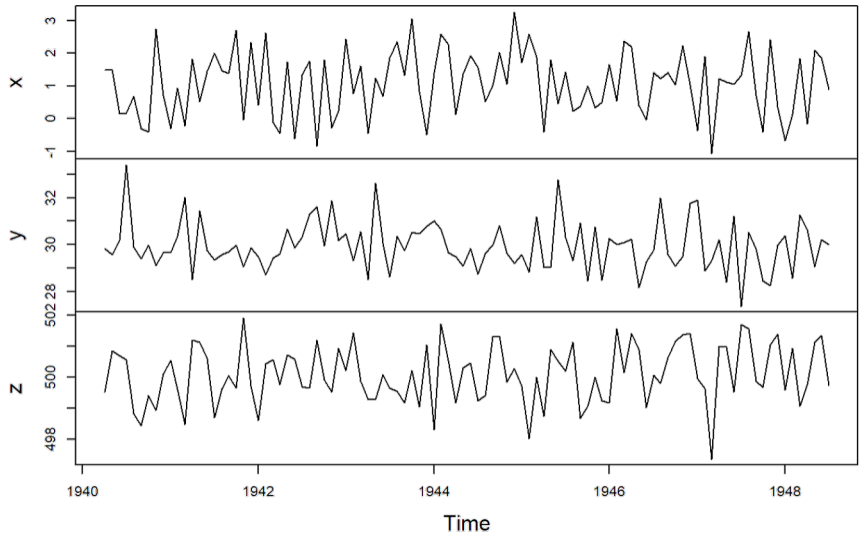




Human Microbiome



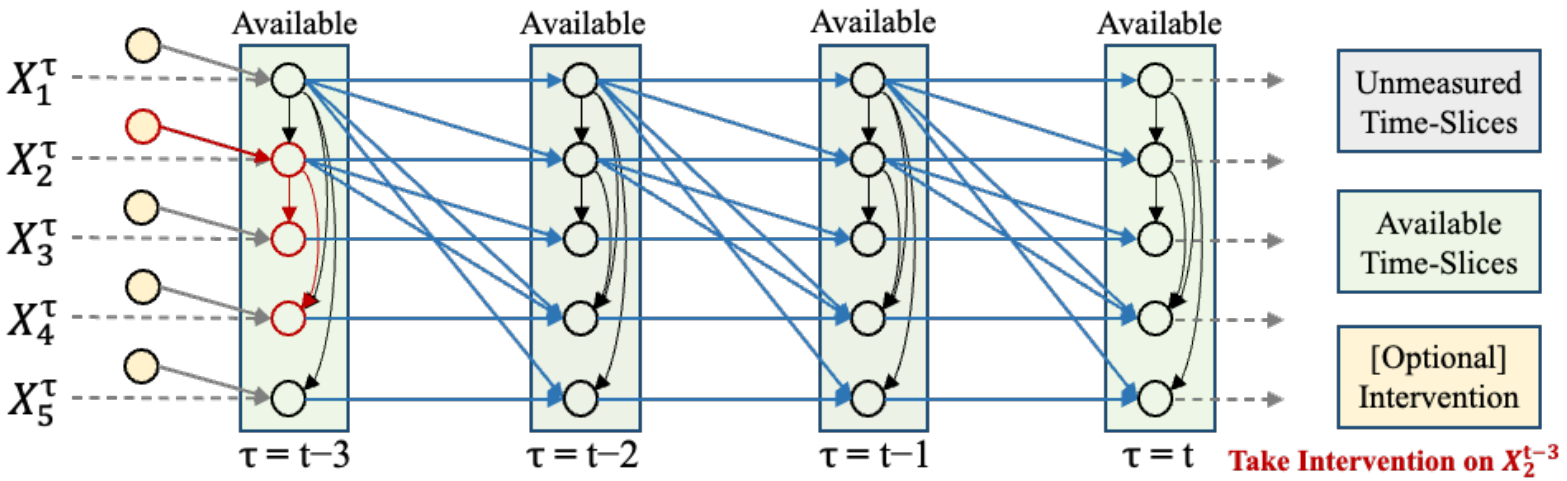
EEG Electrodes



Panel: Multivariable Time Series

Time-Series Data is ubiquitous in real-applications.

Climate



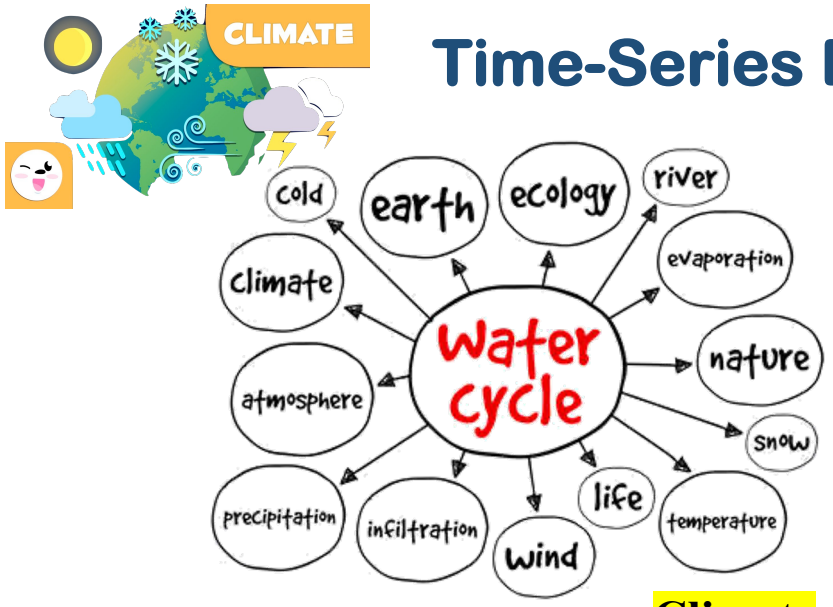
Gold Method: Granger causality, which builds an autoregressive to model causal relations.

Advantages: Granger causality does not require specific assumptions about functional forms or distributions, making it applicable to various types of time series data.

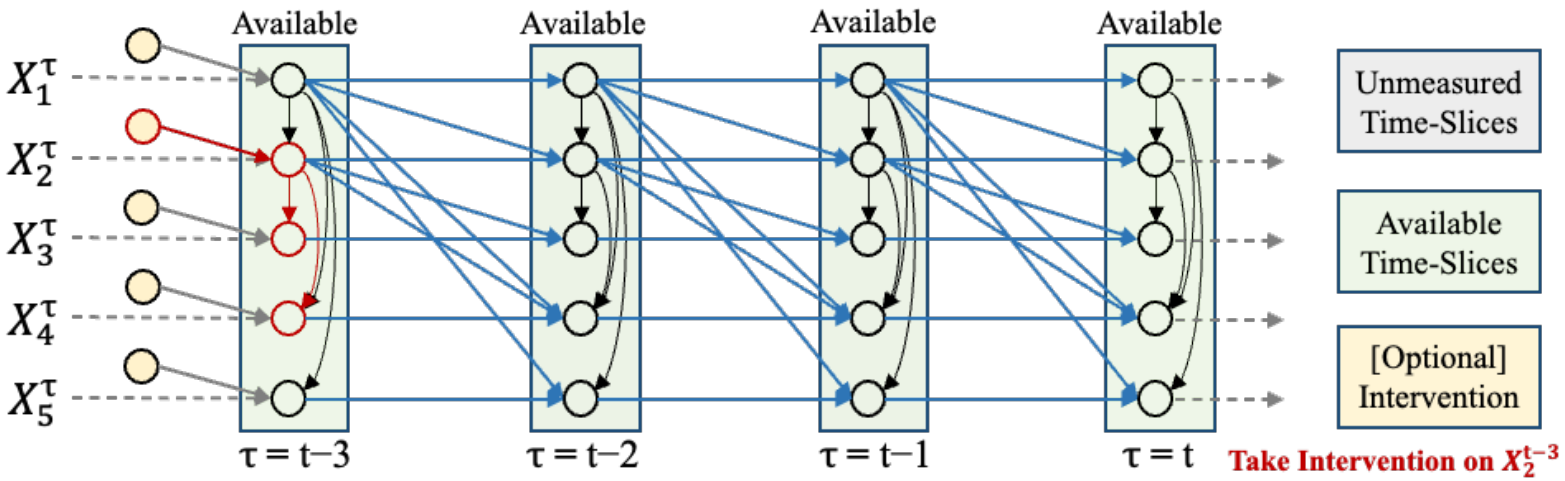
Disadvantages: Granger causality requires

- **Correct Time Lags:** Granger causality relies on lagged terms to model causal relationships.
- **Numerous Time Slices:** Sufficient time slices are needed for accurate inference, and inadequate data length or large time intervals may lead to unreliable results.
- **Without Instantaneous Effects:** Granger causality assumes causal relationships based on lagged terms and cannot capture instantaneous causality, such as simultaneous changes between variables at the same time point.

Time-Series Data is ubiquitous in real-applications.



Climate



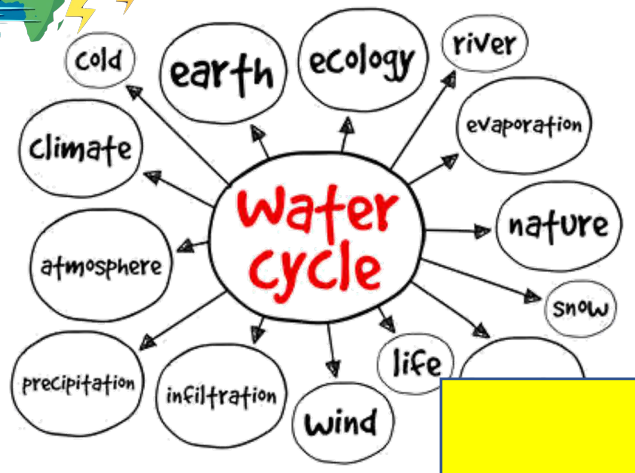
Stational Methods.

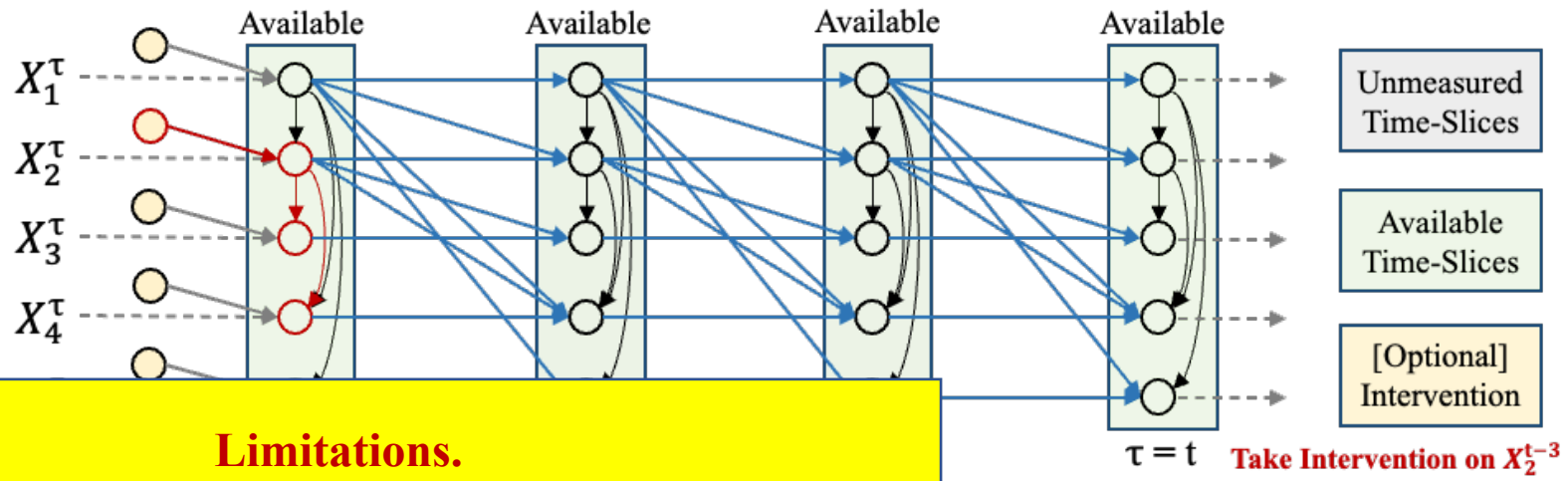
- **Constraint-based methods**:
  - PC, FCI, SGS and ICPs.
- **Score-based methods**:
  - GES and GIES.
- **Continuous-optimization methods**:
  - GraNDAG, GOLEM, NOTEARS, and ReScore.
- **Hybrid methods:**
  - GSP and IGSP.

Temporal Methods.

- **Granger causality and Autoregressive:**
  - PWGC, MVGC, TCDF.
- **Variants of Stational Methods**:
  - PCMCI, oCSE, ANLTSM, tsFCI, SVAE-FCI, VarLiNGAM, DYNOTEARS.
- **Others**:
  - CD-NOD, VarLiNGAM, TiMINo.

Time-Series Data is ubiquitous in real-applications.





Limitations.

- **Some works rely on Linear Assumption;**
- **Some works rely on No Instantaneous Effects;**
- **All works rely on Modeling Causal Structures at the System Timescale.**
- **All works rely on Causal Sufficiency Assumption;**
- **All works rely on Numerous Time Slices Data;**

Stational Methods.

- **Constraint-based**
  - PC, FCI, SGS and
- **Score-based meth**
  - GES and GIES.
- **Continuous-optimization methods**:
  - GraNDAG, GOLEM, NOTEARS, and ReScore.
- **Hybrid methods:**
  - GSP and IGSP.

and Autoregressive:

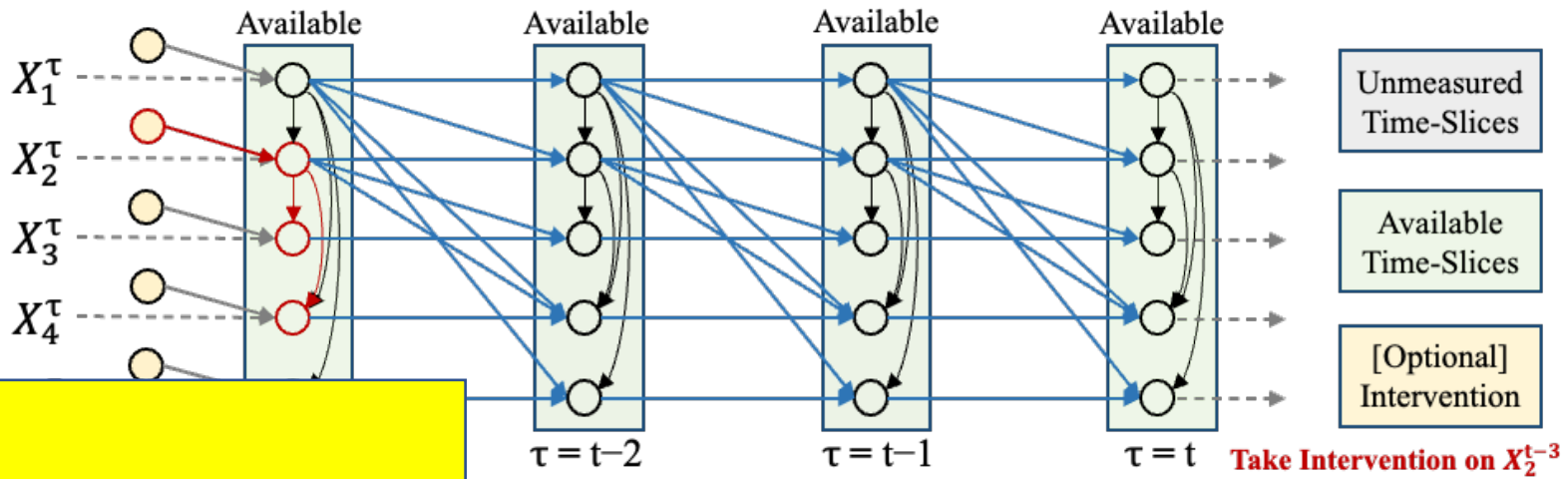
TCDF.

al Methods:

- PCMCI, oCSE, ANLTSM, tsFCI, SVAE-FCI, VarLiNGAM, DYNOTEARS.

- **Others**:
  - CD-NOD, VarLiNGAM, TiMINo.

Subsampled Time-Series Data is ubiquitous in real-applications.



Limitations.

- **Some works rely on Linear Assumption;**
- **Some works rely on No Instantaneous Effects;**
- **All works rely on Modeling Causal Structures at the System Timescale.**
- **All works rely on Causal Sufficiency Assumption;**
- **All works rely on Numerous Time Slices Data;**

Subsampled Time Series.

Measurements are sparse and sampled at a coarser timescale than the causal timescale of the underlying system

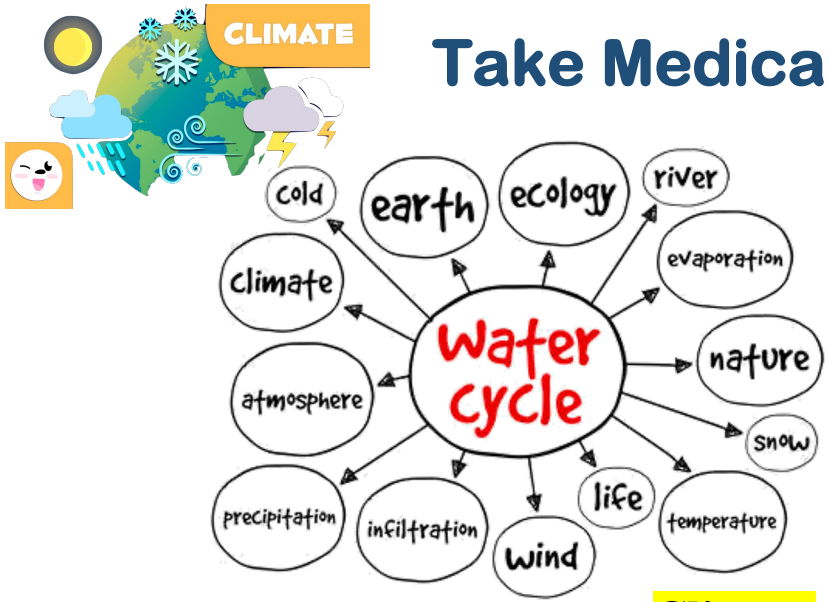
- **Continuous-optimization methods**:
  - GraNDAG, GOLEM, NOTEARS, and ReScore.
- **Hybrid methods:**
  - GSP and IGSP.

Variants

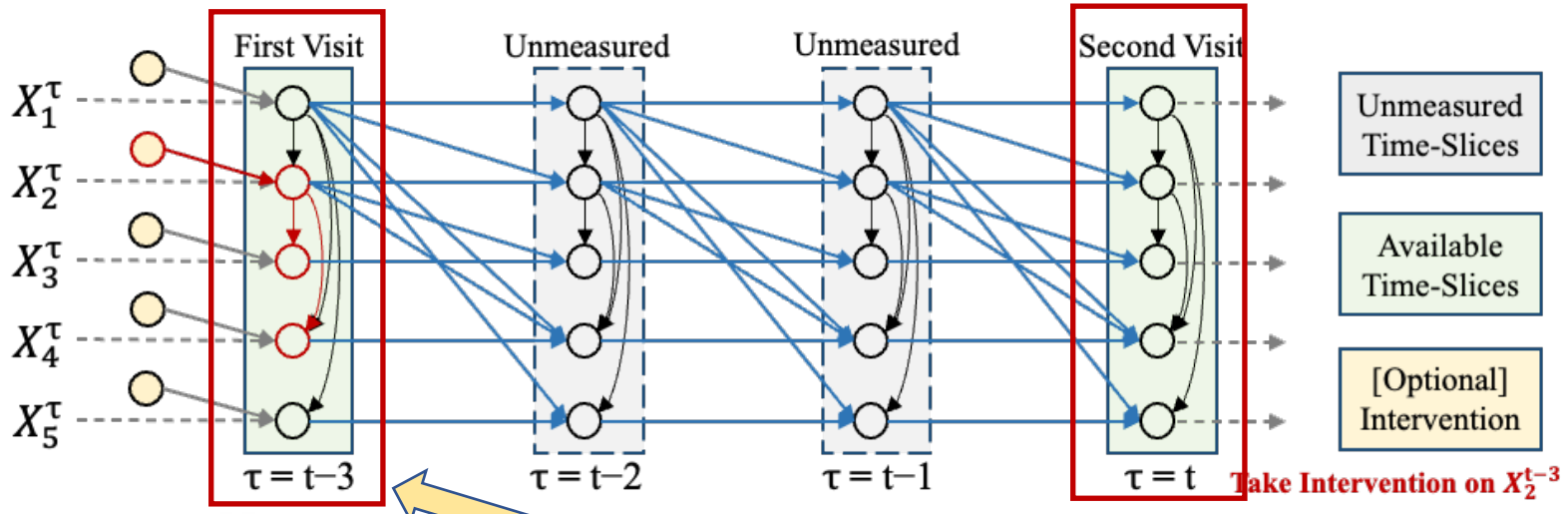
- PCMCI, oCSE, ANLTSM, tsFCI, SVAE-FCI, VarLiNGAM, DYNOTEARS.

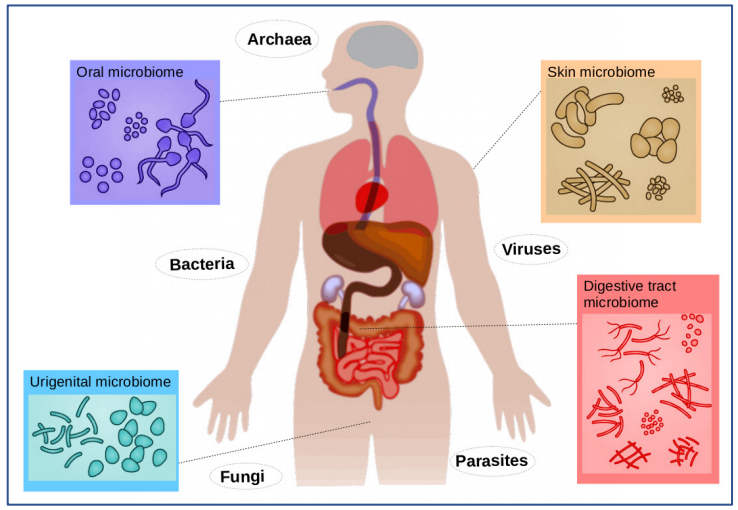
- **Others**:
  - CD-NOD, VarLiNGAM, TiMINo.

Take Medical Scenarios as an Example.



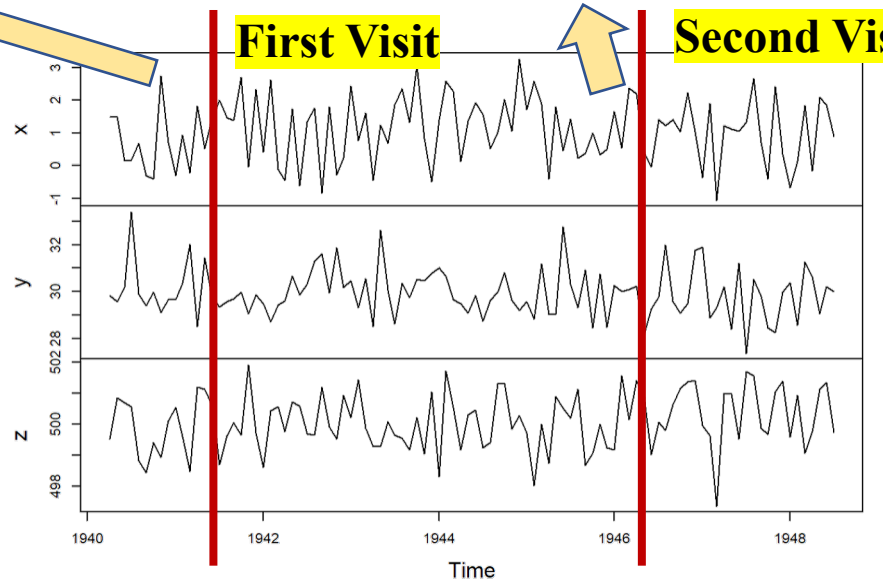
Climate





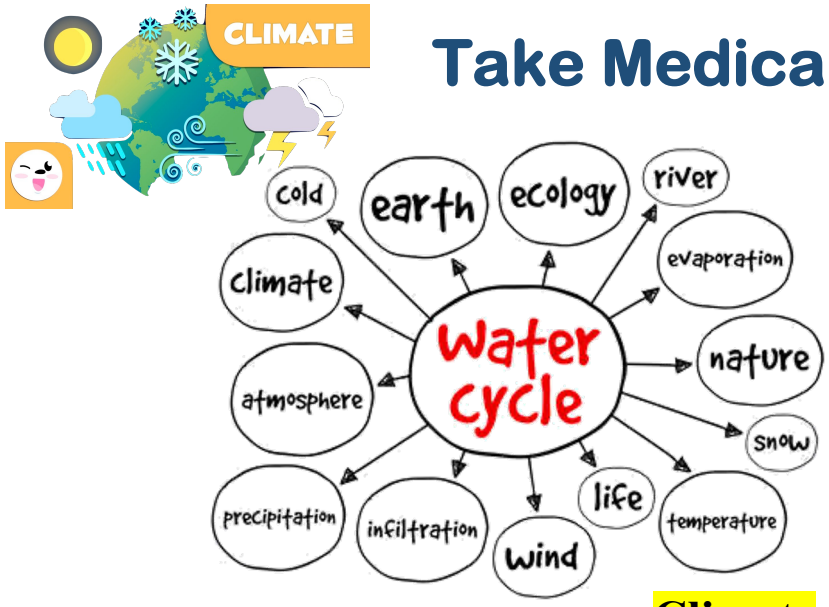
Human Microbiome

EEG Electrodes

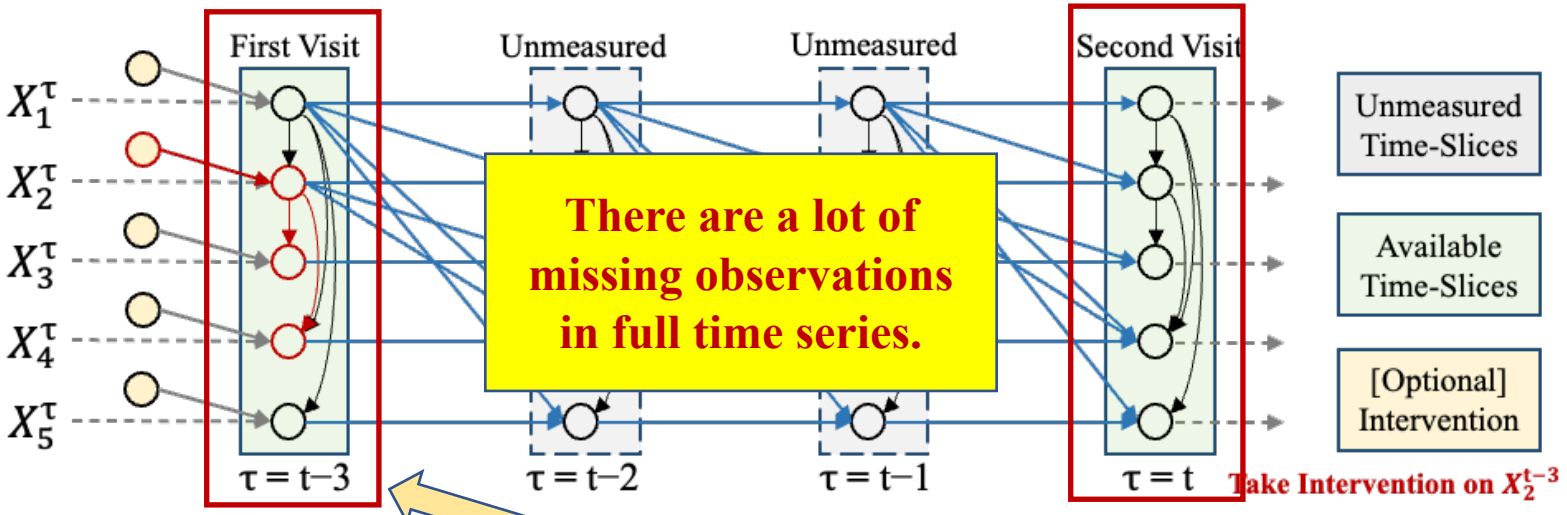


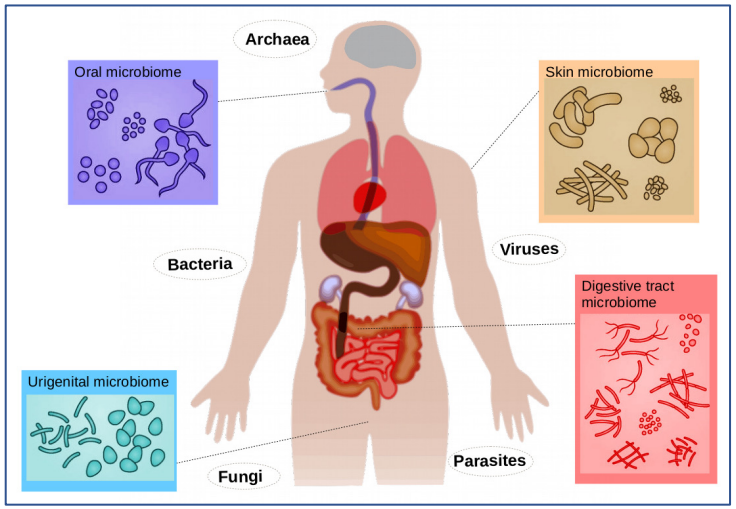
Panel: Multivariable Time Series

Take Medical Scenarios as an Example.



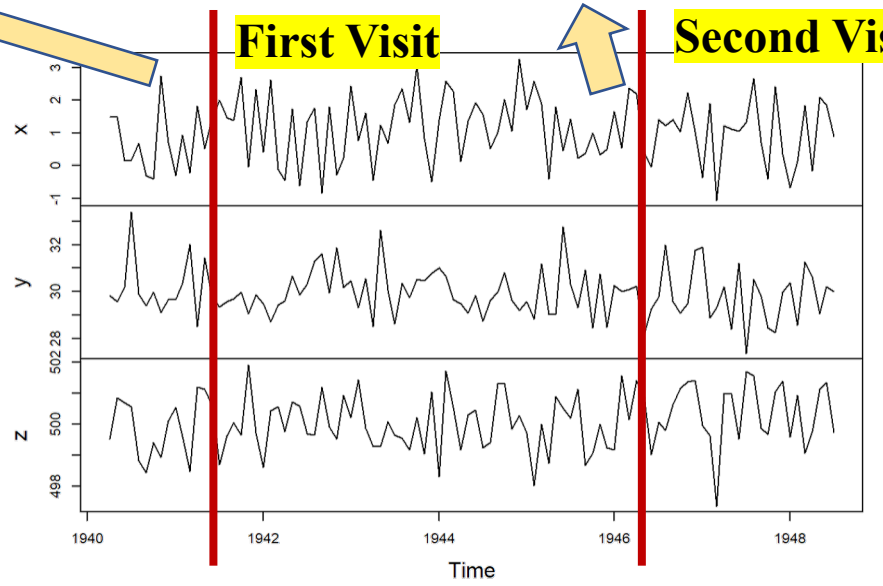
Climate





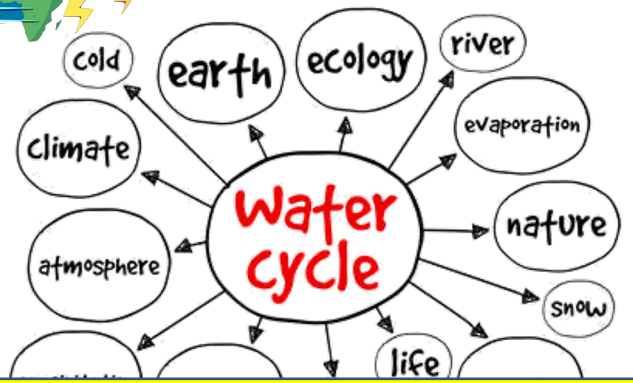
Human Microbiome

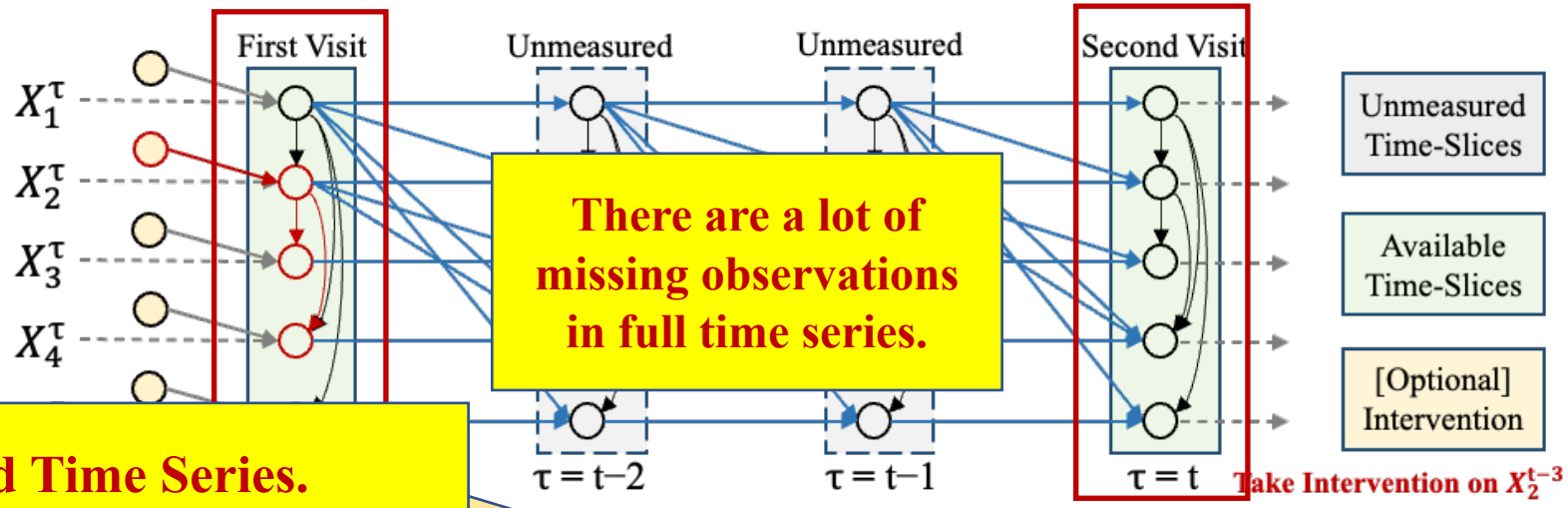
EEG Electrodes



Panel: Multivariable Time Series

Take Medical Scenarios as an Example.

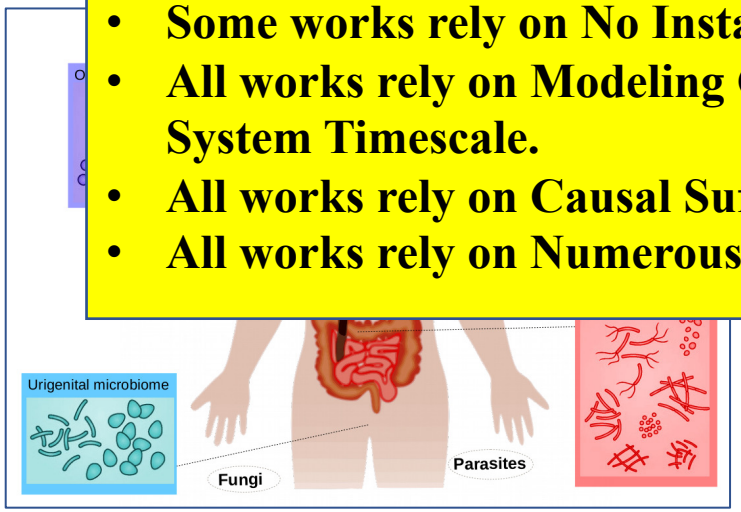




There are a lot of missing observations in full time series.

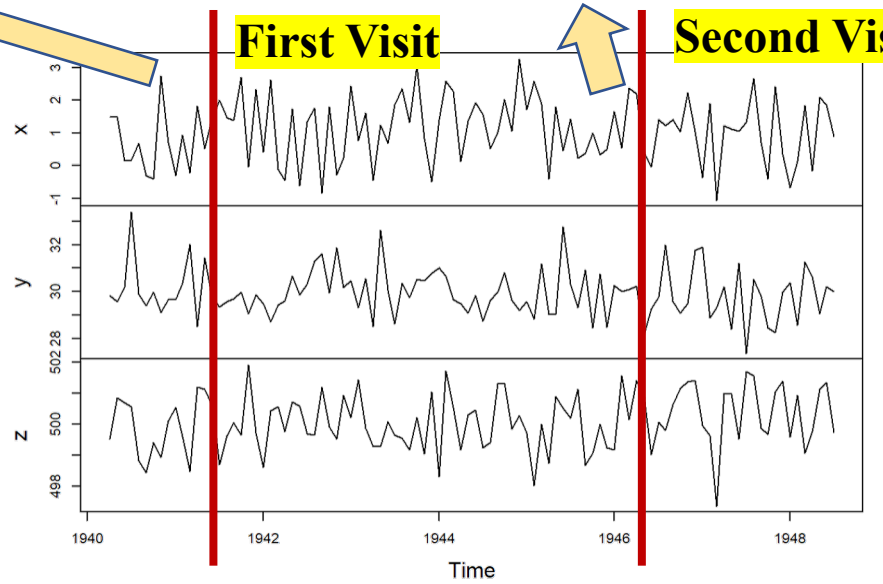
Challenges in Subsampled Time Series.

- Some works rely on Linear Assumption;
- Some works rely on No Instantaneous Effects;
- All works rely on Modeling Causal Structures at the System Timescale.
- All works rely on Causal Sufficiency Assumption;
- All works rely on Numerous Time Slices Data;



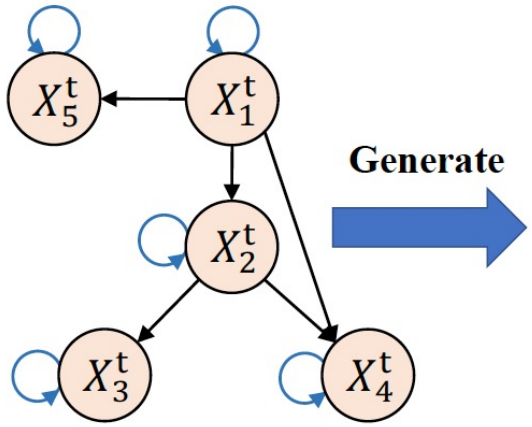
Human Microbiome

EEG Electrodes

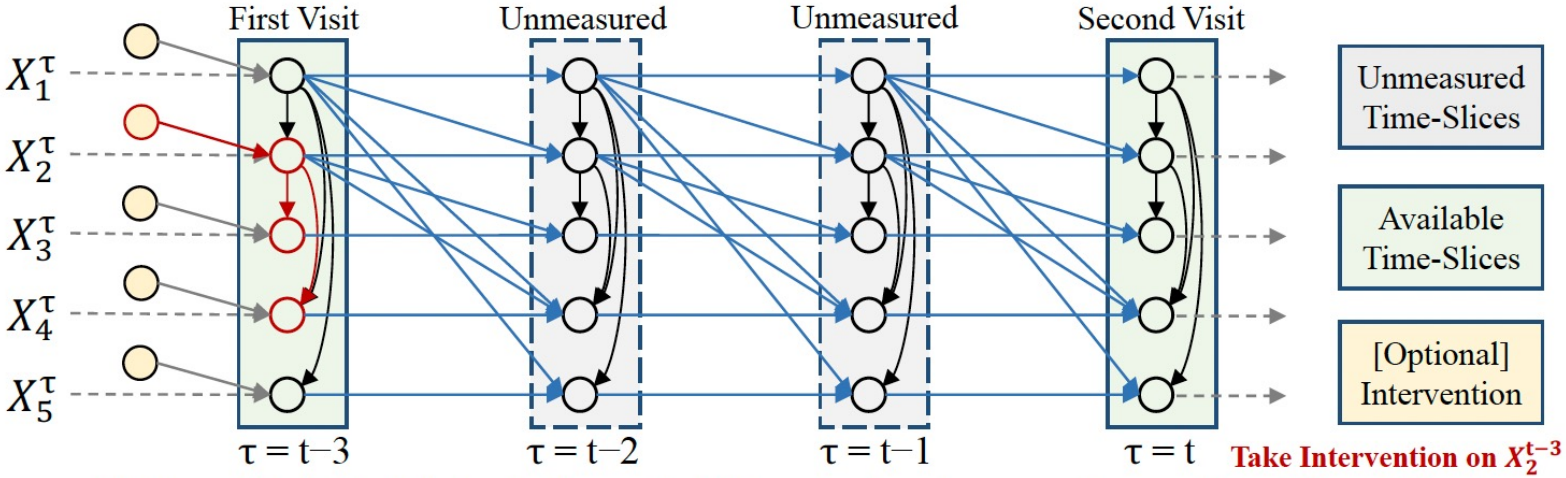


Panel: Multivariable Time Series

Abstract Figure.

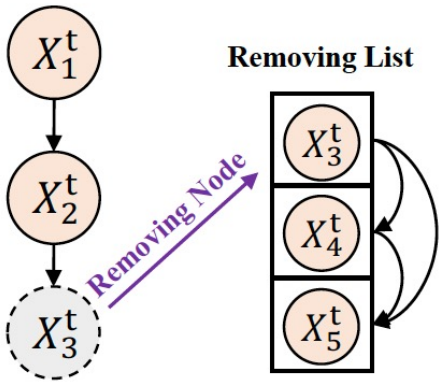


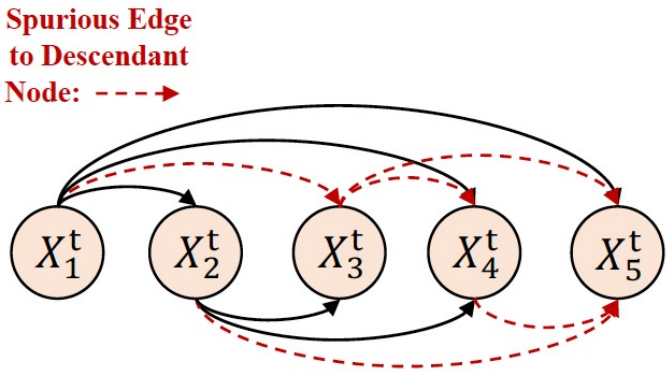
Generate



(a) Summary Causal Graph

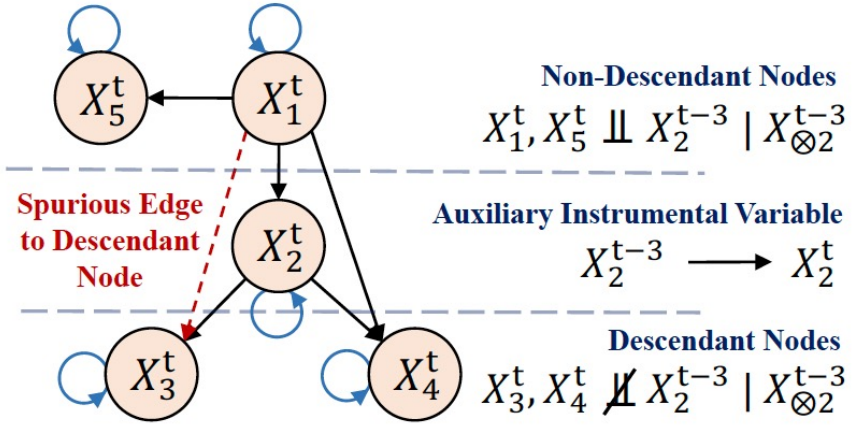
(b) Graphical Models for Subsampled Time-Series (Interventions or Two Time-Slices)





(I) Sequentially Identifying and Removing the Leaves (II) Complete Topological Ordering

(c) Identifying Leave Nodes for Complete Topological Ordering (SCORE)

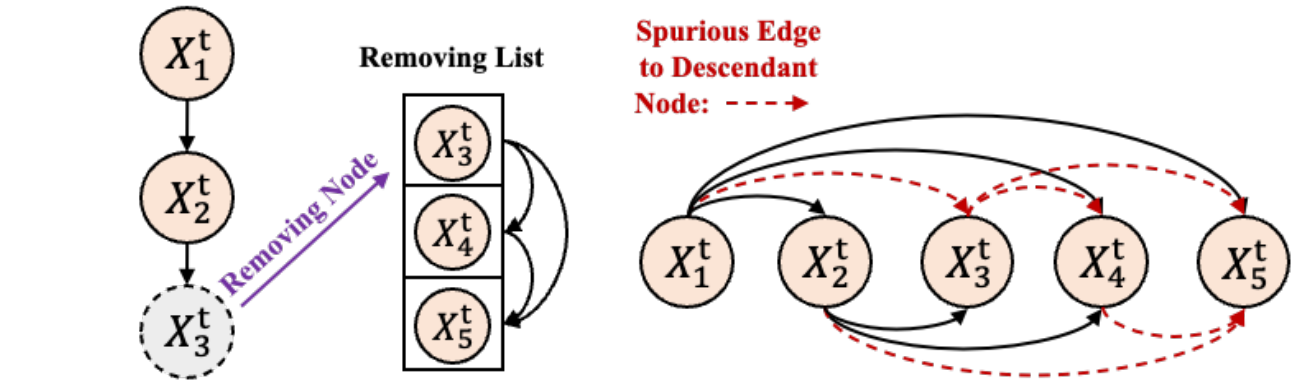


(d) Descendant Hierarchical Topology (Ours)

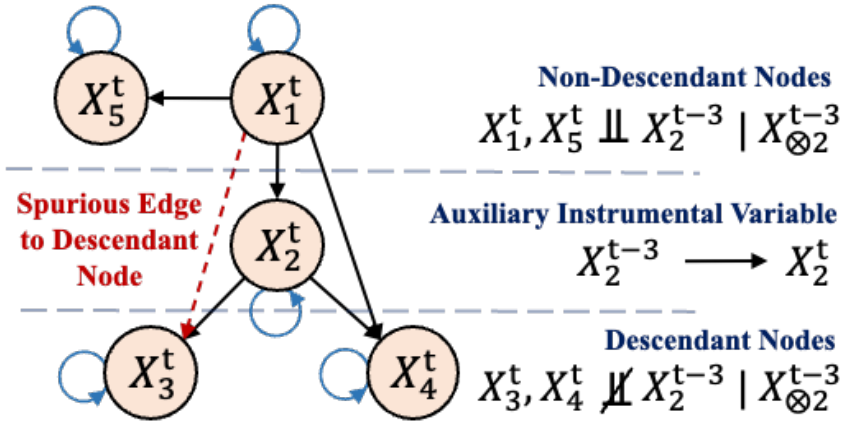
Descendant Hierarchical Topology

Definition 4.1 (Complete Topological Ordering). The complete topological ordering ($\pi(\boldsymbol{X}) = (X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_d})$, π_i is the reordered index of node) is a sorting of all nodes in a DAG such that for any pair of nodes X_{π_i} and X_{π_j}, if there exists a directed edge from X_{π_i} to X_{π_j}, then $i > j$.

Definition 4.2 (Hierarchical Topological Ordering). In the hierarchical topological ordering e.g., $\Pi(\boldsymbol{X}) = (\{X_{\pi_1}\}_{\boldsymbol{L}_1}, \{X_{\pi_2}, X_{\pi_3}\}_{\boldsymbol{L}_2}, \cdots)$, each layer is denoted by $\boldsymbol{L}_i$ and the located layer of X_j are represented as l_j. If there is a directed edge from X_{π_i} to X_{π_j}, then $l_{\pi_i} > l_{\pi_j}$.



(c) Identifying Leave Nodes for Complete Topological Ordering



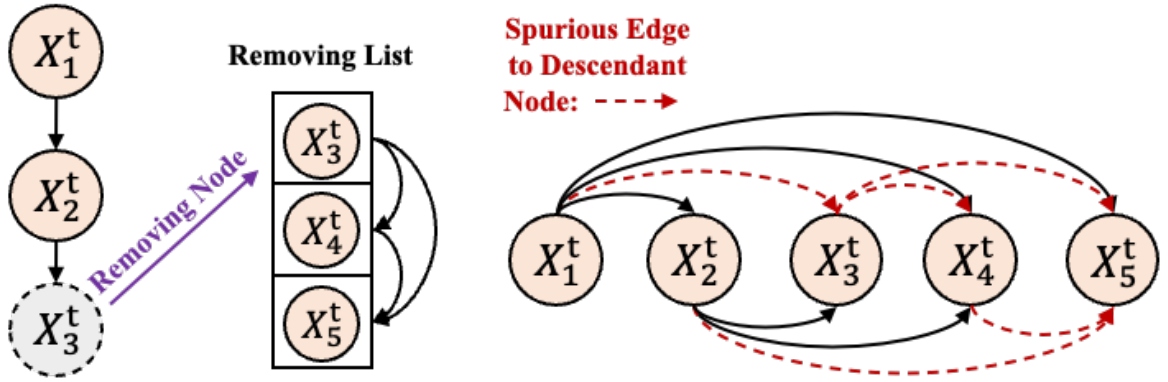
(d) Descendant Hierarchical Topology (Ours)

Descendant Hierarchical Topology

Definition 4.1 (Complete Topological Ordering). The complete topological ordering ($\pi(\boldsymbol{X}) = (X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_d})$, π_i is the reordered index of node) is a sorting of all nodes in a DAG such that for any pair of nodes X_{π_i} and X_{π_j}, if there exists a directed edge from X_{π_i} to X_{π_j}, then $i > j$.

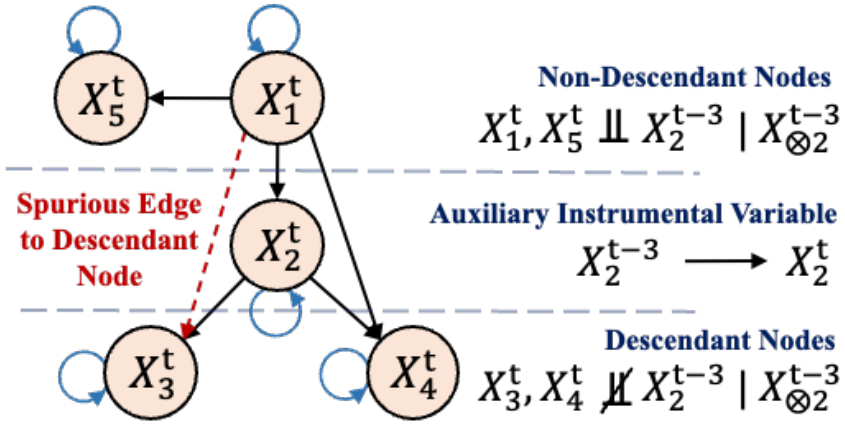
Definition 4.2 (Hierarchical Topological Ordering). In the hierarchical topological ordering e.g., $\Pi(\boldsymbol{X}) = (\{X_{\pi_1}\}_{\boldsymbol{L}_1}, \{X_{\pi_2}, X_{\pi_3}\}_{\boldsymbol{L}_2}, \cdots)$, each layer is denoted by $\boldsymbol{L}_i$ and the located layer of X_j are represented as l_j. If there is a directed edge from X_{π_i} to X_{π_j}, then $l_{\pi_i} > l_{\pi_j}$.

Definition 4.3 (Descendant Hierarchical Topology). In the descendant hierarchical topology, each node X_i^t identifies other nodes as either non-descendant nodes or descendant nodes, and each node X_i^t establishes direct edges pointing to its descendants $\mathbf{de}_i^t$, i.e., $X_i^t \rightarrow \mathbf{de}_i^t, i \in \{1, 2, \cdots, d\}$.



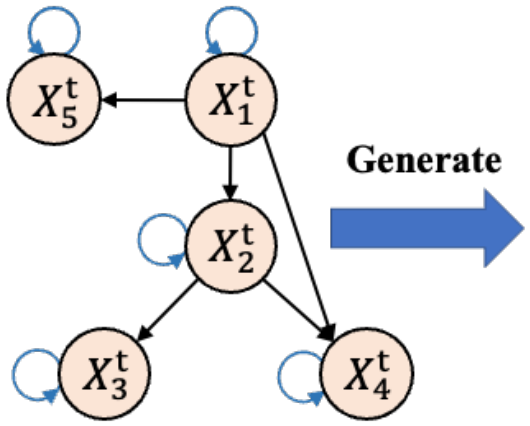
(I) Sequentially Identifying and Removing the Leaves (II) Complete Topological Ordering

(c) Identifying Leave Nodes for Complete Topological Ordering

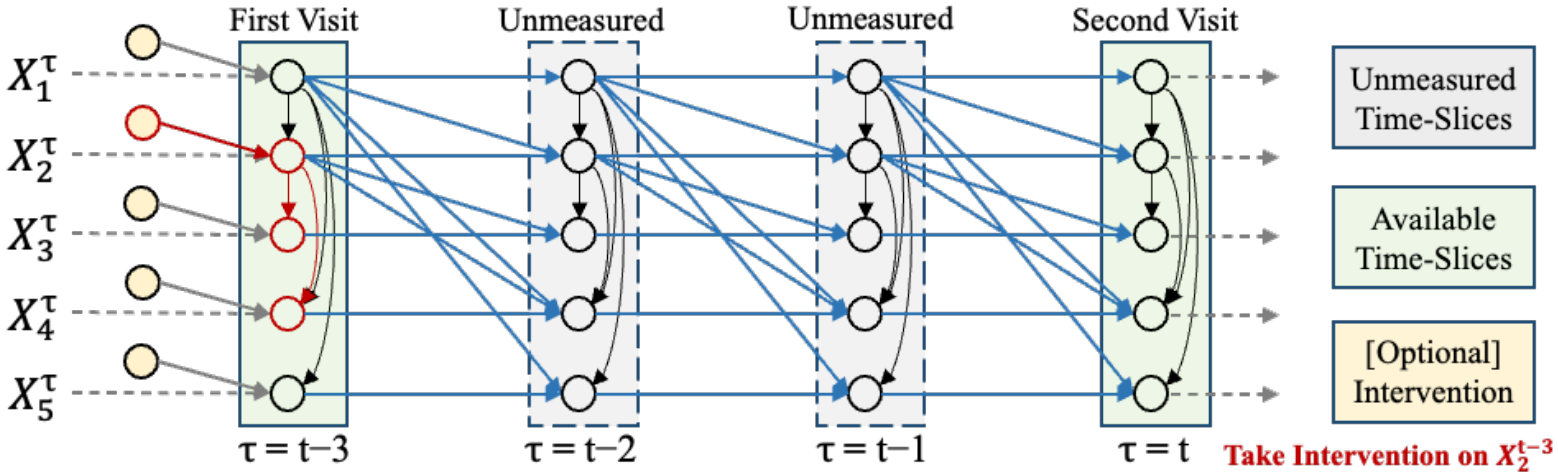


(d) Descendant Hierarchical Topology (Ours)

Using Conditional Instrumental Variables to Replace Interventions.



Generate



(a) Summary Causal Graph **(b) Graphical Models for Subsampled Time-Series (Interventions or Two Time-Slices)**

Assumption 3.1 (Markov Property)**.** The Markov property of time series assumes the future slice $\boldsymbol{X}^{t+1}$ depends on current state $\boldsymbol{X}^t$ but does not depend on history $\boldsymbol{X}^{1\cdots t-1}$.

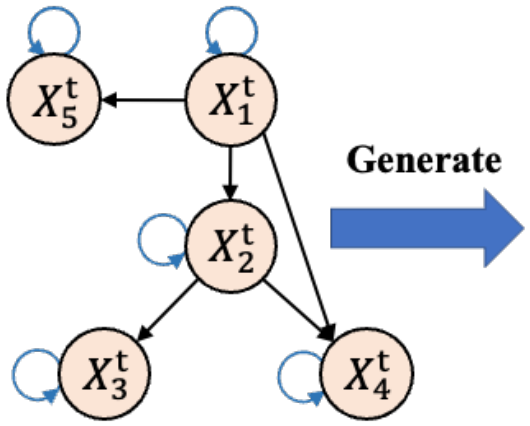
Assumption 3.2 (Acyclic Summary Causal Graph, Section 5.2.1 in Assaad et al. (2022)))**.** The summary causal graph of a time series is considered acyclic if the lagged effect of each variable solely affects its own value and its descendants, without any influence on its non-descendants.

Assumption 3.3 (Consistency Throughout Time, Definition 7 in Assaad et al. (2022)))**.** A causal graph $\mathcal{G}$ for a multivariate time series $\boldsymbol{X}$ is said to be consistent throughout time if all the causal relationships remain constant throughout time, also referred to as stationary full-time graph.

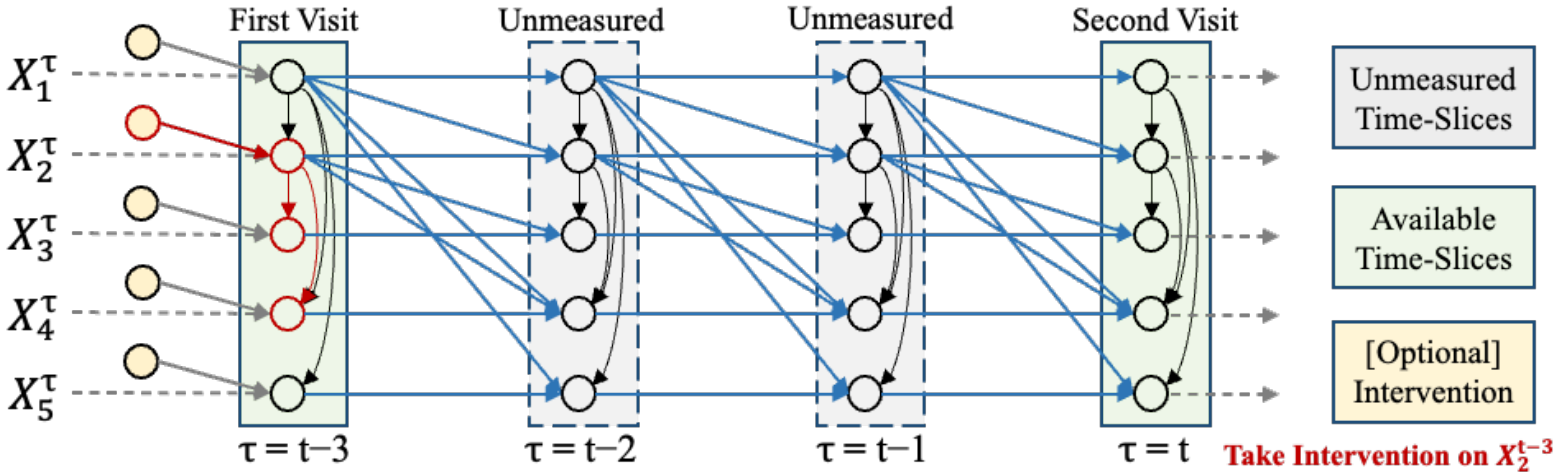
$$X_i^\tau = f_i\left(\mathbf{pa}_i^\tau, X_i^{\tau-1}, \mathbf{pa}_i^{\tau-1}\right) + \epsilon_i^\tau, \qquad (1)$$

where $f_i(\cdot)$ is a twice continuously differentiable function, which embeds the instantaneous effects from its parents $\mathbf{pa}_i^\tau$ at time τ and non-zero time-lagged effects from previous variable $\boldsymbol{X}^{\tau-1}$; and ϵ_i^τ denotes the *Additive Noise* term at time τ. In the generation function $f_i(\cdot)$, we require that the time-lagged effect of $X_i^{\tau-1}$ on X_i^τ is non-zero.

Using Conditional Instrumental Variables to Replace Interventions.



Generate



(a) Summary Causal Graph **(b) Graphical Models for Subsampled Time-Series (Interventions or Two Time-Slices)**

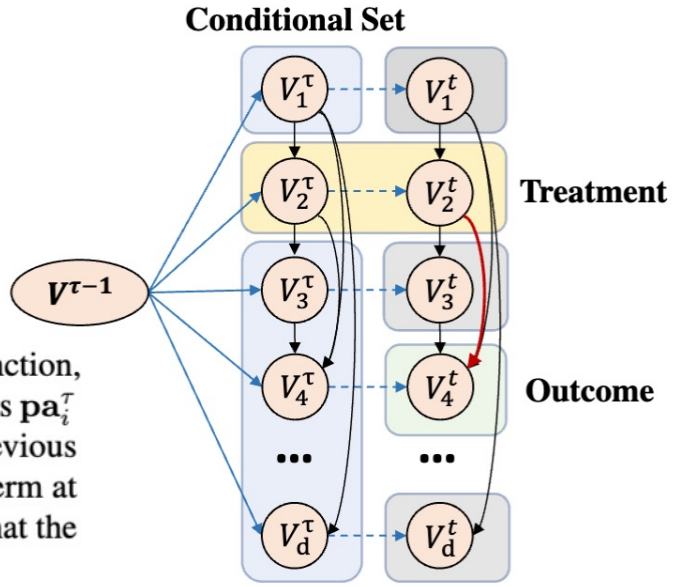
Assumption 3.1 (Markov Property). The Markov property of time series assumes the future slice $\boldsymbol{X}^{t+1}$ depends on current state $\boldsymbol{X}^t$ but does not depend on history $\boldsymbol{X}^{1\cdots t-1}$.

Assumption 3.2 (Acyclic Summary Causal Graph, Section 5.2.1 in Assaad et al. (2022))). The summary causal graph of a time series is considered acyclic if the lagged effect of each variable solely affects its own value and its descendants, without any influence on its non-descendants.

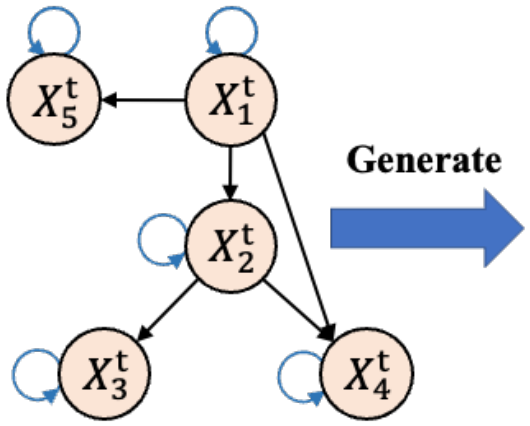
Assumption 3.3 (Consistency Throughout Time, Definition 7 in Assaad et al. (2022))). A causal graph $\mathcal{G}$ for a multivariate time series $\boldsymbol{X}$ is said to be consistent throughout time if all the causal relationships remain constant throughout time, also referred to as stationary full-time graph.

$$X_i^\tau = f_i\left(\mathbf{pa}_i^\tau, X_i^{\tau-1}, \mathbf{pa}_i^{\tau-1}\right) + \epsilon_i^\tau,$$

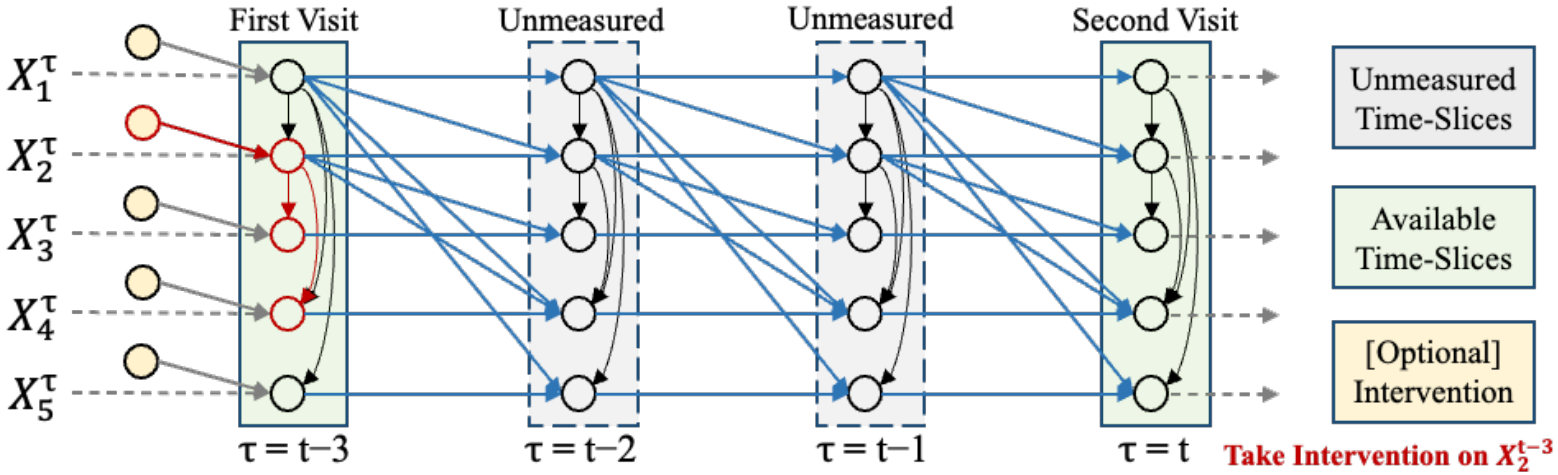
where $f_i(\cdot)$ is a twice continuously differentiable function, which embeds the instantaneous effects from its parents $\mathbf{pa}_i^\tau$ at time τ and non-zero time-lagged effects from previous variable $\boldsymbol{X}^{\tau-1}$; and ϵ_i^τ denotes the *Additive Noise* term at time τ. In the generation function $f_i(\cdot)$, we require that the time-lagged effect of $X_i^{\tau-1}$ on X_i^τ is non-zero.



Descendant-Oriented Conditional Independence Criteria



Generate



(a) Summary Causal Graph

(b) Graphical Models for Subsampled Time-Series (Interventions or Two Time-Slices)

Theorem 4.4 (Descendant-Oriented Conditional Independence Criteria). *Given observations* $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$ *satisfying Assumptions 3.1, 3.2, and 3.3, for variables* $X_i^{t_a}$ *and* $X_i^{t_b}$, *where* $i \in \{1, 2, \cdots, d\}$, *we can conclude that* $X_j^{t_b}$ *is a descendant node of* $X_i^{t_b}$ *iff* $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$.

Descendant-Oriented Conditional Independence Criteria

Theorem 4.4 (Descendant-Oriented Conditional Independence Criteria). *Given observations* $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$ *satisfying Assumptions 3.1, 3.2, and 3.3, for variables* $X_i^{t_a}$ *and* $X_i^{t_b}$, *where* $i \in \{1, 2, \cdots, d\}$, *we can conclude that* $X_j^{t_b}$ *is a descendant node of* $X_i^{t_b}$ *iff* $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$.

Corollary 4.5. *Given observations* $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$, *for variables* X_i *and* X_j *where* $i, j \in \{1, 2, \cdots, d\}$, X_j *is a descendant node of* X_i *iff* $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \boldsymbol{X}_{\otimes i}^{t_a}$.

Proof. From the the non-zero *time-lagged effect* and Assumptions 3.1, 3.2, and 3.3, we can infer that:
(a) The effect of $X_i^{t_a}$ on $X_i^{t_b}$ is non-zero, i.e., $X_i^{t_a} \dashrightarrow X_i^{t_b}$;
(b) Under Markov property, $\boldsymbol{X}^{\tau} \not\dashrightarrow X_i^{t_b}$ for $\tau < t_a < t_b$;
(c) Under acyclic assumption, $X_i^{t_a} \not\dashrightarrow \mathbf{an}_i^{t_b}$ for $t_a < t_b$;
(d) Under stationary time series, $X_i^{t_a} \dashrightarrow X_j^{t_a} \dashrightarrow X_j^{t_b}$.
Under conditions (a), (b), (c) and (d), if $X_j^{t_b} \in \mathbf{an}_i^{t_b}$, then there are only two causal paths between $X_i^{t_a}$ and $X_j^{t_b}$: $X_i^{t_a} \dashleftarrow \mathbf{an}_i^{t_a} \dashrightarrow X_j^{t_b}$ and $X_i^{t_a} \dashrightarrow \{X_i^{t_b}, \mathbf{de}_i^{t_b}\} \dashleftarrow X_j^{t_b}$. Hence, once we cut off all backdoor paths by controlling the conditional set $\mathbf{an}_i^{t_a}$, then the confounding effect between $X_i^{t_a}$ and $X_j^{t_b}$ would be eliminated, leading to $X_i^{t_a} \perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$. Similarity, if $X_j^{t_b} \in \mathbf{sib}_i^{t_b}$, then the summary backdoor path is $X_i^{t_a} \dashleftarrow \mathbf{an}_i^{t_a} \dashrightarrow \mathbf{an}_j^{t_b} \dashrightarrow X_j^{t_b}$. In summary, if $X_j^{t_b}$ is a non-descendant node of $X_i^{t_b}$, then $X_i^{t_a} \perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$. In turn, given the condition $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$, $X_j^{t_b}$ is a descendant node of $X_i^{t_b}$. □

DHT-CIT Algorithm

Algorithm 1 DHT-CIT: Descendant Hierarchical Topology with Conditional Independence Test

Input: Two time-slices $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$ with d nodes; two significance threshold $\alpha = 0.01$ and $\beta = 0.001$ for conditional independence test and pruning process; the layer index $k = 0$.
Output: One adjacency matrix of descendant hierarchical topology $\boldsymbol{A}^{TP}$, one DAG $\mathcal{G}$.
Components: Conditional independence test $\mathbf{HSIC}(\dots)$; and pruning process $\mathbf{CAM}(\cdots)$.
Stage 1 - Identifying Descendant Hierarchical Topology:
for $i = 1$ **to** d **do**
  Construct the conditional set $\boldsymbol{X}^{t_a}_{\otimes i}$ via an independence test $\boldsymbol{X}^{t_a}_{\otimes i} = \{X^{t_a}_j \mid X^{t_a}_j \perp\!\!\!\perp X^{t_a}_i\}$
  for $j = 1$ **to** d **do**
    $p_{i,j} = \mathbf{HSIC}(X^{t_a}_i, X^{t_b}_j \mid \boldsymbol{X}^{t_a}_{\otimes i})$
    $a^{TP}_{i,j} = \mathbb{I}(p_{i,j} \leq \alpha)$
  end for
end for
We obtain $\boldsymbol{P} = \{p_{i,j}\}_{d\times d}$ and $\boldsymbol{A}^{TP} = \{a^{TP}_{i,j}\}_{d\times d}$
Stage 2 - Adjusting the Topological Ordering:
while The causal relationship between the unprocessed nodes is a directed cyclic graph **do**
  $k := k + 1$
  $X_{M_{i,k}} = \{X^{t_a}/X^{t_a}_i, \boldsymbol{L}_{1:k-1}\}$
  $X^{t_b}_i \in \boldsymbol{L}_k$, if $a^{TP}_{i,j} = 0$ for all $j \in M_{i,k}$
  while $\boldsymbol{L}_k = \emptyset$ **do**
    $p_{i^*,j^*} := 2\alpha$ and $a^{TP}_{i^*,j^*} = 0$, $(i^*, j^*) = \arg\max_{i,j}(p_{i,j} \leq \alpha)$
    $X^{t_b}_i \in \boldsymbol{L}_k$, if $a^{TP}_{i,j} = 0$ for all $j \in M_{i,k}$
  end while
  We obtain $\boldsymbol{P} = \{p_{i,j}\}_{d\times d}$ and $\boldsymbol{A}^{TP} = \{a^{TP}_{i,j}\}_{d\times d}$
end while
Stage 3 - Pruning Spurious Edges:
We obtain $\mathcal{G} = \mathbf{CAM}(\mathcal{D}, \boldsymbol{A}^{TP}, \beta)$
Return: $\boldsymbol{A}^{TP}$ **and** $\mathcal{G}$

Data-nodes-edges with different complex functions: Sin-10-10, Sin-20-20, Sigmoid-10-10, Poly-10-10

Table 1. The results (mean$_{\pm std}$) on Sin-*d*-*e* using observational data ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$).

	Sin-10-10 Graph with Observational Data ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$)					**Sin-20-20 Graph with Observational Data** ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$)				
Method	**SHD↓**	**SID↓**	**F1-Score↑**	**Dis.↓**	**#Prune↓**	**SHD↓**	**SID↓**	**F1-Score↑**	**Dis.↓**	**#Prune↓**
PC	$12.8_{\pm 5.03}$	$43.6_{\pm 9.94}$	$0.56_{\pm 0.12}$	$3.51_{\pm 0.72}$	-	$21.5_{\pm 6.75}$	$98.2_{\pm 31.8}$	$0.61_{\pm 0.11}$	$4.59_{\pm 0.69}$	-
FCI	$15.3_{\pm 3.77}$	$71.0_{\pm 11.5}$	$0.54_{\pm 0.09}$	$3.89_{\pm 0.46}$	-	$30.5_{\pm 4.09}$	$237._{\pm 59.1}$	$0.54_{\pm 0.05}$	$5.51_{\pm 0.37}$	-
GOLEM	$\mathbf{0.50}_{\pm 0.80}$	$1.80_{\pm 2.70}$	$\mathbf{0.97}_{\pm 0.03}$	$\mathbf{0.38}_{\pm 0.59}$	-	$1.30_{\pm 1.10}$	$5.60_{\pm 4.40}$	$0.97_{\pm 0.03}$	$0.93_{\pm 0.66}$	-
NOTEARS	$1.20_{\pm 0.60}$	$2.30_{\pm 1.20}$	$0.94_{\pm 0.02}$	$1.02_{\pm 0.30}$	-	$2.60_{\pm 1.49}$	$6.00_{\pm 3.40}$	$0.94_{\pm 0.03}$	$1.55_{\pm 0.46}$	-
ReScore	$1.00_{\pm 0.63}$	$\mathbf{1.40}_{\pm 1.36}$	$0.95_{\pm 0.03}$	$0.88_{\pm 0.47}$	-	$2.00_{\pm 0.77}$	$5.10_{\pm 2.90}$	$0.95_{\pm 0.01}$	$1.38_{\pm 0.28}$	-
Granger	$31.3_{\pm 11.6}$	$66.8_{\pm 30.8}$	$0.21_{\pm 0.04}$	$5.48_{\pm 1.10}$	-	$104_{\pm 20.7}$	$368_{\pm 8.82}$	$0.10_{\pm 0.03}$	$10.1_{\pm 1.01}$	-
VarLiNGAM	$35.0_{\pm 0.00}$	$69.4_{\pm 3.20}$	$0.36_{\pm 0.00}$	$5.91_{\pm 0.00}$	-	$170_{\pm 0.00}$	$339_{\pm 3.20}$	$0.19_{\pm 0.00}$	$13.0_{\pm 0.00}$	-
CD-NOD	$5.40_{\pm 0.92}$	$15.5_{\pm 4.70}$	$0.74_{\pm 0.04}$	$2.32_{\pm 0.19}$	-	-	-	-	-	-
CAM	$3.70_{\pm 2.95}$	$13.2_{\pm 10.6}$	$0.84_{\pm 0.13}$	$1.79_{\pm 0.74}$	$80.00_{\pm 0.00}$	$10.3_{\pm 6.50}$	$41.6_{\pm 34.7}$	$0.79_{\pm 0.12}$	$3.07_{\pm 0.98}$	$360.0_{\pm 0.00}$
SCORE	$5.60_{\pm 3.92}$	$21.2_{\pm 16.1}$	$0.78_{\pm 0.14}$	$2.25_{\pm 0.78}$	$35.80_{\pm 0.98}$	$7.40_{\pm 2.41}$	$31.3_{\pm 21.7}$	$0.85_{\pm 0.04}$	$2.68_{\pm 0.47}$	$172.1_{\pm 0.22}$
DHT-CIT	$1.00_{\pm 1.22}$	$3.20_{\pm 3.70}$	$0.95_{\pm 0.05}$	$0.68_{\pm 0.72}$	$\mathbf{13.20}_{\pm 4.30}$	$\mathbf{1.00}_{\pm 1.32}$	$\mathbf{3.10}_{\pm 4.40}$	$\mathbf{0.98}_{\pm 0.03}$	$\mathbf{0.51}_{\pm 0.61}$	$\mathbf{30.60}_{\pm 7.70}$

* CD-NOD on **Sin-20-20** takes over 5 hours and **#Prune** on one-stage methods is not meaningful. We don't discuss these results and represent them with '-'.

Table 2. The results (mean$_{\pm std}$) on Sigmoid-10-10 & Poly-10-10 data.

	Sigmoid-10-10 data with Gaussian Noise ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$)					**Poly-10-10 data with Gaussian Noise** ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$)				
Method	**SHD↓**	**SID↓**	**F1-Score↑**	**Dis.↓**	**#Prune↓**	**SHD↓**	**SID↓**	**F1-Score↑**	**Dis.↓**	**#Prune↓**
GOLEM	$4.30_{\pm 2.19}$	$18.4_{\pm 7.92}$	$0.78_{\pm 0.11}$	$2.00_{\pm 0.51}$	-	$19.00_{\pm 4.00}$	$59.4_{\pm 13.6}$	$0.20_{\pm 0.12}$	$4.33_{\pm 0.45}$	-
NOTEARS	$12.5_{\pm 5.40}$	$45.3_{\pm 17.9}$	$0.46_{\pm 0.21}$	$3.44_{\pm 0.78}$	-	$17.8_{\pm 5.36}$	$56.4_{\pm 16.9}$	$0.23_{\pm 0.18}$	$4.16_{\pm 0.64}$	-
ReScore	$12.2_{\pm 4.30}$	$45.6_{\pm 14.4}$	$0.45_{\pm 0.17}$	$3.43_{\pm 0.63}$	-	$17.7_{\pm 4.73}$	$57.3_{\pm 14.1}$	$0.22_{\pm 0.15}$	$4.16_{\pm 0.56}$	-
CAM	$3.70_{\pm 3.43}$	$10.4_{\pm 7.86}$	$0.82_{\pm 0.17}$	$1.55_{\pm 1.20}$	$80.00_{\pm 0.00}$	$8.00_{\pm 4.69}$	$19.8_{\pm 7.88}$	$0.63_{\pm 0.21}$	$2.68_{\pm 0.95}$	$80.00_{\pm 0.00}$
SCORE	$9.90_{\pm 3.81}$	$32.8_{\pm 11.6}$	$0.56_{\pm 0.16}$	$3.09_{\pm 0.61}$	$38.90_{\pm 1.60}$	$18.90_{\pm 4.33}$	$40.4_{\pm 10.9}$	$0.23_{\pm 0.13}$	$4.32_{\pm 0.52}$	$42.20_{\pm 1.48}$
DHT-CIT	$\mathbf{0.67}_{\pm 1.12}$	$\mathbf{1.80}_{\pm 2.99}$	$\mathbf{0.96}_{\pm 0.06}$	$\mathbf{0.46}_{\pm 0.72}$	$\mathbf{8.67}_{\pm 2.92}$	$\mathbf{3.22}_{\pm 3.15}$	$\mathbf{10.8}_{\pm 5.69}$	$\mathbf{0.84}_{\pm 0.15}$	$\mathbf{1.51}_{\pm 1.03}$	$\mathbf{11.33}_{\pm 3.87}$

Data-nodes-edges for denser graphs: Sigmoid-10-20 , Sigmoid-10-30, Sigmoid-10-40 , Sigmoid-20-60, Sigmoid-20-100

Table 4. The experiments (mean$_{\pm std}$) on Sigmoid-d-e using observations $D = \{X^{t_a}, X^{t_b}\}$ with Subsampling Rate $u = t_b - t_a$.

	Sigmoid-10-20 on $\mathcal{D} = \{X^1, X^4\}$			**Sigmoid-10-30 on $\mathcal{D} = \{X^1, X^4\}$**			**Sigmoid-10-40 on $\mathcal{D} = \{X^1, X^4\}$**		
Method	**SHD↓**	**SID↓**	**#Prune↓**	**SHD↓**	**SID↓**	**#Prune↓**	**SHD↓**	**SID↓**	**#Prune↓**
GOLEM	$10.70_{\pm 2.93}$	$63.70_{\pm 11.85}$	-	$26.90_{\pm 4.83}$	$71.40_{\pm 8.59}$	-	$35.20_{\pm 3.25}$	$67.00_{\pm 11.20}$	-
SCORE	$15.10_{\pm 3.65}$	$53.10_{\pm 12.95}$	$31.20_{\pm 1.60}$	$14.80_{\pm 5.71}$	$46.40_{\pm 8.04}$	$20.80_{\pm 1.60}$	$23.60_{\pm 2.24}$	$38.40_{\pm 13.46}$	$10.30_{\pm 2.00}$
DHT-CIT	$\mathbf{6.30}_{\pm 2.90}$	$\mathbf{25.30}_{\pm 13.66}$	$\mathbf{13.50}_{\pm 2.91}$	$\mathbf{14.10}_{\pm 4.46}$	$\mathbf{38.80}_{\pm 11.02}$	$\mathbf{11.10}_{\pm 1.58}$	$\mathbf{23.60}_{\pm 2.24}$	$\mathbf{41.20}_{\pm 6.90}$	$\mathbf{6.80}_{\pm 2.48}$
	Sigmoid-20-20 on $\mathcal{D} = \{X^1, X^4\}$			**Sigmoid-20-60 on $\mathcal{D} = \{X^1, X^4\}$**			**Sigmoid-20-100 on $\mathcal{D} = \{X^1, X^4\}$**		
Method	**SHD↓**	**SID↓**	**#Prune↓**	**SHD↓**	**SID↓**	**#Prune↓**	**SHD↓**	**SID↓**	**#Prune↓**
GOLEM	$26.0_{\pm 5.60}$	$138.0_{\pm 47.15}$	-	$60.10_{\pm 5.49}$	$322.3_{\pm 23.84}$	-	$100.0_{\pm 5.32}$	$336.4_{\pm 16.19}$	-
SCORE	$8.40_{\pm 6.20}$	$39.10_{\pm 38.82}$	$173.2_{\pm 1.99}$	$37.10_{\pm 8.14}$	$257.9_{\pm 34.38}$	$144.7_{\pm 4.27}$	$57.5_{\pm 11.00}$	$266.4_{\pm 48.18}$	$112.4_{\pm 3.10}$
DHT-CIT	$\mathbf{0.70}_{\pm 0.90}$	$\mathbf{3.20}_{\pm 3.49}$	$\mathbf{30.10}_{\pm 9.84}$	$\mathbf{22.10}_{\pm 3.75}$	$\mathbf{173.5}_{\pm 38.71}$	$\mathbf{58.8}_{\pm 6.52}$	$\mathbf{53.5}_{\pm 8.43}$	$\mathbf{233.3}_{\pm 31.78}$	$\mathbf{75.4}_{\pm 6.05}$
	Sigmoid-10-20 on $\mathcal{D} = \{X^2, X^4\}$			**Sigmoid-10-20 on $\mathcal{D} = \{X^2, X^6\}$**			**Sigmoid-10-20 on $\mathcal{D} = \{X^2, X^{10}\}$**		
Method	**SHD↓**	**SID↓**	**#Prune↓**	**SHD↓**	**SID↓**	**#Prune↓**	**SHD↓**	**SID↓**	**#Prune↓**
GOLEM	$17.40_{\pm 4.96}$	$58.40_{\pm 13.81}$	-	$21.60_{\pm 4.50}$	$67.20_{\pm 11.41}$	-	$21.80_{\pm 4.87}$	$69.70_{\pm 10.66}$	-
SCORE	$12.20_{\pm 1.78}$	$47.20_{\pm 5.21}$	$29.20_{\pm 0.40}$	$15.80_{\pm 3.76}$	$54.70_{\pm 10.17}$	$30.00_{\pm 1.20}$	$22.30_{\pm 4.43}$	$67.60_{\pm 10.34}$	$32.30_{\pm 2.40}$
DHT-CIT	$\mathbf{8.30}_{\pm 3.82}$	$\mathbf{26.00}_{\pm 12.77}$	$\mathbf{0.46}_{\pm 0.72}$	$\mathbf{8.30}_{\pm 1.55}$	$\mathbf{38.10}_{\pm 5.19}$	$\mathbf{11.20}_{\pm 2.28}$	$\mathbf{14.60}_{\pm 3.75}$	$\mathbf{36.20}_{\pm 12.86}$	$\mathbf{13.60}_{\pm 0.92}$

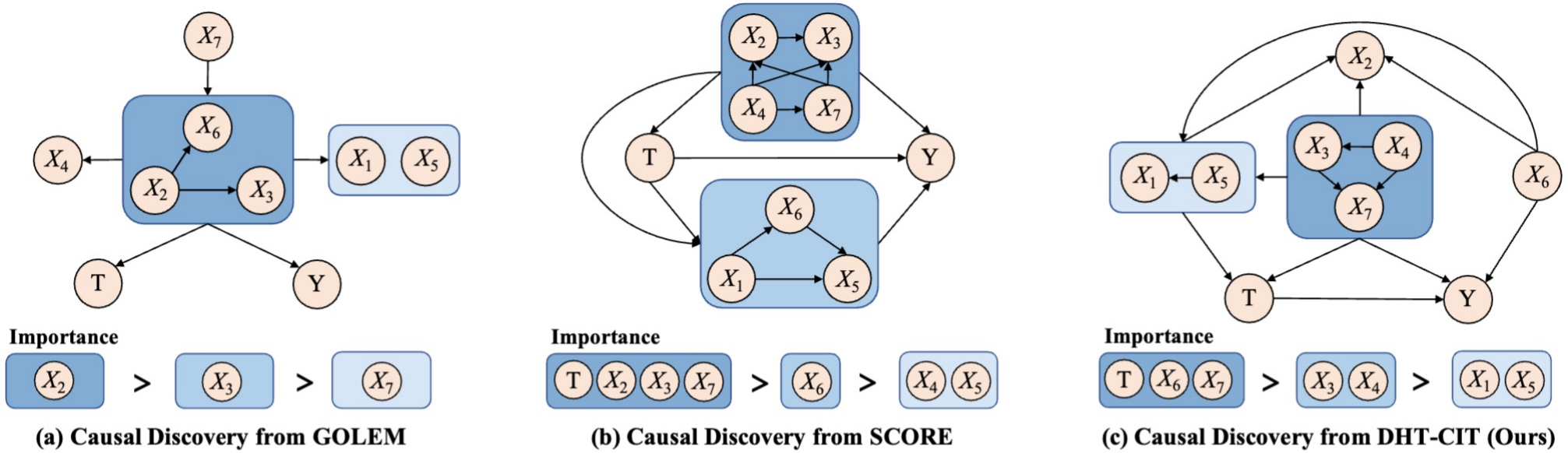
Data-nodes-edges for large graphs: Sin-50-100, Sin-100-100

Table 7. The experiments on Sin-50-100 & Sin-100-100 datasets.

	Sin-50-100 data with Gauss noise ($\mathcal{D} = \{X^1, X^3\}$)			
Method	**SID↓**	**SHD↓**	**#Prune↓**	**Running Time(s)↓**
SCORE	$247.0_{\pm 102.5}$	$23.0_{\pm 8.56}$	$1127._{\pm 2.06}$	1027s
DHT-CIT	$2039._{\pm 84.14}$	$234._{\pm 2.71}$	$397.6_{\pm 25.54}$	3217s
DHT-CIT (50 Intervention)	$203.0_{\pm 61.1}$	$14.8_{\pm 3.90}$	$\mathbf{149.0}_{\pm 27.00}$	**357s**
DHT-CIT+SCORE	$97.4_{\pm 101.6}$	$7.60_{\pm 5.28}$	$352.6_{\pm 23.69}$	1249s
DHT-CIT+SCORE (10 Intervention)	$\mathbf{53.2}_{\pm 20.29}$	$\mathbf{4.80}_{\pm 0.98}$	$284.0_{\pm 26.58}$	1109s
	Sin-100-100 data with Gauss noise ($\mathcal{D} = \{X^1, X^3\}$)			
Method	**SID↓**	**SHD↓**	**#Prune↓**	**Running Time(s)↓**
SCORE	$381._{\pm 156.5}$	$28.67_{\pm 4.5}$	$4850_{\pm 0.2}$	4689s
DHT-CIT	$2377_{\pm 427}$	$218._{\pm 12.9}$	$787.0_{\pm 49.1}$	19655s
DHT-CIT (100 Intervention)	$\mathbf{5.33}_{\pm 7.50}$	$\mathbf{1.00}_{\pm 1.41}$	$\mathbf{347.0}_{\pm 9.10}$	**1074s**
DHT-CIT+SCORE	$28.67_{\pm 9.53}$	$4.67_{\pm 0.47}$	$925.0_{\pm 83.3}$	6342s
DHT-CIT+SCORE (10 Intervention)	$14.67_{\pm 11.9}$	$3.33_{\pm 1.25}$	$797.3_{\pm 29.4}$	6108s

Real Data

Variable	Description
$PM_{2.5}(T)$	Annual county PM2.5 concentration, $\mu g/m^3$
CMR(Y)	Annual county cardiovascular mortality rate, deaths/100,000 person-years
Unemploy(X_1)	Civilian labor force unemployment rate in 2010
Income(X_2)	Median household income in 2009
Female(X_3)	Family households - female householder, no spouse present in 2010 / Family households in 2010
Vacant(X_4)	Vacant housing units in 2010 / Total housing units in 2010
Owner(X_5)	Owner-occupied housing units - percent of total occupied housing units in 2010
Edu(X_6)	Educational attainment - persons 25 years and over - high school graduate (includes equivalency) in 2010
Poverty(X_7)	Families below poverty level in 2009



(a) Causal Discovery from GOLEM (b) Causal Discovery from SCORE (c) Causal Discovery from DHT-CIT (Ours)

Figure 2. Causal Discovery on the PM-CMR Dataset.

Thanks

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62441605, 62376243, 62037001, U20A20387, 623B2002), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010).