Causal Discovery and Inference in Time Series

C. K. Assaad, E. Devijver, *E. Gaussier*, G. Goessler, A. Meynaoui, L. Zan

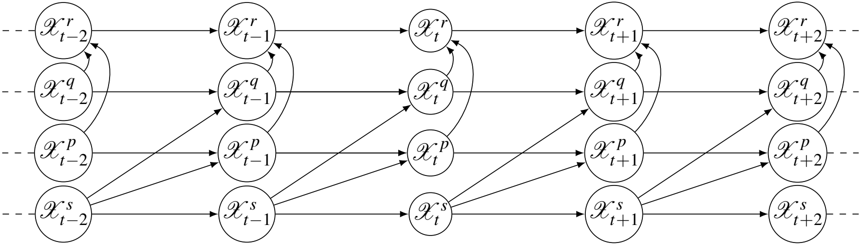
17 April 2023

Table of content

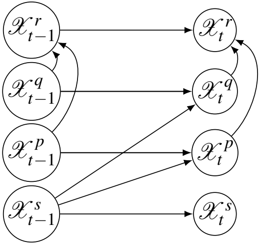
Time series are everywhere

- Time series arise as soon as observations, from sensors or experiments, for example, are collected over time
- They are present in various forms in many different domains
  - Healthcare (through, e.g., monitoring systems)
  - Industry 4.0 (through, e.g., predictive maintenance and industrial monitoring systems)
  - Surveillance systems (from images, acoustic signals, seismic waves, etc.)
  - Energy management (through, e.g. energy consumption data)
  - ...

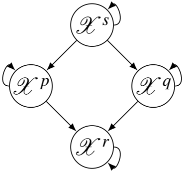
How to represent relations in time series?



Full Time Causal Graph (a)



Window Causal Graph (b)



Summary Causal Graph (c)

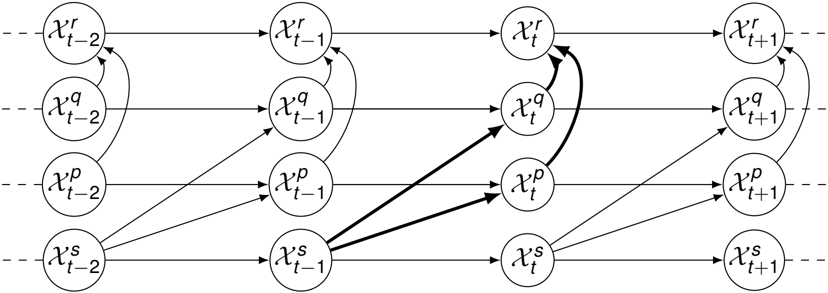
Causal graphs in time series (1)

Assumption 1 - Temporal priority *A cause does not occur after its effects (maximal temporal lag)*

Causal graphs in time series (1)

Assumption 1 - Temporal priority *A cause does not occur after its effects (maximal temporal lag)*

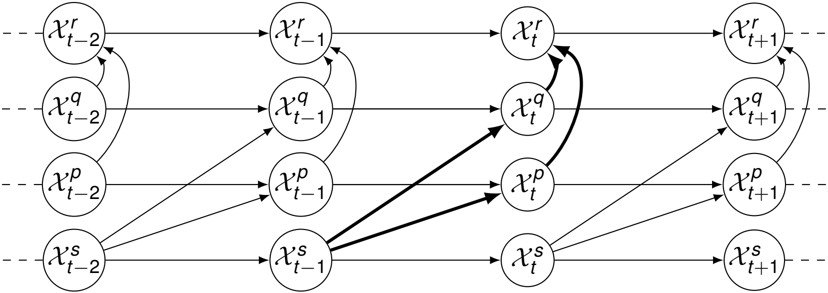
The full story



Causal graphs in time series (1)

Assumption 1 - Temporal priority *A cause does not occur after its effects (maximal temporal lag)*

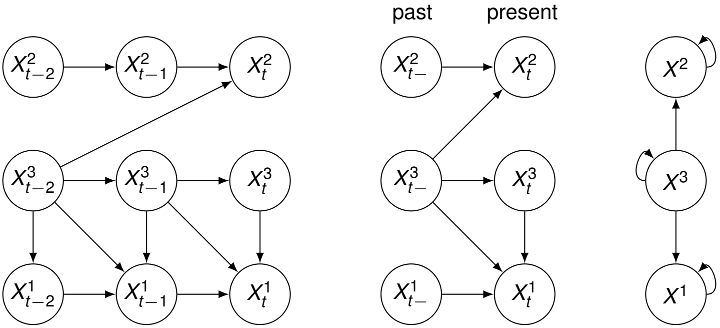
The full story



Assumption 2 - Consistency through time [causal stationarity] *All causal relationships remain constant in direction throughout time*

Causal graphs in time series (2)

Summarizing the full-time window graph with or without loss of information: window causal graph (a), extended summary causal graph (b) and summary causal graph (c)



(a) (b) (c)

Causal graphs in time series (3)

Remarks

- Window causal graphs equivalent to full-time graphs (cons. through time)
- Unique extended summary causal graph for a given window causal graph (reverse not true)
- Unique summary causal graph for a given extended summary causal graph (reverse not true)
- Advantage/disadvantage of summary causal graphs

  $+$ Easier to manipulate by experts

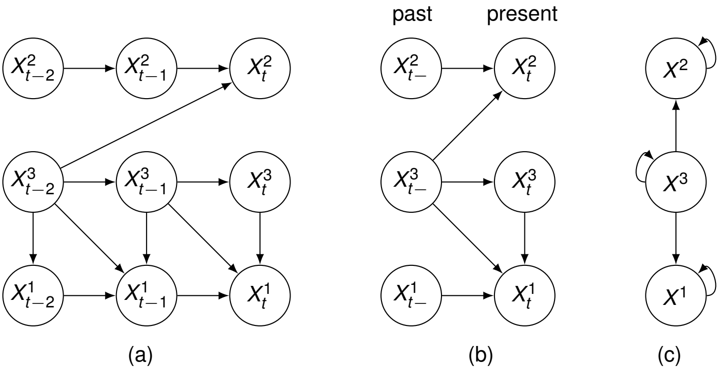
  $-$ Less precise than window causal graphs

Causal graphs in time series (3)

Remarks

- Window causal graphs equivalent to full-time graphs (cons. through time)
- Unique extended summary causal graph for a given window causal graph (reverse not true)
- Unique summary causal graph for a given extended summary causal graph (reverse not true)
- Advantage/disadvantage of summary causal graphs
  + Easier to manipulate by experts
  – Less precise than window causal graphs

Assumption 3 - Acyclicity *All causal graphs are acyclic*



X^2_{t-2}
X^2_{t-1}
X^2_t
X^3_{t-2}
X^3_{t-1}
X^3_t
X^1_{t-2}
X^1_{t-1}
X^1_t
(a)
past
present
X^2_{t-}
X^2_t
X^3_{t-}
X^3_t
X^1_{t-}
X^1_t
(b)
X^2
X^3
X^1
(c)

Causal graphs in time series (3)

Remarks

- Window causal graphs equivalent to full-time graphs (cons. through time)
- Unique extended summary causal graph for a given window causal graph (reverse not true)
- Unique summary causal graph for a given extended summary causal graph (reverse not true)
- Advantage/disadvantage of summary causal graphs
  - + Easier to manipulate by experts
  - – Less precise than window causal graphs

Assumption 3 - Acyclicity *All causal causal graphs are acyclic*

Focus on summary causal graphs (starting with extended summary causal graphs) and continuous time series

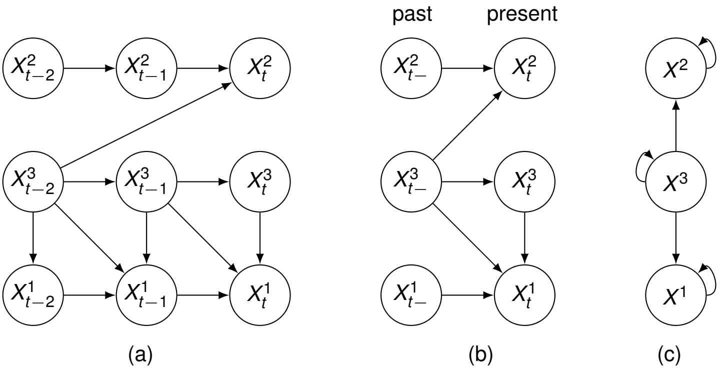
Problem statement, Markov cond. and faithfulness

Problem: *directly* infer extended summary graph from observational data (causal structure learning)

Problem statement, Markov cond. and faithfulness

Problem: *directly* infer extended summary graph from observational data (causal structure learning)

Markov condition *A necessary and sufficient condition for a probability distribution to be compatible with a DAG $\mathcal{G}$ is that every variable be independent of all its nondescendants (in $\mathcal{G}$), conditional on its parents*



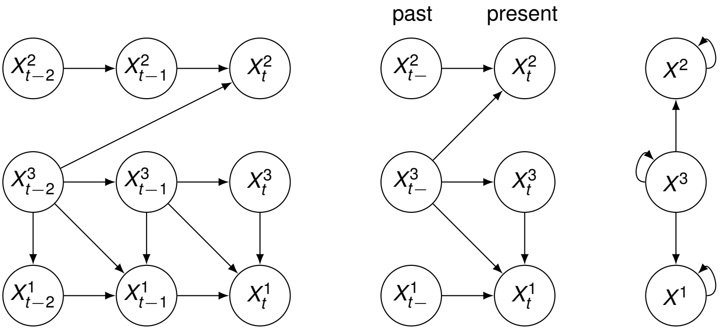
X^2_{t-2}
X^2_{t-1}
X^2_t
X^3_{t-2}
X^3_{t-1}
X^3_t
X^1_{t-2}
X^1_{t-1}
X^1_t
(a)
past
present
X^2_{t-}
X^2_t
X^3_{t-}
X^3_t
X^1_{t-}
X^1_t
(b)
X^2
X^3
X^1
(c)

Problem statement, Markov cond. and faithfulness

Problem: infer extended summary graph from observational data (causal structure learning)

Markov condition *A necessary and sufficient condition for a probability distribution to be compatible with a DAG $\mathcal{G}$ is that every variable be independent of all its nondescendants (in $\mathcal{G}$), conditional on its parents*

Assumption 4 - Faithfulness *We say that a graph $\mathcal{G}$ and a compatible probability distribution P are faithful to one another if all and only the conditional independence relations true in P are entailed by the Markov condition applied to $\mathcal{G}$*


past
present
X^2_{t-2}
X^2_{t-1}
X^2_t
X^3_{t-2}
X^3_{t-1}
X^3_t
X^1_{t-2}
X^1_{t-1}
X^1_t
X^2_{t-}
X^3_{t-}
X^1_{t-}
X^2
X^3
X^1

(a) (b) (c)

Truly causal methods?

General functional model of any potential effect X^q

$$X_t^q = f(\mathcal{C}_t^q(X^{r_1}), \cdots, \mathcal{C}_t^q(X^{r_q}), \xi_t^q)$$

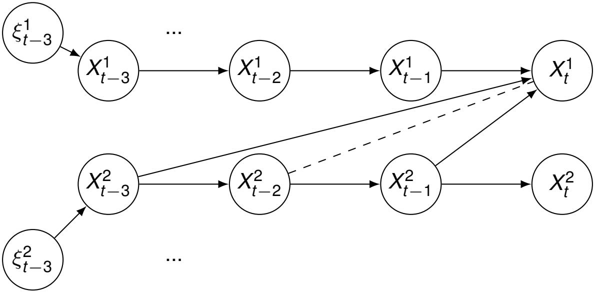
- f: any real-valued multivariate function
- $\mathcal{C}^q$: set of causes, $\mathcal{C}_t^q(X^r)$: past instants of X^r which are actual causes
- ξ_t^q: noise independent from all causes

Truly causal methods?

General functional model of any potential effect X^q

$$X_t^q = f(\mathcal{C}_t^q(X^{r_1}), \cdots, \mathcal{C}_t^q(X^{r_q}), \xi_t^q)$$

- f: any real-valued multivariate function
- $\mathcal{C}^q$: set of causes, $\mathcal{C}_t^q(X^r)$: past instants of X^r which are actual causes
- ξ_t^q: noise independent from all causes

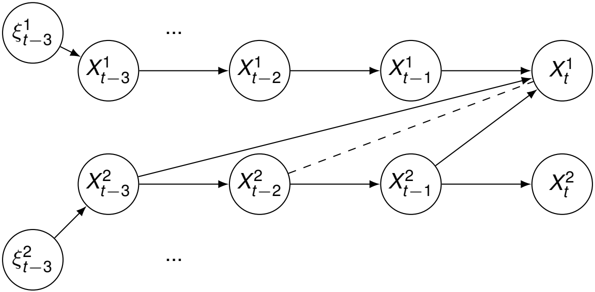


Truly causal methods?

General functional model of any potential effect X^q

$$X_t^q = f(\mathcal{C}_t^q(X^{r_1}), \cdots, \mathcal{C}_t^q(X^{r_q}), \xi_t^q)$$

- f: any real-valued multivariate function
- $\mathcal{C}^q$: set of causes, $\mathcal{C}_t^q(X^r)$: past instants of X^r which are actual causes
- ξ_t^q: noise independent from all causes



Truly causal methods aim at distinguishing causal correlations from spurious correlations

Families of causal methods

Main families of causal methods

- Granger causality, CCM causality, PAI causality
- Constraint-based approaches, noise-based approaches, score-based approaches, logic-based approaches

Families of causal methods

Main families of causal methods

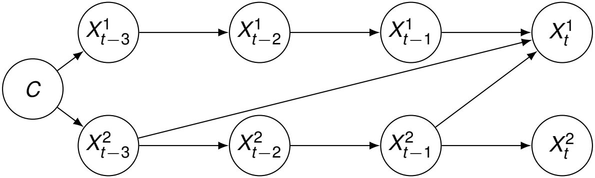
- Granger causality, CCM causality, PAI causality
- Constraint-based approaches, noise-based approaches, score-based approaches, logic-based approaches

Focus on constraint-based approaches (and noise-based approaches): popular, sound and complete, no assumption on f

Table of content

PC with causal sufficiency

Assumption 5 - Causal sufficiency *Observed variables are causally sufficient, i.e., all common causes of all variables are observed*


C
X^1_{t-3}
X^1_{t-2}
X^1_{t-1}
X^1_t
X^2_{t-3}
X^2_{t-2}
X^2_{t-1}
X^2_t

PC with causal sufficiency

Assumption 5 - Causal sufficiency *Observed variables are causally sufficient, i.e., all common causes of all variables are observed*

The PC algorithm (Spirtes-2000; Colombo-2014)

1. Build complete undirected graph U between all pairs of variables
2. Set $n = 0$
3. For each pair of adjacent variables (X, Y) in (current) U such that $ADJ(U, X) - Y$ or $ADJ(U, Y) - X$ has at leat n elements, check conditional independence bet. X and Y with any subset of S of n elements. If $X \perp\!\!\!\perp Y | S$, $S = Sepset(X, Y)$
4. $n = n + 1$
5. Execute orientation rules on obtained skeleton using temporal priority between the past and present slices

Conditional independence at the core of the procedure for constructing the skeleton

Independence measure: mutual information

Mutual information ...

- ... has proven useful in several studies on causal inference (Affeldt-2015, Runge-2019, Runge-2020)
- ... does not require assumptions on the generative model
- ... can be (somewhat) robustly estimated from observational data (k-NN)

Independence measure: mutual information

Mutual information ...

- ... has proven useful in several studies on causal inference (Affeldt-2015, Runge-2019, Runge-2020)
- ... does not require assumptions on the generative model
- ... can be (somewhat) robustly estimated from observational data (k-NN)

Present slice Standard mutual information

Independence measure: mutual information

Mutual information ...

- ... has proven useful in several studies on causal inference (Affeldt-2015, Runge-2019, Runge-2020)
- ... does not require assumptions on the generative model
- ... can be (somewhat) robustly estimated from observational data (k-NN)

Present slice Standard mutual information

Between past and present slices X^p does not directly cause X^q if there exists a subset of time series $\mathcal{X}^R$ such that $\forall K \in \mathbb{Z}^*$, $\forall \{\gamma_1, \cdots, \gamma_K\}$ s.t. $0 \leq \gamma_K < \cdots < \gamma_1$:

$$I(X_t^q; X_{t-\gamma_1}^p, \cdots, X_{t-\gamma_K}^p | \mathcal{X}_{t*}^R) = 0$$

Efficient estimation

One can efficiently identify independence between considering the window in X^p starting at $t - \gamma$ and ending at t, denoted $t - \gamma : t$ where γ is the maximal gap:

Efficient estimation

One can efficiently identify independence between considering the window in X^p starting at $t - \gamma$ and ending at t, denoted $t-\gamma:t$ where γ is the maximal gap:

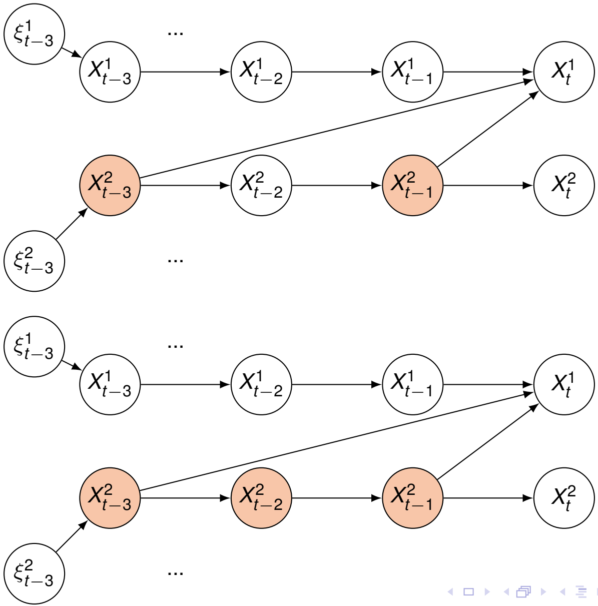
Property

Let γ denote the maximum gap between a cause and its effect. The following two propositions are equivalent:

(a) $I(X_t^q; X_{t-\gamma_1}^p, \cdots, X_{t-\gamma_K}^p | \mathcal{X}_{t*}^R) = 0, \forall K \geq 1, \forall \gamma_1 > \cdots > \gamma_K \geq 0,$

(b) $I(X_t^q; X_{t-\gamma:t}^p | \mathcal{X}_{t*}^R) = 0$

Illustration



Greedy causation entropy

Definition (causal gredy entropy)
The greedy causation entropy, denoted by GCE, from the time series X^p to the time series X^q is defined by:

$$GCE(X^p \rightarrow X^q) = I(X_t^q; X_{t-\gamma:t-1}^p) \tag{1}$$

Denoting by $X^{\boldsymbol{R}}$ a set of m time series $\{X^{r_1}, \cdots, X^{r_m}\}$, the conditional greedy causation entropy takes the form:

$$GCE(X^p \rightarrow X^q | X^{\boldsymbol{R}}) = I(X_t^q; X_{t-\gamma:t-1}^p | X_{t^*}^{r_1}, \cdots, X_{t^*}^{r_m}), \tag{2}$$

where t^* denotes either the present instant t or the time window $t-\gamma:t-1$

Greedy causation entropy

Definition (causal gredy entropy)
The greedy causation entropy, denoted by GCE, from the time series X^p to the time series X^q is defined by:

$$GCE(X^p \rightarrow X^q) = I(X_t^q; X_{t-\gamma:t-1}^p) \quad (1)$$

Denoting by $X^{\mathbf{R}}$ a set of m time series $\{X^{r_1}, \cdots, X^{r_m}\}$, the conditional greedy causation entropy takes the form:

$$GCE(X^p \rightarrow X^q | X^{\mathbf{R}}) = I(X_t^q; X_{t-\gamma:t-1}^p | X_{t^*}^{r_1}, \cdots, X_{t^*}^{r_m}), \quad (2)$$

where t^* denotes either the present instant t or the time window $t-\gamma:t-1$

Estimation k-NN method (Frenzel-2007)

From a skeleton to a causal graph: orientation rules

PC-Rule 0 - Origin of causality

(i) In an unshielded triple $X_t^p - X_t^r - X_t^q$, if $X_t^r \notin \texttt{Sepset}(p \leftrightarrow q)$, then X_t^r is an unshielded collider: $X_t^p \rightarrow X_t^r \leftarrow X_t^q$

(ii) In an unshielded triple $X_{t-}^q \rightarrow X_t^q - X_t^p$, if $X_t^q \notin \texttt{Sepset}(q \rightarrow p)$, then X_t^q is an unshielded collider: $X_{t-}^q \rightarrow X_t^q \leftarrow X_t^p$

From a skeleton to a causal graph: orientation rules

PC-Rule 0 - Origin of causality

(i) In an unshielded triple $X^p_t - X^r_t - X^q_t$, if $X^r_t \notin \texttt{Sepset}(p \leftrightarrow q)$, then X^r_t is an unshielded collider: $X^p_t \rightarrow X^r_t \leftarrow X^q_t$

(ii) In an unshielded triple $X^q_{t-} \rightarrow X^q_t - X^p_t$, if $X^q_t \notin \texttt{Sepset}(q \rightarrow p)$, then X^q_t is an unshielded collider: $X^q_{t-} \rightarrow X^q_t \leftarrow X^p_t$

PC-Rule 1- Propagation of causality In an unshielded triple $X^p_t \rightarrow X^r_t - X^q_t$ (resp. $X^p_{t-} \rightarrow X^r_t - X^q_t$), if $X^r_t \in \texttt{Sepset}(p \leftrightarrow q)$ then orient the unshielded triple as $X^p_t \rightarrow X^r_t \rightarrow X^q_t$ (resp. $X^p_{t-} \rightarrow X^r_t \rightarrow X^q_t$)

From a skeleton to a causal graph: orientation rules

PC-Rule 0 - Origin of causality

(i) In an unshielded triple $X^p_t - X^r_t - X^q_t$, if $X^r_t \notin \texttt{Sepset}(p \leftrightarrow q)$, then X^r_t is an unshielded collider: $X^p_t \rightarrow X^r_t \leftarrow X^q_t$

(ii) In an unshielded triple $X^q_{t-} \rightarrow X^q_t - X^p_t$, if $X^q_t \notin \texttt{Sepset}(q \rightarrow p)$, then X^q_t is an unshielded collider: $X^q_{t-} \rightarrow X^q_t \leftarrow X^p_t$

PC-Rule 1- Propagation of causality In an unshielded triple $X^p_t \rightarrow X^r_t - X^q_t$ (resp. $X^p_{t-} \rightarrow X^r_t - X^q_t$), if $X^r_t \in \texttt{Sepset}(p \leftrightarrow q)$ then orient the unshielded triple as $X^p_t \rightarrow X^r_t \rightarrow X^q_t$ (resp. $X^p_{t-} \rightarrow X^r_t \rightarrow X^q_t$)

PC-Rule 2 If there exist a direct path from X^p_t to X^q_t and an edge between X^p_t and X^q_t, then orient $X^p_t \rightarrow X^q_t$

From a skeleton to a causal graph: orientation rules

PC-Rule 0 - Origin of causality

(i) In an unshielded triple $X_t^p - X_t^r - X_t^q$, if $X_t^r \notin \texttt{Sepset}(p \leftrightarrow q)$, then X_t^r is an unshielded collider: $X_t^p \rightarrow X_t^r \leftarrow X_t^q$

(ii) In an unshielded triple $X_{t-}^q \rightarrow X_t^q - X_t^p$, if $X_t^q \notin \texttt{Sepset}(q \rightarrow p)$, then X_t^q is an unshielded collider: $X_{t-}^q \rightarrow X_t^q \leftarrow X_t^p$

PC-Rule 1- Propagation of causality In an unshielded triple $X_t^p \rightarrow X_t^r - X_t^q$ (resp. $X_{t-}^p \rightarrow X_t^r - X_t^q$), if $X_t^r \in \texttt{Sepset}(p \leftrightarrow q)$ then orient the unshielded triple as $X_t^p \rightarrow X_t^r \rightarrow X_t^q$ (resp. $X_{t-}^p \rightarrow X_t^r \rightarrow X_t^q$)

PC-Rule 2 If there exist a direct path from X_t^p to X_t^q and an edge between X_t^p and X_t^q, then orient $X_t^p \rightarrow X_t^q$

PC-Rule 3 Orient $X_t^p - X_t^q$ as $X_t^p \rightarrow X_t^q$ whenever there are two paths $X_t^p - X_t^r \rightarrow X_t^q$ and $X_t^p - X_t^s \rightarrow X_t^q$

Wrapping-up

- Simple algorithm with causal sufficiency:
  - Standard skeleton construction based on greedy causation entropy
  - More or less standard orientation rules operating on the present slice
- Theoretical guarantees under assumptions made

Theorem

Let the distribution of V be faithful to a DAG $\mathcal{G} = (V, E)$, and assume that we are given perfect conditional independence information about all pairs of variables (X^p, X^q) in V given subsets $X^{\boldsymbol{R}} \subseteq V \backslash \{X^p, X^q\}$. Then the skeleton constructed previously followed by the above orientation rules represents the CPDAG of the extended summary causal graph $\mathcal{G}$

FCI without causal sufficiency

1. Same skeleton construction as before
2. Orientation rules
   - FCI-Rule 0 (origin of causality) is adapted as before
   - FCI-Rules 1, 2, 3 and 4 (Spirtes-2000)
   - FCI-Rules 8, 9 and 10 (Zhang-2008) - rules 5, 6 and 7 deal with selection bias

FCI without causal sufficiency

1. Same skeleton construction as before
2. Orientation rules
    - FCI-Rule 0 (origin of causality) is adapted as before
    - FCI-Rules 1, 2, 3 and 4 (Spirtes-2000)
    - FCI-Rules 8, 9 and 10 (Zhang-2008) - rules 5, 6 and 7 deal with selection bias

Remark Possible D-separation sets vs. separation sets, PAG (MAG) vs CPDAG (DAG)

Table of content

Experimental setting

Datasets, evaluation measures, methods

Experimental setting

Datasets, evaluation measures, methods

- Simple artificial datasets correspond to three different causally sufficient structures and two non causally sufficient structures generated by:

$$X_t^q = a_{t-1}^{qq} X_{t-1}^q + \sum_p a_{t-\gamma}^{pq} f(X_{t-\gamma}^p) + 0.1\xi_t^q$$

with $\gamma \geq 0$, $a_t^{jq} \sim \mathcal{U}([-1; -0.1] \cup [0.1; 1])$, $\xi_t^q \sim \mathcal{N}(0, 1)$ and f chosen at random in $\{$absolute value, tanh, sine, cosine$\}$

Experimental setting

Datasets, evaluation measures, methods

- Simple artificial datasets correspond to three different causally sufficient structures and two non causally sufficient structures generated by:

$$X_t^q = a_{t-1}^{qq} X_{t-1}^q + \sum_p a_{t-\gamma}^{pq} f(X_{t-\gamma}^p) + 0.1\xi_t^q$$

  with $\gamma \geq 0$, $a_t^{jq} \sim \mathcal{U}([-1;-0.1] \cup [0.1;1])$, $\xi_t^q \sim \mathcal{N}(0,1)$ and f chosen at random in $\{$absolute value, tanh, sine, cosine$\}$

- FMRI datasets for 28 different brain networks (Smith-2011)

Experimental setting

Datasets, evaluation measures, methods

- Simple artificial datasets correspond to three different causally sufficient structures and two non causally sufficient structures generated by:

$$X_t^q = a_{t-1}^{qq} X_{t-1}^q + \sum_p a_{t-\gamma}^{pq} f(X_{t-\gamma}^p) + 0.1\xi_t^q$$

with $\gamma \geq 0$, $a_t^{jq} \sim \mathcal{U}([-1; -0.1] \cup [0.1; 1])$, $\xi_t^q \sim \mathcal{N}(0, 1)$ and f chosen at random in $\{$absolute value, tanh, sine, cosine$\}$

- FMRI datasets for 28 different brain networks (Smith-2011)
- F1-score (between and within time series)

Experimental setting

Datasets, evaluation measures, methods

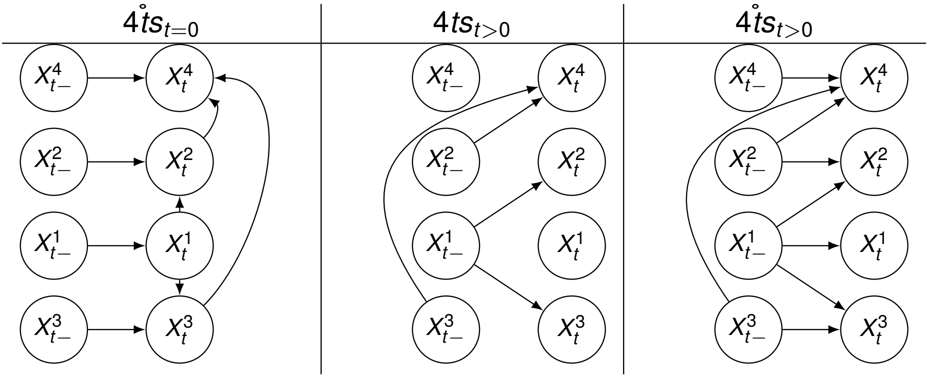
- Simple artificial datasets correspond to three different causally sufficient structures and two non causally sufficient structures generated by:

$$X_t^q = a_{t-1}^{qq} X_{t-1}^q + \sum_p a_{t-\gamma}^{pq} f(X_{t-\gamma}^p) + 0.1\xi_t^q$$

with $\gamma \geq 0$, $a_t^{jq} \sim \mathcal{U}([-1; -0.1] \cup [0.1; 1])$, $\xi_t^q \sim \mathcal{N}(0, 1)$ and f chosen at random in $\{$absolute value, tanh, sine, cosine$\}$

- FMRI datasets for 28 different brain networks (Smith-2011)
- F1-score (between and within time series)
- PCMICI (Runge-2019, Runge-2020), oCSE (Sun-2015), tsFCI (Entner-2010), VarLiNGAM (Hyvärinen-2010), Dynotears (Pamfil-2020), GCMVL (Arnold-2017), TCDF (Nauta-2019)

Artificial datasets (extract)



Results

No hidden common cause

	Perf.	PCGCE	oCSE	PCMCI	VarLiNGAM	Dynotears	TCDF	MVGCL
$4\mathring{t}s_{t=0}$	$F^{p\neq q}$	**0.62** $\pm$ 0.17	–	0.60 $\pm$ 0.12	0.32 $\pm$ 0.13	0.04 $\pm$ 0.12	0.00 $\pm$ 0.00	–
	$F^{p=q}$	0.81 $\pm$ 0.12	–	0.87 $\pm$ 0.12	**0.92** $\pm$ 0.07	0.37 $\pm$ 0.21	0.18 $\pm$ 0.24	–
$4ts_{t>0}$	$F^{p\neq q}$	0.**71** $\pm$ 0.13	0.31 $\pm$ 0.21	0.67 $\pm$ 0.16	0.00 $\pm$ 0.00	0.16 $\pm$ 0.19	0.00 $\pm$ 0.00	0.52 $\pm$ 0.11
$4\mathring{t}s_{t>0}$	$F^{p\neq q}$	**0.81** $\pm$ 0.18	0.78 $\pm$ 0.17	**0.81** $\pm$ 0.12	0.00 $\pm$ 0.00	0.16 $\pm$ 0.19	0.04 $\pm$ 0.12	0.53 $\pm$ 0.09
	$F^{p=q}$	0.94 $\pm$ 0.06	0.82 $\pm$ 0.11	0.97 $\pm$ 0.05	**0.98** $\pm$ 0.04	0.47 $\pm$ 0.15	0.35 $\pm$ 0.27	–

Results

No hidden common cause

	Perf.	PCGCE	oCSE	PCMCI	VarLiNGAM	Dynotears	TCDF	MVGCL
$4\mathring{t}s_{t=0}$	$F^{p\neq q}$	**0.62** $\pm$ 0.17	–	0.60 $\pm$ 0.12	0.32 $\pm$ 0.13	0.04 $\pm$ 0.12	0.00 $\pm$ 0.00	–
	$F^{p=q}$	0.81 $\pm$ 0.12	–	0.87 $\pm$ 0.12	**0.92** $\pm$ 0.07	0.37 $\pm$ 0.21	0.18 $\pm$ 0.24	–
$4ts_{t>0}$	$F^{p\neq q}$	**0.71** $\pm$ 0.13	0.31 $\pm$ 0.21	0.67 $\pm$ 0.16	0.00 $\pm$ 0.00	0.16 $\pm$ 0.19	0.00 $\pm$ 0.00	0.52 $\pm$ 0.11
$4\mathring{t}s_{t>0}$	$F^{p\neq q}$	**0.81** $\pm$ 0.18	0.78 $\pm$ 0.17	**0.81** $\pm$ 0.12	0.00 $\pm$ 0.00	0.16 $\pm$ 0.19	0.04 $\pm$ 0.12	0.53 $\pm$ 0.09
	$F^{p=q}$	0.94 $\pm$ 0.06	0.82 $\pm$ 0.11	0.97 $\pm$ 0.05	**0.98** $\pm$ 0.04	0.47 $\pm$ 0.15	0.35 $\pm$ 0.27	–

No hidden common cause (FMRI)

	Perf.	PCGCE	oCSE	PCMCI	VarLiNGAM	Dynotears	TCDF	MVGCL
FMRI	$F^{p\neq q}$	0.31 $\pm$ 0.2	0.16 $\pm$ 0.19	0.22 $\pm$ 0.18	**0.49** $\pm$ 0.28	0.34 $\pm$ 0.13	0.06 $\pm$ 0.12	0.35 $\pm$ 0.08

Results

No hidden common cause

	Perf.	PCGCE	oCSE	PCMCI	VarLiNGAM	Dynotears	TCDF	MVGCL
$4\mathring{t}s_{t=0}$	$F^{p \neq q}$	**0.62** $\pm$ 0.17	–	0.60 $\pm$ 0.12	0.32 $\pm$ 0.13	0.04 $\pm$ 0.12	0.00 $\pm$ 0.00	–
	$F^{p = q}$	0.81 $\pm$ 0.12	–	0.87 $\pm$ 0.12	**0.92** $\pm$ 0.07	0.37 $\pm$ 0.21	0.18 $\pm$ 0.24	–
$4ts_{t>0}$	$F^{p \neq q}$	**0.71** $\pm$ 0.13	0.31 $\pm$ 0.21	0.67 $\pm$ 0.16	0.00 $\pm$ 0.00	0.16 $\pm$ 0.19	0.00 $\pm$ 0.00	0.52 $\pm$ 0.11
$4\mathring{t}s_{t>0}$	$F^{p \neq q}$	**0.81** $\pm$ 0.18	0.78 $\pm$ 0.17	**0.81** $\pm$ 0.12	0.00 $\pm$ 0.00	0.16 $\pm$ 0.19	0.04 $\pm$ 0.12	0.53 $\pm$ 0.09
	$F^{p = q}$	0.94 $\pm$ 0.06	0.82 $\pm$ 0.11	0.97 $\pm$ 0.05	**0.98** $\pm$ 0.04	0.47 $\pm$ 0.15	0.35 $\pm$ 0.27	–

No hidden common cause (FMRI)

	Perf.	PCGCE	oCSE	PCMCI	VarLiNGAM	Dynotears	TCDF	MVGCL
FMRI	$F^{p \neq q}$	0.31 $\pm$ 0.2	0.16 $\pm$ 0.19	0.22 $\pm$ 0.18	**0.49** $\pm$ 0.28	0.34 $\pm$ 0.13	0.06 $\pm$ 0.12	0.35 $\pm$ 0.08

With hidden common causes

	Perf.	FCIGCE	tsFCI	TCDF
$7t2h_{t>0}$	$F^{p \neq q}$	**0.57** $\pm$ 0.1	0.52 $\pm$ 0.1	0.02 $\pm$ 0.1
$7t\mathring{2}h_{t>0}$	$F^{p \neq q}$	0.33 $\pm$ 0.1	**0.36** $\pm$ 0.1	0.07 $\pm$ 0.1
	$F^{p = q}$	0.83 $\pm$ 0.1	**0.99** $\pm$ 0.1	0.19 $\pm$ 0.2

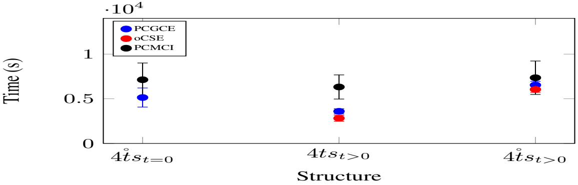
First conclusions

1. Relatively simple extensions of PC and FCI to extended summary graphs (Assaad-2022(b))
2. Method sound and complete with robust behaviour on various datasets
3. Assumptions: temporal priority, consistency through time, acyclicity, faithfulness
4. Remark: possibility to mix approaches, *e.g.* noise-based and constraint-based (Assaad-2021)

First conclusions

1. Relatively simple extensions of PC and FCI to extended summary graphs (Assaad-2022(b))
2. Method sound and complete with robust behaviour on various datasets
3. Assumptions: temporal priority, consistency through time, acyclicity, faithfulness
4. Remark: possibility to mix approaches, *e.g.* noise-based and constraint-based (Assaad-2021)

Time complexity



What about summary graphs?

The situation in summary graphs is slightly more difficult as there is no distinction between past and present so that windows need be considered on each time series. One can nevertheless define a mutual infomation measure based on these windows and follow an appraoch similar to the one above with additional rules for instantaneous relations (Assaad-2022(c)).

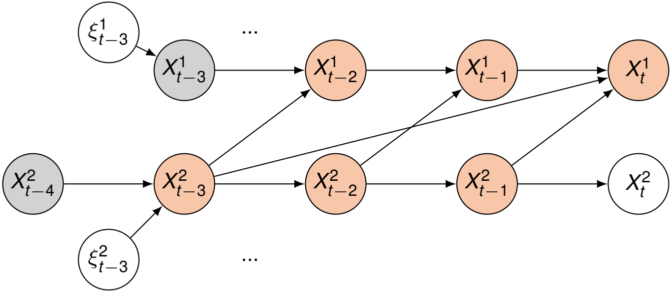


Table of content

Background

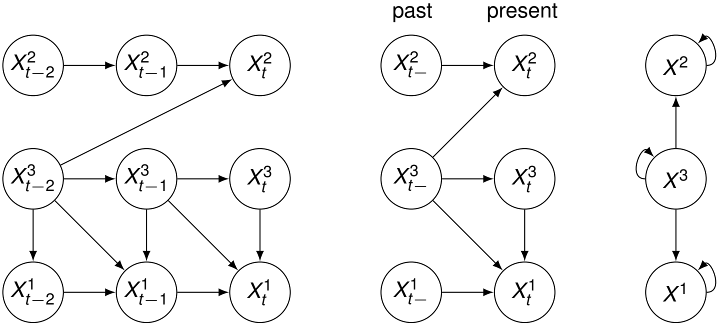
General form of an intervention $P(\mathbf{R}_t = \mathbf{r}_t | do(\mathbf{I}_{\mathbf{t}^i} = \mathbf{i}_{\mathbf{t}^i}))$ where $\mathbf{I}_{\mathbf{t}^i}$ is the intervention set, $\mathbf{R}_t$ the response set and $do()$ the standard intervention operator (Pearl-2009)

Background

$P(X^1_t = x^1_t, X^3_t = x^3_t | do(X^2_{t-3} = x^2_{t-3}))$ is an example of an intervention where the response set is $\{X^1_t, X^3_t\}$ and the intervention set is $\{X^2_{t-3}\}$

Background

$P(X^1_t = x^1_t, X^3_t = x^3_t | do(X^2_{t-3} = x^2_{t-3}))$ is an example of an intervention where the response set is $\{X^1_t, X^3_t\}$ and the intervention set is $\{X^2_{t-3}\}$



(a) (b) (c)

Problem statement

Identifiability If it exists, provide an equivalent do-free formula, or estimand, for an intervention which can be estimated from observational data - *using observational data to compute a causal effect*

Problem statement

Identifiability If it exists, provide an equivalent do-free formula, or estimand, for an intervention which can be estimated from observational data - *using observational data to compute a causal effect*

Example In the above window causal graph, $P(x_t^1|do(x_{t-1}^3)) = \sum_{x_{t-2}^3} P(x_{t-2}^3)P(x_t^1|x_{t-1}^3, x_{t-2}^3)$ as X_{t-2}^3 satisfies the back-door criterion

Problem statement

Identifiability If it exists, provide an equivalent do-free formula, or estimand, for an intervention which can be estimated from observational data - *using observational data to compute a causal effect*

Example In the above window causal graph, $P(x_t^1|do(x_{t-1}^3)) = \sum_{x_{t-2}^3} P(x_{t-2}^3)P(x_t^1|x_{t-1}^3, x_{t-2}^3)$ as X_{t-2}^3 satisfies the back-door criterion

Note that the estimand depends ont the underlying causal graph. Identifiability problem solved in window causal graphs (Blondel-2016). *What about (extended) summary causal graphs?*

Problem statement

Identifiability If it exists, provide an equivalent do-free formula, or estimand, for an intervention which can be estimated from observational data - *using observational data to compute a causal effect*

Example In the above window causal graph,
$P(x_t^1|do(x_{t-1}^3)) = \sum_{x_{t-2}^3} P(x_{t-2}^3)P(x_t^1|x_{t-1}^3, x_{t-2}^3)$ as X_{t-2}^3 satisfies the back-door criterion

Note that the estimand depends ont the underlying causal graph. Identifiability problem solved in window causal graphs (Blondel-2016). *What about (extended) summary causal graphs?*

The difficulty lies in the fact that the (extended) summary graph is not complete (does not contain all causal relations)

Problem statement

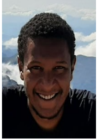
Identifiability If it exists, provide an equivalent do-free formula, or estimand, for an intervention which can be estimated from observational data - *using observational data to compute a causal effect*

Example In the above window causal graph,
$P(x_t^1|do(x_{t-1}^3)) = \sum_{x_{t-2}^3} P(x_{t-2}^3)P(x_t^1|x_{t-1}^3, x_{t-2}^3)$ as X_{t-2}^3 satisfies the back-door criterion

Note that the estimand depends ont the underlying causal graph. Identifiability problem solved in window causal graphs (Blondel-2016). *What about (extended) summary causal graphs?*

The difficulty lies in the fact that the (extended) summary graph is not complete (does not contain all causal relations)

$\rightarrow$ *Anouar's pres. tomorrow*



Equivalence relation

From (extended) summary causal graphs to window causal graphs

Equivalence relation

From (extended) summary causal graphs to window causal graphs

A simple approach to identifiability is to enumerate all window causal graphs compatible with a given (extended) summary causal graph and use a version of the ID algorithm (Shpitser-2006). However not possible in practice when the number of time series considered is large

Equivalence relation

From (extended) summary causal graphs to window causal graphs

A simple approach to identifiability is to enumerate all window causal graphs compatible with a given (extended) summary causal graph and use a version of the ID algorithm (Shpitser-2006). However not possible in practice when the number of time series considered is large

Equivalent window causal graph *Let G_s be a (extended) summary causal graph. Consider an intervention set $\mathbf{I}_{\mathbf{t}^i}$ and a response set $\mathbf{R}_\mathbf{t}$. Two window causal graphs G_{w_1} and G_{w_2} are equivalent for the intervention $P(\mathbf{r}_\mathbf{t}|do(\mathbf{i}_{\mathbf{t}^i}))$ if and only if the intervention is non-identifiable in both graphs, and there is no estimand, or identifiable in both with the same estimand*

Possible approach

1. Identify relevant interventional parents blocking all backdoor paths for a given intervention
2. Compute do-free formula from relevant interventional parents

Preliminary experimental results encouraging with possibility to use maximal equivalence class in all cases for extended summary graphs

Assumptions & data types

Assumptions made

1. Temporal priority, acyclicity, faithfulness
2. *Causal stationarity, (local) statistical stationarity*

Assumptions & data types

Assumptions made

1. Temporal priority, acyclicity, faithfulness
2. *Causal stationarity, (local) statistical stationarity*

Data types From purely continuous to mixed (quantitative and qualitative) data

Table of content

Dealing with mixed data

Once one has a way to estimate the (conditional) mutual information between mixed data samples, then one directly apply the above approach to infer causal graphs from observational data

Dealing with mixed data

Once one has a way to estimate the (conditional) mutual information between mixed data samples, then one directly apply the above approach to infer causal graphs from observational data

Different approaches have been proposed in the literature to estimate the (conditional) entropy of a set of variables, the two prominent ones being based on histograms, particularly adapted to qualitative data, and the k-NN approach, mostly used on quantitative data (Zan-2022)

Dealing with mixed data

Once one has a way to estimate the (conditional) mutual information between mixed data samples, then one directly apply the above approach to infer causal graphs from observational data

Different approaches have been proposed in the literature to estimate the (conditional) entropy of a set of variables, the two prominent ones being based on histograms, particularly adapted to qualitative data, and the k-NN approach, mostly used on quantitative data (Zan-2022)

We propose here an hybrid approach combining both histograms and k-NNs

Decomposing mutual information

Let us consider three mixed random vectors X, Y and Z, where any of their components can be either qualitative or quantitative. Let us denote by X^t (respectively Y^t, Z^t) the sub-vector of X (respectively Y, Z) composed by the quantitative components. Similarly, we denote by X^ℓ (respectively Y^ℓ, Z^ℓ) the sub-vector of qualitative components of X (respectively Y, Z). Then one has:

$$
\begin{aligned}
I(X;Y|Z) &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\
&= H(X^t,X^\ell,Z^t,Z^\ell) + H(Y^t,Y^\ell,Z^t,Z^\ell) \\
&- H(X^t,X^\ell,Y^t,Y^\ell,Z^t,Z^\ell) - H(Z^t,Z^\ell) \\
&= H(X^t,Z^t|X^\ell,Z^\ell) + H(Y^t,Z^t|Y^\ell,Z^\ell) - H(X^t,Y^t,Z^t|X^\ell,Y^\ell,Z^\ell) \\
&- H(Z^t|Z^\ell) + H(X^\ell,Z^\ell) + H(Y^\ell,Z^\ell) - H(X^\ell,Y^\ell,Z^\ell) - H(Z^\ell)
\end{aligned}
$$

Estimating $H(.,.)$ and $H(.|.)$

$H(.,.)$ is estimated with histograms

$H(.^t|.^\ell)$ is estimated with k-NNs

Theorem

Let (X, Y, Z) be a qualitative-quantitative mixed random vector. The obtained estimator $\widehat{I}(X;Y|Z)$ is consistent, meaning that, for all $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\widehat{I}(X;Y|Z) - I(X;Y|Z)| > \varepsilon) = 0$$

In addition, $\widehat{I}(X;Y|Z)$ is asymptotically unbiased, that is

$$\lim_{n\to\infty} \mathbb{E}[\widehat{I}(X;Y|Z) - I(X;Y|Z)] = 0$$

Illustration (1)

Comparison for $I(X;Y)$ and $I(X;Y|Z)$ (Z quantitative) with one state-of-the-art histogram-based, namely LH (Marx-2021), and three state-of-the-art methods based on k-NN, namely FP (Frenzel-2007), MS (Mesner-2020), RAVK (Rahimzamani-2018) on synthetic datasets revealed that (Zan-2022):

- FP performs well in the purely quantitative case with no conditioning but is not competitive in the mixed data case
- MS and RAVK are close to each other and have similar performance. Main drawback is that MS gives the value 0, or close to 0, to the estimator in some particular cases (as when the k-NN is always determined by Z)
- CMIh and LH behave well overall but LH is so slow that it can not be used when the dimension of Z exceeds 2
- The good behaviour of CMIh is further confirmed when used in conjunction with a local a permutation test (Zan-2022)

Illustration (1)

Comparison for $I(X;Y)$ and $I(X;Y|Z)$ (Z quantitative) with one state-of-the-art histogram-based, namely LH (Marx-2021), and three state-of-the-art methods based on k-NN, namely FP (Frenzel-2007), MS (Mesner-2020), RAVK (Rahimzamani-2018) on synthetic datasets revealed that (Zan-2022):

- FP performs well in the purely quantitative case with no conditioning but is not competitive in the mixed data case
- MS and RAVK are close to each other and have similar performance. Main drawback is that MS gives the value 0, or close to 0, to the estimator in some particular cases (as when the k-NN is always determined by Z)
- CMIh and LH behave well overall but LH is so slow that it can not be used when the dimension of Z exceeds 2
- The good behaviour of CMIh is further confirmed when used in conjunction with a local a permutation test (Zan-2022)

$\rightarrow$ *Discussion with Lei*

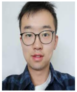


Illustration (2)

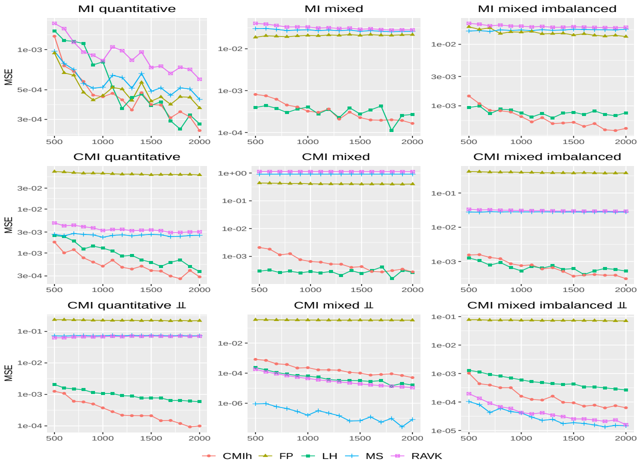


Figure: Synthetic data with known ground truth. MSE (on a log-scale) of each method with respect to the sample size (in abscissa)

Table of content

Conclusion

Exploring the use of (extended) summary graphs ...

1. Operational advantages: easier to infer, validate and come up with
2. Inferring such graphs from observational data
   - Independence estimation is a practical bottleneck (precise and fast)
3. Solving the identifiability problem in such graphs (on-going work)

for mixed time series

A few challenges

- Non stationarity
- Cyclic causal graphs (instantaneous relations)
- Efficient and effective methods with (some) theoretical guarantees with hidden confounders
- ...

Survey and Evaluation of Causal Discovery Methods for Time Series (Assaad-2022(a); JAIR, IJCAI)

Table of content

References (1)

Affeldt-2015 Séverine Affeldt, Hervé Isambert. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. UAI 2015

Arnold-2017 Andrew Arnold, Yan Liu, Naoki Abe. Temporal causal modeling with graphical granger methods. KDD 2007

Blondel-2016 G. Blondel, M. Arias, R. Gavaldà. Identifiability and transportability in dynamic causal networks. International Journal of Data Science and Analytics, 3:131–341 147, 2016

Colombo-2014 Diego Colombo and Marloes H. Maathuis. Order- independent constraint-based causal structure learning. Journal of Machine Learning Research, 15(116):3921– 3962, 2014

Entner-2010 Doris Entner, Patrik Hoyer. On causal discovery from time series data using fci. Proceedings of the 5th Euro- pean Workshop on Probabilistic Graphical Models, PGM 2010

Frenzel-2007 Stefan Frenzel, Bernd Pompe. Partial mutual information for coupling analysis of multivariate time series. Physical review letters, 99:204101, 2007

References (2)

Hyvärinen-2010 Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. J. Mach. Learn. Res., 2010

Marx-2021 Marx, A., Yang, L., van Leeuwen, M. Estimating conditional mutual information for discrete-continuous mixtures using multi-dimensional adaptive histograms. SIAM 2021

Mesner-2020 Mesner, O.C., Shalizi, C.R. Conditional Mutual Information Estimation for Mixed, Discrete and Continuous Data. IEEE Trans. Inf. Theory 2020

Nauta-2019 Meike Nauta, Doina Bucur, Christin Seifert. Causal discovery with attention-based convolutional neural net- works. Machine Learning and Knowledge Extraction, 2019

Pamfil-2020 Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, Bryon Aragam. Dynotears: Structure learning from time-series data. AISTATS 2020

Pearl-2009 J. Pearl. Causal inference in statistics: An overview. Statistics surveys, 2009

Rahimzamani-2018 Rahimzamani, A., Asnani, H., Viswanath, P., Kannan, S. Estimators for multivariate information measures in general probability spaces. NeurIPS 2018

References (3)

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

Runge-2020 Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. UAI 2020

Smith-2011 Stephen M. Smith, Karla L. Miller, Gholamreza Salimi Khorshidi, Matthew A. Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph Ramsey, Mark W. Woolrich. Network modelling methods for fmri. NeuroImage, 54:875–891, 2011

Shpitser-2006 I. Shpitser, J. Pearl. Identification of joint interventional distributions in re- cursive semi-markovian causal models. AAAI 2006

Spirtes-2000 Peter Spirtes, Clark Glymour, Richard Scheines. Causation, Prediction, and Search. MIT press, 2nd edition, 2000

Sun-2015 Jie Sun, Dane Taylor, Erik M. Bollt. Causal network inference by optimal causation entropy. SIAM Journal on Applied Dynamical Systems, 14(1):73–106, 201

Zhang-2008 Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence, 172(16):1873 – 1896, 2008

Publications on presented material

Assaad-2021 K. Assaad, E. Devijver, E. Gaussier, A. Aït-Bachir. A Mixed Noise and Constraint-Based Approach to Causal Inference in Time Series. ECML/PKDD 2021

Assaad-2022(a) C. K. Assaad, E. Devijver, E. Gaussier. Survey and evaluation of causal discovery methods for time series. Journal of Artificial Intelligence Research, 73: 767–819, feb 2022.

Assaad-2022(b) C. K. Assaad, E. Devijver, E. Gaussier. Causal Discovery of Extended Summary Graphs in Time Series. UAI 2022

Assaad-2022(c) C. K. Assaad, E. Devijver, E. Gaussier. Entropy-Based Discovery of Summary Causal Graphs in Time Series. Entropy 24(8), 2022

Zan-2022 L. Zan, A. Meynaoui, C. K. Assaad, E. Devijver, E. Gaussier. A Conditional Mutual Information Estimator for Mixed Data and an Associated Conditional Independence Test. Entropy 24(8), 2022