Causal Discovery in Observational Time Series

E. Devijver,
C. K. Assaad, E. Gaussier

CNRS, Université Grenoble Alpes, France

September 23, 2022

Why causality?

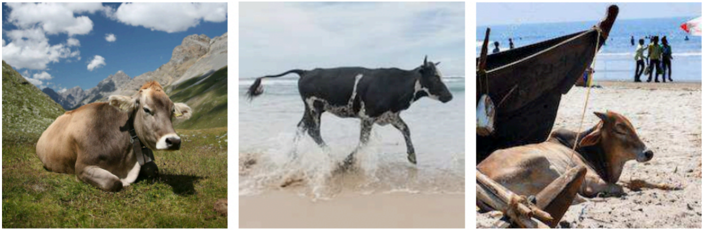
- Machine learning systems lack:

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works
  - react to events different from the training set



(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works
  - react to events different from the training set
  - go beyond correlation relationships

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works
  - react to events different from the training set
  - go beyond correlation relationships
  - capacity to answer what if, intervention, and counterfactual questions

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works
  - react to events different from the training set
  - go beyond correlation relationships
  - capacity to answer what if, intervention, and counterfactual questions
    - Was it the new powder that caused an increasing in porosity?

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works
  - react to events different from the training set
  - go beyond correlation relationships
  - capacity to answer what if, intervention, and counterfactual questions
    - Was it the new powder that caused an increasing in porosity?
    - What if my train had not been late?

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works
  - react to events different from the training set
  - go beyond correlation relationships
  - capacity to answer what if, intervention, and counterfactual questions
    - Was it the new powder that caused an increasing in porosity?
    - What if my train had not been late?
    - How effective is a treatment in preventing a disease?

Why causality?

- Machine learning systems lack:
  - ability to capture how the world works
  - react to events different from the training set
  - go beyond correlation relationships
  - capacity to answer what if, intervention, and counterfactual questions
    - Was it the new powder that caused an increasing in porosity?
    - What if my train had not been late?
    - How effective is a treatment in preventing a disease?
    - Can hiring records prove an employer guilty of gender discrimination?

Why causality for time series?

- Causality is crucial for explanatory purpose, since an effect can be explained by its causes, regardless of the correlations it may have with other variables
- Time series are everywhere but you know that ;)

Why causality for time series?

- Causality is crucial for explanatory purpose, since an effect can be explained by its causes, regardless of the correlations it may have with other variables
- Time series are everywhere but you know that ;)

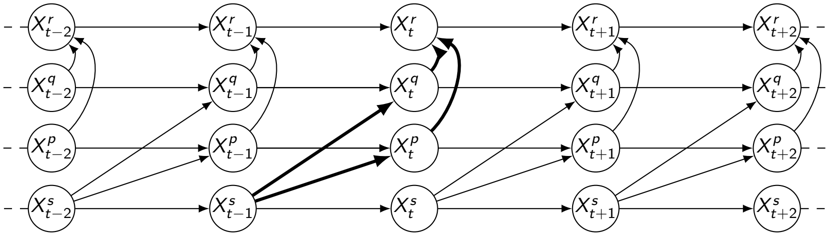


Figure: *Running example*: a diamond structure with self causes.

Outline

Causal graphs for time series

A d-variate time series X of continuous values
For a fixed t, each X_t is a vector $(X_t^1, \ldots, X_t^d)$,
in which X_t^p is the measurement of the pth time series at time t.

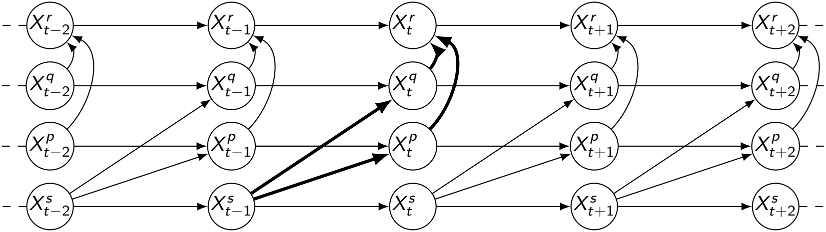
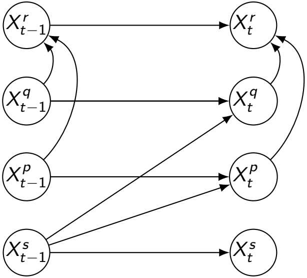


Figure: Full time causal graph.

Causal graphs for time series

A d-variate time series X of continuous values
For a fixed t, each X_t is a vector $(X_t^1, \ldots, X_t^d)$,
in which X_t^p is the measurement of the pth time series at time t.

A causal graph for a multivariate time series X is said to be *consistent throughout time* if all the causal relationships remain constant in direction throughout time.

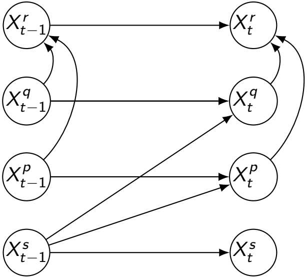


Window Causal Graph

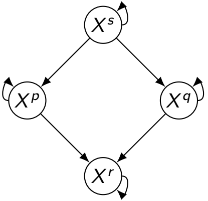
Causal graphs for time series

A d-variate time series X of continuous values
For a fixed t, each X_t is a vector $(X_t^1, \ldots, X_t^d)$,
in which X_t^p is the measurement of the pth time series at time t.

A causal graph for a multivariate time series X is said to be *consistent throughout time* if all the causal relationships remain constant in direction throughout time.



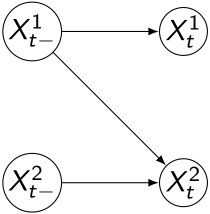
Window Causal Graph



Summary Causal Graph

Classical Assumptions

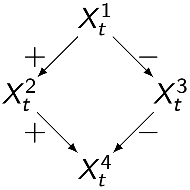
- A set of variables is said to be *causally sufficient* if all common causes of all variables are observed.
- A causal relation between two variables is said to satisfy the *temporal priority* if it is oriented in such a way that the cause occurred before its effect.
- *Causal Markov Condition*: (conditional) independence in the graph leads to (conditional) independence in the data.



$$X^1_t \perp\!\!\!\perp X^2_t \mid X^1_{t-}$$

Classical Assumptions

- A set of variables is said to be *causally sufficient* if all common causes of all variables are observed.
- A causal relation between two variables is said to satisfy the *temporal priority* if it is oriented in such a way that the cause occurred before its effect.
- *Causal Markov Condition*: (conditional) independence in the graph leads to (conditional) independence in the data.
- *Minimality condition*: the graph does not contain dependencies not present in the observational data.
- *Faithfulness*: only the conditional independence relations true in the data are entailed by the Causal Markov condition applied to the graph.



Outline

Granger Causality

X^p Granger-causes X^q if past values of X^p provide unique statistically significant information about future values of X^q.

Granger Causality

X^p Granger-causes X^q if past values of X^p provide unique statistically significant information about future values of X^q.

Pairwise Granger causality

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \xi_t^q, \qquad \text{(Mres)}$$

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \sum_{i=1}^{\tau} a_{p,i} X_{t-i}^p + \xi_t^q, \qquad \text{(Mfull)}$$

Statistical test (e.g. F-test) can be used to determine whether (Mfull) is significantly better than (Mres),

H_0: X^p does not Granger-cause X^q.

Granger Causality

X^p Granger-causes X^q if past values of X^p provide unique statistically significant information about future values of X^q.

Pairwise Granger causality

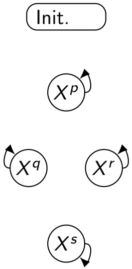


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

Granger Causality

X^p Granger-causes X^q if past values of X^p provide unique statistically significant information about future values of X^q.

Pairwise Granger causality

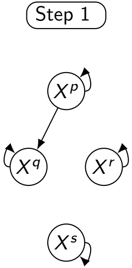


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

Granger Causality

X^p Granger-causes X^q if past values of X^p provide unique statistically significant information about future values of X^q.

Pairwise Granger causality

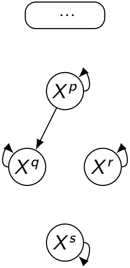


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

Granger Causality

X^p Granger-causes X^q if past values of X^p provide unique statistically significant information about future values of X^q.

Pairwise Granger causality

Step 6

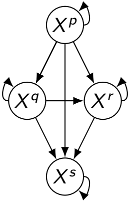


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

Granger Causality

Multivariate Granger causality

$$X_t^q = a_{q,0} + \sum_{\substack{r=1\\r\neq p}}^{d} \sum_{i=1}^{\tau} a_{r,i} X_{t-i}^p + \xi_t^q, \qquad \text{(mvMres)}$$

$$X_t^q = a_{q,0} + \sum_{r=1}^{d} \sum_{i=1}^{\tau} a_{r,i} X_{t-i}^r + \xi_t^q, \qquad \text{(mvMfull)}$$

Extensions

- Non-linear associations
- Nonstationnarity

Constraint-based approaches

Constraint-based approaches exploit conditional independencies to build a skeleton between variables. This skeleton is then oriented according to a set of rules that define constraints on admissible orientations.

Constraint-based approaches

Constraint-based approaches exploit conditional independencies to build a skeleton between variables. This skeleton is then oriented according to a set of rules that define constraints on admissible orientations.

Assumptions

- Causal Markov Condition
- Faithfulness

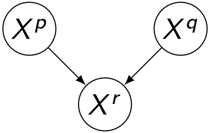
Constraint-based approaches

Constraint-based approaches exploit conditional independencies to build a skeleton between variables. This skeleton is then oriented according to a set of rules that define constraints on admissible orientations.

Assumptions

- Causal Markov Condition
- Faithfulness

v-structures (colliders): only structures which can be oriented without ambiguity.



Constraint-based approaches

Constraint-based approaches exploit conditional independencies to build a skeleton between variables. This skeleton is then oriented according to a set of rules that define constraints on admissible orientations.

Assumptions

- Causal Markov Condition
- Faithfulness

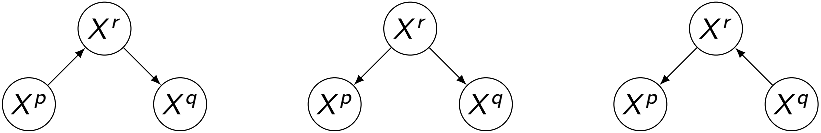


Figure: Three equivalent structures: $X^p \perp\!\!\!\perp X^q | X^r$

Markov equivalence class: set of DAGs that encode the same set of conditional independencies.

Constraint-based approaches

PCMCI

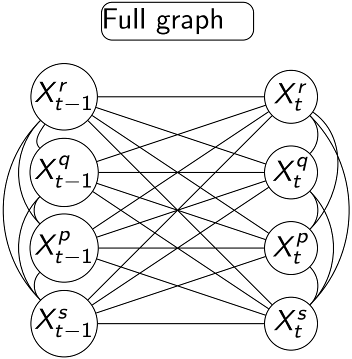


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

Constraint-based approaches
PCMCI

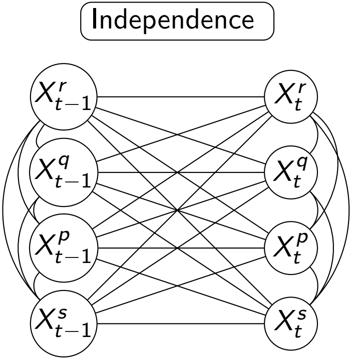


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

Constraint-based approaches

PCMCI

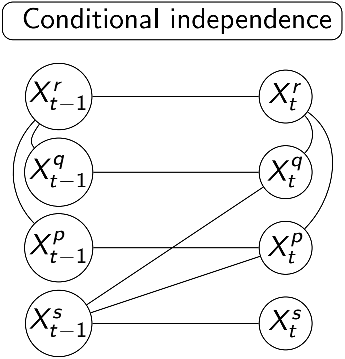


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

Constraint-based approaches
PCMCI

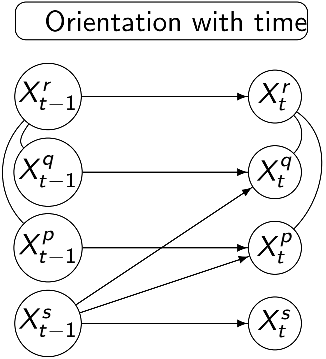


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

Constraint-based approaches

PCMCI

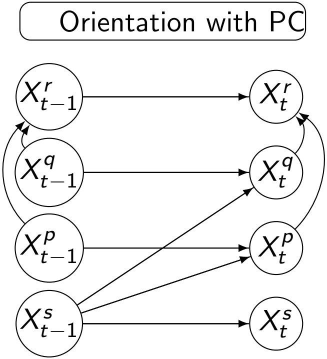


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

Noise-based approaches

Causal system described by a set of equations, where each equation explains one variable of the system in terms of its direct causes and some additional noise.

Assumptions

- Causal Markov Condition
- Minimality

Can deal with 2 variables

Additive noise model with nonlinear functions

$$X^p = \xi^p,$$
$$X^q = f_q(X^p) + \xi^q \qquad \text{with } X^p \perp\!\!\!\perp \xi_q.$$

Noise-based approaches

Additive Noise Models

Additive noise model with nonlinear functions

$$X^p = \xi^p,$$
$$X^q = f_q(X^p) + \xi^q \qquad \text{with } X^p \perp\!\!\!\perp \xi_q.$$

Theorem (Identifiability of ANMs)

Assume that the conditional distribution of $X^q \mid X^p$ admits a smooth ANM, and that there exists $x_q \in \mathbb{R}$ such that, for almost all $x_p \in \mathbb{R}$,

$$(\log p_{\xi^q})''(x_q - f_q(x_p))f_q'(x_p) \neq 0.$$

Then, the set of log densities $\log p_X$ for which the obtained joint distribution P_{X^p,X^q} admits a smooth ANM from X^q to X^p is contained in a 3-dimensional affine space.

Additive noise model with nonlinear functions

$$X^p = \xi^p,$$
$$X^q = f_q(X^p) + \xi^q \qquad \text{with } X^p \perp\!\!\!\perp \xi_q.$$

Principle (Multivariate additive noise principle)

Suppose we are given a joint distribution $P(X^1, \cdots, X^d)$. *If it satisfies an identifiable Additive Noise Model such that* $\{(X^p_{t-j})_{1\leq p\neq q\leq d, 0\leq j\leq \tau}, (X^q_{t-j})_{1\leq j\leq \tau}\} \rightarrow X^q$, *then it is likely that* $\{(X^p_{t-j})_{1\leq p\neq q\leq d, 0\leq j\leq \tau}, (X^q_{t-j})_{1\leq j\leq \tau}\}$ *precedes* X^q *in the causal order.*

Noise-based approaches

VarLINGAM

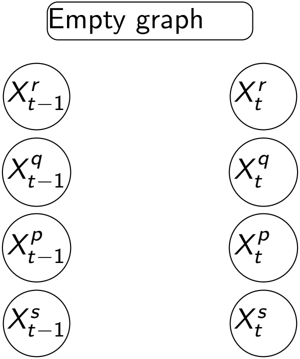


Figure: Running example: structured inferred by VarLiNGAM.

Noise-based approaches

VarLINGAM

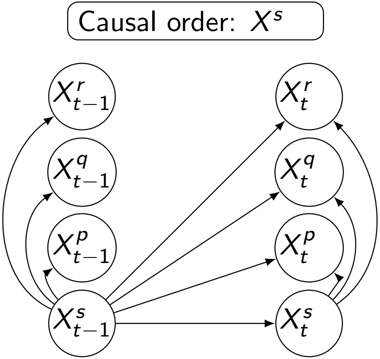


Figure: Running example: structured inferred by VarLiNGAM.

Noise-based approaches

VarLINGAM

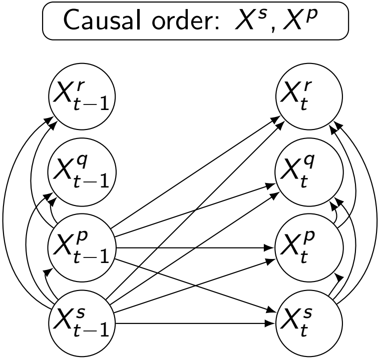


Figure: Running example: structured inferred by VarLiNGAM.

Noise-based approaches

VarLINGAM

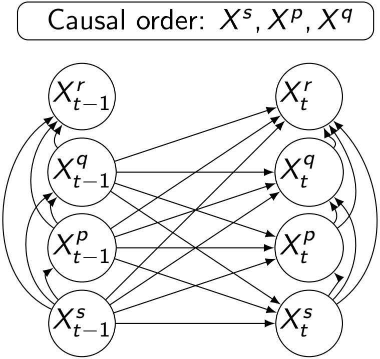


Figure: Running example: structured inferred by VarLiNGAM.

Noise-based approaches
VarLINGAM

Causal order: X^s, X^p, X^q, X^r

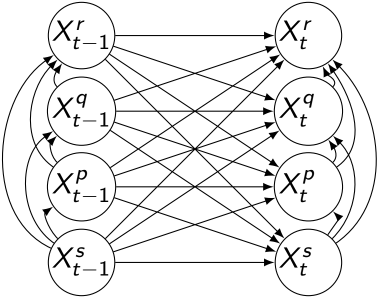


Figure: Running example: structured inferred by VarLiNGAM.

Noise-based approaches

VarLINGAM

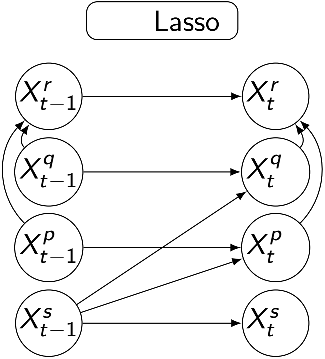


Figure: Running example: structured inferred by VarLiNGAM.

Outline

NBCB[1]

A mix between noise-based and constraint-based approaches

Assumptions

- Causal Markov Condition
- Adjacency faithfulness: if X^p and X^q are adjacent, then they are not conditionally independent given any subset of vertices except Xp, Xq.
- Minimality

[1]ECMLPKDD 2021

NBCB[1]

A mix between noise-based and constraint-based approaches

Assumptions

- Causal Markov Condition
- Adjacency faithfulness: if X^p and X^q are adjacent, then they are not conditionally independent given any subset of vertices except Xp, Xq.
- Minimality

Representation of the time series

Optimal lag γ_{pq} and $(\lambda_{pq}, \lambda_{qp})$ the *optimal* windows:

$$\gamma_{pq}, \lambda_{pq}, \lambda_{qp} = \underset{\gamma \geq 0, \lambda_1, \lambda_2}{\text{argmax}} \ h(X^q_{t:t+\lambda_2} \mid X^q_{t-1}, X^p_{t-\gamma-1}) - h(X^q_{t:t+\lambda_2} \mid X^p_{t-\gamma-1:t-\gamma+\lambda_1}, X^q_{t-1}).$$

where h denotes the entropy.

[1] ECMLPKDD 2021

NBCB

A mix between noise-based and constraint-based approaches

Step 1: causal ordering (additive noise model)
Last place: time series which yields the residuals that are more independent to the other time series.
Step 2: pruning to remove spurious relations

$$\begin{aligned}
&\mathrm{TCE}(X^p \rightarrow X^q \mid X^{\mathbf{R}}) = \\
&\min_{\Gamma_{r_i} \geq 0,\, 1 \leq i \leq K} \; h(X^q_{t:t+\lambda_{qp}} \mid (X^{r_i}_{t-\Gamma_{pq|r_i}})_{1 \leq i \leq K}, X^q_{t-1}, X^p_{t-\gamma_{pq}-1}) \\
&- h(X^q_{t:t+\lambda_{qp}} \mid (X^{r_i}_{t-\Gamma_{pq|r_i}})_{1 \leq i \leq K}, X^p_{t-\gamma_{pq}-1:t-\gamma_{pq}+\lambda_{pq}}, X^q_{t-1}),
\end{aligned}$$

where $\Gamma_{pq|r_1}, \cdots, \Gamma_{pq|r_K}$ are the lags between $X^{\mathbf{R}}$ and X^q.

NBCB

A mix between noise-based and constraint-based approaches

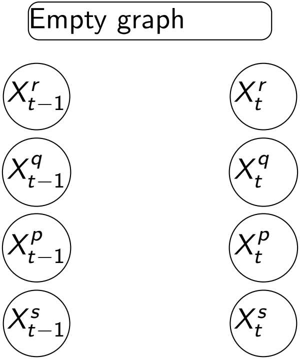


Figure: Running example: structured inferred by NBCB.

NBCB

A mix between noise-based and constraint-based approaches

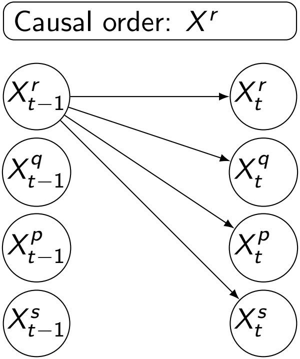


Figure: Running example: structured inferred by NBCB.

NBCB

A mix between noise-based and constraint-based approaches

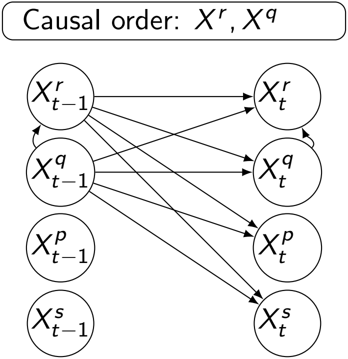


Figure: Running example: structured inferred by NBCB.

NBCB

A mix between noise-based and constraint-based approaches

Causal order: X^r, X^q, X^p

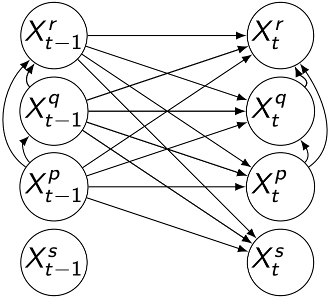


Figure: Running example: structured inferred by NBCB.

NBCB

A mix between noise-based and constraint-based approaches

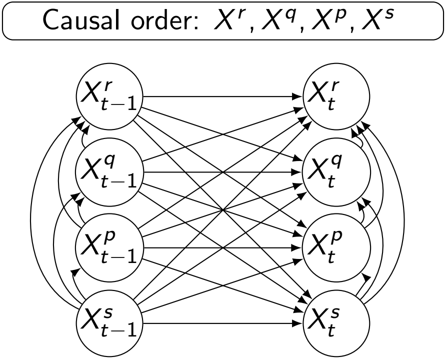


Figure: Running example: structured inferred by NBCB.

NBCB

A mix between noise-based and constraint-based approaches

Conditional independence using TCE

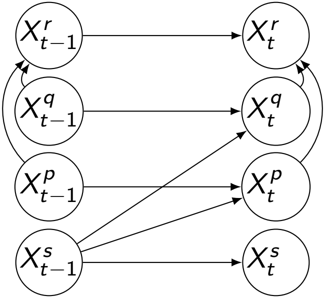
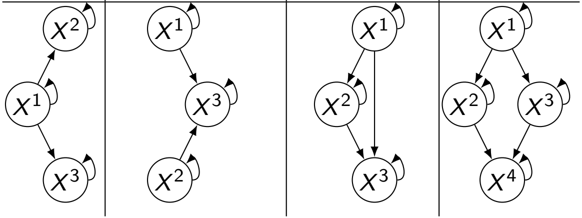


Figure: Running example: structured inferred by NBCB.

Experiments

Table: Structures of simulated data.

Fork	V-structure	Mediator	Diamond



Experiments

Simulated data 1

Table: Results obtained on the simulated data for the different structures with 1000 observations. We report the mean and the standard deviation of the F1 score. The best results are in bold.

	V-struct	Fork	Mediator	Diamond
GC	0.37 ± 0.25	0.44 ± 0.38	0.83 ± 0.22	0.66 ± 0.26
PCMCI	0.67 ± 0.37	0.78 ± 0.17	0.84 ± 0.09	0.82 ± 0.16
VarLiNGAM	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.03 ± 0.08
TiMINo	0.65 ± 0.37	0.52 ± 0.44	0.80 ± 0.19	0.60 ± 0.25
NBCB	0.67 ± 0.28	0.67 ± 0.38	0.66 ± 0.32	0.71 ± 0.16

Table: Results obtained on the unfaithful simulated data for the different structures with 1000 observations. We report the mean and the standard deviation of the F1 score. The best results are in bold.

	unfaith. Mediator	unfaith. Diamond
GC	0.12 ± 0.27	0.14 ± 0.23
PCMCI	0.05 ± 0.15	0.20 ± 0.22
VarLiNGAM	0.0 ± 0.0	0.02 ± 0.06
TiMINo	$\mathbf{0.64} \pm 0.08$	0.49 ± 0.03
NBCB	0.56 ± 0.26	$\mathbf{0.5} \pm 0.31$

Experiments

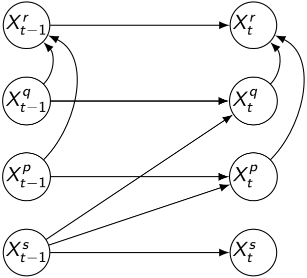
Real datasets

Table: Results for real datasets. We report the mean and the standard deviation of the F1 score.

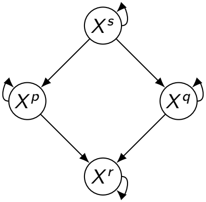
	Temperature	Diary	FMRI
GC	0.66	0.33	0.24 ± 0.18
PCMCI	1	0.5	0.22 ± 0.18
VarLiNGAM	0	0.0	0.49 ± 0.28
TiMINo	0	0.0	0.32 ± 0.11
NBCB	1	0.8	0.40 ± 0.21

Outline

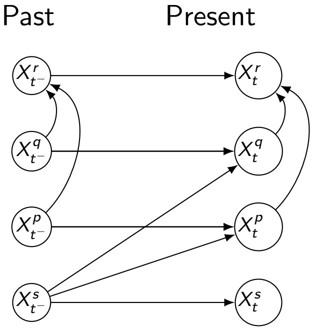
PCGCE[2]: discover an extended summary causal graph



Window Causal Graph



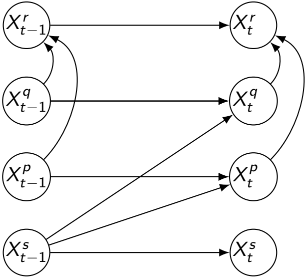
Summary Causal Graph



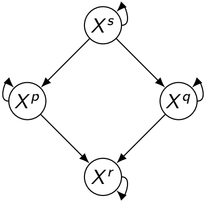
Extended Summary Causal Graph

[2] UAI 2022

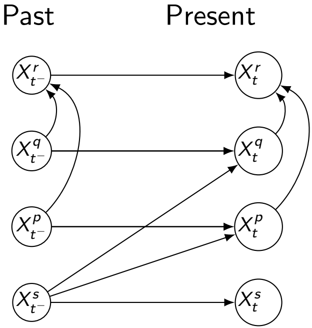
PCGCE[2]: discover an extended summary causal graph



Window Causal Graph



Summary Causal Graph



Extended Summary Causal Graph

Window causal graph

Difficult to be validated and analyzed by experts

Computationally expensive

---

[2] UAI 2022

PCGCE: discover an extended summary causal graph

Assumptions

- Causal Markov condition
- Faithfulness
- Causal sufficiency for PCGCE (but extension to FCIGCE)

PCGCE: discover an extended summary causal graph

Assumptions

- Causal Markov condition
- Faithfulness
- Causal sufficiency for PCGCE (but extension to FCIGCE)

Measure: Greedy Causation Entropy (GCE)

$$\mathrm{GCE}(X^p \to X^q | X^{\mathbf{Pa}}, X^{\mathbf{Pr}}) \\ := I(X_t^q; X_{t-\gamma:t-1}^p | X_{t-}^{Pa_1}, \cdots, X_{t-}^{Pa_l}, X_t^{Pr_1}, \cdots, X_t^{Pr_m})$$

- γ: *maximum gap between a cause and its effect*
- X_{t-}^p do not cause X_t^q iff there exists
  - $X^{\mathbf{Pr}} = \{X_t^{Pr_1}, \cdots, X_t^{Pr_m}\}$ and $X^{\mathbf{Pa}} = \{X_{t-}^{Pa_1}, \cdots, X_{t-}^{Pa_l}\}$ s.t.
  - $\mathrm{GCE}(X^p \to X^q | X^{\mathbf{Pa}}, X^{\mathbf{Pr}}) = 0 \qquad (1)$
- $\mathtt{Sepset}(p \leftrightarrow q)$ = smallest $X^{\mathbf{Pa}}, X^{\mathbf{Pr}}$ that satisfy (1)

PCGCE: discover an extended summary causal graph

Assumptions

- Causal Markov condition
- Faithfulness
- Causal sufficiency for PCGCE (but extension to FCIGCE)

Measure: Greedy Causation Entropy (GCE)

$$\mathrm{GCE}(X^p \to X^q | X^{\mathbf{Pa}}, X^{\mathbf{Pr}})$$
$$:= I(X^q_t; X^p_{t-\gamma:t-1} | X^{Pa_1}_{t-}, \cdots, X^{Pa_l}_{t-}, X^{Pr_1}_t, \cdots, X^{Pr_m}_t)$$

- γ: *maximum gap between a cause and its effect*
- X^p_{t-} do not cause X^q_t iff there exists
  - $X^{\mathbf{Pr}} = \{X^{Pr_1}_t, \cdots, X^{Pr_m}_t\}$ and $X^{\mathbf{Pa}} = \{X^{Pa_1}_{t-}, \cdots, X^{Pa_l}_{t-}\}$ s.t.
  - $\mathrm{GCE}(X^p \to X^q | X^{\mathbf{Pa}}, X^{\mathbf{Pr}}) = 0 \qquad (1)$
- $\mathtt{Sepset}(p \leftrightarrow q)$ = smallest $X^{\mathbf{Pa}}, X^{\mathbf{Pr}}$ that satisfy (1)

Estimation and testing

kNN estimator and local permutation test

PCGCE: discover an extended summary causal graph

Running example

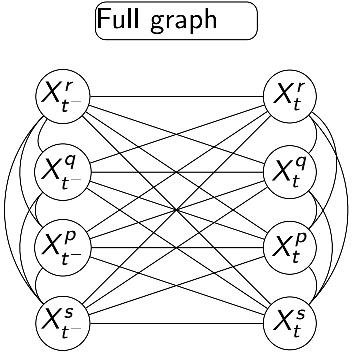


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

PCGCE: discover an extended summary causal graph

Running example

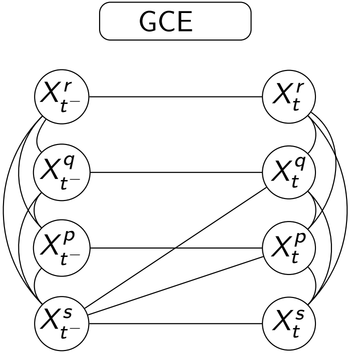


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

PCGCE: discover an extended summary causal graph

Running example

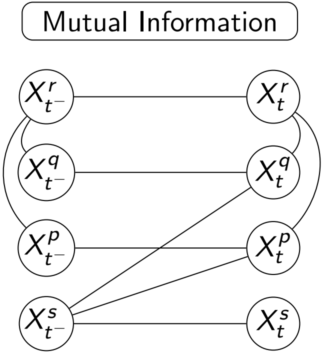


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

PCGCE: discover an extended summary causal graph

Running example

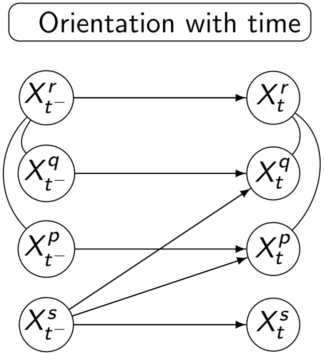


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

PCGCE: discover an extended summary causal graph

Running example

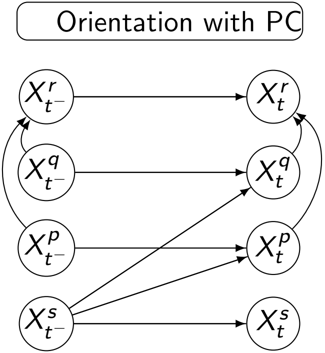


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

PCGCE

Experiments and conclusion

- Several experiments to understand the gain
- Performances comparable to PCMCI, but algorithm much faster

- Several experiments to understand the gain
- Performances comparable to PCMCI, but algorithm much faster
- PCGCE and FCIGCE can loose performance for high maximum time lags, compared to window-approaches
- We can think of reducing the dimension of the past slice (using autoencoders?)

Outline

Conclusion and perspectives

- Many families to discover causal graph for time series (more than the one introduced here: score-based, logic-based, topology-based, difference-based)
- Hybrid methods can take benefit of several worlds
- An important question is: which causal graph do we want to infer?
- The representation of time series is essential (windows - lags)

Conclusion and perspectives

- Many families to discover causal graph for time series (more than the one introduced here: score-based, logic-based, topology-based, difference-based)
- Hybrid methods can take benefit of several worlds
- An important question is: which causal graph do we want to infer?
- The representation of time series is essential (windows - lags)

- Causal discovery on mixed data? Ongoing work with Lei Zan
- Interventions in graphs for time series? Ongoing work with Anouar Meynaoui

Conclusion and perspectives

- Many families to discover causal graph for time series (more than the one introduced here: score-based, logic-based, topology-based, difference-based)
- Hybrid methods can take benefit of several worlds
- An important question is: which causal graph do we want to infer?
- The representation of time series is essential (windows - lags)

- Causal discovery on mixed data? Ongoing work with Lei Zan
- Interventions in graphs for time series? Ongoing work with Anouar Meynaoui

- Some ad: we are organizing a trimester on causality (between Paris, Grenoble and Saclay) in April, May, June 2023!

References

- Charles K. Assaad, Emilie Devijver, and Eric Gaussier. *Survey and evaluation of causal discovery methods for time series*. JAIR, 73, 2022.
- Charles K. Assaad, Emilie Devijver, and Eric Gaussier. *A Mixed Noise and Constraint Based Approach to Causal Inference in Time Series*, ECMLPKDD 2021
- Charles K. Assaad, Emilie Devijver, and Eric Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022
- Jakob Runge. *Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information*. In AISTATS, pages 938–947, 2018.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.