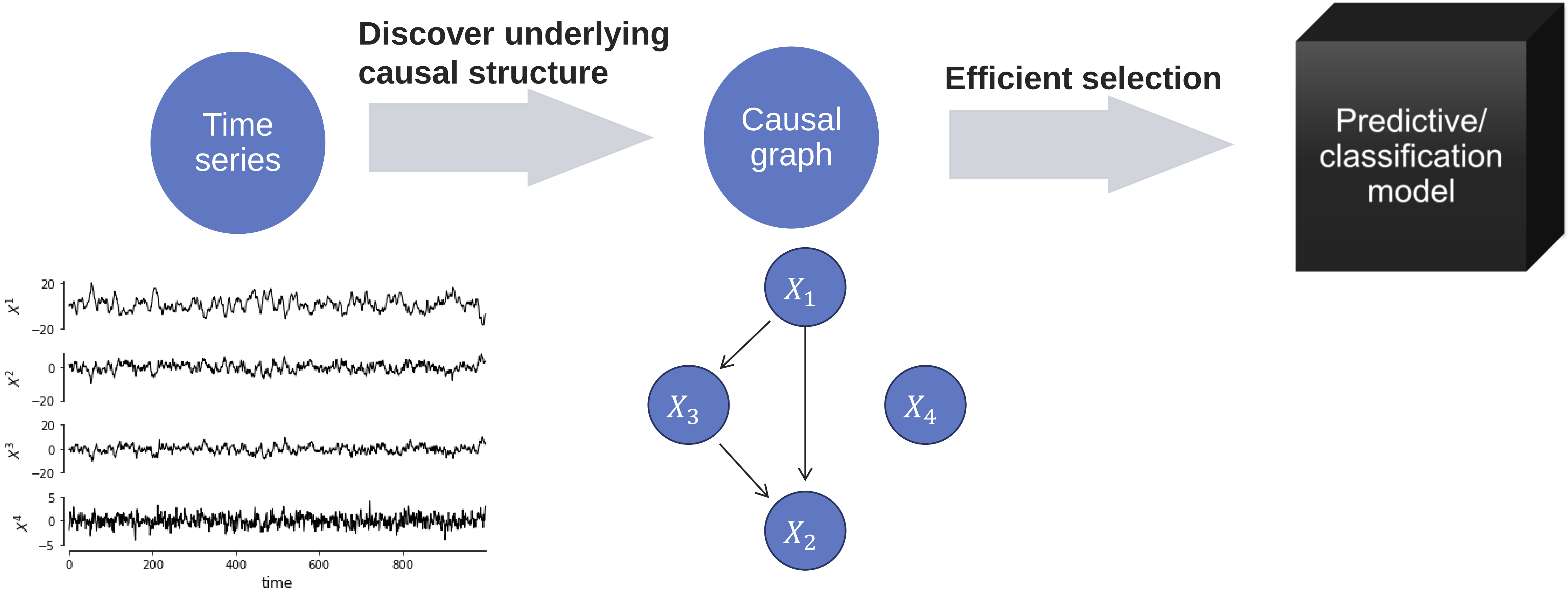
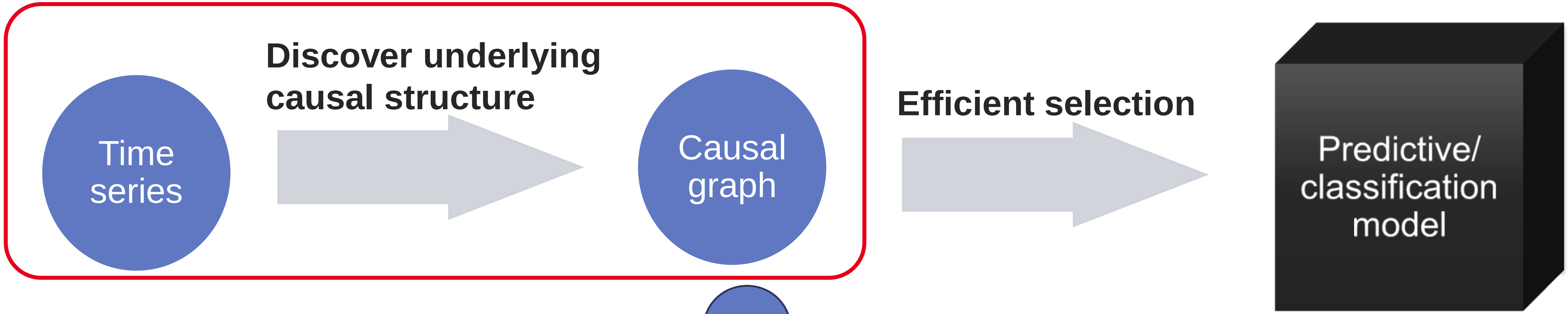
Causal discovery for time series with constraint-based model and PMIME measure

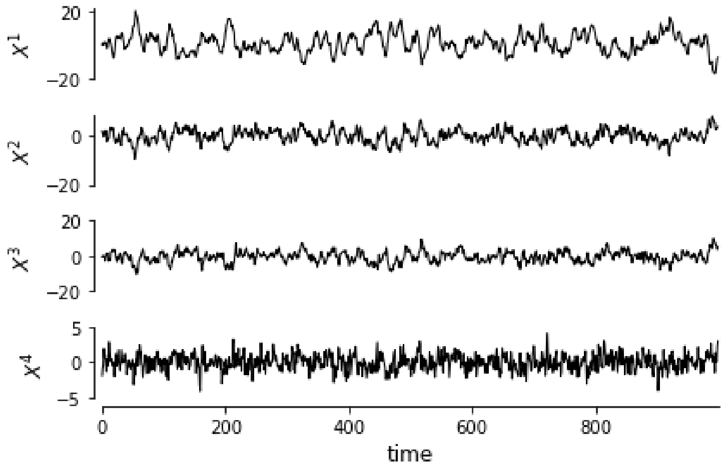
Antonin Arsac, Aurore Lomet, Jean-Philippe Poli

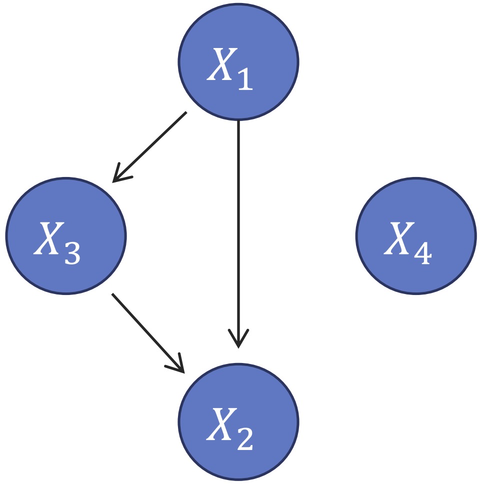
Context


Time series
Discover underlying causal structure
Causal graph
Efficient selection
Predictive/ classification model
20
−20
0
−20
20
0
−20
5
0
−5
0
200
400
600
800
time
X_1
X_3
X_4
X_2

Context


Discover underlying
causal structure
Time
series
Causal
graph
Efficient selection
Predictive/
classification
model


20
−20
0
−20
20
0
−20
5
0
−5
0
200
400
600
800
time


X_1
X_3
X_4
X_2

Causality

- **Making cause and effect relationships is at the basis ouf the human way of thinking**

- **Correlation is not causation:**

  Measuring dependencies between observational data is not enough to fully grasp the causal model

  → Connect statistical dependencies and causation

Causality

- **Making cause and effect relationships is at the basis ouf the human way of thinking**

- **Correlation is not causation:**

  Measuring dependencies between observational data is not enough to fully grasp the causal model

  → Connect statistical dependencies and causation

- **Causation :**

  Variable X causes variable Y if an intervention on X (and only X) can change Y

In the last decades, **causal inference** theory largely developed in *e.g.* [Pearl, 2009, Spirtes et al., 2000, Peters et al. 2017].

Necessary notions

Causal Bayesian networks (CBN) defined by:

- Set of random variables $\boldsymbol{X} = (X^1, X^2, \dots, X^g)$ following distribution $\mathcal{P}$
- A DAG $\mathcal{G} = (\mathcal{V}, E)$, in which each node from $\mathcal{V}$ is associated to a variable in $\boldsymbol{X}$
  - ➢ Arrows connecting two nodes stands for direct dependency
  - ➢ No arrow between two nodes show either independence or conditional independence

Necessary notions

Causal Bayesian networks (CBN) defined by:

- Set of random variables $\boldsymbol{X} = (X^1, X^2, \ldots, X^g)$ following distribution $\mathcal{P}$
- A DAG $\mathcal{G} = (\mathcal{V}, E)$, in which each node from $\mathcal{V}$ is associated to a variable in $\boldsymbol{X}$
  - ➢ Arrows connecting two nodes stands for direct dependency
  - ➢ No arrow between two nodes show either independence or conditional independence

Temporal priority property:

- A causal relationship oriented in a way such that a cause precedes its effect
  - ➢ Causality **asymmetrical** in time

A simple example

Suppose the generative model of the system follows a **Structural Causal Model** (SCM):

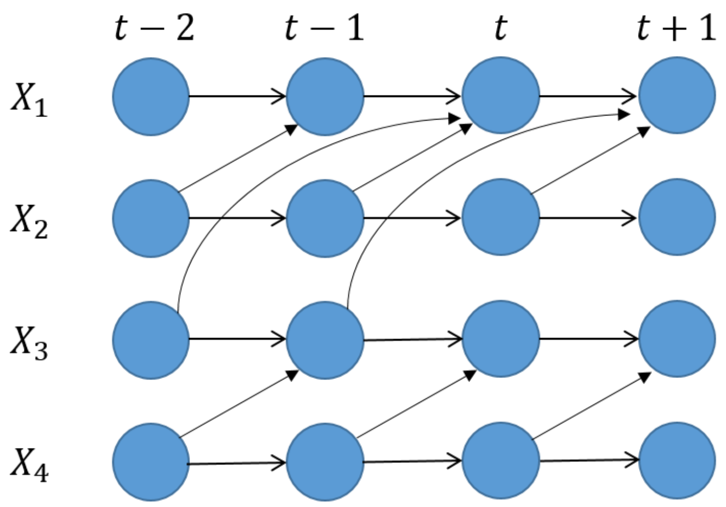
$\boldsymbol{X} = (X^1, X^2, X^3, X^4)$ a multivariate time series

$$X_t^1 = f_1(X_{t-1}^1, X_{t-1}^2, X_{t-2}^3, \epsilon_t^1)$$

$$X_t^2 = f_2(X_{t-1}^2, \epsilon_t^2)$$

$$X_t^3 = f_3(X_{t-1}^3, X_{t-1}^4, \epsilon_t^3)$$

$$X_t^4 = f_4(X_{t-1}^4, \epsilon_t^4)$$



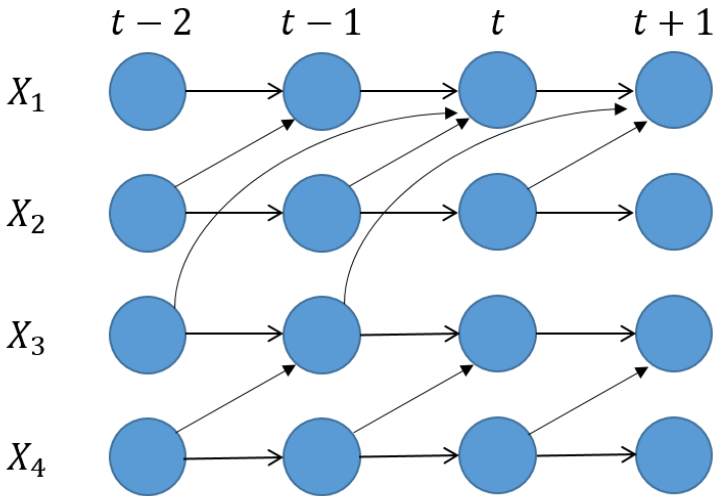
Window causal graph

A simple example

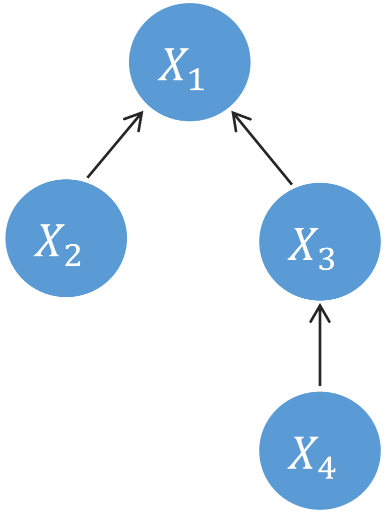
Suppose the generative model of the system follows a **Structural Causal Model** (SCM):

$\boldsymbol{X} = (X^1, X^2, X^3, X^4)$ a multivariate time series

$$
\begin{aligned}
X_t^1 &= f_1(X_{t-1}^1, X_{t-1}^2, X_{t-2}^3, \epsilon_t^1) \\
X_t^2 &= f_2(X_{t-1}^2, \epsilon_t^2) \\
X_t^3 &= f_3(X_{t-1}^3, X_{t-1}^4, \epsilon_t^3) \\
X_t^4 &= f_4(X_{t-1}^4, \epsilon_t^4)
\end{aligned}
$$



Window causal graph



Summary causal graph

Causal graph shows **direct** dependency between each random variables.

Framework

- Observations $\boldsymbol{X} = (X^1, \dots, X^g) \sim \mathcal{P}$
  - ➢ Multivariate time series
  - ➢ Joint probabilities generated by linear or non linear model
  - ➢ No assumption on the probability distribution of the observed model
  - ➢ All causes of each effect observed (causal sufficiency)

Framework

- Observations $\boldsymbol{X} = (X^1, \dots, X^g) \sim \mathcal{P}$
  - ➢ Multivariate time series
  - ➢ Joint probabilities generated by linear or non linear model
  - ➢ No assumption on the probability distribution of the observed model
  - ➢ All causes of each effect observed (causal sufficiency)

Discover causal relationships between time series data

- **Constraint-based** causal discovery algorithm
  - ➢ Need a **conditional independence measure**

PC algorithm

From the sets of conditional independence of observed data, build the causal graph

- **Input:** completed non oriented graph formed from the data
- **Outpout:** a Completed Partially DAG (CPDAG)

PC algorithm

From the sets of conditional independence of observed data, build the causal graph

- **Input:** completed non oriented graph formed from the data
- **Outpout:** a Completed Partially DAG (CPDAG)

- **Three different steps:**

1. Discover the sets of conditional independence in $\mathcal{G}$
   - start with an empty conditional set → increase its size with the parents of the tested variables
2. Find v-structures and orient them
3. Use knowledge given in step 1 and 2 to finish to orient the graph (so called PC rules)

PC algorithm

From the sets of conditional independence of observed data, build the causal graph

- **Input:** completed non oriented graph formed from the data
- **Outpout:** a Completed Partially DAG (CPDAG)

Find a conditional independence test

- **Three different steps:**

1. Discover the sets of conditional independence in $\mathcal{G}$
   - start with an empty conditional set → increase its size with the parents of the tested variables
2. Find v-structures and orient them
3. Use knowledge given in step 1 and 2 to finish to orient the graph (so called PC rules)

Partial Mutual Information from Mixed Embedding

PMIME: Quantify **direct** and **directional** dependencies from stationary multivariate time series (kugiumtzis, 2013)

- Based on **information theory**
  - ➢ Restrict assumptions on the data

- Built for multivariate time series

- Returns a bounded value
  - ➢ Easier to interpret
  - ➢ No additional statistical significance test

Partial Mutual Information from Mixed Embedding

PMIME: Quantify **direct** and **directional** dependencies from stationary multivariate time series (kugiumtzis, 2013)

- Based on **information theory**
  - Restrict assumptions on the data
- Built for multivariate time series
- Returns a bounded value
  - Easier to interpret
  - No additional statistical significance test

Few parameters:
- Stopping criterion
- Maximal lag
- Estimation parameters

Partial Mutual Information from Mixed Embedding

Consider g time series X, Y and $\boldsymbol{Z} = (Z^1, \dots, Z^{g-2})$

- Build iteratively an **embedding vector** $\boldsymbol{w}_t$, with lagged components from X, Y and $\boldsymbol{Z}$ that explain Y the most
- Let $\boldsymbol{w}^x, \boldsymbol{w}^y, \boldsymbol{w}^z$ be the components of X, Y and $\boldsymbol{Z}$ in the embedding vector $\boldsymbol{w_t}$ and $Y_t^T = (Y_{t+1}, \dots, Y_{t+T})$ the future of Y

Partial Mutual Information from Mixed Embedding

Consider g time series X, Y and $\boldsymbol{Z} = (Z^1, \dots, Z^{g-2})$

- Build iteratively an **embedding vector** $\boldsymbol{w}_t$, with lagged components from X, Y and $\boldsymbol{Z}$ that explain Y the most
- Let $\boldsymbol{w}^x, \boldsymbol{w}^y, \boldsymbol{w}^z$ be the components of X, Y and $\boldsymbol{Z}$ in the embedding vector $\boldsymbol{w_t}$ and $Y_t^T = (Y_{t+1}, \dots, Y_{t+T})$ the future of Y

The effect of X on Y, conditional on $\boldsymbol{Z}$:

$$R_{X \to Y|\boldsymbol{Z}} = \frac{I(Y_t^T, \boldsymbol{w}^x | \boldsymbol{w}^y, \boldsymbol{w}^z)}{I(Y_t^T; \boldsymbol{w}_t)}.$$

R **bounded between 0 and 1**

PMIME : building the embedding vector

Select the elements by the (conditional) **mutual information** I

→ Add item to $\boldsymbol{w_t}$ only if it strictly increases the information already included in $\boldsymbol{w_t}$

- Define a maximal lag $\tau_{\max}$
- Set of all lagged components $\boldsymbol{\mathcal{W}} = (X_t, X_{t-1}, \dots, X_{t-\tau_{\max}}, Y_t, Y_{t-1}, \dots, Z^{g-2}_{t-\tau_{\max}})$

PMIME : building the embedding vector

Select the elements by the (conditional) **mutual information** I

→ Add item to $\boldsymbol{w_t}$ only if it strictly increases the information already included in $\boldsymbol{w_t}$

- Define a maximal lag $\tau_{\max}$
- Set of all lagged components $\mathcal{W} = (X_t, X_{t-1}, \dots, X_{t-\tau_{\max}}, Y_t, Y_{t-1}, \dots, Z_{t-\tau_{\max}}^{g-2})$

Principle:

1^{st} embedding cycle

$$w_t^1 = argmax_{w \in \boldsymbol{W}}\, I(Y_t^T, w)$$

$$\boldsymbol{w}_t^1 = (w_t^1)$$

→

2^{nd} embedding cycle

$$w_t^2 = argmax_{w \in \boldsymbol{W}}\, I(Y_t^T, w | \boldsymbol{w}_t^1)$$

$$\boldsymbol{w}_t^2 = (w_t^1, w_t^2)$$

→ … →

$$\boldsymbol{w}_t^j = (w_t^1, w_t^2, \dots, w_t^{j-1})$$

j^{th} embedding cycle

$$w_t^j = argmax_{w \in \boldsymbol{W}}\, I(Y_t^T, w | \boldsymbol{w}_t^{j-1})$$

$$\boldsymbol{w}_t^j = (w_t^1, w_t^2, \dots, w_t^j)$$

PMIME : building the embedding vector

Select the elements by the (conditional) **mutual information** I

→ Add item to $\boldsymbol{w_t}$ only if it strictly increases the information already included in $\boldsymbol{w_t}$

- Define a maximal lag $\tau_{\max}$
- Set of all lagged components $\boldsymbol{\mathcal{W}} = (X_t, X_{t-1}, \ldots, X_{t-\tau_{\max}}, Y_t, Y_{t-1}, \ldots, Z^{g-2}_{t-\tau_{\max}})$

Principle:

1^{st} embedding cycle

$w_t^1 = argmax_{w \in \boldsymbol{\mathcal{W}}} \, I(Y_t^T, w)$

↓

$\boldsymbol{w}_t^1 = (w_t^1)$

→

2^{nd} embedding cycle

$w_t^2 = argmax_{w \in \boldsymbol{\mathcal{W}}} \, I(Y_t^T, w | \boldsymbol{w}_t^1)$

↓

$\boldsymbol{w}_t^2 = (w_t^1, w_t^2)$

→ $\boldsymbol{w}_t^j = (w_t^1, w_t^2, \ldots, w_t^{j-1})$ … →

j^{th} embedding cycle

$w_t^j = argmax_{w \in \boldsymbol{\mathcal{W}}} \, I(Y_t^T, w | \boldsymbol{w}_t^{j-1})$

↓

$\boldsymbol{w}_t^j = (w_t^1, w_t^2, \ldots, w_t^j)$

Stop at iteration j if

$$\frac{I(Y_t^T, \boldsymbol{w}_t^{j-1})}{I(Y_t^T, \boldsymbol{w}_t)} > A$$

Proposed approach

Merging PC, a causal discovery algorithm, with PMIME, a measure of direct links between time series.

> ➢ **PC-PMIME a constraint based method**

Major assumptions :

- Causal sufficiency
- Causal stationarity
- Faithfulness
- Stationary time series

Proposed approach

Merging PC, a causal discovery algorithm, with PMIME, a measure of direct links between time series.

- **PC-PMIME a constraint based method**

Major assumptions :

- Causal sufficiency
- Causal stationarity
- Faithfulness
- Stationary time series

Causal relationships observed at time t remain the same at time $t + \tau$

Proposed approach

Merging PC, a causal discovery algorithm, with PMIME, a measure of direct links between time series.

- ➢ **PC-PMIME a constraint based method**

Major assumptions :

- Causal sufficiency
- Causal stationarity
- Faithfulness
- Stationary time series

$X \perp Y \mid Z \text{ in } \mathcal{P} \Leftrightarrow X \text{ and } Y$ *d-seperated by* Z in $\mathcal{G}$

PC-PMIME

Merging PC, a causal discovery algorithm, with PMIME, a measure of direct links between time series.

1. Start with a graph $\mathcal{G}$ with all vertices connected
2. Remove edges between independent variables
3. For each couple (A, B) linked by an edge and for each C having an edge linked to A or B, remove the edge $A - B$ if $A \perp\!\!\!\perp B \mid C$.
4. For each couple (A, B) linked by an edge and for each set $\{C, D\}$ where C and D are both adjacent to A or both adjacent to B, remove the edge $A - B$ if $A \perp\!\!\!\perp B \mid \{C, D\}$.
5. Go on augmenting the size of the conditioning set until there is no (A, B) with a sufficient amount of adjacent nodes.

PMIME in the bivariate case

⇩

$A \perp\!\!\!\perp B$ if $R_{A \to B} = 0$

PC-PMIME

Merging PC, a causal discovery algorithm, with PMIME, a measure of direct links between time series.

1. Start with a graph $\mathcal{G}$ with all vertices connected
2. Remove edges between independent variables
3. For each couple (A, B) linked by an edge and for each C having an edge linked to A or B, remove the edge $A - B$ if $A \perp\!\!\!\perp B \mid C$.
4. For each couple (A, B) linked by an edge and for each set $\{C, D\}$ where C and D are both adjacent to A or both adjacent to B, remove the edge $A - B$ if $A \perp\!\!\!\perp B \mid \{C, D\}$.
5. Go on augmenting the size of the conditioning set until there is no (A, B) with a sufficient amount of adjacent nodes.

PMIME in the bivariate case

⇓

$A \perp\!\!\!\perp B$ if $R_{A \to B} = 0$

PMIME in the multivariate case : when a conditioning set is involved

⇓

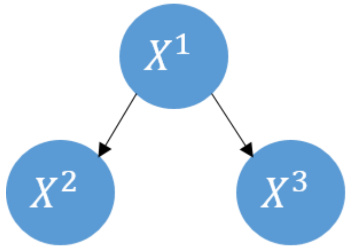
$A \perp\!\!\!\perp B \mid C$ if $R_{A \to B|C} = 0$

⇓

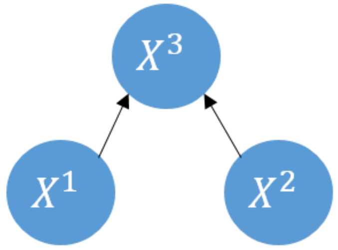
Orient with asymmetry:
$A \to B \; if \; R_{A \to B|C} > 0$

Experiments

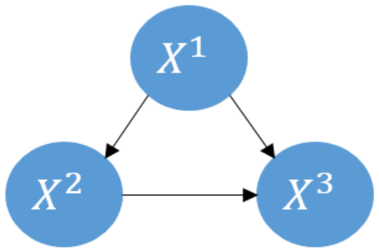
- Four basic causal structures simulated, from [Assaad et al. 2022]
- For each causal structure, 10 simulated datasets
- Linear auto-correlation for each variable and non-linear functions between a variable and parents



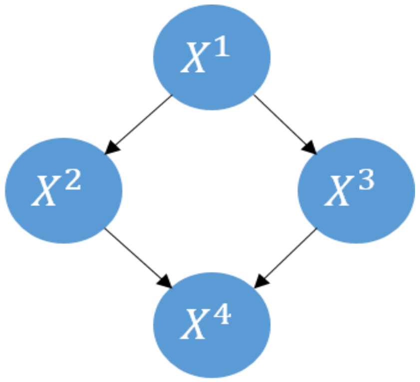
Fork



V-structure



Mediator



Diamond

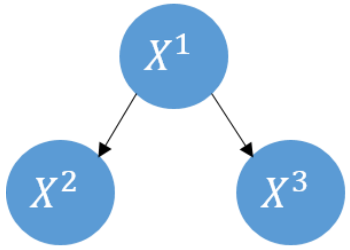
Experiments

- Four basic causal structures simulated, from [Assaad et al. 2022]
- For each causal structure, 10 simulated datasets
- Linear auto-correlation for each variable and non-linear functions between a variable and parents

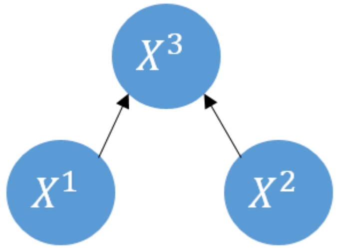
Example: a variable X^j simulated with:

$$\forall t > 0, \quad X_t^j = a_t^j X_{t-1}^j + \sum_{(p,\gamma)} a_{t-\gamma}^p f_p\left(X_{t-\gamma}^p\right) + 0.1\varepsilon_t^j$$

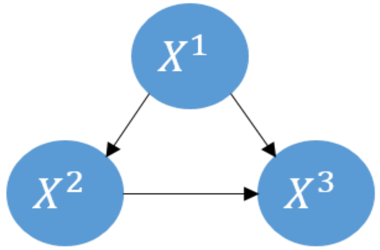
- $\gamma = \{1, \dots, \tau_p\}$ and $X^p \in Pa(X^j, \mathcal{G})$
- a_t^p random coefficients chosen in $\mathcal{U}([-1; -0{,}1] \cup [0.1; 1])$ for $1 \leq j \leq d$
- $\varepsilon_t^{g+1} \sim \mathcal{N}(0, \sigma)$
- f a non-linear function drawn in the list $[absolute\ value, \tanh, \sin e, cosine]$



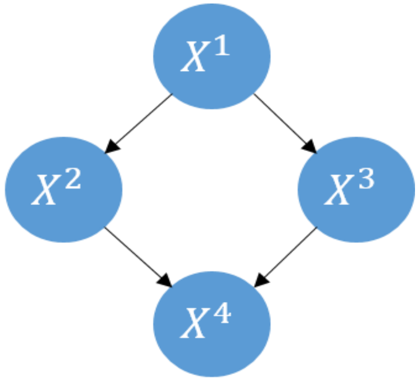
Fork



V-structure



Mediator



Diamond

Experiments

10 datasets of size $\mathrm{n} \in N = [125, 250, 500, 1000, 2000, 4000]$.

In PC-PMIME :

- Maximal lag considered: $\tau_{max} = 4$
- Threshold of the stopping criterion: $A = 0.03$
- Number of nearest neighbors: $k = 0.1n$

Experiments

10 datasets of size $\mathrm{n} \in N = [125, 250, 500, 1000, 2000, 4000]$.

In PC-PMIME :

- Maximal lag considered: $\tau_{max} = 4$
- Threshold of the stopping criterion: $A = 0.03$
- Number of nearest neighbors: $k = 0.1n$

$A \leq 0.01$ more restrictive
$A \geq 0.1$ more permissive

Experiments

10 datasets of size $\mathrm{n} \in N = [125, 250, 500, 1000, 2000, 4000]$.

In PC-PMIME :

- Maximal lag considered: $\tau_{max} = 4$
- Threshold of the stopping criterion: $A = 0.03$
- Number of nearest neighbors: $k = 0.1n$

Estimation of entropy by k-nearest neighbors

Experiments

10 datasets of size $\mathrm{n} \in N = [125, 250, 500, 1000, 2000, 4000]$.

In PC-PMIME :

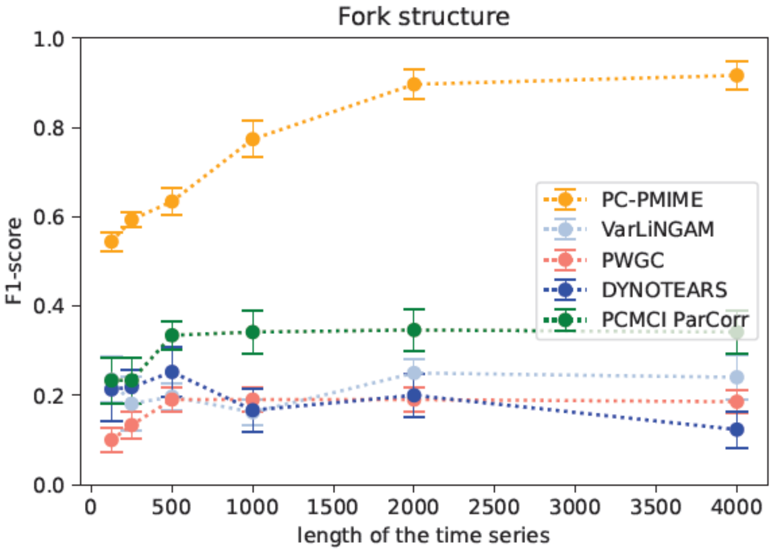
- Maximal lag considered: $\tau_{max} = 4$
- Threshold of the stopping criterion: $A = 0.03$
- Number of nearest neighbors: $k = 0.1n$

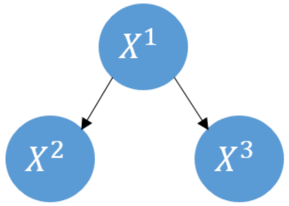
- Four other methods to compare with:
  - VarLiNGAM, PCMCI (Partial Correlation), Pairwise Granger Causality, DYNOTEARS

- Other methods exist but not tested here *e.g.* PCMCI derivatives, oCSE, MTE-MESS, Rhino…

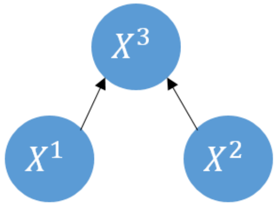
Results on simulations

- To evaluate the method : $F1$-score
- No consideration of auto-correlation in the score

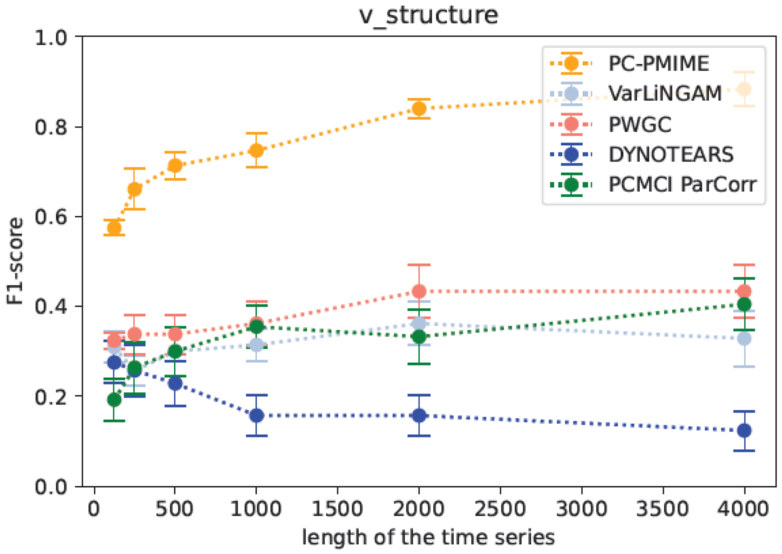




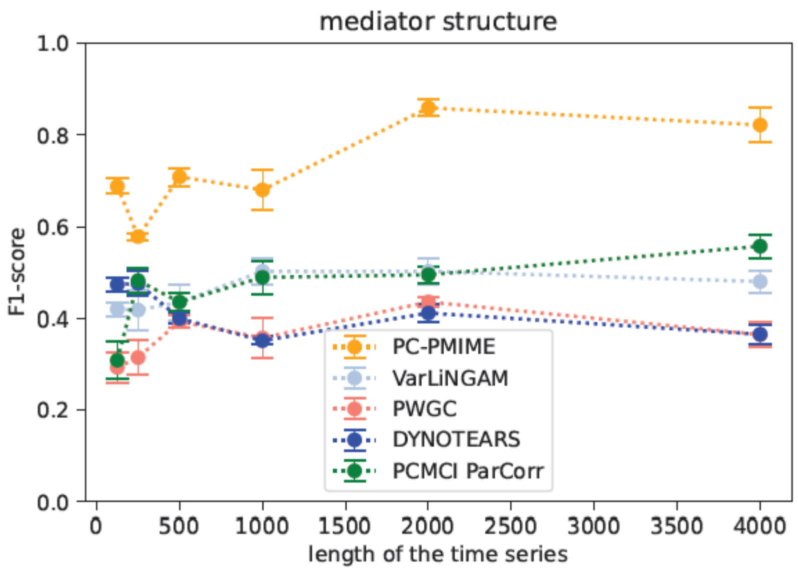
Fork

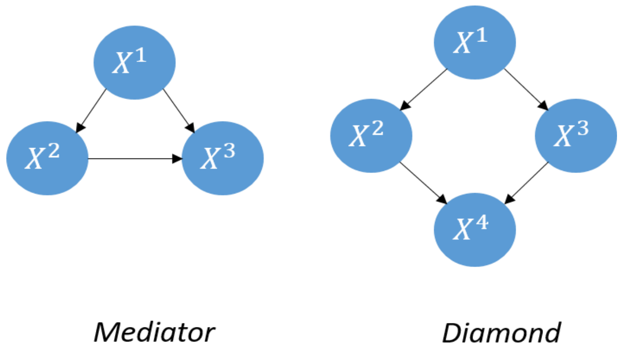


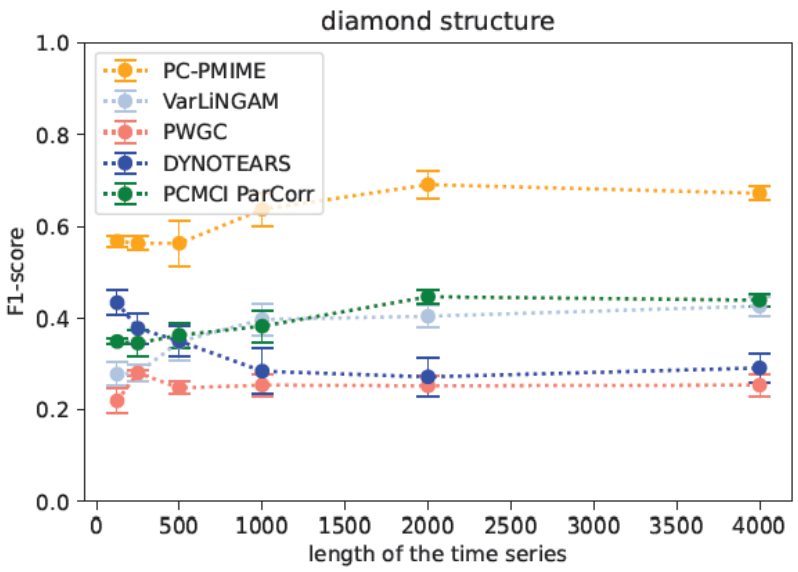
V-structure



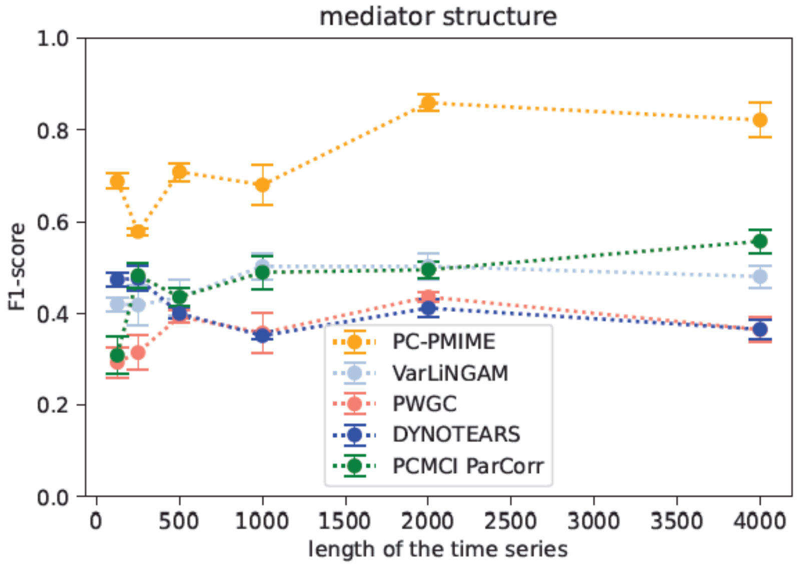
Results on simulations

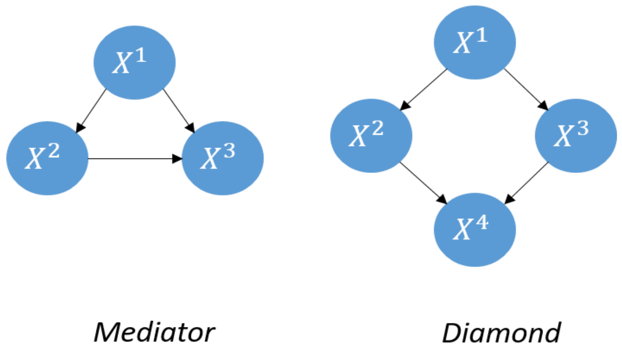


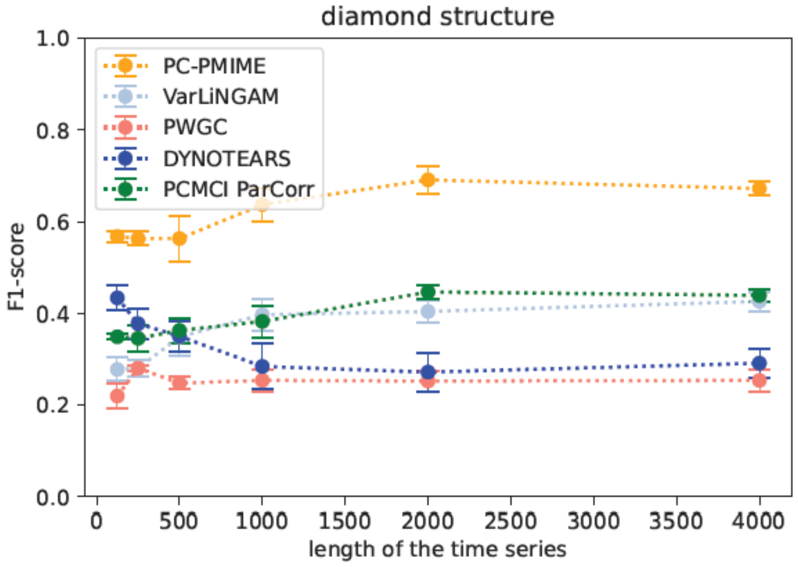




Results on simulations







High score for large size of times series ($n > 1000$)

→ Lack stability due to k-nn estimator

Conclusion and perspectives

- PC-PMIME shows very promising results on simulations
  - ➢ Tests on real data

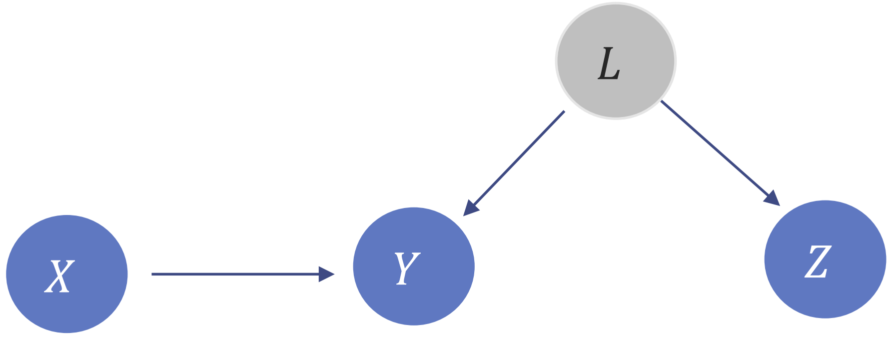
- Several limitations can be removed:
  - ➢ better orientations of edges
  - ➢ computing auto-correlation

Conclusion and perspectives

- PC-PMIME shows very promising results on simulations
  - ➢ Tests on real data

- Several limitations can be removed:
  - ➢ better orientations of edges
  - ➢ computing auto-correlation

- Causal sufficiency is not realistic:
  - ➢ Consider hidden counfounders in future work



Thanks for listening

antonin.arsac@cea.fr

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

References

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

[Kugiumtzis, 2013] Kugiumtzis, D. (2013). Direct-coupling information measure from nonuniform embedding. *Physical Review E*, *87*(6), 062918.

[Pearl, 2009] Pearl, J. (2009). *Causality*. Cambridge university press.

[Peters et al. 2017] Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

[Spirtes et al., 2000] Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

PCMCI: J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, D. Sejdinovic, Detecting and quantifying causal associations in large nonlinear time series datasets. Sci. Adv. 5, eaau4996 (2019).

VarLiNGAM: A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. Journal of Machine Learning Research, 11: 1709-1731, 2010.

Dynotears: Pamfil, R., Sriwattanaworachai, N., Desai, S., et al. Dynotears : Structure learning from time-series data. In International Conference on Artificial Intelligence and Statistics (2020), PMLR, pp. 1595–1605.