Data Imputation with Adversarial Neural Networks for Causal Discovery from Subsampled Time Series

Julio Muñoz-Benítez - jcmunoz@inaoep.mx
L. Enrique Sucar - esucar@inaoep.mx

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).

CaDis

INAOE

Tonantzintla, Puebla
June, 2023

Agenda

- Motivation
- Introduction
- Proposal
- Experimental Results
- Discussion
- Conclusions
- Future Work

Motivation

- One of the main objectives in causal discovery is to identify causal relations in dynamic phenomena (Runge et al., 2019). In this sense, there is a special interest in analyzing causal effects in time series data.

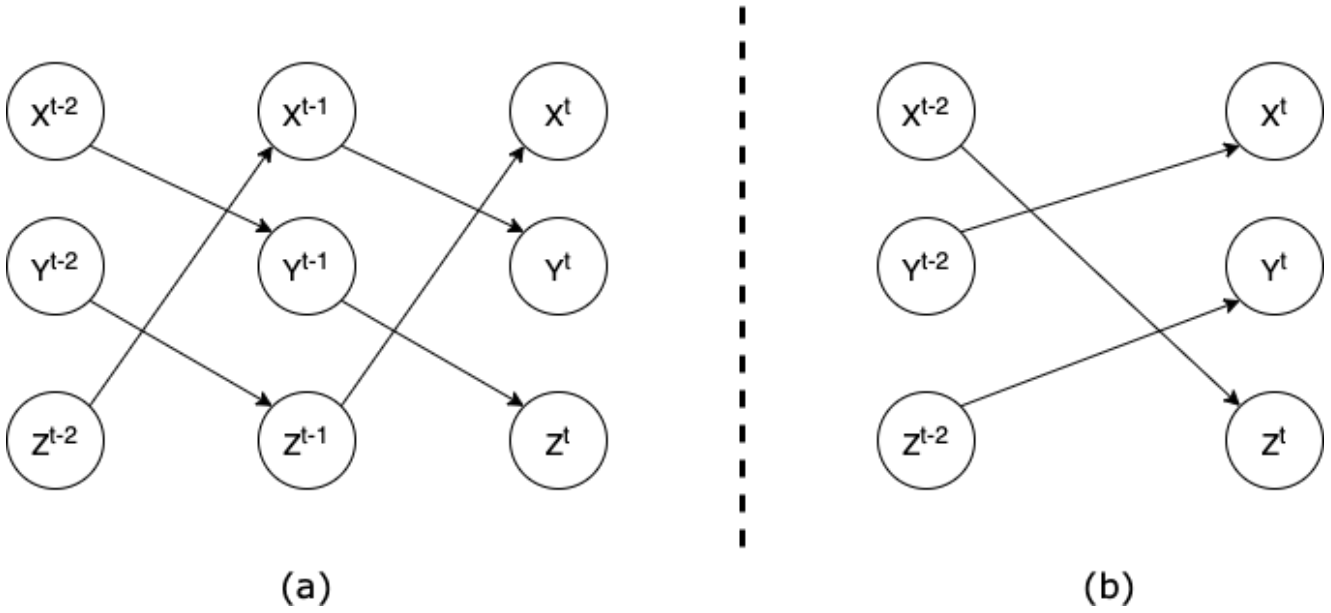
- However, the study of causal relations in time series is still a challenging issue, which is partly due to the complexity and dynamism of real world, inconsistent data, or even missing data

- One of these challenges is that causal interactions may occur on a time scale faster than the frequency of measurement (Hyttinen et al., 2017; Lawrence et al., 2020), this phenomena is known as sub-sampling.

- This can lead to a loss of valuable information to determine the true causal relationships between events.

Introduction (1/2)

In order to model dynamical systems one may use graphical models such as Directed Acyclic Graphs (DAGs), which consist of a series of nodes connected through edges or links directed from parent nodes to child nodes (Malinsky and Danks, 2018).



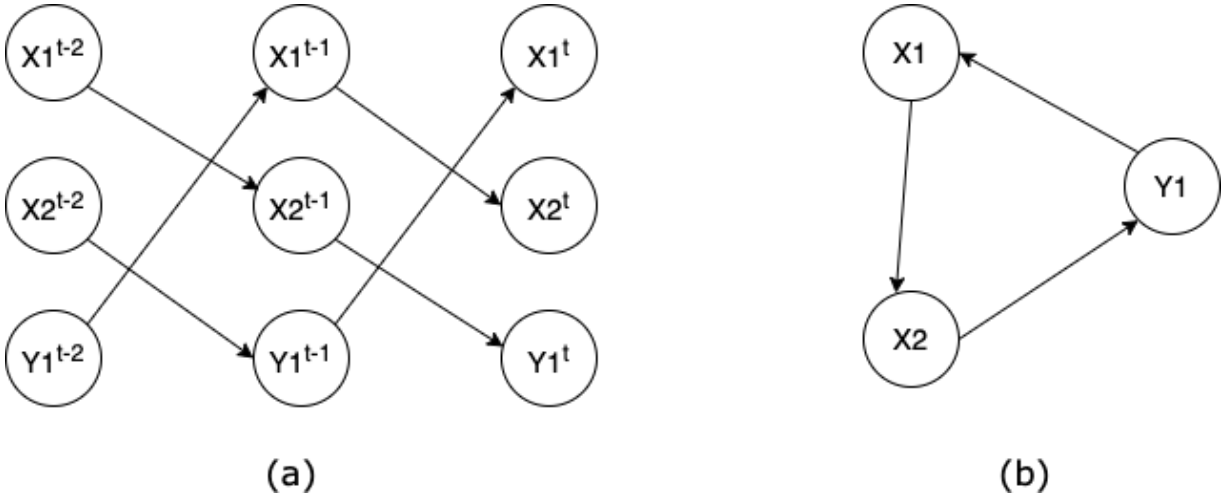
Causal structures for a time series with variables{X, Y, Z}. (a) Original structure. (b) Structure obtained from subsampled data (every two time steps).

Introduction (2/2)

The following assumptions are considered:

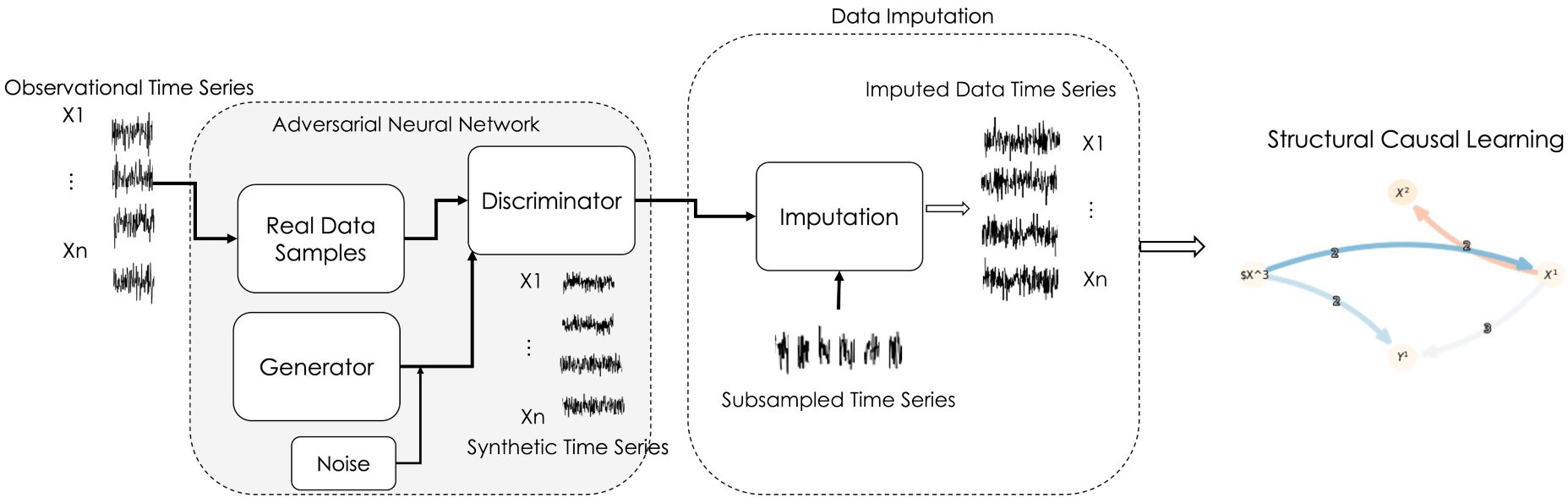
1. The causal process is invariant over time, that is, that the causal links between variables are repeated through time.

2. 2. Causal sufficiency; that is, V^{t-1} includes all common causes of V^t and there are not causal links of the form $X_i^t \rightarrow X_j^t$ (Hyttinen et al., 2017a).



(a) A time series causal structure where the causal links are repeated over time. (b) The same causal structure showed as a rolled graph.

Proposal (1/2)



The generator model generates samples until the discriminator model accepts the synthetic time series as valid.

This synthetic time series complements the missing values so that the output is a time series with imputed data that resembles the original time series. The completed data is fed to a causal structure learning algorithm to obtain the causal structure of the time series.

Proposal (2/2)

Causal Structure Learning

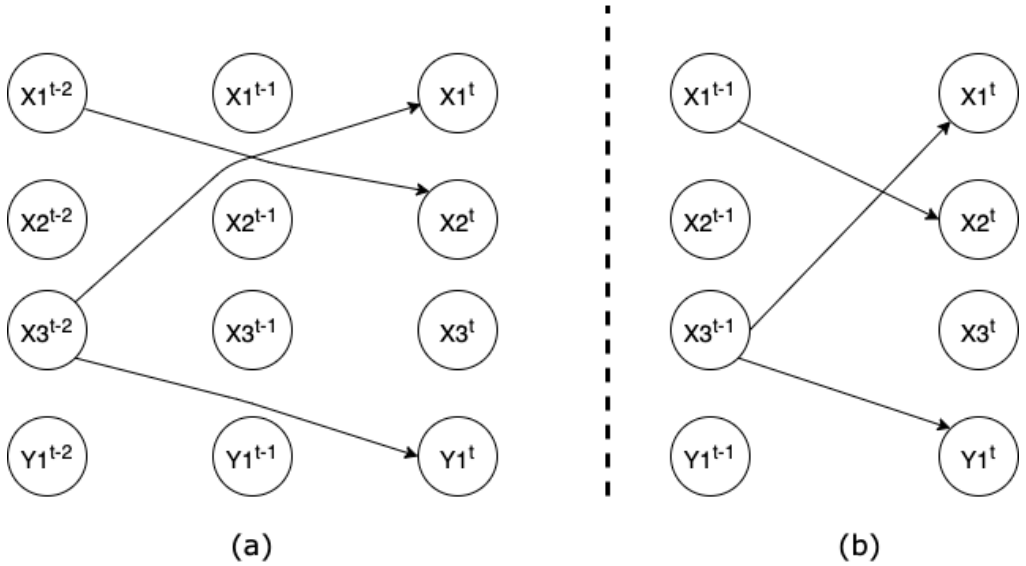
Once the data generated complements the time series, the PCMCI algorithm (Runge et al., 2019) is used to reconstruct its causal structure. This algorithm serves to analyze and compare the original causal structure and the structure resulting from the imputed data time series.

Causal Structure Verification

To verify the causal structure of the time the method proposed by Danks (2016) and Hyttinen et al. (2017b) is used. This approach takes into account a causal structure of the subsampled time series $\mathcal{H}$ is obtained, and used to generate all possible causal structures $\mathcal{G}$ that are consistent with the original causal structure.

Experimental Resutls (1/8)

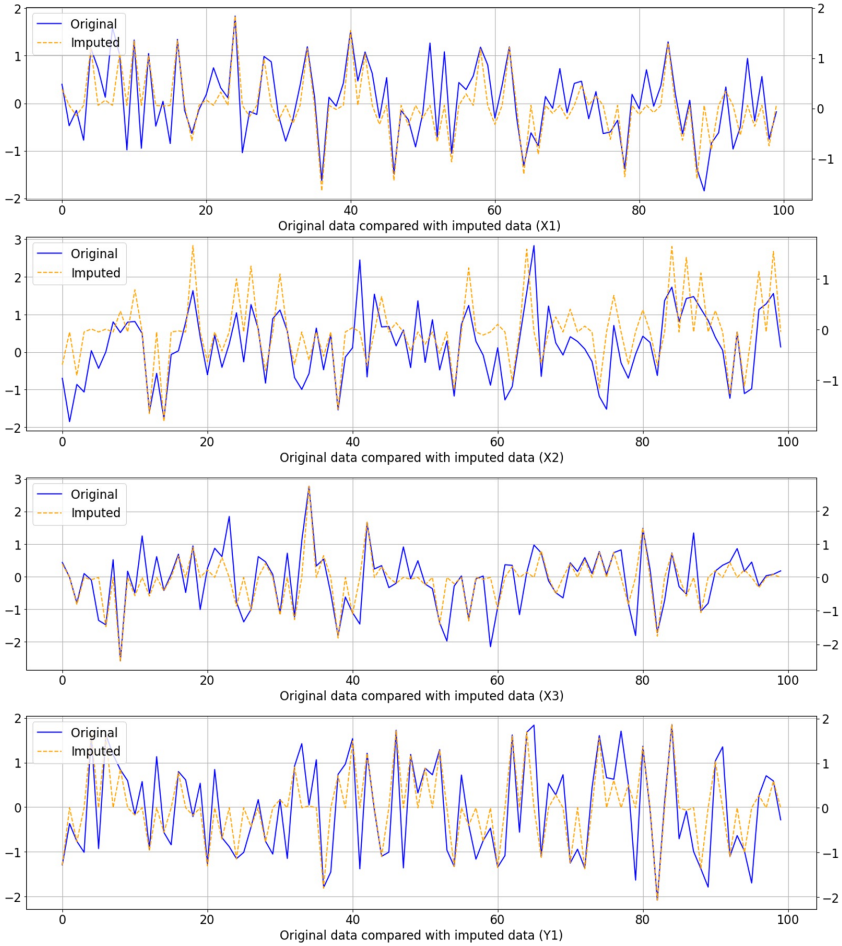
Experiment 1: Subsampled Time Series with the Same Distributions



(a) The original causal structure shows a link from X^1 to X^2, a link from X^3 to X^1 and X^3 to Y^1 at two time steps. (b) The causal structure of the subsampled time series shows a link from X^1 to X^2 at one time step, as well for the links from X^3 to X^1 and to Y^1.

Experimental Resutls (2/8)

Experiment 1: Subsampled Time Series with the Same Distributions

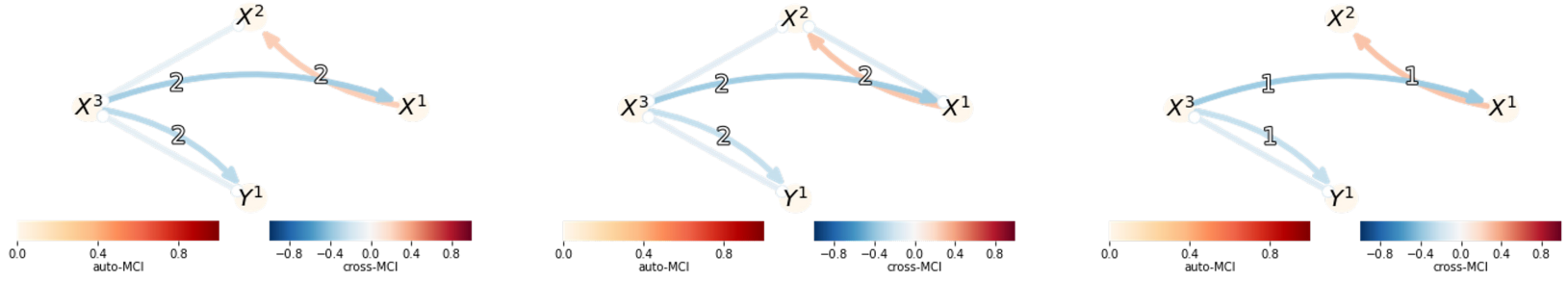


Graphs of the original (blue) and generated (orange) data for variables (from top to bottom) X^1, X^2, X^3 and Y^1.

The generated data for each one of the variables resembles the behavior of of the original values.

Experimental Resutls (3/8)

Experiment 1: Subsampled Time Series with the Same Distributions



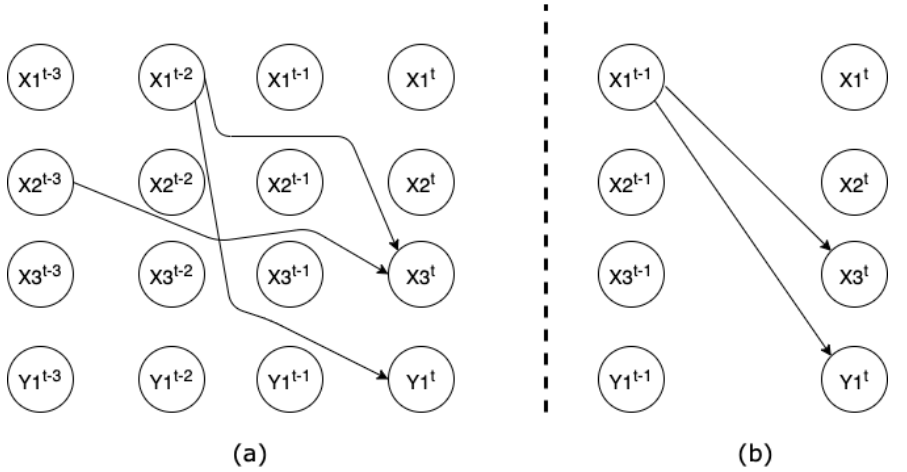
(a) (b) (c)

Representation of the causal structure for (a) the original time series, (b) the time series with imputed data and (c) the subsampled time series.

When evaluating the MAE for the causal structure of the subsampled time series the resulting value is **0.1875**, while the error value for the imputed time series is **0**.

Experimental Resutls (4/8)

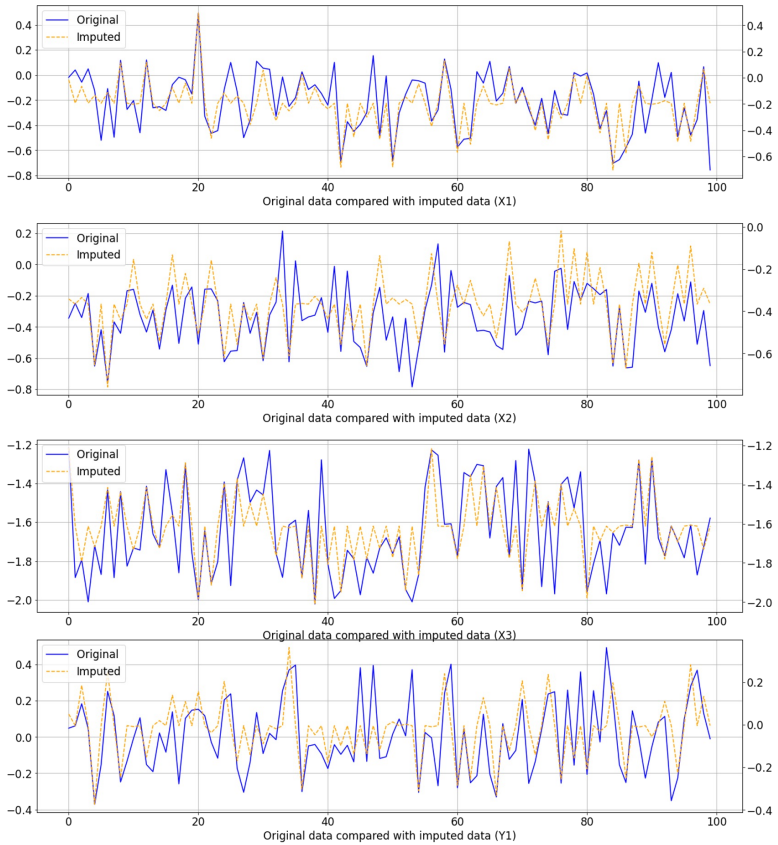
Experiment 2: Subsampled Time Series with Four Variables and Time Lag 3



(a) the original causal structure shows a link from X^1 to X^3 and Y^1 at the second time step; and in a third time step a link from X^2 to X^3. (b) The causal structure of the subsampled times series with two links from X^1 to X^3 and to Y^1 but at one time step; the link at the third time step is lost.

Experimental Resutls (5/8)

Experiment 2: Subsampled Time Series with Four Variables and Time Lag 3

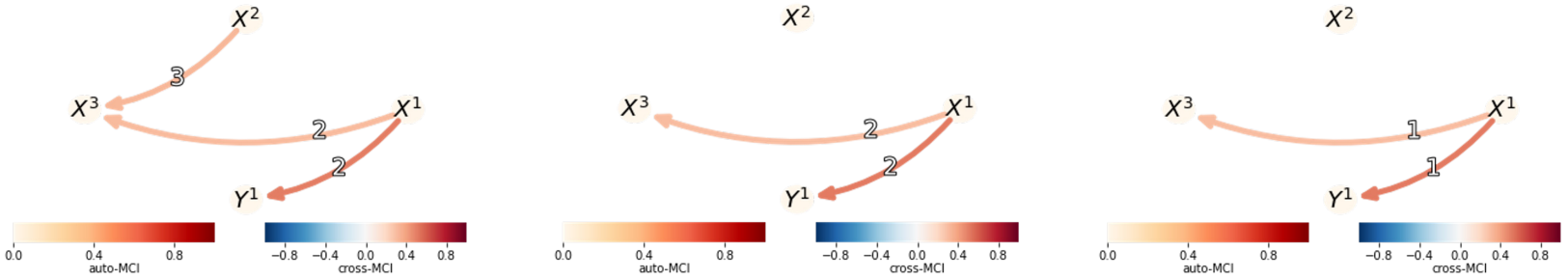


Graphs of the original (blue) and generated (orange) data for variables (from top to bottom) X^1, X^2, X^3 and Y^1.

The generated data for each one of the variables resembles the behavior of of the original values.

Experimental Resutls (6/8)

Experiment 2: Subsampled Time Series with Four Variables and Time Lag 3



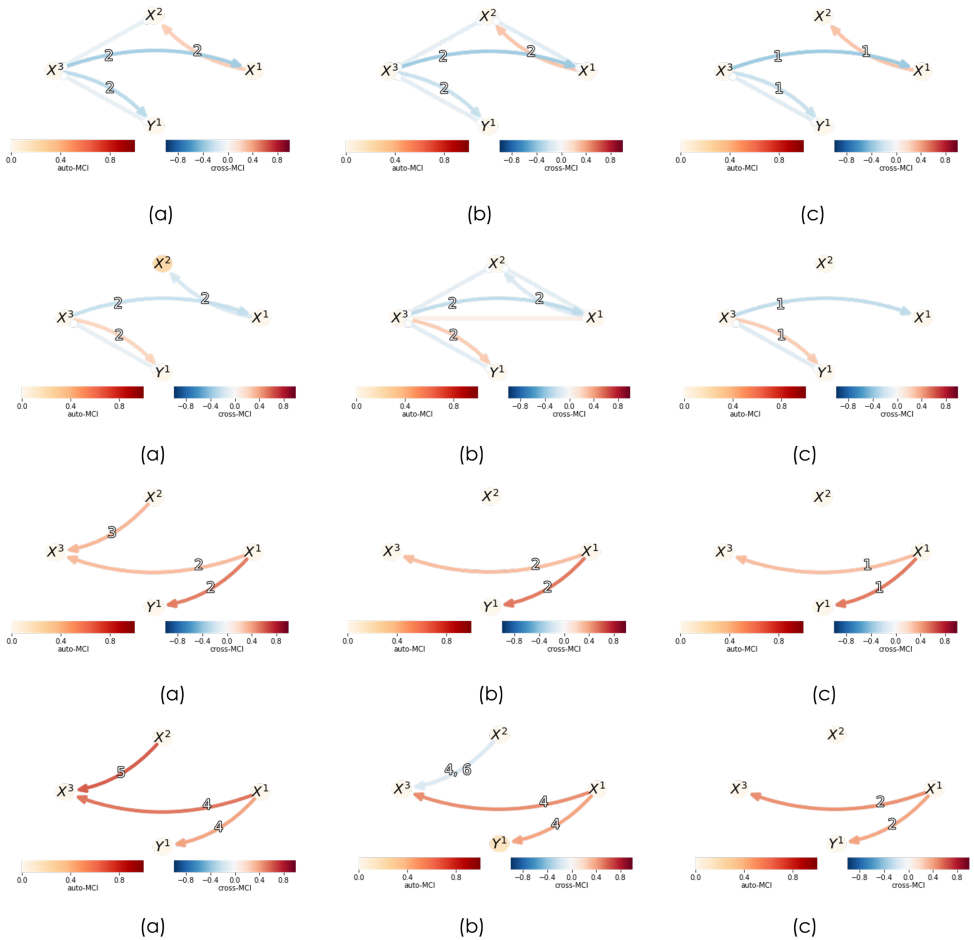
(a) (b) (c)

Representation of the causal structure for (a) the original time series, (b) the time series with imputed data and (c) the subsampled time series.

The MAE for the causal structure of the subsampled time series the resulting value is **0.1041**, while the error value for the imputed time series is **0.0208**.

Experimental Resutls (7/8)

Experiment 2-5: Subsampled Time Series with Different Distributions and Time Steps.



Experimental Resutls (8/8)

Experiment 1: Subsampled Time Series with the Same Distributions

	Number of Variables	Maximum Lag	MAE Subsampled Time series	MAE Imputed Time Series
Scenario 2	4	2	0.1875	0
Scenario 3	4	2	0.156	0
Scenario 4	4	3	0.1041	0.0208
Scenario 5	4	5	0.052	0.0312

Discussion

- The imputed data is in general very close to the original data.
- The causal structure obtained from the imputed data tends to maintain the strong causal links in the original model with the correct time scale.
- Some weak causal links may be deleted or added in the structure derived from the imputed data.

Conclusions

- Causal discovery in time series aims to analyze dynamic events that occur in the real world.
- If the data is affected by subsampling, causal discovery algorithms will generate incorrect structures.
- The imputed data time series presents a similar behavior to the original time series, so causal discovery algorithms can produce a causal structure closer to the true one.
- Experimental results considering a known subsampling rate show promising results.
- We expect that this approach can be extended to consider more complex and realistic scenarios.

Future Work

- Future work includes extending the experiments, testing data imputation in time series with a higher degree of complexity.
- Similarly, we aim to verify the recovery of the causal structure of time series affected by various rates of subsampling. An important assumption is that the subsampling rate is known.
- A possible solution is to train the ANN model for different subsampling rates (within certain range), and at the testing stage choose the most probable and consistent causal structure.

References

C. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3):424–438, 1969. ISSN 00129682, 14680262.

Antti Hyttinen, Sergey Plis, Matti J¨arvisalo, Frederick Eberhardt, and David Danks. A constraint optimization approach to causal discovery from subsampled time series data. International Journal of Approximate Reasoning, 90:208–225, 2017b. ISSN 0888-613X.

Helmut L¨utkepohl. New introduction to multiple time series analysis. Springer Science & Business Media, 2005.

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

Peter Spirtes. Introduction to causal inference. Journal of Machine Learning Research, 11(5), 2010.

Pearl, J. Causality: Models, Reasoning, and Inference. Cambridge University Press; Cambridge: 2000.

Andrew Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020a.

David Danks. Causal search, causal modeling, and the folk. A companion to experimental philosophy, pages 463–471, 2016.

David Danks and Sergey Plis. Learning causal structure from undersampled time series. JMLR: Workshop and Conference Proceedings, 2014.

Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. Philosophy Compass, 13(1):e12470, 2018.

Data Imputation with Adversarial Neural Netwroks for Causal Discovery from Subsampled Time Series

Julio Muñoz-Benítez - jcmunoz@inaoep.mx
L. Enrique Sucar - esucar@inaoep.mx

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).

CaDis

INAOE

Tonantzintla, Puebla
June, 2023