Causal Representation Learning and Causal inference in observational Time Series

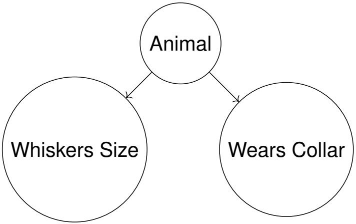
Charles K. Assaad, Emilie Devijver

emilie.devijver@univ-grenoble-alpes.fr

Table of content

Example

An image classifier observes photos of cats and dogs. The variable `Animal` is a cause of variables such as `Wears Collar` and `Has Long Whiskers`. A classifier should understand that putting a collar on a cat does not make it a dag, and making a dog's whiskers longer does not make it a cat. However, the classifier does not have direct access to these variables, besides the label `Animal`. For both robustness and interpretability, we may want the classifier to learn the variables `Wears Collar` and `Has Long Whiskers` from the pixels that it observes.



Problem statement

- **Data:** image, or audio, or **...** high-dimensional
- **Classical modeling:** ML for prediction on the dataset
- **Causal inference:** on the observed variables (pixels, blood pressure, . . .)

The mechanism behind the variables we observed might be meaningless, while there are underlying variables, latent, that may be causally related.

Problem statement

- **Data:** image, or audio, or ... high-dimensional
- **Classical modeling:** ML for prediction on the dataset
- **Causal inference:** on the observed variables (pixels, blood pressure, ...)

The mechanism behind the variables we observed might be meaningless, while there are underlying variables, latent, that may be causally related.

Goal: understand the latent structure, and the mapping between latent variables an observed variables

Problem statement

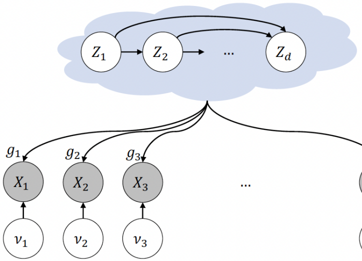
- **Data:** image, or audio, or ... high-dimensional
- **Classical modeling:** ML for prediction on the dataset
- **Causal inference:** on the observed variables (pixels, blood pressure, ...)

The mechanism behind the variables we observed might be meaningless, while there are underlying variables, latent, that may be causally related.

Goal: understand the latent structure, and the mapping between latent variables an observed variables

Robustness, interpretability, transferability!

Causal disentanglement model



Definition: $X \in \mathbb{R}^p$ follows an *(additive-noise) causal disentanglement model* if there exists $Z \in \mathbb{R}^d$ such that Z follows a structural causal model M, and $X = g(Z) + \eta$ for some mixing function g and η following a product distribution.
We call a causal disentanglement model *deterministic* if $X = g(Z)$, i.e., $\eta = 0$ almost surely.

Problem statement

Given X generated from a causal disentanglement model, we have two goals:

- **Goal 1:** Recover the causal graph over Z.
- **Goal 2:** Recover the mixing function g.

Problem statement

Given X generated from a causal disentanglement model, we have two goals:

- **Goal 1:** Recover the causal graph over Z.
- **Goal 2:** Recover the mixing function g.

In the deterministic setting, if g is invertible, then $h := g^{-1}$ disentangles a sample X_i into its causal representation $Z_i = h(X_i)$.

Linear case

Definition: We call a causal disentanglement model *linear* if

1. Z follows a linear structural causal model,
2. g is a linear function.

Linear case

Definition: We call a causal disentanglement model *linear* if

1. Z follows a linear structural causal model,
2. g is a linear function.

Example: let $\varepsilon_1, \varepsilon_2$ be independent random variables. We cannot identify between $Z_1 = \varepsilon_1, Z_2 = \varepsilon_2$ and $Z_1 = \varepsilon_1, Z_2 = A_{12}\varepsilon_1 + \varepsilon_2$ with X being a linear combination of Z_1 and Z_2: the same distribution can be get, if selecting g smartly.

Problem statement in the linear case

How to get identifiable models?

1. Restrict the form of the mixing function;
2. Restrict the form of the latent DAG;
3. Incorporate interventional data.

Restricting the form of the mixing function

Any idea?

Restricting the form of the mixing function

Any idea?

$g = id$: the latent DAG can be inferred up to the Markov equivalence class, using only observational data.

Restricting the form of the mixing function

Any idea?

$g = id$: the latent DAG can be inferred up to the Markov equivalence class, using only observational data.

disappointing ...

Restricting the form of the mixing function

Any idea?

$g = id$: the latent DAG can be inferred up to the Markov equivalence class, using only observational data.

disappointing ...

Less stringent condition: **any idea?**

Restricting the form of the mixing function

Any idea?

$g = id$: the latent DAG can be inferred up to the Markov equivalence class, using only observational data.

disappointing ...

Less stringent condition: **any idea?**

Each latent variable Z_i has a pure child (also called an anchor) - an observed variable that depends only on Z_i and no other latent variables. The pure child assumption essentially says that some submatrix of G is a scaled version of the identity.

Restricting the form of the DAG

A dual approach would restrict the latent DAG instead of the mixing function.

Any idea?

Restricting the form of the DAG

A dual approach would restrict the latent DAG instead of the mixing function.

Any idea?

For example, linear independent component analysis (ICA) assumes that the latent variables Z_i are all independent.

Restricting the form of the DAG

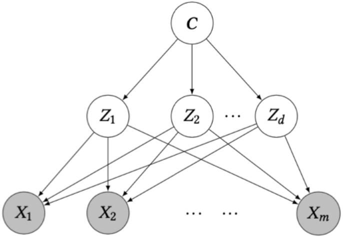
A dual approach would restrict the latent DAG instead of the mixing function.

Any idea?

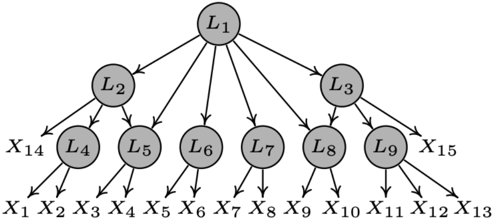
For example, linear independent component analysis (ICA) assumes that the latent variables Z_i are all independent.

Assumption: sparse factor graph joint distribution between latent factors (high-level variables) and inference involves few variables at a time.

Restricting the form of the DAG



Restricting the form of the DAG



Incorporate interventional data

- Usually, most work assumes known-interventional data
- Real world: other agents or environment can intervene: hence, interventions unknown
- How to handle unknown interventions? Infer it!

References

- Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards causal representation learning. arXiv preprint arXiv:2102.11107.
- Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, Kun Zhang Proceedings of the 39th International Conference on Machine Learning, PMLR 162:24370-24387, 2022.
- Ahuja, K., Wang, Y., Mahajan, D., and Bengio, Y. (2022). Interventional causal representation learning. arXiv preprint arXiv:2209.11924.
- Seigal, A., Squires, C., and Uhler, C. (2022). Linear causal disentanglement via interventions. arXiv preprint arXiv:2211.16467.
- Varici, B., Acarturk, E., Shanmugam, K., Kumar, A., and Tajer, A. (2023). Score-based causal representation learning with interventions. arXiv preprint

Table of content

Causal graphs for time series

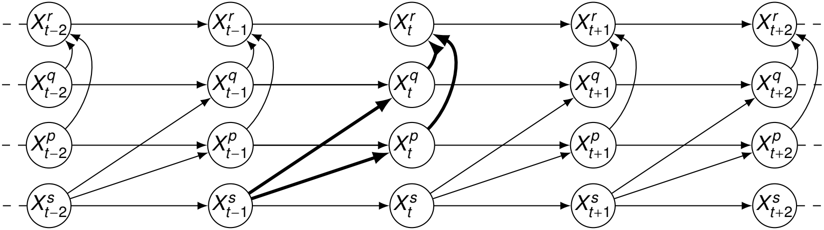


Figure: Full time causal graph.

A d-variate time series X
For a fixed t, each X_t is a vector $(X^1_t, \dots, X^d_t)$,
in which X^p_t is the measurement of the pth time series at time t.

Causal graphs for time series

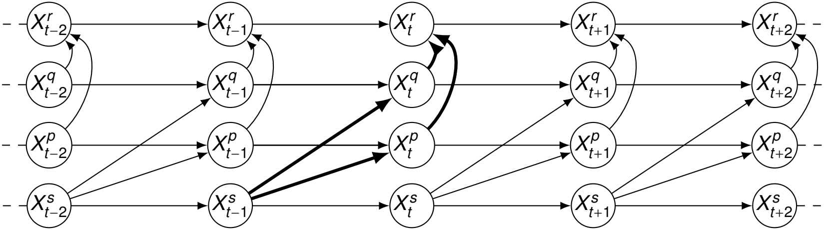
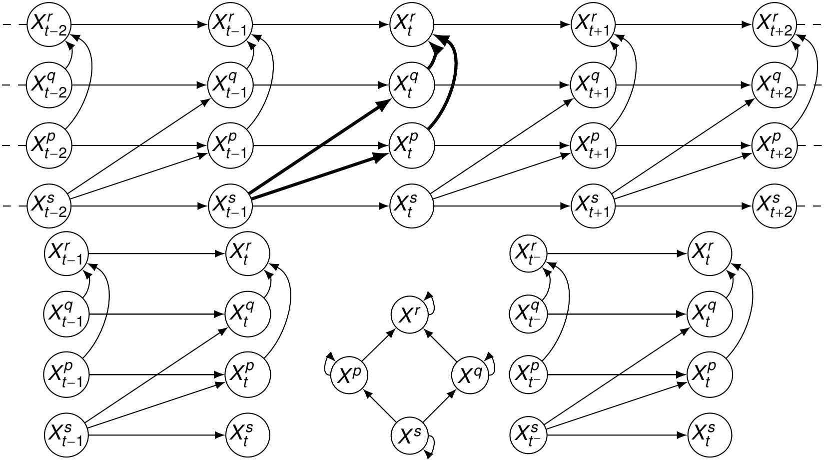


Figure: Full time causal graph.

A d-variate time series X
For a fixed t, each X_t is a vector $(X^1_t, \ldots, X^d_t)$,
in which X^p_t 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.

Causal graphs for time series



Causal discovery

Granger Causality[1]

[1] Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–38.

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

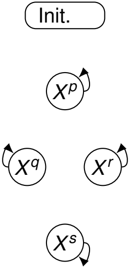


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

[1] Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–38.

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

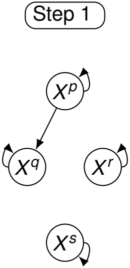


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

[1] Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–38.

Causal discovery

Granger Causality[1]

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

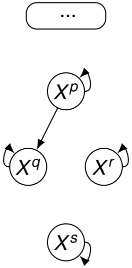


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

[1] Granger, C. (1969). Investigating causal relations by econometric models

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

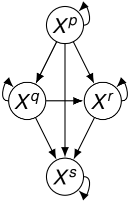


Figure: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

[1] Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–38.

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

Main difficulty when dealing with time series: determine a good measure of (conditional) dependencies

- Measure : (partial) correlation, entropy, mutual information ...
- Estimation
- Type of data (continuous, discrete, mixed)

Representation of the time series

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

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

where h denotes the entropy.

Causal discovery

Constraint-based approaches: PCMCI[2] to discover a window causal graph

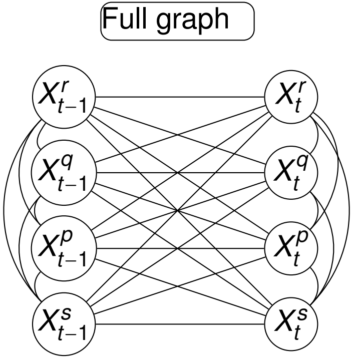


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

[2] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11).

Causal discovery

Constraint-based approaches: PCMCI[2] to discover a window causal graph

Independence

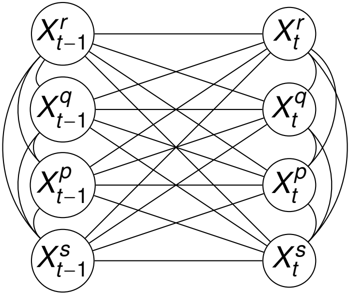


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

[2] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11).

Causal discovery

Constraint-based approaches: PCMCI[2] to discover a window causal graph

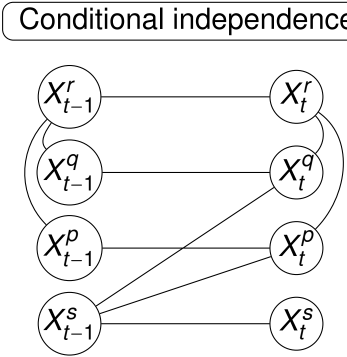


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

[2] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11).

Causal discovery

Constraint-based approaches: PCMCI[2] to discover a window causal graph

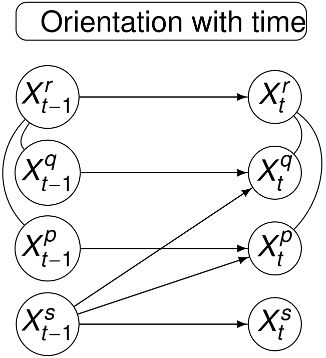


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

---

[2] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11).

Causal discovery

Constraint-based approaches: PCMCI[2] to discover a window causal graph

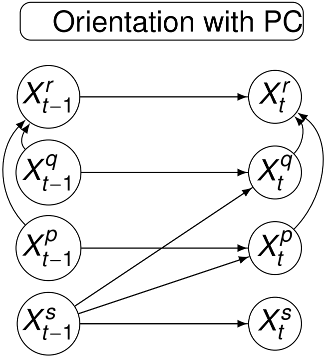


Figure: Running example: structure inferred by PCMCI with instantaneous relations.

[2] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11).

Causal discovery

Constraint-based approaches: PCGCE[3] to discover an extended summary causal graph

Assumptions

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

[3] C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022

Causal discovery

Constraint-based approaches: PCGCE[3] to discover an extended summary causal graph

Assumptions

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

Measure: need a specific one due to the graph structure
Proposed Greedy Causation Entropy (GCE)

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

[3] C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022

Causal discovery

Constraint-based approaches: PCGCE[4] to discover an extended summary causal graph

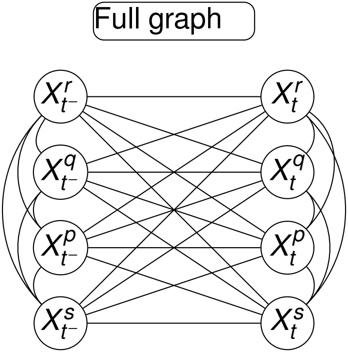


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

[4] C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022

Causal discovery

Constraint-based approaches: PCGCE[4] to discover an extended summary causal graph

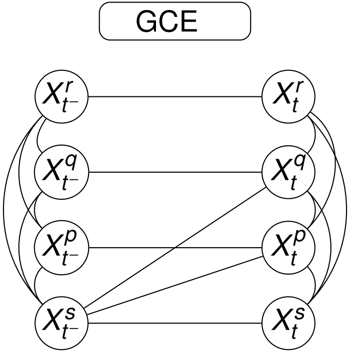


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

[4] C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022

Causal discovery

Constraint-based approaches: PCGCE[4] to discover an extended summary causal graph

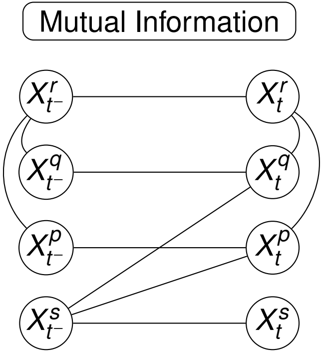


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

[4] C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022

Causal discovery

Constraint-based approaches: PCGCE[4] to discover an extended summary causal graph

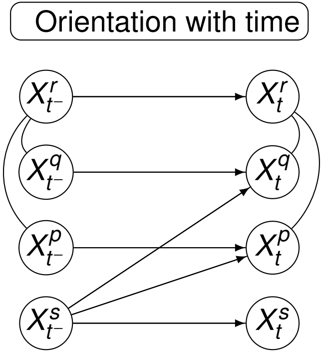


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

[4] C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022

Causal discovery

Constraint-based approaches: PCGCE[4] to discover an extended summary causal graph

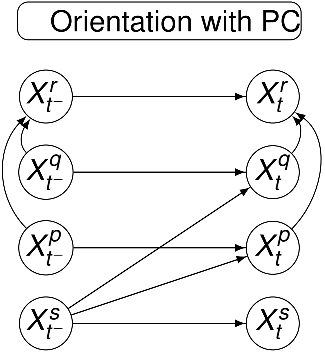


Figure: Running example: structure inferred by PCGCE with instantaneous relations.

[4] C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 2022

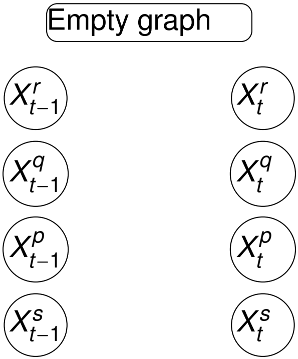


Figure: Running example: structured inferred by VarLiNGAM.

[5] Hyvarinen, A., Zhang, K., Shimizu, S., Hoyer, P. O. *Estimation of a structural vector autoregression model using non-gaussianity*. JMLR 2010

Causal discovery

Noise-based approaches: VarLINGAM[5]

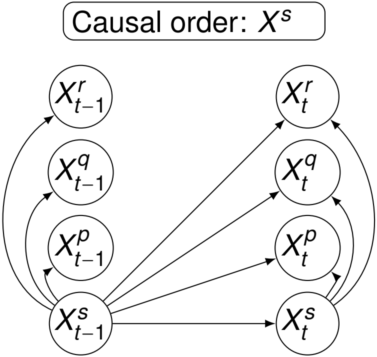


Figure: Running example: structured inferred by VarLiNGAM.

[5] Hyvarinen, A., Zhang, K., Shimizu, S., Hoyer, P. O. *Estimation of a structural vector autoregression model using non-gaussianity*. JMLR 2010

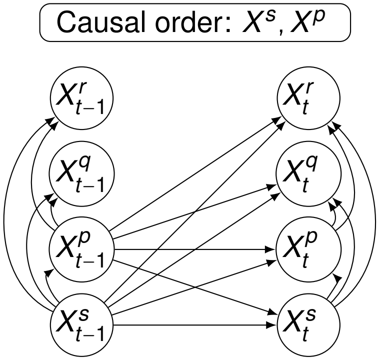


Figure: Running example: structured inferred by VarLiNGAM.

[5]Hyvarinen, A., Zhang, K., Shimizu, S., Hoyer, P. O. *Estimation of a structural vector autoregression model using non-gaussianity*. JMLR 2010

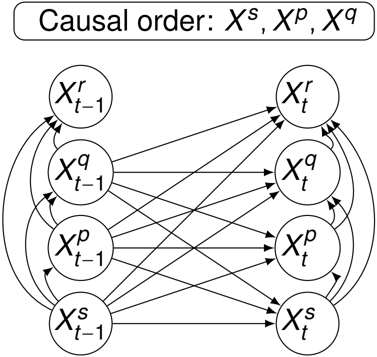


Figure: Running example: structured inferred by VarLiNGAM.

[5]Hyvarinen, A., Zhang, K., Shimizu, S., Hoyer, P. O. *Estimation of a structural vector autoregression model using non-gaussianity*. JMLR 2010

Causal discovery

Noise-based approaches: VarLINGAM[5]

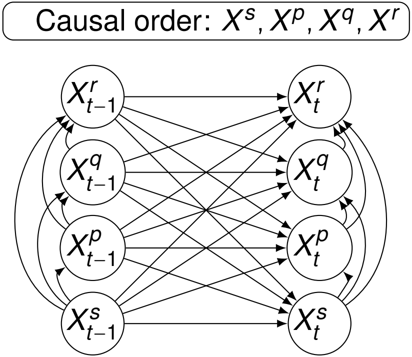


Figure: Running example: structured inferred by VarLiNGAM.

[5]Hyvarinen, A., Zhang, K., Shimizu, S., Hoyer, P. O. *Estimation of a structural vector autoregression model using non-gaussianity*. JMLR 2010

Causal discovery

Noise-based approaches: VarLINGAM[5]

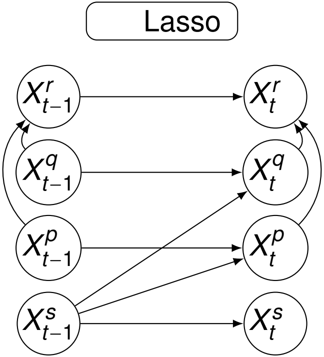


Figure: Running example: structured inferred by VarLiNGAM.

[5]Hyvarinen, A., Zhang, K., Shimizu, S., Hoyer, P. O. *Estimation of a structural vector autoregression model using non-gaussianity*. JMLR 2010

Causal discovery

NBCB[6]: a mix between noise-based and constraint-based approaches

[6] C. K. Assaad, E. Devijver, and E. Gaussier. *A Mixed Noise and Constraint Based Approach to Causal Inference in Time Series*, ECMLPKDD 2021

Causal discovery

NBCB[6]: 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

[6] C. K. Assaad, E. Devijver, and E. Gaussier. *A Mixed Noise and Constraint Based Approach to Causal Inference in Time Series*, ECMLPKDD 2021

NBCB[6]: 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

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 based on (conditional) independence measure.

[6] C. K. Assaad, E. Devijver, and E. Gaussier. *A Mixed Noise and Constraint Based Approach to Causal Inference in Time Series*, ECMLPKDD 2021

Causal discovery

NBCB: a mix between noise-based and constraint-based approaches

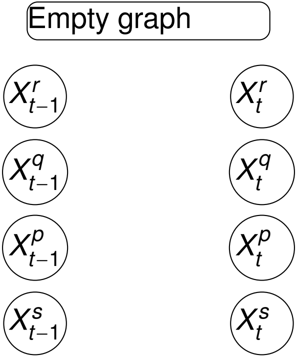


Figure: Running example: structured inferred by NBCB.

Causal discovery

NBCB: a mix between noise-based and constraint-based approaches

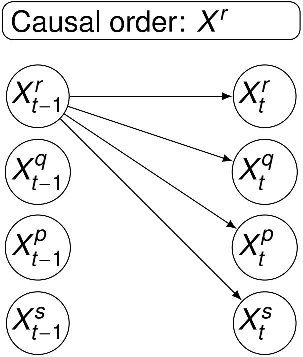


Figure: Running example: structured inferred by NBCB.

Causal discovery

NBCB: a mix between noise-based and constraint-based approaches

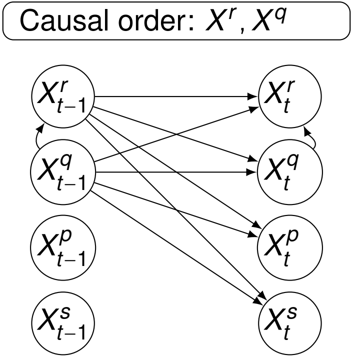


Figure: Running example: structured inferred by NBCB.

Causal discovery

NBCB: a mix between noise-based and constraint-based approaches

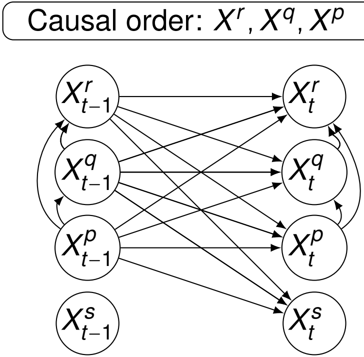


Figure: Running example: structured inferred by NBCB.

Causal discovery

NBCB: a mix between noise-based and constraint-based approaches

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

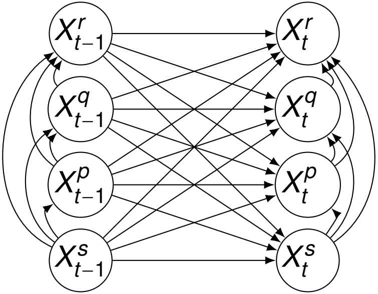


Figure: Running example: structured inferred by NBCB.

Causal discovery

NBCB: a mix between noise-based and constraint-based approaches

Conditional independence using TCE

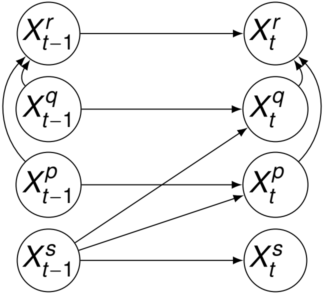


Figure: Running example: structured inferred by NBCB.

Causal reasoning

Identifiability in FTCG and ECG

[7] Blondel et al. 2016, Shpitser et al. 2008

Assumptions: causal sufficiency, consistency throughout time.

Theorem 1[7]: Consider an FTCG $\mathcal{G}^f$ (or equivalently a WCG). The total effect $P(y_t|do(x_{t-\gamma}))$, with $\gamma \geq 0$ is identifiable in $\mathcal{G}^f$.

Theorem 2: Consider an ECG $\mathcal{G}^e$. The total effect $P(y_t|do(x_{t-\gamma}))$, with $\gamma \geq 0$ is identifiable in $\mathcal{G}^e$.

[7] Blondel et al. 2016, Shpitser et al. 2008

Causal reasoning

Identifiability in SCG[8]

Assumptions: causal sufficiency, consistency throughout time.

Theorem 3: Consider an SCG $\mathcal{G}^s = (\mathcal{V}^s, \mathcal{E}^s)$. The total effect $P(y_t|do(x_{t-\gamma}))$, with $\gamma \geq 0$, is not identifiable if and only if $X \in Anc(Y, \mathcal{G}^s)$ and one of the following holds:

- $\gamma \neq 0$ and $Cycles^{>}(X, \mathcal{G}^s \backslash \{Y\}) \neq \emptyset$, or
- there exists a σ-active back-door path $\pi^s = \langle V^1 = X, \cdots, V^n = Y\rangle$ such that $\langle V^2, \cdots, V^{n-1}\rangle \subseteq Desc(X, \mathcal{G}^s)$ and one of the following holds:
  - $n > 2$, ie $\langle V^2, \cdots, V^{n-1}\rangle \neq \emptyset$, or
  - $\gamma \neq 1$, or
  - $\gamma = 1$, $n = 2$ and $Cycles(Y, \mathcal{G}^s \backslash \{X\}) \neq \emptyset$.

[8] A. Meynaoui et al., *Identifiability of total effects from abstractions of time series causal graphs*, submitted

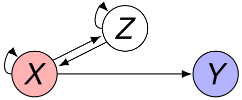


Figure: An SCG $\mathcal{G}_1^s$ and the total effect $P(y_t|do(x_{t-1}))$.

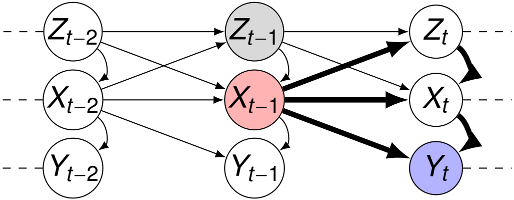


Figure: An FTCG compatible with the SCG $\mathcal{G}_1^s$.

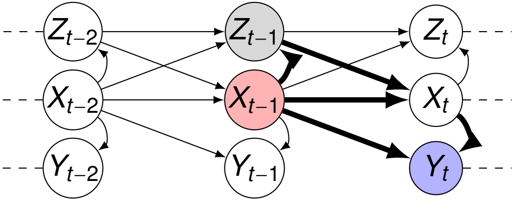


Figure: Another FTCG compatible with the SCG $\mathcal{G}_1^s$.

Causal reasoning

Identifiability in SCG: non identifiable example 2

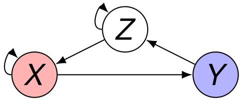


Figure: An SCG $\mathcal{G}_2^s$ and the total effect $P(y_t|do(x_{t-1}))$.

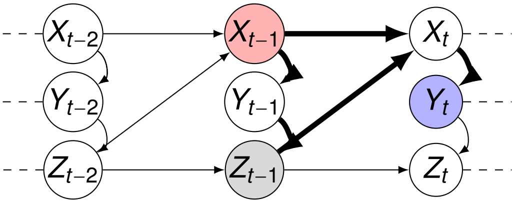


Figure: An FTCG compatible with the SCG $\mathcal{G}_2^s$.

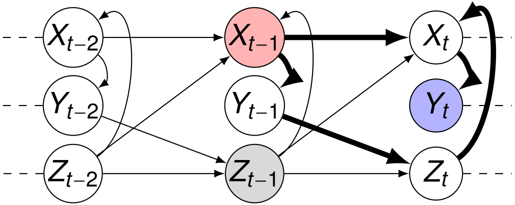


Figure: Another FTCG compatible with the SCG $\mathcal{G}_2^s$.

Causal reasoning

Identifiability in SCG: identifiable examples

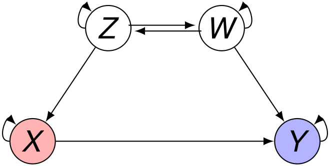


Figure: $P(y_t|do(x_{t-\gamma}))$

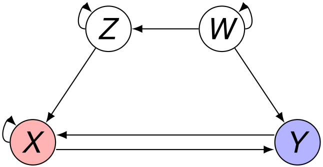


Figure: $P(y_t|do(x_{t-1}))$

Conclusion and perspectives

- Which causal graph do we want to infer?
- The representation of time series is essential (windows - lags)
- Many families to discover causal graph for time series (also score-based, logic-based, topology-based, difference-based)
- Hybrid methods can take benefit of several worlds

References (Causal discovery)

- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.
- C. K. Assaad, E. Devijver, and E. Gaussier. *A Mixed Noise and Constraint Based Approach to Causal Inference in Time Series*, ECMLPKDD 2021
- C. K. Assaad, E. Devijver, and E. Gaussier. *Survey and evaluation of causal discovery methods for time series*. JAIR, 73, 2022.
- C. K. Assaad, E. Devijver, and E. Gaussier. *Causal Discovery of Extended Summary Graphs in Time Series*, UAI 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 2022

References (Causal reasoning)

- G. Blondel, M. Arias, R. Gavalda, *Identifiability and transportability in dynamic causal networks*, International Journal of Data Science and Analytics, 2016
- Perkovic, E., *Identifying causal effects in maximally oriented partially directed acyclic graphs*, UAI 2020
- A. Meynaoui, C. K. Assaad, E. Devijver, E. Gaussier, G. Gössler, *Identifiability of total effects from abstractions of time series causal graphs*, submitted