Causal Inference by Identification of Vector Autoregressive Processes with Hidden Components

Philipp Geiger[a], Kun Zhang[a,b], Mingming Gong[c], Dominik Janzing[a] and Bernhard Schölkopf[a]

[a]Empirical Inference Department, Max Planck Institute for Intelligent Systems, Tübingen, Germany
[b]Information Sciences Institute, University of Southern California, USA
[c]Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia

7th July 2015

Why observational causal inference? An example

Causal question

Does extensive food commodity **speculation cause high food prices**? [Gilbert 2010]

Why this is relevant

If the answer is "yes", it may make sense to **modify the policy** on speculation, given **lower food prices are a goal**.

A first causal answer

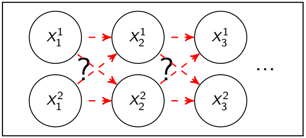
Cheap **non-experimental** economic time series may yield a **first hypothesis**!

(Or at least a priority ordering on the causal search space.)

Problem formulation and previous approaches

Given

A **finite multivariate time series** $X_{1:L} = X_1, \ldots, X_L$ of measurements of a system.



Goal

Infer the **causal structure** of the system.

I.e., **correctly predict the outcome of targeted manipulations** of the system's mechanisms that generate $X_{1:L}$.

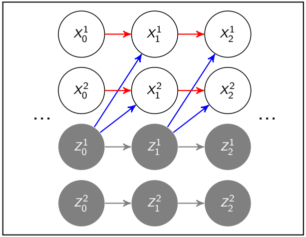
($\rightarrow$ Deeper understanding of the system.)

Previous approaches (selection)

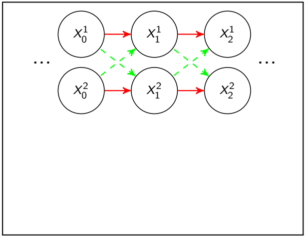
- Regress X_t on X_{t-1}, interpret r. matrix causally [Granger1969]
- Assume linear latent process and sparsity [Jalali2012]

Underlying idea

- If we had **measured** the evolution of the **whole universe**, then we might infer the causal structure underlying $X_{1:L}$ from this "observational data" (using Granger/PC algorithm). [Granger1969, Pearl2000, Spirtes2000]
- Otherwise: problem of potential **hidden confounders**.



A confounder Z^1 in the hidden part of the universe ...



... may introduce spurious association between X^1 and X^2.

Our approach consists of two parts

Part 1: Examination of identifiability

We present **certain model restrictions** that make the crucial parameters of **partially observed processes identifiable**.

(“Identifiable” = mapping from parameter to $P(X)$ is injective.)

Part 2: Two tailored estimators

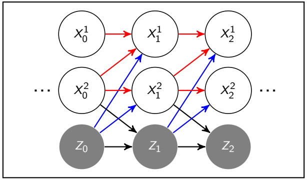
We present two algorithms that work on simulated data under the respective assumptions from part 1.

(Conjecture: (“approximate”) consistency, due to identifiability.)

General model assumptions

Statistical model for $X_{1:L}$

$X = (X_t)_{t\in\mathbb{Z}}$ together with an **unmeasured multivariate** Z with $\dim(Z_t) \leq \dim(X_t)$ is a stable **vector autoregressive (VAR) process**,



$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = \underbrace{\begin{pmatrix} B & C \\ D & E \end{pmatrix}}_{=:A} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + \underbrace{N_t}_{\text{i.i.d.}}.$$

Causal claim

The matrix B can be interpreted causally, i.e., it correctly predicts the outcome of randomized empirical experiments on X.

Theorem 1: conditions for identifiability of B, C

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = \underbrace{\begin{pmatrix} B & C \\ D & E \end{pmatrix}}_{A} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + N_t.$$

If:

1. All noise terms N_t^k are **non-Gaussian**.
2. $N_t^1, N_t^2, \ldots$ are **jointly independent** for all t.
3. Certain **generic** full rank assumptions w.r.t. the transition matrix A and the autocovariance hold.

Then: Given only $P(X)$,

1. the matrix B is **uniquely identifiable**,
2. the columns of C with at least two non-zero entries are identifiable up to scaling and permutation of those columns.

Theorem 2: weaker but including the Gaussian case

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = \underbrace{\begin{pmatrix} B & C \\ D & E \end{pmatrix}}_{A} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + N_t.$$

If:

1. $D = 0$ (i.e. no influence from Z to X).
2. Certain **generic** assumptions regarding A and the noise covariance matrix hold.
   ("Generic" = only excludes Lebesgue null set of parameters.)

Then:

Given only the autocovariance of X, B is identifiable up to $\binom{2\dim(X_t)}{\dim(X_t)}$ possibilities.

($\binom{\cdot}{\cdot}$ = binomial coefficient.)

Two estimation algorithms (sketch)

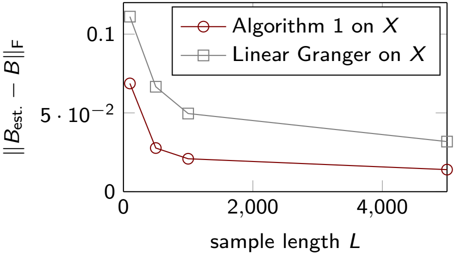
Algorithm 1: based on variational EM

- Assuming a parametric model: the complete VAR model with **mixture of Gaussians (MoG) as noise**.
- Maximizing a **variational lower bound** $\mathcal{L}$ to the marginal log-likelihood based on **mean field assumption** (a factorized approximation).

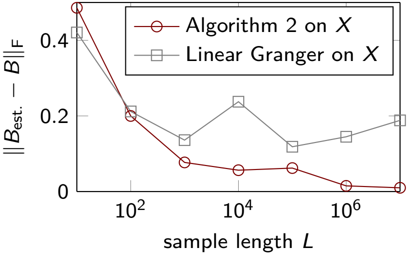
Algorithm 2: using only the autocovariance

- Idea (from Theorem 2): B fulfills a certain equation w.r.t. the autocovariance of X.
- Replace it by empirical autocovariance and calculate solutions.

Evaluation of Algorithms 1 and 2 on simulated data



2-variate observed X,
1-variate hidden Z.



1-variate observed X,
1-variate hidden Z.

Evaluation of Algorithm 1 on real data

Experimental setup

- $W = \begin{pmatrix} X^1 \\ X^2 \\ Z \end{pmatrix} = \begin{pmatrix} \text{cheese price} \\ \text{butter price} \\ \text{milk price} \end{pmatrix}$
- Assumed **ground truth**: linear Granger applied to complete W

Outcome

- Error ($\|\cdot\|_{\mathrm{F}}$) of linear Granger applied to only X: **0.0662**.
- Error of Algorithm 1 applied to X: **0.0753** - slightly worse.

Remark

The full W **does not fulfill our model assumptions**, e.g. the estimated lag is 3. A certain **model check on** X detects this!

Conclusion and references

Future directions

- Finding "linear enough" application domains.
- Identifiability in non-linear latent processes.

Take home message

From partially hidden linear time series,
important "causal" properties often can still be identified.

References

- Paper underlying this talk: Philipp Geiger, Kun Zhang, Mingming Gong, Dominik Janzing and Bernhard Schölkopf. Causal Inference by Identification of Vector Autoregressive Processes with Hidden Components. ICML, 2015.
- [Gilbert 2010]: C. L. Gilbert. How to Understand High Food Prices. Journal of Agricultural Economics, 2010.
- [Granger1969]: C. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 1969.
- [Jalali2012]: A. Jalali and S. Sanghavi. Learning the dependence graph of time series with latent factors. ICML, 2012.
- [Pearl2000]: J. Pearl. Causality. Cambridge University Press, 2000.
- [Spirtes2000]: P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction, and search. MIT, Cambridge, MA, 2nd edition, 2000.

Some additional remarks

Model checking

- Parts of our model assumptions can be checked from only X.
- E.g. using certain independence tests and tests for Gaussianity.

Proof ideas

- Theorem 1: overcomplete ICA.
- Theorem 2: B solves a matrix polynomial defined from the autocovariance of X.