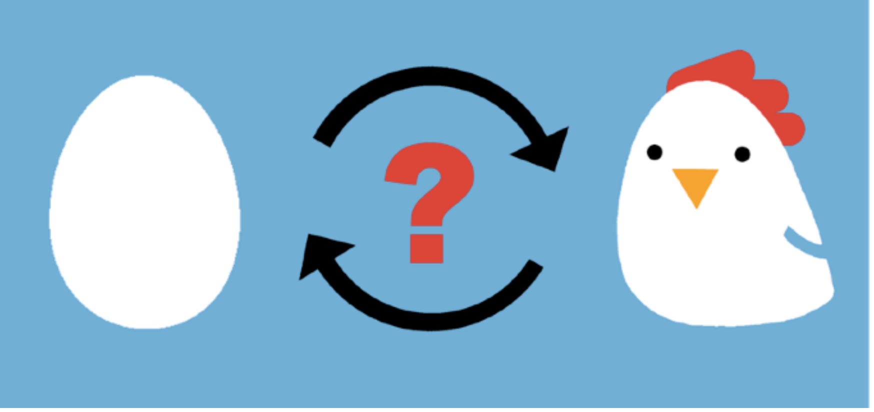


Causality and Machine Learning
(80-816/516)

Classes 13 & 14 (Feb 24 & 26, 2025)

Linear, Non-Gaussian Case: Dealing with Confounders and Cycles

Instructor:

Kun Zhang (kunz1@cmu.edu)

Zoom link: https://cmu.zoom.us/j/8214572323)

Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

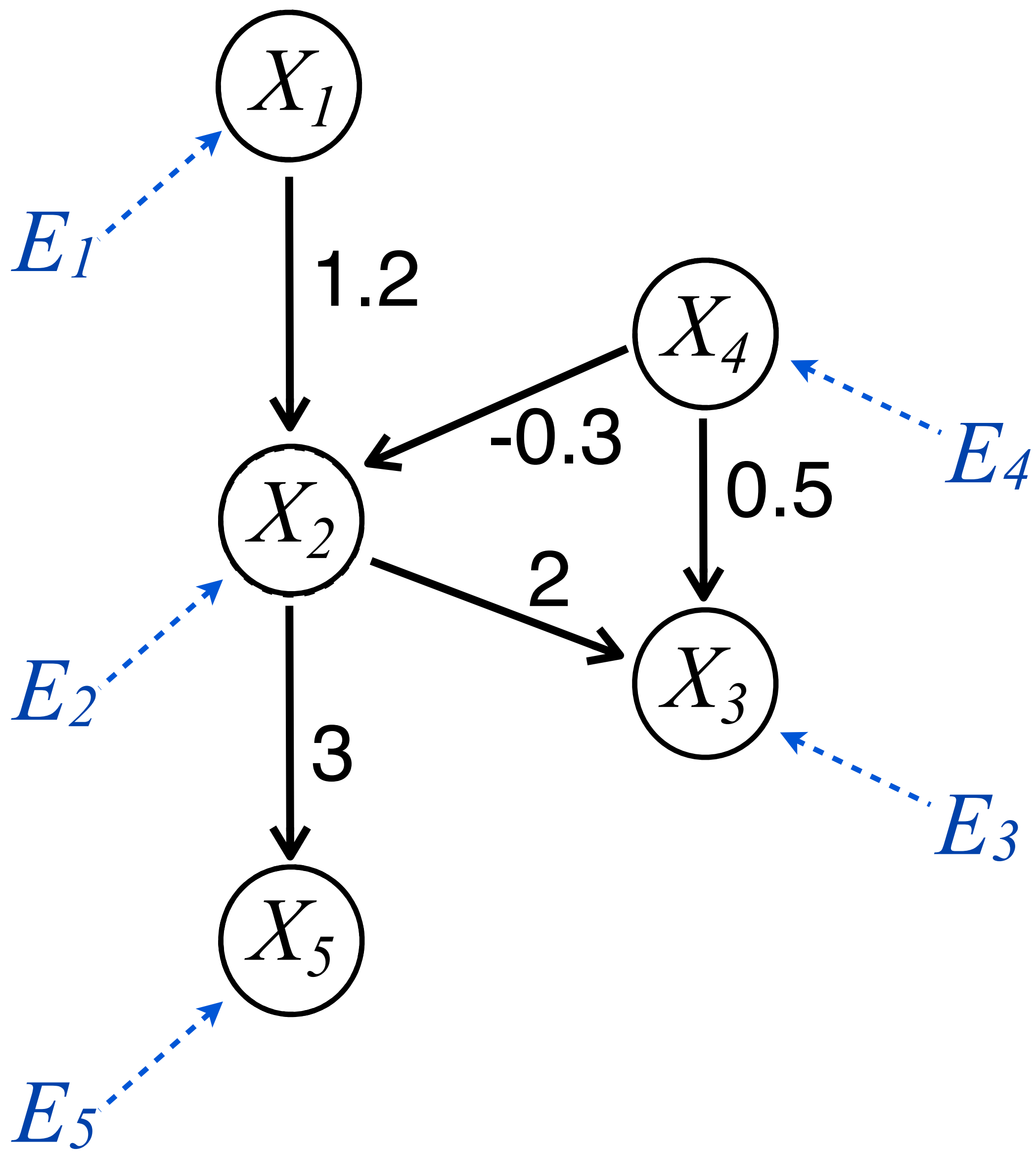
Practical Issues in Causal Discovery...

- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19; Xie et al., NeurIPS'20); causal representation learning (Xie et al., NeurIPS'20; Cai et al., NeurIPS'19...)
- Cycles (Richardson 1996; Lacerda et al., 2008)
- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Spirtes 1995; Zhang et al., UAI'16)
- Missing values (Tu et al., AISTATS'19)
- Causality in time series
    - Time-delayed + instantaneous relations (Hyvarinen ICML'08; Zhang et al., ECML'09; Hyvarinen et al., JMLR'10)
    - Subsampling / temporally aggregation (Danks & Plis, NIPS WS'14; Gong et al., ICML'15 & UAI'17)
    - From partially observable time series (Geiger et al., ICML'15)
- Nonstationary/heterogeneous data (Zhang et al., IJCAI'17; Huang et al, ICDM'17, Ghassami et al., NIPS'18; Huang et al., ICML'19 & NIPS'19; Huang et al., JMLR'20)

With Confounders

- Confounders cause trouble in causal discovery
- Assuming independent confounders:
    - Possible solutions I: Overcomplete ICA for Linear-Non-Gaussian case
- **Assuming causally related confounders!**
    - Possible solutions II: GIN for Linear-Non-Gaussian case
    - Possible solution II: Rank deficiency for Linear-Gaussian case

Are They Confounders ?



X_1?

X_4?

X_2?

X_5?

FCI Allows Confounders

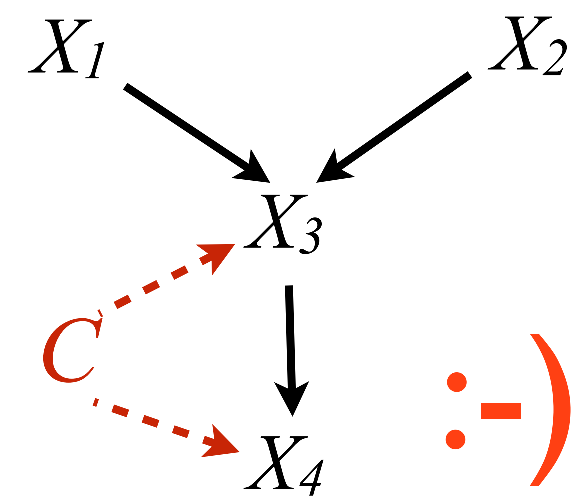
Example I

$X_1 \perp\!\!\!\perp X_2;$

$X_1 \perp\!\!\!\perp X_4 \mid X_3;$

$X_2 \perp\!\!\!\perp X_4 \mid X_3.$

Possible to have confounders behind X_3 and X_4?



E.g., X_1: Raining; X_3: wet ground; X_4: slippery.

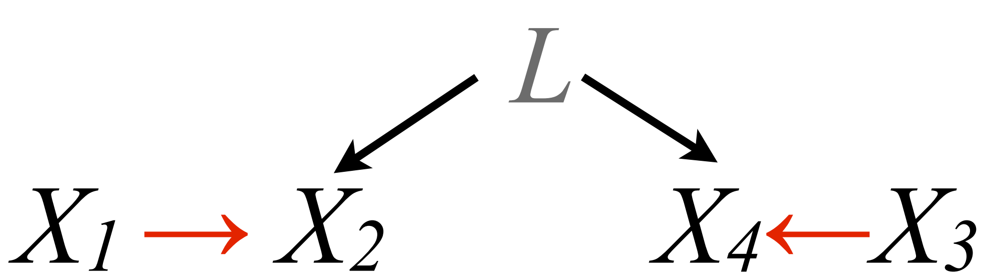
Example II

$X_1 \perp\!\!\!\perp X_3;$

$X_1 \perp\!\!\!\perp X_4;$

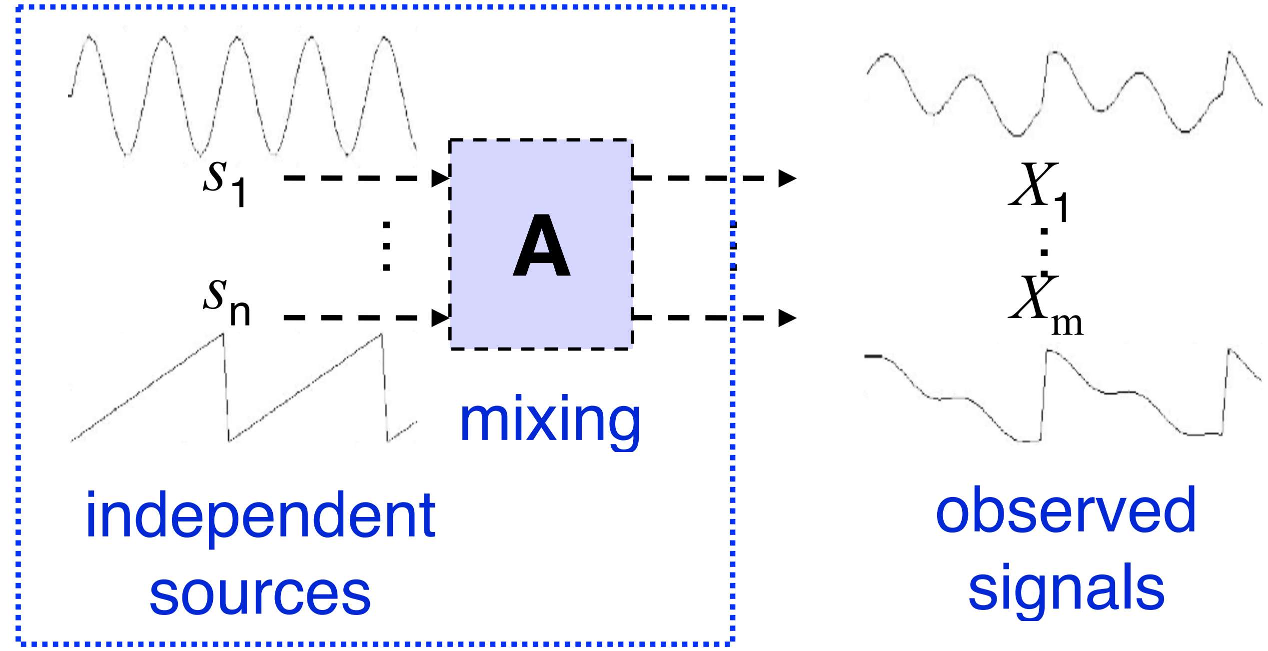
$X_2 \perp\!\!\!\perp X_3.$

Are there confounders behind X_2 and X_4?



E.g., X_1: I am not sick; X_2: I am in this lecture room; X_4: you are in this lecture room; X_3: you are not sick.

Identifiability of Overcomplete ICA



unknown mixing system

$n > m$

- More independent sources than observed variables, i.e., n>m

$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? & ? \\ ? & ? & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\ s_2 \\ s_3 \end{array}$$

A

Theorem: Suppose the random vector $X = (X_1, ..., X_m)^\intercal$ is generated by $X = \mathbf{A}S$, where the components of S, $S_1, ..., S_n$, are statistically independent. Even when $n > m$, the columns of $\mathbf{A}$ are still identifiable up to a scale transformation if

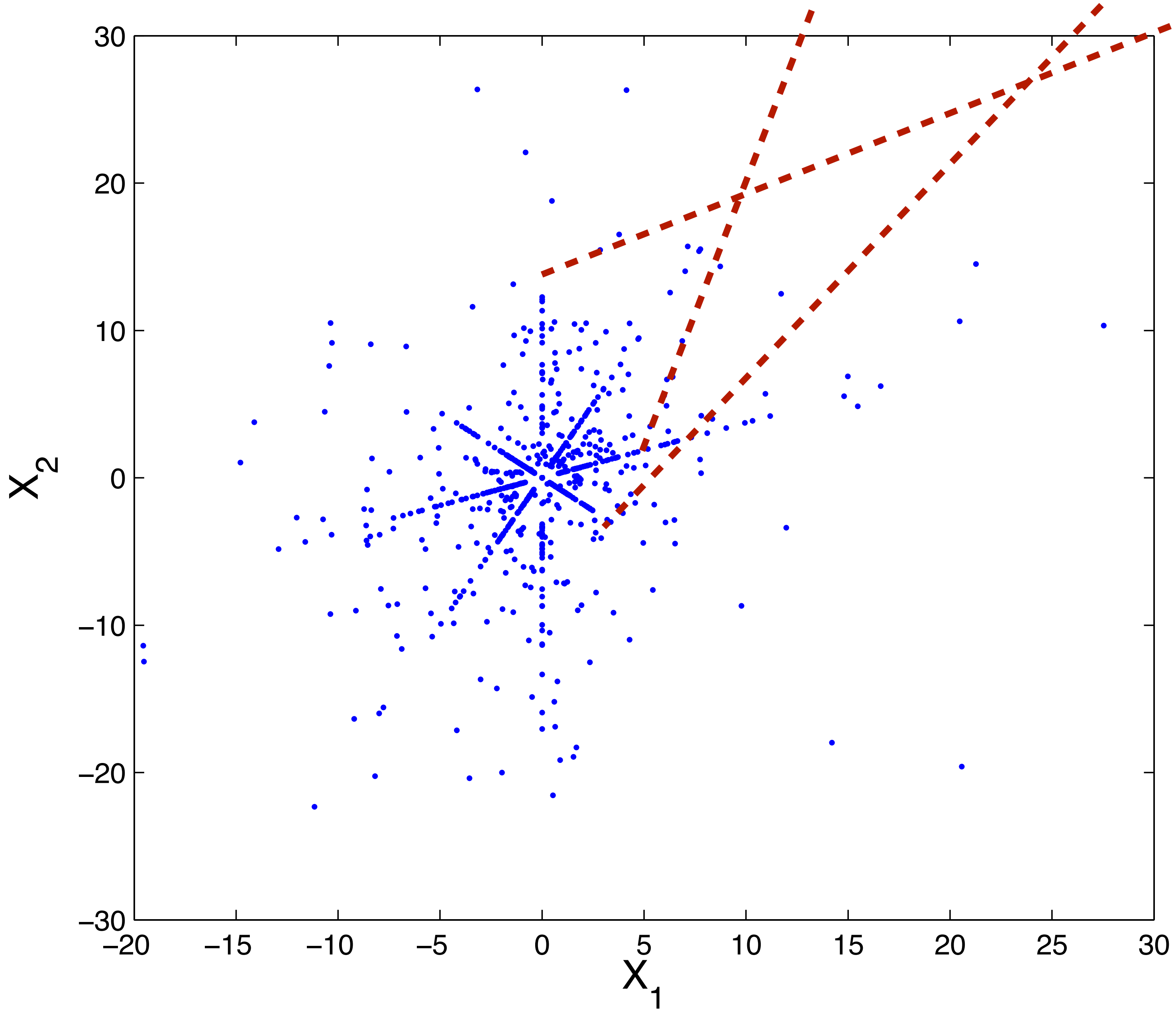
- all S_i are non-Gaussian, or
- $\mathbf{A}$ is of full column rank and at most one of S_i is Gaussian.

Kagan et al., Characterization Problems in Mathematical Statistics. New York:Wiley, 1973

Eriksson and Koivunen (2004). Identifiability, Separability and Uiiiqueness of Linear ICA Models, IEEE Signal Processing Lett.: vol. 11, no. 7, pp. GOI-604, Jul. 2004.

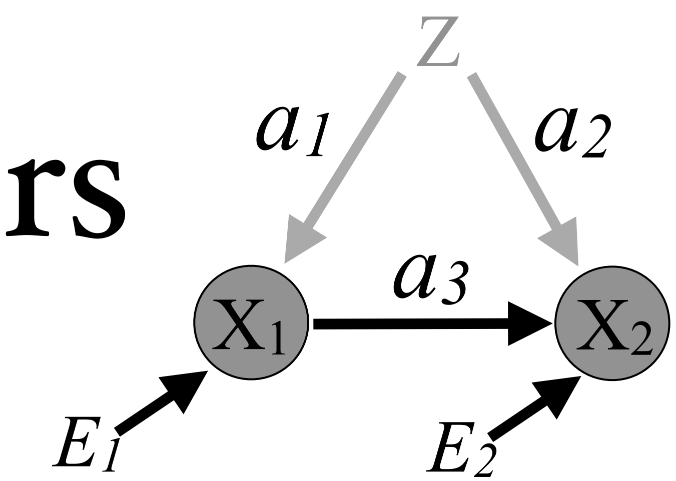
Overcomplete ICA: Illustration

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.4 & -0.9 & 0 \\ 0.3 & 0.8 & 0.8 & 1 \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}$$



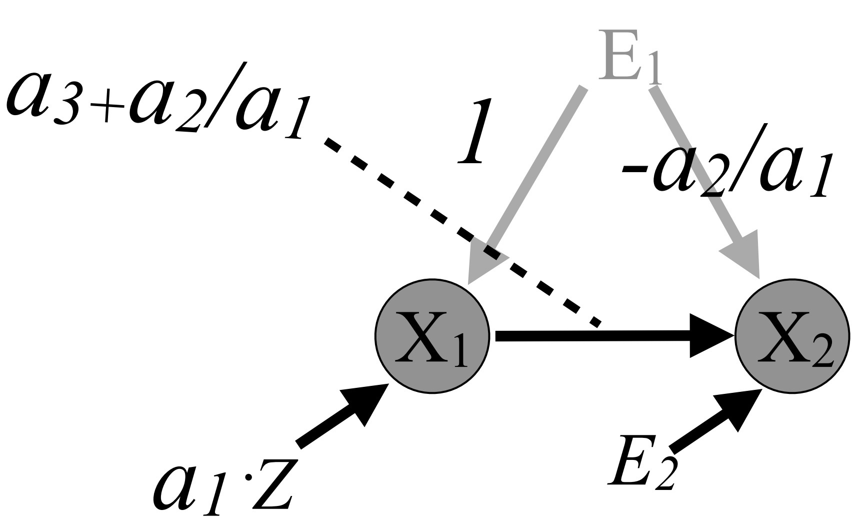
What if they are Gaussian?

Discussions I: Confounders



$$\begin{bmatrix}X_1\\X_2\end{bmatrix} = \begin{bmatrix}1 & 0 & a_1\\ a_3 & 1 & a_1a_3+a_2\end{bmatrix}\cdot\begin{bmatrix}E_1\\E_2\\Z\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ a_3 & 1 & a_3+\frac{a_2}{a_1}\end{bmatrix}\cdot\begin{bmatrix}E_1\\E_2\\a_1Z\end{bmatrix}$$

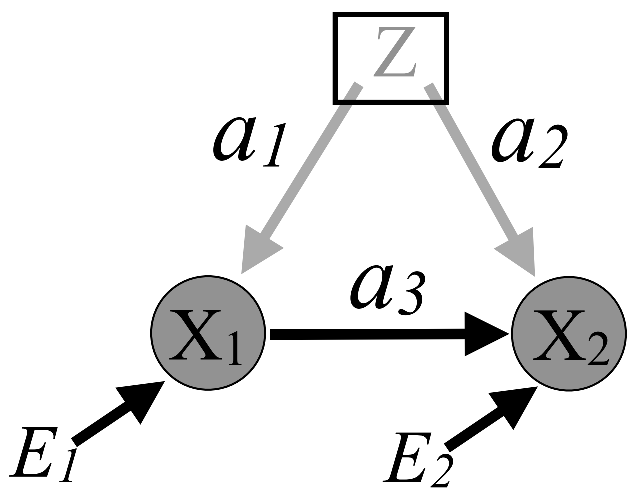
- Can we see the causal direction ?
- Can we determine a_3 ? a_1 and a_2 ?
- Observationally equivalent model:



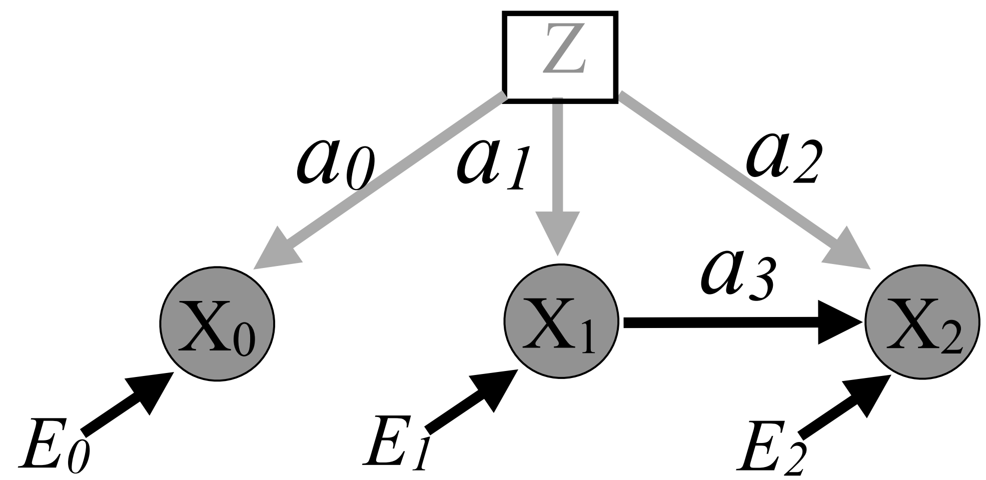
$$\begin{bmatrix}X_1\\X_2\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ \left(a_3+\frac{a_2}{a_1}\right)+\frac{-a_2}{a_1} & 1 & \left(a_3+\frac{a_2}{a_1}\right)\end{bmatrix}\cdot\begin{bmatrix}E_1\\E_2\\a_1Z\end{bmatrix}$$

Hoyer et al. (2008). Estimation of causal effects using linear nonGaussian causal models with hidden variables. IJAR,.
Salehkaleybar, Ghassami, Kiyavash, Zhang (2020), Learning Linear Non-Gaussian Causal Models in the Presence of Latent Variables, JMLR

Two Examples: Causal Effect Identifiable?



Example 1



Example 2

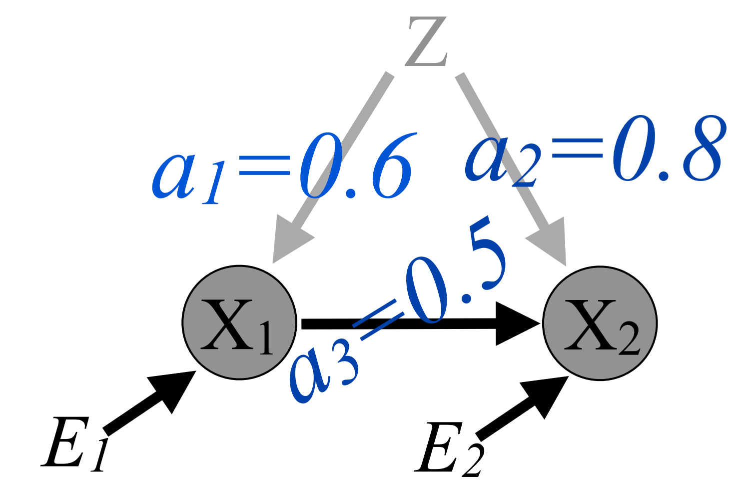
Example 1: $$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & a_1 \\ a_3 & 1 & a_1 a_3 + a_2 \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ Z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ a_3 & 1 & a_3 + \frac{a_2}{a_1} \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ a_1 Z \end{bmatrix}$$

Two possible solutions

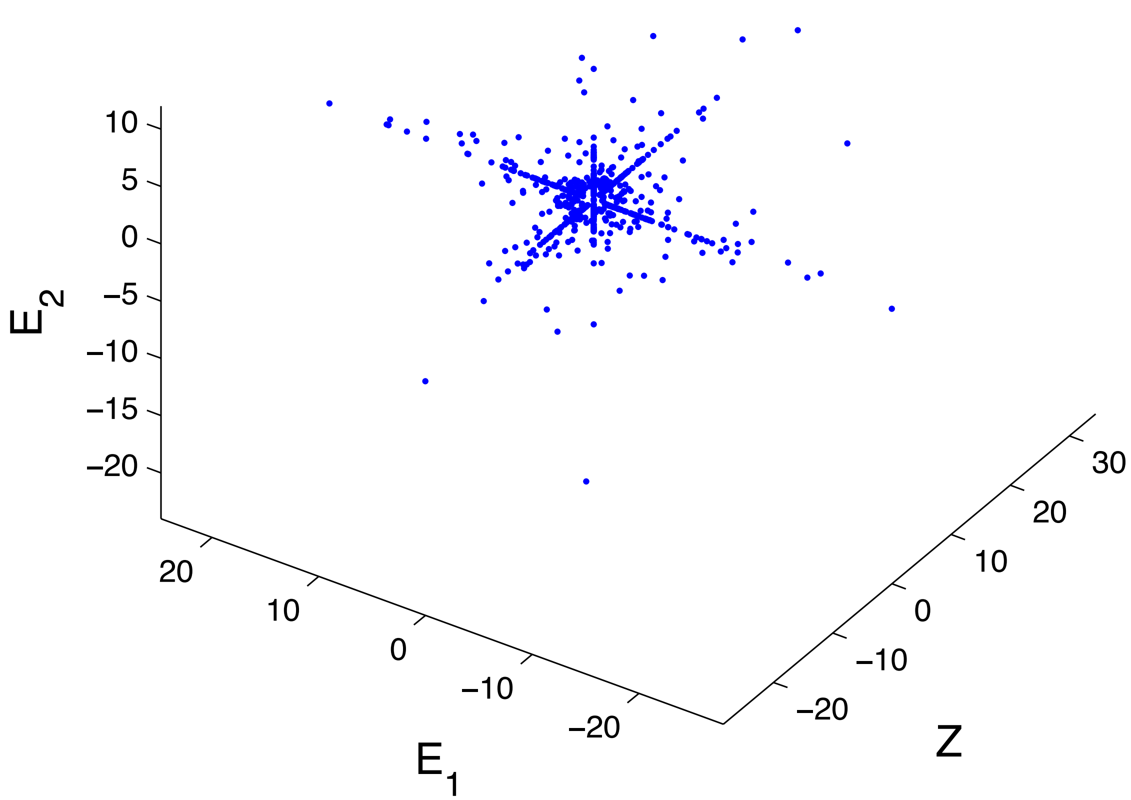
Example 2: $$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & a_0 \\ 0 & 1 & 0 & a_1 \\ 0 & a_3 & 1 & a_1 a_3 + a_2 \end{bmatrix} \cdot \begin{bmatrix} E_0 \\ E_1 \\ E_2 \\ Z \end{bmatrix}$$

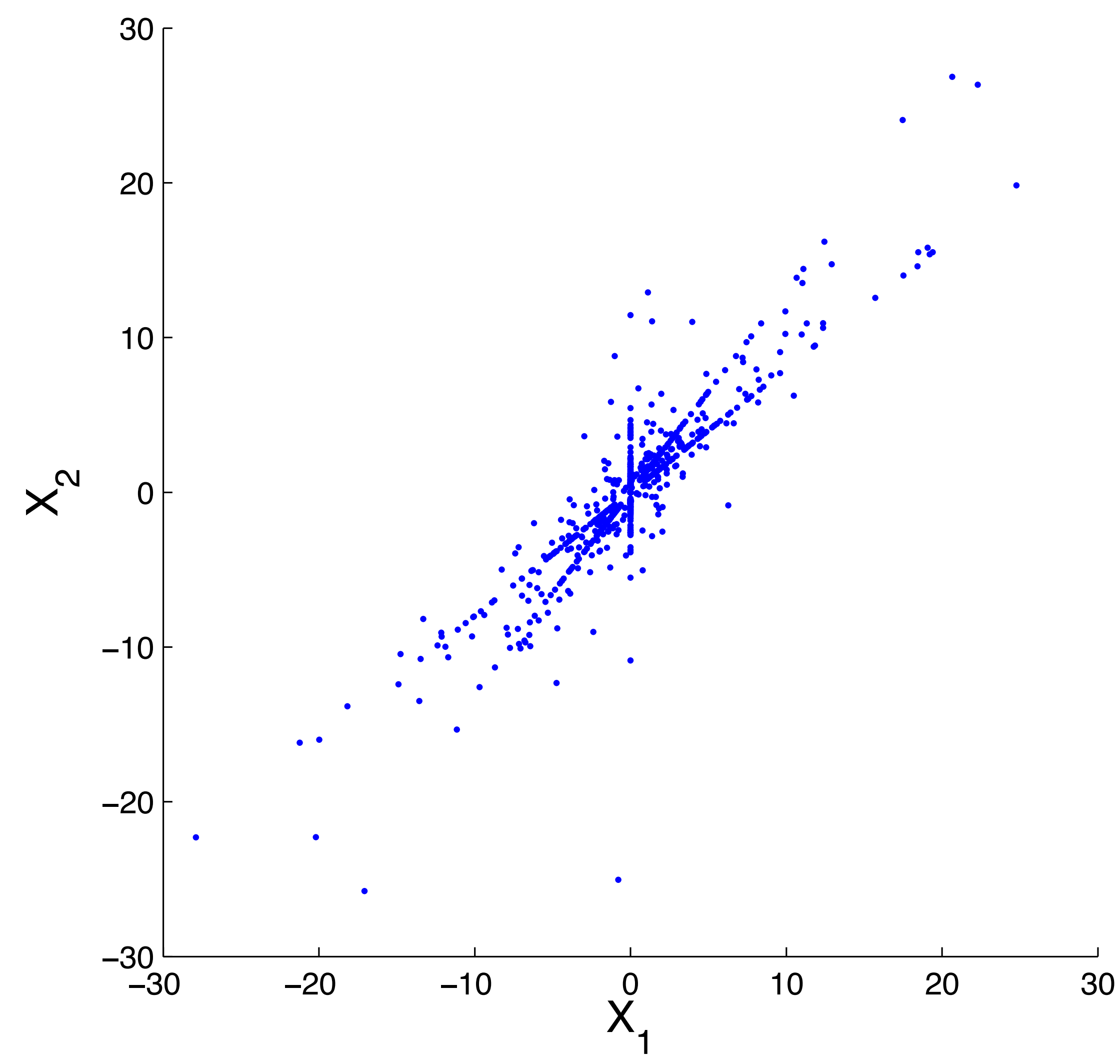
a_3 identifiable!

Confounders: Example



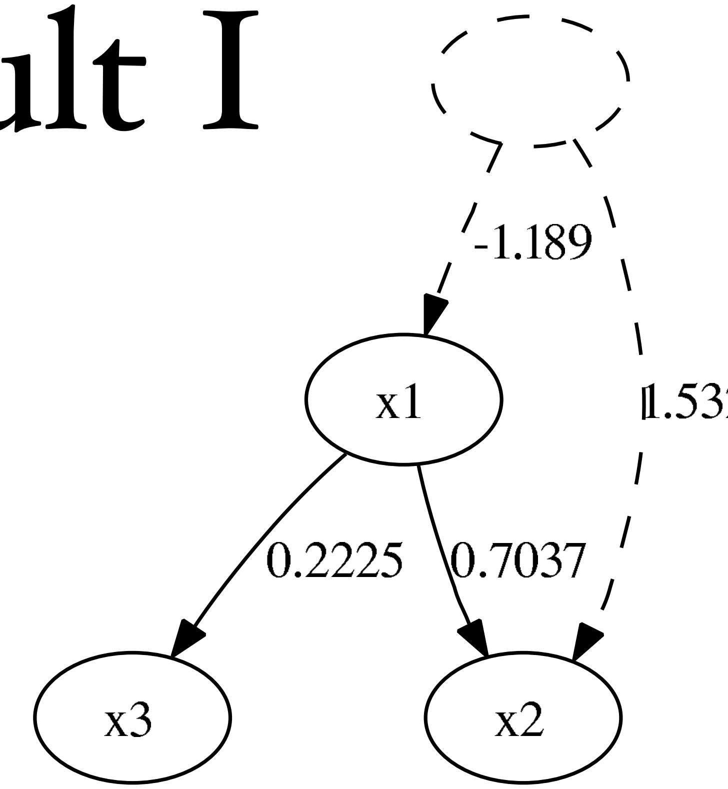
$$\begin{bmatrix}X_1\\X_2\end{bmatrix} = \begin{bmatrix}1 & 0 & a_1\\ a_3 & 1 & a_1a_3+a_2\end{bmatrix}\cdot\begin{bmatrix}E_1\\E_2\\Z\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ a_3 & 1 & a_3+\frac{a_2}{a_1}\end{bmatrix}\cdot\begin{bmatrix}E_1\\E_2\\a_1Z\end{bmatrix}$$



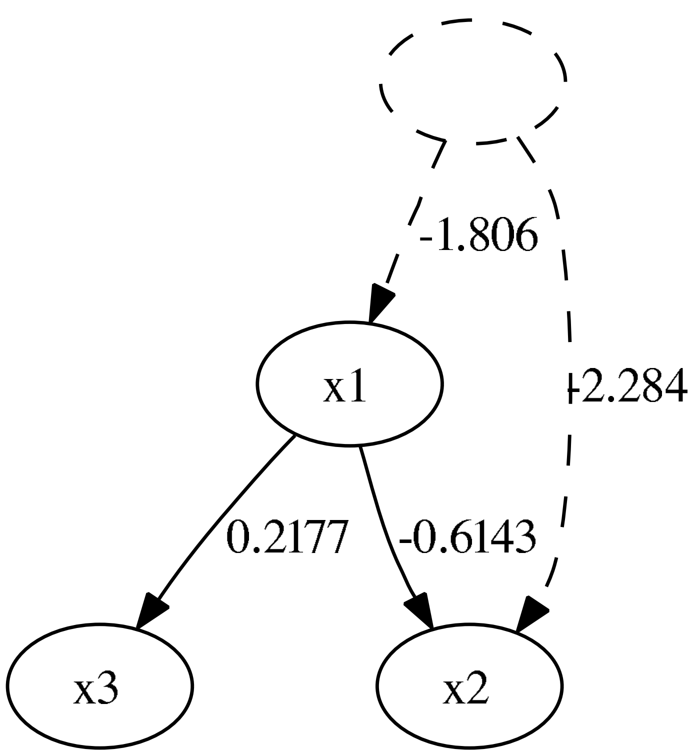


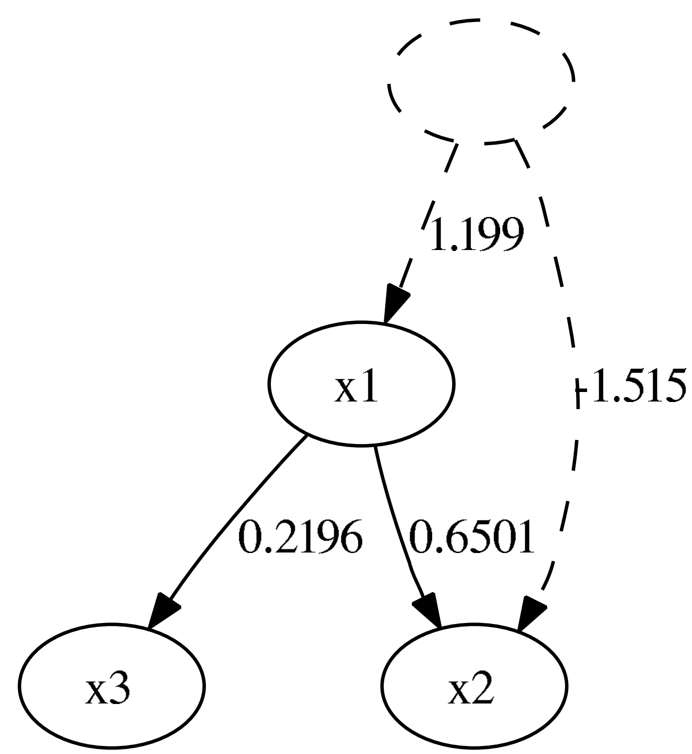
Some Simulation Result I

- Simulate 2500 data points with non-Gaussian noise using this model:



- Output of the algorithm:

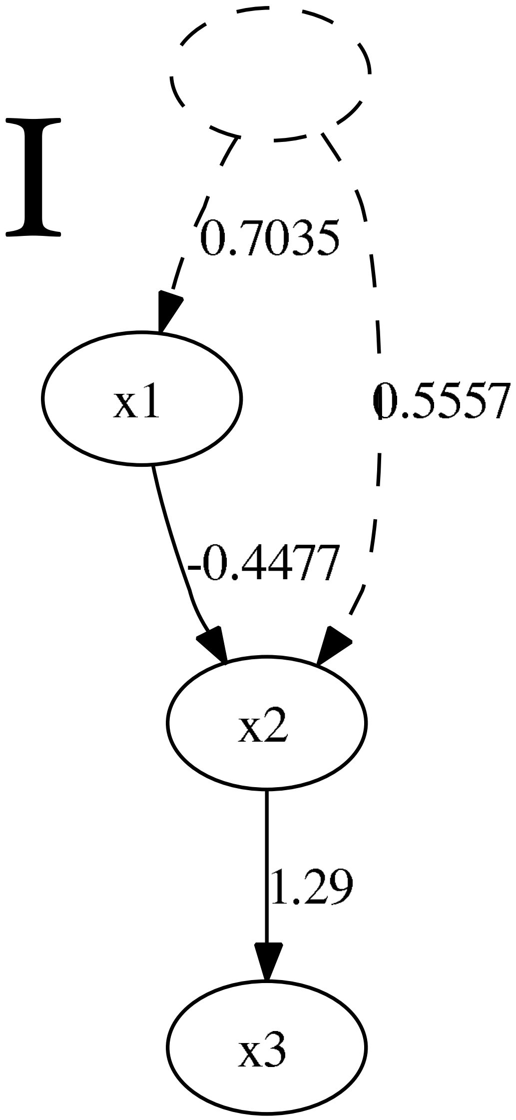




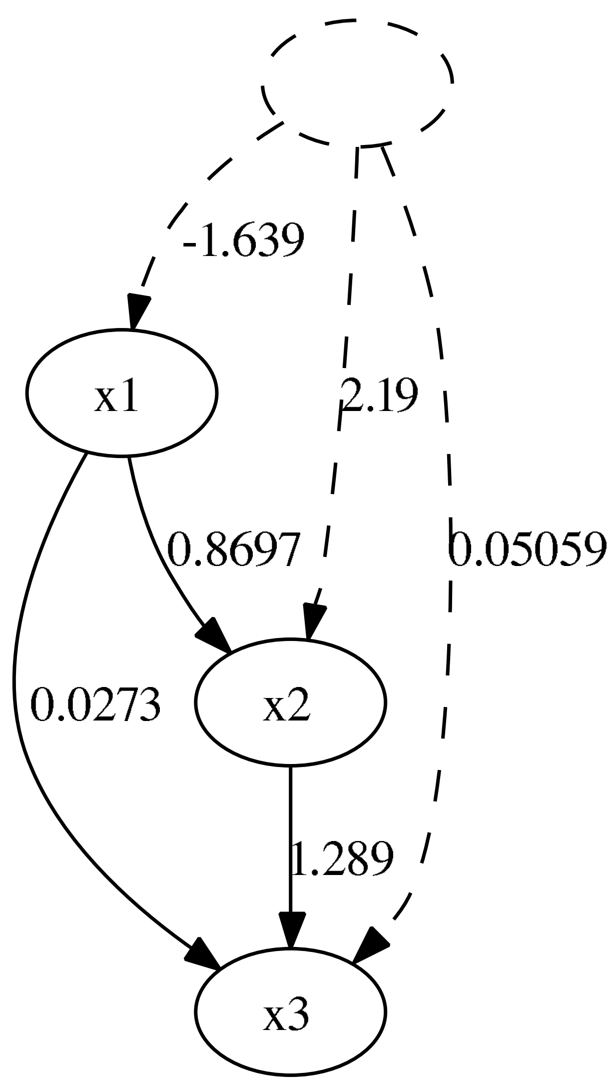
Hoyer et al. (2008). Estimation of causal effects using linear nonGaussian causal models with hidden variables. International Journal of Approximate Reasoning, 49(2):362– 378.

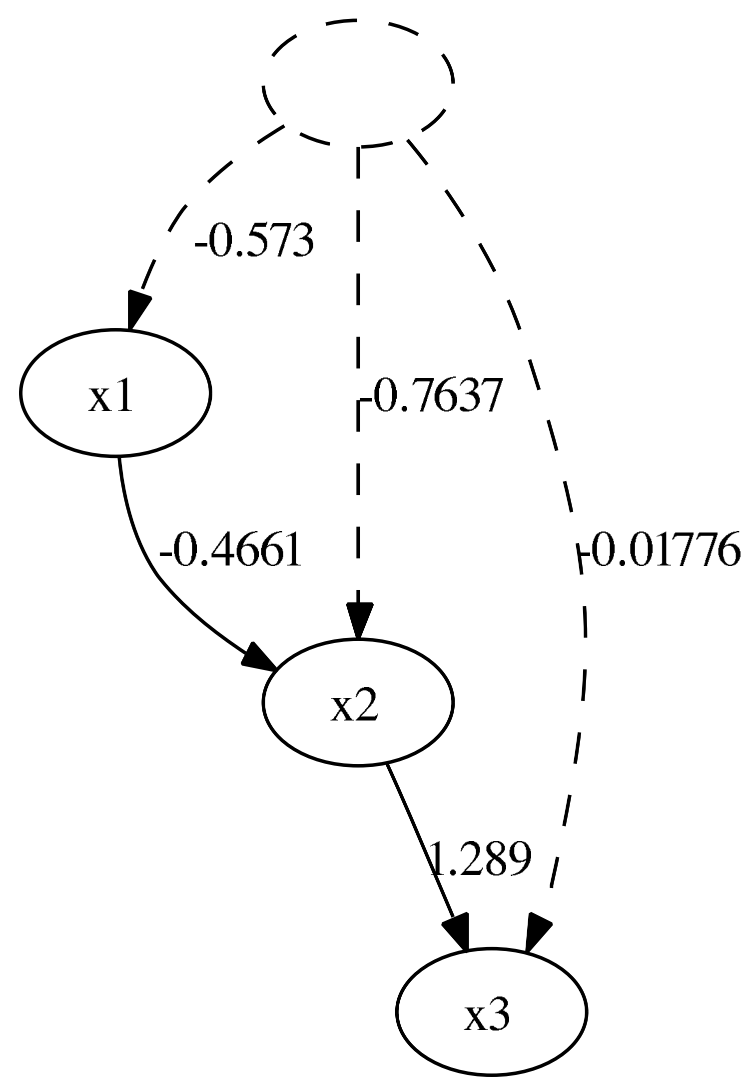
Some Simulation Result II

- Simulate 2500 data points with non-Gaussian noise using this model:



- Output of the algorithm:

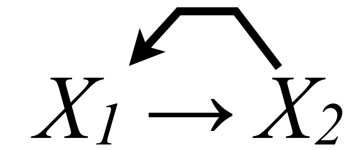




With Cycles

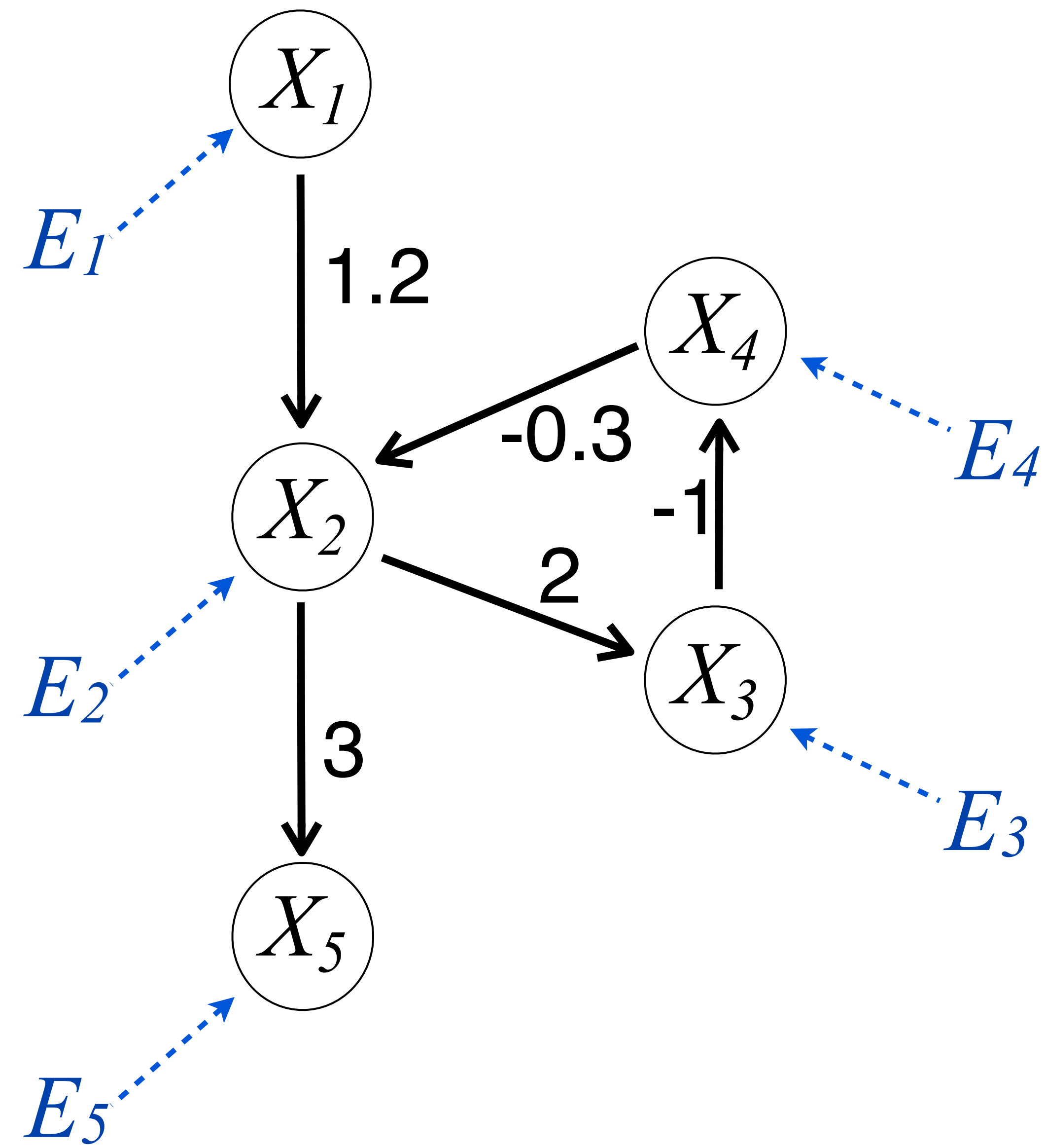
- Interpretation of cyclic causal relations
- ICA-based approach to estimating cyclic causal models

Discussion II: Feedback



- Causal relations may have cycles; Consider an example

$$X_1 = E_1$$
$$X_2 = 1.2X_1 - 0.3X_4 + E_2$$
$$X_3 = 2X_2 + E_3$$
$$X_4 = -X_3 + E_4$$
$$X_5 = 3X_2 + E_5$$



Or in matrix form, $\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$, where

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & 0 & -0.3 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix}$$

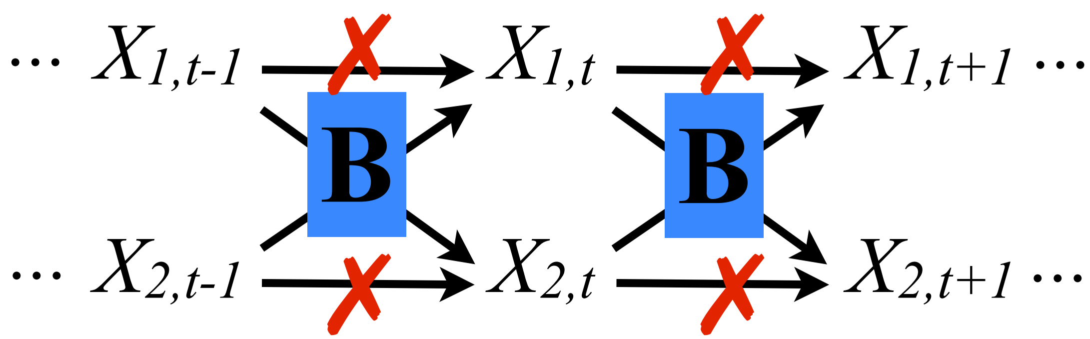
Lacerda, Spirtes, Ramsey and Hoyer (2008). Discovering cyclic causal models by independent component analysis. In Proc. UAI.

A conditional-independence-based method is given in T. Richardson (1996) - A Polynomial-Time Algorithm for Deciding Markov Equivalence of Directed Cyclic Graphical Models. Proc. UAI

Why Feedbacks?

$X_1 \rightleftarrows X_2$

- Some situations where we can recover cycles with ICA:
  - Each process reaches its equilibrium state & we observe the equilibrium states of multiple processes



$$\mathbf{X}_t = \mathbf{B}\mathbf{X}_{t-1} + \mathbf{E}_t.$$

At convergence we have $X_t = X_{t-1}$ for each dynamical process, so

$$\mathbf{X}_t = \mathbf{B}\mathbf{X}_t + \mathbf{E}_t, \quad \text{or} \quad \mathbf{E}_t = (\mathbf{I} - \mathbf{B})\mathbf{X}_t.$$

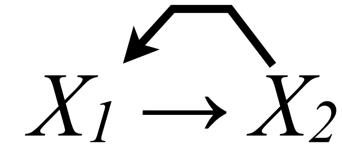
  - On temporally aggregated data

Suppose the underlying process is $\tilde{\mathbf{X}}_t = \mathbf{B}\tilde{\mathbf{X}}_{t-1} + \tilde{\mathbf{E}}_t$, but we just observe $\mathbf{X}_t = \frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{X}}_{t+k}$. Since

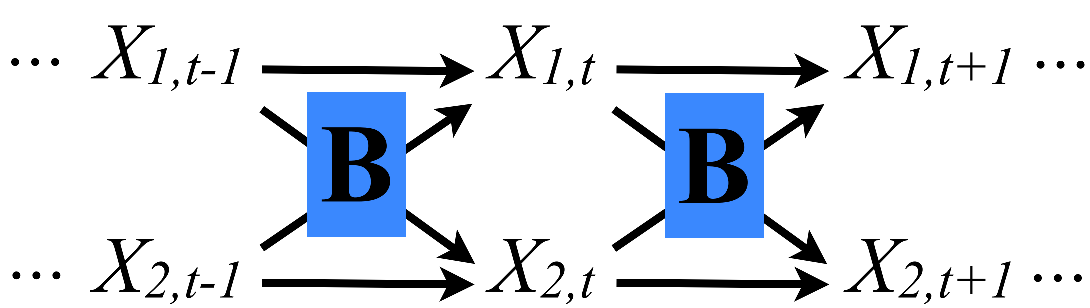
$$\frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{X}}_{t+k} = \mathbf{B}\frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{X}}_{t+k-1} + \frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{E}}_{t+k}.$$

We have $\mathbf{X}_t = \mathbf{B}\mathbf{X}_t + \mathbf{E}_t$ as $L \to \infty$.

Examples



- Some situations where we can recover cycles with ICA:
  - Each process reaches its equilibrium state & we observe the equilibrium states of multiple processes



Consider the price and demand of the same product in different states:

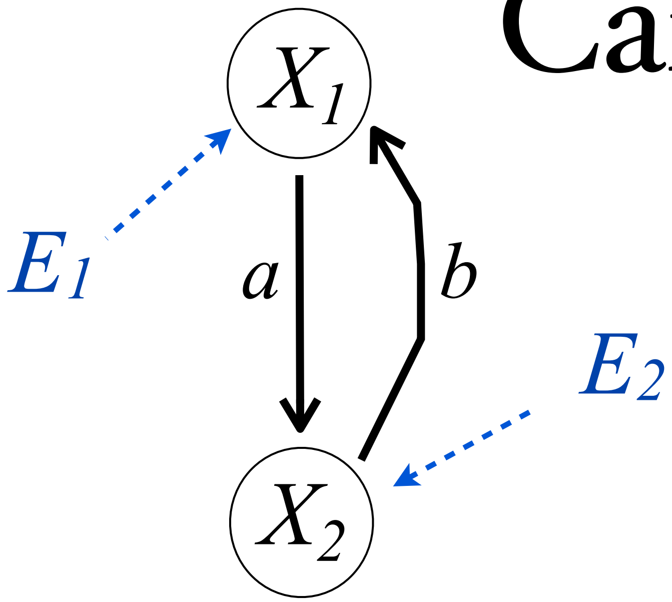
$$\text{price}_t = b_1 \cdot \text{price}_{t-1} + b_2 \cdot \text{demand}_{t-1} + E_1$$
$$\text{demand}_t = b_3 \cdot \text{price}_{t-1} + b_4 \cdot \text{demand}_{t-1} + E_2$$

  - On temporally aggregated data

Suppose the underlying process is $\tilde{\mathbf{X}}_t = \mathbf{B}\tilde{\mathbf{X}}_{t-1} + \tilde{\mathbf{E}}_t$, but we just observe $\mathbf{X}_t = \frac{1}{L}\sum_{k=1}^{L} \tilde{\mathbf{X}}_{t+k}$.

Consider the causal relation between two stocks: the causal influence takes place very quickly ($\sim$ 1-2 minutes) but we only have daily returns.

Can We Recover Cyclic Relations?



Suppose we have the process

$$\mathbf{X}_t = \underbrace{\begin{bmatrix} 0 & b \\ a & 0 \end{bmatrix}}_{\mathbf{B}} \mathbf{X}_t + \mathbf{E}_t.$$

That is,

$$(\mathbf{I} - \mathbf{B})\mathbf{X} = \mathbf{E}, \quad \text{or} \quad \underbrace{\begin{bmatrix} 1 & -b \\ -a & 1 \end{bmatrix}}_{\mathbf{W}} \mathbf{X}_t = \mathbf{E}_t$$

$$\Rightarrow \underbrace{\begin{bmatrix} -a & 1 \\ 1 & -b \end{bmatrix}}_{\mathbf{W'}} \mathbf{X}_t = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \cdot \mathbf{E}_t$$

$$\Rightarrow \begin{bmatrix} 1 & -1/a \\ -1/b & 1 \end{bmatrix} \mathbf{X}_t = \begin{bmatrix} 0 & -1/a \\ -1/b & 0 \end{bmatrix} \cdot \mathbf{E}_t$$

$$\Rightarrow \mathbf{X}_t = \underbrace{\begin{bmatrix} 0 & 1/a \\ 1/b & 0 \end{bmatrix}}_{\mathbf{B'}} \mathbf{X}_t + \begin{bmatrix} 0 & -1/a \\ -1/b & 0 \end{bmatrix} \cdot \mathbf{E}_t.$$

- **E = (I-B)X**; ICA can give **Y = WX**
- Without cycles: unique solution to **B**
- With cycles: solutions to **B** not unique any more; why? :-(
  - A 2-D example?
- Only one solution is stable (assuming no self-loops), i.e., s.t. |*product of coefficients over the cycle*| < 1 :-)

Summary:
1. Still *m* independent components;
2. **W** cannot be permuted to be lower-triangular

Can You Find the Alternative Causal Model ?

$$\begin{tikzcd} X_1 \arrow[r] & X_2 \arrow[l, bend right] \end{tikzcd}$$

- For this example...

$$X_1 = E_1$$
$$X_2 = 1.2X_1 - 0.3X_4 + E_2$$
$$X_3 = 2X_2 + E_3$$
$$X_4 = -X_3 + E_4$$
$$X_5 = 3X_2 + E_5$$

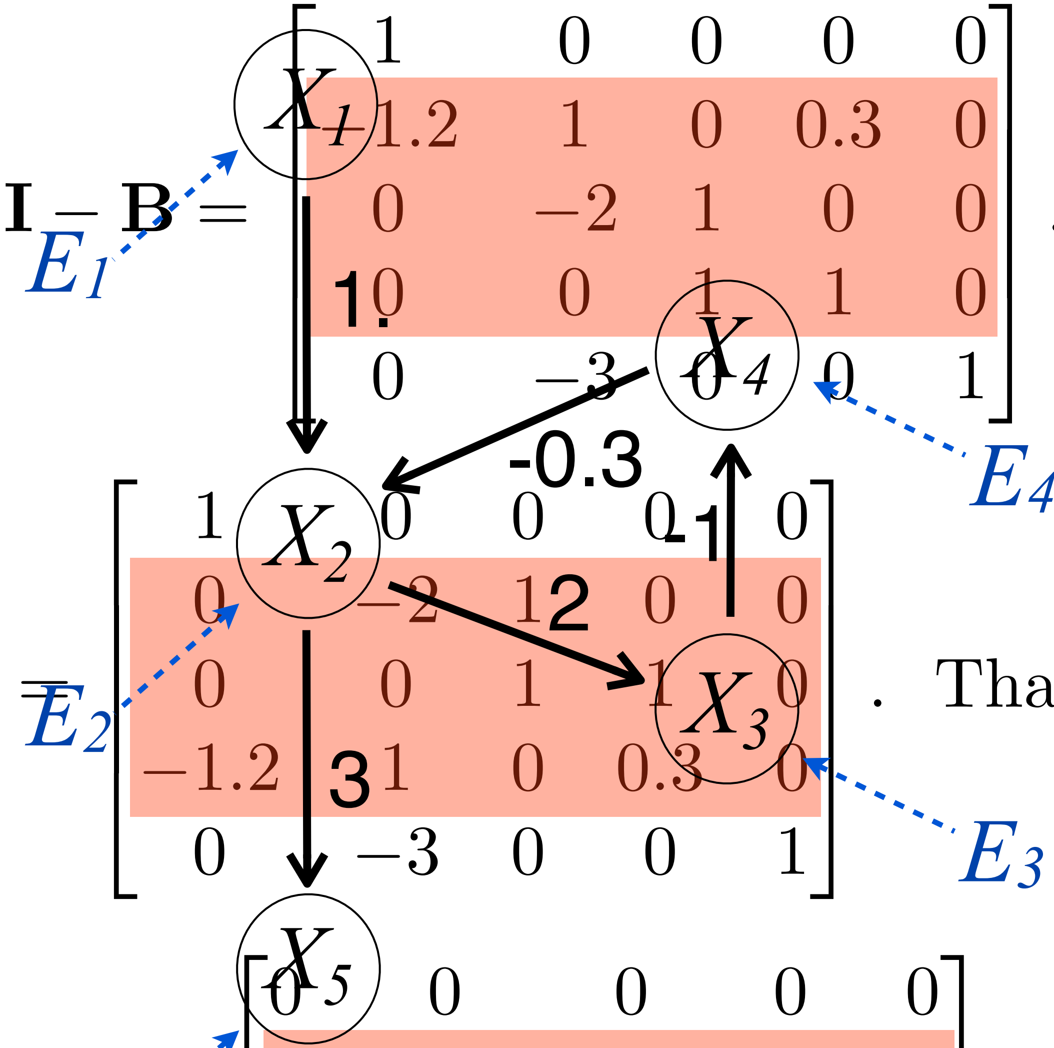
Or in matrix form, $\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$, where

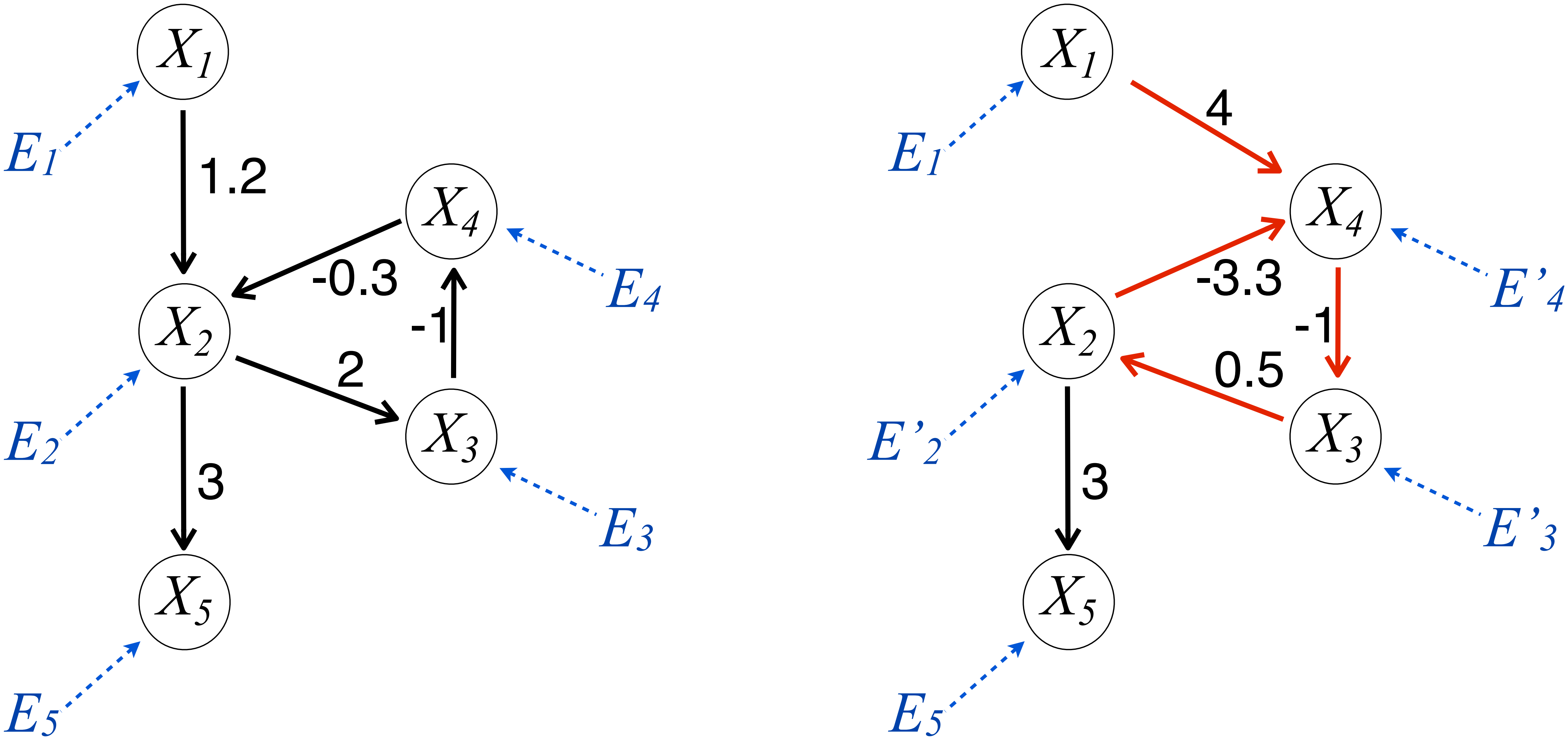
$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & 0 & -0.3 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{I} - \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1.2 & 1 & 0 & 0.3 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & -3 & 0 & 0 & 1 \end{bmatrix}.$$

$$\mathbf{W}' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ -1.2 & 1 & 0 & 0.3 & 0 \\ 0 & -3 & 0 & 0 & 1 \end{bmatrix}. \text{ That is,}$$

$$\mathbf{B}' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 4 & -3.3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix}.$$

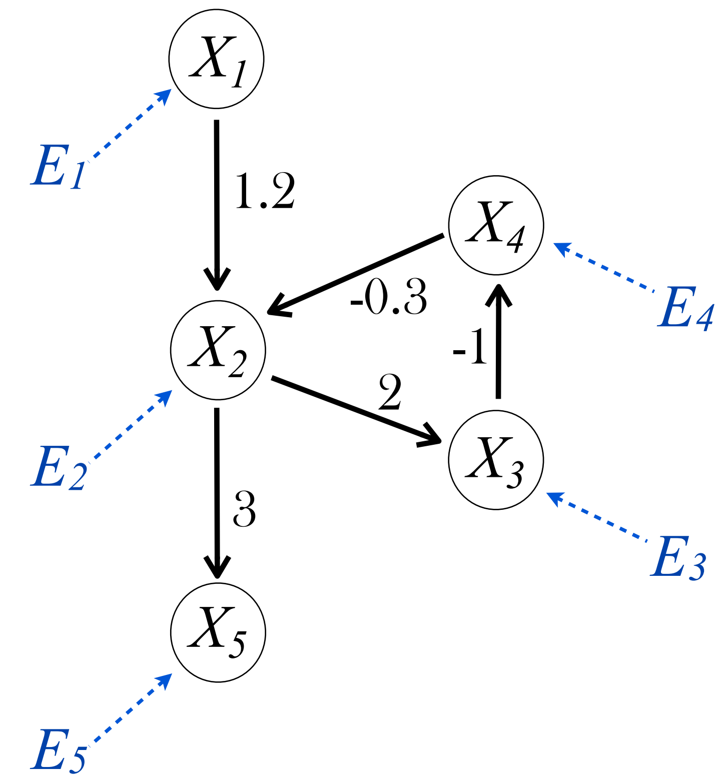




$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & 0 & -0.3 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B}' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 4 & -3.3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix}.$$

Some Simulation Result

- Simulate 15000 data points with non-Gaussian noise using this model:



- Output of the algorithm:

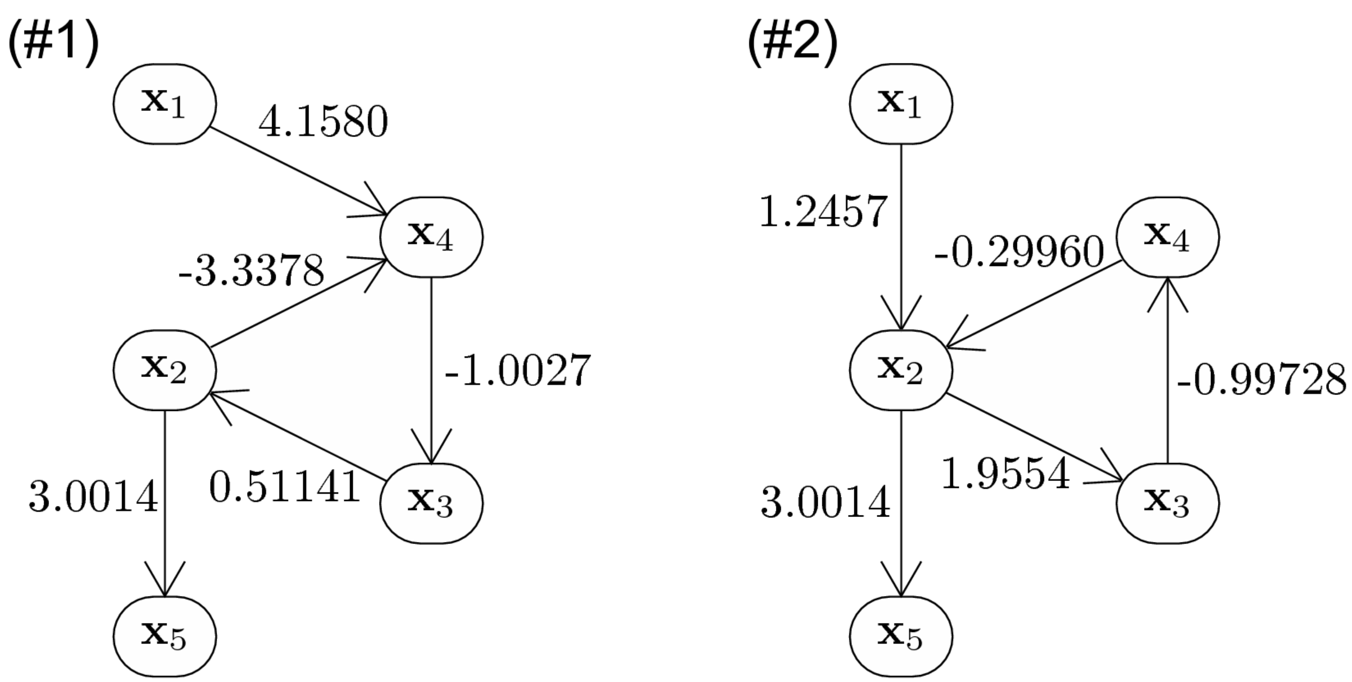


Fig. 3: The output of LiNG-D: Candidate #1 and Candidate #2

Lacerda, Spirtes, Ramsey and Hoyer (2008). Discovering cyclic causal models by independent component analysis. In Proc. UAI.

Summary of the Two Situations

- Can you distinguish between the following situations from ICA result **Y** = **WX** ?
  - cycles:
  - confounders:

1. **Y** still has *m* independent components;
2. **W** cannot be permuted to be lower-triangular

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ a_3 & 1 & a_3 + \frac{a_2}{a_1} \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ a_1 Z \end{bmatrix}$$

Y produced by ordinary ICA does not have independent components

- Either of them makes causal discovery more difficult
- They happen very often, even in the same problem

Take-Home Message

- Constraint-based causal discovery makes use of conditional independence relationships
  - Asymptotically correct, but behavior on finite samples not guaranteed
  - Wide applicability! Worth trying on complex problems
  - Equivalence class!
- Linear non-Gaussian case: Causal model fully identifiable
  - Based on ICA or its variants
- How to tackle practical issues, e.g., confounders, cycles, and error-in-measurements, related to identifiability of the mixing procedure
- Nonlinearities?