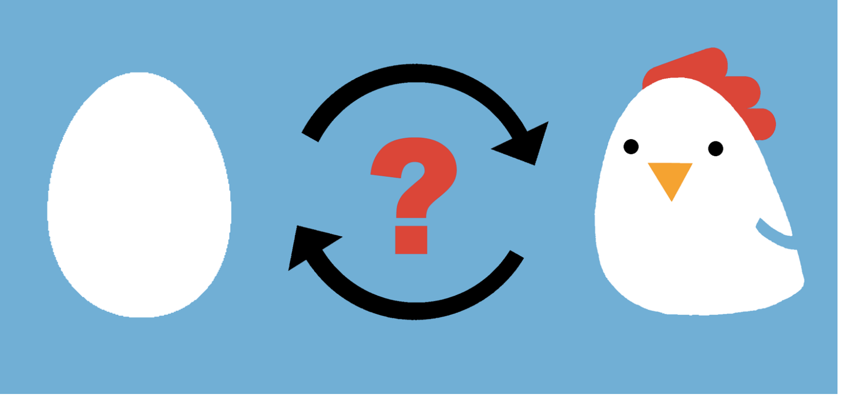


CBMS Conference -- Foundations of Causal Graphical Models and Structure Discovery

Lecture 7

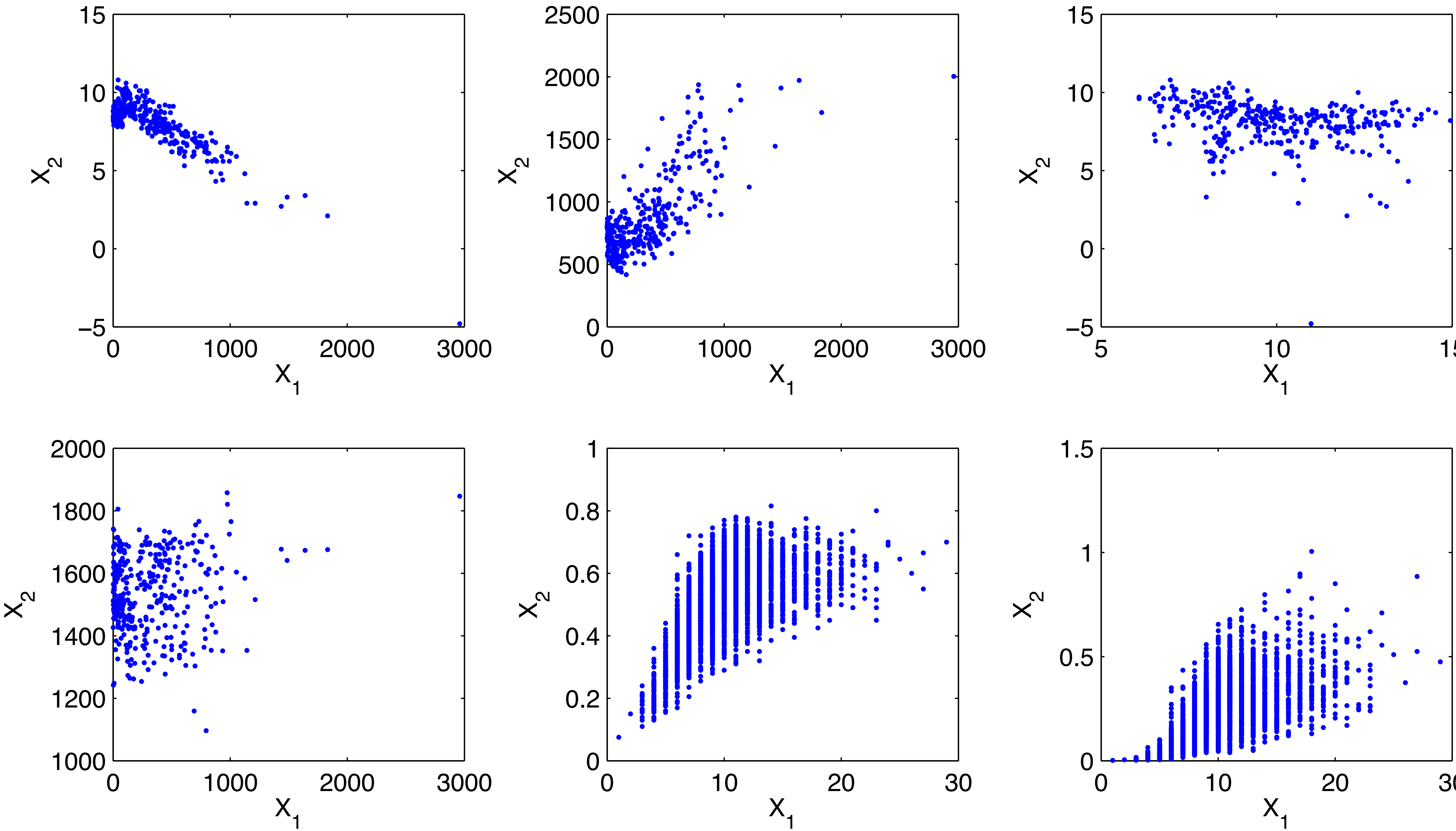
Causal Discovery Based on Linear, Non-Gaussian Models

Instructor: Kun Zhang

Carnegie Mellon University

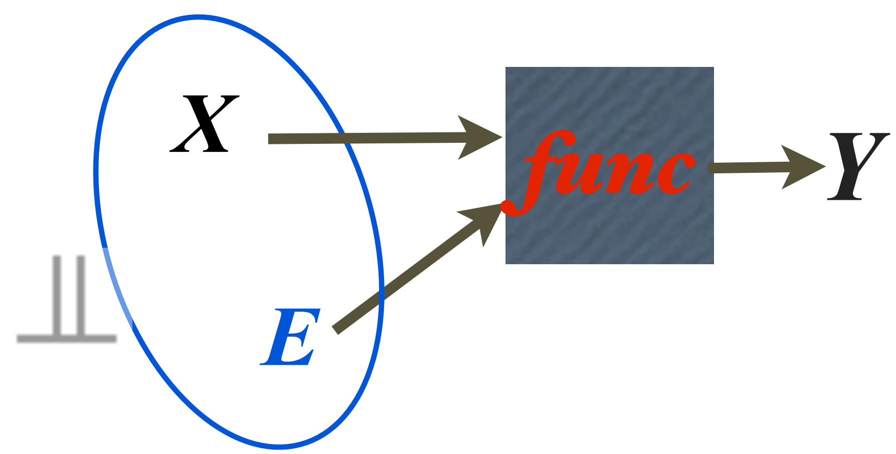
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Distinguishing Cause from Effect:
Examples (Tübingen Cause-Effect Pairs)



A Causal Process

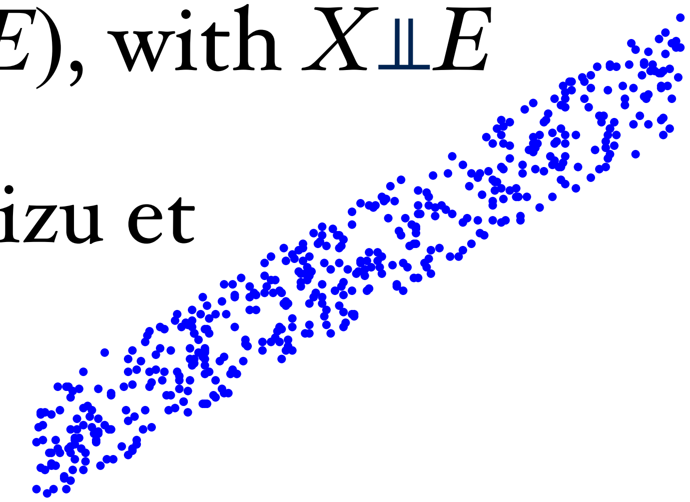
rain $\longrightarrow$ *wet ground*



Functional Causal Model

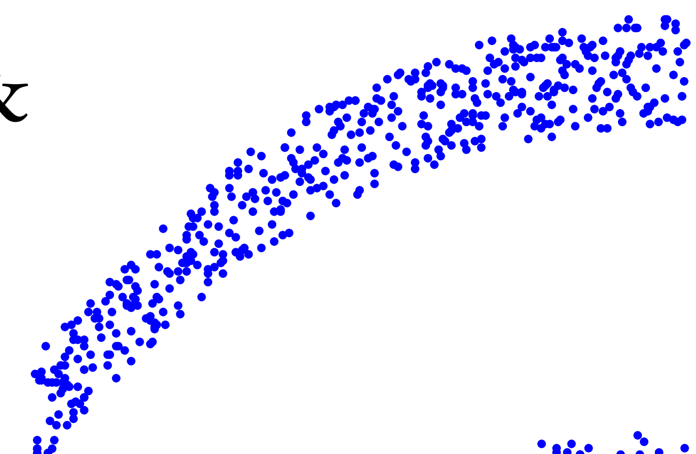
- A **functional causal model** represents effect as a function of direct causes and noise: $Y = f(X, E)$, with $X \perp\!\!\!\perp E$
- Linear non-Gaussian acyclic causal model (Shimizu et al., ‘06)

$$Y = \mathrm{a} \cdot X + E$$



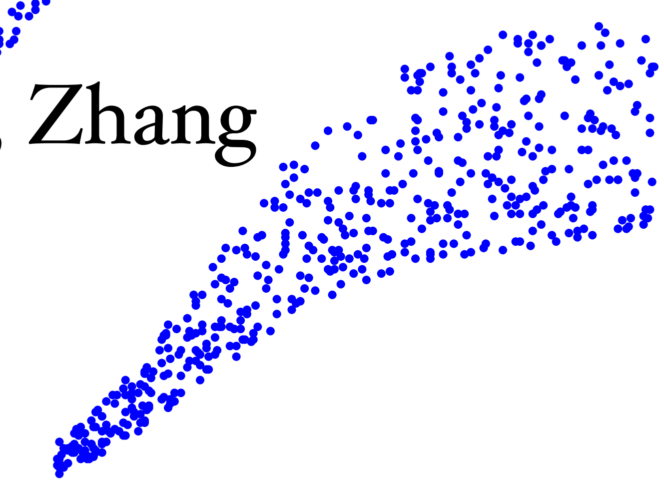
- Additive noise model (Hoyer et al., ’09; Zhang & Hyvärinen, ‘09b)

$$Y = f(X) + E$$



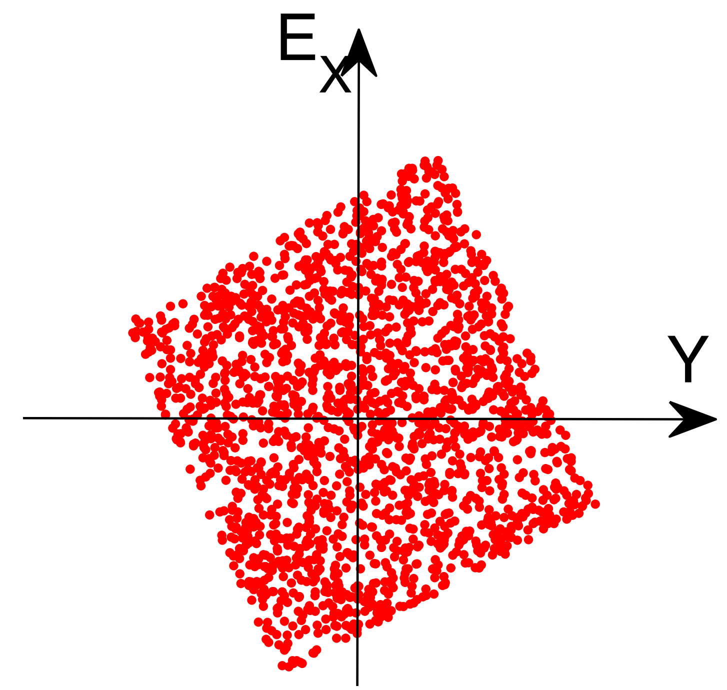
- Post-nonlinear causal model (Zhang & Chan, ’06; Zhang & Hyvärinen, ‘09a)

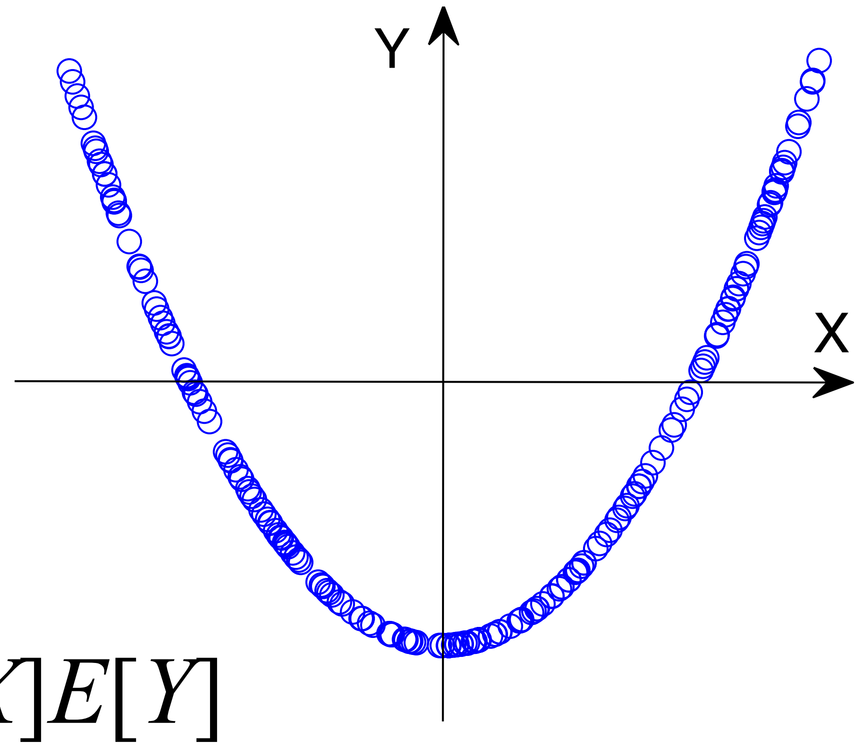
$$Y = f_2\,(\,f_1(X) + E\,)$$



(Conditional) Independence

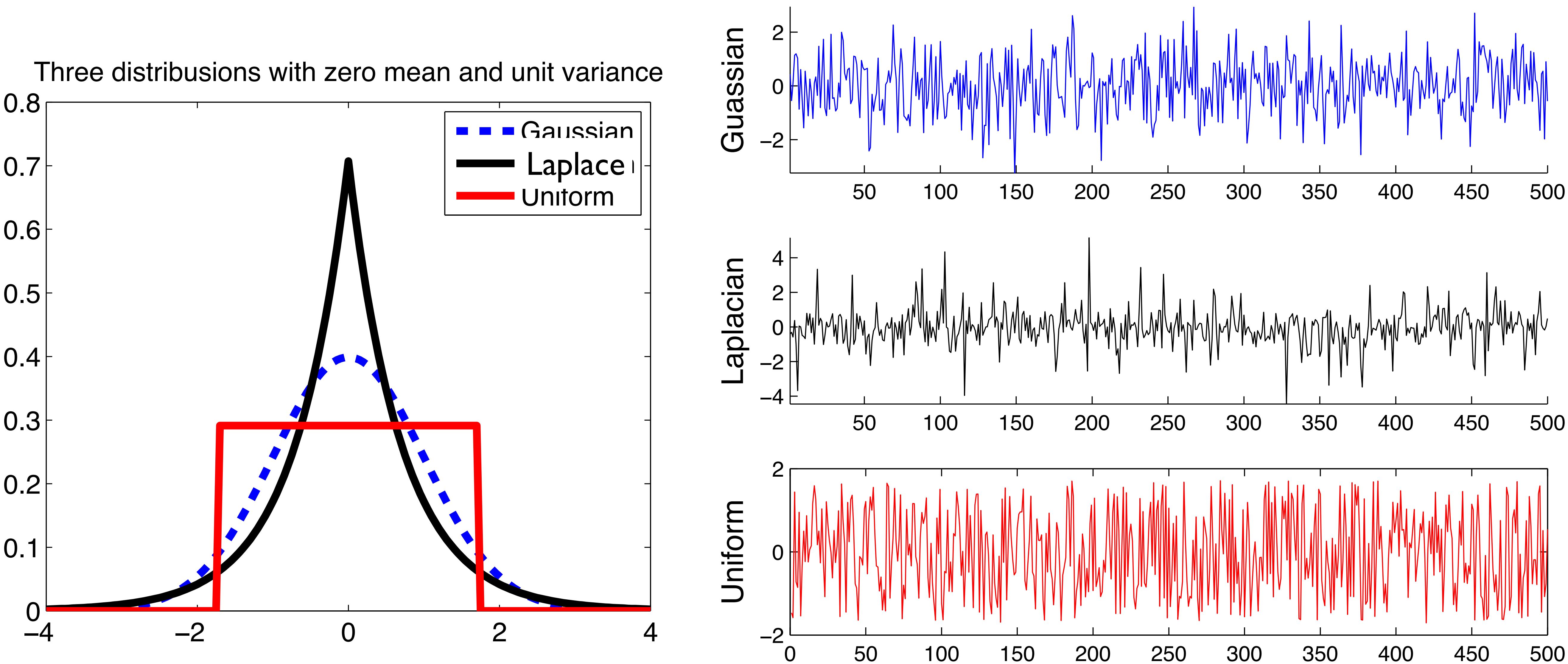
- $X \perp\!\!\!\perp Y$ iff $p(X,Y) = p(X)p(Y)$
  - or $p(X|Y) = P(X)$: Y not informative to X
- $X \perp\!\!\!\perp Y \mid Z$ iff $p(X,Y|Z) = p(X|Z)p(Y|Z)$
  - or, $p(X|Y,Z) = p(X|Z)$: given Z, Y not informative to X
- Divide & conquer, remove irrelevant info...
- By construction, regression residual is uncorrelated (but not necessarily independent !) from the predictor





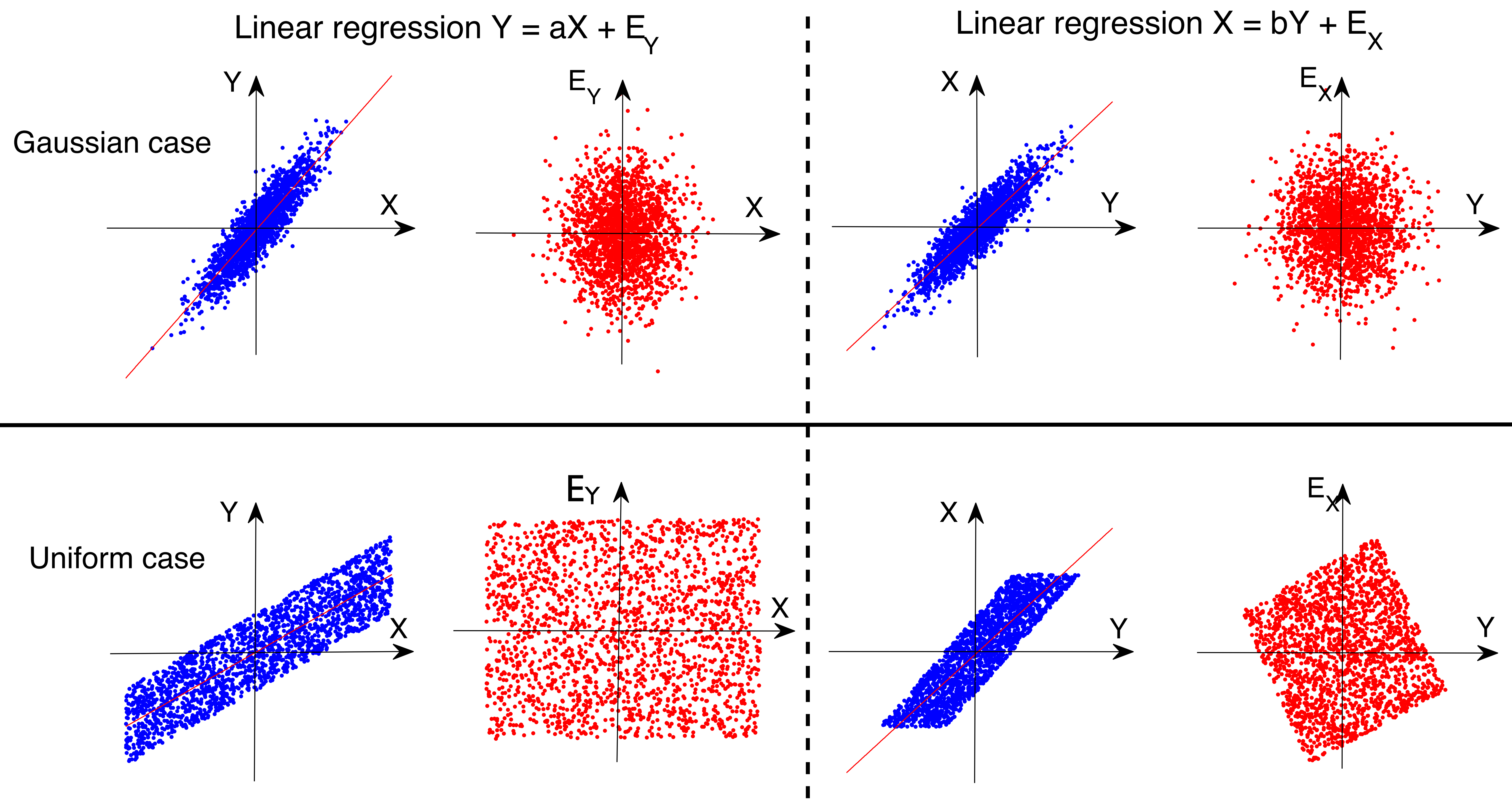
Uncorrelatedness: $E[XY] = E[X]E[Y]$

Gaussian vs. Non-Gaussian Distributions

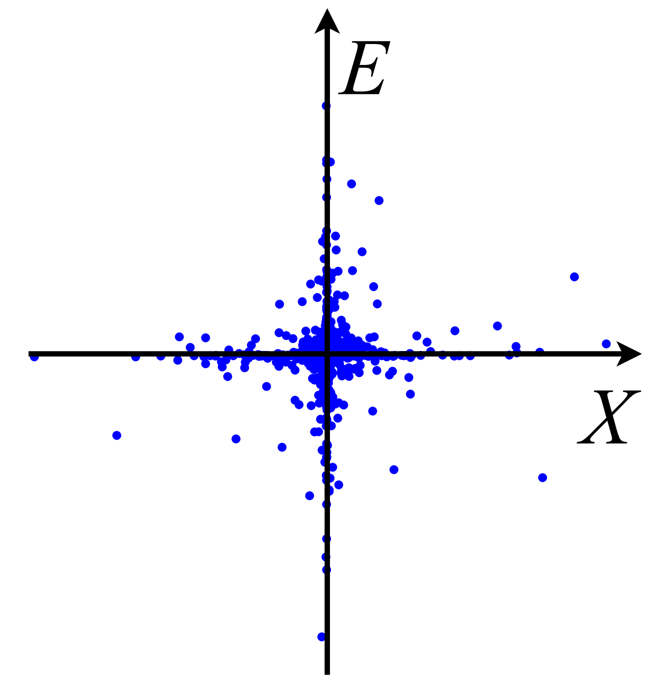


Causal Asymmetry the Linear Case: Illustration

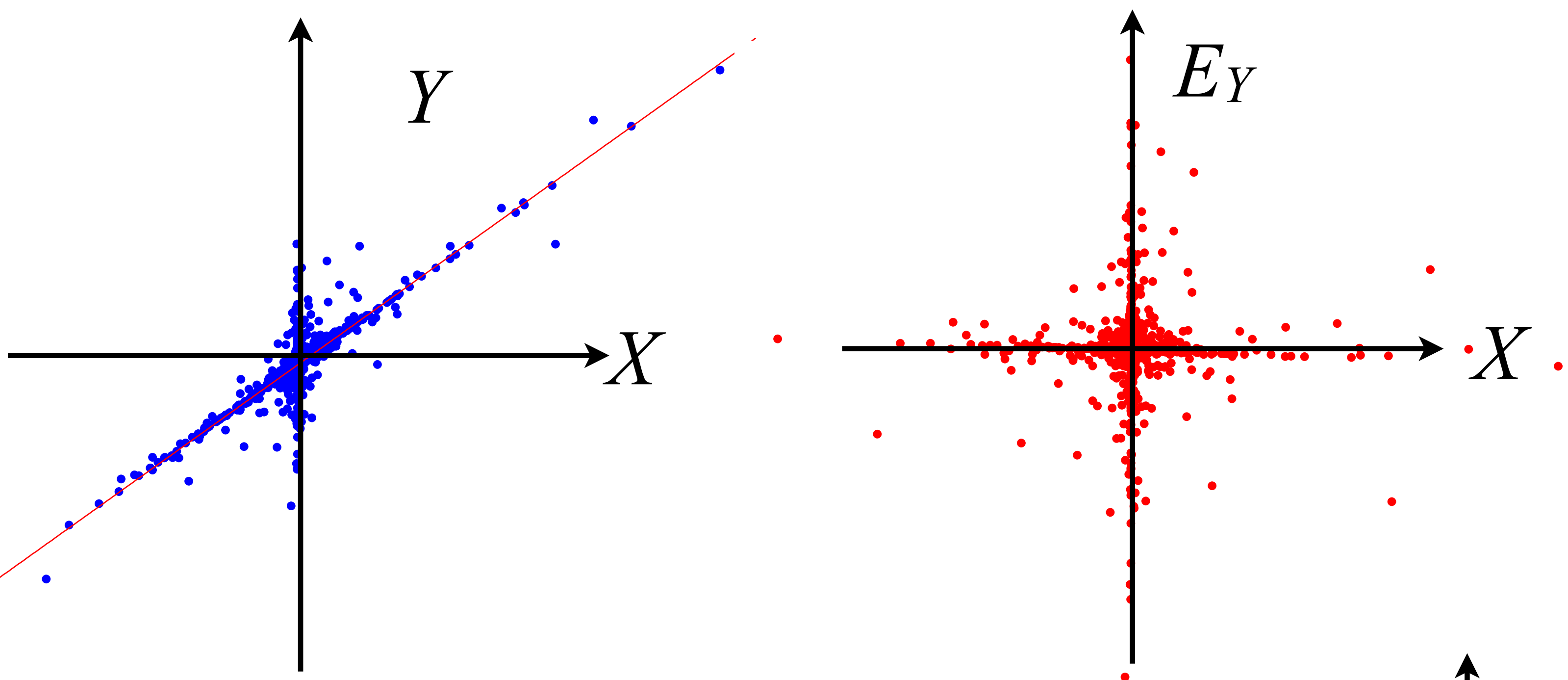
Data generated by $Y = aX + E$ (i.e., $X \rightarrow Y$):

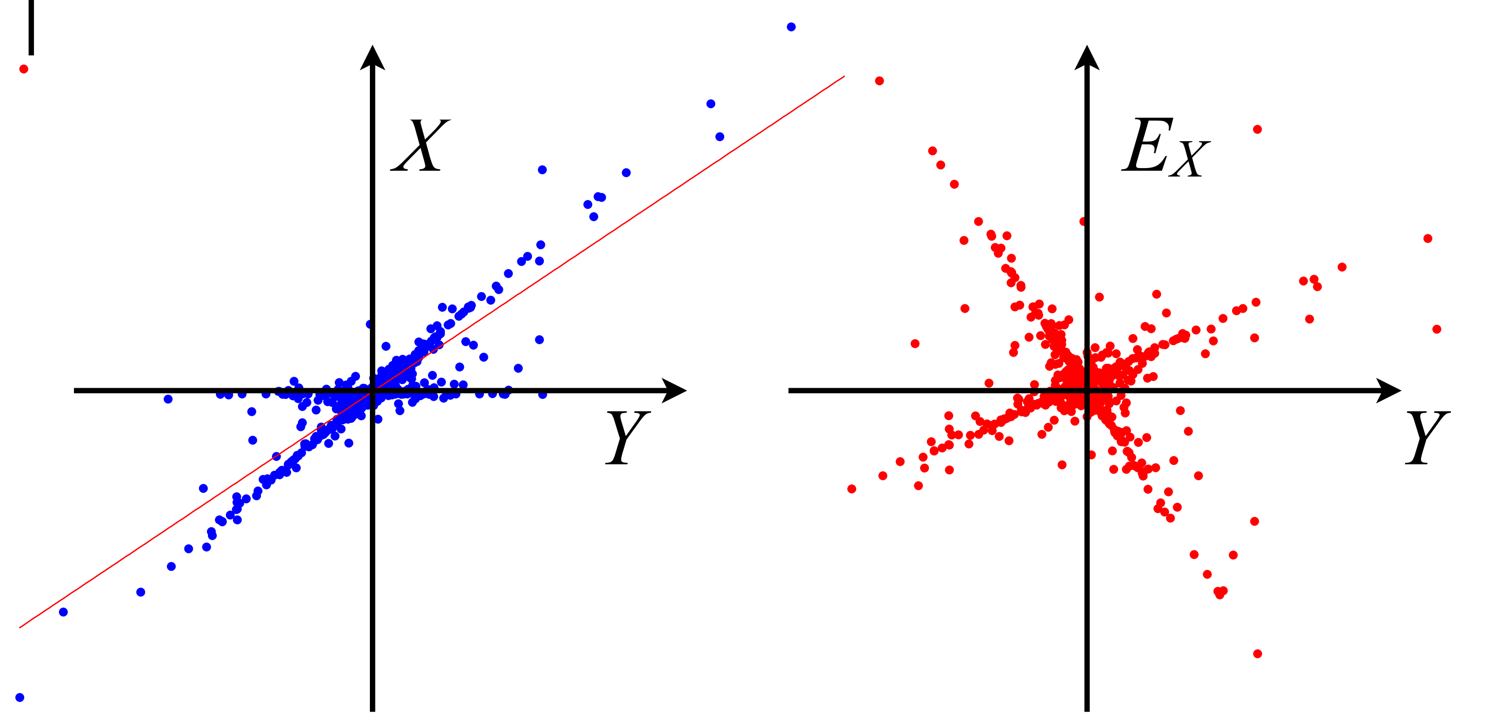


Super-Gaussian Case



Data generated by $Y = aX + E$ ($X \rightarrow Y$):





More Generally, LiNGAM Model

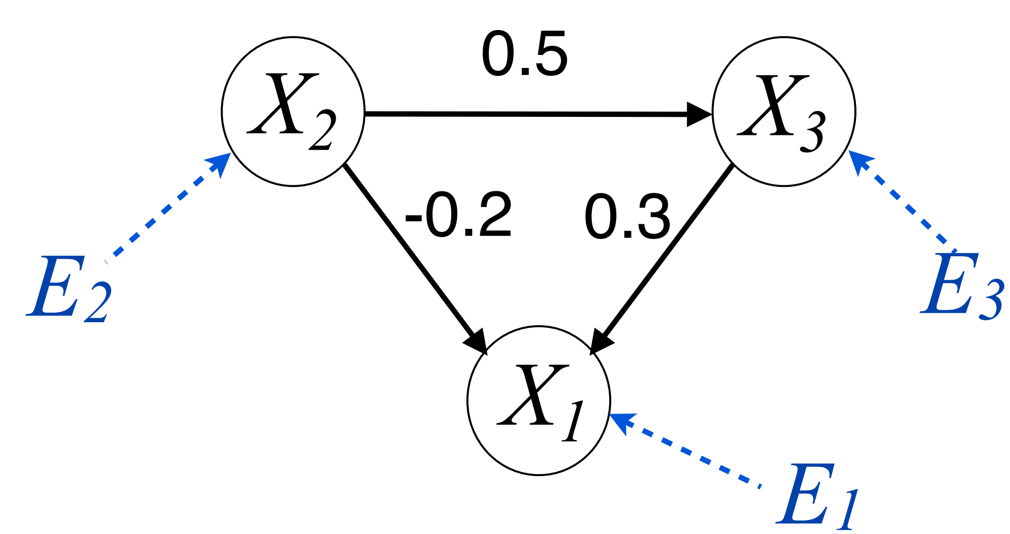
- Linear, non-Gaussian, acyclic causal model (LiNGAM) (Shimizu et al., 2006):

$$X_i = \sum_{j:\ \text{parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$$

  - Disturbances (errors) E_i are non-Gaussian (or at most one is Gaussian) and mutually independent

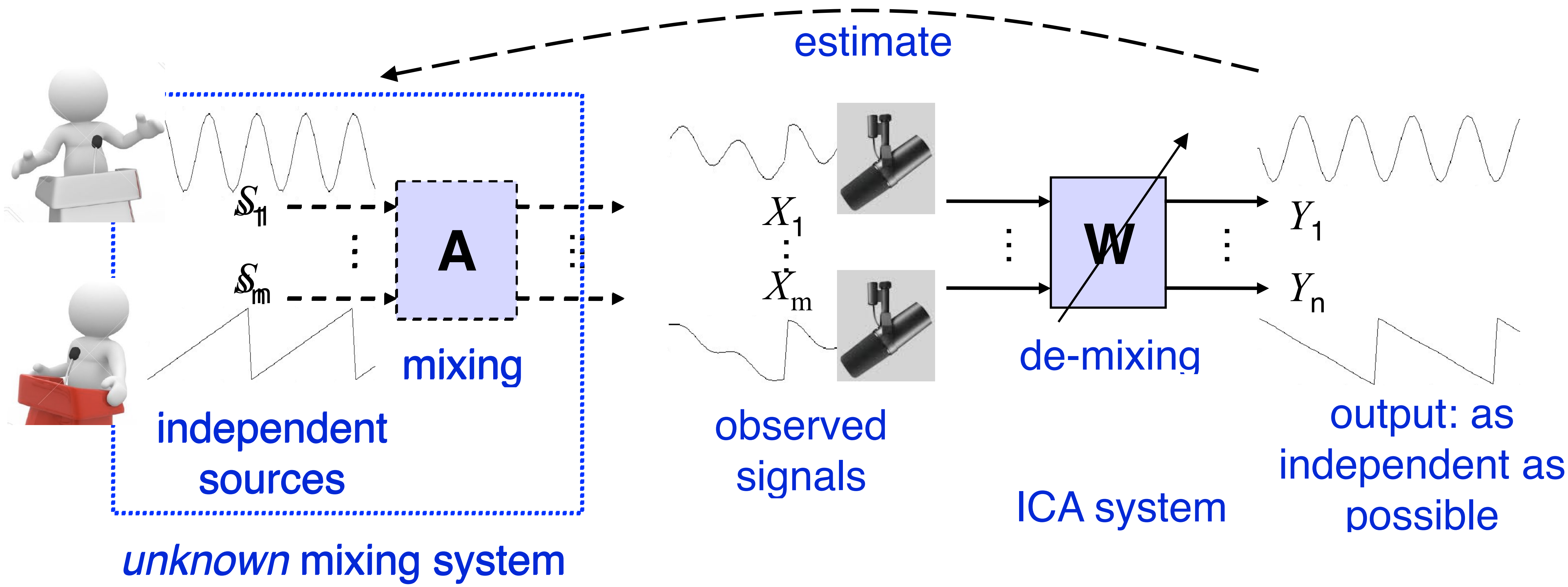
  - Example:

$$X_2 = E_2,$$
$$X_3 = 0.5X_2 + E_3,$$
$$X_1 = -0.2X_2 + 0.3X_3 + E_1.$$



Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.

Independent Component Analysis



$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S}$$

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

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

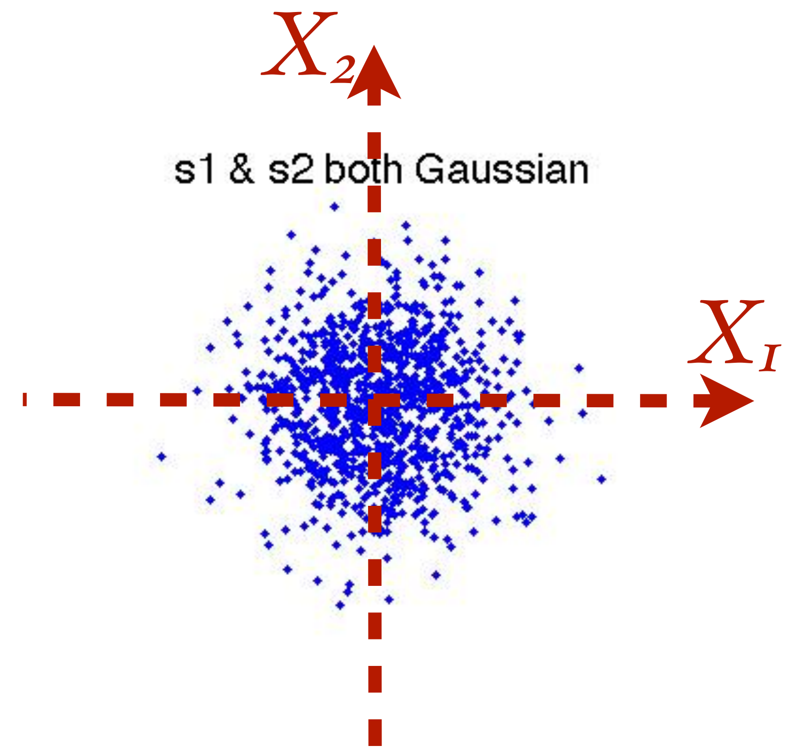
- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

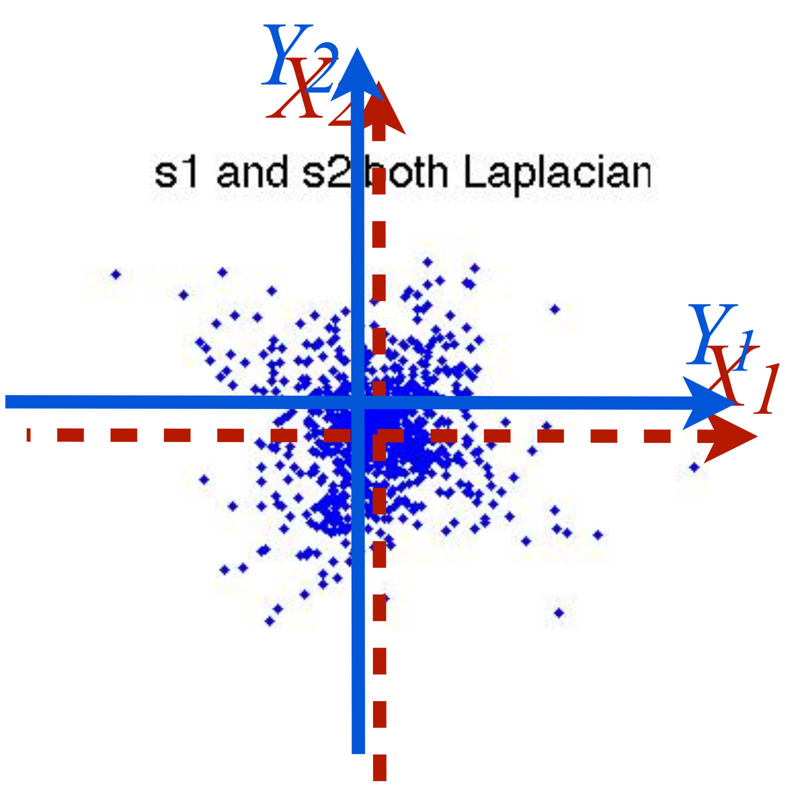
Then A can be estimated up to column scale and permutation indeterminacies

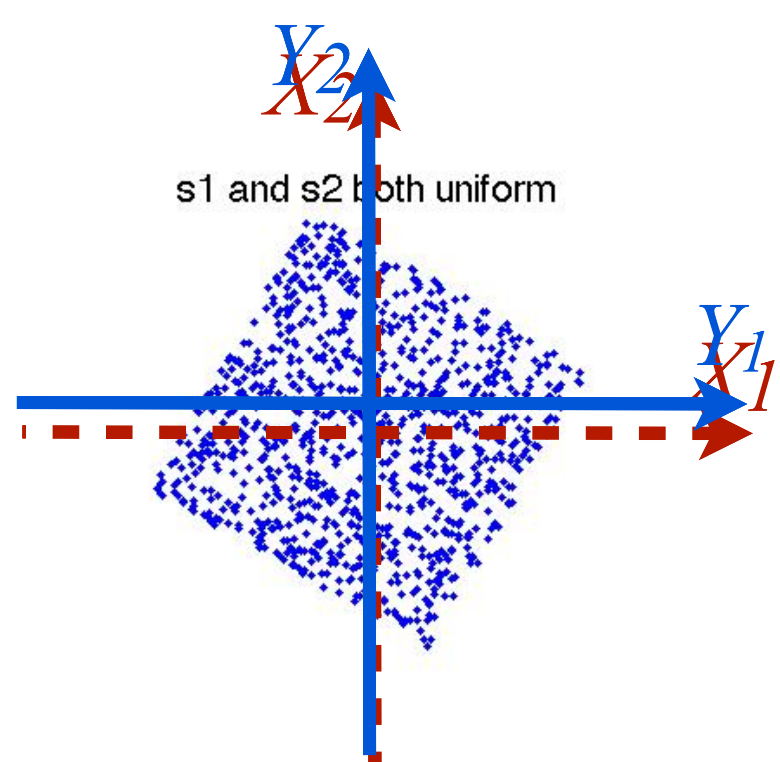
Hyvärinen et al., Independent Component Analysis, 2001

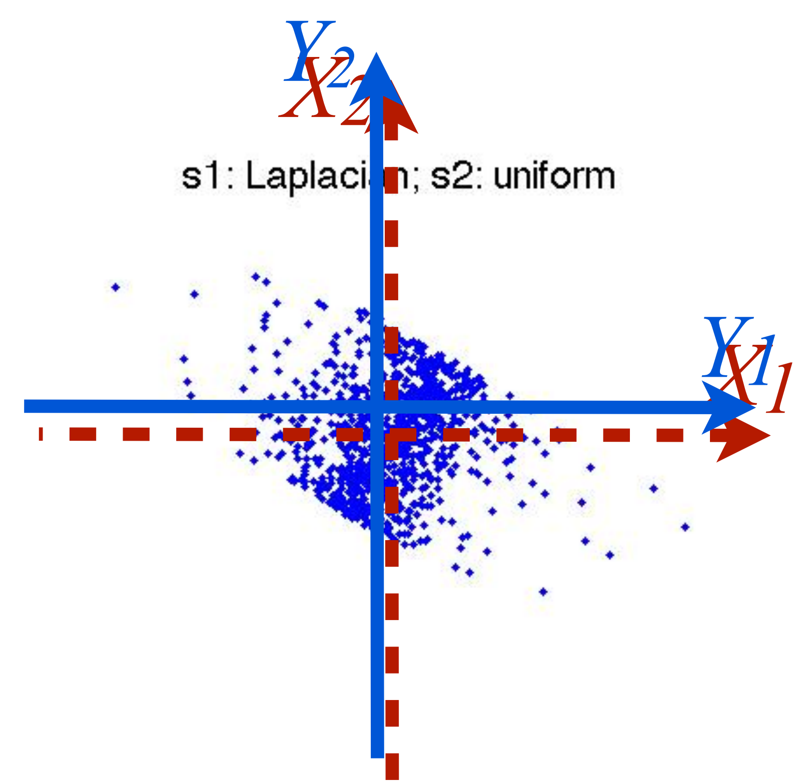
Intuition: Why ICA works?

- (After preprocessing) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$ to making Y_i independent
    - By maximum likelihood $\log p(X|\boldsymbol{A})$, mutual information $MI(Y_1, \ldots, Y_m)$ minimization, infomax...

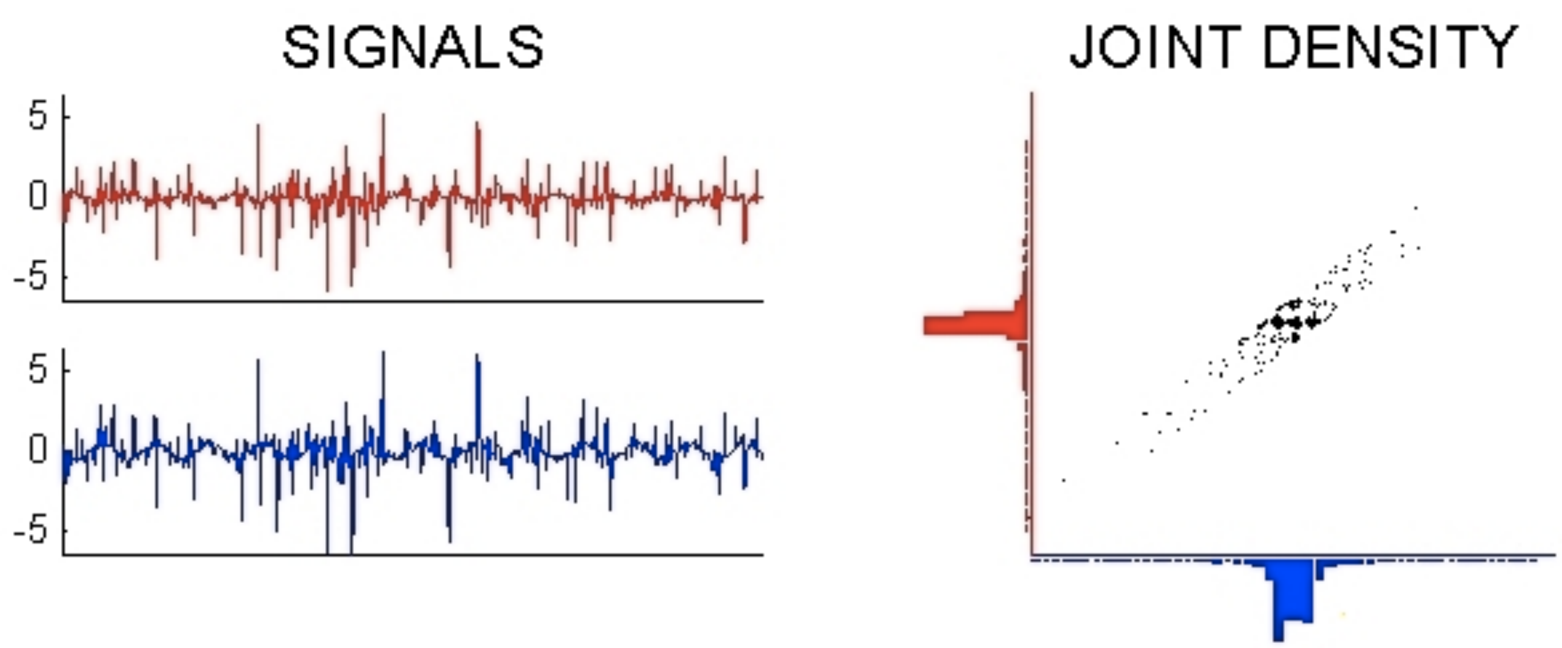




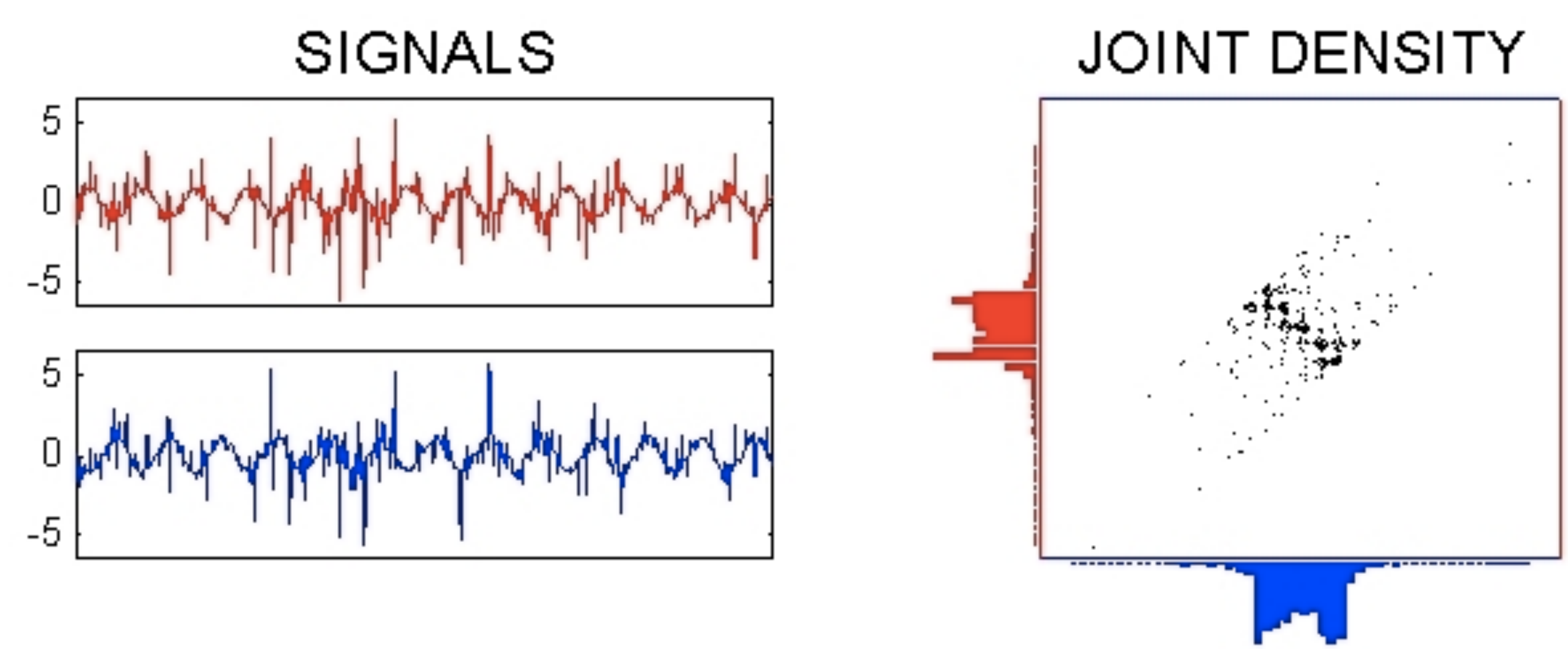




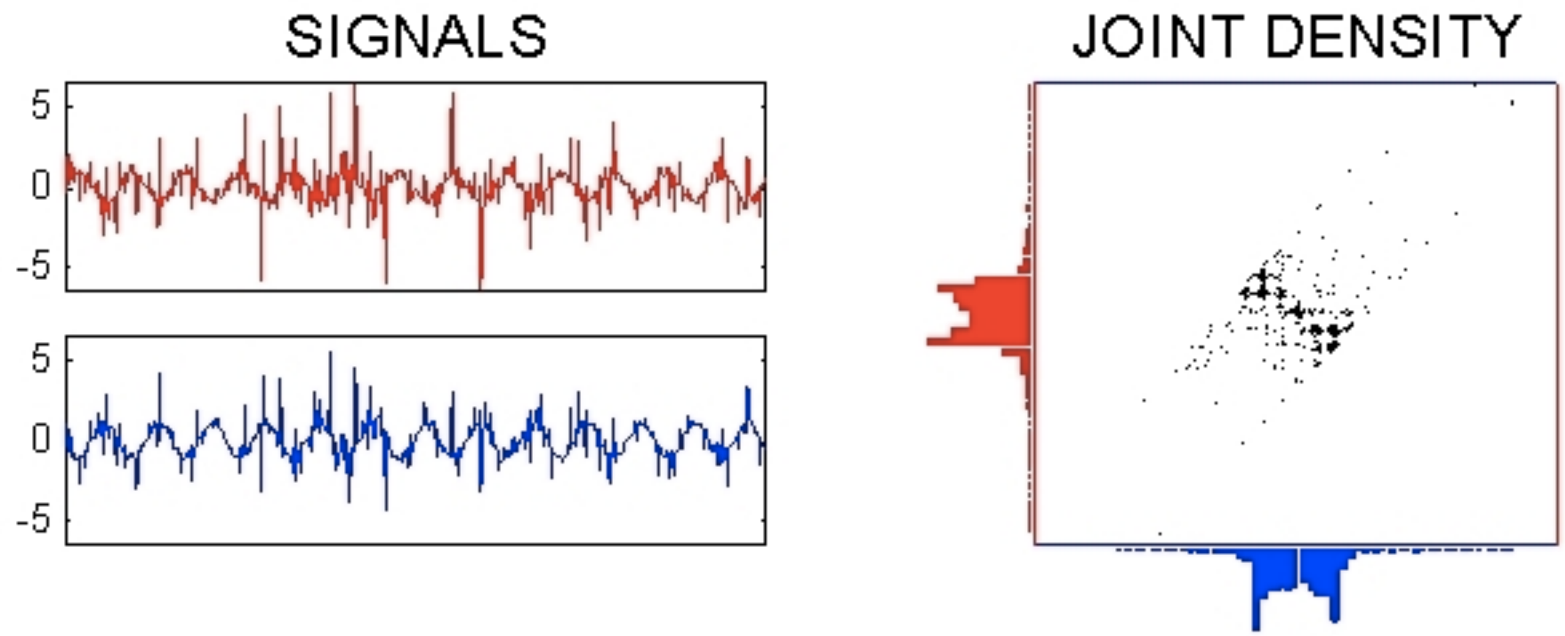
A Demo of the ICA Procedure



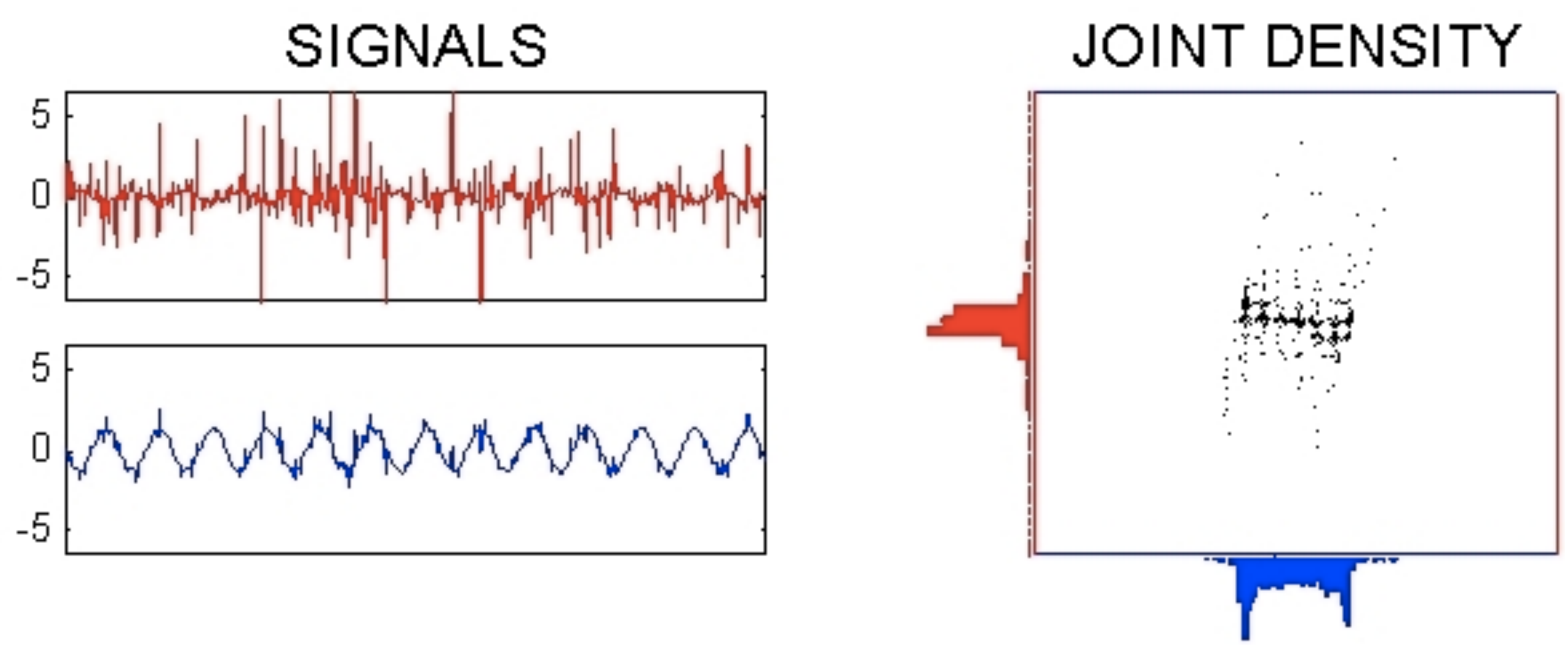
Input signals and density



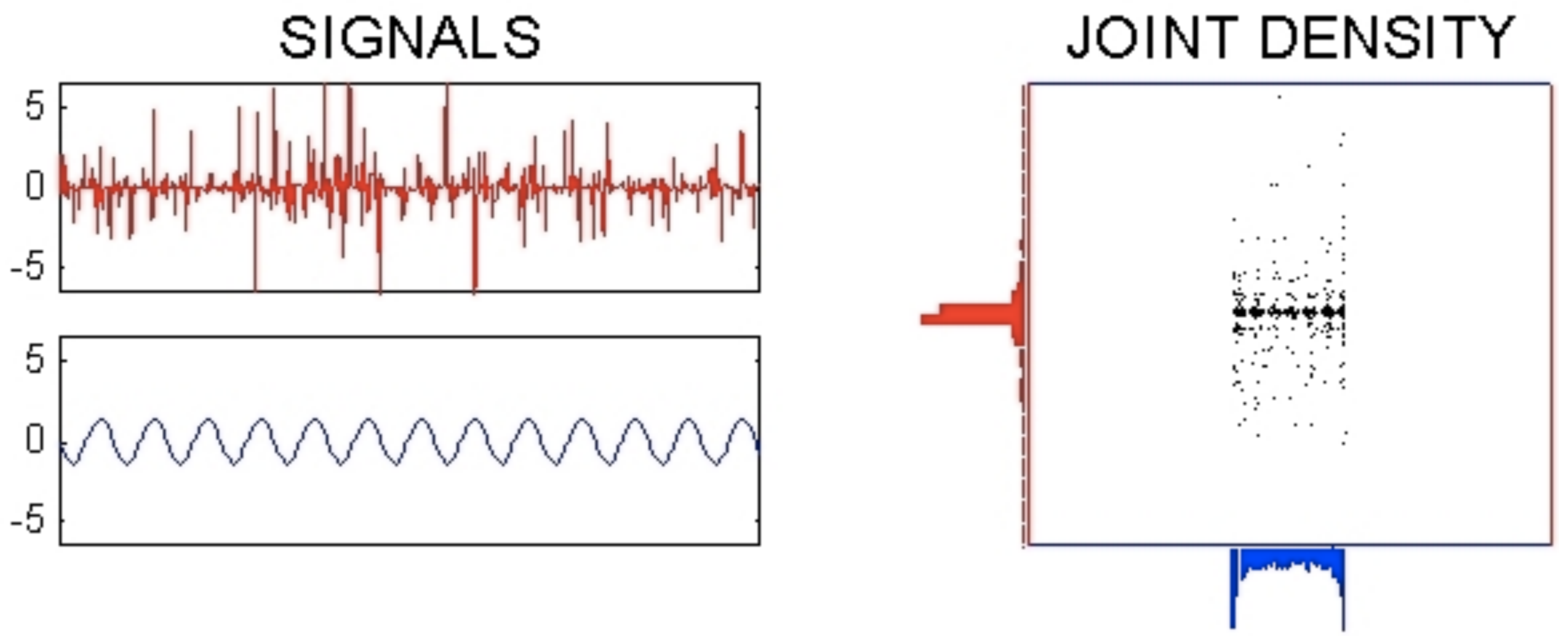
Whitened signals and density



Separated signals after 1 step of FastICA



Separated signals after 3 steps of FastICA



Separated signals after 5 steps of FastICA

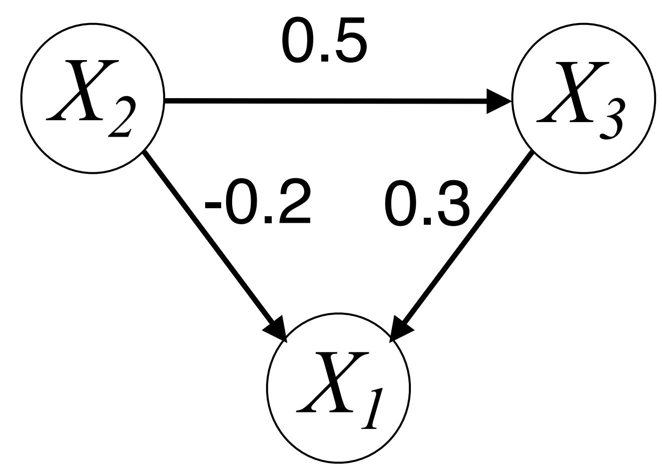
LiNGAM Analysis by ICA

- LiNGAM: $X_i = \sum_{j:\ \text{parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E} \Rightarrow$ **E = (I-B)X**
  - **B** has special structure: acyclic relations
- ICA: **Y = WX**
- **B** can be seen from **W**
  and re-scaling
- Faithfulness assumption avoided
- E.g.,

$$\begin{bmatrix} E_1 \\ E_3 \\ E_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.2 & -0.3 & 1 \end{bmatrix}}_{\mathbf{W}} \cdot \begin{bmatrix} X_2 \\ X_3 \\ X_1 \end{bmatrix}$$

$$\Leftrightarrow \begin{cases} X_2 = E_1 \\ X_3 = 0.5 X_2 + E_3 \\ X_1 = -0.2 X_2 + 0.3 X_3 + E_2 \end{cases}$$

So we have the causal relation:



Question 1. How to find W?

Question 2. How to see B from W?

LiNGAM Analysis by ICA

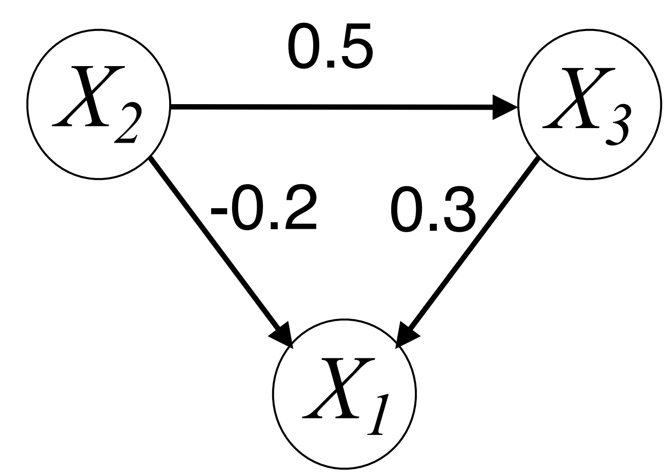
- LiNGAM: $X_i = \sum_{j:\ \text{parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E} \Rightarrow$ **E = (I-B)X**
  - **B** has special structure: acyclic relations
- ICA: **Y = WX**
- **B** can be seen from **W** by permutation and re-scaling
- Faithfulness assumption avoided
- E.g.,

$$\begin{bmatrix} E_1 \\ E_3 \\ E_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.2 & -0.3 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_2 \\ X_3 \\ X_1 \end{bmatrix}$$

$$\Leftrightarrow \begin{cases} X_2 = E_1 \\ X_3 = 0.5 X_2 + E_3 \\ X_1 = -0.2 X_2 + 0.3 X_3 + E_2 \end{cases}$$

1. First permute the rows of **W** to make all diagonal entries non-zero, yielding $\ddot{\mathbf{W}}$.
2. Then divide each row of $\ddot{\mathbf{W}}$ by its diagonal entry, giving $\ddot{\mathbf{W}}$'.
3. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

So we have the causal relation:

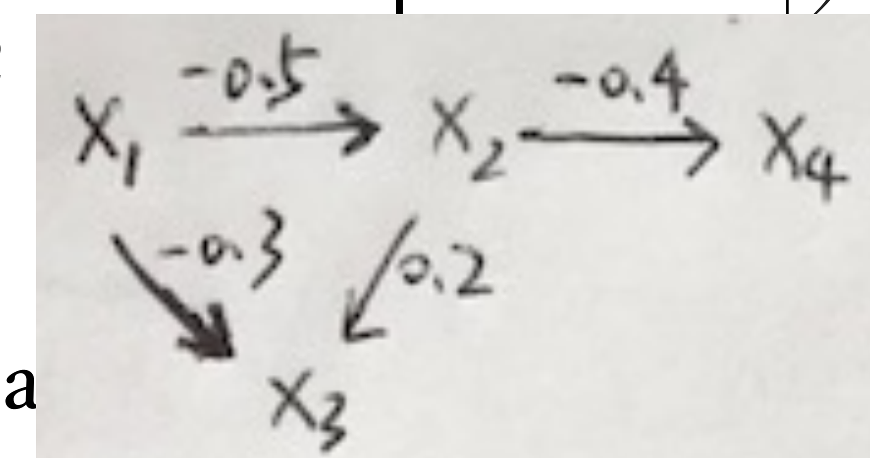


Can You See Causal Relations from $\mathbf{W}$? Example

- ICA gives $\mathbf{Y} = \mathbf{W}\mathbf{X}$ and

$$\mathbf{W} = \begin{bmatrix} 0.6 & -0.4 & 2 & 0 \\ 1.5 & 0 & 0 & 0 \\ 0 & 0.2 & & \\ 1.5 & 3 & & \end{bmatrix}$$

1. First permute the rows of $\mathbf{W}$ to make all diagonal entries non-zero, yielding $\ddot{\mathbf{W}}$.
2. Then divide each row of $\ddot{\mathbf{W}}$
its diagonal entry, giving $\ddot{\mathbf{W}}$'.
$\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

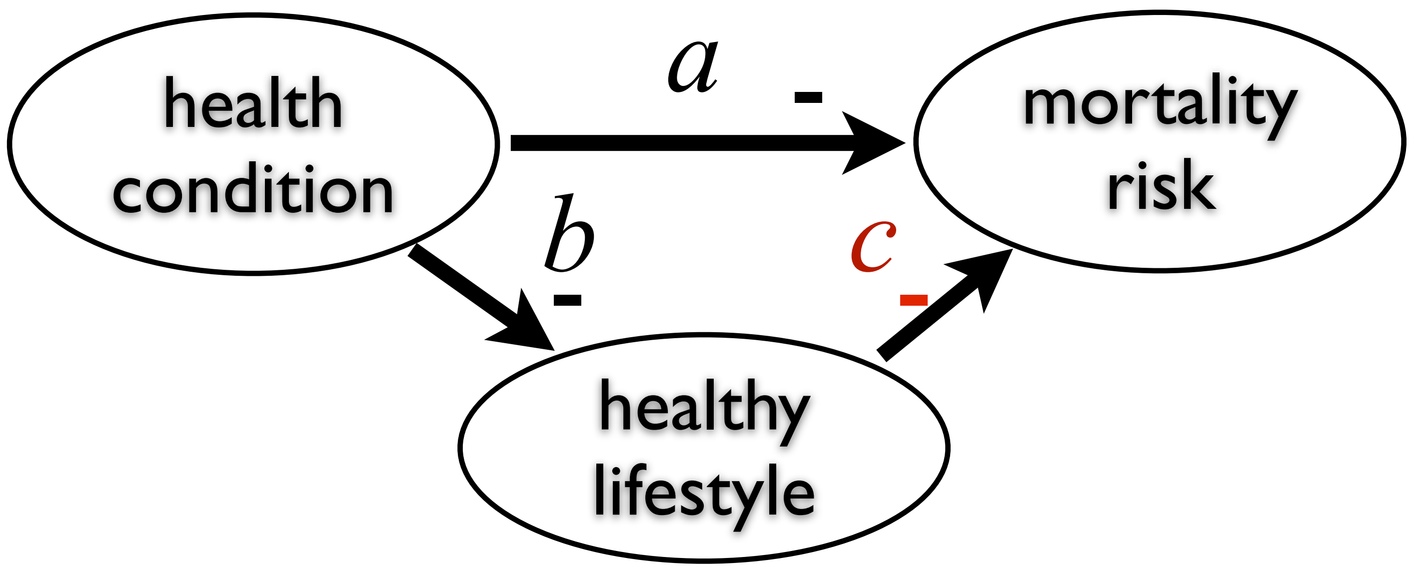


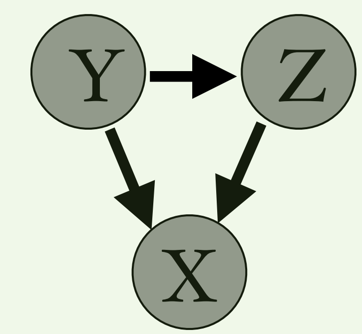
- Can we find the ca

$$\ddot{W} = \begin{pmatrix} 1.5 & 0 & 0 & 0 \\ 1.5 & 3 & 0 & 0 \\ 0.6 & -0.4 & 2 & 0 \\ 0 & 0.2 & 0 & 0.5 \end{pmatrix}, \quad \ddot{W}' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.3 & -0.2 & 1 & 0 \\ 0 & 0.4 & 0 & 1 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -0.5 & 0 & 0 & 0 \\ -0.3 & 0.2 & 0 & 0 \\ 0 & -0.4 & 0 & 0 \end{pmatrix}$$

Faithfulness Assumption Needed?

- One might find independence between health condition & risk of mortality. Why?





Possible to have $Y \perp\!\!\!\perp Z \mid X$?

- E.g., if $a=-bc$, then *health_condition* $\perp\!\!\!\perp$ *mortality_risk*, which cannot by seen from the graph!

- No faithfulness assumption is needed in LiNGAM
  - Minimality (a zero coefficient corresponds to edge absence) is sufficient

Step-by-Step Demo & Application

- Galton family height data
- Result of PC?
- Linear, non-Gaussian methods: let's do causal discovery step by step with 'illust_LiNGAM_Galton.m'

Galton's height data

family	father	mother	Gender	Height
1	78.5	67	0	73.2
1	78.5	67	1	69.2
1	78.5	67	1	69
1	78.5	67	1	69
2	75.5	66.5	0	73.5
2	75.5	66.5	0	72.5
2	75.5	66.5	1	65.5
2	75.5	66.5	1	65.5
3	75	64	0	71
3	75	64	1	68
4	75	64	0	70.5
4	75	64	0	68.5
4	75	64	1	67
4	75	64	1	64.5
...	...	...	...	...

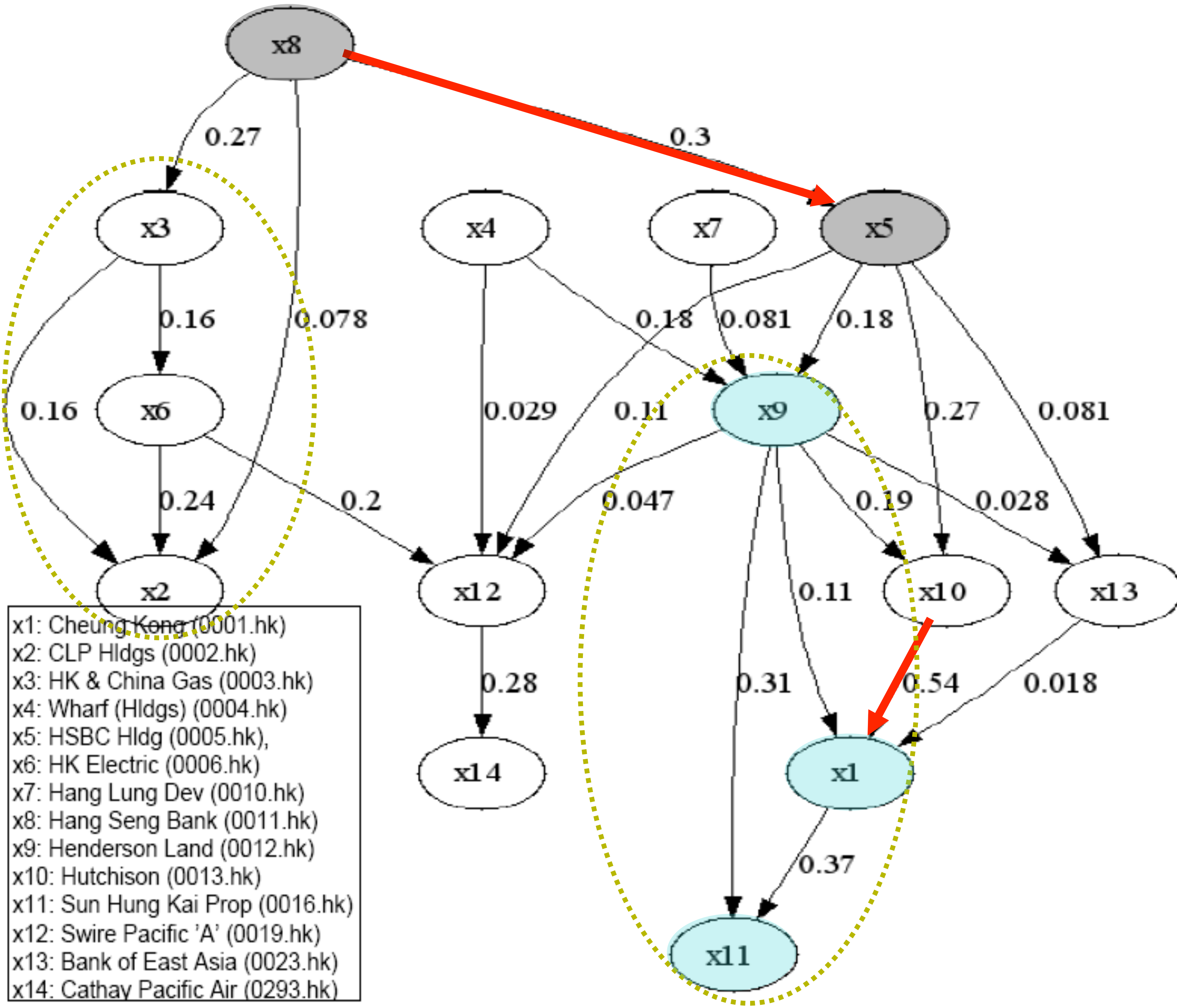
*

Some Estimation Methods for LiNGAM

- ICA-LiNGAM
- ICA with Sparse Connections
- DirectLiNGAM...

Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.
Zhang et al. (2006) ICA with sparse connections: Revisited. Lecture Notes in Computer Science, 5441:195–202, 2009
Shimizu, et al. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 12:1225–1248.

Application: Causal diagram in HK Stock Market (Zhang & Chan, 2006)

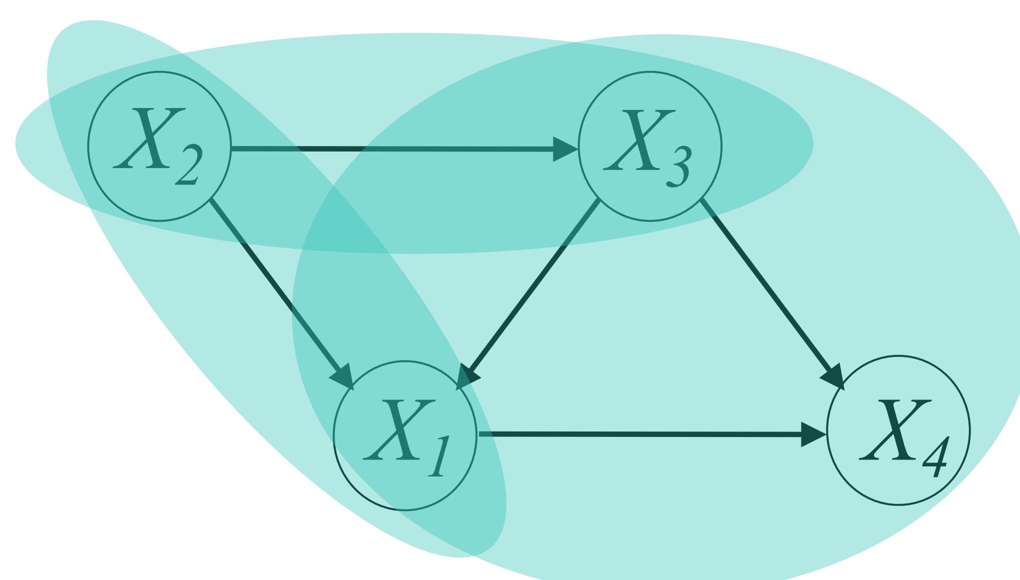


1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.
2. Stocks belonging to the same subindex tend to be connected.
3. Large bank companies (x5 and x8) are the cause of many stocks.
4. Stocks in Property Index (x1, x9, x11) depend on many stocks, while they hardly influence others.

Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), because $(\mathbf{Z}, Y)$ satisfies the IN condition iff
    - All variables in $\mathbf{Z}$ are causally earlier than Y &
    - the common cause for Y and each variable in $\mathbf{Z}$, if there is any, is in $\mathbf{Z}$.
- Can then estimate the LiNGAM (the DirectLiNGAM algorithm)



*

Independence Test / Dependence Measure

- Measure: mutual information $MI(Y_1, Y_2) \geq 0$ with equality holds iff $Y_1 \perp\!\!\!\perp Y_2$

- Statistical test for independence

  - $Y_1 \perp\!\!\!\perp Y_2$ if and only if all functions of them are uncorrelated

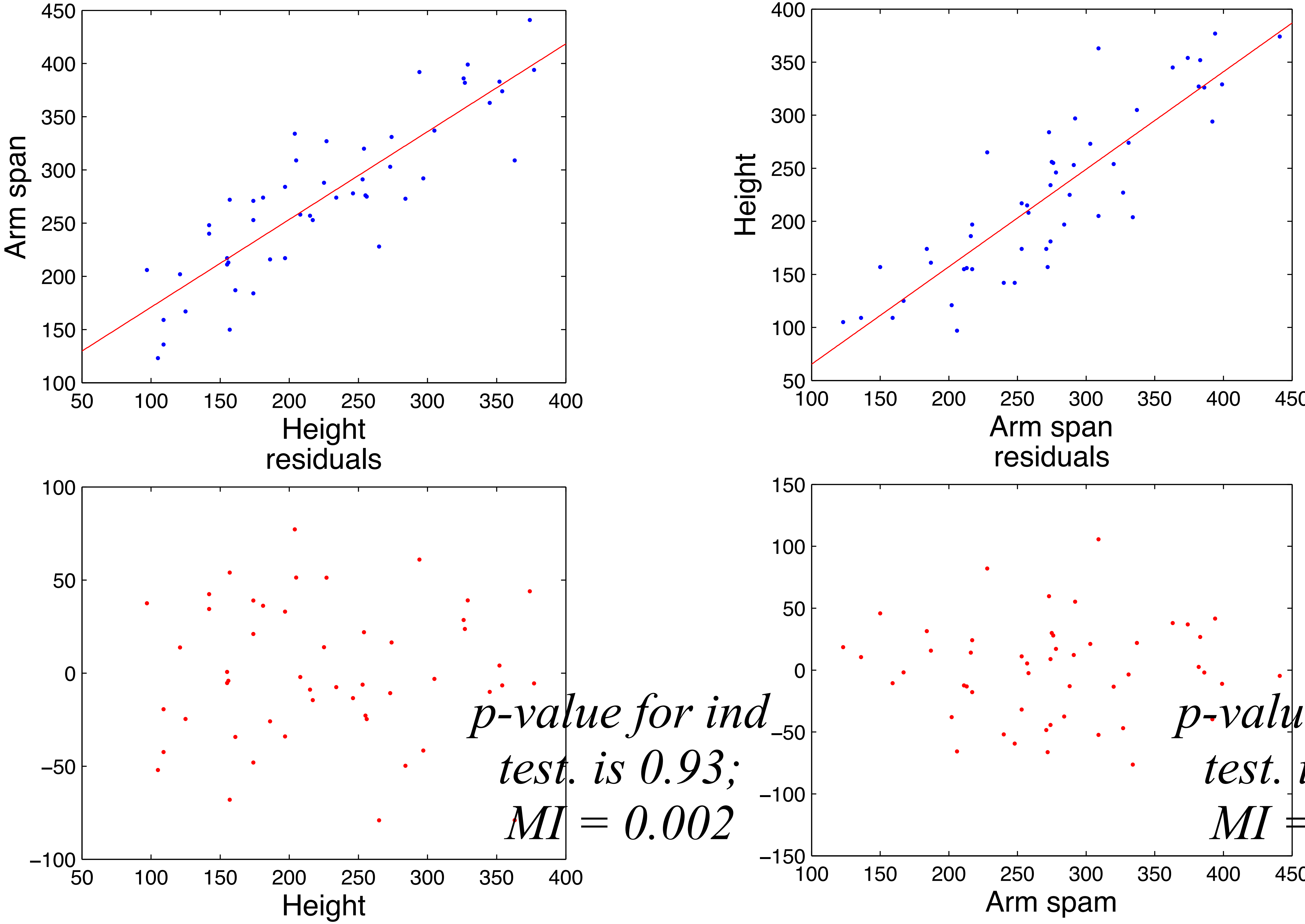
  - The functional space can be narrowed down to the reproducing kernel Hilbert space

  - HSIC independence test; Kernel-based (conditional) independence test; other tests also exist

Gretton et al. (2008). A kernel statistical test of independence. In Advances in Neural Information Processing Systems, 585–592.

Zhang et al. (2011). Kernel-based conditional independence test and application in causal discovery. In Proc. UAI, 804–813.

Real Examples: By Checking Independence in Both Directions

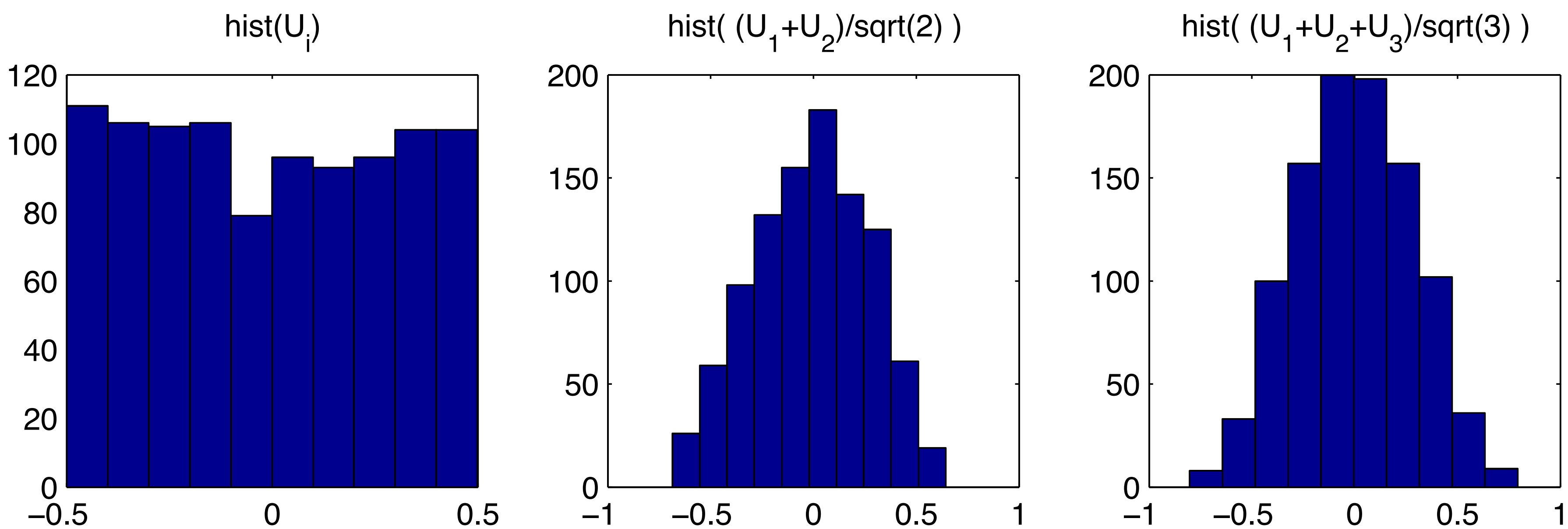


p-value for ind test. is 0.93; MI = 0.002

p-value for ind test. is 0.53; MI = 0.042

Why Was Gaussianity Widely Used?

- Central limit theorem: An illustration



- "Simplicity" of the form; completely characterized by mean and covariance

- Marginal and conditionals are also Gaussian

- Has maximum entropy, given values of the mean and the covariance matrix

E. T. Jaynes. Probability Theory: The Logic of Science. 1994. Chapter 7.

Gaussianity or Non-Gaussianity?

- Non-Gaussianity is actually ubiquitous
  - Linear closure property of Gaussian distribution: If the sum of any finite independent variables is Gaussian, then all summands must be Gaussian (Cramér, 1936)
  - Gaussian distribution is "special" in the linear case
- Practical issue: How non-Gaussian they are?

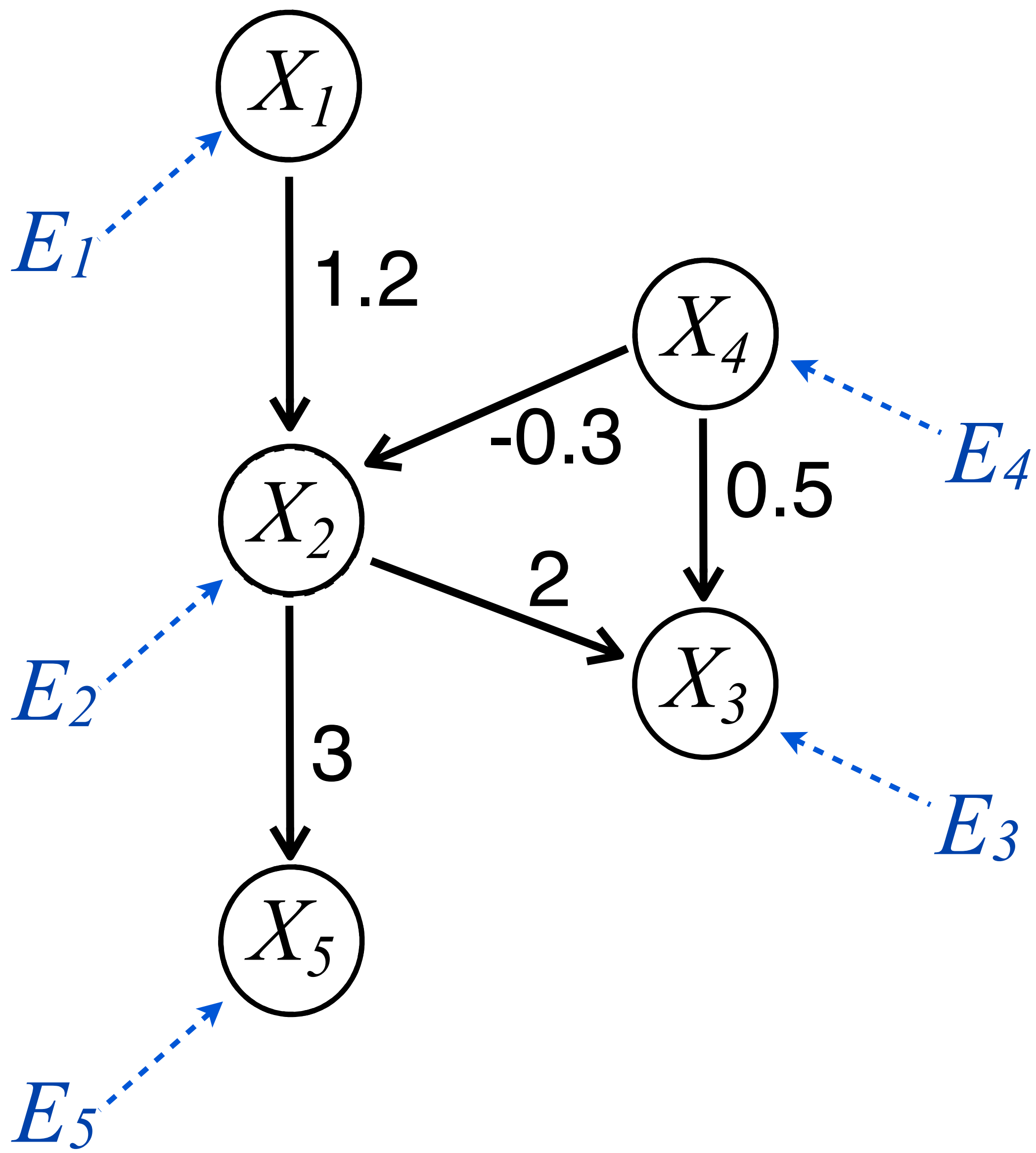
Practical Issues in Causal Discovery…

- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19; Xie et al., NeurIPS'20); latent 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 ?



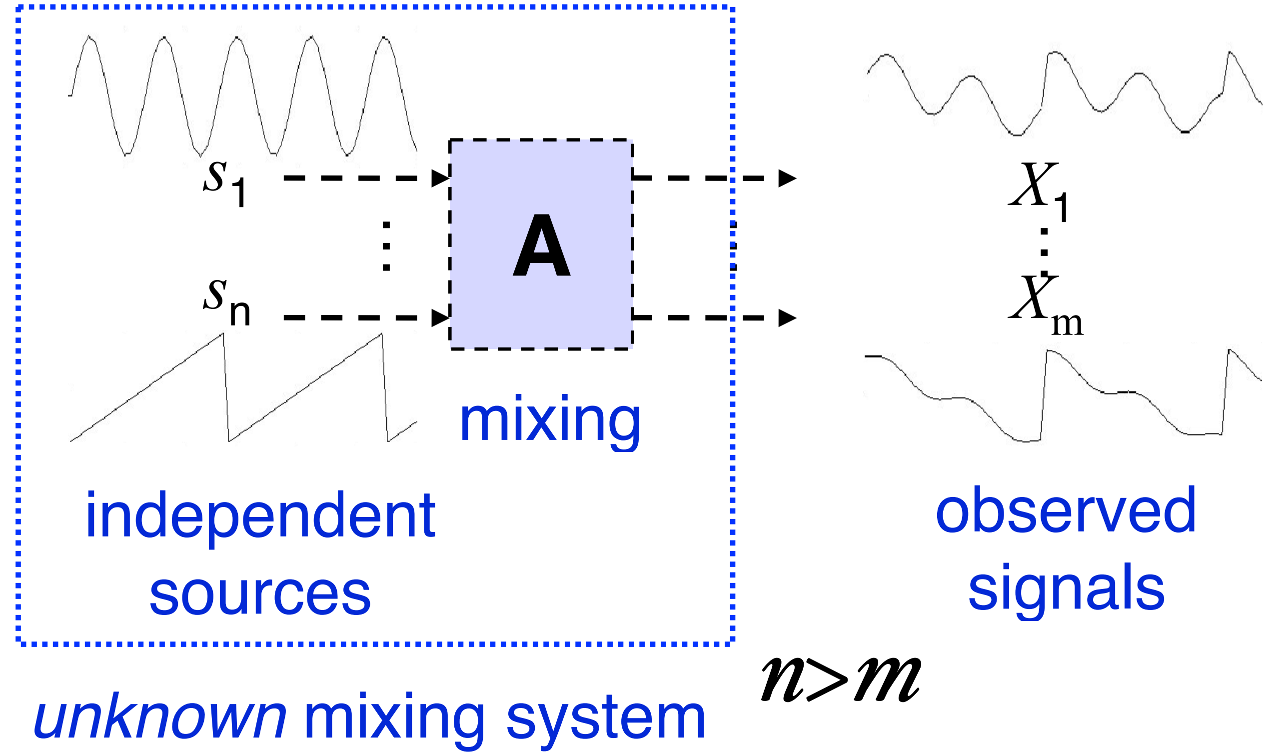
X_1?

X_4?

X_2?

X_5?

Identifiability of Overcomplete ICA



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

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

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

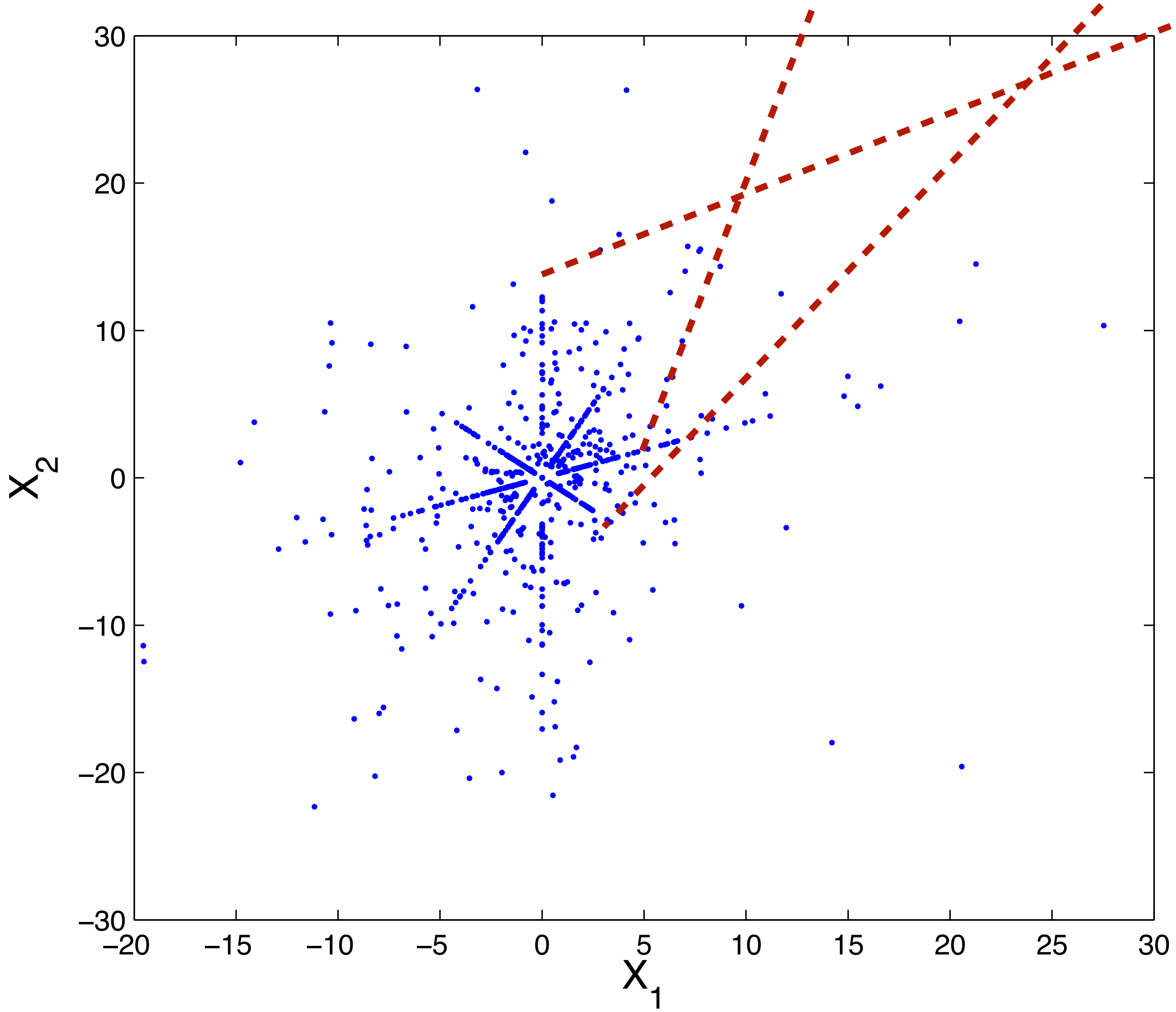
- all S_i are non-Gaussian, or
- $\mathbf{A}$ is of full column rank and at most one of S_i is Gaussian.

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.

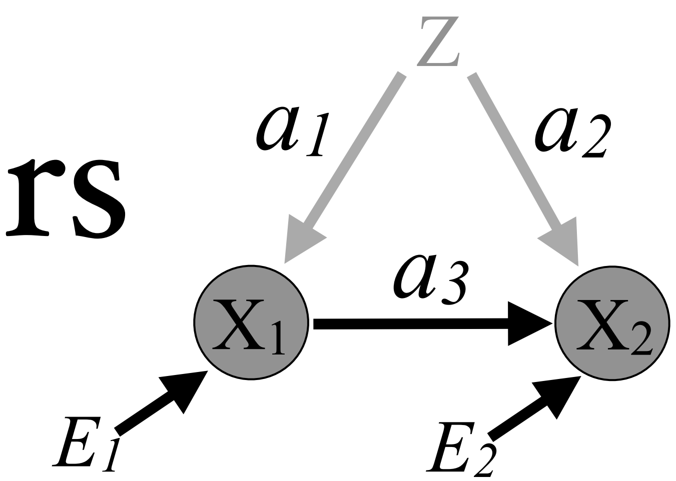
Overcomplete ICA: Illustration

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



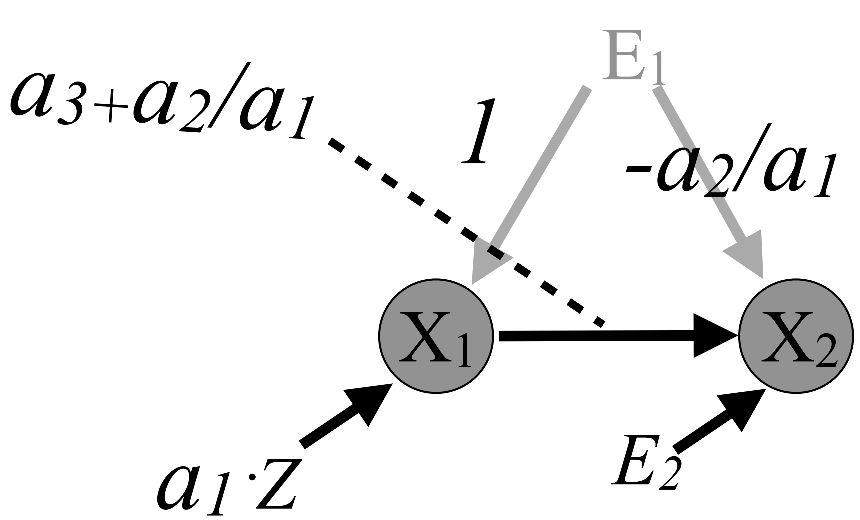
What if they are Gaussian?

Discussions I: Confounders



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

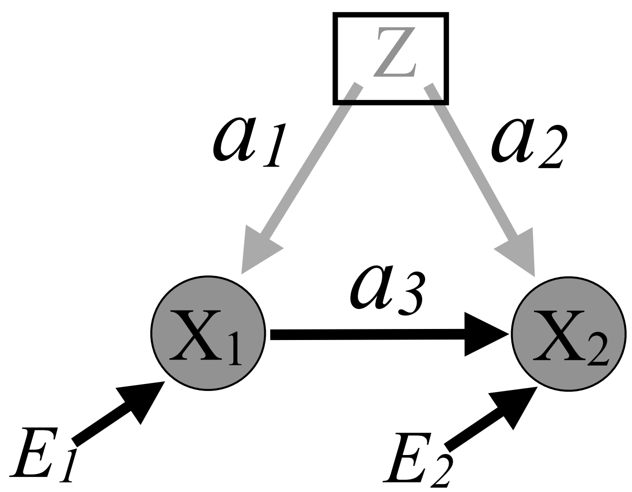
- Can we see the causal direction ?
- Can we determine a_3 ? a_1 and a_2 ?
- Observationally equivalent model:



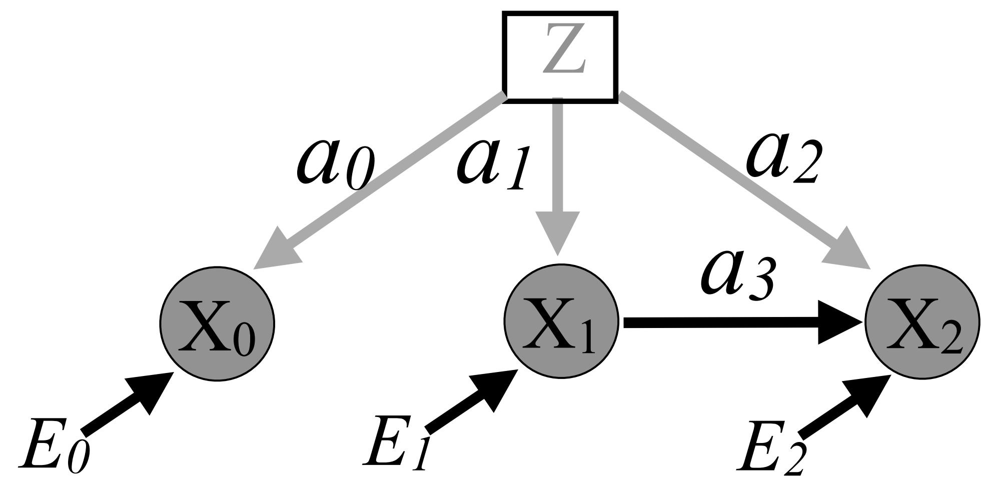
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ \left(a_3 + \frac{a_2}{a_1}\right) + \frac{-a_2}{a_1} & 1 & \left(a_3 + \frac{a_2}{a_1}\right) \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ a_1 Z \end{bmatrix}$$

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

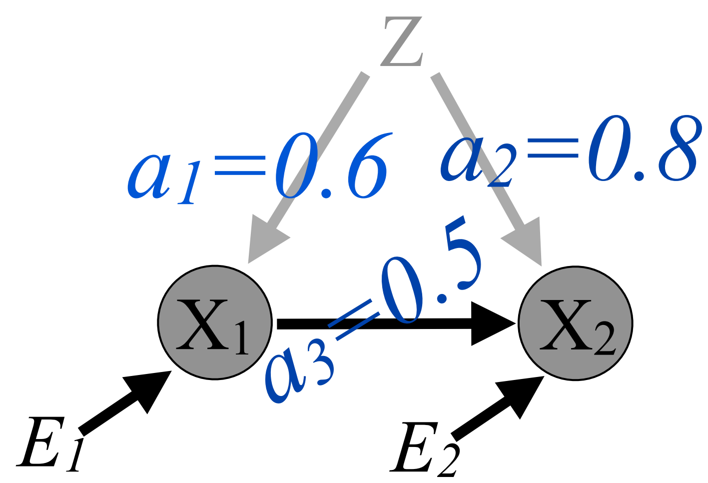
Example 1: $$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & a_1 \\ a_3 & 1 & a_1 a_3 + a_2 \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ Z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ a_3 & 1 & a_3 + \frac{a_2}{a_1} \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ a_1 Z \end{bmatrix}$$

Two possible solutions

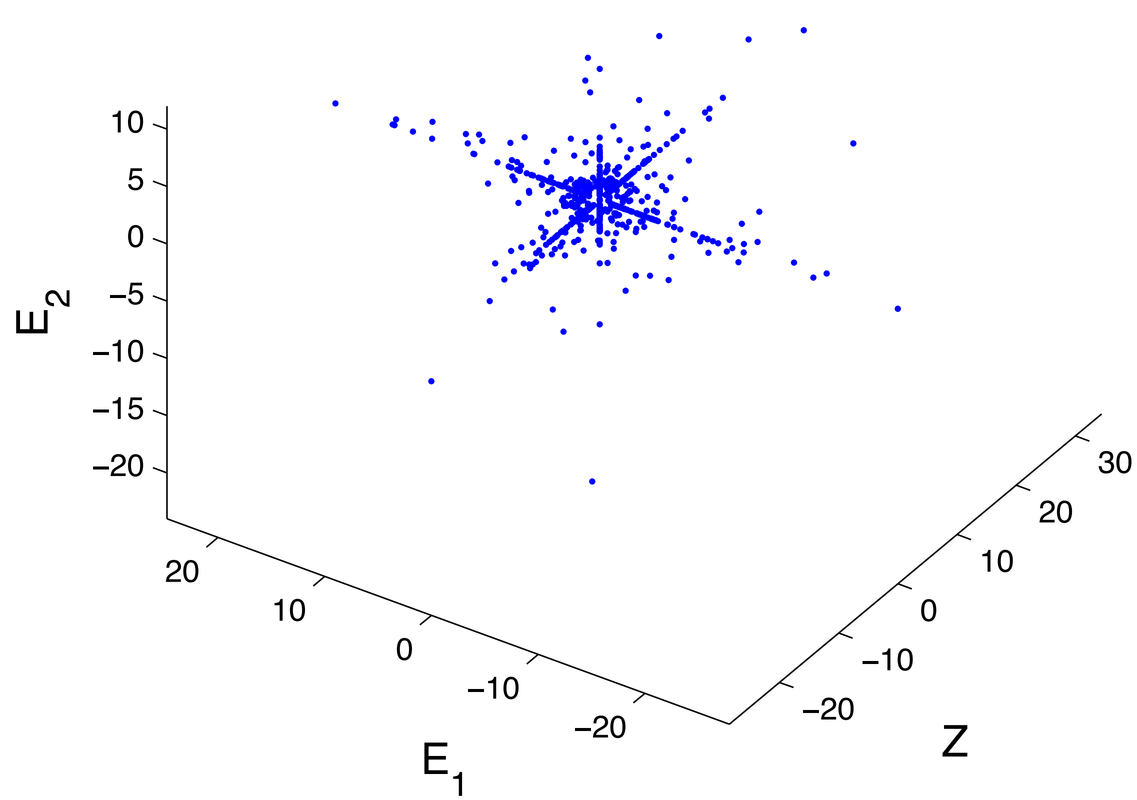
Example 2: $$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & a_0 \\ 0 & 1 & 0 & a_1 \\ 0 & a_3 & 1 & a_1 a_3 + a_2 \end{bmatrix} \cdot \begin{bmatrix} E_0 \\ E_1 \\ E_2 \\ Z \end{bmatrix}$$

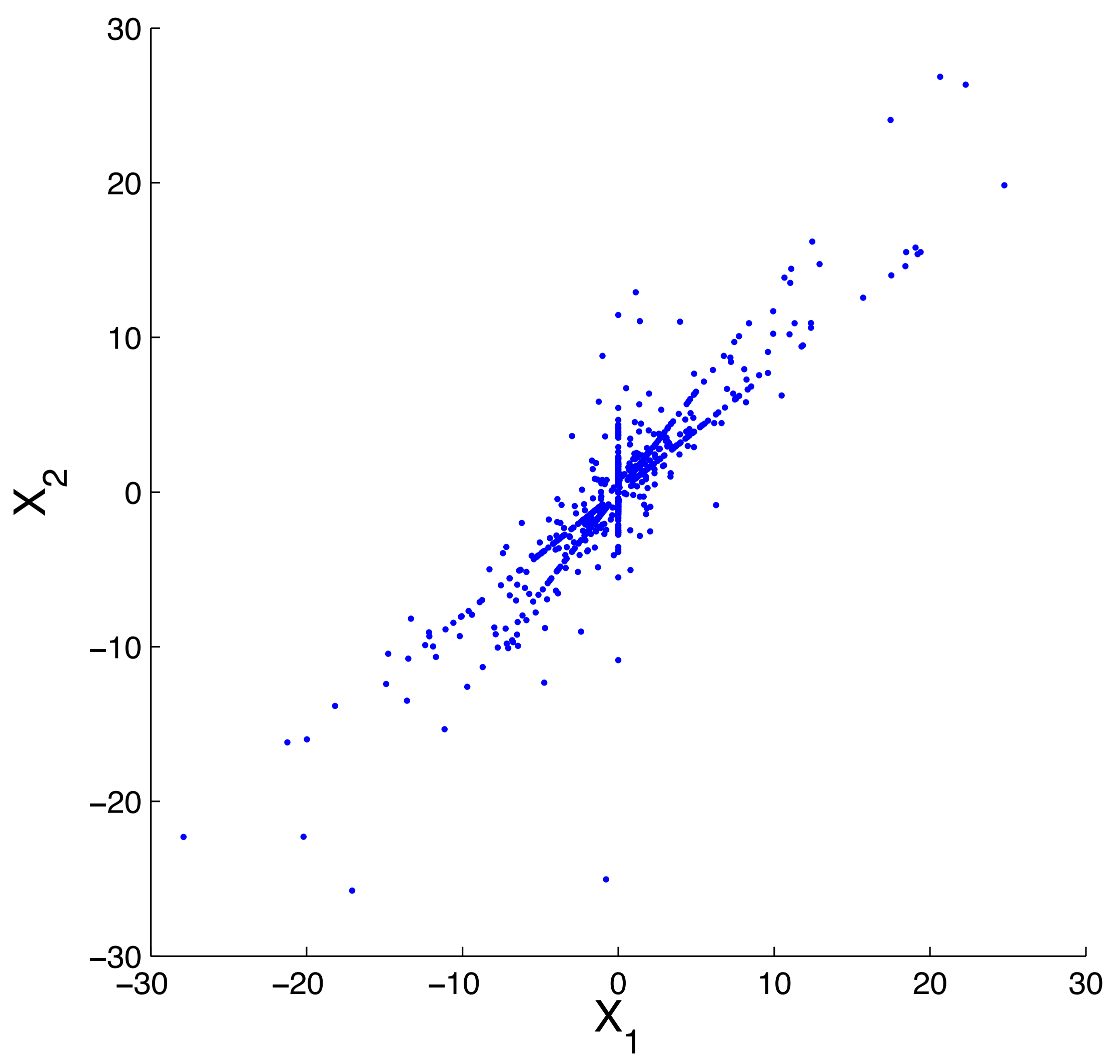
a_3 identifiable!

Confounders: Example

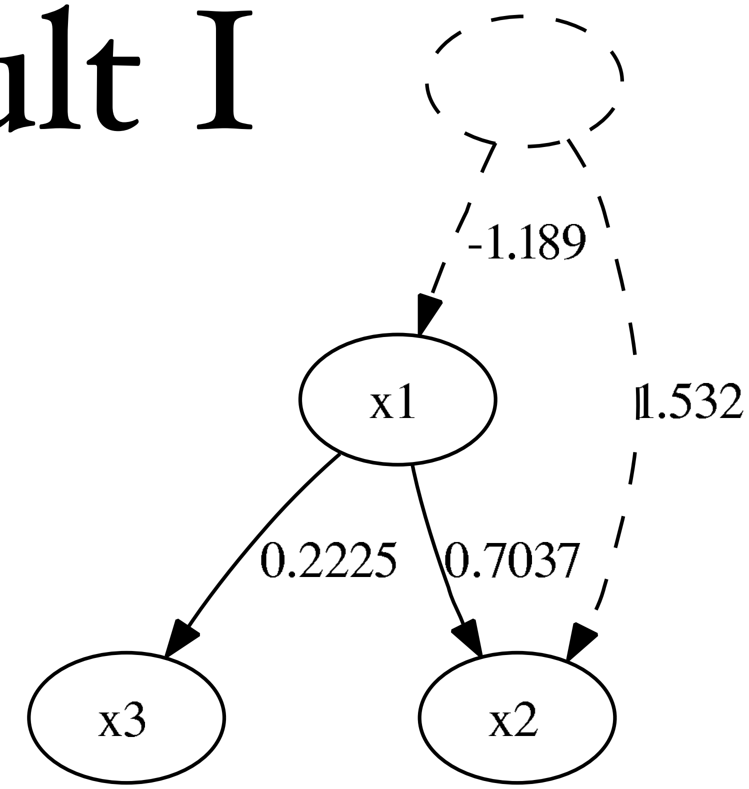


$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & a_1 \\ a_3 & 1 & a_1 a_3 + a_2 \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ Z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ a_3 & 1 & a_3 + \frac{a_2}{a_1} \end{bmatrix} \cdot \begin{bmatrix} E_1 \\ E_2 \\ a_1 Z \end{bmatrix}$$

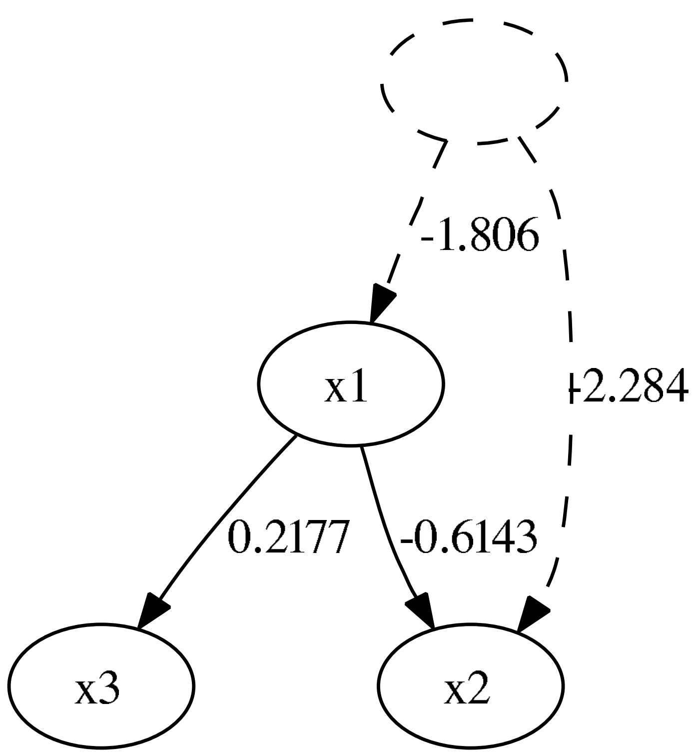


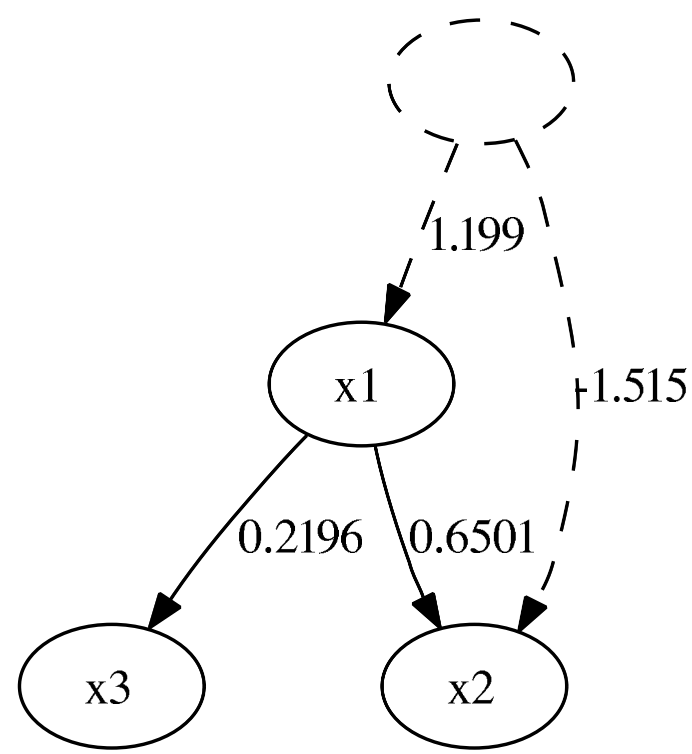


Some Simulation Result I



- Simulate 2500 data points with non-Gaussian noise using this model:
- Output of the algorithm:

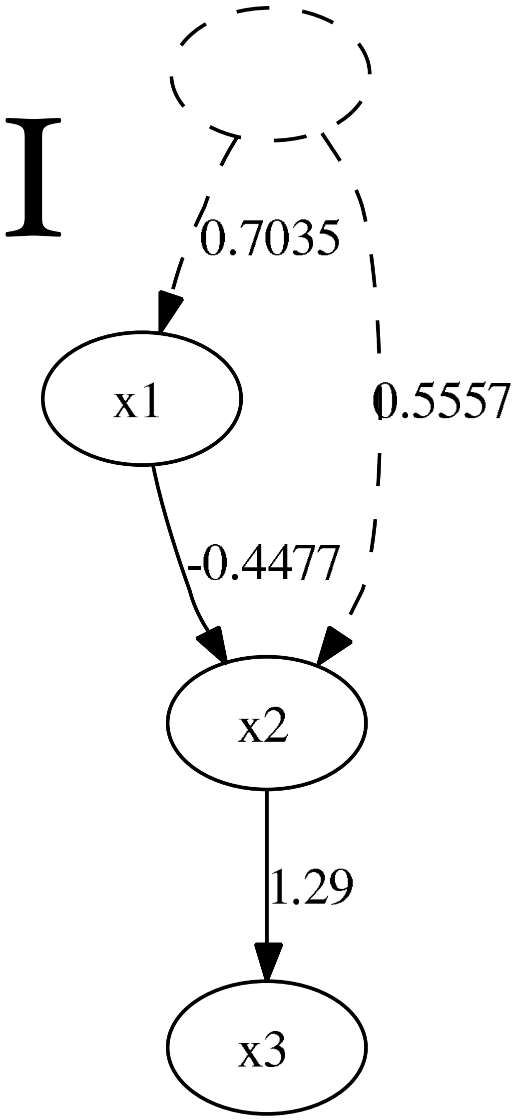




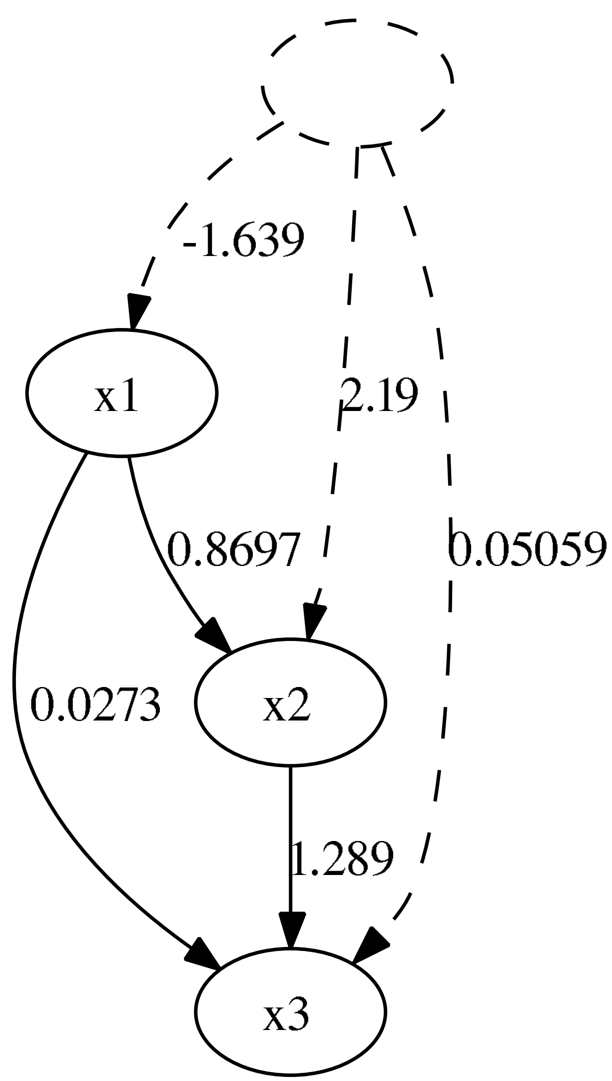
Hoyer et al. (2008). Estimation of causal effects using linear nonGaussian causal models with hidden variables. International Journal of Approximate Reasoning, 49(2):362– 378.

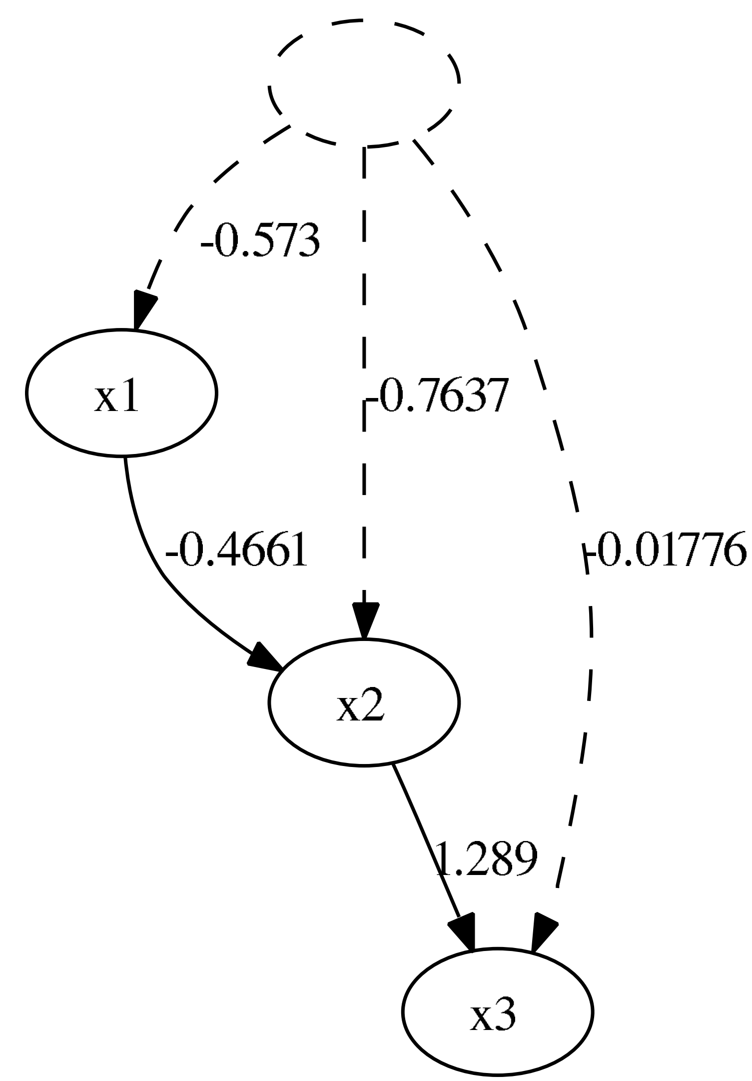
Some Simulation Result II

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



- Output of the algorithm:

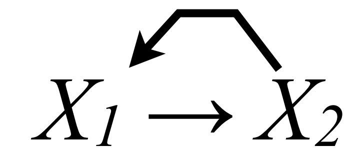




With Cycles

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

Discussion II: Feedback

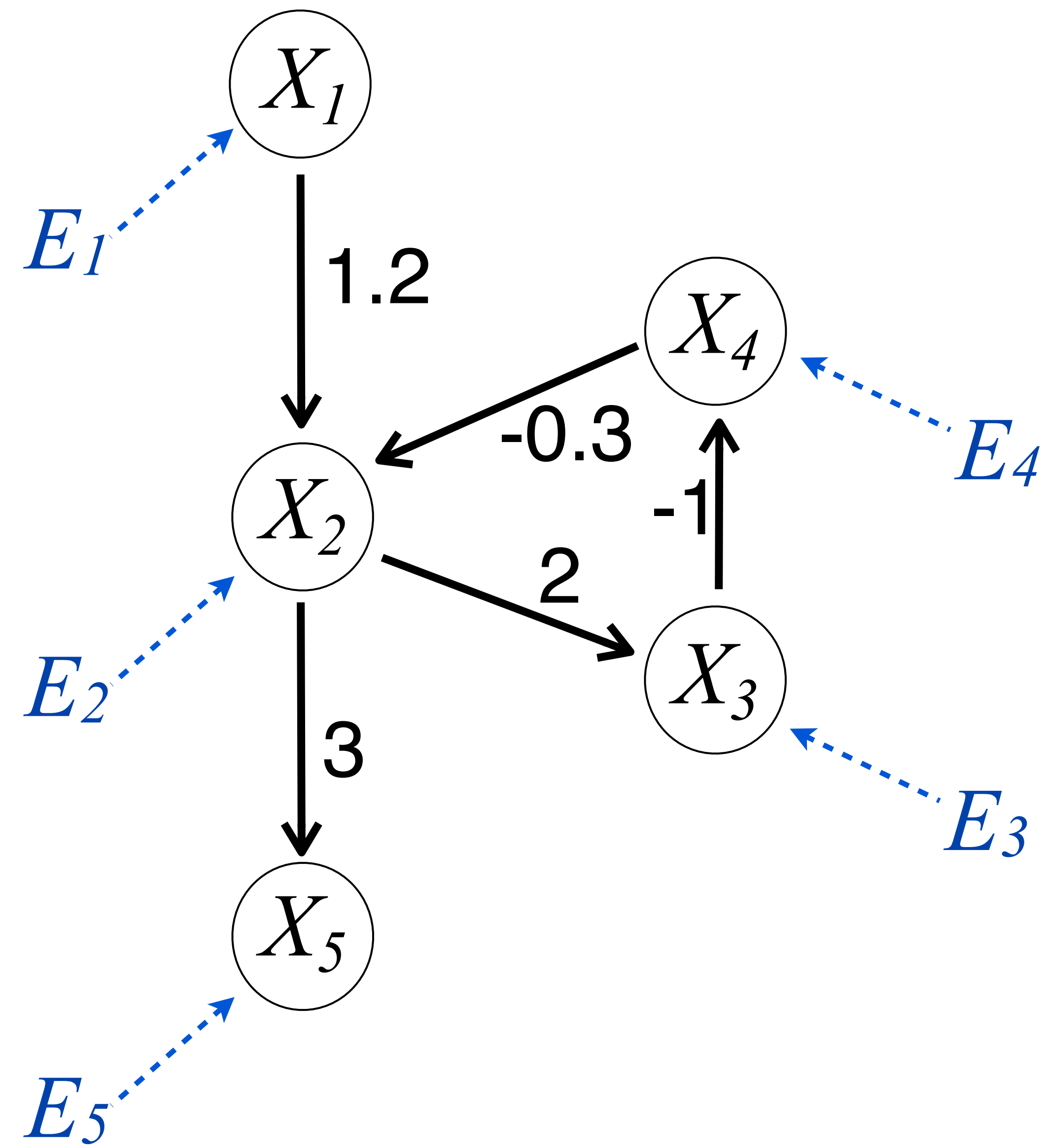


- Causal relations may have cycles; Consider an example

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

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

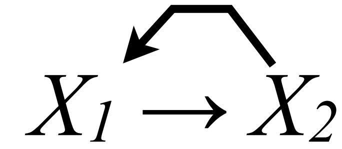
$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & 0 & -0.3 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix}$$



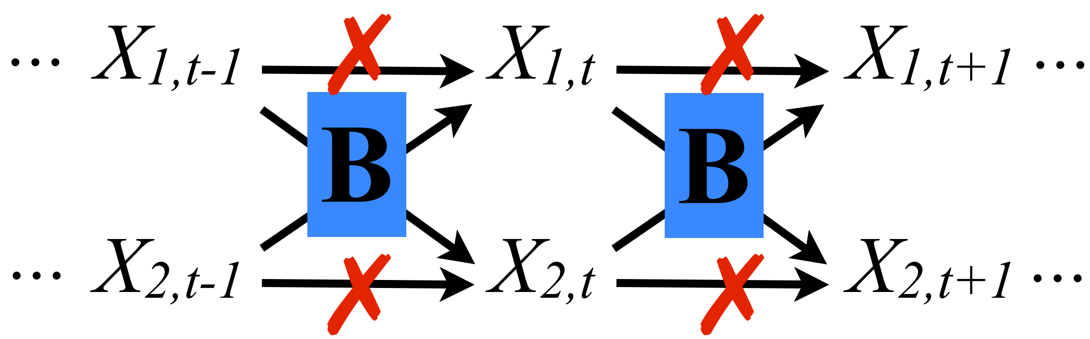
Lacerda, Spirtes, Ramsey and Hoyer (2008). Discovering cyclic causal models by independent component analysis. In Proc. UAI.

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?



- 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

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

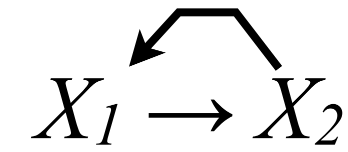
  - On temporally aggregated data

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

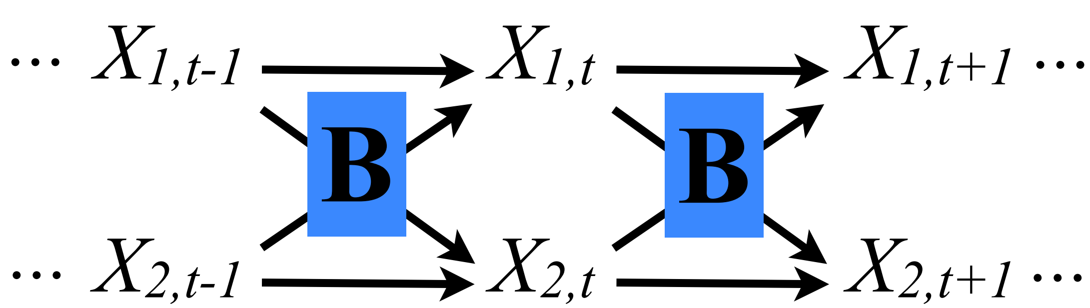
$$\frac{1}{L}\sum_{k=1}^{L} \tilde{\mathbf{X}}_{t+k} = \mathbf{B}\frac{1}{L}\sum_{k=1}^{L} \tilde{\mathbf{X}}_{t+k-1} + \frac{1}{L}\sum_{k=1}^{L} \tilde{\mathbf{E}}_{t+k}.$$

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:

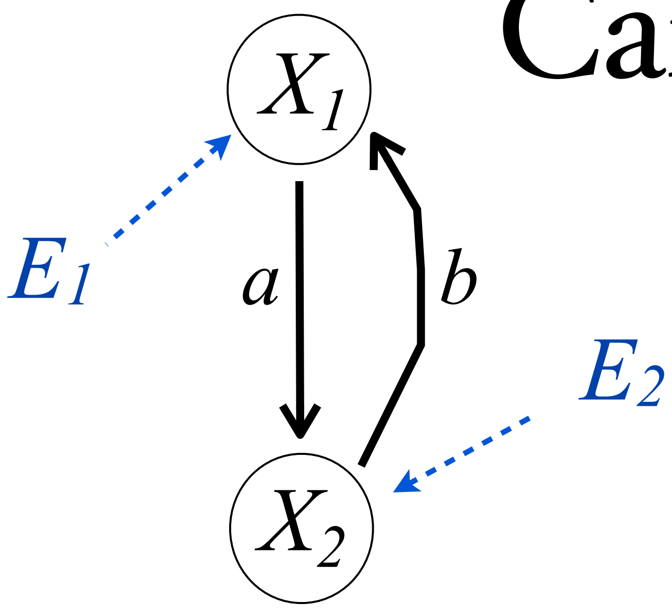
$$\text{price}_t = b_1 \cdot \text{price}_{t-1} + b_2 \cdot \text{demand}_{t-1} + E_1$$
$$\text{demand}_t = b_3 \cdot \text{price}_{t-1} + b_4 \cdot \text{demand}_{t-1} + E_2$$

  - 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,

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

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

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

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

- **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...

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

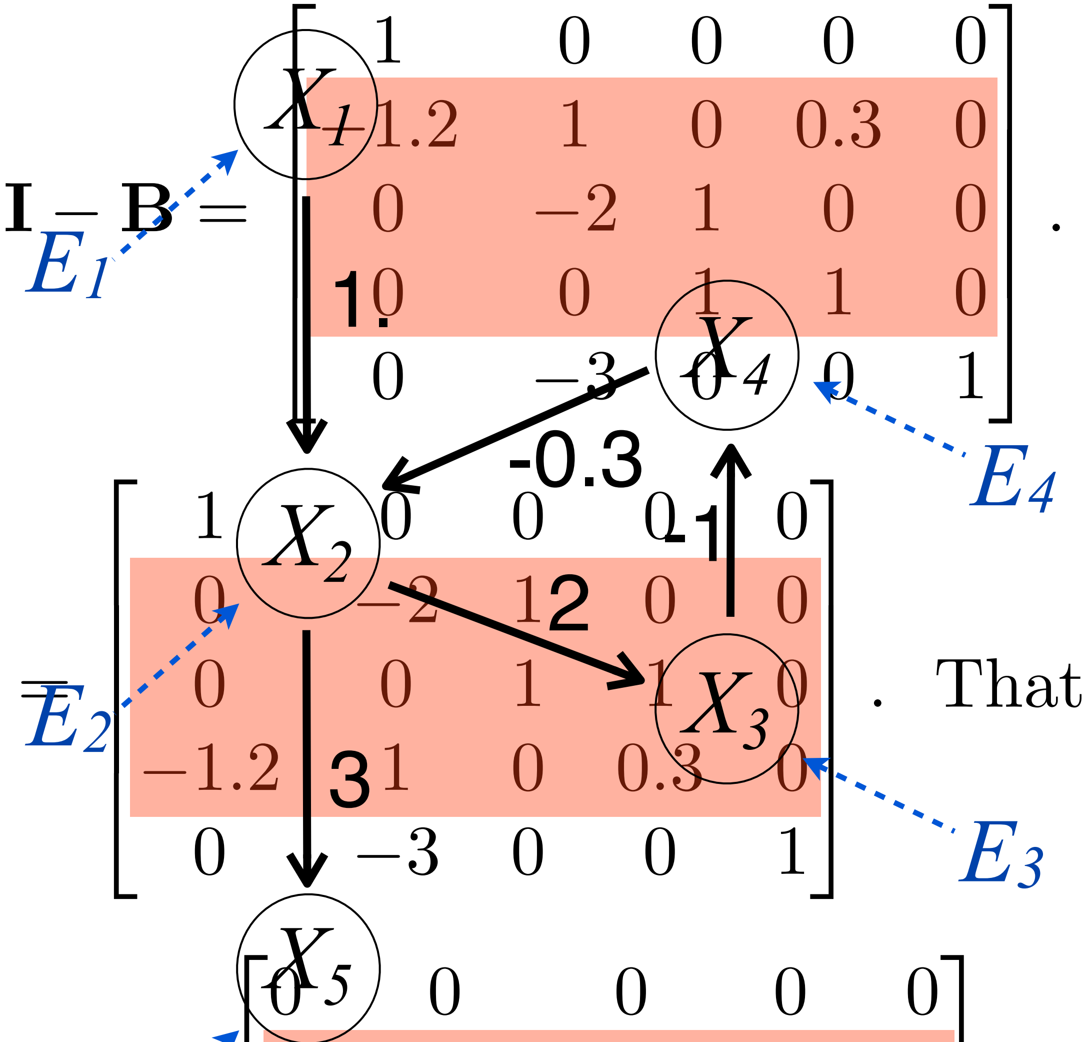
Or in matrix form, $\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$, where

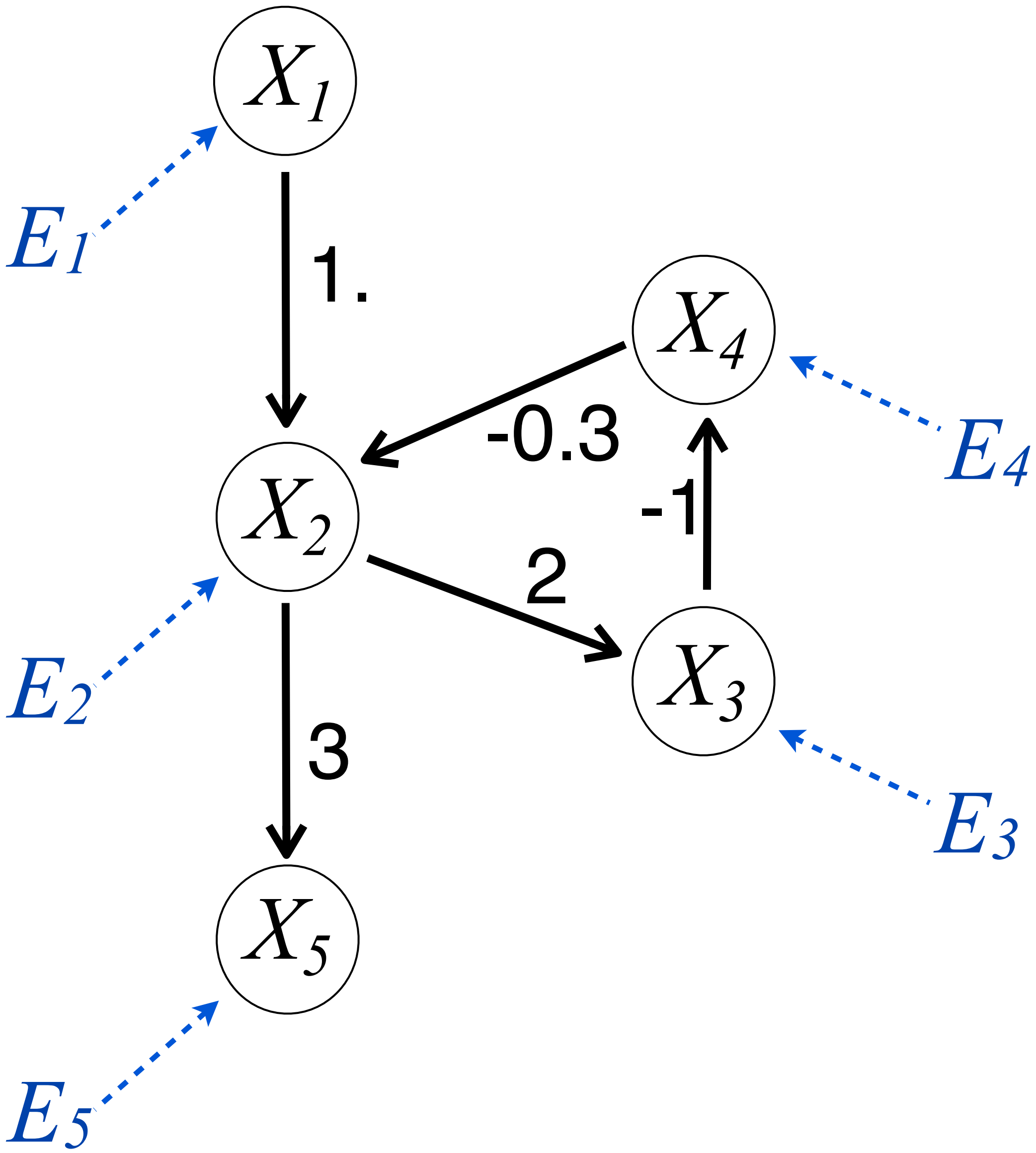
$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & 0 & -0.3 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix}$$

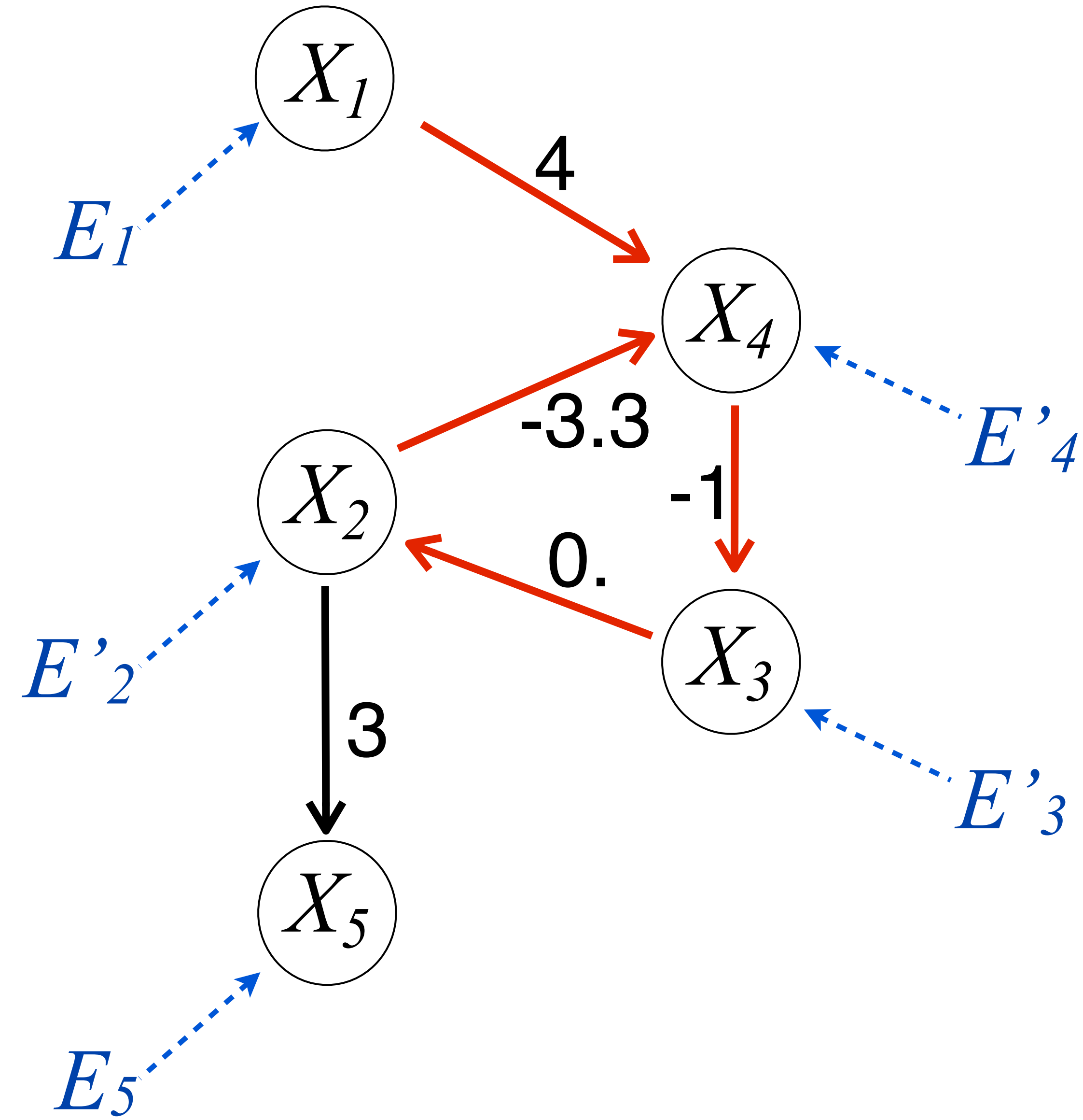
$$\mathbf{I} - \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1.2 & 1 & 0 & 0.3 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & -3 & 0 & 0 & 1 \end{bmatrix}.$$

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

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





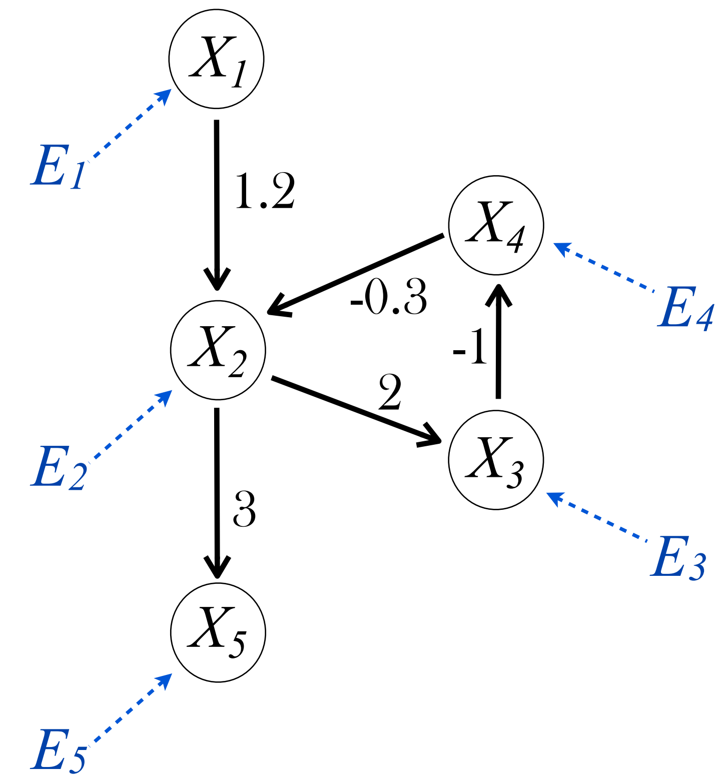


$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & 0 & -0.3 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \end{bmatrix}$$

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

Some Simulation Result

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



- Output of the algorithm:

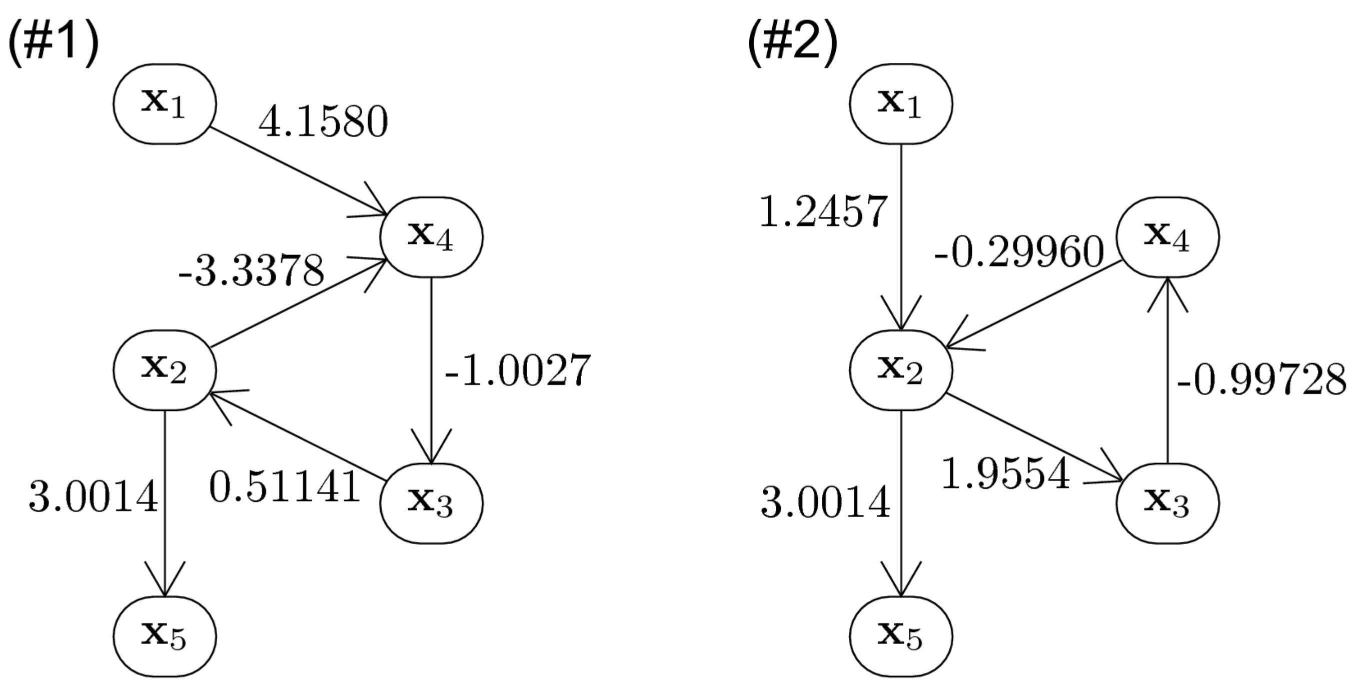


Fig. 3: The output of LiNG-D: Candidate #1 and Candidate #2

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 $\mathbf{Y} = \mathbf{WX}$?
  - 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?