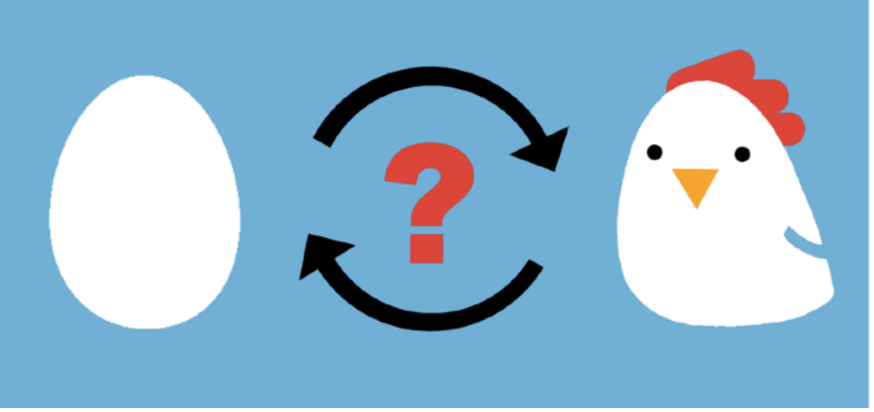


Causality and Machine Learning *(80-816/516)*

Classes 11 & 12 (Feb 18 & 20, 2025)

Linear, Non-Gaussian Causal Models for Causal Discovery (After 2005)

Instructor:

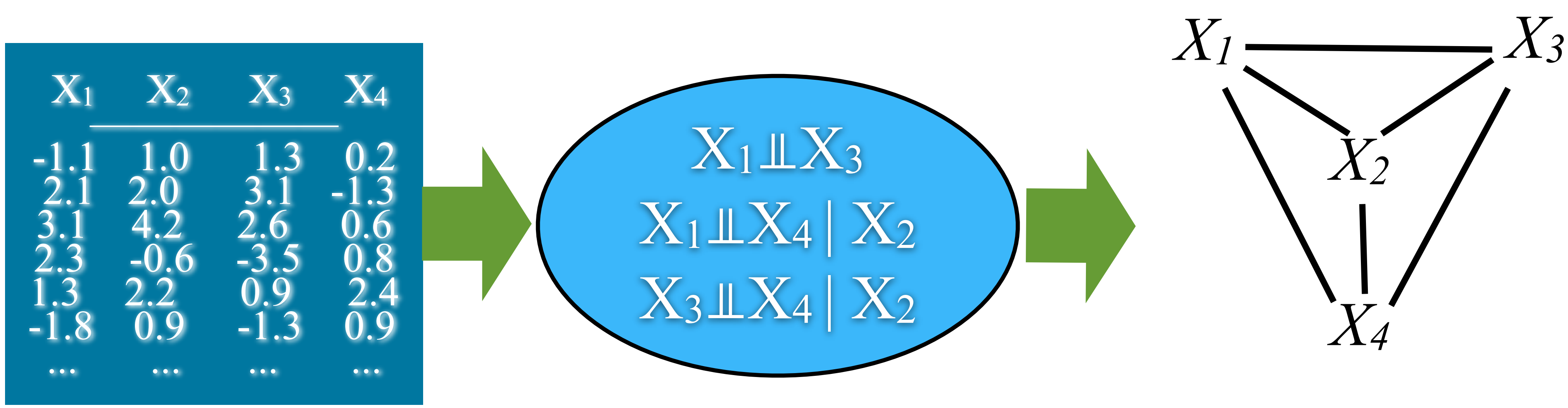
Kun Zhang (kunz1@cmu.edu)

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

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

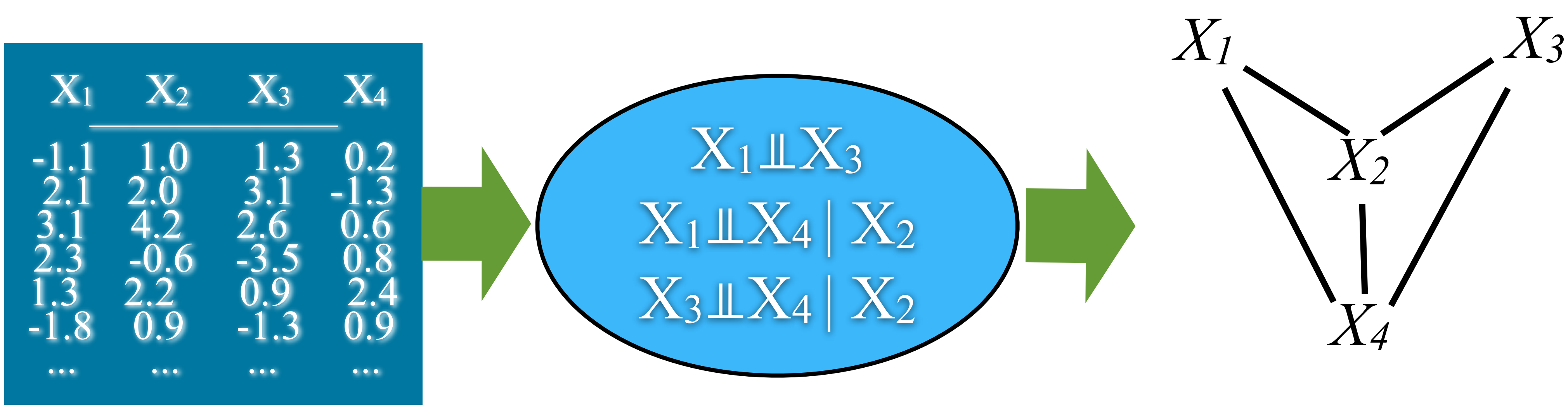
PC Assumes No Confounder

- Make use of conditional independence constraints
- Rely on causal Markov condition + faithfulness assumption
- Output represented by a pattern (CPDAG)



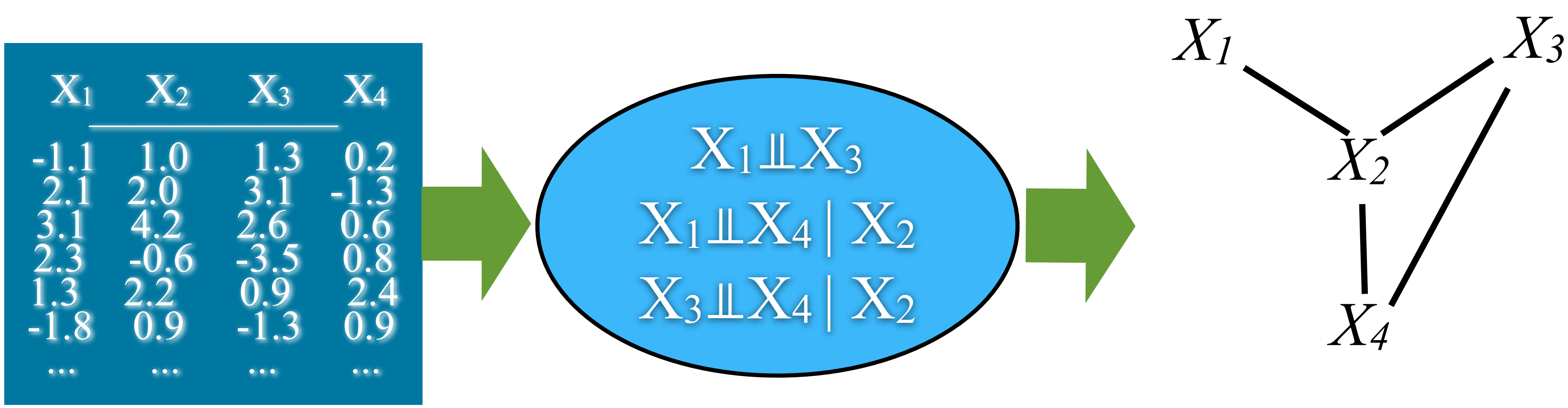
PC Assumes No Confounder

- Make use of conditional independence constraints
- Rely on causal Markov condition + faithfulness assumption
- Output represented by a pattern (CPDAG)



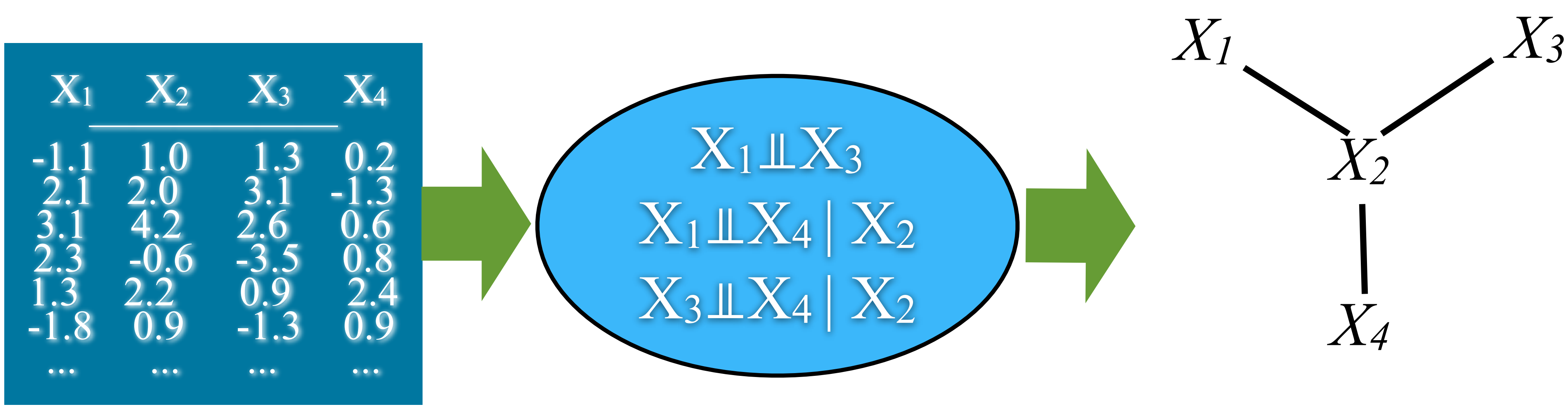
PC Assumes No Confounder

- Make use of conditional independence constraints
- Rely on causal Markov condition + faithfulness assumption
- Output represented by a pattern (CPDAG)



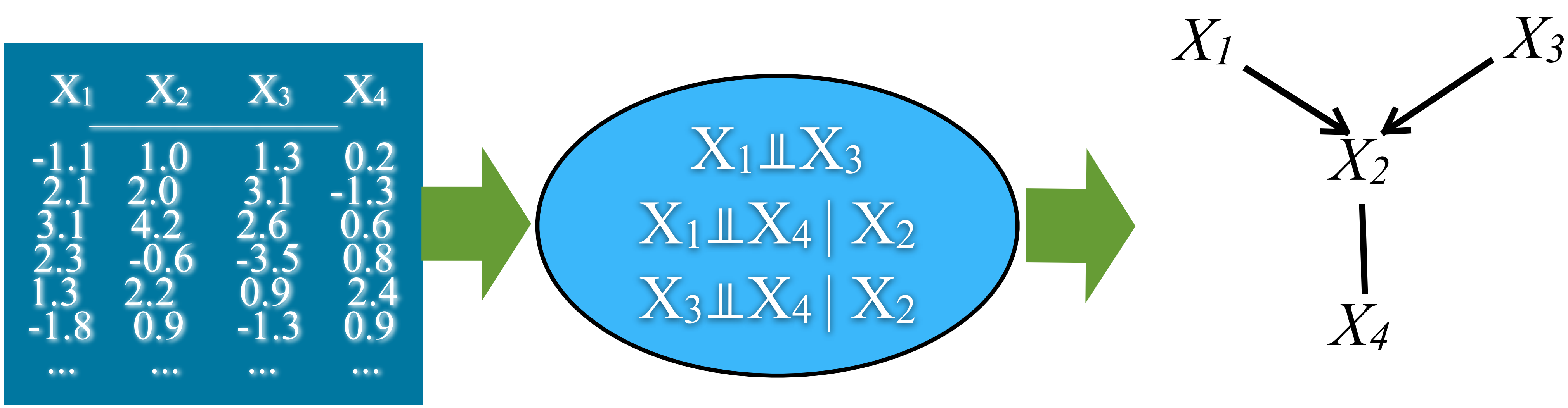
PC Assumes No Confounder

- Make use of conditional independence constraints
- Rely on causal Markov condition + faithfulness assumption
- Output represented by a pattern (CPDAG)



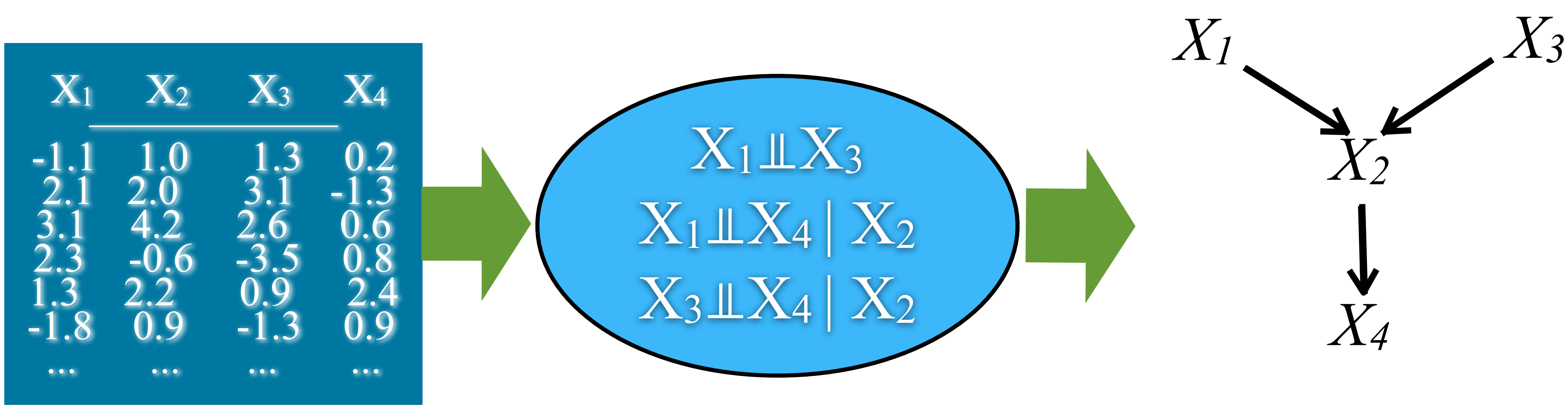
PC Assumes No Confounder

- Make use of conditional independence constraints
- Rely on causal Markov condition + faithfulness assumption
- Output represented by a pattern (CPDAG)



PC Assumes No Confounder

- Make use of conditional independence constraints
- Rely on causal Markov condition + faithfulness assumption
- Output represented by a pattern (CPDAG)



FCI Allows Confounders

Example I

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

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

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

FCI Allows Confounders

Example I

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

$X_1 \perp\!\!\!\perp X_4 \,|\, X_3;$

$X_2 \perp\!\!\!\perp X_4 \,|\, X_3.$

Possible to have confounders behind X_3 and X_4?

FCI Allows Confounders

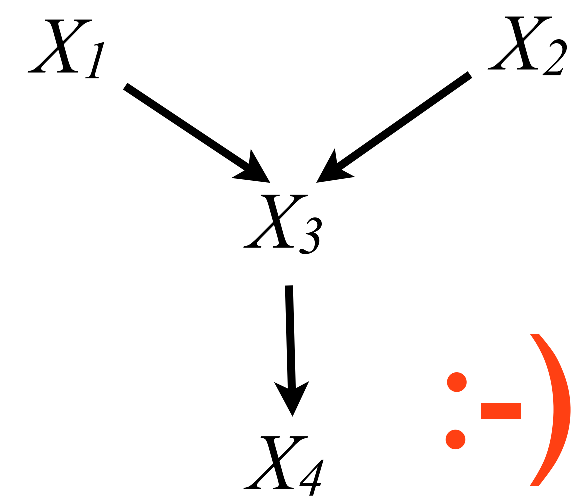
Example I

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

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

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

Possible to have confounders behind X_3 and X_4?



FCI Allows Confounders

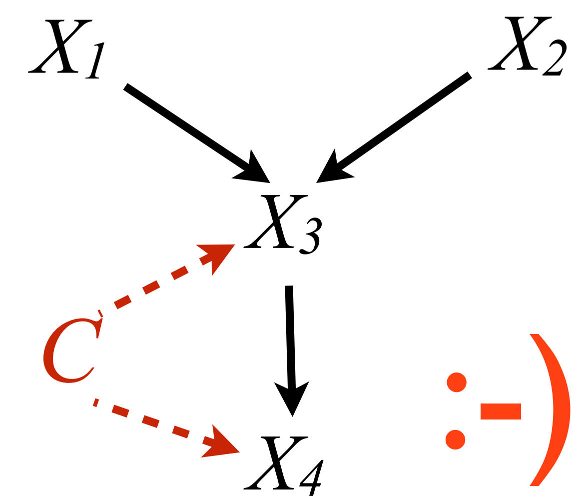
Example I

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

$X_1 \perp\!\!\!\perp X_4 \,|\, X_3;$

$X_2 \perp\!\!\!\perp X_4 \,|\, X_3.$

Possible to have confounders behind X_3 and X_4?



FCI Allows Confounders

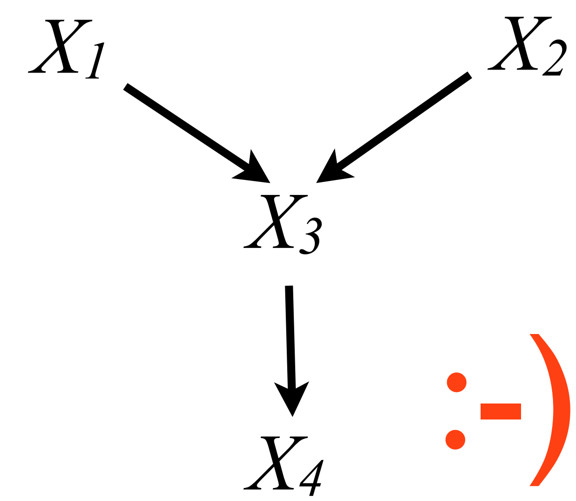
Example I

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

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

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

Possible to have confounders behind X_3 and X_4?



FCI Allows Confounders

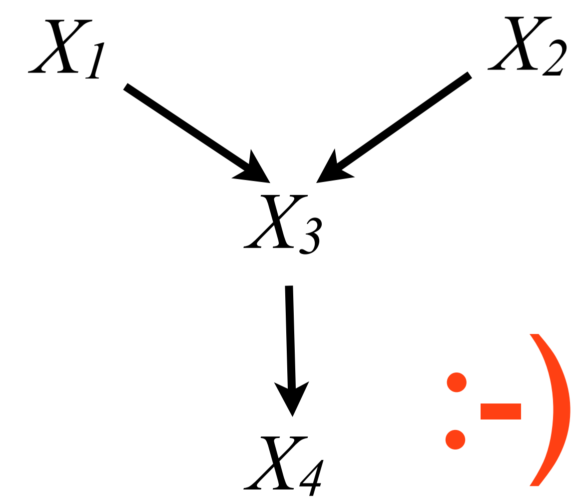
Example I

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

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

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

Possible to have confounders behind X_3 and X_4?



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

FCI Allows Confounders

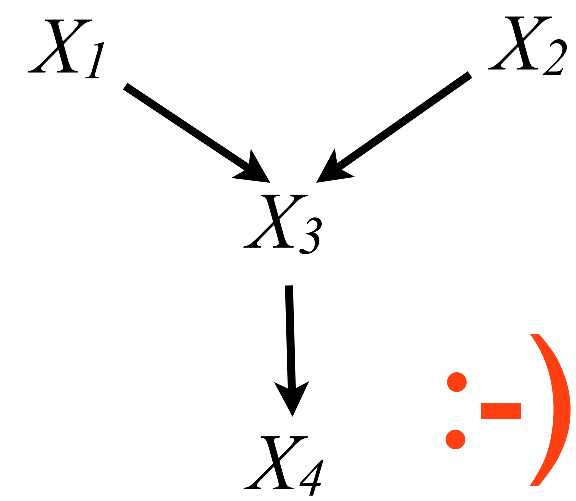
Example I

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

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

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

Possible to have confounders behind X_3 and X_4?



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

Example II

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

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

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

FCI Allows Confounders

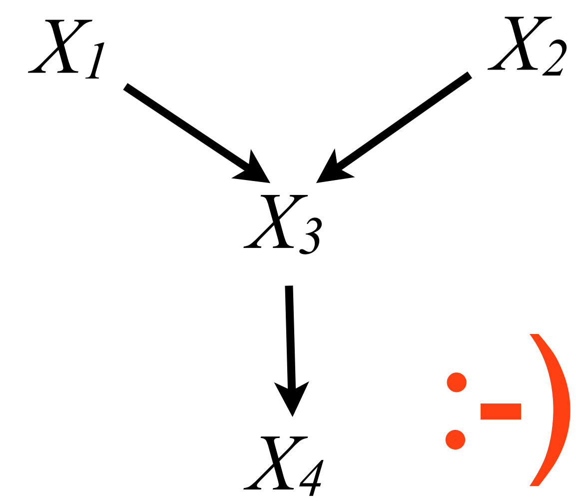
Example I

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

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

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

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?

FCI Allows Confounders

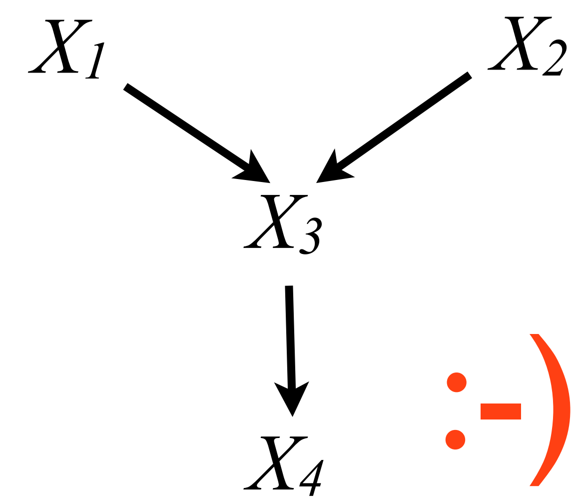
Example I

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

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

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

Possible to have confounders behind X_3 and X_4?



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

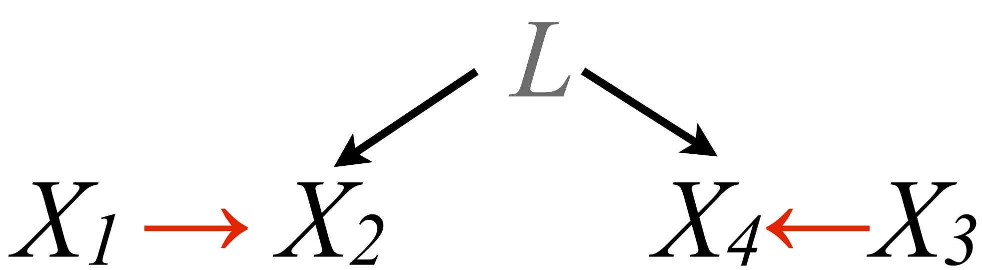
Example II

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

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

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

Are there confounders behind X_2 and X_4?



FCI Allows Confounders

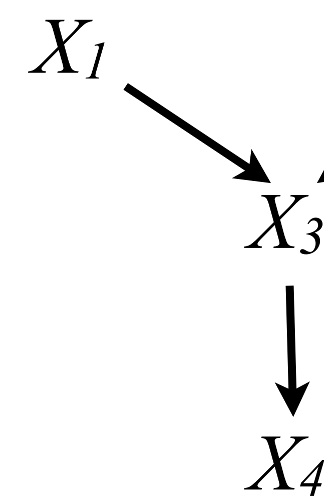
Example I

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

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

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

Possible to have confounders behind X_3 and X_4?



:-)

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

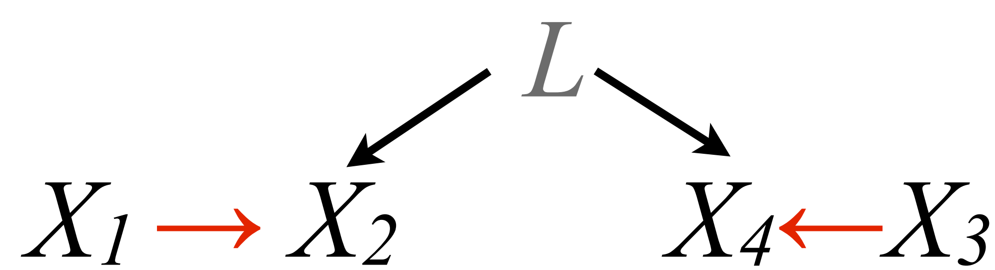
Example II

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

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

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

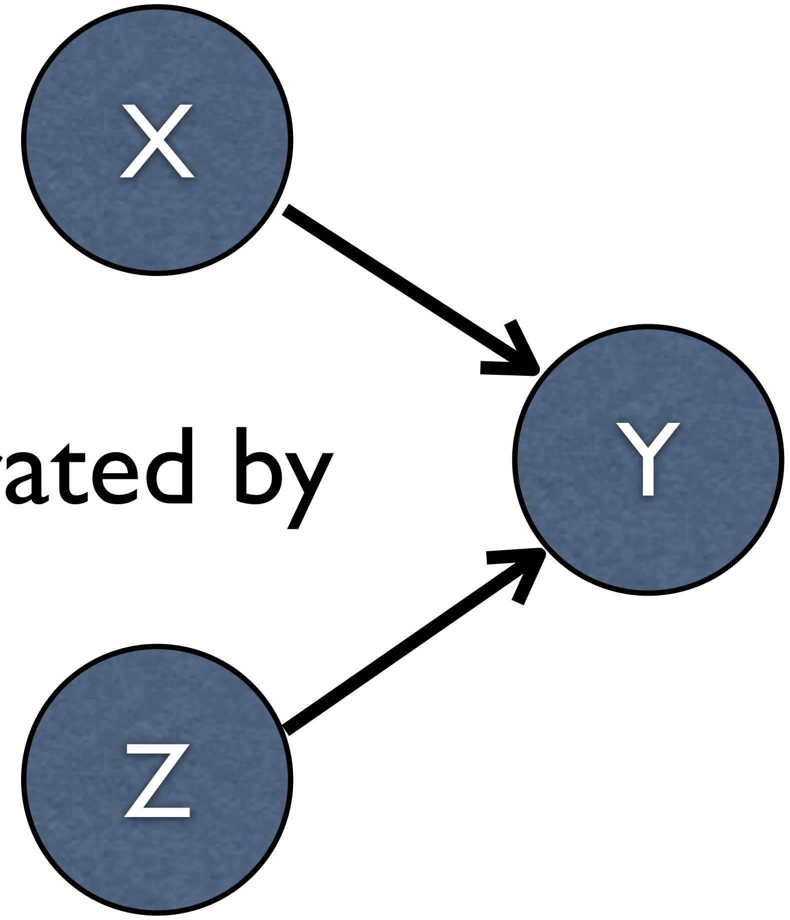
Are there confounders behind X_2 and X_4?



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.

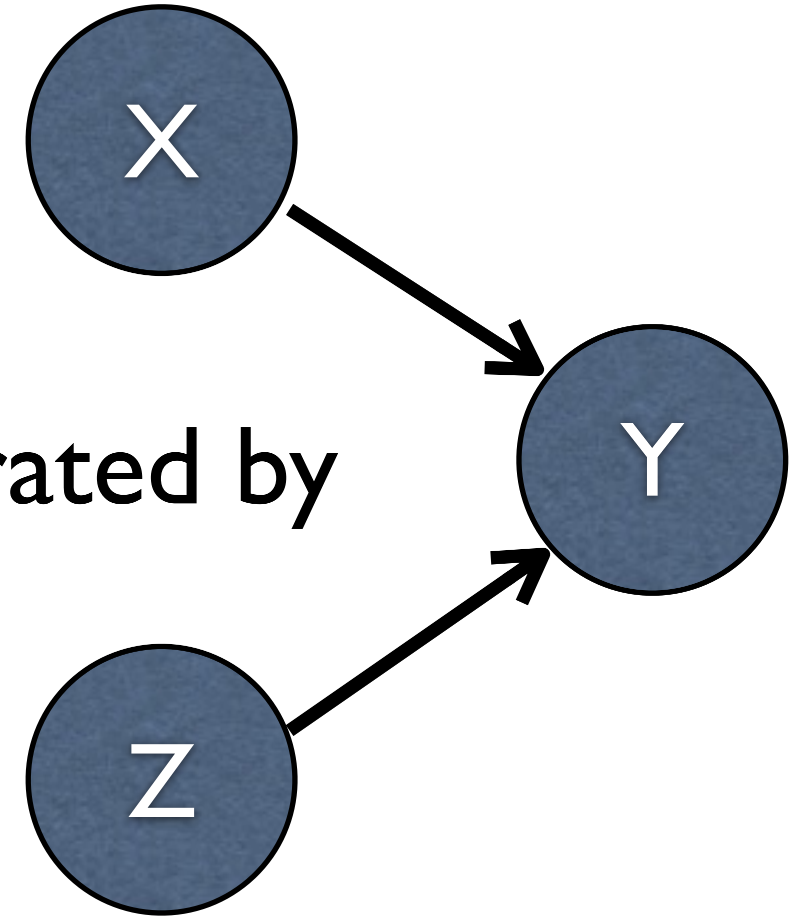
GES Assumes No Confounder

Suppose data were generated by

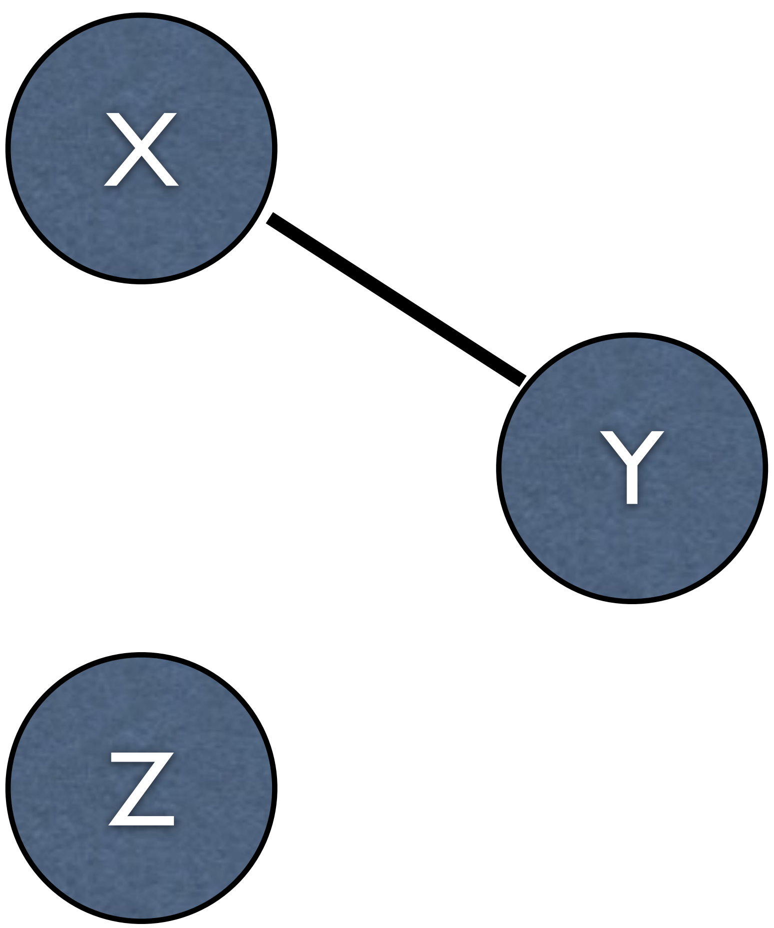


GES Assumes No Confounder

Suppose data were generated by

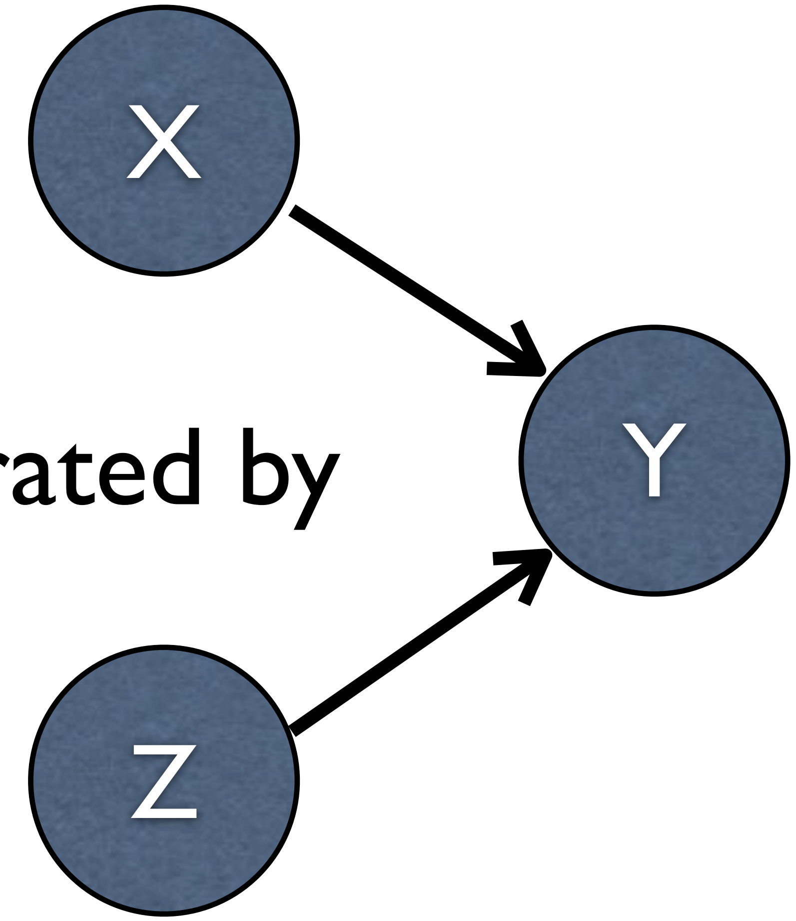


(1)

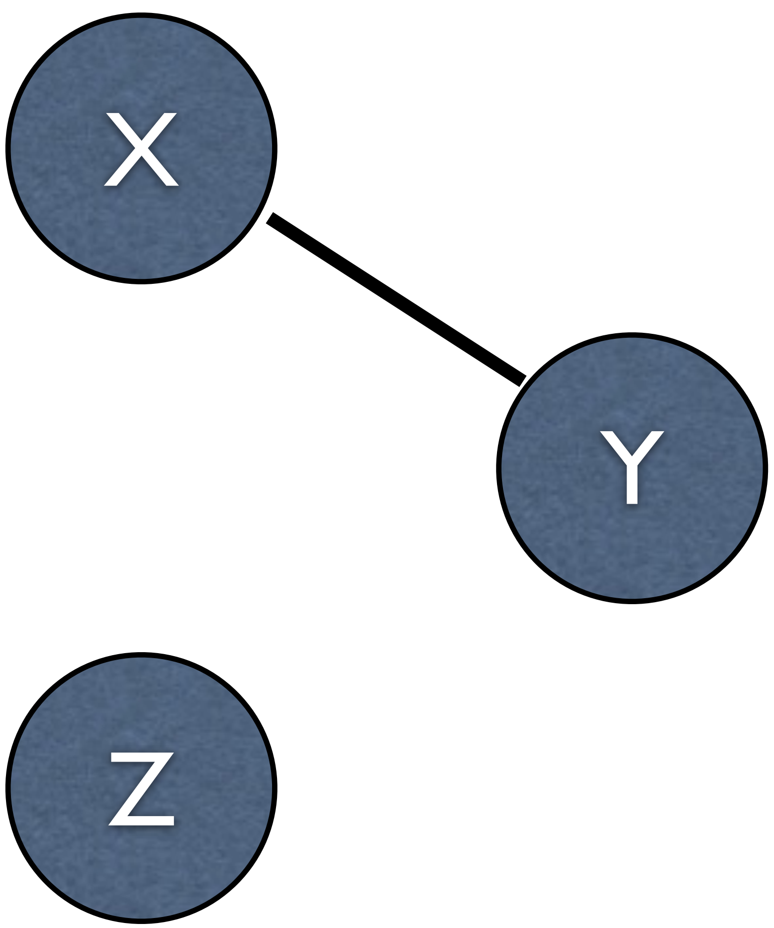


GES Assumes No Confounder

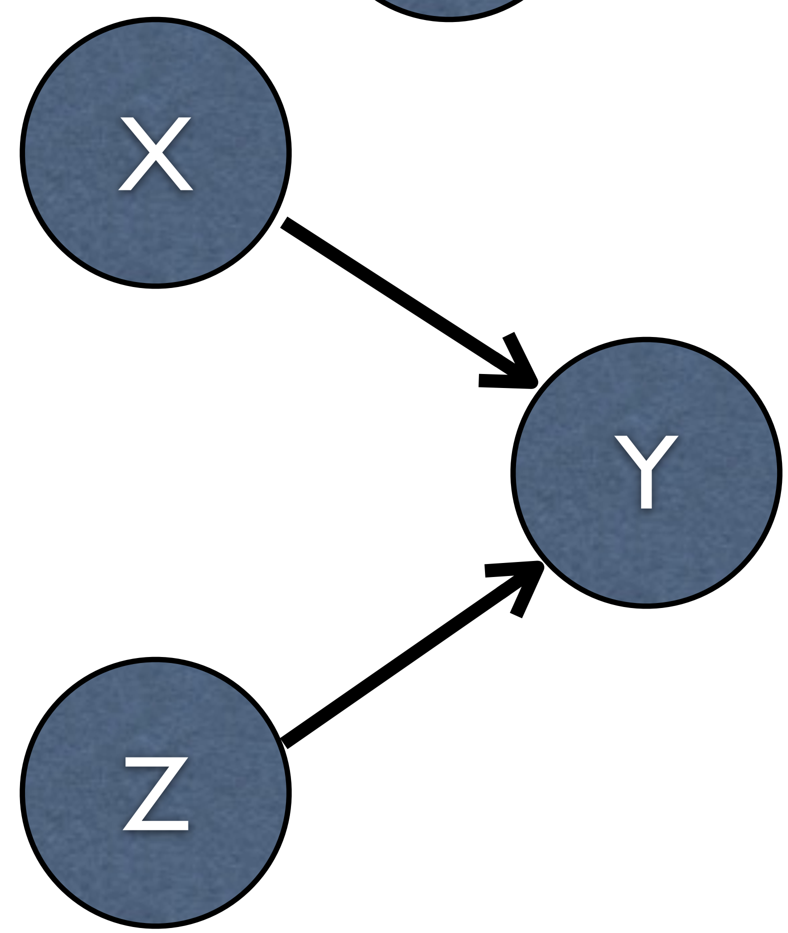
Suppose data were generated by



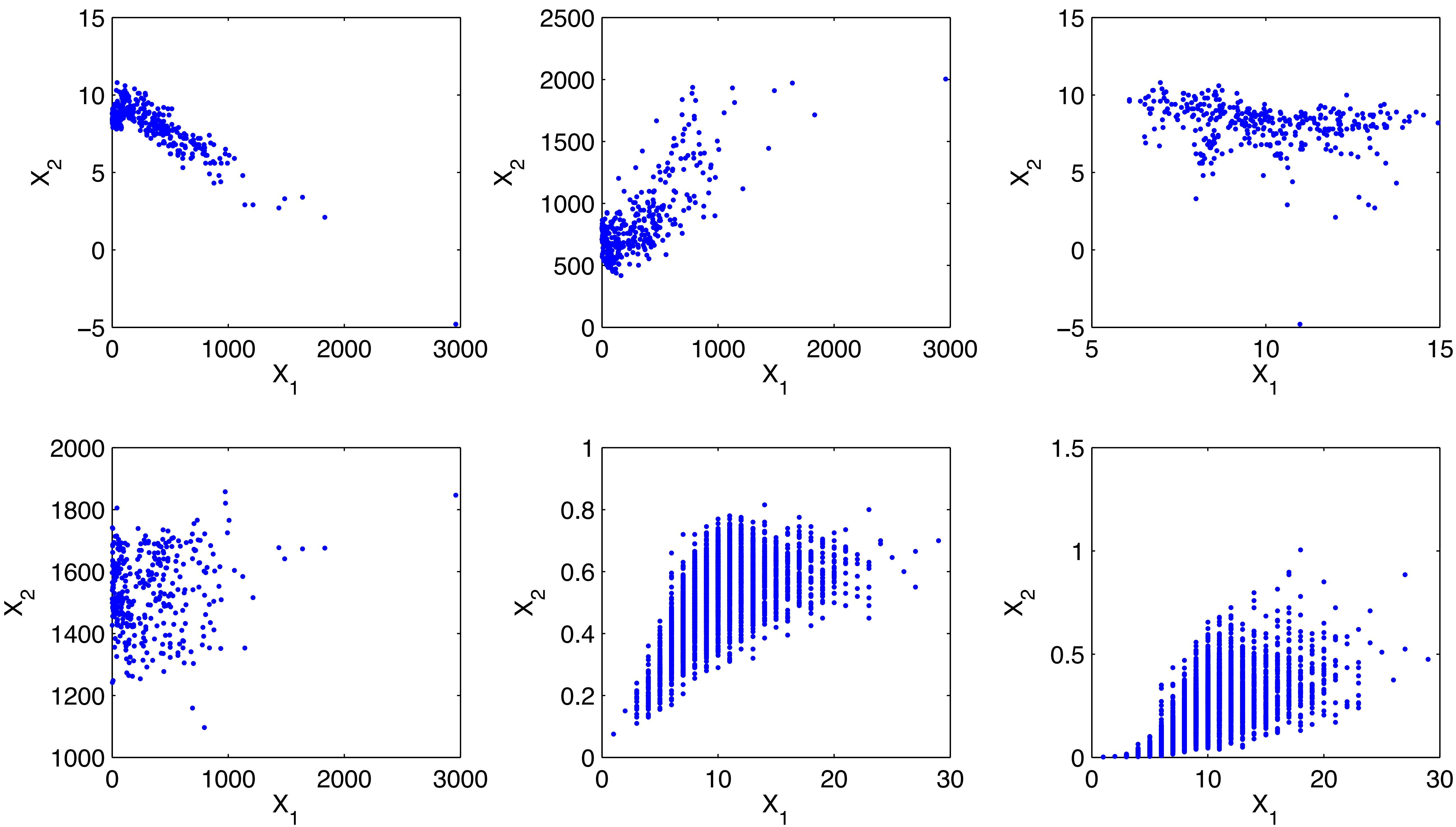
(1)



(2)



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

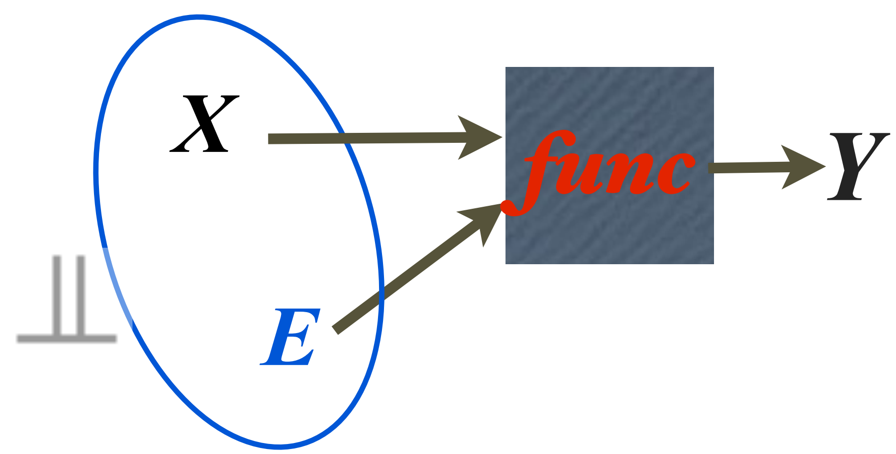


A Causal Process

rain $\longrightarrow$ *wet ground*

A Causal Process

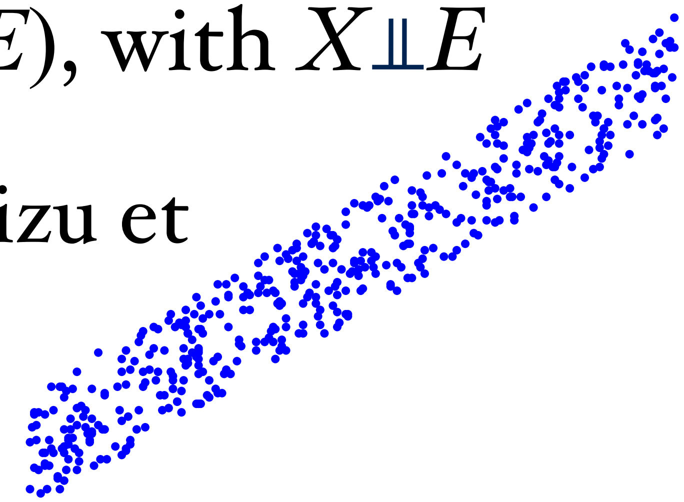
rain ⟶ *wet ground*



Functional Causal Model

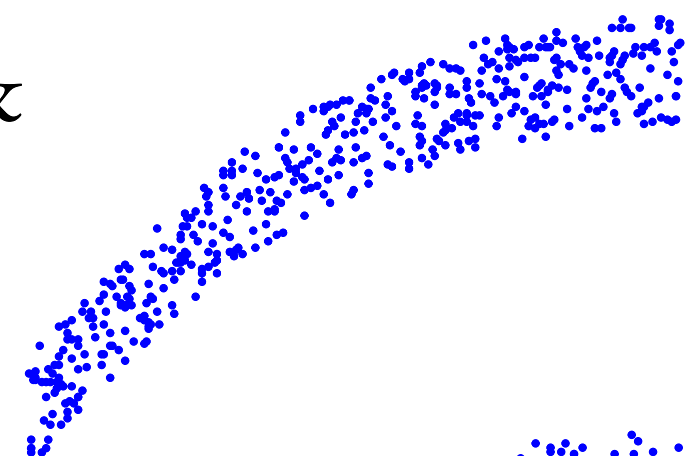
- A **functional causal model** represents effect as a function of direct causes and noise: $Y = f(X, E)$, with $X \perp\!\!\!\perp E$
- Linear non-Gaussian acyclic causal model (Shimizu et al., '06)

$$Y = a \cdot X + E$$



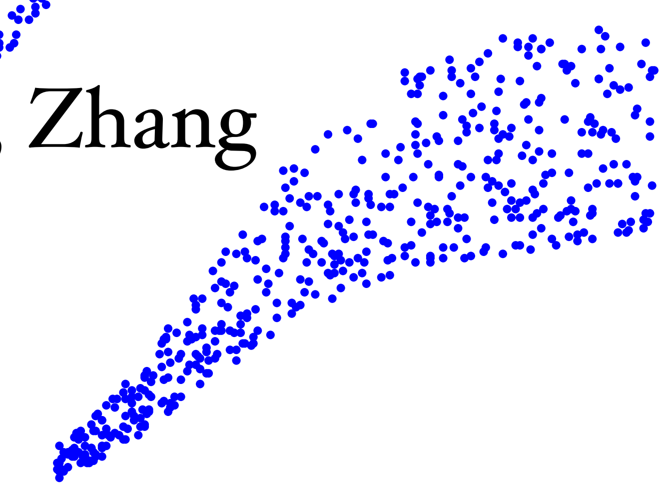
- Additive noise model (Hoyer et al., '09; Zhang & Hyvärinen, '09b)

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

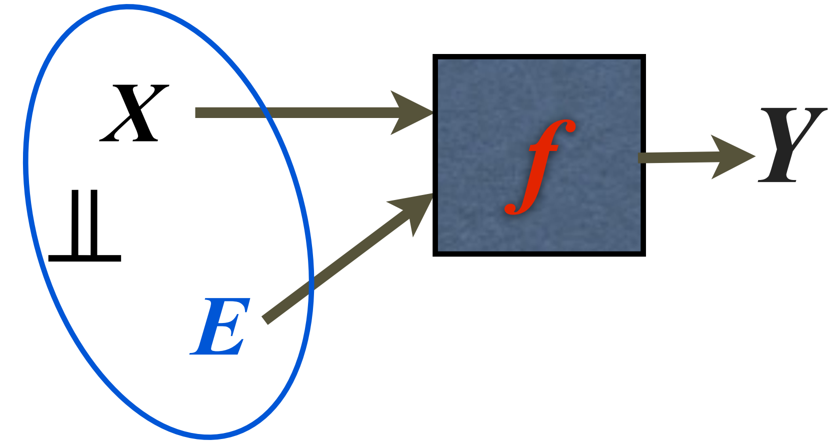


- Post-nonlinear causal model (Zhang & Chan, '06; Zhang & Hyvärinen, '09a)

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



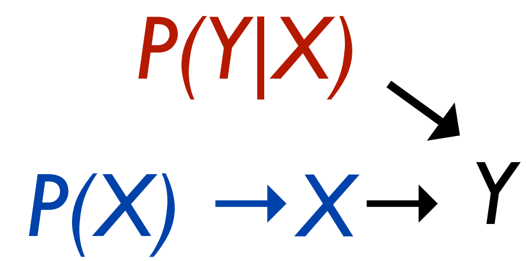
Functional Causal Models



- Effect generated from cause with **independent noise** (Pearl et al.):

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

  - A way to encode the intuition "the generating process for X is 'independent' from that generates Y from X"

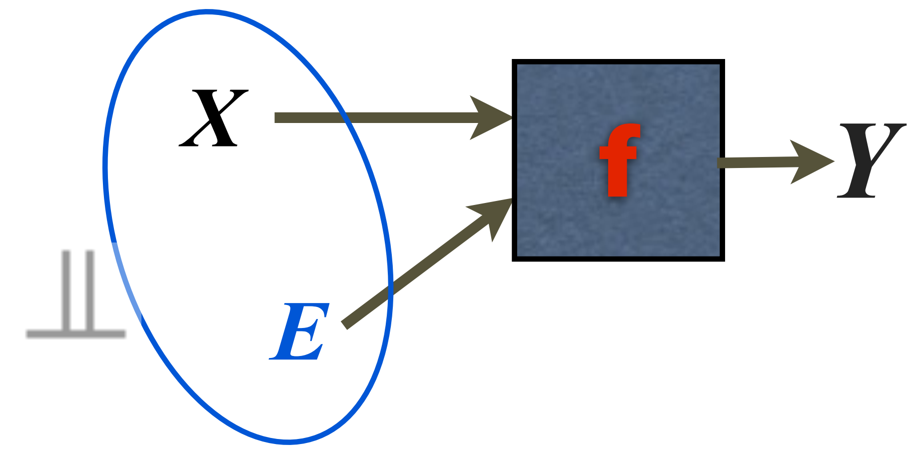


- :-(Without constraints on f, one can find independent noise for both directions (Darmois, 1951; Zhang et al., 2015)

  - Given any X_1 and X2, E' := conditional CDF of $X_2 \mid X_1$ is always independent from X_1 and $X2 = f(X_1, E')$

- :-) Structural constraints on f imply asymmetry

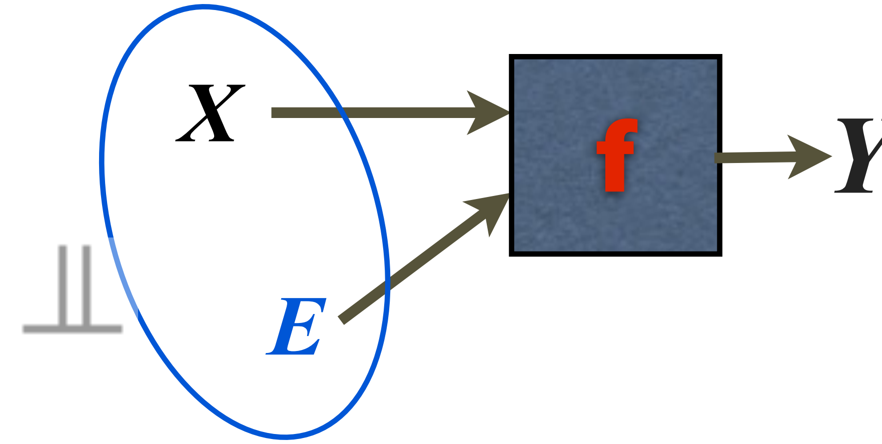
A Way to Construct Independent Error Term



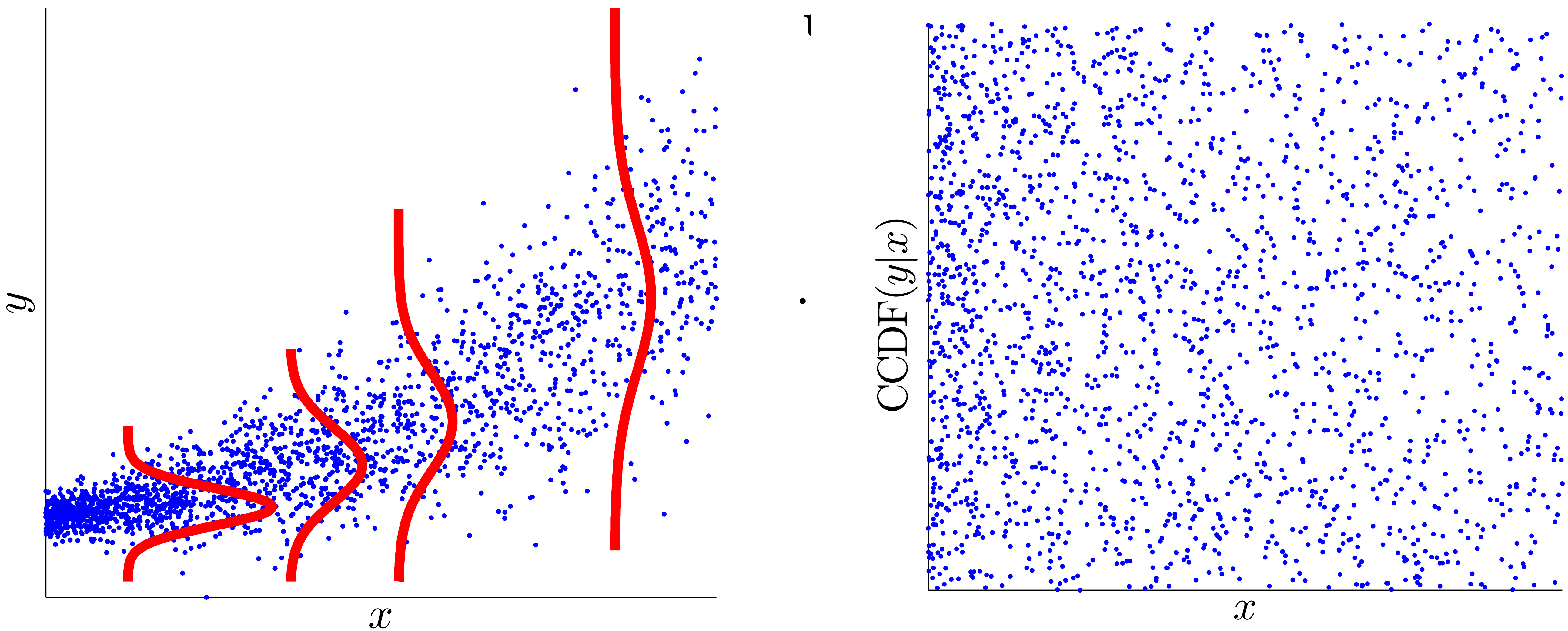
- CDF(Y) is a random variable uniformly distributed over [0,1]
- $E' \triangleq$ Conditional CDF($Y \mid X=x$) is uniformly distributed over [0,1], irrelevant to the value of x
- $E' \perp\!\!\!\perp X$
- Y can be written as $Y = f(X, E')$, i.e., the transformation from (X, Y) to (X, E') is invertible
- Why? The Jacobin !

Zhang et al.(2015), On Estimation of Functional Causal Models: General Results and Application to Post-Nonlinear Causal Model, ACM Transactions on Intelligent Systems and Technology, Forthcoming

A Way to Construct Independent Error Term

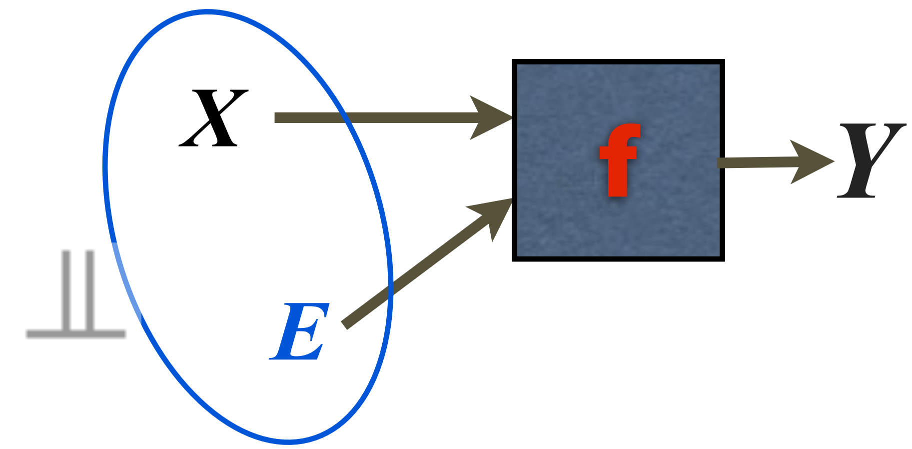


- CDF(Y) is a random variable uniformly distributed over [0,1]



Zhang et al.(2015), On Estimation of Functional Causal Models: General Results and Application to Post-Nonlinear Causal Model, ACM Transactions on Intelligent Systems and Technology, Forthcoming

A Way to Construct Independent Error Term

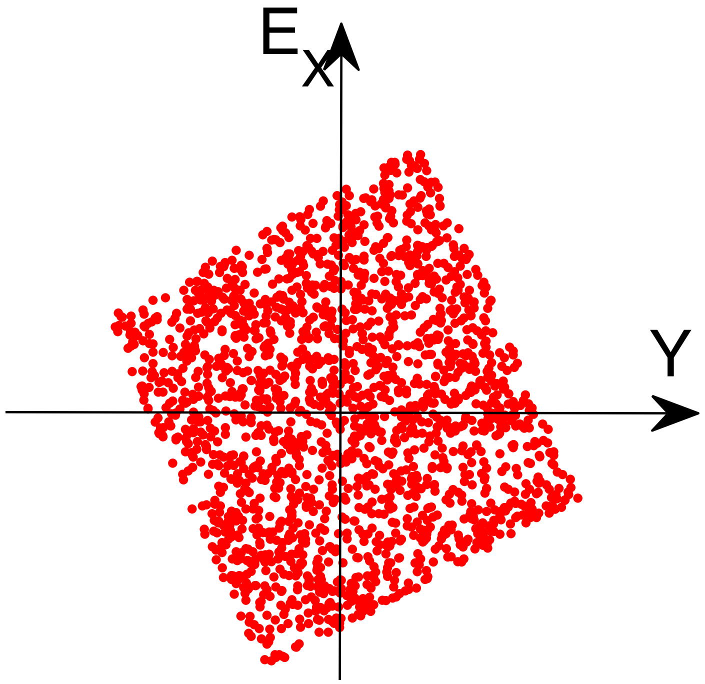


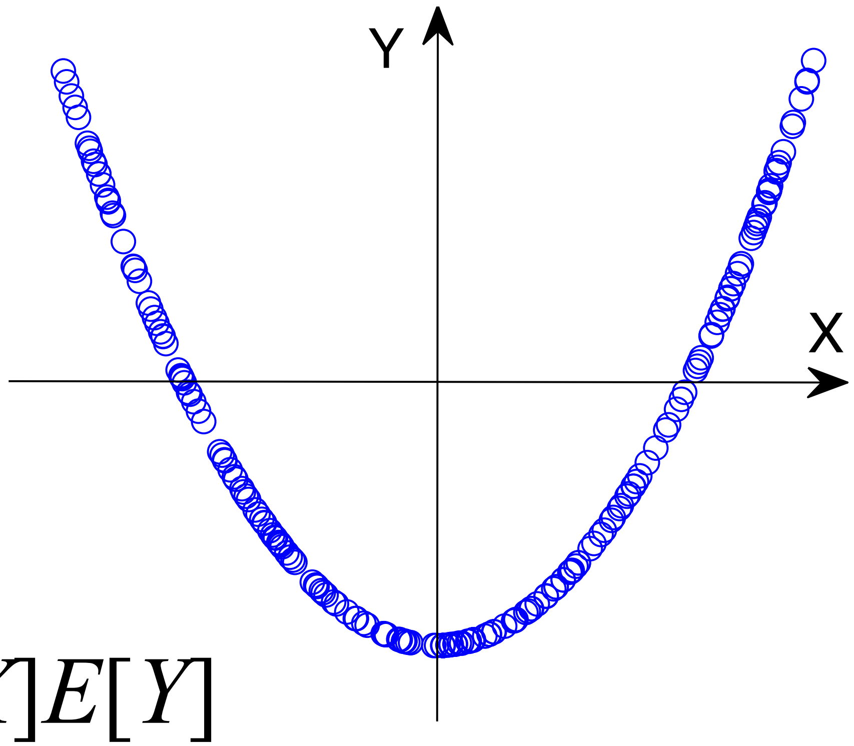
- CDF(Y) is a random variable uniformly distributed over [0,1]
- $E' \triangleq$ Conditional CDF($Y \mid X=x$) is uniformly distributed over [0,1], irrelevant to the value of x
- $E' \perp\!\!\!\perp X$
- Y can be written as $Y = f(X, E')$, i.e., the transformation from (X, Y) to (X, E') is invertible
- Why? The Jacobin !

Zhang et al.(2015), On Estimation of Functional Causal Models: General Results and Application to Post-Nonlinear Causal Model, ACM Transactions on Intelligent Systems and Technology, Forthcoming

(Conditional) Independence

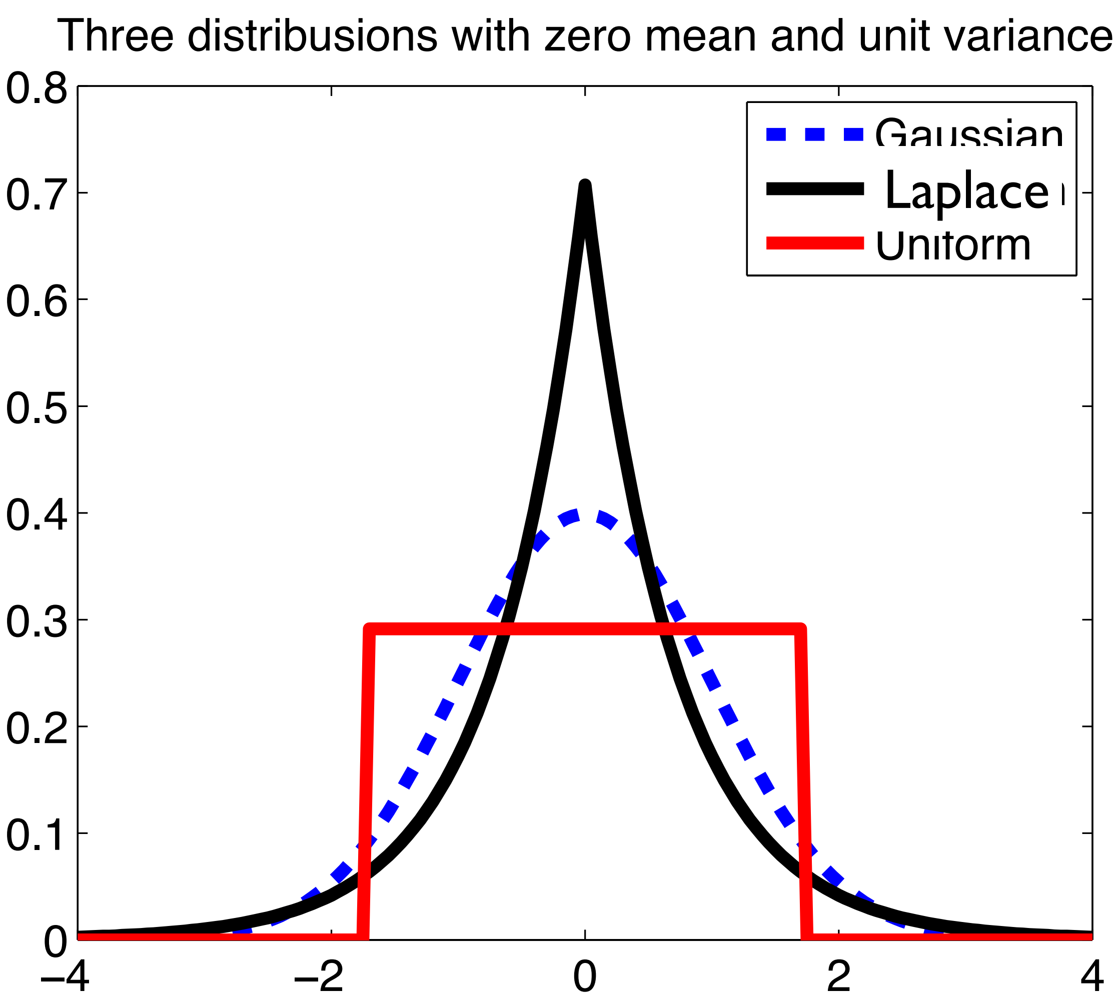
- $X \perp\!\!\!\perp Y$ iff $p(X,Y) = p(X)p(Y)$
  - or $p(X|Y) = P(X)$: Y not informative to X
- $\mathrm{X} \perp\!\!\!\perp \mathrm{Y} \mid \mathrm{Z}$ iff $p(X,Y|Z) = p(X|Z)p(Y|Z)$
  - or, $p(X|Y,Z) = p(X|Z)$: given Z, Y not informative to X
- Divide & conquer, remove irrelevant info…
- By construction,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

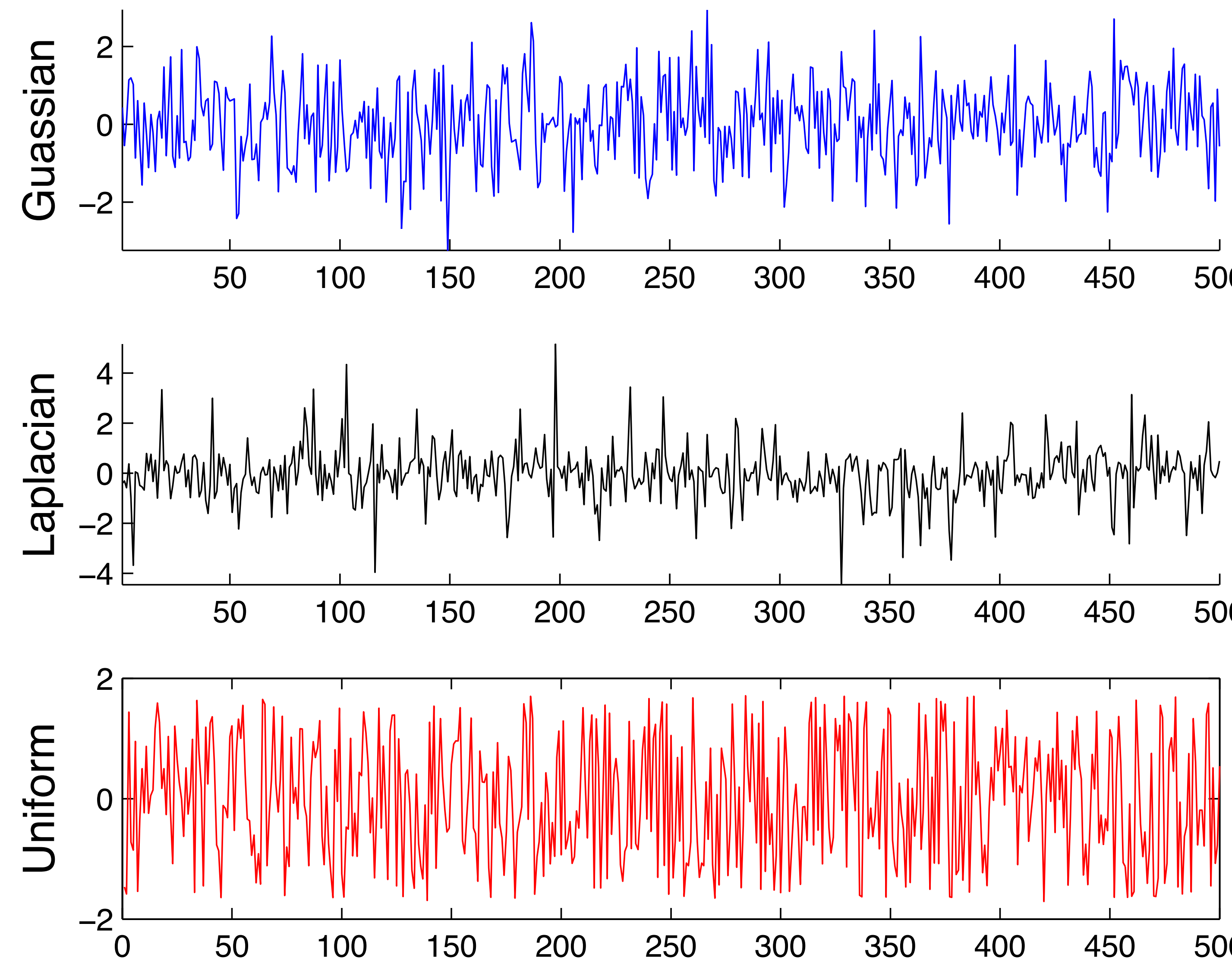




Uncorrelatedness: $E[XY] = E[X]E[Y]$

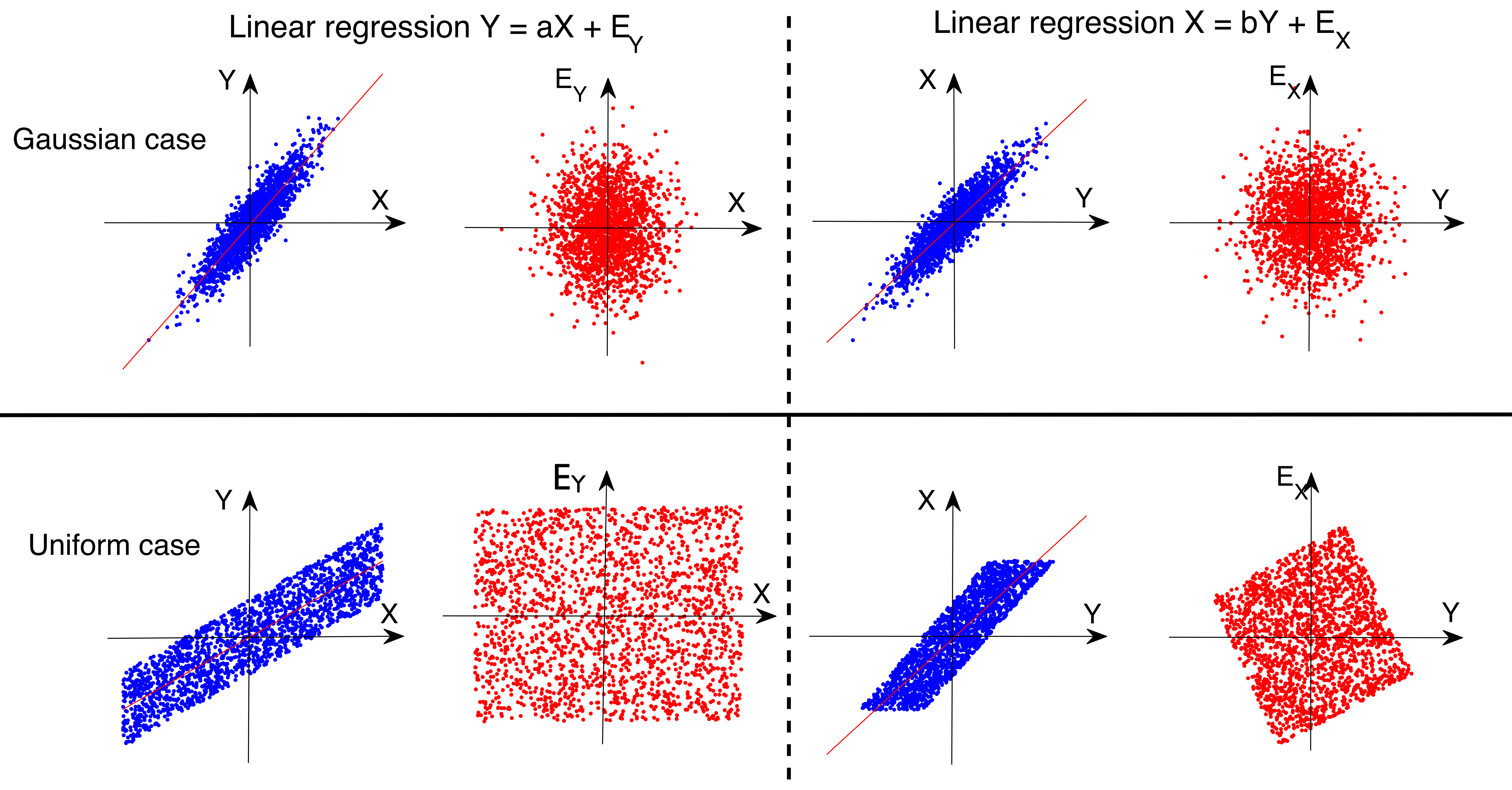
Gaussian vs. Non-Gaussian Distributions



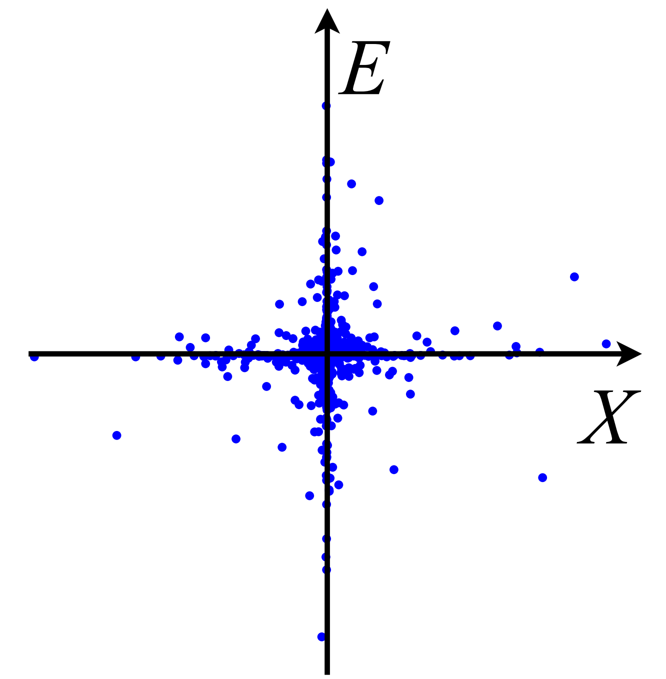


Causal Asymmetry the Linear Case: Illustration

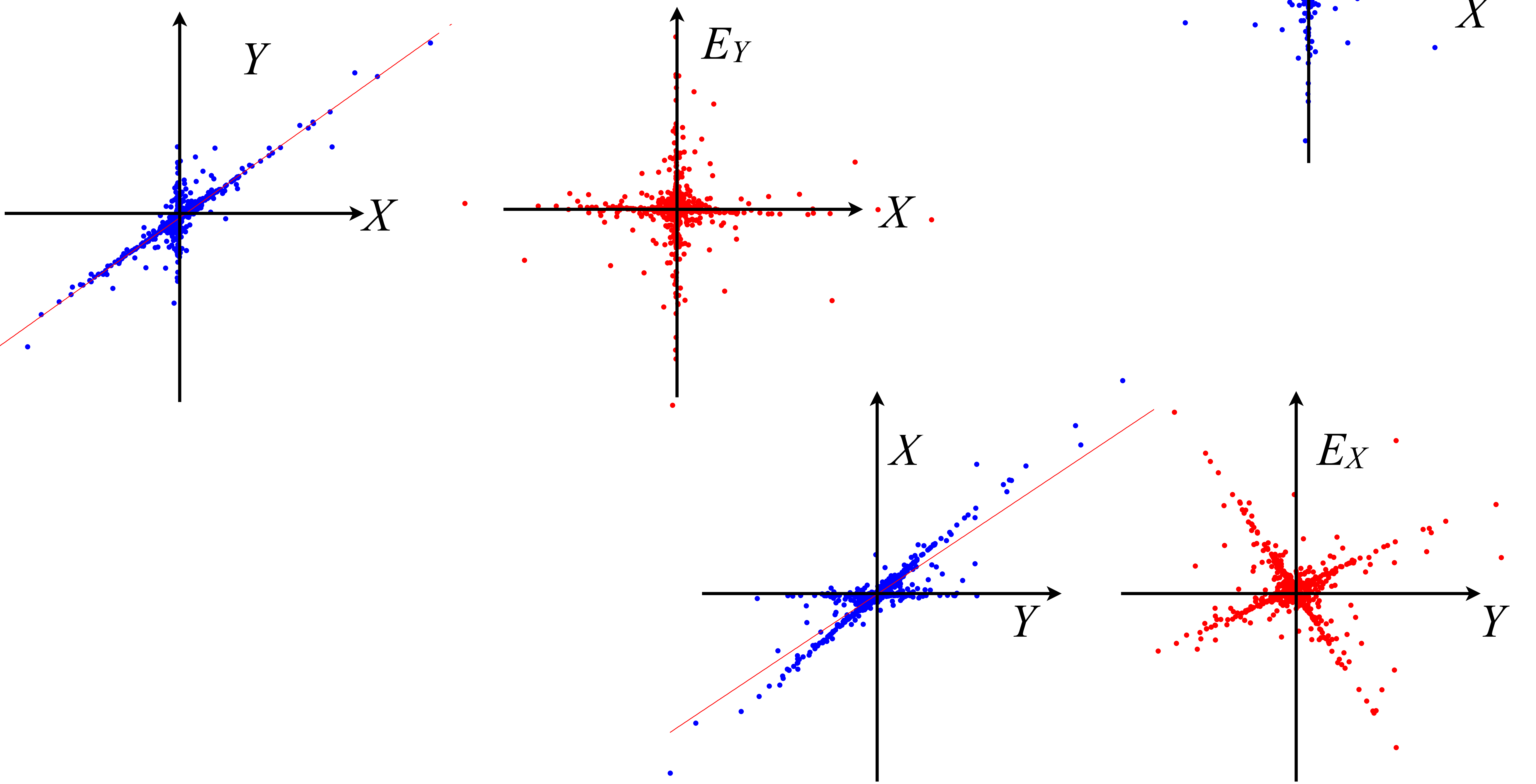
Data generated by $Y = aX + E$ (i.e., $X \to Y$):



Super-Gaussian Case



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



More Generally, LiNGAM Model

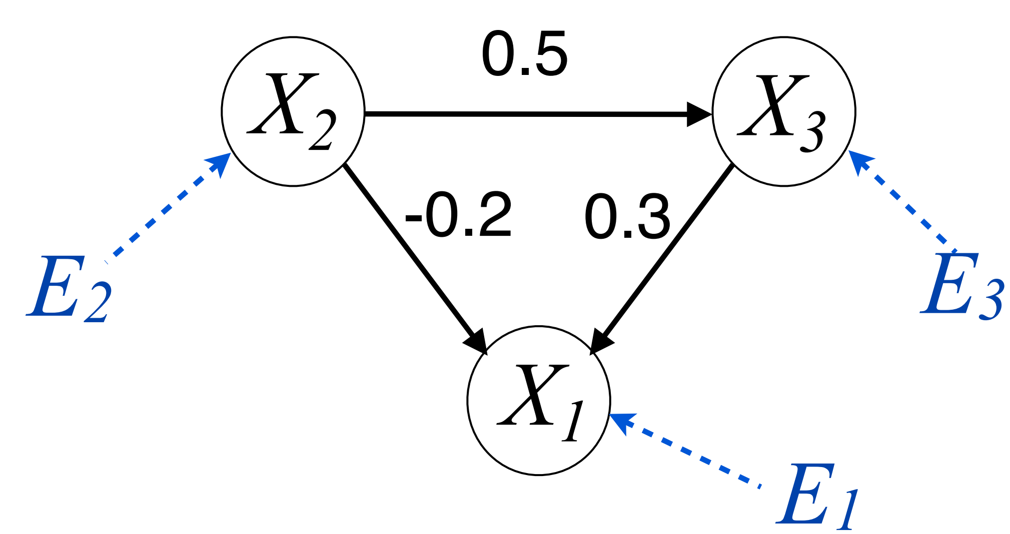
- Linear, non-Gaussian, acyclic causal model (LiNGAM) (Shimizu et al., 2006):

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

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

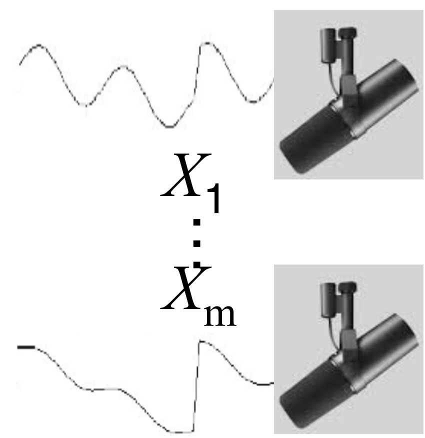
  - Example:

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



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

Independent Component Analysis



observed signals

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

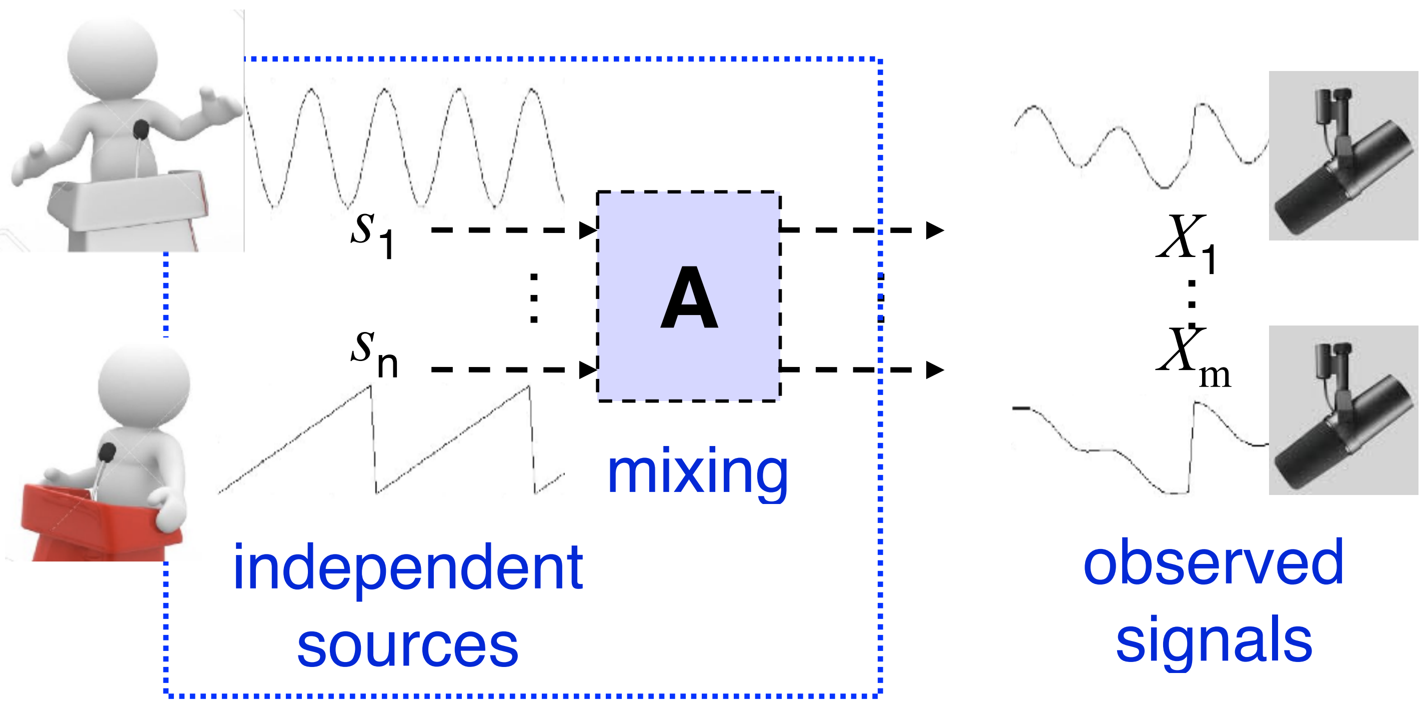
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}_{\mathbf{A}} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\ s_2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

- 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

Independent Component Analysis



unknown mixing system

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

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

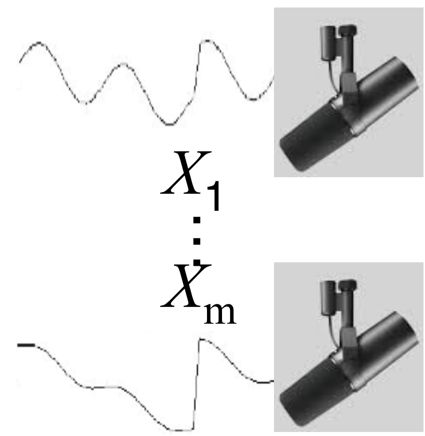
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \underbrace{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\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

- 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

Independent Component Analysis



observed signals

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad\qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

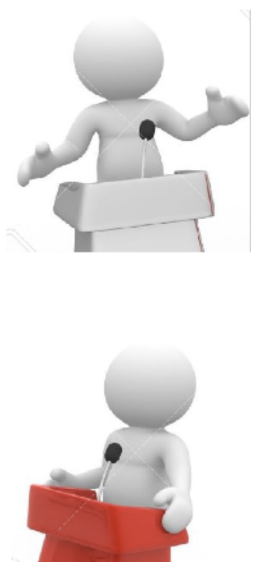
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \underbrace{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\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

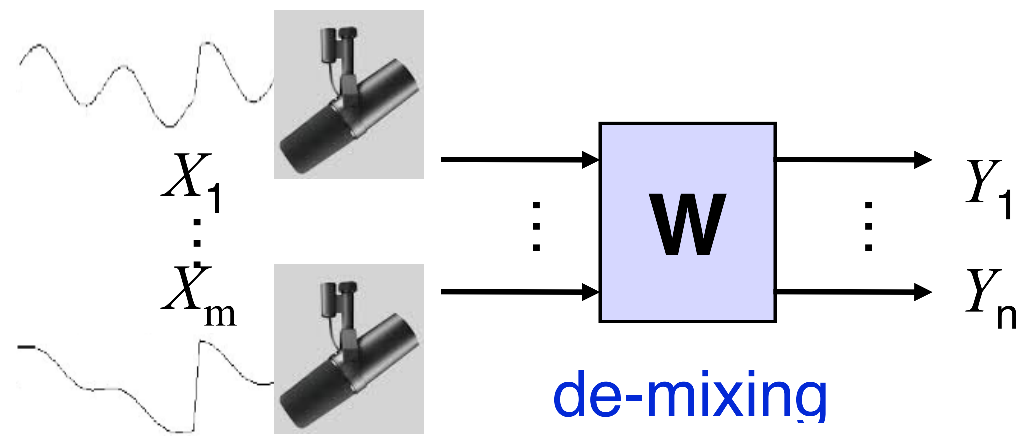
- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

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

Hyvärinen et al., Independent Component Analysis, 2001

Independent Component Analysis





observed signals

ICA system

output: as independent as possible

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad\qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

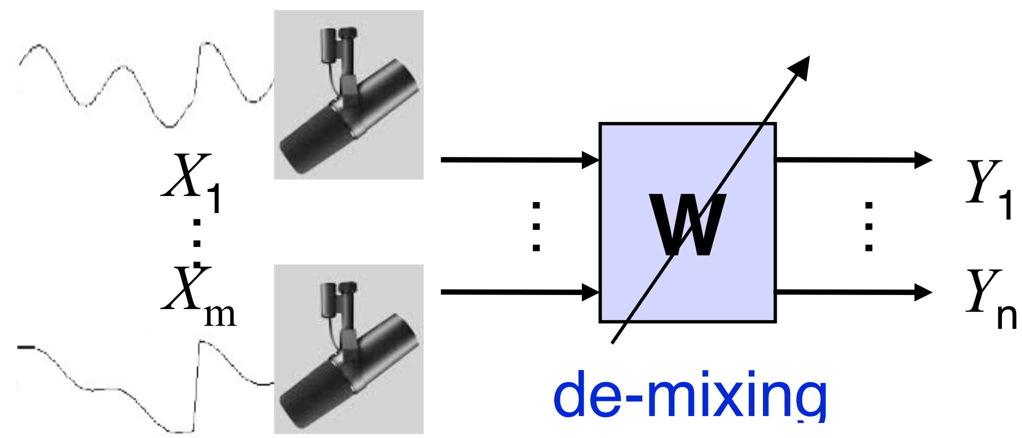
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \overset{A}{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\ s_2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

- 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

Independent Component Analysis



observed signals

ICA system

output: as independent as possible

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad\qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

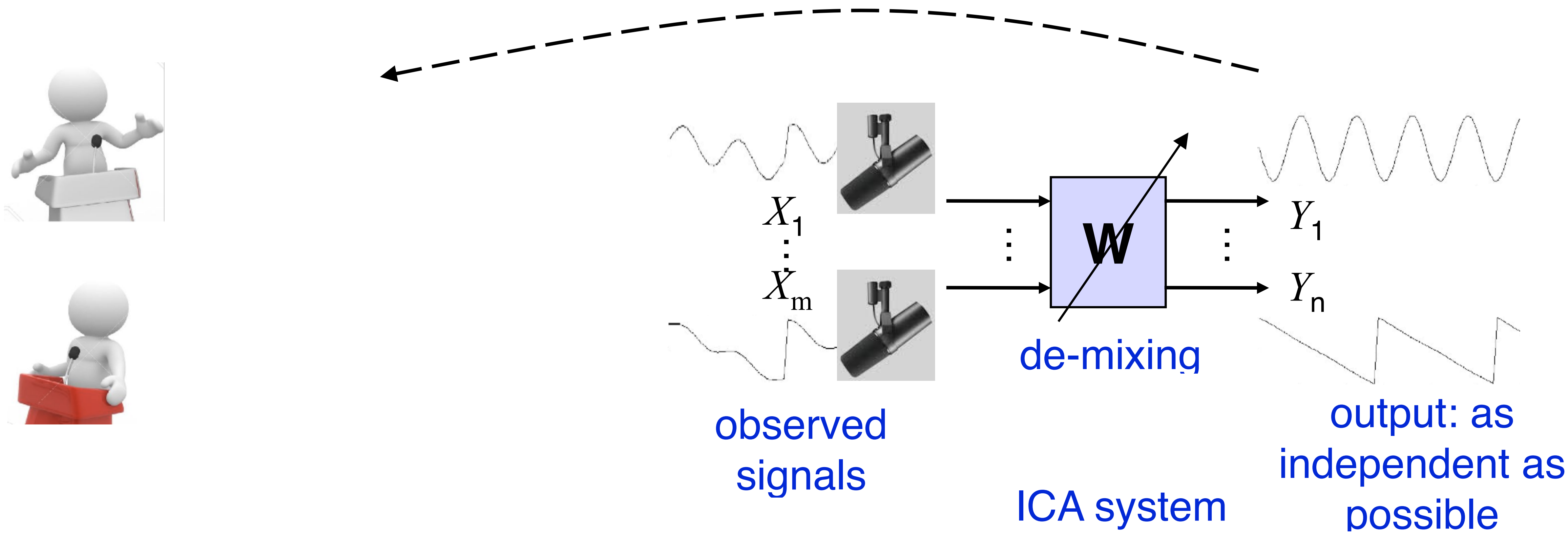
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \underbrace{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\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

- 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

Independent Component Analysis



$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad\qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

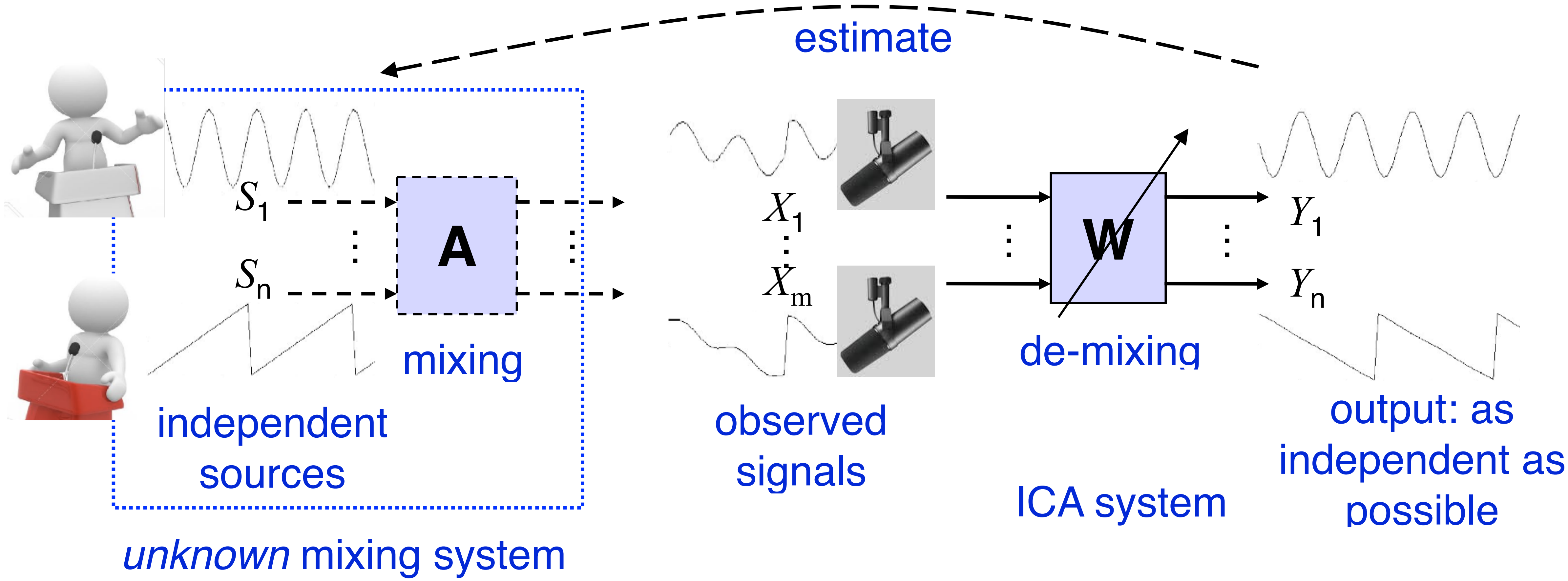
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \underbrace{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\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

- 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

Independent Component Analysis



$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad\qquad \mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

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

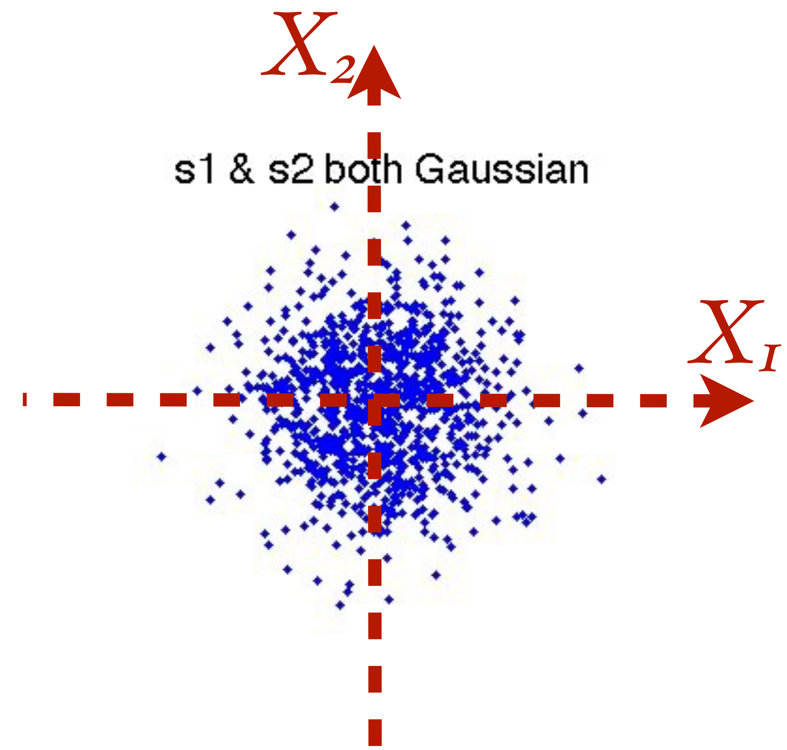
- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

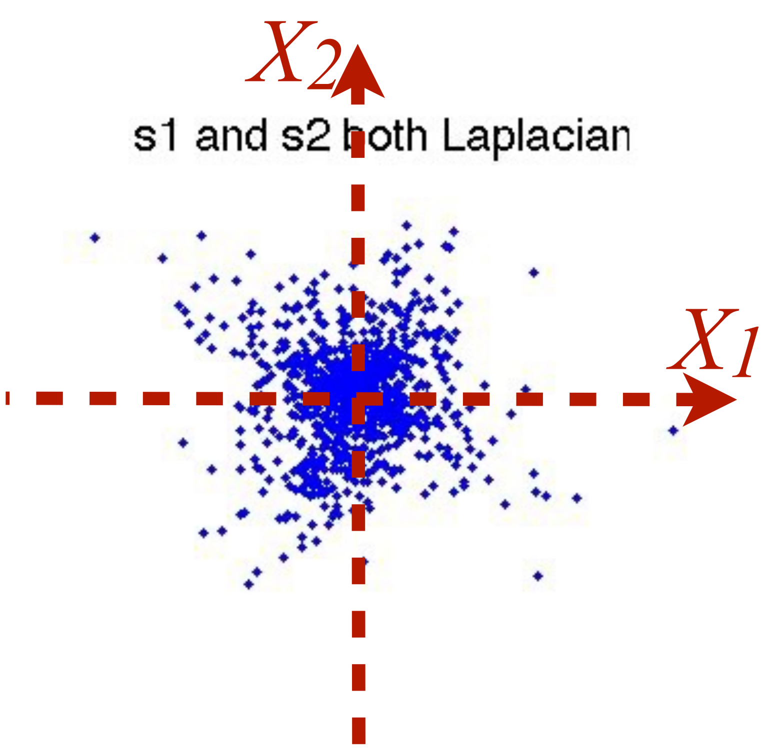
Then A can be estimated up to column scale and permutation indeterminacies

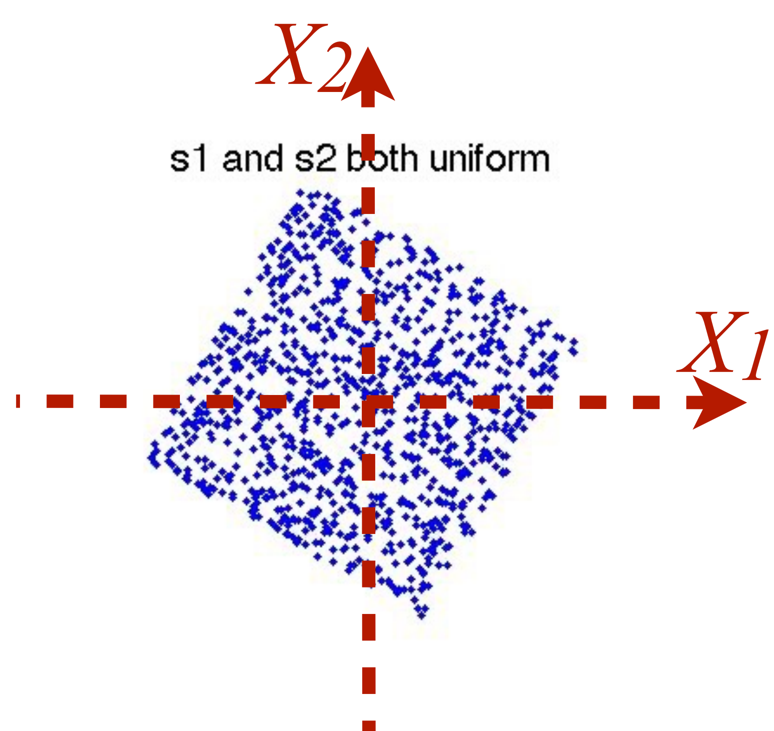
Hyvärinen et al., Independent Component Analysis, 2001

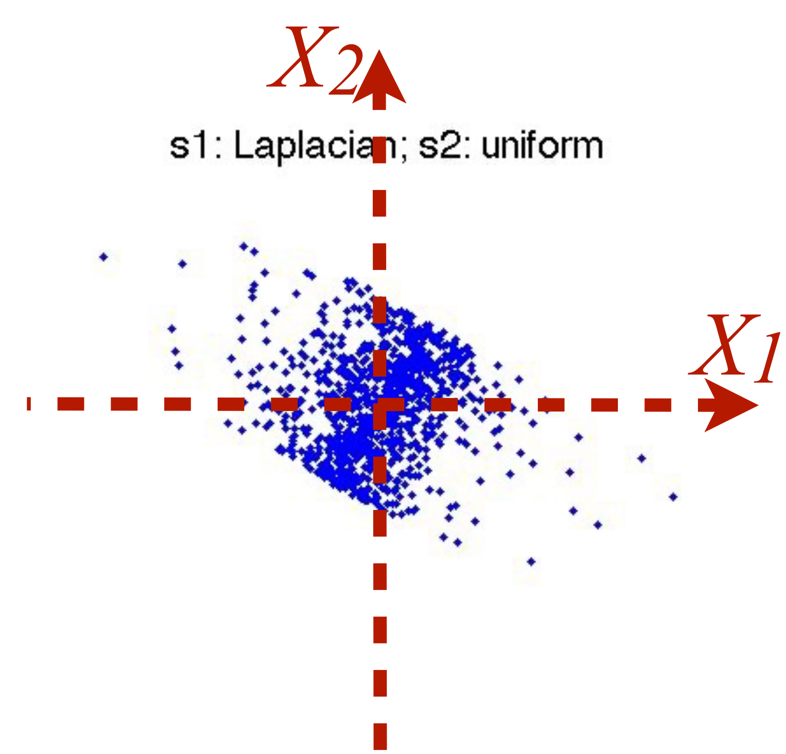
Intuition: Why ICA works?

- (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...



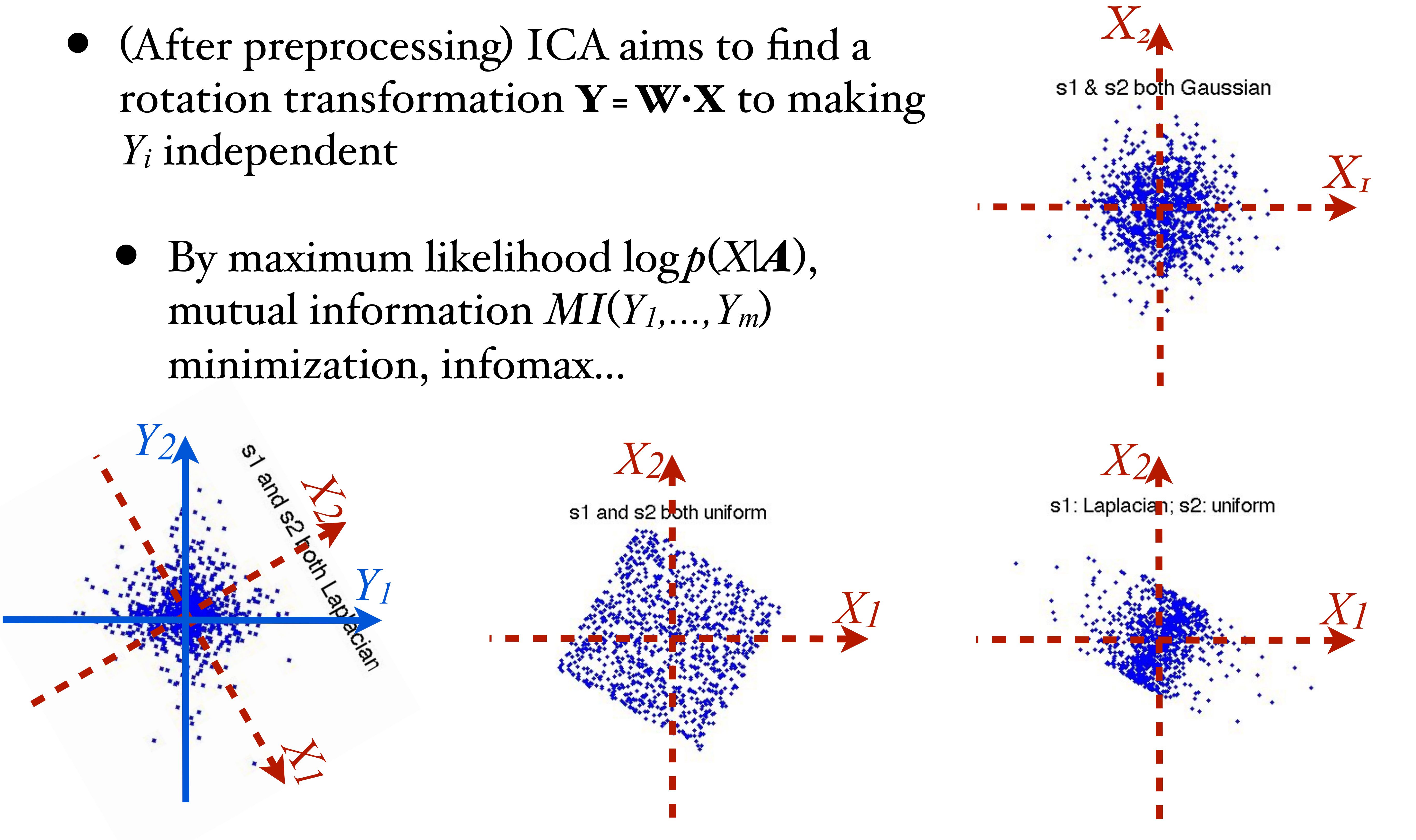






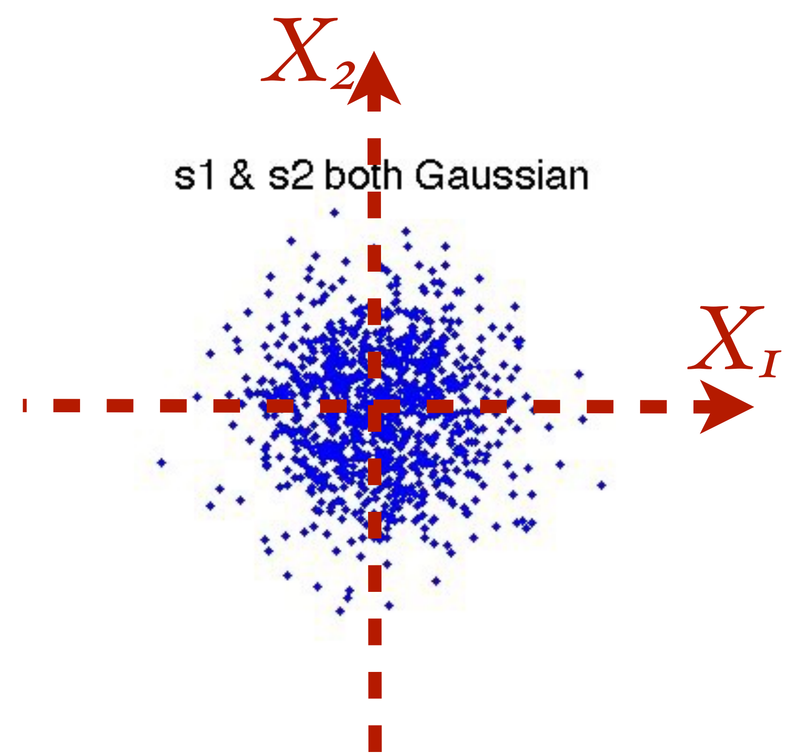
Intuition: Why ICA works?

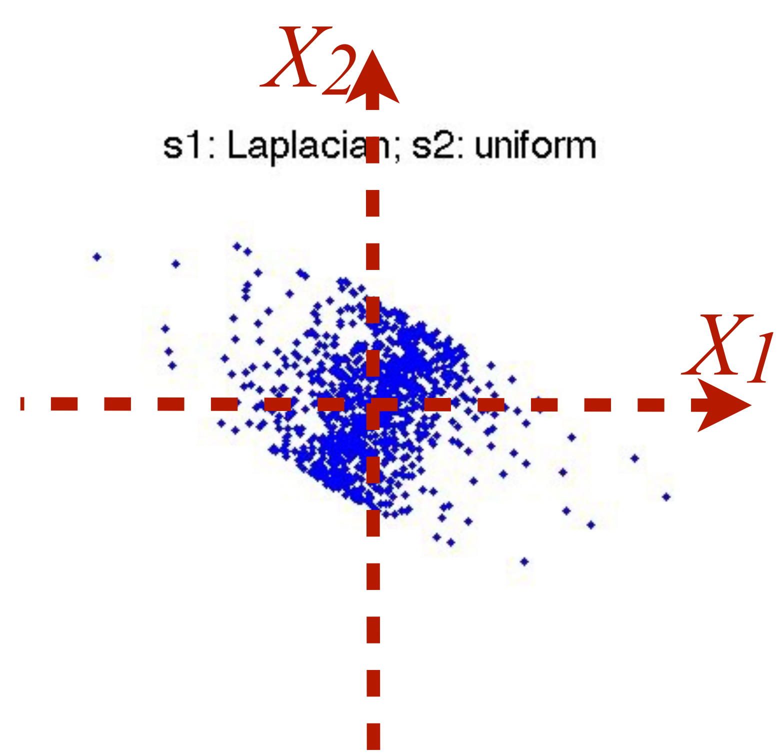
- (After preprocessing) ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$ to making Y_i independent
  - By maximum likelihood $\log p(X|\boldsymbol{A})$, mutual information $MI(Y_1, \ldots, Y_m)$ minimization, infomax...

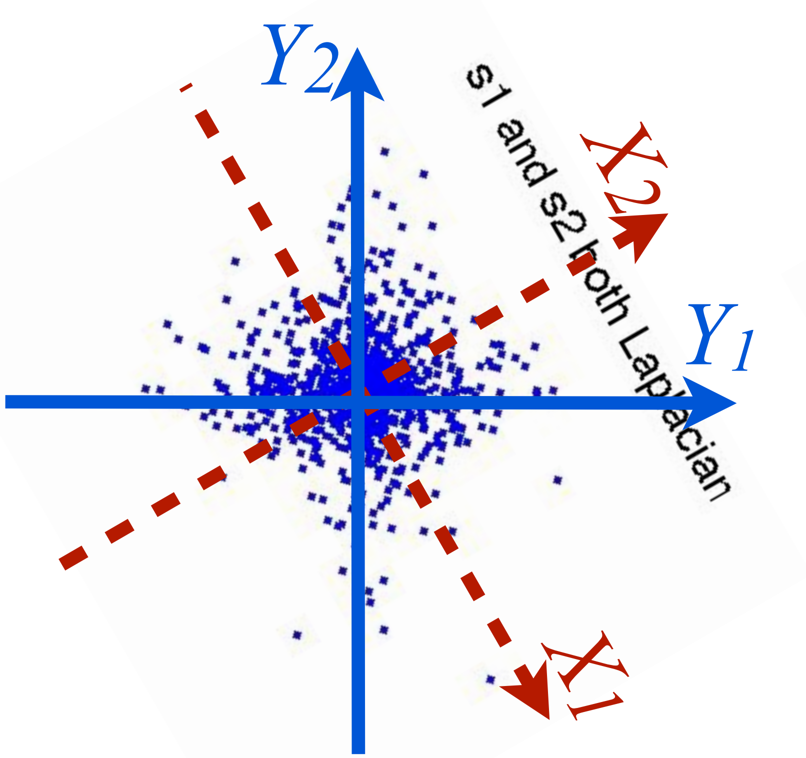


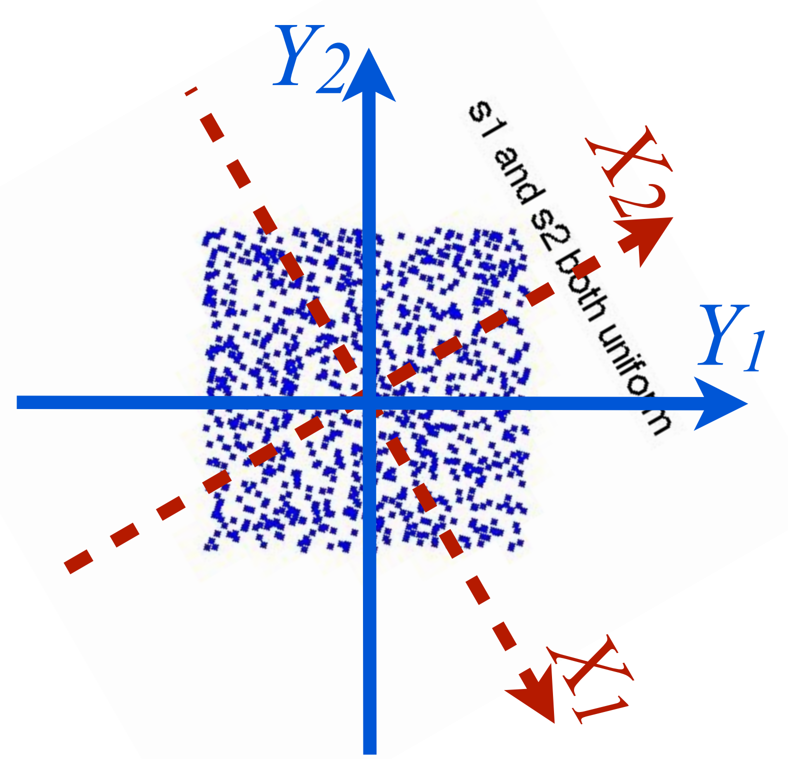
Intuition: Why ICA works?

- (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...



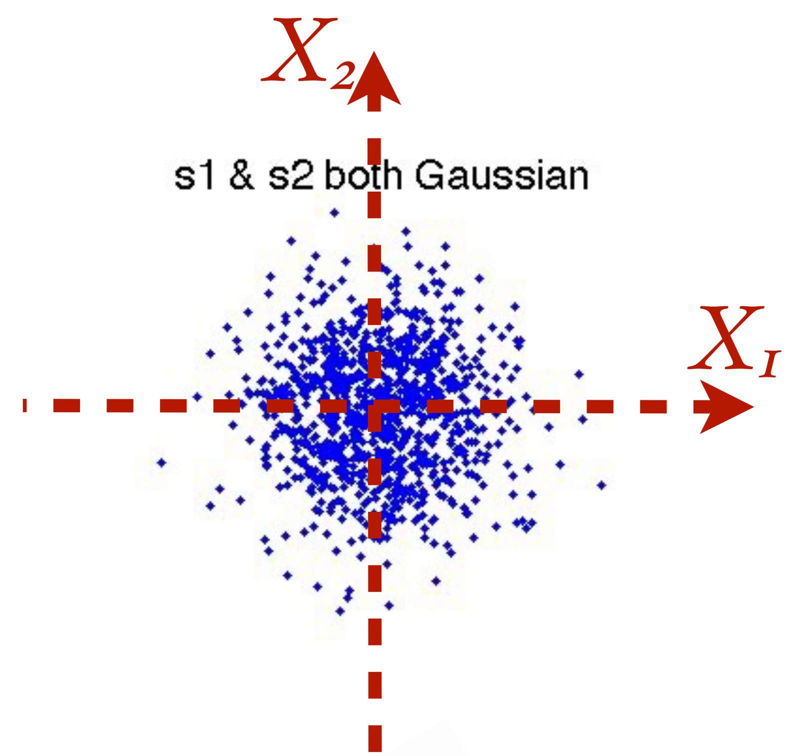


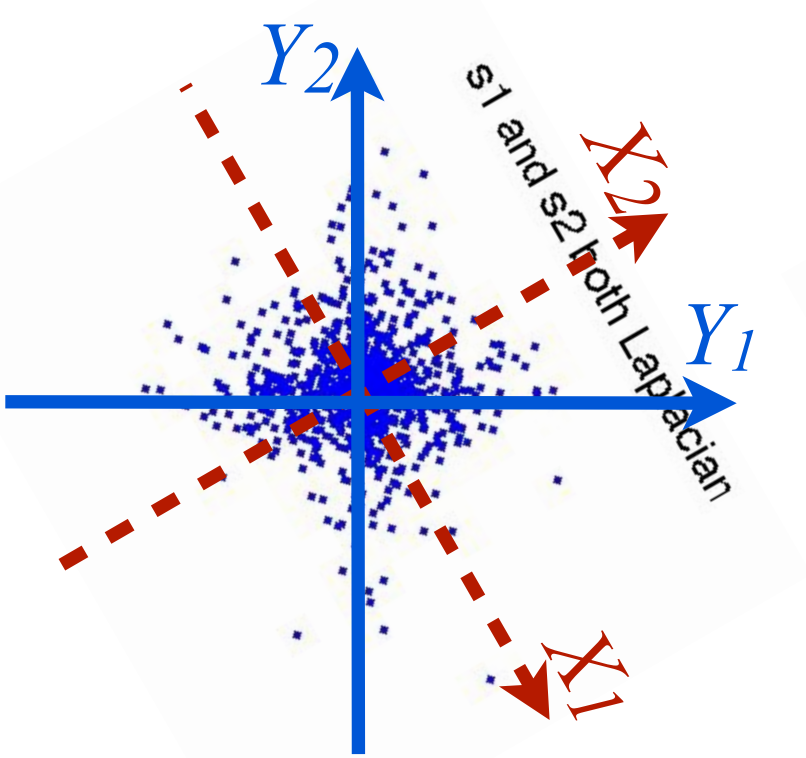


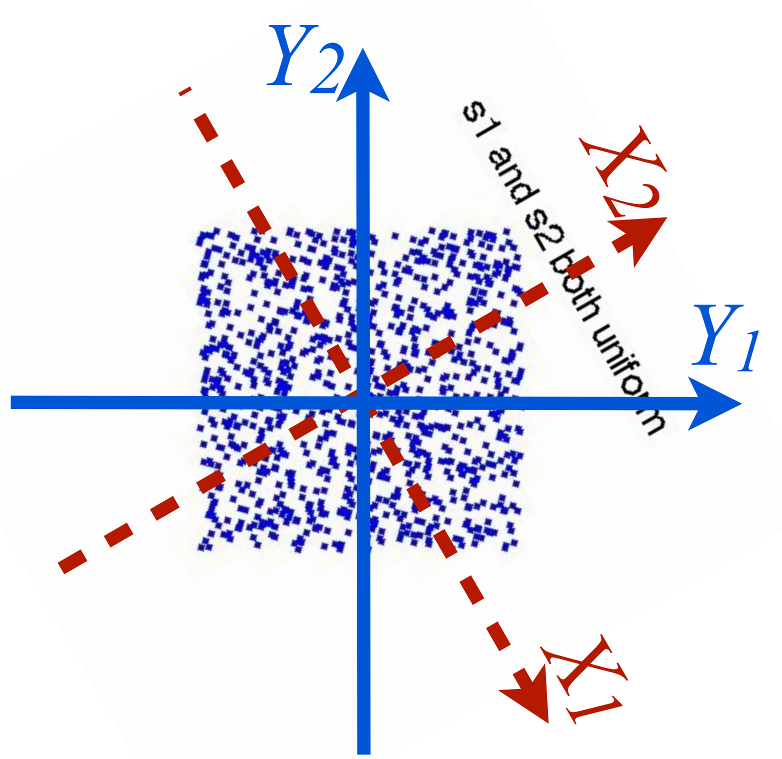


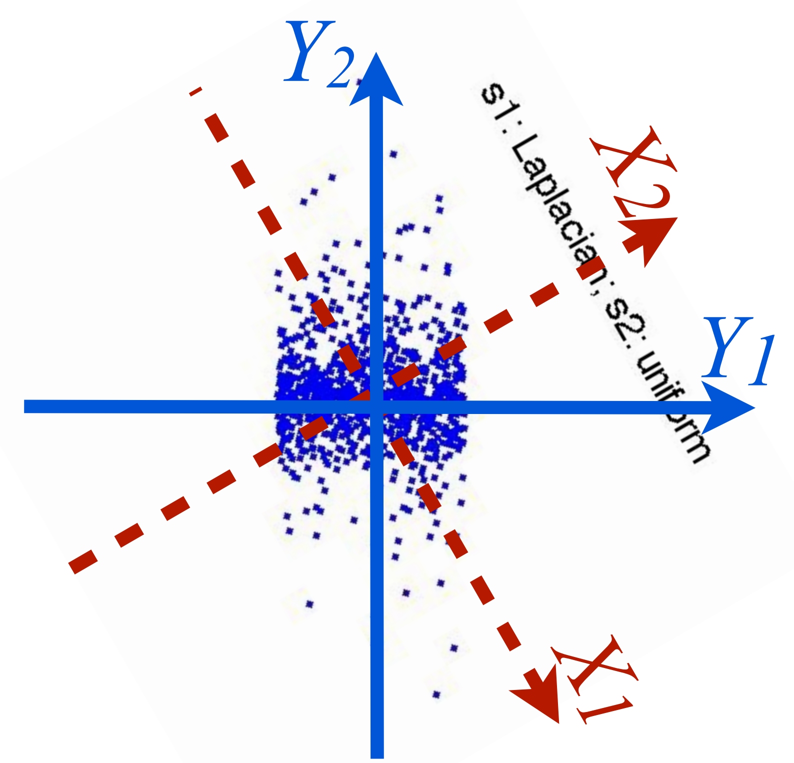
Intuition: Why ICA works?

- (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 $MI(Y_1, \dots, Y_m)$ minimization, infomax...

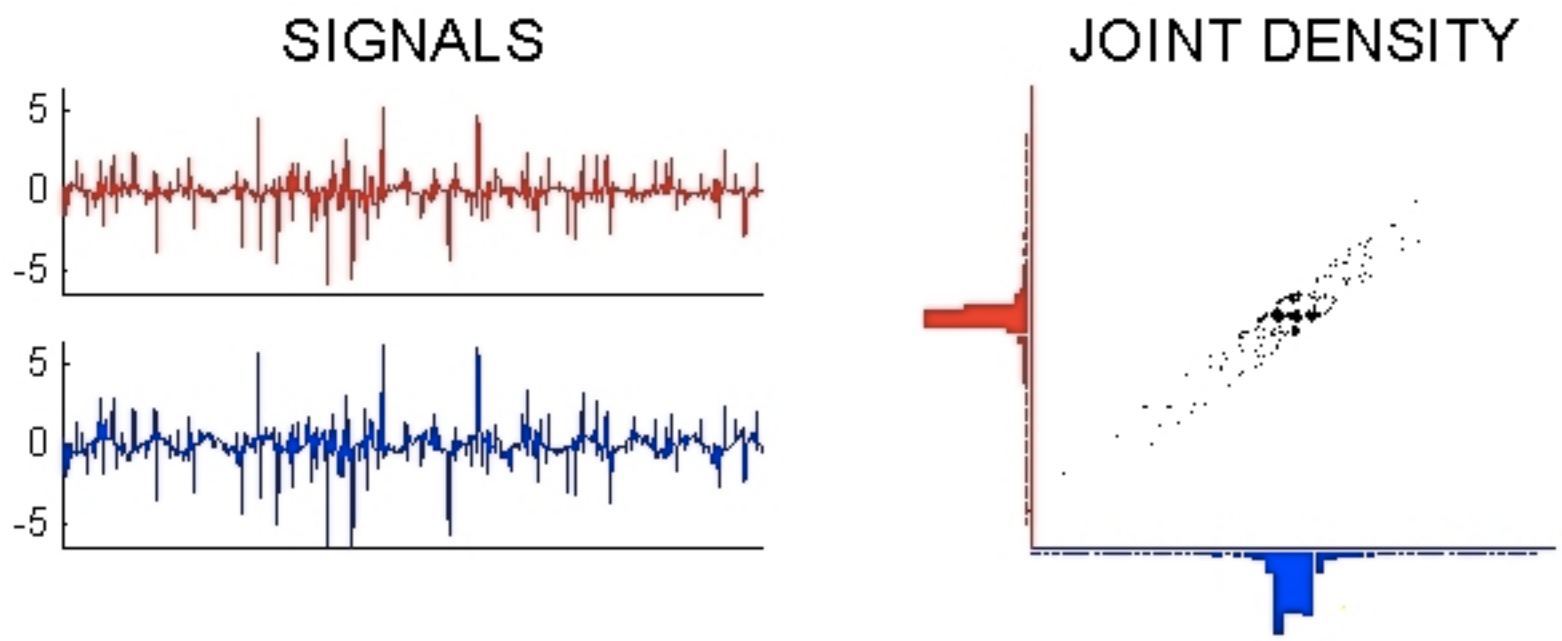




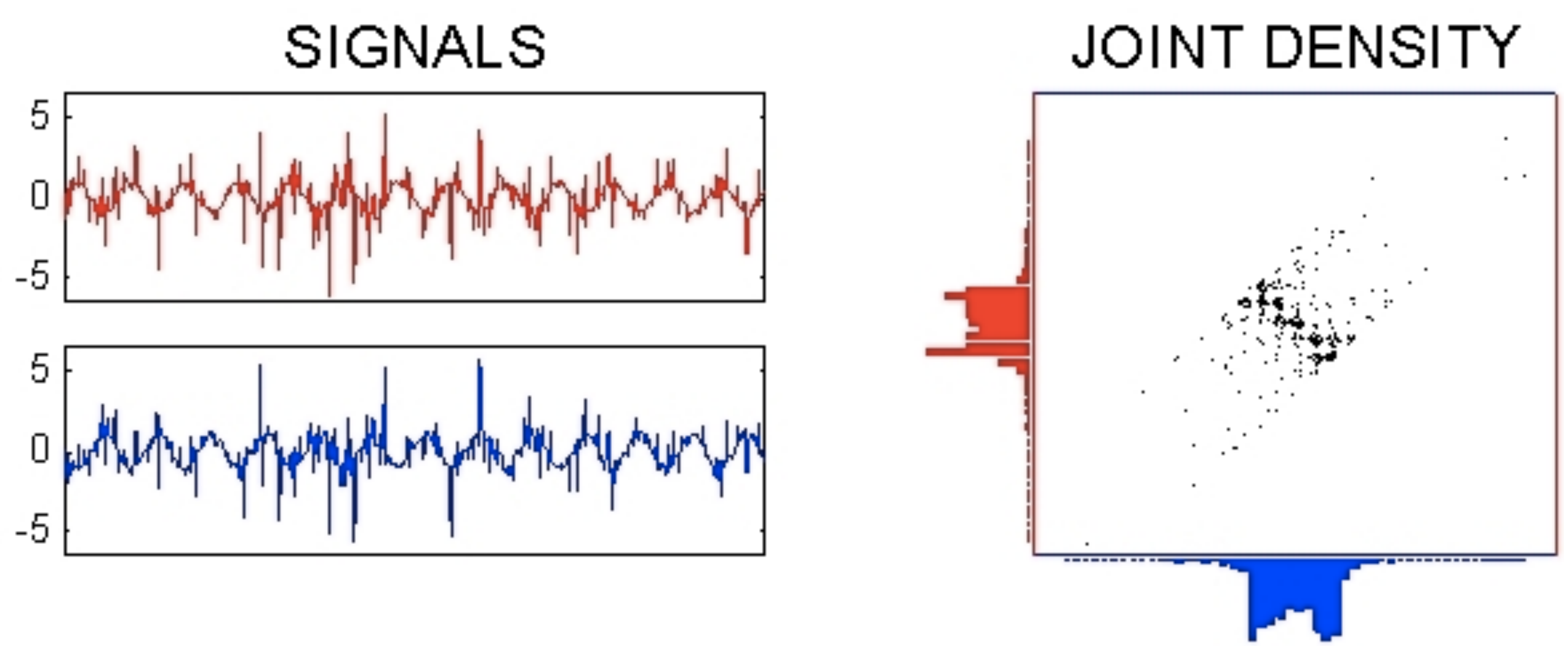




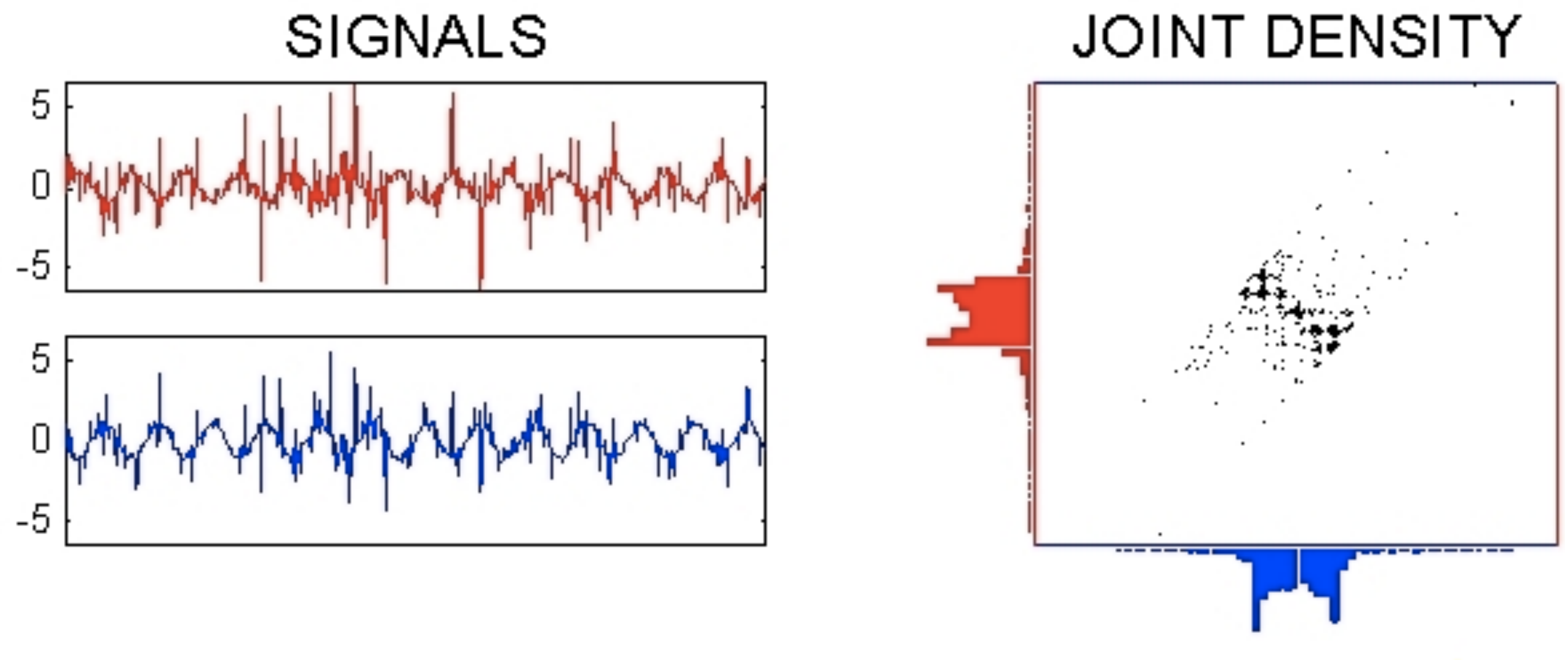
A Demo of the ICA Procedure



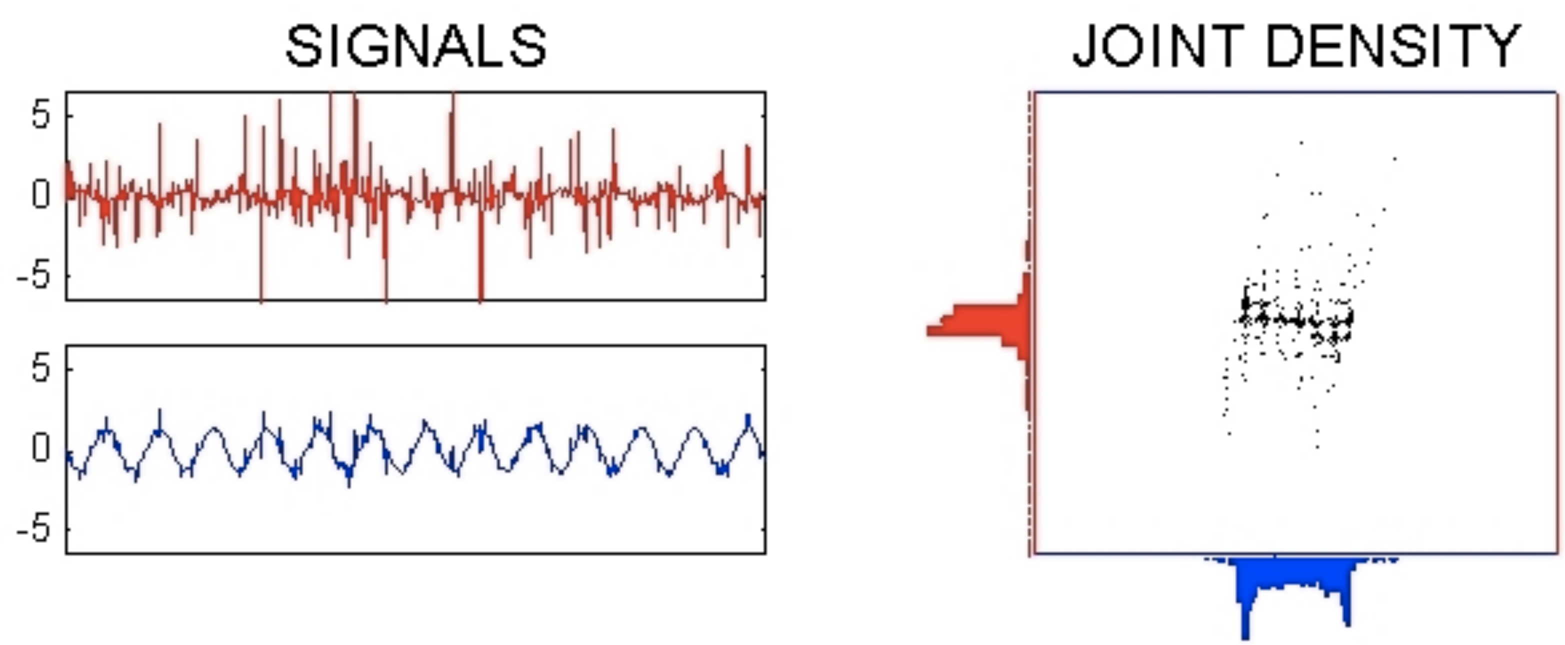
Input signals and density



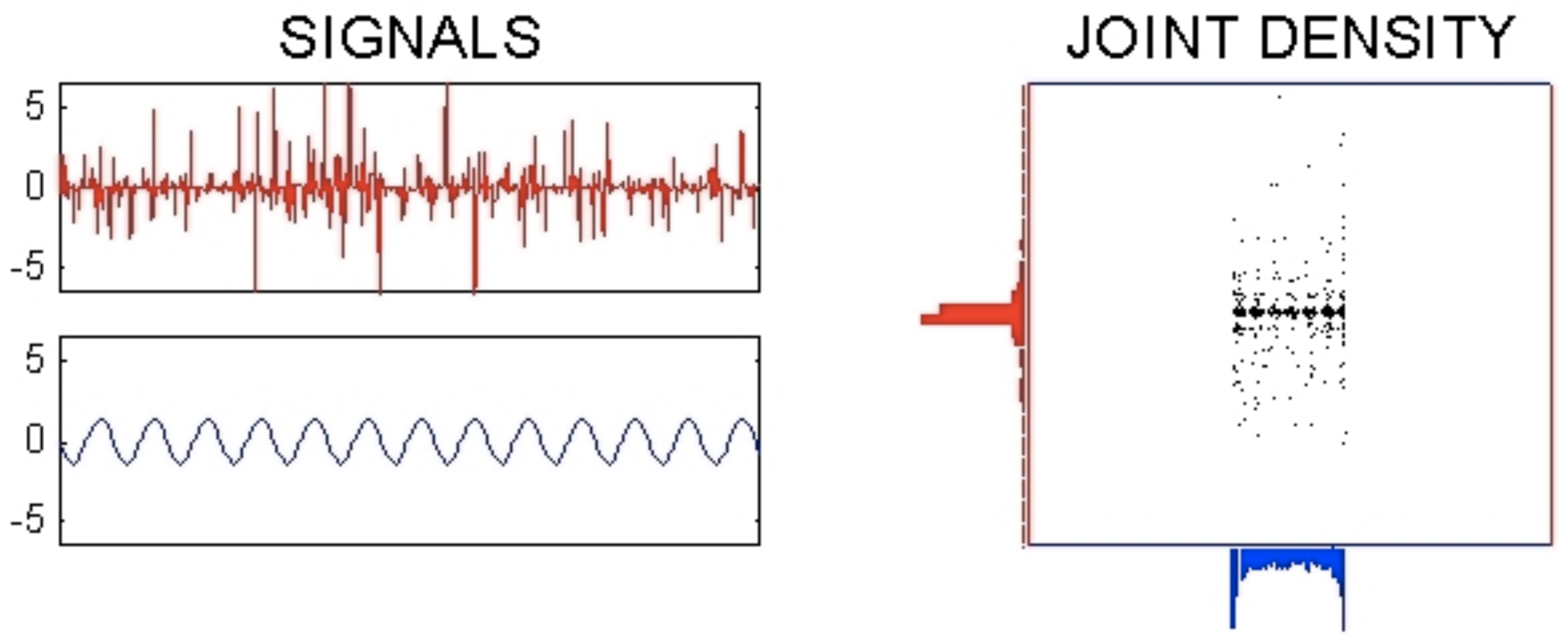
Whitened signals and density



Separated signals after 1 step of FastICA



Separated signals after 3 steps of FastICA



Separated signals after 5 steps of FastICA

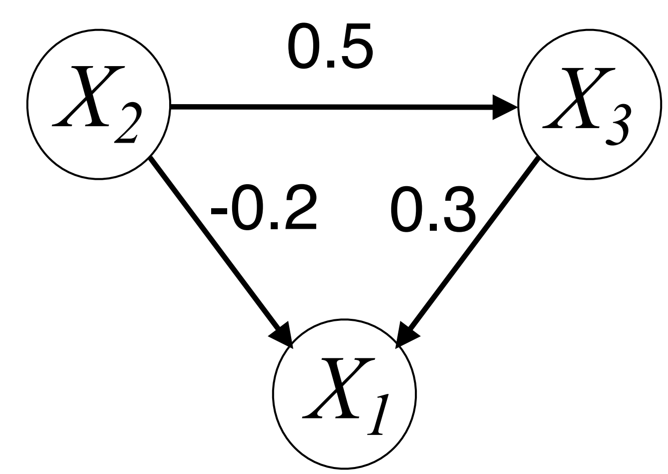
LiNGAM Analysis by ICA

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

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

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

So we have the causal relation:

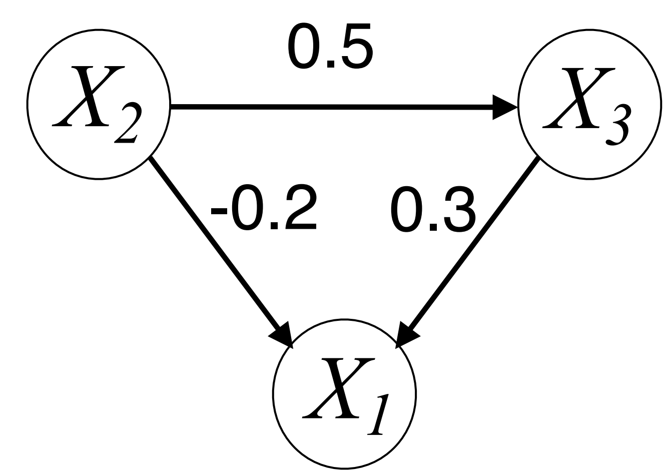


LiNGAM Analysis by ICA

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

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

So we have the causal relation:



Question 1. How to find W?

Question 2. How to see B from W?

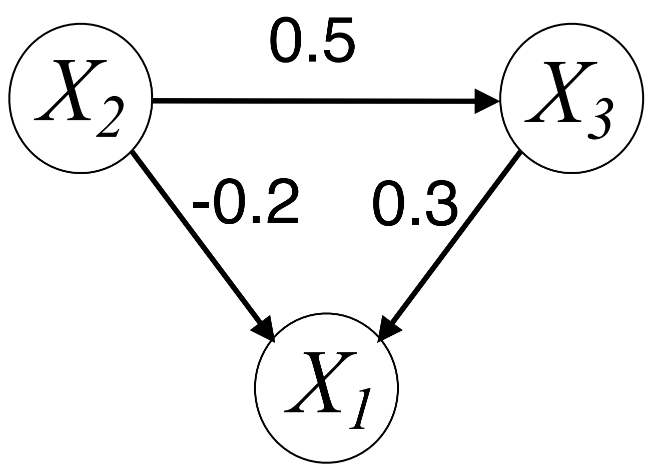
LiNGAM Analysis by ICA

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

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

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

$$\mathbf{W} = \begin{bmatrix} 0.6 & -0.4 & 2 & 0 \\ 1.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.5 \\ 1.5 & 3 & 0 & 0 \end{bmatrix}$$

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 $\ddot{\mathbf{W}}$ by its diagonal entry, giving $\ddot{\mathbf{W}}$'.
3. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

- Can we find the causal model?

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

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

$$\mathbf{W} = \begin{bmatrix} 0.6 & -0.4 & 2 & 0 \\ 1.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.5 \\ 1.5 & 3 & 0 & 0 \end{bmatrix}$$

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 $\ddot{\mathbf{W}}$ by its diagonal entry, giving $\ddot{\mathbf{W}}$'.
3. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

- Can we find the causal model?

$$\ddot{W} = \begin{pmatrix} 1.5 & 0 & 0 & 0 \\ 1.5 & 3 & 0 & 0 \\ 0.6 & -0.4 & 2 & 0 \\ 0 & 0.2 & 0 & 0.5 \end{pmatrix},$$

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

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

$$\mathbf{W} = \begin{bmatrix} 0.6 & -0.4 & 2 & 0 \\ 1.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.5 \\ 1.5 & 3 & 0 & 0 \end{bmatrix}$$

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 $\ddot{\mathbf{W}}$ by its diagonal entry, giving $\ddot{\mathbf{W}}$'.
3. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

- Can we find the causal model?

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

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

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

$$\mathbf{W} = \begin{bmatrix} 0.6 & -0.4 & 2 & 0 \\ 1.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.5 \\ 1.5 & 3 & 0 & 0 \end{bmatrix}$$

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 $\ddot{\mathbf{W}}$ by its diagonal entry, giving $\ddot{\mathbf{W}}$'.
3. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

- Can we find the causal model?

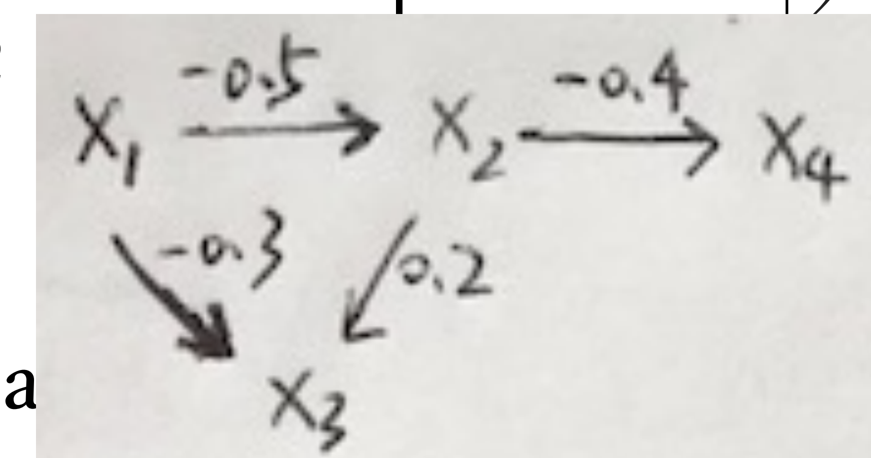
$$\ddot{W} = \begin{pmatrix} 1.5 & 0 & 0 & 0 \\ 1.5 & 3 & 0 & 0 \\ 0.6 & -0.4 & 2 & 0 \\ 0 & 0.2 & 0 & 0.5 \end{pmatrix}, \quad \ddot{W}' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.3 & -0.2 & 1 & 0 \\ 0 & 0.4 & 0 & 1 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -0.5 &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

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 $\ddot{\mathbf{W}}$'.
$\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

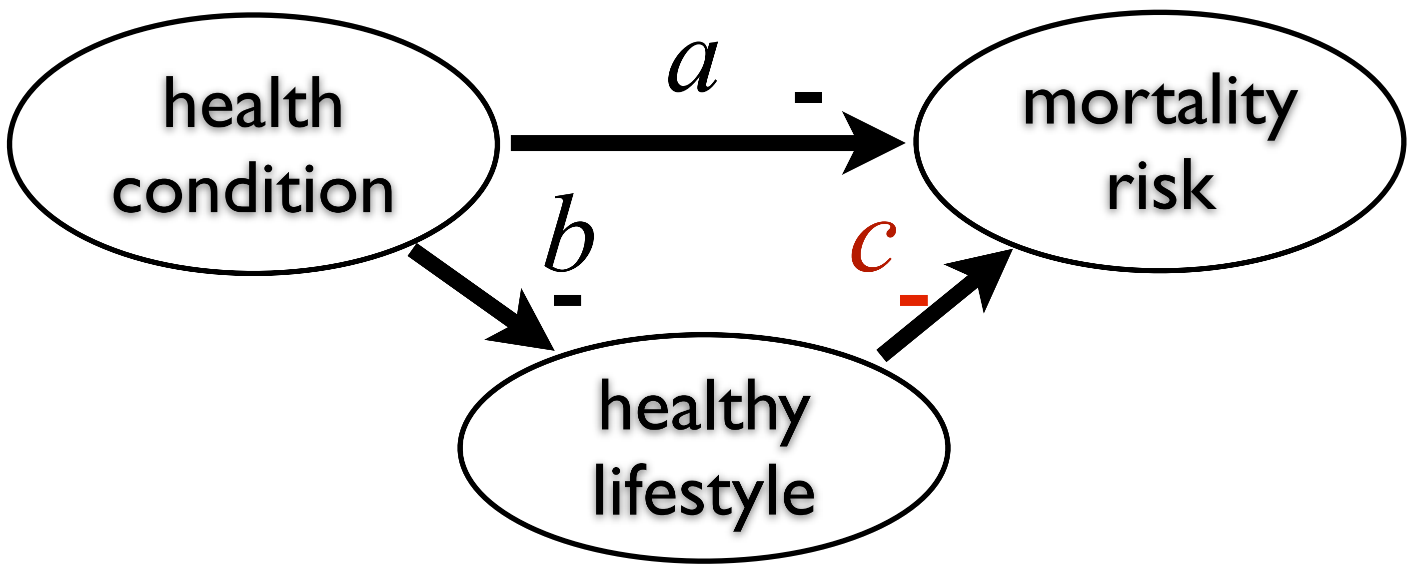


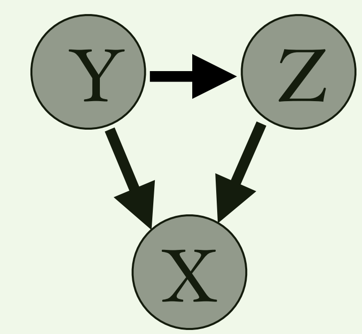
- Can we find the ca

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

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

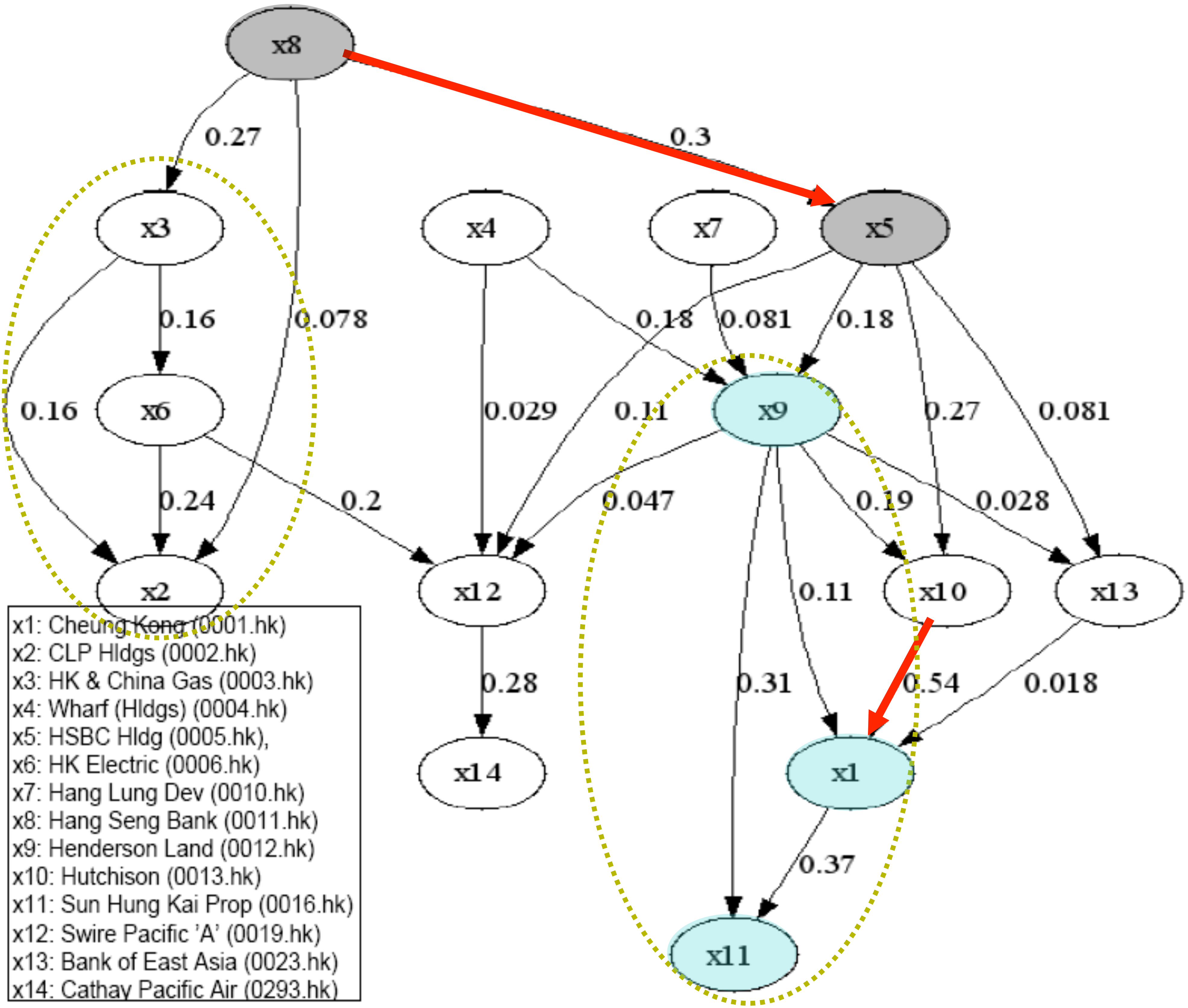
- ICA-LiNGAM
- ICA with Sparse Connections
- DirectLiNGAM...

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

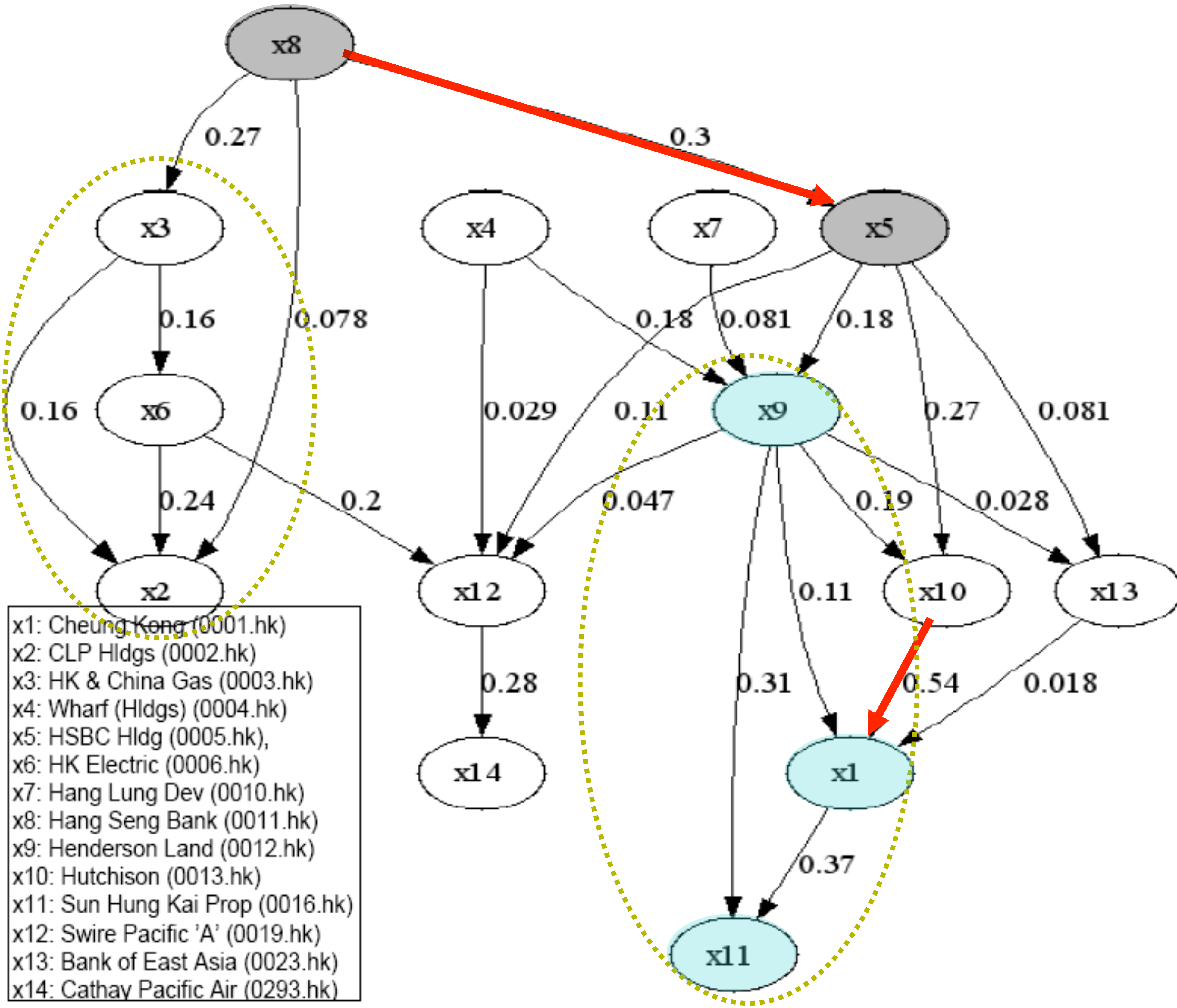
Zhang et al. (2006) ICA with sparse connections: Revisited. Lecture Notes in Computer Science, 5441:195–202, 2009

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)

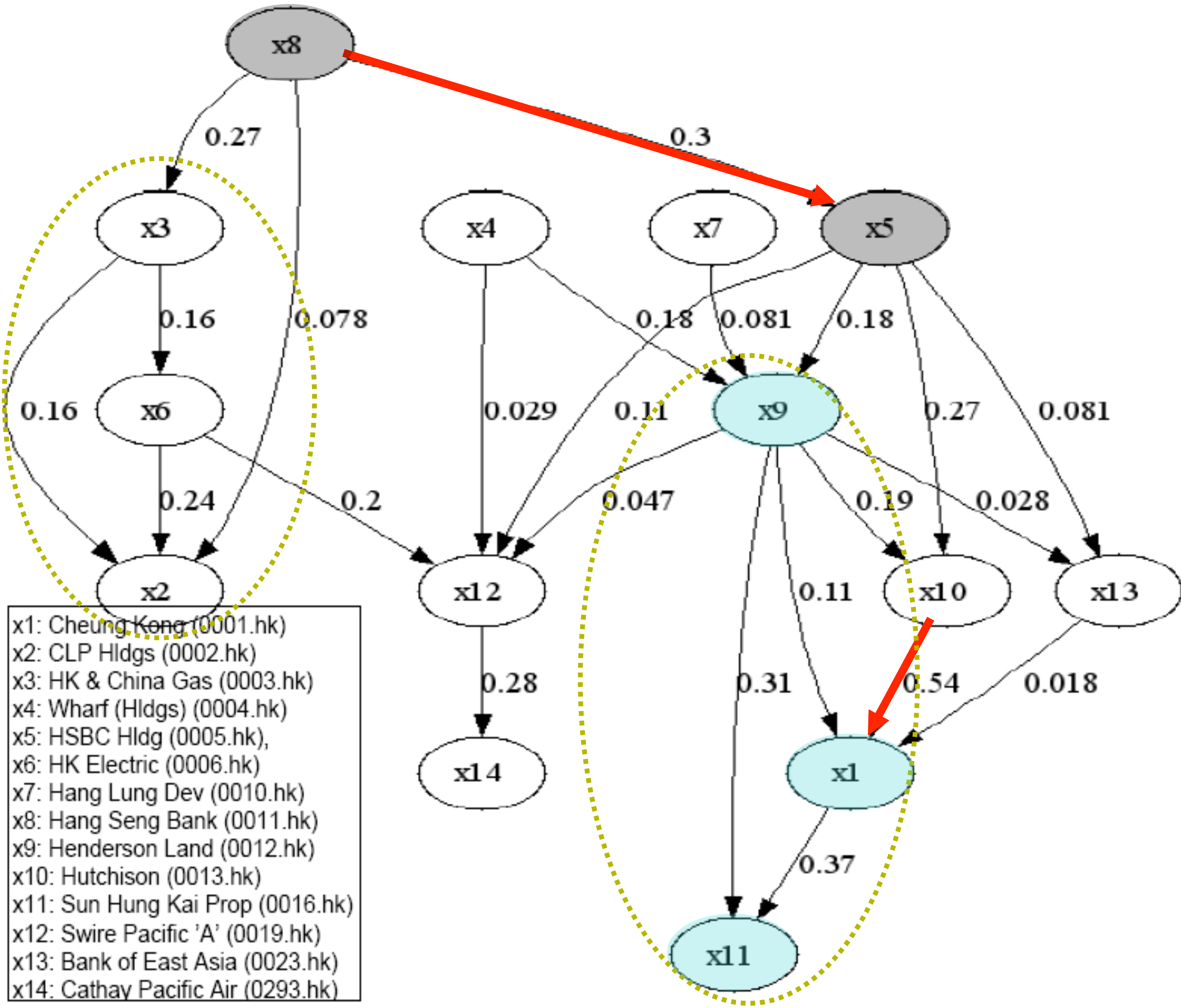


Application: Causal diagram in HK Stock Market (Zhang & Chan, 2006)



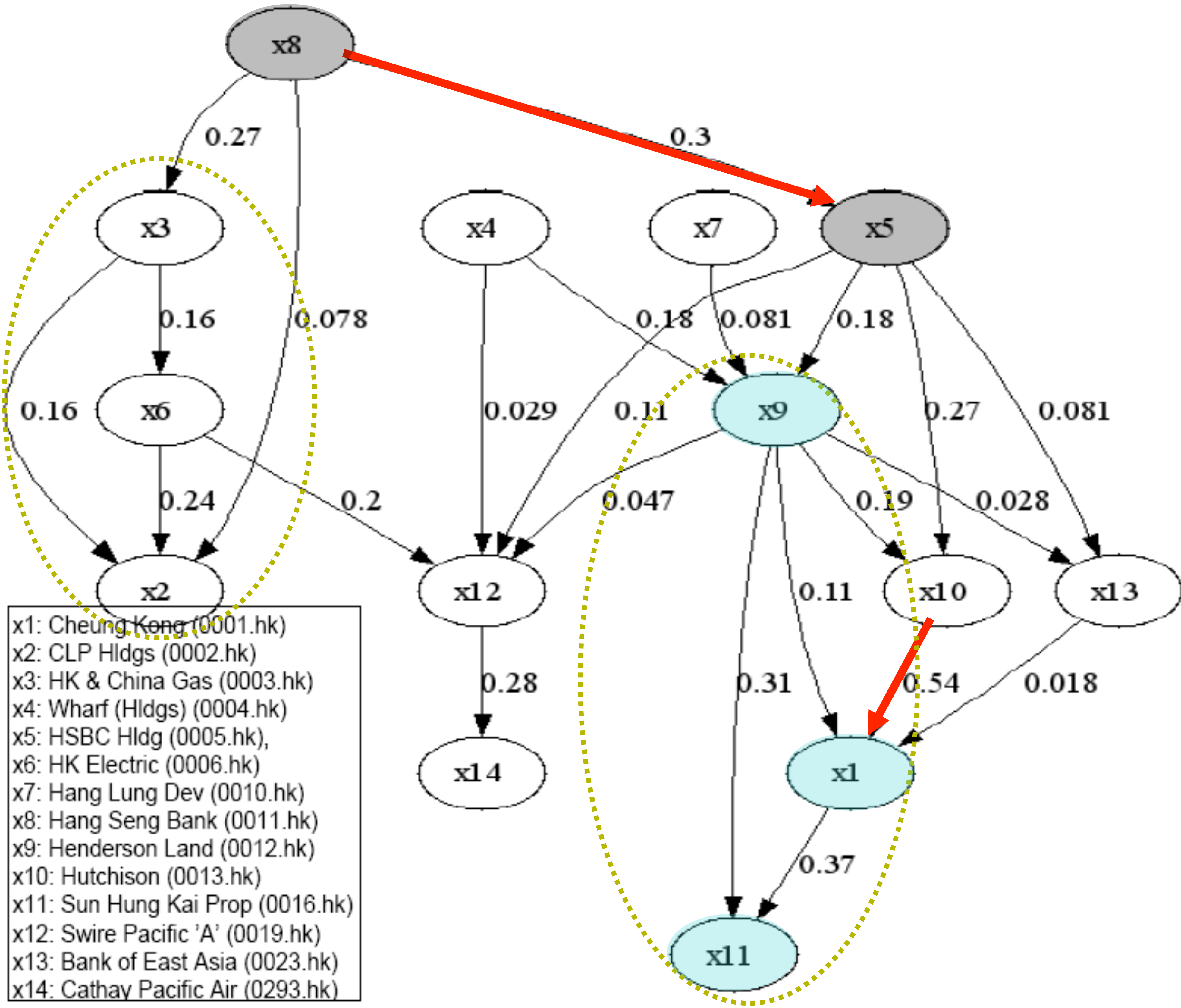
1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.

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



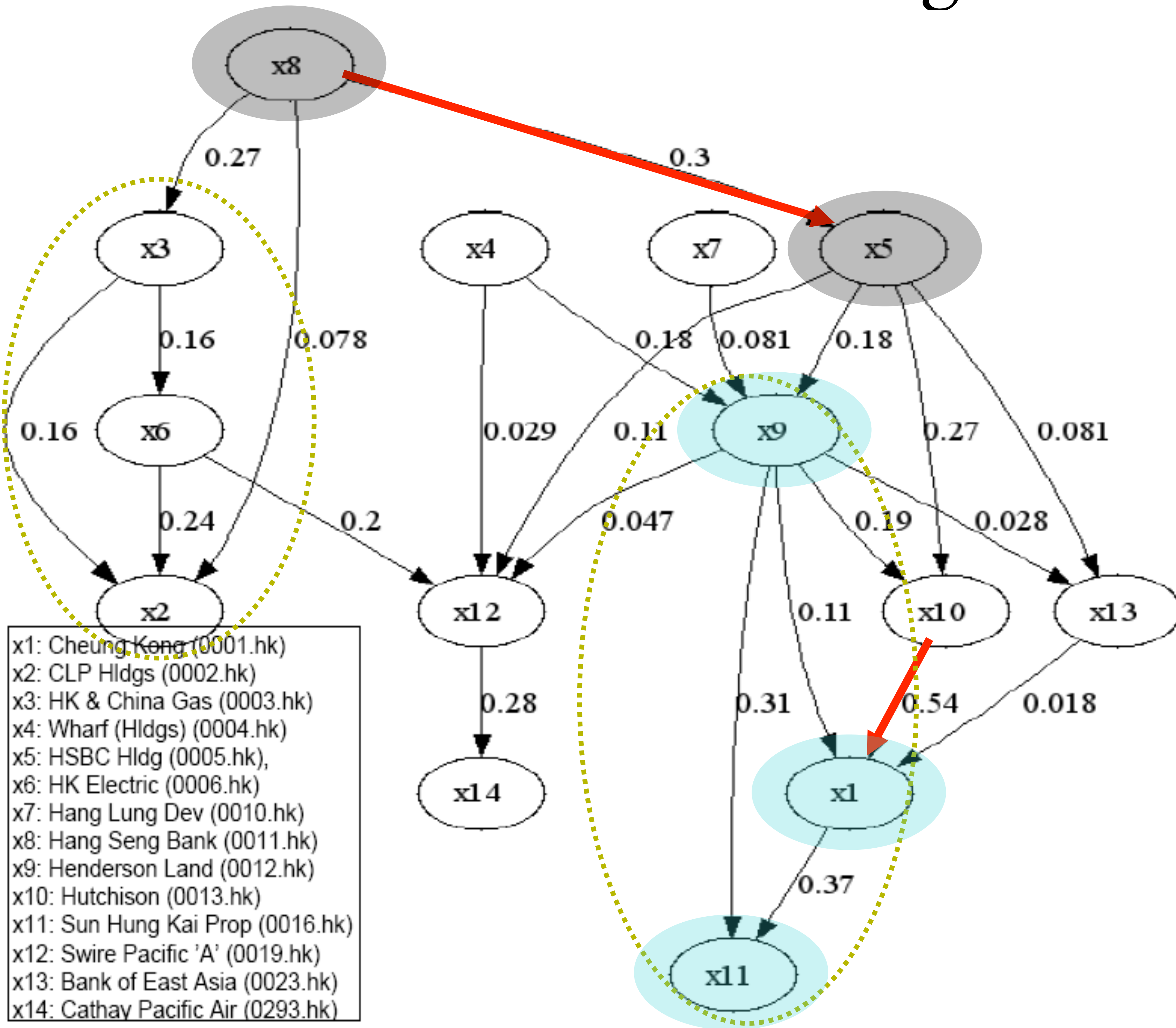
1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.
2. Stocks belonging to the same subindex tend to be connected.

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.

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

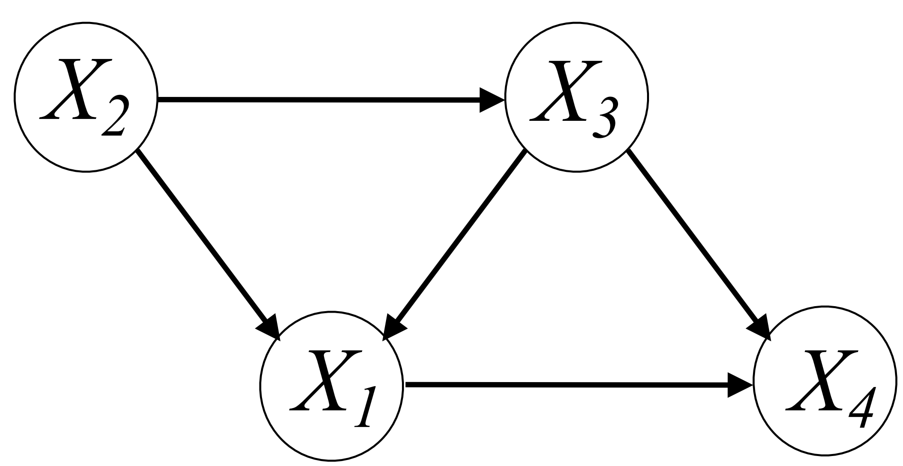


1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.
2. Stocks belonging to the same subindex tend to be connected.
3. Large bank companies (x5 and x8) are the cause of many stocks.
4. Stocks in Property Index (x1, x9, x11) depend on many stocks, while they hardly influence others.

Independent Noise (IN) Condition

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

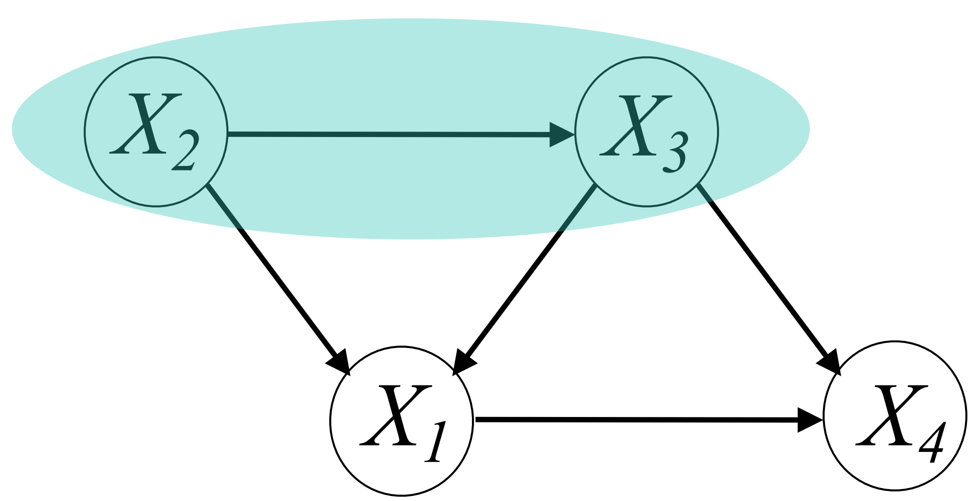
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), because $(\mathbf{Z}, Y)$ satisfies the IN condition iff
  - All variables in $\mathbf{Z}$ 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)



Independent Noise (IN) Condition

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

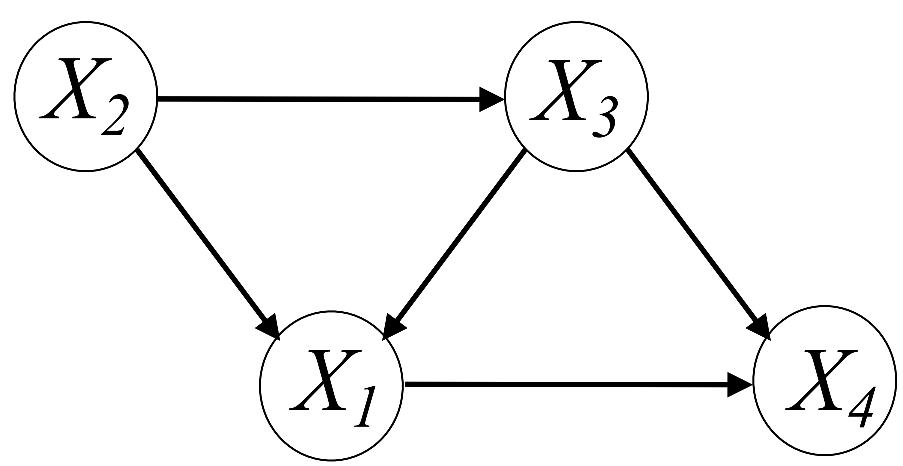
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), because $(\mathbf{Z}, Y)$ satisfies the IN condition iff
  - All variables in $\mathbf{Z}$ 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)



Independent Noise (IN) Condition

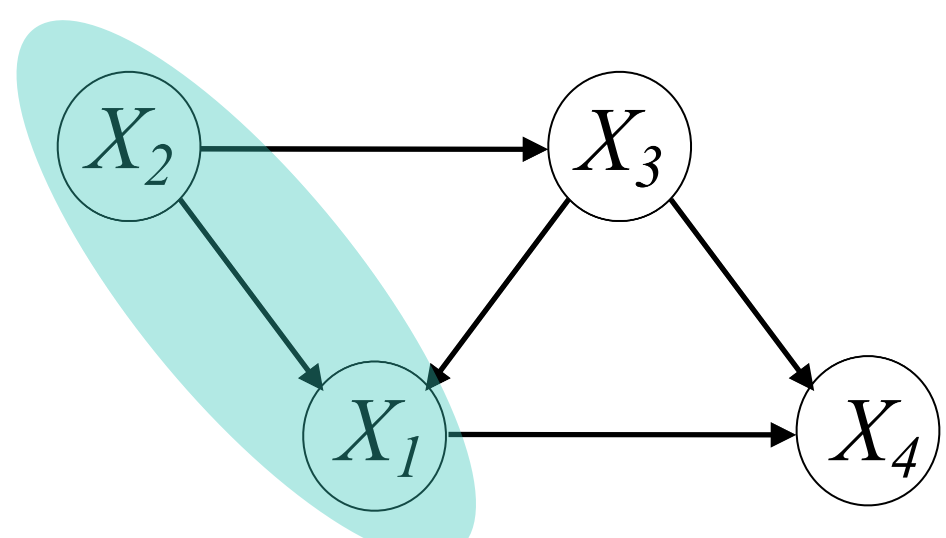
$$\mathbf{Z} \longrightarrow Y$$

- $(\mathbf{Z}, Y)$ 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)



Independent Noise (IN) Condition

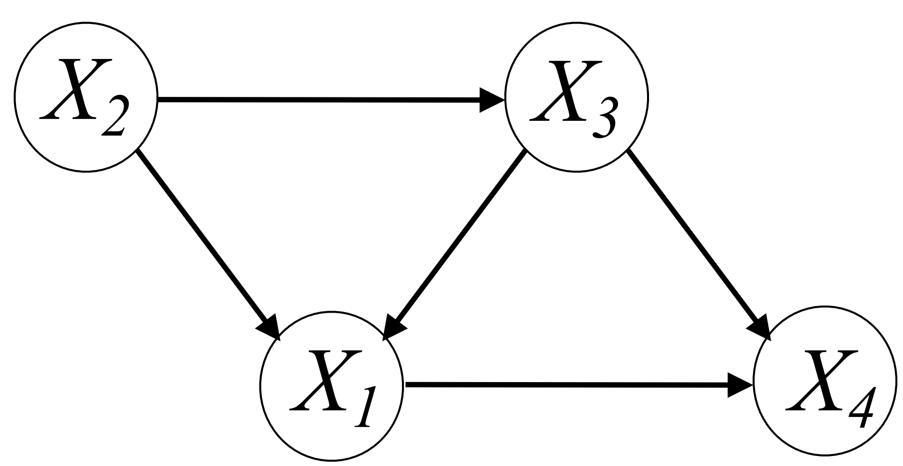
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), because $(\mathbf{Z}, Y)$ satisfies the IN condition iff
  - All variables in $\mathbf{Z}$ 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)



Independent Noise (IN) Condition

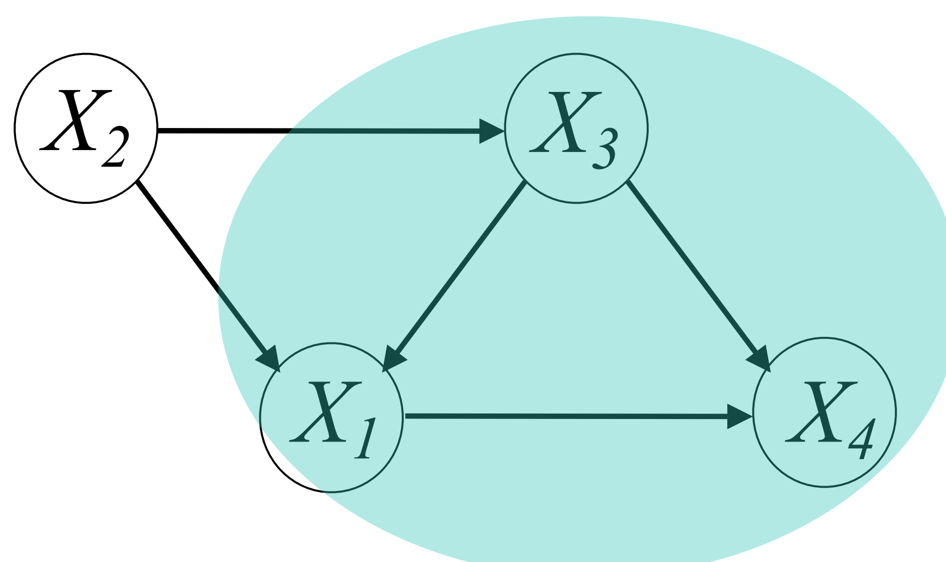
$$\mathbf{Z} \longrightarrow Y$$

- $(\mathbf{Z}, Y)$ 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)



Independent Noise (IN) Condition

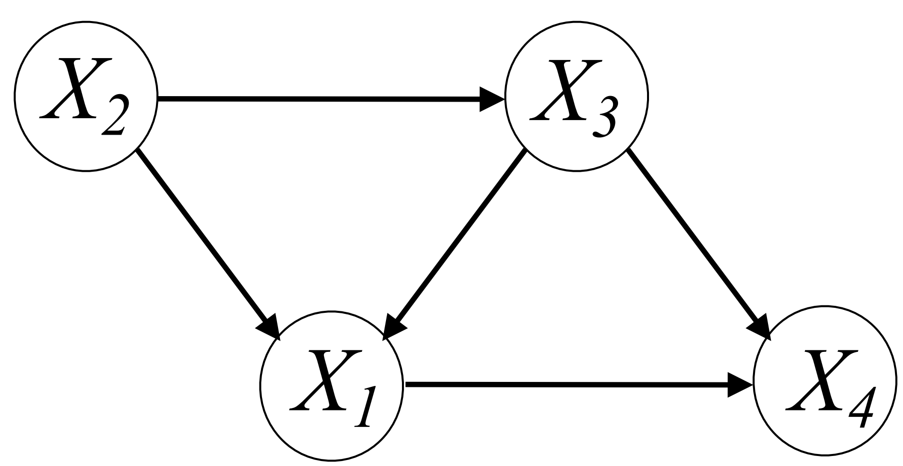
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), because $(\mathbf{Z}, Y)$ satisfies the IN condition iff
  - All variables in $\mathbf{Z}$ 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)



Independent Noise (IN) Condition

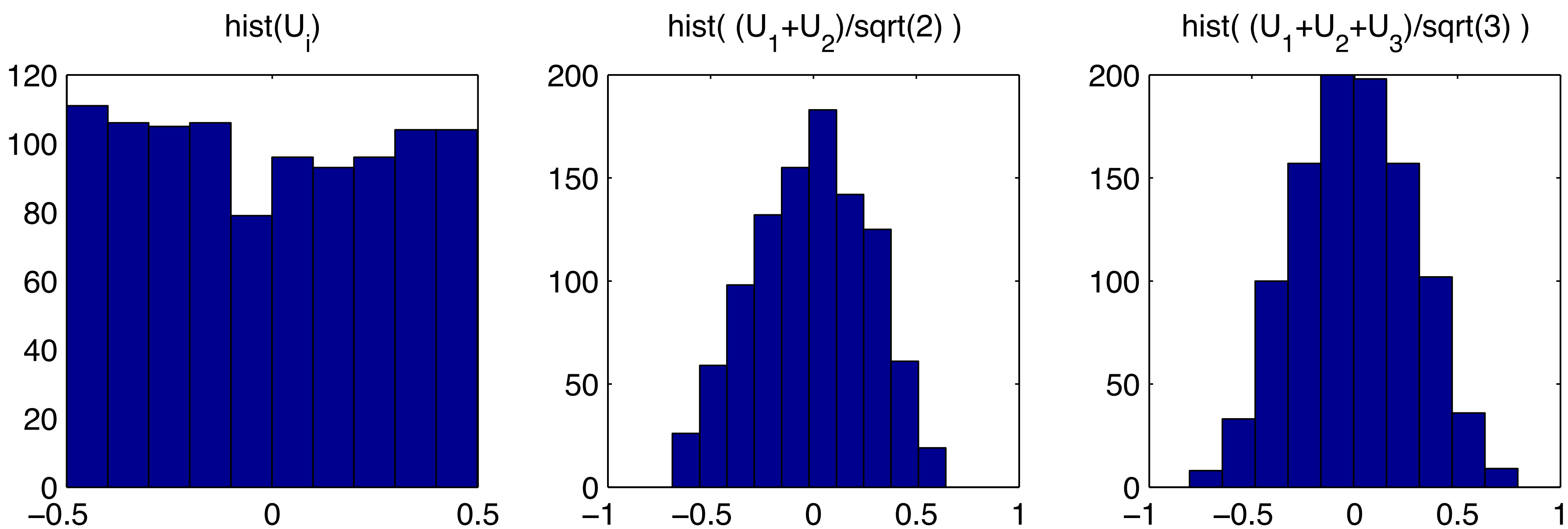
$$\mathbf{Z} \longrightarrow Y$$

- $(\mathbf{Z}, Y)$ 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)



Why Gaussianity Was 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.

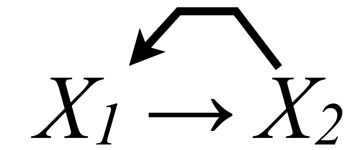
Gaussianity or Non-Gaussianity?

- 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...

- 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)
- 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)
- 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)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- 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)

Issue I: Feedback

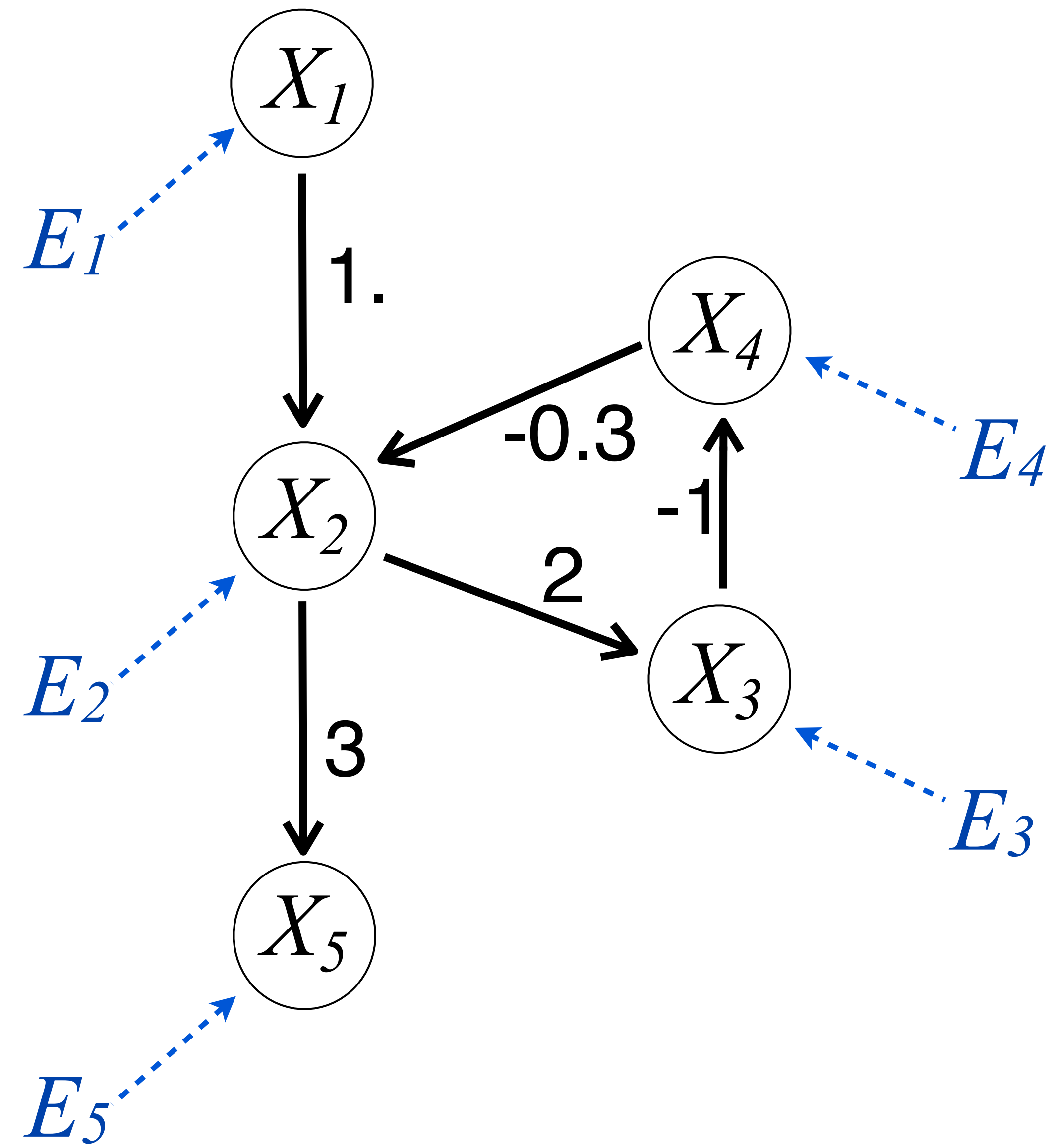


- Causal relations may have cycles; Consider an example

$$
\begin{aligned}
X_1 &= E_1 \\
X_2 &= 1.2X_1 - 0.3X_4 + E_2 \\
X_3 &= 2X_2 + E_3 \\
X_4 &= -X_3 + E_4 \\
X_5 &= 3X_2 + E_5
\end{aligned}
$$

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

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



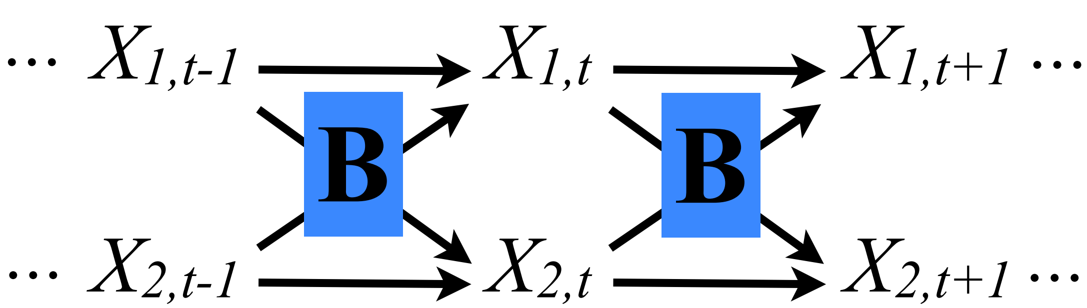
Lacerda, Spirtes, Ramsey and Hoyer (2008). Discovering cyclic causal models by independent component analysis. In Proc. UAI.

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?

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

- 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

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

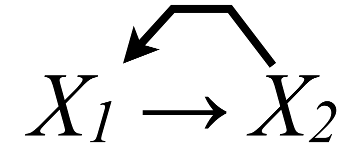
  - On temporally aggregated data

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

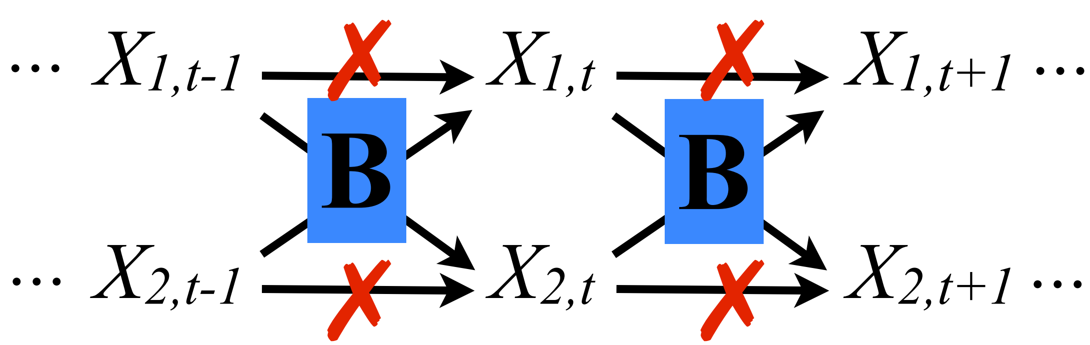
$$\frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{X}}_{t+k} = \mathbf{B}\frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{X}}_{t+k-1} + \frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{E}}_{t+k}.$$

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

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

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

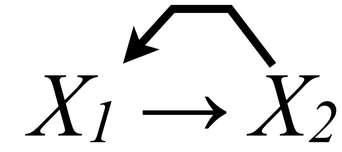
  - On temporally aggregated data

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

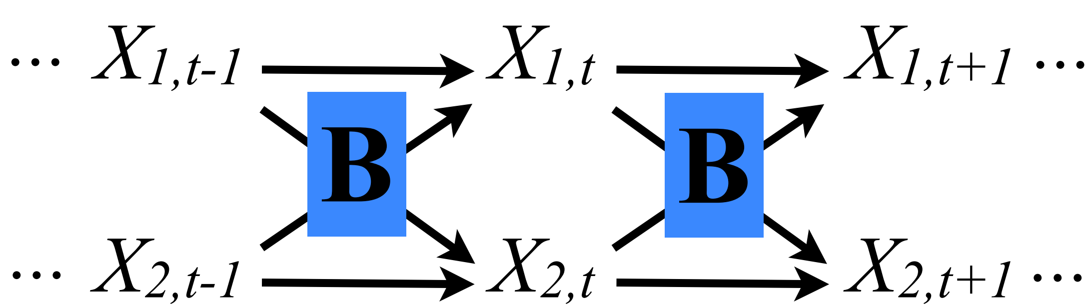
$$\frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{X}}_{t+k} = \mathbf{B}\frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{X}}_{t+k-1} + \frac{1}{L}\sum_{k=1}^{L}\tilde{\mathbf{E}}_{t+k}.$$

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:

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

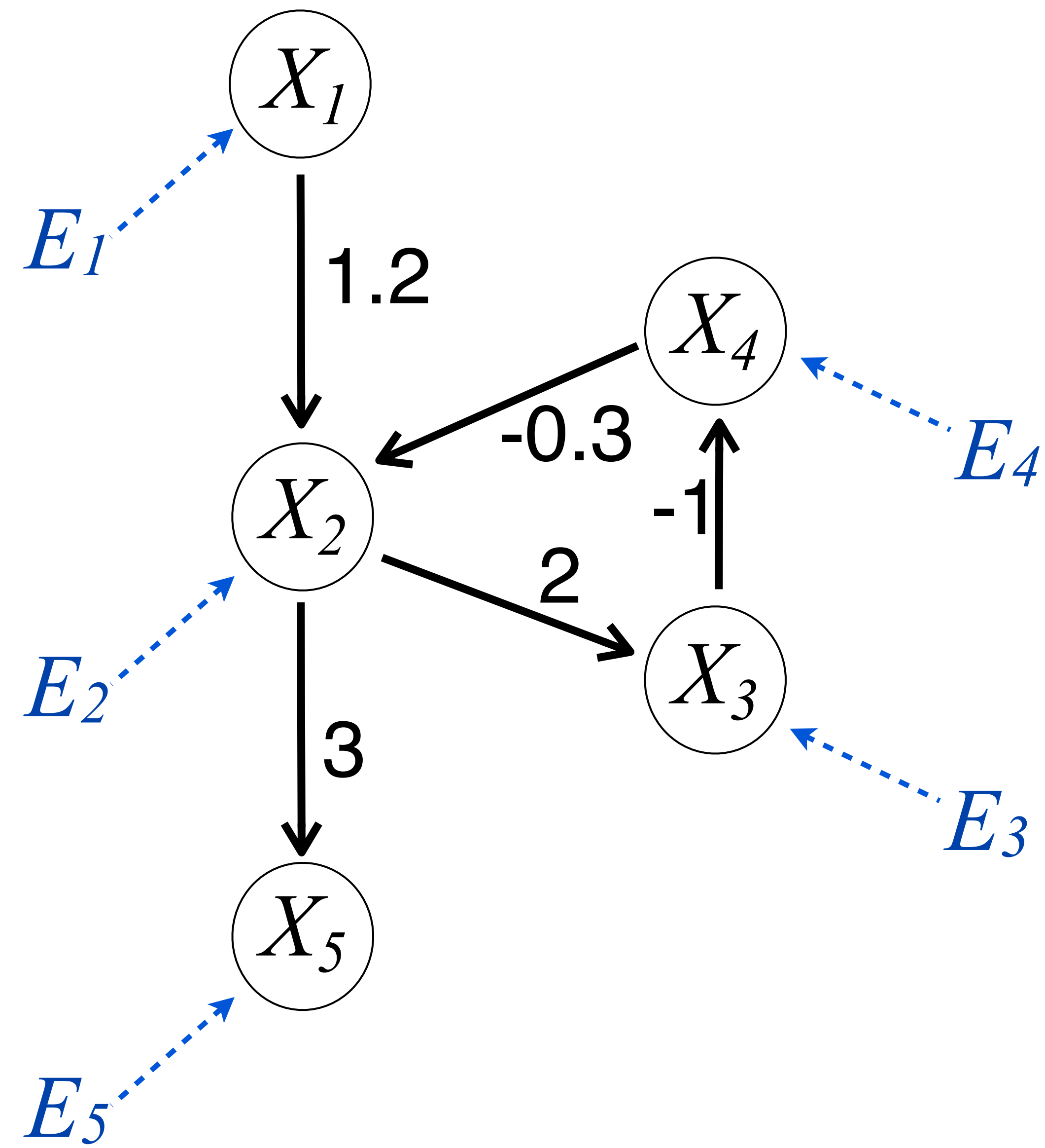
  - On temporally aggregated data

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.

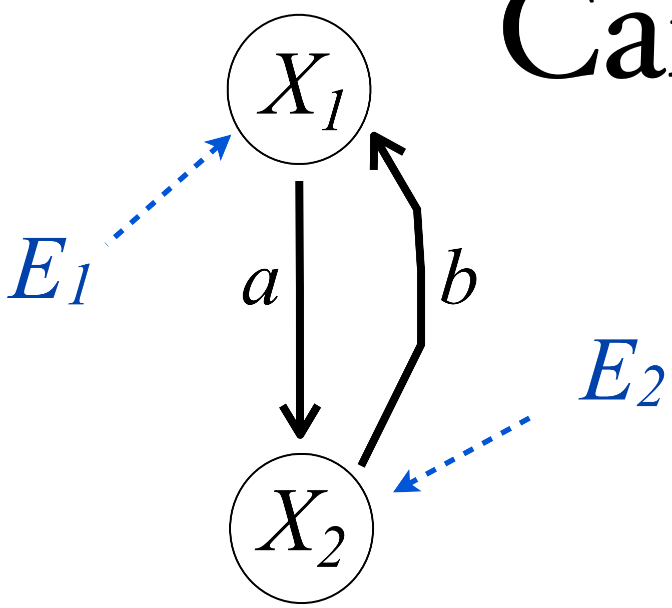
Cyclic Model: Global or Local Markov Condition?

- Local Markov condition?
- Global Markov condition?
  - Linear case?
  - General nonlinear case?



P. Spirtes, Directed Cyclic Graphical Representations of Feedback Models, UAI 1995

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,

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

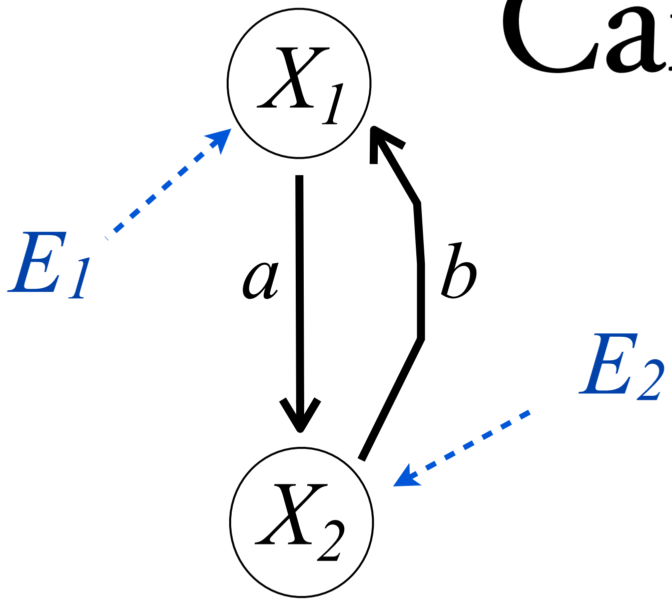
$$\Rightarrow \underbrace{\begin{bmatrix} -a & 1 \\ 1 & -b \end{bmatrix}}_{\mathbf{W'}} \mathbf{X}_t = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \cdot \mathbf{E}_t$$

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

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

- **E = (I-B)X**; ICA gives **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 :-)

Can We Recover Cyclic Relations?



- **E = (I-B)X**; ICA gives **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 :-)

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,

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

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

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

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

Summary:

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

Summary: LiNGAM

- We started making use of additional (plausible?) assumptions about causal mechanisms
  - Linear models with non-Gaussian noise
- Methods for estimating linear non-Gaussian causal models
- Difference between Linear, non-Gaussian and linear-Gaussian models
- Next: Interpretation and estimation of cyclic models