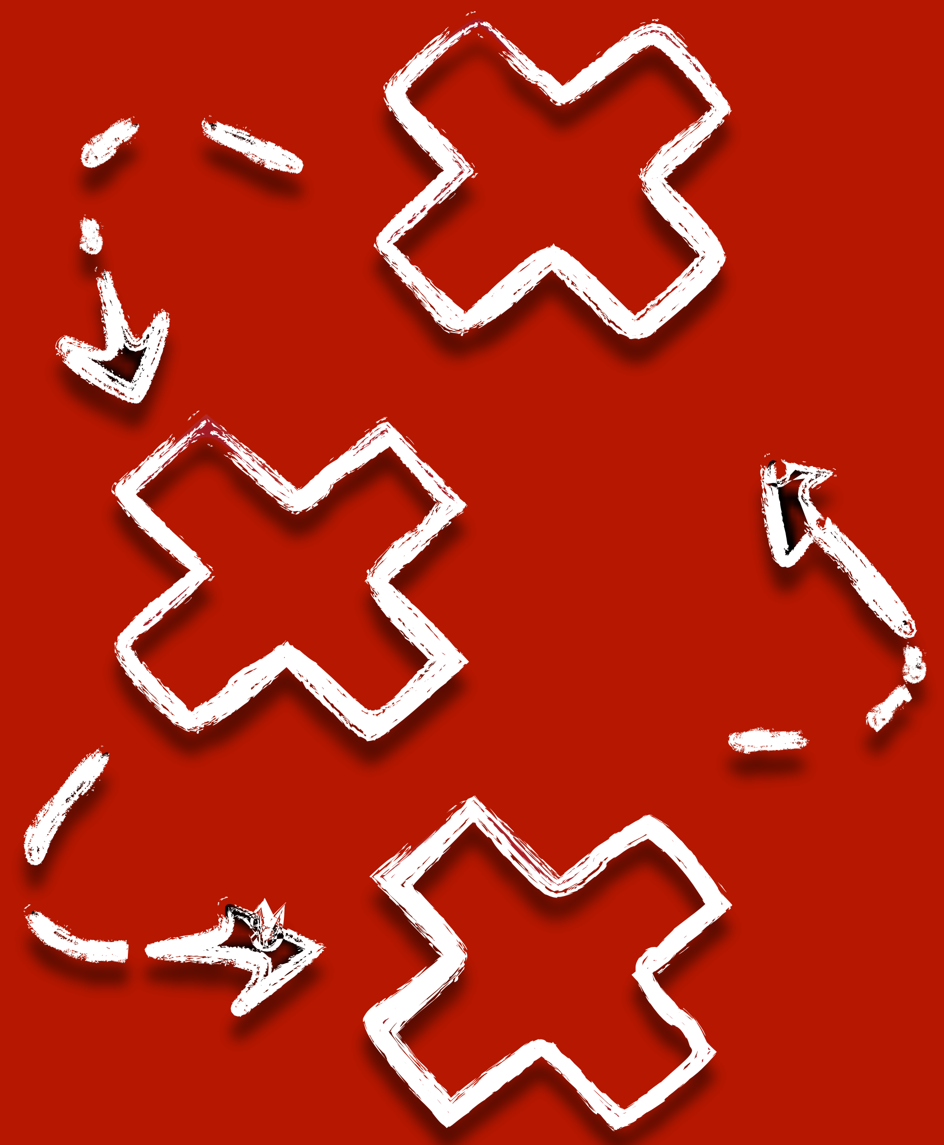


Causal Data Science

Lecture 10.2: Restricted models

Lecturer: Sara Magliacane

UvA - Spring 2024

Score-based causal discovery vs restricted models

- Similarly to constraint-based causal discovery, also GES returns a CPDAG
- Identification strategies are based on a known DAG
    - Add **background knowledge**
    - Add **experimental/interventional data (next week)**
    - Use advanced methods like IDA (Maathuis et al 2009) that combine the identification of each possible DAG in MEC into bounds

- Under some special assumptions, **we can recover the true causal graph:**
    - Nonlinear models with additive noise: Additive Noise Models (ANM)
    - Linear Non-Gaussian Acyclic Models (LINGAM)

Causal discovery simplified overview

Constraint-based causal discovery

- Conditional independence tests
- Observational data
- Output: MEC
- SGS, PC, FCI

Score-based causal discovery

- Penalised likelihood
- Observational data
- Output: MEC
- GES, MMHC

Restricted models

- Nonlinear additive noise, Linear Non-Gaussianity
- Observational data
- Output: DAG
- RESIT, LINGAM

Interventional causal discovery / causal invariance

- Observational and Interventional data
- Output: parents of Y, I-MEC
- ICP, GIES, JCI

Additive noise models (ANMs)

$$\begin{cases} X = \varepsilon_X \\ Y = f(X) + \varepsilon_Y \end{cases}$$

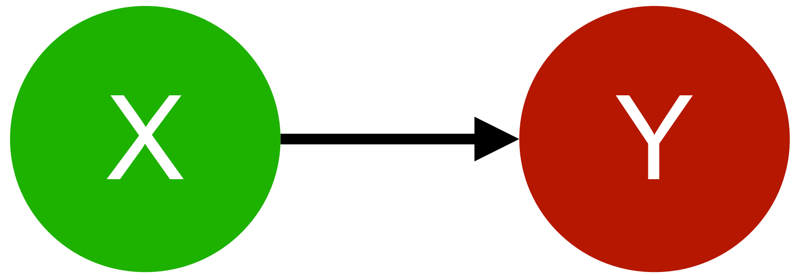
$$\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$$

Here ε_i are not gaussian, but rather uniform

http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf Chapter 4

Additive noise models (ANMs)

$$\begin{cases} X = \varepsilon_X \\ Y = f(X) + \varepsilon_Y \end{cases}$$



$$\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$$

Here ε_i are not gaussian, but rather uniform

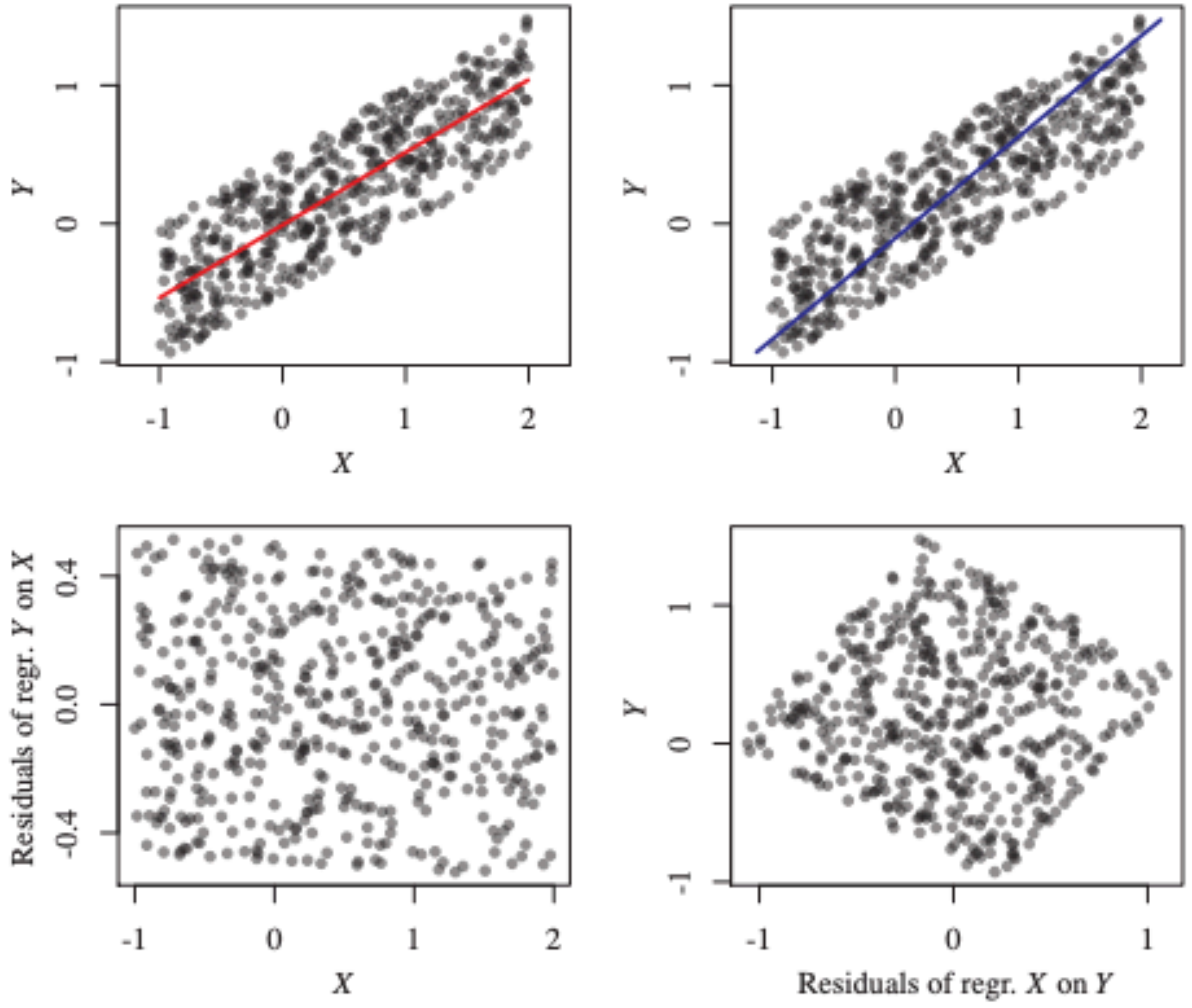


Figure 4.5: We are given a sample from the underlying distribution and perform a linear regression in the directions $X \to Y$ (left) and $Y \to X$ (right). The fitted functions are shown in the top row, the corresponding residuals are shown in the bottom row. Only the direction $X \to Y$ yields independent residuals; see also Figure 4.1.

http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf Chapter 4

Linear models with additive Gaussian noise

- If we have the linear SCM

$$Y = \alpha X + \epsilon_Y \text{ such that } \epsilon_Y \perp\!\!\!\perp X$$

Then there exists a $\beta \in \mathbb{R}$ and random variable ϵ_X such that:

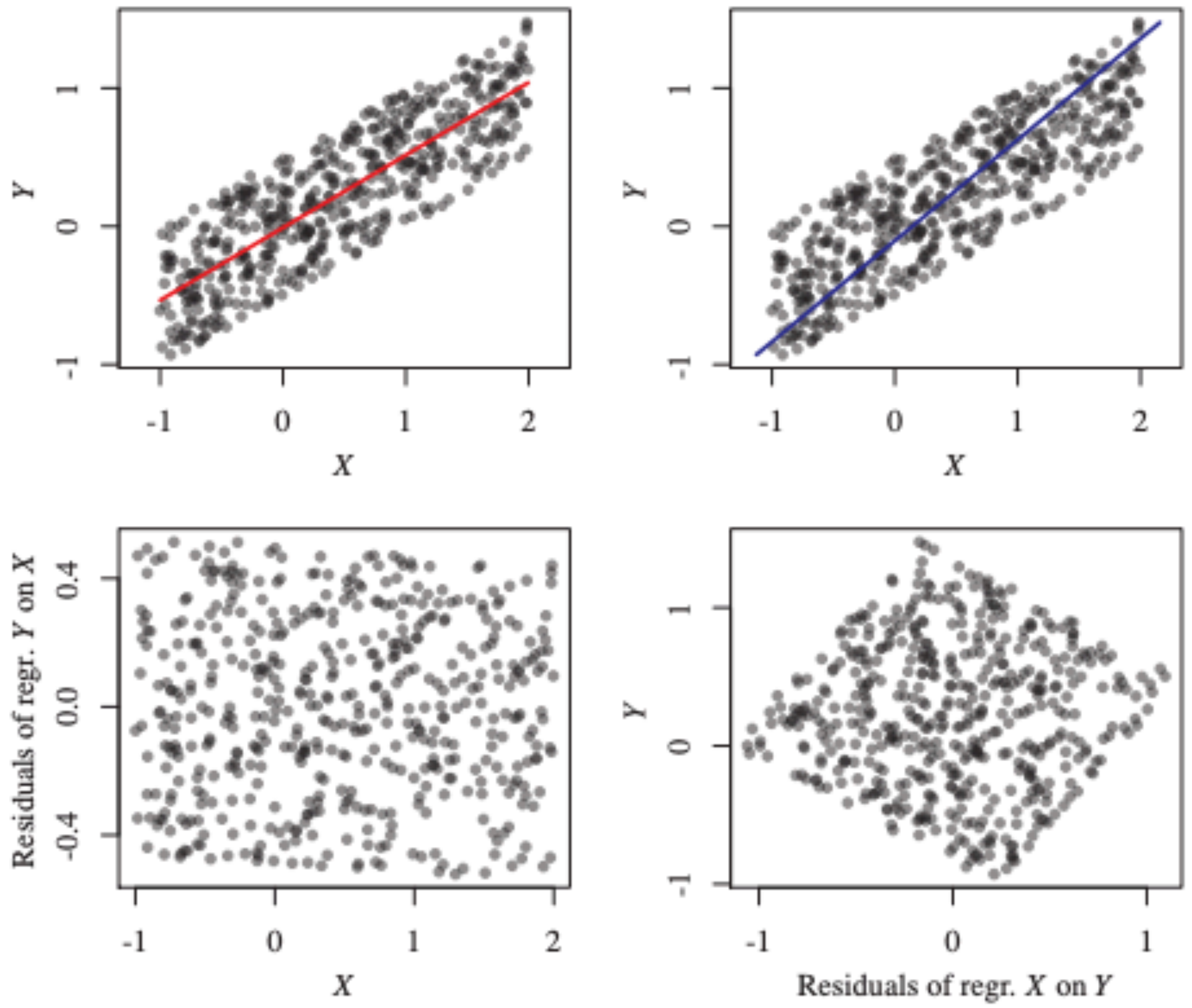
$$X = \beta Y + \epsilon_X \text{ such that } \epsilon_X \perp\!\!\!\perp Y$$

if and only if ϵ_Y and X are Gaussian

This means we cannot distinguish $X \to Y$ or $Y \to X$!

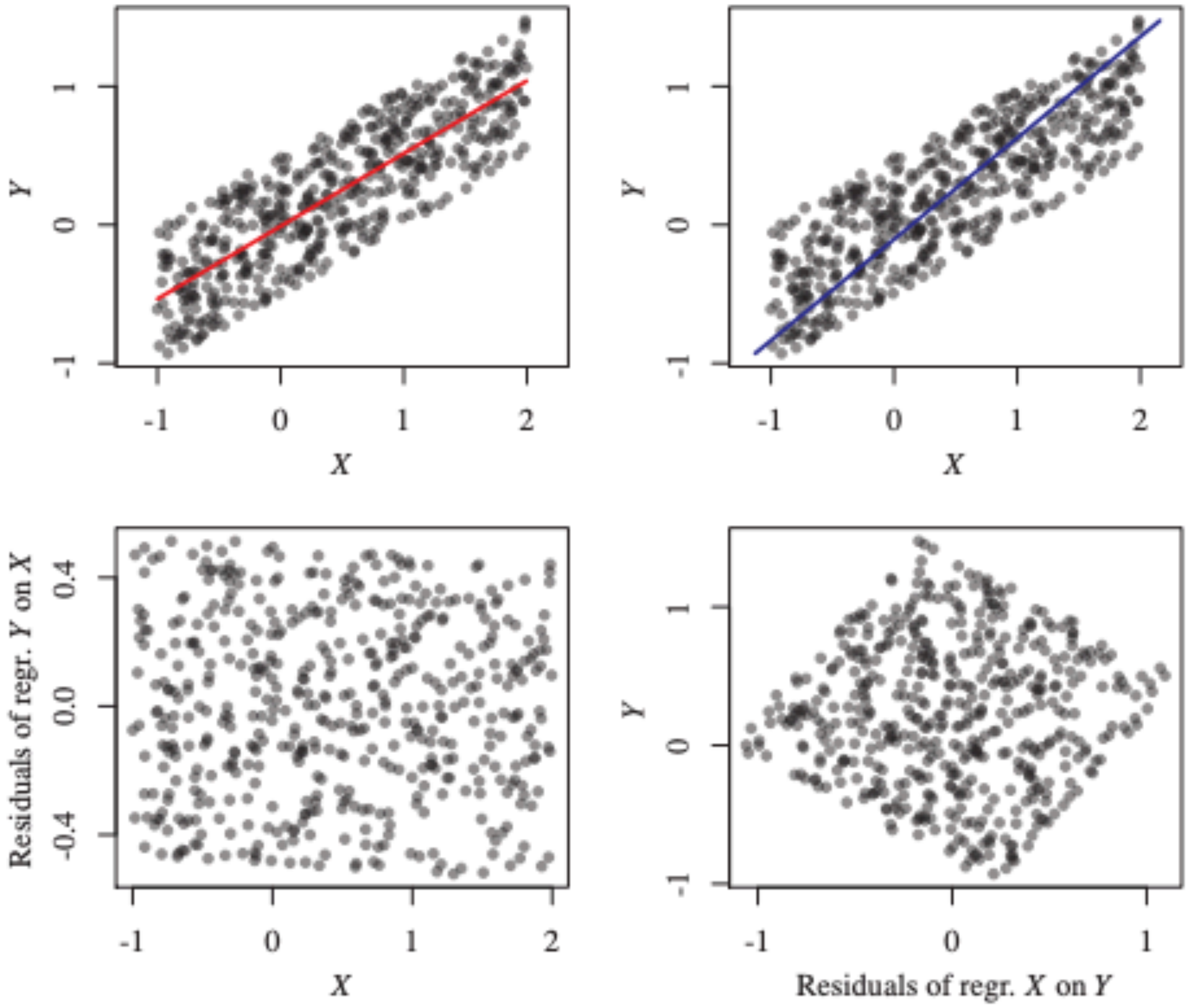
RESIT: regression with subsequent independence test

1. Regress X on Y with (possibly nonlinear) regression and estimate $\hat{f}_Y(X)$
2. Test if $(Y - \hat{f}_Y(X))$ is independent of X



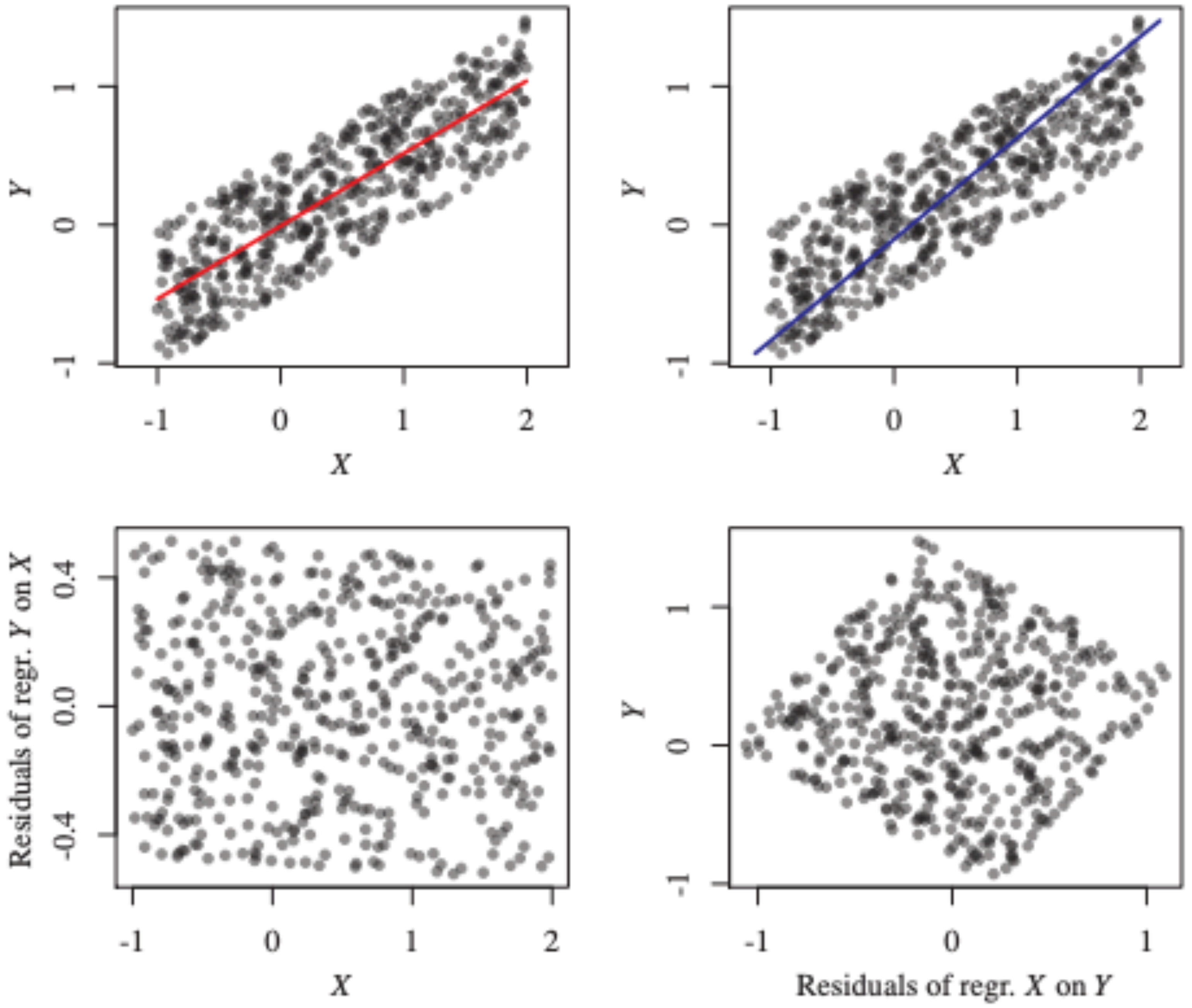
RESIT: regression with subsequent independence test

1. Regress X on Y with (possibly nonlinear) regression and estimate $\hat{f}_Y(X)$
2. Test if $(Y - \hat{f}_Y(X))$ is independent of X
3. Regress Y on X with (possibly nonlinear) regression and estimate $\hat{f}_X(Y)$
4. Test if $(X - \hat{f}_X(Y))$ is independent of Y



RESIT: regression with subsequent independence test

1. Regress X on Y with (possibly nonlinear) regression and estimate $\hat{f}_Y(X)$
2. Test if $(Y - \hat{f}_Y(X))$ is independent of X
3. Regress Y on X with (possibly nonlinear) regression and estimate $\hat{f}_X(Y)$
4. Test if $(X - \hat{f}_X(Y))$ is independent of Y



5. If independence is rejected in only one direction, the other **independent** direction is **causal**

Extensions

$$\begin{cases} X = \varepsilon_X \\ Y = f(X) \underline{+ \varepsilon_Y} \end{cases}$$

$$\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$$

$$\begin{cases} X = \varepsilon_X \\ Y = g(f(X) + \varepsilon_Y) \end{cases}$$

POST-LINEAR

For more details check Chapter 4 in the book: http://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf

Linear Non Gaussian Acyclic Models (LINGAM)

- We can write a linear SCM in matrix notation:

$$X = \mathbf{B}X + \varepsilon \ \text{ with } \mathbf{B} \in \mathbb{R}^{p \times p}, X \in \mathbb{R}^p, \varepsilon \in \mathbb{R}^p$$

- For Gaussian noise, we cannot distinguish the direction, but **for non-Gaussian noises we can**
  - Assume they are mean zero non Gaussian with positive variance
  - **We don't need faithfulness!** (So it can work on cancelling paths, etc)

Linear Non Gaussian Acyclic Models (LINGAM)

"position"

- For a DAG G, a bijective function $\pi : \{1,\ldots,p\} \to \{1,\ldots,p\}$ is a (not necessarily unique) **causal ordering** if, for all $i, j \in \{1,\ldots,p\}$:

$$\pi(i) < \pi(j) \text{ if } j \in \mathrm{Desc}(i)$$

Linear Non Gaussian Acyclic Models (LINGAM)

"position"

- For a DAG G, a bijective function $\pi : \{1,\ldots,p\} \to \{1,\ldots,p\}$ is a (not necessarily unique) **causal ordering** if, for all $i, j \in \{1,\ldots,p\}$:

$$\pi(i) < \pi(j) \text{ if } j \in \mathrm{Desc}(i)$$

$1 \to 2 \to 3$

$\pi(1) = 1$
$\pi(2) = 2$
$\pi(3) = 3$

$3 \to 2 \to 1$

$\pi(3) = 1$
$\pi(2) = 2$
$\pi(1) = 3$

BECAUSE ACYCLICITY AT LEAST ONE CAUSAL ORDERING EXISTS.

Linear Non Gaussian Acyclic Models (LINGAM)

"position"

- For a DAG G, a bijective function $\pi : \{1,\ldots,p\} \to \{1,\ldots,p\}$ is a (not necessarily unique) **causal ordering** if, for all $i, j \in \{1,\ldots,p\}$:

$$\pi(i) < \pi(j) \text{ if } j \in \mathrm{Desc}(i)$$

$1 \to 2 \to 3$

$\pi(1) = 1$
$\pi(2) = 2$
$\pi(3) = 3$

$3 \to 2 \to 1$

$\pi(3) = 1$
$\pi(2) = 2$
$\pi(1) = 3$

$1 \to 2$, $1 \to 3$

1) 1) 2) 3) 3) 2)

$\pi(1) = 1$
$\pi(2) = 2$
$\pi(3) = 3$

$\pi(1) = 1$
$\pi(2) = 3$
$\pi(3) = 2$

BECAUSE ACYCLICITY AT LEAST ONE CAUSAL ORDERING EXISTS.

Linear Non Gaussian Acyclic Models (LINGAM)

- We can write a linear SCM in matrix notation:

$$X = \mathbf{B}X + \varepsilon \text{ with } \mathbf{B} \in \mathbb{R}^{p \times p}, X \in \mathbb{R}^{p}, \varepsilon \in \mathbb{R}^{p}$$

- Because of acyclicity we can show that we can rewrite $\mathbf{B}$ as **strictly lower triangular** by **permuting the variables using a causal ordering**

$$B_\pi = \begin{bmatrix} 0 & & & & & \\ X & 0 & & & 0 & \\ X & X & 0 & & & \\ X & X & X & 0 & & \\ X & X & X & X & 0 & \\ X & X & X & X & X & 0 \end{bmatrix}$$

$$\begin{tikzcd} 1 \arrow[r, "B_{21}"] \arrow[rr, bend right, "B_{32}"'] & 2 \arrow[r, "B_{31}"] & 3 \end{tikzcd}$$

$$B = \begin{bmatrix} 0 & 0 & 0 \\ B_{21} & 0 & 0 \\ B_{31} & B_{32} & 0 \end{bmatrix}$$

Linear Non Gaussian Acyclic Models (LINGAM)

- We can write a linear SCM in matrix notation:

$$X = \mathbf{B}X + \varepsilon \text{ with } \mathbf{B} \in \mathbb{R}^{p \times p}, X \in \mathbb{R}^p, \varepsilon \in \mathbb{R}^p$$

- Because of acyclicity we can show that we can rewrite $\mathbf{B}$ as **strictly lower triangular** by **permuting the variables using a causal ordering**

$$\begin{tikzcd} 3 \arrow[r] \arrow[rr, bend right] & 1 \arrow[r] & 2 \end{tikzcd}$$

$$B = \begin{bmatrix} 0 & 0 & B_{13} \\ B_{21} & 0 & B_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

$$\pi(3) = 1$$
$$\pi(1) = 2$$
$$\pi(2) = 3$$

$$B_\pi = \begin{array}{cc} \begin{array}{ccc} X_3 & X_1 & X_2 \end{array} & \\ \begin{bmatrix} 0 & 0 & 0 \\ B_{13} & 0 & 0 \\ B_{23} & B_{21} & 0 \end{bmatrix} & \begin{array}{c} X_3 \\ X_1 \\ X_2 \end{array} \end{array}$$

Linear Non Gaussian Acyclic Models (LINGAM)

- We can write a linear SCM in matrix notation:

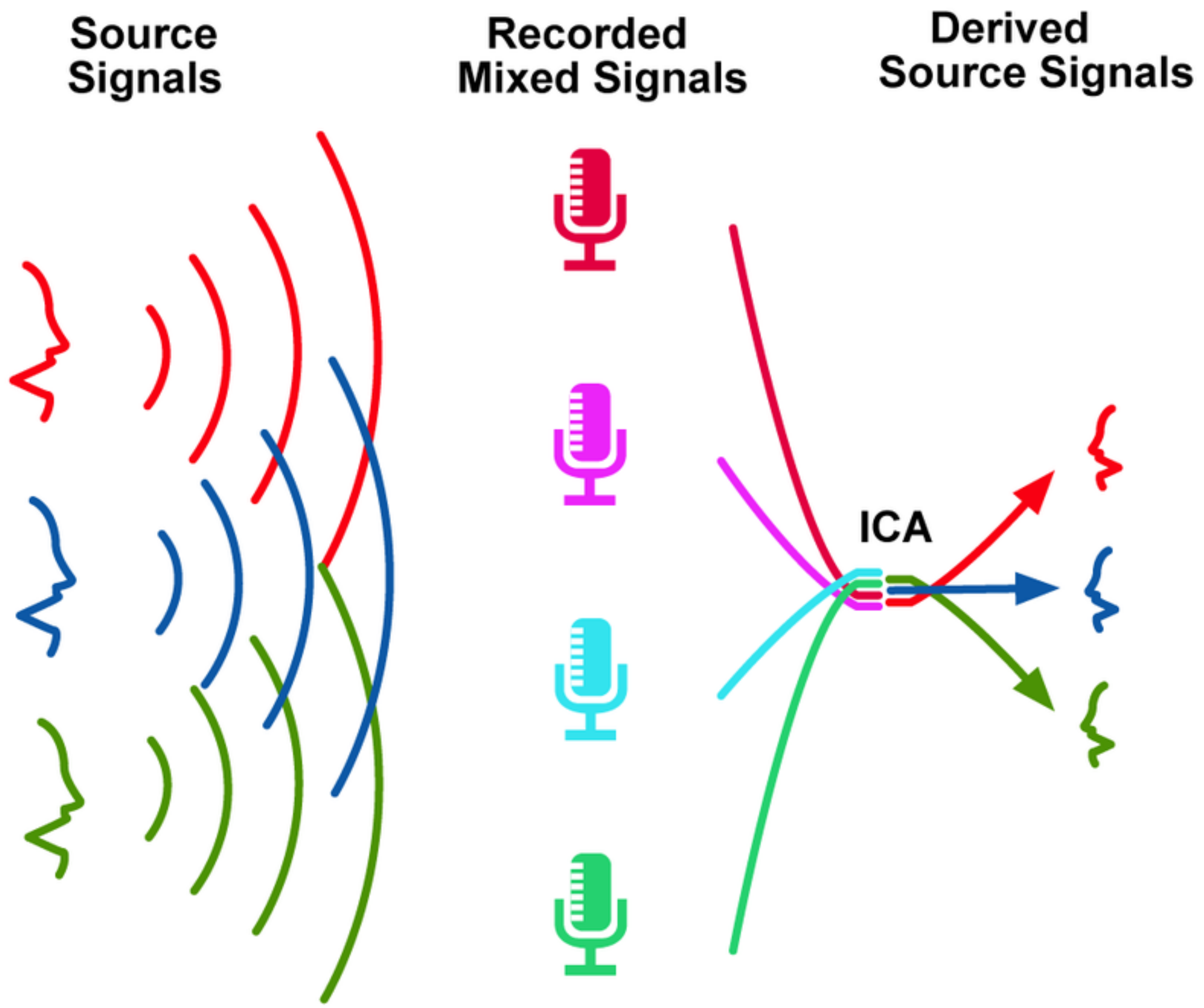
$$X = \mathbf{B}X + \varepsilon \text{ with } \mathbf{B} \in \mathbb{R}^{p \times p}, X \in \mathbb{R}^p, \varepsilon \in \mathbb{R}^p$$

- Because of acyclicity we can show that we can rewrite $\mathbf{B}$ as **strictly lower triangular** by **permuting the variables using a causal ordering**

- **Goal:** estimate B from data (which also identifies the DAG)

- ICA-LINGAM, DirectLINGAM (and many others)

Independent Component Analysis (ICA)



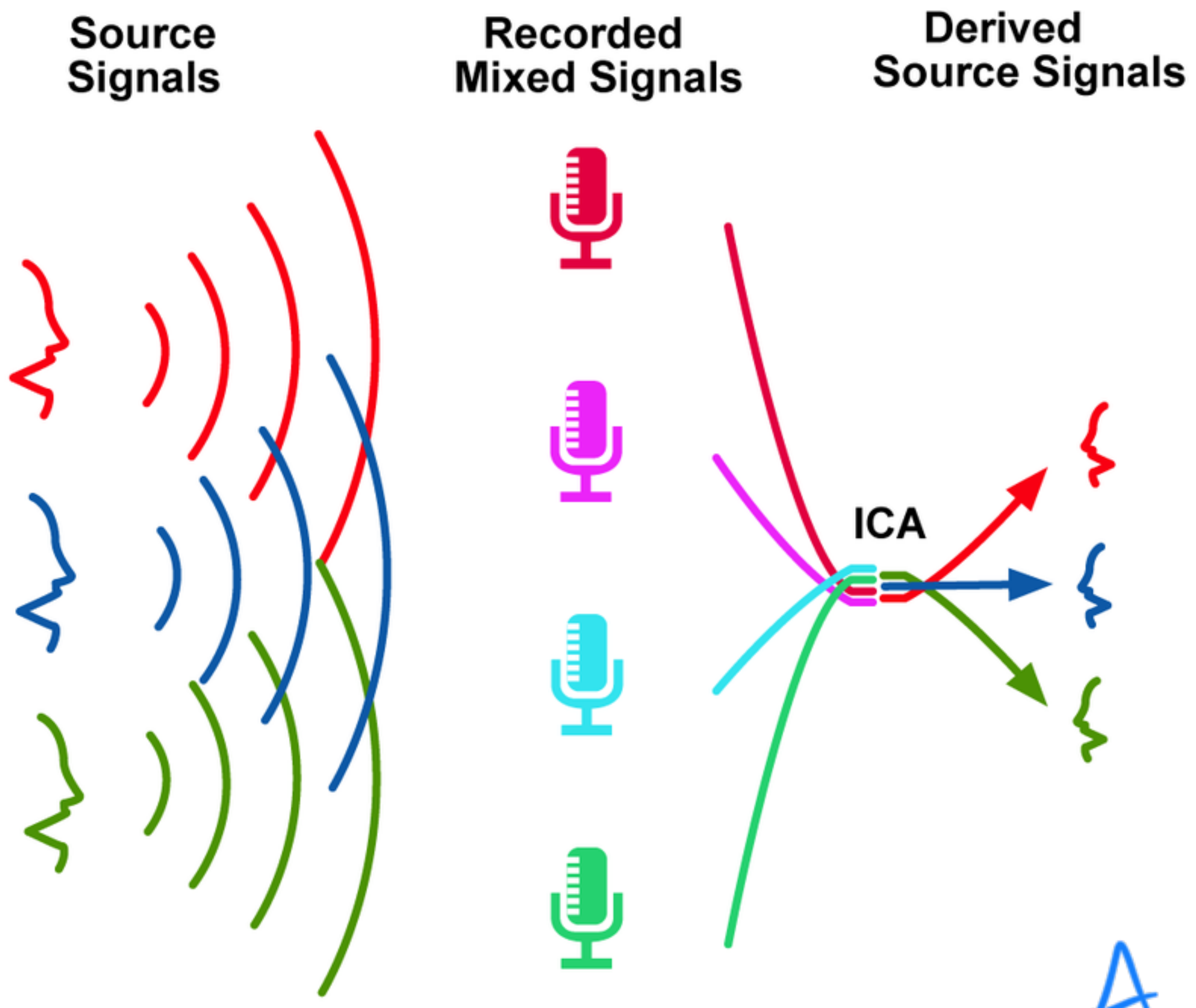
$$X = AS$$

$\in \mathbb{R}^p$

$\in \mathbb{R}^{p \times p}$ mixing matrix

$\in \mathbb{R}^p$ "sources" INDEPENDENT

Independent Component Analysis (ICA)



$$X = AS$$

$X \in \mathbb{R}^p$

$A \in \mathbb{R}^{p \times p}$ mixing matrix

$S \in \mathbb{R}^p$ "sources" INDEPENDENT

A is identifiable up to permutation, scaling, sign } of columns

ICA-LINGAM

- A linear SCM $X = \mathbf{B}X + \varepsilon$ can we rewritten as $(I - \mathbf{B})X = \varepsilon$ and $X = (I - \mathbf{B})^{-1}\varepsilon$

ICA-LINGAM

- A linear SCM $X = \mathbf{B}X + \varepsilon$ can we rewritten as $(I - \mathbf{B})X = \varepsilon$ and $X = \underbrace{(I - \mathbf{B})^{-1}}_{\text{mixing matrix } A}\varepsilon$

Sources S

ICA-LINGAM

- A linear SCM $X = \mathbf{B}X + \varepsilon$ can we rewritten as $(I - \mathbf{B})X = \varepsilon$ and $X = \underbrace{(I - \mathbf{B})^{-1}}_{\text{mixing matrix } A}\varepsilon$

mixing matrix
A

↓

up to permutation, scaling...

Sources S

LINGAM: independent non Gaussian

≡ ICA

ICA-LINGAM

1. Given dataset $D = \{x^1_{\mathbf{V}}, x^2_{\mathbf{V}}, \ldots, x^n_{\mathbf{V}}\}$ use ICA to estimate $W = A^{-1} = (I - \mathbf{B})$
2. Find **unique permutation** of rows of W such that $\tilde{W}$ does not have zeros on diagonal
3. Divide each row in $\tilde{W}$ by its diagonal element (so we get all 1 on the diagonal)
4. Compute $\hat{\mathbf{B}} = I - \tilde{W}$
5. Find **causal ordering** described by the permutation matrix P by making $\tilde{\mathbf{B}} = P\hat{\mathbf{B}}P^T$ as close as possible to strictly lower triangular

Next week: using interventional data

- All of the methods we saw until now use only **observational data**
- For restricted models this works well, since if the assumptions they make are true, then they can recover the true causal graph
- For score-based and constraint-based models, there are more advanced methods that can also use interventional data
  - For example for GES there is GIES
- If we don’t know the targets of the interventions -> Joint Causal Inference