Causal Discovery

Kun Zhang

Carnegie Mellon University

Causal Discovery from Data: Examples

Science AAAS March, 2014

RESEARCH ARTICLES

Large-Scale Psychological Differences Within China Explained by Rice Versus Wheat Agriculture

T. Talhelm,[1*] X. Zhang,[2,3] S. Oishi,[1] C. Shimin,[4] D. Duan,[2] X. Lan,[5] S. Kitayama[5]

Cross-cultural psychologists have mostly contrasted East Asia with the West. However, this study shows that there are major psychological differences within China. We propose that a history of farming rice makes cultures more interdependent, whereas farming wheat makes cultures more independent, and these agricultural legacies continue to affect people in the modern world. We tested 1162 Han Chinese participants in six sites and found that rice-growing southern China is more interdependent and holistic-thinking than the wheat-growing north. To control for confounds like climate, we tested people from neighboring counties along the rice-wheat border and found differences that were just as large. We also find that modernization and pathogen prevalence theories do not fit the data.

Over the past 20 years, psychologists have cataloged a long list of differences be- more insular and collectivistic (6). Studies have found that historical pathogen prevalence

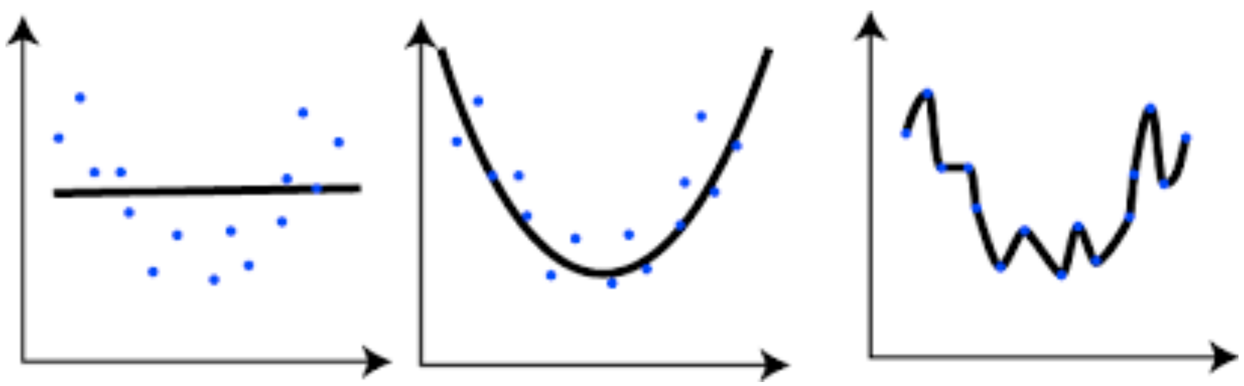
founded with rice—a possibility that prior research did not control for.

X: rice/wheat agriculture;
Y: culture;
Z: climate etc.:

$X \not\perp\!\!\!\perp Y$;
$X \not\perp\!\!\!\perp Y \mid Z$.

Under what conditions can we say $X \rightarrow Y$?

Quick Look at Supervised Learning...

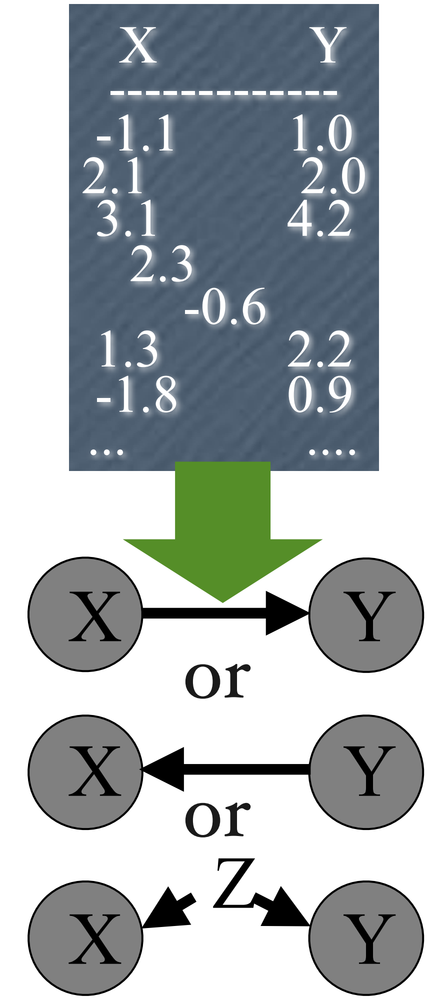


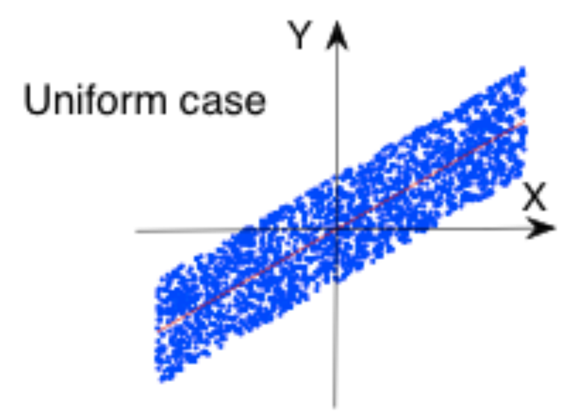
- Given training set $\mathbf{S} = \{(x_1, y_1), ..., (x_n, y_n)\}$
- $\mathcal{H}$: hypothesis space, a space of functions $f : X \rightarrow Y$.
- Learning algorithm looks at $\mathbf{S}$ and selects from $\mathcal{H}$ a function $f_S : x \rightarrow y$ such that $f_S(x) \approx y$ **in a predictive way or f_S generalizes well** (go beyond data!)
- What knowledge helps in causal discovery?

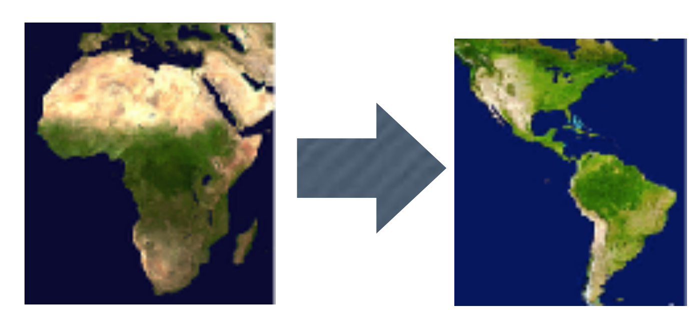
not ill-posed any more...

Outline

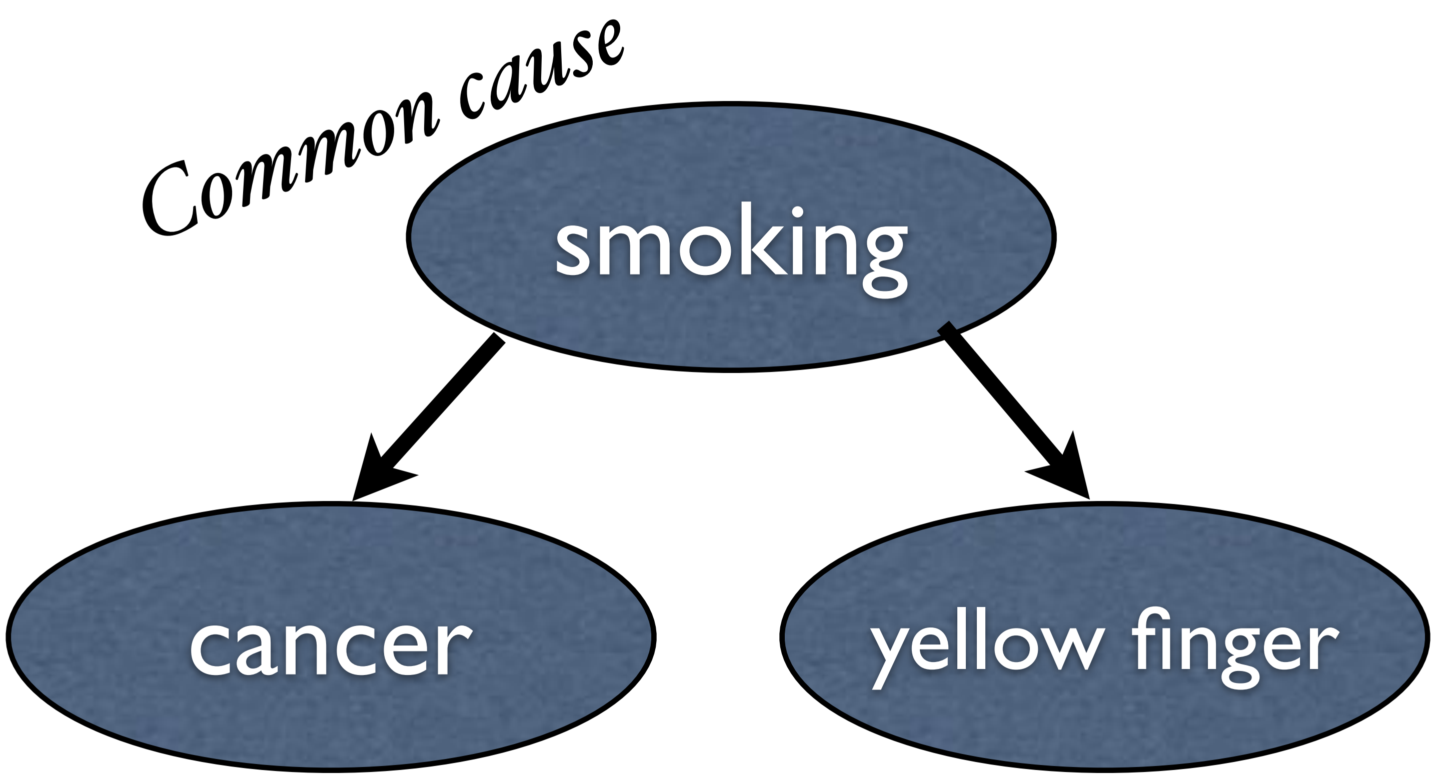
- **Causal discovery**
    - **Constraint-based approach**
    - Score-based approach
    - Functional causal model-based approach
    - Extensions
- Causality-based learning
    - Domain adaptation (transfer learning)

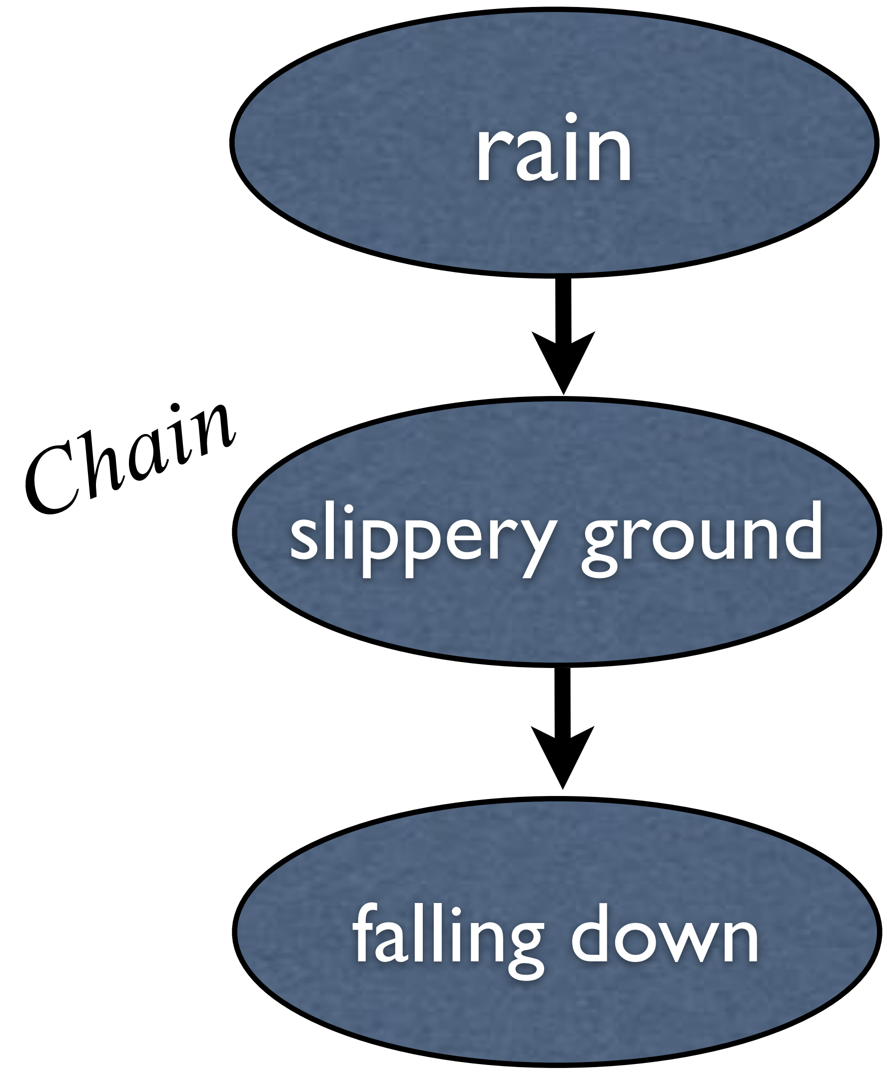






(Local) Causal Markov Condition

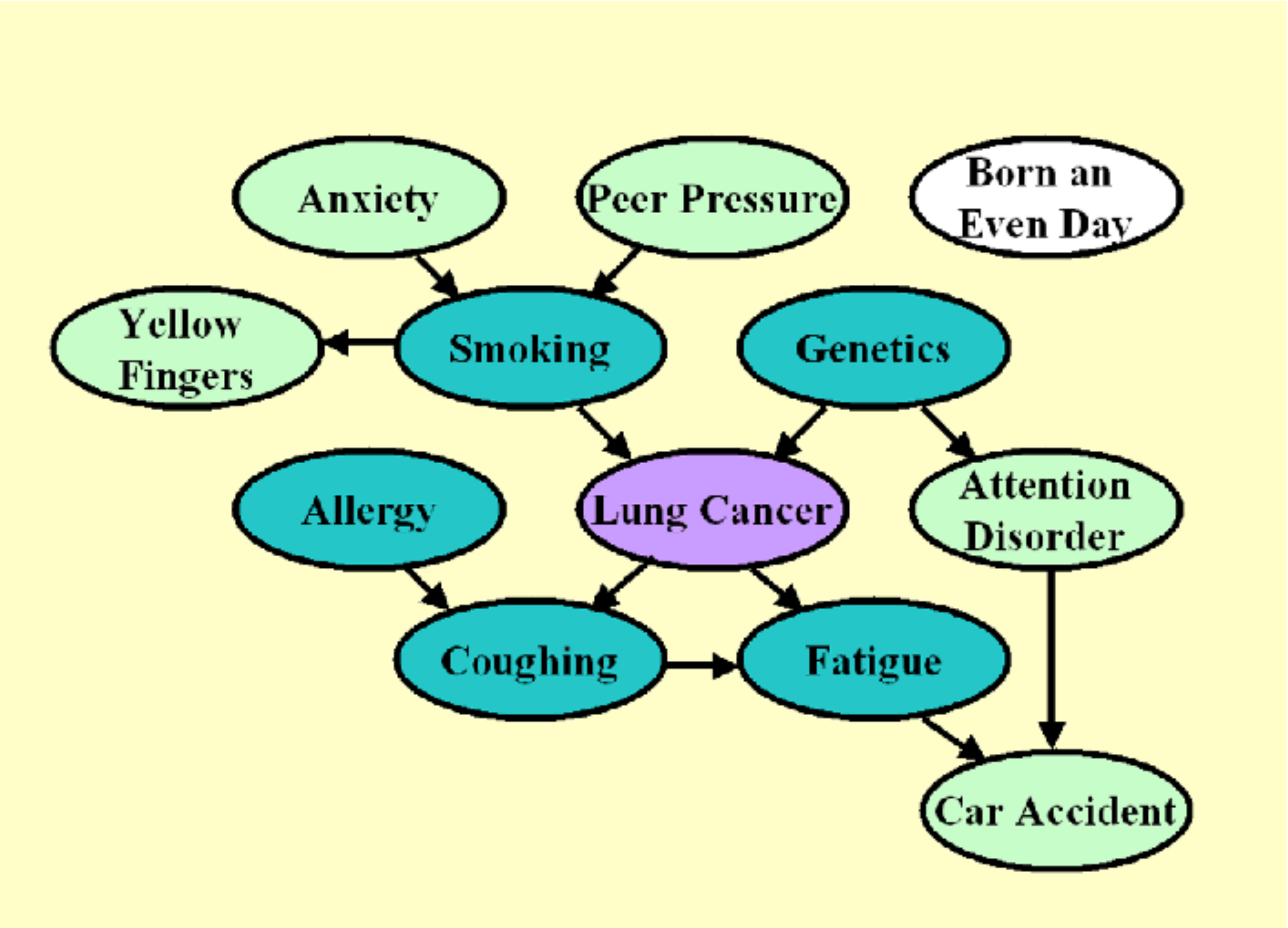




- Each variable is independent from its non-descendants given its parents

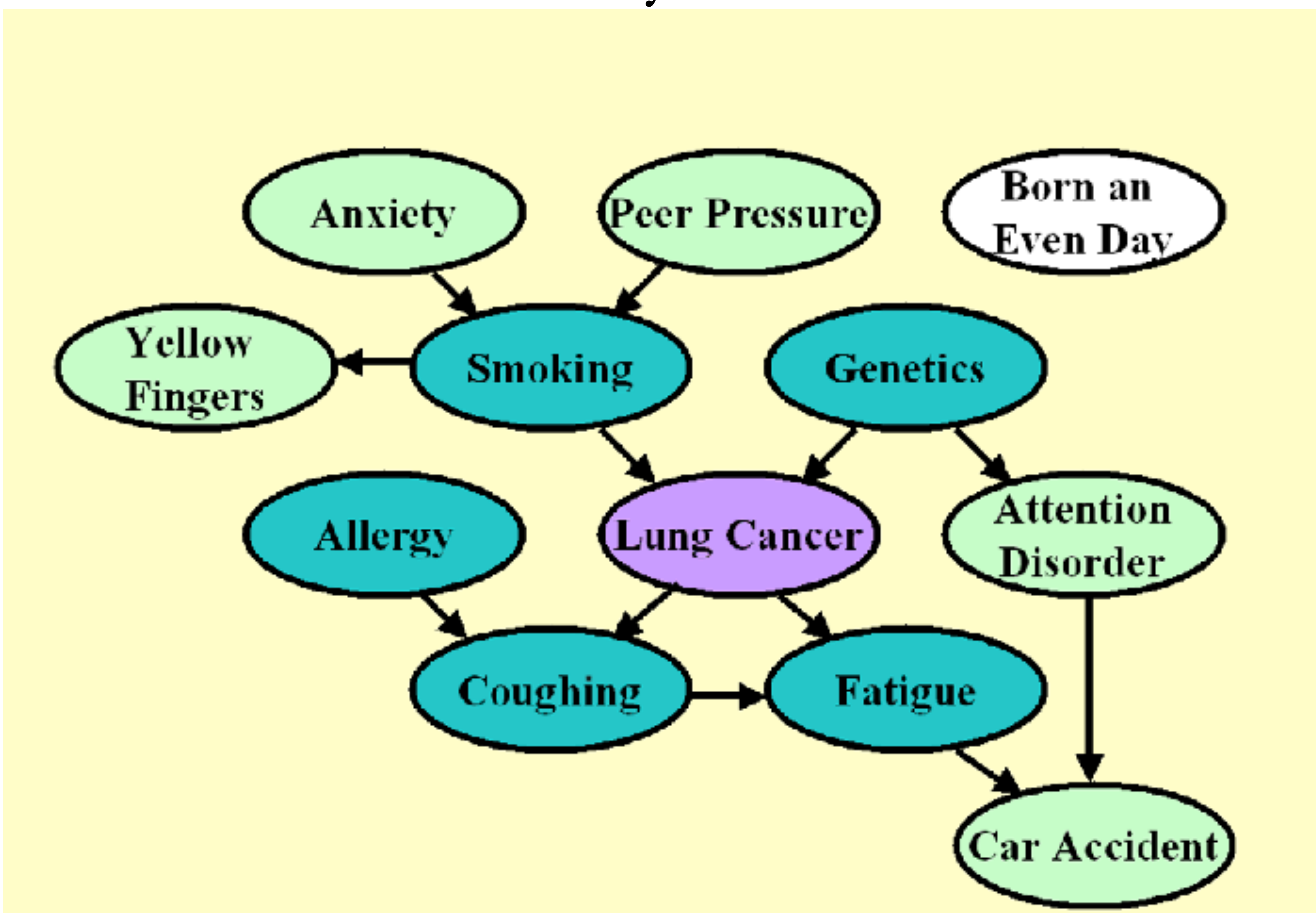
Is Local Causal Markov Condition Enough?

- Can we see whether two arbitrary variables, X and Y, are conditionally independent given an arbitrary set of variables, $\mathbf{Z}$?



D-Separation Tells Conditional Independence

- If every path from a node in **X** to a node in **Y** is d-separated by **Z**, then **X** and **Y** are always conditionally independent given **Z**
- d: directional... You will see why



D-Separation

- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

- A path p is blocked by a set of nodes **Z** if
  - p contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node m is in **Z**, **or**
  - p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is in not **Z** and no descendant of m is in **Z**

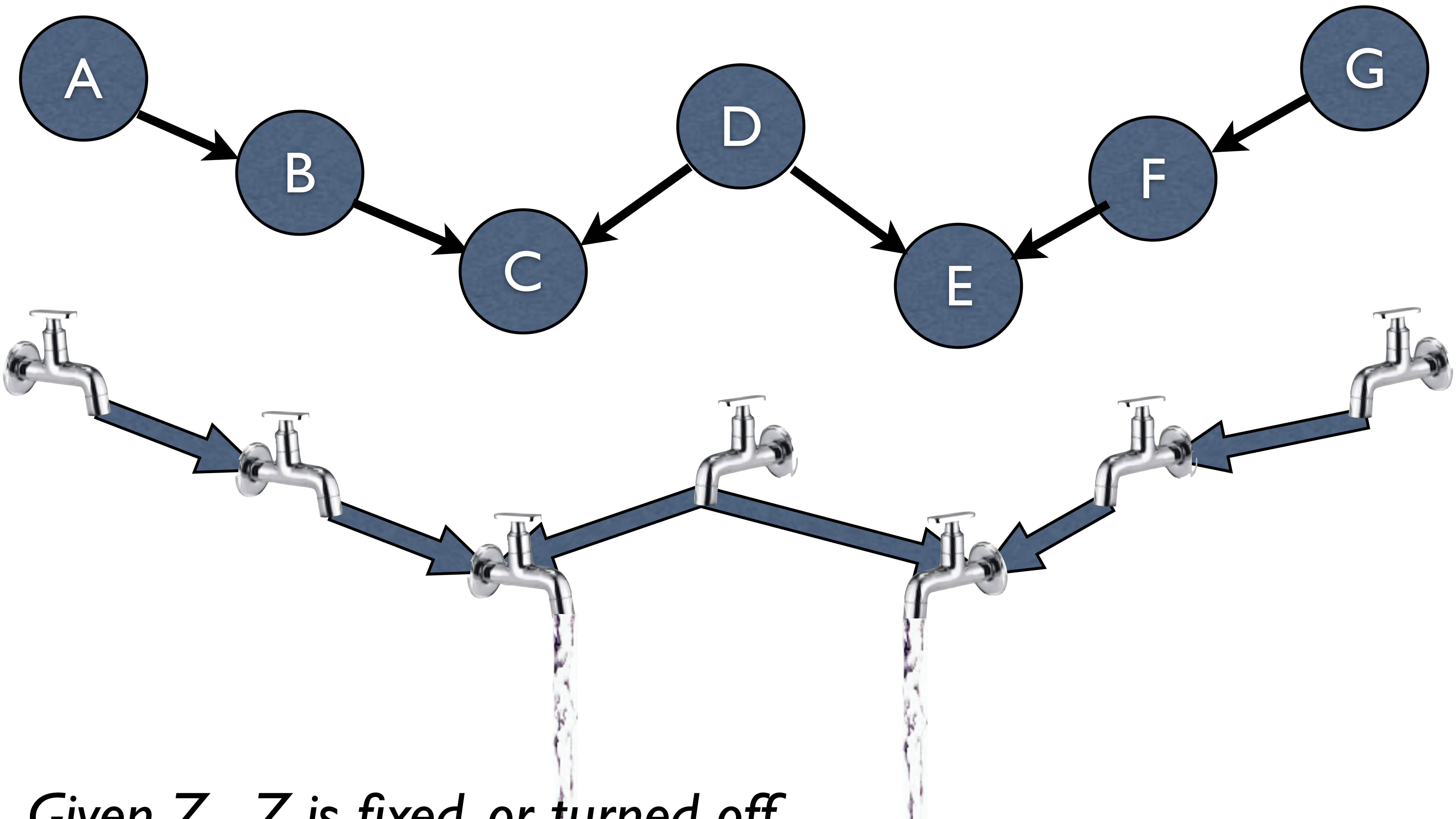
$$X \rightarrow R \rightarrow S \rightarrow T \leftarrow U \leftarrow V \rightarrow Y$$

X and Y d-separated by {R,V}?
S and U d-separated by {R,V}?

$$\begin{tikzcd}
X \arrow[r] & R \arrow[r] \arrow[d] & S \arrow[r] & T \arrow[d] & U \arrow[l] & V \arrow[l] \arrow[r] \arrow[d] & Y \\
 & W & & P & & Q &
\end{tikzcd}$$

X and Y d-separated by {R, P}?

D-Separation: Intuition



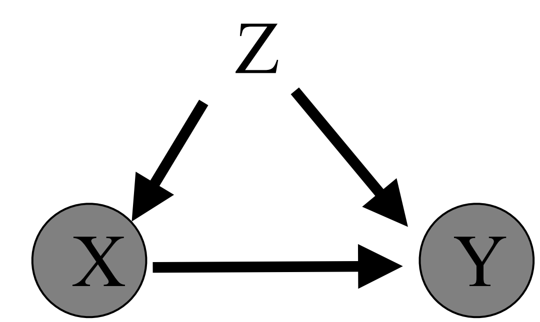
Given Z... Z is fixed, or turned off

Local & Global Markov Conditions

- **Local** Markov condition:
  - In a DAG, a variable X is independent of all its non-descendants given its parents
- **Global** Markov condition:
  - Given a DAG, let X and Y be two variables and $\mathbf{Z}$ be a set of variables that does not contain X or Y. If $\mathbf{Z}$ d-separates X and Y, then $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.
- Actually equivalent on DAGs!

Causal Sufficiency

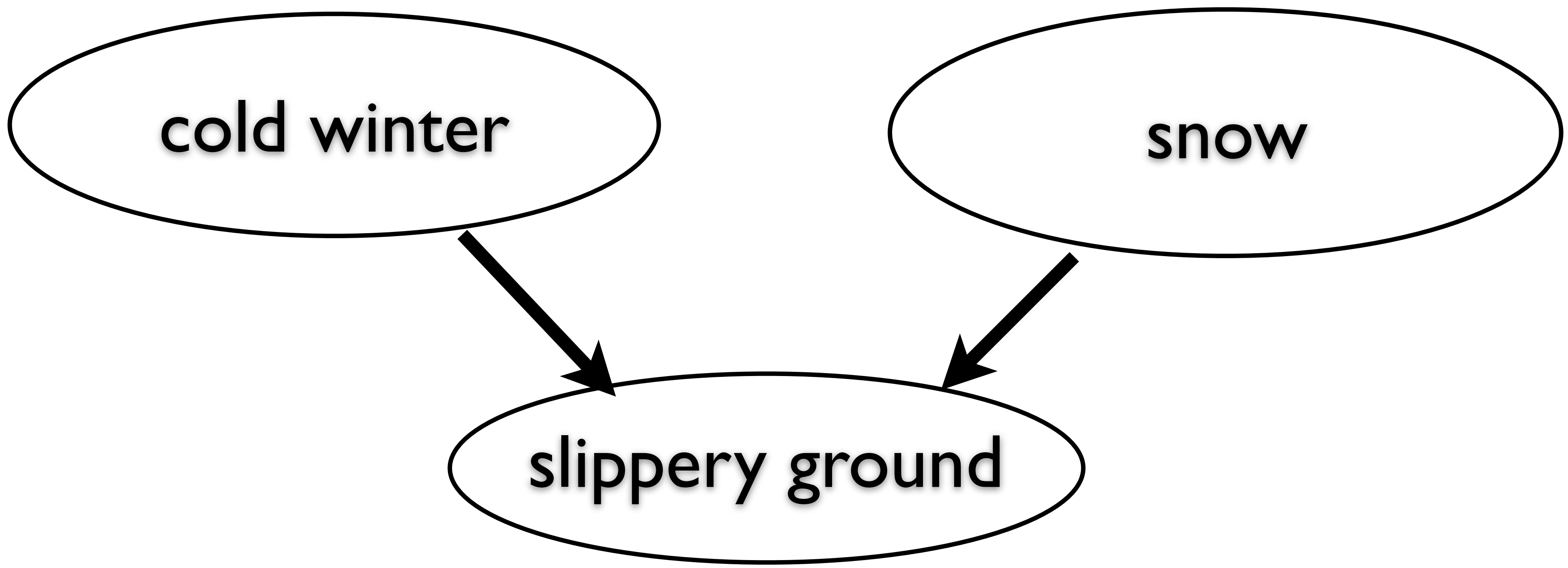
- A set of random variables $\boldsymbol{V}$ is causally sufficient if $\boldsymbol{V}$ contains every direct cause (with respect to $\boldsymbol{V}$) of any pair of variables in $\boldsymbol{V}$



- $\boldsymbol{V} = \{X, Y, Z\}$: causally sufficient
- $\boldsymbol{V} = \{X, Y\}$: causally insufficient
- Methods exist in causally **insufficient** cases, e.g., FCI (*Chapter 6 of the SGS book*)

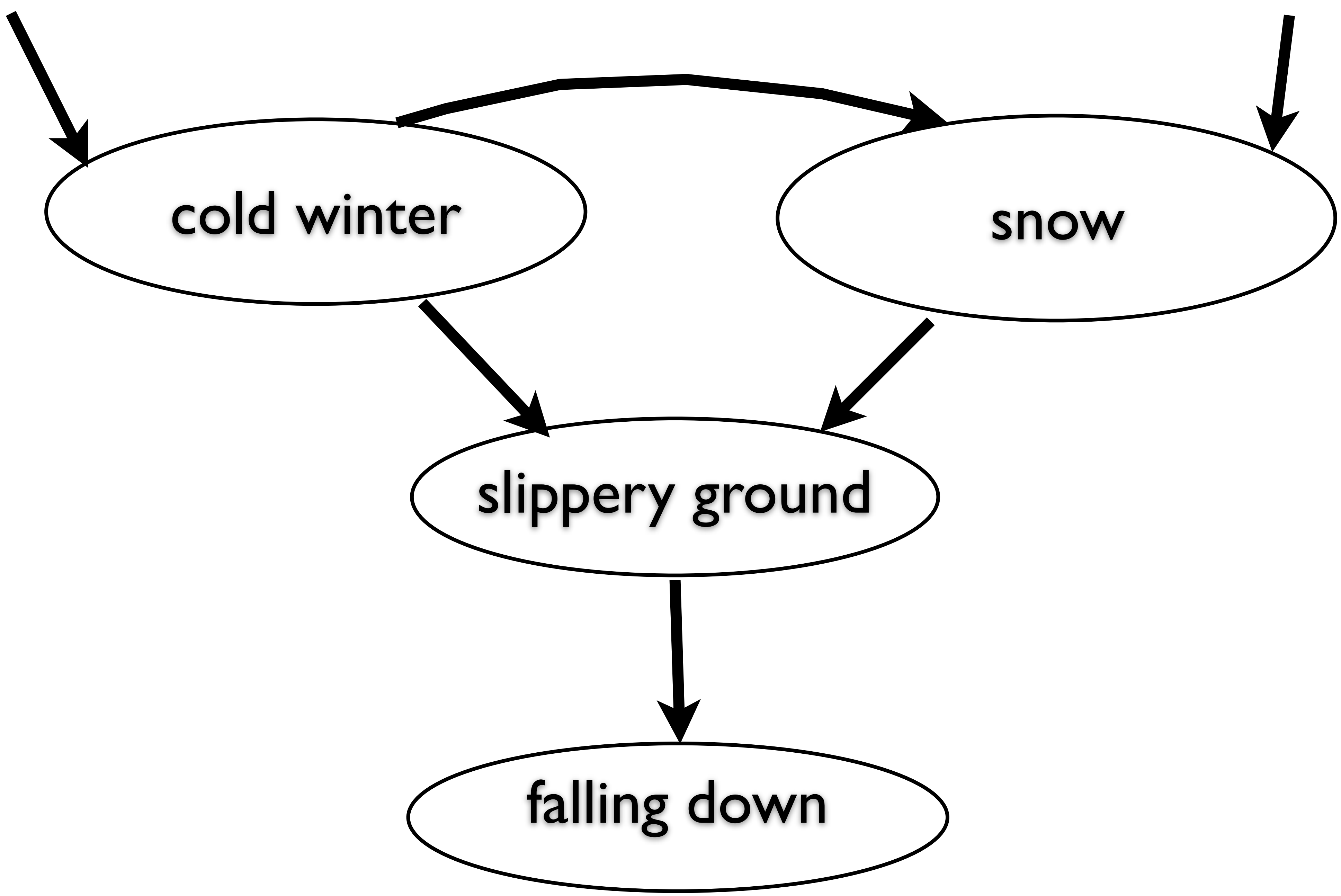
SGS Book, Chapter 5 (for causally sufficient structures); Chapter 6 (without causal sufficiency)

V-Structures

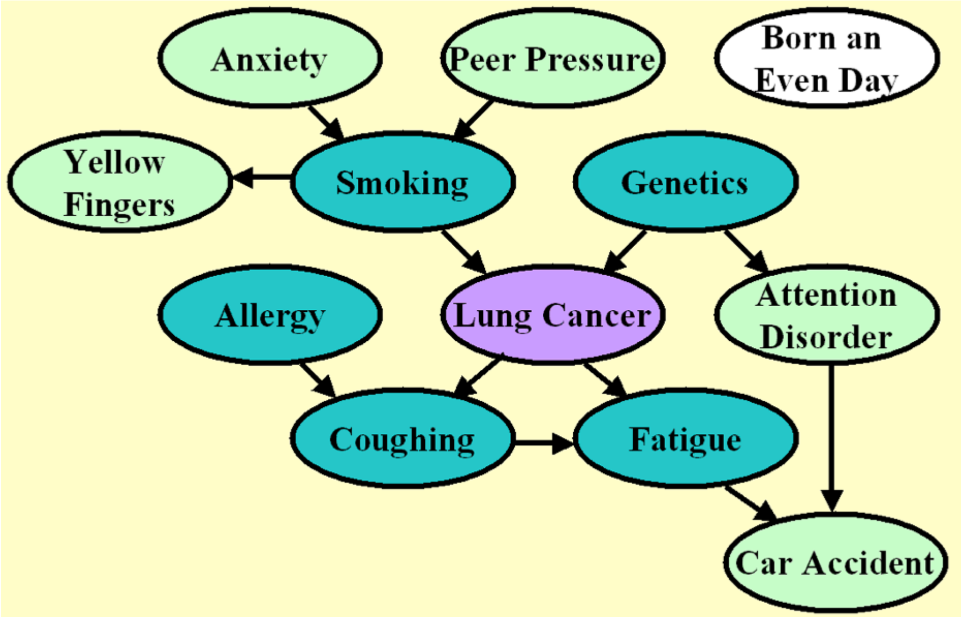


Why so interesting?

Causal Markov Condition



We can See CI Relations from DAGs...



- Local Markov condition
- Global Markov condition
  - d-separation implies conditional independence:

$P(\mathbf{V})$, where $\mathbf{V}$ denotes the set of variables, obeys the global Markov condition (or property) according to DAG $\mathcal{G}$ if for any disjoint subsets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, we have

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}.$$

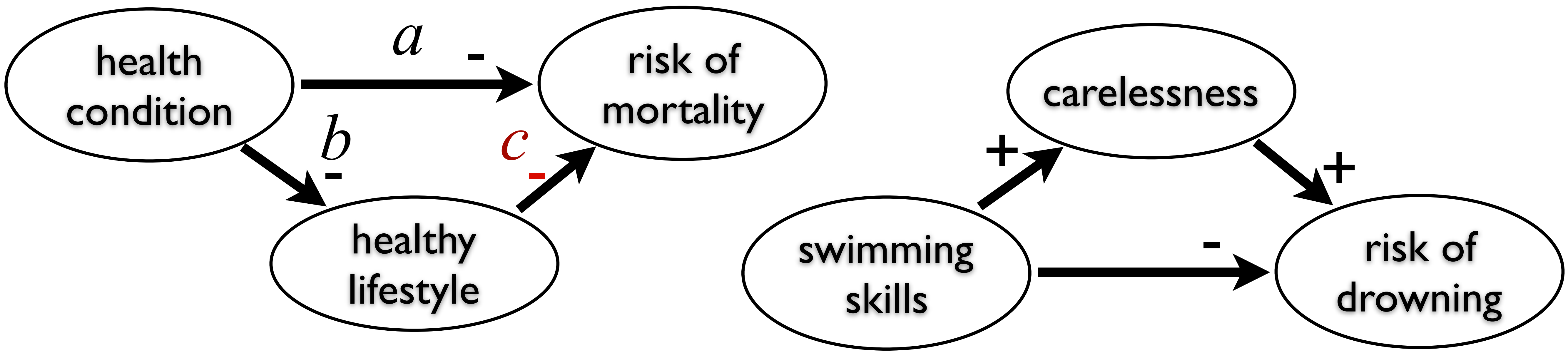
Going from CI to Graph?

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}.$$

- Contrapositive:
  - Conditional dependence implies d-connection
  - What if variables are conditionally independent?
- Can we recover the property of the underlying graph from CI relations with Markov condition?
  - Arbitrary $P(\mathbf{V})$ would satisfy the global Markov condition according to G^{f} *in which there is an edge between each pair of variables*: trivial !
  - Under what assumptions can we have CI ⇒ d-separation?

Faithfulness Assumption

- One may find independence between health condition & risk of mortality and between swimming skills & risk of drowning



- E.g., if they are linear-Gaussian and a=-bc, then *health_condition* ⫫ *risk_mortality*, which cannot by seen from the graph!

- Faithfulness assumption eliminates this possibility!

Causal Structure vs. Statistical Independence
(SGS, et al.)

Causal Markov condition: each variable is ind. of its non-descendants (non-effects) conditional on its parents (direct causes)

causal structure (causal graph)

$Y \rightarrow X \rightarrow Z$

$Y -- X -- Z$ **?**

Statistical independence(s)

$Y \perp\!\!\!\perp Z \mid X$

Faithfulness: all observed (conditional) independencies are entailed by Markov condition in the causal graph

Recall: $Y \perp\!\!\!\perp Z \Leftrightarrow P(Y|Z)=P(Y)$; $Y \perp\!\!\!\perp Z|X \Leftrightarrow P(Y|Z,X)=P(Y|X)$

Constraint-Based Search?

- First, can we find the skeleton of the causal structure? If yes, how?

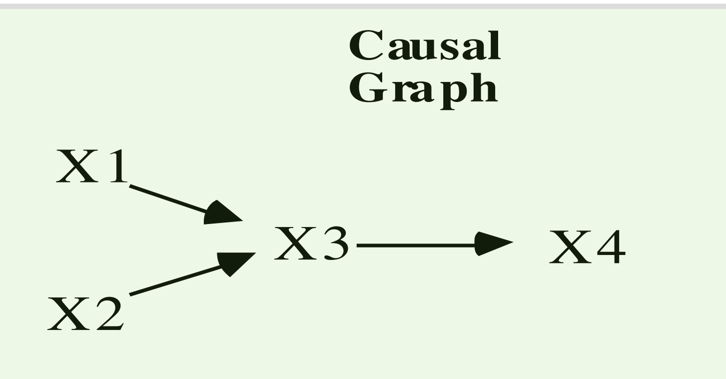
Causal Markov condition + faithfulness

- Second, can we determine the causal direction?

How?

Example I

Step 1: finding skeleton



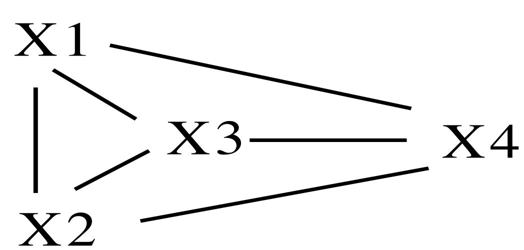
Independcies

$X1 \perp\!\!\!\perp X2$

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

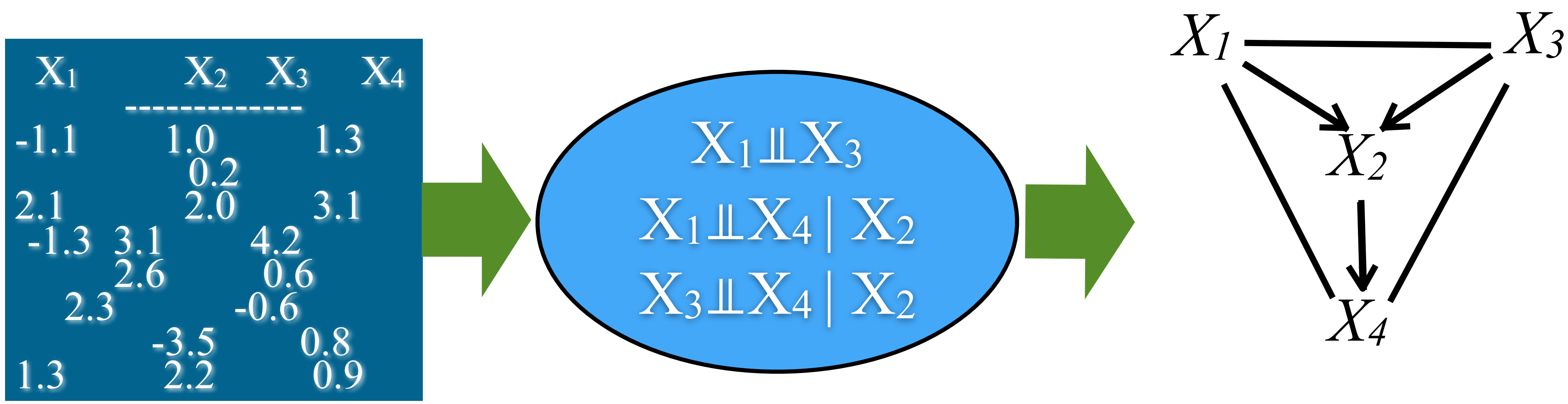
Begin with:



Step II: finding v-structure and doing orientation propagation

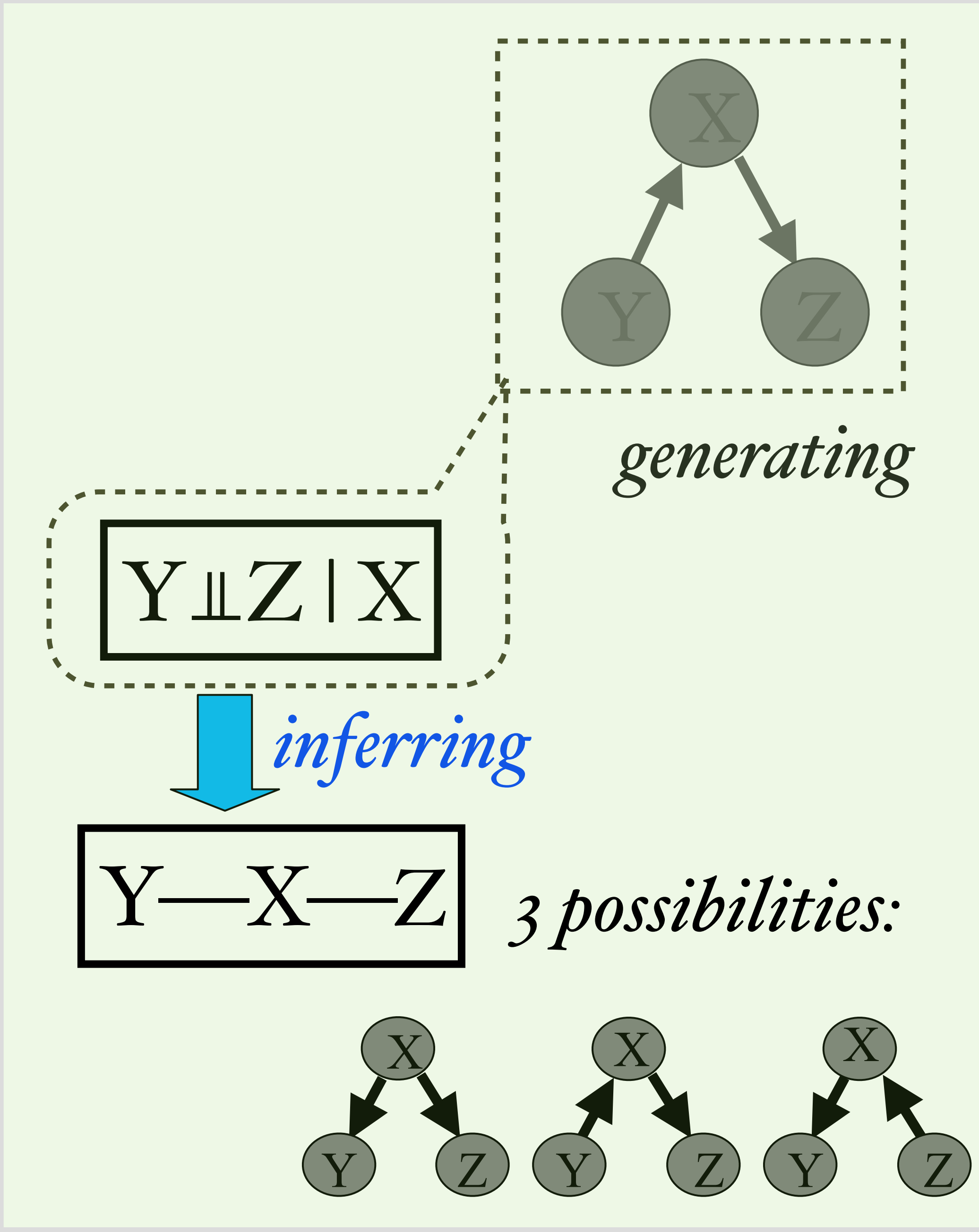
The PC Algorithm: Big Picture

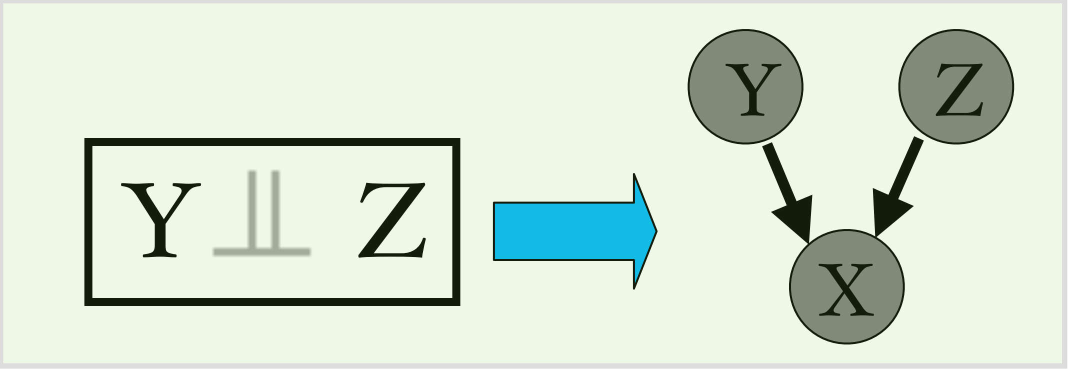
- Make use of conditional independence relations



Constraint-Based Causal Discovery

- (Conditional) independence constraints
  ⇒ candidate causal structures
  - Relies on causal Markov condition & faithfulness assumption
  - PC algorithm (Spirtes & Glymour, 1991)
  - *Step 1*: X and Y are adjacent iff they are dependent conditional on every subset of the remaining variables (SGS, 1990)
  - *Step 2*: Orientation propagation
- v-structure
- Markov equivalence class, with pattern Y—X—Z
  - same adjacencies; → if all agree on orientation; — if disagree

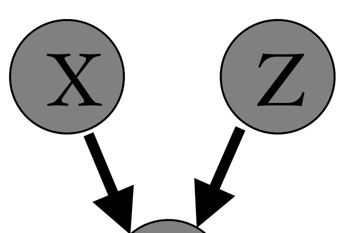




PC Algorithm

Test for (conditional) independence with an increased cardinality of the conditioning set

Finding V-structures



Orient

A.) Form the complete undirected graph C on the vertex set **V**.

B.)

$n = 0$.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset **S** of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given **S** delete edge X - Y from C and record **S** in **Sepset**(X,Y) and **Sepset**(Y,X);

until all ordered pairs of adjacent variables X and Y such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets **S** of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n have been tested for d-separation;

$n = n + 1$;

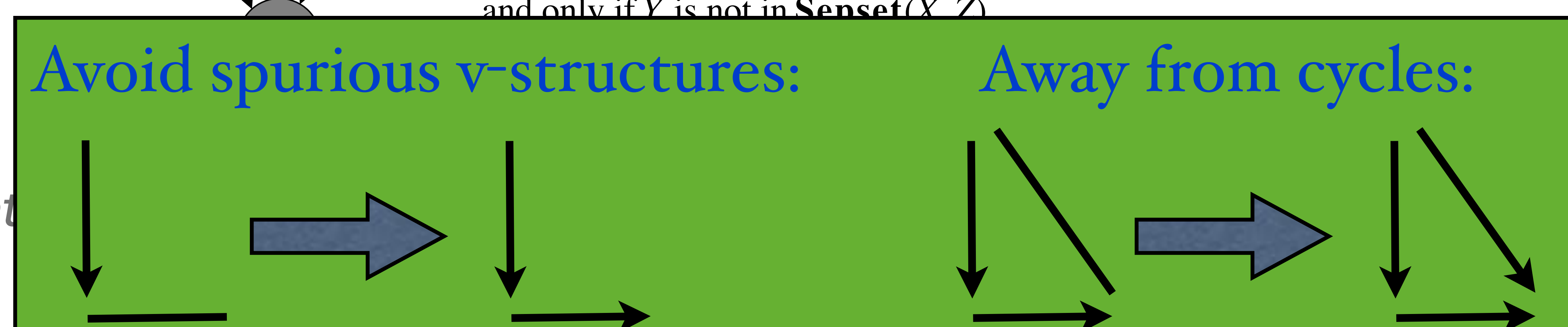
until for each ordered pair of adjacent vertices X, Y, **Adjacencies**$(C,X)\backslash\{Y\}$ is of cardinality less than n.

C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as X -> Y <- Z if and only if Y is not in **Sepset**(X,Z).

here is no

hen orient

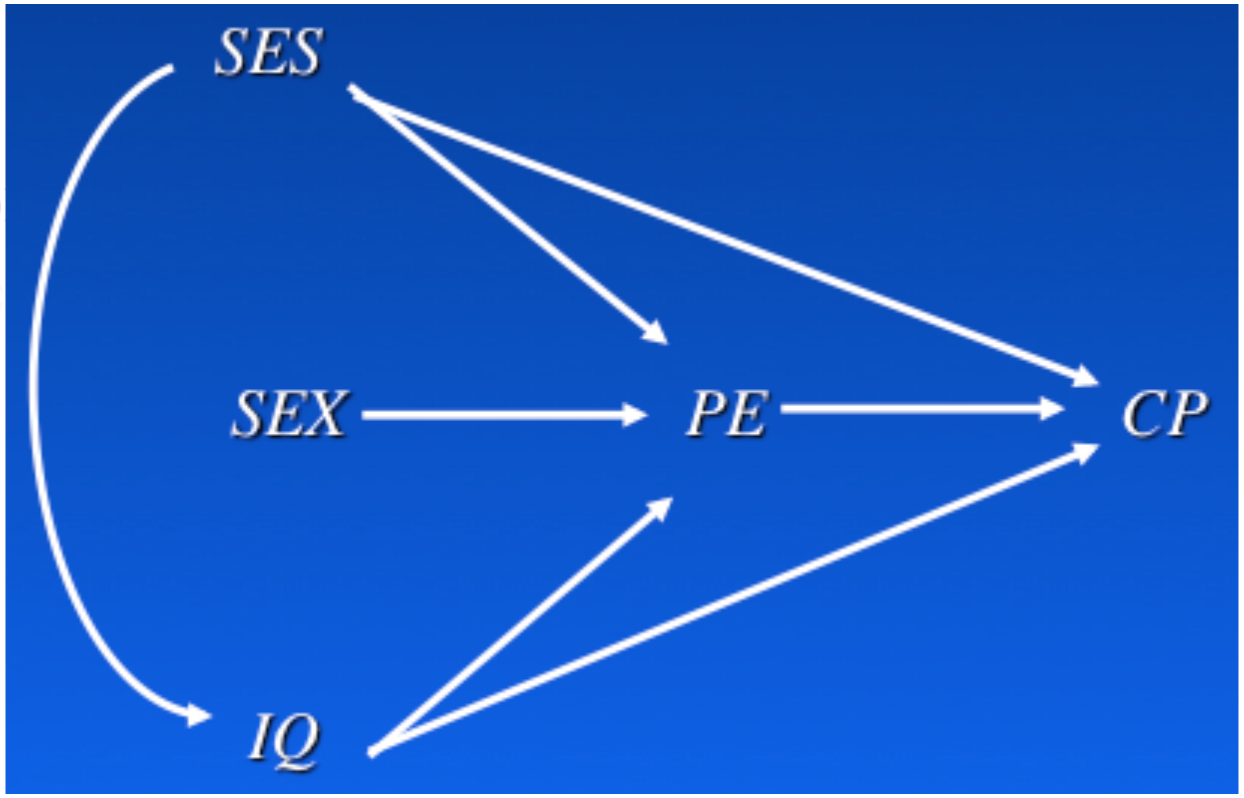
until no more edges can be oriented.



Example 1: College Plans

Sewell and Shah (1968) studied five variables from a sample of 10,318 Wisconsin high school seniors.

SEX [male = 0, female = 1]
IQ = Intelligence Quotient [lowest = 0, highest = 3]
CP = college plans [yes = 0, no = 1]
PE = parental encouragement [low = 0, high = 1]
SES = socioeconomic status [lowest = 0, highest = 3]



Example II: Causal analysis of archeology data

Thanks to collaborator Marlijn Noback

- 8 variables of 250 skeletons collected from different locations

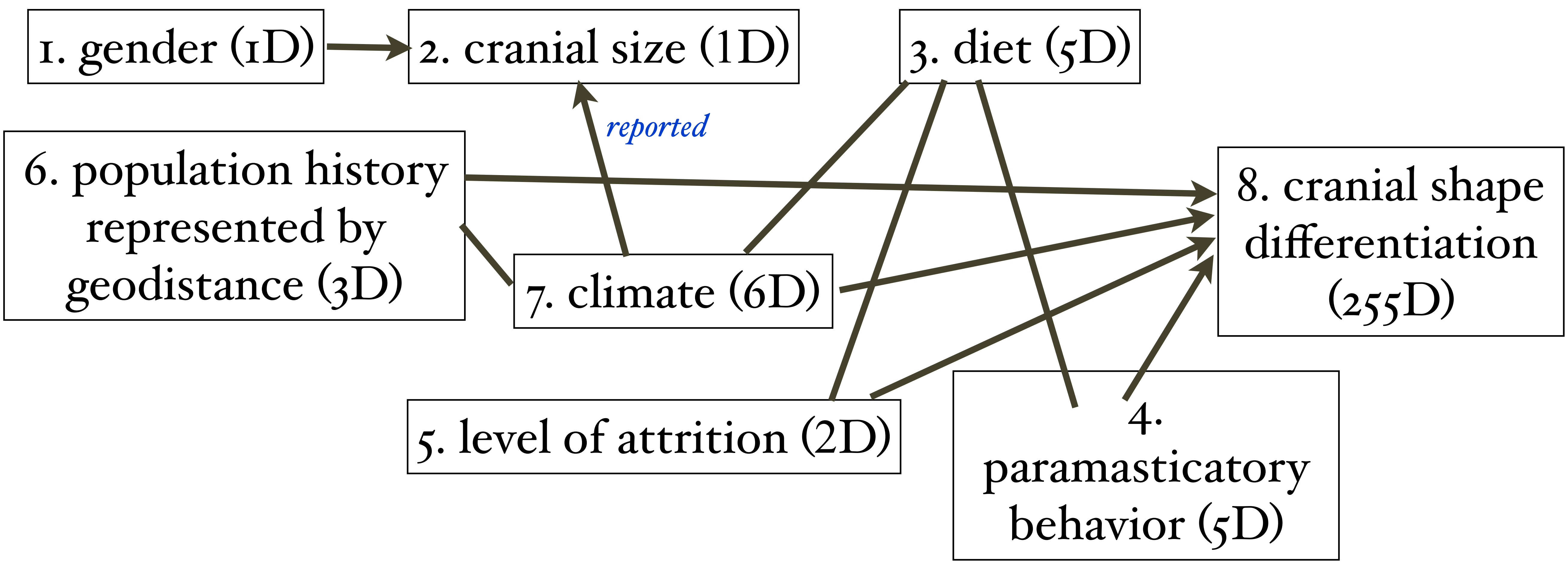
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	Population	Sex	Cranial size	Diet or subsistence					Paramastic	Dental wear		Geographic location per population			Climate per population					
2			(Male, fem	(Centroid S	Gathering	Hunting	Fishing	Pastoralism	Agriculture	Yes=1, no=	Average att	Attrition pe	Distance to	Longitude	Latitude	Tmean	Tmin	Tmax	Vpmean	Vpmin	Vpmax
3	AINU31_1	Ainu	Unknown	713.2542	2	3	4	0	1	0	1.5	2	15464	43.548548	142.539159	2.86	-11.19	17.01	7.43	2.27	15.83
4	AINU7_1	Ainu	Unknown	676.148	2	3	4	0	1	0	1.5	1	15464	43.548548	142.539159	2.86	-11.19	17.01	7.43	2.27	15.83
5	AINU7_2	Ainu	Unknown	675.4924	2	3	4	0	1	0	1.5	1	15464	43.548548	142.539159	2.86	-11.19	17.01	7.43	2.27	15.83
6	AINU_1016	Ainu	Male	684.3304	2	3	4	0	1	0	1.5	2.5	15464	43.548548	142.539159	2.86	-11.19	17.01	7.43	2.27	15.83
7	AINU_1016	Ainu	Female	686.285	2	3	4	0	1	0	1.5	4	15464	43.548548	142.539159	2.86	-11.19	17.01	7.43	2.27	15.83
8	AUSM245	Australia	Male	673.8749	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
9	AUSM246	Australia	Male	647.4586	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
10	AUSM8217	Australia	Male	658.6616	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
11	AUSM8177	Australia	Male	667.5444	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
12	AUSM8173	Australia	Male	629.7138	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
13	AUSM8173	Australia	Male	648.7064	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
14	AUSM8171	Australia	Male	643.0478	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
15	AUSM8165	Australia	Male	616.55	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
16	AUSM8154	Australia	Male	635.0605	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
17	AUSM8153	Australia	Male	650.6539	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
18	AUSF1412	Australia	Female	618.4781	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
19	AUSF8179	Australia	Female	634.3122	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
20	AUSF8175	Australia	Female	605.1759	6	4	0	0	0	1	2.5	1.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
21	AUSF8172	Australia	Female	613.8424	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
22	AUSF8169	Australia	Female	619.1206	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
23	AUSF8157	Australia	Female	628.2819	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
24	AUSF8155	Australia	Female	628.4609	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
25	AUSF1578	Australia	Female	640.6311	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
26	AUSF243	Australia	Female	606.164	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
27	AUSF8158	Australia	Female	631.6258	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.515234	22.46	13.33	30.27	11.10	7.55	15.96
28	DENM1432	Denmark	Male	653.6198	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
29	DENM1011	Denmark	Male	651.4847	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
30	DENM1705	Denmark	Male	636.9841	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
31	DENM116	Denmark	Male	642.9192	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
32	DENM116_	Denmark	Male	646.6609	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
33	DENM116	Denmark	Male	674.9799	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
34	DENM7_77	Denmark	Male	666.53	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
35	DENML_58	Denmark	Male	627.4583	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
36	DENM903	Denmark	Male	662.5553	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
37	DENM901	Denmark	Male	677.8408	0	0	1	3	6	0	2.1	NaN	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
38	DENF1550	Denmark	Female	624.4664	0	0	1	3	6	0	2.1	0.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27

INFO | Shape_rawcoordinates | Shape_symmProcCoord. | Factors | Background | Geodistance_extra | +

Example II: Result

Thanks to collaborator Marlijn Noback

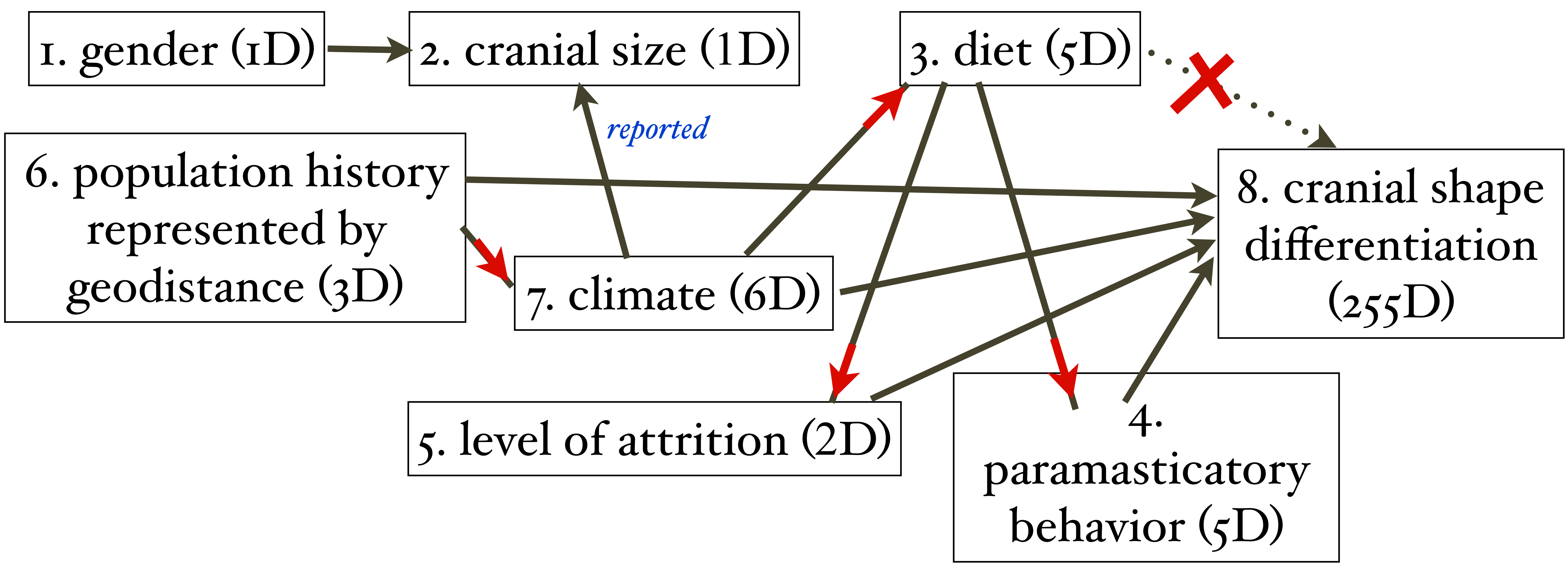
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- PC + kernel-based conditional ind. test (Zhang et al., 2011) seems to be a good choice



Example II: Result

Thanks to collaborator Marlijn Noback

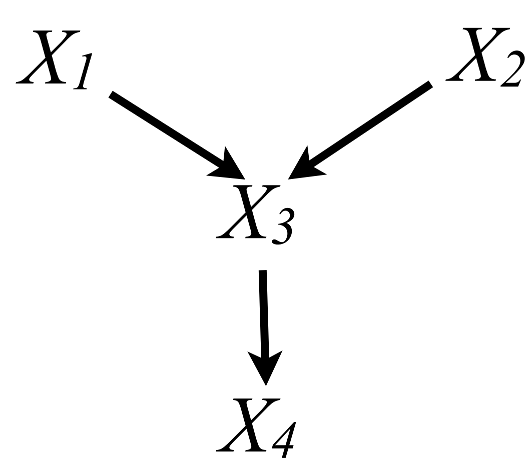
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- PC + kernel-based conditional ind. test (Zhang et al., 2011) seems to be a good choice



How about This Case?

$$X_1 \perp\!\!\!\perp X_2;$$
$$X_1 \perp\!\!\!\perp X_4 \mid X_3;$$
$$X_2 \perp\!\!\!\perp X_4 \mid X_3.$$

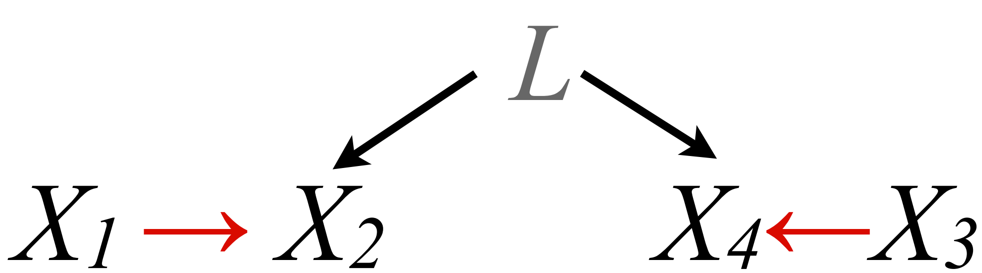
What is corresponding causal structure? Possible to have confounders behind X_3 and X_4?



:-)

How about This Case?

$$X_1 \perp\!\!\!\perp X_3;$$
$$X_1 \perp\!\!\!\perp X_4;$$
$$X_2 \perp\!\!\!\perp X_3.$$



L: a latent variable

- Patterns: a description of a class of causal processes described by various DAGs.
- In the presence of latent variables, the causal process over measured variables **O** is not necessarily a DAG. How can we represent (independence) equivalence classes over **O** ?

FCI (Fast Causal Inference) Allows Confounders

- Assume the distribution over measured variables **O** is the marginal of a distribution satisfying the Markov and faithfulness conditions for the true graph
- Results represented by PAGs

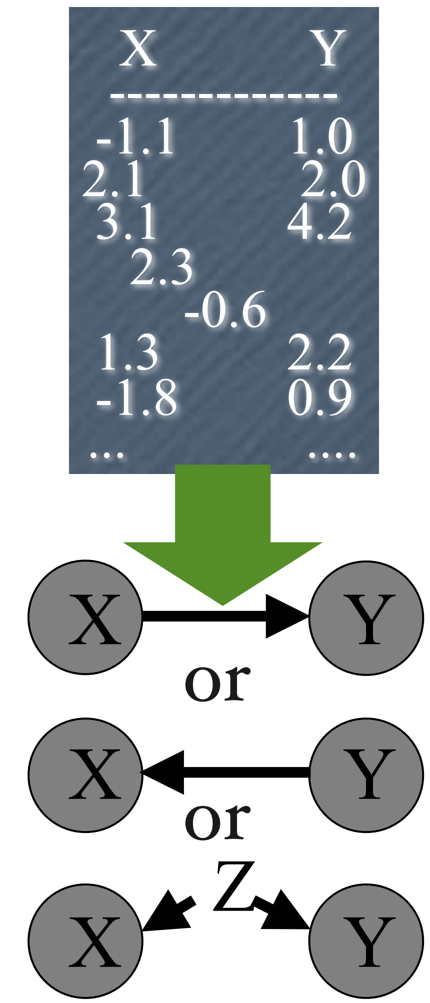
$$
\begin{tikzcd}
 & & L \arrow[dl] \arrow[dr] & & \\
X_1 \arrow[r] & X_2 & & X_4 & X_3 \arrow[l]
\end{tikzcd}
$$

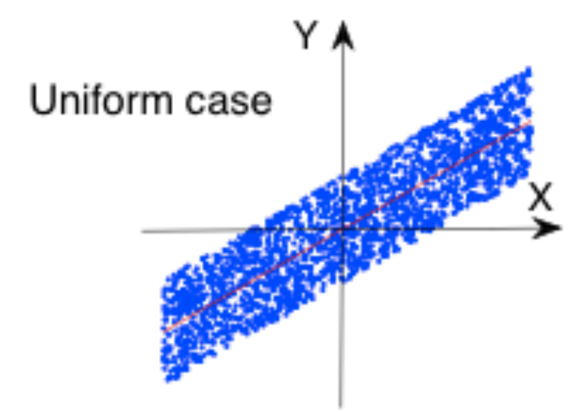
What's FCI's output?

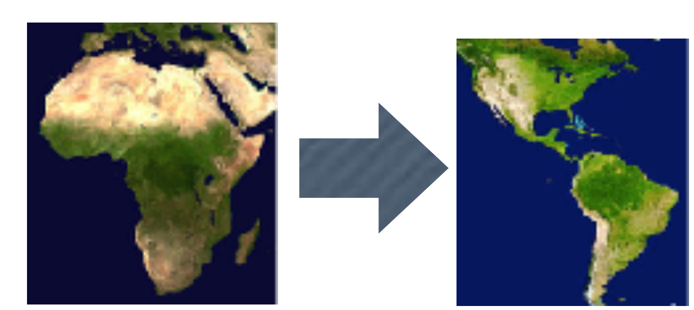
Spirtes et al., Causal inference in the presence of latent variables and selection bias, 1997

Outline

- **Causal discovery**
  - Constraint-based approach
  - **Score-based approach**
  - Functional causal model-based approach
  - Extensions
- Causality-based learning
  - Domain adaptation (transfer learning)







Key Issues

- What score to use?
  - Bayesian scoring: Allows informative prior probabilities of causal structure & parameters
  - Non-Bayesian scoring
- How to traverse the search space of the graph?
  - DAGs? Equivalence classes?
  - How to do optimization?

GES (Greedy Equivalence Search): Score Function

- Assumptions: The score is
  - score equivalent (i.e., assigning the same score to equivalent DAGs)
  - locally consistent: score of a DAG increases (decreases) when adding any edge that eliminates a false (true) independence constraint
  - decomposable: $Score(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} Score(X_i, \mathbf{Pa}_i^{\mathcal{G}})$
- E.g., BIC: $S_B(\mathcal{G}, \mathbf{D}) = \log p(\mathbf{D}|\hat{\boldsymbol{\theta}}, \mathcal{G}^h) - \frac{d}{2} \log m$

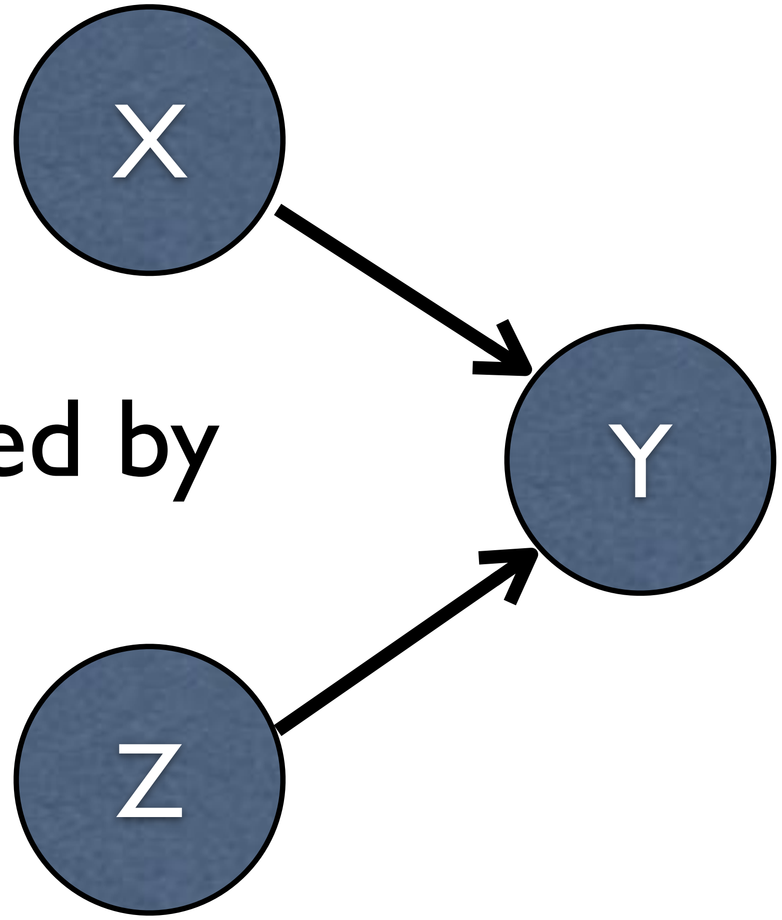
Chickering, Optimal Structure Identification With Greedy Search, Journal of Machine Learning Research, 2002

GES: Search Procedure

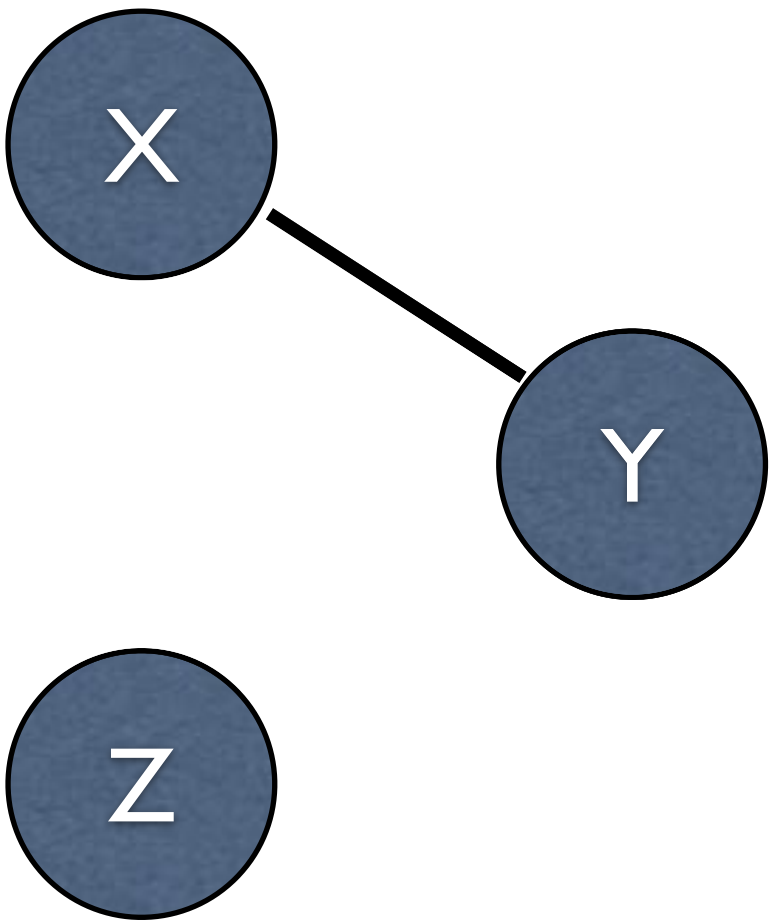
- Performs forward (addition) / backward (deletion) equivalence search through the space of DAG equivalence classes
  - Forward Greedy Search (FGS)
    - Start from some (sparse) pattern (usually the empty graph)
    - Evaluate all possible patterns with one more adjacency that entail strictly fewer CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum
  - Backward Greedy Search (BGS)
    - Start from the output of the Forward Stage
    - Evaluate all possible patterns with one fewer adjacency that entail strictly more CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum

GES

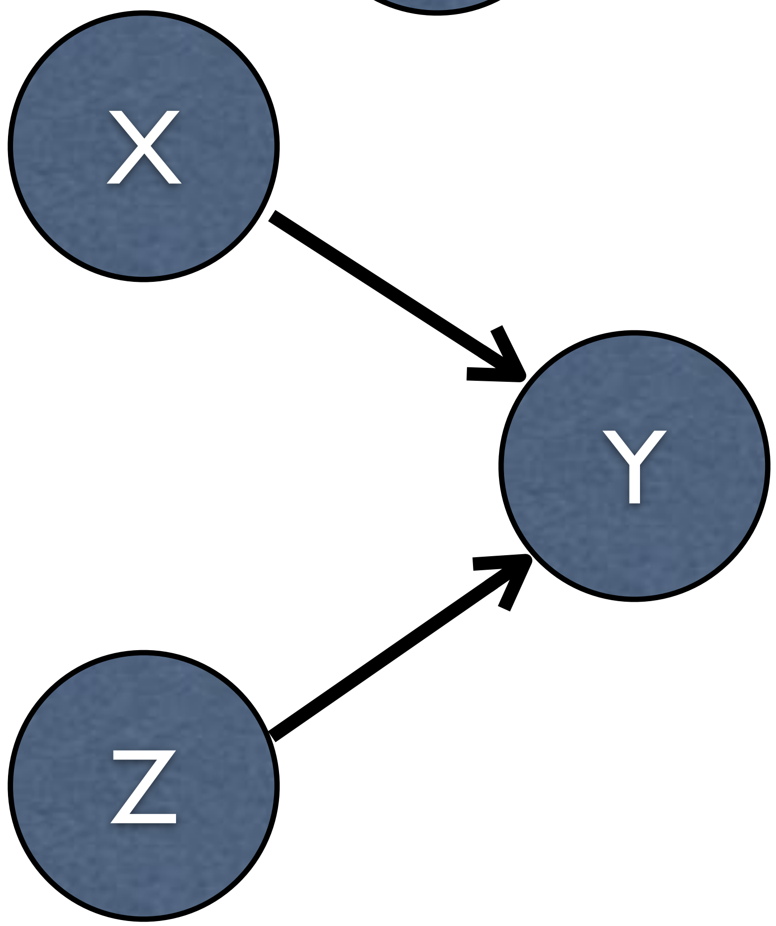
Suppose data were generated by



(1)

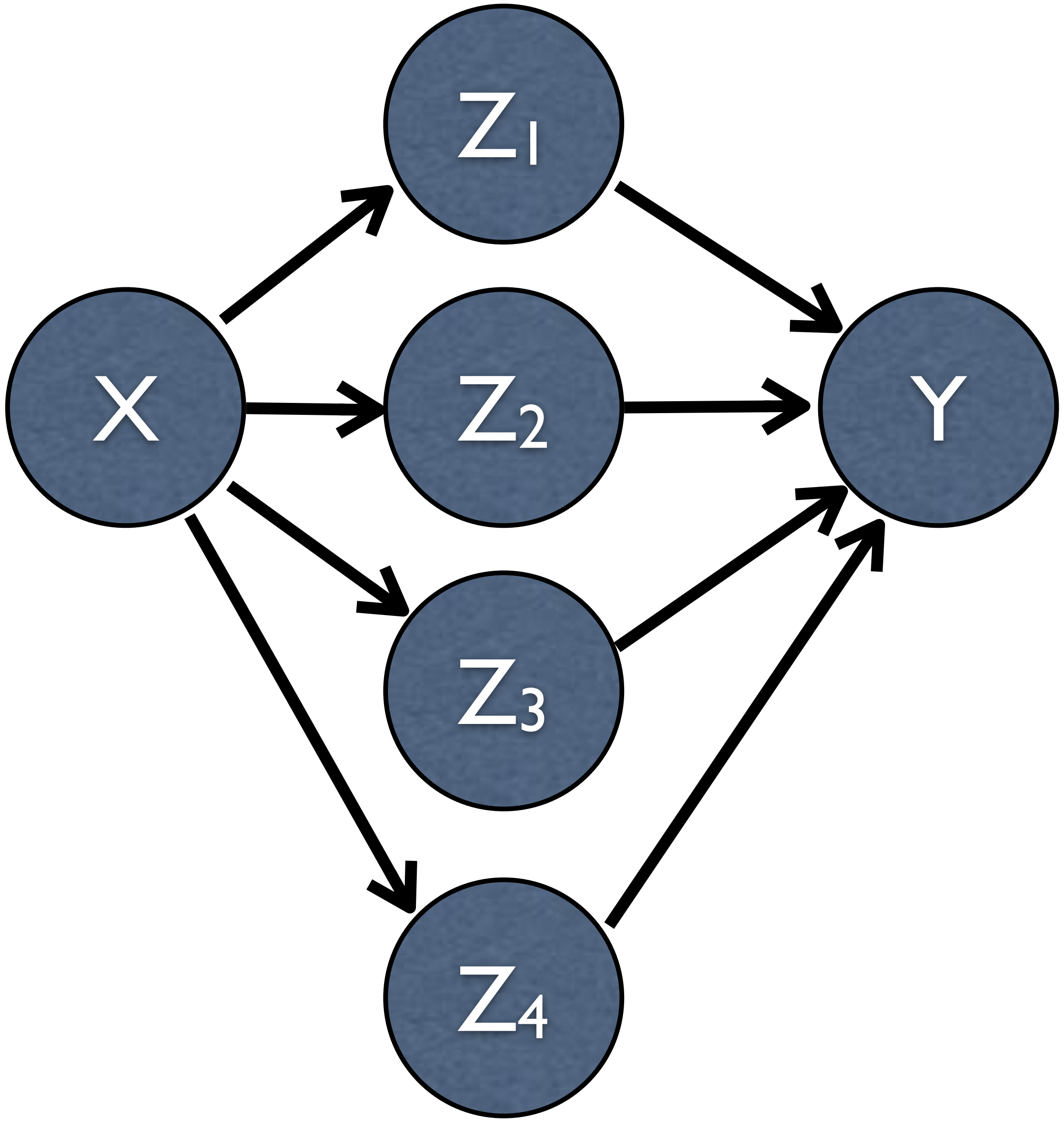


(2)



GES

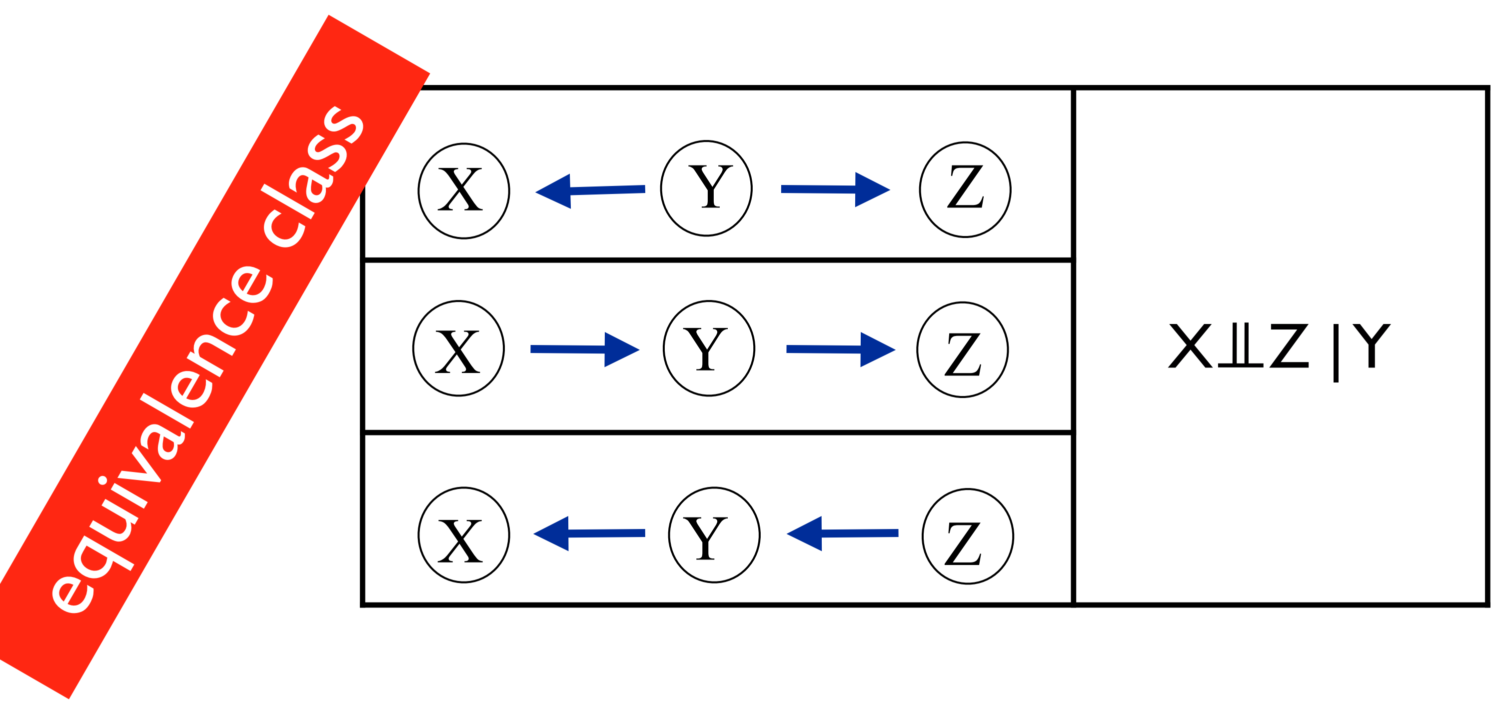
Suppose data were generated by

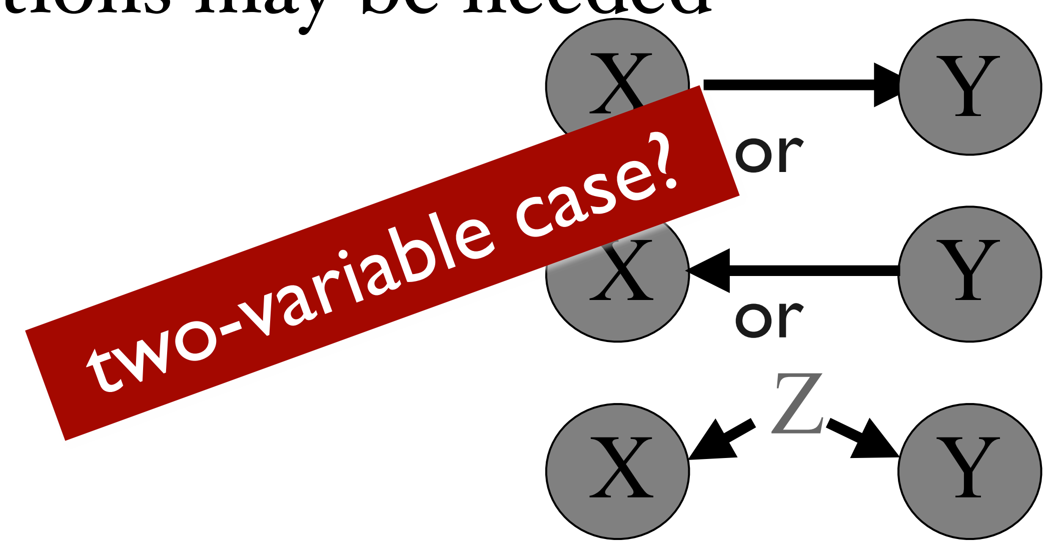


Imagine the GES procedure...

Constraint-based Causal Discovery: Advantages and Limitations

- Nonparametric; widely applicable given reliable conditional independence tests
- Recovering {causal relations} from {conditional independences}: bounded by the equivalence class
- Directly characterize and recover cause-effect relationships?
  - additional weak and reasonable assumptions may be needed

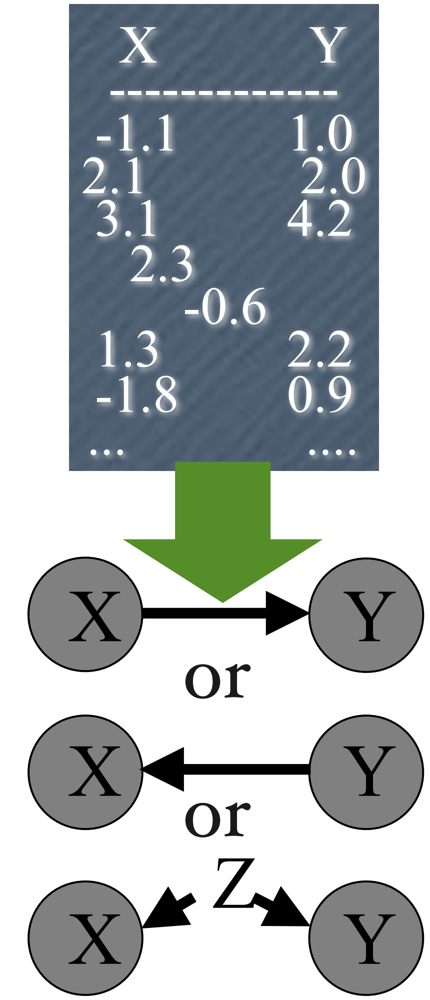


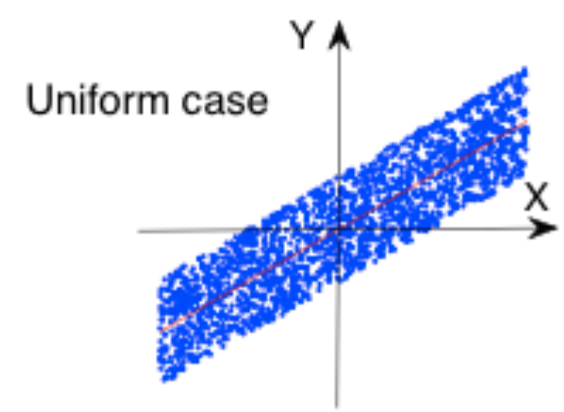


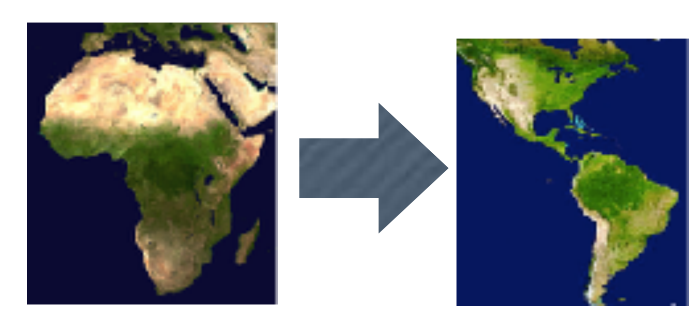
- Instead, try to directly identify local causal structures with functional causal models/structural equation models

Outline

- **Causal discovery**
  - Constraint-based approach
  - Score-based approach
  - **Functional causal model-based approach**
  - Extensions
- Causality-based learning
  - Domain adaptation (transfer learning)



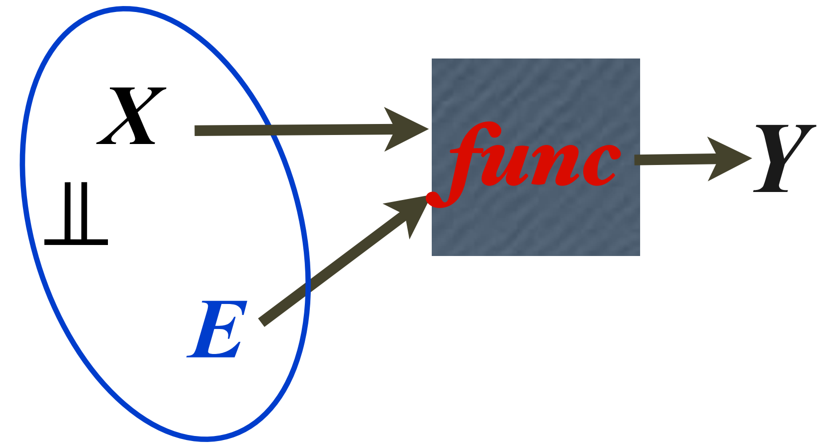




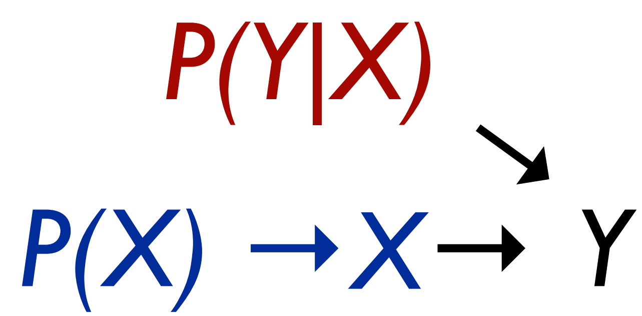
Fully Identifiable Causal Structure? Two-Variable Case.

- Structural equation model / functional causal model

  $Y = f(X, E)$, where $E \perp\!\!\!\perp X$



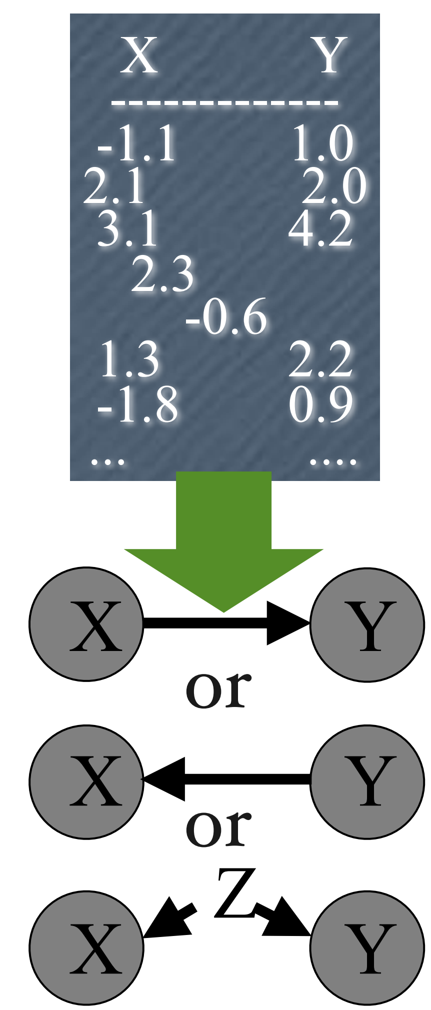
  - Related to this type of "independence":



- Start with the linear case

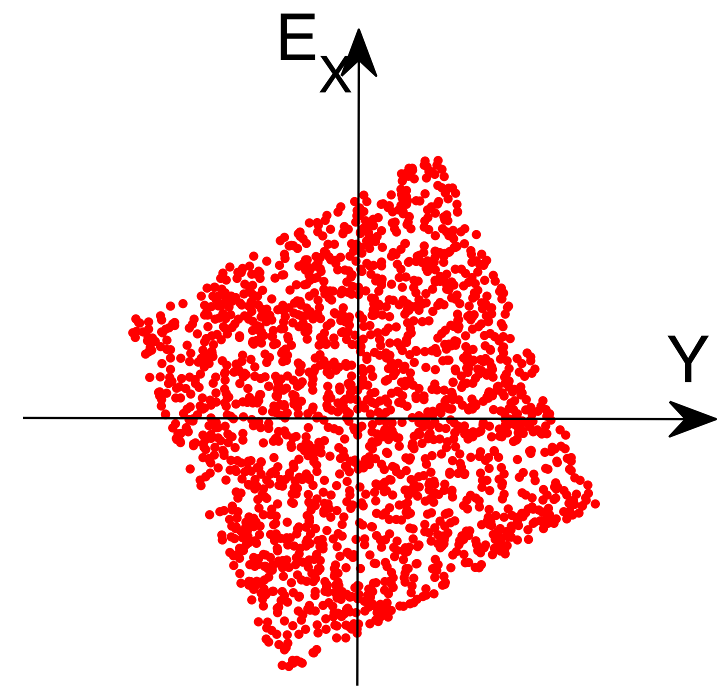
  $Y = aX + E$, where $E \perp\!\!\!\perp X$

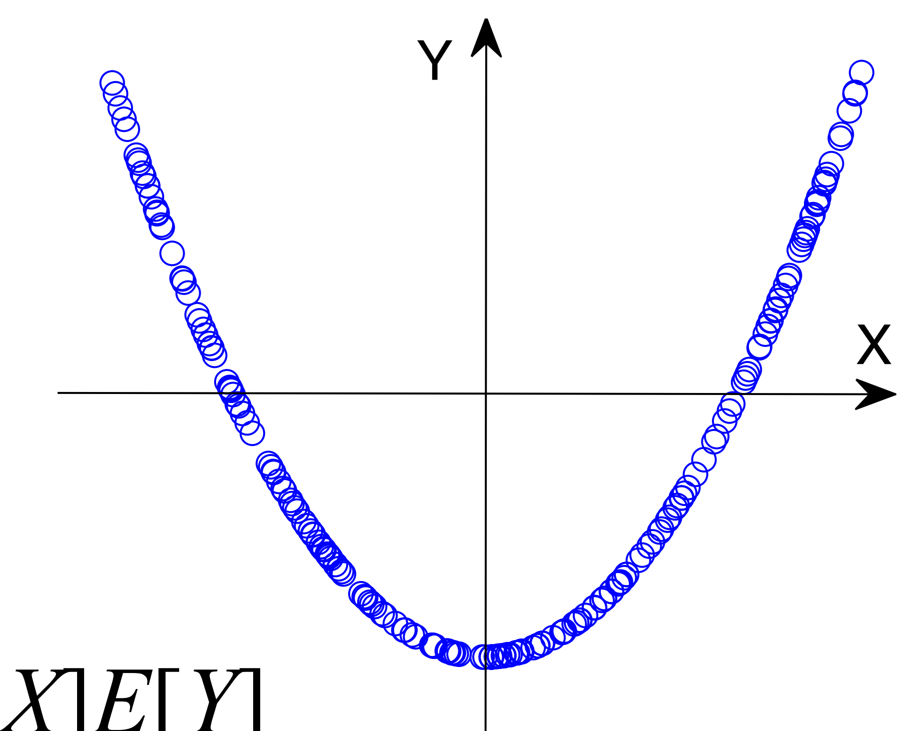
- Determine causal direction in the two-variable case? Identifiability!



(Conditional) Independence

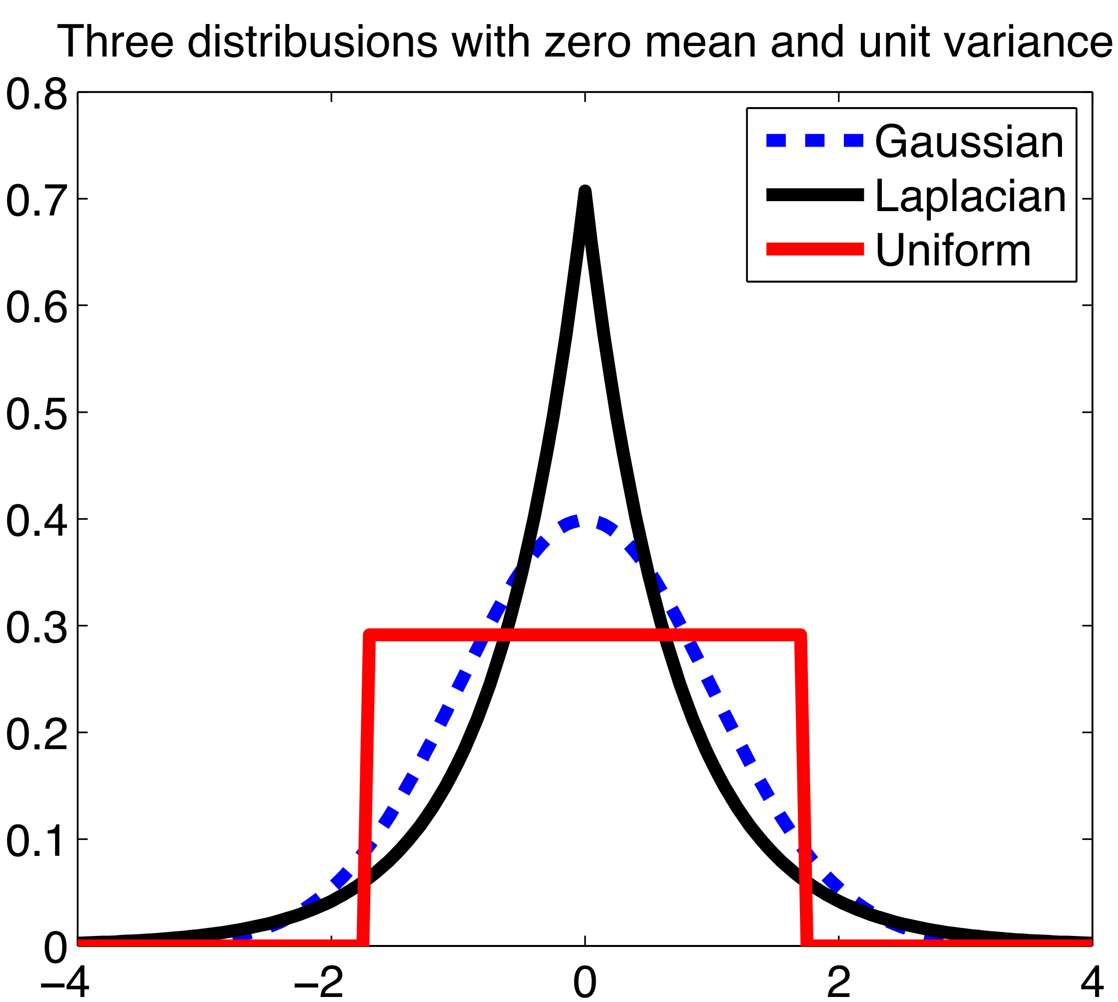
- $X \perp\!\!\!\perp Y$ iff $p(X, Y) = p(X)p(Y)$
  - or $p(X|Y) = P(X)$: Y not informative to X
- $X \perp\!\!\!\perp Y \mid Z$ iff $p(X, Y|Z) = p(X|Z)p(Y|Z)$
  - or, $p(X|Y,Z) = p(X|Z)$: given Z, Y not informative to X
- Divide & conquer, remove irrelevant info…
- By construction, regression residual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

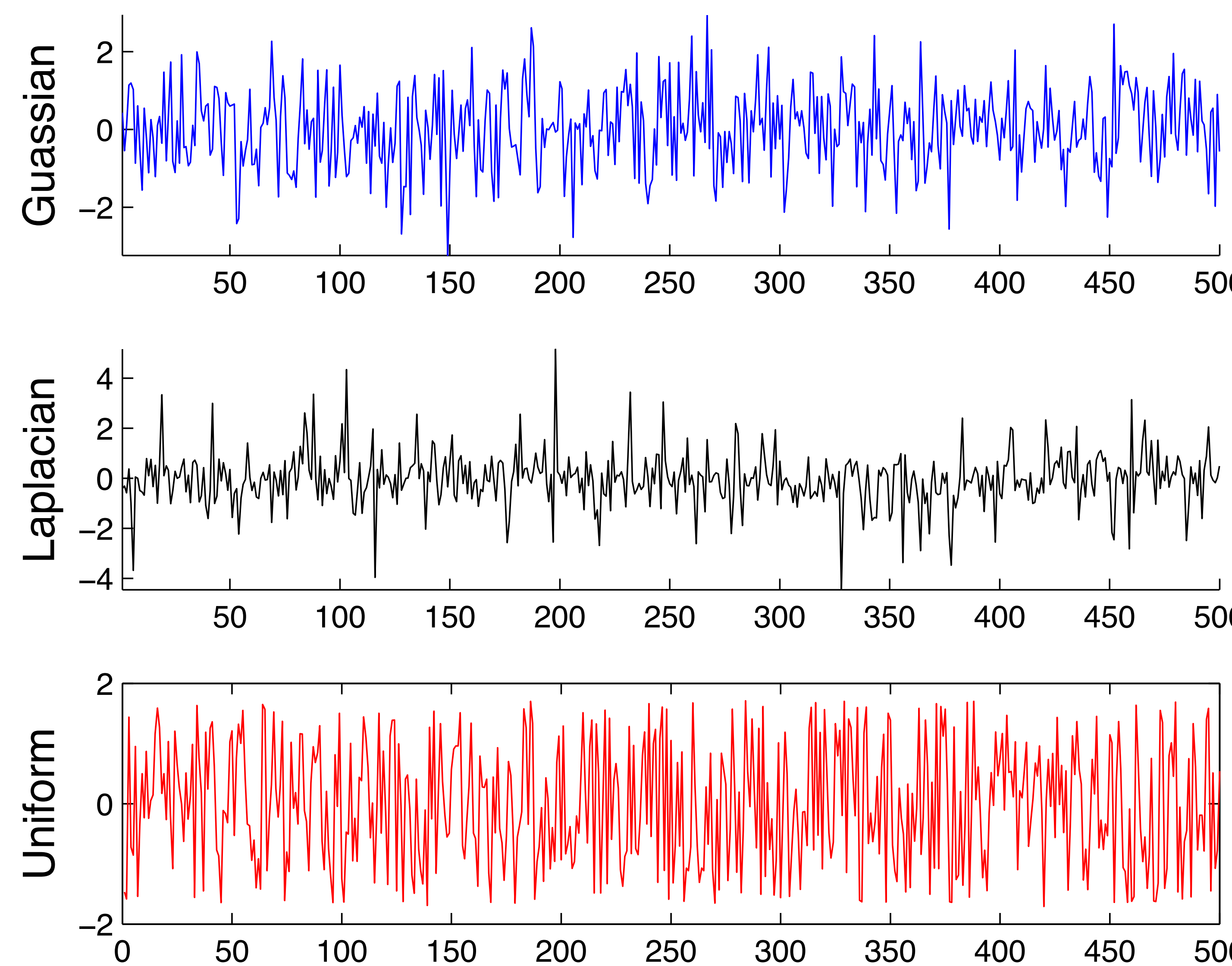




Uncorrelatedness: $E[XY] = E[X]E[Y]$

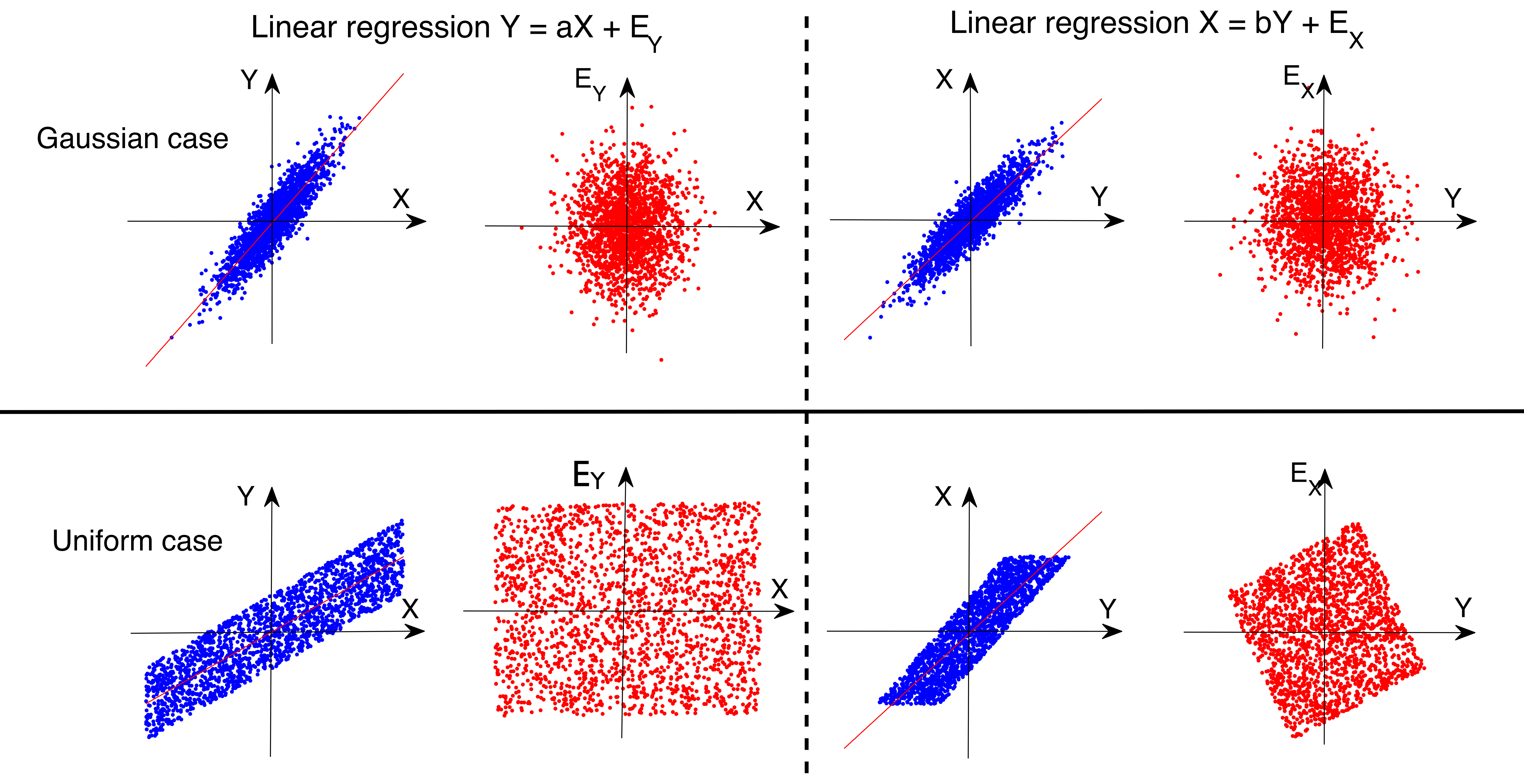
Gaussian vs. Non-Gaussian Distributions





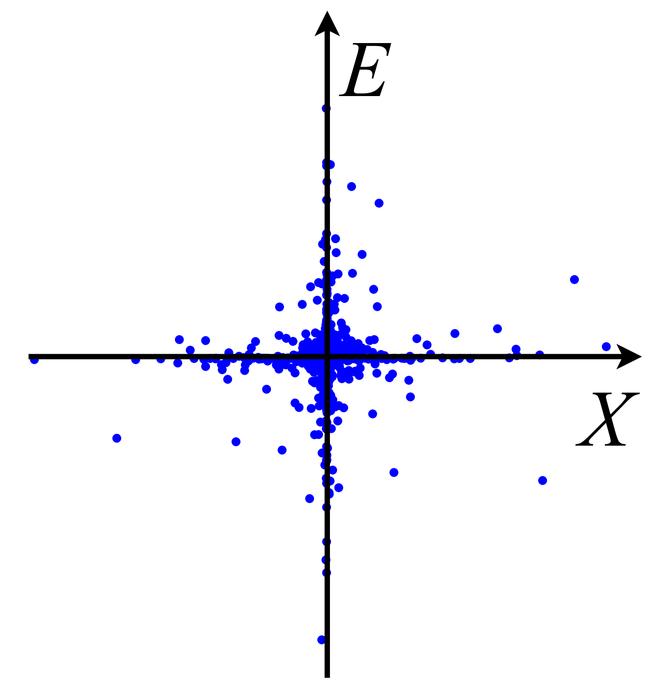
Causal Asymmetry the Linear Case: Illustration

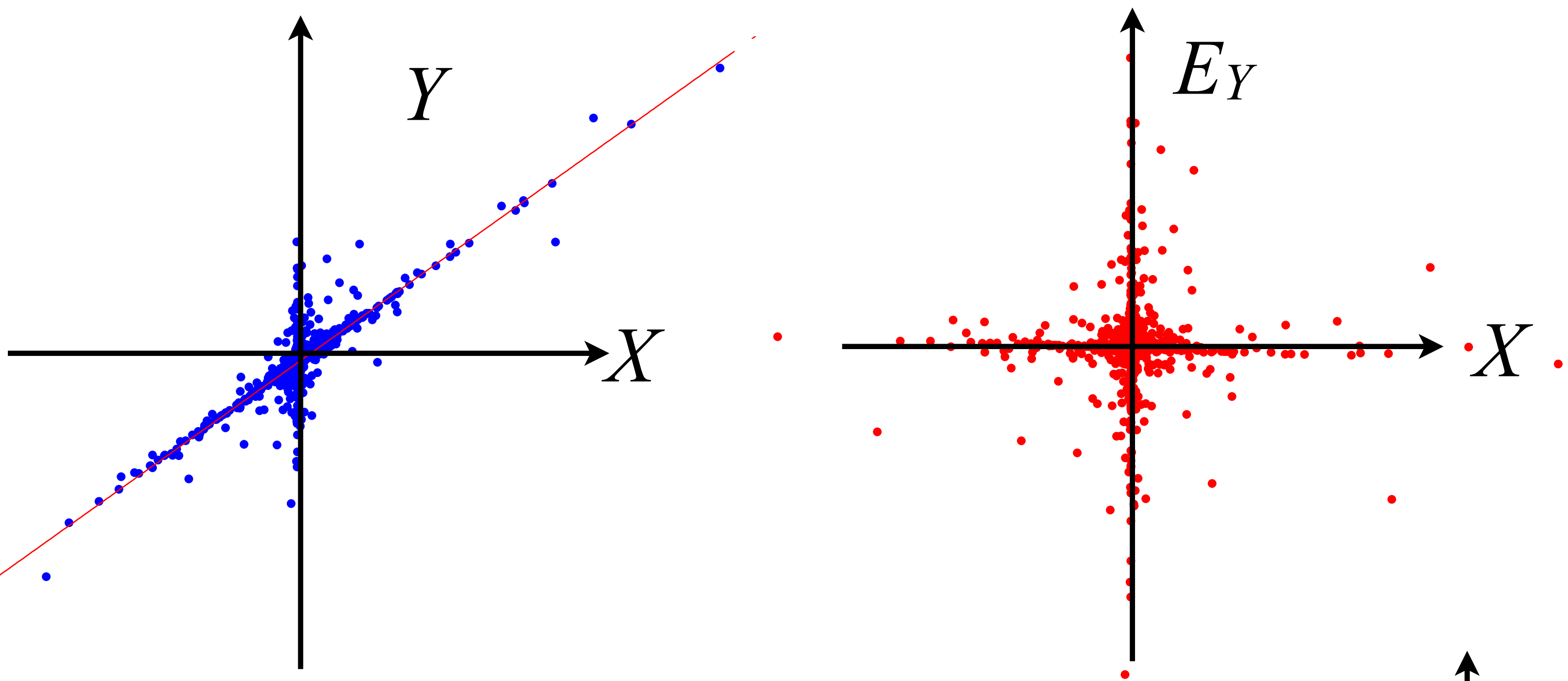
Data generated by $Y = aX + E$ (i.e., $X \rightarrow Y$):

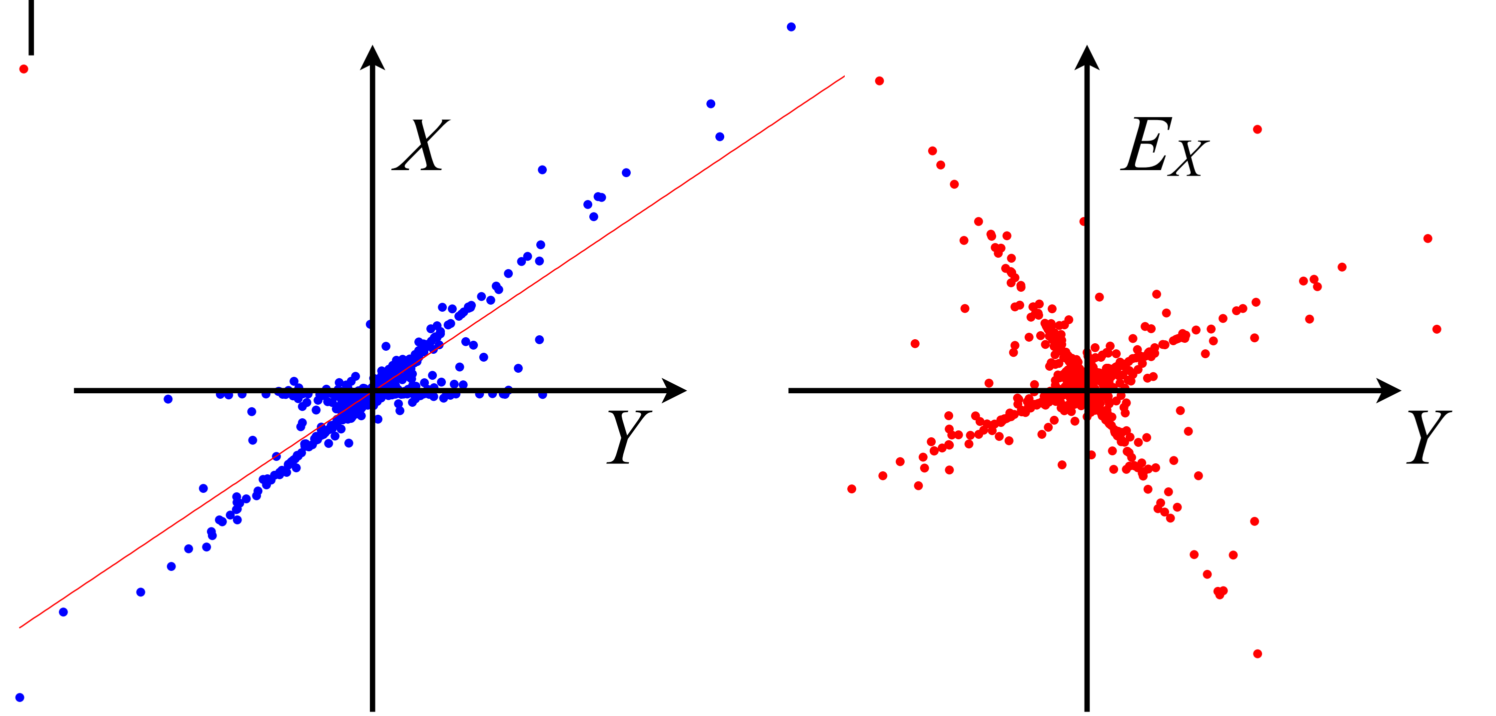


Super-Gaussian Case

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







More Generally, LiNGAM Model

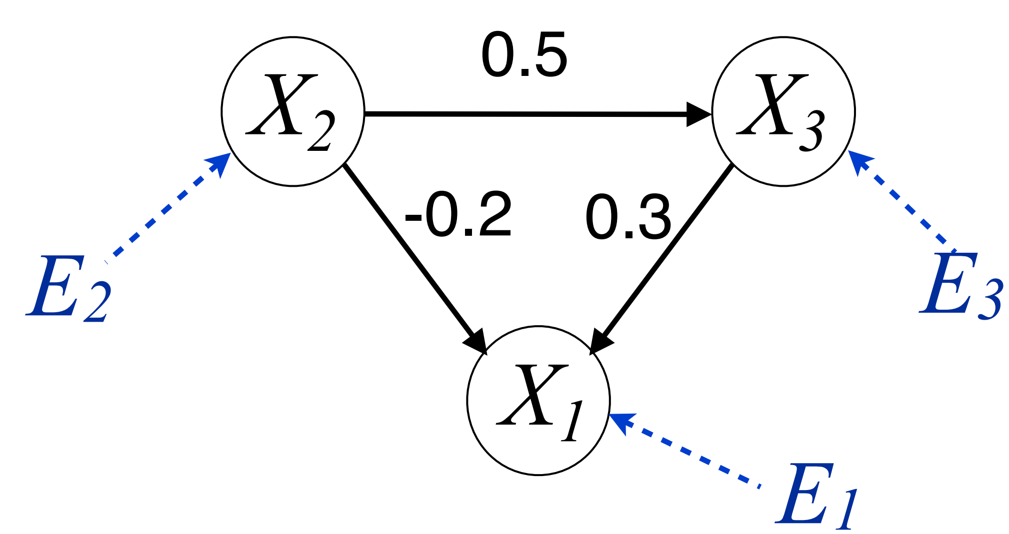
- Linear, non-Gaussian, acyclic causal model (LiNGAM) (Shimizu et al., 2006):

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

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

  - Example:

$$\begin{aligned} X_2 &= E_2, \\ X_3 &= 0.5X_2 + E_3, \\ X_1 &= -0.2X_2 + 0.3X_3 + E_1. \end{aligned}$$

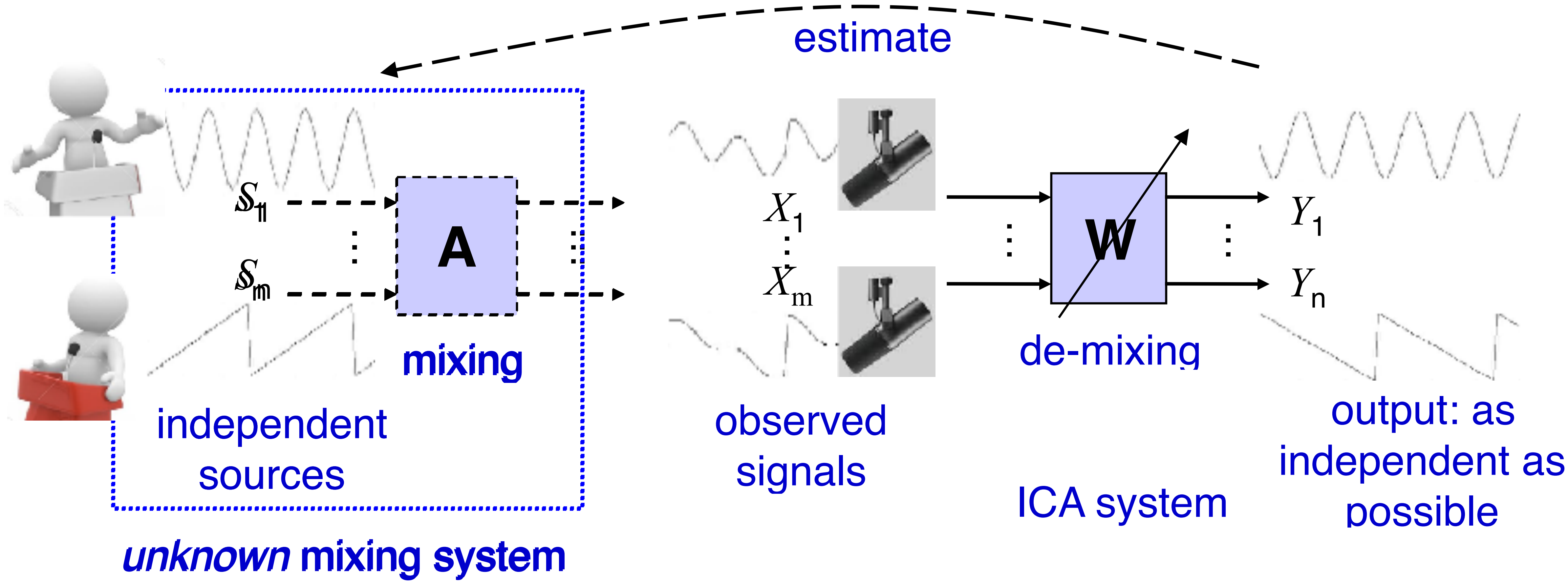


Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.

Identifiability of Causal Direction in the Linear Case

- Supported by the "independent component analysis" theory
- Later will consider a more general nonlinear setting, and you'll see the linear-Gassian case is one of the few non-identifiable situations

Independent Component Analysis



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

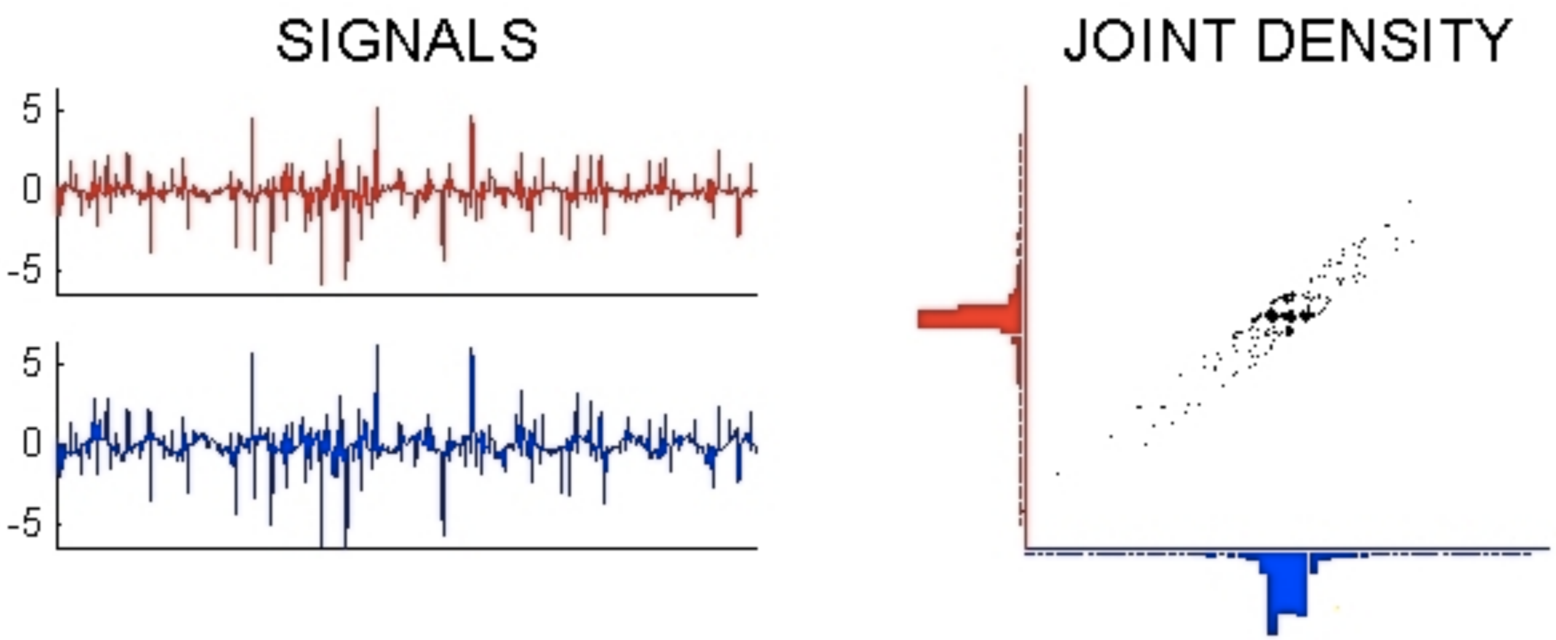
$$\begin{array}{l} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{l} s_1 \\ s_2 \end{array}$$

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

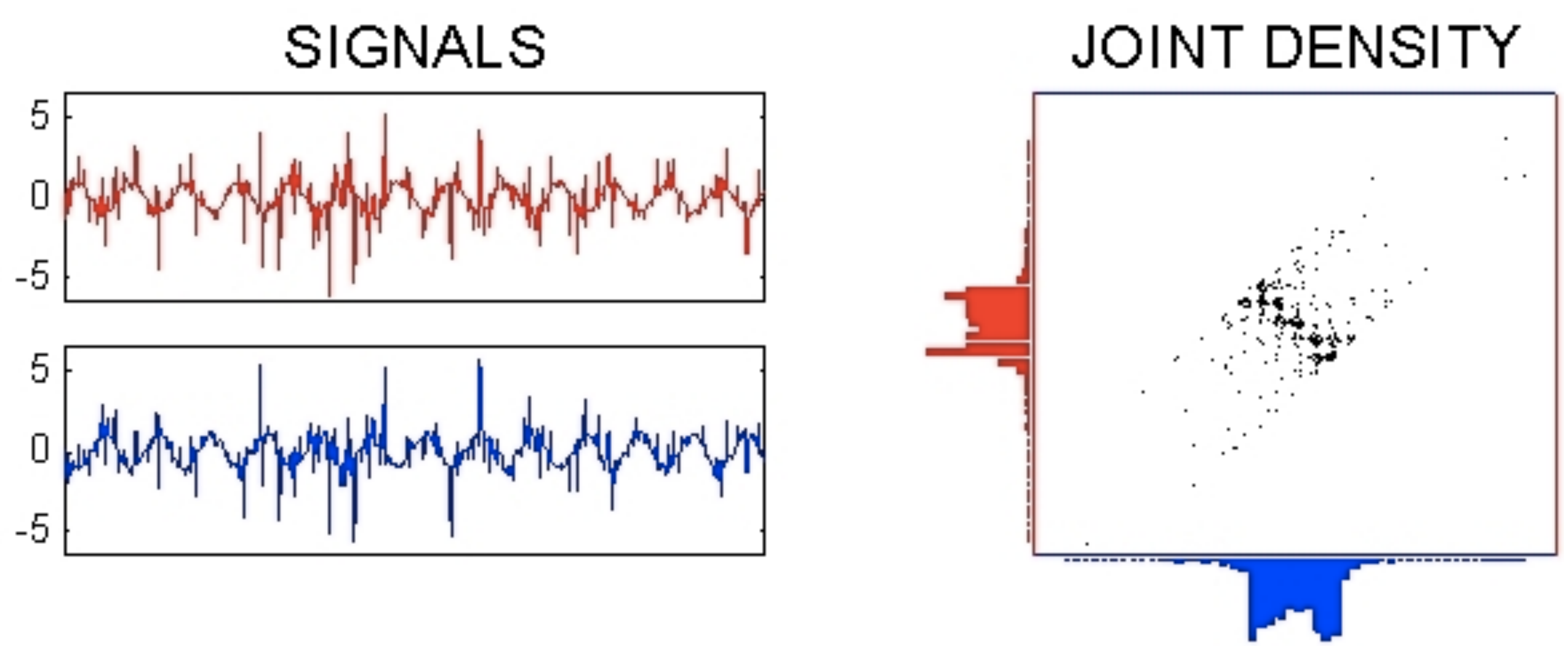
Then A can be estimated up to column scale and permutation indeterminacies

Hyvärinen et al., Independent Component Analysis, 2001

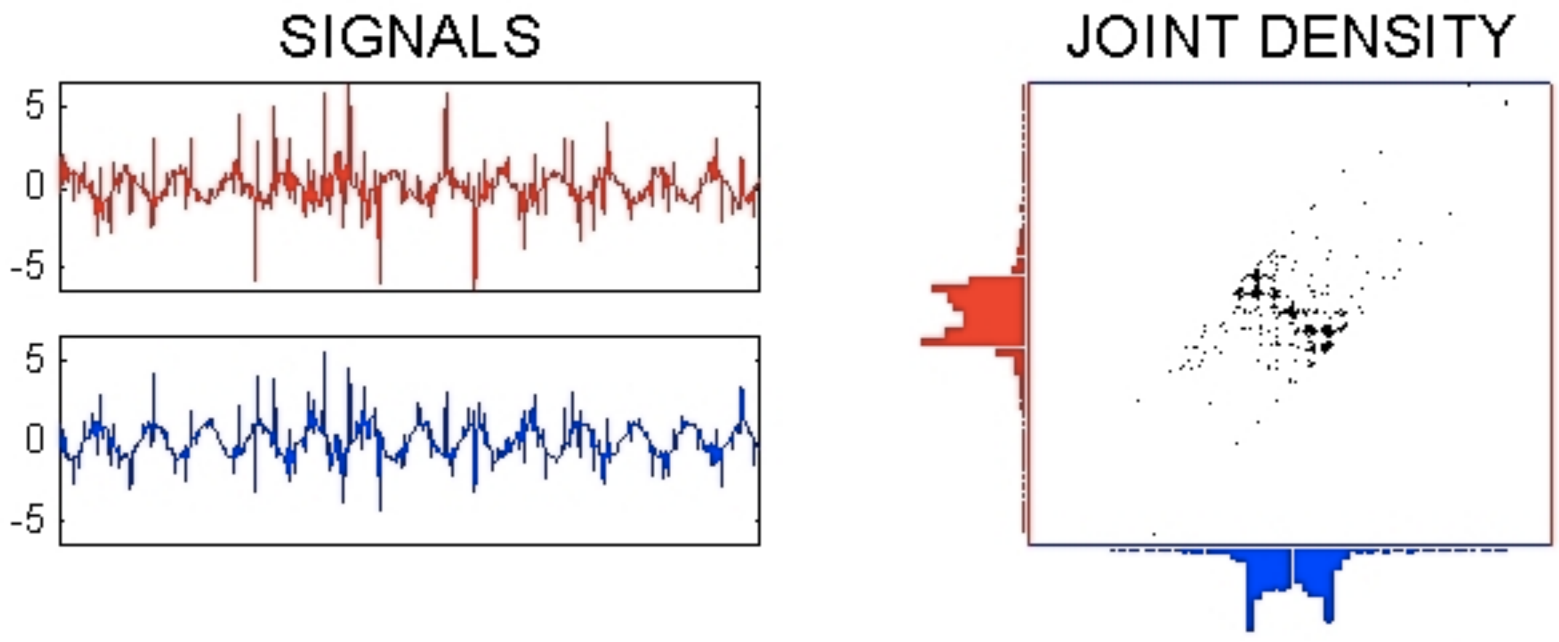
A Demo of the ICA Procedure



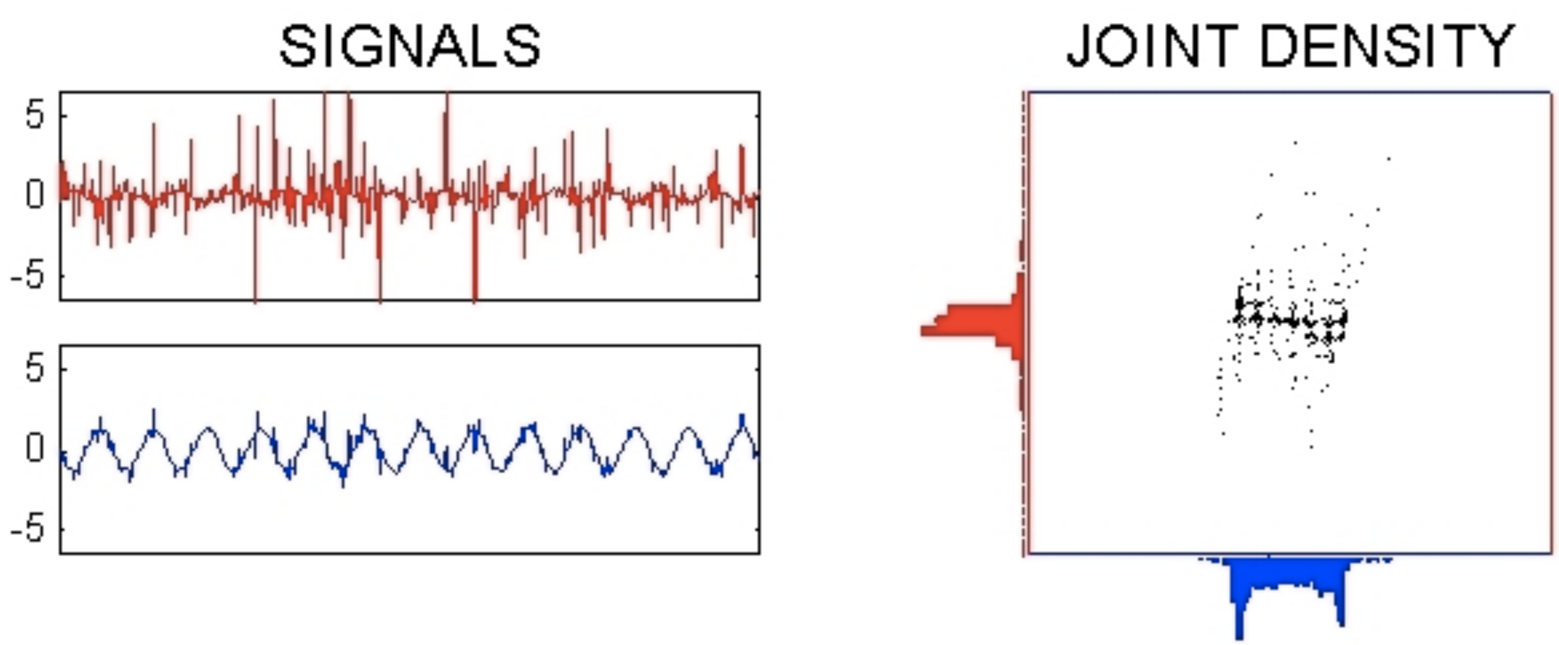
Input signals and density



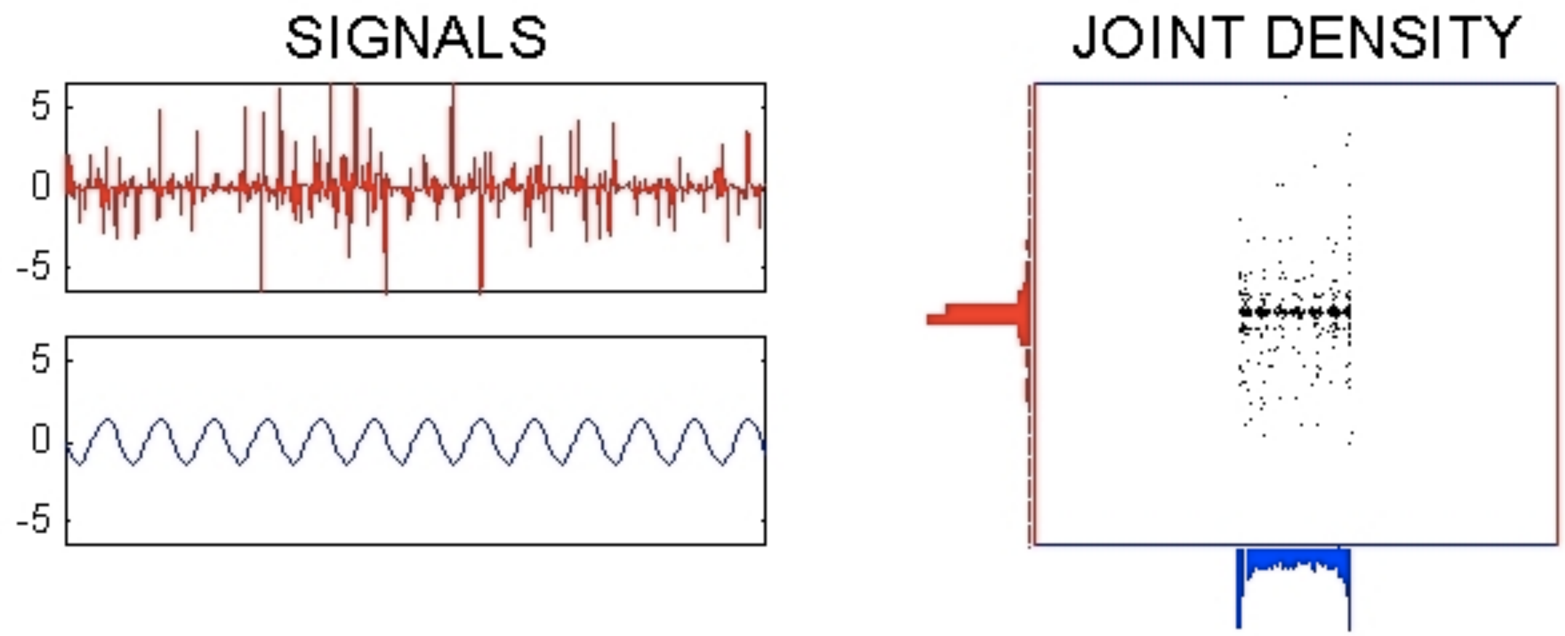
Whitened signals and density



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

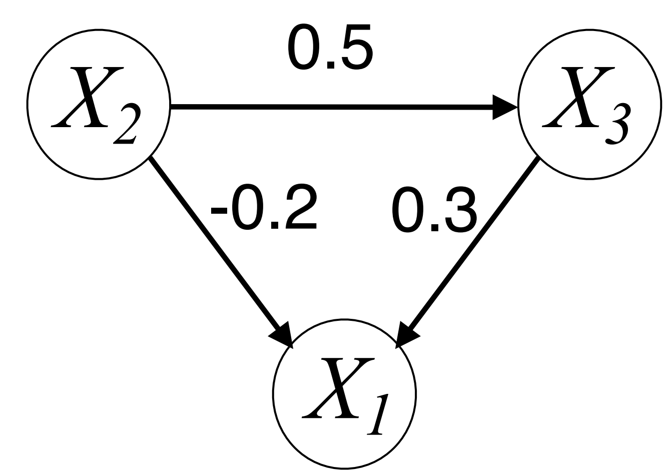
- LiNGAM: $X_i = \sum_{j:\text{ parents of } i} b_{ij} X_j + E_i$ *or* $\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E} \Rightarrow$ $\mathbf{E} = (\mathbf{I}-\mathbf{B})\mathbf{X}$
  - $\mathbf{B}$ has special structure: acyclic relations
- ICA: $\mathbf{Y} = \mathbf{W}\mathbf{X}$
- $\mathbf{B}$ can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

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

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

1. First permute the rows of $\mathbf{W}$ 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. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

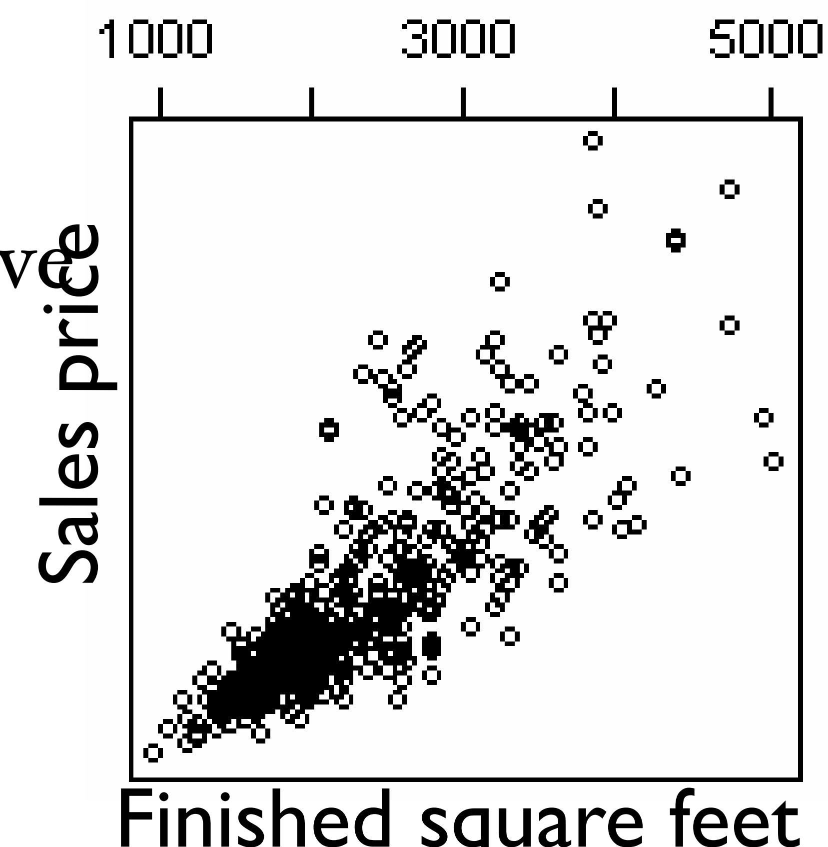
So we have the causal relation:



Limitations of LiNGAM

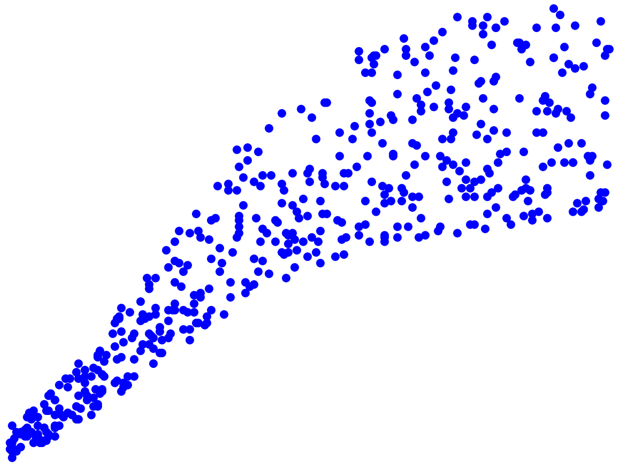
- Confounders
- Measurement noise
- Feedbacks $X_1 \rightarrow X_2$
- Selection bias $X \rightarrow Y \rightarrow S$
- Various nonlinearities
  - Nonlinear function with independent additive noise
  - Sensor / measurement distortion
  - With heteroscedastic noise
  - More general forms

linear or nonlinear?



Zhang et al., Learning causality and causality-related learning, National Science Review, 2018

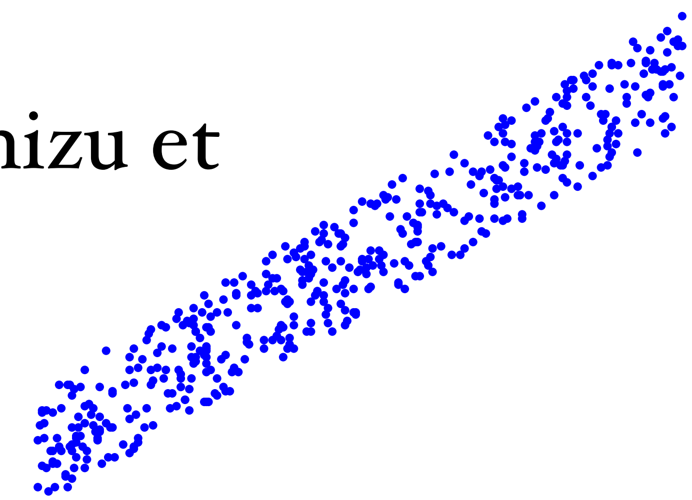
More General Functional Causal Models



FCMs with Which Causal Direction is Generally Identifiable

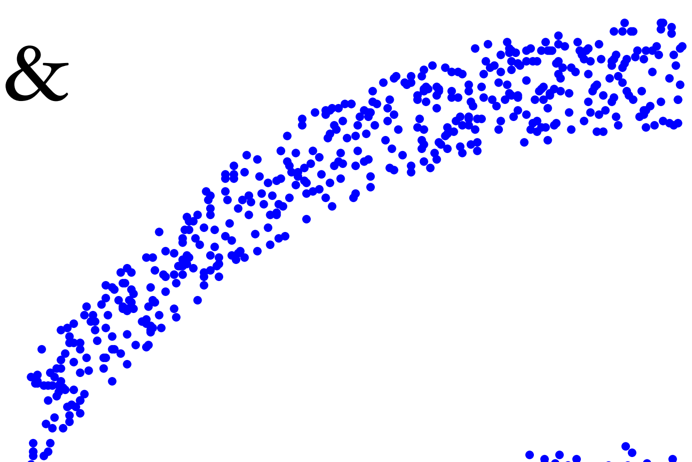
- Linear non-Gaussian acyclic causal model (Shimizu et al., '06)

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



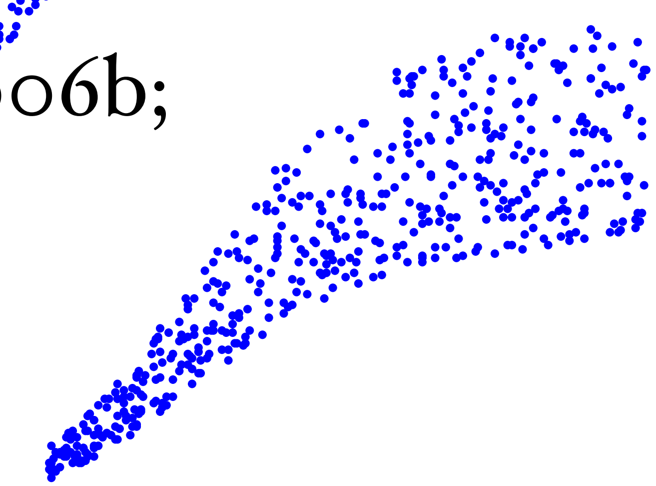
- Additive noise model (Hoyer et al., '09; Zhang & Hyvärinen, '09b)

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



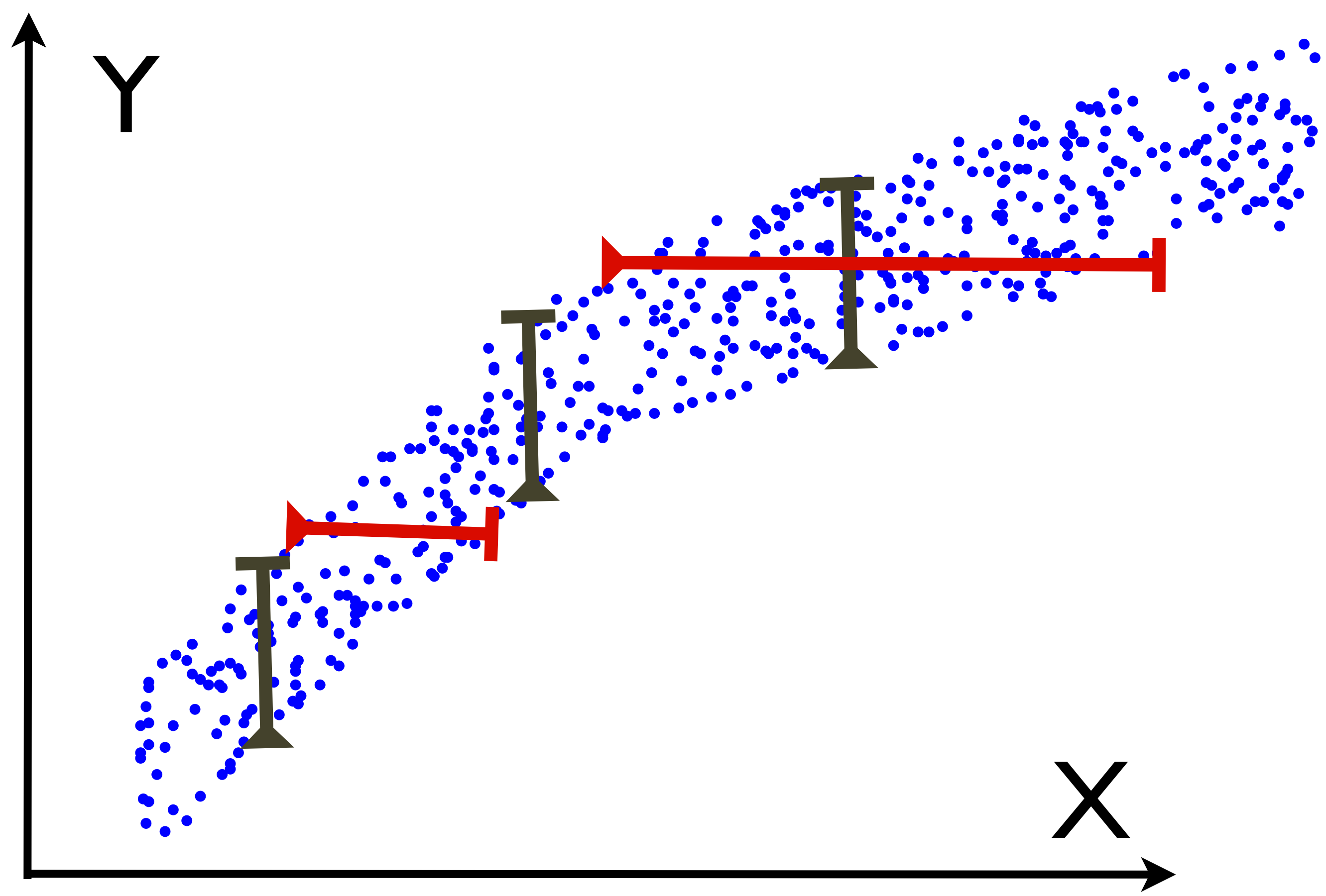
- Post-nonlinear causal model (Zhang & Chen, 2006b; Zhang & Hyvärinen, '09a)

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



Causal Asymmetry with Nonlinear Additive Noise: Illustration

$$Y = f(X) + E \text{ with } E \perp\!\!\!\perp X$$

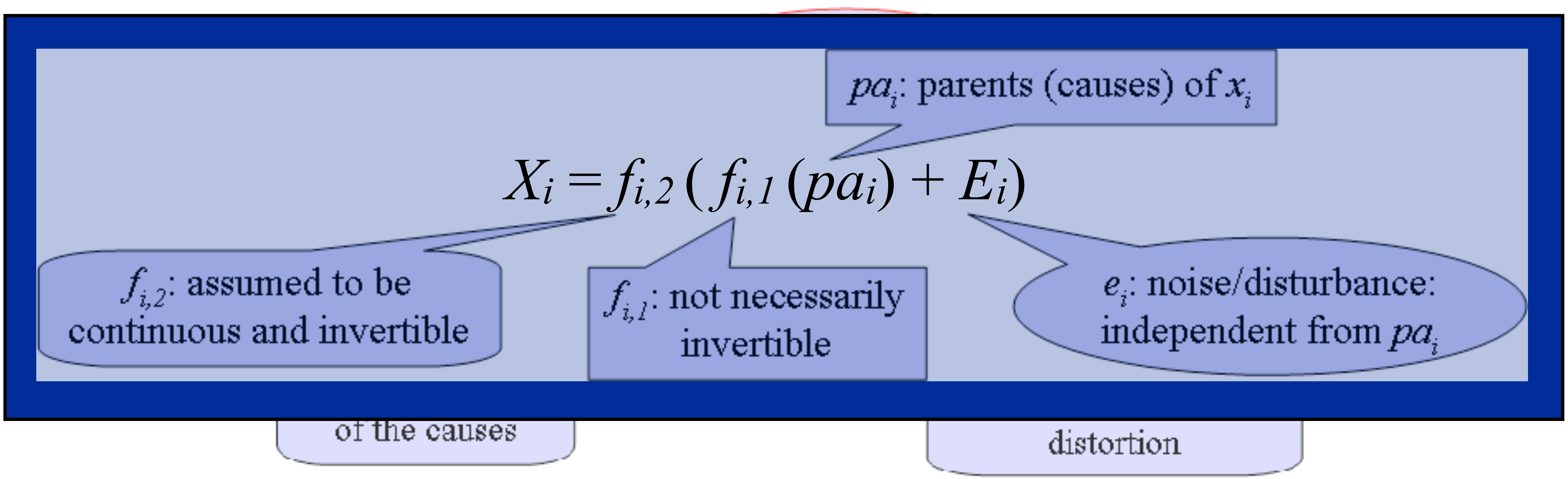


(Hoyer et al., 2009)

Post-Nonlinear (PNL) Causal Model

(Zhang & Chan, 2006; Zhang & Hyvärinen, ‘09a)

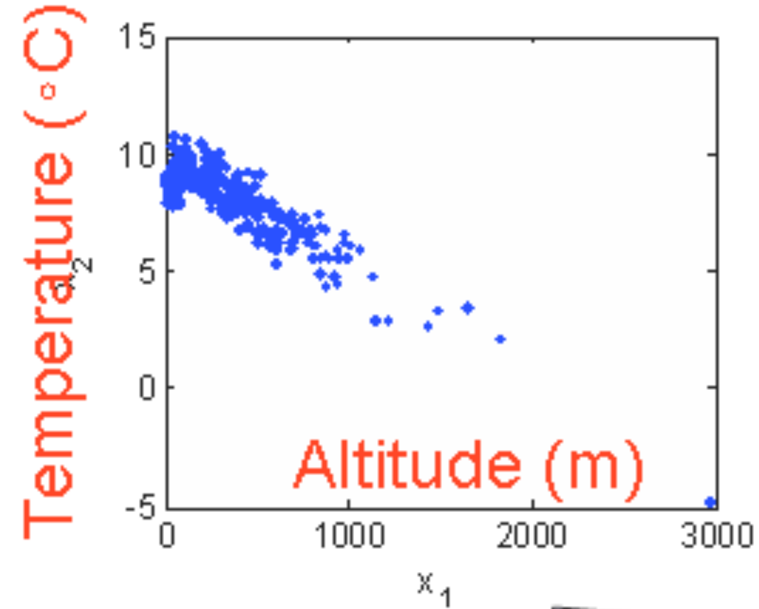
- Without prior knowledge, the assumed model is expected to be
  - general enough: adapt to approximate the true generating process
  - identifiable: asymmetry in causes and effects

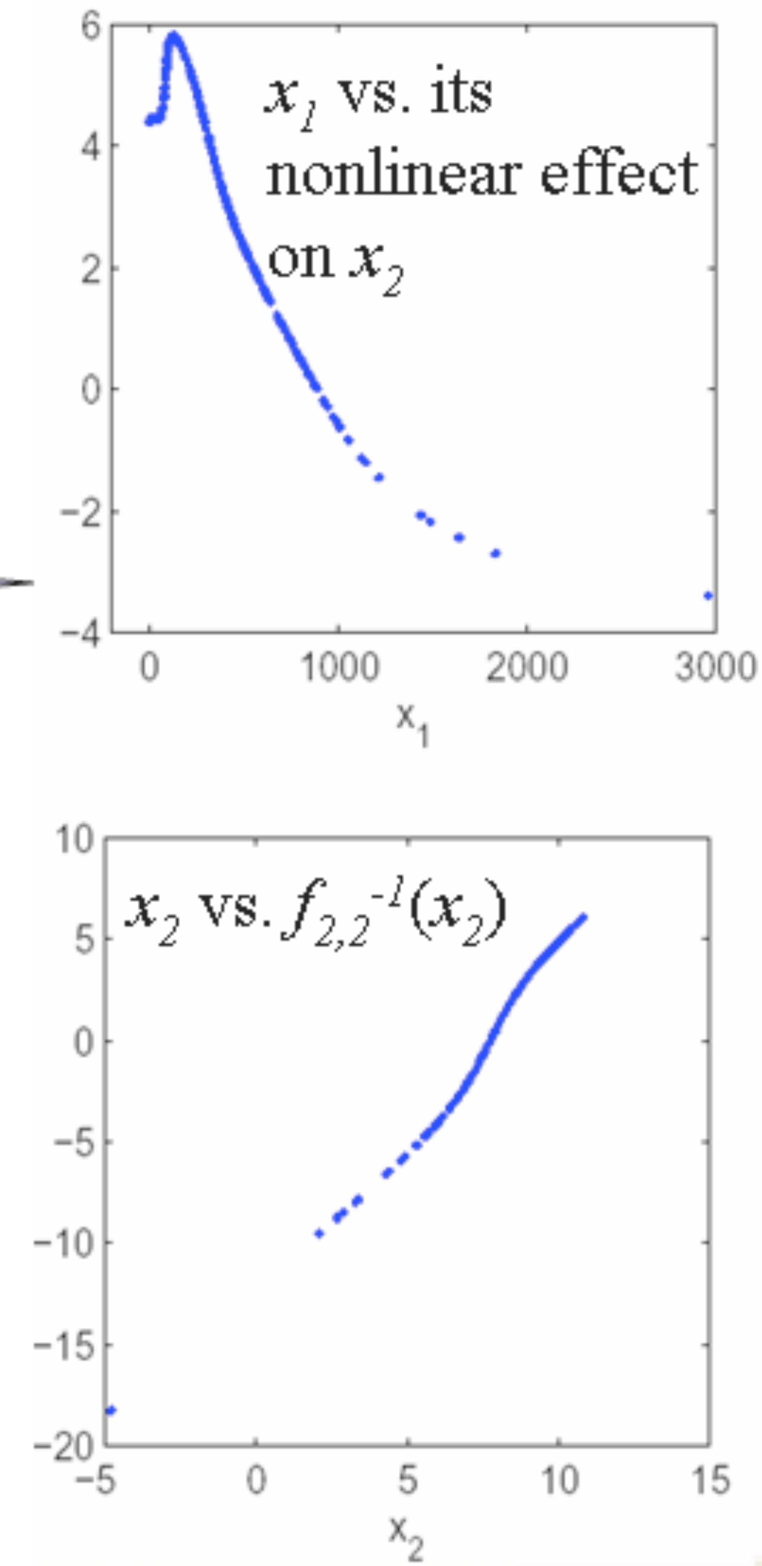


- Special cases: linear models; nonlinear additive noise models; multiplicative noise models: $Y = X \cdot E = \exp\big(\log(X) + \log(E)\big)$

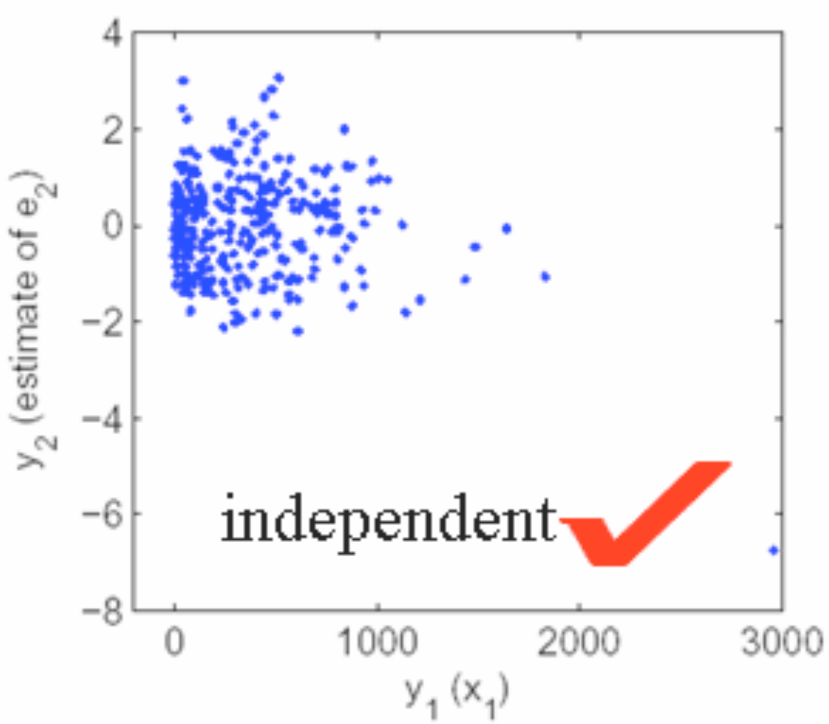
Data Set 1

with PNL Model

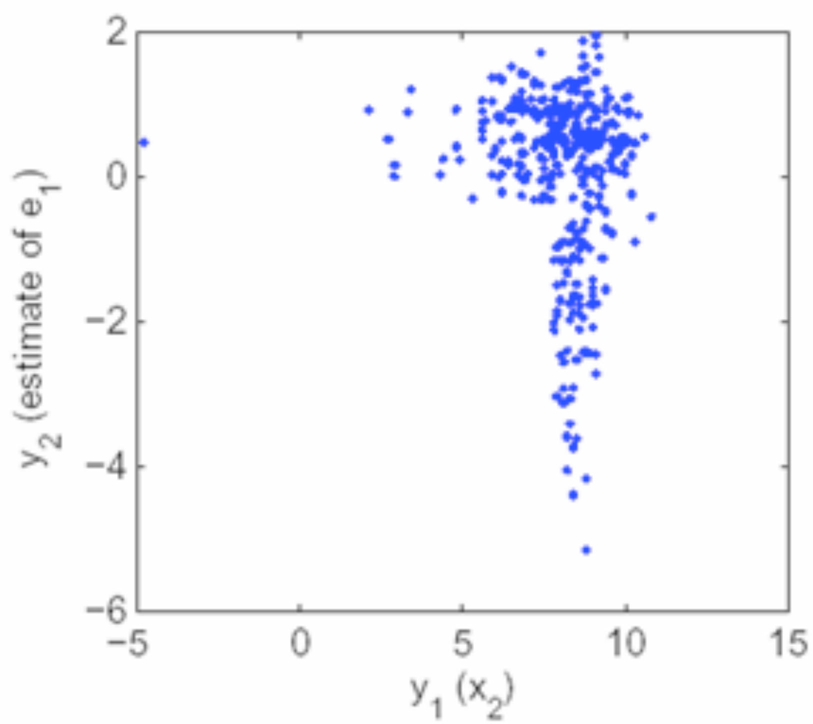




(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$



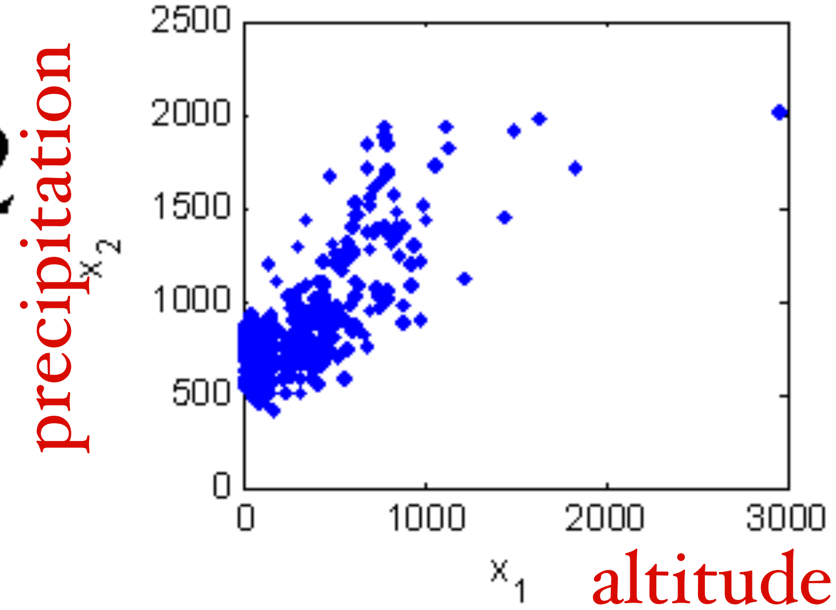
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$



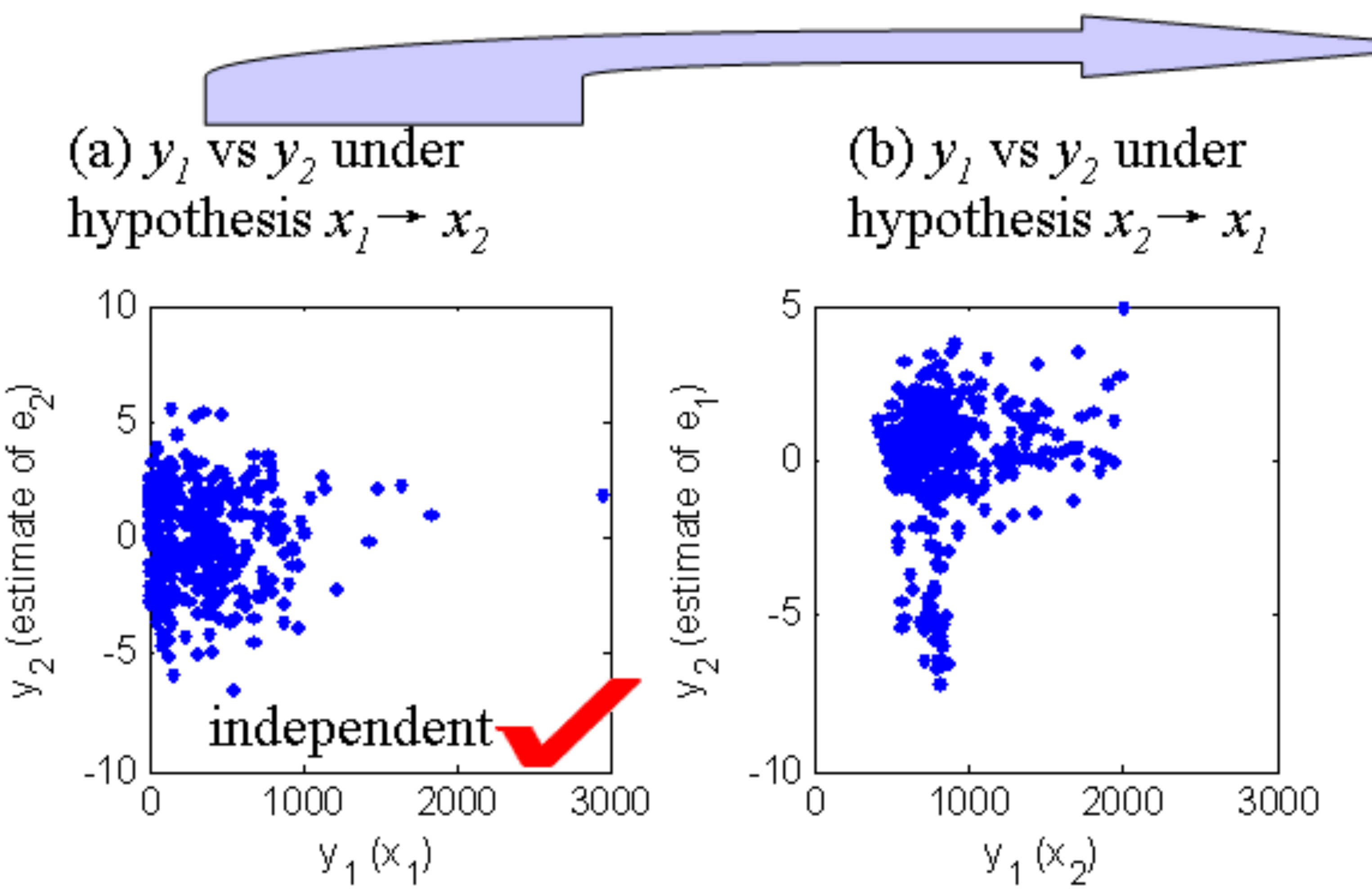
Independence test results on y_1 and y_2 with different assumed causal relations

Data Set	$x_1 \rightarrow x_2$ assumed		$x_2 \rightarrow x_1$ assumed	
	Threshold ($\alpha = 0.01$)	Statistic	Threshold ($\alpha = 0.01$)	Statistic
#1	2.3×10^{-3}	1.7×10^{-3}	2.2×10^{-3}	6.5×10^{-3}

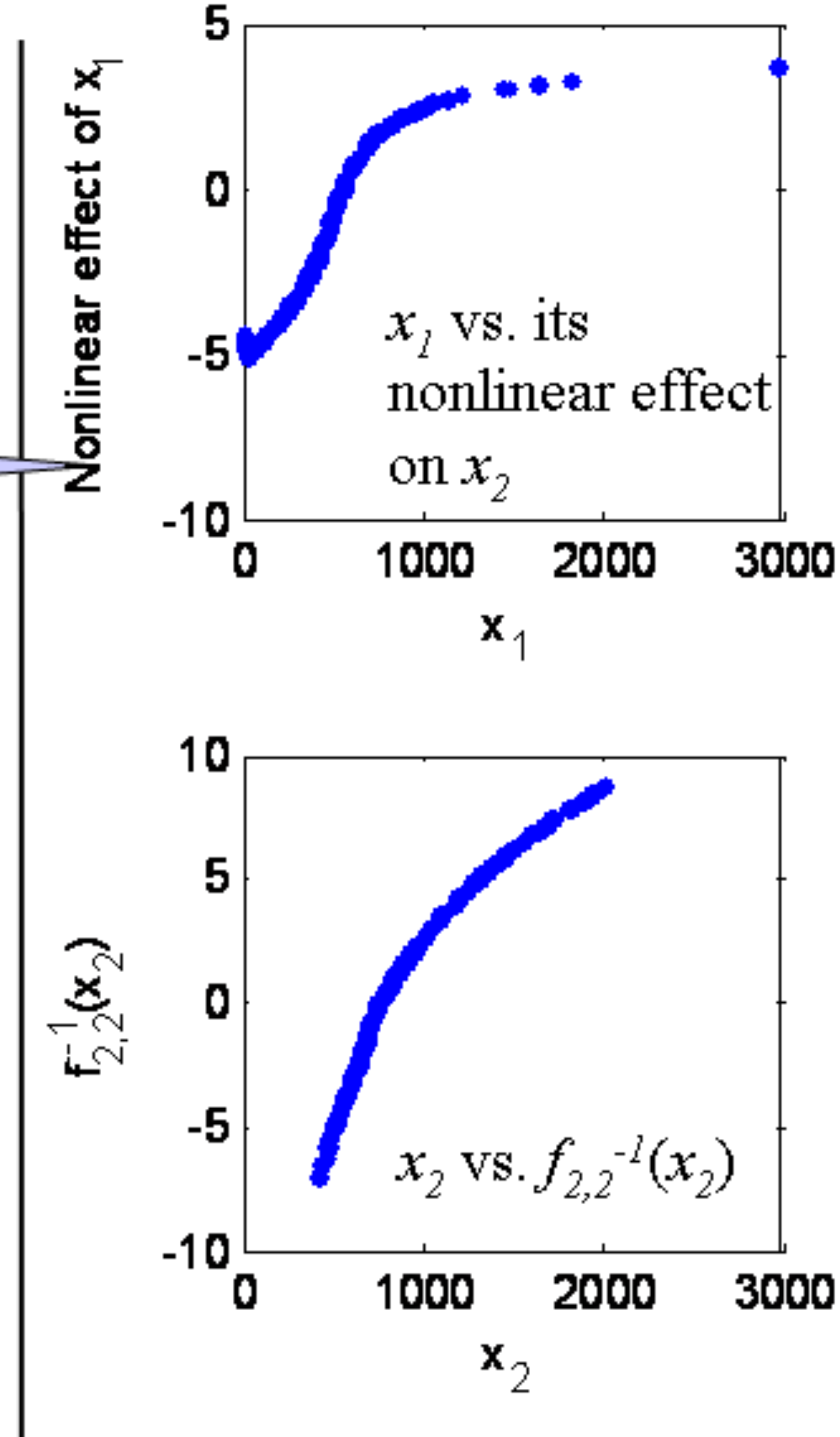
Data Set 2



precipitation
x 2
2500
2000
1500
1000
500
0
0
1000
2000
3000
x 1
altitude

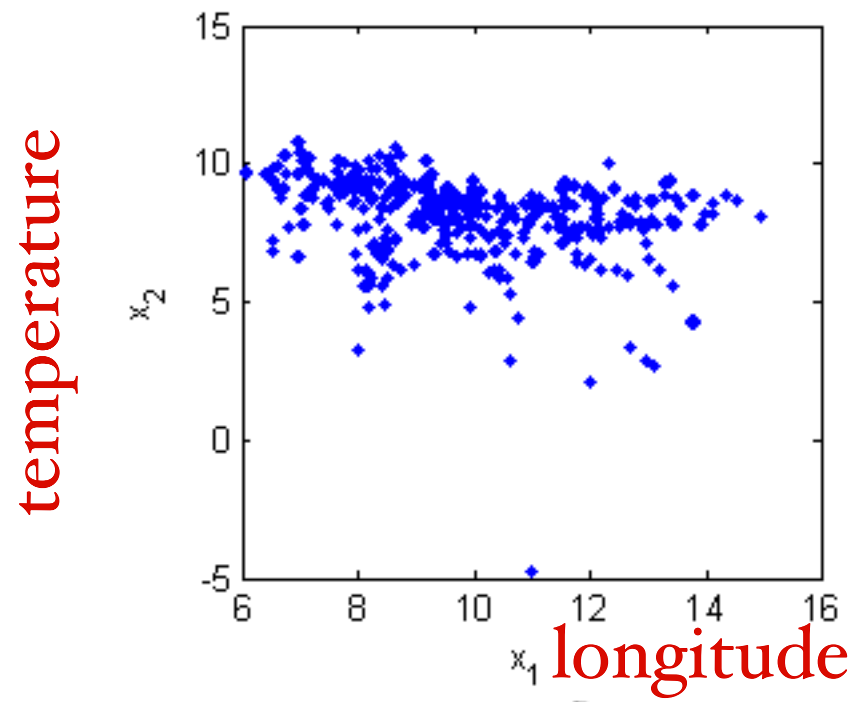


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$
y 2 (estimate of e 2)
10
5
0
-5
-10
0
1000
2000
3000
y 1 (x 1)
independent
y 2 (estimate of e 1)
5
0
-5
-10
0
1000
2000
3000
y 1 (x 2)



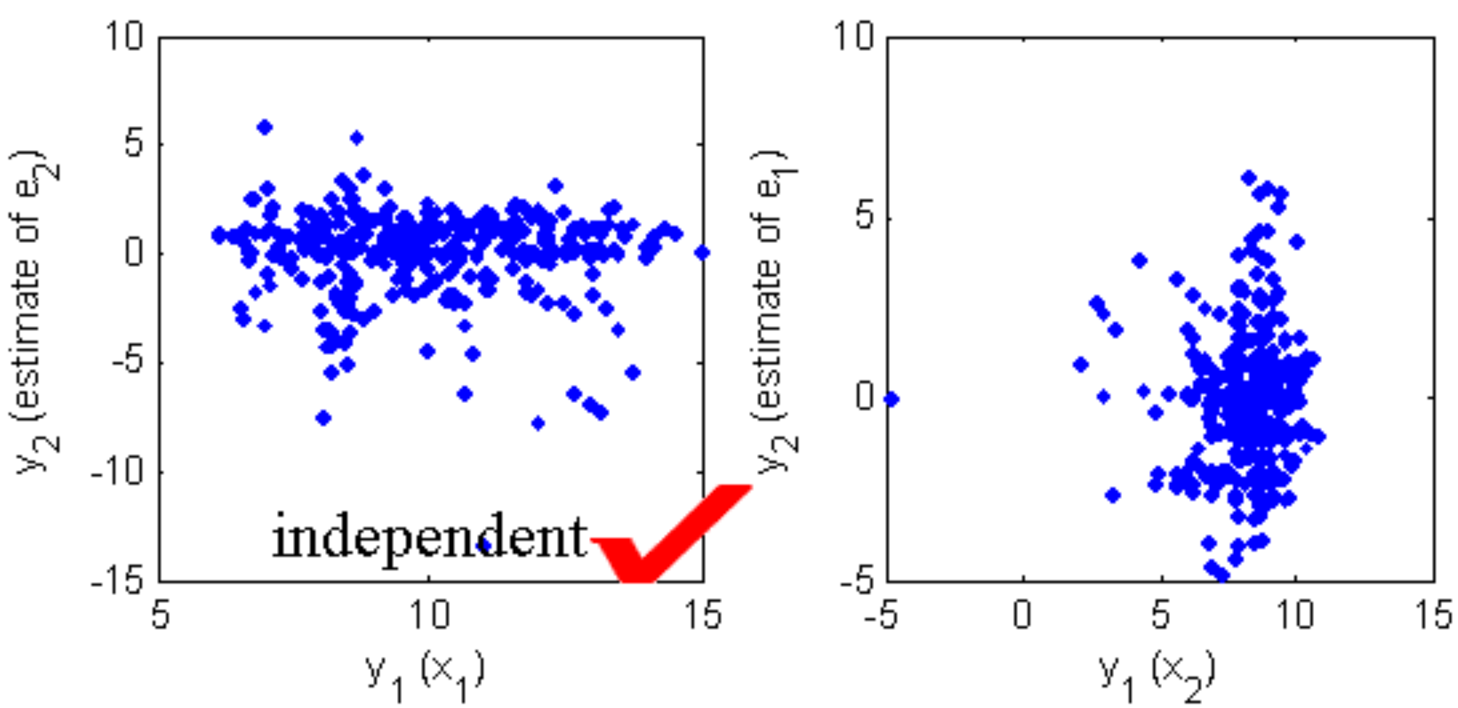
Nonlinear effect of x 1
5
0
-5
-10
0
1000
2000
3000
x 1
x_1 vs. its nonlinear effect on x_2
$f_{2,2}^{-1}(x_2)$
10
5
0
-5
-10
0
1000
2000
3000
x 2
x_2 vs. $f_{2,2}^{-1}(x_2)$

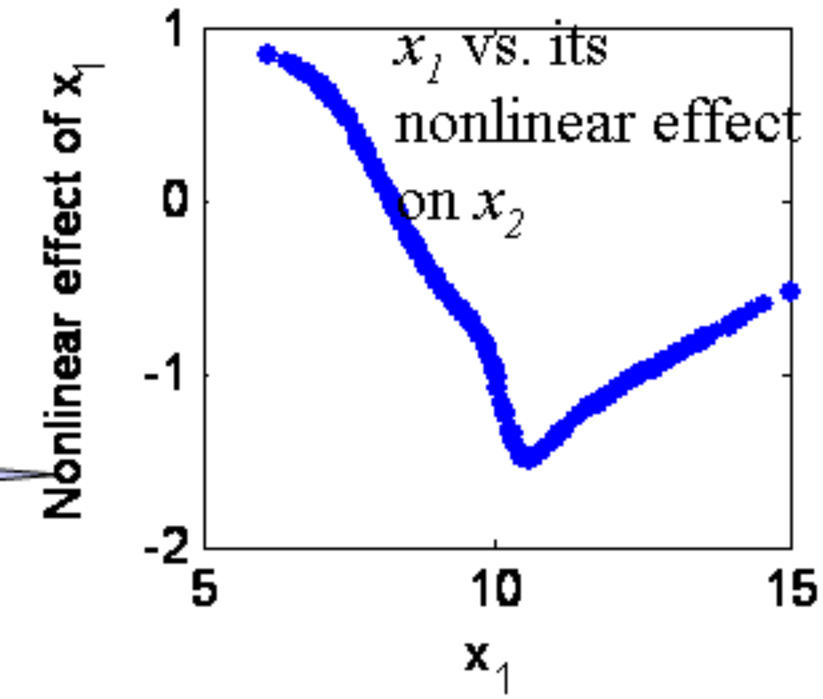
Data Set 3

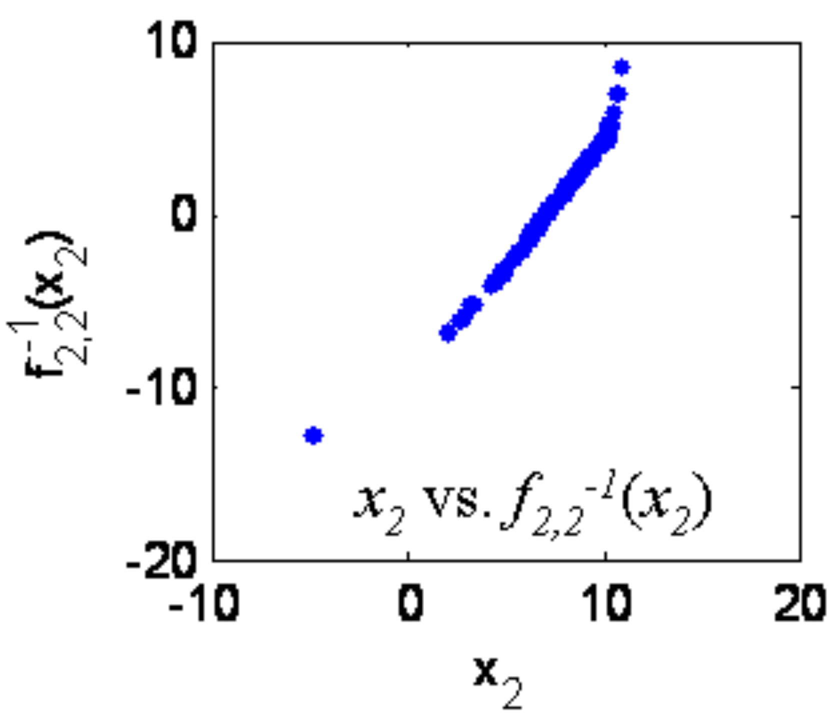


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

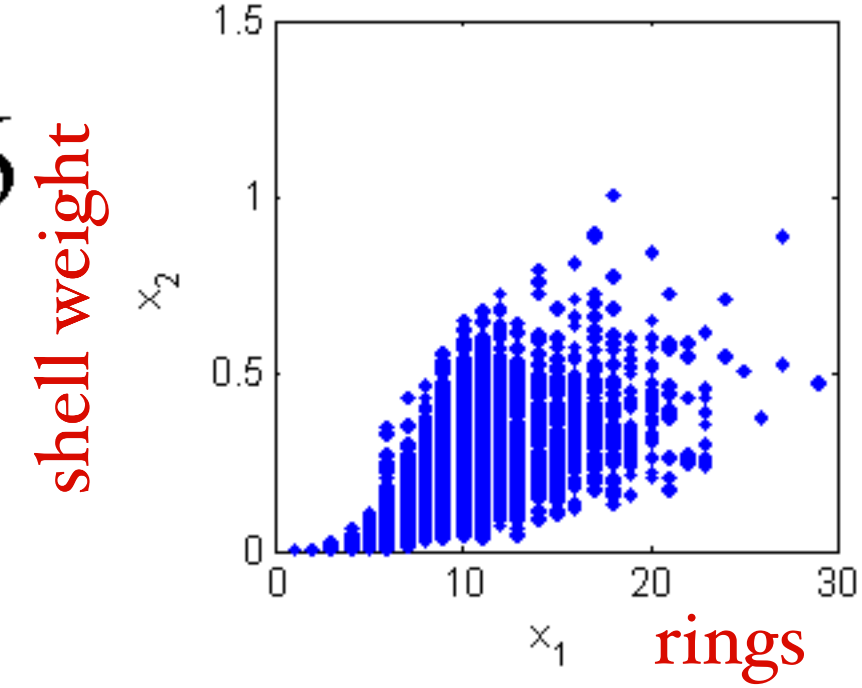
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$

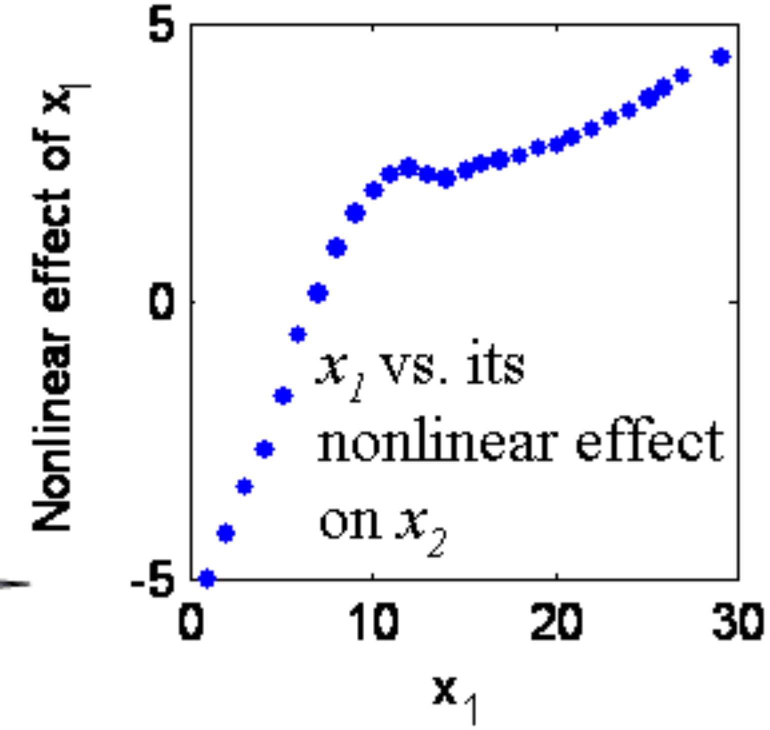






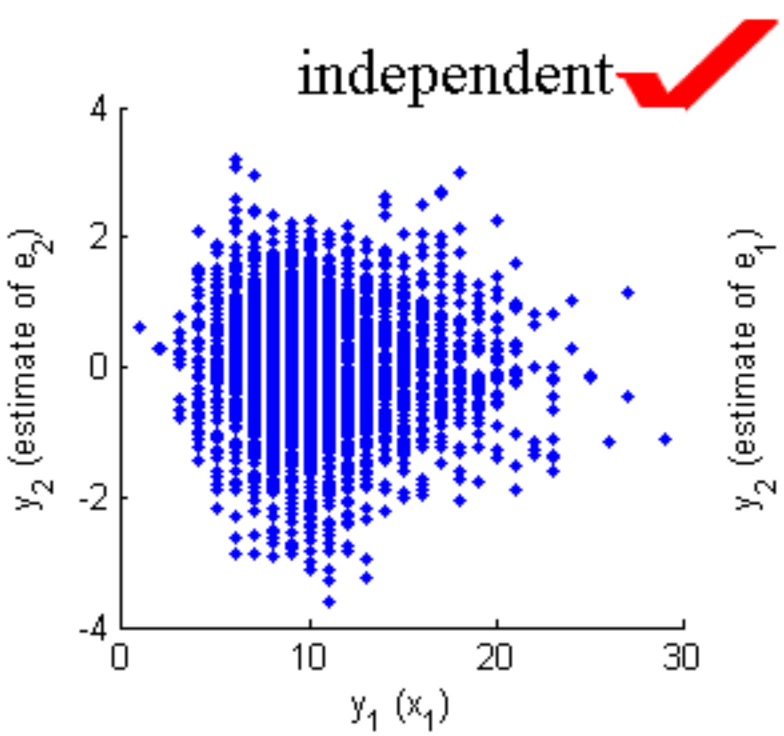
Data Set 6

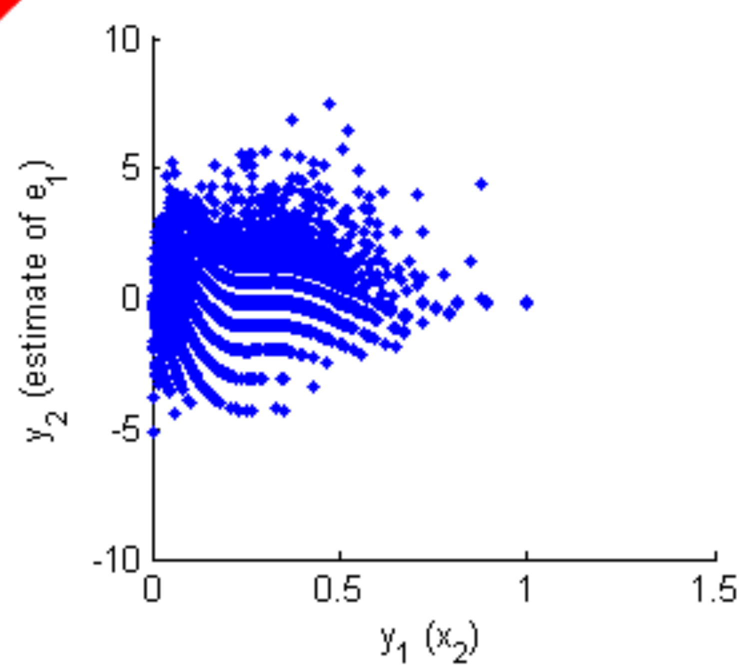


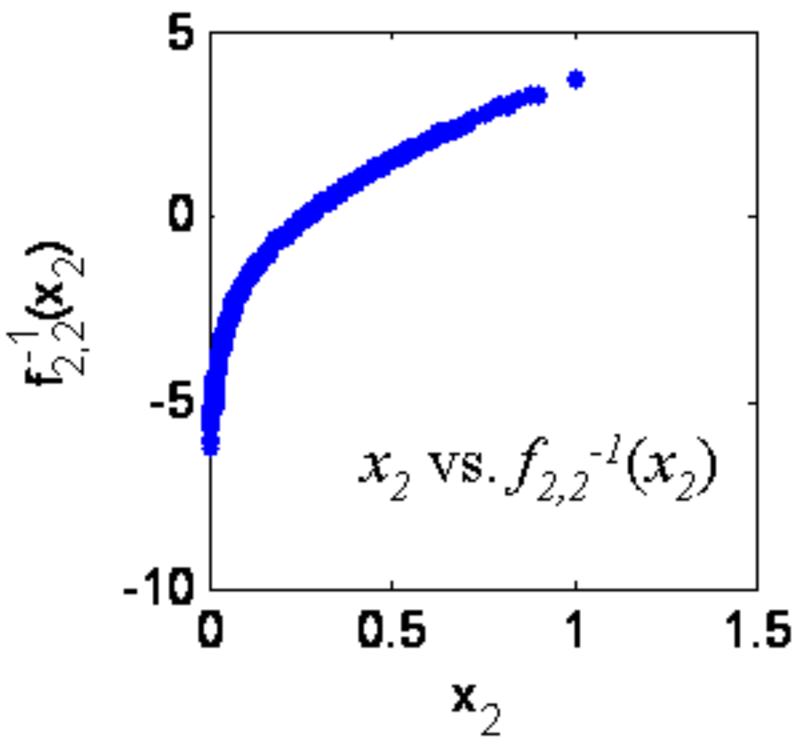


(a) y_1 vs y_2 under hypothesis $x_1 \to x_2$

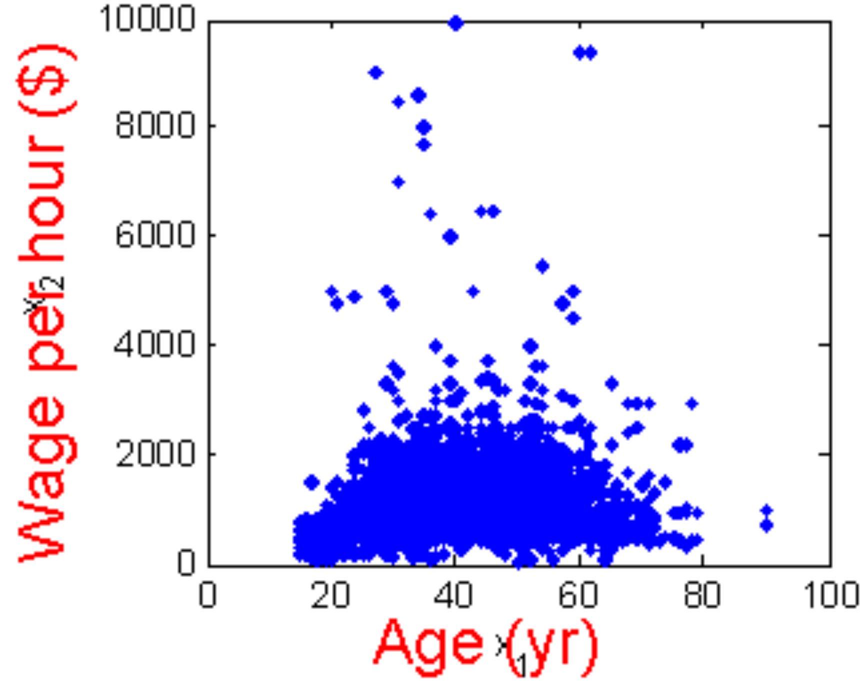
(b) y_1 vs y_2 under hypothesis $x_2 \to x_1$

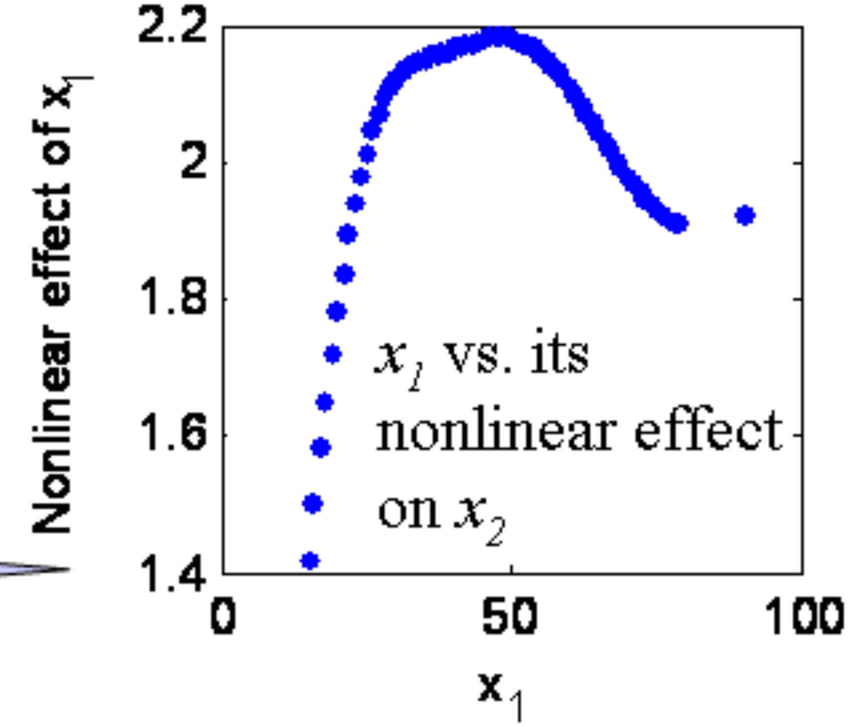






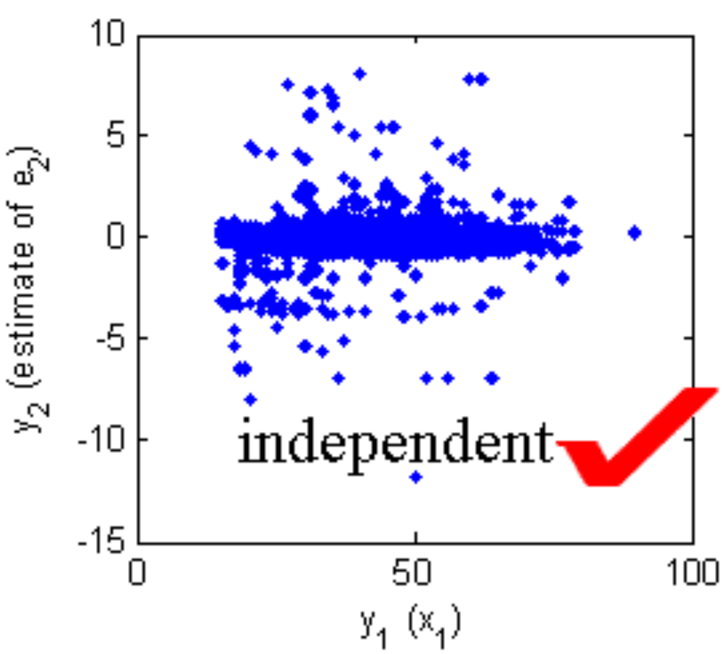
Data Set 8

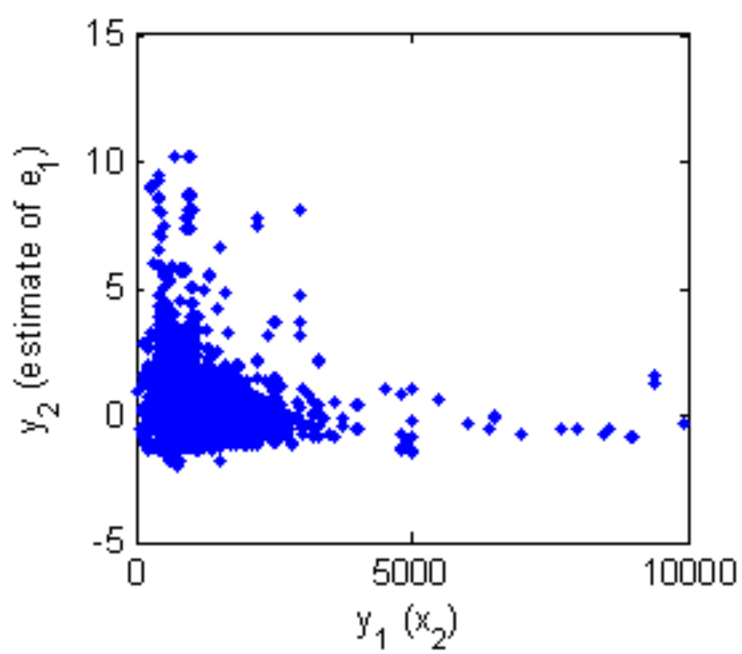


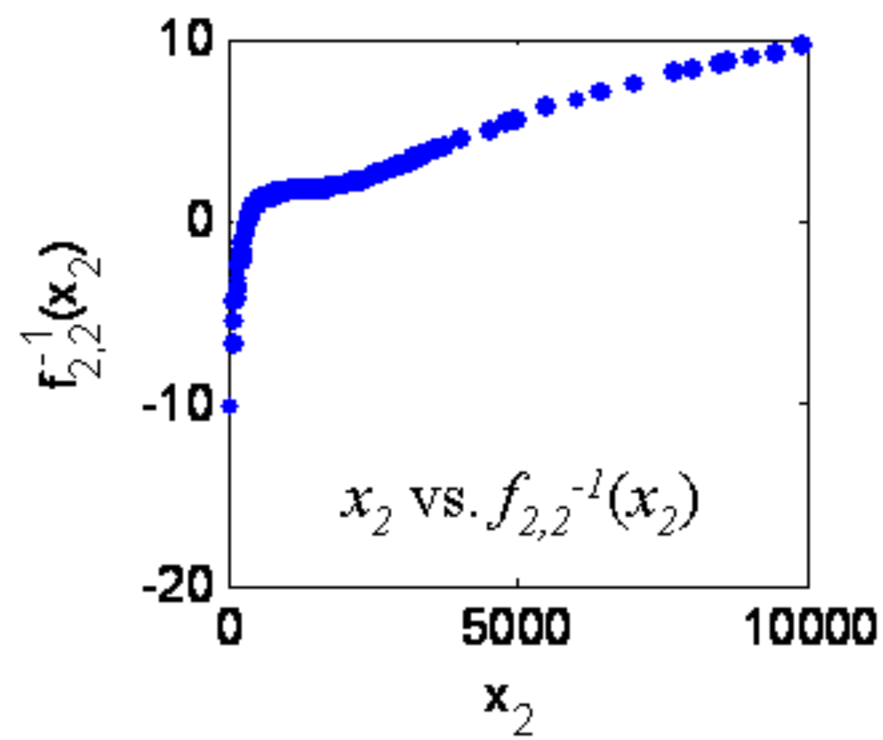


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$







Identifiability in Two-variable Case: Theoretical Results

pa_i: parents (causes) of x_i

$$X_i = f_{i,2}(f_{i,1}(pa_i) + E_i)$$

$f_{i,2}$: assumed to be continuous and invertible

$f_{i,1}$: not necessarily invertible

e_i: noise/disturbance: independent from pa_i

- Two-variable case: if $X_1 \rightarrow X_2$, then $X_2 = f_{2,2}(f_{2,1}(X_1) + E_2)$
- Is the causal direction implied by the model unique?
- By a proof of contradiction
  - Assume both $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_1$ satisfy PNL model (i.e., both directions admit independent noise)
  - One can then find all non-identifiable cases

Identifiability: A Mathematical Result

Notation
$t_1 \triangleq g_2^{-1}(x_1), \quad z_2 \triangleq f_2^{-1}(x_2),$
$h \triangleq f_1 \circ g_2, \quad h_1 \triangleq g_1 \circ f_2.$
$\eta_1(t_1) \triangleq \log p_{t_1}(t_1), \quad \eta_2(e_2) \triangleq \log p_{e_2}(e_2).$

- **Theorem 1**
  - Assume $x_2 = f_2(f_1(x_1) + e_2)$,
    $x_1 = g_2(g_1(x_2) + e_1)$,
  - Further suppose that involved densities and nonlinear functions are third-order differentiable, and that p_{e2} is unbounded,
  - For every point satisfying $\eta_2'' h' \neq 0$, we have

$$\eta_1''' - \frac{\eta_1'' h''}{h'} = \left(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2''\right) \cdot h' h'' - \frac{\eta_2'''}{\eta_2''} \cdot h' \eta_1'' + \eta_2' \cdot \left(h''' - \frac{h''^2}{h'}\right).$$

- Obtained by using the fact that the Hessian of the logarithm of the joint density of independent variables is diagonal everywhere (Lin, 1998)
- It is not obvious if this theorem holds in practice…

List of All Non-Identifiable Cases

Log-mixed-linear-and-exponential:
$\log p_v = c_1 e^{c_2 v} + c_3 v + c_4$

$(\log p_v)' \to c \; (c \neq 0)$,
as $v \to -\infty$ or as $v \to +\infty$

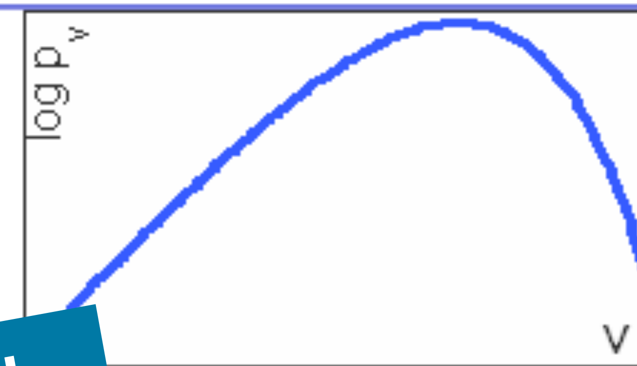
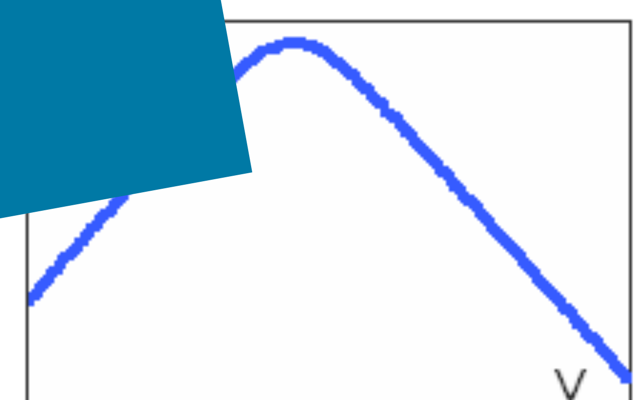


Table 1: All situations i... not identifiable.

	p_{e_2}	...	Remark
I	Gaussian		h_1 also linear
II	log-mix-lin-exp		h_1 strictly monotonic, and $h_1' \to$ ...0, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
III	log-mix-lin-exp		—
IV	log-mix-lin-exp		
V	generalized mixture of two exponentials		

$p_v \propto (c_1 e^{c_2 v} + c_3 e^{c_4 v})^{c_5}$

... $\to c_2 \; (c_2 \neq 0)$,
as $v \to +\infty$

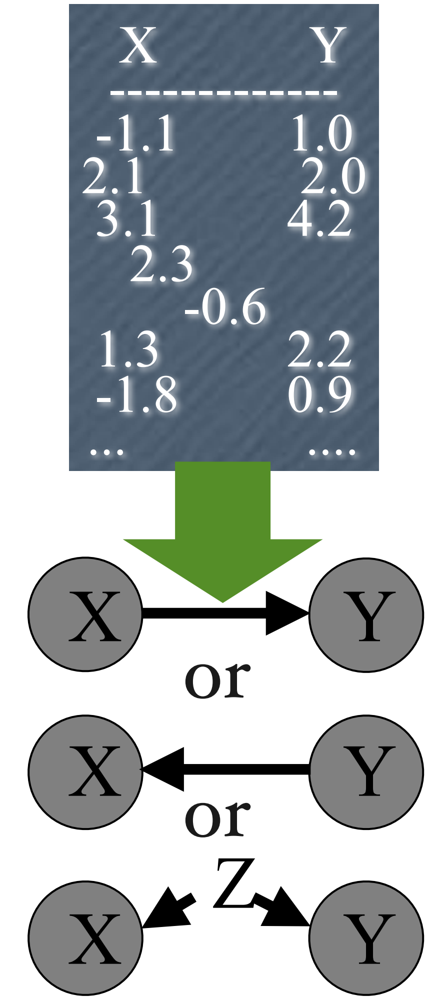


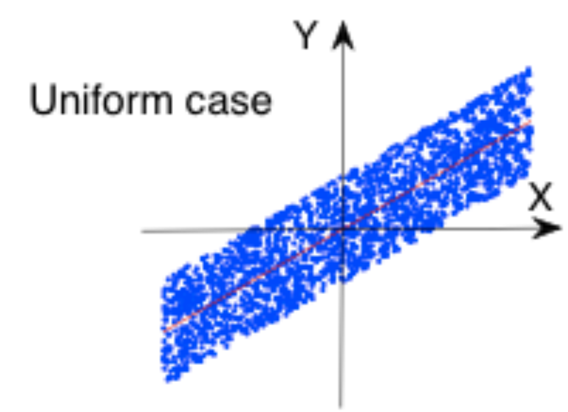
Causal direction is generally identifiable if the data were generated according to
$X_2 = f_2(f_1(X_1) + E)$.

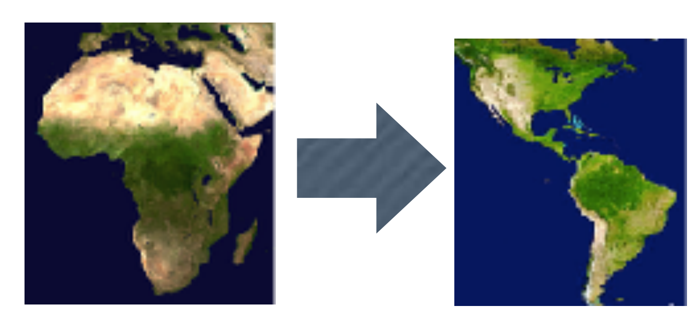
Linear models and nonlinear additive noise models are special cases.

Outline

- **Causal discovery**
  - Constraint-based approach
  - Score-based approach
  - Functional causal model-based approach
  - **Extensions**
- Causality-based learning
  - Domain adaptation (transfer learning)

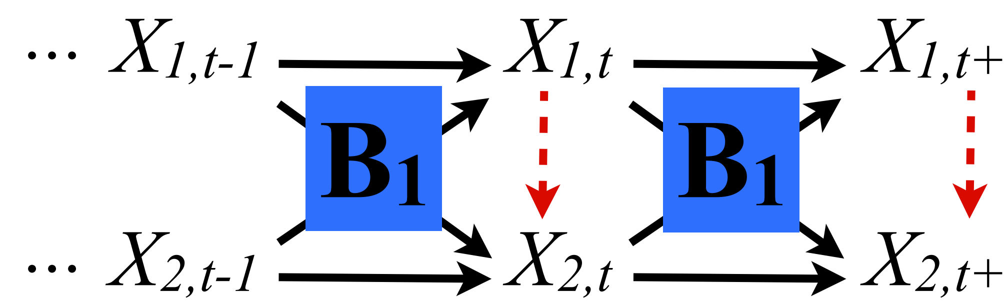




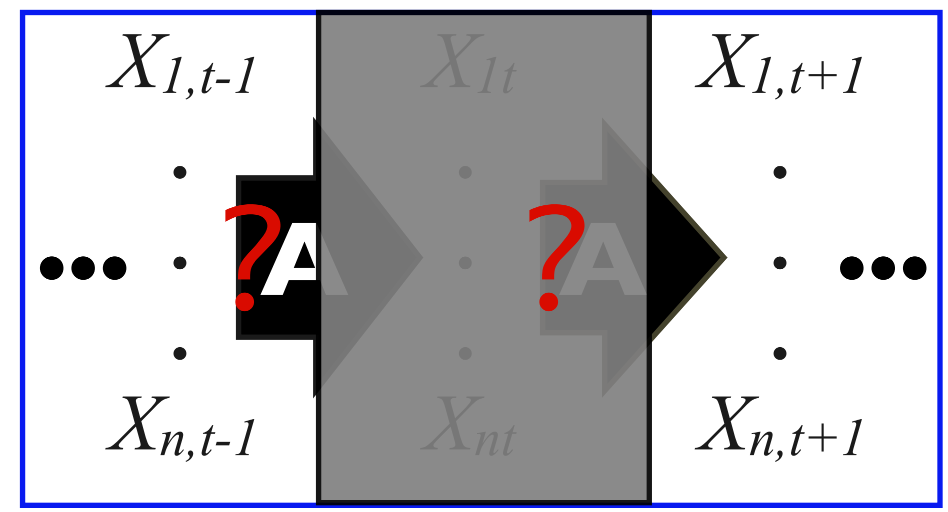


Extension I: Causality in Time Series

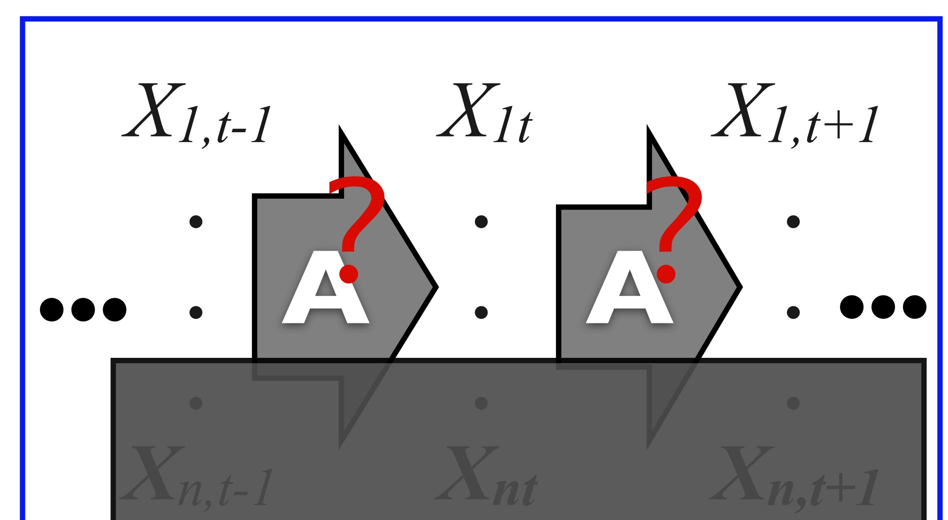
- Functional causal models in time series
  - Time-delayed causality + instantaneous relations



  - Causal discovery from subsampled or temporally aggregated data



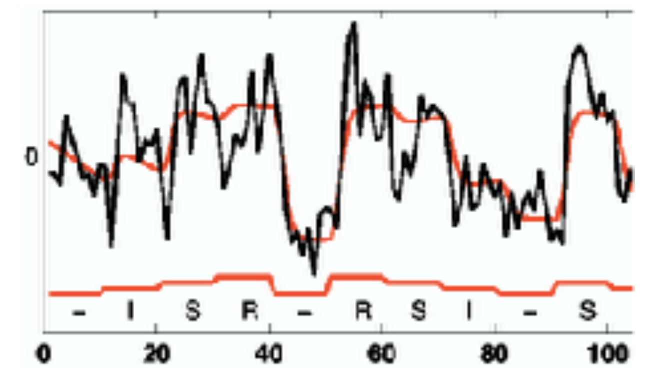
  - From partially observable time series



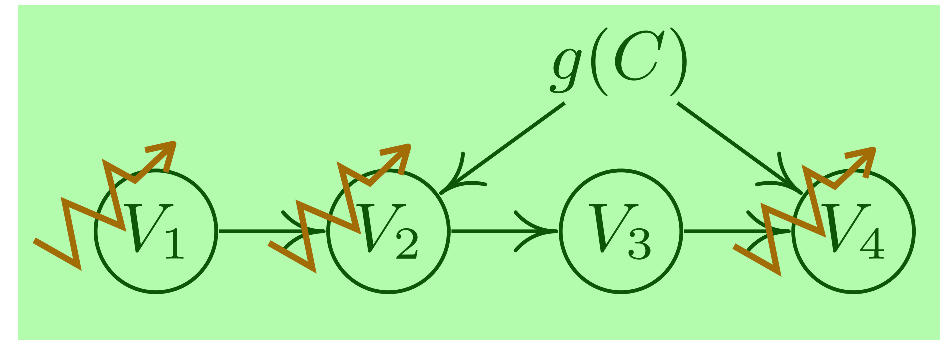
Zhang & Hyvärinen, ECML 2009;
Hyvärinen , Zhang et al., JMLR 2010;
Gong, Zhang, Schölkopf, Tao, Geigere, ICML 2015; UAI 2017;
Geiger, Zhang, Gong, Janzing, Schölkopf, ICML 2015

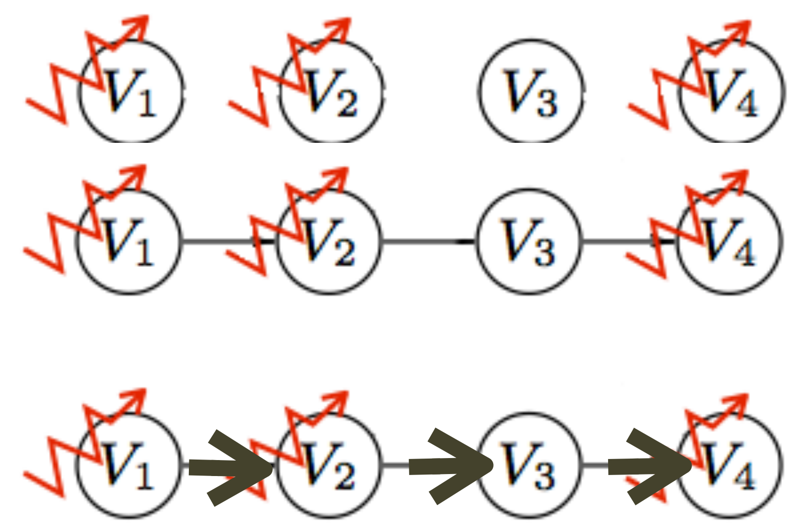
Nonstationary/Heterogeneous Data and Causality

- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data...)
  - Multiple data sets under different observational or experimental conditions
- Causal modeling and distribution shift heavily coupled
- Benefit from nonstationarity/heterogeneity!



Extension 2: Causal Discovery from Nonstationary/Heterogeneous Data





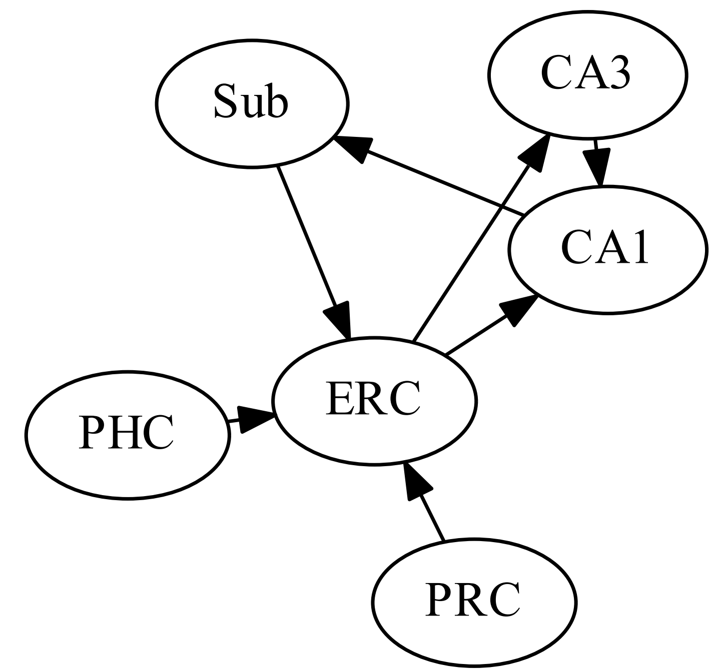
- Method to determine changing causal modules & estimate skeleton
- Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
- How do the nonstationary modules change over time / across data sets?
- Detection of nonstationary confounders

Kernel nonstationary driving force estimation

Zhang et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Zhang et al., Causal discovery in the presence of nonstatioarity/heterogeneity: Skeleton estimation and orientation determination, IJCAI 2017

On fMRI Hippocampus

- Compared our method and original constraint-based method on 10 data sets
- FP rate reduced from 62.9% to 17.1%
- Accuracy of direction determination is 85.7%

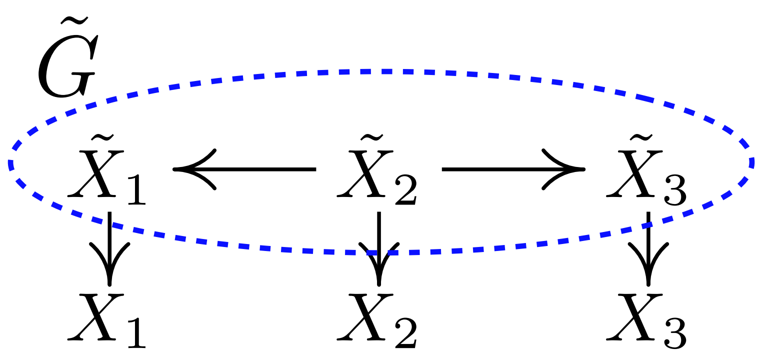


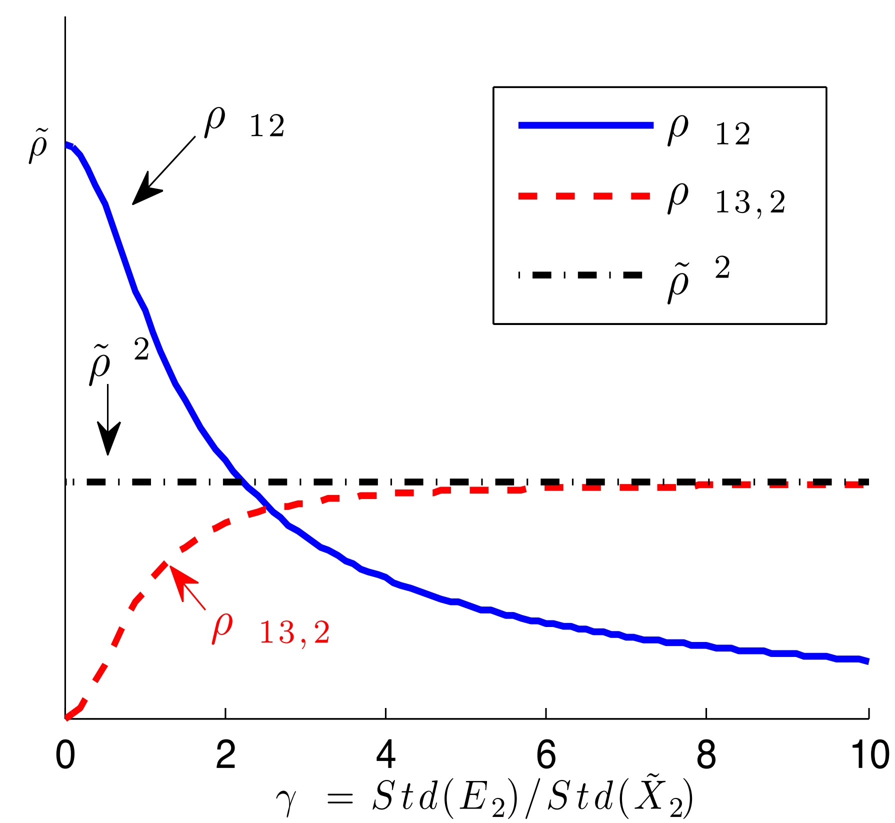
Anatomical connections

Extension 3: Causal Discovery in the Presence of Measurement Error

- To estimate $\tilde{G}$ over variables $\tilde{X}_i$ from noisy observations $X_i = \tilde{X}_i + E_i$.
- Conditional independence/dependence relations among X_i different from those among $\tilde{X}_i$
- Illustration: Correlation(X_1, X_2) & partial_correlation(X_1, X_3 | X_2)

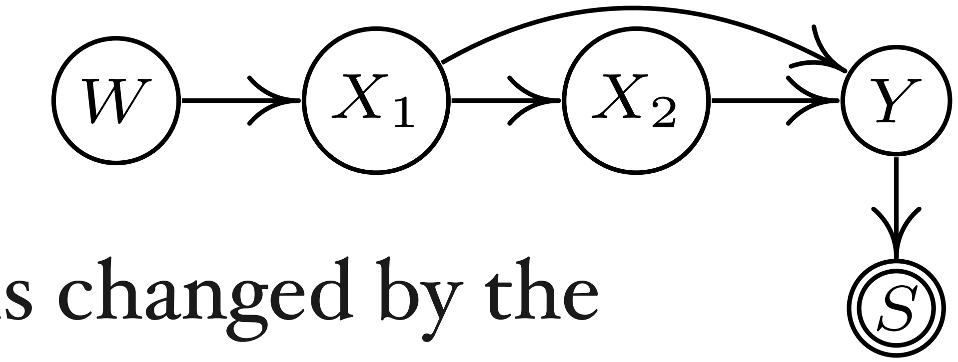
Measurement error changes causal discovery results!





Zhang, Gong, Ramsey, Batmanghelich, Spirtes, Glymour, "Causal Discovery in the Presence of Measurement Error: Identifiability Conditions," UAI 2017 Workshop on Causality

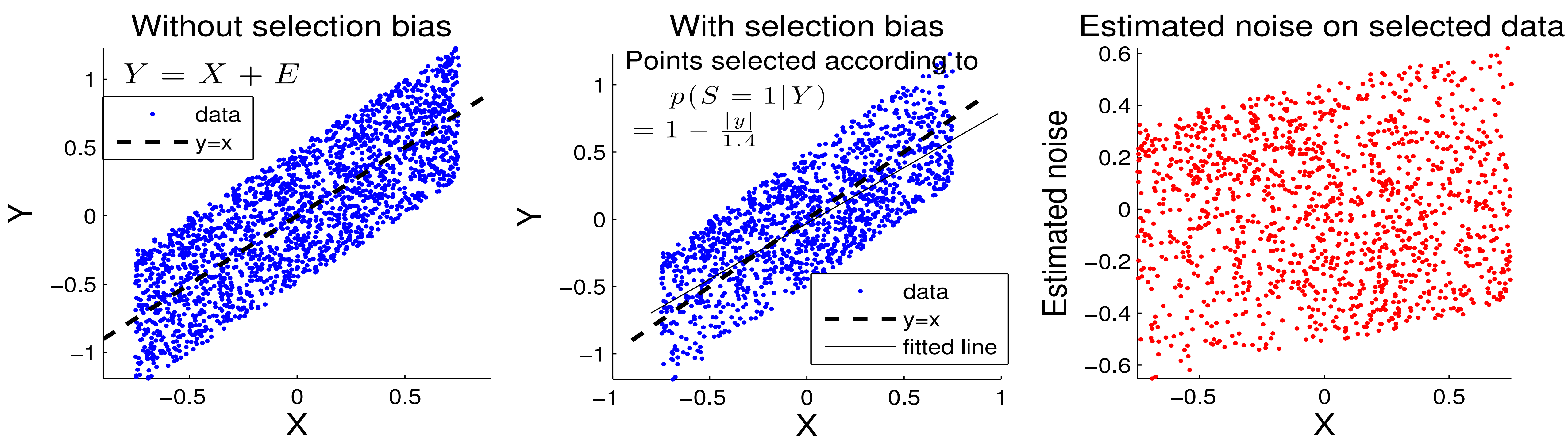
Effect of Output-Dependent Selection Bias



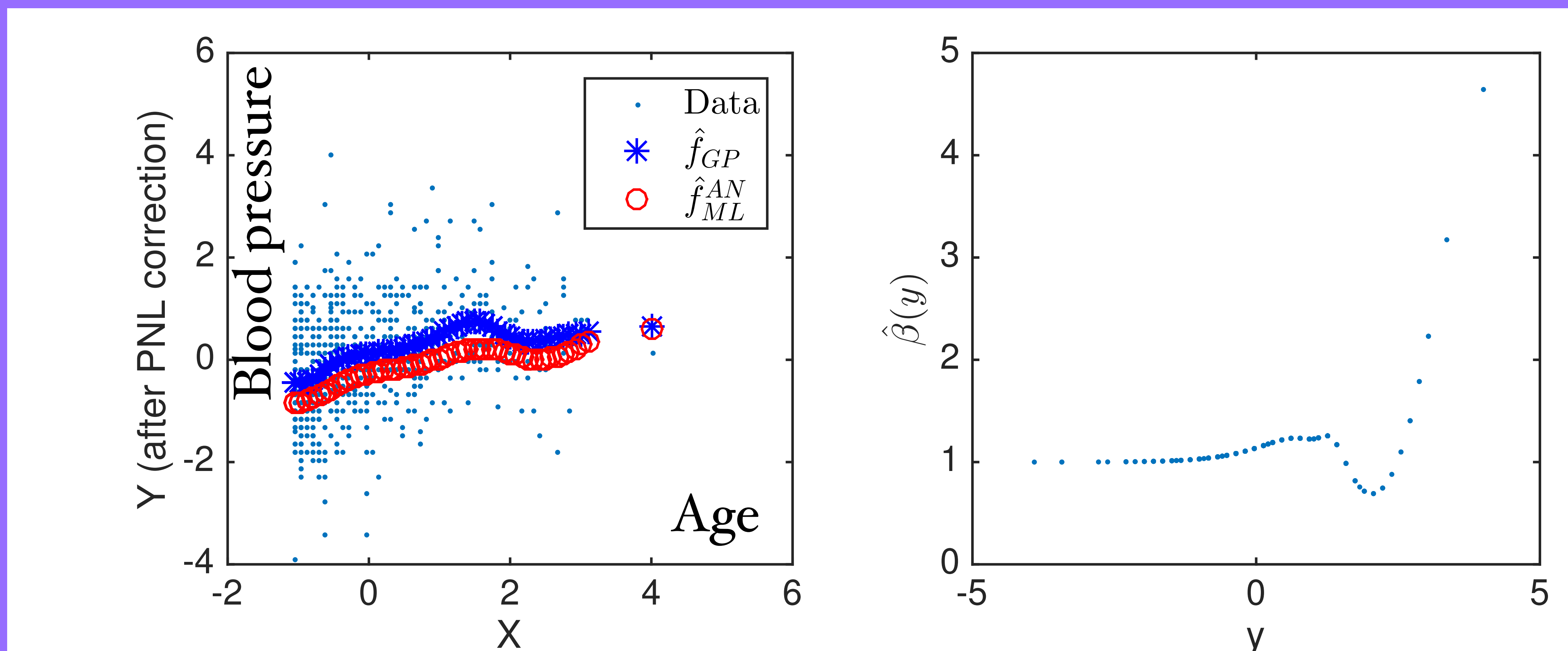
- The distribution of the observed sample is changed by the selection process

$$P_{X,Y \mid S=1} = \beta(y) P_{X,Y}$$

- An illustration: Error is not independent any more from cause



Extension 4: Causal Discovery and Inference under Output-Dependent Selection

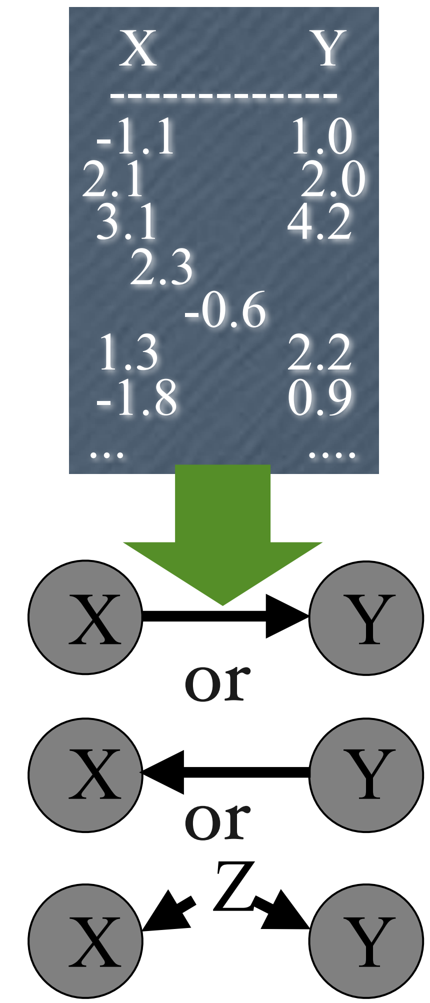


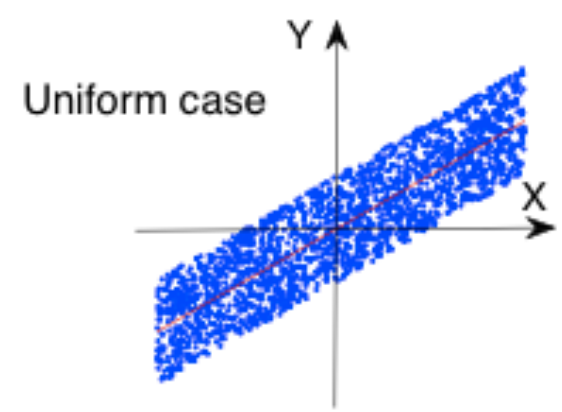
(a) Data & estimated functions. (b) $\hat{\beta}(y)$.

Zhang, Zhang, Schölkopf, Glymour, "On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection Bias," UAI 2016

Outline

- Causal thinking
- Learning causality
  - Constraint-based approach
  - Functional causal model-based approach
  - Some extensions
- **Causality-based learning**
  - **Domain adaptation (transfer learning)**





Domain Adaptation (or Transfer Learning)

$(\mathbf{x}^{tr}, \mathbf{y}^{tr})$ source (training) → target (test) $\mathbf{x}^{te}$

- Traditional supervised learning: $P^{te}_{XY} = P^{tr}_{XY}$
- Might not be the case in practice:

1. Causal relations are stable;

2. Causal relations imply higher-level independence (modularity), allowing separate parameterization

3. Causal models are usually easier to learn

Causal

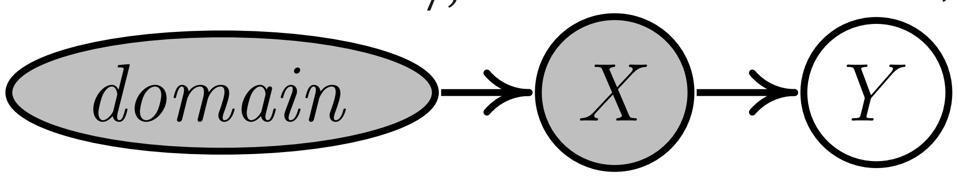
Prob. model $P^{(1)}(X,Y)$, $P^{(2)}(X,Y)$, $P^{(3)}(X,Y)$, ... $P^{(k)}(X,Y)$...

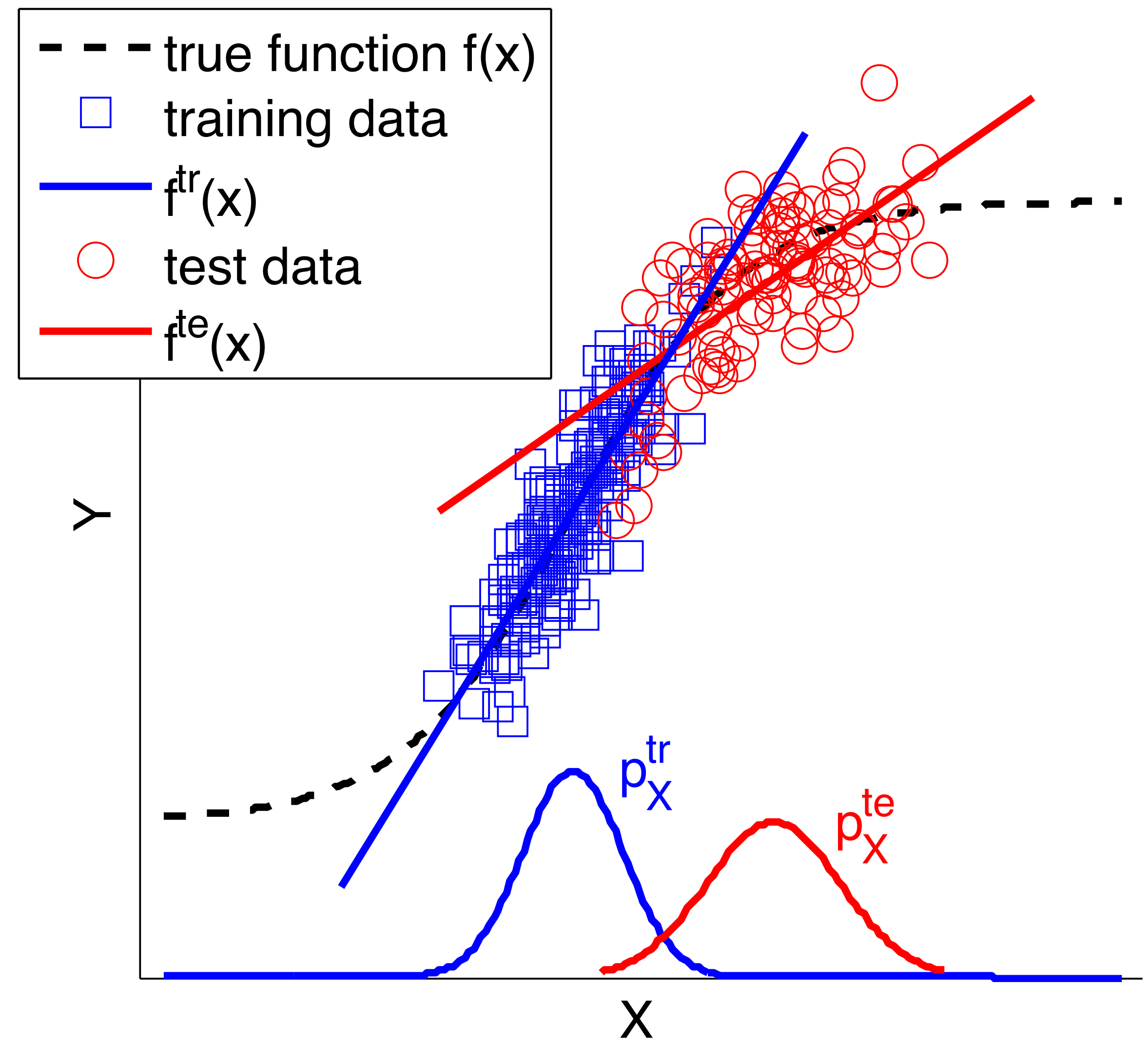
Knowing Effect may Be More Informative

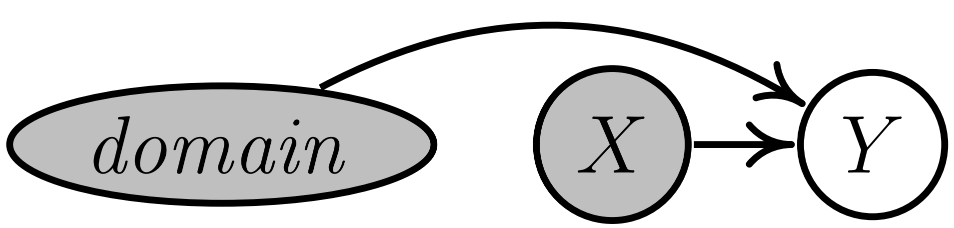
Possible Situations for Domain Adaptation: When **X**→Y

covariate shift

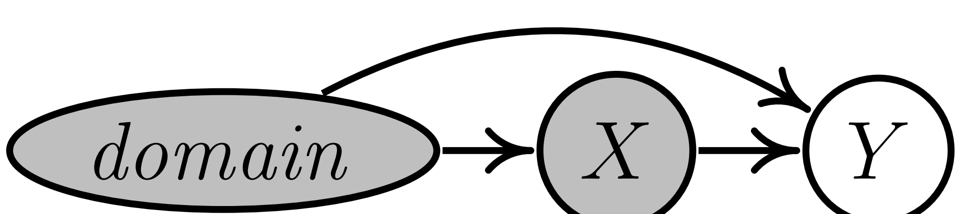
(Shimodaira00; Sugiyama etal.08; Huang etal. 07, Gretton etal.08...)





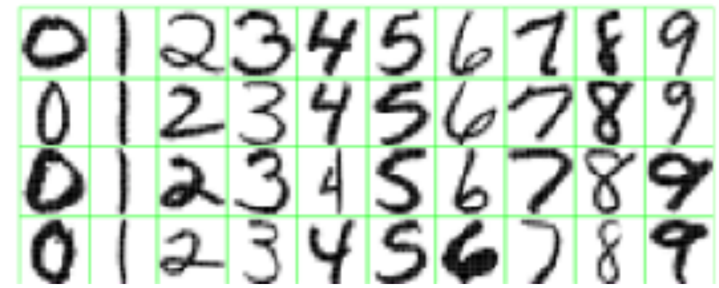


no clue as to find $P^{te}_{Y|X}$ (with one source domain)

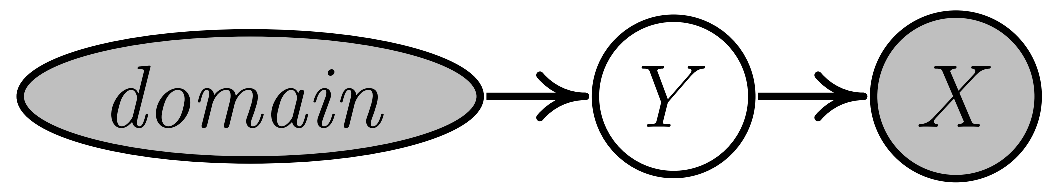


Simple Situations for Domain Adaptation: When Y→**X** (Zhang et al., 2013)

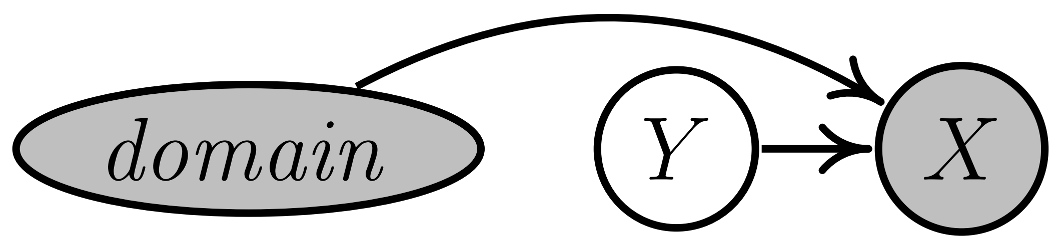
- Y is usually the cause of X
  (especially for classification)



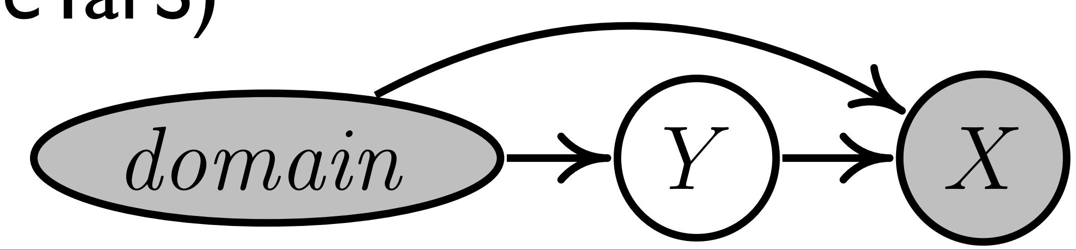
- Target shift (TarS)



- Conditional shift (ConS)



- Generalized target shift (GeTarS)



involved parameters estimated by matching P_X

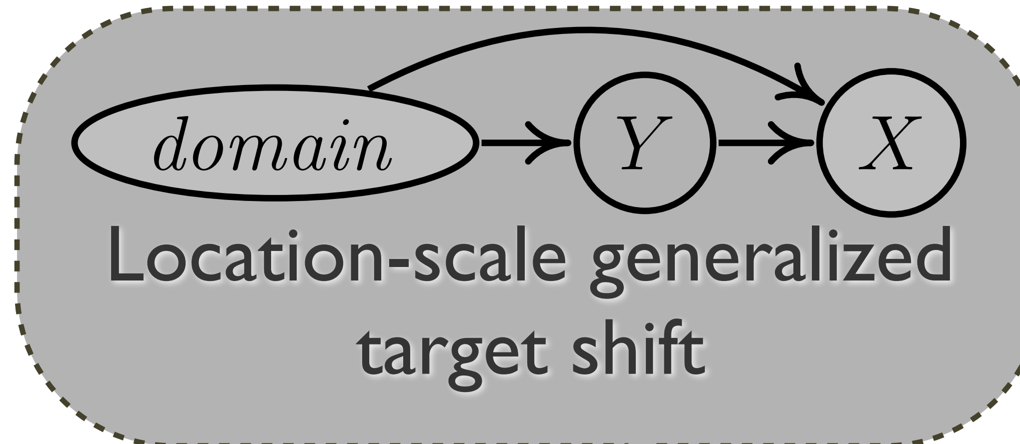
P_X^{te} helps find $P_{Y|X}^{te}$

*Zhang et al., Domain adaptation under **target and conditional Shift**, ICML 2013*
*Zhang et al., **Multi-source domain adaptation**: A causal view, AAAI 2015*
*Gong, Zhang, et al., Domain adaptation with **conditionally transferable components**, ICML 2016*

Application: Remote Sensing Image Classification

- Two domains (area 1 & area 2)
- 14 classes

Class	Number of patterns			
	Area 1		Area 2	
	TR_1	TS_1	TR_2	TS_2
Water	69	57	213	57
Hippo grass	81	81	83	18
Floodplain grasses1	83	75	199	52
Floodplain grasses2	74	91	169	46
Reeds1	80	88	219	50
Riparian	102	109	221	48
Firescar2	93	83	215	44
Island interior	77	77	166	37
Acacia woodlands	84	67	253	61
Acacia shrublands	101	89	202	46
Acacia grasslands	184	174	243	62
Short mopane	68	85	154	27
Mixed mopane	105	128	203	65
Exposed soil	41	48	81	14
Total	1242	1252	2621	627

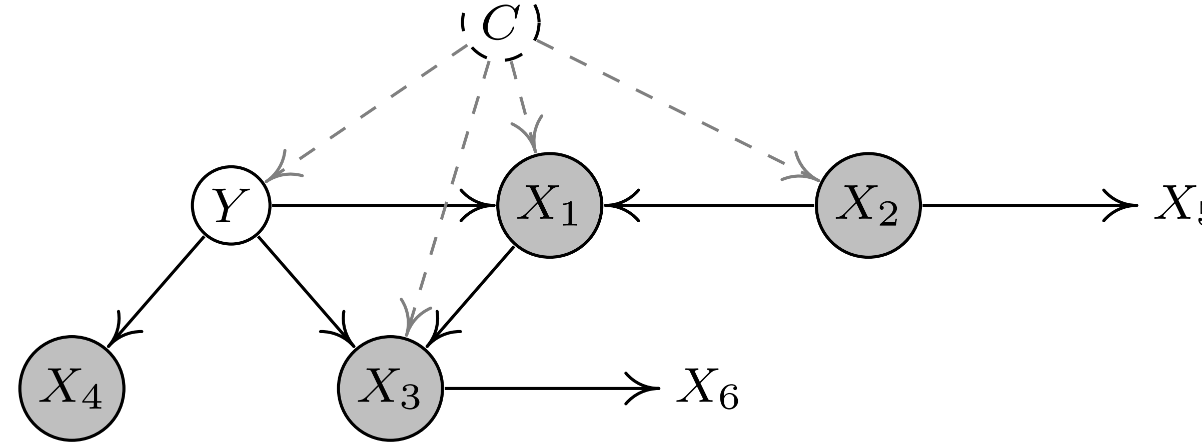


Misclassification rates by different methods

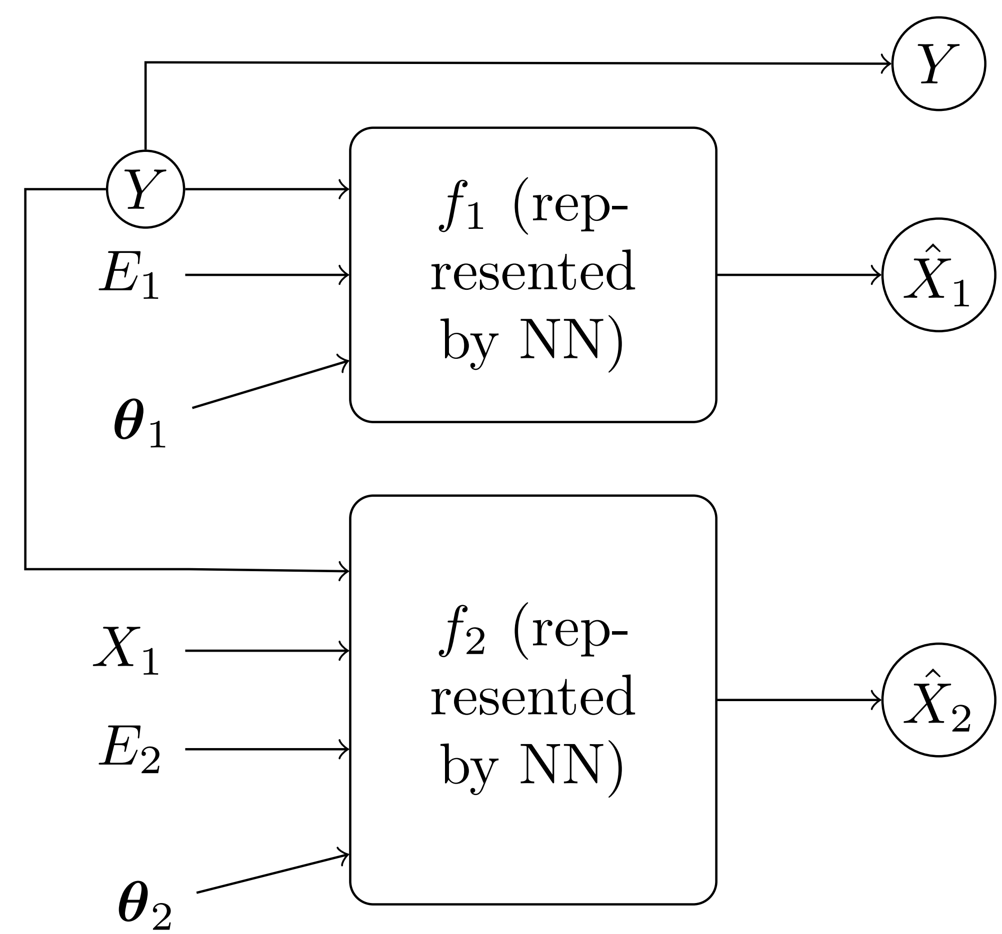
Problem	Unweight	CovS	TarS	LS-GeTarS
$TR_1 \to TS_2$	20.73%	20.73%	20.41%	**11.96%** ✓
$TR_2 \to TS_1$	26.36%	25.32%	26.28%	**13.56%** ✓

Zhang et al., Domain adaptation under **target and conditional Shift**, ICML 2013

Causal Domain Adaptation Networks



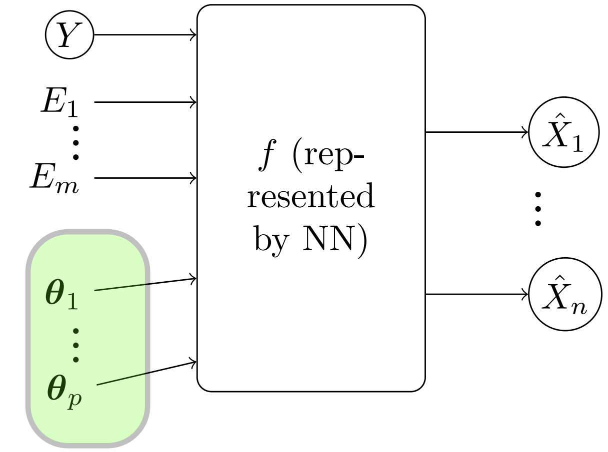
- Which variables should be considered for adaptation?



- How to model and understand the changes in causal modules and make prediction?

Gong, Zhang, ..., Batmanghelich, Causal Domain Adaptation Networks, available upon request

On MNIST Data



- One source domain:  ...
- Target domain:  ...
- Learned parameter values θ: -0.297 (source, 0^o); 0.458 (target, 45^o)
- Generate new data with

For new values of θ:
- -0.3
- -0.1
- 0.1
- 0.3
- 0.46
- 0.6
- 0.7

Summary

- Different types of "independence" helps in causal discovery:
  - Conditional independence: constraint-based approach
  - Cause $\perp\!\!\!\perp$ noise in constrained FCMs $\Rightarrow$ causal asymmetry
  - Independent changes in P(cause) and P(effect | cause)
- Machine learning/data analysis benefit from causal modeling
  - Go beyond the data!

Thanks to

- Biwei Huang, Mingming Gong, Jiji Zhang
- Aapo Hyvarinen, Bernhard Schölkopf, Clark Glymour, Peter Spirtes
- Judea Pearl, Lei Xu, Laiwan Chan, Dominik Janzing, Shohei Shimizu
- Zhikun Wang, Philipp Geiger, Jonas Peters, Joris Mooij, Patrik Hoyer