Causality II

Guest lecture for “Probabilisitic Graphical Models”

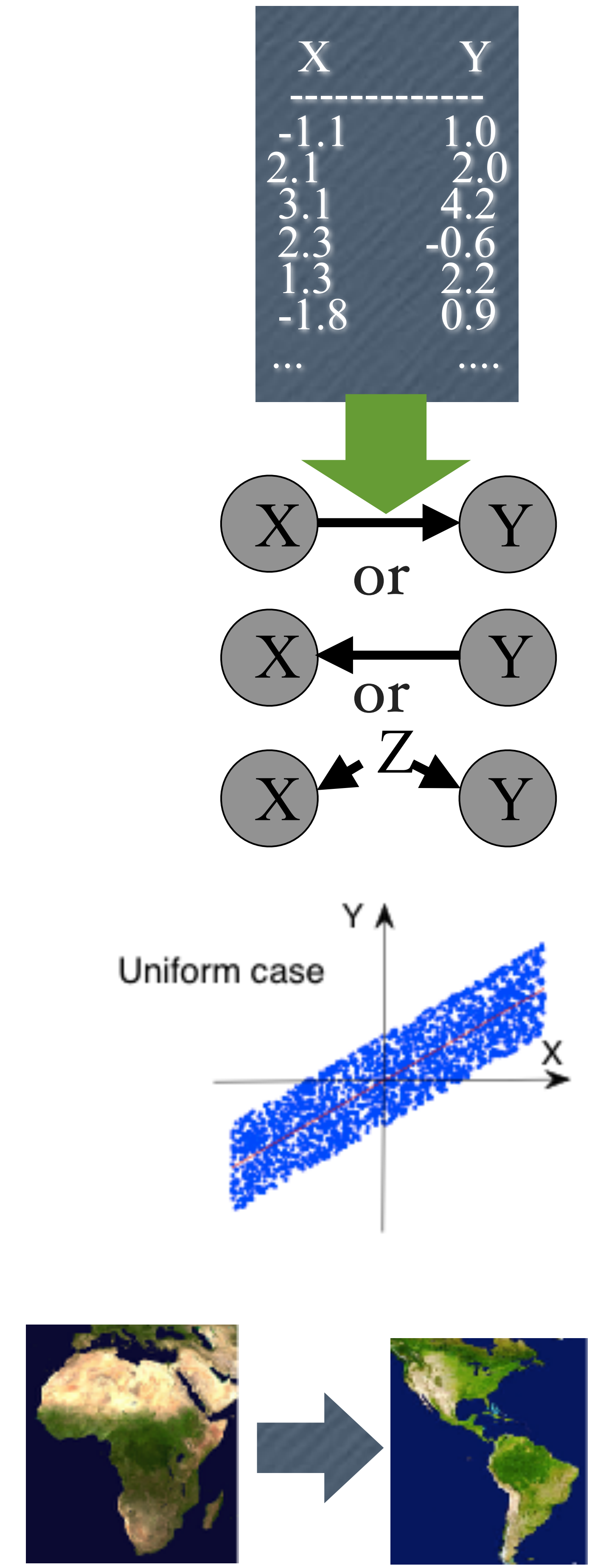
Kun Zhang

kunz1@cmu.edu

Carnegie Mellon University

Outline

- Causality? Interventions? Causal thinking
- Causal graphical models
- Identification of causal effects
- Counterfactual reasoning
- **Causal discovery**
- Implications in machine learning



Finding Causal Relations: Example 1

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 $X \rightarrow Y$?

subsistence crops—rice and wheat—are very dif-

Finding Causal Relations: Example 1

Science AAAS

March, 2014

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 $X \rightarrow Y$?

subsistence crops—rice and wheat—are very dif-

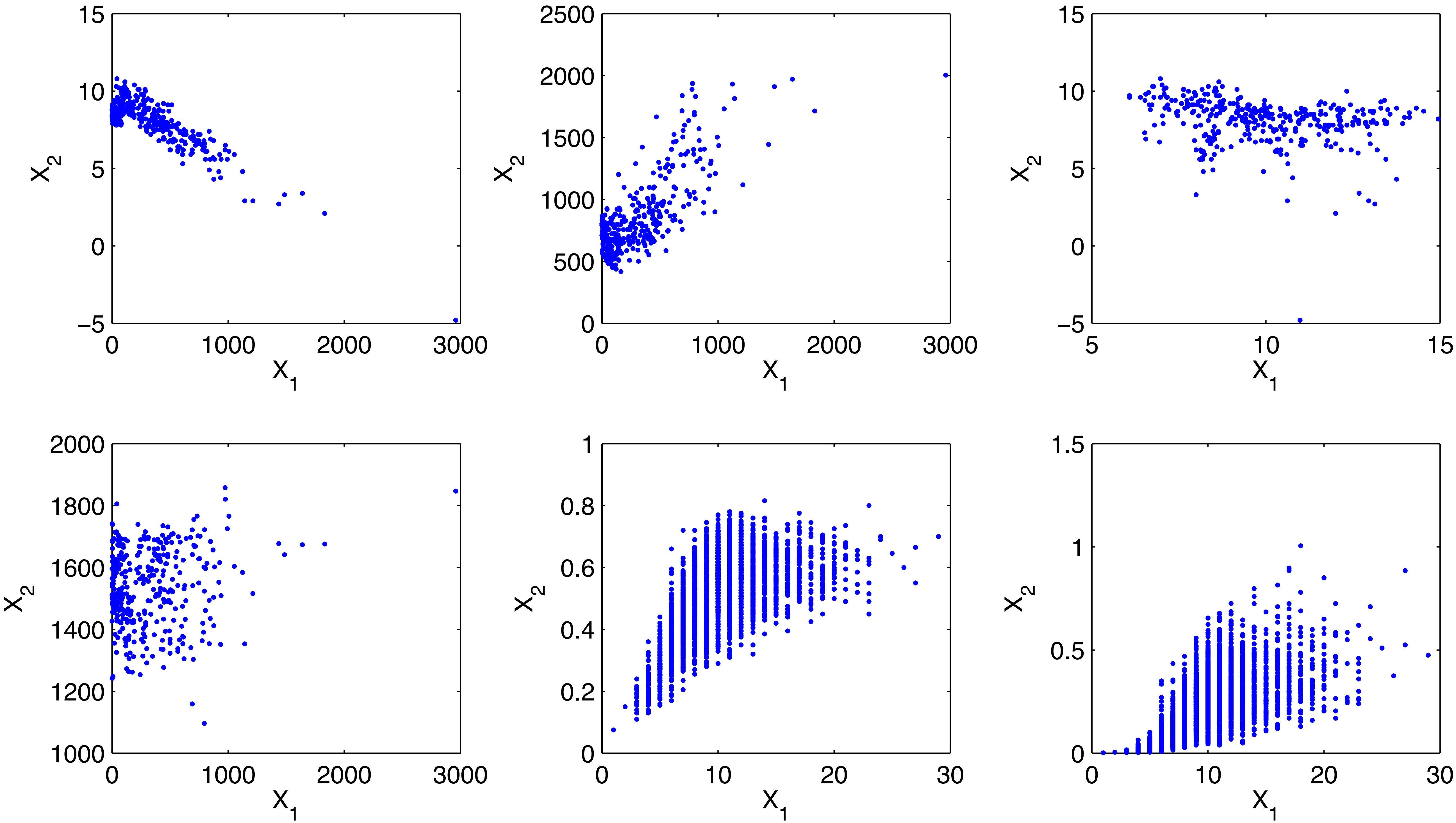
Find Causal Relations: Example 2

Thanks to collaborator Marlijn Noback

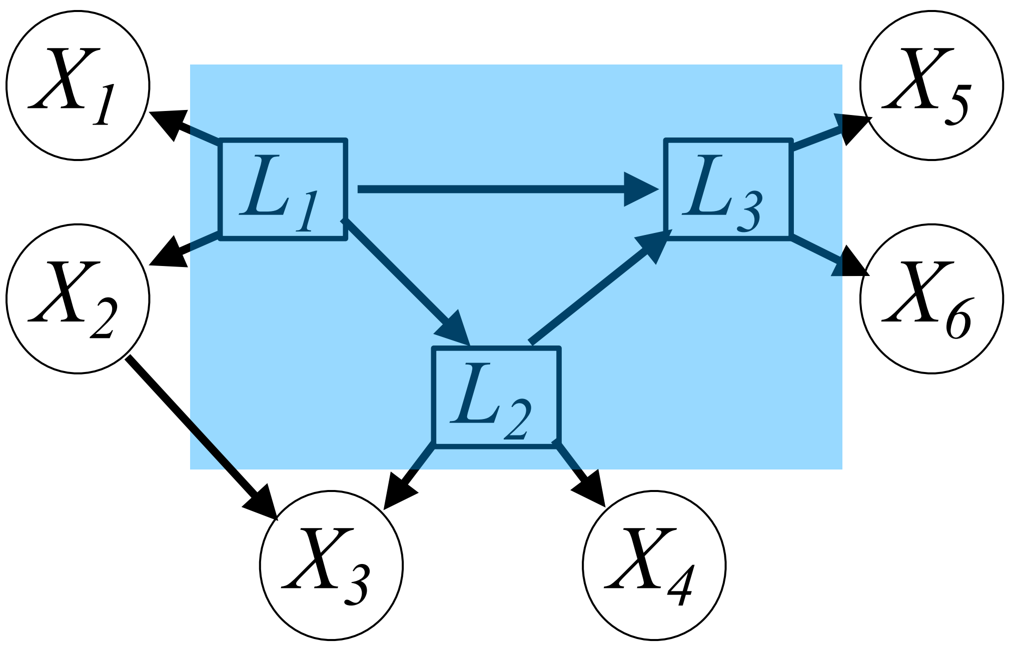
- 8 variables of 250 skeletons collected from different locations

	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 pc	Distance to	Longitude	Latitude	Tmean	Tmin	Tmax	Vpmean	Vpmin	Vpmax
3	AINU31_1	Ainu	Unknown	713.2942	2	3	4	0	1	0	1.5	2	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
4	AINU7_1	Ainu	Unknown	676.148	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
5	AINU7_2	Ainu	Unknown	675.4924	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
6	AINU_1016	Ainu	Male	684.3304	2	3	4	0	1	0	1.5	2.5	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
7	AINU_1016	Ainu	Female	686.285	2	3	4	0	1	0	1.5	4	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
8	AUSM245	Australia	Male	673.8749	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	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	135.615234	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	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
12	AUSM8175	Australia	Male	629.7138	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
14	AUSM8171	Australia	Male	643.0378	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	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	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
17	AUSM8153	Australia	Male	650.6959	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	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	135.615234	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	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
21	AUSF8172	Australia	Female	613.8324	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
28	DENM1432	Denmark	Male	663.6198	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	DENM1205	Denmark	Male	636.9831	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

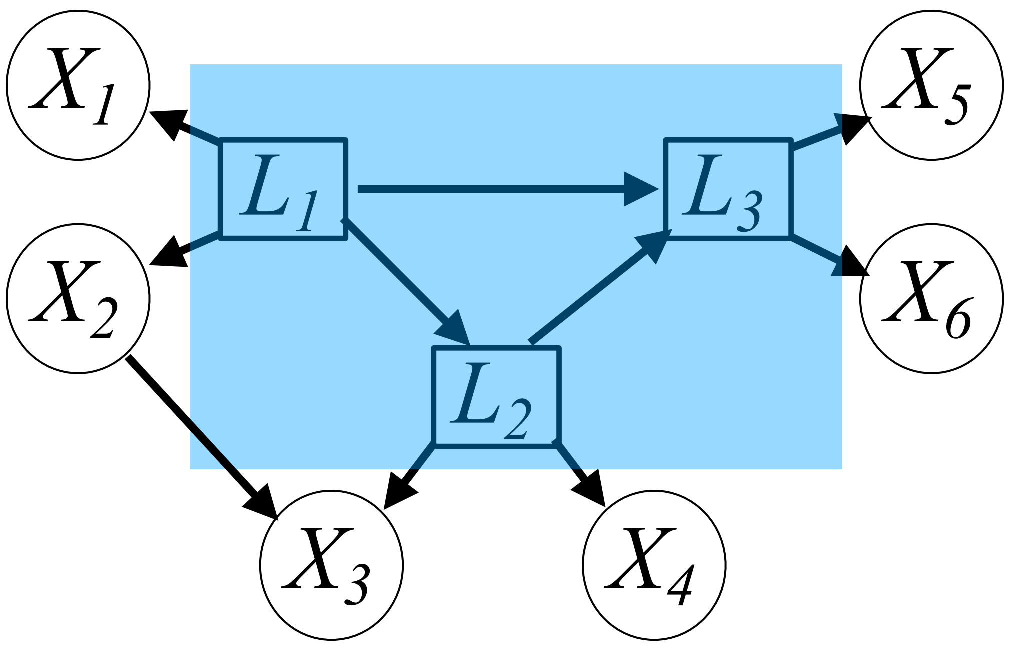
Example III: Distinguishing Cause from Effect

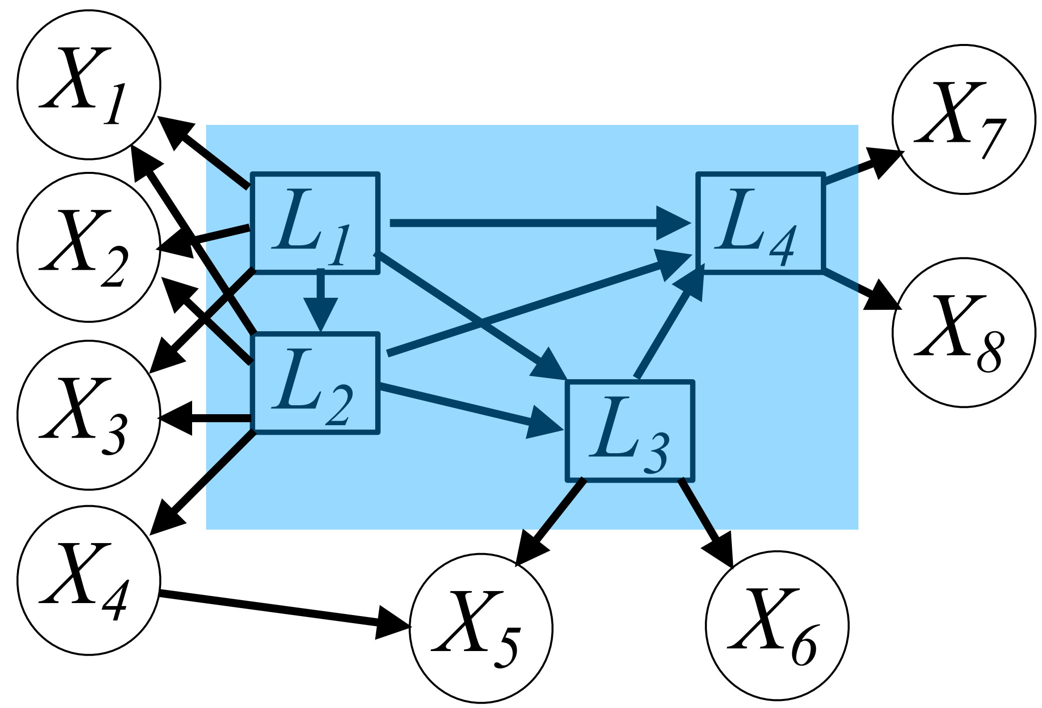


Example IV: Finding the Latent World?



Example IV: Finding the Latent World?





Causal Discovery 1:
Conditional Independence-Based Methods

- Constraint-based methods: PC and FCI

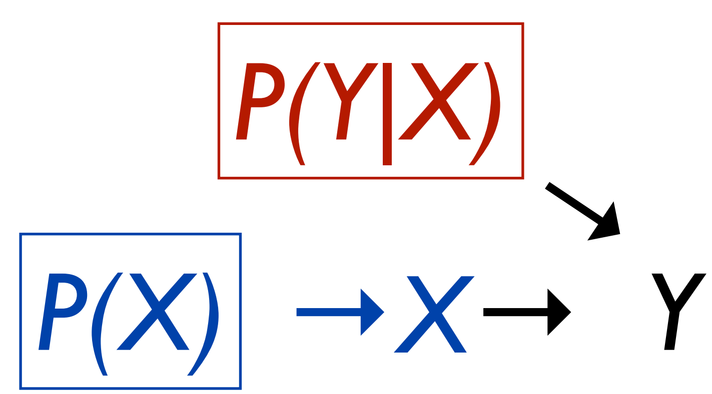
- Score-based approach: GES

What Information Helps Find Causality?

- Connection between causal structure and statistical properties of the data under *suitable assumptions* ?

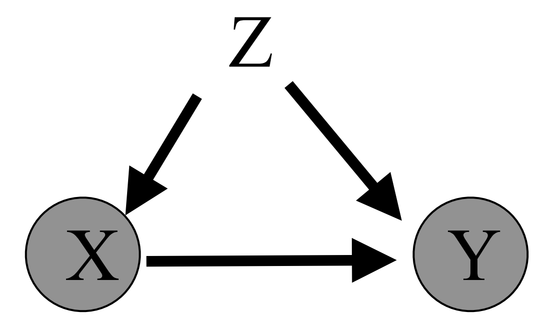
- Properties of causal systems: **modularity**

If there is no common cause of *X* and *Y*, **the generating process for cause *X*** is irrelevant to **that generates effect *Y* from *X***



Causal Sufficiency

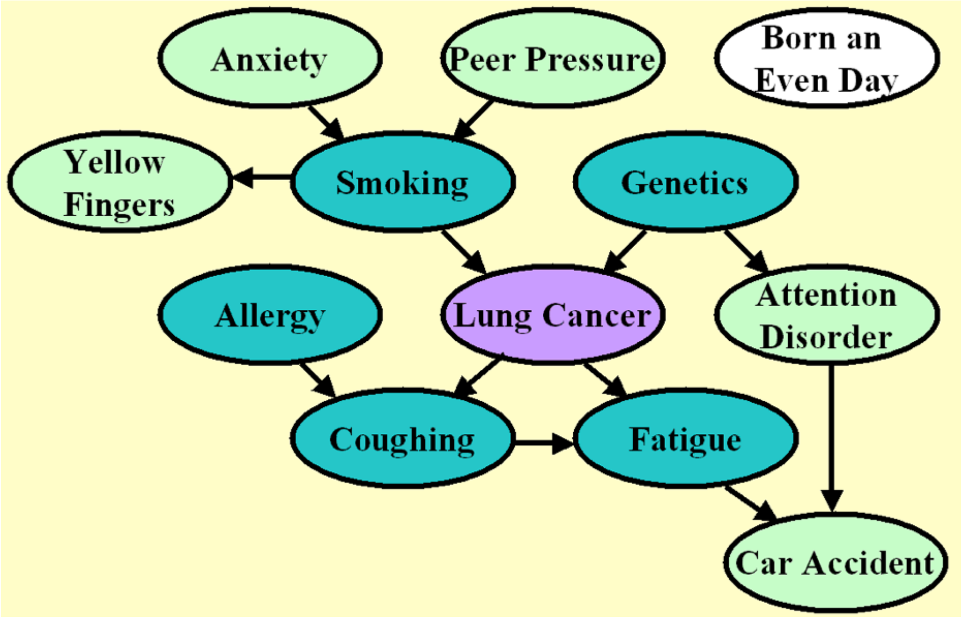
- A set of random variables $\boldsymbol{V}$ is causally sufficient if $\boldsymbol{V}$ contains every direct cause (with respect to $\boldsymbol{V}$) of any pair of variables in $\boldsymbol{V}$



- $\boldsymbol{V} = \{X, Y, Z\}$: 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)

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**?

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

causal structure
(causal graph)
$Y \rightarrow X \rightarrow Z$

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

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)
$Y \rightarrow X \rightarrow Z$

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

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)
$Y \to X \to Z$

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

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)
$Y \rightarrow X \rightarrow Z$

Statistical independence(s)

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

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

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)
$Y \rightarrow X \rightarrow Z$

Statistical independence(s)

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

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

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

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)

$Y \rightarrow X \rightarrow Z$

Statistical independence(s)

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

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

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

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)

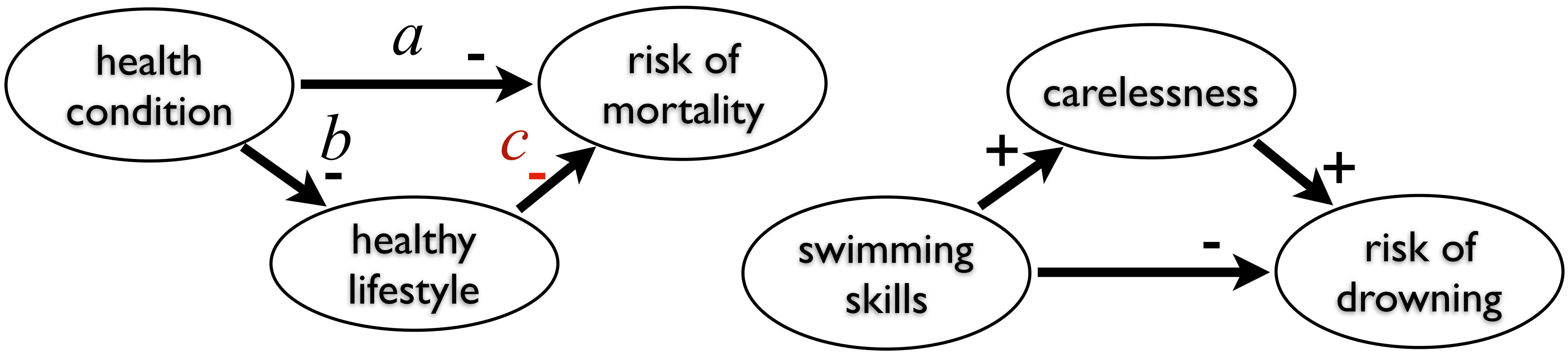
$Y \perp\!\!\!\perp Z \mid X$

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

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

Faithfulness Assumption

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

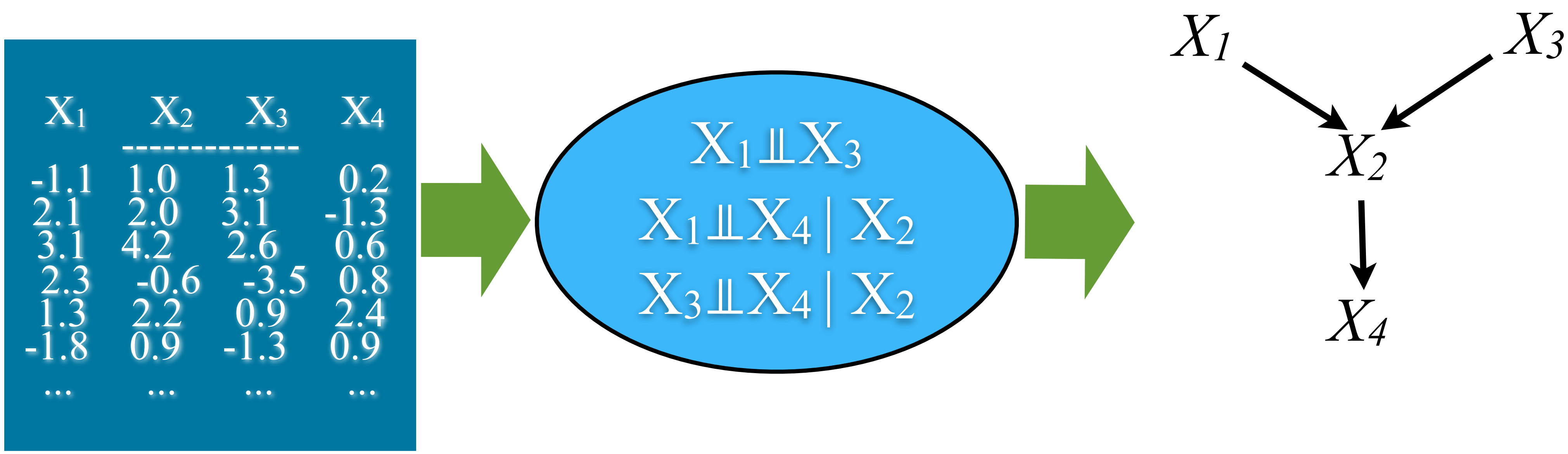


- E.g., if they are linear-Gaussian and a=-bc, then *health_condition* ⫫ *risk_mortality*, which cannot by seen from the graph!

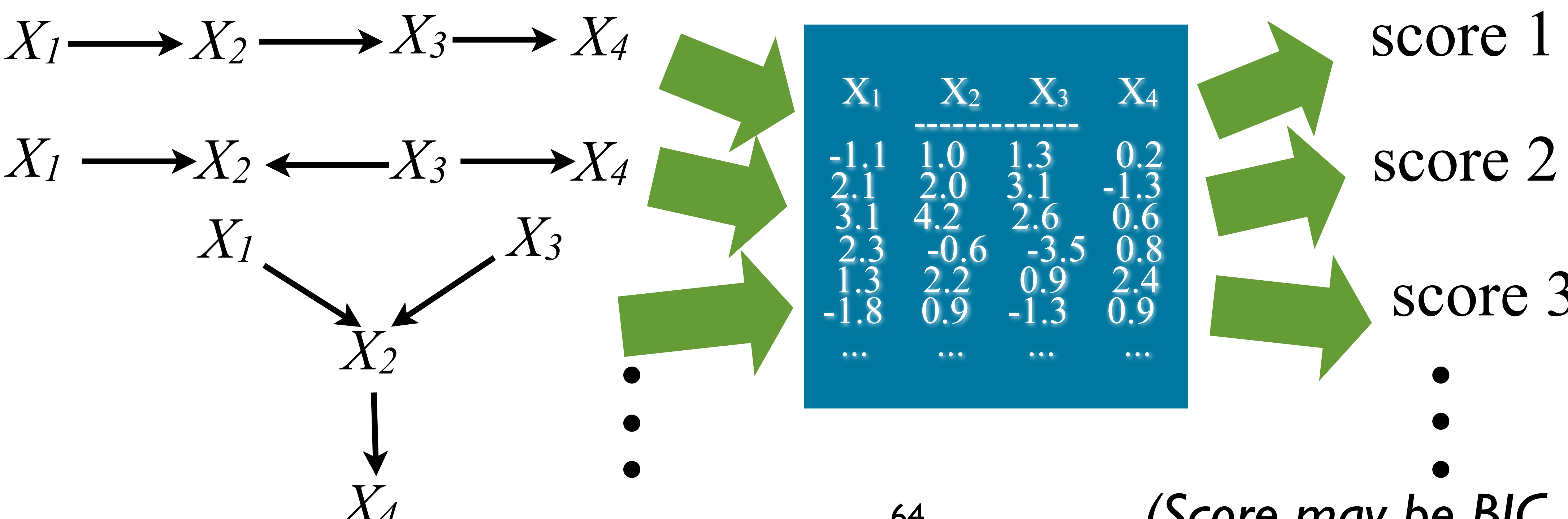
- Faithfulness assumption eliminates this possibility!

Constraint-Based vs. Score-Based

- Constraint-based methods



- Score-based methods



Which one is the best?

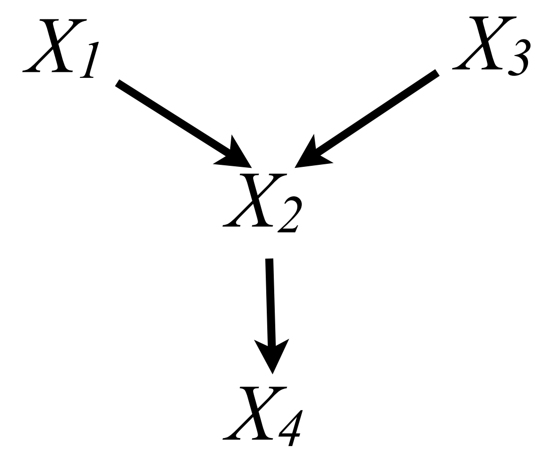
(Score may be BIC, AIC, etc.)

Constraint-Based vs. Score-Based

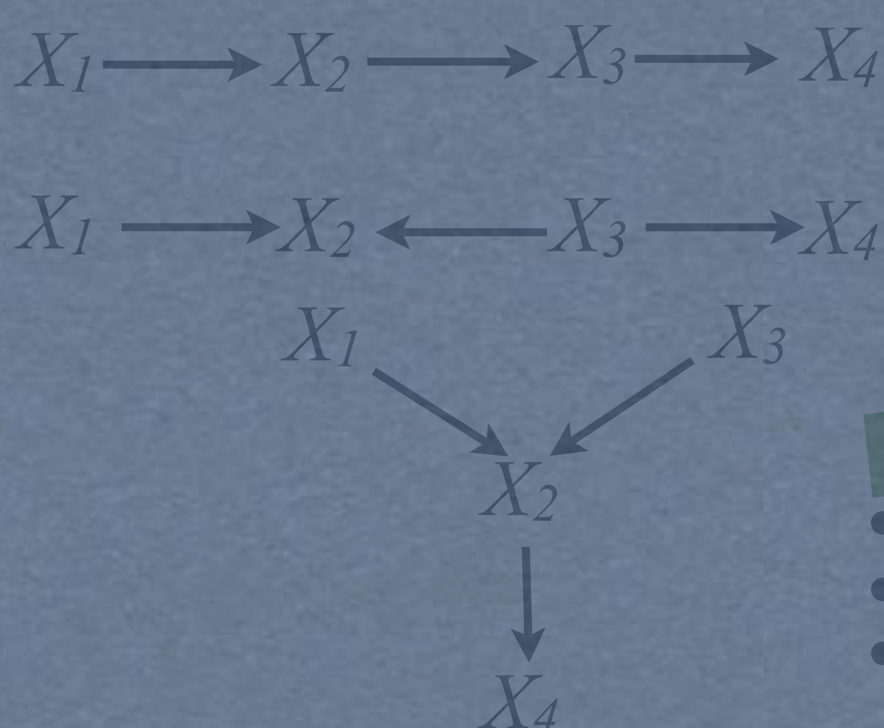
- Constraint-based methods

X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...

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



- Score-based methods



X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...

score 1
score 2
score 3
⋮

Which one is the best?

(Score may be BIC, AIC, etc.)

Discussion

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

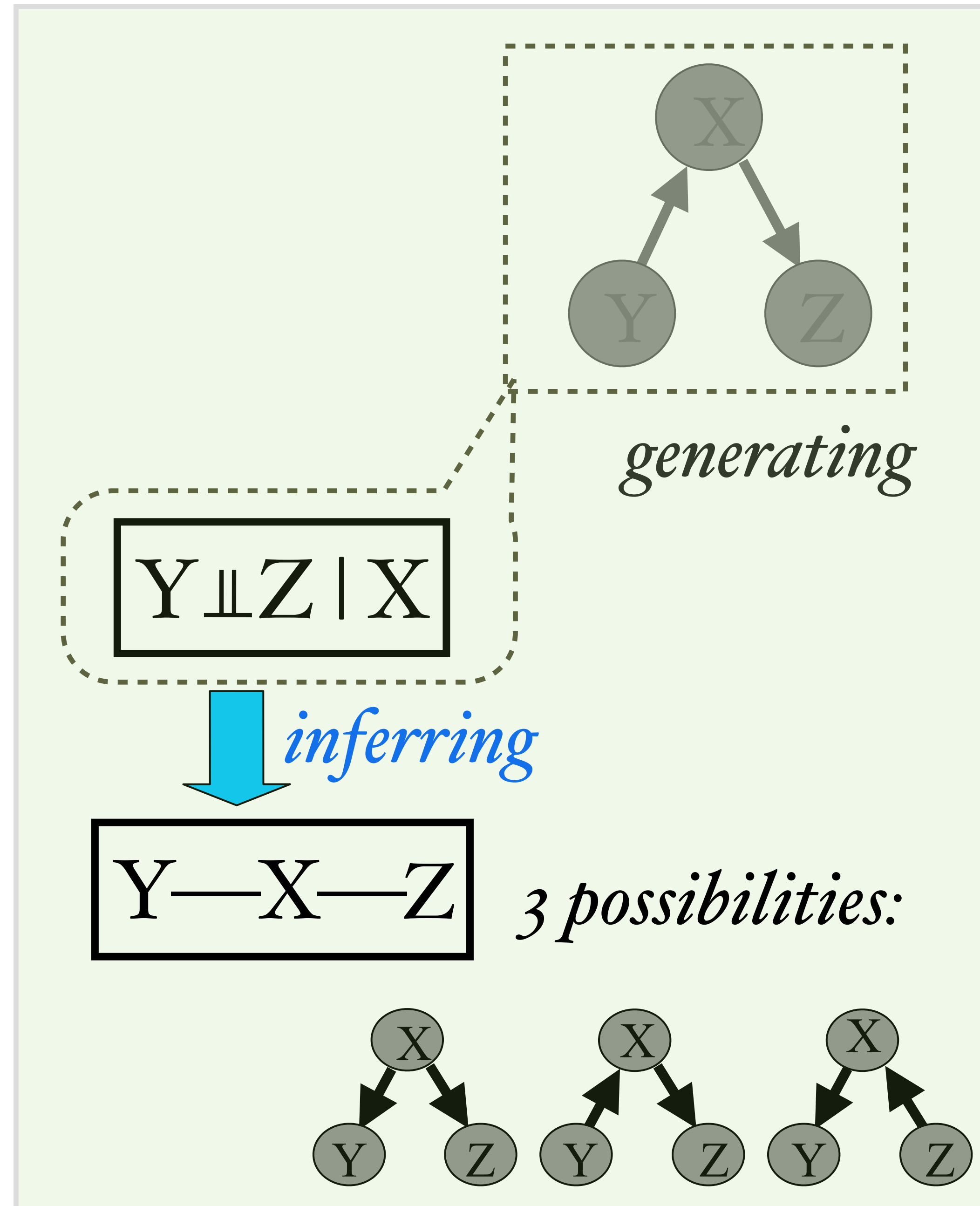
Causal Markov condition + faithfulness

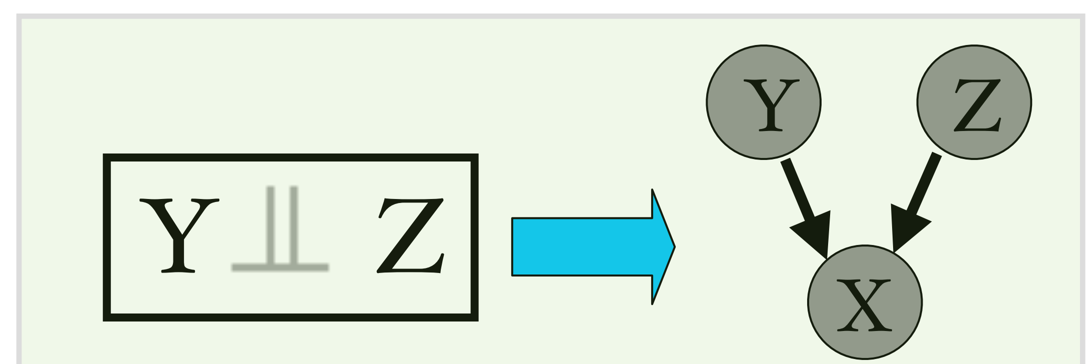
- Second, can we determine the causal direction?

How?

Constraint-Based Causal Discovery

- (Conditional) independence <u>constraints</u> ⇒ 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

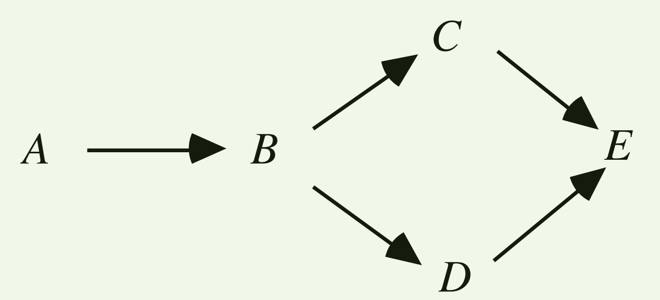




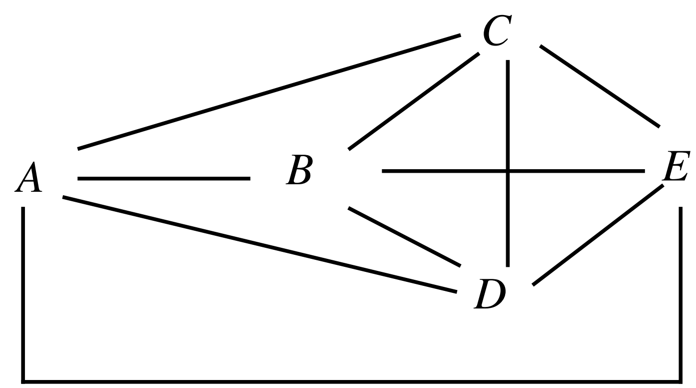
Example (From SGS Book)

Step I

Step II



True Graph

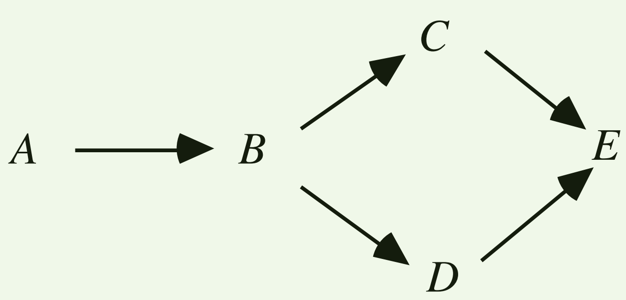


Complete Undirected Graph

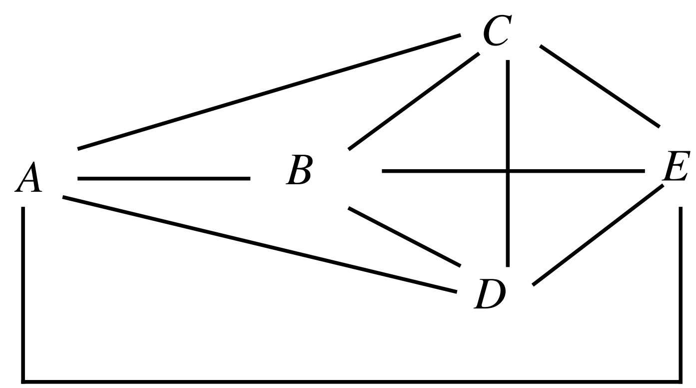
Example (From SGS Book)

Step I

Step II



True Graph



Complete Undirected Graph

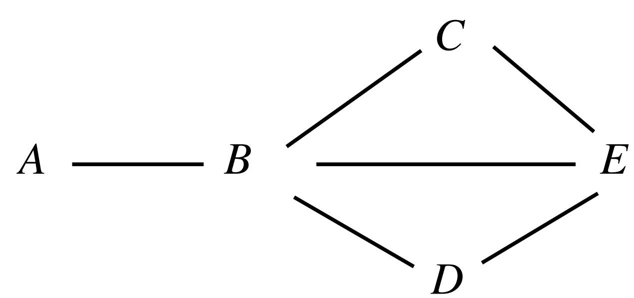
n = 0 No zero order independencies

n = 1 First order independencies

$A \perp\!\!\!\perp C \mid B$ $A \perp\!\!\!\perp D \mid B$

$A \perp\!\!\!\perp E \mid B$ $C \perp\!\!\!\perp D \mid B$

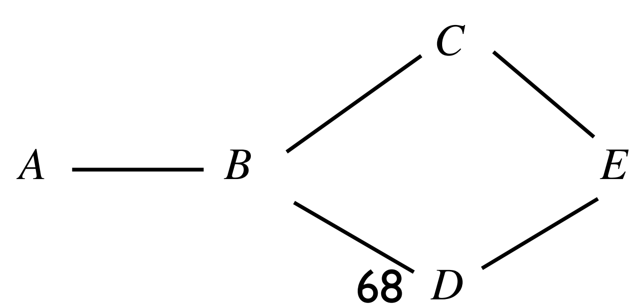
Resulting Adjacencies



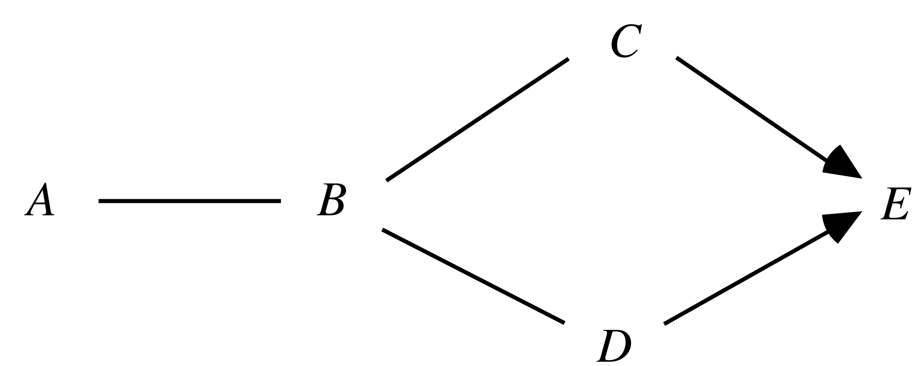
n = 2: Second order independencies

$B \perp\!\!\!\perp E \mid \{C,D\}$

Resulting Adjacencies



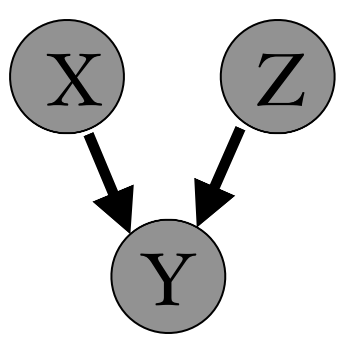
Pattern



PC Algorithm

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

Finding V-structures



Orientation propagation

A.) Form the complete undirected graph C on the vertex set $\mathbf{V}$.

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 $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given $\mathbf{S}$ delete edge X - Y from C and record $\mathbf{S}$ 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 $\mathbf{S}$ 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).

D. repeat

If A -> B, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as B -> C.

If there is a directed path from A to B, and an edge between A and B, then orient A - B as A -> B.

until no more edges can be oriented.

PC Algorithm

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

A.) Form the complete undirected graph C on the vertex set $\mathbf{V}$.

B.)

$n = 0$.

repeat

repeat

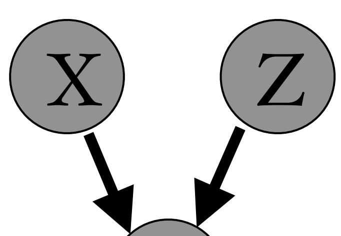
select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset **S** of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given **S** delete edge X - Y from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

Finding V-structures



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).

Orient

...here is no

...hen orient

until no more edges can be oriented.

Avoid spurious v-structures: Away from cycles:

(Independence) Equivalent Classes: Patterns

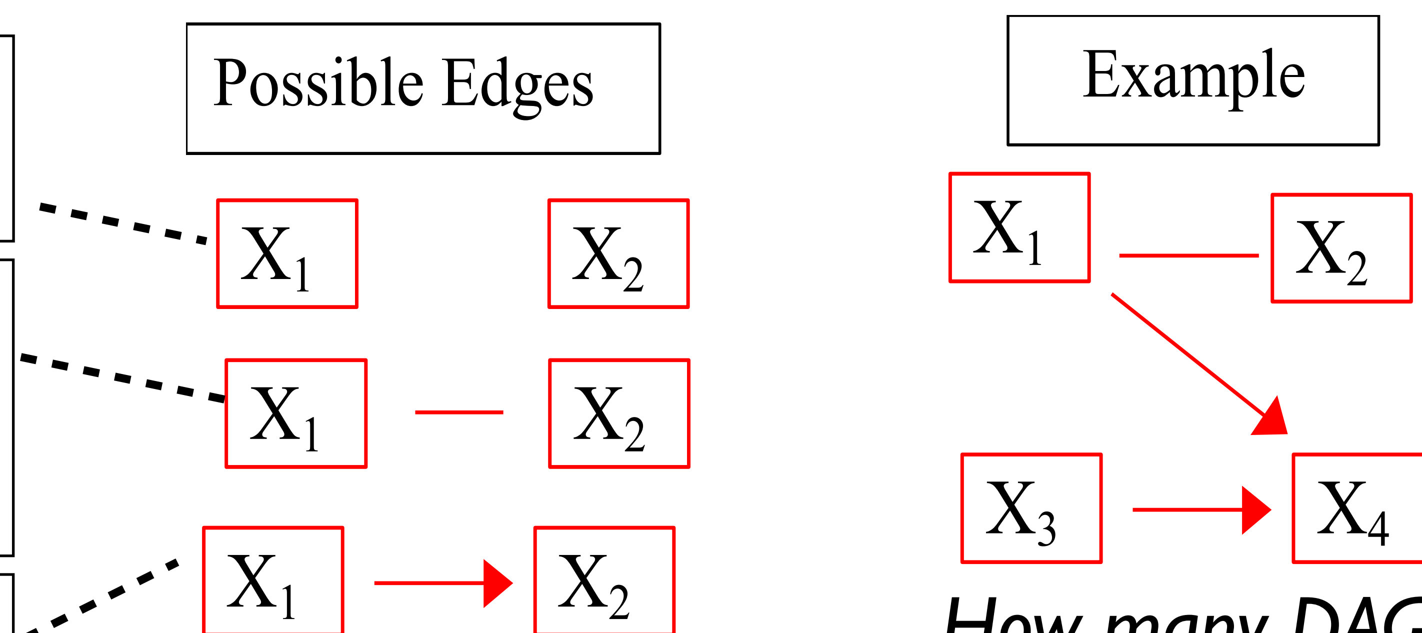
- Two DAGs are (independence) equivalent if and only if they have the same skeletons and the same v-structures (Verma & Pearl, 1991)

- Patterns or CPDAG (Completed Partially Directed Acyclic Graph): graphical representation of (conditional) independence equivalence among models with no latent common causes (i.e., causally sufficient models)

X_1 and X_2 are not adjacent in any member of the equivalent class

$X_1 \rightarrow X_2$ in some members of the equivalent class, and $X_1 \leftarrow X_2$ in some others

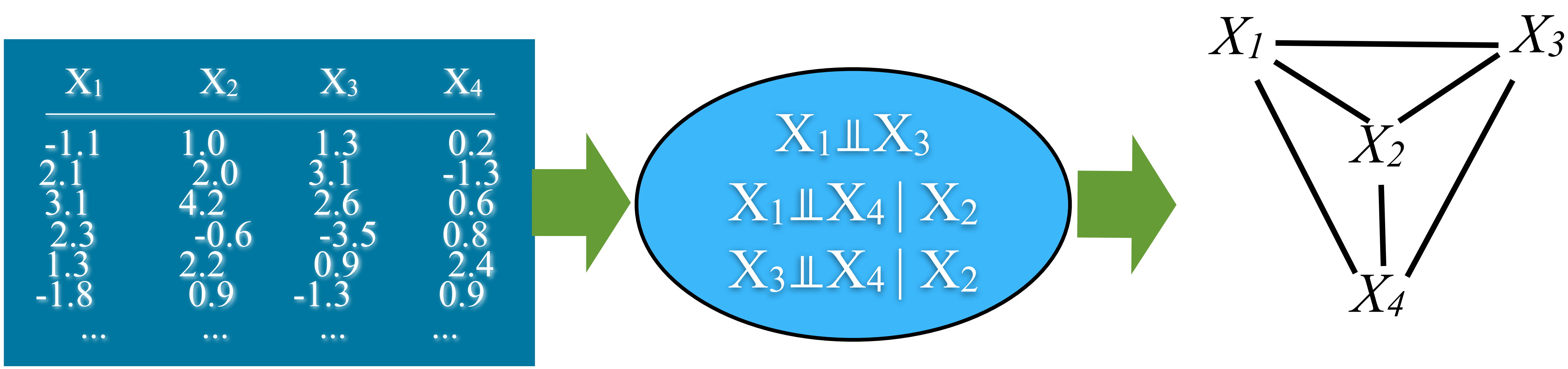
$X_1 \rightarrow X_2$ in every member of the equivalent class



How many DAGs in this class?

The PC Algorithm: Big Picture

- Make use of conditional independence relations

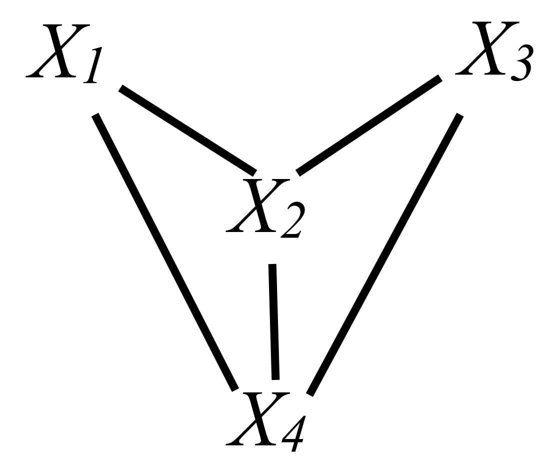


The PC Algorithm: Big Picture

- Make use of conditional independence relations

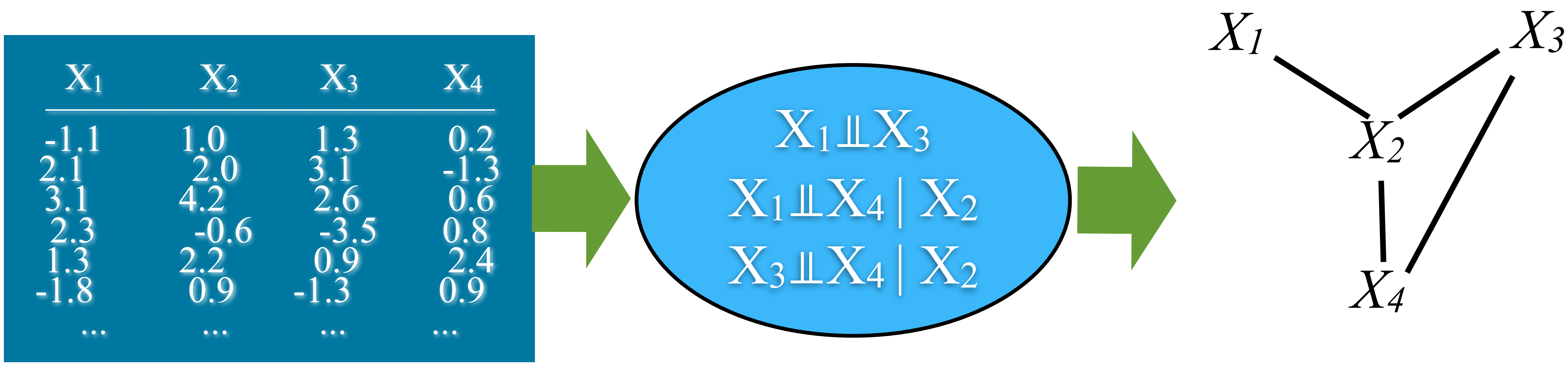
X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...

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



The PC Algorithm: Big Picture

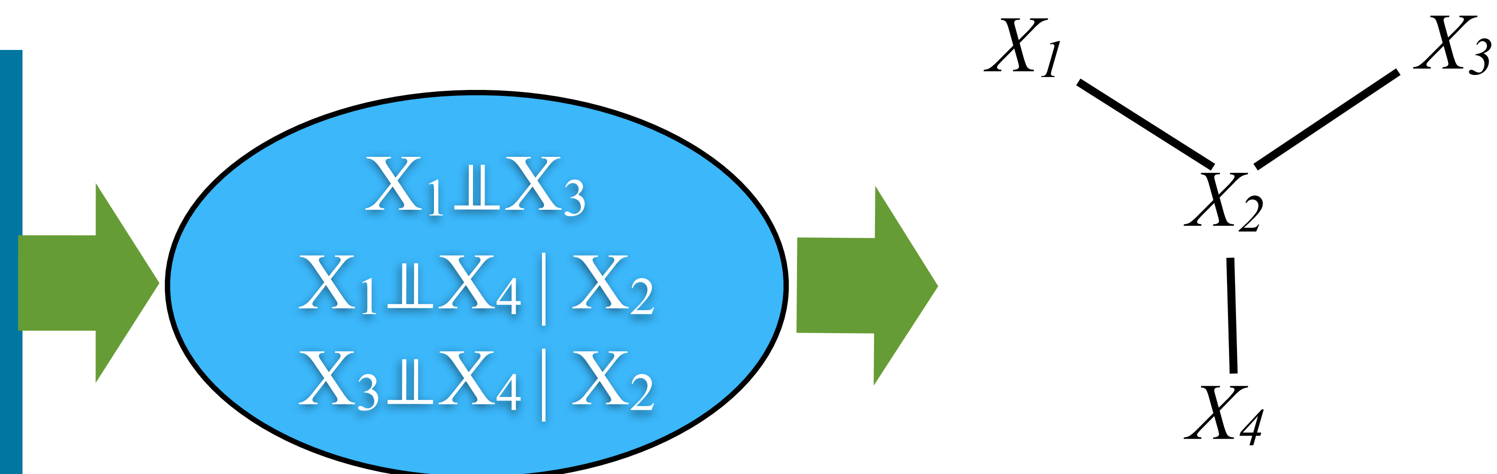
- Make use of conditional independence relations



The PC Algorithm: Big Picture

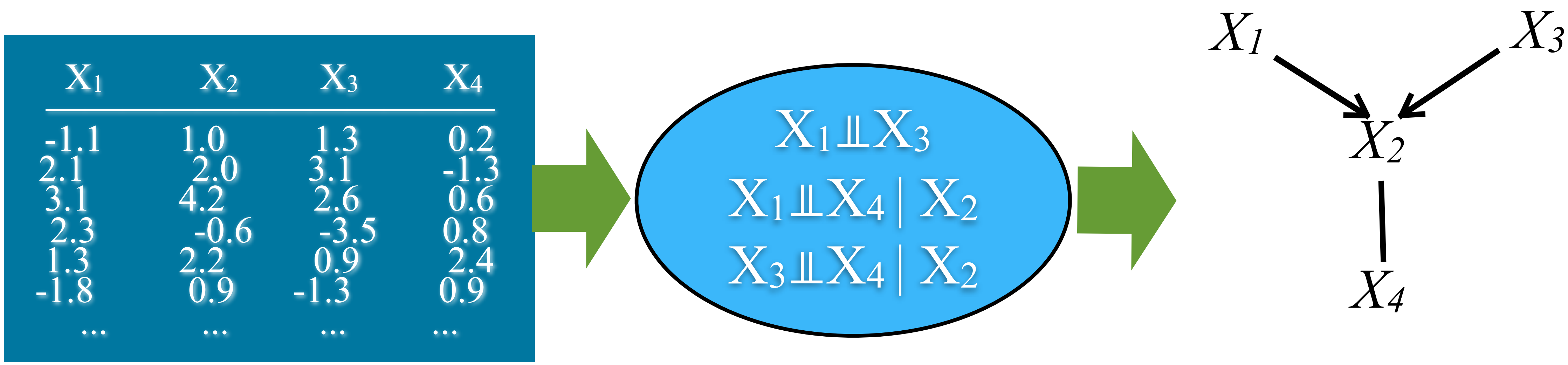
- Make use of conditional independence relations

X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...



The PC Algorithm: Big Picture

- Make use of conditional independence relations



The PC Algorithm: Big Picture

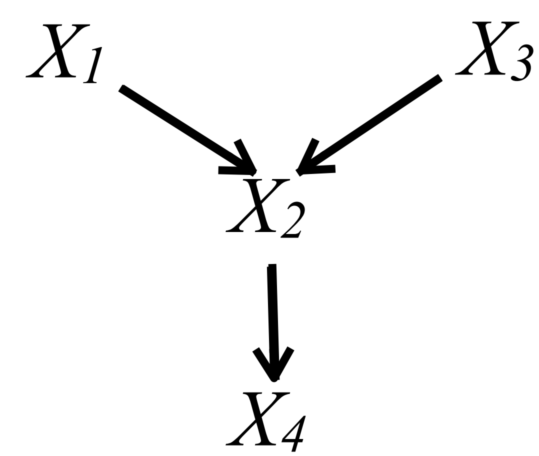
- Make use of conditional independence relations

X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...

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

$X_1 \perp\!\!\!\perp X_4 \mid X_2$

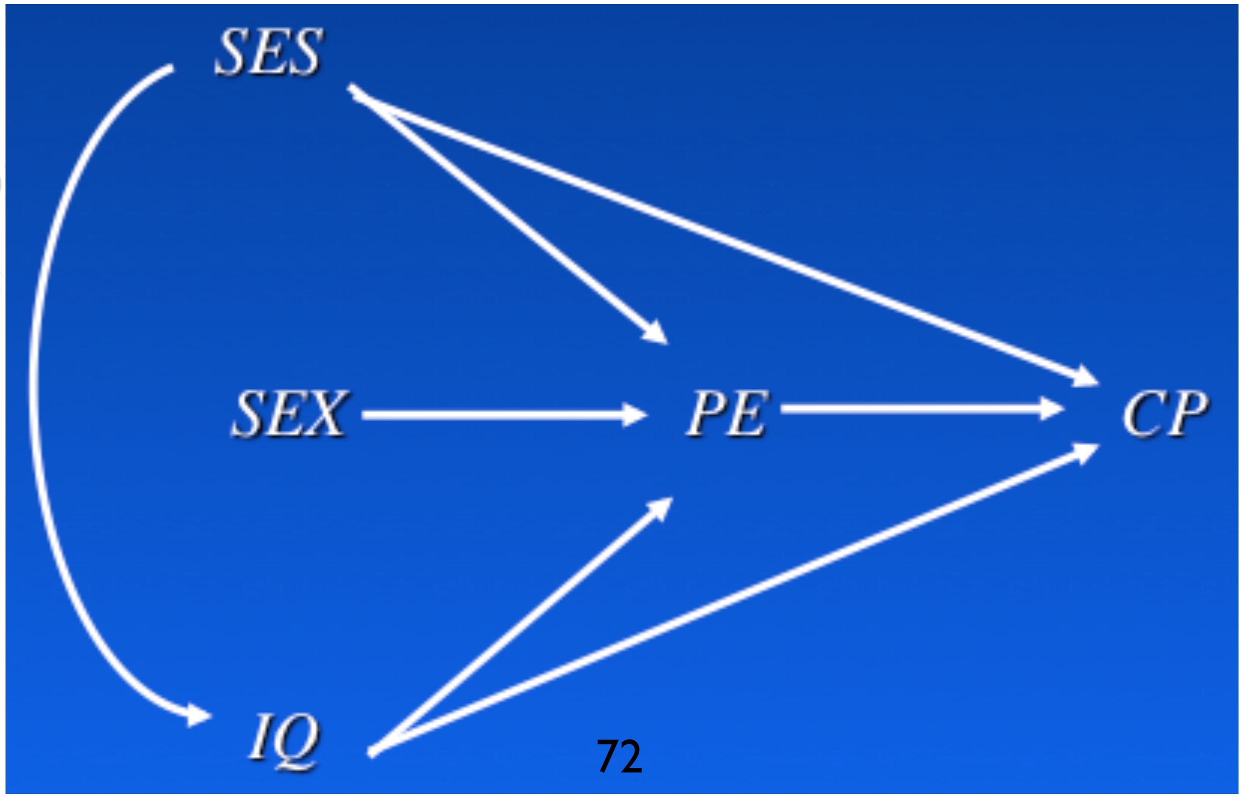
$X_3 \perp\!\!\!\perp X_4 \mid X_2$



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

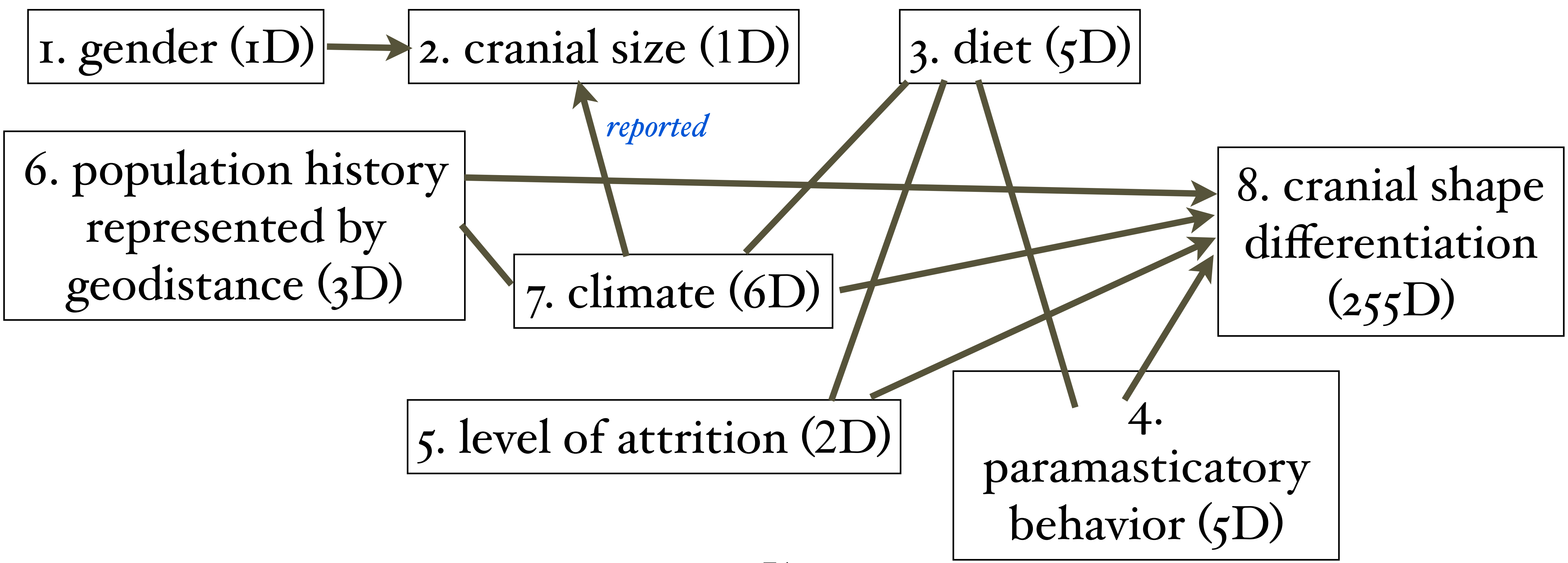
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	Population	Sex	Cranial size	Diet or subsistence					Paramastic	Dental wear		Geographic location per population			Climate per population					
2			(Male, fem	(Centroid S	Gathering	Hunting	Fishing	Pastoralism	Agriculture	Yes=1, no=	Average att	Attrition pe	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	634.3304	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	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
17	AUSM8153	Australia	Male	650.6559	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
19	AUSFB179	Australia	Female	634.3122	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
20	AUSFB175	Australia	Female	605.1759	6	4	0	0	0	1	2.5	1.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
21	AUSFB177	Australia	Female	613.8424	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
22	AUSFB169	Australia	Female	610.1206	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
23	AUSFB157	Australia	Female	628.2819	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
24	AUSFB155	Australia	Female	628.4609	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	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	135.615234	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	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
27	AUSFB158	Australia	Female	631.6258	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	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	646.9841	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	604.4664	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

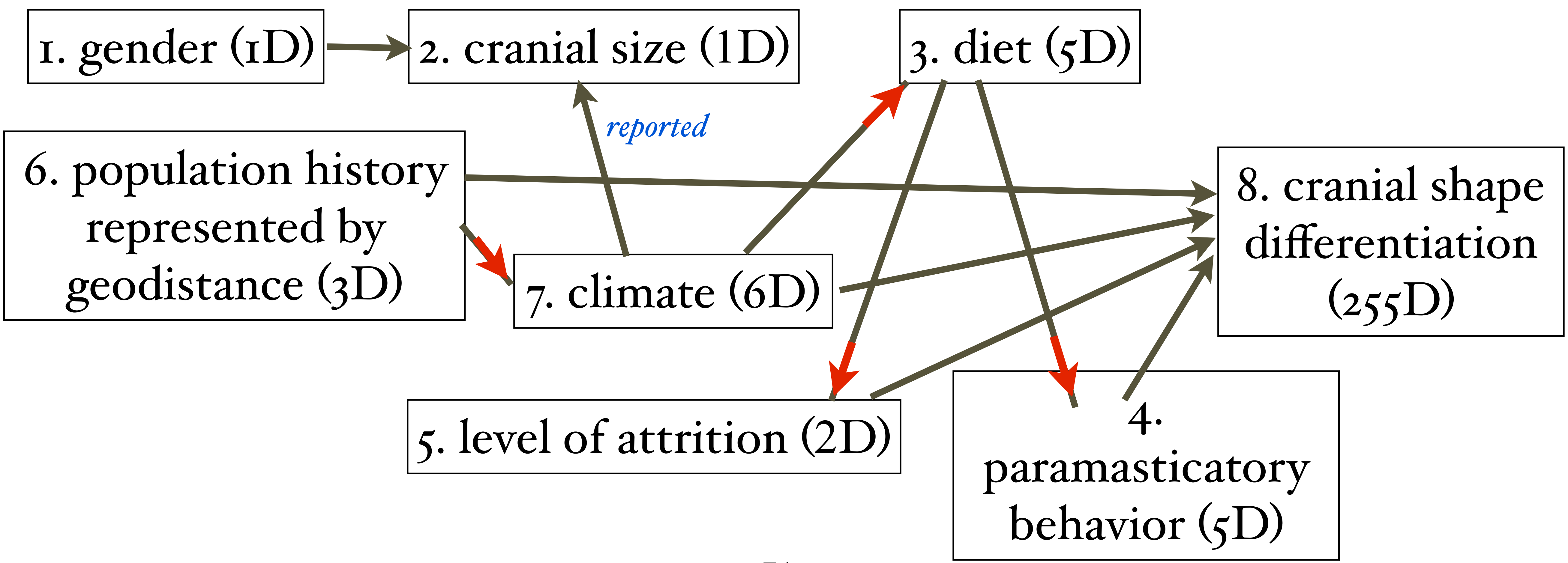
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- PC + kernel-based conditional ind. test seems to be a good choice



Example II: Result

Thanks to collaborator Marlijn Noback

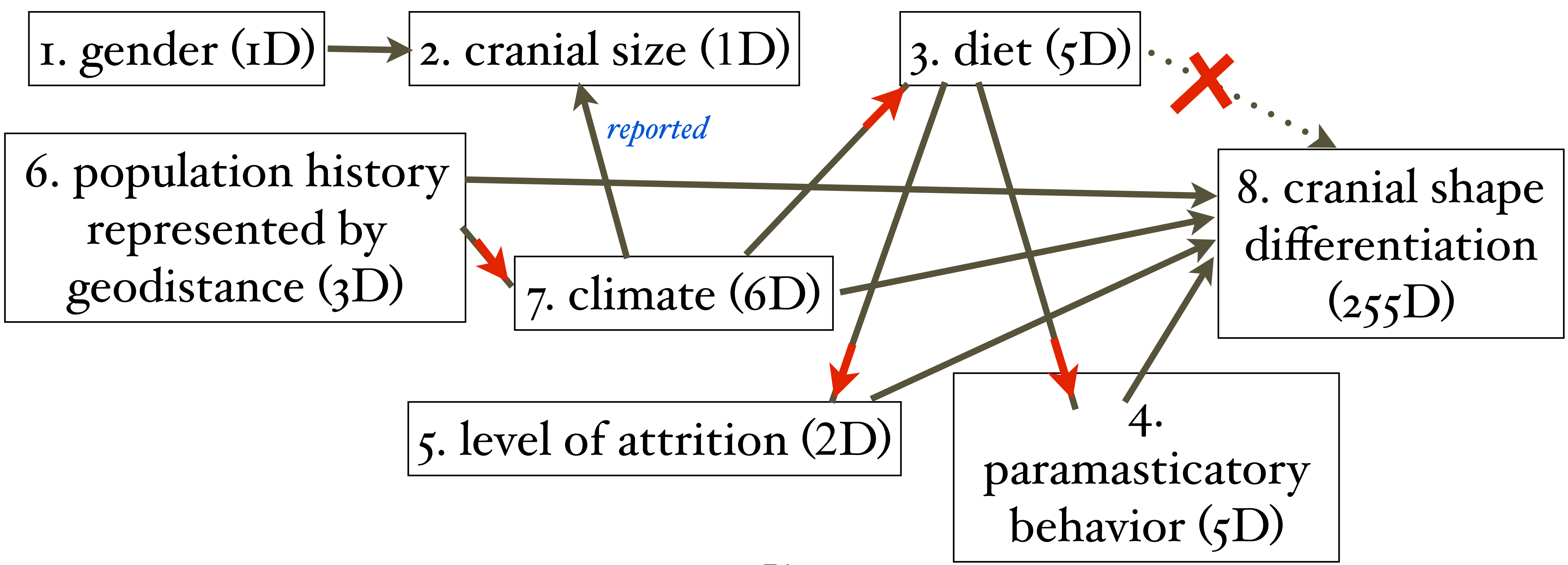
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- PC + kernel-based conditional ind. test seems to be a good choice



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 seems to be a good choice



Confounders? How about This Case?

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

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

Confounders? How about This Case?

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

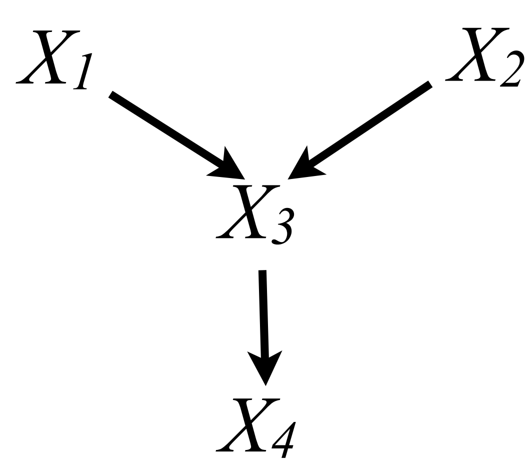
What is the corresponding causal structure? Possible to have confounders behind X_3 and X_4?

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

Confounders? How about This Case?

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

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



:-)

I Can Discover There Is No Confounder: Example

- In the 1970s, the Edison Electric Company in North Carolina was concerned about the effects on plant growth of acid rain produced by emissions from its electric generators.

- The investigators chose samples from the Cape Fear estuary, where the Cape Fear River flows into the Atlantic Ocean.

- obtained 45 samples of Spartina grass up and down the estuary, and measured 13 variables in the samples, including **concentrations of various minerals, acidity (pH), salinity, and the outcome variable, the biomass of each sample**

- The PC algorithm found that among the measured variables the only *direct* cause of biomass was pH.

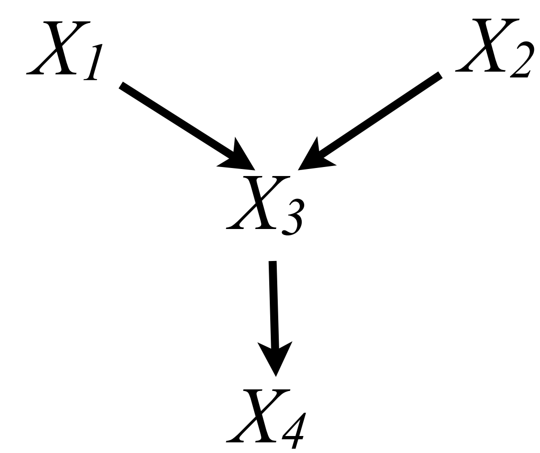
- Y-structure: no confounder!

- Later verified by intervention-based analysis

Other Examples

- A: Raining; B: slippery ground; C: falling down
- A: Geographical background (continental/maritime country); B: economic conditions (agriculture/commerce); C: emergence of science

Other Examples



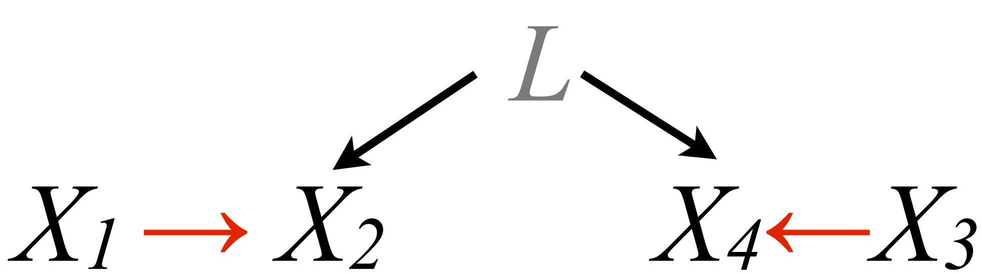
- A: Raining; B: slippery ground; C: falling down
- A: Geographical background (continental/maritime country); B: economic conditions (agriculture/commerce); C: emergence of science

Confounders? How about This Case?

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

Confounders? 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

Confounders? How about This Case?

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

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

L: a latent variable

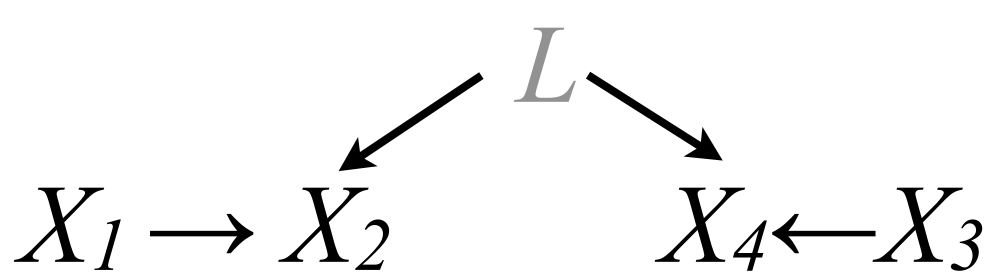
- For example, X_1: I am not sick; X_2: I am in class; X_3: you are in class; X_4: you are not sick

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

- The causal process over measured variables **O** is not necessarily a DAG. How can we represent (independence) equivalence classes over **O** ?

- Results represented by PAGs

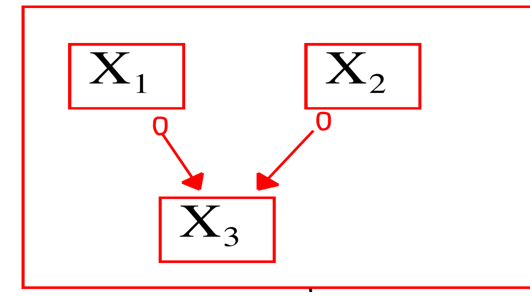


What's FCI's output?

Data available in 'data3_FCI.txt'

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

PAGs (Output of FCI): What Edges Mean?

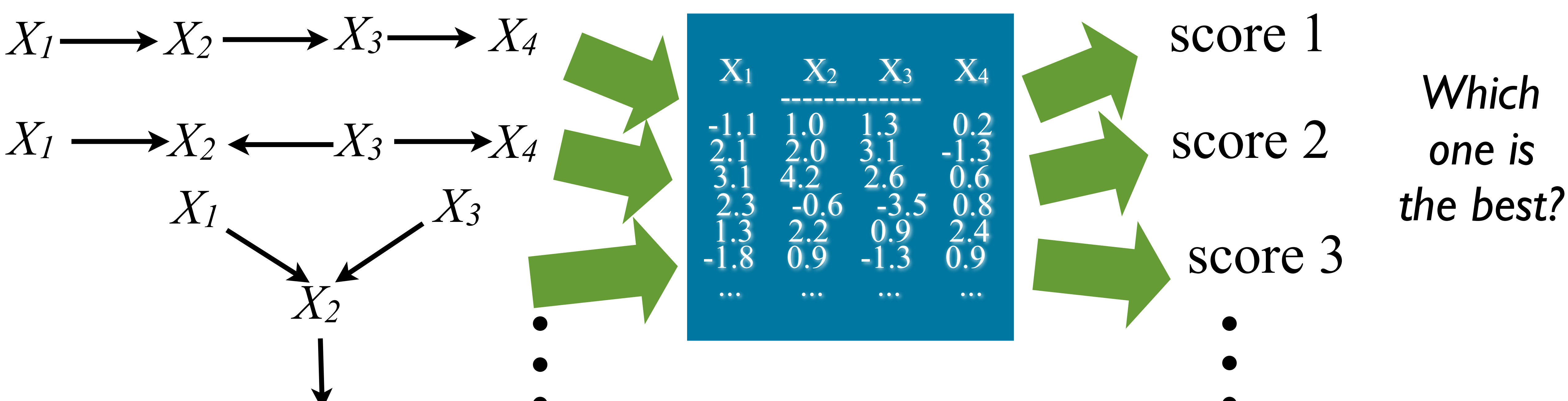


X_1 X_2	X_1 and X_2 are not adjacent
X_1 o→ X_2	X_2 is not an ancestor of X_1
X_1 o—o X_2	No set d-separates X_2 and X_1
$X_1 \rightarrow X_2$	X_1 is a cause of X_2
$X_1 \leftrightarrow X_2$	There is a latent common cause of X_1 and X_2

Constraint-Based vs. Score-Based

- Constraint-based methods

- Score-based methods



(Score may be BIC, AIC, etc.)

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: $S_B(\mathcal{G}, \mathbf{D}) = \log p(\mathbf{D}|\hat{\boldsymbol{\theta}}, \mathcal{G}^h) - \frac{d}{2} \log m$

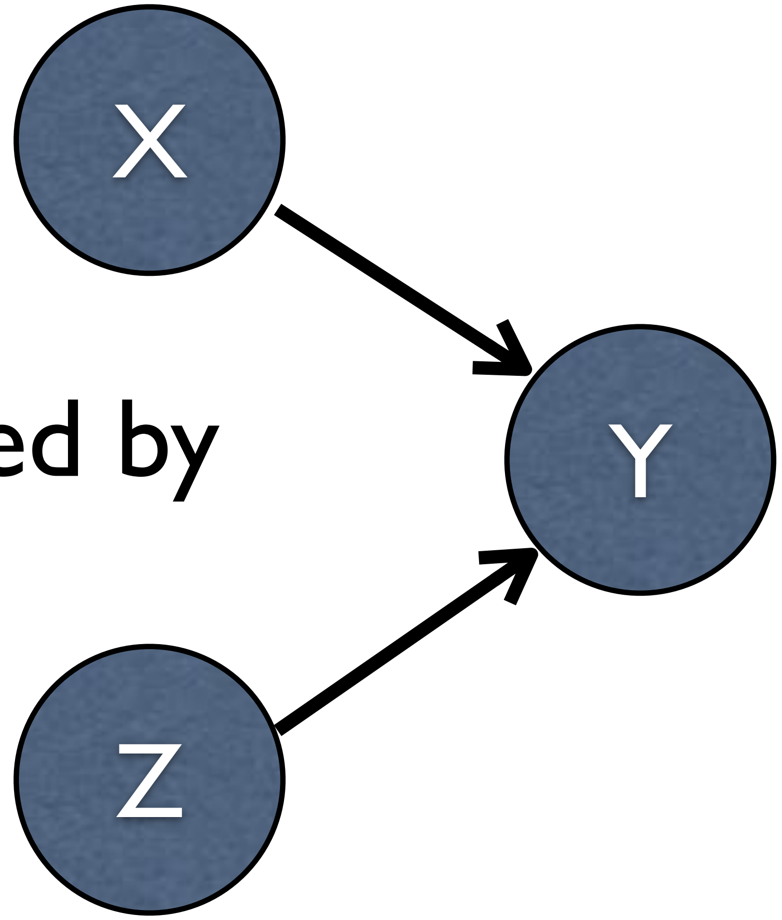
Chickering, Optimal Structure Identification With Greedy Search, Journal of Machine Learning Research, 2002

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

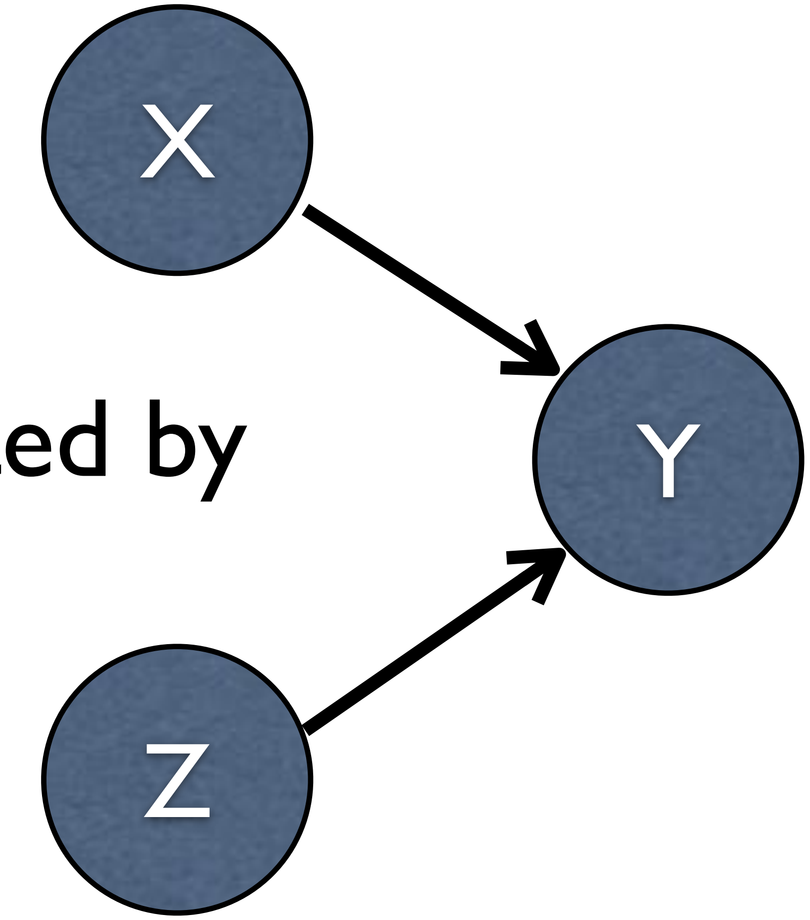
GES

Suppose data were generated by

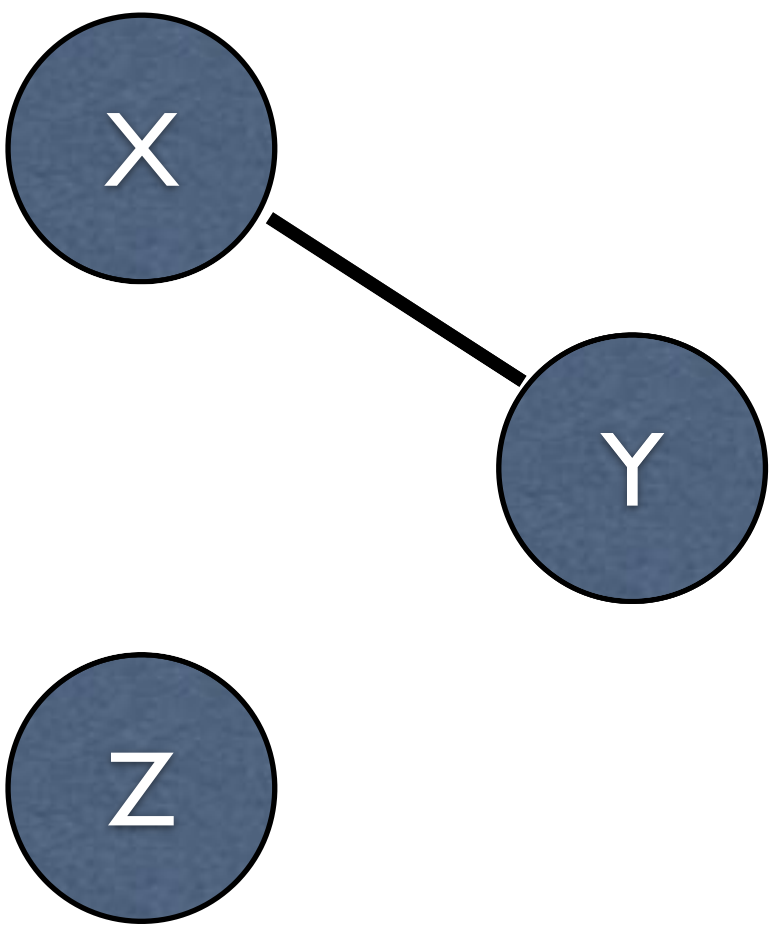


GES

Suppose data were generated by

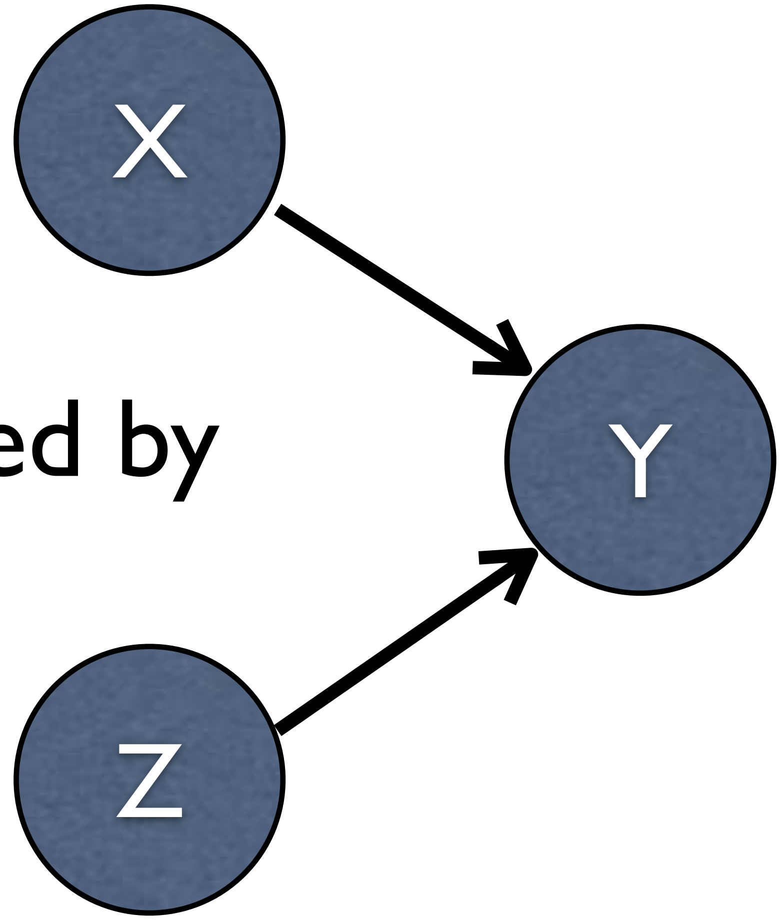


(1)

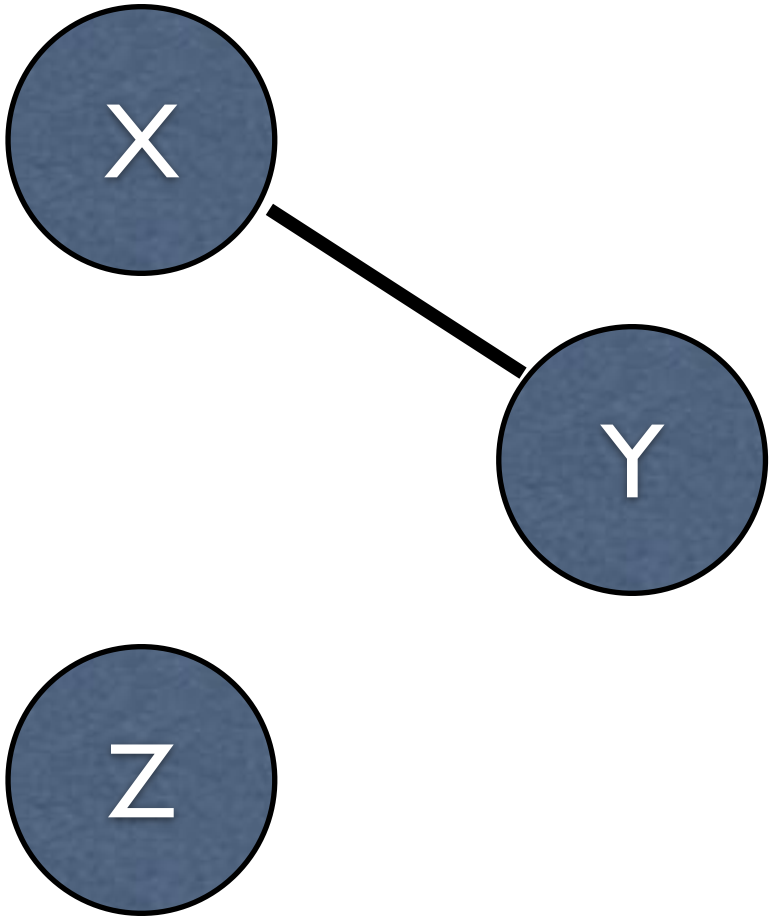


GES

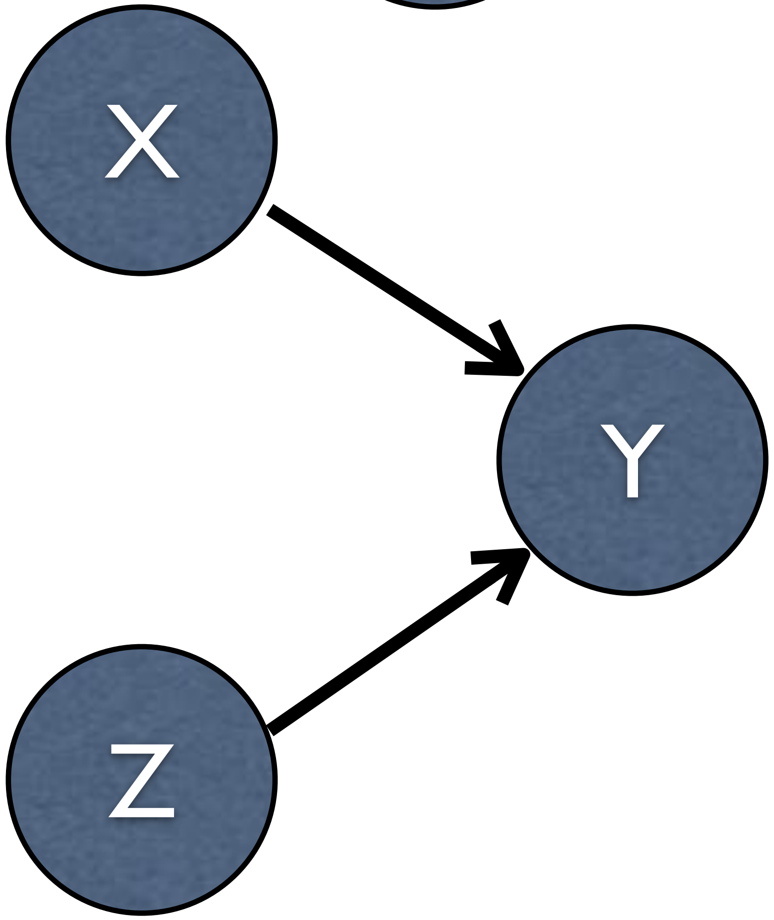
Suppose data were generated by



(1)

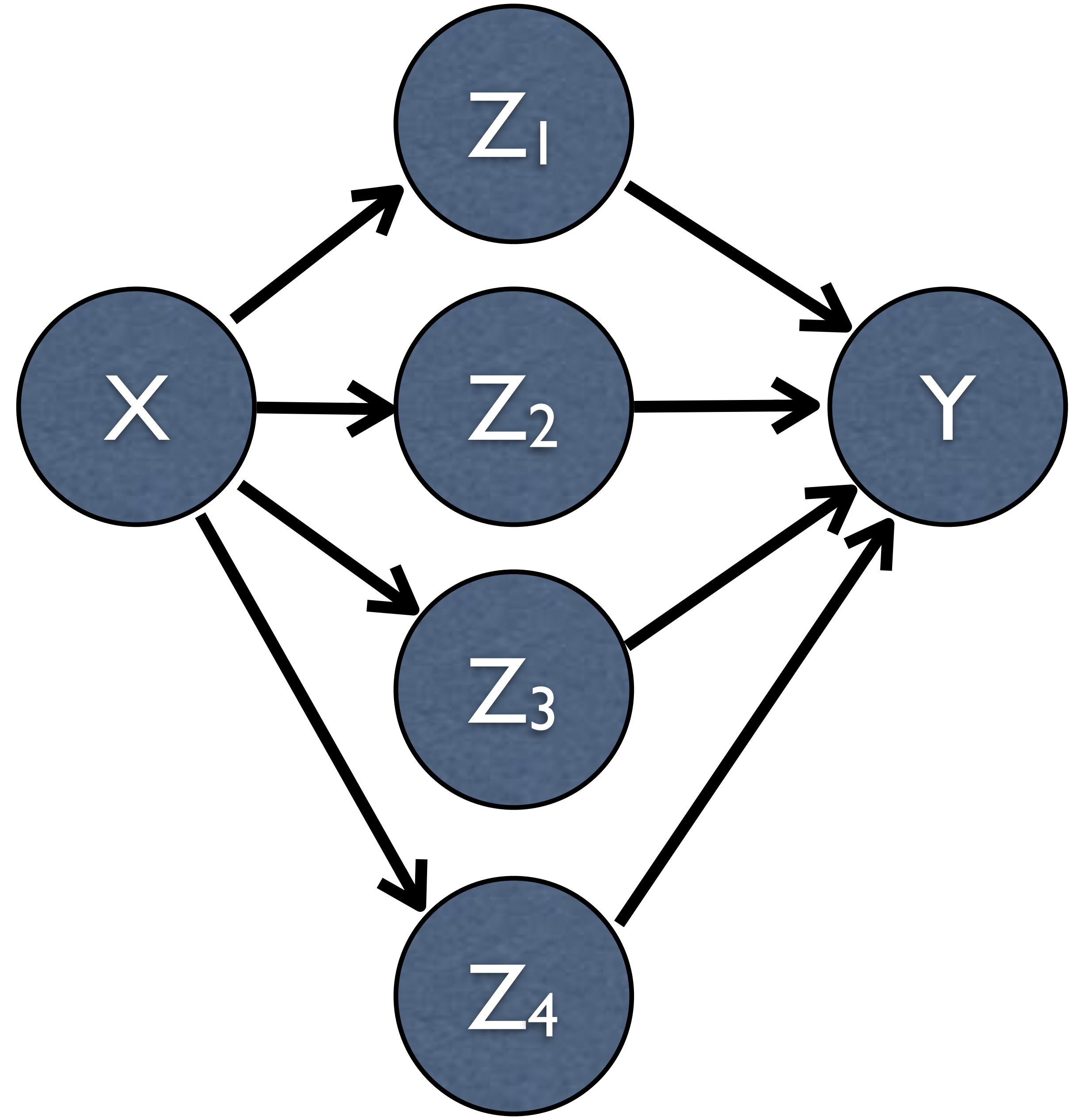


(2)



GES

Suppose data were generated by



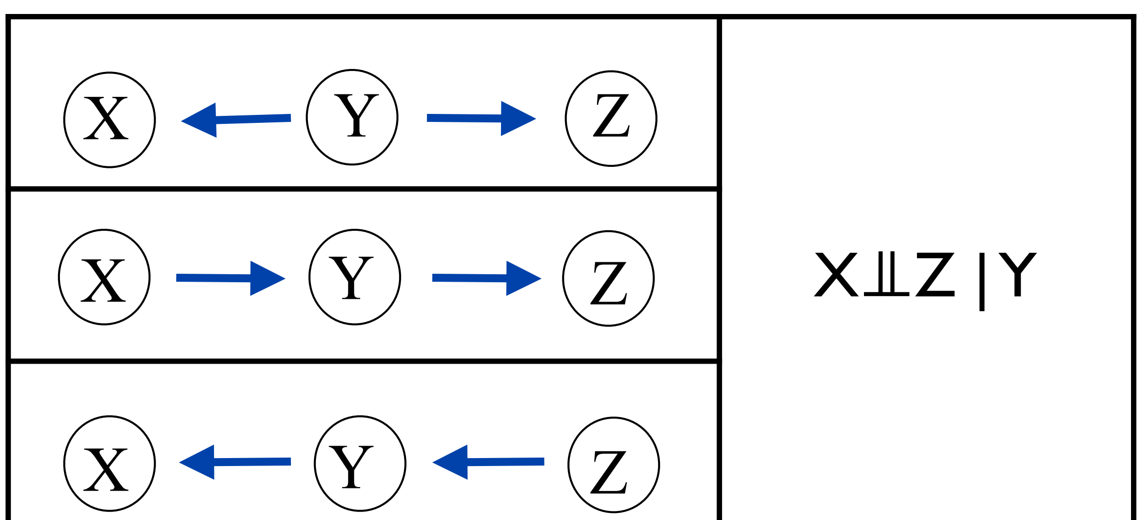
Imagine the GES procedure...

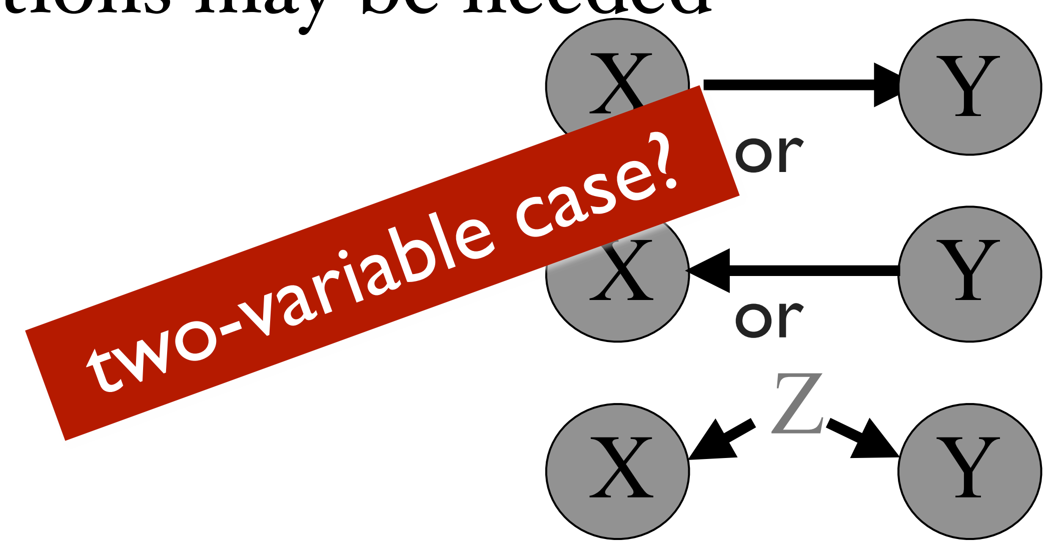
Causal Discovery 2:
Linear, Non-Gaussian Models

- Independent noise condition
- Causal discovery based on structural equation models: linear non-Gaussian case

Constraint-based Causal Discovery: Advantages and Limitations

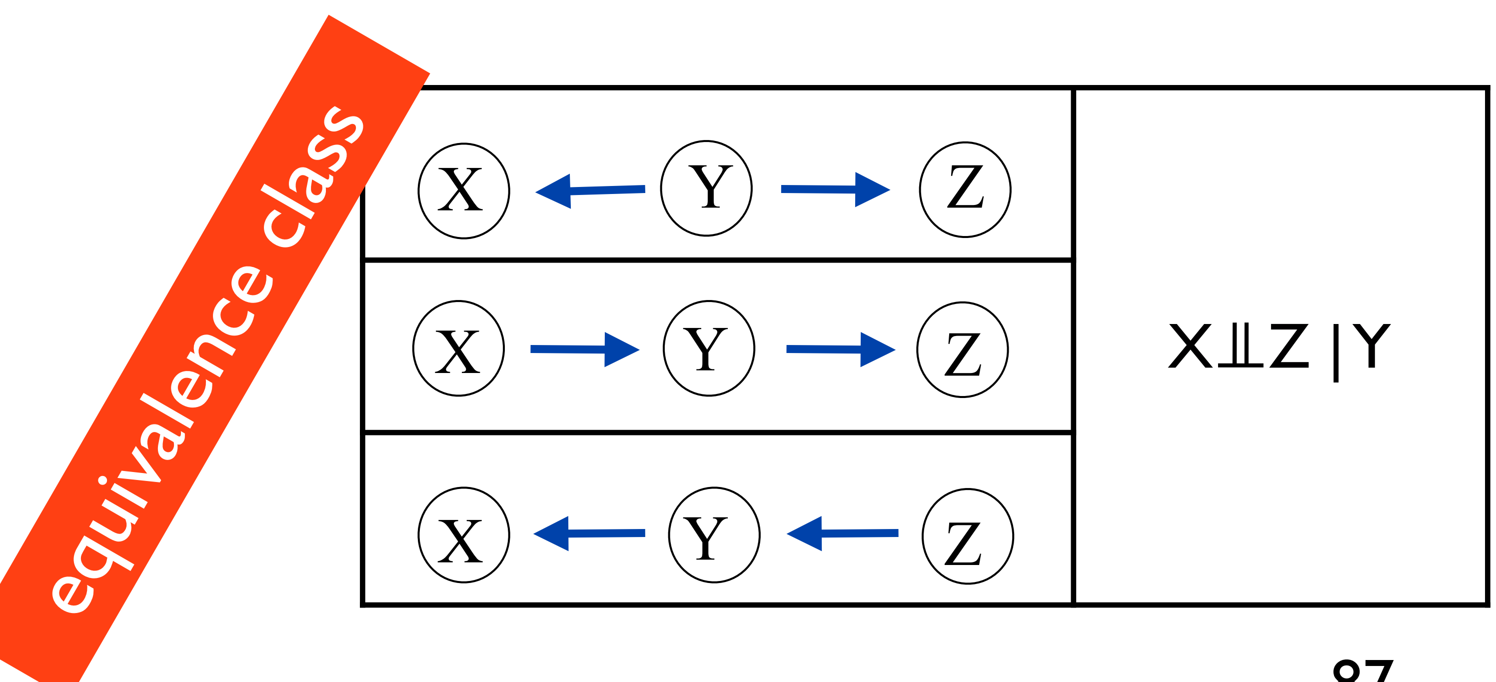
- Nonparametric; widely applicable given reliable conditional independence tests
- Recovering {causal relations} from {conditional independences}: bounded by the equivalence class
- Directly characterize and recover cause-effect relationships?
    - additional weak and reasonable assumptions may be needed

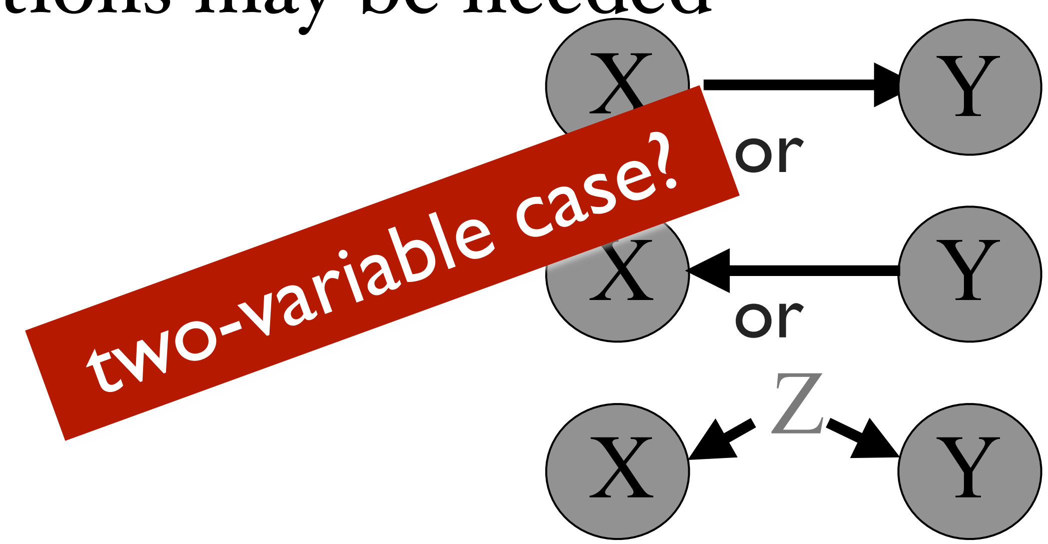




Constraint-based Causal Discovery: Advantages and Limitations

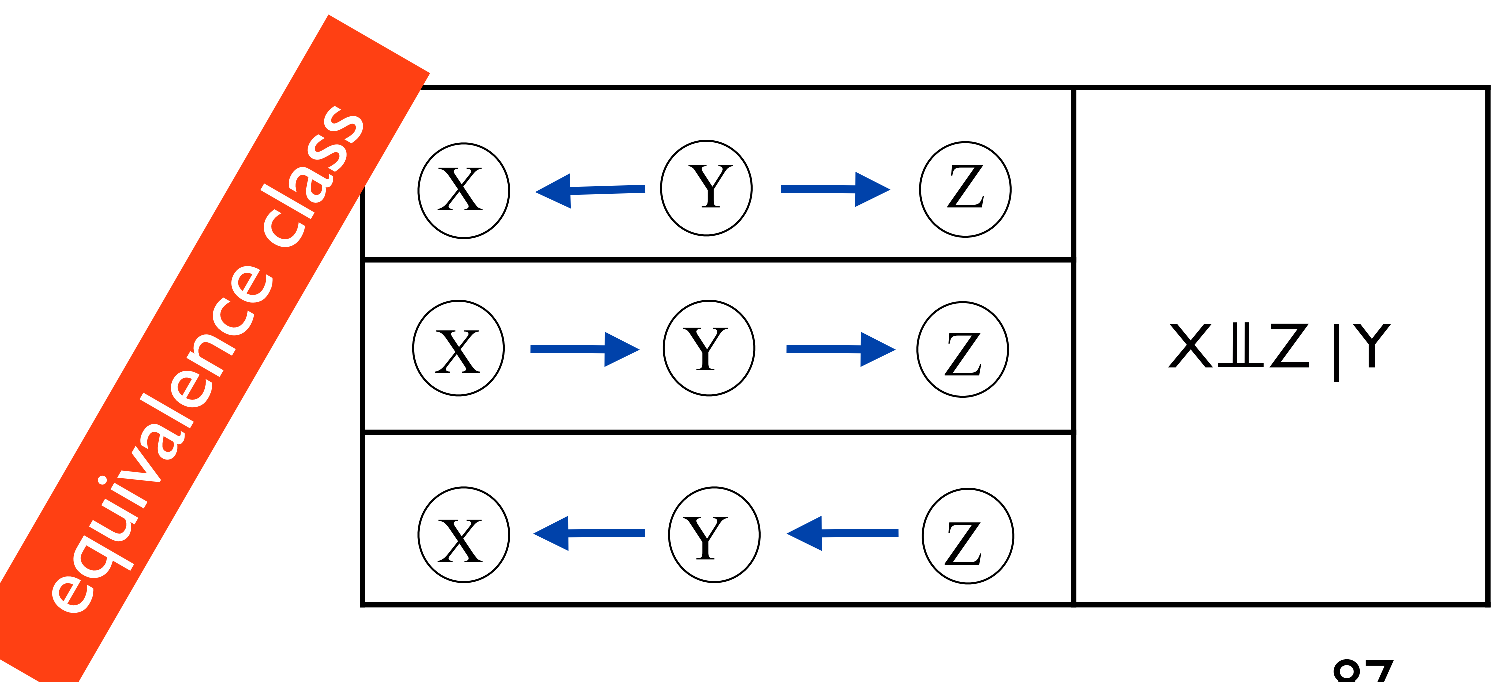
- Nonparametric; widely applicable given reliable conditional independence tests
- Recovering {causal relations} from {conditional independences}: bounded by the equivalence class
- Directly characterize and recover cause-effect relationships?
  - additional weak and reasonable assumptions may be needed

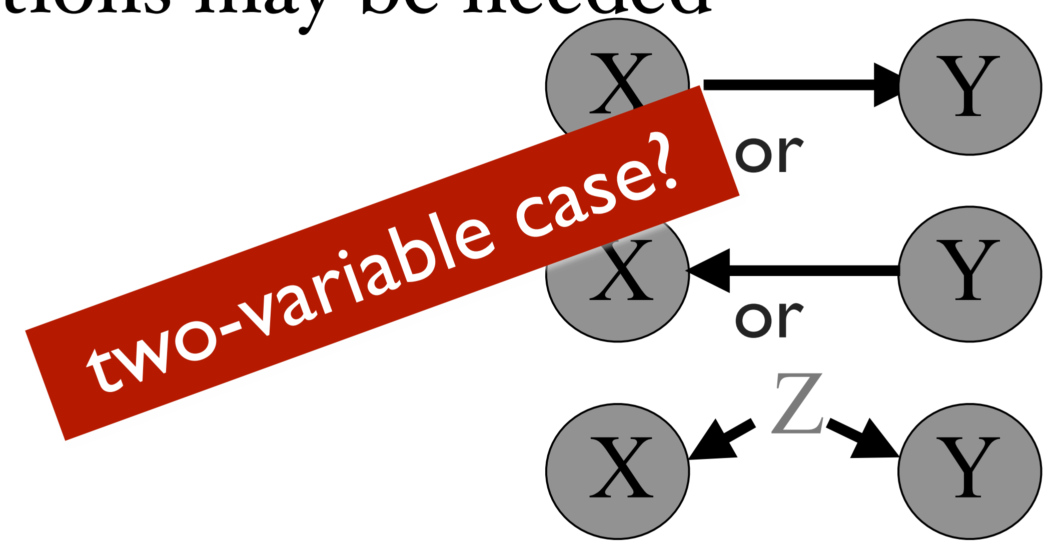




Constraint-based Causal Discovery: Advantages and Limitations

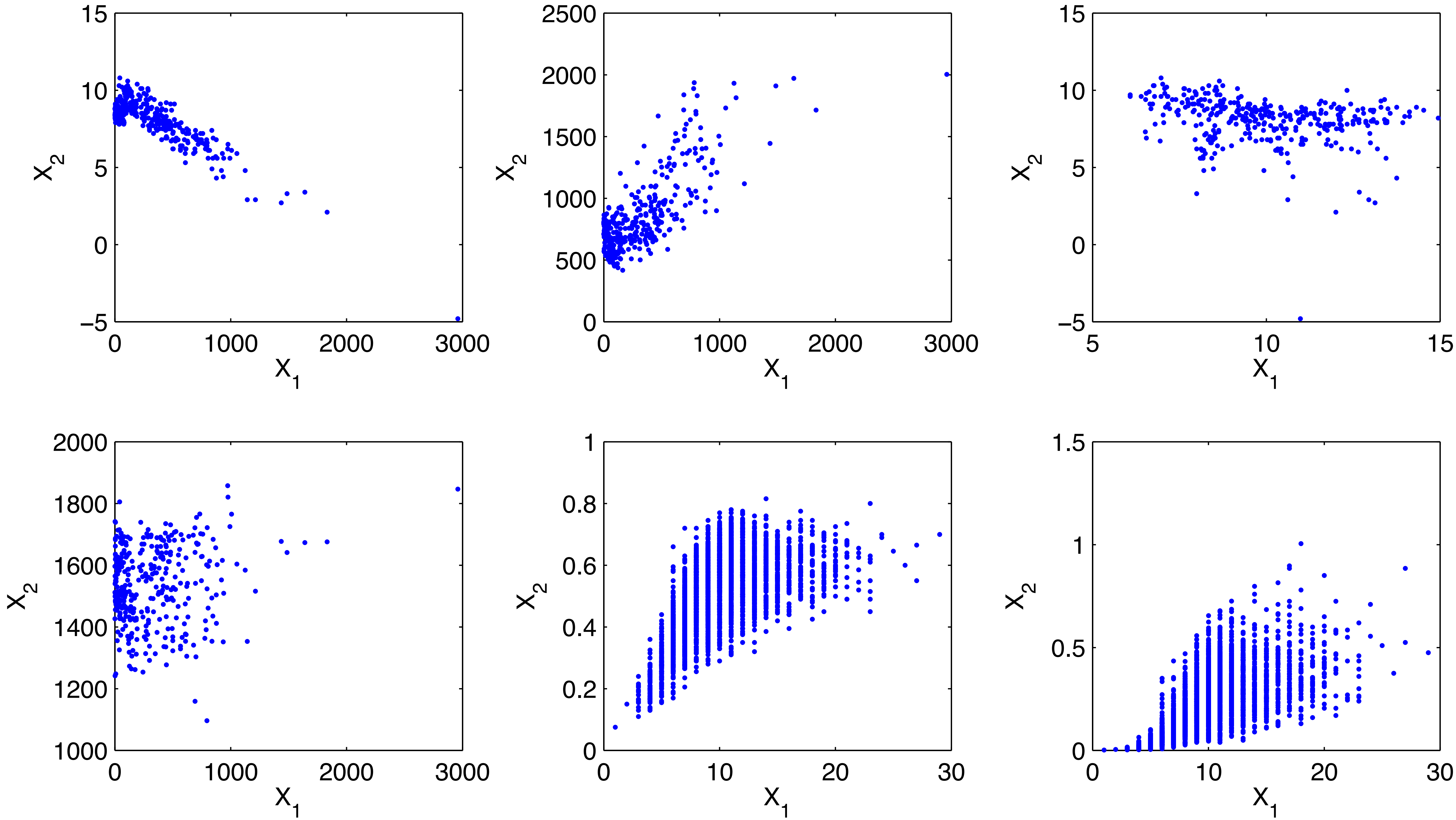
- Nonparametric; widely applicable given reliable conditional independence tests
- Recovering {causal relations} from {conditional independences}: bounded by the equivalence class
- Directly characterize and recover cause-effect relationships?
  - additional weak and reasonable assumptions may be needed





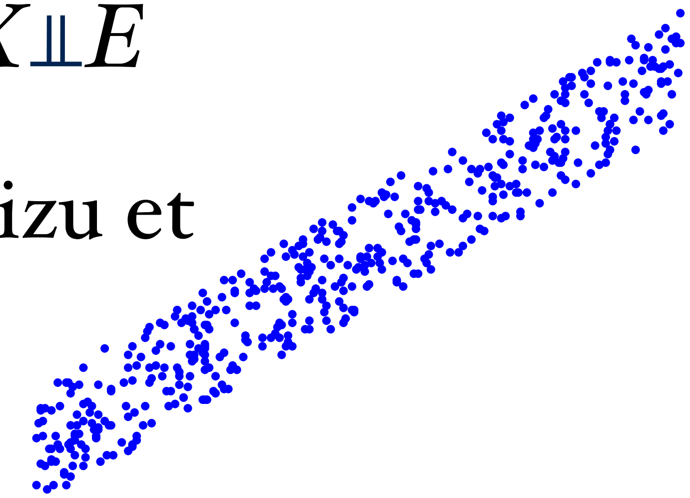
- Instead, try to directly identify local causal structures with functional causal models/structural equation models

Distinguishing Cause from Effect?



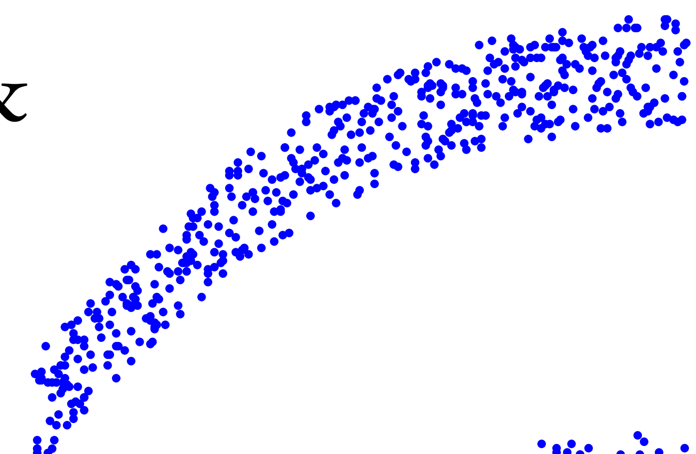
Fully Identify Causal Structure? FCMs!

- A **functional causal model** represents effect as a function of direct causes and noise: $Y = f(X, E)$, with $X \perp\!\!\!\perp E$



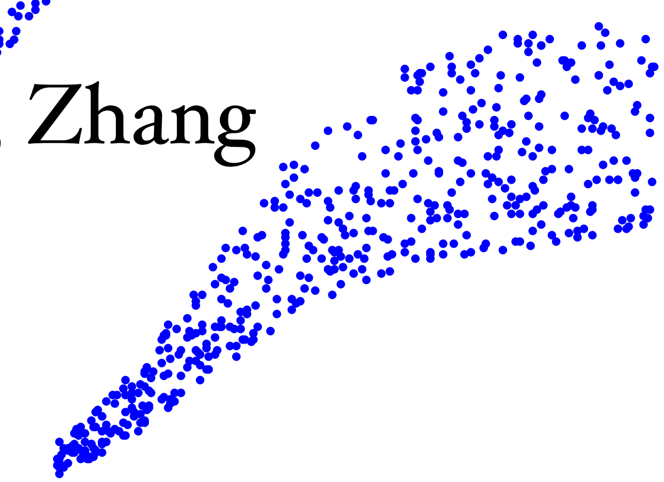
- Linear non-Gaussian acyclic causal model (Shimizu et al., ‘06)

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



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

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

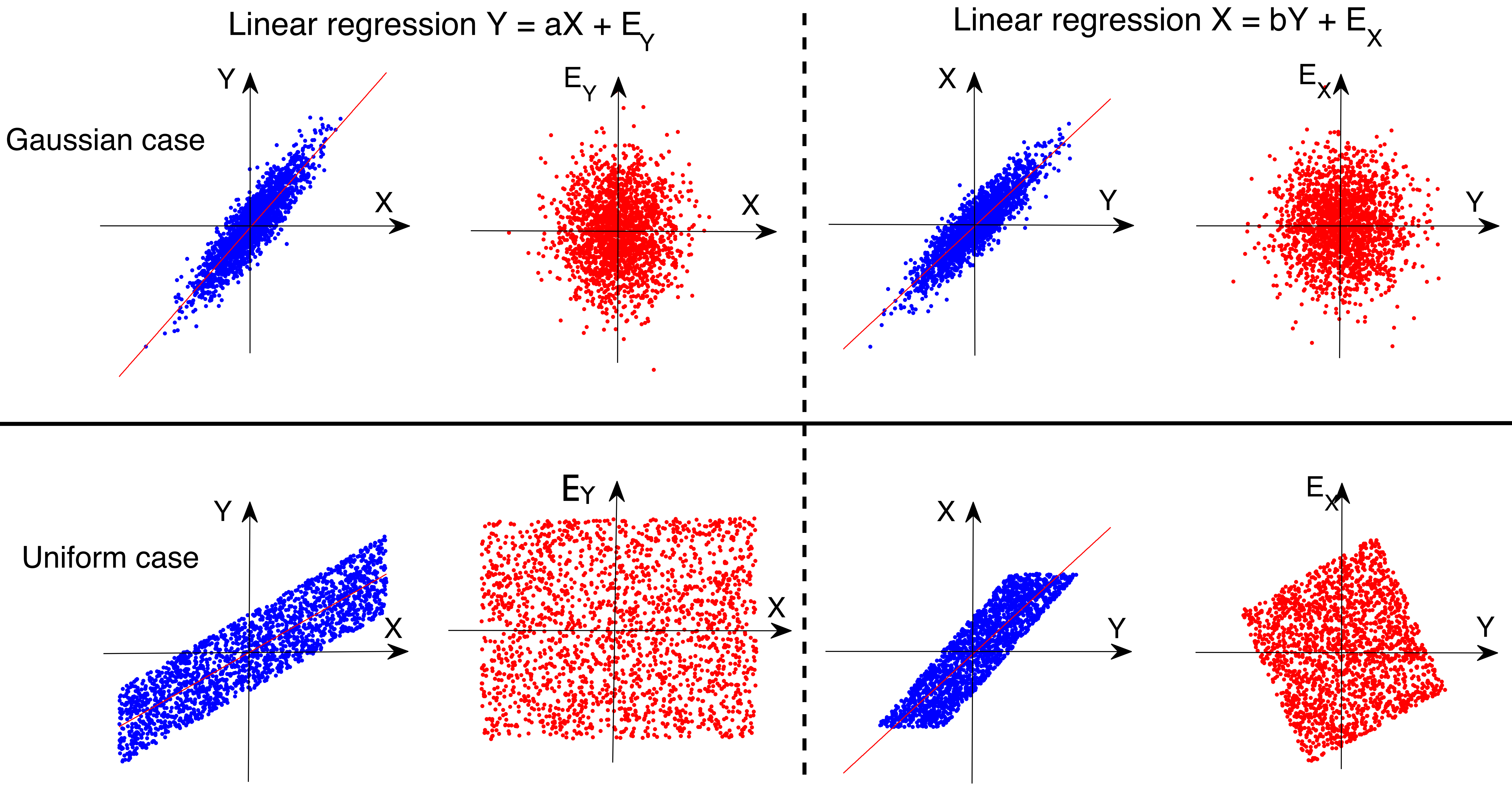


- Post-nonlinear causal model (Zhang & Chan, ’06; Zhang & Hyvärinen, ‘09a)

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

Causal Asymmetry the Linear Case: Illustration

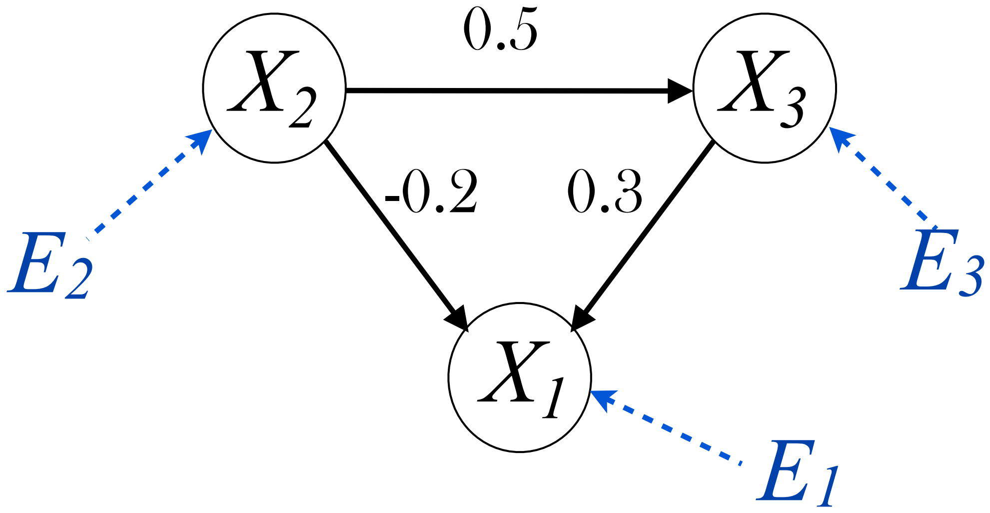
Data generated by $Y = aX + E$ (i.e., $X \rightarrow Y$):



More Generally, LiNGAM Model

- Example:

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

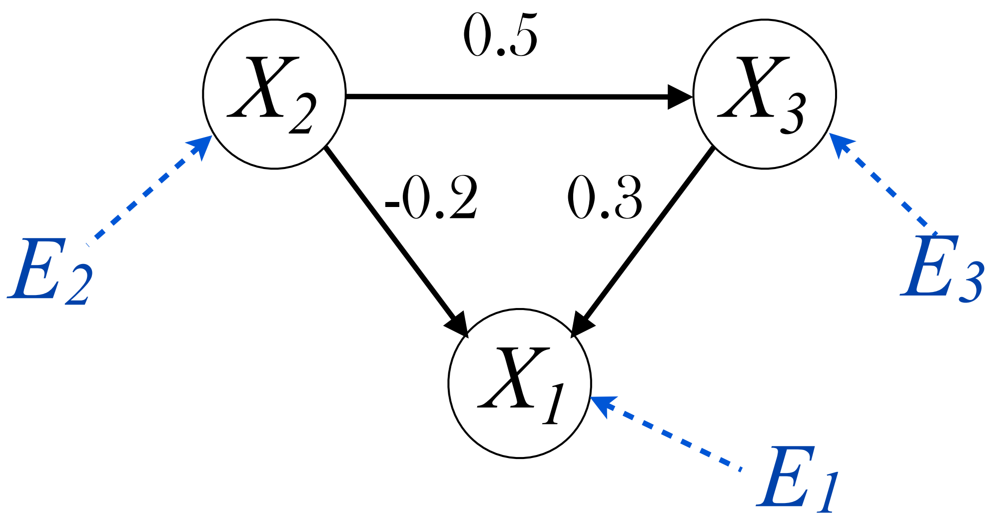


Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.

More Generally, LiNGAM Model

- Example:

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



Matrix form:

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

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

More Generally, LiNGAM Model

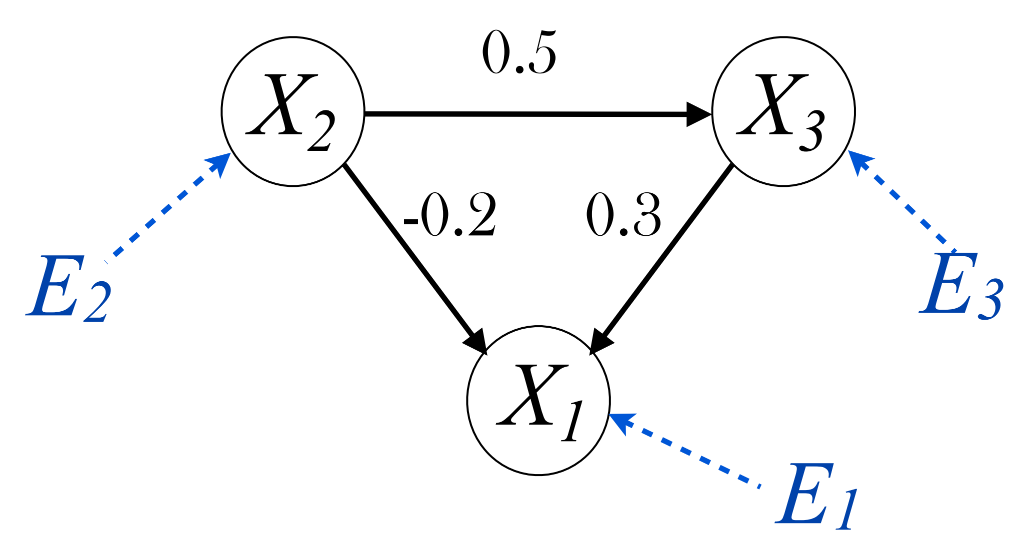
- Linear, non-Gaussian, acyclic causal model (LiNGAM) (Shimizu et al., 2006):

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

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

  - Example:

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



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

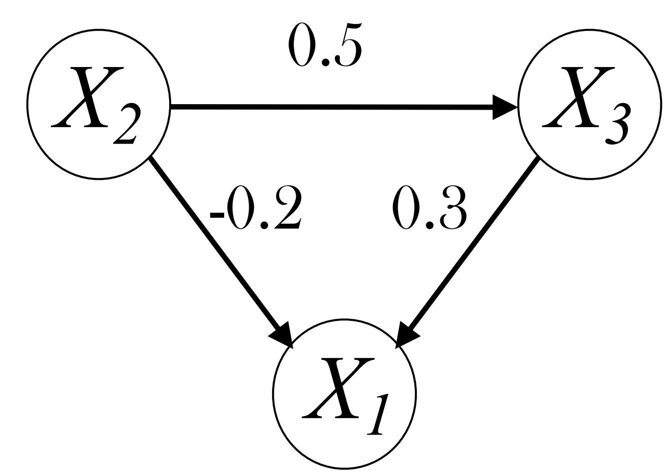
LiNGAM Analysis by ICA

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

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

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

So we have the causal relation:

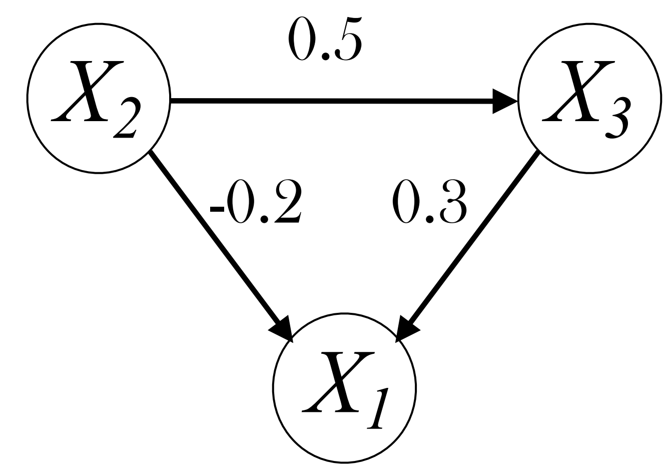


LiNGAM Analysis by ICA

- LiNGAM: $X_i = \sum_{j:\text{ parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{BX} + \mathbf{E} \Rightarrow$ $\mathbf{E} = \mathbf{(I\text{-}B)X}$
  - $\mathbf{B}$ has special structure: acyclic relations
- ICA: $\mathbf{Y} = \mathbf{WX}$
- $\mathbf{B}$ can be seen from $\mathbf{W}$ 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?

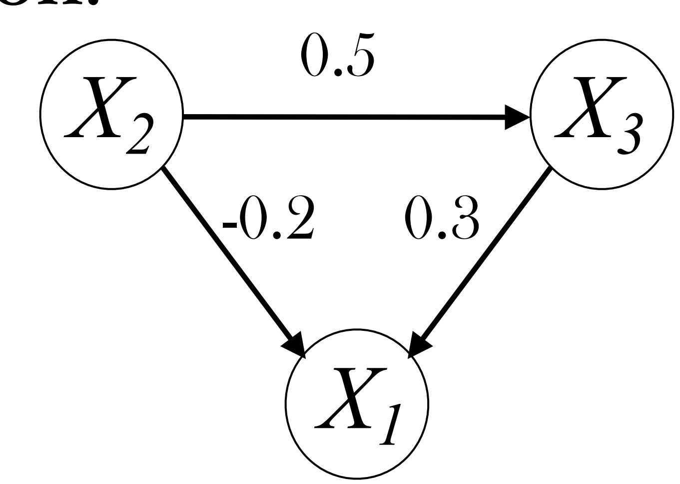
LiNGAM Analysis by ICA

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

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

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

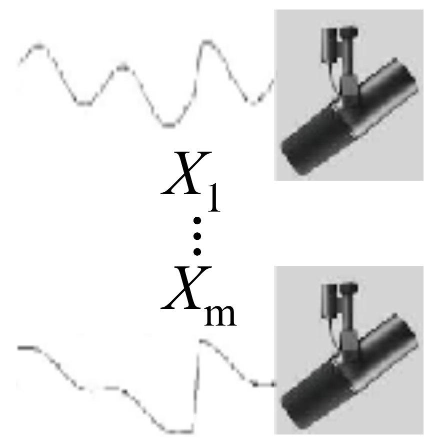
- ICA gives $\mathbf{Y} = \mathbf{W}\mathbf{X}$ and

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

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?

Independent Component Analysis



observed signals

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

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

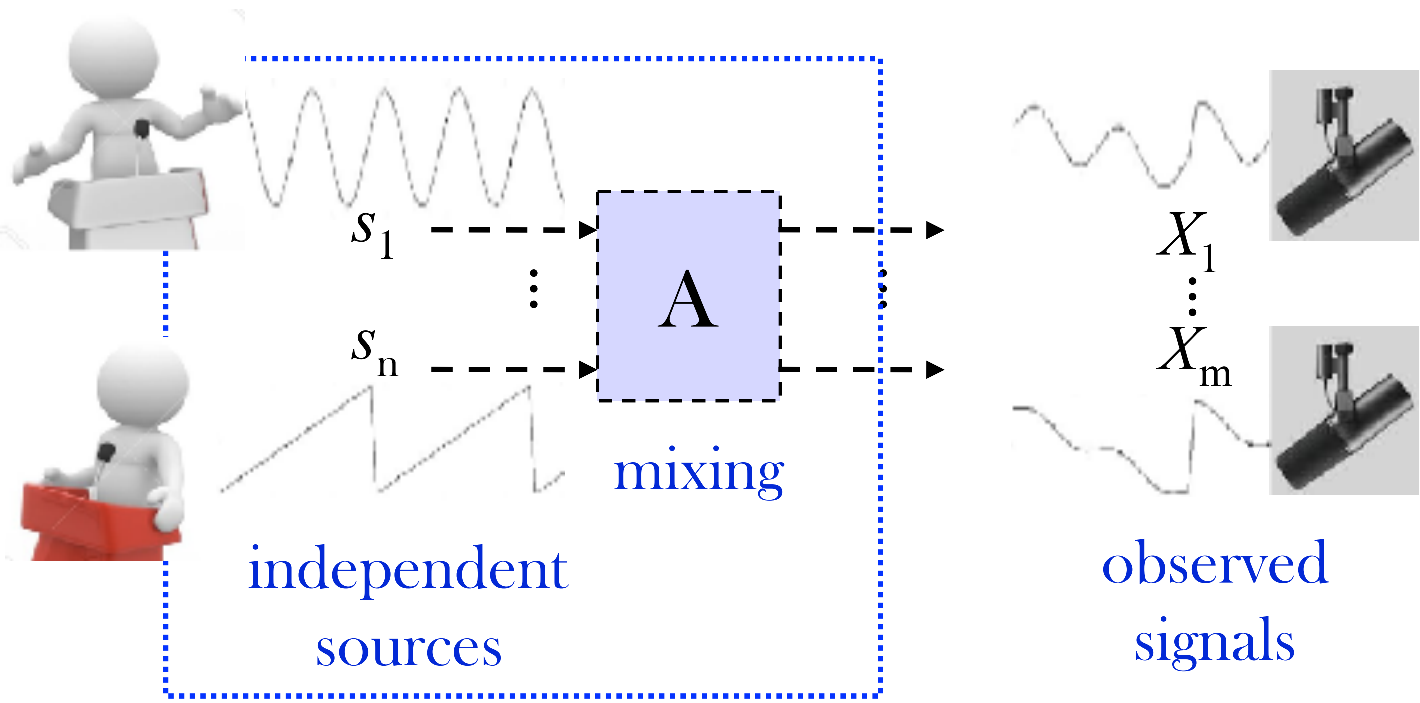
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \underbrace{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\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

- 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



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

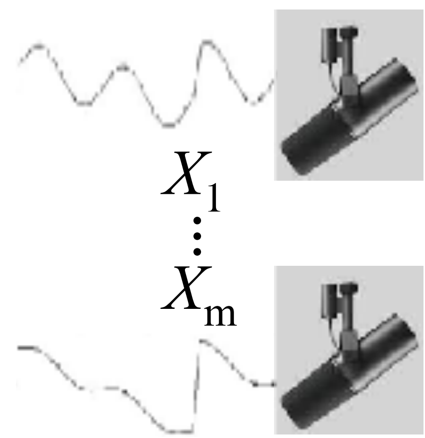
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \ldots \\ .8 & -.7 & .3 & .5 & \ldots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \ldots \\ ? & ? & ? & ? & \ldots \end{bmatrix} \begin{array}{c} s_1 \\ s_2 \end{array}$$

- 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

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

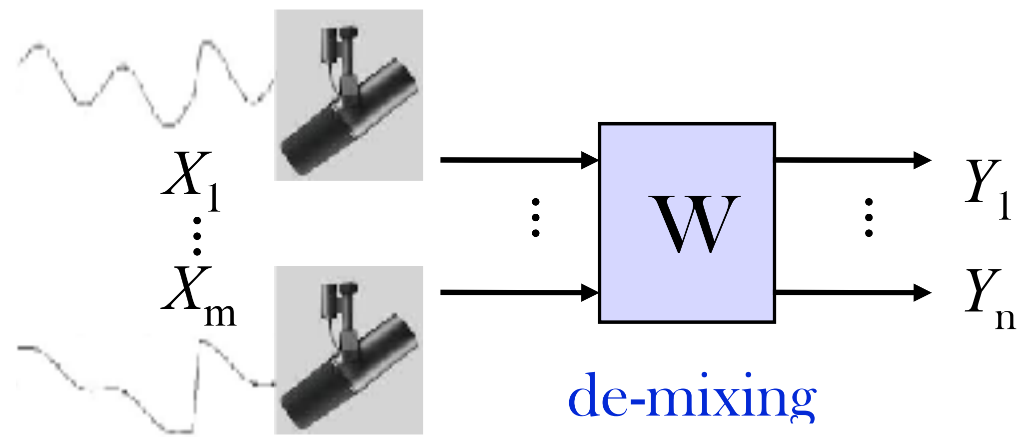
$$\boxed{\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \underbrace{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix}} \begin{array}{c}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

- 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

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

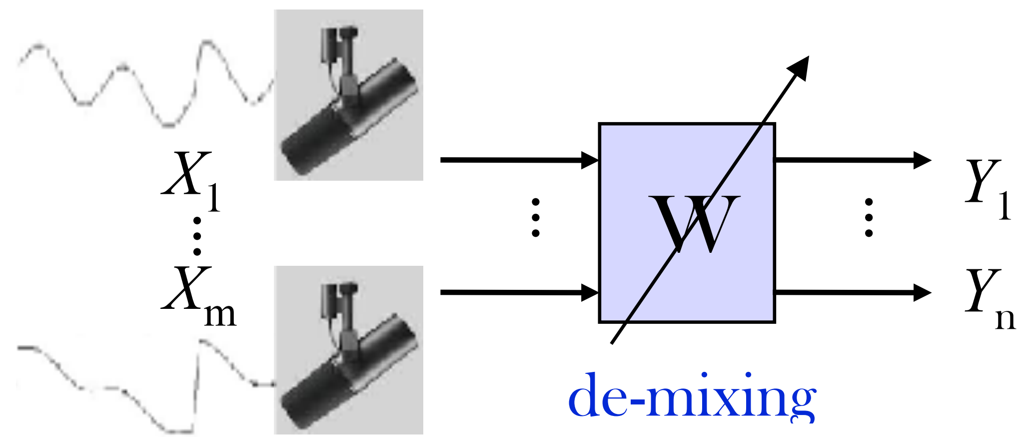
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\ s_2 \end{array}$$

- 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

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

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

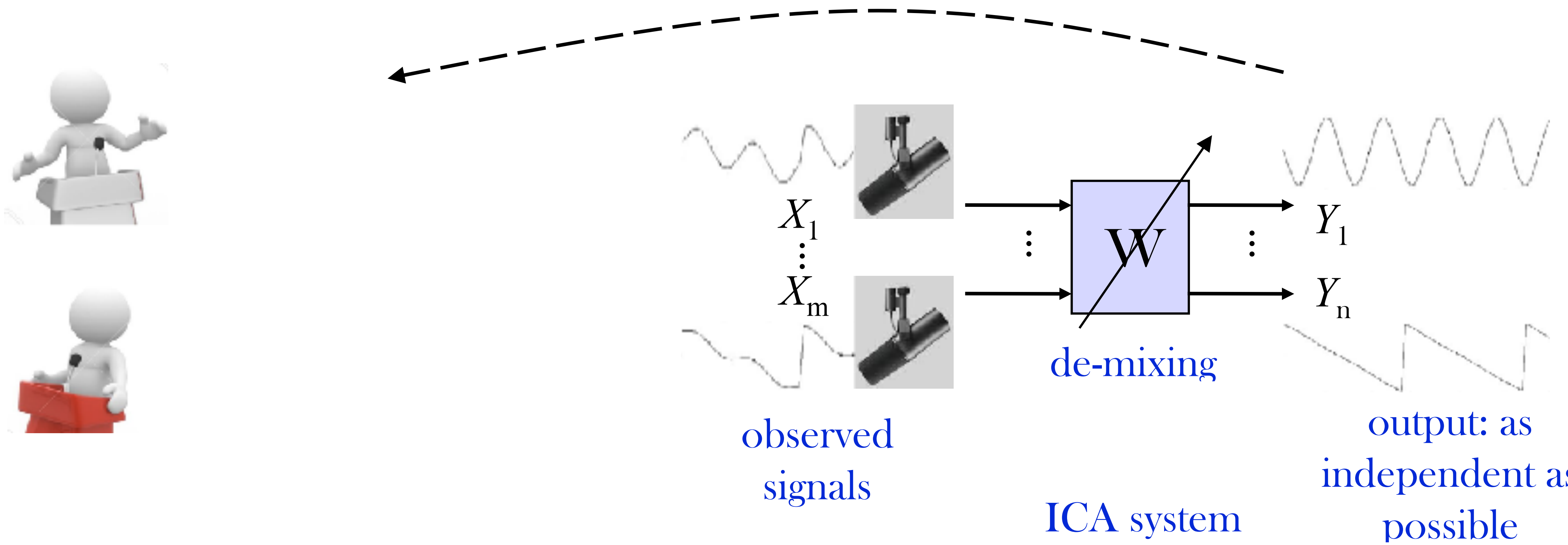
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\ s_2 \end{array}$$

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

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

Hyvärinen et al., *Independent Component Analysis*, 2001

Independent Component Analysis



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

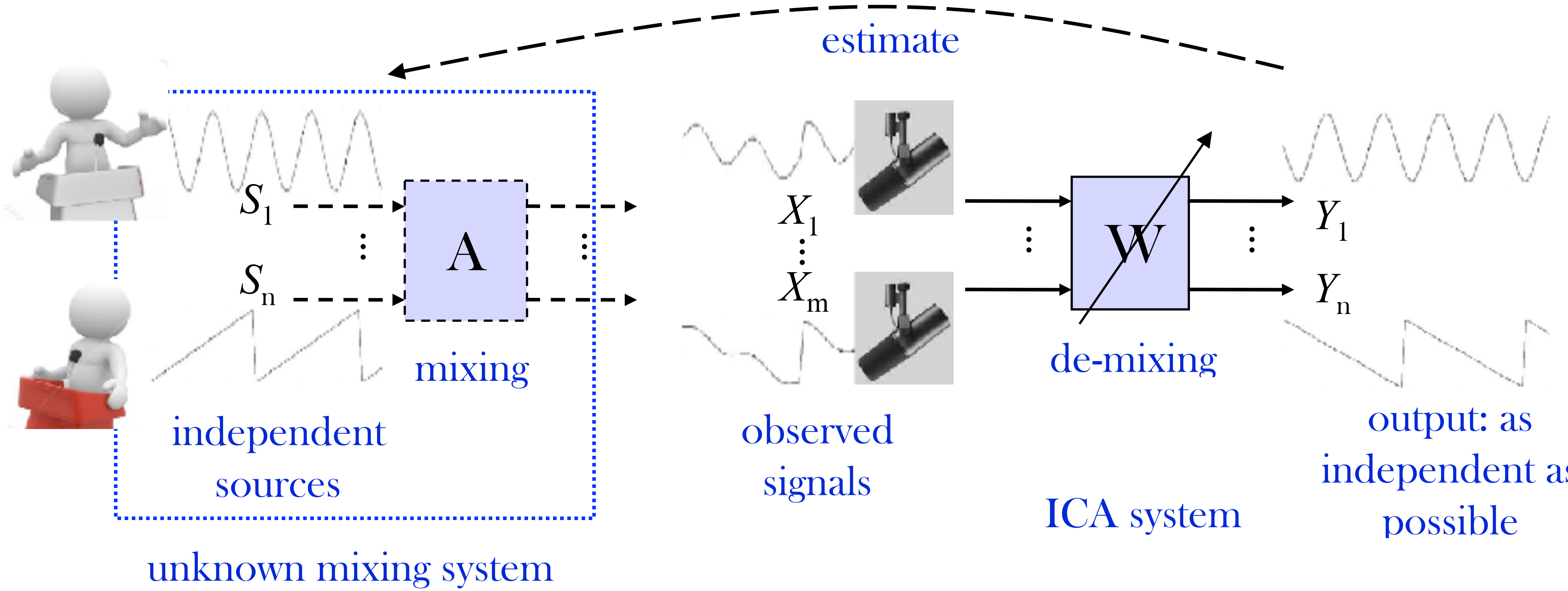
$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \underbrace{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}}_{A} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\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

- 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



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

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

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

- 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

Darmois-Skitovich Theorem

Darmois-Skitovitch theorem: Define two random variables, Y_1 and Y_2, as linear combinations of independent random variables S_i, $i = 1, ..., n$:

$$\begin{aligned} Y_1 &= \alpha_1 S_1 + \alpha_2 S_2 + ... + \alpha_n S_n, \\ Y_2 &= \beta_1 S_1 + \beta_2 S_2 + ... + \beta_n S_n. \end{aligned}$$

If Y_1 and Y_2 are statistically independent, then all variables S_j for which $\alpha_j \beta_j \neq 0$ are Gaussian.

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

How ICA works? By Mutual Information Minimization (or ML)

- Mutual information $I(Y_1,\dots,Y_n)$ is the Kullback-Leiber divergence from P_Y to $\prod_i P_{Yi}$:

$$
\begin{aligned}
I(Y_1,...,Y_n) &= \int \dots \int p_{Y_1,...,Y_n} \log \frac{P_{Y_1,...,Y_n}}{p_{Y_1}...p_{Y_n}} dy_1...dy_n \\
&= \int \dots \int p_{Y_1,...,Y_n} \log P_{Y_1,...,Y_n} dy_1...dy_n - \int p_{Y_1,...,Y_n} \sum_{i=1}^{n} \log p_{Y_i} dy_i \\
&= \sum_i H(Y_i) - H(Y) \\
&= \sum_i H(Y_i) - H(X) - \log |\mathbf{W}| \qquad \text{because } \mathbf{Y} = \mathbf{W}\mathbf{X}
\end{aligned}
$$

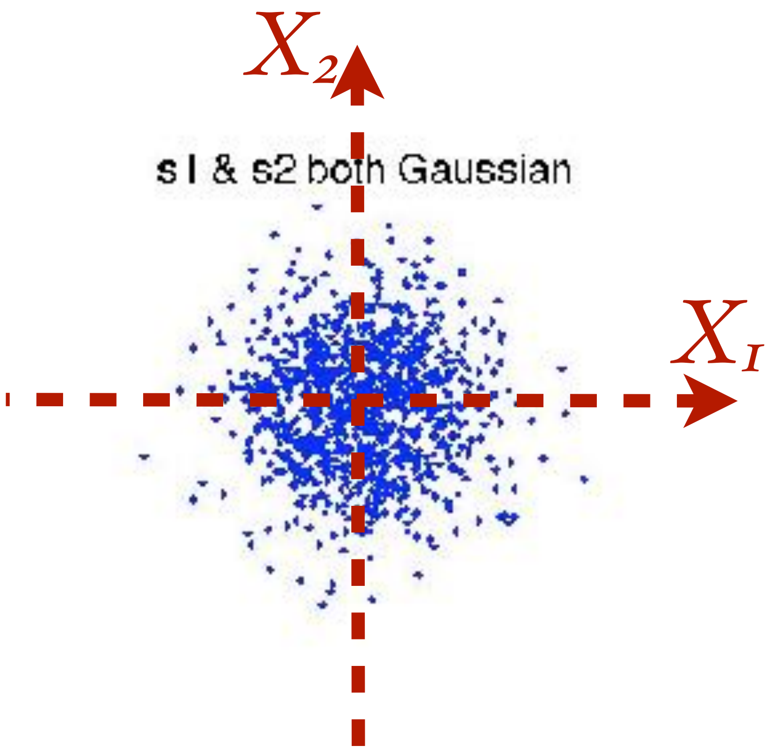
- Nonnegative and zero iff Y_i are independent

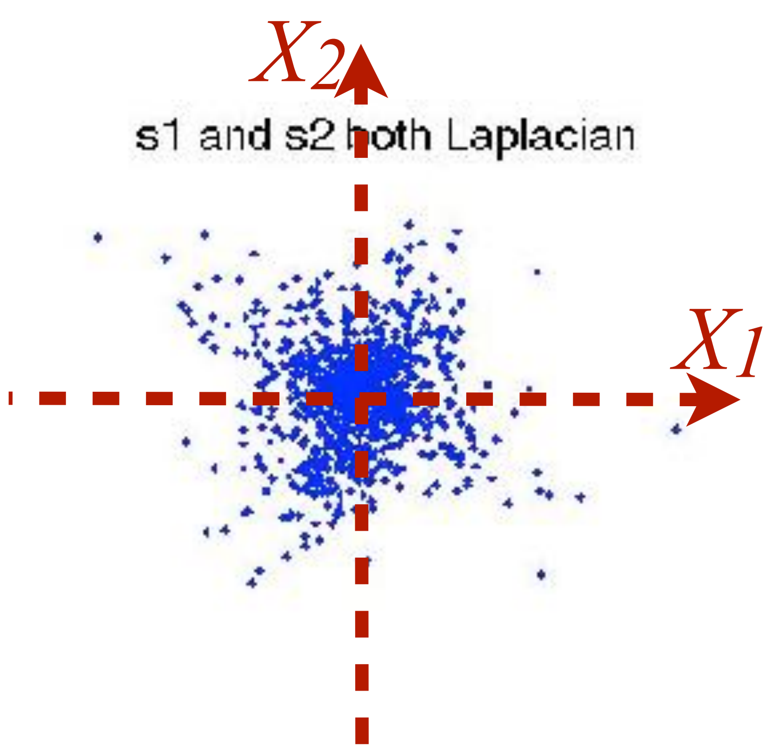
- $H(\cdot)$: differential entropy--how random the variable is?

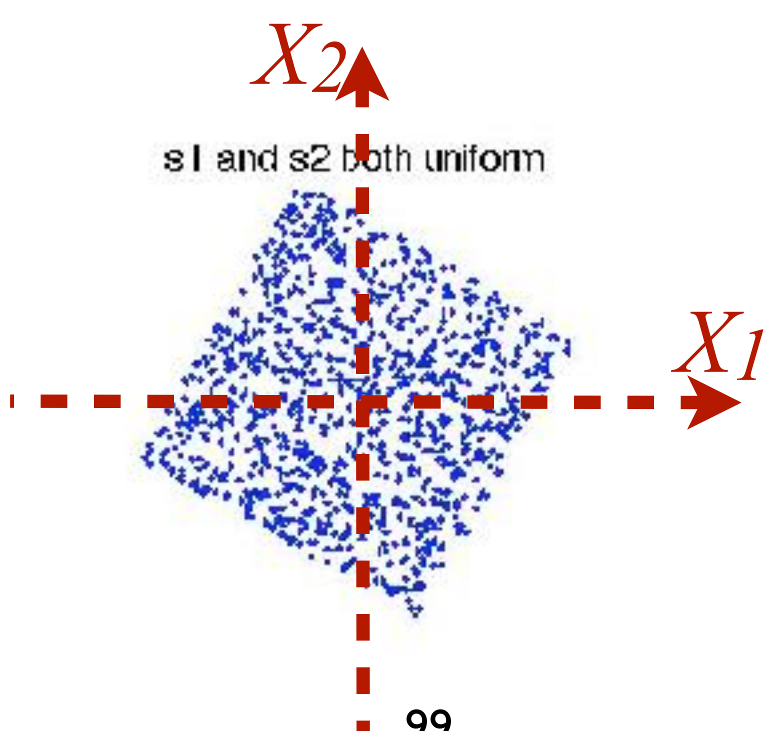
Hyvärinen et al., Independent Component Analysis...

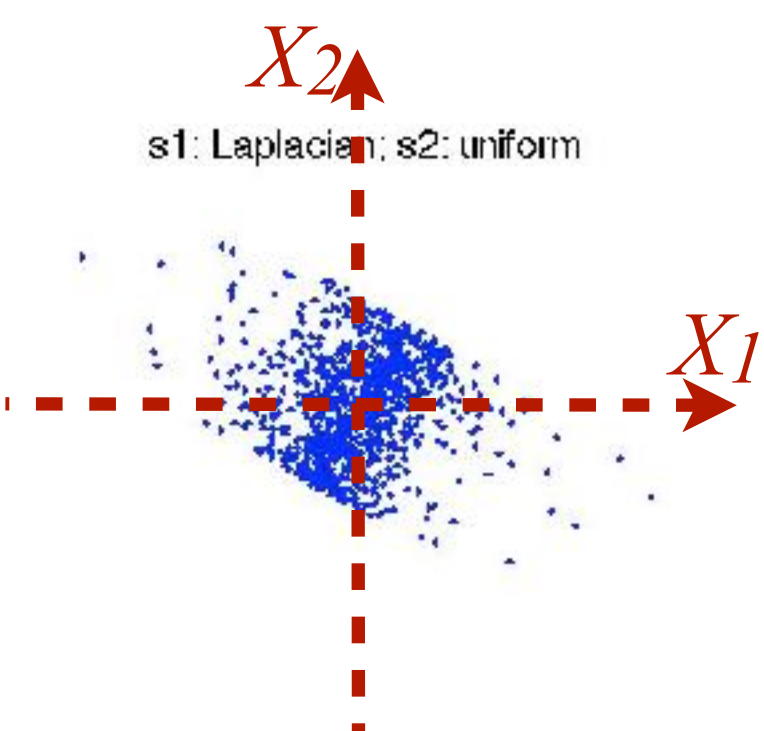
Intuition: Why ICA works?

- (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...



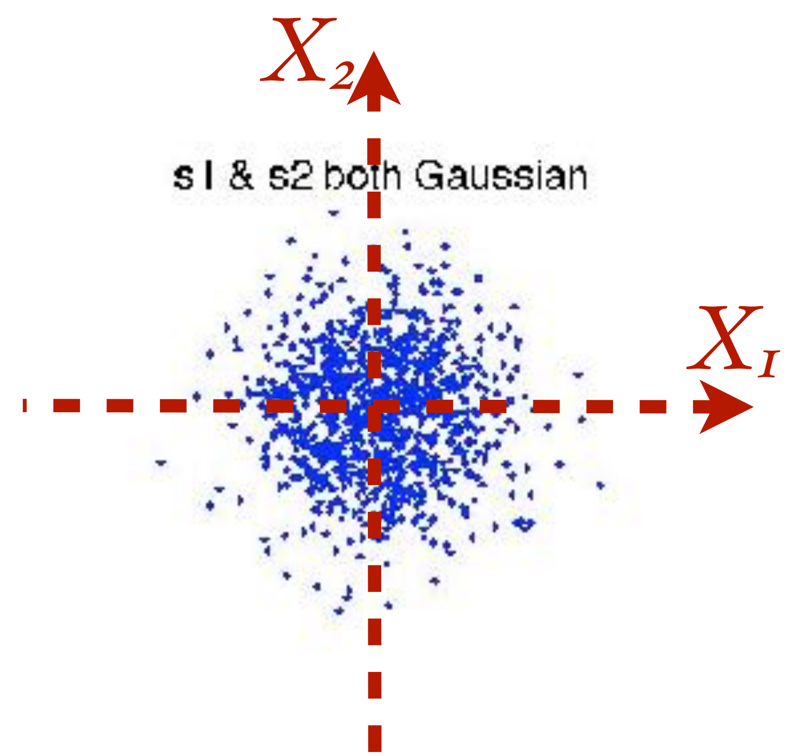


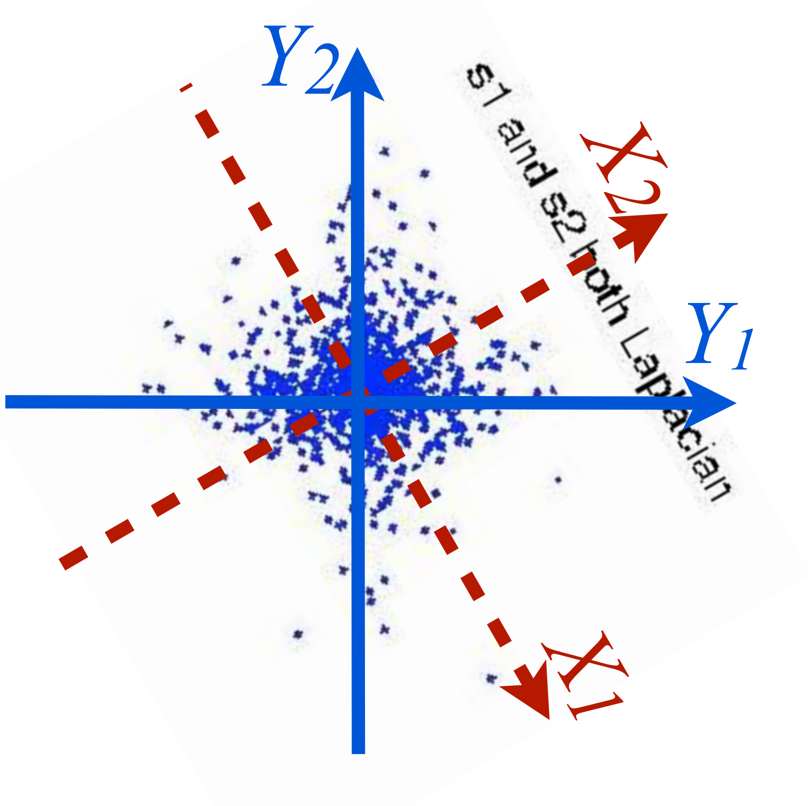


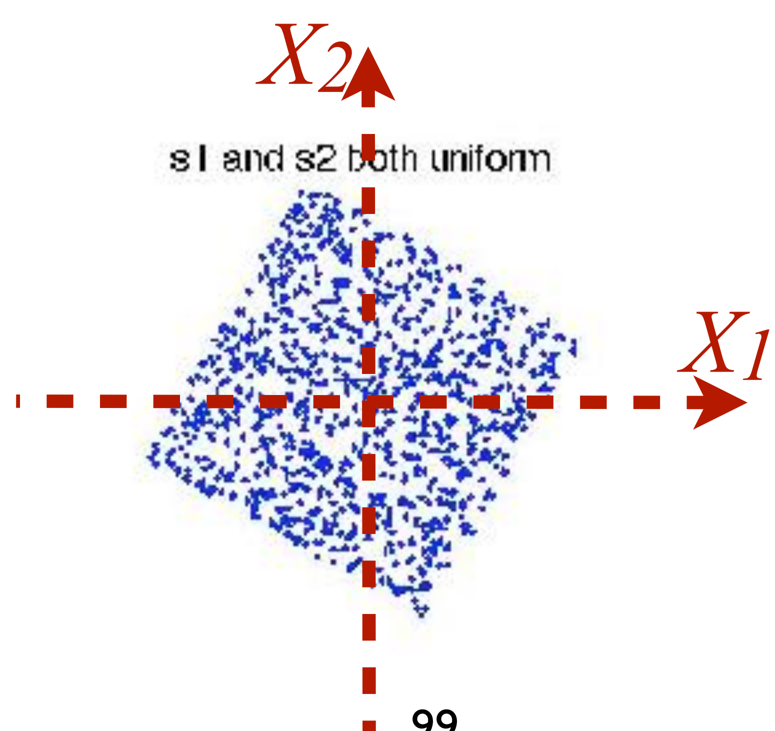


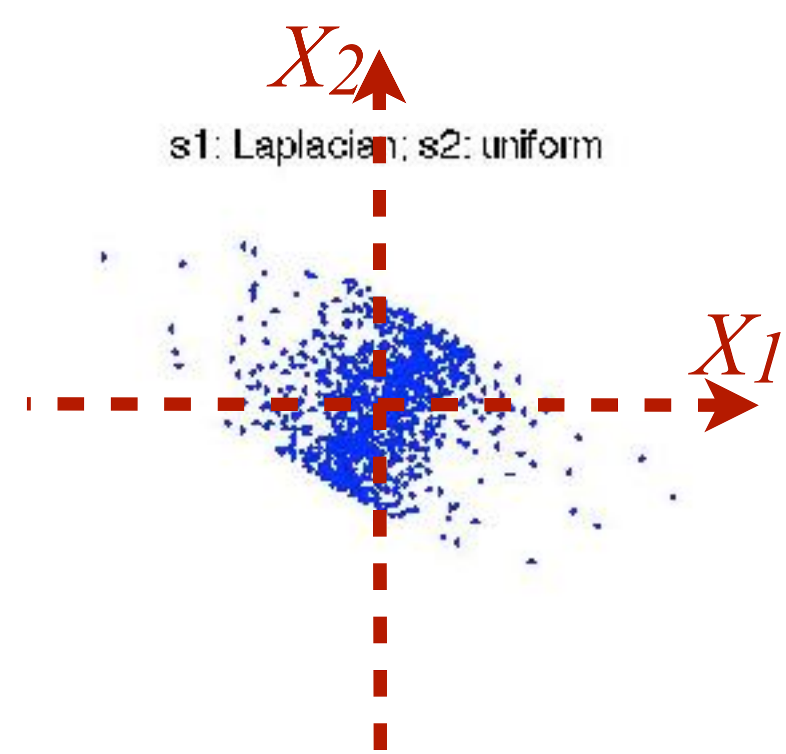
Intuition: Why ICA works?

- (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...



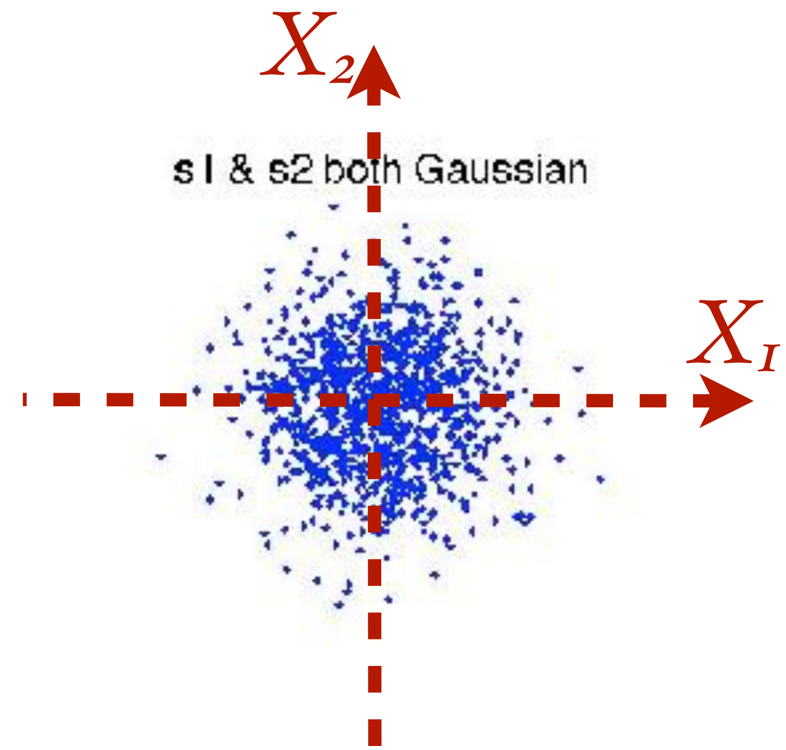


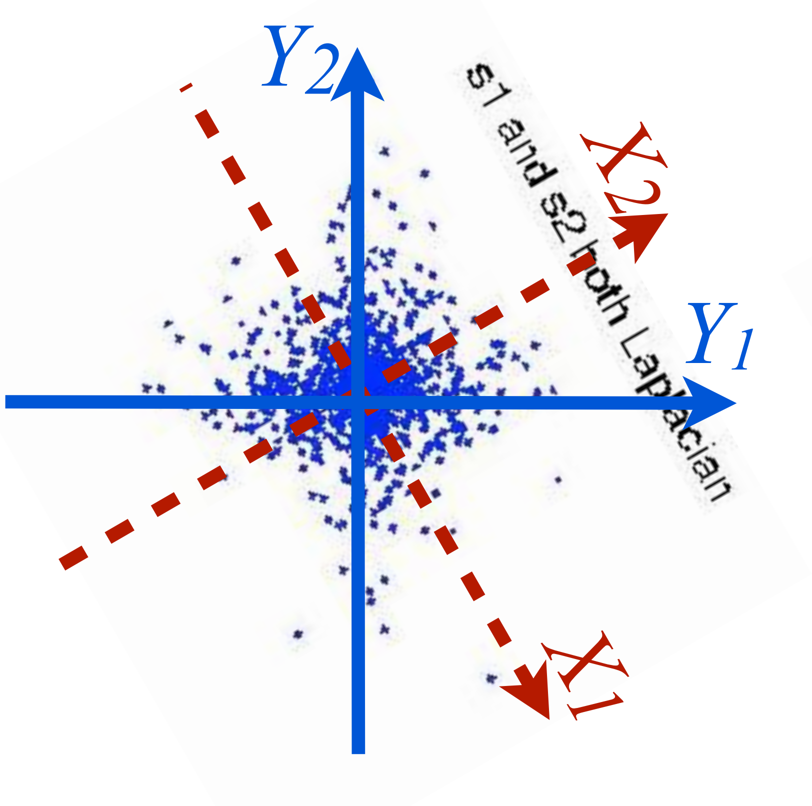


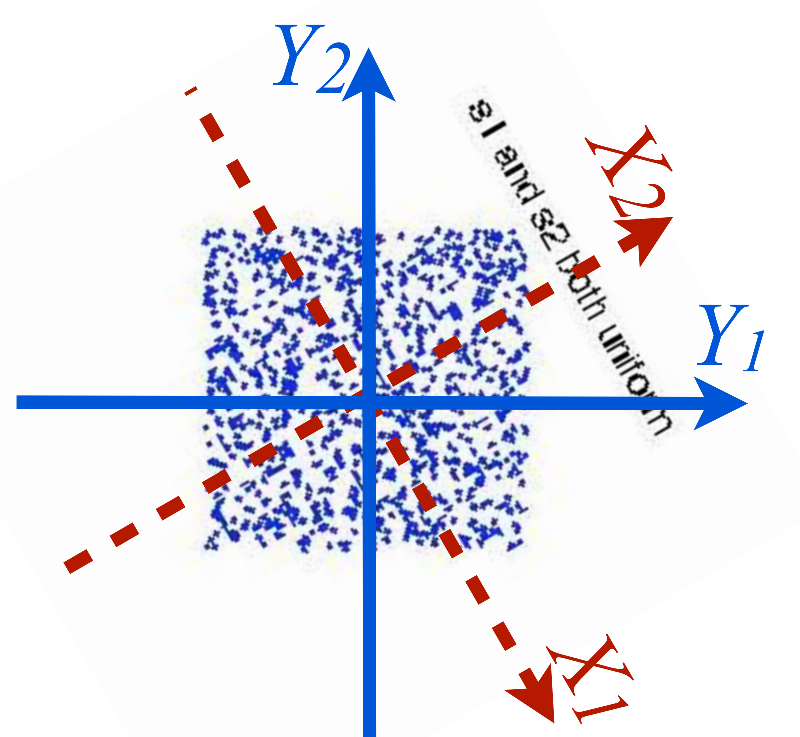


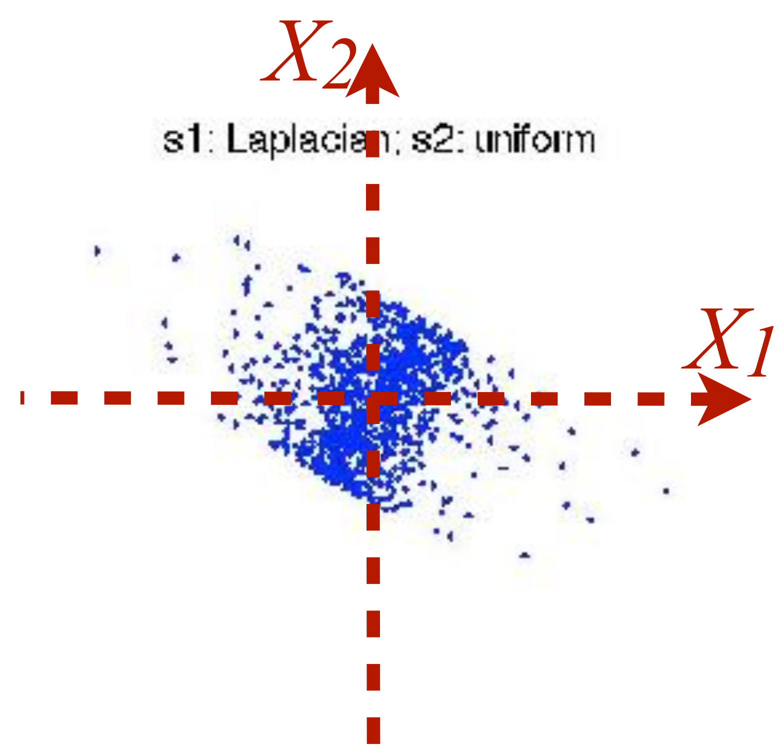
Intuition: Why ICA works?

- (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...



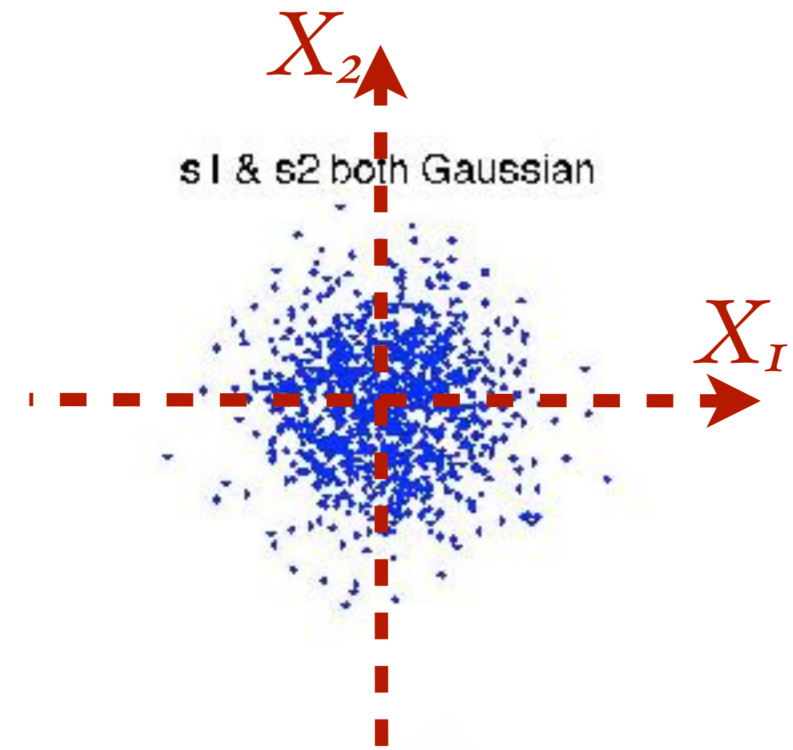


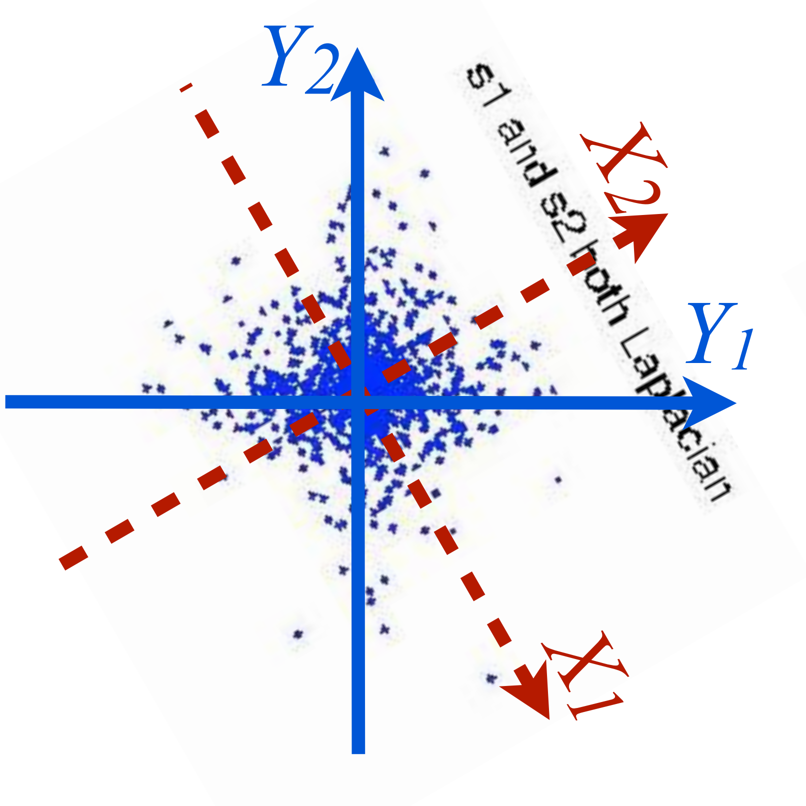


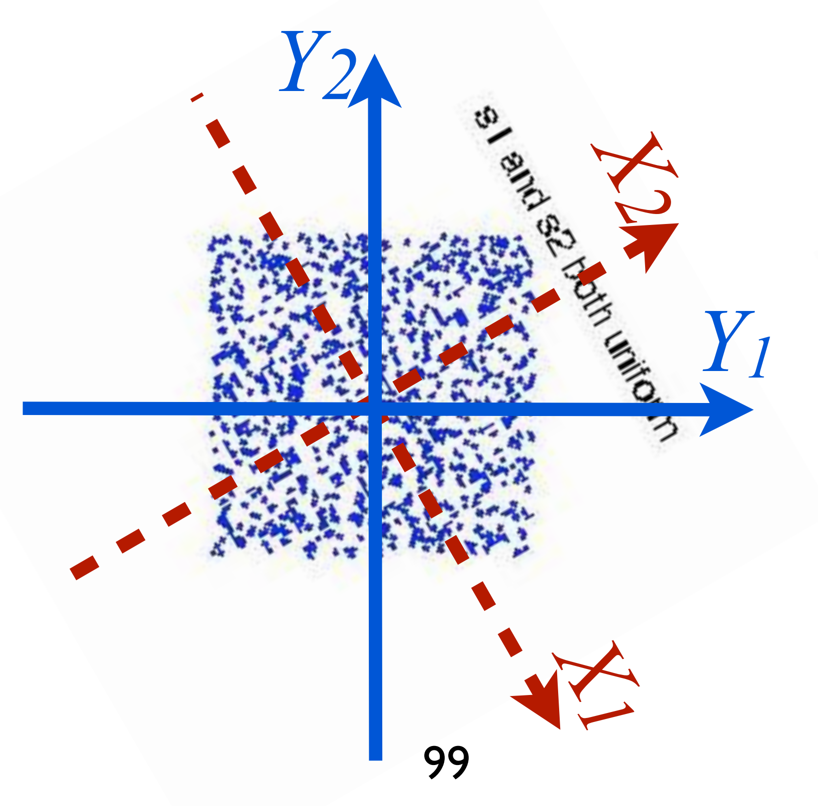


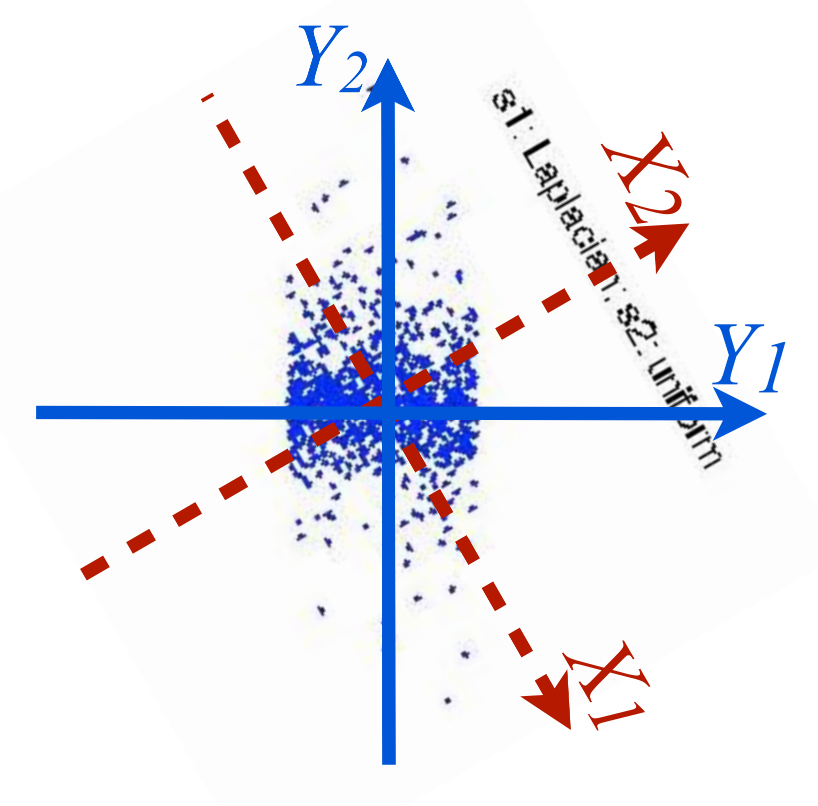
Intuition: Why ICA works?

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

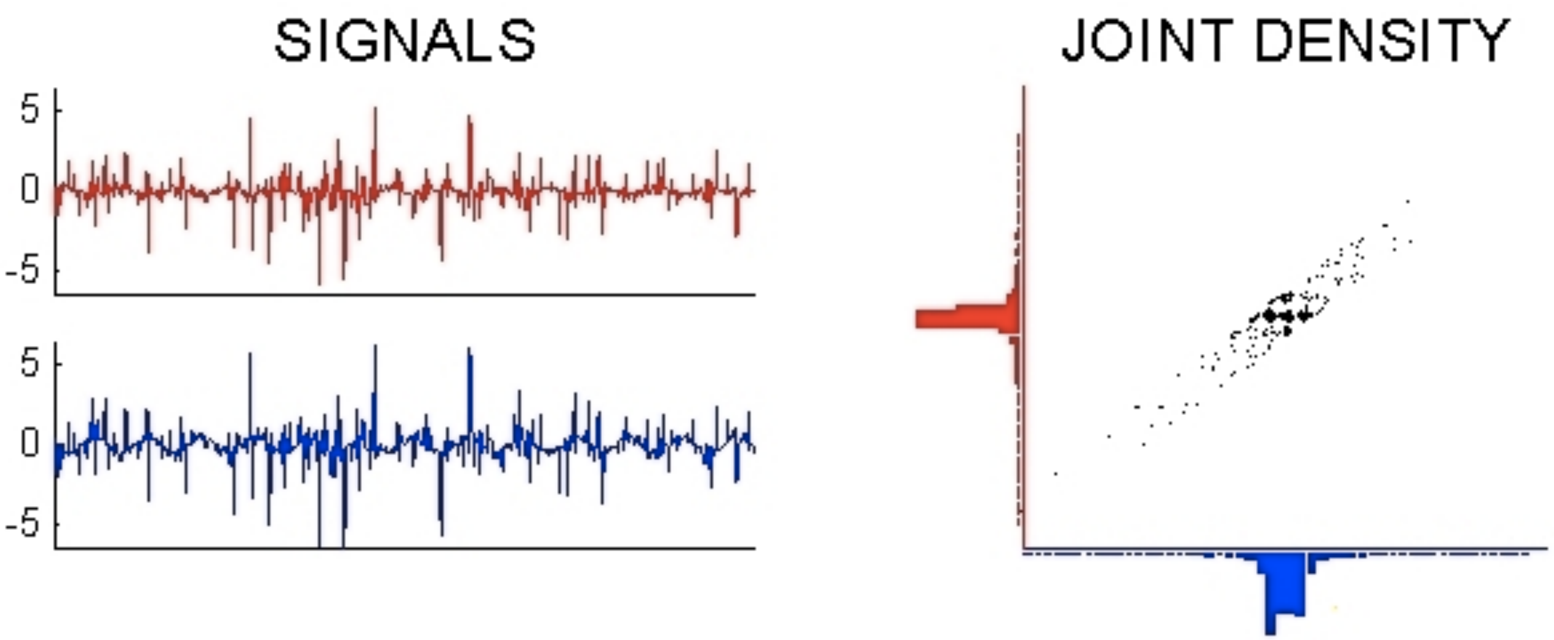




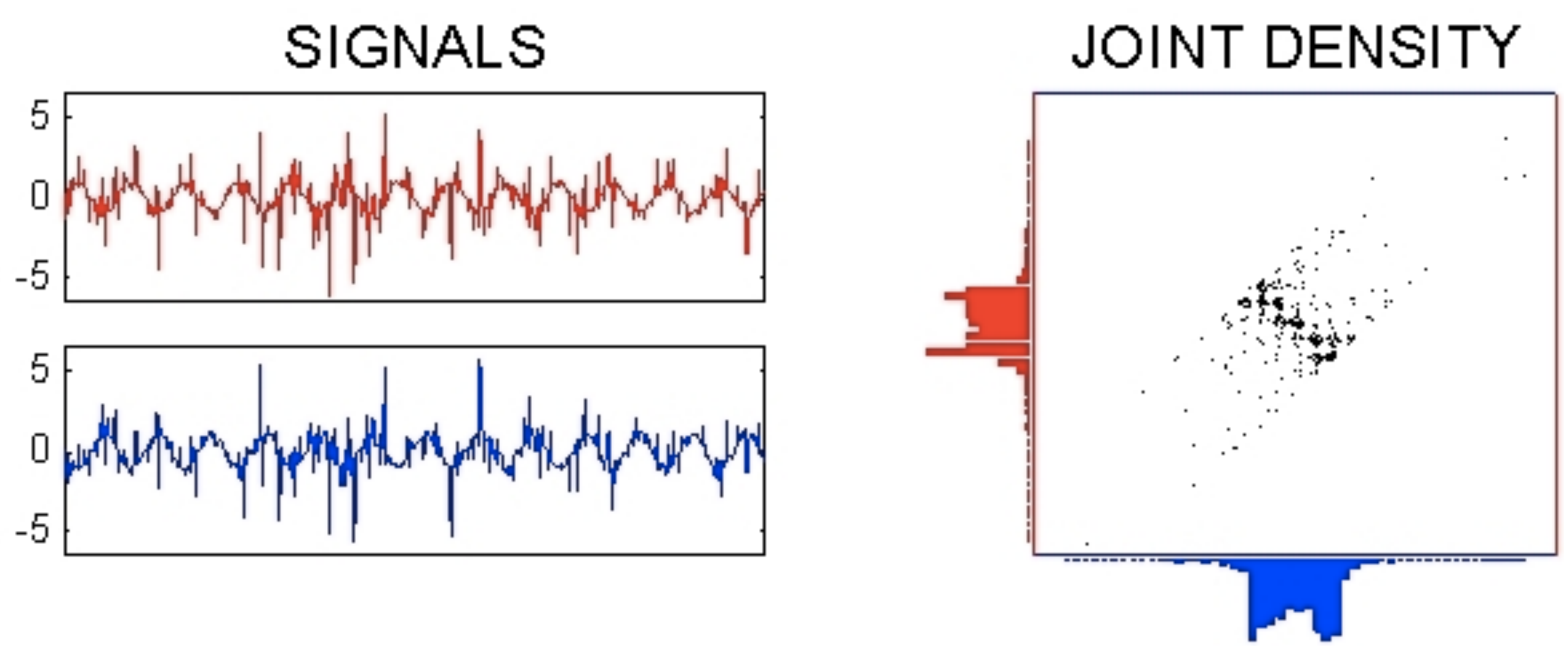




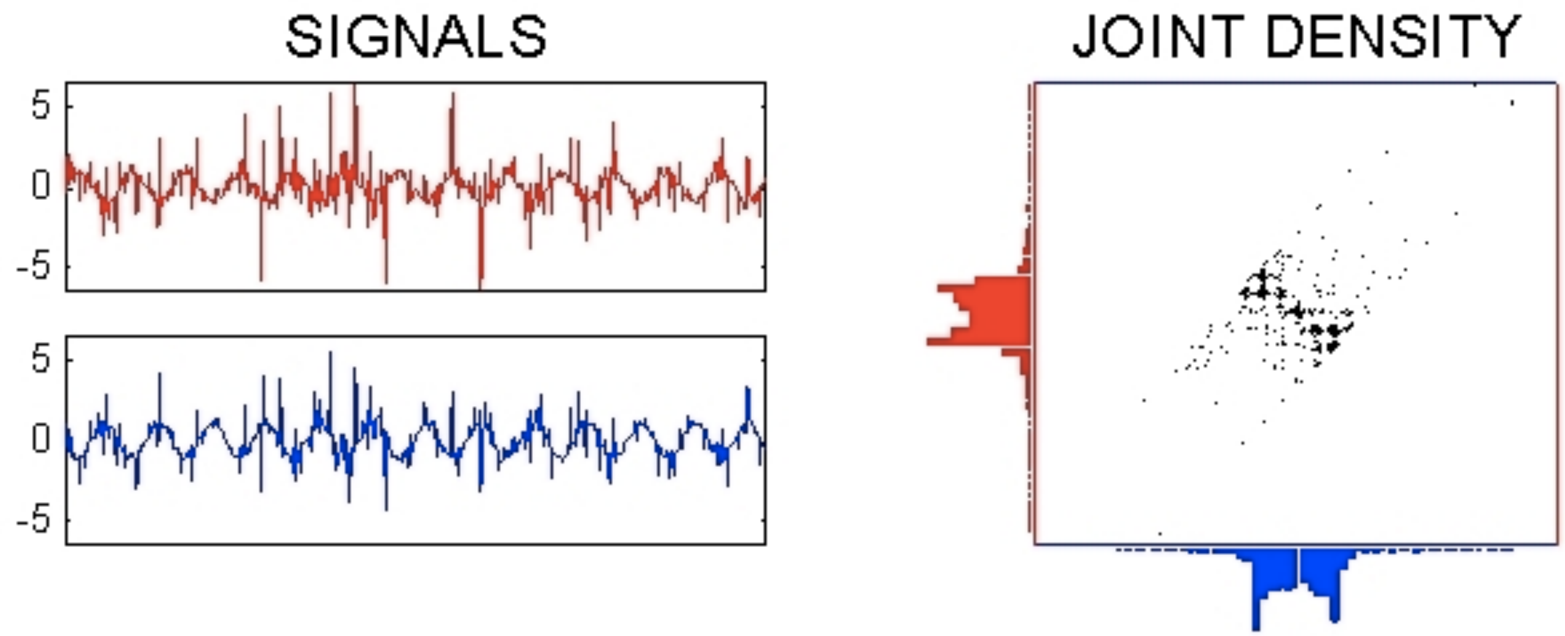
A Demo of the ICA Procedure



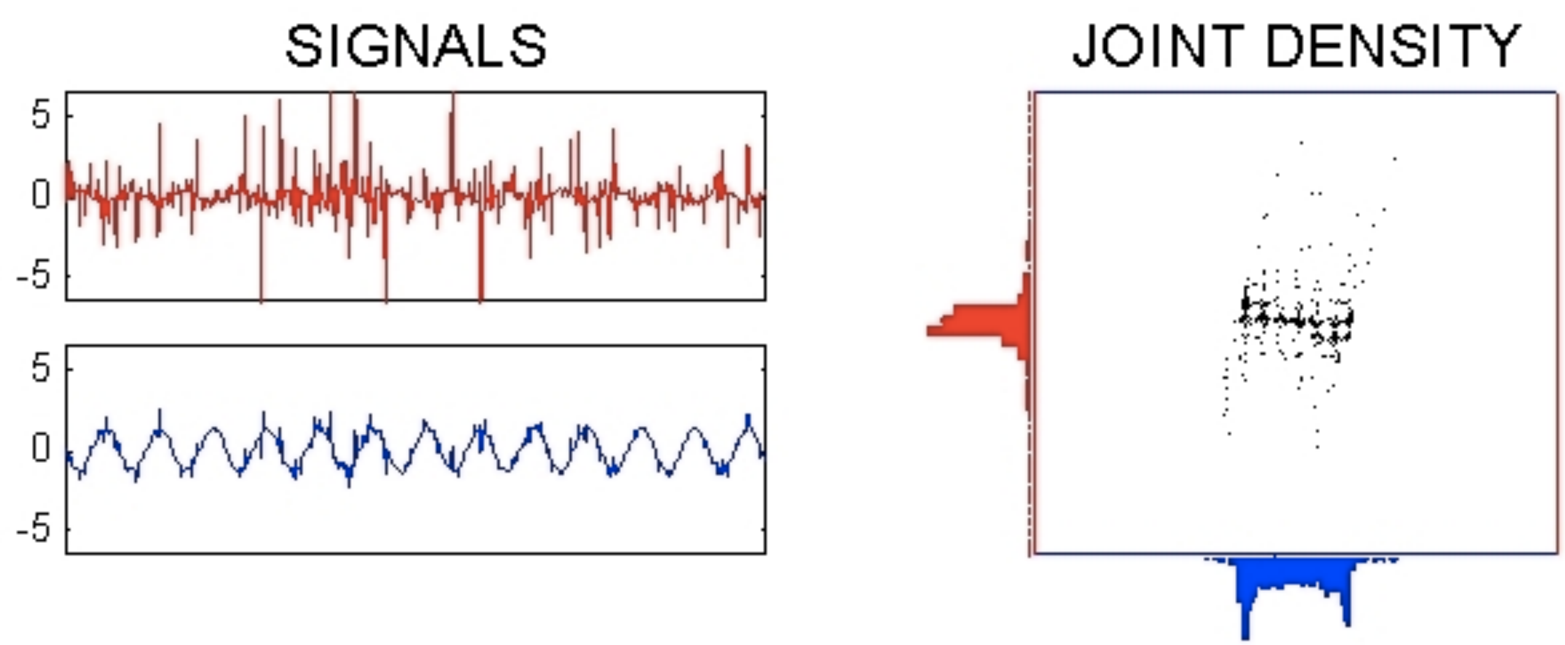
Input signals and density



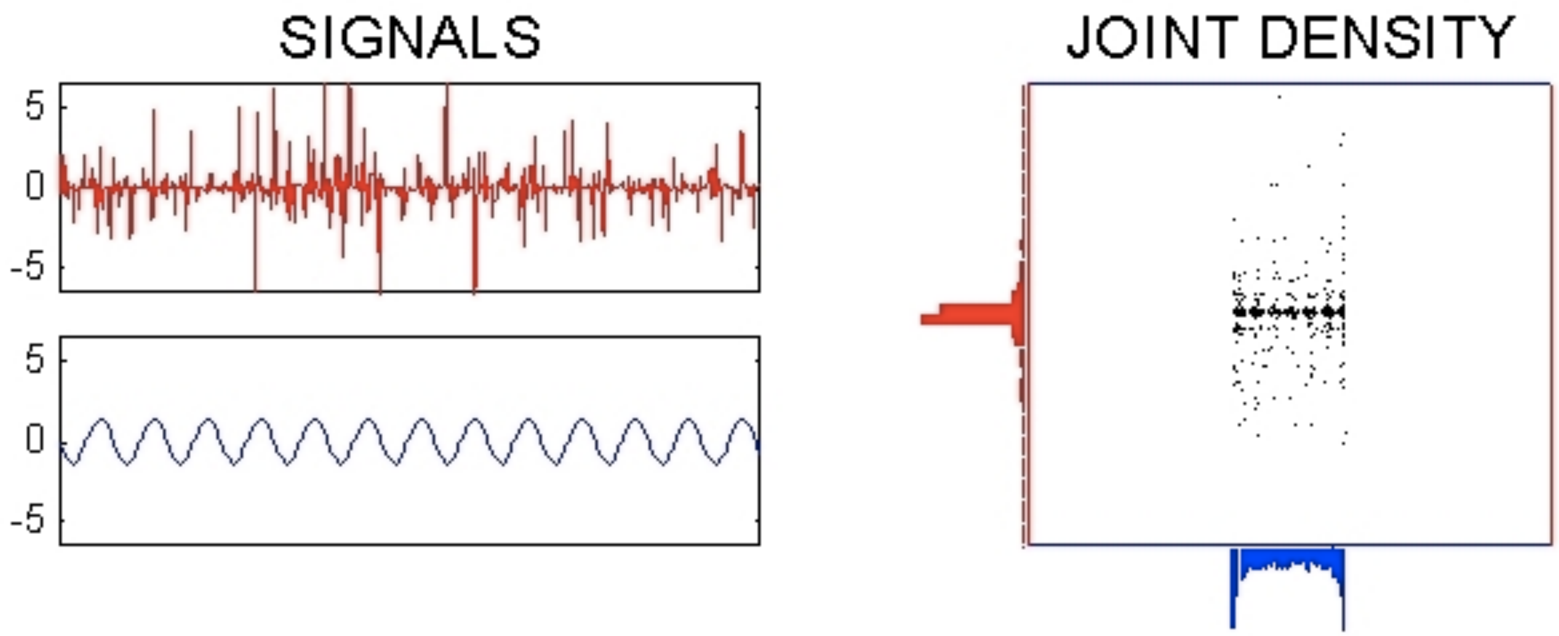
Whitened signals and density



Separated signals after 1 step of FastICA



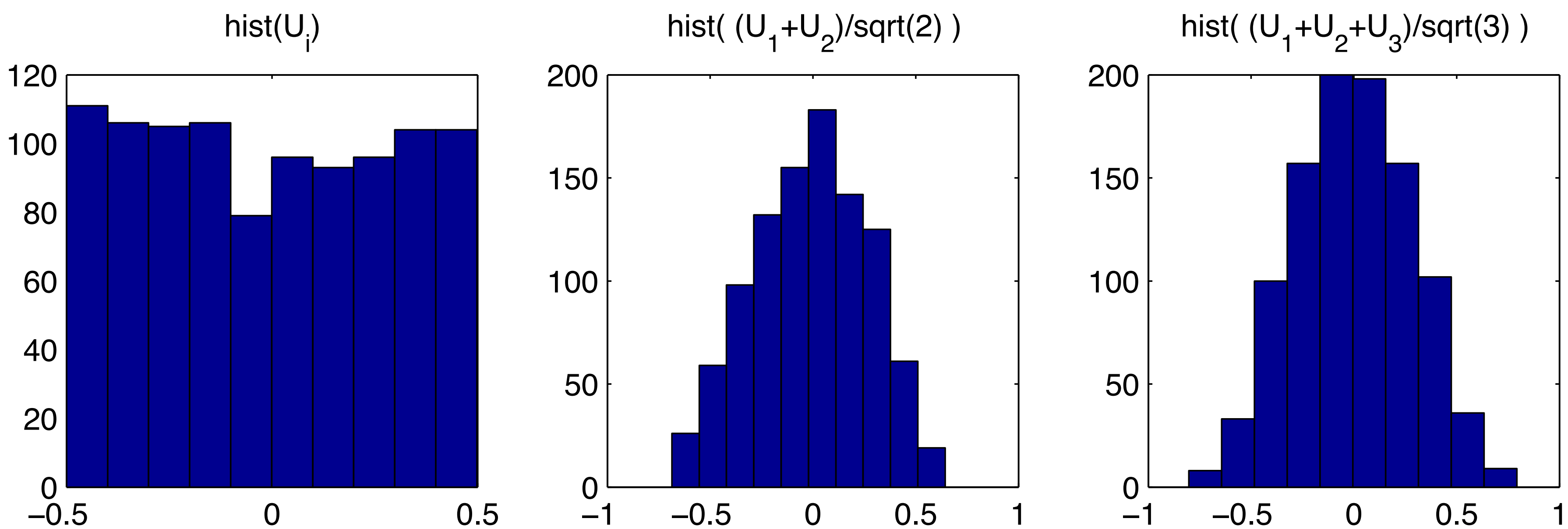
Separated signals after 3 steps of FastICA



Separated signals after 5 steps of FastICA

Why Gaussianity Was Widely Used?

- Central limit theorem: An illustration



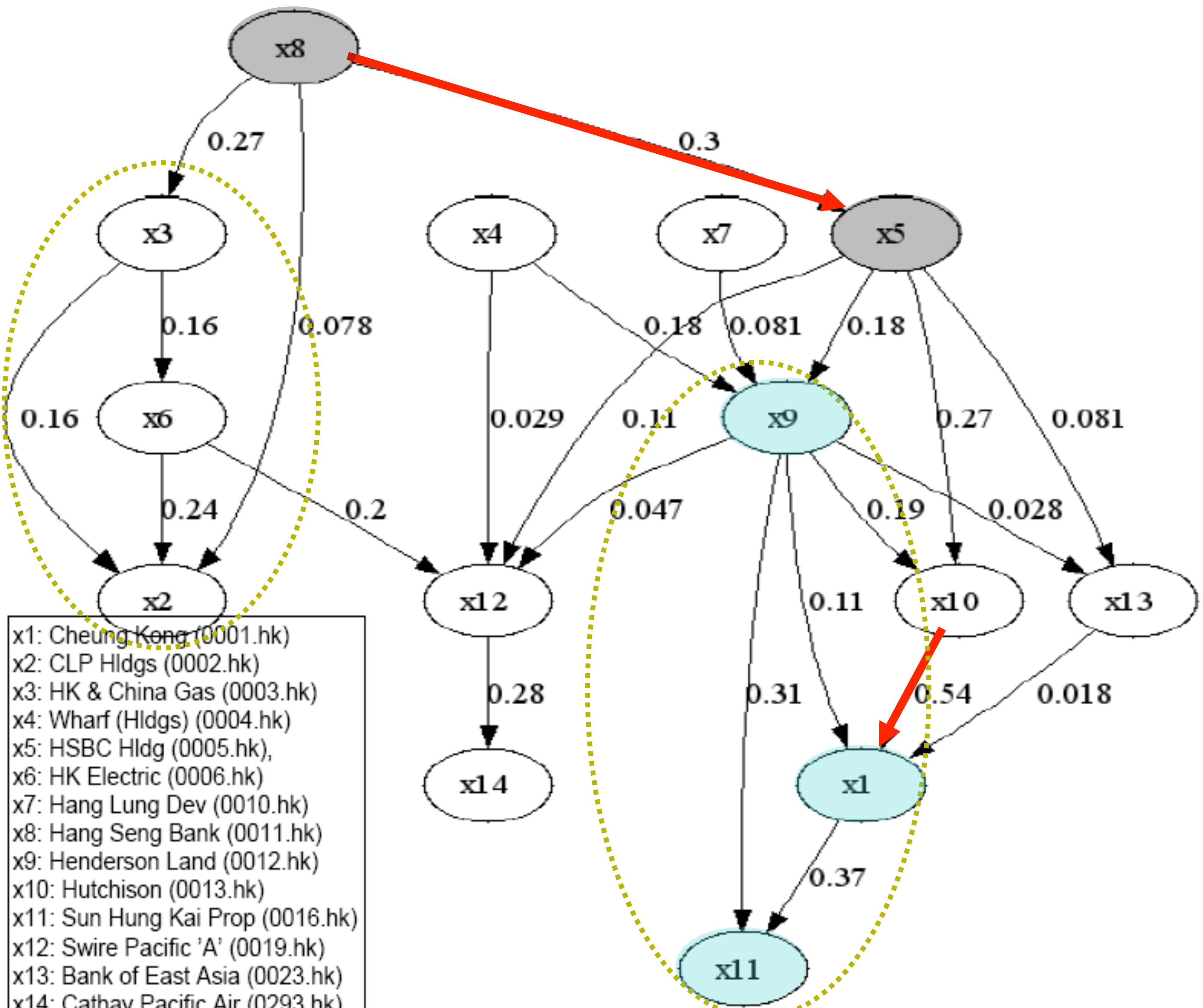
- “Simplicity” of the form; completely characterized by mean and covariance
- Marginal and conditionals are also Gaussian
- Has maximum entropy, given values of the mean and the covariance matrix

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

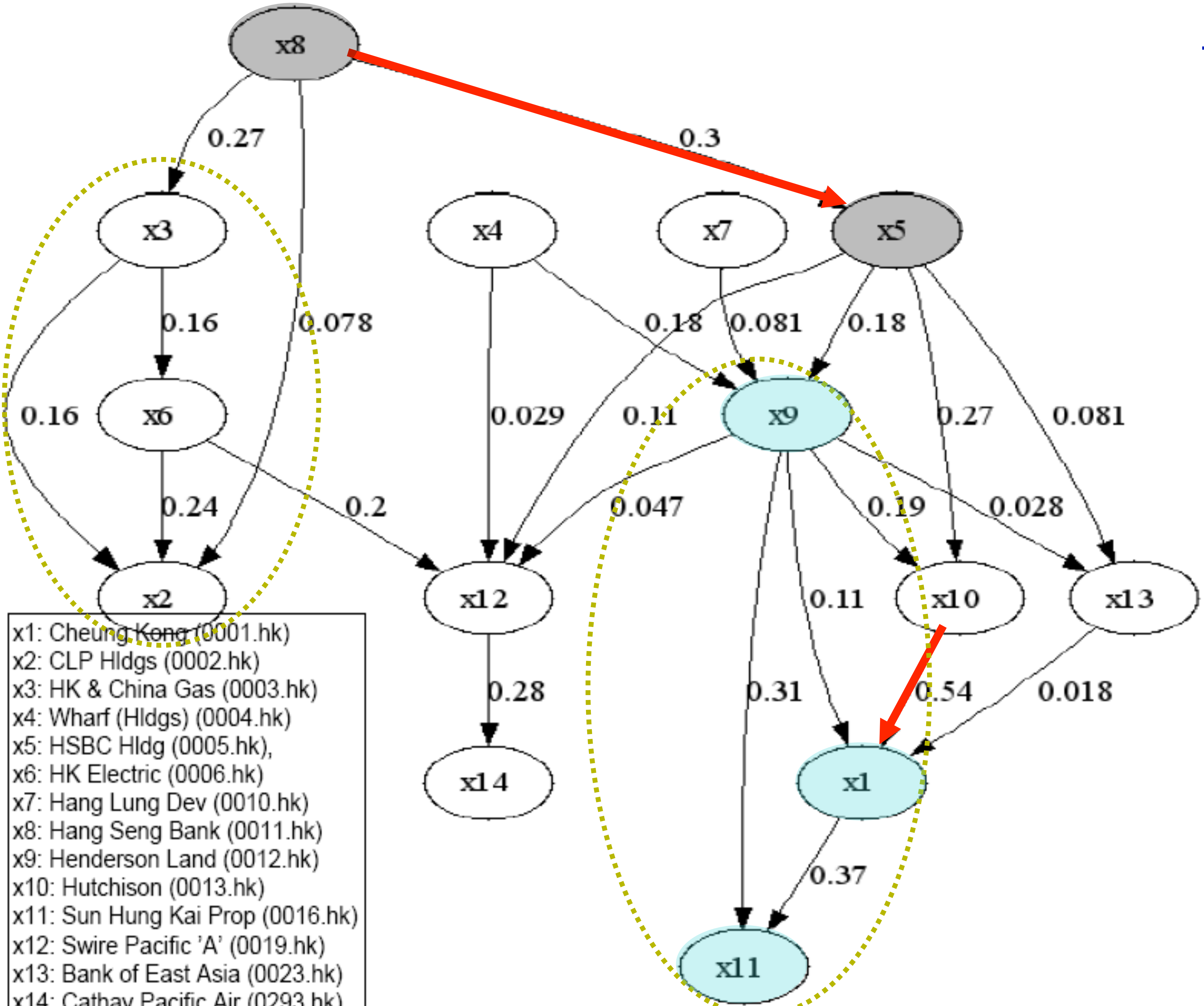
Gaussianity or Non-Gaussianity?

- 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?

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

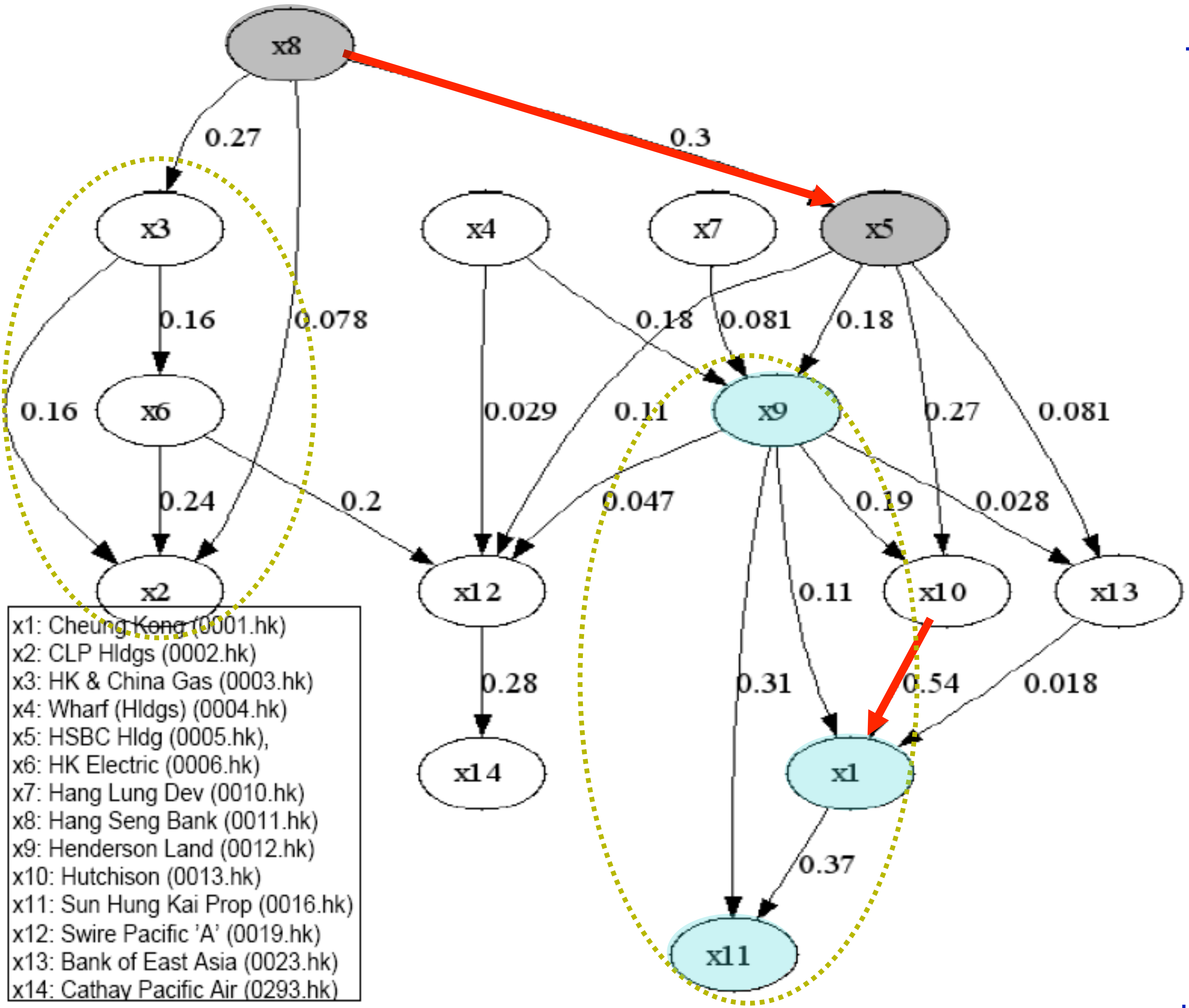


Application: Causal diagram in HK Stock Market (Zhang & Chan, 2006)



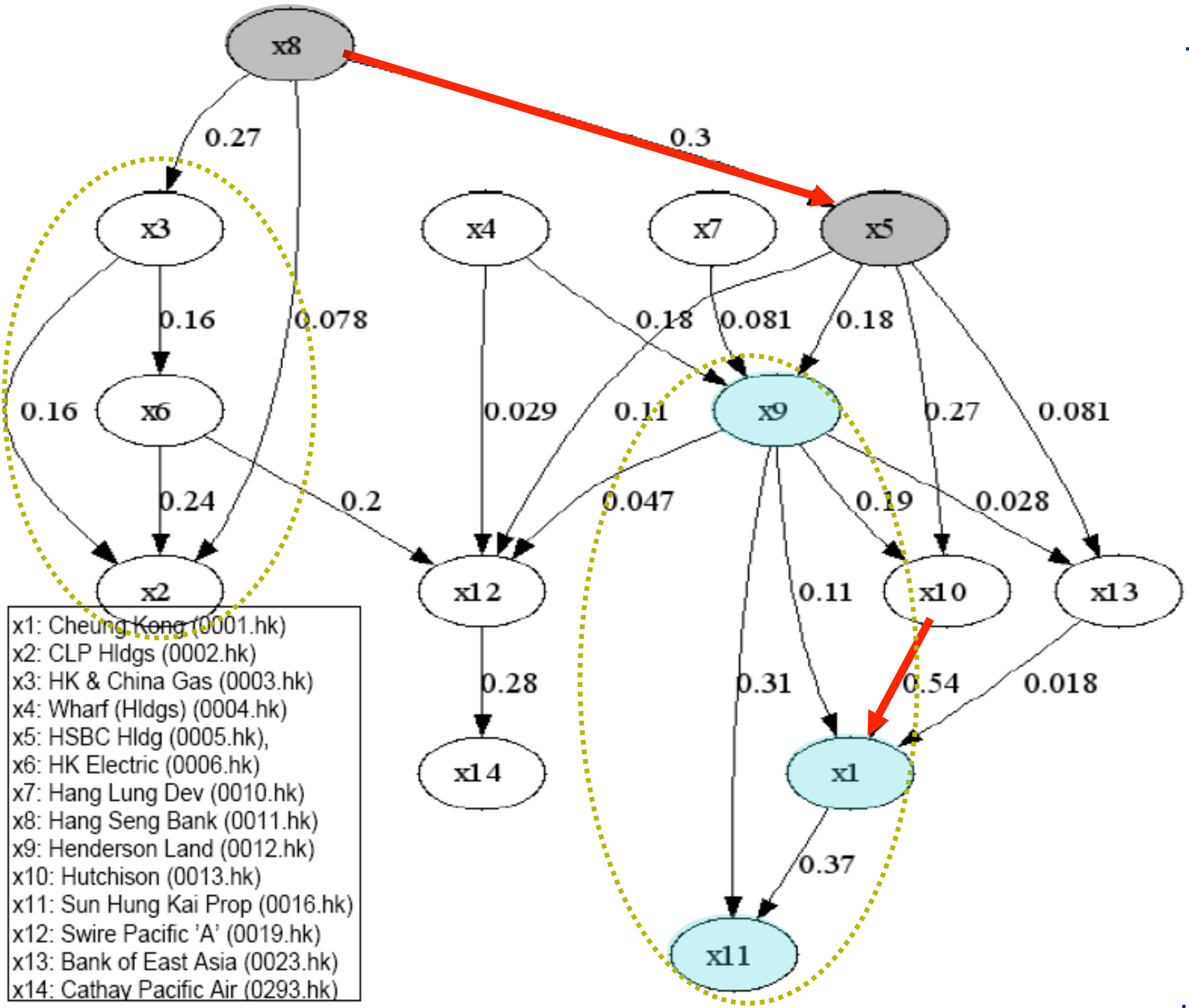
1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.

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



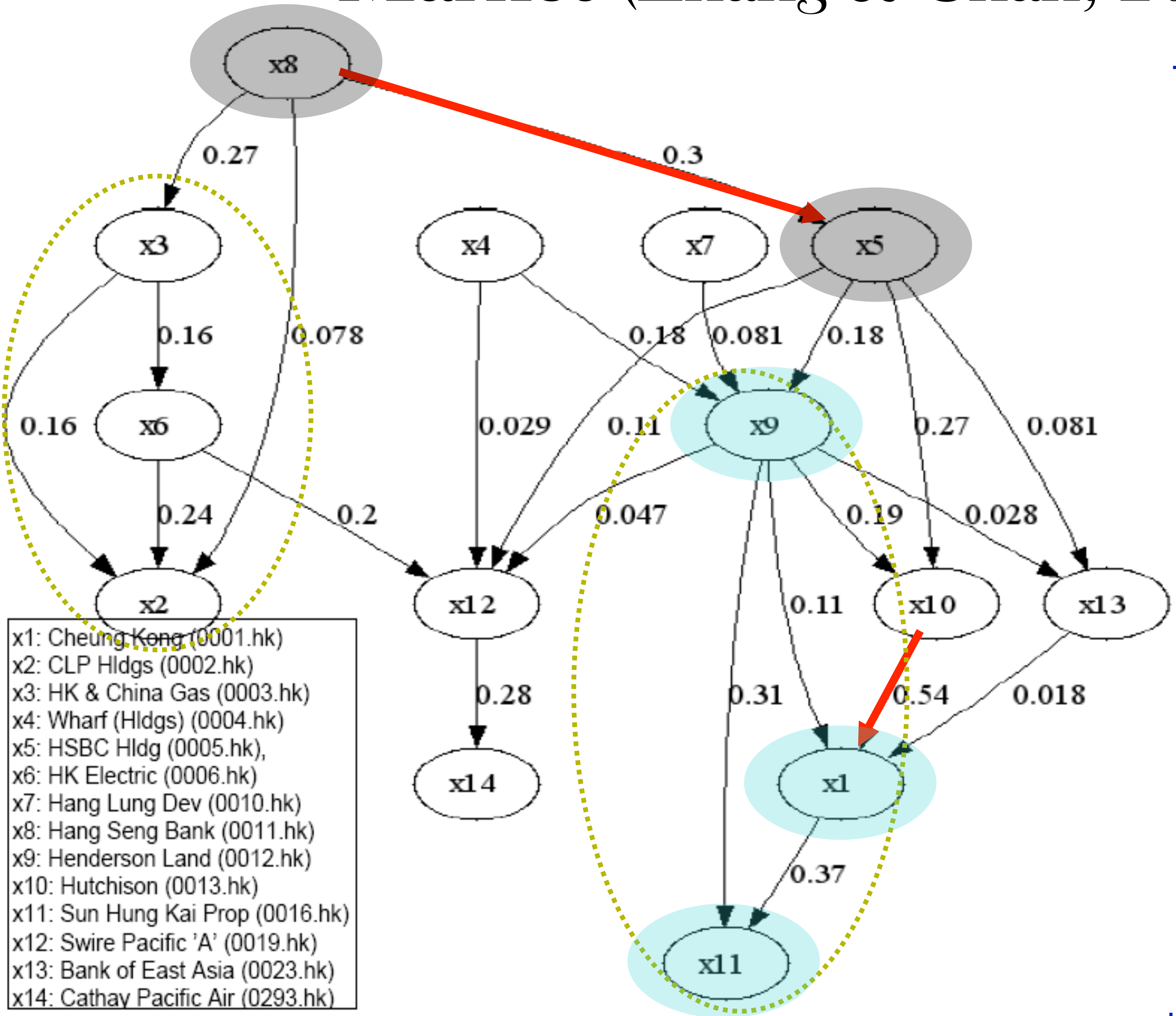
1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.
2. Stocks belonging to the same subindex tend to be connected.

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.

Causal Discovery 3:
Nonlinearity, confounding, missing data, confounding, time series...

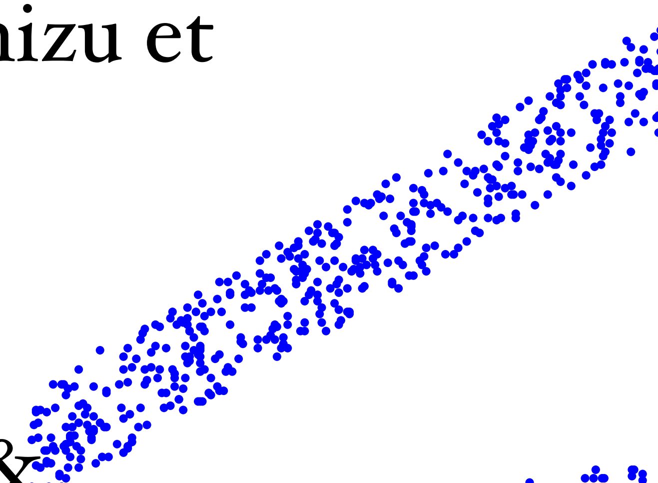
Practical Issues in Causal Discovery...

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)

FCMs with Which Causal Direction is Generally Identifiable

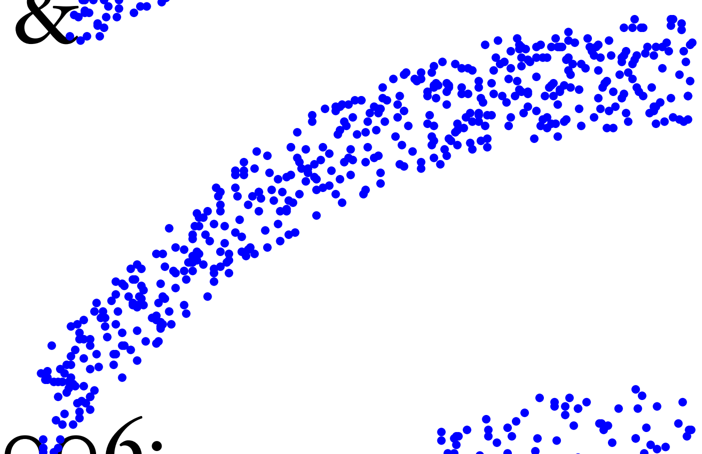
- Linear non-Gaussian acyclic causal model (Shimizu et al., '06)

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



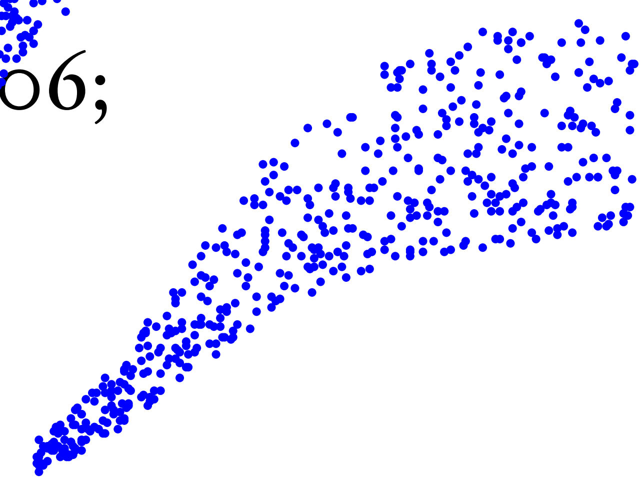
- Additive noise model (Hoyer et al., '09; Zhang & Hyvärinen, '09b)

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



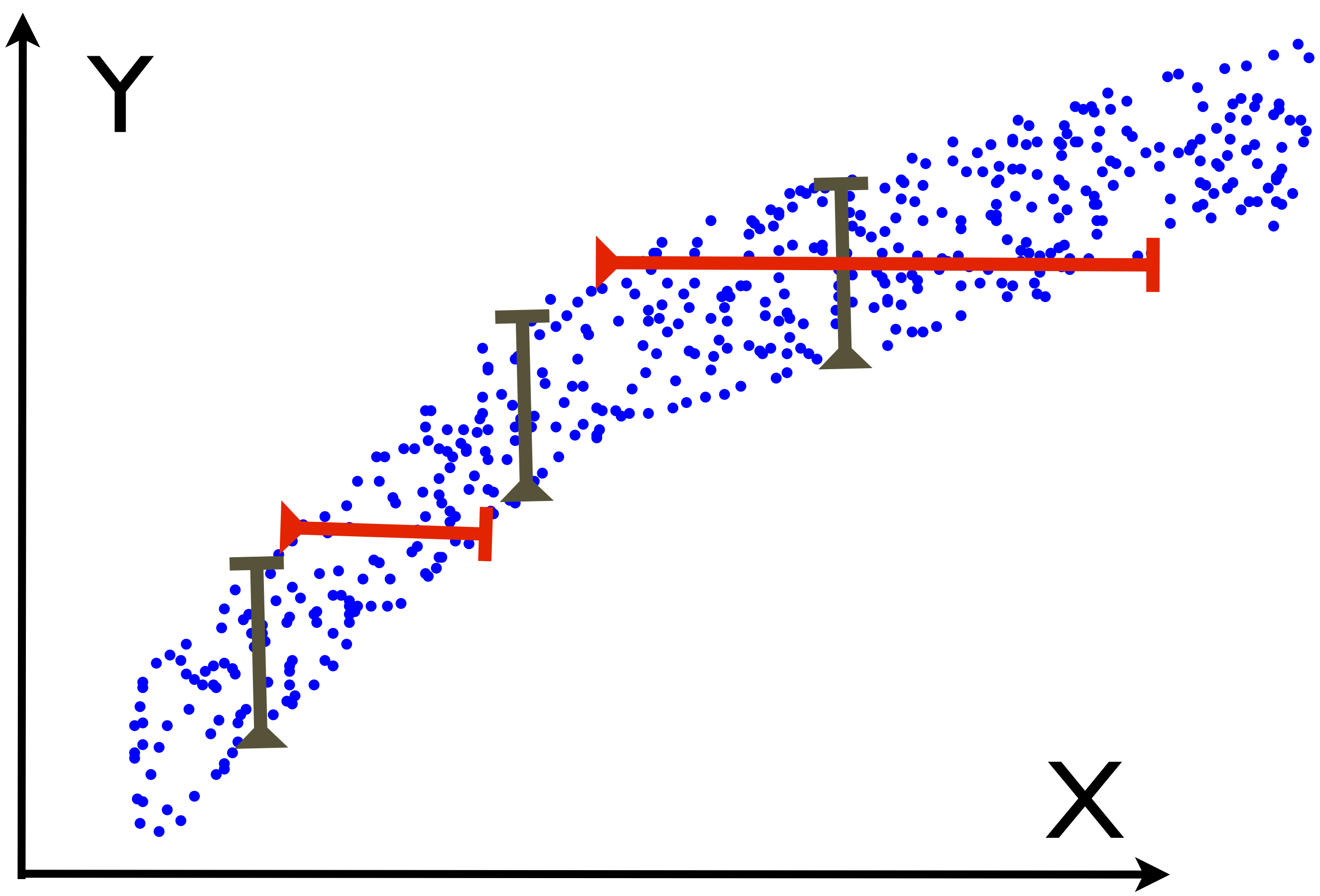
- Post-nonlinear causal model (Zhang & Chen, 2006; Zhang & Hyvärinen, '09a)

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



Causal Asymmetry with Nonlinear Additive Noise: Illustration

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

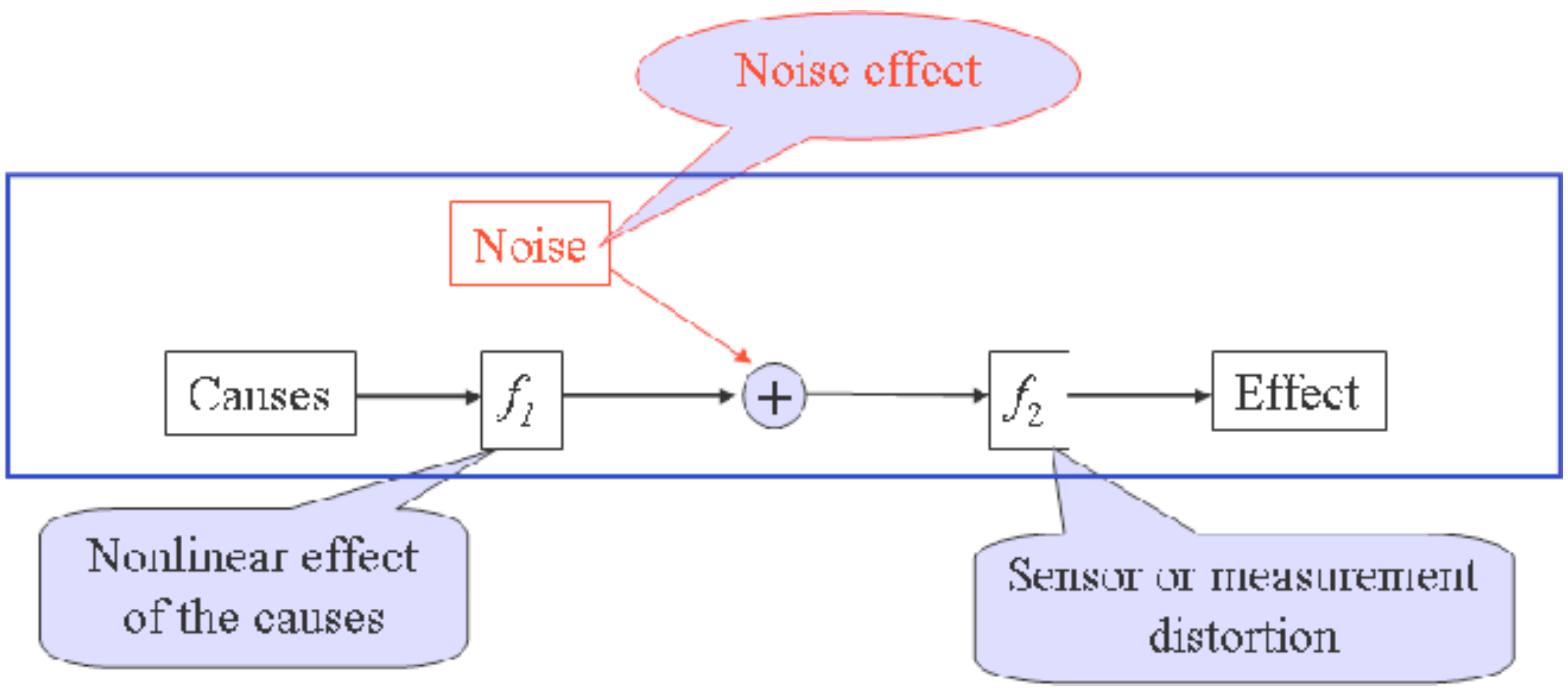


(Hoyer et al., 2009)

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

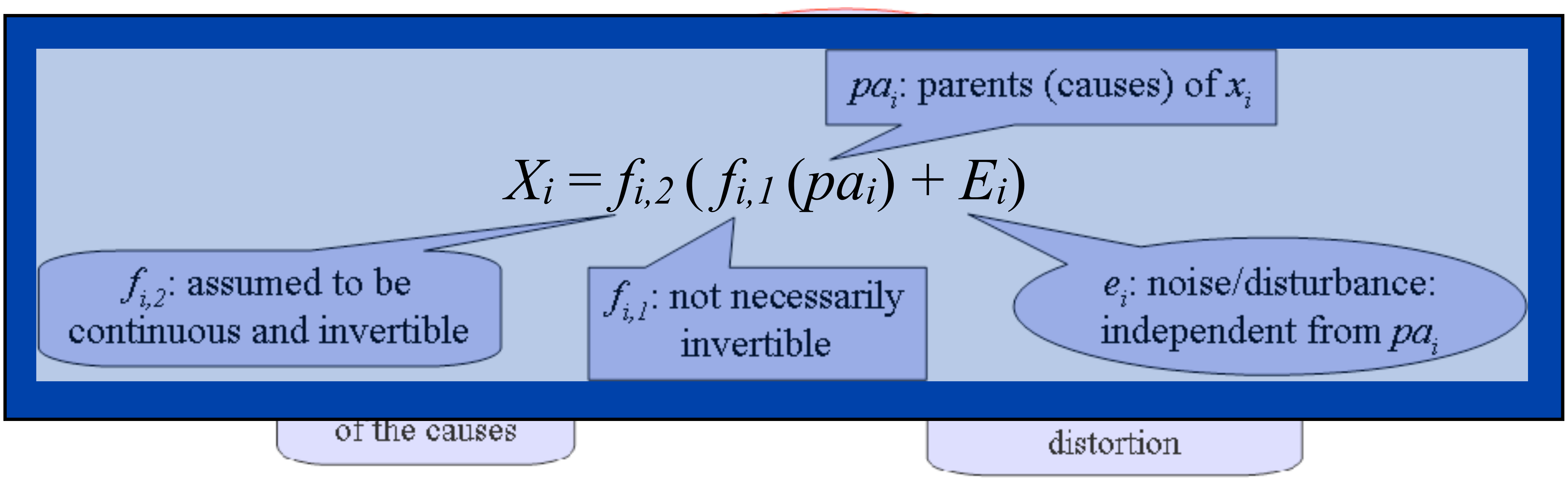


- Special cases: linear models; nonlinear additive noise models; multiplicative noise models: $Y = X \cdot E = \exp\big(\log(X) + \log(E)\big)$

Post-Nonlinear (PNL) Causal Model

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

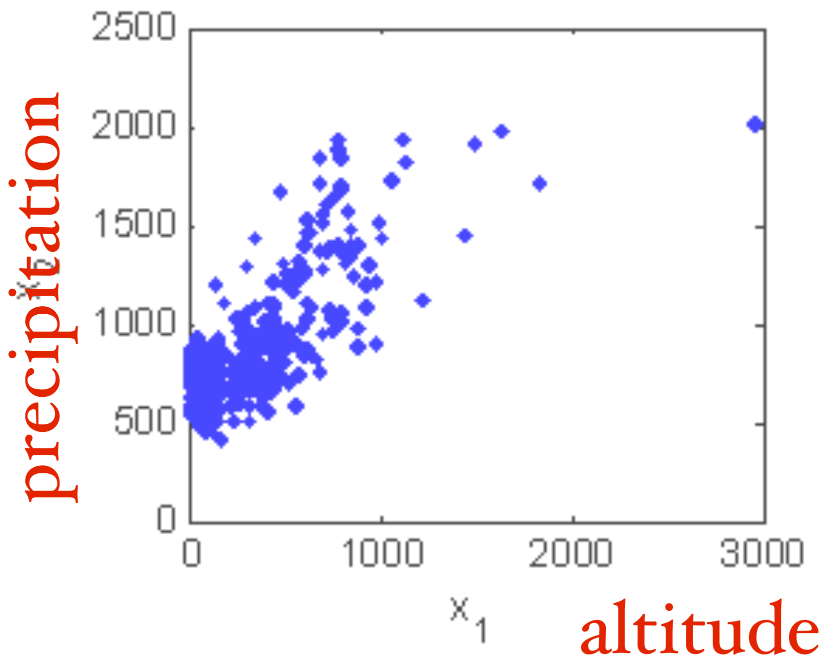
- Without prior knowledge, the assumed model is expected to be
  - general enough: adapt to approximate the true generating process
  - identifiable: asymmetry in causes and effects

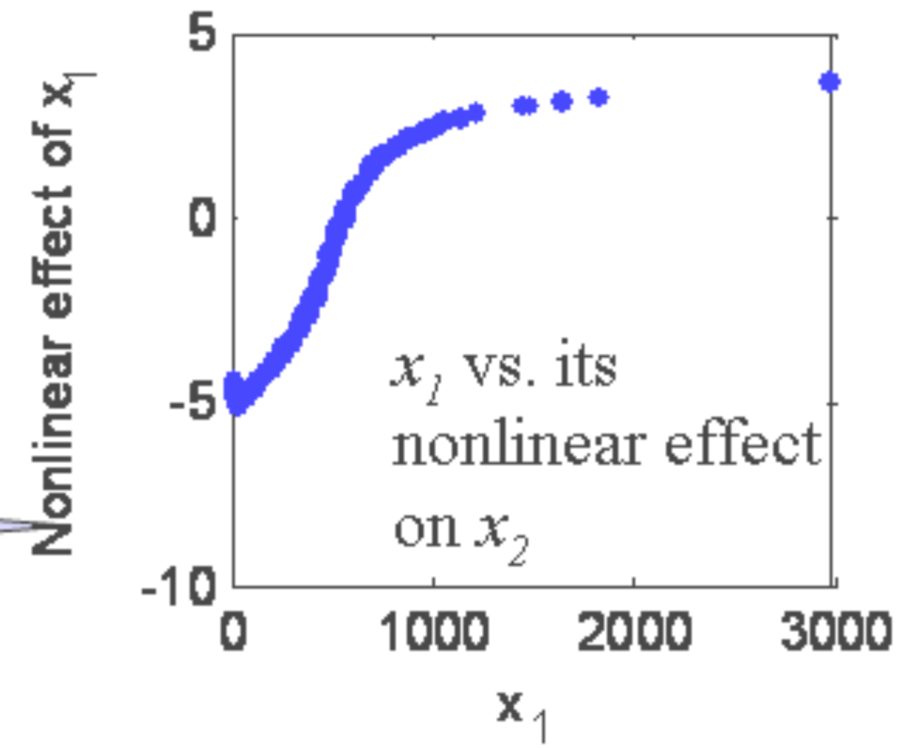


- Special cases: linear models; nonlinear additive noise models; multiplicative noise models: $Y = X \cdot E = \exp\big(\log(X) + \log(E)\big)$

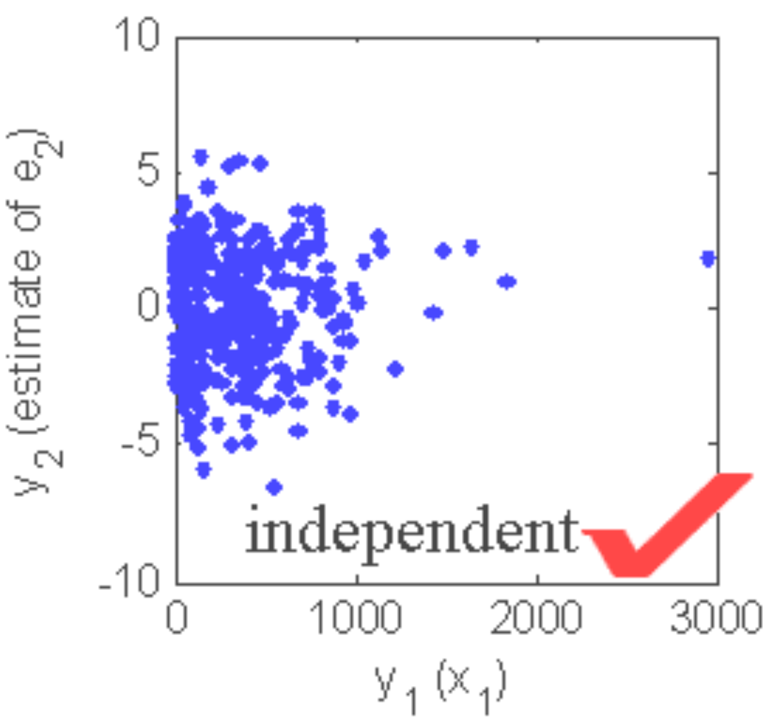
Data Set 2

with PNL Model

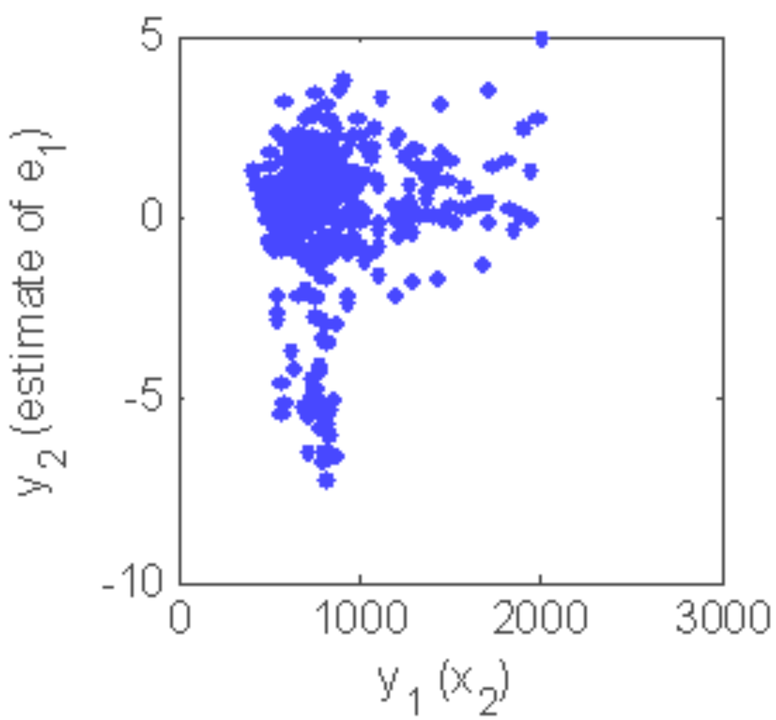


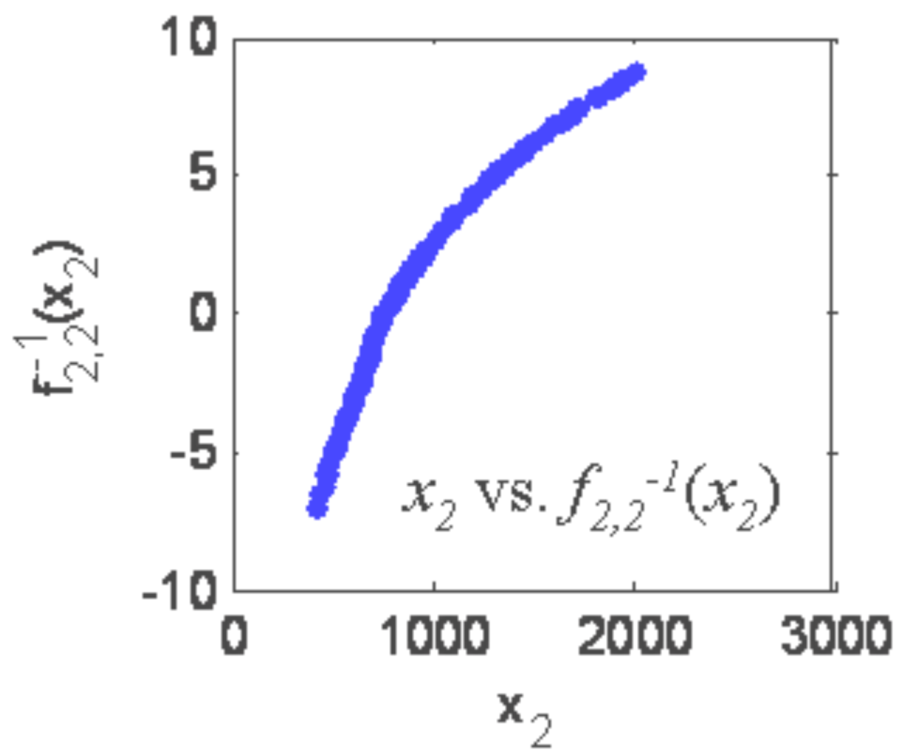


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

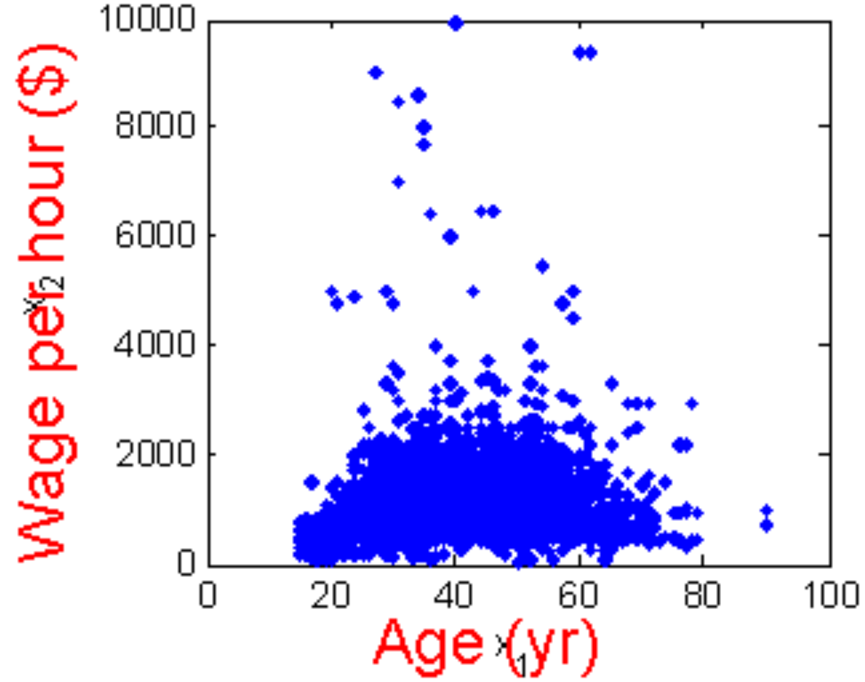


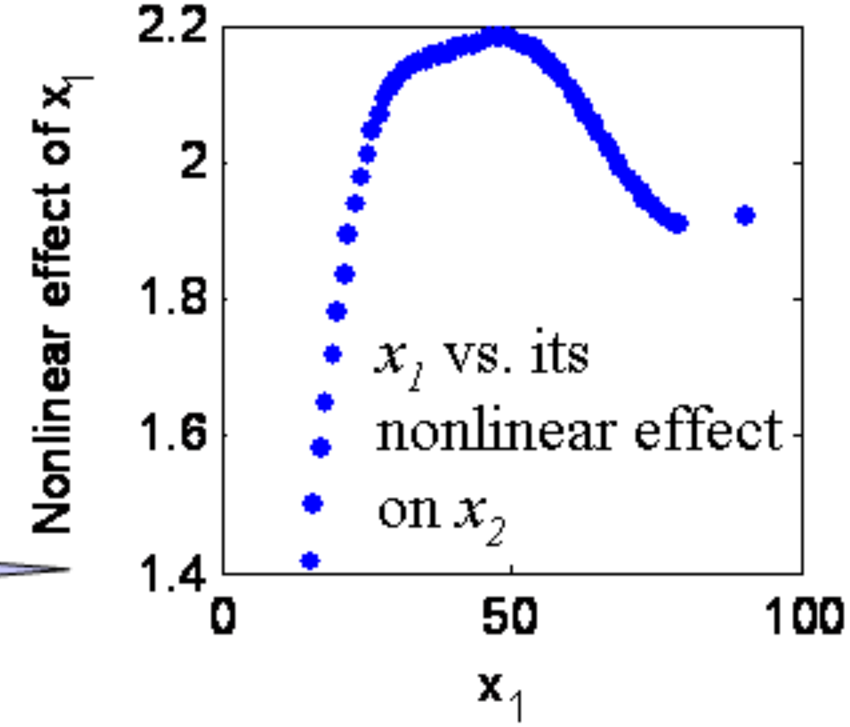
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$





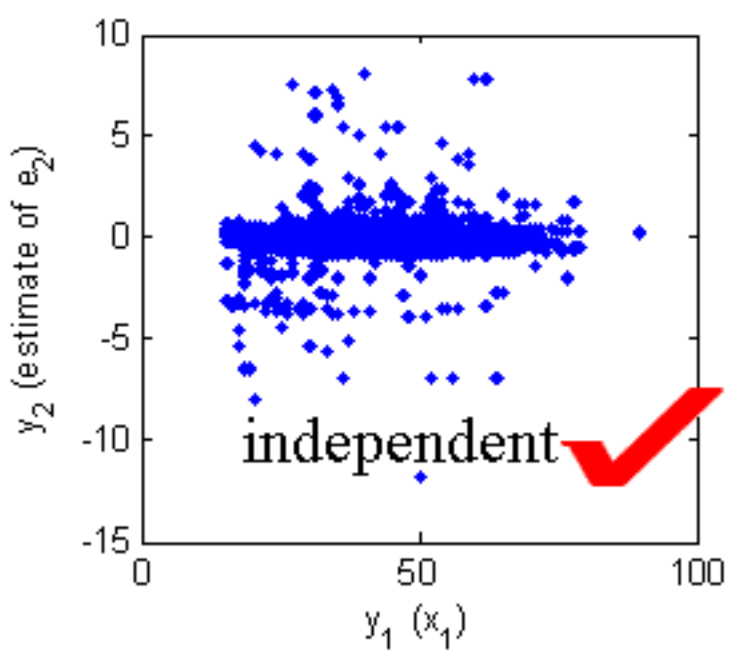
Data Set 8

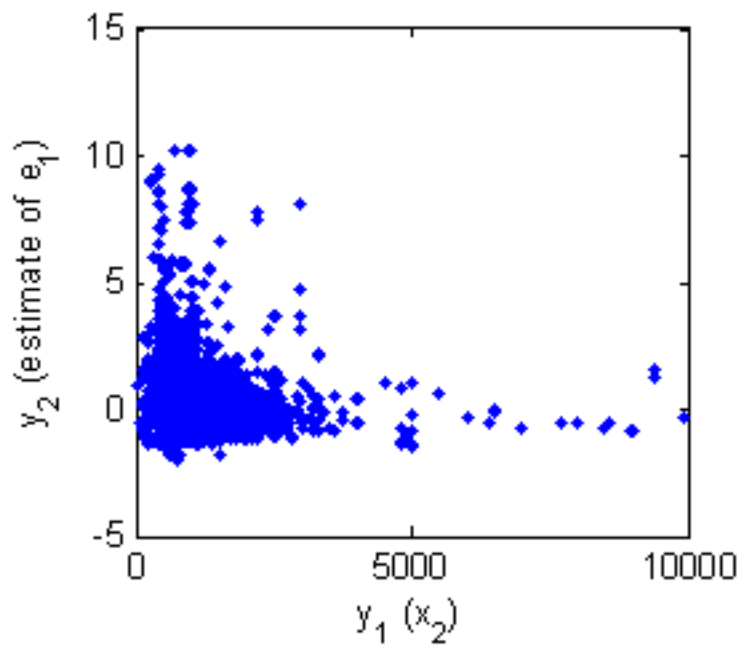


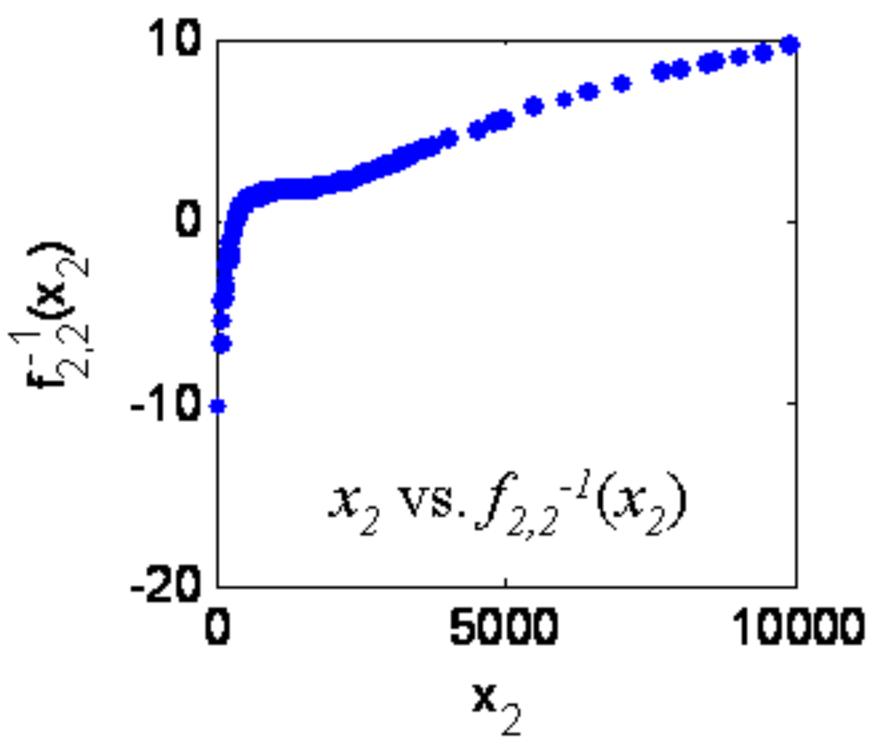


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

(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

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

$f_{i,2}$: 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
  - One can then find all non-identifiable cases

Identifiability: A Mathematical Result

- **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

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

Notation

$t_1 \triangleq g_2^{-1}(x_1), \quad z_2 \triangleq f_2^{-1}(x_2),$

$h \triangleq f_1 \circ g_2, \quad h_1 \triangleq g_1 \circ f_2.$

$\eta_1(t_1) \triangleq \log p_{t_1}(t_1), \quad \eta_2(e_2) \triangleq \log p_{e_2}(e_2).$

- 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…

All Non-Identifiable Cases

(Zhang and Hyvärinen, 2009)

$x_2 = f_2(f_1(x_1) + e_2)$

$x_1 = g_2(g_1(x_2) + e_1)$

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 $v \to +\infty$

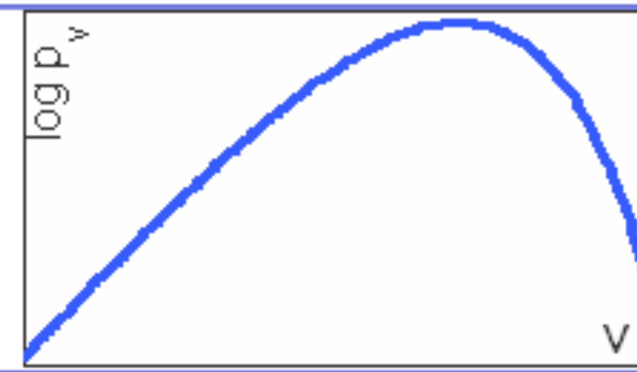


Table 1: All situations in which the PNL causal model is not identifiable.

	p_{e_2}	p_{t_1} $(t_1 = g_2^{-1}(x_1))$	$h = f_1 \circ g_2$	Remark
I	Gaussian	Gaussian	linear	h_1 also linear
II	log-mix-lin-exp	log-mix-lin-exp	linear	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	one-sided asymptotically exponential (but not log-mix-lin-exp)	h strictly monotonic, and $h' \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$	—
IV	log-mix-lin-exp	generalized mixture of two exponentials	Same as above	—
V	generalized mixture of two exponentials	two-sided asymptotically exponential	Same as above	—

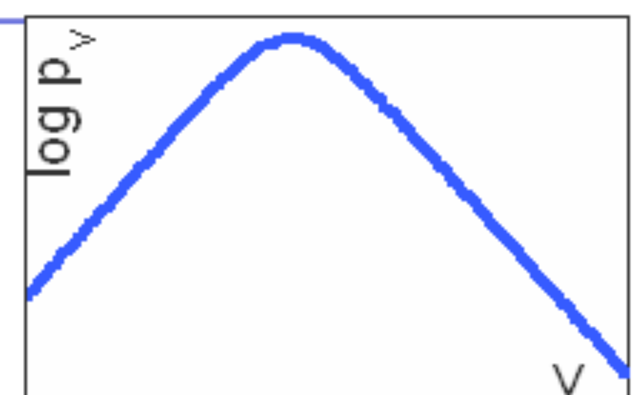
$p_v \propto (c_1 e^{c_2 v} + c_3 e^{c_4 v})^{c_5}$

$(\log p_v)' \to c_1 \ (c_1 \neq 0)$,

as $v \to -\infty$ and

$(\log p_v)' \to c_2 \ (c_2 \neq 0)$,

as $v \to +\infty$



All Non-Identifiable Cases

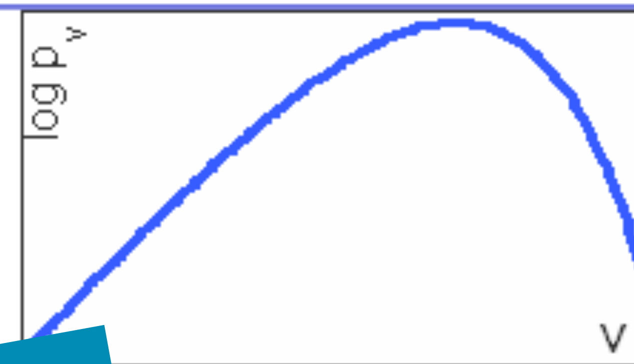
(Zhang and Hyvärinen, 2009)

$x_2 = f_2(f_1(x_1) + e_2)$

$x_1 = g_2(g_1(x_2) + e_1)$

$(\log p_v)' \to c \; (c \neq 0)$,

as $v \to -\infty$ or as $v \to +\infty$



Log-mixed-linear-and-exponential:

$\log p_v = c_1 e^{c_2 v} + c_3 v + c_4$

Table 1: All situations in whi... not identifiable.

	p_{e_2}	...	Remark
I	Gaussian		h_1 also linear
II	log-mix-lin-exp		h_1 strictly monotonic, and $h_1' \to$ ..., 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		

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

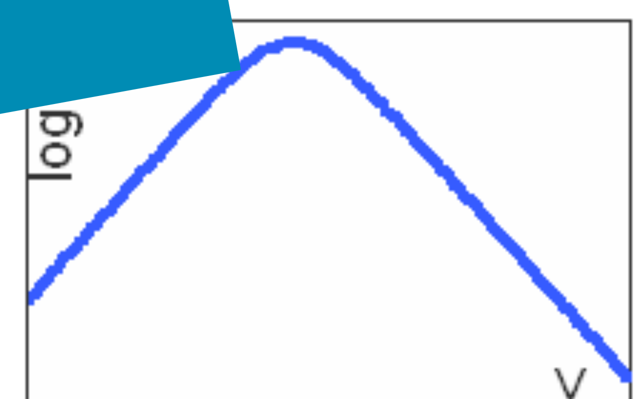
Causal direction is generally identifiable if the data were generated according to $X_2 = f_2(f_1(X_1) + E)$.

Special cases: $X_2 = a \cdot X_1 + E$ and $X_2 = g(X) + E$.

... and

$(\log p_v)' \to c_2 \; (c_2 \neq 0)$,

as $v \to +\infty$

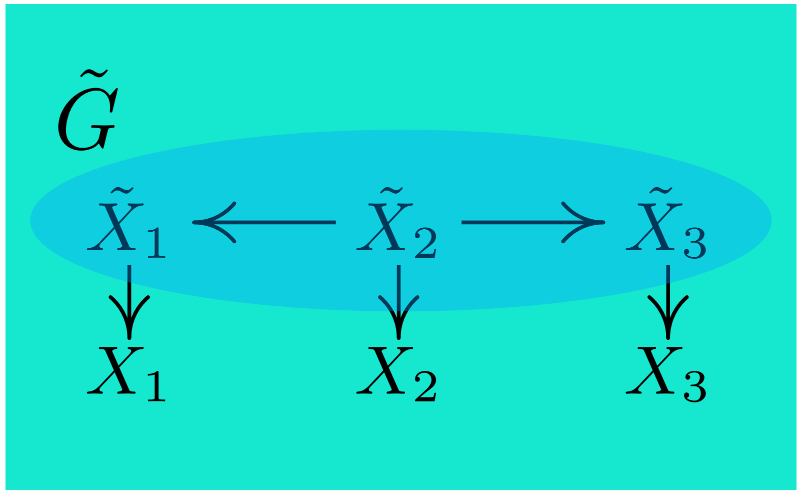


Practical Issues in Causal Discovery...

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)

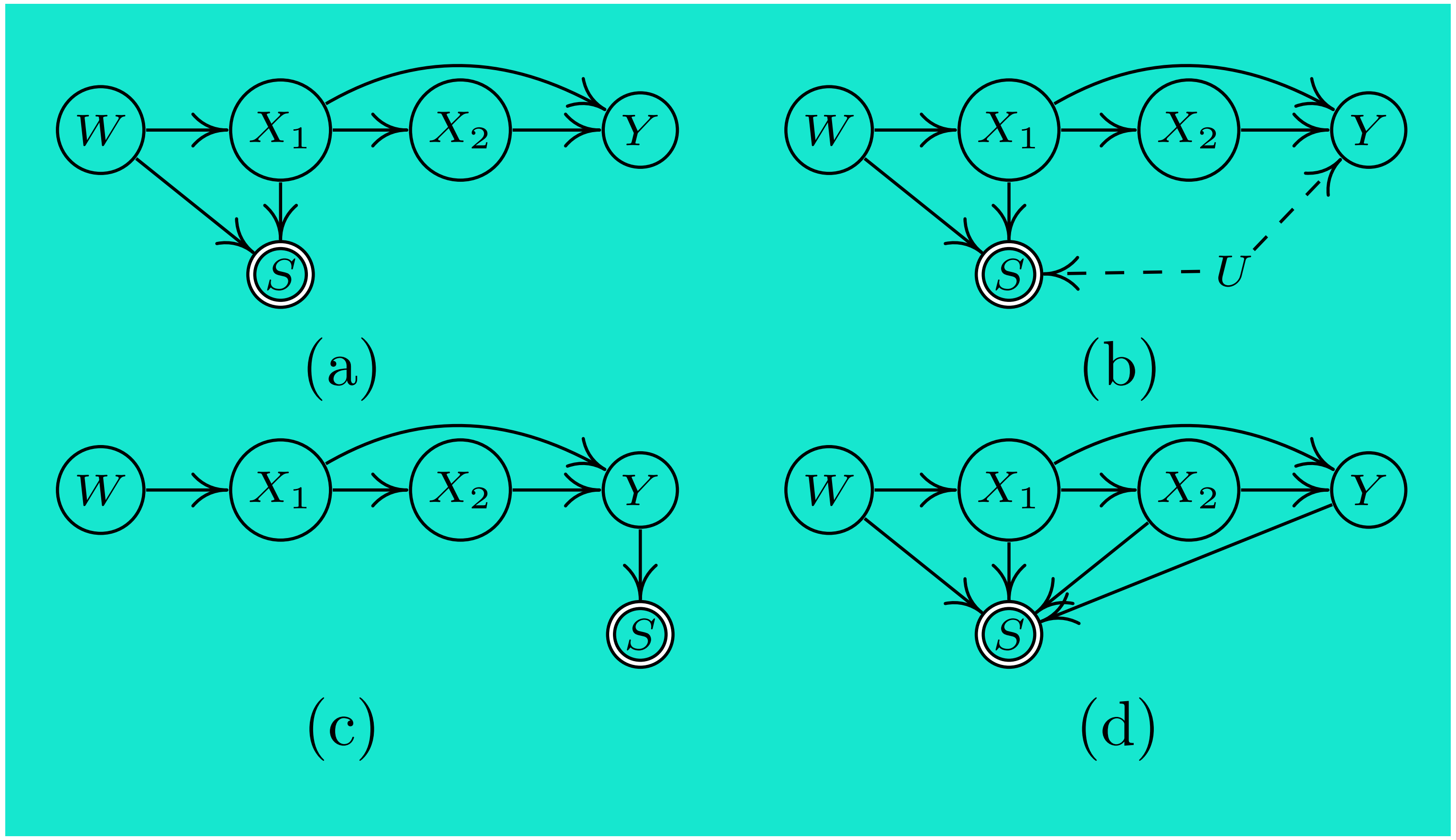
Practical Issues in Causal Discovery…

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- **Measurement error** (Zhang et al., UAI'18; PSA'18)



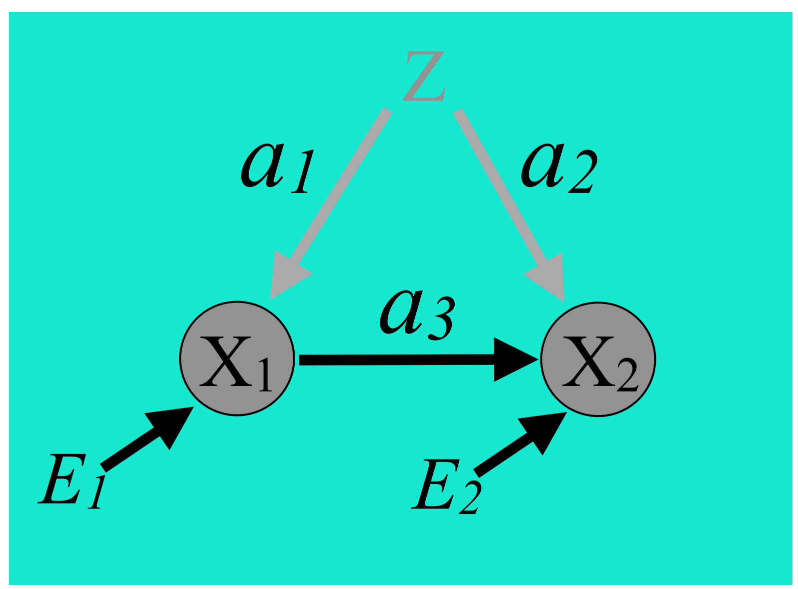
Practical Issues in Causal Discovery...

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- **Selection bias** (Zhang et al., UAI'16)



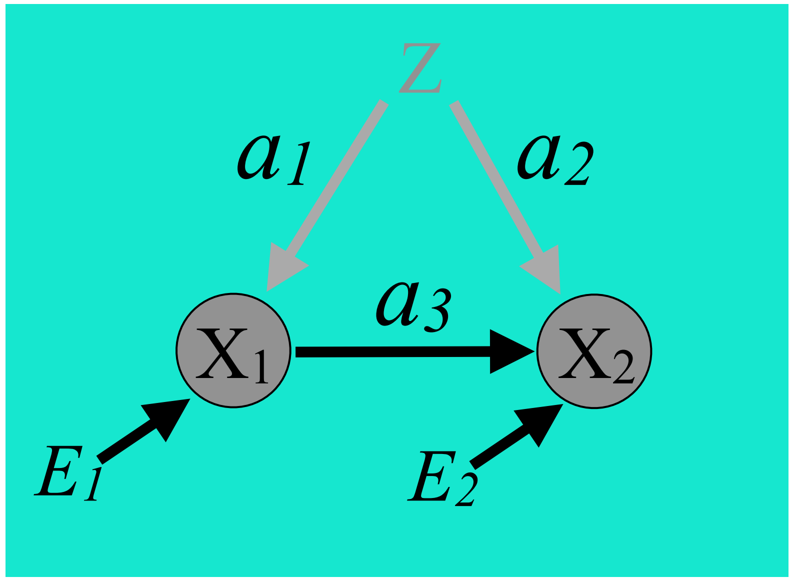
Practical Issues in Causal Discovery...

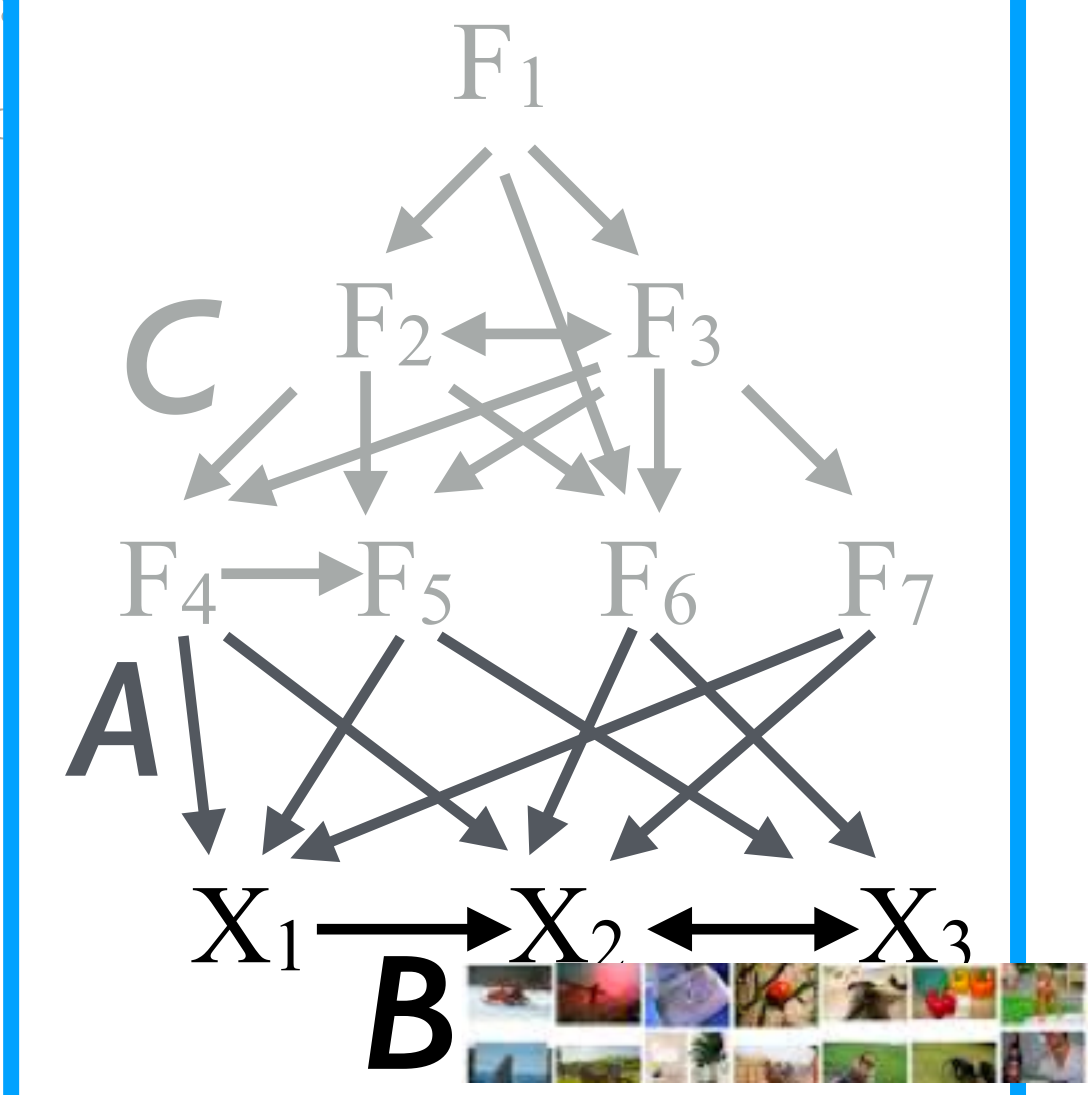
- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Zhang et al., UAI'16)
- **Confounding** (SGS 1993; Hoyer et al., 2008; Cai et al., NIPS'19...)



Practical Issues in Causal Discovery…

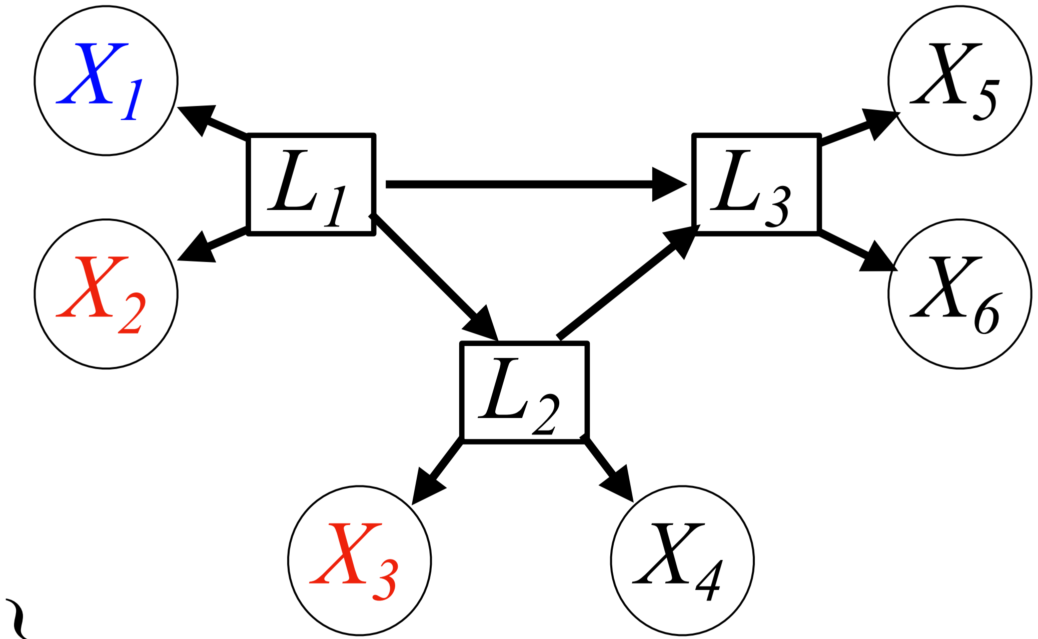
- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et
- Measurement error (Zhang et al., UAI'18; PSA'1
- Selection bias (Zhang et al., UAI'16)
- **Confounding** (SGS 1993; Hoyer et al., 2008; Cai



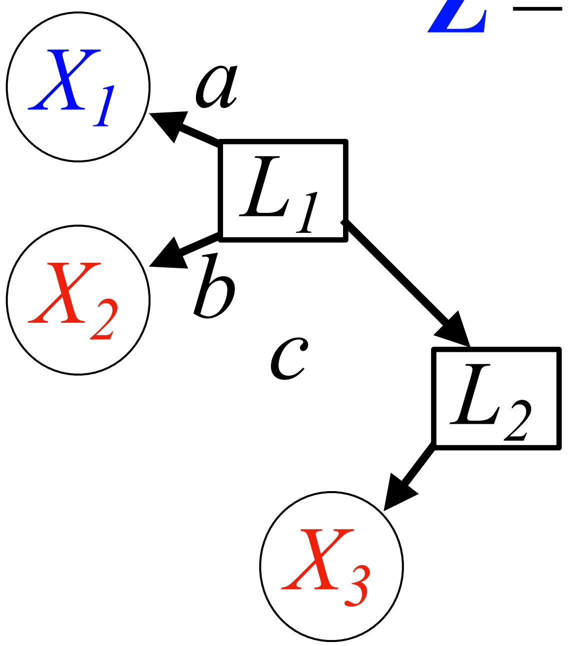


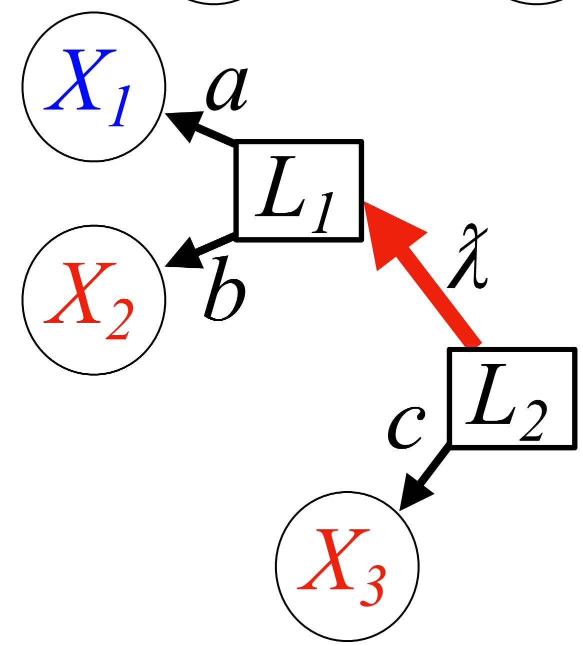
$$\mathbf{F} = C\mathbf{F} + \mathbf{E}_F;$$
$$\mathbf{X} = B\mathbf{X} + A\mathbf{F} + \mathbf{E}_X$$

Generalized Independent Noise (GIN) Condition



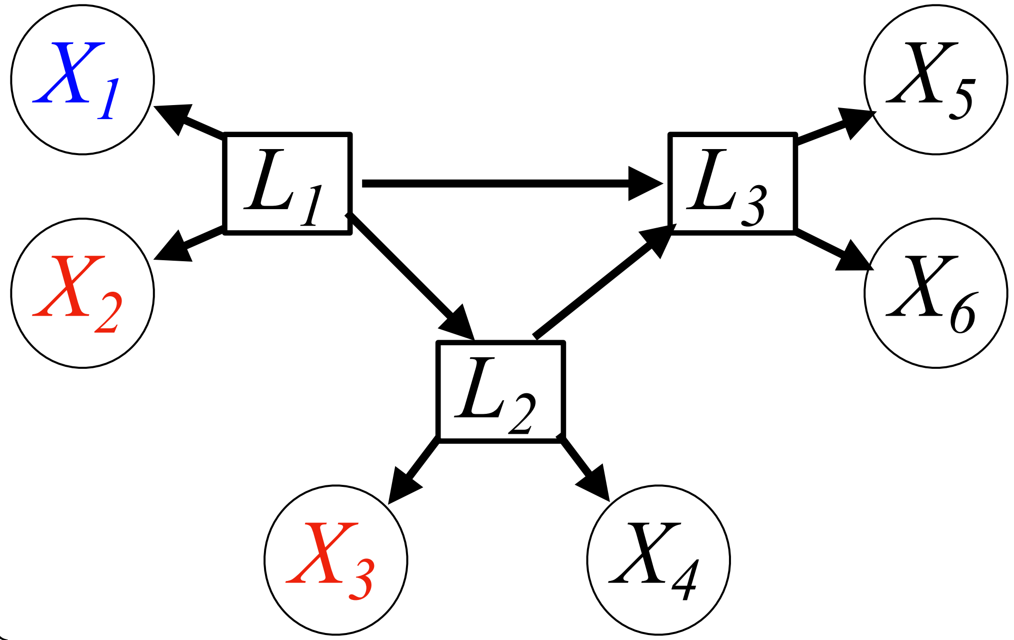
$\mathbf{Z} = \{X_1\}$ $\mathbf{Y} = \{X_2, X_3\}$



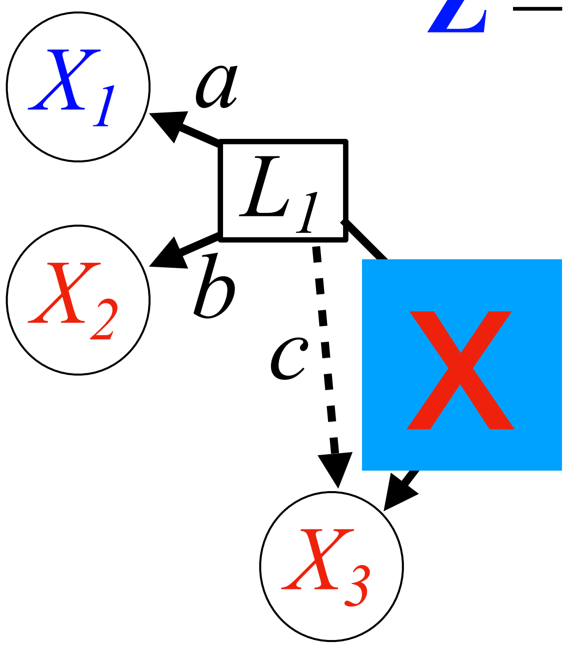


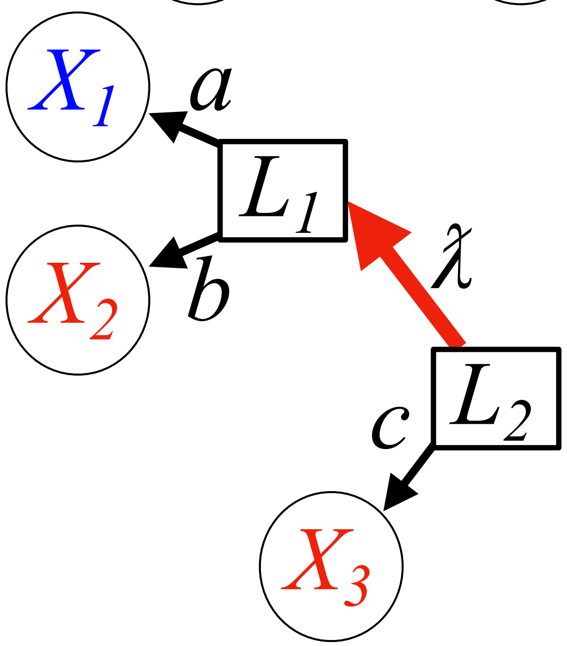
Cai, Xie, Glymour, Hao, Zhang, "Triad Constraints for Learning Causal Structure of Latent Variables," NeurIPS'19

Generalized Independent Noise (GIN) Condition



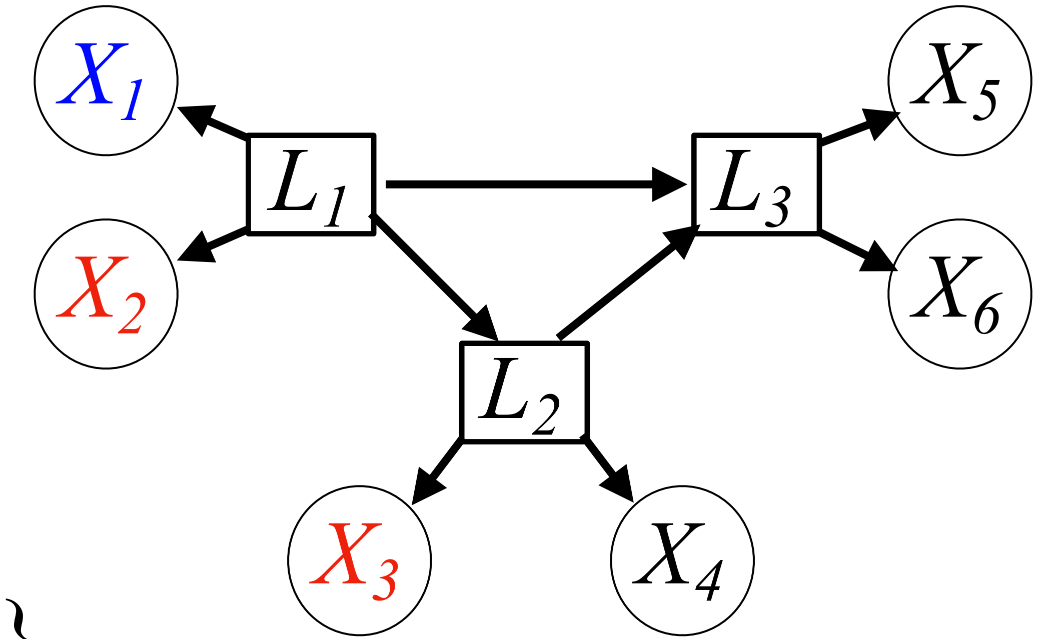
$\mathbf{Z} = \{X_1\}$ $\mathbf{Y} = \{X_2, X_3\}$



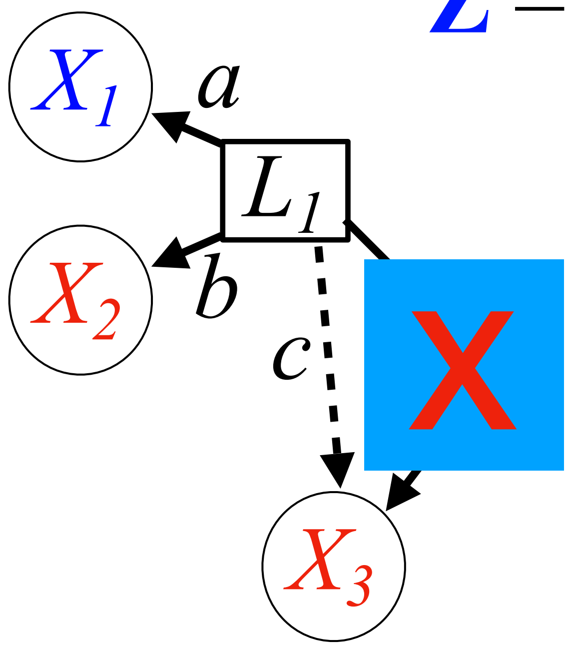


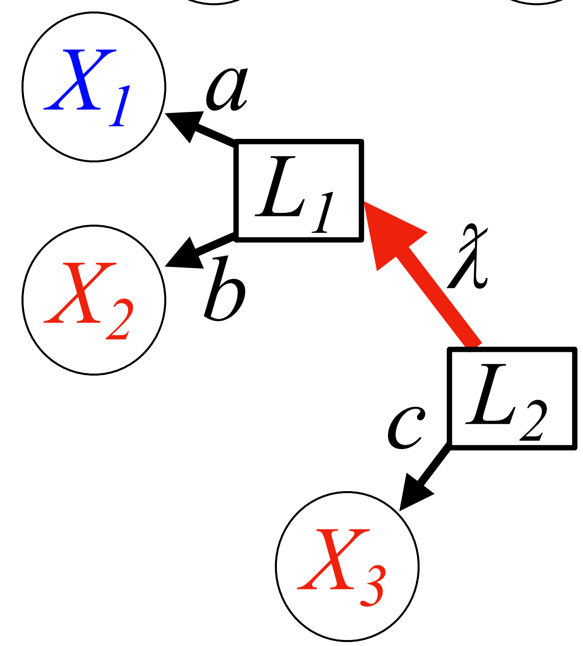
Cai, Xie, Glymour, Hao, Zhang, "Triad Constraints for Learning Causal Structure of Latent Variables," NeurIPS'19

Generalized Independent Noise (GIN) Condition



$\mathbf{Z} = \{X_1\}$ $\qquad$ $\mathbf{Y} = \{X_2, X_3\}$





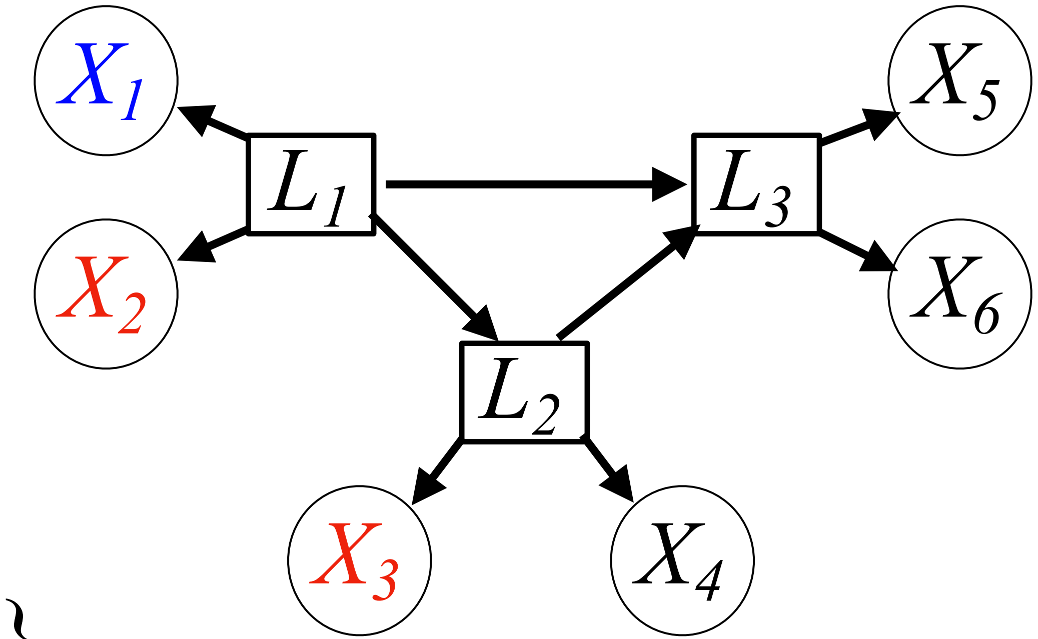
$$
\begin{aligned}
&c \cdot X_2 - b \cdot X_3 \\
=&c(bL1 + E_2) - b(cL_1 + E_3) \\
=&cE_2 - bE_3,
\end{aligned}
$$

independent from L_1 and from X_1,

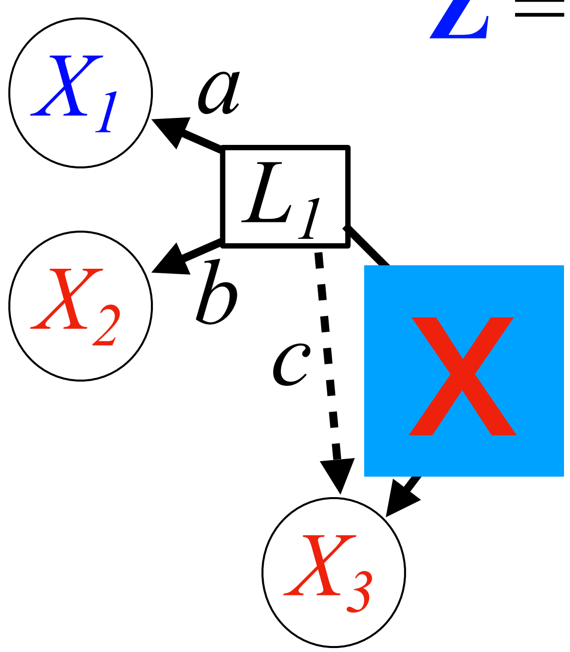
and we know $\frac{b}{c} = \frac{Cov(X_2, X_3)}{Cov(X_1, X_3)}$

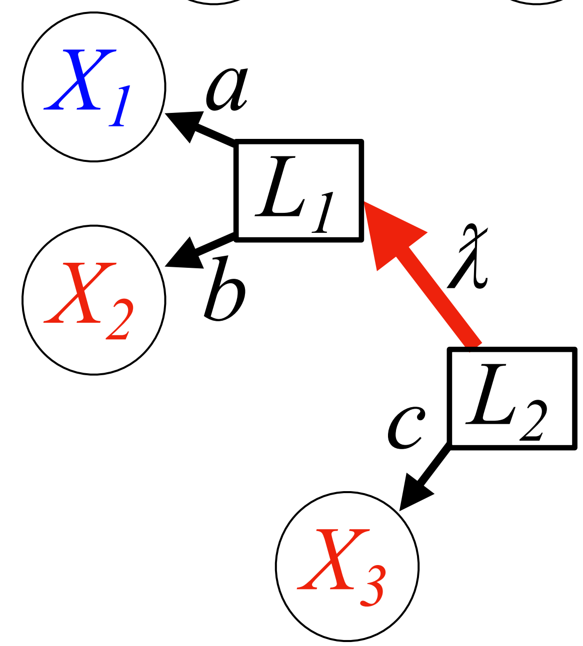
Cai, Xie, Glymour, Hao, Zhang, “Triad Constraints for Learning Causal Structure of Latent Variables,“ NeurIPS’19

Generalized Independent Noise (GIN) Condition



$\mathbf{Z} = \{X_1\}$ $\mathbf{Y} = \{X_2, X_3\}$





$$
\begin{aligned}
& c \cdot X_2 - b \cdot X_3 \\
=& c(bL1 + E_2) - b(cL_1 + E_3) \\
=& cE_2 - bE_3,
\end{aligned}
$$

independent from L_1 and from X_1,

and we know $\dfrac{b}{c} = \dfrac{Cov(X_2, X_3)}{Cov(X_1, X_3)}$

Nontrivial linear combination of X_2 and X_3 will involve the noise term in L_1, hence dependent on X_1

Cai, Xie, Glymour, Hao, Zhang, "Triad Constraints for Learning Causal Structure of Latent Variables," NeurIPS'19

Practical Issues in Causal Discovery...

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19)
- **Missing values (Tu et al., AISTATS'19)**

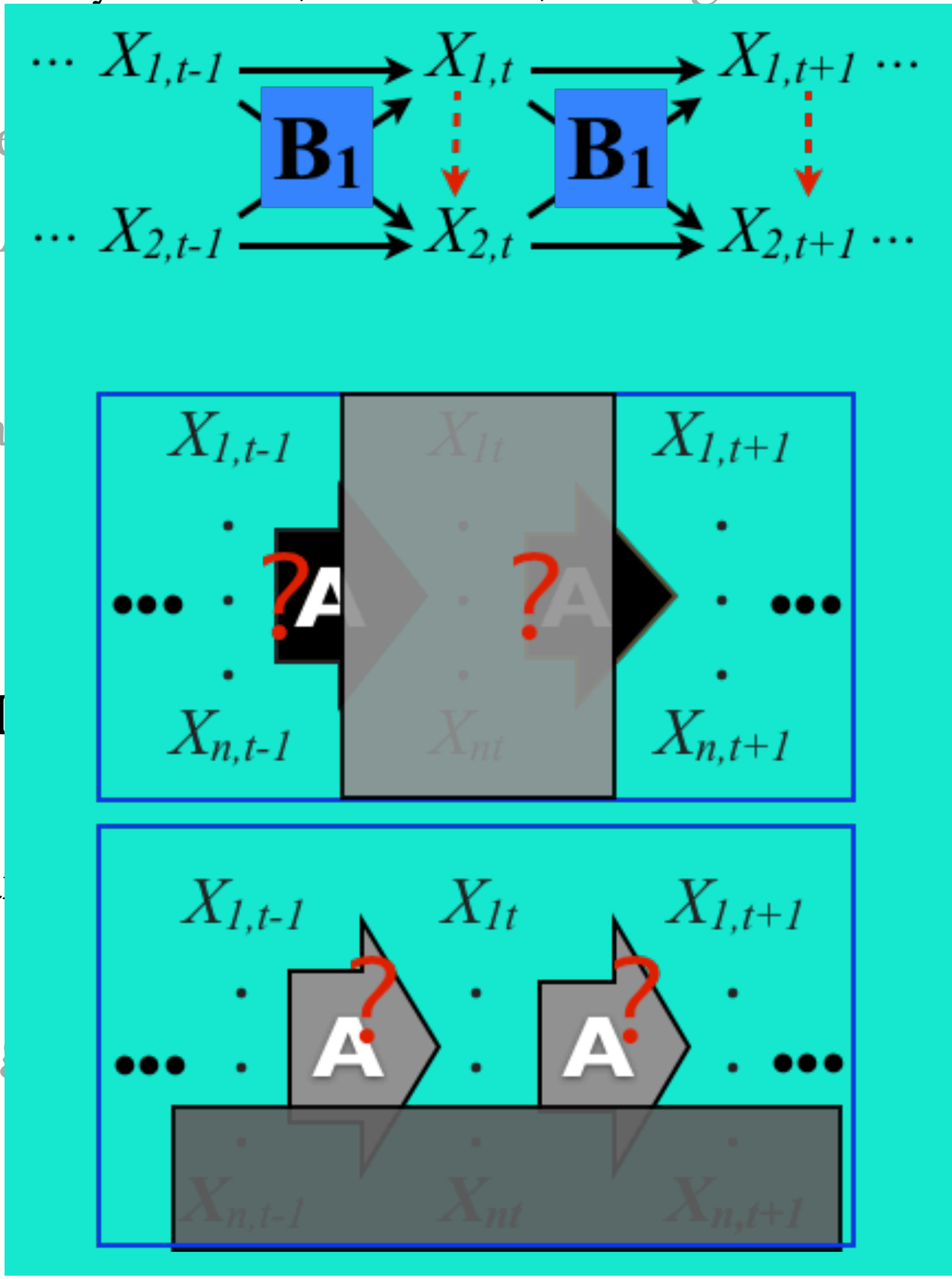
X1	X2	X3	X4	X5	X6
-9.4653403e-01	6.6703495e-01	8.2886922e-01	-1.3695521e+00	-3.2675465e-02	1.8634806e-01
-9.4895568e-01			-4.6381657e-01	-1.8280031e+00	
	5.1435422e-01	6.7338326e-01	4.3403559e-01	9.4535076e-01	7.5164028e-01
7.2489037e-01		5.1325341e-01	8.3567780e-01	2.9825903e-01	7.7796018e-02
		-1.3440612e+00			-7.3325009e-01
1.3261794e+00	-6.1971037e-01	-1.0498756e-01	1.4171149e+00	1.6251026e+00	3.7478050e-01
-2.1128404e+00	1.3359744e-02	-2.0209600e+00	-1.7172659e+00	-2.4746799e+00	-2.8026586e+00
1.5453163e+00	-5.3986972e-01	4.5157367e-01	1.5566262e+00	9.3882105e-01	-4.3382982e-01
6.5974086e-02	5.5826895e-01	6.5247930e-01	-5.7895322e-01	5.0062743e-01	1.0183537e+00
8.9772858e-01	2.6752870e-01	-4.9204975e-01	7.7933358e-02	8.3467624e-01	9.2744311e-01
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Practical Issues in Causal Discovery...

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19)
- Missing values (Tu et al., AISTATS'19)
- **Causality in time series**
    - Time-delayed + instantaneous relations (Hyvarinen ICML'08; Zhang et al., ECML'09; Hyvarinen et al., JMLR'10)
    - Subsampling / temporally aggregation (Danks & Plis, NIPS WS'14; Gong et al., ICML'15 & UAI'17)
    - From partially observable time series (Geiger et al., ICML'15)

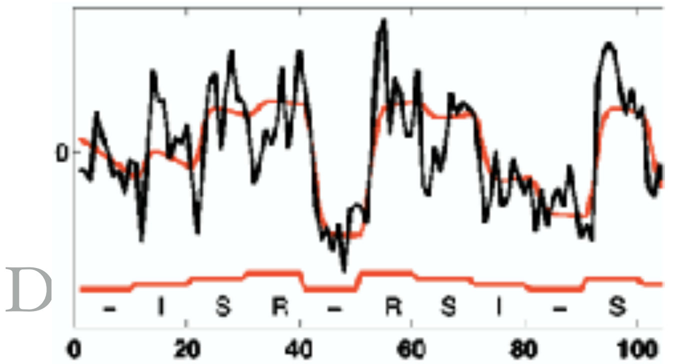
Practical Issues in Causal Discovery…

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang e
- Measurement error (Zhang et al., UAI'18; PS
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Ca
- Missing values (Tu et al., AISTATS'19)
- **Causality in time series**
    - Time-delayed + instantaneous relations (H ECML'09; Hyvarinen et al., JMLR'10)
    - Subsampling / temporally aggregation (Da ICML'15 & UAI'17)
    - From partially observable time series (Gei



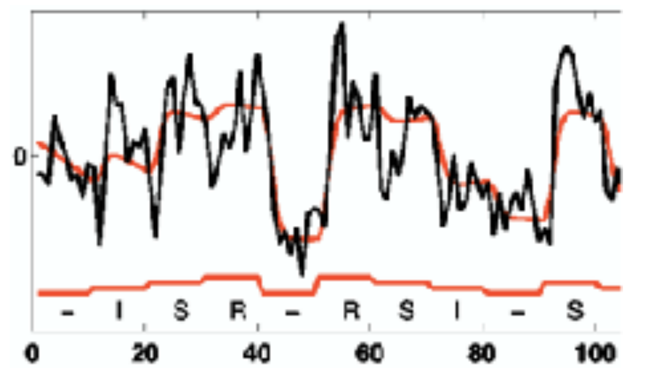
Practical Issues in Causal Discovery...

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; D
- Missing values (Tu et al., AISTATS'19)
- Causality in time series
    - Time-delayed + instantaneous relations (Hyvarinen ICML'o ECML'09; Hyvarinen et al., JMLR'10)
    - Subsampling / temporally aggregation (Danks & Plis, NIPS ICML'15 & UAI'17)
    - From partially observable time series (Geiger et al., ICML'15)
- Application in recommender systems (Wang et al., AAAI'18; Wang et al., NIPS'18)
- **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)**



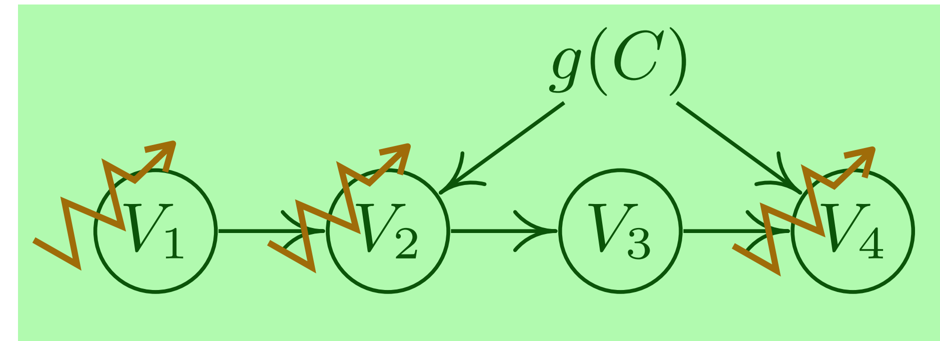
Nonstationary/Heterogeneous Data and Causality

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



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
Ghassami, et al., Multi-Domain Causal Structure Learning in Linear Systems, NIPS 2018

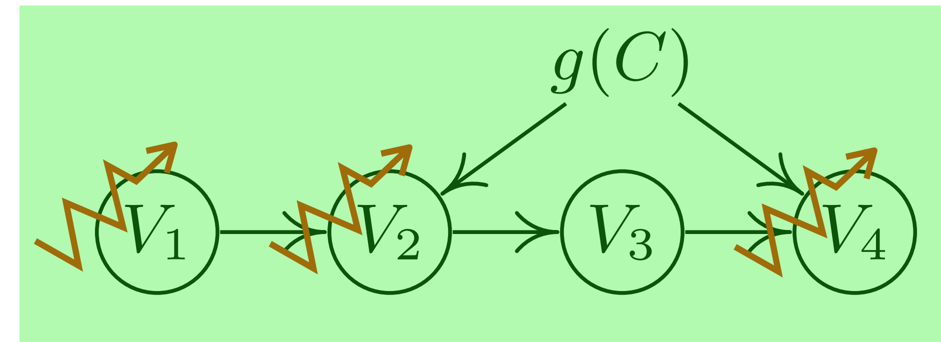
Causal Discovery from Nonstationary/ Heterogeneous Data

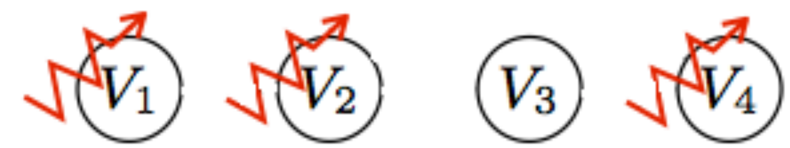


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in P(cause) and P(effect | cause)
  - How do the nonstationary modules change over time / across data sets?

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

Causal Discovery from Nonstationary/ Heterogeneous Data

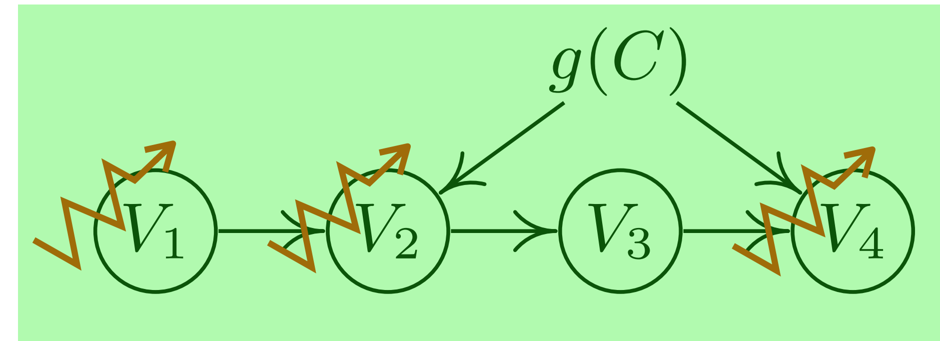




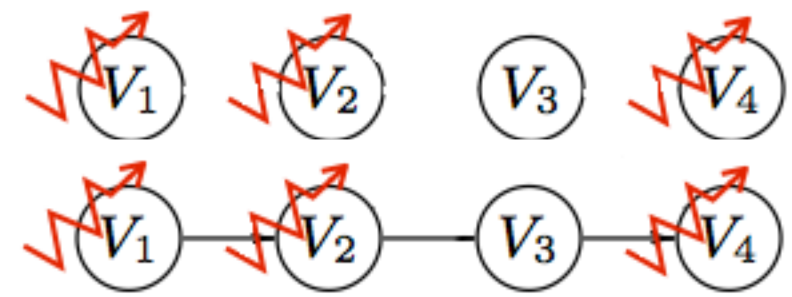
- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in P(cause) and P(effect | cause)
  - How do the nonstationary modules change over time / across data sets?

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

Causal Discovery from Nonstationary/ Heterogeneous Data

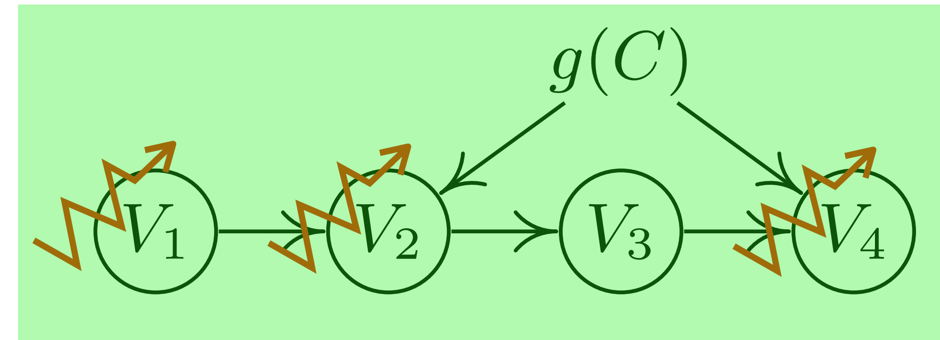


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in P(cause) and P(effect | cause)
  - How do the nonstationary modules change over time / across data sets?

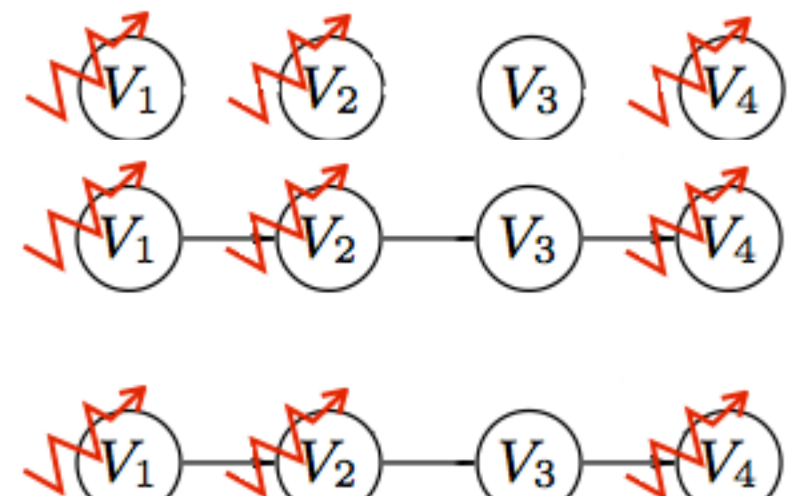


Zhang et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Zhang et al., Causal discovery in the presence of nonstatioarity/heterogeneity: Skeleton estimation and orientation determination, IJCAI 2017

Causal Discovery from Nonstationary/ Heterogeneous Data

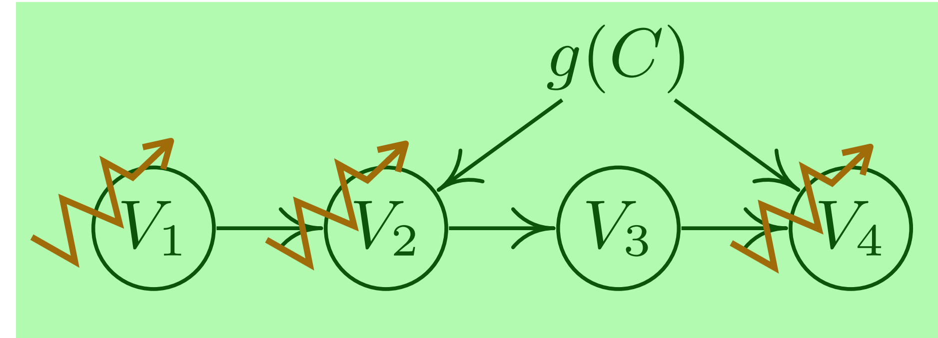


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
  - How do the nonstationary modules change over time / across data sets?

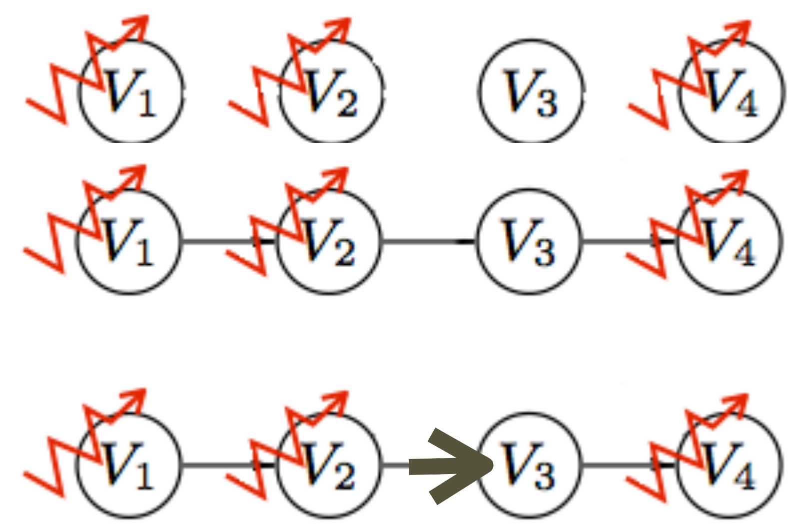


Zhang et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Zhang et al., Causal discovery in the presence of nonstatioarity/heterogeneity: Skeleton estimation and orientation determination, IJCAI 2017

Causal Discovery from Nonstationary/ Heterogeneous Data

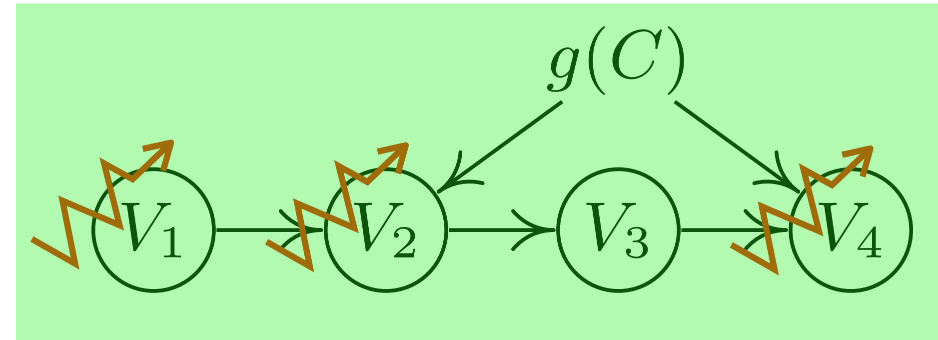


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
  - How do the nonstationary modules change over time / across data sets?

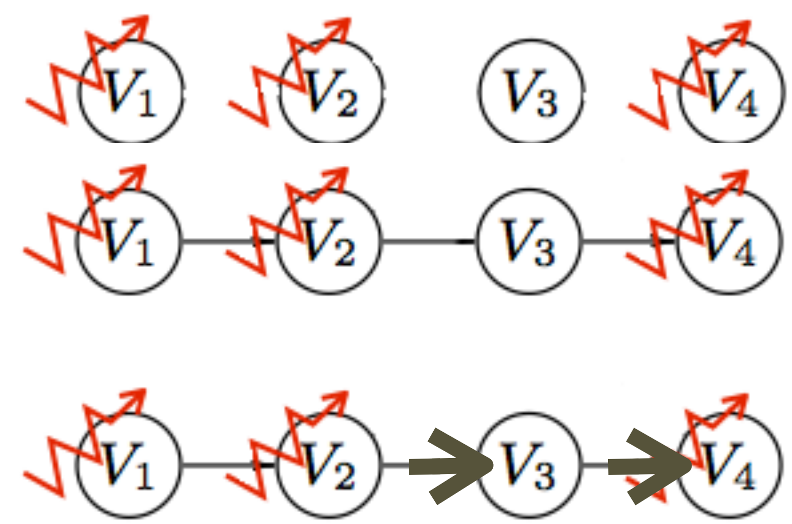


Zhang et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Zhang et al., Causal discovery in the presence of nonstatioarity/heterogeneity: Skeleton estimation and orientation determination, IJCAI 2017

Causal Discovery from Nonstationary/ Heterogeneous Data

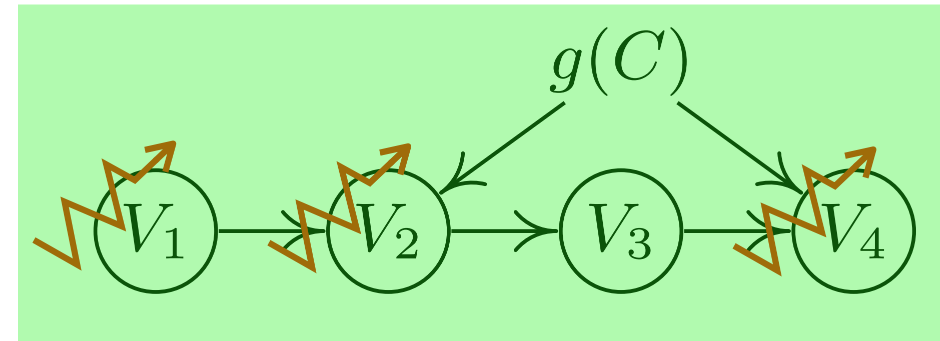


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
  - How do the nonstationary modules change over time / across data sets?

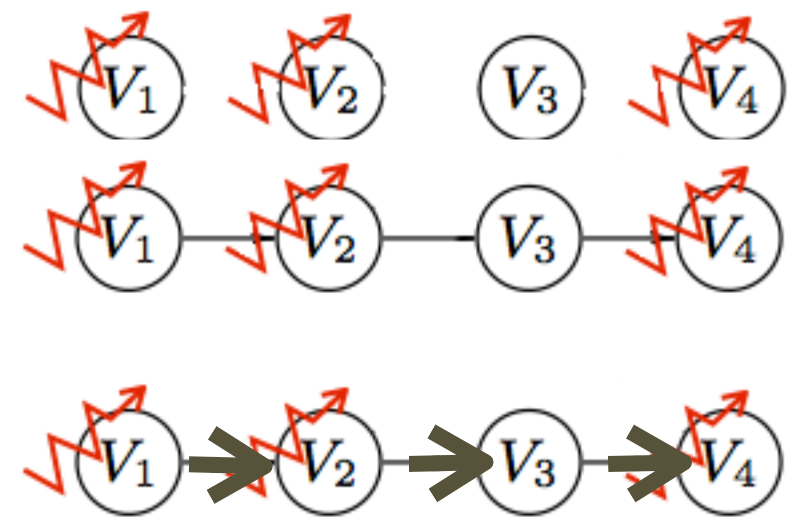


Zhang et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Zhang et al., Causal discovery in the presence of nonstatioarity/heterogeneity: Skeleton estimation and orientation determination, IJCAI 2017

Causal Discovery from Nonstationary/ Heterogeneous Data

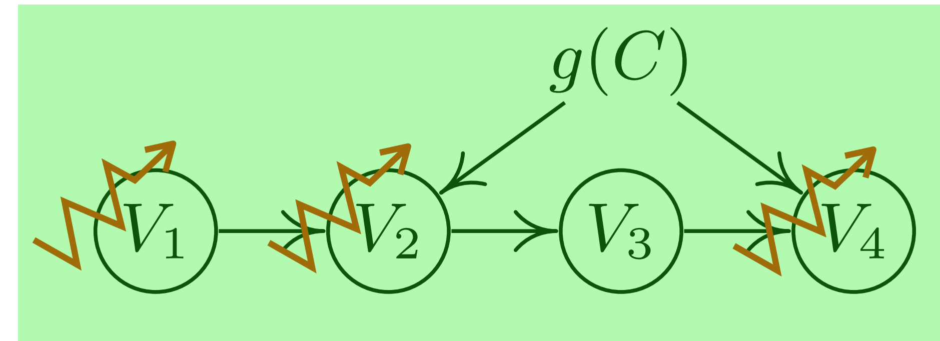


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
  - How do the nonstationary modules change over time / across data sets?

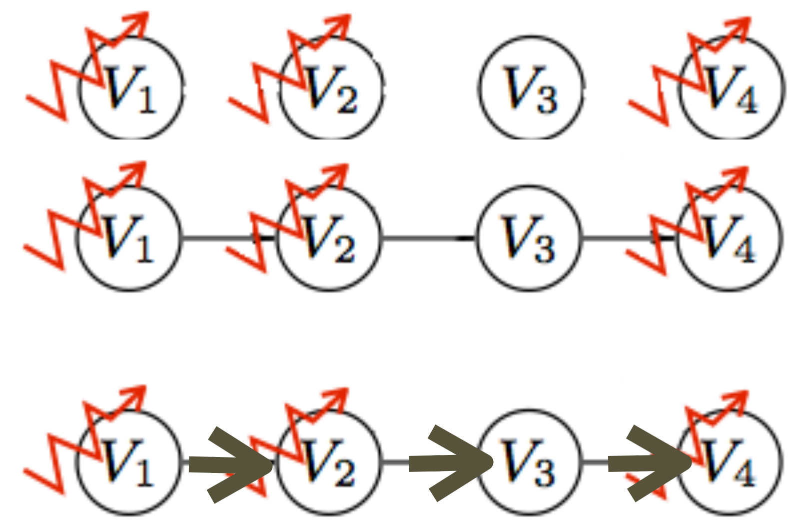


Zhang et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Zhang et al., Causal discovery in the presence of nonstatioarity/heterogeneity: Skeleton estimation and orientation determination, IJCAI 2017

Causal Discovery from Nonstationary/ Heterogeneous Data



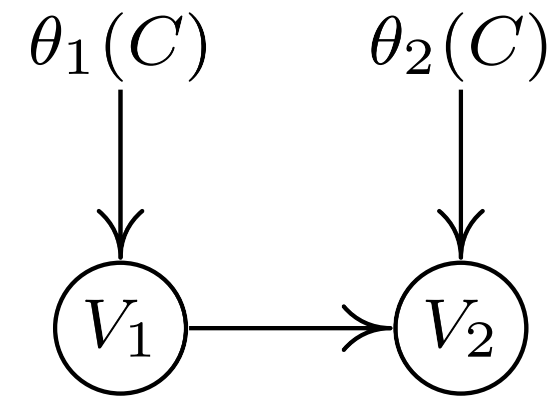
- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
  - How do the nonstationary modules change over time / across data sets?



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

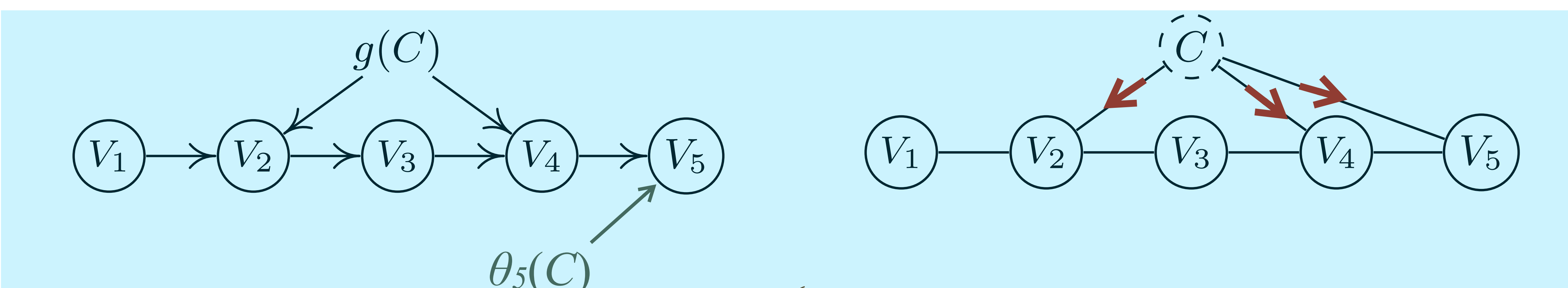
Nonstationarity Helps Determine Causal Direction



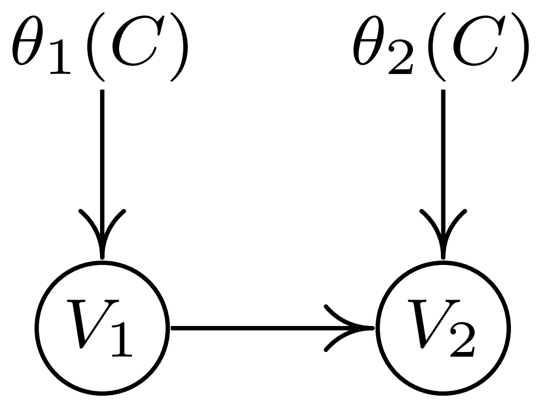
- Independent changes in *P*(cause) and *P*(effect | cause); <u>generally</u> violated for wrong directions

- Special cases: if $C - V_k - V_l$, since $C \to V_k$, we known
  - $C \to V_k \leftarrow V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **excluding** V_k Invariant cause
  - $C \to V_k \to V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **including** V_k Invariant mechanism

Hoover. The logic of causal inference. Economics and Philosophy, 6:207–234, 1990.



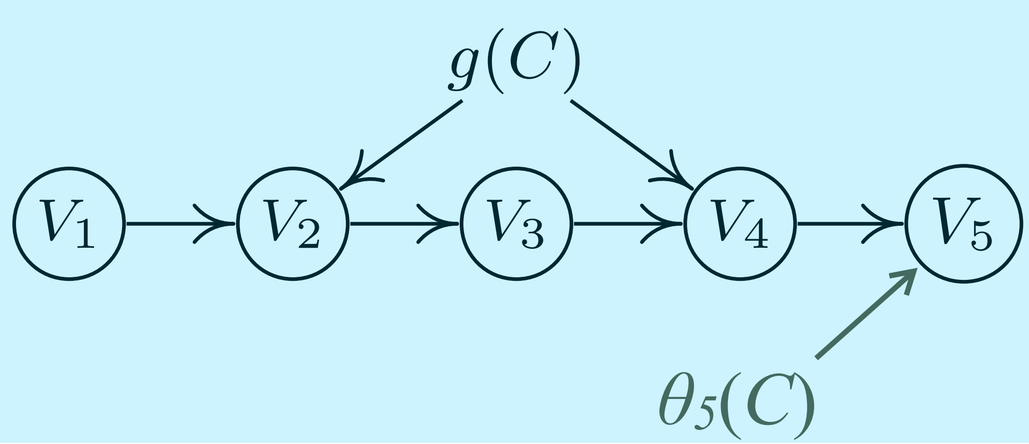
Nonstationarity Helps Determine Causal Direction

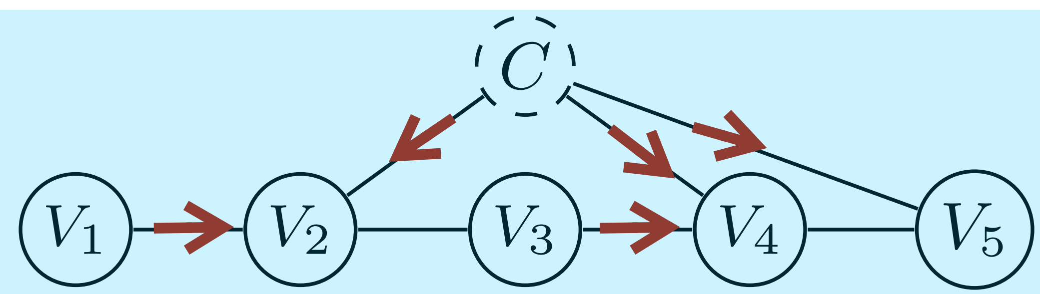


- Independent changes in P(cause) and P(effect | cause); generally violated for wrong directions

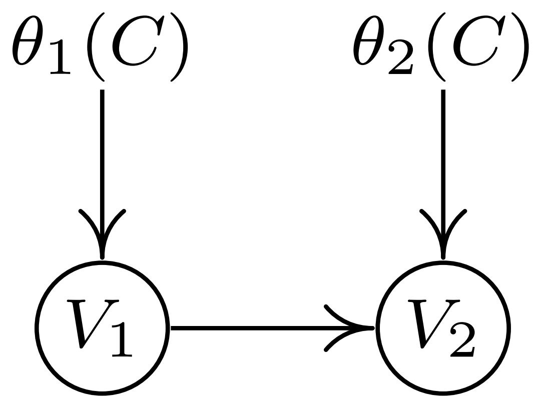
- Special cases: if $C - V_k - V_l$, since $C \to V_k$, we known
    - $C \to V_k \leftarrow V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **excluding** V_k Invariant cause
    - $C \to V_k \to V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **including** V_k Invariant mechanism

Hoover. The logic of causal inference. Economics and Philosophy, 6:207–234, 1990.





Nonstationarity Helps Determine Causal Direction



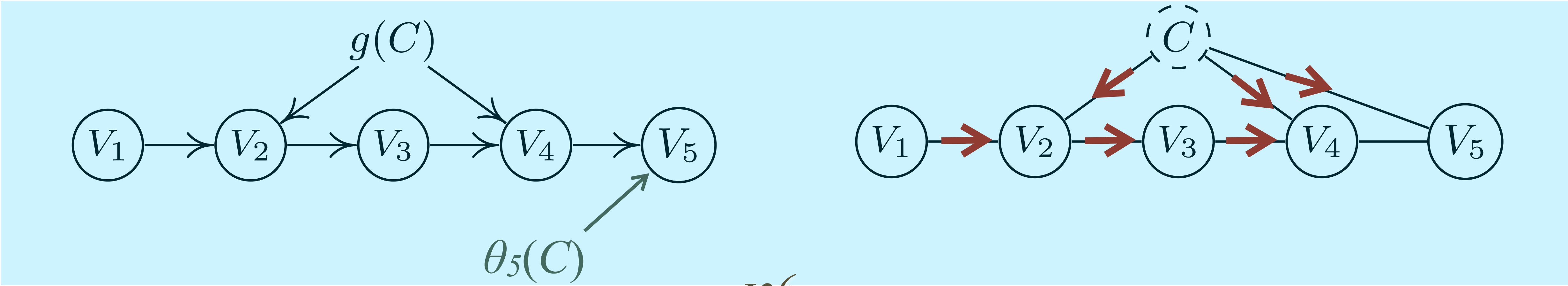
- Independent changes in P(cause) and P(effect | cause); generally violated for wrong directions

- Special cases: if $C - V_k - V_l$, since $C \to V_k$, we known
  - $C \to V_k \leftarrow V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **excluding** V_k
  - $C \to V_k \to V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **including** V_k

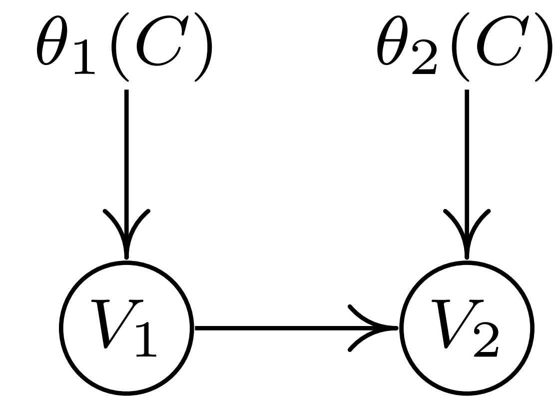
Invariant cause

Invariant mechanism

Hoover. The logic of causal inference. Economics and Philosophy, 6:207–234, 1990.



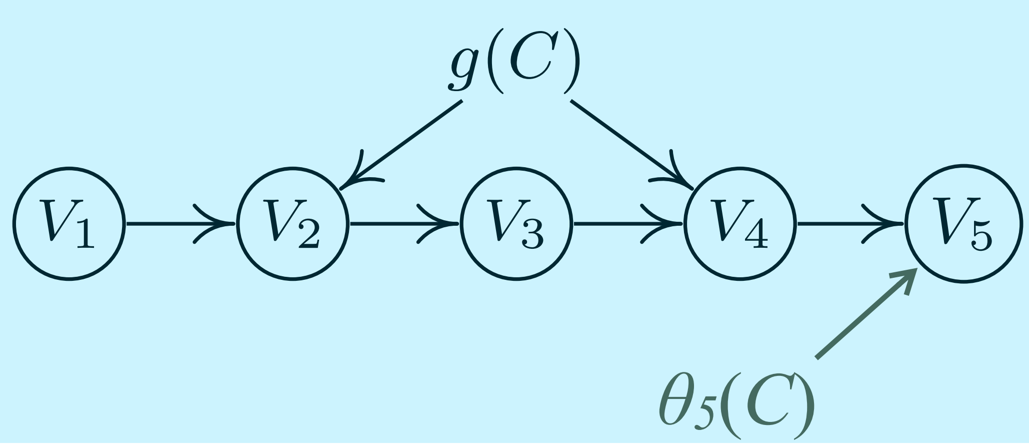
Nonstationarity Helps Determine Causal Direction

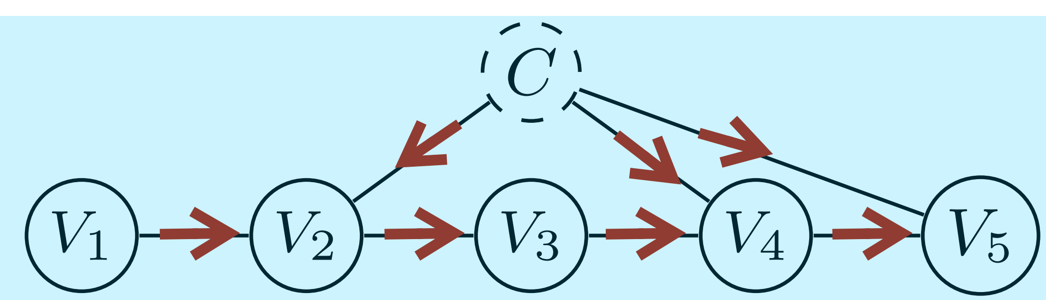


- Independent changes in P(cause) and P(effect | cause); generally violated for wrong directions

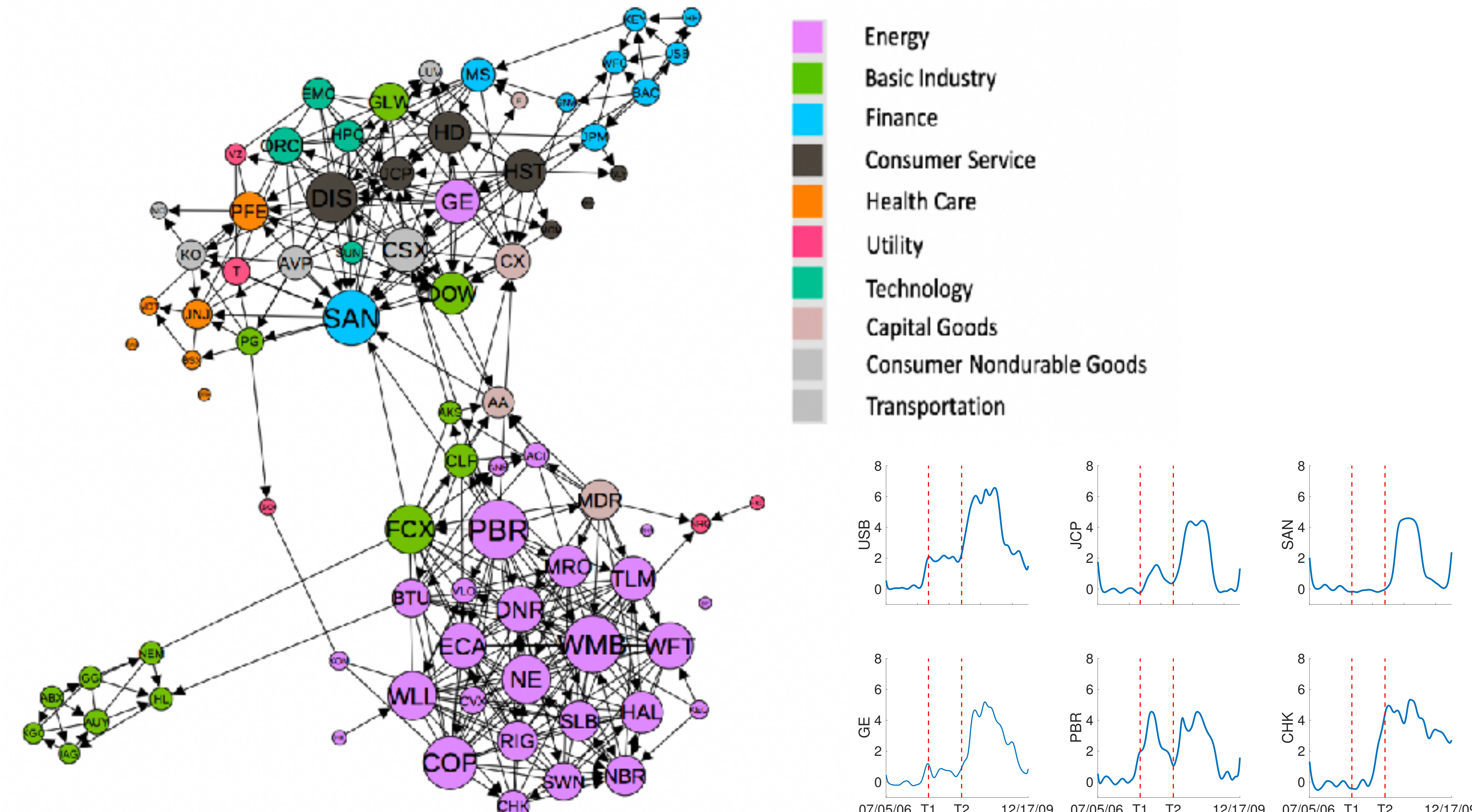
- Special cases: if $C - V_k - V_l$, since $C \to V_k$, we known
  - $C \to V_k \leftarrow V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **excluding** V_k Invariant cause
  - $C \to V_k \to V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **including** V_k Invariant mechanism

Hoover. The logic of causal inference. Economics and Philosophy, 6:207–234, 1990.





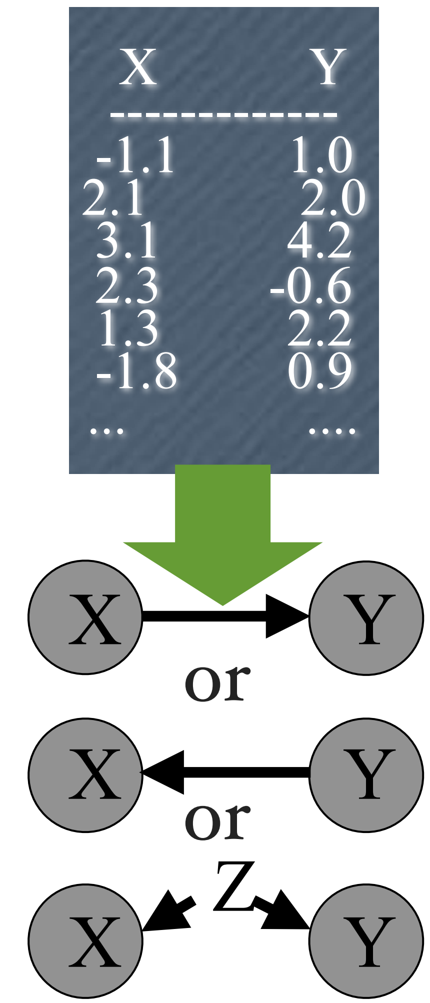
Causal Analysis of Major Stocks in NYSE (07/05/2006 - 12/16/2009)

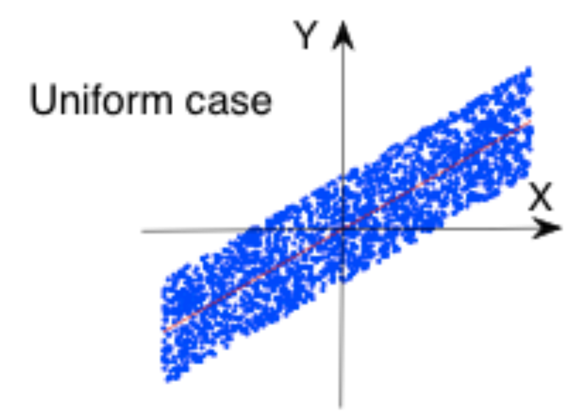


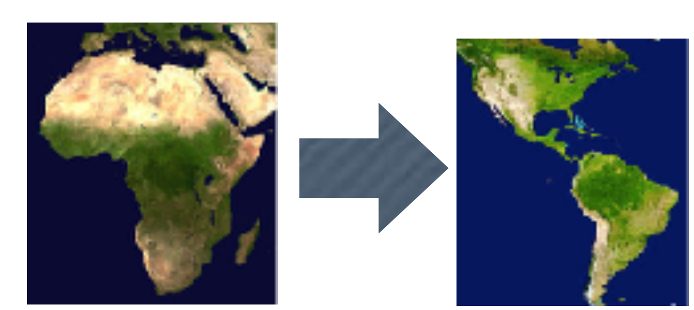
Huang, Zhang, Zhang, Romero, Glymour, Schölkopf, Behind Distribution Shift: Mining Driving Forces of Changes and Causal Arrows," ICDM 2017

Outline

- Causality? Interventions? Causal thinking
- Causal graphical models
- Identification of causal effects
- Counterfactual reasoning
- Causal discovery
- **Implications in machine learning**







Domain Adaptation (or Transfer Learning)

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

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

Causal model $Y \rightarrow X$

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

"Causality" Matters in Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

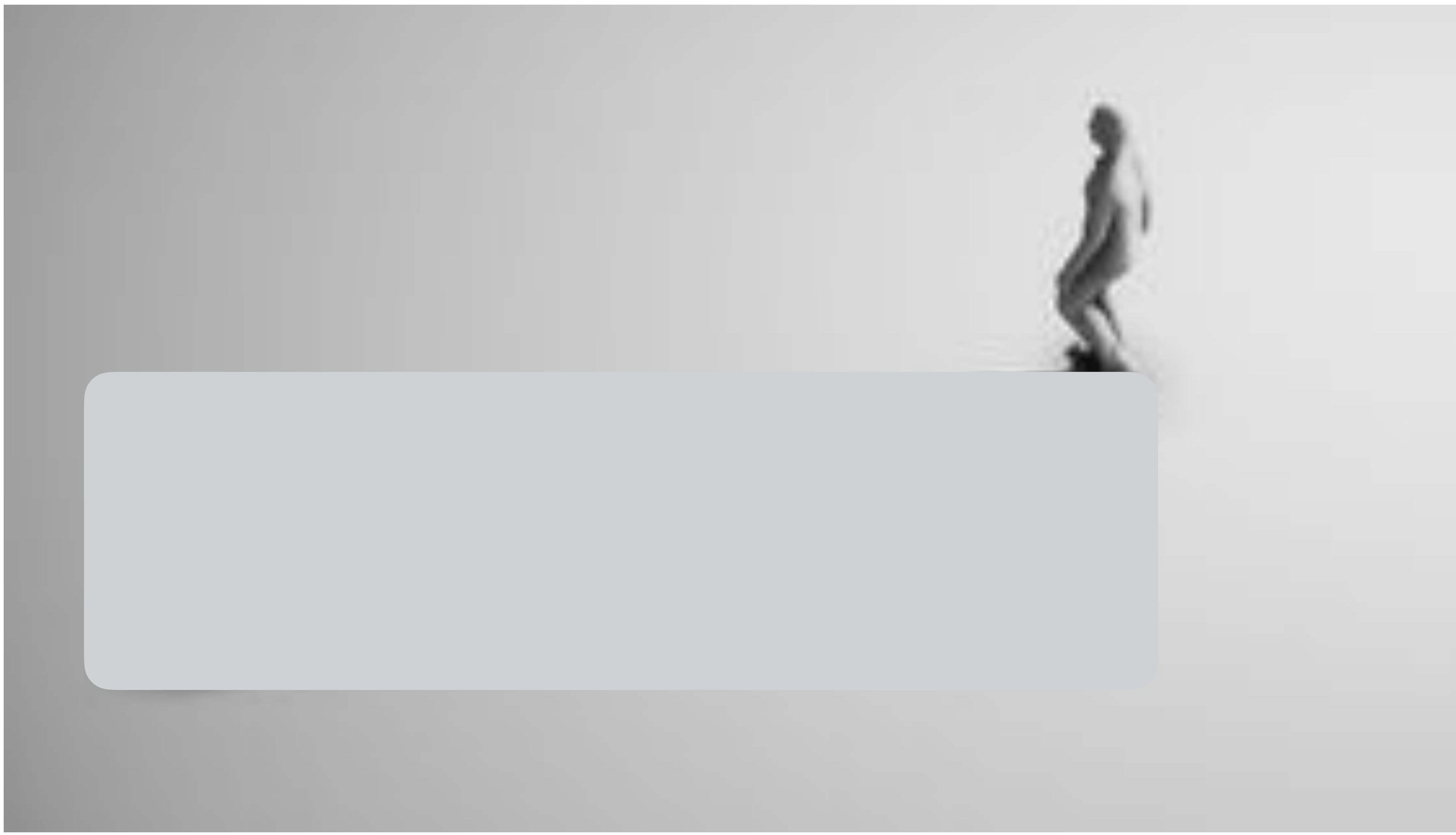
“Causality” Matters in Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

“Causality” Matters in Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

“Causality” Matters in Prediction: An Illustration

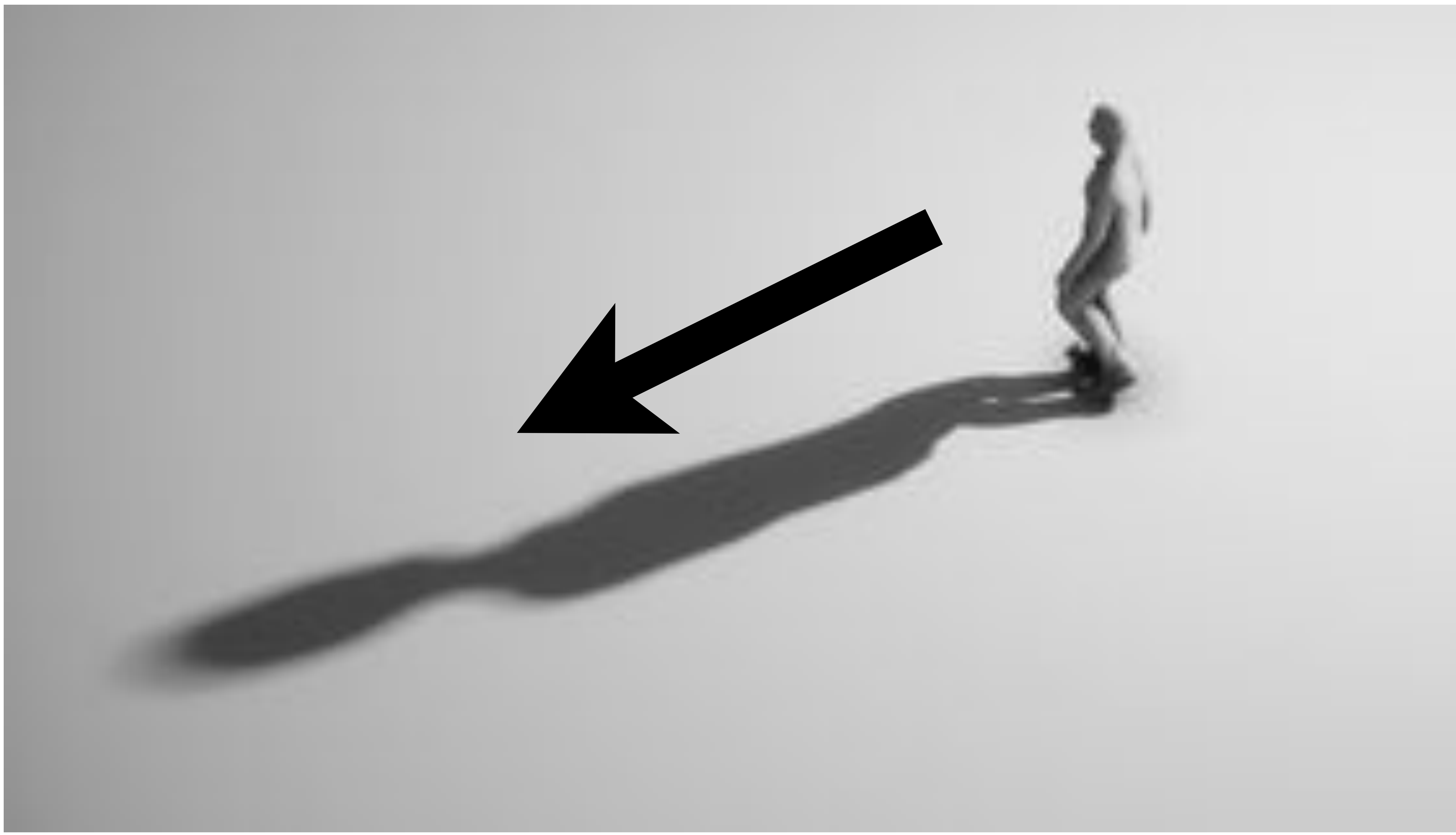


Understanding connections between different scenarios & *modeling* differences

“Causality” Matters in Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

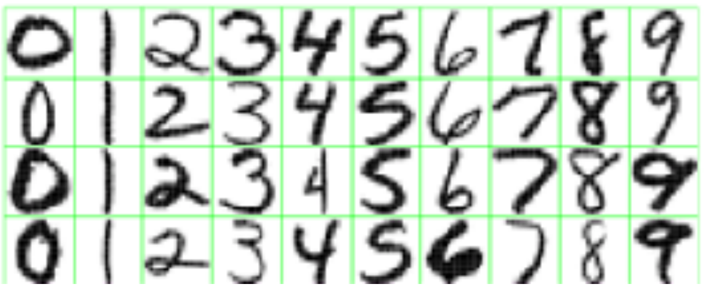
“Causality” Matters in Prediction: An Illustration



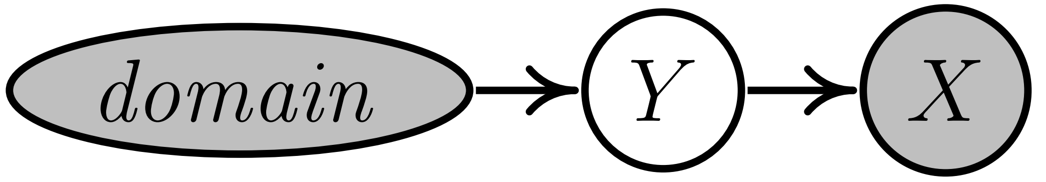
Understanding connections between different scenarios & *modeling* differences

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

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



- Target shift (TarS)

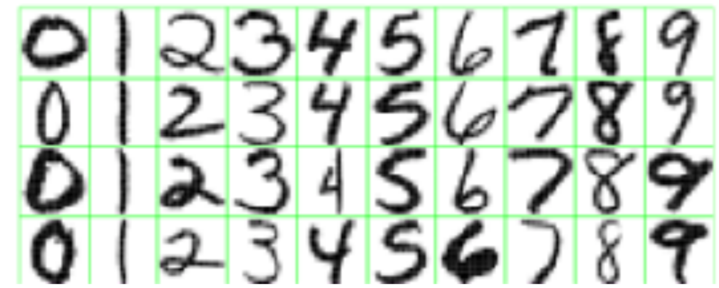


involved parameters estimated by matching P_X

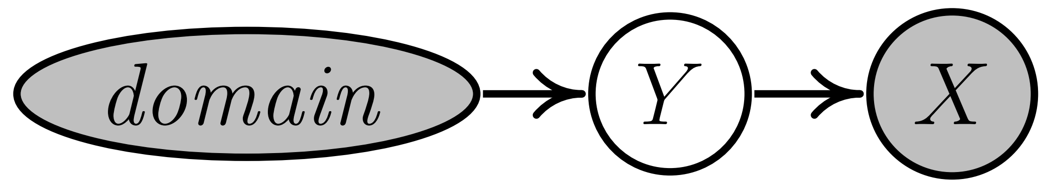
Zhang et al., ICML 2013; Zhang et al., AAAI 2015; Gong et al., ICML 2016; Stojanov et al., AISTATS 2018; Zhao et al., ICML 2019; Fu et al., CVPR 2019…

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

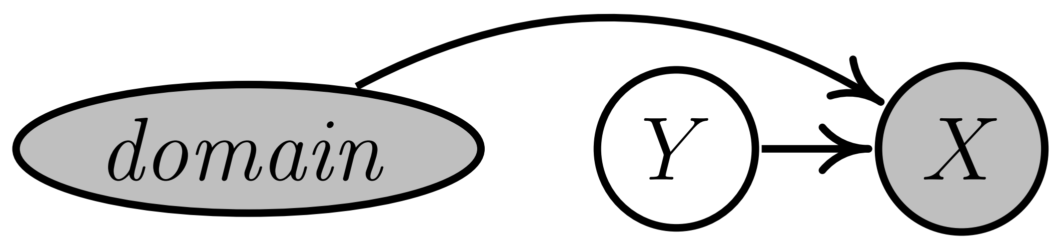
- Y is usually the cause of X
  (especially for classification)



- Target shift (TarS)



- Conditional shift (ConS)

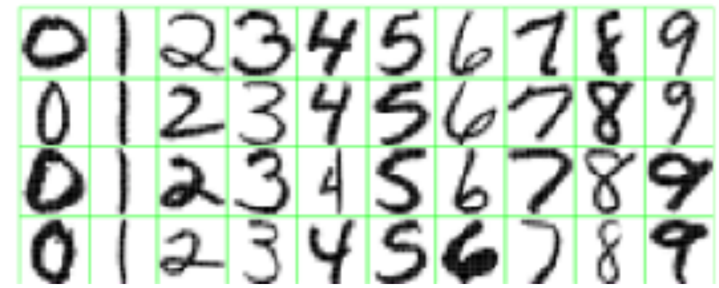


involved parameters estimated by matching P_X

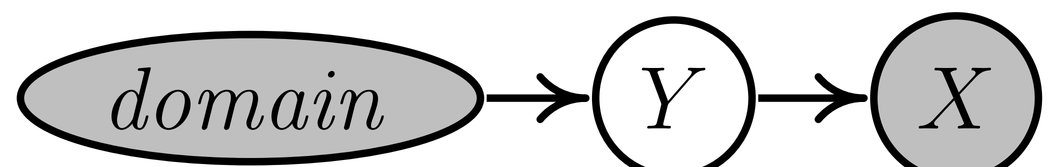
Zhang et al., ICML 2013; Zhang et al., AAAI 2015; Gong et al., ICML 2016; Stojanov et al., AISTATS 2018; Zhao et al., ICML 2019; Fu et al., CVPR 2019…

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

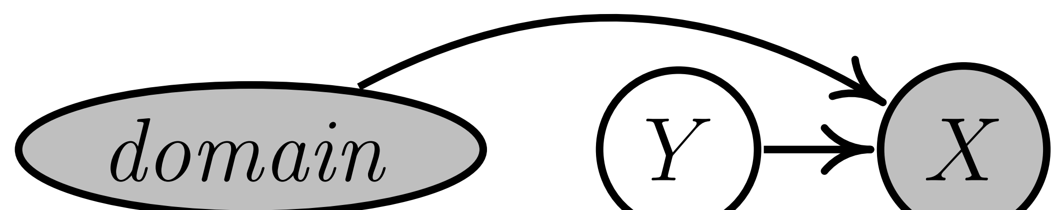
- Y is usually the cause of X
  (especially for classification)



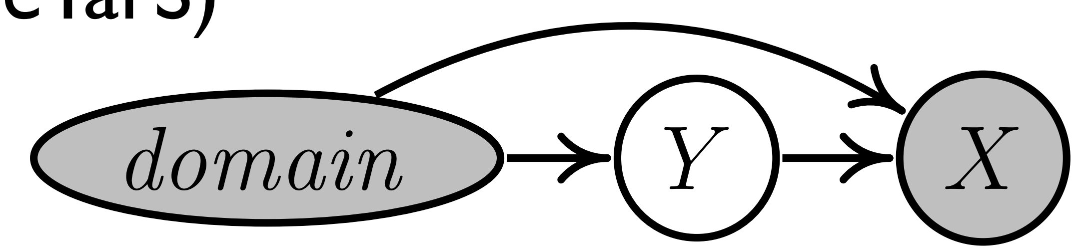
- Target shift (TarS)



- Conditional shift (ConS)



- Generalized target shift (GeTarS)



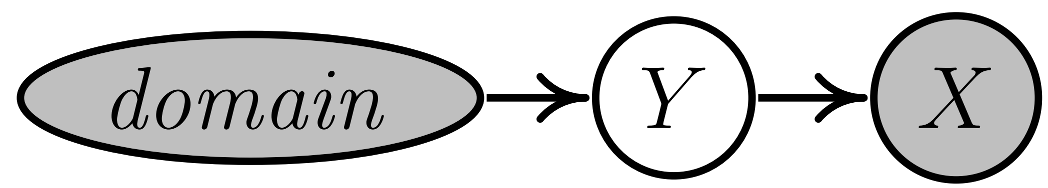
involved parameters estimated by matching P_X

Zhang et al., ICML 2013; Zhang et al., AAAI 2015; Gong et al., ICML 2016; Stojanov et al., AISTATS 2018; Zhao et al., ICML 2019; Fu et al., CVPR 2019…

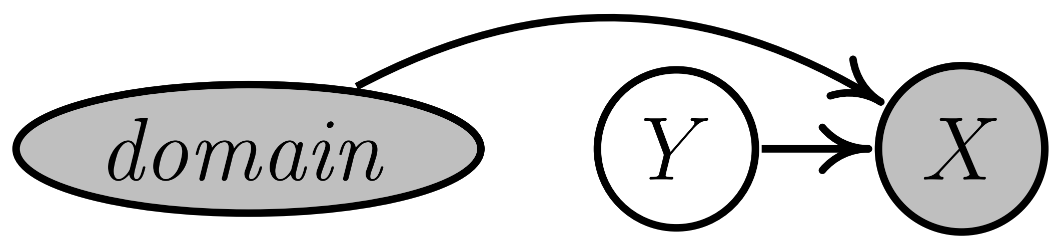
Possible Situations for Domain Adaptation: When Y→**X** (Zhang et al., 2013)

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

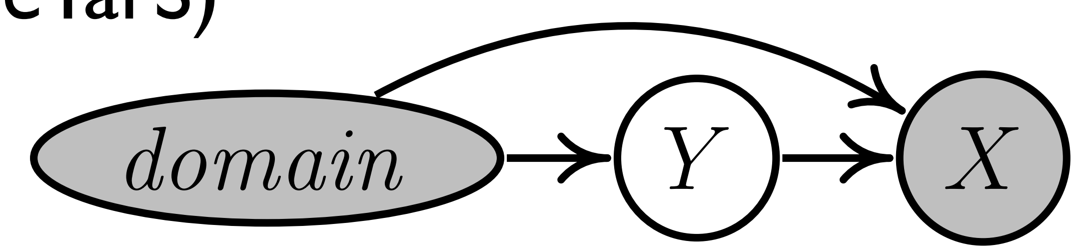

- Target shift (TarS)



- Conditional shift (ConS)



- Generalized target shift (GeTarS)



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

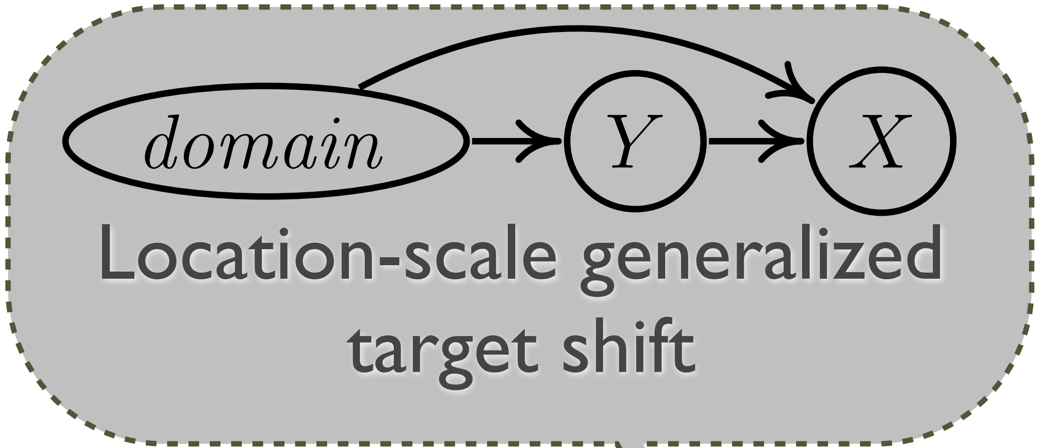
involved parameters estimated by matching P_X

Zhang et al., ICML 2013; Zhang et al., AAAI 2015; Gong et al., ICML 2016; Stojanov et al., AISTATS 2018; Zhao et al., ICML 2019; Fu et al., CVPR 2019…

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

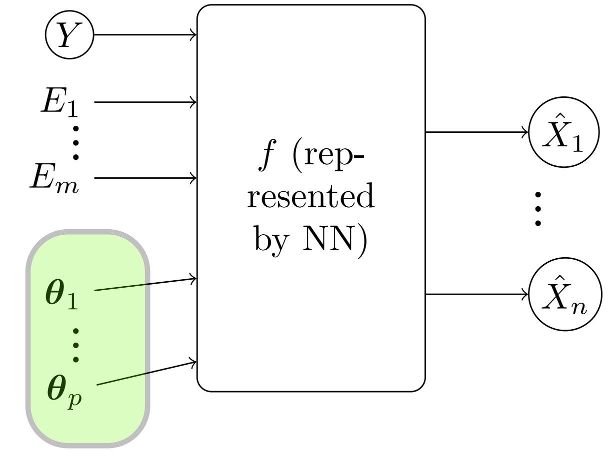
Problem	Unweight	CovS	TarS	LS-GeTarS
$TR_1 \to TS_2$	20.73%	20.73%	20.41%	**11.96%** ✓
$TR_2 \to TS_1$	26.36%	25.32%	26.28%	**13.56%** ✓

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

What Features/Components to Transfer?

- Invariant/transferrable causal mechanism (Zhang et al., 2013; 2014; Gong et al, 2016): invariance of $P(X^{ct} | Y)$
- Nonparametric transfer learning (Stojanov et al.2018a,b; Gong et al., 2018 & 2020; Zhang et al., 2020)
  - *Detect, model, utilize* changes

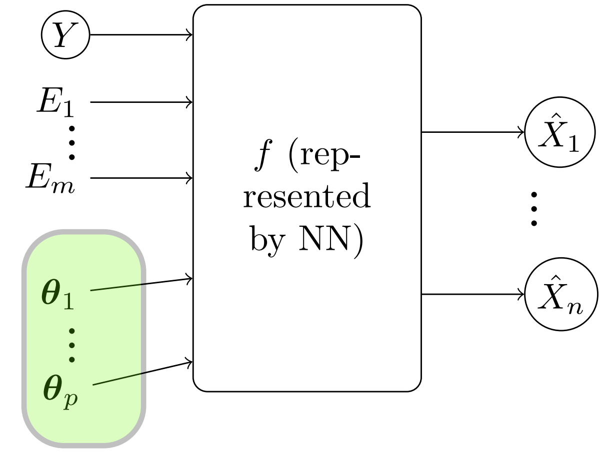
Gong, Zhang et al., Domain adaptation with **conditionally transferable components**, ICML 2016



On MNIST Data

- One source domain:  ...
- Target domain:  ...
- Learned parameter values θ: -0.24 (source, 0^{o}); 0.46 (target,45^{o})
- Generate new data with

On MNIST Data

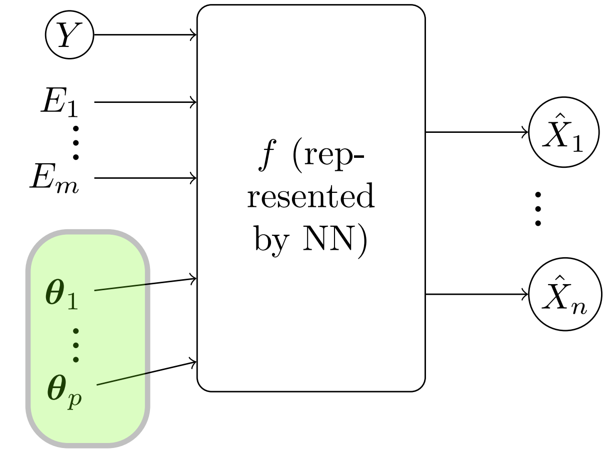


- One source domain:  …
- Target domain:  …
- Learned parameter values θ: -0.24 (source, 0°); 0.46 (target,45°)
- Generate new data with

θ values:

- 0.46
- 0.28
- 0.10
- -0.07
- -0.24

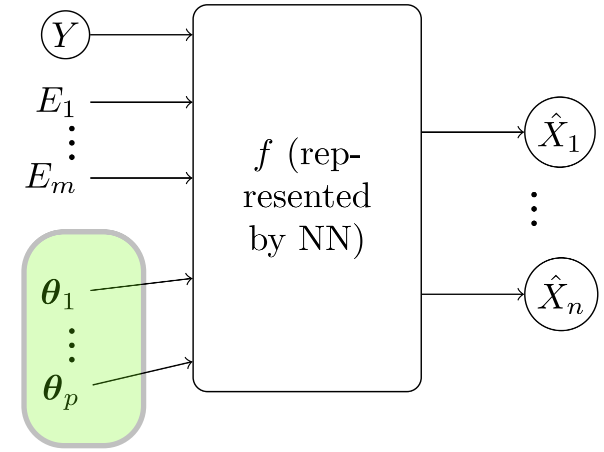
On MNIST Data



- One source domain:  ...
- Target domain:  ...
- Learned parameter values θ: -0.24 (source, 0^o); 0.46 (target,45^o)
- Generate new data with

θ values:
- 0.46
- 0.28
- 0.10
- -0.07
- -0.24

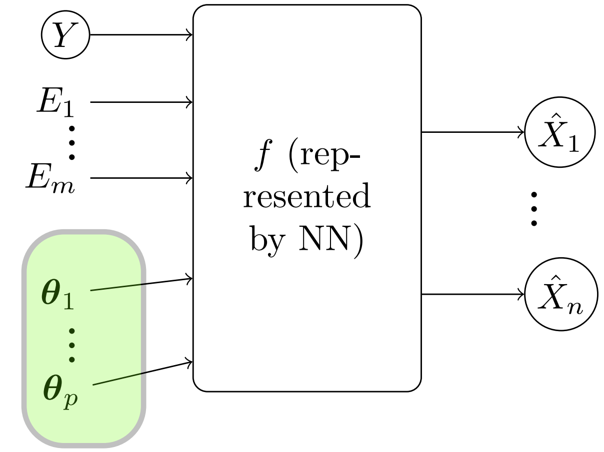
On MNIST Data



- One source domain:  ...
- Target domain:  ...
- Learned parameter values θ: -0.24 (source, 0^{o}); 0.46 (target,45^{o})
- Generate new data with

θ values:

- 0.46
- 0.28
- 0.10
- -0.07
- -0.24



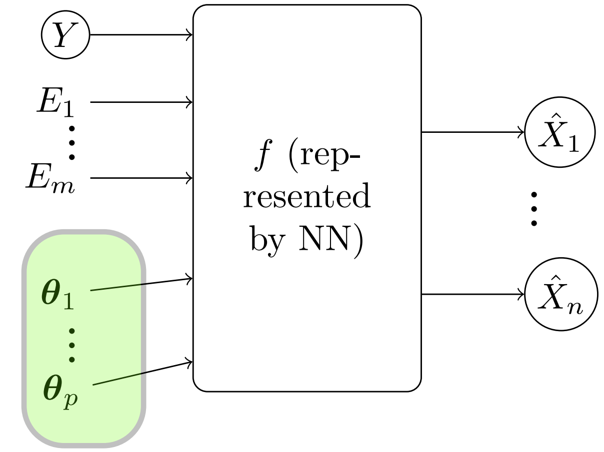
On MNIST Data

- One source domain:  ...
- Target domain:  ...
- Learned parameter values θ: -0.24 (source, 0°); 0.46 (target,45°)
- Generate new data with

θ values:

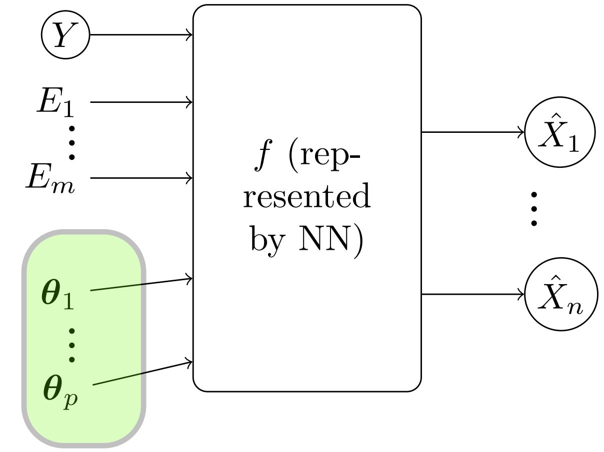
- 0.46
- 0.28
- 0.10
- -0.07
- -0.24

On MNIST Data



- One source domain:  …
- Target domain:  …
- Learned parameter values θ: -0.24 (source, 0º); 0.46 (target,45º)
- Generate new data with

θ values:
- 0.46
- 0.28
- 0.10
- -0.07
- -0.24



On MNIST Data

- One source domain:  ...
- Target domain:  ...
- Learned parameter values θ: -0.24 (source, 0°); 0.46 (target,45°)
- Generate new data with

θ values:

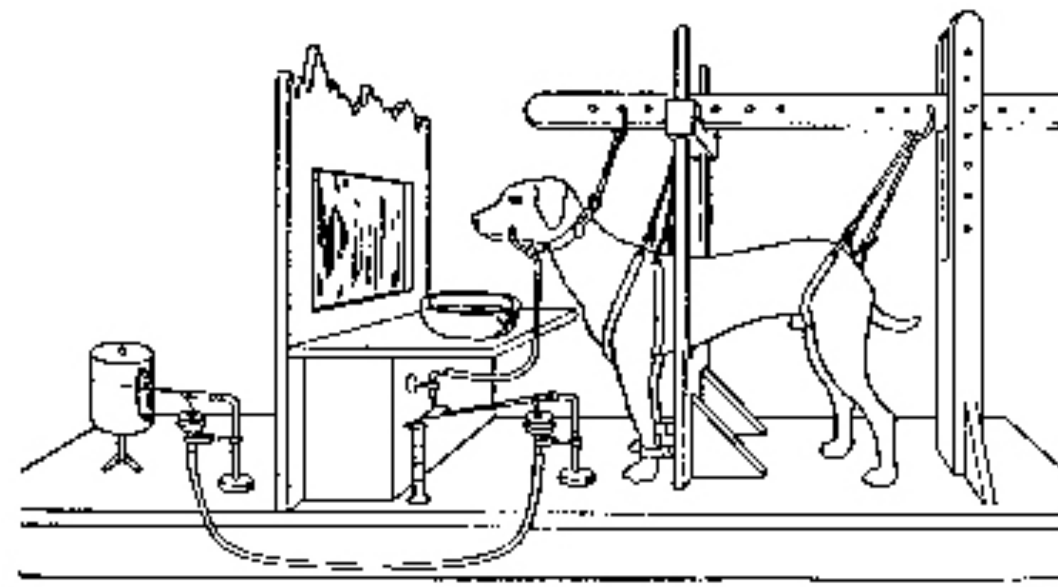
- 0.46
- 0.28
- 0.10
- -0.07
- -0.24

Causality & Transferability

- Causality helps
- One may find causal structure under rather **strong** assumptions
- But do we have to go to the causal level to achieve transferability?
  - Think about classical conditioning

Causality & Transferability

- Causality helps
- One may find causal structure under rather **strong** assumptions
- But do we have to go to the causal level to achieve transferability?
  - Think about classical conditioning



- *"If a particular stimulus in the dog's surroundings was present when the dog was given food then that stimulus could become associated with food and cause salivation on its own."*

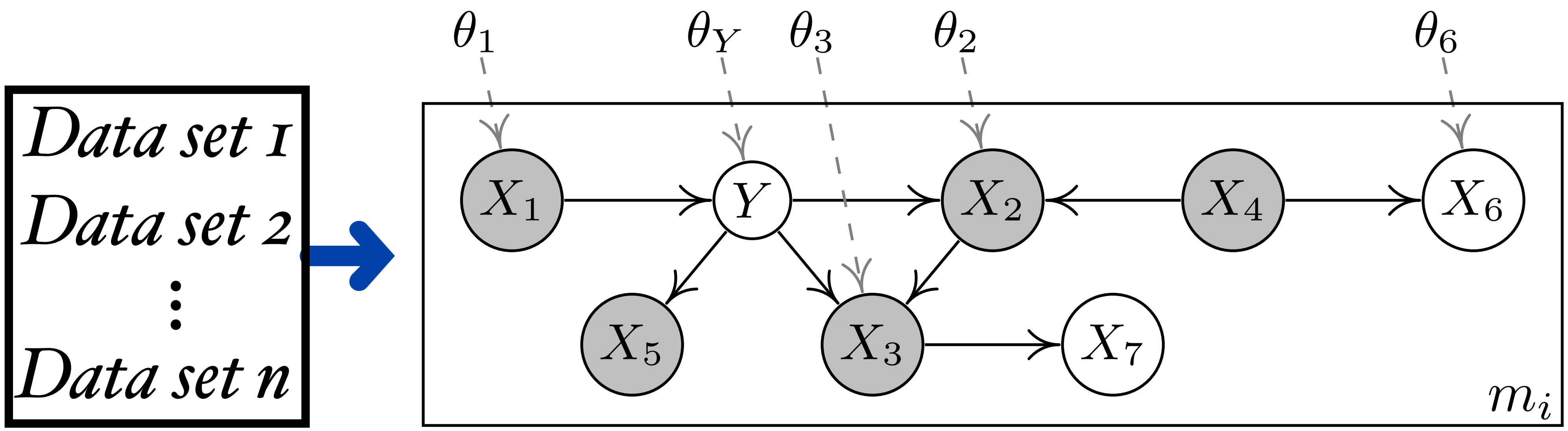
Automated Domain Adaptation

Data set 1
Data set 2
⋮
Data set n

- Discover properties of changes from source domain
- Represent them with an augmented graph
- Domain adaption is just a problem of inference on this graphical model

Zhang, Gong, et al., Domain adaptation as a problem of inference on graphical models, Arxiv 2020

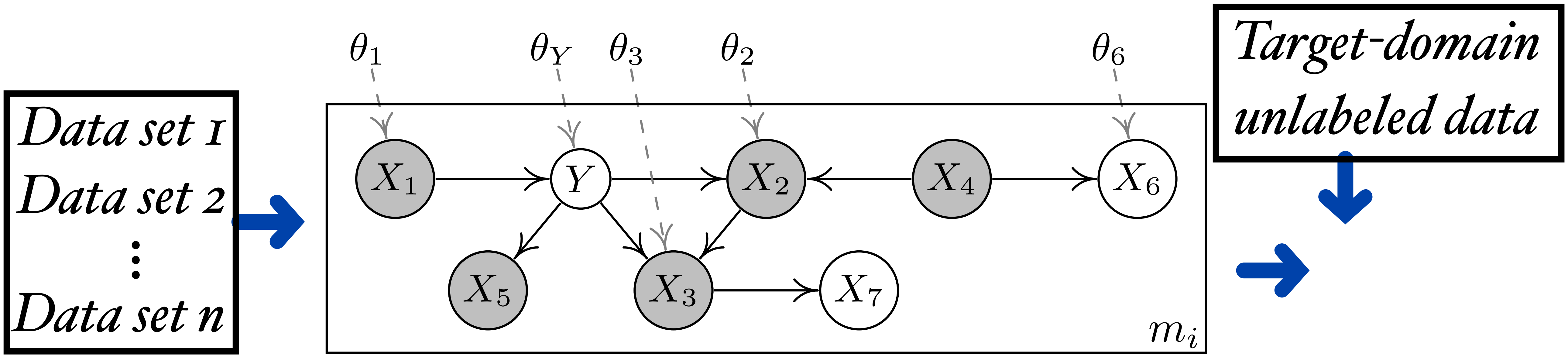
Automated Domain Adaptation



- Discover properties of changes from source domain
- Represent them with an augmented graph
- Domain adaption is just a problem of inference on this graphical model

Zhang, Gong, et al., Domain adaptation as a problem of inference on graphical models, Arxiv 2020

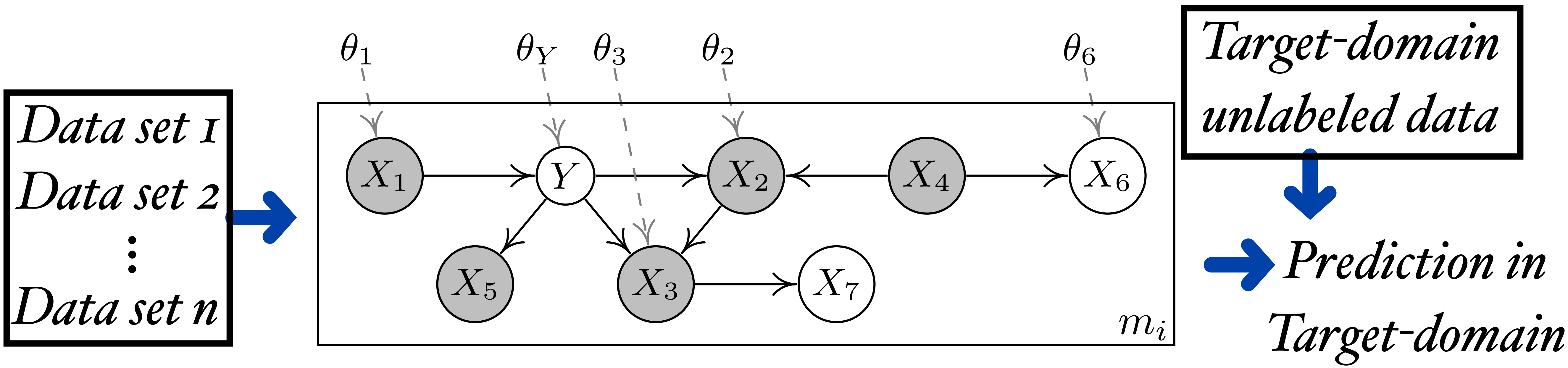
Automated Domain Adaptation



- Discover properties of changes from source domain
- Represent them with an augmented graph
- Domain adaption is just a problem of inference on this graphical model

Zhang, Gong, et al., Domain adaptation as a problem of inference on graphical models, Arxiv 2020

Automated Domain Adaptation



- Discover properties of changes from source domain
- Represent them with an augmented graph
- Domain adaption is just a problem of inference on this graphical model

Zhang, Gong, et al., Domain adaptation as a problem of inference on graphical models, Arxiv 2020

Summary

- Why causality? Why causality?
- Causal inference
- 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)
- Confounding, selection bias, temporal info…
- Transfer learning: compact description of changes
  - Modularity, independent changes…