Causality

Evgeny Burnaev
Rodrigo Rivera

Skoltech

BMJ

BMJ 2011;343:d7679 doi: 10.1136/bmj.d7679 (Published 15 December 2011)

RESEARCH

CHRISTMAS 2011: DEATH'S DOMINION

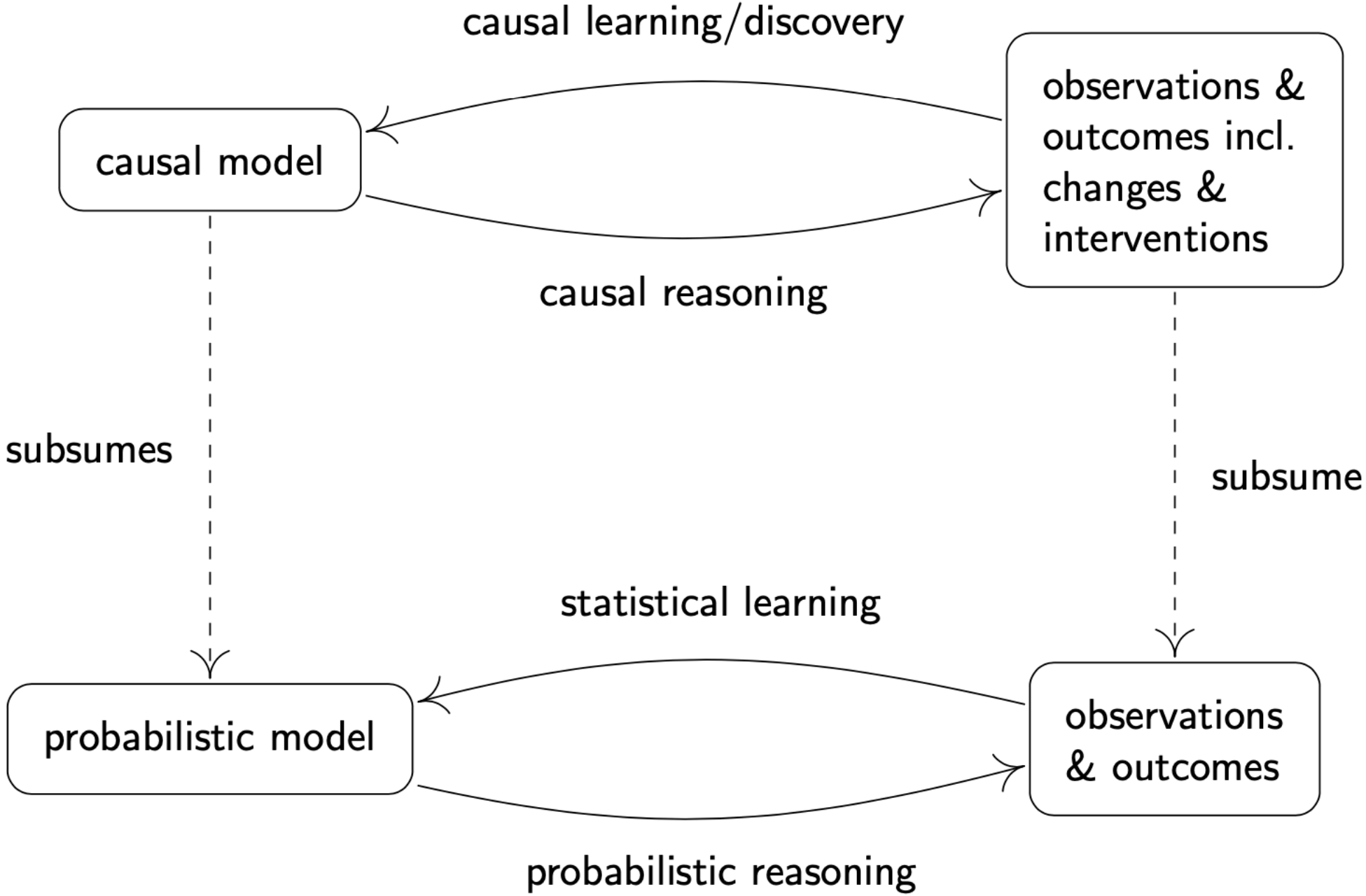
How fast does the Grim Reaper walk? Receiver operating characteristics curve analysis in healthy men aged 70 and over

Fiona F Stanaway *research fellow*[1], Danijela Gnjidic *research fellow*[2 3 4], Fiona M Blyth *deputy*

answer: $0.82 m/s$

(picture by Belle Mellor)

Part II: Causal Discovery


causal learning/discovery
causal model
observations & outcomes incl. changes & interventions
causal reasoning
subsumes
subsume
statistical learning
probabilistic model
observations & outcomes
probabilistic reasoning

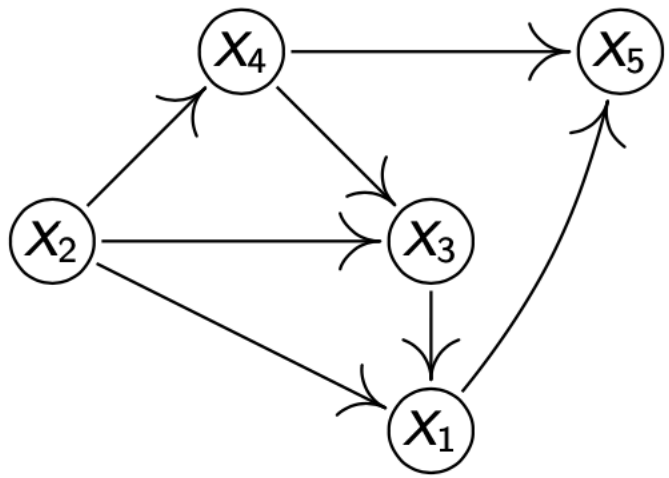
The Problem of Causal Discovery:

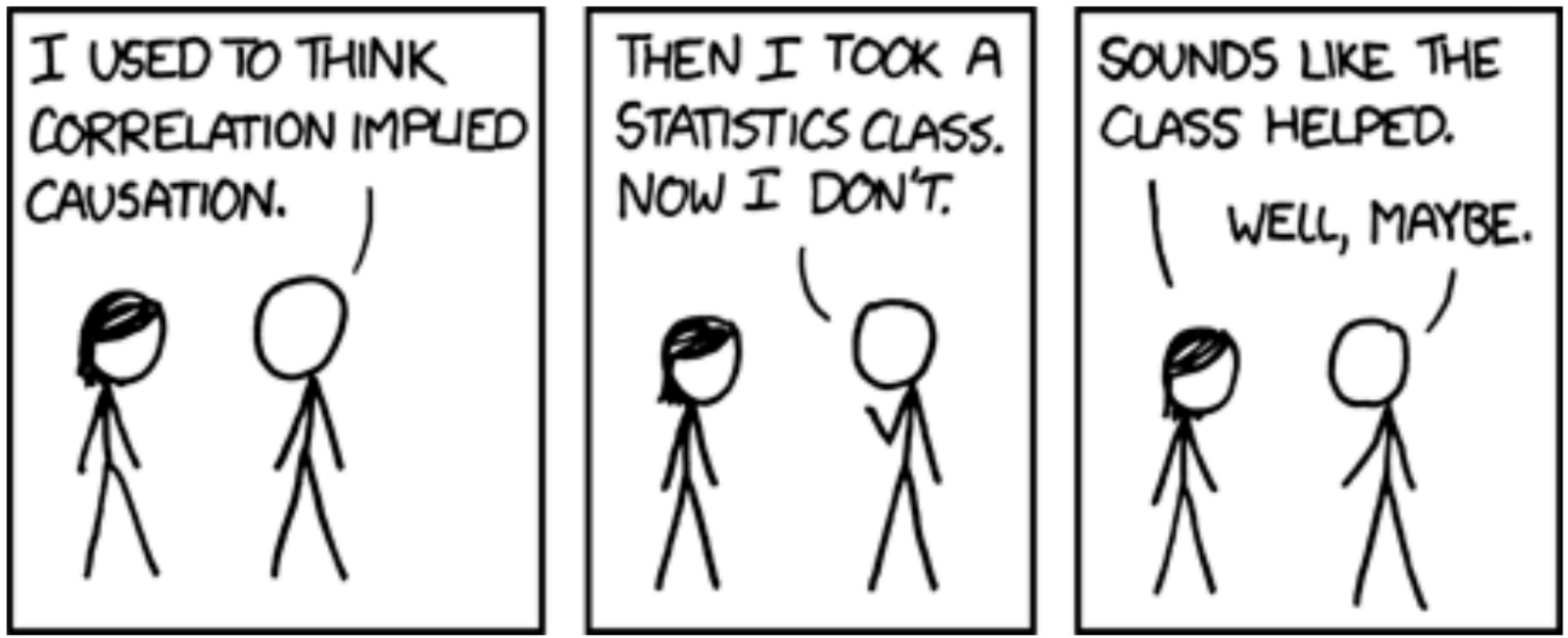
observed iid data
from $P(X_1, \ldots, X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮

$\xrightarrow{?}$

causal model, e.g. DAG $\mathcal{G}$




I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Correlation (Dependence) does not imply causation

Correlation (Dependence) does not imply causation ... but:

Correlation (Dependence) does not imply causation ... but:

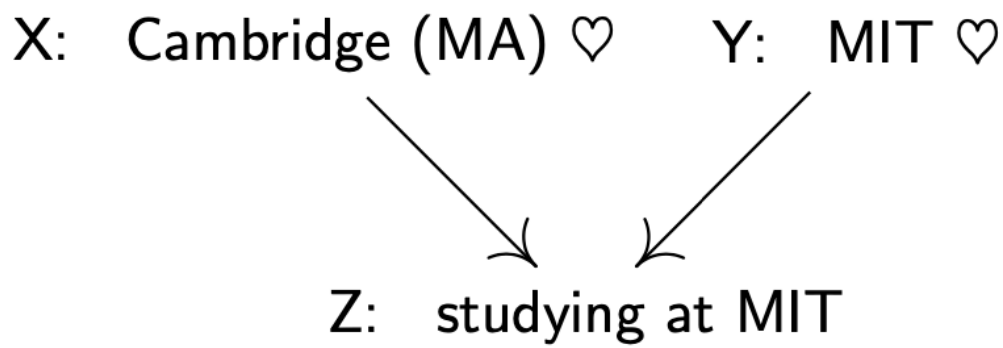
Reichenbach's common cause principle.
Assume that $X \not\perp\!\!\!\perp Y$. Then

- X "causes" Y,
- Y "causes" X,
- there is a hidden common "cause" or
- combination of the above.

Correlation (Dependence) does not imply causation ... but:

Reichenbach's common cause principle.
Assume that $X \not\perp\!\!\!\perp Y$. Then

- X "causes" Y,
- Y "causes" X,
- there is a hidden common "cause" or
- combination of the above.
- (In practice implicit conditioning also happens:

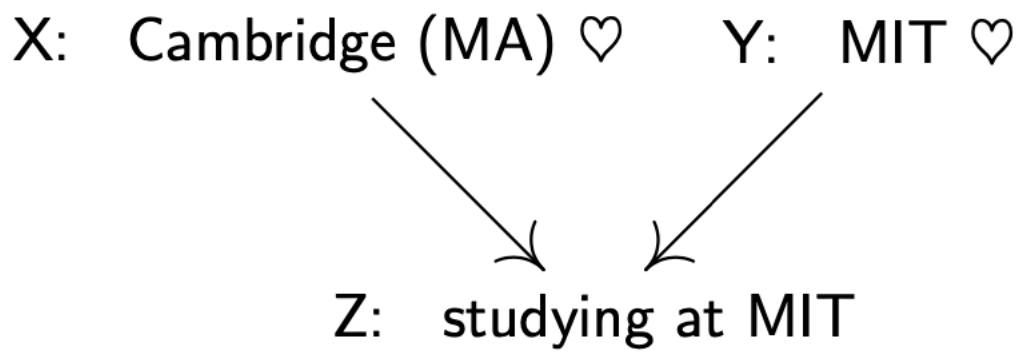


aka "selection bias").

Correlation (Dependence) does not imply causation ... but:

Reichenbach's common cause principle.
Assume that $X \not\perp\!\!\!\perp Y$. Then

- X "causes" Y,
- Y "causes" X,
- there is a hidden common "cause" or
- combination of the above.
- (In practice implicit conditioning also happens:

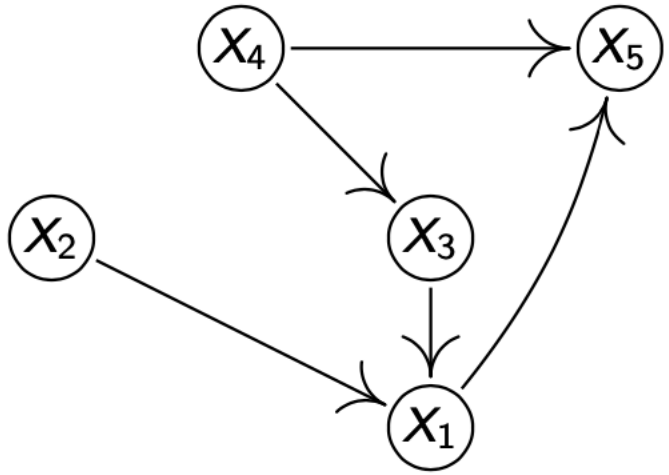


aka "selection bias"). Formalization of this idea...

Definition: graphs

$G = (V, E)$ with $E \subseteq V \times V$. The rest is as in real life!

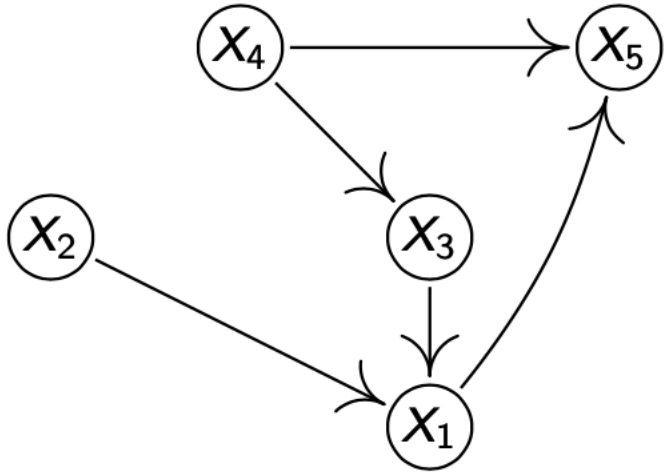
- parents, children, descendants, ancestors, ...
- paths, directed paths
- immoralities (or v-structures)
- d-separation (see next)
- ...



Definition: d-separation

X_i and X_j are *d*-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

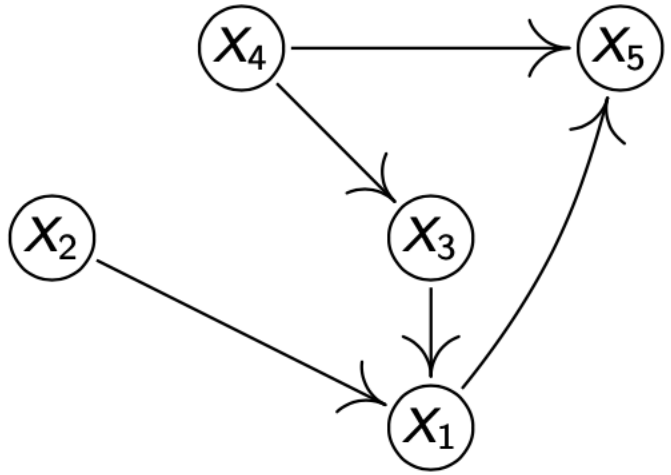
Check, whether all paths blocked!!



Definition: d-separation

X_i and X_j are *d*-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

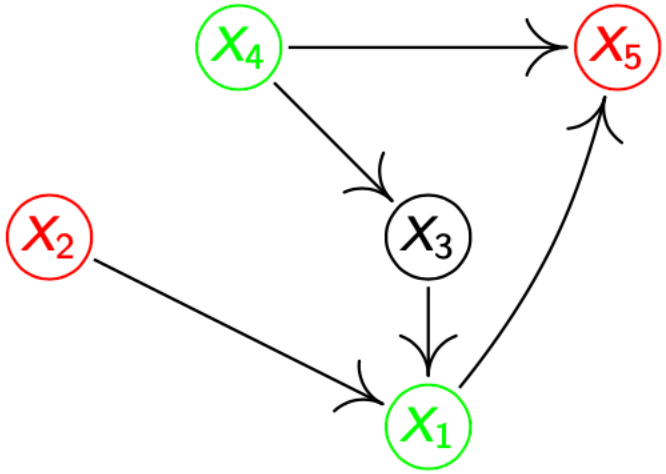
$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

Definition: d-separation

X_i and X_j are *d*-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

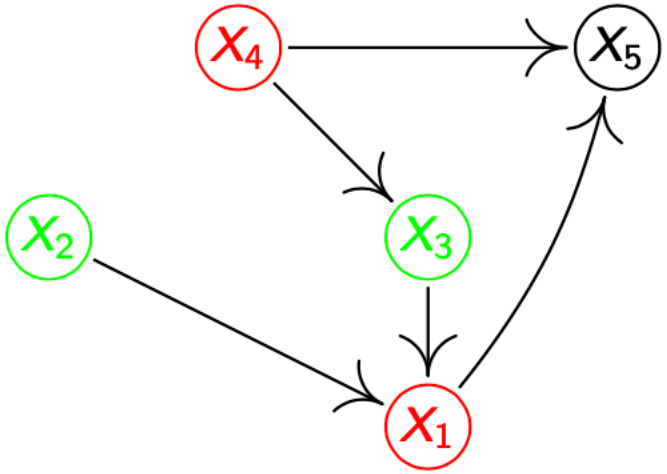
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are *d*-sep. by $\{X_1, X_4\}$

Definition: d-separation

X_i and X_j are *d*-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

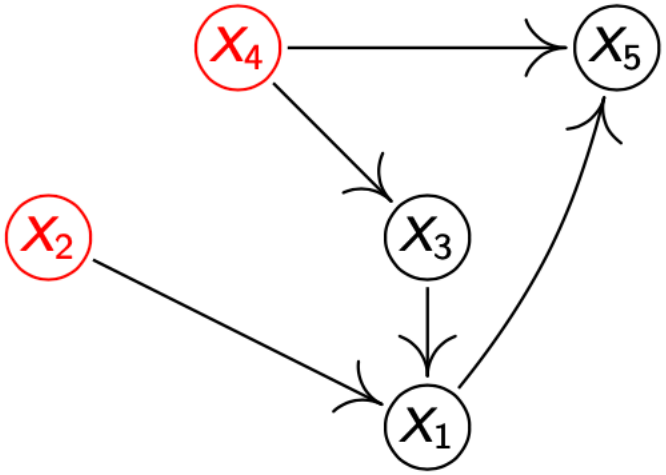
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are *d*-sep. by $\{X_1, X_4\}$
X_4 and X_1 are *d*-sep. by $\{X_2, X_3\}$

Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

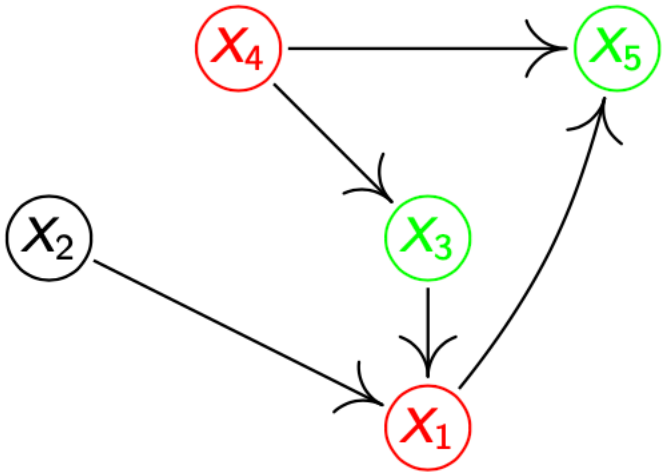
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$

Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

Definition

P is Markov w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Definition

P is Markov w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Proposition

Let the distribution P be Markov wrt a causal graph G. Then, Reichenbach's common cause principle is satisfied.

Proof: dependent variables must be d-connected.

There are three equivalent formulations of the Markov condition.

(i) **global Markov property**:

$$\mathbf{A} \text{ d-sep } \mathbf{B} \,|\, \mathbf{C} \text{ in } G \;\Rightarrow\; \mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$

(ii) **local Markov property**: each variable is independent of its non-descendants given its parents, and

(iii) **Markov factorization property**:

$$p(\mathbf{x}) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j \,|\, \mathbf{pa}_j^{\mathcal{G}}).$$

(assume existence of density)

Definition

P is Markov w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Definition

P is Markov w.r.t. G if

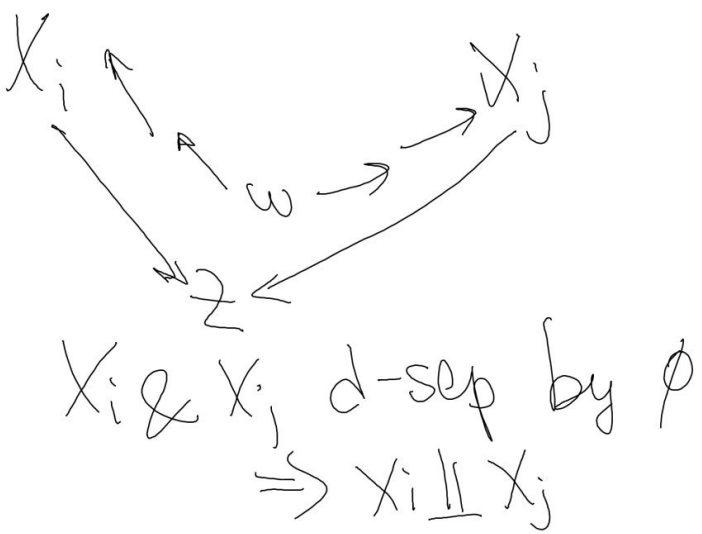
$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Definition

P is faithful w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Leftarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Examples...



Ex. Assume $\underline{P}$ is Markov & faithful wrt. $\mathbb{G}$. The following lists are complete

(i) $X \perp\!\!\!\perp Z$ (var: X, Y, Z)

(ii) $X \perp\!\!\!\perp Y \mid Z$ ($- \perp\!\!\!\perp -$)

(iii) $X \perp\!\!\!\perp Y$, $X \perp\!\!\!\perp W \mid Z$, $X \perp\!\!\!\perp W \mid Z, Y$

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

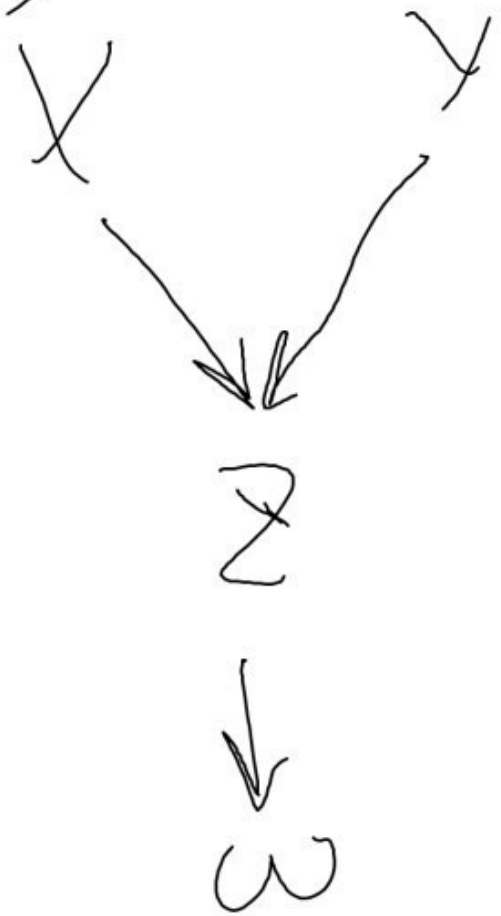
(i) $X \rightarrow Y \rightarrow Z$

(ii) $X \rightarrow Z \rightarrow Y$

OR $X \leftarrow Z \leftarrow Y$

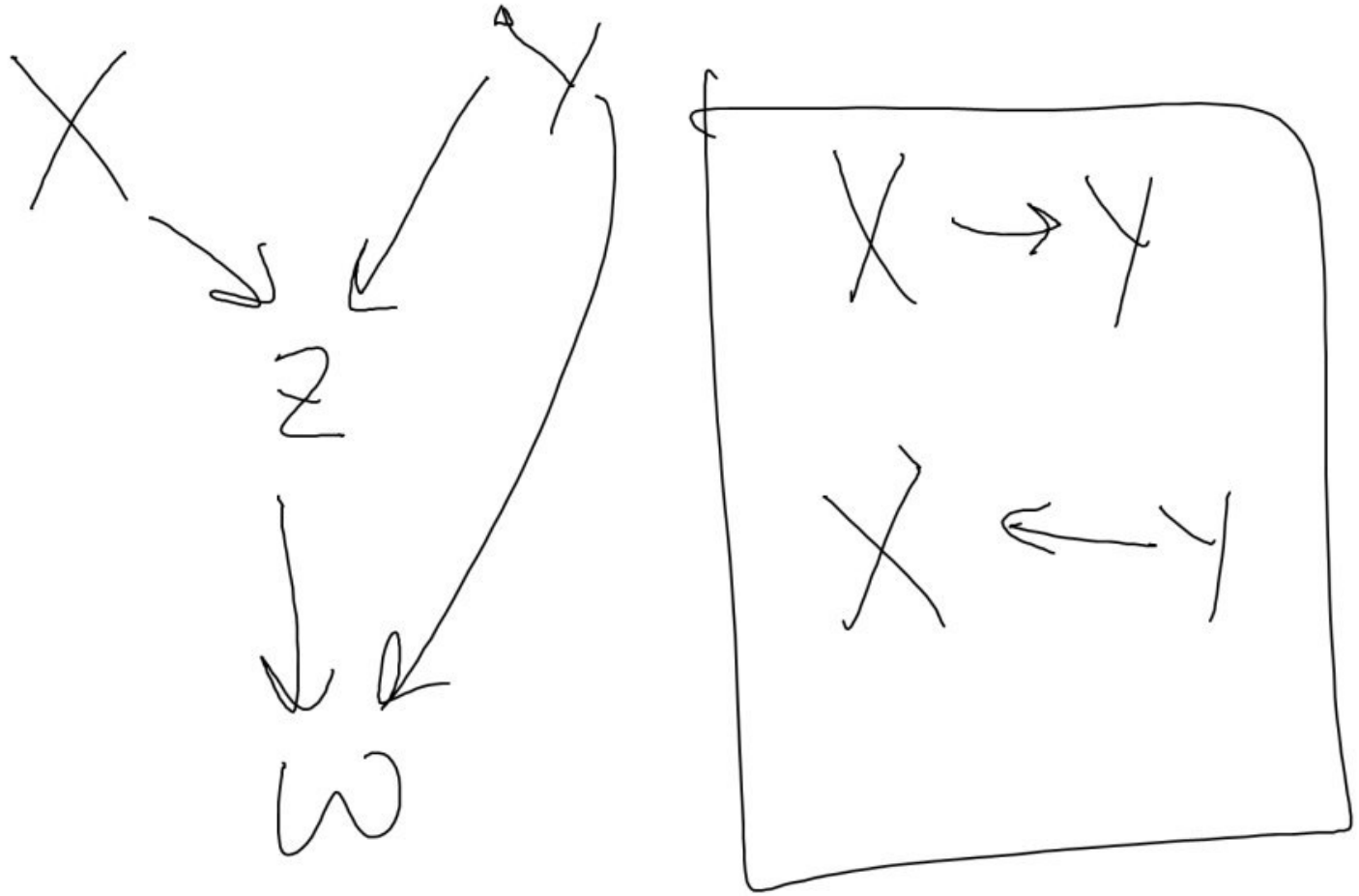
OR $X \leftarrow Z \rightarrow Y$

(iii)

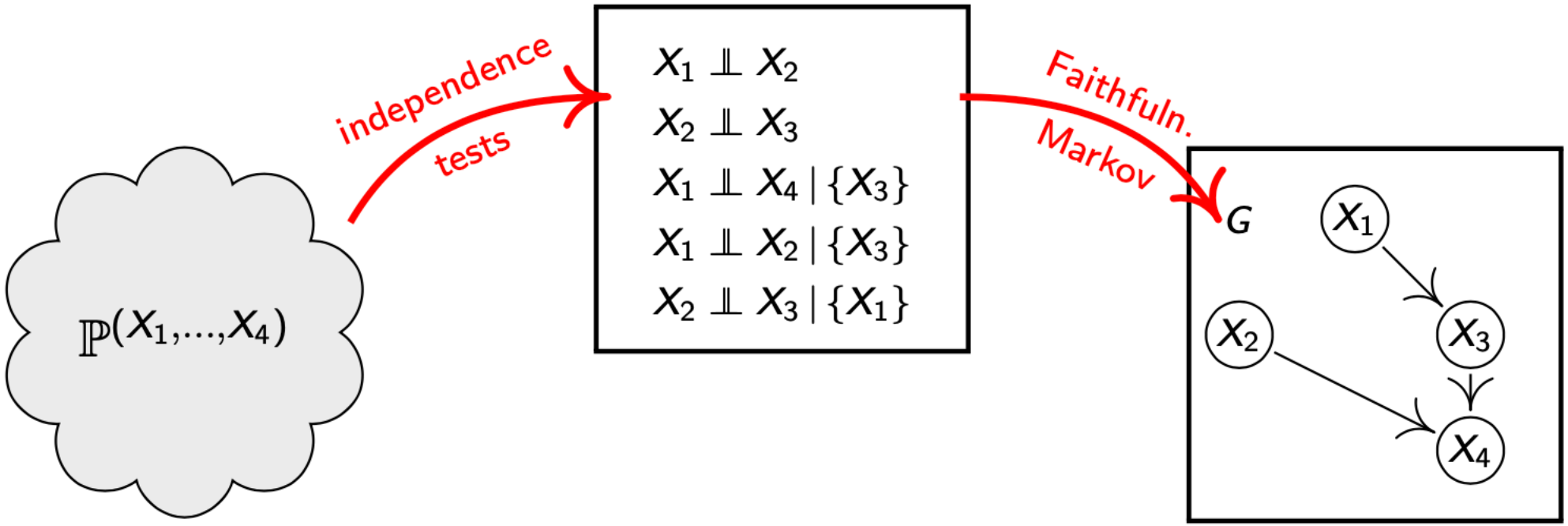


def G, H encoding same set of d-sep $S \rightsquigarrow$ "Markov equivalence" members of the same Markov equivalence class

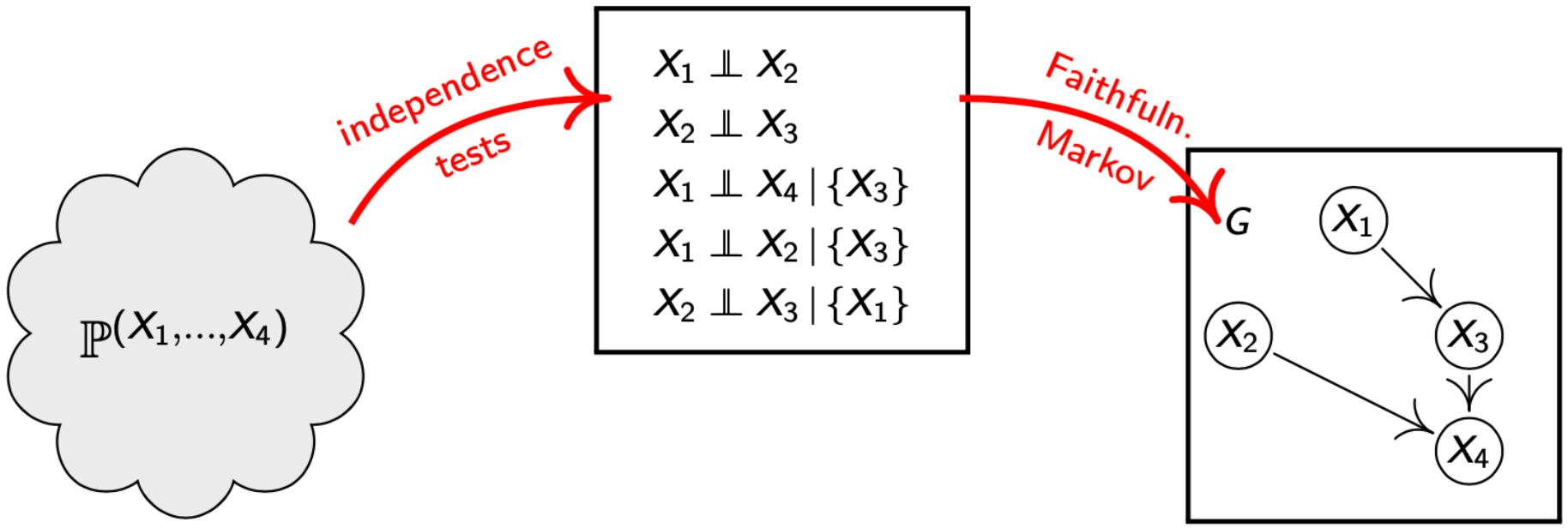
Prop: G, H Markov equivalent $\iff$ they have the same skeletons & the same v-structures


X
Y
Z
W
X → Y
X ← Y

Idea 1: independence-based methods



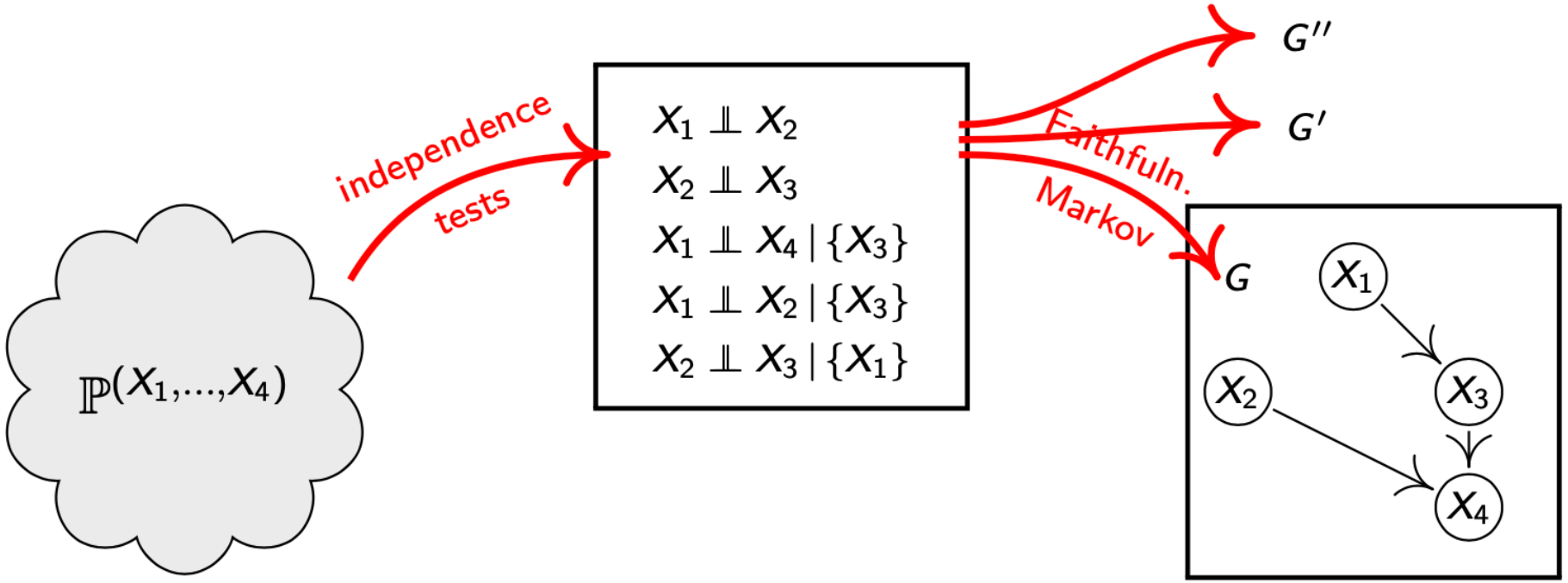
Idea 1: independence-based methods



Method: IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.

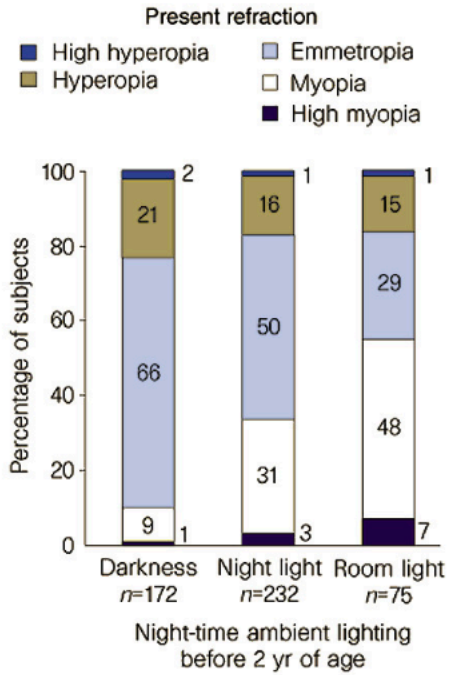
Idea 1: independence-based methods



Method: IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.

Example: myopia



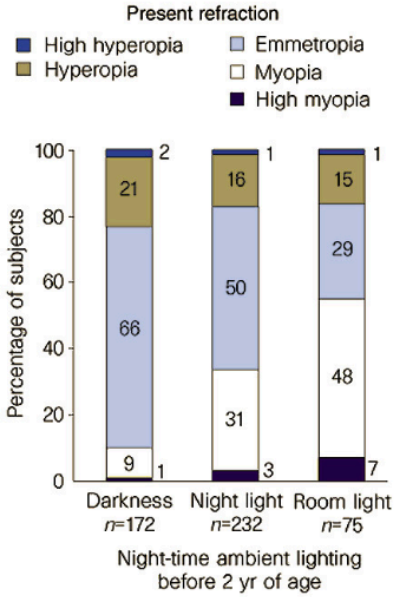
We have

- night light $\not\perp\!\!\!\perp$ child myopia
- night light $\perp\!\!\!\perp$ child myopia $\mid$ parent myopia
- no other independences

Quinn et al.: *Myopia and ambient lighting at night*, Nature 1999
Zadnik et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000
Gwiazda et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000

and therefore ...

Example: myopia

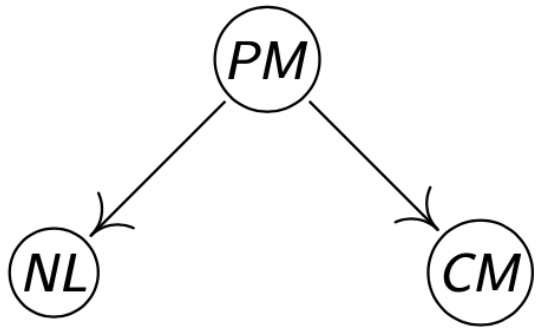


We have

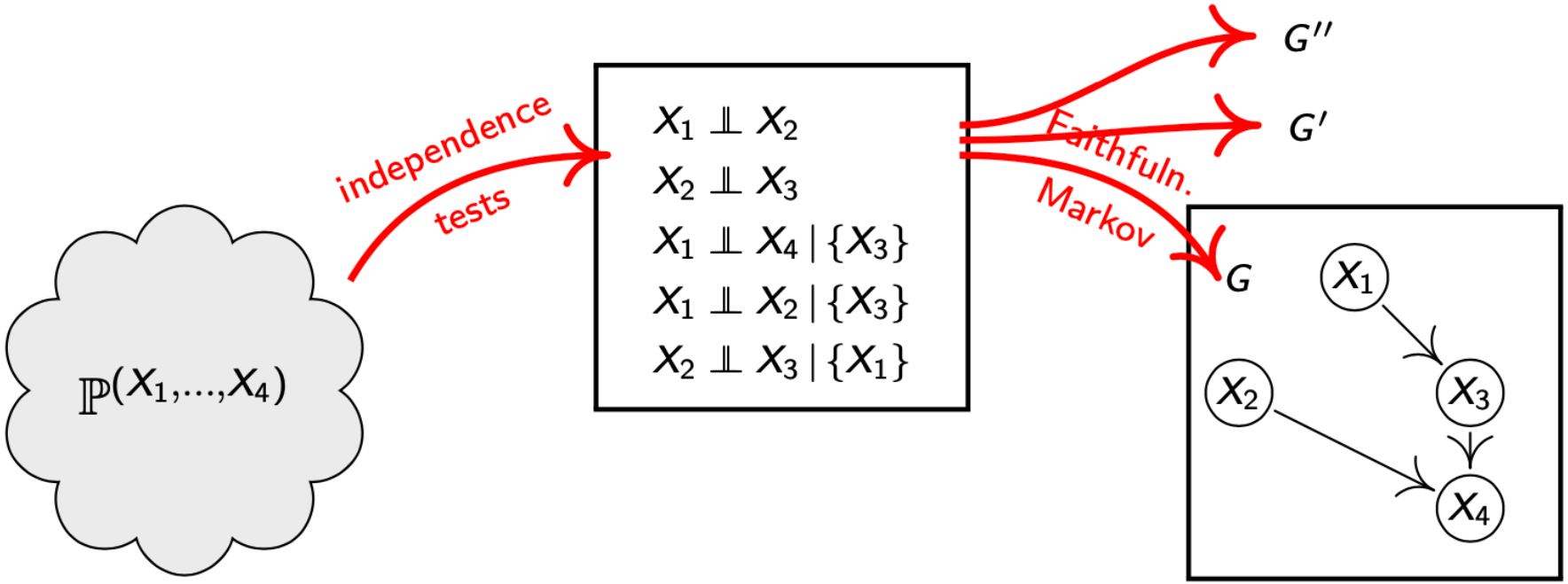
- night light $\not\perp\!\!\!\perp$ child myopia
- night light $\perp\!\!\!\perp$ child myopia $\mid$ parent myopia
- no other independences

Quinn et al.: *Myopia and ambient lighting at night*, Nature 1999
Zadnik et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000
Gwiazda et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000

and therefore ...



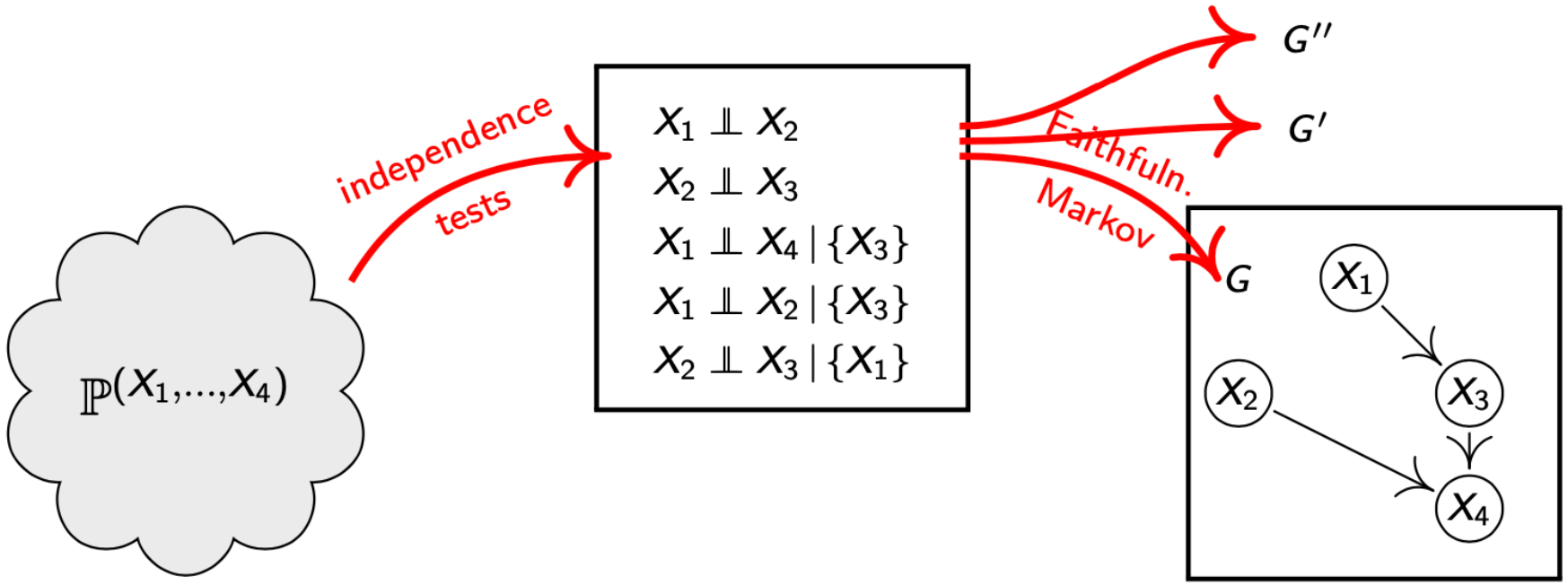
Idea 1: independence-based methods



Method: IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.

Idea 1: independence-based methods



Method: ~~IC (Pearl 2009);~~ PC, FCI (Spirtes et al., 2000)

1. ~~Find all (cond.) independences from the data.~~ Be smart.
2. Select the DAG(s) that corresponds to these independences.

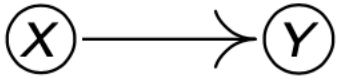
What do we do with two variables? (Nothing is possible in general.)

Idea 2: restricted structural causal models

Consider a distribution generated by

$$Y = \alpha X + N_Y$$

with N_Y, X ind.

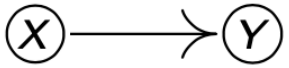


Idea 2: restricted structural causal models

Consider a distribution generated by

$$Y = \alpha X + N_Y$$

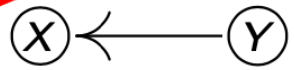
with N_Y, X ind.



Then, if (X, N_Y) is non-Gaussian, there is no

$$X = \beta Y + M_X$$

with M_X, Y ind.



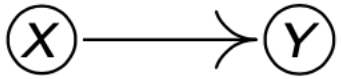
Shimizu et al. 2006

Idea 2: restricted structural causal models

Consider a distribution corresponding to

$$Y = 2X + N_Y$$

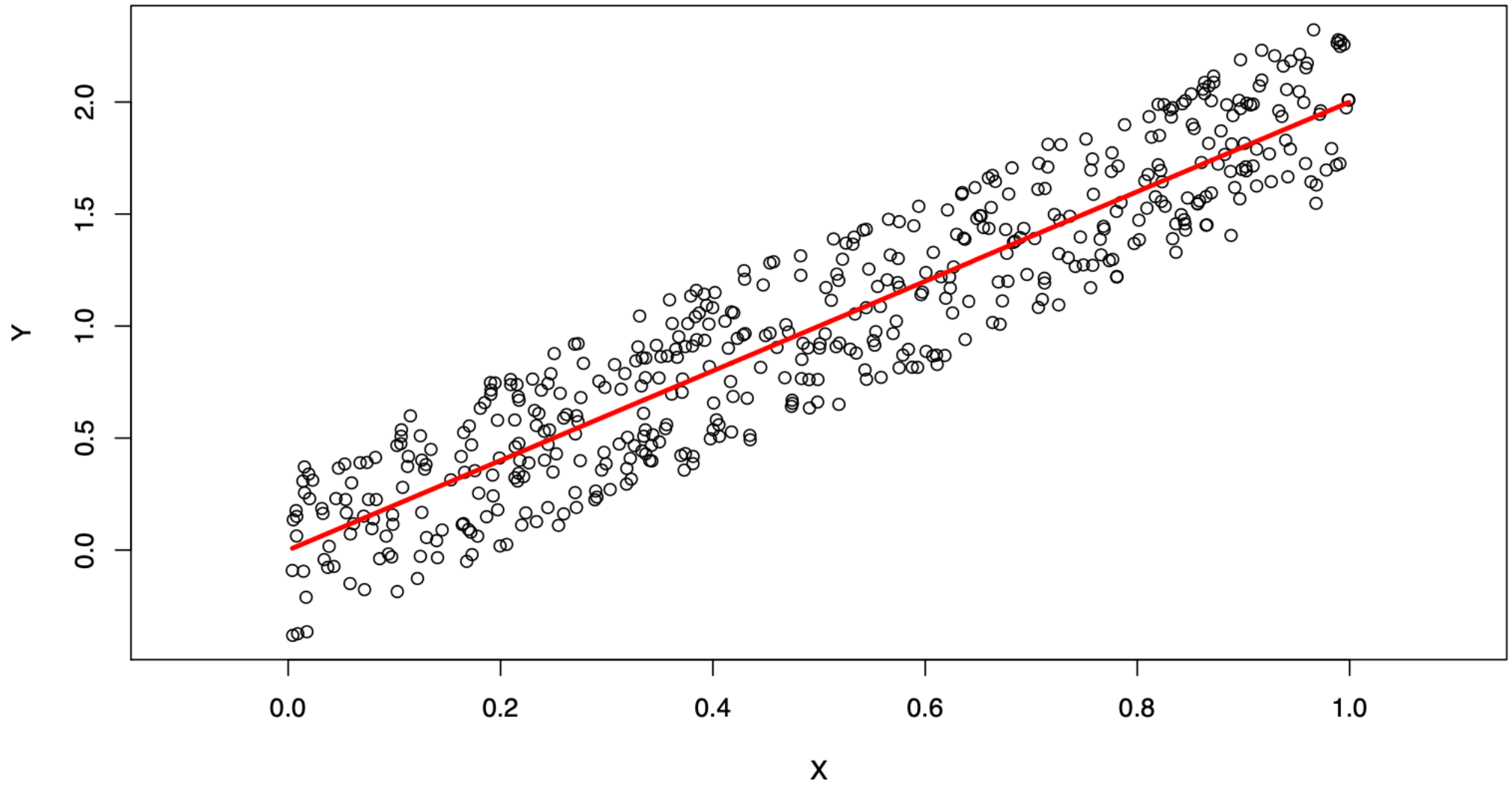
with $N_Y, X \overset{ind}{\sim} \mathcal{U}$



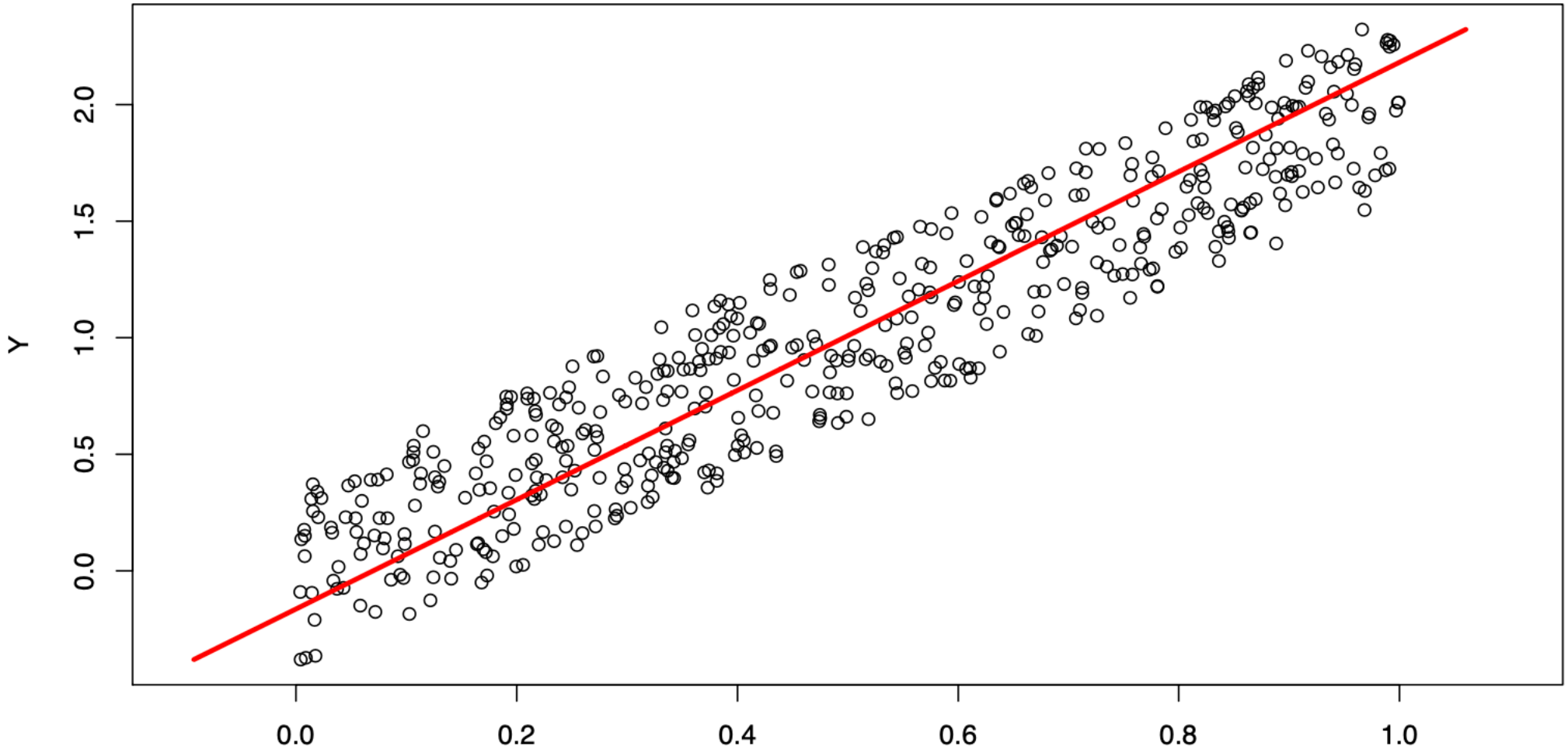
with

$$X \sim \mathcal{U}[-1, 1]$$
$$N_Y \sim \mathcal{U}[-0.4, 0.4]$$

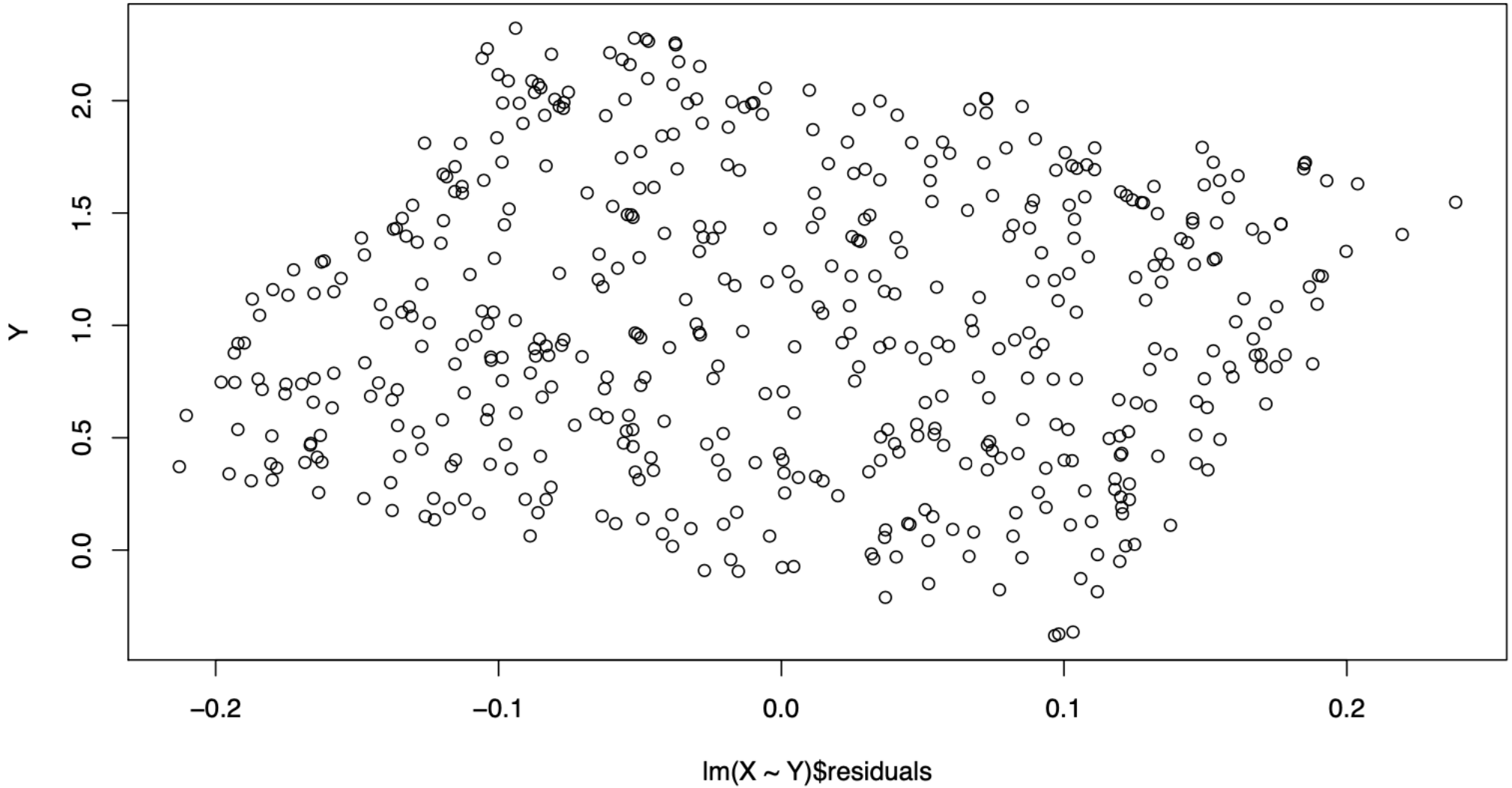
Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Idea 2: restricted structural causal models

Method…

Prop. $\forall P$ on (X, Y)

$\exists g, N_Y$, $\quad Y = g(X, N_Y)$

$N_Y \perp\!\!\!\perp X$

$\exists f, N_X$, $\quad X = f(Y, N_X)$

$N_X \perp\!\!\!\perp Y$

Prop. $\forall P$ on (X, Y)

$\exists g, N_Y$, $\quad Y = g(X, N_Y)$ $\quad N_Y \perp\!\!\!\perp X$

$\exists f, N_X$, $\quad X = f(Y, N_X)$ $\quad N_X \perp\!\!\!\perp Y$

inverse cdf of $X|Y$

uniform

Methods

1) Regress Y on X & check ind. between residuals and X

2) —"— X on Y

—"—

—"— and Y

$X \sim N_X$

$Y = X^3 + N_Y$

HSIC test :

	p-value
residuals vs. X	0.77
residuals vs. Y	0.004

Idea 2: restricted structural causal models

Theory…

Th.: Assume

$$Y = \alpha X + N_Y$$

$\exists\, \beta, N_X$

$$X = \beta Y + N_X$$

$\Rightarrow$ (X, N_Y) is Gaussian

Idea 2: restricted structural causal models

Peters et al ICML 2009 (univariate), Bauer et al ICML 2016 (multivariate)

Theorem

Let $(X_t)_t$ be a causal[a] solution of an $\mathrm{ARMA}(p, q)$ process:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}.$$

Then, X_t is time reversible, i.e., a causal solution of an $\mathrm{ARMA}(\tilde{p}, \tilde{q})$ process with reversed time, if and only if $(Z_t)_t$ is Gaussian.

[a] $(X_t)_t$ causal iff $Z_t \perp\!\!\!\perp X_{t-k}$, $k > 0$.

Idea 2: restricted structural causal models

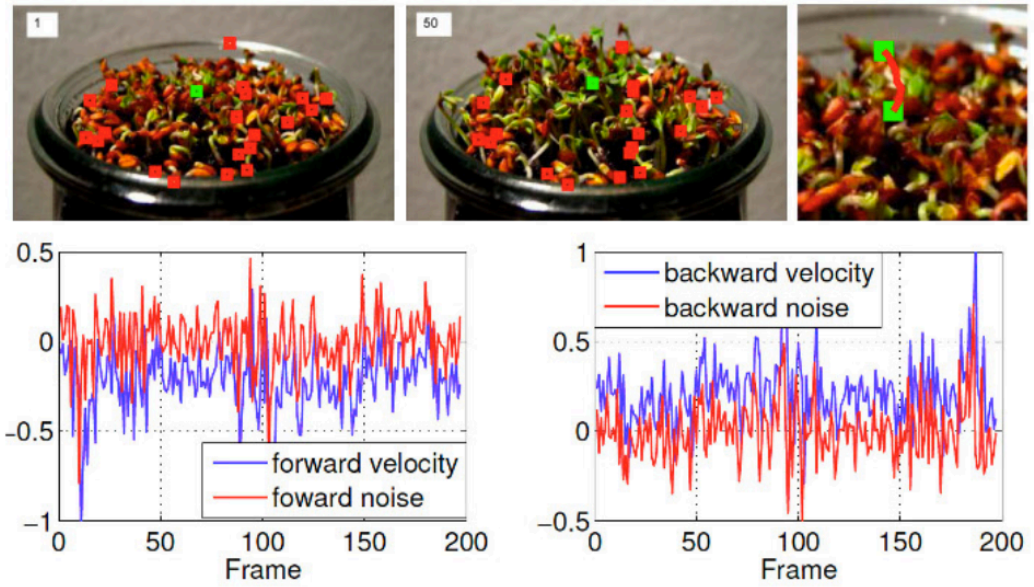
Pickup et al. 2014:

Idea 2: restricted structural causal models

Pickup et al. 2014:

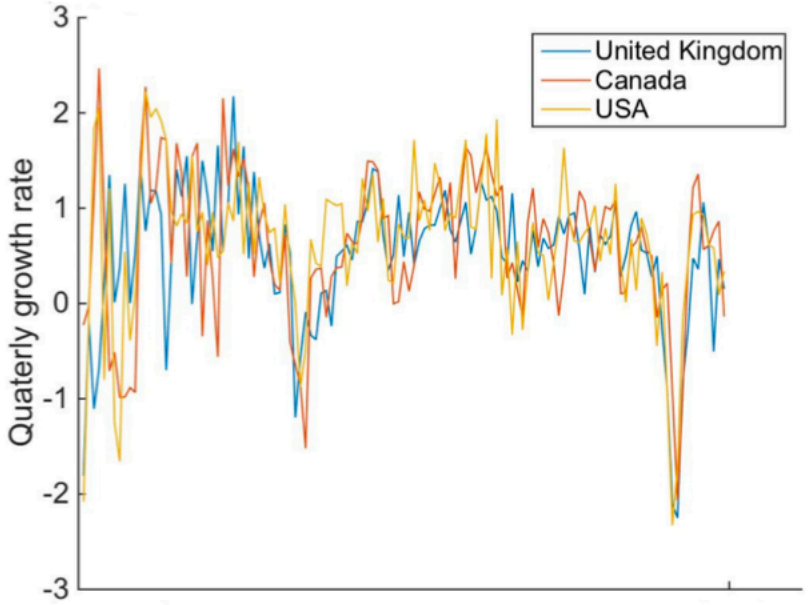
Method #3: Auto-regressive model

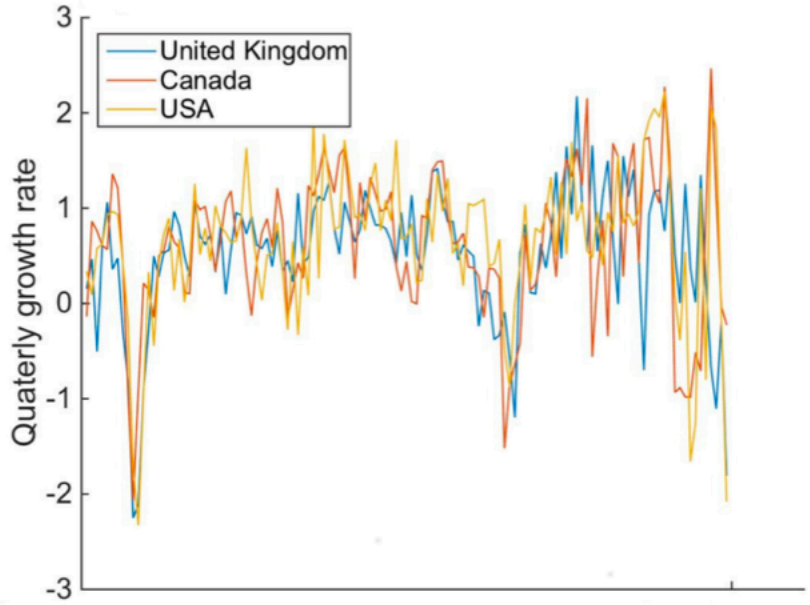
If object motion is linear, then the current velocity of the object should be affected only by the past. Noise on this motion will be asymmetric in the forward and backward directions, and fitting an auto-regressive model to the linear motion ought to yeild independence between the noise and signal only in the forwards-time direction. This method attempts to find the forward direction by looking at the independence of AR fitting error on motion trajectories.



Top: tracked points from a sequence, and an example track. Bottom: Forward-time (left) and backward-time (right) vertical trajectory components, and the corresponding model residuals. Trajectories should be independent from model residuals (noise) in the forward-time direction only. For the example track shown, p-values for the forward and backward directions are 0.52 and 0.016 respectively, indicating that forwards time is more likely.

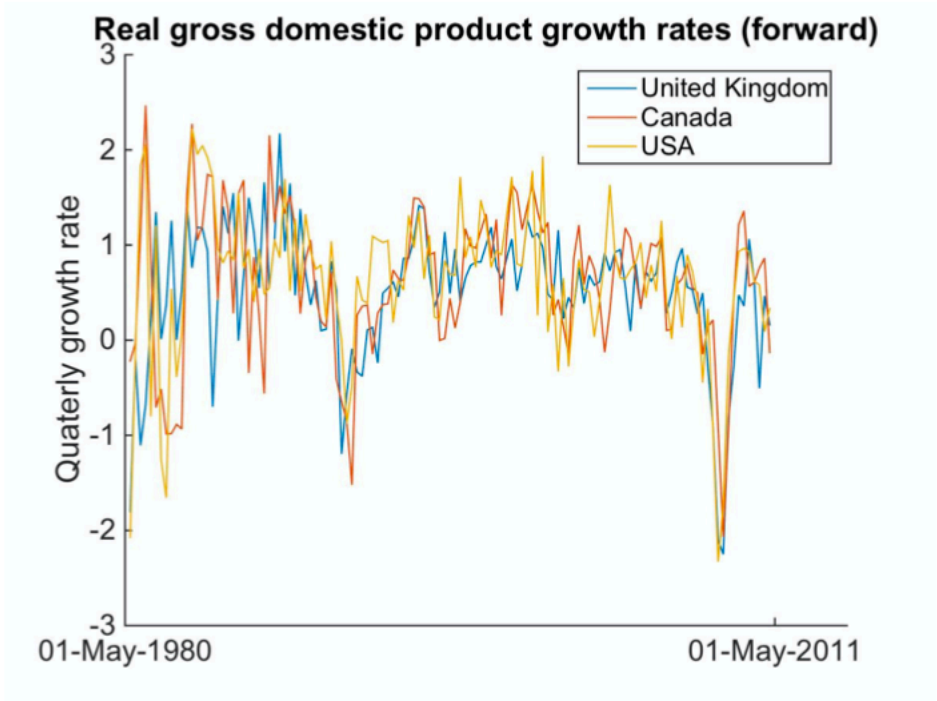
Idea 2: restricted structural causal models

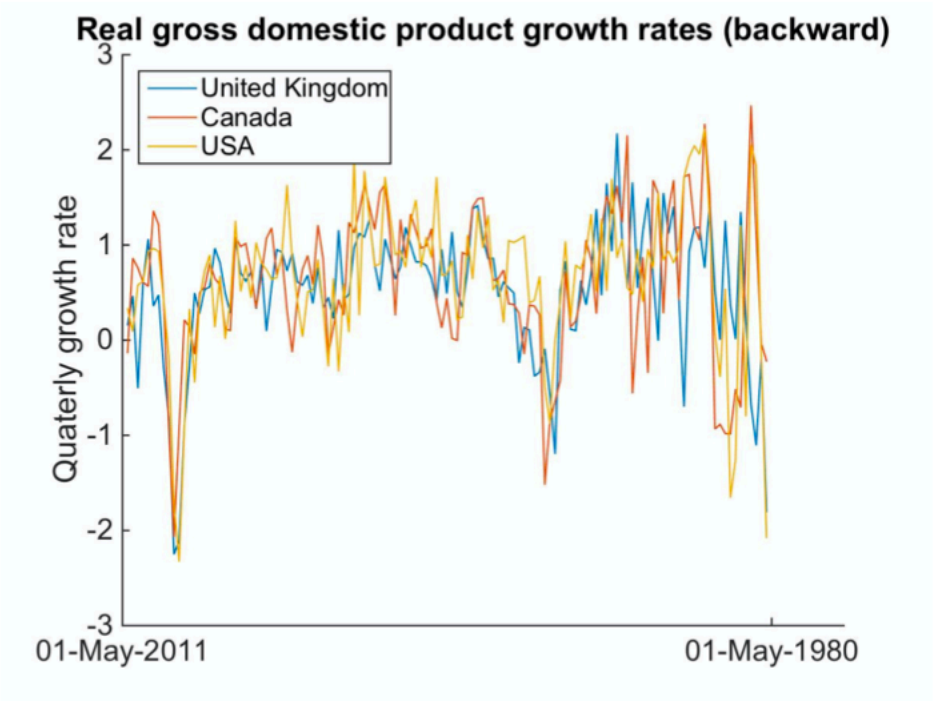




Quarterly growth rates in percentage of GDP for the UK, Canada and USA (Tsay et al 2014).

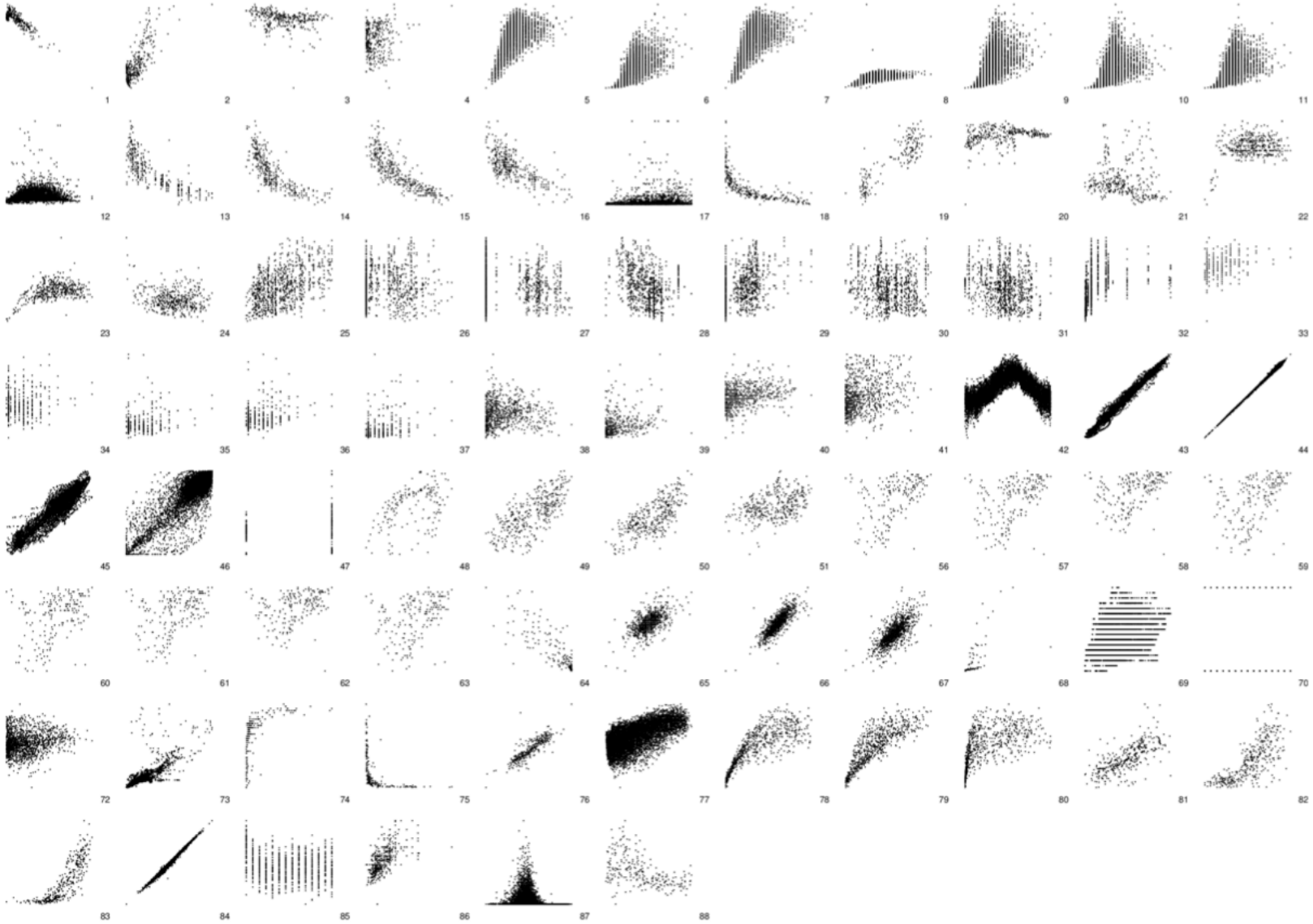
Idea 2: restricted structural causal models





Quarterly growth rates in percentage of GDP for the UK, Canada and USA (Tsay et al 2014).

Idea 2: restricted structural causal models



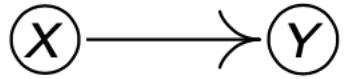
Mooij, JP, Janzing, Zscheischler, Schölkopf: *Disting. cause from effect using obs. data: methods and benchm.*, JMLR 2016

Idea 2: restricted structural causal models

Consider a distribution entailed by

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

with $N_Y, X \overset{ind}{\sim} \mathcal{N}$

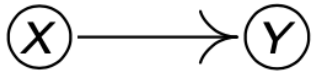


Idea 2: restricted structural causal models

Consider a distribution entailed by

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

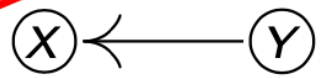
with $N_Y, X \overset{ind}{\sim} \mathcal{N}$



Then, if f is nonlinear, there is no

$$X = g(Y) + M_X$$

with $M_X, Y \overset{ind}{\sim} \mathcal{N}$



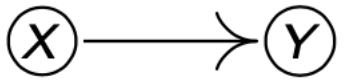
JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*

Idea 2: restricted structural causal models

Consider a distribution corresponding to

$$Y = X^3 + N_Y$$

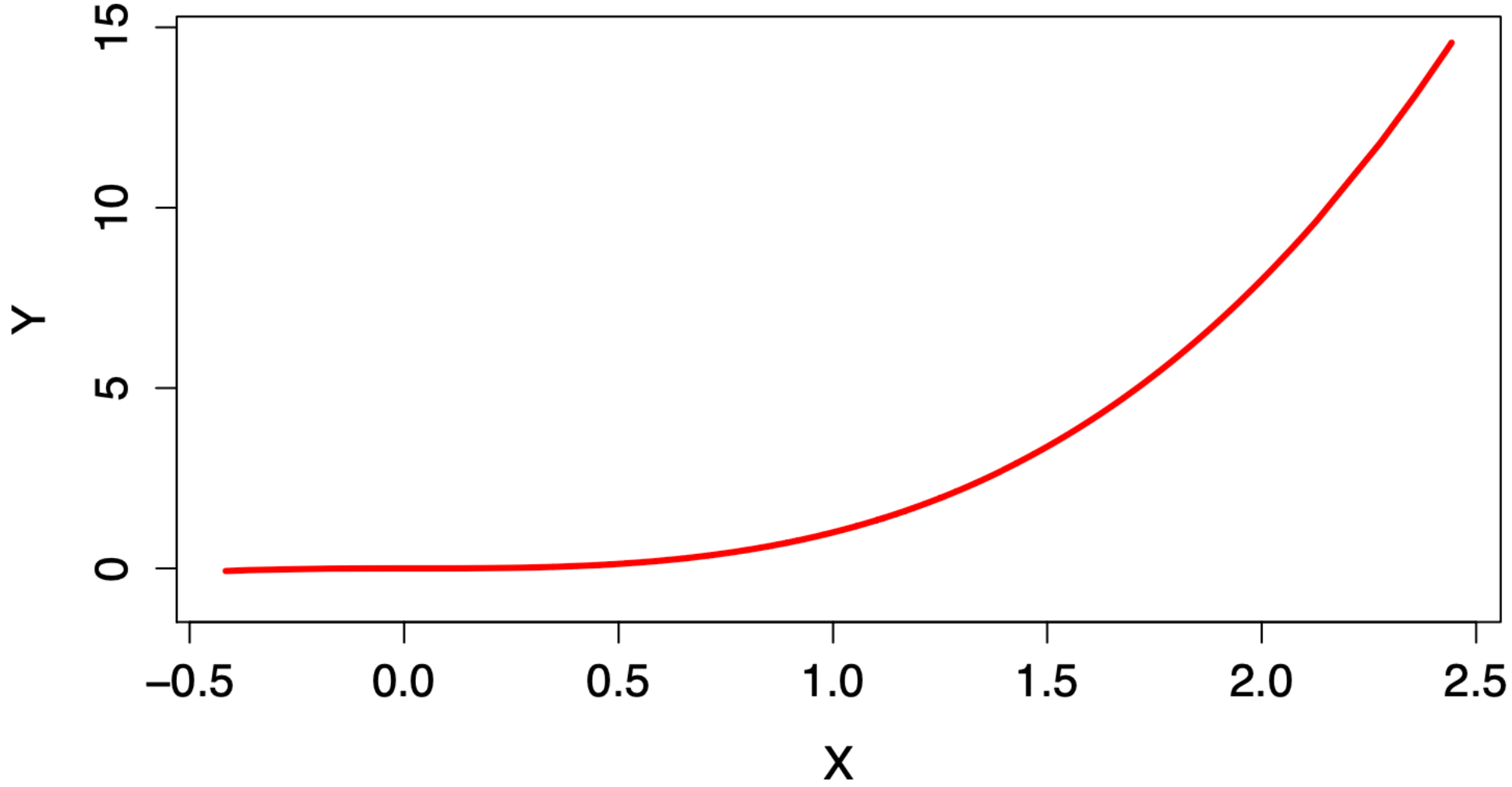
with $N_Y, X \overset{ind}{\sim} \mathcal{N}$



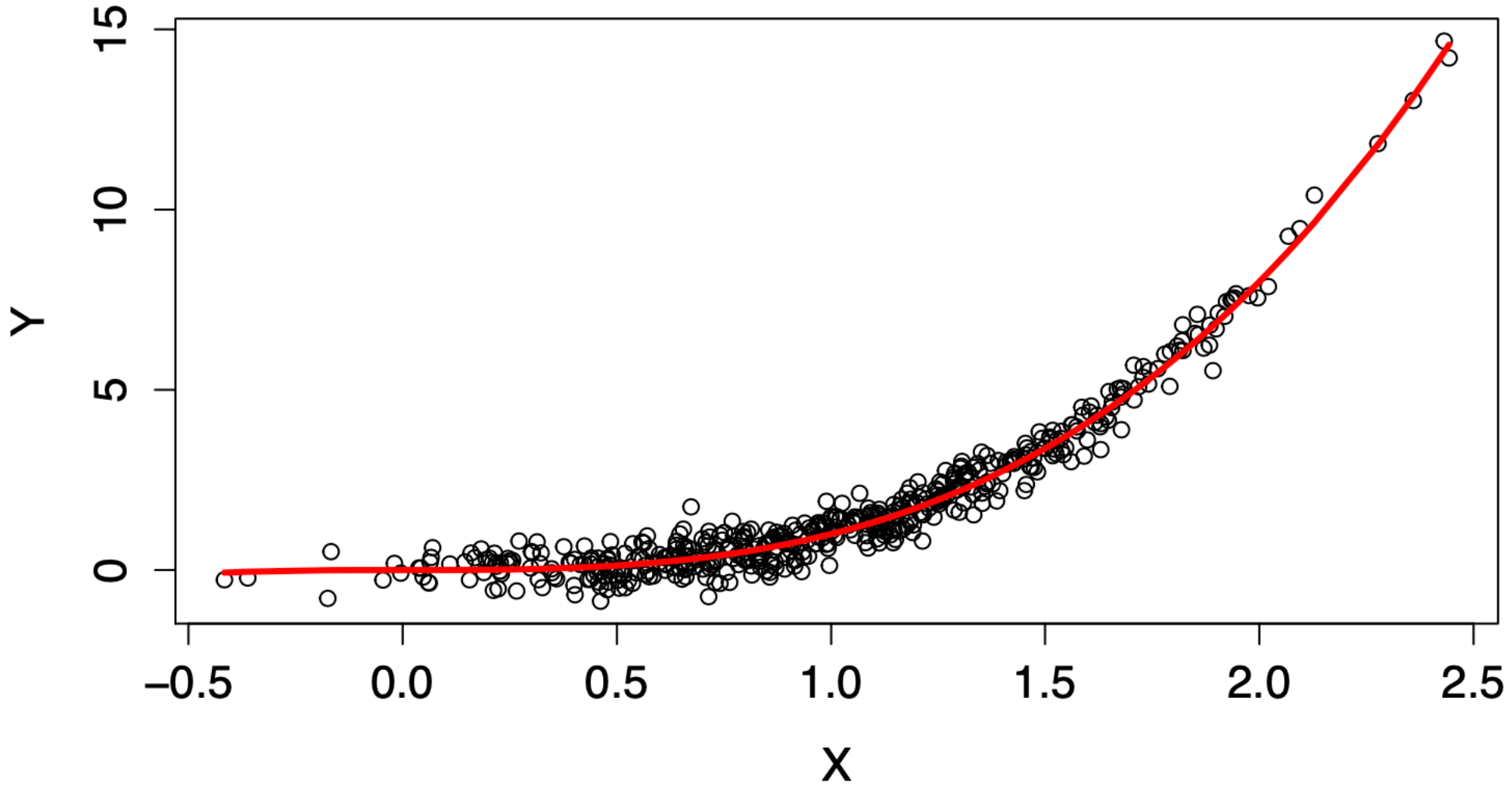
with

$$X \sim \mathcal{N}(1, 0.5^2)$$
$$N_Y \sim \mathcal{N}(0, 0.4^2)$$

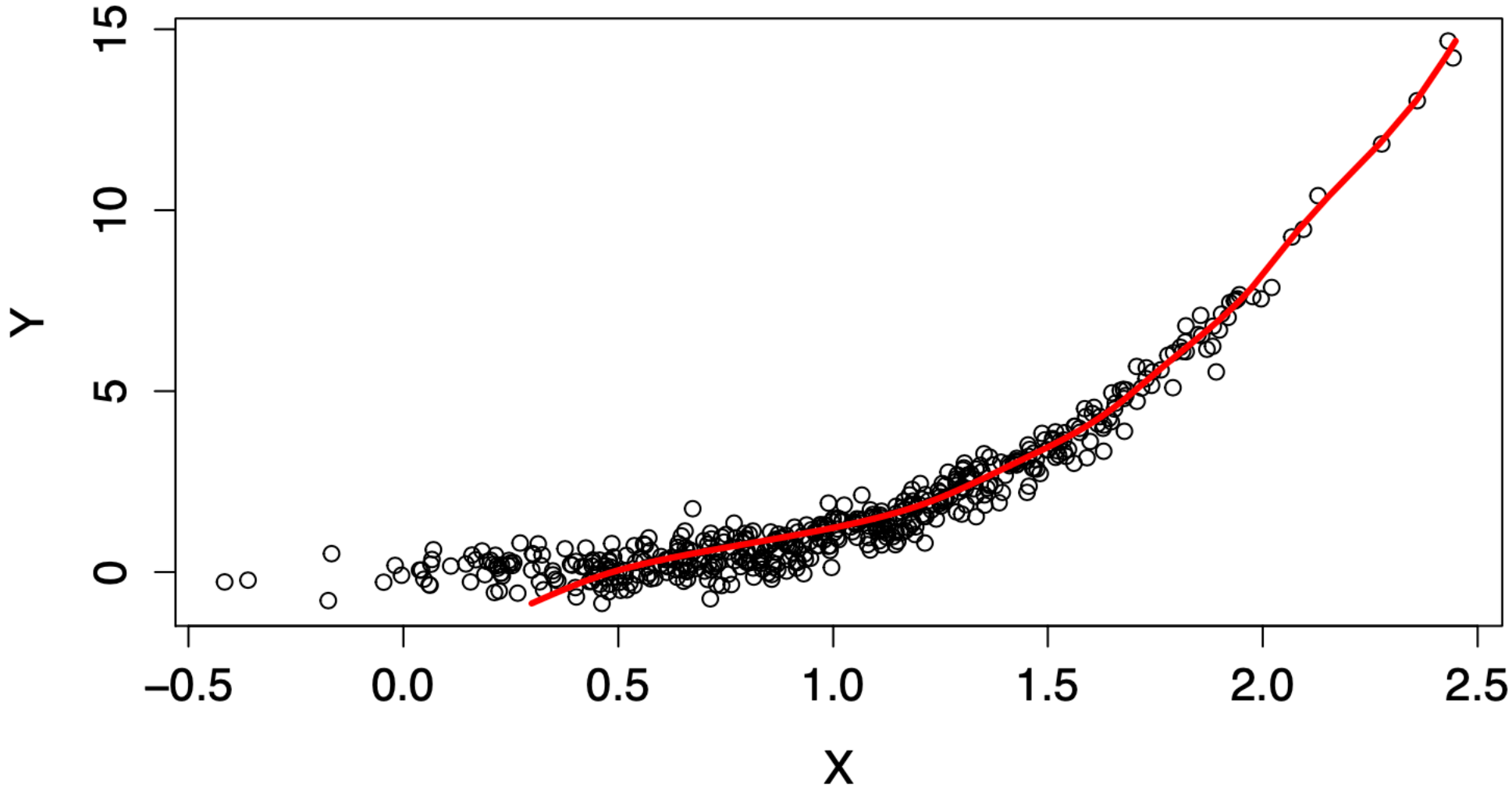
Idea 2: restricted structural causal models



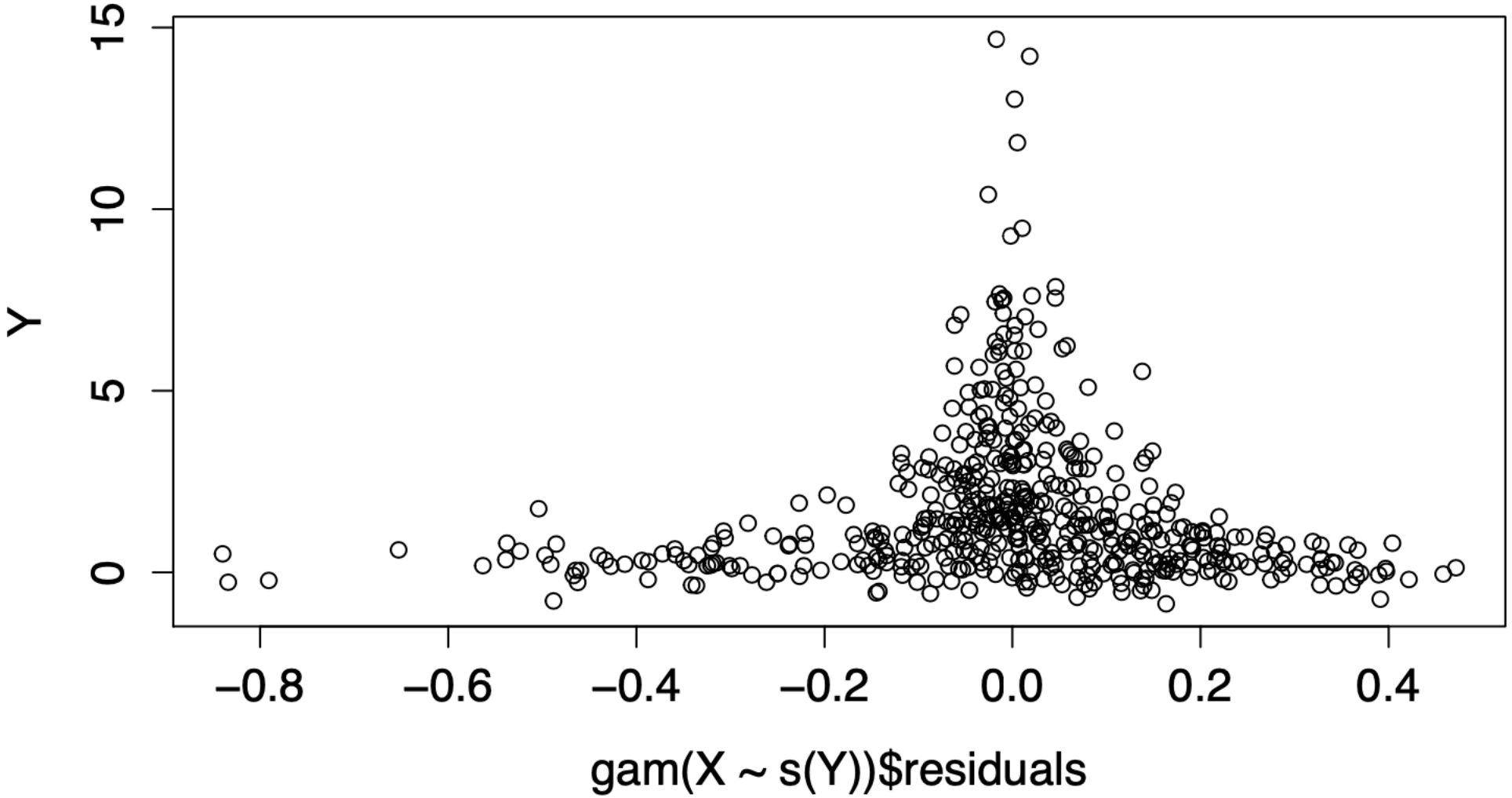
Idea 2: restricted structural causal models



Idea 2: restricted structural causal models

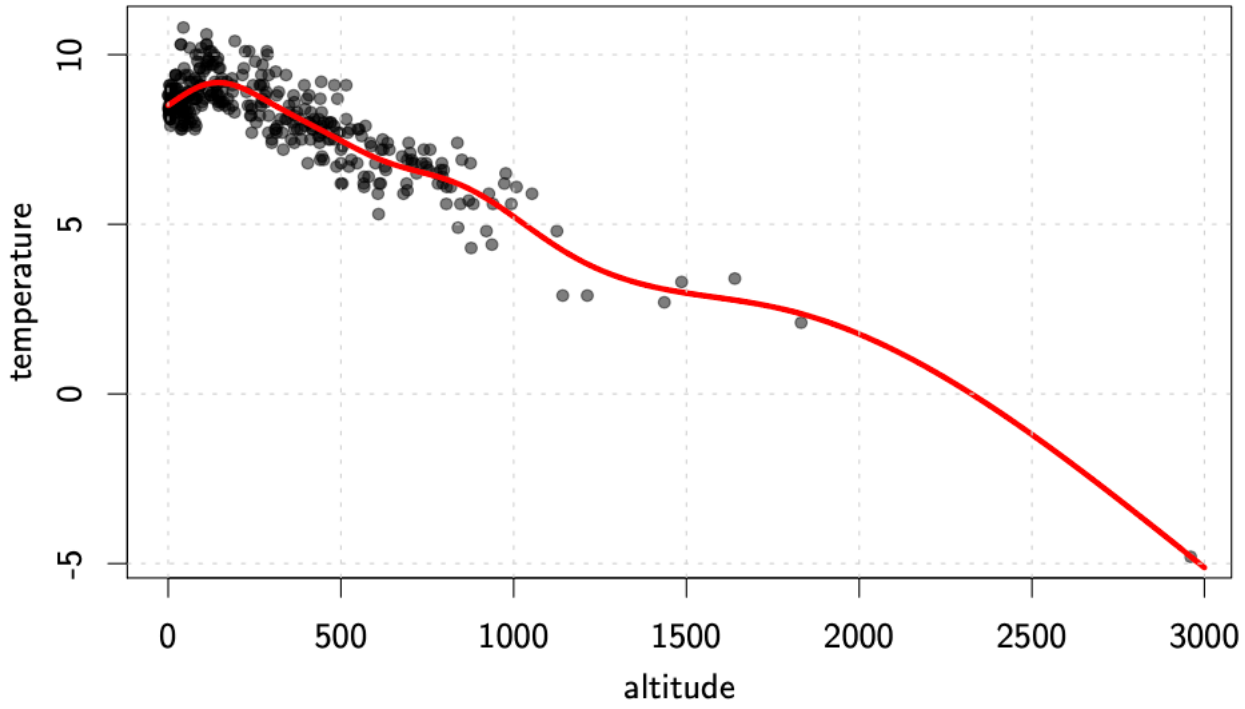


Idea 2: restricted structural causal models

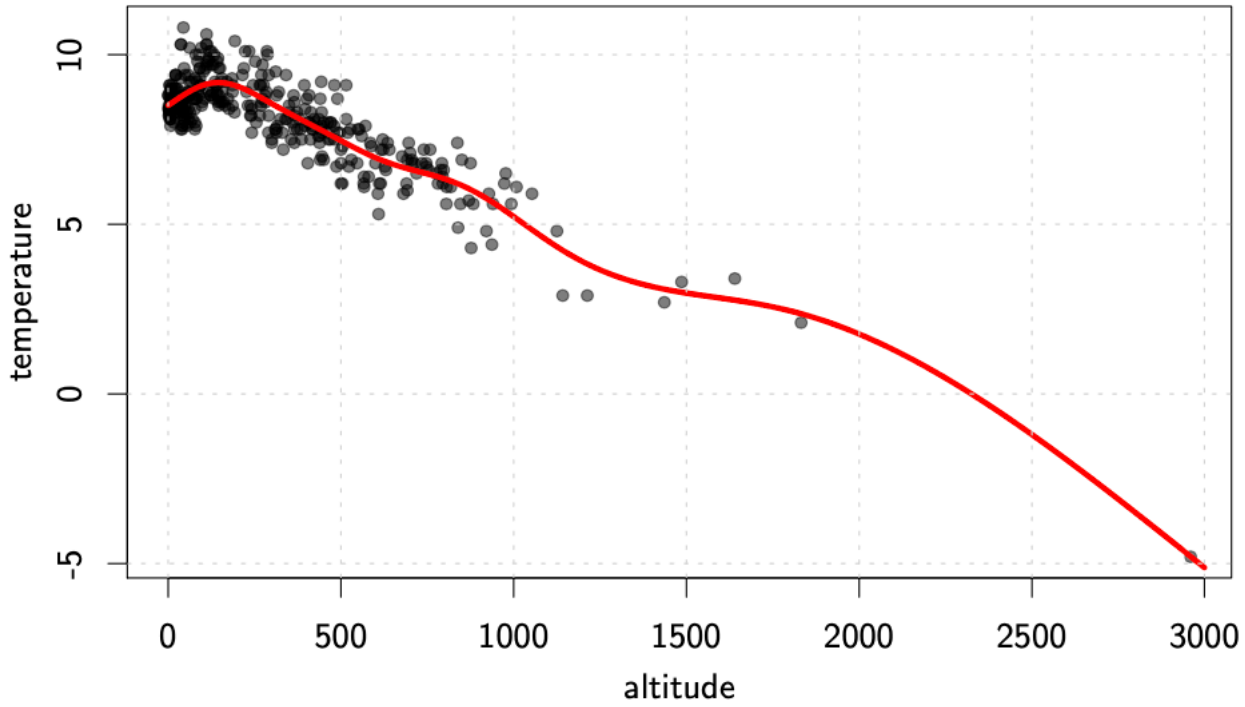


Method...

Real Data: altitude and temperature

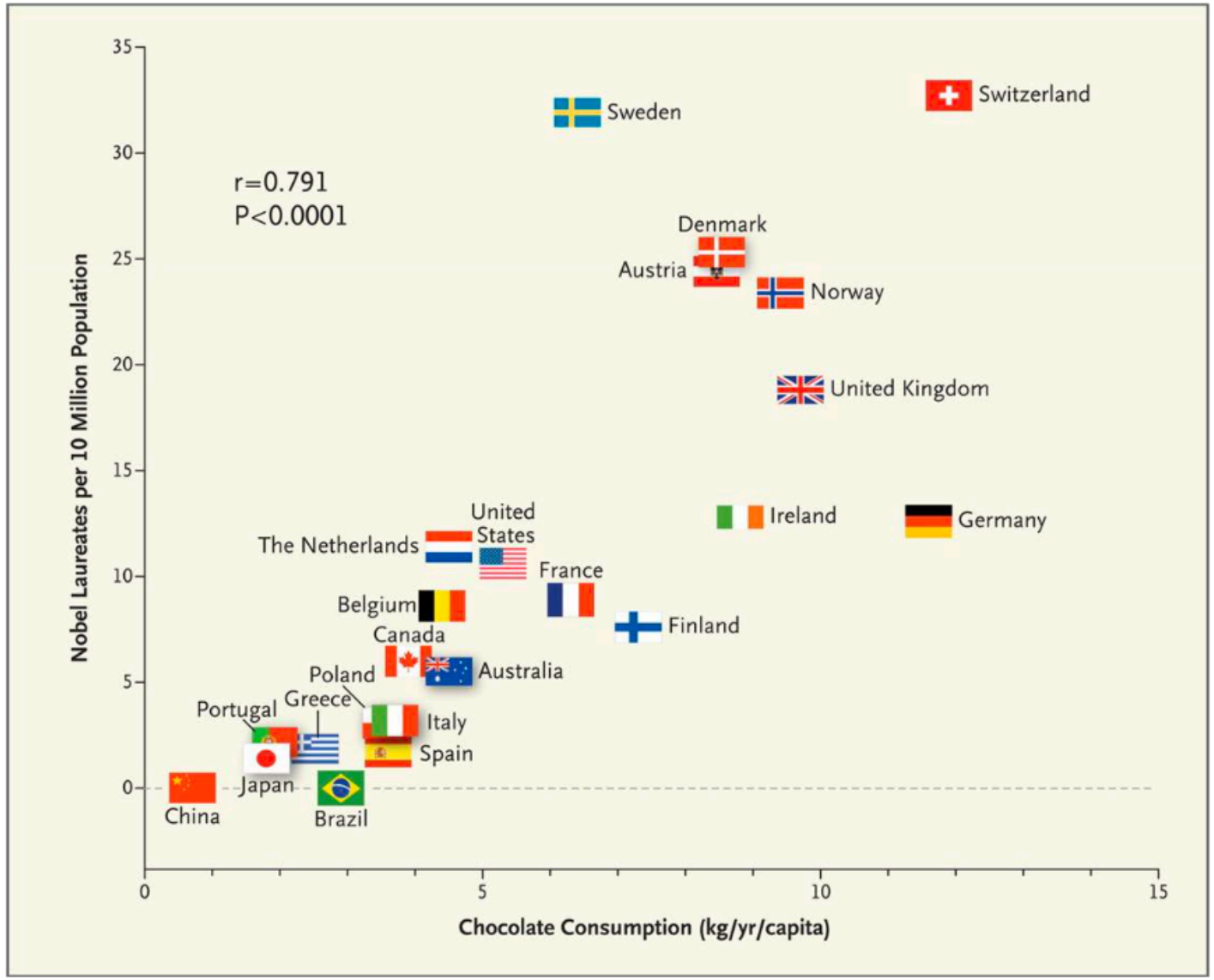


Real Data: altitude and temperature



p-value forward: 0.024
p-value backward: 0.0000000000019

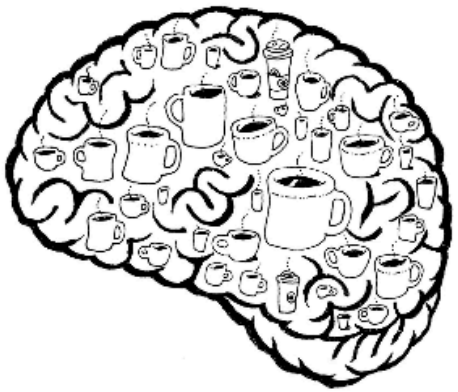
Example: chocolate



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

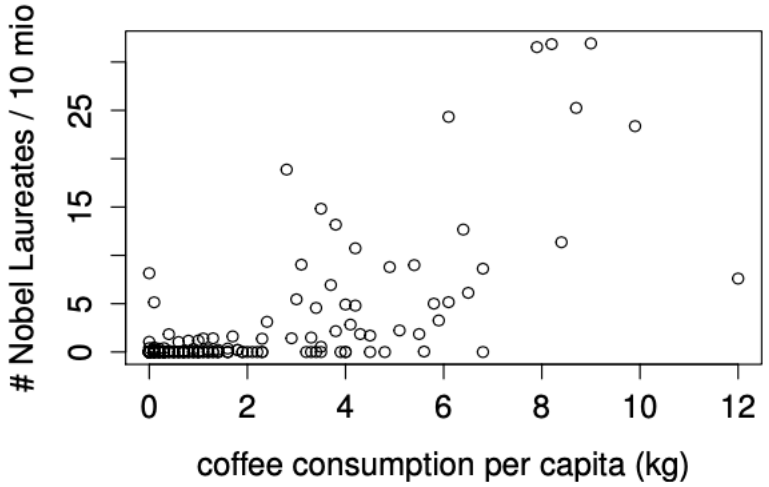
Example: chocolate

No (not enough) data for chocolate



... but we have data for coffee!

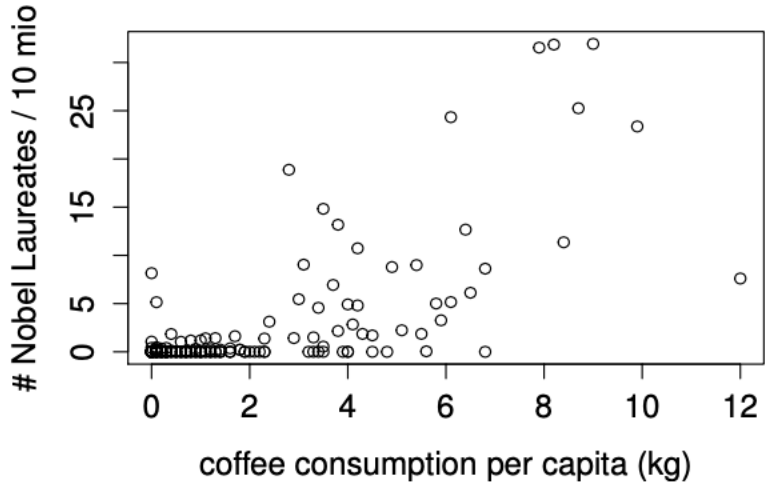
Example: chocolate



Correlation: 0.698

p-value: $< 2.2 \cdot 10^{-16}$

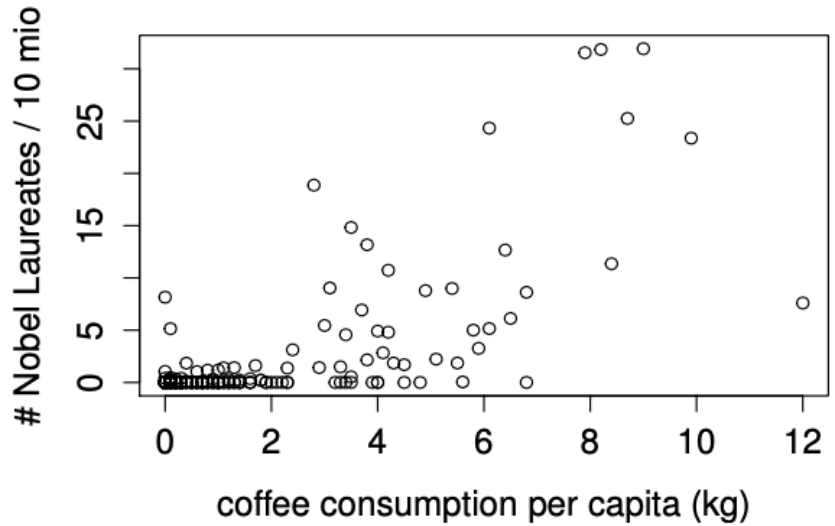
Example: chocolate



Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

Coffee $\rightarrow$ Nobel Prize: Dependent residuals (p-value of $5.1 \cdot 10^{-78}$).
Nobel Prize $\rightarrow$ Coffee: Dependent residuals (p-value of $3.1 \cdot 10^{-12}$).

$\Rightarrow$ Model class too small? Causally insufficient?



Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

Coffee $\rightarrow$ Nobel Prize: Dependent residuals (p-value of $5.1 \cdot 10^{-78}$).
Nobel Prize $\rightarrow$ Coffee: Dependent residuals (p-value of $3.1 \cdot 10^{-12}$).

$\Rightarrow$ Model class too small? Causally insufficient?
Question: When is a p-value too small?