Simplicity of Additive Noise Models

Jonas Peters
ETH Zürich - Marie Curie (IEF)

Workshop on Simplicity and Causal Discovery
Carnegie Mellon University

7th June 2014

ETH zürich

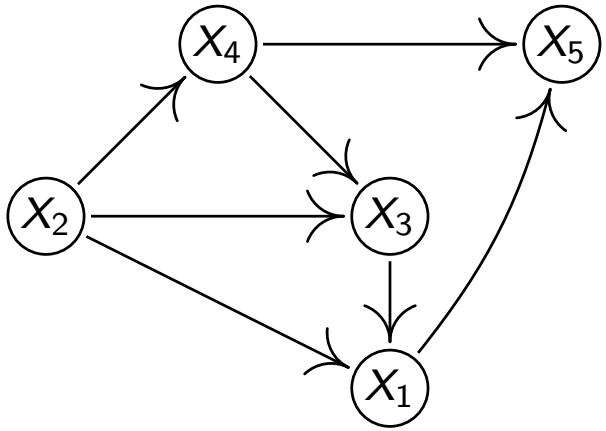
MARIE CURIE ACTIONS

contains joint work with ...

- ETH Zürich:
  Peter Bühlmann, Jan Ernest
- Max-Planck-Institute Tübingen:
  Dominik Janzing, Bernhard Schölkopf
- University of Amsterdam, Radboud University Nijmegen:
  Joris Mooij
- UC Berkeley:
  Sivaraman Balakrishnan, Martin Wainwright

What is the Problem?

Theoretical: $P(X_1, \ldots, X_5)$ $\xrightarrow{?}$ DAG $\mathcal{G}_0$



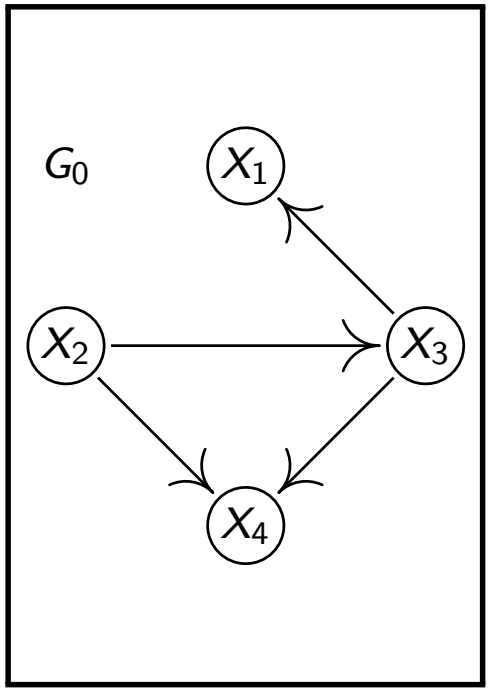
Practical: iid observations from $P(X_1, \ldots, X_5)$ $\xrightarrow{?}$ estimated DAG $\mathcal{G}_0$

Consider the world of Structural Equation Models

Assume $P(X_1, \ldots, X_4)$ has been generated by

$$X_1 = f_1(X_3, N_1)$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2, N_3)$$
$$X_4 = f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles



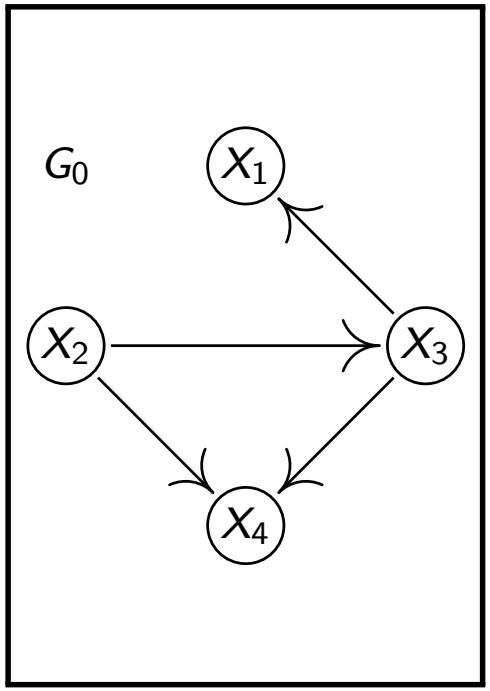
Can the directed acyclic graph be recovered from $P(X_1, \ldots, X_4)$?

Consider the world of Structural Equation Models

Assume $P(X_1, \ldots, X_4)$ has been generated by

$$X_1 = f_1(X_3, N_1)$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2, N_3)$$
$$X_4 = f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles



Can the directed acyclic graph be recovered from $P(X_1, \ldots, X_4)$? **No.**

Consider the world of Structural Equation Models

Proposition

Given a distribution P, we can find a SEM generating P for each graph $\mathcal{G}$, such that P is Markov with respect to $\mathcal{G}$.

JP: *Restricted Structural Equation Models for Causal Inference*, PhD Thesis 2012 (and probably others?)

Consider the world of Structural Equation Models

Proposition

Given a distribution P, we can find a SEM generating P for ~~each graph $\mathcal{G}$, such that P is Markov with respect to $\mathcal{G}$~~ a lot of graphs.

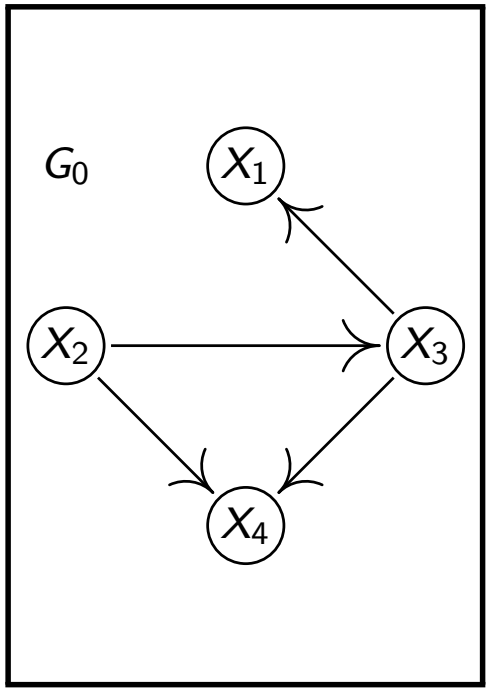
JP: *Restricted Structural Equation Models for Causal Inference*, PhD Thesis 2012 (and probably others?)

Consider the world of Structural Equation Models

Assume $P(X_1, \ldots, X_4)$ has been generated by

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

- N_i jointly independent
- G_0 has no cycles



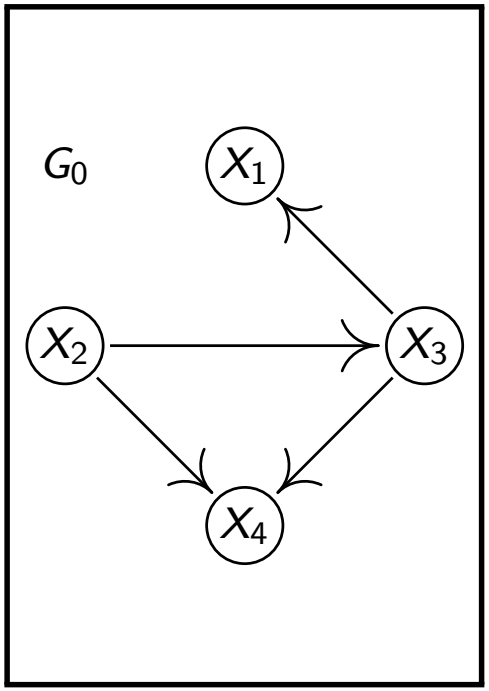
Additive Noise Models.

Consider the world of Structural Equation Models

Assume $P(X_1, \ldots, X_4)$ has been generated by

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

- $N_i \sim \mathcal{N}(0, \sigma_i^2)$ jointly independent
- G_0 has no cycles



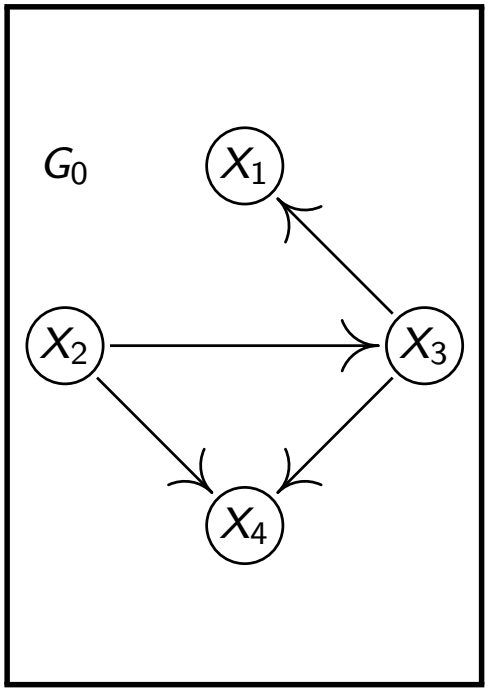
Additive Noise Models with Gaussian noise.

Consider the world of Structural Equation Models

Assume $P(X_1, \ldots, X_4)$ has been generated by

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

- $N_i \sim \mathcal{N}(0, \sigma_i^2)$ jointly independent
- G_0 has no cycles



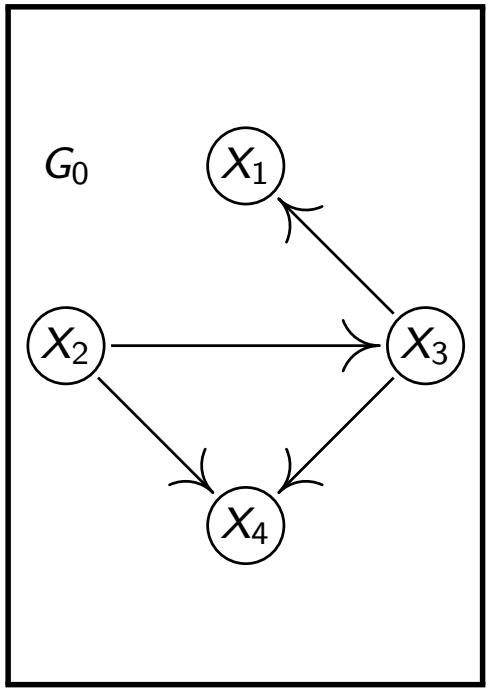
Additive Noise Models with Gaussian noise.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Consider the world of Structural Equation Models

Assume $P(X_1, \ldots, X_4)$ has been generated by

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

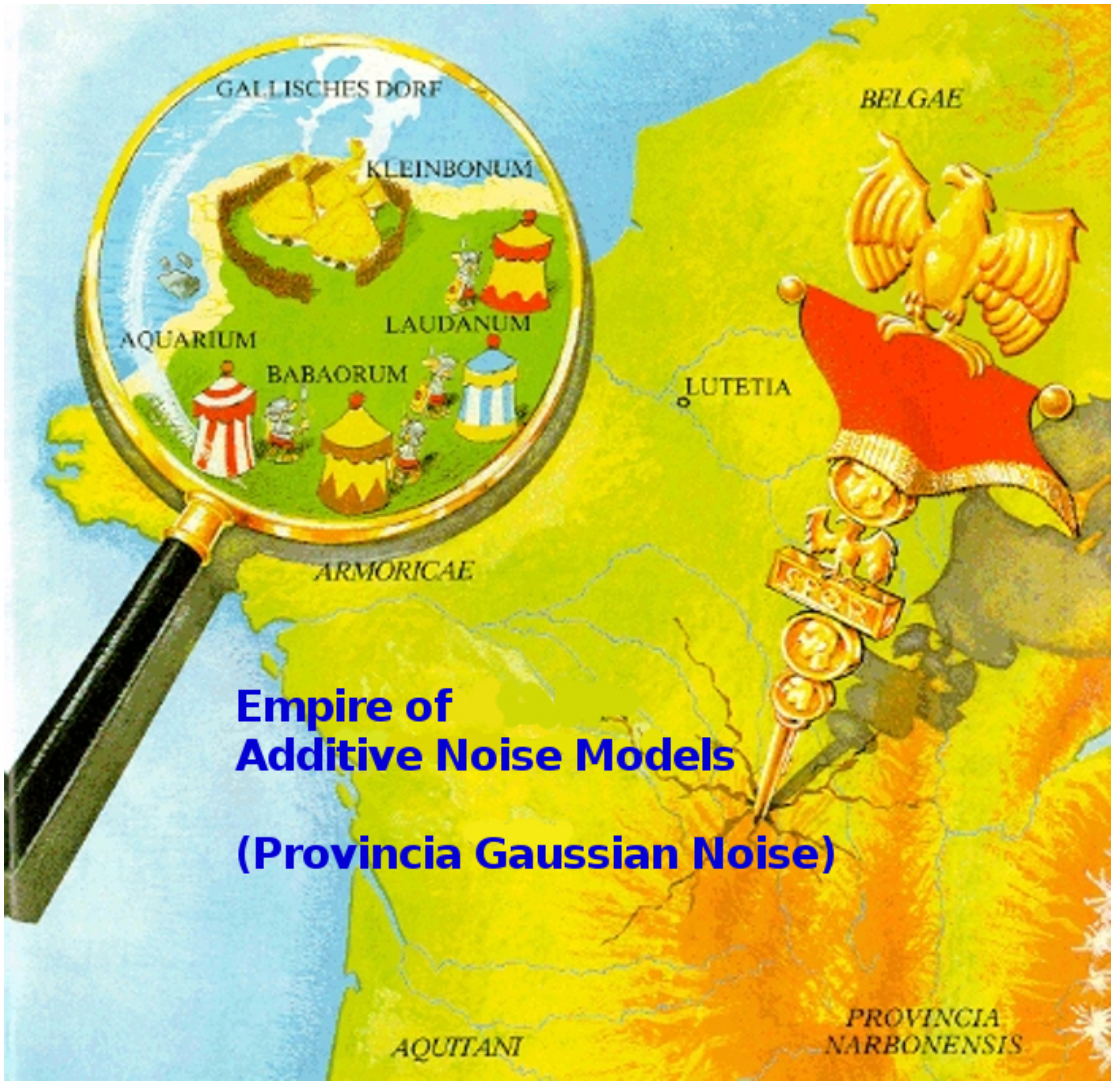
- $N_i \sim \mathcal{N}(0, \sigma_i^2)$ jointly independent
- G_0 has no cycles



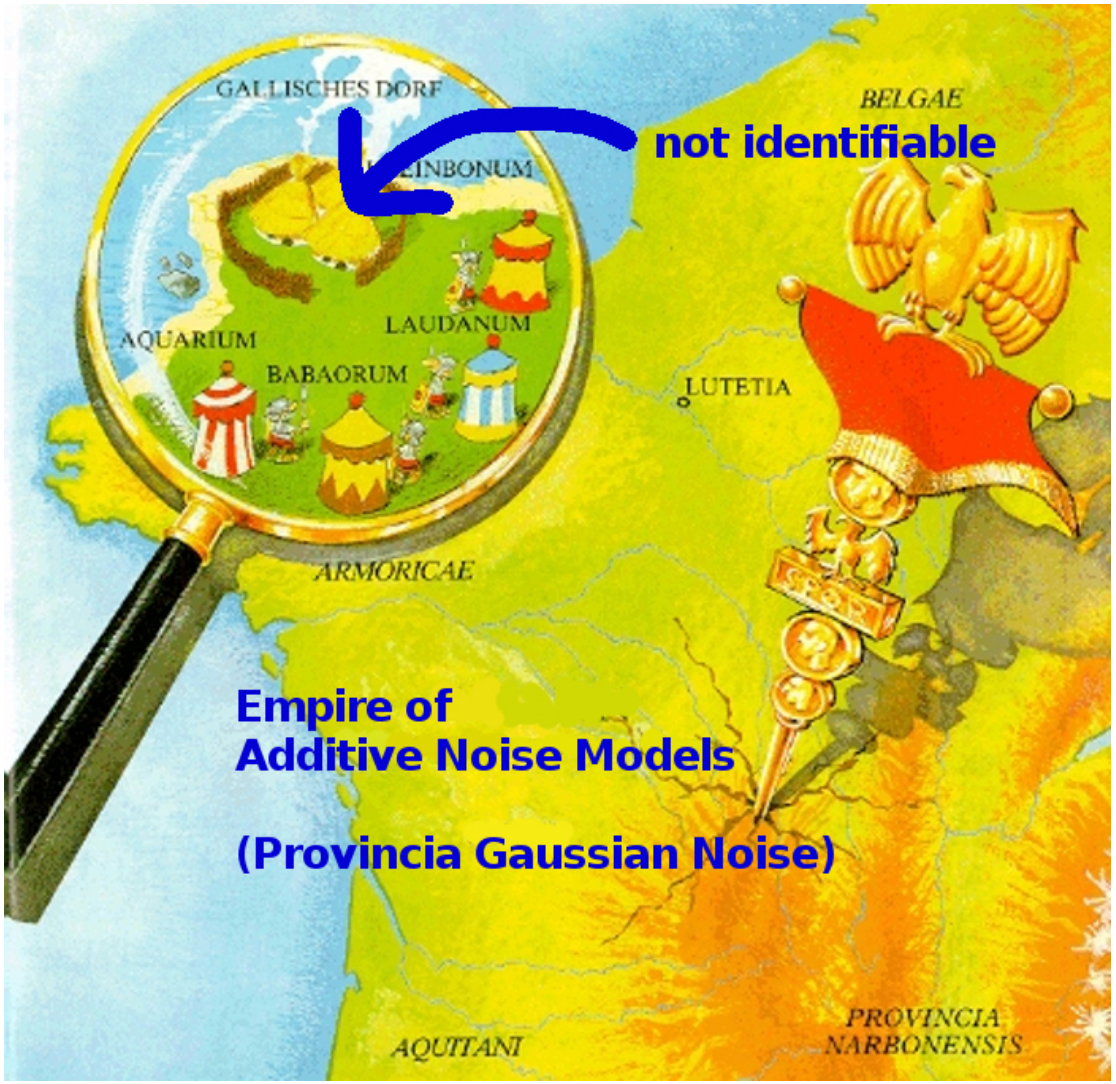
Additive Noise Models with Gaussian noise.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **Yes iff f_i nonlinear.**

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

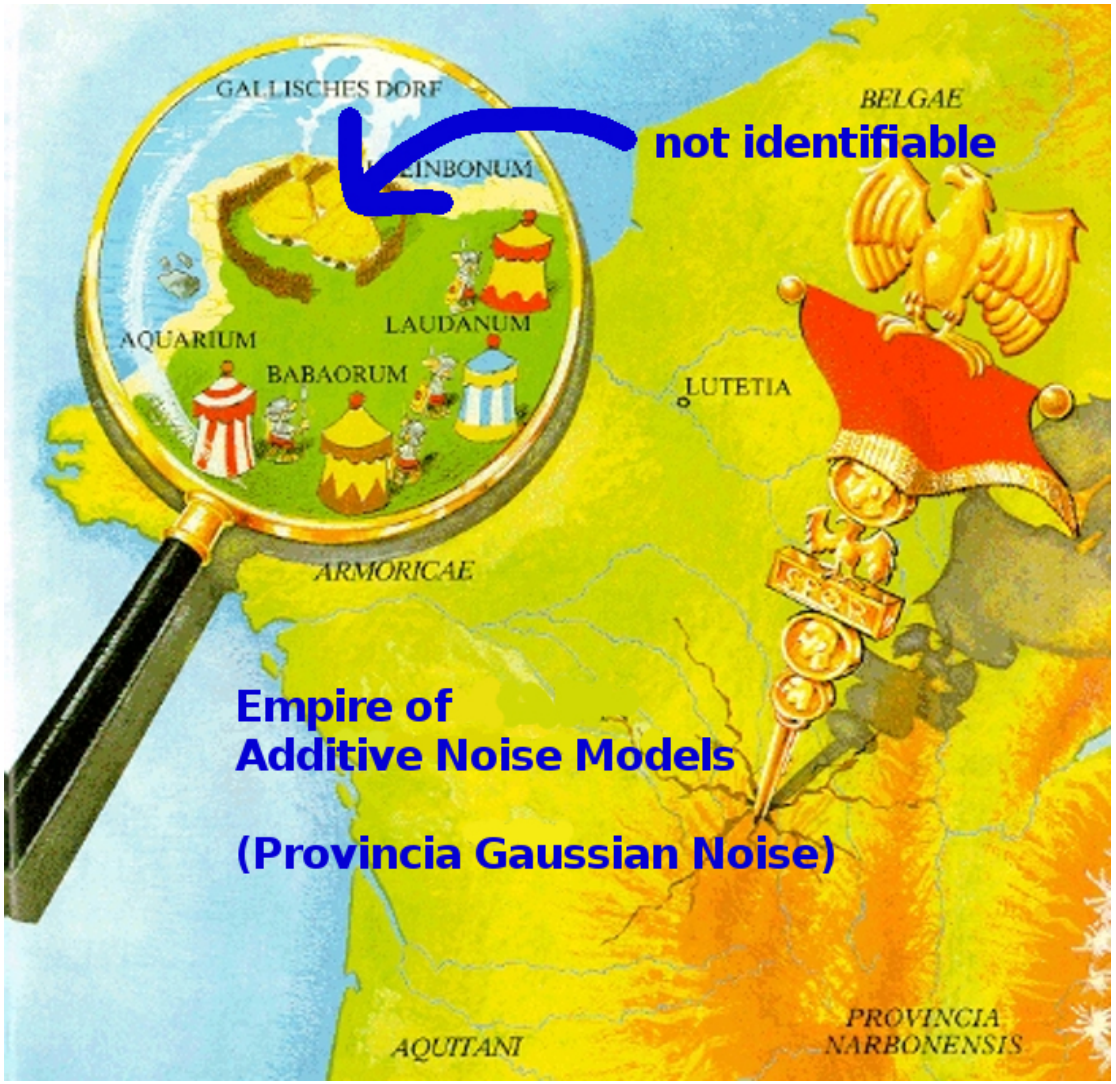
Consider the world of Structural Equation Models



Consider the world of Structural Equation Models



Consider the world of Structural Equation Models



~~GAUL~~ GAUSS
"the LINEAR"

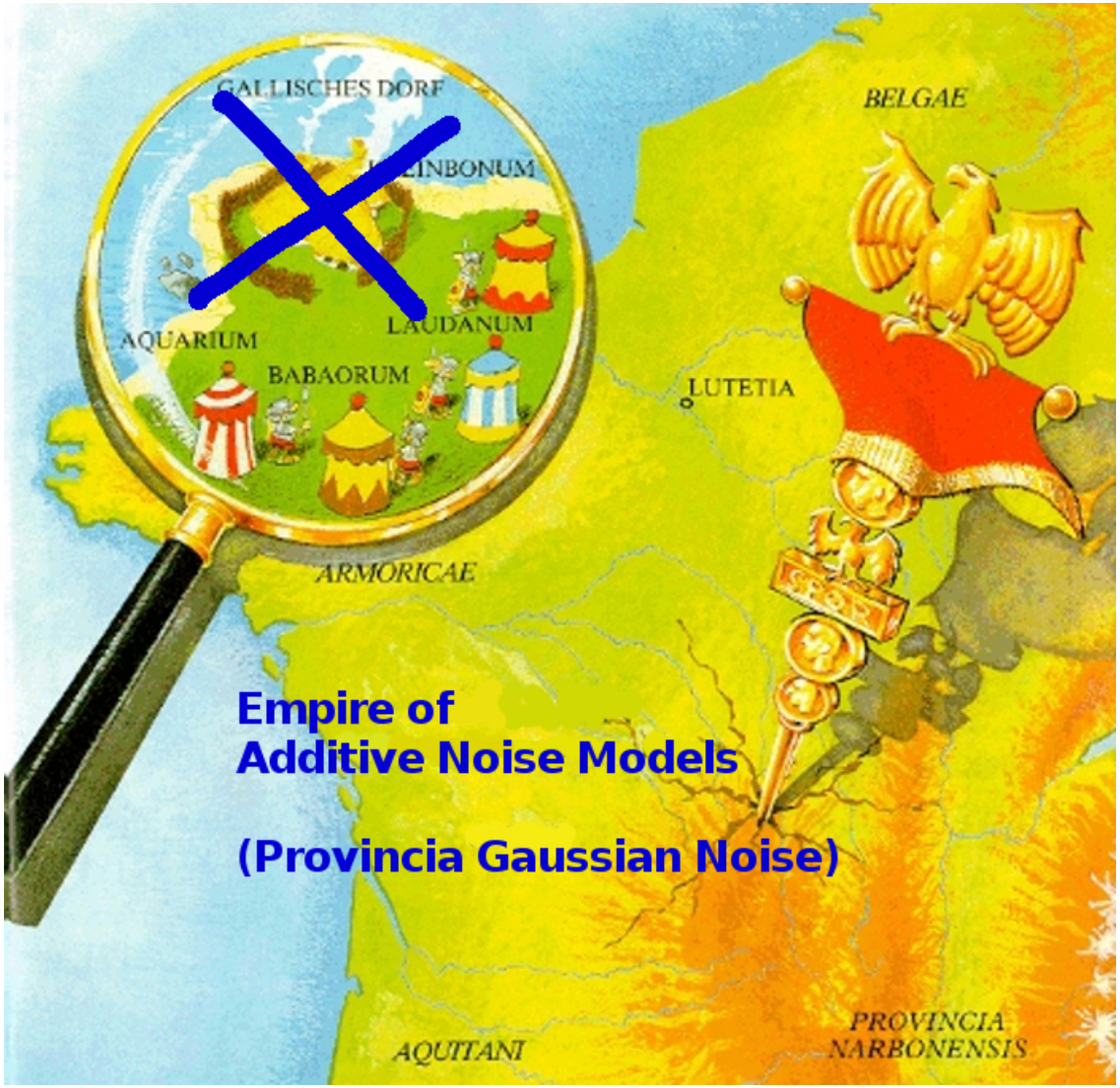
Consider the world of Structural Equation Models

Proposition

Given a Gaussian distribution P, we can find a linear Gaussian SEM generating P for each graph $\mathcal{G}$, such that P is Markov with respect to $\mathcal{G}$.

A. Hauser: *Causal Inference from Interventional Data*, PhD Thesis 2013 (and probably others?)

Consider the world of Structural Equation Models

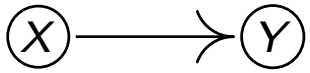


Does X cause Y or vice versa?

Consider a distribution generated by

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

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

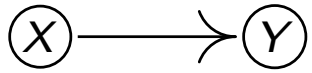


Does X cause Y or vice versa?

Consider a distribution generated by

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

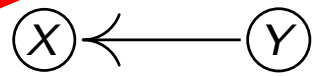
with $N_Y, X \overset{ind}{\sim} \mathcal{N}$



Then, if f is nonlinear, there is no

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

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



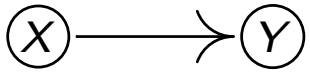
JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, to appear in JMLR*

Does X cause Y or vice versa?

Consider a distribution corresponding to

$$Y = X^3 + N_Y$$

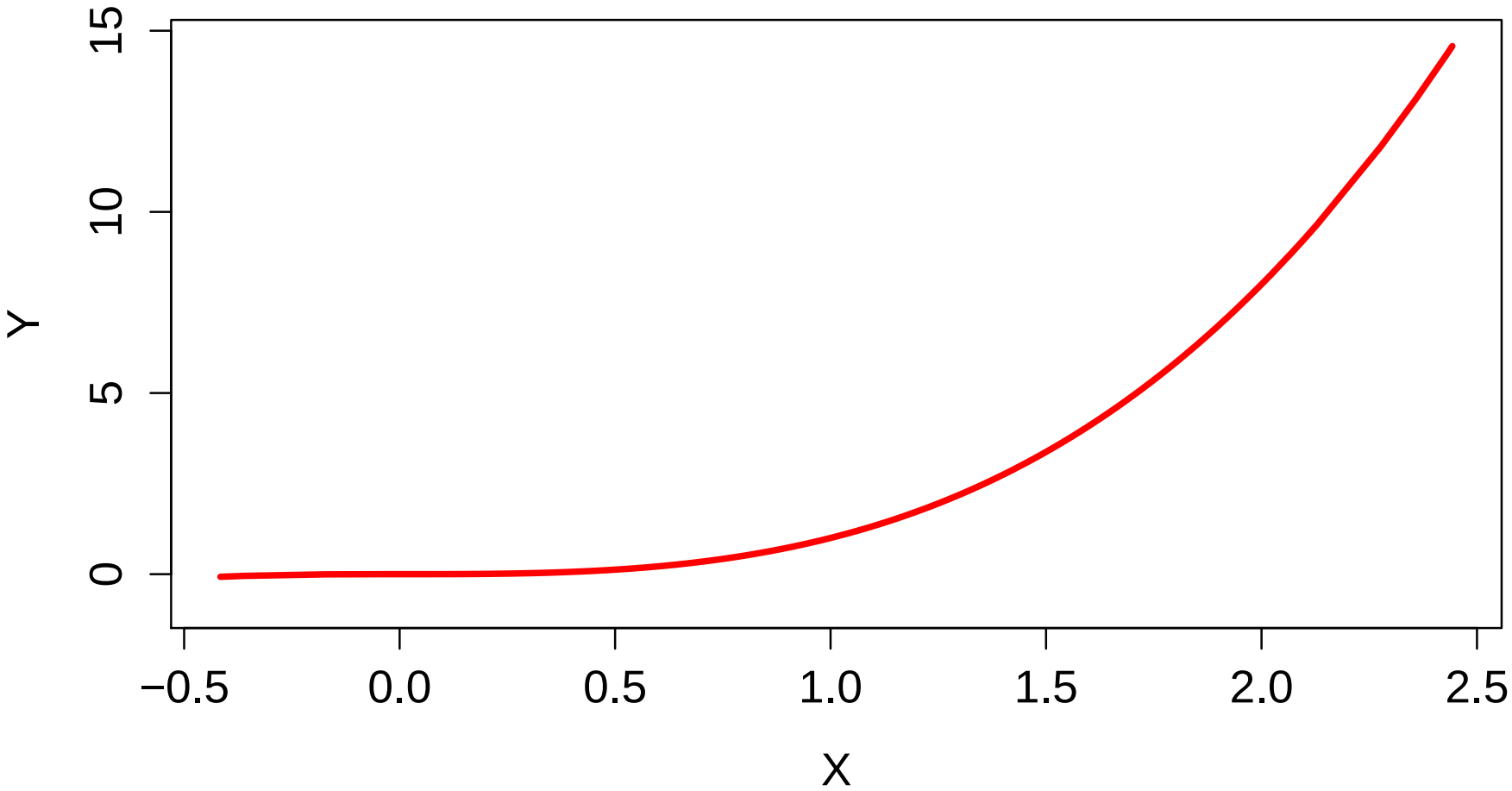
with $N_Y, X \overset{ind}{\sim} \mathcal{N}$



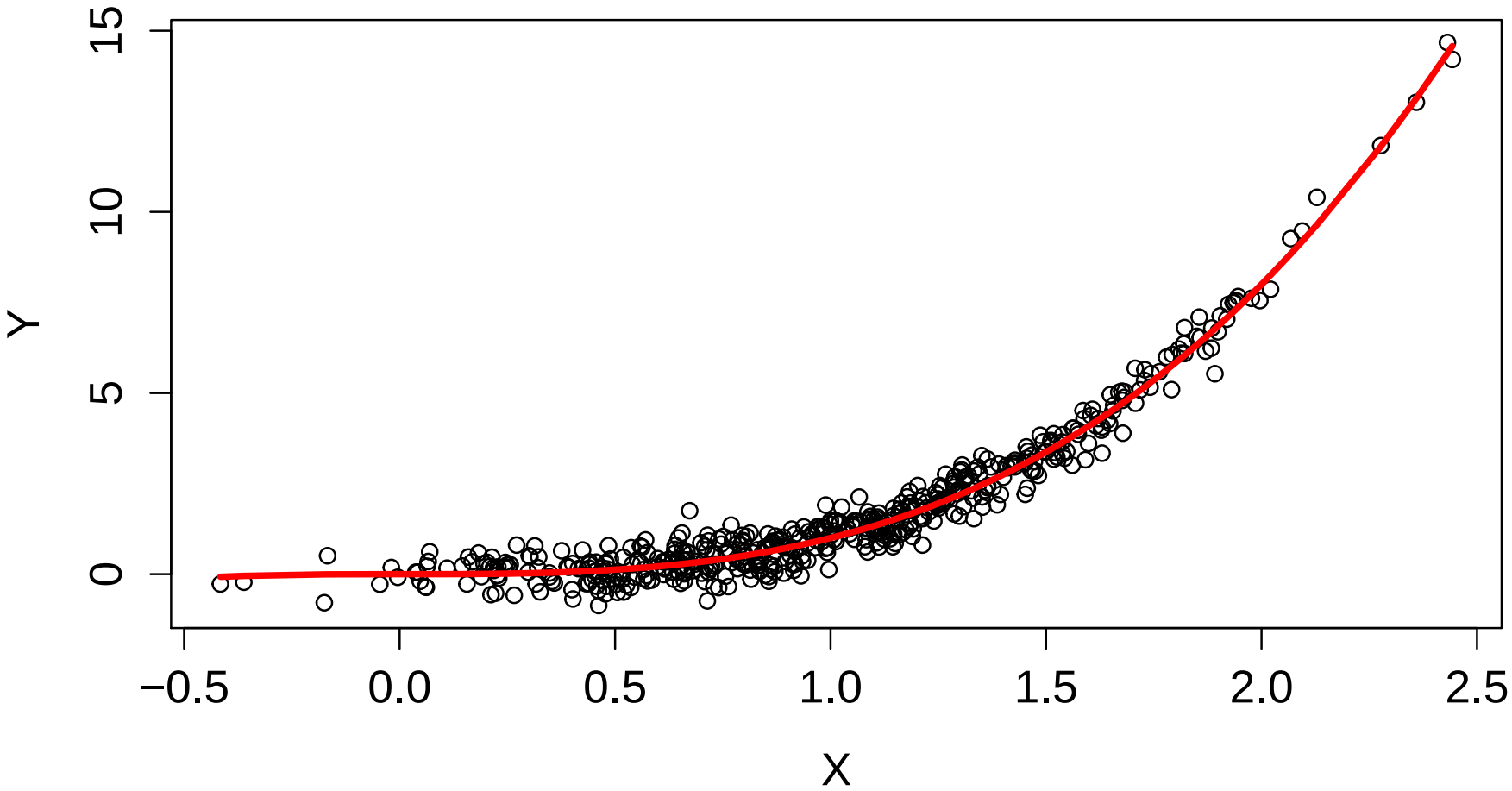
with

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

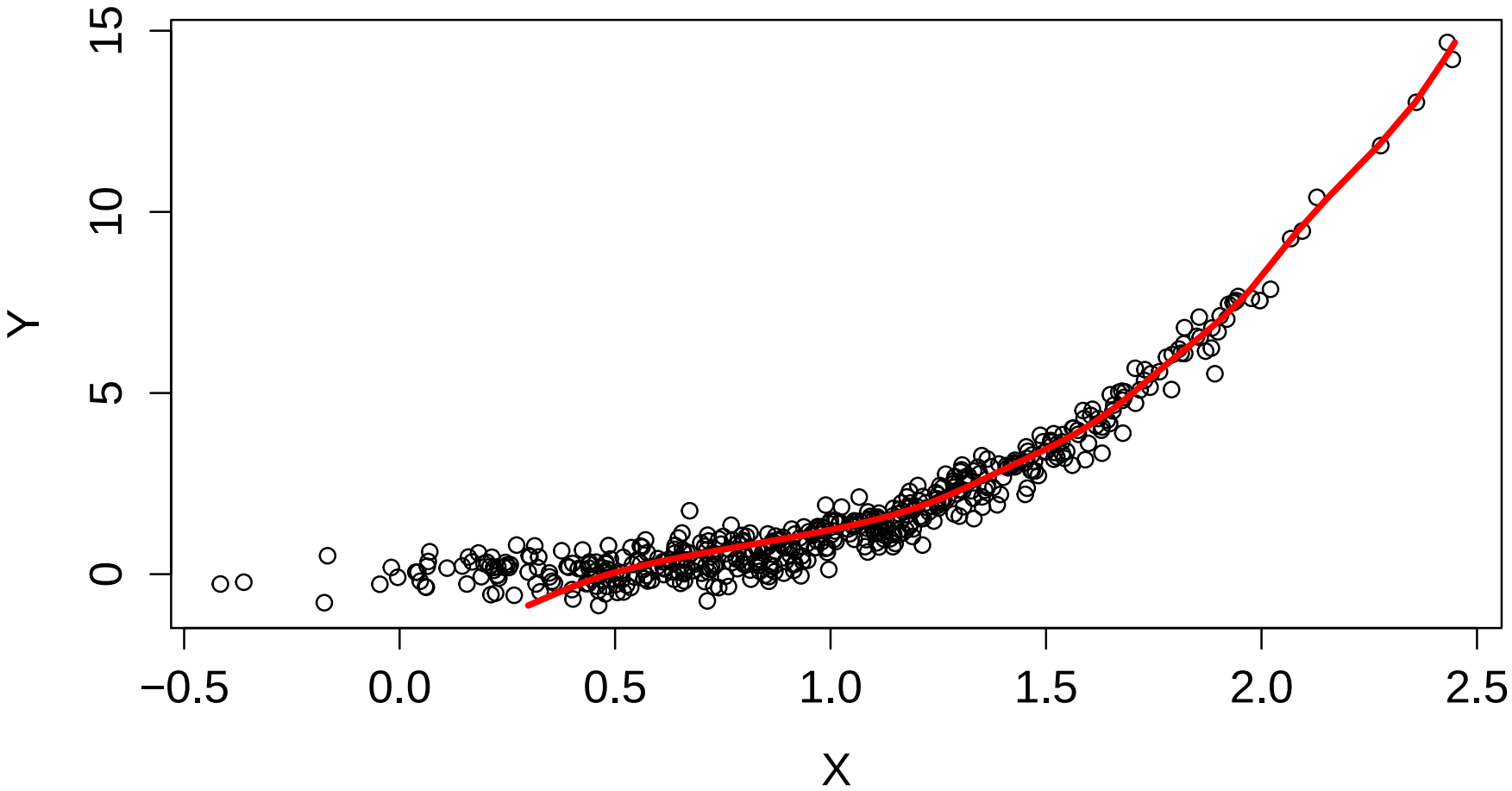
Does X cause Y or vice versa?



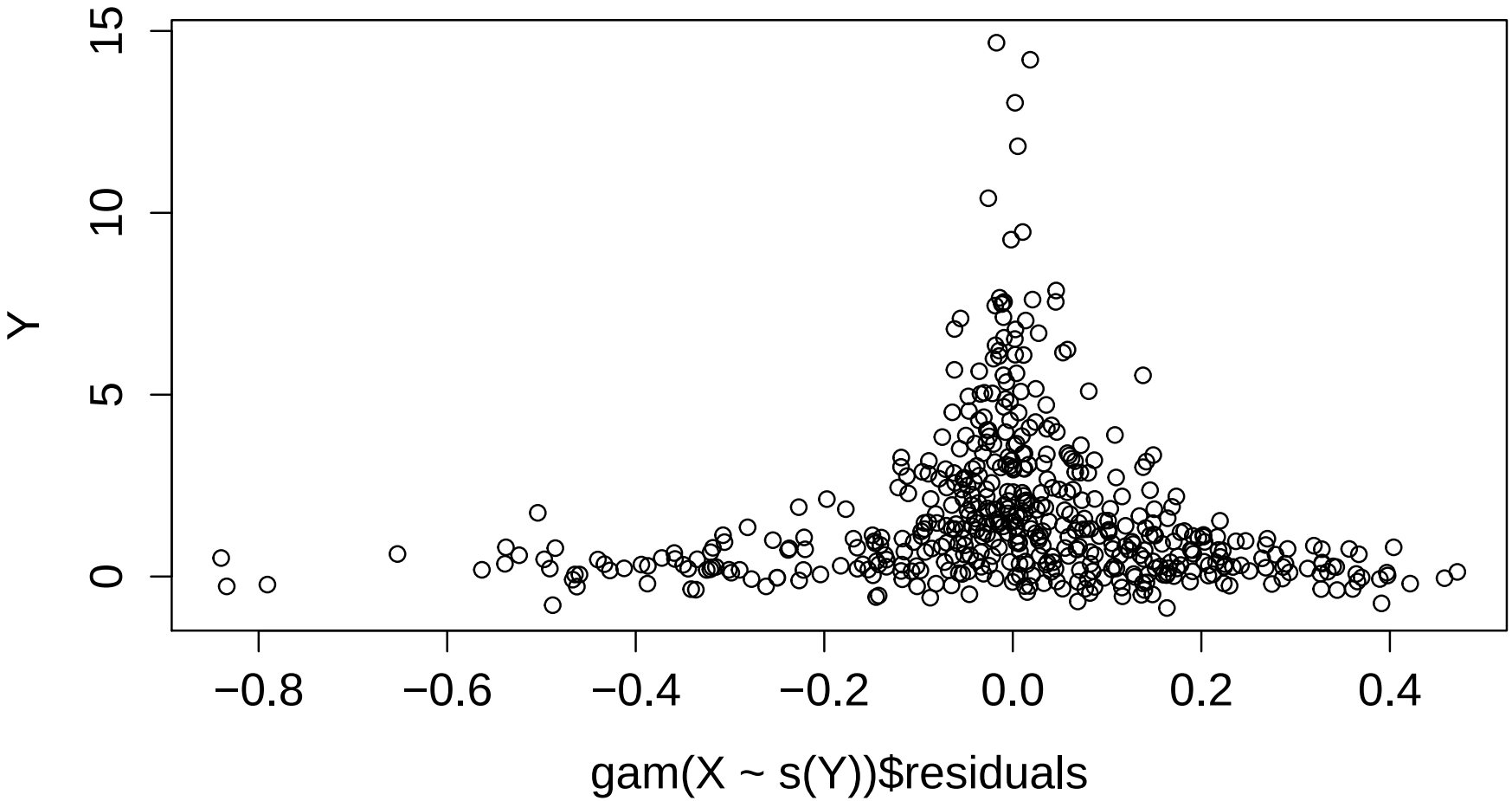
Does X cause Y or vice versa?



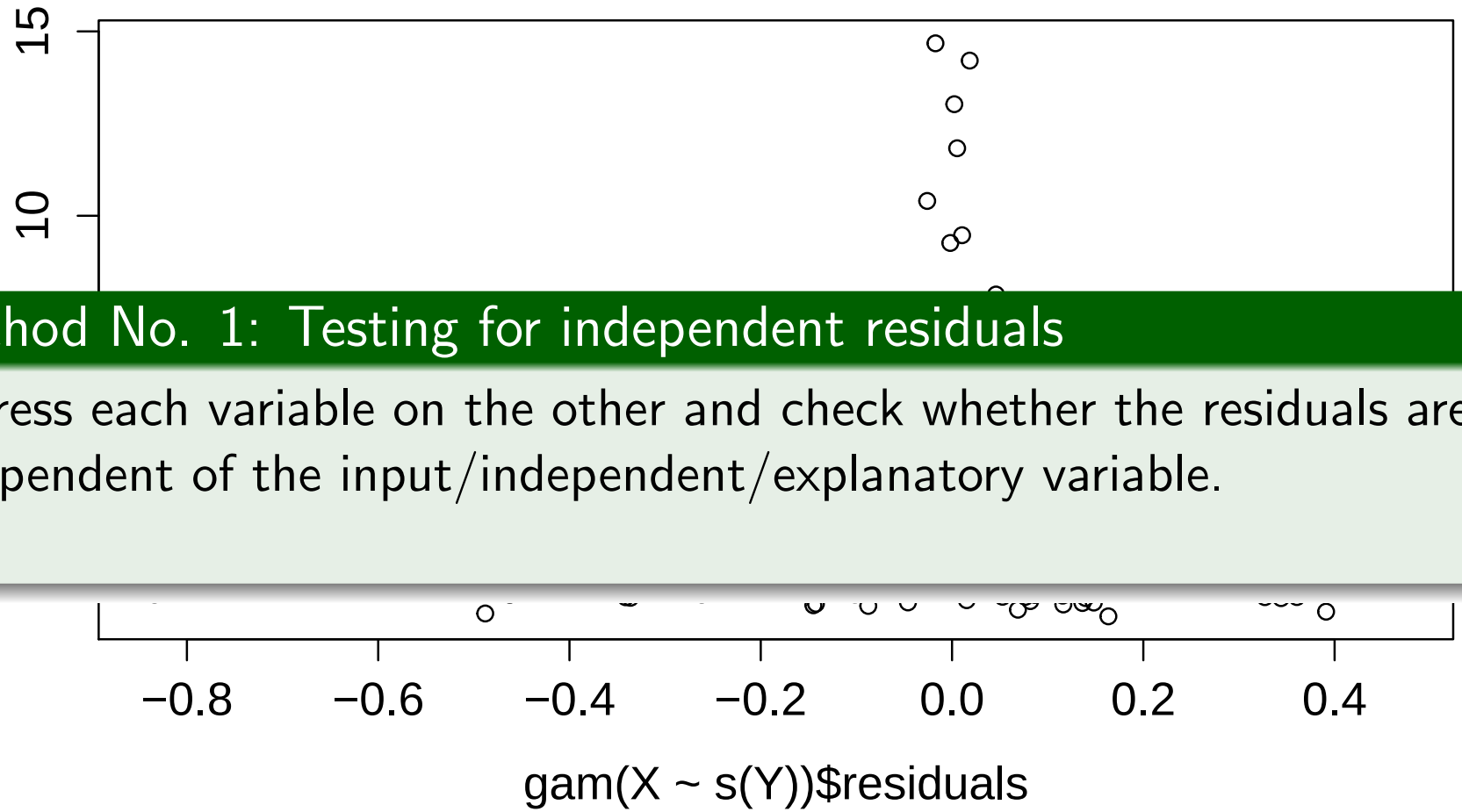
Does X cause Y or vice versa?



Does X cause Y or vice versa?



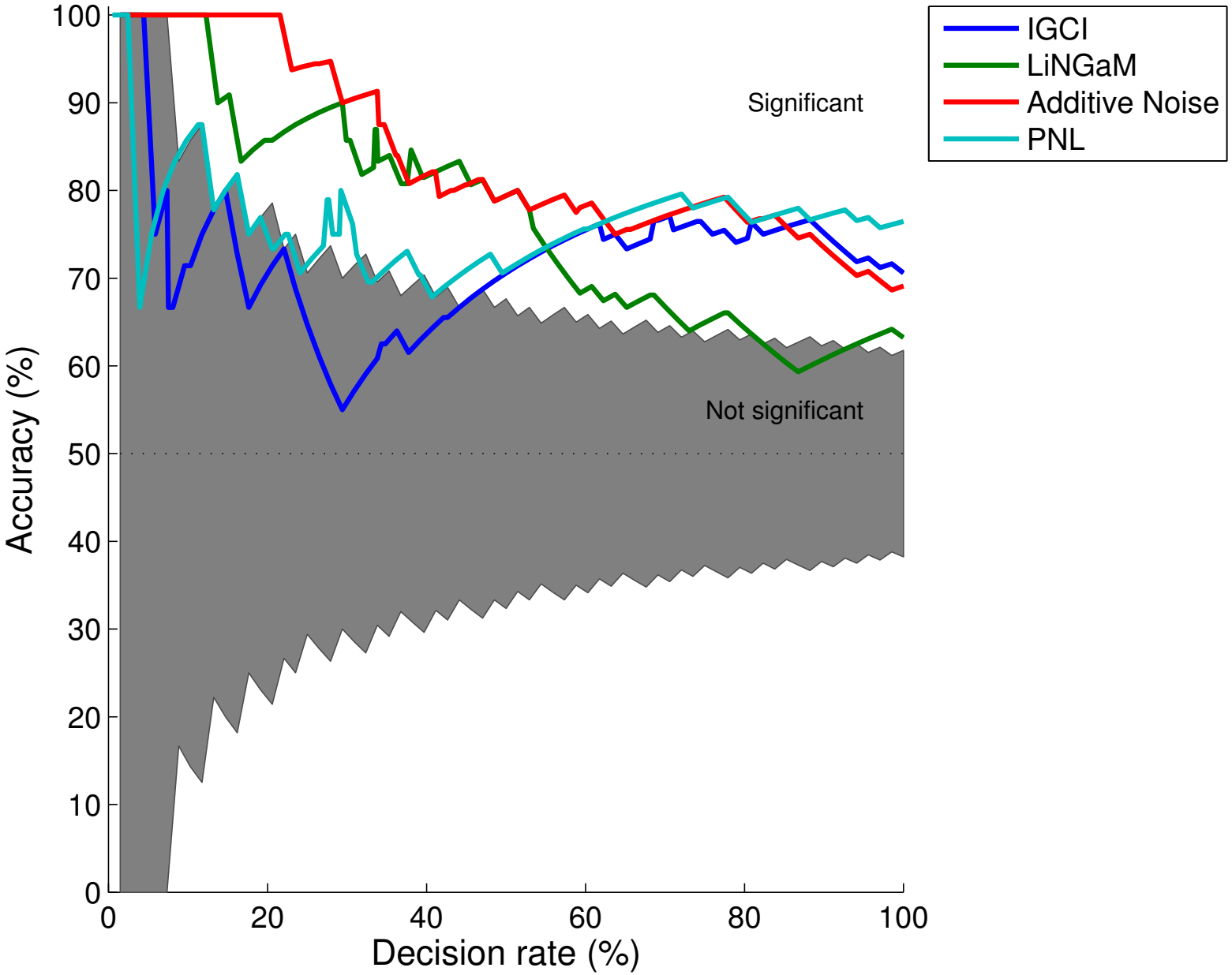
Does X cause Y or vice versa?



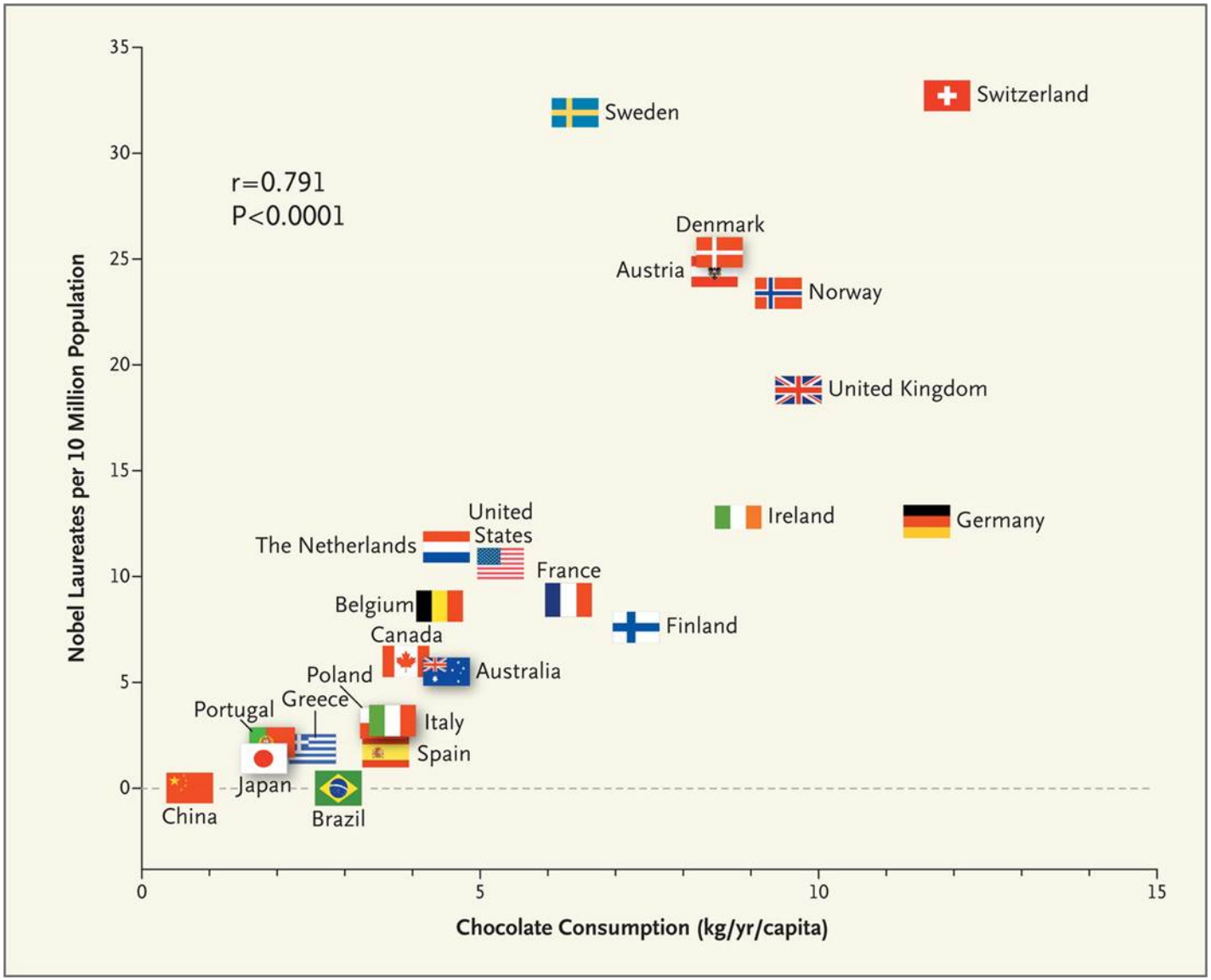
Method No. 1: Testing for independent residuals

Regress each variable on the other and check whether the residuals are independent of the input/independent/explanatory variable.

Does X cause Y or vice versa?



Does X cause Y or vice versa?



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

Does X cause Y or vice versa?

Confectionery news.com

HEADLINES | TRENDS | TECHNOLOGY | PRODUCTS | JOBS | EVENTS | RELATED SITES |

HEADLINES > REGULATION & SAFETY

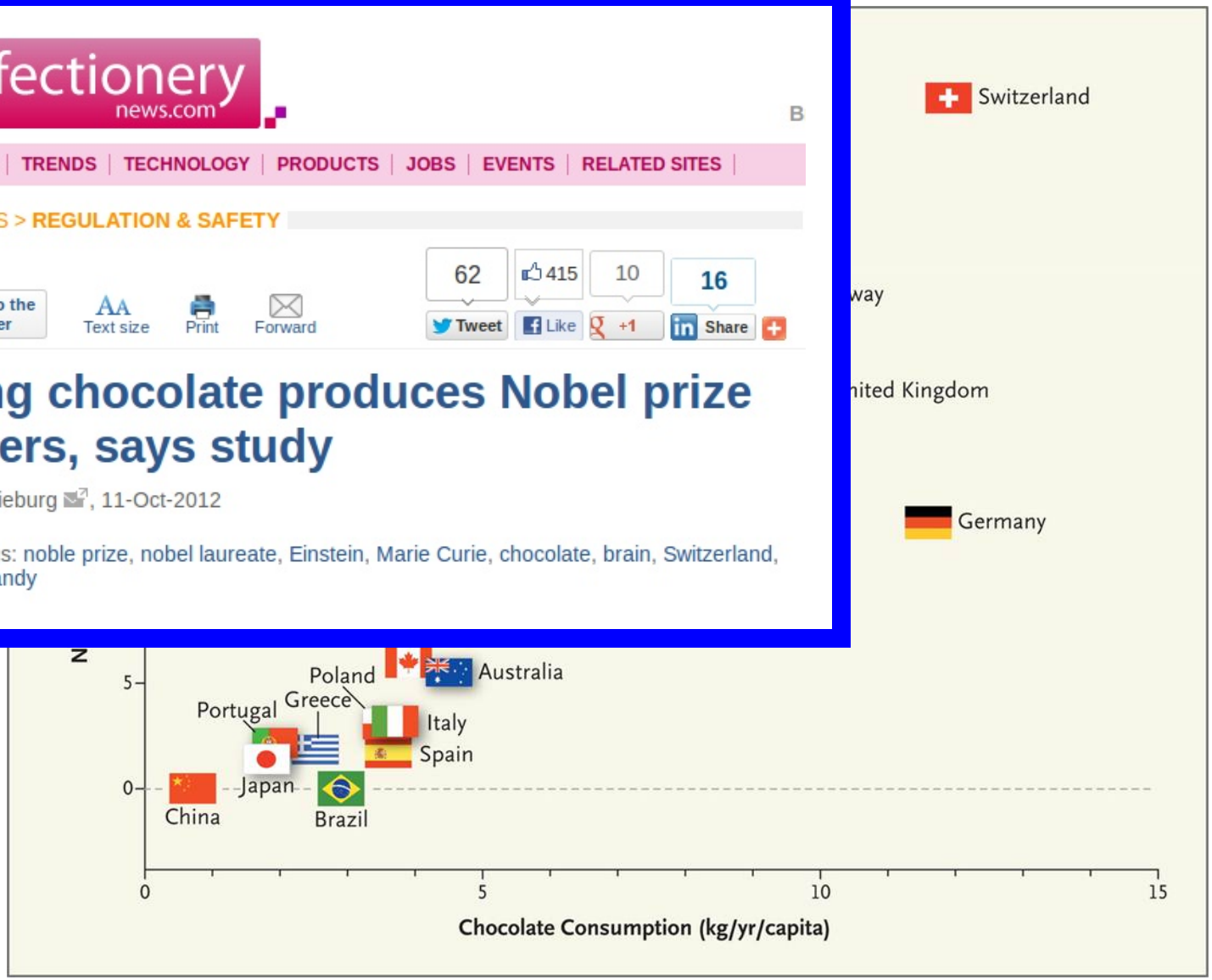
Subscribe to the Newsletter | Text size | Print | Forward

62 Tweet | 415 Like | 10 +1 | 16 Share

Eating chocolate produces Nobel prize winners, says study

By Oliver Nieburg, 11-Oct-2012

Related tags: noble prize, nobel laureate, Einstein, Marie Curie, chocolate, brain, Switzerland, Sweden, candy



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

Does X cause Y or vice versa?

Confectionery

HEADLINES | T

HEADLINES >

Subscribe to the Newsletter

Eating winner

By Oliver Niebu

Related tags: r
Sweden, candy

Forbes

New Posts
+10 posts this hour

Most Popular
Google's Driverless Car

Lis
Busi

+ Follow (73)

PHARMA & HEALTHCARE | 10/10/2012 @ 5:02PM | 14,700 views

Chocolate And Nobel Prizes In Study

4 comments, 2 called-out + Comment Now + Follow Comments

You don't have to be a genius to like chocolate, but geniuses are more likely to eat lots of chocolate, at least according to a new paper published in the August *New England Journal of Medicine*. Franz Messerli reports a highly

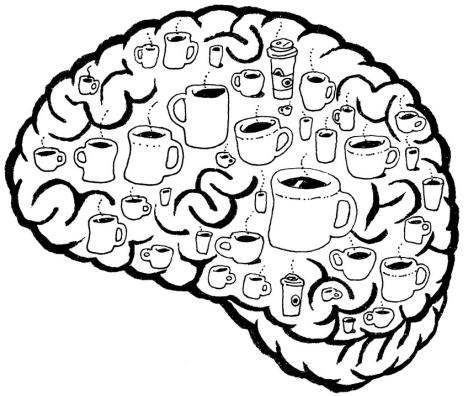
F. H. 12

Does X cause Y or vice versa?

No (not enough) data for chocolate

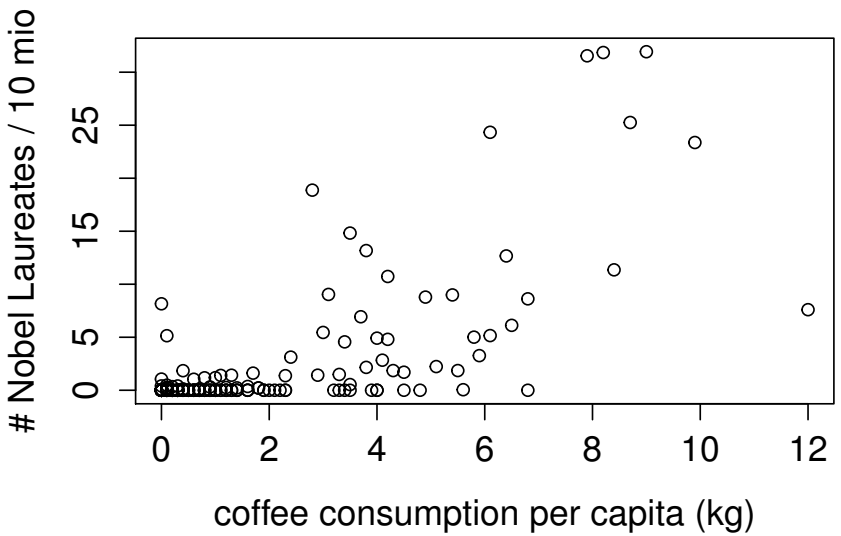
Does X cause Y or vice versa?

No (not enough) data for chocolate



... but we have data for coffee!

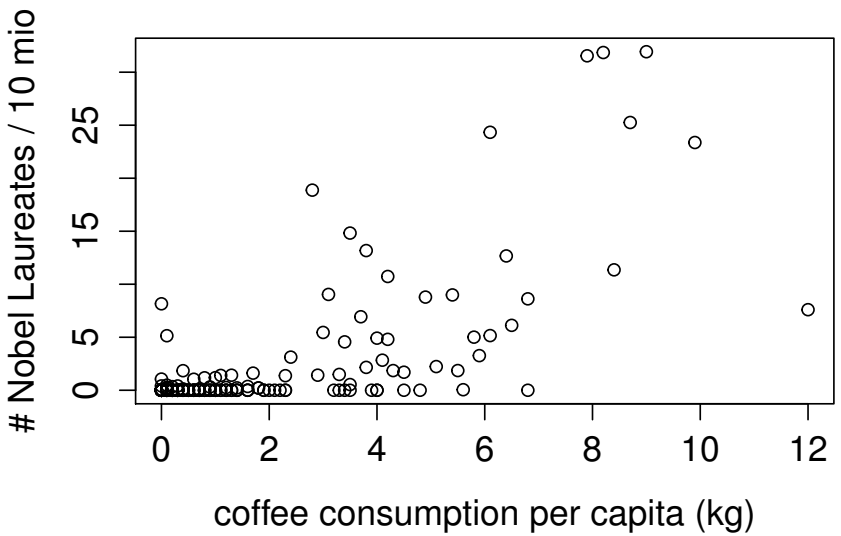
Does X cause Y or vice versa?



Correlation: 0.698

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

Does X cause Y or vice versa?

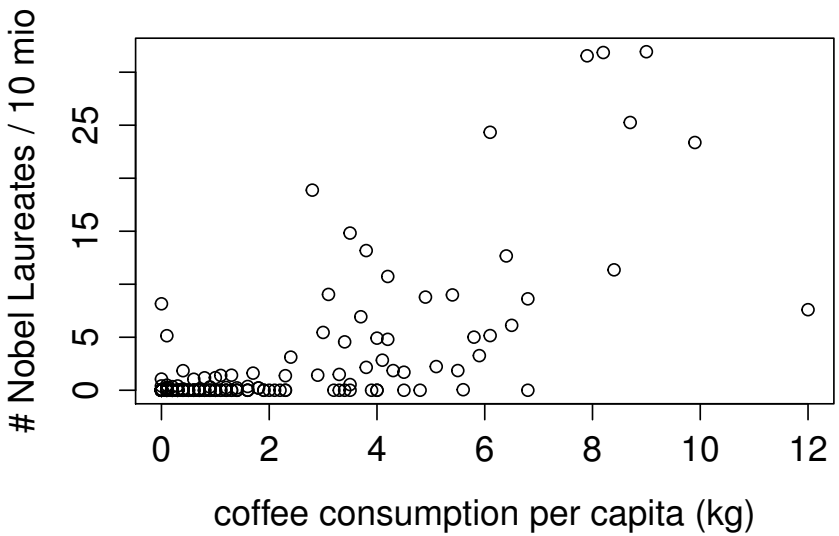


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

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?

Does X cause Y or vice versa?



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

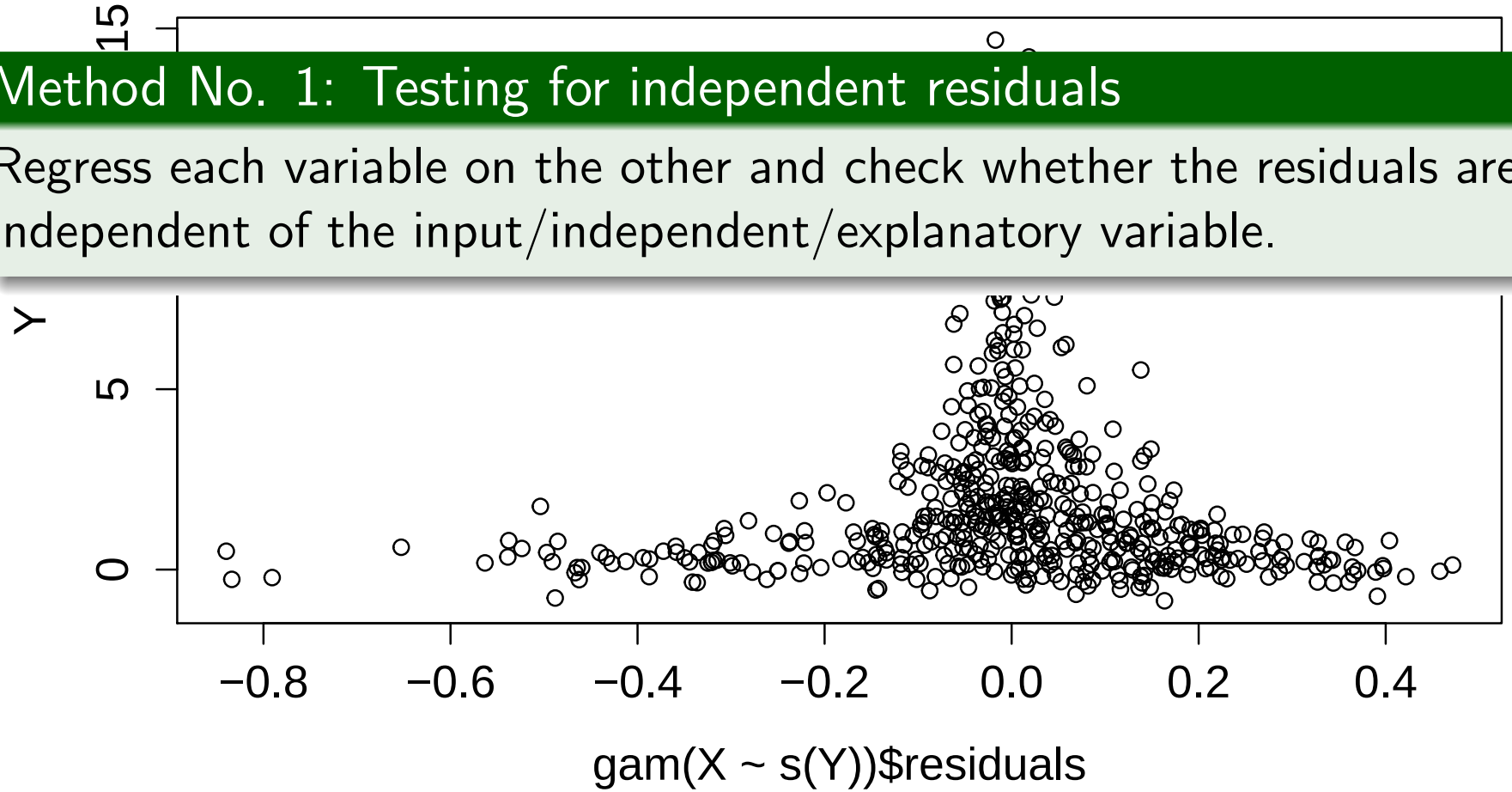
Coffee $\rightarrow$ Nobel Prize: Dependent residuals (p-value of $5.1 \cdot 10^{-78}$).
Nobel Prize $\rightarrow$ Coffee: Dependent residuals (p-value of $3.1 \cdot 10^{-12}$).

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

Does X cause Y or vice versa?

Method No. 1: Testing for independent residuals

Regress each variable on the other and check whether the residuals are independent of the input/independent/explanatory variable.



Does X cause Y or vice versa?

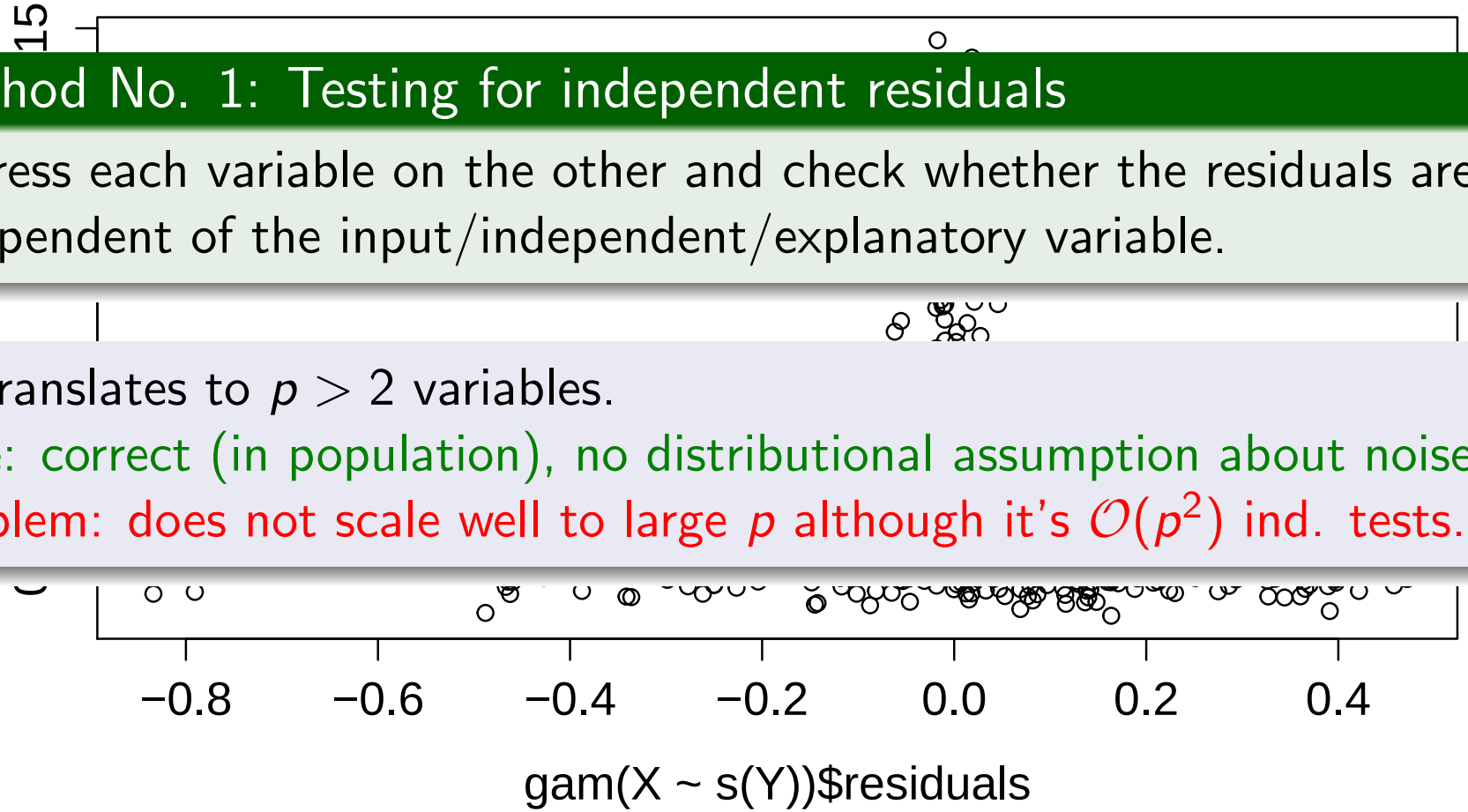
Method No. 1: Testing for independent residuals

Regress each variable on the other and check whether the residuals are independent of the input/independent/explanatory variable.

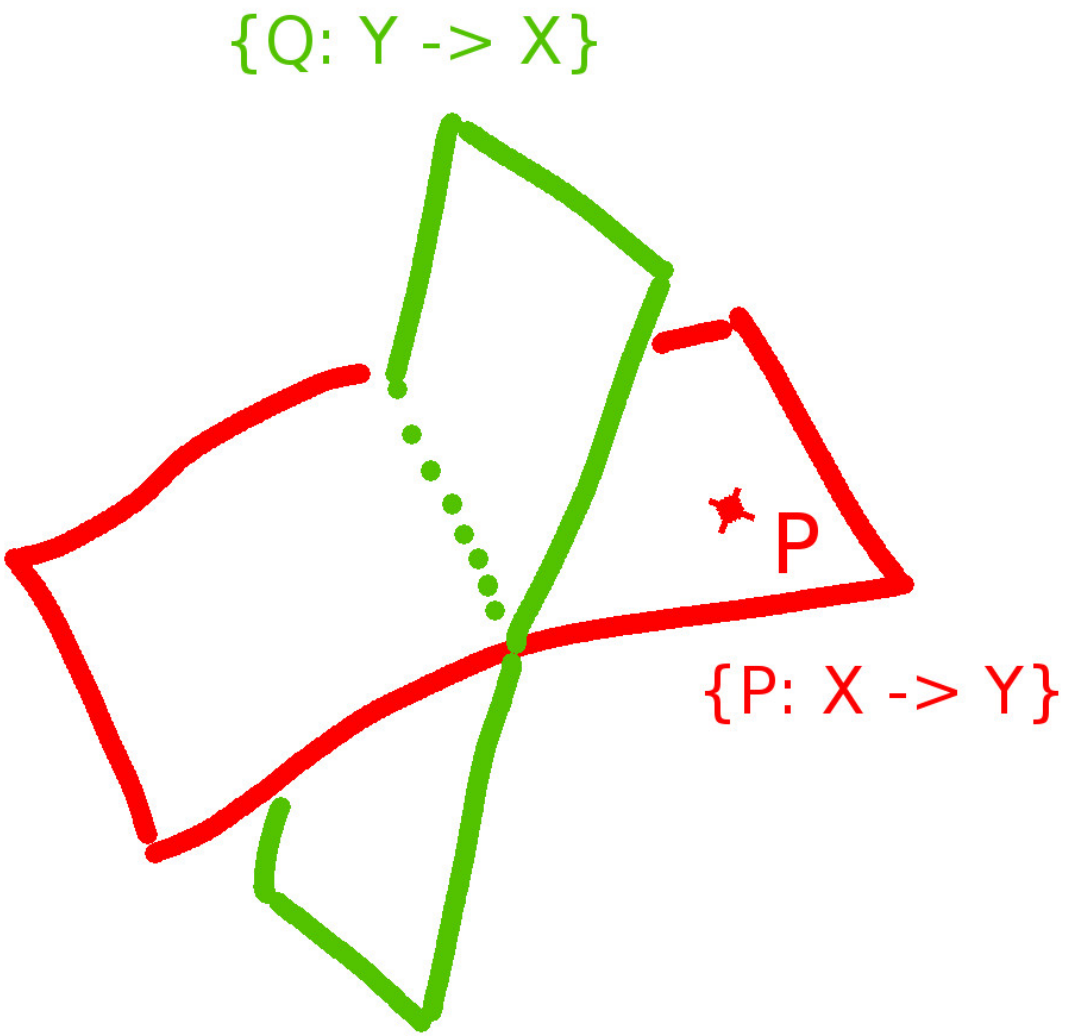
$\rightsquigarrow$ translates to $p > 2$ variables.

Nice: correct (in population), no distributional assumption about noise.

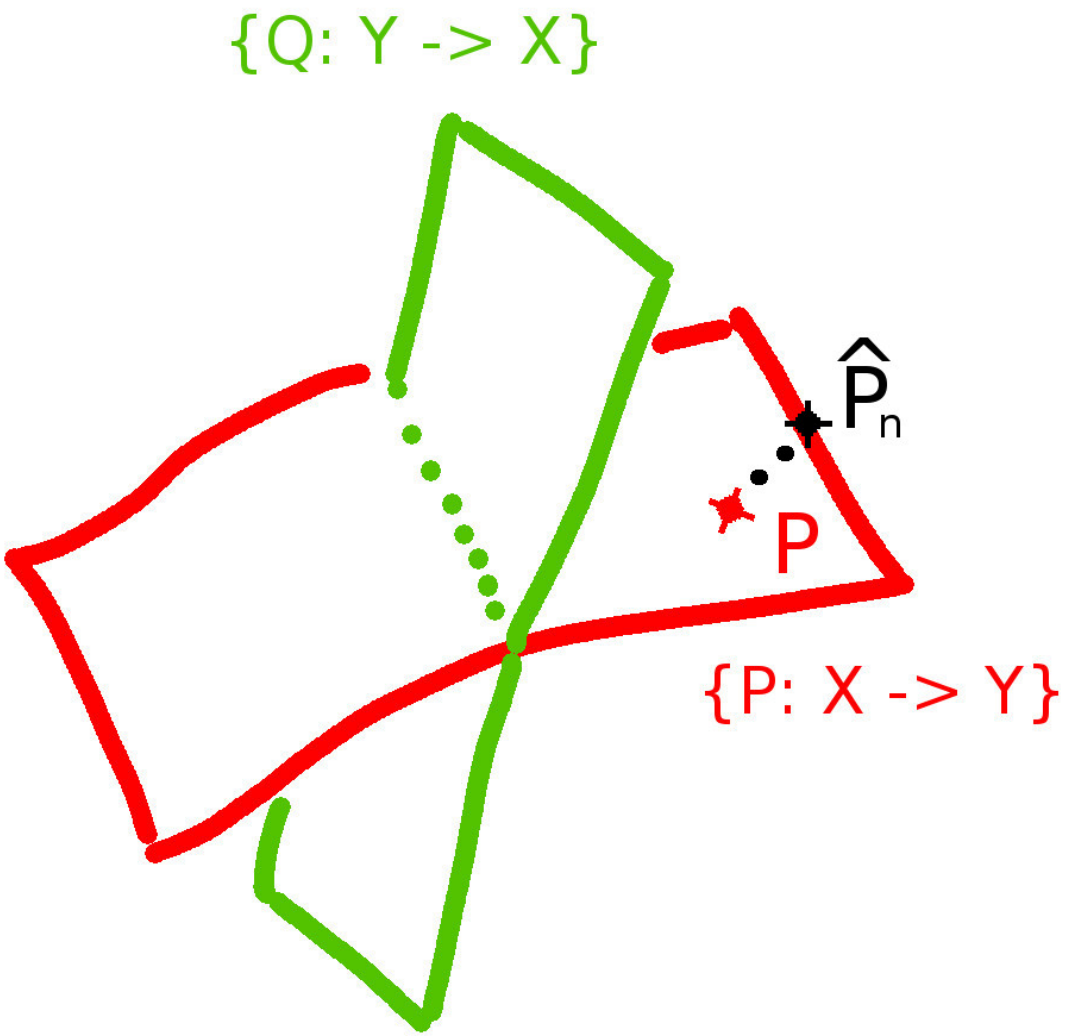
Problem: does not scale well to large p although it's $\mathcal{O}(p^2)$ ind. tests.



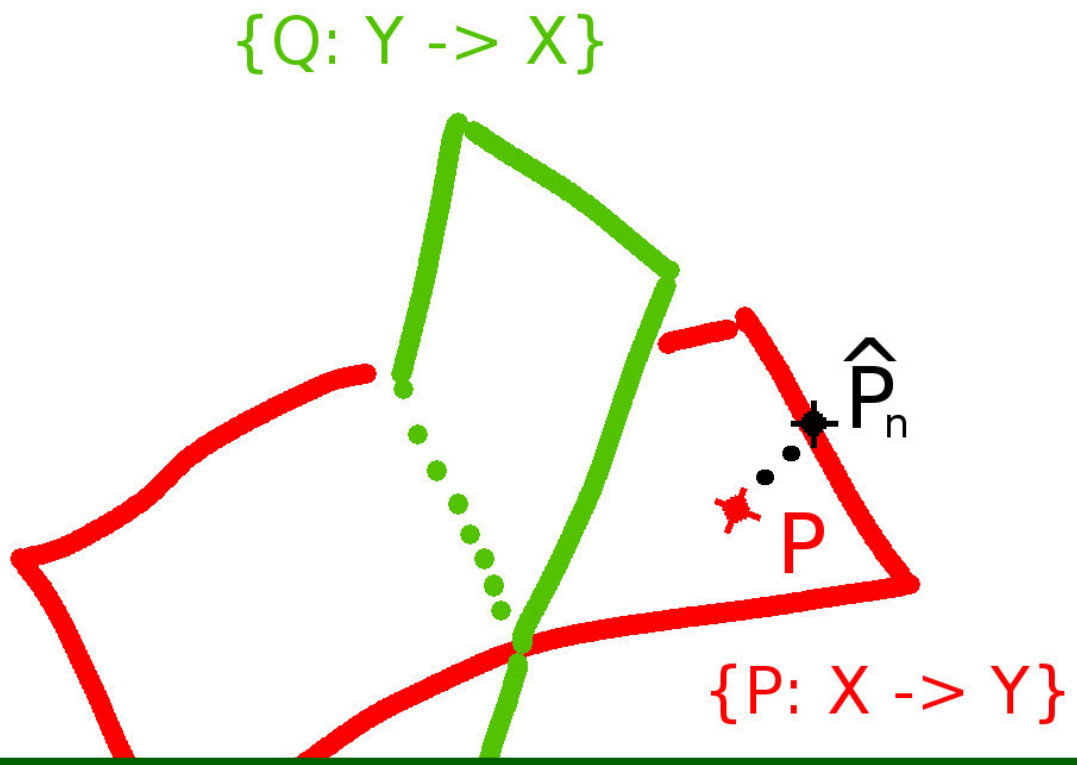
Does X cause Y or vice versa?



Does X cause Y or vice versa?



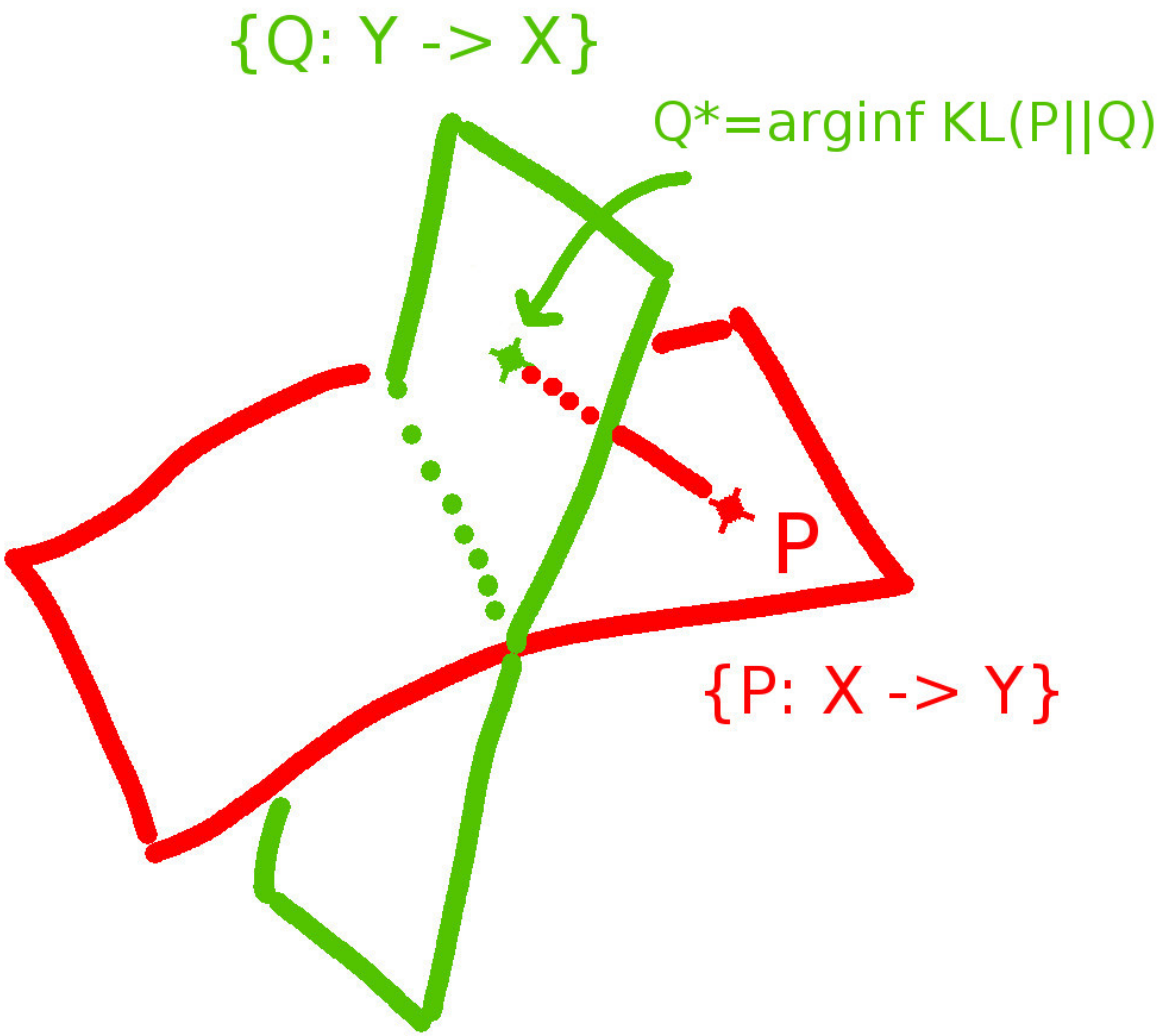
Does X cause Y or vice versa?



Method No. 2: Minimizing KL

Estimate the direction that corresponds to the closest subspace (details follow).

Does X cause Y or vice versa?



Does X cause Y or vice versa?

Proposition

Assume $P(X, Y)$ *is generated by*

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Does X cause Y or vice versa?

Proposition

Assume $P(X, Y)$ *is generated by*

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) > 0 \qquad \text{if } \beta \neq 0\,.$$

Does X cause Y or vice versa?

Proposition

Assume $P(X, Y)$ *is generated by*

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) = \frac{1}{2} \log \left(1 + 2\beta^2 \frac{\sigma_X^4}{\sigma_{N_Y}^2} \right)$$

Does X cause Y or vice versa?

Proposition

Assume $P(X, Y)$ *is generated by*

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

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) \geq$$

Does X cause Y or vice versa?

Proposition

Assume $P(X, Y)$ *is generated by*

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

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) \geq \mathrm{KL}\left(P(Y) \,||\, \mathcal{N}(0, \mathrm{var}\, Y)\right)$$

Does X cause Y or vice versa?

Proposition

Assume $P(X, Y)$ *is generated by*

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

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) \geq \mathrm{KL}\left(P(Y) \,||\, \mathcal{N}(0, \operatorname{var} Y)\right)$$

- gives us finite sample guarantees
- model misspefication: how much non-Gaussianity can we allow for

Does X cause Y or vice versa?

Proposition

Assume $P(X, Y)$ *is generated by*

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

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) \geq \mathrm{KL}\left(P(Y) \,||\, \mathcal{N}(0, \operatorname{var} Y)\right)$$

- gives us finite sample guarantees
- model misspefication: how much non-Gaussianity can we allow for
- Question: $\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) = \ldots$?

Can we reconstruct the whole causal network?

Can we reconstruct the whole causal network?

Theorem

let $P(X_1, \ldots, X_p)$ *be generated by an additive noise model (+Gaussian)*

$$X_i = f_i(X_{\mathbf{PA}_i}) + N_i$$

with jointly independent $N_i \sim \mathcal{N}(0, \sigma_i^2)$ *and differentiable, non-linear* f_i*. Then we can identify the corresponding DAG from* $P(X_1, \ldots, X_p)$*.*

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

Can we reconstruct the whole causal network?

Theorem

let $P(X_1, \ldots, X_p)$ *be generated by an additive noise model (+Gaussian)*

$$X_i = f_i(X_{\mathbf{PA}_i}) + N_i$$

with jointly independent $N_i \sim \mathcal{N}(0, \sigma_i^2)$ *and differentiable, non-linear* f_i*. Then we can identify the corresponding DAG from* $P(X_1, \ldots, X_p)$*.*

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

Surprising: This follows from identifiability in the bivariate case.

Can we reconstruct the whole causal network?

Theorem

let $P(X_1, \ldots, X_p)$ *be generated by a causal additive model (+Gaussian)*

$$X_i = \sum_{k \in \mathbf{PA}_i} f_{i,k}(X_k) + N_i$$

with jointly independent $N_i \sim \mathcal{N}(0, \sigma_i^2)$ *and differentiable, non-linear* $f_{i,k}$*. Then we can identify the corresponding DAG from* $P(X_1, \ldots, X_p)$*.*

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

Surprising: This follows from identifiability in the bivariate case.

Can we reconstruct the whole causal network?

Given $\hat{P}_n(X_1, \ldots, X_4)$. What now?

Can we reconstruct the whole causal network?

Given $\hat{P}_n(X_1, \ldots, X_4)$. What now?

Consider model classes

$$\mathcal{Q}_G := \{Q : Q \text{ generated by a causal additive model CAM with DAG } G\}$$

Optimize

$$\min_{\text{DAG } G} \inf_{Q \in \mathcal{Q}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

Can we reconstruct the whole causal network?

Given $\hat{P}_n(X_1, \ldots, X_4)$. What now?

Consider model classes

$$\mathcal{Q}_G := \{Q : Q \text{ generated by a causal additive model CAM with DAG } G\}$$

Optimize

$$\min_{\text{DAG } G} \inf_{Q \in \mathcal{Q}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{\Longleftrightarrow} \quad \min_{\text{DAG } G} \sum_{i=1}^{p} \log \mathrm{var}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Can we reconstruct the whole causal network?

Given $\hat{P}_n(X_1, \ldots, X_4)$. What now?

Consider model classes

$$\mathcal{Q}_G := \{Q : Q \text{ generated by a causal additive model CAM with DAG } G\}$$

Optimize

$$\min_{\text{DAG } G} \inf_{Q \in \mathcal{Q}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{\Longleftrightarrow} \min_{\text{DAG } G} \sum_{i=1}^{p} \log \mathrm{var}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Wait, there is no penalization on the number of edges!
$\rightsquigarrow$ fully connected graph, i.e. orderings

Can we reconstruct the whole causal network?

There are

$$186766007444320351866648169267 21$$

DAGs with 13 nodes.

Can we reconstruct the whole causal network?

There are

$$186766007444320351866648169267 21$$

DAGs with 13 nodes. There are only :-)

$$6227020800$$

orderings.

Idea: find causal order

Find the order of variables that maximizes likelihood and then perform classical variable selection. This is consistent

Can we reconstruct the whole causal network?

There are

$$1867660074443203518666481692672 1$$

DAGs with 13 nodes. There are only :-)

$$6227020800$$

orderings.

Idea: find causal order

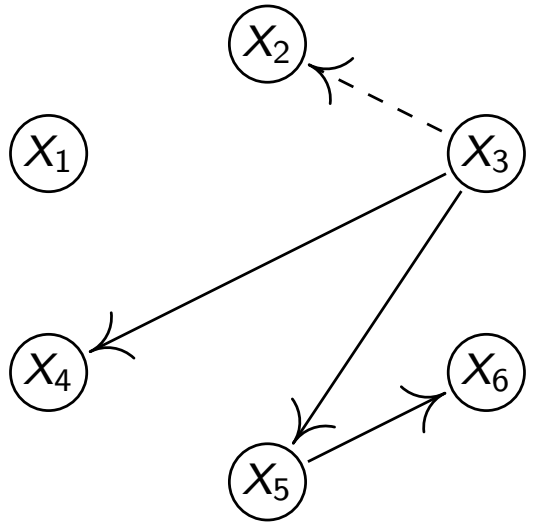
Find the order of variables that maximizes likelihood and then perform classical variable selection. This is consistent (but intractable).

P. Bühlmann, JP and J. Ernest: *CAM: Causal add. models, high-dim. order search and penalized regression, submitted*

Can we reconstruct the whole causal network?

–	0.2	0.1	0.1	0.1	0.3
0.4	–	0.1	0.1	0.1	0.1
0.1	0.6	–	–	–	0.4
0.1	0.1	–	–	0.1	0.1
0.1	0.1	–	0.1	–	–
0.3	0.1	–	0.1	–	–

include best edge $\longrightarrow$ recompute column



STEP 1: Greedy Addition. Include the edge that leads to the largest increase of the log-likelihood.

Can we reconstruct the whole causal network?

Theorem (Some correctness of greedy search for CAM)

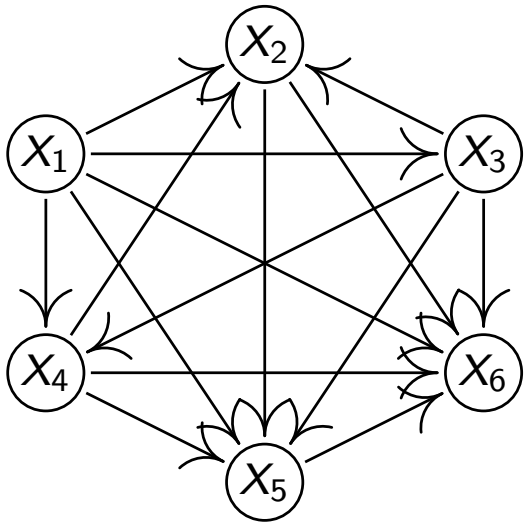
Assume that the skeleton of the correct DAG does not contain any cycles (plus some mild conditions/modifications).

Greedy addition then yields a correct causal order (in population).

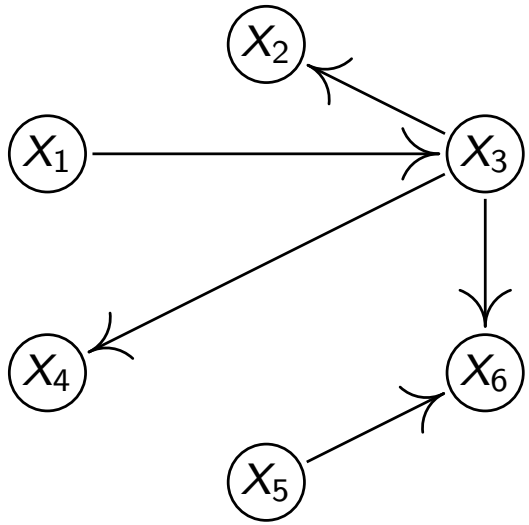
JP, S. Balakrishnan and M. Wainwright: *in progress.*

STEP 1: Greedy Addition. Include the edge that leads to the largest increase of the log-likelihood.

Can we reconstruct the whole causal network?

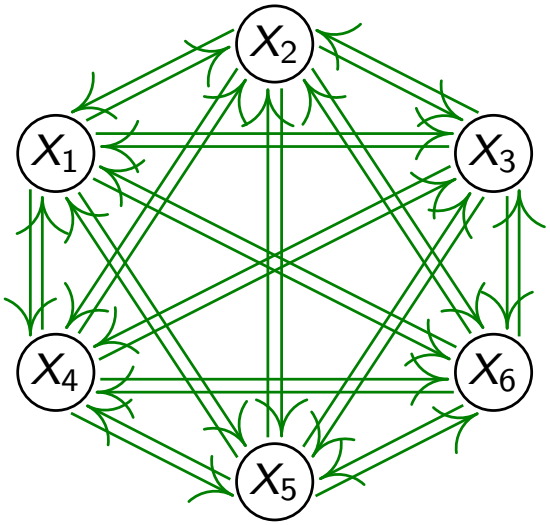


remove edges
by variable selection

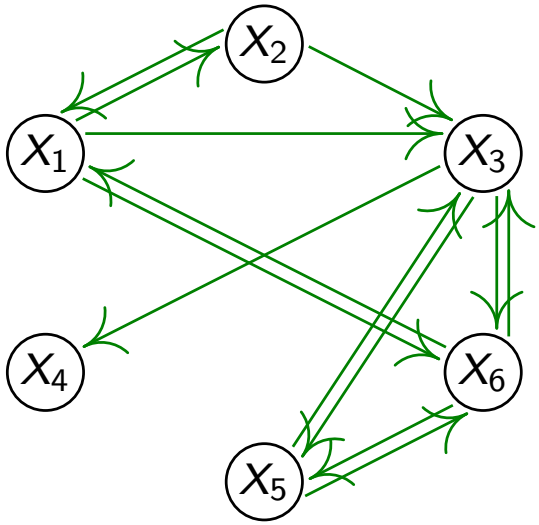


STEP 2: Variable Selection. For each node, remove non-relevant edges.

Can we reconstruct the whole causal network?



$\xrightarrow{\text{PNS}}$
by variable selection



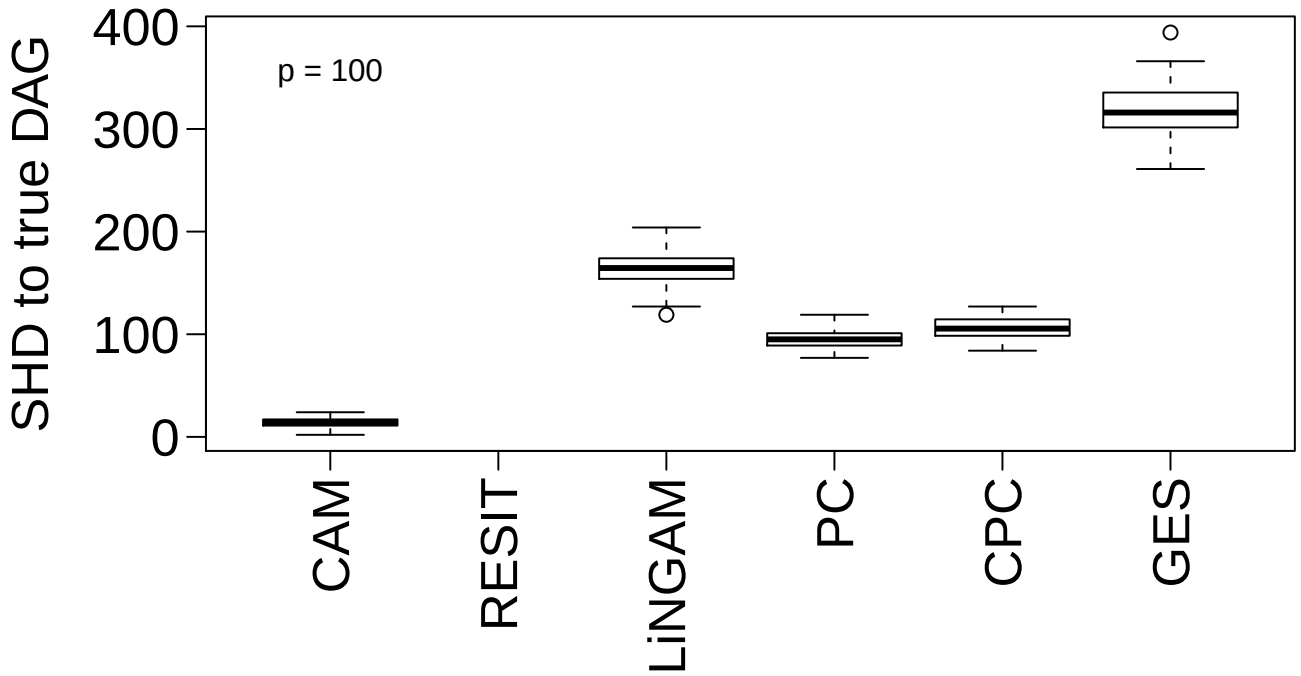
Easy to add for high-dim data:

STEP 0: Preliminary Neighbourhood Selection.

Can we reconstruct the whole causal network?

Simulations

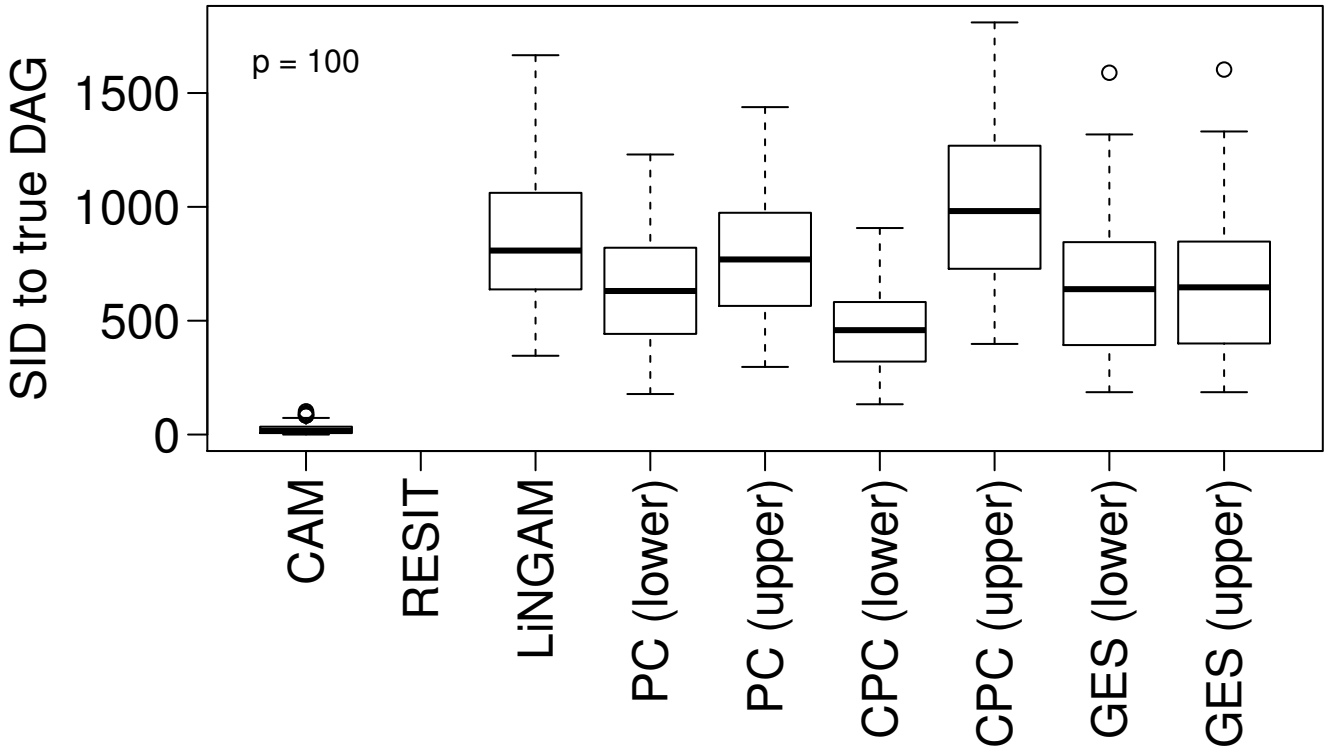
$p = 100, n = 200$, functions drawn from Gaussian Process



Can we reconstruct the whole causal network?

Simulations

$p = 100, n = 200$, functions drawn from Gaussian Process

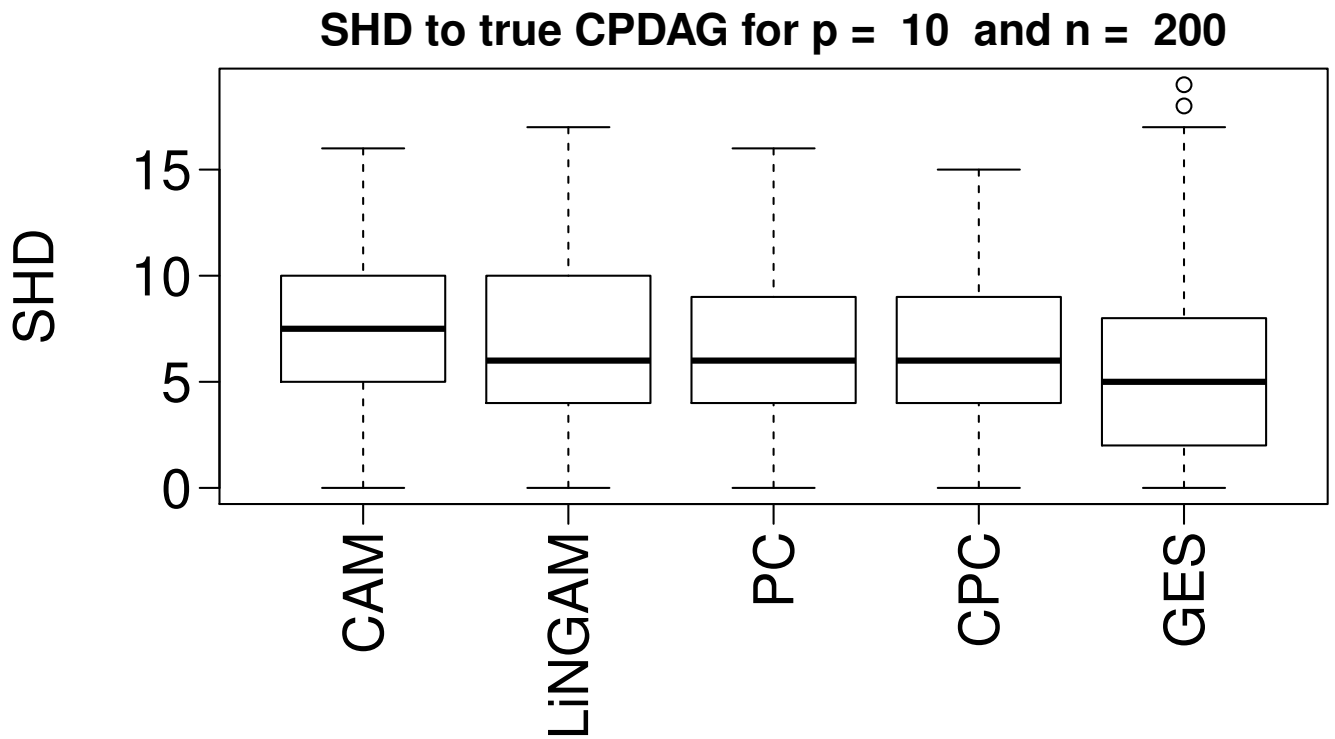


JP and P. Bühlmann: *Structural Intervention Distance (SID) for Evaluating Causal Graphs*, arXiv

Can we reconstruct the whole causal network?

Simulations

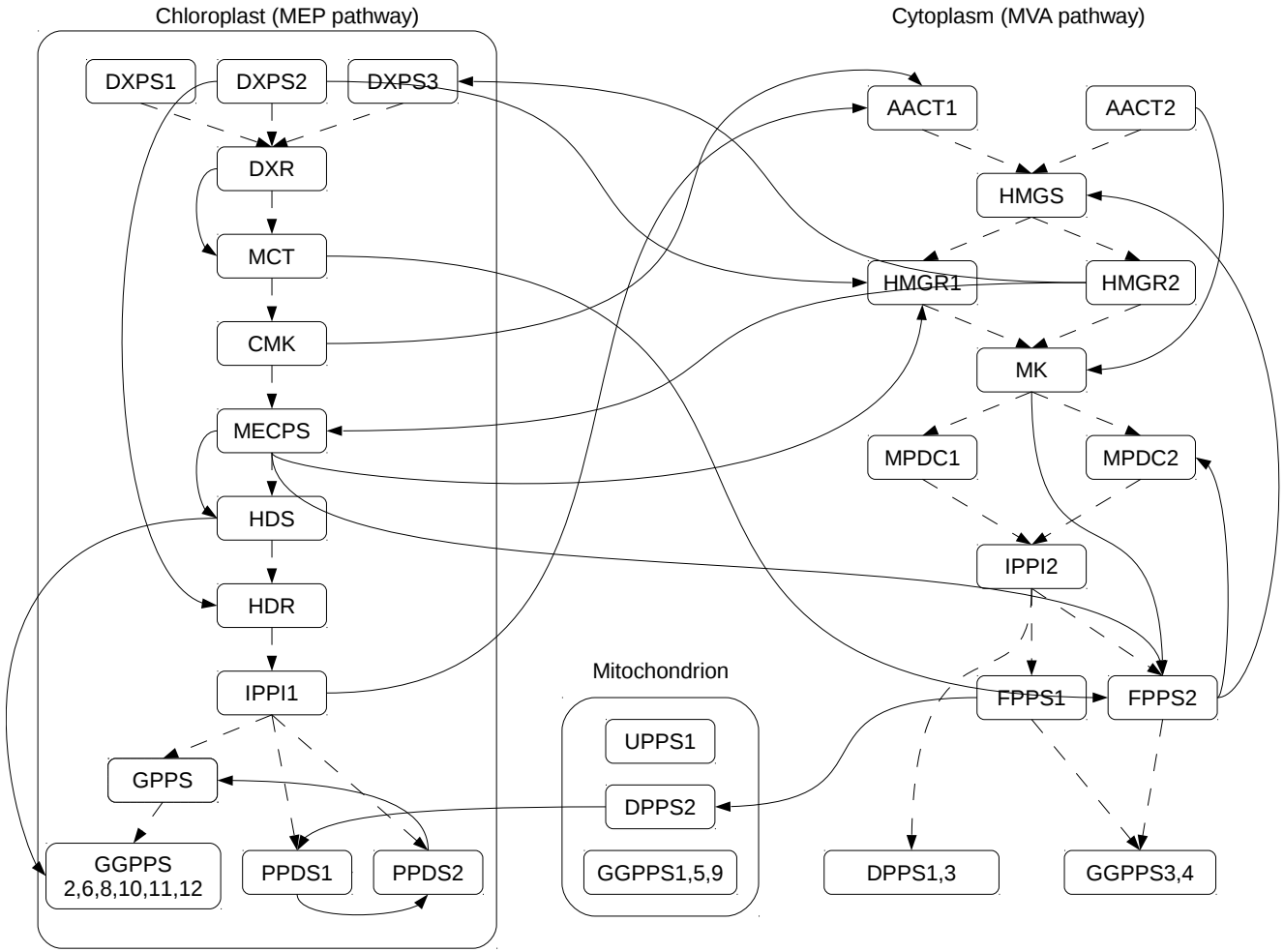
$p = 10, n = 200$, linear functions



Can we reconstruct the whole causal network?

Real Data: Arabidopsis thaliana

$p = 39$, $n = 118$, 20 most significant edges

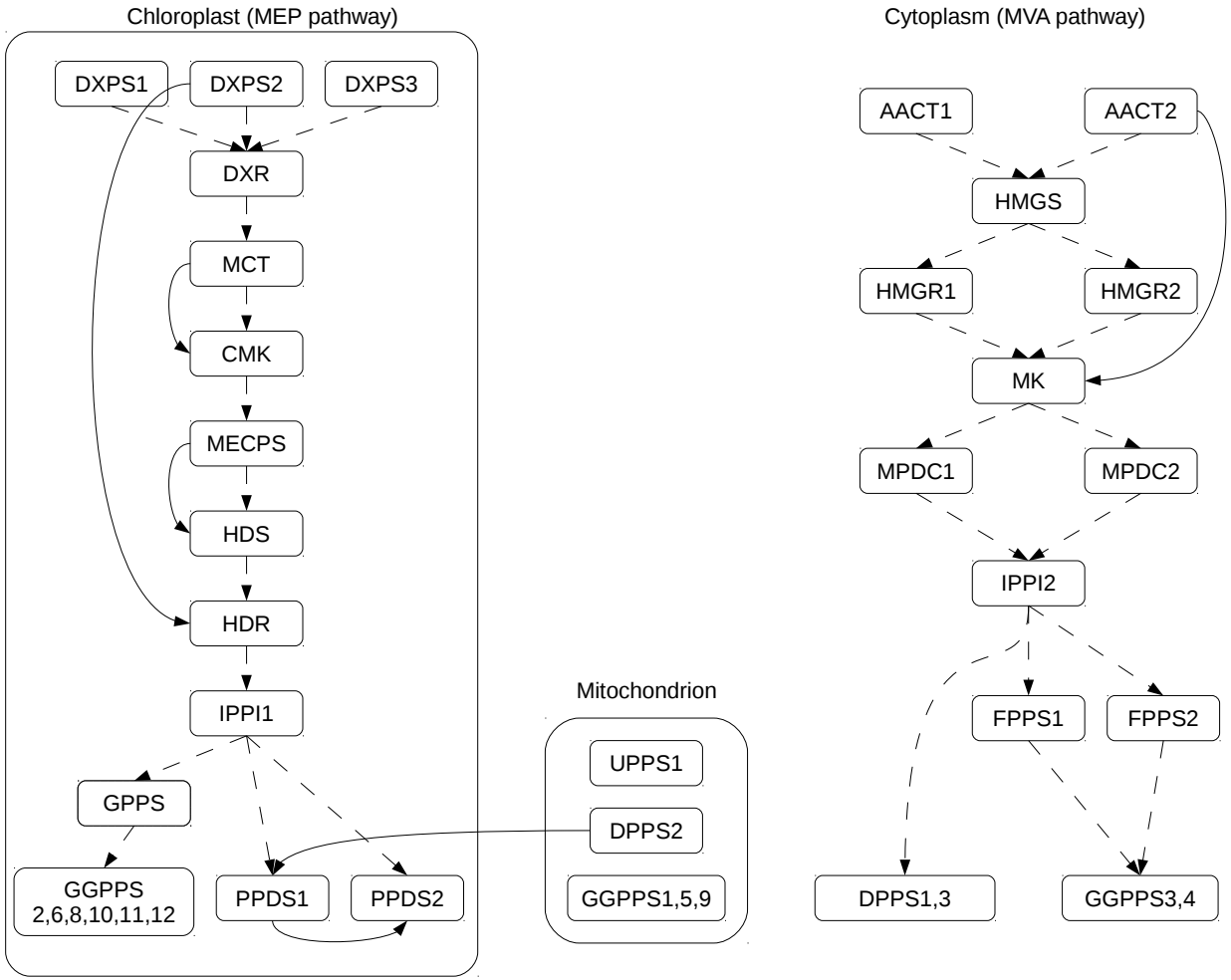


Wille et al.: *Sparse graphical Gaussian modeling of the isoprenoid gene network in arabidopsis thaliana.* Genome Biology, 5(11),

Can we reconstruct the whole causal network?

Real Data: Arabidopsis thaliana

$p = 39$, $n = 118$, stability selection: expected false positives ≤ 2



Wille et al.: *Sparse graphical Gaussian modeling of the isoprenoid gene network in arabidopsis thaliana.* Genome Biology, 5(11),

Can we reconstruct the whole causal network?

Real Data: A ground truth?

Assume we are given **observational** data ($j = 1 \dots 5000;\ k = 1 \dots 120$)

O_{kj} : expression level of gene j in observation k,

Can we reconstruct the whole causal network?

Real Data: A ground truth?

Assume we are given **observational** data ($j = 1 \dots 5000;\ k = 1 \dots 120$)

O_{kj} : expression level of gene j in observation k,

and some **interventional** data ($j = 1 \dots 5000;\ i = \dots$ in total 400 genes)

A_{ij} : expression level of gene j under a knock-down of gene i.

Can we reconstruct the whole causal network?

Real Data: A ground truth?

Assume we are given **observational** data ($j = 1 \ldots 5000;\ k = 1 \ldots 120$)

O_{kj} : expression level of gene j in observation k,

and some **interventional** data ($j = 1 \ldots 5000;\ i = \ldots$ in total 400 genes)

A_{ij} : expression level of gene j under a knock-down of gene i.

How do we evaluate causal inference methods?

Summary

Idea: Additive Noise Models

Structural assumptions like *additive noise models* lead to identifiability:

$$X_i = f_i(X_{pa(i)}) + N_i$$

Idea: causal order + variable selection

Find the order of variables that maximizes likelihood (by greedily adding edges) and then perform classical variable selection.

Summary

Idea: Additive Noise Models

Structural assumptions like *additive noise models* lead to identifiability:

$$X_i = f_i(X_{pa(i)}) + N_i$$

Idea: causal order + variable selection

Find the order of variables that maximizes likelihood (by greedily adding edges) and then perform classical variable selection.

Open Questions

- Identifiability (e.g. in terms of KL-distances) depending on non-linearity, for example.
- Hidden variables
- Do the assumptions (roughly) hold in practice?
- How do we evaluate causal inference methods?

Summary

Idea: Additive Noise Models

Structural assumptions like *additive noise models* lead to identifiability:

$$X_i = f_i(X_{pa(i)}) + N_i$$

Idea: causal order + variable selection

Find the order of variables that maximizes likelihood (by greedily adding edges) and then perform classical variable selection.

Open Questions

- Identifiability (e.g. in terms of KL-distances) depending on non-linearity, for example.
- Hidden variables
- Do the assumptions (roughly) hold in practice?
- How do we evaluate causal inference methods?

Dankeschön!!