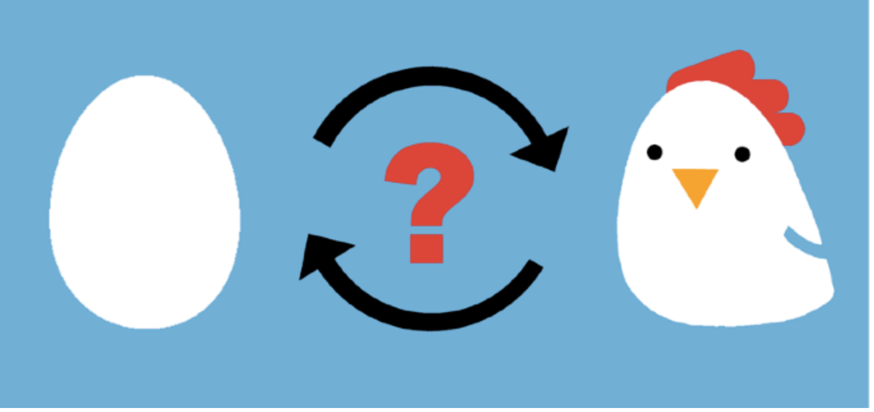


Causality and Machine Learning
(80-816/516)

Classes 15 (March 11, 2025)

Nonlinearity, Nonstationarity & Other Types of "Independence" for Causal Discovery

Instructor:

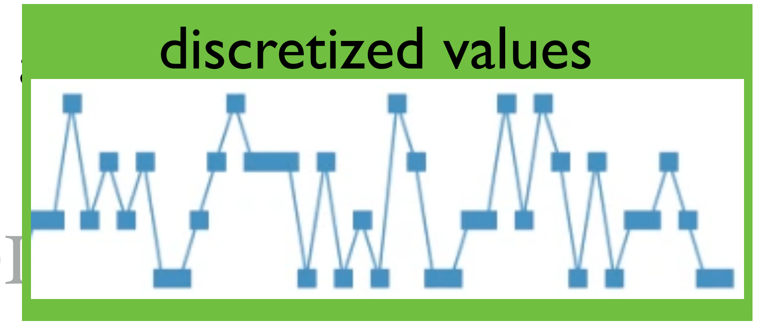
Kun Zhang (kunz1@cmu.edu)

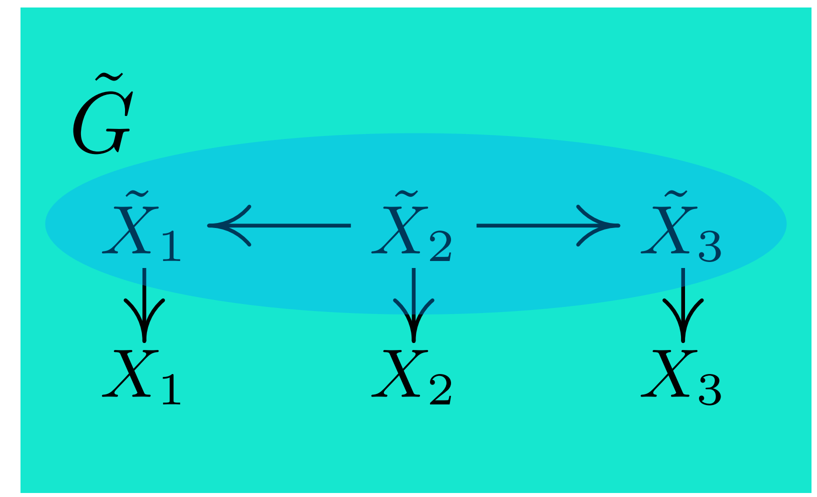
Zoom link: https://cmu.zoom.us/j/8214572323)

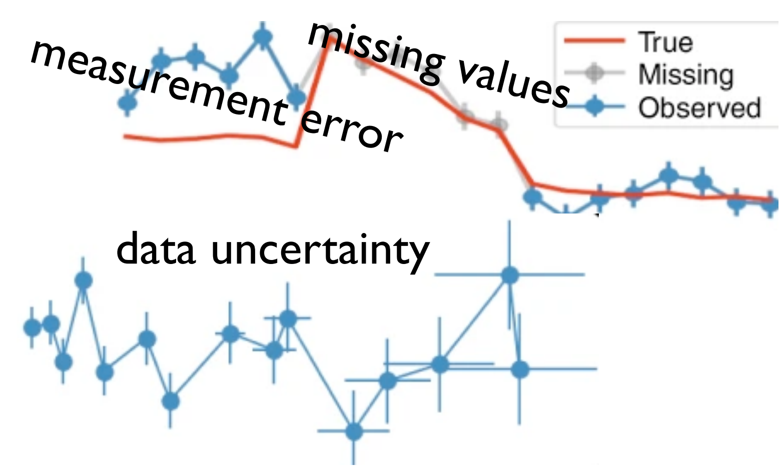
Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

Practical Issues in Causal Discovery...

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

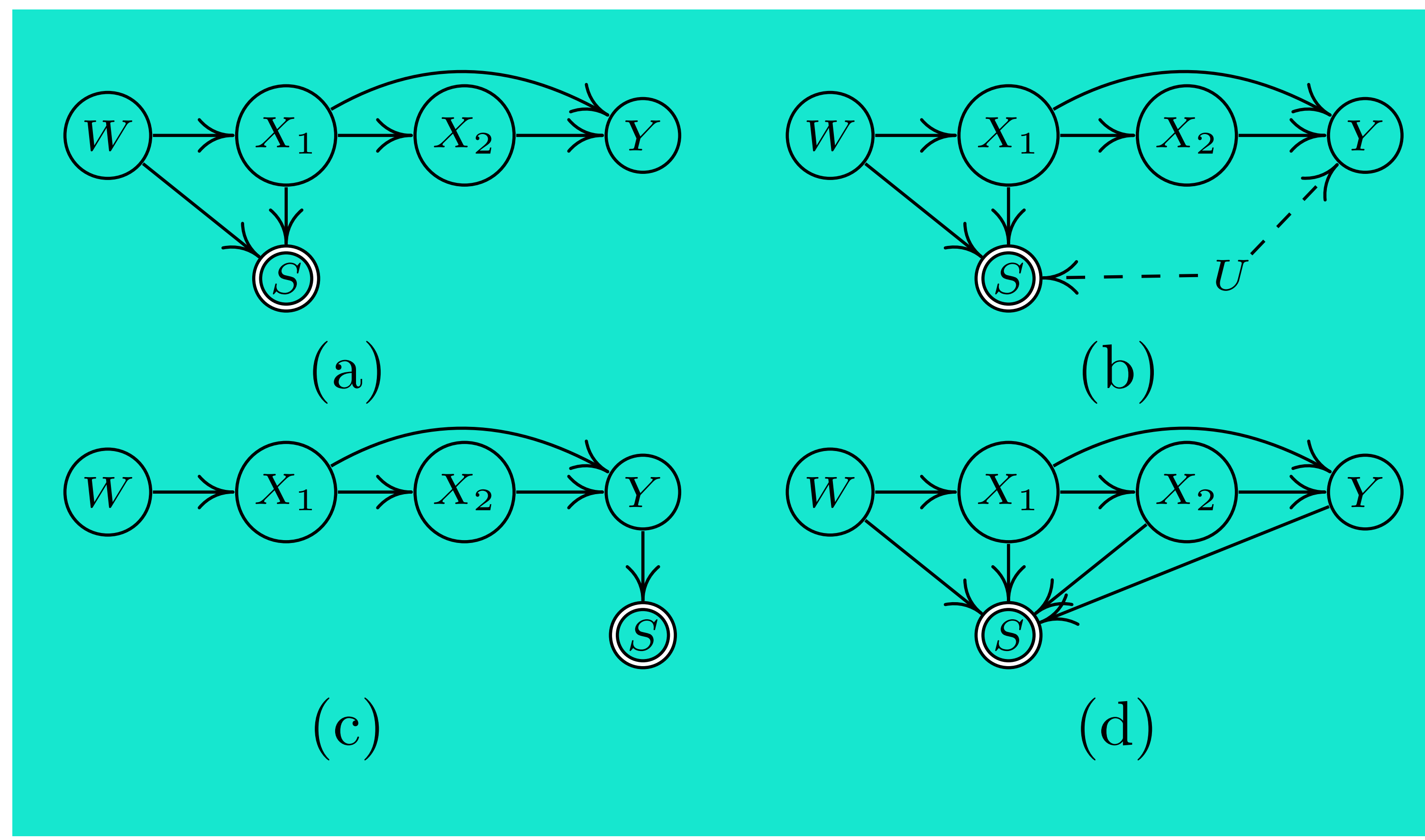






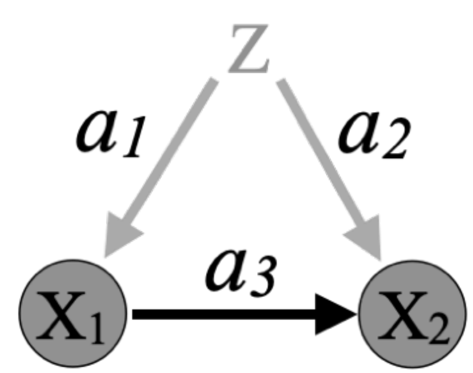
Practical Issues in Causal Discovery...

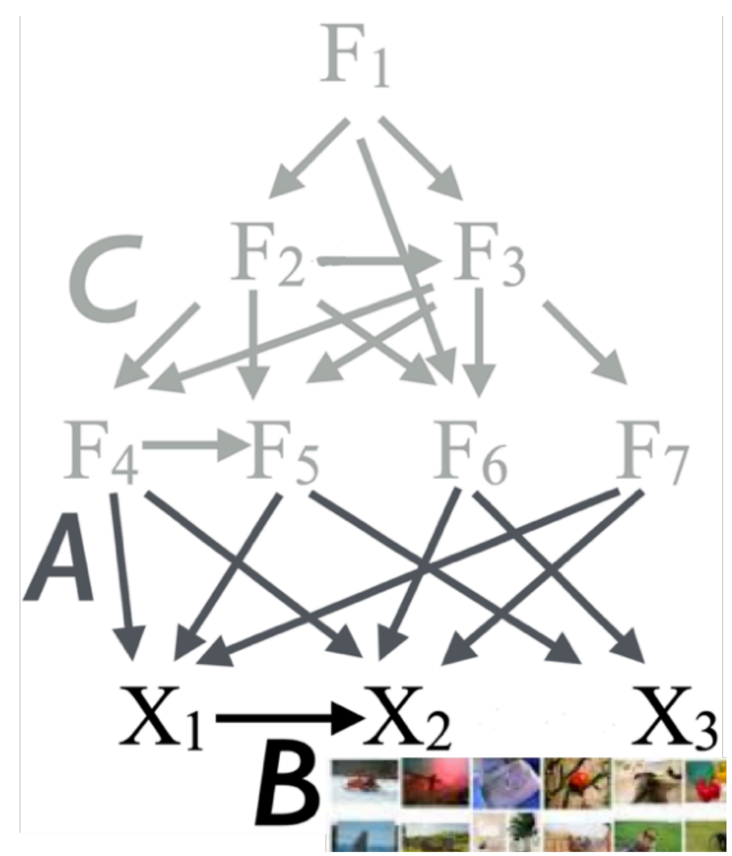
- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- **Selection bias** (Zhang et al., UAI'16)



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; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19); latent causal representation learning (Silva et al., JMLR'06; Xie et al., NeurIPS'20; Cai et al., NeurIPS'19; Adams et al., NeurIPS'21)





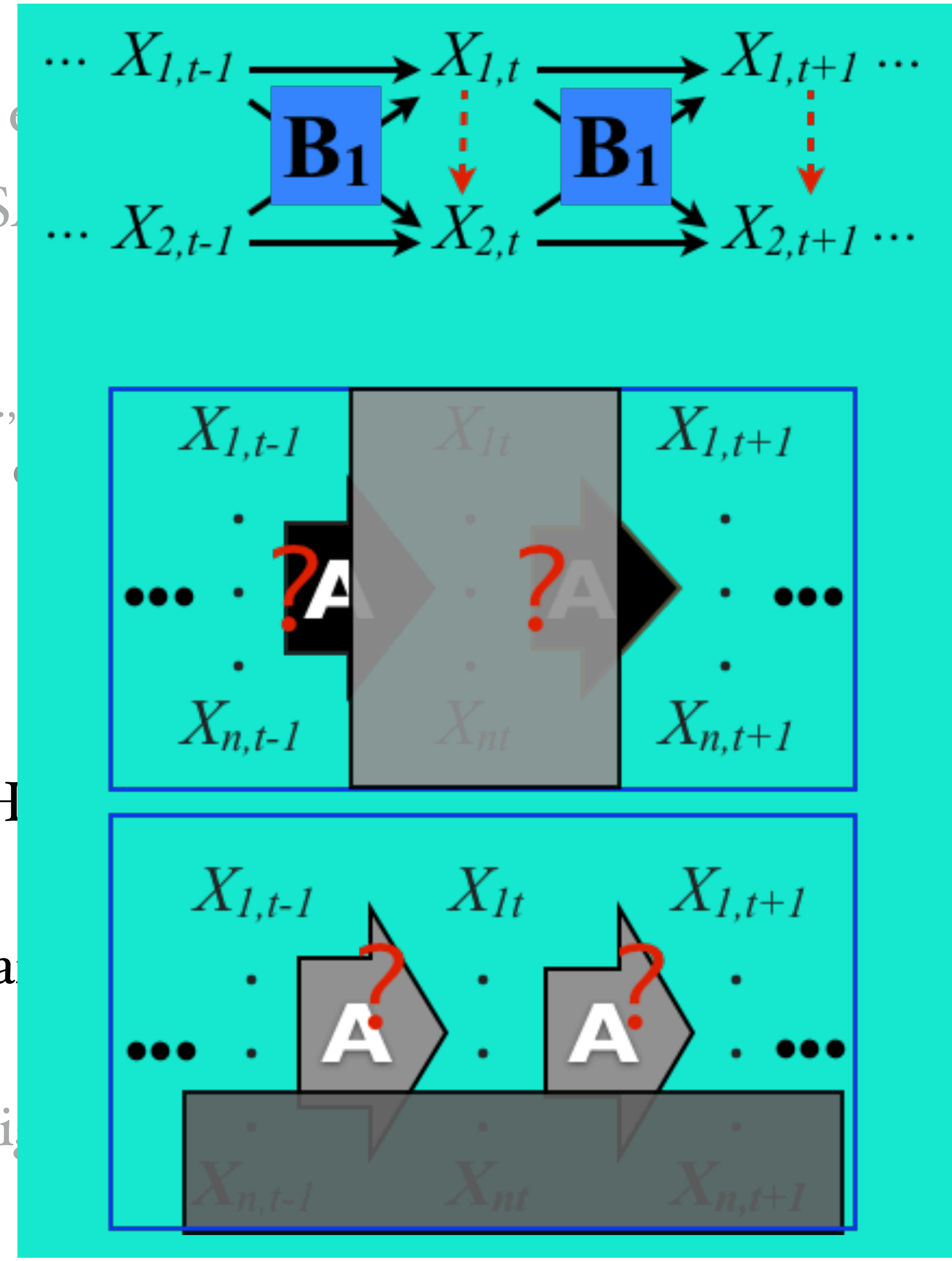
Practical Issues in Causal Discovery…

- 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); latent causal representation learning (Silva et al., JMLR'06; Xie et al., NeurIPS'20; Cai et al., NeurIPS'19; Adams et al., NeurIPS'21)
- **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]

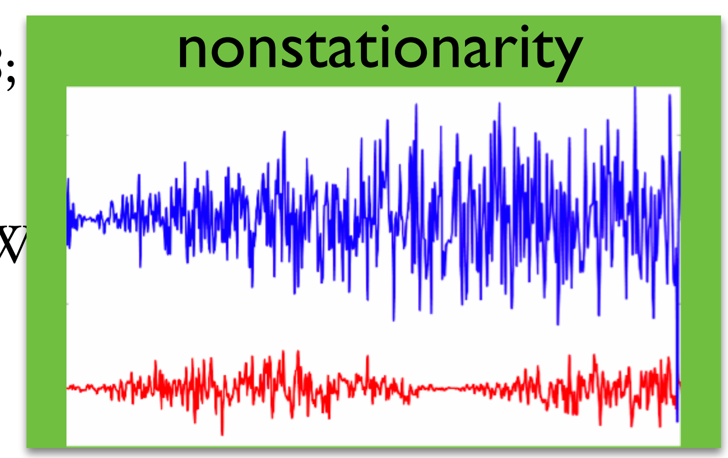
Practical Issues in Causal Discovery...

- 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
- Measurement error (Zhang et al., UAI'18; PS
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., representation learning (Silva et al., JMLR'06; Xie et al., NeurIPS'21)
- 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 (Spirtes 1995; Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19); latent causal representation learning (Silva et al., JMLR'06; Xie et al., NeurIPS'20; Cai et al., NeurIPS'19; Adams et al., NeurIPS'21)
- Missing values (Tu et al., AISTATS'19)
- Causality in time series
  - Time-delayed + instantaneous relations (Hyvarinen ICML'08; Hyvarinen et al., JMLR'10)
  - Subsampling / temporally aggregation (Danks & Plis, NIPS W UAI'17)
  - From partially observable time series (Geiger et al., ICML'15)

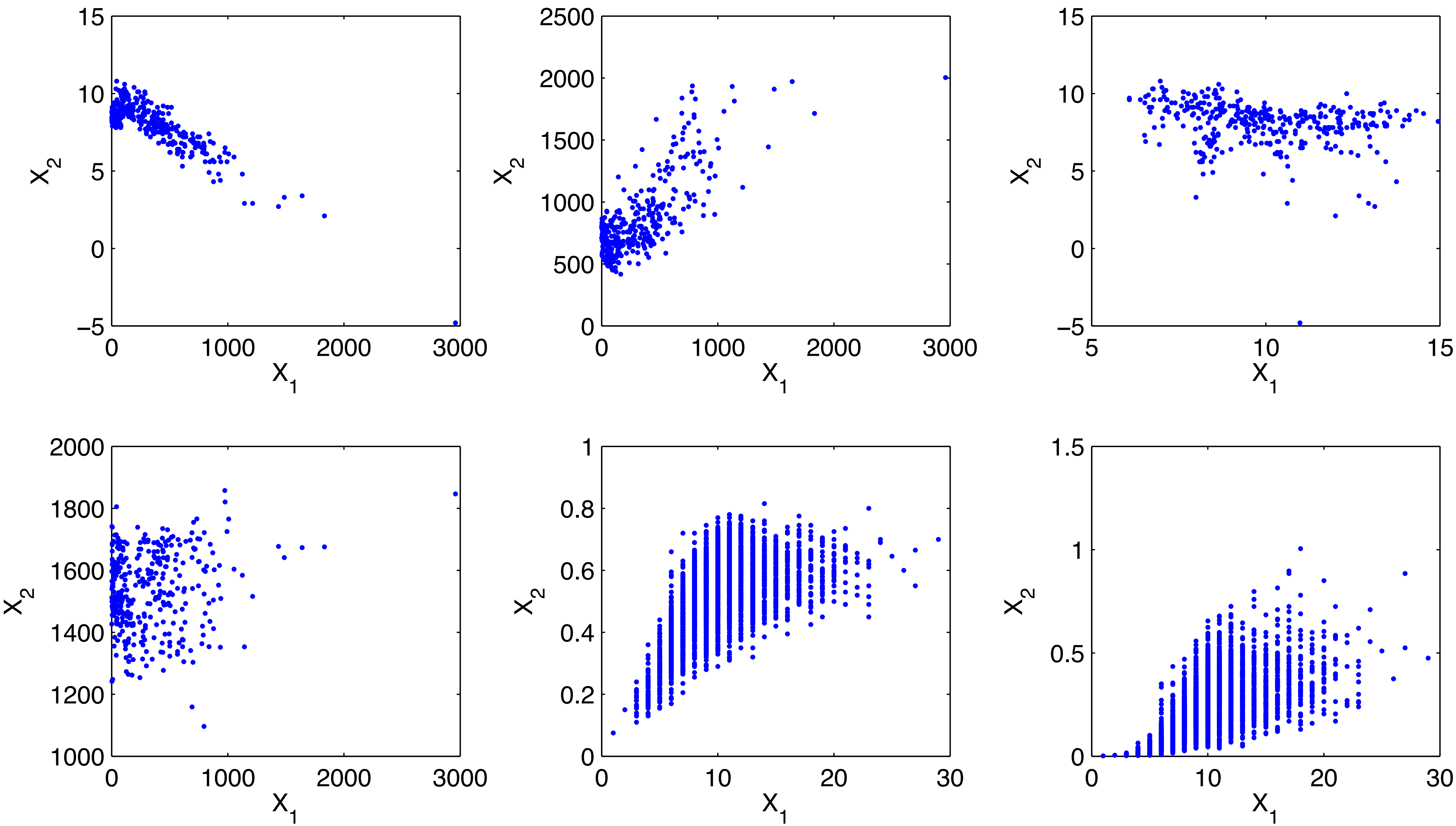


- **Nonstationary/heterogeneous data** (Zhang et al., IJCAI'17; Huang et al, ICDM'17, Ghassami et al., NIPS'18; Huang et al., ICML'19 & NIPS'19; Huang et al., JMLR'20)

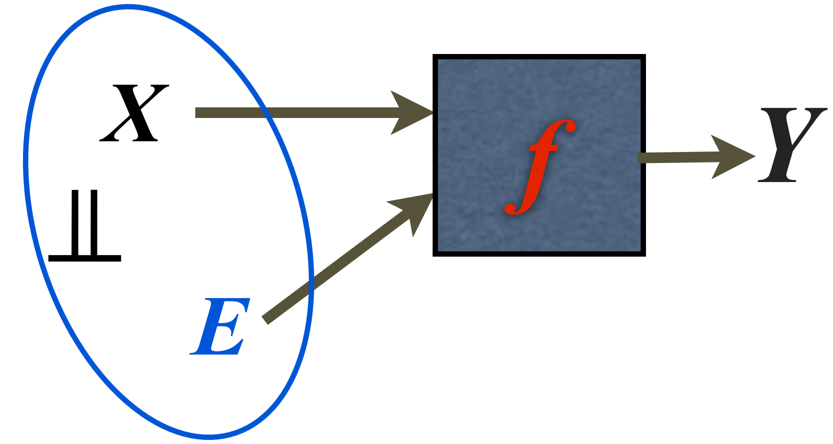
With Nonlinearities

- Model
- Identifiability
- Identification

Some Real Data Sets



Functional Causal Models



- Effect generated from cause with **independent noise** (Pearl et al.):

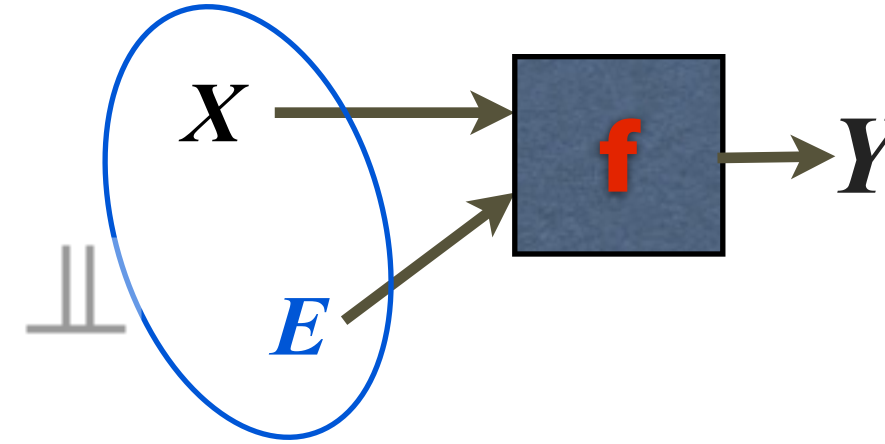
$$Y = f(X, E)$$

  - A way to encode the intuition "the generating process for X is 'independent' from that generates Y from X"

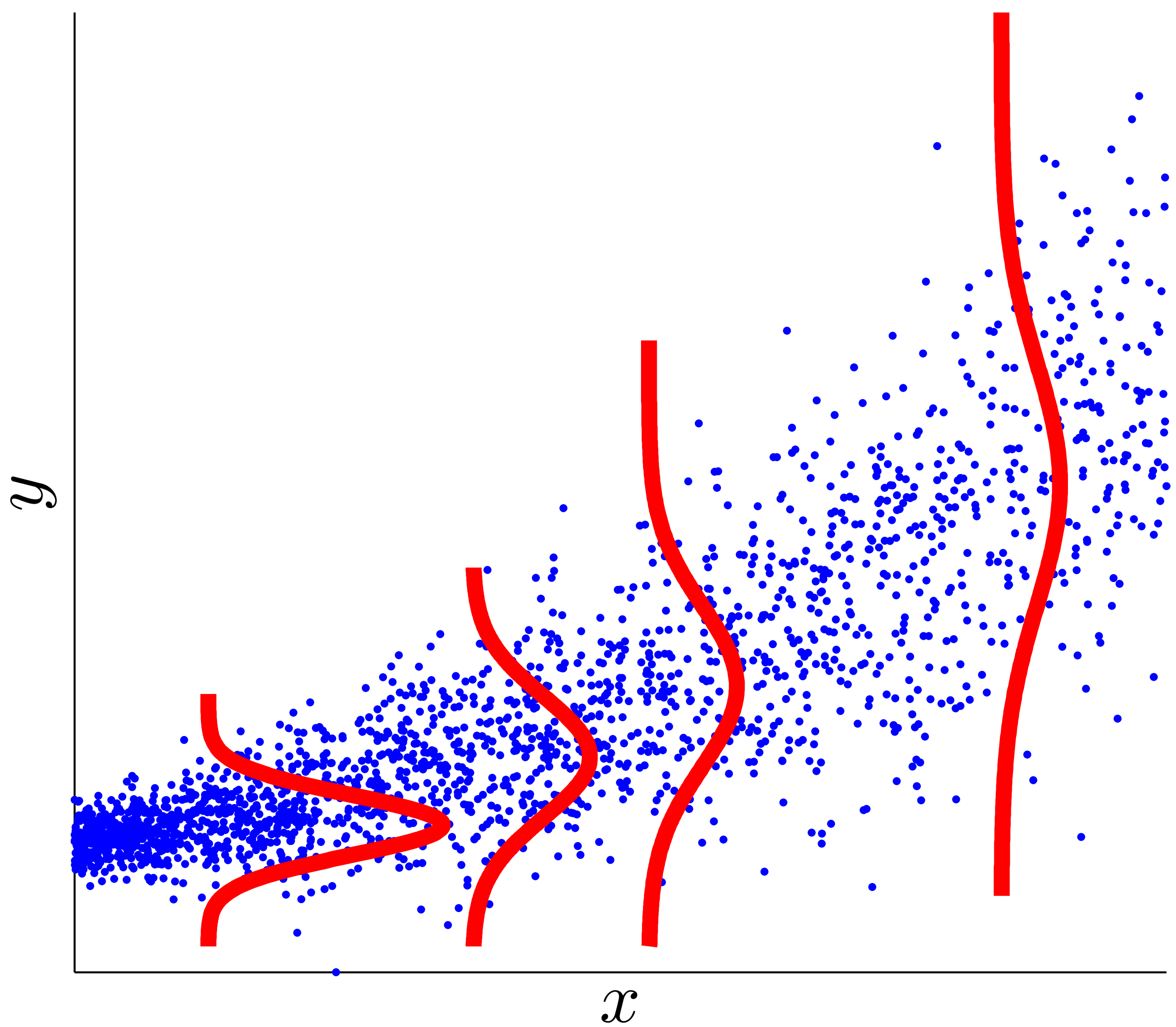
$$\begin{array}{ccccc} & & P(Y|X) & & \\ & & & \searrow & \\ P(X) & \rightarrow & X & \rightarrow & Y \end{array}$$

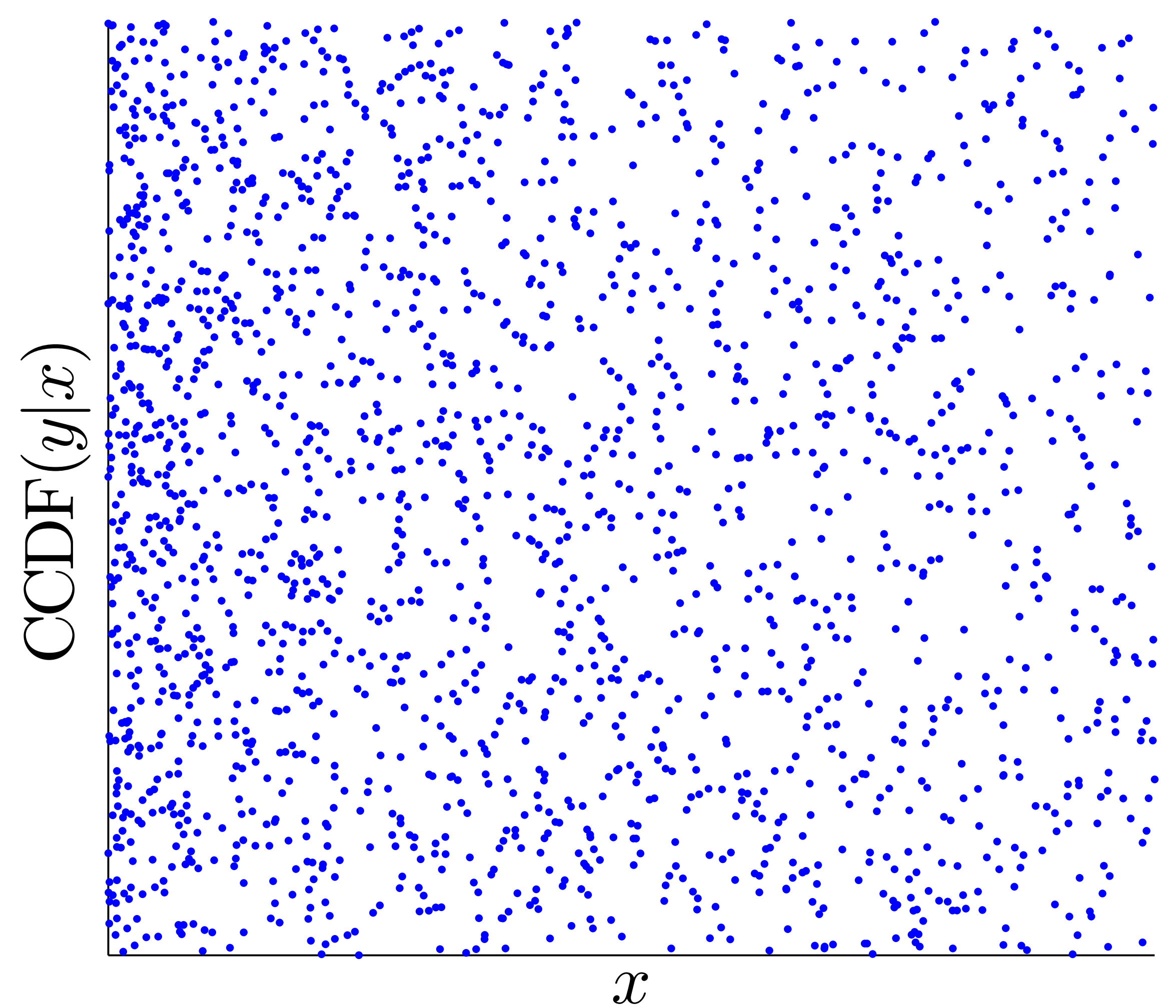
- :-(Without constraints on f, one can find independent noise for both directions (Darmois, 1951; Zhang et al., 2015)
  - Given any X_1 and X_2, E' := conditional CDF of $X_2 \mid X_1$ is always independent from X_1 and $X_2 = f(X_1, E')$
- :-) Structural constraints on f imply asymmetry

A Way to Construct Independent Error Term



- CDF(Y) is a random variable uniformly distributed over [0,1]





Zhang et al.(2015), On Estimation of Functional Causal Models: General Results and Application to Post-Nonlinear Causal Model, ACM Transactions on Intelligent Systems and Technology, Forthcoming

Then What Can We Do?

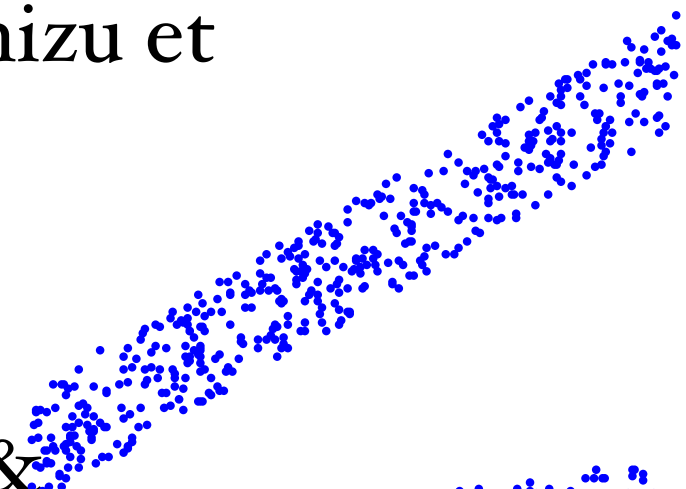
$$Y = f(X, E)$$

- The structure of f should be constrained & be able to approximate the true process...

FCMs with Which Causal Direction is Generally Identifiable

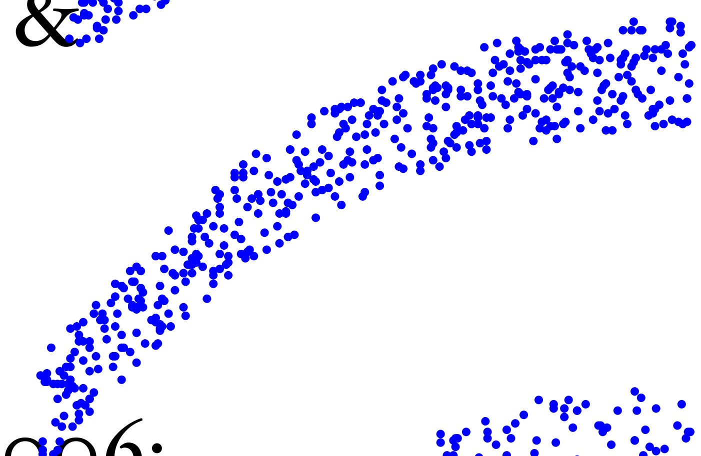
- Linear non-Gaussian acyclic causal model (Shimizu et al., '06)

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



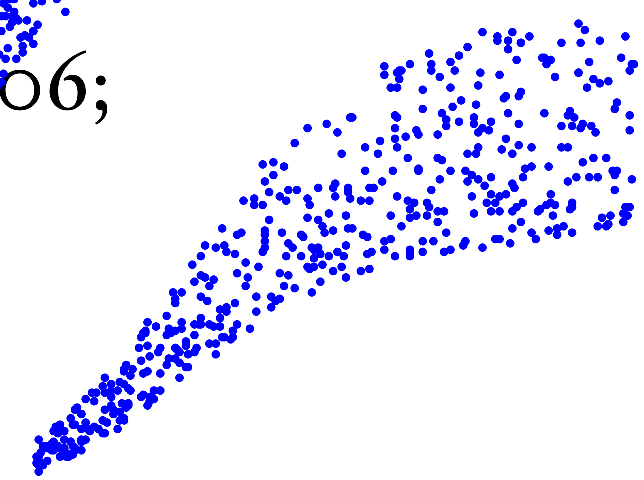
- Additive noise model (Hoyer et al., '09; Zhang & Hyvärinen, '09b)

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



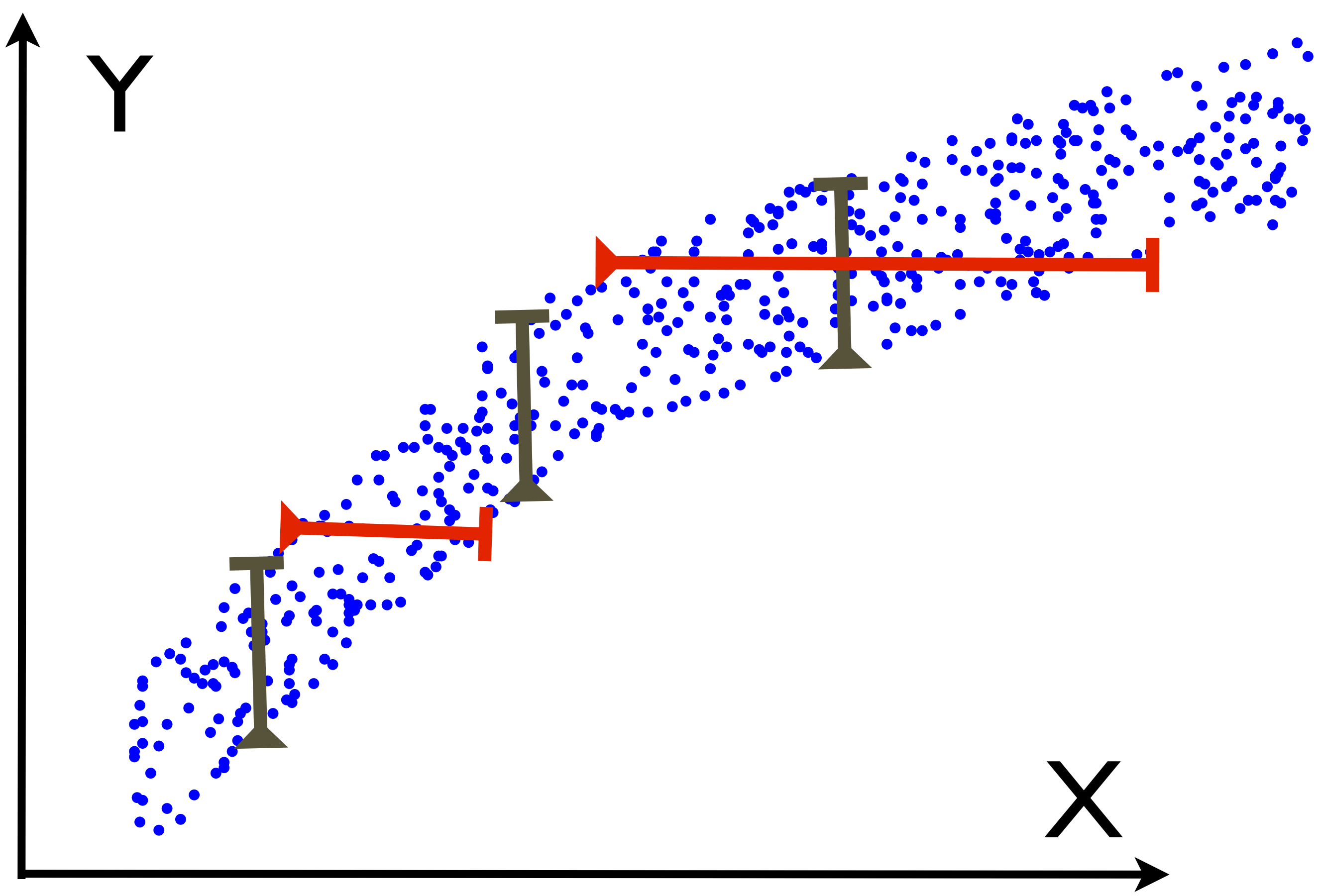
- Post-nonlinear causal model (Zhang & Chen, 2006; Zhang & Hyvärinen, '09a)

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



Causal Asymmetry with Nonlinear Additive Noise: Illustration

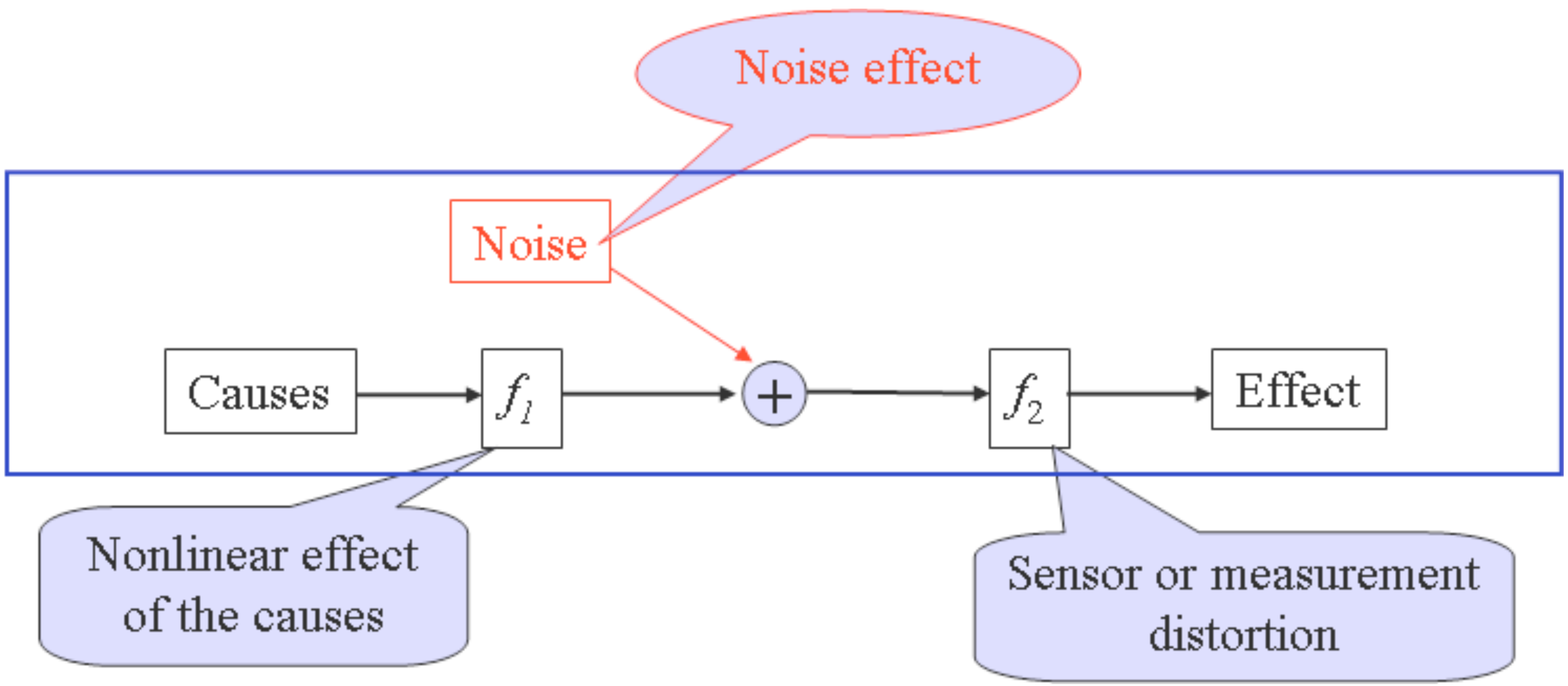
$$Y = f(X) + E \text{ with } E \perp\!\!\!\perp X$$



(Hoyer et al., 2009)

Three Effects usually encountered in a 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



- Represented by <u>post-nonlinear causal model with inner additive noise</u>

PNL Causal Model

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

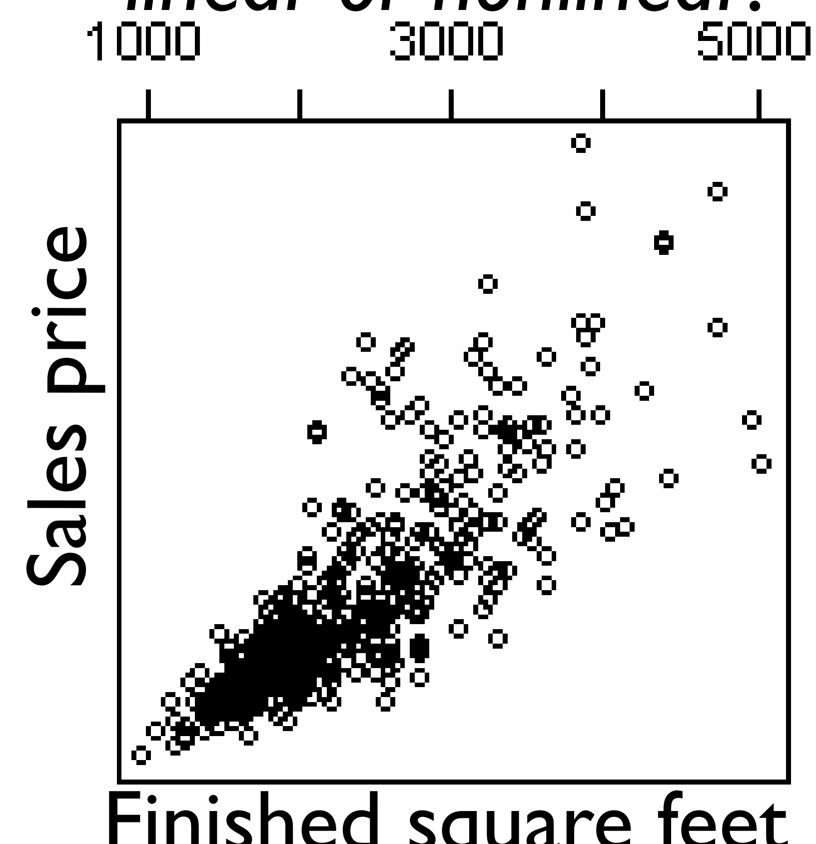
$f_{i,1}$: not necessarily invertible

e_i: noise/disturbance: independent from pa_i

- Special cases:
  - Linear models
  - Nonlinear additive noise models
  - Multiplicative noise models:

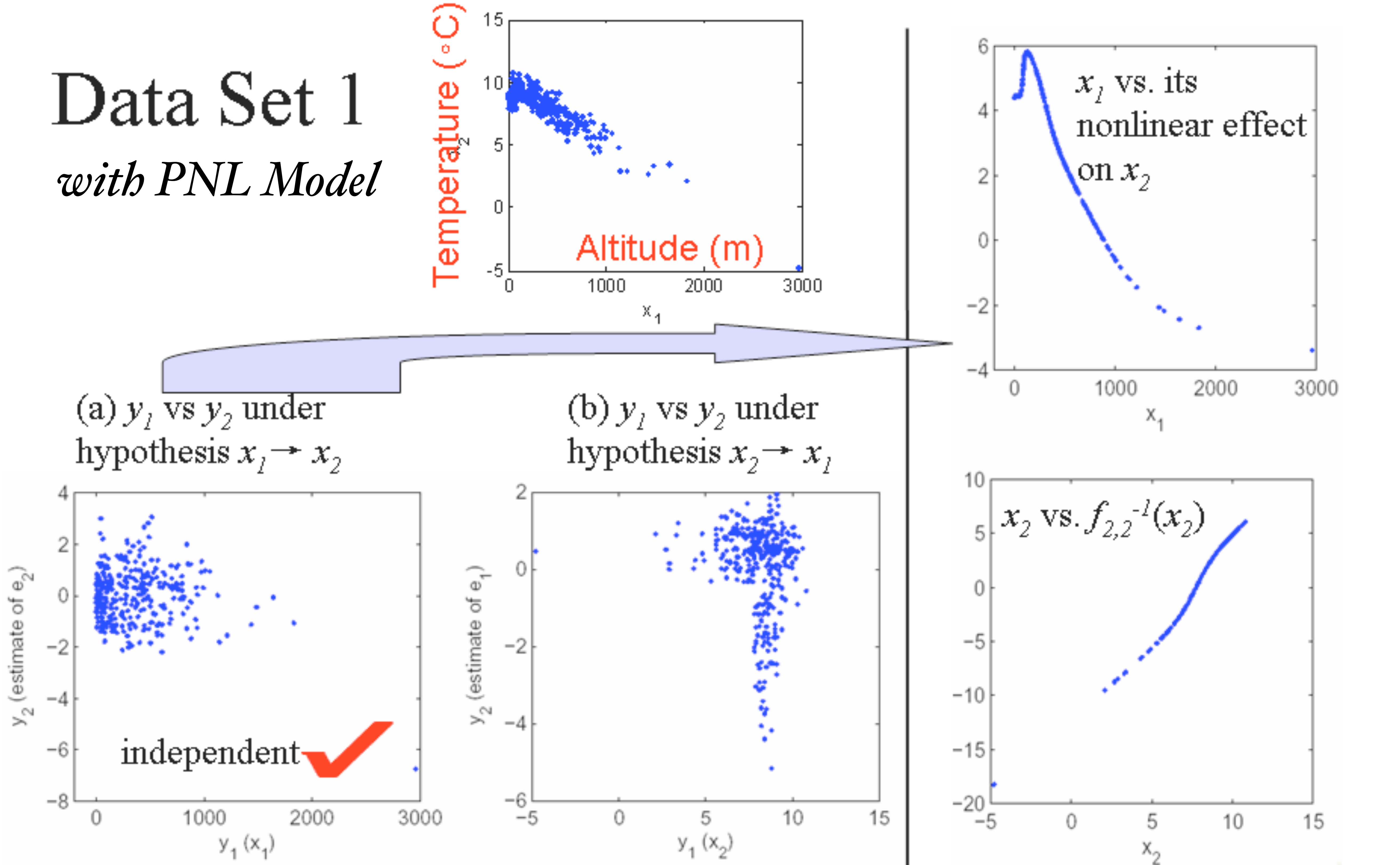
$$Y = X \cdot E = \exp\left(\log(X) + \log(E)\right)$$

linear or nonlinear?



Data Set 1

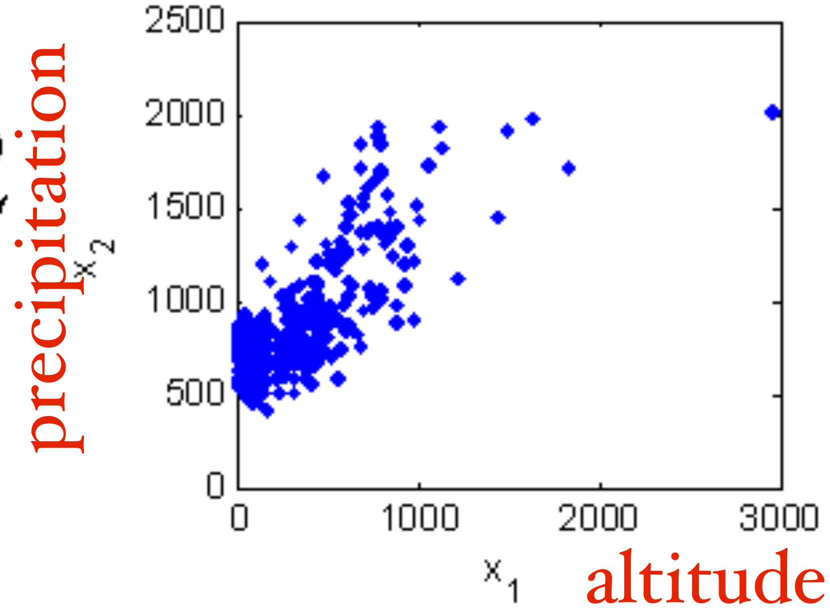
with PNL Model

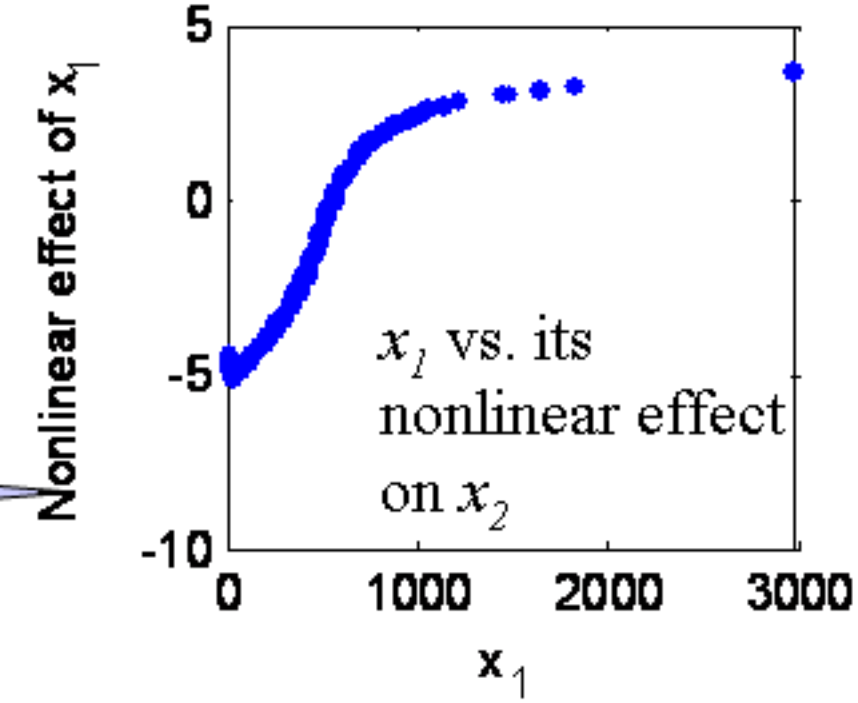


Independence test results on y_1 and y_2 with different assumed causal relations

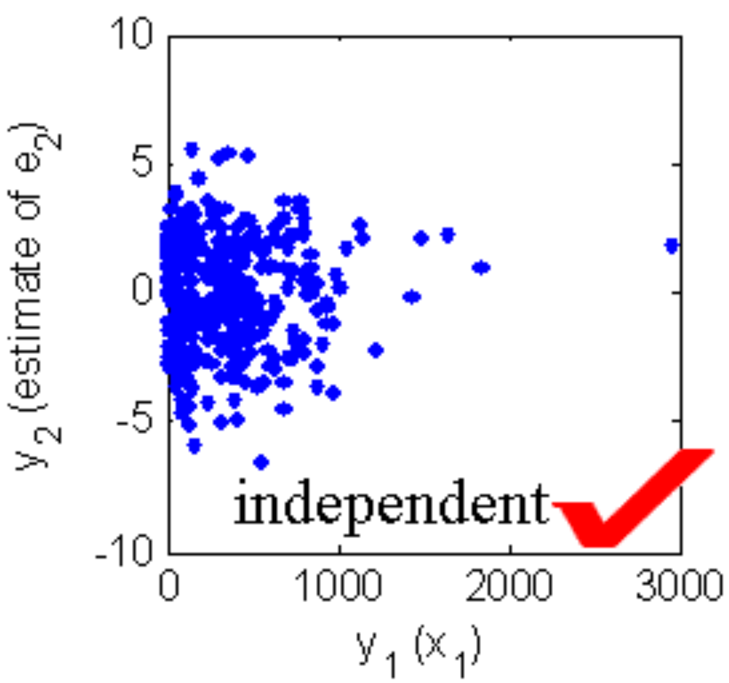
Data Set	$x_1 \rightarrow x_2$ assumed		$x_2 \rightarrow x_1$ assumed	
	Threshold ($\alpha = 0.01$)	Statistic	Threshold ($\alpha = 0.01$)	Statistic
#1	2.3×10^{-3}	1.7×10^{-3}	2.2×10^{-3}	6.5×10^{-3}

Data Set 2

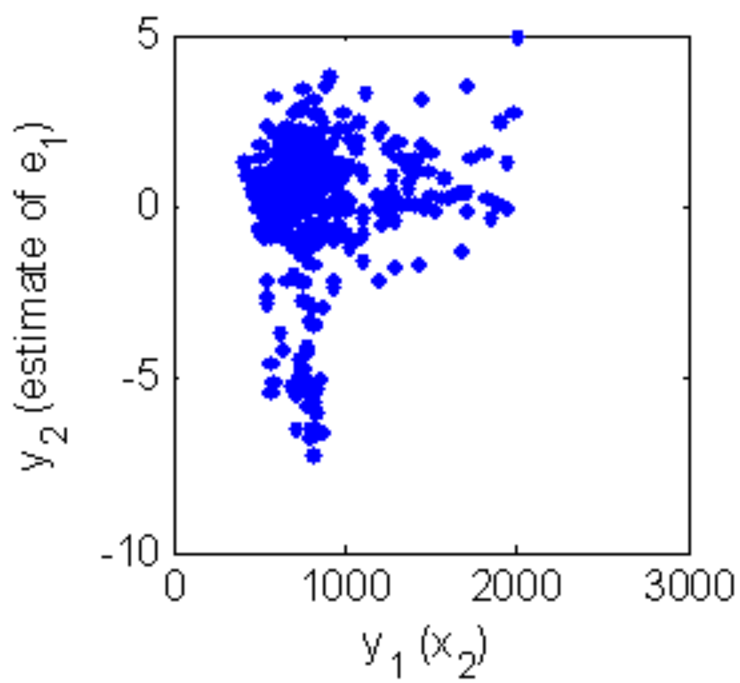


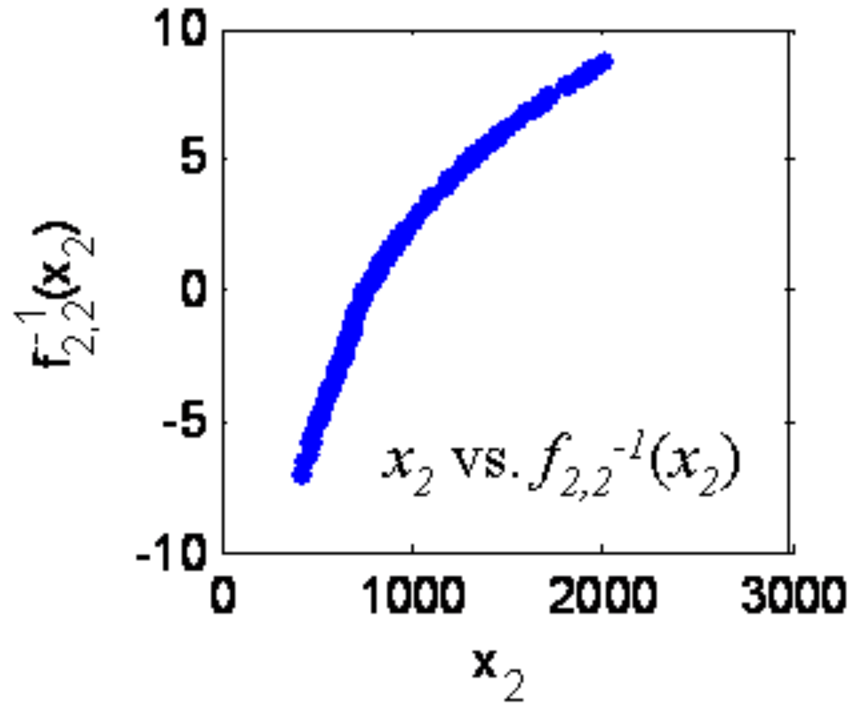


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

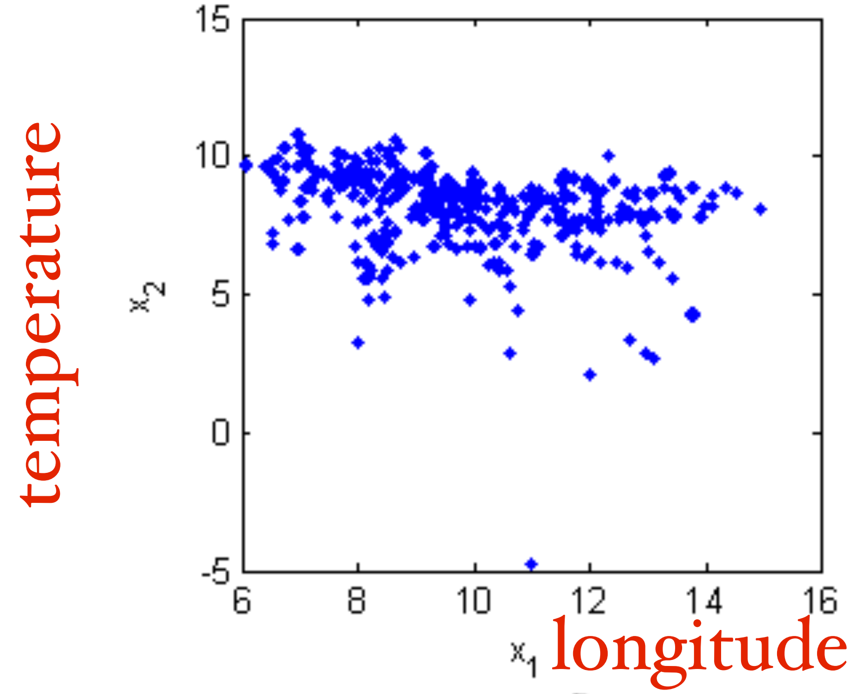


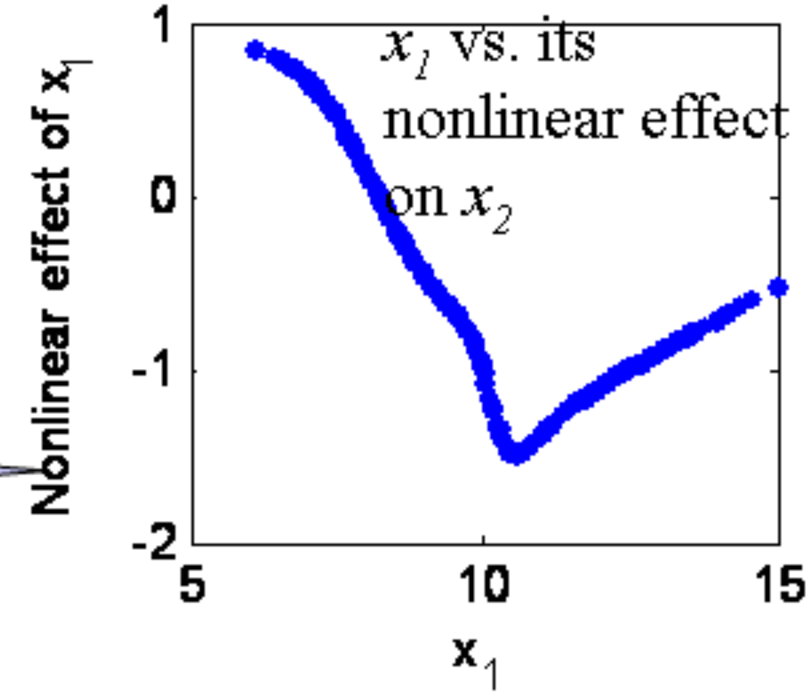
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$





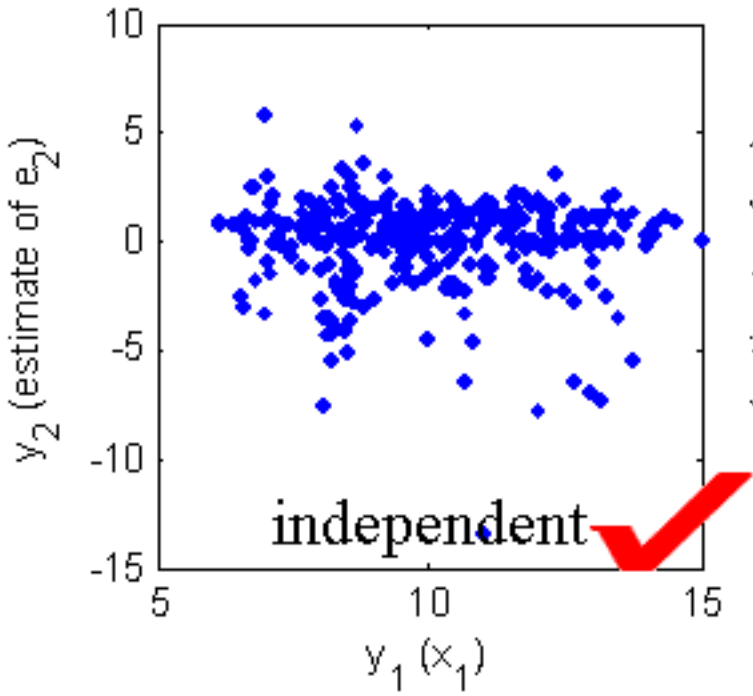
Data Set 3

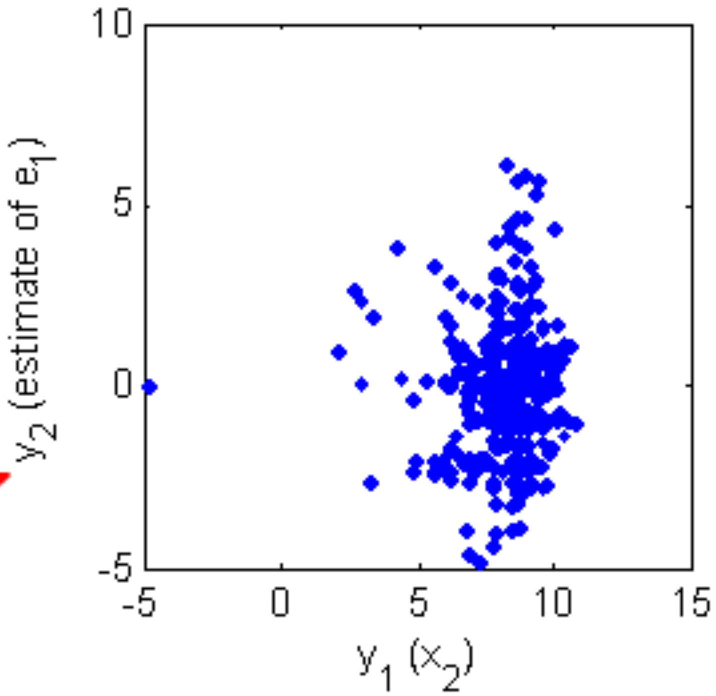


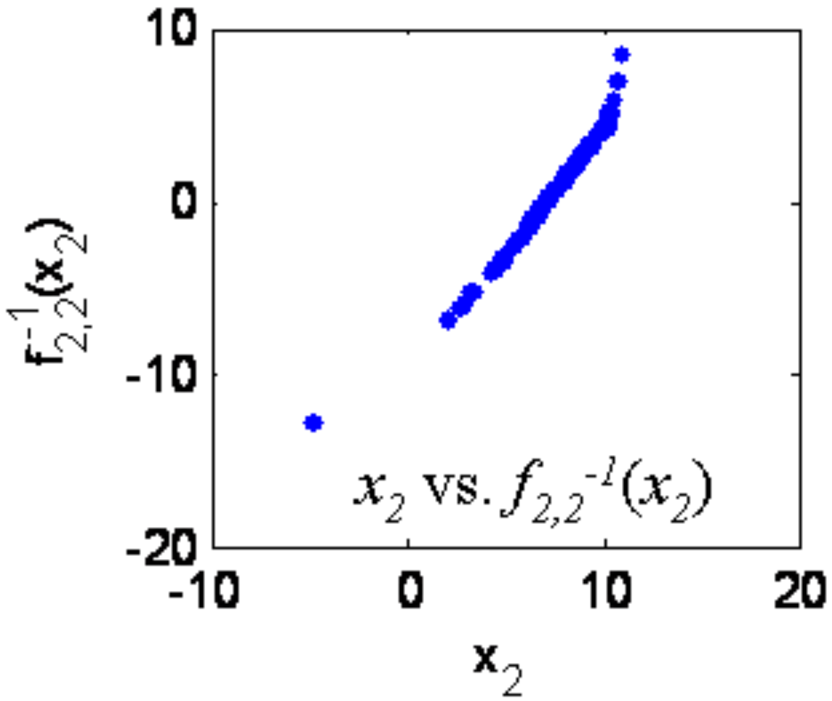


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

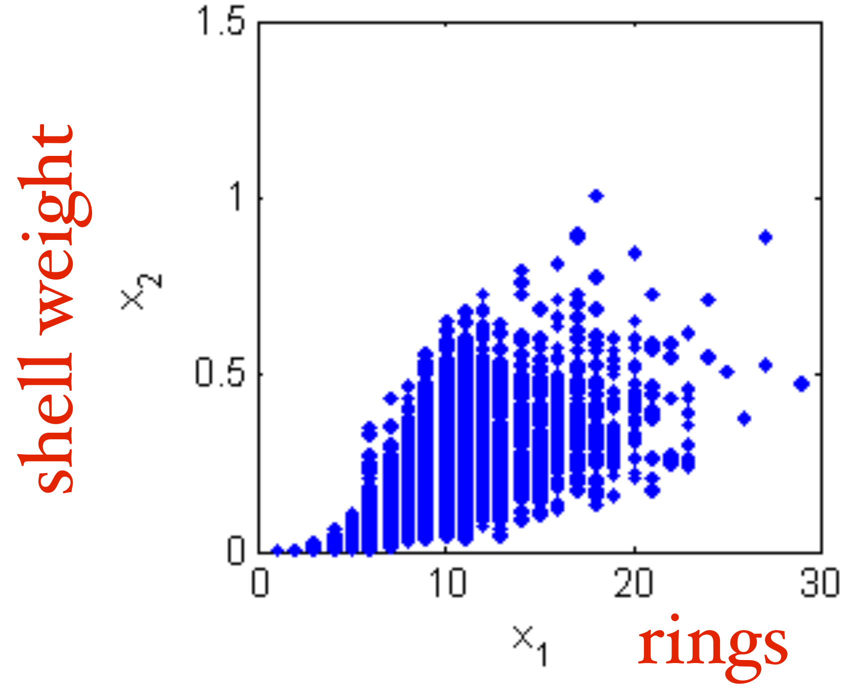
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$

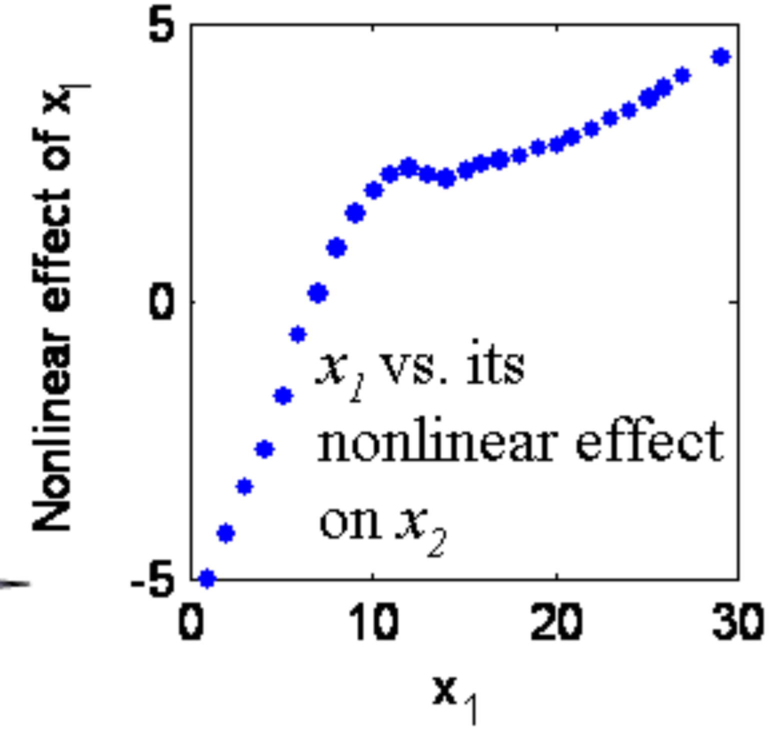






Data Set 6

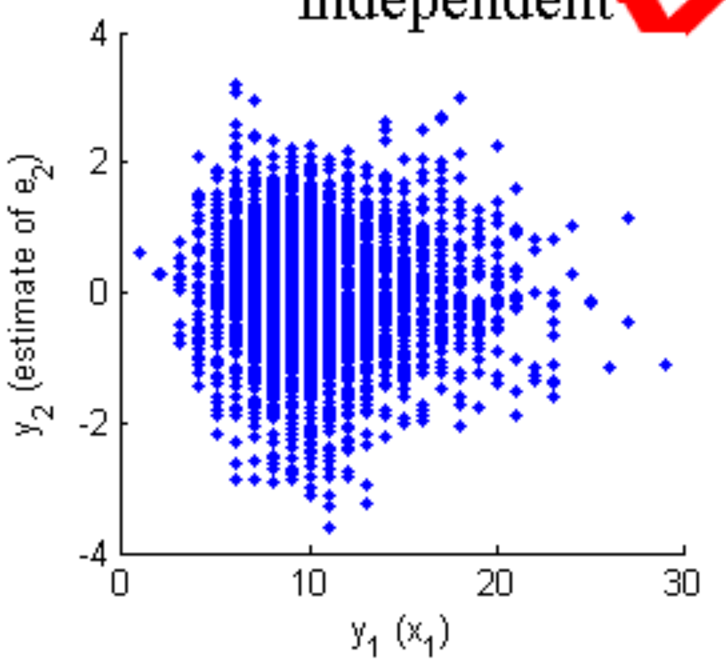


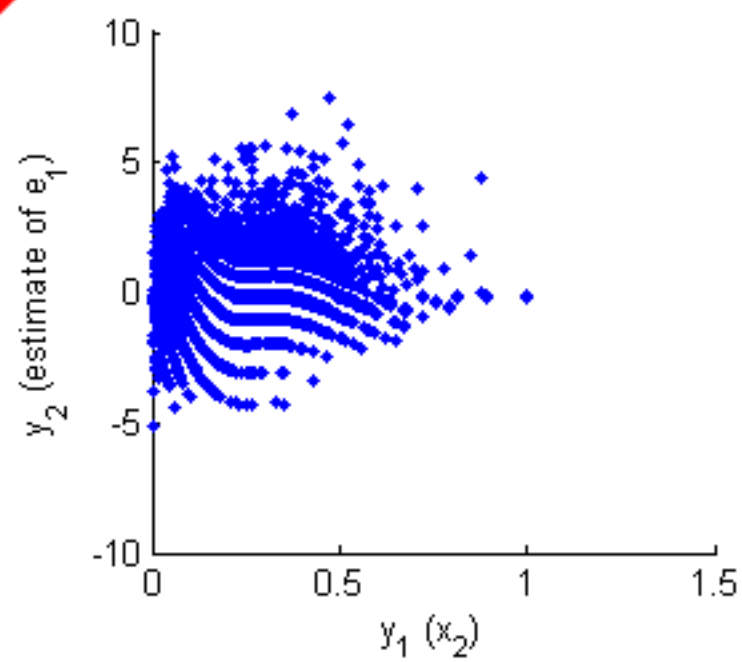


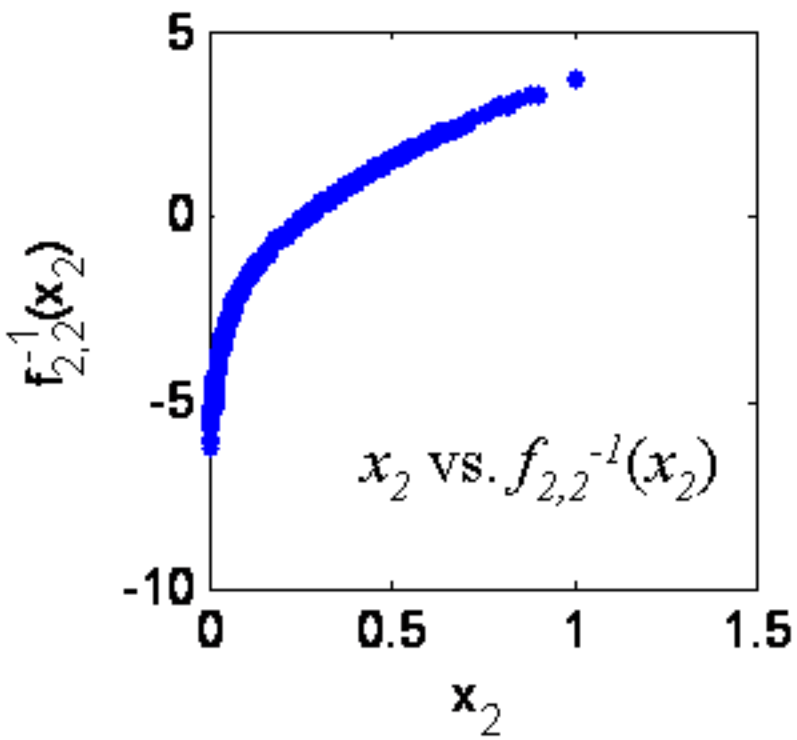
(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

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

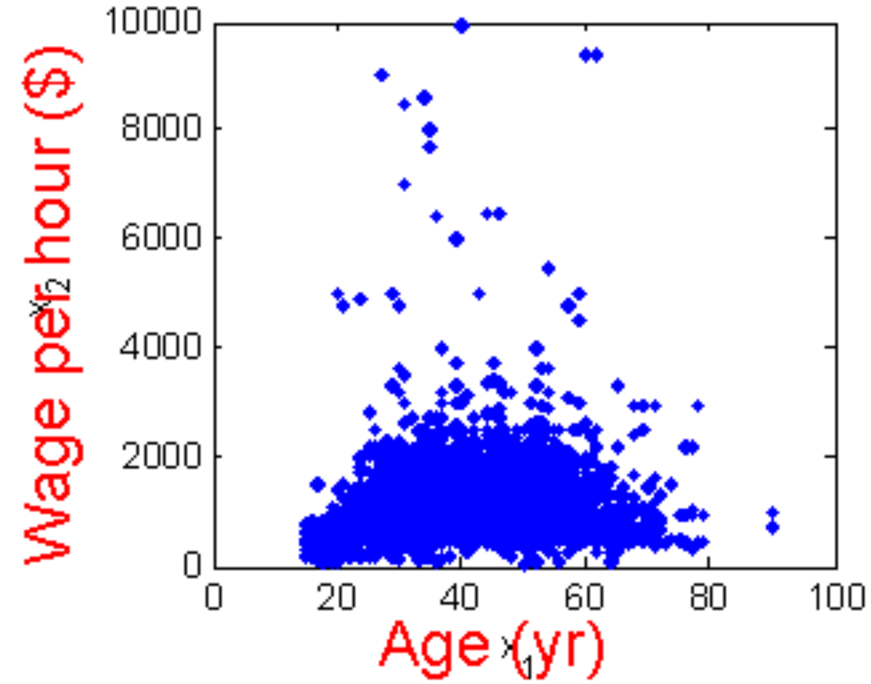
independent

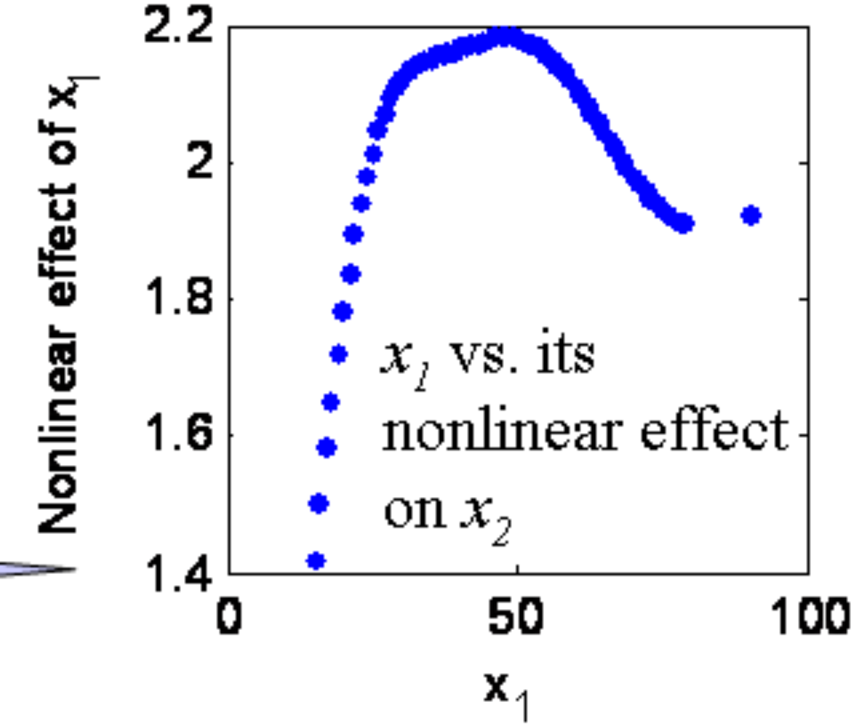




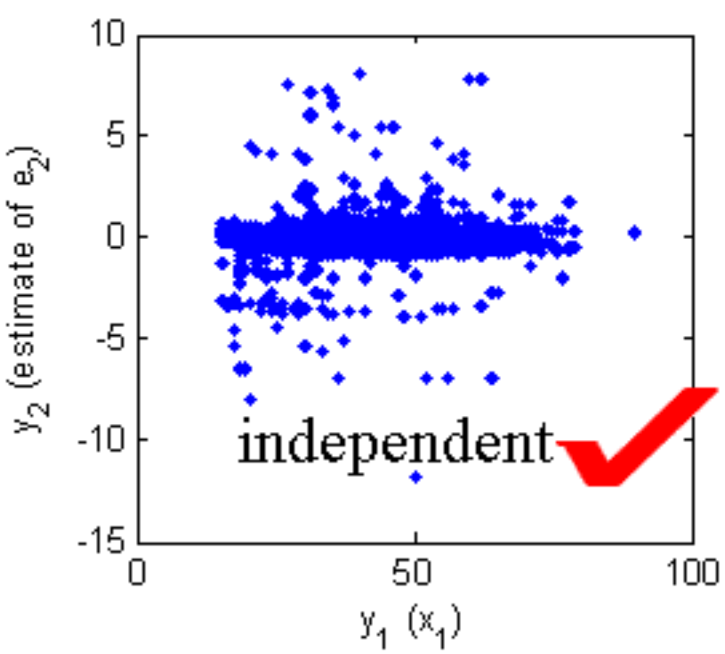


Data Set 8

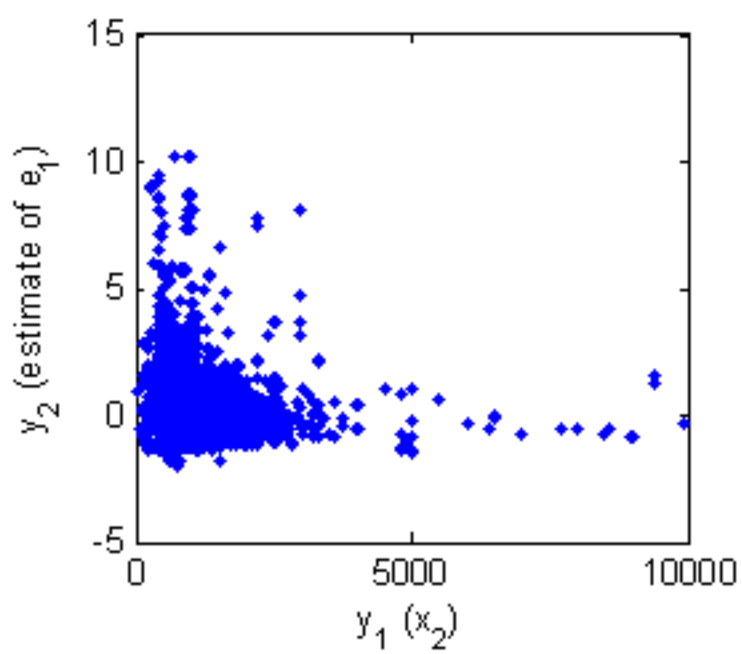


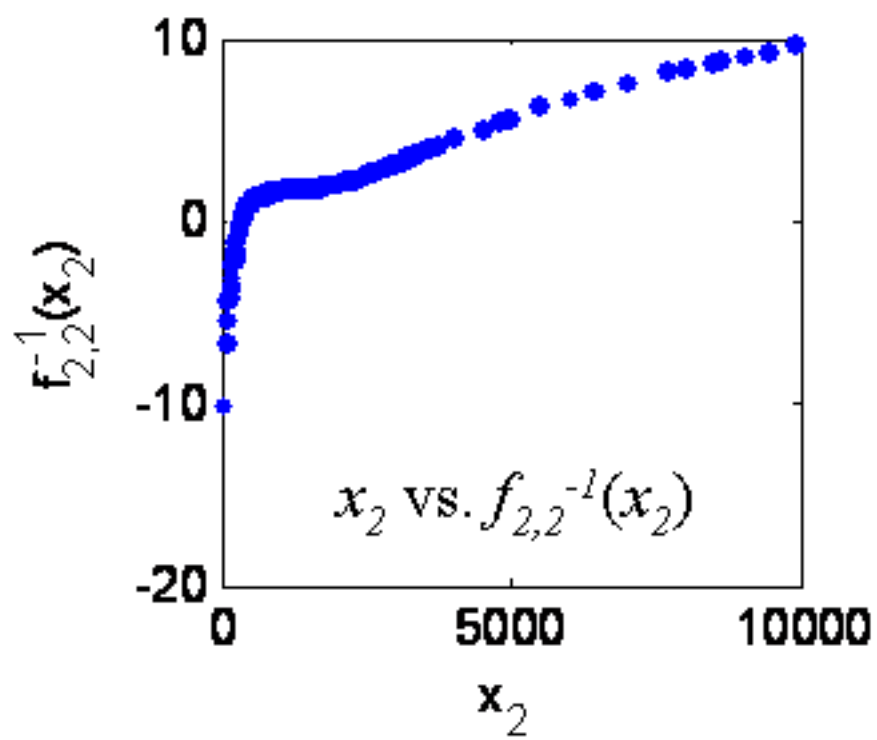


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$



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





Identifiability in Two-variable Case: Theoretical Results

pa_i: parents (causes) of x_i

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

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

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

e_i: noise/disturbance: independent from pa_i

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

Identifiability Establishment: To Be Discussed

- Not Mysterious
- Will explain the basic idea with the identifiability of ICA as an example on March 18

Identifiability: A Mathematical Result

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

- **Theorem 1**
  - Assume $x_2 = f_2(f_1(x_1) + e_2),$
    $x_1 = g_2(g_1(x_2) + e_1),$
  - Further suppose that involved densities and nonlinear functions are third-order differentiable, and that p_{e2} is unbounded,
  - For every point satisfying $\boldsymbol{\eta}_2'' \boldsymbol{h}' \neq 0$, we have

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

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

List of All Non-Identifiable Cases

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

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

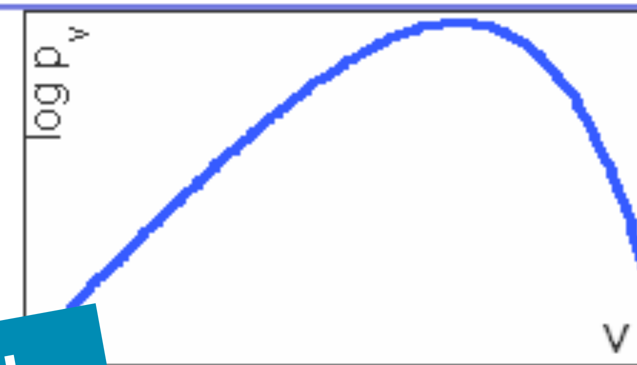


Table 1: All situations i... not identifiable.

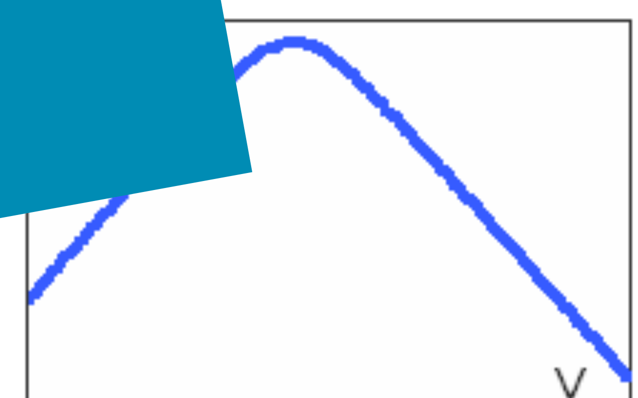
	p_{e_2}	Remark
I	Gaussian	h_1 also linear
II	log-mix-lin-exp	h_1 strictly monotonic, and $h_1' \to$ ...0, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
III	log-mix-lin-exp	—
IV	log-mix-lin-exp	
V	generalized mixture of two exponentials	

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

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

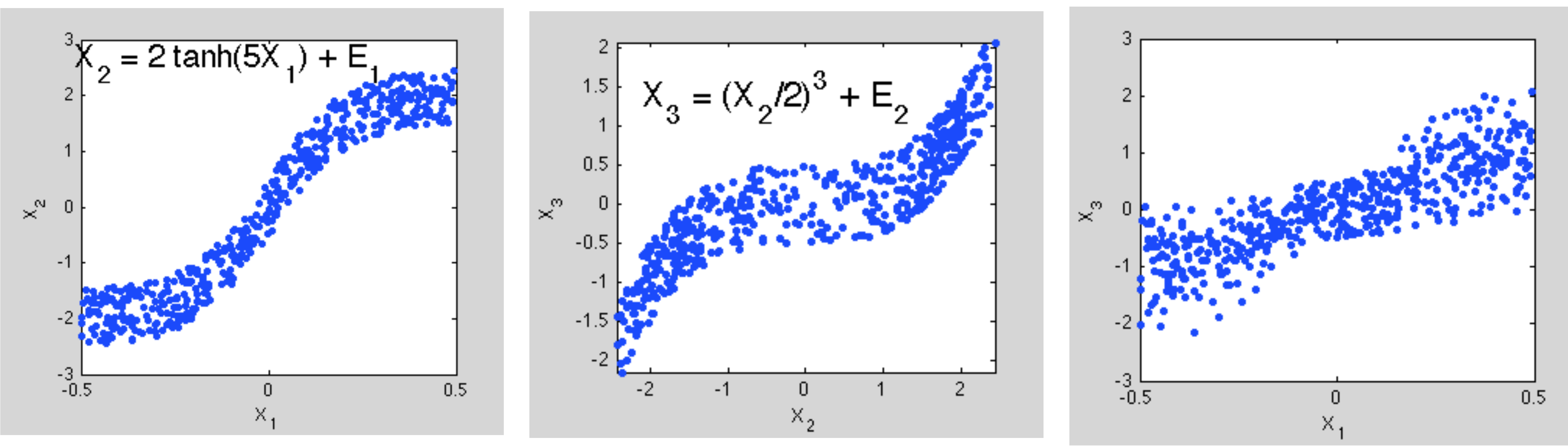
Linear models and nonlinear additive noise models are special cases.

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

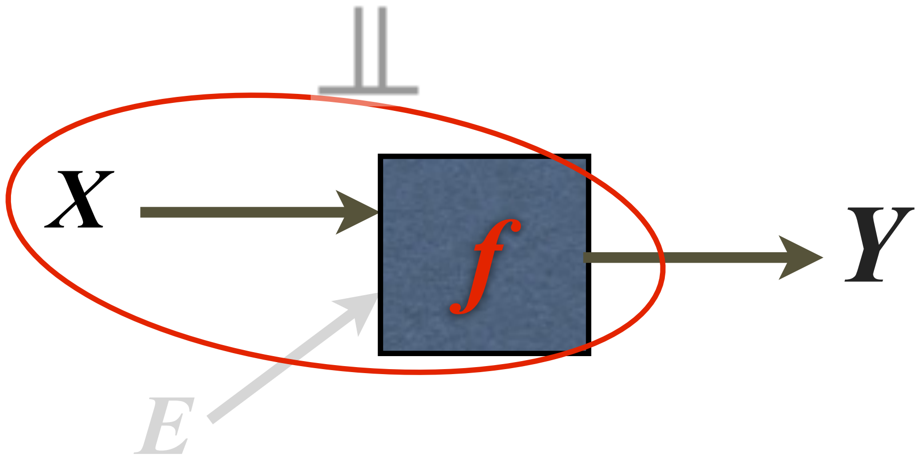


Transitivity of FCMs and Intermediate Causal Variable Recovery

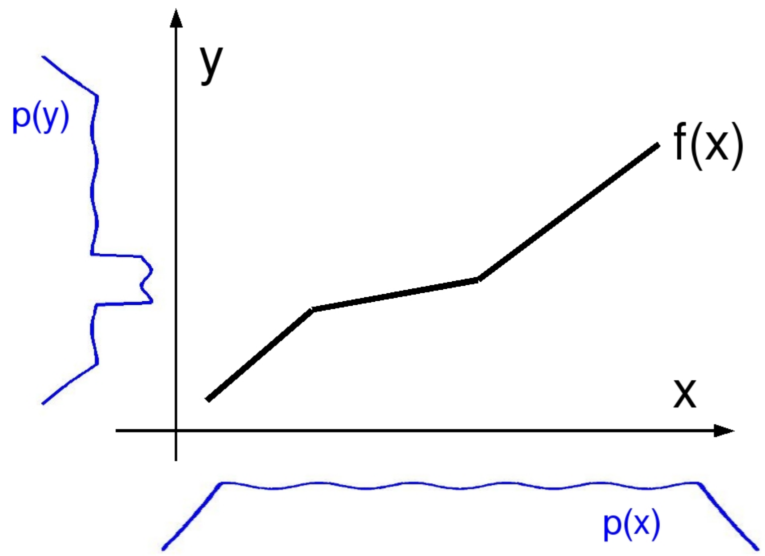
- Transitivity of causal direction violated by FCMs: Intermediate causal variable determination?



Another Type of Method: "Independence" between p(X) and Complex f



- Nonlinear deterministic case:
  - $Y = f(X) \;\Rightarrow\; p(Y) = p(X) / |f'(X)|$
  - $\log f'(X)$ and $p(X)$ uncorrlated w.r.t. a uniform reference; violated for the other direction



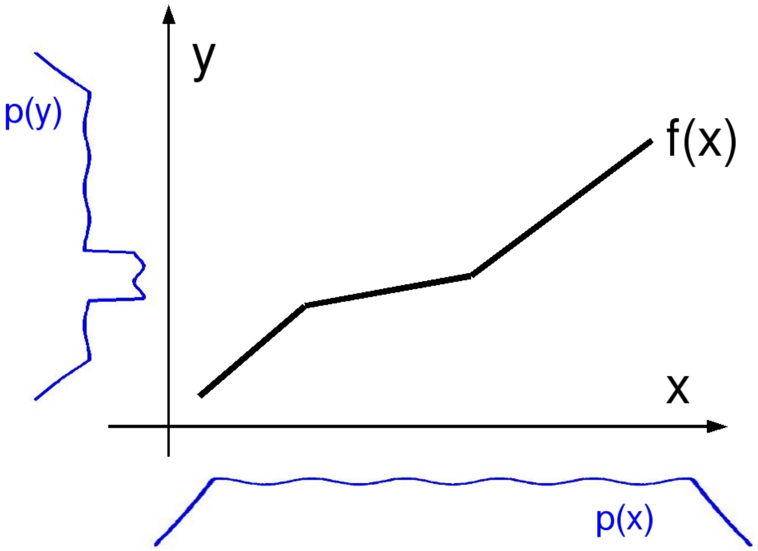
$$\int_0^1 \log f'(x) p(x) dx = \int_0^1 \log f'(x) \frac{p(x)}{p_0(x)} p_0(x) dx$$

$$= \int_0^1 \log f'(x) p_0(x) dx \cdot \int_0^1 p(x) dx = \int_0^1 \log f'(x) dx$$

  - Asymmetry ?

Janzing et al. (2012), Information-geometric approach to inferring causal direction, Artificial Intelligence

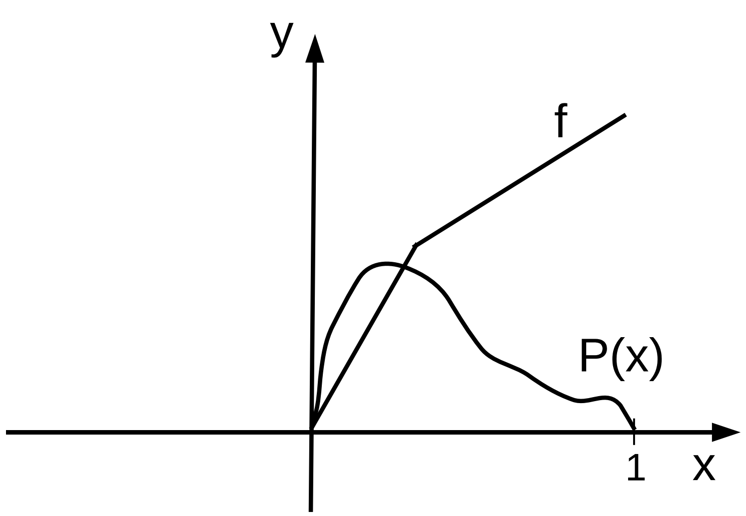
"Independence" between p(X) and Complex f: Asymmetry



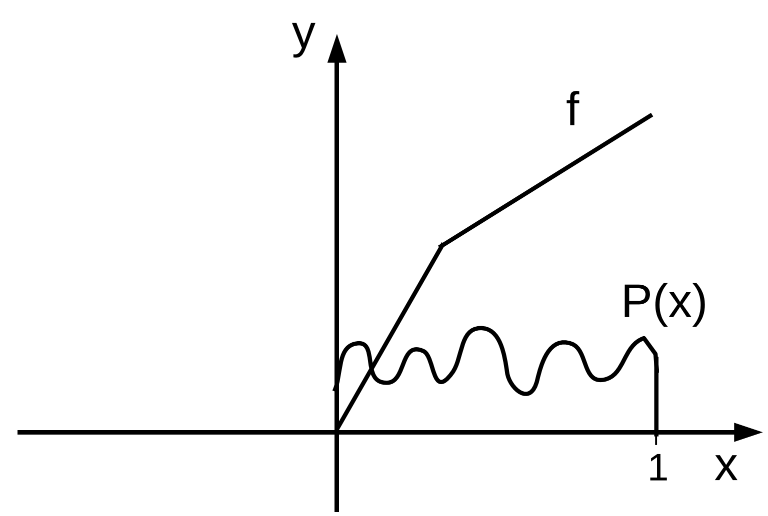
Given $\int_0^1 \log f'(x)p(x)dx = \int_0^1 \log f'(x)dx$, for the other direction

$$\int_0^1 \log(f^{-1})'(y)p(y)dy - \int_0^1 \log(f^{-1})'dy = -\int_0^1 \log f'(x)p(x)dx + \int_0^1 \log(f'(x)f'(x)dx$$

$$= -\int_0^1 \log f'(x)dx + \int_0^1 \log(f'(x)f'(x)dx = \int_0^1 (f'(x)-1)\log f'(x)dx \geq 0$$



Such independence violated



such independence holds

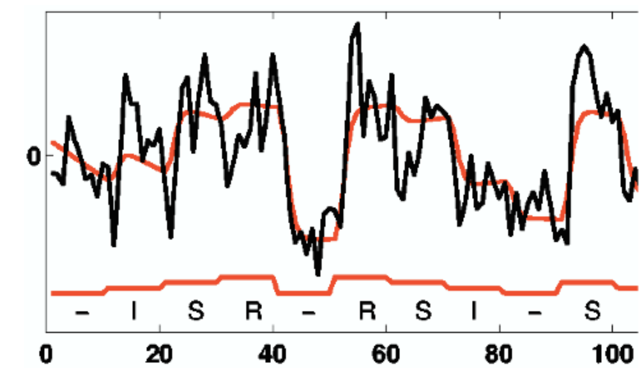
Similarly, "Independence" in Linear Transformations

- Linear high-dimensional deterministic case:
  - $\boldsymbol{Y} = A\boldsymbol{X}$ (causal direction) $\Rightarrow \mathrm{cov}(\boldsymbol{Y}) = A \cdot \mathrm{cov}(\boldsymbol{X}) \cdot A^T$
  - Reverse direction: $\boldsymbol{X} = A^{-1}\boldsymbol{Y}$
  - If A and the covariance matrix of $\boldsymbol{X}$ are chosen independently, then A^{-1} and the covariance matrix of $\boldsymbol{Y}$ will be coupled (in the reverse direction)
  - Asymmetry ?

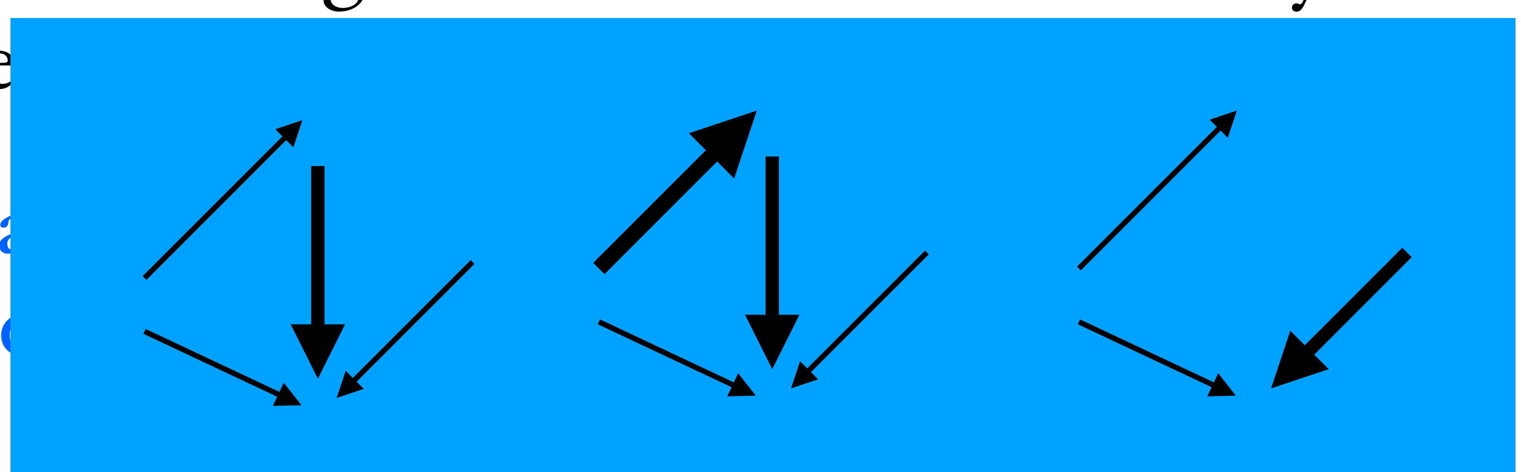
Janzing et al. (2012), Information-geometric approach to inferring causal direction, Artificial Intelligence

Nonstationary/Heterogeneous Data and Causal Modeling

- 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 couple
  - *P*(ca
    inde



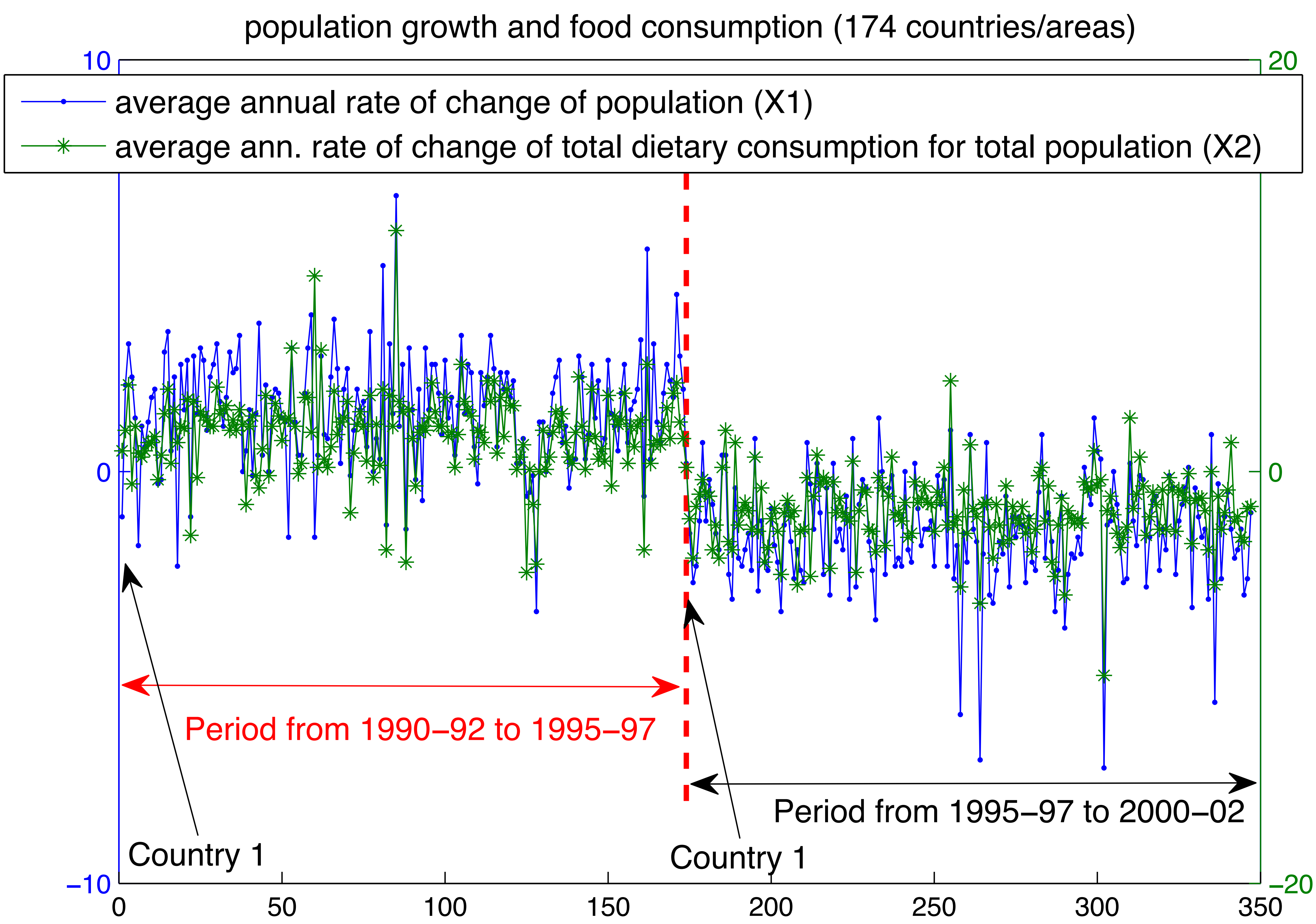
Huang, Zhang, Zhang, Ramsey, Sanchez-Romero, Glymour, Schölkopf, "Causal Discovery from Heterogeneous/ Nonstationary Data," JMLR, 2020

Zhang, Huang, et al., Discovery and visualization of nonstationary causal models, arxiv 2015

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

Causality and Invariance, Robustness, etc.

- Consider prediction (with regression) in different time periods



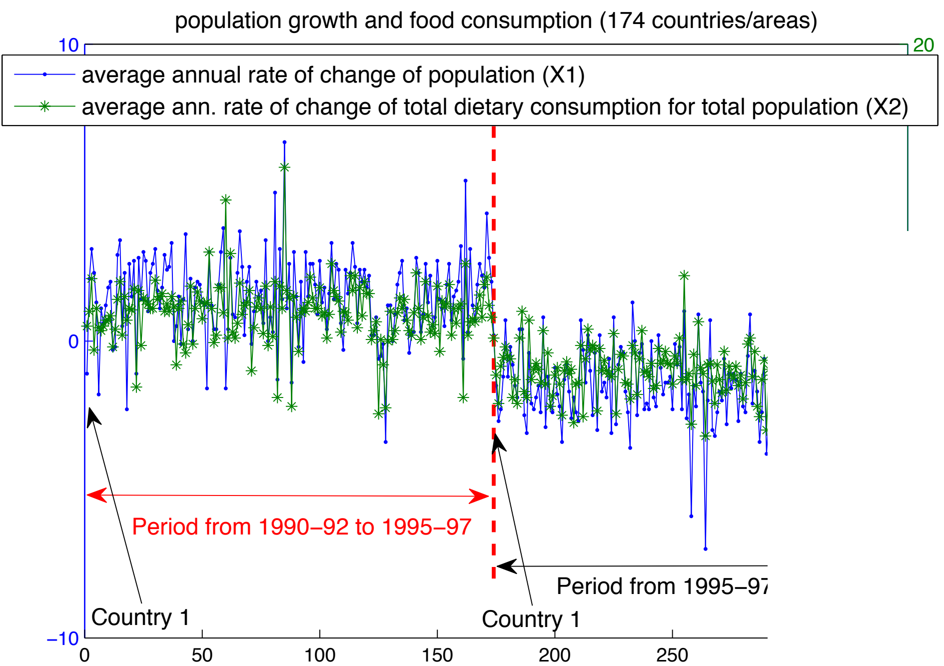
Data

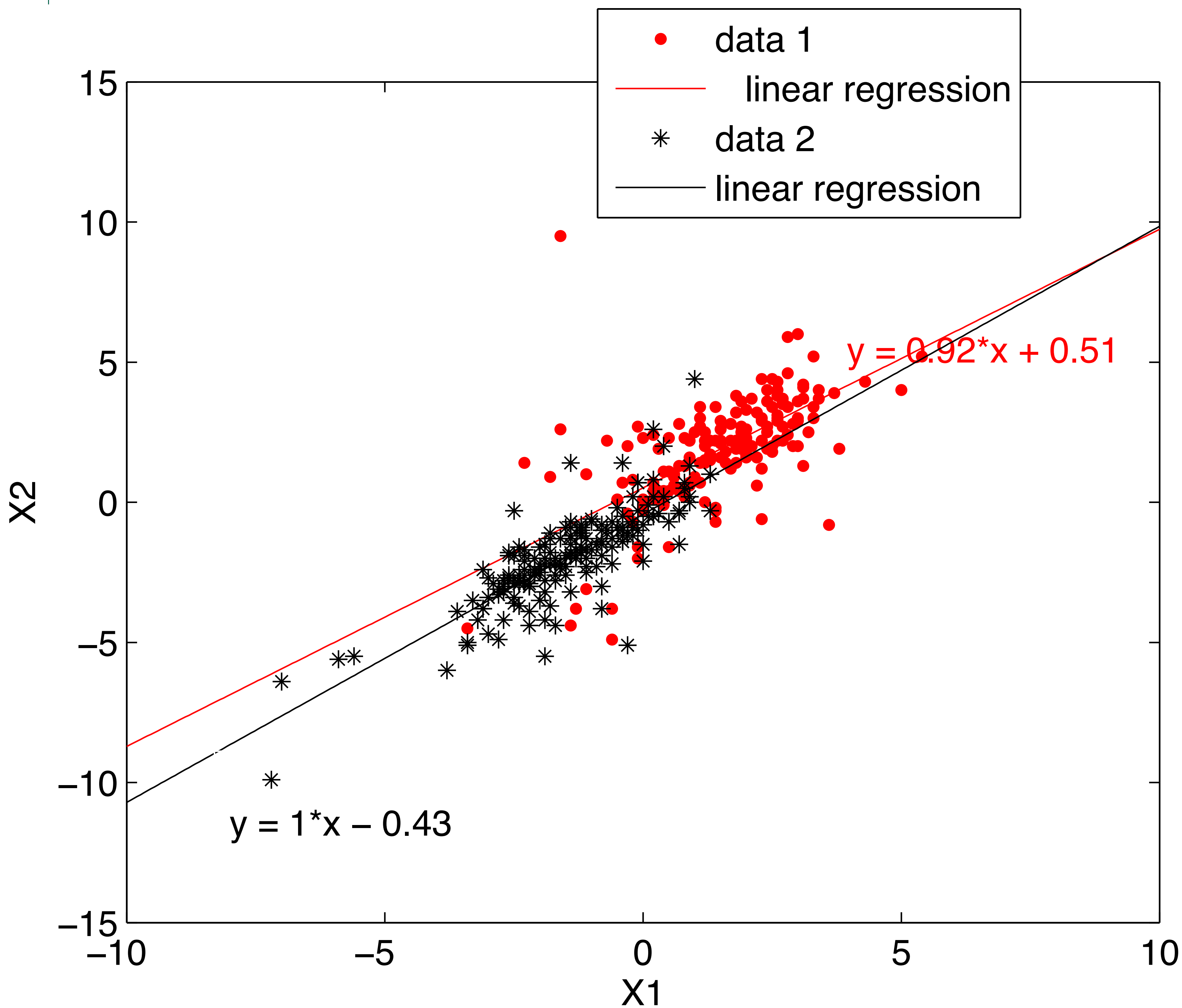
- Population growth and food consumption:
  - data for 174 countries or areas, during the period from 1990-92 to 1995-97 (former 173 data points) and that from 1995-97 to 2000-02 (latter 174 points).
- X1: the average annual rate of change of population; X2: the average annual rate of change of total dietary consumption for total population (kcal/day)

pair0076.txt

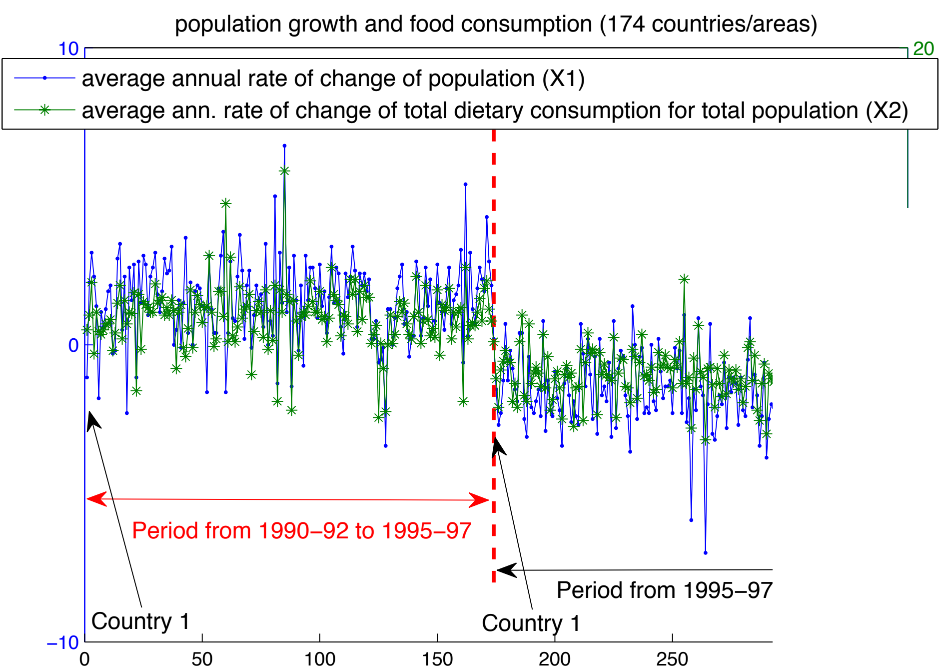
```
-1.1    1.0
2.1     2.0
3.1     4.2
2.3     -0.6
1.3     2.2
-1.8    0.9
1.1     0.7
0.6     1.0
1.2     1.4
1.8     1.4
2.0     1.7
-0.3    -0.4
-0.2    0.8
2.9     2.8
3.4     4.0
0.5     0.4
2.3     3.0
-2.3    1.4
2.6     2.2
1.5     2.1
2.7     3.5
-1.1    -3.1
2.8     3.4
1.4     -0.3
```

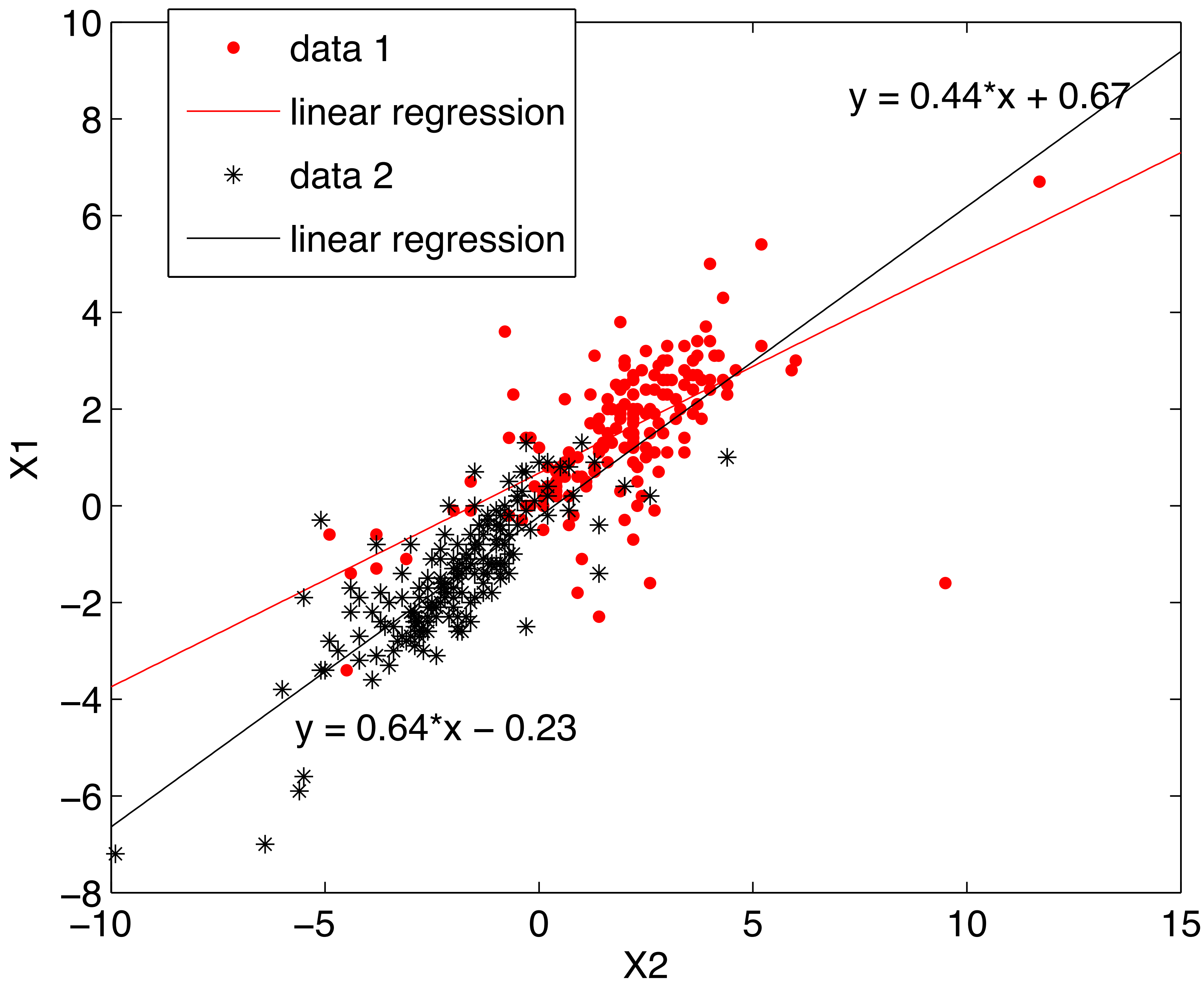
Causality and Invariance, Robustness, etc.



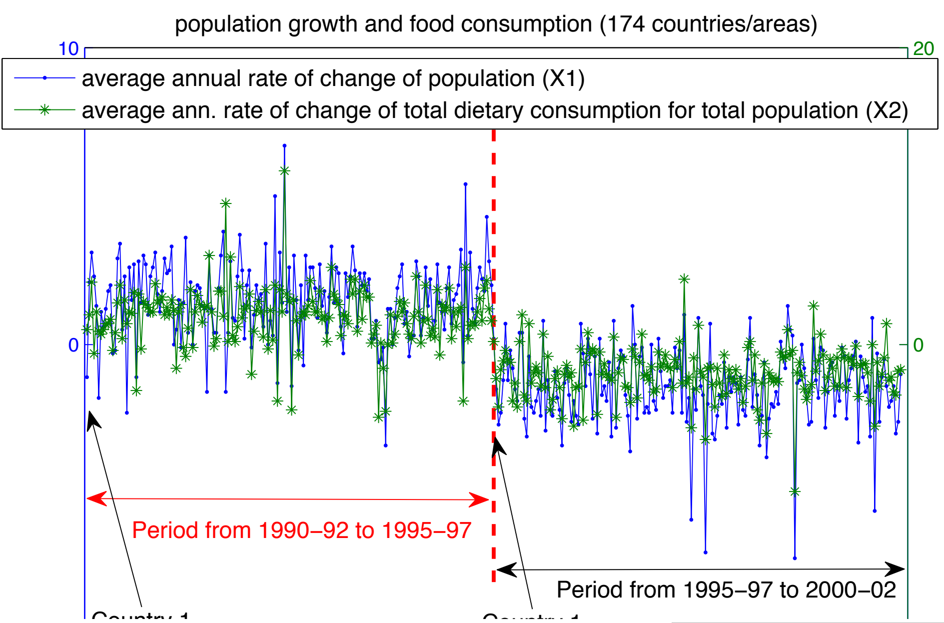


Causality and Invariance, Robustness, etc.

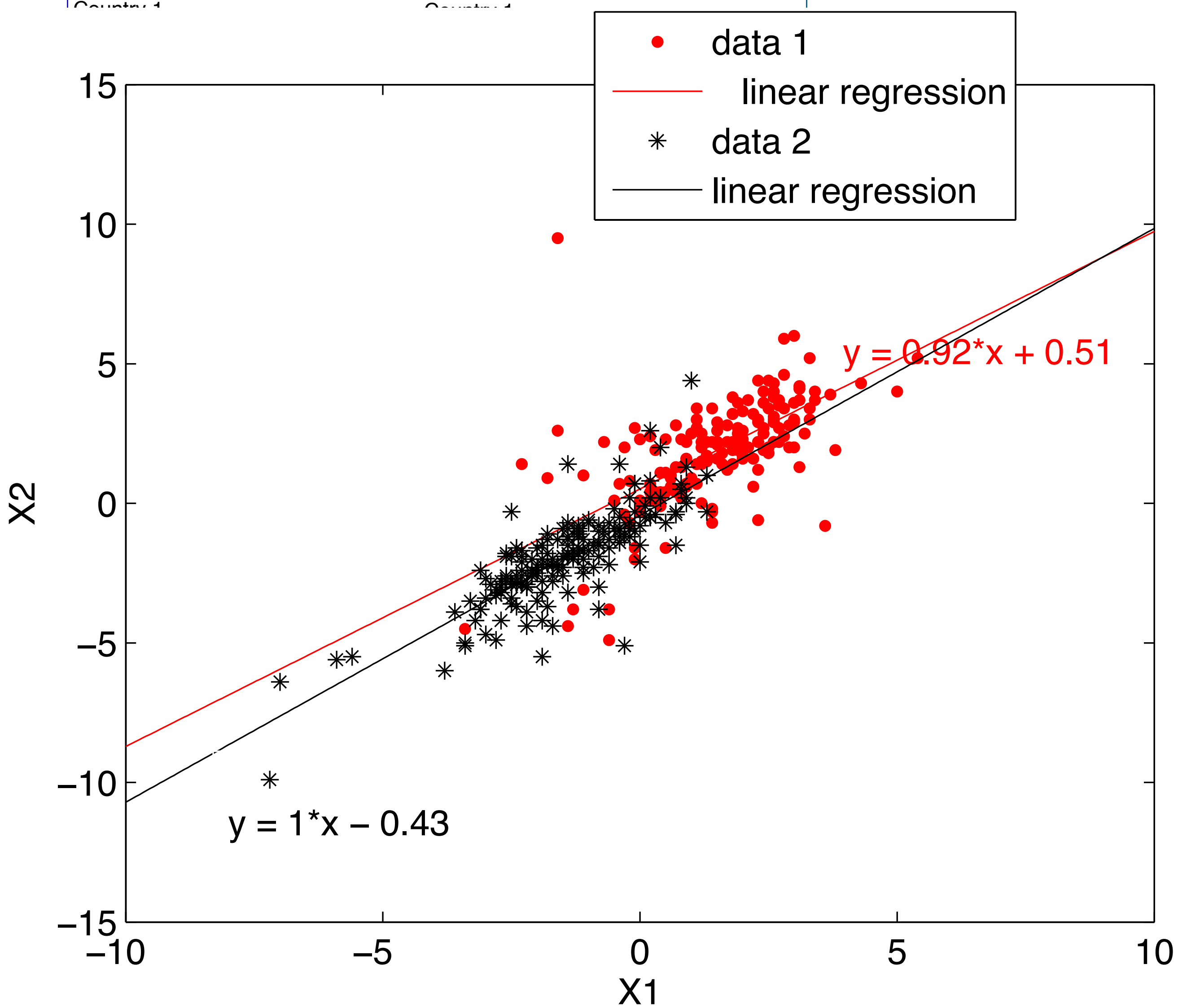


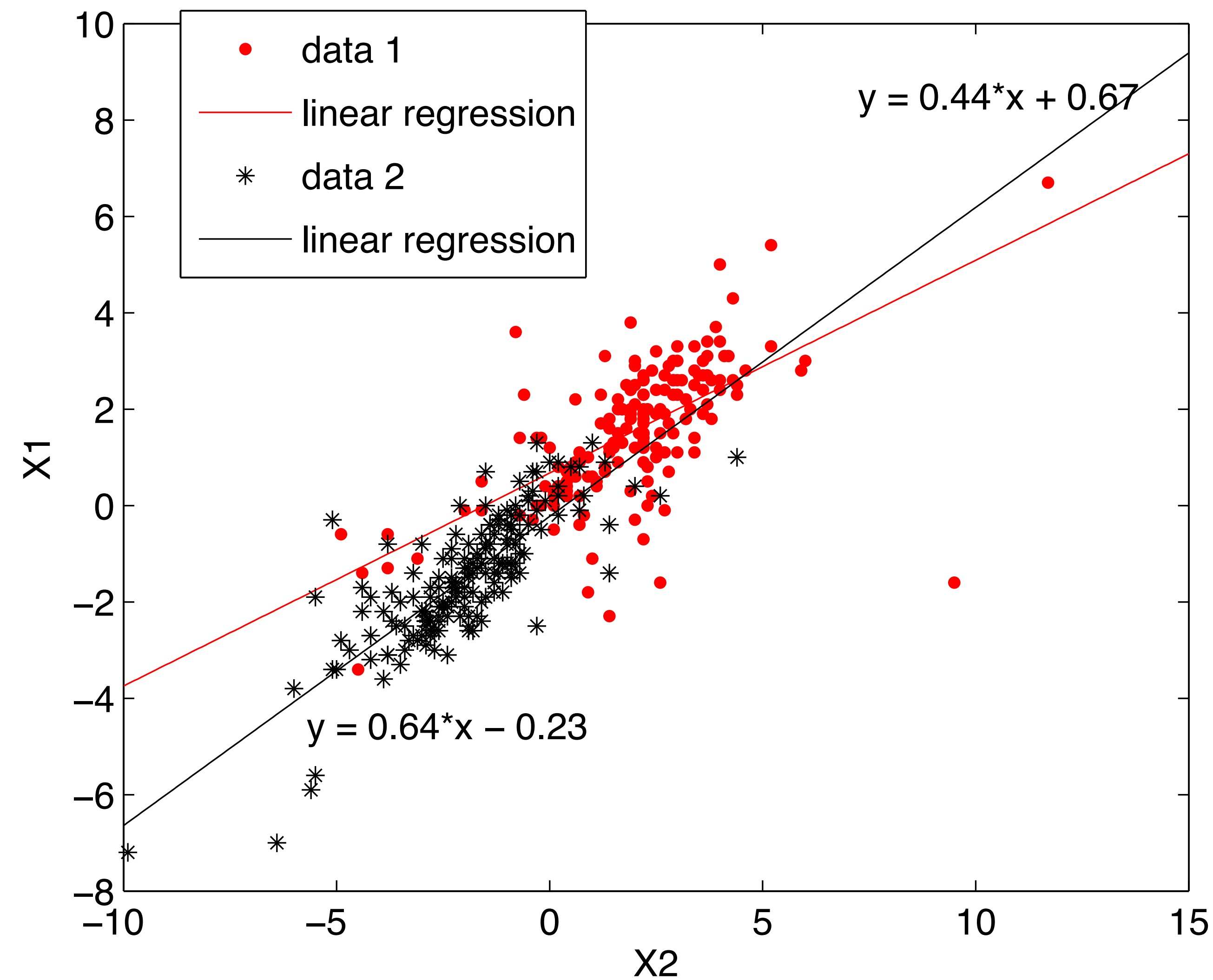


Causality and Invariance, Robustness, etc.



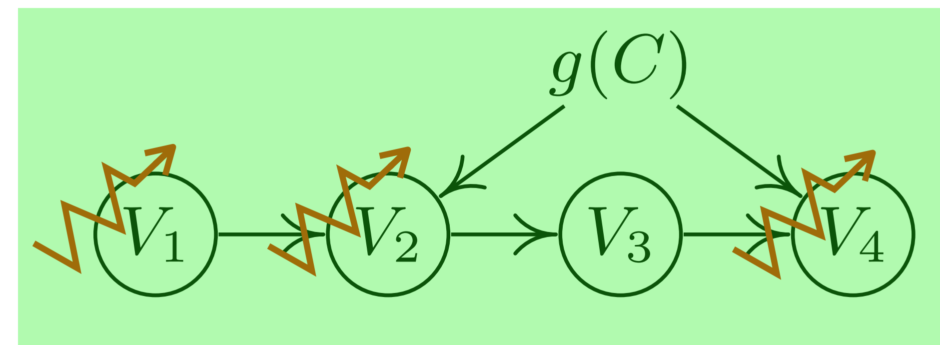
- *Invariance!*
- *More generally, independent changes*



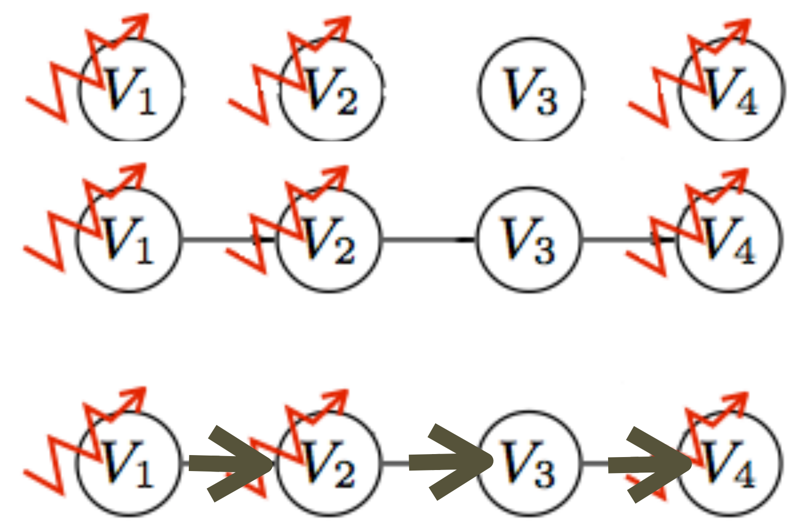


Causal Discovery from Nonstationary/ Heterogeneous Data

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	*Yes*



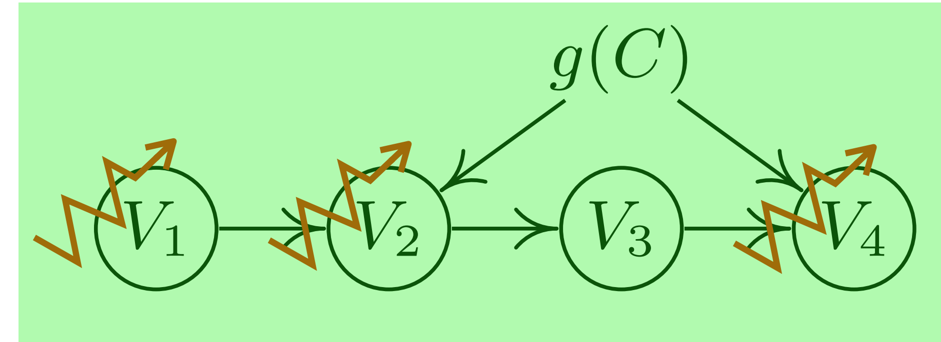
- Task:
  - Determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in P(cause) and P(effect | cause), including invariant mechanism/ cause as special cases
  - Visualization of changing modules over time/ across data sets?



Kernel nonstationary driving force estimation

- *Huang et al., "Causal Discovery from Heterogeneous/Nonstationary Data," JMLR, 2020*
- *Tian, Pearl, "Causal discovery from changes," UAI 2001*
- *Hoover, "The logic of causal inference" Economics and Philosophy, 6:207–234, 1990.*

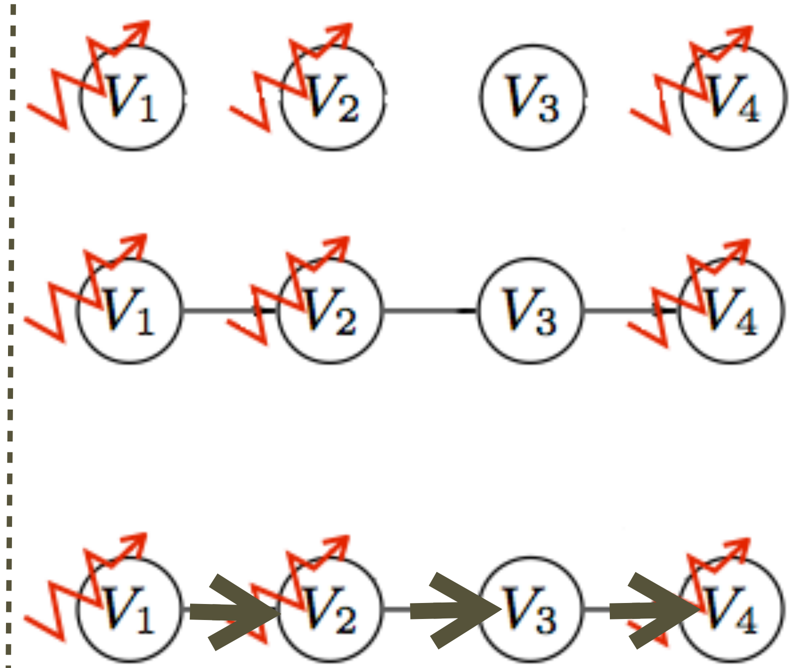
Discovery & Visualization of Changing Causal Modules



* Questions to answer for causal discovery:

With our proposed approach:

- Identify variables with changing causal modules & recover causal skeleton?



- Incorporate time/domain index C as a surrogate + apply constraint-based causal discovery methods

- Identify causal directions by using distribution shifts?

- Independent changes in P(cause) and P(effect | cause)

- Visualize the change in causal modules?

Kernel nonstationarity visualization (KNV)

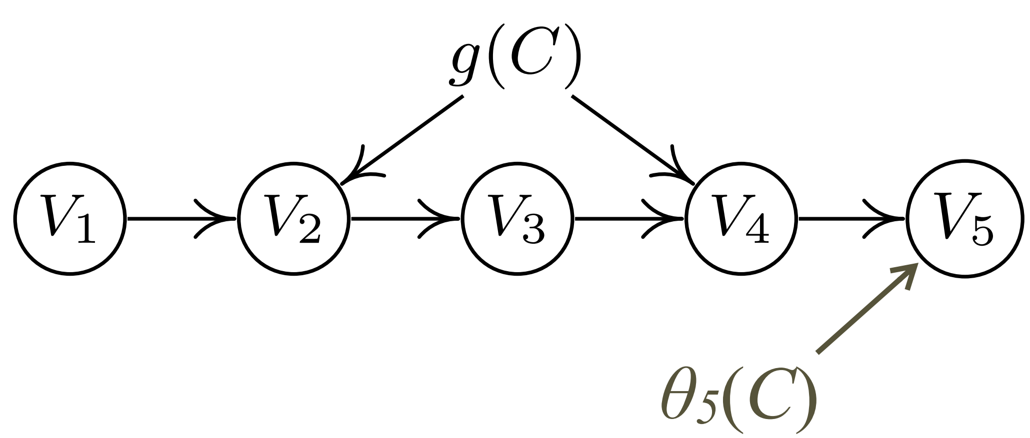
- Find a mapping of $P(V_i \mid PA^i)$ to capture its variability

Method and Its Theoretical Guarantee: Assumptions

- Pseudo causal sufficiency: Confounders as smooth functions of C
  - C: domain or time index; as a surrogate
- Structural equation model:

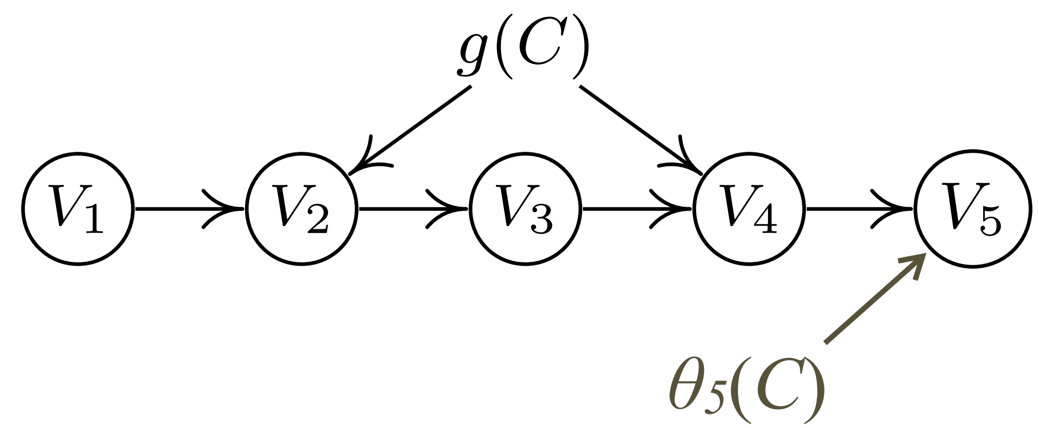
$$V_i = f_i\big(PA^i, \mathbf{g}^i(C), \theta_i(C), \epsilon_i\big)$$

- Causal Markov condition and faithfulness on augmented graph

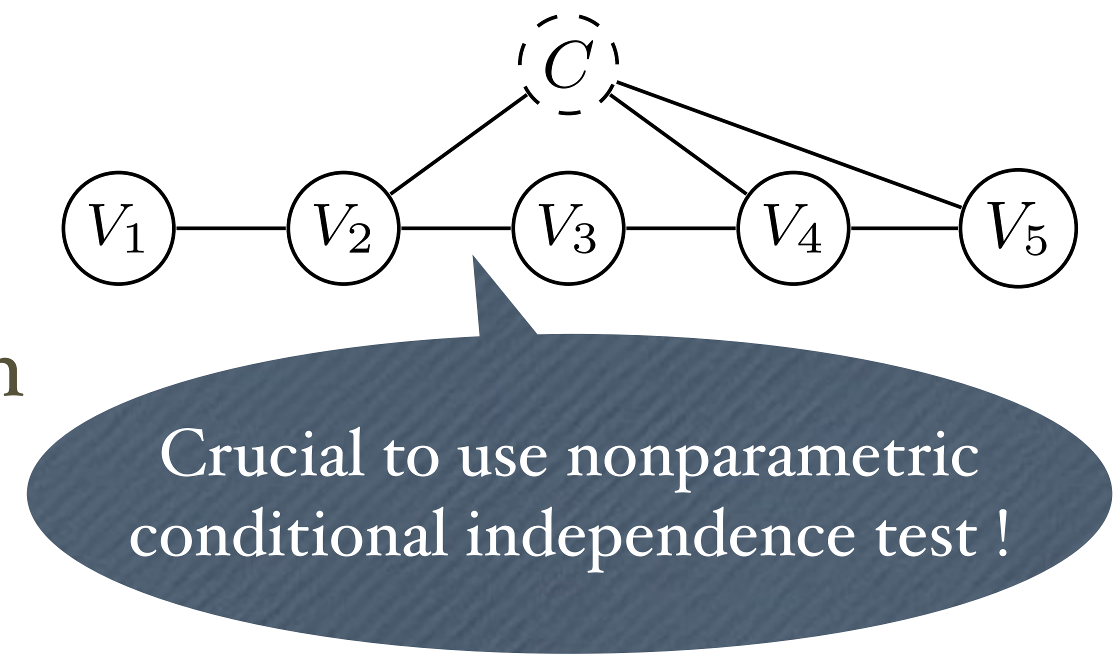


Finding Causal Skeleton and Changing Modules

- Incorporate C into the variable set as a surrogate + apply constraint-based causal discovery



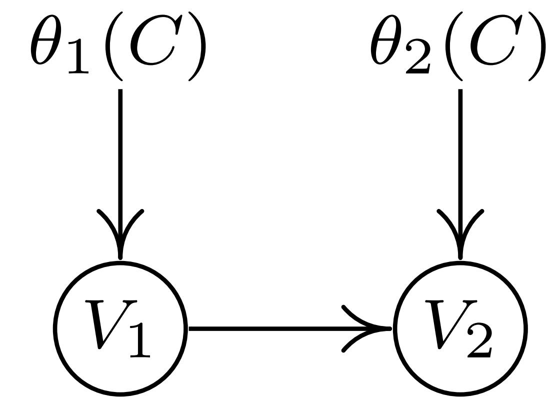
  - Detecting changing causal modules
  - "Robust" causal skeleton discovery



- We can find the correct causal skeleton asymptotically correctly, **as if** the confounders were known

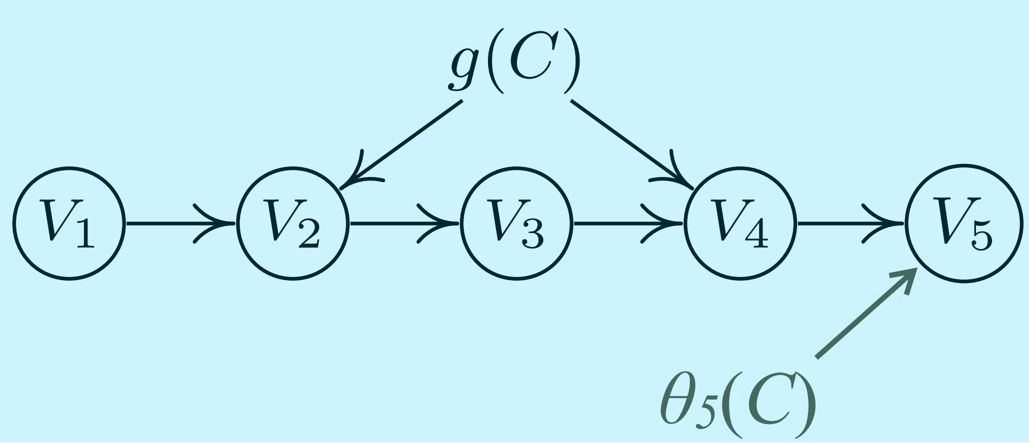
Theorem 1. *Given the previous assumptions, for every $V_i, V_j \in \mathbf{V}$, V_i and V_j are not adjacent in the original causal DAG G if and only if they are independent conditional on some subset of $\{V_k \,|\, k \neq i, k \neq j\} \cup \{C\}$.*

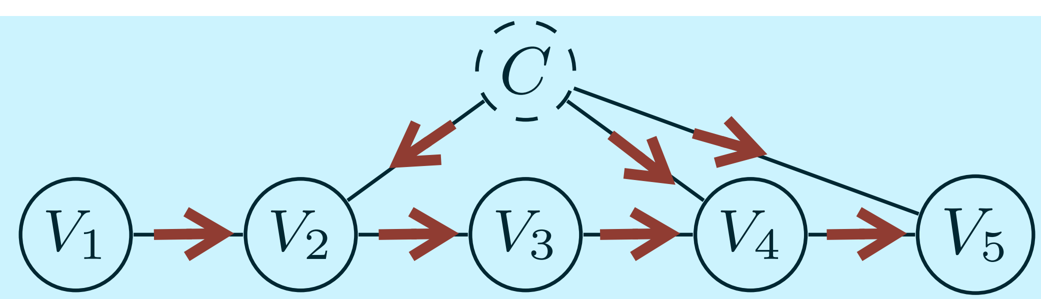
Nonstationarity Helps Determine Causal Direction



- **Independent changes** in P(cause) and P(effect | cause): generalization of invariance; <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.





Kernel Nonstationarity Visualization

$C=1$ $C=2$ $C=3$ …

- Capture the nonstationarity in causal module $PA^i \rightarrow V_i$:

$$\lambda_i(C) = h_i(P(V_i \mid PA^i, C)).$$

  - By maximizing the variability of $\lambda_i(C)$ for all values of C

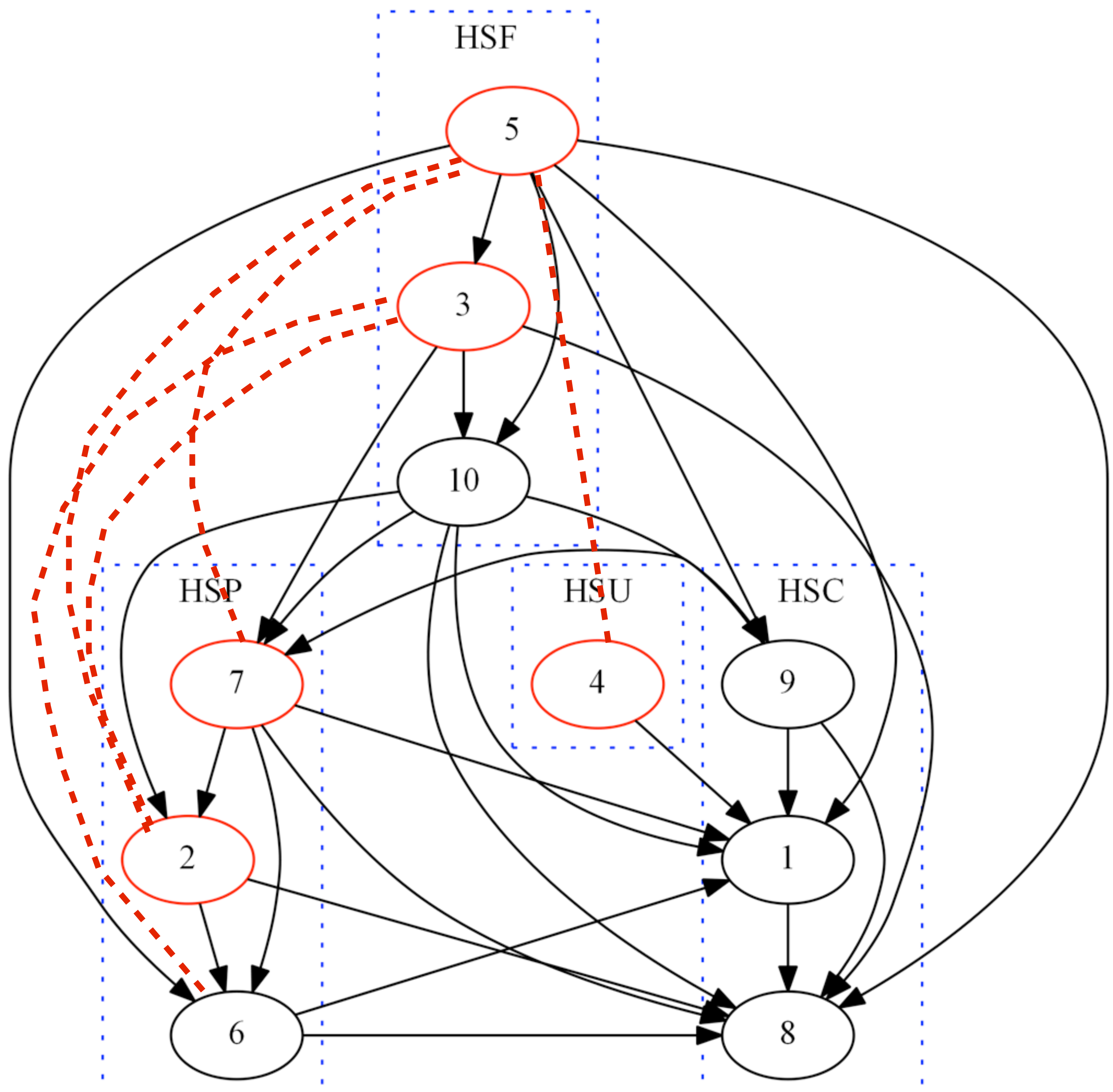
- Kernel nonstationarity visualization (KNV):

  - Kernel embedding of conditional distributions to avoid explicitly estimating them

  - Then borrow the idea of kernel principal component analysis: EVD

Causal Analysis of Major Stocks in Hong Kong Market (10/09/2006 - 08/09/2010)

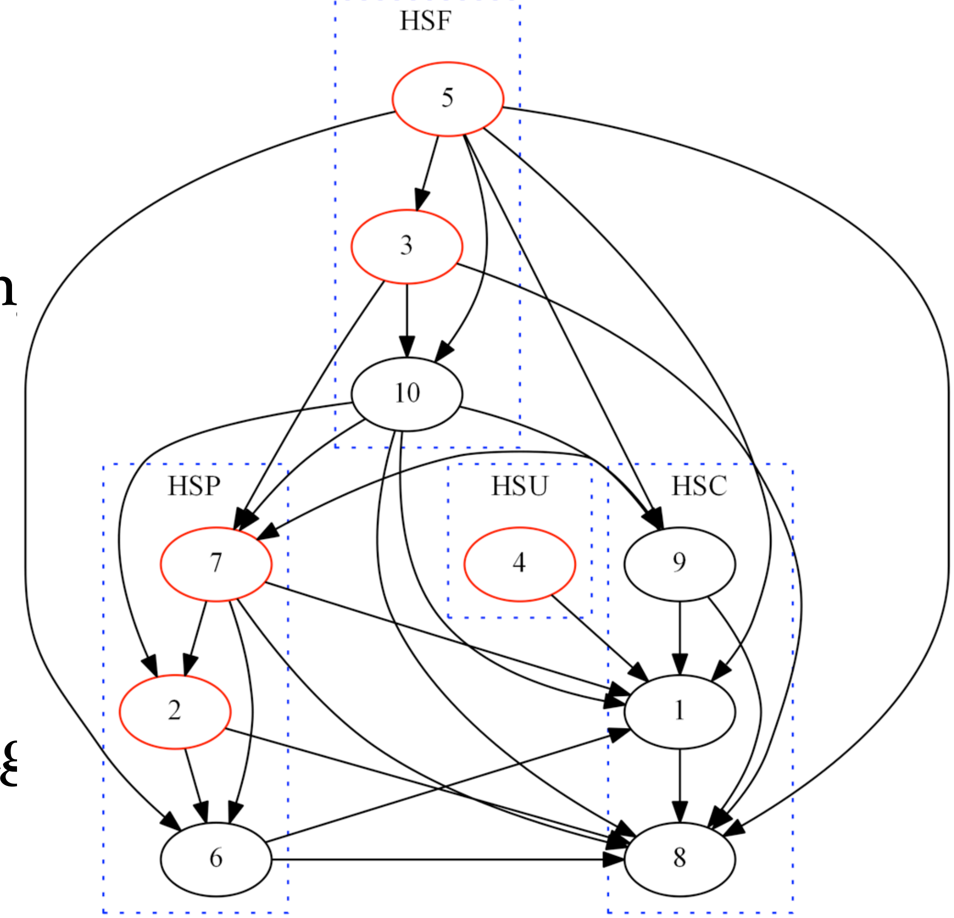
1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4. Hong Kong Electric Holdings,
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong

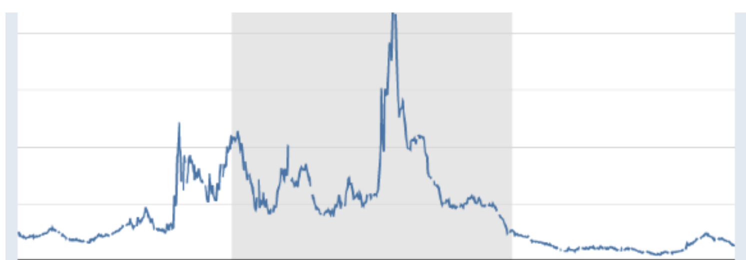


- HSF and HSP usually have nonstationary confounders

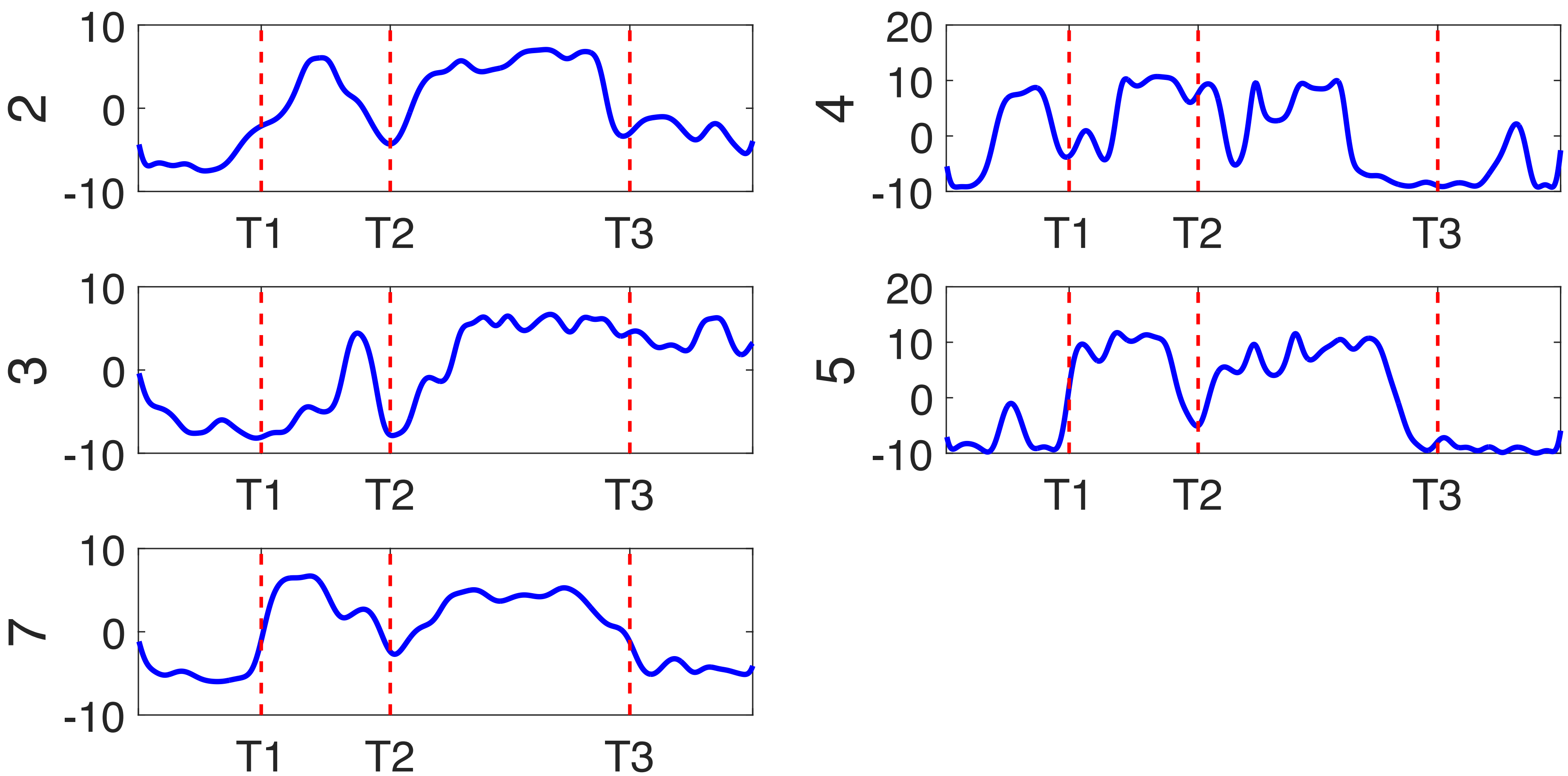
Nonstationarity Driving Force

1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4. Hong Kong Electric Holdin
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong



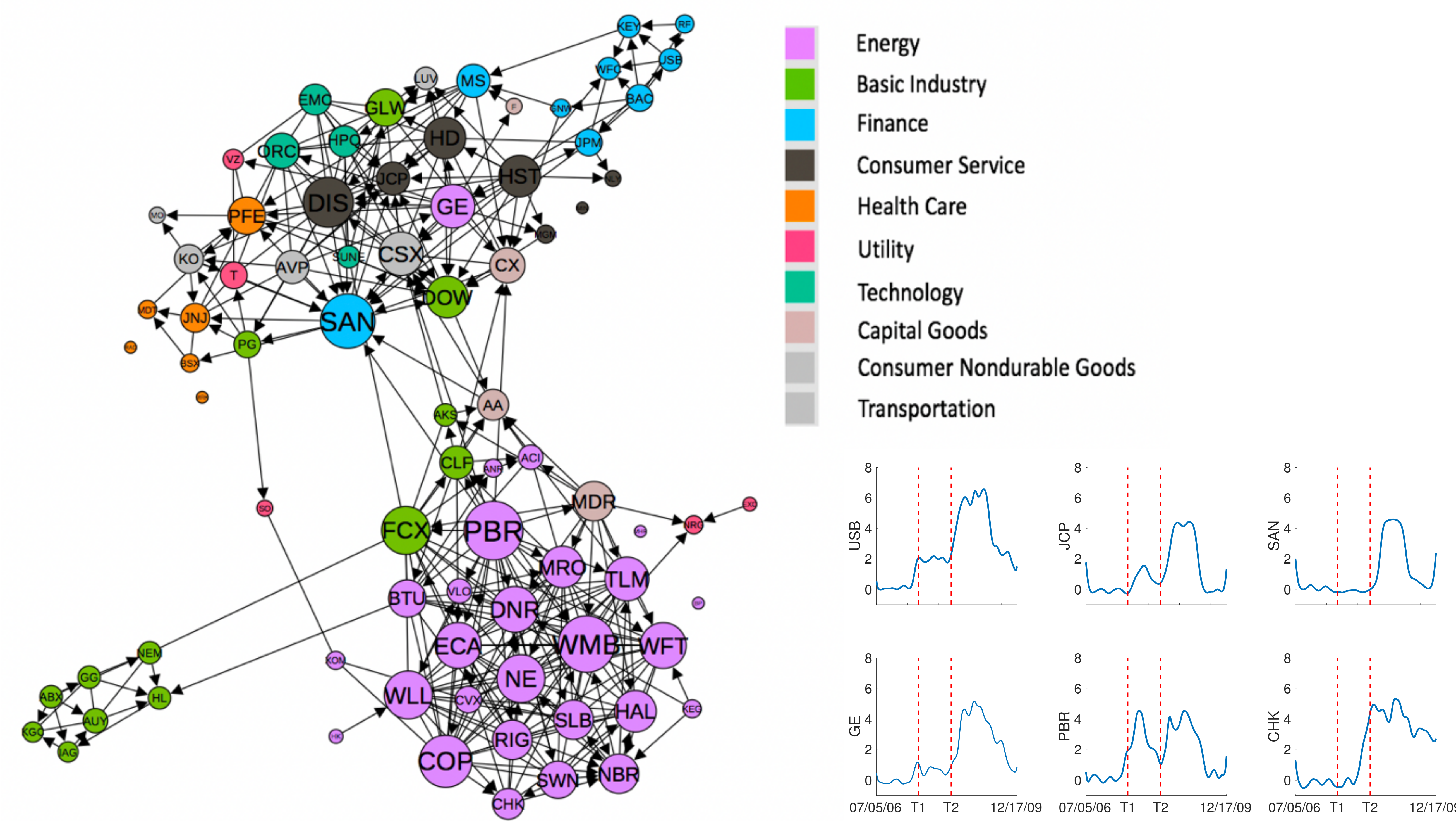


(Curve of TED spread; https://research.stlouisfed.org/fred2/series/TEDRATE)



T_1: 07/16/2007,
T_2: 06/30/2008,
T_3: 02/11/2009

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

Summary

- Nonlinear models with additive noise
    - Just like linear, non-Gaussian models
    - So some people say nonlinear or non-Gaussian methods for causal discovery can recover the DAG uniquely
- Other types of "independence" also help in causal discovery
- Nonstationarity facilitates causal discovery

- Next: Dealing with selection bias, measurement error, missing values, temporal constraints, etc.