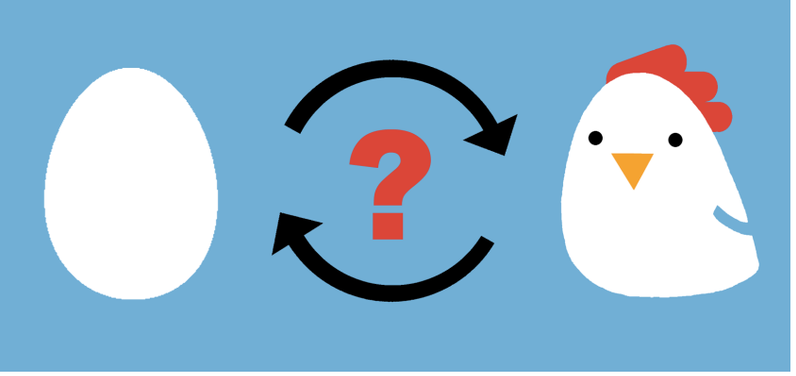


CBMS Conference -- Foundations of Causal Graphical Models and Structure Discovery

Lecture 8

Practical issues: Nonlinear Relations, Measurement Error, Selection Bias, Missing data, and Nonstionarity

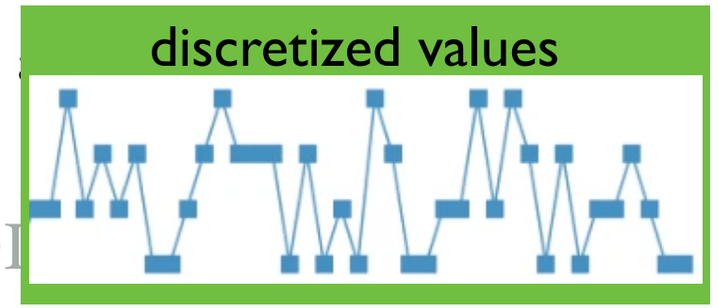
Instructor: Kun Zhang

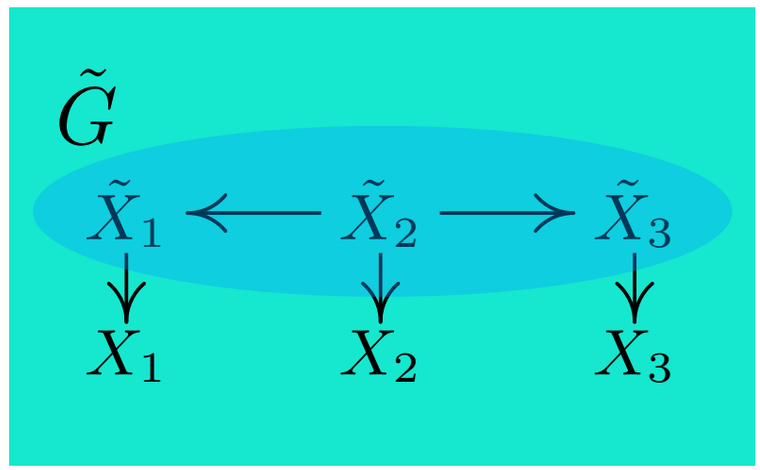
Carnegie Mellon University

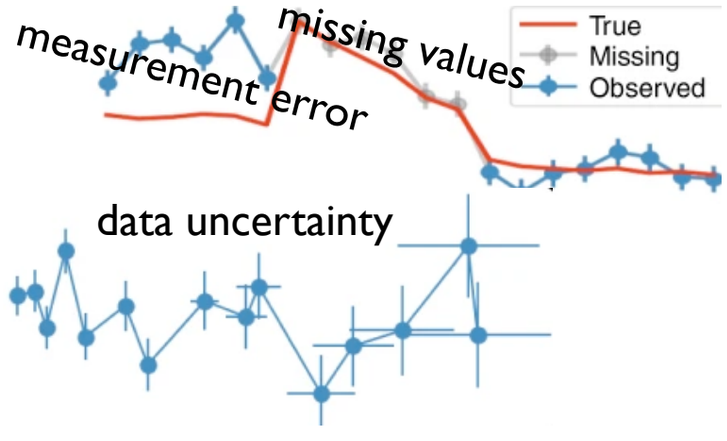
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Practical Issues in Causal Discovery...

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

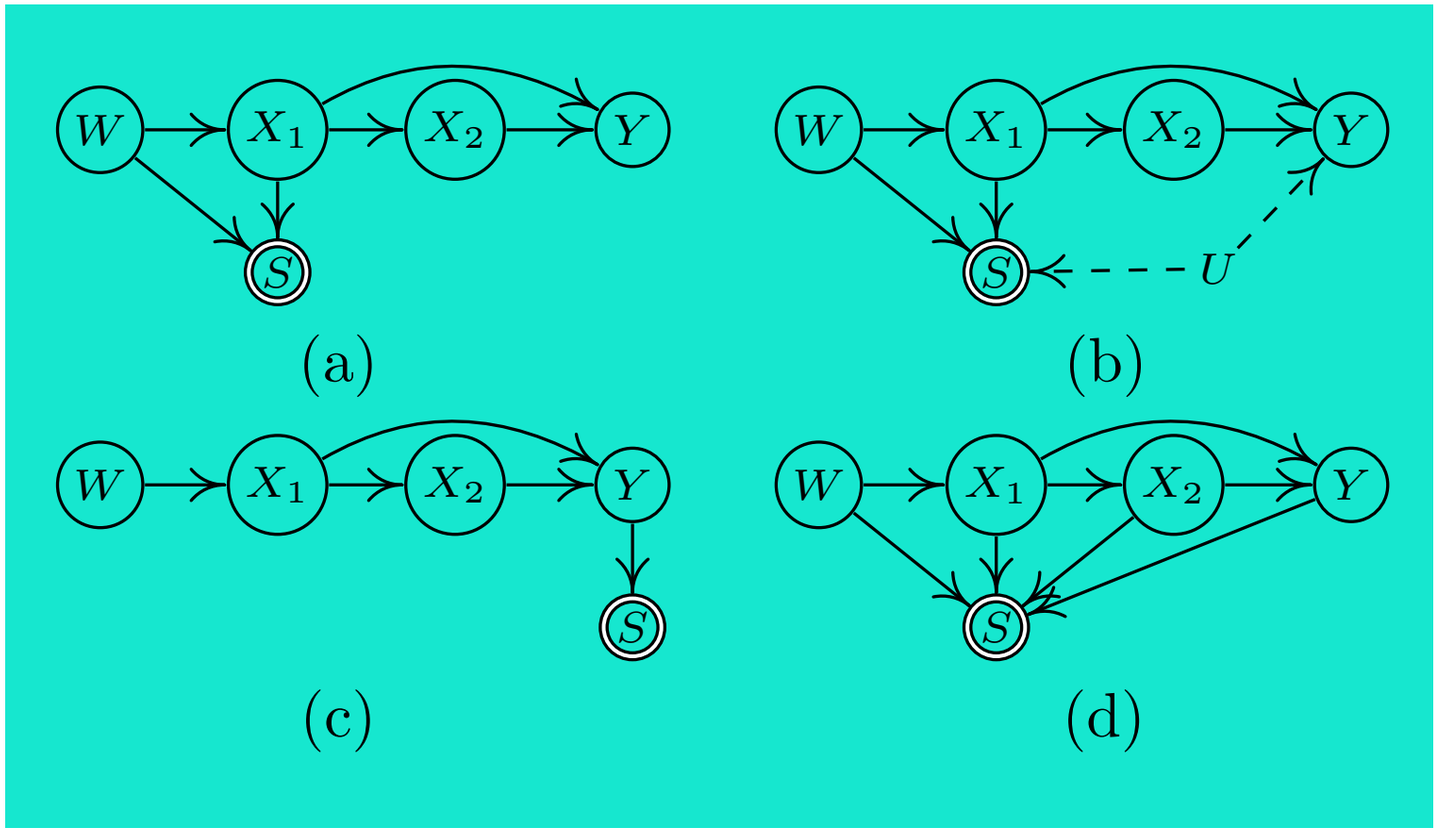






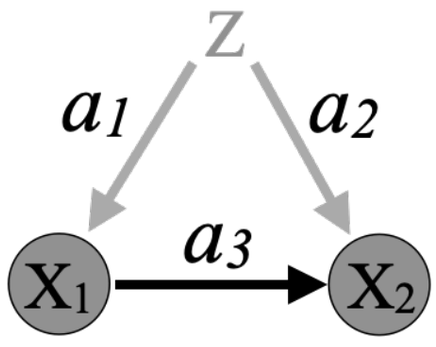
Practical Issues in Causal Discovery...

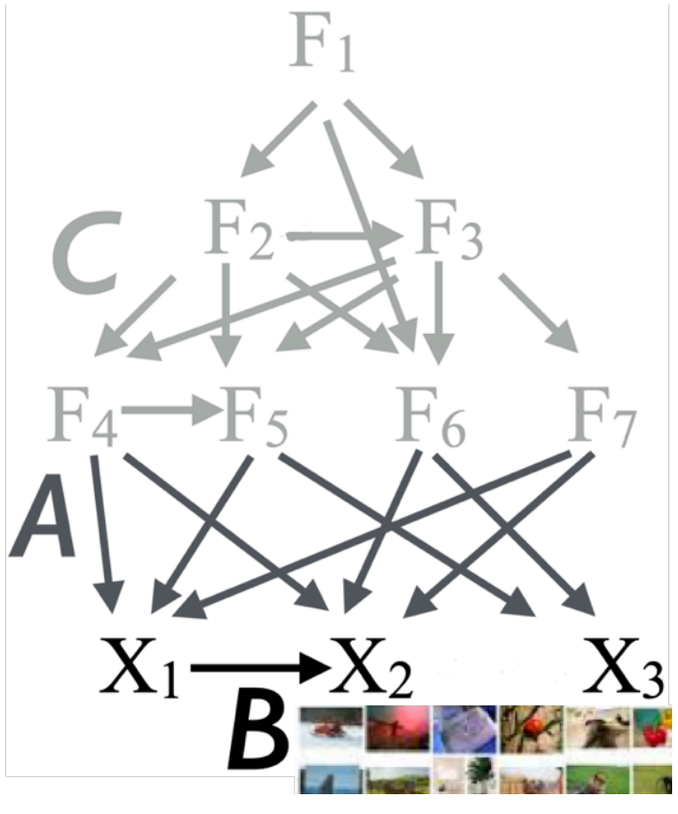
- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- **Selection bias** (Zhang et al., UAI'16)



Practical Issues in Causal Discovery...

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





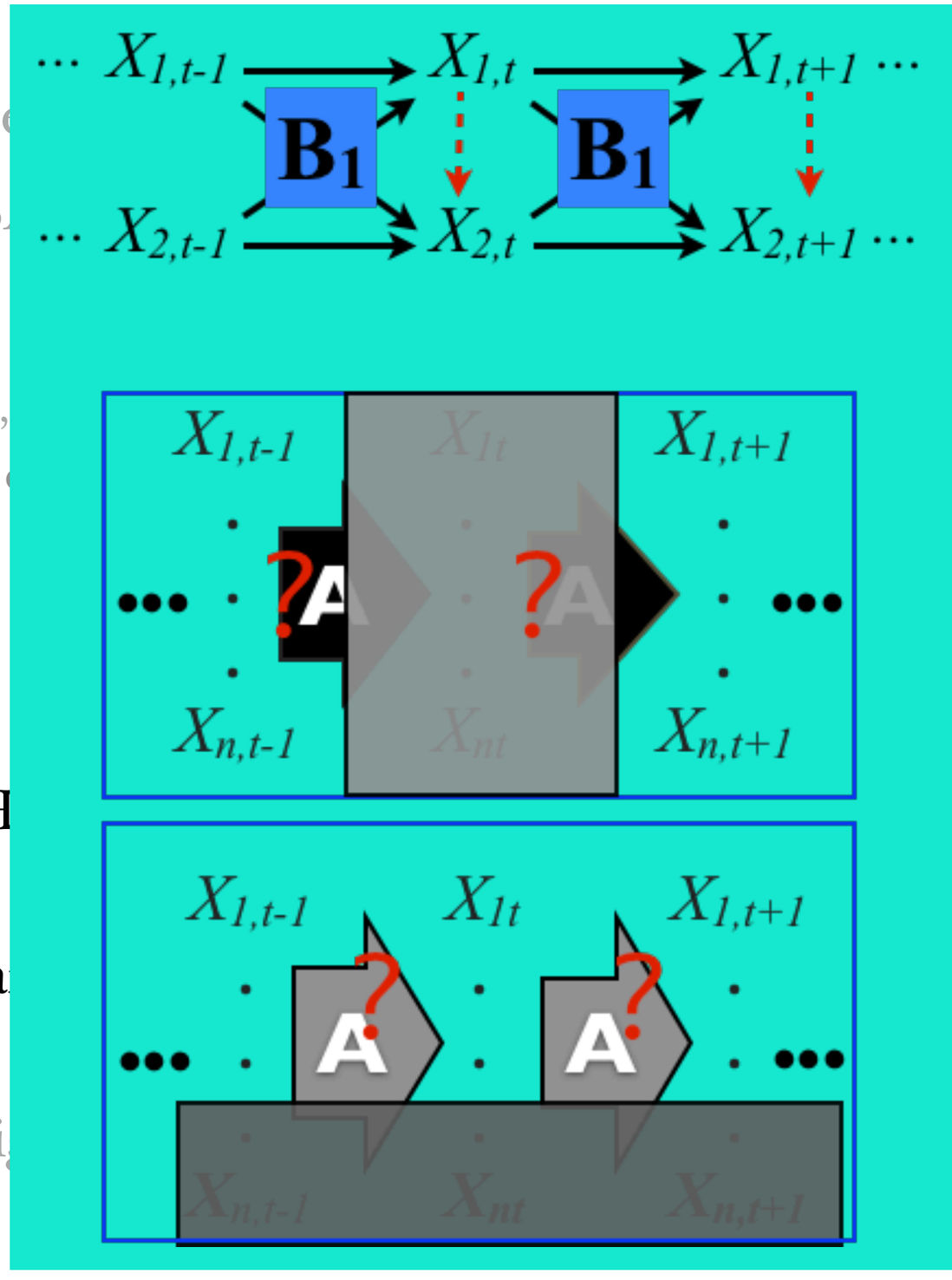
Practical Issues in Causal Discovery...

- 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
-1.1349917e+00	3.5184872e-01	-5.6961669e-01	-4.8335698e-01	9.3747444e-01	2.2763032e-03

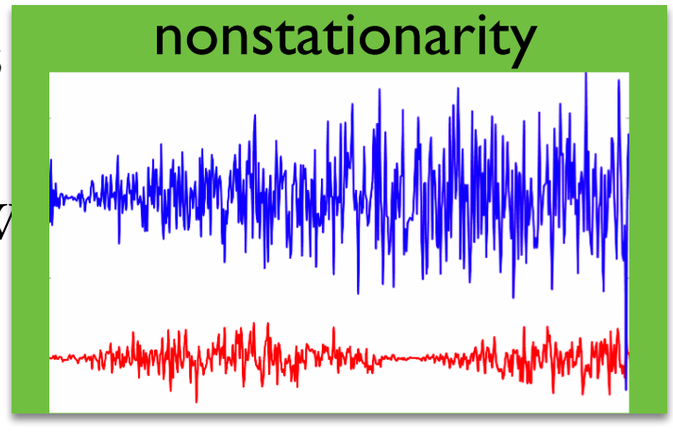
Practical Issues in Causal Discovery...

- 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)

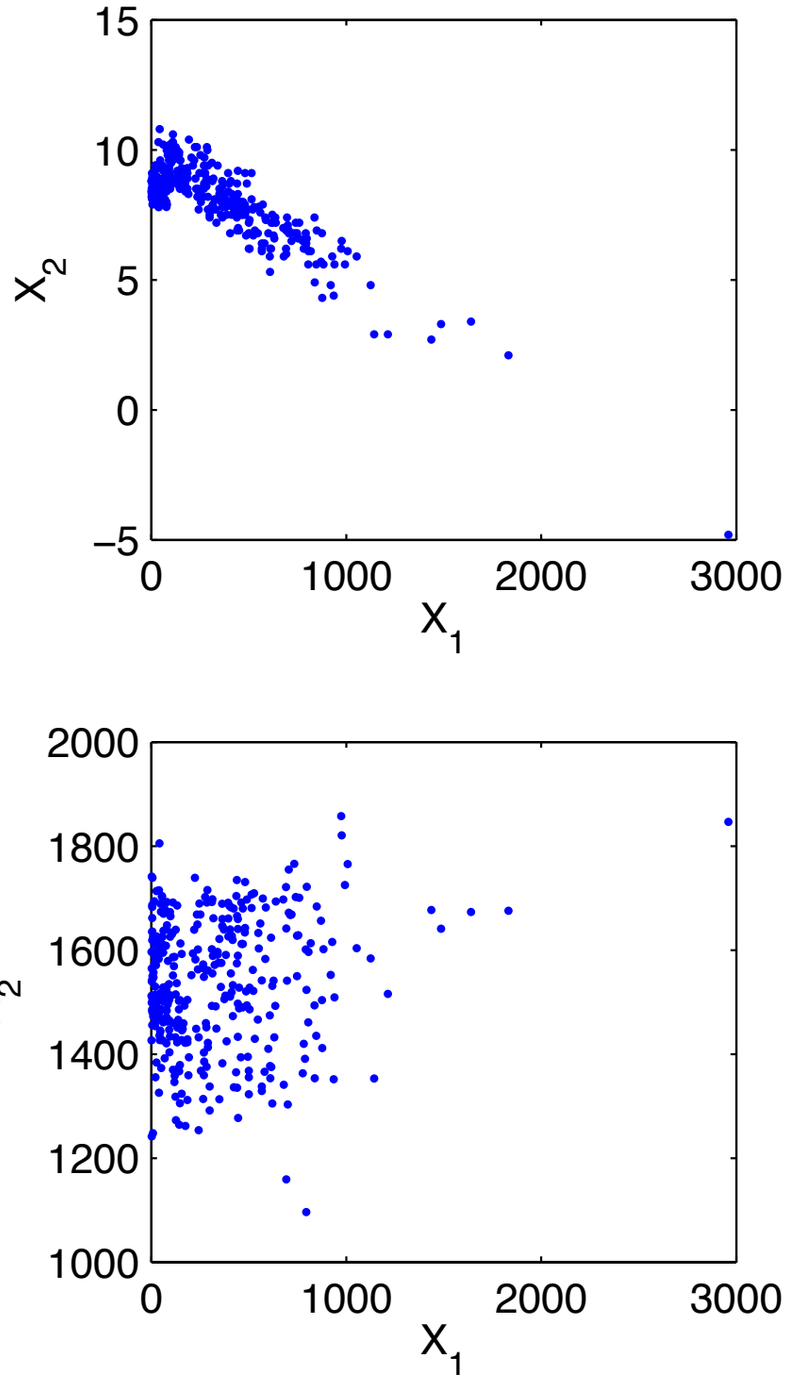


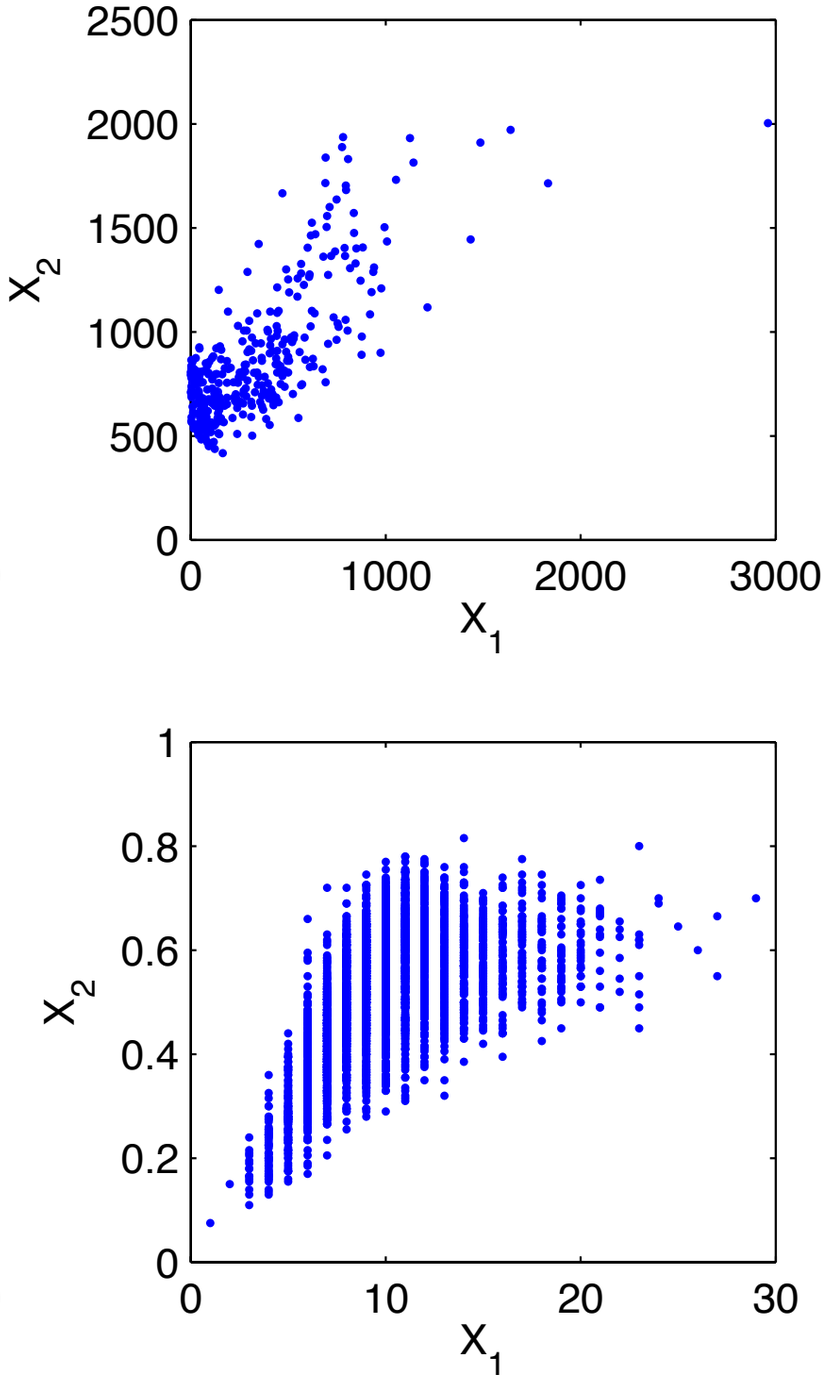
- **Nonstationary/heterogeneous data** (Zhang et al., IJCAI'17; Huang et al, ICDM'17, Ghassami et al., NIPS'18; Huang et al., ICML'19 & NIPS'19; Huang et al., JMLR'20)

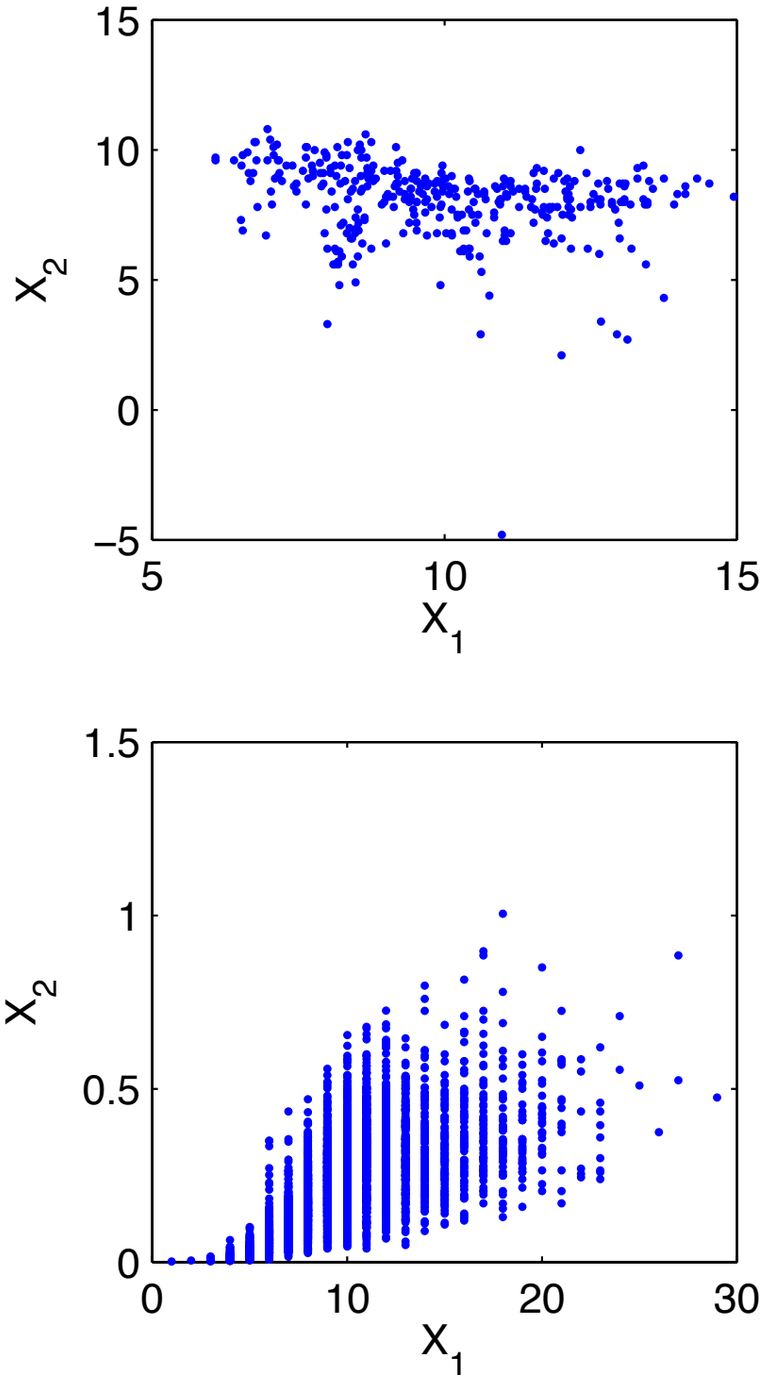
With Nonlinearities

- Model
- Identifiability
- Identification

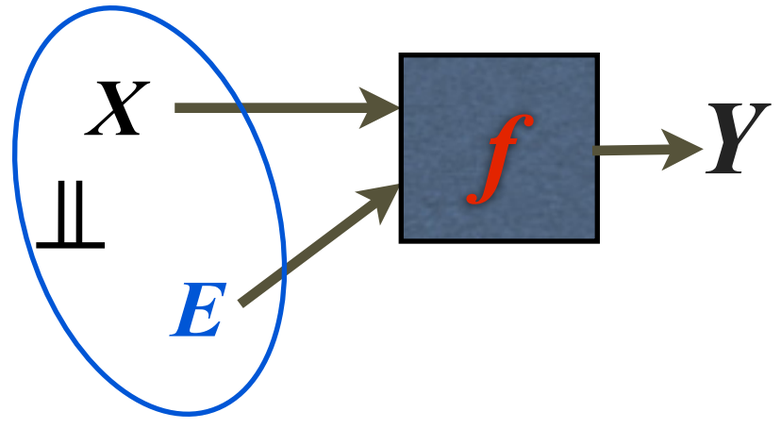
Some Real Data Sets







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"

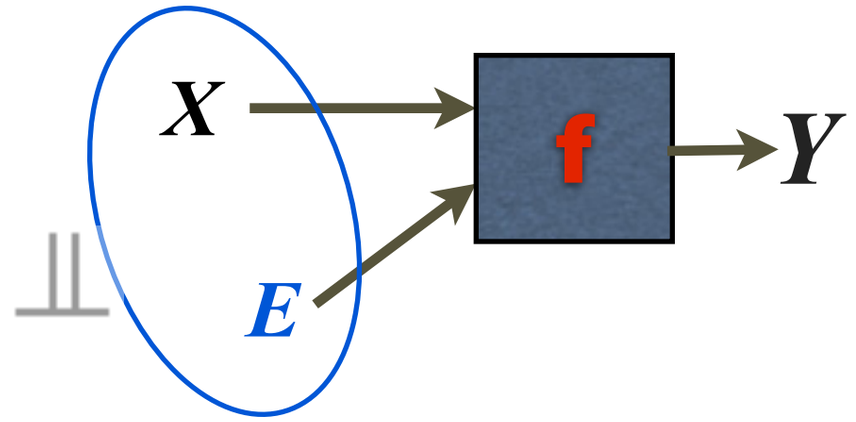
$$P(Y|X)$$
$$P(X) \rightarrow X \rightarrow Y$$

- :-(Without constraints on f, one can find independent noise for both directions (Darmois, 1951; Zhang et al., 2015)

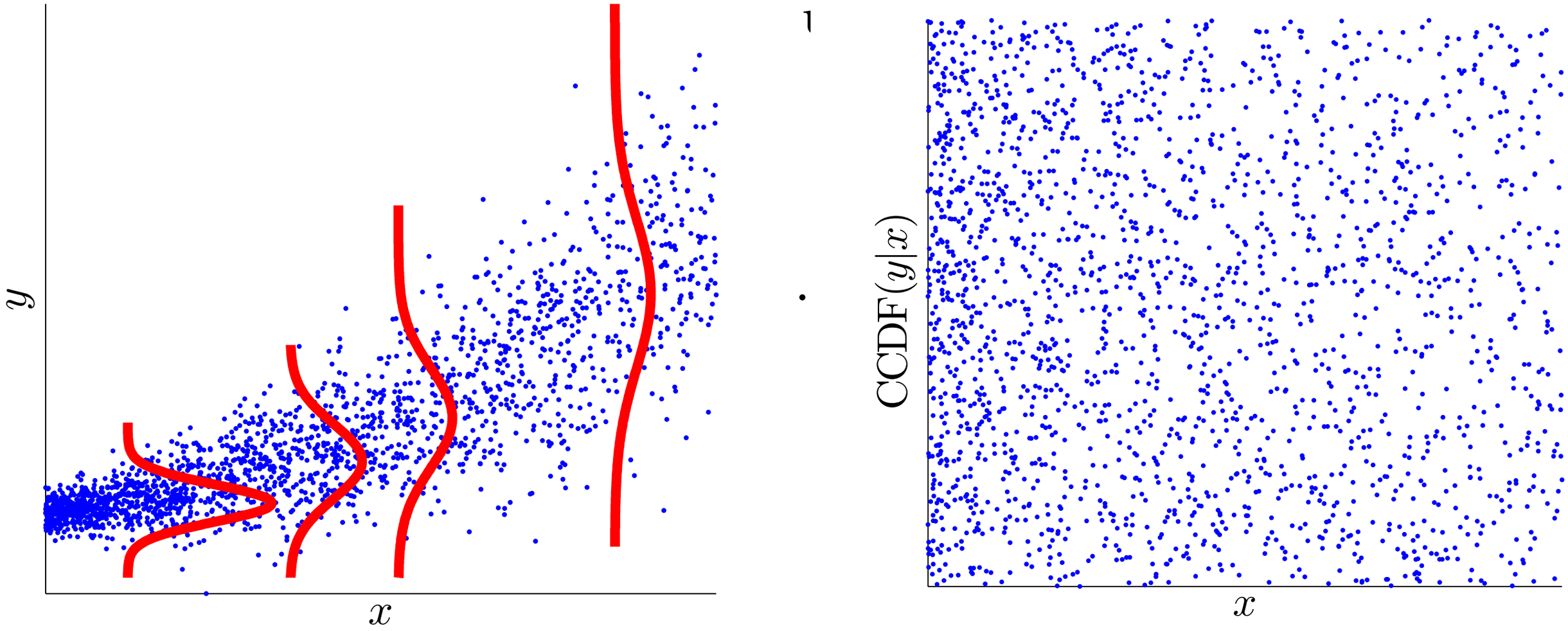
  - Given any X_1 and X_2, E' := conditional CDF of $X_2 \mid X_1$ is always independent from X_1 and $X2 = f(X_1, E')$

- :-) 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)

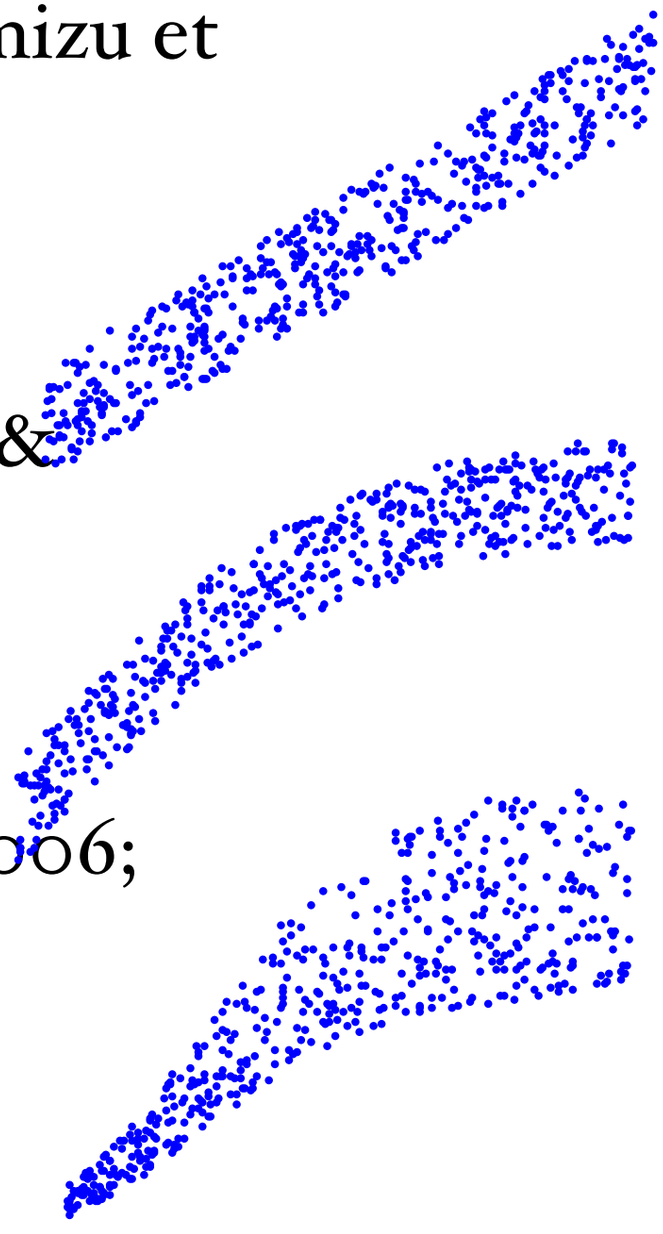
$$Y = a \cdot X + E$$

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

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

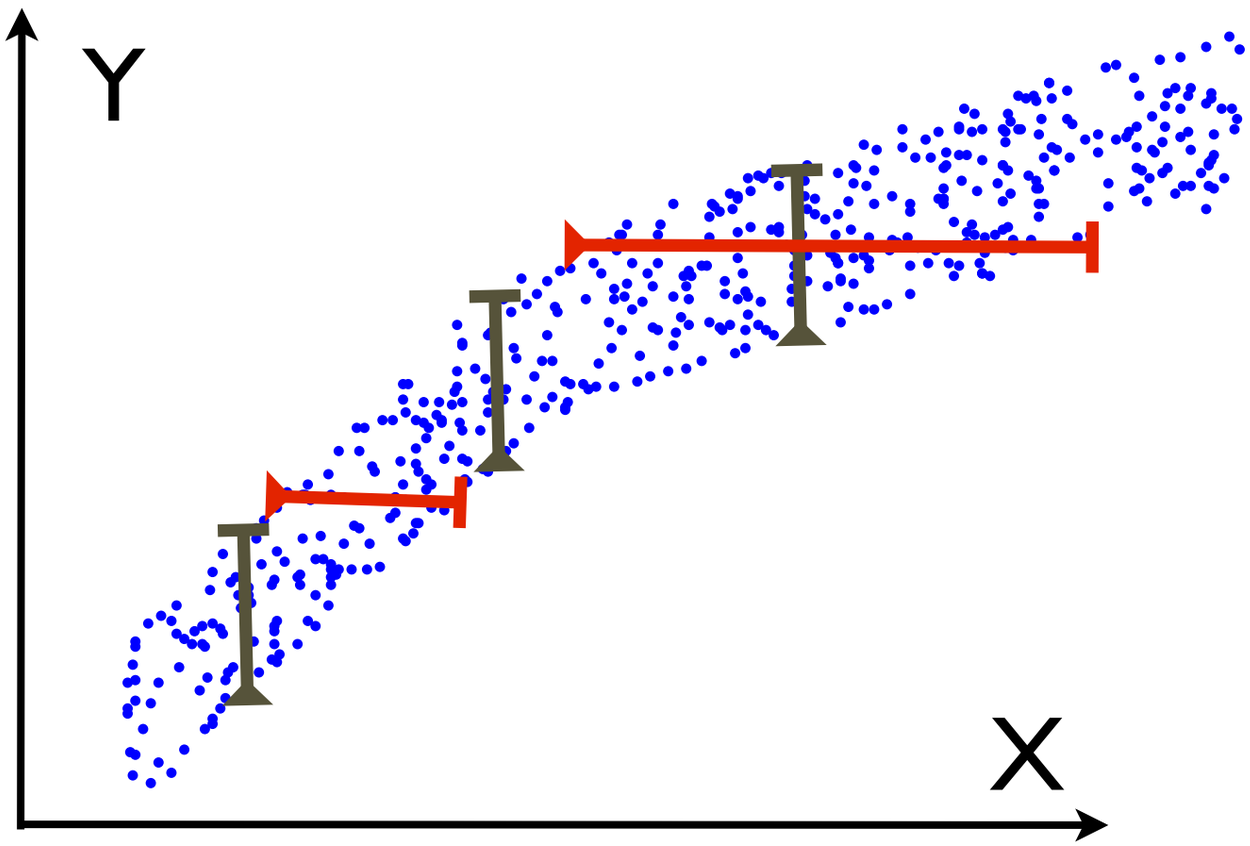
- Post-nonlinear causal model (Zhang & Chen, 2006; Zhang & Hyvärinen, '09a)

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



Causal Asymmetry with Nonlinear Additive Noise: Illustration

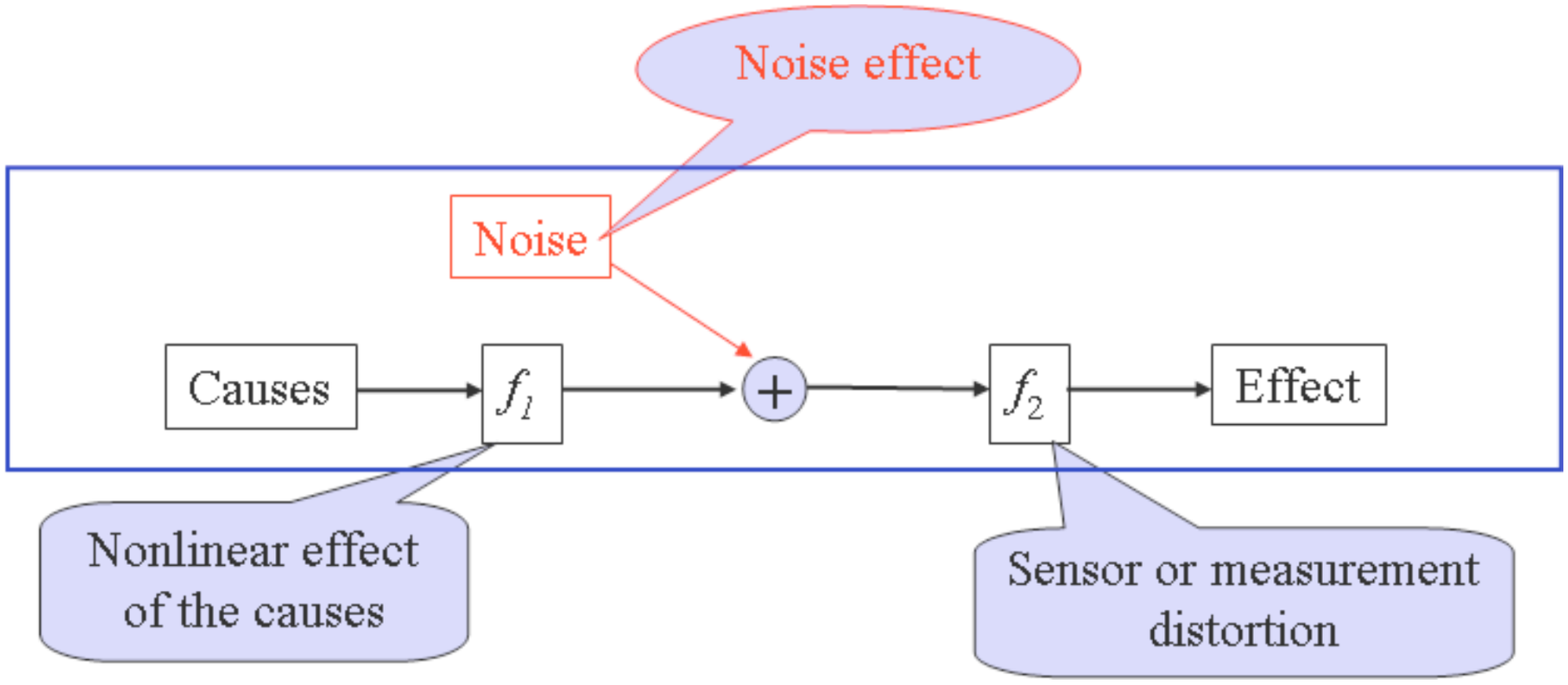
$$Y = f(X) + E \text{ with } E \perp\!\!\!\perp X$$



(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 post-nonlinear causal model with inner additive noise

PNL Causal Model

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

pa_i: parents (causes) of x_i

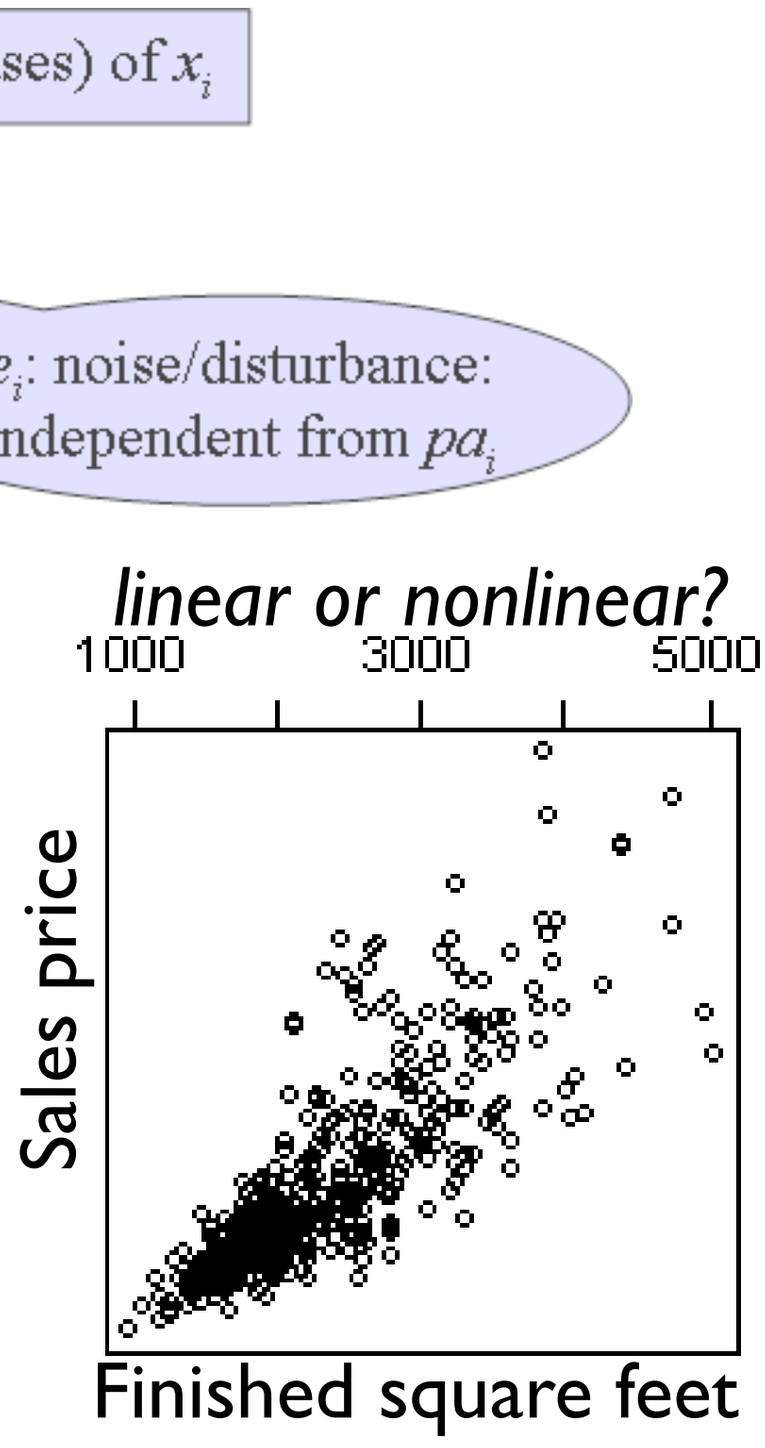
$f_{i,2}$: assumed to be continuous and invertible

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

e_i: noise/disturbance: independent from pa_i

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

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



To Examine If $X_1 \rightarrow X_2$ with MLP Implementation *

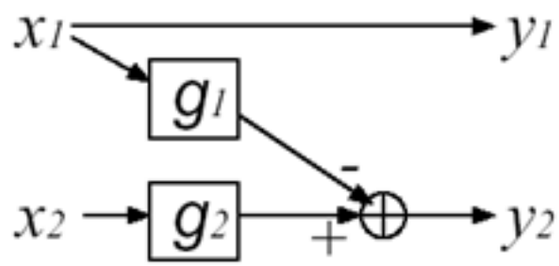
- If $X_1 \rightarrow X_2$, i.e., $X_2 = f_{2,2}(f_{2,1}(X_1) + E_2)$, we have

  $E_2 = f_{2,2}^{-1}(X_2) - f_{2,1}(X_1)$ is independent from X_1

- Two-step procedure to examine if $X_1 \rightarrow X_2$
  - Step 1: constrained nonlinear ICA to estimate E_2

    • $y_2 = g_2(x_2) - g_1(x_1)$; Y_2 and X_1 as independent as possible, such that Y_2 provides $\hat{E}_2$.

    • Parameters learned by minimizing the mutual information (equivalent to negative likelihood):



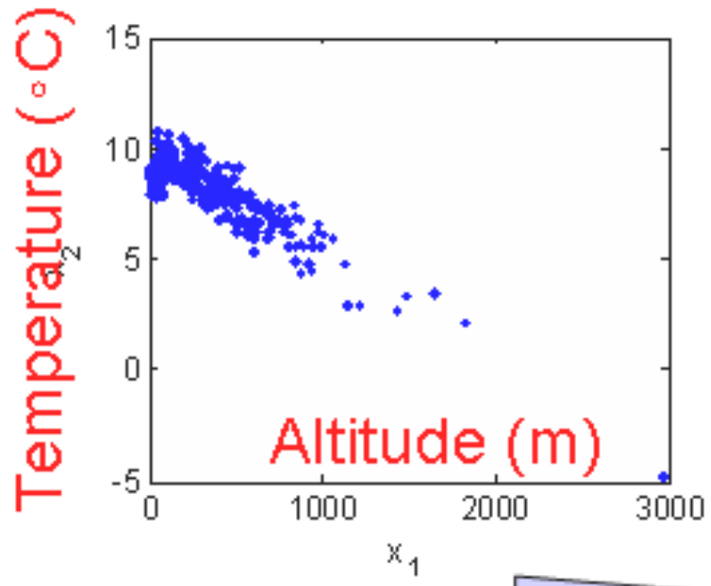
(y_2 produces an estimate of e_2)

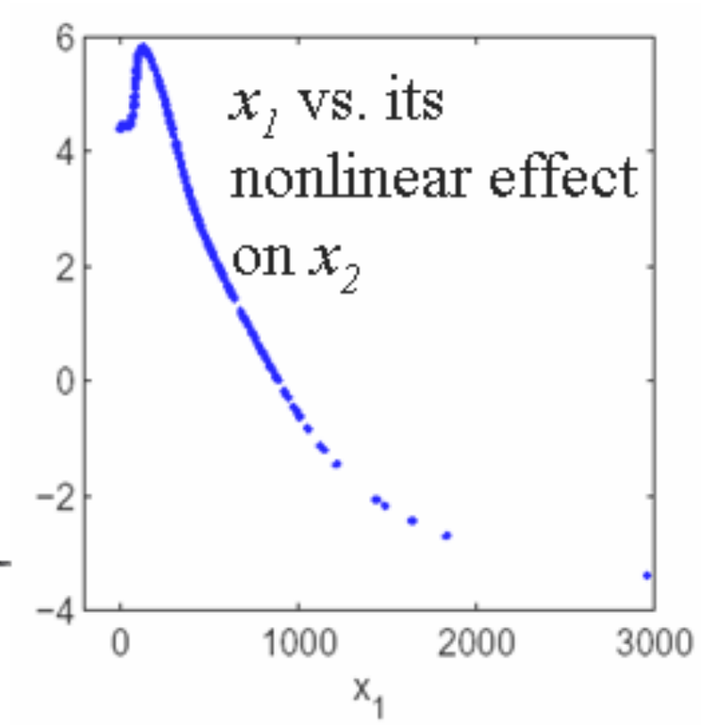
$$I(X_1, Y_2) = H(X_1) + H(Y_2) + E\{\log|\mathbf{J}|\} - H(X_1, X_2)$$
$$= -E\log p_{Y_2} - E\{\log|g_2'(X_2)|\} + \text{const}$$

  - Step 2: uses independence tests to verify if X_1 and $\hat{E}_2$ are independent

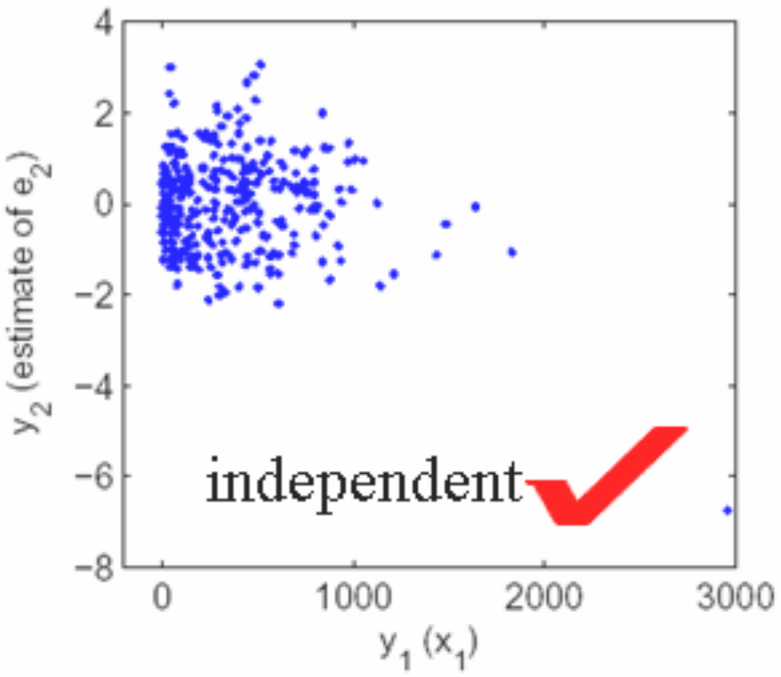
Data Set 1

with PNL Model

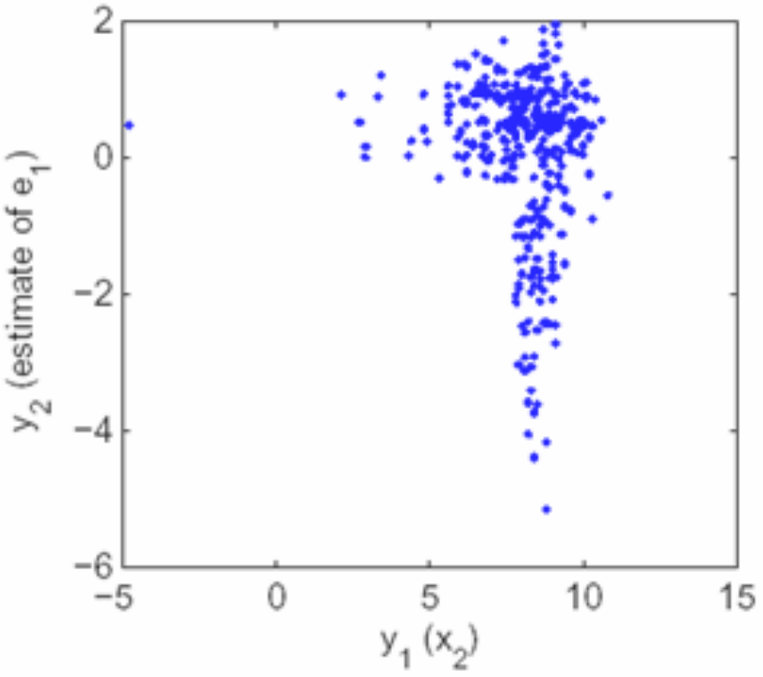


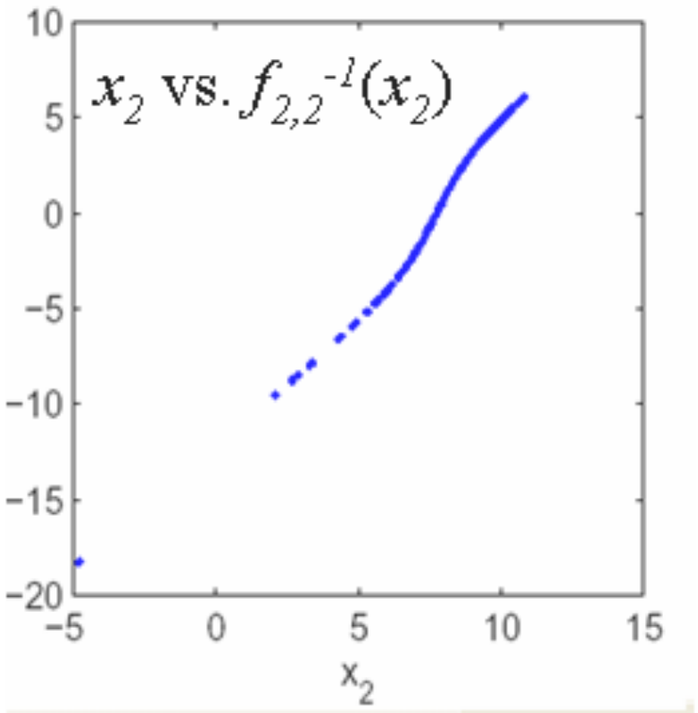


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$



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

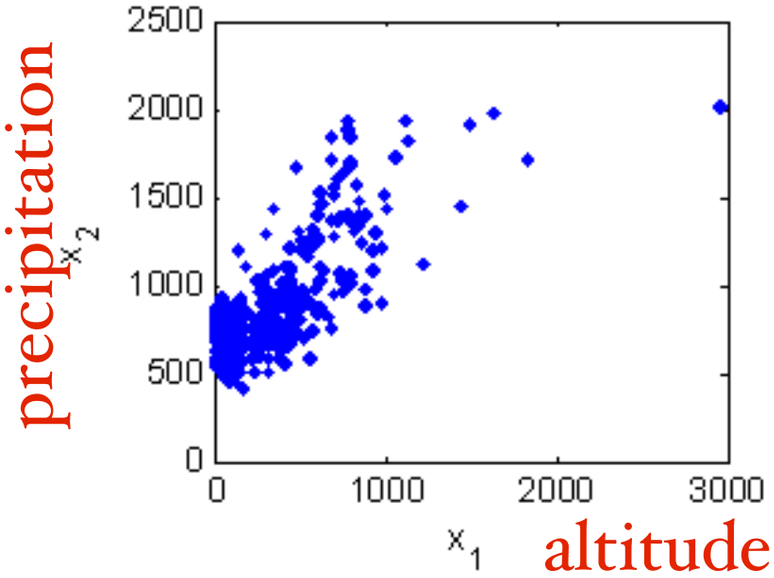




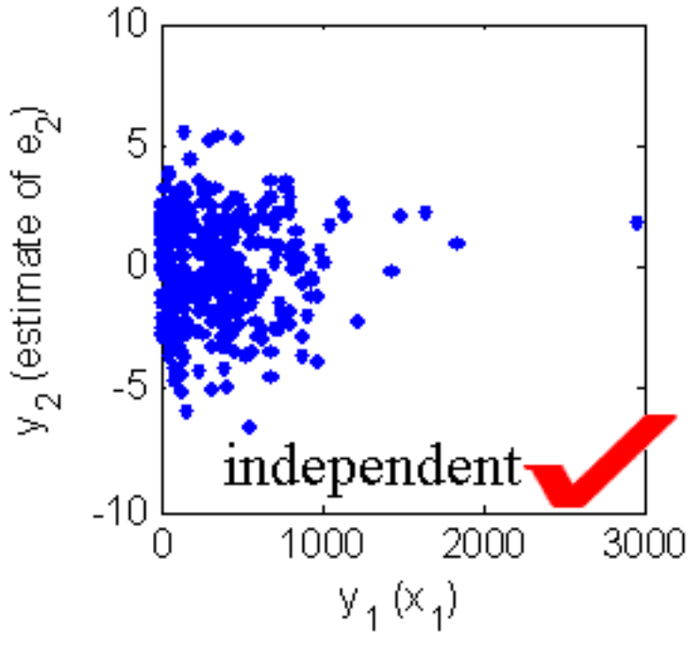
Independence test results on y_1 and y_2 with different assumed causal relations

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

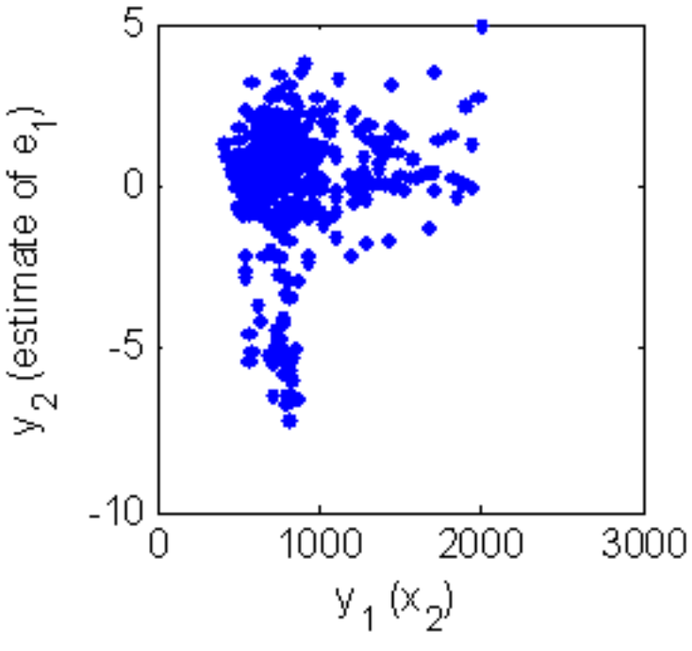
Data Set 2

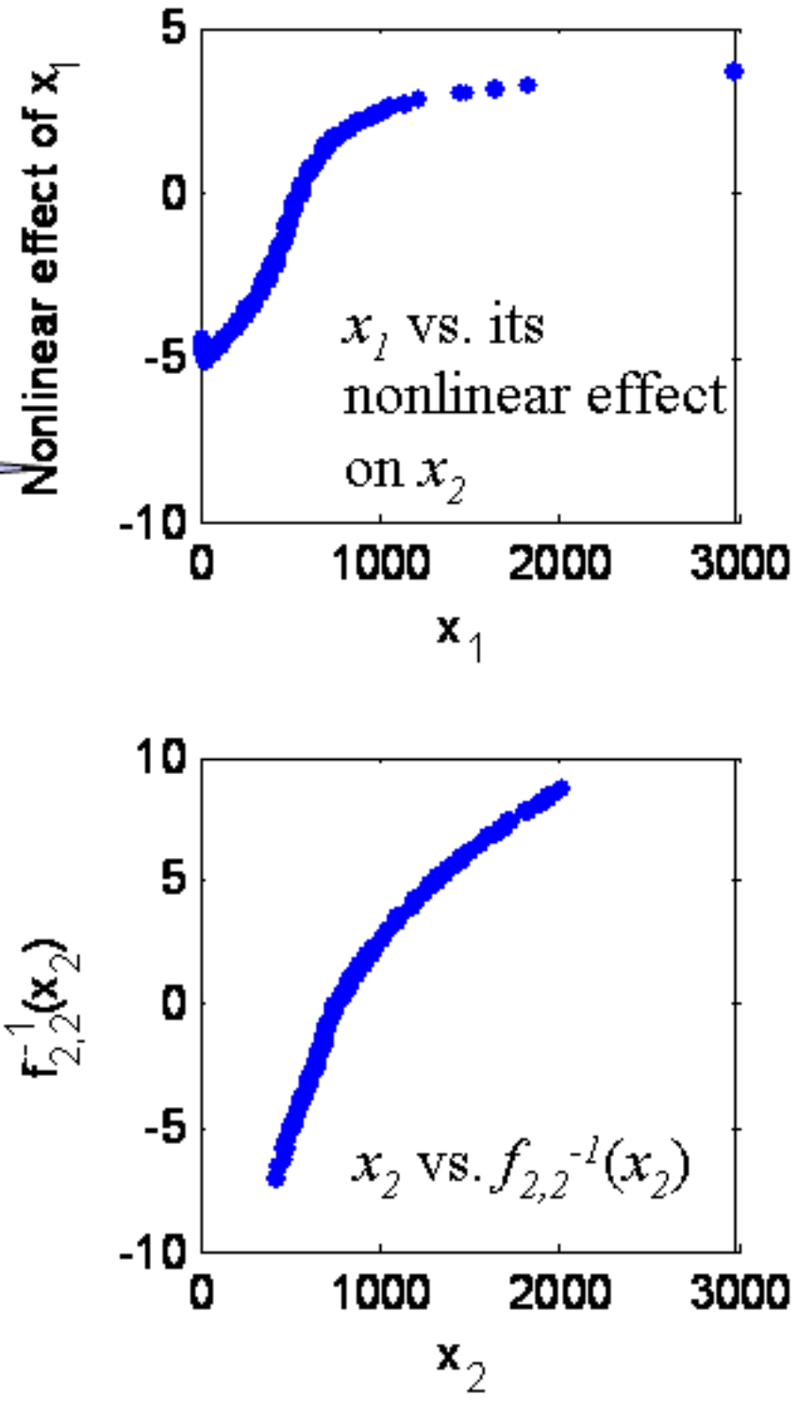


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

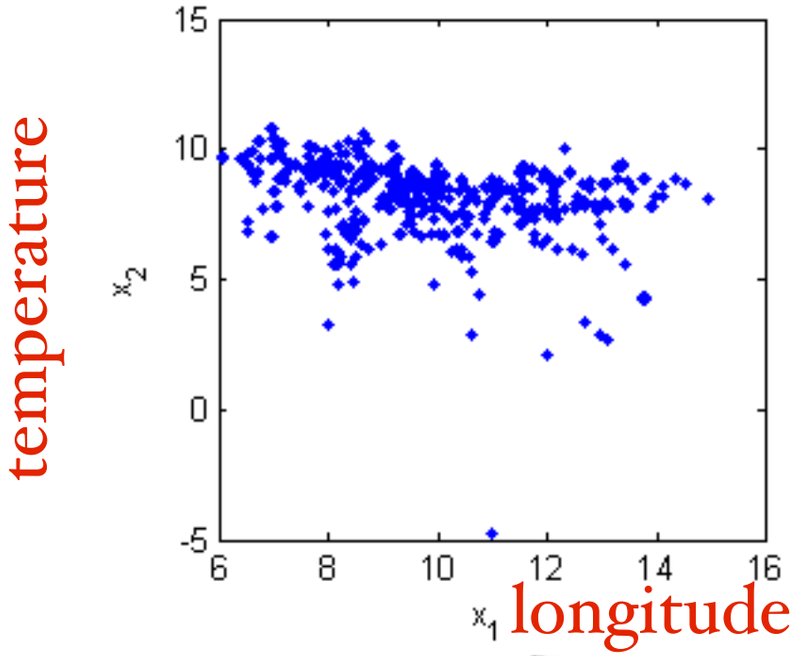


(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$



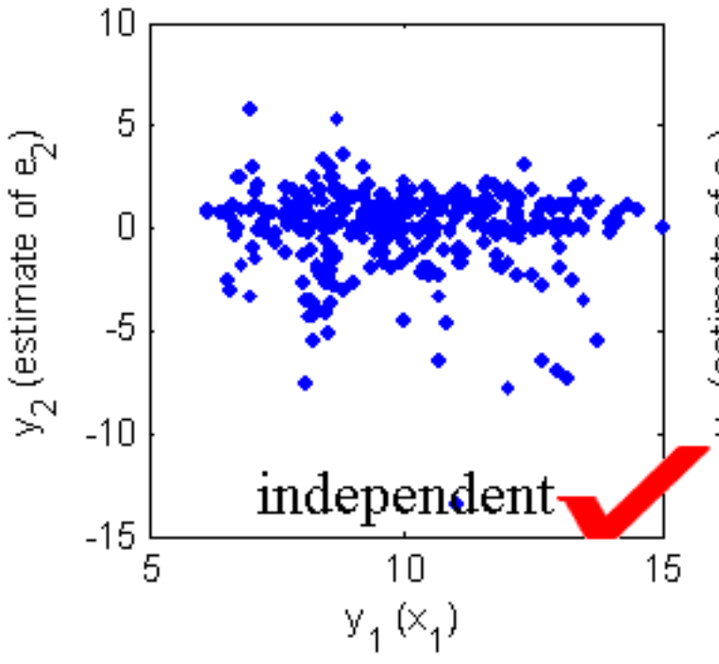


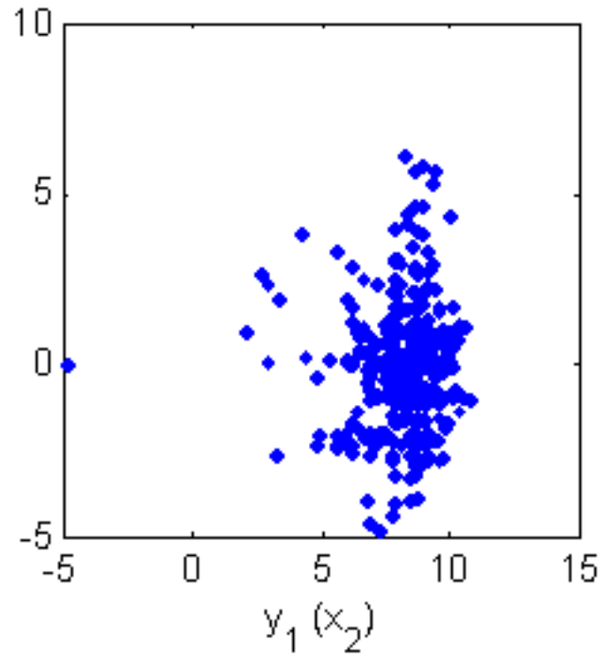
Data Set 3

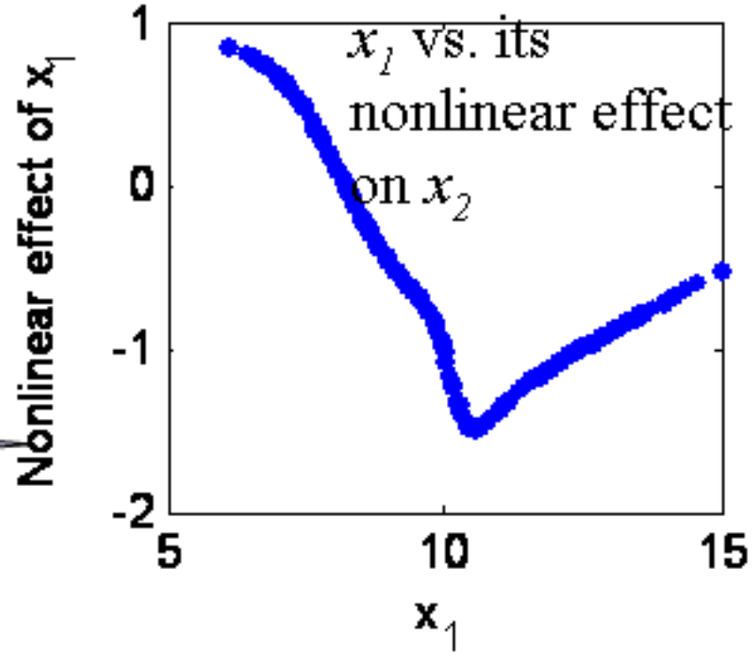


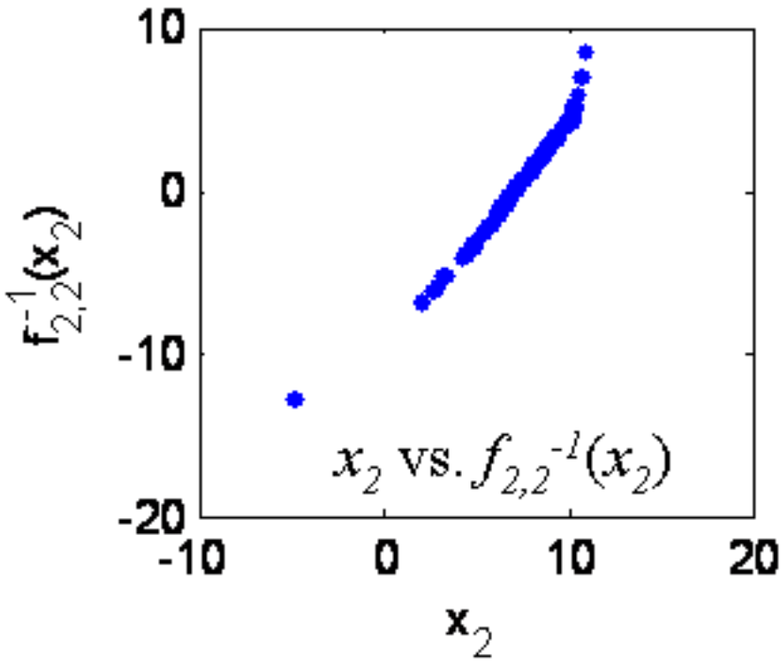
(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

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

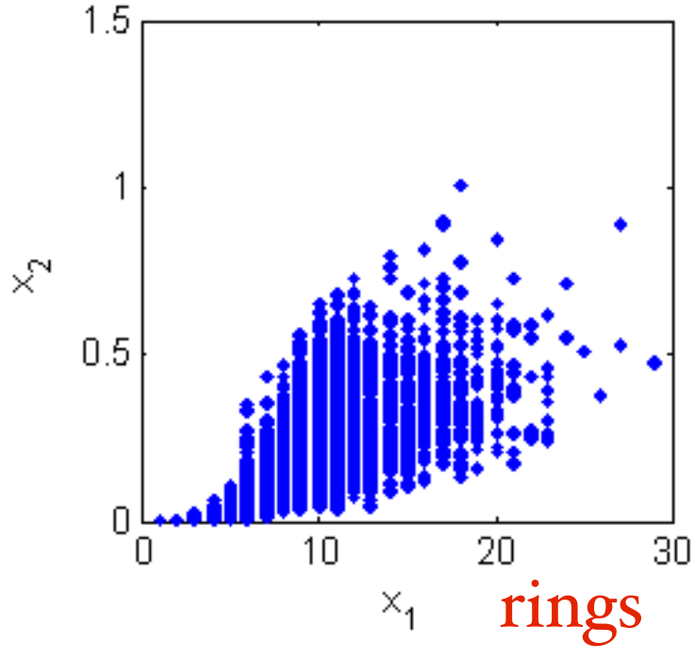


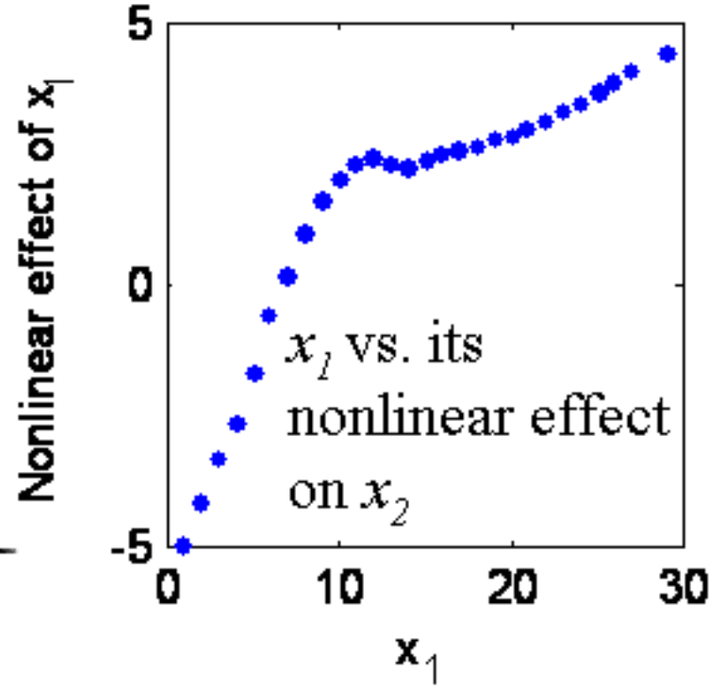






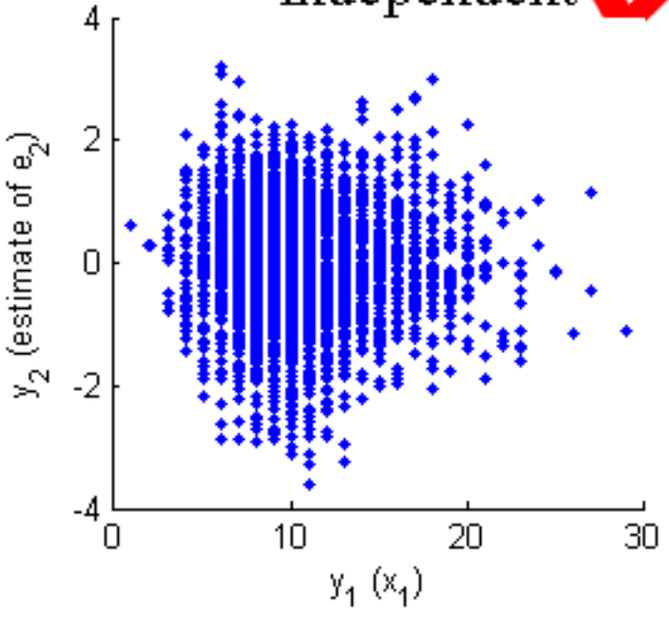
Data Set 6



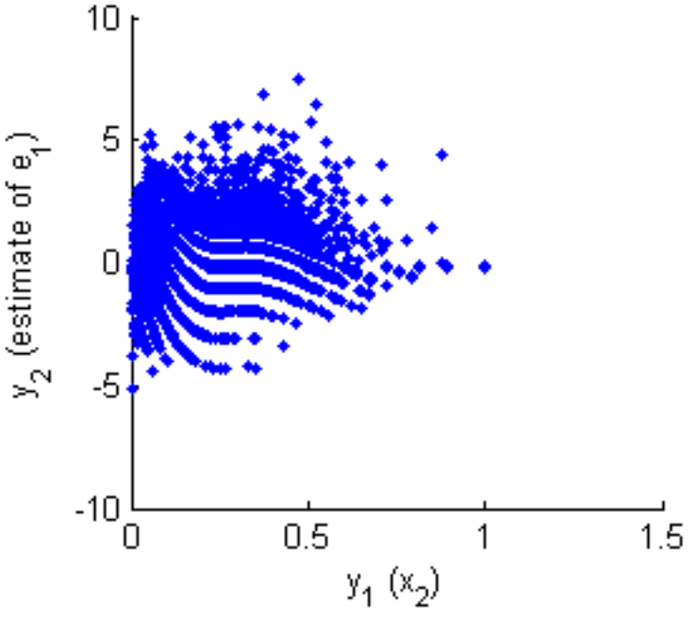


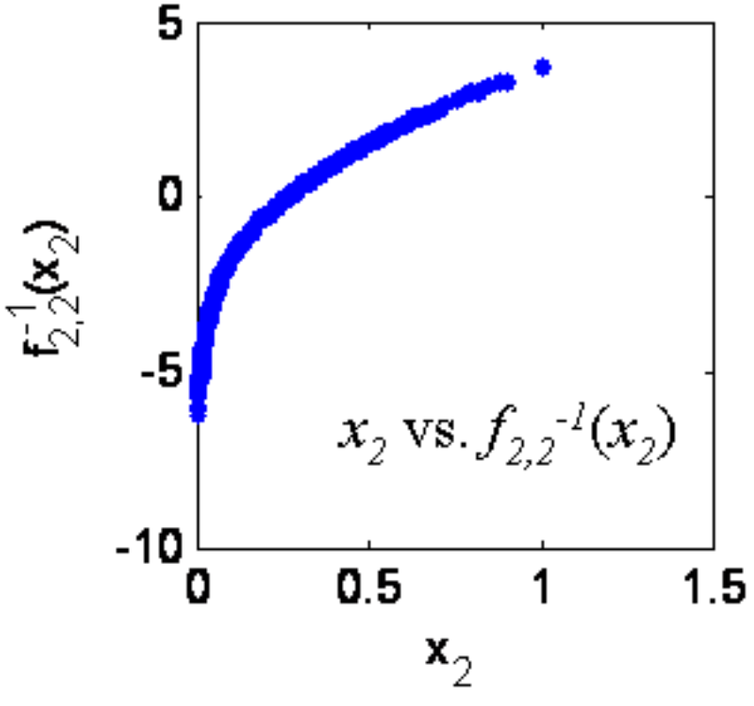
(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

independent ✓

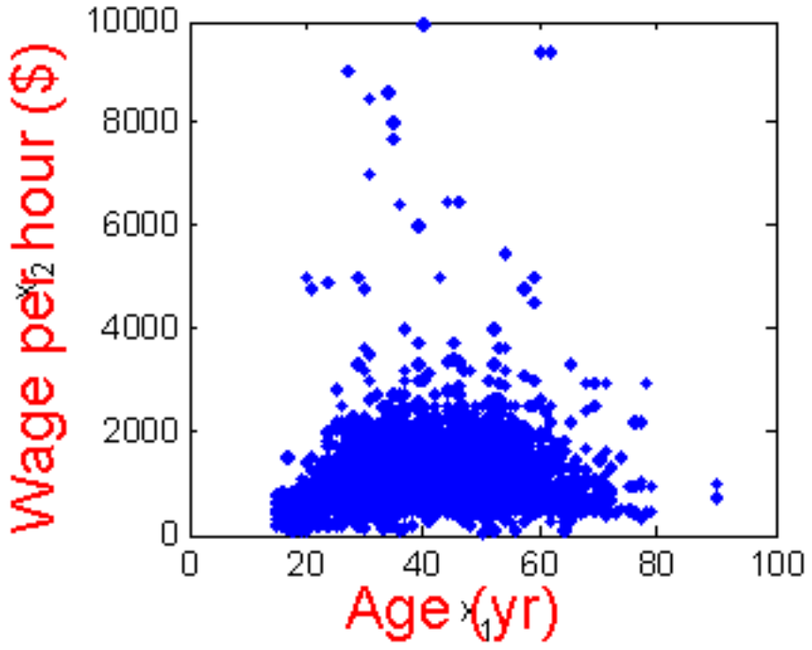


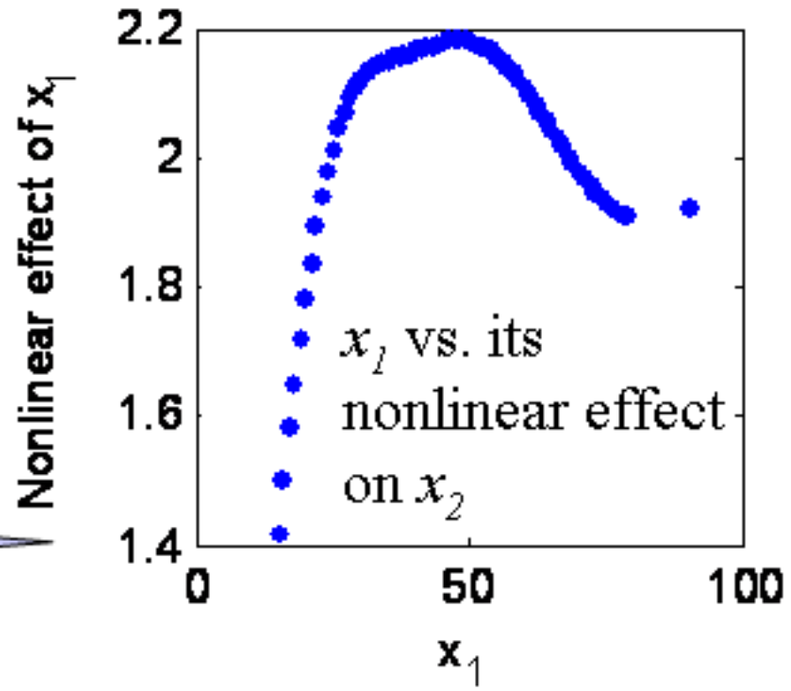
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$



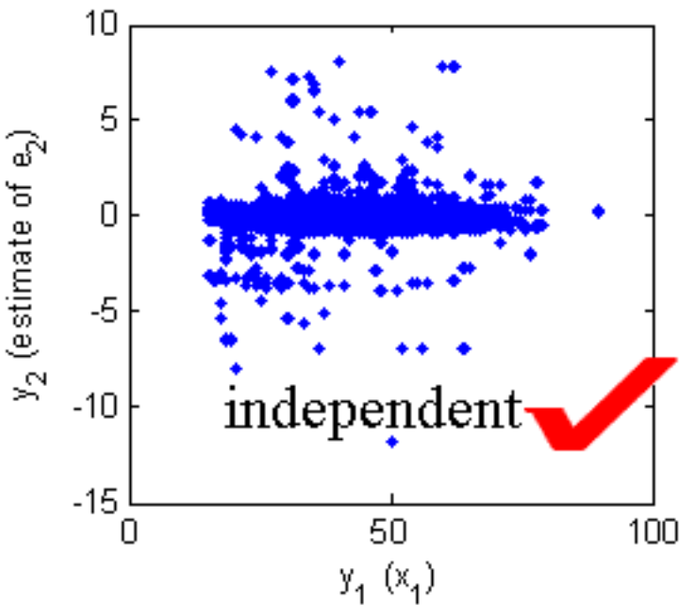


Data Set 8

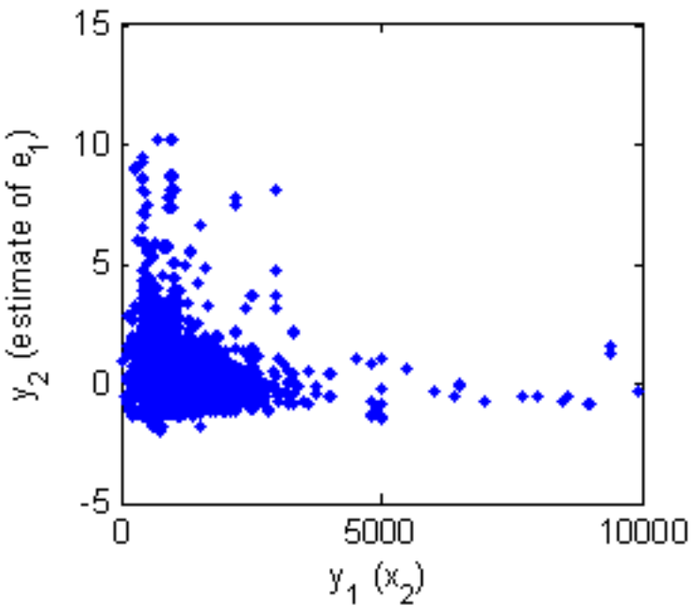


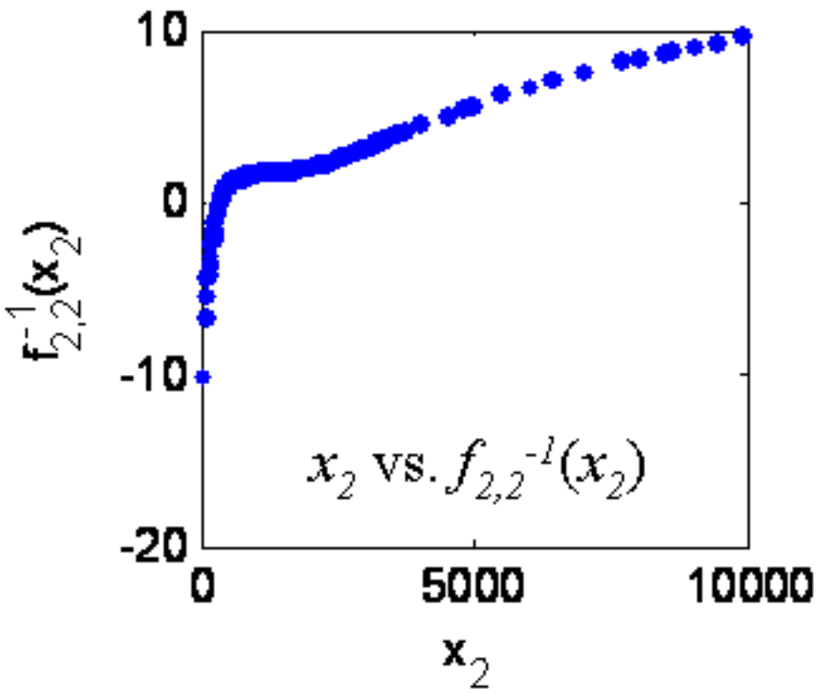


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$



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





Identifiability in Two-variable Case: Theoretical Results

pa_i: parents (causes) of x_i

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

$f_{i,2}$: 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: A Mathematical Result

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

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

Notation

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

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

List of All Non-Identifiable Cases

Log-mixed-linear-and-exponential:

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

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

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

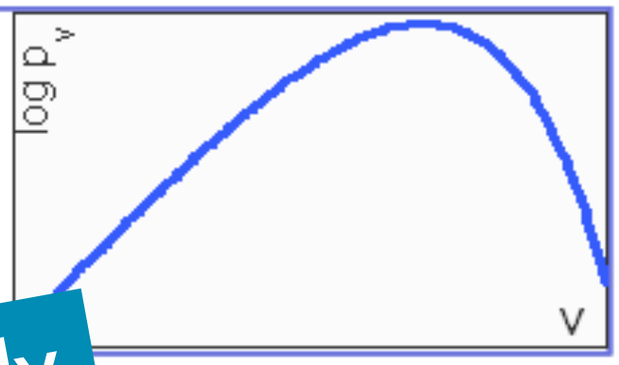


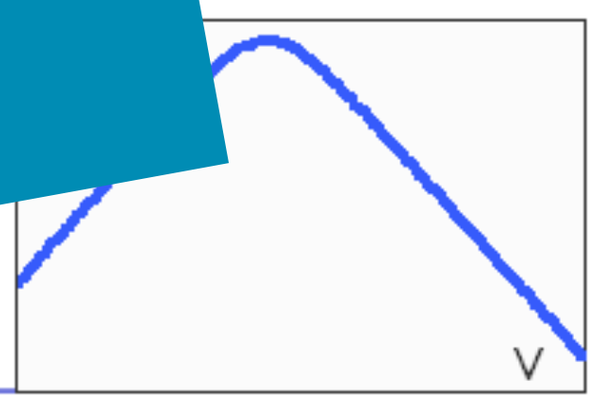
Table 1: All situations i… not identifiable.

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

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

$\to c_2 \ (c_2 \neq 0)$,

as $v \to +\infty$



Causal direction is generally identifiable if the data were generated according to

$$X_2 = f_2(f_1(X_1) + E).$$

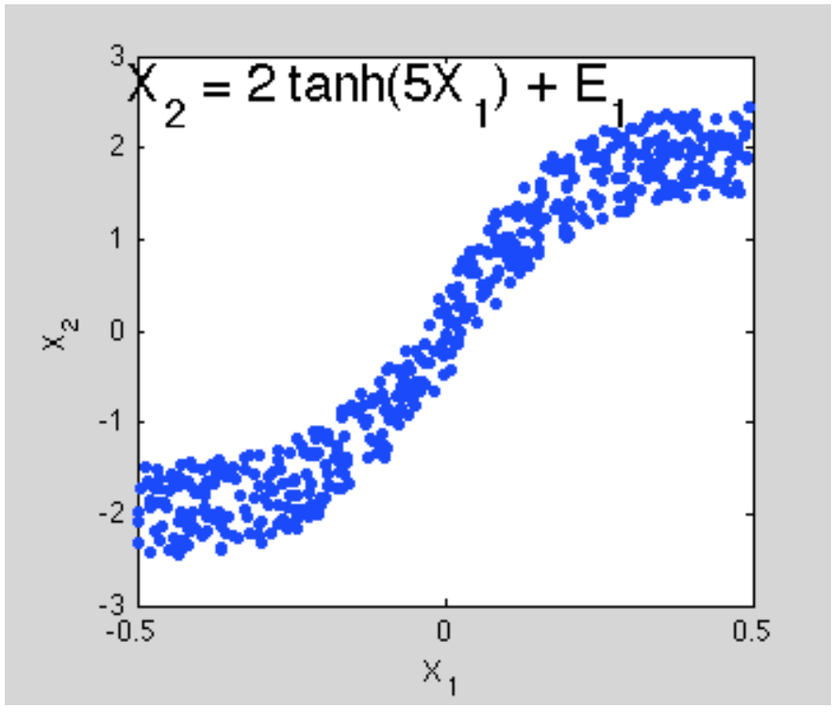
Linear models and nonlinear additive noise models are special cases.

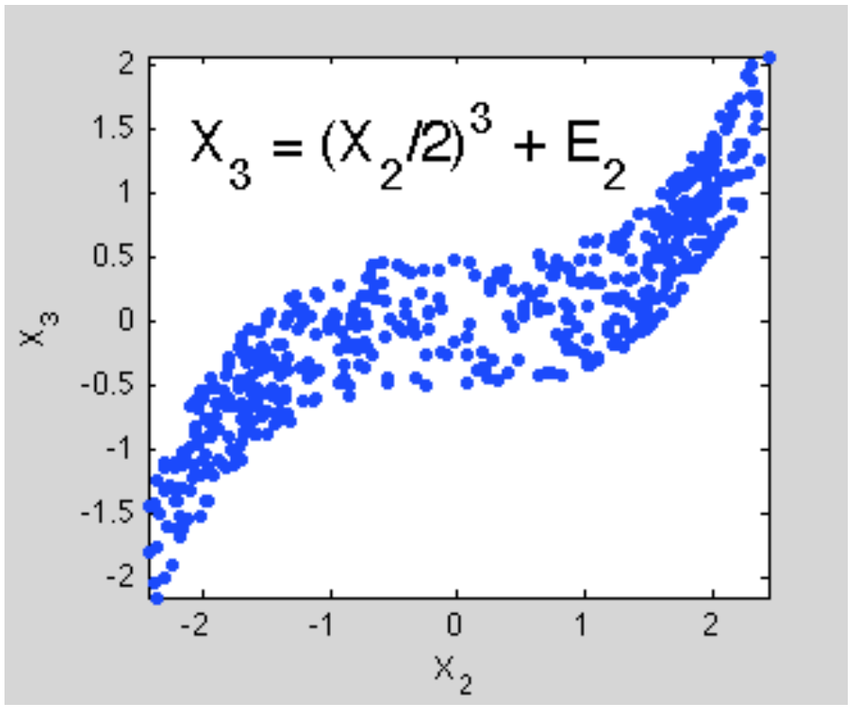
Take-Home Message: Causal Discovery with Nonlinear Functional Causal Models

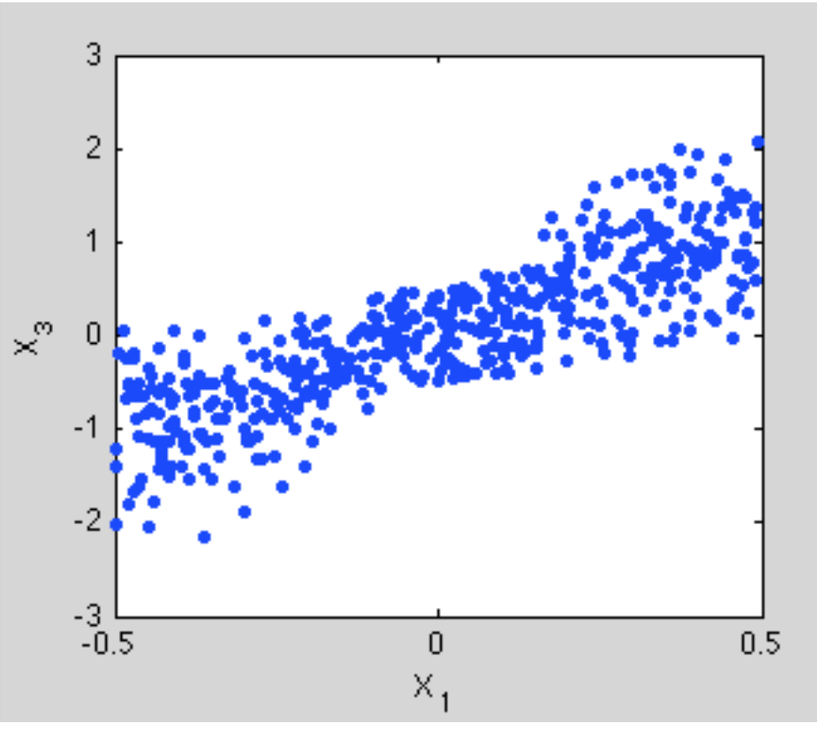
- Functional causal models naturally describe the causal processes
- Can we use them to distinguish cause from effect?
- Certain types of constraints on f are needed to guarantee the identifiability of the causal direction
- Nonlinearities are encountered frequently and should be considered
- Trade-off of generality & identifiability
- Limitation: more than one noise term? large-scale problems?

Transitivity of FCMs and Intermediate Causal Variable Recovery

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

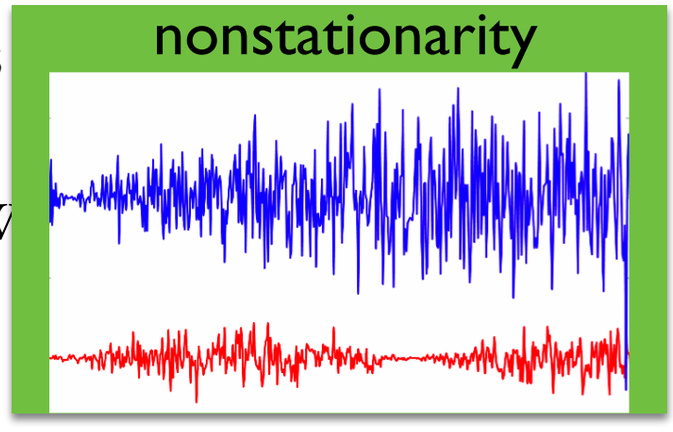






Practical Issues in Causal Discovery...

- 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)



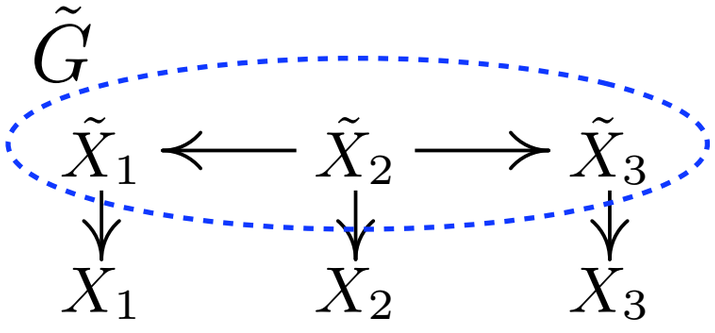
- **Nonstationary/heterogeneous data** (Zhang et al., IJCAI'17; Huang et al, ICDM'17, Ghassami et al., NIPS'18; Huang et al., ICML'19 & NIPS'19; Huang et al., JMLR'20)

Issue 2: Causal Discovery in the Presence of Measurement Error

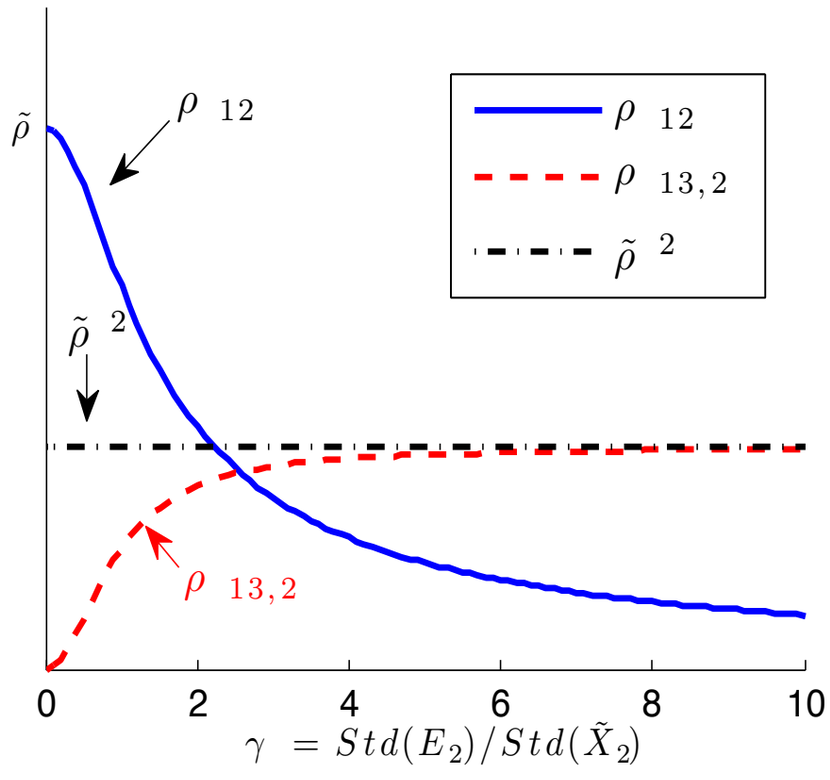
- To estimate $\tilde{G}$ over variables $\tilde{X}_i$ from noisy observations $X_i = \tilde{X}_i + E_i$.
- Conditional independence/dependence relations among X_i different from those among $\tilde{X}_i$

Measurement error changes causal discovery results!

- Illustration: Correlation(X_1, X_2) & partial_correlation(X_1, X_3 | X_2)



$$X_i = \tilde{X}_i + E_i.$$

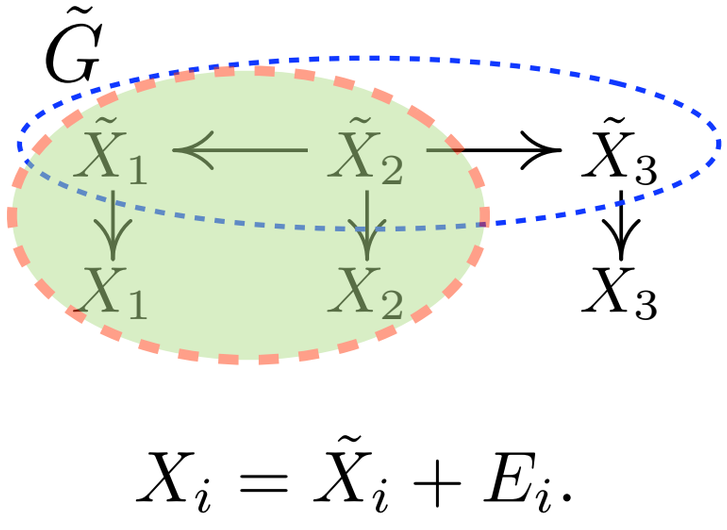


Causal Discovery in the Presence of Measurement Error

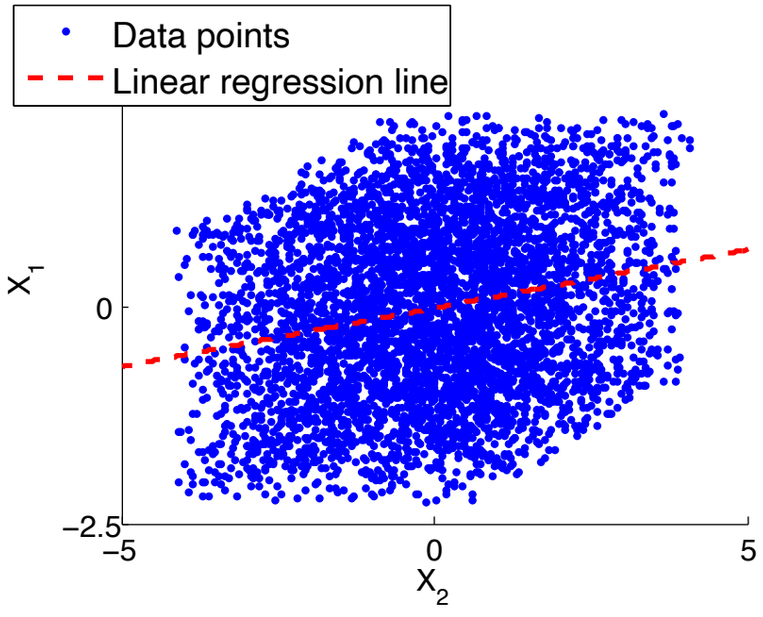
- To estimate $\tilde{G}$ over variables $\tilde{X}_i$ from noisy observations $X_i = \tilde{X}_i + E_i$.
- Conditional independence/dependence relations among X_i different from those among $\tilde{X}_i$

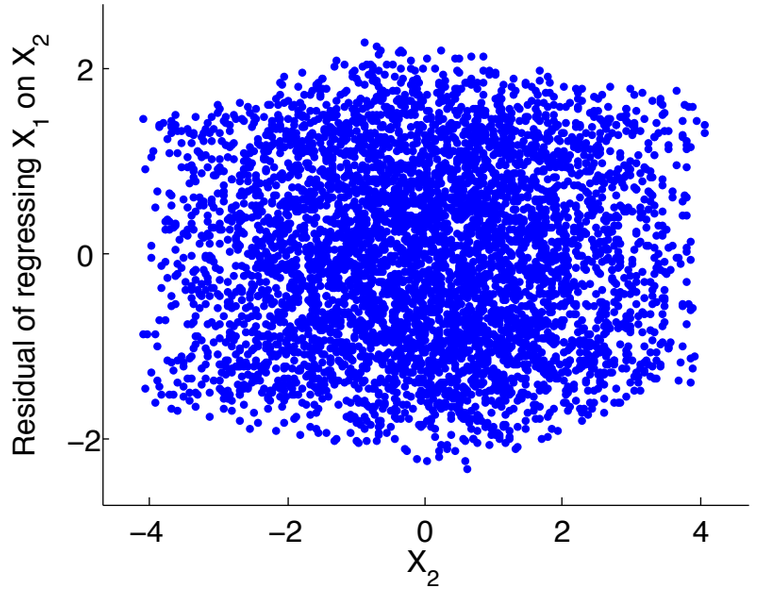
Measurement error changes causal discovery results!

- Illustration: causal model $X_1 \leftarrow X_2$?

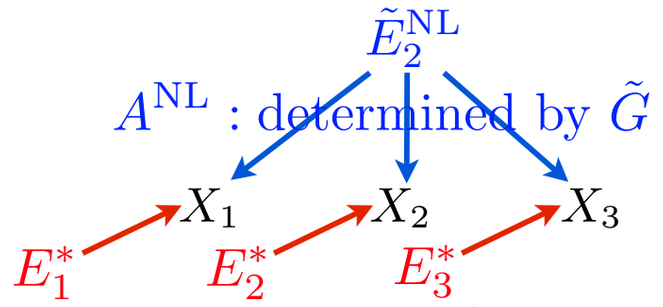


$$X_i = \tilde{X}_i + E_i.$$





Example of CR-CAMME



Factor analysis model: $\mathbf{X} = \tilde{\mathbf{X}}^* + \mathbf{E}^*$

$= A^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}} + \mathbf{E}^*$

Alternatively: $\mathbf{X} = \left[\ \mathbf{A}^{\mathrm{NL}} \mid \mathbf{I}\ \right] \cdot \left[\dfrac{\tilde{\mathbf{E}}^{\mathrm{NL}}}{\mathbf{E}^*}\right]$

Suppose $\tilde{G}$ is $\tilde{X}_1 \xrightarrow{a} \tilde{X}_2 \xleftarrow{b} \tilde{X}_3$:

So $\tilde{\mathbf{X}} = \mathbf{B}\tilde{\mathbf{X}} + \tilde{\mathbf{E}}$, with $\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ a & 0 & b \\ 0 & 0 & 0 \end{bmatrix}$.

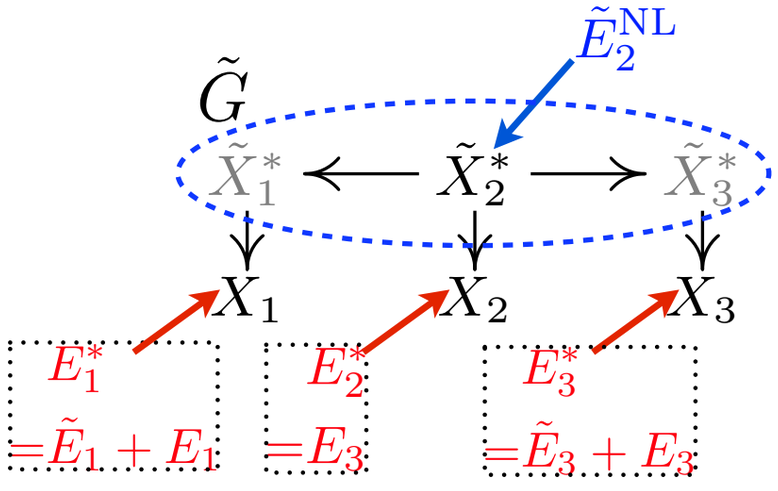
That is, $\tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{E}}$, with $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ a & 1 & b \\ 0 & 0 & 1 \end{bmatrix}$.

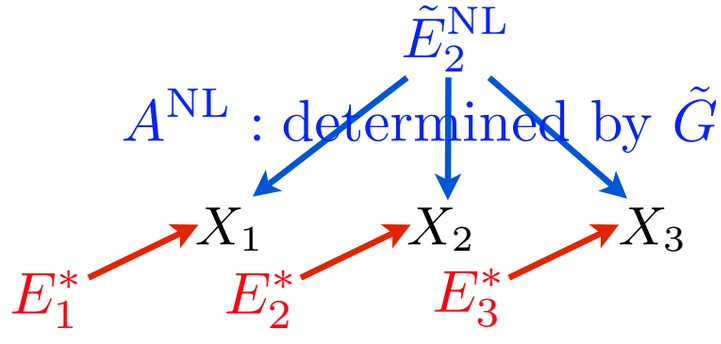
Therefore,

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{E} = \tilde{\mathbf{X}}^* + \mathbf{E}^* = \begin{bmatrix} 1 & 0 \\ a & b \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \tilde{E}_1 \\ \tilde{E}_3 \end{bmatrix} + \begin{bmatrix} E_1 \\ \tilde{E}_2 + E_2 \\ E_3 \end{bmatrix} = \left[\begin{array}{cc|ccc} 1 & 0 & 1 & 0 & 0 \\ a & b & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array}\right] \cdot \begin{bmatrix} \tilde{E}_1 \\ \tilde{E}_3 \\ E_1 \\ \tilde{E}_2 + E_2 \\ E_3 \end{bmatrix}.$$

A^{NL} (of size $n \times (n - l)$)

* Identifiability of CR-CAMME: Second-Order Statistics

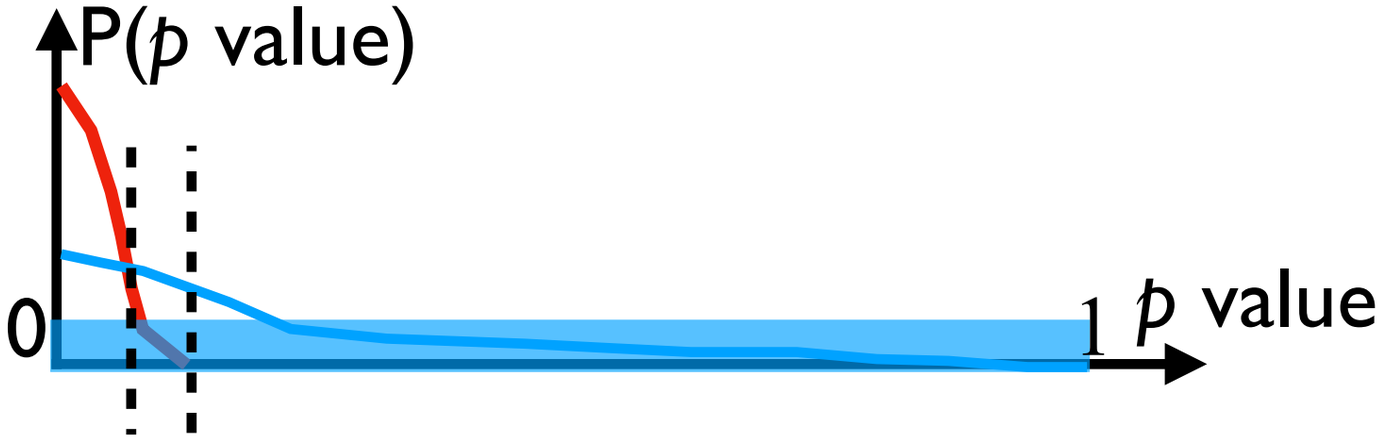




Factor analysis model: $\mathbf{X} = \tilde{\mathbf{X}}^* + \mathbf{E}^*$

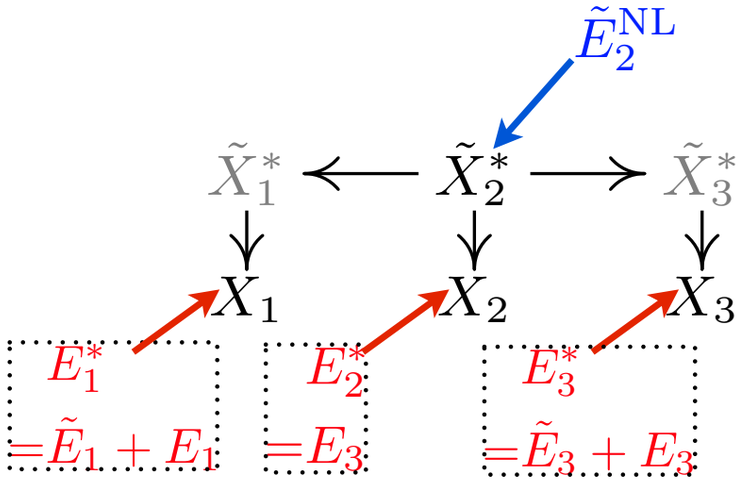
$= A^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}} + \mathbf{E}^*$

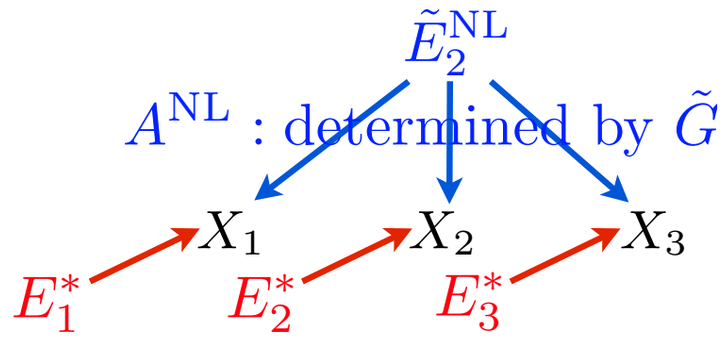
- Identifiability conditions derived based on the <u>factor analysis</u> model: the number of non-leaf nodes has to be small
- Conditions improved if <u>measurement errors have the same variance</u>
- Heuristic correction method: use a small significance level when doing CI tests



Zhang, et al. "Causal Discovery in the Presence of Measurement Error: Identifiability Conditions," UAI 2017 Workshop on Causality

Non-Gaussian Case: Thanks to Over-Complete ICA





Factor analysis model: $\mathbf{X} = \mathbf{X}^* + \mathbf{E}^*$

$$= A^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}} + \mathbf{E}^*$$

$$= \left[\ \mathbf{A}^{\mathrm{NL}} \mid \mathbf{I}\ \right] \cdot \left[\frac{\tilde{\mathbf{E}}^{\mathrm{NL}}}{\mathbf{E}^*}\right]$$

- A^{NL} is identifiable up to permutation and scaling of columns under assumption (Eriksson and Koivunen, 2004):

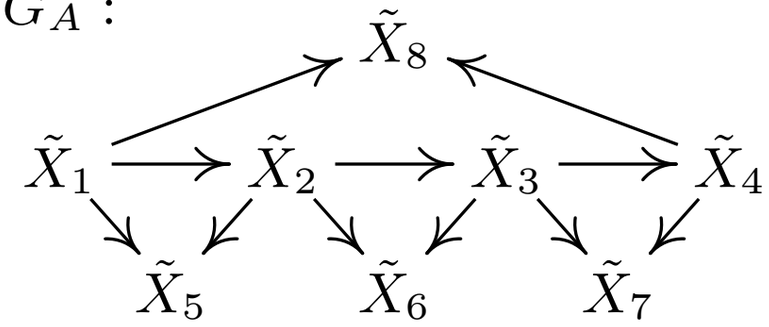
> A1. All $\tilde{E}_i$ are non-Gaussian.

- In original LiNGAM, causal direction can be determined by testing independence between regression residual & predictors
- We cannot estimate the noise terms because it is overcomplete
- Ordered group decomposition is identifiable by analyzing A^{NL}

Ordered Group Decomposition is Identifiable

- Decompose all nodes in $\tilde{G}$ into disjoint groups
- Each group contains a single non-leaf node + its "direct-and-only-direct" effect leaf nodes
- Causal ordering of such groups is identifiable

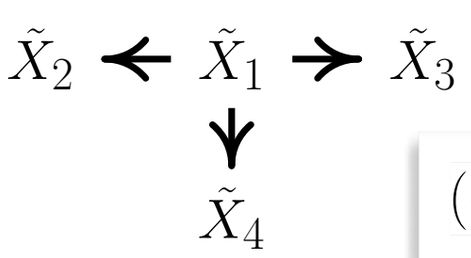
$\tilde{G}_A$:



Ordered group decomposition:

$$(\{\tilde{X}_1^*\} \to \{\tilde{X}_2^*, \tilde{X}_5^*\} \to \{\tilde{X}_3^*, \tilde{X}_6^*\} \to \{\tilde{X}_4^*, \tilde{X}_7^*, \tilde{X}_8^*\})$$

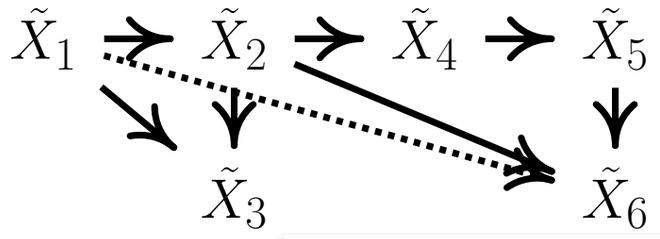
$\tilde{G}_B$:



$$(\{\tilde{X}_1^*, \tilde{X}_2^*, \tilde{X}_3^*, \tilde{X}_4^*\})$$

$\tilde{G}_C$ (solid lines as its edges):

$\tilde{G}_D$ (all lines as its edges):

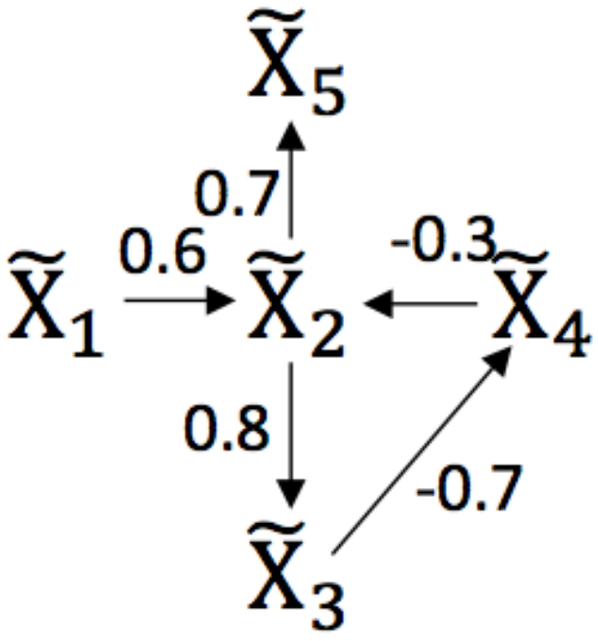


$$(\{\tilde{X}_1^*\} \to \{\tilde{X}_2^*, \tilde{X}_3^*\} \to \{\tilde{X}_4^*\} \to \{\tilde{X}_5^*, \tilde{X}_6^*\})$$

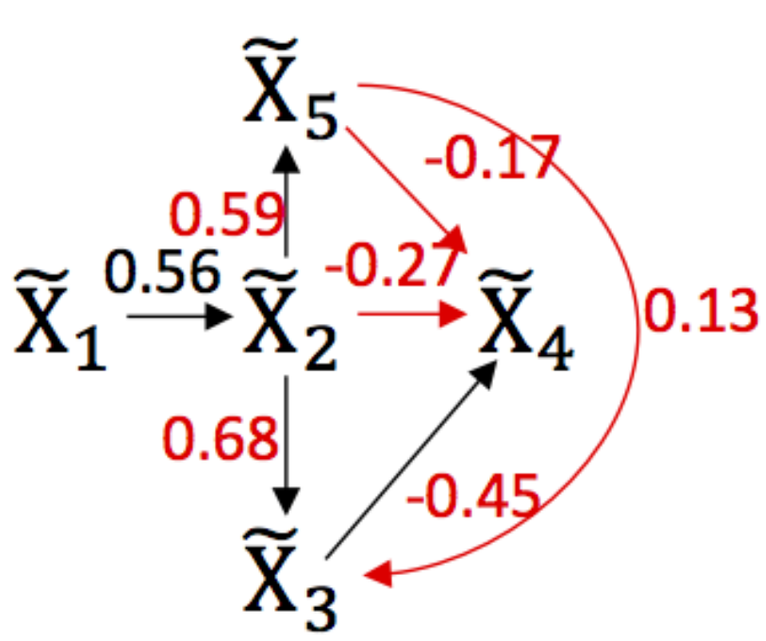
Simulation

- Development of statistically efficient estimation procedures is non-trivial
- Data were generated by the underlying true graph + measurement errors with different variances

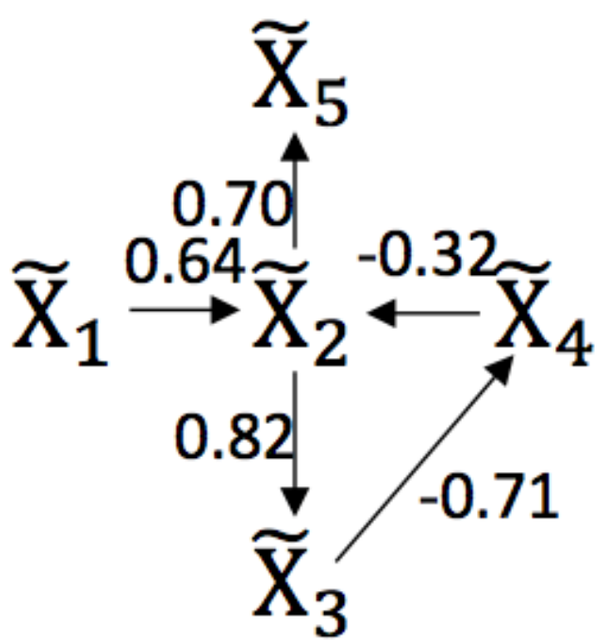
True underlying graph:



Estimated by LiNGAM:



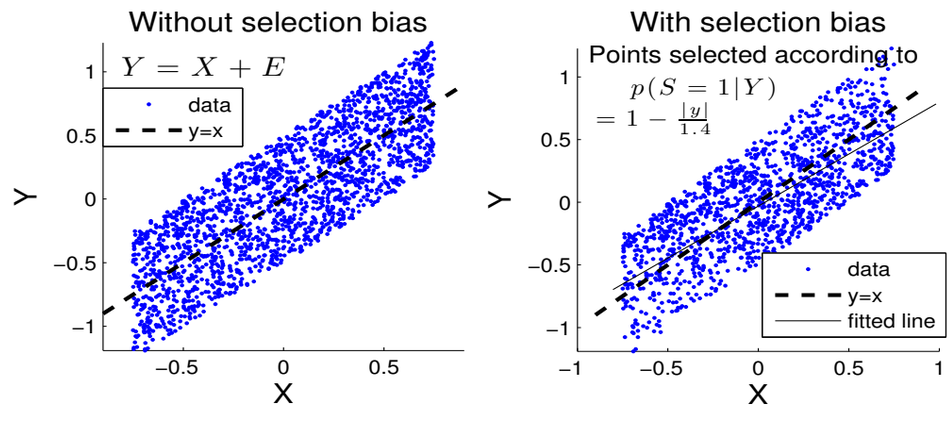
Estimated by our procedure:



Issue 3: Selection Bias

- Examples
  - Hospital-based disease research

- Selection bias: The chance of including a data point in the sample depends on some attributes of the point
- Often distorts the results of statistical analysis

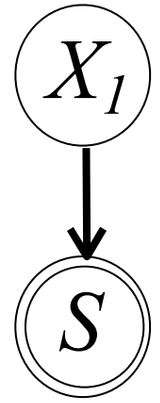


- In causal inference, both learning causal structures and estimating causal mechanisms become more difficult

Selection Bias: Illustration

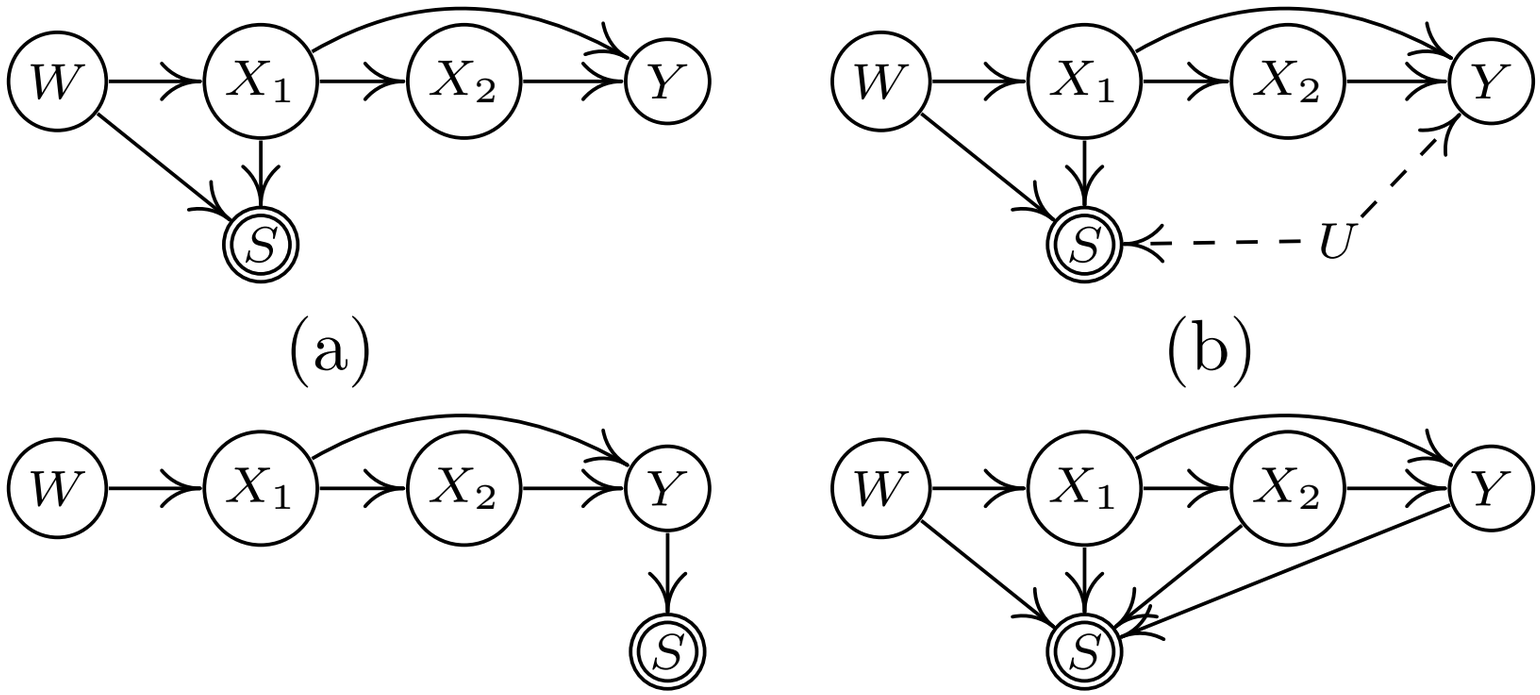
- Suppose the true causal process is

- Connection between the population and the distribution of the selected sample?
  - Section variable S (similar to missingness indicator); the selected sample follows $P(\mathbf{X} \mid S=1)$



- What will be the discovered causal structure if we select data points according to X_1?
  - X_4?
  - X_1 & X_4?
  - Other situations (e.g., X_4 is a common effect)?
- Suppose we work with data collected from patients…

Causal Discovery & Inference under Different Kinds of Selection Bias



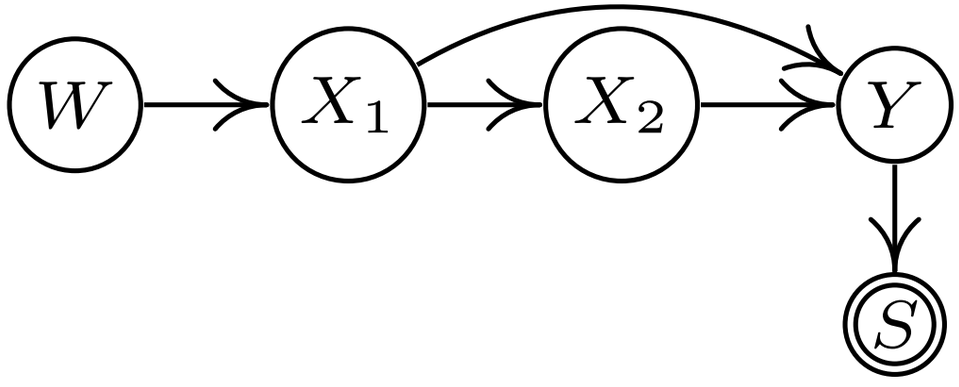
(a) (b)

(c) (d)

Selected sample follows $P_{XY|S=1}$ instead of P_{XY} (dstr in the population)

- Is the causal direction between two variables identifiable?
- Is the causal mechanism as represented by a SEM identifiable?

Causal Discovery & Inference under Different Kinds of Selection Bias



(c)

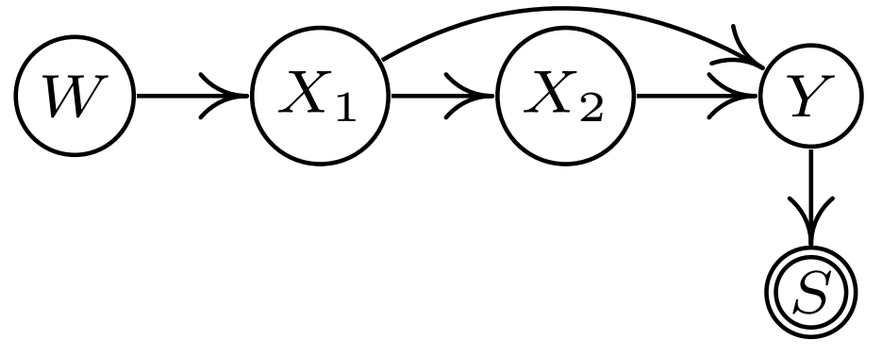
(c) Outcome-dependent selection bias (OSB):
$P_{Y|X,S=1} \neq P_{Y|X}$

Selected sample follows $P_{XY|S=1}$ instead of P_{XY} (dstr in the population)

- Is the causal direction between two variables identifiable?
- Is the causal mechanism as represented by a SEM identifiable?

Zhang, Zhang, Huang, Schölkopf, Glymour, On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection, Proc. UAI 2016, plenary talk

Effect of OSB

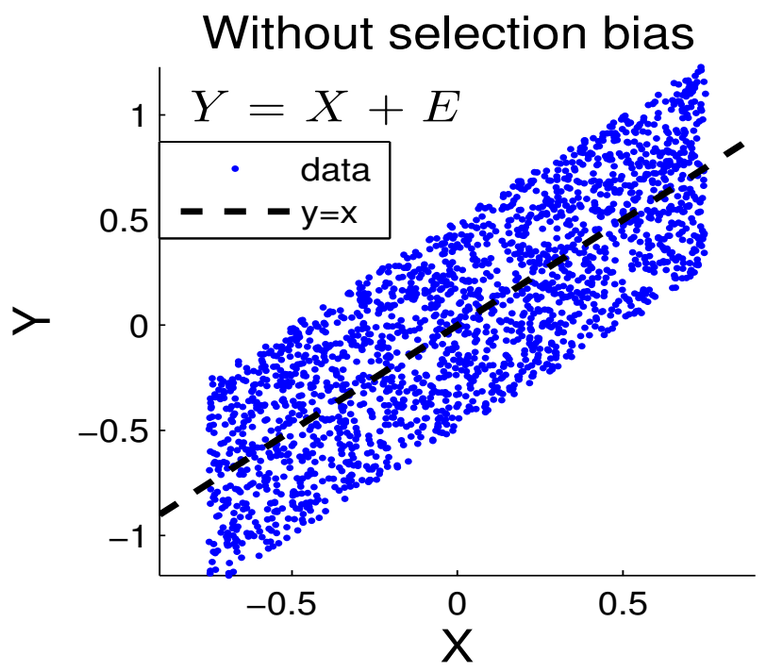


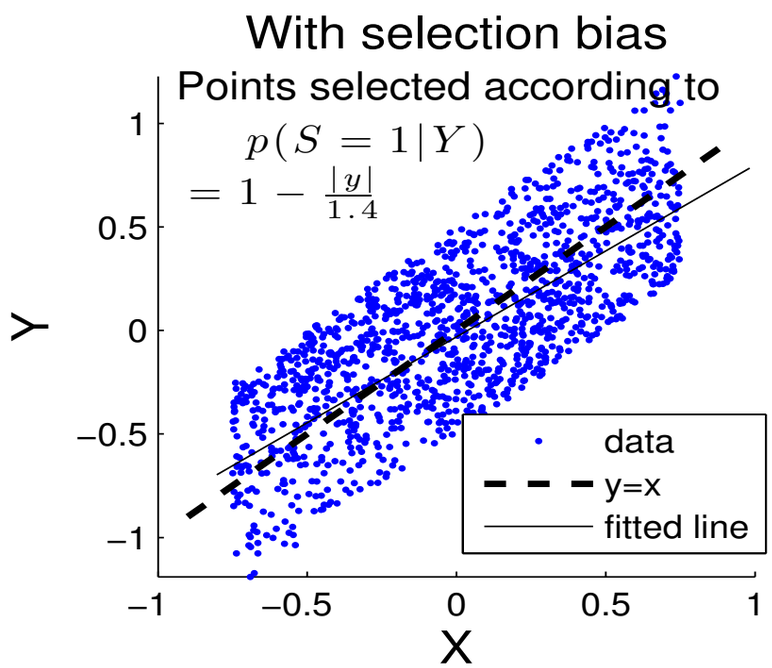
- The distribution of the observed sample is changed by the selection process

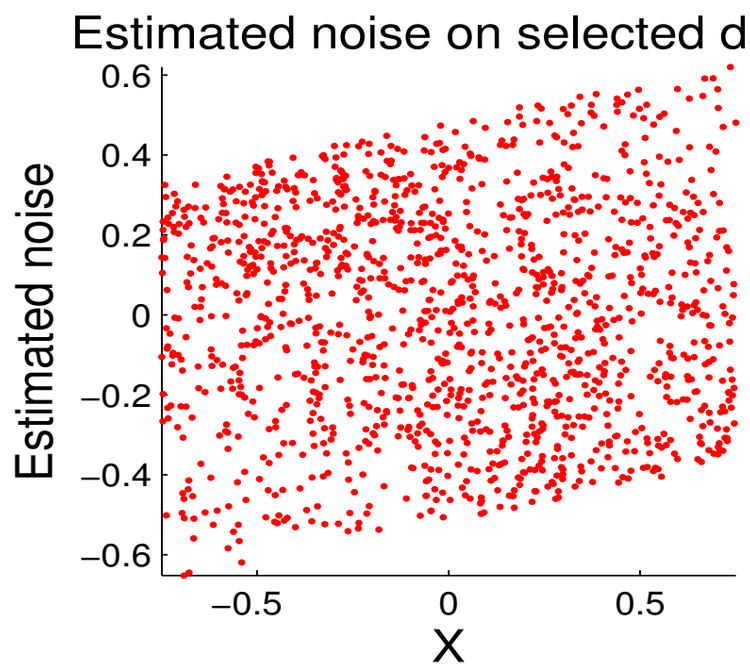
$$p^{\beta}_{XY} \triangleq p_{XY|S=1} = \frac{p_{X,Y,S=1}}{P(S=1)} = p_{XY} \cdot \frac{P(S=1|X,Y)}{P(S=1)}$$

$$= p_{XY} \cdot \frac{P(S=1|Y)}{P(S=1)} = \beta(y) p_{XY}$$

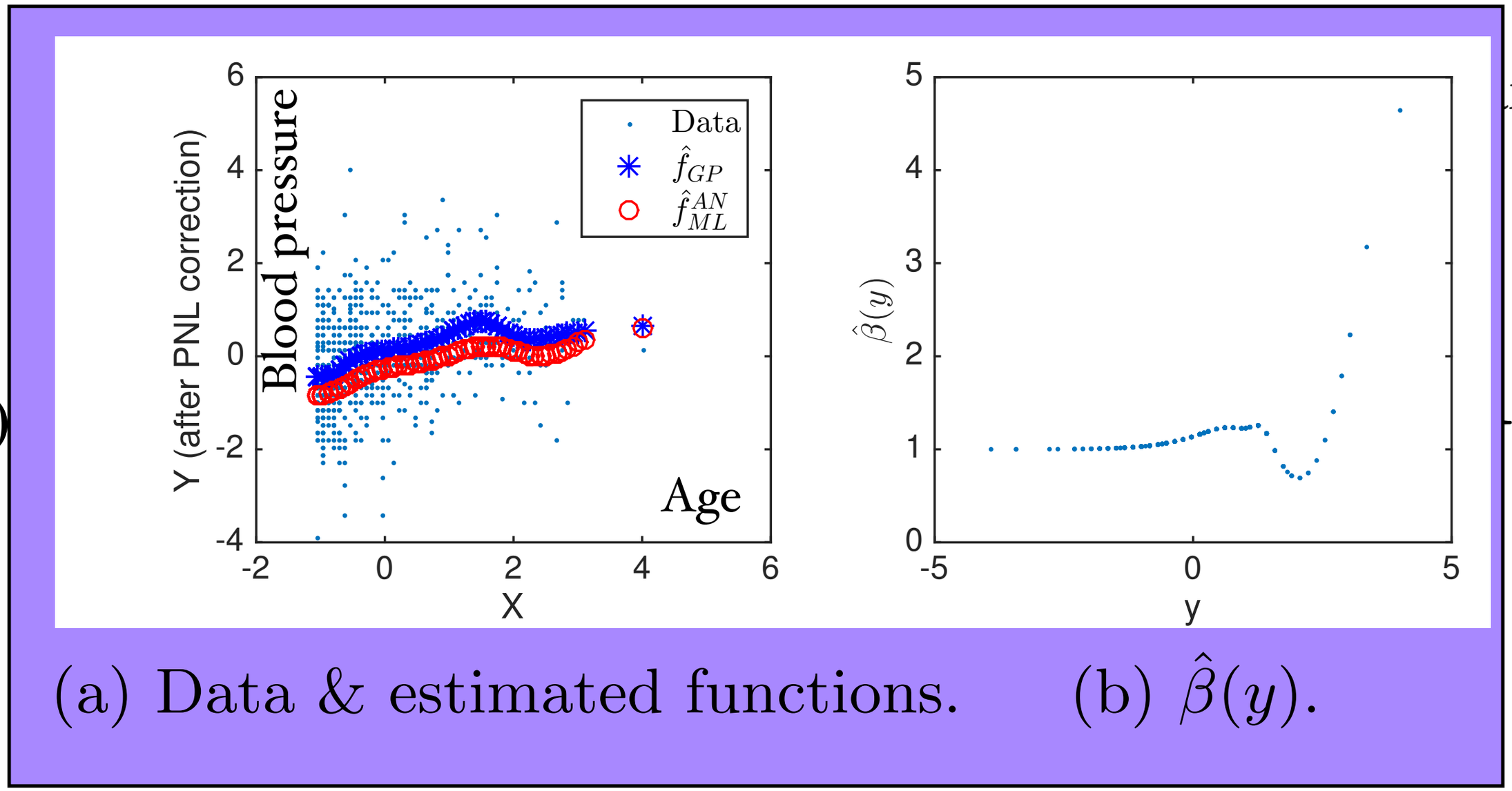
- Illustration: Error is not independent any more from cause







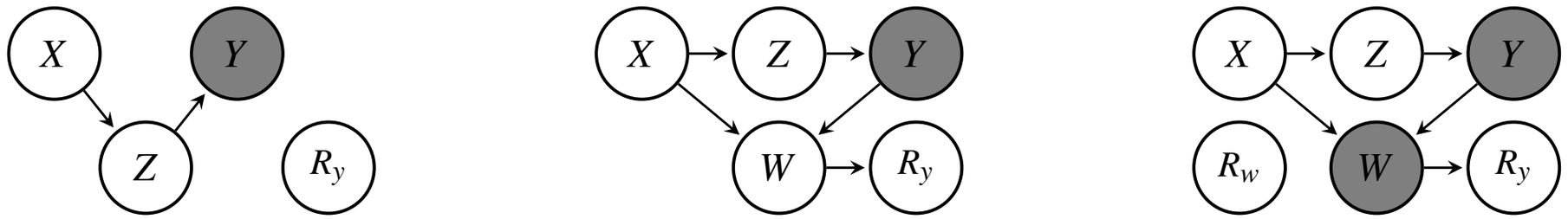
Causal Discovery and Inference under Output-Dependent Selection: An Example



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

Issue 4: Causal Discovery in the Presence of Missing Data

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
1.1349917e+00		3.5184872e-01		-5.6961669e-01	-4.8335698e-01	-9.3747444e-01	2.2763032e-03



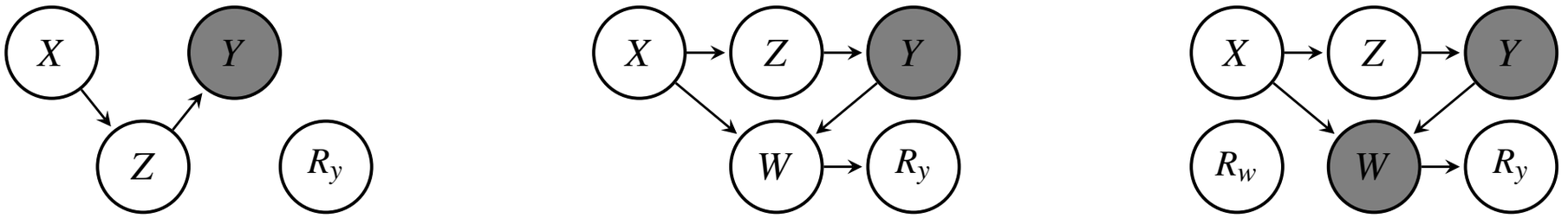
(a) An MCAR graph (b) An MAR graph (c) An MNAR graph

- Conditional independence relations in the data are sensitive to the missingness mechanism
- Key issue: Recover conditional independence relations in the original population from incomplete data

R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, C. Glymour, K. Zhang, "Causal discovery in the presence of missing data," AISTATS 2019

Causal Discovery in the Presence of Missing Data

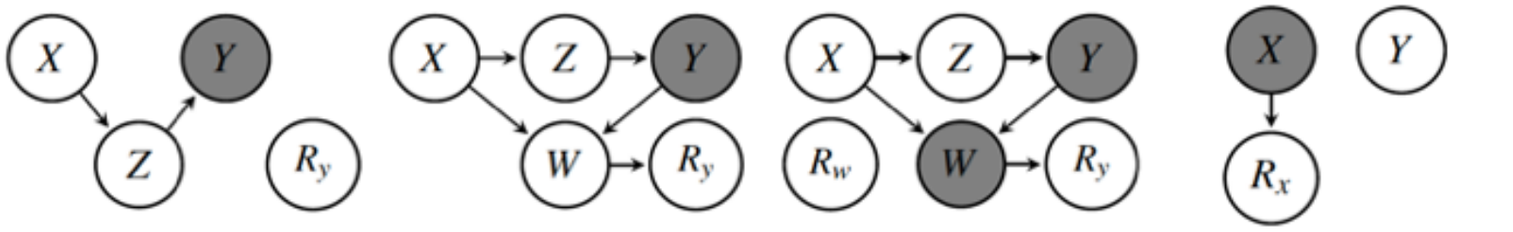
X1	X2	X3	X4	X5	X6
-9.4653403e-01		6.6703495e-01		8.2886922e-01	-1.3695521e+00	-3.2675465e-02	1.8634806e-01
-9.4895568e-01					-4.6381657e-01	-1.8280031e+00	
		5.1435422e-01		6.7338326e-01	4.3403559e-01	9.4535076e-01	7.5164028e-01
7.2489037e-01				5.1325341e-01	8.3567780e-01	2.9825903e-01	7.7796018e-02
				-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



(a) An MCAR graph (b) An MAR graph (c) An MNAR graph

- ***R*** is the set of missingness indicators that represent the status of missingness
- If R_X is 1, the corresponding value of X is missing; if it is 0, it is observed
- Missingness graph

Categories of Missing Data Mechanism



(a) A MCAR graph (b) A MAR graph (c) A MNAR graph (d) Self-masking missingness

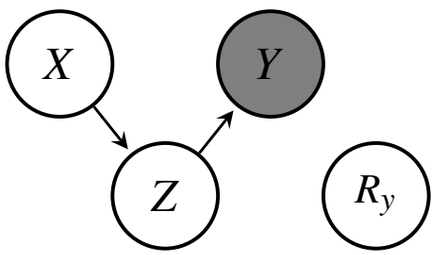
Figure 1: Exemplar missingness graphs in MCAR, MAR, MNAR, and self-masking missingness. X, Y, Z, and W are random variables. In missingness graphs, gray nodes are partially observed variables, and white nodes are fully observed variables. R_x, R_y, and R_w are the missingness indicators of X, Y, and W.

- All missing data mechanisms fall into one of the following three categories (Rubin, 1976):
  - Data are Missing Missing Completely At Random (MCAR) if the cause of missingness is purely random.
  - Data are Missing At Random (MAR) when the direct cause of missingness is fully observed.
  - Data that are neither MAR nor MCAR fall under the Missing Not At Random (MNAR) category.

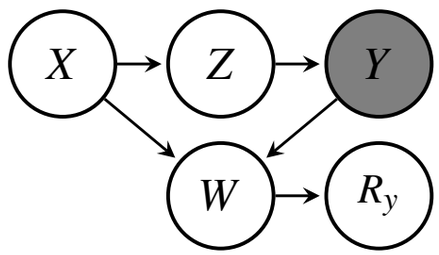
Assumptions for the Method

- Assumption 1 (Missingness indicators are not causes): No missingness indicator can be a cause of any substantive (observed) variable.
- Assumption 2 (Faithful observability): Any conditional independence relation in the observed data also holds in the unobserved data.
- Assumption 3 (No deterministic relation between missingness indicators): No missingness indicator can be a deterministic function of any other missingness indicators.
- Assumption 4 (No self-masking missingness): Self-masking missingness refers to missingness in a variable that is caused by itself.

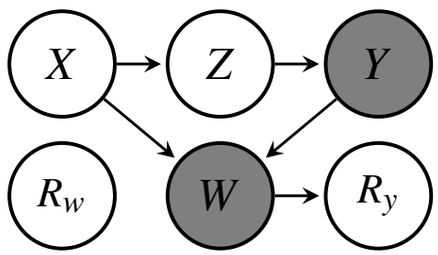
Observations



(a) An MCAR graph



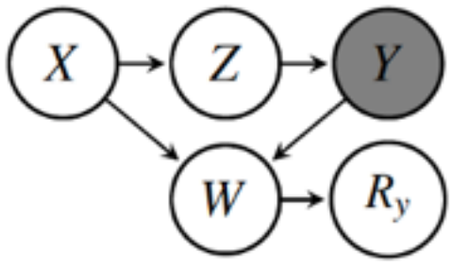
(b) An MAR graph



(c) An MNAR graph

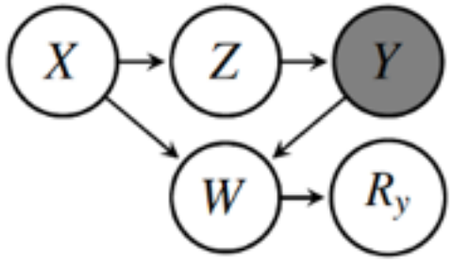
- Trust the testwise deletion conditional independence relations for causal discovery?
- Given Assumptions 1-4, we can prove:
  - If $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ in the testwise-deleted data, then $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ in the full data.
  - If testwise deletion gives extra dependence $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$, compared to the population, then for at least one variable in $\{X\} \cup \{Y\} \cup \mathbf{Z}$, its missingness indicator is either the direct common effect or a descendant of the direct common effect of X and Y.

Missing-Value PC (MVPC)



- Add missingness variables **R** to the dataset with measured variables **V**
- Create knowledge that **R** variables do not cause **V** variables
- Run PC adjacency search over **V**∪**R**
- Identify adjacencies over **V** in triangles over **V**∪**R**-–these might be false positives!
- Try to remove these extra adjacencies using *correction*…
- Finally, do collider orientation and apply the Meek rules to graph *G* over **V**

Essential Step in Missing Value PC



- Goal: see whether X⫫Y | Z by analyzing data with missing values
- Can we recover p(X,Y, Z) when Y has missing values?

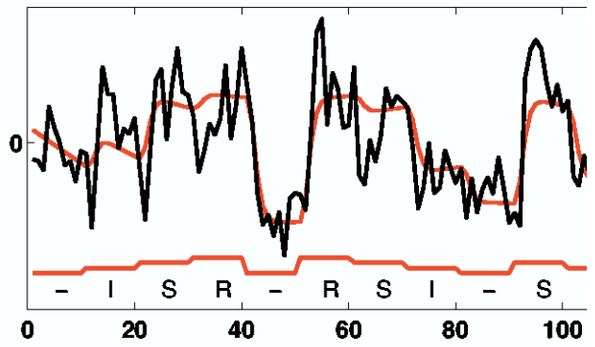
$$P(X,Y,Z) = \int_W P(X,Y,Z \mid W)P(W)dW$$
$$= \int_W P(X,Y^*,Z \mid W,R_y = 0)P(W)dW$$

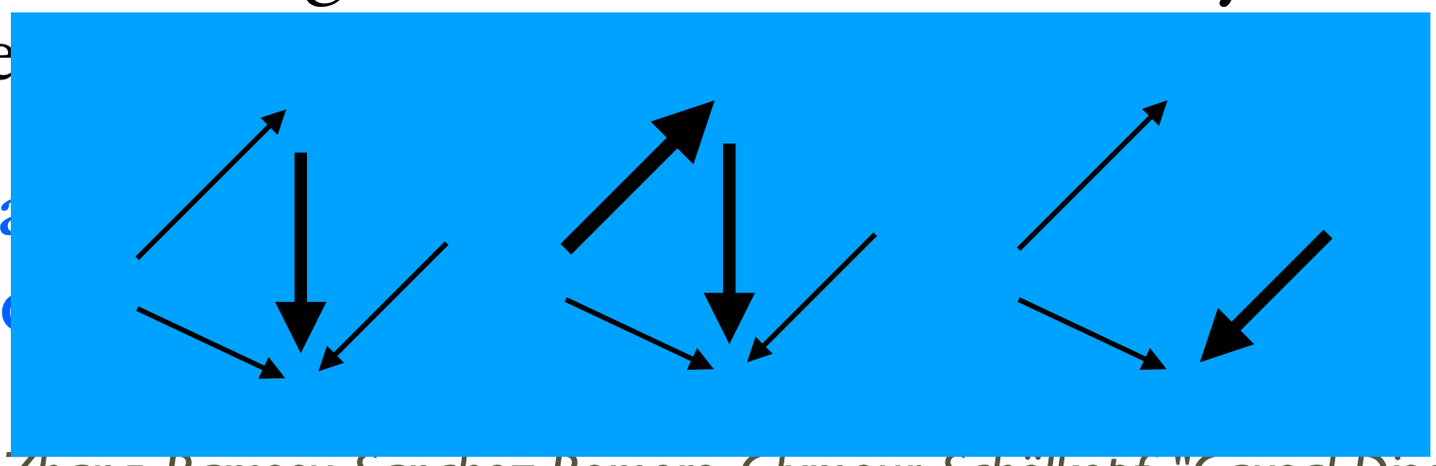
- In the linear-Gaussian or discrete case, permutation test:

$$\widehat{X} := \alpha_1 W^S + \varepsilon_1, \quad \widehat{Y} := \alpha_2 W^S + \varepsilon_2, \quad \widehat{Z} := \alpha_3 W^S + \varepsilon_3,$$

Issue 5: Nonstationary/Heterogeneous Data and Causality

- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data…)
  - Multiple data sets under different observational or experimental conditions
- Causal modeling & distribution shift heavily couple
  - $P(c$ inde



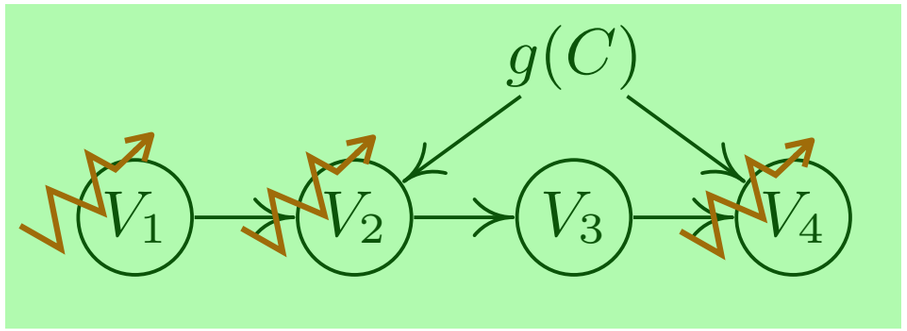


Huang, Zhang, Zhang, Ramsey, Sanchez-Romero, Glymour, Schölkopf, "Causal Discovery from Heterogeneous/ Nonstationary Data," JMLR, 2020

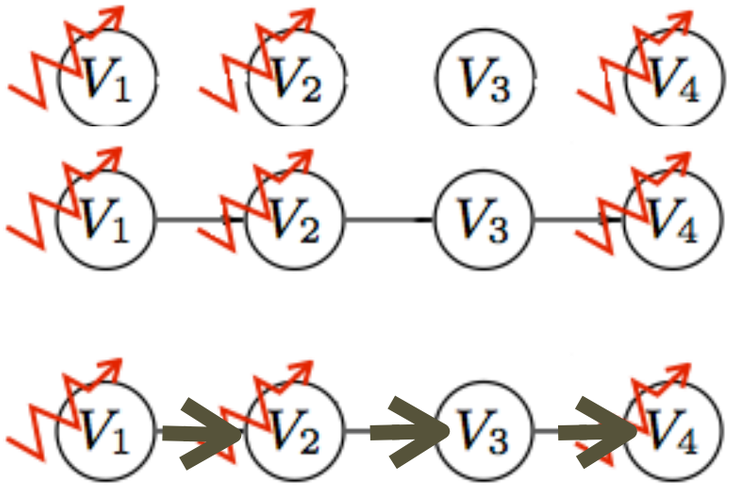
Zhang, Huang, et al., Discovery and visualization of nonstationary causal models, arxiv 2015

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

Causal Discovery from Nonstationary/ Heterogeneous Data

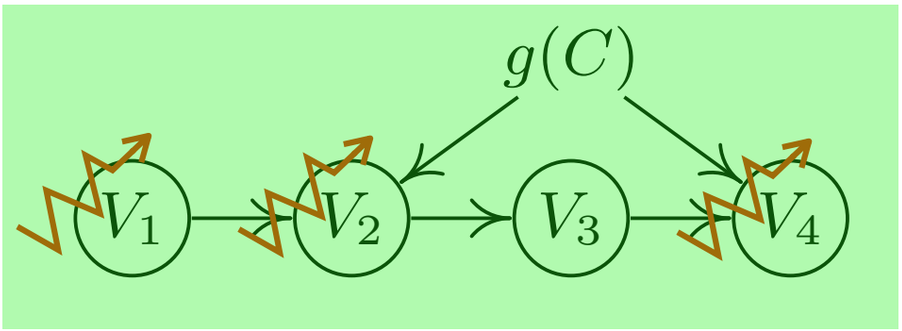


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
  - How do the nonstationary modules change over time / across data sets?



Kernel nonstationary driving force estimation

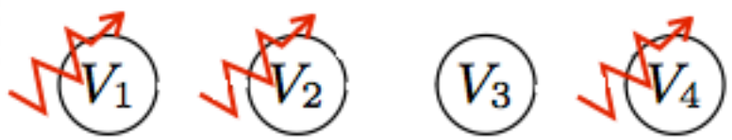
Discovery & Visualization of Changing Causal Modules

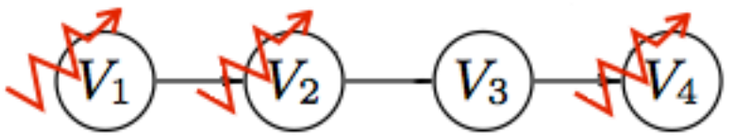


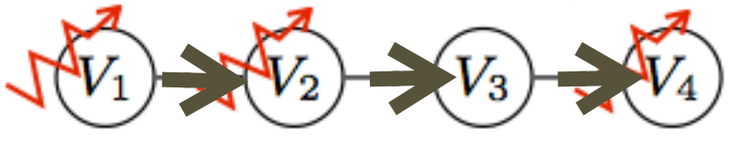
* Questions to answer for causal discovery:

- Identify variables with changing causal modules & recover causal skeleton?
- Identify causal directions by using distribution shifts?
- Visualize the change in causal modules?

With our proposed approach:



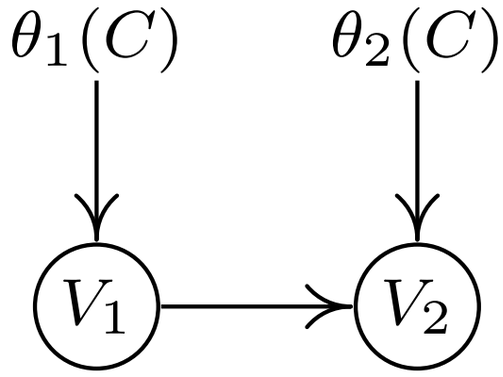




Kernel nonstationarity visualization (KNV)

- Incorporate time/domain index C as a surrogate + apply constraint-based causal discovery methods
- Independent changes in P(cause) and P(effect | cause)
- Find a mapping of $P(V_i \mid PA^i)$ to capture its variability

Nonstationarity Helps Determine Causal Direction



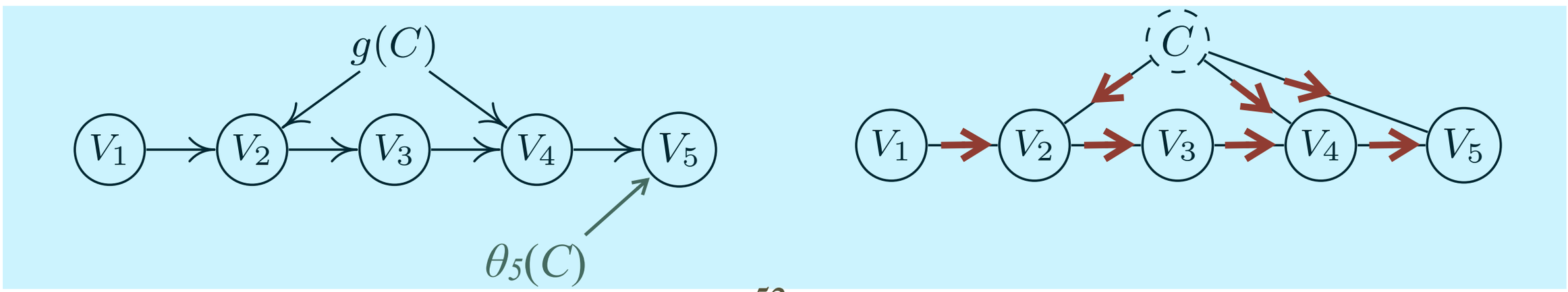
- **Independent changes** in *P*(cause) and *P*(effect | cause): generalization of invariance; generally violated for wrong directions

- Special cases: if $C - V_k - V_l$, since $C \to V_k$, we known

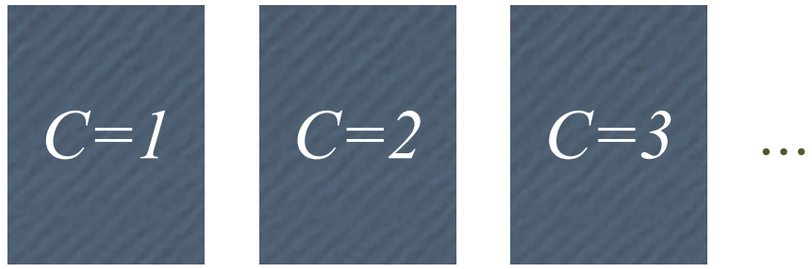
- $C \to V_k \leftarrow V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **excluding** V_k Invariant cause

- $C \to V_k \to V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **including** V_k Invariant mechanism

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



Kernel Nonstationarity Visualization



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

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

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

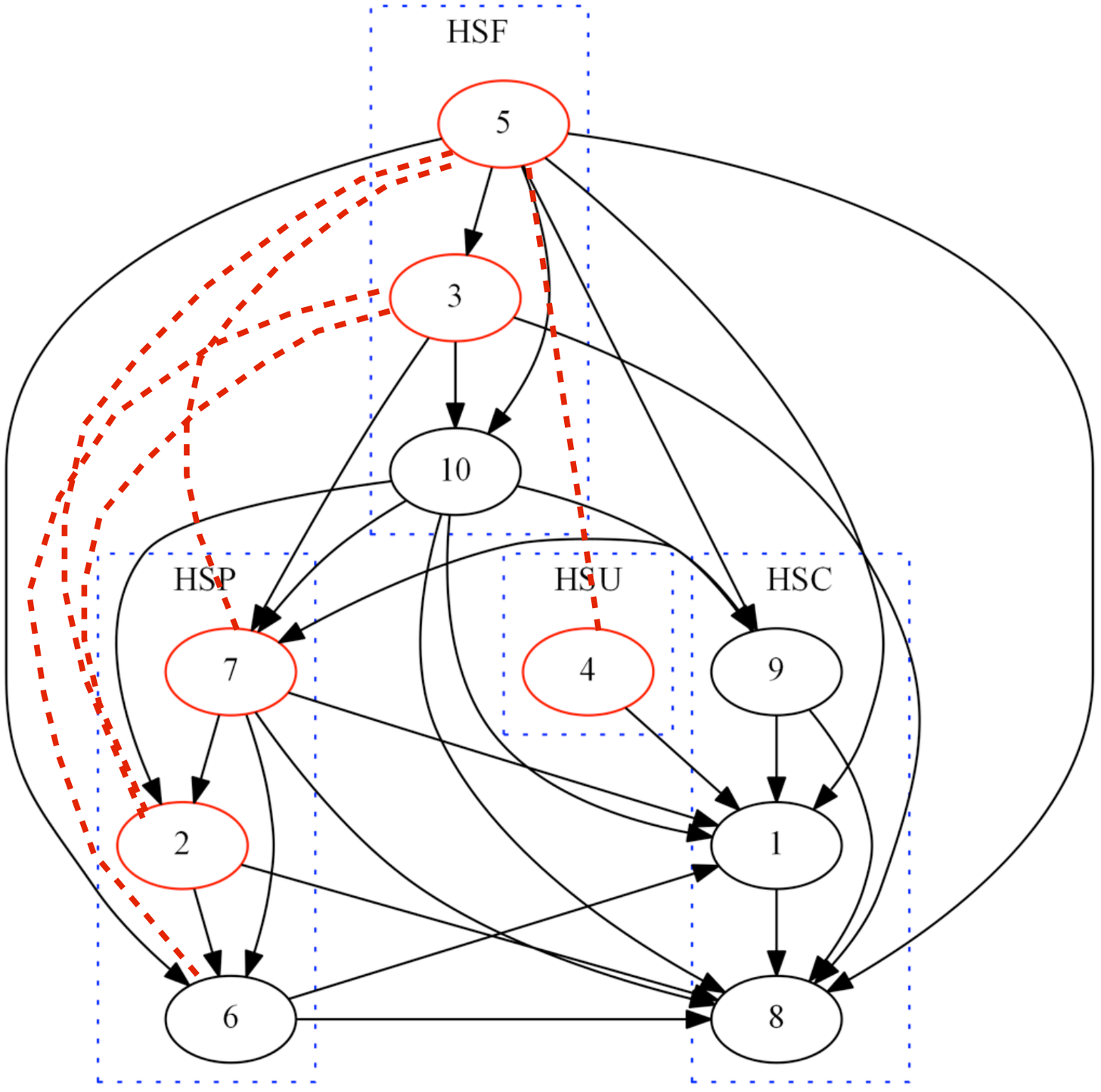
- Kernel nonstationarity visualization (KNV):

  - 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)

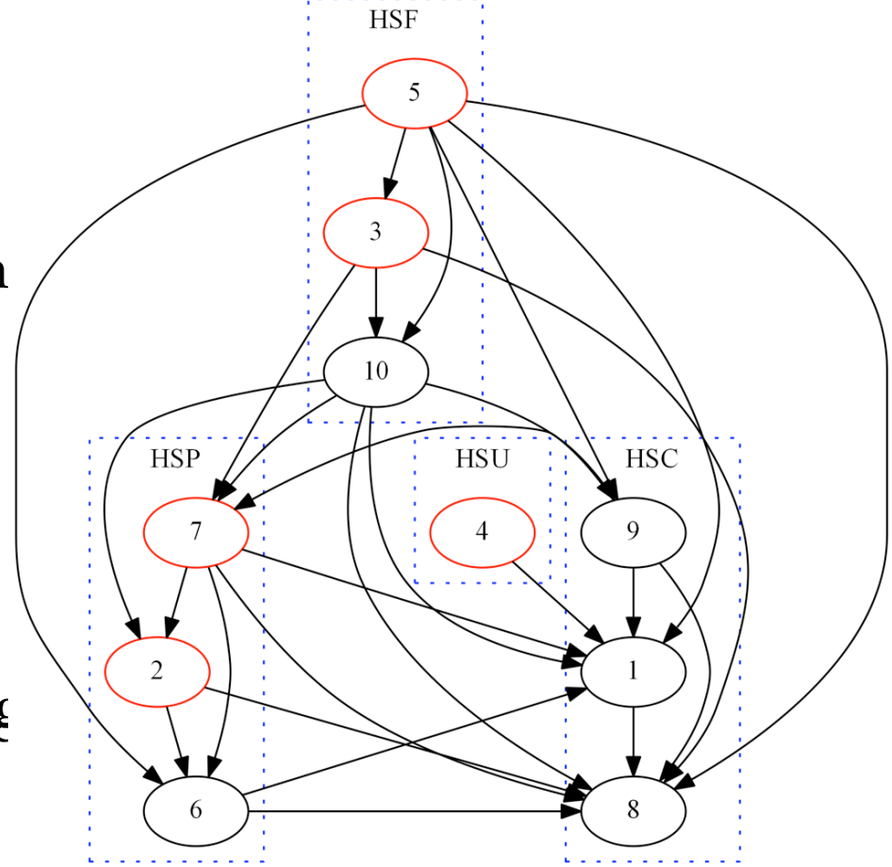
1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4.Hong Kong Electric Holdings,
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong

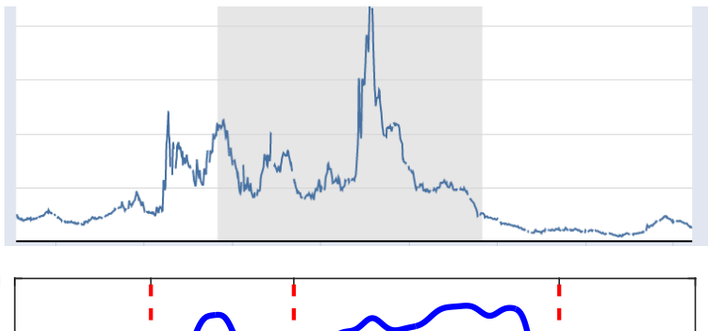


- HSF and HSP usually have nonstationary confounders

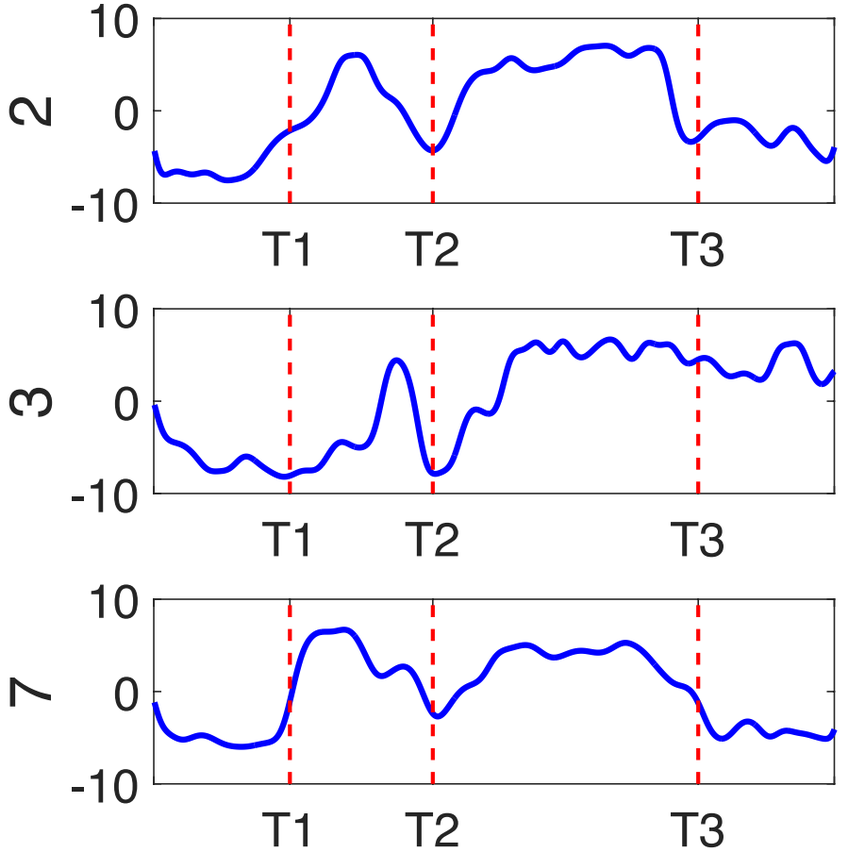
Nonstationarity Driving Force

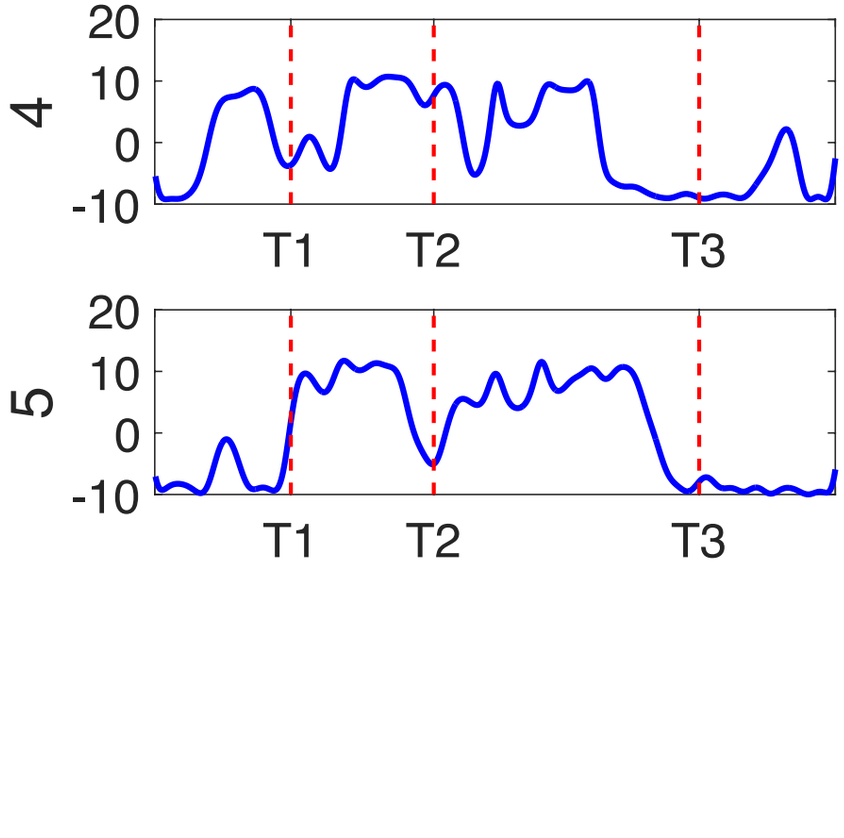
1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4. Hong Kong Electric Holdin
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong





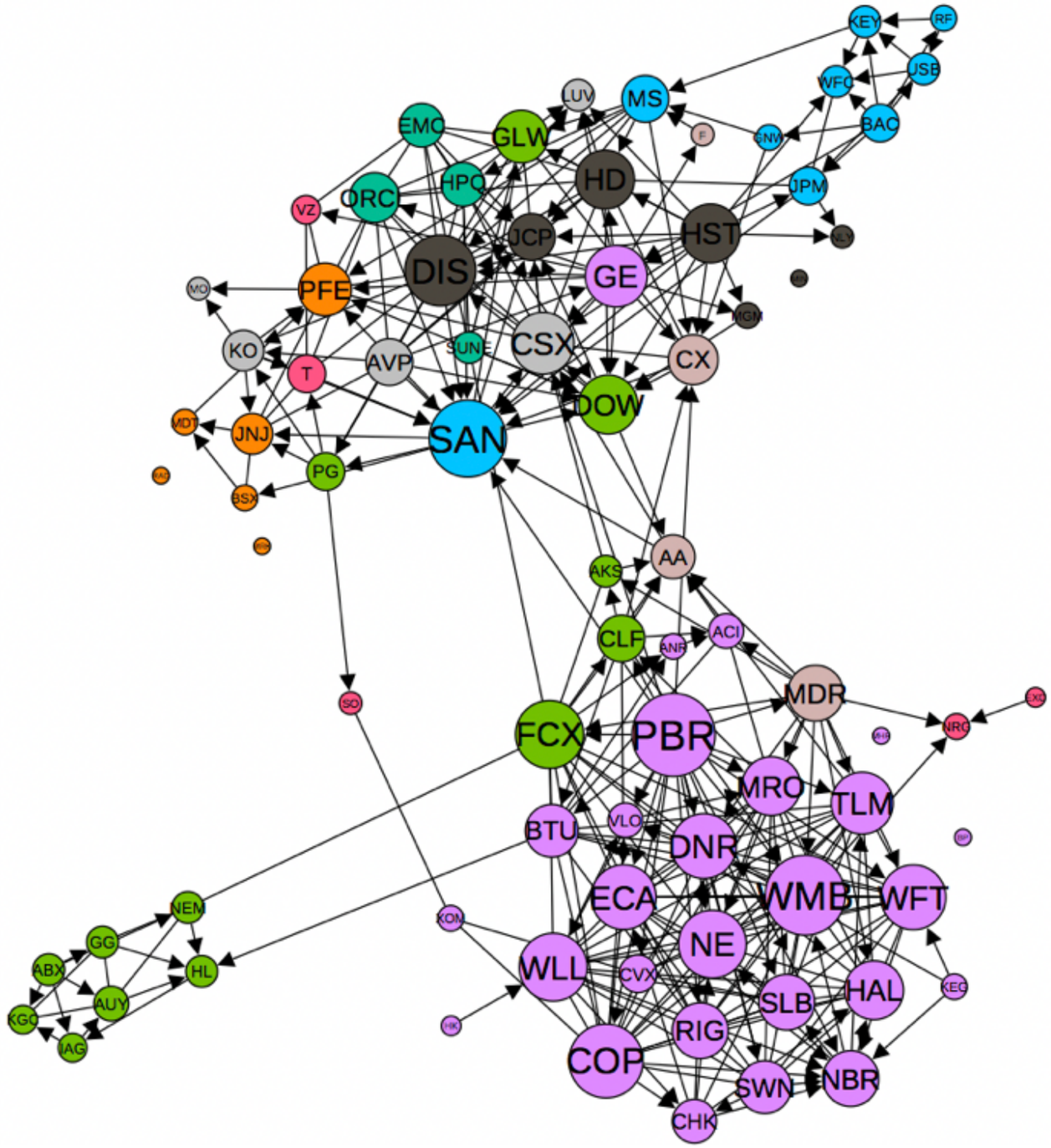
(Curve of TED spread; https://research.stlouisfed.org/fred2/series/TEDRATE)

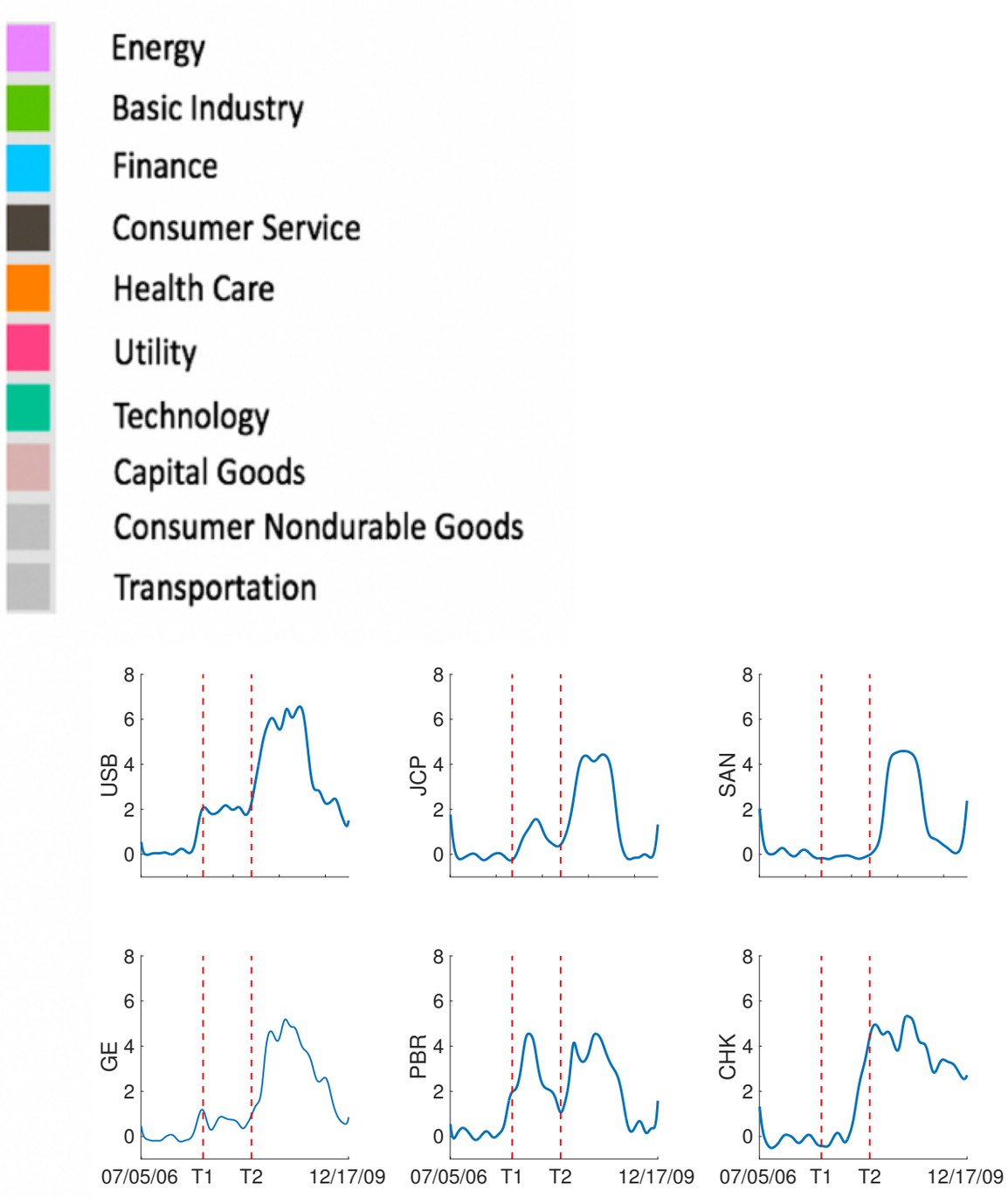




T_1: 07/16/2007,
T_2: 06/30/2008,
T_3: 02/11/2009

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: Practical Issues in Causal Discovery

- Latent confounders, cycles, nonlinearities (and even mixed data types), measurement error, selection bias, missing values, nonstationarity…
- Don’t worry—look into the problems
- Learning latent confounders and their relations!