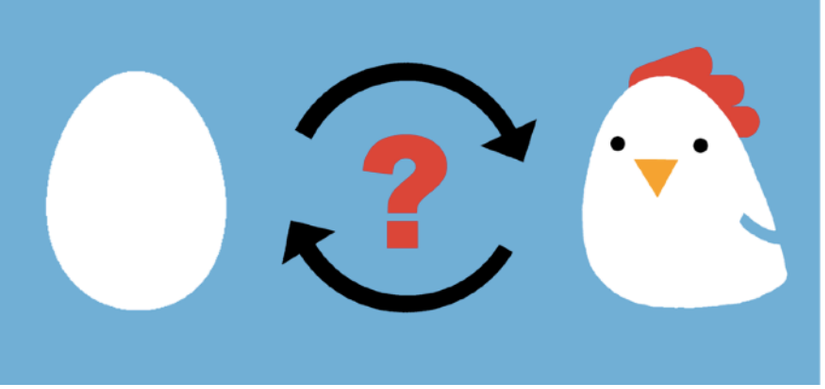


Causality and Machine Learning
(80-816/516)

Classes 16 (March 13, 2025)

Practical Issues in Causal Discovery:
Selection Bias, measurement error, and missing values

Instructor:

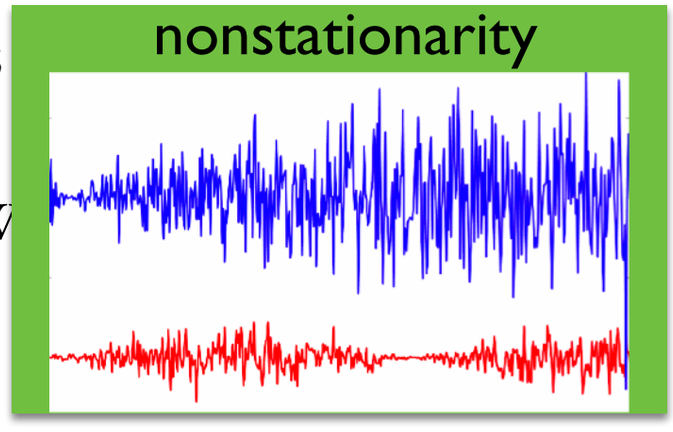
Kun Zhang (kunz1@cmu.edu)

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 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)

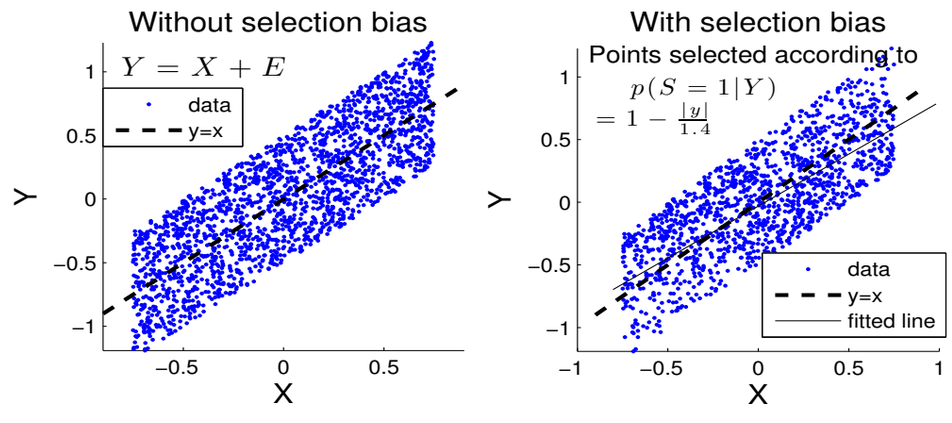


- **Nonstationary/heterogeneous data** (Zhang et al., IJCAI'17; Huang et al, ICDM'17, Ghassami et al., NIPS'18; Huang et al., ICML'19 & NIPS'19; Huang et al., JMLR'20)

Issue 1: 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

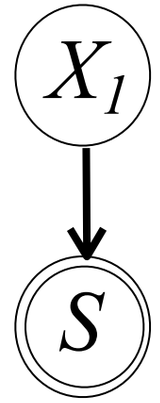


- In causal inference, both learning causal structures and estimating causal mechanisms become more difficult

Selection Bias: Illustration

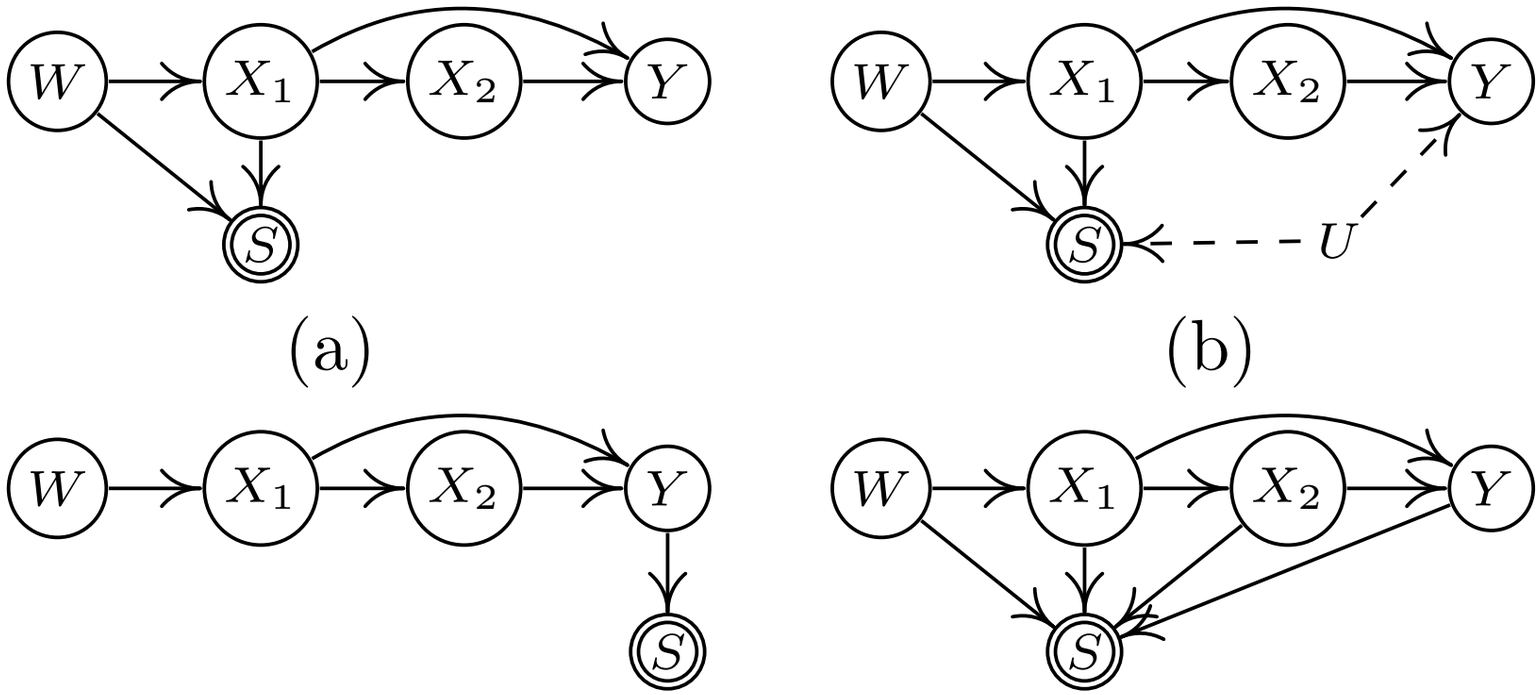
- Suppose the true causal process is

- 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



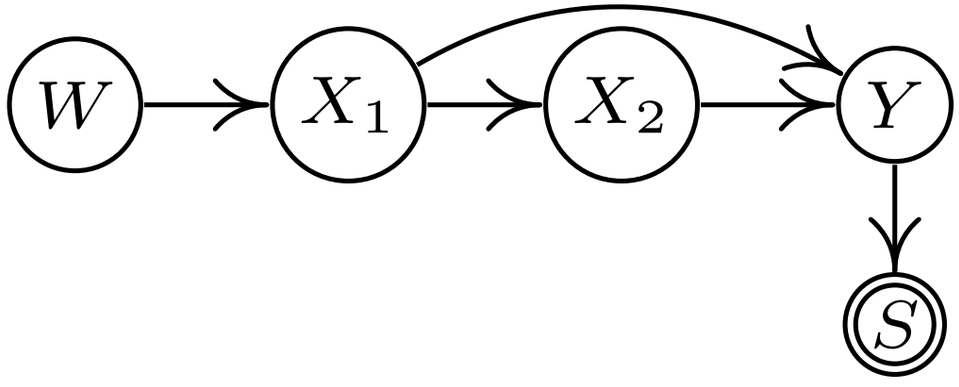
(a) (b)

(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)

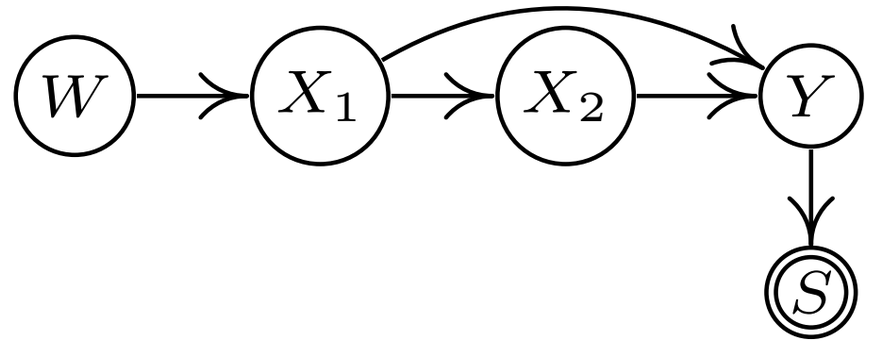
(c) Outcome-dependent selection bias (OSB):
$P_{Y|X,S=1} \neq P_{Y|X}$

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

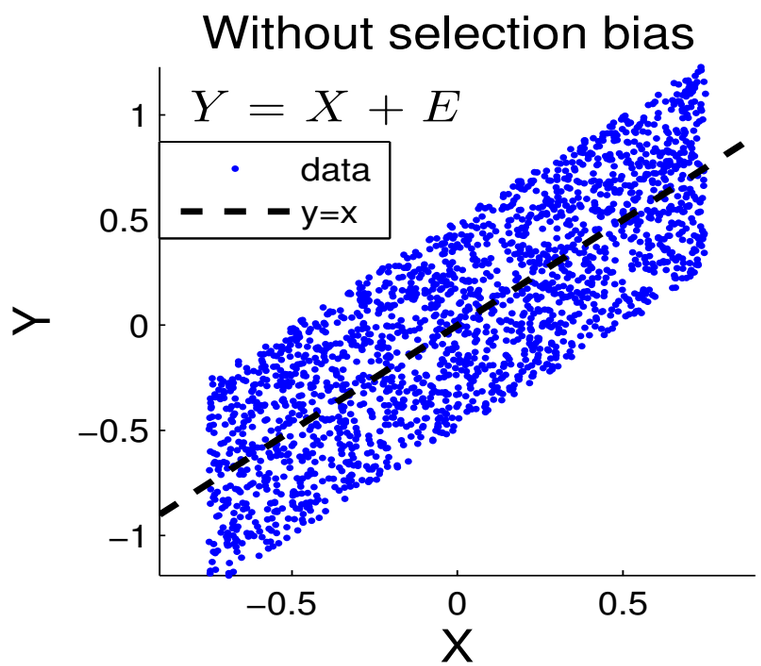


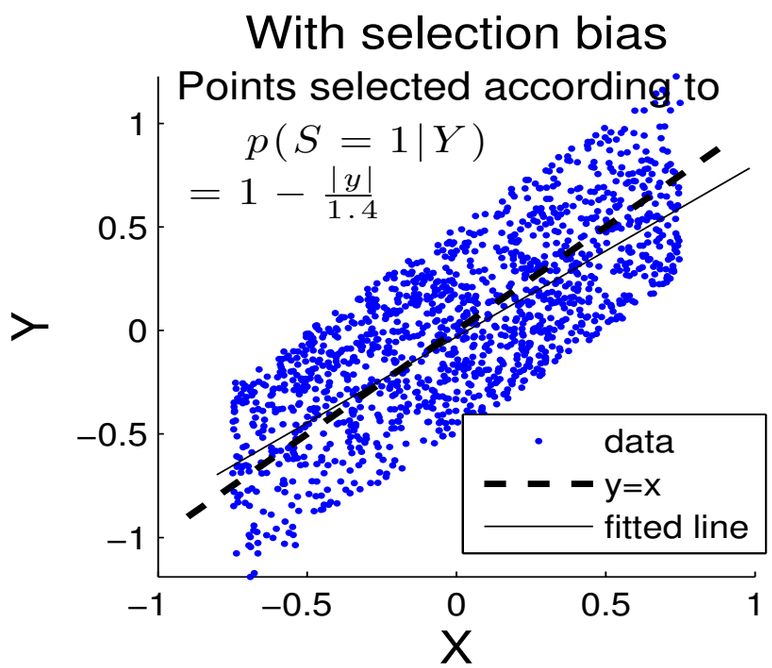
- The distribution of the observed sample is changed by the selection process

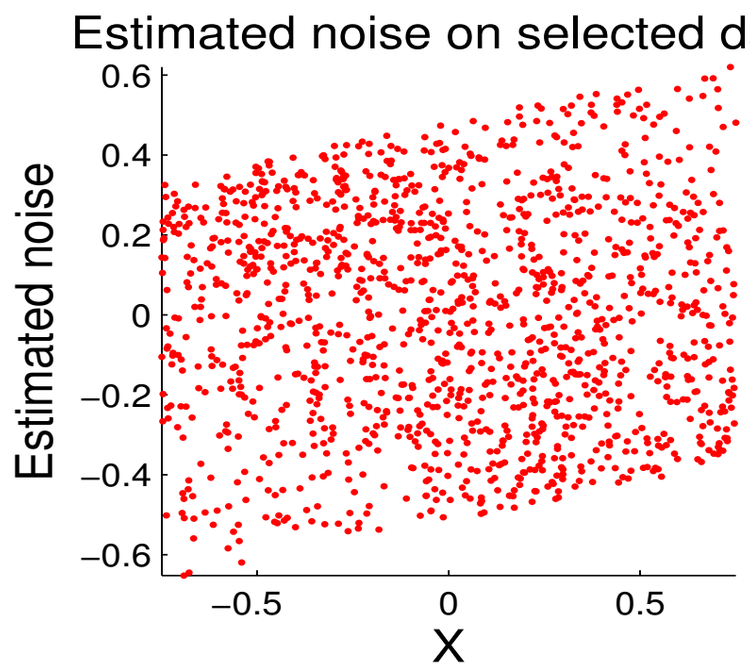
$$p^{\beta}_{XY} \triangleq p_{XY|S=1} = \frac{p_{X,Y,S=1}}{P(S=1)} = p_{XY} \cdot \frac{P(S=1|X,Y)}{P(S=1)}$$

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

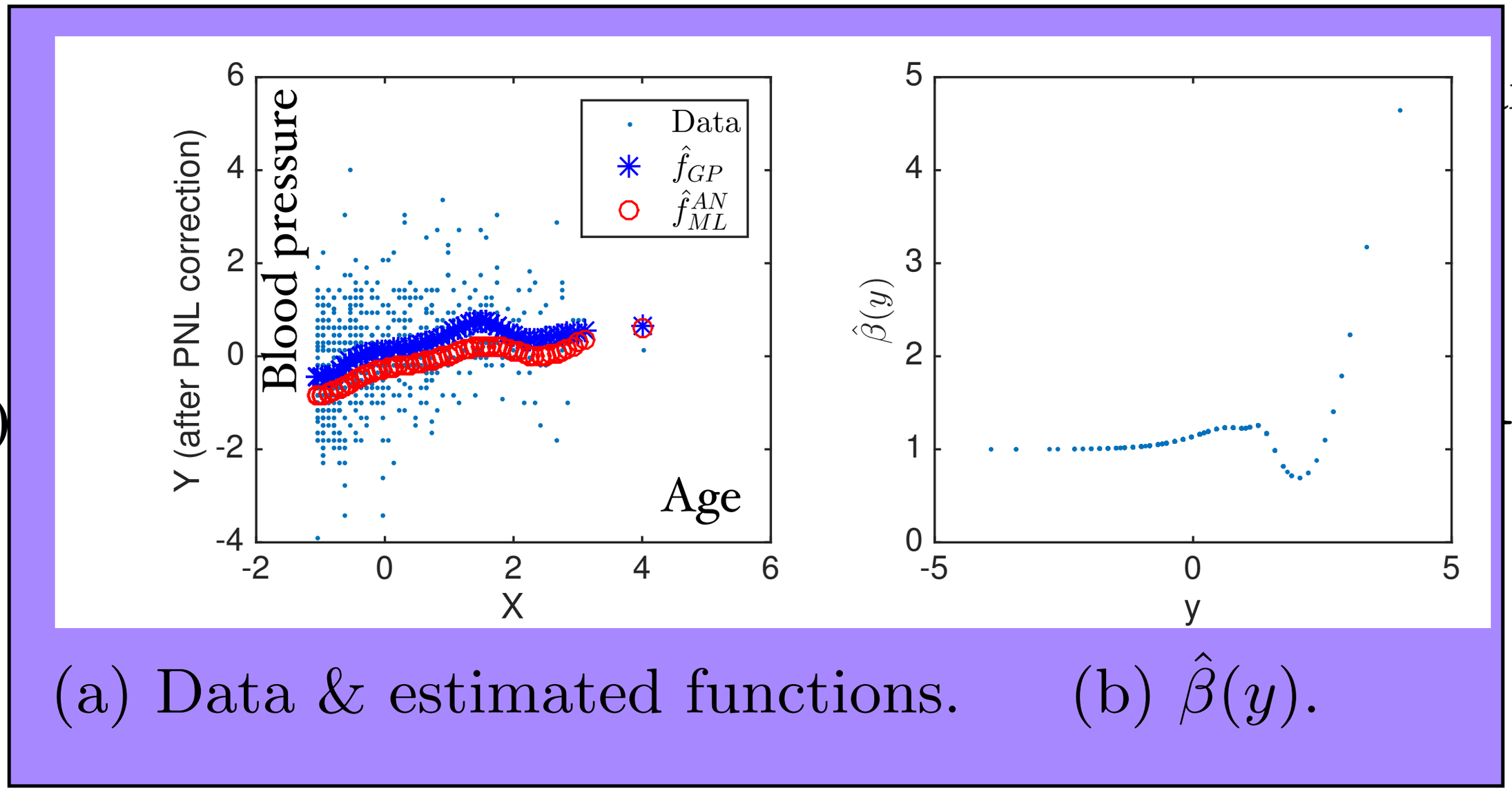
- Illustration: Error is not independent any more from cause







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



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

Autoregresstive Generation in Music, Text, etc.?

Detecting and Identifying Selection Structure in Sequential Data

Yujia Zheng [1] **Zeyu Tang** [1] **Yiwen Qiu** [1] **Bernhard Schölkopf** [2] **Kun Zhang** [1 3]

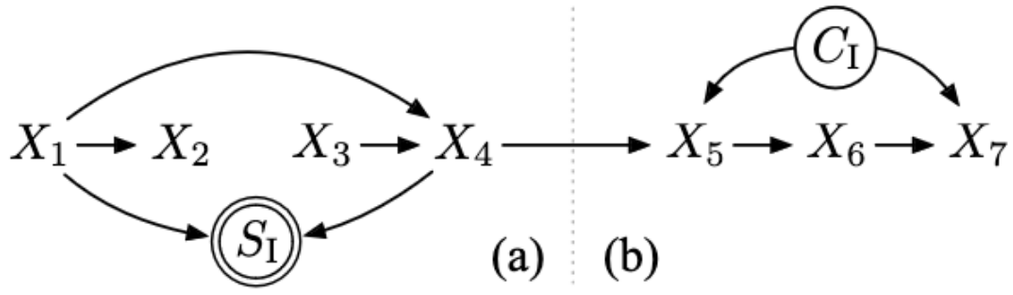
Abstract

We argue that the selective inclusion of data points based on latent objectives is common in practical situations, such as music sequences. Since this selection process often distorts statistical analysis, previous work primarily views it as a bias to be corrected and proposes various methods to mitigate its effect. However, while controlling this bias is crucial, selection also offers an opportunity to provide a deeper insight into the hidden generation process, as it is a fundamental mechanism underlying what we observe. In particular, overlooking selection in sequential data can lead to an incomplete or overcomplicated inductive bias in modeling, such as assuming a universal autoregressive structure for all dependencies. Therefore, rather than merely viewing it as a bias, we explore the causal structure of selection in sequential data to delve deeper into the complete causal process.

generating process in various applications. For instance, in composing music, composers are guided by specific artistic goals or themes, leading them to selectively choose certain patterns of musical combinations (as combinations of basic elements) from their mind, thereby introducing dependencies among the basic elements in the music sequences (Schoenberg et al., 1967). These intentional but unmeasured selections, together with the contextual information, shape the structure of the compositions. A comprehensive understanding of the selection structure is essential for uncovering the underlying causal process and making use of it.

In sequential data, the understanding of selection plays a vital role. One essential question is whether selection leaves unique data dependence patterns that cannot be well explained by direct causal relations or latent confounding. Interestingly, as we will see in this paper, the answer is yes. Consequently, overlooking selection in such data can result in the introduction of incomplete or overcomplicated dependence models for the data. For instance, due to the

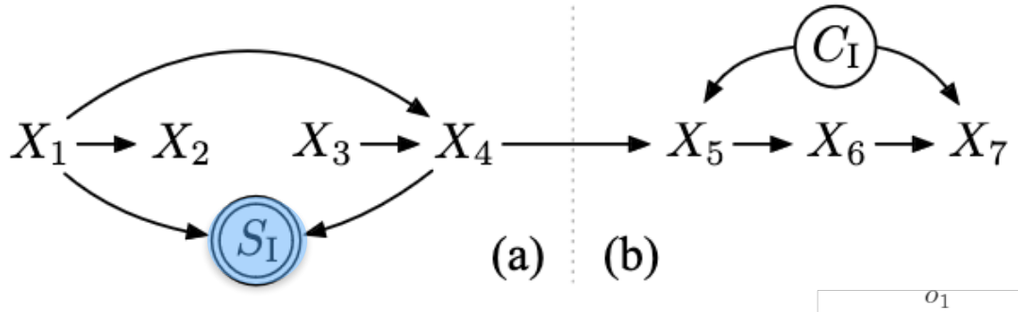
Selection in Sequential Data Identifiable

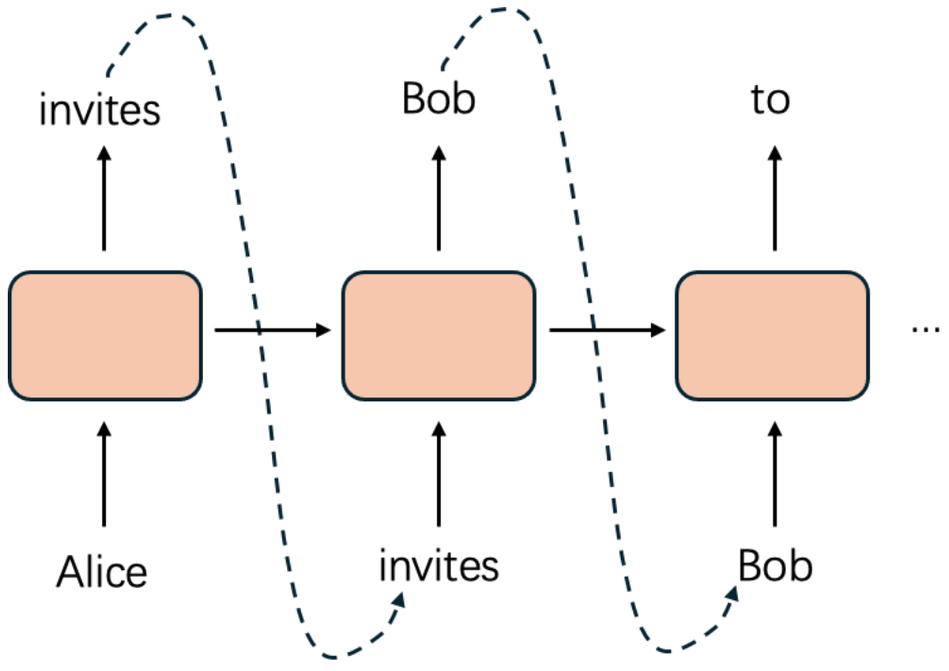


Sequential data with selection and confounders

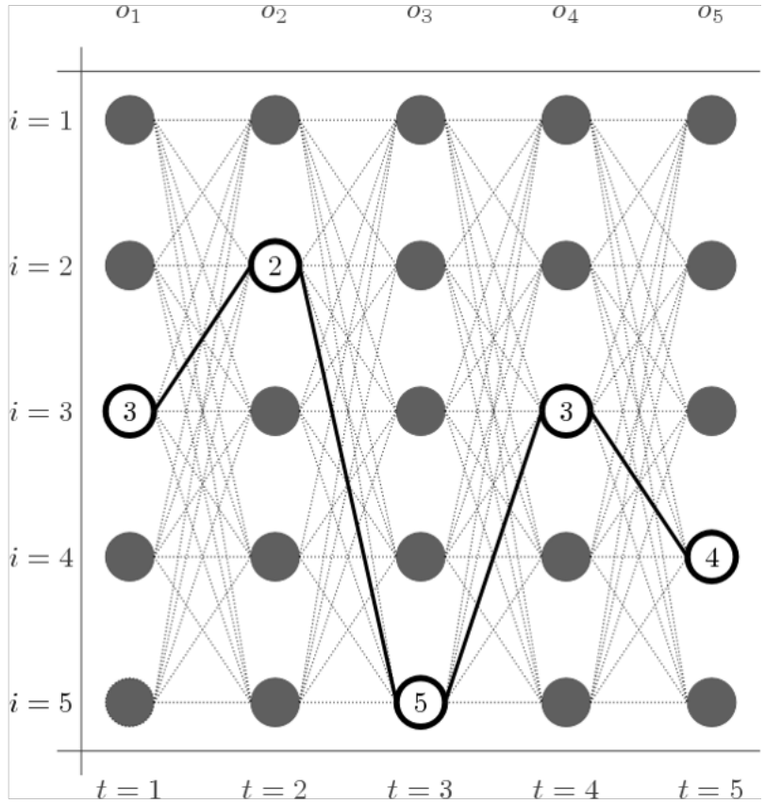
- Theorem: Let the observed data be a large enough sample generated by the above model. Under the faithfulness assumption, Markov condition, and two more technical conditions, all **selection pairs, direct relations, and confounded pairs** in the causal graph are identifiable
- Selection patterns in music verified

How should We Generate Text?





Current LLMs generate text by sampling the next token only.



However, as indicated by the Viterbi algorithm (to compute the most probable path of the hidden variables), the optimal sequence can't be discovered by GREEDILY adding best local solution to the final sequence.

Proposal: Looking forward and backward for Generation

- Two stages of text generation: text generation in an auto-regressive way + detector & rewriter

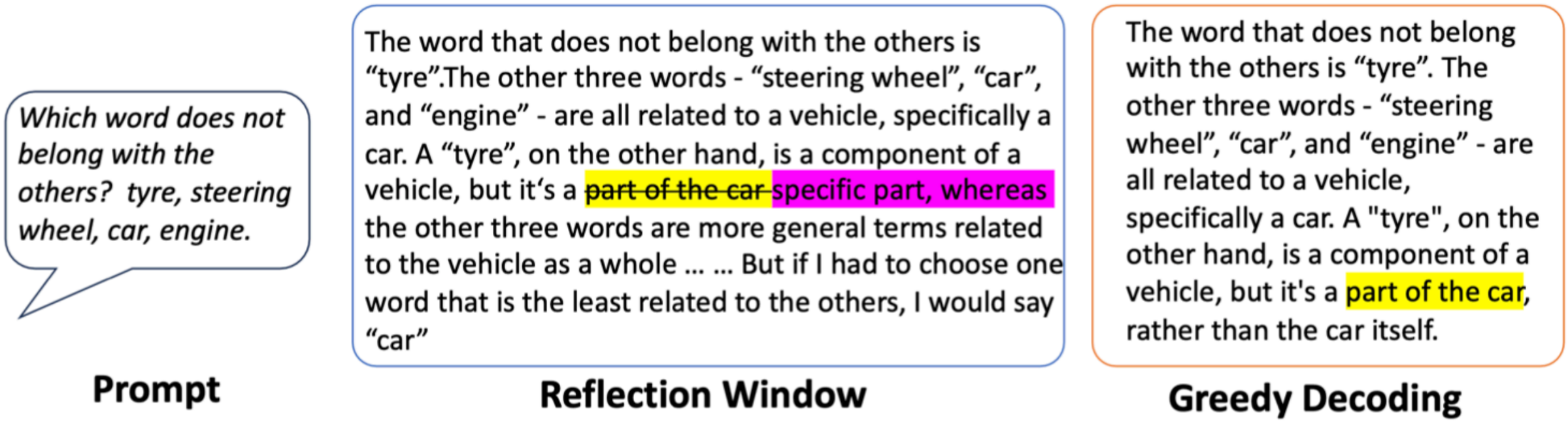


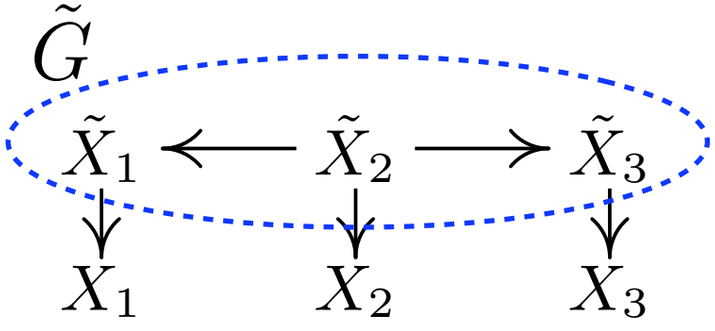
Figure 3: A concrete example demonstrating our propose approach

Issue 2: Causal Discovery in the Presence of Measurement Error

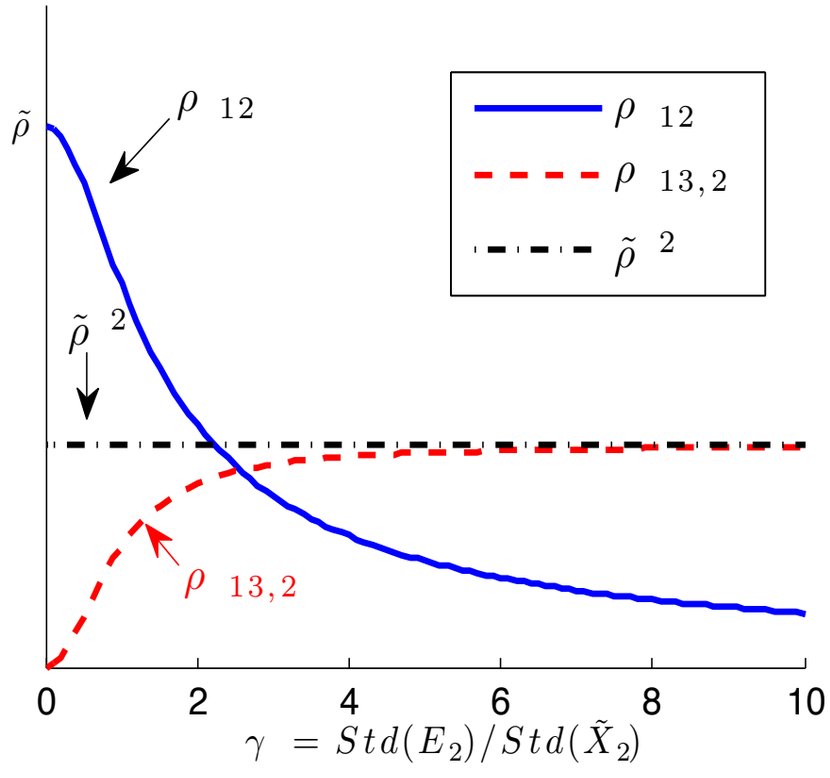
- To estimate $\tilde{G}$ over variables $\tilde{X}_i$ from noisy observations $X_i = \tilde{X}_i + E_i$.
- Conditional independence/dependence relations among X_i different from those among $\tilde{X}_i$

Measurement error changes causal discovery results!

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



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

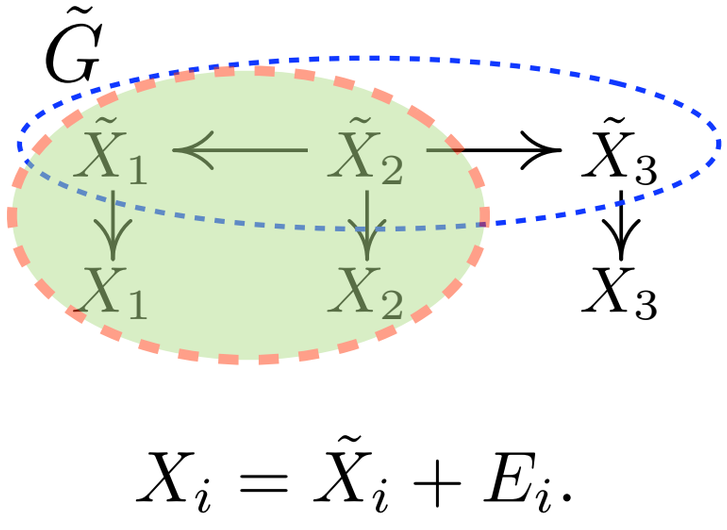


Causal Discovery in the Presence of Measurement Error

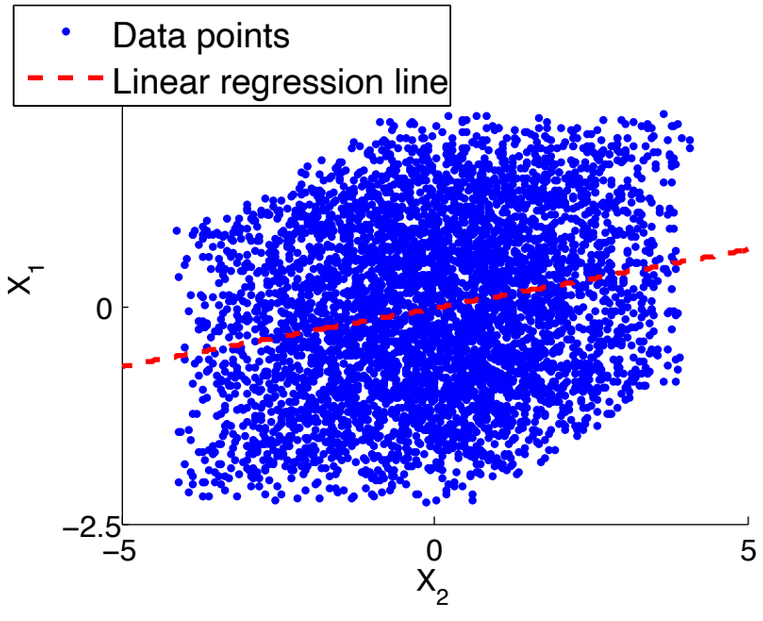
- To estimate $\tilde{G}$ over variables $\tilde{X}_i$ from noisy observations $X_i = \tilde{X}_i + E_i$.
- Conditional independence/dependence relations among X_i different from those among $\tilde{X}_i$

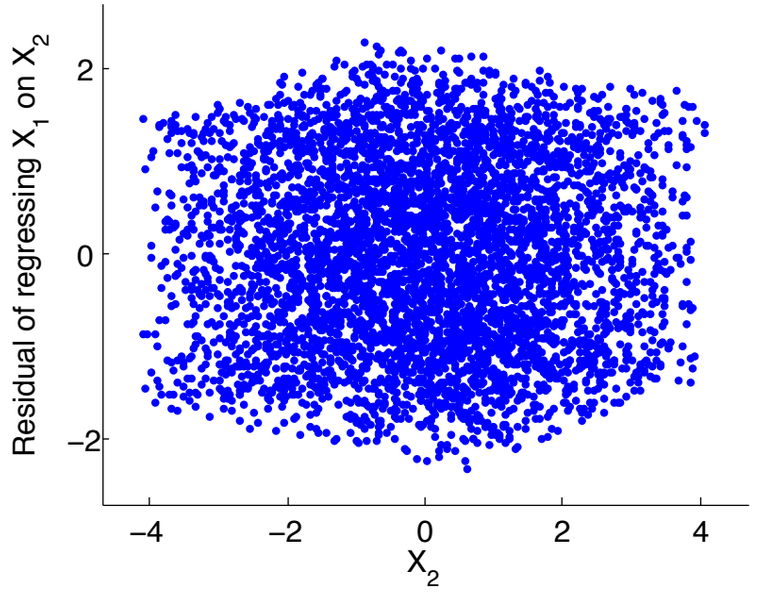
Measurement error changes causal discovery results!

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

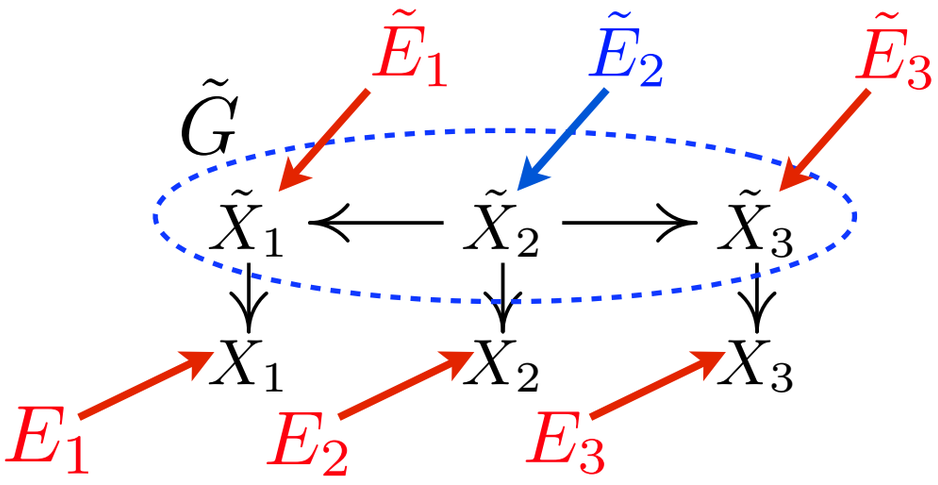


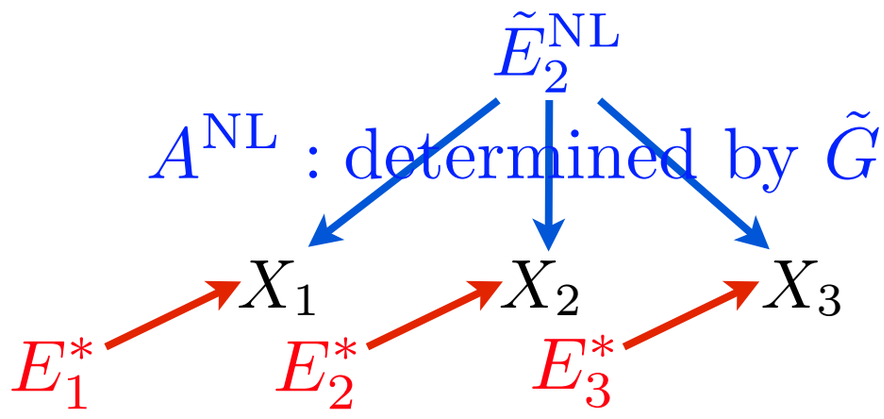
$$X_i = \tilde{X}_i + E_i.$$





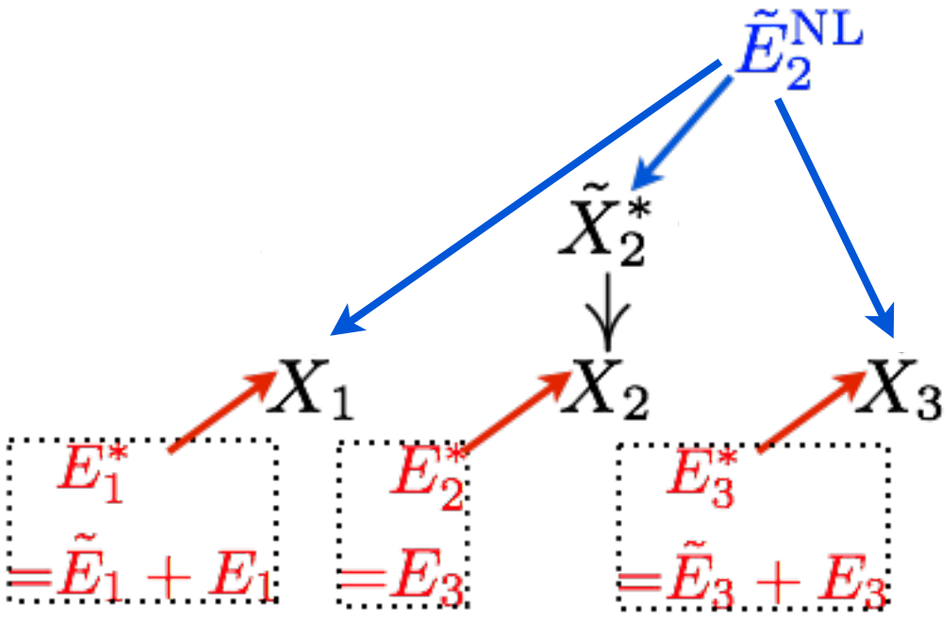
Canonical Representation of CAMME





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

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



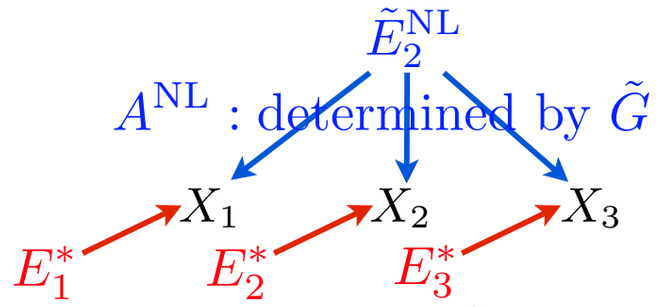
Alternatively: $\mathbf{X} = \left[\ \mathbf{A}^{\mathrm{NL}} \mid \mathbf{I}\ \right] \cdot \left[\begin{array}{c} \tilde{\mathbf{E}}^{\mathrm{NL}} \\ \hline \mathbf{E}^* \end{array}\right]$

- Two issues:
  - Is CR-CAMME identifiable from data?
  - Is $\tilde{G}$ identifiable from the CR-CAMME?

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

*

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 $\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ a & 0 & b \\ 0 & 0 & 0 \end{bmatrix}$.

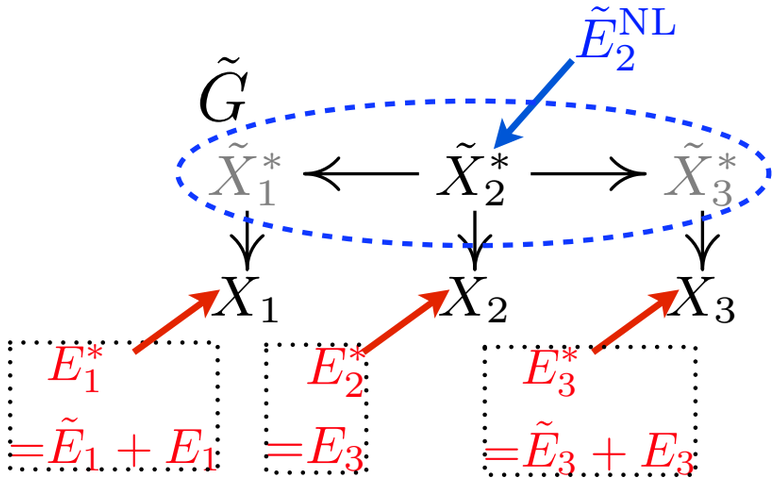
That is, $\tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{E}}$, with $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ a & 1 & b \\ 0 & 0 & 1 \end{bmatrix}$.

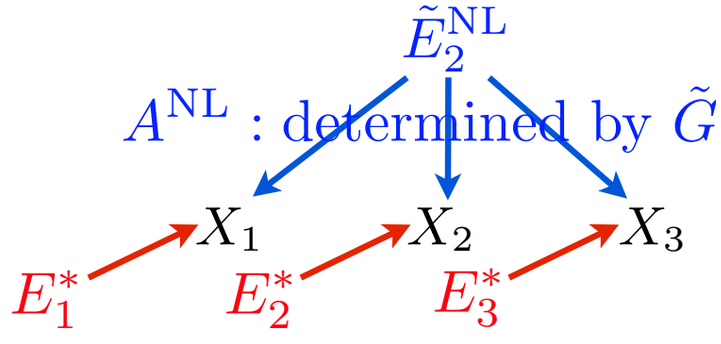
Therefore,

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

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

* Identifiability of CR-CAMME: Second-Order Statistics

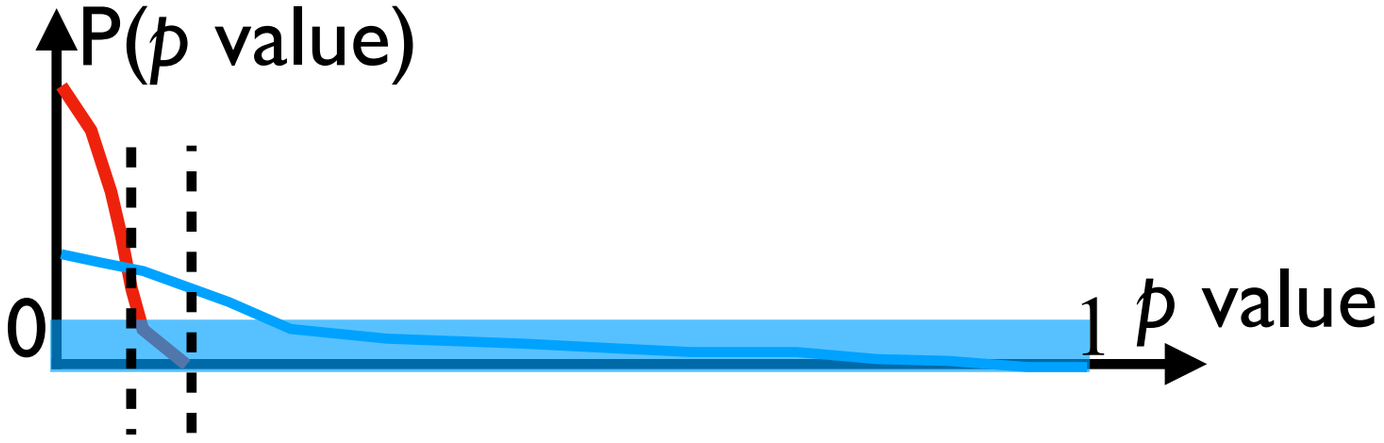




Factor analysis model: $\mathbf{X} = \tilde{\mathbf{X}}^* + \mathbf{E}^*$

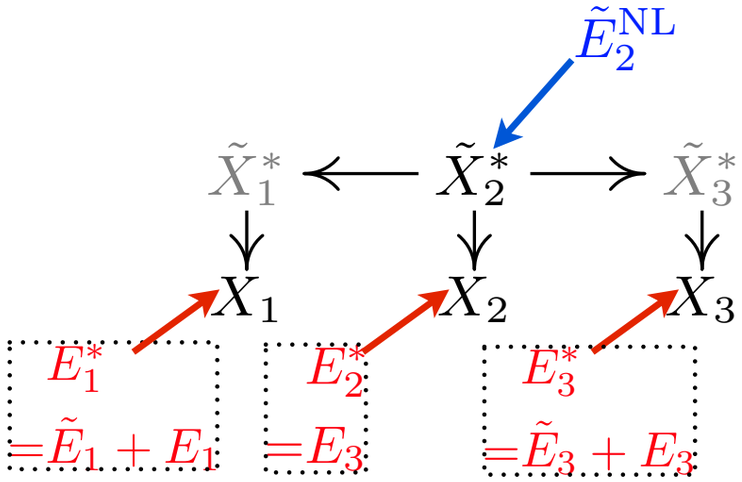
$= A^{\mathrm{NL}}\tilde{\mathbf{E}}^{\mathrm{NL}} + \mathbf{E}^*$

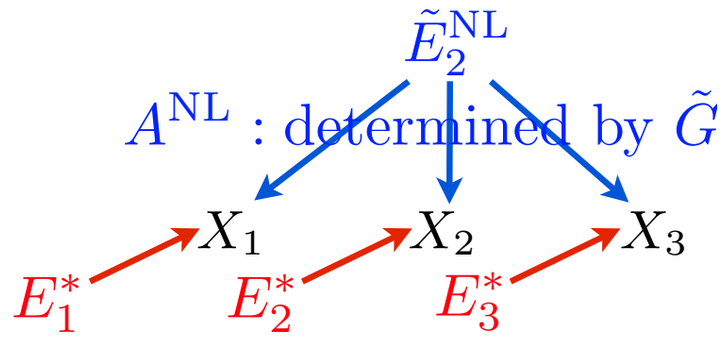
- Identifiability conditions derived based on the <u>factor analysis</u> model: the number of non-leaf nodes has to be small
- Conditions improved if <u>measurement errors have the same variance</u>
- Heuristic correction method: use a small significance level when doing CI tests



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):

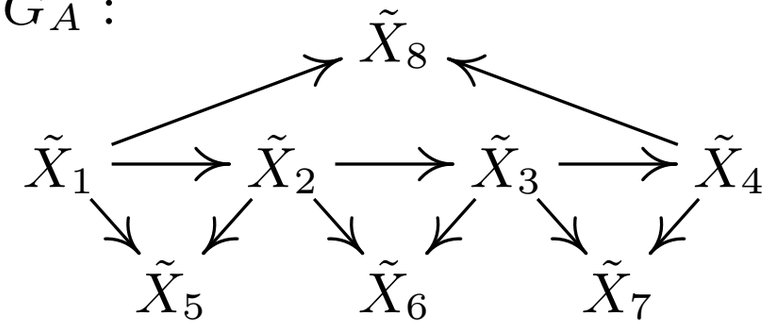
> A1. All $\tilde{E}_i$ are non-Gaussian.

- 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

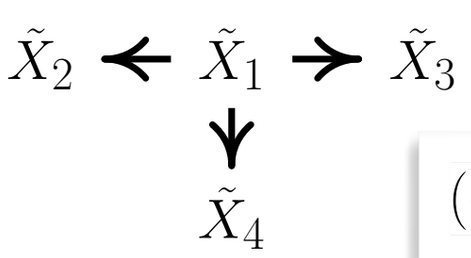
$\tilde{G}_A$:



Ordered group decomposition:

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

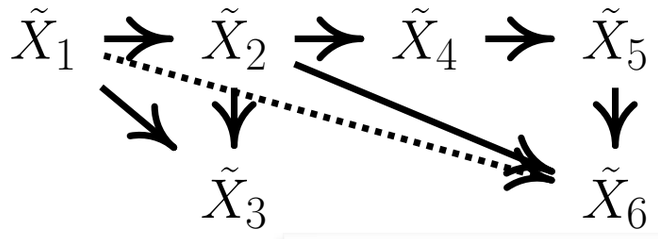
$\tilde{G}_B$:



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

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

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

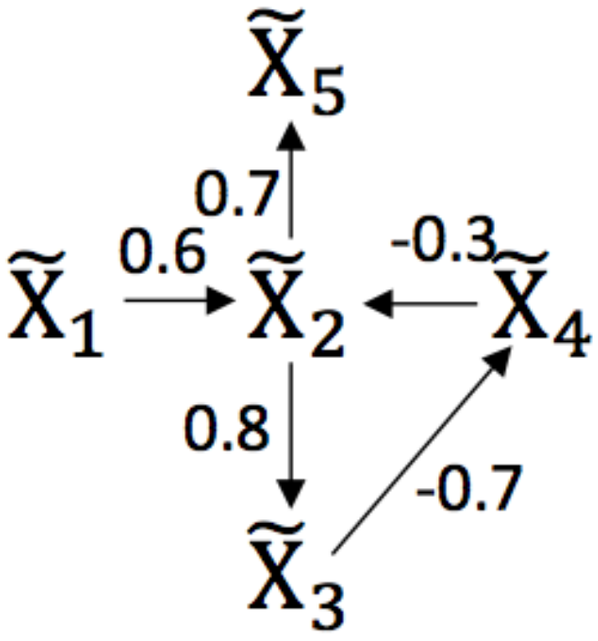


$$(\{\tilde{X}_1^*\} \to \{\tilde{X}_2^*, \tilde{X}_3^*\} \to \{\tilde{X}_4^*\} \to \{\tilde{X}_5^*, \tilde{X}_6^*\})$$

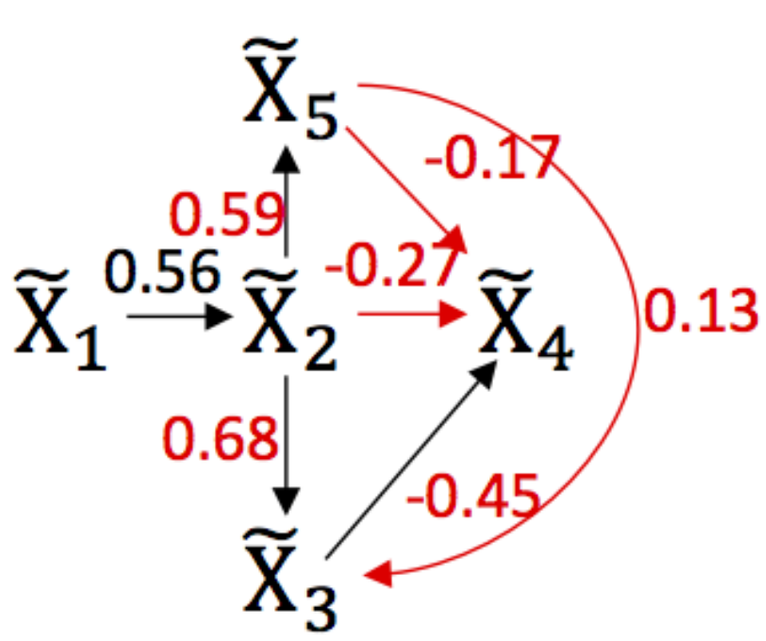
Simulation

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

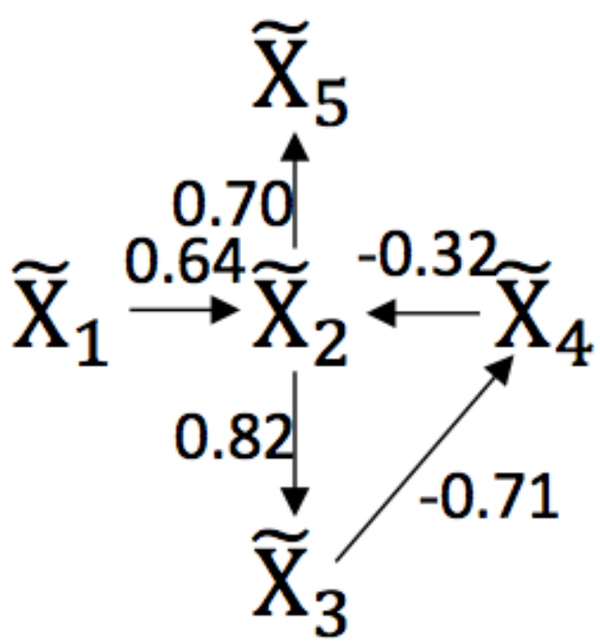
True underlying graph:



Estimated by LiNGAM:

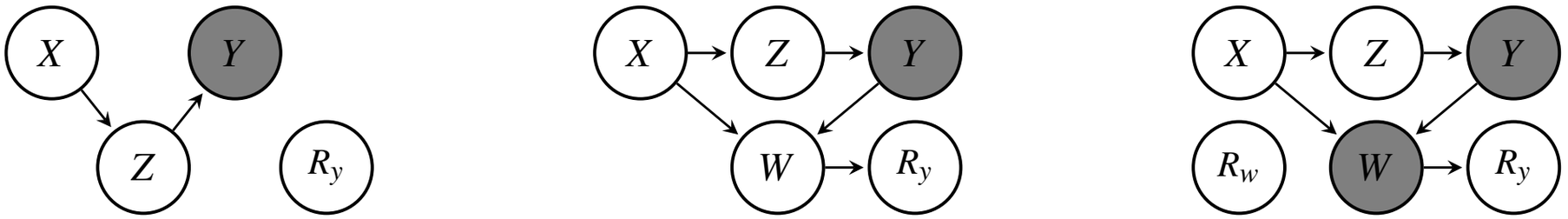


Estimated by our procedure:



Issue 3: 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



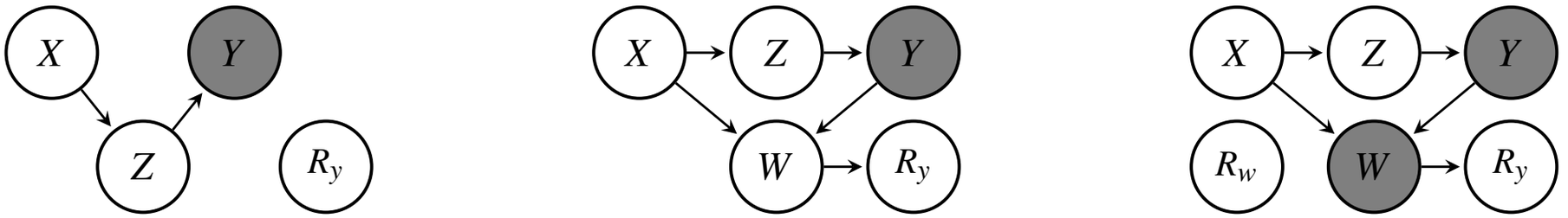
(a) An MCAR graph (b) An MAR graph (c) An MNAR graph

- 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

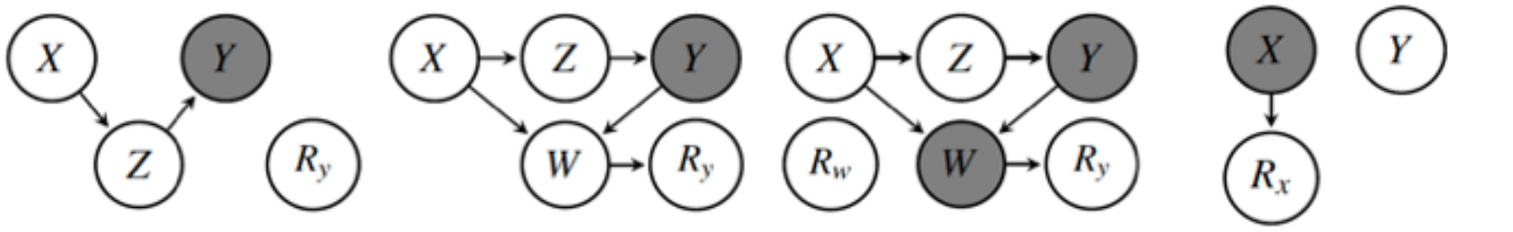
X1	X2	X3	X4	X5	X6
-9.4653403e-01		6.6703495e-01		8.2886922e-01	-1.3695521e+00	-3.2675465e-02	1.8634806e-01
-9.4895568e-01					-4.6381657e-01	-1.8280031e+00	
		5.1435422e-01		6.7338326e-01	4.3403559e-01	9.4535076e-01	7.5164028e-01
7.2489037e-01				5.1325341e-01	8.3567780e-01	2.9825903e-01	7.7796018e-02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

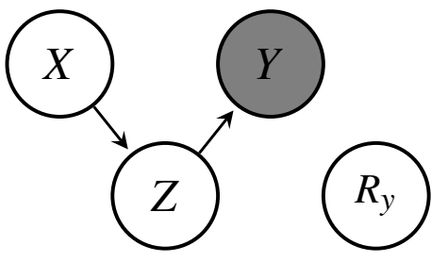
Figure 1: Exemplar missingness graphs in MCAR, MAR, MNAR, and self-masking missingness. X, Y, Z, and W are random variables. In missingness graphs, gray nodes are partially observed variables, and white nodes are fully observed variables. R_x, R_y, and R_w are the missingness indicators of X, Y, and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

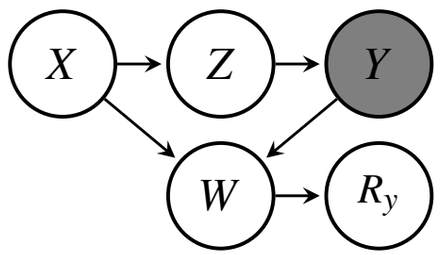
Assumptions for the Method

- 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.

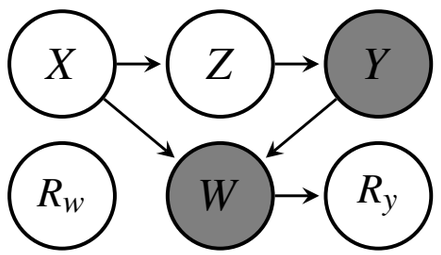
Observations



(a) An MCAR graph



(b) An MAR graph



(c) An MNAR graph

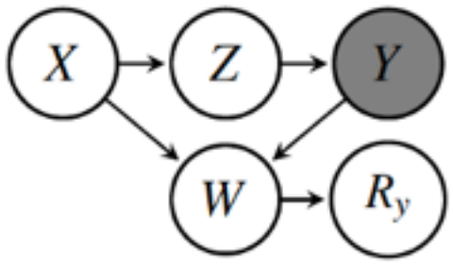
- Trust the testwise deletion conditional independence relations for causal discovery?

- Given Assumptions 1-4, we can prove:

  - If X ⫫ Y | **Z** in the testwise-deleted data, then X⫫Y | **Z** in the full data.

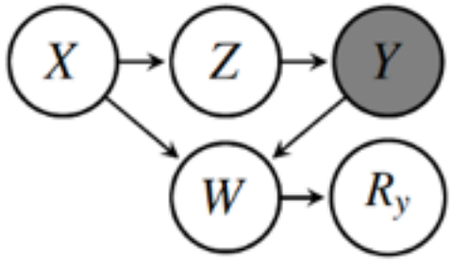
  - If testwise deletion gives extra dependence X⫫̸Y | **Z**, compared to the population, then *for at least one variable in {X}∪{Y}∪**Z**, its missingness indicator is either the direct common effect or a descendant of the direct common effect of X and Y.*

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:

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

Summary: Class 22 & 23

- Practical issues in causal discovery to be considered: They are part of the data-generating process
- Selection bias is ubiquitous
  - Where is it? Finding correct causal model in the presence of selection bias?
- Connection between measurement error and confounders
- Missingness is a causal problem!
  - Missingness graph; causal discovery under missing values