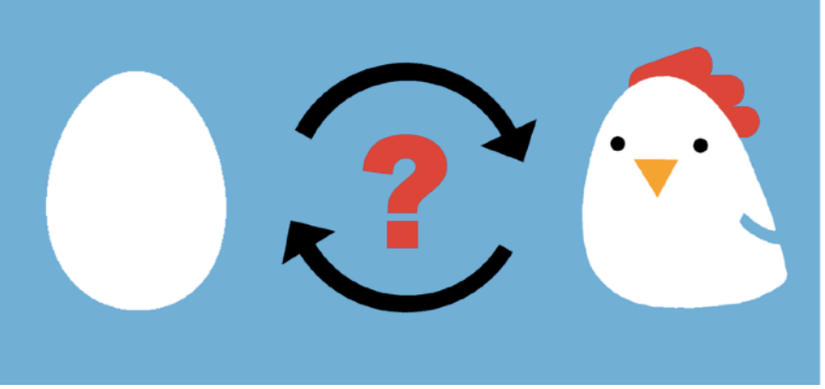


Causality and Machine Learning
(80-816/516)

Classes 17 (March 18, 2025)

Practical Issues in Causal Discovery:
Missing Values and Temporal Constraints
and
Basic Idea of Identifiability Establishment

Instructor:

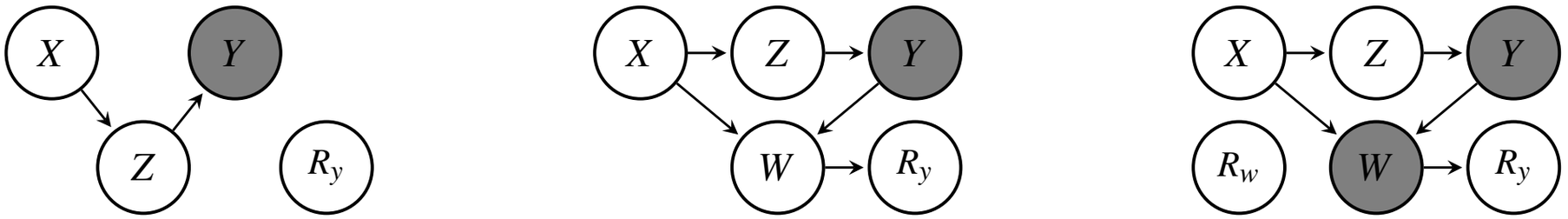
Kun Zhang (kunz1@cmu.edu)

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

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

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



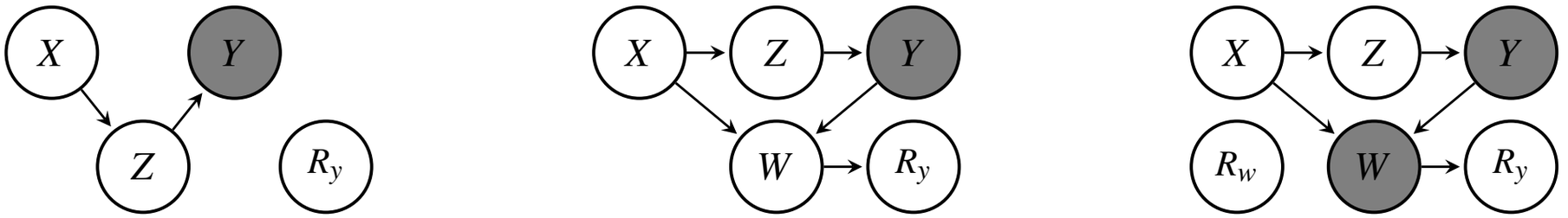
(a) An MCAR graph (b) An MAR graph (c) An MNAR graph

- 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

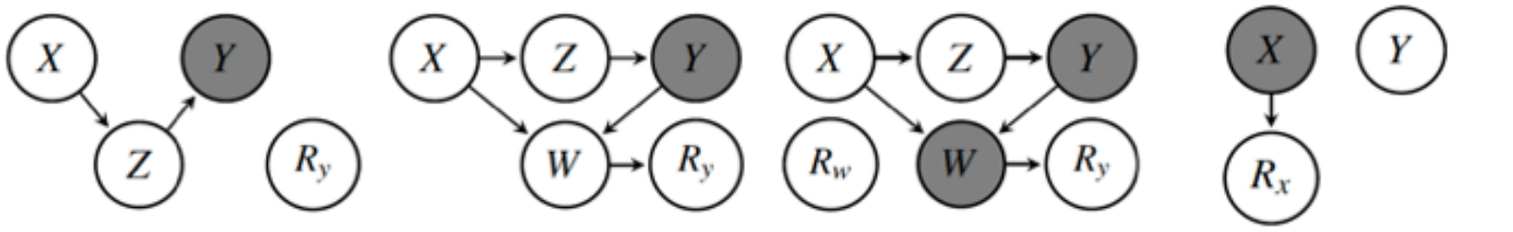
X1	X2	X3	X4	X5	X6
-9.4653403e-01		6.6703495e-01		8.2886922e-01	-1.3695521e+00	-3.2675465e-02	1.8634806e-01
-9.4895568e-01					-4.6381657e-01	-1.8280031e+00	
		5.1435422e-01		6.7338326e-01	4.3403559e-01	9.4535076e-01	7.5164028e-01
7.2489037e-01				5.1325341e-01	8.3567780e-01	2.9825903e-01	7.7796018e-02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

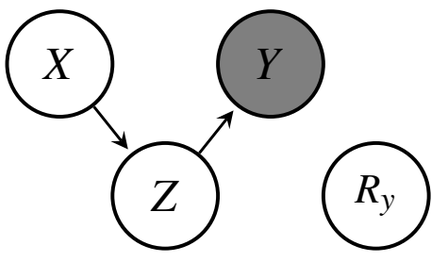
Figure 1: Exemplar missingness graphs in MCAR, MAR, MNAR, and self-masking missingness. X, Y, Z, and W are random variables. In missingness graphs, gray nodes are partially observed variables, and white nodes are fully observed variables. R_x, R_y, and R_w are the missingness indicators of X, Y, and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

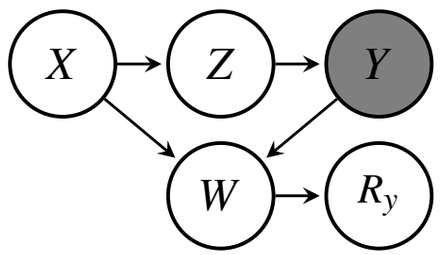
Assumptions for the Method

- 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.

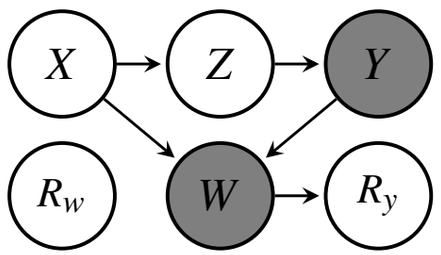
Observations



(a) An MCAR graph



(b) An MAR graph



(c) An MNAR graph

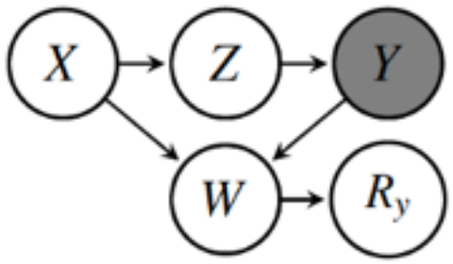
- Trust the testwise deletion conditional independence relations for causal discovery?

- Given Assumptions 1-4, we can prove:

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

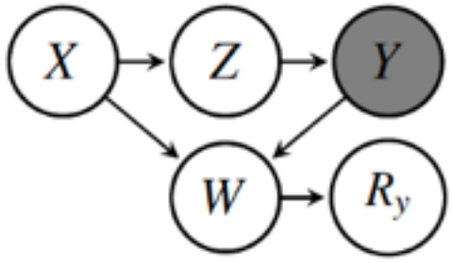
  - If testwise deletion gives extra dependence X⫫̸Y | **Z**, compared to the population, then *for at least one variable in {X}∪{Y}∪**Z**, its missingness indicator is either the direct common effect or a descendant of the direct common effect of X and Y.*

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?

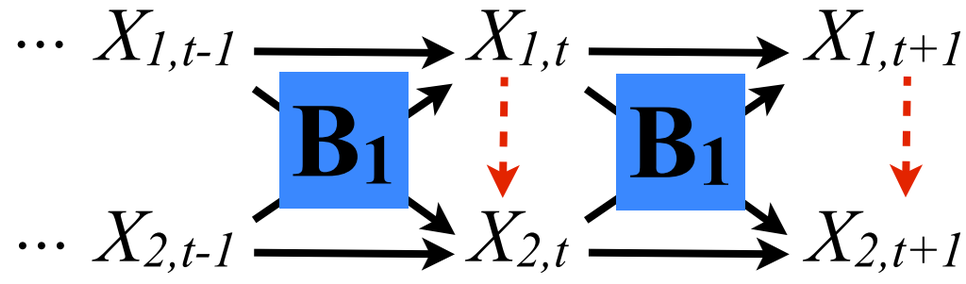
$$P(X,Y,Z) = \int_W P(X,Y,Z \mid W)P(W)dW$$
$$= \int_W P(X,Y^*,Z \mid W,R_y = 0)P(W)dW$$

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

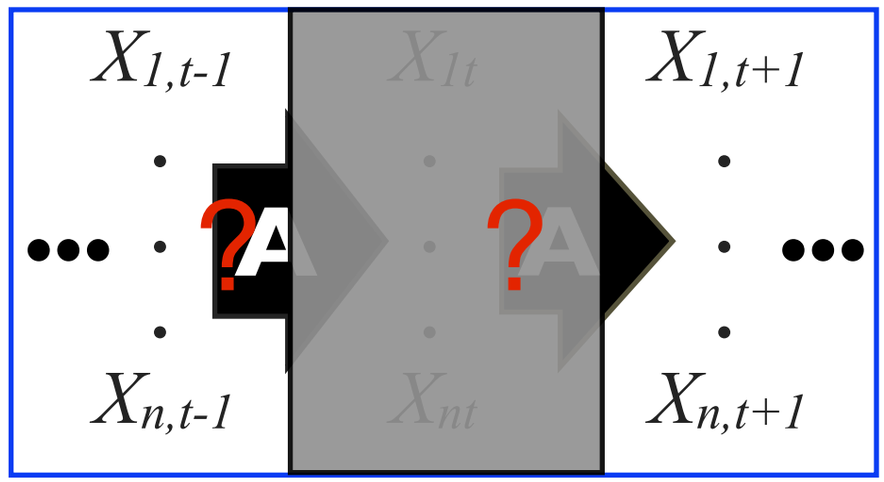
$$\widehat{X} := \alpha_1 W^S + \varepsilon_1, \quad \widehat{Y} := \alpha_2 W^S + \varepsilon_2, \quad \widehat{Z} := \alpha_3 W^S + \varepsilon_3,$$

Issue 4: Causality in Time Series

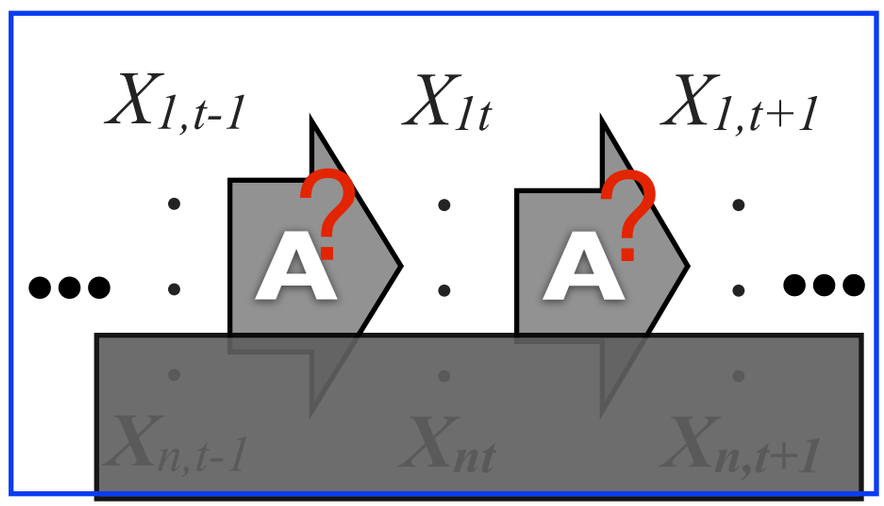
- Functional causal models in time series
  - Time-delayed causality + instantaneous relations



- Causal discovery from subsampled or temporally aggregated data

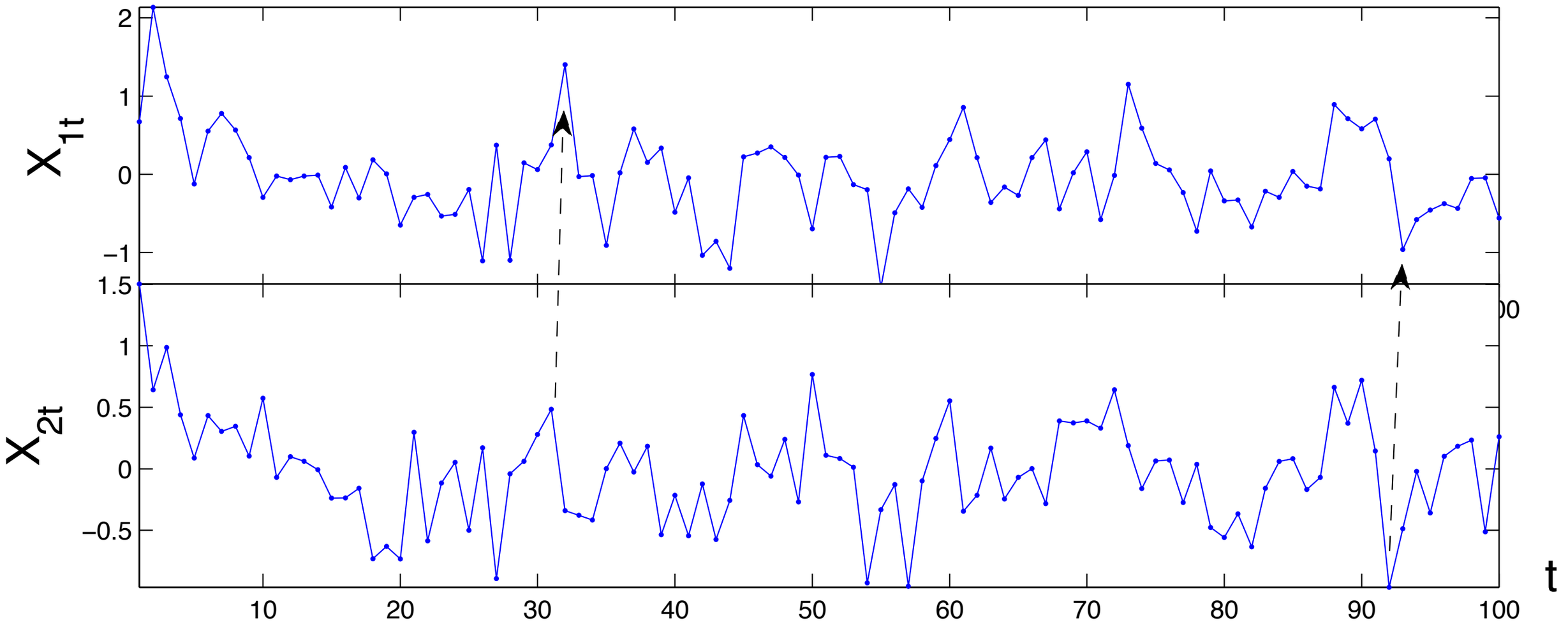


- From partially observable time series



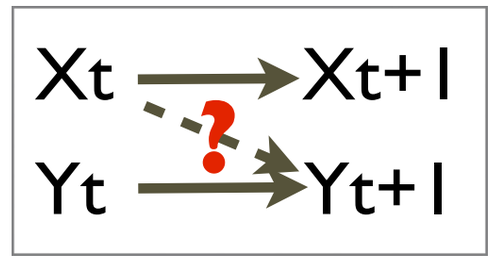
Zhang & Hyvärinen, ECML 2009;
Hyvärinen , Zhang et al., JMLR 2010;
Gong, Zhang, Schölkopf, Tao, Geigere, ICML 2015; UAI 2017;
Geiger, Zhang, Gong, Janzing, Schölkopf, ICML 2015

Granger Causality: Motivation



Granger Causality: Original Definition & Practical Constraints

- Two principles (Granger, '80)
  - Future cannot cause past
  - No redundant info: Cause contains unique information about effect

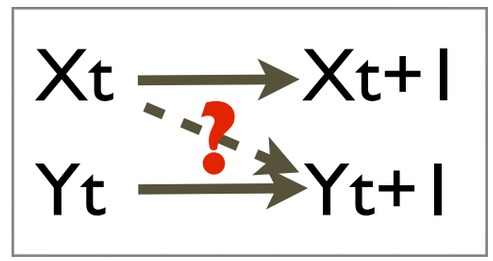


- X causes Y if $P\left(Y_{t+1} \in A \mid \Omega_t\right) \neq P\left(Y_{t+1} \in A \mid \Omega_t^{-X}\right)$
  - Completely nonparametric; $Y_{t+1} \not\perp X_t$ given all the remaining information until time t
- In practice: causality in mean; linear Granger causality

- C.W.J. Granger, Testing for causality: A personal viewpoint. Journal of Economic Dynamics & Control 2: 329–352, 1980

Conditional Independence-Based Method for Causal Discovery from Time Series

- Two principles (Granger, '80)
  - **Future cannot cause past**
  - No redundant info: Cause contains unique information about effect



- X causes Y if $P\left(Y_{t+1} \in A \mid \Omega_t\right) \neq P\left(Y_{t+1} \in A \mid \Omega_t^{-X}\right)$
  - Completely nonparametric; $Y_{t+1} \not\perp\!\!\!\perp X_t$ given all the remaining information until time t
- In practice: causality in mean; linear Granger causality

- The PC algorithm still applies; additional temporal constraints!

Extension of PC for Causal Analysis of Time series

- Unroll the processes
- Apply PC + temporal constraints
- Has been applied to climate analysis

Chu and Glymour, Search for nonlinear time series causal models, JMLR 2008

Application: Ocean Climate Analysis

SOI Southern Oscillation Index: Sea Level Pressure (SLP) anomalies between Darwin and Tahiti

WP Western Pacific: Low frequency temporal function of the 'zonal dipole' SLP spatial pattern over the North Pacific.

AO Arctic Oscillation: First principal component of SLP poleward of 20° N

NAO North Atlantic Oscillation: Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland

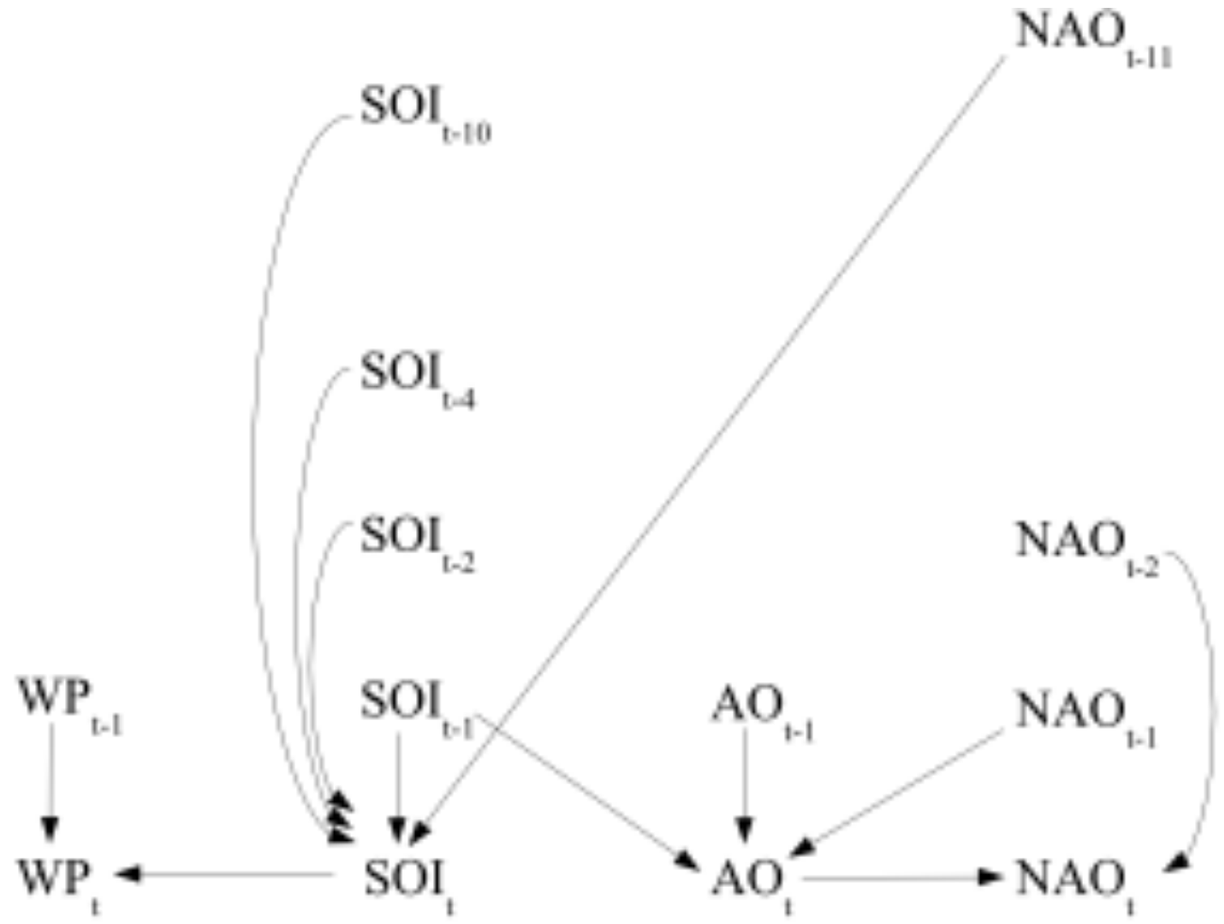
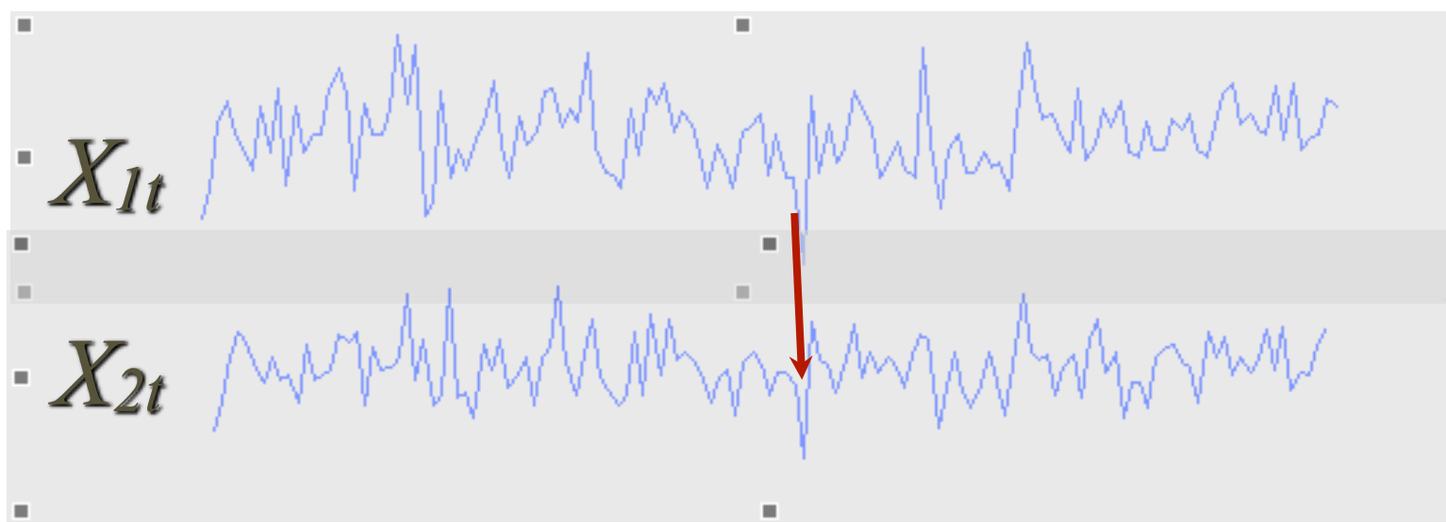


Figure 7: Causal connections among 4 ocean climate indices, using the additive non-linear algorithm

Practical Granger Causality



- X_1: $\{X_{1t}\}$ Granger causes X_2: $\{X_{2t}\}$ if it contains information helping predict $X_{2,t+h}$ (h>0) contained nowhere else (Granger, 1969)

- Temporal constraint: causes must precede effects + linear causal relations

- Vector autoregression (VAR) estimated by multivariate least squares (MLS)

$$\mathbf{X}_t = \sum_{\tau=1}^{p} \mathbf{B}_\tau \mathbf{X}_{t-\tau} + \mathbf{E}_t$$

An Example

- Analyze cheese price (X_1), butter price (X_2), and milk price (X_3);recorded monthly from January 1986 to April 2014
  - http://future.aae. wisc.edu/tab/prices.html
- Estimate $\begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} = \mathbf{B}_1 \cdot \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{3,t-1} \end{bmatrix} + \begin{bmatrix} E_{1t} \\ E_{2t} \\ E_{3t} \end{bmatrix}$
- $\hat{\mathbf{B}}_1 = \begin{pmatrix} 0.8381 & 0.0810 & 0.0375 \\ 0.0184 & 0.9592 & -0.0473 \\ 0.2318 & 0.0522 & 0.7446 \end{pmatrix}$

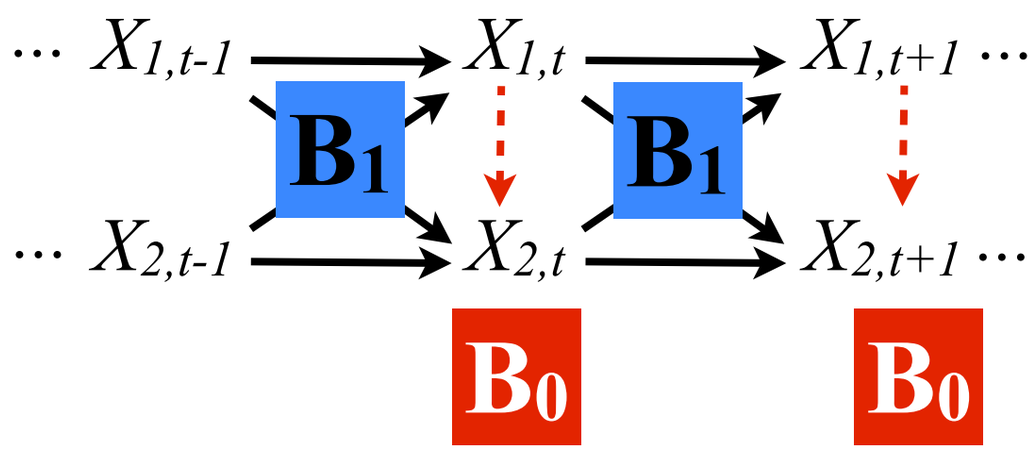
Granger Causality with Instantaneous Relations

$$\mathbf{X}_t = \sum_{\tau=1}^{p} \mathbf{B}_\tau \mathbf{X}_{t-\tau} + \mathbf{E}_t$$

- Are E_{it} independent? ⇒ instantaneous effects between X_{it} (Reale, Wilson et al., 2001)

- **Granger causality with instantaneous effects:**

$$\mathbf{X}_t = \sum_{\tau=1}^{p} \mathbf{B}_\tau \mathbf{X}_{t-\tau} + \mathbf{B}_0 \mathbf{X}_t + \mathbf{E}_t, \quad \text{or} \quad \mathbf{X}_t = \sum_{\tau=0}^{p} \mathbf{B}_\tau \mathbf{X}_{t-\tau} + \mathbf{E}_t$$



What Happens If We Ignore Instantaneous Effects (Hyvärinen et al., ICML 2008)

- Time-delayed "causal relations" will be changed

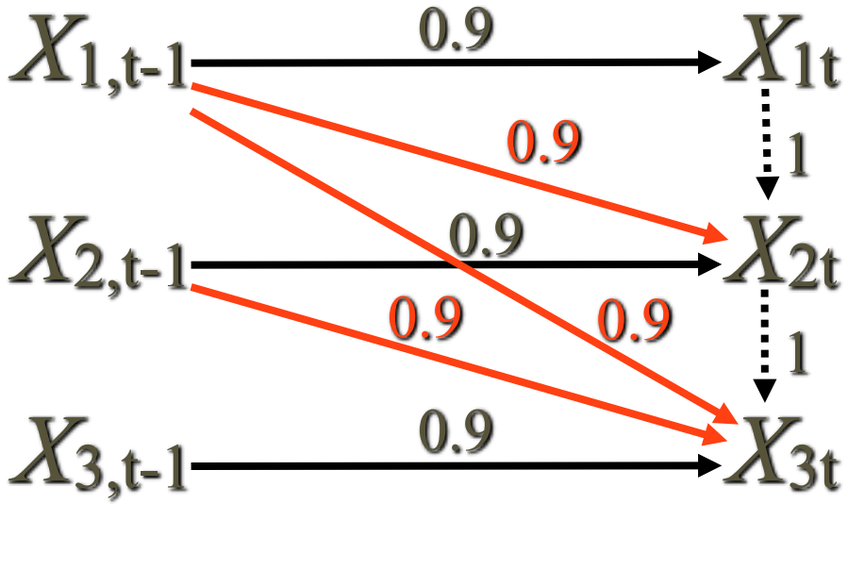
$$\mathbf{X}_t = \sum_{\tau=0}^{p} \mathbf{B}_\tau \mathbf{X}_{t-\tau} + \mathbf{E}_t$$

$$\Rightarrow \mathbf{X}_t = \sum_{\tau=1}^{p} (\mathbf{I} - \mathbf{B}_0)^{-1} \cdot \mathbf{B}_\tau \cdot \mathbf{X}_{t-\tau} + (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{E}_t$$

- Example

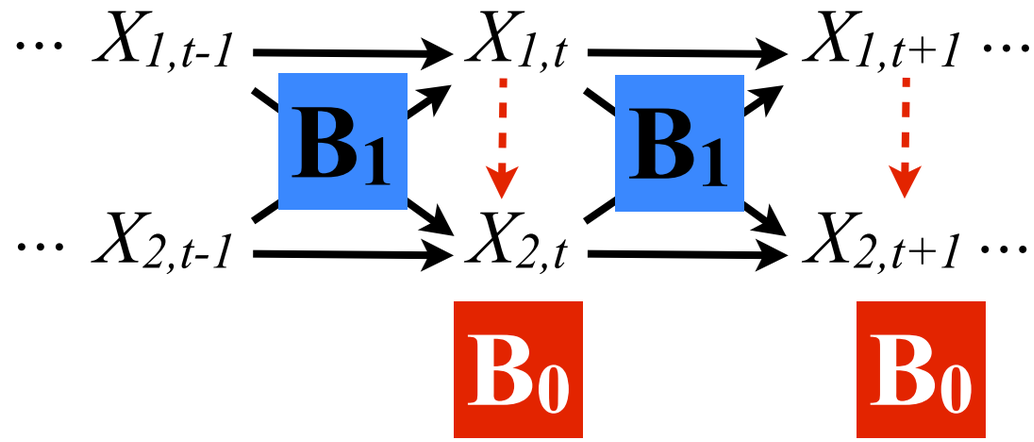
$$\mathbf{B}_0 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 0.9 \end{pmatrix}:$$

$$(\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 & 0 \\ 0.9 & 0.9 & 0 \\ 0.9 & 0.9 & 0.9 \end{pmatrix}:$$



Identification (Zhang & Hyvärinen, ECML 2009)

- E_{it} independent for different i and t, i.e., spatially & temporally independent
- If at most one of E_{it} is Gaussian, it can be solved by multichannel blind deconvolution (MBD) with causal FIR filters
- MBD estimates $\mathbf{W}$ to make $\hat{E}_{it}$ spatially and temporally independent
- $\mathbf{B_\tau}$ can be found from $\mathbf{W_\tau}$, by extending LiNGAM analysis



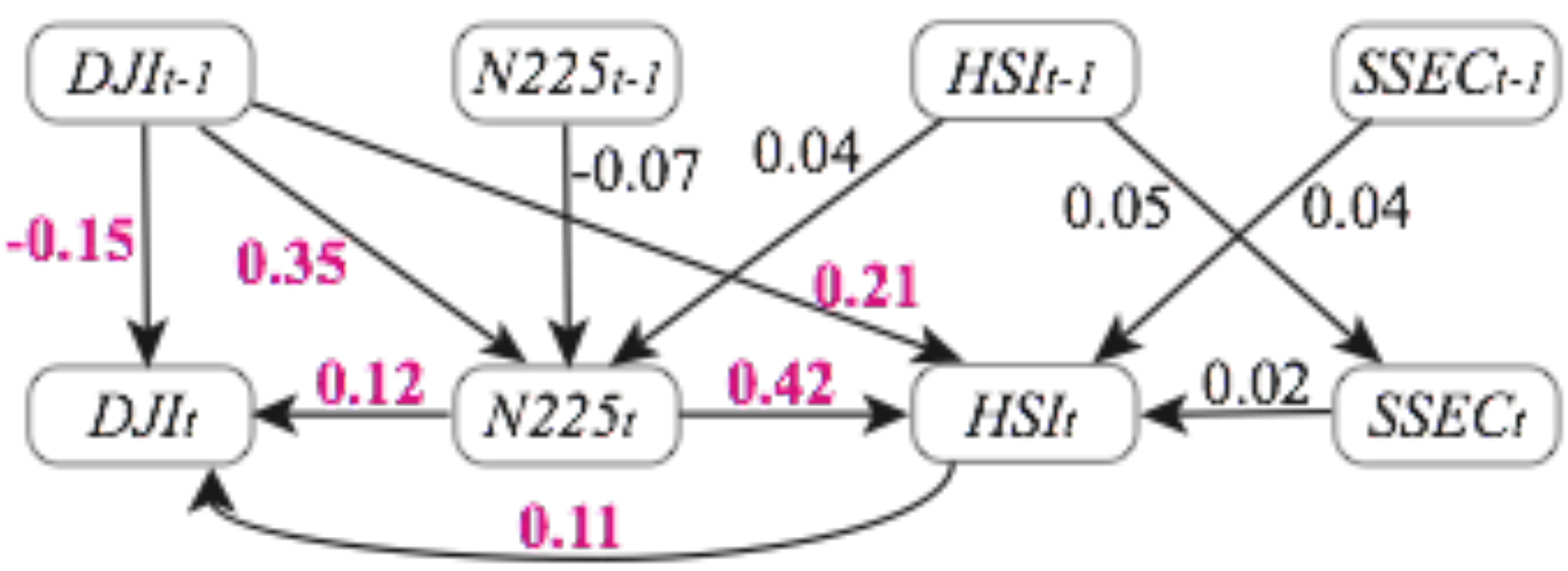
$$\mathbf{X}_t = \sum_{\tau=0}^{p} \mathbf{B}_\tau \mathbf{X}_{t-\tau} + \mathbf{E}_t$$

$$\Rightarrow \mathbf{E}_t = (\mathbf{I} - \mathbf{B}_0)\mathbf{X}_t - \sum_{\tau=1}^{p} \mathbf{B}_\tau \mathbf{X}_{t-\tau}$$

$$= \sum_{\tau=0}^{p} \mathbf{W}_\tau \mathbf{X}_{t-\tau}$$

Experiment on Financial Data

- Extended Granger causality analysis (Granger causality with instantaneous effects) of daily returns of stock indices DJI, N225, HSI, and SSEC, with k = 1 lag (Zhang & Hyvärinen, ECML 2009)



Two Schemes of Temporal Aggregation

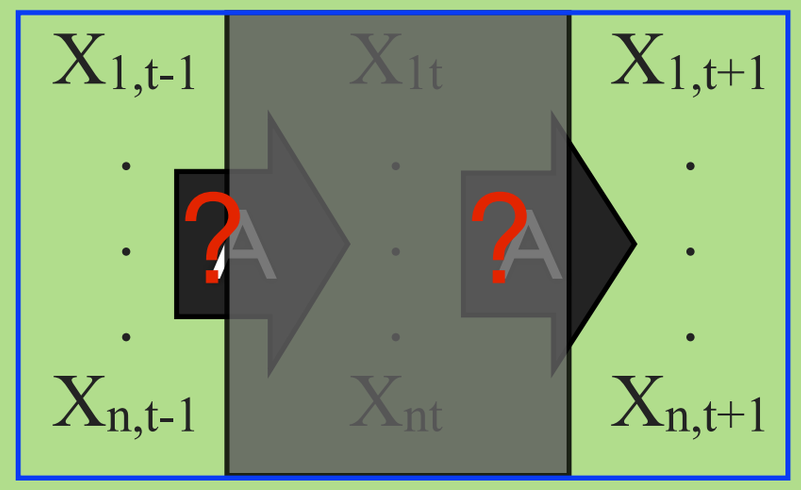
- **Subsampling (syst… sampling)**

$$\mathbf{x}_1, \quad \tilde{\mathbf{x}}_1 = \frac{1}{k} \sum_{l=} \quad \ldots$$

Can we recover the causal influence matrix *A*?

- Examples: temperature data, stock daily returns, GDP, fMRI...

Assume $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{E}_t$



Causal info tends to disappear as $k \to \infty$

…ausal info tends to be …antaneous as $k \to \infty$:

$$\tilde{\mathbf{X}}_t \approx \mathbf{A}\tilde{\mathbf{X}}_t + \tilde{\mathbf{E}}_t$$

Causal Discovery from Subsampled Data: Linear Case

ICML 2015

Discovering Temporal Causal Relations from Subsampled Data

Mingming Gong[*1] MINGMING.GONG@STUDENT.UTS.EDU.AU
Kun Zhang[*2,3] KZHANG@TUEBINGEN.MPG.DE
Bernhard Schölkopf[2] BS@TUEBINGEN.MPG.DE
Dacheng Tao[1] DACHENG.TAO@UTS.EDU.AU
Philipp Geiger[2] PGEIGER@TUEBINGEN.MPG.DE

[1] Centre for Quantum Computation and Intelligent Systems, FEIT, University of Technology, Sydney, NSW, Australia
[2] Max Plank Institute for Intelligent Systems, Tübingen 72076, Germany
[3] Information Sciences Institute, University of Southern California

Abstract

Granger causal analysis has been an important tool for causal analysis for time series in various fields, including neuroscience and economics, and recently it has been extended to include in-

1. Introduction

Granger causal analysis (Granger, 1980) has been widely used to find the temporal causal relations from time series. Time series x_1 is said to cause times series x_2 in the Granger's sense, if and only if the past and current values of x_1 contain useful information to predict the future values of

Causal Discovery from Temporally Aggregated Time Series

UAI 2017

Causal Discovery from Temporally Aggregated Time Series

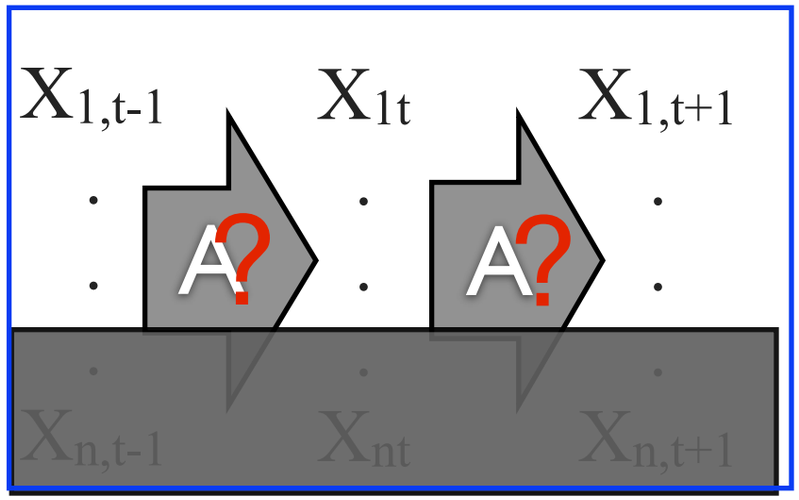
Mingming Gong[*†], **Kun Zhang**[†], **Bernhard Schölkopf**[‡], **Clark Glymour**[†], **Dacheng Tao**[♯]
[*]Centre for Artificial Intelligence, FEIT, University of Technology Sydney, NSW, Australia
[†]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA
[‡]Max Plank Institute for Intelligent Systems, Tübingen, Germany
[♯]School of Information Technologies, FEIT, University of Sydney, NSW, Australia

Abstract

Discovering causal structure of a dynamical system from observed time series is a traditional and important problem. In many practical applications, observed data are obtained by applying subsampling or temporally aggregation to the original causal processes, making it difficult to discover the underlying causal relations. Subsampling refers to the procedure that for every k consecutive observations, one is kept, the rest being skipped, and recently some advances have been

underlying physical process. However, since the true causal frequency is usually unknown, the time series data are often measured at the frequency lower than the causal frequency. For example, some econometric indicators such as GDP and non-farm payroll are usually recorded at quarterly and monthly scales. Causal interactions between the processes, however, may take place at the weekly or fortnightly scales (Ghysels et al., 2016). In neuroscience, imaging technologies have relatively low temporal resolutions, while many high frequency neuronal interactions are important for understanding neuronal dynamics (Zhou et al., 2014). In these situations, the available observations have a lower resolution

Confounding Effect



- What if Z_t is not observable?

$$\begin{bmatrix} X_t \\ Z_t \end{bmatrix} = \overbrace{\begin{bmatrix} B & C \\ D & E \end{bmatrix}}^{\triangleq A} \begin{bmatrix} X_{t-1} \\ Z_{t-1} \end{bmatrix} + E_t$$

- Discovered causal relations sensitive to confounders: Example

$$A = \left(\begin{array}{cc|c} 0.9 & 0 & 0.5 \\ 0.1 & 0.1 & 0.8 \\ \hline 0 & 0 & 0.9 \end{array}\right), B_{\mathrm{pG}} := \mathbb{E}(X_t X_{t-1}^\top)\mathbb{E}(X_t X_t^\top)^{-1} = \begin{pmatrix} 0.89 & 0.35 \\ 0.08 & 0.65 \end{pmatrix}$$

- Can we identify B (as well as C) from X_t ?

- G. Philipp, K. Zhang, M. Gong, D. Janzing, B. Schölkopf. Causal inference by identification of vector autoregressive processes with hidden components, ICML 2015

We tried to find causal relations among the measured time series; what if causal processes are hidden?

Causal representation learning from temporal data (Next week!)

Idea of Identifiability establishment: A Linear, Non-Gaussian Case

(see the notes in PDF)

Identifiability of Parameters in Statistical Models

- Identifiability, in simple words, means that different values of a parameter must produce different probability distributions.
- Mathematically, a parameter θ is said to be identifiable if and only
  $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$, or equivalently $P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$
- Is the mean of a Gaussian distribution identifiable?

Example 1: On the Identifiability of the Post-Nonlinear Causal Model (https://arxiv.org/pdf/1205.2599)

$$x_2 = f_2(f_1(x_1) + e_2), \tag{2}$$

where x_1 and e_2 are independent, function f_1 is non-constant, and f_2 is invertible. If the other causal direction, $x_2 \to x_1$ is true, the data generating process given by the PNL causal model is

$$x_1 = g_2(g_1(x_2) + e_1), \tag{3}$$

where x_2 and e_1 are independent, g_1 is non-constant, and g_2 is invertible.

Notation. The following notations are used hereafter. Suppose that both (2) and (3) hold. Random variables t_1 and z_2 and functions h and h_1 defined as follows:

$$\begin{array}{ll} t_1 \triangleq g_2^{-1}(x_1), & z_2 \triangleq f_2^{-1}(x_2), \\ h \triangleq f_1 \circ g_2, & h_1 \triangleq g_1 \circ f_2. \end{array}$$

That is, $h(t_1) = f_1(g_2(t_1)) = f_1(x_1)$, and similarly, h_1 is a function of z_2. Moreover, we let $\eta_1(t_1) \triangleq \log p_{t_1}(t_1)$, and $\eta_2(e_2) \triangleq \log p_{e_2}(e_2)$.[1] The following theorem gives the constraint that p_{t_1}, p_{e_2}, and h must satisfy to make both (2) and (3) hold.

Theorem 1 *Assume that (x_1, x_2) can be described by both of the causal relations given in (2) and in (3). Further suppose that involved densities and nonlinear functions p_{t_1}, p_{e_2}, f_1, f_2, g_1, and g_2 are third-order differentiable, and that p_{e_2} is positive on $(-\infty, +\infty)$. We then have the following equation for every (x_1, x_2) satisfying $\eta_2'' h' \neq 0$:*

$$\begin{aligned} \eta_1''' - \frac{\eta_1'' h''}{h'} &= \left(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2''\right) \cdot h' h'' - \frac{\eta_2'''}{\eta_2''} \cdot h' \eta_1'' \\ &\quad + \eta_2' \cdot \left(h''' - \frac{h''^2}{h'}\right), \end{aligned} \tag{4}$$

and h_1 depends on η_1, η_2, and h in the following way:

$$\frac{1}{h_1'} = \frac{\eta_1'' + \eta_2'' h'^2 - \eta_2' h''}{\eta_2'' h'}. \tag{5}$$

Proof of Theorem 1: We prove this theorem using the linear separability of the logarithm of the joint density of independent variables, which states the fact that for a set of independent random variables whose joint density is twice differentiable, the Hessian of the logarithm of their density is diagonal everywhere (Lin, 1998). Since g_2 is invertible, the independence between x_1 and e_2 is equivalent to that between t_1 and e_2. Similarly, the independence between x_2 and e_1 is equivalent to that between z_2 and e_1. Combining the two causal models (2) and (3), one can see that the transformation from (z_2, e_1) to (t_1, e_2) is

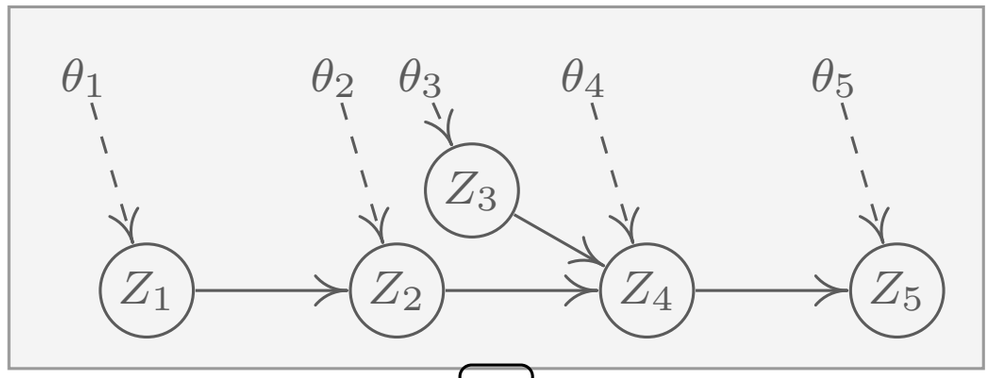
$$\begin{aligned} t_1 &= h_1(z_2) + e_1, & (6) \\ e_2 &= z_2 - h(t_1). & (7) \end{aligned}$$

Denote by $\mathbf{J}$ the Jacobian matrix of this transformation. One can see that $|\mathbf{J}| = 1$. Denote by $p_{(z_2,e_1)}$ the joint density of (z_2, e_1). We then have $p_{t_1} \cdot p_{e_2} = p_{(z_2,e_1)}/|\mathbf{J}| = p_{(z_2,e_1)}$, so, $\log p_{(z_2,e_1)} = \eta_1(t_1) + \eta_2(e_2)$. One can find the (1,2)-th entry of the Hessian matrix of $\log p_{(z_2,e_1)}$ w.r.t. (z_2, e_1): $\frac{\partial^2 \log p_{(z_2,e_1)}}{\partial e_1 \partial z_2} = \eta_1'' \frac{\partial t_1}{\partial z_2} - \eta_2'' h' \frac{\partial e_2}{\partial z_2} - \eta_2' h'' \frac{\partial t_1}{\partial z_2} = \eta_1'' h_1' - \eta_2'' h' + \eta_2'' h'^2 h_1' - \eta_2' h'' h_1'$.

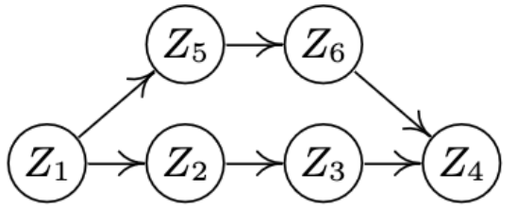
The independence between z_2 and e_1 implies $\frac{\partial^2 \log p_{(z_2,e_1)}}{\partial e_1 \partial z_2} = 0$ for every possible (z_2, e_1). That is, $\eta_1'' h_1' - \eta_2'' h' + \eta_2'' h'^2 h_1' - \eta_2' h'' h_1' = 0$. From this equation one can see that $h_1' = 0$ implies $\eta_2'' h' = 0$. Consequently, the points which satisfy $\eta_2'' h' \neq 0$ also make $h_1' \neq 0$. For such points, dividing both sides of this equation by $h_1' \eta_2'' h'$ finally leads to (5). Furthermore, since h_1 is a functions of z_2 and does not depend on e_1, we have $\partial\left(\frac{1}{h_1'}\right)\Big/\partial e_1 = 0$. According to (5), we have $\partial\left(\frac{\eta_1'' + \eta_2'' h'^2 - \eta_2' h''}{\eta_2'' h'}\right)/\partial e_1 = 0$, which gives $2\eta_2''^2 h'^2 h'' - \eta_2' \eta_2'' h' h''' + \eta_2'' \eta_1''' h' - \eta_2' \eta_2''' h'^2 h'' + \eta_2' \eta_2'' h''^2 + \eta_2''' \eta_1'' h'^2 - \eta_2'' \eta_1'' h'' = 0$. For the points satisfying $\eta_2'' h' \neq 0$, we divide both sides of the above equation by $\eta_2'' h'$. After some simplifications, (4) is obtained. ■

Example 2: Causal Representation Learning from Multiple Distributions: A General Setting (https://arxiv.org/abs/2402.05052)

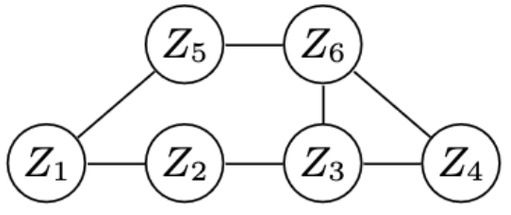
i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



- Goal: Uncovering hidden variables Z_i with changing causal relations from $\mathbf{X}$ in nonparametric settings
- What is identifiable?
  - Markov network of Z_i
  - Each estimated variable $\tilde{Z}_i$ is a function of Z_i and it intimate neighbors
- In this example, each Z_i ($i \neq 4$) can be recovered up to component-wise transformation



(a) $\mathcal{G}_Z$, the DAG over true latent variables Z_i.



(b) The corresponding Markov network $\mathcal{M}_Z$.

Example 2: Causal Representation Learning from Multiple Distributions: A General Setting (https://arxiv.org/abs/2402.05052)

A key ingredient of our results is the Markov network that represents conditional dependencies among random variables via an undirected graph. Let $\mathcal{M}_Z$ be the Markov network over variables Z, i.e., with vertices $\{Z_i\}_{i=1}^n$ and edges $\{Z_i, Z_j\} \in \mathcal{E}(\mathcal{M}_Z)$ if and only if $Z_i \not\perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,j\}}$.[2] Also, we denote by $|\mathcal{M}_Z|$ the number of undirected edges in the Markov network. In Section 3.1, apart from showing how to estimate the underlying latent causal variables up to certain indeterminacies, we also show that such latent Markov network $\mathcal{M}_Z$ can be recovered up to isomorphism. To achieve so, we make use of the following property (assuming that p_Z is twice differentiable):

$$Z_i \perp\!\!\!\perp Z_j \mid Z_{[n]\setminus\{i,j\}} \iff \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j} = 0. \quad (3)$$

Such a connection between pairwise conditional independence and cross derivatives of the density function has been noted by Lin (1997) and utilized in Markov network learning for observed variables (Zheng et al., 2023). With the recovered latent Markov network structure, we provide results in Section 3.2 to show how it relates to the moralized graph of true latent causal DAG $\mathcal{G}_Z$, by exploiting a specific type of faithfulness assumption that is considerably weaker than the standard faithfulness assumption used in the literature of causal discovery (Spirtes et al., 2001).

Proof. Denote by $\operatorname{vol} A$ the volume of matrix A, which is the product of its singular values. Note that $\operatorname{vol} A = \sqrt{\det AA^T}$ when A is of full row rank. In the change-of-variable formula, when the Jacobian is a rectangular matrix, the absolute determinant of the Jacobian can be replaced with the matrix volume (Ben-Israel, 1999; Gemici et al., 2016; Khemakhem et al., 2020a).

Since $X = g(Z)$ and $\hat{X} = \hat{g}(\hat{Z})$, by Eq. (2) and the change-of-variable formula, we have

$$p_{\hat{X}} = p_X \implies p_{\hat{g}(\hat{Z})} = p_{g(Z)} \implies p_{g^{-1}\circ\hat{g}(\hat{Z})} \operatorname{vol} J_{g^{-1}} = p_Z \operatorname{vol} J_{g^{-1}} \implies p_{v(\hat{Z})} = p_Z,$$

where $J_{g^{-1}}$ is the Jacobian matrix of g^{-1} and $v := g^{-1} \circ \hat{g}$ is a composition of diffeomorphisms (and hence also a diffeomorphism). Let J_v be the Jacobian matrix of v. The change-of-variable formula implies

$$\begin{aligned} p(\hat{Z};\hat{\theta})|\det J_{v^{-1}}| &= p(Z;\theta) \\ \log p(\hat{Z};\hat{\theta}) &= \log p(Z;\theta) + \log|\det J_v|. \end{aligned} \quad (8)$$

Suppose $\hat{Z}_k$ and $\hat{Z}_l$ are conditionally independent given $\hat{Z}_{[n]\setminus\{k,l\}}$ i.e., they are not adjacent in the Markov network over $\hat{Z}$. For each $\hat{\theta}$, by Lin (1997), we have

$$\frac{\partial^2 \log p(\hat{Z};\hat{\theta})}{\partial \hat{Z}_k \partial \hat{Z}_l} = 0. \quad (9)$$

To see what it implies, we find the first-order derivative of Eq. (8):

$$\frac{\partial \log p(\hat{Z};\hat{\theta})}{\partial \hat{Z}_k} = \sum_{i=1}^n \frac{\partial \log p(Z;\theta)}{\partial Z_i}\frac{\partial Z_i}{\partial \hat{Z}_k} + \frac{\partial \log|\det J_v|}{\partial \hat{Z}_k}.$$

Let

$$\eta(\theta) := \log p(Z;\theta), \quad \eta_i'(\theta) := \frac{\partial \log p(Z;\theta)}{\partial Z_i}, \quad \eta_{ij}''(\theta) := \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j}, \quad h_{i,l}' := \frac{\partial Z_i}{\partial \hat{Z}_l}, \quad \text{and} \quad h_{i,kl}'' := \frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_l}.$$

We then derive the second-order derivative w.r.t. $\hat{Z}_k$ and $\hat{Z}_l$ and apply Eq. (9):

$$\begin{aligned} 0 &= \sum_{j=1}^n \sum_{i=1}^n \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j}\frac{\partial Z_j}{\partial \hat{Z}_l}\frac{\partial Z_i}{\partial \hat{Z}_k} + \sum_{i=1}^n \frac{\partial \log p(Z;\theta)}{\partial Z_i}\frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_l} + \frac{\partial^2 \log|\det J_v|}{\partial \hat{Z}_k \partial \hat{Z}_l} \\ &= \sum_{i=1}^n \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i^2}\frac{\partial Z_i}{\partial \hat{Z}_l}\frac{\partial Z_i}{\partial \hat{Z}_k} + \sum_{j=1}^n \sum_{i:\{Z_j,Z_i\}\in\mathcal{E}(\mathcal{M}_Z)} \frac{\partial^2 \log p(Z;\theta)}{\partial Z_i \partial Z_j}\frac{\partial Z_j}{\partial \hat{Z}_l}\frac{\partial Z_i}{\partial \hat{Z}_k} \\ &\quad + \sum_{i=1}^n \frac{\partial \log p(Z;\theta)}{\partial Z_i}\frac{\partial^2 Z_i}{\partial \hat{Z}_k \partial \hat{Z}_l} + \frac{\partial^2 \log|\det J_v|}{\partial \hat{Z}_k \partial \hat{Z}_l} \quad (10) \\ &= \sum_{i=1}^n \eta_{ii}''(\theta) h_{i,l}' h_{i,k}' + \sum_{i=1}^n \sum_{j:\{Z_j,Z_i\}\in\mathcal{E}(\mathcal{M}_Z)} \eta_{ij}''(\theta) h_{j,l}' h_{i,k}' + \sum_{i=1}^n \eta_i'(\theta) h_{i,kl}'' + \frac{\partial^2 \log|\det J_v|}{\partial \hat{Z}_k \partial \hat{Z}_l}. \quad (11) \end{aligned}$$

See the notes in PDF

Summary

- 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
- Basic but general idea of identifiability establishment in causal representation learning