Causal inference and causal discovery

Shikui Tu

Department of Computer Science and Engineering, Shanghai Jiao Tong University

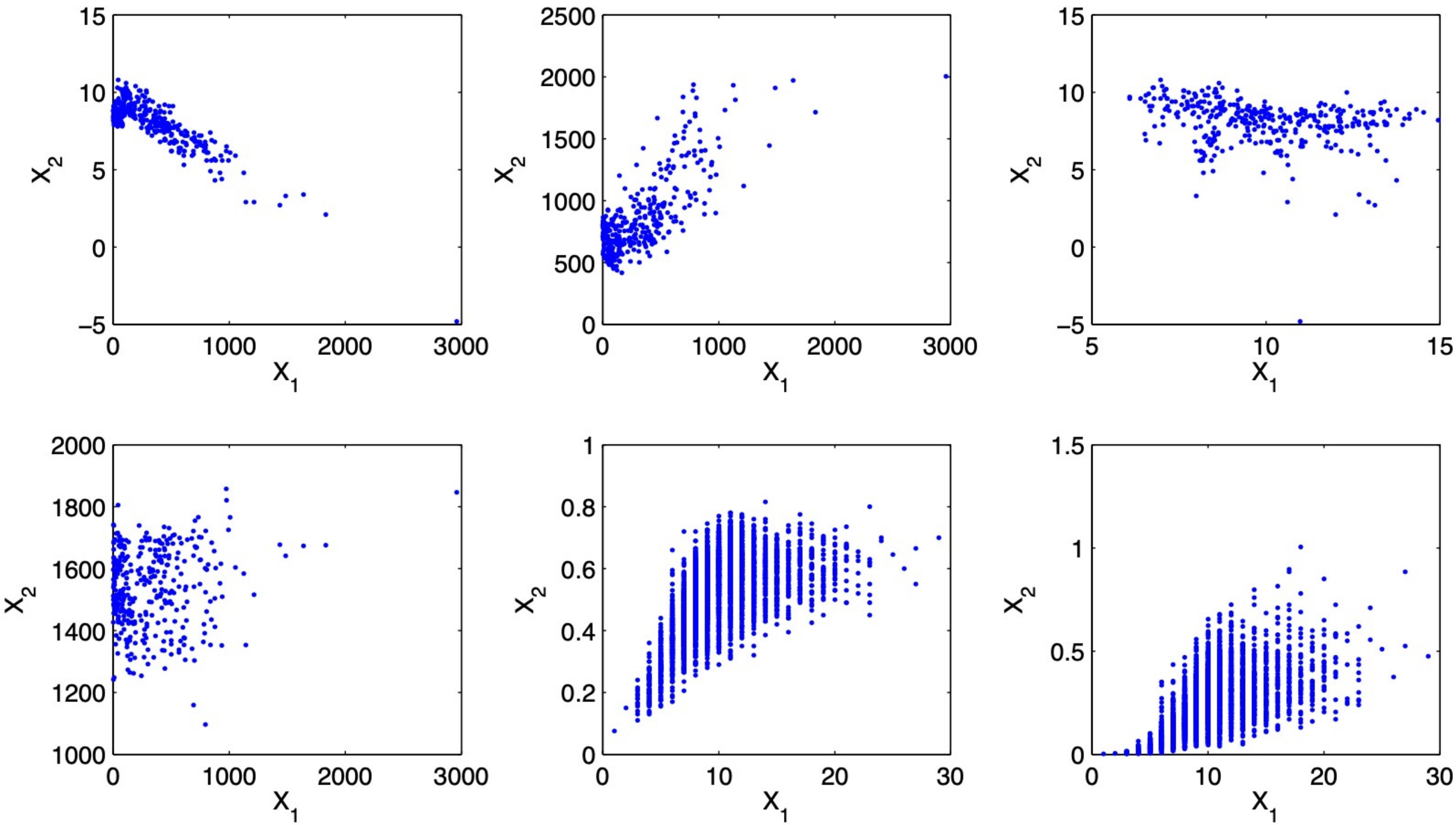
2021-06-08

Outline

- **A linear non-Gaussian model for causal discovery (LiNGAM)**

- Advanced topics

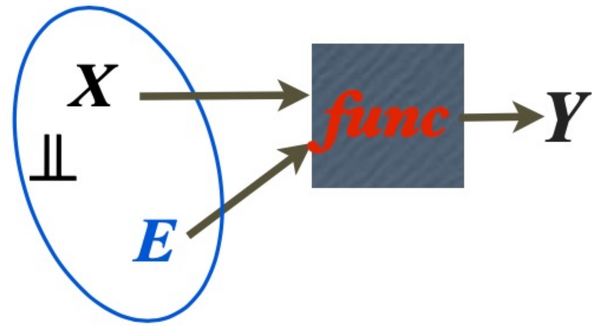
Distinguishing cause from effect



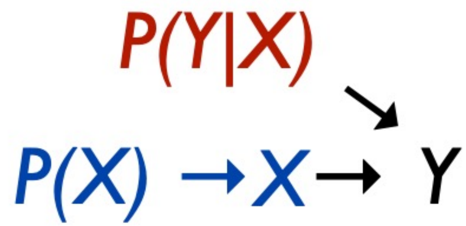
In the two-variable case

- Structural equation model / functional causal model

  $$Y = f(X, E), \text{ where } E \perp\!\!\!\perp X$$



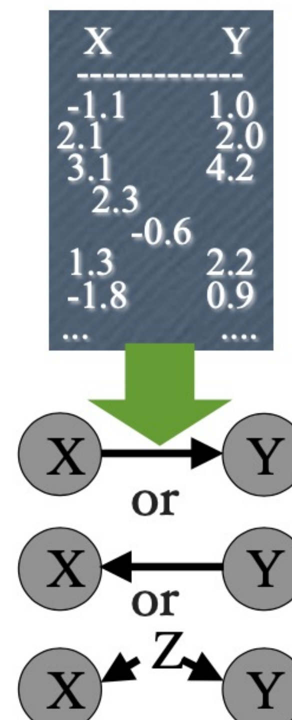
  - Related to this type of "independence":



- Start with the linear case

  $$Y = aX + E, \text{ where } E \perp\!\!\!\perp X$$

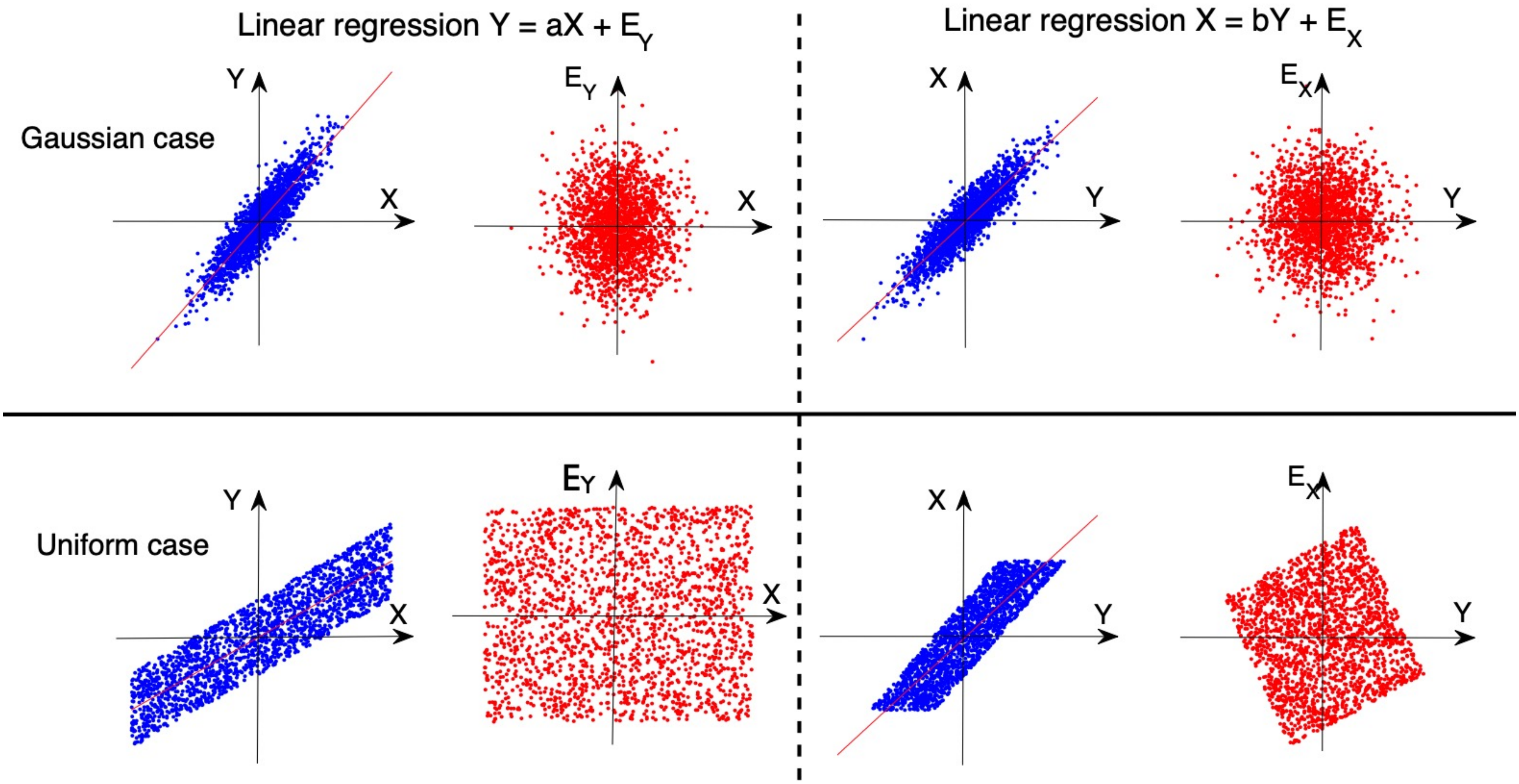
- Determine causal direction in the two-variable case? Identifiability!



Slides from Kun Zhang @CMU

Causal Asymmetry

Data generated by $Y = aX + E$ (i.e., $X \rightarrow Y$):



The linear-Gaussian case is one of the few non-identifiable situations

Darmois-Skitovitch theorem: Define two random variables, Y_1 and Y_2, as linear combinations of independent random variables S_i, $i = 1, \ldots, n$:

$$Y_1 = \alpha_1 S_1 + \alpha_2 S_2 + \ldots + \alpha_n S_n,$$
$$Y_2 = \beta_1 S_1 + \beta_2 S_2 + \ldots + \beta_n S_n.$$

If Y_1 and Y_2 are statistically independent, then all variables S_j for which $\alpha_j \beta_j \neq 0$ are Gaussian.

Generated by $Y = aX + E$ ($X \to Y$): $\begin{bmatrix} 0 & 1 \\ 1 & a \end{bmatrix} \cdot \begin{bmatrix} E \\ X \end{bmatrix} =$

Assuming $Y \to X$ (fitting $X = bY + E_Y$): $\begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} E_Y \\ Y \end{bmatrix} =$

$$\begin{bmatrix} X \\ Y \end{bmatrix} \Rightarrow \begin{bmatrix} E_Y \\ Y \end{bmatrix} = \begin{bmatrix} 1 - ab & -b \\ a & 1 \end{bmatrix} \cdot \begin{bmatrix} E \\ X \end{bmatrix}$$

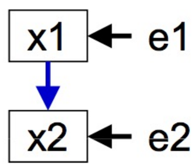
Kagan et al., Characterization Problems in Mathematical Statistics. New York: Wiley, 1973

Linear Structural Equation Models

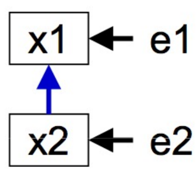
[Wright, 1921; Bollen, 1989]

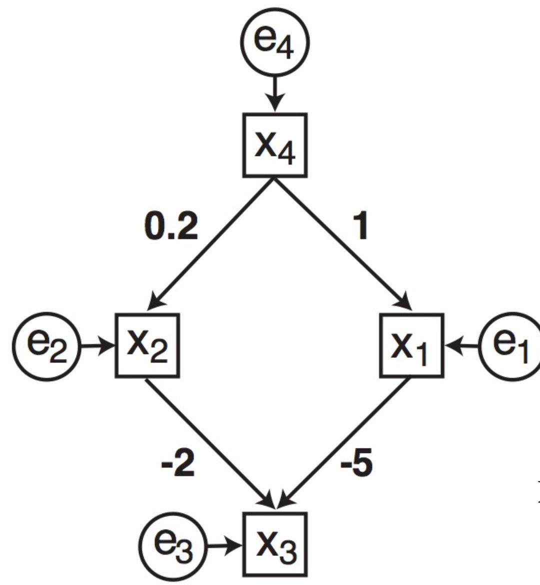
- Use linear SEM to model the data generation process

$$x_1 = e_1$$
$$x_2 = b_{21}x_1 + e_2$$



$$x_1 = b_{12}x_2 + e_1$$
$$x_2 = e_2$$





x4 = e4

x2 = 0.2*x4 +e2

x1 = x4 +e1

x3 = −2*x2 −5*x1 +e3

p(x1,x2,x3,x4) = p(x4) p(x2|x4) p(x1|x4) p(x3|x2,x1)

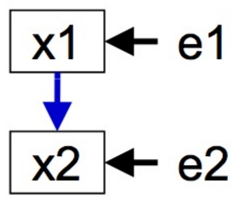
Shimizu et. al., JMLR 2006
Shimizu and Kawahara, UAI2010 Tutorial

Identify Which Model to Generate The Data

- Two models with Gaussian e_1 and e_2 :

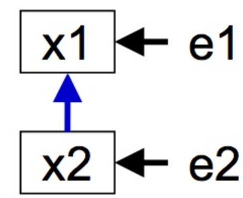
Model 1:

$$x_1 = e_1$$
$$x_2 = 0.8x_1 + e_2$$



Model 2:

$$x_1 = 0.8x_2 + e_1$$
$$x_2 = e_2$$



$$E(e_1) = E(e_2) = 0, \mathrm{var}(x_1) = \mathrm{var}(x_2) = 1$$

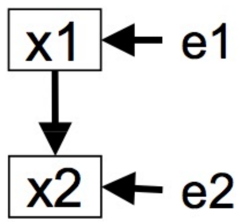
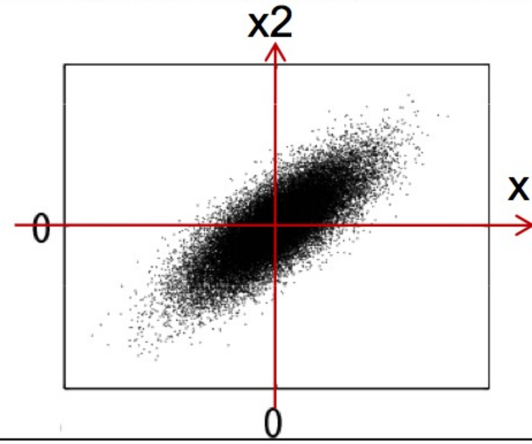
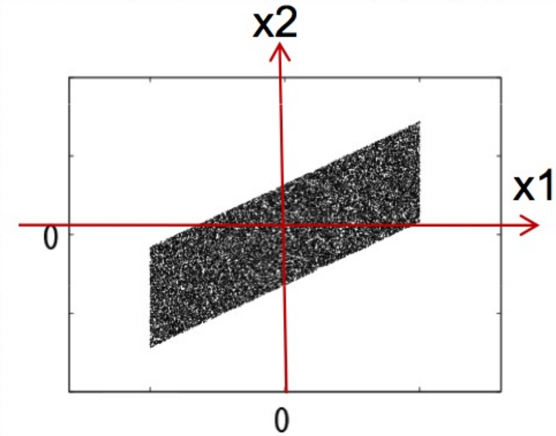
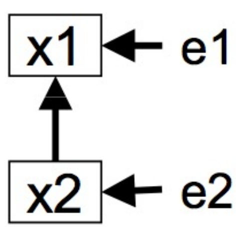
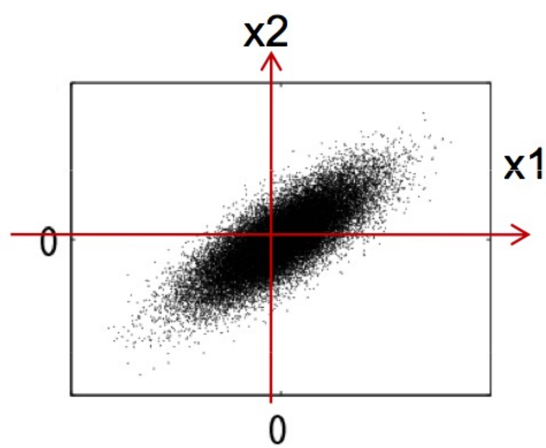
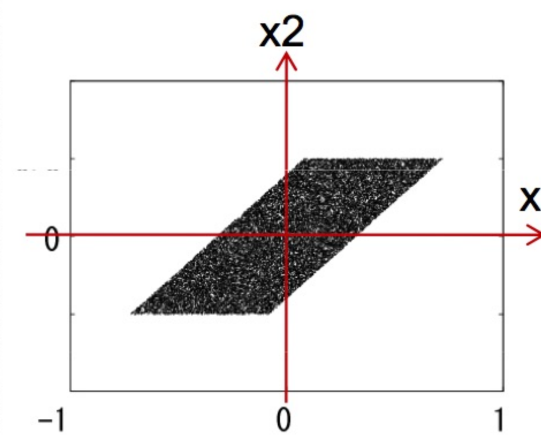
- Both introduce no conditional independence:

$$\mathrm{cov}(x_1, x_2) = 0.8 \neq 0$$

- Both induce the same Gaussian distribution:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right)$$

Gaussian vs. Non-Gaussian

	Gaussian	Non-Gaussian (uniform)
Model 1: $x_1 = e_1$, $x_2 = 0.8x_1 + e_2$ 		
Model 2: $x_1 = 0.8x_2 + e_1$, $x_2 = e_2$ 		

$$E(e_1) = E(e_2) = 0,$$
$$\mathrm{var}(x_1) = \mathrm{var}(x_2) = 1$$

Shimizu and Kawahara, UAI2010 Tutorial

Linear Non-Gaussian Acyclic Model: LiNGAM

- Linear acyclic SEM

[Shimizu, Hyvarinen, Hoyer & Kerminen, JMLR 2006]

$$x_i = \sum_{j:\,\text{parents of } i} b_{ij} x_j + e_i \quad \text{or} \quad \mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

For example:

$$x_1 = 1.5x_3 + e_1$$
$$x_2 = -1.3x_1 + e_2$$
$$x_3 = e_3$$

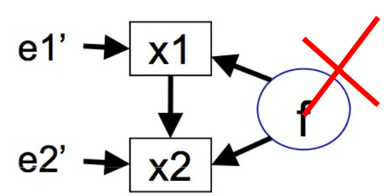
or

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 1.5 \\ -1.3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

Causal Markov condition holds:

$$p(\mathbf{x}) = \prod_{i=1}^{p} p(x_i \mid \text{parents of } x_i)$$

- Assumptions:
  - Directed acyclic graph (DAG): no directed cycles
  - External influences e_i are of non-zero variance, and are independent **non-Gaussian**
  - No latent confounders



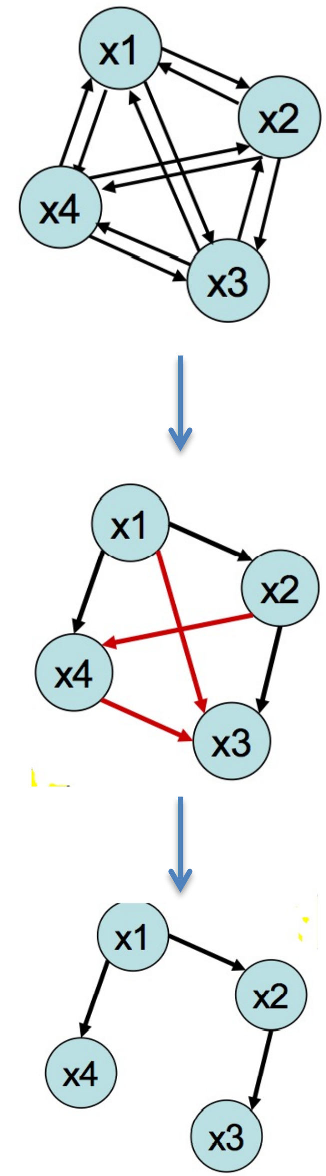
How *B* Is Estimated?

- Step 1: Estimate B by Independent Component Analysis (ICA) with post-processing

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \Leftrightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$
$$= \mathbf{A}\mathbf{e} = \mathbf{W}^{-1}\mathbf{e}$$

- Step 2: Find an order of the variables to get a DAG

- Step 3: Discard non-significant edges



[Shimizu, et. al. Jan 2008]

Performance of the algorithm

- Fast (ICA is fast)
- Possible local optimum problem (ICA is an iterative method)
- A good estimation needs >1000 sample size for >10 variables
- Not scale invariant

ICA-Based LiNGAM: A Real Example

Height	ArmSpan
377	394
374	441
363	309
354	374
352	383
345	363
329	399
327	382
326	386
305	337
294	392
297	292
284	273
274	331
273	303
265	228
256	275
255	276
253	291
254	320
246	278
234	274
227	327
225	288
217	253
215	257
208	258
205	309
204	334
197	217

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \hat{W} \cdot \begin{bmatrix} \text{height} \\ \text{arm span} \end{bmatrix}, \quad \text{where}$$

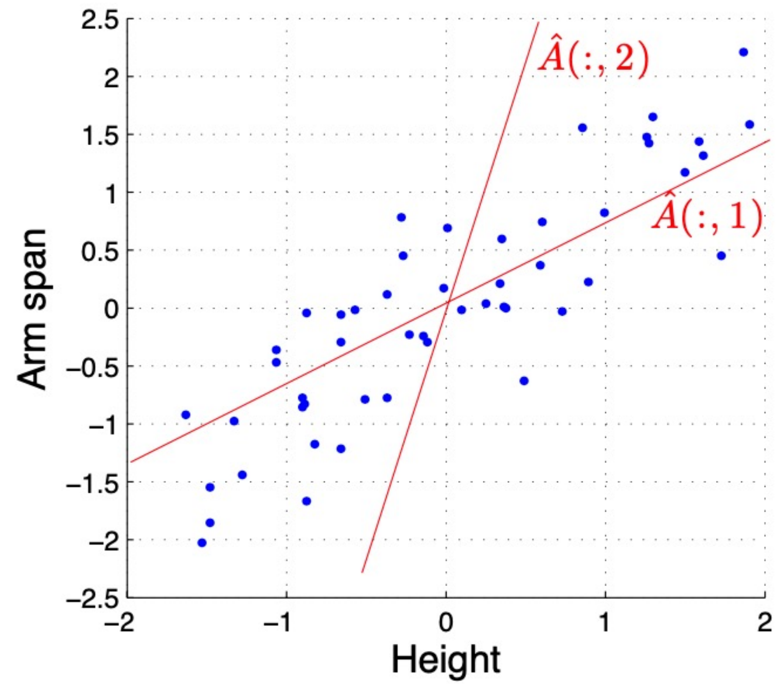
$$\hat{W} = \begin{bmatrix} 1.33 & -0.39 \\ -1.56 & 2.02 \end{bmatrix}, \quad \text{or}$$

$$\hat{A} = \hat{W}^{-1} = \begin{bmatrix} 0.97 & 0.19 \\ 0.75 & 0.64 \end{bmatrix}$$

If $\hat{W}_{1,2} = 0$, then

$$\text{height} = \frac{1}{1.33} Y_1,$$

$$\text{arm span} = \frac{1.56}{2.02}\text{height} + \frac{1}{2.02} Y_2.$$



For small-scale problems, we can compare the dependence between the residual & hypothetical cause in both directions !

Slides from Kun Zhang @CMU

Applications

- Neuroinformatics
  – Brain connectivity analysis (Hyvarinen et al., JMLR, 2010)
- Bioinformatics
  – Gene network estimation (Sogawa et al., ICANN2010)
- Economics(Wan&Tan,2009; Moneta,Entner,Hoyer&Coad,2010)
- Genetics(Ozaki&Ando,2009)
- Environmental sciences(Niyogietal.,2010)
- Physics (Kawahara, Shimizu & Washio, 2010)
- Sociology (Kawahara, Bollen, Shimizu & Washio, 2010)

Outline

- Introduce a linear non-Gaussian model for causal discovery (LinGAM)

- Advanced topics

Some Estimation Methods for LiNGAM

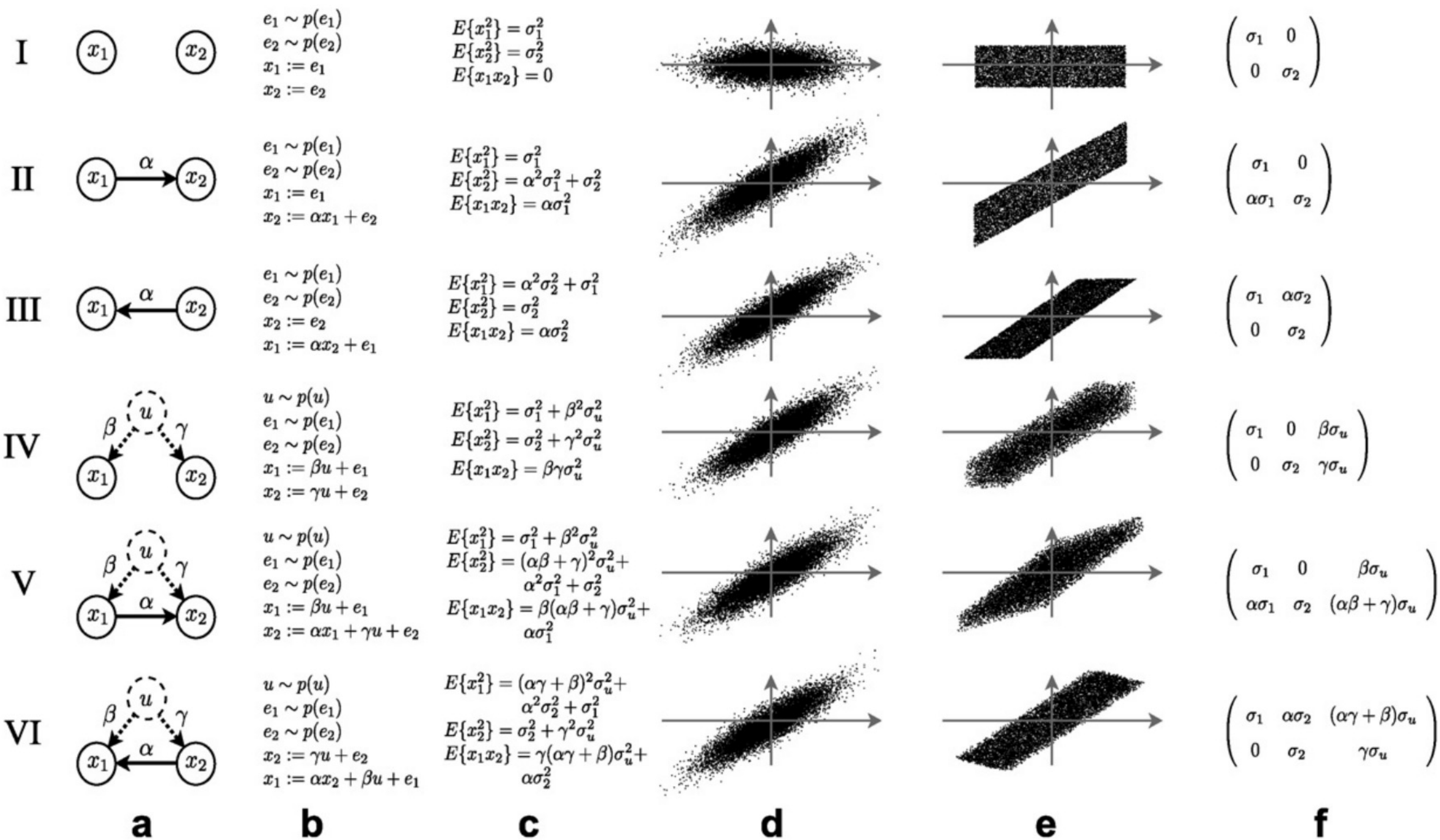
- ICA-LiNGAM
- ICA with Sparse Connections
- DirectLiNGAM...

Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.

Zhang et al. (2006) ICA with sparse connections: Revisited. Lecture Notes in Computer Science, 5441:195– 202, 2009

Shimizu, et al. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 12:1225–1248.

Causal discovery under confounders

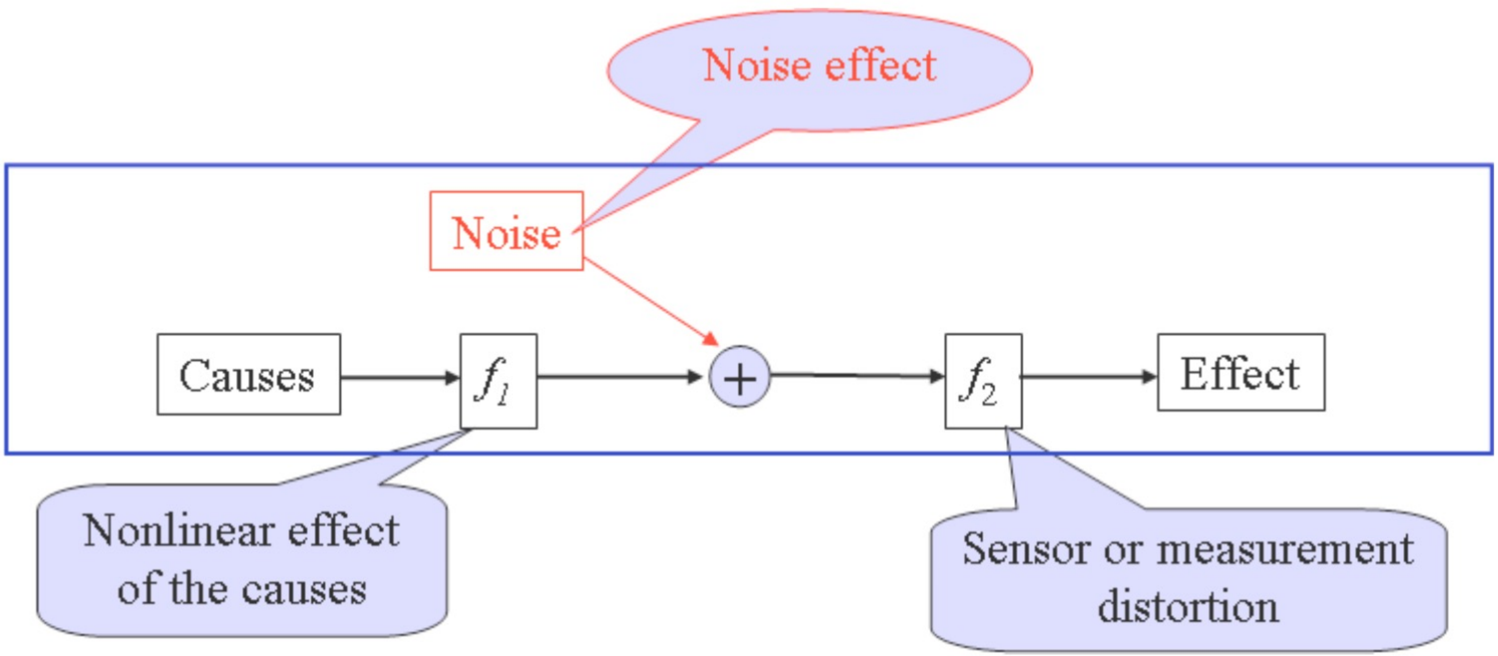


Hoyer et al. (2008). Estimation of causal effects using linear nonGaussian causal models with hidden variables. International Journal of Approximate Reasoning, 49(2):362– 378.

Three Effects usually encountered in a causal model

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

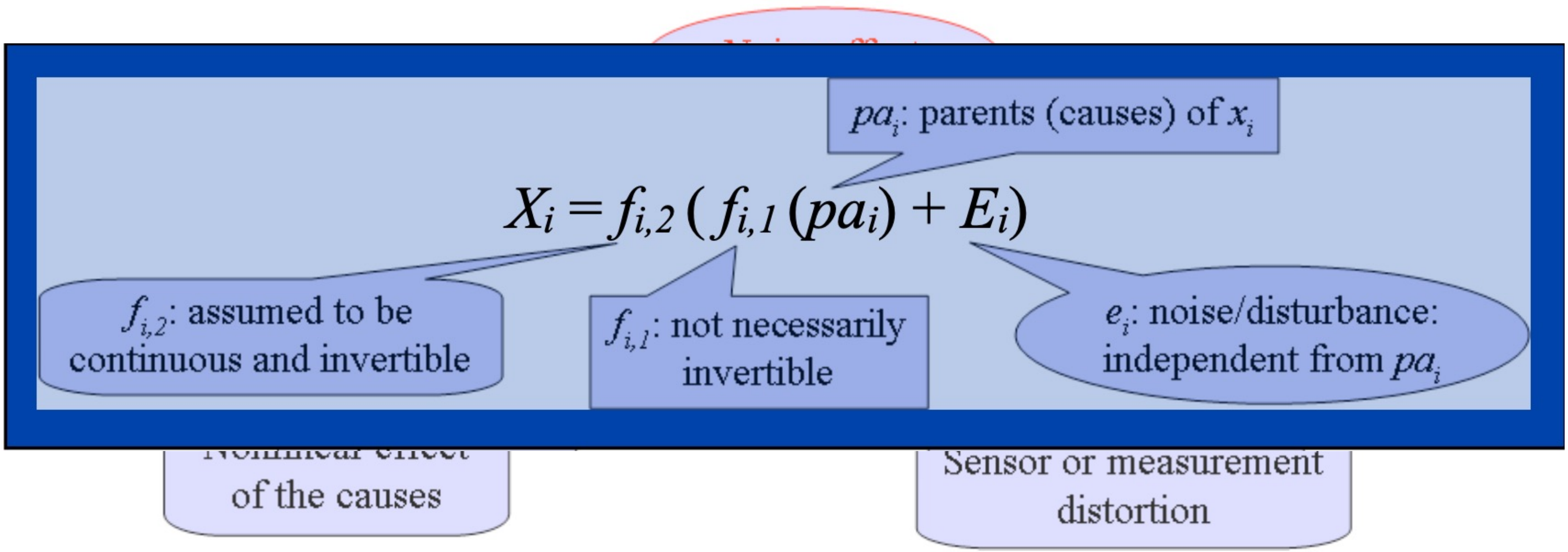


(Zhang & Hyvärinen, '09a)

Post-Nonlinear (PNL) Causal Model

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



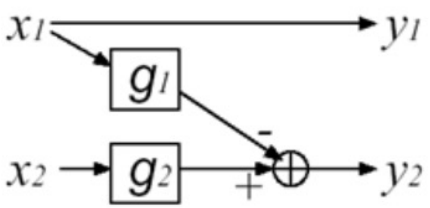
Special cases: linear models; nonlinear additive noise models; multiplicative noise models:

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

(Zhang & Chan, 2006; Zhang & Hyvärinen, '09a)

PNL implemented by MLP

- If $X_1 \to X_2$, 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 $X_1 \to X_2$
  - 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):



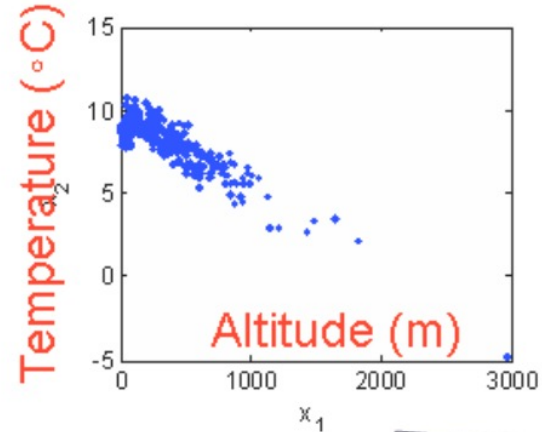
(y_2 produces an estimate of e_2)

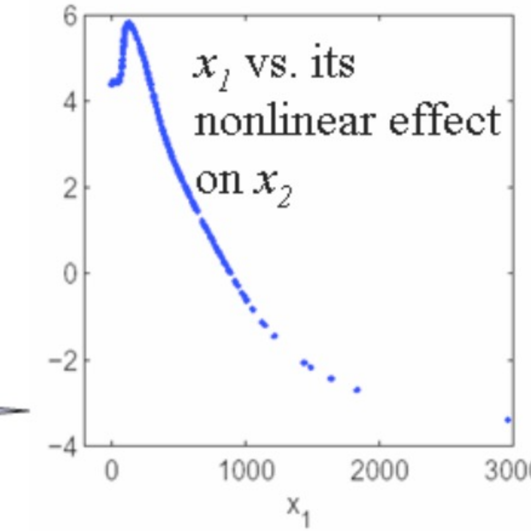
$$
\begin{aligned}
I(X_1, Y_2) &= H(X_1) + H(Y_2) + E\{\log|\mathbf{J}|\} - H(X_1, X_2) \\
&= -E \log p_{Y_2} - E\{\log|g_2'(X_2)|\} + \text{const}
\end{aligned}
$$

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

Data Set 1

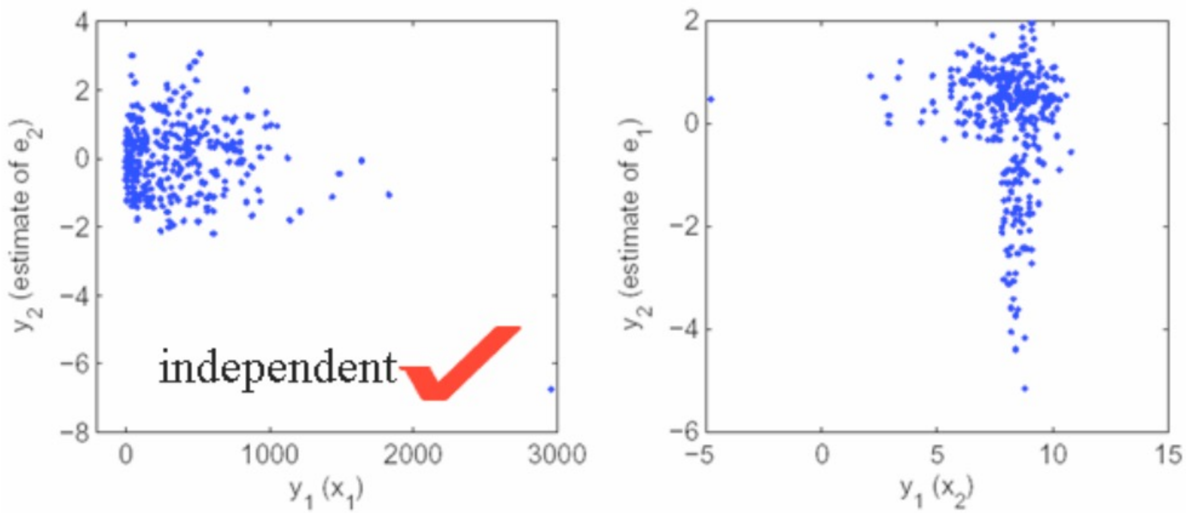
with PNL Model

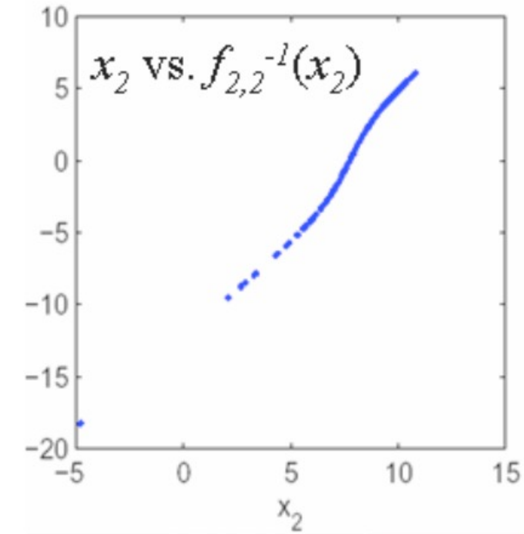




(a) y_1 vs y_2 under hypothesis $x_1 \to x_2$

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





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

Data Set	$x_1 \to x_2$ assumed		$x_2 \to x_1$ 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}

Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa350

CausalMGM: an interactive web-based causal discovery tool

Xiaoyu Ge [1,†], Vineet K. Raghu [1,2,†], Panos K. Chrysanthis [1,*] and Panayiotis V. Benos [1,2,*]

[1]Department of Computer Science, University of Pittsburgh, 4200 Fifth Avenue, Pittsburgh, PA 15260, USA and [2]Department of Computational and Systems Biology, University of Pittsburgh, 3420 Forbes Ave, Pittsburgh, PA 15213, USA

Received March 05, 2020; Revised April 21, 2020; Editorial Decision April 23, 2020; Accepted April 25, 2020

ABSTRACT
High-throughput sequencing and the availability of large online data repositories (e.g. The Cancer Genome Atlas and Trans-Omics for Precision Medicine) have the potential to revolutionize systems biology by enabling researchers to study interactions between data from different modalities (i.e. genetic, genomic, clinical, behavioral, etc.). Currently, data mining and statistical approaches are confined to identifying correlates in these datasets, but re- searchers are often interested in identifying cause- and-effect relationships. Causal discovery methods were developed to infer such cause-and-effect relationships from observational data. Though these algorithms have had demonstrated successes in several biomedical applications, they are difficult to use for non-experts. So, there is a need for web-based tools to make causal discovery methods accessible. Here, we present CausalMGM (http:// causalmgm.org/), the first web-based causal discovery tool that enables researchers to find cause-and- effect relationships from observational data. Web- based CausalMGM consists of three data analysis tools: (i) feature selection and clustering; (ii) automated identification of cause-and-effect relation- ships via a graphical model; and (iii) interactive visualization of the learned causal (directed) graph. We demonstrate how CausalMGM enables an end-to-end exploratory analysis of biomedical datasets, giving researchers a clearer picture of its capabilities.

Causal Discovery from Incomplete Data: A Deep Learning Approach

Yuhao Wang[1], Vlado Menkovski[1], Hao Wang[2], Xin Du[1], Mykola Pechenizkiy[1]
[1]Eindhoven University of Technology, [2]Massachusetts Institute of Technology
{y.wang9, v.menkovski, x.du, m.pechenizkiy}@tue.nl, hoguewang@gmail.com

arXiv:2001.05343v1 [cs.LG] 15 Jan 2020

Abstract
As systems are getting more autonomous with the development of artificial intelligence, it is important to discover the causal knowledge from observational sensory inputs. By encoding a series of cause-effect relations between events, causal networks can facilitate the prediction of effects from a given action and analyze their underlying data generation mechanism. However, missing data are ubiquitous in practical scenarios. Directly performing existing casual discovery algorithms on partially observed data may lead to the incorrect inference. To alleviate this issue, we proposed a deep learning framework, dubbed Imputated Causal Learning (ICL), to perform iterative missing data imputation and causal structure discovery. Through extensive simulations on both synthetic and real data, we show that ICL can outperform state-of-the-art methods under different missing data mechanisms.

A Causal View on Robustness of Neural Networks

Cheng Zhang* [1] Kun Zhang [2] Yingzhen Li* [1]

Abstract

We present a causal view on the robustness of neural networks against input manipulations, which applies not only to traditional classification tasks but also to general measurement data. Based on this view, we design a deep causal manipulation augmented model (deep CAMA) which explicitly models possible manipulations on certain causes leading to changes in the observed effect. We further develop data augmentation and test-time fine-tuning methods to improve deep CAMA's robustness. When compared with discriminative deep neural networks, our proposed model shows superior robustness against unseen manipulations. As a by-product, our model achieves disentangled representation which separates the representation of manipulations from those of other latent causes.

https://arxiv.org/pdf/2005.01095.pdf arXiv:2005.01095v1 [cs.LG] 3 May 2020

Thank you!