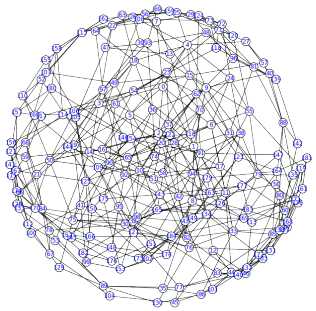
因果推断和因果图模型

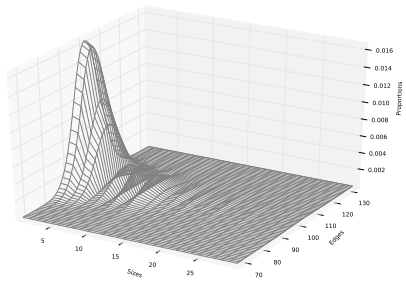
—Causal inference and Causal graphical models

何洋波

北京大学数学科学学院



南京 · 南大
Nov 3, 2018



Outline

Outline

相关未必因果 [1]

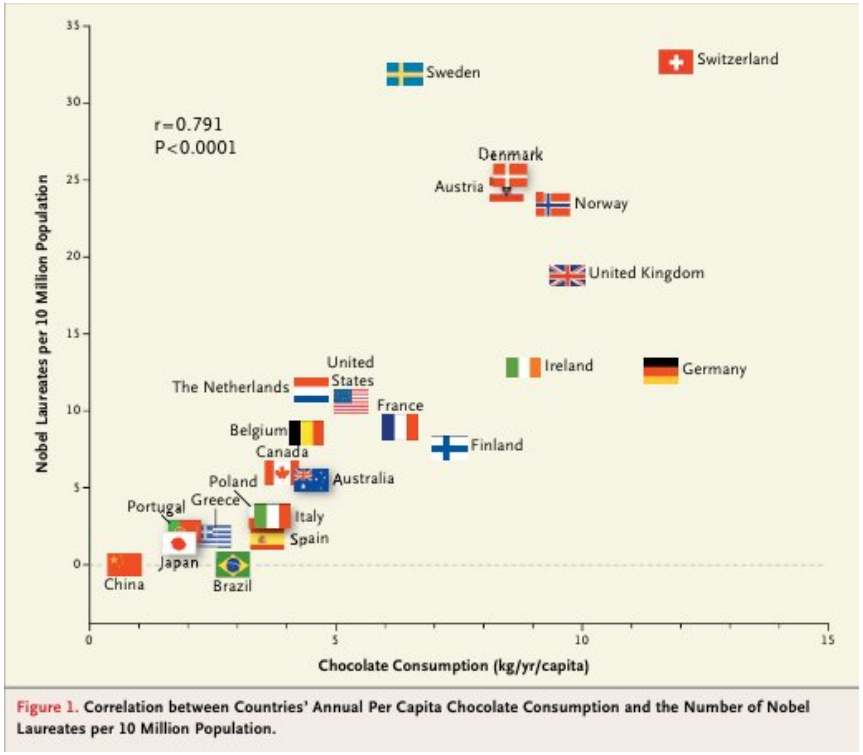
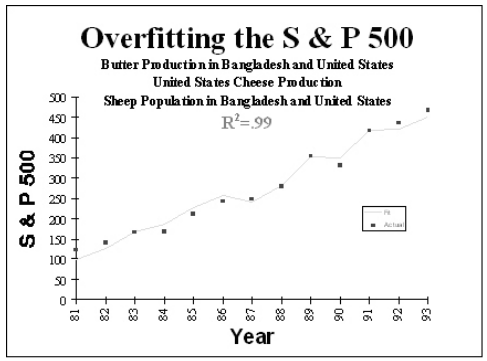
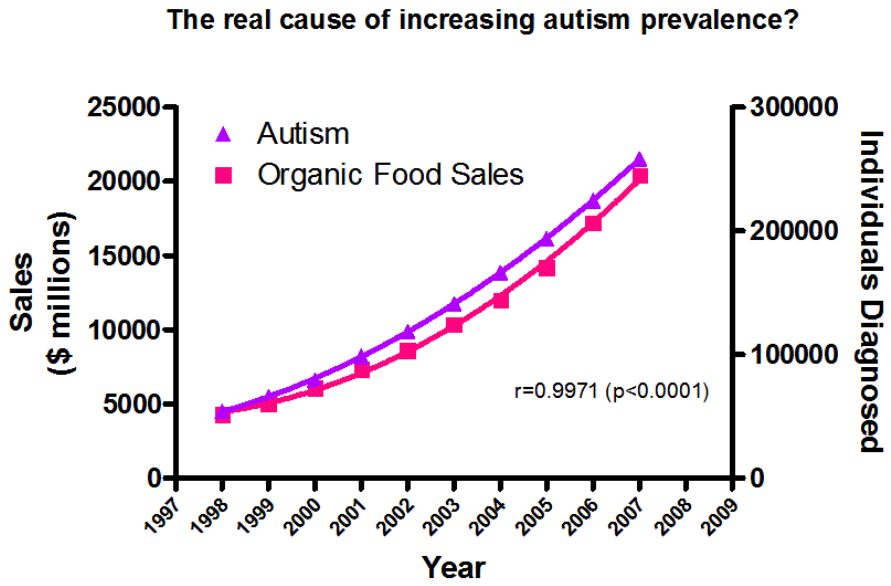


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.



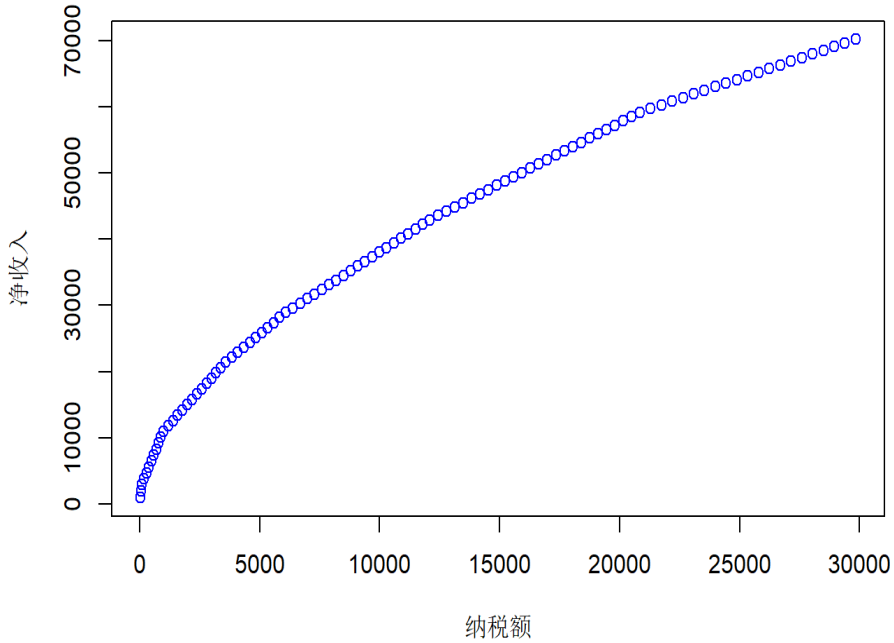


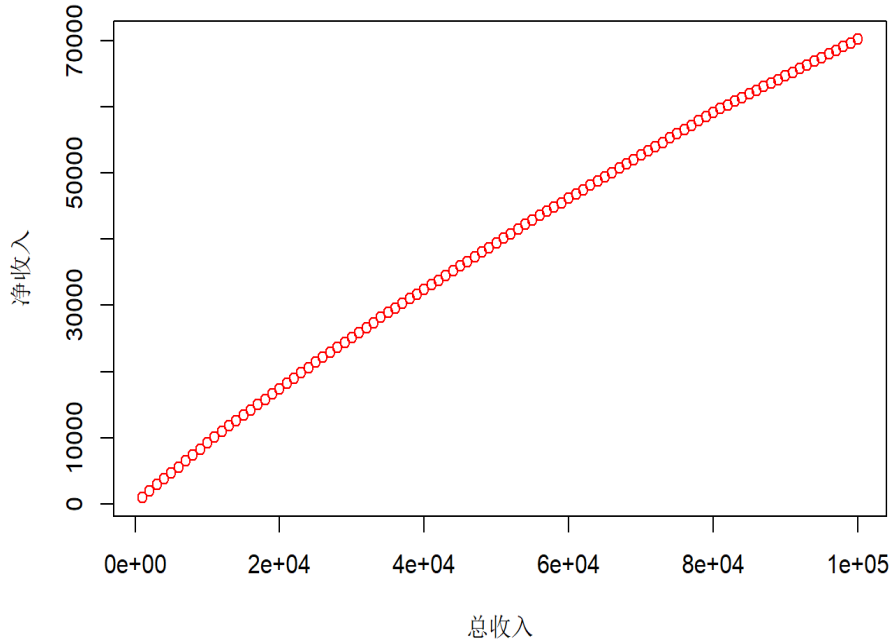
Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

[1] stats.stackexchange.com/questions/36/examples-for-teaching-correlation-does-not-mean-causation

Make a prediction or make a decision

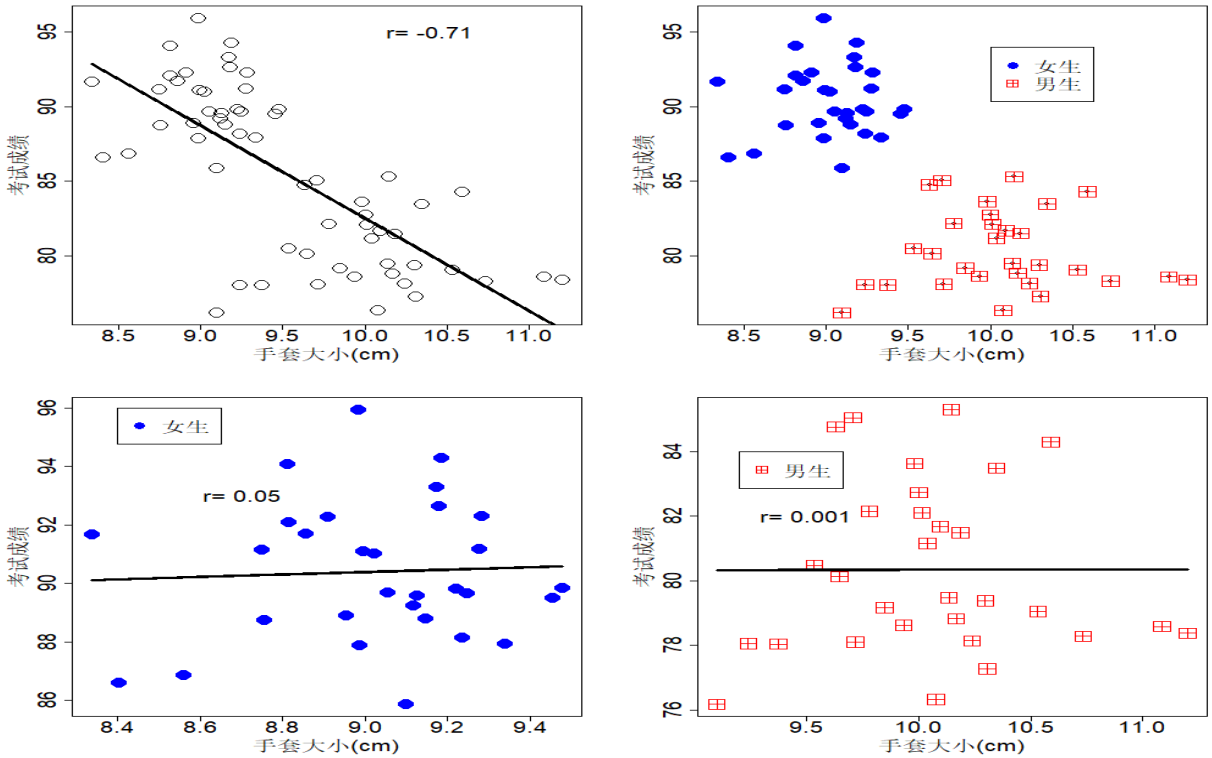
税收可以预测净收入,但是,
提高税收不能提高净收入





Make a prediction or make a decision

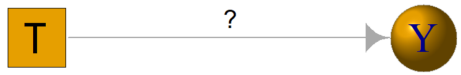
手套大小可以帮助预测成绩,但是,
改变手套大小不太会改变成绩



Causality

- Counterfactuals (Lewis, 1973),Potential Outcomes (Rubin, 1974)
- Intervention (Pearl, 2000)
- Invariance (Hausman & Woodward, 1999; Woodward, 2015a; Peters *et al.*, 2016)

因果–反事实Lewis (1973)，潜在结果模型(Rubin, 1974)



- T为要研究的原因变量（处理）,$T \in \{A, B\}$; Y为相应的结果变量，$Y \in \{0, 1\}$
- T=A为观测(实际发生的),
  T=B为虚拟的（反事实，潜在结果）：
  - 如果不抽烟,
  - 如果接受B治疗,
  - 如果经济危机时不救市,

- Y_A为实际观测的结果，Y_B为虚拟的结果
- 个体水平因果效应

$$Y_{i,B} - Y_{i,A}$$

- 群体平均因果效应

$$E(Y_B) - E(Y_A)$$

因果–随机实验

- 很多情况下，个体因果效应是不可识别的， 而平均因果效应$E(Y_B) - E(Y_A)$可能能识别
- 通常情况下：

$$P(Y_B) \neq P(Y = 1|T = B).$$

- 完全随机实验，从总体中随机选一部分接受处理B，另一部分接受处理A。因为 $T \perp\!\!\!\perp \{Y_A, T_B\}$,有 (Rubin, 1974)

$$E(Y_B) = P(Y = 1|T = B)$$

- 机器学习方法估计因果效应$E(Y_B) - E(Y_A)$(Athey & Imbens, 2016; Pierre & Jean, 2017),在假设$T \perp\!\!\!\perp \{Y_A, T_B\}|X$下，

$$E(Y_B|X) = P(Y = 1|T = B, X)$$

$$E(Y_A|X) = P(Y = 1|T = A, X)$$

Intervention

- 实验，干预
- 因果关系，$P(Y|do(T=A)) \neq P(Y|do(T=B))$ (Pearl, 2000)
- 通常 $P(Y|do(T=A) \neq P(Y|T=A)$
- Intervention and causation

> Woodward (2015b)
>
> C causes E if and only if it is possible to intervene to change the value of C, in such a way that if that intervention were to occur, the value of E or the probability distribution of E would change.

- Types of interventions
  - hard intervention $do(T := A)$
  - soft intervention $do(T := X)$,X 是一个随机变量

Invariance

- 改变（干预）一个变量的因果关系，不会改变其它变量的因果关系 (Hausman & Woodward, 1999)
- 通过多个实验数据集，或不同环境下的数据，利用因果的不变性进行因果推断。
- Causal inference by using invariant prediction (Peters *et al.*, 2016, 2017)

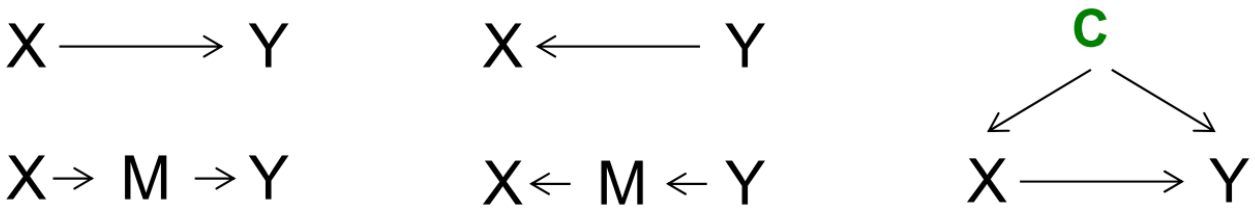
Idea(Peters *et al.*, 2016)

"If we consider all 'direct causes' of a target variable of interest, then the conditional distribution of the target given the direct causes will not change when we interfere experimentally with all other variables in the model except the target itself."

- The Principle of Independent Mechanisms (Peters *et al.*, 2017), 如果$T \to Y$，则改变 $f(T)$, 不会改变 $f(Y|T)$。

观测数据和因果

- 实验数据的不足
  - 实验代价大，比如跨度时间长，费用高
  - 有些问题，比如和人相关的问题或社会问题，也不能进行实验，比如抽烟，有害物质伤害，社会事件等
- 观测数据—相关—因果
- 通过变量之间的相关性，我们可以判断因果可能出现的情形(部分)(Reichenbach, 1956)

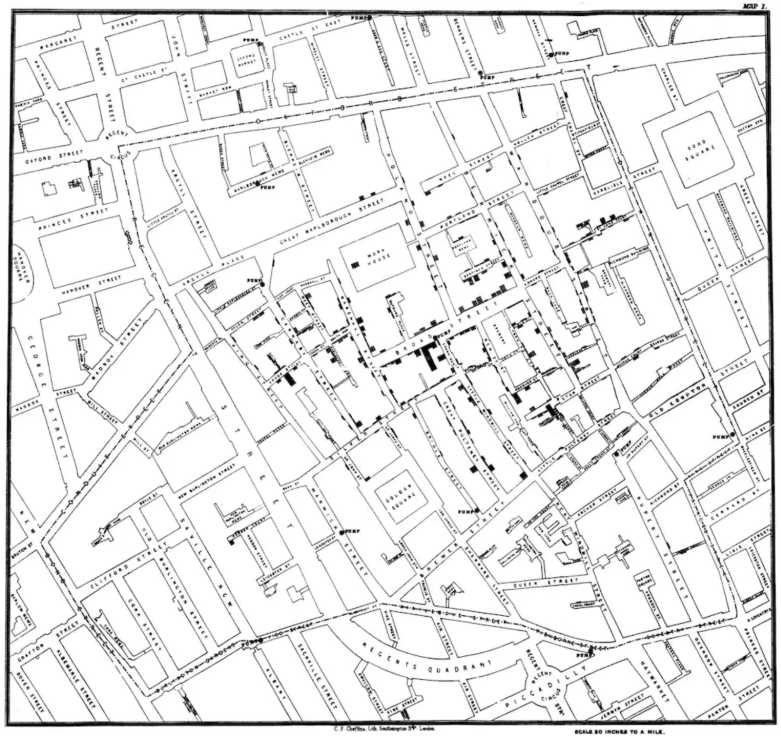


- 判断因果关系存在的条件
  - 相关性存在
  - 原因的改变在结果之前
  - 相关性并不是由别的变量导致
  - 因和果之间有实际的联系和解释，因果发生作用的方式

观测数据和因果

发现相关关系有助于获得因果知识

- 1854年伦敦霍乱，流行的观点是霍乱是通过空气传播的，
- John Snow 统计每户病亡人数， 大多数病例的住所都围绕在Broad Street水泵附近，结合其他证据得出饮用水传播的结论（观测研究，发现相关性）
- 移掉了Broad Street水泵的把手，霍乱最终得到控制 (实验确认)

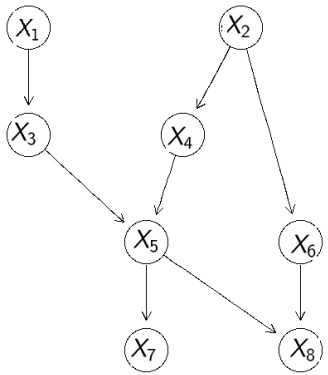


Outline

Causal Graphical Models

A causal Graph + Statistical model

- Vertices represent random variables (processes); an arrow indicates the direction of causation.
- Causal graphical models of DAGs (Causal Bayesian network)



$$X_i = f(pa_i, \varepsilon_i), \quad i = 1, \cdots, n$$

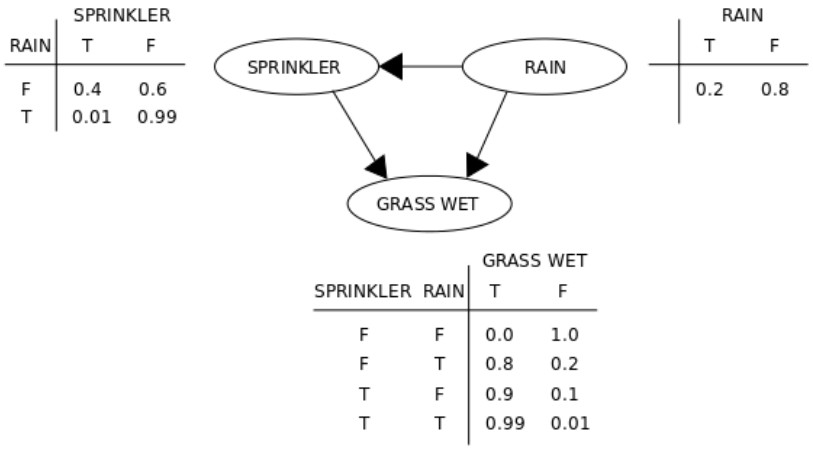
$$\varepsilon_i, \quad \text{independent}$$

- Data generation mechanism; deterministically with omitted variables (Woodward, 2003).
- Compare to potential outcomes, $Y_{X=B}$, $Y_{X=A}$.

Causal Graphical models of DAGs

- Simple statistical analyses.(Spiegelhalter & Lauritzen, 1990; Lauritzen & Spiegelhalter, 1988)
  - Causal Markov: Conditional on its parents, every variable is (probabilistically) independent of every other except its descendants.
  - Convenient recursive factorizations of their joint probability density functions

$$p(x) = \prod_{v \in V} p\left(x_v | x_{pa(v)}\right)$$



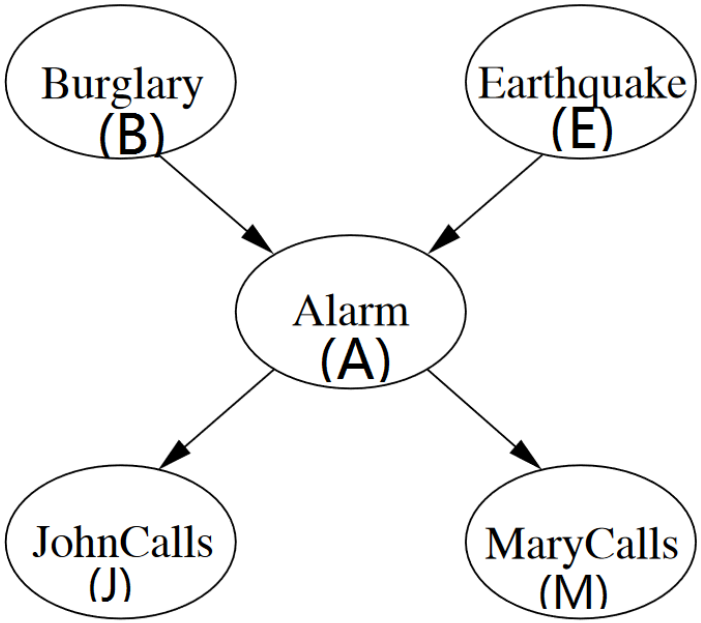
$$P(G, S, R) = P(G|S, R)P(S|R)P(R)$$

$$P(R = T|G = T) = \frac{P(G = T, R = T)}{P(G = T)}$$

DAG and joint distribution—Markov property

- G a DAG over $X = \{X_1, \cdots, X_p\}$; f, a probability distribution of random variables X,
- Conditionally independence: $X_1 \perp\!\!\!\perp X_2 | X_3$
- A probability distribution f satisfies the local Markov property w.r.t. a DAG G

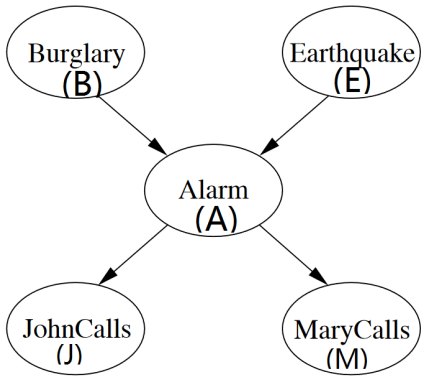
$$\forall i, \quad X_i \perp\!\!\!\perp \{nd(X_i) \setminus pa(X_i)\} | pa(X_i)$$



- $B \perp\!\!\!\perp E$,
- $(B, E) \perp\!\!\!\perp (J, M) | A$
- $J \perp\!\!\!\perp M | A$

DAG and joint distribution —Faithfulness Assumption

- A distribution P is *faithful* to a DAG G if no Conditional Independent relations other than the ones entailed by the Markov property are present



有且仅有如下独立性

- $B \perp\!\!\!\perp E$,
- $(B, E) \perp\!\!\!\perp (J, M) | A$
- $J \perp\!\!\!\perp M | A$

- 理论上，忠实性假定被破坏的概率为0。
- 在考虑样本随机误差的情形下， Uhler *et al.* (2013) 研究了Gaussian 图 模型的忠实性。当变量个数增加时，不可避免出现不忠实的情形。

Dependencies can not identify all causations —Markov equivalent DAGs

A fundamental problem

Different DAGs may determine the same statistical model

- Different DAGs may encode the same conditional independencies

Examples

DAGs	Independencies (with faithfulness assumption)
$x \to y$, $x \leftarrow y$	$x \not\perp\!\!\!\perp y$
$x \leftarrow y \leftarrow z$ $x \to y \to z$ $x \leftarrow y \to z$	$x \perp\!\!\!\perp z \mid y$, $x \not\perp\!\!\!\perp y$, $y \not\perp\!\!\!\perp z$, $x \not\perp\!\!\!\perp z$

- Two DAGs are Markov equivalent iff they entail the same conditional independencies.

Markov equivalence class

Markov equivalence class

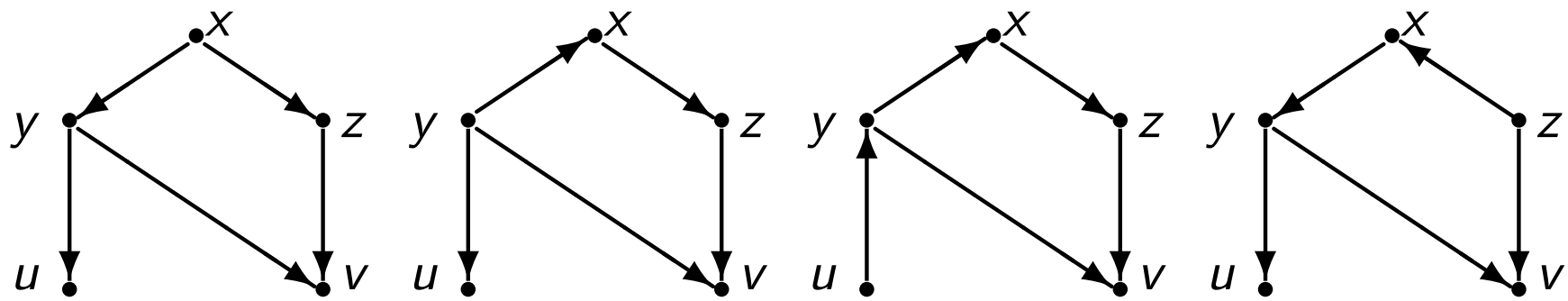
A Markov equivalence class contains all Markov equivalent DAGs.

- Example 1, $x \not\perp\!\!\!\perp y$

$$x \to y, x \leftarrow y$$

- Example 2 the set of independencies:

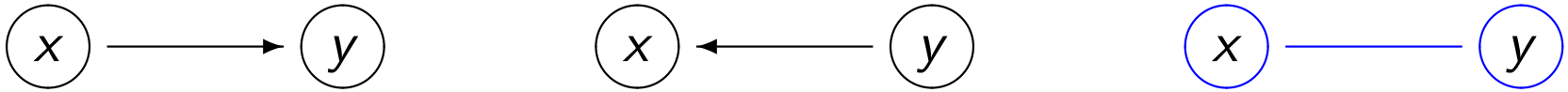
$$\{u \perp\!\!\!\perp (x, z, v) | y, z \perp\!\!\!\perp (y, u) | x, x \perp\!\!\!\perp (u, v) | (y, z), z \not\perp\!\!\!\perp y | (x, v), z \not\perp\!\!\!\perp u | (x, v)\}$$



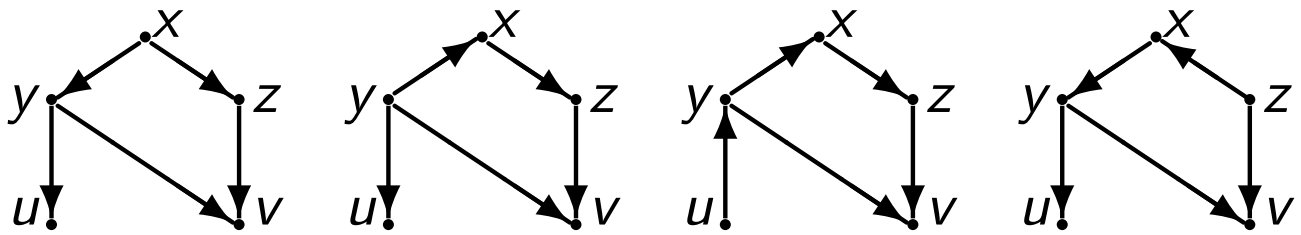
G_0 G_1 G_2 G_3

Representation of Markov equivalence class

- Essential graph, (Andersson *et al.* (1997),AoS)
- Example 1,



- Example 2



Markov equivalence class

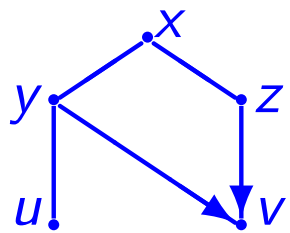
$\{G_1, G_2, G_3, G_4\}$

represent

$\rightleftarrows$

recover

(no v-structure and cycle)



essential graph

Data $\Rightarrow$ Causal graph

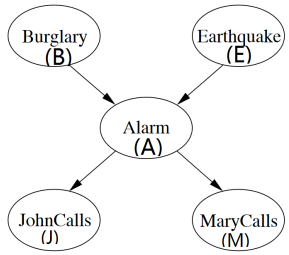
- How to learn (search, estimate) an optimal causal graph from the space of causal structures based on observational data? Non-Bayesian or Bayesian methods

Probabilistic dependencies	DAGs (with several assumptions)	Identifiable equivalence class
$x \perp\!\!\!\perp y$, $y \perp\!\!\!\perp z$, $x \perp\!\!\!\perp z$	$x \quad y \quad z$	$x \quad y \quad z$
$x \not\perp\!\!\!\perp y$	$x \rightarrow y \quad z$ $x \leftarrow y \quad z$	$x - y \quad z$
$\vdots$	$\vdots$	$\vdots$
$x \perp\!\!\!\perp z \mid y$, $x \not\perp\!\!\!\perp y$, $y \not\perp\!\!\!\perp z$, $x \not\perp\!\!\!\perp z$	$x \leftarrow y \leftarrow z$ $x \rightarrow y \rightarrow z$ $x \leftarrow y \rightarrow z$	$x - y - z$
$\vdots$	$\vdots$	$\vdots$

Bayesian network structure learning

- Expert knowledge
- Constrained-based learning, [PC algorithm, Spirtes *et al.* (2000)]
  - $B \perp\!\!\!\perp E$,
  - $(B, E) \perp\!\!\!\perp (J, M) | A$
  - $J \perp\!\!\!\perp M | A$

$\Longrightarrow$



- Score-based algorithms [Chickering (2002)]
  - data $\mathcal{D}$, DAG G, score, space of DAGs $\mathcal{S}$
  - $\text{Score}_{BDeu}(G, \mathcal{D}) = \log \prod_{i=1}^{n} \kappa^{(r_i - 1) q_i} \prod_{j=1}^{q_i} \frac{\Gamma\left(\frac{\eta}{q_i}\right)}{\Gamma\left(\frac{\eta}{q_i} + N_{ij}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(\frac{\eta}{r_i q_i} + N_{ijk}\right)}{\frac{\eta}{r_i q_i}}$
  - $G^* = \text{argmax}_{G \in \mathcal{S}} \text{Score}(G, \mathcal{D})$
- Hybrid algorithms, MMHC algorithm Tsamardinos *et al.* (2006)

Two spaces-DAGs vs Classes

- Space of DAGs
  - The space of DAGs with p vertices contains all DAGs with p vertices, denoted by $\mathcal{B}_p$
  - $\mathcal{B}_2$ is composed of three DAGs:

$$\underline{x_1 \quad x_2}, \qquad \underline{x_1 \rightarrow x_2}, \qquad \underline{x_1 \leftarrow x_2}$$

- Space of Markov equivalence classes
  - The space of Markov equivalence classes contains with p vertices contains all Essential graphs with p vertices, denoted by $\mathcal{E}_p$
  - $\mathcal{E}_2$ is composed of two essential graphs:

$$\underline{x_1 \quad x_2}, \qquad \underline{x_1 - x_2}$$

Two types of model spaces

Table: Graphical models based on DAGs or Markov equivalence classes, (Gillispie & Perlman, 2002; Robinson, 1973, 1977; Steinsky, 2003; Gillispie, 2006)

变量个数	全部DAG空间模型数	全部等价类空间模型数	比例
1	1	1	1.00000
2	3	2	0.66667
3	25	11	0.44000
4	543	185	0.34070
5	29281	8782	0.29992
6	3781518	1067825	0.28238
7	1138762420	312510571	0.27443
8	783705491725	212133402500	0.27068
9	1213426273025934	326266056291213	0.26888
10	4175163455710941184	1118902054495975141	0.26799

Main problems

- The properties of causal graphical model space
- Estimate causal effects or the bound of causal effects from observational data or interventional data,eg.
  - Model averaging of causal Bayesian networks
  - Overall, direct, indirect causal effects
- Learn causal structures from observational data or interventional data,eg.
  - Learn Markov equivalence class from observational data
  - Learn causal DAG from observational data and interventional data
  - Active learning of causal DAG from interventions

Some works about causal learning

Sampling

- He Yangbo, Jia Jinzhu, and Yu Bin. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. The Annals of Statistics, 41(4), 1742-1779, 2013.(He *et al.*, 2013)

Counting

- He, Yangbo and Jia, Jinzhu and Yu, Bin. Counting and Exploring Sizes of Markov Equivalence Classes of Directed Acyclic Graphs. Journal of Machine Learning Research, 16, 2589–2609, 2015.(He *et al.*, 2015)
- Formulas for Counting the Sizes of Markov Equivalence Classes of Directed Acyclic Graphs, Working paper.

Causal Inference and Learning

- Liu Yue, He Yangbo and Geng zhi, Causal inference of direct and indirect effects from observational data, working paper.
- He Yangbo, Geng Zhi. Active learning of causal networks with intervention experiments and optimal designs. Journal of Machine Learning Research, 9, 2523-2547, 2008.(He & Geng, 2008)

Outline

Sampling from a space of Markov equivalence classes

Motivations

- Understanding the model space,
- Bayesian learning of graphical models

- How many DAGs in an equivalence class? (Size)
- How many undirected edges?
- How many undirected subgraphs in the class?
- Posterior distribution of Markov equivalence classes or posterior probability of a given causal structure
- Bayesian model average among Markov equivalence class models

Challenges of Sampling

1. Hard to calculate the sampling distribution
2. Complexity for Markov equivalence classes with (hundreds) thousands of vertices

Our Method

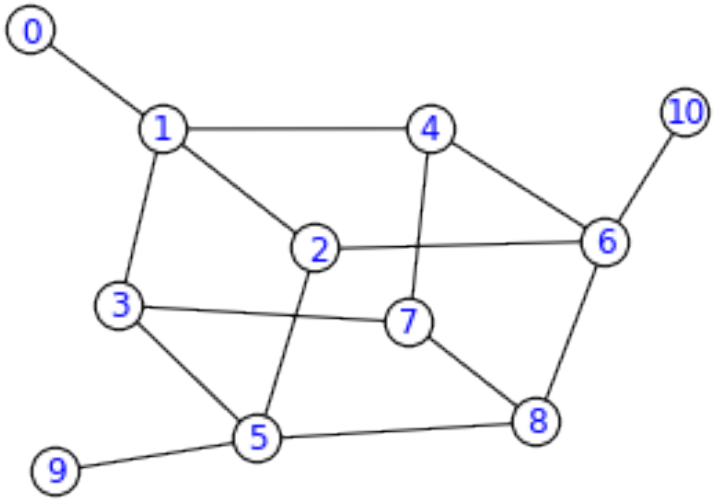
1. Introducing a finite, reversible, and irreducible Markov chain on Markov equivalence classes of interest.
2. Moving among classes via choosing local operators carefully.

Operators and the conditions

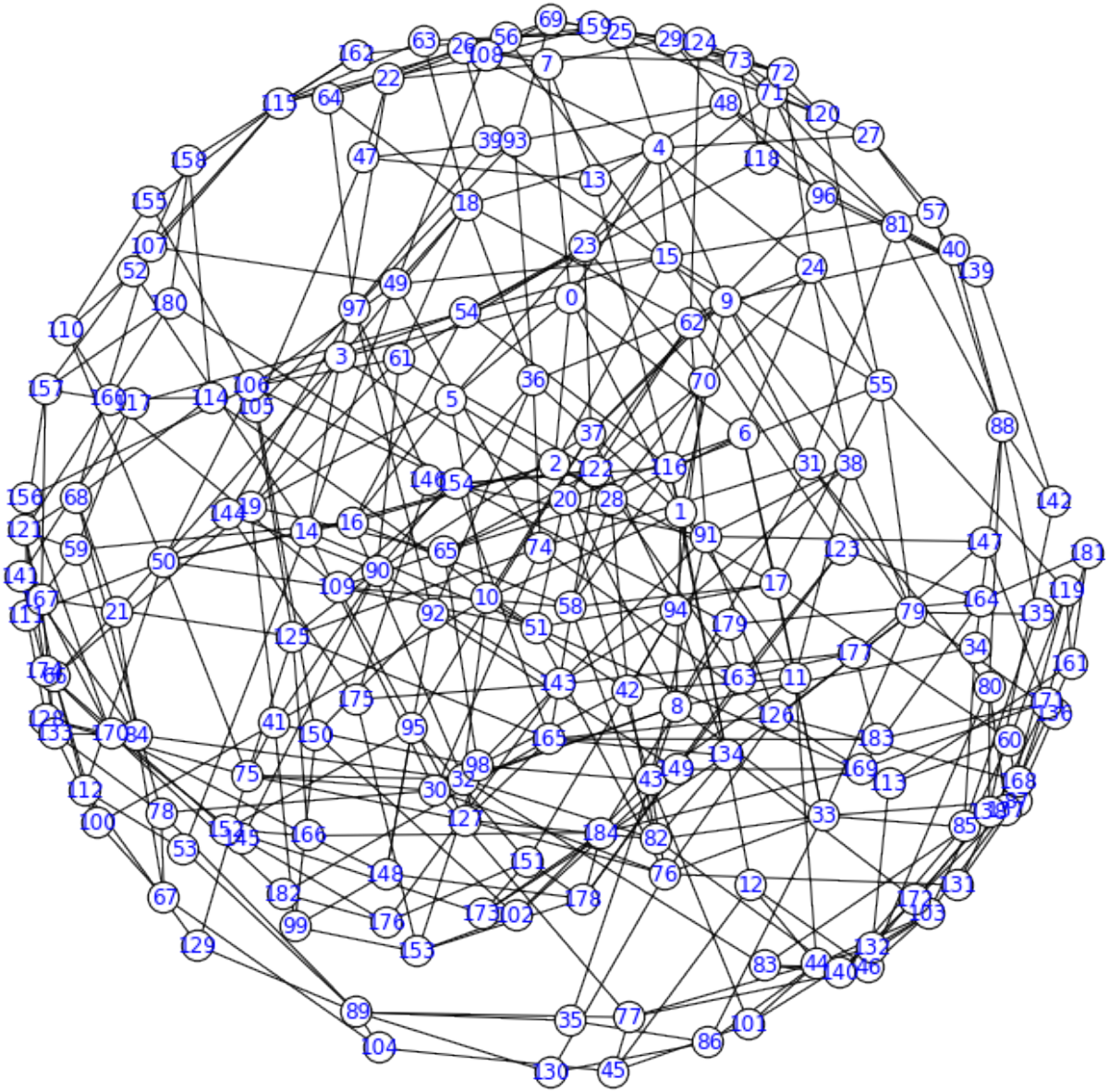
- Let $\mathcal{C}$ be a completed PDAG. six types of operators on $\mathcal{C}$:
  1. inserting an undirected edge, $X \quad Y \Rightarrow X - Y$
  2. deleting an undirected edge, $X - Y \Rightarrow X \quad Y$
  3. inserting a directed edge, $X \quad Y \Rightarrow X \to Y$
  4. deleting a directed edge, $X \to Y \Rightarrow X \quad Y$
  5. making a v-structure, $X - Y - Z \Rightarrow X \to Y \leftarrow Z$
  6. removing a v-structure, $X \to Y \leftarrow Z \Rightarrow X - Y - Z$
- Four conditions of operators
  1. Validity
  2. Distinguishability
  3. Reversibility
  4. Irreducibility

MCMC on $\mathcal{S}_3$

Markov equivalence classes with 3 vertices, 11 EGs



MCMC on $\mathcal{S}_4$, 185 EGs



(Sampling) Size distributions, $\mathcal{E}_6$ and $\mathcal{E}_{10}$

p=6, Steps=10^5, Time:6×10^{-4}s		
Size	GP-value	mean(std)
1	0.28667	0.28588(0.00393)
2	0.25858	0.25897(0.00299)
3	0.17064	0.17078(0.00248)
	...	
28	0.00017	0.00017(0.00004)
30	0.00169	0.00170(0.00017)
32	0.00236	0.00238(0.00017)
36	0.00052	0.00053(0.00008)
38	0.00034	0.00035(0.00004)
40	0.00118	0.00120(0.00010)
42	0.00051	0.00052(0.00009)
48	0.00013	0.00013(0.00004)
50	0.00034	0.00034(0.00007)
52	0.00017	0.00018(0.00003)
54	0.00017	0.00018(0.00004)
60	0.00019	0.00020(0.00004)
72	0.00006	0.00006(0.00002)
88	0.00004	0.00004(0.00001)
144	0.00009	0.00009(0.00003)
156	0.00006	0.00006(0.00003)
216	0.00001	0.00001(0.00002)

p=10, Time: 2×10^{-3}s		
	Steps=10^5	Steps=10^6
Size	mean(std)	mean(std)
1	0.27062(0.00499)	0.2716746(0.00202)
2	0.25931(0.00406)	0.2581884(0.00123)
3	0.16646(0.00306)	0.1656770(0.00161)
4	0.11856(0.00288)	0.1182577(0.00115)
5	0.01613(0.00091)	0.0165379(0.00044)
6	0.04116(0.00128)	0.0416083(0.00061)
7	0.00009(0.00006)	0.0000861(0.00003)
8	0.04765(0.00157)	0.0478519(0.00063)
9	0.00084(0.00015)	0.0008667(0.00006)
10	0.03128(0.00141)	0.0317196(0.00035)
12	0.01173(0.00068)	0.0115982(0.00030)
13	0.00602(0.00045)	0.0060431(0.00013)
14	0.00718(0.00053)	0.0072707(0.00017)
15	0.00085(0.00015)	0.0007971(0.00004)
16	0.00458(0.00032)	0.0045318(0.00013)
17	0.00085(0.00015)	0.0008279(0.00003)
18	0.00187(0.00021)	0.0018126(0.00006)
19	0.00010(0.00004)	0.0000899(0.00001)
20	0.00308(0.00029)	0.0031347(0.00011)
21	0.00005(0.00003)	0.0000601(0.00001)
22	0.00049(0.00010)	0.0004869(0.00004)
23	0.00002(0.00001)	0.0000134(0.00000)
24	0.00142(0.00021)	0.0013617(0.00009)

Table: Size distributions and their estimations for $p = 6$ and 10. Time is the seconds used in each step.

Proportion of directed edges

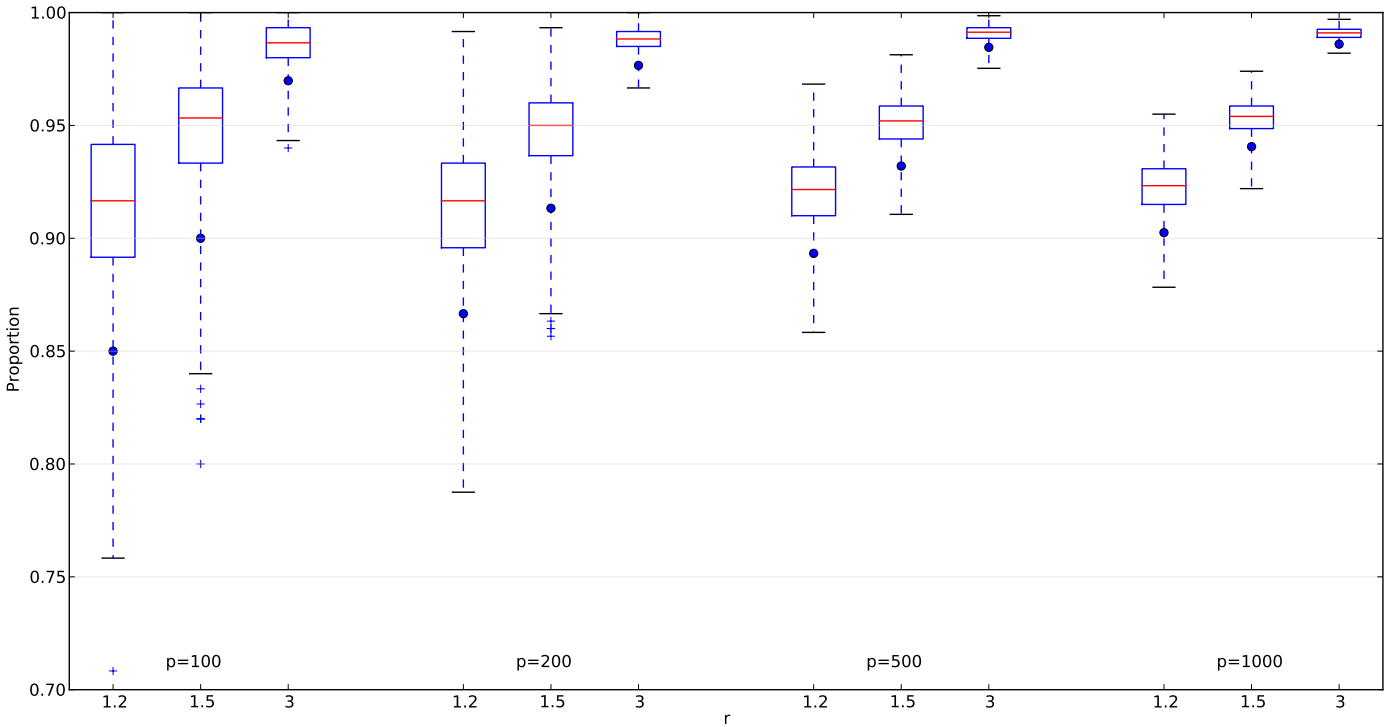


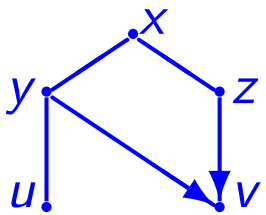
Figure: Directed edges in the spaces of Markov equivalence classes with sparsity constraints.

通过观测数据很大可能识别绝大部分因果关系！

Outline

Size of a Markov equivalence class

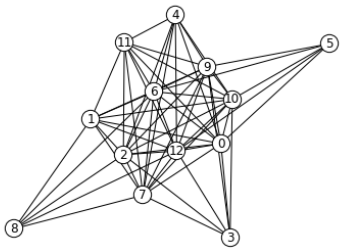
- The size of a Markov equivalence class is the number of DAGs in the class.
- The size only depends on the undirected subgraphs (chain components) of essential graph



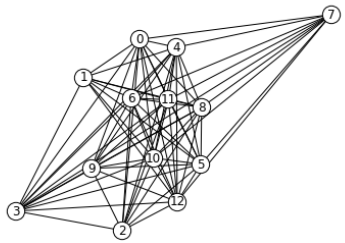
Markov equivalence class $\{G_1, G_2, G_3, G_4\}$ Size=4

essential graph

- Example



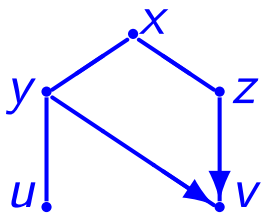
13 nodes, 60 edges, size=768,240



13 nodes, 73 edges, size=52,980,480

Size of a Markov equivalence class

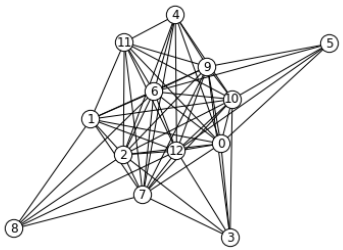
- The size of a Markov equivalence class is the number of DAGs in the class.
- The size only depends on the undirected subgraphs (chain components) of essential graph



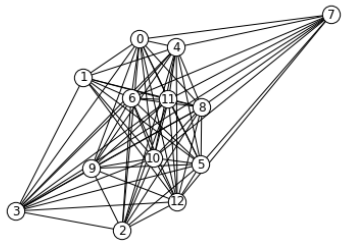
Markov equivalence class $\{G_1, G_2, G_3, G_4\}$ Size=4

essential graph

- Example



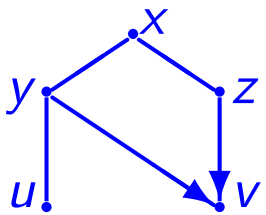
13 nodes, 60 edges, size=768,240



13 nodes, 73 edges, size=52,980,480

Size of a Markov equivalence class

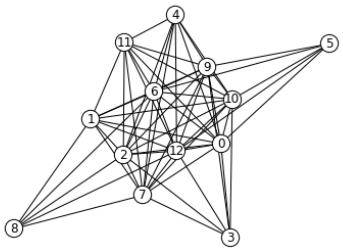
- The size of a Markov equivalence class is the number of DAGs in the class.
- The size only depends on the undirected subgraphs (chain components) of essential graph



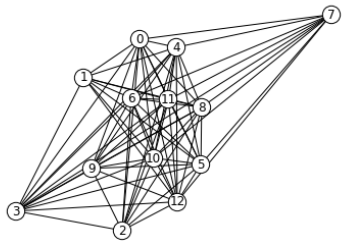
Markov equivalence class $\{G_1, G_2, G_3, G_4\}$ Size=4

essential graph

- Example



13 nodes, 60 edges, size=768,240



13 nodes, 73 edges, size=52,980,480

Motivations

Why Size important?

- Size measures the “uncertainty" or “complexity" of the learning of DAG models
- Used in causal structure learning and causal effect inference (Maathuis *et al.*, 2009; Chickering, 2002)

Counting equivalent DAGs

- Two approaches: Traversal methods or formulas-based methods
  - (Gillispie & Perlman, 2002) lists all DAGs with up to 10 vertices.
  - Robinson (1973, 1977) provide recursive formulas to count DAGs given the number of vertices.
- Counting equivalent DAGs in a class:
  - Listing all might be impractical,
    $1 \leq size \leq p!$,$(10! = 3628800, 20! > 10^{18})$, for a class with p vertices,
  - Hard to get general size formulas

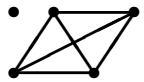
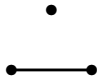
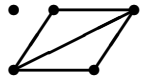
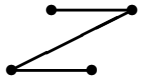
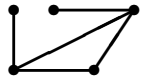
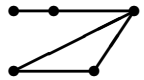
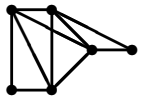
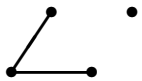
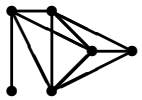
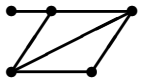
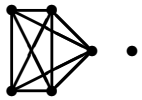
Proposed method

Traverse subclasses + formulas for counting subclass

Counting methods

- Partition a class into smaller rooted sub-classes recursively until the sizes of all sub-classes can be obtained via five closed-form formulas. (He *et al.*, 2015)
- Size Formulas by Symbolic Computations (He & Yu, 2016)

Formulas for counting the sizes of Markov Equivalence Classes

id	(n', p)	$\mathcal{K}$	$f(\mathcal{K}, m)/m!$	id	(n', p)	$\mathcal{K}$	$f(\mathcal{K}, m)/m!$
1	(1, 2)		$2m+1$	9	(4,5)		$24m+(m+4)\cdots(m+1)$
2	(2,3)		m^2+5m+2	10	(5,4)		m^2+7m+2
3	(3,3)		$3m+1$	11	(5,5)		$2m^3+11m^2+29m+10$
4	(3,4)		$3m^2+7m+4$	12	(5,5)		$m^3+10m^2+19m+10$
5	(3,4)		$m^3+6m^2+17m+6$	13	(5,5)		$m^3+10m^2+19m+10$
6	(4,4)		$4m^2+12m+4$	14	(5,6)		$m^4+14m^3+55m^2+82m+40$
7	(4,4)		$2m^2+8m+3$	15	(5,6)		$(m+1)(2m+3)(m^2+7m+16)$
8	(4,5)		$(m+1)(m+4)(2m+3)$	16	(5,6)		$120m+(m+5)\cdots(m+1)$

(Counting) Example

- Example

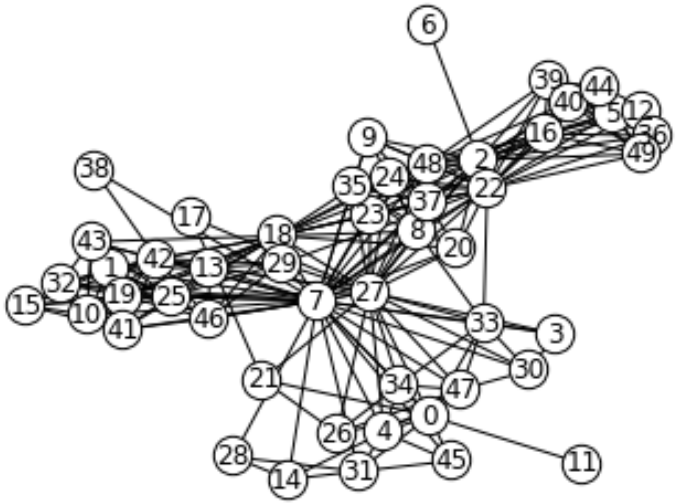


Figure: p=50, n=250, size=90,334,615,633,920

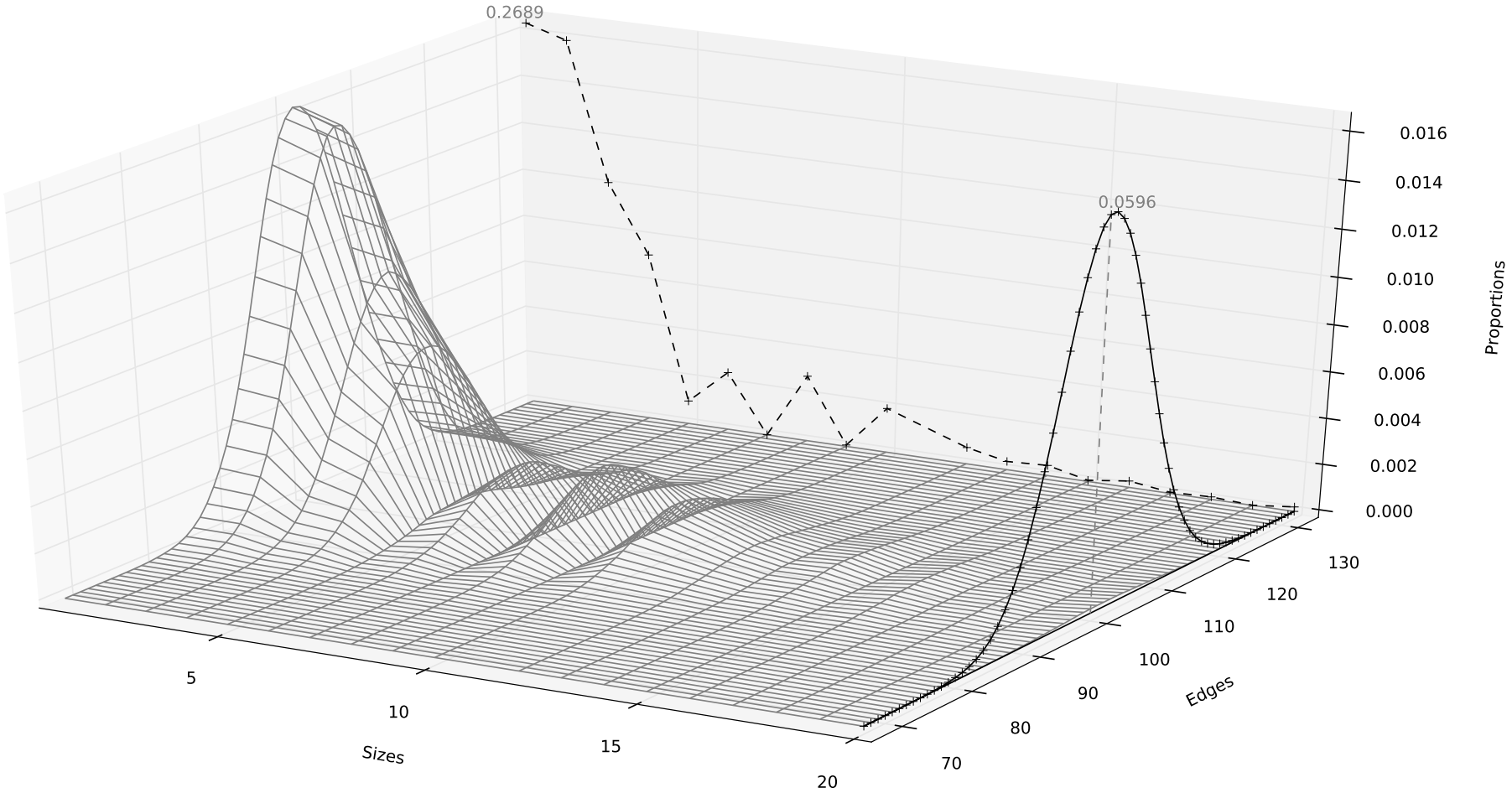


Figure: The surface displays the distribution of the Markov equivalence classes with 20 vertices. Two rescaled marginal distributions are shown in the planes. The black dashed line is the size distribution and the black solid line is the edge distribution of Markov equivalence classes.

(Counting) larger p with a sparsity constraint

Table: The average size of Markov equivalence classes and average counting time are reported based on 10^5 samples from $\mathbb{U}_p^{pr*}$, where p ranges from 15 to 100.

r	p	15	20	30	40	50	60	70	80	90	100
2	Size	7363	6.98e4	4.74e6	6.94e8	1.9e10	1.2e12	1.2e14	1.5e15	1.8e17	2.6e19
3		3.0e5	3.3e6	1.1e10	7.1e12	4.4e15	8.6e18	1.3e21	6.1e23	1.4e27	9.1e27
4		2.7e6	5.4e8	6.7e12	2.8e16	3.5e19	5.9e22	5.8e25	1.3e29	1.3e38	1.5e34
5		4.9e7	6.7e9	8.3e14	5.4e18	1.1e24	2.8e26	2.3e30	4.8e33	5.6e40	3.8e40
2	Time (sec.)	3.2e-3	5.7e-3	1.2e-2	2.3e-2	0.028	0.037	0.059	0.074	0.090	0.15
3		1.7e-2	3.8e-2	8.8e-2	0.15	0.17	0.27	0.42	0.53	0.75	0.86
4		0.19	0.43	0.72	1.37	1.51	2.16	3.35	3.64	6.14	9.03
5		2.89	7.07	7.91	17.49	50.43	82.99	90.37	95.54	127.25	213

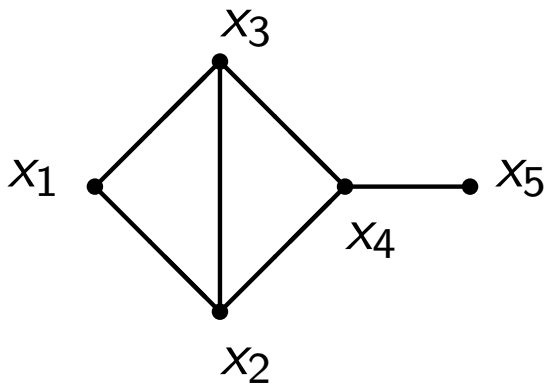
Outline

Motivation of Active Learning

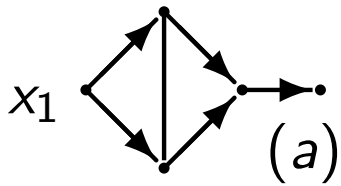
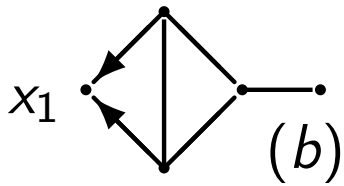
- The different DAGs in a Markov equivalence class cannot be distinguished from observational data (GGM,BN),
- Experiments with external interventions are needed
- Find optimal designs of experiments
- To minimize the number of manipulated variables and the set of candidate structures.

Interventional essential graph

- Candidate pre-intervention structures can be represented by essential graph
- Given the information from additional interventions, the number of candidate structures become smaller, and the corresponding subclass can be represented by a post-intervention essential graph.
- An essential graph of DAGs. There are 12 DAGs in this Markov equivalence class



Post-intervention subclasses and essential graphs

No of subclass	post-intervention result on x_1	number of DAGs	post-intervention essential graphs
(i)	$x_2 \leftarrow x_1 \rightarrow x_3$	2	
(ii)	$x_2 \rightarrow x_1 \rightarrow x_3$	1	
(iii)	$x_2 \rightarrow x_1 \leftarrow x_3$	8	
(iv)	$x_2 \leftarrow x_1 \leftarrow x_3$	1	

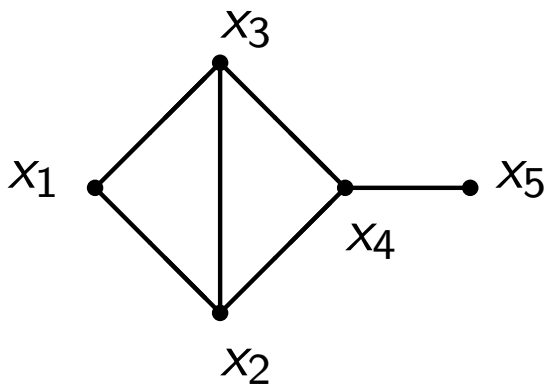
Optimization for sequential interventions

- Let M is the number of all possible orientations of edges adjacent to x obtained by manipulating x,
- l_i denotes the number of possible DAGs with the ith orientation obtained by manipulating v,
- $L = \sum_i l_i$; $l_{max} = \max l_i$
- **Minimax criterion**.
  We choose the variable x that minimizes l_{max}
- **The maximum entropy criterion**.
  We choose a variable x such that the following entropy is maximized for any x

$$H_x = -\sum_{i=1}^{M} \frac{l_i}{L} \log \frac{l_i}{L}, \tag{1}$$

A simulation

- Underlying pre-intervention essential graph



- Simulation

	m^*			
Design	1	2	3	4
Random	268	475	202	55
Minimax	437	563	0	0
Entropy	441	559	0	0

m^* denotes the number of manipulated variables

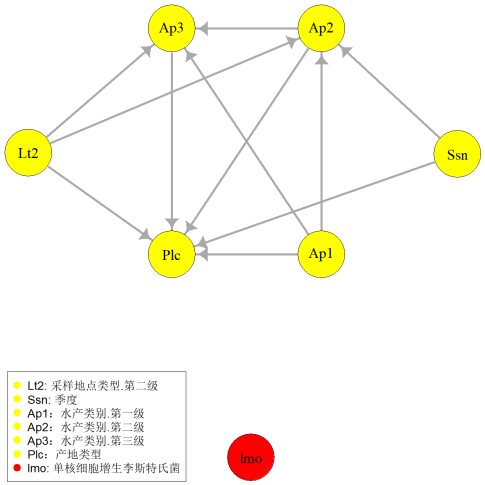
Table: The frequencies of the numbers of interventions

Outline

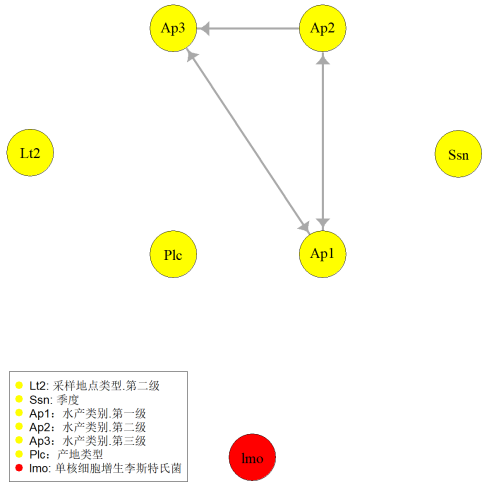
Assumptions revisited and a real data example

- No hidden variables, all common causes are observed
- A variable is determined by its direct causes and a noise,
- The causal graph is acyclic
- The distribution of observational data is faithful to the underlying DAG
- A real data example

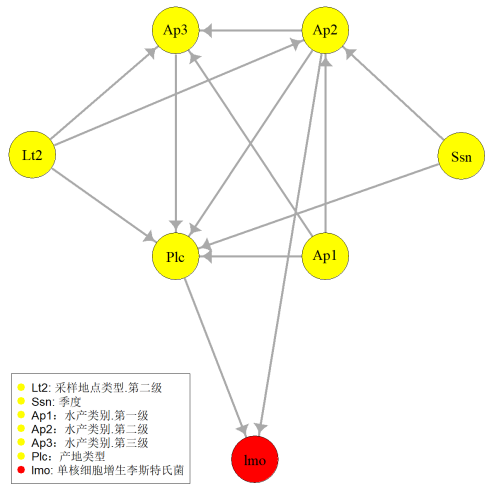
PC algorithm



FCI algorithm



PC(domain knowledge)



Outline

Thanks!

Andersson, S. A., Madigan, D., & Perlman, M. D. 1997. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, **25**(2), 505–541.

Athey, S, & Imbens, G. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(27), 7353.

Chickering, D. M. 2002. Learning equivalence classes of Bayesian-network structures. *The Journal of Machine Learning Research*, **2**, 445–498.

Gillispie, S. B. 2006. Formulas for counting acyclic digraph Markov equivalence classes. *Journal of Statistical Planning and Inference*, **136**(4), 1410–1432.

Gillispie, S.B., & Perlman, M.D. 2002. The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, **141**(1-2), 137–155.

Hausman, DM, & Woodward, J. 1999. Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science*, **50**(4), 521–583.

He, Y., & Yu, B. 2016. Formulas for Counting the Sizes of Markov Equivalence Classes of Directed Acyclic Graphs. *ArXiv e-prints*, Oct.

He, Yangbo, & Geng, Zhi. 2008. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, **9**, 2523–2547.

He, Yangbo, Jia, Jinzhu, & Yu, Bin. 2013. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *The Annals of Statistics*, **41**(4), 1742–1779.

He, Yangbo, Jinzhu, Jia, & Yu, Bin. 2015. Counting and Exploring Sizes of Markov Equivalence Classes of Directed Acyclic Graphs. *Research Reports, No 175,Center for Statistical Science, Peking University*.

Lauritzen, Steffen L, & Spiegelhalter, David J. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 157–224.

Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.

Maathuis, M. H., Kalisch, M., & Bühlmann, P. 2009. Estimating

high-dimensional intervention effects from observational data. *The Annals of Statistics*, **37**(6A), 3133–3164.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge Univ Pr.

Peters, Jonas, Peter, Bühlmann, & Meinshausen, Nicolai. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society*, **78**(5), 947–1012.

Peters, Jonas, Janzing, Dominik, & Bernhard, Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. MIT press.

Pierre, Gutierrez, & Jean, Yves Gerardy. 2017. Causal Inference and Uplift Modelling: A Review of the Literature. *Pages 1–13 of:* Hardgrove, Claire, Dorard, Louis, Thompson, Keiran, & Douetteau, Florian (eds), *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. Proceedings of Machine Learning Research, vol. 67. Microsoft NERD, Boston, USA: PMLR.

Reichenbach, H. 1956. *The direction of Time*. University of California Press, Berkeley, CA.

Robinson, R. 1973. Counting labeled acyclic digraphs. *New Directions in the Theory of Graphs*, 239–273.

Robinson, R. 1977. Counting unlabeled acyclic digraphs. 28–43.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5), 688–701.

Spiegelhalter, David J, & Lauritzen, Steffen L. 1990. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**(5), 579–605.

Spirtes, Peter, Glymour, Clark N, & Scheines, Richard. 2000. *Causation, prediction, and search*. Vol. 81. MIT press.

Steinsky, B. 2003. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete mathematics*, **270**(1-3), 266–277.

Tsamardinos, Ioannis, Brown, Laura E, & Aliferis, Constantin F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, **65**(1), 31–78.

Uhler, Caroline, Raskutti, Garvesh, B ü hlmann, Peter, & Yu, Bin. 2013. Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, **41**(2), 436–463.

Woodward, James. 2003. CRITICAL NOTICE: CAUSALITY BY JUDEA PEARL. *Economics & Philosophy*, **19**(2), 321–340.

Woodward, James. 2015a. Normative Theory and Descriptive Psychology in Understanding Causal Reasoning: The Role of Interventions and Invariance.

Woodward, James. 2015b (August). *Normative Theory and Descriptive Psychology in Understanding Causal Reasoning: The Role of Interventions and Invariance*.