Bayes Optimal Estimation of Intervention Effects and its Approximation*

*Some parts of this presentation will also be presented at CMStatistics 2020

WASEDA University

Shunsuke Horii

Waseda University – Academia Sinica
Data Science Workshop

Agenda

- Introduction to structural causal inference
- Bayes optimal estimator of intervention effects
- Approximation algorithm for the Bayes optimal estimator
- Experimental results

Introduction to structural causal inference

- Will increasing chocolate consumption increase the number of Nobel Prize winners?

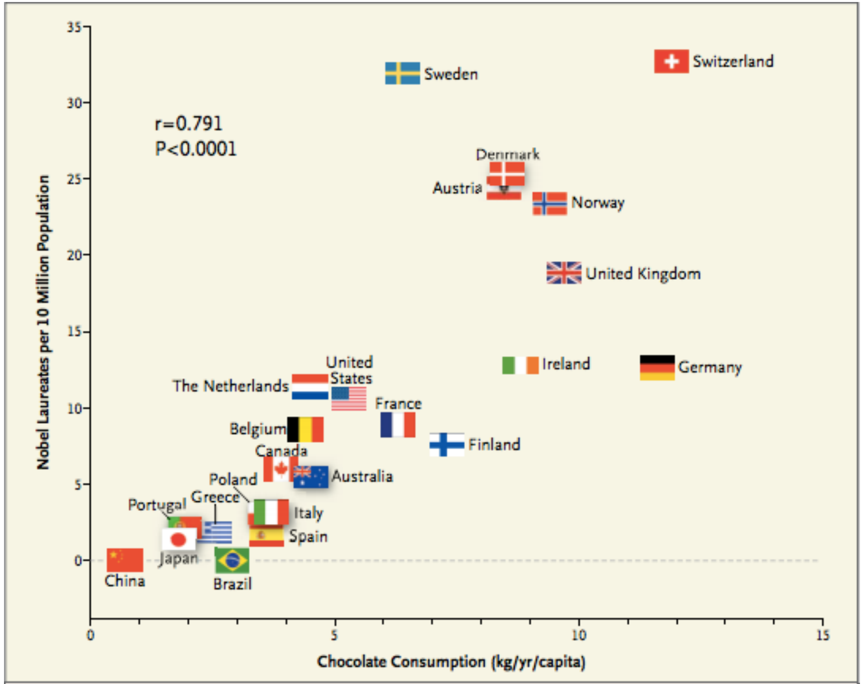


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

F. H. Messerli, "Chocolate Consumption, Cognitive Function, and Nobel Laureates," 2012

Introduction to structural causal inference

- Will increasing chocolate consumption increase the number of Nobel Prize winners?

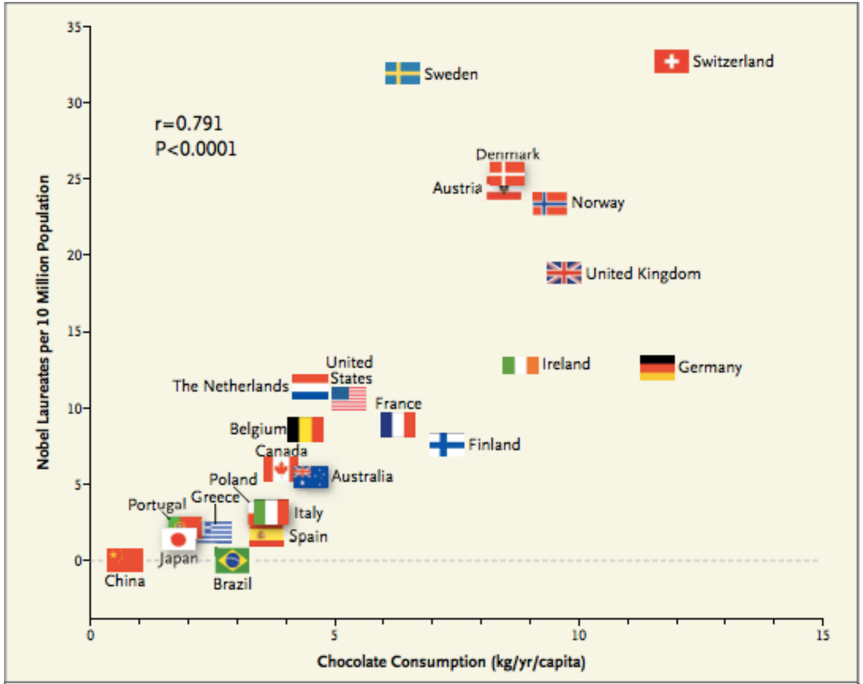
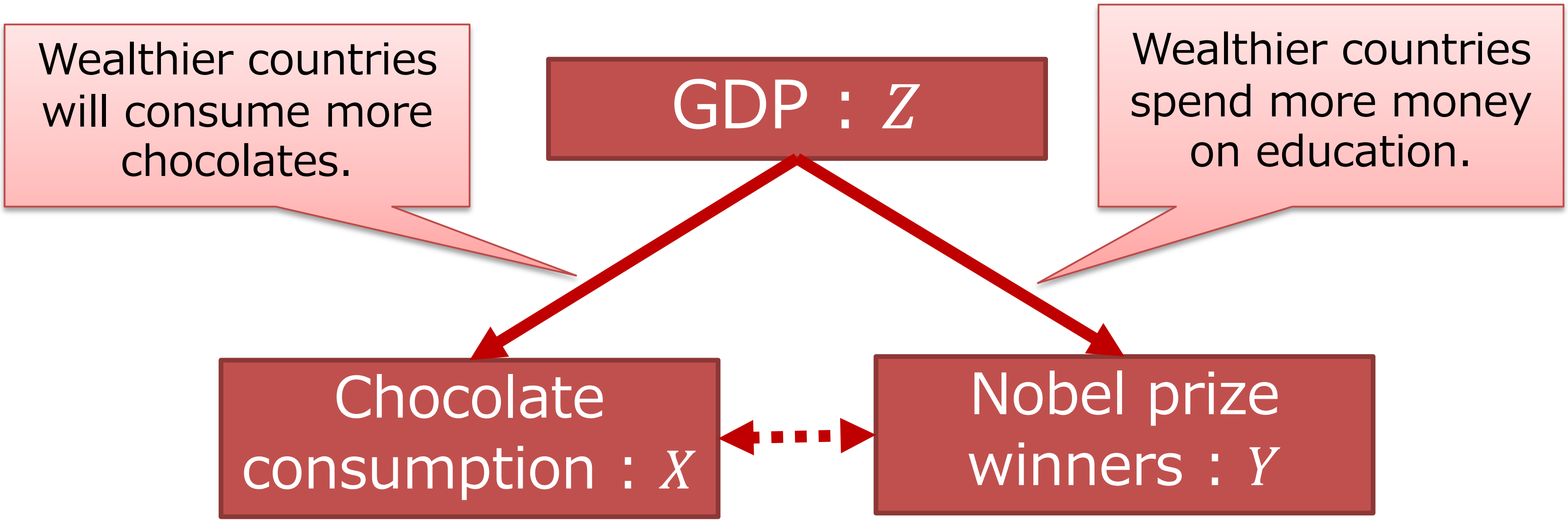


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

F. H. Messerli, "Chocolate Consumption, Cognitive Function, and Nobel Laureates," 2012

Correlation
≠
Causation

Introduction to structural causal inference



Introduction to structural causal inference

- Karl Pearson, "The Grammar of Science," 1911

> No phenomena are causal; all phenomena are contingent, and the problem before us is to measure the degree of this contingency, ⋯

  – There had been no concrete definition of "**causality**" in statistics.

  Exception: R. Fisher's Randomized controlled trial ⇒ limited to experimental studies

- **Statistical** estimation of causality in observational studies has been attracting a lot of attention recently.

We need a ***definition*** of causality

Introduction to structural causal inference

- Simpson's paradox

	Treatment A	Treatment B
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
	$\frac{562}{700} = 0.80$	

Charig et al. "Comparison of treatment of renal calculi by open surgery, (⋯)", British Medical Journal, 1986

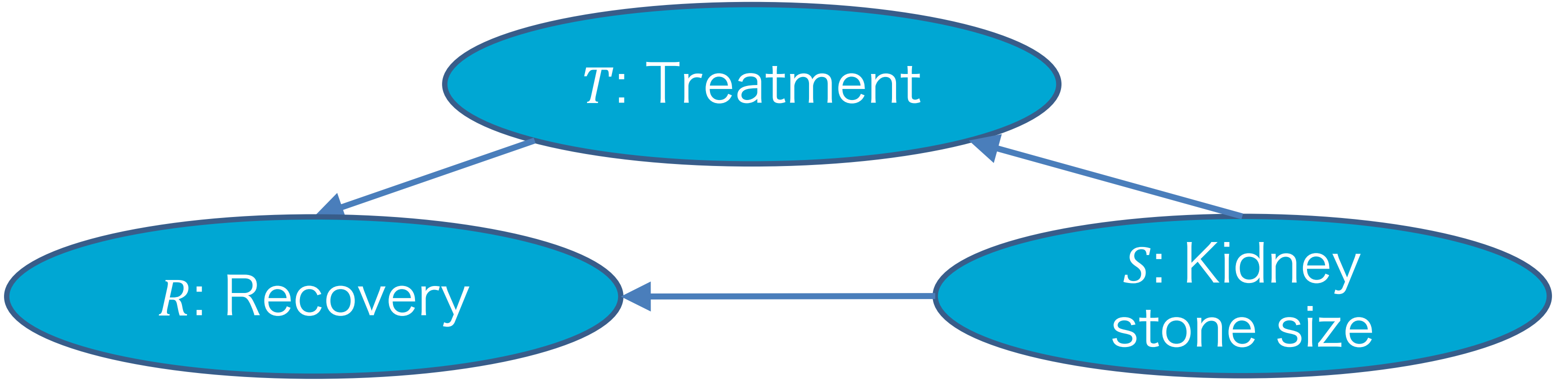
- Simpson's paradox

	Treatment A	Treatment B
kidney stone size: small	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
kidney stone size: large	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
	$\frac{562}{700} = 0.80$	

Charig et al. "Comparison of treatment of renal calculi by open surgery, (⋯)", British Medical Journal, 1986

Introduction to structural causal inference

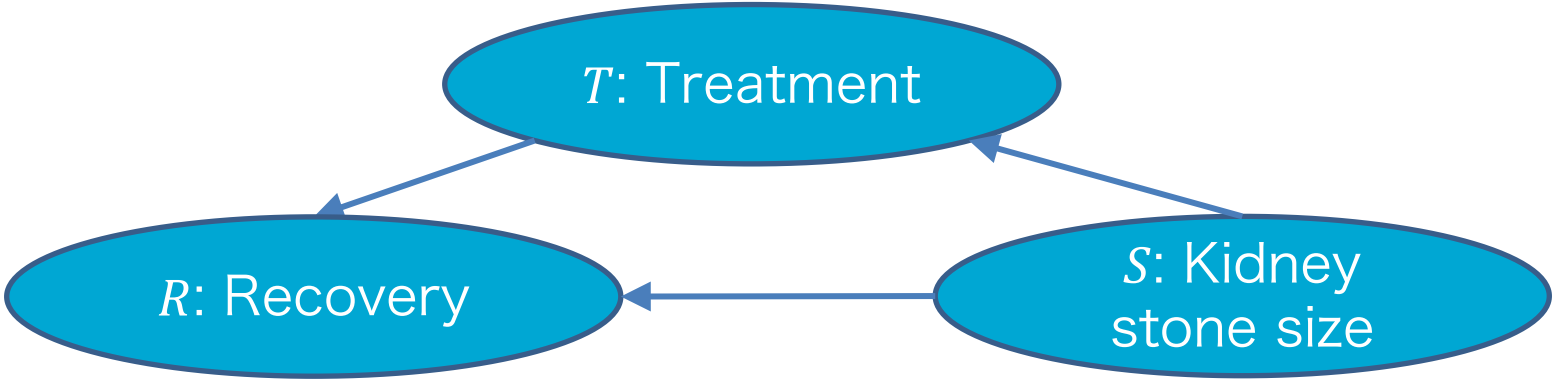
- Structure behind the data



Probability of being recovered when the treatment is A or B:

$$P(R = 1|T = A) = 0.78$$
$$P(R = 1|T = B) = 0.83$$

- Structure behind the data



Probability of being recovered when the treatment **is** A or B:

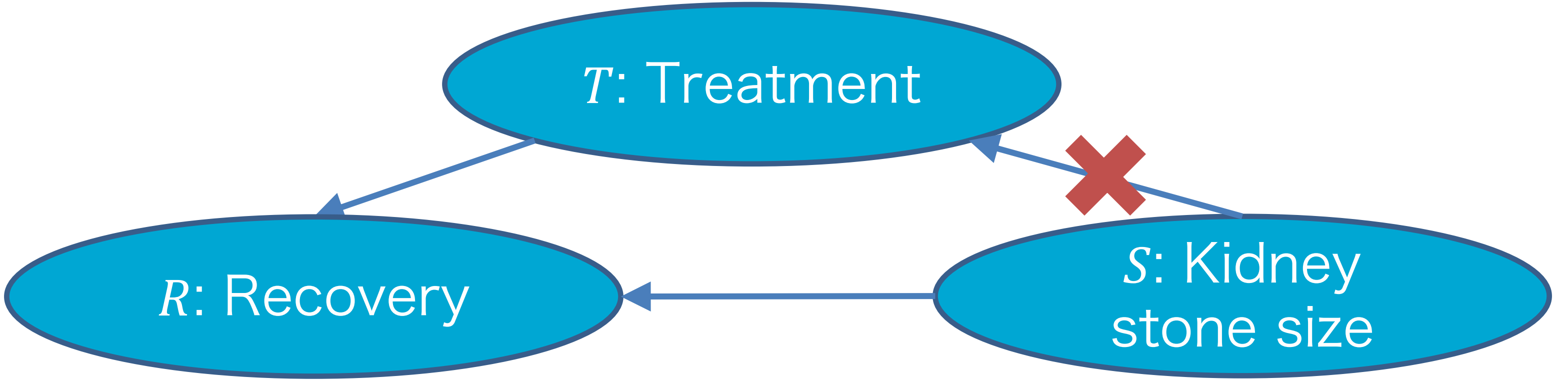
$$P(R = 1|T = A) = 0.78$$
$$P(R = 1|T = B) = 0.83$$

$\neq$

Probability of being recovered when the treatment **is set to** A or B:

$$P_{\mathrm{do}\{T:=A\}}(R = 1)$$
$$P_{\mathrm{do}\{T:=B\}}(R = 1)$$

- Structure behind the data



$P_{\mathrm{do}\{T:=A\}}(R=1)$ is **defined** as follows:

- The probability of $R=1$ when T is set to A **independently of** S
- Assumption: $P(S), P(R|S,T)$ does not change

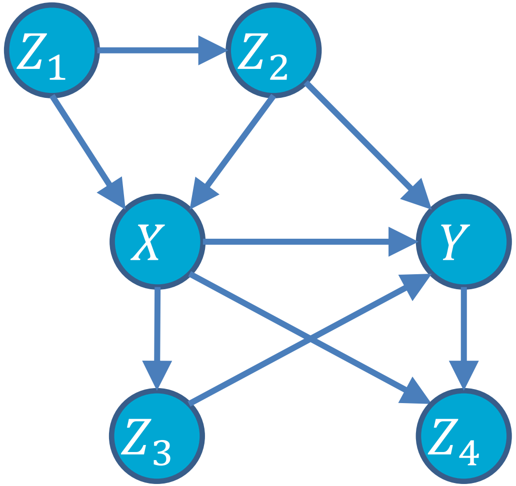
Introduction to structural causal inference

$$
\begin{aligned}
P_{do\{T:=A\}}(R=1) &= \sum_s \sum_t P_{do\{T:=A\}}(R=1, S=s, T=t) \\
&= \sum_s \sum_t P_{do\{T:=A\}}(R=1|S=s, T=t) P_{do\{T:=A\}}(S=s, T=t) \\
&= \sum_s P_{do\{T:=A\}}(R=1|S=s, T=A) P_{do\{T:=A\}}(S=s) \\
&= \sum_s P(R=1|S=s, T=A) P(S=s) \\
&= 0.93 \times 0.51 + 0.73 \times 0.49 \\
&= 0.832
\end{aligned}
$$

$$
P_{do\{T:=B\}}(R=1) = 0.782
$$

Introduction to structural causal inference

- Structural (equation/causal) model (SCM) & Causal graph



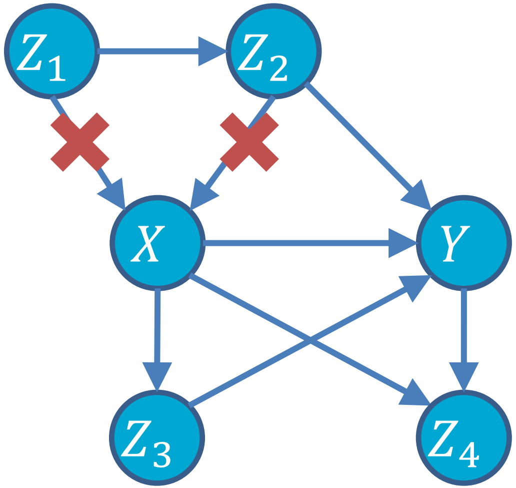
$$Z_2 = f_{Z_2}(Z_1, \epsilon_{Z_2})$$
$$X = f_X(Z_1, Z_2, \epsilon_X)$$
$$Z_3 = f_{Z_3}(X, \epsilon_{Z_3})$$
$$Y = f_Y(Z_2, X, Z_3, \epsilon_Y)$$
$$Z_4 = f_{Z_4}(X, Y, \epsilon_{Z_4})$$

The left-hand side variables are **generated** according to the right-hand side equations.

Stronger assumption than just assuming conditional distributions.

Introduction to structural causal inference

- Definition of **Intervention effect** [J. Pearl 1998]



Distribution of the objective variable when the treatment variable is fixed to a certain value independently of the other variables:

$$P_{do\{X:=x\}}(y) \triangleq \int \cdots \int \frac{p(x, y, z_1, \ldots, z_p)}{p(x|\mathrm{pa}(x))} dz_1 \ldots dz_p$$

Motivation

- Three steps to calculate the intervention effect:
  1. Determine/Estimate the structure of causal graph
     - Use domain knowledge
     - Estimate from data
       - Use independence and conditional independence→Ex: PC algorithm
       - Use posterior probabilities of models→Ex: GES algorithm
       - Use restrictions on models→Ex: LiNGAM
  2. Estimate the conditional distributions
  3. Calculate the intervention effect

If the goal is to estimate the intervention effect, we don't have to fix a single causal graph and conditional distributions.

➡ It is Bayes optimal to average the intervention effects estimated under each model.

Motivation

- Posterior probabilities calculation for all candidate models are necessary for the Bayes optimal estimation

$\Rightarrow$ Computationally hard when the number of candidate models is large

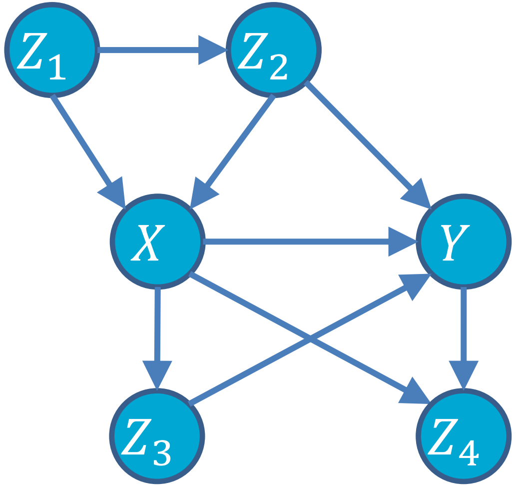
Our research

- Develop the Bayes optimal estimator of mean intervention effect in linear SCM
- Develop an approximation to Bayes optimal estimator by using variational Bayes method
  - Utilize an idea developed in Bayesian sparse modeling literature

Linear SCM

- Structural equations are linear (path analysis)

Ex.



$$
\begin{aligned}
Z_2 &= \theta_{Z_2 Z_1} Z_1 + \epsilon_{Z_2} \\
X &= \theta_{X Z_1} Z_1 + \theta_{X Z_2} Z_2 + \epsilon_X \\
Z_3 &= \theta_{Z_3 X} X + \epsilon_{Z_3} \\
Y &= \theta_{Y Z_2} Z_2 + \theta_{Y X} X + \theta_{Y Z_3} Z_3 + \epsilon_Y \\
Z_4 &= \theta_{Z_4 X} X + \theta_{Z_4 Y} Y + \epsilon_{Z_4}
\end{aligned}
$$

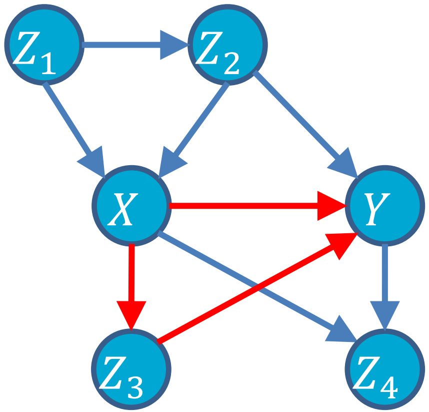
coefficients in RHS:
path coefficients

Linear SCM

- Mean intervention effect (MIE) can be expressed by **total effect** [J. Pearl 1998]

$$\bar{y}_x \triangleq \int y \cdot P_{do\{X:=x\}} \mathrm{d}y = \left(\sum_{l \in \mathcal{P}} \prod_{(i,j) \in l} \theta_{ij} \right) x$$

Ex.



$$Z_2 = \theta_{Z_2 Z_1} Z_1 + \epsilon_{Z_2}$$
$$X = \theta_{X Z_1} Z_1 + \theta_{X Z_2} Z_2 + \epsilon_X$$
$$Z_3 = \theta_{Z_3 X} X + \epsilon_{Z_3}$$
$$Y = \theta_{Y Z_2} Z_2 + \theta_{Y X} X + \theta_{Y Z_3} Z_3 + \epsilon_Y$$
$$Z_4 = \theta_{Z_4 X} X + \theta_{Z_4 Y} Y + \epsilon_{Z_4}$$

$$\bar{y}_x = (\theta_{YX} + \theta_{Z_3 X} \theta_{Y Z_3}) x$$

Bayes optimal estimator of MIE

Assumptions

- Causal graph $G \in \mathcal{G}$ is a random variable with prior $p(G)$
- Path coefficients $\{\theta_{ij}\}$ are random variables with prior $p(\boldsymbol{\theta}_G|G)$
  - $\boldsymbol{\theta}_G$: set of path coefficients under a causal graph G

Bayes optimal estimator

- data: D^N, decision function: $d(D^N)$
- loss function: $\ell(G, \boldsymbol{\theta}_G, d(D^N)) = (\bar{y}_x(G, \boldsymbol{\theta}_G) - d(D^N))^2$

$$d^*(D^N) = \sum_{G \in \mathcal{G}} p(G|D^N) \int \bar{y}_x(G, \boldsymbol{\theta}_G) p(\boldsymbol{\theta}_G|G, D^N) d\boldsymbol{\theta}_G$$

Approximate Bayes optimal estimator

Difficulty in calculation

$$d^*(D^N) = \sum_{G \in \mathcal{G}} p(G|D^N) \int \bar{y}_x(G, \boldsymbol{\theta}_G) p(\boldsymbol{\theta}_G | G, D^N) d\boldsymbol{\theta}_G$$

1. Difficulty in integration

If we assume conjugate prior, we can calculate $p(\boldsymbol{\theta}_G|G, D^N)$ analytically, but even then, this integration is difficult (due to the nonlinearity of $\bar{y}_x(G, \boldsymbol{\theta}_G)$ w.r.t. θ_G)

$$d^*(D^N) \approx \sum_{G \in \mathcal{G}} p(G|D^N) \bar{y}_x(G, \boldsymbol{\theta}_G^{MAP})$$

$$\boldsymbol{\theta}_G^{MAP} = \arg\max_{\boldsymbol{\theta}_G} p(\boldsymbol{\theta}_G|G, D^N)$$

Approximate Bayes optimal estimator

Difficulty in calculation

$$d^*(D^N) = \sum_{G \in \mathcal{G}} p(G|D^N) \int \bar{y}_x(G, \boldsymbol{\theta}_G) p(\boldsymbol{\theta}_G | G, D^N) d\boldsymbol{\theta}_G$$

2. Difficulty in summation over all models

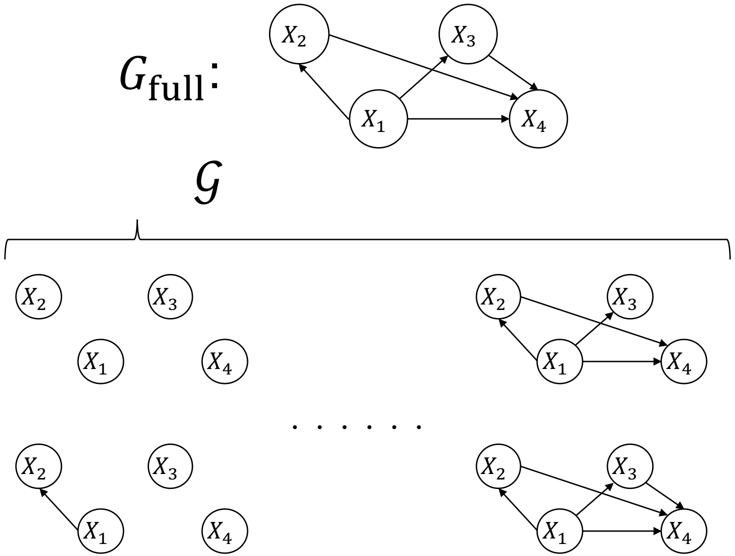
Computationally hard when the number of models is large

Approximate Bayes optimal estimator

Assumptions for approximation

- We know the positions of the possible edges
- We know the orientation of the possible edges (causal order)
- Probability that an edge exist is p
- If an edge (i, j) exists, $p(\theta_{ij}|G) \sim \mathcal{N}(0, \tau)$

$$p(G) = p^{|E_G|}(1-p)^{|E_{\text{full}} \setminus E_G|}$$



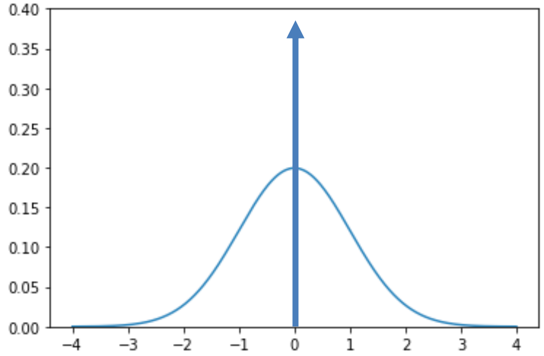
Approximate Bayes optimal estimator

Key idea

- Under the assumptions, we can write $p(\theta_{ij}) = \sum_G p(\theta_{ij}|G)$ as follows:

$$p(\theta_{ij}) = (1-p)\delta_0(\theta_{ij}) + p\mathcal{N}(\theta_{ij}; 0, \tau)$$

$\delta_0(\cdot)$: Dirac delta function



Superficially, we can replace summation with integration (We assume G_{full} as a model, and this distribution as the prior for $\boldsymbol{\theta}_{G_{\mathrm{full}}}$)

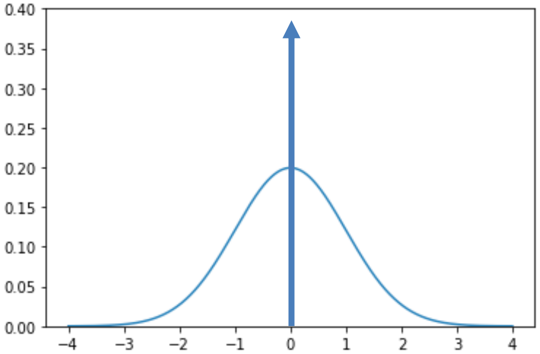
Approximate Bayes optimal estimator

Key idea

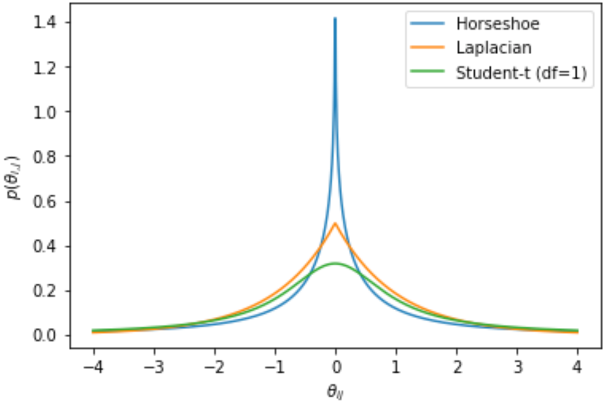
- Under the assumptions, we can write $p(\theta_{ij}) = \sum_G p(\theta_{ij}|G)$ as follows:

$$p(\theta_{ij}) = (1-p)\delta_0(\theta_{ij}) + p\mathcal{N}(\theta_{ij}; 0, \tau)$$

$\delta_0(\cdot)$: Dirac delta function



Approximate it by Gaussian Scale Mixture



Approximate Bayes optimal estimator

Approximation algorithm

- Assume GSM for $p(\boldsymbol{\theta}_{G_{\mathrm{full}}}|G_{\mathrm{full}})$
- Approximately calculate $\widehat{\boldsymbol{\theta}}_{G_{\mathrm{full}}} = \operatorname*{argmax}_{\boldsymbol{\theta}_{G_{\mathrm{full}}}} p(\boldsymbol{\theta}_{G_{\mathrm{full}}}|G_{\mathrm{full}}, D^N)$ using variational Bayes method
- Calculate $\bar{y}_x(G_{\mathrm{full}}, \widehat{\boldsymbol{\theta}}_{G_{\mathrm{full}}})$

Experiments

- Semi-synthetic data
  - Infant Health and Development Program (IHDP) data
  - Linked Birth and Infant Death Data (LBIDD)
- Include counterfactual data generated artificially

y_{cfact}	y_{fact}	x	w_1	w_2	$\cdots$
4.32	5.60	1	-0.53	-0.34	$\cdots$
7.86	6.88	0	-1.74	-1.80	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Experiments

- Semi-synthetic data
  - Infant Health and Development Program (IHDP) data
  - Linked Birth and Infant Death Data (LBIDD)
- Include counterfactual data generated artificially

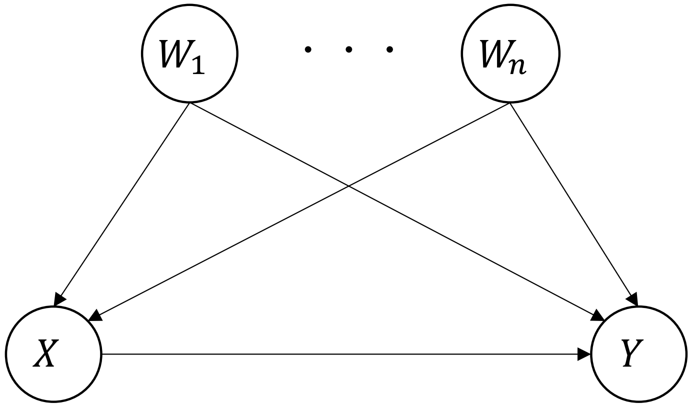
	y_{cfact}	y_{fact}	x	w_1	w_2	⋯
1.58	4.32	5.60	1	-0.53	-0.34	⋯
0.98	7.86	6.88	0	-1.74	-1.80	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Avg. 4.02 ← estimate

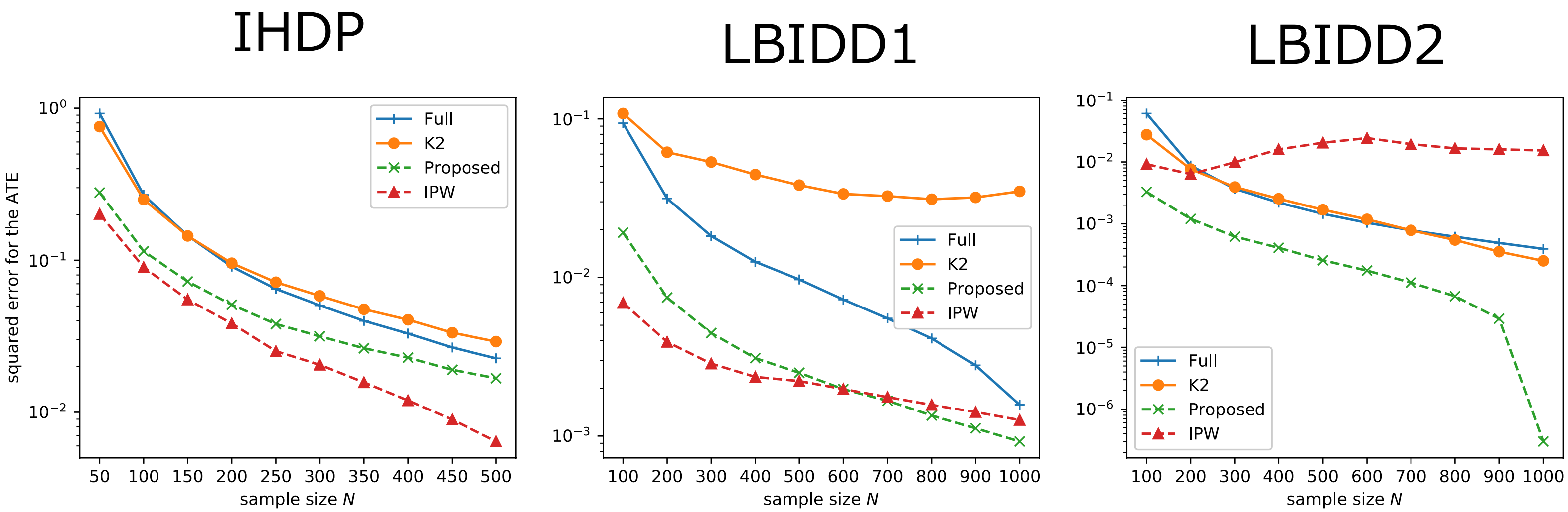
Experiments

y_{cfact}	y_{fact}	x	w_1	w_2	$\cdots$
4.32	5.60	1	-0.53	-0.34	$\cdots$
7.86	6.88	0	-1.74	-1.80	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

G_{full} :



Experiments



- The outputs of IPW estimator and proposed estimator would be comparatively reliable for IHDP and LBIDD1
- The proposed estimator is robust in the data generation process

Summary

- Introduced the Pearl's framework of causality estimation
- Derived the Bayes optimal estimator of the mean intervention effect when the data generating model is an unknown random variable
- Developed an approximation algorithm for the optimal estimator by using a sparse model technique

- Future works:
  - Unknown causal order
  - Unobservable latent variables
  - Non-linear model or Non-Gaussian model