Causality in Biomedicine
Lecture Series: Lecture 7

Ava Khamseh

11 March, 2020

Causal Discovery Methods (Based on Graphical Models)

Class of Algorithm	Name	Assumptions	Short comings	Input
Constraint-based	PC (oldest)	Any distribution, No unobsv. confounders, Markov cond, faithfulness	Causal info only up to equivalence classes, Non bivariate	Complete undirected graph
	FCI	Any distribution, Asymptotically correct with confounders, Markov cond, faithfulness		
Score-based	GES	No unobsv. confounders	Non-bivariate	Empty graph, adds edges, removes some
Functional Causal Models (FCMs)	LinGAM/ ANM	Asymmetry in data	Requires additional assumptions (not general), harder for discrete data	Structural Equation Model

Glymour et al. (2019)

Constraint-based assumptions

- **Markov condition:**
  - Absent edge implies conditional independence (**CI**)
  - Observing conditional dependence implies an edge

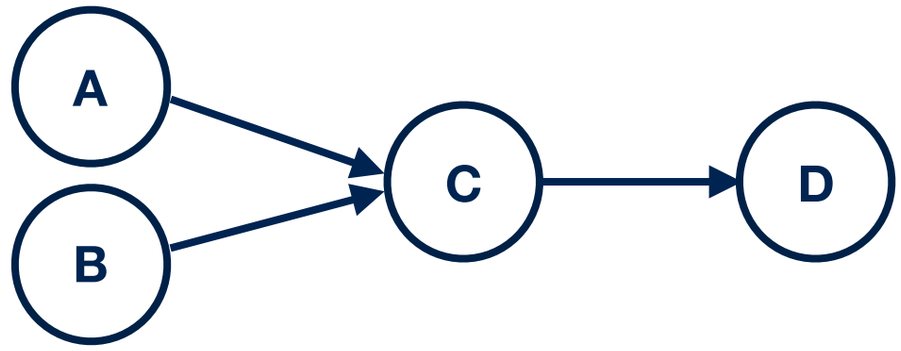
- **Causal sufficiency:** For any pair of variables X, Y, if there exists a variable Z which is a direct of cause of both X and Y, then Z is included in the causal graph (Z may be unobserved)

- **Faithfulness:**
  - Conjugate to the Markov condition
  - Edge implies conditional dependence
  - Observing CI implies absence of an edge

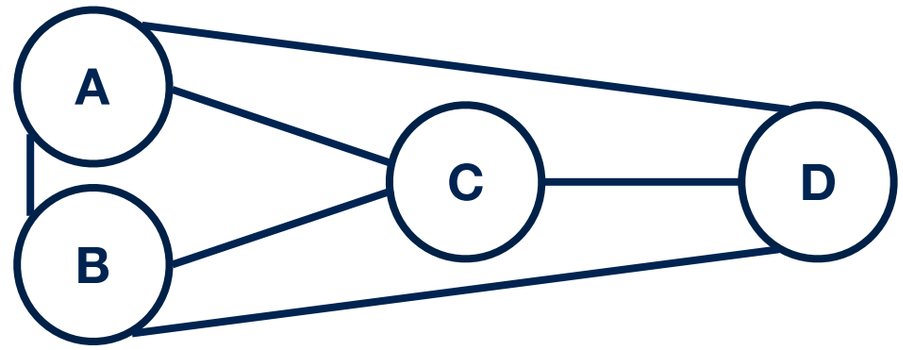
Could fail in regulatory systems, e.g., homeostasis.

Peter-Clark (PC) Algorithm

True causal graph:

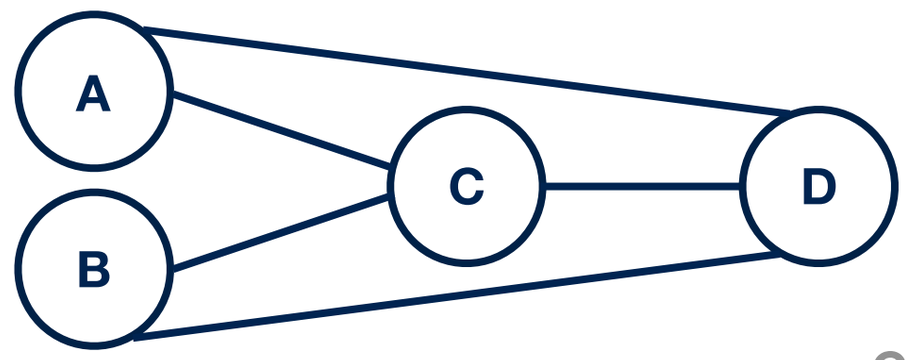


1. Start with the complete graph



2. Zeroth order CI, $A \perp\!\!\!\perp B$, by faithfulness:

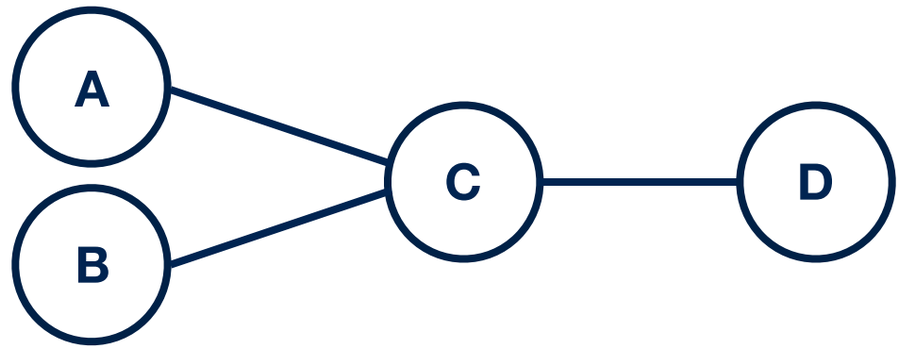
See later for statistical independence tests.



Glymour et al. (2019)

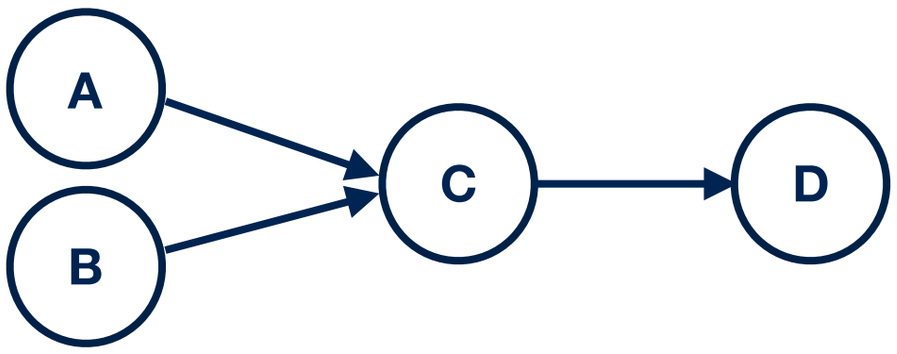
Peter-Clark (PC) Algorithm

3. 1st order CI, $A \perp\!\!\!\perp D|C$, by faithfulness:
$B \perp\!\!\!\perp D|C$



4. No higher order CI observed. Notice that conditioning sets only need to contain **neighbours** for the two nodes due to the Markov condition. We do not know the parents but parents are a subsets of neighbours. As the graph becomes sparser, the number of tests to be performed decrease. This makes CP very efficient.

5. Orient V-structures (colliders): take triplets where 2 nodes are connected to the 3rd: $A \not\perp\!\!\!\perp B|C$ only.



Note $C \to D$ cannot be as it would have been a collider (not detected in 5)

Glymour et al. (2019)

Today's lecture

- **Functional Causal Models (FCMs):** Utilising asymmetry in data for causal discovery

- **LiNGAMs:** Linear non-gaussian acyclic models, allow for new approaches for causal learning from observational data

- **ANM:** Additive noise models and **causal identifiablity**

- **IGCI:** Information Geometric Causal Inference

Jonas Peters et al, Elements of Causal Inference (2017)

Causal Structure Identifiability

- **LiNGAMs:** Linear non-gaussian acyclic models, allow for new approaches for causal learning from observational data.

- Focusing on 2 variables only, we wish to distinguish between:

$$x \to y \text{ or } y \to x$$

from **observational data**.

- Assumption: The effect of E is a linear function of C up to additive noise:

$$E = \alpha C + N_E, \quad N_E \perp\!\!\!\perp C$$

These assumptions are not enough to identify cause/effect.

Theorem: Identifiability of LiNGAMs

i.e., non-identifiability of gaussian Cause and Effect. If:

$$Y = \alpha X + N_Y, \quad N_Y \perp\!\!\!\perp X$$

There exists a β and a random variable N_X s.t.:

$$X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y$$

if and only if X and N_Y are gaussians.

i.e., it is sufficient that for X (Y) or N_Y (N_X) to be **non-gaussian** to render the causal direction identifiable.

Theorem: Identifiability of LiNGAMs

Proof:

1. Theorem (Darmois-Skitvic): Let $x_1, \cdots, x_d$ be independent, non-degenerate random variable. If there exists non-vanishing coefficients $a_1, \cdots, a_d$ and $b_1, \cdots, b_d$ such that the two linear combinations:

$$l_1 = a_1 x_1 + \cdots + a_d x_d$$

$$l_2 = b_1 x_1 + \cdots + b_d x_d$$

$l_1 \perp\!\!\!\perp l_2$ are independent, then each x_i is normally distributed

Theorem: Identifiability of LiNGAMs

Proof:

1. Theorem (Darmois-Skitvic): Let $x_1, \cdots, x_d$ be independent, non-degenerate random variable. If there exists non-vanishing coefficients $a_1, \cdots, a_d$ and $b_1, \cdots, b_d$ such that the two linear combinations:

$$l_1 = a_1 x_1 + \cdots + a_d x_d$$

$$l_2 = b_1 x_1 + \cdots + b_d x_d$$

$l_1 \perp\!\!\!\perp l_2$ are independent, then each x_i is normally distributed

2. Lemma (Peters 2008): Let $X \perp\!\!\!\perp N$. Then $N \not\perp\!\!\!\perp (X + N)$

3. We prove that $Y = \alpha X + N_Y \Rightarrow X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y$ $N_Y \perp\!\!\!\perp X$ **iff** $X, N_Y \sim \mathcal{N}$

Theorem: Identifiability of LiNGAMs

Proof:

3 We prove that if $X, N_Y \sim \mathcal{N}$ and $Y = \alpha X + N_Y, N_Y \perp\!\!\!\perp X$

$$\Longrightarrow X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y$$

Define:

$$\beta := \frac{Cov[X,Y]}{Cov[Y,Y]} = \frac{\alpha Var[X]}{\alpha^2 Var[X] + Var[N_Y]}$$

$$X = \beta Y + N_X \Rightarrow N_X = X - \beta Y$$

Then N_X, Y are uncorrelated by construction,

Moreover, Y is gaussian as it is a convolution of 2 gaussians.

Therefore, N_X is also gaussian.

Hence, N_X, Y are uncorrelated & gaussian, i.e., **independent**.

Theorem: Identifiability of LiNGAMs

Proof:

3 We prove the reverse: If

$$Y = \alpha X + N_Y, \quad N_Y \perp\!\!\!\perp X$$
$$X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y$$

$\Longrightarrow$

$$X, N_Y \sim \mathcal{N}$$

Since $N_X \perp\!\!\!\perp Y$, we have: $N_X = X - \beta(\alpha X + N_Y) = (1 - \alpha\beta)X - \beta N_Y$

There are 3 cases:

(i) $(1 - \alpha\beta) \neq 0 \;\&\; \beta \neq 0$

Then, given $N_X \perp\!\!\!\perp Y$, DS theorem implies $X, N_Y \sim \mathcal{N}$

Theorem: Identifiability of LiNGAMs

Proof:

3 We prove the reverse: If

$$Y = \alpha X + N_Y, \quad N_Y \perp\!\!\!\perp X$$
$$X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y$$

$\Longrightarrow$

$$X, N_Y \sim \mathcal{N}$$

Since $N_X \perp\!\!\!\perp Y$, we have: $N_X = X - \beta(\alpha X + N_Y) = (1 - \alpha\beta)X - \beta N_Y$

There are 3 cases:

(i) $(1 - \alpha\beta) \neq 0 \;\&\; \beta \neq 0$

Then, given $N_X \perp\!\!\!\perp Y$, DS theorem implies $X, N_Y \sim \mathcal{N}$

(ii) $(1 - \alpha\beta) \neq 0 \;\&\; \beta = 0$

Then, since $N_X \perp\!\!\!\perp Y$, and $N_X = X$, then $X \perp\!\!\!\perp \alpha X + N_Y$ in contradiction with Peters' lemma

Theorem: Identifiability of LiNGAMs

Proof:

3 We prove the reverse: If

$$Y = \alpha X + N_Y, \quad N_Y \perp\!\!\!\perp X$$
$$X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y$$

$\Longrightarrow$

$$X, N_Y \sim \mathcal{N}$$

Since $N_X \perp\!\!\!\perp Y$, we have: $N_X = X - \beta(\alpha X + N_Y) = (1 - \alpha\beta)X - \beta N_Y$

There are 3 cases:

(iii) $1 - \alpha\beta = 0 \;\&\; \beta \neq 0$

Then, since $N_X \perp\!\!\!\perp Y$, and $N_X = -\beta N_Y$, $N_Y \perp\!\!\!\perp \alpha X + N_Y$ again in contradiction with Peters' lemma

Therefore, as long as one of X, N_Y, Y, N_X is not gaussian, the causal direction is **identifiable** from **observational data**!

Jonas Peters et al, Elements of Causal Inference (2017)

Linear Additive Noise Models (ANMs)

ANM: The joint distribution $P_{X,Y}$ is said to admit an ANM for $X \to Y$ if there exists a measurable function f_Y and a noise variable N_Y s.t.

$$Y = f_Y(X) + N_Y, N_Y \perp\!\!\!\perp X$$

For this model, using convolution of probabilities we have:

$$p(x,y) = p_{N_Y}(y - f_Y(x))p_X(x)$$

Similarly, **if** a backward model exists:

$$p(x,y) = p_{N_X}(x - f_X(y))p_Y(y)$$

Jonas Peters et al, Elements of Causal Inference (2017)

Theorem: Identifiability of ANMs

Let $p(x, y) = p_{N_Y}(y - f_Y(x))p_X(x)$

if the backward model exists: $p(x, y) = p_{N_X}(x - f_X(y))p_Y(y)$

It must satisfy the following condition: $\nu^{''}(y - f_y(x))f^{'}(x) \neq 0$

$$\xi^{'''} = \xi^{'''}\left(-\frac{\nu^{'''}f^{'}}{\nu^{''}} + \frac{f^{''}}{f^{'}}\right) - 2\nu^{''}f^{''}f^{'} + \nu^{'}f^{''}f^{'} + \nu^{'}f^{'''} + \frac{\nu^{'}\nu^{'''}f^{''}f^{'}}{\nu^{''}} - \frac{\nu^{'}(f^{''})^2}{f^{'}}$$

where $\nu = \log(p_{N_Y}),\ \xi = \log(p_X)$

Jonas Peters et al, Elements of Causal Inference (2017)

Theorem: Identifiability of ANMs

Let $p(x, y) = p_{N_Y}(y - f_Y(x))p_X(x)$

if the backward model exists: $p(x, y) = p_{N_X}(x - f_X(y))p_Y(y)$

It must satisfy the following condition: $\nu''(y - f_y(x))f'(x) \neq 0$

$$\xi''' = \xi''' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'}$$

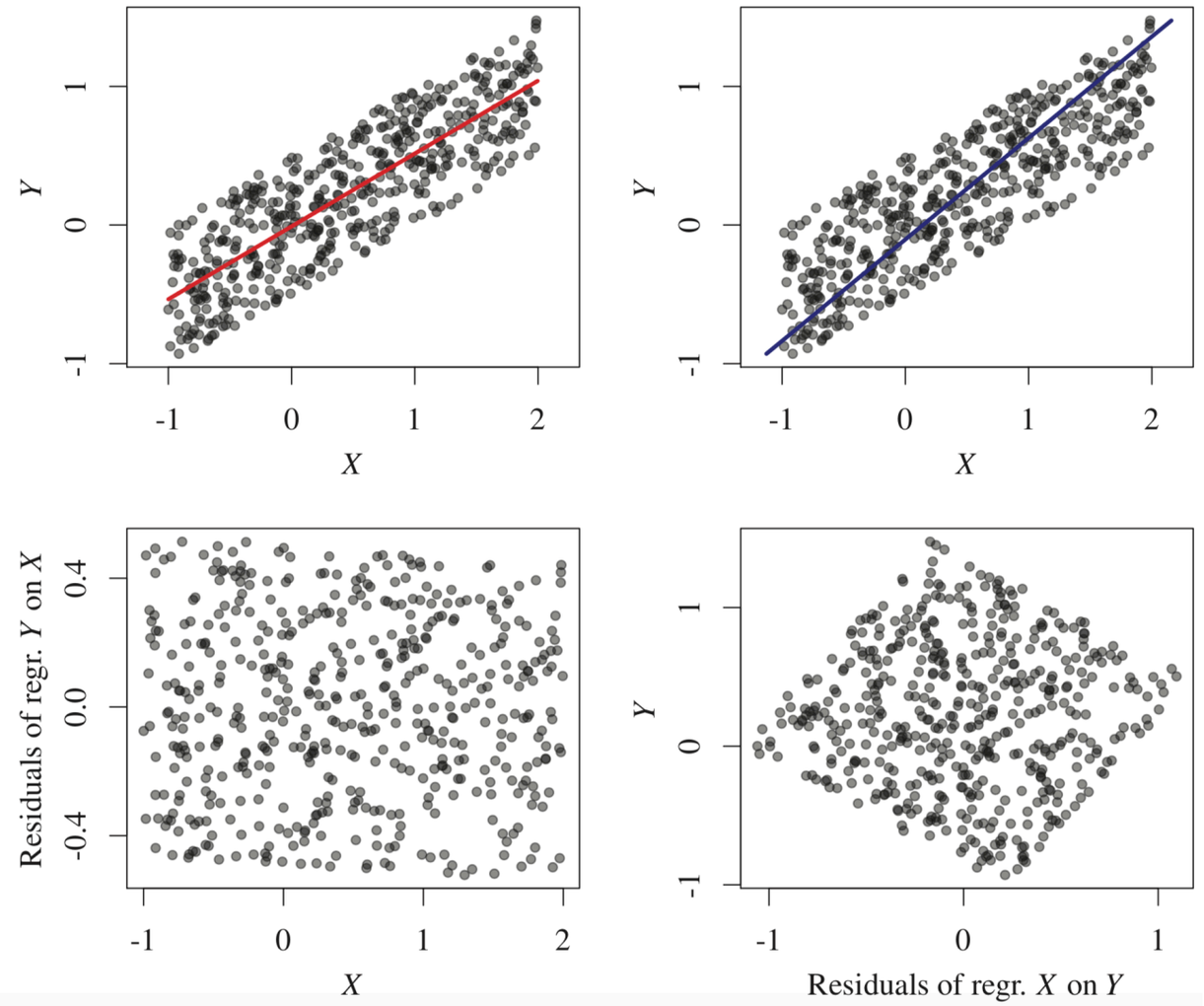
where $\nu = \log(p_{N_Y}),\ \xi = \log(p_X)$

The set of all p_X for which there is a backward model is **contained** in a 3-dim space (**small!**)

Jonas Peters et al, Elements of Causal Inference (2017)

In practice

1. Regress Y on X
2. **Test** whether $Y - \hat{f}_Y$ is independent of X
3. Repeat, swapping X and Y
4. If the independence is accepted for one direction and rejected for the other, infer the former as the causal direction,



Statistical Test of Independence: Choose one the accounts for higher order statistic rather than testing correlations only, e.g. HSIC

Jonas Peters et al, Elements of Causal Inference (2017)

In practice

```
library(dHSIC)
library(mgcv)
#
# generate data set
set.seed(1)
X <- rnorm(200)
Y <- X^3 + rnorm(200)
#
# fit models
modelforw <- gam(Y ~ s(X))
modelbackw <- gam(X ~ s(Y))
#
# independence tests
dhsic.test(modelforw$residuals, X)$p.value
# [1] 0.7628932
dhsic.test(modelbackw$residuals, Y)$p.value
# [1] 0.004221031
#
# computing likelihoods
- log(var(X)) - log(var(modelforw$residuals))
# [1] 0.1420063
- log(var(modelbackw$residuals)) - log(var(Y))
# [1] -1.014013
```

Gaussian noise

Jonas Peters et al, Elements of Causal Inference (2017)

Information Geometric Causal Inference (IGCI)

Provide an idea of how ‘independence’ between $p(E|C)$ and $p(C)$ can be formalised. How much **information** they contain about each other.

Toy model: Oversimplified, deterministic (no noise) Daniusis et al., (2010)

If $X \to Y$ is a causal model, the distribution of X and the function f mapping X to Y are ‘independent’ since they correspond to independent mechanisms of nature.

Jonas Peters et al, Elements of Causal Inference (2017)

Information Geometric Causal Inference (IGCI)

Toy model: Oversimplified, deterministic (no noise)

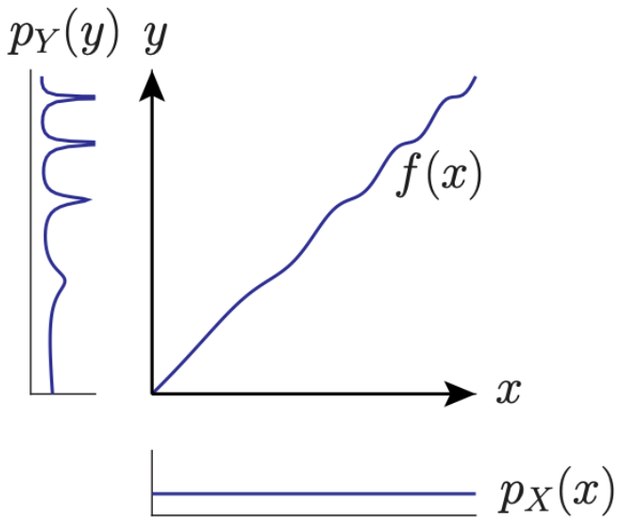
If $X \to Y$ is a causal model, the distribution of X and the function f mapping X to Y are 'independent' since they correspond to independent mechanisms of nature.

Daniusis et al., (2010)

Example: Uniform density as input X.

Let $P_X(x) = 1$ be the uniform density or [0,1] and f diffeomorphism of [0,1] with f(0)=0 and f(1)=1 and inverse $g := f^{-1}$. Then the distribution of $y = f(x)$ is:

$$p_Y(y) = g'(y) = \frac{1}{f'(f^{-1}(y))}$$



i.e. when $f'(x) = 0$ (flat regions), y has a peak.

Jonas Peters et al, Elements of Causal Inference (2017)

Information Geometric Causal Inference (IGCI)

Independence of f and p_X in terms of information geometry:

$$Cov\Big[\log f^{'}(x), p(x)\Big] = \int \log f^{'}(x) \cdot p(x) - \int \log f^{'}(x) \int p(x)$$

1

So **if** f and p_X are independent:

$$Cov\Big[\log f^{'}(x), p(x)\Big] = 0 \Rightarrow \int \log f^{'}(x) \cdot p(x) = \int \log f^{'}(x)$$

Jonas Peters et al, Elements of Causal Inference (2017)

Information Geometric Causal Inference (IGCI)

Independence of f and p_X in terms of information geometry:

$$Cov\Big[\log f^{'}(x), p(x)\Big] = \int \log f^{'}(x) \cdot p(x) - \int \log f^{'}(x) \int p(x)$$

(the last integral $\int p(x)$ is crossed out and equals 1)

So **if** f and p_X are independent:

$$Cov\Big[\log f^{'}(x), p(x)\Big] = 0 \Rightarrow \int \log f^{'}(x) \cdot p(x) = \int \log f^{'}(x)$$

This **will not hold for the opposite direction**: Using $p_Y(y) = g^{'}(y)$

$$Cov\Big[\log g^{'}(y), p(y)\Big] = \int \log g^{'}(y) \cdot p(y) - \int \log g^{'}(y) \int p(y)$$

(the last integral $\int p(y)$ is crossed out and equals 1)

$$= \int (g^{'}(y) - 1) \log(g^{'}(y))$$

$$= D_{KL}(g^{'} || v) + D_{KL}(v || g^{'}) \geq 0 \qquad v = U(0,1) \text{ in } [0,1]$$

Jonas Peters et al, Elements of Causal Inference (2017)

Information Geometric Causal Inference (IGCI)

Previous results can be reformulated in information space

Daniusis et al., (2010)

$$X \to Y$$

$$C_{X \to Y} = D_{KL}(p_X || \mathcal{E}_X) - D_{KL}(p_y || \mathcal{E}_Y) \leq 0$$

$$Y \to X$$

$$C_{Y \to X} = D_{KL}(p_Y || \mathcal{E}_Y) - D_{KL}(p_X || \mathcal{E}_X) \leq 0$$

Reference distributions

For non-zero C (if the function is not ‘too simple’), sign of C determines the direction.

Jonas Peters et al, Elements of Causal Inference (2017)

Information Geometric Causal Inference (IGCI)

Daniusis et al., (2010)

Estimate C:

In case of uniform reference distribution:

$$C_{X\to Y} = D_{KL}(p_X||u) - D_{KL}(p_y||v)$$

$$= \int p_X(x)\log(p_X(x))dx - \int p_Y(y)\log(p_Y(y))dy$$

$$= \int p_X(x)\log(p_X(x))dx - \cancel{\int \log(p_X(x))\cdot p_X(x)dx} + \int p_X(x)\log(|f^{'}(x)|)dx$$

$$= \int p_X(x)\log(|f^{'}(x)|)dx$$

Which can be estimated from the data as:

$$C_{X\to Y} \approx \frac{1}{m-1}\sum_{i=1}^{m-1}\log\left|\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\right|$$

Jonas Peters et al, Elements of Causal Inference (2017)

Gene Perturbation Experiments

- Method: Invariant Causal Prediction (ICP)

- Quantifies confidence probabilities for inferring causal structures, i.e., ’error bars’

- Uses data from different experimental conditions/perturbations, i.e., heterogeneous

- Mix of observational and interventional data

$$Y = \sum_{k \in S^*} \gamma_k^* X_k + \epsilon_Y$$

Heterogeneity
Noise
Asymmetry

Main idea: Look for components of the regression vector that are invariant among various experimental settings.

Nicolai Meinshausen et al., PNAS, 2016

Invariant Causal Prediction (ICP): State-of-the-art

- Automatic identifiability
- Confidence bounds
- Intervention/perturbation do not need to exactly specified
- Avoids typically unstable/complicated estimating with graphs (equivalence classes from data)

Meinshausen et al. (2019)

Overview of the course

- Estimating causal effects
- Randomised trial vs observational data
- **Causal inference (of effects)** [DoWhy and others]
    - Rubin: Potential outcomes framework (observed confounders)
    - Rubin (unobserved confounders)
    - Simulations
    - Pearl: Structural causal models framework (observed and unobserved confounders)
    - Simulations
- **Causal discovery**
    - Constraint-based algorithms (PC)
    - Functional Causal Models

Outcomes of the course

- Be able to find and follow papers that have developed causal techniques
- Understand which area of causal analysis the papers apply to
- Be able to apply causal techniques to a particular problem of interest
- Use causal analysis packages in R and Python (Microsoft DoWhy, CausalGraphicalModels)
- Be able to modify a current technique in such a way that applies to a particular problem of interest
- A foundation to start developing techniques in causal inference and causal discovery

Causality in Biomedicine
Lecture Series: Lecture 7

Ava Khamseh

11 March, 2020