Causal Inference: Classical Approaches

Si Yi (Cathy) Meng

Feb 5, 2020

UBC MLRG

Outline

- Potential Outcomes
- Confounding and Causal DAGs
- Granger Causality
- ICA for Causal Discovery

Associational Inference

- Universe U
- For each unit $u \in U$:
  - Attribute variable X(u)
  - Observed variable $Y(u)$
- Inference:
  - $P(Y = y|X = x)$

Associational Inference

- Universe U
- For each unit $u \in U$:
  - Attribute variable X(u)
  - Observed variable $Y(u)$
- Inference:
  - $P(Y = y|X = x)$

Causal Inference

- Universe U
- For each unit $u \in U$:
  - Treatment variable $T(u) \in \{1,0\}$
  - **Potential outcome** $Y_1(u), Y_0(u)$
- Inference:
  - $Y_1(u) - Y_0(u)$

Rubin's Framework

- For each unit $u \in U$:
  - Treatment variable $T(u) \in \{1,0\}$
  - **Potential outcomes** $Y_1(u), Y_0(u)$
    - the outcome that would be observed if treatment was set to $T = 0$ or 1, on the same unit.
    - (before)

  - If $T(u)$ is set to 1
    - $Y_1(u)$ is the **observed outcome**
    - $Y_0(u)$ is the **counterfactual outcome**
    - (after)

Causal Effects

- $Y_1(u) - Y_0(u)$ is the **causal effect** of treatment 1 (relative to 0) on u.
  - Abbreviated as Y_1 and Y_0

Causal Effects

- $Y_1(u) - Y_0(u)$ is the **causal effect** of treatment 1 (relative to 0) on u.
  - Abbreviated as Y_1 and Y_0
- **Fundamental Problem of Causal Inference**
  - **It is impossible to *observe* both Y_1 and Y_0 on the same unit, and therefore it is impossible to observe the causal effect.**

THE END

Scientific solution to the Fundamental Problem

- Assume temporal stability and causal transience
  - The value of Y_0 does not depend on when $T = 0$ is applied and measured.
  - The effect of $T = 0$ and the measurement process that gives rise to Y_0 does not change u enough to affect Y_1 measured later.

Scientific solution to the Fundamental Problem

- Assume temporal stability and causal transience
  - The value of Y_0 does not depend on when $T = 0$ is applied and measured.
  - The effect of $T = 0$ and the measurement process that gives rise to Y_0 does not change u enough to affect Y_1 measured later.
  - With these two assumptions, we can simply measure both Y_0 and Y_1 by applying $T = 0$ then $T = 1$, taking the measurement after each exposure.
  - Widely used in experiments involving physical devices.

Scientific solution to the Fundamental Problem

- Assume temporal stability and causal transience
  - The value of Y_0 does not depend on when $T = 0$ is applied and measured.
  - The effect of $T = 0$ and the measurement process that gives rise to Y_0 does not change u enough to affect Y_1 measured later.
  - With these two assumptions, we can simply measure both Y_0 and Y_1 by applying $T = 0$ then $T = 1$, taking the measurement after each exposure.
  - Widely used in experiments involving physical devices.
- Assume unit homogeneity
  - For two units u_1 and u_2, we assume $Y_0(u_1) = Y_0(u_2)$ and $Y_1(u_1) = Y_1(u_2)$.

Scientific solution to the Fundamental Problem

- Assume temporal stability and causal transience
  - The value of Y_0 does not depend on when $T = 0$ is applied and measured.
  - The effect of $T = 0$ and the measurement process that gives rise to Y_0 does not change u enough to affect Y_1 measured later.
  - With these two assumptions, we can simply measure both Y_0 and Y_1 by applying $T = 0$ then $T = 1$, taking the measurement after each exposure.
  - Widely used in experiments involving physical devices.
- Assume unit homogeneity
  - For two units u_1 and u_2, we assume $Y_0(u_1) = Y_0(u_2)$ and $Y_1(u_1) = Y_1(u_2)$.
  - Causal effect can then be computed using $Y_1(u_1) - Y_0(u_2)$.
  - Implies the constant effect assumption: $Y_1(u) - Y_0(u)$ is the same for all $u \in U$.

Scientific solution to the Fundamental Problem

- Assume temporal stability and causal transience
  - The value of Y_0 does not depend on when $T = 0$ is applied and measured.
  - The effect of $T = 0$ and the measurement process that gives rise to Y_0 does not change u enough to affect Y_1 measured later.
  - With these two assumptions, we can simply measure both Y_0 and Y_1 by applying $T = 0$ then $T = 1$, taking the measurement after each exposure.
  - Widely used in experiments involving physical devices.
- Assume unit homogeneity
  - For two units u_1 and u_2, we assume $Y_0(u_1) = Y_0(u_2)$ and $Y_1(u_1) = Y_1(u_2)$.
  - Causal effect can then be computed using $Y_1(u_1) - Y_0(u_2)$.
  - Implies the constant effect assumption: $Y_1(u) - Y_0(u)$ is the same for all $u \in U$.
- It's very difficult to argue that these are valid…

Statistical solution to the Fundamental Problem

- *“What would have happened if I had not taken the flu shot” --> “What would the flu rate be if everyone got the flu shot vs if no one did?”*

Statistical solution to the Fundamental Problem

- *“What would have happened if I had not taken the flu shot” --> “What would the flu rate be if everyone got the flu shot vs if no one did?”*
- **Average causal effect** of $T = 1$ (relative to $T = 0$) over U:
  - $\mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y_1) - \mathbb{E}(Y_0)$
  - Imagine parallel universes with the same population…
  - Can’t observe this.
- Observed data can only give us information about the average of the outcome over $u \in U$ exposed to $T = t$.
  - $\mathbb{E}(Y_1 | T = 1) - \mathbb{E}(Y_0 | T = 0)$

Statistical solution to the Fundamental Problem

- *"What would have happened if I had not taken the flu shot" --> "What would the flu rate be if everyone got the flu shot vs if no one did?"*
- **Average causal effect** of $T = 1$ (relative to $T = 0$) over U:
  - $\mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y_1) - \mathbb{E}(Y_0)$
  - Imagine parallel universes with the same population…
  - Can't observe this.
- Observed data can only give us information about the average of the outcome over $u \in U$ exposed to $T = t$.
  - $\mathbb{E}(Y_1|T = 1) - \mathbb{E}(Y_0|T = 0)$
  - In general, $\mathbb{E}(Y_t) \neq \mathbb{E}(Y_t|T = \mathrm{t})$

Statistical solution to the Fundamental Problem

- *“What would have happened if I had not taken the flu shot” --> “What would the flu rate be if everyone got the flu shot vs if no one did?”*
- **Average causal effect** of $T = 1$ (relative to $T = 0$) over U:
  - $\mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y_1) - \mathbb{E}(Y_0)$
  - Imagine parallel universes with the same population…
  - Can’t observe this.
- Observed data can only give us information about the average of the outcome over $u \in U$ exposed to $T = t$.
  - $\mathbb{E}(Y_1|T = 1) - \mathbb{E}(Y_0|T = 0)$
  - In general, $\mathbb{E}(Y_t) \neq \mathbb{E}(Y_t|T = \mathrm{t})$
  - Independence assumption hold via randomized treatment assignment allows equality to hold, which lets us compute the ACE above.

Other assumptions

- Stable Unit Treatment Value Assumption (SUVTA)
  - No interference: units do not interact with each other.
  - One version of treatment.
- Consistency
  - The potential outcome Y_t is equal to the observed outcome if the actual treatment received is $T = t$.
- Positivity
  - $\mathbb{P}(T(u) = t) > 0$ for all t and u.

Other assumptions

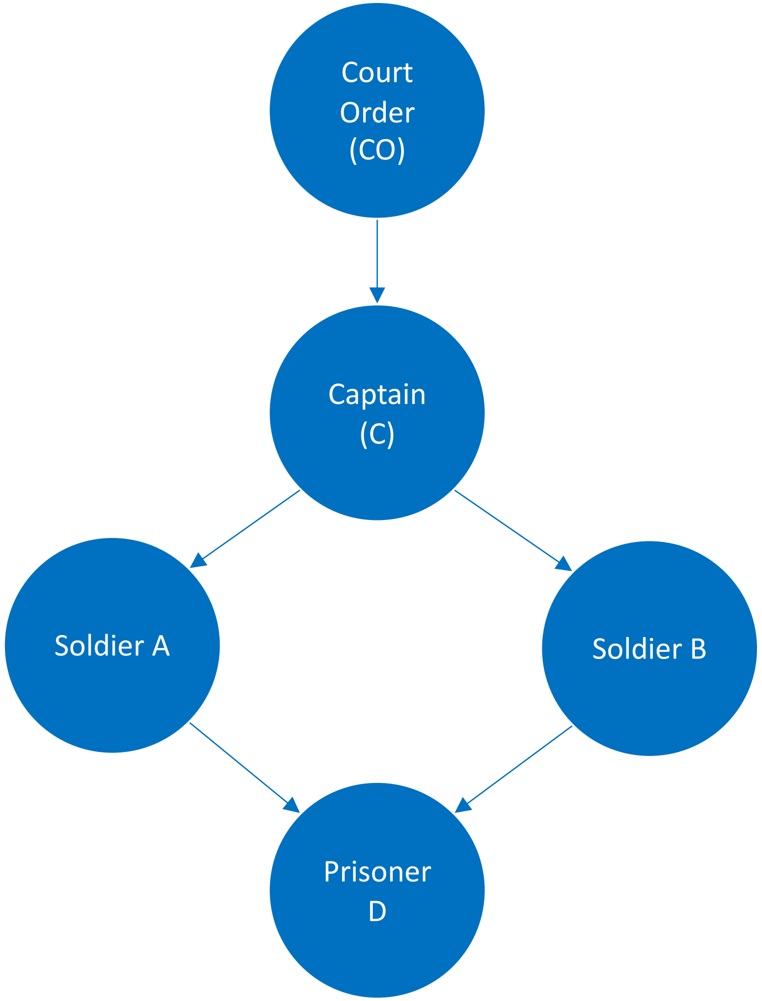
- Stable Unit Treatment Value Assumption (SUVTA)
  - No interference: units do not interact with each other.
  - One version of treatment.
- Consistency
  - The potential outcome Y_t is equal to the observed outcome if the actual treatment received is $T = t$.
- Positivity
  - $\mathbb{P}(T(u) = t) > 0$ for all t and u.
- **Ignorability** (aka no unmeasured confounders assumption)
  - $Y_0, Y_1 \perp \mathrm{T} | \mathrm{X}$
  - Among people with the same features X, we can think of treatment T as being randomly assigned.

Outline

- Potential Outcome
- Confounding and Causal DAGs
- Granger Causality
- ICA for Causal Discovery

DAGs

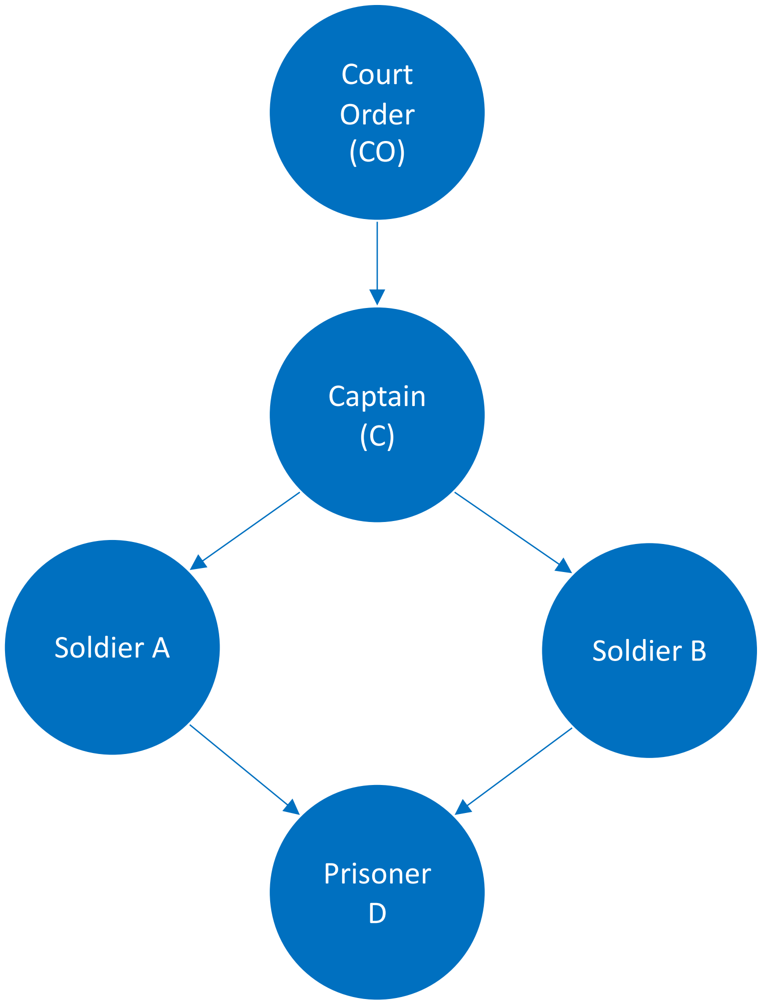
- Useful for identifying dependencies and ways to factor and simplify the joint distribution.
- $p(x_1, \ldots, x_n) = \prod_{\{i=1\}}^{n} p(x_i | x_{\{pa(i)\}})$



Firing squad example [Pearl, 2018]

DAGs

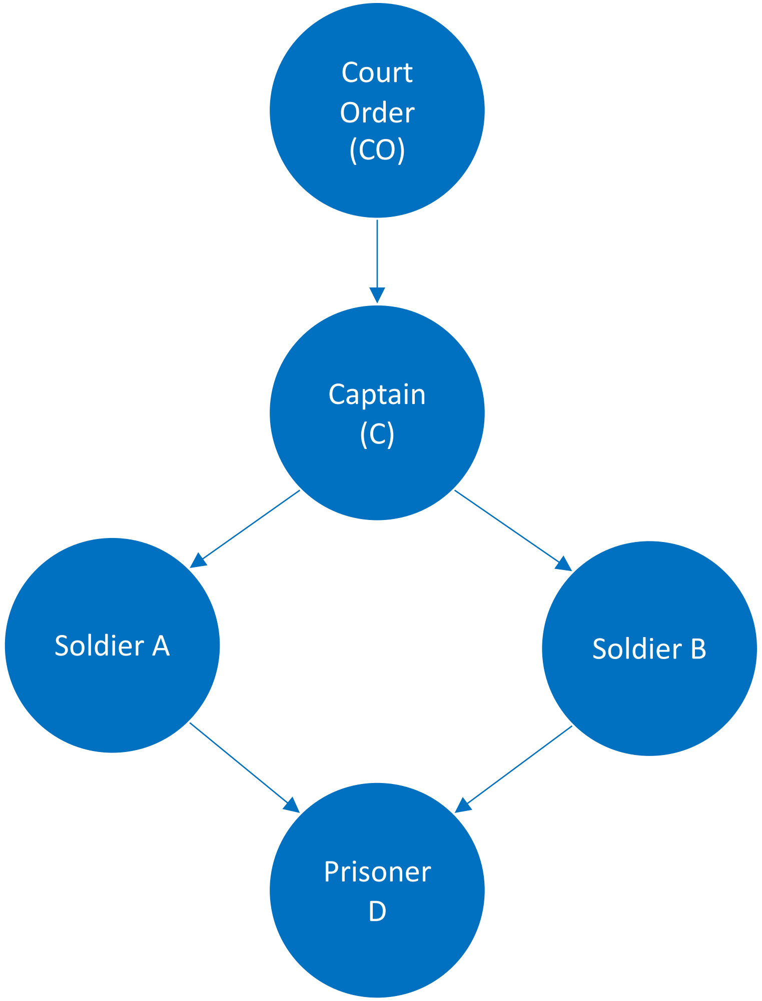
- Useful for identifying dependencies and ways to factor and simplify the joint distribution.
- $p(x_1, \dots, x_n) = \prod_{\{i=1\}}^{n} p(x_i | x_{\{pa(i)\}})$
- Two variables A and B are **d-separated** by a set of variables Z if A and B are conditionally independent given Z.
  - $p(A, B|Z) = p(A|Z)p(B|Z)$
  - Chain
  - Fork
  - Inverted fork



Firing squad example [Pearl, 2018]

Causal DAGs

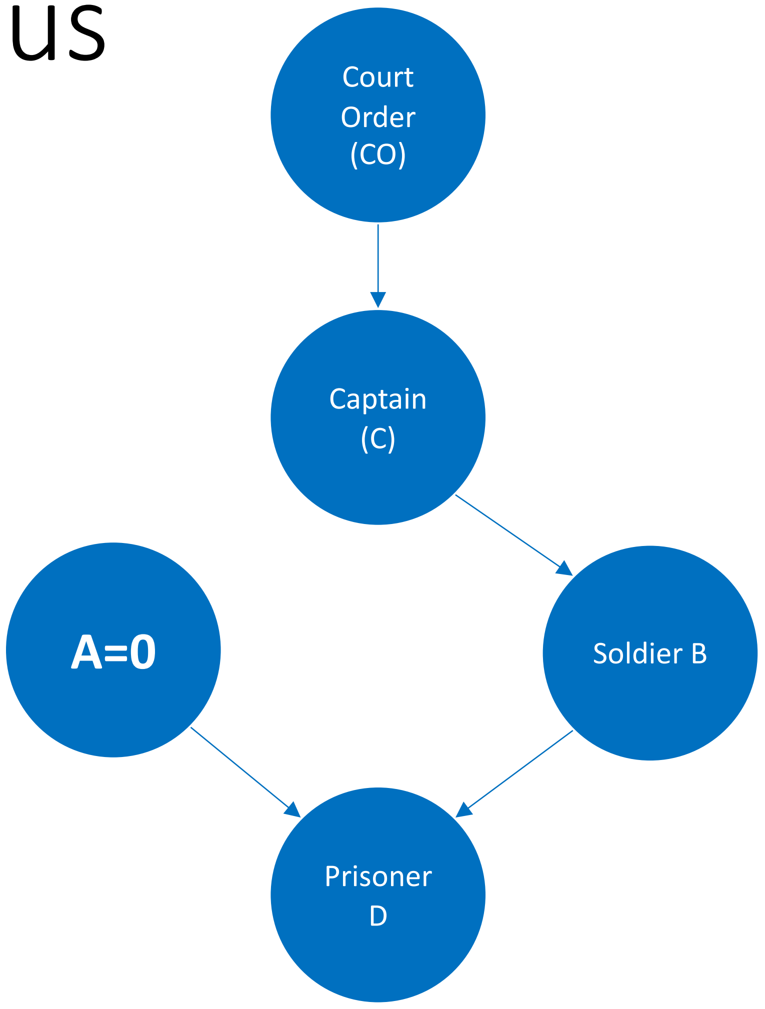
- DAGs where directions of the edges represent causal relationships.
- In contrast to Rubin's potential outcome framework, this is a structural approach to causal inference which Pearl advocates.
  - They are shown to be mathematically equivalent.



Firing squad example [Pearl, 2018]

Intervention and Pearl’s do-calculus

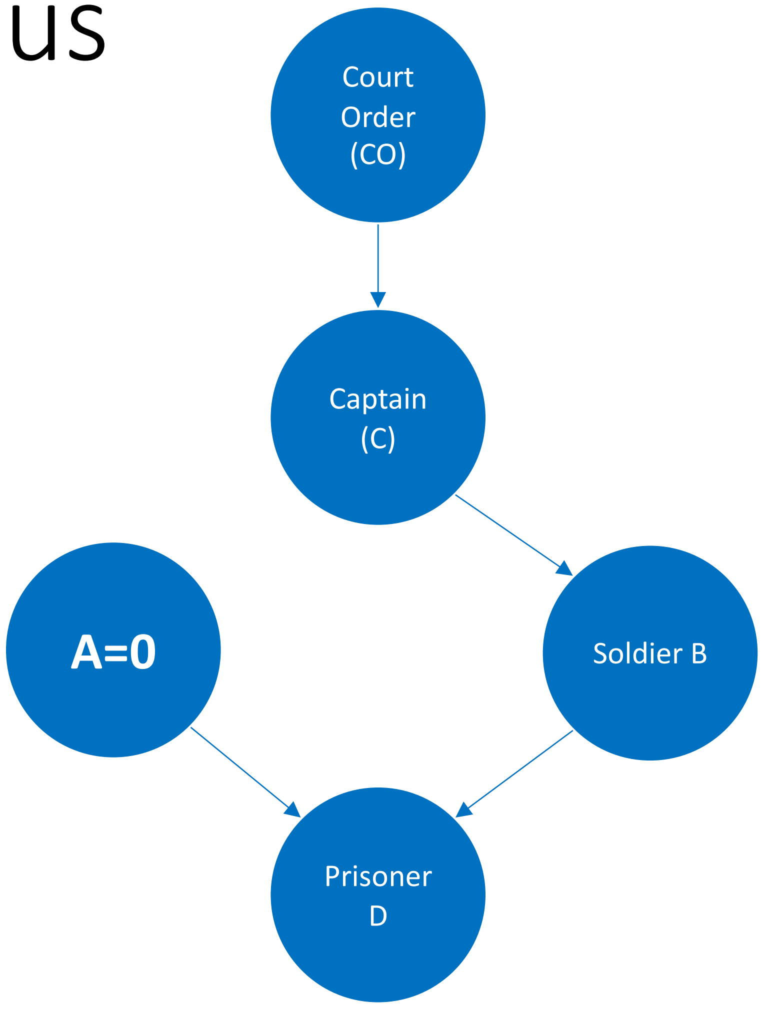
- $do()$ operator signals an intervention on a variable.
  - Replace that variable with the actual value that we assign.
  - Removes all incoming edges to that node.



Firing squad example [Pearl, 2018]

Intervention and Pearl's do-calculus

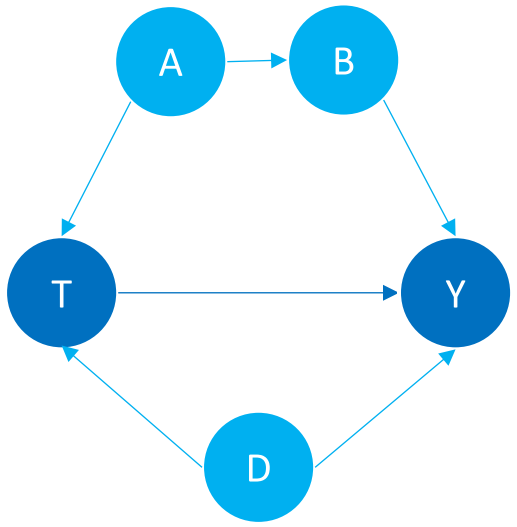
- $do()$ operator signals an intervention on a variable.
  - Replace that variable with the actual value that we assign.
  - Removes all incoming edges to that node.
- Instead of $p(D|A = 0)$
- We want $p(D|do(A = 0))$
  - The causal effect of $A = 0$ on D.



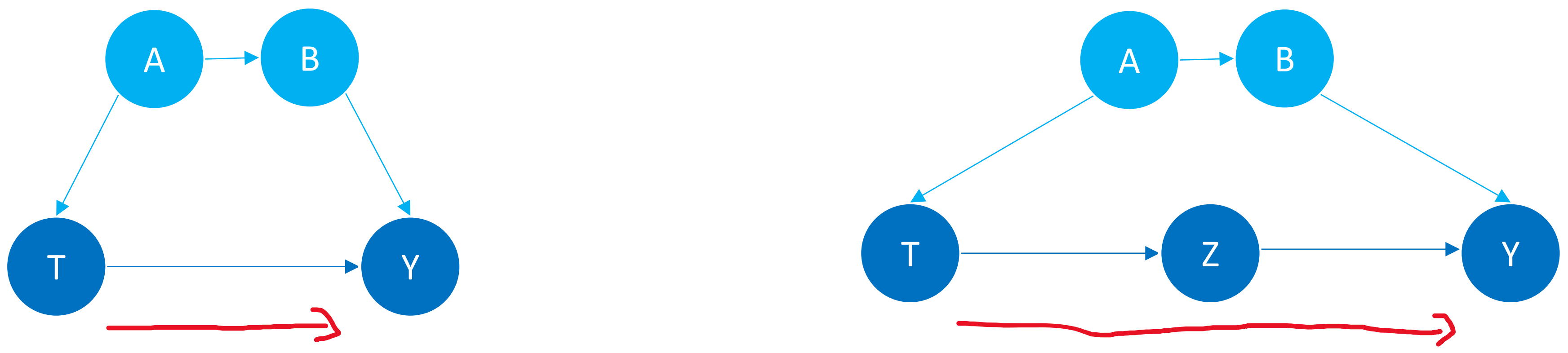
Firing squad example [Pearl, 2018]

Confounding

- **Confounders**: variables that influences both treatment and outcome.
  - **Want:** identify a set of variables so that ignorability holds.
  - We don’t need to identity specific confounders
  - We just need to be able to control for confounding.
- Need to block backdoor paths from T to Y.



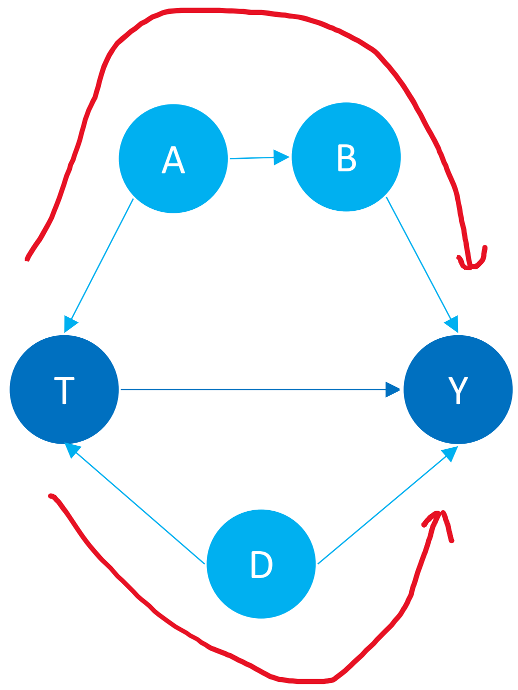
Frontdoor paths



- We are not concerned about frontdoor paths.
- We don't want to control anything along the frontdoor paths.
  - Unless we care about the magnitude of the causal effect…

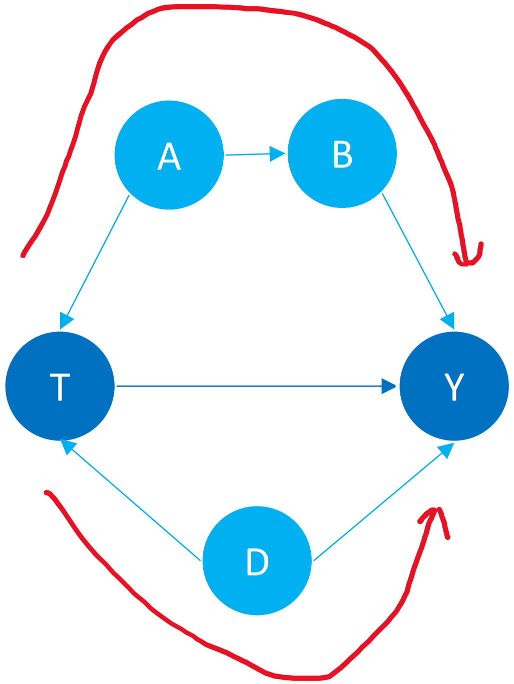
Backdoor paths

- **Begins with a parent of T and ends at Y.**
- Need to control these paths as they confound our causal effect.
- How?
  - Identify the set of variables that blocks all backdoor paths from T to Y.



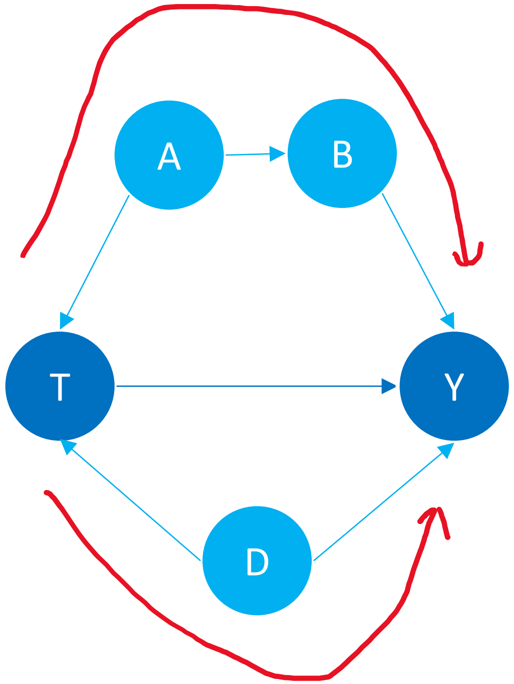
Backdoor criterion

- A set of variables C satisfies the **backdoor criterion** if
  1. it blocks all backdoor paths from T to Y, and
  2. It does not include any descendants of T.



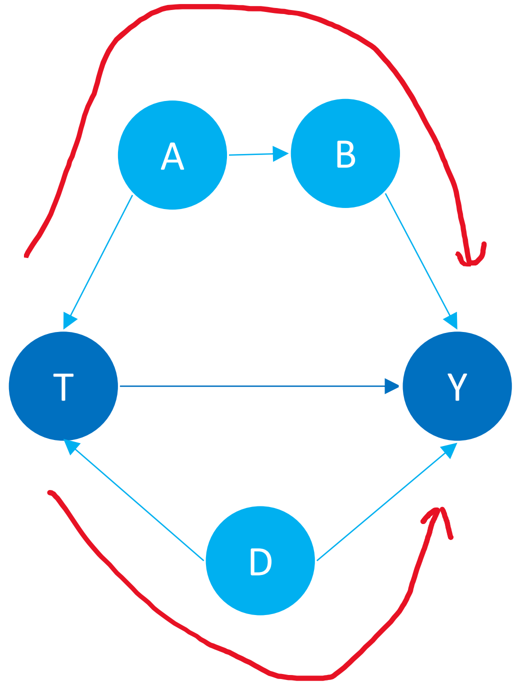
Backdoor criterion

- A set of variables C satisfies the **backdoor criterion** if
  1. it blocks all backdoor paths from T to Y, and
  2. It does not include any descendants of T.
- $C = \{A, D\}$
  - Alternatively, $C = \{B, D\}$, $C = \{A, B, D\}$



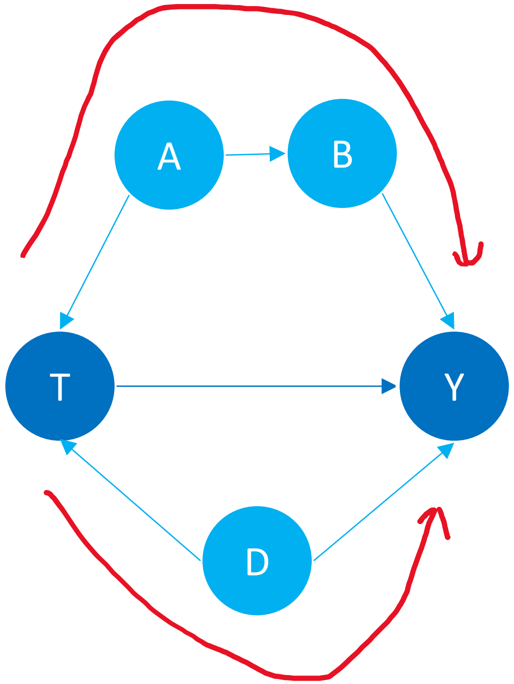
Backdoor criterion

- A set of variables C satisfies the **backdoor criterion** if
  1. it blocks all backdoor paths from T to Y, and
  2. It does not include any descendants of T.
- $C = \{A, D\}$
  - Alternatively, $C = \{B, D\}$, $C = \{A, B, D\}$
- Controlling any of these sets allow us to control for confounding.



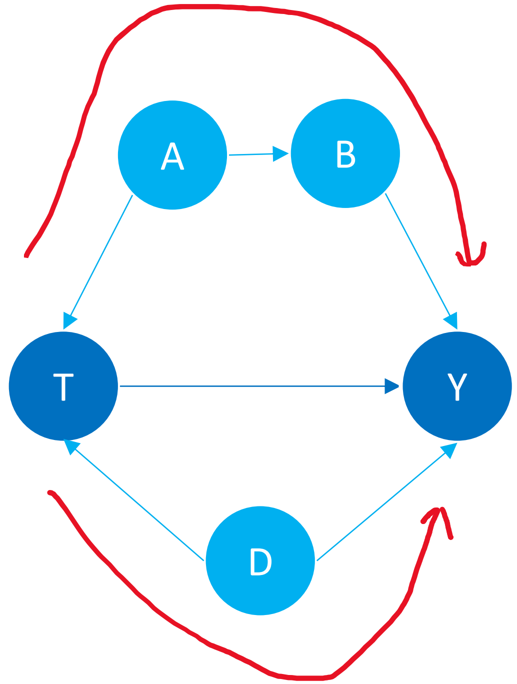
Backdoor criterion

- A set of variables C satisfies the **backdoor criterion** if
  1. it blocks all backdoor paths from T to Y, and
  2. It does not include any descendants of T.
- $C = \{A, D\}$
  - Alternatively, $C = \{B, D\}$, $C = \{A, B, D\}$
- Controlling any of these sets allow us to control for confounding.
- **Backdoor Adjustment:**
  - If a set of variables C satisfies the backdoor criterion relative to T ane Y, then the causal effect of T on Y is given by
  - $\mathbb{P}(Y|do(T=t)) = \sum_{c \in C} \mathbb{P}(Y|T=t,c)\mathbb{P}(c)$.



Backdoor criterion

- A set of variables C satisfies the **backdoor criterion** if
  1. it blocks all backdoor paths from T to Y, and
  2. It does not include any descendants of T.
- $C = \{A, D\}$
  - Alternatively, $C = \{B, D\}$, $C = \{A, B, D\}$
- Controlling any of these sets allow us to control for confounding.
- **Backdoor Adjustment:**
  - If a set of variables C satisfies the backdoor criterion relative to T ane Y, then the causal effect of T on Y is given by
  - $\mathbb{P}(Y|do(T = t)) = \sum_{c \in C} \mathbb{P}(Y|T = t, c)\mathbb{P}(c)$.
- In Rubin's framework, this is equivalent to the ignorability assumption:
  - Treatment assignment is effectively randomized given C.

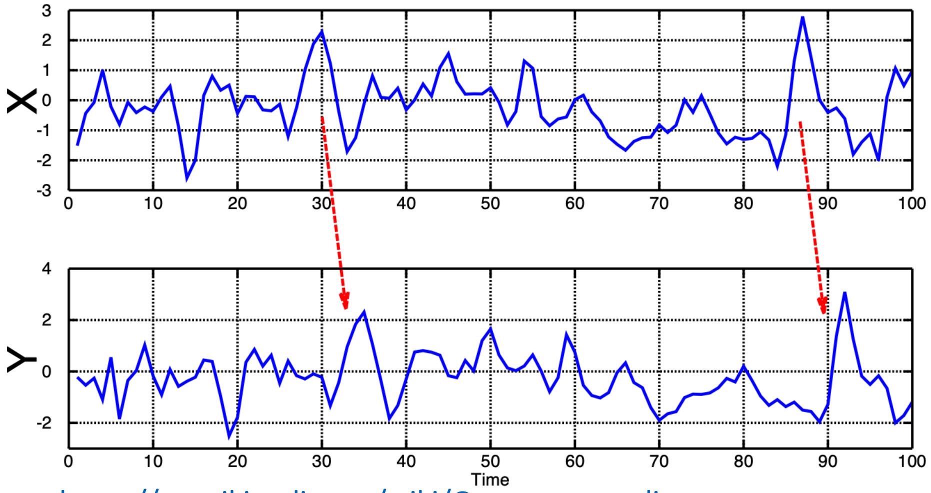


Outline

- Potential Outcomes
- Confounding and Causal DAGs
- Granger Causality
- ICA for Causal Discovery

Granger Causality

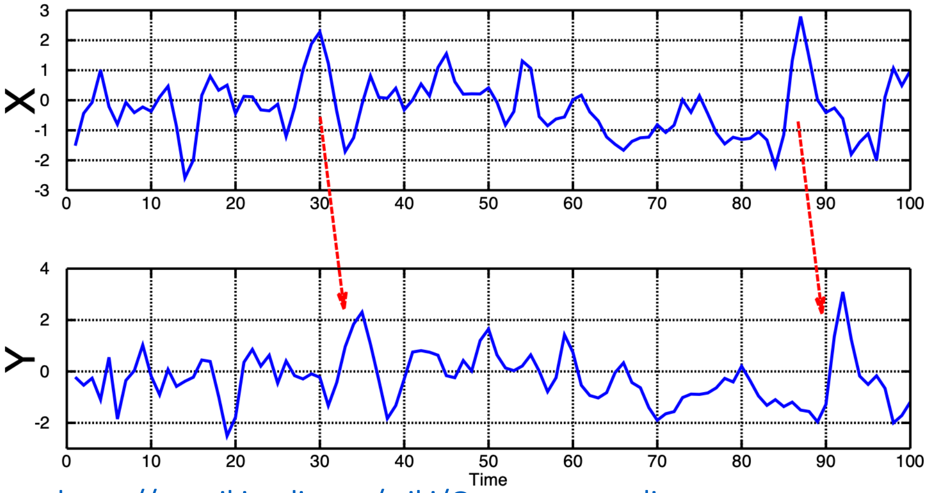
- Relationship between several time series.
- The **Granger causality test** is used to determine if the **past** values of $X(t)$ helps in predicting the **future** values of $Y(t)$.



https://en.wikipedia.org/wiki/Granger_causality

Granger Causality

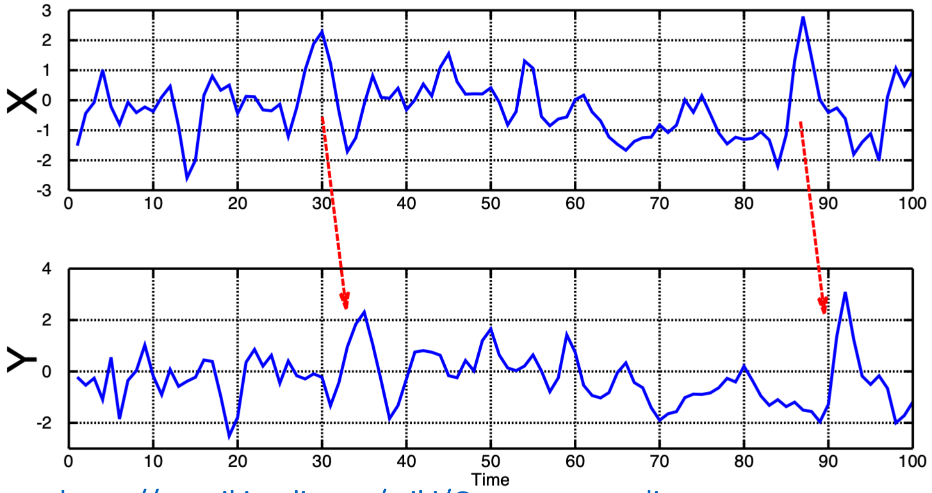
- Relationship between several time series.
- The **Granger causality test** is used to determine if the **past** values of $X(t)$ helps in predicting the **future** values of $Y(t)$.
- Two principles/assumptions:
  1. The cause happens prior to the effect.
  2. The cause has unique information about the future values of its effect.



https://en.wikipedia.org/wiki/Granger_causality

Granger Causality

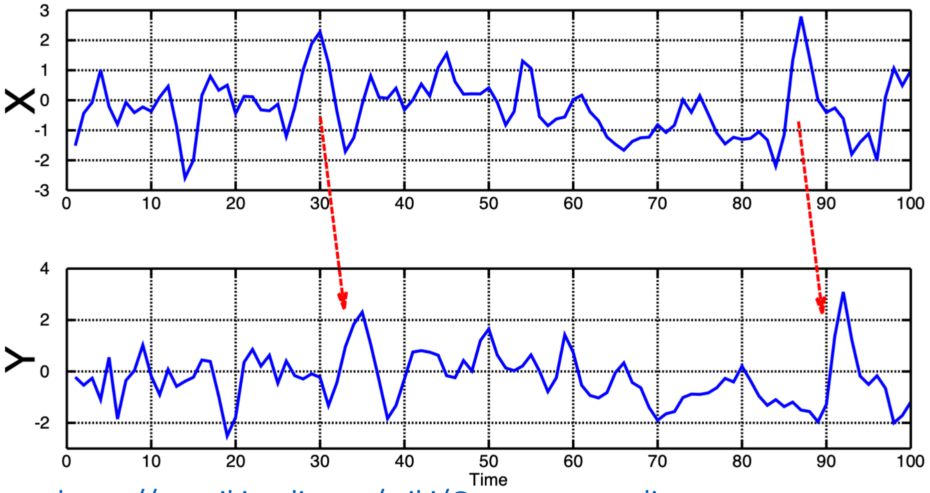
- Relationship between several time series.
- The **Granger causality test** is used to determine if the **past** values of $X(t)$ helps in predicting the **future** values of $Y(t)$.
- Two principles/assumptions:
  1. The cause happens prior to the effect.
  2. The cause has unique information about the future values of its effect.
- Hypothesis test:
  - $\mathbb{P}(Y(t+1)|I(t)) \neq \mathbb{P}\big(Y(t+1)\big|I_{\{-X\}}(t)\big)$
    - $I(t)$ all information up to time t
    - $I_{\{-X\}}(t)$ all information up to time t with X excluded.



https://en.wikipedia.org/wiki/Granger_causality

Granger Causality

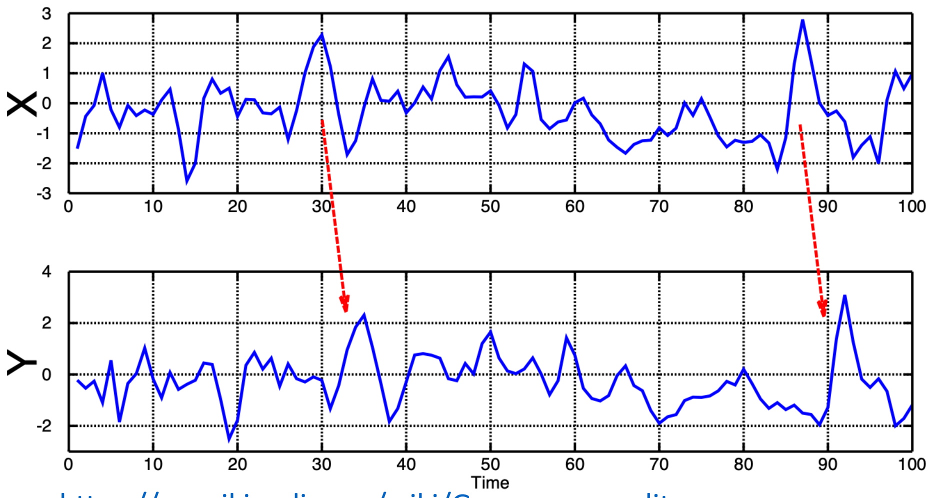
- Hypothesis test:
  - $\mathbb{P}(Y(t+1)|I(t)) \neq \mathbb{P}\big(Y(t+1)\big|I_{\{-X\}}(t)\big)$
    - $I(t)$ all information up to time t
    - $I_{\{-X\}}(t)$ all information up to time t with X excluded.
- Steps:
  - $y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \epsilon_t$



https://en.wikipedia.org/wiki/Granger_causality

Granger Causality

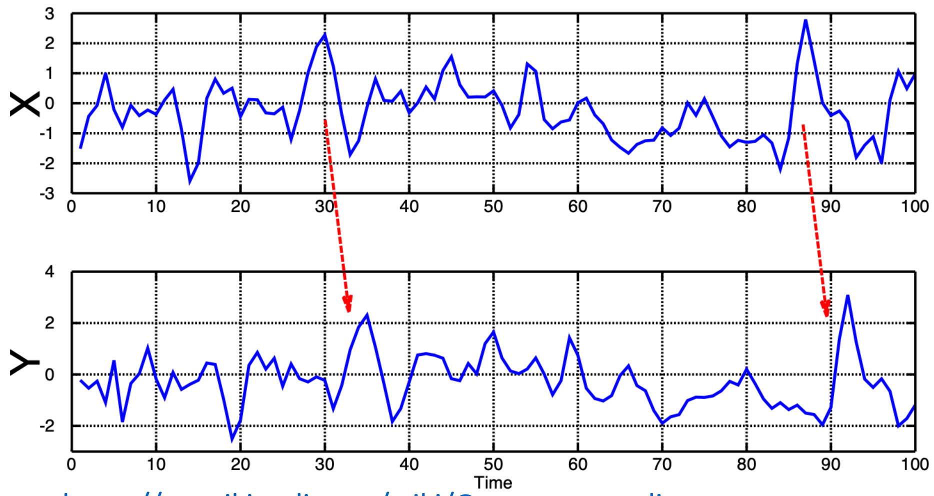
- Hypothesis test:
  - $\mathbb{P}(Y(t+1)|I(t)) \neq \mathbb{P}\big(Y(t+1)\big|I_{\{-X\}}(t)\big)$
    - $I(t)$ all information up to time t
    - $I_{\{-X\}}(t)$ all information up to time t with X excluded.
- Steps:
  - $y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \epsilon_t$
  - $y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + b_1 x_{t-1} + b_2 x_{t-2} + \epsilon_t$



https://en.wikipedia.org/wiki/Granger_causality

Granger Causality

- Hypothesis test:
  - $\mathbb{P}(Y(t+1)|I(t)) \neq \mathbb{P}\big(Y(t+1)\big|I_{\{-X\}}(t)\big)$
    - $I(t)$ all information up to time t
    - $I_{\{-X\}}(t)$ all information up to time t with X excluded.
- Steps:
  - $y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \epsilon_t$
  - $y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + b_1 x_{t-1} + b_2 x_{t-2} + \epsilon_t$
  - H_0: All $b_1 = \cdots = b_p = 0$
  - H_1: At least one is non-zero.
- Null hypothesis: X does not Granger cause Y iff no lagged values of x are retained.



https://en.wikipedia.org/wiki/Granger_causality

Remarks about GC

Holland

- GC formulation does not necessarily require the time-series setting adopted.
  - Time is only used to split up the variables.

Remarks about GC

Holland

- GC formulation does not necessarily require the time-series setting adopted.
  - Time is only used to split up the variables.
- Can be reformulated using Rubin's framework.
  - Conditional independence.
  - In a randomized experiment Granger noncausality implies zero ACE on all subpopulations defined by the values of $\{y_{t-1} \ldots\}$.

Remarks about GC

Holland

- GC formulation does not necessarily require the time-series setting adopted.
  - Time is only used to split up the variables.
- Can be reformulated using Rubin's framework.
  - Conditional independence.
  - In a randomized experiment Granger noncausality implies zero ACE on all subpopulations defined by the values of $\{y_{t-1} \ldots\}$.
- Granger causes are "temporary".
  - Adding more lags may change the overall Granger causes.

Remarks about GC

Holland

- GC formulation does not necessarily require the time-series setting adopted.
  - Time is only used to split up the variables.
- Can be reformulated using Rubin's framework.
  - Conditional independence.
  - In a randomized experiment Granger noncausality implies zero ACE on all subpopulations defined by the values of $\{y_{t-1} \dots\}$.
- Granger causes are "temporary".
  - Adding more lags may change the overall Granger causes.

Pearl

Outline

- Potential Outcomes
- Confounding and Causal DAGs
- Granger Causality
- ICA for Causal Discovery

Independent Component Analysis (ICA)

- PCA: $X = ZW$ where Z is $n \times k$ and W is $k \times d$, $k \leq d$
  - Factor analysis, data compression, etc.
  - Invariant to rotation.

Independent Component Analysis (ICA)

- PCA: $X = ZW$ where Z is $n \times k$ and W is $k \times d$, $k \leq d$
  - Factor analysis, data compression, etc.
  - Invariant to rotation.
- ICA: $X = ZW$ usually with $k = d$
  - **Require the components of each z_i to be independent, and at most one can be normally distributed.**
  - Independence is measured by non-normality.
  - $W = \mathrm{argmax}_{\mathrm{W}} \sum_{i=1}^{n} \sum_{j=1}^{k} kurt(w_j^T x_i)^2$ where $kurt(u) = \mathbb{E}(u^4) - 3\left(\mathbb{E}(u^2)\right)^2$
  - Up to permutation and scaling, we can identify the factors W.

Causal Discovery

- Randomized control trials are not always feasible.
  - Cost, ethics, etc.

Causal Discovery

- Randomized control trials are not always feasible.
  - Cost, ethics, etc.
- Structural equation modeling (SEM): path analysis integrating factor analysis and latent variables for non-experimental data.
  - But requires background knowledge BEFORE collecting and analyzing data.
  - Normality constraints.
  - **Causal directions are often unknown.**
  - Model 1: $x_1 = b_{12}x_2 + \epsilon_1$ and Model 2: $x_2 = b_{21}x_1 + \epsilon_1$, both are saturated with the same covariance matrix.

Causal Discovery

- Randomized control trials are not always feasible.
  - Cost, ethics, etc.
- Structural equation modeling (SEM): path analysis integrating factor analysis and latent variables for non-experimental data.
  - But requires background knowledge BEFORE collecting and analyzing data.
  - Normality constraints.
  - **Causal directions are often unknown.**
  - Model 1: $x_1 = b_{12}x_2 + \epsilon_1$ and Model 2: $x_2 = b_{21}x_1 + \epsilon_1$, both are saturated with the same covariance matrix.
- With non-normality, ICA can be used to find the causal ordering of any number of observed variables based on non-experimental data. (Shimizu et al. 2006)

Causal Order

- Causality (a causal order) from a random variable x_1 to another random variable x_2, denoted by $x_1 \rightarrow x_2$ is confirmed if the following holds:
  - $x_2 = f(x_1, \epsilon_2)$ where **ϵ_2 is some perturbation independently distributed from x_1**.
  - Assume $f(x_1, \epsilon_2) = b_{21}x_1 + \epsilon_2$.
  - Boils down to finding b_{21}.

Causal Order

- Causality (a causal order) from a random variable x_1 to another random variable x_2, denoted by $x_1 \to x_2$ is confirmed if the following holds:
  - $x_2 = f(x_1, \epsilon_2)$ where **ϵ_2 is some perturbation independently distributed from x_1**.
  - Assume $f(x_1, \epsilon_2) = b_{21}x_1 + \epsilon_2$.
  - Boils down to finding b_{21}.
- Note that we can't just have uncorrelatedness from x_1 and ϵ_2:
  - Independence: $\mathbb{E}\big(g(x)h(y)\big) = \mathbb{E}(g(x))\mathbb{E}(h(y))$ for any deterministic functions g and h.
  - Uncorrelatedness (covariance is 0): $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$, weaker assumption.

Causal Order

- Causality (a causal order) from a random variable x_1 to another random variable x_2 , denoted by $x_1 \rightarrow x_2$ is confirmed if the following holds:
  - $x_2 = f(x_1, \epsilon_2)$ where **ϵ_2 is some perturbation independently distributed from x_1**.
  - Assume $f(x_1, \epsilon_2) = b_{21}x_1 + \epsilon_2$.
  - Boils down to finding b_{21}.
- Note that we can't just have uncorrelatedness from x_1 and ϵ_2:
  - Independence: $\mathbb{E}\big(g(x)h(y)\big) = \mathbb{E}(g(x))\mathbb{E}(h(y))$ for any deterministic functions g and h.
  - Uncorrelatedness (covariance is 0): $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$, weaker assumption.
  - If x_1 and ϵ_2 are only uncorrelated, then there could exist unobserved confounder z that affects both x_1 and x_2.
  - Suppose $x_1 = \alpha z + \epsilon_1$ and $x_2 = \beta z + b_{21}x_1 + \epsilon_2$, then $Cov(x_1, x_2) = b_{21}Var(x_1) + \alpha\beta Var(z)$ can be non-zero even if b_{21} is 0.

Finding a Causal Order

- Suppose we have N measurements of x_1 and x_2.
  - $\overline{x_1^2} = \frac{1}{N}\sum_{i=1}^{N} x_{1i}^2$, similarly for $\overline{x_2^2}$ and $\overline{x_1 x_2}$
- Model 1: $x_1 = b_{12}x_2 + \epsilon_1$
- Model 2: $x_2 = b_{21}x_1 + \epsilon_2$

Finding a Causal Order

- Suppose we have N measurements of x_1 and x_2.
  - $\overline{x_1^2} = \frac{1}{N}\sum_{i=1}^{N} x_{1i}^2$, similarly for $\overline{x_2^2}$ and $\overline{x_1 x_2}$
- Model 1: $x_1 = b_{12}x_2 + \epsilon_1$
- Model 2: $x_2 = b_{21}x_1 + \epsilon_2$
- Second order moments of Model 1:

$$E\begin{bmatrix} \overline{x_1^2} \\ \overline{x_1 x_2} \\ \overline{x_2^2} \end{bmatrix} = \begin{bmatrix} b_{12}^2 E(x_2^2) + E(\xi_1^2) \\ b_{12} E(x_2^2) \\ E(x_2^2) \end{bmatrix} \quad \text{which we denote by} \quad E[\boldsymbol{m}_2] = \boldsymbol{\sigma}_2(\boldsymbol{\tau}_2)$$

$$\boldsymbol{\tau}_2 = [E(x_2^2),\ E(\xi_1^2),\ b_{12}]^T$$

- Symmetric for Model 2.
- Undistinguishable.

Finding a Causal Order

- Suppose we have N measurements of x_1 and x_2.
  - $\overline{x_1^2} = \frac{1}{N}\sum_{i=1}^{N} x_{1i}^2$, similarly for $\overline{x_2^2}$ and $\overline{x_1 x_2}$
- Model 1: $x_1 = b_{12}x_2 + \epsilon_1$
- Model 2: $x_2 = b_{21}x_1 + \epsilon_2$
- Second order moments of Model 1:

$$E\begin{bmatrix} \overline{x_1^2} \\ \overline{x_1 x_2} \\ \overline{x_2^2} \end{bmatrix} = \begin{bmatrix} b_{12}^2 E(x_2^2) + E(\xi_1^2) \\ b_{12}E(x_2^2) \\ E(x_2^2) \end{bmatrix}$$

which we denote by $E[\boldsymbol{m}_2] = \boldsymbol{\sigma}_2(\boldsymbol{\tau}_2)$

$\boldsymbol{\tau}_2 = [E(x_2^2),\ E(\xi_1^2),\ b_{12}]^T$

- Symmetric for Model 2.
- Undistinguishable.

- Fourth order moments for Model 1 and assume the residuals are NOT normally distributed:

$$E\begin{bmatrix} \overline{x_1^4} \\ \overline{x_1^3 x_2} \\ \overline{x_1^2 x_2^2} \\ \overline{x_1 x_2^3} \\ \overline{x_2^4} \end{bmatrix} = \begin{bmatrix} b_{12}^4 E(x_2^4) + 6b_{12}^2 E(x_2^2)E(\xi_1^2) + E(\xi_1^4) \\ b_{12}^3 E(x_2^4) + 3b_{12}E(x_2^2)E(\xi_1^2) \\ b_{12}^2 E(x_2^4) + E(x_2^2)E(\xi_1^2) \\ b_{12}E(x_2^4) \\ E(x_2^4) \end{bmatrix}$$

which we denote by $E[\boldsymbol{m}_4] = \boldsymbol{\sigma}_4(\boldsymbol{\tau}_4)$,

$\boldsymbol{\tau}_4 = [\boldsymbol{\tau}_2^T,\ E(x_2^4),\ E(\xi_1^4)]^T.$

Finding a Causal Order

- Suppose we have N measurements of x_1 and x_2.
  - $\overline{x_1^2} = \frac{1}{N}\sum_{i=1}^{N} x_{1i}^2$, similarly for $\overline{x_2^2}$ and $\overline{x_1 x_2}$
- Model 1: $x_1 = b_{12}x_2 + \epsilon_1$
- Model 2: $x_2 = b_{21}x_1 + \epsilon_2$
- Second order moments of Model 1:

$$E\begin{bmatrix}\overline{x_1^2} \\ \overline{x_1 x_2} \\ \overline{x_2^2}\end{bmatrix} = \begin{bmatrix} b_{12}^2 E(x_2^2) + E(\xi_1^2) \\ b_{12}E(x_2^2) \\ E(x_2^2)\end{bmatrix}$$

which we denote by $E[\boldsymbol{m}_2] = \boldsymbol{\sigma}_2(\boldsymbol{\tau}_2)$

$\boldsymbol{\tau}_2 = [E(x_2^2),\ E(\xi_1^2),\ b_{12}]^T$

- Symmetric for Model 2.
- Undistinguishable.

- Fourth order moments for Model 1 and assume the residuals are NOT normally distributed:

$$E\begin{bmatrix}\overline{x_1^4} \\ \overline{x_1^3 x_2} \\ \overline{x_1^2 x_2^2} \\ \overline{x_1 x_2^3} \\ \overline{x_2^4}\end{bmatrix} = \begin{bmatrix} b_{12}^4 E(x_2^4) + 6b_{12}^2 E(x_2^2)E(\xi_1^2) + E(\xi_1^4) \\ b_{12}^3 E(x_2^4) + 3b_{12}E(x_2^2)E(\xi_1^2) \\ b_{12}^2 E(x_2^4) + E(x_2^2)E(\xi_1^2) \\ b_{12}E(x_2^4) \\ E(x_2^4)\end{bmatrix}$$

which we denote by $E[\boldsymbol{m}_4] = \boldsymbol{\sigma}_4(\boldsymbol{\tau}_4)$,

$\boldsymbol{\tau}_4 = [\boldsymbol{\tau}_2^T,\ E(x_2^4),\ E(\xi_1^4)]^T$.

$$T = N\left(\begin{bmatrix}\boldsymbol{m}_2 \\ \boldsymbol{m}_4\end{bmatrix} - \begin{bmatrix}\boldsymbol{\sigma}_2(\hat{\boldsymbol{\tau}}_2) \\ \boldsymbol{\sigma}_4(\hat{\boldsymbol{\tau}}_4)\end{bmatrix}\right)^T \hat{M} \left(\begin{bmatrix}\boldsymbol{m}_2 \\ \boldsymbol{m}_4\end{bmatrix} - \begin{bmatrix}\boldsymbol{\sigma}_2(\hat{\boldsymbol{\tau}}_2) \\ \boldsymbol{\sigma}_4(\hat{\boldsymbol{\tau}}_4)\end{bmatrix}\right),$$

- Measure of model fit $\approx$ distance between data and the model used
- **Compare T for the two models which will imply causal direction.**

Causal ordering - more than 2 variables

- Suppose we have n variables, and we want to find an ordering $i(1), \dots, i(n)$ such that
  - $x_{i(j)} = \sum_{k=1}^{j-1} b_{i(j),i(k)} x_{i(k)} + \epsilon_{i(j)}$ for all $j = 1, \dots, n$, with nonzero coefficients and $\epsilon_{i(j)}$ non-normal, independent from $x_{i(k)}$ for $k < j$.
  - "$x_{i(j)}$ can be written as a linear combination of its preceding variables in that order plus an (mutually) independent error."
  - Causal ordering $x_{i(1)} \to x_{i(2)} \to \cdots \to x_{i(n)}$

Causal ordering - more than 2 variables

- Suppose we have n variables, and we want to find an ordering $i(1), \ldots, i(n)$ such that
  - $x_{i(j)} = \sum_{k=1}^{j-1} b_{i(j),i(k)} x_{i(k)} + \epsilon_{i(j)}$ for all $j = 1, \ldots, n$, with nonzero coefficients and $\epsilon_{i(j)}$ non-normal, independent from $x_{i(k)}$ for $k < j$.
  - "$x_{i(j)}$ can be written as a linear combination of its preceding variables in that order plus an (mutually) independent error."
  - Causal ordering $x_{i(1)} \to x_{i(2)} \to \cdots \to x_{i(n)}$
- Assuming such model exists, we need find the correct mapping $i(j)$ for $j = 1, \ldots, n$.
- $\tilde{x} = B\tilde{x} + \tilde{\epsilon}$ where B is lower triangular and $\tilde{x}$ is a vector of the observed variables with the desired ordering.

Causal ordering - more than 2 variables

- Suppose we have n variables, and we want to find an ordering $i(1), \dots, i(n)$ such that
  - $x_{i(j)} = \sum_{k=1}^{j-1} b_{i(j),i(k)} x_{i(k)} + \epsilon_{i(j)}$ for all $j = 1, \dots, n$, with nonzero coefficients and $\epsilon_{i(j)}$ non-normal, independent from $x_{i(k)}$ for $k < j$.
  - "$x_{i(j)}$ can be written as a linear combination of its preceding variables in that order plus an (mutually) independent error."
  - Causal ordering $x_{i(1)} \to x_{i(2)} \to \cdots \to x_{i(n)}$
- Assuming such model exists, we need find the correct mapping $i(j)$ for $j = 1, \dots, n$.
- $\tilde{x} = B\tilde{x} + \tilde{\epsilon}$ where B is lower triangular and $\tilde{x}$ is a vector of the observed variables with the desired ordering.
- Normalize the above so that the errors have unit variance using
  - $w_{i(j),i(j)} = 1/\sqrt{Var(\epsilon_{i(j)})}$ and $w_{i(j),i(k)} = -b_{i(j),i(k)}/\sqrt{Var(\epsilon_{i(j)})}$ for $k \neq j$
- $\mathrm{diag}(W)\tilde{x} = -\mathrm{offdiag}(W)\tilde{x} + \tilde{\epsilon}^*$, or $W\tilde{x} = \tilde{\epsilon}^*$

Causal ordering - more than 2 variables

- Suppose we have n variables, and we want to find an ordering $i(1), \ldots, i(n)$ such that
  - $x_{i(j)} = \sum_{k=1}^{j-1} b_{i(j),i(k)} x_{i(k)} + \epsilon_{i(j)}$ for all $j = 1, \ldots, n$, with nonzero coefficients and $\epsilon_{i(j)}$ non-normal, independent from $x_{i(k)}$ for $k < j$.
  - "$x_{i(j)}$ can be written as a linear combination of its preceding variables in that order plus an (mutually) independent error."
  - Causal ordering $x_{i(1)} \rightarrow x_{i(2)} \rightarrow \cdots \rightarrow x_{i(n)}$
- Assuming such model exists, we need find the correct mapping $i(j)$ for $j = 1, \ldots, n$.
- $\tilde{x} = B\tilde{x} + \tilde{\epsilon}$ where B is lower triangular and $\tilde{x}$ is a vector of the observed variables with the desired ordering.
- Normalize the above so that the errors have unit variance using
  - $w_{i(j),i(j)} = 1/\sqrt{Var(\epsilon_{i(j)})}$ and $w_{i(j),i(k)} = -b_{i(j),i(k)}/\sqrt{Var(\epsilon_{i(j)})}$ for $k \neq j$
- $\text{diag}(W)\tilde{x} = -\text{offdiag}(W)\tilde{x} + \tilde{\epsilon}^*$, or $W\tilde{x} = \tilde{\epsilon}^*$
- We can use ICA to estimate W!
- And there exists a unique permutation to make W lower triangular if the coefficients in B are non-zero.

Causal ordering - more than 2 variables

- Suppose we have n variables, and we want to find an ordering $i(1), \ldots, i(n)$ such that
  - $x_{i(j)} = \sum_{k=1}^{j-1} b_{i(j),i(k)} x_{i(k)} + \epsilon_{i(j)}$ for all $j = 1, \ldots, n$, with nonzero coefficients and $\epsilon_{i(j)}$ non-normal, independent from $x_{i(k)}$ for $k < j$.
  - "$x_{i(j)}$ can be written as a linear combination of its preceding variables in that order plus an (mutually) independent error."
  - Causal ordering $x_{i(1)} \rightarrow x_{i(2)} \rightarrow \cdots \rightarrow x_{i(n)}$
- Assuming such model exists, we need find the correct mapping $i(j)$ for $j = 1, \ldots, n$.
- $\tilde{x} = B\tilde{x} + \tilde{\epsilon}$ where B is lower triangular and $\tilde{x}$ is a vector of the observed variables with the desired ordering.
- Normalize the above so that the errors have unit variance using
  - $w_{i(j),i(j)} = 1/\sqrt{Var(\epsilon_{i(j)})}$ and $w_{i(j),i(k)} = -b_{i(j),i(k)}/\sqrt{Var(\epsilon_{i(j)})}$ for $k \neq j$
- diag$(W)\tilde{x} = -$offdiag$(W)\tilde{x} + \tilde{\epsilon}^*$, or $W\tilde{x} = \tilde{\epsilon}^*$
- We can use ICA to estimate W!
- And there exists a unique permutation to make W lower triangular if the coefficients in B are non-zero.
- We can also estimate B, depending whether we care about the coefficients or just the causal ordering.

Summary (and References)

- Rubin’s framework
  - Potential outcomes and counterfactuals.
    - Holland, P. W. Statistics and Causal Inference. *Journal of the American statistical Association,* 1986.
    - Roy, J. A Crash course in Causality.
- Pearl’s framework
  - Utilizing causal DAGs, do-operator, backdoor adjustment.
    - Pearl, J. *Causality*. Cambridge University Press, 2009.
    - Pearl, J. and Mackenzie D. *The book of why: the new sciences of cause and effect*. 2018
- Granger Causality
  - Using past values of one variable to predict future values of another.
    - Granger, C. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 1969.
- ICA for Causal Discovery (Shimizu et al.)
  - With the non-normality assumption and independence assumption, can find causal directions of a set of variables using non-experimental data.
    - Shimizu, S., Shimizu, S., Hyvärinen, A., Hoyer, P.O. and Kano, Y. Finding a causal ordering via independent component analysis. *Computational Statistics & Data Analysis*, 2006.

Thank you