Causality: inference under changes

Group 1: Arrigo, Di Maria, Martello, Rubino

April 26, 2017

Outline

1. **What is causality?**
   1.1 Probabilistic vs Deterministic causality
   1.2 Correlation and causation

2. **Bayesian Networks**
   2.1 Probability and DAG
   2.2 The d-separation criterion

3. **Causal Inference**
   3.1 Intervention and do - calculus
   3.2 Identification of causal quantities

4. Structural Equations Models (SEMs)

4.1 Predictions

4.2 Interventions and causal effects in functional models

5. Counterfactuals

6. A practical application

Conclusions

References

1. What is causality?

- What do you understand when you hear the sentence Smoking causes Lung Cancer?
  - If I start smoking (continue smoking), I change (increase or decrease) my chances of getting Lung Cancer, relative to if I stop smoking (continue not smoking)?
  - How do you write this mathematically?

1.1 Probabilistic vs Deterministic causality

Deterministic causality (as typically used in language): A causes B (for binary events A and B) means: If an external mechanism makes A occur, then B will occur.

Probabilistic causality (used in this talk): A causes B means: If an external mechanism makes A occur, the probability that B will occur changes.

Many philosophical problems with this definition...

Deterministic causality (as typically used in language): A causes B (for binary events A and B) means: If an external mechanism makes A occur, then B will occur.

Probabilistic causality (used in this talk): For variables with multiple values: If an external mechanism sets the values of A, the statistical distribution of B may change.

Many philosophical problems with this definition...

1.2 Correlation and causation

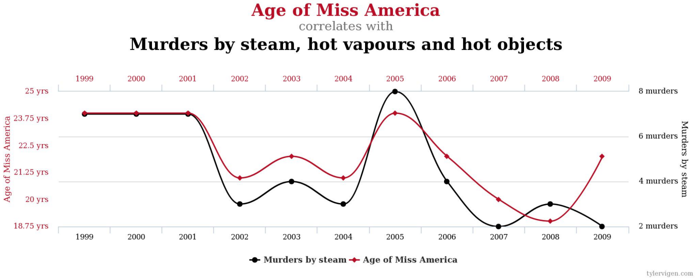


Figure 1: Data sources: Wikipedia and Centers for Disease Control and Prevention. Correlation: 87.01 % (r=0.870127).

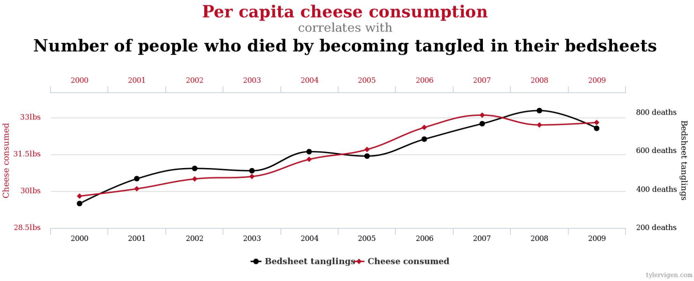


Figure 2: Data sources: U.S. Department of Agriculture and Centers for Disease Control and Prevention.

Correlation: 94.71 % (r=0.947091).

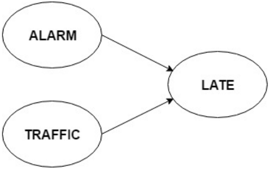


Figure 3

- Assumptions about the world
- Other omitted causes ⟶ NOISE

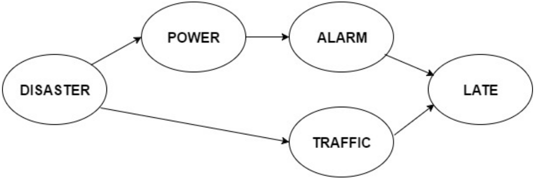


Figure 4

- Correlation does not imply causation
- There is no correlation without causation (If there is a correlation, there must be a **common cause** ⟶ Confounder) ⟶ Simpson's Paradox

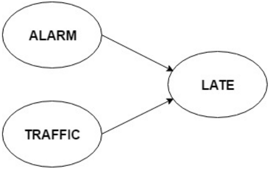


Figure 3

Alarm and traffic are independent but NOT conditionally independent.

Berkson's Paradox: Conditioning on common effects results in two causes becoming correlated, even if they were uncorrelated originally.

Example: Selection Bias

Solution

- Not conditioning on downstream effects
- Conditioning on common causes (confounding factors)

BUT you have to know the right picture of the world (staticity).

2. Bayesian Networks

V $\longrightarrow$ set of vertices (variables)

E $\longrightarrow$ set of links (causal influences).

A DAG (Directed acyclic graph) simulates the causal mechanism which operates in the environment; it represents and responds to changing configurations.

Three kinds of queries:

- associational $\longrightarrow$ probabilistic
- abductive $\longrightarrow$ causal
- control $\longrightarrow$ causal

2.1 Probability and DAG

Definition Given a DAG $\mathcal{G}$ and a joint distribution P over a set $V = \{X_1, ..., X_n\}$ of discrete variables, we say that $\mathcal{G}$ *represents* P if there is a one-to-one correspondence between the variables in X and the nodes of $\mathcal{G}$, such that P admits the recursive product decomposition

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i | \ \mathbf{pa}_i)$$

where $\mathbf{pa}_i$ are the parents of X_i.The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

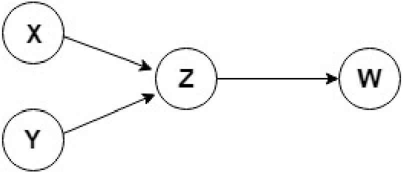


Figure 5

$$P(x, y, z, w) = P(x)P(y)P(z|x, y)P(w|z)$$

Theorem

A distribution $P \in M(\mathcal{G})$ if and only if the following Markov Condition holds: for every variable W,

$$W \perp \overline{W} | \mathbf{pa}_W$$

where $\overline{W}$ denotes all the other variables except the parents and descendants of W.

2.2 The d-separation criterion

The rules of d-separation:

1. In a non-collider X and Z are correlated and d-separated given Y;
2. If X and Z collide at Y they are d-separated but correlated given Y;
3. Conditioning on the descendant of a collider has the same effect as conditioning on the collider.

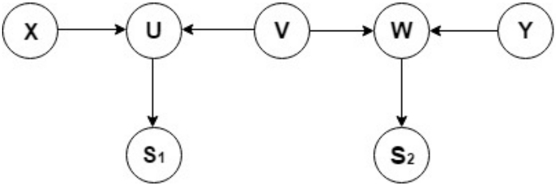


Figure 6

X and Y are d-separated (given the empty set)

X and Y are d-connected given $\{S_1, S_2\}$

X and Y are d-separated given $\{S_1, S_2, V\}$

The central role of the d-separation concept is shown by the following theorem.

Theorem (Spirtes, Glymour, Scheines)
Let A, B and C be disjoint sets of vertices. Then, $A \perp B|C$ if and only if A and B are d-separated by C.

Another crucial concept is that of Markov equivalence: Markov-equivalent DAGs have the same independence relationships.

The skeleton of a DAG $\mathcal{G}$ is the undirected graph obtained by replacing the arrows with undirected edges.

Theorem (Markov equivalence of DAGs)

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if
(i) skeleton($\mathcal{G}_1$) $=$ skeleton($\mathcal{G}_2$);
(ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same colliders.

Remark

Analitically, DAGs can be written as a set of equations of the form

$$x_i = f_i(\mathbf{pa}_i, u_i).$$

3. Causal Inference

Traditional statistical inference paradigm

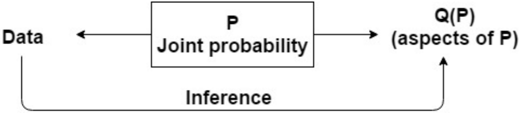


Figure 7

e.g. Infer whether customers who bought product A would also buy product B.
$Q = P(B|A)$

From Statistical to Causal Analysis

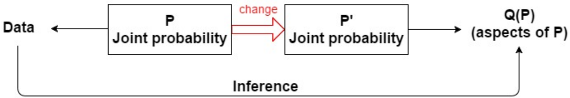


Figure 8

What happens when P changes?

e.g. Infer whether customers who bought product A would still buy A if we were to double the price.

$$P'(v) \neq P(v|2 \cdot price)$$

3.1 Intervention and do - calculus

Definition Let X be a set of variables in V. The action $do(x)$ sets X to constant x regardless of the factors which previously determined X. The action $do(x)$ replaces all functions f_i determining X with the constant functions $X = x$ to create a mutilated model M_x.

$do(X_i = x_i) \equiv$ removing the equation $x_i = f_i(\mathbf{pa}_i, u_i)$ from the model and substituting $X_i = x_i$ in the remaining equations.

The causal effect of X_i on X_j is denoted by $P(x_j|\hat{x}_i)$.

Truncated factorization

$$P(x_1, ..., x_n|\hat{x}_i') = \begin{cases} \prod_{j \neq i} P(x_j|\mathbf{pa}_j) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

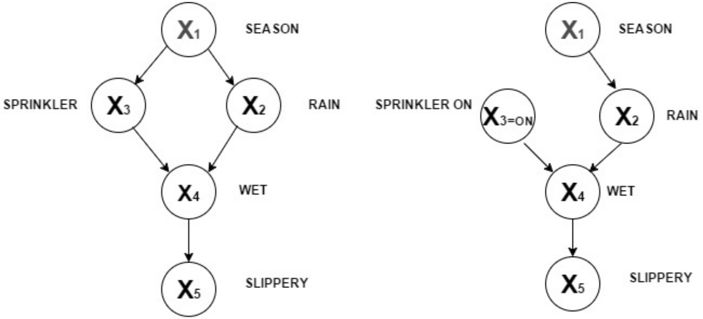

X_1
SEASON
SPRINKLER
X_3
X_2
RAIN
X_4
WET
X_5
SLIPPERY
X_1
SEASON
SPRINKLER ON
$X_{3=ON}$
X_2
RAIN
X_4
WET
X_5
SLIPPERY

Figure 8

Theorem (Adjustment for direct causes)

The effect of the intervention $do(X_i = x'_i)$ on Y is given by

$$P(y|\hat{x}'_i) = \sum_{\mathbf{pa}_i} P(y|\hat{x}'_i, \mathbf{pa}_i)P(\mathbf{pa}_i),$$

where $P(y|\hat{x}'_i, \mathbf{pa}_i)$ and $P(\mathbf{pa}_i)$ represent preintervention probabilities.

3.2 Identification of causal quantities

Causal quantities are defined relative to a causal model and not relative to a joint distribution over the set of the observed variables.

Problems:

- nonexperimental data provide information about the joint distribution alone;
- several models can generate the same distribution;

$$\Downarrow$$

$$\boxed{\text{The desired quantity may not be discernible unambiguously from the data}}$$

Definition (Identifiability) Let $Q(M)$ be any computable quantity of a model M. We say that Q is identifiable in a class M of models if, for any pairs of models M_1 and M_2 from M, $Q(M_1) = Q(M_2)$ whenever $P_{M_1}(v) = P_{M_2}(v)$.

What criterion should one use to decide which variables are appropriate for adjustment?

Identifiability of causal effects: an intuitive solution

Definition $(\mathbf{Back - door\ criterion})$ A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG $\mathcal{G}$ if
(i) no node in Z is a descendant of X_i;
(ii) Z blocks every path between X_i and X_j that contains an arrow into X_i.

Theorem (Back-door adjustment)
If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and it is given by the formula

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z).$$

The Rules of *do*-calculus

$\boldsymbol{Rule}\ \mathbf{1}\ (Insertion/deletion\ of\ observations):$

$P(y \mid \hat{x}, z, w) = P(y \mid \hat{x}, w) \quad if\ (Y \perp Z \mid X, W)_{G_{\overline{X}}}$

$\boldsymbol{Rule}\ \mathbf{2}\ (Action/observation\ exchange):$

$P(y \mid \hat{x}, \hat{z}, w) = P(y \mid \hat{x}, z, w) \quad if\ (Y \perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$

$\boldsymbol{Rule}\ \mathbf{3}\ (Insertion/deletion\ of\ actions):$

$P(y \mid \hat{x}, \hat{z}, w) = P(y \mid \hat{x}, w) \quad if\ (Y \perp Z \mid X, W)_{G_{\overline{X}\overline{Z(W)}}}$

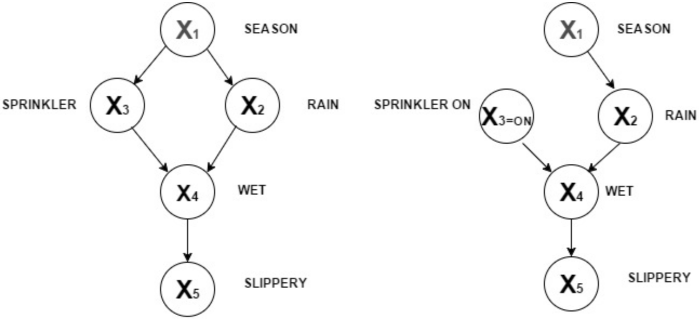

X_1
SEASON
SPRINKLER
X_3
X_2
RAIN
X_4
WET
X_5
SLIPPERY
X_1
SEASON
SPRINKLER ON
$X_{3=ON}$
X_2
RAIN
X_4
WET
X_5
SLIPPERY

Figure 8

4. Structural Equations Models (SEMs)

Definition A model made up of a set of equations

$$x_i = f_i(\mathbf{pa}_i, u_i)$$

is called functional model.

If the error terms are randomly distributed and mutually independent and each variable appears on the left of an equation, the model is called causal structural model.

Operational definitions

Definition An equation $y = \beta x + u$ is said to be structural if it is to be interpreted as follows: In an ideal experiment where we control X to x and any other set Z of variables (not containing X or Y) to z, the value y of Y is given by $\beta x + u$ where u is not a function of the settings x and z.

- All quantities are observable
- Different meaning of the equality symbol $\longrightarrow$ Graphs

Meaning of β

Rate of change (relative to x) of the expectation of Y in an experiment where X is held at x by external control.

$$\beta = \frac{\partial}{\partial x} E[Y|do(x)].$$

Remark

β is different from the regression coefficient, but they coincide under certain conditions (Single-Door criterion for direct effects).

The error term u

The previous equations provide the following operational definition:

$$u = y - E[Y|do(x)].$$

- u measures the deviation of Y from its controlled expectation $E[Y|do(x)]$ (and NOT from its conditional expectation $E[Y|x]$).
- the formula prescribes how errors are measured, not how they originate, but it's important to keep in mind the omitted factors conception.

Three kinds of queries

- predictions
- interventions
- counterfactuals

4.1 Predictions

Advantages of causal-functional specification in predictive tasks:

- Stability of conditional independencies in the causal graph;
- Intuitiveness of functional specification and use of a small number of parameters;
- Possibility of making local changes without changing the entire model.

4.2 Interventions and causal effects in functional models

All features of causal Bayesian networks can be emulated in Markovian functional models.

Procedure

1. (abduction): Update the probability $P(u)$ to obtain $P(u|o)$.
2. (action): Replace the equations corresponding to variables in set X by the equations $X = x$.
3. (prediction): Use the modified model to compute the probability of $Y = y$.

Pros

- The analysis of interventions can be extended to cyclic models;
- interventions involving the modification of equational parameters are more readily comprehended than those described as modifiers of conditional probabilities;
- there are infinitely many conditional probabilities $P(x_i|\mathbf{pa}_i)$ but only a finite number of functions $x_i = f_i(\mathbf{pa}_i, u_i)$ among discrete variables X_i and PA_i.

5. Counterfactuals

Concepts

If A were true, then C would have been true?

A is the counterfactual antecedent; C is the counterfactual consequent.

Typical counterfactual sentences:
If Oswald were not to have shot Kennedy, then Kennedy would still be alive.
If Germany were not punished so severely at the end of World War I, Hitler would not have come to power.
If Reagan did not lower taxes, our deficit would be lower today.

Notation

$V = \{X_1,\ X_2,\ ...,\ X_n\} \longrightarrow$ set of variables describing the world

$x_1, x_2, ..., x_n \longrightarrow$ observed values

x_i^* value in the counterfactual world

$\hat{x}_i^*$ events referenced explicitly in the counterfactual antecedent through an external action

A typical counterfactual query will have the form:

Does $\hat{x}^* \rightarrow c^* |\, x, o$ hold true?

Example

$$c = \begin{cases} 0 \equiv Captain\, gives\, the\, signal\, to\, release\, the\, traitor. \\ 1 \equiv Captain\, gives\, the\, signal\, to\, shoot\, the\, traitor. \end{cases}$$

$$b = \begin{cases} 0 \equiv Bob\, does\, not\, fire\, his\, rifle. \\ 1 \equiv Bob\, fires\, his\, rifle. \end{cases}$$

$$t = \begin{cases} 0 \equiv Traitor\, lives. \\ 1 \equiv Traitor\, dies. \end{cases}$$

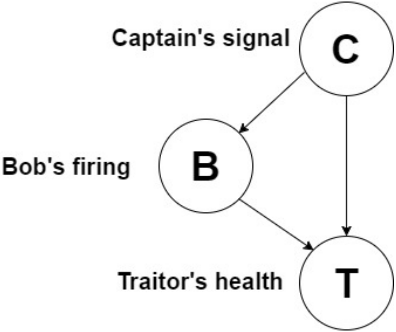


Figure 9

The structural equations relative to the example are

$$B = C$$

$$T = B \vee C$$

Suppose that we observe Bob fire his rifle $(b = 1)$ and the traitor expires $(t = 1)$. If Bob were not to have fired $(\hat{b}^* = 0)$, would the traitor have lived $(t^* = 0)$, i.e, does $\hat{b}^* = 0 \rightarrow t^* = 0|\, b = 1, t = 1$ hold true?

The structural equations relative to the modified model are

$$B = 0$$

$$T = B \vee C$$

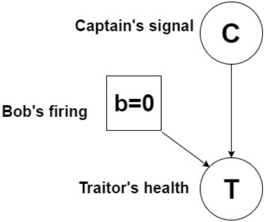


Figure 10

A more realistic model may be given by

$$B = (C \vee ab_{b1}) \wedge \neg ab_{b2}$$

$$T = (B \vee C) \wedge \neg ab_{t1} \vee ab_{t2}$$

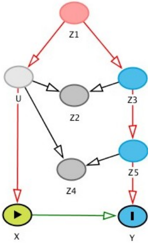
where

$ab_{b1} \equiv$ events that can cause Bob to fire even though the Captain did not give the order to fire;
$ab_{b2} \equiv$ events that can prevent Bob from firing his rifle;
$ab_{t1} \equiv$ events that can prevent the Traitor from expiring even though the riflemen fired;
$ab_{t2} \equiv$ events that can cause the Traitor to die even though Bob did not fire and the Captain did not give the order to fire.

6. A practical application

```
simpson.simulator <- function(N,s,ce)
Z1 <- rnorm(N,0,s)
Z3 <- rnorm(N,0,s) + Z1
Z5 <- rnorm(N,0,s) + Z3
U <- rnorm(N,0,s) + Z1
Z4 <- rnorm(N,0,s) + Z5 + U
Z2 <- rnorm(N,0,s) + Z3 + U
X <- rnorm(N,0,s) + U
Y <- rnorm(N,0,s) + ce*X + 10*Z5
data.frame(Y,X,Z1,Z2,Z3,Z4,Z5)
#1st parameter: sample size
#2nd parameter: noise standard deviation
#3rd parameter: true causal effect
D <- simpson.simulator(1000,0.01,1)
#unadjusted estimate
m0 <- lm(D[,1:2])
summary(m0)

confint(m0,'X')
```



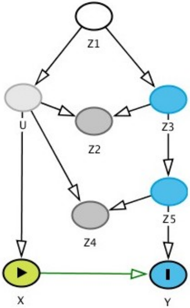
```
                    Coefficients:
             Estimate  Std. Error  t-value Pr(>|t|)
(Intercept)   -0.00685   0.00516 -1.328 0.185
X             4.4309 0.2972 14.907 <2e-16 ***
                 Confidence Interval
                    2.5 % 97.5%
               X 3.847582 5.014162
```

```
#adjusted for Z1
m1 <- lm(D[,c(1,2,3)])
summary(m1)

confint(m1,'X')
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005357 0.004496 -1.191 0.233751
X 1.086259 0.320112 3.393 0.000718 ***
Z1 10.243901 0.575792 17.791 < 2e-16 ***
Confidence Interval
2.5 % 97.5 %

X 0.4580887 1.71443
```



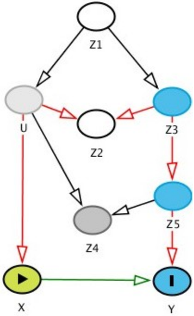
adjusted for Z1, Z2

```
m2 <- lm(D[,c(1,2,3,4)])
summary(m2)

confint(m2,'X')
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.003372 0.003993 -0.844 0.399
X -0.815766 0.306963 -2.658 0.008 **
Z1 3.664976 0.649852 5.640 2.22e-08 ***
Z2 4.148437 0.252998 16.397 < 2e-16 ***
Confidence Interval
2.5 %   97.5 %

X -1.418134 -0.2133971
```



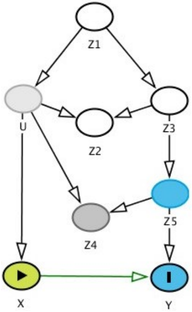
adjusted for Z1, Z2, Z3

```
m3 <- lm(D[,c(1,2,3,4,5)])
summary(m3)

confint(m3,'X')
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001023 0.003123 -0.328 0.743
X 0.983357 0.250373 3.928 9.17e-05 ***
Z1 -0.292680 0.531765 -0.550 0.582
Z2 0.144881 0.253718 0.571 0.568
Z3 10.283137 0.408171 25.193 < 2e-16 ***
Confidence Interval
2.5 % 97.5 %

X 0.4920364 1.474677
```



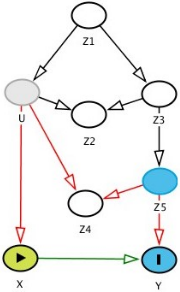
adjusted for Z1, Z2, Z3, Z4

```
m4 <- lm(D[,c(1,2,3,4,5,6)])
summary(m4)

confint(m4,'X')
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.003292 0.002388 -1.379 0.1683
X -0.502946 0.199277 -2.524 0.0118 *
Z1 -1.681000 0.409653 -4.103 4.40e-05 ***
Z2 -0.985001 0.198450 -4.963 8.14e-07 ***
Z3 7.091261 0.334095 21.225 < 2e-16 ***
Z4 4.143695 0.155492 26.649 < 2e-16 ***
Confidence Interval
2.5 % 97.5 %

X -0.8939973 -0.1118943
```



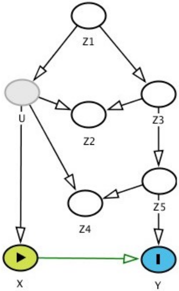
adjusted for Z1, Z2, Z3, Z4, Z5

```
m5 <- lm(D[,c(1,2,3,4,5,6,7)])
summary(m5)

confint(m5,'X')
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0001118 0.0003313 -0.337 0.736
X 0.9782021 0.0283885 34.458 <2e-16 ***
Z1 0.0680691 0.0573002 1.188 0.235
Z2 -0.0132510 0.0278384 -0.476 0.634
Z3 0.0508255 0.0558591 0.910 0.363
Z4 -0.0080416 0.0283535 -0.284 0.777
Z5 10.0063443 0.0444126 225.304 <2e-16 ***
Confidence Interval
2.5 % 97.5 %

X 0.9224937 1.03391
```



Conclusions

Pros...

- Isomorphic reconfiguration of the network topology in response to changing conditions;
- DAGs define sequential procedures of knowledge acquisition and reduce the number of required assessments;
- They allow to build plans and strategies under uncertainty.

...and Cons

- Difficulties in inferring causal structures just from observational data (unmeasured/unknown confounders);
- Insufficient attention to the strong causal assumptions that are part of the SEMs.

Further developments

- Method for unifying causal DAGs and Counterfactuals (SWIG);
- Transportability.

References

[1] Balke, A. and Pearl, J. (1994) Counterfactual probabilities: computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence* 10. R. Lopez deMantaras and D. Poole, Eds., pp. 46-54. Morgan Kaufmann, San Mateo, CA.

[2] Balke, A. and Pearl, J. (1995) Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence* 11. P. Besnard and S. Hanks, Eds., pp. 11-18. Morgan Kaufmann, San Francisco, CA.

[3] Bollen, K. A. Pearl, J. Eight Myths About Causality and Structural Equation Models, in S.L. Morgan(Ed.), *Handbook of Causal Analysis for Social Research*, Chapter 15, 301-328, Springer 2013.

[4] Pearl, J. The Causal Foundations of Structural Equation Modeling in R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling*, Chapter 5, pp. 68-91, New York: Guilford Press 2012.

[5] Pearl, J. Graphical Models for Probabilistic and Causal Reasoning, in Allen B. Tucker (Ed.), *Computer Science Handbook, Second Edition*, Chapter 70, pp. 70-1 - 70-18, CRC Press, 2004.

[6] Pearl, J. (2009) *Causality: Models, Reasoning and Inference* 2nd Ed., Cambridge University Press New York, NY, USA.

[7] Pearl, J. (2013) Structural Counterfactuals: A Brief Introduction, in *Cognitive Science*, pp. 977-985.

[8] Robins, J.M., Wasserman, L. (1999) On the Impossibility of Inferring Causation from Association without Background Knowledge, in Glymour, P. Cooper, G. (Eds.) *Computation, Causation and Discovery*, pp. 305-321, MIT Press.

[9] Richardson, T.S., Robins, J.M. (2013) Single World Intervention Graphs: A Primer.

[10] Wasserman, L. (2004) *All of Statistics: A Concise Course in Statistical Inference*. Springer-Verlag, New York.

Questions?