A Decision Theoretic Approach to Causality

Vanessa Didelez
School of Mathematics
University of Bristol

(based on joint work with Philip Dawid)

Bordeaux, June 2011

Based on:

Dawid & Didelez (2010). Identifying the consequences of dynamic treatment strategies: A decision theoretic overview. *Statistics Surveys*, **4**, 184-231.

Didelez & Dawid (2008). Identifying optimal sequential decisions. *24th UAI*, 113-120.

Didelez, Geneletti & Dawid (2006). Direct and indirect effects of sequential treatments *22nd UAI*, 138-146.

(more references with abstract; and references therein)

all on `http://www.maths.bris.ac.uk/~maxvd/`

Overview

Part 1: Decision theoretic framework for causal inference

Part 2: Graphical models and influence diagrams

Part 3: Sequential decisions and direct effects

Part 1:

Decision theoretic framework for causal inference

- Examples
- Target of inference and assumptions
- Formal frameworks: potential responses vs. decision theory

Examples

Many studies are carried out to inform e.g. public health interventions, doctors' decisions, give advice etc.

– add folic acid to flour to prevent nural tube defects?

– banning smoking in pubs to lower lung cancer risk?

– advice on breast feeding: how long is best and under what conditons?

– HIV patients: start HAART when CD4-counts below what threshold?

– hormone replacement therapy (HRT) beneficial or not?

Want these to be well informed so that decisions not wrong / interventions not useless $\Rightarrow$ distinguish **association** and **causation!**

Target of Inference & Assumptions

Some issues when using data to inform decisions:

Target of inference: what possible decisions do we want to compare for what population?

Assumptions: under what assumptions linking the data to the decision problem at hand do our *methods* give valid conclusions (identifiability)? and can we **justify** these assumptions?

$\Rightarrow$ need to be clear and explicit about both, target and assumptions.

Formal Frameworks

... should allow to distinguish *association* and *causation*.

Association: *observing* X predicts / is informative for Y.

Events occur more often together than expected under independence.

Usual conditional probability notation: $p(Y = y|X = x)$, i.e. probability of $Y = y$ given we happen to *know* that $X = x$ has occurred.

Causation: *intervening* in X predicts / is informative for Y.

We can 'make' event $Y = y$ more likely by manipulating X.

Need some special **notation** for this; conditional probability not enough!

Potential Responses (PRs)

(Rubin, 1974; many others)

Consider binary treatment $X^i \in \{0, 1\}$, individual i

$Y_0^i =$ response if $X^i = 0$

$Y_1^i =$ response if $X^i = 1$ for same subject (at the same time)

$\Rightarrow \{Y_0^i, Y_1^i\}$ can never be observed together $\Rightarrow$ potential responses.

Once a decision has been made, say $X^i = 1$, then Y_1^i can be observed and Y_0^i is *counterfactual*.

Note: PRs only well defined if manipulation of X well defined.

Target(s) of inference

Individual effects: any contrast of Y_1^i and Y_0^i, e.g. individual causal effect (ICE)

$$ICE^i = Y_1^i - Y_0^i \qquad \text{(or ratio or...)}.$$

Note: inference about ICE depends on assumptions about joint distribution of (Y_0, Y_1).

Alternatively: any contrast of the **distributions** $p(Y_1)$ and $p(Y_0)$, e.g. **average** causal effect (ACE)

$$ACE = E(Y_1) - E(Y_0).$$

Note: does not depend on joint distribution of (Y_0, Y_1).

Assumptions

An example for a standard assumption that allows to identify $p(Y_x)$ is the one of **random treatment assignment** or **no unmeasured confounding**: let C be observed pre–treatment covariates and

$$X \perp\!\!\!\perp Y_x \mid C \qquad\qquad x = 1, 0.$$

E.g. in RCT X is randomised and hence independent of 'pre-treatment'.

Then $p(Y|X = x, C = c) = p(Y_x|X = x, C = c) = p(Y_x|C = c)$

and hence e.g.

$$ACE = \sum_c \{E(Y|X = 1, C = c) - E(Y|X = 0, C = c)\} p(C = c)$$

Using also consistency: $Y = Y_X$

Decision Theoretic Approach

(Pearl, 1993; Dawid, 2002)

Indicator: index distributions by **regime** indicator σ_X

$$\sigma_X = \begin{cases} x \in \mathcal{X}, & \text{set } X \text{ to } x \text{ by specified intervention} \\ \emptyset, & \text{let } X \text{ arise 'naturally'}, \end{cases}$$

where $p(X; \sigma_X = x) = I(X = x)$
and $p(X; \sigma_X = \emptyset)$ 'observational' distribution of X.

- If $p(Y; \sigma_X = x)$ depends on x, we consider X causal for Y.
- If $p(Y|X = x; \sigma_X = \emptyset)$ depends on x, then there is an association which could also be due to confounding, reverse causation etc.

Target(s) of inference

In general: contrast intervention distribution $p(Y; \sigma_X = x)$ for different values of x, e.g. **average** causal effect (ACE)

$$ACE = E(Y; \sigma_X = 1) - E(Y; \sigma_X = 0).$$

Cannot formulate **individual** causal parameters!
Could instead consider **conditional effects** in sub-population s, e.g.

$$ACE_s = E(Y|S = s; \sigma_X = 1) - E(Y|S = s; \sigma_X = 0).$$

More **generally**, let $k(\cdot)$ be **loss** function
$\Rightarrow$ want to evaluate $E(k(Y); \sigma_X = x)$ for different x.

Assumptions

Analogous to 'no unmeasured confounders' is assumption of **sufficient covariates** C: (Dawid, 2002)

$$C \perp\!\!\!\perp \sigma_X \quad \text{and} \quad Y \perp\!\!\!\perp \sigma_X | (X, C).$$

Then can identify intervention distribution

$$\begin{aligned} p(Y; \sigma_X = x) &= \sum_{X,C} p(Y|X, C; \sigma_X = x) I(X = x) p(C; \sigma_X = x) \\ &= \sum_{C} p(Y|X, C; \sigma_X = \emptyset) p(C; \sigma_X = \emptyset). \end{aligned}$$

Outlook (Part 1)

More complex targets:

– effect of treatment on the treated
– complier causal effect (only with potential responses)
– conditional strategies: decision is made depending on additional observations $\Rightarrow$ optimal decision
– direct and indirect effects.

Assumptions: most methods of causal inference make some version of 'no unmeasured confounders' assumption.
Exception: instrumental variables (but other assumptions needed).

Methods: instead of adjusting for covariates, can use *propensity scores* or *inverse probability weighting*.
All make same assumptions! Different wrt. robustness and efficiency.

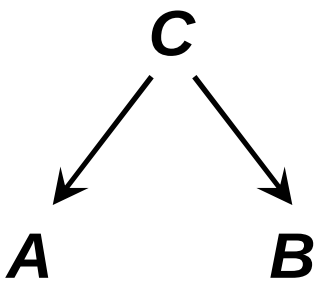
Part 2:

Graphical models and influence diagrams

- Graphical models
- Influence diagrams
- Representing assumptions
- An example

Graphical Models

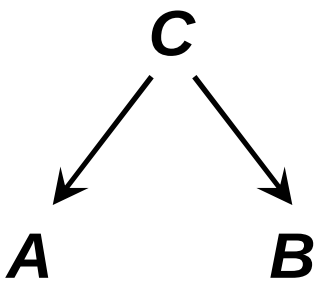
Here: directed acyclic graphs (DAGs), where
vertices = variables and
no edge = some (conditional) independence.



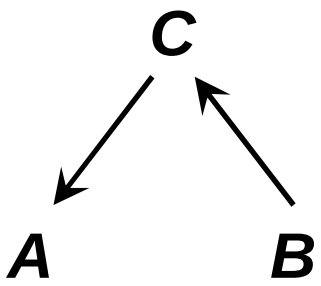
$$A \perp\!\!\!\perp B | C$$

Graphical Models

Here: directed acyclic graphs (DAGs), where
vertices = variables and
no edge = some (conditional) independence.



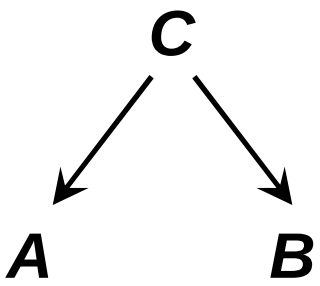
$A \perp\!\!\!\perp B | C$



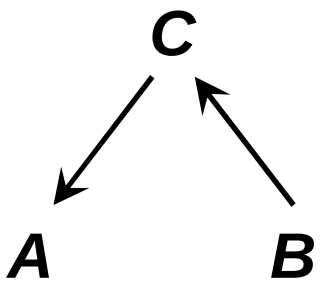
also $A \perp\!\!\!\perp B | C$

Graphical Models

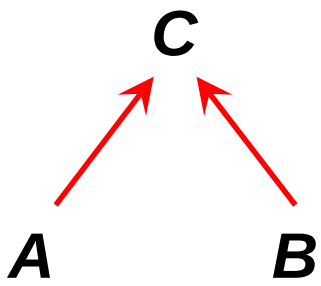
Here: directed acyclic graphs (DAGs), where
vertices = variables and
no edge = some (conditional) independence.



$A \perp\!\!\!\perp B | C$



also $A \perp\!\!\!\perp B | C$

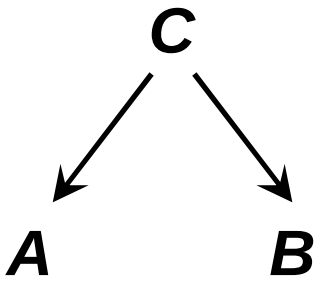


but $A \not\perp\!\!\!\perp B | C$, $A \perp\!\!\!\perp B$

Graphical Models

Factorisation of joint density

$$p(\mathbf{x}) = \prod_i p(x_i|\mathbf{x}_{\mathsf{pa}(i)})$$

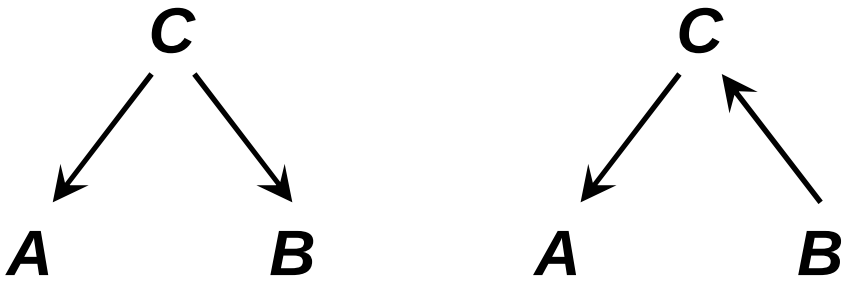


$p(a|c)p(b|c)p(c)$

Graphical Models

Factorisation of joint density

$$p(\mathbf{x}) = \prod_i p(x_i|\mathbf{x}_{\mathsf{pa}(i)})$$

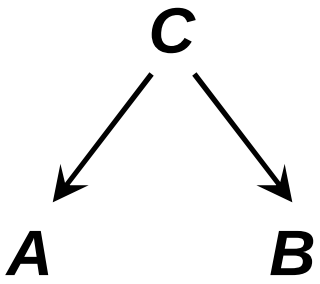


$p(a|c)p(b|c)p(c)$ $\quad$ $p(a|c)p(b)p(c|b)$

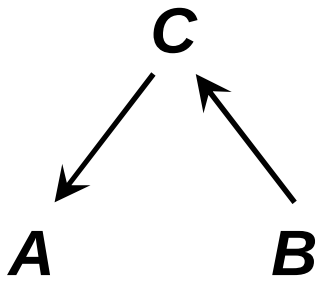
Graphical Models

Factorisation of joint density

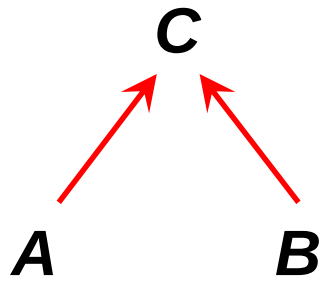
$$p(\mathbf{x}) = \prod_i p(x_i|\mathbf{x}_{\mathsf{pa}(i)})$$



$p(a|c)p(b|c)p(c)$



$p(a|c)p(b)p(c|b)$

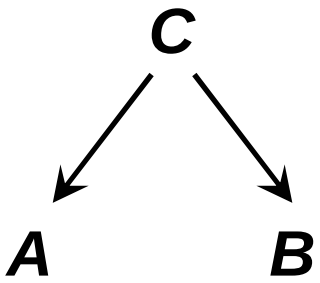


$p(a)p(b)p(c|a,b)$

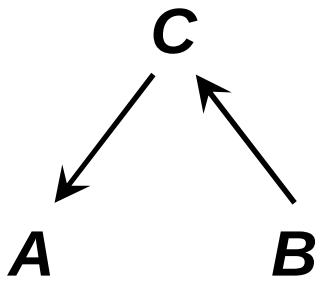
Graphical Models

In general:

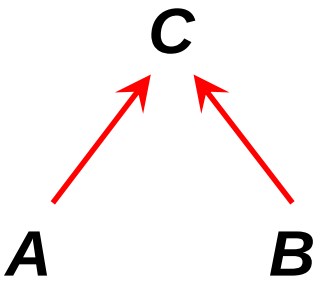
$$X_i \perp\!\!\!\perp \mathbf{X}_{\mathsf{nd}(i)} \mid \mathbf{X}_{\mathsf{pa}(i)}$$



$A \perp\!\!\!\perp B | C$



also $A \perp\!\!\!\perp B | C$



$A \perp\!\!\!\perp B$

Influence Diagrams

(Dawid 2002, 2003)

Include decision node / intervention indicator σ_X

Example:
while the following just means $p(x, y) = p(x)p(y|x)$ (no restriction)

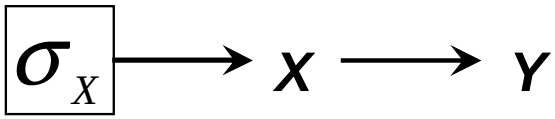
Influence Diagrams

(Dawid 2002, 2003)

Include decision node / intervention indicator σ_X

Example:
while the following just means $p(x, y) = p(x)p(y|x)$ (no restriction)

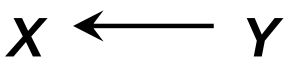
The following implies: $Y \perp\!\!\!\perp \sigma_X | X$ or $p(x, y; \sigma_X) = p(y|x)p(x; \sigma_X)$



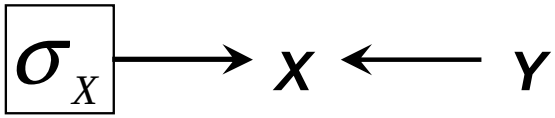
Influence Diagrams

(Dawid 2002, 2003)

Example ctd.:
while the following just means $p(x, y) = p(y)p(x|y)$ (no restriction)



The following implies: $Y \perp\!\!\!\perp \sigma_X$ or $p(x, y; \sigma_X) = p(y)p(x|y; \sigma_X)$

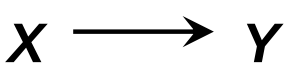


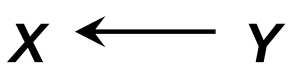
Influence Diagrams

(Dawid 2002, 2003)

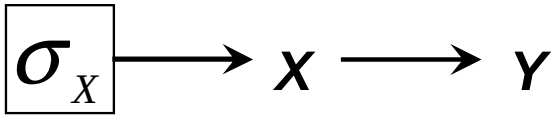
Example ctd.:

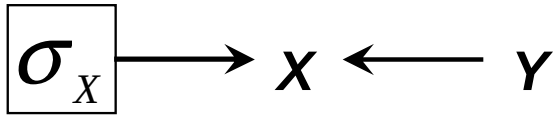
While there is no difference between the models





... we can express different assumptions about interventions e.g. by





Note: indicator σ_X not random $\Rightarrow$ in 'box' and always conditioned on.

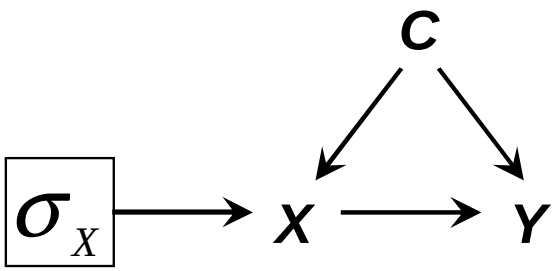
Representing Assumptions

Use influence diagrams to represent assumptions about interventions.

E.g. assumption of **sufficient covariates**

$$C \perp\!\!\!\perp \sigma_X \quad \text{and} \quad Y \perp\!\!\!\perp \sigma_X | (X, C).$$

uniquely represented by influence diagram



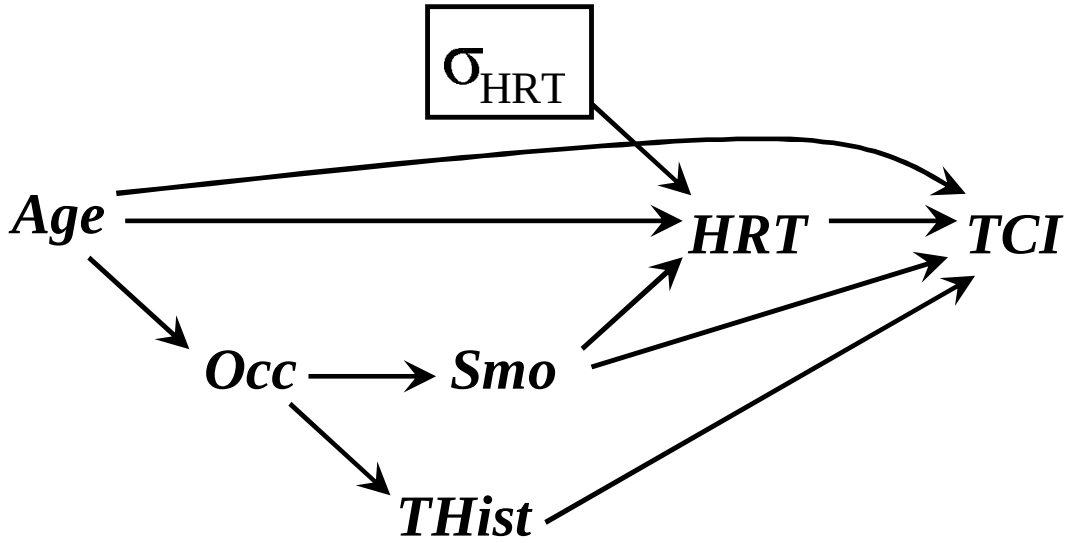
Where Do the Graphs Come From?

Combination of

- subject matter background knowledge (biology, physics...), and e.g. study design, type of intervention considered;
- testable implications: certain conditional independencies can be tested from data.

Example

Study on effect of hormone replacement therapy (HRT) on temporal cerebral ischemia (TCI) (Pedersen et al., 1997; Didelez et al., 2008)



Here: 'Age' & 'Smoking' sufficient covariates for effect of HRT on TCI.

Outlook (Part 2)

Graphical models / influence diagrams

– are used to facilitate *reasoning*
– remember: **justify absence** of further nodes &edges
– help to formulate background knowledge
– provide graphical rules to check for identifiability
– suggest what variables to adjust for (if at all), and how.

Other applications, for example:

– effect of treatment on the treated
– instrumental variables (Mendelian randomisation)
– data situtations with potential for selection bias
– dynamic models, time series / event histories
– and of course Part 3...

Part 3:

Sequential decisions and direct effects

- (Optimal) sequential decisions
- G–computation
- Simple & extended stability
- Direct causal effects

Sequential Decisions

Examples:

(1) Stroke patients receive regular anticoagulant treatment:
— has to be continually monitored and dosage **adjusted**
— depending on blood test results and other health indicators

(2) HIV patients:
— have to decide at what point in time to start HAART
— depending on latest CD4 count

Target:
want to find **optimal** treatment **strategy** from (observational?) data.

Sequential Decisions

Question:
what assumptions do we need to be able to make inferences about

- decisions that may depend on the patient's history — *conditional* interventions / strategies?

- and are these more restrictive when we want to find an *optimal* strategy?

When assumptions are plausible, how do we evaluate effect of a (optimal) strategy?

Some Notation

$A_1, \ldots, A_N$ "action" variables $\rightarrow$ can be 'manipulated'

$L_1, \ldots, L_N$ covariates $\rightarrow$ (available) background information

$Y = L_{N+1}$ response variable

all measured over time, L_i before A_i

$\mathbf{A}^{<i} = (A_1, \ldots, A_{i-1})$ past up to before i; $\mathbf{A}^{\leq i}$, $\mathbf{A}^{>i}$ etc. similarly

Strategies

Strategy $\mathbf{s} = (s_1, \ldots, s_N)$ set of functions assigning an action

$$a_i = s_i(\mathbf{a}^{<i}, \mathbf{l}^{\leq i}) \text{ to each history } (\mathbf{a}^{<i}, \mathbf{l}^{\leq i})$$

(Could be stochastic, then dependence on $\mathbf{a}^{<i}$ relevant.)

Also called: **conditional** / **dynamic** / **adaptive** strategies.

Indicator

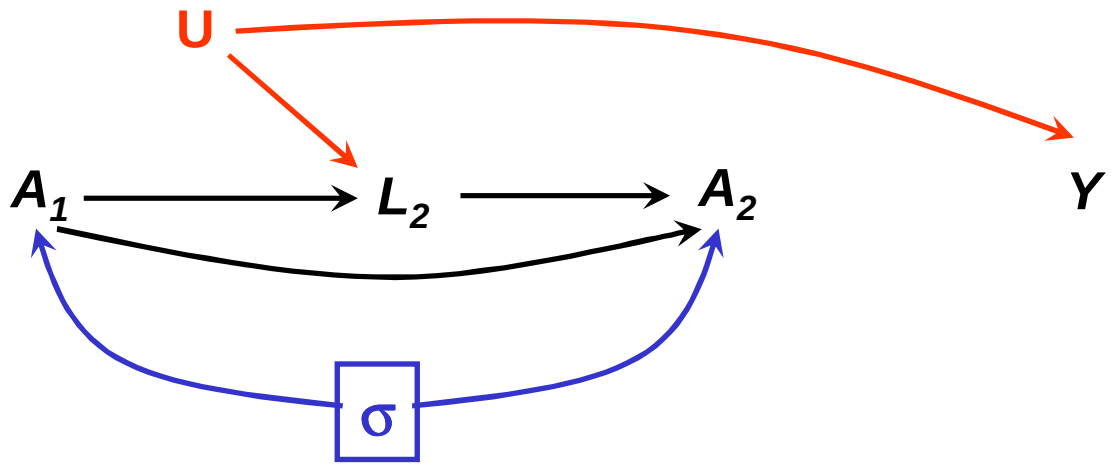
$$\sigma = \begin{cases} \emptyset, & \text{observational regime} \\ s, & \mathbf{s} \in \mathcal{S} = \text{ set of strategies} \end{cases}$$

Denote $p(\cdot; \mathbf{s}) = p(\cdot; \sigma = \mathbf{s})$ distributions under strategy $\mathbf{s}$.

NB: Why Naïve Regression Does Not Work

Naïvely: fit regression model $p(y|\mathbf{a}, \mathbf{l})$, e.g. Cox regression?

Example: Possible scenario under null hypothesis of no causal effect:



But: $Y \not\perp\!\!\!\perp A_1 | (A_2, L_2)$ — conditioning on *collider* L_2 will induce association between Y and A_1! Regression **not consistent** at the null.

However, need to condition on L_2 as possible confounder for A_2.

Evaluation of Strategies

Let $k(\cdot)$ be a loss function. Want to evaluate $E(k(Y); \mathbf{s})$.

Define

$$f(\mathbf{a}^{\leq j}, \mathbf{l}^{\leq i}) := E\{k(Y)|\mathbf{a}^{\leq j}, \mathbf{l}^{\leq i}; \mathbf{s}\} \quad i = 1, \ldots, N; j = i-1, i.$$

Then obtain $f(\emptyset) = E(k(Y); \mathbf{s})$ from $f(\mathbf{a}^{\leq N}, \mathbf{l}^{\leq N})$ iteratively by:

$$f(\mathbf{a}^{<i}, \mathbf{l}^{\leq i}) = \sum_{a_i} p(a_i|\mathbf{a}^{<i}, \mathbf{l}^{\leq i}; \mathbf{s}) \times f(\mathbf{a}^{\leq i}, \mathbf{l}^{\leq i})$$

$$f(\mathbf{a}^{<i}, \mathbf{l}^{<i}) = \sum_{l_i} p(l_i|\mathbf{a}^{<i}, \mathbf{l}^{<i}; \mathbf{s}) \times f(\mathbf{a}^{<i}, \mathbf{l}^{\leq i}).$$

Note: well–known as extensive form analysis.
Also known as **G–computation** (Robins, 1986)

Evaluation of Strategies

Let $k(\cdot)$ be a loss function. Want to evaluate $E(k(Y); \mathbf{s})$.

Define

$$f(\mathbf{a}^{\leq j}, \mathbf{l}^{\leq i}) := E\{k(Y)|\mathbf{a}^{\leq j}, \mathbf{l}^{\leq i}; \mathbf{s}\} \quad i = 1, \ldots, N; j = i-1, i.$$

Then obtain $f(\emptyset) = E(k(Y); \mathbf{s})$ from $f(\mathbf{a}^{\leq N}, \mathbf{l}^{\leq N})$ iteratively by:

$$f(\mathbf{a}^{<i}, \mathbf{l}^{\leq i}) = \sum_{a_i} \underbrace{p(a_i|\mathbf{a}^{<i}, \mathbf{l}^{\leq i}; \mathbf{s})}_{\text{known by } \mathbf{s}} \times f(\mathbf{a}^{\leq i}, \mathbf{l}^{\leq i})$$

$$f(\mathbf{a}^{<i}, \mathbf{l}^{<i}) = \sum_{l_i} p(l_i|\mathbf{a}^{<i}, \mathbf{l}^{<i}; \mathbf{s}) \times f(\mathbf{a}^{<i}, \mathbf{l}^{\leq i}).$$

Evaluation of Strategies

Let $k(\cdot)$ be a loss function. Want to evaluate $E(k(Y); \mathbf{s})$.

Define

$$f(\mathbf{a}^{\leq j}, \mathbf{l}^{\leq i}) := E\{k(Y)|\mathbf{a}^{\leq j}, \mathbf{l}^{\leq i}; \mathbf{s}\} \quad i = 1, \ldots, N; j = i-1, i.$$

Then obtain $f(\emptyset) = E(k(Y); \mathbf{s})$ from $f(\mathbf{a}^{\leq N}, \mathbf{l}^{\leq N})$ iteratively by:

$$f(\mathbf{a}^{<i}, \mathbf{l}^{\leq i}) = \sum_{a_i} \underbrace{p(a_i|\mathbf{a}^{<i}, \mathbf{l}^{\leq i}; \mathbf{s})}_{\text{known by } \mathbf{s}} \times f(\mathbf{a}^{\leq i}, \mathbf{l}^{\leq i})$$

$$f(\mathbf{a}^{<i}, \mathbf{l}^{<i}) = \sum_{l_i} \underbrace{p(l_i|\mathbf{a}^{<i}, \mathbf{l}^{<i}; \mathbf{s})}_{\text{not (?) known}} \times f(\mathbf{a}^{<i}, \mathbf{l}^{\leq i}).$$

Simple Stability

(Dawid & Didelez, 2010)

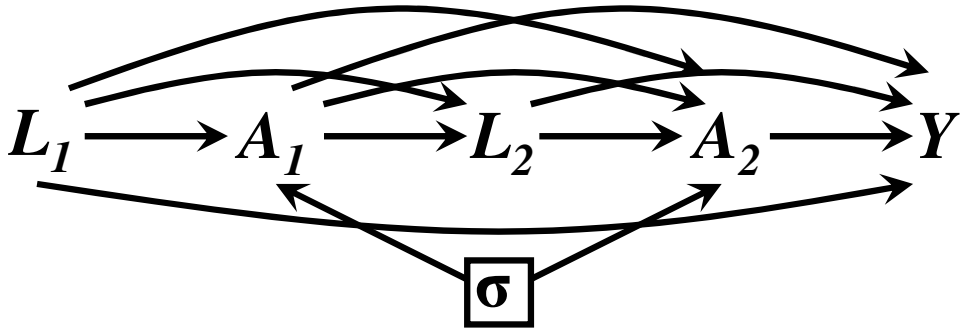
Sufficient for identifiability of any type of strategy is

$$p(l_i|\mathbf{a}^{<i},\mathbf{l}^{<i};\mathbf{s}) = p(l_i|\mathbf{a}^{<i},\mathbf{l}^{<i};\emptyset) \quad \text{for all } i = 1,\ldots,N+1$$

or (via intervention indicator)

$$L_i \perp\!\!\!\perp \sigma | (\mathbf{A}^{<i},\mathbf{L}^{<i}) \quad \text{for all } i = 1,\ldots,N+1$$

Or graphically:

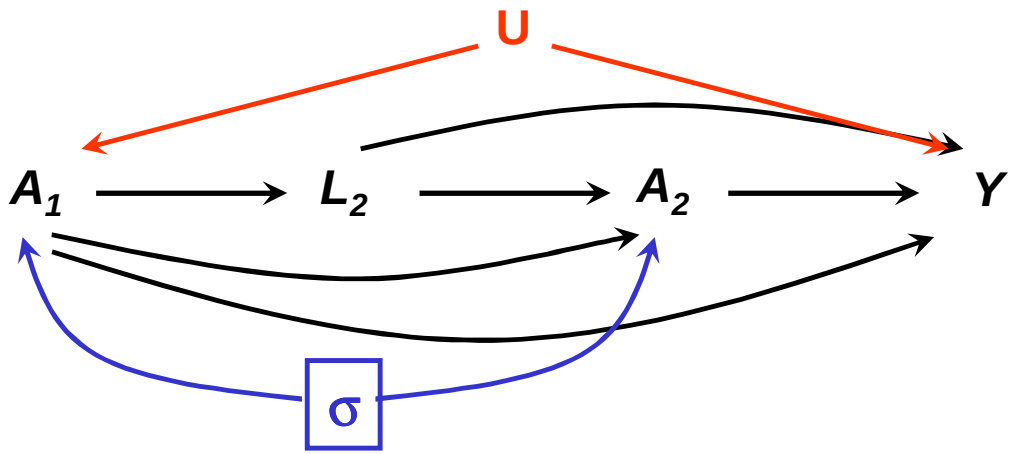


Extended Stability

Might not be able to assess simple stability without taking unobserved variables into account.

$\Rightarrow$ extend covariates $\mathbf{L}$ to include unobserved / hidden variables $\mathbf{U} = (U_1, \ldots, U_N)$ and check if simple stability can be deduced.

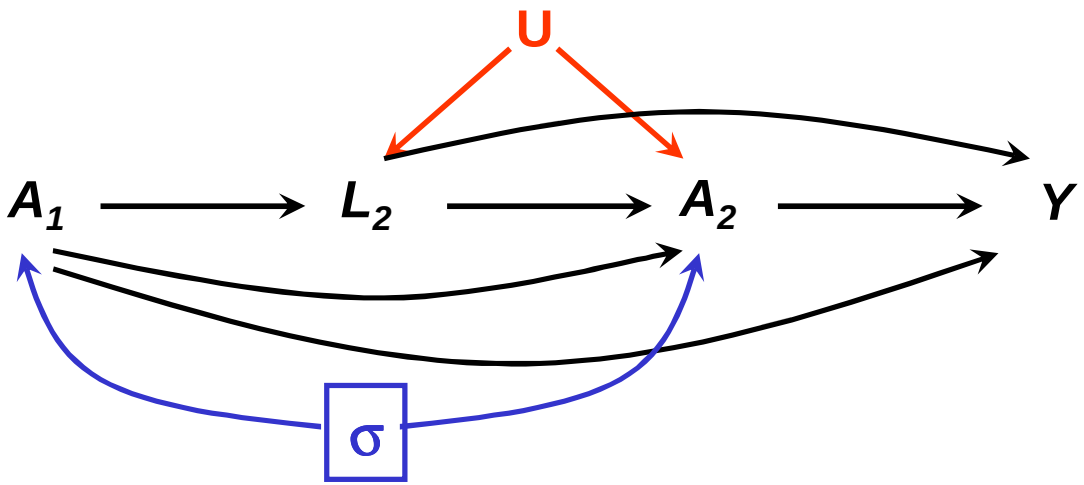
Example 1: particular underlying structure (note: $L_1 = \emptyset$)



Simple stability violated as $Y \not\perp\!\!\!\perp \sigma \mid (A_1, A_2, L_2)$.

Examples

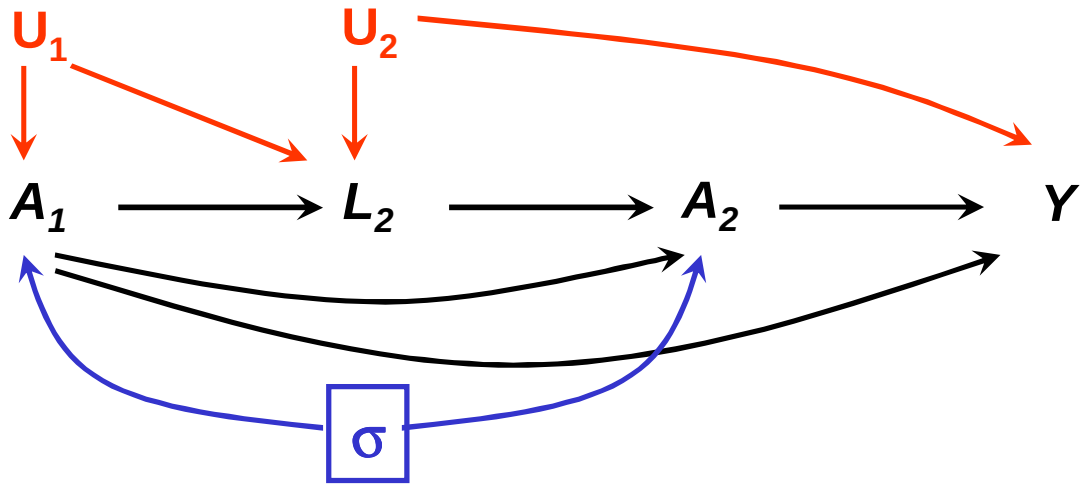
Example 2: different underlying structure



Simple stability satisfied.

Examples

Example 3: another different underlying structure



Simple stability violated: $L_2 \not\perp\!\!\!\perp \sigma \mid A_1$ and $Y \not\perp\!\!\!\perp \sigma \mid (A_1, A_2, L_2)$

Relax Simple Stability?

Specific conditional strategies can be identified under **weaker** conditions than simple stability. (Pearl & Robins, 1995; Dawid & Didelez, 2010)

However, in order to identify an **optimal** strategy, cannot apply these weaker conditions. (Dawid & Didelez, 2008)

Heuristically: optimal strategy has to be allowed to depend on **all** previous observed information.

Note: *necessary* and sufficient conditions have been given (in a more restrictive causal framework) for conditional strategies, but not generalised to optimal ones. (Tian, 2008)

Direct Causal Effects

(Greenland & Robins, 1999; Pearl, 2001)

Examples: Want direct effect of X on Y not mediated by Z

- direct effect of treatment not mediated by mental attitude (no placebo effect);
- direct effect of oral contraception on thrombosis risk not mediated by prevention of pregnancy;
- direct effect of gender on salary not mediated by qualification.

$\Rightarrow$ need to 'block' effect through Z by fixing it $\Rightarrow$ compare interventions in X while intervening in Z to keep it 'constant'.

Controlled Direct Causal Effects

Direct effects can be formalised in different ways.
Most popular: fix Z at z.

Controlled Direct effect of X (binary, say) at $\sigma_Z = z$, e.g.

$$CDE_z = E(Y; \sigma_X = 1, \sigma_Z = z) - E(Y; \sigma_X = 0, \sigma_Z = z).$$

Interpretation:
'force' $Z = z$ for everyone and compare different settings of X.

Note: as before, regression $p(y|x, z)$ not suitable!

General Direct Causal Effects

Standardised Direct Effect: (Geneletti, 2006)

More generally, let Z arise from the same distribution $\mathcal{D}$ under different interventions in X

$$SDE_{\mathcal{D}} = E(Y; \sigma_X = 1, \sigma_Z = \mathcal{D}) - E(Y; \sigma_X = 0, \sigma_Z = \mathcal{D})$$

"Natural" Direct Effect — with POs: $E(Y_{1,Z_0} - Y_0)$

If $\mathcal{D}_0^W = p(z|W; \sigma_X = 0, \sigma_Z = \emptyset)$ then

$$NDE = E(Y; \sigma_X = 1, \sigma_Z = \mathcal{D}_0^W) - E(Y; \sigma_X = 0, \sigma_Z = \mathcal{D}_0^W)$$

Note: only for NDE (and specific W): total=direct+indirect effect!

Sequential Decisions and Direct Effects

(Didelez et al., 2006)

Direct effects essentially the same as sequential decision problem:

Choose $A_1 = X$ and $A_2 = Z \Rightarrow$ same conditions for identifiability and same methods of evaluation.

Can apply conditions that are weaker than simple stability as intervention to fix Z does not depend on previous observations.

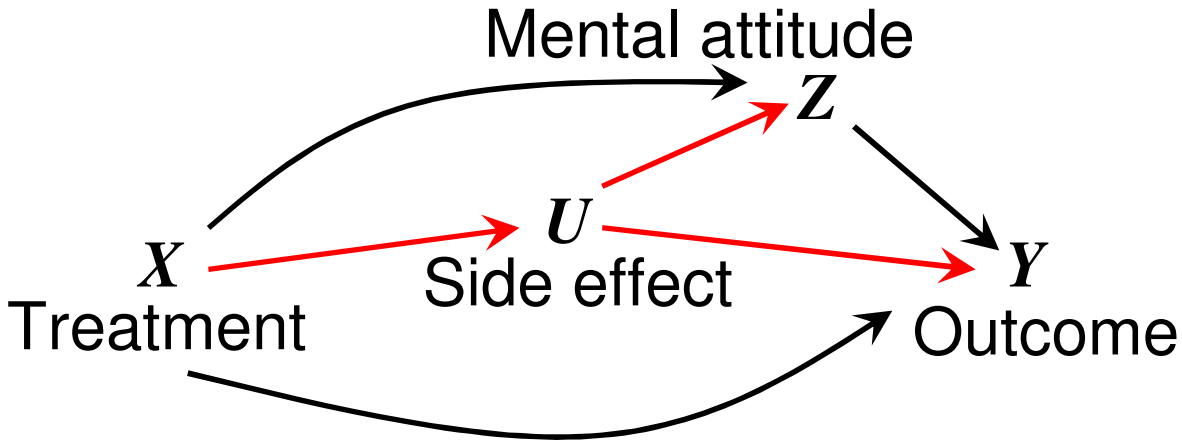
Note: for NDE need to obtain $\mathcal{D}_0^W \Rightarrow$ additional conditions required.

Also, NDE not well defined in some situations where CDE or SDE are.

Example: Double Blind Study

NDE not well defined / not identified:
Side effect will depend on actual treatment and hence mental attitude will not be generated from same distribution for treated and untreated!

Cannot speak of direct effect unless 'side effect' can also be controlled.



Outlook (Part 3)

In principle, can use G–formula, estimating all $p(l_i|\mathbf{a}^{<i}, \mathbf{l}^{<i}; \emptyset)$ from observational data $\Rightarrow$ sensitive to misspecification / 'null–paradox'. Also: curse of high dimensions & dynamic programming becomes infeasible.

1st Alternative: inverse probability of treatment weighting (IPTW). Requires models / estimation of $p(a_i|\mathbf{a}^{<i}, \mathbf{l}^{\leq i}; \emptyset)$.
Recently: advances in using IPTW for finding 'optimal' dynamic treatments. (Orellana, Rotnitzky, Robins, 2011)

2nd Alternative: G–estimation $\Rightarrow$ so far only motivated within PR framework and specific survival models. (Robins, 1992)

Models / methods for finding optimal treatment strategy still need more testing in practice. (Rosthoj et al., 2006)

Summary

Decision theoretic framework brings clarity to

– causal questions

– underlying assumptions

and will alert users to "cross–world" assumptions