Section 7

Lecture 3

Plan

- Factorizations
- Graphoid axioms
- HIV example
- D separation
- Backdoor criterion
- Survival analysis (if time)

Factorisation of the nodes V

Lemma

If V follows a NPSEM-IE, then for any $p(\overline{v}_{j-1})$ with $p(\overline{v}_{j-1}) > 0$ we have that $p(v_j \mid \overline{v}_{j-1}) = p(v_j \mid pa_j)$ and therefore the joint density factorizes as

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

This factorisation is the only restriction that the causal model implies on the law of the observed data.

Thus, in our example from slide 82, the observed law factorizes as

$$p(v) = p(l, a', y) = p(l)p(a' \mid l)p(y \mid a', l),$$

which means that here we put absolutely no restrictions on the law $p(v) \equiv P(V = v)$. `You will prove (part of this lemma) this in your homework.`

No restrictions on $p(v)$ imposed by the NPSEM-IE

We have seen from Slide 71 that the only restriction imposed on the observed law is the factorisation

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

Proof.

Any further restriction must be a restriction on the form of $p(v_j \mid pa_j)$ for any $j \in \{0, \dots, m\}$. But

$$P(V_j = v_j \mid PA_j = pa_j) = P(f_{v_j}(pa_j, U_{v_j}) = v_j),$$

and we have not put any restrictions on the marginal density of U_{v_j}. □

Markov equivalence classes

Definition (Markov equivalence class)

A Markov equivalence class is a set of DAGs that encode the same set of conditional independencies.

Example of markov equivalent DAGs:

$$L \longrightarrow A \longrightarrow Y \quad L \longleftarrow A \longrightarrow Y$$

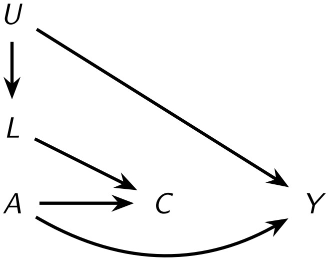
Implication: We cannot use data alone to distinguish between causal graphs.

A clinical story

- Suppose the graph on Slide 75 represents a study of HIV-positive individuals to estimate the effect of an antiretroviral treatment A on 3-year risk of death Y.
- The unmeasured variable $U \in \{0, 1\}$ indicates high level of immunosuppression. Those with $U = 1$ have a greater risk of death.
- Individuals who drop out from the study or are otherwise lost to follow-up are censored ($C = 1$).
- Individuals with $U = 1$ are more likely to be censored because the severity of their disease prevents them from participating in the study.
- The effect of U on censoring C is mediated by the presence of symptoms (fever, weight loss, diarrhea, and so on), CD4 count, and viral load in plasma, all included in L, which could or could not be measured.
- Individuals receiving treatment are at a greater risk of experiencing side effects, which could lead them to dropout, as represented by the arrow from A to C. The square around C indicates that the analysis is restricted to individuals who remained uncensored ($C = 0$) because those are the only ones in which Y can be assessed.

Loss to follow-up example 1

A graph corresponding to the story from Slide 74



Factorisation according to the DAG with ordering $\langle A, U, L, C, Y \rangle$:

$$p(y, c, l, u, a) = p(y \mid u, a)p(c \mid l, a)p(l \mid u)p(u)p(a)$$

But how do we use this factorization to identify causal effects?

Properties of conditional independence

Theorem (Graphoid axioms)

Let X, Y, Z, W be random variables on a Cartesian product space. Conditional independence satsifies

1. $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$ *(Symmetry)*
2. $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$ *(Decomposition)*
3. $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp W \mid Y, Z$ *(Weak union)*
4. $X \perp\!\!\!\perp W \mid Y, Z$ *and* $X \perp\!\!\!\perp Y \mid Z \implies X \perp\!\!\!\perp Y, W \mid Z$ *(Contraction)*
5. *If* $p(x, y, z, w) > 0$, *then* $X \perp\!\!\!\perp W \mid Y, Z$ *and*
   $X \perp\!\!\!\perp Y \mid W, Z \implies X \perp\!\!\!\perp Y, W \mid Z$ *(Intersection)*

`You will study these in your homework.`

Proof of Graphoid axioms

I will not prove all of them here. The fifth identity is part of the homework. I just state a brief proof of the first one.

Proof.

1. Symmetry follows simply because

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z \leftrightarrow p(x \mid z)p(y \mid z) &= p(x, y \mid z) \\ &= p(y \mid z)p(x \mid z) \leftrightarrow Y \perp\!\!\!\perp X \mid Z \ . \end{aligned}$$

□

D separation of a path

Now we will study a beautiful graphical condition on $\mathcal{G}$ that immediately tells if $X \perp\!\!\!\perp Y \mid Z$, where X, Y, Z are disjoint sets of nodes in V, is implied by the Markov factorisation.

Definition (d-separation of a path)

A path r is d-separated by a set of nodes Z iff

1. r contains a chain $V_i \rightarrow V_j \rightarrow V_k$ or a fork $V_i \leftarrow V_j \rightarrow V_k$ such that V_j is in Z, or
2. r contains a collider $V_i \rightarrow V_j \leftarrow V_k$ such that V_j *is not* in Z and such that no descendant of V_j is in Z.

Otherwise the path is d-connected.

D separation of two nodes

Definition (d-separation of two nodes)

Nodes V_i and V_k are d-separated by a set of nodes Z if all trails between V_i and V_k are d-separated by Z. We write d-separation as

$$(V_i \perp\!\!\!\perp V_k \mid Z)_G .$$

If V_i and V_k are not d-separated, they are d-connected and we write

$$(V_i \not\!\perp\!\!\!\perp V_k \mid Z)_G .$$

Theorem (Soundness of d-separation)

$(V_i \perp\!\!\!\perp V_k \mid Z)_G$ *implies the statistical independence*

$$V_i \perp\!\!\!\perp V_k \mid Z.$$

A consequence of soundness is that d-separation in $\mathcal{G}$ implies conditional independence for any distribution that factorizes according to $\mathcal{G}$.

D-separation details and intuition

- D-separation can be shown solely using the Graphoid axioms (but the proof is tedious).
- d-separation allows us to determine independencies of a distribution from the structure of a statistical DAG that represents it.
- Heuristically, two variables are d-separated (independent) if there is no open path between them.

Linear structural equation example

We have not imposed any parametric assumptions so far. However, just for the illustration, suppose we have a (partially) linear structural equation model with two variables satisfying

$$\begin{aligned} A &= f(U_A) \\ Y &= \alpha + \beta A + U_Y \end{aligned} \tag{6}$$

This structural equation model implies that the individual level causal effects is $Y^{a=1} - Y^{a=0} = \beta$!

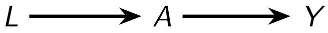
We conclude that the linear equation model relies on extremely strong assumptions that usually will be implausible. In this course, we will not rely on such assumptions.

Modified non-parametric example

A different SEM $\mathcal{M}$

$$\begin{aligned} L &= f_L(U_L) \\ A &= f_A(L, U_A) \\ Y &= f_Y(A, U_Y) \end{aligned} \tag{7}$$

and the graph $\mathcal{G}$,



- Encodes that, changes in L leaves Y unchanged, provided that U_Y and A remain constant.
- Does this graph encode any restrictions on the distribution of (L, A, Y)?
  `We will formally study what kind of restrictions the structural models involve`

Faithfulness and completeness of d-separation

Definition

A law $\mathbb{P}$ is faithful to a DAG $\mathcal{G}$ if for any disjoint set of nodes A, B, C we have that $A \perp\!\!\!\perp C \mid B$ under $\mathbb{P}$ implies $(A \perp\!\!\!\perp C \mid B)_{\mathcal{G}}$.

Theorem (Completeness of d-separation)

In a Bayesian Network with respect to a direct acyclic graph $\mathcal{G}$ there exists a faithful law $\mathbb{P}$.

We will not prove this important result[11].

The completeness of d-separation allows us to use d-separation to represent the conditional independence structure of a multivariate distribution.

You can look at the graph, and read off all independencies that hold in the entire class of distributions factorizing according to the DAG.

[11] Ann Becker, Dan Geiger, and Christopher Meek. "Perfect tree-like markovian distributions". In: *arXiv preprint arXiv:1301.3834* (2013); Pearl, *Causality: Models, Reasoning and Inference 2nd Edition*.

The causal Markov assumption and faithfulness (intuition and interpretation)

- d-separation implies statistical independence, but does not allow one to deduce that d-connection implies statistical dependence.
- However, d-connected variables will be independent only if there is an exact balancing of positive and negative causal effects.
- Because such precise balancing of effects is highly unlikely to occur, we shall henceforth generally assume that d-connected variables are dependent.

Backdoor adjustment

Definition (Backdoor path)

In a DAG $\mathcal{G}$ a backdoor path between two nodes V_i and V_j is a trail that starts in V_i and ends in V_j; and with initial edge being an arrow pointing into V_i

Example backdoor path between V_i and V_j is: $V_i \leftarrow V_k \rightarrow V_j$.

Backdoor theorem

Theorem (Backdoor theorem wrt. to a DAG)

In DAG $\mathcal{G}$ representing a NPSEM-IE, let X, Y and Z be three sets of nodes of $\mathcal{G}$, each comprised of one or more nodes. Suppose that X contains no descendants of Z and it blocks all back-door paths between any node in Z and any node in Y: Suppose that $g = (g_1, \dots, g_t)$ is a regime for $Z = (Z_1, \dots, Z_t)$ (for some $t \geq 1$) such that treatment assignments depend at most on X: Then, for any x in the support of X such that $p(Z = g(x) \mid x) \equiv Pr(Z = g(x) \mid X = x) > 0$; it holds that

$$P(Y^g = y) = \sum_x P(Y = y \mid Z = g(x), X = x) P(X = x)$$

See Pearl[12] for proof (not required). This theorem is very useful, because it allows us to identify causal effects even if certain nodes in the graph are unmeasured.

[12] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.

Implication from the Backdoor theorem

It follows immediately from the backdoor theorem that if $Y^a \perp\!\!\!\perp A \mid L$ then

$$P(Y^a = y) = \sum_l P(Y = y \mid L = l, A = a)P(L = l).$$

However, we can also use it to identify causal effects in much more complicated settings, which also involve unmeasured variables.

Backdoor theorem in the example on loss to follow-up

Consider the example from Slide 75.

- Note that
  - L blocks all backdoor paths between (A, C) and Y.
  - Thus,

$$\mathbb{E}(Y^{a,c=0}) = \sum_{l} \mathbb{E}(Y \mid A = a, C = 0, L = l) P(L = l),$$

which can be estimated simply by standardisation:

- Estimate $\mathbb{E}(Y \mid A = a, C = 0, L = l)$ by $\hat{\mathbb{E}}(Y \mid A = a, C = 0, L = l)$,
- Estimate $P(L = l)$ empirically.
- Standardise

PS: Many causal questions are more difficult

Realistic questions are often more difficult. Consider for example:

- when should we start a treatment?
- How long should we continue treatment?
- When to switch to different treatment?
- What event should guide us to switch treatment?

PS: Many causal questions are more difficult

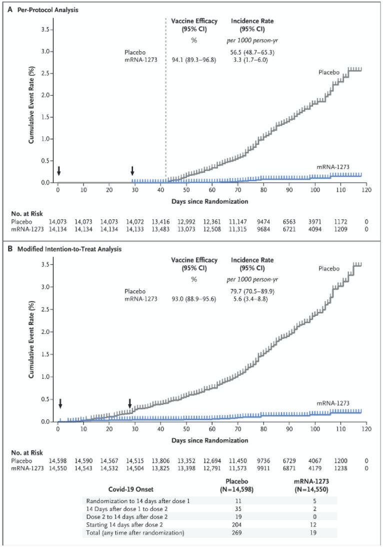
Realistic questions are often more difficult. Consider for example:

- When should we start a treatment?
- How long should we continue treatment?
- When to switch to different treatment?
- What event should guide us to switch treatment?

Section 8

Time-to-events and survival analysis

The Moderna vaccine



A Per-Protocol Analysis

No. at Risk													
Placebo	14,073	14,073	14,073	14,072	13,416	12,992	12,361	11,147	9474	6563	3971	1172	0
mRNA-1273	14,134	14,134	14,134	14,133	13,483	13,073	12,508	11,315	9684	6721	4094	1209	0

B Modified Intention-to-Treat Analysis

No. at Risk													
Placebo	14,598	14,590	14,567	14,515	13,806	13,352	12,694	11,450	9736	6729	4067	1200	0
mRNA-1273	14,550	14,543	14,532	14,504	13,825	13,398	12,791	11,573	9911	6871	4179	1238	0

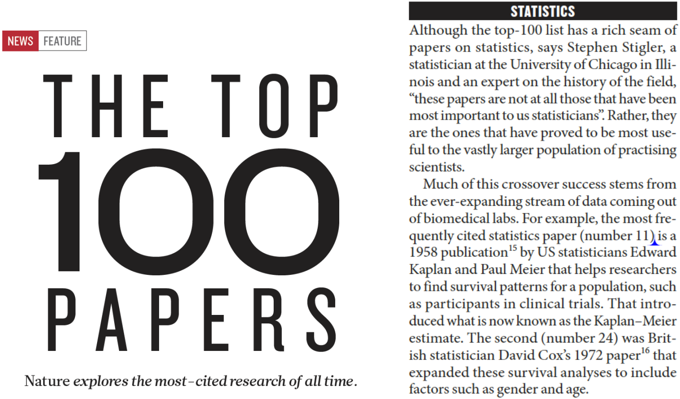
Covid-19 Onset	Placebo (N=14,598)	mRNA-1273 (N=14,550)
Randomization to 14 days after dose 1	11	5
14 Days after dose 1 to dose 2	35	2
Dose 2 to 14 days after dose 2	19	0
Starting 14 days after dose 2	204	12
Total (any time after randomization)	269	19

Figure 3: Survival analysis is e.g. used to present results from vaccine trials.

Time to events are all over the place

- Time from birth to death.
- Time from birth to cancer diagnosis.
- Time from disease onset to death.
- Time from entry to a study to cancer relapse.
- Time from marriage to divorce.
- Time from production until a machine is broken.
- Time from origin of the coronavirus until a stock (marked) crashes.

Take metrics with a grain of salt, but...



NEWS FEATURE

THE TOP 100 PAPERS

Nature *explores the most-cited research of all time.*

STATISTICS

Although the top-100 list has a rich seam of papers on statistics, says Stephen Stigler, a statistician at the University of Chicago in Illinois and an expert on the history of the field, "these papers are not at all those that have been most important to us statisticians". Rather, they are the ones that have proved to be most useful to the vastly larger population of practising scientists.

Much of this crossover success stems from the ever-expanding stream of data coming out of biomedical labs. For example, the most frequently cited statistics paper (number 11) is a 1958 publication[15] by US statisticians Edward Kaplan and Paul Meier that helps researchers to find survival patterns for a population, such as participants in clinical trials. That introduced what is now known as the Kaplan–Meier estimate. The second (number 24) was British statistician David Cox's 1972 paper[16] that expanded these survival analyses to include factors such as gender and age.

Figure 4: The two most cited statistics papers concern survival analysis

Some common questions

- What is survival under treatment A vs B?
- What is the duration of a certain component in the machine?
- How long does it take before a stock marked crashes?

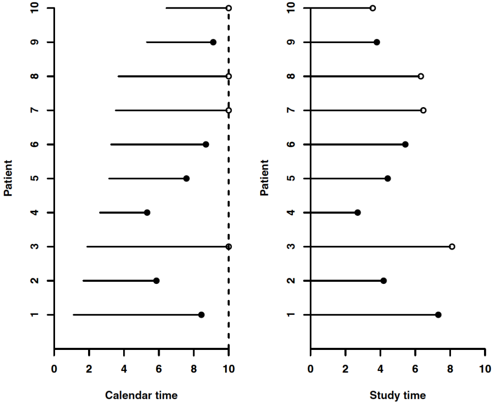
`PS: These questions are very often about causal effects....`

An overview of the time-to-event data structure

- We follow units of over time;
  humans, animals, engines, etc.
- The events of interest may be **the time to** deaths, cancer diagnoses, divorces, child births, engine failures, etc.
- We often stop the study before everyone has experienced the event of interest.

Censored survival times (illustration)

Consider 10 patients with newly diagnosed cancer. Let $T \in (0, \tau]$ be a survival time.



7.32, 4.19, 8.11, 2.70, 4.42, 5.43, 6.46, 6.32, 3.80, 3.50.
How do you estimate $\mathbb{E}(T)$, that is, the mean survival?

One way to define censoring

Definition (Censoring)

A censoring event is any event occurring in the study by time t that ensures the values of all future (possibly counterfactual) outcomes of interest under a regime g are unknown, even for an individual receiving the intervention g.

- This definition covers observational (non-causal) settings as a special case, by considering a regime g which implements exactly the decision rule that was used in the observed data.
- Many other definitions exist in the literature. I will argue why this definition is useful.

Why not use "standard methods"?

- We have incomplete observations.
- Instead of observing the survival time $T_i \in (0, \infty)$ we observe $(\tilde{T}_i, D_i)$,

$$\tilde{T}_i = T_i \text{ if } D_i = 1,$$
$$\tilde{T}_i < T_i \text{ if } D_i = 0.$$

  where D_i is a censoring indicator.

  We want to use our information on $\tilde{T}_i$ to make inference on T_i.
- There is a strong link to causal inference and "what if" questions: What would happen if we observed T_i instead of $\tilde{T}_i$.
- We must make assumptions about the censoring, similarly to assumptions in causal inference.

Let's start with a single outcome process

Assume $T > 0$ is an absolutely continuous random variable.

Definition (Survival function)

The survival function is $S(t) = P(T > t)$, that is, the probability that the survival time T exceeds t.

Definition (Hazard rate)

The hazard rate $\alpha(t) = \lim_{dt \to 0} \frac{1}{dt} P(t + dt > T \geq t \mid T \geq t)$ is the rate of events per unit of time.

Informally, $\alpha(t)dt = P(t + dt > T > t \mid T \geq t)$ is the probability that the event will happen between time t and time $t + dt$ given that it has not happened earlier.[13]

[13] PS: We are going to extend this to multiple events later.

Cumulative hazard and some relations

Define the cumulative hazard,

$$H(t) = \int_0^t \alpha(s) ds.$$

Then,

$$H'(t) = \alpha(t) = \lim_{dt \to 0} \frac{1}{dt} \frac{S(t) - S(t + dt)}{S(t)} = -\frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)}.$$

By integration

$$\int_0^t \alpha(s) ds = -\log\{S(t)\},$$

and thus

$$S(t) = \exp\{-\int_0^t \alpha(s) ds\}.$$

$\alpha(t)$ completely determines the distribution of survival times T.

Illustration of hazards and survival functions

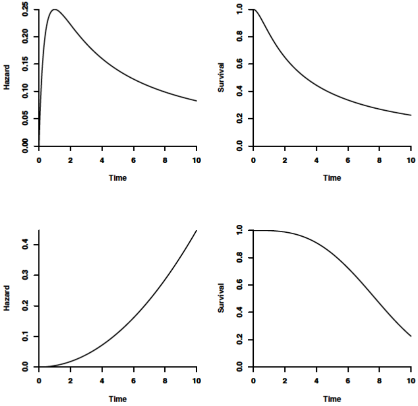


Fig. 1.2 *Illustrating hazard rates and survival curves. The hazard rates on the left correspond to the survival curves on the right.*