Probabilistic and Graphical Models of Causality

Wojciech Niemiro

January 17, 2024

Contents

Chapter 1

Graphs and Conditional Independence

1.1 Introduction

Interventional vs observational conditioning

In probability theory the notion of conditional distribution is defined in terms of a probability measure which describes the joint distribution of all random variables under consideration. In most applications, however, the order is reversed. The joint distribution is usually defined in terms of conditional (and marginal) distributions. In particular, *causal* relations between variables are adequately modelled by *conditional-by-intervention distributions.* As the simplest example let us consider two random variables (random elements) $\mathbf{X}$ and $\mathbf{Y}$. Assume that $\mathbf{X}$ is a cause of $\mathbf{Y}$ ($\mathbf{Y}$ is the effect of $\mathbf{X}$). We shall then write $\mathbf{X} \to \mathbf{Y}$. The joint distribution of both the variables can be defined by $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$, that is by the marginal probability of the cause and the conditional probability of the effect.

Now imagine a controlled experiment in which we *force* variable $\mathbf{X}$ to assume the value $\mathbf{x}$. Denote by $p(\mathbf{y}\|\mathbf{x})$ the probability density of $\mathbf{Y}$ given that the value of $\mathbf{X}$ is set to $\mathbf{x}$. If $\mathbf{X} \to \mathbf{Y}$ and there is no causal influence in the opposite direction, $\mathbf{Y} \nrightarrow \mathbf{X}$, then the conditioning-by-intervention reduces to the standard notion of conditioning: $p(\mathbf{y}\|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$. On the other hand we then have $p(\mathbf{x}\|\mathbf{y}) = p(\mathbf{x}) \neq p(\mathbf{x}|\mathbf{y})$.

However, we would like to study the situation where $\mathbf{X} \to \mathbf{Y}$ *and* $\mathbf{Y} \to \mathbf{X}$, that is allow for a feedback. If $\mathbf{X}$ and $\mathbf{Y}$ are not static random variables but rather stochastic processes evolving in time then possible presence of causal infuences in both directions becomes natural and obvious. The evolution of $\mathbf{Y}$ in the future can depend on the past and the present state of $\mathbf{X}$ and vice versa. The "causal dilemma" diappears: we are not to choose between $\mathbf{X} \to \mathbf{Y}$ and $\mathbf{Y} \to \mathbf{X}$ but rather, taking into account the ordering of events in time, decide separately if each of the two causal relations holds.

Introductory example: 2 variables

- If we assume that $\mathbf{X}$ is a **cause** of $\mathbf{Y}$ ($\mathbf{Y}$ is an **effect** of $\mathbf{X}$) then we write:

$$\begin{tikzcd}\mathbf{X} \arrow[r] & \mathbf{Y}\end{tikzcd}$$

- If $\mathbf{X}$, $\mathbf{Y}$ are random variables (or stochastic processes) then we usually specify the joint probability distribution via:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$$

  (joint probability)=(marginal probability of cause)×(conditional probability of effect)

- Conditioning-by-intervention. Imagine an experiment in which we **control** (set) value $\mathbf{X} = \mathbf{x}$ and **observe** $\mathbf{Y}$:

$$p(\mathbf{y}\|\mathbf{x}) = \mathbb{P}(\mathbf{Y} = \mathbf{y}\|\mathbf{X} = \mathbf{x})$$

- If $\mathbf{X} \rightarrow \mathbf{Y}$ and $\mathbf{X} \not\leftarrow \mathbf{Y}$ then:

$$\begin{aligned} p(\mathbf{y}\|\mathbf{x}) &= p(\mathbf{y}|\mathbf{x}) \\ p(\mathbf{x}\|\mathbf{y}) &= p(\mathbf{x}) \neq p(\mathbf{x}|\mathbf{y}) \end{aligned}$$

Conditional independences for 3 variables

Consider random variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and examine conditional independences between $\mathbf{X}$ and $\mathbf{Z}$ for different sets of causal relations.

- **Chain** connection:

$$\begin{tikzcd}\mathbf{X} \arrow[r] & \mathbf{Y} \arrow[r] & \mathbf{Z}\end{tikzcd}$$

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})p(\mathbf{z}|\mathbf{y})$$

  We have $\mathbf{X} \perp\!\!\!\perp \mathbf{Z}|\mathbf{Y}$ but $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Z}$

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= \underbrace{p(\mathbf{x})p(\mathbf{y}|\mathbf{x})}_{\phi(\mathbf{x},\mathbf{y})}\underbrace{p(\mathbf{z}|\mathbf{y})}_{\psi(\mathbf{z},\mathbf{y})} \\ p(\mathbf{x}, \mathbf{z}) &= \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}, \mathbf{z}) \neq \phi(\mathbf{x})\psi(\mathbf{z}) \quad [\text{ in general }] \end{aligned}$$

  Example: $\mathbf{X}$ - I do not have Xmas gifts, $\mathbf{Y}$ - I go to a busy shopping mall, $\mathbf{Z}$ - I catch COVID 19.

- **Fork** connection:

  $$\mathbf{X} \longleftarrow \mathbf{Y} \longrightarrow \mathbf{Z}$$

  $$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y})p(\mathbf{z}|\mathbf{y})$$

  $\mathbf{X} \perp\!\!\!\perp \mathbf{Z}|\mathbf{Y}$ but $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Z}$

  $$p(\mathbf{x},\mathbf{y},\mathbf{z}) = \underbrace{p(\mathbf{y})p(\mathbf{x}|\mathbf{y})}_{\phi(\mathbf{x},\mathbf{y})}\underbrace{p(\mathbf{z}|\mathbf{y})}_{\psi(\mathbf{z},\mathbf{y})}$$

  $$p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{y}} p(\mathbf{x},\mathbf{y},\mathbf{z}) \neq \phi(\mathbf{x})\psi(\mathbf{z}) \quad [\text{ in general }]$$

  Example (due to J. Noble): **X** - I wake up with a headache, **Z** - I wake up in galoshes, **Y** - I spent the evening before in a pub.

- **Collider** connection:

  $$\mathbf{X} \longrightarrow \mathbf{Y} \longleftarrow \mathbf{Z}$$

  $$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x})p(\mathbf{z})p(\mathbf{y}|\mathbf{x},\mathbf{z})$$

  $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Z}|\mathbf{Y}$ ale $\mathbf{X} \perp\!\!\!\perp \mathbf{Z}$

  $$p(\mathbf{x},\mathbf{y},\mathbf{z}) \neq \phi(\mathbf{x},\mathbf{y})\psi(\mathbf{z},\mathbf{y}) \quad [\text{ in general }]$$

  $$p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{y}} p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x})p(\mathbf{z})$$

  Example: **X** - burglary, **Z** - earthquake, **Y** - alarm is switched on.

SUMMARY:

When trail **X** — **Y** — **Z** is blocked ?

That is: when the flow of information from **X** to **Z** is blocked (ie. possibility of observing **X** does not increase information about **Z** and *vice versa*) ? The answers are the following:

- For a chain : $\mathbf{X} \longrightarrow \boxed{\mathbf{Y}} \longrightarrow \mathbf{Z}$, $\quad \mathbf{X} \perp\!\!\!\perp \mathbf{Z}|\mathbf{Y}$
  (if we observe **Y**)
- For a fork: $\mathbf{X} \longleftarrow \boxed{\mathbf{Y}} \longrightarrow \mathbf{Z}$, $\quad \mathbf{X} \perp\!\!\!\perp \mathbf{Z}|\mathbf{Y}$
  (if we observe **Y**)
- For a collider: $\mathbf{X} \longrightarrow \mathbf{Y} \longleftarrow \mathbf{Z}$, $\quad \mathbf{X} \perp\!\!\!\perp \mathbf{Z}|\emptyset$
  (if we do not observe **Y**)

(observed variable is encased in a “box”)

1.2 Causal models based on DAGs

Directed Acyclic Graphs

Throughout $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a directed graph (DG) without self-loops

- $\mathcal{V}$ is a finite set of nodes (vertices).
- $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$; $\mathcal{E} \not\ni (v, v)$ for any $v \in \mathcal{V}$.
- We write $v \to w$ or, equivalently, $w \leftarrow v$ if $(v, w) \in \mathcal{E}$. An element $e \in \mathcal{E}$ is called an arrow (or a directed edge).
- For $v \in \mathcal{V}$ we put $\mathrm{pa}(v) = \{w : w \to v\}$.

Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is acyclic (**Directed Acyclic Graph**, DAG) if there is no directed path from a node to itself (no sequence $v \to u_1 \to \cdots \to u_k \to v$).

More graph-theoretical definitions in a more general setup will be introduced in Section 1.4.

Factorisation of probability distributions

Let $(\mathbf{X}_v : v \in \mathcal{V})$ be a collection of random variables indexed by nodes of a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For a subset of nodes $\mathcal{A} \subseteq \mathcal{V}$ we write $\mathbf{X}_\mathcal{A} = (\mathbf{X}_v : v \in \mathcal{A})$; in particular $\mathbf{X} = \mathbf{X}_\mathcal{V}$. We also write $p(\mathbf{x}_\mathcal{A}) = \mathbb{P}(\mathbf{X}_\mathcal{A} = \mathbf{x}_\mathcal{A})$ (assuming that the radom variables are discrete; otherwise p is a density wrt. some reference measure denoted by $\mathrm{d}\mathbf{x}_\mathcal{A} = \prod_{v \in \mathcal{A}} \mathrm{d}\mathbf{x}_v$). As usual, we use script letters for the respective spaces eg. $\mathbf{x}_v \in \mathcal{X}_v$ and $\mathbf{x}_\mathcal{A} \in \mathcal{X}_\mathcal{A}$.

A causal model is constructed from a collection of conditional distributions $p(\mathbf{x}_v | \mathbf{x}_{\mathrm{pa}(v)})$ for $v \in \mathcal{V}$:

$$p(\mathbf{x}) = \prod_{v \in \mathcal{V}} p(\mathbf{x}_v | \mathbf{x}_{\mathrm{pa}(v)}) \tag{1.2.1}$$

(if $\mathrm{pa}(v) = \emptyset$ then it is understood that $p(\mathbf{x}_v | \mathbf{x}_{\mathrm{pa}(v)}) = p(\mathbf{x}_v)$ – marginal distribution).

Remark. The factorisation formula (1.2.1) is a *correct* definition of joint probability distribution, because $\mathcal{G}$ is a DAG ! Formally speaking, it must be verified that (a) $\int_\mathcal{X} p(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$ and (b) the factors on the RHS of (1.2.1) indeed satisfy $p(\mathbf{x}_v | \mathbf{x}_{\mathrm{pa}(v)}) = p(\mathbf{x}_{\mathrm{pa}(v) \cup v}) / p(\mathbf{x}_{\mathrm{pa}(v)})$. Since $\mathcal{G}$ is acyclic, the nodes can be ordered in such a way that w always precede v whenever $w \to v$. Statements (a) and (b) easily follow from this.

Remark. In (1.2.1) the conditional distributions can be interpreted as **causal** ie. conditional-by-intervention[1]:

$$p(\mathbf{x}_v|\mathbf{x}_{\mathrm{pa}(v)}) = p(\mathbf{x}_v\|\mathbf{x}_{\mathrm{pa}(v)})$$

Structural Causal Models (SCMs)

We shall present another way of looking at graphical causal models. It is based on the following facts:

- If $\mathbf{Y}$ is an arbitrary random variable with values in space $\mathcal{Y}$[2] than there exists a function $\psi : [0,1] \to \mathcal{Y}$ such that
$$\mathbf{Y} =_d \psi(U)$$
- If $(\mathbf{X},\mathbf{Y})$ is an arbitrary random variable with values in space $\mathcal{X} \times \mathcal{Y}$ than there exists a function $\psi : \mathcal{X} \times [0,1] \to \mathcal{Y}$ such that
$$(\mathbf{X},\mathbf{Y}) =_d \psi(\mathbf{X}, U)$$
where $U \sim \mathrm{U}(0,1)$ (a uniform variate).

The definition of an SCM (governed by a DAG $\mathcal{G} = (\mathcal{V},\mathcal{E})$) is the following:

- We have a collection of functions $(\psi_v : v \in \mathcal{V})$, where $\psi_v : \mathcal{X}_{\mathrm{pa}(v)} \times [0,1] \to \mathcal{X}_v$. (If $\mathrm{pa}(v) = \emptyset$ then $\psi_v : [0,1] \to \mathcal{X}_v$.) Let $(U_v : v \in \mathcal{V})$ be a collection of independent uniforms:
$$U_v \sim_{\text{i.i.d.}} \mathrm{U}(0,1)$$
- Put
$$\mathbf{X}_v = \psi_v(\mathbf{X}_{\mathrm{pa}(v)}, U_v)$$

It is clear that this definition is correct (leads to no contradiction provided that $\mathcal{G}$ is a DAG). It is equivalent to the definition given before ie. specification of conditional distributions $p(\mathbf{x}_v|\mathbf{x}_{\mathrm{pa}(v)})$ and has an appealing interpretation:

- Functions ψ_v describe deterministic *causal mechanisms.*
- U_vs are *noice variables.*

[1] Formal definition of conditioning-by-intervention will be given (in a more general setup) in Section 1.3.

[2] We neglect measure-theoretic issues here. In reality it is required that $\mathcal{Y}$ (and also $\mathcal{X}$ below) are Borel measurable spaces. However, most spaces encountered in applications ($\mathbb{R}$, $\mathbb{R}^d$, discrete spaces) fulfil this requirement).

1.3 Stochastic processes and CREs

Time series and DBNs

Let $\mathbf{X} = (X_v(t))$ be a multivariate stochastic process in discrete time ($t = 0, 1, \ldots, u$), with components indexed by $v \in \mathcal{V}$. As always in this section, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a DG *with possible cycles* but without self-loops. Let $\mathcal{G}' = (\mathcal{V}, \mathcal{E}')$ be a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with added self-loops (arrows $v \to v$) for some (not necessarily all) nodes v. Put $\text{pa}'(v) = \{w : (w, v) \in \mathcal{E}'\}$.

Assume the following:

$$p(x(0)) = \prod_{v \in \mathcal{V}} p(x_v(0)) \tag{1.3.1a}$$

$$p(x(t)|x(t-1), \ldots, x(0)) = \prod_{v \in \mathcal{V}} p(x_v(t)|x_{\text{pa}'(v)}(t-1), \ldots, x_{\text{pa}'(v)}(0)) \tag{1.3.1b}$$

The joint distribution of $\mathbf{X}$ can be defined in terms of initial density $p(x(0))$ and subsequent conditional densities $p(x(t)|x(t-1), \ldots, x(0))$:

$$\begin{aligned} p(\mathbf{x}) &= p(x(0)) \prod_{t=1}^{u} p(x(t)|x(t-1), \ldots, x(0)) \\ &= \text{ under (1.3.1)} \\ &= \left[\prod_{v \in \mathcal{V}} p(x_v(0))\right] \prod_{t=1}^{u} \prod_{v \in \mathcal{V}} p(x_v(t)|x_{\text{pa}'(v)}(t-1) \ldots, x_{\text{pa}'(v)}(0)) \\ &= \prod_{v \in \mathcal{V}} \underbrace{\left[p(x_v(0)) \prod_{t=1}^{u} p(x_v(t)|x_{\text{pa}'(v)}(t-1) \ldots, x_{\text{pa}'(v)}(0))\right]}_{=p(\mathbf{x}_v \| \mathbf{x}_{\text{pa}(v)})} \end{aligned}$$

We thus obtain the following strange looking formula:

$$p(\mathbf{x}) = \prod_{v \in \mathcal{V}} p(\mathbf{x}_v \| \mathbf{x}_{\text{pa}(v)}) \tag{1.3.2}$$

Moreover

$$p(\mathbf{x}_v \| \mathbf{x}_{\text{pa}(v)}) = p(x_v(0)) \prod_{t=1}^{u} p(x_v(t)|x_{\text{pa}'(v)}(t-1) \ldots, x_{\text{pa}'(v)}(0))$$

can indeed be interpreted as conditional-by-intervention distribution. This is the probability distribution of $\mathbf{X}_v$ if the trajectories of $\mathbf{X}_{\text{pa}(v)}$ are set to $\mathbf{x}_{\text{pa}(v)}$. A more general and formal definition of conditioning-by-intervention is given later in this section.

If $\mathbf{X}$ is a Markov chain then

$$p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) = p(x_v(0)) \prod_{t=1}^{u} p(x_v(t) | x_{\mathrm{pa}'(v)}(t-1))$$

Example: For two nodes and the graph $\mathbf{X} \rightleftarrows \mathbf{Y}$ we have

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x} \| \mathbf{y}) p(\mathbf{y} \| \mathbf{x})$$

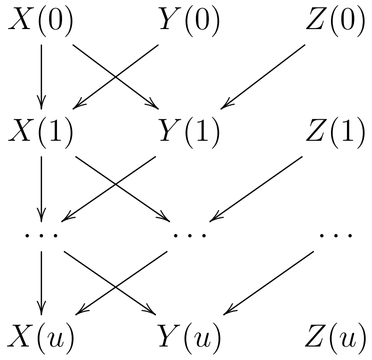
1.3.3 EXAMPLE. Let us consider process $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (X(t), Y(t), Z(t) : t = 1, \dots, u)$. Assume that Markov property holds. Graph $\mathcal{G}$ is depicted below:

$$\begin{tikzcd} \mathbf{X} \arrow[r, bend right] & \mathbf{Y} \arrow[l, bend right] & \mathbf{Z} \arrow[l] \end{tikzcd}$$

The augmented graph $\mathcal{G}'$ is obtained by adding the self-loop $\mathbf{X} \to \mathbf{X}$, indicated below by a dotted arrow:

$$\begin{tikzcd} \mathbf{X} \arrow[loop left, dotted] \arrow[r, bend right] & \mathbf{Y} \arrow[l, bend right] & \mathbf{Z} \arrow[l] \end{tikzcd}$$

The structure of dependence of the $3(u+1)$ random variables is given by the following graph:



We have

$$p(\mathbf{x} \| \mathbf{y}) = p(x(0)) \prod_{t=1}^{u} p(x(t) | x(t-1), y(t-1)),$$
$$p(\mathbf{y} \| \mathbf{x}, \mathbf{z}) = p(y(0)) \prod_{t=1}^{u} p(y(t) | x(t-1), z(t-1)),$$
$$p(\mathbf{z}) = p(z(0)) \prod_{t=1}^{u} p(z(t))$$

and the joint distribution is

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x} \| \mathbf{y}) p(\mathbf{y} \| \mathbf{x}, \mathbf{z}) p(\mathbf{z})$$

Moreover, it is easy to verify that, for example, $p(\mathbf{x}, \mathbf{y} \| \mathbf{z}) = p(\mathbf{x} \| \mathbf{y}) p(\mathbf{y} \| \mathbf{x}, \mathbf{z})$ can be interpreted as the conditional-by-intervention distribution of $(\mathbf{X}, \mathbf{Y})$ given $\mathbf{Z} = \mathbf{z}$. In particular, $\iint p(\mathbf{x}, \mathbf{y} \| \mathbf{z}) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} = 1$. Analogous remark applies to conditioning-by-intervention on other subsets of nodes. △

Note that in the above example we defined an *acyclic* directed graph (DAG) on the set of "**space-and-time**" nodes (v, t), where $v \in \mathcal{V}$ and $t \in \{0, 1, \dots, u\}$. The set of edges is defined as follows: $(v, s) \to (w, t)$ whenever $s < t$ and $(v, w) \in \mathcal{G}'$. (If the process is Markov then we have only edges $(v, t-1) \to (w, t)$ for $(v, w) \in \mathcal{G}'$.)

The collection of random variables $(X_v(t))$ arranged in the"space-and-time" DAG is named a dynamic bayesian network (DBN).

Continuous Time Bayesian Networks (CTBN)

Let $\mathbf{X} = (X(t) : 0 \leqslant t \leqslant u)$ be a Markov process with values in a finite state space $\mathcal{S}$. The distribution of $\mathbf{X}$ is characterised by its initial distribution, say ν, and the transition intensities. Assume that there exist bounded functions $Q(t; x, x')$ such that for $x, x' \in \mathcal{S}$,

$$\mathbb{P}(X(t+h) = x' | X(t) = x) = \begin{cases} Q(t; x, x')h + o(h) & \text{for } x \neq x' \\ 1 + Q(t; x, x)h + o(h) & \text{for } x = x' \end{cases} \tag{1.3.4}$$

as $h \searrow 0$, where $Q(t;\, x, x) = -\sum_{x' \neq x} Q(t; x, x')$. We also have that $\mathbb{P}(x(0) = x) = \nu(x)$.

The density of the process $\mathbf{X}$ (regarded as a random element in the space of trajectories $\mathbf{x}$) is given by the formula

$$p(\mathbf{x}) = \nu(x(0)) \prod_{t : x(t-) \neq x(t)} Q(t;\, x(t-), x(t)) \times \exp\left[\int_0^u Q(t;\, x(t), x(t)) \mathrm{d}t \right] \tag{1.3.5}$$

We provide a derivation of formula (1.3.5) in Appendix A.

The analogue of (1.3.1) in the present setting is the following "infinitesimal independence" condition (accompanied by the independence condition on the initial distribution):

$$\nu(x) = \prod_v \nu(x_v) \tag{1.3.6a}$$

$$\mathbb{P}(X(t+h) = x'|X(t) = x) = \prod_v \left[\mathbb{P}(X_v(t+h) = x'_v|X(t) = x) + o(h)\right] \tag{1.3.6b}$$

$$(h \searrow 0) \text{ for } 0 \leqslant t < u$$

It is clear that the RHS of (1.3.6b) is $o(h)$ whenever $x_v \neq x'_v$ and $x_w \neq x'_w$ for some $v \neq w$[3]. Thus jumps of $\mathbf{X}$ cannot occur at more than one coordinate simultaneously.

Additionally assume that the components are indexed by nodes of a (possibly cyclic) directed graph $\mathcal{G}$ and, for $x'_v \neq x_v$,

$$\mathbb{P}(X_v(t+h) = x'_v|X(t) = x) = \mathbb{P}(X_v(t+h) = x'_v|X_{\mathrm{pa}(v)}(t) = x_{\mathrm{pa}(v)}, X_v(t) = x_v) + o(h)$$

The transition intensities are then of the following form: for $x \neq x'$,

$$Q(t;\, x, x') = \begin{cases} Q_v(t;\, x_{\mathrm{pa}(v)};\, x_v, x'_v) & \text{if } x_v \neq x'_v, x_{-v} = x'_{-v} \\ 0 & \text{if there are } v \neq w \text{ such that } x_v \neq x'_v, x_w \neq x'_w \end{cases}$$

$Q_v(t;\, x_{\mathrm{pa}(v)};\, \cdot, \cdot)$ is a matrix of intensities of transitions at node v at time t if the configuration of parent nodes at this time is $x_{\mathrm{pa}(v)}$. We say that $\mathbf{X}$ is a **Contionuous Time Bayesian Network** (CTBN) wrt. graph $\mathcal{G}$[4].

It is easy to verify that $Q(t;\, x, x) = \sum_v Q_v(t;\, x_{\mathrm{pa}(v)};\, x_v, x_v)$, where we put $Q_v(t;\, x_{\mathrm{pa}(v)};\, x_v, x_v) = -\sum_{x'_v \neq x_v} Q_v(t;\, x_{\mathrm{pa}(v)};\, x_v, x'_v)$

It follows that for a CTBN we have

$$p(\mathbf{x}) = \prod_{v \in \mathcal{V}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)})$$

where

$$p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) = \nu(x_v(0)) \prod_{t: x_v(t-) \neq x_v(t)} Q_v(t, x_{\mathrm{pa}(v)}(t);\, x_v(t-), x_v(t))$$
$$\times \exp\left[\int_0^u Q_v(t;\, x_{\mathrm{pa}(v)}(t);\, x_v(t), x_v(t)) \mathrm{d}t\right]$$

Consider a subset $\mathcal{A}$ of nodes and the expression $\prod_{v \in \mathcal{A}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)})$. A moment of reflection shows that this expression, regarded as a function of $\mathbf{x}_{\mathcal{A}}$, is the probability density of a Markov process $\mathbf{X}_{\mathcal{A}}$ obtained when we set by intervention $\mathbf{X}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}} = \mathbf{x}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}}$. In particular $\int \prod_{v \in \mathcal{A}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) \mathrm{d}\mathbf{x}_{\mathcal{A}} = 1$. Let us illustrate this on the following example.

[3] in view of (1.3.4), $\mathbb{P}(X_v(t+h) = x'_v|X(t) = x) = O(h)$ whenever $x'_v \neq x_v$.

[4] Usual definition of CTBN assumes time-homogeneity, ie. the transition intensities Q_v do not depend on t. For our purposes this assumption is not needed.

1.3.7 EXAMPLE. Let us consider a Markov process $(\mathbf{X},\mathbf{Y},\mathbf{Z}) = (X(t), Y(t), Z(t) : t \in [0,u])$. Assume $\mathbf{X}$ is a CTBN with respect to the following graph $\mathcal{G}$:

$$\begin{tikzcd}\mathbf{X} \arrow[r, bend right] & \mathbf{Y} \arrow[l, bend right] & \mathbf{Z} \arrow[l]\end{tikzcd}$$

Thus $\mathbf{X}$ is defined via the following conditional intensities: $Q_x(t;\, y;\, x, x')$, $Q_y(t;\, (x,z);\, y, y')$ and $Q_z(t;\, z, z')$. The density of $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ is given by the formula

$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x}\|\mathbf{y}) \cdot p(\mathbf{y}\|\mathbf{x},\mathbf{z}) \cdot p(\mathbf{z})$$

where

$$p(\mathbf{x}\|\mathbf{y}) = \nu(x(0)) \prod_{t: x(t-)\neq x(t)} Q_x(t;\, y(t);\, x(t-), x(t)) \cdot \exp\left[\int Q_x(t;\, y(t);\, x(t), x(t))\mathrm{d}t\right]$$

$$p(\mathbf{y}\|\mathbf{x},\mathbf{z}) = \nu(y(0)) \prod_{t: y(t-)\neq y(t)} Q_y(t;\, x(t), z(t);\, y(t-), y(t)) \cdot \exp\left[\int Q_y(t;\, x(t), z(t);\, y(t), y(t))\mathrm{d}t\right]$$

$$p(\mathbf{z}) = \nu(z(0)) \prod_{t: z(t-)\neq z(t)} Q_z(t;\, z(t-), z(t)) \cdot \exp\left[\int Q_z(t;\, z(t), z(t))\mathrm{d}t\right]$$

If for example we fix $\mathbf{z}$, then it is easy to notice that $p(\mathbf{x}\|\mathbf{y})p(\mathbf{y}\|\mathbf{x},\mathbf{z})$ is the probability density of the process $(\mathbf{X},\mathbf{Y})$ with transition intensities

$$Q_x^{|\mathbf{z}}(t;\, y;\, x, x') = Q_x(t;\, y;\, x, x')$$
$$Q_y^{|\mathbf{z}}(t;\, x;\, y, y') = Q_x(t;\, x, z(t);\, y, y')$$

Therefore $\int p(\mathbf{x}\|\mathbf{y})p(\mathbf{y}\|\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = 1$ for every $\mathbf{z}$. Analogously $\int p(\mathbf{y}\|\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{y} = 1$ for every $(\mathbf{x},\mathbf{z})$ and so on. △

Composable Random Elements (CRE)

The considerations in the preceding two subsections motivate the following definition.

1.3.8 Definition. *Assume that we have a DG $\mathcal{G}$ and a collection of transition probabilities $p(\mathbf{x}_v\|\mathbf{x}_{\mathrm{pa}(v)})$*[5]*. We say that $\mathbf{X} = (\mathbf{X}_v : v \in \mathcal{V})$ is a* ***composable random element (CRE)*** *if its probability density p exists, satisfies (1.3.2):*

$$p(\mathbf{x}) = \prod_{v\in\mathcal{V}} p(\mathbf{x}_v\|\mathbf{x}_{\mathrm{pa}(v)}) \tag{1.3.2}$$

and for every $\mathcal{A} \subset \mathcal{V}$, every $\mathbf{x}_{\mathcal{V}\setminus\mathcal{A}}$ we have

$$\int_{\mathcal{X}_{\mathcal{A}}} \prod_{v\in\mathcal{A}} p(\mathbf{x}_v\|\mathbf{x}_{\mathrm{pa}(v)})\mathrm{d}\mathbf{x}_{\mathcal{A}} = 1$$

[5] For every fixed $\mathbf{x}_{\mathrm{pa}(v)}$, $p(\cdot\|\mathbf{x}_{\mathrm{pa}(v)})$ is a probability density ie. $\int p(\mathbf{x}_v\|\mathbf{x}_{\mathrm{pa}(v)})\mathrm{d}\mathbf{x}_v = 1$; plus measurablity requirements.

This definition is rather abstract but it covers at least three classes of models:

- If $\mathcal{G}$ is a DAG and $\mathbf{X}$ fulfils the factorisation condition (1.2.1) wrt. $\mathcal{G}$ then $\mathbf{X}$ is a CRE (with $p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) = p(\mathbf{x}_v | \mathbf{x}_{\mathrm{pa}(v)})$).
- If $\mathcal{G}$ is a general DG and $\mathbf{X}$ is a time series which fulfils the conditions (1.3.1) wrt. $\mathcal{G}$ then $\mathbf{X}$ is a CRE.
- If If $\mathcal{G}$ is a general DG and $\mathbf{X}$ is a Continuous Time Bayesian Network (CTBN) wrt. $\mathcal{G}$ then $\mathbf{X}$ is a CRE.

Let us underline that we regard transition probabilities $p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)})$ as elementary building blocks of a causal model. It turns out (see the next section) that many classical results for models based on DAGs remain valid for CREs with general DGs).

Conditioning-by-intervention

Let $\mathbf{X} = (\mathbf{X}_v : v \in \mathcal{V})$ be a CRE (Definition 1.3.8) wrt. a possibly cyclic graph.

1.3.9 Definition. *We will define conditional-by-intervention distribution in two steps.*

- *Assume that $\mathcal{A} \subseteq \mathcal{V}$ and $\mathcal{C} = \mathcal{V} \setminus \mathcal{A}$. Then the distribution $\mathbf{X}_\mathcal{A} \| \mathbf{X}_\mathcal{C}$ has the density*

$$p(\mathbf{x}_\mathcal{A} \| \mathbf{x}_\mathcal{C}) = \prod_{v \in \mathcal{A}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)})$$

- *For arbitrary disjoint subsets $\mathcal{A}$ and $\mathcal{C}$ of $\mathcal{V}$ we let*

$$p(\mathbf{x}_\mathcal{A} \| \mathbf{x}_\mathcal{C}) = \int\limits_{\mathcal{X}_{\mathcal{V} \setminus (\mathcal{A} \cup \mathcal{C})}} p(\mathbf{x}_{\mathcal{V} \setminus \mathcal{C}} \| \mathbf{x}_\mathcal{C}) \mathrm{d}\mathbf{x}_{\mathcal{V} \setminus (\mathcal{A} \cup \mathcal{C})}$$

$p(\mathbf{x}_\mathcal{A} \| \mathbf{x}_\mathcal{C})$ is interpreted as the probability density of $\mathbf{X}_\mathcal{A}$ given that $\mathbf{X}_\mathcal{C}$ is forced to assume value $\mathbf{x}_\mathcal{C}$[6]. Of course, $p(\mathbf{x}_\mathcal{A} \| \mathbf{x}_\mathcal{C})$ is in general different from the standard, conditional-by-observation probability density $p(\mathbf{x}_\mathcal{A} | \mathbf{x}_\mathcal{C})$ defined by

$$p(\mathbf{x}_\mathcal{A} | \mathbf{x}_\mathcal{C}) = \frac{p(\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C})}{p(\mathbf{x}_\mathcal{C})}$$

We explain, using Example 1.3.3, two ways of presenting the intuitive sense of Definition 1.3.9.

[6] Notation $p(\mathbf{x}_\mathcal{A} \| \mathbf{x}_\mathcal{C})$ is used instead of Pearlian $p(\mathbf{x}_\mathcal{A} | \mathrm{do}(\mathbf{x}_\mathcal{C}))$ or $p(\mathbf{x}_\mathcal{A} | \hat{\mathbf{x}}_\mathcal{C})$.

- Conditioning-by-intervention is dropping arrows leading to nodes intervened on.
- Conditioning-by-intervention is introducing additional ‘controller nodes’ which overrule transition probabilities.

Example: Consider the graph in Example 1.3.3:

$$\begin{tikzcd}\mathbf{X} \arrow[r, bend right] & \mathbf{Y} \arrow[l, bend right] & \mathbf{Z} \arrow[l]\end{tikzcd}$$

which corresponds to the join probability $p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x}\|\mathbf{y})p(\mathbf{y}\|\mathbf{x},\mathbf{z})p(\mathbf{z})$. According to Definition 1.3.9, the conditional-by-intervention distribution of $\mathbf{Y},\mathbf{Z}\|\mathbf{X}=\mathbf{x}$ is

$$p(\mathbf{y},\mathbf{z}\|\mathbf{x}) = p(\mathbf{y}\|\mathbf{x},\mathbf{z})p(\mathbf{z})$$

Notice that this corresponds to the graph with removed the arrow to $\mathbf{X}$ and setting $\mathbf{X}=\mathbf{x}$:

$$\begin{tikzcd}\mathbf{x} \arrow[r, bend right] & \mathbf{Y} & \mathbf{Z} \arrow[l]\end{tikzcd}$$

Alternative way is the following. We can imagine a ficticious ‘controller node’, say $\boldsymbol{\Xi}$, with set of states: $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{idle}\}$ and augmented graph

$$\begin{tikzcd}\boldsymbol{\Xi} \arrow[r] & \mathbf{X} \arrow[r, bend right] & \mathbf{Y} \arrow[l, bend right] & \mathbf{Z} \arrow[l]\end{tikzcd}$$

We now replace $p(\mathbf{x}\|\mathbf{y})$ by $p(\mathbf{x}\|\mathbf{y},\boldsymbol{\xi})$, where

$$p(\mathbf{x}\|\mathbf{y},\boldsymbol{\xi}) = \begin{cases} 1 & \text{if } \mathbf{x}=\boldsymbol{\xi} \text{ and } \boldsymbol{\xi}\neq \mathbf{idle} \\ 0 & \text{if } \mathbf{x}\neq\boldsymbol{\xi} \text{ and } \boldsymbol{\xi}\neq \mathbf{idle} \\ p(\mathbf{x}\|\mathbf{y}) & \text{if } \boldsymbol{\xi}=\mathbf{idle} \end{cases}$$

1.4 Characterisation of conditional independences

Graph-theoretical notions

Let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be a directed graph (DG) **with possible cycles** but without self-loops.

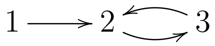
A trail is a sequence of nodes connected by arrows together with the information on the *direction of arrows.* Formally, we define a **trail** between $u\in\mathcal{V}$ and $w\in\mathcal{V}$ as a sequence

$$u=v_0,e_1,v_1,e_2,v_2,e_3,\ldots,e_{k-1},v_{k-1},e_k,v_k=w$$

where $v_0,v_1,\ldots,v_k$ are *distinct* nodes and $e_1,\ldots,e_k$ are edges, $e_i=(v_{i-1}\rightarrow v_i)$ or $e_i=(v_{i-1}\leftarrow v_i)$[7].

[7] Definition of a *trail* requires some caution, because there are subtle differences between several definitions appearing in the literature.

Example: Consider the following graph:



We have two distinct trails between 1 and 3. Trail $1 \to 2 \to 3$ is different from $1 \to 2 \leftarrow 3$.

Remark. Every edge of a trail must have a *uniquely chosen* direction! Every node can appear in a trail *at most once*!

Let v_i be a non-end node in the trail $v_0, e_1, v_1, e_2, v_2, e_3, \ldots, e_{k-1}, v_{k-1}, e_k, v_k$, that is $i \neq 0$ and $i \neq k$. We say that

> There is a *chain* connexion at v_i if we have $v_{i-1} \to v_i \to v_{i+1}$ or $v_{i-1} \leftarrow v_i \leftarrow v_{i+1}$.
>
> There is a *fork* connexion at v_i if we have $v_{i-1} \leftarrow v_i \to v_{i+1}$.
>
> There is a *collider* connexion at v_i if we have $v_{i-1} \to v_i \leftarrow v_{i+1}$.

A **directed path** from u to w is a trail such that all arrows are directed to the right, i.e. $e_i = (v_{i-1} \to v_i)$. Formally we define the set of *ancestors* and the set of *descendants* as follows:

$$\begin{aligned}\mathrm{an}(v) &= \{v\} \cup \{w : \text{ there exists a directed path from } w \text{ to } v\},\\ \mathrm{de}(v) &= \{v\} \cup \{w : \text{ there exists a directed path from } v \text{ to } w\}.\end{aligned}$$

Moreover, for $\mathcal{A} \subseteq \mathcal{V}$ we put $\mathrm{an}(\mathcal{A}) = \bigcup_{v\in\mathcal{A}} \mathrm{an}(v)$, $\mathrm{de}(\mathcal{A}) = \bigcup_{v\in\mathcal{A}} \mathrm{de}(v)$, $\mathrm{pa}(\mathcal{A}) = \bigcup_{v\in\mathcal{A}} \mathrm{pa}(v)$.

Below $\mathcal{A}, \mathcal{B}, \mathcal{C}$ stand for any three disjoint subsets of $\mathcal{V}$ with $\mathcal{C}$ possible empty. A trail (directed path) from $\mathcal{A}$ to $\mathcal{B}$ is a trail (directed path) from an $a \in \mathcal{A}$ to a $b \in \mathcal{B}$.

Interventional (conditional) independence: u-separation

Our definition of *causal conditional independence* is the following. Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be three disjoint subsets of $\mathcal{V}$, with $\mathcal{C}$ possible empty. We say that $\mathbf{X}_\mathcal{B}$ is **causally independent of $\mathbf{X}_\mathcal{A}$ imposing $\mathbf{X}_\mathcal{C}$**, symbolically $\mathbf{X}_\mathcal{A} \not\triangleright \mathbf{X}_\mathcal{B} \| \mathbf{X}_\mathcal{C}$, if

$$p(\mathbf{x}_\mathcal{B} \| \mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_\mathcal{B} \| \mathbf{x}_\mathcal{C}) \tag{1.4.1}$$

where $(\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = \mathbf{x}_{\mathcal{A}\cup\mathcal{C}}$ and the symbol $p(\cdot\|\cdot)$ on both sides of this equation denotes conditional-by-intervention probability defined in the previous section. Symbol $\triangleright$ denotes the negation of $\not\triangleright$. Formula (1.4.1) is an exact counterpart of the analogous property of observational independence: relation $\mathbf{X}_\mathcal{A} \perp\!\!\!\perp \mathbf{X}_\mathcal{B} | \mathbf{X}_\mathcal{C}$ is equivalent to

$$p(\mathbf{x}_\mathcal{B} | \mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_\mathcal{B} | \mathbf{x}_\mathcal{C})$$

Our definition of causal conditional independence via (1.4.1) seems to be very natural and to properly reflect intuitive meaning of this notion. Another definition appearing in the literature (due to Ay and Polani [1, Equation 6]) is the following. They say that "$\mathbf{X}_{\mathcal{B}}$ is causally independent of $\mathbf{X}_{\mathcal{A}}$ imposing $\mathbf{X}_{\mathcal{C}}$" if

$$p(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}}) = \int p(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}'_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}}) p(\mathbf{x}'_{\mathcal{A}} \| \mathbf{x}_{\mathcal{C}}) \mathrm{d}\mathbf{x}'_{\mathcal{A}}. \tag{1.4.2}$$

Clearly (1.4.1) implies (1.4.2). The converse is not true, as illustrated by the following example.

1.4.3 EXAMPLE. Consider four binary random variables $\mathbf{X}_a, \mathbf{X}_b, \mathbf{X}_c, \mathbf{X}_u$ and the following graph of causal relations:

$$\begin{tikzcd}
\mathbf{X}_u \arrow[r] \arrow[d] \arrow[dr] & \mathbf{X}_c \\
\mathbf{X}_a \arrow[r] & \mathbf{X}_b
\end{tikzcd}$$

Let $\mathbb{P}(\mathbf{X}_u = 0) = \mathbb{P}(\mathbf{X}_u = 1) = 1/2$, $\mathbf{X}_a = 1 - \mathbf{X}_u$ and $\mathbf{X}_b = \mathbb{1}(\mathbf{X}_a = \mathbf{X}_u)$. The values of $\mathbb{P}(\mathbf{X}_c \| \mathbf{X}_u)$ are irrelevant (as, in fact is the very presence of node c). Put $\mathcal{A} = \{a\}$, $\mathcal{B} = \{b\}$, $\mathcal{C} = \{c\}$. Clearly we have

$$\mathbb{P}(\mathbf{X}_b = 1 \| \mathbf{X}_a = 0, \mathbf{X}_c) = \mathbb{P}(\mathbf{X}_b = 1 \| \mathbf{X}_a = 1, \mathbf{X}_c) = 1/2$$

so (1.4.2) is true. On the other hand,

$$\mathbb{P}(\mathbf{X}_b = 1 \| \mathbf{X}_c) = 0$$

so (1.4.1) is false.

The difference between (1.4.2) and(1.4.1) is by no means technical and pertains to the fundamental question "what we mean by causality?". Imagine the following story standing behind our example. There are two raiload tracks (marked '1' and '0') between Undover and Andover. A train from Undover to Andover departs just a minute before a train from Andover to Undover. $\mathbf{X}_u = 1$ means that the train from Undover chooses track '1', and $\mathbf{X}_u = 0$ if it chooses track '0'. Analogously, $\mathbf{X}_a$ denotes the choice of the track by the train from Andover. Now, $\mathbf{X}_b = 1$ denotes the event of crash (both trains have chosen the same track). In the "observational regime" we have $\mathbb{P}(\mathbf{X}_b = 1) = 0$, because Undover can let Andover know which track is free. On the other hand, "conditioning-by-intervention" $\mathbb{P}(\cdot \| X_a)$ actually blocks the flow of information from Undover to Andover and thus causes a railroad crash (with probability 1/2). Note that the choice of imposed value $\mathbf{x}_a$ is unimportant but the mere *fact of imposing* a value is important. We claim that, intuitively, we should say that $\mathbf{X}_a$ *does* have causal effect on $\mathbf{X}_b$ (imposing $\mathbf{X}_c$). Our definition is consistent with the intuitive sense of causal independence whereas the definition proposed by Ay and Polani is not. △

Characterisation of causal conditional independence in terms of the underlying graph will be given in Theorem 1.4.4 and requires the notion of u-separation.

Let $\mathcal{G}$ be a (possibly cyclic) graph. We say that $\mathcal{B}$ is **u-separated** (unidirectionally separated) from $\mathcal{A}$ by $\mathcal{C}$ if every directed path from $\mathcal{A}$ to $\mathcal{B}$ has a node belonging to $\mathcal{C}$. We will write $\mathcal{A} \not\longrightarrow_u \mathcal{B}|\mathcal{C}$ (and $\mathcal{A} \longrightarrow_u \mathcal{B}|\mathcal{C}$ if u-separation does not hold).

1.4.4 Theorem (**u-separation**)**.** *Let* $\mathbf{X}$ *be a composable random element (CRE) wrt.* $\mathcal{G}$*. If* $\mathcal{B}$ *is u-separated from* $\mathcal{A}$ *by* $\mathcal{C}$ *then* $\mathbf{X}_\mathcal{B}$ *is causally independent of* $\mathbf{X}_\mathcal{A}$ *imposing* $\mathbf{X}_\mathcal{C}$*:*

$$\mathcal{A} \not\longrightarrow_u \mathcal{B}|\mathcal{C} \textit{ implies } \mathbf{X}_\mathcal{A} \not\triangleright \mathbf{X}_\mathcal{B}\|\mathbf{X}_\mathcal{C}$$

Proof. Let

$$\bar{\mathcal{A}} = \{v : \mathcal{A} \longrightarrow_u v|\mathcal{C}\}$$
$$\bar{\mathcal{B}} = \mathcal{V} \setminus (\bar{\mathcal{A}} \cup \mathcal{C})$$

Obviously, $\bar{\mathcal{B}} \supseteq \mathcal{B}$ and there are no arrows from $\bar{\mathcal{A}}$ to $\bar{\mathcal{B}}$. Consequently, $\prod_{v\in\bar{\mathcal{B}}} p(\mathbf{x}_v\|\mathbf{x}_{\mathrm{pa}(v)})$ does not depend on $\mathbf{x}_{\bar{\mathcal{A}}}$.

We are to show (1.4.1), i.e.

$$p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{C}). \tag{1.4.1}$$

To lighten notation we write $p_v = p(\mathbf{x}_v\|\mathbf{x}_{\mathrm{pa}(v)})$, and omit '$\mathrm{d}\mathbf{x}_\mathcal{W}$' in the integrals $\int_{\mathcal{X}_\mathcal{W}} \cdots \mathrm{d}\mathbf{x}_\mathcal{W}$. In accordance with our Definition 1.3.9, the RHS of (1.4.1) is given by

$$\begin{aligned}\text{RHS} = \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \int_{\mathcal{X}_{\bar{\mathcal{A}}}} \prod_{v\in\bar{\mathcal{B}}} p_v \prod_{v\in\bar{\mathcal{A}}} p_v &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v \int_{\mathcal{X}_{\bar{\mathcal{A}}}} \prod_{v\in\bar{\mathcal{A}}} p_v \\ &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v\end{aligned}$$

because $\prod_{v\in\bar{\mathcal{B}}} p_v$ does not depend on $\mathbf{x}_{\bar{\mathcal{A}}}$ and $\int_{\mathcal{X}_{\bar{\mathcal{A}}}} \prod_{v\in\bar{\mathcal{A}}} p_v = 1$ in view of the definition of CRE. Similarly we compute the LHS of (1.4.1):

$$\begin{aligned}\text{LHS} = \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \int_{\mathcal{X}_{\bar{\mathcal{A}}\setminus\mathcal{A}}} \prod_{v\in\bar{\mathcal{B}}} p_v \prod_{v\in\bar{\mathcal{A}}\setminus\mathcal{A}} p_v &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v \int_{\mathcal{X}_{\bar{\mathcal{A}}\setminus\mathcal{A}}} \prod_{v\in\bar{\mathcal{A}}\setminus\mathcal{A}} p_v \\ &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v\end{aligned}$$

We obtain that RHS=LHS and this completes the proof. □

Ay and Polani [1] proved a result analogous to Theorem 1.4.4 under the assumption that $\mathcal{G}$ is acyclic and they used (1.4.2) as definition of causal independence. Our theorem provides two improvements: generalisation to possibly cyclic graphs and using stronger condition (1.4.1) as definition of causal independence.

The following proposition shows that Theorem 1.4.4 gives in some sense full characterisation of causal conditional independence.

1.4.5 Proposition. *If $\mathcal{B}$ is not u-separated from $\mathcal{A}$ by $\mathcal{C}$ in a directed graph $\mathcal{G}$ then there exists a probability density $p(\mathbf{x})$ which is a CRE with respect to $\mathcal{G}$ such that $\mathbf{X}_{\mathcal{B}}$ is not causally independent of $\mathbf{X}_{\mathcal{A}}$ imposing $\mathbf{X}_{\mathcal{C}}$:*

$$\mathcal{A} \longrightarrow_u \mathcal{B}|\mathcal{C} \textit{ implies } \mathbf{X}_{\mathcal{A}} \triangleright \mathbf{X}_{\mathcal{B}} \| \mathbf{X}_{\mathcal{C}} \textit{ for some } \mathbf{X} \textit{ which is a CRE wrt. } \mathcal{G}.$$

Proof. Let $\tau = (a = v_0 \to v_1, \to \ldots \to v_{n-1} \to v_n = b)$ be a directed path with $a \in \mathcal{A}$, $b \in \mathcal{B}$ which does not intersect $\mathcal{C}$. If we put $\mathcal{E}_\tau = \mathcal{E} \cap \tau$ then $(\mathcal{V}, \mathcal{E}_\tau)$ is a DAG and any probability density which factorises along this DAG defines a CRE wrt. $\mathcal{G}$. Such a density makes all nodes in $\mathcal{V} \setminus \tau$ mutually independent and independent of $\mathcal{V} \cap \tau$, so $p(\mathbf{x}_b \| \mathbf{x}_a, \mathbf{x}_{\mathcal{C}}) = p(\mathbf{x}_b | \mathbf{x}_a)$ and $p(\mathbf{x}_b \| \mathbf{x}_{\mathcal{C}}) = p(\mathbf{x}_b)$. It is therefore enough to choose a density on nodes belonging to τ such that $\mathbf{X}_a$ and $\mathbf{X}_b$ are dependent variables. □

Observational (conditional) independence: d-separation

The concept of d-separation was defined by Judea Pearl for DAGs (see the monograph [8]). The same definition works for possibly cyclic DGs. We say that $\mathcal{B}$ is **d-separated** from $\mathcal{A}$ by $\mathcal{C}$ if every trail from $\mathcal{A}$ to $\mathcal{B}$ contains

- a chain $\leftarrow c \leftarrow$ or a chain $\to c \to$ with $c \in \mathcal{C}$
- or a fork $\leftarrow c \to$ with $c \in \mathcal{C}$
- or a collider $\to v \leftarrow$ with $v \in \mathrm{an}(\mathcal{C})$.

We will then write $\mathcal{A} \perp_d \mathcal{B}|\mathcal{C}$ (and $\mathcal{A} \longrightarrow_d \mathcal{B}|\mathcal{C}$ if d-separability does not hold).

Note that the relation of d-separation is symmetric (with respect to $\mathcal{A}$ and $\mathcal{B}$) whereas u-separation is clearly not.

Pearl proved (in an article published in 1985) that d-separation characterises conditional independence for models governed by DAGs. This remains true for CREs governed by DGs.

1.4.6 Theorem (**d-separation**). *Let $\mathbf{X}$ be a composable random element (CRE) wrt. $\mathcal{G}$. If $\mathcal{B}$ is d-separated from $\mathcal{A}$ by $\mathcal{C}$ then $\mathbf{X}_{\mathcal{B}}$ is conditionally independent of $\mathbf{X}_{\mathcal{A}}$ given $\mathbf{X}_{\mathcal{C}}$:*

$$\mathcal{A} \perp_d \mathcal{B}|\mathcal{C} \textit{ implies } \mathbf{X}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{X}_{\mathcal{B}}|\mathbf{X}_{\mathcal{C}}.$$

We will present a very simple proof based on two facts[8]. Firstly, we use alternative equivalent characterisation of d-separation, Theorem 1.4.7. Secondly, we use characterisation of conditional independence for undirected graphical models, Proposition 1.4.8. Note that the "auxilliary" results: 1.4.7 and 1.4.8 are in fact important theorems of independent interest.

We shall begin with a few necessary definitions. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph with possible cycles. If The *moralised* graph is an *undirected* graph $\mathcal{G}^m = (\mathcal{V}, \mathcal{E}^m)$, where

$$\mathcal{E}^m = \{\{v, w\} : v \to w \text{ or } v \leftarrow w \text{ or there is } u \text{ such that } v \to u \text{ and } w \to u\}$$

If $(\mathcal{W}, \mathcal{K})$ is an undirected graph then a trail (undirected) between $v \in \mathcal{W}$ and $w \in \mathcal{W}$ is a sequence $v = u_0, u_1, \dots, u_k = w$ such that $\{u_{i-1}, u_i\} \in \mathcal{K}$ for $i = 1, \dots, k$. We say that $\mathcal{A}$ and $\mathcal{B}$ are **separated** (adjectivelessly separated) by $\mathcal{C}$ if every trail between $\mathcal{A}$ and $\mathcal{B}$ has a node belonging to $\mathcal{C}$ ($\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are disjoint subsets od $\mathcal{W}$, with $\mathcal{C}$ possibly empty).

1.4.7 Theorem (**equivalence**). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph and let $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{V}$. Then $\mathcal{A}$ and $\mathcal{B}$ are d-separated by $\mathcal{C}$ in $\mathcal{G}$ if and only if $\mathcal{A}$ and $\mathcal{B}$ are separated by $\mathcal{C}$ in $\mathcal{G}^m_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$.*

The proof is in Appendix B. It is based on [7].

The next result is an easy part of the celebrated Hammmersley-Clifford theorem [6].

1.4.8 Proposition. *Let $(\mathcal{W}, \mathcal{K})$ be an undirected graph and assume that the probability density of $\mathbf{X} = \mathbf{X}_{\mathcal{W}}$ factorises over its cliques:*

$$p(\mathbf{x}) = \prod_{\mathcal{C} \textit{ is a clique}} \phi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}}).$$

Consider disjoint subsets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ of $\mathcal{W}$. If $\mathcal{A}$ and $\mathcal{B}$ are separated by $\mathcal{C}$ then $\mathbf{X}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{X}_{\mathcal{B}}|\mathbf{X}_{\mathcal{C}}$.

Proof of Proposition 1.4.8. Let

$$\bar{\mathcal{A}} = \mathcal{A} \cup \{v : v \text{ — } \mathcal{A}\}, \quad \bar{\mathcal{B}} = \mathcal{A} \cup \{v : v \text{ — } \mathcal{B}\},$$

where ' — ' means that there is a connecting trail not passing through $\mathcal{C}$, and

$$\mathcal{D} = \mathcal{W} \setminus (\bar{\mathcal{A}} \cup \bar{\mathcal{B}} \cup \mathcal{C}).$$

[8] An independent but more complicated proof is given in [10, Appendix A]. The idea of this independent proof is taken from John Noble.

Clearly we cannot have neither $\bar{\mathcal{A}} — \bar{\mathcal{B}}$ nor $\bar{\mathcal{A}} — \mathcal{D}$ nor $\bar{\mathcal{B}} — \mathcal{D}$. Thus every clique $\mathcal{C}$ is a subset of either $\bar{\mathcal{A}} \cup \mathcal{C}$ or $\bar{\mathcal{B}} \cup \mathcal{C}$ or $\mathcal{D} \cup \mathcal{C}$. Consequently the density factorises as

$$p(\mathbf{x}) = \psi_1(\mathbf{x}_{\bar{\mathcal{A}}\cup\mathcal{C}})\psi_2(\mathbf{x}_{\bar{\mathcal{B}}\cup\mathcal{C}})\psi_3(\mathbf{x}_{\mathcal{D}\cup\mathcal{C}}).$$

It follows that $\mathbf{X}_{\bar{\mathcal{A}}} \perp\!\!\!\perp \mathbf{X}_{\bar{\mathcal{B}}} | \mathbf{X}_{\mathcal{C}}$. □

Proof of Theorem 1.4.6. Assume that $\mathbf{X}$ is a CRE, ie. $p(\mathbf{x})$ satisfies conditions 1.3.8. We are to prove the following implication:

$$\mathcal{A} \perp_d \mathcal{B}|\mathcal{C} \text{ implies } \mathbf{X}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{X}_{\mathcal{B}} | \mathbf{X}_{\mathcal{C}}.$$

Let $\mathcal{W} = \mathrm{an}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C})$. By 1.3.8 we have

$$\begin{aligned} p(\mathbf{x}_{\mathcal{W}}) &= \int_{\mathcal{X}_{\mathcal{V}\setminus\mathcal{W}}} \prod_{v\in\mathcal{W}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) \prod_{v\in\mathcal{V}\setminus\mathcal{W}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) \mathrm{d}\mathbf{x}_{\mathcal{V}\setminus\mathcal{W}} \\ &= \prod_{v\in\mathcal{W}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) \int_{\mathcal{X}_{\mathcal{V}\setminus\mathcal{W}}} \prod_{v\in\mathcal{V}\setminus\mathcal{W}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}) \mathrm{d}\mathbf{x}_{\mathcal{V}\setminus\mathcal{W}} \\ &= \prod_{v\in\mathcal{W}} p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)}), \end{aligned} \tag{1.4.9}$$

because $\mathcal{W}$ is an ancestral set and thus $p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)})$ does not depend on $\mathbf{x}_{\mathcal{V}\setminus\mathcal{W}}$ for $v \in \mathcal{W}$.

Consider the undirected moral graph $\mathcal{G}^m_{\mathcal{W}}$. Every factor $p(\mathbf{x}_v \| \mathbf{x}_{\mathrm{pa}(v)})$ in (1.4.9) depends on $\mathrm{pa}(v) \cup \{v\}$ which is a *clique* in $\mathcal{G}^m_{\mathcal{W}}$. Therefore the density of $\mathbf{X}_{\mathcal{W}}$ factorises over cliques with respect to $\mathcal{G}^m_{\mathcal{W}}$. By Theorem 1.4.7 we know that $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$ in $\mathcal{G}^m_{\mathcal{W}}$. From Proposition 1.4.8 we infer that $\mathbf{X}_{\mathcal{A}} \perp\!\!\!\perp \mathbf{X}_{\mathcal{B}} | \mathbf{X}_{\mathcal{C}}$. □

Remark. The important theorem published by Hammersley and Clifford in 1971 was really the beginning of the theory of probabilistic graphical models. Theorem 1.4.7 was proved in an article published in 1990 by Stephen Lauritzen et al. in the setting of DAGs.

Independence properties related to time: δ- and ε-separation

In this subsection we tackle independences between the past of one subprocess and the future of another given the past of the third subprocess. More precisely, let $\mathbf{X} = (X_v(t), v \in \mathcal{V})$ be a multivariate stochastic process. We want to find necessary and sufficient conditions for the following independence relation:

$$(X_{\mathcal{B}}(s), s > t) \perp\!\!\!\perp (X_{\mathcal{A}}(s), s \leq t) | (X_{\mathcal{B}\cup\mathcal{C}}(s), s \leq t) \tag{1.4.10}$$

This is a very natural problem clearly related to the task of prediction. If we want to predict the *future* of $\mathbf{X}_{\mathcal{B}}$ knowing the *past* of $\mathbf{X}_{\mathcal{B}\cup\mathcal{C}}$ then relation (1.4.10) tells us that additional information on the past of $\mathbf{X}_{\mathcal{A}}$ is redundant and can be discarded.

To characterise relation (1.4.10) in graph-theoretic terms, we need the following definition.

We say that $\mathcal{B}$ is **ε-separated** from $\mathcal{A}$ by $\mathcal{C}$ if every trail from $\mathcal{A}$ to $\mathcal{B}$ which ends with an arrow $\to b \in \mathcal{B}$ and has no other nodes in $\mathcal{B}$, satisfies at least one of the following conditions: it contains a chain $\leftarrow c \leftarrow$ or a fork $\leftarrow c \to$ with $c \in \mathcal{C}$ or a collider $\to v \leftarrow$ such that $v \notin \mathrm{an}(\mathcal{C})$, or contains a chain $\to c \to$ with $c \in \mathcal{C}$ that occurs earlier than some collider. We will then write $\mathcal{A} \not\rightharpoondown_{\varepsilon} \mathcal{B}|\mathcal{C}$. Negation of this statement is denoted $\mathcal{A} \rightharpoondown_{\varepsilon} \mathcal{B}|\mathcal{C}$.

1.4.11 Theorem (**ε-separation**). *Let $\mathbf{X} = (X_v(t), t = 0, 1, \ldots, u)$ be a discrete time process satisfying (1.3.1) or a Continuous Time Bayesian Network with respect to graph $\mathcal{G}$. If $\mathcal{B}$ is ε-separated from $\mathcal{A}$ by $\mathcal{C}$ then (1.4.10) holds for every $t < u$:*

$$\mathcal{A} \not\rightharpoondown_{\varepsilon} \mathcal{B}|\mathcal{C} \text{ implies } (X_{\mathcal{B}}(s), s > t) \perp\!\!\!\perp (X_{\mathcal{A}}(s), s \leq t) | (X_{\mathcal{B}\cup\mathcal{C}}(s), s \leq t).$$

Proof. Let

$$\mathcal{D} = \mathrm{an}(\mathcal{B}) \setminus (\mathcal{B} \cup \mathcal{C}),$$
$$\mathcal{R} = \mathcal{V} \setminus (\mathcal{A} \cup \mathrm{an}(\mathcal{B}) \cup \mathcal{C}).$$

In the sequel we assume that $\mathcal{D}$ and $\mathcal{R}$ are nonempty. If $\mathcal{D} = \emptyset$ or $\mathcal{R} = \emptyset$ then the reasoning is similar but simpler, and therefore omitted.

We will use the following notation. For a fixed $t = 0, 1, \ldots, u - 1)$ let

- $\mathbf{A} = (X_{\mathcal{A}}(s), 0 \leq s \leq t)$ – the past of $\mathbf{X}_{\mathcal{A}}$,
- $\mathbf{B} = (X_{\mathcal{B}}(s), 0 \leq s \leq t)$ – the past of $\mathbf{X}_{\mathcal{B}}$,
- $\mathbf{C} = (X_{\mathcal{C}}(s), 0 \leq s \leq t)$ – the past of $\mathbf{X}_{\mathcal{C}}$,
- $\mathbf{D} = (X_{\mathcal{D}}(s), 0 \leq s \leq t)$ – the past of $\mathbf{X}_{\mathcal{D}}$,
- $\mathbf{R} = (X_{\mathcal{R}}(s), 0 \leq s \leq t)$ – the past of $\mathbf{X}_{\mathcal{R}}$,
- $\mathbf{F} = (X_{\mathcal{B}}(s), t < s \leq n)$ – the future of $\mathbf{X}_{\mathcal{B}}$,

with $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}$ and $\mathbf{f}$ denoting the values of these random elements. In these notations the conclusion which we are to prove is the following:

$$p(\mathbf{f}|\mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{f}|\mathbf{b}, \mathbf{c}).$$

We first show that $\mathcal{A} \not\rightharpoondown_\varepsilon \mathcal{B}|\mathcal{C}$ implies

$$\mathcal{D} \perp_d \mathcal{A}|\mathcal{B} \cup \mathcal{C}.$$

Indeed, suppose that there exists a trail τ from $a \in \mathcal{A}$ to $d \in \mathcal{D}$ which is not d-separated by $\mathcal{B} \cup \mathcal{C}$. Let $x \leftarrow y$ be the first entry of τ to $\mathrm{an}(\mathcal{B})$ (the arrow must be directed to a because otherwise x would belong to $\mathrm{an}(\mathcal{B})$). If we had $y \in \mathcal{B} \cup \mathcal{C}$ then τ would be d-separated because y is not a collider. If we had $y \in \mathcal{D}$ then we would have $\mathcal{A} \rightharpoondown_\varepsilon \mathcal{B}|\mathcal{C}$. To see this, note that $y \in \mathrm{an}(\mathcal{B}) \setminus \mathcal{B}$ implies that there exists $z \in \mathrm{an}(\mathcal{B})$ such that $y \to z$. Let τ_y be the part of τ from a to y. Clearly, τ_y contains neither $\leftarrow v \leftarrow$ nor $\to v \to$ nor $\leftarrow v \to$ with $v \in \mathcal{B} \cup \mathcal{C}$. If τ_y contains $\to v \leftarrow$ then $v \in \mathrm{an}(\mathcal{B} \cup \mathcal{C})$. Since y is the first entry to $\mathrm{an}(\mathcal{B})$, it follows that $v \in \mathrm{an}(\mathcal{C})$ and thus the collider at v does not ε-block the trail from a to y to z to some $b \in \mathcal{B}$ (by $\mathcal{C}$).

We have obtained contradiction and therefore $\mathcal{D} \perp_d \mathcal{A}|\mathcal{B} \cup \mathcal{C}$. By Theorem 1.4.6, applied to $(X_v(s), 0 \le s \le t)$, this graph separation statement entails conditional independence $\mathbf{D} \perp\!\!\!\perp \mathbf{A}|\mathbf{B}, \mathbf{C}$ and thus

$$p(\mathbf{d}|\mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{d}|\mathbf{b}, \mathbf{c}). \tag{1.4.12}$$

The next step is to show that

$$p(\mathbf{f}|\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = p(\mathbf{f}|\mathbf{b}, \mathbf{c}, \mathbf{d}). \tag{1.4.13}$$

Equation (1.4.13) follows from $\mathcal{B} \cup \mathcal{C} \cup \mathcal{D} \supseteq \mathrm{an}(\mathcal{B})$. Indeed, since $\mathrm{an}(\mathcal{B})$ is an ancestral set, $\mathbf{X}_{\mathrm{an}(\mathcal{B})}$ is a process which evolves independently of the rest of nodes and $(X_{\mathrm{an}(\mathcal{B})}(s), s > t)$ depends on $(X_{\mathcal{V}}(s), s \le t)$ only through $(X_{\mathrm{an}(\mathcal{B})}(s), s \le t)$. Therefore $\mathbf{F}$ is conditionally independent of $\mathbf{A}$ and $\mathbf{R}$ given $(\mathbf{B}, \mathbf{C}, \mathbf{D})$. It follows that $p(\mathbf{f}|\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = p(\mathbf{f}|\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}) = p(\mathbf{f}|\mathbf{b}, \mathbf{c}, \mathbf{d})$ and we obtain (1.4.13).

Finally, to obtain the desired conclusion we combine (1.4.12) with (1.4.13):

$$\begin{aligned} p(\mathbf{f}|\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \int p(\mathbf{f}|\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) p(\mathbf{d}|\mathbf{a}, \mathbf{b}, \mathbf{c}) \mathrm{d}\mathbf{d} \\ &= \int p(\mathbf{f}|\mathbf{b}, \mathbf{c}, \mathbf{d}) p(\mathbf{d}|\mathbf{b}, \mathbf{c}) \mathrm{d}\mathbf{d} = p(\mathbf{f}|\mathbf{b}, \mathbf{c}) \end{aligned}$$

which completes the proof. □

Remark. In the case of discrete time, a different proof of Theorem 1.4.11 is based on the representation of "space-time" graph (mentioned in subsection on discrete time processes, Section 1.3). One just checks that ε-separation $\mathcal{A} \rightharpoondown_\varepsilon \mathcal{B}|\mathcal{C}$ in graph $\mathcal{G}$ implies d-separation $\mathcal{A}(\le t) \perp_d \mathcal{B}(> t)|\mathcal{C}(\le t)$ in the "space-time" graph. We omit details.

The following proposition is a sort of the converse statement.

1.4.14 Proposition. *If $\mathcal{B}$ is not ε-separated from $\mathcal{A}$ by $\mathcal{C}$ then there exists a a discrete time process $\mathbf{X} = (X_v(t), t = 0, 1, \ldots, n)$ satisfying (1.3.1) such that (1.4.10) does not hold.*

The proof of this proposition given in [10, Theorem 3.9(b)] is based on analysing the "space-time" graph and finding a d-open trail in it.

Analogously to (1.4.10) we also consider a conditional independence relation in which *future* is replaced by "*immediate future*". First we focus on discrete time processes. For such a process the immediate future of t is represented by $t+1$. Therefore we we ask when it is true that

$$X_{\mathcal{B}}(t+1) \perp\!\!\!\perp (X_{\mathcal{A}}(s), s \le t) | (X_{\mathcal{B}\cup\mathcal{C}}(s), s \le t) \tag{1.4.15}$$

$(t = 0, 1, \ldots, u-1)$.

To characterise relation (1.4.15) we need the following definition.

We say that $\mathcal{B}$ is **δ-separated** from $\mathcal{A}$ by $\mathcal{C}$ if every trail from $\mathcal{A}$ to $\mathcal{B}$ which ends with an arrow $\to b \in \mathcal{B}$ and has no other nodes in $\mathcal{B}$, satisfies at least one of the following conditions: it contains a chain $\leftarrow c \leftarrow$ or a fork $\leftarrow c \to$ or a chain $\to c \to$ with $c \in \mathcal{C}$ or a collider $\to v \leftarrow$ such that $v \notin \mathrm{an}(\mathcal{C})$. We will then write $\mathcal{A} \not\to_\delta \mathcal{B}|\mathcal{C}$. Negation of this statement is denoted $\mathcal{A} \to_\delta \mathcal{B}|\mathcal{C}$.

1.4.16 Theorem (**δ-separation, discrete case**). *Consider a discrete time process $\mathbf{X} = (X_v(t), t = 0, \ldots u)$ satisfying (1.3.1) with respect to graph $\mathcal{G}$. If $\mathcal{B}$ is δ-separated from $\mathcal{A}$ by $\mathcal{C}$ then (1.4.15) holds for every $t < u$:*

$$\mathcal{A} \not\to_\delta \mathcal{B}|\mathcal{C} \text{ implies } X_{\mathcal{B}}(t+1) \perp\!\!\!\perp (X_{\mathcal{A}}(s), s \le t) | (X_{\mathcal{B}\cup\mathcal{C}}(s), s \le t)$$

Proof. The proof is similar to that of Theorem 1.4.11. We just repalace $\mathrm{an}(\mathcal{B})$ by $\mathrm{pa}(\mathcal{B})$. Let

$$\mathcal{D}' = \mathrm{pa}(\mathcal{B}) \setminus (\mathcal{B} \cup \mathcal{C}),$$
$$\mathcal{R}' = \mathcal{V} \setminus (\mathcal{A} \cup \mathrm{pa}(\mathcal{B}) \cup \mathcal{C}).$$

We will use the same notation for $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ (and $\mathbf{a}, \mathbf{b}, \mathbf{c}$) as in the proof of Theorem 1.4.11. We now put

- $\mathbf{D}' = (X_{\mathcal{D}'}(s), 0 \le s \le t)$ – the past of $\mathbf{X}_{\mathcal{D}'}$,
- $\mathbf{R}' = (X_{\mathcal{R}'}(s), 0 \le s \le t)$ – the past of $\mathbf{X}_{\mathcal{R}'}$,
- $\mathbf{F}' = X_{\mathcal{B}}(t+1)$ – the immediate future of $\mathbf{X}_{\mathcal{B}}$.

We show that $\mathcal{A} \not\to_\delta \mathcal{B}|\mathcal{C}$ implies

$$\mathcal{D}' \perp_d \mathcal{A}|\mathcal{B} \cup \mathcal{C}.$$

Indeed, suppose that there exists a trail τ from $a \in \mathcal{A}$ to $d \in \mathcal{D}'$ which is not d-separated by $\mathcal{B} \cup \mathcal{C}$. By assumption, τ_y contains neither $\leftarrow v \leftarrow$ nor $\to v \to$ nor $\leftarrow v \to$ with $v \in \mathcal{B} \cup \mathcal{C}$. If all colliders in τ belong to an$(\mathcal{C})$ (or if there are no colliders) then τ with added d $\to b \in \mathcal{B}$ is not δ-separated by $\mathcal{C}$, contrary to our assumption. However, τ may contain a collider $\to v \leftarrow$ with $v \in \text{an}\mathcal{B}$ and $v \notin \text{an}(\mathcal{C})$. Assume that v is the first such collider. Then we can compose the part of τ ending at v with a directed path from v to the first element of $\mathcal{B}$ to obtain a δ-open trail from $\mathcal{A}$ to $\mathcal{B}$ (not separated by $\mathcal{C}$).

The obtained contradiction shows that $\mathcal{D}' \perp_d \mathcal{A}|\mathcal{B} \cup \mathcal{C}$. By Theorem 1.4.6 this graph separation statement entails conditional independence $\mathbf{D}' \perp\!\!\!\perp \mathbf{A}|\mathbf{B}, \mathbf{C}$ and thus

$$p(\mathbf{d}'|\mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{d}'|\mathbf{b}, \mathbf{c}) \tag{1.4.17}$$

The next equation,

$$p(\mathbf{f}'|\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = p(\mathbf{f}|\mathbf{b}, \mathbf{c}, \mathbf{d}') \tag{1.4.18}$$

follows from $\mathcal{B} \cup \mathcal{C} \cup \mathcal{D} \supseteq \text{pa}(\mathcal{B})$. Indeed, our basic assumption (1.3.1b) shows that the future of $X_\mathcal{B}(t+1)$ depends on $X(t)$ only through $X_{\text{pa}(\mathcal{B})}(t)$. Thus $\mathbf{F}'$ is conditionally independent of $\mathbf{A}$ and $\mathbf{R}$ given $(\mathbf{B}, \mathbf{C}, \mathbf{D}')$. It follows that $p(\mathbf{f}'|\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}') = p(\mathbf{f}'|\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}', \mathbf{r}') = p(\mathbf{f}'|\mathbf{b}, \mathbf{c}, \mathbf{d}')$ and we obtain (1.4.18). The rest of the proof is exactly the same as the proof of Theorem 1.4.11: from (1.4.17) and (1.4.18) we infer that

$$p(\mathbf{f}'|\mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{f}'|\mathbf{b}, \mathbf{c})$$

□

There also is a sort of converse to Theorem 1.4.6.

1.4.19 Proposition. *If $\mathcal{B}$ is not δ-separated from $\mathcal{A}$ by $\mathcal{C}$ then there exists a a discrete time process $\mathbf{X} = (X_v(t), t = 0, 1, \ldots, n)$ satisfying (1.3.1) such that (1.4.15) does not hold.*

Independences in stationarity: c-separation

Assume that a process $\mathbf{X} = (X_v(t), t = 0, 1, \ldots, u)$ satisfies (1.3.1) and is also a Markov chain with a unique stationary distribution π. Consider a process in equillibrium, ie. $X(t) \sim \pi$. We want to characterise the relation

$$X_\mathcal{A}(t) \perp\!\!\!\perp X_\mathcal{B}(t)|X_\mathcal{C}(t) \tag{1.4.20}$$

To emphasise that (1.4.20) is a statement involving one-dimensional 'time section' of the process, we rephrase this equation as '$X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} | X_{\mathcal{C}}$ with respect to π'.

We need yet another (very strong!) notion of separation.

We say that $\mathcal{A}$ and $\mathcal{B}$ are c-**separated** from by $\mathcal{C}$ if every trail from $\mathcal{A}$ to $\mathcal{B}$ contains a collider $\to v \leftarrow$ such that $v \notin \operatorname{an}(\mathcal{C})$. We will then write $\mathcal{A} \not\frown_c \mathcal{B}|\mathcal{C}$. Negation of this statement is denoted $\mathcal{A} \frown_c \mathcal{B}|\mathcal{C}$.

1.4.21 Theorem (c**-separation).** *Let* $\mathbf{X} = (X_v(t), t = 0, 1, \dots, u)$ *be a discrete time process satisfying* (1.3.1) *or a Continuous Time Bayesian Network with respect to graph* $\mathcal{G}$*. If* $\mathcal{A}$ *and* $\mathcal{B}$ *are c-separated by* $\mathcal{C}$ *then* (1.4.20) *holds for every* $t < u$*:*

$$\mathcal{A} \not\frown_c \mathcal{B}|\mathcal{C} \textit{ implies } X_{\mathcal{B}} \perp\!\!\!\perp X_{\mathcal{A}} | X_{\mathcal{C}} \textit{ wrt. } \pi$$

Proof. We focus on the discrete case (for CTBNs the argument is the same).

Without loss of generality we can assume that $\mathcal{V} = \operatorname{an}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C})$. Let

$$\mathcal{A}_1 = \operatorname{an}(\mathcal{A}), \quad \mathcal{B}_1 = \operatorname{an}(\mathcal{B})$$

We have $\mathcal{A}_1 \cap \mathcal{B}_1 = \emptyset$. Indeed, if we had $v \in \mathcal{A}_1 \cap \mathcal{B}_1$ then a trail composed of two directed paths from v to $a \in \mathcal{A}$ and from v to $b \in \mathcal{B}$ would be c-open, as it contains no colliders. Now observe that every trail in $\mathcal{V} \setminus (\mathcal{A}_1 \cup \mathcal{B}_1)$ is c-open (if it exists), because $\mathcal{V} \setminus (\mathcal{A}_1 \cup \mathcal{B}_1) \subseteq \operatorname{an}(\mathcal{C})$, so every collider must belong to $\operatorname{an}(\mathcal{C})$. It follows that three sets

$$\begin{aligned} \mathcal{A}_2 &= \mathcal{A}_1 \cup \{v : \text{ there is a trail between } v \text{ and } \mathcal{A}_1\} \\ \mathcal{B}_2 &= \mathcal{B}_1 \cup \{v : \text{ there is a trail between } v \text{ and } \mathcal{B}_1\} \\ \mathcal{C}_2 &= \mathcal{V} \setminus (\mathcal{A}_2 \cup \mathcal{B}_2) \end{aligned}$$

are disjoint and ancestral. Therefore

$$\pi(x_{\mathcal{A}_2}, x_{\mathcal{B}_2}, x_{\mathcal{C}_2}) = \pi(x_{\mathcal{A}_2})\pi(x_{\mathcal{B}_2})\pi(x_{\mathcal{C}_2})$$

Consequently the joint distribution of $(X_{\mathcal{A}}, X_{\mathcal{B}}, X_{\mathcal{C}})$ can be expressed as

$$\pi(x_{\mathcal{A} \cup (\mathcal{C} \cap \mathcal{A}_2)}, x_{\mathcal{B} \cup (\mathcal{C} \cap \mathcal{B}_2)}, x_{\mathcal{C} \cap \mathcal{C}_2}) = \pi(x_{\mathcal{A} \cup (\mathcal{C} \cap \mathcal{A}_2)})\pi(x_{\mathcal{B} \cup (\mathcal{C} \cap \mathcal{B}_2)})\pi(x_{\mathcal{C} \cap \mathcal{C}_2})$$

This is a function of $x_{\mathcal{A} \cup \mathcal{C}}$ times a function of $x_{\mathcal{B} \cup \mathcal{C}}$. □

Remark. We conjecture that c-separation is not only a sufficient but also necessary condition in Theorem 1.4.21 in the following sense: if $\mathcal{A} \frown_c \mathcal{B}|\mathcal{C}$ then there exsists a collection of (one-step) transition probabilities satisfying (1.3.1b) such that there exists a unique stationary distribution π and $X_{\mathcal{A}} \not\perp\!\!\!\perp X_{\mathcal{B}} | X_{\mathcal{C}}$ wrt. π.

Chapter 2

Causal Information Theory

2.1 Directed Information Theory

Mutual information

In this section we assume that random variables are discrete (take finite number of values). Symbol $P_{\mathbf{X}}$ denotes probability distribution of random variable $\mathbf{X}$. We will use concise notation as eg. $p(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x})$ whenever this does not lead to ambiguity.

We begin by recalling some equivalent definitions of *mutual information* (see [2]):

$$\begin{aligned} I(\mathbf{X} \wedge \mathbf{Y}) &= H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ &= D(P_{(\mathbf{X},\mathbf{Y})}; P_{\mathbf{X}} \times P_{\mathbf{Y}}) = \sum_{\mathbf{x},\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \\ &= \mathbb{E}_{\mathbf{X}} D(P_{\mathbf{Y}|\mathbf{X}}; P_{\mathbf{Y}}) = \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} \end{aligned}$$

where H denotes entropy and D is the Kullback-Leibler divergence. The *mutual conditional information* is

$$\begin{aligned} I(\mathbf{X} \wedge \mathbf{Y}|\mathbf{Z}) &= H(\mathbf{X}|\mathbf{Z}) + H(\mathbf{Y}|\mathbf{Z}) - H(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = H(\mathbf{Y}|\mathbf{Z}) - H(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \\ &= \mathbb{E}_{\mathbf{Z}} D(P_{(\mathbf{X},\mathbf{Y}|\mathbf{Z})}; P_{\mathbf{X}|\mathbf{Z}} \times P_{\mathbf{Y}|\mathbf{Z}}) = \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{\mathbf{x},\mathbf{y}} p(\mathbf{x}, \mathbf{y}|\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{y}|\mathbf{z})}{p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})} \\ &= \sum_{\mathbf{z}} p(\mathbf{z}) I(\mathbf{X} \wedge \mathbf{Y}|\mathbf{Z} = \mathbf{z}) \\ &= \mathbb{E}_{\mathbf{X},\mathbf{Z}} D(P_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}; P_{\mathbf{Y}|\mathbf{Z}}) = \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{z}) \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \mathbf{z}) \log \frac{p(\mathbf{y}|\mathbf{x}, \mathbf{z})}{p(\mathbf{y}|\mathbf{z})} \end{aligned}$$

The mutual conditional information is symmetric (with respect to $\mathbf{X}, \mathbf{Y}$). It is interpreted as the amount of information (decrease of uncertainty) about $\mathbf{Y}$ if we observe $\mathbf{X}$ (or amount of information about $\mathbf{X}$ if we observe $\mathbf{Y}$, equivalently), provided that $\mathbf{Z}$ is observed. Clearly, mutual information does not distinguish between the situation when $\mathbf{X}$ is the *cause* of $\mathbf{Y}$ from the situation when $\mathbf{X}$ is the *effect* of $\mathbf{Y}$.

Directed information theory (DIT) was created to describe and quantify causal (asymmetric) relations between variables. DIT deals principally with time series and is based on Granger understanding of causality. Causality is considered from a purely *predictive* perspective, without introducing the notion of *intervention*.

Directed information for time series

Consider a bivariate discrete time series $(\mathbf{X}, \mathbf{Y}) = ((X(t), Y(t)) : t = 0, \ldots, u)$. Write

- $\mathbf{X}(< t) = (X(s) : 0 \leqslant s < t)$, and similarly for $\mathbf{x}(< t)$, $\mathbf{Y}(< t)$,... etc.

Directed information $I(\mathbf{X} \to \mathbf{Y})$ is defined as follows:

$$I(\mathbf{X} \to \mathbf{Y}) = \sum_{t=1}^{u} I\big[Y(t) \wedge \mathbf{X}(< t) | \mathbf{Y}(< t)\ \big]$$

The mutual information can be decomposed as follows:

$$\begin{aligned}
I(\mathbf{X} \wedge \mathbf{Y}) &= \mathbb{E} \log \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{X}) p(\mathbf{Y})} \\
&= \mathbb{E} \log \frac{\prod_{t=0}^{u} p(X(t), Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t))}{\prod_{t=0}^{u} p(X(t) | \mathbf{X}(< t)) \prod_{t=0}^{u} p(Y(t) | \mathbf{Y}(< t))} \\
&= \mathbb{E} \sum_{t=0}^{u} \log \frac{p(X(t), Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t))}{p(X(t) | \mathbf{X}(< t), \mathbf{Y}(< t))\, p(Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t))} \times \\
&\qquad \times \frac{p(X(t) | \mathbf{X}(< t), \mathbf{Y}(< t))\, p(Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t))}{p(X(t) | \mathbf{X}(< t))\, p(Y(t) | \mathbf{Y}(< t))} \\
&= \sum_{t=0}^{u} I\big[X(t) \wedge Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t)\big] \\
&\qquad + \sum_{t=1}^{u} I\big[X(t) \wedge \mathbf{Y}(< t) | \mathbf{X}(< t) + \sum_{t=1}^{u} I\big[Y(t) \wedge \mathbf{X}(< t) | \mathbf{Y}(< t)\ \big]
\end{aligned}$$

We have obtained the following nice looking result.

2.1.1 Proposition (Decomposition of mutual info)**.** *For an arbitrary bivariate discrete time series* $(\mathbf{X}, \mathbf{Y})$,

$$I(\mathbf{X} \wedge \mathbf{Y}) = I(\mathbf{X} \to \mathbf{Y}) + I(\mathbf{Y} \to \mathbf{X}) + I(\mathbf{X} \sim \mathbf{Y})$$

where

$$I(\mathbf{X} \sim \mathbf{Y}) = I\big[X(0) \wedge Y(0)\big] + \sum_{t=1}^{u} I\big[X(t) \wedge Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t)\big]$$

2.1.2 Corollary. *If the series satisfies* (1.3.1)*, ie.* $X(t) \perp\!\!\!\perp Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t)$ *then*

$$I(\mathbf{X} \wedge \mathbf{Y}) = I(\mathbf{X} \to \mathbf{Y}) + I(\mathbf{Y} \to \mathbf{X})$$

The quantity

$$T_t(\mathbf{X} \to \mathbf{Y}) = I\big[Y(t) \wedge \mathbf{X}(< t) | \mathbf{Y}(< t)\ \big] \tag{2.1.3}$$

is called *transfer entropy*. This term was introduced in the influential paper of Schreiber [11]. The transfer entropy is an information-theoretic quantification of “Granger causality”.

Now we proceed to the trivariate case, ie. consider time series $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. *Conditional directed information* $I(\mathbf{X} \to \mathbf{Y} | \mathbf{Z})$ is defined as follows:

$$I(\mathbf{X} \to \mathbf{Y} | \mathbf{Z}) = \sum_{t=1}^{u} I\big[Y(t) \wedge \mathbf{X}(< t) | \mathbf{Y}(< t), \mathbf{Z}(< t)\ \big]$$

In the trivariate case, a result analogous to Proposition 2.1.1 is not as satisfactory. Computation is similar and goes as follows:

$$\begin{aligned}
&\mathbb{E} \log \frac{\prod_{t=0}^{u} p(X(t), Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t), \mathbf{Z}(< t))}{\prod_{t=0}^{u} p(X(t) | \mathbf{X}(< t), \mathbf{Z}(< t)) \prod_{t=0}^{u} p(Y(t) | \mathbf{Y}(< t), \mathbf{Z}(< t))} \\
&= \mathbb{E} \sum_{t=0}^{u} \log \frac{p(X(t), Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t), \mathbf{Z}(< t))}{p(X(t) | \mathbf{X}(< t), \mathbf{Y}(< t), \mathbf{Z}(< t))\, p(Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t), \mathbf{Z}(< t))} \times \\
&\qquad \times \frac{p(X(t) | \mathbf{X}(< t), \mathbf{Y}(< t), \mathbf{Z}(< t))\, p(Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t), \mathbf{Z}(< t))}{p(X(t) | \mathbf{X}(< t), \mathbf{Z}(< t))\, p(Y(t) | \mathbf{Y}(< t), \mathbf{Z}(< t))} \\
&= \sum_{t=0}^{u} I\big[X(t) \wedge Y(t) | \mathbf{X}(< t), \mathbf{Y}(< t), \mathbf{Z}(< t)\big] \\
&\qquad + \sum_{t=1}^{u} I\big[X(t) \wedge \mathbf{Y}(< t) | \mathbf{X}(< t), \mathbf{Z}(< t)\ \big] + \sum_{t=1}^{u} I\big[Y(t) \wedge \mathbf{X}(< t) | \mathbf{Y}(< t), \mathbf{Z}(< t)\ \big]
\end{aligned}$$

Let us denote the expression in the top line by $I(\mathbf{X} \wedge \mathbf{Y} \uparrow \mathbf{Z})$[1]. We have obtained the following result.

[1] This notation is not standard; see the remark after Proposition 2.1.4.

2.1.4 Proposition (Decomposition of info, conditional)**.** *For an arbitrary trivariate discrete time series* $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$*,*

$$I(\mathbf{X} \wedge \mathbf{Y} \uparrow \mathbf{Z}) = I(\mathbf{X} \to \mathbf{Y}|\mathbf{Z}) + I(\mathbf{Y} \to \mathbf{X}|\mathbf{Z}) + I(\mathbf{X} \sim \mathbf{Y}|\mathbf{Z})$$

where

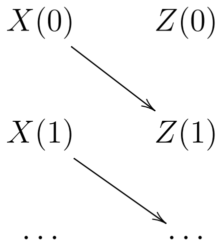
$$I(\mathbf{X} \sim \mathbf{Y}|\mathbf{Z}) = I\big[X(0) \wedge Y(0)|Z(0)\big] + \sum_{t=1}^{u} I\big[X(t) \wedge Y(t)|\mathbf{X}(<t), \mathbf{Y}(<t), \mathbf{Z}(<t)\big]$$

Remark. Unfortunately, in general

$$I(\mathbf{X} \wedge \mathbf{Y} \uparrow \mathbf{Z}) \neq I(\mathbf{X} \wedge \mathbf{Y}|\mathbf{Z})$$

Indeed, we have $I(\mathbf{X} \wedge \mathbf{Y}|\mathbf{Z}) = \mathbb{E} \log p(\mathbf{X}, \mathbf{Y}|\mathbf{Z})/p(\mathbf{X}|\mathbf{Z})p(\mathbf{Y}|\mathbf{Z})$. There is no reason to expect that the three products appearing in the definition of $I(\mathbf{X} \wedge \mathbf{Y} \uparrow \mathbf{Z})$ are equal to $p(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$, $p(\mathbf{X}|\mathbf{Z})$ and $p(\mathbf{Y}|\mathbf{Z})$ respectively.

For example it is not in general true that $p(\mathbf{X}|\mathbf{Z}) = \prod p(X(t)|\mathbf{X}(<t), \mathbf{Z}(<t))$. A counter-example is trivially simple:



Suppose that the variables without parents are distributed Ber(1/2) and arrows correspond to deteministic copying (eg. $\mathbb{P}(Z(1) = X(0)|X(0)) = 1$ etc.). Clearly we have $\mathbb{P}(X(0) = 1)$ $\mathbb{P}(X(1) = 1|Z(0) = 1, X(0) = 1) = 1/4$ and $\mathbb{P}(X(0) = 1, X(1) = 1|Z(0) = 1, Z(1) = 1) = 1/2$.

We introduced the nonstandard notation $I(\mathbf{X} \wedge \mathbf{Y} \uparrow \mathbf{Z})$ to underline that this quantity differs from the usual mutual conditional information $I(\mathbf{X} \wedge \mathbf{Y}|\mathbf{Z})$. In the next section we argue that $I(\mathbf{X} \wedge \mathbf{Y} \uparrow \mathbf{Z})$ has not a good interventional interpretation, either.

Let us emphasize that the definitions and results in this section apply to *arbitrary* time series $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ can well be themselves multivariate sub-series. We will be mostly interested in considering a process $\mathbf{X} = (X_v(t))$ which satisfies conditions (1.3.1) with respect to a DG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (in our terminology such $\mathbf{X}$ is a $\mathcal{G}$-CRE). Then we can consider quantities $I(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B})$ and $I(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}|\mathbf{X}_\mathcal{C})$ for any disjoint $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathcal{V}$. Hovever, we have to bear in mind that in general (1.3.1) *does not imply* $p(x_\mathcal{A}(t), x_\mathcal{B}(t)|\mathbf{x}_{\mathcal{A}\cup\mathcal{B}}(<t)) = p(x_\mathcal{A}(t)|\mathbf{x}_{\mathcal{A}\cup\mathcal{B}}(<t))p(x_\mathcal{B}(t)|\mathbf{x}_{\mathcal{A}\cup\mathcal{B}}(<t))$ and does not imply

$$p(x_\mathcal{A}(t), x_\mathcal{B}(t)|\mathbf{x}_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(<t)) = p(x_\mathcal{A}(t)|\mathbf{x}_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(<t))p(x_\mathcal{B}(t)|\mathbf{x}_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(<t))$$

unless $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C} = \mathcal{V}$.

Below we mention the relations between directed information (its nonzero values) and certain separation properties in the graph $\mathcal{G}$.

- $I(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}|\mathbf{X}_\mathcal{C}) > 0$ is equivalent to $X_\mathcal{B}(t) \not\perp\!\!\!\perp \mathbf{X}_\mathcal{A}(< t)|\mathbf{X}_\mathcal{C}$ and this conditional dependence implies that $\mathcal{A}$ is not d-separated from $\mathrm{pa}(\mathcal{B}) \setminus \mathcal{B}$ by $\mathcal{B} \cup \mathcal{C}$[2], where $\mathrm{pa}(\mathcal{B}) = \bigcup_{v \in \mathcal{B}} \mathrm{pa}(v)$.
- $I(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}|\mathbf{X}_{\mathcal{V} \setminus (\mathcal{A} \cup \mathcal{B})}) > 0$ implies that there is an arrow $(a \to b) \in \mathcal{E}$ for some $a \in \mathcal{A}$, $b \in \mathcal{B}$. This is a particularly simple special case of the previous statement, with $\mathcal{C} = \mathcal{V} \setminus (\mathcal{A} \cup \mathcal{B})$.
- $I(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}) > 0$ implies that there is a directed path from some $a \in \mathcal{A}$ to $b \in \mathcal{B}$ or there exist a node v with directed paths from v to $a \in \mathcal{A}$ and to $b \in \mathcal{B}$. This is also special case of the first statement, with $\mathcal{C} = \emptyset$.

2.2 Interventional conditioning in Information Theory

"Information Flow" of Ay and Polani

We assume the Pearlian causal model based on a DAG and defined by (1.2.1). Let $\mathcal{A}, \mathcal{B}, \mathcal{C}$ be disjoint subsets of $\mathcal{V}$, with $\mathcal{C}$ possibly empty[3]. The following definition appears in [1][4].

The **information flow** from $\mathbf{X}_\mathcal{A}$ to $\mathbf{X}_\mathcal{B}$ imposing $\mathbf{X}_\mathcal{C}$ is given by

$$(2.2.1) \qquad I_{\mathrm{AP}}(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}\|\mathbf{X}_\mathcal{C}) = \sum_{\mathbf{x}_\mathcal{C}} p(\mathbf{x}_\mathcal{C}) \sum_{\mathbf{x}_\mathcal{A}} p(\mathbf{x}_\mathcal{A}\|\mathbf{x}_\mathcal{C}) \sum_{\mathbf{x}_\mathcal{B}} p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) \log \frac{p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C})}{\hat{p}(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{C})}$$

where

$$(2.2.2) \qquad \hat{p}(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{C}) = \sum_{\mathbf{x}_\mathcal{A}} p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) p(\mathbf{x}_\mathcal{A}\|\mathbf{x}_\mathcal{C})$$

If $\mathcal{C} = \emptyset$ then (2.2.1) and (2.2.2) reduce to

$$I_{\mathrm{AP}}(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}) = \sum_{\mathbf{x}_\mathcal{A}} p(\mathbf{x}_\mathcal{A}) \sum_{\mathbf{x}_\mathcal{B}} p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}) \log \frac{p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A})}{\hat{p}(\mathbf{x}_\mathcal{B})}$$

[2] There is a trail from $\mathcal{A}$ to $\mathcal{B}$ which is not δ-separated by $\mathcal{C}$.

[3] We assume the setup considered in the cited paper [1]. Nevertheless, formulas (2.2.1) and (2.2.2) can be applied also if $\mathbf{X}$ is a CRE, that is a random element which satisfies Assumption 1.3.8. If $\mathbf{X}$ is a discrete time series which satisfies (1.3.1) then $\mathbf{X}$ can be equivalently regarded as a CRE or as a collection of variables $(X_v(t))$ arranged in a space-time DAG.

[4] It is rewritten in our notation, in particular using $p(\cdot\|\cdot)$ for conditioning-by-intervention and writing $\hat{p}$ in (2.2.2). We added subscript AP to avoid confusion with the directed information considered in the previous section.

and

$$\hat{p}(\mathbf{x}_{\mathcal{B}}) = \sum_{\mathbf{x}_{\mathcal{A}}} p(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{A}}) p(\mathbf{x}_{\mathcal{A}})$$

It is interesting to examine how this notion of information flow applies to multivariate time series, in particular to Markov chains.

2.2.3 Proposition. *Consider a bivariate time-homogeneous and stationary Markov chain* $(\mathbf{X}, \mathbf{Y}) = ((X(t), Y(t)) : t = 0, \dots, u)$ *which satisfies* (1.3.1b), *ie.*

$$p(x(t), y(t)|x(t-1), y(t-1)) = p(x(t)|x(t-1), y(t-1))\, p(y(t)|x(t-1), y(t-1))$$

Let $\pi(x, y)$ *be the stationary distribution. The following expressions hold:*

$$I(X(t-1) \wedge Y(t)|Y(t-1) = y) = \sum_x \pi(x|y) \sum_{y'} p(y'|x, y) \log \frac{p(y'|x, y)}{p(y'|y)} \tag{2.2.4}$$

where $p(y'|y) = \sum_x \pi(x|y) p(y'|x, y)$, *and*

$$I_{\mathrm{AP}}(X(t-1) \to Y(t) \| Y(t-1) = y) = \sum_x \pi(x) \sum_{y'} p(y'|x, y) \log \frac{p(y'|x, y)}{\hat{p}(y'\|y)} \tag{2.2.5}$$

where $\hat{p}(y'\|y) = \sum_x \pi(x) p(y'|x, y)$.

Of course, x, y and y' above are shorter notations for $x(t-1), y(t-1)$ and $y(t)$, respectively. Note that the differences between the *mutual information* in (2.2.4) and *information flow* in (2.2.5) are marked blue.

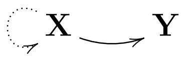
Proof. It is enough to note that $\mathbb{P}(X(t-1) = x|Y(t-1) = y) = \pi(x|y)$ if the chain is in the stationary regime. Equation (2.2.4) follows from the standard definition of mutual information. On the other hand $\mathbb{P}(X(t-1) = x\|Y(t-1) = y) = \pi(x)$. The transition probabilities $p(y'|x, y)$ are also conditional-by-intervention probabilities. Thus (2.2.5) follows from (2.2.1). □

Since $(\mathbf{X}, \mathbf{Y})$ is a Markov chain, $I(X(t-1) \wedge Y(t)|Y(t-1))$ upper bounds the transfer entropy:

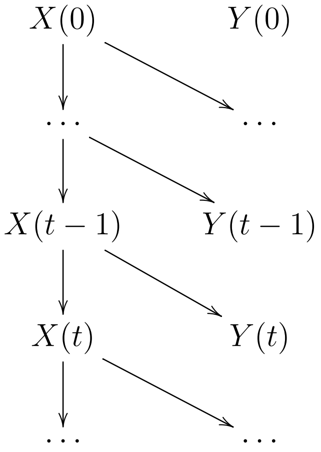
$$\begin{aligned}
T_t(\mathbf{X} \to \mathbf{Y}) &= I(\mathbf{X}(< t) \wedge Y(t)|\mathbf{Y}(< t)) \\
&= \mathbb{E} \log p(Y(t)|\mathbf{X}(< t), \mathbf{Y}(< t)) - \mathbb{E} \log p(Y(t)|\mathbf{Y}(< t)) \\
&= \mathbb{E} \log p(Y(t)|X(t-1), Y(t-1)) - \mathbb{E} \log p(Y(t)|\mathbf{Y}(< t)) \\
&\leqslant \mathbb{E} \log p(Y(t)|X(t-1), Y(t-1)) - \mathbb{E} \log p(Y(t)|Y(t-1)) \\
&= I(X(t-1) \wedge Y(t)|Y(t-1))
\end{aligned}$$

The equality $p(Y(t)|\mathbf{X}(< t), \mathbf{Y}(< t)) = p(Y(t)|X(t-1), Y(t-1))$ is the Markov property. The inequality holds because conditioning decreases entropy, so $-\mathbb{E} \log p(Y(t)|\mathbf{Y}(< t)) = H(Y(t)|\mathbf{Y}(< t)) \leqslant H(Y(t)|Y(t-1)) = -\mathbb{E} \log p(Y(t)|Y(t-1))$.

2.2.6 EXAMPLE. Consider the Markov chain $(\mathbf{X}, \mathbf{Y})$ with the structure of dependences given by the following graph:



Variables $X(t)$ and $Y(t)$ are binary, 0/1-valued. The corresponding space-time graph is the following:



The transition probabilities $\mathbb{P}(X(t) = x', Y(t) = y'|X(t-1) = x, Y(t-1) = y)$ are

$$p(x', y'|x, y) = p(x'|x)p(y'|x)$$

where

$$p(x'|x) = \begin{cases} a & \text{if } x' = x \\ 1-a & \text{if } x' \neq x \end{cases} \qquad\qquad p(y'|x) = \begin{cases} b & \text{if } x' = x \\ 1-b & \text{if } x' \neq x \end{cases}$$

To compute $I_{\mathrm{AP}}(X(t-1) \to Y(t) \| Y(t-1) = y)$, note that

- $p(x\|y) = \pi(x) = 1/2$
- $\hat{p}(y'\|y) = \sum_x p(y'\|x, y)p(x\|y) = \sum_x p(y'|x)\pi(x) = 1/2$

Therefore $\sum_{y'} p(y'|x, y) \log \dfrac{p(y'|x, y)}{\hat{p}(y'\|y)} = b \log \dfrac{b}{1/2} + (1-b) \log \dfrac{1-b}{1/2} = 1 - h(b)$.

Finally

$$I_{\mathrm{AP}}(X(t-1) \to Y(t) \| Y(t-1)) = 1 - h(b) \tag{2.2.7}$$

Elementary but slightly more tedious computations are needed for $I(X(t-1) \wedge Y(t)|Y(t-1))$. Let us introduce the following observations and notations:

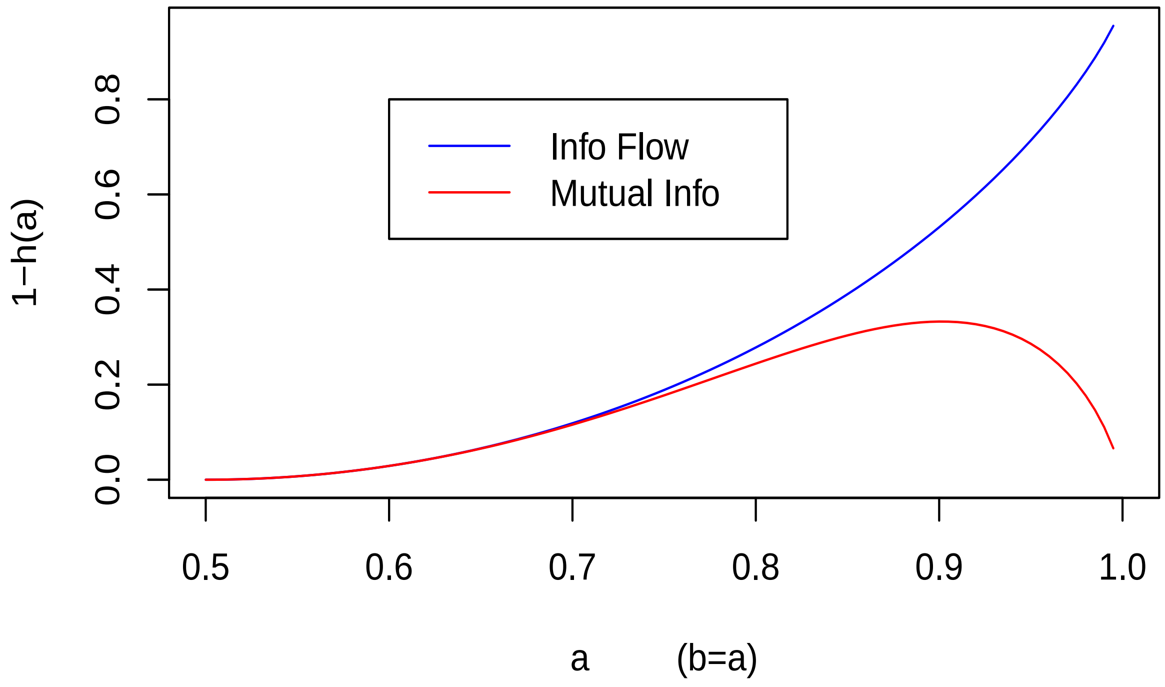
- $p(x|y) = \pi(x|y) = ab + (1-a)(1-b) =: d$ if $x = y$ and $\cdots = a(1-b) + (1-a)b = 1-d$ if $x \neq y$
- $p(y'|y) = ab^2 + a(1-b)^2 + (1-a)b^2 + 2(1-a)b(1-b) =: c$ if $y' = y$ and $\cdots = (1-a)b^2 + (1-a)(1-b)^2 + 2ab(1-b) = 1-c$ if $x \neq y$

Therefore

$$\sum_{y'} p(y'|x,y) \log \frac{p(y'|x,y)}{p(y'|y)} = \begin{cases} b \log \frac{b}{c} + (1-b) \log \frac{1-b}{1-c} & \text{if } x = y \\ (1-b) \log \frac{1-b}{c} + b \log \frac{b}{1-c} & \text{if } x \neq y \end{cases}$$

To compute $I(X(t-1) \wedge Y(t)|Y(t-1) = y)$ we multiply the upper line by d, the lower line by $1-d$, take the sum and rearrange terms. Note that $db + (1-d)(1-b) = c$ and $d(1-b) + (1-d)b = 1-c$. Finally we arrive at the formula

$$I(X(t-1) \wedge Y(t)|Y(t-1)) = h(c) - h(b) \tag{2.2.8}$$



The figure shows $I(X(t-1) \wedge Y(t)|Y(t-1))$ (Mutual Info) and $I_{\mathrm{AP}}(X(t-1) \to Y(t)\|Y(t-1))$ (Info Flow) as functions of a, assuming that $b = a$.

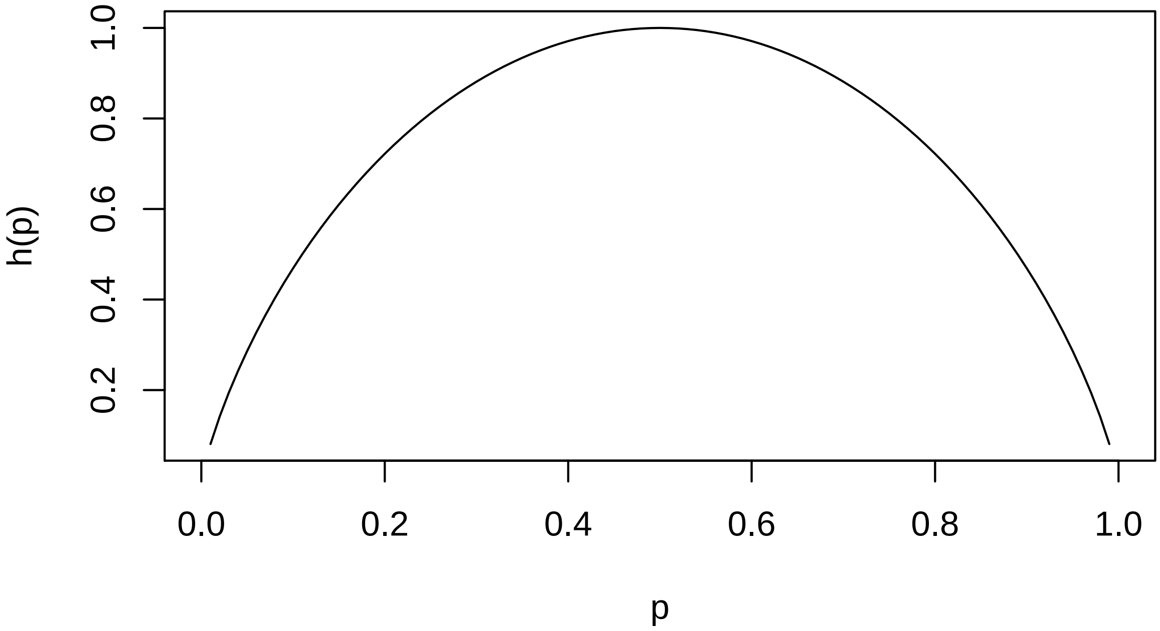
For a close to 0.5 we have almost independent variables and thus both curves approach 0 ($\mathbf{X}$ has neither predictive power for $\mathbf{Y}$ nor interventional effect on $\mathbf{Y}$). For a close to 1 the conditional predictive power of $\mathbf{X}$ vanishes because the past of $\mathbf{Y}$ is sufficient to predict its present ($a = 1$ means deterministic copying $X(t-1)$ to both $X(t)$ and $Y(t)$). On the other hand, intervention on $\mathbf{X}$ has a strong (almost deterministic) effect on $\mathbf{Y}$.

Racall that $T_t(\mathbf{X} \to \mathbf{Y}) \leqslant I(X(t-1) \wedge Y(t)|Y(t-1))$ Consequently in our example the transfer entropy goes to 0 while the the "information flow" goes to 1 as $a = b \to 1$. This shows the essential difference between the two information measures. △

Remark. There is a nice way to interpret formulas (2.2.7) and (2.2.8) (as well as a few next expressions in this section). Recall that $h(p)$ denotes the entropy of the probabilty distribution $\mathrm{Ber}(p)$:

$$h(p) = -p \log p - (1-p)\log(1-p)$$

The behaviour of function h is shown in the following figure:



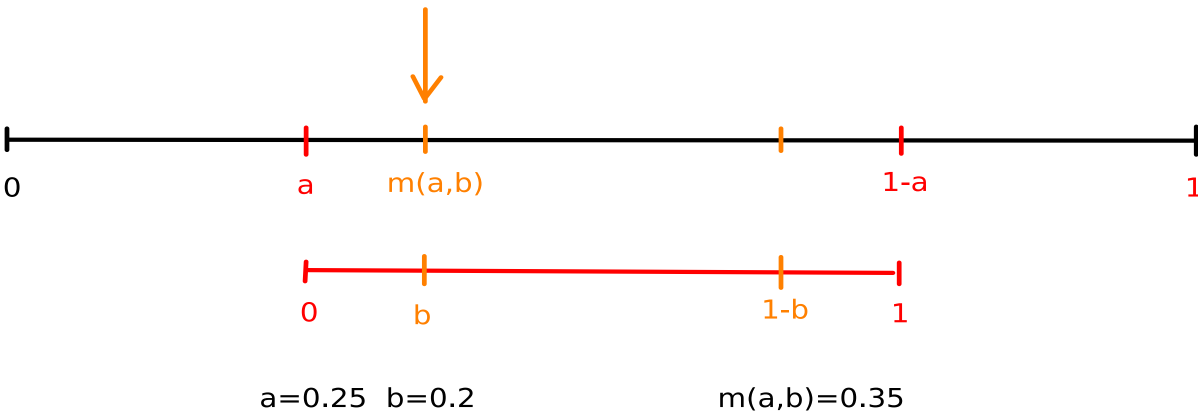
Function h is symmetric and has maximum at $1/2$. Let us introduce the following notation: for $0 < a < 1$ and $0 < b < 1$,

$$m(a, b) = a + b - 2ab$$

It is easy to verify that

- $m(a,b) = m(b,a)$
- if $m(0,a) = m(1,a) = a$
- $|m(a,b) - 0.5| \leqslant \min(|a-0.5|, |b-0.5|)$, so $h(m(a,b)) \geqslant \max(h(a), h(b))$
- $m(a, m(b,c)) = m(m(a,b),c)$, so we can define $m(a,b,c)$

The geometric interpretation of function m is shown below:



It is funny and puzzling that expressions with function m keep appearing in several formulas.

In Example 2.2.6 we can express c and d as $d = m(a,b)$, $c = m(d,b) = m(a,b,b)$ and obtain

$$I(X(t-1) \wedge Y(t)|Y(t-1)) = h(m(a,b,b)) - h(b)$$

Problems with the definition of "information flow"

Firstly, the information flow defined by (2.2.1) and (2.2.2) is supposed to measure the effect of $\mathbf{X}_{\mathcal{A}}$ on $\mathbf{X}_{\mathcal{B}}$ and not *vice versa*. This idea is captured by the Kullback-Liebler (K-L) divergence,

$$\sum_{\mathbf{x}_{\mathcal{B}}} p(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{A}}, x_{\mathcal{C}}) \log \frac{(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}})}{\hat{p}(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{C}})}, \tag{2.2.9}$$

which is “at the heart of” equation (2.2.1). But the definition of $I_{\mathrm{AP}}(\mathbf{X}_{\mathcal{A}} \to \mathbf{X}_{\mathcal{B}} \| \mathbf{X}_{\mathcal{C}})$ involves averaging of the above K-L divergence with respect to $p(\mathbf{x}_{\mathcal{C}})p(\mathbf{x}_{\mathcal{A}} \| \mathbf{x}_{\mathcal{C}})$. If $\mathcal{A}$ contains only a single node, there is no problem with interpreting the average. However, if there are arrows leading from $\mathcal{B}$ to $\mathcal{A}$ then $p(\mathbf{x}_{\mathcal{A}} \| \mathbf{x}_{\mathcal{C}})$ can well depend on $\mathbf{x}_{\mathcal{B}}$. Consequently, $I_{\mathrm{AP}}(\mathbf{X}_{\mathcal{A}} \to \mathbf{X}_{\mathcal{B}} \| \mathbf{X}_{\mathcal{C}})$ is not free from causal effect of $\mathbf{X}_{\mathcal{B}}$ on $\mathbf{X}_{\mathcal{A}}$ (in the “wrong” direction), even if $\mathcal{C}$ is empty.

We illustrate this by the following example.

2.2.10 EXAMPLE. Consider the following DAG: $Y_1 \to X \to Y_2$. Let $Y + (Y_1, Y_2)$. All three nodes are binary (0/1) variables. The joint probability distribution is specified as follows: $Y_1 \sim \mathrm{Ber}(d)$, $\mathbb{P}(X = Y_1 | Y_1) = a$ and $\mathbb{P}(Y_2 = X | X) = b$. Thus d, a and b are probabilities of copying values along the arrows in the following schematic picture:

$$\begin{tikzcd} 1 \arrow[r, "d"] & Y_1 \arrow[r, "a"] & X \arrow[r, "b"] & Y_2 \end{tikzcd}$$

Intuitively, a measure of causal effect such as $I_{\mathrm{AP}}(Y \to X)$ should depend on b but not on a (in contrast with $I(Y \wedge X)$). This is not the case. In fact,

- $I(X \wedge Y) = -h(a) - h(b) + h(m(a,b)) + h(m(a,d))$,
- $I_{\mathrm{AP}}(X \to Y) = -h(b) + h(m(d,a,b))$,

where functions h and m are defined in the Remark which followed Example 2.2.6. These facts can be shown by elementary but tedious computations. We see that I_{AP} does depend on a. When, for example, $a \to 1$ and $d \to 1$ then $I_{\mathrm{AP}}(X \to Y) \to 0$ even if $b \to 1$. This is clearly not in agreement with the natural interpretation of b as the “strength of causal effect of X on Y”. △

A more fundamental problem with the approach proposed by Ay and Polani in [1] is related to their understanding of (conditional) causal independence. Clearly,

$$I_{\mathrm{AP}}(\mathbf{X}_{\mathcal{A}} \to \mathbf{X}_{\mathcal{B}} \| \mathbf{X}_{\mathcal{C}}) = 0 \text{ iff } p(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}}) = \hat{p}(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{C}})$$

This fact, stated in [1, Proposition 2], is apparently an analogue of the well-known characterisation of probabilistic conditional independence:

$$I(\mathbf{X}_{\mathcal{A}} \wedge \mathbf{X}_{\mathcal{B}} | \mathbf{X}_{\mathcal{C}}) = 0 \text{ iff } p(\mathbf{x}_{\mathcal{B}} | \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}}) = p(\mathbf{x}_{\mathcal{B}} | \mathbf{x}_{\mathcal{C}})$$

that is iff $\mathbf{X}_{\mathcal{B}}$ is conditionally independent of $\mathbf{X}_{\mathcal{A}}$ given $\mathbf{X}_{\mathcal{C}}$.

The analogy is emphasised by the definition of “causal conditional independence” given in the cited article of Ay and Polani. They say that “$\mathbf{X}_{\mathcal{B}}$ is causally independent of $\mathbf{X}_{\mathcal{A}}$ imposing $\mathbf{X}_{\mathcal{C}}$ if $p(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}})$ does not depend on $\mathbf{x}_{\mathcal{A}}$”, see (1.4.2).

However, in Section 1.4 we showed that there is an essential difference between $\hat{p}(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{C}) = \sum_{\mathbf{x}_\mathcal{A}} p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C})p(\mathbf{x}_\mathcal{A}\|\mathbf{x}_\mathcal{C})$ and $p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{C})$, the latter being the genuine conditional-by - intervention distribution in the sense of Definition 1.3.9. We argued that $\mathbf{X}_\mathcal{B}$ should be said to be *causally independent of* $\mathbf{X}_\mathcal{A}$ *imposing* $\mathbf{X}_\mathcal{C}$ if equation (1.4.1) holds, ie.

$$p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{C})$$

This condition is not implied by $I_{\mathrm{AP}}(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}\|\mathbf{X}_\mathcal{C}) = 0$, as demonstrated by Example 1.4.3. Let us reconsider this example with a slight modification, to show difficulties in applying information-theoretic measures in the interventional setting. We will consider the simple case when $\mathcal{C} = \emptyset$, that is focus on $I_{\mathrm{AP}}(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B})$.

2.2.11 EXAMPLE (Modified version of Example 1.4.3). Consider three binary random variables $\mathbf{X}_a, \mathbf{X}_b, \mathbf{X}_u$ and the following graph of causal relations:

$$\begin{tikzcd}
\mathbf{X}_u \arrow[d] \arrow[dr] & \\
\mathbf{X}_a \arrow[r] & \mathbf{X}_b
\end{tikzcd}$$

Let $\mathbb{P}(\mathbf{X}_u = 0) = \mathbb{P}(\mathbf{X}_u = 1) = 1/2$, $\mathbb{P}(\mathbf{X}_a = \mathbf{X}_u|\mathbf{X}_u) = \varepsilon$ and $\mathbf{X}_b = \mathbb{1}(\mathbf{X}_a = \mathbf{X}_u)$. Put $\mathcal{A} = \{a\}$ and $\mathcal{B} = \{b\}$. Clearly we have

$$\mathbb{P}(\mathbf{X}_b = 1\|\mathbf{X}_a = 0) = \mathbb{P}(\mathbf{X}_b = 1\|\mathbf{X}_a = 1) = \frac{1}{2}$$

so

$$\hat{\mathbb{P}}(\mathbf{X}_b = 1) = \mathbb{P}(\mathbf{X}_b = 1\|\mathbf{X}_a = 0)\frac{1}{2} + \mathbb{P}(\mathbf{X}_b = 1\|\mathbf{X}_a = 1)\frac{1}{2} = \frac{1}{2}$$

On the other hand,

$$\mathbb{P}(\mathbf{X}_b = 1) = \varepsilon$$

Consequently $I_{\mathrm{AP}}(\mathbf{X}_a \to \mathbf{X}_b) = 0$ but $\mathbb{P}(\mathbf{X}_b = 1\|\mathbf{X}_a) \neq \mathbb{P}(\mathbf{X}_b = 1)$. We insist that one should *not* say "$\mathbf{X}_b$ is causally independent of $\mathbf{X}_a$". In the sense defined by our equation (1.4.1), it is not. Recall that the particular instance of our definition is the following: $\mathbf{X}_b$ is causally independent of $\mathbf{X}_a$ if

$$\mathbb{P}(\mathbf{X}_b = 1\|\mathbf{X}_a) = \mathbb{P}(\mathbf{X}_b = 1)$$

△

Our above discussion would suggest the following alternative definition of "information flow":

$$I_{\mathrm{WN}}(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}) = \sum_{\mathbf{x}_\mathcal{A}} p(\mathbf{x}_\mathcal{A}) \sum_{\mathbf{x}_\mathcal{B}} p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}) \log \frac{p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A})}{p(\mathbf{x}_\mathcal{B})}$$

Clearly,

$$I_{\mathrm{WN}}(\mathbf{X}_\mathcal{A} \to \mathbf{X}_\mathcal{B}) = 0 \text{ iff } p(\mathbf{x}_\mathcal{B}\|\mathbf{x}_\mathcal{A}) = p(\mathbf{x}_\mathcal{B})$$

ie. iff $\mathbf{X}_{\mathcal{B}}$ is causally independent of $\mathbf{X}_{\mathcal{A}}$. Therefore, I_{WN} seems to be consistent with our definition of causal independence. Unfortunately, I_{WN} does not have properties we expect that a measure of "amount of information" should have.

2.2.12 EXAMPLE (Continuation of Example 2.2.11). Let us compute $I_{\mathrm{WN}}(\mathbf{X}_a \to \mathbf{X}_b)$ in the previous example. Since $p(\mathbf{x}_b \| \mathbf{x}_a) = 1/2$ and $\mathbb{P}(\mathbf{X}_b = 1) = \varepsilon$, $\mathbb{P}(\mathbf{X}_b = 0) = 1 - \varepsilon$,

$$\sum_{\mathbf{x}_b} p(\mathbf{x}_{\mathcal{B}} \| \mathbf{x}_a) \log \frac{p(\mathbf{x}_b \| \mathbf{x}_a)}{p(\mathbf{x}_b)} = \frac{1}{2} \log \frac{1/2}{\varepsilon} + \frac{1}{2} \log \frac{1/2}{1-\varepsilon}$$

It follows that

$$I_{\mathrm{WN}}(\mathbf{X}_a \to \mathbf{X}_b) = -1 - \frac{1}{2}\big[\log \varepsilon + \log(1-\varepsilon)\big]$$

For $\varepsilon \to 0$, we see that $I_{\mathrm{WN}}(\mathbf{X}_a \to \mathbf{X}_b) \to +\infty$. If we are to interpret $I_{\mathrm{WN}}(\mathbf{X}_a \to \mathbf{X}_b)$ as "a measure of information on $\mathbf{X}_b$ gained by setting $\mathbf{X}_a$ to a value $\mathbf{x}_a$ by intervention" then this result is nonsense. For a binary variable, the maximum information on it is 1 bit. △

Conclusion of this section (and of the entire chapter) is rather disappointing. In our view, it is difficult to reconcile the information-theoretic approach with the idea of intervention.

Bibliography

[1] Nihat Ay and Daniel Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(1): 17–41, 2008.

[2] Thomas M. Cover, Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 9. edition, 1991.

[3] Vanessa Didelez. Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics*, 34(1): 169–185, 2007.

[4] Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. In *23rd Annual Conference on Uncertainty in Artifical Intelligence*, 2007.

[5] Clive W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438, 1969.

[6] John Hammersley and Peter Clifford. Markov fields on finite graphs and lattices. 1971.

[7] Steffen Lauritzen, A. Philip Dawid, B. Larsen and H. Leimer. Independence Properties of Directed Markov Fields. *Networks* 20: 491–505, 1990.

[8] Judea Pearl. *Causality: Models, Reasoning and Inference: Models, Reasoning and Inference.* Cambridge University Press, 2. edition, 2009.

[9] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

[10] Wojciech Niemiro and Łukasz Rajkowski. Local Dependence Graphs for Discrete Time Processes. *Proceedings of the Second Conference on Causal Learning and Reasoning, PMLR* 213:772–790, 2023.

[11] Thomas Schreiber. Measuring Information Transfer. *Physical Review Letters* 85(2), 461–464, 2000.

[12] Tore Schweder. Composable Markov processes. *Journal of Applied Probability*, 7(2): 400–410, 1970.

Appendix A

Derivation of (1.3.5)

We consider continuous time Markov process $\mathbf{X} = (X(t) : 0 \leqslant t \leqslant u)$ with values in a finite state space $\mathcal{S}$. The aim is to derive a formula for the density $p(\mathbf{x})$, where $\mathbf{x}$ is a trajectory of $\mathbf{X}$.

An infromal derivation goes as follows. First consider the time of the first jump, ie.

$$T = \inf\{t > 0 : X(t) \neq x\}$$

under the condition $X(0) = x$. This condition will be tacitly assumed and omitted from notation. Let

$G(t) = \mathbb{P}(T > t)$ and let $g(t)$ be the density of random variable T.

Recall that $\mathbb{P}(X(t+h) = x | X(t) = x) = 1 + Q(t; x, x)h + o(h) = 1 - \sum_{x' \neq x} Q(t; x, x') + o(h)$. Using our notations we obtain[1]

$$G(t+h) = \mathbb{P}(T > t)\mathbb{P}(T > t+h | T > t) = G(t)[1 + Q(t; x, x)h + o(h)]$$

and consequently

$$\frac{G(t+h) - G(t)}{G(t)} = Q(t; x, x)h + o(h)$$

hence

$$\frac{\mathrm{d}}{\mathrm{d}t} \log G(t) = \lim_{h \searrow 0} \frac{G(t+h) - G(t)}{hG(t)} = Q(t; x, x)$$

[1] The second equality in this expression is intuitively rather obvious but perhaps needs an explanation. We have $\mathbb{P}(T > t+h | T > t) = \mathbb{P}(X(s) = x$ for all $s \in [t, t+h] | X(t) = x)$ by the Markov property. This probability is equal to $\mathbb{P}(X(t+h) = x | X(t) = x) + o(h)$, because two or more jumps can occur in a short interval with probability $o(h)$.

hence

$$G(t) = \exp\left\{\int_0^t Q(s;x,x)\mathrm{d}s\right\} \tag{A.0.1}$$

and

$$g(t) = -Q(t;x,x)\exp\left\{\int_0^t Q(s;x,x)\mathrm{d}s\right\} \tag{A.0.2}$$

Let us now turn to the state, $X(T)$, assumed by the process at the first jump (by convention, the trajectories are right-continuous). Using the Markov property and the Bayes rule we obtain for $x' \neq x$

$$\mathbb{P}(X(T) = x'|T = t) = \mathbb{P}(X(t) = x'|X(t-) = x, X(t) \neq x) = \frac{Q(t;x,x')}{-Q(t;x,x)}. \tag{A.0.3}$$

Let $0 < T_1 < T_2 < \cdots < T_N < u < T_{N+1}$ be moments of consecutive jumps of the process $\mathbf{X}$ and put $X_i = X(T_i)$ (by convention, the trajectories are right-continuous). Trajectory $\mathbf{x} = (x(t) : 0 \leqslant t \leqslant u)$ can be encoded as a "space-time skeleton":

$$\mathbf{x} \equiv \begin{pmatrix} 0 & t_1, & \cdots & t_n \\ x_0 & x_1, & \cdots & x_n \end{pmatrix} \tag{A.0.4}$$

where t_is and x_is are realisations of random variables T_i and X_i. Using (A.0.2), (A.0.3) and finally (A.0.1) we see that (informally written) the density of the skeleton is equal to

$$\begin{aligned}
&p(x_0, t_1, x_1, \dots, t_n, n) \\
&= \mathbb{P}(X_0 = x_0, T_1 = t_1, X_1 = x_1, \dots, T_n = t_n, X_n = x_n, N = n) \\
&= \mathbb{P}(X_0 = x_0, T_1 = t_1, X_1 = x_1, \dots, T_n = t_n, X_n = x_n, T_{n+1} > u) \\
&= \nu(x_0)Q(t_1;x_0,x_0)\exp\left\{\int_0^{t_1} Q(s;x_0,x_0)\mathrm{d}s\right\}\frac{Q(t_1;x_0,x_1)}{Q(t_1;x_0,x_0)} \\
&\qquad \times Q(t_2;x_1,x_1)\exp\left\{\int_{t_1}^{t_2} Q(s;x_1,x_1)\mathrm{d}s\right\}\frac{Q(t_2;x_1,x_2)}{Q(t_2;x_1,x_1)} \qquad \text{(A.0.5)} \\
&\cdots \quad \cdots \\
&\qquad \times \qquad \exp\left\{\int_{t_n}^{u} Q(s;x_n,x_n)\mathrm{d}s\right\} \\
&= Q(t_1;x_0,x_1)Q(t_2;x_1,x_2)\cdots Q(t_n;x_{n-1},x_n)\exp\left\{\int_0^u Q(s;x(s),x(s))\mathrm{d}s\right\}
\end{aligned}$$

Appendix B

Proof of Theorem (1.4.7)

Proof of Theorem 1.4.7. For the convenience of formulation, we show that

$$\Big(\mathcal{A} \text{ and } B \text{ are } \textbf{not } d\text{-separated by } \mathcal{C} \text{ in } \mathcal{G}\Big) \iff \Big(\mathcal{A} \text{ and } \mathcal{B} \text{ are } \textbf{not} \text{ separated by } \mathcal{C} \text{ in } \mathcal{G}^m_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}\Big)$$

($\Rightarrow$) Consider an $\mathcal{C}$-active trail $\tau = (u_0, u_1, \ldots, u_k)$ from $u_0 = a \in \mathcal{A}$ to $u_k = b \in \mathcal{B}$. Note that all vertices on that path belong to $\mathrm{an}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C})$. Indeed, let $u_p \to u_{p+1}$ be the left-most $\to$ connection, similarly let $u_{q-1} \leftarrow u_q$ be the right-most $\leftarrow$ connection. Then $u_0, \ldots, u_{p-1} \in \mathrm{an}(\mathcal{A})$ (for $p = 0$ this is an empty condition) and $u_{q+1}, \ldots, u_k \in \mathrm{an}(\mathcal{B})$. Moreover the trail from u_p to u_q can be split into 'sinks' of colliders[1] on that trail. The trail τ was $\mathcal{C}$-active, which means that every collider belongs to $\mathrm{an}(\mathcal{C})$ (and so do all vertices in its sink), therefore $u_p, u_{p+1}, \ldots, u_q \in \mathrm{an}(\mathcal{C})$. This means that τ prevails in $\mathcal{G}_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$. Let $u_{n_1}, \ldots, u_{n_l}$ be all vertices on this trail that belong to $\mathcal{C}$ (in the order of appearance). Since τ is $\mathcal{C}$-active, all those vertices must be colliders, which also means that the vertices $u_{n_i \pm 1}$ are not in $\mathcal{C}$. As u_{n_i} are colliders, edges $u_{n_i-1} — u_{n_i+1}$ are present in $\mathcal{G}^m_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$ for all $1 \leqslant i \leqslant l$. This proofs that we can remove from τ all vertices that belong to $\mathcal{C}$ and obtain a valid path in $\mathcal{G}^m_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$ that does not intersect $\mathcal{C}$.

($\Leftarrow$) Consider a path $\tau = (u_0, u_1, \ldots, u_k)$ in $\mathcal{G}^m_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$ that does not intersect $\mathcal{C}$ (i.e. $v_i \notin \mathcal{C}$ for $0 \leqslant i \leqslant k$). Some of the edges on this path were created by the moralization procedure. Let $u_{m_i} — u_{m_{i+1}}$ $(0 \leqslant i \leqslant k-1)$ be all edges on τ that are not present in the undirected version of $\mathcal{G}_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$ (arrows replaced by undirected edges and double edges removed). This means that there exist $v_{m_i} \in \mathrm{an}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C})$ $(i \leqslant k-1)$ such that $u_{m_i} \to v_{m_i} \leftarrow u_{m_i+1}$ in $\mathcal{G}_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$. Consider the trail τ' obtained from τ by inserting v_{m_i}s in-between all connections

[1] by 'sink of a collider on trail' we mean the collider together with its ancestral set in the graph restricted to the trail.

u_{m_i} — $u_{m_{i+1}}$. This trail is present in $\mathcal{G}_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$, moreover

$$(\star) \qquad \text{all its vertices that belong to } \mathcal{C} \text{ must be colliders}$$

(as $u_0, \ldots, u_k \notin \mathcal{C}$ and all v's are colliders). If all colliders on τ' belong to $\mathrm{an}(\mathcal{C})$ then ($\star$) implies that τ' is $\mathcal{C}$-active in $\mathcal{G}$. Otherwise let $c \notin \mathrm{an}(\mathcal{C})$ be a collider on τ'. As $\tau' \subseteq \mathcal{A}\cup\mathcal{B}\cup\mathcal{C}$ we get $c \in \mathrm{an}(\mathcal{A}\cup\mathcal{B})$. Without loss of generality assume that $c \in \mathrm{an}(\mathcal{A})$. This means that there is a directed path from c to some $a' \in \mathcal{A}$ which does not intersect $\mathcal{C}$ (since $c \notin \mathrm{an}(\mathcal{C})$). As a consequence $a' \leftarrow \ldots \leftarrow c \leftarrow [\cdots]b$ is a trail in $\mathcal{G}_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$ that satisfies ($\star$) (here by '$[\cdots]b$' we mean that we loose track of the edges' directions). Note that c is no longer a collider and no colliders have been created by this procedure. Hence the number of colliders decreased. By iteration we obtain a trail in $\mathcal{G}_{\mathrm{an}(\mathcal{A}\cup\mathcal{B}\cup\mathcal{C})}$ that satisfies ($\star$) and has no colliders outside $\mathrm{an}(\mathcal{C})$, which proves that $\mathcal{A}$ is not d-separated by $\mathcal{C}$ from $\mathcal{B}$ in $\mathcal{G}$. □