

Decision theoretic foundations for statistical causal modelling

David Johnston

A thesis submitted for the degree of Doctor of Philosophy

College of Engineering and Computer Science



Australian
National
University

September, 2022

Declaration

This dissertation is an account of research undertaken between January 2018 and September 2022 at the College of Engineering and Computer Sciences, The Australian National University, Canberra, Australia.

The work presented in this thesis is that of the candidate alone, except where indicated by due literature reference and acknowledgements in the text. It has not been submitted in whole or in part for any other degree at this or any other university.

The development of ideas and research was undertaken with guidance from my primary supervisors Robert Williamson and Cheng Soon Ong, and the thesis was written by myself. The overall direction of the research was developed in collaboration with my supervisors, who also provided a lot of detailed feedback on my work along the way. The original results presented here were primarily my work.

David Johnston
15 September 2022

Acknowledgements

Acknowledgements

Abstract

Update abstract

Keywords: causal inference, decision theory

Contents

Acknowledgements	ii
Abstract	iii
List of Symbols	v
1 Other causal modelling frameworks	1
1.1 What is a Causal Bayesian Network?	1
1.2 What is a Potential Outcomes model?	14
1.3 Individual-level response functions	16
1.4 Conclusion	29
Bibliography	30
List of notation	

List of Symbols

Name	Notation	Meaning	Reference
Probability theory			
Iverson bracket	$\llbracket \cdot \rrbracket$	Function equal to 1 if \cdot is true, false otherwise	
Variable	X	Measurable function $(\Omega, \mathcal{F}) \rightarrow (X, \mathcal{X})$	Definition ??
Trivial variable	$*$	Single-valued variable, maps $(\Omega, \mathcal{F}) \rightarrow (\{*\}, \{\emptyset, \{*\}\})$	Definition ??
Variable sequence	(X, Y)	The variable given by $\omega \mapsto (X(\omega), Y(\omega))$	Definition ??
Probability measure	$\mathbb{P} \in \Delta(\Omega)$	Countably additive measure on (Ω, \mathcal{F}) with $\mathbb{P}(\Omega) = 1$	Definition ??
Set of probability measures	$\Delta(\Omega)$	Set of probability measures on (Ω, \mathcal{F})	Notation ??
Markov kernel	$\mathbb{K} : X \rightarrow Y$	Measurable map from (X, \mathcal{X}) to probability measures on (Y, \mathcal{Y})	Definition ??
Dirac measure	δ_x	Probability measure where $\delta_x(A) = 1$ if $x \in A$, 0 otherwise	Definition ??
Markov kernel associated with a function	\mathbb{F}_f	Markov kernel associated with $f : X \rightarrow Y$ that maps $x \mapsto \delta_{f(x)}$	Definition ??
Marginal distribution	\mathbb{P}^X	$\mathbb{P}\mathbb{F}_X$	Definition ??
Conditional distribution	$\mathbb{P}^{Y X}$	Arbitrary Markov kernel $X \rightarrow Y$ such that $\mathbb{P}^{XY}(A \times B) = \int_A \mathbb{P}^{Y X}(B x) \mathbb{P}^X(dx)$	Definition ??
Conditional independence	$X \perp\!\!\!\perp_{\mathbb{P}} Y Z$	$\mathbb{P}^{X YZ}(A y, z)$ does not depend on z	Definition ??

Name	Notation	Meaning	Reference
Uniform conditional probability	$\mathbb{P}_A^{Y X}$	Arbitrary Markov kernel $X \rightarrow Y$ that is a conditional distribution for every $\alpha \in A$	Definition ??
Kernel product	$\mathbb{K}\mathbb{L}$	The Markov kernel given by $(A x) \mapsto \int_Y \mathbb{L}(A y)\mathbb{K}(\mathrm{d}y x)$	Definition ??
Semidirect product	$\mathbb{K} \odot \mathbb{L}$	The Markov kernel given by $(A \times B x) \mapsto \int_A \mathbb{L}(B) y)\mathbb{K}(\mathrm{d}y x)$	Definition ??

String diagrams

Identity map	Id_X	Markov kernel associated with the identity function $X \rightarrow X$	Definition ??
Erase map	Del_X, \uparrow^*	Markov kernel associated with the trivial variable $*_X : X \rightarrow \{*\}$	Definition ??
Swap map	Swap_{XY}, \times	Markov kernel associated with the function that swaps its inputs $(x, y) \mapsto (y, x)$	Definition ??
Swap according to permutation	Swap_ρ	Markov kernel that swaps inputs in a manner specified by permutation ρ	
Copy map	Copy_X, \cup	Markov kernel associated with the function that makes two copies of its inputs	Definition ??

Probability sets and decision models

Probability set	\mathbb{P}_A	A collection of probability measures $\{\mathbb{P}_\alpha \alpha \in A\}$ on a common sample space	Definition ??
Decision model	$(\mathbb{P}_C, (\Omega, \mathcal{F}), C)$	An option set C , a sample space (Ω, \mathcal{F}) and a probability measure \mathbb{P}_α for each option	Definition ??

Name	Notation	Meaning	Reference
Option set	C	Interpreted as the set of options available to a decision maker	Definition ??
Nonstochastic variable	ϕ	Function defined on the option set $C \rightarrow A$	Definition ??
Complementary variables	(ϕ, ξ)	Sequence of non-stochastic variables that induces an invertible function	Definition ??
Extended conditional independence	$X \perp\!\!\!\perp_{\mathbb{P}_C}^e (Y, \phi) (Z, \xi)$	Generalisation of conditional independence to decision models	Definition ??
Choice variable	id_C	Identity function on option set C ; corresponds to the choice made by decision maker	
Tabular conditional	Y^X	Variable with the property that $Y = \sum_{x \in X} \llbracket X = x \rrbracket Y^x$; not necessarily interpretable as potential outcomes	Definition ??

Chapter 1

Other causal modelling frameworks

In this chapter, we examine the types of decision models that can be constructed from causal Bayesian networks and potential outcomes models. Neither of these popular approaches to causal inference yields a fully specified decision making model. Causal Bayesian networks are usually specified in a “rolled up” form, and certain judgements must be made about how this should be unrolled to a sequential model. Potential outcomes models, on the other hand, do not feature a native notion of “choices”, and a judgement must be made about what the relevant collection of choices in a potential outcomes model is.

1.1 What is a Causal Bayesian Network?

1.1.1 Definition of a Causal Bayesian Network

We follow the definition of a Causal Bayesian Network on [Pearl \(2009, page 23-24\)](#). There are a couple of technical differences: we require that interventional models are a measurable map from interventions to probability distributions, and we assume that there is a common sample space for every interventional distribution. There are also some non-technical differences: the notation is adapted for compatibility with the rest of the work in this thesis, and we separate the definition into two parts for clarity (Definitions [1.1.10](#) and [1.1.11](#)).

An interventional model is a *Causal Bayesian Network* with respect to a directed acyclic graph if it satisfies a number of compatibility requirements. The following definitions are standard, and reproduced here for convenience. The definitions here are terse, readers should refer to [Pearl \(2009, chap. 1\)](#) for a more intuitive explanation.

Definition 1.1.1 (Directed graph). A directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ is a set of nodes $\tilde{\mathcal{V}}$ and edges, which are ordered pairs of nodes $\tilde{\mathcal{E}} \subset \tilde{\mathcal{V}} \times \tilde{\mathcal{V}}$. Nodes are written using the font $\tilde{\mathcal{V}}$.

The parents of a target node are all nodes with an edge ending at the target node.

Definition 1.1.2 (Parents). Given a directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ and $\tilde{V}_i \in \tilde{\mathcal{V}}$, the parents of \tilde{V}_i are $\text{Pa}_{\tilde{\mathcal{G}}}(\tilde{V}_i) := \{\tilde{V}_j \mid (\tilde{V}_j, \tilde{V}_i) \in \tilde{\mathcal{E}}\}$.

A path is a sequence of edges such that the i th edge and the $i + 1$ th edge share exactly one node.

Definition 1.1.3 (Path). Given a directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, a path is a sequence of edges $(E_i)_{i \in A}$ (where A is either $[n]$ or \mathbb{N}) such that for any i , E_i and E_{i+1} share exactly one node.

A directed path is a sequence of edges such that the end of the i th edge is the beginning of the $i + 1$ th edge.

Definition 1.1.4 (Directed path). Given a directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, a directed path is a sequence of edges $(E_i)_{i \in A}$ (where A is either $[n]$ or \mathbb{N}) such that for any i , $E_i = (\tilde{V}_k, \tilde{V}_l)$ implies $E_{i+1} = (\tilde{V}_l, \tilde{V}_m)$ for some $\tilde{V}_m \in \tilde{\mathcal{V}}$.

In an acyclic graph, directed paths never reach to the same node more than once.

Definition 1.1.5 (Directed acyclic graph). A directed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ is acyclic if, for every path, each node appears at most once. Directed acyclic graph is abbreviated to “DAG”.

D-separation is a key property of directed acyclic graphs for defining causal Bayesian networks. It is defined with respect to undirected paths.

Definition 1.1.6 (Blocked path). Given a DAG $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, a path p is blocked by $\tilde{V}_A \subset \tilde{\mathcal{V}}$ iff

1. $(\tilde{V}_i, \tilde{V}_j) \in p$ and $(\tilde{V}_j, \tilde{V}_k) \in p$ while $\tilde{V}_j \in \tilde{V}_A$
2. $(\tilde{V}_j, \tilde{V}_i) \in p$ and $(\tilde{V}_j, \tilde{V}_k) \in p$ while $\tilde{V}_j \in \tilde{V}_A$
3. $(\tilde{V}_i, \tilde{V}_j) \in p$ and $(\tilde{V}_k, \tilde{V}_j) \in p$ while $\tilde{V}_j \cup \text{De}_{\tilde{\mathcal{G}}}(\tilde{V}_j) \cap \tilde{V}_A = \emptyset$

Definition 1.1.7 (d-separation). Given a DAG $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, \tilde{V}_A is d -separated from \tilde{V}_B by \tilde{V}_C (all subsets of $\tilde{\mathcal{V}}$) if \tilde{V}_C blocks every path starting at \tilde{V}_A and ending at \tilde{V}_B . This is written $\tilde{V}_A \perp_{\tilde{\mathcal{G}}} \tilde{V}_B | \tilde{V}_C$.

Definition 1.1.8 (Variable-node association). Given a graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ and a sequence of variables $(V_i)_{i \in A}$, if $|A| = |\tilde{\mathcal{V}}|$ we can associate a variable with each node of the graph with an invertible map $m : \{V_i | i \in A\} \rightarrow \tilde{\mathcal{V}}$. By convention, we give associated variables and nodes corresponding indices, and graphical operations are defined on variables through m , i.e. $\text{Pa}(V_i) := m(\text{Pa}(\tilde{V}_i))$.

Definition 1.1.9 (Compatibility). Given a measurable space (Ω, \mathcal{F}) , a Markov kernel $\mathbb{P} : \mathcal{C} \rightarrow \Omega$ and a sequence of variables $(V_i)_{i \in A}$ with $V_i : \Omega \rightarrow V_i$ and a DAG \mathcal{G} with nodes $\{\tilde{V}_i\}_{i \in A}$ and the variable-node association $m : V_i \mapsto \tilde{V}_i$, \mathbb{P} is compatible with \mathcal{G} relative to m if for all $I, J, K \subset A$, $\tilde{V}_I \perp_{\tilde{\mathcal{G}}} \tilde{V}_J | \tilde{V}_K$ implies $V_I \perp_{\mathbb{P}} V_J | (V_K, \text{id}_{\mathcal{C}})$.

The following definition is reproduced from [Pearl \(2009\)](#) with the differences mentioned: notation has been matched to ours, the interventional model is assumed to be measurable and the interventional distributions are assumed to be defined on a common sample space.

Definition 1.1.10 (Interventional model). An interventional model is a tuple $(\mathbb{P}_{\mathcal{C}}, \Omega, (V_i)_{i \in A})$ where (Ω, \mathcal{F}) is a measurable space, $\mathcal{V} := (V_i)_{i \in A}$ a sequence of variables with $V_i : \Omega \rightarrow V_i$, V_i denumerable, where \mathcal{C} the choice set

$$\mathcal{C} := \{\text{do}_{\emptyset}\} \cup \{(\text{do}_B, v_B) | B \subset A, v_B \in \text{Range}(V_B)\}$$

That is, we take every subsequence \mathcal{V}_B of \mathcal{V} and add to \mathcal{C} every element of the range of \mathcal{V}_B , each labeled with the symbol do_B .

Definition 1.1.11 (Causal Bayesian network). Given an interventional model $(\mathbb{P}_{\mathcal{C}}, \Omega, (V_i)_{i \in A})$ and a directed acyclic graph $\tilde{\mathcal{G}}$ with nodes $\tilde{\mathcal{V}}$, $(\mathbb{P}_{\mathcal{C}}, \Omega, (V_i)_{i \in A}, \tilde{\mathcal{G}})$ is a *causal Bayesian network* with respect the node-variable association $m : \tilde{V}_i \mapsto V_i$ if:

1. \mathbb{P} . is compatible with $\tilde{\mathcal{G}}$ with respect to m
2. $B \neq \emptyset \implies \mathbb{P}_{(\text{do}_B, v_B)}^{\mathbf{V}_B} = \delta_{v_B}$
3. $\mathbb{P}_{(\text{do}_B, v_B)}^{\mathbf{V}_i | \text{Pa}(\mathbf{V}_i)} \stackrel{\cong}{=} \mathbb{P}_{\text{do}_\emptyset}^{\mathbf{V}_i | \text{Pa}(\mathbf{V}_i)}$ for all $i \notin B$

One might go about constructing a causal Bayesian network by starting with a joint probability distribution and then “adding in” the interventions. When constructing a causal Bayesian network this way, one has to make sure the chosen sequence of variables $(\mathbf{V}_i)_{i \in A}$ is “interventionally compatible”. In particular, we require $v_{\text{Pa}(i)} \mapsto \delta_{v_i}$ be a $\mathbf{V}_i | \text{Pa}(\mathbf{V}_i)$ -valid conditional for all i (Definition ??, or else the probability set for an intervention on \mathbf{V}_i will be empty).

For continuously valued variables, the ability to pick a version of the conditional probability for each intervention is problematic. Suppose $\tilde{\mathbf{V}}_i$ is a parent of $\tilde{\mathbf{V}}_j$, and the associated variable \mathbf{V}_i is continuously valued and $\mathbb{P}_{\text{do}_\emptyset}^{\mathbf{V}_i}(\{v_i\}) = 0$ for all singletons $v_i \in V_i$. Then for every intervention $\text{do}_{\{i\}}(v_i)$, we can choose a version of $\mathbb{P}_{\text{do}_\emptyset}^{\mathbf{V}_j | \mathbf{V}_i}$ that takes an arbitrary value at the point v_i (because this point has measure 0), so property (3) is satisfied trivially.

1.1.2 Unrolling a causal Bayesian network

Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and an independent and identically distributed (IID) sequence $\mathbf{X} := (\mathbf{X}_i)_{i \in [n]}$, it is common to “roll up” the joint distribution $\mathbb{P}^{\mathbf{X}} \in \Delta(X^n)$ to a single representative distribution $\mathbb{P}^{\mathbf{X}_0} \in \Delta(X)$ and say something like “the \mathbf{X}_i are IID according to $\mathbb{P}^{\mathbf{X}_0}$ ”. Because of the IID assumption, the full joint distribution $\mathbb{P}^{\mathbf{X}} \in \Delta(X^n)$ can be unambiguously reconstructed from a statement like this.

A causal Bayesian network is similarly a rolled-up representation of a model of some sequence of variables. Unlike an IID sequence, it isn’t completely unambiguous how to unroll it. We propose the following method: first, posit a sequence of variables $\mathbf{V} := (\mathbf{V}_{ij})_{i \in A, j \in [n]}$, and extend the set C to be the set of sequences of interventions

$$\{(\text{do}_{B_{ij}}(v_B))_{j \in [n]} | \forall j : B_j \subset A, v_{B_j} \in \text{Range}(\mathbf{V}_{B_j})\}$$

i.e. C now consists of all sequences of separate interventions to each subsequence of variables $\mathbf{V}_{A_j} := (\mathbf{V}_{ij})_{i \in A}$, understood to refer to variables arising from a particular iteration of the decision procedure.

Given a graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, we now have a collection of variable-node association maps $m_j : \{\mathbf{V}_{ij} | i \in A\} \rightarrow \tilde{\mathcal{V}}$ such that $m_j(\mathbf{V}_{ij}) = \tilde{\mathbf{V}}_i$.

We now need to specify how variables in an unrolled causal Bayesian network are distributed, given some sequence of interventions. By analogy with the original case of IID variables, we conclude that the $\mathbf{V}_{A_j} := (\mathbf{V}_{ij})_{i \in A}$ are mutually independent given any particular sequence of interventions. Furthermore, Definition 1.1.11 constrains the distribution of each variable given a particular sequence of interventions from C . For a sequence of interventions $\alpha \in C$, let $\pi_j(\alpha)$ be the j th intervention in the sequence. We might posit the following analogue of condition (3):

$$3' \quad \pi_j(\alpha) = (\text{do}_{B_j}, v_{B_j}) \text{ implies } \mathbb{P}_\alpha^{\mathbf{V}_{ij} | \text{Pa}(\mathbf{V}_{ij})} \stackrel{\cong}{=} \mathbb{P}_{\text{do}_\emptyset}^{\mathbf{V}_{i1} | \text{Pa}(\mathbf{V}_{i1})} \text{ for all } i \notin B$$

Where $\text{do}_{\mathcal{O}}^n$ is a sequence of n $\text{do}_{\mathcal{O}}$ interventions. This is a combination of an assumption that variables in the sequence are conditionally identically distributed given appropriate interventions and condition (3) from Definition 1.1.11. However, it's not quite satisfactory. Take $B := \text{Pa}(V_{i1})$, and suppose $\mathbb{P}_{\text{do}_{\mathcal{O}}^n}^{V_{B1}}(\{x\}) = 0$. Then (3') would be satisfied by a model for which

$$\begin{aligned}\mathbb{P}_{(\text{do}_{B_1}, x, \text{do}_{B_2}, x)}^{V_{i1}|B}(U|x) &= \delta_0(U) \\ \mathbb{P}_{(\text{do}_{B_1}, x, \text{do}_{B_2}, x)}^{V_{i2}|B}(U|x) &= \delta_1(U)\end{aligned}$$

that is, if the empty intervention is unsupported over some element of the range of a variable, then (3') allows models that assign different consequences to repetitions of the same intervention on this variable, if those intervention forces the variable into the region that originally had no support.

We propose instead the restricted assumption of identical response functions: for any pair V_{ij} and V_{ik} , unless i is intervened on by $\pi_j(\alpha)$ and not intervened on by $\pi_k(\alpha)$, then then the conditional probability of V_{ij} given its parents is equal to the conditional probability of V_{ik} given its parents. This is condition [3*].

In order to be able to “roll up” a sequence of interventions, we also require that the response to the j th intervention does not depend on any of the interventions other than the j th. If this were not the case, then even if the restricted assumption of identical response functions were satisfied, different sequences of interventions would “roll up” to different interventional models. Condition 4* is the formalisation of this requirement. In Chapter ??, we showed that conditionally independent and identical response functions allow for the estimation of conditional probabilities from previous data, but noted that this did not necessarily imply that estimating conditional probabilities under a particular fixed choice was sufficient for decision making. Condition 4* is the assumption that, once we have the interventional conditional probabilities for any sequence of interventions, nothing else is needed.

Condition 5* is the requirement that observations are mutually independent.

Definition 1.1.12 (Unrolled causal Bayesian network). Given an interventional model $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]})$ and a directed acyclic graph $\tilde{\mathcal{G}}$ with nodes $\tilde{\mathcal{V}}$, $(\mathbb{P}, C, \Omega, (V_i)_{i \in A}, \tilde{\mathcal{G}})$ is an *unrolled causal Bayesian network* with respect the node-variable association maps $m_j : \tilde{\mathcal{V}}_{ij} \mapsto V_i$ if, for all $j, k \in [n]$:

- 1* $\mathbb{P}^{V_{Aj}}$ is compatible with $\tilde{\mathcal{G}}$ with respect to m_j for all $j \in [n]$
- 2* $\pi_j(\alpha) = (\text{do}_{B_j}, v_{B_j})$ and $B_j \neq \emptyset$ implies $\mathbb{P}_{(\text{do}_{B_j}, v_{B_j})}^{V_{Bj}} = \delta_{v_{B_j}}$
- 3* If $\pi_j(\alpha) = (\text{do}_{B_j}, v_{B_j})$, $\pi_k(\alpha) = (\text{do}_{B_k}, v_{B_k})$ and $i \notin B_j \cup B_k$ then $\mathbb{P}_{\alpha}^{V_{ij}|\text{Pa}(V_{ij})} \stackrel{\mathbb{P}_{\alpha}}{\cong} \mathbb{P}_{\text{do}_{\mathcal{O}}}^{V_{ik}|\text{Pa}(V_{ik})}$
- 4* $\pi_j(\alpha) = \pi_j(\alpha')$ implies $\mathbb{P}_{\alpha}^{V_{Aj}} = \mathbb{P}_{\alpha'}^{V_{Aj}}$
- 5* $V_{Aj} \perp\!\!\!\perp_{\mathbb{P}_C}^e V_{A[n] \setminus \{j\}} | \text{id}_C$

1.1.3 Uncertainty in an unrolled causal Bayesian network

Condition 3* of Definition 1.1.12 establishes that, depending on the precise sequence of interventions chosen, certain conditionals are identical. In Chapter ??, we considered conditionals (or “response functions”) that were identical *conditional on some hypothesis* H . Problems

addressed with causal Bayesian networks are also usually problems where these conditional distributions are initially unknown (and, in some cases, they may remain unknown even after examining an arbitrarily large amount of data). We propose to use the same method to represent uncertainty over conditional distributions in a causal Bayesian network. In particular, an *uncertain unrolled causal Bayesian network* is an interventional model $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]})$ with some variable H such that, conditional on any $h \in H$, the result is an unrolled causal Bayesian network.

Definition 1.1.13 (Uncertain unrolled causal Bayesian network). Given an interventional model $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]})$ and a directed acyclic graph $\tilde{\mathcal{G}}, (\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]}, H, \tilde{\mathcal{G}})$ is an *uncertain unrolled causal Bayesian network* with respect to some variable $H : \Omega \rightarrow H$ if for each $h \in H$, defining $\mathbb{P}_{\cdot, h} := \alpha \mapsto \mathbb{P}_{\alpha}^{\text{id}_{\Omega}|H}(\cdot|h)$, $(\mathbb{P}_{\cdot, h}, C, \Omega, (V_i)_{i \in A}, \tilde{\mathcal{G}})$ is an unrolled causal Bayesian network.

Recalling the discussion in Section ??, Definition 1.1.13 associates each intervention with a unique probability distribution. One could imagine therefore calling uncertain unrolled causal Bayesian networks “Bayesian causal Bayesian networks”, although this is obviously a bit of a confusing name.

An uncertain unrolled causal Bayesian network is *almost* a conditionally independent and identical response function model. Due to 3*, such a model features conditionally independent and identical response functions wherever α consists of a sequence of interventions none of which target i . This leads us to the key result of this section: considering a subset of the interventions in C , an uncertain unrolled causal Bayesian network is IO contractible (with respect to some parameters) by application of Theorem ??.

Theorem 1.1.14 (IO contractibility of CBNs). *Given an uncertain unrolled causal Bayesian network $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in A, j \in [n]}, H, \tilde{\mathcal{G}})$, take $C' \subset C$ to be sequences of interventions that, for some $j \in [n]$, do not target a particular V_{ij} for any $i \in A$. Then $V_{i[n]} \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e \text{id}_C | (H, \text{Pa}(V_{i[n]}))$ and $\mathbb{P}_C^{V_{i[n]}|H\text{Pa}(V_{i[n]})}$ is IO contractible over H .*

Proof. First we will prove $V_{i[n]} \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e \text{id}_C | (H, \text{Pa}(V_{i[n]}))$. This is equivalent to the claim that $\mathbb{P}_{\alpha}^{V_{i[n]}|H\text{Pa}(V_{i[n]})}$ is the same as $\mathbb{P}_{\alpha'}^{V_{i[n]}|H\text{Pa}(V_{i[n]})}$ for any α, α' . By assumption 5* of Definition 1.1.12, for each $h \in H$

$$V_{Aj} \perp\!\!\!\perp_{\mathbb{P}_{\alpha, h}}^e V_{A[n] \setminus \{j\}} | \text{id}_C$$

which implies

$$\begin{aligned} V_{Aj} &\perp\!\!\!\perp_{\mathbb{P}_{\alpha}}^e V_{A[n] \setminus \{j\}} | (\text{id}_C, H) \\ \implies V_{ij} &\perp\!\!\!\perp_{\mathbb{P}_{\alpha}}^e V_{A[n] \setminus \{j\}} | (H, \text{Pa}(V_{i[n]}), \text{id}_C) \end{aligned} \tag{1.1}$$

thus it is sufficient to show that, for any $\alpha, \alpha' \in C'$ and $j \in [n]$

$$\mathbb{P}_{\alpha}^{V_{ij}|H\text{Pa}(V_{ij})} = \mathbb{P}_{\alpha'}^{V_{ij}|H\text{Pa}(V_{ij})}$$

By assumption, if $\pi_j(\alpha) =: (\text{do}_{B_j}, v_{B_j})$ and $\pi_j(\alpha') =: (\text{do}_{B'_j}, v'_{B'_j})$, $i \notin B_j \cup B'_j$, and similarly replacing the j s with k s for any $k \in [n]$. Define α'' such that, for some k , $\pi_k(\alpha'') = \pi_k(\alpha')$ and

$\pi_j(\alpha'') = \pi_j(\alpha)$. Then by 4*, for all $h \in H$

$$\begin{aligned}
\mathbb{P}_\alpha^{\mathbf{V}_{ij}|\text{HPa}(\mathbf{V}_{ij})}(A|h, y) &= \mathbb{P}_{\alpha''}^{\mathbf{V}_{ij}|\text{HPa}(\mathbf{V}_{ij})}(A|h, y) \\
&= \mathbb{P}_{\alpha''}^{\mathbf{V}_{ik}|\text{HPa}(\mathbf{V}_{ik})}(A|h, y) && \text{by 3*} \\
&= \mathbb{P}_{\alpha'}^{\mathbf{V}_{ik}|\text{HPa}(\mathbf{V}_{ik})}(A|h, y) && \text{by 4*} \\
&= \mathbb{P}_{\alpha'}^{\mathbf{V}_{ij}|\text{HPa}(\mathbf{V}_{ij})}(A|h, y) && \text{by 3*} \\
\implies \mathbb{P}_\alpha^{\mathbf{V}_{ij}|\text{HPa}(\mathbf{V}_{ij})} &= \mathbb{P}_{\alpha'}^{\mathbf{V}_{ij}|\text{HPa}(\mathbf{V}_{ij})}
\end{aligned}$$

Next, IO contractibility of $\mathbb{P}_C^{\mathbf{V}_{i[n]}|\text{HPa}(\mathbf{V}_{i[n]})}$ over H . By Eq. (1.1)

$$\mathbf{V}_{ij} \perp\!\!\!\perp_{\mathbb{P}_\alpha}^e (\mathbf{V}_{i[1,j]}, \text{Pa}(\mathbf{V}_{i[1,j]})) | (H, \text{Pa}(\mathbf{V}_{i[n]}), \text{id}_C)$$

furthermore, by 3* and the assumption that no intervention $\alpha \in C'$ targets \mathbf{V}_{ij} for any j , for any $\alpha \in C'$

$$\mathbb{P}_\alpha^{\mathbf{V}_{ij}|\text{HPa}(\mathbf{V}_{ij})}(A|h, y) = \mathbb{P}_\alpha^{\mathbf{V}_{ik}|\text{HPa}(\mathbf{V}_{ik})}(A|h, y)$$

thus \mathbb{P}_C has independent and identical response functions conditional on H and by Theorem ??, $\mathbb{P}_C^{\mathbf{V}_{i[n]}|\text{HPa}(\mathbf{V}_{i[n]})}$ is IO contractible over H . \square

1.1.4 Probabilistic Graphical Models

Lattimore and Rohde (2019a,b) have previously published work in which they demonstrated how to “unroll” causal Bayesian networks into what they call “Probabilistic Graphical Models”. Their work goes into more detail than this thesis on how identifiability results transfer from causal Bayesian networks to their unrolled forms.

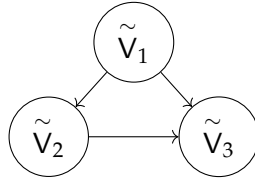
Illustrating the fact that some choices must be made in order to work out what kind of rolled-up model corresponds to a given causal Bayesian network, Rohde and Lattimore consider an unrolling where the empty intervention is always made in conjunction with a particular intervention on a particular node in the DAG. Where we explicitly write down a model of an entire sequence of observations, Probabilistic Graphical Models can be assumed to represent a sequence of an arbitrary number of empty interventions in conjunction with an arbitrary number of particular interventions on particular nodes in the DAG. Such compact representations are of course very useful when the extra details are redundant. The difference underscores the approach taken to causal modelling in this thesis – we proceed cautiously, aiming to explicitly represent all relevant assumptions that go into building a particular type of causal model, and approach that can lead to relatively verbose model definitions and representations.

Precisely, a probabilistic graphical model is a map \mathbb{P} . from the set of single-node interventions C to probability distributions \mathbb{P}_α defined on (Ω, \mathcal{F}) . A probabilistic graphical model is typically associated with a causal Bayesian network $(\mathbf{Q}, C, \Omega', (\mathbf{V}_i)_{i \in A}, \tilde{\mathcal{G}})$ where, for each $\mathbf{V}_i : \Omega \rightarrow V_i$ in the original causal Bayesian network, two variables \mathbf{V}_i and \mathbf{V}_i^* are defined on (Ω, \mathcal{F}) . The probabilistic graphical model also adds a “parameter” W_i for each variable pair $(\mathbf{V}_i, \mathbf{V}_i^*)$ such that, taking C' to be interventions not targeting \mathbf{V}_i^* , for any $\alpha \in C'$, $\mathbb{P}_\alpha^{\mathbf{V}_i|W_i\text{Pa}(\mathbf{V}_i)} = \mathbb{P}_\alpha^{\mathbf{V}_i^*|W_i\text{Pa}(\mathbf{V}_i^*)}$ and $\mathbf{V}_i \perp\!\!\!\perp_{\mathbb{P}_{C'}}^e (\mathbf{V}_A^*, \text{id}_C) | (W_i)$ (where parents are assessed relative to the graph $\tilde{\mathcal{G}}$). This should look familiar - it is specifying, in a very similar manner to

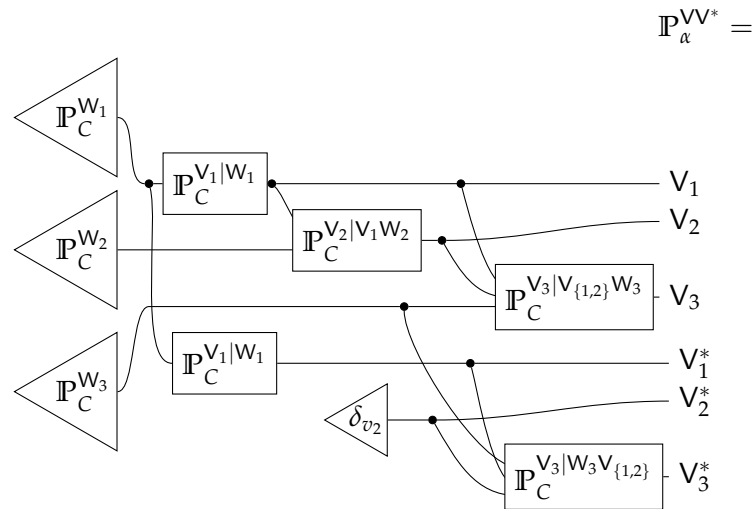
Theorem 1.1.14, that a Probabilistic Graphical Model constructed from a causal Bayesian network $(\mathbf{Q}, \mathbf{C}, \Omega', (V_i)_{i \in A}, \tilde{\mathcal{G}})$ features independent and identical response functions for each node given its parents conditional on the parameter W_i .

A depiction of probabilistic graphical models and uncertain unrolled causal Bayesian networks using string diagrams gives some intuition regarding the structure of these different types of models, as well as some of the “off-page” assumptions of ordinary causal Bayesian networks.

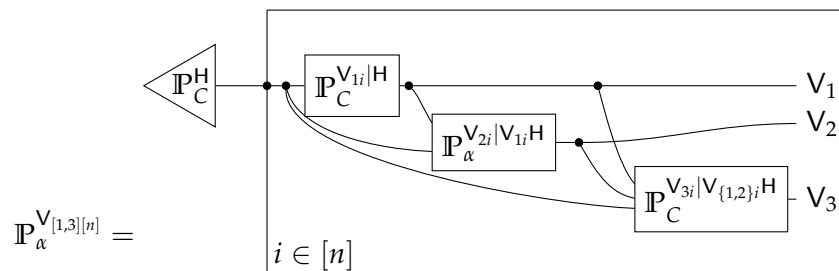
Here is the original graph $\tilde{\mathcal{G}}$ associated with $(\mathbf{Q}, \mathbf{C}, \Omega', (V_i)_{i \in A}, \tilde{\mathcal{G}})$:



Here is the probabilistic graphical model associated with the intervention (do_2, v_2)



and here is the uncertain unrolled CBN associated with the restricted set of interventions \mathbf{C}' that consists of, for each element of the sequence, either the empty intervention or some intervention targeting V_2



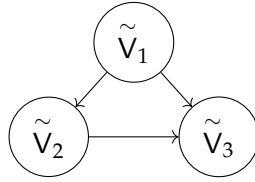
where

$$\mathbb{P}_\alpha^{V_{2i}|V_{1i}H} = \begin{cases} \delta_v & \pi_i(\alpha) = (\text{do}_2, v) \\ \mathbb{P}_{\text{mathrm{do}_\emptyset}^{V_{2i}|V_{1i}H}} & \text{otherwise} \end{cases}$$

1.1.5 Hidden confounders and precedents

One of the particularly interesting questions in causal inference is how to infer consequences of actions from observational data. A particular question of interest for problems of this type is the question of what kinds of inductive assumptions are applicable to this problem.

A common assumption in the causal Bayesian network tradition applicable to this kind of problem is the assumption of *hidden confounders*. Suppose we have an uncertain unrolled causal Bayesian network $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in [3], j \in [n]}, H, \tilde{\mathcal{G}})$ where the graph G is as follows:



and we consider the subset $C' \subset C$ of interventions that are either empty or target V_2 only. We note that Theorem 1.1.14 implies that $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ is IO contractible, but not $\mathbb{P}_C^{V_{3[n]}|HV_{2[n]}}$. We can specify somewhat informally that $V_{1[n]}$ is not observed – that is, it is not associated with a measurement procedure.

The assumption of a hidden confounder often implies that, for any choice, the consequences of “interventions” have been anticipated by *some* fraction of the observations. Specifically, the IO contractibility of $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ implies that $\mathbb{P}_C^{V_{3[n]}|HV_{2[n]}}$ is unchanged by swaps that leave $V_{1[n]}$ unchanged.

Theorem 1.1.15. *Given $(\mathbb{P}, C, \Omega, (V_{ij})_{i \in [3], j \in [n]}, H, \tilde{\mathcal{G}})$ with $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ IO contractible over H , V_i discrete for all $i \in [3]$ and $\mathbb{P}_C^{V_{2i}|H}(v_1, v_2|h) > 0$ for all v_1, v_2, h , let $Q : \Omega \rightarrow [n]^n$ be a permutation of $[n]$ such that $V_{1[n]} = V_{1Q([n])}$. Then*

$$\mathbb{P}_C^{V_{3[n]}|HV_{2[n]}} = \mathbb{P}_C^{V_{3Q([n])}|HV_{2Q([n])}}$$

Proof. By IO contractibility of $\mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}}$ over H

$$\begin{aligned} \mathbb{P}_C^{V_{3[n]}|HV_{1[n]}V_{2[n]}} &= \mathbb{P}_C^{V_{3Q([n])}|HV_{1Q([n])}V_{2Q([n])}} \\ &= \mathbb{P}_C^{V_{3Q([n])}|HV_{1[n]}V_{2Q([n])}} \end{aligned}$$

To formalise this, we say there is a variable D that is like an “action” at the $i = 0$ index in the sense that a choice α that leads to a distribution of actions D_0 identical to some mixture of other choices α' and α'' induces consequences equal to the same mixture of α' and α'' . We require that the model is IO contractible over the sequence of actions $(D_i)_{-i \in \{0\} \cup \mathbb{N}}$, and that each action has positive probability in the “historical” indices $i < 0$. The reason that this assumption is weaker than IO contractibility is that D_i for $i < 0$ is unobserved – that is, rules for choosing the action D_0 cannot depend on D_i for $i < 0$.

Definition 1.1.16 (Precedent). Given a probability set \mathbb{P}_C on (Ω, \mathcal{F}) and variables $Y := (Y_i)_{-i \in \{0\} \cup \mathbb{N}}$ and $D := (D_i)_{-i \in \{0\} \cup \mathbb{N}}$, D discrete, we say (\mathbb{P}_C, D, Y) is preceded if for directing random measure H , $\mathbb{P}_\alpha^{Y_i | H D_i}$ are independent and identical responses for all i ,

$$\begin{aligned} \mathbb{P}_\alpha^{D_0 | Y_{-\mathbb{N}}} &= a \mathbb{P}_{\alpha'}^{D_0 | Y_{-\mathbb{N}}} + b \mathbb{P}_{\alpha''}^{D_0 | Y_{-\mathbb{N}}} \\ \implies \mathbb{P}_\alpha &= a \mathbb{P}_{\alpha'} + b \mathbb{P}_{\alpha''} \end{aligned}$$

as well as $\mathbb{P}_\alpha^{D_{-N}}$ is exchangeable and, defining G to be the directing random measure of $(\mathbb{P}_C, *, D)$, $\mathbb{P}_\alpha^{D_i | G}(\alpha | g) > 0$ for all $\alpha \in C$, almost all $g \in G$, $i < 0$.

The assumption of preemption as given in Definition 1.1.16 can, under some side conditions, yield nontrivial conclusions. Theorem 1.1.18 comes with a lot of complicated conditions, so it is worth explaining with an example first.

Suppose we have a collection of doctors Z_i who each see a series of patients and offer a treatment X_i and report their results Y_i . Each doctor may decide whether or not to prescribe based on any number of unobserved factors, and may offer additional unrecorded treatments, vary in their bedside manner and so forth, and there may be stochastic variation in any of these. The decision maker is also a doctor, and is reviewing the data contained in the sequences $(X_i, Y_i, Z_i)_{i \in [n]}$. The decision maker supposes that whatever overall treatment plan they adopt (which could also depend on features not listed in this set of variables), the same thing has probably been done at least sometimes by some of these prior doctors – that is, their treatment protocol is preempted. They also assume that the doctor’s identity has no bearing on outcomes over and above the treatment protocol, they assume that doctors don’t all select the same mixture of treatment protocols. It then stands to reason that the doctors who choose different treatment plans will see slightly different results *if the different treatment plans actually lead to different results*. Conversely, if there is *no* variation in results after conditioning on whether patients were treated this suggests that whether or not treatment occurred is the *only* important feature of any treatment plan.

One way that this story could fail is if the doctors all knew exactly the long-run probabilistic outcomes of different treatment plans and coordinated with one another, they could (in principle) each pick different mixtures of treatment plans just so that the variation in outcomes is masked – that is, for example, doctor 1 picks a medium effectiveness plan 100% of the time, while doctor 2 picks a highly effective plan 50% of the time and a low effectiveness plan 50% of the time leading to the same distribution over outcomes. The conditions in Theorem 1.1.18 requiring domination by the uniform measure on $[0, 1]$ are assumptions that this kind of thing does not happen, either because the doctors don’t coordinate or because, even if they did coordinate, they would not know the long-run averages of outcomes associated with each plan precisely enough to completely mask the variation.

Nxample 1.1.17 (Matrix notation). Given a sequential input-output model (\mathbb{P}_C, D, Y) with D, Y discrete, the directing random measure H takes values in the set of Markov kernels $D \rightarrow Y$, which can be identified with a subset of matrices in $\mathbb{R}^{|D| \times |Y|}$. We can therefore refer

to elements of H as matrices $(h_d^y)_{d \in D, y \in Y}$ with $\sum_{y \in Y} h_d^y = 1$ for all d , and $h_d^y \stackrel{\mathbb{P}_\alpha}{\cong} \mathbb{P}_\alpha^{Y|HD}(y|h, d)$. We can also define H_d^y as the d, y -th projection of H .

Theorem 1.1.18. *Suppose we have a probability set \mathbb{P}_C on (Ω, \mathcal{F}) with variables $Y := (Y_i)_{-i \in \{0\} \cup \mathbb{N}}$, $D := (D_i)_{-i \in \{0\} \cup \mathbb{N}}$, $X := (X_i)_{-i \in \{0\} \cup \mathbb{N}}$ and $Z_i := (Z_i)_{-i \in \mathbb{N}}$, with D, X, Y, Z discrete. Suppose further that $(\mathbb{P}_C, (D, Z), (X, Y))$ is preempted. Let G be the directing random conditional of (\mathbb{P}_C, Z, D) and H the directing random conditional of $(\mathbb{P}_C, (D, Z), (X, Y))$. Suppose for all i , $Y_i \perp\!\!\!\perp_{\mathbb{P}_C} Z_i | (H, X_i, id_C)$. Finally suppose for each $h \in H$ and $d \in D$ and $z, z' \in Z$,*

$$\mathbb{P}_\alpha^{G_z^d | HG_{z'}^d}(\cdot | h, g_{z'}^d) \ll U_{[0,1]} \quad (1.2)$$

where $U_{[0,1]}$ is the uniform probability measure on $([0,1], \mathcal{B}([0,1]))$ – that is, the Lebesgue measure on $[0,1]$ restricted to the Borel sets.

Then $\mathbb{P}_\alpha^{Y_i | X_i, H}$ are independent and identical responses for all $-i \in \{0\} \cup \mathbb{N}$.

Proof. First, define matrices k and l by

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i | HD_i Z_i X_i}(y|h, d, z, x) &\stackrel{\mathbb{P}_\alpha}{\cong} k_{dzz}^y \\ \mathbb{P}_\alpha^{X_i | HD_i Z_i}(x|h, d, z) &\stackrel{\mathbb{P}_\alpha}{\cong} l_{dz}^x \end{aligned}$$

noting that both k and l are almost surely deterministic functions of h .

The assumption $Y_i \perp\!\!\!\perp_{\mathbb{P}_C} Z_i | (H, X_i, id_C)$ implies, for \mathbb{P}_α -almost all $k, l, \alpha, z, z', x, y$

$$\sum_{d \in D} k_{dzz}^y \frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d'z}^x g_{d'}^d} = \sum_{d \in D} k_{dzz}^y \frac{l_{dz}^x g_{z'}^d}{\sum_{d' \in D} l_{d'z'}^x g_{d'}^d} \quad (1.3)$$

Fixing k, l and g_z^d , Eq. (1.3) defines a polynomial constraint on $g_{z'}^d$. We will show that, unless $k_{dzz}^y = k_{d'zz}^y$ for all d, d' and z then this constraint is nontrivial for some z' . For arbitrary d , without loss of generality, assume $k_{dzz}^y > k_{d'zz}^y$ for some $d <$.

Then either $l_{dz'}^x = l_{d'z'}^x$, $l_{dz'}^x < l_{d'z'}^x$ or $l_{dz'}^x = l_{d'z'}^x$. Consider the first case, and take g' such that $g'd_{z'} = g_{z'}^d - \epsilon$ and $g'd_{z'}^< = g_{z'}^d + \epsilon$ and equal to g otherwise. There is almost surely some ϵ such that g' is a Markov kernel, as $g_{z'}^d > 0$ almost surely. Then

$$\begin{aligned} \frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d'z}^x g_{d'}^d} &> \frac{l_{dz}^x g_{z'}^d}{\sum_{d' \in D} l_{d'z}^x g_{d'}^d} \\ \frac{l_{d'z}^x g_z^d}{\sum_{d' \in D} l_{d'z}^x g_{d'}^d} &< \frac{l_{d'z}^x g_{z'}^d}{\sum_{d' \in D} l_{d'z}^x g_{d'}^d} \end{aligned}$$

because by assumption the denominator remains the same. But then

$$\sum_{d \in D} k_{dzz}^y \frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d'z}^x g_{d'}^d} > \sum_{d \in D} k_{dzz}^y \frac{l_{dz}^x g_{z'}^d}{\sum_{d' \in D} l_{d'z}^x g_{d'}^d} \quad (1.4)$$

because on the left side a larger term in the sum receives more weight, a smaller term receives less weight and all other terms are weighted equally.

Consider $l_{dz'}^x > l_{d< z'}^x$. Then we still have

$$\frac{l_{dz}^x g_z^d}{\sum_{d' \in D} l_{d'z}^x g_z^{d'}} > \frac{l_{dz}^x g_z^{d'}}{\sum_{d' \in D} l_{d'z}^x g_z^{d'}}$$

$$\frac{l_{d< z}^x g_z^{d'}}{\sum_{d' \in D} l_{d'z}^x g_z^{d'}} < \frac{l_{d< z}^x g_z^{d'}}{\sum_{d' \in D} l_{d'z}^x g_z^{d'}}$$

For the first inequality, both the numerator and the denominator shrink. For the second, note that

$$\frac{l_{d< z}^x g_z^{d'}}{\sum_{d' \in D} l_{d'z}^x g_z^{d'}} = \frac{l_{d< z}^x (g_z^{d'} + \epsilon)}{(1 + \frac{\epsilon}{g_z^{d'}}) \sum_{d' \in D} l_{d'z}^x g_z^{d'}}$$

$$< \frac{l_{d< z}^x (g_z^{d'} + \epsilon)}{\sum_{d' \in D} l_{d'z}^x g_z^{d'}}$$

and so the conclusion in Eq. (1.4) follows for the same reasons. Finally considering $l_{dz'}^x < l_{d< z'}^x$, analogous reasoning implies Eq. (1.4) once again.

Thus, unless $k_{dzz}^y = k_{d'zx}^y$ for all d, d' and z , Eq. (1.3) implies a nontrivial constraint on $g_{z'}^d$ for some z' . Thus, for some z', x, d, d', y the set of solutions $A := \{g_{z'}^d | \text{Eq. (1.3) is satisfied} \wedge k_{dzz}^y \neq k_{d'zx}^y\}$ has Lebesgue measure 0 in the set $[0, 1]^D$ (Okamoto, 1973), and so finally

$$\mathbb{P}_\alpha^{\mathbf{G}_z^d | \mathbf{H}\mathbf{G}_{z'}^d}(A | h, g_z^d) = 0$$

by the assumption that this probability is dominated by the Lebesgue measure. On the other hand, by assumption, the set $B := \{g_{z'}^d | \text{Eq. (1.3) is satisfied}\}$ has measure 1. Thus we conclude that $k_{dzz}^y = k_{d'zx}^y$ with probability 1. That is, $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e D_i | (H, Z_i, X_i, \text{id}_C)$. By contraction with $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Z_i | (H, X_i, \text{id}_C)$, we have $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (Z_i, D_i) | (H, X_i, \text{id}_C)$. Because we already know $\mathbb{P}_\alpha^{Y_i | \mathbf{H}\mathbf{X}_i Z_i}$ are identical for all i , this implies $\mathbb{P}_\alpha^{Y_i | \mathbf{H}\mathbf{X}_i}$ are independent and identical response functions. \square

We will now specify the medical example in more detail

Example 1.1.19. Let $Y := (Y_i)_{-i \in \{0\} \cup \mathbb{N}}$ be associated with patient outcomes, $D := (D_i)_{-i \in \{0\} \cup \mathbb{N}}$ with treatment plans (including, for example, what assessments are made, what other treatments are used and so forth), some of which the decision maker can consider, $X := (X_i)_{-i \in \{0\} \cup \mathbb{N}}$ to be the recommendation of a particular treatment and $Z_i := (Z_i)_{-i \in \mathbb{N}}$ to be other doctor's identifiers in the dataset.

Assume D_i screens off Z_i from X_i and Y_i – that is, the latent treatment plan is sufficiently detailed to screen off the relevance of the doctor's identity (this is a causal assumption), and the stochastic response to treatment plans is identical for each patient with positive support in the past data for each plan the decision maker is considering. That is, assume $(\mathbb{P}_C, D, (X, Y))$ is preempted. By the assumption that D_i screens off Z_i , we can also conclude that $(\mathbb{P}_C, (D, Z), (X, Y))$ is preempted.

Suppose that each doctor makes choices \mathbf{G}_z^d by some deterministic function of their beliefs of what every other doctor does $\mathbf{G}_{z'}^e$ and of the effect of treatment plans \mathbf{H} , but they estimate both $\mathbf{G}_{z'}^e$ and \mathbf{H} with continuously distributed noise. Then their beliefs, and hence (by supposition) their choices end up dominated by the uniform measure on $[0, 1]$.

If the decision maker is then told by an oracle that for all i , $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e Z_i | (H, X_i, \text{id}_C)$, she may then conclude that $\mathbb{P}_\alpha^{Y_i | H X_i}$ are independent and identical responses for all $-i \in \{0\} \cup \mathbb{N}$.

This story makes a number of assumptions with a causal character. First, the assumption of preemption, which we’ve acknowledged isn’t perfectly worked out, is taken here quite literally – we are not just supposing that the problem can be posed “as if” the consequences of our choices had been previously realised, but we are actually taking the D_i s to literally stand for unobserved choices that actual doctors make. Secondly, the assumption that the unobserved D_i s screen off the doctor’s identities is reminiscent of the idea from the causal graphical models tradition that variables have complete sets of causal parents, some of which may be unobserved, but all of which together screen that variable off from other nondescendant variables. Note that this assumption isn’t required by Theorem 1.1.18, but it supports the particular story being told here.

Finally, the idea that the long run frequencies of each doctor’s choices are generic conditioned on the other doctor’s choices and the long run input-output relationship seems closely related to the idea of “independent causal mechanisms”. This comes up in two forms in the literature. First, it is used to justify the assumption of *causal faithfulness*: here, it is shown that *if* one makes an assumption that conditional probabilities in a causal model are generic with respect to one another in a manner similar to Equation (1.2), then causal faithfulness holds with probability 1 (Meek, 1995). However, it’s been noted that conditional probabilities routinely do line up in “non-generic” ways in an anti-causal direction.

Interestingly, Theorem 1.1.18 itself doesn’t depend on a notion of causal direction, and merely shows that a conclusion of independent and identical response functions follows from an assumption of preemption and an assumption of “generic mechanism association”. Example 1.1.19 shows one way that this generic association could be argued for.

Note that in that example, there is no reason not to expect that each doctor doesn’t select a mixture of treatment plans that is *close* to having support at a special singleton – in fact, it is assumed that doctors try to take into account the response of patient outcomes to treatment plans and the actions of other doctors, and that they simply fail to do this perfectly. We also cheat by having an oracle tell the decision maker that the key conditional independence holds. In particular, we ask the decision maker to conclude something precise about H (namely, the key conditional independence) while also assuming that none of the other doctors are able to do this.

There is a substantial literature that aims to draw causal conclusions from observational data by first applying a graph learning algorithm to a sequence of observational data, and then using the graph obtained as a DAG for a causal Bayesian network. Earlier examples treat the graph learning problem as a discrete optimization problem and include the PC algorithm and the Causal Inference Algorithm Spirtes et al. (2000, Ch. 5& 6) and Greedy Equivalent Search Chickering (2002, 2003). More recent examples pose graph learning as a continuous optimization problem Ng et al. (2019); Zheng et al. (2018). Underpinning all of these approaches are a number of key assumptions, which include the assumption of *faithfulness* – that missing edges in the learned graph correspond to missing edges in the appropriate causal DAG – and often also the assumption of *causal sufficiency*, which is the assumption that there are “no relevant unobserved variables”. Together, these assumptions imply that certain conditional independences in the observational data sequence imply the same conditional independences in the data produced under intervention. One open question we raise is: can this implication also be understood as a special case of the interventional data being preempted by the observational data?

On unobservability

The fact that we are offering the assumption that covariates are unobservable as an informal assumption is due to the fact that we are limiting our attention to data-independent models (recall Definition ??). In these models, actions never depend on the available data, and choosing some action based on observations must happen outside the model. If we were considering some data-dependent variation of a causal Bayesian network, the fact that $V_{1[n]}$ is unobserved would have formal implications for our model. For example, if V_{1i} is unobserved for all i while V_{2i} is directly controlled for all i , then we should require that, under every choice α , V_{2i} is independent of $V_{1[1,i]}$ given $V_{\{2,3\}[1,i]}$ – that is, there is no choice that induces the controlled variable V_{2i} to be dependent on the history of unobserved variables $V_{1[1,i]}$, given the history of the observed variables.

1.2 What is a Potential Outcomes model?

Potential outcomes is another popular framework for modelling causal problems. There are two key differences between the potential outcomes approach and the causal Bayesian network approach: potential outcomes models are “unrolled by default” and they feature no notion of “intervention”. A third difference relates to the possibility of expressing “counterfactual” statements, although this difference seems to be contingent on the particular manner we use to unroll a causal Bayesian network - see Section ??, and recall from Section 1.1.2 that we had to make some choices in our construction of unrolled causal Bayesian networks, Definition 1.1.12.

Thus, to formulate a decision making model from a potential outcomes model, we do have to make a judgement about what the “choices” are (while CBNs provide the notion of “intervention” for this role), while we do not need to make any judgements about how to unroll a potential outcomes model, because this is already given. For the following, we rely on Rubin (2005) for the definition of a potential outcomes model.

Our definition of potential outcomes has a lot in common with the tabulated conditional distribution (Definition ??). However, it is different: in particular, $\mathbb{P}_\alpha^{Y_i|Y_i^D D_i}(y^d|y^D, d) = 1$, which is usually false for Definition ??.

Definition 1.2.1 (Potential outcomes). Given $(\mathbb{P}_C, \Omega, \mathcal{F})$ and, for some i , variables $D_i : \Omega \rightarrow D$ (D denumerable), $Y_i : \Omega \rightarrow Y$ and $Y_i^D : \Omega \rightarrow Y^D$, Y_i^D is a vector of *potential outcomes* with respect to D_i for all α

$$\mathbb{P}_\alpha^{Y_i|Y_i^D D_i} = \begin{array}{c} Y_i^D \\ \text{---} \end{array} \begin{array}{c} \text{---} \\ \text{---} \end{array} \boxed{\mathbb{F}_{\text{lus}}} \text{---} Y_i$$

Where \mathbb{F}_{lus} is the Markov kernel associated with the single-shot lookup map

$$\begin{aligned} \text{lus} : Y^D \times D &\rightarrow Y \\ (d, (y_i)_{i \in D}) &\mapsto y_d \end{aligned}$$

Note that $|D|$ copies of Y_i ($Y_i, Y_i, Y_i, \dots, Y_i$) always satisfies Definition 1.2.1. This definition is not the sole constraint on potential outcomes, but the additional constraints come from what we want them to model, and are therefore not able to be formally stated.

A “potential outcomes model” is simply a probability map with potential outcomes. Traditionally, potential outcomes models did not feature any choices. That is, a “traditional”

potential outcomes model is a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ rather than a probability function. We extend this to probability functions in what we think is the obvious way - to replace the probability space with a probability function space.

Definition 1.2.2 (Potential outcomes model). $(\mathbb{P}_C, \Omega, \mathcal{F})$ is a potential outcomes model with respect to $Y^D := (Y_i^D)_{i \in A}$, $Y := (Y_i)_{i \in A}$ and $(D_i)_{i \in A}$ if Y_i^D is a vector of potential outcomes with respect to D_i and Y_i for all $i \in A$.

Theorem 1.2.3. A potential outcomes model $(\mathbb{P}_C, \Omega, \mathcal{F})$ with respect to $D_i : \Omega \rightarrow D$, $Y_i : \Omega \rightarrow Y$ and $Y_i^D : \Omega \rightarrow Y^D$, Y_i^D has $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e \text{id}_C | (D, Y^D)$ and $\mathbb{P}_C^{Y|Y^D D}$ is IO contractible (with respect to $*$).

Proof. IO contractibility of $\mathbb{P}_C^{Y|Y^D D}$ follows from the fact that Y_i is deterministic given Y_i^D and D_i , and thus $Y_i \perp\!\!\!\perp_{\mathbb{P}_C}^e (D_{\{i\}^c}, Y_{\{i\}^c}, \text{id}_C) | (Y_i^D, D_i)$. Furthermore, for all i, j

$$\mathbb{P}_C^{Y_i|Y_i^D D_i} = \mathbb{P}_C^{Y_j|Y_j^D D_j}$$

hence the $\mathbb{P}_C^{Y_i|Y_i^D D_i}$ are independent and identical response functions conditional on $*$.

From Definition 1.2.1, $\mathbb{P}_\alpha^{Y_i|Y_i^D D_i}$ is the same for all $\alpha \in C$, and by the argument above,

$$\mathbb{P}_C^{Y_i|Y_i^D D_i Y_{\{i\}^c}^D D_{\{i\}^c}} = \mathbb{P}_C^{Y_i|Y_i^D D_i} \otimes \text{Del}_{Y|D \times A \setminus \{i\} \times D|A|}$$

hence

$$\mathbb{P}_C^{Y|Y^D D} = \bigotimes_{i \in A} \mathbb{P}_C^{Y_i|Y_i^D D_i}$$

hence $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e \text{id}_C | (D, Y^D)$. □

A key theorem of potential outcomes is that, if D is “strongly ignorable” with respect to Y^D , then the average treatment effect is identified. “Strong ignorability” here means that the probability $\mathbb{P}_\alpha^{D_i}(d) > 0$ for each d and for each choice α the inputs D are independent of the potential outcomes Y^D given the covariates and the choice. We reproduce this theorem in terms of IO contractibility. Note that Theorem 1.2.4 applies to potential outcomes models with sets of choices, rather than simply to single probability distributions.

Theorem 1.2.4 (Potential outcomes identifiability). If $(\mathbb{P}_C, \Omega, \mathcal{F})$ is a potential outcomes model with respect to $Y^D := (Y_i^D)_{i \in \mathbb{N}}$, $Y := (Y_i)_{i \in \mathbb{N}}$ and $(D_i)_{i \in \mathbb{N}}$, each value of D occurs infinitely often with probability 1, there is some $X := (X_i)_{i \in \mathbb{N}}$ such that $\mathbb{P}_\alpha^{Y^D X}$ is exchangeable for all α and $D \perp\!\!\!\perp_{\mathbb{P}_C}^e Y^D | (X, Y, \text{id}_C)$ and for each α $\mathbb{P}_\alpha^{D_i}$ is absolutely continuous with respect to some exchangeable distribution in $\Delta(D^{\mathbb{N}})$, then there is some W such that for all α $\mathbb{P}_\alpha^{Y|W X D}$ is IO contractible over W .

Proof. By exchangeability of $\mathbb{P}_\alpha^{Y^D X}$, $\mathbb{P}_\alpha^{Y^D | X}$ commutes with exchange. Because Y is deterministic given D and Y^D , $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e (X, \text{id}_C) | (Y^D, D)$. Thus, for some finite permutation ρ , by IO

contractibility of $\mathbb{P}_C^{Y|Y^D D}$

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y|XD} &= \begin{array}{c} \begin{array}{c} X \\ D \end{array} \begin{array}{|c} \hline \mathbb{P}_\alpha^{Y^D|X} \\ \hline \end{array} \begin{array}{|c} \hline \mathbb{P}_C^{Y|Y^D D} \\ \hline \end{array} \begin{array}{c} \text{---} Y \end{array} \end{array} \\
 &= \begin{array}{c} \begin{array}{c} X \\ D \end{array} \begin{array}{|c} \hline \mathbb{P}_\alpha^{Y^D|X} \\ \hline \end{array} \begin{array}{|c} \hline \text{Swap}_\rho \\ \hline \end{array} \begin{array}{|c} \hline \mathbb{P}_C^{Y|Y^D D} \\ \hline \end{array} \begin{array}{|c} \hline \text{Swap}_{\rho^{-1}} \\ \hline \end{array} \begin{array}{c} \text{---} Y \end{array} \end{array} \\
 &= \begin{array}{c} \begin{array}{c} X \\ D \end{array} \begin{array}{|c} \hline \text{Swap}_\rho \\ \hline \end{array} \begin{array}{|c} \hline \mathbb{P}_\alpha^{Y^D|X} \\ \hline \end{array} \begin{array}{|c} \hline \mathbb{P}_C^{Y|Y^D D} \\ \hline \end{array} \begin{array}{|c} \hline \text{Swap}_{\rho^{-1}} \\ \hline \end{array} \begin{array}{c} \text{---} Y \end{array} \end{array}
 \end{aligned}$$

IO contractibility of $\mathbb{P}_\alpha^{Y|WXD}$ over some W follows from Theorem ??.

□

1.3 Individual-level response functions

Exchangeability of potential outcomes, a key assumption in Theorem 1.2.4, is hard to explain in terms of symmetries of experiments. Given some experiment producing a sequence of pairs $(D_i, Y_i)_{i \in \mathbb{N}}$, say where D_i s are treatment administrations and Y_i s are health outcomes, there's no obvious generic way to design a related experiment whose model is the same as the original except with potential outcomes Y_i^D interchanged. This is in sharp contrast to the assumption of exchangeability of observed outcomes - say, instead of the potential outcomes being exchangeable, we hold that the pairs (D_i, Y_i) are exchangeable in the original experiment. Then we commit ourselves to the proposition that an alternative experiment which proceeds exactly as the first except, before being “committed to memory”, the experimental results are interchanged should be modeled exactly as the first.

One could propose that exchangeability of potential outcomes in our example experiment corresponds to an *exchangeability of patients*; perhaps, if we believe the model should be unchanged after we swap the order in which patients are seen, then we should accept that the model has exchangeable potential outcomes. First, note that this isn't a generic transformation like swapping labels – it depends on the experiment featuring a sequence of patients who can be swapped. Secondly, this proposition depends on some assumption that ties patients to potential outcomes. For example, if each patient were assumed to have a fixed but unknown vector of potential outcomes that is unchanged by the swapping operation, then swapping patients does indeed correspond to swapping potential outcomes.

We formalise the idea of “potential outcomes attached to individuals” as *individual-level response functions*. We offer a formal definition of the assumption of individual-level response functions, but like exchangeability of potential outcomes it is difficult to understand fundamentally what this assumption entails, or what it might be motivated by. Nevertheless, it does allow us to separate the assumption of “exchangeable potential outcomes” into the assumption of individual level response functions and the assumption of exchangeability of individuals. We also use this notion to prove Theorem 1.3.8. At a high level, it plays a similar role to Theorem 1.2.4: it seems to justify causal identifiability in certain kinds of controlled experiments. However, the content of the two theorems is very different. While Theorem 1.2.4 concerns independence of the inputs and potential outcomes along with exchangeability of the potential outcomes, Theorem 1.3.8 says (informally) if:

- There are individual-level response functions
- Individuals can be swapped without meaningfully altering the experiment

- Inputs are deterministically controlled by the choice
- There is only one choice for each value of the inputs

then the model is also IO contractible with respect to the inputs and the outputs only (ignoring the individual identifiers). In our view, this comes closer to a set of assumptions that are directly applicable to a controlled experiment than those in Theorem 1.2.4, and reflects Kasy (2016)’s dictum that, for the identifiability of causal effects, a “controlled experiment” is sufficient.

1.3.1 References to individual-level IO contractibility

The role of individuals has often been mentioned in literature on causal inference. For example, Greenland and Robins (1986) explain

Equivalence of response type may be thought of in terms of exchangeability of individuals: if the exposure states of the two individuals had been exchanged, the same data distribution would have resulted.

Here, the idea of “exchangeable individuals” plays a role in the author’s reasoning about model construction, but “individuals” are not actually referenced by the resulting model, and “exchanging individuals” does not correspond to a model transformation.

Dawid (2020) suggests (with some qualifications) that “post-treatment exchangeability” for a decision problem regarding taking aspirin to treat a headache may be acceptable if the data are from

A group of individuals whom I can regard, in an intuitive sense, as similar to myself, with headaches similar to my own.

As in the previous work, the similarity of individuals involved in an experiment is raised when justifying particular model constructions, but the individuals are not referenced by the model.

Pearl (2009, pg. 98) writes

Although the term unit in the potential-outcome literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterize that individual, the experimental conditions under study, the time of day, and so on – all of which are represented as components of the vector u in structural modeling.

Once again, the idea of an individual (or a particular set of conditions) is raised in the context of explaining modelling choices. Unlike the previous authors, Pearl introduces a vector u to stand for the “unit”. However, he subsequently assumes that u is a sequence of *independent samples* from some distribution. This seems to contradict an important feature of “individuals” or “units”: individuals are typically supposed to be unique, a property that will usually not be satisfied by independently sampling from some distribution (at least, as long as the distribution is discrete).

Finally, Rubin (2005) writes:

Here there are N units, which are physical objects at particular points in time (e.g., plots of land, individual people, one person at repeated points in time).

Note that Rubin’s explanation of *units* guarantees that they are unique: they are particular things at particular times. These units are associated with input-output functions (the *potential outcomes*), which are later assumed to be exchangeable:

the indexing of the units is, by definition, a random permutation of $1, \dots, N$, and thus any distribution on the science must be row-exchangeable

Our proposition is: can the intuition that unique individuals are an important for the motivation for causal models, be captured by considering models that feature “unique identifier” variables referencing these unique individuals?

1.3.2 Unique identifiers

A sequence of *unique identifiers* is a vector of finite or infinite length such that no two coordinates are equal. We are interested in models that assign positive probability to any particular coordinate having any particular value. This is straightforward in the finite case. In the infinite case, note that a vector of $|\mathbb{N}|$ unique values with an arbitrary entry k in the j th coordinate can be obtained by starting with $(i)_{i \in \mathbb{N}}$ and then transposing j with k . More generally, we consider infinite length sequences of unique identifiers to be elements of the set of finite permutations $\mathbb{N} \rightarrow \mathbb{N}$.

Definition 1.3.1 (Measurable space of unique identifiers). The measurable space of unique identifiers (I, \mathcal{I}) is the set I of finite permutations $\mathbb{N} \rightarrow \mathbb{N}$ with the discrete σ -algebra \mathcal{I} .

The set I is countable, as it is the countable union of finite subsets (i.e. the permutations that leave all but the first n numbers unchanged for all n).

Definition 1.3.2 (Unique identifier). Given a sample space (Ω, \mathcal{F}) , a *sequence of unique identifiers* $\mathcal{I} : \Omega \rightarrow I$ is a variable taking values in I .

The values of each coordinate of sequence of unique identifiers is just called an identifier (for obvious reasons, we don’t call it an identity).

Definition 1.3.3 (Identification). Given \mathbf{l} , define the i -th *identifier* $l_i := \text{ev}(i, \mathbf{l})$, where $\text{ev} : \mathbb{N} \times I \rightarrow \mathbb{N}$ is the evaluation map $(i, f) \mapsto f(i)$.

For *any* sample space (Ω, \mathcal{F}) we can define a trivial \mathcal{I} that maps every $\omega \in \Omega$ to $(1, 2, 3, \dots) =: (\mathbb{N})$. In this case, the identifiers are all known by the modeller at the outset. Using this sequence of identifiers renders exchange commutativity trivial.

Example 1.3.4. Given a sequential input-output model $(\mathbb{P}_C, (D, \mathbf{l}), Y)$ where \mathbf{l} is the identifier variable $\omega \mapsto (\mathbb{N})$, $\mathbb{P}_\alpha^{Y|D\mathbf{l}}$ commutes with exchange.

This is because for any permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$ except the identity, $\mathbb{P}_\alpha^{Y|D\mathbf{l}}$ and $\text{Swap}_\rho \mathbb{P}_\alpha^{Y|D\mathbf{l}}$ will have no common support; the first will be supported on $\mathbf{l} \bowtie (\mathbb{N})$ only, and the second only on $\mathbf{l} \bowtie \rho(\mathbb{N})$.

We are particularly interested in models where exchange commutativity is not trivial, so we focus on the case where each identifier l_i has some non-zero probability of taking any value in \mathbb{N} .

1.3.3 Identification with individual-level response functions

The key result of this section is Theorem 1.3.8. A key assumption for this theorem is the assumption of “individual-level response functions”. That is the assumption that, given a sequential input-output model $(\mathbb{P}_C, (D, \mathbf{l}), Y)$, there is some J such that $\mathbb{P}_\alpha^{Y_i|D_i \mathbf{l}_i^J}$ are mutually independent and identical responses for all α, i, J , unlike the directing random conditional H (Definition ??), is not necessarily a function of the inputs and outputs. We also assume that each individual identifier l_i has positive probability of taking on any particular value.

This assumption is somewhat difficult to understand. If we imagine that the identifiers l_i are, for example, patient names in some medical experiment, the pair of assumptions rule out certain possibilities. For example, these assumption is incompatible with the idea that, if Tina is first in line, then she will be deterministically cured by the treatment while if she is second in line she will be deterministically not-cured by it. If we model the problem such that the variable J specifies a stochastic response for each individual independent of the order of treatment or the treatment of any other individual, then the assumption of IO contractibility over J holds and J could be interpreted as a complete specification of potential outcomes. There may be other variables that motivate an assumption of individual level response functions, but we do not know of them at this.

Because every individual has a unique identifier and so, in general, the outcome of individual i 's experiment need not constrain the outcome of individual j even if they receive the same inputs, it seems that it may very often be possible to construct a model with individual level response functions. Finding a simple set of conditions sufficient to enable such a construction is an open question.

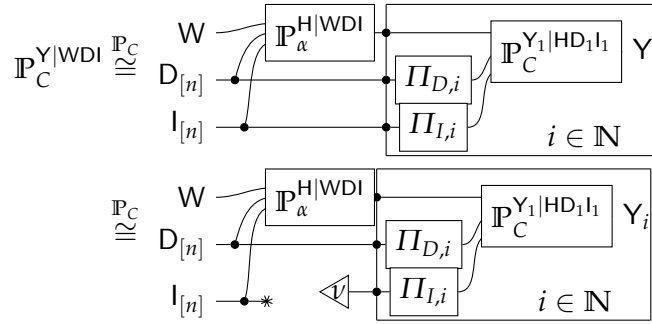
Before proving Theorem 1.3.8, we prove a number of lemmas and a preliminary theorem. Lemma 1.3.5 and Theorem 1.3.6 do *not* require that l be a sequence of unique identifiers, they hold just as well if it is a sequence of non-unique labels; that is, if $l_i \bowtie l_j$ had positive measure for some $i \neq j$. The reason why we are interested mainly in the case where l is a sequence of unique identifiers is that the assumption $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e l | id_C$ is substantially more limiting in the case that l is a non-unique sequence of labels. In particular, it implies that the conditional probability of (Y_1, Y_2) given $(l_1 = 1, l_2 = 1)$ is exactly the same as the conditional probability of (Y_1, Y_2) given $(l_i = 1, l_2 = 2)$; observations associated with equal labels are no more relevant than observations associated with different labels.

In the following, it is helpful to assume that each sub-experiment has a “unique identifier” l_i , with the sequence of all sub-experiment labels given by l . With this, if $\mathbb{P}_C^{Y|Dl}$ is assumed IO contractible, then it's possible to talk about the individual response functions $\mathbb{P}_C^{Y_i | l_i, HD_i}$. These plays a role very similar to the i th vector of potential outcomes Y_i^D . Because l_i is unique (i.e. never equal to l_j for $j \neq i$), only one observation of any individual is ever given, just like only one element of a vector of potential outcomes is ever observed.

Theorem 1.3.8 can also be extended to the case where D is a function of the choice α and a “random signal” R , as in Theorem 1.3.9.

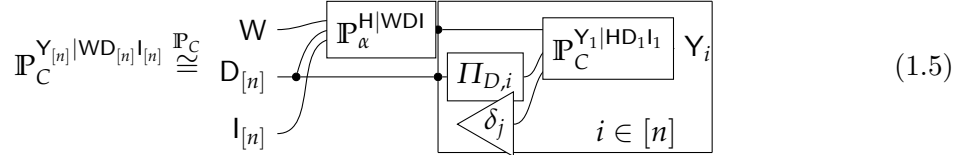
Lemma 1.3.5. *Given sequential input-output model $(\mathbb{P}_C, (D, l), Y)$ with $\mathbb{P}_\alpha^{Y|WDl}$ IO contractible over W , if $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e (l, id_C) | (W, D)$ and for any $j \in I$, $\sum_{\alpha \in C} \mathbb{P}_\alpha^{l_i}(j) > 0$, then there is some W' such that $\mathbb{P}_\alpha^{Y|W'D}$ is also IO contractible over W .*

Proof. Fix arbitrary $\nu \in \Delta(I^{\mathbb{N}})$ such that $\sum_{\alpha \in C} \mathbb{P}_{\alpha}^I \gg \nu$. By assumption of IO contractibility of $\mathbb{P}_C^{Y|WDI}$ and Theorem ??

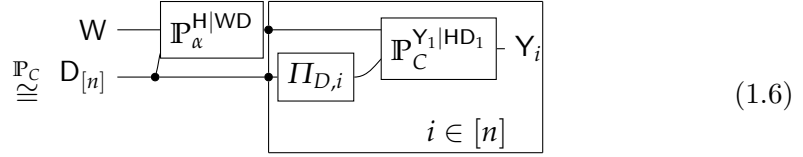


Where $\Pi_{D,i} : D^{\mathbb{N}} \rightarrow D$ projects the i th coordinate, and similarly for $\Pi_{Y,i}$.

In particular, for any $i \in \mathbb{N}$, $j \in I$, this holds for some ν such that $\nu(\Pi_{Y,i}^{-1}(j)) = 1$ and by extension for any finite $A \subset \mathbb{N}$ we can find ν such that $\nu(\Pi_{Y,i}^{-1}(j)) = 1$ for all $i \in A$, $j \in I$. Thus, for any $n \in \mathbb{N}$



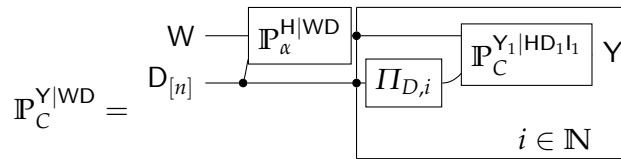
(1.5)



(1.6)

where Equation (1.5) follows from Theorem ?? and Equation (1.6) follows from the fact that Equation (1.5) holds for arbitrary $j \in I$.

Thus by Lemma ??



Applying Theorem ??, $\mathbb{P}_C^{Y|WD}$ is IO contractible over W . □

Theorem 1.3.6. Given a sequential input-output model $(\mathbb{P}_C, (D, I), Y)$ on (Ω, \mathcal{F}) with Y standard measurable and C countable, if there is some J such that for each α

$$\begin{aligned} \mathbb{P}_{\alpha}^{Y_i|J I_i D_i} &= \mathbb{P}_{\alpha}^{Y_i|J I_i D_i} \\ Y_i &\perp\!\!\!\perp_{\mathbb{P}_C}^{\epsilon} (I_{\{i\}^c}, D_{\{i\}^c}) | (J, I_i, D_i) \end{aligned} \quad \forall i, j \in \mathbb{N}$$

and

$$\begin{aligned}
 Y &\perp\!\!\!\perp_{\mathbb{P}_C}^e I | id_C \\
 YIJ &\perp\!\!\!\perp_{\mathbb{P}_C}^e D | id_C \\
 YIJ &\perp\!\!\!\perp_{\mathbb{P}_C}^e id_C | D \\
 \forall i, j \in \mathbb{N} : \sum_{\alpha \in C} \mathbb{P}_\alpha^{I_i}(j) &> 0
 \end{aligned}$$

then $\mathbb{P}_C^{Y|JD}$ is IO contractible over J.

Proof. For any $\alpha \in C$

$$\begin{aligned}
 \mathbb{P}_\alpha^{YJ|I} &= \begin{array}{c} I \text{ --- } \bullet \\ \quad \downarrow \\ \begin{array}{|c|} \hline \mathbb{P}_\alpha^{D|I} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{YJ|ID} \\ \hline \end{array} \text{ --- } (Y, J) \end{array} \\
 &= \begin{array}{c} I \text{ --- } \triangleleft \mathbb{P}_\alpha^D \\ \quad \downarrow \\ \begin{array}{|c|} \hline \mathbb{P}_C^{YJ|ID} \\ \hline \end{array} \text{ --- } (Y, J) \end{array}
 \end{aligned}$$

Define \mathbf{Q} by $\alpha \mapsto \mathbb{P}_\alpha$ and $\mathbf{Q}^{\cdot|id_C}$ by $\alpha \mapsto \mathbb{P}_\alpha^*$ and \mathbf{Q}^{id_C} is an arbitrary distribution in $\Delta(C)$ with full support. Note that the support of \mathbf{Q}^{IDYJ} is the union of the support of \mathbb{P}_α^{IDYJ} for all α . Then

$$\mathbf{Q}^{YJ|IC} \stackrel{\mathbf{Q}}{\cong} id_C \text{ --- } \begin{array}{|c|} \hline \mathbf{Q}^{D|id_C} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{YJ|ID} \\ \hline \end{array} \text{ --- } (Y, J)$$

By assumption $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e D | id_C$, it is also the case that

$$\begin{aligned}
 \mathbf{Q}^{Y|ID} &\stackrel{\mathbf{Q}}{\cong} \begin{array}{c} I \text{ --- } \bullet \\ \quad \downarrow \\ \begin{array}{|c|} \hline \mathbf{Q}^{id_C|ID} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{Q}^{Y|IC} \\ \hline \end{array} \text{ --- } Y \\ D \text{ --- } \end{array} \\
 &\stackrel{\cong}{\cong} \begin{array}{c} I \text{ --- } \bullet \\ \quad \downarrow \\ \begin{array}{|c|} \hline \mathbf{Q}^{id_C|D} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{Q}^{YJ|IC} \\ \hline \end{array} \text{ --- } (Y, J) \\ D \text{ --- } \end{array} \\
 &\stackrel{\cong}{\cong} \begin{array}{c} I \text{ --- } \bullet \\ \quad \downarrow \\ \begin{array}{|c|} \hline \mathbf{Q}^{id_C|D} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{Q}^{D|id_C} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{YJ|ID} \\ \hline \end{array} \text{ --- } (Y, J) \\ D \text{ --- } \end{array}
 \end{aligned}$$

But

$$\begin{aligned}
 \mathbf{Q}^{Y|ID} &= \sum_{\alpha \in C} \mathbb{P}_\alpha^{Y|ID} \mathbf{Q}^{id_C}(\alpha) \\
 &= \mathbb{P}_C^{Y|ID} \\
 \Rightarrow \begin{array}{c} I \text{ --- } \bullet \\ \quad \downarrow \\ \begin{array}{|c|} \hline \mathbf{Q}^{id_C|D} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{Q}^{D|id_C} \\ \hline \end{array} \begin{array}{|c|} \hline \mathbb{P}_C^{YJ|ID} \\ \hline \end{array} \text{ --- } (Y, J) \\ D \text{ --- } \end{array} &= \mathbb{P}_C^{Y|ID}
 \end{aligned}$$

Furthermore, by assumption $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e I | id_C$, so there is some $\mathbb{K} : C \rightarrow Y \times W$ such that

$$\begin{aligned}
 Q^{YJ|C} &\stackrel{Q}{\cong} I \overset{*}{\underset{D}{\curvearrowright}} \boxed{\mathbb{K}} \text{---} (Y, J) \\
 \implies \mathbb{P}_C^{YJ|D} &= I \text{---} \boxed{\mathbb{F}_\rho} \text{---} \boxed{\mathbb{P}_C^{YJ|D}} (Y, J) \\
 &\quad \underset{D}{\curvearrowright} \boxed{Q^{id_C|D}} \boxed{Q^{D|id_C}} \text{---} \\
 &= I \text{---} \boxed{\mathbb{P}_C^{YJ|id_C}} \text{---} (Y, J) \\
 &\quad \underset{D}{\curvearrowright}
 \end{aligned}$$

Then by Lemma 1.3.5, $\mathbb{P}_C^{YJ|D}$ is IO contractible over J . \square

Lemma 1.3.7 is used to apply Theorem 1.3.6 to models where I is a sequence of unique identifiers. Only in this case, exchangeability of the unique identifiers implies the identifiers are independent of the outcomes Y .

Lemma 1.3.7. *Given any probability set \mathbb{P}_C where $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e id_C | (D, I)$ and $I : \Omega \rightarrow I$ is an infinite sequence of unique identifiers, if for each finite permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$*

$$\mathbb{P}_\alpha^{Y|I} = (Swap_{\rho(I)} \otimes Id_X) \mathbb{P}_\alpha^{Y|I}$$

then $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e I | id_C$.

Proof. By definition of the set I of finite permutations, for every $\rho \in I$, $B \in \mathcal{Y}^{\mathbb{N}}$, $d \in D^{\mathbb{N}}$ there is a finite permutation $\rho^{-1} \in I$ such that $\rho \circ \rho^{-1} = id_{\mathbb{N}}$. Then

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y|I}(B|\rho) &= (\mathbb{F}_{\rho^{-1}} \otimes Id_X) \mathbb{P}_\alpha^{Y|I}(B|\rho) \\
 &= \mathbb{P}_\alpha^{Y|I}(B|id_{\mathbb{N}})
 \end{aligned}$$

Therefore

$$\mathbb{P}_\alpha^{Y|I} \stackrel{\mathbb{P}_C}{\cong} \text{erase}_I \otimes \mathbb{P}_\alpha^Y$$

\square

Theorem 1.3.8 presents a set of sufficient conditions for $\mathbb{P}_C^{Y_i|HD_i}$ to be conditionally independent and identical response functions with respect to the standard directing random conditional H :

1. There exist variables I representing “unique identifiers” which satisfy the assumption that $\mathbb{P}_C^{Y_i|JD_i I_i}$ are a sequence of independent and identical response functions for some J
2. The identifiers I can be swapped without altering the model of the consequences Y
3. The inputs D and the choice id_C are substitutable with respect to Y and I : $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e id_C | D$ and $YI \perp\!\!\!\perp_{\mathbb{P}_C}^e D | id_C$

Theorem 1.3.8. *Given a sequential input-output model $(\mathbb{P}_C, (D, I), Y)$, on (Ω, \mathcal{F}) with Y standard measurable and C and D countable, I an infinite sequence of unique identifiers, if*

there is some J such that

$$\begin{aligned} \mathbb{P}_\alpha^{Y|I} &= (\text{Swap}_{\rho(I)} \otimes \text{Id}_X) \mathbb{P}_\alpha^{Y|I} && \forall \text{ finite permutations } \rho \\ Y|J &\perp\!\!\!\perp_{\mathbb{P}_C}^e D | \text{id}_C \\ Y|J &\perp\!\!\!\perp_{\mathbb{P}_C}^e \text{id}_C | D \\ \forall i, j \in \mathbb{N} : &\sum_{\alpha \in C} \mathbb{P}_\alpha^i(j) > 0 \end{aligned}$$

and for each α $\mathbb{P}_\alpha^{Y|JD}$ is IO contractible over J , then we have conditionally independent and identical responses $\mathbb{P}_\alpha^{Y_i|D_iH}$ for all i , where H is the directing random conditional with respect to (D, Y) .

Proof. Apply lemma 1.3.7 to get $Y \perp\!\!\!\perp_{\mathbb{P}_C}^e I | \text{id}_C$, then apply Theorem 1.3.6 for $\mathbb{P}_C^{Y|JD}$ IO contractible. The result follows from Theorem ?? \square

Theorem 1.3.8 can be extended to the case where decisions D are a one-to-one deterministic function of the choice, or a random mixtures of one-to-one deterministic functions of the choice. This extension is applicable a randomised controlled trial, where the treatments are deterministically controlled and randomly assigned.

Theorem 1.3.9. Consider a sequential input-output model $(\mathbb{P}_{C'}, D, Y)$ where $\mathbb{P}_{C'}^{Y|WD}$ is IO contractible over W , and construct a second model (\mathbb{P}_C, D, Y) where \mathbb{P}_C is the union of $\mathbb{P}_{C'}$ and its convex hull. Then $\mathbb{P}_C^{Y|WD}$ is also IO contractible.

Proof. For all $\alpha \in C$, there is some probability measure $\mu : C' \rightarrow [0, 1]$ such that

$$\begin{aligned} \mathbb{P}_\alpha^{Y|WD} &= \sum_{\beta \in C'} \mu(\beta) \mathbb{P}_\beta^{Y|WD} \\ &= \mathbb{P}_{C'}^{Y|WD} \end{aligned}$$

thus

$$\mathbb{P}_C^{Y|WD} = \mathbb{P}_{C'}^{Y|WD}$$

and in particular, $\mathbb{P}_C^{Y|WD}$ is IO contractible. \square

Theorem 1.3.9 can be used to argue that, given a sequence of experiments IO contractible under deterministic choices, adding random mixtures of these choices also yields a IO contractible sequence. Kasy (2016) argues that as long as the experimenter controls the treatment assignment, causal effects are identified (i.e. the randomisation step is not strictly necessary). Example 1.3.10 shows that this argument might be supported, but Example ?? shows that there are subtle ways that might lead to this argument failing.

We assume an infinite sequence, which is clearly unreasonable. Extending the representation theorems to the case of finite sequences, using for example the result of Diaconis and Freedman (1980) with establishes that finite exchangeable distributions are approximately mixtures of independent and identically distributed sequences, would allow some implausible assumptions in the following example to be removed.

Theorem 1.3.8 is used in the following example to argue that, under certain conditions, a controlled experiment supports a IO contractible model.

Example 1.3.10. A sequential experiment is modeled by a probability set \mathbb{P}_C with binary treatments $D := (D_i)_{i \in \mathbb{N}}$ and binary outcomes $Y := (Y_i)_{i \in \mathbb{N}}$. The set of choices C is the set of all probability distributions $\Delta(D^{\mathbb{N}})$ for some $N \subset \mathbb{N}$ (this is to ensure C is countable).

Each treatment D_i is given to a patient, and each patient provides a unique identifier I_i which for simplicity we assume is a number in \mathbb{N} (instead of, say, a driver's license number and state of issue), and that (implausibly) there is a positive probability for I_i to take any value in \mathbb{N} for any choice α .

The treatments are decided as follows: the analyst consults the model \mathbb{P}_C , and, according to \mathbb{P}_C and some previously agreed upon decision rule, comes up with a (possibly stochastic) sequence of treatment distributions $\alpha := (\mu_i)_{i \in \mathbb{N}}$ with each μ_i in $\Delta(\{0, 1\})$. If μ_i is deterministic – that is, it puts probability 1 on some treatment d_i , the experiment administrator will assign patient i the treatment d_i . Otherwise, if μ_i is nondeterministic, the administrator will consult a random number generator that yields treatment assignments according to μ_i , and treatment will then be assigned deterministically according to the result. Letting $C' \subset C$ be the deterministic elements of C , this scheme is assumed by the analyst to support the assumptions $Y|J \perp\!\!\!\perp_{\mathbb{P}_{C'}} D|id_C$ and $Y|J \perp\!\!\!\perp_{\mathbb{P}_{C'}} id_C|D$ for any J , and the randomisation procedure is deemed sufficient to ensure that for any mixed $\alpha \in C$ where $\alpha = \sum_{\beta \in C'} \mu(\beta)\beta$, $\mathbb{P}_\alpha = \sum_{\beta \in C'} \mathbb{P}_\beta$.

Furthermore, assume $\mathbb{P}_C^{Y|D|I}$ is IO contractible. Then by Theorem ??, there is some J such that, conditional on J , $\mathbb{P}_C^{Y_i|J D_i I_i}$ are conditionally independent and identical response functions. The analyst constructing the model has no particular knowledge about any identifier, and so for any choice the associated model is assumed invariant to permutations of identifiers – that is $Y \perp\!\!\!\perp_{\mathbb{P}_C} I|id_C$ (see Lemma 1.3.6). The assumption that this holds given any choice can be tricky – not only must the identifiers appear symmetric to the analyst constructing the model, but nothing breaking this symmetry may be learned from the choice α (see the Example 1.3.4). One reason supporting this assumption is that the decision maker selects α according to a rule known in advance, so they do not “learn” anything upon picking a particular α .

Then, for the deterministic subset $C' \subset C$, application of Theorem 1.3.8 yields $\mathbb{P}_{C'}^{Y|HD}$ is IO contractible over H , and by application of Corollary 1.3.9, so is $\mathbb{P}_C^{Y|D}$.

Permutability of identifiers can fail when the rule for selecting α is not known in advance. The following example is extreme in order to illustrate the issue clearly. The distinction between the analyst and the administrator is also intended to make the example easier to parse. The key point is that, when the rule for selecting α is not known in advance, symmetries that are apparent at the time of model construction do not necessarily hold for every choice α , and this remains true if e.g. the selection of choices leads to less extreme confounding or the analyst and the administrator are actually the same person.

The following example involves the choice α depending on some covariate U . It is not straightforward to express the idea that “ α depends on U ” in a probability set model \mathbb{P}_C , and they are intended to apply to situations where the choice doesn't depend on anything not already expressed in the model (as in Example 1.3.10). However, the fact that probability sets don't work well in situations where the choice depends on something not expressed in the model doesn't mean that you can't use a probability set to model such a situation, it just means that you shouldn't do it. This is what the following example shows.

The condition $Y|J \perp\!\!\!\perp_{\mathbb{P}_C} id_C|D$ without also having $Y|J \perp\!\!\!\perp_{\mathbb{P}_C} D|id_C$ does *not* imply the conclusion of Theorem 1.3.8. Informally, if D gives some “extra information” over and above id_C , then any symmetry that holds before we observe D might not hold after D has been observed. We have argued in Section 1.3.4 somewhat informally that the choice id_C should

be completely under the decision maker's control – for Theorem 1.3.8, this perfect control has to extend to the sequence of inputs D . Constructing the following example requires the hypotheses that any given identifier $i \in \mathbb{N}$ could be associated with one of two input-output maps $D \rightarrow Y$. Thus the space of hypotheses is a sequence of binary values $H = \{0, 1\}^{\mathbb{N}}$. Equipped with the product topology, H is a countable product of separable, completely metrizable spaces and is therefore also separable and completely metrizable (Willard, 1970, Thm. 16.4, Thm. 24.11). Thus $(H, \mathcal{B}(H))$ is a standard measurable space and, because it is uncountable, it is isomorphic to $([0, 1], \mathcal{B}([0, 1]))$.

Example 1.3.11. Take $Y = C = D = \{0, 1\}$ and take (H, \mathcal{H}) to be $\{0, 1\}^{\mathbb{N}}$ equipped with the product topology. For any $i \neq 1$, $Y_i | D_i \perp\!\!\!\perp_{\mathbb{P}_C} \text{id}_C$, while $\mathbb{P}_\alpha^{D_1} = \delta_\alpha$ and $I_i \perp\!\!\!\perp_{\mathbb{P}_C} \text{id}_C$.

$YI \perp\!\!\!\perp_{\mathbb{P}_C} \text{id}_C | D$ follows from the fact that id_C can be (almost surely) written as a function of D .

For all $i \in \mathbb{N}$, $y, d \in \{0, 1\}$, $h \in H$ set

$$\mathbb{P}_C^{Y_i | H I_i D_i}(y | h, j, d) = \delta_1(p(j, h))\delta_d(y) + \delta_0(p(j, h))\delta_{1-d}(y)$$

where $p(j, h)$ projects the j -th component of h . That is, if h maps j to 1, Y goes with D while if h maps j to 0, Y goes opposite D . Suppose also

$$Y_i \perp\!\!\!\perp_{\mathbb{P}_C} (X_{<i}, Y_{<i}, I_{<i}, \text{id}_C) | (X_i, Y_i, H)$$

Then $\mathbb{P}_C^{Y | D I}$ is IO contractible. Set \mathbb{P}_C^H to be the uniform measure on (H, \mathcal{H}) and for $i > 1$

$$\mathbb{P}_C^{D_i | I_i H}(d | j, h) = \delta_{p(j, h)}(d)$$

that is, if h maps j to 1, D is 1 while if h maps j to 0, D is 0. This also implies

$$\mathbb{P}_C^{I_i | D_i H}(p(\cdot, h)^{-1}(d) | d, h) = 1 \quad (1.7)$$

Then, for $i > 1$

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i | H D_i}(y | h, d) &= \sum_{j \in \mathbb{N}} \delta_1(p(j, h))\delta_d(y)\mathbb{P}_C^{I_i | D_i H}(j | d, h) + \delta_0(p(j, h))\delta_{1-d}(y)\mathbb{P}_C^{I_i | D_i H}(j | d, h) \\ &= \sum_{j \in \mathbb{N}} \delta_1(d)\delta_d(y)\mathbb{P}_C^{I_i | D_i H}(j | d, h) + \delta_0(d)\delta_{1-d}(y)\mathbb{P}_C^{I_i | D_i H}(j | d, h) \quad \text{by Eq (1.7)} \\ &= \delta_1(y) \end{aligned}$$

$$\implies \mathbb{P}_\alpha^{Y_i | D_i}(y | d) = \delta_1(y)$$

For $q \in I$, set

$$\mathbb{P}_C^{I | H}(q | h) = \begin{cases} 0.5 & q = (1, 2, 3, 4, \dots) \text{ or } (1, 3, 2, 4, \dots) \\ 0 & \text{otherwise} \end{cases}$$

and set

$$\mathbb{P}_C^{H | D}(h) = \begin{cases} 0.5 & h = (0, 1, 0, 1, 1, \dots) \text{ or } h = (0, 0, 1, 1, 1, \dots) \\ 0 & \text{otherwise} \end{cases}$$

Let \bar{H} be the support of $\mathbb{P}_C^{H|D}(h)$.

Then for $i = 1$

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y_1|D_1}(y|h, d) &= \sum_{h \in H} \sum_{j \in \mathbb{N}} \mathbb{P}_\alpha^{I_1|D_1H}(j|d, h) \mathbb{P}_C^{H|D_1}(h|d) (\delta_1(p(j, h))\delta_d(y) + \delta_0(p(j, h))\delta_{1-d}(y)) \\
 &= \sum_{h \in \bar{H}} 0.5(\delta_1(p(1, h))\delta_d(y) + \delta_0(p(1, h))\delta_{1-d}(y)) \\
 &= \delta_{1-d}(y) \\
 &\neq \mathbb{P}_\alpha^{Y_i|D_i}(y|h, d) \quad i \neq 1
 \end{aligned}$$

Thus $\mathbb{P}_C^{Y|D}$ is not IO contractible by Theorem ??.

However, given any finite permutation $\rho : \mathbb{N} \rightarrow \mathbb{N}$

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y|I}(y|q) &= \sum_{h \in \bar{H}} \sum_{d \in \{0,1\}^{\mathbb{N}}} \prod_{i \in \mathbb{N}} \mathbb{P}_C^{Y_i|I_iD_iH}(y_i|q_i, d_i, h) \mathbb{P}_\alpha^{D_i|I_iH}(d_i|q_i, h) \mathbb{P}_C^H(h) \\
 &= \delta_{1-\alpha}(y_1) \delta_{(1)_{i \in \mathbb{N}}}(y_{>1}) \\
 &= \mathbb{P}_\alpha^{Y|I}(y|\rho^{-1}(q)) \\
 &= \mathbb{F}_\rho \mathbb{P}_\alpha^{Y|I}(y|q)
 \end{aligned}$$

1.3.4 Other examples

Example 5: Backdoor adjustment The “backdoor adjustment” formula is a fundamental tool for many kinds of causal inference. This is a short example to show the conditions under which it’s applicable, stated in terms of IO contractibility. Suppose a sequential input-output model $(\mathbb{P}_C, (D, X), Y)$ where $(\mathbb{P}_C^{Y|WDX})$ is IO contractible, and:

- $i > n \implies X_i \perp\!\!\!\perp_{\mathbb{P}_C} D_i | (H, \text{id}_C)$
- $\mathbb{P}_\alpha^{X_i|H} \cong \mathbb{P}_\alpha^{X_1|H}$ for all α

Then the model exhibits a kind of “backdoor adjustment”. Specifically, for $i > n$

$$\begin{aligned}
 \mathbb{P}_\alpha^{Y_i|D_iH}(A|d, h) &= \int_X \mathbb{P}_\alpha^{Y_i|X_iD_iH}(A|d, x, h) \mathbb{P}_\alpha^{X_i|D_iH}(dx|d, h) \\
 &= \int_X \mathbb{P}_\alpha^{Y_1|X_1D_1H}(A|d, x, h) \mathbb{P}_\alpha^{X_i|H}(dx|h) \\
 &= \int_X \mathbb{P}_\alpha^{Y_1|X_1D_1H}(A|d, x, h) \mathbb{P}_\alpha^{X_1|H}(dx|h) \tag{1.8}
 \end{aligned}$$

Equation (1.8) is identical to the backdoor adjustment formula (Pearl, 2009, Chap. 1) for an intervention on D_1 targeting Y_1 where X_1 is a common cause of both.

Example 6: the provenance of the choice variable

Use individual-level ccont

The point of this example is to clarify the idea of a “choice” variable. If we say that some value is the outcome of a choice, a straightforward interpretation of this term suggests that

this value was chosen by someone, somewhere. However, for the purposes of decision making models, there are important differences between:

- Values chosen by someone, somewhere
- Values chosen by the decision maker, using the decision making model

We call this example the “I choose vs you choose” problem. Suppose we have a decision maker (“DM”) and an administrator (“admin”) cooperating to collect data to support DM to make a choice.

First, consider the “I choose” condition. Here, DM’s choice $\alpha \in \{0,1\}^2$ deterministically sets the value of binary inputs D_1, D_2 , and the decision maker is interested in evaluating the corresponding binary outputs Y_1, Y_2 . The decision maker assesses that their knowledge of the real-world mechanisms that gives rise to each output Y_i in the context on an input D_i render these mechanisms indistinguishable. From their point of view, the input-output relations for each step are indistinguishable. In particular, they assess that the marginal probabilities of the outputs are the same given a corresponding input, and that the evidence that the first experiment brings to bear on the second is equivalent to the evidence that the second brings to bear on the first. Thus, they assess that exchange commutativity is appropriate; for all α :

$$\mathbb{P}_\alpha^{Y_1 Y_2 | D_1 D_2} = \mathbb{P}_\alpha^{Y_2 Y_1 | D_2 D_1}$$

but this example suggests another reason one might want to avoid deterministic treatment assignments. If the choices α are a deterministic sequence of assignments for each index i , this means that there is an enormous set of possible choices, and many degrees of freedom if the choices “aren’t actually chosen” in the sense of the example above. In contrast, if the set of choices is a single parameter in $[0,1]$ which is then used to assign all treatments according to a random procedure depending only on this parameter, there are many fewer degrees of freedom to exploit if the choice “isn’t actually chosen”.

A particular concern arises when the choice variable id_C is not associated with the output of a decision procedure involving the model \mathbb{P}_C . In this situation, the value of id_C can affect the model in potentially unexpected ways. “Potentially unexpected” is a vague notion, and we can’t say whether id_C being completely under the decision maker’s control avoids “unexpected” dependence on id_C , but it seems to be less problematic.

We set this up in terms of an “analyst” and an “administrator” who have responsibility for different parts of the procedure. They don’t strictly need to be different people, but it helps make the issue clearer. The analyst’s job is to construct a model \mathbb{P}_C , evaluate different options $\alpha \in C$ and offer advice regarding the choice. The administrator’s job is to choose some $\alpha \in C$ satisfying the analyst’s requirements and to carry out any procedure arising from this.

This separation of concerns gives the administrator a degree of freedom in their choice, and they can potentially use this to choose α with access to information that the analyst lacks.

In particular, suppose an experiment is modeled by a sequential input-output model $(\mathbb{P}_C, (D, U), Y)$ and the set of choices $C = [0,1]^{\mathbb{N}}$ is a length \mathbb{N} sequence of probability distributions in $\Delta(\{0,1\})$. The analyst, based on their knowledge of the experiment, constructs \mathbb{P}_C such that $\mathbb{P}_C^{Y_i | U_i D_i}(1 | \cdot, \cdot)$ is given by:

	$D_i = 0$	$D_i = 1$
$U_i = 0$	0	0
$U_i = 1$	1	1

and the triples (D_i, U_i, Y_i) are mutually independent given id_C . This makes $\mathbb{P}_C^{Y|UD}$ IO contractible over $*$. Suppose also

$$\mathbb{P}_\alpha^{D_i}(1) = \alpha_i$$

where $\alpha = (\alpha_i)_{i \in \mathbb{N}}$. From the analyst's point of view, both before and after making their recommendations the U_i are also IID. This will be expressed with a probability distribution \mathbf{Q} representing the analyst's prior knowledge:

$$\mathbf{Q}^{U_i}(1) = 0.5$$

one might be tempted to reason that, if \mathbf{Q} is the analyst's state of knowledge after making any recommendation, then we should take $\mathbb{P}_C^U = \mathbf{Q}^U$. Call the resulting model \mathbb{P}'_C . Together with the other assumptions above, this would imply

$$\mathbb{P}_C^{Y_i|D_i}(1|d) = 0.5 \quad \forall d \in \{0, 1\}$$

Thus $\mathbb{P}_C^{Y|D}$ is also IO contractible.

However, the analyst's recommendation *does not* fix the value of id_C . Suppose analyst actually recommends any α such that $\lim_{n \rightarrow \infty} \sum_i^n \frac{\alpha_i}{n} = 0.5$ (acknowledging that, in this contrived example, there's no obvious reason to do so). Suppose that the administrator operates by the following rule: *first* they observe the value of U_i , then they choose α_i equal to whatever they saw with an ϵ sized step towards 0.5. That is, if they see $U_i \bowtie 1$, they choose $\alpha_i = 1 - \epsilon$, where $\epsilon < 0.5$.

Then the analyst should instead adopt the model

$$\mathbb{P}_\alpha^{U_i}(1) = \mathbb{1}_{\alpha_i > 0.5}$$

Take α such that $\alpha_i = 1 - \epsilon$ and $\alpha_j = \epsilon$. Then

$$\begin{aligned} \mathbb{P}_\alpha^{Y_i|D_i}(1|1) &= 1 \\ &\neq \mathbb{P}_\alpha^{Y_j|D_j}(1|1) \\ &= 0 \end{aligned}$$

everything has been assumed IID, so $\mathbb{P}_C^{Y|HD}$ is not IO contractible.

The original justification for having a set of choices C is that C is the set of things that, after deliberation aided by the model \mathbb{P}_C , the decision maker might select. The present example does not conform to this understanding of the meaning of the set C , and it suggests that one should be cautious when modelling “decision problems” with “choices” that are not actually the things that are being chosen.

This point is related to the question of why experimenters randomise. [Kasy \(2016\)](#) argues that “randomised controlled trials are not needed for causal identifiability, only controlled trials”, and suggests that experiments should sometimes be designed with deterministic assignments

of patients to treatment and control groups, optimised according to the experiment designer’s criteria. Following this, [Banerjee et al. \(2020\)](#) suggested that deterministic rules might falter when one can’t pick a function to balance covariates in a way that satisfies everyone in a panel of reviewers.

Without solving the problem, we observe that the terms “control” and “choice” here subsume both different kinds of choice indicated above, each of which has different implications for the construction of decision making models. We offer a speculative alternative explanation for randomisation: perhaps that the same model may be appropriate for both notions of “choice” under randomised choices, but not under nonrandomised choices.



[Sävje \(2021\)](#) argues that random assignment (under his definition) does not imply unconfoundedness

1.4 Conclusion

We review the decision making models implied by causal Bayesian networks and potential outcomes models. We find that these kinds of models have complementary “missing pieces” needed to induce the relevant decision making model – while causal Bayesian networks already have interventions that provide a kind of “choice set”, they need to be unrolled to a sequential model. On the other hand potential outcomes models can already be specified in an unrolled form, but need some notion of “choice set” to induce a decision making model. Common formulations of both models feature conditionally independent and identical response functions. We explore individual-level response functions as a means of establishing the widely accepted result that randomised trials. We note that the assumption of individual-level response functions seems to be a missing step in the often cited idea that exchangeability of individuals implies exchangeability of potential outcomes, and we show that with individual-level response functions, exchangeable individuals and completely controllable inputs, causal relationships are identified. The need for completely controllable inputs is also widely accepted, but to our knowledge it only appears as a formal assumption in Theorem ??.

Bibliography

- Banerjee, A. V., Chassang, S., Montero, S. and Snowberg, E. (2020). ‘A Theory of Experimenters: Robustness, Randomization, and Balance’. *American Economic Review*, 110(4), pp. 1206–1230. DOI: [10.1257/aer.20171634](https://doi.org/10.1257/aer.20171634) (cited on p. 29).
- Chickering, D. M. (2002). ‘Learning Equivalence Classes of Bayesian-Network Structures’. *Journal of Machine Learning Research*, 2(Feb), pp. 445–498. [↗](#) (visited on 14 October 2018) (cited on p. 13).
- Chickering, D. M. (2003). ‘Optimal Structure Identification with Greedy Search’. *J. Mach. Learn. Res.*, 3, pp. 507–554. DOI: [10.1162/153244303321897717](https://doi.org/10.1162/153244303321897717) (cited on p. 13).
- Dawid, A. P. (2020). ‘Decision-theoretic foundations for statistical causality’. *arXiv:2004.12493 [math, stat]*. arXiv: 2004.12493. [↗](#) (visited on 23 September 2020) (cited on p. 17).
- Diaconis, P. and Freedman, D. (1980). ‘Finite Exchangeable Sequences’. *The Annals of Probability*, 8(4). Publisher: Institute of Mathematical Statistics, pp. 745–764. DOI: [10.1214/aop/1176994663](https://doi.org/10.1214/aop/1176994663) (cited on p. 23).
- Greenland, S. and Robins, J. M. (1986). ‘Identifiability, Exchangeability, and Epidemiological Confounding’. *International Journal of Epidemiology*, 15(3), pp. 413–419. DOI: [10.1093/ije/15.3.413](https://doi.org/10.1093/ije/15.3.413) (cited on p. 17).
- Holy Bible : Contemporary English Version* (1995). New York : American Bible Society, [1995] ©1995. [↗](#) (cited on p. 9).
- Kasy, M. (2016). ‘Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead’. *Political Analysis*, 24(3). Publisher: Cambridge University Press, pp. 324–338. DOI: [10.1093/pan/mpw012](https://doi.org/10.1093/pan/mpw012) (cited on pp. 17, 23, 28).
- Lattimore, F. and Rohde, D. (2019a). ‘Causal inference with Bayes rule’. *arXiv:1910.01510 [cs, stat]*. arXiv: 1910.01510. [↗](#) (visited on 30 September 2021) (cited on p. 6).
- Lattimore, F. and Rohde, D. (2019b). ‘Replacing the do-calculus with Bayes rule’. *arXiv:1906.07125 [cs, stat]*. arXiv: 1906.07125. [↗](#) (visited on 23 September 2020) (cited on p. 6).
- Meek, C. (1995). ‘Strong Completeness and Faithfulness in Bayesian Networks’. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI’95. event-place: Montréal, Qué, Canada. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 411–418. ISBN: 978-1-55860-385-1. [↗](#) (visited on 19 March 2019) (cited on p. 13).
- Ng, I., Zhu, S., Chen, Z. and Fang, Z. (2019). *A Graph Autoencoder Approach to Causal Structure Learning*. Number: arXiv:1911.07420 arXiv:1911.07420 [cs, stat]. DOI: [10.48550/arXiv.1911.07420](https://doi.org/10.48550/arXiv.1911.07420) (cited on p. 13).
- Okamoto, M. (1973). ‘Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample’. *The Annals of Statistics*, 1(4). Publisher: Institute of Mathematical Statistics, pp. 763–765. [↗](#) (visited on 5 July 2022) (cited on p. 12).
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press (cited on pp. 1, 2, 17, 26).

- Rubin, D. B. (2005). ‘Causal Inference Using Potential Outcomes’. *Journal of the American Statistical Association*, 100(469), pp. 322–331. DOI: [10.1198/016214504000001880](https://doi.org/10.1198/016214504000001880) (cited on pp. [14](#), [17](#)).
- Sävje, F. (2021). *Randomization does not imply unconfoundedness*. arXiv:2107.14197 [stat]. DOI: [10.48550/arXiv.2107.14197](https://doi.org/10.48550/arXiv.2107.14197) (cited on p. [29](#)).
- Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. and Richardson, T. (2000). *Causation, prediction, and search*. MIT press (cited on p. [13](#)).
- Willard, S. (1970). *General topology*. Reading, Mass., Addison-Wesley Pub. Co. ISBN: 978-0-201-08707-9.  (visited on 2 May 2022) (cited on p. [25](#)).
- Zheng, X., Aragam, B., Ravikumar, P. K. and Xing, E. P. (2018). ‘DAGs with NO TEARS: Continuous Optimization for Structure Learning’. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.  (visited on 15 June 2022) (cited on p. [13](#)).