Causal Inference - A Probabilistic Modelling Perspective

Abstract

The topic of causality has caused heated debates within the machine learning and statistics communities. Following the release of "the Book of Why" [1], many have come to question how the methods advocated in the book relate to pre-existing approaches in those fields. For instance, many dispute whether one *needs* a bespoke set of tools, such as the *do*-notation or the *do*-calculus, to tackle causal problems. In this note, we attempt to get to the bottom of these questions, and present a clear view on how causal inference problems can be approached using machine learning techniques. To this end, we introduce the general concepts of causal inference from a probabilistic modelling perspective, link methods developed under the umbrella of causal inference to those familiar to the machine learning audience, and clarify various claims about the insufficiencies of machine learning methods for tackling causal inference problems.

1 Introduction

In recent years, we have been struck by the fact that eminent researchers in the area of causality and statistical inference appear to hold contradictory views on the technical foundations of their field. The broader disagreement is well exemplified by the argument between Judea Pearl and Andrew Gelman; whereas Judea Pearl holds the position that *"there is no way to answer causal questions without snapping out of statistical vocabulary"* [2], Andrew Gelman states that *"I find it baffling that Pearl and his colleagues keep taking statistical problems and, to my mind, complicating them by wrapping them in a causal structure"* [3].

> *Q*: People are excited about the possibilities for AI. You're not?
> *"As much as I look into what's being done with deep learning, I see they're all stuck there on the level of associations. Curve fitting."*
>
> — Judea Pearl [2]

The debate largely followed after the release of "The Book of Why" [1] and an interview with its author Judea Pearl [2], an excerpt of which is shown above. In the book, the author introduces causal thinking into statistics and machine learning. Simultaneously, he also presents a rather dim view on these fields. In a broad sweep Judea Pearl paints modern statistics, for example, as "a model-blind data-reduction enterprise".

Both parties are highly esteemed academics on the topic of causality; Andrew Gelman has authored many statistics books which cover the topic of causal inference. Judea Pearl is a Turing award winner who made many fundamental developments to the graphical approach to causality and has written multiple books on the topic. As such, it is particularly striking that they struggle to reach a consensus on the technical foundations of their fields.

These statements consequently propelled a larger discourse on the topic of causality and its relation to machine learning and statistics. The debate ranges across topics such as: do you need bespoke methods advocated by Pearl to do causal inference; how does previous work related to causality within machine learning and statistics fit in with Pearl's framework; and — a question of a particularly semantic nature — whether deep learning can perform causal inference.

There is a lack of consensus about the micro-level technical details that echoes the landscape of the broader disagreement. For instance, some contend whether graphical models alone can be used to answer certain types of causal questions - counterfactuals. In their book, Jonas Peters, Dominik Janzing and Bernhard Schölkopf write that *"causal graphical models are not rich enough to predict counterfactuals"*. Instead, they claim that a class of models called structural causal models (SCMs) is necessary: *"Formally, SCMs contain strictly more information than their corresponding graph and law (e.g., counterfactual statements)."* [4] On the other hand, Kusner, Loftus and Russell et al., for instance, use causal graphical models to reason about questions of counterfactual nature in their paper on counterfactual fairness [5].

The lack of clear literature and a unified consensus in the field has induced confusion in those less familiar with the topic of causality. As an outsider to the field it's particularly hard to discern the meaning behind grand-sounding statements like "deep learning can't do causal inference". This confusion is further amplified by a seeming a lack of concrete examples illustrating how to utilise statistical methods to answer causal questions. We aim to address this short-coming, and clearly illustrate and explain how Pearl's causal framework relates to problems and methods encountered in machine learning. Specifically, our objective is to get to the bottom of various claims made by Judea Pearl and others using the language of probabilistic modelling and inference.

Firstly, we aim to introduce the fundamental concepts and challenges relating to causal inference. In section 2 we endeavour to show examples of what constitutes a causal problem, and clearly describe the obstacles and common pitfalls inherent to those problems. Rather than introducing various causal frameworks such as Potential Outcomes or Causal Graphical Models from the start, we first explain the intricacies of causal inference through a unifying and familiar toolset of probabilistic modelling. Readers familiar with causal inference may wish to skip this section with the exception of example 1, which introduces a concrete model we'll refer to extensively throughout. We'll then show concrete examples of causal questions such as *interventions* (section 3) and *counterfactuals* (section 7), and in each case describe how one can approach tackling these within the probabilistic modelling framework.

Secondly, we'll address the claim that you *need* a bespoke causal framework or notation to do causal inference. Pearl and others have alleged that without the *do*-operator, attempting to answer causal questions is all but hopeless. In this note, we'll show that you *can* in fact approach causal inference without resorting to these tools; the probabilistic modelling approach is sufficient. We'll relate the probabilistic modelling approach introduced in sections 3 and 7 to other causal inference frameworks. Namely, we'll introduce the methodology advocated in Pearl's "The Book of Why", including Causal Graphical Models (section 4) and Structural Causal Models (section 7.3), and show how these can be made equivalent to the probabilistic modelling approach. We'll see under what conditions they are a special case of said approach. In the context of causal inference within the realm of machine learning, we'll discuss the nature and the proposed utility of the *do*-notation, and demystify the purpose of the *do*-calculus (section 5).

Lastly, we'll demonstrate how deep learning tools can be used within a probabilistic modelling programme to tackle causal inference problems. We'll see concrete illustrative examples of inference problems on which methods from supervised and unsupervised learning — such as deep neural networks and amortised variational inference — can be applied.

Hopefully, by the end of this note, the reader should come away with a clear understanding of how the methods from machine learning and statistics can be used to tackle causal problems, and a sense of appreciation for the challenges inherent in causal inference.

PREREQUISITES This note assumes familiarity with probabilistic machine learning and with the basics of graphical models; in particular, knowledge of Bayesian Networks — on which much of Pearl's causal framework is based — and *d*-separation will be useful. For an introduction to graphical models and Bayesian Networks see e.g. [6, §8][7][8]. For an introduction to probabilistic machine learning see [9].

NOTATION We refer to a random variable with a capital letter (e.g., X), the value it obtains as a lowercase letter (e.g., x), and a collection or a vector of random variables with boldface font (e.g., $\mathbf{X} = \{X_1, \ldots, X_n\}$). We will write $p(x, y)$ to denote probability density/mass function $p(X = x, Y = y)$ whenever it's apparent from the context which random variables the function is referring to. Sometimes we'll also write $p_{\boldsymbol{\theta}}(\cdot)$ in place of $p(\,\cdot\,|\boldsymbol{\theta})$ to signify that this distribution is modelled with a statistical model parameterised by $\boldsymbol{\theta}$.

2 Key characteristics of a causal inference problem

In this section, we want to begin by highlighting the key characteristics and challenges underlying causal inference. We will give a concrete example of a causal inference problem, illustrate the pitfalls of a naïve approach, and demonstrate a need to go beyond modelling just the probability distribution from which the samples are drawn to answer the intended causal question. We'll start by discussing how causal inference relates to the tasks most commonly encountered in machine learning, and show that the standard approach to solving these problems is insufficient for causal inference.

Typically in probabilistic machine learning, the aim is to model a probability distribution given samples from that distribution. For instance, in supervised learning, the typical objective is to model a conditional distribution $p(\mathbf{y}|\mathbf{x})$ given i.i.d. samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}$ therefrom. In generative modelling, on the other hand, the goal is to model (or sample from) the joint distribution $p(\mathbf{y})$. More broadly, in those tasks, the general aim is to train a model that will generalise predictions to new samples from the same *observed* distribution.

In many other machine learning tasks we might want to go beyond modelling the observed distribution — we want our model to generalise to other data distributions. For instance, in domain adaptation, the goal is to generalise what the model has learnt in one domain to another domain. This, of course, necessitates further assumptions about how these two settings are related; we need an inductive bias that connects the two. As we'll discuss below, causal inference problems fall into this latter category.

To make this apparent, let's examine a concrete question of causal nature. Consider the task of inferring the efficacy of aspirin on the duration of a headache; we want to answer: "what will be the expected duration of my headache Y on average if I decide take a given dose of aspirin T"? This question would be considered an instance of a causal inference question pertaining to an interpretation of the *causal effect* of the aspirin dose on the headache duration.

One straightforward approach would be to fit a supervised model to predict Y given T — whether this be a linear regression model, a neural network or otherwise. If it's a probabilistic model, it would estimate $p(Y\!=\!y|T\!=\!t)$[1]. When the model is fit with observed data, however, the conditional distribution $p(Y\!=\!y|T\!=\!t)$ might not necessarily correspond to the desired answer to the aforementioned question.

Let's consider what would go wrong with the naïve approach: predicting Y given T. In the observed setting, the headache duration Y presumably depends on the initial headache severity Z. Initial headache severity, however, likely also affects the choice of the treatment T — what dose of aspirin to take. Namely, people suffering from a more severe headache are more likely to take a higher dose of aspirin. Likewise, knowing that someone has taken a low dose means that in all likelihood their headache was not that severe. It should be evident that conditioning on T affects our belief about the initial headache severity Z, as the decision to take a certain dose of aspirin is itself based on the headache severity. Because Z also directly affects the headache duration Y, we could say that the headache severity Z *confounds* the relationship between T and Y.

This is a problem as we are interested in the isolated effect of the aspirin dose. If we intervened to administer someone a higher dose, we do not believe that would affect their initial headache severity. On the other hand, conditioning on a higher value of T would increase our expectation over what the initial headache severity might have been. This illustrates the difference between *observing* that a given aspirin dose was taken, and contemplating the effects of an *action* to take a certain dose. The questions relating to the former can be answered by conditioning on a dose $T = t$. We'll illustrate how to tackle questions relating to the latter in the next section. Before we do so, we'll give a concrete example to give some intuition for the effects of confounding and what might go wrong if use the naïve approach to esimate the effects of a hypothesised action.

The below example makes the effect of confounding quantitatively clear with a concrete model that we'll refer to frequently throughout.

[1] Alternatively, it might predict the most likely Y given the observed T, or a risk minimising value.

Example 1: Aspirin's Efficacy on Headache Duration

Inference question

We're interested in the question of the efficacy of aspirin. Specifically, we want to know: by how much does doubling the aspirin dose reduce the headache duration on average? To answer this question, we can use observational data coming from e.g. a survey on aspirin's effect on headache duration. In the data, the subjects suffering from a headache who took aspirin self-reported their initial headache severity Z, the ingested aspirin dose T, and the subsequent headache duration Y.

Model

Let's start by specifying a model for the data-generating process:

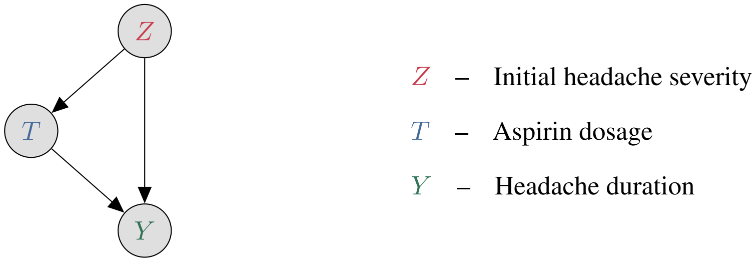


Figure 1: Graphical model for the observational data on aspirin's effect in a population.

Remark. ***Log-normal Distribution.*** *A variable X is distributed according to a log-normal distribution with parameters μ, σ^2 when its logarithm is distributed according to a Gaussian distribution. In other words:*

$$X \sim \log \mathcal{N}\left(\mu, \sigma^2\right) \qquad \textit{iff} \qquad \log X \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

A notable property is that product of two log-normal variables is still log-normal, and a log-normal variable raised to a constant power is also log-normal.

In the model, we assume that every variable is distributed according to a log-normal distribution like so:

$$\begin{aligned} Z &\sim \log \mathcal{N}\left(\mu_Z, \sigma_Z^2\right) & & \\ T &= Z^a \varepsilon_T & &\text{where } \varepsilon_T \sim \log \mathcal{N}\left(0, \sigma_T^2\right) \\ Y &= \frac{Z^b}{T^c} \varepsilon_Y & &\text{where } \varepsilon_Y \sim \log \mathcal{N}\left(0, \sigma_Y^2\right) \end{aligned} \tag{1}$$

Here, $\theta = \{a, b, c, \mu_z, \sigma_Z, \sigma_T, \sigma_Y\}$ are the parameters to be inferred from observed data.

As can be seen in the set of equations 1, the parameter c captures the strength of the effect of aspirin dosage on headache duration. The larger the value of c, the more effective aspirin is. In the special case when $c{=}0$: $Y{=}Z^b \varepsilon_T$; i.e. the amount of aspirin taken has no effect on the final headache duration. Similarly, a dictates how the headache severity affects the decision to take a given dose of aspirin, and b controls how headache severity affects headache duration. We would reasonably expect $a > 0$ and $b > 0$, as a higher initial headache severity usually makes people take more aspirin, and we would expect it also makes the headache last for longer.

Using the properties of log-normal distributions, it can then be seen that the variables are jointly distributed according to a multivariate log-normal (see Appendix A.1):

$$\begin{bmatrix} Z \\ T \\ Y \end{bmatrix} \sim \log \mathcal{N}\left(\begin{bmatrix} \mu_Z \\ a\mu_Z \\ (b-ac)\mu_Z \end{bmatrix}, \begin{bmatrix} \sigma_Z^2 & a\sigma_Z^2 & (b{-}ac)\sigma_Z^2 \\ a\sigma_Z^2 & a^2\sigma_Z^2{+}\sigma_T^2 & a(b{-}ac)\sigma_Z^2{-}c\sigma_T^2 \\ (b{-}ac)\sigma_Z^2 & a(b{-}ac)\sigma_Z^2{-}c\sigma_T^2 & (b{-}ac)^2\sigma_Z^2{+}c^2\sigma_T^2{+}\sigma_Y^2 \end{bmatrix} \right) \tag{2}$$

Inference

Let's go through the steps of the naïve approach of estimating the efficacy of aspirin by estimating $p(Y{=}\ y|T{=}\ t)$. Assume that we have an accurate estimate for all the model parameters $\boldsymbol{\theta}$ from the observed data. We can then obtain the expected value of Y given T from the joint distribution (see Appendix A.2):

$$\mathbb{E}[Y|T] = \exp\left(\frac{a(b-c)\sigma_Z^2 - c\sigma_T^2}{\sigma_Z^2 + c^2\sigma_T^2 + \sigma_Y^2}(T - a\mu_z) + \underbrace{\text{const.}}_{\substack{\text{independent}\\ \text{of } T}}\right) \tag{3}$$

Let's investigate this expression. In the special case when $c = 0$, this would reduce to:

$$\mathbb{E}[Y|T] = \exp\left(\underbrace{\frac{ab\sigma_Z^2}{\sigma_Z^2 + \sigma_Y^2}}_{\text{positive whenever } ab>0}(T - a\mu_z) + \text{const.}\right)$$

In this case, when $a > 0$ and $b > 0$, we will observe that the expected headache duration goes up with the ingested aspirin dose T, even though we specified that the aspirin dose has no effect on the headache by setting $c = 0$.

Depending on the parameters of the model, we might observe that people who take a higher dose of aspirin have a longer headache on average even when $c > 0$ (i.e. when aspirin has a remedial effect on the headache duration). From eq. 3 it can be seen that this trend will arise whenever $a(b-c)\sigma_Z^2 - c\sigma_T^2 > 0$. The longer headache in individuals who had taken a higher dose is of course due to a more severe initial headache, rather than due to adverse workings of aspirin. Figure 2 illustrates this phenomenon, which is often referred to as Simpson's paradox[a] [10][7, §6.1].

Hence, it should be clear from this example that by estimating the conditional distribution in a supervised learning fashion we would not obtain the desired answer about the efficacy of aspirin.

[a] Although hopefully given the probabilistic exposition of the problem it doesn't appear like a paradox at all.

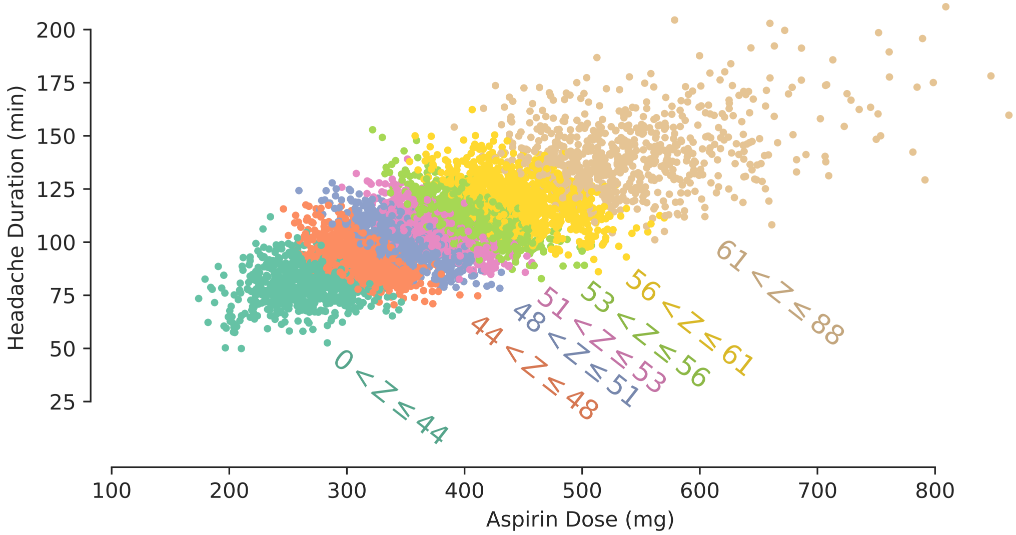


Figure 2: The distribution of samples of the headache duration against aspirin dose generated from the log-normal aspirin model. The plot shows headache duration increasing with aspirin dose taken; however, for any group of people with a narrow range of headache severities, the trend is opposite. Parameters for the model are $a{=}1.5$, $b{=}2.68$, $c{=}1.0$, $\mu_Z{=}3.95$, $\sigma_Z = 0.15$, $\sigma_T = 0.07$ and $\sigma_Y = 0.05$.

Why is this confounding an issue? In a decision-making context — which is the context in which many machine-learning algorithms are deployed — we can essentially change the mechanism by which the dose is selected. For instance, we can give someone a recommendation based on the insights from the data, or deploy a machine learning system that prescribes its users a given personalised dose. Hence, our insights ought to generalise to the settings where the current dose-determining mechanism is no longer in place.

To actually infer aspirin's efficacy, we would have to decouple the dependence of the ingested dose on the initial headache severity. We could do so by obtaining data from an experimental setting instead. Alternatively, building on the model of the observed scenario, we could model the setting in which everyone is administered a fixed dose regardless of their headache severity. Then, using the data from the observed setting, we would hopefully be able to answer questions in the other setting. We'll show in the section below how we could specify such a model. It should, however, be clear that we would be dealing with two different settings overall: the observational world in which the subjects make the decision about their aspirin dose based on the headache severity, and another one in which the dose-deciding policy has been altered — we are manipulating the system to decide someone's dose. It should be evident that data in these two settings, in general, is distributed according to different distributions.

The above preliminary discussion highlights that inference corresponding to causal questions often requires generalising beyond the observed data distribution. As we'll see in the following sections, significant modelling assumptions are necessary to answer causal questions, and we'll discuss how we can specify these assumptions by defining a joint model over all the tasks/settings/distributions that we care about.

3 Interventions

In this section we will demonstrate a particular type of a causal inference question: interventions. The question about the efficacy of aspirin from the section above is an instance thereof. We will discuss what modelling assumptions would allow us to answer such questions, and how to approach interventional inference using probabilistic modelling. We will also introduce various frameworks relating to interventional inference, such as Pearl's Causal Bayesian Networks and the *do*-calculus, and relate them to the probabilistic modelling paradigm.

Returning to the aspirin problem from the previous section, we saw that using the conditional distribution $p(Y{=}y|T{=}t)$ did not correspond to the intended question about the efficacy of aspirin. The mechanism by which the subjects decided their aspirin dose resulted in headache severity confounding the relationship between T and Y. To isolate the effect of aspirin, we could ask what would happen if we hypothetically *intervened* to assign people a higher dose. Such an intervention would change the mechanism by which a person's dose is determined; the probability distribution over the dose conditioned on the headache severity in the hypothetical intervened-upon setting would be different.

Let's consider how we could model such an intervention. In a probabilistic framework, the first step in any inference procedure would be to write down all the assumptions about the problem. From these assumptions, a joint distribution over all variables and settings under consideration should follow. That joint distribution can then be queried for any quantities of interest. Below, we're going to follow this probabilistic modelling approach, and write a joint distribution over the original *unintervened* world and the new *intervened* world by first specifying exactly how they are related.

***Example 2:* Interventions on Aspirin Dose**

INFERENCE QUESTION

In the observed world, in which the subjects decide their own aspirin dosage, we've surveyed people to collect a dataset $\mathcal{D} = \{(z_i, t_i, y_i)\}_{i=1}^{N}$. We want to use this dataset to answer the question: what is the expected effect of intervening to assign someone a given dose of aspirin T on their headache duration Y on average? By how much would doubling someone's assigned aspirin dose reduce their headache duration?

MODEL AND ASSUMPTIONS

Let's begin by setting up a model in which the two settings of interest — the observed and the interventional — are modelled explicitly. In addition to the random variables corresponding to the observed dataset $\{(Z_i, T_i, Y_i)\}_{i=1}^{N}$, which we assume has been generated as described in eq. 1 in the previous section, we also define variables Z^*, T^*, Y^*, which correspond to the intervened-upon setting. We can then make assumptions about how these two settings are related and consequently write down a joint distribution over those.

First, we're going to make the assumptions embodied in the graphical model in figure 3.

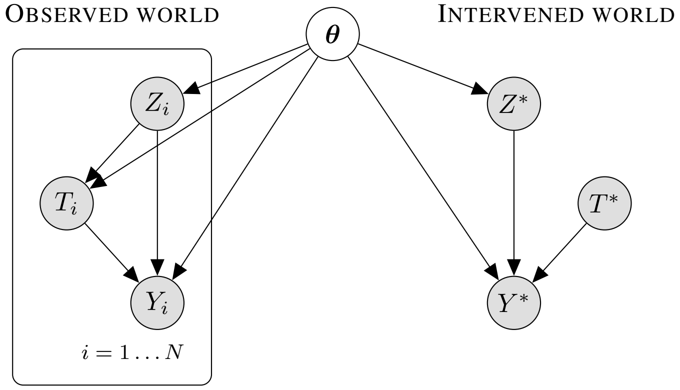


Figure 3: Graphical model for interventional inference in the aspirin example. The nodes Z_i, T_i, Y_i on the left correspond to the cases in the observed world, whereas the nodes Z^*, T^*, Y^* correspond to the hypothetical world in which we have intervened on T^* — it's no longer dependent on Z^*. Parameters $\boldsymbol{\theta}$ are shared between the two worlds allowing observations in the real world to inform inference in the intervened-upon one.

Based on this graphical model, we can write down a full joint distribution for the model (we'll show how the graphical model embodies these assumptions later in this section). For notational convenience, we'll write $q(\cdot)$ in place of $p(\cdot)$ when referring to distribution functions over interventional variables Z^*, T^*, Y^* when using the machine learning short-hand; for instance, we'll write $q_{\boldsymbol{\theta}}(y|t, z)$ in place of $p_{\boldsymbol{\theta}}(Y^* = y|T^* = t, Z^* = z)$.

$$p(z^*, t^*, y^*, \boldsymbol{\theta}, \mathcal{D}) = \tag{4}$$

$$= p(z^*, t^*, y^* | \boldsymbol{\theta}, \cancel{\mathcal{D}}) p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{5}$$

$$= q_{\boldsymbol{\theta}}(z^*, t^*, y^*) \left(\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(z_i, t_i, y_i) \right) p(\boldsymbol{\theta}) \tag{6}$$

$$= q_{\boldsymbol{\theta}}(z^*) q_{\boldsymbol{\theta}}(t^* | \cancel{z^*}) q_{\boldsymbol{\theta}}(y^* | z^*, t^*) \left(\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(z_i) p_{\boldsymbol{\theta}}(t_i|z_i) p_{\boldsymbol{\theta}}(y_i|z_i, t_i) \right) p(\boldsymbol{\theta}) \tag{7}$$

$$= \underbrace{q_{\boldsymbol{\theta}}(z^*) q_{\boldsymbol{\theta}}(t^*) q_{\boldsymbol{\theta}}(y^* | z^*, t^*)}_{\text{Interventional world}} \underbrace{\left(\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(z_i) p_{\boldsymbol{\theta}}(t_i|z_i) p_{\boldsymbol{\theta}}(y_i|z_i, t_i) \right)}_{\text{Observed world}} p(\boldsymbol{\theta}) \tag{8}$$

Let's go through and motivate the assumptions we exploited in the above derivation.

Firstly, in lines 5 and 6 we used the assumption given by the graphical model that all the observed examples (Z_i, T_i, Y_i) and the interventional outcome (Z^*, T^*, Y^*) are independent given the model parameters $\boldsymbol{\theta}$. I.e. once we know the parameters, the duration of one person's headache, its severity and the dose they took is independent of how much aspirin someone else had taken, how long their headache lasted for, and its severity.

Secondly, in line 7 we used the fact that the aspirin dose T^* in the interventional setting is independent of the headache severity Z^*, hence $q_{\boldsymbol{\theta}}(t^*|z^*) = q_{\boldsymbol{\theta}}(t^*)$. Let's consider why this assumption makes sense. We're interested in a hypothetical intervention where we *assign* people a higher dose. To isolate the effect of aspirin, we are enquiring about an assignment of a given dose regardless of the subject's initial headache severity. Hence, the dose T^* in the interventional setting ought to be independent of the headache severity Z^*.

These are all the assumptions we used to get us to eq. 8. However, the exact form of the distributions $q_{\boldsymbol{\theta}}(z^*)$, $q_{\boldsymbol{\theta}}(t^*)$, $q_{\boldsymbol{\theta}}(y^*|z^*, t^*)$ is yet to be defined.

Let's first consider $q_{\boldsymbol{\theta}}(y^*|z^*, t^*)$. We believe that the physical mechanism determining the headache duration Y^* based on the initial headache severity Z^* and aspirin dose ingested T^* remains unchanged in the interventional setting; in other words, the physiological properties of the patients in the hypothetical setting are the same as in the observed one. Hence, in the intervened-upon world, we assume that conditional distribution of the headache duration Y^* conditioned on T^* and Z^* — $p_{\boldsymbol{\theta}}(Y^* = y^*|T^* = t^*, Z^* = z^*)$ — remains the same as in the observed world — $p_{\boldsymbol{\theta}}\ (Y_i = y|T_i = t, Z_i = z)$. Hence, we can write our assumption as:

$$p_{\boldsymbol{\theta}}(y|t, z) = q_{\boldsymbol{\theta}}(y|t, z) \tag{9}$$

Next, to represent the belief that we are manipulating the system to assign a specific dose t^* independently of Z^*, we can specify:

$$T^* = t^* \qquad \iff \qquad p(T^* = t) = \delta(t^* - t) \tag{10}$$

where $\delta(\cdot)$ is a Dirac delta function. Setting T^* to be delta distributed results in the property:

$$p(\cdot) = \int p(\cdot, T^* = t)\ dt = \int p(\cdot|T^* = t)\,\underbrace{p(T^* = t)}_{\delta(t^* - t)}\,dt = p(\cdot\ |T^* = t^*) \tag{11}$$

I.e. marginalising out T^* is equivalent to conditioning on $T^* = t^*$. This identity will come in useful below.

Lastly, we can specify the distribution on the headache severity Z^* in the intervened-upon setting. In this case, we can assume that we are interested in the effect in a person drawn from the same population as that in the observed data. Hence, we can specify that Z^* is distributed in the same way as Z:

$$q_{\boldsymbol{\theta}}(z) = p_{\boldsymbol{\theta}}(z) \tag{12}$$

We can then fully specify the joint in terms of probability density function which we know:

$$p(z^*, t^*, y^*, \boldsymbol{\theta}, \mathcal{D}) = \underbrace{p_{\boldsymbol{\theta}}(z^*) q_{\boldsymbol{\theta}}(t^*) p_{\boldsymbol{\theta}}(y^*|z^*, t^*)}_{\text{Interventional world}} \underbrace{\left(\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(z_i) p_{\boldsymbol{\theta}}(t_i|z_i) p_{\boldsymbol{\theta}}(y_i|z_i, t_i)\right)}_{\text{Observed world}} p(\boldsymbol{\theta}) \tag{13}$$

Given the joint, probability theory tells us how to find all the marginal distributions, conditional distributions and expectations of interest. How to actually compute the numerical values for these expressions (possibly approximately) then falls right within the realms of machine learning and statistics.

Inference

Returning to the inference question, we wanted to know what's the expected headache duration in the intervened world in response to a given dose t^* (which is a parameter of the model). Hence, we can compute the expected value of Y^* given the observed data $\mathcal{D}$:

$$\begin{aligned}
\mathbb{E}[Y^*|\mathcal{D}] &= \int y^* p(y^*|\mathcal{D}) dy^* && (14)\\
&= \int y^* p(y^*|t^*, \mathcal{D}) dy^* \qquad \text{(using eq. 11)} && (15)\\
&= \iiint y^* p(y^*, z^*, \boldsymbol{\theta}|t^*, \mathcal{D}) dz^* dy^* d\boldsymbol{\theta} && (16)\\
&= \iiint y^* p(y^*|z^*, t^*, \boldsymbol{\theta}, \cancel{\mathcal{D}}) p(z^*|\boldsymbol{\theta}, \cancel{t^*}, \cancel{\mathcal{D}}) p(\boldsymbol{\theta}|\cancel{t^*}, \mathcal{D}) dz^* dy^* d\boldsymbol{\theta} && (17)\\
&= \iiint \underbrace{y^* q_{\boldsymbol{\theta}}(y^*|z^*, t^*) q_{\boldsymbol{\theta}}(z^*) dz^* dy^*}_{\substack{\text{Calculate expectation in the intervened}\\ \text{world conditioned on posterior over } \boldsymbol{\theta}}} \quad \underbrace{p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}}_{\substack{\text{Infer model parameters}\\ \text{using observed data } \mathcal{D}}} && (18)
\end{aligned}$$

where the cancellations in eq. 17 follow from the conditional independencies in the model.

Note that on line 18, all the distribution functions in the inner integral are the same as these in the observed distribution. For the log-normal aspirin model, we can actually calculate the inner expectation exactly (Appendix A.3):

$$\iint y^* p_{\boldsymbol{\theta}}(y^*|z^*, t^*) p_{\boldsymbol{\theta}}(z^*) dz^* dy^* = (t^*)^{-c} \exp\left(b\mu_Z + \frac{b^2\sigma_Z^2 + \sigma_Y^2}{2} \right) \tag{19}$$

As expected, the effect of assigning a different dose of aspirin t^* is only determined by the parameter c; as long as $c > 0$, intervening to administer a higher dose will in expectation shorten the headache duration.

In this example, for an appropriate choice of a prior on $\boldsymbol{\theta}$, the integral in eq. 18 could be calculated exactly, as the model is just a Linear Gaussian model in the log-space[a]. Alternatively, we could approximate $p(\boldsymbol{\theta}|\mathcal{D})$ with, for instance, the maximum-a-posteriori (MAP) estimate $p(\boldsymbol{\theta}|\mathcal{D}) \approx \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})$ where $\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

We can then consider the expected value of Y^* under two different interventions $t^* = t_1^*$ and $t^* = t_2^*$. Assuming a point estimate of the parameters, and using eq. 19, we can calculate the effect of doubling the aspirin dose by setting $t_2^* = 2t_1^*$ and considering the ratio of the expected headaches (also know as the *risk ratio*) under the two interventions:

$$\frac{(t_2^*)^{-c} \exp\left(b\mu_Z + \frac{b^2\sigma_Z^2 + \sigma_Y^2}{2} \right)}{(t_1^*)^{-c} \exp\left(b\mu_Z + \frac{b^2\sigma_Z^2 + \sigma_Y^2}{2} \right)} = \left(\frac{t_2^*}{t_1^*} \right)^{-c} = \left(\frac{2t_1^*}{t_1^*} \right)^{-c} = 2^{-c} \tag{20}$$

Hence, according to the model, doubling the aspirin dose has the effect of reducing the headache duration by a factor of 2^{-c}. This factor turned out to be independent of the aspirin dose being doubled in this specific model.

[a] Specifically, when a, b, c are Gaussian distributed, and $\sigma_Z, \sigma_T, \sigma_Y$ follow an inverse-gamma distribution, $p(\boldsymbol{\theta})$ becomes a conjugate prior for this model.

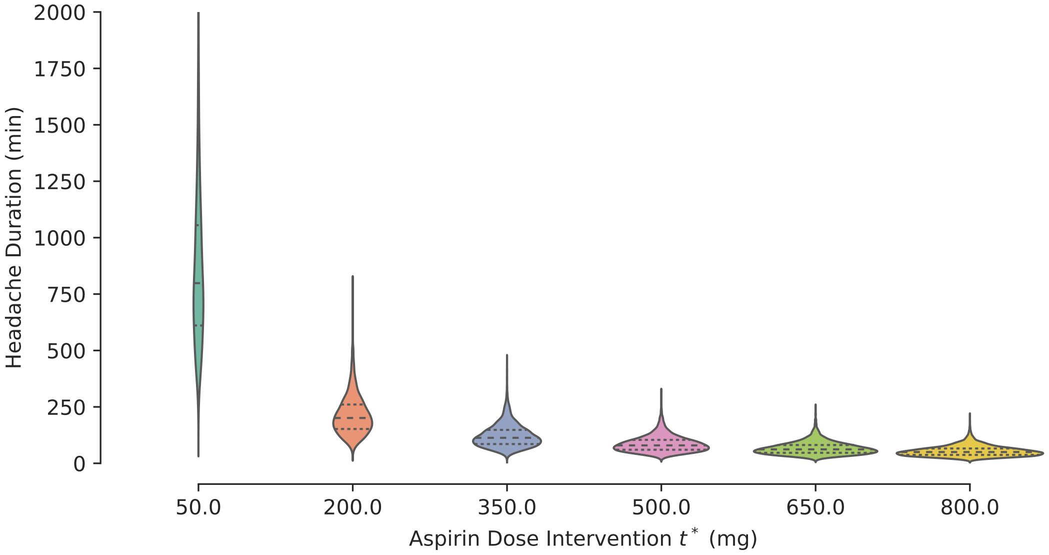


Figure 4: Violin plot of the distributions of the interventional headache duration given different interventions on the aspirin dose t^* in the log-normal aspirin model. The dotted lines demarcate quartiles. The parameters of the model are the same as those in figure 2.

In the example above, we demonstrated a probabilistic modelling approach to solving a causal inference problem. We showed how, by specifying a joint model over all the settings of interest and writing down the relevant assumptions for these settings, we can answer a question of a causal nature. We didn't need any bespoke causal framework or notation to do so.

In the example, we looked at inferring an interventional quantity of interest under one of many possible interventions. The intervention detailed above, in which the treatment (aspirin dose) is independent of the headache severity, allowed us to determine the efficacy of aspirin. We could conceive of different, possibly probabilistic, interventions. For instance, what if we were wondering what would happen if the national department for health issued a new recommendation to increase aspirin dosage? In this interventional setting, the aspirin dose would still likely be dependent on the headache severity; however, as a result of the intervention, $p(T^* = t^* | Z^* = z^*)$ would differ from the observed world. We could specify a similar joint model as above by designating how we believe the decision mechanism for choosing a dose based on the headache severity would change subject to the intervention.

Effects of such "soft" probabilistic interventions are of interest in many domains, especially with regards to policy-making. For example, consider predicting the effect of a regulation banning cigarette advertisement on the national life expectancy. We could construct an equivalent model with T representing cigarette usage, Y the life expectancy, and Z the underlying health attitude. In this case, the regulation is likely to alter the conditional distribution of cigarette usage given someone's health attitude — $p(T^* = t^* | Z^* = z^*)$ — hopefully lowering the expected number of cigarettes consumed for any given z^*; nevertheless, T^* would still be dependent on Z^* after this intervention.

Another thing worth remarking upon is: what if we had data available from both the interventional and the observational settings? For instance, we might have some limited data available from an experimental setting in which patients were assigned a dose at random, but, in addition, we would like to use the data from population surveys where many more examples are available. Again, in this case, specifying a joint model over the two settings as we did before allows for combining the data from the two sources to infer the shared model parameters, which can then again be used for an inference task of interest.

3.1 Specifying the interventional model

We have used a rather informal method to write down our probabilistic model in the previous section. More formally, the methodology employed a graphical model called a *Bayesian Network*. There has been a debate in the literature as to whether Bayesian Networks are sufficient in themselves to support causal inference. In this section, we introduce Bayesian Networks and show how they can be used for modelling interventions. In sections 4 and 7 we will link this approach to the approaches advocated by Pearl and others, hopefully resolving any confusion surrounding their sufficiency for causal inference.

3.1.1 Bayesian Networks

Bayesian Networks provide a way to specify the generative model for the data. We employed a Bayesian Network over both the observed and intervened-upon settings, as shown in figure 3, to deduce the set of conditional independencies in the aspirin intervention example (e.g. in equation 17). Formally, a Bayesian Network can be defined as:

Definition 1. ***Bayesian Network (BN).*** *Random variables $\mathbf{X} = (X_1, \ldots, X_d)$ are a Bayesian Network with respect to a directed acyclic graph $\mathcal{G}$ if the random variables satisfy the following set of independencies[2]:*

$$X_i \perp\!\!\!\perp \mathbf{X}_{\mathbf{ND}_i^{\mathcal{G}}} | \mathbf{X}_{\mathbf{PA}_i^{\mathcal{G}}}$$

where $\mathbf{X}_{\mathbf{ND}_i^{\mathcal{G}}}$ are the non-descendants of X_i in $\mathcal{G}$, and $\mathbf{X}_{\mathbf{PA}_i^{\mathcal{G}}}$ are its parents. In other words, the non-descendants of X_i are independent of X_i given its parents.

The graph $\mathcal{G}$ in a Bayesian Network simply describes the set of conditional independencies between variables in $\mathbf{X}$. This set of conditional independencies leads to a simplified factorisation of the joint distribution:

$$p\left(x_1, \ldots, x_d\right) = \prod_{i=1}^{d} p\left(x_i | \mathbf{x}_{\mathbf{PA}_i^{\mathcal{G}}}\right) = \prod_{i=1}^{d} g_i\left(x_i, \mathbf{x}_{\mathbf{PA}_i^{\mathcal{G}}}\right) \tag{21}$$

[2] When the random variables satisfy this condition, it is often said said that they satisfy the *Markov property* with respect to the graph $\mathcal{G}$.

where $g_i(x_i, x_{\mathbf{PA}_i^{\mathcal{G}}}) = p(x_i|x_{\mathbf{PA}_i^{\mathcal{G}}})$ is the conditional probability density function of x_i given its parents[3]. This factorisation and the set of conditional independencies are two equivalent ways of defining a Bayesian Network; one necessarily implies the other.

3.1.2 Modelling interventions with Bayesian Networks

Let's briefly summarise how we *could* approach interventional inference using Bayesian Networks by defining a joint model over the observed and the intervened-upon setting. This approach summarises the general procedure that we used to model the effects of an intervention in the preceding aspirin example:

1. Specify a Bayesian Network over the observed variables $\boldsymbol{X}$.
2. Define the form of the conditional distributions $p_{\boldsymbol{\theta}}(x_i|\boldsymbol{x}_{\mathbf{PA}_i^{\mathcal{G}}})$ parameterised by some (unknown) parameters $\boldsymbol{\theta}$.
3. Specify a Bayesian Network over the corresponding variables $\boldsymbol{X}^*$ in the intervened-upon setting by considering which dependencies have been altered/removed compared to the observed setting.
4. Define the form of the conditional distributions $q_{\boldsymbol{\theta}}(x_i^*|\boldsymbol{x}^*_{\mathbf{PA}_i^{\mathcal{G}}})$ in the intervened setting by specifying which are to remain the same as in the observed setting, and detailing the specifc form for the remaining ones.
5. The complete joint model over the two settings is constructed from the two Bayesian Networks over $\mathbf{X}$ and $\mathbf{X}^*$ by drawing arrows from $\boldsymbol{\theta}$ to variables in both world, reflecting that the parameters $\boldsymbol{\theta}$ are the same in both settings, and that $\mathbf{X}$, $\mathbf{X}^*$ are independent once the parameter vector is known.

Bayesian Networks are only one of the many tools we could use to define a joint model over the interventional and observed settings; we could specify our assumptions in a variety of other equivalent ways. The key part is to recognise that the variables in the interventional and observational settings might be distributed differently, and then choose a method to define how the two settings are related by defining a joint distribution over $\mathbf{X}, \mathbf{X}^*$. In fact, as we'll later see and discuss in section 9, Bayesian Networks – and graphical models in general – are somewhat limited in that they cannot describe certain settings that, for instance, probabilistic programming languages can.

3.1.3 Challenges of specifying an interventional model

In the aspirin example, it was fairly easy to see how to model the intervention; we could quite intuitively deduce which parts of the intervened-upon world to keep shared with the real world and which ones to alter. However, it is worth noting that in general there is significant nuance to it.

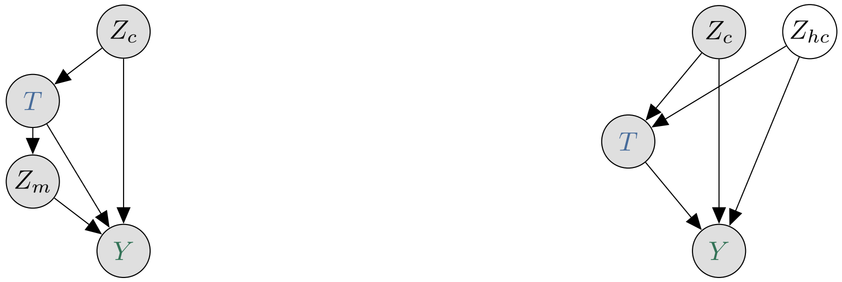
In a more general setting, consider having an outcome variable Y (e.g. headache duration), a treatment variable T (e.g. aspirin dose), and a set of remaining variables $\boldsymbol{Z}$ in the observed world. The initial instinct, based on the aspirin example, might be to factorise the joint distribution $p(\mathbf{z}, t, y) = p(\mathbf{z})p(t|\mathbf{z})p(y|t, \mathbf{z})$ and simply replace the 'treatment-determining' mechanism $p(t|\mathbf{z})$ with a new 'intervention' mechanism $q(t|\mathbf{z})$ to get the joint over the interventional variables. For instance, in the intervened-upon setting, we could specify T to be independent of Z — $q(t|\mathbf{z}) = \delta(t - t^*)$ — mirroring what we did in the aspirin example. Although this was adequate in said example, it is not necessarily so in the general case.

For instance, what if some of the variables in $\mathbf{Z}$ mediate the effect of the treatment T rather than confound it? If $\boldsymbol{Z}$ contained a variable representing measurements of post-ingestion prostaglandin levels, for instance, making prostaglandin levels independent of the aspirin dose could "block" the main mechanism by which the aspirin affects the headache duration. This is illustrated in figure 5a. When we want to infer the effects of an *action* to assign a given aspirin dose, we would imagine that in the intervened-upon world the physical effect of aspirin on the body is still the same as in the observed-world; that is, we would believe that prostaglandin levels are still dependent (in the same way as in the observed setting) on the aspirin dose ingested[4].

[3] In the above definition of Bayesian Networks we considered the case when each node corresponds to one random variable. The definition can be trivially extended to the case in which we multiple random-variables correspond to each node. This could, for instance, occur when dealing with image data.

[4] Note that in this example, which variables in $\boldsymbol{Z}$ mediate and which confound requires a very simple judgement call. Any measurments taken before the treatment decision was made are potential confounders, and any measurements taken after the treatment decision couldn't have possibly influenced the decision to take a given treatment. This is, however, complicated by the fact that there could be other unobserved variables that confound the treatment decision and the measurement taken after that decision was made.

Another important consideration is: what if the there is another hidden confounder that we haven't accounted for? For instance, the subjects in the survey possibly based their aspirin dose decision on whether they had a fever; however, a fever might be indicative of the headache type (migrane, tension headache, etc.), which in turn affects the effectiveness of aspirin. We saw in section 2 how not accounting for a confounder could yield an inference outcome different from the one we intended. In the intervened setting, if we are asking about hypothetically administering someone a given dose regardless of their condition, we implicitly assume the dose assignment to be independent of the fever. However, if the model of the interventional setting doesn't reflect that assumption, as it doesn't account for some hidden confounder, the output of the inference procedure could be heavily misleading. In other words, there would be a mismatch between our interpretation of the output of the procedure, and what it actually corresponds to. In fact, this is one of the fundamental problems of causal inference: the hope that you have accounted for all the factors that could confound the relationship between the quantities of interest.



(a) Graphical model for an aspirin example with a mediator variable Z_m.

(b) Graphical model with an unobserved hidden confounder Z_{hc}.

Figure 5

Judea Pearl's "The Book of Why" [1] and "Causality" [7] do an excellent job showcasing how an incorrect treatment of such scenarios could lead to a misinterpretation of the outcome of an inference procedure. The takeaway message advocated by Pearl is that there is no universal way to generalise from the observed distribution to what we would interpret as an intervened-upon distribution; you need to make assumptions about the generative process for the data, and how an intervention would change it, to be able to do so.

4 Interventions in Bayesian Networks - a concise graphical representation

So far, we've used BNs to specify a joint distribution over the intervened and non-intervened settings encapsulating our assumptions about both worlds and how they are related. We'll refer to this probabilistic modelling approach as the *twin model* approach[5] in this section. This method works. The assumptions conveyed are transparent, and as long as the assumptions are correct, it's an adequate way to approach causal inference.

An alternative approach would be to define a graphical model over a single set of variables in the observed world only — like the one over Z, T, Y in figure 1 — in a way that encompasses the assumptions about how the distribution would change in the interventional setting. Namely, we could specify a *rule* for obtaining a joint distribution in the intervened setting from a graphical model of the observed setting.

Having such a rule can make everything notationally more compact. In such a framework, we'd only have to define the model of the observed setting (adequately), and how to carry out interventional inference would then be implicit in that model. In this subsection, we will discuss how to specify such models with Bayesian Networks and give a rule that would automatically give us the joint in the interventional setting. It is worth mentioning that this approach seems to be more prevalent in Pearl's work, likely as dealing with defining a model over the observed world variables only is more intuitive for practitioners outside the fields of machine learning and statistics. This approach of course has its own caveats, which we'll discuss later in this section.

The key idea to using a single Bayesian Network over the observed world to specify the model of the interventional setting is that, if the graphical model has been constructed in such a way that the directed edges correspond to some autonomous mechanisms in the generative process, then it is conceptually feasible to alter some of these mechanisms

[5] This name is chosen given the fact that the method is analogous to the so called *twin network* method presented by Balke and Pearl [11].

while keeping the others intact. This gives us a clue as to how we could specify a rule for obtaining the intervened-upon distribution from the BN for the observed world. An intervention could be described as an operation in which all the edges going into a node that is being intervened upon are removed and replaced with a new mechanism that determines its value.

Formally, an intervention in a Bayesian Network can be defined as follows[6]:

Definition 2. ***Interventions in Bayesian Networks.*** *Consider a Bayesian Network on* $\mathbf{X} = \{X_1, \ldots, X_d\}$ *with a graph* $\mathcal{G}$ *that entails the factorisation:*

$$p(\mathbf{x}) = \prod_{i=1}^{d} g_i(x_i, \mathbf{x}_{\mathbf{PA}_i^{\mathcal{G}}})$$

where $g_i(x_i, \mathbf{x}_{\mathbf{PA}_i^{\mathcal{G}}}) = p(x_i|\mathbf{x}_{\mathbf{PA}_i^{\mathcal{G}}})$ *is the conditional probability density function of* x_i *given its parents.*

An intervention on node j *yields a new BN on* $\mathbf{X}^* = \{X_1^*, \ldots, X_d^*\}$ *with a graph* $\mathcal{G}^*$*. The graph* $\mathcal{G}^*$ *is identical to* $\mathcal{G}$ *with the exception that the edges going into* j *are replaced with a new set of edges from a new set of parents* $\mathbf{PA}_j^*$ *(specified as part of the intervention). The intervened-upon BN has a new entailed joint distribution:*

$$q(\mathbf{x}) = g^*(x_j, \mathbf{x}_{\mathbf{PA}_j^*}) \prod_{i \in \{1,\ldots,d\} \setminus \{j\}} g_i(x_i, \mathbf{x}_{\mathbf{PA}_i^{\mathcal{G}}}) \tag{22}$$

Here, the new conditional distribution function $g^*(x_j, \mathbf{x}_{\mathbf{PA}_j^*}) = q(x_j|x_{\mathbf{PA}_j^*})$ *replaces the previous one* $g(x_j, \mathbf{x}_{\mathbf{PA}_j^{\mathcal{G}}})$ *in the factorisation, while the remaining conditionals are kept the same.*

Figure 6 illustrates the result of an application of this rule.

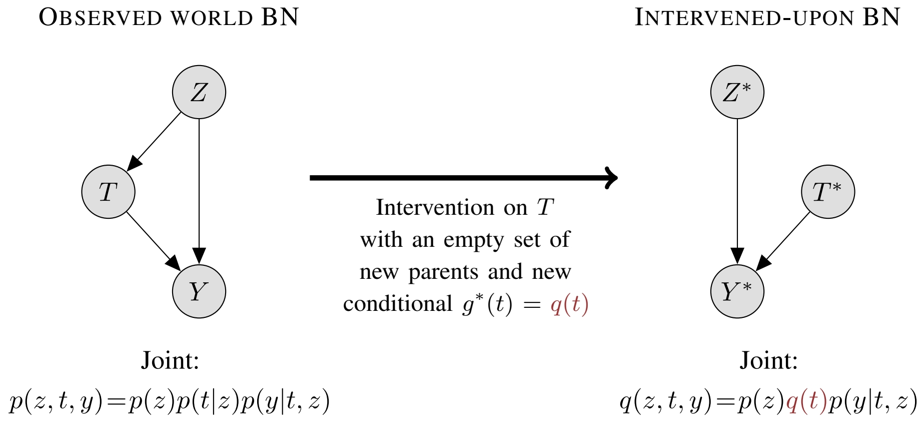

OBSERVED WORLD BN
INTERVENED-UPON BN
Z
T
Y
Intervention on T with an empty set of new parents and new conditional $g^*(t) = q(t)$
Z^*
T^*
Y^*
Joint:
$p(z,t,y)=p(z)p(t|z)p(y|t,z)$
Joint:
$q(z,t,y)=p(z)q(t)p(y|t,z)$

Figure 6: Illustration of the rule in def. 2 for obtaining the joint on the intervened-upon variables from the BN over the observed variables.

An intervention is a purely mathematical operation on a Bayesian Network and its graph. Intuitively, the intervention simply results in a new distribution where we've changed a mechanism determining one of the variables. The hope is that, if we ascribe some real-world interpretation to the edges as mechanisms determining the value of a variable in the first place — you may choose to call this interpretation causal — the altered Bayesian Network hopefully has some feasible and realisable interpretation; in the aspirin example, this operation applied to T had the interpretation of changing the aspirin-dosage recommendations, or administering someone a given dose independently of the headache severity.

The most prevalent type of an intervention is an *atomic intervention*. In an atomic intervention, no new edges are added in the modified model ($\mathbf{PA}_j^* = \emptyset$) and the random variable at node j is set to a constant value c_j, i.e. $g^*(x_j) = \delta(x_j - c_j)$[7]. Intuitively, the modification to the graphical model compromises removing edges going into

[6] Here, we continue to use the notation $q(\cdot)$ when referring to probability density/mass functions over the variables in the intervened setting, and $p(\cdot)$ in the observed setting.

[7] In the discrete case, a Kronecker delta function can equivalently be used.

j and forcing the variable X_j to take on some fixed value. This corresponds exactly to the type of intervention we considered in the aspirin example. The post-intervention distribution for the case of an atomic intervention can be written as[8]:

$$q(\mathbf{x}) = \delta(x_j - c_j) \prod_{i \in \{1,\ldots,d\} \setminus \{j\}} g_i(x_i, \mathbf{x}_{\mathbf{PA}_i^g})$$

For atomic interventions, and in fact for any intervention with an empty set of parents ($\mathbf{PA}_j^* = \emptyset$), the conditional distribution $q(\mathbf{x}_{-j}|x_j){=}q(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_d|x_j)$ simplifies to:

$$q(\mathbf{x}_{-j}|x_j) = \left(g^*(x_j) \prod_{i \in \{1,\ldots,d\} \setminus \{j\}} g(x_i, \mathbf{x}_{\mathbf{PA}_i^g}) \right) \frac{1}{g^*(x_j)} = \prod_{i \in \{1,\ldots,d\} \setminus \{j\}} g(x_i, \mathbf{x}_{\mathbf{PA}_i^g}) \tag{23}$$

whenever x_j is in the support of $g^*(x_j)$. I.e., the conditional distribution given $X_j = x_j$ is the same independently of the choice of $g^*(x_j)$, or whether the intervention is atomic or probabilistic.

4.1 Interventions and randomised control trials

Linking interventional distributions obtained through this rule to experimental studies helps to establish some intuition for its workings. Specifically, atomic interventions have an intuitive relation to randomised control trials (RCTs).

In RCT studies, one randomly assigns the value (treatment) to a variable of interest X_j irrespective of other factors. For instance, patients would be administered placebo or the real drug based solely on 'the flip of a coin'. This can of course be represented by an intervention in a Bayesian Network (assuming original BN represents a non-experimental setting), where the new set of parents of X_j is an empty set, and its value is determined purely at random, according to some distribution $g^*(x_j)$ (e.g. $X_j \sim \text{Bern}(\frac{1}{2})$ if $X_j \in \{0,1\}$ is picked based on a coinflip). As such, it can be seen that in the BN corresponding to the randomised control trial, the conditional distribution $p_{RCT}(\mathbf{x}_{-j}|x_j)$ is the same as the conditional distribution for a BN resulting from an atomic intervention fixing the value of X_j to x_j:

$$p_{RCT}(\mathbf{x}_{-j}|x_j) = \frac{p_{RCT}(\mathbf{x})}{g*(x_j)} = \left(g^*(x_j) \prod_{i \in \{1,\ldots,d\} \setminus \{j\}} g(x_i, x_{\mathbf{PA}_i^g}) \right) \frac{1}{g^*(x_j)} = \prod_{i \in \{1,\ldots,d\} \setminus \{j\}} g(x_i, x_{\mathbf{PA}_i^g})$$

4.2 *do*-notation for interventions

The *do*-notation can be used as a notational tool for describing these alterations to a Bayesian Network. For instance, one could write $q(\mathbf{x}) = p_{do(X_k := g^*(\cdot|x_{\mathbf{PA}_\mathbf{k}^*}))}(\mathbf{x})$ to denote an intervention on node k with a new set of parents $\mathbf{PA}_k^*$ and a new conditional distribution given parents $g^*(\cdot|x_{\mathbf{PA}_\mathbf{k}^*})$. For atomic interventions, the notation is often further simplified to:

$$p(\mathbf{x}|do(X_k = c_k)) = p_{do(X_k := \delta(\cdot\ -\ c_k))}(\mathbf{x}|X_k = c_k)$$

The conditioning operator in this notation signifies that this joint is equal to the distribution in an RCT conditioned on X_k.

4.2.1 Confounding

The above definition of interventions, and atomic interventions specifically, leads to a succinct definition of *confounding*. The effect of T on Y is said to be confounded whenever $p(y|t) \neq p(y|do(T = t))$.

4.3 Markov Equivalence Class

We mentioned that there are caveats to defining a single Bayesian Network on the observed world, and the BN on the interventional world implicitly from it. Equivalence classes of BNs make doing so ambiguous.

Namely, there can be multiple different Bayesian Networks with different graphs that give the same set of conditional independencies, and hence entail the same factorisation of the joint distribution. They are said to belong to the same

[8] For the discrete case, when $\delta(x_j - c_j) \in \{0,1\}$ is a Kronecker delta, this neatly simplifies to: $q(\mathbf{x}) = \prod_{i \in \{1,\ldots,d\} \setminus \{j\}} g(x_i, \mathbf{x}_{\mathbf{PA}_i^g})$ if $x_j = c_j$ and 0 otherwise.

Markov Equivalence Class. Recall that in the definition of a Bayesian Network, the graph was solely used to define said conditional independencies. Hence, all members of a Markov Equivalence Class imply the same restrictions on the joint distribution.

To give a concrete example of a Markov Equivalence Class, consider the graph in fig. 1 for the aspirin example. That graph does not entail any conditional independencies; the factorisation that we obtain from that Bayesian Network — $p(y, t, z) = p(y|t, z)p(t|z)p(z)$ — is valid for *any* probability distribution by the product rule of probability functions. Hence, any permutation of the edges, as shown in fig. 7, would yield the same set of conditional independencies.

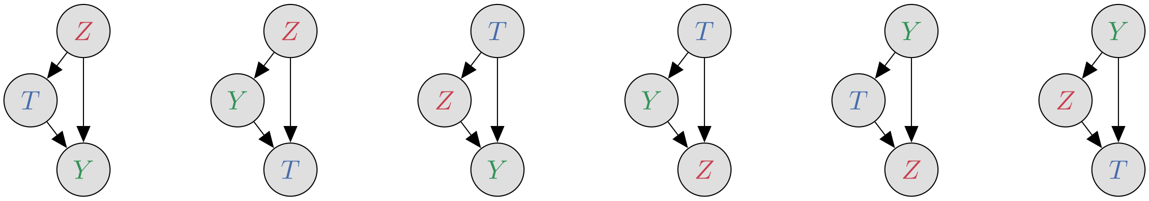


Figure 7: Bayesian Networks entailing the same set of conditional independencies (and consequently the same factorisation of the joint distribution).

For the purposes of fitting the observed distribution only, it makes no difference which of the Bayesian Networks in the Markov equivalence class is being considered, as they all imply the same restrictions (in terms of conditional independencies) on the observed distribution. Hence, if we tried to infer the graph by estimating the conditional independencies[9] in the observed data, we could only distinguish between different Markov Equivalence Classes, but not between the members of each class. Even in the limit of infinite data — there is no statistical test on the observed data to uniquely determine a member from within the equivalence class.

Despite the fact that the exact graph non-uniquely defines these independence assumptions, different graphs in the equivalence class can entail different interventional distributions through applications of the rule in def. 2. Which graph from the Markov Equivalence Class the rule is being applied *does* make a difference. For instance, an atomic intervention on T in the 6th graph in figure 7 would make T^* independent of Z^* and Y^* in the intervened-upon setting, whereas same intervention on T for the 3rd graph wouldn't introduce any independencies.

Hence, if we use a single BN on the observed variables and intend to use it for interventional inference, we have to be careful to select the right graph that will give us the interventional distribution that we desire. In other words, when specifying a BN on the observed world in this way, we have to additionally make the assumption that for any intervention considered, applying the rule in def. 2 to the BN will yield the interventional distribution that we are interested in. We could fuse the definition of Bayesian Networks with this extra assumption to define a new model class — a *Causal Bayesian Network* [7, p. 23]. By defining a Causal Bayesian Network, we are explicitly saying that for any intervention considered, applying the rule in definition 2 will yield the *right* interventional distribution.

Although this corresponds to the definition given by Pearl in his book "Causality"[7, p. 23], it differs from the definitions given by other authors. For instance Schölkopf, Janzing, and Peters define a Causal Graphical Model (their name for a Causal Bayesian Network) as a Bayesian Network in which each directed edge is assumed to correspond to a 'causal' relationship [4]. However, as 'causal' is an overloaded and ambiguous term, this definition somewhat obfuscates the true assumptions given above. Namely, that with a Causal Bayesian Network we are implicitly defining an interventional distribution for every intervention possible [10]

Note that this is a complete non-issue when using a joint model over all the settings of interest. By specifying a complete joint model over all the relevant tasks — like we did in the aspirin example — we are forced to explicitly make the assumptions about how these tasks are related. Hence, you can either choose to specify a single model over the observed setting while being mindful of the implicit assumptions and the Markov Equivalence Classes, or you can specify a joint model where the assumptions are crystal clear.

[9] Which is a challenging statistical problem in itself in the finite sample regime.

[10] Pearl would motivate the definition of the term *causal* by saying that one variable causally affects the other when an intervention on that variable would change the distribution on the other variable. This interpretation of the word causal is later used to define a Causal Bayesian Network/Causal Graphical Model by some authors, and to state that in a Causal Bayesian Network the rule in def. 2 will give the desired interventional distribution. This partly circular definition obfuscates where and in what form subjective expert opinion comes in to define these assumptions.

5 Inference in interventional models and the *do*-calculus

Once the model is defined, we can use standard probabilistic inference to obtain any quantities of interest from the joint. However, in the context of causal inference, there are some common operations and shortcuts that might be of use, which we'll discuss below.

5.1 The *do*-calculus

In the aspirin example, in equations 14 - 18 we were able to obtain an expression for the conditional $q(y^*|t^*, \boldsymbol{\theta})$ in the interventional part of the model in terms of density functions that were the same between observed and interventional worlds: $q(y^*|t^*, \boldsymbol{\theta}) = \int p_{\boldsymbol{\theta}}(y^*|t^*, z^*)p_{\boldsymbol{\theta}}(z^*)dz^*$. Specifically, conditioning on Z^* allowed us to do so. This is typically referred to as *adjusting* for Z.

The mathematics were simple for this three-variable model, however, what if the model was more complex? In Bayesian Networks, are there any shortcuts to expressing the conditional of interest in the intervened-upon world in terms of conditionals of the variables from the observed world only?

The definition for the intervention operation in BNs def. 2 gives an expression for the joint in term of conditionals from the observed setting and $g^*(x_j, \mathbf{x}_{\mathbf{PA}_j^*})$, which is specified as part of the intervention (eq. 22). From the joint, we can obtain any desired conditional of interest in the interventional setting by marginalising out; say, without loss of generality, we're interested in $q(x_1, \ldots, x_m|x_{m+1} \ldots x_n)$ for a BN on d variables:

$$q(x_1, \ldots, x_m|x_{m+1} \ldots x_n) = \frac{q(x_1, \ldots, x_n)}{\int q(x_1, \ldots, x_n)d\mathbf{x}_{m+1:n}} = \frac{\int q(x_1, \ldots, x_d)d\mathbf{x}_{n+1:d}}{\int q(x_1, \ldots, x_d)d\mathbf{x}_{m+1:d}}$$

Where we can express $q(x_1, \ldots, x_d)$ in terms of densities from the observed distribution as desired. For complex models, such as neural networks, the fraction of the two integrals could be estimated with e.g. sampling, however, it is not necessarily trivial, and usually a simpler expression could be obtained.

For instance, in the case of an atomic intervention on x_t (or any intervention with an empty set of new parents), adjusting using the parents of x_t in the observed BN graph $\mathcal{G}$ is sufficient. Consider estimating $q(x_y)$ under such an intervention when $X_y \notin \mathbf{X}_{\mathbf{PA}_t^{\mathcal{G}}}$. Then, we can obtain a potentially much simpler expression:

$$q(x_y) = q(x_y|x_t) = \int p(x_y|x_t, \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}})p(\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}})d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}} \tag{24}$$

The last expression in fact matches the ones we obtained for the aspirin intervention earlier: $\int p_{\boldsymbol{\theta}}(y^*|t^*, z^*)p_{\boldsymbol{\theta}}(z^*)dz^*$. Pearl and others have proposed other "adjustment sets" of this sorts, including the *backdoor criterion* and *towards necessity* [4, Proposition 6.41].

Beyond just obtaining expressions for probability distributions of interest, another consideration comes into play when some of the variables in $\mathbf{X}$ are unobserved. This would mean that, for conditional probability functions in the observed setting $p(\mathbf{x}_A|\mathbf{x}_B)$, the ones for which a subset of variables in either $\mathbf{x}_A$ or $\mathbf{x}_B$ correspond to unobserved variables would not be easily estimable. Hence, one may ask: can we express a conditional probability function in the interventional setting using only the probability density/mass functions of the observed variables (in the observational setting) only? This would clearly be desirable, as any such probability function can be fitted via standard supervised learning; but when is that possible? And how could we obtain such an expression (potentially automatically from the graph)?

The *do*-calculus has been developed to address these questions. We've put its rules in the appendix A.5, or you can find them in [7, §3.4] or [4, §6.7]. A particular probability distribution in the interventional world is called *identifiable* if it can be expressed in terms of distribution functions of *observed* variables in the observed world only [4, p. 119] [7, Corollary 3.4.2]. The *do*-calculus allows for finding expressions for all identifiable intervention distributions with a repeated application of its rules[11]. Using the *do*-calculus, Tian and Pearl have developed an algorithm that is guaranteed to find all identifiable intervention distributions [12], and Schipster and Pearl have developed a graphical criterion on $\mathcal{G}$ for determining identifiability of an interventional distribution [13].

[11] Which you'd hope seeing as you can do this by using definitions 1 and 2 and the rules of probability theory.

The choice to include ‘calculus’ in the name of the *do*-calculus adds a ring of profoundness to it, and inspires curiosity. This has possibly instilled a somewhat warped expectation of what it actually is in some. The *do*-calculus simply builds on the rules of probability theory, the definition of a BN and the definition of interventions in BNs. Its rules are derived from these, and hence any result obtained using the *do*-calculus can be obtained from these underlying rules as well.

5.2 Identifiability from a probabilistic perspective*

Above, we briefly introduced the concept of identifiability: an interventional conditional probability function (the "causal effect") is *identifiable* whenever it can be expressed in terms of probability functions of the observed variables only [7, p. 77]. If the quantities we wish to infer are *identifiable*, this results in a major simplification when modelling; for the purpose of inferring the quantities of interest, we can leave the form of any dependencies on the unobserved variables in the model unspecified. These variables do not need to appear in the final expression for what we need to infer. This criterion conforms to a frequentist notion of identifiability wherein we would require that a quantity of interest can be uniquely determined in the infinite data limit. However, when adopting a probabilistic modelling perspective, there is no *requirement* for identifiability for learning to occur.

In particular, Bayesian learning occurs whenever the posterior distributions can differ from the prior distributions, reflecting that we have updated our beliefs based on the data [14]. From a Bayesian perspective, as long as we specify the model in full, there is nothing stopping us conceptually from computing a posterior over any quantity of interest in our model. If we specify the (possibly uncertain) form of the dependence of the observed variables on the latent ones, we can apply the rules of probability theory to update our belief about any other part of the model. We can still concern ourselves with whether observations reduce the (epistemic) uncertainty in the quantity we wish to infer, but this is a far less stringent requirement than demanding Pearl’s identifiability.

6 Interventional inference with machine learning models

The aspirin example presented in section 3 is a very simple three-variable task. We could have in principle derived analytical expressions for all the conditional distributions. In machine learning, however, we are often dealing with high-dimensional and highly non-linear data. For instance, in a medical setting we might be dealing with DNA sequences or CT images rather than something as simple as a scalar headache severity rating. How could we approach modelling interventions in these more complex settings?

We saw previously in eq. 22 that we can factorise the joint distribution in the intervened-upon world into conditional distribution functions shared with the observed world and those explicitely specified as part of the intervention. Each of the conditional distribution functions had the form $g_{\boldsymbol{\theta}}(\mathbf{x_k}, \mathbf{x}_{\mathbf{PA}_k^{\mathcal{G}}}) = p_{\boldsymbol{\theta}}(\mathbf{x_k}|\mathbf{x}_{\mathbf{PA}_k^{\mathcal{G}}})$. We can use a probabilistic model of choice to model these conditional distributions in a supervised-learning fashion using data from the observed world. For the variables without any parents[12], we could use unsupervised learning with, e.g. Variational Autoencoders (VAEs) [15] or Normalising Flows [16], to model $g_{\boldsymbol{\theta}}(\mathbf{x_k}) = p_{\boldsymbol{\theta}}(\mathbf{x_k})$. A fairly general framework for specifying such systems using Gaussian Processes has been proposed by Silva et al. in [17].

Running inference in these models can be very application-dependent. For instance, Louizos et al. [18] consider a specific architecture based on VAEs for estimating Individual Treatment Effect for a specific graphical model with a partially-observable confounder. In the more general case, however, one can imagine using e.g. ancestral sampling to estimate any integrals of interest — first sample the variables that don’t have any parents from $p_{\boldsymbol{\theta}}(\mathbf{x_k})$, then their children from $p_{\boldsymbol{\theta}}(\mathbf{x_k}|\mathbf{x}_{\mathbf{PA}_k^{\mathcal{G}}})$ conditioned on the sampled values, their children’s children and so on.

Hence, it should be clear that you *can* do causal inference using machine learning and deep learning methods. In a later section, we’ll see a specific example of using deep learning in the more complex case of a *counterfactual question*, but in essence the same principles apply.

[12] Also called *exogenous*.

7 Counterfactuals

In the previous section, we described modelling interventions. With those, we sought to infer the future effects of a contemplated manipulation in a system. We could, for instance, answer questions like: “What’s the effectiveness of aspirin on reducing headache duration?”, “How would changing regulations on cigarette companies impact population’s life-expectancy?”, or “How effective is a lock-down at suppressing the spread of a pandemic?”

A different type of questions we might also want to ask is about what might have happened in previously observed cases had something been done differently. What would have happened in an alternative world, were some values and/or mechanisms in our model altered? Some examples of such *counterfactual* questions would be: “Would I still have a headache now had I taken a larger dose of aspirin?”, or “How many disease cases would there be in a population had a lock-down not been put in place?”[13]

Counterfactual analysis can be used when trying to assign blame for an outcome, or when assessing the cause of an event [20]. This bears significance in, for instance, legal fields, where the court may seek to assign responsibility for something that has happened [21]. For example, a judge might wish to conclude that “were it not for the defendant’s actions, the chemical spill would in all likelihood not have occurred”. With probabilistic modelling, we can endeavour to quantify how likely or unlikely different outcomes would have been had different actions been taken.

How could we approach modelling counterfactuals? Intuitively, when talking about things that might have happened had some factor been different, we might consider a hypothetical world that is the same as the observed one, with the exception of the alteration of interest. A common expression in economics for this line of thinking is *ceteris paribus* — ‘all other things being equal’. Stating *ceteris paribus* as an assumption, however, is fairly ambiguous in itself. We might prefer to specify what exactly is being held equal and what isn’t [7, §7.2.2]. One way to do so with probabilistic modelling is to share the desired variables between two settings — the observed and the counterfactual — making modelling assumptions explicit.

We’ll start again by giving a concrete example of counterfactual-like analysis, discuss the mathematical machinery behind it, and attempt to get to the bottom of some commonly raised discussion points such as “can you do counterfactual inference with Bayesian Networks?”.

***Example 3:* Aspirin Model Counterfactual**

INFERENCE PROBLEM

A friend tells you that she had a headache before her exam. She says that she took a given dose of aspirin t, and the headache lasted for y hours, disrupting her exam performance. She doesn’t recall the initial headache severity. You have access to the survey dataset $\mathcal{D} = \{(z_i, t_i, y_i)\}_{i=1}^N$ from the previous section, and you are now wondering: had your friend taken a larger dose of aspirin, would her headache had gone away before the exam?

MODEL AND ASSUMPTIONS

Let’s begin again by setting up a model over the two settings of interest — the observed and the counterfactual. In addition to the random variables Z, T, Y in the observed setting, where Z is now latent, we also define variables T^* and Y^*, which correspond to the counterfactual dose choice and headache duration respectively. We assume that Z — the headache severity — is the same in the counterfactual and the observed world. In addition, we define the variables for the dataset of previous fully-observed cases $\{(Z_i, T_i, Y_i)_{i=1}^N\}$, which we can use to infer the parameters of our model.

Again, we can specify a graphical model to define our assumptions as shown in figure 8.

[13] This latter counterfactual question and the question about the effectiveness of non-pharmaceutical interventions on pandemic spread have been addressed in the context of COVID-19 by Flaxman et al. [19].

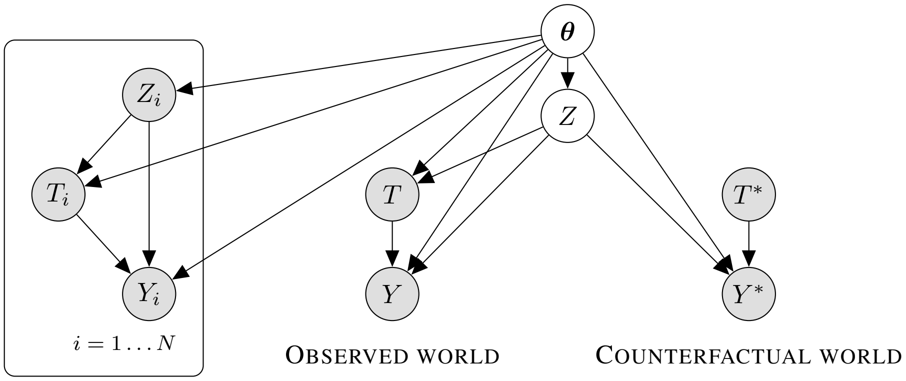


Figure 8: Graphical model for counterfactual inference in the aspirin example. The nodes Z, T, Y correspond to the observed case of interest. The headache severity Z is unobserved, and is assumed to be shared with the counterfactual setting in which we've intervened on the treatment T^*. The remaining fully observed examples $\{(Z_i, T_i, Y_i)\}_{i=1}^N$ can be used to estimate the model parameters $\boldsymbol{\theta}$.

Based on the graphical model, we can again write down a full joint distribution over the settings of interest. We'll use $q(\cdot)$ to refer to conditionals in the counterfactual world, e.g. $p_{\boldsymbol{\theta}}(Y^* {=} y | T^* {=} t, Z {=} z) {=} q_{\boldsymbol{\theta}}(y|t,z)$ and $p(T^* {=} t) {=} q(t)$.

$$\begin{aligned}
&p(z, t, y, z, t^*, y^*, \boldsymbol{\theta}, \mathcal{D}) = \\
&\quad = p(z, t, y, t^*, y^* | \boldsymbol{\theta}, \cancel{\mathcal{D}}) p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) && (25)\\
&\quad = \Big(p(z|\boldsymbol{\theta}) p(t|z, \boldsymbol{\theta}) p(y|t, z, \boldsymbol{\theta}) p(t^* | \cancel{y}, \cancel{t}, \cancel{z}, \cancel{\boldsymbol{\theta}}) p(y^* | t^*, \cancel{y}, \cancel{t}, z, \boldsymbol{\theta}) \Big) p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) && (26)\\
&\quad = \left(p_{\boldsymbol{\theta}}(z) p_{\boldsymbol{\theta}}(t|z) p_{\boldsymbol{\theta}}(y|t, z) q(t^*) q_{\boldsymbol{\theta}}(y^*|t^*, z) \right) \left(\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(z_i) p_{\boldsymbol{\theta}}(t_i|z_i) p_{\boldsymbol{\theta}}(y_i|z_i, t_i) \right) p(\boldsymbol{\theta}) && (27)\\
&\quad = \left(\underbrace{p_{\boldsymbol{\theta}}(t|z) p_{\boldsymbol{\theta}}(y|t, z)}_{\text{Observed world}} \underbrace{p_{\boldsymbol{\theta}}(z)}_{\substack{\text{Shared} \\ \text{in both}}} \underbrace{q(t^*) p_{\boldsymbol{\theta}}(y^*|t^*, z)}_{\text{Counterfactual world}} \right) \underbrace{\left(\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(z_i) p_{\boldsymbol{\theta}}(t_i|z_i) p_{\boldsymbol{\theta}}(y_i|z_i, t_i) \right)}_{\text{Dataset likelihood}} p(\boldsymbol{\theta}) && (28)
\end{aligned}$$

Let's again unpack the assumptions made. We used the conditional independence assumptions embodied in the graphical model to cancel terms in lines 25 and 26. In line 28, we make the additional assumption that $q_{\boldsymbol{\theta}}(y|t,z) = p_{\boldsymbol{\theta}}(y|t,z)$; in other words, the distribution of the counterfactual headache duration Y^* conditioned on Z and T^* is distributed in the same way as the observed world counterpart Y conditioned on Z and T.

We might also want to justify why we've removed the edge from Z to T^* in the counterfactual setting. This is because we're changing the dose-determining mechanism; in this hypothetical world, our friend no longer decides what dose to take based on the headache severity — we've intervened to decide for her. Note that if we didn't remove the $Z \to T^*$ edge, the choice of a hypothetical dose T^* in the counterfactual world would influence our belief over the headache severity Z in the observed world, which would be somewhat paradoxical. We could further specify the exact form of the distribution on T^*, which, as we believe it corresponds to an intervention, can be set to:

$$q(t) = \delta(t^* - t) \tag{29}$$

Again, using the properties of log-normal distributions, it can be seen that, conditioned on $\boldsymbol{\theta}$, variables $[Z, T, Y, Y^*]^\top$ are jointly distributed according to:

$$\log \mathcal{N} \left(\begin{bmatrix} \mu_Z \\ a\mu_Z \\ (b-ac)\mu_Z \\ b\mu_Z - c \log t^* \end{bmatrix}, \begin{bmatrix} \sigma_Z^2 & a\sigma_Z^2 & (b-ac)\sigma_Z^2 & b\sigma_Z^2 \\ a\sigma_Z^2 & a^2\sigma_Z^2 {+} \sigma_T^2 & a(b-ac)\sigma_Z^2 {-} c\sigma_T^2 & ab\sigma_Z^2 \\ (b-ac)\sigma_Z^2 & a(b-ac)\sigma_Z^2 {-} c\sigma_T^2 & (b-ac)^2\sigma_Z^2 {+} c^2\sigma_T^2 {+} \sigma_Y^2 & b(b-ac)\sigma_Z^2 \\ b\sigma_Z^2 & ab\sigma_Z^2 & b(b-ac)\sigma_Z^2 & b^2\sigma_Z^2 + \sigma_Y^2 \end{bmatrix} \right) \tag{30}$$

INFERENCE

Returning to the inference question, we wanted to infer the likely headache duration in the counterfactual world had we given our friend t^* milligrams of aspirin. More concretely, we are interested in the conditional probability distribution $p(y^*|t, y, \mathcal{D})$ in our model — the distribution over the hypothetical headache duration in response to aspirin dose t^*, given that we've observed that the actual headache lasted for y minutes after our friend took a dose t, and given the dataset $\mathcal{D}$:

$$
\begin{aligned}
p(y^*|t, y, \mathcal{D}) &= \int p(y^*|t, y, \boldsymbol{\theta}, \cancel{\mathcal{D}}) p(\boldsymbol{\theta}|t, y, \mathcal{D}) d\boldsymbol{\theta} && (31)\\
&= \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D},t,y)} \left[p_{\boldsymbol{\theta}}(y^*|t, y)\right] && (32)\\
&= \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D},t,y)} \left[\int p_{\boldsymbol{\theta}}(y^*|t^*, z, \cancel{t}, \cancel{y}) p(z|t, y, \boldsymbol{\theta}) dz\right] && (33)\\
&= \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D},t,y)} \left[\int p_{\boldsymbol{\theta}}(y^*|t^*, z) p_{\boldsymbol{\theta}}(z|t, y) dz\right] && (34)
\end{aligned}
$$

The expression in eq. 34 can be interpreted as performing inference in the counterfactual part of the model by marginalising over the posterior on Z conditioned on the observed data. We'll discuss this observation shortly.

Alternatively, we can simply obtain the joint on $[T, Y, Y^*]^\top$ from equation 30 by marginalising out Z. This would yield:

$$
\begin{bmatrix} T \\ Y \\ Y^* \end{bmatrix} \sim \log\mathcal{N}\left(\begin{bmatrix} a\mu_Z \\ d\mu_Z \\ b\mu_Z - c\log t^* \end{bmatrix}, \begin{bmatrix} a^2\sigma_Z^2{+}\sigma_T^2 & ad\sigma_Z^2{-}c\sigma_T^2 & ab\sigma_Z^2 \\ ad\sigma_Z^2{-}c\sigma_T^2 & d^2\sigma_Z^2{+}c^2\sigma_T^2{+}\sigma_Y^2 & bd\sigma_Z^2 \\ ab\sigma_Z^2 & bd\sigma_Z^2 & b^2\sigma_Z^2 + \sigma_Y^2 \end{bmatrix}\right) \tag{35}
$$

It is the straightforward to obtain the conditional $p_{\boldsymbol{\theta}}(y^*|t, y)$ from a log-normal joint over $[Z, T, Y^*]^\top$, which we could then put into eq. 32. Using a MAP estimate of the parameters, we could obtain an analytical (albeit lengthy) expression for the probability distribution $p(y^*|t, y, \mathcal{D})$ that we were seeking to infer.

In the above example, we demonstrated again that by defining a probabilistic model — a Bayesian Network in this case — over all the settings of interest, we can answer questions of *counterfactual* nature. To emphasise this point once again: no further causal-specific framework or notation was required.

7.1 An alternative perspective

There is an alternative perspective on the procedure we outlined above. Namely, just as we could equivalently frame our twin model procedure in the interventional setting as inference in two separate models (the observed and the interventional), an analogue perspective holds for counterfactuals. Recall from the graphical model that the counterfactual variables T^*, Y^* are independent of the variables in the observed setting conditioned on Z (and model parameters $\boldsymbol{\theta}$). We could define a model over a single set of counterfactual variables — Z^*, T^*, Y^* — and set the prior on Z^* in that model to equal the posterior on Z in the observed world:

$$
p_{\boldsymbol{\theta}}(Z^* = z) = p_{\boldsymbol{\theta}}(z|t, y)
$$

If we then run inference in that model, say conditioned on $T^* = t^*$, we would obtain:

$$
\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D},t,y)} \left[p_{\boldsymbol{\theta}}(y^*|t^*)\right] = \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D},t,y)} \left[\int p_{\boldsymbol{\theta}}(y^*|t^*, z^*) \underbrace{p_{\boldsymbol{\theta}}(Z^* = z)}_{=p_{\boldsymbol{\theta}}(z|t,y)} dz\right] \tag{36}
$$

This is the same expression as the one we obtained using the joint model in eq. 34. Hence, we can completely equivalently view the joint model approach we have outlined for the aspirin example as the posterior in the observed world becoming the prior for a counterfactual world model.

7.2 Which latent variables should we share?

In the example above, we considered a counterfactual world in which our friend had the same headache severity as in the observed world. However, you may recall that in our original formulation of the model in eq. 1 we specified $Y = \frac{Z^b}{T^c}\varepsilon_Y$. We used this equation as a short-hand to define that Y conditioned on T, Z is log-normal distributed: $Y|Z,T \sim \log\mathcal{N}(\frac{Z^b}{T^c}, \sigma_Y^2)$. So far in this note, we've only made use of this conditional distribution, and in no way relied or considered the dependence on ε_Y. Should we have shared it between the counterfactual and the observed world? And what would it mean to share it?

If we assume that ε_Y captures some latent person-specific health attributes (weight, age, BMI, etc.) then it would certainly make sense to share it in our example. After all, our intent was to predict the counterfactual headache duration for *the same person* as accurately as possible. Alternatively, the variable could potentially represent the effect of environmental factors: the atmospheric pressure, which could deteriorate or ameliorate the headache, or the randomness in the manufacturing process of aspirin, which makes the actual amount of the active substance in aspirin vary from batch to batch. If said friend had a 20% higher active substance content in her tablet by chance, would we believe this to be the case if we intervened to give her a higher-dose tablet too?

This gets even more ambiguous once we consider noise variables that have been added to account for modelling error, rather than to represent actual stochasticity in the data. This is a common practice in machine learning; even if the problem is known or assumed to be deterministic, we would often specify a small observational noise to: **1)** deal with models of finite capacity, or **2)** make the likelihood continuous with respect to model parameters to allow for gradient-based optimisation. If that is the case, sharing the noise variable between the observed and counterfactual worlds is hard to link to an intuitive interpretation.[14]

Although the meaning of holding ε_Y fixed across the two worlds might be enigmatic, we can consider what happens when we share it. For the particular parameterisation on ε_Y that we have chosen, $Y = \frac{Z^b}{T^c}\varepsilon_Y$, Y is a deterministic function of Z, T and ε_Y (conditioned on $\boldsymbol{\theta}$). Hence, by sharing both Z and ε_Y we could uniquely identify the counterfactual value of Y^* for any given T^*. To see this, we can write (assuming Y^* has the same functional dependence on ε_Y, Z, T^*):

$$Y^* = \frac{Z^b}{(T^*)^c}\varepsilon_Y = \left(\frac{Z^b}{T^c}\varepsilon_Y\right)\frac{T^c}{(T^*)^c} = Y\frac{T^c}{(T^*)^c} \tag{37}$$

which, assuming point-estimate of model parameters $\boldsymbol{\theta}$, uniquely identifies the value of Y^* for any given set of Y, T, T^*.

It is worth noting that only when ε_Y is shared across the two worlds would the inference outcome be considered a "proper" counterfactual in Pearl's framework [1][7]. Per his definition, which we'll introduce later in this section, only when *all* sources of randomness have been shared does the result deserve the title of a counterfactual[15]. This does not detract from the conclusion that you *can* do counterfactual inference with Bayesian Networks; we could have just as easily represented sharing all the sources of randomness by specifying a Bayesian Network over all the settings of interest.

7.2.1 Multiple latent representations can yield different counterfactuals

There is another complication with trying to share an unobserved variable between the counterfactual and the observed settings. Let's say that we hypothesise the existence of a latent variable that represents some person or situation-specific properties relating to how the headache duration will be affected by a given dose of aspirin. There are potentially infinitely many parameterisations on such a latent that will give the same observed distribution, but different counterfactual outcomes.

[14] It is worth noting, however, that for the additive noise case, sharing the noise variable between the observed and counterfactual worlds can reduce bias of our estimator under some conditions (e.g. when marginal distribution on T^* equals that on T, and T is exogenous).

[15] Although, Pearl admits that even when some latent variables aren't shared, the inference outcome can still be considered a counterfactual if these latent variables represent intrinsic randomness - one that is never observable, such as that due to quantum effects [7, §7.2.2]. Nonetheless, in the probabilistic modelling approach above, we clearly specified what variables are shared between the observed and the counterfactual setting. The inference outcome has a clear interpretation in light of the assumptions used, and as such whether it is labelled a counterfactual or not can be considered primarily a terminological issue.

For a concrete example, consider what would happen if the headache duration's (Y) dependence on ε_Y had instead been defined as:

$$Y = \frac{Z^b}{T^c}(\varepsilon_Y)^{\operatorname{sign}(\log T)} \tag{38}$$

The marginal distribution on (Z, T, Y) would be left completely unchanged by this revision; we've left the distributions on Z and T unchanged, and the conditional distribution on Y given Z, T is the same, because both $(\varepsilon_Y)^1$ and $(\varepsilon_Y)^{-1}$ are distributed as $\log \mathcal{N}(0, 1)$[16].

Now, consider what would then happen if we shared ε_Y with the counterfactual world (let's for convenience assume that Z is observed this time). From eq. 38, we could infer the value of the latent: $\varepsilon_Y = \left(\frac{YT^c}{Z^b}\right)^{\operatorname{sign}(\log T)}$. The equation for the counterfactual headache duration would then take the form:

$$Y^* = \frac{Z^b}{(T^*)^c}\left(\varepsilon_Y\right)^{\operatorname{sign}(\log T^*)} = \begin{cases} \frac{Z^b}{(T^*)^c}\varepsilon_Y^1 \\ \frac{Z^b}{(T^*)^c}\varepsilon_Y^{-1} \end{cases} = \begin{cases} \frac{Z^b}{(T^*)^c}\left(\frac{YT^c}{Z^b}\right)^{\operatorname{sign}(\log T)} & \text{if } \log T^* \geq 0 \\ \frac{Z^b}{(T^*)^c}\left(\frac{YT^c}{Z^b}\right)^{-\operatorname{sign}(\log T)} & \text{if } \log T^* < 0 \end{cases}$$

This is compared to the case for the standard formulation ($Y = \frac{Z^b}{T^c}\varepsilon_Y$), where the counterfactual outcome takes the form $Y\frac{T^c}{(T^*)^c}$ as shown in eq. 37. Hence, clearly, the inferred counterfactual quantity would be different for these two formulations.

Without additional assumptions, there are infinitely many ways in which the observed variables can depend on the latent variables that could give different counterfactual outcomes. Hence, returning to the recurring theme, we need to specify these assumptions to be able to do counterfactual inference.

7.3 Structural Causal Models

So far in this section, we have shown that you can use Bayesian Networks, without introducing any additional notation or distinctive causal concepts, to answer counterfactual questions. Nevertheless, there appears to be, somewhat surprisingly, a debate as to whether Bayesian Networks are sufficient. Many claim that you need *Structural Causal Models* to predict counterfactuals. Below, we'll introduce Structural Causal Models, relate them to our approach above, and lay to rest concerns about using Bayesian Networks and probabilistic modelling for causal inference.

Informally, a Structural Causal Model is a Bayesian Network on random variables $\mathbf{X}$ in which each variable X_k is a deterministic function of its parents and a latent noise variable N_j. Formally, we can define SCMs as follows [4, §6.2]:

Definition 3. ***Structural Causal Model (SCM).*** *A structural causal model $\mathfrak{C}$ on a set of random variables $\mathbf{X}$ consists of a tuple $\langle \mathcal{S}, \mathcal{G}, p_{\mathbf{N}} \rangle$ where **1)** $\mathcal{G}$ is a directed acyclic graph on $\mathbf{X}$, **2)** $\mathcal{S}$ is a collection of d structural assignments:*

$$X_j := f_j\left(\mathbf{X}_{\mathbf{PA}_j^{\mathcal{G}}}, N_j\right), \quad j = 1, \ldots, d$$

*in which $\mathbf{X}_{\mathbf{PA}_j^{\mathcal{G}}} \subseteq \{X_1, \ldots, X_d\} \setminus \{X_j\}$ are the parents of X_j in graph $\mathcal{G}$ and N_j are the latent noise variables, and **3)** $p_{\mathbf{N}}$ is a probability distribution over the noise variables $\{N_1, \ldots, N_d\}$ which are assumed to be jointly independent, i.e. $p_{\mathbf{N}}$ factorises as $p_{\mathbf{N}}(\mathbf{n}) = \prod_{j=1}^{d} p_{N_j}(n_j)$.*

A Structural Causal Model implies a Bayesian Network (or a Causal Graphical Model if you will) on $\mathbf{X}$ with the same graph $\mathcal{G}$. This is because each functional assignment $X_j := f_j\left(\mathbf{X}_{\mathbf{PA}_j^{\mathcal{G}}}, N_j\right)$ and the corresponding distribution function over the noise variable N_j yield a conditional distribution on X_j given its parents:

$$p(x_j | \mathbf{x}_{\mathbf{PA}_j^{\mathcal{G}}}) = \int \delta\left(x_j - f_j(\mathbf{x}_{\mathbf{PA}_j^{\mathcal{G}}}, n_j)\right) p(n_j) dn_j$$

Since the functional assignments entail the same set of conditional independencies as a Bayesian Network with a graph $\mathcal{G}$, the SCM also yields the same factorisation of the joint $p(\mathbf{x}) = \prod_j p(x_j | \mathbf{x}_{\mathbf{PA}_j^{\mathcal{G}}})$. In that sense, an SCM can be viewed as a Bayesian Network on $\mathbf{X}$ with additional assumptions.

[16] To see this, we can write $(\varepsilon_Y)^{-1} = \exp\left(-1 \log \varepsilon_Y\right)$ and recall that $\log \varepsilon_Y \sim \mathcal{N}(0, 1)$. A standard normal distributed variable multiplied by -1 is, however, also standard normal distributed.

On the other hand, a Structural Causal Model is just a special case of a Bayesian Network defined on both variables $\mathbf{N}$ and $\mathbf{X}$. There are additional constraints on the graph and the form of conditional distributions which the BN on $\mathbf{N}$ and $\mathbf{X}$ has to satisfy to be equivalent to an SCM, however, a Structural Causal Model *can* be represented with a Bayesian Network.

Let's consider how a counterfactual would be defined in an SCM:

Definition 4. ***Counterfactuals in Structural Causal Models****. Given an SCM $\mathfrak{C}{=}\langle\mathcal{S},\mathcal{G},p_{\mathbf{N}}\rangle$ and an observation of the variables $\mathbf{X}{=}\mathbf{x}$, we can define a counterfactual SCM $\mathfrak{C}_{CF}$ by replacing the distribution of the noise variables:*

$$\mathfrak{C}_{CF} = \langle\mathcal{S},\mathcal{G},p_{\mathbf{N}^*}\rangle$$

The altered distribution on the noise variables is given by: $p_{\mathbf{N}^*}(\mathbf{n}){=}p_{\mathbf{N}}(\mathbf{n}|\mathbf{X}{=}\mathbf{x})$. In other words, the posterior over the noise variables given $\mathbf{X}{=}\mathbf{x}$ in the original SCM $\mathfrak{C}$ becomes the prior for the new counterfactual SCM $\mathfrak{C}_{CF}$. The noise variables in this new counterfactual SCM need not be independent anymore.

Counterfactual statements can then be computed by intervening (in the sense of the rule given in def. 2) on variables in the counterfactual SCM $\mathfrak{C}_{CF}$.[17]

This definition results in an inference procedure that is no different than what we've been doing in a joint model over both settings so far. Following this definition, we would first compute the posterior over the unobserved variables $\mathbf{n}$ given the observed data, and then run inference using that posterior as a prior in a second *counterfactual* model (which might have been altered to represent that we've intervened on the generative process). However, we've already shown that this is equivalent to running inference in a joint model where the noise variables are shared between the two settings. Pearl would likely agree with that point, seeing as he himself proposed such a 'twin network' model as a way to do probabilistic inference in SCMs [11][7, §7.1.4]. Hence, computing counterfactuals using the SCM framework *is* just a special case of the probabilistic modelling approach we have outlined at the start of this section.

As such, Structural Causal Models are sufficient for causal inference, but not necessary. There are many equivalent formulations that can be used to tackle interventional and counterfactual problems, and an SCM is just one of them.

8 Counterfactual inference with machine learning models

So far, we've seen how probabilistic modelling principles can be used to answer counterfactual and interventional questions, and discussed how machine learning methods can be used in the context of the latter (section 6). Here, we'll show that in the counterfactual case, in essence the same principles apply. Once we factorise the joint distribution in the model over all the settings into conditional distribution functions shared with the observed world and those explicitly specified as part of the counterfactual, we can use a probabilistic model of choice — such as a neural network — to model these.

To explicitly illustrate how this could be done, below we showcase a more complex, highly non-linear counterfactual inference problem, and demonstrate a machine learning approach to solving it. We chose a problem that is grounded on simulated data, which allows us to compare our counterfactual estimates to ground truth.

Example 4: **Predator-prey counterfactuals**

INFERENCE PROBLEM

Imagine that fishermen in some region of the world decided to cull their local seal population, driven by anecdotal fears about the danger to the fish population on which the fishermen rely. Some months later, that seal becomes an endangered species. There has since been a court case looking into the culling practices and the jury is wondering if the fishermen are responsible for the species' endangerment. They want to know: in retrospective, what would the seal population have been had it not been culled earlier?

To concretely formulate the counterfactual problem: we have observations of the immediate pre-cull, post-cull and current seal populations. These are denoted s_1^*, s_1 and s_2 respectively. We do not, however, have observations of either the initial, or the final fish populations in your region, as there are plenty of fish in the sea.

[17] In that regard, the SCM also implicitly entails analogous assumptions to a Causal Graphical Model on $\mathbf{X}$.

The initial fish population is a latent, unobserved variable. From these observations, and the available data on fish and seal population elsewhere, we want to answer as accurately as possible what the seal population would have been now, had it been kept at the pre-cull level earlier. The task is illustrated in figure 9.

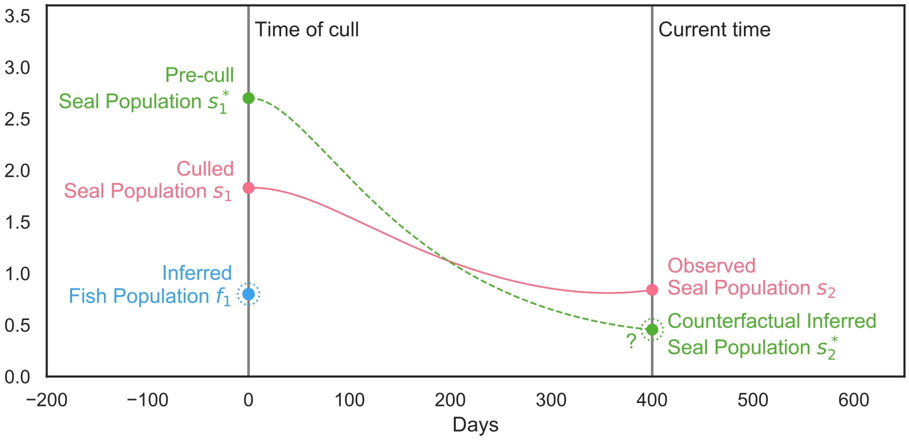


Figure 9: Illustration of the causal inference problem.

Data

The data in this example is based on a dynamics model of predator and prey populations. Specifically, the populations' evolution over time is simulated with the Lotka-Volterra differential equations (also known as the predator-prey equations):

$$\frac{df}{dt} = \gamma_1 f - \gamma_2 f s \qquad\qquad \frac{ds}{dt} = \gamma_3 f s - \gamma_4 s \tag{39}$$

Here, s represents the seal (predator) population and f represents the fish (prey) population. $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \mathbb{R}_{>0}$ are the parameters of the equations that determine the population dynamics. Examples of population oscillations generated from these equations are shown in figure 10.

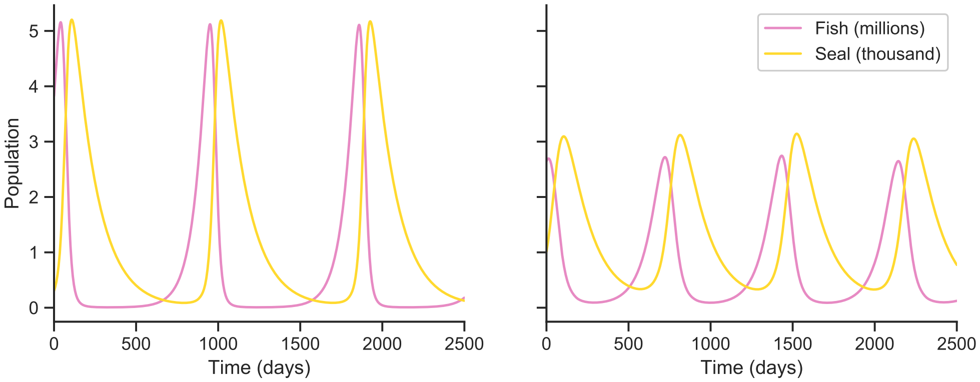


Figure 10: Example population oscillations generated by the Lotka-Volterra equations.

In a practical scenario, a practitioner would not have access to the simulator itself, but rather only to some data generated from it. As such such, they would have to resolve to using statistical techniques to estimate the population dynamics. For this example we assume that the fish and seal population dynamics are dictated by the pair of equations 39 with a fixed set of parameters $\gamma_1 = 0.015$, $\gamma_2 = 0.012$, $\gamma_3 = 0.007$, $\gamma_4 = 0.009$. The dataset available consists of tuples $\mathcal{D} = \{(f_{1,i}, s_{1,i}, \tau_i, f_{2,i}, s_{2,i})\}_{i=1}^{N}$ where $f_{1,i}, s_{1,i}$ are the initial fish and seal populations, and $f_{2,i}, s_{2,i}$ are the populations after time τ_i. A model can then be fitted to this data to predict the seal and fish population after some time τ given the initial fish and seal populations: $p_\theta(s_2, f_2|s_1, f_1, \tau)$. For simplicity, we'll assume a fixed $\tau = 400$.

Model and assumptions

We can again represent our assumptions by setting up a joint model over the variables in the observed world S_1, S_2 and F_1 (where F_1 is unobserved), and the corresponding variables in the counterfactual world S_1^* and S_2^*. We assume that the initial fish population F_1 is the same in both the counterfactual and the observed setting. We can then specify further assumptions about the problem with a graphical model over all the settings of interest, as shown in figure 11.

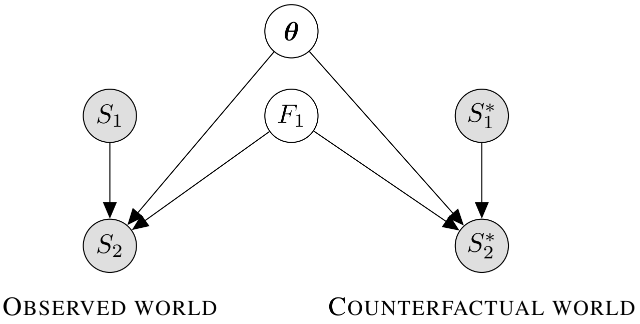


Figure 11: Graphical model corresponding to the inference problem. The dependence of the fully observed dataset $\mathcal{D}$ on $\boldsymbol{\theta}$ has been omitted for conciseness.

From there on, the joint distribution over all the variables follows:

$$\begin{aligned}
&p(s_1, s_2, f_1, s_1^*, s_2^*, \mathcal{D}) = p(s_1, s_2, f_1, s_1^*, s_2^*|\boldsymbol{\theta}, \cancel{\mathcal{D}})p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \Big(p(f_1|\cancel{\boldsymbol{\theta}})p(s_1|\cancel{f_1}, \boldsymbol{\theta})p(s_2|s_1, f_1, \boldsymbol{\theta})p(s_1^*|\cancel{s_2}, \cancel{s_1}, \cancel{f_1}, \cancel{\boldsymbol{\theta}})p(s_2^*|s_1^*, \cancel{s_2}, \cancel{s_1}, f_1, \boldsymbol{\theta})\Big)p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \Big(\underbrace{p(s_1)p_{\boldsymbol{\theta}}(s_2|s_1, f_1)}_{\text{Observed world}}\underbrace{p(f_1)}_{\substack{\text{Shared}\\\text{in both}}}\underbrace{q(s_1^*)q_{\boldsymbol{\theta}}(s_2^*|s_1^*, f_1)}_{\text{Counterfactual world}}\Big)p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \qquad (40)
\end{aligned}$$

In the derivation, we again exploited the conditional independencies in the graphical model. We can further assume that the mechanism determining the final seal population with the observed world is the same as that in the counterfactual, i.e. $q_{\boldsymbol{\theta}}(s_2|s_1, f_1) = p_{\boldsymbol{\theta}}(s_2|s_1, f_1)$.

We further assume that we know $S_1 \sim \log\mathcal{N}(0, 1)$ and $F_1 \sim \log\mathcal{N}(0, 1)$.

Hence, the only distribution we need to learn the parameters for is $p_{\boldsymbol{\theta}}(s_2|s_1, f_1)$. As the distribution on S_2 conditioned on S_1, S_2 is highly nonlinear, we can choose to represent it with a deep neural network. Specifically, we will use a neural network that parameterises a Gaussian distribution: $p_{\boldsymbol{\theta}}(s_2|s_1, f_1) = \mathcal{N}(s_2; f_\theta(s_1, f_1), \sigma^2)$ where the mean $f_\theta(s_1, f_1)$ is modelled with a neural network, and the standard deviation $\sigma = 10^{-2}$ is fixed.

We could set the the distribution $q(s_1^*)$ to a delta function as we did before; however, recall that setting S_1^* to be delta-distributed and marginalising over it is equivalent to conditioning on $S_1^* = s_1^*$, as we saw in eq. 11. Hence, we'll leave the distribution on S_1^* unspecified and just condition on $S_1^* = s_1^*$ instead.

Inference

Our inference objective was to compute the probability distribution over the current counterfactual seal population S_2^*, given that we observed the true initial seal population s_1, the true final seal population s_2, and in the counterfactual world the initial seal population is set to the pre-cull level $S_1^* = s_1$. To again mathematise the problem: we want to compute the probability distribution $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$. This particular graphical model structure allows for the decomposition:

$$\begin{aligned}p(s_2^*|s_1, s_2, s_1^*, \mathcal{D}) &= \iint p(s_2^*, f_1, \boldsymbol{\theta}|s_1, s_2, s_1^*, \mathcal{D}) dz d\boldsymbol{\theta} \\ &= \iint p(s_2^*|s_1, s_2, \cancel{s_1^*}, \boldsymbol{\theta}, \cancel{\mathcal{D}}) p(f_1|\boldsymbol{\theta}, s_1, s_2, \cancel{s_1^*}, \boldsymbol{\theta}, \cancel{\mathcal{D}}) p(\boldsymbol{\theta}|s_1, s_2, \cancel{s_1^*}, \mathcal{D}) dz d\boldsymbol{\theta} \\ &= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}, s_1, s_2)} \Big[\int p_{\boldsymbol{\theta}}(s_2^*|s_1^*, f_1) p(f_1|s_1, s_2, \boldsymbol{\theta}) df_1 \Big] \end{aligned} \tag{41}$$

where:

$$p(f_1|s_1, s_2, \boldsymbol{\theta}) \propto p_{\boldsymbol{\theta}}(s_2|s_1, f_1) p(f_1) \tag{42}$$

We can put this back into eq. 41 to obtain the following expression for the quantity of interest:

$$p(s_2^*|s_1, s_2, s_1^*, \mathcal{D}) \propto \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}, s_1, s_2)} \Big[\mathbb{E}_{p(f_1)} [p_{\boldsymbol{\theta}}(s_2^*|s_1^*, f_1) p_{\boldsymbol{\theta}}(s_2|s_1, f_1)] \Big] \tag{43}$$

The above expression gives us a way to estimate $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$. Firstly, the outer expectation over $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathcal{D}, s_1, s_2)$ can be approximated with a point estimate $\boldsymbol{\theta}_{ML}$ by likelihood maximisation on the dataset $\mathcal{D}$. In principle, the posterior takes the form $p(\boldsymbol{\theta}|\mathcal{D}, s_1, s_2)$, but we can assume that the influence of the additional sample (s_1, s_2) is negligible, and take the maximum likelihood estimate w.r.t. just $\mathcal{D}$ instead. The inner expectation can be estimated by standard Monte-Carlo sampling. Thanks to s_2^* being one-dimensional, this simple sampling procedure will suffice.

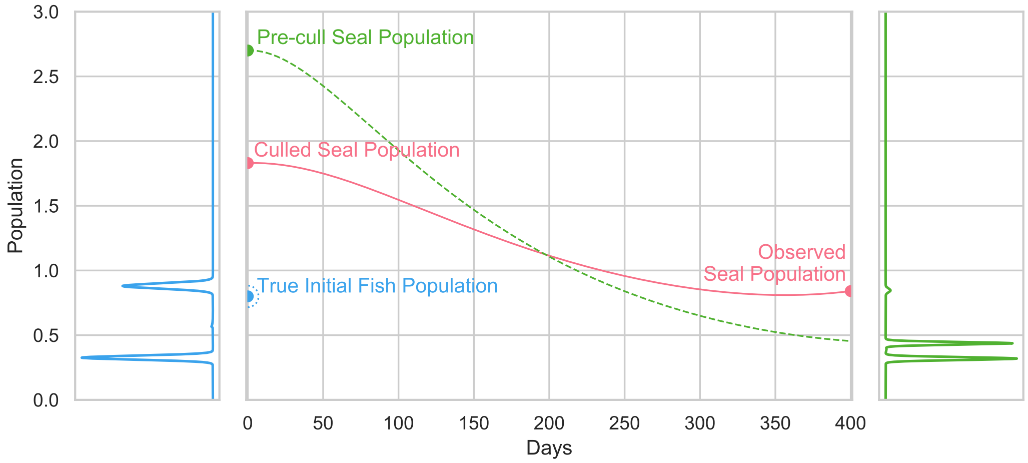


Figure 12: **Left:** the inferred distribution over the initial fish population. **Right:** the counterfactual seal population had the initial fish population been kept at a pre-cull level. **Centre:** the actual simulated seal population dynamics for both the actual and the counterfactual initial conditions.

The outcome of this inference procedure for particular values of f_1, s_1, s_2, s_1^* is shown in figure 12. As can be seen, the posterior distribution over F_1 is bimodal, indicating that two different initial fish populations could have explained the observed example well. Consequently, the posterior over the counterfactual seal population S_2^* is also bimodal. In either case, we can conclude from our model that had the fishermen not culled the seal population, in all likelihood the seal population would have been lower at the present date.

8.1 Counterfactual inference with deep latent variable models

In both the counterfactual examples considered so far, we conveniently had access to a dataset of fully observed cases. In the aspirin example, we assumed that we had a dataset $\mathcal{D} = \{z_i, t_i, y_i\}_{i=1}^N$ where Z was observed. Consequently, we were able to obtain an estimate for $p_{\boldsymbol{\theta}}(y|z, t)$. Similarly, in the predator-prey example we had a fully observed dataset where F_1 was observed, so that we could fit a neural network to model $p_{\boldsymbol{\theta}}(s_2|s_1, f_1)$. However, what if we wanted to do counterfactual inference when the variable we want to share is never observed? Here, we'll consider a possible approach using tools from deep learning.

One viable approach to estimate the dependence of a variable on a latent would be to use a conditional VAE. These have originally been introduced in the context of semi-supervised learning [22]. Here, we'll give an illustrative example to demonstrate how a similar procedure could be applied in the context of counterfactual inference.

Example 5: Predator-prey counterfactuals with a latent variable model

Continuing on from the predator-prey example, we assume that we are facing the same counterfactual problem as before. This time, however, we only have access to a dataset of seal populations at times 0 and τ: $\mathcal{D} = \{(s_{1i}, s_{2i}, \tau_i)\}_{i=1}^N$; we do not get to observe the corresponding initial fish populations f_1.

MODEL

The graphical model for this problem is nearly equivalent, with the only difference being that F_{1i} variables corresponding to the initial fish populations in the dataset are now latent. We'll further assume that we know F_1 has a log-normal prior.

We can specify our assumptions on the conditional distribution on S_2 given S_1 and F_1 by parameterising it with a neural network $p_{\boldsymbol{\theta}}(s_2|s_1, f_1) = \mathcal{N}(s_2; f_{\boldsymbol{\theta}}(s_1, f_1), \sigma^2)$ exactly as we did before. In that sense, our model specification is equivalent to that for the case when F_1 was observed in the dataset. However, as we never get to observe F_1, we no longer can fit the model parameters with a standard supervised learning approach.

ESTIMATING MODEL PARAMETERS

To infer the posterior over the parameters, we can resort to the maximum-likelihood approach: $\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$. To turn this into an optimisation objective, we can make use of the variational lower-bound [15]:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] - KL\left[q_\phi(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right]$$

Here, $q_\phi(\mathbf{z}|\mathbf{x})$ is parameterised by another neural network with parameters ϕ. Putting $\mathbf{x} := (s_1, s_2)$ and $\mathbf{z} := f_1$ into the bound gives:

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(s_1, s_2) &\geq \mathbb{E}_{q_\phi(f_1|s_1,s_2)}\left[\log p_{\boldsymbol{\theta}}(s_1, s_2|f_1)\right] - \text{KL}\left[q_\phi(f_1|s_1, s_2))\|p_{\boldsymbol{\theta}}(f_1)\right] \\ &= \mathbb{E}_{q_\phi(f_1|s_1,s_2)}\left[\log p_{\boldsymbol{\theta}}(s_2|s_1, f_1)p_{\boldsymbol{\theta}}(s_1|\cancel{f_1})\right] - \text{KL}\left[q_\phi(f_1|s_1, s_2))\|p_{\boldsymbol{\theta}}(f_1)\right] && (44) \\ &= \mathbb{E}_{q_\phi(f_1|s_1,s_2)}\left[\log p_{\boldsymbol{\theta}}(s_2|s_1, f_1)\right] - \text{KL}\left[q_\phi(f_1|s_1, s_2))\|p(f_1)\right] + \log p(s_1) && (45)\end{aligned}$$

Here, in line 44 we used the independence assumption from the graphical model for the problem. Note that the objective on line 45 is now something we can optimise via stochastic gradient ascent [15][22].

INFERENCE

With $p_{\boldsymbol{\theta}}(s_2|s_1, f_1)$ being computable with a neural network, and having obtained an estimate for the model parameters $\boldsymbol{\theta}_{\text{ML}}$ by maximising the variational objective, we can proceed to do counterfactual inference in a manner equivalent to what we did when F_1 was observed. Namely, we can utilise the expression we obtained in eq. 43:

$$p(s_2^*|s_1, s_2, s_1^*, \mathcal{D}) \propto \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}, s_1, s_2)}\left[\mathbb{E}_{p(f_1)}\left[p_{\boldsymbol{\theta}}(s_2^*|s_1^*, f_1)p_{\boldsymbol{\theta}}(s_2|s_1, f_1)\right]\right]$$

We can again estimate this quantity via Monte-Carlo sampling. Figure 13 shows the outcome of this procedure for the same counterfactual query as we considered previously for the regular deep neural network approach in fig. 12.

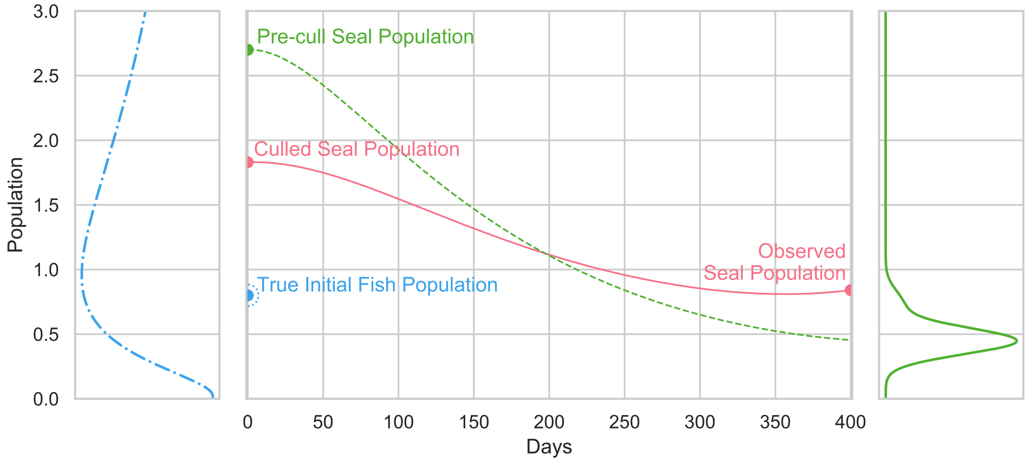


Figure 13: **Left:** the variational approximation to the initial fish population posterior $q_\phi(\log f_1 | s_1, s_2)$. **Right:** the posterior on the counterfactual seal population from the Conditional VAE model. **Centre:** the simulated ground-truth.

We previously discussed in section 7.2.1 that there are infinitely many possible parameterisations on a latent that would result in the same observed distribution, but different counterfactual inference outcomes. As such, we noted that one has to specify assumptions on the form of the dependence on the latent. Above, we did so by assuming the dependence has the form given by a deep feedforward neural network. It is safe to say that the properties of the inductive bias introduced through such an assumption in a counterfactual context are not yet fully understood, and possibly an interesting avenue for future research.

9 Causal inference beyond graphical models.

So far, we have shown how one can approach causal inference with the use of Bayesian Networks to represent manipulations on a system. Bayesian Networks, however, do not constitute the only such tool capable of addressing causal problems. In fact, it is possible to conceive of examples where the use of Bayesian Networks would feel somewhat restrictive.

Concretely, consider a system in which the number of random variables for one observation depends on another random variable. Take as an example a casino game in which the player throws a dice, and depending on its outcome D, they draw D cards $\{C_i\}_{i=1}^D$ from the deck. Then, the sum of the values of the cards drawn is compared to that of a dealer to determine the payout. For a simple six-sided die, we 'could' specify six distinct card nodes $\hat{C}_i$ in a Bayesian Network graph, and augment their sample space to include a special symbol when they aren't drawn. But what if D is an infinite-sided die (a discrete random variable taking on values in $\{0, 1, 2, \ldots\}$ with some probability). Specifying a graph on an infinite number of nodes seems cumbersome — the Bayesian Network representation for this system seems somewhat contrived.

On the other hand, one could for instance, specify a probabilistic model for this setting using probabilistic programming. A probabilistic programming framework that deals with a potentially infinite number of random variables has, for instance, been proposed by Milch et al. [23]. It should be clear that we can extend the ideas from the previous sections to answer counterfactual or interventional queries in such a framework. The key principle was to specify our assumptions through a model over all the tasks and settings of interest. What tool we use to specify those assumptions is secondary.

10 Causal-statistical dichotomy

A quick glance at the causality debate reveals one of the major recurring themes: the proposition that causality is in some sense strictly distinct from statistical analysis. Judea Pearl has repeatedly insisted that these two differ in its goals and approaches, even coming as far as to say *"If I am remembered for no other contribution except for insisting on the causal-statistical distinction, I would consider my scientific work worthwhile"* [7, p. 332]. In this section, we summarise and comment on the arguments for such a distinction, hopefully addressing and resolving the confusion that has arisen from this assertion.

The rationale for this claim is grounded in the argument that standard statistical analysis' only aim is to assess properties of a *static* distribution. Judea Pearl proclaims that causal analysis *"goes one step further; its aim is to infer not only the likelihood of events under static conditions, but also the dynamics of events under changing conditions, for example, changes induced by treatments or external interventions, or by new policies or new experimental designs"* [7, p. 332]. He maintains that there is no more to statistics than inferring or answering questions about the properties of the observational distribution $p(\mathbf{x})$. Once any desirable characteristic of $p(\mathbf{x})$ cam be tractably computed with no uncertainty, there is nothing more to do within the realm of statistics.

This argument of course rests on imposing the above constraining definition onto statistics; it relies on accepting that statistics does in fact only deal with observational distributions and static conditions. Many researchers who consider themselves statisticians would, however, disagree with the claim that statistics reduces to inference within the observational distribution, as there is plenty of work within the statistical realm that extends to non-static domains as well. For instance, research in covariate and dataset shift, domain adaption, off-policy reinforcement learning, or robustness is all about dealing with changes in the distribution from which the variables are drawn.

If one were to consider Judea Pearl's causal-statistical distinction as a new terminological proposition rather than a claim, the nature of the argument becomes clearer. As there is nothing in a distribution function that tells us how it would change if the external conditions were to change, there is a need for additional assumptions to answer these claims. Andrew Gelman, who has notably criticised Pearl's claims in his review of "The Book of Why" [3], admits that he "agree[s] [with Judea Pearl] that data analysis alone cannot solve any causal problems. Substantive assumptions are necessary too". One might then argue that these extra assumptions should be called *causal assumptions*, and analysis that goes beyond inferring the observational distribution should be referred to as *causal analysis*. However, with respect to the necessity of assumptions, there is nothing new here: you always have to make assumptions when doing probabilistic inference. In his book, David MacKay writes: "You can't do inference – or data compression – without making assumptions" [24]. These assumptions are going to be subjective, but once the human input enters through the design of the hypothesis space, in a probabilistic modelling programme the inference is mechanical. Furthermore, as the term *statistical analysis* has not been used in a scope constrained to static distributions in the past, the insistence on the existence of a dichotomy between statistics and causal analysis is somewhat confusing.

In the paper "Causal inference in statistics: an overview" Judea Pearl tempers the terminological separation above by suggesting a distinction between *associational* and causal concepts instead [25, p. 99]. There, he defines the demarcation line by saying that an associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, whereas a causal concept is any relationship that cannot be defined from the observed distribution alone. Although this definition does not rely on unfaithfully limiting the scope of statistics, it does impose the causal trademark onto a broad class of concepts.

11 Discussion

In this note, we gave an overview of the fundamental concepts of causal inference — the nature of various causal problems and the inherent difficulties in tackling those. We related these causal concepts to problems encountered in machine learning and statistics, and demonstrated how you can approach causal inference with methods from these fields. Concretely, we demonstrated that you can do causal inference within the framework of probabilistic modelling by defining a model over all the settings and tasks of interest, and specifying assumptions on how they are related. We gave concrete examples of causal inference problems on which we illustrated this approach, hopefully resolving the confusion surrounding whether you can do causal inference with machine learning techniques.

We further linked the probabilistic modelling approach to the bespoke methods from within the field of causality. Namely, we introduced various frameworks such as *Causal Graphical Models*, *Structural Causal Models* and the *do*-calculus, discussed their properties and proposed utility, and related them to the probabilistic modelling methodology.

Lastly, we have shown how the tools from modern deep learning, such as deep neural networks and variational inference, can be used to tackle causal problems. We followed with a discussion of how researchers in the field of machine learning have been tackling problems similar to those considered causal by Pearl and others, hopefully putting a final nail in the coffin on the claim that machine learning researchers are stuck on the level of curve-fitting.

To conclude, if we wanted the reader to take away one key point from this note, it would be the following:

> *"Always write down the probability of everything."*
>
> — Steve Gull [24]

Or to expand upon that "one key takeaway", we want to firmly recommend: *always write down the joint probability over all the settings you are interested in.* Crucially, we would not be in this mess where we need bespoke causal frameworks if people wrote down a complete model and a joint probability of everything they are intending to do inference over.

References

[1] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[2] Kevin Hartnett. To build truly intelligent machines, teach them cause and effect, May 2018. [Online; posted 15-May-2018]. URL: `https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/`.

[3] Andrew Gelman. Review of "the book of why" by pearl and mackenzie, January 2019. [Online; posted 9-January-2018]. URL: `https://statmodeling.stat.columbia.edu/2019/01/08/book-pearl-mackenzie/`.

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

[5] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017. URL: `http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf`.

[6] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, January 2006. URL: `https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/`.

[7] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

[8] Kevin Murphy. An introduction to graphical models. 96:1–19, 2001. URL: `https://www.cs.ubc.ca/~murphyk/Papers/intro_gm.pdf`.

[9] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[10] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[11] Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, page 230–237. AAAI Press, 1994.

[12] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. *arXiv preprint arXiv:1301.0608*, 2012.

[13] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[14] Jesus Palomo, David B Dunson, and Ken Bollen. Bayesian structural equation modeling. In *Handbook of latent variable and related models*, pages 163–188. Elsevier, 2007.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[17] Ricardo Silva and Robert B Gramacy. Gaussian process structural equation models with latent variables. *arXiv preprint arXiv:1002.4802*, 2010.

[18] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[19] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Unwin, Helen Coupland, T Mellan, Harisson Zhu, T Berah, J Eaton, P Perez Guzman, et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries. 2020.

[20] Judea Pearl. Causes of effects and effects of causes. *Sociological Methods & Research*, 44(1):149–164, 2015. `arXiv:https://doi.org/10.1177/0049124114562614`, `doi:10.1177/0049124114562614`.

[21] A. Philip Dawid, David L. Faigman, and Stephen E. Fienberg. Fitting science into legal contexts: Assessing effects of causes or causes of effects? *Sociological Methods & Research*, 43(3):359–390, 2014. `arXiv:https://doi.org/10.1177/0049124113515188`, `doi:10.1177/0049124113515188`.

[22] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[23] Brian Milch, Bhaskara Marthi, Stuart Russell, David Sontag, Daniel L Ong, and Andrey Kolobov. Blog: Probabilistic models with unknown objects. *Statistical Relational Learning*, page 373, 2007. URL: `https://people.eecs.berkeley.edu/~russell/papers/srlbook-blog.pdf`.

[24] David JC MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge university press, 2003.

[25] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009. URL: `https://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf`.

[26] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

[27] A Philip Dawid. Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application*, 2:273–303, 2015.

[28] Daphne Koller, Nir Friedman, Sašo Džeroski, Charles Sutton, Andrew McCallum, Avi Pfeffer, Pieter Abbeel, Ming-Fai Wong, David Heckerman, Chris Meek, et al. *Introduction to statistical relational learning*. MIT press, 2007.

[29] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[30] Wikipedia contributors. Multivariate normal distribution — Wikipedia, the free encyclopedia, 2020. [Online; accessed 26-May-2020]. URL: `https://en.wikipedia.org/w/index.php?title=Multivariate_normal_distribution&oldid=953478542`.

A Appendix

A.1 Joint log-normal distribution derivation for the aspirin example

To show that $[Z, T, Y]^\top$ are jointly log-normal as depicted in eq. 2, we can use the definition in eq. 1 to write:

$$\log \begin{bmatrix} Z \\ T \\ Y \end{bmatrix} = \begin{bmatrix} \log Z \\ a \log Z + \log \varepsilon_T \\ (b - ac) \log Z - c \log \varepsilon_T + \log \varepsilon_Y \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ (b-ac) & (-c) & 1 \end{bmatrix}}_{A} \begin{bmatrix} \log Z \\ \log \varepsilon_T \\ \log \varepsilon_Y \end{bmatrix}$$

Since $\log[Z, \varepsilon_T, \varepsilon_Y]^\top$ are jointly normal distributed as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [\mu_Z, 0, 0]^\top$ and $\boldsymbol{\Sigma} = \text{diag}([\sigma_Z^2, \sigma_T^2, \sigma_y^2]^\top)$, an affine transformation of these variables is also a Gaussian with mean $A\boldsymbol{\mu}$ and covariance $A\Sigma A^\top$. Hence, $[Z, T, Y]^\top$ are jointly distributed as $\log \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^\top)$. Evaluating the matrix multiplications yields the expression in eq. 2.

A.2 Expectation of Y conditioned on T in the aspirin example

To obtain the expression in eq. 3, first denote the mean and the covariance matrix for the joint on $\log[Z,T,Y]^\top$ as $\overline{\boldsymbol{\mu}}$ and $\overline{\Sigma}$, i.e. $\log[Z,T,Y]^\top \sim \mathcal{N}(\overline{\boldsymbol{\mu}}, \overline{\Sigma})$. Using the marginalisation and conditioning properties of the multivariate normal [30] we get:

$$\log \begin{bmatrix} T \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \overline{\mu}_2 \\ \overline{\mu}_3 \end{bmatrix}, \begin{bmatrix} \overline{\Sigma}_{22} & \overline{\Sigma}_{23} \\ \overline{\Sigma}_{32} & \overline{\Sigma}_{33} \end{bmatrix} \right)$$

$$\log Y | \log T \sim \mathcal{N}\Big(\underbrace{\overline{\mu}_3 + \overline{\Sigma}_{32}\overline{\Sigma}_{22}^{-1} (\log T - \overline{\mu}_2)}_{\mu_{Y|T}}, \underbrace{\overline{\Sigma}_{33} - \overline{\Sigma}_{32}\overline{\Sigma}_{22}^{-1}\overline{\Sigma}_{23}}_{\sigma_{Y|T}} \Big)$$

Again, since $\log Y | \log T$ is Gaussian, $Y|T$ is log-normal distributed. We can then use the property that for a log-normal distribution $\log\mathcal{N}(\mu, \sigma^2)$ the expected value is $\exp(\mu + 0.5\sigma^2)$. Plugging in the values for $\overline{\boldsymbol{\mu}}$ and $\overline{\Sigma}$ into equations for $\mu_{Y|T}$ and $\sigma^2_{Y|T}$ then yields the expression in eq. 3.

A.3 Marginal on Y^* in the aspirin example

In the interventional part of the model, $Y^*|Z^*, T^*$ is distributed in the same way as $Y|Z,T$ — $Y^*|Z^*, T^* \sim \log\mathcal{N}\big(\frac{(Z^*)^b}{(t^*)^c}, 1\big)$. We can equivalently specify $Y^* = \frac{(Z^*)^b}{(t^*)^c}\varepsilon^*_Y$ where $\varepsilon^*_Y \sim \log\mathcal{N}(0,1)$. Hence, we can write:

$$\log \begin{bmatrix} Z^* \\ Y^* \end{bmatrix} = \begin{bmatrix} \log Z^* \\ b\log Z^* + \log\varepsilon^*_Y - c\log t^* \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}}_{B} \begin{bmatrix} \log Z \\ \log\varepsilon^*_Y \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ -c\log t^* \end{bmatrix}}_{\mathbf{d}}$$

Again, as this is an affine transformation of normal-distributed variables, $\log[Z^*, Y^*]^\top$ is Gaussian:

$$\log \begin{bmatrix} Z^* \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_Z \\ b\mu_Z - c\log t^* \end{bmatrix}, \begin{bmatrix} \sigma^2_Z & b\sigma^2_Z \\ b\sigma^2_Z & b^2\sigma^2_Z + \sigma^2_Y \end{bmatrix} \right)$$

Consequently, by marginalising out Z^* we get $Y^* \sim \log\mathcal{N}(b\mu_Z - c\log t^*, b^2\sigma^2_Z + \sigma^2_Y)$. Using the property that for a log-normal distribution $\log\mathcal{N}(\mu, \sigma^2)$ the expected value is $\exp(\mu + 0.5\sigma^2)$, we obtain:

$$\mathbb{E}[Y^*] = \exp\left(b\mu_Z - c\log t^* + \frac{b^2\sigma^2_Z + \sigma^2_Y}{2} \right) = (t^*)^{-c} \exp\left(b\mu_Z + \frac{b^2\sigma^2_Z + \sigma^2_Y}{2} \right)$$

A.4 Proof of the parent adjustment formula

In this appendix, we give a proof of the parent adjustment formula.

The formula says that, in the case of an intervention on X_t with an empty set of new parents, adjusting using the parents of X_t in the observed Bayesian Network graph $\mathcal{G}$ is sufficient, i.e.:

$$q(x_y) = q(x_y|x_t) = \int p(x_y|x_t, \mathbf{x}_{\mathbf{PA}^{\mathcal{G}}_t}) p(\mathbf{x}_{\mathbf{PA}^{\mathcal{G}}_t}) d\mathbf{x}_{\mathbf{PA}^{\mathcal{G}}_t} \tag{46}$$

Where $q(\cdot)$ are the distribution functions in the intervened setting, and $p(\cdot)$ correspond to the observed setting. We can show that the parent adjustment formula holds by expanding:

$$\begin{aligned} q(x_y|x_t) &= \frac{q(x_y, x_t)}{q(x_t)} = \frac{q(x_y, x_t)}{g^*(x_t)} \\ &= \int \left(\int \frac{q(\mathbf{x})}{g^*(x_t)} d\mathbf{x}_{other} \right) d\mathbf{x}_{\mathbf{PA}^{\mathcal{G}}_t} \end{aligned}$$

where the inner integral over $\mathbf{x}_{other}$ is taken with respect to all nodes $\{1, \dots, d\}$ other than $\{t, y\} \cup \mathbf{PA}_t^{\mathcal{G}}$. Expanding the joint $q(\mathbf{x})$:

$$
\begin{aligned}
&= \int \left(\int \frac{\cancel{g^*(x_t)} \prod_{i \neq t} g\left(x_i, x_{\mathbf{PA}_i^{\mathcal{G}}}\right)}{\cancel{g^*(x_t)}} d\mathbf{x}_{other} \right) d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}} \\
&= \int \left(\int \frac{g_t(x_t, \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}) \prod_{i \neq t} g\left(x_i, x_{\mathbf{PA}_i^{\mathcal{G}}}\right)}{g_t(x_t, \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}})} d\mathbf{x}_{other} \right) d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}} \\
&= \int \left(\int \underbrace{\prod_i g\left(x_i, x_{\mathbf{PA}_i^{\mathcal{G}}}\right)}_{p(\mathbf{x})} d\mathbf{x}_{other} \right) \frac{1}{g_t(x_t, \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}})} d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}} \\
&= \int \left(\int p(\mathbf{x}) d\mathbf{x}_{other} \right) \frac{1}{p(x_t | \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}})} d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}} \\
&= \int p(x_y, x_t, \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}) \frac{1}{p(x_t | \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}})} d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}} \\
&= \int p(x_y | x_t, \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}) p(x_t | \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}) p(\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}) \frac{1}{p(x_t | \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}})} d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}} \\
&= \int p(x_y | x_t, \mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}) p(\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}) d\mathbf{x}_{\mathbf{PA}_t^{\mathcal{G}}}
\end{aligned}
$$

where the last line is the same as the right-hand side of equation 46.

A.5 The rules of the *do*-calculus

In this section we give the rules of the *do*-calculus for reader's reference. It is helpful to define some extra notation before doing so, however. We'll write $\mathcal{G}_{\overline{\mathbf{X}}}$ for the graph obtained by deleting all the incoming edges from the nodes in set $\mathbf{X}$ in graph $\mathcal{G}$. Similarly, we'll write $\mathcal{G}_{\underline{\mathbf{X}}}$ for the graph obtained by deleting all the outgoing edges from the nodes in set $\mathbf{X}$ in graph $\mathcal{G}$

For a Causal Graphical Model (or a Structural Causal Model) with a graph $\mathcal{G}$ and any disjoint subsets of variables X, Y, Z and W, the rules of the *do*-calculus are as follows:

1. **Insertion/deletion of observations**

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

if $\mathbf{Y}$ and $\mathbf{Z}$ are *d*-separated by $\mathbf{X}, \mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$

2. **Action/observation exchange**

$$p(\mathbf{y} | \mathbf{w}, do(\mathbf{X} := \mathbf{x}, \mathbf{Z} = \mathbf{z})) = p(\mathbf{y} | \mathbf{z}, \mathbf{w}, do(\mathbf{X} := \mathbf{x}))$$

if $\mathbf{Y}$ and $\mathbf{Z}$ are *d*-separated by $\mathbf{X}, \mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$.

3. **Insertion/deletion of actions**

$$p(\mathbf{y} | \mathbf{w}, do(\mathbf{X} := \mathbf{x}, \mathbf{Z} = \mathbf{z})) = p(\mathbf{y} | \mathbf{w}, do(\mathbf{X} := \mathbf{x}))$$

if $\mathbf{Y}$ and $\mathbf{Z}$ are *d*-separated by $\mathbf{X}, \mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}\,\overline{\mathbf{Z}_{(\mathbf{W})}}}$ where $\mathbf{Z}_{(\mathbf{W})}$ is the subset of nodes in $\mathbf{Z}$ that are not ancestors of any nodes in $\mathbf{W}$ in graph $\mathcal{G}_{\overline{\mathbf{X}}}$.