Causal Inference using Probabilistic Modelling

Word Count: 3975

27th January 2021

1 Introduction

1.1 Motivation

Probabilistic modelling is a framework for understanding and inferring probability distributions, whilst handling uncertainty in a principled way. It was initially developed to answer questions of an associative nature, an example of which might be "What is a patient's probability of recovery if we observe that they have taken a particular drug compared with if we observe that they have not?". It has been argued in recent years, most notably by Judea Pearl in [6], that the traditional tools of probabilistic modelling are not equipped to deal with problems of a causal nature, such as "What will be the effect on a patients probability of recovery if we intervene and give them a certain dose of the drug?".

The reason that these two questions differ, and that an approach to answering the second question that is based on comparing the observational rates of recovery among those that have taken the drug and those that have not can fail catastrophically is a subtle one. The key is that there can often be lurking confounders. In our example it could be the case that the drug is administered preferentially to the most vulnerable, meaning that old people are more likely to receive the treatment as young people are believed to have a high probability of recovery regardless. This means that even though intervening and administering the drug could increase recovery probability for all age groups, the rate of recovery could be lower among those who received the drug in the observational world, because those who received it already had a much lower probability of recovery. This counter-intuitive result is known as Simpson's paradox and is often pointed to as an example of the dangers of assigning causal interpretations to models based on observational data without careful consideration of the confounders.

In this report, we aim to understand how to conduct causal inference and answer counterfactual questions using probabilistic models. To illustrate the core ideas we consider a simple example in which we answer a counterfactual question in the context of a prey-predator problem: "what would the seal population be now, had fishermen not culled it some time ago?". In Section 2, we replicate the counterfactual analysis performed in Sections 4.4 and 4.5 of [4]. We note that this basic scenario does not involve confounding variables and, as such, cannot adequately represent the challenges of causal inference. Therefore, in Section 3 and 4 we consider a more complex scenario where there is a confounder in the prey-predator population data generating process. Section 5 concludes the report, and Section 6 details a possible direction for future work

1.2 Counterfactual analysis using probabilistic modelling

Causal inference can be approached using different frameworks, such as graphical models, structural equations, or potential outcomes. Inspired by the probabilistic modelling framework used in [4] and hierarchical Bayesian models used in [1, 2, 3], we will focus on the study of causal inference using probabilistic modelling.

In general, the aim of probabilistic modelling is to learn the parameters θ of a model that defines the posterior distribution $p(y|\boldsymbol{x}, \theta)$ over the variables of interest y, given some observed variables $\boldsymbol{x}$. In the supervised learning setting, we have access to some training data $\mathcal{D}$, consisting of paired observations of $\boldsymbol{x}$ and y, which can be used to estimate the parameters θ. In the case where some variables in $\boldsymbol{x}$ are not observed, there exist unsupervised learning methods such as variational auto-encoders (VAEs), which can help us estimate a posterior distribution over the latent variables and use this to estimate the posterior distribution over the variables of interest.

In Judea Pearl's *The Book of Why* [6] and also in his more technical book [5], he advocates for approaching causal inference problems using causal diagrams and his three rules of *do-calculus*, which he proposes give a sufficient framework for answering causal questions of any kind, so long as it is possible to do so. However, throughout this report we take an alternative approach proposed in [4], where a Bayesian Network[1] (BN) illustrates the relationship between variables in the observational and counterfactual worlds. More specifically, all nodes in the BN are shared, but the intervened-upon variable and output variable nodes, which are unique for observational and counterfactual scenarios. This allows us to define a set of conditional independence assumptions which can then be used to factorise a joint distribution over all variables of interest. We then define some further assumptions that relate distributions in the observational and counterfactual worlds, which is necessary to allow for use of observational data in causal inference. The next step is to define an appropriate probabilistic model that represents the data generating process and to learn the model parameters from the training set. Finally, we can infer the posterior distributions over the variables in the counterfactual world using the rules of probability.

2 Counterfactual analysis with no confounder

In this section, we write our own code to replicate the counterfactual analysis for the prey-predator problem performed in Sections 4.4 and 4.5 of [4]. We also provide more detailed derivations, explanations and analysis for the methods we use.

2.1 The prey-predator counterfactual problem

The prey-predator counterfactual problem setting is as follows. Suppose that the fishermen had culled the local seal population on day 0, in the hope that this would help the growth of the fish population they rely on. Unfortunately, the seal became an endangered species on day 400. We would like to understand whether the fishermen are responsible for this. That is, we want to answer the counterfactual question "What would the seal population

[1] In a BN nodes represent variables and directed edges indicate dependencies between variables.

have been if the fishermen did not cull it earlier?".

The mathematical formulation of this problem is as follows. Suppose that the pre-cull seal population s_1^* ("$*$" denotes quantities in the counterfactual world), post-cull seal population s_1, and current seal population s_2 are all observed. The goal is to infer the posterior distribution $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$, where s_2^* is the current seal population in the counterfactual world (where seal was not initially culled) and $\mathcal{D}$ is the observed dataset we have access to. We will need to learn the model $p(s_2|s_1, f_1, \theta)$ from the dataset $\mathcal{D}$, which will be necessary for causal inference. During inference, the initial fish population f_1 is unobserved, as it is usually not easily countable.

2.2 Simulator

The simulator we use is based on the Lotka-Volterra Differential Equations. The Lotka-Volterra Model is a system of two ordinary differential equations (ODEs) that describes the prey-predator population dynamics as time evolves:

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

where f is the fish population (in millions), s is the seal population (in thousands), and $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ (real positive numbers) are the parameters of the simulator:

- γ_1: the natural growth rate of fish;
- γ_2: the dying rate of fish due to seal predation;
- γ_3: the seal growth rate due to fish abundance;
- γ_4: the natural dying rate of seal.

We set $\gamma_1 = 0.015$, $\gamma_2 = 0.012$, $\gamma_3 = 0.007$, and $\gamma_4 = 0.009$ and use this model as the "ground truth" data generating process. For the inference problem, we assume the pre-cull seal population $s_1^* = 2.7$, the post-cull seal population $s_1 = 2.2$, and the initial fish population $f_1 = 0.8$. Figure 1 shows the ground truth dynamics in both observed and counterfactual worlds. Note that the training set $\mathcal{D}$ is also sampled from this simulator.

2.3 Modelling and inference

2.3.1 Supervised learning setting

To gradually increase the complexity of the analysis, we first consider the training set $\mathcal{D} = \{(f_{1,n}, s_{1,n}, s_{2,n})\}_{n=1}^N$ that contains the initial fish population $f_{1,n}$. Figure 2 shows the dependencies between the variables in the training set and in both observed and counterfactual worlds. The first step is to write down the joint distribution of everything according to the conditional independence assumptions illustrated by the BN:

$$p(s_1, s_2, f_1, s_1^*, s_2^*, \mathcal{D}, \theta) = p(s_1)p(s_2|f_1, s_1, \theta)p(f_1)q(s_1^*)q(s_2^*|f_1, s_1^*, \theta)p(\mathcal{D}|\theta)p(\theta),$$

where q denotes the distribution of variables in the counterfactual world. Here we assume that model parameters θ encapsulate all information about the dataset $\mathcal{D}$. Therefore, the

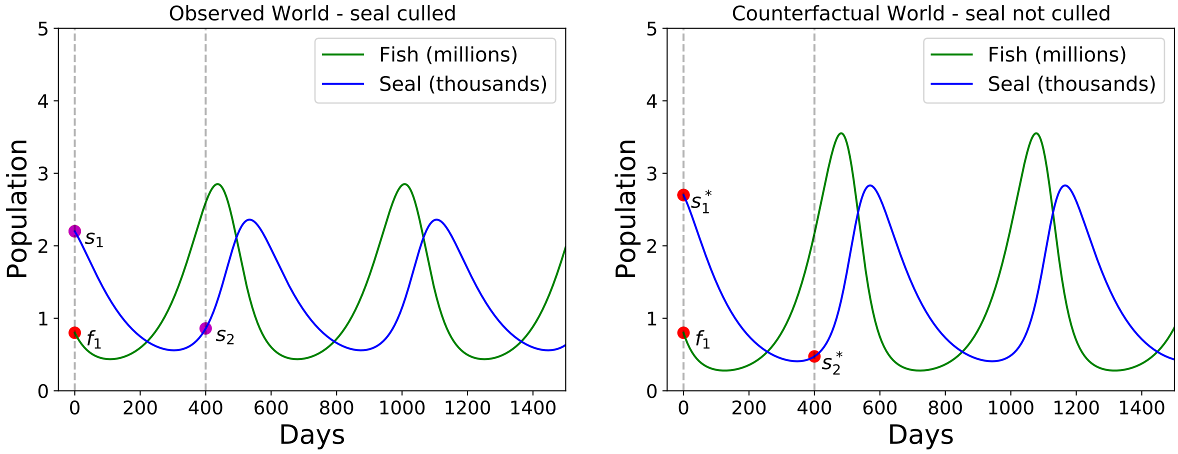


Figure 1: The ground truth seal and fish population in the observed and counterfactual worlds on day 0-1499, given by the simulator.

current and initial observations we consider are independent given model parameters. The likelihood is defined as:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(s_{2,n}, s_{1,n}, f_{1,n}|\theta) \propto \prod_{n=1}^{N} p(s_{2,n}|s_{1,n}, f_{1,n}, \theta).$$

We can approximate the posterior $p(s_2|f_1, s_1, \theta)$ with a Gaussian distribution

$$p(s_2|f_1, s_1, \theta) \approx \mathcal{N}(s_2; g_{\boldsymbol{w}}(f_1, s_1), \sigma^2)$$

whose mean is given by the output of a neural network $g_{\boldsymbol{w}}(f_1, s_1)$ and whose variance, σ^2, is another parameter to be learned. The neural network architecture we use is a two-hidden-layer MLP with 100 hidden units and ReLU activation.

The distributions $p(s_1)$, $p(f_1)$, $q(s_1^*)$, and $q(s_2^*|f_1, s_1^*, \theta)$ are still unspecified. We define them using further assumptions listed below. In fact, it is impossible to conduct causal inference without stating the assumptions of how the observed and counterfactual distributions relate to each other (inductive bias).

1. The initial fish population is assumed to be distributed as $p(f_1) = \log \mathcal{N}(f_1; 0, 1)$;

2. The physical mechanism determining the dynamics of the fish and seal population in the observed world is assumed to be the same as that in the counterfactual world. That is $p(s_2|f_1, s_1, \theta) = q(s_2|f_1, s_1, \theta)$;

3. The initial seal distribution in the counterfactual world is assumed to be $q(s_1^*) = \log \mathcal{N}(s_1^*; 0, 1)$. In the observed world, we assume $p(s_1)$ to be a delta function, indicating that marginalizing over s_1 is equivalent to conditioning on $S_1 = s_1$.

The main inference objective for answering the counterfactual question is the posterior distribution $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$. Using the conditional independence assumptions specified in

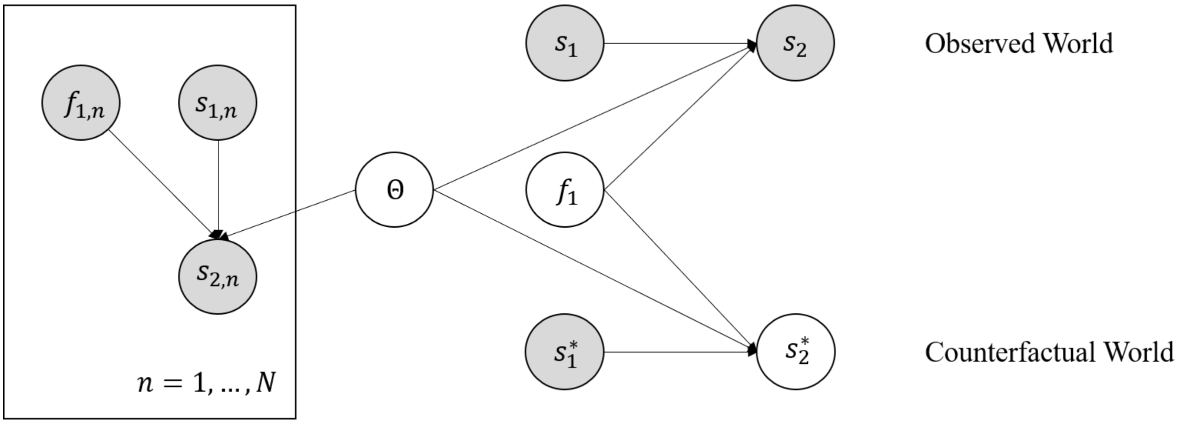


Figure 2: The Bayesian network of the model used for answering the prey-predator counterfactual problem.

the BN and rules of probability, we have

$$\begin{aligned} p(s_2^*|s_1, s_2, s_1^*, \mathcal{D}) &= \iint p(s_2^*, f_1, \theta|s_1, s_2, s_1^*, \mathcal{D}) df_1 d\theta \\ &= \iint p(s_2^*|s_1^*, f_1, \theta) p(f_1|s_1, s_2, \theta) p(\theta|s_1, s_2, \mathcal{D}) df_1 d\theta \\ &= \mathbb{E}_{p(\theta|s_1, s_2, \mathcal{D})} \left[\int p(s_2^*|s_1^*, f_1, \theta) p(f_1|s_1, s_2, \theta) df_1 \right] \\ &\propto \mathbb{E}_{p(\theta|\mathcal{D}, s_1, s_2)} \left[\mathbb{E}_{p(f_1)} \left[p(s_2^*|s_1^*, f_1, \theta) p(s_2|s_1, f_1, \theta) \right] \right], \end{aligned}$$

where the last line follows by Bayes rule

$$p(f_1|s_1, s_2, \theta) \propto p(s_2|s_1, f_1, \theta) p(f_1).$$

For the outer expectation over the posterior distribution of the parameters θ, it is reasonable to assume that the impact of s_1 and s_2 is negligible compared to the training set $\mathcal{D}$, which gives $p(\theta|s_1, s_2, \mathcal{D}) \approx p(\theta|\mathcal{D})$. Furthermore, we use maximum likelihood estimation to approximate $p(\theta|\mathcal{D}) \approx \delta(\theta - \theta_{\text{ML}})$, where θ_{ML} is given by

$$\begin{aligned} \theta_{\text{ML}} &= \underset{\theta}{\operatorname{argmax}} \log p(\mathcal{D}|\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(s_{2,n}|f_{1,n}, s_{1,n}, \theta) \\ &= \underset{\boldsymbol{w}, \sigma^2}{\operatorname{argmax}} \sum_{n=1}^{N} \log \mathcal{N}(s_{2,n}; g_{\boldsymbol{w}}(f_{1,n}, s_{1,n}), \sigma^2) \\ &= \underset{\boldsymbol{w}, \sigma^2}{\operatorname{argmin}} \frac{1}{2\sigma^2} \sum_{n=1}^{N} (g_{\boldsymbol{w}}(s_{1,n}, f_{1,n}) - s_{2,n})^2 + \frac{N}{2} \log \sigma^2 + \frac{N}{2} \log(2\pi) \\ \implies \boldsymbol{w}_{\text{ML}} &= \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{n=1}^{N} (g_{\boldsymbol{w}}(s_{1,n}, f_{1,n}) - s_{2,n})^2 \\ \sigma^2_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^{N} (g_{\boldsymbol{w}_{\text{ML}}}(s_{1,n}, f_{1,n}) - s_{2,n})^2. \end{aligned}$$

This involves training the neural network using a mean squared error (MSE) loss function, and setting the value of σ^2 to the MSE after training. Interestingly this differed from the approach in [4] where the variance is set to an arbitrary value.

The inner expectation over the prior distribution $p(f_1)$ can then be approximated by Monte Carlo sampling:

$$\begin{aligned}
p(s_2^*|s_1, s_2, s_1^*, \mathcal{D}) &\propto \mathbb{E}_{p(\theta|\mathcal{D},s_1,s_2)}\left[\mathbb{E}_{p(f_1)}\left[p(s_2^*|s_1^*, f_1, \theta)p(s_2|s_1, f_1, \theta)\right]\right] \\
&\approx \mathbb{E}_{p(f_1)}\left[p(s_2^*|s_1^*, f_1, \theta_{\text{ML}})p(s_2|s_1, f_1, \theta_{\text{ML}})\right] \\
&\approx \frac{1}{L}\sum_{l=1}^{L} p(s_2^*|s_1^*, f_1^{(l)}, \theta_{\text{ML}})p(s_2|s_1, f_1^{(l)}, \theta_{\text{ML}}), \quad f_1^{(l)} \sim p(f_1) \\
&\propto \sum_{l=1}^{L} w^{(l)} p(s_2^*|s_1^*, f_1^{(l)}, \theta_{\text{ML}}), \quad w^{(l)} := p(s_2|s_1, f_1^{(l)}, \theta_{\text{ML}}) \\
\implies p(s_2^*|s_1, s_2, s_1^*, \mathcal{D}) &\approx \sum_{l=1}^{L} \pi^{(l)} p(s_2^*|s_1^*, f_1^{(l)}, \theta_{\text{ML}}), \quad \pi^{(l)} := \frac{w^{(l)}}{\sum_{j=1}^{L} w^{(j)}}.
\end{aligned}$$

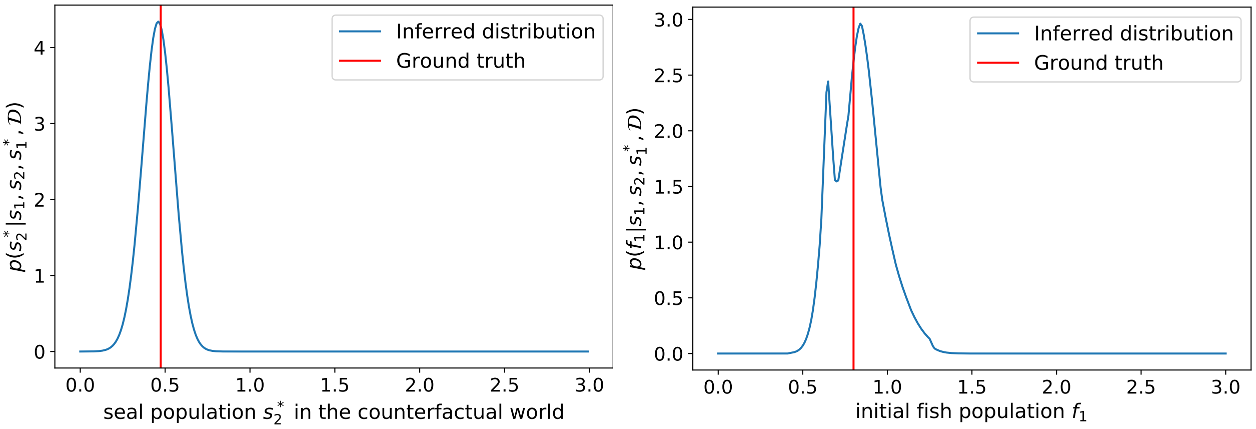
We also infer the posterior distribution of the (unobserved) initial fish population f_1 for the completeness of this inference problem:

$$\begin{aligned}
p(f_1|s_1, s_2, s_1^*, \mathcal{D}) &= \mathbb{E}_{p(\theta|\mathcal{D},s_1,s_2)}[p(f_1|s_1, s_2, \theta)] \\
&\approx p(f_1|s_1, s_2, \theta_{\text{ML}}) \\
&\propto p(s_2|s_1, f_1, \theta_{\text{ML}})p(f_1),
\end{aligned}$$

for which the normalizing constant $\int p(s_2|s_1, f_1, \theta_{\text{ML}})p(f_1)df_1$ is computed by numerical integration.

Figure 3 shows the inferred posterior distributions $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$ and $p(f_1|s_1, s_2, s_1^*, \mathcal{D})$ when the training set $\mathcal{D}$ contains $f_{1,n}$. In probabilistic modelling we are interested in the posterior distribution as a whole, but The maximum a posteriori (MAP) estimate for variables of interest can sometimes be a good indication of whether the distribution is centred in the correct place. The MAP estimate for the value of s_2^* is 0.460 which is close to the ground truth value given by the output of the simulator, which was 0.474. This accuracy, along with the narrow posterior distribution indicating low uncertainty, shows that the inference methods used were capable of accurately predicting s_2^*. The MAP estimate of the value of f_1 is 0.84 is also close to it's ground truth value of 0.80. We note that the the inferred posterior of f_1 is bimodal. The uncertainty in f_1 is also larger than that of s_2^*, which is reflected in the wider posterior distribution.

Interestingly, whilst the inference does seem to be accurate, the posterior distribution over s_2^* that we obtained has a different shape to that from Section 4.4 of [4], which was bimodal. There are a few reasons that this may be the case, such as differences in neural network architecture, differences in the quantity of training data, the fact that we used the ML estimate of σ^2, or possibly differences in the MC sampling procedure.



(a) $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$

(b) $p(f_1|s_1, s_2, s_1^*, \mathcal{D})$

Figure 3: Inferred posterior distributions of s_2^* and f_1 when $\mathcal{D}$ contains $f_{1,n}$.

2.3.2 Unsupervised learning setting

In a realistic scenario, the training set $\mathcal{D} = \{(s_{1,n}, s_{2,n})\}_{n=1}^N$ will not contain the initial fish population $f_{1,n}$. The likelihood is defined as:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \int p(s_{2,n}, s_{1,n}, f_{1,n}|\theta) df_{1,n} \propto \prod_{n=1}^{N} \int p(s_{2,n}|s_{1,n}, f_{1,n}, \theta) p(f_{1,n}) df_{1,n}.$$

In this case, we need a flexible latent variable model to discover the distribution $p(s_2|s_1, f_1, \theta)$. We choose VAEs which are a kind of deep latent variable model that learns a parametric function mapping from the latent space to the observed space by maximizing the mariginal likelihood of the observed variables.

For any given variational distribution $q_\phi(f_1|s_1, s_2)$ with variational parameters ϕ, the objective of VAEs is the variational lower bound on the marginal likelihood $\log p_\theta(s_1, s_2)$:

$$\begin{aligned}
\log p_\theta(s_1, s_2) &= \mathbb{E}_{f_1 \sim q_\phi(f_1|s_1,s_2)} [\log p_\theta(s_1, s_2)] \\
&= \mathbb{E}_{f_1} \left[\log \frac{p_\theta(s_1, s_2|f_1) p_\theta(f_1)}{p_\theta(f_1|s_1, s_2)} \right] \\
&= \mathbb{E}_{f_1} \left[\log \frac{p_\theta(s_1, s_2|f_1) p_\theta(f_1)}{p_\theta(f_1|s_1, s_2)} \frac{q_\phi(f_1|s_1, s_2)}{q_\phi(f_1|s_1, s_2)} \right] \\
&= \mathbb{E}_{f_1} [\log p_\theta(s_1, s_2|f_1)] - \mathbb{E}_{f_1} \left[\log \frac{q_\phi(f_1|s_1, s_2)}{p_\theta(f_1)} \right] + \mathbb{E}_{f_1} \left[\log \frac{q_\phi(f_1|s_1, s_2)}{p_\theta(f_1|s_1, s_2)} \right] \\
&= \mathbb{E}_{f_1} [\log p_\theta(s_1, s_2|f_1)] - \mathrm{KL}(q_\phi(f_1|s_1, s_2) || p_\theta(f_1)) + \mathrm{KL}(q_\phi(f_1|s_1, s_2) || p_\theta(f_1|s_1, s_2)) \\
&\geq \mathbb{E}_{f_1} [\log p_\theta(s_1, s_2|f_1)] - \mathrm{KL}(q_\phi(f_1|s_1, s_2) || p_\theta(f_1)),
\end{aligned}$$

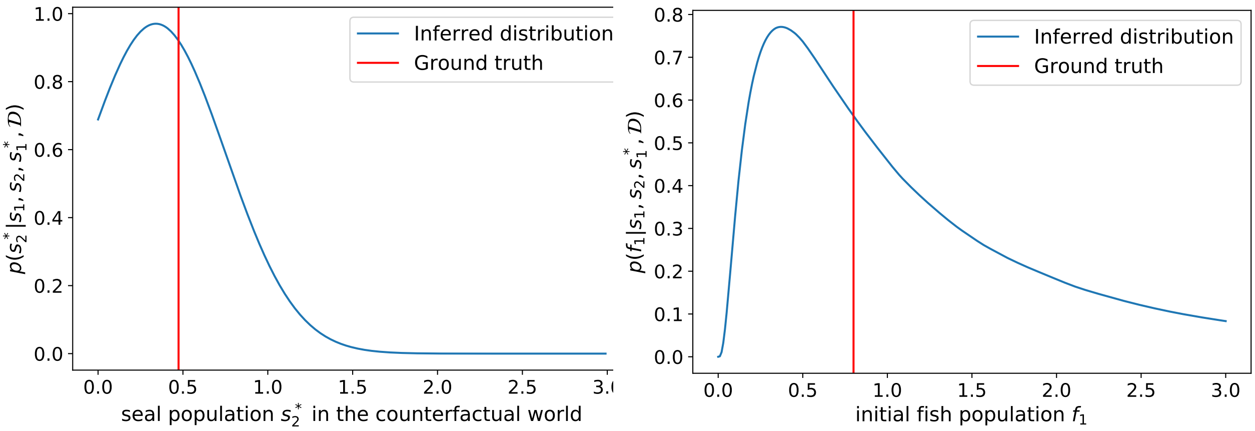
where the last line follows by the non-negativity of KL divergence. This gives us the VAE objective:

$$\max_{\theta,\phi} \mathcal{F}(\theta, \phi) = \mathbb{E}_{f_1 \sim q_\phi(f_1|s_1,s_2)} [\log p_\theta(s_2|s_1, f_1)] - \mathrm{KL}(q_\phi(f_1 \mid s_1, s_2) || p(f_1)),$$

where, from an auto-encoder point of view, the first term on the right hand side can be viewed as the negative expected reconstruction error, and the second term a regularizer.

This is called amortized variational inference – the variational distribution $q_\phi\left(f_1|s_1, s_2\right)$ is assumed to be log-normal with both mean and variance parameterized by a two-hidden-layer MLP with 100 hidden units and ReLU activation and is used to approximate the intractable true posterior distribution $p_\theta\left(f_1|s_1, s_2\right)$ (inference network). The generative model $p_\theta\left(s_2|s_1, f_1\right)$ is approximated by a Gaussian with mean parameterized by a neural network with the same architecture as the inference network, and variance σ^2. They are jointly trained by maximizing the variational free energy objective using stochastic gradient optimization methods.

Once we have learned the model parameters θ_{ML} using a VAE, we can apply the same inference as we did in the previous section to answer the counterfactual question. Figure 4 shows the inferred posterior distributions $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$ and $p(f_1|s_1, s_2, s_1^*, \mathcal{D})$ when the training set $\mathcal{D}$ does not contain $f_{1,n}$. The MAP estimates are slightly further from the ground truth values, with the MAP estimate for s_2^* being 0.340, and the MAP estimate for f_1 being 0.370. These distributions are much wider than the corresponding distributions in the case where $f_{1,n}$ is observed in the training set. This is due to the inability to learn the distribution $p(s_2|s_1, f_1)$ in a supervised setting, and thus there could be many more plausible explanations of the observed data. Hence, the unsupervised approach leading to a broader estimate, and therefore more uncertainty in the final inferences of s_2^* and f_1.



(a) $p(s_2^*|s_1, s_2, s_1^*, \mathcal{D})$

(b) $p(f_1|s_1, s_2, s_1^*, \mathcal{D})$

Figure 4: Inferred posterior distributions of s_2^* and f_1 when $\mathcal{D}$ does not contain $f_{1,n}$.

3 Counterfactual analysis with confounders

3.1 Confounding

In Section 2, we introduced how the probabilistic modelling framework can be used to answer counterfactual questions when the covariates are fully observed or latent. However, as such, the studied example did not go beyond what can be achieved with traditional probabilistic inference. The observed data was fully generalisable to the scenario where the initial seal population s_1 was intervened-upon by culling. This is because initial seal population was an independent factor and the distribution was not affected by an intervention – there are no incoming links to the variables s_1 or s_1^*, so instead of phrasing the problem as a counterfactual question, it could be thought of as enquiring about the output values given a different set of input values in a standard probabilistic model.

In this section we extend the BN to include an additional confounding variable z, which has effect on both initial seal population (treatment) and current seal population (outcome). This more complex scenario will enable us to study how the presence of confounders affects the inference. Most notably, the fact that the node s_1 in the observed world represents an intervention as opposed to an observation of the same value is now a crucial consideration when performing the inference:

$$p(s_2|f_1, s_1) \coloneqq P(s_2|f_1, do(S_1 = s_1)).$$

The potential confounders could include anything that affects the generating process of both $s_{1,n}$ and $s_{2,n}$, such as migration or seasonal variation in seal behaviour. For clarity of argument, we consider a general confounding case in this section and will explore a more specific example of migration confounder later.

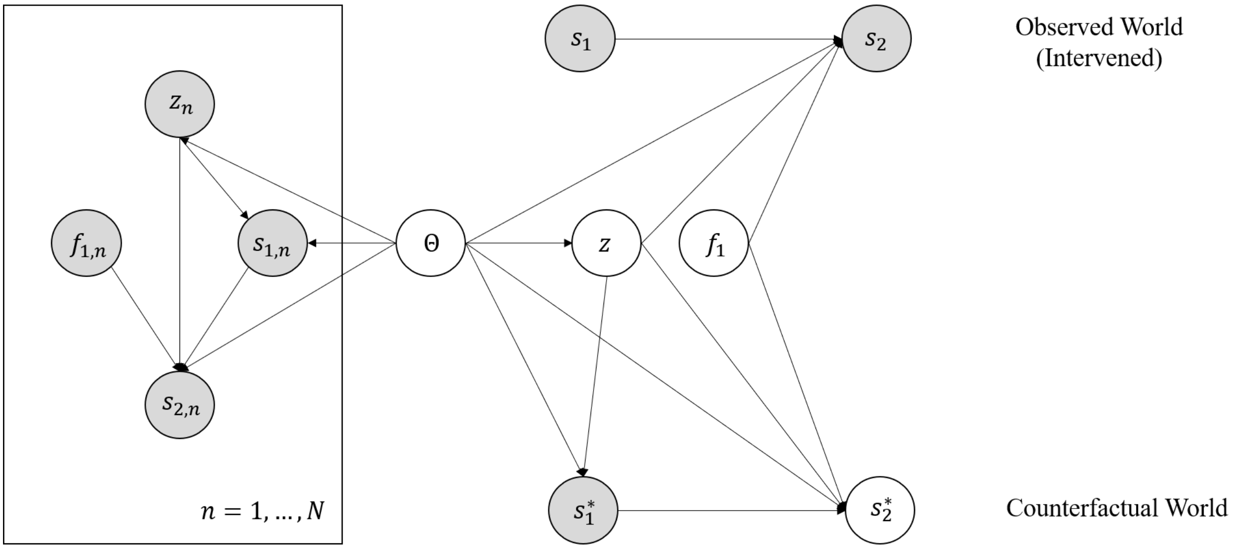


Figure 5: Bayesian Network from Figure 2 extended to include simple dependency of seal population on confounder z.

Figure 5 shows that the counterfactual pre-cull initial population s_1^* and current seal population s_2 and s_2^* are both dependent on some common variable z, which determines how seal population evolves. In the post-cull, observed world, we assume that the fishermen culled the seal population to some predefined value (which the fishermen thought was optimal for fishing), so s_1 becomes independent of any pre-cull levels and the confounder z.

We assume $p(s_1)$ to be an atomic intervention such that $p(s_1) = \delta(s_1 - s_1')$. As a result, the physical mechanism that determines the initial seal population is no longer the same as $q(s_1^*|z)$ in the counterfactual world. It is important to note that we assumed a simplistic intervention, where s_1 has no parent nodes. In reality, a general intervention $p(s_1|z)$ could be of some arbitrary form and could still dependent on z, as we will show later.

As a result of the confounding, the posterior distribution of s_2 will not be the same across observed and counterfactual worlds. That is

$$p(s_2|f_1, s_1) \coloneqq P(s_2|f_1, do(S_1 = s_1)) \neq q(s_2|f_1, s_1).$$

However, for the simple intervention, where s_1 has no parent nodes, it is sufficient to simply adjust the confounding variable z to obtain the conditional distribution $p(s_2|f_1, s_1)$ for the observed (intervened) world using the adjustment formula [7] (see Section 3.3 for derivation and a more general treatment of intervention):

$$\begin{aligned} p(s_2|f_1, s_1') &:= P(s_2|f_1, do(S_1 = s_1')) \\ &= \int q(s_2|f_1, s_1', z)q(z)dz. \end{aligned}$$

3.2 Probabilistic modelling approach

Following the framework presented in Section 2, to answer counterfactual questions using probabilistic modelling, we start by stating and factorising the joint distribution over variables of interest:

$$\begin{aligned} & p(s_1, s_2, f_1, s_1^*, s_2^*, z, \mathcal{D}, \theta) \\ = \; & p(s_1, s_2, f_1, s_1^*, s_2^*, z|\theta)p(\mathcal{D}|\theta)p(\theta) \\ = \; & p(s_1)p(s_2|f_1, s_1, z, \theta)p(f_1)p(z|\theta)q(s_1^*|z, \theta)q(s_2^*|f_1, s_1^*, z, \theta)p(\mathcal{D}|\theta)p(\theta). \end{aligned}$$

The joint distribution above was factorised according to BN network in Figure 5. Especially, the assumption that $p(z|\theta)$ and $p(f_1)$ are shared between observed and counterfactual worlds allows us to focus on effect of the seal culling intervention (according to the definition in [5], all sources of randomness have to be shared for the outcome to be a counterfactual). As in Section 2, we also assume the generating process of the current seal population s_2 from (s_1, f_1, z) to be identical in observed and counterfactual worlds:

$$p_\theta(s_2|f_1, s_1, z) = q_\theta(s_2|f_1, s_1, z).$$

This distribution can be learned using supervised learning algorithms if we have access to $\mathcal{D} = \{(f_{1,n}, z_n, s_{1,n}, s_{2,n})\}_{n=1}^N$ or modelled with VAEs if $f_{1,n}$ and/or z_n are not available in $\mathcal{D}$. The likelihood is defined as:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^N p(s_{2,n}, s_{1,n}, f_{1,n}, z_n|\theta) \propto \prod_{n=1}^N p(s_{2,n}|s_{1,n}, f_{1,n}, z_n, \theta)$$

or

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{n=1}^N \iint p(s_{2,n}, s_{1,n}, f_{1,n}, z_n|\theta)dz_n df_{1,n} \\ &\propto \prod_{n=1}^N \iint p(s_{2,n}|s_{1,n}, f_{1,n}, z_n, \theta)p(z_n|\theta)p(f_{1,n})dz_n df_{1,n}. \end{aligned}$$

For the BN in Figure 5, we assume that during inference both z and f_1 are latent, and we are still interested in the posterior $p(s_2^*|s_1, s_2, s_1^*, D)$ for answering the counterfactual question. That is, "given the observational data $\mathcal{D}$, pre-cull and post-cull initial seal populations s_1^* and s_1, and current post-cull seal population s_2, what would the current seal population s_2^* be had it not been culled earlier?". This can be inferred using similar techniques presented in Section 2:

$$\begin{aligned}
p(s_2^*|s_1,s_2,s_1^*,\mathcal{D}) &= \iiint p(s_2^*,f_1,z,\theta|s_1,s_2,s_1^*,\mathcal{D})df_1dzd\theta \\
&= \iiint p(s_2^*|s_1^*,f_1,z,\theta)p(f_1|s_1,s_2,z,\theta)p(z|s_1,s_2,\theta)p(\theta|s_1,s_2,\mathcal{D})df_1dzd\theta \\
&\approx \iint p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*,f_1,z)p_{\theta_{\mathrm{ML}}}(f_1|s_1,s_2,z)p_{\theta_{\mathrm{ML}}}(z|s_1,s_2)df_1dz \\
&\propto \iint p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*,f_1,z)p_{\theta_{\mathrm{ML}}}(s_2|s_1,f_1,z)p(f_1)p_{\theta_{\mathrm{ML}}}(s_2|s_1,z)p_{\theta_{\mathrm{ML}}}(z)df_1dz \\
&= \mathbb{E}_{p(z)}\left\{\mathbb{E}_{p(f_1)}\left[p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*,f_1,z)p_{\theta_{\mathrm{ML}}}(s_2|s_1,f_1,z)\right]\mathbb{E}_{p(f_1)}\left[p_{\theta_{\mathrm{ML}}}(s_2|s_1,f_1,z)\right]\right\},
\end{aligned}$$

where the expectations over $p(f_1)$ and $p(z)$ can be approximated by MC sampling. In Section 4, we will consider a concrete example with migration confounder.

3.3 Beyond atomic intervention

We now derive the adjustment formula for the atomic intervention $p(s_1) = \delta(s_1 - s_1')$. In the counterfactual world, we have the following causal factorization:

$$q(s_2,s_1,z|f_1) = q(z)q(s_1|z)q(s_2|f_1,s_1,z).$$

Applying the atomic intervention, we have

$$\begin{aligned}
P(s_2,s_1,z|f_1,do(S_1=s_1')) &:= q(z)p(s_1)q(s_2|f_1,s_1,z) \\
&= q(z)\delta(s_1-s_1')q(s_2|f_1,s_1,z).
\end{aligned}$$

Integrating over s_1 and z gives us the adjustment formula:

$$\begin{aligned}
p(s_2|f_1,s_1') := P(s_2|f_1,do(S_1=s_1')) &= \iint P(s_2,s_1,z|f_1,do(S_1=s_1'))ds_1dz \\
&= \iint q(z)\delta(s_1-s_1')q(s_2|f_1,s_1,z)ds_1dz \\
&= \int q(z)q(s_2|f_1,s_1',z)dz.
\end{aligned}$$

However, what if, the intervention was not to cull seal to a predefined value (making S_1 have no parent nodes) but by a predefined value C? (i.e. $p(s_1|z) = q(s_1|z) - C$ which still depends on z, where C is a constant). In a general case, the adjustment formula becomes

$$\begin{aligned}
P(s_2|f_1,do(s_1\sim p(s_1|z))) &= \iint P(s_2,s_1,z|f_1,do(s_1\sim p(s_1|z)))ds_1dz \\
&= \iint q(z)p(s_1|z)q(s_2|f_1,s_1,z)ds_1dz.
\end{aligned}$$

Performing non-atomic intervention would be difficult in general, since we need 1) domain knowledge to decide which type of intervention $p(s_1|z)$ to use in order to gain as much information about the mechanism of the physical world as possible to better answer the counterfactual question; 2) to solve the non-trivial integral over s_1.

4 Counterfactual analysis with migration confounder

In Figure 5 we introduced a confounding variable z. However, we did not specify how this confounder affects s_1 and s_2. In fact, in order to implement a confounding effect we need to modify our data generating process. The new data generating process can be summarised as follows:

1. Sample the initial conditions: $f_{0,n}$, $s_{0,n}$, z_n;
2. Run the ODE simulator for X steps to obtain $f_{1,n}$ and $s_{1,n}$;
3. Run the ODE simulator for another Y steps to obtain $s_{2,n}$.

In Section 3 we mentioned two potential confounders: seasonal variation of prey-predator dynamics and a variation in animal population due to migration. We could introduce seasonal variation in seal and fish population evolution, by adjusting the dying and growth rates of animals based on a season. For example, we could assume seals to have a higher growth rate during the breeding season and fish could have a higher death rate during winter when seals need the extra energy. However, such a seasonal confounder would be a categorical variable and requires a different latent variable modeling technique to the one we used so far, since VAE can only deal with continuous latent variables. The paper [8] proposed an approach to dealing with discrete latent variable, but this is out of scope of our work.

4.1 Simulator with migration effect

The migration confounder offers a more straightforward way of altering seal and fish populations. An animal migration constant m can be added to the right hand side of the Lotka-Voltera ODEs:

$$\frac{df}{dt} = \gamma_1 f - \gamma_2 f s + m_f$$
$$\frac{ds}{dt} = \gamma_3 f s - \gamma_4 s + m_s.$$

In our example, we set $m_f = 0$ and consider only positive migration of seals m_s sampled from $p(m_s) = \mathcal{N}(m_s; 0.001, 0.0001^2)$. Figure 6 shows how the marginal distributions of s_1 and s_2 change as a function of the confounding variable m_s.

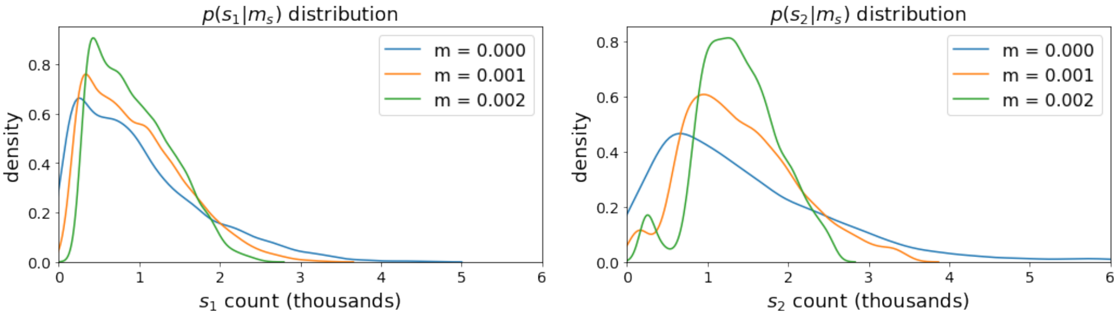


Figure 6: The effect of the migration confounder m_s on seal counts s_1 and s_2.

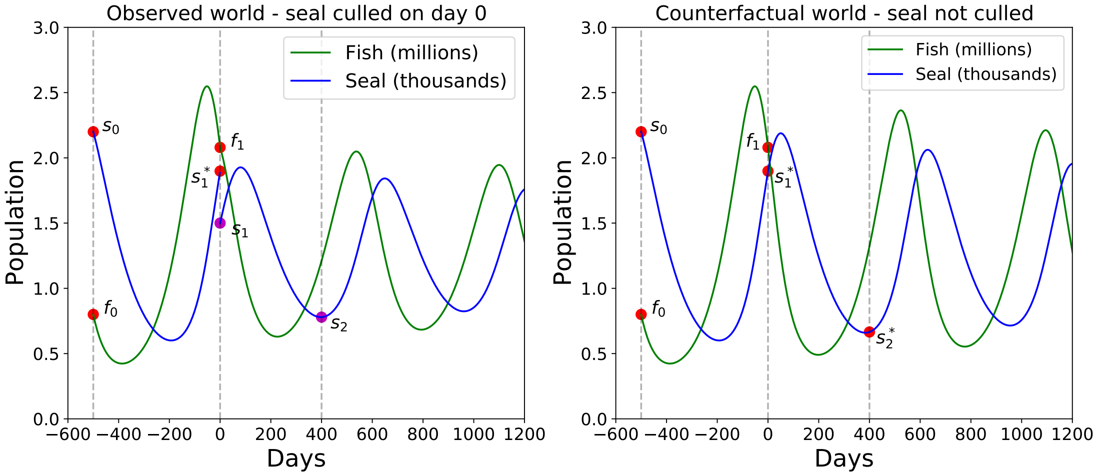


Figure 7: The ground truth seal and fish population in the observed and counterfactual worlds with the migration confounder included, given by the simulator.

For the inference problem, we assume that the initial conditions are $f_0 = 0.8$ and $s_0 = 2.2$ (on day -500) and set $m_s = 0.0005$. In the counterfactual world, we run the simulator for $X = 500$ days to obtain f_1 and s_1^* (on day 0), followed by another $Y = 400$ days to obtain the ground truth value of s_2^*. In the observed world, the fishermen culled the seal population to $s_1 = 1.5$ on day 0. Figure 7 shows the ground truth dynamics in both observed and counterfactual worlds with the migration confounder included. Note that culling the seal population on day 0 did not change the fish population f_1 but affected the dynamic of it after day 0 as indicated by the non-smoothness of the green curve on day 0.

4.2 Modelling and inference

To answer the counterfactual question, we follow the inference procedure for the posterior distribution of s_2^* presented in Section 3. The modified hierarchical BN is presented in Figure 8. There are a new pair of shared initial fish and seal populations f_0 and s_0, both of which are assumed to be standard log-normally distributed. s_1 is no longer log-normal distributed but generated from s_0. As a result, any changes to the ODE dynamics due to confounding will affect both s_1 and s_2. With s_2^* being conditionally independent of f_0 and s_0 given s_1^* and f_1, the presence of extra initial animal population variables f_0 and s_0 does not complicate the problem too much (given f_1 and s_1 are observed in the training set $\mathcal{D}$):

$$\begin{aligned}
& p(s_0, f_0, s_1, s_2, f_1, s_1^*, s_2^*, m_s, \mathcal{D}, \theta) \\
= \; & p(s_0, f_0, s_1, s_2, f_1, s_1^*, s_2^*, m_s|\theta)p(\mathcal{D}|\theta)p(\theta) \\
= \; & p(s_1)p(s_2|s_1, f_1, m_s, \theta)p(f_1|s_0, f_0, m_s, \theta)p(m_s)p(s_0)p(f_0)p(s_1^*|s_0, f_0, m_s, \theta)p(s_2^*|s_1^*, f_1, m_s, \theta)p(\mathcal{D}|\theta)p(\theta).
\end{aligned}$$

We assume that we have access to a dataset $\mathcal{D}$ with all variables of interest being observed, from which we need to learn the parameters for $p_\theta(f_1|f_0, s_0, m_s)$ and $p_\theta(s_2|s_1, f_1, m_s)$.

For inference, the only implication of the extended BN in Figure 8 is that we need to take expectation with respect to $p(f_1|s_0, f_0, m_s)$ rather than with respect to the prior $p(f_1)$. If all $s_{0,n}$, $f_{0,n}$, and $m_{s,n}$ are observed in $\mathcal{D}$, then we can again approximate $p(f_1|s_0, f_0, m_s)$ with a Gaussian with mean parameterized by a neural network. Then, to sample from

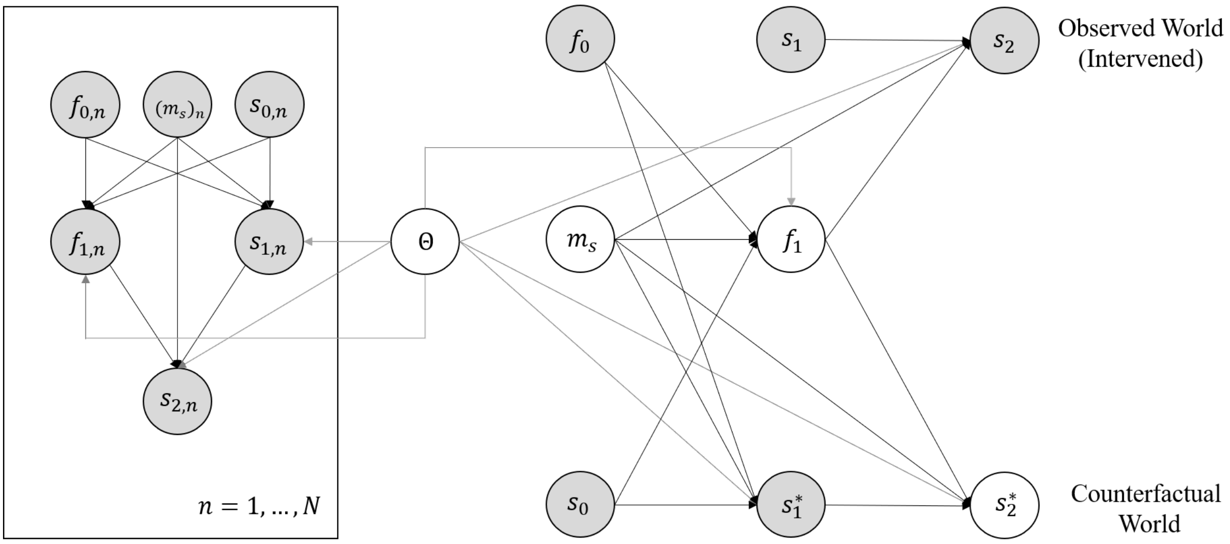


Figure 8: Bayesian Network from Figure 5 extended to include dependency of seal population on the migration confounder m_s.

$p(f_1|s_0, f_0, m_s)$, we need to first sample $m_s^{(v)}$ from $p(m_s)$ and then sample $f_1^{(l,v)}$ from $p(f_1|s_0, f_0, m_s^{(v)})$. The revised inference equation is

$$\begin{aligned}
& p(s_2^*|f_0, s_0, s_1, s_2, s_1^*, \mathcal{D}) \\
= & \iiint p(s_2^*, f_1, m_s, \theta|f_0, s_0, s_1, s_2, s_1^*, \mathcal{D}) df_1 dm_s d\theta \\
= & \iiint p(s_2^*|s_1^*, f_1, m_s, \theta) p(f_1|f_0, s_0, s_1, s_2, m_s, \theta) p(m_s|s_1, s_2, f_0, s_0, \theta) p(\theta|s_1, s_2, f_0, s_0, \mathcal{D}) df_1 dm_s d\theta \\
\approx & \iint p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*, f_1, z) p_{\theta_{\mathrm{ML}}}(f_1|f_0, s_0, s_1, s_2, m_s) p_{\theta_{\mathrm{ML}}}(m_s|s_1, s_2, f_0, s_0) df_1 dm_s \\
\propto & \iint p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*, f_1, m_s) p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1, m_s) p_{\theta_{\mathrm{ML}}}(f_1|f_0, s_0, m_s) p_{\theta_{\mathrm{ML}}}(s_2|s_1, m_s, f_0, s_0) p(m_s) df_1 dm_s \\
= & \mathbb{E}_{p(m_s)} \left\{ \mathbb{E}_{p(f_1|f_0, s_0, m_s)} \left[p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*, f_1, m_s) p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1, m_s) \right] \mathbb{E}_{p(f_1|f_0, s_0, m_s)} \left[p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1, m_s) \right] \right\}.
\end{aligned}$$

Note that we assumed f_0 to be an observed variable during inference. This simplification is rather unrealistic, since if we assume f_1 to be uncountable during inference, we should also assume the same for f_0. If f_0 was latent, however, then during inference we would need to sample from $p(f_1|s_0, m_s)$ rather then $p(f_1|s_0, f_0, m_s)$, for which we would have to marginalise out f_0, leading to yet another layer of computation, which makes the inference problem complicated. Hence, for simplicity we assume f_0 to be known.

To approximate the posterior distribution of s_2^* we apply MC sampling, following the same principles as in Section 2:

$$
\begin{aligned}
&p(s_2^*|f_0, s_0, s_1, s_2, s_1^*, \mathcal{D}) \\
\propto\ & \mathbb{E}_{p(m_s)} \left\{ \mathbb{E}_{p(f_1|f_0,s_0,m_s)} \left[p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*, f_1, m_s) p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1, m_s) \right] \mathbb{E}_{p(f_1|f_0,s_0,m_s)} \left[p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1, m_s) \right] \right\} \\
\approx\ & \frac{1}{V} \sum_{v=1}^{V} \left[\frac{1}{L} \sum_{l=1}^{L} \left(p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*, f_1^{(l,v)}, m_s^{(v)}) p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1^{(l,v)}, m_s^{(v)}) \right) \frac{1}{L} \sum_{l=1}^{L} p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1^{(l,v)}, m_s^{(v)}) \right], \\
& \left(m_s^{(v)} \sim p(m_s), \quad f_1^{(l,v)} \sim p_{\theta_{\mathrm{ML}}}(f_1|f_0, s_0, m_s^{(v)}) \right) \\
\propto\ & \frac{1}{V} \sum_{v=1}^{V} \left[\left(\frac{1}{L} \sum_{l=1}^{L} w^{(l,v)} p_{\theta_{\mathrm{ML}}}(s_2^*|s_1^*, f_1^{(l,v)}, m_s^{(v)}) \right) \left(\frac{1}{L} \sum_{l=1}^{L} w^{(l,v)} \right) \right], \\
& \left(w^{(l,v)} := p_{\theta_{\mathrm{ML}}}(s_2|s_1, f_1^{(l,v)}, m_s^{(v)}) \right).
\end{aligned}
$$

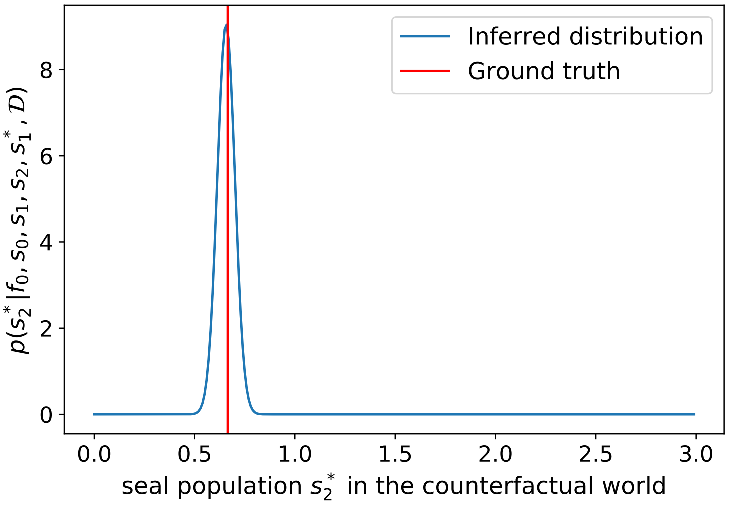


Figure 9: Inferred posterior distributions of s_2^*.

Figure 9 shows the inferred posterior distribution $p(s_2^*|f_0, s_0, s_1, s_2, s_1^*, \mathcal{D})$. The MAP estimate for s_2^* from the posterior distribution is 0.660, which is very close to the ground truth value of 0.666. The posterior also seems relatively narrow, indicating relatively low uncertainty in this value. This accurate inference shows us that we can answer counterfactual questions with the use of probabilistic modelling, even in the presence of confounders. Nevertheless, we acknowledge that the quality of causal inference may depend on the strength of the confounding effect. A potential extension to our work could include analysis of how our ability to answer counterfactual questions changes as we vary the prior distribution of migration $p(m_s)$.

5 Conclusions

In this project, we investigated the use of probabilistic modelling for causal inference, and asking counterfactual questions. We started by replicating a simple non-confounded example from [4], involving prey-predator populations in the context of seals and fish. The problem was to infer the posterior distribution over the counterfactual population of the seals had a cull that took place in the observed world not taken place. We then extended this to another example from [4] in which the latent variable of the initial fish population was not observed in the training instances. This made the problem more realistic, as it was slightly unusual that in the initial example the fish population was not observable in the test scenario, but we had access to many observed training instances. Because of this lack of fully observed training instances, we had to make use of a VAE to model the latent fish population.

Whilst this problem was phrased as a counterfactual problem, we noticed that it is in fact a classical probabilistic machine learning problem, and that it could be answered with no particular attention payed to confounding. In order to investigate more sofisticated counterfactual reasoning, we extended the original problem to one in which there was an additional confounding variable of migration, as well as correlation between the inputs. We showed that similar methods as were used in the earlier examples could be applied here, as long as careful consideration was payed to the causal mechanisms involved.

6 Further work: extended prey-predator simulator

We can devise a straightforward way to extend the Lotka-Voltera simulator to include an arbitrary number of predators and prey animals as follows. The population of animal a_i is dependent on animal a_j if $\gamma_{i,j} \neq 0$. If $\gamma_{i,j} > 0$, then a_j is a prey of a_i; if $\gamma_{i,j} < 0$, then a_j is a predator of a_i:

$$\begin{aligned}
\frac{da_1}{dt} &= \gamma_{1,1}a_1 + \gamma_{1,2}a_1a_2 + \cdots + \gamma_{1,n}a_1a_n \\
\frac{da_2}{dt} &= \gamma_{2,2}a_2 + \gamma_{2,1}a_2a_1 + \cdots + \gamma_{2,n}a_2a_n \\
&\vdots \\
\frac{da_n}{dt} &= \gamma_{n,n}a_n + \gamma_{n,1}a_na_1 + \cdots + \gamma_{n,n-1}a_na_{n-1}.
\end{aligned}$$

This raises interesting research questions, such as "How does our ability to answer counterfactual questions, in the prey-predator problem, change as more (potentially latent) variables are included?".

References

[1] N. Banholzer, E. van Weenen, B. Kratzwald, A. Seeliger, D. Tschernutter, P. Bottrighi, A. Cenedese, J. Puig Salles, W. Vach, S. Feuerriegel, *Impact of non-pharmaceutical interventions on documented cases of COVID-19*, medRxiv 2020.04.16.20062141v3 [Preprint], 28 April 2020.

[2] J.M. Brauner, S. Mindermann, M. Sharma, D. Johnston, J. Salvatier, T. Gavenčiak, A.B. Stephenson, G. Leech, G. Altman, V. Mikulik, A.J. Norman, J.T. Monrad, T. Besiroglu, H. Ge, M.A. Hartwick, Y.W. Teh, L. Chindelevitch, Y. Gal, J. Kulveit, *Inferring the effectiveness of government interventions against COVID-19. Science*, Science, https://science.sciencemag.org/content/early/2020/12/15/science.abd9338, Dec. 2020.

[3] S. Flaxman, S. Mishra, A. Gandy, H.J.T. Unwin, T.A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J.W. Eaton, M. Monod, A.C. Ghani, C.A. Donnelly, S. Riley, M.A.C. Vollmer, N.M. Ferguson, L.C. Okell, S. Bhatt, *Estimating the effects of non-pharmaceutical interventions on COVID19 in Europe*, Nature 584, 257–261, doi:10.1038/s41586-020-2405-7 Medline, 2020.

[4] B.K. Mlodozeniec, *Causal Inference, Probabilistic Modelling and Bayesian Inference*, May 2020.

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

[6] J. Pearl, D. Mackenzie, *The book of why: the new science of cause and effect*, Basic Books, 2018.

[7] J. Peters, D. Janzing, B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*, MIT press, 2017.

[8] J.T. Rolfe, *Discrete variational autoencoders*, arXiv preprint arXiv:1609.02200, 2016.