Bivariate Causal Discovery using Bayesian Model Selection

Anish Dhir [1] **Samuel Power** [2] **Mark van der Wilk** [3]

Abstract

Much of the causal discovery literature prioritises guaranteeing the identifiability of causal direction in statistical models. For structures within a Markov equivalence class, this requires strong assumptions which may not hold in real-world datasets, ultimately limiting the usability of these methods. Building on previous attempts, we show how to incorporate causal assumptions within the Bayesian framework. Identifying causal direction then becomes a Bayesian model selection problem. This enables us to construct models with realistic assumptions, and consequently allows for the differentiation between Markov equivalent causal structures. We analyse why Bayesian model selection works in situations where methods based on maximum likelihood fail. To demonstrate our approach, we construct a Bayesian non-parametric model that can flexibly model the joint distribution. We then outperform previous methods on a wide range of benchmark datasets with varying data generating assumptions.

1. Introduction

Many fields use statistical models to predict the response to actions (interventions), e.g. health outcomes after treatment. Such predictions can not be made based on correlations gained from purely observational data, but require access to causal structure (Pearl, 2009). In machine learning, causal structure also helps in many prediction tasks ranging from domain adaptation (Wang and Veitch), robustness (Bühlmann, 2020), and generalisation (Scherrer et al., 2022). Conditional independencies in data can be used to infer causal structures, but only up to a Markov equivalence class (MEC) (app. A) (Pearl, 2009). Further assumptions are required to completely recover a causal structure. In this paper, we aim to identify causal direction within a Markov equivalence class. Hence, for simplicity, we focus solely on distinguishing $\mathbf{X} \to Y$ and $X \leftarrow \mathbf{Y}$[1].

In the bivariate case, much of the causal literature assumes restrictions on a model, which when they match reality, guarantee identifiabilility of the causal direction. For example, for data generated by an additive noise model (ANM), an ANM model achieves a higher likelihood in the causal direction, as the anti-causal direction violates the ANM assumptions (Hoyer et al., 2008; Zhang et al., 2015). A method's overall usefulness is then empirically verified on a wider array of data generating distributions, including datasets where the rigid restrictions of models such as ANMs hamper their performance, as the models are misspecified. Moreover, in these settings, causal identifiability is no longer guaranteed. The goal of our work is to investigate whether causality can be identified in models without hard restrictions, reducing misspecification, even if we lose *strict* guarantees of identifiability.

We replace hard restrictions on models with softer ones encoded by Bayesian priors. Each causal direction is treated as a separate Bayesian model, after which causal discovery can be formulated as Bayesian model selection. The causal models encode the *independent causal mechanisms* (ICM) assumption in their respective causal directions. We show that if the anti-causal factorisation violates the ICM assumption, Bayesian model selection can discern the causal direction. This is true even for model classes whose flexibility prevents likelihood based discrimination. We then analyse the probability of error of the method, both when the model is correct and when it is misspecified. Our work shows that when faced with datasets for which no restricting assumptions can reasonably be made, our framework allows for use of an appropriate model that can better enable identification of causal direction, relative to a more restricted model with inappropriate assumptions.

To demonstrate the usefulness of our approach, we use a Gaussian process latent variable model (GPLVM) (Titsias and Lawrence, 2010), which has the ability to model a wide range of densities. We test this on a range of benchmark datasets with various data generating assumptions. We also compare against previously proposed methods, both those

[1]Imperial College London [2]University of Bristol [3]University of Oxford. Correspondence to: Anish Dhir <anish.dhir13@imperial.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]For clarity, we colour the cause blue throughout.

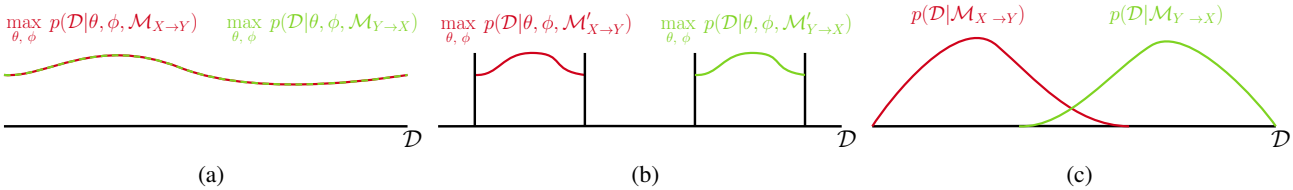


Figure 1: Toy figure with datasets on the x axis and values of densities on the y axis (a) With a sufficiently flexible model, maximising the likelihood for each dataset will give the same value for both causal models in Section 2.3. (b) This has been solved by making restrictions on the datasets they can model. (c) Bayesian model selection retains the ability to identify causal direction, while allowing flexibility. This may lead to some probability of error (overlap).

which rely on strict restrictions, and those which are more flexible, but lack formal identifiability guarantees. Whereas most methods perform well on the types of datasets where their assumptions hold, we find that our method performs well across all datasets. Our findings show that causal discovery is possible without losing the ability to model datasets well, a property that is desirable in real world cases.

2. Preliminaries & Assumptions

In this paper, we focus on the problem of bivariate causal discovery, under the assumption that there are no hidden confounders. We first introduce notation.

Throughout, given a space $\mathcal{Z}$, we write $\mathcal{P}(\mathcal{Z})$ for the space of probability measures over $\mathcal{Z}$. Similarly, given a pair of spaces $\mathcal{Z}, \mathcal{Z}'$, we write $\mathcal{K}(\mathcal{Z}, \mathcal{Z}')$ for the space of Markov kernels from $\mathcal{Z}$ to $\mathcal{Z}'$. In all models considered, individual bivariate observations take the form $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y}$ are (potentially distinct) spaces. We will view (x, y) as realisations of random variables, using capital letters (e.g. P) to denote the associated probability measure in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ (which we will refer to as a 'data-generating process'), and lower-case letters (e.g. p) to denote the associated probability density with respect to a suitable reference measure. Our 'datasets' will then be modelled as $N \in \mathbf{N}$ iid observations from this probability measure, which we abbreviate as $\mathcal{D}^N = \left(\mathbf{x}^N, \mathbf{y}^N\right) = \{(x_i, y_i) : i \in [N]\}$. Our assumption is that our data-generating process admits a causal interpretation, whereby either i) X causes Y ($\mathbf{X} \to Y$), or ii) Y causes X ($X \leftarrow \mathbf{Y}$). Our goal is to then determine which of these causal directions underlies the 'true' data-generating process, based on our observed dataset.

2.1. Data Generation through Structural Causal Models

We can express causal relationships by a Structural Causal Model (SCM), which describes the hierarchical ordering of variable generation from causes to effects (Pearl, 2009). Each SCM can be represented as a *Directed Acyclic Graph* (DAG) with a vertex for each variable. We denote the parents of a vertex j in a DAG $\mathcal{G}$ as $\mathrm{pa}_{\mathcal{G}}(j)$. In our bivariate setting, the causal direction $\mathbf{X} \to Y$ corresponds to a data-generating process of the form

$$X_i := f_X\left(N_i^X\right), \quad Y_i := f_Y\left(X_i, N_i^Y\right) \tag{1}$$

for $i = 1, \ldots, N$, where f_X and f_Y are deterministic 'generation functions', and $\left\{N_i^X : i \in [N]\right\} \overset{\text{iid}}{\sim} \nu^X$, $\left\{N_i^Y : i \in [N]\right\} \overset{\text{iid}}{\sim} \nu^Y$ are iid realisations of some mutually independent 'noise' random variables. This naturally induces a factorisation $P_{X,Y}(\mathrm{d}(x, y)) = P_X(\mathrm{d}x) \cdot P_{Y|X}(\mathrm{d}y \mid x)$, which we term the 'causal factorisation' of this SCM. By disintegration of measure, one can equally factorise $P_{X,Y}(\mathrm{d}(x, y)) = P_Y(\mathrm{d}y) \cdot P_{X|Y}(\mathrm{d}x \mid y)$, which we term the 'anti-causal factorisation' of the same SCM. The SCM for $X \leftarrow \mathbf{Y}$ is analogously defined with the causal factorisation being $P_Y(\mathrm{d}y) \cdot P_{X|Y}(\mathrm{d}x \mid y)$.

2.2. Interventions and Independent Causal Mechanisms (ICM)

An intervention on a variable is an action which alters its value, generation function, or noise input, while leaving those of all other variables unchanged. In the causal factorisation, an intervention on a cause means that only the distribution of the cause is changed, while the conditional distribution of effect given cause remains unchanged. For example, if our model assumes the causal direction $\mathbf{X} \to Y$, then an intervention on X can be achieved by changing either f_X or N_i^X in Equation (1). Such an action changes only changes one term in the causal factorisation $P_X(\mathrm{d}x)$, but will leave $P_{Y|X}(\mathrm{d}y \mid x)$ invariant. By contrast, the same action will often induce a change in both terms of the anti-causal factorisation, $P_Y(\mathrm{d}y)$ and $P_{X|Y}(\mathrm{d}x \mid y)$. Given this invariance, we thus say that the causal factorisation satisfies the Independent Causal Mechanisms (ICM) assumption (Peters et al., 2017, ch. 2). The ICM assumption implies a fundamental asymmetry for the impact of interventions on different factorisations of the same joint distribution.

2.3. Causal Models

A causal model which encodes the ICM assumption should directly specify the causal factorisation. Specifying terms

of the causal factorisation ensures that parts of the model remain well specified under interventions. Conversely, a model which specifies the anti-causal factorisation will often find both of its components to be mismatched after an intervention (Schölkopf et al., 2012). We define a causal model in line with the ICM assumption.

Definition 2.1. A **causal model** is a tuple $\mathcal{M}_{\mathcal{G}} = (\mathcal{G}, \mathcal{C}, \mathcal{F})$, where $\mathcal{G}$ is a DAG with vertex set $\mathcal{V}$, and $\mathcal{C}$ is a set of conditional distributions that specifies the causal factorisation

$$\mathcal{C} = \prod_{i \in \mathcal{V}} \mathcal{C}_{i|\mathrm{pa}_{\mathcal{G}}(i)} \subset \prod_{i \in \mathcal{V}} \mathcal{K}(\mathcal{X}_{\mathrm{pa}_{\mathcal{G}}(i)} \to \mathcal{X}_i). \tag{2}$$

Given a tuple of conditionals, one for each variable, $P = (P_i : i \in \mathcal{V}) \in \mathcal{C}$, define $\delta_{\mathcal{C}} : \mathcal{C} \to \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ as the map that assembles P into the corresponding joint

$$\delta_{\mathcal{C}}(P)(\mathrm{d}x_{\mathcal{V}}) = \prod_{i \in \mathcal{V}} P_i(\mathrm{d}x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)}). \tag{3}$$

Finally, define $\mathcal{F}$ as the set of induced joint distributions $\mathcal{F} = \{\delta_{\mathcal{C}}(P) : P \in \mathcal{C}\}$.

For example, $\mathcal{M}_{\mathbf{X}\to Y}$ specifies $\mathcal{C} = \mathcal{C}_X \times \mathcal{C}_{Y|X}$, with an induced joint $\delta_{\mathcal{C}}(P_X, P_{Y|X})(\mathrm{d}x, \mathrm{d}y) = P_X(\mathrm{d}x)P_{Y|X}(\mathrm{d}y)$ for $P_X \in \mathcal{C}_X, P_{Y|X} \in \mathcal{C}_{Y|X}$. Note that the mapping $\delta_{\mathcal{C}}$ is injective. Throughout, we work directly with distributions for generality. In practice, the elements of $\mathcal{C}$ are constructed as a parameterised family (though the parameter may in principle be infinite-dimensional). When relevant, we denote the parameterisation by parameter spaces Φ and Θ for X and Y respectively, as in Figure 2. To emphasise the direction of the causal model, we will often append $\mathcal{M}_{\mathbf{X}\to Y}$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ to distributions and densities of interest.

We are interested in the case where the causal structure is *not* known in advance and we seek to determine it from data. In the case that both $\mathcal{F}_{\mathbf{X}\to Y}$ and $\mathcal{F}_{X\leftarrow\mathbf{Y}}$ are sufficiently expressive such that there are few restrictions on which joint distributions on $\mathcal{X} \times \mathcal{Y}$ can be learned (in principle), then model selection based on maximum likelihood will fail to identify causal direction, as it assigns equal scores to the 'best' model compatible with each causal direction (Zhang et al., 2015). The notion of distribution-equivalence (Geiger and Heckerman, 2002) helps to capture this principle.

Definition 2.2. Two causal models $\mathcal{M}_{\mathbf{X}\to Y} = (\mathbf{X} \to Y, \mathcal{C}_{\mathbf{X}\to Y}, \mathcal{F}_{\mathbf{X}\to Y})$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}} = (X \leftarrow \mathbf{Y}, \mathcal{C}_{X\leftarrow\mathbf{Y}}, \mathcal{F}_{X\leftarrow\mathbf{Y}})$ are **distribution-equivalent** if $\mathcal{F}_{\mathbf{X}\to Y} = \mathcal{F}_{X\leftarrow\mathbf{Y}}$. Equivalently, there exists a unique translating bijection $\gamma : \mathcal{C}_{\mathbf{X}\to Y} \to \mathcal{C}_{X\leftarrow\mathbf{Y}}$ such that for any $P \in \mathcal{C}_{\mathbf{X}\to Y}$, there holds an equality of (joint) measures $\delta_{\mathcal{C}_{\mathbf{X}\to Y}}(P) = \delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}}(\gamma(P))$.

In short, for every $(m, c) \in \mathcal{C}_X \times \mathcal{C}_{Y|X}$, there exists $(m', c') \in \mathcal{C}_Y \times \mathcal{C}_{X|Y}$ such that $m(\mathrm{d}x) \cdot c(\mathrm{d}y \mid x) = m'(\mathrm{d}y) \cdot c'(\mathrm{d}x \mid y)$ as probability measures, and vice versa.

A solution to restore identifiability is to restrict $\mathcal{C}_{\mathbf{X}\to Y}, \mathcal{C}_{X\leftarrow\mathbf{Y}}$, but this comes at the cost of not being able to learn some distributions (Figure 1 (a,b), app. A.2), i.e. a loss in modelling flexibility. Our aim in the following sections is to allow for discovering causal relations, even with distribution-equivalent models.

3. Related Work

One class of methods makes hard restrictions on the set of distributions ($\mathcal{C}$) which induce the causal factorisation. For example, linear function with non-Gaussian noise (LiNGAM) (Shimizu et al., 2006), non-linear functions with additive noise (ANM) (Hoyer et al., 2008), and post non-linear models (PNL) (Zhang and Hyvarinen, 2012), among others (Immer et al., 2022). These restrictions can crudely be thought of as controlling the complexity of $\mathcal{F}$. Identifiability is proven by showing that the more complex anti-causal factorisation lies outside of the core model. Zhang et al. (2015) showed that the likelihood can be used to identify the causal direction in these models (Figure 1(b)), but if the dataset is generated by a model without these restrictions, it is possible for both causal directions to achieve similar likelihoods (Zhang et al., 2015).

Another class of methods assumes more flexibility but try and control a measure of complexity. Marx and Vreeken (2019) (SLOPPY) build on previous non-identifiable methods (RECI (Blöbaum et al., 2018), QCCD (Tagasovska et al., 2020)) and assume that the causal factorisation has been generated by a model with fewer parameters than the anti-causal factorisation. Balancing mean squared error along with the number of parameters can then identify the causal direction. However, measuring complexity by the number of parameters can be parametrisation-dependent. Such conceptual problems are amplified in the setting of non-parametric or over-parametrised models. Additionally, SLOPPY also assumes low noise and additive noise. CGNN (Goudet et al., 2018) forego strict identifiability and try to learn a causal generative model of the data using neural networks. Their methods works with small networks but with no clear way of mitigating overfitting, their method easily achieves the same score for both causal factorisations.

Other methods either try to i) directly measure the dependence of the factorisation, based on the ICM principle, or ii) measure the complexity of a proposed direction. CURE (Sgouritsa et al., 2015) and IGCI (Daniusis et al., 2012) try to measure the dependence of the factorisations. CDCI (Duong and Nguyen, 2021), CDS (Fonollosa, 2019) and KCDC (Mitrovic et al., 2018) try to measure the stability of the conditional distributions under different input values, arguing that the more stable conditional is more likely the cause. Of the above, only IGCI has proven identifiability.

We base our approach on Bayesian model selection, which has automatic mechanisms of balancing model fit and complexity. The prescribed procedure is straightforward (Section 4.1), and was first investigated in the 90s in the context of finding Bayesian network structure (Heckerman et al., 1995; 2006; Heckerman, 1995; Chickering, 2002; Geiger and Heckerman, 2002). However, while we argue that Bayesian model selection is helpful for causal discovery *within* a Markov Equivalence Class (MEC), these early papers restricted their focus to finding network structure up to an MEC. This is due to a focus on linear causal relationships, which is one key setting where Bayesian model selection does not provide much benefit (see app. D.1 for a discussion). Indeed, the wider benefits of Bayesian methods to infer causality within an MEC has been touched upon in MacKay (2003, ch. 35). Friedman and Nachman (2000) are first to apply Bayesian model selection to do bivariate causal discovery (i.e. within an MEC). They compare two Gaussian process regression models, and attempt to determine which variable should be used as the input (cause). However, since this model was effectively an ANM, Bayesian model selection provided little added value, since causal direction is already identifiable by the likelihood alone (Zhang et al., 2015) (Section 2.3). A similar approach was used by Kurthen and Enßlin (2019). Stegle et al. (2010) highlighted these issues by noting that Friedman and Nachman (2000) worked only due to model fit. Like us, they acknowledge the larger benefit when $\mathcal{C}$ is not restricted. While their method is similar to Bayesian model selection, it is heuristically justified by Kolmogorov complexity (Janzing and Steudel, 2010) (see app. B); as such, the impact of both model and prior on their procedure is unclear. This is explicit in our work (Section 4.2, Section 4.3).

Our contributions follow in the path of Friedman and Nachman (2000) and Stegle et al. (2010). Specifically, we provide a more complete view of why and under what conditions Bayesian model selection can identify causal direction (Section 4.2). Our work gives insight into the performance when a chosen model correctly encodes the assumptions on a dataset (Section 4.3), and when it does not (Section 4.4).

4. Bayesian Inference of Causal Direction

We explain how causal discovery can be viewed as a Bayesian model selection problem, outlining the requisite assumptions. We then give conditions under which Bayesian model selection discriminates even distribution-equivalent causal models. Correctness then depends on the exact assumptions made in the model, and how well they match reality. We analyse the case when the assumptions are correct, providing a statistical test that can be used to quantify the probability of error inherent in the procedure. We also provide analysis when the assumptions might be wrong.

4.1. Bayesian Model Selection for Causal Inference from First Principles

The maximum likelihood approach outlined in Section 2.3 was restricted in its ability to simultaneously i) estimate model parameters and ii) infer causal direction. The core issue is that for richly-parameterised models, both causal models can express distributions which explain the observed data equally well. Bayesian inference provides a general framework for inferring unknown quantities in statistical models (Gelman et al., 2013; MacKay, 2003; Bernardo and Smith, 2009) which offers solutions to this problem. In this section, we will describe how Bayesian inference allows us to directly infer our quantity of interest — the causal direction — and the additional assumptions which need to be specified in order for this approach to succeed.

Inferring causal direction can be seen as a *Bayesian model selection* problem, where we seek to distinguish which causal model (Section 2.3) is more likely to have generated a dataset. The evidence for each causal direction is quantified in the posterior

$$P(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})P(\mathcal{M}_{\mathbf{X}\to Y})}{P(\mathcal{D})}. \quad (4)$$

We can summarise the balance of evidence for both causal directions with the log ratio:

$$\log \frac{P(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D})}{P(\mathcal{M}_{X\leftarrow \mathbf{Y}}|\mathcal{D})} = \log \frac{P(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})P(\mathcal{M}_{\mathbf{X}\to Y})}{P(\mathcal{D}|\mathcal{M}_{X\leftarrow \mathbf{Y}})P(\mathcal{M}_{X\leftarrow \mathbf{Y}})}. \quad (5)$$

Bayesian inference requires us to specify a *prior* on which causal direction is more likely. To represent our lack of specific knowledge, we set these prior probabilities to be equal $P(\mathcal{M}_{\mathbf{X}\to Y}) = P(\mathcal{M}_{X\leftarrow \mathbf{Y}}) = 0.5$. Thus, the above log posterior ratio will determined only by the ratio $P(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})/P(\mathcal{D}|\mathcal{M}_{X\leftarrow \mathbf{Y}})$.

To find the ratio in Equation (5), we are required again to specify prior measures, this time on $\mathcal{C}$. We denote prior over distributions by π (as opposed to distributions over observations). This augments a causal model to form a Bayesian Causal Model (BCM).

Definition 4.1. A **Bayesian causal model** (BCM) is a causal model $\mathcal{M}_{\mathcal{G}}$ equipped with a prior distribution over $\mathcal{C}$, that is a tuple $(\mathcal{G}, \mathcal{C}, \mathcal{F}, \pi)$, shortened to $(\mathcal{M}_{\mathcal{G}}, \pi)$.

While we will discuss the consequences of selecting a particular prior later, we can now already determine some of its properties from our problem setting. A strict view of the ICM assumption implies that information about the distribution on causes should not provide information on the distribution of effect given cause. This implies that the prior should be separable, with respect to the causal factorisation, for both causal models.[2]. We formally define separability of

[2]Guo et al. (2022) provide the most direct argument for this,

priors as follows.

Definition 4.2. Given a BCM $(\mathcal{M}_{\mathcal{G}}, \pi)$, we call prior π **separable** with respect to $\mathcal{C}$ if it factorises as $\prod_{i\in\mathcal{V}} \pi_i$ for some $(\pi_i : i \in \mathcal{V}) \in \prod_{i\in\mathcal{V}} \mathcal{P}(\mathcal{C}_{i|\mathrm{pa}_{\mathcal{G}}(i)})$.

For $\mathcal{M}_{\mathbf{X}\to Y}$, this amounts to saying that the prior factorises as $\pi_X(P_X)\pi_Y(P_{Y|X})$ with $P_X \in \mathcal{C}_X$ and $P_{Y|X} \in \mathcal{C}_{Y|X}$. For parametrised models, our BCMs can be represented as the graphical models in Figure 2.

Given a prior, any BCM $(\mathcal{M}_{\mathcal{G}}, \pi)$ naturally induces *data distribution*, i.e. a probability measure over sequences of bivariate observations, $P(\cdot \mid \mathcal{M})$. This is found by integrating the joint over the prior distribution.

Definition 4.3. The **data distribution** of $(\mathcal{M}_{\mathcal{G}}, \pi)$ is

$$P(\mathrm{d}x, \mathrm{d}y|\mathcal{M}_{\mathcal{G}}) = \int_{P'\in\mathcal{C}} \delta_{\mathcal{C}}(P')(\mathrm{d}x, \mathrm{d}y|\mathcal{M}_{\mathcal{G}})\pi(\mathrm{d}P'|\mathcal{M}_{\mathcal{G}}).$$

The density of this measure with respect to a suitable reference measure is known as the **marginal likelihood**.

Given a dataset, and under the above procedure, the most likely model $\mathcal{M}^*$ can be chosen as the one with the higher marginal likelihood

$$\mathcal{M}^* = \begin{cases} \mathcal{M}_{\mathbf{X}\to Y} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) > p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \\ \mathcal{M}_{X\leftarrow\mathbf{Y}} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) < p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \end{cases} . \quad (6)$$

This is straightforward Bayesian model selection, but applied to models that incorporate the causal assumptions of Section 2.3. Following this procedure from first principles, we see that Bayesian inference *prescribes* that we must specify priors π. In specifying these priors, one implicitly constrains the sets distributions which will be used to explain the observed data. This is consistent with earlier work, where hard constraints on these sets are imposed to ensure identifiability (Zhang et al., 2015). While these hard constraints can be represented in the priors by only assigning non-zero mass to the desired regions, priors also allow the specification of softer constraints (Figure 1(c)).

While Bayesian inference prescribes the procedure which our method will follow, it is not clear this will produce correct answers. The performance of all Bayesian methods, including ours, depends on how compatible our prior assumptions are with reality. This is also the case for previous causal discovery methods that provide identifiability guarantees, since the assumptions made have to hold in reality. In

by proving a "causal de Finetti" theorem, based on the requirement that additional information on the cause mechanism does not give information on the effect mechanism. Earlier, Janzing and Schölkopf (2010) also argue that this must be the case, but from a more heuristic argument based on Kolmogorov complexity. The assumption has also been made in earlier methods (Stegle et al., 2010; Sgouritsa et al., 2015; Heckerman, 1995).

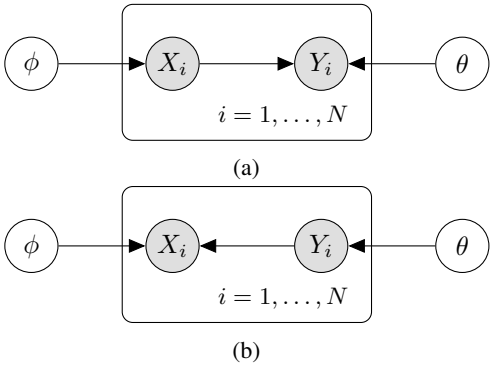


Figure 2: Graphical models for parametrised Bayesian causal models $\mathcal{M}_{\mathbf{X}\to Y}$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}}$. The causal direction indicates the factorisation that encodes ICM.

the next sections, we investigate the influence of prior design on the performance of Bayesian model selection. Specifically, we will analyse conditions under which choices of priors cannot imply the same data distributions for the two BCMs, the guarantees we can get when the assumptions made are correct, and when they are incorrect.

4.2. Asymmetry Between Dataset Densities of Bayesian Causal Models

In Section 2.3 we discussed that the maximum likelihood score is indifferent to causal direction when the causal models are distribution-equivalent. In general, we can expect this to cause difficulties when the distributions specified ($\mathcal{C}$) are sufficiently flexible. By contrast, Bayesian inferences prescribes using the *marginal likelihood* to guide model selection. Here, we show that Bayesian model selection can be sensitive to causal direction, even when using distribution-equivalent causal models. This shows that the marginal likelihood is capable of providing a preference between factorisations in situations where maximum likelihood cannot. While this result does not, by itself, imply that Bayesian model selection will identify the correct causal direction (which we discuss later), it does offer insight into the difference between the two approaches. All proofs are in app. C.

We will find a necessary condition under which marginal likelihood cannot discriminate BCMs. If the condition does not hold, the BCMs can surely be discerned. We thus define a notion of equivalence, similar to Definition 2.2, but for BCMs and marginal likelihoods.

Definition 4.4. Given two BCMs $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$, $(\mathcal{M}_{X\leftarrow\mathbf{Y}}, \pi_{X\leftarrow\mathbf{Y}})$, say they are **Bayesian distribution-equivalent** if $P(\cdot \mid \mathcal{M}_{\mathbf{X}\to Y}) = P(\cdot \mid \mathcal{M}_{X\leftarrow\mathbf{Y}})$, i.e. for all $N \in \mathbf{N}$, and for all $(\mathbf{x}^N, \mathbf{y}^N) \in (\mathcal{X}\times\mathcal{Y})^N$, it holds that $p(\mathbf{x}^N, \mathbf{y}^N \mid \mathcal{M}_{\mathbf{X}\to Y}) = p(\mathbf{x}^N, \mathbf{y}^N \mid \mathcal{M}_{X\leftarrow\mathbf{Y}})$.

In the following, we show that if two causal models are

not distribution-equivalent, then there exist Bayesian causal models that are not Bayesian-distribution equivalent. This demonstrates that if maximum likelihood can distinguish causal models, then suitably constructed Bayesian causal models can be differentiated using the marginal likelihood.

Proposition 4.5. *Given two BCMs* $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$, $(\mathcal{M}_{X\leftarrow \mathbf{Y}}, \pi_{X\leftarrow \mathbf{Y}})$, *suppose that there exists a subset* $\mathcal{C}_\Delta \subset \mathcal{C}_{\mathbf{X}\to Y}$ *such that* $\pi_{\mathbf{X}\to Y}(\mathcal{C}_\Delta) > 0$, *and* $\delta_{\mathcal{C}_{\mathbf{X}\to Y}}(\mathcal{C}_\Delta) \cap \mathcal{F}_{X\leftarrow \mathbf{Y}}$ *is empty. Then the two Bayesian causal models are not Bayesian distribution-equivalent.*

We are ultimately interested in the case where the underlying causal models are distribution-equivalent. To help with this, we introduce *separable-compatibility*.

Definition 4.6. Let $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$, $(\mathcal{M}_{X\leftarrow \mathbf{Y}}, \pi_{X\leftarrow \mathbf{Y}})$ be two BCMs where the underlying causal models are distribution-equivalent, denoting γ as the corresponding translation mapping $\gamma : \mathcal{C}_{\mathbf{X}\to Y} \to \mathcal{C}_{X\leftarrow \mathbf{Y}}$ (in Definition 2.2). Say the two are **separable-compatible** if: i) the push-forward $\gamma\#\pi_{\mathbf{X}\to Y}$ is separable with respect to $\mathcal{C}_{X\leftarrow \mathbf{Y}}$, ii) $\gamma^{-1}\#\pi_{X\leftarrow \mathbf{Y}}$ is separable with respect to $\mathcal{C}_{\mathbf{X}\to Y}$.

Under the above definition, we see that the ICM assumption is preserved in the anti-causal factorization. We now show that if two Bayesian causal models are not separable-compatible, they cannot be Bayesian distribution-equivalent.

Proposition 4.7. *Given two BCMs* $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$, $(\mathcal{M}_{X\leftarrow \mathbf{Y}}, \pi_{X\leftarrow \mathbf{Y}})$, *where the underlying causal models are distribution-equivalent, the two Bayesian causal models are Bayesian distribution-equivalent only if they are separable-compatible.*

For parametrised models, we require that the parametrisation is injective for the same result to hold. Being separable-compatible then implies that each BCM can factorise as both graphical models in Figure 2.

Corollary 4.8. *Given two injectively parametrised BCMs,* $\mathcal{B}^1 := (\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$ *that factorises as Figure 2(a), and* $\mathcal{B}^2 := (\mathcal{M}_{X\leftarrow \mathbf{Y}}, \pi_{X\leftarrow \mathbf{Y}})$ *that factorises as Figure 2(b), assume the underlying causal models are distribution-equivalent. The two Bayesian causal models are Bayesian distribution-equivalent only if: i)* $\mathcal{B}^1$ *also factorises as Figure 2 (b), ii)* $\mathcal{B}^2$ *also factorises as Figure 2 (a).*

This result can be interpreted as follows: given models corresponding to each causal direction, suppose that both models are distribution-equivalent. Supposing that the priors for each causal direction are specified in line with the ICM assumption, one can compute the anti-causal factorisation of each model, and see which prior it induces in the converse model. If this induced prior is incompatible with the ICM assumption (i.e. it is not separable with respect to the anti-causal factorisation), then we can deduce that the two models cannot be Bayesian distribution-equivalent. In short, if ICM does not hold in the anti-causal factorisation, there exist datasets for which the marginal likelihood of the two models is not equal. This is a necessary condition for distinguishing causal directions. We analyse specific models in app. D. Next, we analyse correctness and provide a test to quantify the level overlap between the models.

4.3. Correctness of Bayesian Model Selection

Distinguishing between models is necessary but not sufficient for correctly identifying the causal direction. Correctness depends on how well the assumptions made in the method match reality. We follow the causal literature by asking when Bayesian model selection will identify the correct causal direction, when the model is correctly specified. The causal literature commonly aims to prove a strict notion of identifiability, where the correct causal direction is *always* be recovered as $N \to \infty$ when assumptions hold. Here, we follow Guyon et al. (2019) in quantifying the probability of making an error in identifying the causal direction.

The assumptions made in a Bayesian model are specified by the model structure and prior densities. In this section, we thus assume that our assumptions hold, and that the data is generated from the chosen BCMs. Given the true causal direction and the decision rule in Equation (6), we find the probability of error by integrating over the region where the wrong model would be selected, i.e.

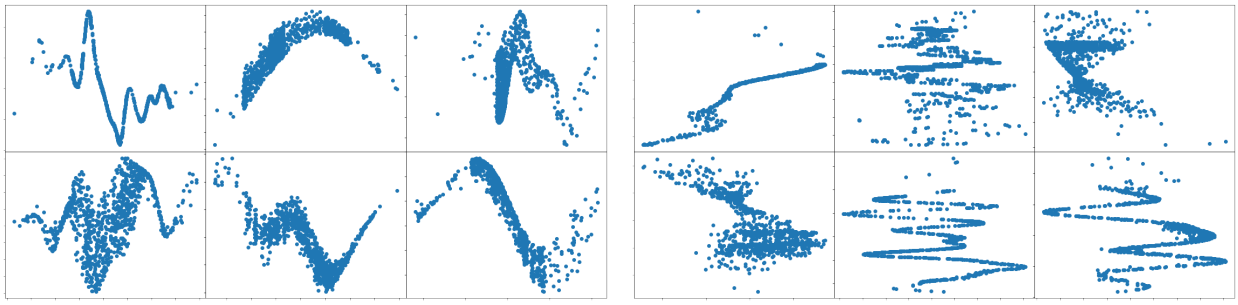
$$P(E|\mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{R}_Y} p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\mathrm{d}\mathcal{D}\,, \tag{7}$$

where $\mathcal{R}_Y = \{\mathcal{D} \mid p(\mathcal{D}|\mathcal{M}_{X\leftarrow \mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\}$. The total probability of error can be written as (app. E.3)

$$P(E) = \frac{1}{2}(1 - \mathrm{TV}[P(\cdot|\mathcal{M}_{\mathbf{X}\to Y}), P(\cdot|\mathcal{M}_{X\leftarrow \mathbf{Y}})]), \tag{8}$$

where $\mathrm{TV}(P, Q) \in [0, 1]$ is the total variation distance. We can see that the probability of making an error falls as the distance between the data distributions increases. For identifiable models where the data distributions are completely separate, $\mathrm{TV} = 1$ and hence $P(E) = 0$, showing that the same guarantees hold for Bayesian model selection as in these strictly identifiable settings. For models without hard restrictions, this probability of error may be positive (overlap in Figure 1(c)). However, this may still be preferable to the error incurred by using an overly-restricted model.

Statistical Test of Asymmetry: We can estimate Equation (8) to find the error under the assumption that a model is correct. This can be used as a statistical test for whether causality can be discerned between two BCMs. The procedure is to sample multiple datasets from the two BCMs, and classify the causal direction by using Equation (6). We



(a) (b)

Figure 3: Samples of datasets from our chosen GPLVM model. ALL figures have the variable X on the x-axis and the variable Y on the y-axis. (a) Shows 6 datasets sampled from GPLVM with $\mathcal{M}_{\mathbf{X}\to Y}$. (b) Shows 6 datasets sampled from GPLVM with $\mathcal{M}_{X\leftarrow\mathbf{Y}}$. The figures show that the data distribution varies between the two Bayesian causal models.

do this here for our model choice: the GPLVM[3]. For 100 datasets of 1000 samples each, we get an approximate probability of error of 0 with a standard deviation upper bounded by 0.05 (app. E). To aid intuition, we plot samples (Figure 3) from the GPLVM prior for the two causal assumptions. Here, we can see that while datasets from $\mathcal{M}_{\mathbf{X}\to Y}$ are multimodal when conditioned on Y, datasets from $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ are multimodal when conditioned on X. This shows that the likely datasets are strongly influenced by the causal assumptions. When the GPLVM is a good description of the data generating process, we thus expect to obtain the correct causal direction with high probability.

4.4. Model Misspecification

The previous result relied on assuming that our model describes the data generating process well. When our models deviate from the true data generating process, our estimate for the probability of error is incorrect. In this case, we can bound the difference between the true probability of error and the one of our models. We denote the true data distributions as $Q(\cdot|\mathbf{X}\to Y)$ and $Q(\cdot|X\leftarrow\mathbf{Y})$, but we use our models to decide on the causal direction, i.e. the decision rule in Equation (6). We can then find a bound (app. E.4)

$$2|Q(\mathrm{E})-P(\mathrm{E})| \leq \mathrm{TV}[Q(\cdot|\mathbf{X}\to Y), P(\cdot|\mathcal{M}_{\mathbf{X}\to Y})]+ \mathrm{TV}[Q(\cdot|X\leftarrow\mathbf{Y}), P(\cdot|\mathcal{M}_{X\leftarrow\mathbf{Y}})].$$

Thus, the difference between the model error and the true error is bounded by the total variation between the true and our model's data distribution, i.e. when assumptions are violated only mildly, we can still accurately estimate the probability of error. However, large deviations from the assumptions will result in an inaccurate probability of error. We test the GPLVM under different data generating assumptions in Section 6.1.

[3]We use an approximation to the marginal likelihood that we describe later.

5. Model Choice

The practical advantage of our approach is the ability to specify causal models with a large $\mathcal{C}$, which may help with reducing model misspecification. To use this advantage, we choose to use the (conditional) GPLVM (Dutordoir et al., 2018) as a prior. Densities for this model are constructed by warping a Gaussian random (latent) variable by a non-linear function (as in a VAE (Kingma and Welling, 2014)), with a Gaussian Process (GP) prior. The latent and flexible prior on f, g allows highly non-Gaussian and heteroskedastic distributions for $\mathcal{C}$. For $\mathcal{M}_{\mathbf{X}\to Y}$, $\mathcal{C}_{\mathbf{X}\to Y}$ takes the form

$$p(y_i|x_i, f, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(y_i|f(x_i, w_i), x_i, w_i)p(w_i)\mathrm{d}w_i\,,$$

$$p(x_i|g, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(x_i|g(v_i), v_i)p(v_i)\mathrm{d}v_i\,,$$

where w_i, v_i are standard Gaussian distributed and priors $f|\boldsymbol{\lambda} \sim \mathcal{GP}$, $g|\boldsymbol{\lambda} \sim \mathcal{GP}$, with $\boldsymbol{\lambda}$ collecting all hyperparameters. The marginal likelihoods are found by integrating over the priors of f, g. We use existing variational inference schemes (Dutordoir et al., 2018; Lalchand et al., 2022) to approximate the marginal likelihoods (app. F). This can cause additional loss in performance compared to ideal Bayesian model comparison (Blei et al., 2017). However, we found our model to have a near zero probability of error with variational inference (Section 4.3). Due to the symmetry of the problem, we assume the same prior on the cause, and effect given cause in the two BCMs. We expand on integrating out hyperparameters in app. G.

6. Experiments

Having laid out our method, we now test it on a mixture of real and synthetic datasets. We test our method on benchmark datasets of a wide variety of dataset-generating distributions, showcasing the benefits of removing model restrictions. These datasets are not sampled from our model, and

Table 1: Performance comparisons. Numbers convey the ROC AUC metric (higher is better). Best results are in bold while the best results from the baselines are underlined. Our method (GPLVM) outperforms competing methods. CDCI contains multiple methods, we show the result of the best method on each dataset.

Methods	CE-Cha	CE-Multi	CE-Net	CE-Gauss	CE-Tueb
LiNGAM	57.8	62.3	3.3	72.2	31.1
ANM	43.7	25.5	87.8	90.7	63.9
PNL	78.6	51.7	75.6	84.7	73.8
IGCI	55.6	77.8	57.4	16.0	63.1
RECI	59.0	94.7	66.0	71.0	70.5
SLOPPY	60.1	95.7	79.3	71.4	65.3
CGNN	76.2	94.7	86.3	89.3	76.6
GPI	71.5	73.8	88.1	90.2	70.6
CDCI (best method reported)	72.2	96.0	94.3	**91.8**	58.7
GPLVM	**81.9**	**97.7**	**98.9**	89.3	**78.3**

hence provide empirical verification of Section 4.4, i.e. that good performance can be obtained with imperfect models. The details of our method are in app. H.2.

Following previous work (Mooij et al., 2016), we use the *Area under the curve* (AUC) of the *Receiver Characteristic Operator* (ROC) metric, ensuring there is 50% of each causal direction present to avoid directional bias. We normalise all datasets following Reisach et al. (2021). Additional experiments are in app. I.

6.1. Real and Synthetic Data

We test the GPLVM under model misspecification on a wide variety of data-generating mechanisms, not all generated from known identifiable models, the full details of which are in app. H.1. The results of our method (GPLVM) are shown in Table 1, along with competing methods. Results for SLOPPY were obtained by rerunning the author's code (app. H.3), results for CDCI are from (Duong and Nguyen, 2021), and the rest are taken from (Guyon et al., 2019).

Patterns to note here are that methods with restrictive assumptions, in exchange for strict identifiability, do not perform very well. Methods with weaker assumptions and without strict identifiability perform better. The poor performance of LiNGAM (Shimizu et al., 2006) can be attributed to the fact that few of the datasets contain linear functions. ANM is seen to perform well when the datasets contain additive noise (CE-Gauss). PNL is less restrictive than ANM, which contributes to its better performance on most datasets. Methods dependent on low noise, such as IGCI, RECI, and SLOPPY only perform well on CE-Multi. RECI and SLOPPY also rely on the additive noise assumption, explaining their similar performance; we observe that SLOPPY performs better in most cases due to its better complexity control. More flexible methods based informally on the ICM assumption such as CGNN, GPI, and CDCI tend to perform better across all datasets. Although CGNN uses neural networks, it requires additional datasets to tune its complexity. GPI uses a similar model as us, but their inference method differs. CDCI is a class of methods and the reported results are the best of 5 different methods. Our approach, labelled GPLVM, performs well on datasets regardless of the data generating assumptions, owing to its ability to model flexible densities. These results demonstrate the strength of our approach in identifying causal direction for more realistic assumptions.

7. Conclusion

In this work, we show that causal discovery with Bayesian model selection allows for removing restrictions on modelling capability that may hamper performance on real world datasets. Starting from first principles, we show how to view causal discovery as Bayesian model selection, encoding important causal assumptions. We then show that Bayesian model selection can infer causality in cases where flexible model choices inhibit likelihood based discovery. We exhibit the reversibility of the ICM assumption as the key underlying mechanism for this. We also discuss the correctness of our method and provide a statistical method for quantifying the probability of error of chosen Bayesian priors. We show that under mild model misspecification, the estimated probability of error can retain its accuracy. We significantly outperform previous methods on a wide range of data generating processes, owing to the removal of restrictions on the model. Such an approach is vital for expanding the use of causal discovery to real world datasets. While we provide a statistical test for quantifying cross-model overlap, an open question is to find universal, practical conditions for model equivalence. An avenue of further interest would be to use deeper models (Damianou and Lawrence, 2013).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 2017.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3), 2020.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov), 2002.

Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*. PMLR, 2013.

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.

Bao Duong and Thin Nguyen. Bivariate causal discovery via conditional divergence. In *First Conference on Causal Learning and Reasoning*, 2021.

Vincent Dutordoir, Hugh Salimbeni, James Hensman, and Marc Deisenroth. Gaussian process conditional density estimation. *Advances in neural information processing systems*, 31, 2018.

José AR Fonollosa. Conditional distribution variability measures for causality detection. In *Cause Effect Pairs in Machine Learning*. Springer, 2019.

Nir Friedman and Iftach Nachman. Gaussian process networks. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, 2000.

Dan Geiger and David Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5), 2002.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018.

Peter Grunwald and Paul Vitányi. Shannon information and kolmogorov complexity. *arXiv preprint cs/0410002*, 2004.

Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *arXiv preprint arXiv:2203.15756*, 2022.

Isabelle Guyon, Olivier Goudet, and Diviyan Kalainathan. Evaluation methods of cause-effect pairs. In *Cause Effect Pairs in Machine Learning*, chapter 2. Springer, 2019.

David Heckerman. A bayesian approach to learning causal networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 285–295, 1995.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3), 1995.

David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. *Innovations in Machine Learning: Theory and Applications*, 2006.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Edwin Hewitt and Leonard J Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2), 1955.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. *arXiv preprint arXiv:2210.09054*, 2022.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10), 2010.

Dominik Janzing and Bastian Steudel. Justifying additive noise model-based causal discovery via algorithmic information theory. *Open Systems & Information Dynamics*, 17(02), 2010.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Maximilian Kurthen and Torsten Enßlin. A bayesian model for bivariate causal inference. *Entropy*, 22(1), 2019.

Vidhi Lalchand, Aditya Ravuri, and Neil D Lawrence. Generalised gaussian process latent variable models (gplvm) with stochastic variational inference. *arXiv preprint arXiv:2202.12979*, 2022.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15 (1), 2014.

David JC MacKay. Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11 (5), 1999.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Alexander Marx and Jilles Vreeken. Identifiability of cause and effect using regularized regression. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

Alexander Marx and Jilles Vreeken. Formally justifying mdl-based inference of cause and effect. In *AAAI Workshop on Information-Theoretic Causal Inference and Discovery (ITCI'22)*, 2022.

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1), 2016.

Frank Nielsen. Generalized bhattacharyya and chernoff upper bounds on bayes error using quasi-arithmetic means. *Pattern Recognition Letters*, 42, 2014.

Sebastian W Ober, Carl E Rasmussen, and Mark van der Wilk. The promises and pitfalls of deep kernel learning. In *Uncertainty in Artificial Intelligence*. PMLR, 2021.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (5), 2016.

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

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*. Springer, 2003.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

Nino Scherrer, Anirudh Goyal, Stefan Bauer, Yoshua Bengio, and Nan Rosemary Ke. On the generalization and adaption performance of causal models. *arXiv preprint arXiv:2206.04620*, 2022.

B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*. International Machine Learning Society, 2012.

Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Artificial intelligence and statistics*. PMLR, 2015.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *Advances in neural information processing systems*, 23, 2010.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*. PMLR, 2020.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*. PMLR, 2009.

Michalis Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, 1990.

Zihao Wang and Victor Veitch. A unified causal view of domain invariant representation learning. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2), 2015.

A. Background

A.1. Markov Equivalence Class

Definition A.1. Given a directed acyclic graph (DAG) $\mathcal{G}$ and a joint distribution $P(X_1, \ldots, X_D)$, the distribution is said to satisfy the **Markov property** if for disjoint sets $\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k$

$$\mathbf{X}_i \perp\!\!\!\perp_{\mathcal{G}} \mathbf{X}_j | \mathbf{X}_k \implies \mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j | \mathbf{X}_k, \tag{9}$$

where $\perp\!\!\!\perp_{\mathcal{G}}$ denotes d-separation in $\mathcal{G}$ and $\perp\!\!\!\perp$ denotes independence (Peters et al., 2017).

There can be more than one distribution that is Markov with respect to a DAG $\mathcal{G}$. If there exists a set of distributions that is Markov with respect to two distinct DAGs, $\mathcal{G}_1$ and $\mathcal{G}_2$, then $\mathcal{G}_1$ and $\mathcal{G}_2$ are *Markov equivalent*. The set of all Markov equivalent DAGs is called a *Markov equivalence class*. Verma and Pearl (1990) provide a graphical criteria for determining whether two graphs are Markov equivalent. Simply stated, two DAGs are Markov equivalent if they share the same adjacencies and the same colliders. This effectively states that the respective Markov distributions must obey the same set of conditional independences.

A.2. Distribution-equivalence and causal identification

In this section, we expand on the relation between distribution-equivalence and causal identifiability. For readability, we state the relevant definitions again.

A.2.1. Distribution-equivalence

For clarity, we reintroduce the definition of a causal model.

Definition A.2. A causal model is a tuple $\mathcal{M}_{\mathcal{G}} = (\mathcal{G}, \mathcal{C}, \mathcal{F})$, where $\mathcal{G}$ is a DAG with vertex set $\mathcal{V}$, and $\mathcal{C}$ is a set of conditional distributions

$$\mathcal{C} = \prod_{i \in \mathcal{V}} \mathcal{C}_{i|\mathrm{pa}_{\mathcal{G}}(i)} \subset \prod_{i \in \mathcal{V}} \mathcal{K}(\mathcal{X}_{\mathrm{pa}_{\mathcal{G}}(i)} \to \mathcal{X}_i) \tag{10}$$

that are Markov with respect to $\mathcal{G}$. Given $P = (P_i : i \in \mathcal{V}) \in \mathcal{C}$, define $\delta_{\mathcal{C}} : \mathcal{C} \to \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ as the map that assembles P into the corresponding joint

$$\delta_{\mathcal{C}}(P)(\mathrm{d}x_{\mathcal{V}}) = \prod_{i \in \mathcal{V}} P_i(\mathrm{d}x_i \mid x_{\mathrm{pa}_{\mathcal{G}(i)}}). \tag{11}$$

Finally, define $\mathcal{F}$ as the set of induced joint distributions $\mathcal{F} = \{\delta_{\mathcal{C}}(P) : P \in \mathcal{C}\}$.

We restate the definition of distribution-equivalence from (Geiger and Heckerman, 2002).

Definition A.3. Two causal models $\mathcal{M}_{\mathbf{X} \to Y} = (\mathbf{X} \to Y, \mathcal{C}_{\mathbf{X} \to Y}, \mathcal{F}_{\mathbf{X} \to Y})$ and $\mathcal{M}_{X \leftarrow \mathbf{Y}} = (X \leftarrow \mathbf{Y}, \mathcal{C}_{X \leftarrow \mathbf{Y}}, \mathcal{F}_{X \leftarrow \mathbf{Y}})$ are **distribution-equivalent** if $\mathcal{F}_{\mathbf{X} \to Y} = \mathcal{F}_{X \leftarrow \mathbf{Y}}$. Equivalently, there exists a unique translating bijection $\gamma : \mathcal{C}_{\mathbf{X} \to Y} \to \mathcal{C}_{X \leftarrow \mathbf{Y}}$ such that for any $P \in \mathcal{C}_{\mathbf{X} \to Y}$, there holds an equality of (joint) measures $\delta_{\mathcal{C}_{\mathbf{X} \to Y}}(P) = \delta_{\mathcal{C}_{X \leftarrow \mathbf{Y}}}(\gamma(P))$.

Distribution equivalence implies that for every $(m, c) \in \mathcal{C}_X \times \mathcal{C}_{Y|X}$, there exists $(m', c') \in \mathcal{C}_Y \times \mathcal{C}_{X|Y}$ such that

$$m(\mathrm{d}x) \cdot c(\mathrm{d}y \mid x) = m'(\mathrm{d}y) \cdot c'(\mathrm{d}x \mid y).$$

Thus, given a dataset $\mathcal{D}^N = \left(\mathbf{x}^N, \mathbf{y}^N\right)$ (and for sufficiently large N), maximum likelihood cannot distinguish distribution-equivalent causal models. Hence, causal models that are distribution-equivalent, are not identifiable by maximum likelihood. Note that Markov equivalent graphs need not imply distribution-equivalent causal models (Geiger and Heckerman, 2002).

A.2.2. Causal identifiability.

If two causal models, $\mathcal{M}_{\mathbf{X} \to Y}$ and $\mathcal{M}_{X \leftarrow \mathbf{Y}}$, are not distribution equivalent, then there exists a $(m, c) \in \mathcal{C}_X \times \mathcal{C}_{Y|X}$ such that for all $(m', c') \in \mathcal{C}_Y \times \mathcal{C}_{X|Y}$,

$$m(\mathrm{d}x) \cdot c(\mathrm{d}y \mid x) \neq m'(\mathrm{d}y) \cdot c'(\mathrm{d}x \mid y).$$

In this case, there exists some dataset $\mathcal{D}^N$ such that (for large enough N) maximum likelihood can be used to identify the causal model.

We retain the traditional notion of causal identifiability if we force $\mathcal{F}_{\mathbf{X}\to Y}$ and $\mathcal{F}_{X\leftarrow\mathbf{Y}}$ to be disjoint.

Definition A.4. (adapted from Guyon et al. (2019, ch. 2)) Given two causal models $\mathcal{M}_{\mathbf{X}\to Y}$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}}$, the models are said to be **identifiable** if for every $(m, c) \in \mathcal{C}_X \times \mathcal{C}_{Y|X}$, there is no $(m', c') \in \mathcal{C}_Y \times \mathcal{C}_{X|Y}$ such that

$$m(\mathrm{d}x) \cdot c(\mathrm{d}y \mid x) = m'(\mathrm{d}y) \cdot c'(\mathrm{d}x \mid y).$$

The above definition ensures that the two causal models will never get the same maximum likelihood score, regardless of the dataset.

An example of an identifiable causal model is when $\mathcal{C}$ is restricted to an additive noise form (Hoyer et al., 2008); these are collectively called additive noise models (ANM). In this case, we can define the conditional family that induces $\mathcal{F}_{\mathbf{X}\to Y}$ as

$$\mathcal{C}_{Y|X} = \left\{P_{\theta_{Y|X}} : P_{\theta_{Y|X}}(Y|X) = \theta_{Y|X}(X) + N, \theta_{Y|X} \in \Theta_{Y|X}\right\},$$

where $\Theta_{Y|X}$ is an arbitrary space of non-linear functions, N is sampled from some arbitrary distribution and $N \perp\!\!\!\perp Y$. For this choice of conditional distribution family, Hoyer et al. (2008, thm. 1) show that, unless some complicated conditions hold, the backward factorisation of an ANM causal model is not expressible as an ANM model. Simply put, suppose $\mathcal{C}_{Y|X}$ and $\mathcal{C}_{X|Y}$ are additive non-linear models, then if $P \in \mathcal{F}_{X\leftarrow\mathbf{Y}}$, then generally $P \notin \mathcal{F}_{\mathbf{X}\to Y}$. Hence, additive noise models are identifiable.

Zhang et al. (2015) demonstrated that for a Structural Causal Model (SCM) assuming independence of noise on the effect from the cause, certain restrictions on the SCM are necessary to ensure identifiability. If the restrictions imply that the backward factorisation does not admit independence between the noise and the input, the causal model is identifiable. This assumption on the conditional family for $\mathcal{F}_{\mathbf{X}\to Y}$ can be written as

$$\mathcal{C}_{Y|X} = \left\{P_{\theta_{Y|X}} : P_{\theta_{Y|X}}(Y|X) = \theta_{Y|X}(X, N), Y \perp\!\!\!\perp N, \theta_{Y|X} \in \Theta_{Y|X}\right\},$$

with an analogous family $\mathcal{C}_{X\leftarrow\mathbf{Y}}$, with the same parameter space, $\Theta := \Theta_{Y|X} = \Theta_{X|Y}$ and the same noise distribution. The causal models with this family are then identifiable if Θ is restricted appropriately such that $P \in \mathcal{F}_{\mathbf{X}\to Y} \implies P \notin \mathcal{F}_{X\leftarrow\mathbf{Y}}$. Zhang et al. (2015) also formally show that maximum likelihood can be used to identify such models.

B. Discussion on Kolmogorov Complexity, Causality and Bayes

B.1. Kolmogorov Complexity and Causality

The **Kolmogorov complexity** of a string x, denoted $K(x)$, is the length of the shortest computer program that prints x and halts (Grunwald and Vitányi, 2004). This computer program can be written in any universal language, the complexity will change based on the universal language by a constant factor not depending on x. We can equally define a conditional version of Kolmogorov complexity, given an input string y as the shortest program that generates x from y and halts — $K(x|y)$. We can then think of the Kolmogorov complexity of a function, for a given input x, as the shortest program that generates the output $f(x)$ up to a certain precision. The definition of the Kolmogorov complexity of a probability distribution follows.

Janzing and Schölkopf (2010) propose using the Kolmogorov complexity of factorisations of the joint to infer causality. Given a causal graph $\mathbf{X} \to Y$, they formalised the assumption of *Independent Causal Mechanisms* (ICM) in terms of Kolmogorov complexity by stating that the *algorithmic mutual information* of the causal factorisation is zero,

$$I(P_X : P_{Y|X}) \stackrel{+}{=} 0 \tag{12}$$

$$\implies K(P_X, P_{Y|X}) \stackrel{+}{=} K(P_X) + K(P_{Y|X}) \tag{13}$$

$$\implies K(P_X) + K(P_{Y|X}) \stackrel{+}{\leq} K(P_Y) + K(P_{X|Y}), \tag{14}$$

where $I(\cdot : \cdot)$ is the algorithmic mutual information (Grunwald and Vitányi, 2004), Equation (13) follows by definition and Equation (14) follows from the fact that the Kolmogorov complexity of the anti-causal factorisation cannot be less than that of the joint. The addition symbol above the inequality relations symbolises the fact they only hold up to an additive constant.

Equation (14) suggests that causality can be inferred by finding the factorisation with the lowest Kolmogorov complexity. However, in addition to the fact that Equation (14) requires access to the actual distributions, and that the relation only holds up to unknown additive constants, the Kolmogorov complexity is also uncomputable (Grunwald and Vitányi, 2004). Equation (14) has thus been used informally to try and infer causality from data (Goudet et al., 2018; Mitrovic et al., 2018; Duong and Nguyen, 2021; Stegle et al., 2010). These methods only use Equation (14) as a philosophical foundation, and Equation (14) does not necessarily provide guarantees that their method will return the correct causal direction.

B.2. Minimum Description Length relaxations

Recently there have been attempts to find an analogous inequality to Equation (14) using the *Minimum Description length* (MDL) principle (Grünwald, 2007). First, MDL allows for reasoning about finite data that has to be used to estimate the relevant probability distributions. Marx and Vreeken (2022) use relations between Shannon entropy and Kolmogorov complexity to find a formulation in terms of the Kolmogorov complexity of the model and data given the probability distribution (Marx and Vreeken, 2022)

$$K_{\mathbf{X}\to Y} := K(P_X) + K(x|P_X) + K(P_{Y|X}) + K(y|x, P_{Y|X}). \tag{15}$$

In expectation, above equation will equal the left hand side of the inequality Equation (14),

$$\mathbb{E}_{P(x,y)}[K_{\mathbf{X}\to Y}] = K(P_X) + K(P_{Y|X}). \tag{16}$$

Hence, the inequality in Equation (14) only holds in expectation for finite data. Second, MDL restricts the definition of Kolmogorov complexity from the set of all programs to only those that can be computed, usually specified by a model class. Marx and Vreeken (2022) further make the assumption of a model class, $\mathcal{M}$, and assume that the data is generated from that model class. In this case Equation (15) can be written as

$$L_{\mathcal{M}_{\mathbf{X}\to Y}} := L(\mathcal{M}_X) + L(\mathbf{x}|\mathcal{M}_X) + L(\mathcal{M}_{Y|X}) + L(\mathbf{y}|\mathbf{x}, \mathcal{M}_{Y|X}), \tag{17}$$

where L is some encoding scheme. As such, the above approach can be considered as balancing the fit of the model (encoding of data given the model) and the complexity of the model class (encoding of the model). However, here the exact performance will depend on the encoding scheme used.

The Bayesian approach we have considered can be seen as a variant of the MDL principle. Here, the data given the model and model are note encoded separately. Specifically, due to the fact that the marginal likelihood has to normalise over datasets, it has an in built complexity penalty. It thus also balances model fit along with a complexity penalty. To see this clearly, consider a model $\mathcal{M}$ with parameter ρ and prior $p(\rho|\mathcal{M})$, we can write the marginal likelihood as

$$p(\mathbf{x}|\mathcal{M}) = \mathbb{E}_{p(\rho|\mathbf{x},\mathcal{M})}[p(\mathbf{x}|\rho, \mathcal{M})] - \mathrm{KL}[p(\rho|\mathbf{x}, \mathcal{M})\|p(\rho|\mathcal{M})], \tag{18}$$

where the first term is the model fit (expectation of the likelihood under the posterior), and the second term is the complexity penalty (distance from the posterior to the prior). Our approach can thus also be justified by using MDL arguments, though the MDL view does not provide insight into why the Bayesian approach works, nor into the consequences of the choice of priors and models. The choice of the prior is subjective and equivalent to choosing a normalised luckiness function in refined MDL (Grünwald, 2007).

C. Proofs

We re-state the propositions and provide their proofs in this section. We being by proving a lemma that will help us prove the propositions.

Lemma C.1. *Given two Bayesian causal models* $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$ *and* $(\mathcal{M}_{X\leftarrow \mathbf{Y}}, \pi_{X\leftarrow \mathbf{Y}})$*, they are Bayesian distribution equivalent precisely when*

$$\delta_{\mathcal{C}_{\mathbf{X}\to Y}} \# \pi_{\mathbf{X}\to Y} = \delta_{\mathcal{C}_{X\leftarrow \mathbf{Y}}} \# \pi_{X\leftarrow \mathbf{Y}}. \tag{19}$$

Proof. To prove this, we make use of the Hewitt-Savage representation theorem in Hewitt and Savage (1955, thm. 9.4). From this, it holds that for each model, the data distribution is uniquely expressible in the form

$$P(\cdot \mid \mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mu_{\mathbf{X}\to Y}(\mathrm{d}Q) \cdot Q^{\otimes\infty}\,, \tag{20}$$

$$P(\cdot \mid \mathcal{M}_{X\leftarrow\mathbf{Y}}) = \int_{\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mu_{X\leftarrow\mathbf{Y}}(\mathrm{d}Q) \cdot Q^{\otimes\infty}, \tag{21}$$

where $\mu_{\mathbf{X}\to Y}$ and $\mu_{X\leftarrow\mathbf{Y}}$ are probability measures over the space $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$, that is probability measures over the space of probability measures over $\mathcal{X}\times\mathcal{Y}$. In particular $P(\cdot \mid \mathcal{M}_{\mathbf{X}\to Y}) = P(\cdot \mid \mathcal{M}_{X\leftarrow\mathbf{Y}})$ precisely when $\mu_{\mathbf{X}\to Y} = \mu_{X\leftarrow\mathbf{Y}}$ (Hewitt and Savage, 1955). By construction we can take $\mu_{\mathbf{X}\to Y} = \delta_{\mathcal{C}_{\mathbf{X}\to Y}}\#\pi_{\mathbf{X}\to Y}$ and $\mu_{X\leftarrow\mathbf{Y}} = \delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}}\#\pi_{X\leftarrow\mathbf{Y}}$. The result follows. □

In words, the substance of this result is that in order to check the property of Bayesian distribution-equivalence, it suffices to compare the marginal distribution of a single bivariate observation under each model. Note that this result is agnostic to the model parametrisation and its identifiability, as it concerns only marginal data distributions.

Proposition C.2. *Given two BCMs $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$, $(\mathcal{M}_{X\leftarrow\mathbf{Y}}, \pi_{X\leftarrow\mathbf{Y}})$, suppose that there exists a subset $\mathcal{C}_\Delta \subset \mathcal{C}_{\mathbf{X}\to Y}$ such that $\pi_{\mathbf{X}\to Y}(\mathcal{C}_\Delta) > 0$, and $\delta_{\mathcal{C}_{\mathbf{X}\to Y}}(\mathcal{C}_\Delta) \cap \mathcal{F}_{X\leftarrow\mathbf{Y}}$ is empty. Then the two Bayesian causal models are not Bayesian distribution-equivalent.*

Proof. Let $\mathcal{F}_\Delta := \delta_{\mathcal{C}_{\mathbf{X}\to Y}}(\mathcal{C}_\Delta)$ noting that $\mathcal{F}_\Delta \in \mathcal{F}_{\mathbf{X}\to Y}$ but $\mathcal{F}_\Delta \notin \mathcal{F}_{X\leftarrow\mathbf{Y}}$. There is no choice of prior $\pi_{X\leftarrow\mathbf{Y}}$ such that $(\delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}}\#\pi_{X\leftarrow\mathbf{Y}})(\mathcal{F}_\Delta) = (\delta_{\mathcal{C}_{\mathbf{X}\to Y}}\#\pi_{\mathbf{X}\to Y})(\mathcal{F}_\Delta)$. The result immediately follows from Lemma C.1. □

Note that the same result as above holds if we assume that $\delta_{\mathcal{C}_{\mathbf{X}\to Y}}(\mathcal{C}_\Delta) \cap \mathcal{F}_{X\leftarrow\mathbf{Y}}$ is non-empty, but $\mathcal{C}_\Delta$ no mass under $\pi_{X\leftarrow\mathbf{Y}}$. In this respect, the effect of placing hard restrictions on the set of conditionals can be mimicked by suitable prior design.

For distribution-equivalent models, we show that the ICM assumption must also hold in the anti-causal factorisation for the models to be Bayesian distribution-equivalent. We formalise the reversibility of the ICM assumption by introducing separable-compatibility.

Definition C.3. Let $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$, $(\mathcal{M}_{X\leftarrow\mathbf{Y}}, \pi_{X\leftarrow\mathbf{Y}})$ be two Bayesian causal models where the underlying causal models are distribution-equivalent, denoting γ as the corresponding translation mapping $\gamma : \mathcal{C}_{\mathbf{X}\to Y} \to \mathcal{C}_{X\leftarrow\mathbf{Y}}$ (in Definition 2.2). Say the two are **separable-compatible** if: 1) the pushforward $\gamma\#\pi_{\mathbf{X}\to Y}$ is separable with respect to $\mathcal{C}_{X\leftarrow\mathbf{Y}}$, 2) $\gamma^{-1}\#\pi_{X\leftarrow\mathbf{Y}}$ is separable with respect to $\mathcal{C}_{\mathbf{X}\to Y}$.

Proposition C.4. *Fix two Bayesian causal models $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$, $(\mathcal{M}_{X\leftarrow\mathbf{Y}}, \pi_{X\leftarrow\mathbf{Y}})$, where the underlying causal models are distribution-equivalent. The two Bayesian causal models are Bayesian distribution-equivalent only if they are separable-compatible.*

Proof. Assume that $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$ and $(\mathcal{M}_{X\leftarrow\mathbf{Y}}, \pi_{X\leftarrow\mathbf{Y}})$ are Bayesian distribution-equivalent. Take $\mu_{\mathbf{X}\to Y} = \delta_{\mathcal{C}_{\mathbf{X}\to Y}}\#\pi_{\mathbf{X}\to Y}$ and $\mu_{X\leftarrow\mathbf{Y}} = \delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}}\#\pi_{X\leftarrow\mathbf{Y}}$. By Lemma C.1 $\mu_{\mathbf{X}\to Y} = \mu_{X\leftarrow\mathbf{Y}}$. Recalling that $\mathcal{M}_{\mathbf{X}\to Y}$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ are distribution-equivalent, the translation mapping γ satisfies $\delta_{\mathcal{C}_{\mathbf{X}\to Y}} = \delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}} \circ \gamma$, also that the mapping $\delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}}$ is injective. We can argue

$$\begin{aligned}
\delta_{\mathcal{C}_{\mathbf{X}\to Y}}\#\pi_{\mathbf{X}\to Y} &= \delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}}\#\pi_{X\leftarrow\mathbf{Y}},\\
(\delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}} \circ \gamma)\#\pi_{\mathbf{X}\to Y} &= \delta_{\mathcal{C}_{X\leftarrow\mathbf{Y}}}\#\pi_{X\leftarrow\mathbf{Y}},\\
\gamma\#\pi_{\mathbf{X}\to Y} &= \pi_{X\leftarrow\mathbf{Y}},
\end{aligned}$$

that is, the pushforward $\gamma\#\pi_{\mathbf{X}\to Y}$ is equal to $\pi_{X\leftarrow\mathbf{Y}}$, and hence is separable with respect to $\mathcal{C}_{X\leftarrow\mathbf{Y}}$. To conclude, $(\mathcal{M}_{\mathbf{X}\to Y}, \pi_{\mathbf{X}\to Y})$ and $(\mathcal{M}_{X\leftarrow\mathbf{Y}}, \pi_{X\leftarrow\mathbf{Y}})$ are separable-compatible. □

The proof of Corollary 4.8 directly follows from the above. For parametrised models, we can define δ as the map from the parameter space $\Theta \times \Phi$ to the space of probability measures $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$. From the above proof, we can see that this will require the parametrisation to be injective, that is, the map $\delta : \Theta \times \Phi \to \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ to be injective.

D. Analysis of Models

In Section 4.2, we showed that two models being separable-compatible is a necessary condition for the models to be Bayesian distribution-equivalent. Thus, if the models are not separable-compatible, they will not be Bayesian distribution-equivalent. In this section, we analyse specific models and argue that Bayesian causal models being separable-compatible is a strong condition that does not hold very often.

D.1. Unnormalised linear Gaussian model

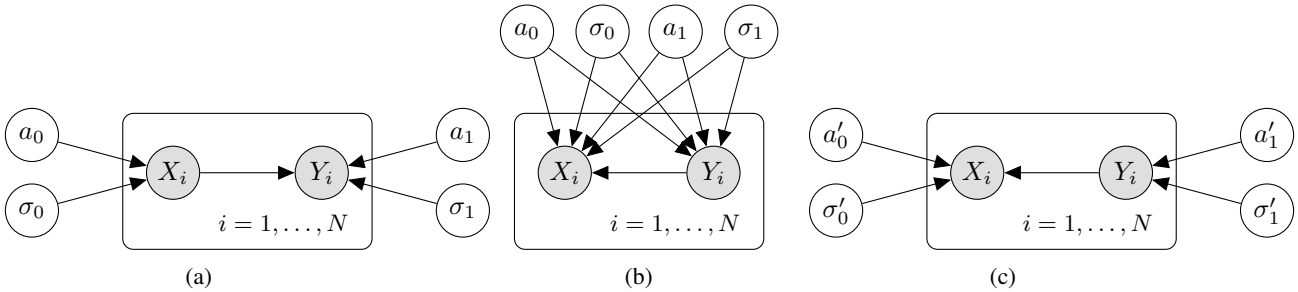


Figure 4: Graphical models for: (a) The linear Gaussian causal model $\mathcal{M}_{\mathbf{X}\to Y}$ in Equation (22). (b) The anti-causal factorisation of $\mathcal{M}_{\mathbf{X}\to Y}$ in Equation (24). (c) The causal model for $\mathcal{M}_{X\leftarrow\mathbf{Y}}$, where ICM holds in the factorisation $P(Y)P(X|Y)$.

Here, we consider the Linear Gaussian model of the form

$$\begin{aligned} P(X|a_0, \sigma_0, \mathcal{M}_{\mathbf{X}\to Y}) &= \mathcal{N}(a_0, \sigma_0^2), \\ P(Y|X, a_1, \sigma_1, \mathcal{M}_{\mathbf{X}\to Y}) &= \mathcal{N}(a_1 X, \sigma_1^2). \end{aligned} \tag{22}$$

We can compute the anti-causal factorisation of this model as

$$P(Y|a_0, \sigma_0, a_1, \sigma_1, \mathcal{M}_{\mathbf{X}\to Y}) = \mathcal{N}(a_1 a_0, \sigma_1^2 + a_1^2\sigma_0^2), \tag{23}$$

$$P(X|Y, a_0, \sigma_0, a_1, \sigma_1, \mathcal{M}_{\mathbf{X}\to Y}) = \mathcal{N}\left(\Sigma\left(\frac{a_1}{\sigma_1^2}Y + \frac{a_0}{\sigma_0^2}\right), \Sigma\right) \tag{24}$$

with $\Sigma = \frac{\sigma_0^2\sigma_1^2}{\sigma_1^2+\sigma_0^2 a_1^2}$. We can see that while the causal factorisation factorises as Figure 4 (a), the anti-causal factorisation, in general, factorises as Figure 4 (b). The only way for the linear Gaussian Bayesian causal models to be separable-compatible, is if the priors on the parameters are chosen such that

$$a_1 a_0 \perp\!\!\!\perp \Sigma\left(\frac{a_1}{\sigma_1^2}Y + \frac{a_0}{\sigma_0^2}\right), \tag{25}$$

$$\sigma_1^2 + a_1^2\sigma_0^2 \perp\!\!\!\perp \Sigma. \tag{26}$$

If the above holds, we can reparametrise the anti-causal factorisation to factorise as in Figure 4 (c). Geiger and Heckerman (2002) show that the **only** prior that satisfies this property is the normal-Wishart prior (the Wishart prior can be scaled by a real function for the bivariate case, see Appendix of Geiger and Heckerman (2002)). This is known as the BGe model.

BGe Model: We illustrate how the prior in the BGe model gives the independence required in Equation (23). We only show this for the mean and defer to the results in Geiger and Heckerman (2002) for the complete argument. In this model, a Normal-Wishart prior is placed directly on the joint Gaussian

$$P\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}^{-1}\right). \tag{27}$$

A normal-Wishart prior on $\boldsymbol{\mu}$ is of the form $P(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\eta}, \gamma\mathbf{W}^{-1})$. We can factorise the distribution for X and Y, for example

$$P(X) = \mathcal{N}\big(\mu_1, (W_{11} - W_{12}W_{22}^{-1}W_{21})^{-1}\big), \tag{28}$$

$$P(Y|X) = \mathcal{N}(\mu_2 - W_{21}^{-1}W_{22}X + W_{21}W_{22}^{-1}\mu_1, W_{22}^{-1}). \tag{29}$$

We show that the random variable μ_1 is independent of $\mu_2 + W_{21}W_{22}^{-1}\mu_1$. Hence, separability of priors holds. To see this, we can factorise the prior

$$p(\mu_1) = \mathcal{N}(\eta_1, (W_{11} - W_{12}W_{22}^{-1}W_{21})^{-1}),$$
$$p(\mu_2|\mu_1) = \mathcal{N}(\eta_2 - W_{21}W_{22}^{-1}(\mu_1 - \eta_1), W_{22}^{-1})$$

The implied distribution of $\mu_2' = \mu_2 + W_{21}W_{22}^{-1}\mu_1$ is thus

$$p_{\mu_2|\mu_1}(\mu_2 + W_{21}W_{22}^{-1}\mu_1) = \mathcal{N}(\eta_2 + W_{21}W_{22}^{-1}\eta_1, W_{22}^{-1}),$$

which is readily seen to be independent of μ_1. The same can be shown for the Wishart distribution on $\mathbf{W}$, i.e. that $W_{11} - W_{12}W_{22}^{-1}W_{21}$ is independent of W_{22} (Geiger and Heckerman, 2002).

Thus for any other choice of prior, if the prior matches the data generating distribution, we expect Bayesian model selection to find the correct causal direction in this case. However, this requires knowledge of the true variance of the cause and effect in this case. This matches known results in Loh and Bühlmann (2014). The above may not be desirable as simple scaling of the data, changing the variance, may alter inference of the causal direction, see discussion in Reisach et al. (2021). We thus suggest normalising all datasets which will render the method mean and scale invariant.

We can also visually verify that two Bayesian Linear Gaussian models imply different data distributions. We draw parameters from priors — an inverse gamma for the scales. For ease of exposition, we consider the case where the means are 0. Of course, the datasets will differ more if we do not include this constraint. For the same set of drawn parameters, we also find the parameters of the joint for the causal model $\mathcal{M}_{X \leftarrow \mathbf{Y}}$. We then plot the contours of the resulting joint distributions as generated by the two causal models. This shows the likely joint distributions that a model generates. Figure 5 shows contours of such Gaussians. This shows distributions with the same mean and variances for the cause and effect but with different ground truth causal directions $\mathbf{X} \rightarrow Y$ (red) and $X \leftarrow \mathbf{Y}$ (blue). Figure 5 (a) shows 1 such joint, and (b) shows 10 such joints. From these figures, it is clear to see that the causal directions alone imply different joint distributions.

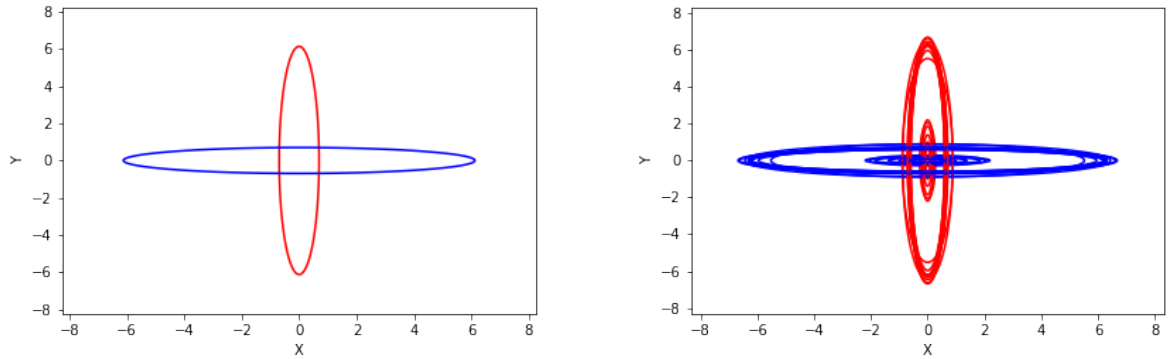


Figure 5: Shows samples of joint distributions from the same priors on parameters of the cause and effect, but with different causal models. Red shows the joint of the causal model $\mathbf{X} \rightarrow Y$ and blue shows the joint of the causal model $X \leftarrow \mathbf{Y}$. The contours are plotted for the same draw of parameters, showing that the different causal models will explain different joint distributions well.

D.2. Normalised linear Gaussian model

As we suggest normalising the dataset in line with Reisach et al. (2021), we show here that for this model, if the same prior is used for both causal models (as we recommend for our model choice), then the causal models will be Bayesian distribution-equivalent.

We conduct the same procedure as above, but ensure that the marginal distributions of X and Y are normalised to $\mathcal{N}(0, 1)$. Figure 6 (a) shows the contours of one such joint distribution for the two causal models $\mathbf{X} \rightarrow Y$ and $X \leftarrow \mathbf{Y}$, while (b) shows

10 such joints. We can see that the joints completely overlap for the two causal models, hence the Bayesian model selection will have no opinion on the true causal model, and will convey this uncertainty.

We can show this mathematically by assuming the data samples has been generated as follows

$$\Pi(X|a_0, \sigma_0) = \mathcal{N}(a_0, \sigma_0^2), \tag{30}$$

$$\Pi(Y|X, a_1, \sigma_1) = \mathcal{N}(a_1 X, \sigma_1^2). \tag{31}$$

On normalisation, we create two new variables X' and Y' that have a standard normal distribution. Accordingly, we set $X' = \frac{X - a_0}{\sigma_0}$ and $Y' = \frac{Y - a_0 a_1}{\sqrt{\sigma_1^2 + \sigma_0^2 a_1^2}}$, with distributions

$$\Pi(X') = \mathcal{N}(0, 1), \tag{32}$$

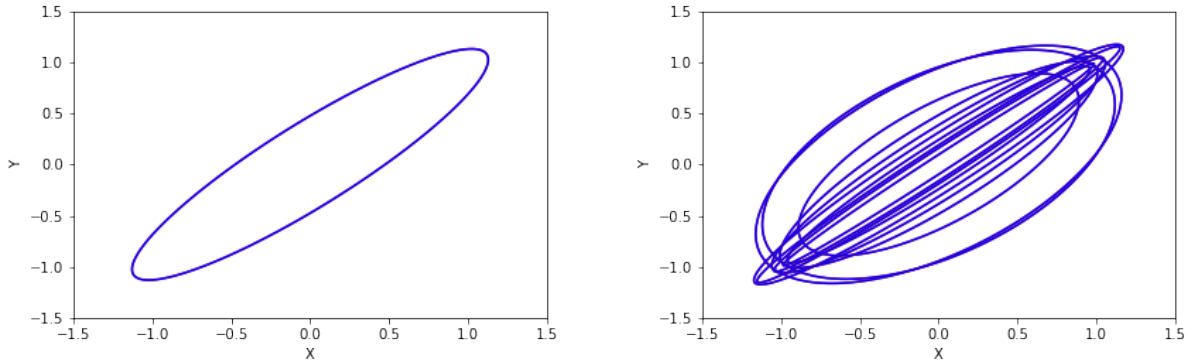
$$\Pi(Y'|X', a_1, \sigma_0, \sigma_1) = \mathcal{N}\left(\frac{\sigma_0 a_1}{\sqrt{\sigma_1^2 + \sigma_0^2 a_1^2}} X', \frac{\sigma_1^2}{\sigma_1^2 + \sigma_0^2 a_1^2}\right). \tag{33}$$

We can see that both $\Pi(X')$ and $\Pi(Y')$ are standard normal. The anti-causal factorisation in this case results in the exact same distribution. Using simple algebra we get

$$\Pi(Y') = \mathcal{N}(0, 1), \tag{34}$$

$$\Pi(X'|Y', a_1, \sigma_0, \sigma_1) = \mathcal{N}\left(\frac{\sigma_0 a_1}{\sqrt{\sigma_1^2 + \sigma_0^2 a_1^2}} Y', \frac{\sigma_1^2}{\sigma_1^2 + \sigma_0^2 a_1^2}\right). \tag{35}$$

For any prior on the parameters of $\Pi(X')\Pi(Y' \mid X' a_1, \sigma_0, \sigma_1)$, the same prior on $\Pi(Y')\Pi(X' \mid Y' a_1, \sigma_0, \sigma_1)$ will result in the two causal models having the same data distribution. Hence, there exist priors such that the Bayesian causal models constructed out of these causal models cannot be discriminated by the marginal likelihood. Note, we can clearly see here that for any choice of priors, the two models are separable-compatible.



(a) (b)

Figure 6: Best viewed in colour. Joint contours for normalised cause and effect. The joint contours completely overlap for the two causal models. Hence, both causal models will explain both datasets equally well.

D.3. Gaussian Process Latent variable model

We refer the reader to Rasmussen (2003) for an introduction to Gaussian processes. Here, we argue that the GPLVM prior does not satisfy separability with respect to the anti-causal factorisation. We only condition on the model ($\mathcal{M}_{\mathbf{X} \to Y}$ or $\mathcal{M}_{X \leftarrow \mathbf{Y}}$) on the left hand side of equations to avoid clutter of notation. Note that the final data density (density implied by the data distribution) is found by integrating over the hyperparameters as laid out in app. G. For clarity, we omit this additional step from this section.

Model $\mathcal{M}_{\mathbf{X}\to Y}$: In $\mathcal{M}_{\mathbf{X}\to Y}$, $\mathbf{x}$ is modelled directly, and $\mathbf{y}$ is modelled conditional on $\mathbf{x}$. For the estimation of $\mathbf{x}$ we have

$$p(\mathbf{x}, \mathbf{f}_X, \mathbf{w}_X|\lambda_X, \mathcal{M}_{\mathbf{X}\to Y}) = p(\mathbf{x}|\mathbf{f}_X)p(\mathbf{f}_X|\mathbf{w}_X, \lambda_X)p(\mathbf{w}_X), \tag{36}$$

where λ_X collects all the hyperparameters for modelling $\mathbf{x}$. The data density can be found by integrating over the priors

$$p(\mathbf{x}|\lambda_X, \mathcal{M}_{\mathbf{X}\to Y}) = \iint p(\mathbf{x}|\mathbf{f}_X)p(\mathbf{f}_X|\mathbf{w}_X, \lambda_X)p(\mathbf{w}_X)d\mathbf{f}_X d\mathbf{w}_X \tag{37}$$

$$= \int p(\mathbf{x}|\mathbf{w}_X, \lambda_X)p(\mathbf{w}_X)d\mathbf{w}_X, \tag{38}$$

where $p(\mathbf{x}|\mathbf{w}_X, \lambda_X) = \mathcal{N}(0, K_{\lambda_X}(\mathbf{w}_X, \mathbf{w}'_X) + \sigma_X^2)$, σ_X^2 is the likelihood noise hyperparameter and K_{λ_X} is the chosen kernel with hyperparameters λ_X. All modelling choices are as laid out in section 5. $p(\mathbf{w}_X)$ is usually chosen to be standard Gaussian distribution. From this, we can see that the dataset density is a mixture of Gaussians, mixed by Gaussian-distributed weights. Similarly, for the conditional model $\mathbf{y}|\mathbf{x}$ we have

$$p(\mathbf{y}|\mathbf{x}, \lambda_Y, \mathcal{M}_{\mathbf{X}\to Y}) = \iint p(\mathbf{y}|\mathbf{x}, \mathbf{f}_Y)p(\mathbf{f}_Y|\mathbf{w}_Y, \lambda_Y)p(\mathbf{w}_Y)d\mathbf{f}_Y d\mathbf{w}_Y \tag{39}$$

$$= \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}_Y, \lambda_Y)p(\mathbf{w}_Y)d\mathbf{w}_Y, \tag{40}$$

where $p(\mathbf{y}|\mathbf{x}, \mathbf{w}_Y, \lambda_Y) = \mathcal{N}(0, K_{\lambda_Y}((\mathbf{x}, \mathbf{w}_Y), (\mathbf{x}, \mathbf{w}_Y)') + \sigma_Y^2)$, and $p(\mathbf{w}_Y)$ is a standard Gaussian. Hence

$$p(\mathbf{x}|\lambda_X, \mathcal{M}_{\mathbf{X}\to Y})p(\mathbf{y}|\mathbf{x}, \lambda_Y, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(\mathbf{x}|\mathbf{w}_X, \lambda_X)p(\mathbf{w}_X)d\mathbf{w}_X \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}_Y, \lambda_Y)p(\mathbf{w}_Y)d\mathbf{w}_Y. \tag{41}$$

Model $\mathcal{M}_{X\leftarrow\mathbf{Y}}$: Similarly, for the causal model $\mathcal{M}_{X\leftarrow\mathbf{Y}}$, we have

$$p(\mathbf{y}|\lambda_X, \mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathbf{x}|\mathbf{y}, \lambda_Y, \mathcal{M}_{X\leftarrow\mathbf{Y}}) = \int p(\mathbf{y}|\mathbf{w}_Y, \lambda_Y)p(\mathbf{w}_Y)d\mathbf{w}_Y \int p(\mathbf{x}|\mathbf{y}, \mathbf{w}_X, \lambda_X)p(\mathbf{w}_X)d\mathbf{w}_X, \tag{42}$$

where

$$\begin{aligned} p(\mathbf{y}|\mathbf{w}_Y, \lambda_Y, \mathcal{M}_{X\leftarrow\mathbf{Y}}) &= \mathcal{N}(0, K_{\lambda_Y}(\mathbf{w}_Y, \mathbf{w}'_Y) + \sigma_Y^2), \\ p(\mathbf{x}|\mathbf{y}, \mathbf{w}_X, \lambda_X, \mathcal{M}_{X\leftarrow\mathbf{Y}}) &= \mathcal{N}(0, K_{\lambda_X}((\mathbf{y}, \mathbf{w}_X), (\mathbf{y}, \mathbf{w}_X)') + \sigma_X^2). \end{aligned} \tag{43}$$

To compare against the model $\mathcal{M}_{\mathbf{X}\to Y}$, we must find the anti-causal factorisation of $\mathcal{M}_{X\leftarrow\mathbf{Y}}$. First, we can see that if we factorise Equation (42), then the priors on the distributions are no longer separable and hence the integrals cannot be estimated separately

$$\int p(\mathbf{y}|\mathbf{w}_Y, \lambda_Y)p(\mathbf{w}_Y)d\mathbf{w}_Y \int p(\mathbf{x}|\mathbf{y}, \mathbf{w}_X, \lambda_X)p(\mathbf{w}_X)d\mathbf{w}_X \tag{44}$$

$$= \iint p(\mathbf{x}|\mathbf{w}_Y, \mathbf{w}_X, \lambda_X, \lambda_Y)p(\mathbf{y}|\mathbf{x}, \mathbf{w}_Y, \mathbf{w}_X, \lambda_X, \lambda_Y)p(\mathbf{w}_Y)p(\mathbf{w}_X)d\mathbf{w}_Y d\mathbf{w}_X. \tag{45}$$

Here,

$$p(\mathbf{x}|\mathbf{w}_X, \mathbf{w}_Y, \lambda_X, \lambda_Y, \mathcal{M}_{X\leftarrow\mathbf{Y}}) = \int p(\mathbf{y}|\mathbf{w}_Y, \lambda_Y)p(\mathbf{x}|\mathbf{y}, \mathbf{w}_X, \lambda_X)d\mathbf{y}, \tag{46}$$

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}_X, \mathbf{w}_Y, \lambda_X, \lambda_Y, \mathcal{M}_{X\leftarrow\mathbf{Y}}) = \frac{p(\mathbf{y}|\mathbf{w}_Y, \lambda_Y)p(\mathbf{x}|\mathbf{y}, \mathbf{w}_X, \lambda_X)}{p(\mathbf{x}|\mathbf{w}_X, \mathbf{w}_Y, \lambda_X, \lambda_Y)}. \tag{47}$$

As the kernel K is generally complicated non-linear function, these terms are not Gaussian distributed, in contrast with the terms of the causal factorisation of $\mathcal{M}_{\mathbf{X}\to Y}$.

Directly comparing the marginals over $\mathbf{x}$ in the two causal models,

$$p(\mathbf{x}|\lambda_X, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(\mathbf{x}|\mathbf{w}_X, \lambda_X)p(\mathbf{w}_X)d\mathbf{w}_X, \tag{48}$$

$$p(\mathbf{x}|\lambda_X, \lambda_Y, \mathcal{M}_{X\leftarrow\mathbf{Y}}) = \iint \left(\int p(\mathbf{y}|\mathbf{w}_Y, \lambda_Y)p(\mathbf{x}|\mathbf{y}, \mathbf{w}_X, \lambda_X)d\mathbf{y}\right)p(\mathbf{w}_Y)p(\mathbf{w}_X)d\mathbf{w}_Y d\mathbf{w}_X,$$

while the term for $\mathcal{M}_{\mathbf{X}\to Y}$ is a mixture of Gaussians, the term for $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ is clearly not a mixture of Gaussians. The same holds for the conditionals

$$p(\mathbf{y}|\mathbf{x},\lambda_Y,\mathcal{M}_{\mathbf{X}\to Y}) = \int p(\mathbf{y}|\mathbf{x},\mathbf{w}_Y,\lambda_Y)p(\mathbf{w}_Y)d\mathbf{w}_Y, \tag{49}$$

$$p(\mathbf{y}|\mathbf{x},\lambda_X,\lambda_Y,\mathcal{M}_{X\leftarrow\mathbf{Y}}) = \iint \frac{p(\mathbf{y}|\mathbf{w}_Y,\lambda_Y)p(\mathbf{x}|\mathbf{y},\mathbf{w}_X,\lambda_X)}{p(\mathbf{x}|\mathbf{w}_X,\mathbf{w}_Y,\lambda_X,\lambda_Y)}p(\mathbf{w}_Y)p(\mathbf{w}_X)d\mathbf{w}_Y d\mathbf{w}_X.$$

As there is no analytical form for the terms of the anti-causal factorisation of $\mathcal{M}_{X\leftarrow\mathbf{Y}}$, we proceed by explicitly showing that the variance of $p(\mathbf{x}|\mathbf{w}_X,\mathbf{w}_Y,\lambda_X,\lambda_Y,\mathcal{M}_{X\leftarrow\mathbf{Y}})$ depends on $\mathbf{w}_X,\mathbf{w}_Y$ for a certain choice of kernel (thus that the term actually depends on $\mathbf{w}_X,\mathbf{w}_Y$). We then contend that it is unlikely that the variance of $p(\mathbf{y}|\mathbf{x},\mathbf{w}_X,\mathbf{w}_Y,\lambda_X,\lambda_Y,\mathcal{M}_{X\leftarrow\mathbf{Y}})$ also does not depend on $\mathbf{w}_X,\mathbf{w}_Y$. If this holds, then the induced prior over $\mathbf{w}_X,\mathbf{w}_Y$ is not separable with respect to the anti-causal factorisation.

Explicit Derivation for RBF kernels. From Equation (47), we can see that both terms of the anti-causal factorisation of $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ depend on $\mathbf{w}_X$, $\mathbf{w}_Y$ and hence the prior over these terms is not separable. That is, while the causal factorisation factorises as Figure 2(b), the anti-causal factorisation does not factorise as Figure 2(a). We can show the explicit dependence on $\mathbf{w}_X$, $\mathbf{w}_Y$ to one of the terms in Equation (47) for certain kernel choices — specifically the ARD RBF kernel. We can find the variance of

$$p(\mathbf{x}|\mathbf{w}_X,\mathbf{w}_Y,\lambda_X,\lambda_Y,\mathcal{M}_{X\leftarrow\mathbf{Y}}) = \int p(\mathbf{y}|\mathbf{w}_Y,\lambda_Y)p(\mathbf{x}|\mathbf{y},\mathbf{w}_X,\lambda_X)d\mathbf{y}. \tag{50}$$

Here (suppressing the λ and model notation for clarity),

$$\mathbb{E}[X^2 \mid W_X, W_Y] = \int p(\mathbf{y}|\mathbf{w}_Y)\int x^2 p(\mathbf{x}|\mathbf{y},\mathbf{w}_X)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \tag{51}$$

$$= \int p(\mathbf{y}|\mathbf{w}_Y)\big[\mathbf{k}(\mathbf{y},\mathbf{y}')\mathbf{k}(\mathbf{w}_X,\mathbf{w}_X') + \sigma_X^2\big]\mathrm{d}\mathbf{y}. \tag{52}$$

We can calculate the kernel expectation

$$\mathbf{\Psi} := \int p(\mathbf{y}|\mathbf{w}_Y)\mathbf{k}(\mathbf{y},\mathbf{y}')\mathrm{d}\mathbf{y}, \tag{53}$$

where $\mathbf{\Psi}\in\mathbb{R}^{N\times N}$. Let $\mathbf{k}_{nm}$ be the 2×2 matrix of the form (covariance of $p(y_n,y_m|w_{Y_n},w_{Y_m})$)

$$\mathbf{k}_{nm} = \begin{pmatrix} k(w_{Y_n},w_{Y_n}) & k(w_{Y_n},w_{Y_m}) \\ k(w_{Y_m},w_{Y_n}) & k(w_{Y_m},w_{Y_m}) \end{pmatrix} = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}. \tag{54}$$

We can write the nm^{th} element of $\mathbf{\Psi}$ as

$$[\mathbf{\Psi}]_{nm} = \iint p(y_n,y_m|w_{Y_n},w_{Y_m})k(y_n,y_m)\mathrm{d}y_n\mathrm{d}y_m \tag{55}$$

$$= \int p(y_m|w_{Y_m})\int p(y_n|y_m,w_{Y_n},w_{Y_m})k(y_n,y_m)\mathrm{d}y_m\mathrm{d}y_n, \tag{56}$$

where

$$p(y_m|w_{Y_m}) = \mathcal{N}(0,k_{22}) \tag{57}$$

$$p(y_n|y_m,w_{Y_n},w_{Y_m}), = \mathcal{N}(k_{12}k_{22}^{-1}y_m, k_{11}-k_{12}k_{22}^{-1}k_{21}). \tag{58}$$

We let $\mu = k_{12}k_{22}^{-1}$ and $\Sigma = k_{11}-k_{12}k_{22}^{-1}k_{21}$ in the following. We can write the terms inside the exponential of $p(y_n|y_m,w_{Y_n},w_{Y_m})k(y_n,y_m)$ as (ignoring the lengthscale of the RBF kernel)

$$\propto \exp\left(-\frac{1}{2}\big((y_n-\mu y_m)^2\Sigma^{-1} + (y_n-y_m)^2\big)\right) \tag{59}$$

$$= \exp\left(-\frac{1}{2}c\right)\exp\left(-\frac{1}{2}\big((y_n-\nu)^2\eta^{-1}\big)\right), \tag{60}$$

where $\eta^{-1} = (\Sigma^{-1} + 1)$, $\nu = \eta(\mu y_m \Sigma^{-1} + y_m)$, and $c = \eta(y_m^2 \mu^2 \Sigma^{-1} + y_m^2 \Sigma^{-1} - 2\mu y_m^2 \Sigma^{-1})$. This gives that

$$[\mathbf{\Psi}]_{nm} = \frac{\sqrt{\eta}}{|\mathbf{k}_{nm}|^{0.5}} \int \exp\left(-\frac{1}{2}c\right) p(y_m | w_{Y_m}) \mathrm{d}y_m. \tag{61}$$

We again expand the terms inside the exponential of the integrand,

$$\propto \exp\left(-\frac{1}{2} y_m^2 (k_{22}^{-1} + \eta\mu^2\Sigma^{-1} + \eta\Sigma^{-1} - 2\mu\eta\Sigma^{-1})\right). \tag{62}$$

Integrating this gives

$$[\mathbf{\Psi}]_{nm} = \frac{\sqrt{2\eta}}{|\mathbf{k}_{nm}|^{0.5} \sqrt{k_{22}^{-1} + \eta\mu^2\Sigma^{-1} + \eta\Sigma^{-1} - 2\mu\eta\Sigma^{-1})}}. \tag{63}$$

Hence, we can explicitly see that the variance of $p(\mathbf{x}|\mathbf{w}_X, \mathbf{w}_Y, \lambda_X, \lambda_Y, \mathcal{M}_{X \leftarrow \mathbf{Y}})$ depends on both $\mathbf{w}_Y$ and $\mathbf{w}_X$. The variance of $p(\mathbf{y}|\mathbf{x}, \mathbf{w}_X, \mathbf{w}_Y, \lambda_X, \lambda_Y, \mathcal{M}_{X \leftarrow \mathbf{Y}})$ is much harder to calculate. We view it as highly unlikely that it would marginally depend on neither of $\mathbf{w}_Y$ and $\mathbf{w}_X$. If it depends on either $\mathbf{w}_Y$ or $\mathbf{w}_X$, then the anti-causal factorisation shares parameters, and the prior is thus unlikely to be separable in the anti-causal direction.

E. Derivations for section 4.3

E.1. Derivation of the probability of error of both models being the same under similar priors

Here we show that the probability of error of both the models is exactly the same. Assuming a parametrisation and denoting parameters as ϕ for X and θ for Y, we get

$$p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X} \to Y}) = \int p(\mathbf{x}|\phi, \mathcal{M}_{\mathbf{X} \to Y}) \pi(\phi|\mathcal{M}_{\mathbf{X} \to Y}) \mathrm{d}\phi \int p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}_{\mathbf{X} \to Y}) \pi(\theta|\mathcal{M}_{\mathbf{X} \to Y}) \mathrm{d}\theta\,, \tag{64}$$

$$p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X \leftarrow \mathbf{Y}}) = \int p(\mathbf{y}|\theta, \mathcal{M}_{X \leftarrow \mathbf{Y}}) \pi(\theta|\mathcal{M}_{X \leftarrow \mathbf{Y}}) \mathrm{d}\phi \int p(\mathbf{x}|\mathbf{y}, \phi, \mathcal{M}_{X \leftarrow \mathbf{Y}}) \pi(\phi|\mathcal{M}_{X \leftarrow \mathbf{Y}}) \mathrm{d}\theta\,. \tag{65}$$

The marginal and conditional densities in the above are chosen from the same families. If we assume similar priors on the cause and effect in both models, then $\pi(\phi|\mathcal{M}_{\mathbf{X} \to Y}) = \pi(\theta|\mathcal{M}_{X \leftarrow \mathbf{Y}})$ and $\pi(\theta|\mathcal{M}_{\mathbf{X} \to Y}) = \pi(\phi|\mathcal{M}_{X \leftarrow \mathbf{Y}})$. Thus, $p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X} \to Y}) = p(\mathbf{y}, \mathbf{x}|\mathcal{M}_{X \leftarrow \mathbf{Y}})$. That is, swapping $\mathbf{x}$ and $\mathbf{y}$ in one model will give the dataset density of the other model. We can use this to show

$$P(E|\mathcal{M}_{\mathbf{X} \to Y}) = \int_{\mathcal{R}} p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X} \to Y}) \mathrm{d}(\mathbf{x}, \mathbf{y})\,, \qquad \mathcal{R} = \{(\mathbf{x}, \mathbf{y}) \mid p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X \leftarrow \mathbf{Y}}) > p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X} \to Y})\} \tag{66}$$

$$= \int_{\mathcal{R}} p(\mathbf{y}, \mathbf{x}|\mathcal{M}_{X \leftarrow \mathbf{Y}}) \mathrm{d}(\mathbf{x}, \mathbf{y})\,, \qquad \mathcal{R} = \{(\mathbf{x}, \mathbf{y}) \mid p(\mathbf{y}, \mathbf{x}|\mathcal{M}_{\mathbf{X} \to Y}) > p(\mathbf{y}, \mathbf{x}|\mathcal{M}_{X \leftarrow \mathbf{Y}})\} \tag{67}$$

$$= P(E|\mathcal{M}_{X \leftarrow \mathbf{Y}}). \tag{68}$$

E.2. Upper bound on standard deviation of probability of error

The probability of error is

$$P(\text{Error}) = P(\text{Error}|\mathcal{M}_{\mathbf{X} \to Y}) \tag{69}$$

$$= \int \mathbb{I}[\mathcal{D} \in \mathcal{R}] \cdot p(\mathcal{D}|\mathcal{M}_{\mathbf{X} \to Y}) \mathrm{d}\mathcal{D} \tag{70}$$

$$\hat{I} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}[\mathcal{D}_t \in \mathcal{R}], \;\; \mathcal{D}_t \sim p(\mathcal{D}|\mathcal{M}_{\mathbf{X} \to Y}), \tag{71}$$

where $\mathbb{I}[\cdot]$ is the indicator function and $\hat{I}$ is the Monte Carlo estimator of $P(\text{Error})$. We can bound the variance by using the fact that $\text{Var}(\mathbb{I}[\mathcal{D}_t \in \mathcal{R}]) \leq 0.25$, upon viewing $\mathbb{I}[\cdot]$ as a Bernoulli random variable. We hence obtain that

$$\text{Var}(\hat{I}) \leq \frac{1}{4T}. \tag{72}$$

This can be calculated numerically.

E.3. Derivation of total probability of error

The total probability of error is (see Nielsen (2014) for more details)

$$P(E) = P(E|\mathcal{M}_{\mathbf{X}\to Y})P(\mathcal{M}_{\mathbf{X}\to Y}) + P(E|\mathcal{M}_{X\leftarrow\mathbf{Y}})P(\mathcal{M}_{X\leftarrow\mathbf{Y}}) \tag{73}$$

$$= \int_{\mathcal{R}_Y} p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})\mathrm{d}\mathcal{D} + \int_{\mathcal{R}_X} p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})\mathrm{d}\mathcal{D} \tag{74}$$

$$= \int \min(p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y}), p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}}))\mathrm{d}\mathcal{D}, \tag{75}$$

where $\mathcal{R}_Y = \{\mathcal{D} \mid p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\}$, and $\mathcal{R}_X = \{\mathcal{D} \mid p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) < p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\}$. We now use $\min(a,b) = \frac{a+b-|b-a|}{2}$ to obtain

$$P(E) = \frac{1}{2} - \frac{1}{2}\int |p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) - p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})|\mathrm{d}\mathcal{D} \tag{76}$$

$$= \frac{1}{2}(1 - \mathrm{TV}[P(\cdot|\mathcal{M}_{\mathbf{X}\to Y}), P(\cdot|\mathcal{M}_{X\leftarrow\mathbf{Y}})]). \tag{77}$$

E.4. Misspecification

We can write

$$|Q(E) - P(E)| = \frac{1}{2}|\mathrm{TV}(P(\cdot|\mathcal{M}_{\mathbf{X}\to Y}), P(\cdot|\mathcal{M}_{X\leftarrow\mathbf{Y}})) - \mathrm{TV}(Q(\cdot|\mathcal{M}_{\mathbf{X}\to Y}), Q(\cdot|\mathcal{M}_{X\leftarrow\mathbf{Y}})| \tag{78}$$

$$= \frac{1}{2}\left|\int (|p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) - p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})| - |q(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) - q(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})|)\mathrm{d}\mathcal{D}\right| \tag{79}$$

$$\leq \frac{1}{2}\left|\int (|p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) - q(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) + q(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) - p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})|)\mathrm{d}\mathcal{D}\right| \tag{80}$$

$$\leq \frac{1}{2}\left|\int (|p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) - q(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})| + |q(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) - p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})|)\mathrm{d}\mathcal{D}\right| \tag{81}$$

$$= \frac{1}{2}|\mathrm{TV}(P(\cdot|\mathcal{M}_{\mathbf{X}\to Y}), Q(\cdot|\mathcal{M}_{\mathbf{X}\to Y})) + \mathrm{TV}(P(\cdot|\mathcal{M}_{X\leftarrow\mathbf{Y}}), Q(\cdot|\mathcal{M}_{X\leftarrow\mathbf{Y}}))|. \tag{82}$$

Where we use the reverse triangle inequality and the triangle inequality. We can get rid of the absolute value by noticing that TV is bounded by 0 and 1.

F. Model Details

Here, we provide a more in depth introduction to our model and approximations.

F.1. Latent variable Gaussian Processes with inducing points

Gaussian processes (GPs) (Rasmussen, 2003) are non-parametric Bayesian models that define a prior over functions. The form of the prior is controlled by choice of a kernel function, $\mathbf{K}$. Specifically, the kernel defines a covariance over outputs for the function. The kernels are parametrised by continuous hyperparameters. Adding kernels together and varying their hyperparameters allows for construction of flexible priors which support a wide range of functions.

Latent variable Gaussian Processes (GPLVM) consider a latent noise term $\mathbf{w}$ as an input with an associated prior. Integrating over the noise term allows for modelling heteroscedastic noise as well as non-Gaussian likelihoods. GPs have a well-known computational cost of $\mathcal{O}(N^3)$ where N is the number of samples, which prohibits their direct application to large datasets. To allow for scalability, we use an inducing point approximation (Titsias, 2009; Hensman et al., 2013) to the posterior. Here, we approximate the inputs $\mathbf{x}$ and latents $\mathbf{w}$ with $M < N$ 'inducing' inputs $\mathbf{z}$, and their corresponding outputs with $\mathbf{u}$. This formulation now has a cost of $\mathcal{O}(M^3)$.

We collectively denote all hyperparameters of the model with $\boldsymbol{\lambda}$. The latent variable Gaussian process for modelling

$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda})$ has the following form

$$\begin{aligned} &\mathbf{y}|\mathbf{x}, \mathbf{w} \sim \mathcal{N}(\mathbf{f}(\mathbf{x}, \mathbf{w}), \sigma^2), \quad &&\mathbf{F}|\mathbf{u} \sim \mathcal{N}(\mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}\mathbf{u}, \boldsymbol{\Sigma}), \\ &\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{K_{uu}}), &&[\mathbf{K_{ff}}]_{nn'} = \mathbf{K}(x_n, x_{n'}), \end{aligned}$$

where $\Sigma = \mathbf{K_{ff}} - \mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}\mathbf{K_{uf}}$, and σ denotes the likelihood noise hyperparameter. We posit an approximate posterior over the latents $q(\mathbf{w})$ and inducing outputs $q(\mathbf{u})$ following the variational inference framework. This yields a lower bound to the log marginal likelihood that can then be maximised with respect to the variational distributions q and the inducing inputs $\mathbf{z}$. The lower bound for the conditional $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda})$ has the form

$$\mathcal{L}_{\mathbf{y}|\mathbf{x}}(q, \mathbf{z}, \boldsymbol{\lambda}) := \sum_n \mathbb{E}_{q(w_n)q(\mathbf{u})p(\mathbf{f}|\mathbf{u})}[\log p(y_n|x_n, w_n)] + KL[q(\mathbf{w})||p(\mathbf{w})] + KL[q(\mathbf{u})||p(\mathbf{u})]. \tag{83}$$

For smaller datasets, we follow the steps in Titsias and Lawrence (2010) and analytically integrate over $p(\mathbf{f}|\mathbf{u})$ first and then find the closed form solution for $q(\mathbf{u})$ — denoted GPLVM-closed form. For certain kernels, for example the RBF and linear kernels, we can analytically find the expectation under $q(w_n)$ of the remaining first term in Equation (83). For larger datasets, we need to use stochastic variational inference (Hensman et al., 2013) as finding the closed form solutions are prohibitive — we denote this GPLVM-stochastic. This requires using doubly stochastic variational inference to calculate the expectations (Lalchand et al., 2022). We assume a standard Gaussian distribution for $p(\mathbf{w})$ and a Gaussian for $q(\mathbf{w})$, with variational parameters to be trained. The final KL term is between two Gaussians and is analytically tractable. We use the evidence approximation for the hyperparameters, which we detail in app. G.

The bound for all the marginal and conditional models, $p(\mathbf{y}|\mathbf{x})$, $p(\mathbf{x})$, $p(\mathbf{x}|\mathbf{y})$, and $p(\mathbf{y})$, follows the form of Equation (83). We assume the same kernels and priors for all the models as well.

F.2. Final Score

To model each distribution, we maximise the lower bound with respect to the variational distributions q and inducing inputs $\mathbf{z}$ to tighten the bound. Simultaneously, we maximise the lower bound with respect to the kernel hyperparameters and the likelihood noise, collectively denoted as $\boldsymbol{\lambda}$ (see app. G). The final scores we calculate are

$$\mathcal{F}_{X\to Y} = \mathcal{L}_{\mathbf{x}}(\hat{q}, \hat{\mathbf{z}}, \hat{\boldsymbol{\lambda}}) + \mathcal{L}_{\mathbf{y}|\mathbf{x}}(\hat{q}, \hat{\mathbf{z}}, \hat{\boldsymbol{\lambda}}), \tag{84}$$

$$\mathcal{F}_{Y\to X} = \mathcal{L}_{\mathbf{y}}(\hat{q}, \hat{\mathbf{z}}, \hat{\boldsymbol{\lambda}}) + \mathcal{L}_{\mathbf{x}|\mathbf{y}}(\hat{q}, \hat{\mathbf{z}}, \hat{\boldsymbol{\lambda}}), \tag{85}$$

where $(\hat{q}, \hat{\mathbf{z}}, \hat{\boldsymbol{\lambda}})$ denote the values the maximise the corresponding lower bound. We finally infer the predicted causal model as $\mathcal{M}_{X\to Y}$ if $\mathcal{F}_{X\to Y} > \mathcal{F}_{Y\to X}$, $\mathcal{M}_{Y\to X}$ if $\mathcal{F}_{Y\to X} > \mathcal{F}_{X\to Y}$, and undecided otherwise.

G. Justifying MAP estimation of Hyperparameters

In this section, we give details on the approximation we consider to integrate over the prior over hyperparameters. This is standard practice in Gaussian process training (Titsias, 2009; Damianou and Lawrence, 2013; Dutordoir et al., 2018; Rasmussen, 2003).

We need to integrate out all model hyperparameters to get an accurate value of the marginal likelihood. Otherwise, the actual quantity being compared is the posterior of the model with a specific hyperparameter setting. Due to non-linearity of kernels, the integral over priors over hyperparameters tend to be intractable for our method, the GPLVM. As such, we use the Laplace Approximation to approximate these integrals (MacKay, 1999). Taking the conditional $p(\mathbf{y}|\mathbf{x})$ as an example (leaving out terms in the conditional for simplicity), we wish to calculate

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x}, \lambda)p(\boldsymbol{\lambda})d\boldsymbol{\lambda}, \tag{86}$$

where $p(\boldsymbol{\lambda})$ is a prior over the hyperparameters. The justification for this approximation is that integral in Equation (86) is simply the normalising constant of the posterior over the hyperparameters. This posterior tends to be highly peaked, and even more so as the number of datapoints increases and the number of hyperparameters are few (Rasmussen, 2003). Thus, as most of the volume is around the MAP solution of the posterior, we can assume a Gaussian distribution around this point and

approximate the integral Equation (86) as the normalising constant of this Gaussian. The approximation of Equation (86) is

$$\log p(\mathbf{y}|\mathbf{x}) \approx \log p(\mathbf{y}|\mathbf{x}, \hat{\boldsymbol{\lambda}})p(\hat{\boldsymbol{\lambda}})\left|\frac{1}{2\pi}\mathbf{A}\right|^{-\frac{1}{2}}, \tag{87}$$

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda}), \tag{88}$$

$$\mathbf{A} = -\nabla^2_{\boldsymbol{\lambda}} \log p(\mathbf{y}, \boldsymbol{\lambda}|\mathbf{x}). \tag{89}$$

We ignore the value $p(\hat{\boldsymbol{\lambda}})$ as we don't actually take a prior over the hyperparameters, this can be thought of assuming the same density over all hyperparameter values (Rasmussen, 2003). We further ignore $\mathbf{A}$; this is based on the expectation that in the large sample limit, the posterior of the hyperparameters will concentrate around a single point. We can thus safely ignore $\mathbf{A}$, and simply approximate $\log P(\mathbf{y}|\mathbf{x}) \approx \log P(\mathbf{y}|\mathbf{x}, \hat{\boldsymbol{\lambda}})$. We find that this works well in practice.

It is possible to 'overfit' with this approximation (Ober et al., 2021). This is not a major issue in our chosen model, as the number of hyperparameters is very low compared to the number of data samples.

As we lower bound $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda})$ using $\mathcal{L}_{\mathbf{y}|\mathbf{x}}$ in Equation (83), we can carry out the procedure described above by considering $\mathcal{L}_{\mathbf{y}|\mathbf{x}}$ instead. Thus, our approximation of the integral over the hyperparameters will involve finding the values of the hyperparameters $\hat{\boldsymbol{\lambda}}$ that maximise $\mathcal{L}_{\mathbf{y}|\mathbf{x}}$ and using $\mathcal{L}_{\mathbf{y}|\mathbf{x}}(q, \hat{\boldsymbol{\lambda}})$.

H. Experiment Details

We outline the experimental details of our method. As we implement SLOPPY, we also outline the details of the settings we used. The results for the rest of the baselines were taken from (Guyon et al., 2019) and (Duong and Nguyen, 2021).

H.1. Dataset details

CE-Cha: A mixture of synthetic and real world data. Taken from the cause-effect pairs challenge (Guyon et al., 2019).

CE-Multi (Goudet et al., 2018): Synthetic data with effects generated with varying noise relationships. The noise relationships are pre-additive $(f(X + E))$, post-additive $(f(X) + E)$, pre-multiplicative $(f(X \times E))$, or post-multiplicative $(f(X) \times E)$. The function is linear or polynomial.

CE-Net (Goudet et al., 2018): Synthetic data with randomly initialised neural networks for functions and random exponential family distributions chosen for the cause.

CE-Gauss (Mooij et al., 2016): Synthetic data generated with random noise distributions E_1, E_2 defined in (Mooij et al., 2016). The cause and effect are generated according to $X = f_x(E_1)$ and $Y = f_y(X, E_2)$, where f_x, f_y are sampled from Gaussian processes.

CE-Tueb (Mooij et al., 2016): Contains 105 pairs of real cause effect pairs taken from the UCI dataset. We use the version dating August 22, 2016. We also remove high dimensional datasets leaving 99 datasets in total.

H.2. GPLVM details

We use the GPLVM-closed form for all datasets except for CE-Tueb where we GPLVM-stochastic due to the high number of variables in some of the datasets.

For GPLVM-closed form, we use the sum of an RBF and linear kernels. $q(w_n)$ has an analytical expectation for these kernels, as discussed in app. F.1. As detailed in app. F.1, we find the optimal form of $q(\boldsymbol{u})$ following the procedure of (Titsias and Lawrence, 2010). We use 200 inducing points for all experiments. The model was first trained using Adam with a learning rate of 0.1. After 20,000 epochs, the model was trained using BFGS. We found that this greatly helped the numerical instability of BFGS, but found better ELBO (variational approximation to the marginal likelihood) values than simply using Adam.

For GPLVM-stochastic, as expectations are calculated by sampling, it was possible to use a larger number of kernels. We used a sum of RBF, Linear, Matern32 and Rational Quadratic kernels. 10 samples were used to calculate the expectations. The model was trained with Adam with a learning rate of 0.05. The model stopped training if the value of the ELBO plateaued, else it ran for a maximum of 100,000 epochs. In our experiments, we only use GPLVM-stochastic for CE-Tueb

as it had a few datasets that had a large number of samples.

The approximate posteriors for both the models $q(w_n)$ are initialised with low variance and the mean equal to 0.01 times the output. This reduced the instability during optimisation.

As GPLVMs are known to suffer from local optima issues, we use 20 random restarts of hyperparameter initialisations, and choose the highest estimate of the approximate marginal likelihood as the final score. For the various hyperparameters, the sampling procedures were:

1. The kernel variances were always set to 1.
2. The likelihood variances were sampled by first sampling $\kappa \sim \text{Uniform}(10, 100)$, and then $\sigma^2_{\text{Likelihood}} = 1/\kappa^2$.
3. The kernel lengthscales were sampled by first sampling $\psi \sim \text{Uniform}(1, 100)$, then set $\lambda_{\text{Lengthscale}} = 1/\psi$.

H.3. SLOPPY details

For benchmarking the SLOPPY method, we use the author's code (Marx and Vreeken, 2019). We use the spline estimator as it performs better on all the dataset. For this estimator, we select the best performing regularisation metric between the AIC and BIC.

I. Additional Experiments

We carry out some additional experiments that give us insight into our method. In app. I.1, we show that the GPLVM performs well on ANM data, despite being more fleixble than an ANM. In app. I.2, we show that importance of modelling the joint instead of just the conditional or marginal densities.

I.1. ANM Data

Table 2: ROC AUC scores for identifying causal direction of datastes generated by an ANM (higher is better).

Methods	ANM
Gaussian Process	100.0
GPLVM	100.0

ANM is an example of a strictly identifiable model. Here we show that our added flexibility does not result in a loss in performance when compared to identifiable models.

The GPLVM model contains the hyperparameters $\boldsymbol{\lambda}$ of the GP priors, which makes the GPLVM a hierarchical model. Depending on which $\boldsymbol{\lambda}$ is inferred, the GPLVM can learn to behave in different ways. For example, for datasets that follow ANM assumptions, the effect of the latent variable w_i can be ignored, making the model behave as an ANM. In this section, we show that the added flexibility of the GPLVM model (over ANM) does not lead to a reduction in performance when tested on data from an ANM.

We use datasets generated from an ANM (taken from Tagasovska et al. (2020)). A straightforward Gaussian process (GP) model satisfies the conditions of an ANM which have been shown to identify causal direction using the likelihood only (Zhang et al., 2015). Table 2 shows that the marginal likelihood also perfectly identifies causal direction. Furthermore, even though a GPLVM is a more flexible model than a GP, the added flexibility does not suffer from a loss of performance.

I.2. Only modelling the conditional or marginal

Methods such that ANM, PNL, SLOPPY, RECI only model the conditional densities to find the causal direction. In fact, methods such as SLOPPY base their theory on modelling the joint, but make the assumption that the cause is always Gaussian distributed, and hence only consider the conditional. The Kolmogorov complexity formalisation of the ICM principle (Janzing and Schölkopf, 2010; Peters et al., 2016) also considers the whole joint.

We show that modelling the joint is crucial to our approach, and that we suffer a degradation in performance when only considering one component - the marginal or conditional. In Table 3, we show that results of the same model, but making

the decision on the predicted causal model with the joint, the conditional, or with the marginal. The results corroborate with our theory.

Table 3: Results of making the decision on the predicted causal model with the full joint, or just with the marginal or conditional densities. In accordance with our theory, modelling the joint is important. The numbers are ROCAUC (higher is better).

Methods	CE-Cha	CE-Multi	CE-Net	CE-Gauss	CE-Tueb
GPLVM - Joint	81.9	97.7	98.9	89.3	78.3
GPLVM - Conditional	61.5	89.7	70.3	21.7	36.2
GPLVM - Marginal	44.5	23.3	42.6	83.1	75.7