Causal Discovery using Marginal Likelihood

Anonymous Author(s)
Affiliation
Address
email

Abstract

Causal discovery is an important problem in many fields such as medicine, epidemiology, or economics. Here, causal structure is necessary to relay information about the effectiveness of treatments. Recently, causal structure has also been linked with generalisation and out of distribution generalisation in prediction tasks. This problem however, is only solvable upto a Markov equivalence class without strong assumptions. Previous work has made assumptions on the data generation process to render the causal graph identifiable. These methods fail when the data generation assumptions no longer hold. In this work, we directly algorithmise the independence of causal mechanism (ICM) assumption to achieve a flexible causal discovery algorithm. In the bivariate case, this is done by showing that independent parametrisation with independent priors encodes an ICM assumption. We show that this implies different marginal likelihoods for models of different causal directions. Using a Bayesian model selection procedure to take advantage of this, we show that our method outperforms competing methods.

1 Introduction

Having access to a causal structure allows for answering questions beyond predictions, opening the possibilities to answer interventional questions [31] with only observational data, that is, data where no interventions have taken place [3, 36]. Knowing the causal structure also has consequences for prediction. Conditional distributions corresponding to the causal generative process remain invariant as other variables in the system are intervened on [32, 1]. This is a particularly useful for domain adaptation [41, 7, 2], but also impacts the robustness [6, 23], and adaptation speed under distributional shifts [4, 34] of machine learning models. It is also possible to take advantage of these properties when causal variables are not given by learning causal representations [35].

Causal relations can be inferred reliably from interventional data, however obtaining this can be financially burdensome or ethically problematic. This motivates the need to learn causal relations using observational data. In this regime, conditional independences can only recover a causal structure upto its Markov equivalence class [31]. However, identifying the causal structure within an equivalence class is necessary to take advantage of causal insights. For example, while the causal structures $X \to Y$, and $Y \to X$ are in the same Markov equivalence class, an intervention on any of the variables in these two graphs will have different causal conclusions.

Previous methods make assumptions on the noise distribution or functions to identify the causal direction. These assumptions may not always hold in practice. *Independence of causal mechanisms* (ICM) has been proposed as a foundational principle for causal discovery [20]. This states that the conditional distributions corresponding to the causal generative process are mutually independent. This implies that a change in any one of these distributions, should leave the rest invariant. For example, if the altitude A is the cause and the temperature T is the effect, changing the distribution of altitudes $p(A)$ will not change how the altitude effects the temperature, $p(T|A)$. Although this is an assumption on the data generating process, it does not restrict the functions or noise distributions that induce the joint.

In this work, we tackle the problem of learning bivariate causal relations with access to observational data by using Bayesian model selection [27]. We show that models with causal direction can encode an ICM condition if the causal conditionals have separate parametrisations and the parameters have independent priors. This retains the intuitive implications of the ICM principle. We show that the models cannot be reversed, in general, in a way where ICM holds in the reverse direction. Hence, with the same priors, our models of causal direction imply different densities over datasets. Given that a dataset is generated according to our model, the correct causal model will obtain a higher log *marginal likelihood* that allows for identification of the causal model. To calculate the marginal likelihood, we use Gaussian Process latent variable models that allows us to model flexible densities [40, 10, 24]. We test our method with synthetic and real data that encode different data generation assumptions. Our results show that we not only out perform methods that make specific data generating assumptions, but also other methods inspired by the ICM principle.

2 Related Work

One direction of work has relied on strict assumptions about the data generation process to infer the causal structure within an equivalence class. [18] shows that causal discovery is possible when the functional relationship between cause and effect is non linear and the effect has additive noise (ANM). It is not possible in general to reverse a causal model with these assumptions and stay within the model class. This intuition has been extended to post non linear noise relationships (PNL) [42]. Assuming a linear relationship with non Gaussian noise allows for recovery of the independent noise terms. This procedure, known as LiNGAM [37, 38], relies on the statistical dependences between these recovered noise terms and the cause and effect. This has also been extended to non linear relationships [29] by using non linear ICA [19]. RECI [5] assumes low effect noise and shows that the test error can identify the causal direction.

Attempts to formalise the ICM principle have depended on algorithmic information theory [13, 20], with work even showing that this may subsume other methods of causal discovery [21]. This relies on the notion of Kolmogorov complexity [25], with the implication that the complexity of the causal factorisation is minimal. The Kolmogorov complexity however, is uncomputable in general. Other ways of viewing this principle attempt to alleviate the problem of uncomputability. IGCI [8] tries to infer the dependence between the mechanisms by using information geometry. A near zero measure of dependence infers the causal direction. However, this requires a low effect noise assumption as well as invertibility of the cause to effect function. CGNN [12] try to learn a generative model of the data with competing causal generative structures. With limited complexity, the causal direction with independent components should be easier to fit than the anticausal direction. Overfitting is tackled by using validation datasets as increasing the complexity will lead to an equally good fit in both directions. CDCI [9] tries to measure the complexity of the conditional distributions by measuring the stability of the conditional under different input values. Similar attempts have been made by using the norm of kernel mean embeddings to define variability [28]. The method closest to ours is the GPI [39]. They extend previous methods using Gaussian processes [11] to using latent variable Gaussian processes. Our method differs from theirs as we show that there is an explicit model asymmetry, and our approximation for the marginal likelihood leads to empirically better results.

3 Preliminaries

In this section we describe the *Structural Causal Model* (SCM) [31], which is the main framework we utilise. We also outline the main assumption underpinning this work, mainly that of *Independent Causal Mechanisms* (ICM) [20].

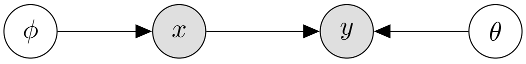


Figure 1: Graphical model for the model $\mathcal{M}_{X\to Y}$. The causal direction indicates which factorisation has independent parameters.

3.1 Structural Causal Models (SCM)

In the bivariate case, and with the following data generating process, we say that X causes Y, written as $X \to Y$:

$$\begin{aligned} X :=& f_x(N_x), \\ Y :=& f_y(X, N_y), \end{aligned} \tag{1}$$

where N_x and N_y are independent noise variables that are sampled from some arbitrary distribution. The equations above induce a joint probability and we refer to the factorisation corresponding to terms in the SCM ($P(X)P(Y|X)$ in the above) as the *causal factorisation*. The factorisation found by applying Bayes rule to the causal factorisation is referred to as the *anticausal factorisation*.

3.2 Independent causal mechanisms (ICM)

This assumption follows directly from the form of the SCM. Assuming there are no confounders, the ICM assumption states that the distribution of the cause, P_{Cause}, and the distribution of the effect given the cause, $P_{\text{Effect|Cause}}$, are independent. These two components are independent in the sense that a change in one of them leaves the other invariant. Changes in the distribution in the SCM correspond to changing either the functions or the noise terms. Hence changing f_x or N_x in equation 1 will result in a change in the distribution $P(X)$, but will leave the form of $P(Y|X)$ invariant. This is because $P(Y|X)$ is only determined by f_y and N_y. Shifting the values of X will change the values of Y that are observed, but not the distribution $P(Y|X)$ itself. This intuition does not necessarily hold for the anticausal factorisation. As $P(Y) = \int P(Y|X)P(X)dX$ and $P(X|Y) \propto P(Y|X)P(X)$, we can easily see that changing f_x or N_x can result in a change in both $P(Y)$ as well as $P(X|Y)$. This is a fundamental asymmetry implied by assuming that variables are generated by an SCM. ICM is a flexible assumption for causal discovery in the sense that no assumptions on the functional or noise terms of the SCM have been made.

4 Causal discovery using marginal likelihood

We cast the problem of causal discovery in the bivariate case as a Bayesian model selection problem. We show that a model that directly parametrises an ICM condition has an asymmetry in its causal and anticausal factorisation, depdendent on the choice of the prior. Bayesian model selection then gives us the assurance that if the data is generated according to a model, it will have a higher log *marginal likelihood* than the competing model.

4.1 Asymmetry between causal and anticausal models

In the bivariate case, causal discovery can be reframed as a model selection problem between two models, $\mathcal{M}_{X\to Y}$ and $\mathcal{M}_{Y\to X}$. The arrow in the model subscript indicates the causal direction that the model postulates. We directly parametrise the causal factorisation of each model, with a prior over the parameters. We assume the same parametrisation and same priors for both the causal models. Figure 1 shows the graphical model for the corresponding model $\mathcal{M}_{X\to Y}$. The joint factorises into the causal factorisation as the following here,

$$P(x, y, \theta, \phi|\mathcal{M}_{X\to Y}) = P(y|x, \theta, \mathcal{M}_{X\to Y})P(x|\phi, \mathcal{M}_{X\to Y})P(\theta|\mathcal{M}_{X\to Y})P(\phi|\mathcal{M}_{X\to Y}). \tag{2}$$

The below analysis always considers the model $\mathcal{M}_{X\to Y}$ and we leave out $\mathcal{M}_{X\to Y}$ from here on for succinctness. As we don't observe ϕ and θ, the observed distribution for the conditional and marginal postulated by this model is the following

$$\begin{aligned} P(x, y) &= P(y|x)P(x) && (3) \\ &= \int P(y|x, \theta)P(\theta)d\theta \int P(x|\phi)P(\phi)d\phi && (4) \end{aligned}$$

ICM here is encoded in equation 4 as the two components $P(y|x,\theta)$ and $P(x|\phi)$ have different parameters, and the parameters have independent priors, $P(\phi,\theta) = P(\phi)P(\theta)$. Effectively these conditions imply that the distributions $P(y|x)$ and $P(x)$ are independent in the sense discussed in section 3.2; changing the distribution of θ and hence $P(y|x)$ will not effect $P(x)$, and vice versa.

We are interested in the case where ICM holds in the anticausal direction. This is interesting as it may lead to cases where two causal models that postulate ICM in different causal direction, end up implying the same distribution over the joint. This analysis closely follows [16, 15] where it is used as a starting assumption, but for our case guides identifiability.

Theorem 4.1 *Assume a given a model $\mathcal{M}_{X\to Y}$. Assume that the model factorises as figure 1. if there exists an $\eta := f_1(\theta,\phi)$ and $\gamma := f_2(\theta,\phi)$, such that*

$$\int\int P(x|y,\theta,\phi,\mathcal{M}_{X\to Y})P(y|\theta,\phi,\mathcal{M}_{X\to Y})P(\theta|\mathcal{M}_{X\to Y})P(\phi|\mathcal{M}_{X\to Y})d\theta d\phi \tag{5}$$

$$= \int P(x|y,\eta,\mathcal{M}_{X\to Y})P(\eta|\mathcal{M}_{X\to Y})d\eta \int P(y|\gamma,\mathcal{M}_{Y\to X})P(\gamma|\mathcal{M}_{X\to Y})d\gamma. \tag{6}$$

Then the following are true:

1. *For every (θ,ϕ), $P(x|y,\theta,\phi) = P(x|y,\eta)$, and $P(y|\theta,\phi) = P(y|\gamma)$.*
2. *The implied priors*

$$P(\eta,\gamma) = P(\theta)P(\phi)\left|\begin{bmatrix} \frac{\partial f_1(\theta,\phi))}{\partial\theta} & \frac{\partial f_1(\theta,\phi)}{\partial\phi} \\ \frac{\partial f_2(\theta,\phi)}{\partial\theta} & \frac{\partial f_2(\theta,\phi)}{\partial\phi}\end{bmatrix}\right|^{-1} \tag{7}$$

   are independent.

If the above is true, we say that the anticausal factorisation of a causal model satisfies ICM.

The proof of this is in appendix B.1. The constraint in condition 1 in the above is required so that the anticausal factorisation can be expressed using the chosen independent parametrisation. The implied priors over the parameters also need to be independent, this is trivially true if the Jacobian in equation 7 is diagonal. If these two conditions hold, the observed anticausal densities can be split into separate integrals in the same way as equation 4.

If ICM does not hold in the anticausal direction for a causal model, two causal models with opposite causal directions will imply different densities over the data for any choice of priors.

Theorem 4.2 *Assume two given causal models $\mathcal{M}_{X\to Y}$ and $\mathcal{M}_{Y\to X}$. If the anticausal factorisation of $\mathcal{M}_{X\to Y}$ and $\mathcal{M}_{Y\to X}$ do not satisfy ICM, then there exist x, y such that*

$$P(x,y|\mathcal{M}_{X\to Y}) \neq P(x,y|\mathcal{M}_{Y\to X}). \tag{8}$$

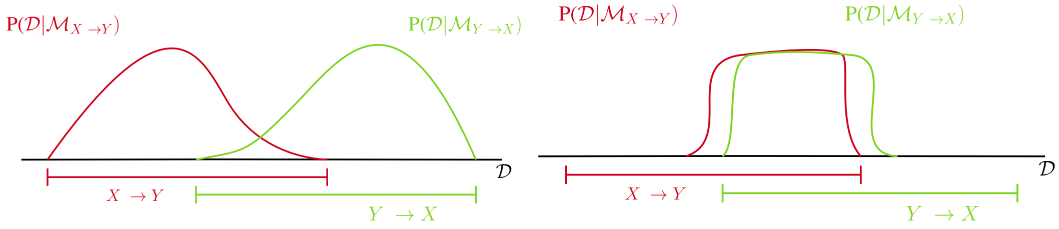
The proof of the above is in appendix B.2. An important insight is that for the densities to be the same, we usually require different priors on the two causal models.

To summarise, a causal model postulates the direction in which ICM holds. This implies independent priors and an overall structure as shown in figure 1. Hence, two models, that postulate ICM in opposite directions, will imply different densities on a dataset. In general, the above shows that $P(x,y|\mathcal{M}_{X\to Y})$ does not equal $P(x,y|\mathcal{M}_{Y\to X})$ for all x, y. We can use this insight to design a Bayesian model selection procedure such that data generated by a causal model is likely to have a higher marginal likelihood under the true causal model. Furthermore, the above also gives us conditions when we would expect the Bayesian model selection procedure to fail to distinguish between causal directions.

4.2 Causal Discovery as Model selection

Based on the previous section, we propose using Bayesian model selection [22, 27] to select between causal models. We denote the models in the bivariate case as $\mathcal{M}_{X\to Y}$ and $\mathcal{M}_{Y\to X}$. To choose between the two models, and given data $\mathcal{D} = (X,Y)$, we need to compare their log posteriors,

$$\log P(\mathcal{M}_i|\mathcal{D}) = \log\frac{P(\mathcal{D}|\mathcal{M}_i)P(\mathcal{M}_i))}{P(\mathcal{D})}. \tag{9}$$



(a)

(b)

Figure 2: Datasets can have the ICM assumption hold in one direction (red and green), or both. A causal model effectively encodes the direction in which it expects ICM to hold. The log marginal likelihood is then higher for the correct causal model as it encodes the correct assumptions. (b) Changing the prior on the parameters changes the shape of the marginal likelihood distribution. There are cases where the data can be described by both causal directions (overlap between green and red).

Assuming a uniform prior over models, we can then simply choose the model with the highest log *marginal likelihood*

$$\mathcal{M}^{*} = \arg\max_{i}\{\log P(\mathcal{D}|\mathcal{M}_i)\}_i. \tag{10}$$

The marginal likelihood for a model is calculated by integrating over all the parameters and latents in a model. Figure 2 shows the intuition behind our model selection procedure. If a data is generated according to a causal model, it will have a higher log marginal likelihood under that causal model. This allows for identifiability of the correct causal model. Changing the prior will change these densities, but as long as ICM is encoded in one causal direction, the log marginal likelihood should be higher for the correct causal model. There are cases where the models are indistinguishable, for example when the prior for the second causal model is chosen using 7. A case where this happens is linear Gaussian models as shown in appendix A. For cases that are unidentifiable, the right prior should lead to a similar marginal likelihood value for the two models.

5 Method

It is necessary to choose flexible models to work with a wide range of data. We use latent variable Gaussian Process models to do this [10, 40, 24].

Causal Score To perform model selection, we calculate the log marginal likelihood by modelling the causal factorisations for both models. Thus, for $\mathcal{M}_{X\rightarrow Y}$ we model $\log P(\mathbf{x},\mathbf{y}|\mathcal{M}_{X\rightarrow Y}) = \log P(\mathbf{y}|\mathbf{x},\mathcal{M}_{X\rightarrow Y}) + \log P(\mathbf{x}|\mathcal{M}_{X\rightarrow Y})$, where we directly calculate the two terms by modelling the conditional and marginal distributions separately. The model with the higher log marginal likelihood is chosen as the most likely model.

Latent variable Gaussian Processes Gaussian processes (GPs) [33] are non-parametric Bayesian models that directly define a prior over functions. The form of the prior is controlled by a choice of a kernel function. Specifically, the kernel defines a covariance over outputs for the function $\mathbf{f}$, $\mathbf{f} \sim \mathcal{N}(\mathbf{0},\mathbf{K}_{\rho})$. The kernels are parametrised by continuous hyperparameters, ρ. Changing the prior simply amounts to choosing a different kernel or changing the values of the hyperparameters — allowing the ability to model different distributions. 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 of heteroscedastic noise as well as non Gaussian likelihoods. The likelihood for the conditional distribution for this model is $P(\mathbf{y}|\mathbf{x},\mathbf{f},\mathbf{w},\sigma) = \mathcal{N}\left(\mathbf{f}(\mathbf{x},\mathbf{w}),\sigma^2\right)$, with σ denoting the likelihood noise hyperparameter. The final log marginal likelihood for the conditional distribution is

$$P(\mathbf{y}|\mathbf{x}) = \int\int\int\int P(\mathbf{y}|\mathbf{x},\mathbf{f},\mathbf{w},\sigma)P(\mathbf{f}|\mathbf{x},\mathbf{w},\rho)P(\mathbf{w})P(\rho)P(\sigma)d\mathbf{f}d\mathbf{w}d\rho d\sigma. \tag{11}$$

The marginal likelihood for the marginal distribution $P(\mathbf{x})$ is analogous.

Methods	CE-Cha	CE-Multi	CE-Net	CE-Gauss
CGNN [12]	76.2	94.7	86.3	89.3
GPI [39]	71.5	73.8	88.1	90.2
PNL [42]	78.6	51.7	75.6	84.7
ANM [18]	43.7	25.5	87.8	90.7
IGCI [8]	55.6	77.8	57.4	16.0
LiNGAM [37]	57.8	62.3	3.3	72.2
RECI [5]	59.0	94.7	66.0	71.0
CDCI [9]	72.2	97.6	94.3	**91.8**
GPLVM	**82.1**	**97.7**	**98.8**	90.2

Table 1: Performance comparisons. Results for the baselines taken from [14]. Numbers convey the ROC AUC metric. Best results are in bold. Our method (GPLVM) outperforms competing methods.

Priors We use a standard normal prior for the latent term $\mathbf{w}$ and uniform priors over hyperparameters. The priors are the same for the marginal and conditional distributions, and for both causal models.

Integration The integration over the function and latent term is done by following the procedure in [40]. Here, an inducing point approximation is also used for scalability and variational inference [17] is used to tackle the intractability of the integrals. The integral over the hyperparameters ρ and σ is done by using the evidence approximation [26]. This simply involves maximising the log marginal likelihood with respect to the hyperparameters. This is motivated by the observation that the log marignal likelihood tends to be peaked for low dimensional hyperparameters and high data [33].

6 Experiments

We wish to test our method on a wide variety of data generating distributions. As we use flexible density approximators (latent variable GPs), we expect our method (labelled GPLVM) to work for a wide variety of functional and noise assumptions.

Datasets The following datasets are used to measure the performance of the proposed method. Each dataset contains 300 pairs with relationships $X \to Y$ and $Y \to X$ of 1500 samples each:

- **CE-Cha**: A mixture of synthetic and real world data. Taken from the cause-effect pairs challenge [14].
- **CE-Multi** [12]: 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** [12]: Synthetic data with randomly initialised neural networks for functions and random exponential family distributions chosen for the cause.
- **CE-Gauss** [30]: Synthetic data generated with random noise distributions E_1, E_2 defined in [30]. 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.

Metrics We use the *Area under ROC curve* (AUC) metric to analyse the performance of the methods. This takes the confidence of classifying a causal model into account as well.

Results Table 1 shows the results of our method (GPLVM), along with competing methods. Our method outperforms previous methods in a wide range of data generating assumptions. Methods that explicitly put assumptions on the data generation process (ANM, LiNGAM, RECI), only seem to do well on certain datasets. Methods based on ICM (CGNN, GPI, CDCI) do better on multiple datasets, however our method outperforms.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27), 2016.

[3] Alexis Bellot, Anish Dhir, and Giulia Prando. Generalization bounds and algorithms for estimating conditional average treatment effect of dosage. *arXiv preprint arXiv:2205.14692*, 2022.

[4] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020.

[5] 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.

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

[7] Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *arXiv preprint arXiv:2010.15764*, 2020.

[8] 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.

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

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

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

[12] 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.

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

[14] Isabelle Guyon, Alexander Statnikov, and Berna Bakir Batu. *Cause effect Pairs in machine learning*. Springer, 2019.

[15] David Heckerman and Dan Geiger. Likelihoods and parameter priors for bayesian networks. 1995.

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

[17] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

[18] 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.

[19] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

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

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

[22] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430), 1995.

[23] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020.

[24] 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.

[25] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

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

[27] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

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

[29] Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, 2020.

[30] 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.

[31] Judea Pearl. *Causality*. Cambridge university press, 2009.

[32] 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.

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

[34] 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.

[35] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[36] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 2017.

[37] 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.

[38] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12, 2011.

[39] 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.

[40] 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.

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

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

A Non identifiability in linear additive Gaussian noise model example

We first look at the model $\mathcal{M}_{X\to Y}$. As it is common in causal discovery settings to normalise the input and outputs [30], we assume the following generative model of our data,

$$P(x|\mathcal{M}_{X\to Y}) = \mathcal{N}(x|0,1), \tag{12}$$

$$P(y|x,b,\mathcal{M}_{X\to Y}) = \mathcal{N}(y|bx, 1-b^2). \tag{13}$$

The above chosen model ensures that the marginals $P(x)$ and $P(y)$ are standard normal distributions. Using Bayes rule, we can find the backward model in closed form in this case

$$P(y|b,\mathcal{M}_{X\to Y}) = \mathcal{N}\left(y|0,1\right), \tag{14}$$

$$P(x|y,b,\mathcal{M}_{X\to Y}) = \mathcal{N}\left(x|by, 1-b^2\right). \tag{15}$$

Clearly ICM holds in the backward direction here. The causal factorisation for the model $\mathcal{M}_{Y\to X}$ must follow the same parametrisation as the causal factorisation for $\mathcal{M}_{X\to Y}$

$$P(y|b,\mathcal{M}_{Y\to X}) = \mathcal{N}\left(y|0,1\right), \tag{16}$$

$$P(x|y,b,\mathcal{M}_{Y\to X}) = \mathcal{N}\left(x|by, 1-b^2\right). \tag{17}$$

With any prior on the parameter b, it is clear to see that

$$P(x,y|\mathcal{M}_{Y\to X}) = P(x,y|\mathcal{M}_{X\to Y}). \tag{18}$$

Normalisation forces an ICM condition here that makes the linear additive Gaussian noise model non identifiable. In a lot of cases, this is desirable as a shift in the mean or scale can be semantically meaningless. For example, a shift from Celsius to Fahrenheit should not affect the causal conclusions. It is interesting to note that in this case, the marginal likelihoods are equal and the posteriors for the two models will be roughly equal — effectively conveying uncertainty over the model choice.

B Proofs

B.1 Proof of theorem 4.1

Proof. For a causal model $\mathcal{M}_{X\to Y}$, the anticausal factorisation is

$$\int\int P(x|y,\theta,\phi,\mathcal{M}_{X\to Y})P(y|\theta,\phi,\mathcal{M}_{X\to Y})P(\theta|\mathcal{M}_{X\to Y})P(\phi|\mathcal{M}_{X\to Y})d\theta d\phi, \tag{19}$$

where θ, ϕ are the parameters of the causal factorisation. If condition 1 holds, this is equal to

$$\int\int P(x|y,\eta,\mathcal{M}_{X\to Y})P(y|\gamma,\mathcal{M}_{Y\to X})P(\eta,\gamma|\mathcal{M}_{X\to Y})d\eta d\gamma, \tag{20}$$

where $P(\eta,\gamma|\mathcal{M}_{X\to Y})$ is given by equation 7. If condition 2 must be satisfied, $P(\eta,\gamma|\mathcal{M}_{X\to Y}) = P(\eta|\mathcal{M}_{Y\to X})P(\gamma|\mathcal{M}_{X\to Y})$. Hence the anticausal factorisation is

$$\int P(x|y,\eta,\mathcal{M}_{X\to Y})P(\eta|\mathcal{M}_{X\to Y})d\eta \int P(y|\gamma,\mathcal{M}_{X\to Y})P(\gamma|\mathcal{M}_{X\to Y})d\gamma. \tag{21}$$

ICM holds in equation 21 as a change in the distribution of η changes $P(x|y,\mathcal{M}_{X\to Y})$ but does not affect $P(y|\mathcal{M}_{X\to Y})$. The same intuition holds for changing the distribution of γ.

B.2 Proof of theorem 4.2

Note that we assume the same parametrisation of causal factors for both causal models. We prove this for the model $\mathcal{M}_{X\to Y}$. The anticausal factorisation of $\mathcal{M}_{X\to Y}$ is

$$\int\int P(x|y,\theta,\phi,\mathcal{M}_{X\to Y})P(y|\theta,\phi,\mathcal{M}_{X\to Y})P(\theta|\mathcal{M}_{X\to Y})P(\phi|\mathcal{M}_{X\to Y})d\theta d\phi, \tag{22}$$

where ϕ and θ are the parameters for the causal factorisation. The causal factorisation for $\mathcal{M}_{Y\to X}$ is

$$\int P(x|y,\zeta,\mathcal{M}_{X\to Y})P(\zeta|\mathcal{M}_{Y\to X})d\zeta\int P(y|\rho,\mathcal{M}_{Y\to X})P(\rho|\mathcal{M}_{Y\to X})d\rho, \tag{23}$$

where ζ and ρ are the parameters for the causal factorisation for this causal model.

It is instructive to see the case where two causal models imply the same joint. It is trivial to see that if ICM holds in the anticausal direction for $\mathcal{M}_{X\to Y}$, that is theorem 4.1 holds, the anticausal factorisation can be written as

$$\int P(x|y,\eta,\mathcal{M}_{X\to Y})P(\eta|\mathcal{M}_{X\to Y})d\eta\int P(y|\gamma,\mathcal{M}_{X\to Y})P(\gamma|\mathcal{M}_{X\to Y})d\gamma, \tag{24}$$

with η and γ defined in theorem 4.1. With the right choice of priors for $\mathcal{M}_{Y\to X}$, namely $P(\rho|\mathcal{M}_{Y\to X}) = P(\gamma|\mathcal{M}_{X\to X})$ and $P(\zeta|\mathcal{M}_{X\to Y}) = P(\eta|\mathcal{M}_{Y\to X})$, we get

$$P(x,y|\mathcal{M}_{X\to Y}) = P(x,y|\mathcal{M}_{Y\to X}). \tag{25}$$

Note that this will usually require a different prior for the two causal models. The choice of prior for $\mathcal{M}_{Y\to X}$ for equality to hold will depend on equation 7. For the same prior to give equality in two models, the Jacobian of the implied prior in equation 7 needs to be identity. A case where this happens is discussed in appendix A.

Proof. Assume that $\mathcal{M}_{X\to Y}$ does not satisfy the ICM principle in the anticausal direction. According to theorem 4.1, this is due to the two conditions not being satisfied. If condition 1 is not satisfied, there exists some θ,ϕ such that there is no $\eta := f_1(\theta,\phi)$ and $\gamma := f_2(\theta,\phi)$ that gives $P(x|y,\theta,\phi,\mathcal{M}_{X\to Y}) = P(x|y,\eta,\mathcal{M}_{X\to Y})$, and $P(y|\theta,\phi,\mathcal{M}_{X\to Y}) = P(y|\gamma,\mathcal{M}_{X\to Y})$. This implies that the anticausal factorisation for $\mathcal{M}_{X\to Y}$ cannot be expressed in the chosen parametrisation. Hence, we have that for some θ,ϕ

$$\begin{aligned} P(x|y,\theta,\phi,\mathcal{M}_{X\to Y}) \neq& P(x|y,\zeta,\mathcal{M}_{Y\to X}), && (26)\\ P(y|\theta,\phi,\mathcal{M}_{X\to Y}) \neq& P(y|\rho,\mathcal{M}_{Y\to X}). && (27)\end{aligned}$$

If condition 1 does hold, but condition 2 does not, then the anticausal factorisation for $\mathcal{M}_{X\to Y}$ can be written as

$$\int\int P(x|y,\eta,\mathcal{M}_{X\to Y})P(y|\gamma,\mathcal{M}_{X\to Y})P(\eta,\gamma|\mathcal{M}_{X\to Y})d\eta d\gamma, \tag{28}$$

where

$$P(\eta,\gamma|\mathcal{M}_{X\to Y}) = P(\theta|\mathcal{M}_{X\to Y})P(\phi|\mathcal{M}_{X\to Y})\left|\begin{bmatrix}\frac{\partial f_1(\theta,\phi))}{\partial\theta} & \frac{\partial f_1(\theta,\phi)}{\partial\phi}\\ \frac{\partial f_2(\theta,\phi)}{\partial\theta} & \frac{\partial f_2(\theta,\phi)}{\partial\phi}\end{bmatrix}\right|^{-1}. \tag{29}$$

Due to the parametrisations being the same of the two causal models, clearly for every (η,γ), there is some (ζ,ρ) such that $P(x|y,\zeta,\mathcal{M}_{Y\to X}) = P(x|y,\eta,\mathcal{M}_{X\to Y})$ and $P(y|\rho,\mathcal{M}_{Y\to X}) = P(y|\gamma,\mathcal{M}_{X\to Y})$. As the priors for the parameters are dependent (as a consequence of condition 2 not holding), it cannot be expressed as the product of two distributions and hence equation 28 cannot equal equation 23 for any choice of prior in equation 23. To conclude, there must be some x,y such that

$$P(x,y|\mathcal{M}_{X\to Y}) \neq P(x,y|\mathcal{M}_{Y\to X}). \tag{30}$$