2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–21

Bayesian Model Selection for Identifying Markov Equivalent Causal Graphs

M. Burak Kurutmaz[*] BURAK.KURUTMAZ@BOUN.EDU.TR
Melih Barsbey[*] MELIH.BARSBEY@BOUN.EDU.TR
A. Taylan Cemgil TAYLAN.CEMGIL@BOUN.EDU.TR
Boğaziçi University, İstanbul, Turkey
Sinan Yıldırım SINANYILDIRIM@SABANCIUNIV.EDU
Sabancı University, İstanbul, Turkey
Umut Şimşekli UMUT.SIMSEKLI@TELECOM-PARISTECH.FR
LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France

Abstract

Many approaches to causal discovery are limited by their inability to discriminate between Markov equivalent graphs given only observational data. We formulate causal discovery as a marginal likelihood based Bayesian model selection problem. We adopt a parameterization based on the notion of the independence of causal mechanisms which renders Markov equivalent graphs distinguishable. We complement this with an empirical Bayesian approach to setting priors so that the actual underlying causal graph is assigned a higher marginal likelihood than its alternatives. Adopting a Bayesian approach also allows for straightforward modeling of unobserved confounding variables, for which we provide a variational algorithm to approximate the marginal likelihood, since this desirable feat renders the computation of the marginal likelihood intractable. We believe that the Bayesian approach to causal discovery both allows the rich methodology of Bayesian inference to be used in various difficult aspects of this problem and provides a unifying framework to causal discovery research. We demonstrate promising results in experiments conducted on real data, supporting our modeling approach and our inference methodology.

1. Introduction

Causal networks (CNs) are special Bayesian networks where all edges reflect causal relations (Pearl, 2009). The aim of causal structure learning is identifying the CN underlying the observed data. In this paper, we focus on the problem of scoring causal graphs using marginal likelihood in a way that identifies the unique causal generative graph. Succeeding to do so is very valuable, since once the correct CN is selected, various causal inference tasks such as estimating causal effects or examining confounder distributions becomes straightforward in a Bayesian framework. A central challenge in such an attempt, however, is adopting a prior selection policy that not only allows discriminating between Markov equivalent graphs but also assigns higher marginal likelihood score to the actual underlying CN.

The key notion underlying our solution to first part of this challenge is the widely accepted principle of *independence of the cause-effect mechanisms* (Janzing et al., 2012),

[*] Equal contribution

that is, the natural mechanisms that generate the cause and the effect (based on cause) must be independent of each other. We embody this assumption by assuming the *mutual independence* of the parameters pertaining to cause and effect distributions in a Bayesian model, a line of reasoning that is natural to this modeling perspective, where parameters are modeled as *random variables* (Spiegelhalter et al., 1993; Heckerman et al., 1995; Geiger et al., 1997; Blei et al., 2003). By assigning independent priors to the cause and effect variables, we render them statistically independent. Critically, this assignment of independent priors also breaks the likelihood equivalence between Markov equivalent graphs. This is contrast to other ways of selecting independent priors such as the *BDeu* prior, which leads to assigning equal marginal likelihood to Markov equivalent graphs (Heckerman et al., 1995).

As mentioned above, though breaking likelihood equivalence does not necessarily lead to assigning a higher marginal likelihood to the *actual* underlying CN, it is a prerequisite for doing so[1]. The second part of the problem is adapting a prior selection policy that leads to assigning a higher marginal likelihood to the actual CN compared to its alternatives. In this work, we use an empirical Bayesian approach in selecting the hyperparameters of the independent priors described above, as we learn the priors that lead to assigning higher marginal likelihood to the actual CN from labeled data.

The current approach is in the intersection of various other approaches in the literature, thereby combining many of their respective advantages (Spirtes and Zhang, 2016; Glymour et al., 2019). It is based on the notion of mechanism independence similar to Janzing et al. (2012); Zhang et al. (2015), does not assume causal sufficiency similar to Silva et al. (2006); Shimizu et al. (2009); Janzing et al. (2009, 2012); Zhang et al. (2015); Schölkopf et al. (2016), can theoretically work on arbitrary graph structures that possibly include latent variables similar to Spirtes et al. (1993), and can discriminate between Markov equivalent structures similar to Shimizu et al. (2006); Zhang and Hyvärinen (2008); Hoyer et al. (2009); Janzing et al. (2012); Zhang et al. (2015). Our approach diverges from other Bayesian methods (Stegle et al., 2010; Shimizu and Bollen, 2014; Zhang et al., 2016) in various dimensions such as by being able to distinguish between Markov equivalent causal graphs, using marginal likelihood (or approximations thereof) instead of surrogate scores such as BIC, or being able to model non-linear relationships.

In Section 2, we introduce an example model for continuous observations and latent categorical confounders. To approximate the marginal likelihood in graphs which include latent confounders, we present a variational inference algorithm in Section 3. After testing our approach on various real data sets in Section 4, we present our conclusions and further avenues of research in Section 5.

2. A Mixture of Linear Basis Functions Model

A general causal graph $\mathcal{G}(V_\mathcal{G}, E_\mathcal{G})$ is a combination of a vertex set $V_\mathcal{G}$, which is the set of observed and latent random variables, and a set of directed edges $E_\mathcal{G} \subseteq V_\mathcal{G} \times V_\mathcal{G}$ where directed edges imply immediate cause-effect relationships between these variables. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \ldots, \boldsymbol{x}_N\} \subseteq V_\mathcal{G}$ denote the set of continuous random variables, and similarly

[1] We conduct a more in-depth exploration of identifiability conditions in Bayesian networks, and discuss the relationship among the concepts *Markov equivalence*, *distribution equivalence*, and *likelihood equivalence* in Appendix A.

$\{\boldsymbol{r}_1 \ldots, \boldsymbol{r}_k, \ldots, \boldsymbol{r}_K\} \subseteq V_\mathcal{G}$ denote the discrete latent variables of the network where each $\boldsymbol{x}_n$ and each $\boldsymbol{r}_k$ are defined in the domains $\mathcal{X}_n$ and $\mathcal{R}_k$, respectively. The set of parent vertices of a vertex $\boldsymbol{v} \in V_\mathcal{G}$ is denoted by $\pi(\boldsymbol{v})$, while we denote its continuous parents by $\boldsymbol{x}_{\pi(\boldsymbol{v})}$, and discrete parents by $\boldsymbol{r}_{\pi(\boldsymbol{v})}$.

For the scope of this text, we specify conditional distributions for the graphs as follows: we assume *categorical* distributions on the discrete variables $\boldsymbol{r}_{1:K}$ and linear basis functions models with *Gaussian* noise on the continuous variables $\boldsymbol{x}_{1:N}$. Though these choices are by no means mandatory for our framework, we define latent variables as categorical. Furthermore, we restrict our attention to the graphical structures that do not include a continuous variable as a parent of a categorical variable for inferential convenience (Heckerman et al., 1995), and construct the following generative model for T independent and identically distributed observations from the network $\mathcal{G}$:

$$\forall k, t: \qquad r_k^t \mid r_{\pi(\boldsymbol{r}_k)}^t \sim \text{Categorical}(\theta_{k|r^t_{\pi(\boldsymbol{r}_k)}}) \tag{1}$$

$$\forall n, t: \qquad x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, r_{\pi(\boldsymbol{x}_n)}^t \sim \mathcal{N}(w_{n|r^t_{\pi(\boldsymbol{x}_n)}}{}^{\mathrm{T}} \phi(x^t_{\pi(\boldsymbol{x}_n)}), \rho_{n|r^t_{\pi(\boldsymbol{x}_n)}}{}^{-1}) \tag{2}$$

where $1 \leq t \leq T$, ϕ is an arbitrary basis function with the convention $\phi(\{\}) = 1$, and $\theta_{k|r^t_{\pi(\boldsymbol{r}_k)}}$, $w_{n|r^t_{\pi(\boldsymbol{x}_n)}}$, $\rho_{n|r^t_{\pi(\boldsymbol{x}_n)}}$'s are the parameters of the conditional distributions. Namely, θ_k is the conditional distribution table of $\boldsymbol{r}_k$, w_n is the weights of the basis functions, and ρ_n is the precision parameter of the conditional distribution of $\boldsymbol{x}_n$.

Notice that declaring parameters as random variables simplifies the notion of independent cause-effect mechanisms as follows: Since the conditional distributions are the functions of the parameters, independence of the conditional distributions boils down to the independence of the parameters. Therefore, we complete our generative model by defining independent conjugate prior distributions on the parameters

$$\forall k, r_{\pi(\boldsymbol{r}_k)}: \qquad \theta_{k|r_{\pi(\boldsymbol{r}_k)}} \sim \text{Dirichlet}(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}) \tag{3}$$

$$\forall n, r_{\pi(\boldsymbol{x}_n)}: \qquad w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \sim \mathcal{NG}(m_{n|r_{\pi(\boldsymbol{x}_n)}}, \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}, a_{n|r_{\pi(\boldsymbol{x}_n)}}, b_{n|r_{\pi(\boldsymbol{x}_n)}}) \tag{4}$$

where $\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}$, $m_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}$, $a_{n|r_{\pi(\boldsymbol{x}_n)}}$, $b_{n|r_{\pi(\boldsymbol{x}_n)}}$ are the prior parameters, i.e. hyperparameters, of our generative model.

3. Mean-Field Variational Bayes

Variational Bayesian inference (VB) (Beal et al., 2006) is a technique where an intractable posterior distribution $\mathcal{P}$ is approximated by a variational distribution $\mathcal{Q}$ via minimizing *Kullback-Leibler* divergence $\text{KL}(\mathcal{Q}||\mathcal{P})$. In the context of Bayesian model selection, minimization of the $\text{KL}(\mathcal{Q}||\mathcal{P})$ corresponds to establishing a tight lower bound for the marginal log-likelihood, which we refer to as *evidence lower bound* (ELBO). This correspondence is due to the following decomposition of marginal log-likelihood

$$\log p(x_{1:N}^{1:T}) = \text{KL}(\mathcal{Q}\|\mathcal{P}) + \mathcal{B}_\mathcal{P}[\mathcal{Q}] \geq \mathcal{B}_\mathcal{P}[\mathcal{Q}] \tag{5}$$

where $\mathcal{P} = p(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N} \mid x_{1:N}^{1:T})$ is the full posterior distribution, and ELBO is denoted by $\mathcal{B}_\mathcal{P}[\mathcal{Q}]$. In a typical scenario of VB, $\mathcal{Q}$ is assumed to be a member of a restricted family of distributions. In its most common form, also known as *mean-field* approximation, $\mathcal{Q}$ is assumed to factorize over some partition of the latent variables, in a way that is reminiscent to a rank-one approximation in the space of distributions

$$\mathcal{Q}(\boldsymbol{r}_{1:K}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) = q(\boldsymbol{r}_{1:K}^{1:T})\; q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})$$

ELBO is then maximized with respect to $\mathcal{Q}$ which is restricted to the class of factorized distributions. Due to conjugacy, maximization of $\mathcal{Q}$ results in further factorized variational distributions which also belong to the same family as the prior

$$\begin{aligned} q(\boldsymbol{r}_{1:K}^{t}) &= \text{Categorical}(\boldsymbol{r}_{1:K}^{t}; \hat{\theta}^{t}) \\ q(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}}) &= \text{Dirichlet}(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}}; \hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}) \\ q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}) &= \mathcal{NG}(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}) \end{aligned}$$

Here $\hat{\theta}^t$, $\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}$, $\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}$, $\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}$ represent the *variational parameters*. To calculate variational parameter updates, we need to calculate the *expected sufficient statistics*. In its final form, our variational algorithm becomes equivalent to iteratively calculating the expected sufficient statistics and updating the parameters. The explicit forms for the variational parameters and ELBO can be found in Appendix C.

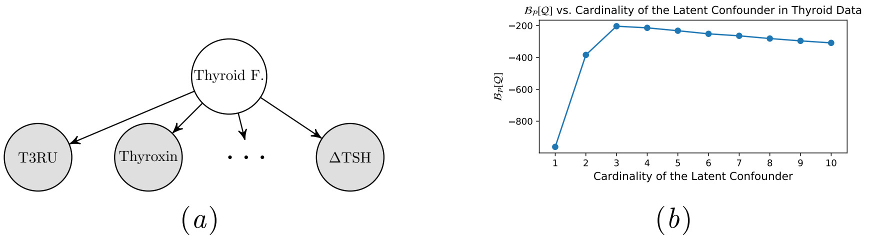
4. Experiments

In Section 4.1 we test the performance of our approach in bivariate causal discovery. Then in Section 4.2 we identify the cardinality and distribution of a latent confounder in a multivariate data set, exemplifying the versatility of a Bayesian approach to causality.

4.1. Finding Causal Direction: The CauseEffectPairs Data Set

In the first part we measured the accuracy of VB for the causal direction determination problem. The data set in this part is CEP (Mooij et al., 2016), frequently used in causal discovery research, which includes 100 data sets, vast majority of which is bivariate. For the hyperparameters of the model, we created 36 different settings by varying the critical hyperparameters systematically. We detail this hyperparameter creation process in the Appendix D.1. In making a decision between two causal directions in a given hyperparameter setting, we choose the model which obtains a higher ELBO[2]. We tested our algorithm on the data set by using 10 × 3 cross-validation. That is, for each test, we separated the data set into three, detected the hyperparameter setting (of 36) that obtained the best accuracy score on the first two thirds, and tested our model on the last third of the data set, which corresponds to an empirical Bayesian approach to prior selection. We conducted the same process two more times, each fold becoming the test set once. We conducted this split and tested randomly 10 times. We report the accuracy and AUC values according to these 10 runs. the CEP data set, we obtained a mean accuracy of .78±.09 and AUC score of .84±.13 (the values following the mean values correspond to 68% CI) where the accuracy and AUC calculations are performed by using the weights mentioned by Mooij et al. (2016). Mooij et al. (2016) also compared most recent methods on their performance on the data set; our results correspond to a state-of-the-art performance in bivariate causality detection.

[2] The two models compared in this experiment are depicted in Figures 2(b) and 2(c) in Appendix A.1.



(a) (b)

Figure 1: (a) Causal graph of thyroid data and (b) ELBO with respect to $|\mathcal{R}_1|$.

4.2. Inferring the Latent Confounder: The Thyroid Data Set

Using a different data set, we next examine the ability of our approach to identify a latent confounder. For this purpose, we use the smallest database in the Thyroid data set from the UCI repository (Dheeru and Karra Taniskidou, 2017). This data involves five different diagnostic measurements from patients with *low*, *normal*, and *high* thyroid activity. This being a diagnostic data set, the causal structure is known, where the thyroid activity is the cause of the rest of the variables (Figure 1(a)). In our experiments we ignore the thyroid activity variable, thus it becomes a latent confounder. This way we can test how well our approach identifies the latent confounder.

To assess our method's performance, we first examine whether the latent variable cardinality our method favors corresponds to the cardinality of the actual variable that we held out. Figure 1(b) shows that the ELBO of the model is maximized at the latent cardinality which corresponds to the *actual* cardinality of thyroid activity variable (which is 3). Then, to ascertain that the inferred latent variable indeed corresponds to thyroid activity variable, we compare the assignments of our model to actual patient thyroid activity levels. The results demonstrate an accuracy of .93, thus we conclude that our method accurately identified the latent causal variable.

5. Conclusion

Overall, we show that Bayesian model selection is a promising framework that can facilitate causal research significantly both through conceptual unification and increased performance. Given that Bayesian modeling is agnostic to specific variable types, conditional distributions, and to approximate inference methodology, the value of a successful Bayesian modeling approach for causal research is immense.

Though our empirical Bayesian approach to setting priors can be useful in various contexts (e.g. in data sets where only *some* of the bivariate causal directions are known), finding other principled ways of assigning (or integrating out) priors that do not require labeled data is an important direction for future research. Conducting causal discovery with different variable types, and/or different distributions would also be beneficial for demonstrating current approach's viability in various contexts.

Acknowledgments

The work was partly supported as a bilateral Turkish-French research programme by the French National Research Agency under Grant ANR-16-CE23-0014 (FBIMATRIX) and by TUBITAK Turkish Science, Technology, and Research Foundation under Grant 116E580 (FBIMATRIX).

References

Matthew J Beal, Zoubin Ghahramani, et al. Variational bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1(4):793–831, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

A Philip Dawid, Steffen L Lauritzen, et al. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dan Geiger, David Heckerman, et al. A characterization of the dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, 25(3):1344–1369, 1997.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, June 2019. ISSN 1664-8021. doi: 10/gf7jkg. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00524/full. 5/5_2019_09_09.

Daniel M Hausman and James Woodward. Independence, invariance and the causal markov condition. *The British journal for the philosophy of science*, 50(4):521–583, 1999.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 249–257. AUAI Press, 2009.

Dominik Janzing, Eleni Sgouritsa, Oliver Stegle, Jonas Peters, and Bernhard Schölkopf. Detecting low-complexity unobserved causes. *arXiv preprint arXiv:1202.3737*, 2012.

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):1103–1204, 2016.

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

Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016.

Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *The Journal of Machine Learning Research*, 15(1):2629–2652, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (Oct):2003–2030, 2006.

Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

Ricardo Silva, Richard Scheine, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

David J Spiegelhalter, A Philip Dawid, Steffen L Lauritzen, and Robert G Cowell. Bayesian analysis in expert systems. *Statistical science*, pages 219–247, 1993.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, page 3. SpringerOpen, 2016.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 1993.

Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in neural information processing systems*, pages 1687–1695, 2010.

Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*, pages 157–164. JMLR. org, 2008.

Kun Zhang, Jiji Zhang, and Bernhard Schölkopf. Distinguishing cause from effect based on exogeneity. *arXiv preprint arXiv:1504.05651*, 2015.

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):13, 2016.

Appendix A. Identifiability of Markov Equivalent Graphs

When constructing a generative model for causal inference, our aim is making Markov equivalent graph structures identifiable. However, the model that is described only by Equations (1) and (2) is not necessarily identifiable (Shimizu et al., 2006; Hoyer et al., 2009). To be more precise, consider the case where we have two continuous variables and no latent categorical variable, which is equivalent to the following structural equation model:

$$\begin{aligned} x_1 &= w_1(1) + \rho_1^{-1/2}\epsilon_1 & \epsilon_1 &\sim \mathcal{N}(0,1) \\ x_2 &= w_2(1)x_1 + w_2(2) + \rho_2^{-1/2}\epsilon_2 & \epsilon_2 &\sim \mathcal{N}(0,1) \end{aligned}$$

One can also construct the following equivalent structural equation model in which the dependence structure is reversed:

$$\begin{aligned} x_2 &= w_1(1)w_2(1) + w_2(2) + \hat{\rho}_2^{-1/2}\hat{\epsilon}_2 = \hat{w}_2(1) + \hat{\rho}_2^{-1/2}\hat{\epsilon}_2 & \hat{\epsilon}_2 &\sim \mathcal{N}(0,1) \\ x_1 &= \frac{1}{w_2(1)}x_2 - \frac{w_2(2)}{w_2(1)} + \hat{\rho}_1^{-1/2}\hat{\epsilon}_1 = \hat{w}_1(1)x_2 - \hat{w}_1(2) + \hat{\rho}_1^{-1/2}\hat{\epsilon}_1 & \hat{\epsilon}_1 &\sim \mathcal{N}(0,1) \end{aligned}$$

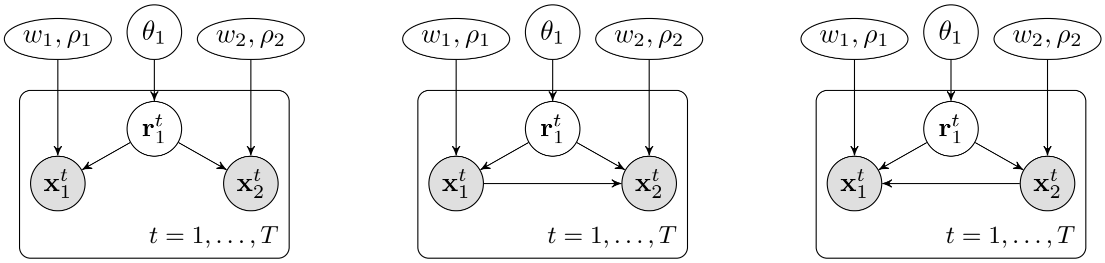
These two models are not identifiable with the descriptions above, since they both correspond to linear models with Gaussian noise. However, by assuming priors on the parameters we can break the symmetry and make these Markov equivalent models identifiable. For instance, assuming Gaussian priors on the weights of the first model implies non-Gaussian priors on the second model, which makes these two models *distribution inequivalent* (Spirtes and Zhang, 2016). Moreover, even when two Markov equivalent models are also distribution equivalent, choosing appropriate prior parameters that violate *likelihood equivalence* still makes them identifiable (Heckerman et al., 1995). Indeed, for a model with a parameterization as described, only a very specific choice of priors leads to likelihood equivalence between the Markov equivalent models (Geiger et al., 1997; Dawid et al., 1993), and we will avoid following such a constraint. Choosing arbitrary priors almost always leads to likelihood inequivalent, hence identifiable models.

A.1. Identifiable Graphical Models for Bivariate Causality

In this section, we define the appropriate graphical structures for causal structure learning in the bivariate case. As we stated in Section 1, we do not assume causal sufficiency and allow the existence of possibly many exogenous variables. Luckily, we can combine the effects of exogenous variables into a single latent variable with an arbitrary cardinality. As a result, the relationship between two observable dependent variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ boils down to one of three cases due to causal Markov condition (Hausman and Woodward, 1999):

1. $\boldsymbol{x}_1$ causes $\boldsymbol{x}_2$,
2. $\boldsymbol{x}_2$ causes $\boldsymbol{x}_1$,
3. they do not cause each other, but a latent variable $\boldsymbol{r}_1$ causes both of them.

Associated causal networks corresponding to each of these hypotheses are depicted in Figure 2, where latent variable $\boldsymbol{r}_1$ represents the overall effect of the all unobserved variables. For the spurious relationship (Figure 2(a)), marginally correlated variables $\boldsymbol{x}_1$ and



(a) Spurious correlation. (b) $\mathbf{x}_1$ causes $\mathbf{x}_2$. (c) $\mathbf{x}_2$ causes $\mathbf{x}_1$.

Figure 2: Graphical models for bivariate causality.

$\boldsymbol{x}_2$ become independent once the latent common cause variable $\boldsymbol{r}_1$ is known. However in direct causal relationships (Figures 2(b) and 2(c)), even when the latent common cause is known, two variables are still dependent and the direction of cause-effect relationship is implicit in the parameterization of the models.

The identifiability of these models resides in the fact that modelling parameters explicitly as random variables makes these graphs Markov inequivalent. If we were considering only the marginal models of the observed variables, then we would end up with three Markov equivalent graphs. However, including latent variables and independent parameters renders distinctive conditional independence properties for each graph. For instance, when $\boldsymbol{x}_2$ and $\boldsymbol{r}_1$ are known, $\boldsymbol{x}_1$ and the parameters of $\boldsymbol{x}_2$ are dependent only in the case of $\boldsymbol{x}_1 \to \boldsymbol{x}_2$, or knowing $\boldsymbol{r}_1$ makes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ independent only if they have a spurious relationship. These distinctive conditional independence properties are the underlying reasons making all of these graphs identifiable.

Appendix B. Exponential Family

B.1. Basic Distributions

In this section, we supply the brief descriptions of the basic distributions that we mentioned in the main part of the manuscript.

B.1.1. Gamma Distribution

1. Gamma function:

$$\Gamma(z) \equiv \int_0^\infty x^{z-1} e^{-z} \, \mathrm{d}x$$

which is equal to $(z-1)!$ for nonnegative integer z.

2. Gamma density:

$$\mathrm{Gamma}(\rho; a, b) = \exp((a-1)\log\rho - b\rho - \log\Gamma(a) + a\log b)$$

where a is the *shape* and b is the *rate* parameter.

3. Expected sufficient statistics:

$$\mathsf{E}\left\{\rho\right\} = a/b, \quad \mathsf{E}\left\{\log\rho\right\} = \psi(a) - \log(b)$$

4. Cross entropy:

$$\begin{aligned}\mathsf{E}_{\mathrm{Gamma}(\rho;\hat{a},\hat{b})}&\left\{-\log \mathrm{Gamma}(\rho;a,b)\right\}\\&= -(a-1)\mathsf{E}\left\{\log\rho\right\} + b\mathsf{E}\left\{\rho\right\} + \log\Gamma(a) - a\log b\\&= -(a-1)(\psi(\hat{a}) - \log(\hat{b})) + \frac{\hat{a}b}{\hat{b}} + \log\Gamma(a) - a\log b\end{aligned}$$

Here, $\psi(x)$ is the *digamma* function which is defined as $\psi(x) = \frac{\mathrm{d}\log\Gamma(x)}{\mathrm{d}x}$.

B.1.2. Dirichlet Distribution

1. Multivariate Beta function:

$$B(\gamma) = \frac{\prod_r \Gamma(\gamma_r)}{\Gamma(\sum_r \gamma_r)}$$

2. Dirichlet density:

$$\mathrm{Dirichlet}(\theta;\gamma) = \frac{1}{B(\gamma)}\exp\big(\sum_r(\gamma_r - 1)\log\theta_r\big)$$

3. Expected sufficient statistics:

$$\mathsf{E}\left\{\theta_r\right\} = \frac{\gamma_r}{\sum_m \gamma_m}, \quad \mathsf{E}\left\{\log\theta_r\right\} = \psi(\gamma_r) - \psi(\sum_m \gamma_m)$$

4. Cross entropy:

$$\begin{aligned}\mathsf{E}_{\mathrm{Dirichlet}(\theta;\hat{\gamma})}\left\{-\log\mathrm{Dirichlet}(\theta;\gamma)\right\} &= \log B(\gamma) - \sum_r(\gamma_r - 1)\mathsf{E}\left\{\log\theta_r\right\}\\&= \log B(\gamma) - \sum_r(\gamma_r - 1)\big(\psi(\gamma_r) - \psi(\sum_m\gamma_m)\big)\end{aligned}$$

B.1.3. Categorical Distribution

1. Categorical density:

$$\mathrm{Categorical}(r;\theta) = \prod_{k=1}^{K}\theta_k^{\mathbb{1}_{\{r=k\}}}$$

2. Expected sufficient statistics:

$$\mathsf{E}\left\{\mathbb{1}_{\{r=k\}}\right\} = \theta_k$$

3. Cross entropy:

$$\mathsf{E}_{\mathrm{Categorical}(r;\hat{\theta})}\left\{-\log\mathrm{Categorical}(r;\theta)\right\} = -\sum_k \hat{\theta}_k\log\theta_k$$

B.1.4. Normal Distribution

1. Normal density:

$$x \sim \mathcal{N}(\mu, \rho^{-1}) \quad = \quad \frac{1}{\sqrt{2\pi}} \exp\Big(\frac{1}{2}\log\rho - \frac{1}{2}\rho(x-\mu)^2\Big)$$

where μ is the *mean* parameter and ρ is the *precision* parameter, i.e. ρ^{-1} is the variance.

2. Expected sufficient statistics

$$\mathsf{E}\left\{x\right\} = \mu \qquad\qquad \mathsf{E}\left\{x^2\right\} = \mu^2 + \rho^{-1}$$

B.1.5. Multivariate Normal Distribution

1. Multivariate Normal density:

$$x \sim \mathcal{N}(\mu, \Lambda^{-1}) \quad = \quad \frac{1}{(2\pi)^{K/2}} \exp\Big(\frac{1}{2}\log\det(\Lambda) - \frac{1}{2}(x-\mu)^T\Lambda(x-\mu)\Big)$$

where μ is the *mean vector* and Λ is the *precision matrix*, i.e. Λ^{-1} is the covariance matrix.

2. Expected sufficient statistics:

$$\mathsf{E}\left\{x\right\} = \mu \qquad\qquad \mathsf{E}\left\{x^{\mathrm{T}}Ax\right\} = \mu^{\mathrm{T}}A\mu + \mathrm{tr}(\Lambda^{-1}A)$$

for any symmetric matrix A.

B.1.6. Normal-Gamma Distribution

1. Normal-Gamma density:

$$\mu, \rho \sim \mathcal{NG}(m, \lambda, a, b) \quad = \quad \frac{b^a\sqrt{\lambda}}{\Gamma(a)\sqrt{2\pi}} \exp\Big(\big(a - \frac{1}{2}\big)\log\rho - b\rho - \frac{\lambda}{2}\rho(\mu - m)^2\Big)$$

which can be equivalently decomposed into a marginal Gamma distribution and a conditional Normal distribution:

$$\rho \sim \mathrm{Gamma}(a, b) \qquad\qquad x \mid \rho \sim \mathcal{N}(m, (\lambda\rho)^{-1})$$

2. Expected sufficient statistics:

$$\mathsf{E}\left\{\log\rho\right\} = \psi(a) - \log b \qquad \mathsf{E}\left\{\rho\right\} = \frac{a}{b} \qquad \mathsf{E}\left\{\rho\mu\right\} = m\frac{a}{b} \qquad \mathsf{E}\left\{\rho\mu^2\right\} = \frac{1}{\lambda} + m^2\frac{a}{b}$$

3. Cross entropy:

$$\begin{aligned} &\mathsf{E}_{\mathcal{NG}(\hat{m},\hat{\lambda},\hat{a},\hat{b})}\left\{-\log\mathcal{NG}(\mu,\rho;m,\lambda,a,b)\right\} = \\ &\qquad -a\log b + \log\Gamma(a) - \frac{1}{2}\log\lambda + \frac{\lambda}{2\hat{\lambda}} + \frac{1}{2}\log 2\pi \\ &\qquad -\Big(a - \frac{1}{2}\Big)(\psi(\hat{a}) - \log\hat{b}) + \frac{\hat{a}b}{\hat{b}} + \frac{\hat{a}}{2\hat{b}}\lambda(\hat{m} - m)^2 \end{aligned}$$

B.1.7. Multivariate Normal-Gamma Distribution

1. Multivariate Normal-Gamma density:

$$\begin{aligned} w, \rho &\sim \mathcal{NG}(m, \Lambda, a, b) \\ &= \frac{b^a \sqrt{\det(\Lambda)}}{(2\pi)^{M/2}\Gamma(a)} \exp\left(\left(a + \frac{M}{2} - 1\right) \log \rho - b\rho - \frac{1}{2}\rho(w-m)^{\mathrm{T}}\Lambda(w-m)\right) \end{aligned}$$

   which can be equivalently decomposed into a marginal Gamma distribution and a conditional Multivariate Normal distribution:

$$\rho \sim \text{Gamma}(a, b) \qquad\qquad x \mid \rho \sim \mathcal{N}(m, (\rho\Lambda)^{-1})$$

2. Expected sufficient statistics:

$$\begin{aligned} &\mathsf{E}\{\log \rho\} = \psi(a) - \log b \qquad \mathsf{E}\{\rho\} = \frac{a}{b} \qquad \mathsf{E}\{\rho w\} = \frac{a}{b}m \\ &\mathsf{E}\left\{\rho\ w^{\mathrm{T}}Aw\right\} = \mathrm{tr}(\Lambda^{-1}A) + \frac{a}{b}m^{\mathrm{T}}Am \end{aligned}$$

   for any symmetric matrix A.

3. Cross entropy:

$$\begin{aligned} &\mathsf{E}_{\mathcal{NG}(\hat{m},\hat{\Lambda},\hat{a},\hat{b})}\{-\log \mathcal{NG}(w, \rho; m, \Lambda, a, b)\} = \\ &\quad -a \log b + \log \Gamma(a) - \frac{1}{2}\log\det(\Lambda) + \frac{1}{2}\mathrm{tr}(\hat{\Lambda}^{-1}\Lambda) + \frac{M}{2}\log 2\pi \\ &\quad -\left(a + \frac{M}{2} - 1\right)(\psi(\hat{a}) - \log \hat{b}) + \frac{\hat{a}b}{\hat{b}} + \frac{\hat{a}}{2\hat{b}}(\hat{m} - m)^{\mathrm{T}}\Lambda(\hat{m} - m) \end{aligned}$$

B.2. Basic Conjugate Models

In this section we summarize the basic conjugate models that are closely related to our example model.

B.2.1. Dirichlet-Categorical Model

1. Generative model:

$$\begin{aligned} \theta &\sim \text{Dirichlet}(\gamma) \\ r^1, \ldots, r^T &\sim \text{Categorical}(\theta) \end{aligned}$$

2. Posterior of θ:

$$\theta \mid r^1, \ldots, r^T \sim \text{Dirichlet}(\gamma^*)$$

   where $\gamma^*_r = \gamma_r + \sum_{t=1}^{T} \mathbb{1}_{\{r=r^t\}}$

B.2.2. Normal-Gamma-Normal Model

1. Generative model:

$$\begin{aligned} \mu, \rho &\sim \mathcal{NG}(m, \lambda, a, b) \\ x^1, \ldots, x^T &\sim \mathcal{N}(\mu, \rho^{-1}) \end{aligned}$$

2. Posterior of μ and ρ:

$$\mu, \rho \mid x^1, \ldots, x^T \;\sim\; \mathcal{NG}(m^*, \lambda^*, a^*, b^*)$$

where

$$\begin{aligned} \lambda^* &\equiv \lambda + T & m^* &\equiv \frac{\lambda m + \sum_t x^t}{\lambda^*} \\ a^* &\equiv a + \frac{T}{2} & b^* &\equiv b + \frac{1}{2}\big(\lambda m^2 - \lambda^* m^{*2} + \sum_t (x^t)^2\big) \end{aligned}$$

B.2.3. Bayesian Linear Regression

1. Generative model:

$$y^t = w^{\mathrm{T}} x^t + \rho^{-1/2} \epsilon^t \qquad\qquad \epsilon^t \sim \mathcal{N}(0, 1)$$

An equivalent description with Normal-Gamma priors is

$$\begin{aligned} w, \rho &\sim \mathcal{NG}(m, \Lambda, a, b) \\ y^t \mid x^t &\sim \mathcal{N}(w^{\mathrm{T}} x^t, \rho^{-1}) \end{aligned}$$

2. Posterior of w and ρ:

$$w, \rho \mid (x^1, y^1), \ldots, (x^T, y^T) \;\sim\; \mathcal{NG}(m^*, \Lambda^*, a^*, b^*)$$

where

$$\begin{aligned} \Lambda^* &\equiv \Lambda + \sum\nolimits_t x^t x^{tT} & m^* &\equiv \Lambda^{*-1}\big(\Lambda m + \sum\nolimits_t y^t x^t\big) \\ a^* &\equiv a + \frac{T}{2} & b^* &\equiv b + \frac{1}{2}\Big(m^{\mathrm{T}} \Lambda m - m^{*\mathrm{T}} \Lambda^* m^* + \sum_t (x^t)^2\Big) \end{aligned}$$

Appendix C. Variational Bayes

Minimization of $\mathrm{KL}(\mathcal{Q} \| \mathcal{P})$ ends up with the following marginal variational distributions:

$$q(\boldsymbol{r}_{1:K}^{1:T}) \propto \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})} \left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \tag{6}$$

$$q(\boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) \propto \exp\big(\mathsf{E}_{q(\boldsymbol{r}_{1:K}^{1:T})} \left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \tag{7}$$

In this section, we will explicitly evaluate these equations to derive closed form expressions for the variational posteriors:

1. We first simplify the (6) via factorization property of the joint distribution and removing the multiplicative constants

$$\begin{aligned} q(\boldsymbol{r}_{1:K}^{1:T}) &\propto \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K},\boldsymbol{\rho}_{1:N},\boldsymbol{w}_{1:N})}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\ &\propto \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K},\boldsymbol{\rho}_{1:N},\boldsymbol{w}_{1:N})}\left\{\log p(\boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\ &= \prod_{t=1}^{T} \exp\big(\mathsf{E}_{q(\boldsymbol{\theta}_{1:K},\boldsymbol{\rho}_{1:N},\boldsymbol{w}_{1:N})}\left\{\log p(\boldsymbol{r}_{1:K}^{t}, x_{1:N}^{t} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\ &\propto \prod_{t=1}^{T} q(\boldsymbol{r}_{1:K}^{t}) \end{aligned}$$

In order to keep the notation uncluttered, from now on we will omit the implicit subscripts in expectation operators. So each individual factor $q(\boldsymbol{r}_{1:K}^{t})$ above is equal to

$$\begin{aligned} & q(r_{1:K}^{t}) \propto \exp\big(\mathsf{E}\left\{\log p(r_{1:K}^{t}, x_{1:N}^{t} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\}\big) \\ &\quad = \exp\big(\sum_{k=1}^{K} \mathsf{E}\left\{\log p(r_k^t \mid r_{\pi(\boldsymbol{r}_k)}^t, \boldsymbol{\theta}_k)\right\} + \sum_{n=1}^{N} \mathsf{E}\left\{\log p(x_n^t \mid r_{\pi(\boldsymbol{x}_n)}^t, x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_{\boldsymbol{n}}, \boldsymbol{\rho}_{\boldsymbol{n}})\right\}\big) \\ &\quad \propto \exp\Big(\sum_{k=1}^{K} \mathsf{E}\left\{\log \boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}^t}(r_k^t)\right\} + \frac{1}{2}\sum_{n=1}^{N} \mathsf{E}\left\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}^t}\right\} \\ &\quad - \frac{1}{2}\sum_{n=1}^{N} \mathsf{E}\left\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}^t}\big(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}^t}{}^{\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t) - x_n^t\big)^2\right\} \\ &\quad \propto \text{Categorical}(r_{1:K}^t; \hat{\theta}^t) \end{aligned}$$

2. We now pursue the same strategy for the expression in (7)

$$\begin{aligned} q(\theta_{1:K}, \rho_{1:N}, w_{1:N}) \propto \exp\big(&\mathsf{E}_{q(\boldsymbol{r}_{1:K}^{1:T})}\left\{\log p(\theta_{1:K}, \rho_{1:N}, w_{1:N} \mid \boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T})\right\}\big) \\ &= \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} \exp\big(\mathsf{E}\left\{\log p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}} \mid \boldsymbol{r}_{1:K}^{1:T})\right\}\big)\Big) \\ &\Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} \exp\big(\mathsf{E}\left\{\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \mid \boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T})\right\}\big)\Big) \\ &\propto \Big(\prod_{k=1}^{K} \prod_{r_{\pi(\boldsymbol{r}_k)}} q(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\Big)\Big(\prod_{n=1}^{N} \prod_{r_{\pi(\boldsymbol{x}_n)}} q(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) \end{aligned}$$

where each individual factor turns out to be

$$\begin{aligned}
q(\theta_{k|r_{\pi(\boldsymbol{r}_k)}}) &\propto \exp\Big(\mathsf{E}\left\{\log p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}} \mid \boldsymbol{r}_{1:K}^{1:T})\right\}\Big) \\
&\propto \exp\left(\sum_{r_k}\left(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}} + \sum_{t=1}^{T} \mathsf{E}\left\{\mathbb{1}_{\{r_k^t=r_k\}}\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{r}_k)}=r_{\pi(\boldsymbol{r}_k)}\right\}}\right\} - 1\right)\log\theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right) \\
&\propto \text{Dirichlet}(\theta_{k|r_{\pi(\boldsymbol{r}_k)}}; \hat\gamma_{k|r_{\pi(\boldsymbol{r}_k)}})
\end{aligned}$$

$$\begin{aligned}
q(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}) &\propto \exp\Big(\mathsf{E}\left\{\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}} \mid \boldsymbol{r}_{1:K}^{1:T}, x_{1:N}^{1:T})\right\}\Big) \\
&\propto \exp\Big(\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}) + \mathsf{E}\left\{\log p(x_{1:N}^{1:T} \mid \boldsymbol{r}_{1:K}^{1:T}, w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})\right\}\Big) \\
&\propto \exp\Big(\log p(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}) \\
&\quad + \sum_{t=1}^{T}\mathsf{E}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)}=r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}\log p(x_n^t \mid x^t_{\pi(\boldsymbol{x}_n)}, w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) \\
&\propto \mathcal{NG}(w_{n|r_{\pi(\boldsymbol{x}_n)}}, \rho_{n|r_{\pi(\boldsymbol{x}_n)}}; \hat m_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat a_{n|r_{\pi(\boldsymbol{x}_n)}}, \hat b_{n|r_{\pi(\boldsymbol{x}_n)}})
\end{aligned}$$

Finally, we match the coefficients of the sufficient statistics in above equations with the natural parameters and find the following variational parameters in terms of the expected sufficient statistics:

$$\begin{aligned}
\log\hat\theta^t(r^t_{1:K}) &=^{+} \sum_{k=1}^{K}\mathsf{E}_{\mathcal{Q}}\left\{\log\theta_{k|r^t_{\pi(\boldsymbol{r}_k)}}(r_k^t)\right\} - \frac{1}{2}\sum_{n=1}^{N}\phi(x^t_{\pi(\boldsymbol{x}_n)})^{\mathrm{T}}\hat\Lambda^{-1}_{n|r^t_{\pi(\boldsymbol{x}_n)}}\phi(x^t_{\pi(\boldsymbol{x}_n)}) \\
&\quad + \frac{1}{2}\sum_{n=1}^{N}\mathsf{E}_{\mathcal{Q}}\left\{\log\rho_{n|r^t_{\pi(\boldsymbol{x}_n)}}\right\} - \frac{1}{2}\sum_{n=1}^{N}\big(\hat m^{\mathrm{T}}_{n|r^t_{\pi(\boldsymbol{x}_n)}}\phi(x^t_{\pi(\boldsymbol{x}_n)}) - x_n^t\big)^2\mathsf{E}_{\mathcal{Q}}\left\{\rho_{n|r^t_{\pi(\boldsymbol{x}_n)}}\right\} \\
\hat\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k) &= \gamma_{k|r_{\pi(\boldsymbol{r}_k)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{r}_k)}=r_{\pi(\boldsymbol{r}_k)}\right\}}\mathbb{1}_{\{r_k^t=r_k\}}\right\} \\
\hat\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} &= \Lambda_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)}=r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}\phi(x^t_{\pi(\boldsymbol{x}_n)})\phi(x^t_{\pi(\boldsymbol{x}_n)})^{\mathrm{T}} \\
\hat m_{n|r_{\pi(\boldsymbol{x}_n)}} &= \hat\Lambda^{-1}_{n|r_{\pi(\boldsymbol{x}_n)}}\big(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}m_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)}=r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}x_n^t\phi(x^t_{\pi(\boldsymbol{x}_n)})\big) \\
\hat a_{n|r_{\pi(\boldsymbol{x}_n)}} &= a_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2}\sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)}=r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \\
\hat b_{n|r_{\pi(\boldsymbol{x}_n)}} &= b_{n|r_{\pi(\boldsymbol{x}_n)}} + \frac{1}{2}\Big(m_{n|r_{\pi(\boldsymbol{x}_n)}}{}^{\mathrm{T}}\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}m_{n|r_{\pi(\boldsymbol{x}_n)}} \\
&\qquad - \hat m^{\mathrm{T}}_{n|r_{\pi(\boldsymbol{x}_n)}}\hat\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}\hat m_{n|r_{\pi(\boldsymbol{x}_n)}} + \sum_{t=1}^{T}\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)}=r_{\pi(\boldsymbol{x}_n)}\right\}}\right\}(x_n^t)^2\Big)
\end{aligned}$$

Algorithm 1 VB-CN: Variational inference for causal networks

```
Require: $\mathcal{G}$,$x_{1:N}^{1:T}$
  Initialize $\hat{\gamma}_{1:K}$, $\hat{m}_{1:N}$, $\hat{\Lambda}_{1:N}$, $\hat{a}_{1:N}$, $\hat{b}_{1:N}$
  repeat
    Update expected sufficient statistics
      • $\mathsf{E}_{\mathcal{Q}}\left\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \leftarrow \psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}) - \log \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}$
      • $\mathsf{E}_{\mathcal{Q}}\left\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\right\} \leftarrow \frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}}{\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}}$
      • $\mathsf{E}_{\mathcal{Q}}\left\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\right\} \leftarrow \psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\big(\sum_{r'_k} \hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)$
    for $t = 1, \ldots, T$ do
      Update $\log \hat{\theta}^t$
      Update expected sufficient statistics
        • $\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{r}_k)} = r_{\pi(\boldsymbol{r}_k)}\right\}} \mathbb{1}_{\left\{r^t_k = r_k\right\}}\right\} \leftarrow \sum_{r_{\neg U}} \hat{\theta}^t(r_{1:K})$ where $U = \pi(\boldsymbol{r}_k) \cup \{k\}$
        • $\mathsf{E}_{\mathcal{Q}}\left\{\mathbb{1}_{\left\{r^t_{\pi(\boldsymbol{x}_n)} = r_{\pi(\boldsymbol{x}_n)}\right\}}\right\} \leftarrow \sum_{r_{\neg U}} \hat{\theta}^t(r_{1:K})$ where $U = \pi(\boldsymbol{x}_n)$
    end for
    Update $\hat{\gamma}_{1:K}$, $\hat{m}_{1:N}$, $\hat{\Lambda}_{1:N}$, $\hat{a}_{1:N}$, $\hat{b}_{1:N}$ w.r.t. the expected sufficient statistics.
    Update $\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$ via Equation (8)
  until $\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$ converges
  return Variational parameters $\hat{\theta}^{1:T}$, $\hat{\gamma}_{1:K}$, $\hat{m}_{1:N}$, $\hat{\Lambda}_{1:N}$, $\hat{a}_{1:N}$, $\hat{b}_{1:N}$
  return The evidence lower bound $\mathcal{B}_{\mathcal{P}}[\mathcal{Q}]$.
```

A simplified sketch of our variational inference algorithm *VB-CN* is also presented in Algorithm 1.

C.1. Evidence Lower Bound

ELBO can be expressed as a sum of expectation terms most of which are in the form of negative cross entropy or negative entropy:

$$\mathcal{B}_{\mathcal{P}}[\mathcal{Q}] \equiv \mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}^{1:T}_{1:K}, x^{1:T}_{1:N}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N}) - \log \mathcal{Q}(\boldsymbol{r}^{1:T}_{1:K}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\rho}_{1:N}, \boldsymbol{w}_{1:N})\right\} \tag{8}$$

$$= \sum_{t=1}^{T}\sum_{n=1}^{N} \mathsf{E}_{\mathcal{Q}}\left\{\log p(x^t_n \mid x^t_{\pi(\boldsymbol{x}_n)}, \boldsymbol{r}^t_{\pi(\boldsymbol{x}_n)}, \boldsymbol{w}_n, \boldsymbol{\rho}_n)\right\} \tag{9}$$

$$+ \sum_{t=1}^{T}\Big(\sum_{k=1}^{K} \mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{r}^t_k \mid \boldsymbol{r}^t_{\pi(\boldsymbol{r}_k)}, \boldsymbol{\theta}_k)\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{r}^t_{1:K})\right\}\Big) \tag{10}$$

$$+ \sum_{k=1}^{K}\sum_{r_{\pi(\boldsymbol{r}_k)}} \Big(\mathsf{E}_{\mathcal{Q}}\left\{\log p(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\theta_{k|r_{\pi(\boldsymbol{r}_k)}})\right\}\Big) \tag{11}$$

$$+ \sum_{n=1}^{N}\sum_{r_{\pi(\boldsymbol{x}_n)}} \Big(\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\} - \mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\}\Big) \tag{12}$$

In this section we will evaluate each of those expectations explicitly. We start our derivation with the trickier Gaussian log-likelihood term, then the rest of the expectations will correspond to negative cross entropy values of standard exponential family distributions:

$$
\begin{aligned}
&\mathsf{E}_{\mathcal{Q}}\Big\{\log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_n, \boldsymbol{\rho}_n)\Big\} \\
&\quad= \sum_{r_{\pi(\boldsymbol{x}_n)}} \mathsf{E}\bigg\{\mathbb{1}_{\big\{\boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\big\}}\bigg\}\, \mathsf{E}\Big\{\log p(x_n^t \mid x_{\pi(\boldsymbol{x}_n)}^t, \boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}, \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\Big\} \\
&\quad= \frac{1}{2}\sum_{r_{\pi(\boldsymbol{x}_n)}} \mathsf{E}\bigg\{\mathbb{1}_{\big\{\boldsymbol{r}_{\pi(\boldsymbol{x}_n)}^t = r_{\pi(\boldsymbol{x}_n)}\big\}}\bigg\} \Big(\mathsf{E}\Big\{\log \boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\Big\} \\
&\quad -\mathsf{E}\Big\{\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}}\big(x_n^t - \boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}}{}^{\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t)\big)^2\Big\} - \log 2\pi\Big) \\
&\quad= \frac{1}{2}\sum_{r_{\pi(\boldsymbol{x}_n)}}\sum_{r_{\neg\pi(\boldsymbol{x}_n)}} \hat{\theta}^t(r_{\pi(\boldsymbol{x}_n)}, r_{\neg\pi(\boldsymbol{x}_n)})\Big(\psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}) - \log \hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}} \\
&\quad -\frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}}{\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}}\big(x_n^t - \hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}^{\mathrm{T}}\phi(x_{\pi(\boldsymbol{x}_n)}^t)\big)^2 - \phi(x_{\pi(\boldsymbol{x}_n)}^t)^{\mathrm{T}}\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}^{-1}\phi(x_{\pi(\boldsymbol{x}_n)}^t) - \log 2\pi\Big)
\end{aligned}
$$

Variational distribution $\mathcal{Q}$ treats $r_{1:K}^t$ and $\theta_{1:K}$ as independent variables. So, the expectations of the categorical log-likelihood terms admit the following form

$$
\begin{aligned}
&\mathsf{E}_{\mathcal{Q}}\Big\{\log p(\boldsymbol{r}_k^t \mid \boldsymbol{r}_{\pi(\boldsymbol{r}_k)}^t, \boldsymbol{\theta}_k)\Big\} = \sum_{r_k}\sum_{r_{\pi(\boldsymbol{r}_k)}} \mathsf{E}\bigg\{\mathbb{1}_{\{\boldsymbol{r}_k^t = r_k\}}\mathbb{1}_{\big\{\boldsymbol{r}_{\pi(\boldsymbol{r}_k)}^t = r_{\pi(\boldsymbol{r}_k)}\big\}}\bigg\}\, \mathsf{E}\Big\{\log \theta_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\Big\} \\
&= \sum_{r_{1:K}} \hat{\theta}^t(r_{1:K})\Big(\psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \psi\big(\sum_{r'_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)\Big)
\end{aligned}
$$

The rest of the terms are related to cross entropy or entropy of the well-known exponential family distributions, and closed form expressions for them are supplied in Appendix B. So here, we only modify these expressions by changing their parameters with the appropriate variational parameters.

1. By using the negative cross entropy formulation in Appendix B.1.3 for categorical distributions:

$$
\mathsf{E}_{\mathcal{Q}}\big\{\log q(\boldsymbol{r}_{1:K}^t)\big\} = \sum_{r_{1:K}} \hat{\theta}^t(r_{1:K}) \log \hat{\theta}^t(r_{1:K})
$$

2. By using the Dirichlet negative cross entropy formulation in Appendix B.1.2:

$$\begin{aligned}\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}})\right\} &= \log\Gamma\big(\sum_{r_k}\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \sum_{r_k}\log\Gamma\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)\\ &+\sum_{r_k}\big(\gamma_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)-1\big)\Big(\psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)-\psi\big(\sum_{r'_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)\Big)\end{aligned}$$

$$\begin{aligned}\mathsf{E}_{\mathcal{Q}}\left\{\log q(\boldsymbol{\theta}_{k|r_{\pi(\boldsymbol{r}_k)}})\right\} &= \log\Gamma\big(\sum_{r_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big) - \sum_{r_k}\log\Gamma\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)\\ &+\sum_{r_k}\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)-1\big)\Big(\psi\big(\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r_k)\big)-\psi\big(\sum_{r'_k}\hat{\gamma}_{k|r_{\pi(\boldsymbol{r}_k)}}(r'_k)\big)\Big)\end{aligned}$$

3. Finally, by using the Multivariate Normal-Gamma negative cross entropy formulation in Appendix B.1.7:

$$\begin{aligned}&\mathsf{E}_{\mathcal{Q}}\left\{\log p(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}},\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\} =\\ &a_{n|r_{\pi(\boldsymbol{x}_n)}}\log b_{n|r_{\pi(\boldsymbol{x}_n)}} - \log\Gamma(a_{n|r_{\pi(\boldsymbol{x}_n)}}) + \frac{1}{2}\log\det(\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}) - \frac{1}{2}\mathrm{tr}(\hat{\Lambda}^{-1}_{n|r_{\pi(\mathbf{x}_n)}}\Lambda_{n|r_{\pi(\mathbf{x}_n)}})\\ &-\frac{M}{2}\log 2\pi + \Big(a_{n|r_{\pi(\boldsymbol{x}_n)}}+\frac{M}{2}-1\Big)\Big(\psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}})-\log\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) - b_{n|r_{\pi(\boldsymbol{x}_n)}}\frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}}{\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}}\\ &-\frac{\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}}{2\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}}}(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}-m_{n|r_{\pi(\boldsymbol{x}_n)}})^{\mathrm{T}}\Lambda_{n|r_{\pi(\boldsymbol{x}_n)}}(\hat{m}_{n|r_{\pi(\boldsymbol{x}_n)}}-m_{n|r_{\pi(\boldsymbol{x}_n)}})\end{aligned}$$

$$\begin{aligned}&\mathsf{E}\left\{\log q(\boldsymbol{w}_{n|r_{\pi(\boldsymbol{x}_n)}},\boldsymbol{\rho}_{n|r_{\pi(\boldsymbol{x}_n)}})\right\} =\\ &\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}\log\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}} - \log\Gamma(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}) + \frac{1}{2}\log\det(\hat{\Lambda}_{n|r_{\pi(\boldsymbol{x}_n)}}) - \frac{M}{2}\\ &-\frac{M}{2}\log 2\pi + \Big(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}+\frac{M}{2}-1\Big)\Big(\psi(\hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}})-\log\hat{b}_{n|r_{\pi(\boldsymbol{x}_n)}})\Big) - \hat{a}_{n|r_{\pi(\boldsymbol{x}_n)}}\end{aligned}$$

Appendix D. About Experiments

D.1. Hyperparameter settings

Given that some of our experiments are computationally demanding, certain parameters were fixed for all our experiments when it was reasonable to do so, in order to avoid excessive computational costs. For all experiments, the basis function were allowed to be of linear, quadratic, and cubic order, and the cardinality $|\mathcal{R}_1|$ of the latent variable was allowed to range between 1 and 5. For the bivariate models in Figure 2, the cardinality $|\mathcal{R}_1|$ of the latent variable $\boldsymbol{r}_1$ was allowed to range between 1 and 5, in each case the cardinality and basis function order that leads to the highest marginal likelihood was selected. As the parameters we fixed before inference: for both values of $n \in \{1, 2\}$ and for all values of

$r_1 \in \mathcal{R}_1$, $m_{n|r_1}$'s were set to 0, and $\Lambda_{n|r_1}$'s were set to $\frac{1}{10}I$ each; while for all values of $r_1 \in \mathcal{R}_1$, $\gamma_1(r_1)$'s were set to 10.

We next describe the remaining hyperparameters with respect to the causal graph in Figure 2(b) in which $\boldsymbol{x}_1$ causes $\boldsymbol{x}_2$. Their adaptation to other two graphs is straightforward due to symmetry. The hyperparameters of the Gamma distributions, (a_1, b_1, a_2, b_2), from which the precision of the observed variables were drawn, were allowed to take different values with the condition that $a_{n|r_1} \geq b_{n|r_1}$ at all times, but again every element of these vectors corresponding to different values of r_1 assumed to be constant within the vector. This is because the mean of a Gamma distribution Gamma(a, b) is a/b and its variance is a/b^2, therefore when b is allowed to take a greater value than a, this results in a close to zero precision value for the relevant distribution for the observed variable. Obeying the constraint, the a and b's were allowed to take values among 1, 10, and 100 each. The a parameter was not allowed to be larger than 100 since this leads to an equivalent sample size much larger than the sample size of certain data sets used in experiments, effectively rendering the observations unimportant. The b parameter was not allowed to be smaller than 1 since this again implies extremely imprecise Gaussian distributions for the observed variables to which the Gamma distribution provided the precision variable. The combinations with these constraints lead to a total of 36 sets of hyperparameters.

While doing model comparison in a hyperparameter setting, we expect several criteria to be satisfied for maintaining consistency. For instance, in the spurious model (Figure 2(a)) there is no reason to assign different priors on variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Otherwise, just by permuting the labels of the pairs, we would obtain inconsistent marginal likelihoods. Likewise, when the labels of a pair are permuted, e.g. $\hat{x}_1^{1:T} \equiv x_2^{1:T}$ and $\hat{x}_2^{1:T} \equiv x_1^{1:T}$, we expect the marginal likelihood of the pair $(x_1^{1:T}, x_2^{1:T})$ given the relation $\boldsymbol{x}_1 \rightarrow \boldsymbol{x}_2$ to be equal to the marginal likelihood of the permuted pair $(\hat{x}_1^{1:T}, \hat{x}_2^{1:T})$ given the relation $\hat{\boldsymbol{x}}_2 \rightarrow \hat{\boldsymbol{x}}_1$. The rule we used to solve inconsistency issues in such situations is the following: the prior parameters of two variables must be identical whenever the parental graphs of them are *homomorphic*. So, if we are calculating the marginal likelihood of the relation $\boldsymbol{x}_1 \rightarrow \boldsymbol{x}_2$ with a particular hyperparameter setting, say ($a_1 = 100$, $b_1 = 10$, $a_2 = 10$, $b_2 = 1$), then the corresponding consistent hyperparameter setting for $\boldsymbol{x}_2 \rightarrow \boldsymbol{x}_1$ should be ($a_1 = 10$, $b_1 = 1$, $a_2 = 100$, $b_2 = 10$), whereas the corresponding consistent hyperparameters for the spurious relationship should be ($a_1 = 100$, $b_1 = 10$, $a_2 = 100$, $b_2 = 10$).

D.2. Synthetic Data Experiments

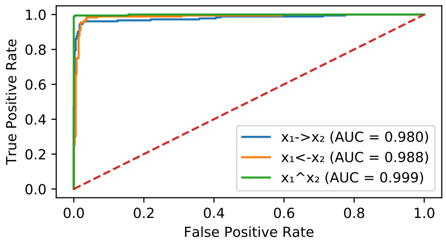


Figure 3: The ROC curves for synthetic data experiments.

For this experiment, for each of 36 hyperparameter combinations, and for each rank values of $|\mathcal{R}_1| = 1$ to 5 for the linear model, a total of 3 different data pairs (one for each different graphical model) with 2000 observations were generated. This amounted to a total of 540 data pairs. For each synthetic data pair, the corresponding hyperparameters were used to compare the three hypotheses demonstrated in Figure 2 using the marginal likelihood estimate of the variational Bayes algorithm. The resulting ROC curves can be seen in the Figure 3. With an overall accuracy of .961 and AUC of .998, the results demonstrate that our method can identify the data generating graph comfortably, given the correct hyperparameter settings.

D.3. Detecting Spurious Relationships in the CEP Data Set

The CEP data set is not labeled as to the spurious relationships, therefore it is not possible to conduct hyperparameter selection with cross-validation. However, we ran the experiments again, this time including the spurious relationship hypothesis in the experiments, for all 36 parameter settings, and recorded the pairs for which the marginal likelihood of the spurious hypothesis was the highest. We observed that, using the hyperparameter setting that achieved the highest accuracy in the previous experiment, these four data sets were found to be spurious: 19, 91, 92, and 98. The scatter plots of these data sets are presented in Figure 4.

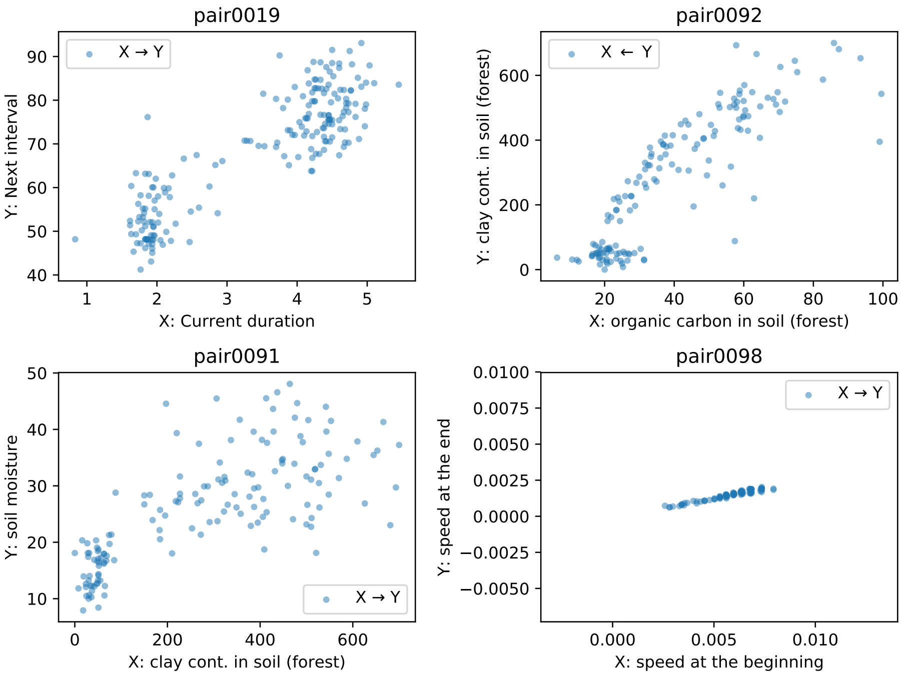


Figure 4: Scatter plots of spurious pairs found in Cause Effect Pairs.

Visual examination of the first three pairs reveals that, although each of these pairs are correlated, they can be separated into two clusters in which X and Y axes become independent. In other words, once the confounding variables governing the cluster affiliations are decided, then the variables X and Y generated independently, so their correlation is indeed spurious. As we lack the expertise, we do not know what these confounding variables correspond in reality, but the existence of such variables is evident from the scatter plots. The case of the fourth spurious pair is slightly different than other correlated pairs. The fourth pair consists of the measurements of initial and final speeds of a ball on a ball track where initial speed is thought as the cause of final speed. However, our variational algorithm selected the spurious model with a latent variable having cardinality $|\mathcal{R}_1| = 1$, which actually corresponds to the marginal independence of X and Y. Such an explanation makes sense considering the plot in Figure 4, as the initial speed of the ball does not seem related to its final speed.