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, allowing the identification of a unique causal graph. Adopting a Bayesian approach also allows for straightforward modeling of unobserved confounding variables, omission of which might engender misleading causal deductions. However, this comes at the cost of rendering the computation of marginal likelihood intractable, therefore we present a variational algorithm to approximate marginal likelihood values in the presence of latent confounding variables. We demonstrate state-of-the-art results in applications of our approach, validating both our modeling approach and our inference methodology. We believe that the Bayesian approach to causal discovery both allows the rich methodology of approximate Bayesian inference to be used in various difficult aspects of this problem and provides a unifying framework to the fragmented scenery of causal discovery research.

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 in a way that identifies the unique causal generative graph. One of the main factors limiting the usefulness of Bayesian model selection based approaches in this case is that most such methods assign equal scores to Markov equivalent graphs, thereby limiting themselves to detecting causal graphs up to Markov equivalence (Huang et al., 2018; Heckerman, 1995). This inability can be most iconically exemplified by such methods not being able to discriminate between the graphs \( X \rightarrow Y \) and \( Y \rightarrow X \).

The key notion underlying our solution to this problem the widely accepted principle of independence of the cause-effect mechanisms (Janzing et al., 2012), that is, the natural

∗ Equal contribution

© M. Kurutmaz, M. Barsbey, A. Cemgil, S. Yıldırım & U. Şimşekli.
mechanisms that generate the cause and the effect based on cause must be independent of each other. For example, that more people use the drug should not affect the effect of the drug on people who use them. We embody this assumption by assuming the mutual independence of the parameters pertaining to cause and effect distributions (Zhang et al., 2015) in a Bayesian model, a line of reasoning that is natural to this modeling perspective, where parameters are modeled as random variables (Heckerman et al., 1995; Geiger et al., 1997; Spiegelhalter et al., 1993; Blei et al., 2003; Geiger et al., 1997). 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 among Markov equivalent graphs, leading to the identification of the a-posteriori most likely CN.

In Section 2, we introduce an example model that combines the advantages of many others in the literature such as handling causal insufficiency, identifying a unique causal graph, being able to model nonlinear relations, and theoretically being able to work on arbitrary causal graphs structures1. To approximate the marginal likelihood in graphs which include latent confounders, we present a variational inference algorithm in Section 3. Finally, we test our approach on various real data sets in Section 4.

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 $\{x_1, \ldots, x_n, \ldots, x_N\} \subseteq V_\mathcal{G}$ denote the set of continuous random variables, and similarly $\{r_1, \ldots, r_k, \ldots, r_K\} \subseteq V_\mathcal{G}$ denote the discrete latent variables of the network where each $x_n$ and each $r_k$ are defined in the domains $\mathcal{X}_n$ and $\mathcal{R}_k$, respectively. The set of parent vertices of a vertex $v \in V_\mathcal{G}$ is denoted by $\pi(v)$, while we denote its continuous parents by $x_{\pi(v)}$, and discrete parents by $r_{\pi(v)}$.

For the scope of this text, we specify conditional distributions for the graphs as follows: we assume categorical distributions on the discrete variables $r_{1,K}$ and linear basis functions with Gaussian noise on the continuous variables $x_{1,N}$. 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 $\mathcal{G}$:

\begin{align}
\forall k, t: \quad r_k^t \mid r_{\pi(r_k)} &\sim \text{Categorical}(\theta_k | r_{\pi(r_k)}) \\
\forall n, t: \quad x_n^t \mid x_{\pi(x_n)}, r_{\pi(x_n)} &\sim N(w_n | r_{\pi(x_n)}^t, \rho_n | r_{\pi(x_n)}^{-1})
\end{align}

where $1 \leq t \leq T$, $\phi$ is an arbitrary basis function with the convention $\phi(\{\}) = 1$, and $\theta_k | r_{\pi(r_k)}$, $w_n | r_{\pi(x_n)}$, $\rho_n | r_{\pi(x_n)}$’s are the parameters of the conditional distributions. Namely, $\theta_k$ is the conditional distribution table of $r_k$, $w_n$ is the weights of the basis functions, and $\rho_n$ is the precision parameter of the conditional distribution of $x_n$.

---

1. See Glymour et al. (2019) and Spirtes and Zhang (2016) for recent reviews of the literature.
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

\[ \forall k, \gamma_k : \theta_{k|\gamma_k} \sim \text{Dirichlet}(\gamma_k) \]  
\[ \forall n, r_n : w_{n|r_n}, \rho_{n|r_n} \sim \mathcal{N}(m_{n|r_n}, \Lambda_{n|r_n}, a_{n|r_n}, b_{n|r_n}) \]

where \(\gamma_k, m_{n|r_n}, \Lambda_{n|r_n}, a_{n|r_n}, b_{n|r_n}\) are the prior parameters, i.e. hyper-parameters, 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 \(P\) is approximated by a variational distribution \(Q\) via minimizing Kullback-Leibler divergence \(\text{KL}(Q||P)\). In the context of Bayesian model selection, minimization of the \(\text{KL}(Q||P)\) 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

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

where \(P = p(r_{1:T}^{1:K}, \theta_{1:K}, \rho_{1:N}, w_{1:N} | x_{1:T}^{1:N})\) is the full posterior distribution, and ELBO is denoted by \(\mathcal{B}_P[Q]\). In a typical scenario of VB, \(Q\) is assumed to be a member of a restricted family of distributions. In its most common form, also known as mean-field approximation, \(Q\) 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

\[ Q(r_{1:T}^{1:K}, \theta_{1:K}, \rho_{1:N}, w_{1:N}) = q(r_{1:T}^{1:K}) q(\theta_{1:K}) q(\rho_{1:N}, w_{1:N}) \]

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

\[ q(r_{1:K}^t) = \text{Categorical}(r_{1:K}^t; \hat{\theta}^t) \]
\[ q(\theta_{k|r_k}) = \text{Dirichlet}(\theta_{k|r_k}; \hat{\gamma}_k) \]
\[ q(w_{n|r_n}, \rho_{n|r_n}) = \mathcal{N}(\hat{m}_{n|r_n}, \hat{\Lambda}_{n|r_n}, \hat{a}_{n|r_n}, \hat{b}_{n|r_n}) \]

Here \(\hat{\theta}^t, \hat{\gamma}_k, \hat{m}_{n|r_n}, \hat{\Lambda}_{n|r_n}, \hat{a}_{n|r_n}, \hat{b}_{n|r_n}\) 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. Experimental Results and Discussion

We illustrate the performance of our approach on two real world examples\textsuperscript{2}.

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 CauseEffectPairs (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. We tested our algorithm on the data set by using $10 \times 3$ 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. 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. For the CEP data set, we obtained a mean accuracy of $0.78 \pm 0.09$ and AUC score of $0.84 \pm 0.13$ (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.

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 Thyroid data set from the UCI repository (Dheeru and Karra Taniskidou, 2017). This data set 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 in this case 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 approximated posterior distribution of the cardinality of the latent variable is maximized at 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.

Overall, we show that Bayesian model selection is a promising framework that can facilitate causal research significantly both through increased performance and conceptual unification. Given that Bayesian modeling is agnostic to specific variable types, conditional

\textsuperscript{2} We also test our approach on synthetic data; these experiments can be found in Appendix D.2
Bayesian Model Selection for Identifying Markov Equivalent Causal Graphs

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

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


Bayesian Model Selection for Identifying Markov Equivalent Causal Graphs


Appendix A. Identifiability of Markov Equivalent Graphs

Remember that 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:

\[ x_1 = w_1(1) + \frac{\rho_1}{2} \epsilon_1 \quad \epsilon_1 \sim N(0, 1) \]
\[ x_2 = w_2(1) x_1 + w_2(2) + \frac{\rho_2}{2} \epsilon_2 \quad \epsilon_2 \sim N(0, 1) \]

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

\[ x_2 = w_1(1) w_2(1) + w_2(2) + \frac{\rho_2}{2} \hat{\epsilon}_2 = \hat{w}_2(1) + \frac{\rho_2}{2} \hat{\epsilon}_2 \quad \hat{\epsilon}_2 \sim N(0, 1) \]
\[ x_1 = \frac{w_2(2)}{w_2(1)} x_2 - \frac{w_2(2)}{w_2(1)} + \frac{\rho_1}{2} \hat{\epsilon}_1 = \hat{w}_1(1) x_2 - \hat{w}_1(2) + \frac{\rho_1}{2} \hat{\epsilon}_1 \quad \hat{\epsilon}_1 \sim N(0, 1) \]

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 \( x_1 \) and \( x_2 \) boils down to one of three cases due to causal Markov condition (Hausman and Woodward, 1999):

1. \( x_1 \) causes \( x_2 \),
2. \( x_2 \) causes \( x_1 \),
3. they do not cause each other, but a latent variable \( r_1 \) causes both of them.
Bayesian Model Selection for Identifying Markov Equivalent Causal Graphs

Figure 2: Graphical models for bivariate causality.

Associated causal networks corresponding to each of these hypotheses are depicted in Figure 2, where latent variable $r_1$ represents the overall effect of the all unobserved variables. For the spurious relationship (Figure 2(a)), marginally correlated variables $x_1$ and $x_2$ become independent once the latent common cause variable $r_1$ 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 $x_2$ and $r_1$ are known, $x_1$ and the parameters of $x_2$ are dependent only in the case of $x_1 \to x_2$, or knowing $r_1$ makes $x_1$ and $x_2$ 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:

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

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

2. Gamma density:

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

where $a$ is the shape and $b$ is the rate parameter.
3. Expected sufficient statistics:
\[ E\{\rho\} = a/b, \quad E\{\log \rho\} = \psi(a) - \log(b) \]

4. Cross entropy:
\[
E_{\text{Gamma}(\rho; \hat{a}, \hat{b})} \{-\log \text{Gamma}(\rho; a, b)\} = -(a - 1)E\{\log \rho\} + bE\{\rho\} + \log \Gamma(a) - a \log b
\]
\[
= -(a - 1)(\psi(\hat{a}) - \log(\hat{b})) + \frac{\hat{a}b}{b} + \log \Gamma(a) - a \log b
\]

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

B.1.2. Dirichlet Distribution

1. Multivariate Beta function:
\[ B(\gamma) = \prod_r \frac{\Gamma(\gamma_r)}{\Gamma(\sum_r \gamma_r)} \]

2. Dirichlet density:
\[ \text{Dirichlet}(\theta; \gamma) = \frac{1}{B(\gamma)} \exp\left(\sum_r (\gamma_r - 1) \log \theta_r\right) \]

3. Expected sufficient statistics:
\[ E\{\theta_r\} = \frac{\gamma_r}{\sum_m \gamma_m}, \quad E\{\log \theta_r\} = \psi(\gamma_r) - \psi(\sum_m \gamma_m) \]

4. Cross entropy:
\[
E_{\text{Dirichlet}(\theta; \hat{\gamma})} \{-\log \text{Dirichlet}(\theta; \gamma)\} = \log B(\gamma) - \sum_r (\gamma_r - 1)E\{\log \theta_r\}
\]
\[
= \log B(\gamma) - \sum_r (\gamma_r - 1)(\psi(\gamma_r) - \psi(\sum_m \gamma_m))
\]

B.1.3. Categorical Distribution

1. Categorical density:
\[ \text{Categorical}(r; \theta) = \prod_{k=1}^K \theta_{k}^{1_{\{r=k\}}} \]

2. Expected sufficient statistics:
\[ E\{1_{\{r=k\}}\} = \theta_k \]

3. Cross entropy:
\[
E_{\text{Categorical}(r; \hat{\theta})} \{-\log \text{Categorical}(r; \theta)\} = -\sum_k \hat{\theta}_k \log \theta_k
\]
B.1.4. Normal Distribution

1. Normal density:

\[ x \sim \mathcal{N}(\mu, \rho^{-1}) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2} \log \rho - \frac{1}{2} \rho (x - \mu)^2\right) \]

where \( \mu \) is the mean parameter and \( \rho \) is the precision parameter, i.e. \( \rho^{-1} \) is the variance.

2. Expected sufficient statistics

\[ \mathbb{E}\{x\} = \mu \quad \mathbb{E}\{x^2\} = \mu^2 + \rho^{-1} \]

B.1.5. Multivariate Normal Distribution

1. Multivariate Normal density:

\[ x \sim \mathcal{N}(\mu, \Lambda^{-1}) = \frac{1}{(2\pi)^{K/2}} \exp\left(\frac{1}{2} \log \det(\Lambda) - \frac{1}{2} (x - \mu)^T \Lambda (x - \mu)\right) \]

where \( \mu \) is the mean vector and \( \Lambda \) is the precision matrix, i.e. \( \Lambda^{-1} \) is the covariance matrix.

2. Expected sufficient statistics:

\[ \mathbb{E}\{x\} = \mu \quad \mathbb{E}\{x^T A x\} = \mu^T A \mu + \text{tr}(\Lambda^{-1} A) \]

for any symmetric matrix \( A \).

B.1.6. Normal-Gamma Distribution

1. Normal-Gamma density:

\[ \mu, \rho \sim \mathcal{NG}(m, \lambda, a, b) = \frac{\lambda^{a/2}}{\Gamma(a) \sqrt{2\pi}} \exp\left(\frac{1}{2} \log \lambda - \frac{1}{2} \rho (\mu - m)^2\right) \]

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

\[ \rho \sim \text{Gamma}(a, b) \quad x \mid \rho \sim \mathcal{N}(m, (\lambda \rho)^{-1}) \]

2. Expected sufficient statistics:

\[ \mathbb{E}\{\log \rho\} = \psi(a) - \log b \quad \mathbb{E}\{\rho\} = \frac{a}{b} \quad \mathbb{E}\{\rho \mu\} = m \frac{a}{b} \quad \mathbb{E}\{\rho \mu^2\} = \frac{1}{\lambda} + m^2 \frac{a}{b} \]

3. Cross entropy:

\[ \mathbb{E}_{\mathcal{NG}(m, \lambda, a, b)} \{-\log \mathcal{NG}(\mu, \rho; m, a, b)\} = \]

\[ -a \log b + \log \Gamma(a) - \frac{1}{2} \log \lambda + \frac{\lambda}{2\lambda} + \frac{1}{2} \log 2\pi \]

\[ -\left(a - \frac{1}{2}\right)(\psi(\hat{a}) - \log \hat{b}) + \frac{\hat{a}}{2b} \lambda (\hat{m} - m)^2 \]
B.1.7. Multivariate Normal-Gamma Distribution

1. Multivariate Normal-Gamma density:

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

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

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

2. Expected sufficient statistics:

\[
\begin{aligned}
E\{\log \rho\} &= \psi(a) - \log b \\
E\{\rho\} &= \frac{a}{b} \\
E\{\rho w\} &= \frac{a}{b} m \\
E\{\rho w^T A w\} &= \text{tr}(\Lambda^{-1} A) + \frac{a}{b} m^T A m \\
\end{aligned}
\]

for any symmetric matrix \(A\).

3. Cross entropy:

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

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 \(\theta\):

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

where \(\gamma^*_r = \gamma_r + \sum_{t=1}^T \mathbb{1}_{\{r=r^t\}}\)
B.2.2. Normal-Gamma-Normal Model

1. Generative model:

\[
\mu, \rho \sim \mathcal{NG}(m, \lambda, a, b)
\]
\[
x^1, \ldots, x^T \sim \mathcal{N}(\mu, \rho^{-1})
\]

2. Posterior of \( \mu \) and \( \rho \):

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

where

\[
\lambda^* \equiv \lambda + T
\]
\[
m^* \equiv \frac{\lambda m + \sum_t x^t}{\lambda^*}
\]
\[
a^* \equiv a + \frac{T}{2}
\]
\[
b^* \equiv b + \frac{1}{2} \left( \lambda m^2 - \lambda^* m^* + \sum_t (x^t)^2 \right)
\]

B.2.3. Bayesian Linear Regression

1. Generative model:

\[
y^t = w^T x^t + \rho^{-1/2} e^t
\]
\[
e^t \sim \mathcal{N}(0, 1)
\]

An equivalent description with Normal-Gamma priors is

\[
w, \rho \sim \mathcal{NG}(m, \Lambda, a, b)
\]
\[
y^t \mid x^t \sim \mathcal{N}(w^T x^t, \rho^{-1})
\]

2. Posterior of \( w \) and \( \rho \):

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

where

\[
\Lambda^* \equiv \Lambda + \sum_t x^t x^T
\]
\[
m^* \equiv \Lambda^{-1} + \sum_t y^t x^t
\]
\[
a^* \equiv a + \frac{T}{2}
\]
\[
b^* \equiv b + \frac{1}{2} \left( \Lambda m^2 - \Lambda^* m^* + \sum_t (x^t)^2 \right)
\]

Appendix C. Variational Bayes

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

\[
q(r^T_{1:K}) \propto \exp \left( E_{q(\theta_{1:K}, \rho_{1:N}, w_{1:N})} \left\{ \log p(r^T_{1:K}, x^T_{1:N}, \theta_{1:K}, \rho_{1:N}, w_{1:N}) \right\} \right)
\]
\[
q(\theta_{1:K}, \rho_{1:N}, w_{1:N}) \propto \exp \left( E_{q(r^T_{1:K})} \left\{ \log p(r^T_{1:K}, x^T_{1:N}, \theta_{1:K}, \rho_{1:N}, w_{1:N}) \right\} \right)
\]

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

\[ q(r^{1:T}_{1:K}) \propto \exp(E_q(\theta_{1:K}, \rho_{1:N}, w_{1:N}) \{ \log p(r^{1:T}_{1:K}, x^{1:T}_{1:1:N}, \theta_{1:K}, \rho_{1:N}, w_{1:N}) \}) \]

\[ \propto \exp(E_q(\theta_{1:K}, \rho_{1:N}, w_{1:N}) \{ \log p(r^{1:T}_{1:K}, x^{1:T}_{1:1:N} | \theta_{1:K}, \rho_{1:N}, w_{1:N}) \}) \]

\[ = \prod_{t=1}^{T} \exp(E_q(\theta_{1:K}, \rho_{1:N}, w_{1:N}) \{ \log p(r^{t}_{1:K}, x^{t}_{1:1:N} | \theta_{1:K}, \rho_{1:N}, w_{1:N}) \}) \]

\[ \propto \prod_{t=1}^{T} q(r^{t}_{1:K}) \]

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

\[ q(r^{t}_{1:K}) \propto \exp(E \{ \log p(r^{t}_{1:K}, x^{t}_{1:1:N} | \theta_{1:K}, \rho_{1:N}, w_{1:N}) \}) \]

\[ = \exp(\sum_{k=1}^{K} E \{ \log p(r^{t}_{k} | \pi^{t}_{(r_{k})}, \theta_{k}) \}) + \sum_{n=1}^{N} E \{ \log p(x^{t}_{n} | \pi^{t}_{(x_{n})}, x^{t}_{n}, w_{n}, \rho_{n}) \}) \]

\[ \propto \exp(\sum_{k=1}^{K} E \{ \log \theta_{k|r_{s}(r_{k})} \} + \frac{1}{2} \sum_{n=1}^{N} E \{ \log \rho_{n|r_{s}(x_{n})} \} ) \]

\[ - \frac{1}{2} \sum_{n=1}^{N} E \{ \rho_{n|r_{s}(x_{n})} (w_{n|r_{s}(x_{n})} T \phi(x^{t}_{n}) - x^{t}_{n})^{2} \} \]

\[ \propto \text{Categorical}(r^{t}_{1:K}; \hat{\theta}^{t}) \]

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

\[ q(\theta_{1:K}, \rho_{1:N}, w_{1:N}) \propto \exp(E_{q(r^{1:T}_{1:K})} \{ \log p(\theta_{1:K}, \rho_{1:N}, w_{1:N} | r^{1:T}_{1:K}, x^{1:T}_{1:1:N}) \}) \]

\[ = \left( \prod_{k=1}^{K} \prod_{r_{s}(r_{k})} \exp\left( E \left\{ \log p(\theta_{k|r_{s}(r_{k})} | r^{1:T}_{1:K}) \right\} \right) \right) \]

\[ \left( \prod_{n=1}^{N} \prod_{r_{s}(x_{n})} \exp\left( E \left\{ \log p(w_{n|r_{s}(x_{n})}, \rho_{n|r_{s}(x_{n})} | r^{1:T}_{1:K}, x^{1:T}_{1:1:N}) \right\} \right) \right) \]

\[ \propto \left( \prod_{k=1}^{K} \prod_{r_{s}(r_{k})} q(\theta_{k|r_{s}(r_{k})}) \right) \left( \prod_{n=1}^{N} \prod_{r_{s}(x_{n})} q(w_{n|r_{s}(x_{n})}, \rho_{n|r_{s}(x_{n})}) \right) \]

13
Bayesian Model Selection for Identifying Markov Equivalent Causal Graphs

where each individual factor turns out to be

\[ q(\theta_k | r_{\pi(r_k)}) \propto \exp\left(\mathbb{E}\left\{\log p(\theta_k | r_{\pi(r_k)}) \mid r_{1:T}^1:K\right\}\right) \]

\[ \propto \exp\left(\sum_{r_k} (\gamma_k | r_{\pi(r_k)}) + \sum_{t=1}^T \mathbb{E}\left\{\mathbb{I}\left\{r_0 = r_k\right\}\mathbb{I}\left\{r_{\pi(r_k)} = r_{\pi(r_k)}\right\}\right\} - 1\right) \log q(\theta_k | r_{\pi(r_k)}(r_k)) \]

\[ \propto \text{Dirichlet}(\theta_k | r_{\pi(r_k)}; \gamma_k | r_{\pi(r_k)}) \]

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:

\[ \log \hat{\gamma}_k | r_{\pi(r_k)}(r_k) = \gamma_k | r_{\pi(r_k)}(r_k) + \sum_{t=1}^T \mathbb{E}\left\{\mathbb{I}\left\{r_{\pi(r_k)} = r_{\pi(r_k)}\right\}\mathbb{I}\left\{r_0 = r_k\right\}\right\} \]

\[ \hat{\Lambda}_n | r_{\pi(a_n)}(a_n) = \Lambda_n | r_{\pi(a_n)} + \sum_{t=1}^T \mathbb{E}\left\{\mathbb{I}\left\{r_{\pi(a_n)} = r_{\pi(a_n)}\right\}\phi(x_{\pi(a_n)}^t)\right\} \phi(x_{\pi(a_n)}^T) \]

\[ \hat{m}_n | r_{\pi(a_n)}(a_n) = \hat{\Lambda}_n^{-1} | r_{\pi(a_n)}(a_n) \left(\Lambda_n | r_{\pi(a_n)}(a_n) m_n | r_{\pi(a_n)} + \sum_{t=1}^T \mathbb{E}\left\{\mathbb{I}\left\{r_{\pi(a_n)} = r_{\pi(a_n)}\right\}\right\} x_{\pi(a_n)}^t \phi(x_{\pi(a_n)}^T) \right) \]

\[ \hat{a}_n | r_{\pi(a_n)}(a_n) = a_n | r_{\pi(a_n)} + \frac{1}{2} \sum_{t=1}^T \mathbb{E}\left\{\mathbb{I}\left\{r_{\pi(a_n)} = r_{\pi(a_n)}\right\}\right\} \]

\[ \hat{b}_n | r_{\pi(a_n)}(a_n) = b_n | r_{\pi(a_n)} + \frac{1}{2} \left( m_n | r_{\pi(a_n)}^T \Lambda_n | r_{\pi(a_n)} m_n | r_{\pi(a_n)} - \hat{m}_n | r_{\pi(a_n)}(a_n) \hat{m}_n | r_{\pi(a_n)}(a_n) \right) \]
Algorithm 1 VB-CN: Variational inference for causal networks

Require: $G, x_{1:N}^{T}$
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
\begin{itemize}
    \item $E_Q\{\log \rho_n|\pi(x_n)\} \leftarrow \psi(\hat{a}_n|\pi(x_n)) - \log \hat{b}_n|\pi(x_n)$
    \item $E_Q\{\rho_n|\pi(x_n)\} \leftarrow \frac{\hat{a}_n|\pi(x_n)}{\hat{b}_n|\pi(x_n)}$
    \item $E_Q\{\log \theta_k|\pi(r_k)\} \leftarrow \psi(\hat{\gamma}_k|\pi(r_k)) - \psi(\sum_{r_k'} \hat{\gamma}_k|\pi(r_k') (r_k'))$
\end{itemize}
for $t = 1, \ldots, T$ do
Update log $\hat{\theta}$
Update expected sufficient statistics
\begin{itemize}
    \item $E_Q\left\{\mathbb{1}_{r_k^t=r_k(r_k')} \mathbb{1}_{r_k(r_k')}\right\} \leftarrow \sum_{r_k \cup k} \hat{\theta}^t(r_k)$ where $U = \pi(r_k) \cup \{k\}$
    \item $E_Q\left\{\mathbb{1}_{r_k^t=r_k(r_k')}\right\} \leftarrow \sum_{r_k \cup k} \hat{\theta}^t(r_k)$ where $U = \pi(x_n)$
\end{itemize}
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 $B_P[Q]$ via Equation (8)
until $B_P[Q]$ 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 $B_P[Q]$.

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:

$$B_P[Q] = E_Q \left\{ \log p(r_{1:T}^{1:K}, x_{1:N}^{1:T}, \theta_{1:K}, \rho_{1:N}, w_{1:N}) - \log Q(r_{1:T}^{1:K}, \theta_{1:K}, \rho_{1:N}, w_{1:N}) \right\}$$  \hspace{1cm} (8)

$$= \sum_{t=1}^{T} \sum_{n=1}^{N} E_Q \left\{ \log p(x_n^t | x_{1:n}^{t-1}, r_{1:n}^{t-1}, w_n, \rho_n) \right\}$$  \hspace{1cm} (9)

$$+ \sum_{t=1}^{T} \sum_{k=1}^{K} \left( E_Q \left\{ \log p(r_k^t | r_{1:k}^t, \theta_k) \right\} - E_Q \left\{ \log q(r_{1:k}^t) \right\} \right)$$  \hspace{1cm} (10)

$$+ \sum_{k=1}^{K} \sum_{r_{k}(x_k)} \left( E_Q \left\{ \log p(\theta_k|r_{k}(x_k)) \right\} - E_Q \left\{ \log q(\theta_k|\pi(r_{k})) \right\} \right)$$  \hspace{1cm} (11)

$$+ \sum_{n=1}^{N} \sum_{r_{\pi(x_n)}} \left( E_Q \left\{ \log p(w_n|r_{\pi(x_n)}, \rho_n|r_{\pi(x_n)}) \right\} - E_Q \left\{ \log q(w_n|r_{\pi(x_n)}, \rho_n|r_{\pi(x_n)}) \right\} \right)$$  \hspace{1cm} (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:

\[
E_Q \left\{ \log p(x^t_n \mid x^t_{\pi(x_n)}, r^t_{\pi(x_n)}, w_n, \rho_n) \right\} \\
= \sum_{r^t_{\pi(x_n)}} E \left\{ 1_{\{r^t_{\pi(x_n)} = r^t_{\pi(x_n)}\}} \right\} E \left\{ \log p(x^t_n \mid x^t_{\pi(x_n)}, w_n, r^t_{\pi(x_n)}, \rho_n \mid r^t_{\pi(x_n)}) \right\} \\
= \frac{1}{2} \sum_{r^t_{\pi(x_n)}} E \left\{ 1_{\{r^t_{\pi(x_n)} = r^t_{\pi(x_n)}\}} \right\} \left( E \left\{ \log \rho_n \mid r^t_{\pi(x_n)} \right\} \right) \\
- E \left\{ \rho_n \mid r^t_{\pi(x_n)} \right\} (x^t_n - w_n, r^t_{\pi(x_n)})^T \phi(x^t_{\pi(x_n)})^2 \right) - \log 2\pi \\
= \frac{1}{2} \sum_{r^t_{\pi(x_n)}} \sum_{r^t_{\pi(x_n)}} \hat{\theta}^t(r^t_{\pi(x_n)}, \theta_{\pi(x_n)}) \left( \psi(\hat{\theta}^t_{\pi(x_n)}) - \log \hat{b}^t_{r^t_{\pi(x_n)}} \right) \left( \phi(x^t_{\pi(x_n)})^2 - \phi(x^t_{\pi(x_n)})^T \hat{\Lambda}^{-1}_{n, r^t_{\pi(x_n)}} \phi(x^t_{\pi(x_n)}) \right) - \log 2\pi \\

\]

Variational distribution \( Q \) treats \( r^t_{1:t} \) and \( \theta_{1:t} \) as independent variables. So, the expectations of the categorical log-likelihood terms admit the following form

\[
E_Q \left\{ \log p(r^t_k \mid r^t_{\pi(k)}, \theta_k) \right\} = \sum_{r^t_k} \sum_{r^t_{\pi(k)}} E \left\{ 1_{\{r^t_k = r^t_k\}} 1_{\{r^t_{\pi(k)} = r^t_{\pi(k)}\}} \right\} E \left\{ \log \theta_k \mid r^t_{\pi(k)} \right\} \\
= \sum_{r^t_{1:t}} \hat{\theta}^t(\theta_{1:t}) \left( \psi(\hat{\theta}^t_{r^t_{1:t}}) - \psi(\sum_{r^t_k} \hat{\theta}^t_{r^t_{1:t}}) \right) \\

\]

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 those 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:

\[
E_Q \left\{ \log q(r^t_{1:t}) \right\} = \sum_{r^t_{1:t}} \hat{\theta}^t(\theta_{1:t}) \log \hat{\theta}^t(\theta_{1:t}) \\

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

\[
\mathbb{E}_{\Theta} \left\{ \log p (\theta_k | r_{\pi(r_k)}) \right\} = \log \Gamma \left( \sum \gamma_k | r_{\pi(r_k)} (r_k) \right) - \sum \log \Gamma \left( \gamma_k | r_{\pi(r_k)} (r_k) \right) \\
+ \sum \left( \gamma_k | r_{\pi(r_k)} (r_k) - 1 \right) \left( \psi (\hat{\gamma}_k | r_{\pi(r_k)} (r_k)) - \psi \left( \sum \hat{\gamma}_k | r_{\pi(r_k)} (r_k') \right) \right)
\]

\[
\mathbb{E}_{\Theta} \left\{ \log q (\theta_k | r_{\pi(r_k)}) \right\} = \log \Gamma \left( \sum \hat{\gamma}_k | r_{\pi(r_k)} (r_k) \right) - \sum \log \Gamma \left( \hat{\gamma}_k | r_{\pi(r_k)} (r_k) \right) \\
+ \sum \left( \hat{\gamma}_k | r_{\pi(r_k)} (r_k) - 1 \right) \left( \psi (\hat{\gamma}_k | r_{\pi(r_k)} (r_k)) - \psi \left( \sum \hat{\gamma}_k | r_{\pi(r_k)} (r_k') \right) \right)
\]

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

\[
\mathbb{E}_{\Theta} \left\{ \log p (w_n | r_{\pi(x_n)} ; \rho_n | r_{\pi(x_n)}) \right\} =
\]

\[
a_n | r_{\pi(x_n)} \log b_n | r_{\pi(x_n)} - \log \Gamma (a_n | r_{\pi(x_n)}) + \frac{1}{2} \log \det (\Lambda_n | r_{\pi(x_n)}) - \frac{1}{2} \text{tr} (\hat{\Lambda}_n | r_{\pi(x_n)} \Lambda_n | r_{\pi(x_n)}) \\
- \frac{M}{2} \log 2\pi + \left( a_n | r_{\pi(x_n)} + \frac{M}{2} - 1 \right) \left( \psi (\hat{a}_n | r_{\pi(x_n)}) - \log \hat{b}_n | r_{\pi(x_n)} \right) - \frac{\hat{a}_n | r_{\pi(x_n)}}{\hat{b}_n | r_{\pi(x_n)}}
\]

\[
\mathbb{E} \left\{ \log q (w_n | r_{\pi(x_n)} ; \rho_n | r_{\pi(x_n)}) \right\} =
\]

\[
\hat{a}_n | r_{\pi(x_n)} \log \hat{b}_n | r_{\pi(x_n)} - \log \Gamma (\hat{a}_n | r_{\pi(x_n)}) + \frac{1}{2} \log \det (\hat{\Lambda}_n | r_{\pi(x_n)}) - \frac{M}{2} \log 2\pi + \left( \hat{a}_n | r_{\pi(x_n)} + \frac{M}{2} - 1 \right) \left( \psi (\hat{a}_n | r_{\pi(x_n)}) - \log \hat{b}_n | r_{\pi(x_n)} \right) - \hat{a}_n | r_{\pi(x_n)}
\]

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 \(|R_1|\) of the latent variable was allowed to range between 1 and 5. For the bivariate models in Figure 2, the cardinality \(|R_1|\) of the latent variable \(r_1\) was allowed to range between 1 and 5, in each case the cardinality 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 \(r_1 \in R_1\), \(m_{n|r_1}\)'s were set to 0, and \(\Lambda_{n|r_1}\)'s were set to \(\frac{1}{10} I\) each; while for all values of \(r_1 \in R_1\), \(\gamma_1(r_1)'s\) were set to 10.
We next describe the remaining hyperparameters with respect to the causal graph in Figure 2(b) in which \( x_1 \) causes \( x_2 \). Though 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 \( a_{n\mid r_1} \geq b_{n\mid r_1} \) 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 \( \text{Gamma}(a, b) \) 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.

On the other hand, 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 \( x_1 \) and \( x_2 \). 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. \( x_1^{1:T} = x_2^{1:T} \) and \( x_2^{1:T} = x_1^{1:T} \), we expect the marginal likelihood of the pair \((x_1^{1:T}, x_2^{1:T})\) given the relation \( x_1 \rightarrow x_2 \) to be equal to the marginal likelihood of the permuted pair \((x_2^{1:T}, x_1^{1:T})\) given the relation \( x_2 \rightarrow x_1 \). 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 \( x_1 \rightarrow x_2 \) 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 \( x_2 \rightarrow x_1 \) 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.](image)

For this experiment, for each of 36 hyperparameter combinations, and for each rank values of \(|R_1| = 1 \text{ to } 5\), a total of 3 different data pairs (one for each different graphical
Bayesian Model Selection for Identifying Markov Equivalent Causal Graphs

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 the hyperparameter setting that achieved the highest accuracy in the previous experiment found these four data sets were 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 $|R_1| = 1$, which actually corresponds to the marginal independence of $X$ and $Y$. Such an explanation makes sense considering the plot in Figure 4, since the changes in the initial speed of the ball do not seem to affect the final speed, and therefore they must be independent. In the light of these results, we can also conclude that our model and algorithm are able to detect spurious relationships, even in the absence of previously labelled spurious examples.