Bayesian Causal Discovery Networks for Linear Mixed Data

Moabi Mokhoro MOABI.MOKHORO@RU.NL
Institute for Computing and Information Sciences, Radboud University Nijmegen

Ioan Gabriel Bucur GABRIEL.BUCUR@RU.NL
Institute for Computing and Information Sciences, Radboud University Nijmegen

Tom Heskes TOM.HESKES@RU.NL
Institute for Computing and Information Sciences, Radboud University Nijmegen

Jildau Bouwman JILDAU.BOUWMAN@TNO.NL
Digital Health, TNO, The Netherlands

Tom Claassen TOMC@CS.RU.NL
Institute for Computing and Information Sciences, Radboud University Nijmegen

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Causal discovery from observational data is challenging due to limited sample sizes and noise, motivating probabilistic approaches that represent uncertainty over causal structures and parameters. Bayesian Causal Discovery Networks (BCD Nets) and related methods approximate posterior distributions over causal structures and parameters, but existing approaches primarily focus on continuous data, with limited support for mixed discrete–continuous settings common in healthcare and economics. In this work, we extend BCD Nets to handle linear mixed data. Our approach incorporates an appropriate likelihood function for mixed data into the BCD Nets framework, enabling it to jointly model discrete and continuous variables. Experiments on synthetic and real-world datasets show that our method significantly outperforms a state-of-the-art causal discovery model for mixed data in both structural and causal effects accuracy.

Keywords: Bayesian Causal Discovery, Uncertainty, Variational Inference, Mixed data

1. Introduction

Learning causal structure from purely observational data is often preferable to alternatives such as randomized controlled trials, which can be unethical, costly, or infeasible in many real-world scenarios. This is especially true in domains such as economics and medicine, where practitioners may seek to use causal models to predict the effects of interventions on outcomes of interest. In such settings, observed data is often limited and includes a mix of continuous and discrete variables. However, most existing causal discovery methods, ranging from score-based approaches (Chickering, 2002) to constraint-based algorithms (Spirtes et al., 2000) and to the recent gradient-based techniques (Zheng et al., 2018), are typically designed for datasets composed solely of either continuous or discrete variables, limiting their applicability to mixed data.

Zeng et al. (2022) introduce their approach, LiM, to address the challenge of causal discovery from linear mixed data, avoiding common techniques such as discretizing continuous variables, which can result in information loss. They propose a hybrid score-based learning method consisting of two steps. In the first step, they solve a constrained continuous optimization problem where the objective function is based on the negative log-likelihood of the mixed data. In the second step, they

perform a discrete search over graphs whose skeleton matches the output of the first step, continuing the search until the score can no longer be improved. However, despite showing good results on mixed data, their method does not capture the uncertainty in the learned causal relationships, focusing mainly on point estimation. In fact, a lot of attention in causal discovery research focuses on learning a single point estimate of the underlying causal graph (Spirtes et al. (2000), Chickering (2002), Zheng et al. (2018), Yu et al. (2019)), thereby failing to capture the epistemic uncertainty inherent in finite data and to express the level of confidence in the learned model. A Bayesian approach is appealing for causal discovery, as it enables quantification of uncertainty for downstream decision-making (Heckerman et al. (2006), Friedman and Koller (2003),Viinikka et al. (2020)).

A key challenge in Bayesian causal discovery is the intractability of directly computing the posterior over a super-exponential number of possible causal graphs (Zhou and Chang, 2023). To address this limitation, existing methods rely on approximation strategies, including Markov Chain Monte Carlo (MCMC) as implemented in Gadget (Viinikka et al., 2020), and variational inference as used in BCD Nets (Cundy et al., 2021). The latter approach is scalable to high-dimensional settings, but is limited to continuous data coming from a linear-Gaussian structural equation model.

In this work, we tackle the challenge of capturing the epistemic uncertainty of causal discovery from mixed data by adopting a Bayesian variational inference framework. Specifically, we leverage the mixed-data log-likelihood proposed by Zeng et al. (2022) and adapt the model design of BCD Nets (Cundy et al., 2021) to enable Bayesian causal discovery on mixed data[1]. Our contributions can be summarized as follows:

1. We extend Bayesian Causal Discovery Networks (BCD Nets) to handle linear mixed data.
2. We empirically demonstrate that this approach improves structural accuracy and causal effects accuracy across varying levels of graph sparsity, particularly as the dimensionality increases, compared to the baseline.
3. We show that the model converges to a stationary point at which the skeletons of the sampled graphs from the learned approximate posterior are more consistent with the ground truth, while the edge orientations remain less accurate. To address this issue, we apply the second-stage refinement procedure of Zeng et al. (2022), which iteratively changes edge orientations in the samples using greedy discrete search, thereby improving the samples.

The remainder of the paper is organized as follows. Section 2 provides the necessary background information on the data-generating process, i.e., the structural equation model for linear mixed data, causal discovery in the presence of mixed data and Bayesian causal discovery. Section 3 details our procedure for modifying BCD Nets to handle linear mixed data by incorporating the likelihood for mixed data into the model design. Finally, in Section 4 we evaluate our model's performance on synthetic data and a protein-signaling dataset by comparing it to a baseline model.

2. Background

2.1. Linear Structural Equation Models

A structural equation model (SEM) defines the relationships among a collection of random variables $\{x_1, \ldots, x_d\}$ associated with a directed acyclic graph (DAG) G, having d nodes through a set of

1. Code Repository: https://gitlab.com/moabi_mok/bcd-lim.git

structural assignments of the form $x_j = f_j(\mathbf{Pa}(x_j), \epsilon_j)$, where each f_j is a deterministic function, $\mathbf{Pa}(x_j)$ denotes the set of parent variables of x_j, and ϵ_j is an independent noise variable. The graph G encodes the causal (parental) relationships among the variables.

Cundy et al. (2021) consider the case of continuous data generated by a linear-Gaussian SEM, where each f_j is a linear function and ϵ_j is Gaussian. The model can be written in vector form as $\mathbf{X} = \mathbf{W}^\top \mathbf{X} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal noise covariance matrix. On the other hand, Zeng et al. (2022) consider an SEM for a mixture of continuous and binary variables, where each continuous variable is a linear combination of its parent variables plus an exogenous error term and an intercept term:

$$x_j = e_j + a_j + \sum_{i \in \mathbf{Pa}(j)} w_{ij} x_i \tag{1}$$

here the independent noise variables e_j are non-Gaussian, while w_{ij} and a_j are fixed coefficients and intercepts, respectively. For binary discrete variables, x_j is determined by thresholding a linear function of its parents plus a logistic noise term:

$$x_j = \begin{cases} 1, & \text{if } e_j + a_j + \sum_{i \in \mathbf{Pa}(j)} w_{ij} x_i > 0, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{where } e_j \sim \text{Logistic}(0, 1). \tag{2}$$

2.2. Causal Discovery for Linear Mixed Data

Zeng et al. (2022) use a hybrid score-based learning method to infer the causal structure from linear mixed data. The aim is to minimize a score function that is based on the negative of the following log-likelihood of mixed data:

$$\begin{aligned} \log p(\mathbf{X} \mid \mathbf{W}, \mathbf{a}) = \sum_{t=1}^{n} \sum_{j=1}^{d} z_j \Big\{ x_{j,t} \log[\sigma_{\mathbf{W}}(x_{j,t})] + (1 - x_{j,t}) \log[1 - \sigma_{\mathbf{W}}(x_{j,t})] \Big\} \\ + (1 - z_j) \log p_j \Big(x_{j,t} - \big(a_j + \sum_{k \in \mathbf{Pa}(j)} w_{kj} x_{k,t} \big) \Big) \end{aligned} \tag{3}$$

where $\mathbf{W}$ is the weighted adjacency matrix, $x_{j,t}$ is the t-th sample of variable x_j, and z_j is an indicator variable ($z_j = 1$ if x_j is discrete, and $z_j = 0$ otherwise). The function $\sigma_{\mathbf{W}}$ is the sigmoid activation

$$\sigma_{\mathbf{W}}(x_{j,t}) = \frac{1}{1 + \exp\Big[- \Big(a_j + \sum_{k \in \mathbf{Pa}(j)} w_{kj} x_{k,t} \Big) \Big]},$$

while p_j denotes the density of the non-Gaussian error term e_j. Zeng et al. (2022) use the Laplace distribution, although the framework allows for other choices of non-Gaussian distributions. To estimate the causal structure (DAG), they first minimize the negative log-likelihood augmented with a sparsity term, subject to an acyclicity (DAGness) constraint, $h(\mathbf{W}) = \text{trace}(e^{\mathbf{W} \circ \mathbf{W}}) - d = 0$ (Bhattacharya et al., 2021). This constrained problem is reformulated into an unconstrained problem using the Quadratic Penalty Method (QPM) (Nocedal and Wright, 2006), which is then solved with the L-BFGS-B algorithm (Zhu et al., 1997). Zeng et al. (2022) then refine the graph estimate by

searching over graphs whose skeleton matches that produced by the constrained optimization stage, continuing the search until the score can no longer be improved. This refinement step is included because the continuous optimization often gets stuck in a local optimum, where the recovered skeleton is consistent with the ground truth but the edge orientations are not.

2.3. Bayesian Causal Discovery

Cundy et al. (2021) introduce BCD Nets to approximate the posterior distribution over the parameters of a linear-Gaussian SEM, i.e., to infer $p(\mathbf{W}, \mathbf{\Sigma} \mid \mathbf{X})$, where $\{\mathbf{W}, \mathbf{\Sigma}\}$ are the SEM parameters. In this SEM, the data is generated as $\mathbf{X} = (\mathbf{I} - \mathbf{W})^{-\top}\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Consequently, $\mathbf{X}$ follows a multivariate normal distribution:

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}^{-1}), \quad \text{where } \mathbf{\Theta} = (\mathbf{I} - \mathbf{W})\mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{W})^{\top}. \tag{4}$$

The corresponding log-likelihood for n samples is:

$$\log p(\mathbf{X} \mid \mathbf{\Sigma}, \mathbf{W}) = \frac{n}{2}(\log\det \mathbf{\Theta} - d\log(2\pi)) - \frac{1}{2}\sum_{t=1}^{n} \mathbf{X}_t^{\top}\mathbf{\Theta}\mathbf{X}_t. \tag{5}$$

To make inference tractable, Cundy et al. (2021) first parameterize the weighted adjacency matrix as $\mathbf{W} = \mathbf{P}\mathbf{L}\mathbf{P}^{\top}$, where $\mathbf{L}$ is a strictly lower triangular weight matrix for a canonical DAG with fixed node ordering, and $\mathbf{P}$ is a permutation matrix that modifies the ordering of the nodes. The matrix $\mathbf{L}$ is parameterized by a vector $\mathbf{l} \in \mathbb{R}^{\frac{d(d-1)}{2}}$, and its lower triangular property ensures that $\mathbf{W}$ defines a valid DAG. The inference task then becomes approximating the posterior distribution $p(\mathbf{P}, \mathbf{L}, \mathbf{\Sigma} \mid \mathbf{X})$. Since this posterior is intractable and cannot be computed analytically, Cundy et al. (2021) resort to variational inference (Blei et al., 2017). Specifically, they define a variational distribution $q_\phi(\mathbf{P}, \mathbf{L}, \mathbf{\Sigma})$ which serves as an approximation to the true posterior distribution and optimize its parameters ϕ by maximizing the Evidence Lower Bound (ELBO):

$$\text{ELBO}(\phi) = \mathbb{E}_{(\mathbf{P},\mathbf{L},\mathbf{\Sigma})\sim q_\phi}\left[\log p(\mathbf{X} \mid \mathbf{P}, \mathbf{L}, \mathbf{\Sigma}) - \log\frac{q_\phi(\mathbf{P}, \mathbf{L}, \mathbf{\Sigma})}{p(\mathbf{P}, \mathbf{L}, \mathbf{\Sigma})}\right]. \tag{6}$$

3. Bayesian Causal Discovery Networks for Linear Mixed Data

The goal of this work is to approximate the posterior distribution of a causal structure given mixed data by combining the methodologies of Cundy et al. (2021) and Zeng et al. (2022). In contrast to Cundy et al. (2021), who focus on Gaussian SEMs, we extend the model to handle a mixture of both continuous and discrete binary data by leveraging the mixed-data log-likelihood in (3). This framework naturally accommodates k-way categorical variables, which can be represented as binary indicators through one-hot (1-of-k) encoding.

Formally, let $\mathbf{W} = \mathbf{P}\mathbf{L}\mathbf{P}^{T}$ denote the weighted adjacency matrix of a DAG, parameterized as the product of a permutation matrix $\mathbf{P}$ and a strictly lower-triangular weight matrix $\mathbf{L}$. To handle mixed data, we explicitly incorporate the prior and posterior distributions of the intercepts a_j (as defined in (1) and (2)) into the variational model for ELBO optimization. Unlike Cundy et al. (2021) and Zeng et al. (2022), we find this treatment is necessary because real-world continuous variables rarely have zero mean, and binary variables often exhibit unbalanced class distributions. By explicitly accounting for these baseline offsets, we ensure the model does not erroneously force

the data through the origin. Furthermore, we assume a Laplace$(0, 1)$ distribution for the noise term in (1), as it is known to be robust to misspecification; we demonstrate this empirically in Appendix A. Let $\mathbf{a} = \{a_1, \ldots, a_d\}$ denote the vector of intercepts for our model. Our aim is to learn the posterior

$$p(\mathbf{P}, \mathbf{L}, \mathbf{a} \mid \mathbf{X})$$

via a variational approximation

$$q_\phi(\mathbf{P}, \mathbf{L}, \mathbf{a}).$$

3.1. Posterior and Prior Distributions

A key challenge is selecting an expressive variational family for $q_\phi(\mathbf{P}, \mathbf{L}, \mathbf{a})$ that allows differentiable sampling and density estimation, which are required for stochastic ELBO optimization (Dhaka et al., 2021). Following Cundy et al. (2021), we factorize the approximate posterior as

$$q_\phi(\mathbf{P}, \mathbf{L}, \mathbf{a}) = q_\phi(\mathbf{P} \mid \mathbf{L}, \mathbf{a})\, q_\phi(\mathbf{L}, \mathbf{a}),$$

so that $\mathbf{P}$ is sampled conditionally on $\mathbf{L}$ and $\mathbf{a}$, while $\mathbf{L}$ and $\mathbf{a}$ are sampled jointly.

The corresponding ELBO is

$$\mathbb{E}_{\mathbf{L},\mathbf{a}\sim q_\phi}\left[\mathbb{E}_{\mathbf{P}\sim q_\phi(\cdot\mid\mathbf{L},\mathbf{a})}\left[\log p(\mathbf{X} \mid \mathbf{L}, \mathbf{P}, \mathbf{a}) - \log \frac{q_\phi(\mathbf{P} \mid \mathbf{L}, \mathbf{a})}{p(\mathbf{P} \mid \mathbf{L}, \mathbf{a})}\right] - \log \frac{q_\phi(\mathbf{L}, \mathbf{a})}{p(\mathbf{L}, \mathbf{a})}\right]. \tag{7}$$

For the continuous weights $\mathbf{L}$ and intercepts $\mathbf{a}$, we choose $q_\phi(\mathbf{L}, \mathbf{a})$ as a diagonal-covariance Gaussian, where ϕ parameterizes the mean and variance. This choice is motivated by the expectation that under the identifiability assumptions in Zeng et al. (2022) and Proposition 1, asymptotically, the true posterior distribution, $p(\mathbf{L}, \mathbf{a} \mid \mathbf{X})$, converges to a unimodal multivariate Gaussian. If these assumptions do not hold, more expressive posteriors, such as Normalizing Flows (NFs) (Caterini et al., 2021) can be used.

Proposition 1 *Consider an identifiable structural equation model (SEM) with true parameters $\mathbf{P}_0, \mathbf{L}_0, \mathbf{a}_0$, where $\mathbf{P}_0$ is known, and let $\boldsymbol{\beta}_0 = \{\mathbf{L}_0, \mathbf{a}_0\}$. Under standard regularity conditions, as the sample size $n \to \infty$, the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{X})$ converges to a unimodal distribution.*

Proof By the Bernstein–von Mises theorem (van der Vaart, 1998), for the true parameter $\boldsymbol{\beta}_0$, if the following conditions hold:

- The observed data $\mathbf{X}_{1:n}$ are i.i.d. samples from $p_{\boldsymbol{\beta}_0}(\mathbf{X})$;
- The likelihood function $p(\mathbf{X} \mid \boldsymbol{\beta})$ is smooth and identifiable;
- The prior $p(\boldsymbol{\beta})$ assigns positive mass to $\boldsymbol{\beta}_0$, i.e., $p(\boldsymbol{\beta}_0) > 0$,

then the posterior distribution satisfies

$$\|p(\boldsymbol{\beta} \mid \mathbf{X}_{1:n}) - \mathcal{N}(\hat{\boldsymbol{\beta}},\, n^{-1}\mathcal{I}(\boldsymbol{\beta}_0)^{-1})\|_{TV} \to 0,$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate and $\mathcal{I}(\boldsymbol{\beta}_0)$ is the Fisher information matrix. Since a multivariate normal distribution is unimodal, the posterior converges in total variation to a unimodal (normal) distribution. ■

3.1.1. DISTRIBUTION OVER PERMUTATIONS

The set of d-dimensional permutation matrices $\boldsymbol{\mathcal{P}}_d$ is discrete and grows combinatorially with d, making it challenging to specify a conditional posterior $q_\phi(\mathbf{P} \mid \mathbf{L}, \mathbf{a})$ that permits differentiable exact sampling and density evaluation, which are both required for continuous optimization of the ELBO in (7). Following Cundy et al. (2021), the reference (base) distribution over permutations is defined as the Boltzmann distribution, parameterized by $\mathbf{T} \in \mathbb{R}^{d \times d}$ as follows:

$$P_{\mathbf{T}}(\mathbf{P}) \propto \exp\langle \mathbf{T}, \mathbf{P} \rangle, \quad \mathbf{P} \in \boldsymbol{\mathcal{P}}_d.$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product.

Exact density evaluation from this distribution is challenging because the partition function, $\sum_{\mathbf{P} \in \boldsymbol{\mathcal{P}}_d} P_{\mathbf{T}}(\mathbf{P})$, is intractable especially in high dimensions as it involves summing over an exponential number of terms (Haddadan et al., 2021). Since it is equal to the matrix permanent $\mathrm{perm}(\exp(\mathbf{T}))$, which is #P-hard to compute (Li et al., 2021), Cundy et al. (2021) tractably approximate it using the Bethe permanent (Anari and Rezaei, 2025).

Exact sampling from the Boltzmann distribution is NP-hard (Pochart et al., 2022). To address this challenge, Cundy et al. (2021) employ the Gumbel-Sinkhorn relaxation (Mena et al., 2018) for sampling, which produces a continuous, differentiable approximation of $\mathbf{P}$, enabling gradient-based optimization during training. Specifically, a sample from the Gumbel-Sinkhorn distribution with parameters $\mathbf{T}$ is obtained as follows:

$$\tilde{\mathbf{P}} = \boldsymbol{\mathcal{S}}\big((\mathbf{T} + \gamma)/\tau\big),$$

where $\boldsymbol{\mathcal{S}}$ denotes the Sinkhorn operator, which for a square matrix $\mathbf{T}$ transforms it into a doubly stochastic matrix by first exponentiating it, $\exp(\mathbf{T})$, and then iteratively normalizing its rows and columns for a fixed number of iterations (Mena et al., 2018). Furthermore, $\boldsymbol{\gamma}$ is a matrix of i.i.d. Gumbel noise entries, and $\tau > 0$ is a temperature hyperparameter. As $\tau \to 0$, the relaxed samples converge to permutation matrices $\mathbf{P}$ in $\boldsymbol{\mathcal{P}}_d$. For more details, see Cundy et al. (2021) and Mena et al. (2018). During training, gradients are taken with respect to the relaxed $\tilde{\mathbf{P}}$ using a straight-through estimator (Bengio et al., 2013), while the forward pass uses a hard permutation $\mathbf{P}$ obtained via a transformation of $\tilde{\mathbf{P}}$ using the Hungarian algorithm (Kuhn, 1955).

3.1.2. PRIOR DISTRIBUTIONS

The prior is factorized as

$$p(\mathbf{P}, \mathbf{L}, \mathbf{a}) = p(\mathbf{P})\, p(\mathbf{L})\, p(\mathbf{a}),$$

with a uniform prior over permutations $p(\mathbf{P})$, an uninformative Gaussian prior for $p(\mathbf{a})$, and a horseshoe prior for $p(\mathbf{L})$, where l_i has a horseshoe distribution if it is obtained by first sampling $\lambda_i \sim C^+(0, 1)$ from a half-Cauchy distribution, and then sampling $l_i \sim \mathcal{N}(0, \lambda_i^2 \tau^2)$. Here, τ is a hyperparameter which encodes prior belief about the sparsity of the true data generating DAG.

Algorithm 1 summarizes the full procedure, where $h_\phi(\mathbf{L}, \mathbf{a})$ is a neural network that computes the parameters $\mathbf{T}$ which are used to sample from the Gumbel–Sinkhorn distribution $q_\phi(\mathbf{P} \mid \mathbf{L}, \mathbf{a})$, thereby capturing the conditional dependence of $\mathbf{P}$ on $\mathbf{L}$ and $\mathbf{a}$.

Algorithm 1: Bayesian Causal Discovery for Linear Mixed Data

Input: Data $\mathbf{X}$, step size η, temperature τ
Output: Learned parameters ϕ
Initialize parameterized distribution q_ϕ and neural network $h_\phi(\mathbf{L}, \mathbf{a})$
while *not converged* **do**
  Draw $\mathbf{L}, \mathbf{a} \sim q_\phi(\mathbf{L}, \mathbf{a})$
  Compute logits $\mathbf{T} = h_\phi(\mathbf{L}, \mathbf{a})$
  Draw $\gamma \in \mathbb{R}^{d \times d}$ i.i.d. from standard Gumbel
  Compute soft $\tilde{\mathbf{P}} = \boldsymbol{\mathcal{S}}\big((\mathbf{T} + \gamma)/\tau\big)$
  Compute hard $\mathbf{P} = \text{Hungarian}(\tilde{\mathbf{P}})$
  Compute $g = \nabla_\phi[\text{ELBO}(\phi)]$ using $\mathbf{P}$ (forward pass) and $\tilde{\mathbf{P}}$ (backward pass)
  Update $\phi \leftarrow \phi - \eta g$
end

4. Experiments

In this section, we conduct experiments on both synthetic datasets and a real-world protein-signaling dataset to evaluate the performance of our model relative to a baseline method for Bayesian causal discovery in the presence of mixed data.

4.1. Synthetic Data

In order to conduct experiments on synthetic data, the synthetic data used to train the models was generated by first creating unweighted graphs using the Erdős–Rényi procedure (Erdős et al., 1960), with the average degree set to $\{1, 2, 3\}$. Edge weights were then sampled uniformly from the intervals $[-2, -1] \cup [1, 2]$ to form the weighted adjacency matrix. The non-Gaussian noise terms for the continuous variables (1) were generated from a Laplace$(0, 1)$ distribution.

We considered varying numbers of variables, $d \in \{5, 10, 20\}$, and our main objective was to evaluate how model performance changes as the number of discrete variables varies. For $d = 5$, the number of discrete variables was set to $\{1, 2, 3, 4\}$; for $d = 10$, to $\{2, 5, 8\}$; and for $d = 20$, to $\{4, 10, 16\}$. We randomly selected which variables were discrete in each case, and then generated data according to (1) for continuous variables and (2) for discrete variables.

Following the first stage of continuous optimization, Zeng et al. (2022) apply a second-stage refinement procedure to the weighted adjacency matrix $\hat{\mathbf{W}}$ obtained in that stage, in order to further improve its accuracy with respect to the ground truth. They refine $\hat{\mathbf{W}}$ by solving the following combinatorial optimization problem:

$$\mathbf{W}^* = \arg \min_{\mathbf{W} \in \text{Ske}(\hat{\mathbf{W}}),\, h(\mathbf{W}) < \omega} \mathcal{L}(\mathbf{W}), \tag{8}$$

where $\text{Ske}(\hat{\mathbf{W}})$ denotes the set of DAGs sharing the same skeleton as $\hat{\mathbf{W}}$, $h(\mathbf{W})$ is an acyclicity constraint ensuring that $\mathbf{W}$ is a DAG, and $\omega = 10^{-8}$ is a tolerance parameter.

Based on this, we report results for two variants of our approach, as well as for a baseline, bootstrapped LiM, which generates samples by bootstrapping the point-estimate model of Zeng et al. (2022). This is in contrast to our method that learns an approximate posterior over weighted

adjacency matrices. The first version of both models evaluates performance using samples from the continuous optimization stage: `mBCD Nets`, our model, which uses samples from the learned approximate posterior $q_\phi(\mathbf{W})$, and `LiM`, which uses bootstrapped samples from the first stage of LiM. The second version incorporates the additional combinatorial optimization refinement stage; these variants are denoted as `mBCD Nets+`, where samples from $q_\phi(\mathbf{W})$ are refined using (8), and `LiM+`, where bootstrapped samples are similarly refined using (8).

We evaluate both models using the expected normalized Structural Hamming Distance (SHD) over sampled weighted adjacency matrices compared to the ground truth:

$$\mathbb{E}[\text{SHD}] = \frac{1}{T}\sum_{i=1}^{T} \text{SHD}(G^{(i)}, G_{\text{GT}})$$

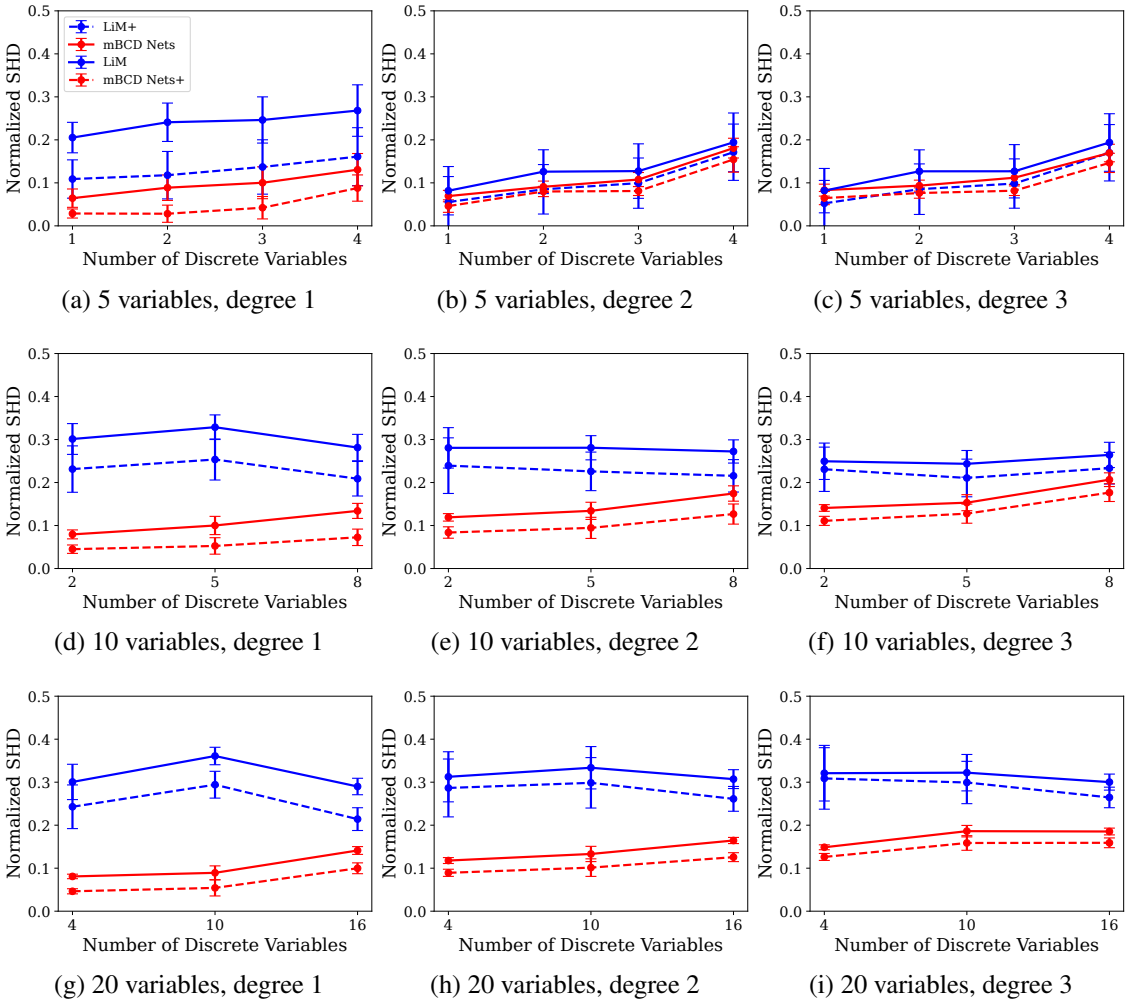
where $G^{(i)} \sim q_\phi(G)$ for our model, and $G^{(i)}$ is drawn from the bootstrapped DAG samples of LiM. Here, G_{GT} denotes the ground-truth DAG.

We provide the results in Figure 1. Overall, our models (`mBCD Nets` and `mBCD Nets+`) obtain lower normalized SHD than the baselines (`LiM` and `LiM+`), particularly for graphs with 10 and 20 variables. For $d = 5$, however, the baselines perform similarly to our models, especially as the graph density increases, i.e., for degree $\in \{2, 3\}$. In all cases, `mBCD Nets` and `LiM` performance improves when the samples are refined using (8), as indicated by the improvement in normalized SHD from the solid to the dashed lines. We also provide additional results in Appendix C, which show that the weighted adjacency matrices sampled from `mBCD Nets` consistently attain higher log-likelihoods than those from `LiM`.

Causal Effect Accuracy: We were also interested in evaluating how well samples from the learned approximate posterior $q_\phi(\mathbf{W})$ perform in terms of causal effects (weights) accuracy. Accuracy is measured by how close the sampled weights are to the ground-truth weights. We compare samples from `mBCD Nets` to those obtained via bootstrapped `LiM`, where the refinement step in Equation 8 is omitted. The ground-truth DAGs and edge weights are shown in Figure 2, with the number of discrete variables set to 2 and 3. Data were generated according to the shown ground truth models. The results are shown in Figure 2 as histograms of sampled edge weights, indicating that in both discrete variable settings, our approach produces samples with means closer to the ground truth (higher accuracy) and with smaller variance compared to the baseline.

We also provide additional experimental results in the Appendix. In Appendix D, we report results for the setting with three discrete variables, using different choices for which variables are discrete. These results show that the position of discrete variables in the graph influences performance. Furthermore, we found that performance can be improved by tuning the hyperparameters batch size, learning rate, and sparsity prior for each configuration, as each setting requires its own optimal hyperparameter values. In Appendix E, we also present results for experiments with varying sample sizes and class imbalance.

Log-Likelihood and SHD during training: In real-world settings, the ground-truth structure is usually unknown, making SHD-based validation infeasible. Since the model maximizes the ELBO during training, which involves maximizing the log-likelihood of the data given samples from the approximate posterior, we investigated whether improvements in log-likelihood correspond to decreases in SHD. This is particularly important for hyperparameter tuning, as we found that batch size, learning rate, and the sparsity prior on the weights all influence performance. Consequently,



(a) 5 variables, degree 1 (b) 5 variables, degree 2 (c) 5 variables, degree 3

(d) 10 variables, degree 1 (e) 10 variables, degree 2 (f) 10 variables, degree 3

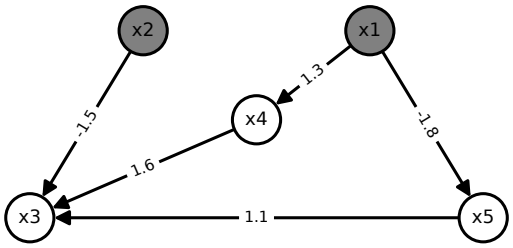
(g) 20 variables, degree 1 (h) 20 variables, degree 2 (i) 20 variables, degree 3

Figure 1: Expected normalized DAG SHD between sampled DAGs and the ground truth DAGs for $d = \{5, 10, 20\}$, varying sparsity levels (degree $= \{1, 2, 3\}$), and different numbers of discrete variables. Results are shown for `mBCD Nets` (red) and `LiM` (blue). Lower normalized SHD, which lies in the range $[0, 1]$, indicates higher accuracy.

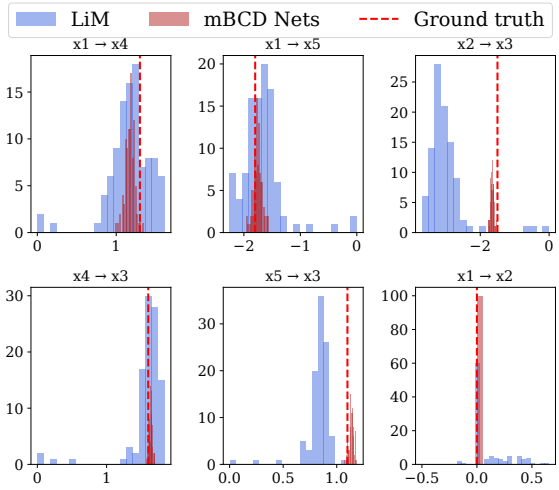
an effective tuning strategy, such as grid search, is required to identify their optimal values. The results, shown in Figure 3, indicate that for both 5- and 10-variable settings, the log-likelihood generally increases while SHD decreases. This suggests that log-likelihood can serve as a useful proxy for SHD when the ground truth is unknown.

4.2. Real World Data

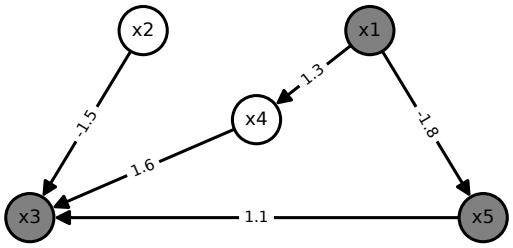
To evaluate our model on real-world data, we compared its performance against the baseline on the Sachs protein-signaling dataset (Sachs et al., 2005) and its widely used consensus network. This dataset contains measurements of 11 continuous protein variables, together with a consensus causal



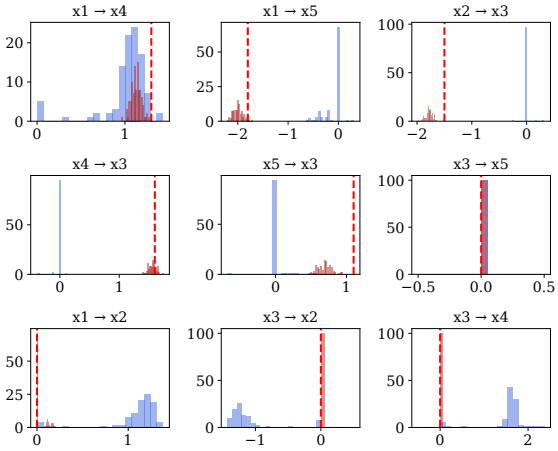
(a) Ground truth: 2 discrete variables.



(b) Histograms of sample weights.



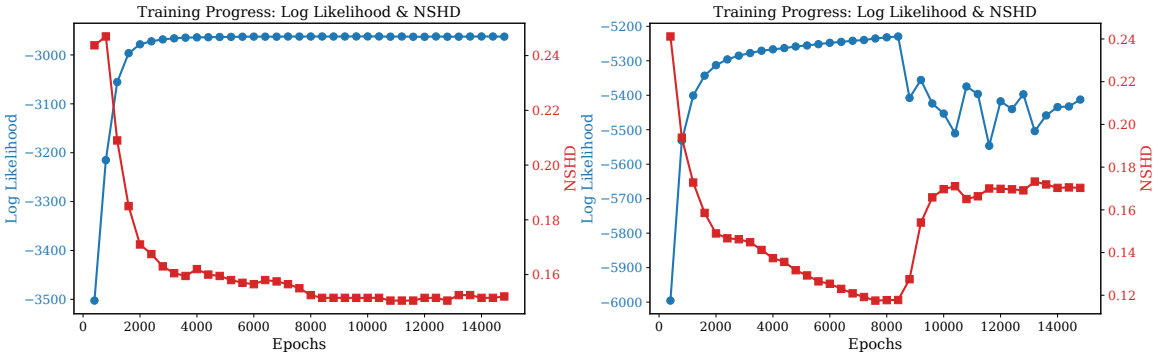
(c) Ground truth: 3 discrete variables.



(d) Histograms of sample weights.

Figure 2: Ground-truth DAGs with edge weights (a) and (c) for 2 and 3 discrete variables (grey nodes), respectively. Histograms of sampled edge weights for our model `mBCD Nets` (pink) and the baseline `LiM` (blue) are shown in (b) for 2 discrete variables and (d) for 3 discrete variables. In both cases, our model's samples have means closer to the ground truth (red dashed line) and smaller variance compared to the baseline.

network describing their causal relationships. Since our focus is on data containing both continuous and binary variables, we assessed model performance under two conditions: (i) when all variables are continuous, and (ii) when a subset of the variables was discretized via median-split binarization. Specifically, we considered configurations with $\{0, 3, 6, 9, 11\}$ binary variables out of the 11 total. The results, summarized in Table 1, report the expected normalized SHD between the consensus network and the graphs sampled from each model. As shown, `LiM+` and `mBCD Nets+` achieve the best performance when all variables remain continuous, whereas `mBCD Nets+` performs best as the number of binary variables increases.



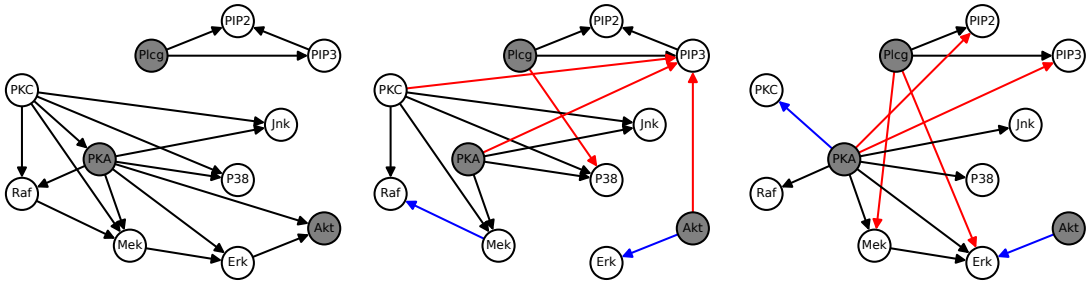
(a) 5 Variables (b) 10 Variables

Figure 3: Log-Likelihood (blue) and normalized SHD (red) during training for (a) 5 variables and (b) 10 variables. Overall, as log-likelihood increases, normalized SHD decreases, and vice versa in plot (b), indicating that improvements and deteriorations in log-likelihood correspond to similar changes in SHD.

	mBCD Nets	**mBCD Nets+**	**LiM**	**LiM+**
Sachs_0	0.17 ± 0.02	$\mathbf{0.16 \pm 0.01}$	0.17 ± 0.01	$\mathbf{0.16 \pm 0.01}$
Sachs_3	0.18 ± 0.02	$\mathbf{0.17 \pm 0.02}$	0.35 ± 0.05	0.34 ± 0.07
Sachs_6	0.17 ± 0.01	$\mathbf{0.15 \pm 0.01}$	0.35 ± 0.06	0.32 ± 0.09
Sachs_9	0.2 ± 0.02	$\mathbf{0.14 \pm 0.02}$	0.34 ± 0.04	0.31 ± 0.06
Sachs_11	0.26 ± 0.02	$\mathbf{0.19 \pm 0.03}$	0.29 ± 0.04	0.23 ± 0.04

Table 1: The expected normalized Structural Hamming Distance (SHD) between the consensus network and sampled graphs from each model on the Sachs dataset. Results are shown for all-continuous variables and for configurations with 3, 6, 9, and 11 binary variables obtained via median-split binarization. Lower SHD indicates better alignment with the consensus network.

We provide estimated graphs constructed by displaying the most frequent consensus graph edges (black), that is, edges that occur in at least 60% of the sampled graphs from both models and that also appear in the consensus graph, where three variables are binarized in Figure 4. Edges that are misoriented in at least 60% of the sampled graphs for each model are shown in blue, and the top four most frequent extra edges that do not appear in the consensus network are shown in red. When comparing `mBCD Nets` and `LiM`, `mBCD Nets` identifies more edges present in the consensus network (10 vs. 8), while both models misorient two edges, including the edge between *Erk* and *Akt*. Among the most frequent non-consensus edges for each model, both methods recover the same extra edge, *PKA* $\rightarrow$ *PIP3*. We note that because the Sachs consensus network contains disputed edges (Brouillard et al. (2025), Mooij et al. (2020)), performance differences may partly reflect agreement with this particular reference rather than true causal validity.



(a) Consensus Network: 3 discrete variables

(b) mBCD Nets estimate

(b) LiM estimate

Figure 4: The consensus network from Sachs et al. (2005) is provided in (a), with three binarized variables shown in grey; sample estimates for `mBCD Nets` in (b) and `LiM` in (c). For each model, we display consensus graph edges that appear in at least 60% of sampled graphs in black. Misoriented edges that appear in at least 60% of sampled graphs are shown in blue, and the four most frequent discovered edges that do not appear in the consensus graph are shown in red.

5. Conclusion

We presented an approach for integrating linear mixed data into BCD Nets, a model originally introduced for posterior approximation in causal discovery under continuous data. Experiments on synthetic and real-world datasets demonstrate that our method outperforms an existing causal discovery model for mixed data in both structural and causal effects accuracy.

In future work, we aim to extend the model to support nonlinear causal relationships, following approaches proposed by Annadani et al. (2023), since linearity cannot be guaranteed in real-world settings. We also plan to incorporate additional discrete variable types, such as ordinal variables. Finally, our model currently assumes causal sufficiency, which can not be guaranteed in real world settings; addressing latent confounding is therefore an important direction for future research (Ashman et al. (2023); Bhattacharya et al. (2021)).

Acknowledgments

This work has been supported by the PersOn project (P21-03), which has received funding from Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). Ioan Gabriel Bucur was financially supported via the Radboud Healthy Data program.

References

Nima Anari and Alireza Rezaei. A tight analysis of Bethe approximation for permanent. *SIAM Journal on Computing*, 54(4):FOCS19–81, 2025.

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. Bayesdag: Gradient-based posterior inference for causal discovery. *Advances in Neural Information Processing Systems*, 36:1738–1763, 2023.

Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal Reasoning in the Presence of Latent Confounders via Neural ADMG Learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2021.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Philippe Brouillard, Chandler Squires, Jonas Wahl, Konrad P. Körding, Karen Sachs, Alexandre Drouin, and Dhanya Sridhar. The landscape of causal discovery data: Grounding causal discovery in real-world applications. In Biwei Huang and Mathias Drton, editors, *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pages 834–873. PMLR, 2025.

Anthony Caterini, Rob Cornish, Dino Sejdinovic, and Arnaud Doucet. Variational inference with continuously-indexed normalizing flows. In *Uncertainty in Artificial Intelligence*, pages 44–53. PMLR, 2021.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Chris Cundy, Aditya Grover, and Stefano Ermon. BCD Nets: Scalable Variational Approaches for Bayesian Causal Discovery . *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael R Andersen, Jonathan Huggins, and Aki Vehtari. Challenges and opportunities in high dimensional variational inference. *Advances in Neural Information Processing Systems*, 34:7787–7798, 2021.

Paul Erdős, Alfréd Rényi, et al. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

Nir Friedman and Daphne Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1):95–125, 2003.

Shahrzad Haddadan, Yue Zhuang, Cyrus Cousins, and Eli Upfal. Fast doubly-adaptive MCMC to estimate the gibbs partition function with weak mixing time bounds. *Advances in Neural Information Processing Systems*, 34:25760–25772, 2021.

David Heckerman, Christopher Meek, and Gregory Cooper. A Bayesian approach to causal discovery. In *Innovations in Machine Learning: Theory and Applications*, pages 1–28. Springer, 2006.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Xuanlin Li, Brandon Trabucco, Dong Huk Park, Michael Luo, Sheng Shen, Trevor Darrell, and Yang Gao. Discovering Non-monotonic Autoregressive Orderings with Variational Inference. In *International Conference on Learning Representations (ICLR)*, 2021.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning Latent Permutations with Gumbel-Sinkhorn Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, 2006.

Thomas Pochart, Paulin Jacquot, and Joseph Mikael. On the challenges of using D-Wave computers to sample Boltzmann Random Variables. In *2022 IEEE 19th International Conference on Software Architecture Companion (ICSA-C)*, pages 137–140. IEEE, 2022.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 1998. Section 10.2 discusses the Bernstein–von Mises Theorem.

Jussi Viinikka, Antti Hyttinen, Johan Pensar, and Mikko Koivisto. Towards scalable Bayesian learning of causal DAGs. *Advances in Neural Information Processing Systems*, 33:6584–6594, 2020.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

Yan Zeng, Shohei Shimizu, Hidetoshi Matsui, and Fuchun Sun. Causal discovery for linear mixed data. In *Conference on Causal Learning and Reasoning*, pages 994–1009. PMLR, 2022.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Quan Zhou and Hyunwoong Chang. Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes. *The Annals of Statistics*, 51(3):1058–1085, 2023.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

Appendix A. Laplace Noise Robustness

In this section, we empirically demonstrate that the Laplace$(0, 1)$ noise distribution assumed for continuous variables in the SEM (1) is robust to misspecification. We evaluate this by generating datasets with 5 variables where the continuous noise follows: (1) Laplace$(0, 1)$, (2) Laplace$(0, 2)$, (3) Gaussian$(0, 1)$, and (4) Gaussian$(0, 2)$. The results, shown in Figure 5, demonstrate that `mBCD-Nets` maintains stable performance across all noise specifications. The lack of performance degradation on other noise distributions indicates that the Laplace$(0, 1)$ assumption is robust to misspecification.

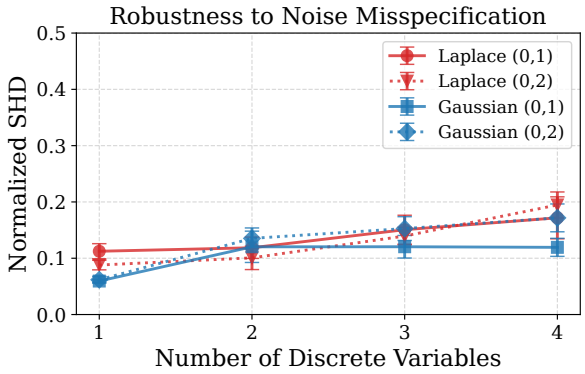


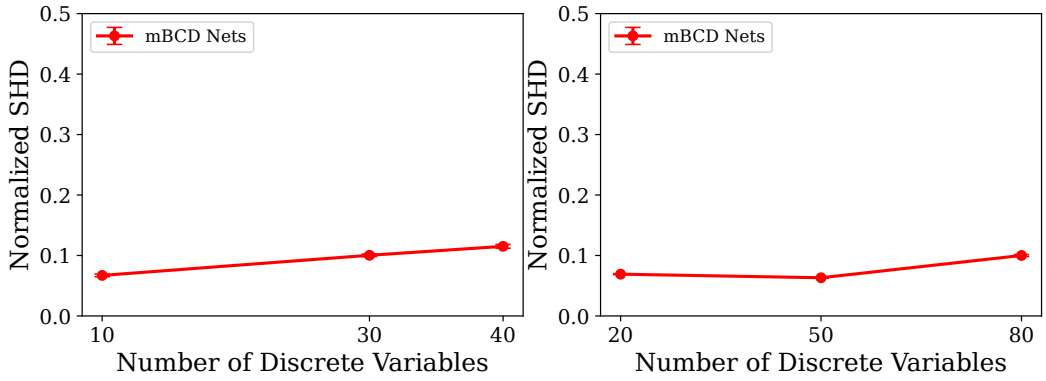
Figure 5: Evaluation of robustness to noise misspecification. Normalized SHD across varying numbers of discrete variables. While our model assumes Laplace$(0, 1)$ noise, it shows stable performance even when the ground-truth data is generated from Gaussian distributions or different Laplace scale, demonstrating the robustness to noise misspecification.

Appendix B. Scalability Analysis in High-Dimensional Settings

We evaluate the scalability of our proposed framework by conducting experiments on high-dimensional synthetic data with $d = 50$ and $d = 100$ variables. As shown in Figure 6, `mBCD Nets` is scalable to high dimensions. Specifically, for the $d = 50$ setting, we evaluated discrete variable counts of $\{10, 30, 40\}$, while for the $d = 100$ setting, we used $\{20, 50, 80\}$.

Appendix C. Likelihood Comparison

As shown in Figure 1, `mBCD Nets` outperforms `LiM` in terms of expected normalized SHD. In Figure 7, we report the negative log-likelihood versus the number of variables $d \in \{5, 10, 20\}$. The results show that (1) the ground-truth data-generating weighted adjacency matrix always achieves the highest log-likelihood, and (2) `mBCD Nets` attains a higher log-likelihood than `LiM`, particularly as d increases, corresponding to a lower expected normalized SHD. Lower negative log-likelihood values indicate better performance, as they correspond to higher data log-likelihood.



(a) 50 Variables. (b) 100 Variables.

Figure 6: Normalized SHD for mBCD Nets on high-dimensional graphs ($d \in \{50, 100\}$).

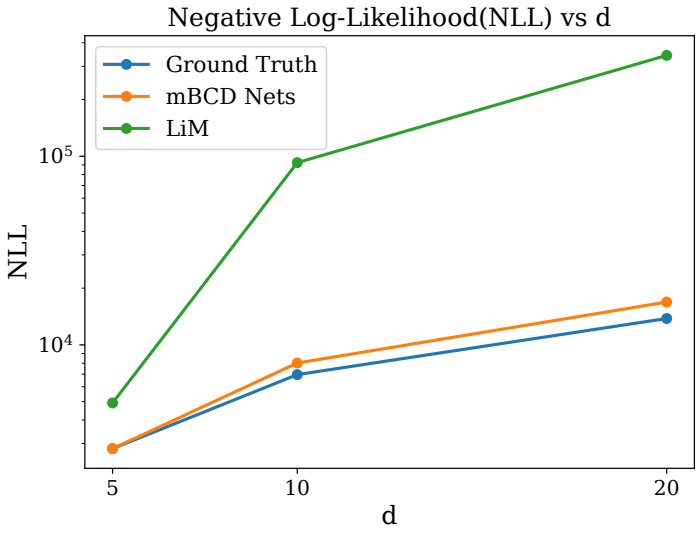
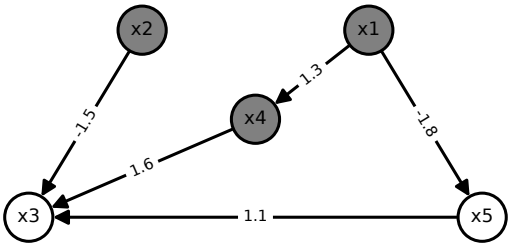


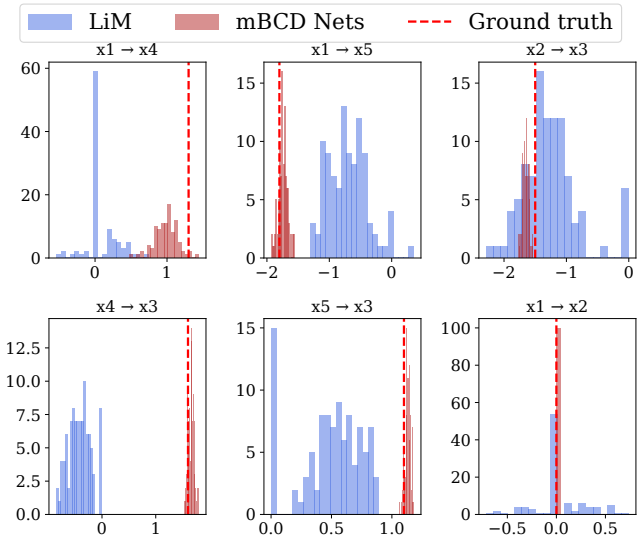
Figure 7: Plot of negative log-likelihood versus the number of variables d, where lower values indicate better performance. In all cases, the ground truth achieves the minimum, followed by `mBCD Nets`, which consistently outperforms `LiM` as d increases, consistent with the results in Figure 1.

Appendix D. Causal Effect Accuracy

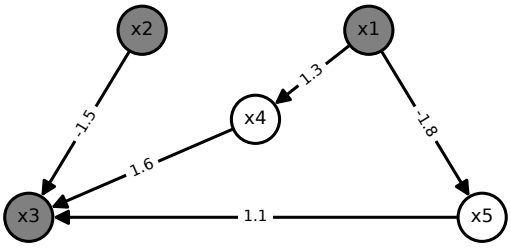
We were also interested in evaluating how well samples from the learned approximate posterior $q_\phi(\mathbf{W})$ perform in terms of causal effect (weights) accuracy. Accuracy is measured by how close the sampled weights are to the ground-truth weights. In this section, we present additional results obtained under different configurations of discrete variables and compare our method to samples derived from bootstrapped LiM. The ground-truth DAGs are shown in Figure 8, where the number of discrete variables is set to 3, with varying choices of which variables are discrete. The corresponding ground-truth edge weights are also displayed in the figure. Data were generated from



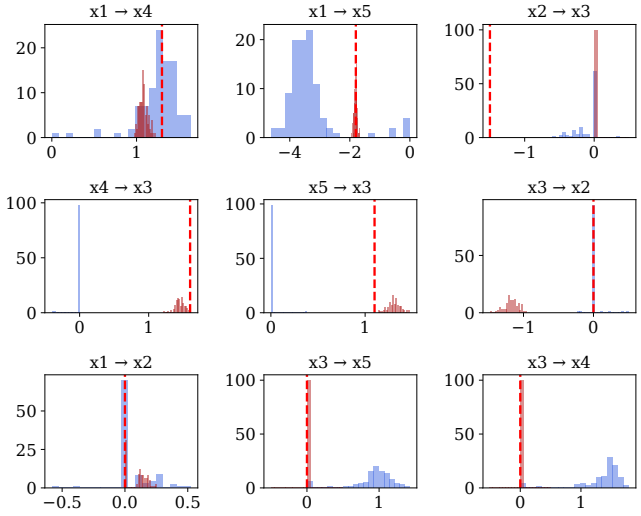
(a) Ground truth: 3 discrete variables.



(b) Histograms of sample weights.



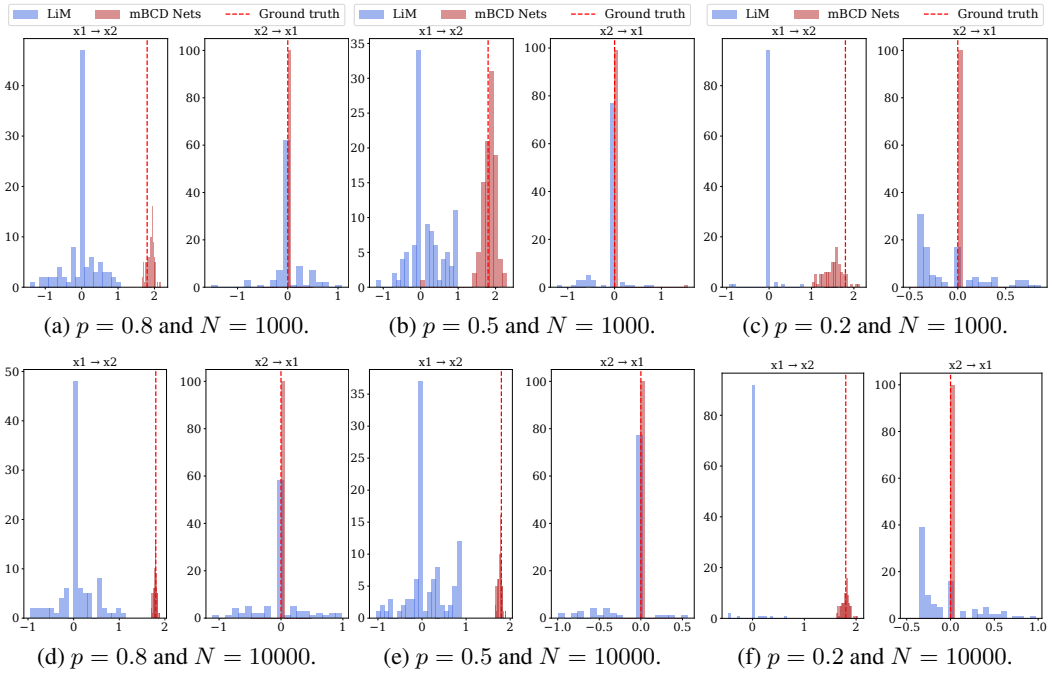
(c) Ground truth: 3 discrete variables.



(d) Histograms of sample weights.

Figure 8: Ground-truth DAGs with edge weights (a) and (c) for 3 discrete variables with different choices of discrete variables. Histograms of sampled edge weights for our model (pink) and the baseline (blue) are shown in (b) and (d) different settings of 3 discrete variables in grey nodes.

this model with a sample size of $N = 1000$. The resulting posterior samples are visualized in Figure 8 as histograms of the sampled edge weights. While `mBCD Nets` overall outperforms the baseline, the results in these figures indicate that performance is sensitive to the choice of discrete variables. In addition, we found that tuning the hyperparameters—batch size, learning rate, and sparsity prior—leads to improved performance for different choices of discrete variables, with no single configuration dominating across all settings.



(a) $p = 0.8$ and $N = 1000$. (b) $p = 0.5$ and $N = 1000$. (c) $p = 0.2$ and $N = 1000$.

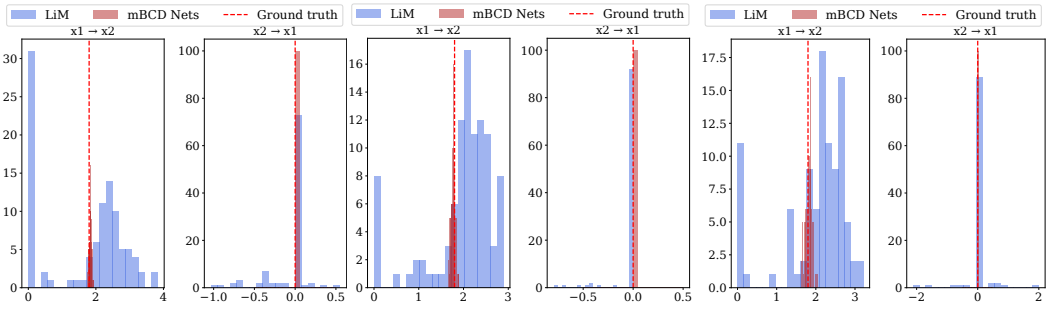
(d) $p = 0.8$ and $N = 10000$. (e) $p = 0.5$ and $N = 10000$. (f) $p = 0.2$ and $N = 10000$.

Figure 9: Weight histograms under varying levels of class imbalance for the parent variable x_1, for the ground-truth graph $x_1 \to x_2$, where both variables are binary and the true weight is 1.8. The parent variable x_1 is generated with $P(x_1 = 1) = p$ and $P(x_1 = 0) = 1 - p$, for imbalance levels $p \in \{0.8, 0.5, 0.2\}$ and sample sizes $N \in \{1000, 10000\}$.

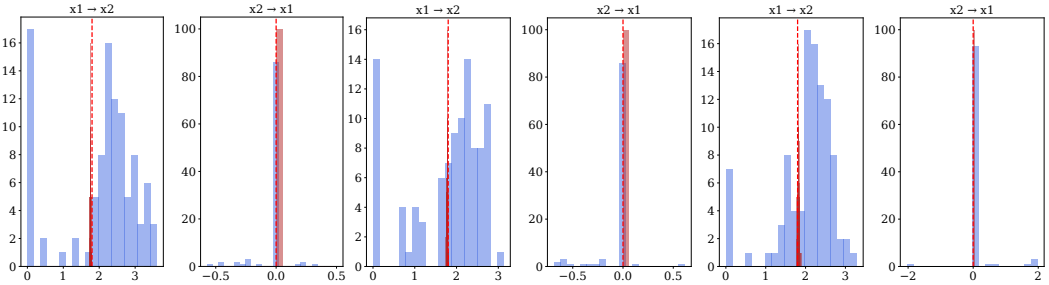
Appendix E. Class Imbalance and Sample Size Effects

We also investigated the impact of varying sample sizes and class imbalance in a bivariate setting, where class imbalance corresponds to different values of p in $P(x1 = 1) = p$ used to generate the data, and $x1$ is the parent variable. The ground-truth structure is $x1 \to x2$ with a true edge weight of 1.8. The sample size was set to $N \in \{1000, 10000\}$. For each value of N, the parent distribution was varied with $p \in \{0.8, 0.5, 0.2\}$.

The results are reported in Figures 9 ($x1$ and $x2$ are binary) and 10 ($x1$ is binary and $x2$ is continuous). Overall, the results indicate that `mBCD Nets` outperforms `LiM` across different sample sizes and levels of class imbalance in terms of causal effect accuracy. Moreover, class imbalance in the parent variable does not appear to substantially affect performance, which remains stable across different values of p. Additionally, Figure 11 reports results for two continuous variables under varying sample sizes N, indicating that `mBCD Nets` outperforms `LiM` across the considered values of N.

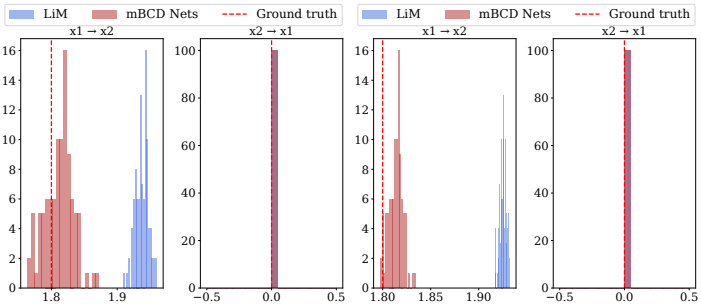


(a) $p = 0.8$ and $N = 1000$. (b)$p = 0.5$ and $N = 1000$. (c)$p = 0.2$ and $N = 1000$.



(a) $p = 0.8$ and $N = 10000$. (b)$p = 0.5$ and $N = 10000$. (c)$p = 0.2$ and $N = 10000$.

Figure 10: Weight histograms under varying levels of class imbalance for the parent variable x_1, for the ground-truth graph $x_1 \rightarrow x_2$, where x_1 is binary and x_2 is continuous. The true weight is 1.8. The parent variable x_1 is generated with $P(x_1 = 1) = p$ and $P(x_1 = 0) = 1 - p$, for imbalance levels $p \in \{0.8, 0.5, 0.2\}$ and sample sizes $N \in \{1000, 10000\}$.



(a)$N = 1000$. (b)$N = 10000$.

Figure 11: Weight histograms for $N = \{1000, 10000\}$. The ground truth graph is $x_1 \rightarrow x2$, for two continuous variables, with ground truth weight of 1.8. The histograms of sampled weights from both models are provided in (b) N = 1000, and (c) N = 10000.