ECAI 2023
K. Gal et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230487

Specifying Prior Beliefs over DAGs in Deep Bayesian Causal Structure Learning

Simon Rittel[a,b;*] **and Sebastian Tschiatschek**[a]

[a]University of Vienna, Faculty of Computer Science, Vienna, Austria
[b]University of Vienna, UniVie Doctoral School Computer Science, Vienna, Austria

Abstract. We consider the principled incorporation of prior knowledge in deep learning based Bayesian approaches to causal structure learning via the prior belief. In particular, we investigate how to include knowledge about individual edges and causal dependencies in the prior over the underlying directed acyclic graph (DAG). While conceptually simple, substantial challenges arise because the acyclicity of a DAG limits the modeling choices of the marginal distributions over its edges. Specifying the marginals iteratively unveils their dependencies and ensures a sound formulation of the probability distribution over DAGs. We provide recipes for formulating valid priors over DAGs for two recent deep learning based Bayesian approaches to causal structure learning and demonstrate empirically that using this prior knowledge can enable significantly more sample-efficient causal structure search.

1 Introduction

Causal structure learning. In the last decades, the field of machine learning has made remarkable advances, in particular with respect to predictive power. The broad usage of modern ML algorithms calls for further research targeting explainability and interpretability as well as fairness and robustness. Causal machine learning promises to play a crucial role in these research directions as it goes beyond purely associative relations between variables and allows for interventional and counterfactual queries [31, 2]. Causal inference typically assumes that the causal model can be represented by *Directed Acyclic Graphs* (DAGs) and *Causal Structure Learning* (CSL) addresses the problem of revealing these DAGs in order to deploy them for causal inference.

In the absence of randomized experiments, it is still possible to infer some causal directions between variables from purely observational data. In a non-parametric setting, (conditional) independencies and dependencies allow to infer the class of graphs that are Markov equivalent under the assumption of faithfulness. This equivalence class can be represented by a *Completely Partially Directed Acyclic Graph* (CPDAG), an acyclic mixed graph in which some edges of the true causal graph remain undirected. However, testing conditional independence is not only a hard statistical problem [32], but its combinatorial nature renders it computationally very demanding, already for moderate numbers of observed variables.

In addition to constraints imposed by independences between variables, some causal directions may be oriented based on the asymmetry of the causal model. Several functional causal models exist that are known to be identifiable [33, 16, 39, 20] and do not require the strong assumption of faithfulness. We refer the reader to [11, 35] for a more comprehensive review of modern CSL.

Uncertainty & Bayesian CSL. Uncertainty quantification is a major challenge for CSL. The outcome of a single statistical hypothesis test tells the probability that the observed data was generated under the null hypothesis, a significance level may then be specified to reject the null hypothesis. To allow for reasonable scaling, independence-based CSL algorithms, e.g. the PC- and FCI algorithm [34], typically restrict their pairwise independence tests repetitively based on the outcome of previous tests. Therefore, a combination of the probabilities attached to the hypothesis tests is non-trivial and impractical even for graphs with a moderate number of edges.

This motivates a probabilistic treatment of the underlying causal graph instead of computing a single point estimate. Analyzing multiple i.i.d. data sets ,generated by the same causal mechanism or obtained by bootstrapping, directly enables a statistical analysis [8] for any CSL algorithm. Alternatively, Bayesian inference allows the incorporation of a prior belief about the graph structure and the computation of the posterior over that structure accounting for made observations.

Recent deep Bayesian CSL algorithms can be distinguished by how they enforce the acyclicity of the graph. The authors of [21, 10, 22] introduce a differentiable DAG constraint [40, 38] in combination with a prefactor into the prior and apply annealing in order to restrict the directed graphs to acyclic ones at the end of the training. By contrast, the authors of [1, 5, 4, 6] constrain the generative model such that no cyclic DAG may be sampled at any stage. For both lines of research, there already exists several extensions that include interventions that are either already present in the data set or appear in the context of an active learning setting. The clear split into the first [37, 13] and second group [30] persists.

Contributions. The effectiveness of Bayesian inference for CSL strongly depends on the considered prior distribution and the therein incorporated prior knowledge. Nevertheless, the role of edge-wise different priors is considerably less studied. In particular, when the amount of data available for CSL is small, the uncertainty about the underlying causal graph may be substantial, and incorporating a prior probability becomes highly influential. Practitioners and domain experts may not exclusively impose hard constraints on the graphical model but additionally may provide probabilistic beliefs about the structure. To the best of our knowledge only global priors, that apply likewise to every single edge or node, are considered in existing work. This does not reflect the modeling reality as the graphs reflect

* Corresponding Author. Email: simon.rittel@univie.ac.at .

the structure of a causal model over distinct entities and are, hence, labeled. In Section 4 of this work, we outline how to define different priors over individual edges in Bayesian structure learning for DAGs and how to incorporate them into the two major approaches regarding enforcing acyclicity. Our major contribution lies in the novel parametrization of a distribution over DAGs, DPM-DAG, that we introduced before in Secion 3. It allows to analytically derive the marginal edge probabilities from the model parameter . Lastly, we demonstrate in our experiments in Section 5 that the with DPM-DAG induced probabilistic priors improve learning in comparison to VI-DP-DAG [4] in terms of sample efficiency when the prior hints information about the true causal graph.

2 Background

In this section, we present the relevant background for our work and review Bayesian structure learning with DiBS [21] and VI-DP-DAG [4].

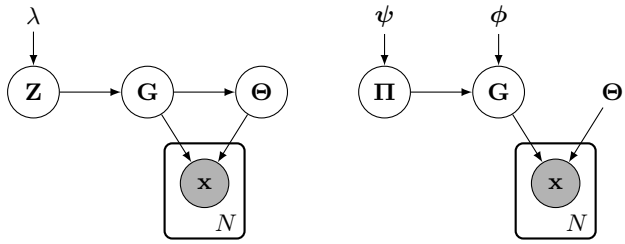
2.1 Structural Causal Model

A *Structural Causal Model* (SCM) or *Functional Causal Model* (FCM) is a triple of a set of endogenous variables $\mathbf{x} := (\mathrm{x}_1, \dots, \mathrm{x}_D)$, a set of exogenous noise variables $\epsilon := (\varepsilon_1, \dots, \varepsilon_D)$ and a set of functions $\boldsymbol{f} := \{f_1, \dots, f_D\}$, one to generate each endogenous variables x_d as a function of $\mathbf{x}$ and ε_d: $x_d = f_d(\boldsymbol{x}_{\sim d}, \varepsilon_d)$, where $\sim d$ denotes the index set $\{1, \dots, D\} \setminus \{d\}$. Typically the structure induced by the direct functional dependencies is restricted to be acyclic and, hence, can be represented by a DAG or equivalently its adjacency matrix $\boldsymbol{G} \in \{0,1\}^{D\times D}$. Then the parents of a node x_d, i.e. the subset of $\mathbf{x}_{\sim d}$ that have a direct influence on x_d over f_d are encoded by the d-th column of $\boldsymbol{G}$. Using a DAG's adjacency matrix $\boldsymbol{G}$ as a mask for $\mathbf{x}$, the generally nonlinear functions $\boldsymbol{f}$ can be approximated by parameterized functions like neural networks. For the remainder of this work, we summarize the parameters of these parameterized functions as $\boldsymbol{\Theta}$. Moreover, we assume causal sufficiency, i.e. all endogenous variables $\mathbf{x}$ are observable and the exogenous noise variables are mutually independent. This states that all dependencies and independencies between the observed random variables $\mathbf{x}$ are only due to their interactions and are not due to unobserved common causes or conditioning on unobserved confounders.

2.2 Enforcing Acyclicity

Acyclicity via permutation sampling. The adjacency matrix $\boldsymbol{G} \in \{0,1\}^{D\times D}$ of any DAG admits a representation by a permutation matrix $\boldsymbol{\Pi} \in \{0,1\}^{D\times D}$ and an upper-triangular matrix $\boldsymbol{U} \in \{0,1\}^{D\times D}$ as shown in Eq. (1). This illustrates that $\boldsymbol{G}$ may have at most $E := \binom{D}{2} = D(D-1)/2$ edges. Consequently, there are 2^E different upper-triangular matrices and $D!$ different permutations. Since $\boldsymbol{U}$ may induce only a partial order between the variables modeled by $\boldsymbol{G}$, the matrix factorization is not unique in general. Yet, the number of labeled DAGs with D variables, n_D, still grows exponentially in D, e.g. for $D = 7$ there exist already 29 281 different labeled DAGs [27], asymptocially $n_D \asymp C_1/C_2^D\, D!\, 2^E$ with $C_1 \approx 1.741$ and $C_2 \approx 1.488$ [36]. The surjective function $g(\boldsymbol{U}, \boldsymbol{\Pi}) = \boldsymbol{\Pi}^T \boldsymbol{U} \boldsymbol{\Pi}$ provides a generative model to sample a random DAG $\boldsymbol{G}$ from distributions over $\boldsymbol{\Pi}$ and $\mathbf{U}$ parameterized by $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ [4]:

$$\boldsymbol{\Pi} \sim p_{\boldsymbol{\psi}}(\boldsymbol{\Pi})\,, \qquad \mathbf{U} \sim p_{\boldsymbol{\phi}}(\boldsymbol{U})\,, \qquad \mathbf{G} = \boldsymbol{\Pi}^T \mathbf{U} \boldsymbol{\Pi}\,. \tag{1}$$



(a) DiBS (b) VI-DP-DAG

Figure 1: Considered graphical models.

Acyclicity via DAG constraint. Alternatively, the acyclicity may be enforced by a non-negative, differentiable function $h(\boldsymbol{G})$ that equals to zero iff $\boldsymbol{G}$ is a DAG [40, 38]. In Bayesian structure learning this constraint is typically incorporated as an exponential factor of an unnormalized Gibbs prior over all graphs $\boldsymbol{G}$ with D nodes, i.e.

$$p_\lambda(\boldsymbol{G}) \propto \exp\left(-\lambda h(\boldsymbol{G})\right)\,. \tag{2}$$

For a sufficiently high prefactor λ, the exponential term in Eq. (2) allocates negligible probability mass to any cyclic graph $\boldsymbol{G}_{\text{cyclic}}$. According to Eq. (2), all DAGs $\boldsymbol{G}_{\text{acyclic}}$ would receive a uniform probability:

$$p_\lambda(\boldsymbol{G}_{\text{cyclic}}) \xrightarrow[\lambda\to\infty]{} 0\,, \quad p_\lambda(\boldsymbol{G}_{\text{acyclic}}) \xrightarrow[\lambda\to\infty]{} \frac{1}{|\mathbb{G}_{\text{acyclic}}|}\,. \tag{3}$$

We denote by $\mathbb{G}_{\mathbb{C}}$ the subset of all directed graphs with D nodes that fulfill the condition in the subscript $\mathbb{C}$, here being acyclic. Note that without any additional factor in Eq. (2), the normalization constant $C_{\mathbb{G}} := \sum_{\boldsymbol{G}'\in\mathbb{G}} \exp(-\lambda h(\boldsymbol{G}'))$ is independent of the respective argument $\boldsymbol{G}$ of the prior, therefore the gradient of the logarithm of the RHS in Eq. (2) can be evaluated in a sampled graph $\mathbf{G}$ without having to compute $C_{\mathbb{G}}$ explicitly.

2.3 Bayesian Structure Learning with SVDG (DiBS)

Generative model & graph posterior. The authors of DiBS [21] translate the discrete structure learning problem into an inference problem over continuous latent variables $\mathbf{Z} = [\mathbf{V}, \mathbf{W}] \in \mathbb{R}^{2\times D\times K}$. They enforce the acyclicity of $\boldsymbol{G}$ by a Gibbs prior over the latent variables $\mathbf{Z}$ via the expectation value of $h(\mathbf{G})$over $\mathbf{G}$ given $\mathbf{Z}$:

$$p_\lambda(\boldsymbol{Z}) \propto \exp\left(-\lambda \mathbb{E}_{\mathbf{G}|\mathbf{Z}}\, h(\mathbf{G})\right)\,. \tag{4}$$

The probability of a directed, loopless graph $\boldsymbol{G}$ given $\boldsymbol{Z}$ is modeled by applying the sigmoid function σ element-wise to the inner product of two latent embedding matrices $\boldsymbol{V}$ and $\boldsymbol{W} \in \mathbb{R}^{D\times K}$:

$$p(\boldsymbol{G}|\boldsymbol{Z}) = \prod_{i=1}^{D}\prod_{j\neq i}^{D} p(G_{ij}|\boldsymbol{v}_i, \boldsymbol{w}_j)$$
$$\text{with} \quad p(G_{ij}|\boldsymbol{v}_i, \boldsymbol{w}_j) = \sigma(\boldsymbol{v}_i \boldsymbol{w}_j^T)\,. \tag{5}$$

Denoting the training data set by $\boldsymbol{X} := \{\mathbf{x}^{(n)}\}_{n=1}^N$, their joint generative model displayed in Fig. 1a factorizes as

$$p(\boldsymbol{Z}, \boldsymbol{G}, \boldsymbol{\Theta}, \boldsymbol{X}) = p_\lambda(\boldsymbol{Z})\, p(\boldsymbol{G}|\boldsymbol{Z})\, p(\boldsymbol{\Theta}|\boldsymbol{G})\, p(\boldsymbol{X}|\boldsymbol{G}, \boldsymbol{\Theta})\,. \tag{6}$$

The posterior over the graph $\boldsymbol{G}$ and the SCM implied by $\boldsymbol{G}$ and $\boldsymbol{\Theta}$ can be obtained by marginalization over $\mathbf{Z}$ as

$$p(\boldsymbol{G}, \boldsymbol{\Theta}|\boldsymbol{X}) \propto p(\boldsymbol{G}) p(\boldsymbol{\Theta}|\boldsymbol{G}) p(\boldsymbol{X}|\boldsymbol{G}, \boldsymbol{\Theta})\,. \tag{7}$$

Differentiable likelihoods. Using the factorization from Eq. (6) analytical expressions for the gradients of the logarithm of the posterior density $p(\boldsymbol{Z}, \boldsymbol{\Theta}|\boldsymbol{X})$ w.r.t. $\mathbf{Z}$ and $\boldsymbol{\Theta}$ can be derived, which can then be evaluated by pathwise gradients using the Gumbel-Softmax trick [24, 17]. In case the data generation mechanism $p(\boldsymbol{\Theta}|\boldsymbol{G})\, p(\boldsymbol{X}|\boldsymbol{G}, \boldsymbol{\Theta})$ is only defined for discrete $\boldsymbol{G}$, discrete Gumbel-max samples can be drawn in the forward pass and their continuous softmax versions can then be used for backpropagation. Alternatively to this straight-through gradient estimator [3], the score function estimator can be applied.

Particle variational inference. To approximate the intractable posterior $p(\boldsymbol{Z}, \boldsymbol{\Theta}|\boldsymbol{X})$ DiBS employs *Stein Variational Gradient Descent* (SVGD) [19] which yields a particle approximation of the joint density in Eq. (6). At each iteration, a set of particles with fixed, preassigned size is mapped to match the target distribution and gets updated using the derived gradients.

Prior formulation. The authors of DiBS [21] choose in their implementation $p(G_{ij}) = 4(D-1)^{-1}$ as a global prior for every single edge. Their prior knowledge about the causal graph states that the edges are independent and every node of the DAG has 2 (incident or outgoing) edges in expectation. This matches the Erdös-Renyi DAGs in their evaluation which they generated by the permutation sampling strategy described in Section 2.4 with the very same probability for each entry of $\boldsymbol{U}$ and a uniformly chosen permutation $\boldsymbol{\Pi}$. While for any DAG this global prior evaluates to the same, intended marginal prior probability for every single edge, it depends on the fact that vertices of the graph are unlabelled and that any permutation is assumed to be equally likely. This constitutes a very generic case and limits the type of prior knowledge of modelers can be incorporated. Since the prior over the complete DAG does not differentiate between different edges, it is better understood as a prior over the number of edges in the DAG.

2.4 DAG Sampling over Permutations (DP-DAG)

Generative model. The DP-DAG model [4] enforces acyclicity by Eq. (1) and also employs a straight-through estimator with Gumbel-softmax samples perturbed by $\log(\sigma(\boldsymbol{\phi}))$ with $\phi_{ij} \in \mathbb{R}$. Note that in the implementation[1] the authors of DP-DAG model a full adjacency matrix $\mathbf{A}$ instead. Omitting to model $\mathbf{U}$ or $\mathbf{A}$ explicitly, the generative model depicted in Fig.1b induces the following factorization of the joint distribution:

$$p_{\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{\Theta}}(\boldsymbol{\Pi}, \boldsymbol{G}, \boldsymbol{X}) = p_{\boldsymbol{\psi}}(\boldsymbol{\Pi})\, p_{\boldsymbol{\phi}}(\boldsymbol{G}|\boldsymbol{\Pi})\, p_{\boldsymbol{\Theta}}(\boldsymbol{X}|\boldsymbol{G}) \,. \tag{8}$$

Variational DP-DAG loss. In contrast to DiBS (see Eq. (6)), VI-DP-DAG [4], a variational method for CSL based on DP-DAG, does not infer a joint probability distribution that also includes the SCM parameters $\boldsymbol{\Theta}$ and is hence not as fully Bayesian as DiBS [21]. Instead it only yields a point estimate for $\boldsymbol{\Theta}$ by maximizing the ELBO $\mathcal{L}$:

$$\max_{\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{\Theta}} \mathcal{L} = \max_{\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{\Theta}} \underbrace{\mathbb{E}_{\mathbf{G}\sim p_{\boldsymbol{\psi},\boldsymbol{\phi}}(\boldsymbol{G})}\left[\log p_{\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{G})\right]}_{(i)} - \underbrace{\beta\, D_{\mathrm{KL}}\left(p_{\boldsymbol{\phi}}(\boldsymbol{A}) \,||\, p_{\boldsymbol{\gamma}}(\boldsymbol{A})\right)}_{(ii)} \,. \tag{9}$$

[1] https://github.com/sharpenb/Differentiable-DAG-Sampling

This loss $\mathcal{L}$ consists of a reconstruction term (i) and the Kullback-Leibler divergence D_{KL} (ii) with prefactor β and prior $p_{\boldsymbol{\gamma}}(\boldsymbol{A})$ acting as a regularization on the unmasked adjacency matrix $\boldsymbol{A}$. For $0 \leq \beta$ it constitutes a lower bound to the joint probability $p(\boldsymbol{x}, \boldsymbol{G})$. In contrast to *Variational Auto Encoders* (VAEs) it does not employ an amortization network for $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. We emphasize that the calculation of the posterior from Eq. (7) depends on $p(\boldsymbol{X}, \boldsymbol{G})$ instead. Assuming i.i.d. samples $\boldsymbol{x}^{(n)} \in \boldsymbol{X}$, $\prod_n^N p(\boldsymbol{x}^{(n)}|\boldsymbol{G})$ provides an unbiased estimate of $p(\boldsymbol{X}|\boldsymbol{G})$, therefore $1/N$ should be incorporated into β to recover the dependence of the posterior on the size of the data set $\boldsymbol{X}$.

Prior formulation. The authors of VI-DP-DAG [4] apply the same marginal prior on all edges, $p(A_{ij}) =: \gamma$, and assume their independence. The resulting KL term (ii) then becomes $\sum_{i,j} \mathrm{KL}\left(p_{\phi}(A_{ij})||\gamma\right)$. Note that during training they evaluate this term for all edges, even for edges masked by the sampled permutation matrix Π. They abstain from defining a prior over the permutation, although for hard Gumbel-softmax there exists a closed form for the KL-divergence as we show in Section 4.2.

3 DAG Sampling by Masking (DPM-DAG)

In the following we introduce DPM-DAG as an alternative to DP-DAG and derive an analytical expression for the marginal edge probabilities analytical which facilitates the specification of edge-wise priors in Section 4.

Generative model. Under different permutations, the entries of the upper triangular matrix $\boldsymbol{U}$ in Eq. (1) correspond to different edges. Learning the entries U_{ij} directly can become unstable, because they can represent different functional causal dependencies and, thus, receive alternating gradient updates. This would imply a mean-field approximation of $p(\boldsymbol{G})$ by $p(\boldsymbol{U})p(\boldsymbol{\Pi})$ that severely limits the expressiveness of the distribution [5]. Following the implementation of DP-DAG, we model a full adjacency matrix $\mathbf{A}$ and mask it by element-wise multiplication with a permuted upper-triangular adjacency matrix of a fully connected DAG $\boldsymbol{M}$[2]:

$$\boldsymbol{\Pi} \sim p_{\boldsymbol{\psi}}(\boldsymbol{\Pi}) \,, \quad \mathbf{A} \sim p_{\boldsymbol{\phi}}(\boldsymbol{A}) \,, \quad \mathbf{G} = \mathbf{A} \circ \underbrace{\left(\boldsymbol{\Pi}^T \boldsymbol{M}\, \boldsymbol{\Pi}\right)}_{:=\mathbf{M}^{(\boldsymbol{\Pi})}} \,. \tag{10}$$

We refer to this model as *differentiable probabilistically masked DAG* (DPM-DAG), since the key difference to Eq. (1) consists in implicitly modeling a distribution over the acyclicity mask $\mathbf{M}^{(\boldsymbol{\Pi})}$ by probabilistically modeling the permutation $\boldsymbol{\Pi}$. In combination with the distribution over $\boldsymbol{A}$ that has independent entries, it yields a more expressive distribution over DAGs than DP-DAG. Omitting to model $\mathbf{A}$ explicitly, the generative model is then still equivalent to the one described by Eq. (8) and Fig.1b. For details on the influence of $\mathbf{A}$ in $p_{\boldsymbol{\phi}}(\boldsymbol{G}|\boldsymbol{\Pi})$ we refer to our full paper version.

Variational DPM-DAG loss. Additionally formulating a prior over the permutation $p_{\boldsymbol{\omega}}(\boldsymbol{\Pi})$ for the generative model in Eq. (8) yields a different regularization term than (ii) in Eq. (9):

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\psi},\boldsymbol{\phi}}(\boldsymbol{\Pi}, \boldsymbol{G}) \,||\, p_{\boldsymbol{\psi},\boldsymbol{\omega}}(\boldsymbol{\Pi}, \boldsymbol{G})\right) = = \mathbb{E}_{\boldsymbol{\Pi},\mathbf{G}\sim p_{\boldsymbol{\psi},\boldsymbol{\phi}}}\left[\log\left(\frac{p_{\boldsymbol{\psi}}(\boldsymbol{\Pi})}{p_{\boldsymbol{\omega}}(\boldsymbol{\Pi})}\frac{p_{\boldsymbol{\phi}}(\mathbf{G}|\boldsymbol{\Pi})}{p_{\boldsymbol{\gamma}}(\mathbf{G}|\boldsymbol{\Pi})}\right)\right] \,. \tag{11}$$

[2] The random upper-triangular matrix $\mathbf{U}$ in Eq. (1) is then defined by $\left(\boldsymbol{\Pi}\mathbf{A}\boldsymbol{\Pi}^T\right) \circ \boldsymbol{M}$.

Differentiable sampling. In order to generate a discrete random permutation matrix $\mathbf{\Pi}$ that is path-wise differentiable, we perturb a feature vector of log-probabilities $\boldsymbol{\psi} \in \mathbb{R}^D$ by a random vector $\mathbf{g} \in \mathbb{R}^D$ of i.i.d. Gumbel random variables g_i. As proposed in [4] we then apply the (one-hot) *argsort* operator row-wise in the forward pass, but the SoftSort operator [29], its continuous approximation, for the gradient computations in the backward pass. This again yields a straight-through estimator.

Taking a single *argmax* of perturbed Gumbel samples is equal in probability to drawing discrete samples from a categorical distribution with weights $\{w_i := \exp \psi_i\}_{i=1}^D$, this is known as the Gumbel trick [12, 25]. Applying the *argsort* operator row-wise to the perturbed Gumbel samples $\psi + \mathrm{g}$ is equivalent to repetitively taking the *argmax* of the sample vector where we ignore already selected rows. Hence, *argsort* of perturbed Gumbel samples is equal in probability to sampling from a categorical distribution over the index set $\mathbb{I} := \{1, \dots, D\}$ without replacement until each category was selected once. Both approaches sample a permutation π over the index set $\mathbb{I}$ and can be described by the *Plackett-Luce* (PL) distribution [28, 23]. Denoting the set of not yet selected indices by $\mathbb{S} \subset \mathbb{I}$, the probability of selecting index $i \in \mathbb{S}$ as the next index in the sequence that defines the permutation π by its total order is coupled to the Gumbel samples $\mathbf{g}$ by

$$\arg\max_{i \in \mathbb{S}} (\psi_i + g_i) \sim p\left(\frac{\exp(\psi_i)}{\sum_{j \in \mathbb{S}} \exp(\psi_j)}\right) . \tag{12}$$

Marginal edge probability. The marginal probability that the edge G_{ij} is masked by the randomly permuted matrix $\mathbf{M}^{(\mathbf{\Pi})}$ as defined in Eq. (10) equals the probability of selecting index j before i in the Plackett-Luce distribution, $m_{ij} = p(i \prec j)$,[3] The DPM-DAG model allows to analytically read off the marginal edge probabilities from the model parameters w_j, w_j and $a_{ij} := p_{\boldsymbol{\phi}}(A_{ij} = 1) = \sigma(\phi_{ij})$.

Proposition 1 (Marginal edge probabilities) *The marginal probability of sampling edge G_{ij} in the DPM model is given by the product of m_{ij} and $a_{ij} := p(A_{ij} = 1)$.*

Proof 1 *The proof for Proposition 1 directly follows from Eq (10).*

Proposition 2 (Marginal probability of pairwise preference) *The marginal preference probability of the Plackett-Luce distribution $p_{\mathbb{I}}^{(PL)}(i \prec j)$, i.e. the probability that i precedes j in a sampled permutation π sampled from the Plackett-Luce distribution over a set of integers $\mathbb{I} \subseteq \{i, j\}$, evaluates to $\frac{w_i}{w_i + w_j}$.*

Proof 2 *Due to page limitation we provide the proof in the full version of our paper.*

4 Incorporating Prior Knowledge

The authors of DiBS and VI-DP-DAG, as well as their extensions, consider a unitary prior marginal probability for all edges, while the considered Bayesian frameworks would allow for more flexible edge-wise priors as we demonstrate in the following. This is important as a single marginal prior for all edges does not reflect realistic practical scenarios for several reasons. Firstly, conjugate distributions are very convenient, but special cases. Having observed some data $\boldsymbol{X}$, our initial prior belief $p(\boldsymbol{G})$ gets updated to a posterior $p(\boldsymbol{G}|\boldsymbol{X})$ that typically deviates from the initial prior. Secondly, the same marginal prior over all edges results in a graph prior term that does not distinguish the contribution of the labeled edges but rather depends on the number of edges in a graph. By contrast, the modeled variables are not unlabelled, anonymous entities. Domain experts typically possess different prior knowledge of the marginal distributions over different edges. The acyclicity of the graph renders these marginals dependent. In this section, we address how to specify valid priors over individual edges in models with acyclicity via a Gibbs prior like DiBS and with our model DPM-DAG that is based on DP-DAG.

4.1 Gibbs Prior

Recall from Eq. (3) that for a sufficiently high λ, the Gibbs prior with the exponential acyclicity constraints allocates a negligible probability to cyclic graphs and equal probability to all acyclic ones. Therefore, we can limit our subsequent analysis to DAGs. To enable priors for different edge combinations, our basic idea is to divide the class of all DAGs by binary splits based on the existence of specific edges and allocate a prior probability to the resulting groups of graphs. In the following, we explain how to incorporate prior probabilities starting with a prior over a single edge (Example 1). In the full version of our paper, we then extend our approach to two edges, multiple combinations of edges and outline how to specify marginals.

Example 1 (Prior over single edge) *The prior knowledge of the marginal probability p_{ij} for the existence of a specific edge G_{ij} can be expressed by including the following term to the exponential acyclicity factor in Eq. (2):*

$$p(\boldsymbol{G}) \propto \left(r_{ij}\, G_{ij} + r_{\overline{ij}}\,(1 - G_{ij})\right) ,$$
$$\text{with} \quad r_{ij} := \frac{p_{ij}}{|\mathbb{G}_{ij}|} \quad \text{and} \quad r_{\overline{ij}} := \frac{1 - p_{ij}}{|\mathbb{G}_{\overline{ij}}|} . \tag{13}$$

Intuitively, this splits the set of DAGs, $\mathbb{G}_{\text{acyclic}}$, in two groups. One group contains all DAGs where the edge G_{ij} is present, i.e. $\mathbb{G}_{ij} = \{\boldsymbol{G} \in \mathbb{G}_{\text{acyclic}} \,|\, G_{ij} = 1\}$, the other consists of all DAGs that do not contain a link between those two nodes at all or the edge in its reverse direction, i.e. $\mathbb{G}_{\overline{ij}} = \{\boldsymbol{G} \in \mathbb{G}_{\text{acyclic}} \,|\, G_{ij} = 0\}$. The probabilities to sample a DAG from $\mathbb{G}_{ij}$ and $\mathbb{G}_{\overline{ij}}$ are then p_{ij} and $1 - p_{ij}$ respectively. Each distinct graph within a group is assigned a uniform portion of the group's probability, e.g. for $\mathbb{G}_{ij}$:

$$p(\boldsymbol{G} \in \mathbb{G}_{ij}) = \frac{r_{ij}}{\sum\limits_{\mathbb{G}_{\text{acyclic}}} r_{ij}\, G_{ij} + r_{\overline{ij}}\,(1 - G_{ij})} \tag{14}$$

$$= \frac{r_{ij}}{p_{ij} + (1 - p_{ij})} = \frac{p_{ij}}{|\mathbb{G}_{ij}|} , \tag{15}$$

$$\sum_{\boldsymbol{G} \in \mathbb{G}_{ij}} p(\boldsymbol{G}) = p_{ij} . \tag{16}$$

Directly adding multiple factors does not work, as the edges in a DAG are not independent. In order to specify the probability for an edge combination, the modeler has to assign a probability for the very edge combination and the remaining probability to all deviations from it. To the best of our knowledge, there is no known formula to calculate the cardinally of the sets $\mathbb{G}_{\mathbb{C}}$ without labeling all DAGs for a predefined number of nodes.

[3] $m_{ij} := \sum_{\mathbf{\Pi} \in \mathcal{P}_D(\boldsymbol{G})} p(\mathbf{\Pi})\, p(M_{ij}^{(\mathbf{\Pi})} = 1)$.

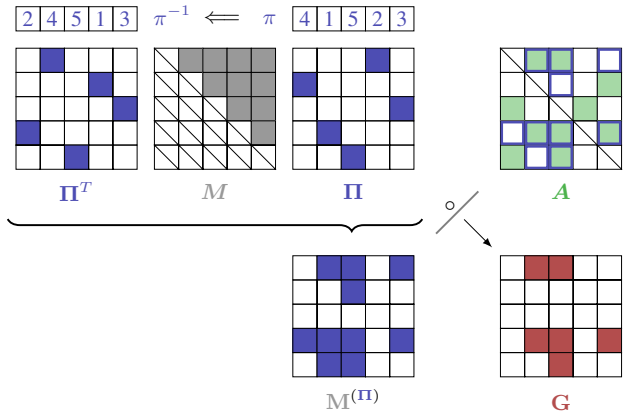


Figure 2: Sampling procedure of DPM-DAG. (1) A random permutation vector π is drawn by Eq. (10) using Gumbel-SoftSort samples and then mapped to its matrix representation $\mathbf{\Pi}$; (2) a strictly upper-triangular matrix $\boldsymbol{M}$ with all-ones above the diagonal is permuted by multiplication with $\mathbf{\Pi}^T$ and $\mathbf{\Pi}$; (3) the resulting random mask $\mathbf{M}^{(\mathbf{\Pi})}$ then constrains the random matrix $\mathbf{A}$ (that is also drawn using Gumbel-Softmax samples) to the adjacency matrix of a DAG $\mathbf{G}$.

4.2 DAG Sampling by Permutation

4.2.1 Prior over permutation

Using the DPM-DAG model, a prior probability over the permutation π can be formulated over the positive weights $w_i := \exp(\psi_i)$ by fixing their ratios. The Kullback-Leibler divergence in Eq. (11) provides then an incentive for the model distribution $p_{\boldsymbol{\psi}}(\mathbf{\Pi})$ to match the specified prior distribution $p_{\boldsymbol{\omega}}(\mathbf{\Pi})$.

Example 2 (Prior over precedence of an index over another index) *Recall from Proposition 2 that the probability of selecting index i before j under the* Plackett-Luce *(PL) distribution, m_{ij}, equals to $\frac{w_i}{w_i+w_j}$ and sets the following constraint $w_i = \frac{m_{ij}}{1-m_{ij}} w_j$. Note that this does not yet affect the probability of edges between any other nodes, since their weights and their probabilities of the entries of $\mathbf{A}$ are not coupled. The probability for the edge in the reverse direction is upper bounded by $m_{ji} = 1 - m_{ij}$, but due to the independence of A_{ji} and A_{ij} only lower bounded by 0.*

Formulating a probability for an additional pairwise order that contains at least one variable whose order is already constrained directly implies probabilities of precedence of newly linked indices by transitivity.

Example 3 (Adding a prior over precedences) *Given two specified prior probabilities over precedences, m_{ij} and m_{kl}, neither m_{ik} nor m_{jl} are coupled. A prior probability over a total order is initially only specified independently for the two subsets (i, j) and (k, l). Adding the prior probability $m_{jk} := p(j \prec k)$ establishes a prior probability over the strict total order on (i, j, k, l) and sets a prior probability for all relations among them. In total, there are four weights, w_i, w_j, w_k and w_l, but before specifying m_{jk} there are only two pairwise constraints. Taking into account the normalization constraint, one degree of freedom remains. This example also highlights the dependencies of the marginal probabilities. The two marginals p_{ij} and p_{kl} can be chosen independently since they are modeled by different parameters, namely w_i, w_j, a_{ij} and w_k, w_l, a_{kl}. After stating the third precedence prior, $m_{jk} = \frac{w_j}{w_j+w_k}$, the probabilities for all pairwise precedence between the variables in $\{i, j, k, l\}$ are defined and the marginal prior probabilities of p_{ik}, p_{il}, p_{jk} and p_{jl} are constrained by the probability over the total order.*

Following the authors of [17] we apply a categorical prior $p_{\boldsymbol{\omega}}(\mathbf{\Pi})$, where we denote by $\boldsymbol{\omega}$ the categorical prior probabilities to distinguish them from the (normalized) weights $\tilde{\boldsymbol{w}}$ derived from the learnable log-probabilities $\boldsymbol{\psi}$:

$$D_{\mathrm{KL}}\left(p_{\boldsymbol{\psi}}(\mathbf{\Pi}) \,||\, p_{\boldsymbol{\omega}}(\mathbf{\Pi})\right) = \sum_i \tilde{\mathrm{w}}_i \left(\log \tilde{\mathrm{w}}_i - \log \omega_i\right) . \tag{17}$$

While this does not constitute a valid lower bound on the ELBO for Gumbel-softmax samples [24] and continuous relaxations of the permutation matrix [26], it works well in practice [17] and is a reasonable approximation since we use hard Gumbel samples in the forward pass.

4.2.2 Prior over marginal edge distributions

For the DPM-DAG model, Proposition 1 states that the marginal prior probabilities p_{ij} are modeled as the product of the precedence probability m_{ij} and the conditional probability a_{ij}. Therefore marginal priors are coupled over the probability on the total order (see Example 3) and modelers can iteratively specify them without formulating invalid marginal probabilities that are not sound for a DAG. Note that the choice of a_{ij} does not affect the probability of edges between any other nodes, since the conditional probabilities are not shared among different edges, but tied to a specific one, i.e. here G_ij. Denoting the parameters of the Bernoulli prior over the edges by $\boldsymbol{\gamma}$ the respective regularization terms from Eq. (11) is

$$\begin{aligned} &D_{\mathrm{KL}}\left(p_{\boldsymbol{\phi}}(\boldsymbol{G}|\mathbf{\Pi}) \,||\, p_{\boldsymbol{\gamma}}(\boldsymbol{G}|\mathbf{\Pi})\right) = \\ &\quad = \sum_{\substack{i \prec j \\ \mathbf{\Pi}}} a_{ij} \frac{\log a_{ij}}{\log \gamma_{ij}} + (1 - a_{ij}) \frac{\log(1 - a_{ij})}{\log(1 - \gamma_{ij})} . \end{aligned} \tag{18}$$

5 Experiments

In this section, we demonstrate empirically that the proposed priors indeed achieve the desired behavior and can, in the case of encoding knowledge about the correct graph, improve sample efficiency of learning. We focus on the permutation sampling strategy outlined in Section 4.2 and adapt the authors' implementation[4] of VI-DP-DAG [4] accordingly. We retain the authors' training procedure and

[4] https://github.com/sharpenb/Differentiable-DAG-Sampling

model architecture unless stated otherwise. Since our aim is not to advance the overall performance but to evaluate our proposed priors, we do not perform any hyperparameter tuning and use the default parameters.

5.1 Experimental Setup

Data set. For our evaluation, we use the synthetic data set provided by the authors of VI-DP-DAG [4]. Each data subset for a random graph model with some specified characteristics contains 50 randomly sampled DAGs. For each randomly sampled DAG with the stated characteristics, 1000 data samples were generated according to a different SCM with independent zero-mean Gaussian noise variables ϵ. The functional dependencies $\boldsymbol{f}$ in the SCM were generated from a nonlinear Gaussian process with RBF kernels with bandwidth 1. We focus on graphs with 10 nodes generated by the Erdös-Rényi model with 10 edges per graph in expectation and use the first 20 of the 50 provided subsets. The samples from each subset are split into train/validation/test sets according to a 80%/10%/10% ratio, normalized and randomly shuffled afterward.

Models. We use the same model architecture to approximate the SCMs as proposed in VI-DP-DAG [4]. In particular, for each function of the SCM, we use a 3-layere MLP with 16 hidden units in each layer and ReLU activation functions except for the output layer which does not have an activation function. The input dimension equals to the number of variables in the data set. The adjacency matrix of a DAG sampled from our probabilistic model determines which inputs are masked (set to zero) and which ones can influence the generation of other variables.

Model training. Training is performed w.r.t. the DPM-DAG loss that replaces the regularization term (ii) of the DP-DAG loss in Eq. (9) by Eq. (11). We employ two Adam optimizers [18]: one with a learning rate of 10^{-3} for the parameters of the MLPs ($\boldsymbol{\theta}$) and one with a learning rate of 10^{-2} for the parameters of the DAG (ψ and ϕ). We train for a maximum number of 500 epochs and apply early stopping with a patience of 10 based on the improvement of the validation loss of our adjusted variational loss evaluated after every second epoch While in the training of VI-DP-DAG a permutation is sampled approximately from the PL distribution with learnable permutation weights $\boldsymbol{w}$, for early stopping on the validation set and for the final evaluation of the test set, the *Maximum Posterior Probability* (MAP) estimate of the permutation is evaluated instead to ensure comparability to [4]. The mode of the PL distribution is the permutation that results from deterministically sorting the weights in decreasing order. We neither apply any pre-processing, such as preliminary neighbor search, nor post-processing pruning steps, nor any additional sparsity regularization.

Evaluation metrics. For the evaluation of our approach, we report the *Area Under the Receiver Operator Characteristic* (AUROC) and the *Area Under the Curve of Precision-Recall* (AUCPR). Higher scores for both are preferred. In contrast to the *Structural Hamming Distance* (SHD), both metrics are independent of the choice of a threshold parameter and therefore are more suitable for the Bayesian setting. To evaluate the predictive performance, the AUCPR should be preferred over the AUROC, since for sparse graphs the prediction of only very few edges leads is beneficial w.r.t. the true and false positive rate of the AUROC [13].

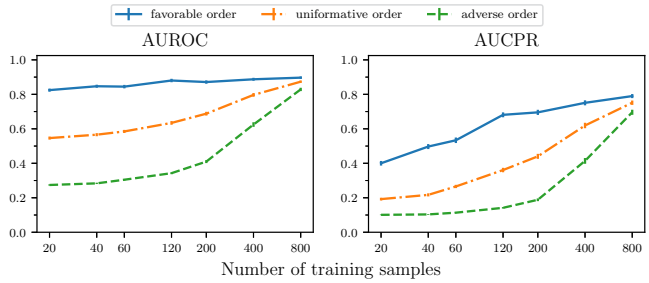


Figure 3: Influence of the prior over the order of the nodes $p_{\boldsymbol{\omega}}(\boldsymbol{\Pi})$ on AUROC and AUCPR, both evaluated using the MAP over the order. The prior over the conditional edges is fixed to $p_{\boldsymbol{\gamma}}(A_{ij}) = 0.5\ \forall i \neq j$. Means and standard errors are calculated over 20 random Erdös-Rényi graphs with 10 nodes and 10 edges in expectation.

5.2 Experimental Results

In the Bayesian setting, the influence of the prior on the posterior typically diminishes with the number of samples that are used for the Bayesian updates. To validate our parametrization qualitatively, we train our parameterized DPM-DAG model for different sample sizes $N \in \{20, 40, 60, 120, 200, 400, 800\}$.

Prior over the permutation. In the first experiment, we demonstrate the effect of the prior on the order of the variables. We compare three different settings. In the *favorable* setting, we compute for each DAG in the training set a total order of its nodes and assign to the weights in the categorical prior distribution over the permutation the values $[100, 81, 64, 49, 36, 25, 16, 9, 4, 1]$ accordingly. Recall from Ex. 2 that under this prior, the probability of selecting the first node before the sixth equals $\frac{4}{5}$. In the *adverse* setting, we reverse the total order calculated for the favorable setting and assign the same weights, e.g. the probability of selecting the first node before the sixth then equals $\frac{36}{37}$. In the *uninformative* setting, we assign uniform weights to each node, such that every order is equally likely under the prior. For all three settings, we employ an uninformative prior over the conditional edges, i.e. $\forall i \neq j : p_{\boldsymbol{\gamma}}(A_{ij} = 1) = 0.5$. We emphasize that DAGs may only have a partial order and the computed order may not be unique. This is even more likely for sparse DAGs but does not weaken our qualitative analysis. Our results depicted in Fig. 3 demonstrate empirically that for all three settings, the curves for the AUROC and AUCPR start to converge towards high values with increasing sample size. The consistency of the different priors is guaranteed by the scaling of the reconstruction loss (ii) for a single sample with the sample size N. For up to 400 samples our experiment shows a strong effect of the prior on the performance of both metrics and motivates the specification of a prior over the order of variables. This finding is in line with the theoretical insight that a provided order of the nodes reduces the search space drastically.

Prior over the conditional edge probabilities. We investigate the influence of the prior on the conditional edges by applying the *informative* prior on the permutation and test three different settings. In the *favorable* setting, we set the conditional prior probability of an edge $p(A_{ij})$ to 0.7, if the edge appears in the true graph, and to 0.3 otherwise. In the *adverse* setting, we reverse this procedure and provide misguiding information in our prior over the conditional edges. For the intermediate regime, we provide the *uninformative* setting, in

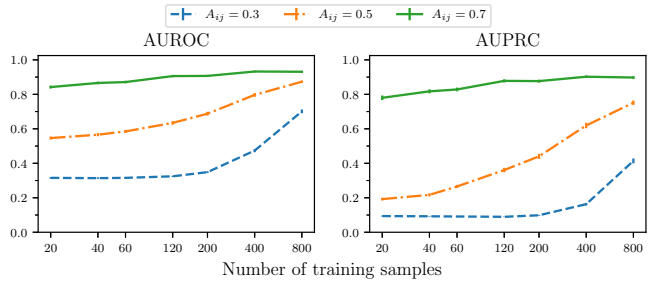


Figure 4: Influence of the prior over the conditional edges $p_{\boldsymbol{\gamma}}(\boldsymbol{A})$ on AUROC and AUCPR evaluated using the MAP over the order. The prior over the order of the nodes is fixed to $p_{\boldsymbol{\omega}}(i \prec j) = 0.5 \; \forall i \neq j$. Means and standard errors are calculated over 20 random Erdös-Rényi graphs with 10 nodes and 10 edges in expectation.

which we assign a prior conditional probability of 0.5 to all nodes. For even up to 800 samples in the training set our experimental results in Fig. 4 exhibit again a strong dependence on the prior. While the onset of convergence cannot be observed, consistency is again still guaranteed to hold.

Comparison with DP-DAG as baseline. Lastly, we compare in Fig. 5 three settings where some correct information about the true underlying graph is specified in the DAG prior to DP-DAG with $\beta = 0.01$ and a global edge prior $p_{\boldsymbol{\gamma}}(A_{ij}) = 0.01$, but otherwise identical parameters. The setting in which favorable information about the order of the nodes as well as the conditional edges is included in the prior performs best in terms of AUROC as well as AUCPR. Note that these best curves both slightly decrease with increasing number of training samples N, since the reconstruction term in our loss grows in relation to the KL-regularization term with N and dominates the training loss on the right side of the plot. The uncertainty arising from the finite training data set then explains the deviation from the higher value before. The second setting consists of the *favorable* prior over the conditional edge, $p_{\boldsymbol{\gamma}}(A_{ij}) = 0.7$ paired with an *uninformative* prior over the permutation, $p_{\boldsymbol{\omega}}(\boldsymbol{\Pi})$. The third setting pairs the *favorable* prior over the permutation with an *uninformative* one over the conditional edges. We observe that all three settings of DPM-DAG clearly outperform DP-DAG w.r.t. sample-efficiency, although we do not induce any additional sparsity regularization and DP-DAG optimizes the hyperparameter β and $p_{\boldsymbol{\gamma}}(A_{ij})$. we can empirically observe that DP-DAG still acts as assigning a global prior to every single edge since our DPM-DAG model with an uninformative prior over the order and over the conditional edges performs very similar.

6 Related Work

Priors for Bayesian CSL. To the best of our knowledge, the literature on priors for Bayesian CSL focuses on modular priors that are decomposable [7]. It can be divided into priors over the number of parents [9, 7], priors penalizing deviations from a specific edge pattern [15, 10], or priors over the expected number of edges stemming from random graph models [14, 21]. We consider the original formulation of VI-DP-DAG [4] rather as a CSL algorithm with a sparsity constraint than a Bayesian one since the authors do not provide any further probabilistic reasoning about the underlying true graph and treat the prior as a single hyperparamter that they tune during training.

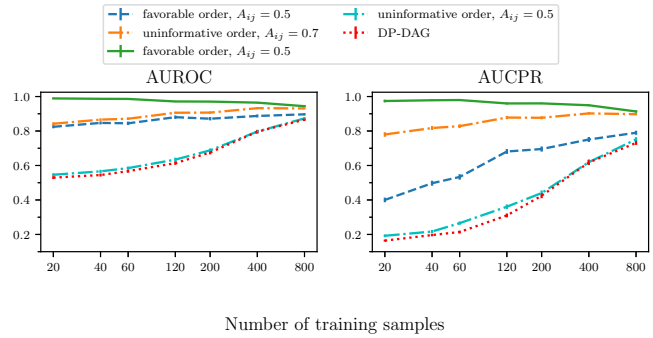


Figure 5: Comparison of three *favorable* settings in which the prior comprises information about the underlying true graph and one *uninformative* setting with DP-DAG as the baseline. A *favorable* prior either refers to a *favorable* order of the variables, a conditional edge probability $p_{\boldsymbol{\gamma}}(A_{ij}) = 0.7$ for any true edge, or both. The *uninformative* setting consists of an *uninformative* order and $p_{\boldsymbol{\gamma}}(A_{ij}) = 0.5$ Means and standard errors are calculated over 20 random Erdös-Rényi graphs with 10 nodes and 10 edges in expectation.

Different parametrization. Alternatively to the generative model in Eq. (10), the authors of BCD-Nets [5] propose to initially generate unlabelled data $\tilde{\mathbf{X}}$ according to an upper-triangular matrix $\mathbf{U}$ with random binary entries and map it to the observed data samples $\mathbf{X}$ by multiplication with a permutation matrix $\boldsymbol{\Pi}$ that is learned using $\mathbf{U}$ and $\tilde{\mathbf{X}}$. While it seems more natural to model only a triangular matrix $\mathbf{U}$ instead of a full adjacency matrix $\mathbf{A}$ and hence fewer SCM parameters $\boldsymbol{\Theta}$, it constrains specifying a prior distribution over the DAG $\boldsymbol{G}$. This becomes evident since the $E = D(D-1)/2$ elements of $p(\boldsymbol{U})$ are now unlabelled and each of them contributes to the probability of a single edge, $p(G_{ij})$, under a different permutation. By contrast, our parametrization allows us to specify marginal distributions over edges in a more transparent and concise manner.

7 Conclusion.

In this paper, we outlined how to incorporate probabilistic beliefs about individual edges in Bayesian structure learning for DAGs and how to incorporate them into the two currently dominating models for deep-learning-based Bayesian CSL, DiBS [21] and VI-DP-DAG [4]. Our adaption of the Gibbs prior formulation demonstrates that is possible to incorporate edge-wise priors. Nevertheless, this requires knowledge of the cardinality of the group of DAGs that feature a specific edge pattern which might be prohibitive in practical applications. Therefore, we advocate our model, DPM-DAG, i.e. modeling DAGs probabilistically by sampling a permutation π first and using it to mask a full adjacency matrix $\boldsymbol{A}$ to express edge-specific priors in Bayesian structure learning. In our experiments, we empirically verified that our proposed probabilistic priors speed up learning in comparison to VI-DP-DAG [4] in terms of sample efficiency when the probabilistic belief contains correct information about the underlying graph. In future work, we aim to investigate the expressivity and scalability of DPM-DAG and to apply it in interactive settings in combinations with interventions.

Acknowledgements

This work was partially funded by the Federal Ministry of Education, Science and Research (BMBWF) of Austria within the interdisciplinary project "Digitize! Computational Social Science in the Digital and Social Transformation". We would like to thank the anonymous reviewers for their thoughtful comments.

References

[1] Yashas Annadani, Jonas Rothfuss, Alexandre Lacoste, Nino Scherrer, Anirudh Goyal, Yoshua Bengio, and Stefan Bauer, 'Variational causal networks: Approximate bayesian inference over causal structures', in *SIGKDD Conference on Knowledge Discovery and Data Mining*, (2021).

[2] Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard, 'On Pearl's hierarchy and the foundations of causal inference', in *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, 507–556, Association for Computing Machinery, New York, NY, USA, 1 edn., (2022).

[3] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville, 'Estimating or propagating gradients through stochastic neurons for conditional computation', *arXiv:1308.3432*, (2013).

[4] Bertrand Charpentier, Simon Kibler, and Stephan Günnemann, 'Differentiable DAG sampling', in *International Conference on Learning Representations*, (2022).

[5] 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).

[6] Tristan Deleu, António Góis, Chris Chinenye Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio, 'Bayesian structure learning with generative flow networks', in *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, (2022).

[7] Ralf Eggeling, Jussi Viinikka, Aleksis Vuoksenmaa, and Mikko Koivisto, 'On structure priors for learning bayesian networks', in *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1687–1695, (2019).

[8] Nir Friedman and Daphne Koller, 'Being bayesian about network structure', in *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, pp. 201–210, (2000).

[9] Nir Friedman and Daphne Koller, 'Being bayesian about network structure. A bayesian approach to structure discovery in bayesian networks', *Machine Learning*, **50**(1-2), 95–125, (2003).

[10] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang, 'Deep end-to-end causal inference', in *NeurIPS Workshop on Causality for Real-world Impact*, (2022).

[11] Clark Glymour, Kun Zhang, and Peter Spirtes, 'Review of causal discovery methods based on graphical models', *Frontiers in Genetics*, **10**, 524, (2019).

[12] Emil Julius Gumbel, 'Statistical theory of extreme values and some practical applications', *NBS Applied Mathematics Series*, **33**, (1954).

[13] Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause, 'BacaDI: Bayesian causal discovery with unknown interventions', in *UAI Workshop on Causal Representation Learning*, (2022).

[14] David Heckerman, 'A bayesian approach to learning causal networks', in *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pp. 285–295, (1995).

[15] David Heckerman, Dan Geiger, and David Maxwell Chickering, 'Learning bayesian networks: The combination of knowledge and statistical data', in *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 293–301, (1994).

[16] Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf, 'Nonlinear causal discovery with additive noise models', in *Advances in Neural Information Processing Systems*, pp. 689–696, (2008).

[17] Eric Jang, Shixiang Gu, and Ben Poole, 'Categorical reparameterization with gumbel-softmax', in *International Conference on Learning Representations*, (2017).

[18] Diederik P. Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', in *International Conference on Learning Representations*, (2015).

[19] Qiang Liu and Dilin Wang, 'Stein variational gradient descent: A general purpose bayesian inference algorithm', in *Advances in Neural Information Processing Systems*, pp. 2370–2378, (2016).

[20] Po-Ling Loh and Peter Bühlmann, 'High-dimensional learning of linear causal networks via inverse covariance estimation', *Journal of Machine Learning Research*, **15**(1), 3065–3105, (2014).

[21] Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause, 'Dibs: Differentiable bayesian structure learning', in *Advances in Neural Information Processing Systems*, volume 34, pp. 24111–24123, (2021).

[22] Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf, 'Amortized inference for causal structure learning', in *Advances in Neural Information Processing Systems*, (2022).

[23] R Duncan Luce, *Individual choice behavior: A theoretical analysis*, Courier Corporation, 2005.

[24] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh, 'The concrete distribution: A continuous relaxation of discrete random variables', in *International Conference on Learning Representations*, (2017).

[25] Chris J. Maddison, Daniel Tarlow, and Tom Minka, 'A* sampling', in *Advances in Neural Information Processing Systems*, pp. 3086–3094, (2014).

[26] Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek, 'Learning latent permutations with gumbel-sinkhorn networks', in *International Conference on Learning Representations*, (2018).

[27] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2023. Published electronically at http://oeis.org/A003024.

[28] RL Plackett, 'The analysis of permutations', *Journal of the Royal Statistical Society Series C: Applied Statistics*, **24**(2), 193–202, (1975).

[29] Sebastian Prillo and Julian Eisenschlos, 'Softsort: A continuous relaxation for the argsort operator', in *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7793–7802, (2020).

[30] Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke, 'Learning neural causal models with active interventions', in *NeurIPS Workshop on Causal Inference & Machine Learning: Why now?*, (2021).

[31] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio, 'Towards causal representation learning', *arXiv:2102.11107*, (2021).

[32] Rajen D Shah and Jonas Peters, 'The hardness of conditional independence testing and the generalised covariance measure', *The Annals of Statistics*, **48**(3), 1514–1538, (2020).

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

[34] Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, Prediction, and Search, Second Edition*, Adaptive Computation and Machine learning, MIT Press, 2000.

[35] Chandler Squires and Caroline Uhler, 'Causal structure learning: a combinatorial perspective', *Foundations of Computational Mathematics*, 1–35, (2022).

[36] Richard P. Stanley, 'Acyclic orientations of graphs', *Discret. Math.*, **5**(2), 171–178, (1973).

[37] Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius Von Kügelgen, 'Active bayesian causal inference', in *Advances in Neural Information Processing Systems*, (2022).

[38] Yue Yu, Jie Chen, Tian Gao, and Mo Yu, 'DAG-GNN: DAG structure learning with graph neural networks', in *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7154–7163, (2019).

[39] Kun Zhang and Aapo Hyvärinen, 'On the identifiability of the post-nonlinear causal model', in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 647–655. AUAI Press, (2009).

[40] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing, 'Dags with NO TEARS: continuous optimization for structure learning', in *Advances in Neural Information Processing Systems*, pp. 9492–9503, (2018).