On Structure Priors for Learning Bayesian Networks

Eggeling, Ralf

2019

Eggeling , R , Viinikka , J , Vuoksenmaa , A I & Koivisto , M 2019 , On Structure Priors for Learning Bayesian Networks . in K Chaudhuri & M Sugiyama (eds) , The 22nd International Conference on Artificial Intelligence and Statistics, 16-18 April 2019 . Proceedings of Machine Learning Research , vol. 89 , Journal of Machine Learning Research , Cambridge, MA , The 22nd International Conference on Artificial Intelligence and Statistics , Naha, Okinawa , Japan , 16/04/2019 . < http://proceedings.mlr.press/v89/eggeling19a/eggeling19a.pdf >

http://hdl.handle.net/10138/310103

publishedVersion

On Structure Priors for Learning Bayesian Networks

Ralf Eggeling[1,2] **Jussi Viinikka**[1] **Aleksis Vuoksenmaa**[1] **Mikko Koivisto**[1]
[1]Department of Computer Science
University of Helsinki, Finland
[2]Department of Computer Science
University of Tübingen, Germany

Abstract

To learn a Bayesian network structure from data, one popular approach is to maximize a decomposable likelihood-based score. While various scores have been proposed, they usually assume a uniform prior, or "penalty," over the possible directed acyclic graphs (DAGs); relatively little attention has been paid to alternative priors. We investigate empirically several structure priors in combination with different scores, using benchmark data sets and data sets generated from benchmark networks. Our results suggest that, in practice, priors that strongly favor sparsity perform significantly better than the uniform prior or even the informed variant that is conditioned on the correct number of parents for each node. For an analytic comparison of different priors, we generalize a known recurrence equation for the number of DAGs to accommodate modular weightings of DAGs, a result that is also of independent interest.

1 INTRODUCTION

Learning a graphical model consists of learning the model structure and inferring the model parameters (Koller and Friedman 2009). It is popular to formulate structure learning as a *model selection* problem, which is solved by some model selection *criterion*. Common criteria take a form of a scoring function that associates every admissible structure a score; one selects a structure that maximizes the score.

To learn the structure of a Bayesian network (BN), namely a directed acyclic graph (DAG), various scoring functions have been derived under different statistical paradigms. For instance, taking a Bayesian approach, every DAG G is assigned a prior probability $P(G)$, the relationship to data set D is captured by the *marginal likelihood* $P(D \mid G)$, and the score is the posterior probability $P(G \mid D)$ (or its monotone transformation) obtained by the Bayes rule. The marginal likelihood is an average over a parameter prior, different choices of which result in different scoring functions, such as the Bayesian Dirichlet (BD) family for categorical data (Heckerman et al. 1995). The Bayesian Information Criterion (BIC) (Schwarz 1978), in turn, is a large-sample approximation of the marginal likelihood; the BIC score also has an interpretation in the (two-part) minimum description length (MDL) framework (Rissanen 1978). Likewise, the more recent fNML and qNML scores (Silander et al. 2010; 2018), which approximate the so-called normalized maximum likelihood (Barron et al. 1998), can be viewed as substitutes to the marginal likelihood.

In contrast to the large variety in instantiating, approximating, or replacing the marginal likelihood $P(D \mid G)$, the structure prior $P(G)$ has received less attention in the literature. While some alternative forms of structure priors have been proposed (Heckerman et al. 1995, Friedman and Koller 2003, Angelopoulus and Cussens 2008), there has been no systematic empirical comparison and none of the priors has reached universal acclaim. Indeed, the commonly employed scoring functions simply ignore the structure prior; in Bayes scores this is equivalent to assigning a uniform prior. A potential concern with a uniform prior is that it assigns a large probability mass on complex (i.e., dense) structures, for they vastly outnumber simpler structures. While the effect of the prior vanishes as the size of the data grows (Koller and Friedman 2009; p. 804), one might benefit from preferring simpler models when there are little data.

This paper investigates the role of structure priors in score-based structure learning of BNs. Unlike some previous work, we do not address the issues of eliciting and expressing a modeller's, possibly sophisticated, background knowledge (Angelopoulus and Cussens

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

2008). Instead, we study what kind of a structure prior is a robust choice in a setting where no actual prior knowledge is available, beyond maybe some vague idea about the sparsity of the "true DAG." We will measure the accuracy of structure learning using the popular *structural Hamming distance (SHD)* (Tsamardinos et al. 2006). The recent Intersection–Validation method (Viinikka et al. 2018) allows us to estimate SHD not only using data generated from benchmark BNs, but also on large benchmark data sets.

The specific questions we address revolve around the concepts of uniformness and sparsity. From a practitioner's point of view, it is relevant that the available software for BN structure learning typically allow the user to set the maximum indegree of the DAG and to choose the scoring function, but not a non-uniform structure prior. The usual procedure is to set the maximum indegree as high as is computationally feasible, and then search for the highest scoring DAG. We ask, whether in this black-box setting, one can achieve better structure learning accuracy by deviating from the usual routine. Our answer is in the affirmative: we give such a procedure, *search space penalization (SSP)*, which also has an interpretation as a structure prior.

As SSP still has a flavor of uniform priors, it is fair to ask, whether the non-uniform priors proposed in the literature further enhance structure learning. To answer this question, we conduced an extensive empirical study to compare different priors. The results support the view that the uniform structure prior is inferior to those that favor simpler structures. Furthermore, we show that it would not not help, but rather harm, if the prior (unrealistically) concentrated on DAGs with the correct indegree for each node.

To shed light on the empirical results, we also present an analytical study. Specifically, we examine the marginal probability distribution on the indegree of a fixed node under different structure priors; to calculate these distributions, we generalize a recurrence formula known for counting unweighted DAGs (Robinson 1973). One might suspect that under the uniform prior, the marginal distribution concentrates at large indegrees (i.e., at the smaller of the maximum indegree and a half of the number of nodes). We confirm this hypothesis for the case of bounded indegree, but refute it for unbounded indegree. We discuss the results further at the end of the paper.

2 MODULAR PRIORS

This section gives a brief review of some structure priors presented in the literature. We exclusively focus on so-called modular priors, which are composed as a product of local terms. This sacrifice in generality is motivated by the fact that modular priors cover a large class of priors. Furthermore, modularity is needed for obtaining a decomposable scoring function, a requirement of most state-of-the-art learning algorithms. (See Supplement for the definition and examples of decomposable scoring functions.)

We will consider graphs on a node set $\{1, 2, \ldots, n\}$, denoted by $[n]$ for short, for some natural number n. If G is a DAG on $[n]$, we write G_i for the set of parents of i in G, i.e.,

$$G_i = \left\{ j : G \text{ contains an arc from } j \text{ to } i \right\}.$$

For brevity, we usually call $|G_i|$ the *indegree* rather than the number of parents of i. We denote by $\mathcal{G}_n$ the set of all DAGs on $[n]$ and by $\mathcal{G}_n^d$ the set of DAGs $G \in \mathcal{G}_n$ whose maximum indegree is at most d, i.e., $|G_i| \leq d$ for all $i \in [n]$.

We say a probability distribution P on $\mathcal{G}_n$ is *modular* if, for each $i \in [n]$, there exists a set function ρ_i from the subsets of $[n] \setminus \{i\}$ to nonnegative reals such that

$$P(G) = c \prod_{i=1}^{n} \rho_i(G_i) \qquad \text{for all } G \in \mathcal{G}_n\,,$$

with some normalizing constant c. We call the set functions *factors*. The reader may note that introducing the constant c is redundant in the definition, for the constant could be absorbed into the factors. However, the formulation is convenient, as it allows us to specify the factors without the trouble of ensuring that the normalizing constant equals unity.

Even if modular priors can express node-specific preferences, such as inclusion or exclusion certain nodes as parents, there is an interest in general-purpose priors that treat all nodes uniformly. Then a prior is specified by giving the maximum indegree d and an expression of ρ_i that only depends on the indegree $s := |G_i|$. Table 1 collects several forms of priors proposed in the literature; we assign each prior a name that captures some distinctive characteristic of the prior.

Some remarks on the structure priors listed in Table 1 are in order; see Angelopoulus and Cussens (2008) for historical notes, some variants, and discussion. Clearly, *Unif* is a special case of *Edge* with $\beta = 1$. Equivalent to *Edge* is the random graph model that contains an arc from node i to node j with probability p independently for all (i, j), however, disregarding graphs with a directed cycle (Madigan and Raftery 1994); the parameters are related by $\beta = p/(1-p)$. Buntine (1991) and Cooper and Herskovits (1992) considered variants of the *Edge* model, where the selection of arcs is conditional on a given node ordering. Friedman and Koller (2003) introduced the prior we

Table 1: Forms of modular structure priors over DAGs on n nodes.

Name	Factor, indegree s	Notes	Exemplary reference
Unif	1	Uniform over DAGs	—
Edge	β^s	Equivalent to the random graph model $p^s(1-p)^{n-1-s}$	Heckerman et al. (1995)
Fair	$1/\binom{n-1}{s}$	Balances the probabilities of different indegree	Friedman and Koller (2003)
Data	$\exp[-(1+\tau)^s \ln N]$	Depends on the data size N; by default $\tau = 0.5$	Pensar et al. (2016)

dub *Fair* similarly conditionally on a node ordering. When averaged over all orderings it results in a prior that is not modular but order-modular (Koivisto and Sood 2004). The prior stems from the idea of "fair" allocation of probability mass to different *numbers* of parents, whence the name. Apparently, *Fair* has rarely been included in empirical studies on score-and-search algorithms; an exception is a recent work on local structures (i.e., context-specific independence) by Talvitie et al. (2018). The *Data* prior is introduced in another work on local structures (Pensar et al. 2016). Unlike the other priors, *Data* is not a Bayesian prior as it depends on the sample size N, whence the name.

3 SEARCH SPACE PENALTY

Not all state-of-the-art software packages (Scutari 2010) allow the user to specify a structure prior; they offer a limited number of pre-implemented scoring functions, which assume a uniform prior. They do allow the user to control the maximum indegree, however. We next show how that enables implementing a nontrivial prior we call *search space penalty (SSP)*.

Recall that we denote by $\mathcal{G}_n^d$ the set of DAGs on $[n]$ with maximum indegree at most d. Clearly, these potential *search spaces* are nested:

$$\mathcal{G}_n^0 \subset \mathcal{G}_n^1 \subset \cdots \subset \mathcal{G}_n^{n-1} = \mathcal{G}_n \,.$$

For a DAG $G \in \mathcal{G}_n$, let $d(G)$ denote the maximum indegree of G. Now, define a prior P_{SSP} by letting the probability of G be inversely proportional to the size of the smallest search space $\mathcal{G}_n^d$ that contains G:

$$P_{SSP}(G) \propto 1/\left|\mathcal{G}_n^{d(G)}\right| .$$

In contrast to the variants discussed in the previous section, this structure prior is not modular.[1] However, the non-modularity of the prior does not cause a computational obstacle: finding a DAG G that maximizes a scoring function $f(G)$ under $P_{SSP}(G)$ reduces to maximizing the score under the uniform prior separately for each possible maximum indegree d:

[1] Alternatively, but with little difference in practice, one could consider a prior that is proportional to the inverse of $|\mathcal{G}_n^{d(G)} \setminus \mathcal{G}_n^{d(G)-1}|$. Neither this prior is modular.

Algorithm *Search Space Penalization*

S1 For each $d = 0, 1, \ldots, n-1$, let

$$\hat{G}(d) \in \arg\max \left\{ f(G) : G \in \mathcal{G}_n^d \right\} .$$

"Find an optimal DAG for each subclass."

S2 Let

$$\hat{d} \in \arg\max \left\{ f(\hat{G}(d))/|\mathcal{G}_n^d| : d = 0, 1, \ldots, n-1 \right\} .$$

"Penalize large subclasses."

S3 Output $\hat{G}(\hat{d})$.

We get the following (the proof is left to the reader):

Proposition 1. *Search Space Penalization outputs a DAG G that maximizes $P_{SSP}(G)f(G)$.*

From an MDL (Rissanen 1978) point of view, the *SSP* prior corresponds to encoding a DAG G by first encoding the maximum indegree $d(G)$ with about $\log_2 n$ bits, and then encoding G using about $\log_2 |\mathcal{G}_n^{d(G)}|$ bits.

The algorithm can be slower than the standard procedure by a factor of n in the worst case, a seemingly significant additional computational burden. However, in many practical settings, the complexity of step S1 is dominated by the time needed to search through the largest search space; and, furthermore, for that largest search space one could set the maximum indegree d to a value much smaller than $n-1$, e.g., $d = 5$.

It remains to show how we obtain the numbers $|\mathcal{G}_n^d|$. In the next section we give a recurrence (Corollary 4) that enables efficient computation of these numbers.

SSP readily applies for any baseline scoring function using exact or heuristic search algorithms. We also have the following result concerning *equivalent* DAGs, i.e., DAGs that encode exactly the same set of conditional independence relations:

Proposition 2. *SSP assigns the same score to equivalent DAGs, if the baseline scoring function does so.*

We omit the proof, which is straightforward and uses the fact that equivalent DAGs have the same maximum indegree (Chickering 1995; Thm. 9).

4 ANALYTICAL RESULTS

A modular prior can express, separately for each node i, which of the possible parents are *a priori* preferred as the actual parents of the node. The expression captured by the corresponding factor ρ_i in the modular function is, however, only approximate because the parents of different nodes are not independent: the actual parents must yield an acyclic graph. For this reason, e.g., a constant factor results in a non-uniform prior over parent sets for each node. We next investigate more closely the relationship of the factors and the resulting marginal prior probabilities.

We focus on a setting that is more special than the framework of modular priors, yet general enough to cover all the concrete priors listed in the previous section. Specifically, we assume that factors are symmetric in the sense that they are invariant under relabelling of the nodes. Equivalently, we assume that for each node i and potential parent set G_i we have that

$$\rho_i(G_i) = w(|G_i|)\,,$$

for some function w. It is easy to see that the factors listed in the previous section indeed are of this form.

How does a given weight function w map to a prior $P(|G_i|)$ on the indegree of node i? Due to the symmetry in w, these distributions are identical for all i. We have that, for $r = 0, 1, \ldots, n-1$, the probabilities $P(|G_i| = r)$ are proportional to the weighted sum of DAGs G where $|G_i| = r$. More precisely, by defining

$$Z_n(w) = \sum_{G \in \mathcal{G}_n} w(G) \quad \text{and} \quad Z_{n,r}(w) = \sum_{\substack{G \in \mathcal{G}_n \\ G_1 = \{2,3,\ldots,r+1\}}} w(G)\,,$$

we have that $P(|G_i| = r) = \binom{n-1}{r} Z_{n,r}(w)/Z_n(w)$.

Write $S_t := \sum_{s=0}^{t} \binom{t}{s} w(s)$. We find the following:

Theorem 3 (Recurrence)**.** *Let $Z_0 = Z_{0,0} = 1$ and $Z_{0,r} = 0$ for $r \geq 1$. For all $n \geq 1$ we have that*

$$\begin{aligned} Z_n &= \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} S_{n-k}^{k}\, Z_{n-k}\,, \\ Z_{n,r} &= \sum_{k=1}^{n-r} (-1)^{k-1} \binom{n-r-1}{k-1} w(r)\, S_{n-k}^{k-1}\, Z_{n-k} \\ &\quad + \sum_{k=1}^{n-r-1} (-1)^{k-1} \binom{n-r-1}{k} S_{n-k}^{k}\, Z_{n-k,r}\,. \end{aligned}$$

The proof (Supplement) uses the inclusion–exclusion method (Robinson 1973, Stanley 1973) and exploits the symmetry in w.

Corollary 4. *Let $a_n(d) := |\mathcal{G}_n^d|$ be the number of labelled DAGs with $n \geq 1$ nodes and maximum indegree $d \geq 0$. Let $a_0(d) = 1$. We have that*

$$a_n(d) = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} \left(\sum_{s=0}^{d} \binom{n-k}{s} \right)^{k} a_{n-k}(d)\,.$$

Proof. Apply Theorem 1 for the function w defined by $w(s) = 1$ if $s \leq d$ and $w(s) = 0$ otherwise. □

The recurrence formulas provide us with a means for efficient computation of the marginal prior distributions even for large numbers of nodes n. If the number of parents is bounded above by d (i.e., $w(s)$ vanishes for $s > d$), then the values $P(|G_1| = r)$, for $0 \leq r \leq d$, can be computed with $O(n^2 d)$ arithmetic operations.

Figure 1 shows the marginal distribution of the number of parents under different modular priors, for $n = 32$ and $n = 128$ nodes with maximum indegree $d = 5$ and with unbounded indegree. We see that *Unif* yields a non-uniform marginal. If we set the maximum indegree to 5, then having five parents is more probable than having zero parents; for $n = 32$ the ratio is about 10 and for $n = 128$ about 100. In contrast, if the indegree is unbounded, then the distribution is nearly uniform up to around $n/2$ parents, after which the probabilities rapidly decrease close to zero. Indeed, if we wished to support large parent sets, we should assign larger weights to larger numbers of parents; this is demonstrated in Fig. 1 by the *Fact* prior.

However, it is not possible to choose the factors so that the distribution would be *exactly* uniform:

Proposition 5. *For every modular distribution $P(G)$ on $\mathcal{G}_n$ with symmetric factors, the distribution $P(|G_i|)$ is non-uniform on $\{0, 1, \ldots, n-1\}$ for each node i.*

The proof, by the probabilistic method (Supplement), is based on the observation that a uniform distribution would imply existence of a DAG with so a large average number of parents that it contradicts acyclicity.

The rest of the priors, *Edge*, *Fair*, and *Data*, favor smaller numbers of parents. Under *Data*, five (or more) parents is several orders of magnitude less probable than zero or one parent. *Edge* is sensitive to the product of β and n: if the product is large (say, at least 10), the prior favors larger indegrees, up to around five, whereas if it is small, the prior renders larger indegrees very unlikely. *Fair* differs from the others in that it exhibits a mild preference for smaller indegrees in all scenarios. In summary, the priors *Edge*, *Fair*, and *Data* are similar in that they assign a relatively large probability to the smallest indegrees, from 0 to 3, unlike *Unif*; however, at larger indegrees the priors differ from each other significantly.

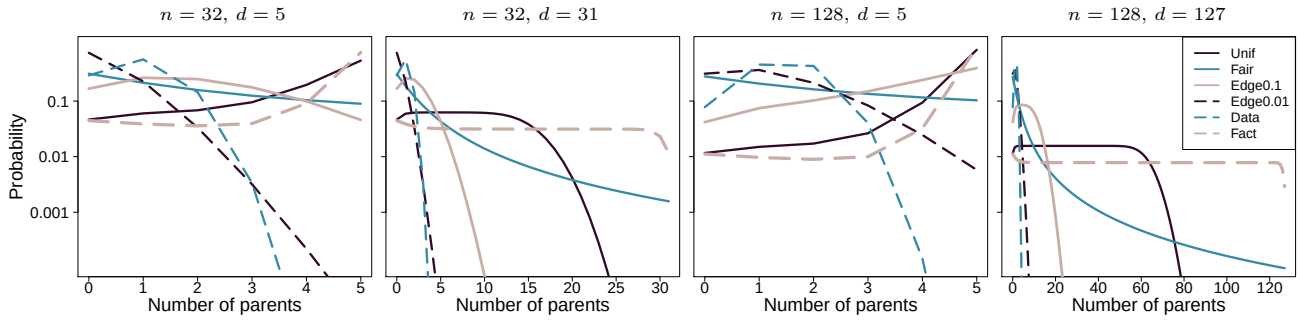

Figure 1: Probability distribution of the number of parents of a fixed node under different modular structure priors. In the *Data* prior the data size N is set to 200. The *Fact* prior is defined by letting $w(s) = s!$.

5 EMPIRICAL STUDIES

We investigate the practical effect of structure priors on structure learning by using both benchmark networks and real data. For searching through the space of DAGs, we generally prefer a globally optimal approach due to superior performance (Malone et al. 2015), but also employ a heuristic search algorithm when the network size demands it (Section 5.5).

Unless stated explicitly otherwise, we use throughout this section as baseline scoring function the BDeu score with an equivalent sample size of 1, which is despite recent criticism (Suzuki 2017) one of the most commonly used scoring functions. For structure prior *Edge*, we choose $\beta = 0.1$, since it performed best in preliminary studies, whereas for *Data*, we follow Pensar et al. (2016) and set $\tau = 0.5$.

In Sections 5.1–5.4, we evaluate structure priors based on four popular benchmark networks of a size that allows finding a globally optimal DAG (Table 2), under a maximum indegree of 5. For each network and sample size we generate ten data sets. For all data sets, we then learn a DAG for each method, i.e., for each combination of baseline score and structure prior. We compute all local scores using `bene` (Silander and Myllymäki 2006), add the penalties arising from the structure priors, and compute the globally optimal DAG using `GOBNILP` (Cussens 2011, Bartlett and Cussens 2013). For each resulting DAG, we compute the structural Hamming distance (SHD) of Tsamardinos et al. (2006) to the ground truth and average the SHDs for each method over the ten independent samples.

Table 2: Benchmark networks used in this study. *MaxIn* is the maximum indegree, *Param* is the number of free parameters of the network, and *Opt* indicates whether the search guarantees global optimality.

Network	Nodes	Arcs	MaxIn	Param	Opt
Child	20	25	2	230	Yes
Insurance	27	52	3	984	Yes
Water	32	66	5	10083	Yes
Alarm	37	46	4	509	Yes
Hailfinder	56	66	4	2656	No
Hepar2	70	123	6	1453	No
Win95pts	76	112	7	574	No
Andes	223	338	6	1157	No

5.1 Structure Priors for BDeu

First, we compare the effect of the different structure priors using BDeu as scoring function (Fig. 2). All non-uniform variants improve on *Unif* when the sample size is small in relation to the number of variables, which confirms that the BDeu score is indeed unsuitable for relatively small data sets. We also observe that *Fair*, *Data*, and *Edge* perform better than SSP and similar in direct comparison, which is in agree-

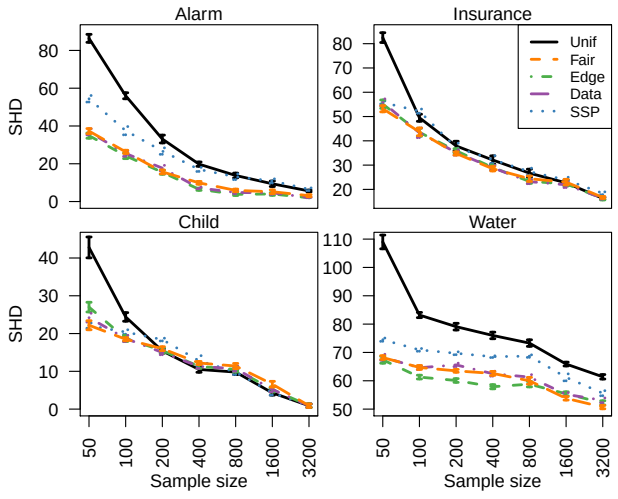

Figure 2: Effect of structure priors on benchmark network recovery for BDeu baseline score.

ment with the results of Section 4: unlike *Unif*, these priors assign relatively large probabilities to small indegrees—even if the priors diverge at larger indegrees (i.e., four or more), the differences have little effect at small data sets, because small data sets are insufficient for learning larger parent sets regardless of the prior. Justified by these observations, we now focus on structure prior *Fair* in what follows.

Next, we decompose the SHD into the individual contributions of spurious edges, missing edges, and incorrect edge orientations. We find that *Fair* dramatically reduces spurious edges; see Fig. 3 for one example. This error reduction comes at a cost, as missing edges are slightly increased. While these two effects are predictable, we also observe that structure priors reduce incorrect edge orientations. This can be explained as a side-effect of the generally reduced number of edges, which entails less acyclicity constraints.

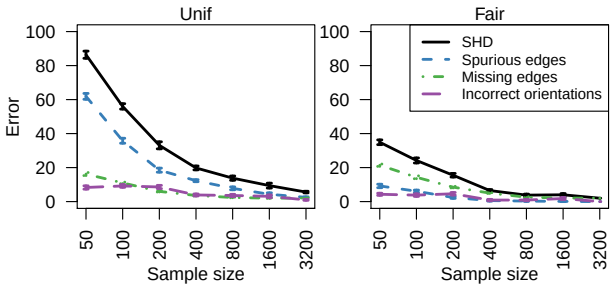

Figure 3: Error types for Alarm.

5.2 Utopian Priors

As complexity penalization can be viewed as a trade-off between spurious and missing edges, it is natural to ask whether the studied structure priors penalize complexity optimally already or whether there is substantial room for improvement. To study this question, we consider, in contrast to the general scope of this article, informative priors that we obtain directly from the ground-truth. Suppose we know the correct indegree d_i for each node i. We can set it as hard constraint to obtain a "utopian" prior (*Utop*) defined as

$$P_{Utop}(G) \propto \begin{cases} 1 & \text{if } |G_i| = d_i \text{ for all nodes } i\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

The practical effect of this prior may surprise at first glance: guiding the learning algorithm towards the correct solution may yield a substantial increase in SHD (Fig. 4). It can be explained by the fact that *Utop* forces each node to a (possibly large) indegree, whereas selecting the correct parent nodes remains challenging when the sample size is small. From the perspective of SHD, choosing a wrong parent counts twice: once as a spurious edge and once as a missing edge.

A less harsh constraint is obtained by just bounding the indegree from above, i.e., replacing the equalities in the definition of *Utop* by the inequalities $|G_i| \leq d_i$. This improved utopian prior, dubbed *Utop+*, is more conservative as it allows for learning smaller indegrees in the case of doubt, i.e., when only little data is available. However, even this variant never performs substantially better than *Fair*, suggesting relatively little room for improvement over *Fair* and the similarly performing *Edge* and *Data* priors.

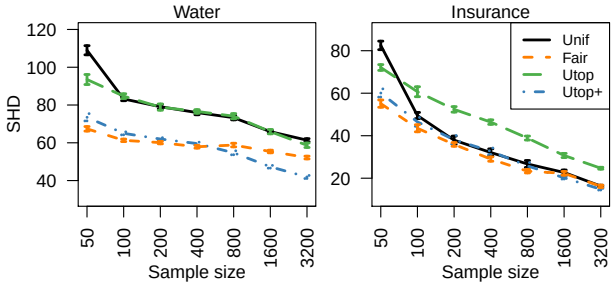

Figure 4: Effect of utopian priors.

5.3 Different Baseline Scores

We now study, in addition to BDeu, also BIC (Schwarz 1978), fNML (Silander et al. 2010), and qNML (Silander et al. 2018), as baseline scores, show summarizing results for comparing *Unif* and *Fair* in Fig. 5, and provide the full results with all structure prior variants in the Supplement. We observe that adding a structure prior decreases SHD dramatically for fNML, moderately for qNML, and has no visible effect for BIC. Moreover, we find that BIC with or without structure prior is not optimal, especially for Alarm. Interestingly, the errors of BIC in spurious edges and missing edges are comparable to the other well-performing methods such as BDeu+*Fair*. The difference is that BIC produces a much larger number of incorrectly oriented edges. This is due to the heavy penalty for the number of free parameters, which strongly prefers a chain or a common cause over a v-structure. Hence, there are comparably many undirected edges in the equivalence class representation of the learned DAG.

5.4 Other Aspects

The results of two further studies are shown only in the Supplement due to space constraints; we here briefly summarize the main findings. First, we also investigated the effects of structure priors on the recently proposed structural intervention distance (Peters and Bühlmann 2015). For this evaluation metric, penalizing complexity with structure priors or SSP gives no benefits. Second, we also studied different hyperparameter choices for priors *Edge* and *Data*. We observe

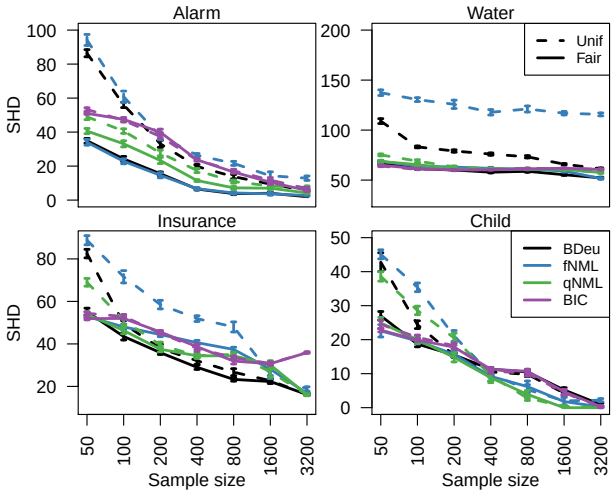

Figure 5: Comparison of different baseline scores.

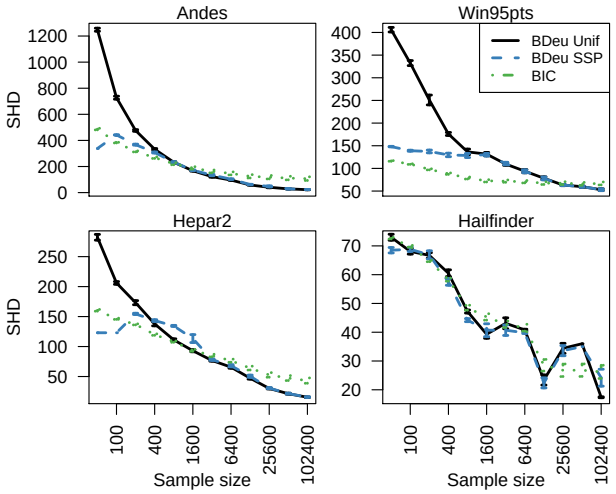

Figure 6: SSP for heuristic search.

that our default values perform well, only in the case of Water we could have obtained a better performance with favoring sparsity more (lower β and higher τ).

5.5 SSP for Heuristic Search

All previous studies used a globally optimal algorithm for searching through the space of DAGs that can take pre-computed local scores as input and thus easily permits to add modular structure priors to arbitrary baseline scoring functions. Now, we the setting, where the user does not have the flexibility of specifying own scoring functions, but can choose only from a set of pre-implemented variants. For this purpose, we use the tabu search algorithm from the `bnlearn` software package (Scutari 2010) and apply it to four common benchmark networks that are too large for finding the globally optimal DAG (Table 2) and thus require heuristic search for structure recovery.

We compare *Unif* with SSP for the BDeu baseline score, and also include plain BIC in the comparison (Fig. 6). We observe that SSP is effective for BDeu, yielding an equal or lower SHD than *Unif* with the exception of two sample sizes for Hepar2. We also find that BIC is often more effective than BDeu+SSP at small sample sizes, unless the lowest SHD is achieved by the empty network. However, BIC has the downside of a much slower convergence, whereas SSP behaves at large sample sizes identical to BDeu.

5.6 Evaluation on Real Data

All previous studies did rely on existing benchmark networks that can be treated as ground-truth. We now evaluate the effect of structure priors on the structure learning performance based on real data.

Here, SHD cannot be used for evaluation as no ground truth network is available. Hence, we employ the Intersection-Validation (InterVal) method (Viinikka et al. 2018). It allows to approximate the SHD by the so-called Partial Hamming Distance (PHD), which can be computed even if no ground-truth DAG is available. The central idea of InterVal is to consider only structural features (presence/absence and orientation of edges) that all methods agree upon when learning from the full data set. Treating these features as surrogate ground truth then allows to compare the methods at smaller subsamples of the original data set.

For the following study, we use data sets from Malone et al. (2018), which originate from the UCI machine learning repository [2] and are already processed for learning discrete BNs. We choose eight data sets that are large enough for applying the InterVal method (at least 1000 data points), but still permit learning a globally optimal DAG with `GOBNILP` when assuming an maximum indegree of three (Splice, Mushroom, Kr-vs-Kp, Optdigits) or four (remaining data sets).

Fig. 7 shows average PHDs for all data sets at $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, etc., of the full sample size. The observations from the the benchmark network study (Fig. 2) are confirmed: Adding a structure prior reduces errors at small sample sizes. Moreover, we find evidence that *Fair* is on real data slightly superior to the other modular priors. The PHDs for the other baseline scoring functions are shown in the Supplement, and confirm the previously observed trend: structure priors improve on *Unif* dramatically for fNML, moderately for qNML, and imperceptibly for BIC. Since they also never perform substantially worse, using a structure prior is a robust choice that can be generally recommended.

[2] `http://archive.ics.uci.edu/ml`

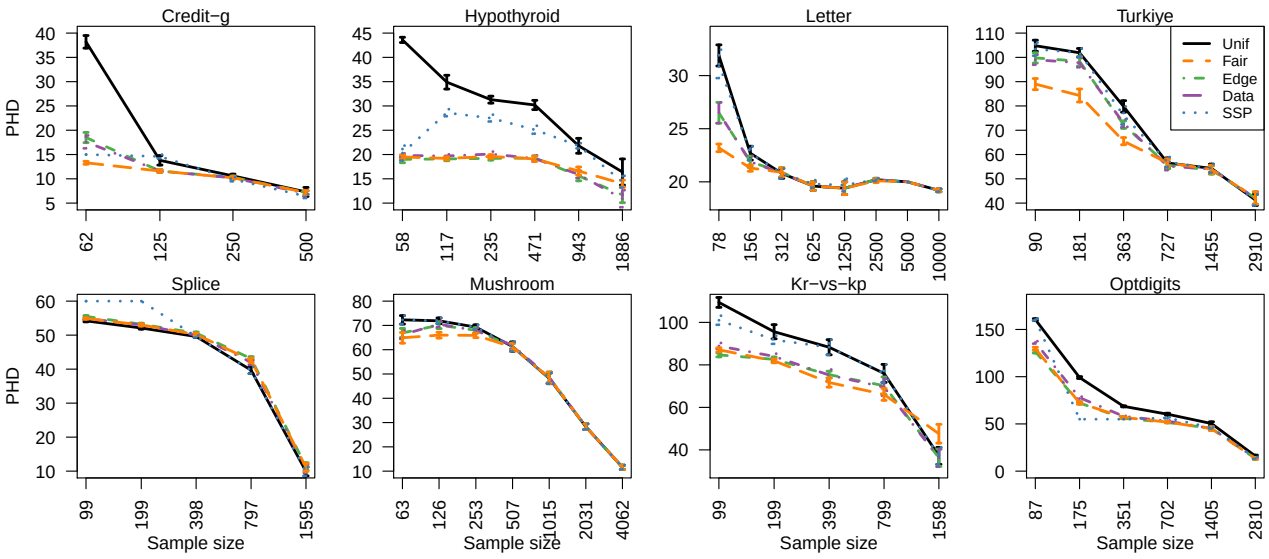

Figure 7: Evaluation of structure priors on real world data sets using the InterVal method.

6 CONCLUSIONS

We have studied the popular score-and-search approach to structure learning in BNs. Previous research has shown that the chosen scoring function matters when one seeks a DAG that is structurally similar to the ground truth DAG, as measured by SHD. Our observations concur with that (Fig. 5). However, most previous studies have effectively assumed a structure prior that is uniform over all DAGs, restricted by some maximum indegree.

In this work we challenged the uniform prior. Our empirical findings suggest that the uniform prior is inferior to various sparsity favoring priors; such priors have so far been employed mainly in the full Bayesian averaging framework, which deviates from the score-and-search approach. Adding a sparsity-favoring prior boosts the performance of all tested scores, except BIC, and render their differences relatively small (Fig. 5). In this light, it is more important to choose an appropriate structure prior than tuning the likelihood part of the score.

A tempting explanation for the inferiority of the uniform prior is that sparsity-favoring priors better match the actual sparsity of the ground truth DAG. This explanation is flawed. Indeed, our study (with the utopian prior) showed that conditioning on the correct indegree for each node rather harms. Another explanation could be that SHD simply favors the sparsest DAG: it is best to output the DAG with no arcs. Also this conclusion is flawed. The empty DAG, for which SHD equals the number of arcs in the ground truth (cf. Table 2), would perform well only at very small data samples, but no longer on moderate size samples where sparsity-favoring priors still show a significant advantage over the uniform prior.

The correct explanation appears to stem from balancing the two extremes. Sparsity-favoring priors are superior because they assign a relatively large prior probability on small indegrees, and yet allow learning larger indegrees as the size of the data grows. It is crucial to avoid assigning too large prior on large indegrees, which would force a highest-scoring DAG include arcs that are likely to be spurious at smaller data sizes.

Among the alternative sparsity-favoring structure priors, our empirical results on data generated from benchmark BNs show little difference. However, the analytical results suggest that the *Fair* prior is perhaps the most robust; the empirical results on benchmark data sets also provide some support for this conclusion.

Motivated by the fact that many existing software packages do not allow the user to choose a non-uniform structure prior, we also introduced the search space penalization (SSP) method. SSP appears to perform better than the uniform prior, even if not being competitive to the sparsity-favoring priors. It is worth noting that SSP is not only applicable to BNs, but to any model family that is nested in relation to some natural complexity parameter.

Acknowledgements

The authors thank the anonymous reviewers for valuable suggestions that helped to improve the presentation. This work was supported in part by the Academy of Finland, Grant 276864.

References

N. Angelopoulus and J. Cussens. Bayesian learning of Bayesian networks with informative priors. *Ann. Math. Artif. Intell.*, 54(1–3):53–98, 2008.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6): 2743–2760, 1998.

M. Bartlett and J. Cussens. Advances in Bayesian network learning using integer programming. In *Proc. UAI*, pages 182–191, 2013.

W. L. Buntine. Theory refinement on Bayesian networks. In *Proc. UAI*, pages 52–60, 1991.

D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proc. UAI*, pages 87–98, 1995.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

J. Cussens. Bayesian network learning with cutting planes. In *Proc. UAI*, pages 153–160, 2011.

N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1-2):95–125, 2003.

D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, May 2004.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

B. Malone, M. Järvisalo, and P. Myllymäki. Impact of learning strategies on the quality of Bayesian networks: An empirical evaluation. In *Proc. UAI*, pages 562–571, 2015.

B. Malone, K. Kangas, M. Järvisalo, M. Koivisto, and P. Myllymäki. Empirical hardness of finding optimal Bayesian network structures: algorithm selection and runtime prediction. *Machine Learning*, 107 (1):247–283, 2018.

J. Pensar, H. Nyman, J. Lintusaari, and J. Corander. The role of local partial independence in learning of Bayesian networks. *International Journal of Approximate Reasoning*, 69:91–105, 2016.

J. Peters and P. Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3):771–799, 2015.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

R. Robinson. Counting labeled acyclic digraphs. In F. Harary, editor, *New Directions in the Theory of Graphs*, pages 239–273. 1973.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 2:461–464, 1978.

M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35 (3):1–22, 2010.

T. Silander and P. Myllymäki. A simple approach for finding the globally optimal Bayesian network structure. In *Proc. UAI*, pages 445–452, 2006.

T. Silander, T. Roos, and P. Myllymäki. Learning locally minimax optimal Bayesian networks. *International Journal of Approximate Reasoning*, 51:544–557, 2010.

T. Silander, J. Leppä-aho, E. Jääsaari, and T. Roos. Quotient normalized maximum likelihood criterion for learning Bayesian network structures. In *Proc. AISTATS*, pages 948–957, 2018.

R. Stanley. Acyclic orientations of graphs. *Discrete Mathematics*, 5(2):171–178, 1973.

J. Suzuki. A theoretical analysis of the BDeu scores in Bayesian network structure learning. *Behaviormetrika*, 44(1):97–116, 2017.

T. Talvitie, R. Eggeling, and M. Koivisto. Finding optimal Bayesian networks with local structure. In *Proc. PGM*, pages 451–462, 2018.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

J. Viinikka, R. Eggeling, and M. Koivisto. Intersection-validation: A method for evaluating structure learning without ground truth. In *Proc. AISTATS*, pages 1570–1578, 2018.