Beyond Uniform Priors in Bayesian Network Structure Learning

(for Discrete Bayesian Networks)

Marco Scutari

scutari@stats.ox.ac.uk
Department of Statistics
University of Oxford

April 5, 2017

Bayesian Network Structure Learning

Learning a BN $\mathcal{B} = (\mathcal{G}, \Theta)$ from a data set $\mathcal{D}$ is performed in two steps:

$$\underbrace{\mathrm{P}(\mathcal{B} \mid \mathcal{D}) = \mathrm{P}(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}.$$

In a Bayesian setting structure learning consists in finding the DAG with the best $\mathrm{P}(\mathcal{G} \mid \mathcal{D})$ (BIC [5] is a common alternative) with some heuristic search algorithm. We can decompose $\mathrm{P}(\mathcal{G} \mid \mathcal{D})$ into

$$\mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{G})\,\mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \mathrm{P}(\mathcal{G}) \int \mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta)\,\mathrm{P}(\Theta \mid \mathcal{G}) d\Theta$$

where $\mathrm{P}(\mathcal{G})$ is the prior distribution over the space of the DAGs and $\mathrm{P}(\mathcal{D} \mid \mathcal{G})$ is the marginal likelihood of the data given $\mathcal{G}$ averaged over all possible parameter sets Θ; and then

$$\mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \prod_{i=1}^{N} \left[\int \mathrm{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i})\,\mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i}) d\Theta_{X_i} \right].$$

where Π_{X_i} are the parents of X_i in $\mathcal{G}$.

The Bayesian Dirichlet Marginal Likelihood

If $\mathcal{D}$ contains no missing values and assuming:

- a Dirichlet conjugate prior ($X_i \mid \Pi_{X_i} \sim \mathit{Multinomial}(\Theta_{X_i} \mid \Pi_{X_i})$ and $\Theta_{X_i} \mid \Pi_{X_i} \sim \mathit{Dirichlet}(\alpha_{ijk})$, $\sum_{jk} \alpha_{ijk} = \alpha_i$ the imaginary sample size);
- positivity (all conditional probabilties $\pi_{ijk} > 0$);
- parameter independence (π_{ijk} for different parent configurations are independent) and modularity (π_{ijk} in different nodes are independent);

Heckerman *et al.* [2] derived a closed form expression for $\mathrm{P}(\mathcal{D} \mid \mathcal{G})$:

$$\mathrm{BD}(\mathcal{G}, \mathcal{D}; \boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathrm{BD}(X_i, \Pi_{X_i}; \alpha_i) =$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \left[\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right]$$

where r_i is the number of states of X_i; q_i is the number of configurations of Π_{X_i}; $n_{ij} = \sum_k n_{ijk}$; and $\alpha_{ij} = \sum_k \alpha_{ijk}$.

Bayesian Dirichlet Equivalent Uniform (BDeu)

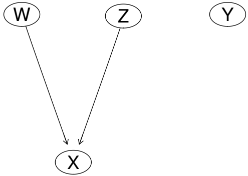
The most common implementation of BD assumes $\alpha_{ijk} = \alpha/(r_i q_i)$, $\alpha_i = \alpha$ and is known from [2] as the Bayesian Dirichlet equivalent uniform (BDeu) marginal likelihood. The uniform prior over the parameters was justified by the lack of prior knowledge and widely assumed to be non-informative.

However, there is ample evidence that this is a problematic choice:

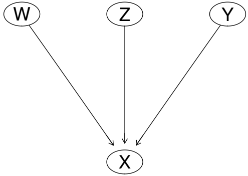
- The prior is actually not uninformative.
- MAP DAGs selected using BDeu are highly sensitive to the choice of α and can have markedly different number of arcs even for reasonable α [8].
- In the limits $\alpha \to 0$ and $\alpha \to \infty$ it is possible to obtain both very simple and very complex DAGs, and model comparison may be inconsistent for small $\mathcal{D}$ and small α [8, 10].
- The sparseness of the MAP network is determined by a complex interaction between α and $\mathcal{D}$ [10, 13].
- There are formal proofs of all this in [12, 13].

Exhibits A and B

$\mathcal{G}^-$



$\mathcal{G}^+$



$\mathcal{D}_1$

X	Z	W	Y
1	0	0	0
0	0	0	0
0	0	0	0
0	0	1	0
1	0	1	0
1	0	1	0
0	1	0	0
1	1	0	0
1	1	0	0
1	1	1	1
0	1	1	1
0	1	1	1

$\mathcal{D}_2$

X	Z	W	Y
0	0	0	0
0	0	0	0
0	0	0	0
1	0	1	0
1	0	1	0
1	0	1	0
1	1	0	0
1	1	0	0
1	1	0	0
0	1	1	1
0	1	1	1
0	1	1	1

Exhibit A

The sample frequencies (n_{ijk}) for $X \mid \Pi_X$ are:

		Z, W			
		0, 0	1, 0	0, 1	1, 1
X	0	2	1	1	2
	1	1	2	2	1

and those for $X \mid \Pi_X \cup Y$ are as follows.

		Z, W, Y							
		0, 0, 0	1, 0, 0	0, 1, 0	1, 1, 0	0, 0, 1	1, 0, 1	0, 1, 1	1, 1, 1
X	0	2	1	1	0	0	0	0	2
	1	1	2	2	0	0	0	0	1

Even though $X \mid \Pi_X$ and $X \mid \Pi_X \cup Y$ have the same entropy,

$$\mathrm{H}(X \mid \Pi_X) = \mathrm{H}(X \mid \Pi_X \cup Y) = 4\left[-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right] = 2.546\ldots$$

Exhibit A

... $\mathcal{G}^-$ has a higher entropy than $\mathcal{G}^+$ *a posteriori* ...

$$\begin{aligned}
\mathrm{H}(X \mid \Pi_X; \alpha) &= 4\left[-\frac{1 + 1/8}{3 + 1/4}\log\frac{1 + 1/8}{3 + 1/4} - \frac{2 + 1/8}{3 + 1/4}\log\frac{2 + 1/8}{3 + 1/4}\right] \\
&= 2.580 \\
\mathrm{H}(X \mid \Pi_X \cup Y; \alpha) &= 4\left[-\frac{1 + 1/16}{3 + 1/8}\log\frac{1 + 1/16}{3 + 1/8} - \frac{2 + 1/16}{3 + 1/8}\log\frac{2 + 1/16}{3 + 1/8}\right] \\
&= 2.564
\end{aligned}$$

... and BDeu with $\alpha = 1$ chooses accordingly, and things fortunately work out:

$$\begin{aligned}
\mathrm{BDeu}(X \mid \Pi_X) &= \left(\frac{\Gamma(1/4)}{\Gamma(1/4 + 3)}\left[\frac{\Gamma(1/8 + 2)}{\Gamma(1/8)} \cdot \frac{\Gamma(1/8 + 1)}{\Gamma(1/8)}\right]\right)^4 \\
&= 3.906 \times 10^{-7}, \\
\mathrm{BDeu}(X \mid \Pi_X \cup Y) &= \left(\frac{\Gamma(1/8)}{\Gamma(1/8 + 3)}\left[\frac{\Gamma(1/16 + 2)}{\Gamma(1/16)} \cdot \frac{\Gamma(1/16 + 1)}{\Gamma(1/16)}\right]\right)^4 \\
&= 3.721 \times 10^{-8}.
\end{aligned}$$

Exhibit B

The sample frequencies for $X \mid \Pi_X$ are:

		Z, W			
		0, 0	1, 0	0, 1	1, 1
X	0	3	0	0	3
	1	0	3	3	0

and those for $X \mid \Pi_X \cup Y$ are as follows.

		Z, W, Y							
		0, 0, 0	1, 0, 0	0, 1, 0	1, 1, 0	0, 0, 1	1, 0, 1	0, 1, 1	1, 1, 1
X	0	3	0	0	0	0	0	0	3
	1	0	3	3	0	0	0	0	0

The conditional entropy of X is equal to zero for both $\mathcal{G}^+$ and $\mathcal{G}^-$, since the value of X is completely determined by the configurations of its parents in both cases.

Exhibit B

Again, the posterior entropies for $\mathcal{G}^+$ and $\mathcal{G}^-$ differ:

$$\mathrm{H}(X \mid \Pi_X; \alpha) = 4\left[-\frac{0 + 1/8}{3 + 1/4}\log\frac{0 + 1/8}{3 + 1/4} - \frac{3 + 1/8}{3 + 1/4}\log\frac{3 + 1/8}{3 + 1/4}\right] = 0.652,$$

$$\mathrm{H}(X \mid \Pi_X \cup Y; \alpha) = 4\left[-\frac{0 + 1/16}{3 + 1/8}\log\frac{0 + 1/16}{3 + 1/8} - \frac{3 + 1/16}{3 + 1/8}\log\frac{3 + 1/16}{3 + 1/8}\right] = 0.392.$$

However, BDeu with $\alpha = 1$ yields

$$\mathrm{BDeu}(X \mid \Pi_X) = \left(\frac{\Gamma(1/4)}{\Gamma(1/4 + 3)}\left[\frac{\Gamma(1/8 + 3)}{\Gamma(1/8)} \cdot \cancel{\frac{\Gamma(1/8)}{\Gamma(1/8)}}\right]\right)^4 = 0.032,$$

$$\mathrm{BDeu}(X \mid \Pi_X \cup Y) = \left(\frac{\Gamma(1/8)}{\Gamma(1/8 + 3)}\left[\frac{\Gamma(1/16 + 3)}{\Gamma(1/16)} \cdot \cancel{\frac{\Gamma(1/16)}{\Gamma(1/16)}}\right]\right)^4 = 0.044,$$

preferring $\mathcal{G}^+$ over $\mathcal{G}^-$ even though the additional arc $Y \to X$ does not provide any additional information on the distribution of X, and even though 4 out of 8 conditional distributions in $X \mid \Pi_X \cup Y$ are not observed at all in the data.

Better Than BDeu: Bayesian Dirichlet Sparse (BDs)

If the positivity assumption is violated or the sample size n is small, there may be configurations of some Π_{X_i} that are not observed in $\mathcal{D}$.

$$\mathrm{BDeu}(X_i, \Pi_{X_i}; \alpha) =$$
$$= \prod_{j:n_{ij}=0} \left[\cancel{\frac{\Gamma(r_i\alpha^*)}{\Gamma(r_i\alpha^*)} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha^*)}{\Gamma(\alpha^*)}} \right] \prod_{j:n_{ij}>0} \left[\frac{\Gamma(r_i\alpha^*)}{\Gamma(r_i\alpha^* + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha^* + n_{ijk})}{\Gamma(\alpha^*)} \right],$$

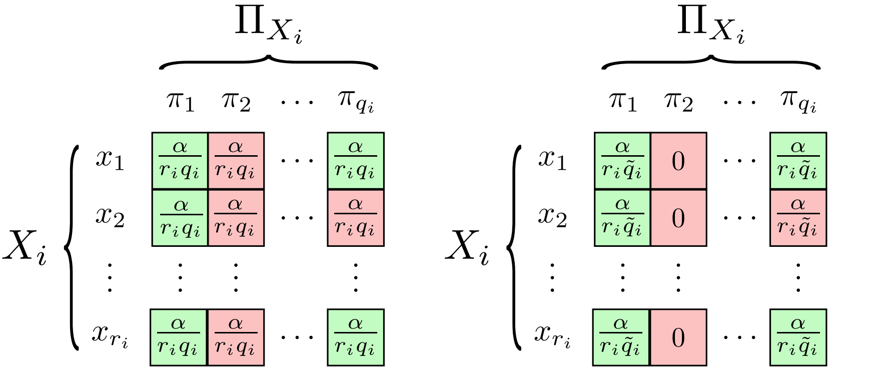
so the effective imaginary sample size decreases as the number of unobserved parents configurations increases. We can prevent that by replacing α_{ijk} with

$$\tilde{\alpha}_{ijk} = \begin{cases} \alpha/(r_i\tilde{q}_i) & \text{if } n_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}, \quad \tilde{q}_i = \{\text{number of } \Pi_{X_i} \text{ such that } n_{ij} > 0\}$$

and we plug it in BD instead of $\alpha_{ijk} = \alpha/(r_i q_i)$ to obtain BDs.

Then $\mathrm{BDs}(X_i, \Pi_{X_i}; \alpha) = \mathrm{BDeu}(X_i, \Pi_{X_i}; \alpha q_i/\tilde{q}_i)$.

BDeu and BDs Compared



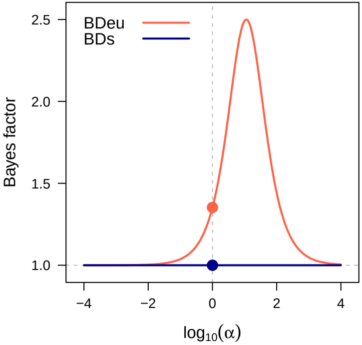
Cells that correspond to $(\mathbf{X}_i, \Pi_{X_i})$ combinations that are not observed in the data are in red, observed combinations are in green.

Exhibits A and B, Once More

BDs does not suffer from the bias arising from $\tilde{q}_i < q_i$ and it correctly assigns the same score to $\mathcal{G}^-$ and $\mathcal{G}^+$ in both examples,

$$\mathrm{BDs}(X \mid \Pi_X) = \mathrm{BDs}(X \mid \Pi_X \cup Y) = 3.906 \times 10^{-7}.$$
$$\mathrm{BDs}(X \mid \Pi_X) = \mathrm{BDs}(X \mid \Pi_X \cup Y) = 0.03262.$$

following the maximum entropy principle.



Entropy and BDeu

In a Bayesian setting, the conditional entropy $\mathrm{H}(\cdot)$ of $X \mid \Pi_X$ given a uniform Dirichlet prior with imaginary sample size α over the cell probabilities is

$$\mathrm{H}(X \mid \Pi_X; \alpha) = -\sum_{j : n_{ij} > 0} \sum_{k=1}^{r_i} p_{ij|k}^{(\alpha_i^*)} \log p_{ij|k}^{(\alpha_i^*)} \quad \text{with} \quad p_{ij|k}^{(\alpha_i^*)} = \frac{\alpha_i^* + n_{ijk}}{r_i \alpha_i^* + n_{ij}}.$$

and $\mathrm{H}(X \mid \Pi_X; \alpha) > \mathrm{H}(X \mid \Pi_X; \beta)$ if $\alpha > \beta$ and $X \mid \Pi_X$ is not a uniform distribution.

Let $\alpha/(r_i q_i) \to 0$ and let $\alpha > \beta > 0$. Then

$$\begin{cases} \mathrm{BDeu}(X \mid \Pi_X; \alpha) > \mathrm{BDeu}(X \mid \Pi_X; \beta) & \text{if } d_{\mathrm{EP}}^{(X_i, \mathcal{G})} > 0, \\ \mathrm{BDeu}(X \mid \Pi_X; \alpha) = \left(\frac{1}{r_i}\right)^{\tilde{q}_i} & \text{if } d_{\mathrm{EP}}^{(X_i, \mathcal{G})} = 0. \end{cases}$$

To Sum It Up in a Theorem

Let $\mathcal{G}^+$ and $\mathcal{G}^-$ be two DAGs differing from a single arc $X_j \to X_i$, and let $\alpha/(r_i q_i) \to 0$. Then the Bayes factor computed using BDs corresponds to the Bayes factor computed using BDeu weighted by the following implicit prior ratio:

$$\frac{\mathrm{P}(\mathcal{G}^+)}{\mathrm{P}(\mathcal{G}^-)} = \frac{(q_i/\tilde{q}_i)^{d_{\mathrm{EP}}^{(X_i,\mathcal{G}^+)}}}{(q'_i/\tilde{q}'_i)^{d_{\mathrm{EP}}^{(X_i,\mathcal{G}^-)}}}.$$

and can be written as

$$\frac{\mathrm{BDs}(X_i, \Pi_{X_i} \cup X_j; \alpha)}{\mathrm{BDs}(X_i, \Pi_{X_i}; \alpha)} = \frac{(q_i/\tilde{q}_i)^{d_{\mathrm{EP}}^{(X_i,\mathcal{G}^+)}} \alpha^{d_{\mathrm{EP}}^{(\mathcal{G}^+)}}}{(q'_i/\tilde{q}'_i)^{d_{\mathrm{EP}}^{(X_i,\mathcal{G}^-)}} \alpha^{d_{\mathrm{EP}}^{(\mathcal{G}^-)}}}$$

$$\to \begin{cases} 0 & \text{if } d_{\mathrm{EDF}} > -\log_\alpha(\mathrm{P}(\mathcal{G}^+)/\,\mathrm{P}(\mathcal{G}^-)) \\ +\infty & \text{if } d_{\mathrm{EDF}} < -\log_\alpha(\mathrm{P}(\mathcal{G}^+)/\,\mathrm{P}(\mathcal{G}^-)) \end{cases}.$$

The Uniform (U) Graph Prior

The most common choice for $\mathrm{P}(\mathcal{G})$ is the uniform (U) distribution because it is extremely difficult to specify informative priors [1, 3]. Assuming a uniform prior is problematic because:

- Score-based structure learning algorithms typically generate new candidate DAGs by a single arc addition, deletion or reversal, e.g.

$$\frac{\mathrm{P}(\mathcal{G} \cup \{X_j \to X_i\} \mid \mathcal{D})}{\mathrm{P}(\mathcal{G} \mid \mathcal{D})} = \cancel{\frac{\mathrm{P}(\mathcal{G} \cup \{X_j \to X_i\})}{\mathrm{P}(\mathcal{G})}} \frac{\mathrm{P}(\mathcal{D} \mid \mathcal{G} \cup \{X_j \to X_i\})}{\mathrm{P}(\mathcal{D} \mid \mathcal{G})}.$$

  U always simplifies, and that implies $\overrightarrow{p_{ij}} = \overleftarrow{p_{ij}} = \mathring{p}_{ij} = 1/3$ favouring the inclusion of new arcs as $\overrightarrow{p_{ij}} + \overleftarrow{p_{ij}} = 2/3$ for each possible arc a_{ij}.

- Two arcs are correlated if they are incident on a common node [7], so false positives and false negatives can potentially propagate through $\mathrm{P}(\mathcal{G})$ and lead to further errors in learning $\mathcal{G}$.

- DAGs that are completely unsupported by the data have most of the probability mass for large enough N.

Better Than U: the Marginal Uniform (MU) Graph Prior

In our previous work [7], we showed that

$$\overrightarrow{p_{ij}} = \overleftarrow{p_{ij}} \approx \frac{1}{4} + \frac{1}{4(N-1)} \to \frac{1}{4} \quad \text{and} \quad \mathring{p}_{ij} \approx \frac{1}{2} - \frac{1}{2(N-1)} \to \frac{1}{2},$$

so each possible arc is present in $\mathcal{G}$ with marginal probability $\approx 1/2$ and, when present, it appears in each direction with probability $1/2$. We can use that as a starting point, and assume an independent prior for each arc with the same marginal probabilities as U (hence the name MU).

- MU does not favour arc inclusion as $\overrightarrow{p_{ij}} + \overleftarrow{p_{ij}} = 1/2$.
- MU does not favour the propagation of errors in structure learning because arcs are independent from each other.
- MU computationally trivial to use: the ratio of the prior probabilities is $1/2$ for arc addition, 2 for arc deletion and 1 for arc reversal, for all arcs.

We can also assume $\overrightarrow{p_{ij}} + \overleftarrow{p_{ij}} = \beta$ with $\beta = \frac{2}{N-1}$ to have $O(N)$ expected arcs in the prior, which often works even better.

Design of the Simulation Study

We evaluated BIC and U+BDeu, U+BDs, MU+BDeu, MU+BDs with $\alpha = 1, 5, 10$ on:

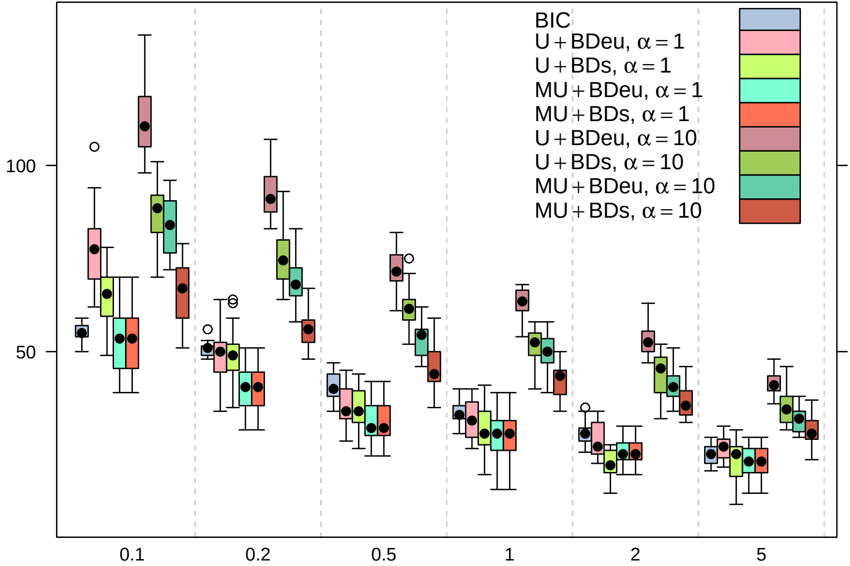
- 10 reference BNs covering a wide range of N (8 to 442), $p = |\Theta|$ (18 to 77K) and number of arcs $|A|$ (8 to 602).
- 20 samples of size $n/p = 0.1$, 0.2, 0.5, 1.0, 2.0, and 5.0 (to allow for meaningful comparisons between BNs with such different N and p) for each BN and n/p.

with performance measures for:

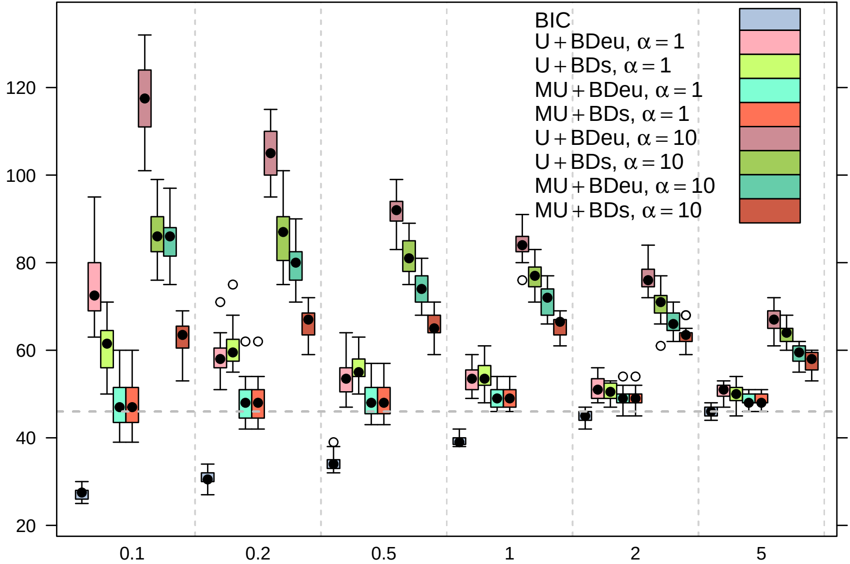
- the quality of the learned DAG using the SHD distance [11] from the reference BN;
- the number of arcs compared to the reference BN;
- the log-likelihood on a separate test set of size 10K, as an approximation of Kullback-Leibler distance.

using hill-climbing and the **bnlearn** R package [6].

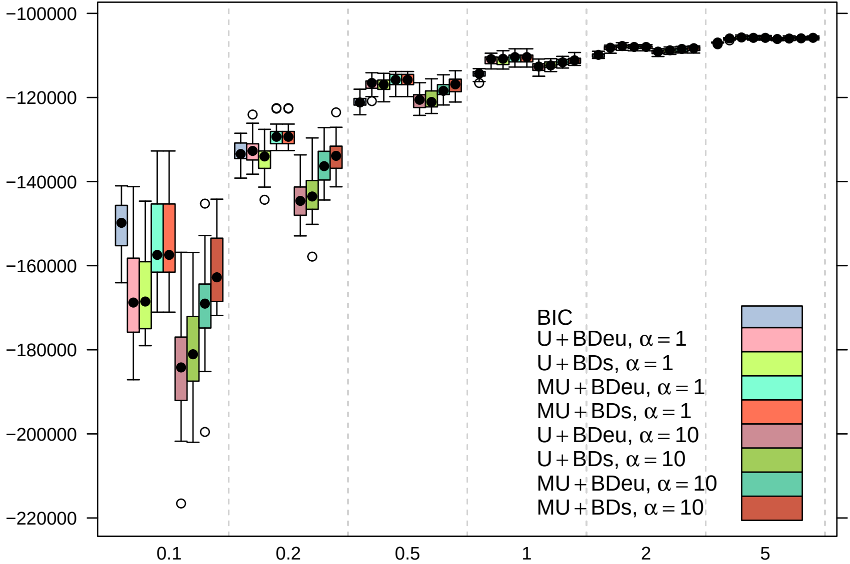
Results: ALARM, SHD



Results: ALARM, Number of Arcs



Results: ALARM, Log-likelihood on the Test Set



Conclusions

- We propose a new default posterior score for discrete BN structure learning, defined it as the combination of a new prior over the space of DAGs, the marginal uniform (MU) prior, and of a new empirical Bayes marginal likelihood, which we call Bayesian Dirichlet sparse (BDs).
- In an extensive simulation study using 10 reference BNs we find that MU+BDs outperforms U+BDeu for all combinations of BN and sample sizes, both in the quality of the learned DAGs and in predictive accuracy. Other proposals in the literature improve one at the expense of the other [4, 9, 13, 14].
- This is achieved without increasing the computational complexity of the posterior score, since MU+BDs can be computed in the same time as U+BDeu.

Thanks!

References

References I

R. Castelo and A. Siebes.
Priors on Network Structures. Biasing the Search for Bayesian Networks.
International Journal of Approximate Reasoning, 24(1):39–57, 2000.

D. Heckerman, D. Geiger, and D. M. Chickering.
Learning Bayesian Networks: The Combination of Knowledge and Statistical Data.
Machine Learning, 20(3):197–243, 1995.
Available as Technical Report MSR-TR-94-09.

S. Mukherjee and T. P. Speed.
Network Inference Using Informative Priors.
Proceedings of the National Academy of Sciences, 105(38):14313–14318, 2008.

M. Scanagatta, C. P. de Campos, and M. Zaffalon.
Min-BDeu and Max-BDeu Scores for Learning Bayesian Networks.
In *Proceedings of the 7th Probabilistic Graphical Model Workshop*, pages 426–441, 2014.

G. Schwarz.
Estimating the Dimension of a Model.
The Annals of Statistics, 6(2):461–464, 1978.

References II

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

M. Scutari.
On the Prior and Posterior Distributions Used in Graphical Modelling (with discussion).
Bayesian Analysis, 8(3):505–532, 2013.

T. Silander, P. Kontkanen, and P. Myllymäki.
On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter.
In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 360–367, 2007.

H. Steck.
Learning the Bayesian Network Structure: Dirichlet Prior versus Data.
In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 511–518, 2008.

H. Steck and T. S. Jaakkola.
On the Dirichlet Prior and Bayesian Regularization.
In *Advances in Neural Information Processing Systems 15*, pages 713–720. 2003.

References III

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.

M. Ueno.
Learning Networks Determined by the Ratio of Prior and Data.
In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 598–605, 2010.

M. Ueno.
Robust Learning of Bayesian Networks for Prior Belief.
In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 698–707, 2011.

M. Ueno and M. Uto.
Non-Informative Dirichlet Score for Learning Bayesian Networks.
In *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*, pages 331–338, 2012.