

An Empirical-Bayes Score for Discrete Bayesian Networks



UNIVERSITY OF
OXFORD

Marco Scutari

scutari@stats.ox.ac.uk

Department of Statistics
University of Oxford

September 8, 2016

Bayesian Network Structure Learning

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

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

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

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

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

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

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

[2] derived a closed form expression for $P(\mathcal{D} | \mathcal{G})$:

$$\begin{aligned} \text{BD}(\mathcal{G}, \mathcal{D}; \boldsymbol{\alpha}) &= \prod_{i=1}^N \text{BD}(X_i, \Pi_{X_i}; \alpha_i) = \\ &= \prod_{i=1}^N \prod_{j=1}^{q_i} \left[\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right] \end{aligned}$$

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 $\alpha \rightarrow 0$ and $\alpha \rightarrow \infty$ 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].

The Uniform (U) Graph Prior

The most common choice for $P(\mathcal{G})$ 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.

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

U always simplifies, and that implies $\overrightarrow{p}_{ij} = \overleftarrow{p}_{ij} = p_{ij}^\circ = 1/3$ **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 $P(\mathcal{G})$** 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 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} .

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

So the **effective imaginary sample size decreases as the number of unobserved parents configurations increases**, and the MAP estimates of π_{ijk} gradually converge to the ML and favour overfitting.

To address these two undesirable features of BDeu we replace α^* with

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

and we plug it in BD instead of $\alpha^* = \alpha / (r_i q_i)$ to obtain BDs.

BDeu and BDs Compared

$$\begin{array}{c}
 \underbrace{\hspace{10em}}_{\Pi_{X_i}} \\
 \pi_1 \quad \pi_2 \quad \dots \quad \pi_{q_i} \\
 \left. \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_{r_i} \end{array} \right\} X_i \quad
 \begin{array}{ccc}
 \begin{array}{|c|c|} \hline \frac{\alpha}{r_i q_i} & \frac{\alpha}{r_i q_i} \\ \hline \end{array} & \dots & \begin{array}{|c|} \hline \frac{\alpha}{r_i q_i} \\ \hline \end{array} \\
 \begin{array}{|c|c|} \hline \frac{\alpha}{r_i q_i} & \frac{\alpha}{r_i q_i} \\ \hline \end{array} & \dots & \begin{array}{|c|} \hline \frac{\alpha}{r_i q_i} \\ \hline \end{array} \\
 \vdots & & \vdots \\
 \begin{array}{|c|c|} \hline \frac{\alpha}{r_i q_i} & \frac{\alpha}{r_i q_i} \\ \hline \end{array} & \dots & \begin{array}{|c|} \hline \frac{\alpha}{r_i q_i} \\ \hline \end{array}
 \end{array}
 \end{array}
 \quad
 \begin{array}{c}
 \underbrace{\hspace{10em}}_{\Pi_{X_i}} \\
 \pi_1 \quad \pi_2 \quad \dots \quad \pi_{q_i} \\
 \left. \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_{r_i} \end{array} \right\} X_i \quad
 \begin{array}{ccc}
 \begin{array}{|c|c|} \hline \frac{\alpha}{r_i \tilde{q}_i} & 0 \\ \hline \end{array} & \dots & \begin{array}{|c|} \hline \frac{\alpha}{r_i \tilde{q}_i} \\ \hline \end{array} \\
 \begin{array}{|c|c|} \hline \frac{\alpha}{r_i \tilde{q}_i} & 0 \\ \hline \end{array} & \dots & \begin{array}{|c|} \hline \frac{\alpha}{r_i \tilde{q}_i} \\ \hline \end{array} \\
 \vdots & & \vdots \\
 \begin{array}{|c|c|} \hline \frac{\alpha}{r_i \tilde{q}_i} & 0 \\ \hline \end{array} & \dots & \begin{array}{|c|} \hline \frac{\alpha}{r_i \tilde{q}_i} \\ \hline \end{array}
 \end{array}
 \end{array}$$

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.

Some Notes on BDs

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

- BDeu is score-equivalent, meaning it takes the same value for DAGs that represent the same probability distribution. **BDs is not score-equivalent for finite samples** that have unobserved parents configurations. Asymptotically, $\text{BDs} \rightarrow \text{BDeu}$ as $n \rightarrow \infty$ if the positivity assumption holds.
- The $\tilde{\alpha}$ is a piece-wise uniform **empirical Bayes** prior because it depends on \mathcal{D} .
- We always have $\sum_{j:n_{ij}>0} \sum_k \tilde{\alpha} = \alpha$, so **the effective imaginary sample size is the same for all DAGs**. Therefore DAG comparisons are consistent with BDs, which is not the case with BDeu.

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

In our previous work [7], we explored the first- and second-order properties of U and we showed that

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

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.

Design of the Simulation Study

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

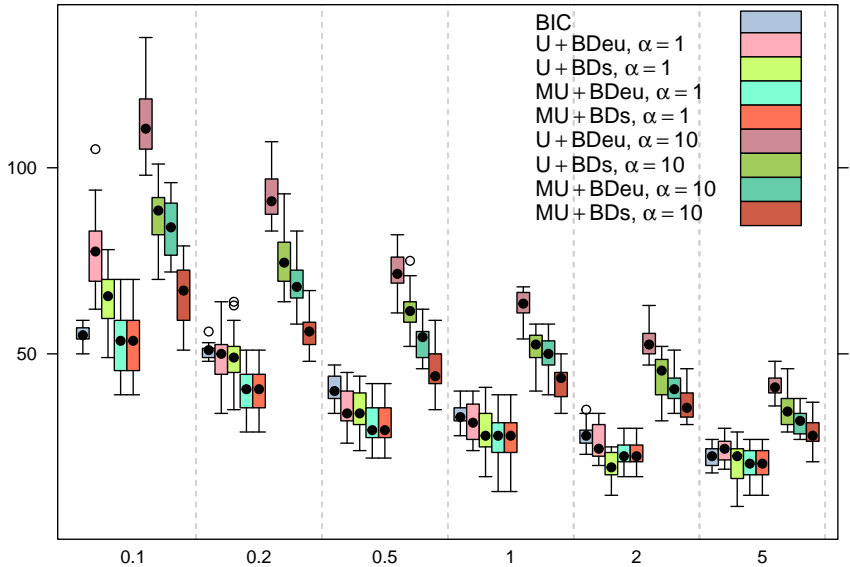
- **10 reference BNs** covering a wide range of N (8 to 442), $p = |\Theta|$ (18 to 77K) and number of arcs $|A|$ (8 to 602).
- **20 samples** of size $n/p = 0.1, 0.2, 0.5, 1.0, 2.0,$ and 5.0 (to allow for meaningful comparisons between BNs with such different N and p) **for each BN and n/p .**

with performance measures for:

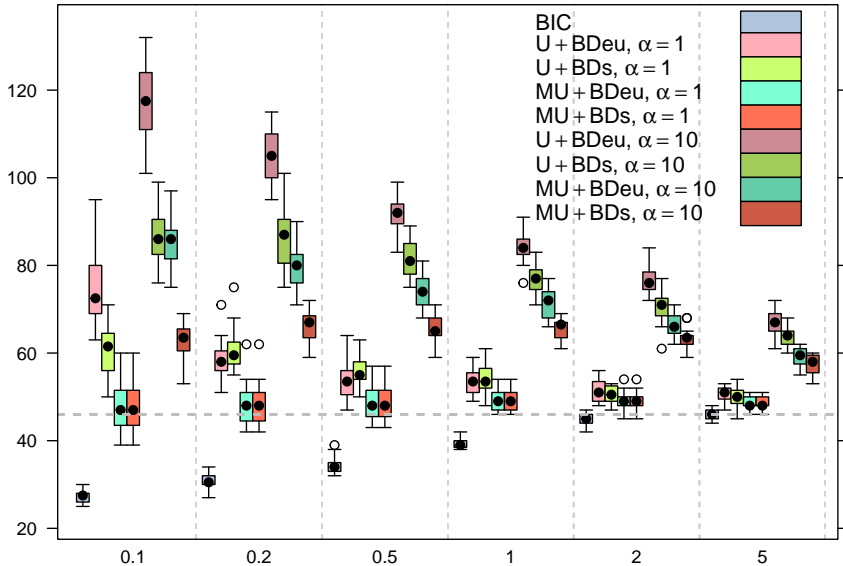
- the quality of the learned DAG using the **SHD distance** [11] from the reference BN;
- the **number of arcs** compared to the reference BN;
- the **log-likelihood on a separate test set** of size 10K, as an approximation of Kullback-Leibler distance.

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

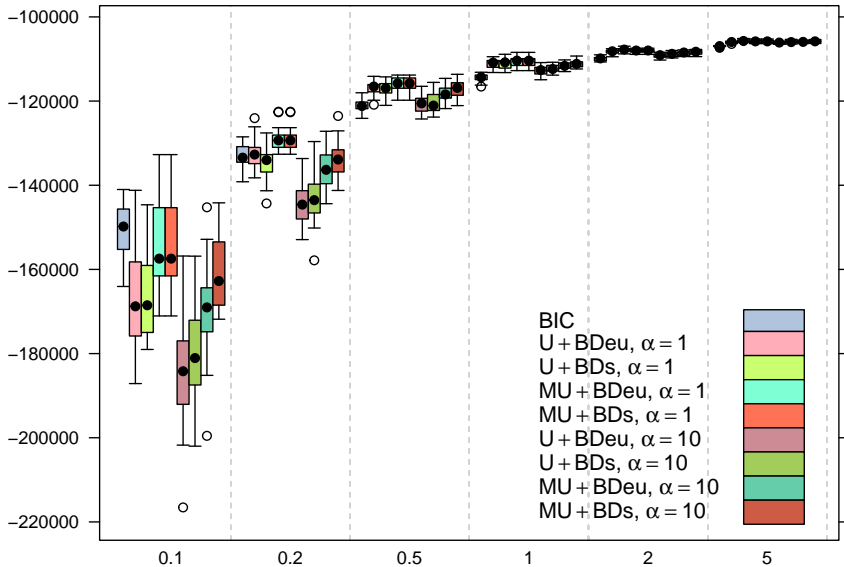
Results: ALARM, SHD



Results: ALARM, Number of Arcs



Results: ALARM, Log-likelihood on the Test Set



Conclusions

- We propose a new 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.