Dirichlet Bayesian Network Scores

and the Maximum Entropy Principle

Marco Scutari

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

September 21, 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 [6] 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 ($X_i \mid \Pi_{X_i} \sim Mult(\Theta_{X_i} \mid \Pi_{X_i})$ and $\Theta_{X_i} \mid \Pi_{X_i} \sim Dir(\alpha_{ijk})$, $\sum_{jk} \alpha_{ijk} = \alpha_i$ 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);

Heckerman *et al.* [4] 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 $\alpha_{ij} = \sum_k \alpha_{ijk}$.

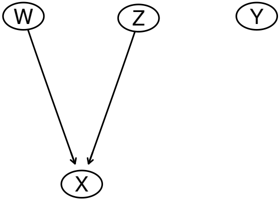
Bayesian Dirichlet Equivalent Uniform (BDeu)

The most common implementation of BD assumes $\alpha_{ijk} = \alpha/(r_i q_i)$, $\alpha = \alpha_i$ and is known from [4] as the Bayesian Dirichlet equivalent uniform (BDeu) marginal likelihood. However, there is evidence that assuming a flat prior over the parameters can be problematic:

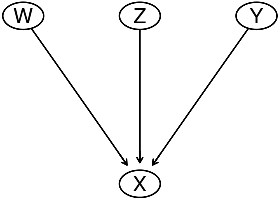
- The prior is actually not uninformative [5].
- MAP DAGs selected using BDeu are highly sensitive to the choice of α and can have markedly different number of arcs even for reasonable α [8].
- In the limits $\alpha \to 0$ and $\alpha \to \infty$ it is possible to obtain both very simple and very complex 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, 12].
- There are formal proofs of all this in [11, 12].

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 empirical 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 empirical 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:

$$\begin{aligned}
\mathrm{H}(X \mid \Pi_X; \alpha) &= 4\left[-\frac{0 + 1/8}{3 + 1/4}\log\frac{0 + 1/8}{3 + 1/4} - \frac{3 + 1/8}{3 + 1/4}\log\frac{3 + 1/8}{3 + 1/4}\right] \\
&= 0.652, \\
\mathrm{H}(X \mid \Pi_X \cup Y; \alpha) &= 4\left[-\frac{0 + 1/16}{3 + 1/8}\log\frac{0 + 1/16}{3 + 1/8} - \frac{3 + 1/16}{3 + 1/8}\log\frac{3 + 1/16}{3 + 1/8}\right] \\
&= 0.392.
\end{aligned}$$

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 $\mathcal{D}$. And then

$$\mathrm{BDeu}(X_i, \Pi_{X_i}; \alpha) =$$

$$\prod_{j:n_{ij}=0} \left[\cancel{\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ijk})}} \right] \prod_{j:n_{ij}>0} \left[\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right],$$

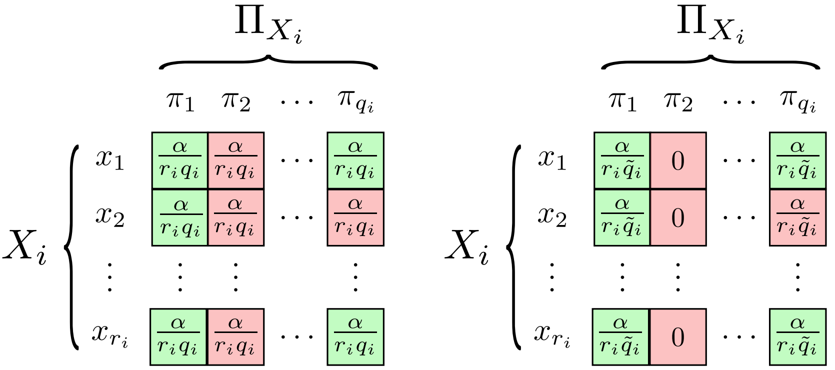
so the effective imaginary sample size decreases as the number of unobserved parents configurations increases. We can prevent that by replacing α_{ijk} with

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

and plugging 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 (X_i, Π_{X_i}) combinations that are not observed in the data are in red, observed combinations are in green.

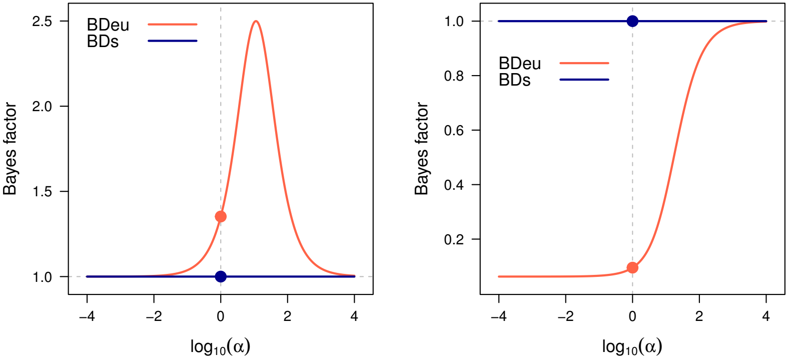
Exhibits A and B, Once More

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

Exhibit A: $\mathrm{BDs}(X \mid \Pi_X) = \mathrm{BDs}(X \mid \Pi_X \cup Y) = 3.9 \times 10^{-7}$,

Exhibit B: $\mathrm{BDs}(X \mid \Pi_X) = \mathrm{BDs}(X \mid \Pi_X \cup Y) = 0.032$.

It also avoids giving wildly different Bayes factors depending on the value of α.



This Left Me with a Few Questions...

The obvious one being:

1. The behaviour of BDeu is certainly undesirable, but it is it wrong?

Followed by:

2. Posterior entropy and BDeu rank $\mathcal{G}^-$ and $\mathcal{G}^+$ in the same order for Exhibit A, but they do not for Exhibit B. Why is that?

And the reason why I found that surprising is that:

3. Maximum (relative) entropy [7, 9, 1] represents a very general approach that includes Bayesian posterior estimation as a particular case [3]; it can also be seen as a particular case of MDL [2].

Hence, unless something is wrong with BDeu I would expect the two to agree. Especially because we can use MDL (using BIC), MAP (using BDeu/BDs),

Bayesian Statistics and Information Theory (I)

The derivation of Bayesian posterior as a particular case of maximum (relative) entropy is made clear in Giffin and Caticha [3]. The selected joint posterior $\mathrm{P}(X, \Theta)$ is that which maximises the relative entropy

$$S(\mathrm{P}, \mathrm{P}_{old}) = -\int \mathrm{P}(X, \Theta) \log \frac{\mathrm{P}(X, \Theta)}{\mathrm{P}_{old}(X, \Theta)} \, dX \, d\Theta.$$

The family of posteriors that reflects the fact that X is now known to take value x' is such that

$$\mathrm{P}(X) = \int \mathrm{P}(X = x', \Theta) \, d\Theta = \delta(X - x')$$

which amounts to a (possibly infinite) number of constraints on $\mathrm{P}(X, \Theta)$: for each possible value of X there is one constraint.

Bayesian Statistics and Information Theory (II)

Maximising $S(\mathrm{P}, \mathrm{P}_{old})$ subject to those constraints using Lagrange multipliers means solving

$$S(\mathrm{P}, \mathrm{P}_{old}) + \lambda_0 \underbrace{\left[\int \mathrm{P}\; dX\, d\Theta - 1\right]}_{\text{normalising constraint}} +$$

$$\int \lambda(x) \underbrace{\left[\int \mathrm{P}(X, \Theta) - \delta(X - x')\, d\Theta\right]}_{\text{constraint for each value of } X} dX$$

and yields the familiar Bayesian update rule:

$$\mathrm{P}_{new}(X, \Theta) = \frac{\mathrm{P}_{old}(X, \Theta)\delta(X - x')}{\mathrm{P}_{old}(X)} = \mathrm{P}_{old}(\Theta \mid X)\delta(X - x').$$

Bayesian Statistics and Information Theory (III)

In particular, the updated distribution for Θ is

$$\mathrm{P}_{new}(\Theta) = \int \mathrm{P}_{new}(X, \Theta)\, dX = \mathrm{P}_{old}(\Theta \mid X = x')$$

which means that the posterior distribution is that in which we only update those aspects of our beliefs for which corrective new evidence (in this case, the data) has been supplied. However, we use all the available information (as opposed to just what is in the empirical entropy):

- the information encoded in the distributional assumptions for the prior distribution over Θ;
- the information encoded in the distributional assumptions for the random variable X;
- the information encoded in the observed data.

Back to BNs: the Posterior Expected Entropy

Starting from the Markov property, for a BN we can write

$$\mathrm{H}^{\mathcal{G}}(\mathbf{X}; \Theta) = \sum_{i=1}^{N} \mathrm{H}^{\mathcal{G}}(X_i; \Theta_{X_i}).$$

where $\mathrm{H}^{\mathcal{G}}(X_i; \Theta_{X_i})$ is the entropy of X_i given its parents Π_{X_i} in $\mathcal{G}$.

The marginal posterior expectation of $\mathrm{H}^{\mathcal{G}}(X_i; \Theta_{X_i})$ with respect to Θ_{X_i} given the data can then be expressed as

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \mid \mathcal{D}\right) = \int \mathrm{H}^{\mathcal{G}}(X_i; \Theta_{X_i})\, \mathrm{P}(\Theta_{X_i} \mid \mathcal{D})\, d\Theta_{X_i}$$

where we use $\mathcal{D}$ to refer specifically to the observed values for X_i and $\Pi^{\mathcal{G}}_{X_i}$ with a slight abuse of notation.

Adding the Dirichlet Prior

We can then introduce a $Dirichlet(\alpha_{ijk})$ prior over Θ_{X_i} with

$$\mathrm{P}(\Theta_{X_i} \mid \mathcal{D}) = \int \mathrm{P}(\Theta_{X_i} \mid \mathcal{D}, \alpha_{ijk}) \, \mathrm{P}(\alpha_{ijk} \mid \mathcal{D}) \, d\alpha_{ijk},$$

which leads to

$$\begin{aligned} &\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \mid \mathcal{D}\right) \\ &\quad = \iint \mathrm{H}^{\mathcal{G}}(X_i; \Theta_{X_i}) \, \mathrm{P}(\Theta_{X_i} \mid \mathcal{D}, \alpha_{ijk}) \, \mathrm{P}(\alpha_{ijk} \mid \mathcal{D}) \, d\alpha_{ijk} \, d\Theta_{X_i} \\ &\quad \propto \int \mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right) \mathrm{P}(\mathcal{D} \mid \alpha_{ijk}) \, \mathrm{P}(\alpha_{ijk}) \, d\alpha_{ijk}, \end{aligned}$$

where $\mathrm{P}(\alpha_{ijk})$ is a hyper-prior distribution over the space of the Dirichlet priors, identified by their parameter sets $\{\alpha_{ijk}\}$.

Components of the Posterior Expected Entropy

$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right)$ is the posterior expected value of the entropy of $X_i \mid \Pi_{X_i}$ given α_{ijk}, and has closed form

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right) = \sum_{j=1}^{q_i}\left[\psi_0(\alpha_{ij} + n_{ij} + 1) - \sum_{k=1}^{r_i} \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \psi_0(\alpha_{ijk} + n_{ijk} + 1)\right].$$

$\mathrm{P}(\mathcal{D} \mid \alpha_{ijk})$ follows a Dirichlet-multinomial distribution, so

$$\mathrm{P}(\mathcal{D} \mid \alpha_{ijk}) = \left[\prod_{j=1}^{q_i} \frac{n_{ij}!}{\prod_{k=1}^{r_i} n_{iijk}!}\right] \cdot \left[\prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(n_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})}\right] \propto \mathrm{BD}(X_i \mid \Pi^{\mathcal{G}}_{X_i}; \alpha_{ijk})$$

making the link between BD scores and entropy explicit.

BDeu and the Maximum Entropy Principle

In the case of BDeu, $P(\alpha_{ijk} = \alpha/(r_i q_i)) = 1$ and learning DAGs based on sparse data following the maximum (relative) entropy means

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \mid \Pi_{X_i}^{\mathcal{G}^-}; \alpha_{ijk}\right) \leqslant \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \mid \Pi_{X_i}^{\mathcal{G}^+}; \alpha_{ijk}(\tilde{q}_i/q_i)\right)$$

whereas it should be

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \mid \Pi_{X_i}^{\mathcal{G}^-}; \alpha_{ijk}\right) \leqslant \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \mid \Pi_{X_i}^{\mathcal{G}^+}; \alpha_{ijk}\right)$$

so structure learning with BDeu may deviate from the maximum (relative) entropy principle when computed from sparse data. BDs does not.

Exhibit A, One Last Time

Exhibit B, One Last Time

Combining BDeu with $\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \mid \mathcal{D}, \alpha_{ijk}\right)$ gives

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(X) \mid \mathcal{D}\right) = 0.3931 \cdot 0.0326 = 0.0128 <$$
$$0.0252 = 0.5707 \cdot 0.0441 = \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(X) \mid \mathcal{D}\right)$$

while BDs gives

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(X) \mid \mathcal{D}\right) = 0.3931 \cdot 0.0326 = 0.0128 =$$
$$0.0128 = 0.3931 \cdot 0.0326 = \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(X) \mid \mathcal{D}\right).$$

Summary and Conclusions

- BDeu can be problematic for small/large values of the imaginary sample size; we found that BDeu can also be problematic regardless if the data are sparse.
- Then we proposed BDs as a minimalistic fix which prevents the imaginary sample size from partially vanishing when there are unobserved parent configurations.
- But is BDeu just from not working very well, or is it methodologically wrong to use it with sparse data? (Many statistical methods that are methodologically correct but do not work very well on sparse data.)
- One way of looking at this problem is in the context or maximum (relative) entropy. Given the same information in the prior, and the same information from the data, the assumptions behind BDeu can give a rank more complex, singular BNs over simpler ones.

References I

A. Caticha.
Relative entropy and inductive inference.
In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 75–96, 2004.

M. Feder.
Maximum Entropy as a Special Case of the Minimum Description Length Criterion.
IEEE Transactions on Information Theory, 32(6):847–849, 1986.

A. Giffin and A. Caticha.
Updating Probabilities with Data and Moments.
In *Proceedings of the 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 74–84, 2007.

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.

I. Nemenman, F. Shafee, and W. Bialek.
Entropy and Inference, Revisited.
In *Proceedings of the 14th Advances in Neural Information Processing Systems (NIPS) Conference*, pages 471–478, 2002.

References II

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

J. E. Shore and R. W. Johnson.
Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy.
IEEE Transactions on Information Theory, IT-26(1):26–37, 1980.

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.

J. Skilling.
The Axioms of Maximum Entropy.
In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, pages 173–187, 1988.

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

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.