Behaviormetrika
https://doi.org/10.1007/s41237-018-0048-x

INVITED PAPER

Dirichlet Bayesian network scores and the maximum relative entropy principle

Marco Scutari[1]

Received: 22 January 2018 / Accepted: 29 March 2018

Abstract A classic approach for learning Bayesian networks from data is to identify a *maximum a posteriori* (MAP) network structure. In the case of discrete Bayesian networks, MAP networks are selected by maximising one of several possible Bayesian–Dirichlet (BD) scores; the most famous is the *Bayesian–Dirichlet equivalent uniform* (BDeu) score from Heckerman et al. (Mach Learn 20(3):197–243, 1995). The key properties of BDeu arise from its uniform prior over the parameters of each local distribution in the network, which makes structure learning computationally efficient; it does not require the elicitation of prior knowledge from experts; and it satisfies score equivalence. In this paper we will review the derivation and the properties of BD scores, and of BDeu in particular, and we will link them to the corresponding entropy estimates to study them from an information theoretic perspective. To this end, we will work in the context of the foundational work of Giffin and Caticha (Proceedings of the 27th international workshop on Bayesian inference and maximum entropy methods in science and engineering, pp 74–84, 2007), who showed that Bayesian inference can be framed as a particular case of the maximum relative entropy principle. We will use this connection to show that BDeu should not be used for structure learning from sparse data, since it violates the maximum relative entropy principle; and that it is also problematic from a more classic Bayesian model selection perspective, because it produces Bayes factors that are sensitive to the value of its only hyperparameter. Using a large simulation study, we found in our previous work [Scutari in J Mach Learn Res (Proc Track PGM 2016) 52:438–448, 2016] that the Bayesian–Dirichlet sparse (BDs) score seems to provide better accuracy in structure learning; in this paper we further show that BDs does not suffer

Communicated by Joe Suzuki.

✉ Marco Scutari
scutari@stats.ox.ac.uk

[1] Department of Statistics, University of Oxford, 24-29 St. Giles', Oxford OX1 3LB, UK

Published online: 07 April 2018

from the issues above, and we recommend to use it for sparse data instead of BDeu. Finally, will show that these issues are in fact different aspects of the same problem and a consequence of the distributional assumptions of the prior.

Keywords Bayesian networks · Structure learning · Bayesian posterior estimation · Maximum relative entropy principle · Discrete data

1 Introduction and background

Bayesian networks (BNs; Pearl 1988; Koller and Friedman 2009) are probabilistic graphical models based on a directed acyclic graph (DAG) $\mathcal{G}$ whose nodes are associated with a set of random variables $\mathbf{X} = \{X_1, \ldots, X_N\}$ following some distribution $P(\mathbf{X})$. (The two are referred to interchangeably.) Formally, $\mathcal{G}$ is defined as an *independency map* of $P(\mathbf{X})$ such that:

$$\mathbf{X}_A \perp\!\!\!\perp_G \mathbf{X}_B | \mathbf{X}_C \Longrightarrow \mathbf{X}_A \perp\!\!\!\perp_P \mathbf{X}_B | \mathbf{X}_C,$$

where $\mathbf{X}_A$, $\mathbf{X}_B$ and $\mathbf{X}_C$ are disjoint subsets of $\mathbf{X}$. In other words, *graphical separation* (denoted $\perp\!\!\!\perp_G$, and called *d-separation* in this context) between two nodes in $\mathcal{G}$ implies the *conditional independence* (denoted $\perp\!\!\!\perp_P$) of the corresponding variables in $\mathbf{X}$. Two nodes linked by an arc cannot be graphically separated; hence the arcs of $\mathcal{G}$ represent *direct dependencies* between the variables they are incident on, as opposed to *indirect dependencies* that are mediated by one or more nodes in $\mathcal{G}$.

A consequence of this definition is the Markov property (Pearl 1988): in the absence of missing data the *global distribution* of $\mathbf{X}$ decomposes into:

$$P(\mathbf{X}|\mathcal{G}) = \prod_{i=1}^{N} P\left(X_i | \Pi_{X_i}^{\mathcal{G}}\right) \tag{1}$$

where the *local distribution* of each node X_i depends only on the values of its parents Π_{X_i} in $\mathcal{G}$ (denoted $\Pi_{X_i}^{\mathcal{G}}$). In this paper we will focus on discrete BNs (Heckerman et al. 1995), in which both $\mathbf{X}$ and the X_i are multinomial random variables; in particular:

$$X_i | \Pi_{X_i}^{\mathcal{G}} \sim \text{Multinomial}\left(\Theta_{X_i} | \Pi_{X_i}^{\mathcal{G}}\right)$$

where each parameter set Θ_{X_i} comprises the conditional probabilities:

$$\pi_{ik|j} = P\left(X_i = k | \Pi_{X_i}^{\mathcal{G}} = j\right)$$

of each value k of X_i given each possible configuration of the values of $\Pi_{X_i}^{\mathcal{G}}$. Other possibilities include Gaussian BNs (Geiger and Heckerman 1994) and conditional linear Gaussian BNs (Lauritzen and Wermuth 1989). In Gaussian BNs, $\mathbf{X}$ is multivariate normal and the (conditional) dependencies between the X_i are assumed to be linear, leading to

$$X_i | \Pi_{X_i}^{\mathcal{G}} \sim N\left(\Theta_{X_i} | \Pi_{X_i}^{\mathcal{G}}\right)$$

which can be written as a linear regression model of the form

$$X_i = \mu_{X_i} + \Pi_{X_i}^{\mathcal{G}} \beta_{X_i} + \varepsilon, \quad \varepsilon \sim N(0, \sigma_{X_i}^2)$$

or using the partial correlations between X_i and each parent given the rest; the two parameterisations are equivalent (Weatherburn 1961). Conditional linear Gaussian BNs combine multinomial and normal variables using mixture of normals, with discrete variables identifying the components of the mixture.

It is important to note that the decomposition in (1) does not uniquely identify a BN; different DAGs can encode the same global distribution, thus grouping BNs into equivalence classes (Chickering 1995) characterised by the skeleton of $\mathcal{G}$ (its underlying undirected graph) and its v-structures (patterns of arcs of the type $X_j \rightarrow X_i \leftarrow X_k$, with no arc between X_j and X_k). Intuitively, the direction of arcs that are not part of v-structures can be reversed without changing the global distribution, just factorising it in different ways, as long as the new arc directions do not introduce additional v-structures or cycles in the DAG. As a simple example, consider,

$$\begin{aligned} \underbrace{P(X_i)P(X_j|X_i)P(X_k|X_j)}_{X_i \rightarrow X_j \rightarrow X_k} &= P(X_j, X_i)P(X_k|X_j) \\ &= \underbrace{P(X_i|X_j)P(X_j)P(X_k|X_j)}_{X_i \leftarrow X_j \rightarrow X_k}. \end{aligned}$$

The task of specifying BNs is called *learning* and can be performed using either data, prior expert knowledge on the phenomenon being modelled or both. The latter has been shown to provide very good results in a variety of applications, and should be preferred if it is feasible to elicit prior information from experts (Castelo and Siebes 2000; Mukherjee and Speed 2008). Learning BNs from data is usually performed in an inherently Bayesian fashion by maximising,

$$\underbrace{P(\mathcal{B}|\mathcal{D}) = P(\mathcal{G}, \Theta|\mathcal{D})}_{\text{learning}} = \underbrace{P(\mathcal{G}|\mathcal{D})}_{\text{structure learning}} \cdot \underbrace{P(\Theta|\mathcal{G}, \mathcal{D})}_{\text{parameter learning}}, \tag{2}$$

where $\mathcal{D}$ is a sample from $\mathbf{X}$ and $\mathcal{B} = (\mathcal{G}, \Theta)$ is a BN with DAG $\mathcal{G}$ and parameter set $\Theta = \bigcup_{i=1}^{N} \Theta_{X_i}$. Structure learning consists in finding the DAG $\mathcal{G}$ that encodes the dependence structure of the data; parameter learning involves the estimation of the parameters Θ given $\mathcal{G}$. Expert knowledge can be incorporated in either or both these steps through the use of informative priors for $\mathcal{G}$ or Θ.

Structure learning can be implemented in several ways, based on many results from probability, information and optimisation theory; algorithms for this task can be broadly grouped into constraint-based, score-based and hybrid.

Constraint-based algorithms (Aliferis et al. 2010a, b) use statistical tests to learn conditional independence relationships from the data and to determine if the corresponding arcs should be included in $\mathcal{G}$. In order to do that they assume that $\mathcal{G}$ is *faithful* to $P(\mathbf{X})$, meaning

$$\mathbf{X}_A \perp\!\!\!\perp_G \mathbf{X}_B | \mathbf{X}_C \Longleftrightarrow \mathbf{X}_A \perp\!\!\!\perp_P \mathbf{X}_B | \mathbf{X}_C;$$

this is a strong assumption that does not hold in a number of real-world scenarios, which are reviewed in Koski and Noble (2012). Depending on the nature of the data, conditional independence tests in common use are the mutual information (G^2) and Pearson's χ^2 tests for contingency tables (for discrete BNs); and Fisher's Z and the exact t tests for partial correlations (for Gaussian BNs); an overview is provided in Scutari and Denis (2014).

Score-based algorithms are closer to model selection techniques developed in classic statistics and information theory. Each candidate network is assigned a score reflecting its goodness of fit, which is then taken as an objective function to maximise. This is often done using heuristic optimisation algorithms, from local search to genetic algorithms (Russell and Norvig 2009); but the availability of computational resources and advances in learning algorithms have recently made exact learning possible (Cussens 2012). Common choices for the network score include the Bayesian Information Criterion (BIC) and the marginal likelihood $P(\mathcal{G}|\mathcal{D})$ itself; for an overview see again Scutari and Denis (2014). We will cover both in more detail for discrete BNs in Sect. 2.

Hybrid algorithms use both statistical tests and score functions, combining the previous two families of algorithms. Their general formulation is described for BNs in Friedman et al. (1999); they have proved to be some of the top performers up to date (see for instance MMHC inTsamardinos et al. 2006).

As for parameter learning, the parameters Θ_{X_i} can be estimated independently for each node following (1) since its parents are assumed to be known from structure learning. Both maximum likelihood and Bayesian posterior estimators are in common use, with the latter being preferred due to their smoothness and superior predictive power (Koller and Friedman 2009).

In this paper we focus on score-based structure learning in a Bayesian framework, in which we aim to identify a *maximum a posteriori* (MAP) DAG $\mathcal{G}$ that directly maximises $P(\mathcal{G}|\mathcal{D})$. For discrete BNs, this means maximising a Bayesian–Dirichlet (BD) marginal likelihood: the most common choice is the *Bayesian–Dirichlet equivalent uniform* (BDeu) score from Heckerman et al. (1995). We will show that the uniform prior distribution over each Θ_{X_i} that underlies BDeu can be problematic from a Bayesian perspective, resulting in wildly different Bayes factors (and thus structure learning

outcomes) depending on the value of its only hyperparameter, the imaginary sample size. We will further investigate this problem from an information theoretic perspective, on the grounds that Bayesian posterior inference can be framed as a particular case of the maximum relative entropy principle (ME; Shore and Johnson 1980; Skilling 1988; Caticha 2004). We find that BDeu is not a reliable network score when applied to sparse data because it can select overly complex networks over simpler ones given the same information in the prior and in the data; and that in the process it violates the maximum relative entropy principle. That does not appear to be the case for other BD scores, which arise from different priors.

The paper is organised as follows: In Sect. 2 we will review Bayesian score-based structure learning and BD scores. In Sect. 3 we will focus on BDeu, covering its underlying assumptions and issues reported in the literature. In particular, we will show with simple examples that BDeu can produce Bayes factors that are sensitive to the choice of its only hyperparameter, the imaginary sample size. In Sect. 4 we will derive the posterior expected entropy associated with a DAG $\mathcal{G}$, which we will further explore in Sect. 5. Finally, in Sect. 6 we will analyse BDeu using ME, and we will compare its behaviour with that of other BD scores.

2 Bayesian–Dirichlet marginal likelihoods

Score-based structure learning, when performed in a Bayesian framework, aims at finding a DAG $\mathcal{G}$ that has the highest posterior probability $P(\mathcal{G}|\mathcal{D})$. Starting from (2), using Bayes' theorem we can write,

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

where $P(\mathcal{G})$ is the prior distribution over the space of the DAGs spanning the variables in $\mathbf{X}$ and $P(\mathcal{D}|\mathcal{G})$ is the marginal likelihood of the data given $\mathcal{G}$ averaged over all possible parameter sets Θ. $P(\mathcal{G})$ is often taken to be uniform so that it simplifies when comparing DAGs; we will do the same in this paper for simplicity while noting that other default priors may lead to more accurate structure learning of sparse DAGs (e.g. Scutari 2016). Using (1) we can then decompose $P(\mathcal{D}|\mathcal{G})$ into one component for each node as follows:

$$P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{N} P\left(X_i|\Pi_{X_i}^{\mathcal{G}}\right) = \prod_{i=1}^{N} \left[\int P\left(X_i|\Pi_{X_i}^{\mathcal{G}}, \Theta_{X_i}\right) P\left(\Theta_{X_i}|\Pi_{X_i}^{\mathcal{G}}\right) \mathrm{d}\Theta_{X_i}\right]. \quad (3)$$

In the case of discrete BNs, we assume $X_i|\Pi_{X_i}^{\mathcal{G}} \sim \text{Multinomial}(\Theta_{X_i}|\Pi_{X_i}^{\mathcal{G}})$ where the parameters $\Theta_{X_i}|\Pi_{X_i}^{\mathcal{G}}$ are the conditional probabilities $\pi_{ik|j} = P(X_i = k|\Pi_{X_i}^{\mathcal{G}} = j)$. We then assume a conjugate prior $\Theta_{X_i}|\Pi_{X_i}^{\mathcal{G}} \sim \text{Dirichlet}(\alpha_{ijk})$, $\sum_{jk} \alpha_{ijk} = \alpha_i > 0$ to obtain the closed-form posterior $\text{Dirichlet}(\alpha_{ijk} + n_{ijk})$ which we use to estimate the $\pi_{ik|j}$ from the counts n_{ijk}, $\sum_{ijk} n_{ijk} = n$ observed in $\mathcal{D}$. α_i is known as the *imaginary* or

equivalent sample size and determines how much weight is assigned to the prior in terms of the size of an imaginary sample supporting it.

Further assuming *positivity* ($\pi_{ik|j} > 0$), *parameter independence* ($\pi_{ik|j}$ for different parent configurations are independent), *parameter modularity* ($\pi_{ik|j}$ associated with different nodes are independent) and *complete data*, Heckerman et al. (1995) derived a closed form expression for (3), known as the *Bayesian–Dirichlet* (BD) family of scores:

$$\begin{aligned} \mathrm{BD}(\mathcal{G}, \mathcal{D}; \boldsymbol{\alpha}) &= \prod_{i=1}^{N} \mathrm{BD}\left(X_i | \Pi_{X_i}^{\mathcal{G}}; \alpha_i\right) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \left[\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right] \end{aligned} \tag{4}$$

where:

- r_i is the number of states of X_i
- q_i is the number of possible configurations of values of $\Pi_{X_i}^{\mathcal{G}}$, taken to be equal to 1 if X_i has no parents;
- $n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$;
- $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$;
- and $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_N\}$, $\alpha_i = \sum_{j=1}^{q_i} \alpha_{ij}$ are the imaginary sample sizes associated with each X_i.

Various choices for α_{ijk} produce different priors and the corresponding scores in the BD family:

- for $\alpha_{ijk} = 1$ we obtain the K2 score from Cooper and Herskovits (1991);
- for $\alpha_{ijk} = 1/2$ we obtain the BD score with Jeffrey's prior (BDJ; Suzuki 2016);
- for $\alpha_{ijk} = \alpha/(r_i q_i)$ we obtain the BDeu score from Heckerman et al. (1995), which is the most common choice in the BD family and has $\alpha_i = \alpha$ for all X_i;
- for $\alpha_{ijk} = \alpha/(r_i \tilde{q}_i)$, where $\tilde{q}_i$ is the number of $\Pi_{X_i}^{\mathcal{G}}$ such that $n_{ij} > 0$, we obtain the BD sparse (BDs) score recently proposed in Scutari (2016);
- for the set $\alpha_{ijk}^s = s/(r_i q_i)$, $s \in S_L = \{2^{-L}, 2^{-L+1}, \ldots, 2^{L-1}, 2^L\}$, $L \in \mathbb{N}$ we obtain the locally averaged BD score (BDla) from Cano et al. (2013).

BDeu is the only score-equivalent BD score (Chickering 1995), that is, it is the only BD score that takes the same value for DAGs in the same equivalence class. This property makes BDeu the preferred score when arcs are given causal interpretation (Pearl 2009), and their directions have a meaningful interpretation beyond allowing to decompose the $P(\mathbf{X})$ into the $P(X_i | \Pi_{X_i}^{\mathcal{G}})$. BDs is only asymptotically score-equivalent because

it converges to BDeu when $n \to \infty$ and the positivity assumption holds. The BIC score, defined as:

$$\mathrm{BIC}(\mathcal{G}, \mathcal{D}) = \sum_{i=1}^{N} \mathrm{BIC}\left(X_i | \Pi_{X_i}^{\mathcal{G}}\right) = \sum_{i=1}^{N} \left[\log P\left(X_i | \Pi_{X_i}^{\mathcal{G}}\right) - \frac{\log n}{2} q_i(r_i - 1)\right] \quad (5)$$

is also score-equivalent and it converges to log BDeu as $n \to \infty$. In the case of discrete BNs, maximising BIC corresponds to selecting the BN with the *minimum description length* (MDL; Rissanen 1978).

3 BDeu and Bayesian model selection

The uniform prior associated with BDeu has been justified by the lack of prior knowledge on the Θ_{X_i}, as well as its computational simplicity and score equivalence; and it was widely assumed to be non-informative (e.g. Silander et al. 2007; Heckerman et al. 1995).

However, there is increasing evidence that this prior is far from non-informative and that it has a strong impact on the accuracy of the learned DAGs, making its use on real-world data problematic. Silander et al. (2007) showed via simulation that the MAP DAGs selected using BDeu are highly sensitive to the choice of α. Even for "reasonable" values such as $\alpha \in [1, 20]$, they obtained DAGs with markedly different number of arcs, and they showed that large values of α tend to produce DAGs with more arcs. This is counter-intuitive because a larger α would normally imply stronger regularisation and would be expected to produce sparser DAGs. Steck and Jaakkola (2003) similarly showed that the number of arcs in the MAP DAG is determined by a complex interaction between α and $\mathcal{D}$; in the limits $\alpha \to 0$ and $\alpha \to \infty$ it is possible to obtain both very sparse and very dense DAGs. (We informally define $\mathcal{G}$ to be *sparse* if $|A| = O(N)$, typically with $|A| < 5N$; a *dense* $\mathcal{G}$, on the other hand, has a relatively large $|A|$ compared to N.) In particular, for small values of α and/or sparse data (that is, discrete data for which we observe a small subset of the possible combinations of the values of the X_i), $\alpha_{ijk} \to 0$ and

$$\lim_{\alpha_{ijk} \to 0} \mathrm{BDeu}(\mathcal{G}, \mathcal{D}; \alpha) - \alpha^{d_{\mathrm{EP}}^{(\mathcal{G})}} = 0 \quad (6)$$

where $d_{\mathrm{EP}}^{(\mathcal{G})}$ is the effective number of parameters of the model, defined as

$$d_{\mathrm{EP}}^{(\mathcal{G})} = \sum_{i=i}^{N} d_{\mathrm{EP}}^{(X_i, \mathcal{G})} = \sum_{i=i}^{N} \left[\sum_{j=1}^{q_i} \tilde{r}_{ij} - \tilde{q}_i\right];$$

$\tilde{r}_{ij}$ is the number of positive counts n_{ijk} in the jth configuration of $\Pi_{X_i}^{\mathcal{G}}$ and $\tilde{q}_i$ is the number of configurations in which at least one n_{ijk} is positive.

This was then used to prove that the Bayes factor

$$\frac{P(\mathcal{D}|\mathcal{G}^+)}{P(\mathcal{D}|\mathcal{G}^-)} = \frac{\mathrm{BDeu}(X_i|\Pi_{X_i}^{\mathcal{G}^+};\alpha)}{\mathrm{BDeu}(X_i|\Pi_{X_i}^{\mathcal{G}^-};\alpha)} \to \begin{cases} 0 & \text{if } d_{\mathrm{EDF}} > 0 \\ +\infty & \text{if } d_{\mathrm{EDF}} < 0 \end{cases} \tag{7}$$

for two DAGs $\mathcal{G}^-$ and $\mathcal{G}^+$ that differ only by the inclusion of a single parent for X_i. The effective degrees of freedom d_{EDF} are defined as $d_{\mathrm{EP}}^{(\mathcal{G}^+)} - d_{\mathrm{EP}}^{(\mathcal{G}^-)}$. The practical implication of this result is that, when we compare two DAGs using their BDeu scores, a large number of zero counts will force d_{EDF} to be negative and favour the inclusion of additional arcs (since $\mathrm{BDeu}(X_i|\Pi_{X_i}^{\mathcal{G}^+};\alpha) \gg \mathrm{BDeu}(X_i|\Pi_{X_i}^{\mathcal{G}^-};\alpha)$). But that in turn makes d_{EDF} even more negative, quickly leading to overfitting. Furthermore, Steck and Jaakkola (2003) argued that BDeu can be rather unstable for "medium-sized" data and small α, which is a very common scenario.

Steck (2008) approached the problem from a different perspective and derived an analytic approximation for the "optimal" value of α that maximises predictive accuracy, further suggesting that the interplay between α and $\mathcal{D}$ is controlled by the skewness of the Θ_{X_i} and by the strength of the dependence relationships between the nodes. Skewed Θ_{X_i} result in some $\pi_{ik|j}$ being smaller than others, which in turn makes sparse data sets more likely; hence the problematic behaviour described in Steck and Jaakkola (2003) and reported above. Most of these results have been analytically confirmed more recently by Ueno (2010, 2011). The key insight provided by the latter paper is that we can decompose the BDeu into a *likelihood term* that depends on the data and a *prior term* that does not:

$$\log \mathrm{BDeu}(X_i|\Pi_{X_i}^{\mathcal{G}};\alpha) = \underbrace{\sum_{j=1}^{q_i}\left(\log\Gamma(\alpha_{ij}) - \sum_{k=1}^{r_i}\log\Gamma(\alpha_{ijk})\right)}_{\text{prior term}} + \underbrace{\sum_{j=1}^{q_i}\left(\sum_{k=1}^{r_i}\log\Gamma(\alpha_{ijk}+n_{ijk}) - \log\Gamma(\alpha_{ij}+n_{ij})\right)}_{\text{likelihood term}}.$$

Then if $\alpha_{ijk} < 1$ (that is, $\alpha < r_i q_i$) the prior term can be approximated by

$$\sum_{j=1}^{q_i}\left(\log\Gamma(\alpha_{ij}) - \sum_{k=1}^{r_i}\log\Gamma(\alpha_{ijk})\right) \approx q_i(r_i - 1)\log\alpha_{ijk}$$

and quickly dominates the likelihood term, penalising complex BNs as the number of parameters increases, which explains why BDeu selects empty DAGs in the limit of $\alpha_{ijk} \to 0$. On the other hand, if all $\alpha_{ijk} > 1$ then the prior term can be approximated by

$$\sum_{j=1}^{q_i} \left(\log \Gamma(\alpha_{ij}) - \sum_{k=1}^{r_i} \log \Gamma(\alpha_{ijk}) \right) \approx \alpha \log r_i + \frac{1}{2} q_i (r_i - 1) \log \frac{\alpha_{ijk}}{2\pi},$$

leading BDeu to select a complete DAG when $\alpha_{ijk} \to \infty$ (and therefore $\alpha \to \infty$) as previously reported in Silander et al. (2007).

As for the likelihood term, Ueno (2011) notes that if $\alpha + n$ is sufficiently small, that is, for sparse samples and small imaginary sample sizes,

$$\sum_{j=1}^{q_i} \left(\sum_{k=1}^{r_i} \log \Gamma(\alpha_{ijk} + n_{ijk}) - \log \Gamma(\alpha_{ij} + n_{ij}) \right) \approx -q_i (r_i - 1) \log \alpha_{ijk}.$$

Hence if some $n_{ijk} = 0$, the change of the likelihood term dominates the prior term and BDeu adds extra arcs, leading to dense DAGs. On the other hand, if $\alpha + n$ is sufficiently large, α actually acts as an imaginary sample supporting the uniform distribution of the parameters assumed in the prior. This explains the observations in Steck (2008): the optimal α should be large when the empirical distribution of the Θ_{X_i} is uniform because the prior is correct; and it should be small when the empirical distribution of Θ_{X_i} is skewed so that the prior can be quickly dominated. This is also the source of the sensitivity of BDeu to the choice of α reported in Steck and Jaakkola (2003).

Finally, Suzuki (2016) studied the asymptotic properties of BDeu by contrasting it with BDJ. He found that BDeu is not *regular* in the sense that it may learn DAGs in a way that is not consistent with either the MDL principle (through BIC) or the ranking of those DAGs given by their entropy. Whether this happens depends on the values of the underlying $\pi_{ik|j}$, even if the positivity assumption holds and if n is large. This agrees with the observations in Ueno (2010), who also observed that BDeu is not necessarily consistent for any finite n, but only asymptotically for $n \to \infty$.

Around the same time, a possible solution to these problems was proposed by Scutari (2016) in the form of BDs. Scutari (2016) starts from the consideration that if the sample $\mathcal{D}$ is sparse, some configurations of the variables will not be observed; it may be that the sample size is small and those configurations have low probability, or it may be that $\mathbf{X}$ violates the positivity assumption ($\pi_{ik|j} = 0$ for some i, j, k). As a result, we may be unable to observe all the configurations of (say) $\Pi_{X_i}^{\mathcal{G}^+}$ in the data. Then the corresponding $n_{ij} = 0$ and

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

The effective imaginary sample size, defined as the sum of the α_{ijk} appearing in terms that do not simplify (and thus contribute to the value of BDeu), decreases to $\sum_{j:n_{ij}>0} \alpha_{ijk} = \alpha(\tilde{q}_i/q_i) < \alpha$, where $\tilde{q}_i < q_i$ is the number of parent configurations that are actually observed in $\mathcal{D}$. In other words, BDeu is computed with an imaginary sample size of $\alpha(\tilde{q}_i/q_i)$ instead of α, and the remaining $\alpha(q_i - \tilde{q}_i)/q_i$ is effectively lost. This may lead to comparing DAGs with marginal likelihoods computed from different priors, which is incorrect from a Bayesian perspective. In order to prevent this from happening, Scutari (2016) replaced the prior of BDeu with

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

thus obtaining

$$\mathrm{BDs}(X_i|\Pi_{X_i}^{\mathcal{G}^+};\alpha) = \prod_{j:n_{ij}>0} \left[\frac{\Gamma(r_i\alpha_{ijk})}{\Gamma(r_i\alpha_{ijk}+n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})} \right]. \tag{8}$$

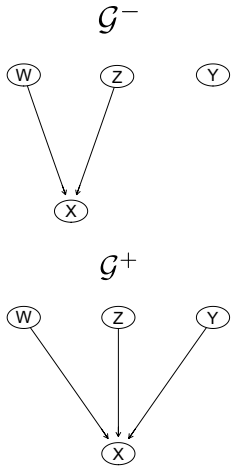
A large simulation study showed BDs to be more accurate than BDeu in learning BN structures without any loss in predictive power.

In addition to all these issues, we also find that BDeu produces Bayes factors that are sensitive to the choice of α. (The fact that BDeu is sensitive to the value of α does not necessarily imply that the Bayes factor is sensitive itself.) In order to illustrate this instability and the other results presented in the section we consider the simple examples below.

Example 1 Consider the DAGs $\mathcal{G}^-$ and $\mathcal{G}^+$ and the data set $\mathcal{D}_1$ in Fig. 1, originally from Suzuki (2016). The sample frequencies n_{ijk} for $X|\Pi_X^{\mathcal{G}^-} = \{Z, W\}$ 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 $X|\Pi_X^{\mathcal{G}^+} = X|\Pi_X^{\mathcal{G}^-} \cup \{Y\}$ are as follows.



$\mathcal{D}_1$

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

$\mathcal{D}_2$

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

Fig. 1 DAGs and data sets used in Examples 1 and 2. The DAGs $\mathcal{G}^-$ and $\mathcal{G}^+$ are used in both examples. Example 1 uses data set $\mathcal{D}_1$, while Example 2 uses $\mathcal{D}_2$; the latter is a modified version of the former, which is originally from Suzuki (2016)

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 distributions of $X|\Pi_X^{\mathcal{G}^-}$ and $X|\Pi_X^{\mathcal{G}^+}$ are both singular, and in $X|\Pi_X^{\mathcal{G}^+}$ we only observe 4 parent configurations out of 8. Furthermore, the observed conditional distributions for those parent configurations are identical to the 4 conditional distributions in $X|\Pi_X^{\mathcal{G}^-}$, since the n_{ijk} are the same. We can then argue that $X|\Pi_X^{\mathcal{G}^+}$ does not fit $\mathcal{D}_1$ any better than $X|\Pi_X^{\mathcal{G}^-}$, and it does not capture any additional information from the data.

However, if we take $\alpha = 1$ in BDeu, then $\alpha_{ijk} = 1/8$ for $\mathcal{G}^-$ and $\alpha_{ijk} = 1/16$ for $\mathcal{G}^+$, leading to

$$\mathrm{BDeu}\left(X \mid \Pi_X^{\mathcal{G}^-}; 1\right) = \left(\frac{\Gamma(1/4)}{\Gamma(1/4+3)}\left[\frac{\Gamma(1/8+3)}{\Gamma(1/8)} \cdot \cancel{\frac{\Gamma(1/8)}{\Gamma(1/8)}}\right]\right)^4 = 0.0326,$$

$$\mathrm{BDeu}\left(X \mid \Pi_X^{\mathcal{G}^+}; 1\right) = \left(\frac{\Gamma(1/8)}{\Gamma(1/8+3)}\left[\frac{\Gamma(1/16+3)}{\Gamma(1/16)} \cdot \cancel{\frac{\Gamma(1/16)}{\Gamma(1/16)}}\right]\right)^4 = 0.0441.$$

If we choose the DAG with the highest BDeu score, we prefer $\mathcal{G}^+$ to $\mathcal{G}^-$ despite all the considerations we have just made on the data. This is not the case if we use BDs, which does not show a preference for either $\mathcal{G}^-$ or $\mathcal{G}^+$ because $\alpha_{ijk} = 1/8$ for both $X|\Pi_X^{\mathcal{G}^-}$ and $X|\Pi_X^{\mathcal{G}^+}$:

$$\mathrm{BDs}\left(X \mid \Pi_X^{\mathcal{G}^-}; 1\right) = \mathrm{BDs}\left(X \mid \Pi_X^{\mathcal{G}^+}; 1\right)$$
$$= \left(\frac{\Gamma(1/4)}{\Gamma(1/4+3)}\left[\frac{\Gamma(1/8+3)}{\Gamma(1/8)} \cdot \cancel{\frac{\Gamma(1/8)}{\Gamma(1/8)}}\right]\right)^4 = 0.0326.$$

The same holds for BDJ, and in general for any BD score with a constant α_{ijk}. Comparing the expressions above, it is apparent that the only difference between them is the value of α_{ijk}, which is a consequence of the different number of configurations of $\Pi_X^{\mathcal{G}^-}$ and $\Pi_X^{\mathcal{G}^+}$.

The Bayes factors for BDeu and BDs are shown for $\alpha \in [10^{-4}, 10^4]$ in the left panel of Fig. 2. The former converges to 1 for both $\alpha_{ijk} \to 0$ and $\alpha_{ijk} \to \infty$, but varies between 1 and 2.5 for finite α; whereas the latter is equal to 1 for all values of α, never showing a preference for either $\mathcal{G}^-$ or $\mathcal{G}^+$. The Bayes factor for BDeu does not diverge nor converge to zero, which is consistent with (7) from Steck and Jaakkola (2003) as $d_{\mathrm{EP}}^{(\mathcal{G}^+)} - d_{\mathrm{EP}}^{(\mathcal{G}^-)} = 0 - 0 = 0$. However, it varies most quickly for $\alpha \in [1, 10]$, exactly the range of the most common values used in practical applications. This provides further evidence supporting the conclusions of Steck and Jaakkola (2003), Steck (2008) and Silander et al. (2007).

Finally, if we consider which DAG would be preferred according to the MDL principle, we can see that BIC (unlike BDeu, like BDs) does not express a preference for either DAG:

$$\mathrm{BIC}\left(X|\Pi_X^{\mathcal{G}^-}\right) = \log P\left(X|\Pi_X^{\mathcal{G}^-}\right) - 0 = \log P\left(X|\Pi_X^{\mathcal{G}^+}\right) - 0 = \mathrm{BIC}\left(X|\Pi_X^{\mathcal{G}^+}\right)$$

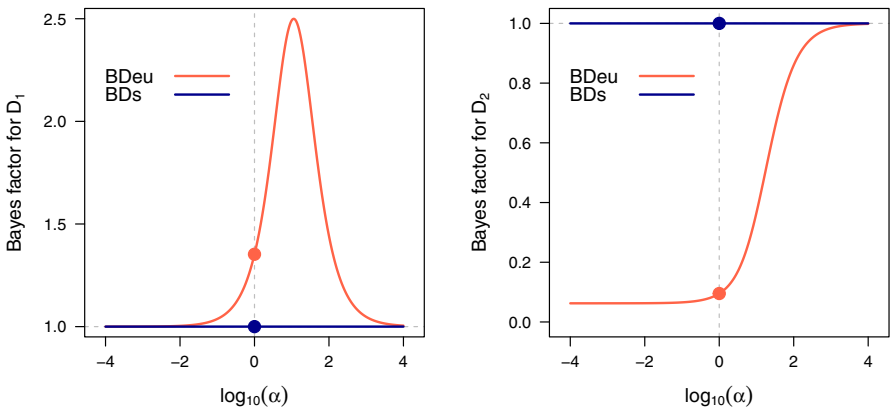


Fig. 2 The Bayes factors for $\mathcal{G}^-$ versus $\mathcal{G}^+$ computed using BDeu and BDs for Example 1 (left panel) and Example 2 (right panel in orange) and dark blue, respectively. The bullet points correspond to the values observed for $\alpha = 1$

which agrees with Suzuki (2016)'s observation that BDeu violates the MDL principle. □

Example 2 Consider another simple example, inspired by Example 1, based on the data set $\mathcal{D}_2$ and the DAGs $\mathcal{G}^-$, $\mathcal{G}^+$ shown in Fig. 1. The sample frequencies (n_{ijk}) for $X|\Pi_X^{\mathcal{G}^-}$ 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 $X|\Pi_X^{\mathcal{G}^+}$ 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

As in Example 1, 4 parent configurations out of 8 are not observed in $\mathcal{G}^+$ and the other 4 have n_{ijk} that are the same as those arising from $\mathcal{G}^-$. The resulting conditional distributions, however, are not singular. If we take again $\alpha = 1$, the BDeu scores for $\mathcal{G}^-$ and $\mathcal{G}^+$ are different but this time $\mathcal{G}^-$ has the highest score:

$$\mathrm{BDeu}\left(X|\Pi_X^{\mathcal{G}^-};1\right) = \left(\frac{\Gamma(1/4)}{\Gamma(1/4+3)}\left[\frac{\Gamma(1/8+2)}{\Gamma(1/8)}\cdot\frac{\Gamma(1/8+1)}{\Gamma(1/8)}\right]\right)^4 = 3.906\times 10^{-7},$$
$$\mathrm{BDeu}\left(X|\Pi_X^{\mathcal{G}^+};1\right) = \left(\frac{\Gamma(1/8)}{\Gamma(1/8+3)}\left[\frac{\Gamma(1/16+2)}{\Gamma(1/16)}\cdot\frac{\Gamma(1/16+1)}{\Gamma(1/16)}\right]\right)^4 = 3.721\times 10^{-8}.$$

On the other hand, in BDs $\alpha_{ijk} = 1/8$ for both DAGs, so they have the same score:

$$\begin{aligned}\mathrm{BDs}\left(X \mid \Pi_X^{\mathcal{G}^-};1\right) &= \mathrm{BDs}\left(X \mid \Pi_X^{\mathcal{G}^+};1\right)\\ &= \left(\frac{\Gamma(1/4)}{\Gamma(1/4+3)}\left[\frac{\Gamma(1/8+3)}{\Gamma(1/8)}\cdot\cancel{\frac{\Gamma(1/8)}{\Gamma(1/8)}}\right]\right)^4 = 3.906\times 10^{-7}.\end{aligned}$$

BDeu once more assigns different scores to $\mathcal{G}^-$ and $\mathcal{G}^+$ despite the fact that the observed conditional distributions in $X|\Pi_X^{\mathcal{G}^-}$ and $X|\Pi_X^{\mathcal{G}^+}$ are the same, while BDs does not.

The Bayes factors for BDeu and BDs are shown in the right panel of Fig. 2. BDeu results in wildly different values depending on the choice of α, with Bayes factors that vary between 0.05 and 1 for small changes of $\alpha \in [1, 10]$; BDs always gives a Bayes factor of 1. Again $d_{\text{EP}}^{(\mathcal{G}^+)} - d_{\text{EP}}^{(\mathcal{G}^-)} = 4 - 4 = 0$, which agrees with the fact that the Bayes factor for BDeu does not diverge or converge to zero; and $\mathcal{G}^-$ and $\mathcal{G}^+$ have the same BIC score, so BDeu (but not BDs) violates the MDL principle in this example as well. □

4 Bayesian structure learning and entropy

Shannon's classic definition of entropy for a multinomial random variable $X \sim \text{Multinomial}(\boldsymbol{\pi})$ with a fixed, finite set of states (alphabet) $\mathcal{A}$ is:

$$H(X;\boldsymbol{\pi}) = E(-\log P(X)) = -\sum_{a \in \mathcal{A}} \pi_a \log \pi_a$$

where the probabilities π_a are typically estimated with the empirical frequency of each a in $\mathcal{D}$, leading to the *empirical entropy estimator*. Its properties are detailed in canonical information theory books such as Mackay (2003) and Rissanen (2007), and it has often been used in BN structure learning (Lam and Bacchus 1994; Suzuki 2015). However, in this paper we will focus on Bayesian entropy estimators, for two reasons. Firstly, they are a natural choice when studying the properties of BD scores since they are Bayesian in nature; and having the same probabilistic assumptions (including the choice of prior distribution) for the BD scores and for the corresponding entropy estimators makes it easy to link their properties. Secondly, Bayesian entropy estimators have better theoretical and empirical properties than the empirical estimator (Hausser and Strimmer 2009; Nemenman et al. 2002).

Starting from (1), for a BN we can write:

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

where $H^{\mathcal{G}}(X_i;\Theta_{X_i})$ is the entropy of X_i given its parents $\Pi_{X_i}^{\mathcal{G}}$. The marginal posterior expectation of $H^{\mathcal{G}}(X_i;\Theta_{X_i})$ with respect to Θ_{X_i} given the data can then be expressed as:

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

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. We can then introduce a Dirichlet(α_{ijk}) prior over Θ_{X_i} with:

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

which leads to

$$\begin{aligned} E\big(H^{\mathcal{G}}(X_i)|\mathcal{D}\big) &= \iint H^{\mathcal{G}}(X_i;\Theta_{X_i})P(\Theta_{X_i}|\mathcal{D}, \alpha_{ijk})P(\alpha_{ijk}|\mathcal{D})\,\mathrm{d}\alpha_{ijk}\,\mathrm{d}\Theta_{X_i} \\ &\propto \int E(H^{\mathcal{G}}(X_i)|\mathcal{D}, \alpha_{ijk})P(\mathcal{D}|\alpha_{ijk})P(\alpha_{ijk})\,\mathrm{d}\alpha_{ijk}, \end{aligned} \tag{9}$$

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

The first term on the right hand-side of (9) is the posterior expectation of:

$$H^{\mathcal{G}}(X_i|\mathcal{D}, \alpha_{ijk}) = -\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} p_{ik|j}^{(\alpha_{ijk})} \log p_{ik|j}^{(\alpha_{ijk})} \quad \text{with } p_{ik|j}^{(\alpha_{ijk})} = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \tag{10}$$

and has closed form

$$\begin{aligned} &E(H^{\mathcal{G}}(X_i)|\mathcal{D}, \alpha_{ijk}) \\ &= \sum_{j=1}^{q_i}\left[\psi_0(\alpha_{ij} + n_{ij} + 1) - \sum_{k=1}^{r_i}\frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}}\psi_0(\alpha_{ijk} + n_{ijk} + 1)\right] \end{aligned} \tag{11}$$

as shown in Nemenman et al. (2002) and Archer et al. (2014), with $\psi_0(\cdot)$ denoting the digamma function. The second term follows a *Dirichlet-multinomial distribution* (also known as *multivariate Polya*; Johnson et al. 1997) with density

$$P(\mathcal{D}|\alpha_{ijk}) = \prod_{j=1}^{q_i}\frac{n_{ij}!\,\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ijk})^{r_i}\Gamma(n_{ij} + \alpha_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(n_{ijk} + \alpha_{ijk})}{n_{ijk}!}, \tag{12}$$

since,

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

where $P(\mathcal{D}|\Theta_{X_i})$ follows a multinomial distribution and $P(\Theta_{X_i}|\alpha_{ijk})$ is a conjugate Dirichlet prior. Rearranging terms in (12) we find that,

$$P(\mathcal{D}|\alpha_{ijk}) = \prod_{j=1}^{q_i} \frac{n_{ij}!}{\prod_{k=1}^{r_i} n_{iijk}!} \cdot \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij}+\alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(n_{ijk}+\alpha_{ijk})}{\Gamma(\alpha_{ijk})} \propto \mathrm{BD}\Big(X_i|\Pi_{X_i}^{\mathcal{G}};\alpha\Big) \tag{13}$$

making the link between BD scores and entropy explicit. Unlike (13), BD has a prequential formulation (Dawid 1984) which focuses on the sequential prediction of future events (Chickering and Heckerman 2000). For this reason it considers observations as coming in a specific temporal order and it does not include a multinomial coefficient, which we will drop in the remainder of the paper. Therefore,

$$E\big(H^{\mathcal{G}}(X_i)|\mathcal{D}\big) = \int E\big(H^{\mathcal{G}}(X_i)|\mathcal{D},\alpha_{ijk}\big)\mathrm{BD}\Big(X_i|\Pi_{X_i}^{\mathcal{G}};\alpha\Big)P(\alpha_{ijk})\,\mathrm{d}\alpha_{ijk}, \tag{14}$$

and is determined by three components: the posterior expected entropy of $X_i|\Pi_{X_i}^{\mathcal{G}}$ under a Dirichlet(α_{ijk}) prior, the BD score term for $X_i|\Pi_{X_i}^{\mathcal{G}}$, and the hyperprior over the space of the Dirichlet priors.

This definition of the expected entropy associated with the structure $\mathcal{G}$ of a BN is very general and encompasses the entropies associated with all the BD scores as special cases. In particular, the entropy associated with K2, BDJ, BDeu and BDs can be obtained by giving $P(\alpha_{ijk}) = 1$ to the single set of α_{ijk} associated with the corresponding prior, leading to,

$$E\big(H^{\mathcal{G}}(X_i)|\mathcal{D}\big) = E\big(H^{\mathcal{G}}(X_i)|\mathcal{D},\alpha_{ijk}\big)\mathrm{BD}\Big(X_i|\Pi_{X_i}^{\mathcal{G}};\alpha\Big);$$

and similarly for BDla,

$$E\big(H^{\mathcal{G}}(X_i)|\mathcal{D}\big) = \frac{1}{|S_L|}\sum_{s\in S_L} E\Big(H^{\mathcal{G}}(X_i)|\mathcal{D},\alpha_{ijk}^{s}\Big)\mathrm{BD}\Big(X_i|\Pi_{X_i}^{\mathcal{G}};s\Big).$$

5 The posterior marginal entropy

The posterior expectation of the entropy for a given Dirichlet(α_{ijk}) prior in (11), despite having a form that looks very different from the marginal posterior entropy in (10), can be written in terms of the latter as we show in the following lemma.

Lemma 1

$$E\big(H^{\mathcal{G}}(X_i)|\mathcal{D};\alpha_{ijk}\big) \approx H^{\mathcal{G}}(X_i|\mathcal{D},\alpha_{ijk}) - \sum_{j=1}^{q_i}\frac{r_i-1}{2(\alpha_{ij}+n_{ij})}.$$

Proof of Lemma 1 Combining $\psi_0(z+1) = \psi_0(z) + 1/z$ with $\psi_0(z) = \log(z) - 1/(2z) + o(z^{-2})$ from Anderson and Qiu (1997), we can write $\psi_0(z+1) = \log(z) + 1/(2z) + o(z^{-2})$. Dropping the remainder term $o(z^{-2})$ we approximate $\psi_0(z+1) \approx \log(z) + 1/(2z)$, which leads to,

$$
\begin{aligned}
&E\big(H^{\mathcal{G}}(X_i)|\mathcal{D}, \alpha_{ijk}\big) \\
&= \sum_{j=1}^{q_i} \left[\psi_0(\alpha_{ij} + n_{ij} + 1) - \sum_{k=1}^{r_i} \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \psi_0(\alpha_{ijk} + n_{ijk} + 1) \right] \\
&\approx \sum_{j=1}^{q_i} \left[\log(\alpha_{ij} + n_{ij}) + \frac{1}{2(\alpha_{ij} + n_{ij})} \right. \\
&\quad \left. - \sum_{k=1}^{r_i} \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \left(\log(\alpha_{ijk} + n_{ijk}) + \frac{1}{2(\alpha_{ijk} + n_{ijk})} \right) \right] \\
&= -\sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \log \left(\frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \right) - \sum_{j=1}^{q_i} \frac{r_i - 1}{2(\alpha_{ij} + n_{ij})} \\
&= H^{\mathcal{G}}(X_i|\mathcal{D}, \alpha_{ijk}) - \sum_{j=1}^{q_i} \frac{r_i - 1}{2(\alpha_{ij} + n_{ij})}.
\end{aligned}
$$

□

Therefore, $E(H^{\mathcal{G}}(X_i)|\mathcal{D};\alpha_{ijk})$ is well approximated by the marginal posterior entropy $H^{\mathcal{G}}(X_i|\mathcal{D}, \alpha_{ijk})$ from (10) plus a bias term that depends on the augmented counts $\alpha_{ij} + n_{ij}$ for the q_i configurations of $\Pi^{\mathcal{G}}_{X_i}$. A similar result was derived in Miller (1955) for the empirical entropy estimator and is the basis of the Miller–Madow entropy estimator.

6 BDeu and the principle of maximum entropy

The *maximum relative entropy* principle (ME; Shore and Johnson 1980; Skilling 1988; Caticha 2004) states that we should choose a model that is consistent with our knowledge of the phenomenon we are modelling and that introduces no unwarranted information. In the context of probabilistic learning this means choosing the model that has the largest possible entropy for the data, which will encode the probability distribution that best reflects our current knowledge of $\mathbf{X}$ given by $\mathcal{D}$. In the Bayesian setting in which BD scores are defined, we then prefer a DAG $\mathcal{G}^+$ over a second DAG $\mathcal{G}^-$ if

$$
E\big(H^{\mathcal{G}^-}(\mathbf{X})|\mathcal{D}\big) \leqslant E\big(H^{\mathcal{G}^+}(\mathbf{X})|\mathcal{D}\big) \tag{15}
$$

because these estimates of entropy incorporate all our knowledge including that encoded in the prior and in the hyperprior. The resulting formulation of ME

represents a very general approach that includes Bayesian posterior estimation as a particular case (Giffin and Caticha 2007); which is intuitively apparent since the expected posterior entropy in (14) is proportional to BD. Furthermore, ME can also be seen as a particular case of the MDL principle (Feder 1986).

Suzuki (2016) defined *regular* those BD scores that prefer simpler BNs that have smaller empirical entropies and few arcs:

$$H\left(X_i|\Pi_{X_i}^{\mathcal{G}^-};\pi_{ik|j}\right) \leqslant H\left(X_i|\Pi_{X_i}^{\mathcal{G}^+};\pi_{ik|j}\right), \quad \Pi_{X_i}^{\mathcal{G}^-} \subset \Pi_{X_i}^{\mathcal{G}^+} \Rightarrow \\ \mathrm{BD}\left(X_i|\Pi_{X_i}^{\mathcal{G}^-};\alpha_i\right) \geqslant \mathrm{BD}\left(X_i|\Pi_{X_i}^{\mathcal{G}^+};\alpha_i\right).$$

For large sample sizes, the probabilities $p_{ik|j}^{(\alpha_{ijk})}$ from (10) used in the posterior entropy estimators converge to the empirical frequencies used in the empirical entropy estimator, making the above asymptotically equivalent to,

$$H\left(X_i|\Pi_{X_i}^{\mathcal{G}^-};\alpha_{ijk}\right) \leqslant H\left(X_i|\Pi_{X_i}^{\mathcal{G}^+};\alpha_{ijk}\right), \quad \Pi_{X_i}^{\mathcal{G}^-} \subset \Pi_{X_i}^{\mathcal{G}^+} \Rightarrow \\ \mathrm{BD}\left(X_i|\Pi_{X_i}^{\mathcal{G}^-};\alpha_i\right) \geqslant \mathrm{BD}\left(X_i|\Pi_{X_i}^{\mathcal{G}^+};\alpha_i\right)$$

and connecting DAGs with the highest BD scores with those that minimise the marginal posterior entropy from (10). However, we prefer to study BDeu and its prior using ME as defined in (15) for two reasons. Firstly, posterior expectations are widely considered to be superior to MAP estimates in the literature (Berger 1985), as has been specifically shown for entropy in Nemenman et al. (2002). Secondly, ME directly incorporates the information encoded in the prior and in the hyperprior, without relying on large samples to link the empirical entropy (which depends on $\mathbf{X}, \boldsymbol{\Theta}$) with the BD scores (which depend on $\mathbf{X}, \boldsymbol{\alpha}$ and integrate $\boldsymbol{\Theta}$ out).

Without loss of generality, we consider again the simple case in which $\mathcal{G}^-$ and $\mathcal{G}^+$ differ by a single arc, so that only one local distribution differs between the two DAGs. For BDeu, $\alpha_{ijk} = \alpha/(r_i q_i)$ and substituting (14) in (15) we get,

$$E\left(H^{\mathcal{G}^-}(X_i)|\mathcal{D},\alpha_{ijk}\right)\mathrm{BDeu}\left(X_i|\Pi_{X_i}^{\mathcal{G}^-};\alpha\right) \\ \leqslant E\left(H^{\mathcal{G}^+}(X_i)|\mathcal{D},\alpha_{ijk}\right)\mathrm{BDeu}\left(X_i|\Pi_{X_i}^{\mathcal{G}^+};\alpha\right). \tag{16}$$

If the sample $\mathcal{D}$ is sparse, some configurations of the variables will not be observed and the effective imaginary sample size may be smaller for $\mathcal{G}^+$ than for $\mathcal{G}^-$ like in Examples 1 and 2. As a result, when we compare a $\mathcal{G}^-$ for which we observe all configurations of $\Pi_{X_i}^{\mathcal{G}^-}$ with a $\mathcal{G}^+$ for which we do not observe some configurations of $\Pi_{X_i}^{\mathcal{G}^+}$, instead of (16) we are actually using,

$$
\begin{aligned}
&E\left(H^{\mathcal{G}^-}(X_i)|\mathcal{D},\alpha_{ijk}\right)\mathrm{BDeu}\left(X_i|\Pi^{\mathcal{G}^-}_{X_i};\alpha\right)\\
&\quad\leqslant E\left(H^{\mathcal{G}^+}(X_i)|\mathcal{D},\alpha_{ijk}\right)\mathrm{BDeu}\left(X_i|\Pi^{\mathcal{G}^+}_{X_i};\alpha(\tilde{q}_i/q_i)\right)
\end{aligned}
\tag{17}
$$

which is different from (15) and thus not consistent with ME. It is not correct from a Bayesian perspective either, because $\mathcal{G}^-$ and $\mathcal{G}^+$ are compared using marginal likelihoods arising from different priors; as expected since we know from Giffin and Caticha (2007) that Bayesian posterior inference is a particular case of ME. From the perspective of ME, those priors carry different information on the Θ_{X_i}. They incorrectly express different levels of belief in the uniform prior underlying BDeu as a consequence of the difference in their effective imaginary sample sizes, even though they were meant to express the same level of belief for all possible DAGs. This may bias the entropy (which is, after all, the expected value of the information on $\mathbf{X}$ and on the Θ_{X_i}) of $\mathcal{G}^+$ compared to that of $\mathcal{G}^-$ in (17) and lead to choosing DAGs which would not be chosen by ME. We would like to stress that this scenario is not uncommon; on the contrary, such a model comparison is almost guaranteed to take place when the data are sparse. As structure learning explores more and more DAGs to identify an optimal one, it will inevitably consider DAGs with unobserved parent configurations, either because they are too dense or because those parent sets are not well supported by the few observed data points.

In particular, if some $n_{ij} = 0$ then the posterior expected entropy of $X_i|\Pi^{\mathcal{G}^+}_{X_i}$ becomes,

$$
\begin{aligned}
&E\left(H^{\mathcal{G}^+}(X_i)\,|\,\mathcal{D},\alpha_{ijk}\right)\\
&\quad=\sum_{j:n_{ij}=0}\left[\psi_0(r_i\alpha_{ijk}+1)-\sum_{k=1}^{r_i}\frac{\cancel{\alpha_{ijk}}}{r_i\cancel{\alpha_{ijk}}}\psi_0(\alpha_{ijk}+1)\right]\\
&\qquad+\sum_{j:n_{ij}>0}\left[\psi_0(r_i\alpha_{ijk}+n_{ij}+1)-\sum_{k=1}^{r_i}\frac{\alpha_{ijk}+n_{ijk}}{r_i\alpha_{ijk}+n_{ij}}\psi_0(\alpha_{ijk}+n_{ijk}+1)\right]
\end{aligned}
$$

where the first term collects the conditional entropies corresponding to the $q_i - \tilde{q}_i$ unobserved parent configurations, for which the posterior distribution coincides with the prior:

$$
\begin{aligned}
&\sum_{j:n_{ij}=0}\left[\psi_0(r_i\alpha_{ijk}+1)-\sum_{k=1}^{r_i}\frac{1}{r_i}\psi_0(\alpha_{ijk}+1)\right]\approx\\
&\qquad-\sum_{j:n_{ij}=0}\sum_{k=1}^{r_i}\frac{\cancel{\alpha_{ijk}}}{r_i\cancel{\alpha_{ijk}}}\log\left(\frac{\cancel{\alpha_{ijk}}}{r_i\cancel{\alpha_{ijk}}}\right)-\sum_{j=1}^{q_i}\frac{r_i-1}{2\alpha_{ij}}=(q_i-\tilde{q}_i)\left[-\log\frac{1}{r_i}-\frac{r_i-1}{2\alpha_{ij}}\right].
\end{aligned}
\tag{18}
$$

By definition, the uniform distribution has the maximum possible entropy; the posteriors we would estimate if we could observe samples for those configurations of the $\Pi_{X_i}^{\mathcal{G}^+}$ would almost certainly have a smaller entropy. At the same time, the entropies in the second term are smaller than what they would be if we only considered the $\tilde{q}_i$ observed parent configurations, because $\alpha_{ijk} = \alpha/(r_i q_i) < \alpha/(r_i \tilde{q}_i)$ means that posterior densities deviate more from the uniform distribution. These two effects, however, do not necessarily balance each other out; as we can see by revisiting Examples 1 and 2 below it is possible to incorrectly choose $\mathcal{G}^+$ over $\mathcal{G}^-$.

Example 1 (*Continued*) The empirical posterior entropies for $\mathcal{G}^-$ and $\mathcal{G}^+$ are:

$$H(X|\Pi_X^{\mathcal{G}^-}) = H(X|\Pi_X^{\mathcal{G}^+}) = 4\big[-0\log 0 - 1\log 1\big] = 0$$

by convention, but the posterior entropies differ:

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

The expected posterior entropies for $\mathcal{G}^-$ and $\mathcal{G}^+$ are:

$$\begin{aligned}
&E\Big(H^{\mathcal{G}^-}(X)|\mathcal{D}, \frac{1}{8}\Big) \\
&\quad = 4\left[\psi_0(1/4+3+1) - \frac{0+1/8}{3+1/4}\psi_0(1/8+0+1) - \frac{3+1/8}{3+1/4}\psi_0(1/8+3+1)\right] \\
&\quad = 0.3931, \\
&E\Big(H^{\mathcal{G}^+}(X)|\mathcal{D}, \frac{1}{16}\Big) \\
&\quad = 4\left[\psi_0(1/8+3+1) - \frac{0+1/16}{3+1/8}\psi_0(1/16+0+1) - \frac{3+1/16}{3+1/8}\psi_0(1/16+3+1)\right] \\
&\qquad + 4\left[\psi_0(1/8+0+1) - \frac{0+1/16}{0+1/8}\psi_0(1/16+0+1) - \frac{3+1/16}{0+1/8}\psi_0(1/16+0+1)\right] \\
&\quad = 0.5707.
\end{aligned}$$

Therefore, substituting these values in (17),

$$
\begin{aligned}
E(H^{\mathcal{G}^-}(X)|\mathcal{D}) &= 0.3931 \cdot 0.0326 = 0.0128 \\
&< 0.0252 = 0.5707 \cdot 0.0441 = E(H^{\mathcal{G}^+}(X)|\mathcal{D});
\end{aligned}
$$

and we would choose $\mathcal{G}^+$ over $\mathcal{G}^-$ even though we only observe $\tilde{q}_i = 4$ configurations of $\Pi_X^{\mathcal{G}^+}$ out of 8, and the sample frequencies are identical for those configurations. The data contribute the same information to the posterior expected entropies; both $X|\Pi_X^{\mathcal{G}^-}$ and $X|\Pi_X^{\mathcal{G}^+}$ have empirical entropy equal to zero. The difference must then arise because of the priors: both $\Theta_X|\Pi_X^{\mathcal{G}^-}$ and $\Theta_X|\Pi_X^{\mathcal{G}^+}$ follow a uniform Dirichlet prior, but in the former $\alpha = 1$ and in the latter $\alpha = 1/2$ because $\tilde{q}_i = 4 < 8 = q_i$. A consistent model comparison requires that all models are evaluated with the same prior, which clearly is not the case in this example. □

Example 2 (*Continued*) The conditional distributions of $X|\Pi_X^{\mathcal{G}^-}$ and $X|\Pi_X^{\mathcal{G}^+}$ both have the same (positive) empirical entropy:

$$
H(X|\Pi_X^{\mathcal{G}^-}) = H(X|\Pi_X^{\mathcal{G}^+}) = 4\left[-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right] = 2.546.
$$

However, their posterior entropies are different:

$$
\begin{aligned}
H(X|\Pi_X^{\mathcal{G}^-};\alpha) =& 4\left[-\frac{1+1/8}{3+1/4}\log\frac{1+1/8}{3+1/4} - \frac{2+1/8}{3+1/4}\log\frac{2+1/8}{3+1/4}\right] \\
=& 2.580, \\
H(X|\Pi_X^{\mathcal{G}^+};\alpha) =& 4\left[-\frac{1+1/16}{3+1/8}\log\frac{1+1/16}{3+1/8} - \frac{2+1/16}{3+1/8}\log\frac{2+1/16}{3+1/8}\right] \\
=& 2.564;
\end{aligned}
$$

and the respective posterior expected entropies are:

$$
\begin{aligned}
&E\left(H^{\mathcal{G}^-}(X)|\mathcal{D}, \frac{1}{8}\right) \\
&\quad = 4\left[\psi_0(1/4+3+1) - \frac{1+1/8}{3+1/4}\psi_0(1/8+1+1) - \frac{2+1/8}{3+1/4}\psi_0(1/8+2+1)\right] \\
&\quad = 2.066, \\
&E\left(H^{\mathcal{G}^+}(X)|\mathcal{D}, \frac{1}{16}\right) \\
&\quad = 4\left[\psi_0(1/8+3+1) - \frac{0+1/16}{3+1/8}\psi_0(1/16+1+1) - \frac{3+1/16}{3+1/8}\psi_0(1/16+2+1)\right] \\
&\qquad + 4\left[\psi_0(1/8+0+1) - \frac{0+1/16}{0+1/8}\psi_0(1/16+0+1) - \frac{3+1/16}{0+1/8}\psi_0(1/16+0+1)\right] \\
&\quad = 4.069.
\end{aligned}
$$

Therefore, substituting these values in (17) leads to,

$$E(H^{\mathcal{G}^-}(X)|\mathcal{D}) = 2.066 \cdot 3.906 \times 10^{-7} = 8.071 \times 10^{-7}$$
$$> 1.514 \times 10^{-7} = 4.069 \cdot 3.721 \times 10^{-8} = E(H^{\mathcal{G}^+}(X)|\mathcal{D}).$$

Even though we choose $\mathcal{G}^-$ over $\mathcal{G}^+$, we still express a preference for one of the DAGs even though the information in the data is the same; which confirms that the difference is attributable to the prior. □

Based on these results and the examples above, we state the following theorem.

Theorem 1 *Using BDeu and the associated uniform prior over the parameters of a BN for structure learning violates the maximum relative entropy principle if any candidate parent configuration of any node is not observed in the data.*

This is not the case for BDs, because its piecewise uniform prior preserves the imaginary sample size even when $\tilde{q}_i < q_i$; and because it prevents the posterior entropy from inflating by allowing the $\tilde{q}_i$ terms corresponding to the $n_{ij} = 0$ to simplify. Assuming $\alpha_{ijk} = 0$ in (18) implies:

$$\sum_{j:n_{ij}=0} \left[\psi_0(1) - \sum_{k=1}^{r_i} \frac{1}{r_i} \psi_0(1) \right] = \psi_0(1) - \psi_0(1) = 0.$$

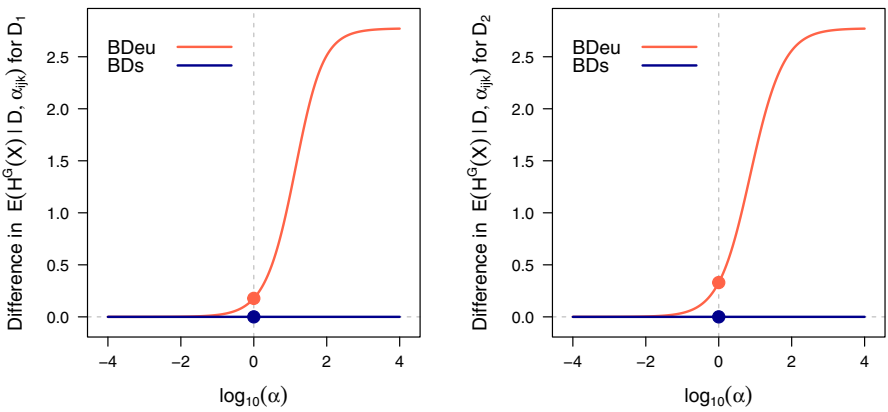


Fig. 3 The difference between $E(H^{\mathcal{G}^-}(X)|\mathcal{D}, \alpha_{ijk})$ and $E(H^{\mathcal{G}^+}(X)|\mathcal{D}, \alpha_{ijk})$ computed using BDeu and BDs for Example 1 (left panel) and Example 2 (right panel) in orange and dark blue, respectively. The bullet points correspond to the values observed for $\alpha = 1$

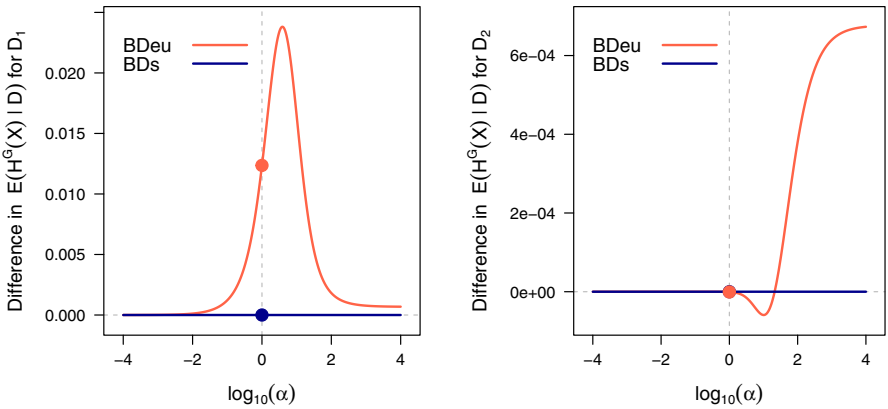


Fig. 4 The difference between $E(H^{\mathcal{G}^-}(X)|\mathcal{D})$ and $E(H^{\mathcal{G}^+}(X)|\mathcal{D})$ computed using BDeu and BDs for Example 1 (left panel) and Example 2 (right panel) in orange and dark blue, respectively. The bullet points correspond to the values observed for $\alpha = 1$

Example 1 (*Continued*) If we compare $X|\Pi_X^{\mathcal{G}^-}$ and $X|\Pi_X^{\mathcal{G}^+}$ under the prior assumed by BDs, we have that $\alpha_{ijk} = 1/8$ for both $X|\Pi_X^{\mathcal{G}^-}$ and the $\tilde{q}_i$ observed parent configurations in $X|\Pi_X^{\mathcal{G}^+}$. Then their posterior expected entropies are:

$$\begin{aligned} E\left(H^{\mathcal{G}^-}(X)|\mathcal{D}, \frac{1}{8}\right) &= E\left(H^{\mathcal{G}^+}(X)|\mathcal{D}, \frac{1}{8}\right) \\ &= 4\left[\psi_0(1/4+3+1) - \frac{0+1/8}{3+1/4}\psi_0(1/8+0+1) - \frac{3+1/8}{3+1/4}\psi_0(1/8+3+1)\right] \\ &= 0.3931 \end{aligned}$$

and substituting these values in (16)

$$\begin{aligned} E(H^{\mathcal{G}^-}(X)|\mathcal{D}) &= 0.3931 \cdot 0.0326 = 0.0128 \\ &= 0.0128 = 0.3931 \cdot 0.0326 = E(H^{\mathcal{G}^+}(X)|\mathcal{D}). \end{aligned}$$

ME does not express a preference for either $\mathcal{G}^-$ or $\mathcal{G}^+$; since we have observed above that the data contribute exactly the same information for both DAGs, the same must be true for the prior associated with BDs.

A side effect of not violating ME is that the choice between $\mathcal{G}^-$ and $\mathcal{G}^+$ is no longer sensitive to the value of α; we can see from the left panels of Figs. 3 and 4 that both the difference between $E(H^{\mathcal{G}^-}(X)|\mathcal{D}, \frac{1}{8})$ and $E(H^{\mathcal{G}^+}(X)|\mathcal{D}, \frac{1}{8})$ and the difference between $E(H^{\mathcal{G}^-}(X)|\mathcal{D})$ and $E(H^{\mathcal{G}^+}(X)|\mathcal{D})$ are equal to zero for all α. □

Example 2 (*Continued*) Again $\alpha_{ijk} = 1/8$ for both $X|\Pi_X^{\mathcal{G}^-}$ and the $\tilde{q}_i$ observed parent configurations in $X|\Pi_X^{\mathcal{G}^+}$, so

$$\begin{aligned} E\left(H^{\mathcal{G}^-}(X)|\mathcal{D}, \frac{1}{8}\right) &= E\left(H^{\mathcal{G}^+}(X)|\mathcal{D}, \frac{1}{8}\right) \\ &= 4\left[\psi_0(1/4+3+1) - \frac{1+1/8}{3+1/4}\psi_0(1/8+1+1) - \frac{2+1/8}{3+1/4}\psi_0(1/8+2+1)\right] \\ &= 2.066 \end{aligned}$$

which leads to

$$\begin{aligned} E(H^{\mathcal{G}^-}(X)|\mathcal{D}) &= 2.066 \cdot 3.906 \times 10^{-7} = 8.071 \times 10^{-7} \\ &= 8.071 \times 10^{-7} = 2.066 \cdot 3.906 \times 10^{-7} = E(H^{\mathcal{G}^+}(X)|\mathcal{D}). \end{aligned}$$

Again we can see from the right panels of Figs. 3 and 4 that the choice between $\mathcal{G}^-$ and $\mathcal{G}^+$ is no longer sensitive to the choice of α; and $\mathcal{G}^+$ is never preferred to $\mathcal{G}^-$. This contrasts especially the behaviour of BDeu in Fig. 4, where $E(H^{\mathcal{G}^+}(X)|\mathcal{D})$ can be both larger and smaller than $E(H^{\mathcal{G}^-}(X)|\mathcal{D})$ for different values of α. □

It is easy to show that the theorem we just stated does not apply to K2 or BDJ, since under their priors α_{ijk} is not a function of q_i; but it does apply to BDla since its formulation is essentially a mixture of BDeu scores.

7 Conclusions and discussion

Bayesian network learning follows an inherently Bayesian workflow in which we first learn the structure of the DAG $\mathcal{G}$ from a data set $\mathcal{D}$, and then we learn the values of the parameters Θ_{X_i} given $\mathcal{G}$. In this paper we studied the properties of the Bayesian posterior scores used to estimate $P(\mathcal{G}|\mathcal{D})$ and to learn the $\mathcal{G}$ that best fits the data. For discrete Bayesian networks, these scores are Bayesian–Dirichlet (BD) marginal likelihoods that assume different Dirichlet priors for the Θ_{X_i} and, in the most general formulation, a hyperprior over the hyperparameters α_{ijk} of the prior. We focused on the most common BD score, BDeu, which assumes a uniform prior over each Θ_{X_i}; and we studied the impact of that prior on structure learning from a Bayesian and an information theoretic perspective. After deriving the form of the posterior expected entropy for $\mathcal{G}$ given $\mathcal{D}$, we found that BDeu may select models in a way that violates the maximum relative entropy principle. Furthermore, we showed that it produces Bayes factors that are very sensitive to the choice of the imaginary sample size. Both issues are related to the uniform prior assumed by BDeu for the Θ_{X_i}, and can lead to the selection of overly dense DAGs when the data are sparse. In contrast, the BDs score proposed in Scutari (2016) does not, even though it converges to BDeu asymptotically; and neither do other BD

scores in the literature. In the simulation study we performed in Scutari (2016), we found that BDs leads to more accurate structure learning; hence we recommend its use over BDeu for sparse data.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD (2010a) Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. J Mach Learn Res 11:171–234

Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD (2010b) Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. J Mach Learn Res 11:235–284

Anderson G, Qiu S (1997) A monotonicity property of the gamma function. Proc Am Math Soc 125(11):3355–3362

Archer E, Park IM, Pillow JW (2014) Bayesian entropy estimation for countable discrete distributions. J Mach Learn Res 15:2833–2868

Berger JO (1985) Statistical decision theory and Bayesian analysis, 2nd edn. Springer, Berlin

Cano A, Gómez-Olmedo M, Masegosa AR, Moral S (2013) Locally averaged Bayesian Dirichlet metrics for learning the structure and the parameters of Bayesian networks. Int J Approx Reason 54(4):526–540

Castelo R, Siebes A (2000) Priors on network structures. Biasing the search for Bayesian networks. Int J Approx Reason 24(1):39–57

Caticha A (2004) Relative entropy and inductive inference. In: Bayesian inference and maximum entropy methods in science and engineering. AIP, New York, pp 75–96

Chickering DM (1995) A transformational characterization of equivalent Bayesian network structures. In: Proceedings of the 11th conference on uncertainty in artificial intelligence. Morgan Kaufmann, San Francisco, pp 87–98

Chickering DM, Heckerman D (2000) A comparison of scientific and engineering criteria for Bayesian model selection. Stat Comput 10:55–62

Cooper GF, Herskovits E (1991) A Bayesian method for constructing Bayesian belief networks from databases. In: Proceedings of the 7th conference on uncertainty in artificial intelligence. Morgan Kaufmann, San Francisco, pp 86–94

Cussens J (2012) Bayesian network learning with cutting planes. In: Proceedings of the 27th conference on uncertainty in artificial intelligence. AUAI Press, pp 153–160

Dawid AP (1984) Present position and potential developments: some personal views: statistical theory: the prequential approach. J R Stat Soc Ser A 147(2):278–292

Feder M (1986) Maximum entropy as a special case of the minimum description length criterion. IEEE Trans Inf Theory 32(6):847–849

Friedman N, Nachman I, Peér D (1999) Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In: Proceedings of the 15th conference on uncertainty in artificial intelligence. Morgan Kaufmann, San Francisco, pp 206–221

Geiger D, Heckerman D (1994) Learning Gaussian networks. In: Proceedings of the 10th conference on uncertainty in artificial intelligence. Morgan Kaufmann, San Francisco, pp 235–243

Giffin A, Caticha A (2007) Updating probabilities with data and moments. In: Proceedings of the 27th international workshop on Bayesian inference and maximum entropy methods in science and engineering. AIP, New York, pp 74–84

Hausser J, Strimmer K (2009) Entropy inference and the James–Stein estimator, with application to nonlinear gene association networks. J Mach Learn Res 10:1469–1484

Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 20(3):197–243 **(available as Technical Report MSR-TR-94-09)**

Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge

Koski TJT, Noble JM (2012) A review of Bayesian networks and structure learning. Math Appl 40(1):53–103

Lam W, Bacchus F (1994) Learning Bayesian belief networks: an approach based on the MDL principle. Comput Intell 10:269–293

Lauritzen SL, Wermuth N (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 17(1):31–57

Mackay DJC (2003) Information theory, inference and learning algorithms. Cambridge University Press, Cambridge

Miller GA (1955) Note on the bias of information estimates. In: Information theory in psychology II-B. Free Press, Glencoe, Illinois, New York, pp 95–100

Mukherjee S, Speed TP (2008) Network inference using informative priors. Proc Natl Acad Sci 105(38):14313–14318

Nemenman I, Shafee F, Bialek W (2002) Entropy and inference, revisited. In: Proceedings of the 14th advances in neural information processing systems (NIPS) conference. MIT Press, Cambridge, Massachusetts, pp 471–478

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, Burlington

Pearl J (2009) Causality: models, reasoning and inference, 2nd edn. Cambridge University Press, Cambridge

Rissanen J (1978) Modeling by shortest data description. Automatica 14(5):465–568

Rissanen J (2007) Information and complexity in statistical models. Springer, Berlin

Russell SJ, Norvig P (2009) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall, Upper Saddle River

Scutari M (2016) An empirical-Bayes score for discrete Bayesian networks. J Mach Learn Res (Proc Track PGM 2016) 52:438–448

Scutari M, Denis JB (2014) Bayesian networks with examples in R. Chapman & Hall, London

Shore JE, Johnson RW (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans Inf Theory IT–26(1):26–37

Silander T, Kontkanen P, Myllymäki P (2007) 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. AUAI Press, pp 360–367

Skilling J (1988) The axioms of maximum entropy. In: Maximum-entropy and Bayesian methods in science and engineering. Springer, Berlin, pp 173–187

Steck H (2008) Learning the Bayesian network structure: dirichlet prior versus data. In: Proceedings of the 24th conference on uncertainty in artificial intelligence. AUAI Press, pp 511–518

Steck H, Jaakkola TS (2003) On the Dirichlet prior and Bayesian regularization. Adv Neural Inf Process Syst 15:713–720

Suzuki J (2015) Consistency of learning Bayesian network structures with continuous variables: an information theoretic approach. Entropy 17:5752–5770

Suzuki J (2016) A theoretical analysis of the BDeu scores in Bayesian network structure learning. Behaviormetrika 44:97–116

Tsamardinos I, Brown LE, Aliferis CF (2006) The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 65(1):31–78

Ueno M (2010) Learning networks determined by the ratio of prior and data. In: Proceedings of the 26th conference on uncertainty in artificial intelligence. AUAI Press, pp 598–605

Ueno M (2011) Robust learning of Bayesian networks for prior belief. In: Proceedings of the 27th conference on uncertainty in artificial intelligence. AUAI Press, pp 698–707

Weatherburn CE (1961) A first course in mathematical statistics. Cambridge University Press, Cambridge