Dirichlet Bayesian Network Scores and the Maximum Entropy Principle

Marco Scutari SCUTARI@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford, United Kingdom

Abstract

A classic approach for learning Bayesian networks from data is to select the *maximum a posteriori* (MAP) network. In the case of discrete Bayesian networks, the MAP network is 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. (1995). The key properties of BDeu arise from its underlying uniform prior, which makes structure learning computationally efficient; does not require the elicitation of prior knowledge from experts; and satisfies score equivalence.

In this paper we will discuss the impact of this uniform prior on structure learning from an information theoretic perspective, showing how BDeu may violate the maximum entropy principle when applied to sparse data and how it may also be problematic from a Bayesian model selection perspective. On the other hand, the BDs score proposed in Scutari (2016) arises from a piecewise prior and it does not appear to violate the maximum entropy principle, even though it is asymptotically equivalent to BDeu.

Keywords: Bayesian networks, structure learning, Bayesian posterior estimation, maximum 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 $\mathbf{X} = \{X_1, \ldots, X_N\}$. The arcs in $\mathcal{G}$ represent direct dependence relationships between the variables in $\mathbf{X}$; and graphical separation of two nodes implies the conditional independence of the corresponding X_i. 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; other possibilities include Gaussian BNs (Geiger and Heckerman, 1994) and conditional linear Gaussian BNs (Lauritzen and Wermuth, 1989).

The task of learning a BN from data is usually performed in an inherently Bayesian fashion by maximising

$$\underbrace{\mathrm{P}(\mathcal{B} \mid \mathcal{D}) = \mathrm{P}(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}, \quad (1)$$

where $\mathcal{D}$ is a sample from $\mathbf{X}$, and $\mathcal{B} = (\mathcal{G}, \Theta)$ is a BN with parameter set Θ. 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 $\mathcal{G}$. Both are computationally feasible for large data thanks to the Markov property (Pearl, 1988): in the absence of missing data the *global distribution* of $\mathbf{X}$ decomposes into

$$\mathrm{P}(\mathbf{X} \,|\, \mathcal{G}) = \prod_{i=1}^{N} \mathrm{P}(X_i \,|\, \Pi_{X_i}^{\mathcal{G}}) \tag{2}$$

where the *local distribution* of each node X_i depends only on the configurations of its parents Π_{X_i} in $\mathcal{G}$. This decomposition 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 $X_j \to X_i \leftarrow X_k$, with no arc between X_j and X_k).

Bayesian score-based structure learning uses goodness-of-fit scores to identify a *maximum a posteriori* (MAP) DAG $\mathcal{G}$ that maximises $\mathrm{P}(\mathcal{G} \,|\, \mathcal{D})$. For discrete BNs, these scores are Bayesian Dirichlet (BD) marginal likelihoods, the most common being the Bayesian Dirichlet equivalent uniform (BDeu) score from Heckerman et al. (1995). We will show that the prior uniform distribution that underlies BDeu is problematic from a Bayesian and information theoretic perspective because it may lead to violations of the maximum (relative) entropy principle (ME; Shore and Johnson, 1980; Skilling, 1988; Caticha, 2004).

The paper is organised as follows. In Section 2 we will review BD scores; and in particular BDeu, its underlying assumptions and its problems as reported in the literature. In Section 3 we will derive the posterior expected entropy associated with a DAG $\mathcal{G}$, which we will further explore in Section 4. Finally, in Section 5 we will analyse BDeu using ME, and we will compare its behaviour with that of the BDs score proposed in Scutari (2016).

2. Bayesian Dirichlet Marginal Likelihoods

Starting from (1), we can write

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

where $\mathrm{P}(\mathcal{G})$ is the prior distribution over the space of the DAGs spanning the variables in $\mathbf{X}$ and $\mathrm{P}(\mathcal{D} \,|\, \mathcal{G})$ is the marginal likelihood of the data given $\mathcal{G}$ averaged over all possible parameter sets Θ. $\mathrm{P}(\mathcal{G})$ is often taken to be a uniform $\mathrm{P}(\mathcal{G}) \propto 1$ so that it simplifies when comparing DAGs; we will do the same in this paper for simplicity. Using (2) we can then decompose $\mathrm{P}(\mathcal{D} \,|\, \mathcal{G})$ into one component for each node as follows:

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

In the case of discrete BNs, we assume $X_i \,|\, \Pi_{X_i}^{\mathcal{G}} \sim \mathit{Multinomial}(\Theta_{X_i} \,|\, \Pi_{X_i}^{\mathcal{G}})$ where the parameters $\Theta_{X_i} \,|\, \Pi_{X_i}^{\mathcal{G}}$ are the conditional probabilities $\pi_{ik\,|\,j} = \mathrm{P}(X_i = k \,|\, \Pi_{X_i}^{\mathcal{G}} = j)$. We then assume a conjugate prior $\Theta_{X_i} \,|\, \Pi_{X_i}^{\mathcal{G}} \sim \mathit{Dirichlet}(\alpha_{ijk})$, $\sum_{jk} \alpha_{ijk} = \alpha_i > 0$ to obtain the closed-form posterior $\mathit{Dirichlet}(\alpha_{ijk} + n_{ijk})$ 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:

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

where r_i is the number of states of X_i; q_i is the number of configurations of $\Pi_{X_i}^{\mathcal{G}}$; $n_{ij} = \sum_k n_{ijk}$; $\alpha_{ij} = \sum_k \alpha_{ijk}$; and $\boldsymbol{\alpha}$ is the set of the α_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;
- and 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).

BDeu is the only score-equivalent BD score (Chickering, 1995), that is, it is the only score that takes the same value for DAGs in the same equivalence class. BDs is asymptotically score-equivalent because it converges to BDeu when $n \to \infty$ and the positivity assumption holds.

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. However, Steck and Jaakkola (2003) and Silander et al. (2007) showed that DAGs selected using BDeu may have very different numbers of arcs depending on complex interactions between α and $\mathcal{D}$, controlled by the skewness of the $X_i \,|\, \Pi_{X_i}^{\mathcal{G}}$ and by the strength of the dependence relationships in $\mathbf{X}$. These results have been confirmed analytically more recently by Ueno (2010, 2011). Suzuki (2016) found that BDeu is not regular in the sense that it may learn DAGs in a way that contradicts the respective empirical entropies, even if the positivity assumption holds and 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 $n \to \infty$. A possible solution to these problems was proposed by Scutari (2016) in the form of BDs, which was shown to be both more accurate than BDeu in learning the $\mathcal{G}$ and competitive in predictive power. Its piecewise empirical Bayes prior, which assigns a constant $\alpha_{ijk} > 0$ to observable values of $X_i \,|\, \Pi_{X_i}$ (e.g., $n_{ij} > 0$) and $\alpha_{ijk} = 0$ otherwise ($n_{ij} = 0$), was shown empirically to rank DAGs consistently with the corresponding Bayesian posterior estimate of entropy, reported in (6) in the following section.

3. Bayesian Structure Learning and Entropy

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

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

where the probabilities π_a are typically estimated with the empirical frequencies 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 prior distributions) for the BD scores and for the 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 (2), 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

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

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

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

which leads to

$$\begin{aligned} \mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \,|\, \mathcal{D}\right) &= \iint \mathrm{H}^{\mathcal{G}}(X_i; \Theta_{X_i}) \,\mathrm{P}(\Theta_{X_i} \,|\, \mathcal{D}, \alpha_{ijk}) \,\mathrm{P}(\alpha_{ijk} \,|\, \mathcal{D})\, d\alpha_{ijk}\, d\Theta_{X_i} \\ &\propto \int \mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) \mathrm{P}(\mathcal{D} \,|\, \alpha_{ijk}) \,\mathrm{P}(\alpha_{ijk})\, d\alpha_{ijk}, \end{aligned} \tag{5}$$

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

The first term on the right hand-side of (5) is the posterior expectation of

$$\mathrm{H}^{\mathcal{G}}(X_i|\mathcal{D}, \alpha_{ijk}) = -\sum_{j=1}^{q_i} \sum_{k=1}^{r_i} p_{ik\,|\,j}^{(\alpha_{ijk})} \log p_{ik\,|\,j}^{(\alpha_{ijk})} \qquad \text{with} \qquad p_{ik\,|\,j}^{(\alpha_{ijk})} = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \tag{6}$$

and has closed form

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) = \sum_{j=1}^{q_i} \left[\psi_0(\alpha_{ij} + n_{ij} + 1) - \sum_{k=1}^{r_i} \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \psi_0(\alpha_{ijk} + n_{ijk} + 1) \right] \quad (7)$$

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

$$\mathrm{P}(\mathcal{D} \,|\, \alpha_{ijk}) = \prod_{j=1}^{q_i} \frac{n_{ij}!\,\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ijk})^{r_i}\Gamma(n_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(n_{ijk} + \alpha_{ijk})}{n_{ijk}!}, \quad (8)$$

since

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

where $\mathrm{P}(\mathcal{D} \,|\, \Theta_{X_i})$ follows a multinomial distribution and $\mathrm{P}(\Theta_{X_i} \,|\, \alpha_{ijk})$ is a conjugate Dirichlet prior. Rearranging terms in (8) we find that

$$\mathrm{P}(\mathcal{D} \,|\, \alpha_{ijk}) = \prod_{j=1}^{q_i} \frac{n_{ij}!}{\prod_{k=i}^{r_i} n_{iijk}!} \cdot \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(n_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \propto \mathrm{BD}(X_i \,|\, \Pi^{\mathcal{G}}_{X_i}; \alpha_{ijk}) \quad (9)$$

making the link between BD scores and entropy explicit. Unlike (9), BD has a prequential formulation (Dawid, 1984) which focuses on sequential probabilistic prediction of future events; hence it does not include a multinomial coefficient, which we will drop in the remainder of the paper. Therefore,

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \,|\, \mathcal{D}\right) = \int \mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) \mathrm{BD}(X_i \,|\, \Pi^{\mathcal{G}}_{X_i}; \alpha_{ijk})\, \mathrm{P}(\alpha_{ijk})\, d\alpha_{ijk}, \quad (10)$$

and is determined by three components: the posterior expected entropy of $X_i \,|\, \Pi^{\mathcal{G}}_{X_i}$ under a *Dirichlet*(α_{ijk}) prior, the BD score component for $X_i \,|\, \Pi^{\mathcal{G}}_{X_i}$, and the hyper-prior 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 each of the BD scores in Section 2 can be obtained by giving $\mathrm{P}(\alpha_{ijk}) = 1$ to the α_{ijk} associated with the corresponding prior, leading to

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \,|\, \mathcal{D}\right) = \mathrm{E}\left(\mathrm{H}^{\mathcal{G}}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) \mathrm{BD}(X_i \,|\, \Pi^{\mathcal{G}}_{X_i}; \alpha_{ijk}).$$

4. The Posterior Marginal Entropy

The posterior expectation of the entropy for a given *Dirichlet*(α_{ijk}) prior in (7), despite having a form that looks very different from the marginal posterior entropy in (6), can be written in terms of the latter as we show in the following lemma.

Lemma 1

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

Proof of Lemma 1 Combining $\psi_0(z+1) = \psi_0(z) + 1/z$ with $\psi_0(z) = \log(z) - 1/(2z) + \mathcal{O}(z^{-2})$ from Anderson and Qiu (1997), we can write $\psi_0(z+1) \approx \log(z) + 1/(2z)$ which leads to

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

■

Therefore, $\mathrm{E}(\mathrm{H}^{\mathcal{G}}(X_i)\,|\,\mathcal{D};\alpha_{ijk})$ is just the marginal posterior entropy $\mathrm{H}^{\mathcal{G}}(X_i|\mathcal{D},\alpha_{ijk})$ from (6) 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.

5. 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 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

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(\mathbf{X})\,|\,\mathcal{D}\right) \leqslant \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(\mathbf{X})\,|\,\mathcal{D}\right) \tag{11}$$

because these estimates of entropy incorporate all our knowledge including that encoded in the prior and in the hyper-prior. 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 (10) is proportional to BD. Furthermore, ME can also be seen as a particular case of the *minimum description length* principle (Feder, 1986, MDL).

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

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

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

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

which connects DAGs with the highest BD scores with those that minimise the marginal posterior entropy from (6), satisfying the MDL principle using a MAP estimator instead of the empirical one. However, we prefer to study BDeu and its prior using ME as defined in (11) 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 hyper-prior, without relying on large samples to link the empirical entropy (which depends on $\mathbf{X}, \Theta$) with the BD scores (which depend on $\mathbf{X}, \boldsymbol{\alpha}$ and integrate Θ out).

Without loss of generality, we consider now the simple case in which $\mathcal{G}^-$ and $\mathcal{G}^+$ differ by a single arc, so that only the local distribution of X_i differs between the two DAGs. For BDeu, $\alpha_{ijk} = \alpha/(r_i q_i)$ and substituting (3) in (11) we get

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \,|\, \Pi_{X_i}^{\mathcal{G}^-}; \alpha_{ijk}\right) \leqslant \\ \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \,|\, \Pi_{X_i}^{\mathcal{G}^+}; \alpha_{ijk}\right). \quad (12)$$

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 $\Pi_{X_i}^{\mathcal{G}^-}, \Pi_{X_i}^{\mathcal{G}^+}$ in the data. Then the corresponding $n_{ij} = 0$ and Scutari (2016) found that

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

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, we are computing BDeu with an imaginary sample size of $\alpha(\tilde{q}_i/q_i)$ instead of α, since the remaining $\alpha(q_i - \tilde{q}_i)/q_i$ is effectively lost. 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 (12) we are actually using

$$\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \,|\, \Pi_{X_i}^{\mathcal{G}^-}; \alpha_{ijk}\right) \leqslant \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(X_i) \,|\, \mathcal{D}, \alpha_{ijk}\right) \mathrm{BDeu}\left(X_i \,|\, \Pi_{X_i}^{\mathcal{G}^+}; \alpha_{ijk}(\tilde{q}_i/q_i)\right) \quad (13)$$

which is different from (11) and thus not consistent with ME. It is not consistent from a Bayesian perspective either, because $\mathcal{G}^-$ and $\mathcal{G}^+$ are compared with marginal likelihoods arising from different priors; as expected since we know from Giffin and Caticha (2007) that the Bayesian posterior estimation can be derived as a particular case of ME.

As for the posterior expected entropy of $X_i \,|\, \Pi_{X_i}^{\mathcal{G}^+}$, if some $n_{ij} = 0$ then the posterior expected entropy for the uniform prior associated with BDeu becomes

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

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 uniform prior:

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

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 $\Pi_{X_i}^{\mathcal{G}^+}$ almost certainly would 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 are farther from the uniform distribution. This, however, does not necessarily compensate for the overestimation of the entropy of the $q_i - \tilde{q}_i$ unobserved distributions, as we can see in the example below.

Example 1 *Consider a simple example taken from Suzuki (2016), where the sample frequencies n_{ijk} for $X \,|\, \Pi_X^{\mathcal{G}^-}$ 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 $X \,|\, \Pi_X^{\mathcal{G}^+}$ *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

If we take $\alpha = 1$*, then in BDeu* $\alpha_{ijk} = 1/8$ *for* $\mathcal{G}^-$ *and* $\alpha_{ijk} = 1/16$ *for* $\mathcal{G}^+$*, so*

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

and

$$\begin{aligned}
\mathrm{BDeu}\left(X \,|\, \Pi_X^{\mathcal{G}^-}; \frac{1}{8}\right) &= \left(\frac{\Gamma(1/4)}{\Gamma(1/4 + 3)}\left[\frac{\Gamma(1/8 + 3)}{\Gamma(1/8)} \cdot \frac{\cancel{\Gamma(1/8)}}{\cancel{\Gamma(1/8)}}\right]\right)^4 = 0.0326, \\
\mathrm{BDeu}\left(X \,|\, \Pi_X^{\mathcal{G}^+}; \frac{1}{16}\right) &= \left(\frac{\Gamma(1/8)}{\Gamma(1/8 + 3)}\left[\frac{\Gamma(1/16 + 3)}{\Gamma(1/16)} \cdot \frac{\cancel{\Gamma(1/16)}}{\cancel{\Gamma(1/16)}}\right]\right)^4 = 0.0441.
\end{aligned}$$

Therefore we choose $\mathcal{G}^+$ *over* $\mathcal{G}^-$ *both by BDeu alone and by ME 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:*

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

Indeed 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 with BDeu requires that all models are evaluated with the same prior and thus with the same effective imaginary sample size, which clearly is not the case in this example.*

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

Theorem 2 *BDeu and the associated uniform prior over the parameters of the BN may lead to violations of ME if any parent configuration of any node is not observed in the data.*

Neither of the problems described above affects 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 $\alpha_{ijk} = 0$ implies

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

Example 1 (Continued) *If we compare* $X \,|\, \Pi_X^{\mathcal{G}^-}$ *and* $X \,|\, \Pi_X^{\mathcal{G}^+}$ *with 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*

$$\begin{aligned}
&\mathrm{E}\left(\mathrm{H}^{\mathcal{G}^-}(X) \,|\, \mathcal{D}, \frac{1}{8}\right) = \mathrm{E}\left(\mathrm{H}^{\mathcal{G}^+}(X) \,|\, \mathcal{D}, \frac{1}{8}\right) = \\
&\quad = 4\left[\psi_0(1/4 + 3 + 1) - \frac{0 + 1/8}{3 + 1/4}\psi_0(1/8 + 0 + 1) - \frac{3 + 1/8}{3 + 1/4}\psi_0(1/8 + 3 + 1)\right] = 0.3931, \\
&\mathrm{BDs}\left(X \,|\, \Pi_X^{\mathcal{G}^-}; \frac{1}{8}\right) = \mathrm{BDs}\left(X \,|\, \Pi_X^{\mathcal{G}^+}; \frac{1}{8}\right) = \left(\frac{\Gamma(1/4)}{\Gamma(1/4 + 3)}\left[\frac{\Gamma(1/8 + 3)}{\Gamma(1/8)} \cdot \frac{\cancel{\Gamma(1/8)}}{\cancel{\Gamma(1/8)}}\right]\right)^4 = 0.0326;
\end{aligned}$$

which leads to

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

Neither BDeu nor ME express a preference for $\mathcal{G}^-$ *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.*

6. 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 given $\mathcal{G}$. In this paper we studied the properties of the Bayesian posterior scores used to estimate $\mathrm{P}(\mathcal{G} \,|\, \mathcal{D})$ 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 Θ_{X_i} and, in the most general formulation, a hyper-prior over the hyper-parameters α_{ijk} of the prior. We concentrated on the most common BD score, BDeu, which assumes a uniform prior over Θ_{X_i}; and we studied the impact of that prior on structure learning from an information theoretic perspective, looking at possible violations of the maximum entropy principle (ME) when $\mathcal{D}$ is sparse and some parent configurations for at least one node are not observed. After deriving the form of the posterior expected entropy for $\mathcal{G}$ given $\mathcal{D}$, we found that the BD equivalent uniform (BDeu) score may select models in a way that violates ME and that is problematic from a Bayesian perspective. In contrast, the BDs score proposed in Scutari (2016) does not, even though it converges to BDeu asymptotically.

References

G. Anderson and S. Qiu. A Monotonicity Property of the Gamma Function. *Proceedings of the American Mathematical Society*, 125(11):3355–3362, 1997.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 2nd edition, 1985.

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

D. M. Chickering. A Transformational Characterization of Equivalent Bayesian Network Structures. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 87–98, 1995.

G. F. Cooper and E. Herskovits. A Bayesian Method for Constructing Bayesian Belief Networks from Databases. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pages 86–94, 1991.

A. P. Dawid. Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A*, 147(2):278–292, 1984.

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

D. Geiger and D. Heckerman. Learning Gaussian Networks. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 235–243, 1994.

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.

J. Hausser and K. Strimmer. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *Journal of Machine Learning Resesearch*, 10:1469–1484, 2009.

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.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Wiley, 1997.

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

W. Lam and F. Bacchus. Learning Bayesian Belief Networks: an Approach Based on the MDL Principle. *Computational Intelligence*, 10:269–293, 1994.

S. L. Lauritzen and N. Wermuth. Graphical Models for Associations Between Variables, Some of Which are Qualitative and Some Quantitative. *The Annals of Statistics*, 17(1): 31–57, 1989.

D. J. C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

G. A. Miller. Note on the Bias of Information Estimates. In *Information Theory in Psychology II-B*, pages 95–100. Free Press, 1955.

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.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

J. Rissanen. *Information and Complexity in Statistical Models*. Springer, 2007.

M. Scutari. An Empirical-Bayes Score for Discrete Bayesian Networks. *Journal of Machine Learning Research (Proceedings Track, PGM 2016)*, 52:438–448, 2016.

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.

J. Suzuki. Consistency of Learning Bayesian Network Structures with Continuous Variables: An Information Theoretic Approach. *Entropy*, 17:5752–5770, 2015.

J. Suzuki. A Theoretical Analysis of the BDeu Scores in Bayesian Network Structure Learning. *Behaviormetrika*, 44:97–116, 2016.

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.