# Non-Informative Dirichlet Score for learning Bayesian networks

Maomi Ueno
University of Electro-Communications, Japan
ueno@ai.is.uec.ac.jp

Masaki Uto
University of Electro-Communications, Japan
uto_masaki@ai.is.uec.ac.jp

## Abstract

Learning Bayesian networks is known to be highly sensitive to the chosen equivalent sample size (ESS) in the Bayesian Dirichlet equivalence uniform (BDeu). This sensitivity often engenders unstable or undesired results because the prior of BDeu does not represent ignorance of prior knowledge, but rather a user's prior belief in the uniformity of the conditional distribution. This paper presents a proposal for a non-informative Dirichlet score by marginalizing the possible hypothetical structures as the user's prior belief. Some numerical experiments demonstrate that the proposed score improves learning accuracy. The results also suggest that the proposed score might be effective especially for small samples.

## 1 Introduction

Marginal likelihood (ML) (using a Dirichlet prior) that ensures likelihood equivalence, the most popular learning score for Bayesian networks, finds the maximum a posteriori (MAP) structure (Buntine, 1991; Heckerman *et al.*, 1995). This score is known as "Bayesian Dirichlet equivalence (BDe)" (Heckerman *et al.*, 1995). Given no prior knowledge, the Bayesian Dirichlet equivalence uniform (BDeu), as proposed earlier by Buntine (1991), is often used. Actually, BDe(u) requires an "equivalent sample size (ESS)", which reflects the degree of a user's prior belief. Moreover, recent studies have demonstrated that learning Bayesian networks is highly sensitive to the chosen equivalent sample size (ESS) (Steck and Jaakkola, 2002; Silander, Kontkanen and Myllymaki, 2007).

To clarify the BDe(u) mechanism, Ueno (2010) analyzed log-BDe(u) asymptotically, obtaining the result that it is decomposed into the log-likelihood and the penalty term of the complexity, which reflects the difference between the learned structure from data and the hypothetical structure from the user's knowledge. As the two structures become equivalent, the penalty term is minimized with the fixed ESS. Conversely, the term increases to the degree that the two structures become different. Furthermore, the result suggests that a tradeoff exists between the role of ESS in the log-likelihood (which helps to block extra arcs) and its role in the penalty term (which helps to add extra arcs). That tradeoff might cause the BDeu score to be highly sensitive to ESS. It might make it more difficult to determine an approximate ESS.

Moreover, Ueno (2011) showed that the prior of BDeu does not represent ignorance of prior knowledge, but rather a user's prior belief in the uniformity of the conditional distribution. This fact is particularly surprising because it had been believed that BDeu has a non-informative prior. The results further imply that the optimal ESS becomes large/small when the empirical conditional distribution becomes uniform/skewed. This main factor underpins the sensitivity of BDeu to ESS.

To solve this problem, Silander, Kontkanen and Myllymaki (2007) proposed a learning method to marginalize the ESS of BDeu. They averaged BDeu values with increasing ESS from 1 to 100 by 1.0 . Moreover, to decrease the computation costs of the marginalization, Cano et al. (2011) proposed averaging a wide range of different chosen ESSs: $ESS > 1.0$, $ESS >> 1.0$, $ESS < 1.0$, $ESS << 1.0$. They reported that the proposed method performed robustly and efficiently.

However, such approximated marginalization does not always guarantee robust results when the optimal ESS is extremely small or large. In addition, the exact marginalization of ESS is difficult because ESS is a continuous variable in domain $(0, \infty)$.

Our proposal in this paper is a full non-informative Dirichlet score. We assume all possible hypothetical structures as a prior belief of BDe because the problem of BDeu is that it assumes only a uniform distribution as a prior belief. This paper presents a proposal for a non-informative Dirichlet score by averaging the hypothetical structures as a user's prior belief.

The optimal ESS of BDeu becomes large/small when the empirical conditional distribution becomes uniform/skewed because its hypothetical structure assumes a uniform conditional probabilities distribution and the ESS adjusts the magnitude of the user's belief for a hypothetical structure. However, the ESS of full non-informative prior is expected to work effectively as actual pseudo-samples to augment the data, especially when the sample size is small, regardless of the uniformity of empirical distribution. This is a unique feature of the proposed method because the previous non-informative methods exclude the ESS from the score(Averaged BDeu(Silander, Kontkanen and Myllymaki, 2007,Cano et al., 2011), Normalized Maximum Likelihood (NML)(Silander, Roos, and Myllymaki, 2010) ).

Some numerical experiments demonstrate that the proposed score improves learning accuracy. The results also suggest that the proposed score might be effective especially for small samples.

## 2 Learning Bayesian networks

Let $\{x_1, x_2, \cdots, x_N\}$ be a set of $N$ discrete variables, each of which can take a value in the set of states $\{1, \cdots, r_i\}$. Here, $x_i = k$ means that an $x_i$ is state $k$. According to the Bayesian network structure $g \in G$, the joint probability distribution is given as

$$p(x_1, x_2, \cdots, x_N \mid g) = \prod_{i=1}^{N} p(x_i \mid \Pi_i, g), \qquad (1)$$

where $G$ signifies the possible set of Bayesian network structures, and where $\Pi_i$ denotes the parent variables set of $x_i$.

Next, we introduce the problem of learning a Bayesian network. Let $\theta_{ijk}$ be a conditional probability parameter of $x_i = k$ when the $j$-th instance of the parents of $x_i$ is observed (We write $\Pi_i = j$). Buntine (1991) assumed the Dirichlet prior and used an expected a posteriori (EAP) estimator as the parameter estimator $\widehat{\Theta} = (\hat{\theta}_{ijk}), (i = 1, \cdots, N, j = 1, \cdots, q_i, k = 1, \cdots, r_i - 1)$:

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}}, (k = 1, \cdots, r_i - 1), \qquad (2)$$

where $n_{ijk}$ represents the number of samples of $x_i = k$ when $\Pi_i = j$ and $n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$, and where $\alpha_{ijk}$ denotes the hyperparameters of the Dirichlet prior distributions. ($\alpha_{ijk}$ is a pseudo-sample corresponding to $n_{ijk}$), $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$, and $\hat{\theta}_{ijr_i} = 1 - \sum_{k=1}^{r_i - 1} \hat{\theta}_{ijk}$.

The marginal likelihood is obtained as

$$p(\mathbf{X} \mid \alpha, g) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}. \qquad (3)$$

Here, $q_i$ signifies the number of instances of $\Pi_i$ in which $q_i = \prod_{x_l \in \Pi_i} r_l$. In addition, $\mathbf{X}$ is a dataset. The problem of learning a Bayesian network is to find the MAP structure that maximizes the score (3).

Particularly, Heckerman *et al.* (1995) presented a sufficient condition for satisfying the likelihood equivalence assumption, which says that data should not help to discriminate network structures that represent the same assertions of conditional independence, in the form of the following constraint related to hyperparameters of (3):

$$\sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \alpha_{ijk} = const. \qquad (4)$$

Furthermore, they proposed a marginal likelihood that reflects a user's prior knowledge as shown below.

$$p(\mathbf{X} \mid \alpha, g, g^h) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij}^{g^h})}{\Gamma(\alpha_{ij}^{g^h} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}^{g^h} + n_{ijk})}{\Gamma(\alpha_{ijk}^{g^h})} \tag{5}$$

$$\alpha_{ijk}^{g^h} = \alpha p(x_i = k, \Pi_i^g = j \mid g^h) \tag{6}$$

Here, $\alpha$ is the user-determined equivalent sample size (ESS), $\Pi_i^g$ denotes the parent variable sets of $x_i$ according to $g$ and $g^h$ is the hypothetical Bayesian network structure that reflects a user's prior knowledge. This metric was designated as the Bayesian Dirichlet equivalence (BDe) score metric.

As Buntine (1991) described, $\alpha_{ijk}^{g^h} = \frac{\alpha}{(r_i q_i)}$ is regarded as a special case of the BDe metric. Heckerman *et al.* (1995) designated this special case as "BDeu".

For cases of which we have no prior knowledge, BDeu is often used in practice. Heckerman *et al.* (1995) reported, as a result of their comparative analyses of BDeu and BDe, that BDeu is better than BDe unless the user's beliefs are close to the true model. BDeu requires an "equivalent sample size (ESS)", which is the value of a free parameter specified by the user. Recent reports have described that ESS in BDeu plays an important role in learning Bayesian networks (Steck and Jaakkola, 2002; Silander, Kontkanen and Myllymaki, 2007).

Especially, Ueno (2011) reported that the prior of BDeu does not represent ignorance of prior knowledge but rather a user's prior belief in the uniformity of the conditional distribution. This result is particularly surprising because it had been believed that BDeu has a non-informative prior. In addition, he explained the mechanism by which the optimal ESS becomes large/small when the empirical conditional distribution becomes uniform/skewed because ESS determines the magnitude of the user's prior belief for a hypothetical structure. This mechanism causes the BDeu score to be highly sensitive to ESS.

## 3 Non-informative Dirichlet Score

This section presents an alternative non-informative prior Dirichlet score because BDeu has no non-informative prior.

To clarify the mechanism of BDe, Ueno (2010) analyzed the log-BDe asymptotically and derived the following theorem.

**Theorem 1.** *(Ueno 2010) When $\alpha + nt$ is sufficiently large, log-BDe converges to*

$$\log p(\boldsymbol{X} \mid \alpha, g, g^h) = \log p(\widehat{\Theta^g} \mid \mathbf{X}, \alpha, g, g^h) \tag{7}$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} \frac{r_i - 1}{r_i} \log\left(1 + \frac{n_{ijk}}{\alpha_{ijk}^{g^h}}\right) + const,$$

*where* $\log p(\widehat{\Theta^g} \mid \mathbf{X}, \alpha, g, g^h) =$

$$\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} (\alpha_{ijk}^{g^h} + n_{ijk}) \log \frac{(\alpha_{ijk}^{g^h} + n_{ijk})}{(\alpha_{ij}^{g^h} + n_{ij})},$$

*const* is independent terms of the number of parameters, and $\widehat{\Theta^g} = \{\widehat{\theta_{ijk}^g}\}, (i = 1, \cdots, , N, j = 1, \cdots, q_i, k = 1, \cdots, r_i - 1)$,

$$\widehat{\theta_{ijk}^g} = \frac{\alpha_{ijk}^{g^h} + n_{ijk}}{\alpha_{ij}^{g^h} + n_{ij}}. \tag{8}$$

From (7), log-BDe can be decomposed into two factors: 1. a log-posterior term $\log p(\widehat{\Theta^{g^h}} \mid \mathbf{X}, \alpha, g, g^h)$ and 2. a penalty term $\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} \frac{r_i-1}{r_i} \log\left(1 + \frac{n_{ijk}}{\alpha_{ijk}^{g^h}}\right)$. $\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} \frac{r_i-1}{r_i}$ is the number of parameters.

This well known model selection formula is generally interpreted 1. as reflecting the fit to the data and 2. as signifying the penalty that blocks extra arcs from being added.

Ueno (2010) described that the term $\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} \log\left(1 + \frac{n_{ijk}}{\alpha_{ijk}^{g^h}}\right)$ in (7) reflects the difference between the learned structure from data and the hypothetical structure $g^h$ from the user's knowledge in BDe. To the degree that the two structures are equivalent, the penalty term is minimized with the fixed ESS. Conversely, to the degree that the two structures differ, the term is larger. Moreover, from (7), $\alpha$ determines the magnitude of the user's prior belief for a hypothetical structure $g^h$.

On the other hand, BDeu, of which the prior distribution assumes an uniform distribution of conditional probabilities, had been believed to employ a non-informative prior. However, from (7), the optimal ESS of BDeu becomes large/small when the empirical conditional distribution becomes uniform/skewed because its hypothetical structure $g^h$ assumes a uniform conditional probabilities distribution and the ESS adjusts the magnitude of the user's belief for a hypothetical structure. Namely, the prior of BDeu does not represent ignorance of prior knowledge but rather a user's prior belief in the uniformity of the conditional distribution.

The main purpose of this paper is to develop a Dirichlet score that has a full non-informative prior. For this purpose, we assume all possible hypothetical structures as a prior belief of BDe because a problem of BDeu is that it assumes only a uniform distribution as a prior belief.

That is, we marginalize ML over the hypothetical structures $g^h \in G$ as follows:

**Definition 1.** $NIP - BDe$ (Non-informative prior Bayesian Dirichlet equivalence) is defined as

$$p(\mathbf{X} \mid g, \alpha) = \sum_{g^h \in G} p(g^h) p(\mathbf{X} \mid \alpha, g, g^h) \tag{9}$$

$$= \sum_{g^h \in G} p(g^h) \left( \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij}^{g^h})}{\Gamma(\alpha_{ij}^{g^h} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}^{g^h} + n_{ijk})}{\Gamma(\alpha_{ijk}^{g^h})} \right),$$

where $\sum_{g^h \in G}$ is the summation over the possible hypothetical structures, and where $p(g^h)$ is a uniform distribution.

To estimate the marginal ML, it is necessary to calculate $p(x_i = k, \Pi_i^g = j \mid g^h)$ automatically for all the structures. For this purpose, we use an empirical estimation of $\widehat{p}(x_i = k, \Pi_i^g = j \mid g^h, \Theta^{g^h})$ using data. Here, we estimate the joint probability estimate $\widehat{p}(x_i = k, \Pi_i^g = j \mid g^h, \Theta^{g^h})$ which is transformed to the joint probability of $x_i$ and its parents variables in $g$ from the estimated conditional probability parameters $\Theta^{g^h}$ given $g^h$.

Our method has the following two steps:

1. Estimate the conditional probability parameters set $\Theta^{g^h} = \{\theta^{g^h}_{ijk}\}, (i = 1, \cdots, , N, j = 1, \cdots, q_i, k = 1, \cdots, r_i - 1)$ given $g^h$ from data

2. Estimate the joint probability $\widehat{p}(x_i = k, \Pi_i^g = j \mid g^h, \Theta^{g^h})$

First, we estimate the conditional probability parameters set given $g^h$ as

$$\widehat{\theta^{g^h}_{ijk}} = \frac{\frac{1}{r_i q_i^{g^h}} + n^{g^h}_{ijk}}{\frac{1}{q_i^{g^h}} + n^{g^h}_{ij}}, \tag{10}$$

where $n^{g^h}_{ijk}$ represents the number of samples of $x_i = k$ when $\Pi_i^{g^h} = j$ (parent variables set of $x_i$ given $g^h$) and $n^{g^h}_{ij} = \sum_{k=1}^{r_i} n^{g^h}_{ijk}$, and $q_i^{g^h}$ denotes the number of parent variables of $\Pi_i^{g^h}$.

Next, we calculate the estimated joint probability $\widehat{p}(x_i = k, \Pi_i^g = j \mid g^h, \Theta^{g^h})$ as shown below.

$$\widehat{p}(x_i = k, \Pi_i^g = j \mid g^h, \Theta^{g^h}) = \tag{11}$$
$$\sum_{x_l \notin \ x_i \cup \Pi_i^g} p(x_1, \cdots, x_i, \cdots, x_N \mid g^h, \Theta^{g^h})$$

For the computation of (11), we can employ various marginalization algorithms (e.g. variable elimination, factor elimination, random sampling elimination: see, for example, (Darwiche, 2009)).

In practice, however, the Expected log-BDe is difficult to calculate because the product of multiple probabilities suffers serious computational problems. To avoid this, we propose an alternative method, Expected log-BDe, as described below.

**Definition 2.** Expected log-BDe is defined as

$$E_{g^h \in G} \log p(\mathbf{X} \mid \alpha, g, g^h) = \tag{12}$$
$$\sum_{g^h \in G} p(g^h) \log p(\mathbf{X} \mid \alpha, g, g^h)$$
$$= \sum_{g^h \in G} p(g^h) \left( \sum_{i=1}^{N} \sum_{j=1}^{q_i} \log \frac{\Gamma(\alpha_{ij}^{g^h})}{\Gamma(\alpha_{ij}^{g^h} + n_{ij})} \right.$$
$$\left. \sum_{k=1}^{r_i} \log \frac{\Gamma(\alpha_{ijk}^{g^h} + n_{ijk})}{\Gamma(\alpha_{ijk}^{g^h})} \right),$$

Compared to NIP-BDe, the Expected log-BDe is practical for computation because it can be calculated by the sum of log-BDe.

Although a similar Bayesian model averaging criterion has already been proposed (Chickering and Heckerman, 2000), (Tian, He, and Ram, 2010), its purpose is not to predict the true structure but to seek the optimal structure which maximizes the inference prediction of a new data $x_{i+1}$. Therefore, their purpose is not the same as that of this study.

## 4 Learning algorithm using dynamic programming

Learning Bayesian networks is known to be an NP complete problem (Chickering, 1996). Recently, however, the exact solution methods can produce results in reasonable computation time if the variables are not prohibitively numerous (e.g.(Silander and Myllymaki, 2006), Malone et al., 2011).

We employ the learning algorithm of (Silander and Myllymaki, 2006) using dynamic programming. Our algorithm comprises four steps:

1. Compute the local Expected log-BDe scores for all possible $N2^{N-1}$ $(x_i, \Pi_i^g)$ pairs.

2. For each variable $x_i \in \boldsymbol{x}$, find the best parent set in parent candidate set $\{\Pi_i^g\}$ for all $\Pi_i^g \subseteq \boldsymbol{x} \setminus \{x_i\}$, $(\forall g \in G)$.

3. Find the best sink for all $2^N$ variables and the best ordering.

4. Find the optimal Bayesian network

Only Step 1 is different from the procedure described by Silander *et al.*, 2006. First, to obtain the local Expected log-BDe score for a variable and its parent variables set pair, we should average log-BDe for all possible hypothetical structures. To compute the local Expected log-BDe score for each pair $(x_i, \Pi_i^g)$ in Step 1, conditional frequency tables $cft(x_i, \Pi_i^{g^h})$ for all possible hypothesis parent sets $\{\Pi_i^{g^h}\}$ for $(\forall g^h \in G)$are needed.

Algorithm 1 gives a pseudo-code for computing the local Expected log-BDe scores, $LS[x_i][\Pi_i^g]$ , for all possible $(x_i, \Pi_i^g)$ pairs. The pseudo-code assumes some helper functions: $getCft(x_i, \Pi_i^g)$ produces the conditional frequency table $cft(x_i, \Pi_i^g)$, and $getTl(\Pi_i^g, cft[x_i][\Pi_i^{g^h}])$ translates the joint probability estimate $\widehat{p}(x_i = k, \Pi_i^g = j \mid g^h, \Theta^{g^h})$ from $cft[x_i][\Pi_i^{g^h}]$, and the function $score_i(\Pi_i^g, Tl[x_i][\Pi_i^g][\Pi_i^{g^h}])$ calculates the local log-BDe score of $g$ given a hypothetical structure $g^h$.

First, $getCft(x_i, \Pi_i^g)$ produces the conditional frequency tables $cft(x_i, \Pi_i^g)$ for all possible $\Pi_i^g$. They are stored on the memory or the disk as soon as they are produced. This procedure is the same as that of (Silander and Myllymaki, 2006). Next, $getTl(\Pi_i^g, cft[x_i][\Pi_i^{g^h}])$ calculates the joint probability estimate $\widehat{p}(x_i = k, \Pi_i^g = j \mid g^h, \Theta^{g^h})$ using the stored conditional frequency tables by marginalizing the product of factors in $\Theta^{g^h}$ including $\{x_i, \Pi_i^g\}$. This procedure runs in $O(|\{x_i \cup \Pi_i^g\}| \exp(w))$ given an elimination order of width $w$. The local Expected log-BDe, $LS[x_i][\Pi_i^g]$ , can be calculated as the summation of local log-BDe values for all local possible structures $\Pi_i^{g^h}$.

**Algorithm 1** *getLocalScore*$(\boldsymbol{x})$.

```
for all $x_i \in \boldsymbol{x}$ do
    for all $\{\Pi_i^g\} \subseteq \boldsymbol{x} \setminus \{x_i\}$ do
        if $cft[x_i][\Pi_i^g] = null$ then
            $cft[x_i][\Pi_i^g] \leftarrow getCft(x_i, \Pi_i^g)$
        end if
    end for
    for all $\{\Pi_i^g\} \subseteq \boldsymbol{x} \setminus \{x_i\}$ do
        for all $\{\Pi_i^{g^h}\} \subseteq \boldsymbol{x} \setminus \{x_i\}$ do
            $Tl[x_i][\Pi_i^g][\Pi_i^{g^h}] \leftarrow getTl(\Pi_i^g, cft[x_i][\Pi_i^{g^h}])$
            $LS[x_i][\Pi_i^g] \leftarrow LS[x_i][\Pi_i^g] + score_i(\Pi_i^g, Tl[x_i][\Pi_i^g][\Pi_i^{g^h}])$
        end for
        $LS[x_i][\Pi_i^g] \leftarrow LS[x_i][\Pi_i^g] / |\{\Pi_i^g \subseteq \boldsymbol{x} \setminus \{x_i\}\}|$
    end for
end for
if $|\boldsymbol{x}| > 1$ then
    $getLocalScore(\boldsymbol{x} \setminus \{x_i\})$
end if
```

The time complexity of the Algorithm 1 is $O(N^2 2^{2(N-1)} \exp(w))$. After computing the local scores, we find the optimal Bayesian network in Steps 2, 3 and 4 of our algorithm. Steps 2–4 are the same as those proposed by Silander *et al.* (2006). Although the traditional scores (BDeu, AIC, BIC, and so on) run in $O(N2^{(N-1)})$, the proposed method requires greater computation costs. However, the required memory for the proposed computation is equal to that of the traditional scores.

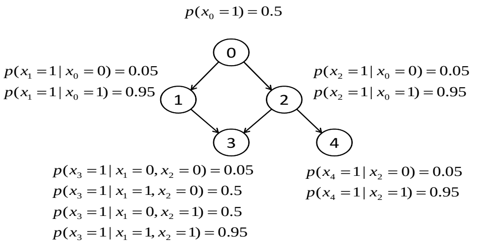
$p(x_0 = 1) = 0.5$
0
$p(x_1 = 1 | x_0 = 0) = 0.05$
$p(x_1 = 1 | x_0 = 1) = 0.95$
1
2
$p(x_2 = 1 | x_0 = 0) = 0.05$
$p(x_2 = 1 | x_0 = 1) = 0.95$
3
4
$p(x_3 = 1 | x_1 = 0, x_2 = 0) = 0.05$
$p(x_3 = 1 | x_1 = 1, x_2 = 0) = 0.5$
$p(x_3 = 1 | x_1 = 0, x_2 = 1) = 0.5$
$p(x_3 = 1 | x_1 = 1, x_2 = 1) = 0.95$
$p(x_4 = 1 | x_2 = 0) = 0.05$
$p(x_4 = 1 | x_2 = 1) = 0.95$

Figure 1: g1: Strongly skewed distribution.

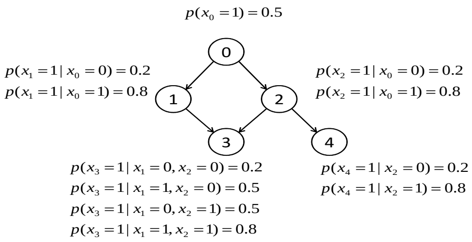
$p(x_0 = 1) = 0.5$
0
$p(x_1 = 1 | x_0 = 0) = 0.2$
$p(x_1 = 1 | x_0 = 1) = 0.8$
1
2
$p(x_2 = 1 | x_0 = 0) = 0.2$
$p(x_2 = 1 | x_0 = 1) = 0.8$
3
4
$p(x_3 = 1 | x_1 = 0, x_2 = 0) = 0.2$
$p(x_3 = 1 | x_1 = 1, x_2 = 0) = 0.5$
$p(x_3 = 1 | x_1 = 0, x_2 = 1) = 0.5$
$p(x_3 = 1 | x_1 = 1, x_2 = 1) = 0.8$
$p(x_4 = 1 | x_2 = 0) = 0.2$
$p(x_4 = 1 | x_2 = 1) = 0.8$

Figure 2: g2: Skewed distribution.

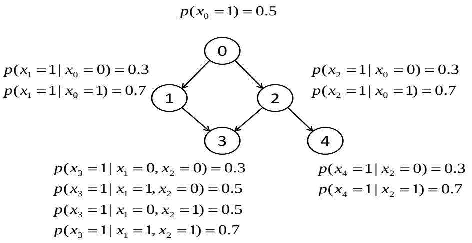
$p(x_0 = 1) = 0.5$
0
$p(x_1 = 1 | x_0 = 0) = 0.3$
$p(x_1 = 1 | x_0 = 1) = 0.7$
1
2
$p(x_2 = 1 | x_0 = 0) = 0.3$
$p(x_2 = 1 | x_0 = 1) = 0.7$
3
4
$p(x_3 = 1 | x_1 = 0, x_2 = 0) = 0.3$
$p(x_3 = 1 | x_1 = 1, x_2 = 0) = 0.5$
$p(x_3 = 1 | x_1 = 0, x_2 = 1) = 0.5$
$p(x_3 = 1 | x_1 = 1, x_2 = 1) = 0.7$
$p(x_4 = 1 | x_2 = 0) = 0.3$
$p(x_4 = 1 | x_2 = 1) = 0.7$

Figure 3: g3: Uniform distribution.

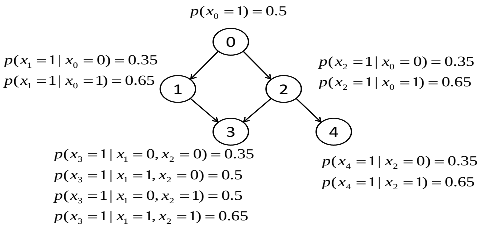
$p(x_0 = 1) = 0.5$
0
$p(x_1 = 1 | x_0 = 0) = 0.35$
$p(x_1 = 1 | x_0 = 1) = 0.65$
1
2
$p(x_2 = 1 | x_0 = 0) = 0.35$
$p(x_2 = 1 | x_0 = 1) = 0.65$
3
4
$p(x_3 = 1 | x_1 = 0, x_2 = 0) = 0.35$
$p(x_3 = 1 | x_1 = 1, x_2 = 0) = 0.5$
$p(x_3 = 1 | x_1 = 0, x_2 = 1) = 0.5$
$p(x_3 = 1 | x_1 = 1, x_2 = 1) = 0.65$
$p(x_4 = 1 | x_2 = 0) = 0.35$
$p(x_4 = 1 | x_2 = 1) = 0.65$

Figure 4: g4: Strongly uniform distribution.

## 5 Simulation experiments

We conducted simulation experiments to compare Expected log-BDe with BDeu according to the procedure described by Ueno (2011). We used small network structures with binary variables in Figs. 1, 2, 3, and 4, in which the distributions are changed from skewed to uniform. Figure 1 presents a structure in which the conditional probabilities differ greatly because of the parent variable states (g1: Strongly skewed distribution). By gradually reducing the difference of the conditional probabilities from Fig. 1, we generated Fig. 2 (g2: Skewed distribution), Fig. 3 (g3: Uniform distribution), and Fig. 4 (g4: Strongly uniform distribution).

Table 1: Learning performance of BDeu

| g1 | BDeu($\alpha = 0.1$) | | | | | | |
|---|---|---|---|---|---|---|---|
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 5 | 233 | 97 | 21 | 12 | 81 | 22 |
| 500 | 76 | 67 | 21 | 11 | 5 | 19 | 11 |
| 1000 | 98 | 5 | 0 | 2 | 0 | 2 | 1 |
| g2 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 5 | 211 | 97 | 9 | 2 | 91 | 12 |
| 500 | 82 | 52 | 8 | 0 | 9 | 19 | 16 |
| 1000 | 99 | 4 | 0 | 0 | 1 | 1 | 2 |
| g3 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 0 | 205 | 131 | 3 | 0 | 69 | 2 |
| 500 | 9 | 194 | 85 | 0 | 2 | 95 | 12 |
| 1000 | 72 | 74 | 16 | 1 | 13 | 27 | 17 |
| g4 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 0 | 253 | 208 | 5 | 0 | 38 | 2 |
| 500 | 0 | 208 | 111 | 1 | 3 | 86 | 7 |
| 1000 | 16 | 184 | 77 | 1 | 3 | 88 | 15 |
| g1 | BDeu($\alpha = 1.0$) | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 34 | 173 | 53 | 17 | 19 | 56 | 28 |
| 500 | 84 | 52 | 5 | 11 | 6 | 16 | 14 |
| 1000 | 96 | 10 | 0 | 4 | 0 | 4 | 2 |
| g2 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 32 | 177 | 54 | 10 | 11 | 69 | 33 |
| 500 | 86 | 45 | 0 | 2 | 11 | 13 | 19 |
| 1000 | 99 | 4 | 0 | 0 | 1 | 1 | 2 |
| g3 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 3 | 227 | 101 | 8 | 5 | 86 | 27 |
| 500 | 33 | 189 | 32 | 1 | 18 | 79 | 59 |
| 1000 | 85 | 49 | 1 | 0 | 13 | 16 | 19 |
| g4 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 1 | 233 | 158 | 7 | 2 | 56 | 10 |
| 500 | 7 | 210 | 84 | 1 | 9 | 93 | 23 |
| 1000 | 41 | 175 | 28 | 0 | 15 | 75 | 57 |
| g1 | BDeu($\alpha = 10$) | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 1 | 458 | 17 | 223 | 11 | 123 | 84 |
| 500 | 24 | 252 | 2 | 112 | 2 | 100 | 36 |
| 1000 | 47 | 176 | 0 | 73 | 2 | 72 | 29 |
| g2 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 40 | 180 | 18 | 43 | 16 | 60 | 43 |
| 500 | 74 | 73 | 0 | 17 | 10 | 26 | 20 |
| 1000 | 87 | 32 | 0 | 14 | 1 | 8 | 9 |
| g3 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 6 | 299 | 46 | 23 | 36 | 87 | 107 |
| 500 | 44 | 200 | 4 | 14 | 32 | 66 | 84 |
| 1000 | 84 | 50 | 0 | 3 | 13 | 15 | 19 |
| g4 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 5 | 290 | 94 | 26 | 19 | 77 | 74 |
| 500 | 22 | 249 | 36 | 10 | 30 | 82 | 91 |
| 1000 | 51 | 174 | 4 | 3 | 24 | 60 | 83 |

Procedures used for the simulation experiments are described below.

1. We generated 200, 500, and 1,000 samples

Table 2: Learning performance of Expected log-BDe

| g1 | Expected log-BDe($\alpha = 0.1$) | | | | | | |
|---|---|---|---|---|---|---|---|
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 5 | 264 | 93 | 40 | 19 | 78 | 34 |
| 500 | 62 | 98 | 19 | 31 | 4 | 32 | 12 |
| 1000 | 91 | 19 | 0 | 9 | 0 | 8 | 2 |
| g2 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 10 | 203 | 91 | 9 | 3 | 86 | 14 |
| 500 | 86 | 41 | 4 | 1 | 8 | 14 | 14 |
| 1000 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| g3 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 0 | 209 | 122 | 4 | 0 | 78 | 5 |
| 500 | 20 | 175 | 70 | 0 | 6 | 81 | 18 |
| 1000 | 81 | 55 | 8 | 1 | 12 | 18 | 16 |
| g4 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 0 | 247 | 192 | 6 | 0 | 45 | 4 |
| 500 | 2 | 204 | 102 | 1 | 3 | 91 | 7 |
| 1000 | 23 | 179 | 66 | 1 | 6 | 82 | 24 |
| g1 | Expected log-BDe($\alpha = 1.0$) | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 34 | 191 | 52 | 30 | 22 | 49 | 38 |
| 500 | 84 | 55 | 5 | 18 | 3 | 22 | 7 |
| 1000 | 92 | 17 | 0 | 8 | 0 | 8 | 1 |
| g2 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 45 | 147 | 36 | 12 | 11 | 56 | 32 |
| 500 | 88 | 37 | 0 | 2 | 9 | 11 | 15 |
| 1000 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| g3 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 4 | 241 | 91 | 12 | 10 | 85 | 43 |
| 500 | 50 | 144 | 20 | 1 | 20 | 55 | 48 |
| 1000 | 87 | 45 | 0 | 1 | 12 | 14 | 18 |
| g4 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 2 | 244 | 141 | 10 | 4 | 62 | 27 |
| 500 | 11 | 223 | 70 | 3 | 18 | 89 | 43 |
| 1000 | 45 | 180 | 19 | 0 | 20 | 69 | 72 |
| g1 | Expected log-BDe($\alpha = 10$) | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 6 | 447 | 24 | 215 | 13 | 87 | 108 |
| 500 | 68 | 112 | 6 | 49 | 2 | 34 | 21 |
| 1000 | 96 | 8 | 0 | 4 | 0 | 4 | 0 |
| g2 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 51 | 126 | 14 | 42 | 11 | 34 | 25 |
| 500 | 82 | 44 | 0 | 10 | 8 | 12 | 14 |
| 1000 | 91 | 22 | 0 | 9 | 0 | 8 | 5 |
| g3 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 14 | 285 | 30 | 42 | 34 | 77 | 102 |
| 500 | 58 | 132 | 1 | 20 | 21 | 38 | 52 |
| 1000 | 84 | 49 | 0 | 6 | 12 | 14 | 17 |
| g4 | | | | | | | |
| n | ○ | SHD | ME | EE | WO | MO | EO |
| 200 | 4 | 324 | 75 | 47 | 25 | 74 | 103 |
| 500 | 25 | 250 | 27 | 19 | 33 | 73 | 98 |
| 1000 | 55 | 158 | 2 | 7 | 22 | 52 | 75 |

from the four figures.

2. Using BDeu and Expected log-BDe by changing $\alpha$ (0.1, 1.0, 10), Bayesian network structures were estimated, respectively, based on 200, 500, and 1,000 samples. We searched for the exactly true structure.

3. The times the estimated structure was the true structure (when Structural Hamming Distance (SHD) is zero; Tsamardinos, Brown, and Aliferis, 2006) were counted by repeating procedure 2 for 100 iterations.

We employ the variable elimination algorithm (Darwiche, 2009) for the marginalization in (11) because our experiments use only a five-node network. Its computational cost is not burdensome.

Table 1 presents the results for BDeu. Table 2 presents results for Expected log-BDe. Column "○ " shows the number of correct-structure estimates in 100 trials. Column "SHD "shows the Structure Hamming Distance (SHD) . Column "ME " shows the total number of missing arcs, column "EE " shows the total number of extra arcs, column "WO " shows the total number of arcs with wrong orientation, column "MO " shows the total number of arcs with missing orientation, and column "EO " shows the total number of arcs with extra orientation in the case of likelihood equivalence. The results presented in Table 1 reveal that BDeu is highly sensitive to $\alpha$.

In Table 1, the optimum value of $\alpha$ becomes small as the conditional distribution becomes skewed. In contrast, the optimum value of $\alpha$ becomes large as the conditional distribution becomes uniform.

As shown in Table 2, the performances of Expected log-BDe are better than those of BDeu in almost all cases. The results also show that the BDeu performances are highly sensitive to $\alpha$ but those of Expected log-BDe are robust for $\alpha$.

Additionally, it is noteworthy that the performance of BDeu is extremely worse than those of Expected log-BDe for g4 with $\alpha = 0.1$ and g1 with $\alpha = 10$. The reason is that the optimal $\alpha$ becomes large/small for uniform/skewed conditional probabilities because BDeu assumes a uniform conditional prior. Although the optimal $\alpha$ for BDeu is highly sensitive to the uniformity of conditional probabilities distribution, the optimal $\alpha$ of the proposed method is robust for the uniformity.

Especially for small samples, the large $\alpha$ setting for the proposed method works effectively because the learning performances of $\alpha = 10$ for $n = 200$ are better than those of $\alpha = 0.1$, which means that the ESS of Expected log-BDe might be affected only by the sample size. The results also suggest that the optimal ESS increases as the sample size becomes small. Therefore, the ESS of the proposed method works effectively as actual pseudo-samples to augment the data, especially when the sample size is small.

## 6 Conclusions

This paper presented a proposal for a non-informative Dirichlet score. The results suggest that the proposed method is effective especially for a small sample size. The future tasks are the following: 1. Analysis of the proposed method using various simulation data. 2. Improvement of the learning algorithm. 3. Determination of the optimal $\alpha$ for a given data size. Furthermore, NML (Silander, Roos, and Myllymaki, 2010) is known as an alternative information-theoretic approach for circumventing the problem with ESS. It is also an important future task to compare the performances of these non-informative learning scores.

## References

W.L. Buntine. 1991. Theory refinement on Bayesian networks. In B. D'Ambrosio, P. Smets and P. Bonissone, (eds.), *Proc. of the Seventh Int. Conf. Uncertainty in Artificial Intelligence*, pages 52–60.

E. Castillo, A.S. Hadi and C. Solares. 1997. Learning and updating of uncertainty in Dirichlet models. *Machine Learning*, **26**, pages 43–63.

D. Chickering. 1995. Learning Bayesian networks is NP-complete. In *Learning from Data – Artificial Intelligence and Statistics*, **V**, pages 121–130.

D.M. Chickering and D. Heckerman. 2000. A comparison of scientific and engineering criteria for Bayesian model selection. *Statistics and Computing*, **10**, pages 55–62.

A. Cano, M. Gomez-Olmedo, A.R. Masegosa, and S. Moral. 2011. Locally Averaged Bayesian Dirichlet Metrics. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty Symbolic and Quantitative Approaches to Reasoning with Uncertainty(Ecsqaru 2011)*, pages 217-228.

A. Darwiche. 2009. Modeling and reasoning with Bayesian networks. Cambridge University Press.

D. Heckerman, D. Geiger and D. Chickering. 1995. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, **20**, pages 197–243.

B.M. Malone, C. Yuan, E.A. Hansen, and S. Bridges. 2011. Improving the Scalability of Optimal Bayesian Network Learning with External-Memory Frontier Breadth-First Branch and Bound Search. In A. Preffer and F.G. Cozman (eds.) *Proc. the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 479-488.

T. Silander and P. Myllymaki. 2006. A simple approach for finding the globally optimal Bayesian network structure. In R. Dechter and T. Richardson (eds.), *Proc. 22nd Conf. on Uncertainty in Artificial Intelligence*, pages 445-452.

T. Silander, P. Kontkanen and P. Myllymaki. 2007. On sensitivity of the MAP Bayesian network structure to the equipment sample size parameter, In K.B. Laskey, S.M. Mahoney and J. Goldsmith (eds.), *Proc. 23rd Conference on Uncertainty in Artificial Intelligence*, pages 360-367.

T. Silander, T. Roos and P. Myllymaki. 2010. Learning Locally Minimax Optimal Bayesian Networks. *International Journal of Approximate Reasoning*, **51**(5): 544-557.

H. Steck and T.S. Jaakkola. 2002. On the Dirichlet Prior and Bayesian Regularization. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems (NIPS)*, pages 697–704.

H. Steck. 2008. Learning the Bayesian network structure: Dirichlet Prior versus Data. D.A. McAllester and P. Myllymaki (eds.), *Proc. 24th Conference on Uncertainty in Artificial Intelligence*, pages 511–518.

J. Tian, R. He and L. Ram. 2010. Bayesian Model Averaging Using the k-best Bayesian Network Structures. In P. Grunwald and P. Spirtes (eds.), *Proc. 26th Conference on Uncertainty in Artificial Intelligence*, pages 589–597.

I. Tsamardinos, L.E. Brown, and C.F. Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, **65**(1), pages 1–78.

M. Ueno. 2010. Learning networks determined by the ratio of prior and data. In P. Grunwald and P. Spirtes (eds.), *Proc. 26th Conference on Uncertainty in Artificial Intelligence*, pages 598–605.

M. Ueno. 2011. Robust learning Bayesian networks for prior belief. In A. Preffer and F.G. Cozman (eds.) *Proc. 27th Conference on Uncertainty in Artificial Intelligence*, pages 698–707.