Branch and Bound for Regular Bayesian Network Structure Learning

Joe Suzuki* **and Jun Kawahara**†
∗ Osaka University, Japan. j-suzuki@sigmath.es.osaka-u.ac.jp
† Nara Institute of Science and Technology, Japan. jkawahara@is.naist.ac.jp

Abstract

We consider efficient Bayesian network structure learning (BNSL) based on scores using branch and bound. Thus far, as a BNSL score, the Bayesian Dirichlet equivalent uniform (BDeu) has been used most often, but it is recently proved that the BDeu does not choose the simplest model even when the likelihood is maximized whereas Jeffreys' prior and MDL satisfy such regularity. Although the BDeu has been preferred because it gives Markov equivalent models the same score, in this paper, we introduce another class of scores (quotient scores) that satisfies the property, and propose a pruning rule for the quotient score based on Jeffreys' prior. We find that the quotient score based on Jeffreys' prior is regular, and that the proposed pruning rule utilizes the regularity, and is applied much more often than that of the BDeu, so that much less computation is required in BNSL. Finally, our experiments support the hypothesis that the regular scores outperform the non-regular ones in the sense of computational efficiency as well as correctness of BNSL.

1 INTRODUCTION

We consider learning stochastic relations among variables from data. If we mean by the relations conditional independence (CI) among variables, and if we express them via a directed acyclic graph (DAG), then such a graphical model will be a Bayesian network (BN) (Pearl, 1988). The factorization of the distribution determines its Bayesian network. Figure 1 shows four BNs consisting of three nodes and two edges. For example, if the distribution is factorized as $P(X)P(Y|X)P(Z|Y)$ and $P(Y)P(X|Y)P(Z|Y)$, then the BN will be (a) and (b), respectively. However, if we express them as $P(X,Y)P(Y,Z)/P(Y)$, we find that they share the same factorization and that both imply the same CI statement: "X and Z are conditionally independent given Y". In this paper, we say the two BNs are Markov equivalent, and do not distinguish them. In a broad sense, the BN is defined in terms of the structure and parameters, i.e., its topology of nodes and edges and the conditional probabilities of variables given other variables.

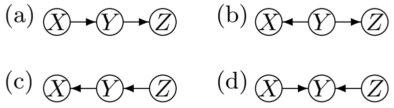

Figure 1: Four BNs with three nodes and two edges.

There are several approaches for Bayesian network structure learning (BNSL). We may test each CI statement between two variable sets given another variable set with the three sets exclusive each other based on an existing statistical method (PC algorithm (Spirtes et al., 1993)). In this paper, however, we focus on the score based approach such as maximizing the posterior probability of a selected structure based on the prior probability and data, or minimizing the description length (MDL (Rissanen, 1978)) of data w.r.t. a selected structure: given data, we compute its score for each structure and select a structure with the optimal value.

We consider four criteria of giving a score to each structure. For many years, the Bayesian Dirichlet equivalent uniform (BDeu) (Buntine, 1991; Heckerman et al., 1995) has been used most often as a criterion that maximizes the posterior probability. The main reason is that the BDeu assures Markov equivalent BNs to have the same score value. Suppose that given n examples w.r.t. X, Y, we assign scores $Q^n(X)$, $Q^n(Y)$, $Q^n(X|Y)$, and $Q^n(Y|X)$ to factors $P(X)$, $P(Y)$, $P(X|Y)$, and $P(Y|X)$, respectively. Then, under the BDeu, $Q^n(X|Y)Q^n(Y) = Q^n(Y|X)Q^n(X)$ holds. The same property holds for any number of variables and any factors. This paper says such BNSL to be normal.

The discussion in this paper is motivated by the paper (Suzuki, 2017) that claims that the BDeu leads to fatal situations in BNSL (see Section 3.1).

In any model selection, given data, simplicity of a model and fitness of the data to the model should be balanced. Any scientific discovery in the history has been found in this way. For example, there would have been many theories that explain mechanics as well as Newton's laws of motion does, but Newton's was selected because it contains only three laws. It is reasonable to think that any belief made by an intelligent activity satisfies such regularity. However, the BDeu violates it. Suppose that given data, we choose a parent set of X from $\{Y\}$ and $\{Y, Z\}$ by comparing $Q^n(X|Y)$ and $Q^n(X|Y, Z)$. Then, the BDeu always chooses $\{Y, Z\}$ when the (empirical) conditional entropy of X given Y is zero (see (Suzuki, 2017) for the proof). In fact, some claim (Silander, 2016) that for small n, the BDeu chooses an over fitted structure.

In this paper, we claim another merit of using the regular scores over the non-regular ones, i.e, BNSL based on the regular scores can be computed more efficiently than BNSL based on the non-regular ones, as explained below.

BNSL consists of finding the optimal parent sets and ordering the variables (Silander and Myllymaki, 2006; Singh and Moore, 2005). We note that as the number of variables grows, the computation exponentially increases (Chickering et al., 2003). For many years, many authors of BNSL have been considering pruning the computation when searching the optimal parent sets in a depth first manner. (Suzuki, 1996) proposed a pruning rule for the MDL principle to reduce the computation; (Tian, 2000) proposed variants of the procedure (Suzuki, 1996); (Campos and Ji, 2011) pointed out that finding the optimal parent sets w.r.t. the MDL principle takes at most polynomial time of p when the sample size n is a constant; (Campos and Ji, 2011) also proposed a pruning rule for the BDeu; and recently, (Suzuki, 2016) proposed a pruning rule for maximizing the posterior probability based on Jeffreys' prior (see Section 3.2 for the details).

Table 1: Normality and regularity (see Sections 2.3 and 3.1 for the definitions, respectively) in the conditional and quotient scores.

Score	BDeu	Jeffreys'
Conditional	normal NOT regular	NOT normal regular
Quotient	N/A	normal regular

In this paper, we point out that BNSL based on the MDL (Suzuki, 1996) and Jeffreys' prior (Suzuki, 2016) are regular, and claims that the pruning rules for those criteria are efficient by utilizing regularity whereas BNSL that maximizes the posterior probability based on Jeffreys' prior (Suzuki, 2016) is not normal.

In order to avoid such inconvenience (not being normal), we consider another framework of constructing scores. For two variables, we only assign four scores $Q^n(\cdot) = 1$, $Q^n(X)$, $Q^n(Y)$, $Q^n(X, Y)$ and compute $Q^n(X|Y) = Q^n(X, Y)/Q^n(Y)$ and $Q^n(Y|X) = Q^n(X, Y)/Q^n(X)$ (quotient scores) rather than assign $Q^n(\cdot) = 1$, $Q^n(X)$, $Q^n(Y)$, $Q^n(X|Y)$, and $Q^n(Y|X)$ (conditional scores). The same idea can be extended into p variables. For the quotient scores, the resulting BNSL will be normal.

In this paper, we propose a pruning rule for the quotient score based on Jeffreys' prior. We prove (Theorem 1) that the resulting BNSL is regular. Although (Suzuki, 2016) has already proposed a pruning rule for the score, we find from Theorem 3 and experiments that the proposed pruning rule significantly improves the existing one (Suzuki, 2016).

Besides, we prove that the minus logarithms of the scores and bounds of the proposed procedure are close to those of the MDL procedure (Theorem 4). While it is known (Campos and Ji, 2011) that the pruning rule for the MDL applies much more often than that for the BDeu, we find that the proposed procedure is much faster than the BDeu as well.

Our contribution is to establish that the regular scores including the quotient ones based on Jeffreys' prior outperform the non-regular ones in the sense of computational efficiency as well as correctness of BNSL.

This paper is organized as follows: Section 2 overviews related matters to the results in Section 3, in particular, Sections 2.1,2.2,2.3 and 2.4 explain BNSL, BDeu, branch and bound, and conditional and quotient scores, respectively; Section 3.1 explains why and how the BDeu faces fatal situation and proves Theorem 1; Section 3.2 claims that the existing pruning rules except the BDeu utilize the fact that the score is regular; Section 3.3 proposes the new bound (Theorem 2), and proves Theorem 3 (it is tighter than the existing one) and Theorem 4; Section 4 shows the results of experiments using the Alarm database (Beinlich et al., 1989); and Section 5 concludes the discussion and raises future works.

2 BACKGROUND

In this section, we overview the notions of Bayesian network structure learning, BDeu, branch and bound for finding the parent sets, and conditional and quotient scores to understand the results in Section 3.

2.1 BAYESIAN NETWORK STRUCTURE LEARNING

Suppose that given n tuples of examples

$$X^{(1)} = x_{i,1}, X^{(2)} = x_{i,2}, \cdots, X^{(p)} = x_{i,p} \ , \qquad (1)$$

$i = 1, 2, \cdots, n$, w.r.t. p variables $X^{(1)}, X^{(2)}, \cdots, X^{(p)}$, we wish to estimate its BN structure that have generated those np examples. We find a BN structure with the maximum posterior probability given the np examples and prior probabilities over structures and parameters (Cooper and Herskovits, 1992).

If X is a binary variable with unknown probability $\theta := P(X=1)$, the probability of a sequence $X = x_1, X = x_2, \cdots, X = x_n$ with c ones and $n-c$ zeros can be expressed by

$$Q^n(X) := \int_0^1 \theta^c (1-\theta)^{n-c} w(\theta) d\theta \qquad (2)$$

using a prior probability $w(\theta)$ over parameter $0 \le \theta \le 1$. It is known (Krichevsky and Trofimov, 1981) that if we assume $w(\theta)$ is proportional to $\theta^{a-1}(1-\theta)^{b-1}$ for real constants $a, b > 0$, (2) can be expressed by

$$Q^n(X) = \prod_{i=1}^{n} \frac{c_{i-1}(x_i) + a(x_i)}{i - 1 + \sum_x a(x)} \ , \qquad (3)$$

where $c_{i-1}(x)$ is the number of occurrences of x in $(x_1, \cdots, x_{i-1}) \in \{0,1\}^{i-1}$, and $a(x) := a$ and $a(x) := b$ for $x = 1$ and for $x = 0$, respectively. For example, if $a = 0.1$, $b = 0.2$, $(x_1, \cdots, x_5) = (0,1,0,1,1)$, then we have

$$Q^5(X) = \frac{0+0.2}{0+0.3} \cdot \frac{0+0.1}{1+0.3} \cdot \frac{1+0.2}{2+0.3} \cdot \frac{1+0.1}{3+0.3} \cdot \frac{2+0.1}{4+0.3} \ .$$

In a similar way, if X, Y are binary variables whose probability is unknown, then the conditional probability of $X = x_1, X = x_2, \cdots, X = x_n$ given $Y = y_1, Y = y_2, \cdots, Y = y_n$ can be expressed by

$$Q^n(X|Y) = \prod_{i=1}^{n} \frac{c_{i-1}(x_i, y_i) + a(x_i, y_i)}{c_{i-1}(y_i) + \sum_x a(x, y_i)} \ , \qquad (4)$$

where $c_{i-1}(x, y)$ is the number of occurrences of $(X, Y) = (x, y)$ in $(X, Y) = (x_1, y_1), \cdots, (x_{i-1}, y_{i-1})$, and $a(x, y) > 0$ is a constant associated with $(X, Y) = (x, y)$. Note that the quantities $Q^n(X)$ and $Q^n(X|Y)$ can be defined even if X and Y are not binary.

If an additional variable Z is available, we can construct $Q^n(X|Y, Z)$ as well as $Q^n(X|Z)$ by regarding (Y, Z) as a single variable that takes a finite number of values. In the same way, we further construct $Q^n(Y|\cdot)$ and $Q^n(Z|\cdot)$, which enables us to express the probabilities of the examples (1) with $p = 3$ and $(X^{(1)}, X^{(2)}, X^{(3)}) = (X, Y, Z)$ given the structures among X, Y, Z. In Figure 1 (a), (b), (c), and (d), those probabilities (scores) are $Q^n(X)Q^n(Y|X)Q^n(Z|Y)$, $Q^n(Y)Q^n(X|Y)Q^n(Z|Y)$, $Q^n(Z)Q^n(Y|Z)Q^n(X|Y)$, and $Q^n(X)Q^n(Z)Q^n(Y|X, Z)$, respectively.

Furthermore, if we have the prior probabilities over the structures, we obtain a BN structure with the maximum posterior probability, where the theory can be extended to p variables each of which takes a finite number of values.

2.2 BDeu

From Section 2.2, we find that $Q^n(X)Q^n(Y|X)Q^n(Z|Y)$, $Q^n(Y)Q^n(X|Y)Q^n(Z|Y)$, and $Q^n(Z)Q^n(Y|Z)Q^n(X|Y)$ should have the same value, which implies

$$Q^n(X)Q^n(Y|X) = Q^n(Y)Q^n(X|Y) \qquad (5)$$

and $Q^n(Y)Q^n(Z|Y) = Q^n(Z)Q^n(Y|Z)$ are required in order for Figure 1 (a)(b)(c) to have the same score. In this paper, we say scores and their BNSL to be *normal* if any Markov equivalent structures share the same score. In particular, in this subsection, we specify $Q^n(\cdot)$ and $Q^n(\cdot|\cdot)$ such that the BNSL is normal.

Let α and β be the numbers of values that X and Y take, respectively. Then, one can check (Buntine, 1991; Heckerman et al., 1995) from (3) and (4) that (5) requires $a(x) = \delta/\alpha$, $a(y) = \delta/\beta$, and $a(x, y) = \delta/\alpha\beta$ for some positive constant δ, if $a(x)$ and $a(x, y)$ take the same values for each of x and for each of (x, y), respectively. Then, we see that (5) has the value

$$\prod_{i=1}^{n} \frac{c_{i-1}(x_i, y_i) + \delta/\alpha\beta}{i - 1 + \delta} \ .$$

Thus, the BDeu gives normal BNSL.

In this paper, we refer to setting $a(\cdot)$ and $a(\cdot, \cdot)$ in this way as the Bayesian Dirichlet equivalent uniform (BDeu) (Buntine, 1991; Ueno, 2008). On the other hand, some use $a(\cdot) = a(\cdot, \cdot) = 0.5$ while the setting does not give normal BNSL as $Q^n(Y)Q^n(X|Y) \neq Q^n(X)Q^n(Y|X)$ for the case. We say that the latter approach is based on Jeffreys' prior (Jeffreys, 1939; Krichevsky and Trofimov, 1981).

2.3 BRANCH AND BOUND FOR FINDING THE PARENT SETS

In order to obtain a BN structure that maximizes the posterior probability, we need to find a subset U of S that maximizes $Q^n(X|U)$ for each $X \in V$ and $S \subseteq V\backslash\{X\}$, where $V := \{X^{(1)}, X^{(2)}, \cdots, X^{(p)}\}$. For computing $R^n_X(S) := \max_{U\subseteq S} Q^n(X|U)$ and $\pi^n(S) := \text{argmax}_{U\subseteq S} Q^n(X|U)$ (the parent set of X w.r.t. S) for $S \subseteq V\backslash\{X\}$, we put $R^n_X(\{\}) = Q^n(X)$, and recursively compute

$$R^n_X(S) = \max_{Y\in S}\{Q^n(X|S), R^n_X(S\backslash\{Y\})\} \qquad (6)$$

for $S \neq \{\}$. Then, we note that the original computation based on dynamic programming takes much time for computing $Q^n(X|S)$, $S \subseteq V\backslash\{X\}$. However, for example, in Figure 2, if we know for $S = \{Y, Z\}$,

$$R^n_X(S) \geq \sup_{T\supseteq S} Q^n(X|T)$$

in some way, we can avoid computing $Q^n(X|Y, Z)$ and $Q^n(X|Y, Z, W)$.

For the BDeu scores, (Campos and Ji, 2011) and (Cussens and Bartlett, 2015) derived

$$\sup_{T\supseteq S} Q^n(X|T) \leq R^n_{X,*}(S) := \alpha^{-d(S)} \ , \qquad (7)$$

where α is the number of values that X takes and $d(S)$ is the number of different states S that actually have occurred in the n tuples of examples. The bound (7) can be applied to branch and bound technique: if $\max_{Y\in S} R^n_X(S\backslash\{Y\}) \geq \alpha^{-d(S)}$ in (6), neither $Q^n(X|S)$ in (6) nor $Q^n(X|T)$ for $T \supsetneq S$ have to be computed.

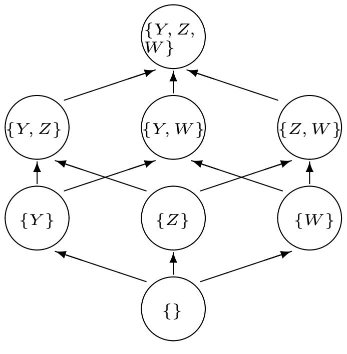

Figure 2: The ordered graph from $\{\}$ to $\{Y, Z, W\}$: compute $\pi^n(S)$ and its maximum value $R^n_X(S)$ in a bottom-up manner.

2.4 CONDITIONAL AND QUOTIENT SCORES

There are some alternative ways to evaluate the probabilities (scores) for the structures. One of the most promising methods (Silander, 2016; Suzuki, 2017) is to compute the local scores $Q^n(S)$ for the $2^p - 1$ subsets S of V to compute global scores by dividing those values (Silander, 2016; Suzuki, 2017). For example, if $S = \{X, Y\}$, then similar to (3), we can compute the score

$$Q^n(X, Y) = \prod_{i=1}^{n} \frac{c_{i-1}(x_i, y_i) + a(x_i, y_i)}{i - 1 + \sum_x a(x, y_i)} \ . \qquad (8)$$

Then, if we have three variables X, Y, Z ($p = 3$), we can compute seven local scores $Q^n(X)$, $Q^n(Y)$, $Q^n(Z)$, $Q^n(Y, Z)$, $Q^n(Z, X)$, $Q^n(X, Y)$, $Q^n(X, Y, Z)$ and eleven global scores such as

$$\frac{Q^n(X, Y)Q^n(Y, Z)}{Q^n(Y)} \ , \ \frac{Q^n(X)Q^n(Z)Q^n(X, Y, Z)}{Q^n(X, Z)}$$

for (a)(b)(c) and (d) in Figure 1, respectively.

We refer the scores in Section 2.1 and the current subsection as the conditional and quotient scores, respectively. For the latter score, we define $Q^n(X|Y)$ by $Q^n(X|Y) := Q^n(X, Y)/Q^n(Y)$, so that the conditions such as (5) are automatically satisfied for any constants $a(\cdot, \cdot)$ (normal BNSL). In particular, we assume $a(\cdot) = a(\cdot, \cdot) = 0.5$ (Jeffreys' prior (Jeffreys, 1939; Krichevsky and Trofimov, 1981)) for the latter score. See Table 1.

Table 2: Dataframe consisting of X, Y, Z with $n = 8$, $\alpha = \gamma = 2$, and $\beta = 4$.

X	0	1	0	1	1	0	1	0
Y	0	1	2	3	3	2	1	0
Z	0	0	1	1	1	1	0	0

It is known (Suzuki, 2016) that the quantities (3), (4), and (8) satisfy

$$\frac{Q^n(X,Y)}{Q^n(Y)} \leq Q^n(X|Y) \tag{9}$$

when $a(\cdot) = a(\cdot,\cdot)$. The quotient score requires to assume its equality in (9).

3 BRANCH AND BOUND FOR REGULAR BNSL

In this section, we propose an improved BNSL pruning rule for the quotient score based on Jeffreys' prior.

3.1 REGULAR BNSL

In this subsection, we illustrate and define the notion of regular BNSL that will be discussed throughout this section.

Suppose that we estimate the parent sets $\pi(X,S)$ of $X \in V$ w.r.t. $S \subseteq V\backslash\{X\}$ from examples. Then, the values of variables in $\pi(X,S)$ determine states, and each example is classified into one of the states. For example, if $\pi(X,S) = \{Y,Z\}$, each example is classified into one of the $\beta\gamma$ states based on the values of (Y,Z) when Y and Z take β and γ values, respectively. Then, the conditional probability of each $X = x$ given each state $(Y,Z) = (y,z)$ is to be estimated. In this sense, as the parent set $\pi(X,S)$ contains more variables, the states are divided into smaller ones that contain fewer examples. In this sense, if X takes only one value in each of the states, it is reasonable to stop dividing the states.

We can see that the BDeu violates such regularity. For example, suppose that for $S = \{Y,Z\}$, we determine either $\pi(X,S) = \{Y\}$ or $\pi(X,S) = \{Y,Z\}$ from the examples in Table 2. We observe that $X = 0$ and $X = 1$ for $Y = 0,2$ and $Y = 1,3$, respectively, so that the states have been fully explained by Y, which means that we should not add any more variable to $\pi(X,S) = \{Y\}$. However, the BDeu further adds Z to $\pi(X,S) = \{Y\}$. In fact, we have for the BDeu with $\delta = 1$,

$$Q^n(X|Y) = (\frac{1/8}{1/4} \cdot \frac{1+1/8}{1+1/4})^4 = (\frac{9}{10})^4 \cdot 2^{-4} \text{ , and}$$

$$Q^n(X|Y,Z) = (\frac{1/16}{1/8} \cdot \frac{1+1/16}{1+1/8})^4 = (\frac{17}{18})^4 \cdot 2^{-4} \text{ ,}$$

so that the latter is larger. However, if we use the quotient score based on Jeffreys' prior, we have $Q^n(Y) = \frac{(3/4)^4}{2\cdot 3 \cdots 9}$, $Q^n(X,Y) = Q^n(Y,Z) = \frac{(3/4)^4}{4\cdot 5 \cdots 11}$, and $Q^n(X,Y,Z) = \frac{(3/4)^4}{8\cdot 9 \cdots 15}$, which means

$$\begin{aligned} \frac{3}{55} = Q^n(X|Y) &= \frac{Q^n(X,Y)}{Q^n(Y)} \\ > \frac{Q^n(X,Y,Z)}{Q^n(Y,Z)} &= Q^n(X|Y,Z) = \frac{1}{39} \text{ .} \end{aligned}$$

Let $(x_1, \cdots, x_n)$ and $(s_1, \cdots, s_n)$ be n realizations of $X \in V$ and $S \subseteq V\backslash\{X\}$, respectively. For example, a realization of $S = \{Y,Z\}$ is that of (Y,Z).

Definition 1 If $Q^n(X|S) \geq Q^n(X|T)$ for any $S \subseteq T$ whenever each $X = x_i$ is uniquely determined by the $S = s_i$, we say scores and their BNSL to be *regular*.

We have seen that the BDeu does not give regular BNSL. On the other hand, for the quotient score based on Jeffreys' prior, we have a positive result:

Theorem 1 The BNSL based on the quotient score based on Jeffreys' prior is regular.

Proof: See Appendix A.

Now we consider the procedure based on the minimum description length (MDL) (Rissanen, 1978) principle (the Bayesian information criterion (Schwarz, 1978)): if we define the empirical conditional entropy of X given Y by

$$H(X|Y) = \sum_x \sum_y \frac{c_n(x,y)}{n} \log \frac{c_n(x,y)}{c_n(y)}$$

with $c_n(y) = \sum_{x'} c_n(x',y)$, then the MDL procedure chooses $\pi(X,S)$ by comparing

$$L^n(\{Y\}) = H(X|Y) + \frac{(\alpha-1)\beta}{2n} \log n \text{ and}$$

$$L^n(\{Y,Z\}) = H(X|Y,Z) + \frac{(\alpha-1)\beta\gamma}{2n} \log n$$

when X, Y, Z take α, β, γ values, respectively: choose either $\pi(X,S) = \{Y\}$ or $\pi(X,S) = \{Y,Z\}$ depending on which description length is smaller. The BNSL satisfies regularity. In fact, among the structures with the empirical conditional entropy zero, it chooses a structure with the smallest number of states.

Table 3: The scores and their pruning bounds in BNSL, where $\delta > 0$ is a constant, $\sigma(S)$ is the number of different values that S takes, and $d(S)$ $(\le \sigma(S))$is the number of values that S actually occurs at least once in the n examples.

Prior	Score $Q^n(X\|S)$	Bound $R^n_*(S)$
BDeu (Campos and Ji, 2011)	$\prod_{i=1}^n \frac{c_{i-1}(x_i,s_i)+\delta/\alpha\sigma(S)}{c_{i-1}(s_i)+\delta/\sigma(S)}$	$\alpha^{-d(S)}$
Conditional Jeffreys' (Suzuki, 2016)	$\prod_{i=1}^n \frac{c_{i-1}(x_i,s_i)+0.5}{c_{i-1}(s_i)+0.5\alpha}$	$\prod_{i=1}^n \frac{c_{i-1}(x_i,s_i)+0.5}{c_{i-1}(x_i,s_i)+0.5\alpha}$
Quotient Jeffreys' (PROPOSED)	$\prod_{i=1}^n \frac{c_{i-1}(x_i,s_i)+0.5}{c_{i-1}(s_i)+0.5} \cdot \prod_{i=1}^n \frac{i-1+0.5\sigma(S)}{i-1+0.5\alpha\sigma(S)}$	$\prod_{i=1}^n \frac{i-1+0.5\sigma(S)}{i-1+0.5\alpha\sigma(S)}$

3.2 BRANCH AND BOUND FOR REGULAR BNSL

In this subsection, we illustrate the branch and bound procedures in Section 2.4 for regular BNSL.

It has been pointed out that the branch and bound procedure based on (7) is not so efficient (Silander, 2016). In particular, for irregular BNSL, even when the empirical conditional entropy of X given S is zero, the procedure continues to seek candidate states with more variables, as we have seen in Section 3.1.

On the other hand, for the MDL procedure in Section 3.1 that gives a regular BNSL, the pruning works as follows (Suzuki, 1996): suppose

$$H(X|\{Y\}) + \frac{(\alpha-1)\beta}{2n}\log n \le \frac{(\alpha-1)\beta\gamma}{2n}\log n \ . \tag{10}$$

Then, we do not have to compute the values of $L^n(T)$ for $T \supseteq \{Y,Z\}$ because $H(X|\{Y,Z\}) \ge 0$, and the number of values that T takes is no less than $\beta\gamma$, so that $L^n(T)$ exceeds the both sides when (10) holds. This means that we can utilize the fact for branch and bound:

$$H(X|S\backslash\{Y\})+\frac{(\alpha-1)\sigma(S)/\beta}{2n}\log n \le \frac{(\alpha-1)\sigma(S)}{2n}\log n$$

for some $Y \in S$ implies

$$\inf_{T\supseteq S}\{H(X|T)+\frac{(\alpha-1)\sigma(T)}{2}\log n\} \ge \frac{(\alpha-1)\sigma(S)}{2}\log n \ ,$$

where $\sigma(T)$ is the number of values that T takes. Note that the bound is obtained by assuming that $X = x_i$ is determined by $S = s_i$, so that $c_{i-1}(s_i) = c_{i-1}(x_i, s_i)$, $i = 1, \cdots, n$ (see Table 3).

Furthermore, for the conditional score based on Jeffreys' prior $a(x, y) = 0.5$, we have $Q^n(X|Y) = \prod_{i=1}^n \frac{c_{i-1}(x_i,y_i)+0.5}{c_{i-1}(y_i)+0.5\alpha}$, which is upperbounded by

$$\sup_{T\supseteq S} Q^n(X|T) \le R^n_{X,*}(S) := \prod_{i=1}^n \frac{c_{i-1}(x_i,s_i)+0.5}{c_{i-1}(x_i,s_i)+0.5\alpha} \tag{11}$$

for $S \subseteq V\backslash\{X\}$.

Note that the bound is obtained by assuming that $X = x_i$ is determined by $S = s_i$, so that $c_{i-1}(s_i) = c_{i-1}(x_i, s_i)$, $i = 1, \cdots, n$ (see Table 3).

3.3 BRANCH AND BOUND FOR THE QUOTIENT SCORES BASED ON JEFFREYS' PRIOR

In this subsection, we seek a similar scenario for the quotient score based on Jeffreys' prior as in Section 3.2.

First of all, the reference (Suzuki, 2016) pointed out that $R^n_{X,*}(S)$ in (11) can be used for pruning unnecessary computation for the quotient score based on Jeffreys' prior as well. In fact, from (9), $R^n_{X,*}(S)$ in (11) is an upperbound for all quotient scores $Q^n(X|T)$ with $T \supseteq S$, so that if $\max_{Y\in S} Q^n(X|S\backslash\{Y\})$ exceeds $R^n_{X,*}(S)$ in (11), no values of $Q^n(X|T)$ with $T \supseteq S$ have to be computed.

In this paper, however, we propose a different bound that will be found to be tighter. From (3) and (8) with $a(\cdot) = a(\cdot,\cdot) = 0.5$, the quantity $Q^n(X|Y) = Q^n(X,Y)/Q^n(Y)$ will be

$$\prod_{i=1}^n \frac{i-1+0.5\beta}{i-1+0.5\alpha\beta}\prod_{i=1}^n \frac{c_{i-1}(x_i,y_i)+0.5}{c_{i-1}(y_i)+0.5} \ ,$$

from which $Q^n(X|S)$ can be expressed by

$$\prod_{i=1}^n \frac{i-1+0.5\sigma(S)}{i-1+0.5\alpha\sigma(S)}\prod_{i=1}^n \frac{c_{i-1}(x_i,s_i)+0.5}{c_{i-1}(s_i)+0.5} \ . \tag{12}$$

We obtain a pruning rule for the quotient score based on Jeffreys' prior:

Theorem 2 For $S \subseteq V\backslash\{X\}$, we have

$$\sup_{T \supseteq S} Q^n(X|T) \le R^n_*(S) := \prod_{i=1}^{n} \frac{i-1+0.5\sigma(S)}{i-1+0.5\alpha\sigma(S)} \ . \tag{13}$$

Note that the bound is obtained by assuming that $X = x_i$ is determined by $S = s_i$, so that $c_{i-1}(s_i) = c_{i-1}(x_i, s_i)$, $i = 1, \cdots, n$ (see Table 3).

Proof of Theorem 2: Checking the inequality $Q^n(X|S) \le R^n_*(S)$ is straightforward. From Proposition 1, when $c_{i-1}(s_i) = c_{i-1}(x_i, s_i)$, $i = 1, \cdots, n$, if we divide the states, no larger value of $Q^n(X|\cdot)$ than $Q^n(X|S)$ is obtained, which completes the proof.

We compare the values of $R^n_{*,X}(S)$ in (11) and (13):

Theorem 3 The value of $R^n_{X,*}(S)$ in (13) is no more than that of (11):

$$\prod_{i=1}^{n} \frac{i-1+0.5\sigma(S)}{i-1+0.5\alpha\sigma(S)} \le \prod_{i=1}^{n} \frac{c_{i-1}(x_i, s_i)+0.5}{c_{i-1}(x_i, s_i)+0.5\alpha}$$

Proof: When $c_{i-1}(x_i, s_i) = c_{i-1}(s_i)$, $i = 1, \cdots, n$, the left hand side is $Q^n(X|Y)$ for the quotient score that is upperbounded by $Q^n(X|Y)$ for the conditional score. On the other hand, the right hand side is an upperbound of $Q^n(X|Y)$ for the conditional score, which completes the proof.

Theorem 3 implies that the obtained bound improves the existing one, and we expect that the computation is reduced for obtaining the optimal parent sets w.r.t. the quotient score based on Jeffreys' prior, unless the overhead for computing the bound $R^n_*(S)$ is too large. In the next section, we evaluate the total computation time as well as the number of subsets S for which $Q^n(X|S)$ are actually computed, and examine that both of them are much improved.

Theorem 4 For (12), we have two equations:

$$-\log\{\prod_{i=1}^{n} \frac{i-1+0.5\sigma(S)}{i-1+0.5\alpha\sigma(S)}\} = \frac{(\alpha-1)\sigma(S)}{2}\log n + \log\frac{\Gamma(\frac{\sigma(S)}{2})}{\Gamma(\frac{\alpha\sigma(S)}{2})} + O(\frac{1}{n}) \tag{14}$$

and

$$-\log\{\prod_{i=1}^{n} \frac{c_{i-1}(x_i, s_i)+0.5}{c_{i-1}(s_i)+0.5}\} = nH(X|S) + O(1) \tag{15}$$

where $\Gamma(\cdot)$ is the Gamma function, and $f = g + O(1)$ implies $f - g$ is bounded by constants from above and below.

Proof: See Appendix B.

Theorem 4 implies that maximizing $Q^n(X|S)$ in (12) is equivalent to minimizing its description length $L^n(S)$ plus $O(1)$ terms, which means that the computation of the quotient score based on Jeffreys' prior is close to that of the MDL procedure. Although many authors of BNSL know that the MDL is an approximation of maximizing the posterior probability w.r.t. some prior probability over the parameters, no efficient pruning rule has been found for maximizing the posterior probability. For example, a pruning rule was proposed by (Campos and Ji, 2011) for the BDeu, the search is not efficient, which will be seen in Section 4 as well. Although (Suzuki, 2016) proposed a pruning rule for the conditional score based on Jeffreys' prior, the BNSL is not normal while the BDeu gives normal BNSL.

4 EXPERIMENTS

In this section, we compare the proposed procedure with the existing ones using the Alarm database (Beinlich et al., 1989) that is a standard benchmark for BNSL.

The algorithm is executed via Rcpp (Eddelbuettel, 2013): each compiled Rcpp procedure runs as an R function almost as fast as when the same procedure runs as a C++ function. The CPU we used in the experiments was Core M-5Y10(Broadwell)/800MHz/2.

The correctness of the BNSL procedures is guaranteed because branch and bound only removes unnecessary computations. Thus, we evaluate for each procedure with pruning computations: the number m of subsets that were actually computed among the 2^{p-1} subsets of $V\backslash\{X\}$; and the execution time for the prepared machine. For the first index, we compare the ratio $r := m/2^{p-1}$; the second one depends on the used machine but it takes into account how large the overhead is: if the computation of the bound was heavy and r is close to one, the execution would take longer than without pruning. However, it does not seem that the relative difference between the procedures depends on the used machine so much.

Because we execute many times we restrict the dataframe to that consisting of the first $n = 200$ and $n = 1000$ rows and first $p = 20$ columns.

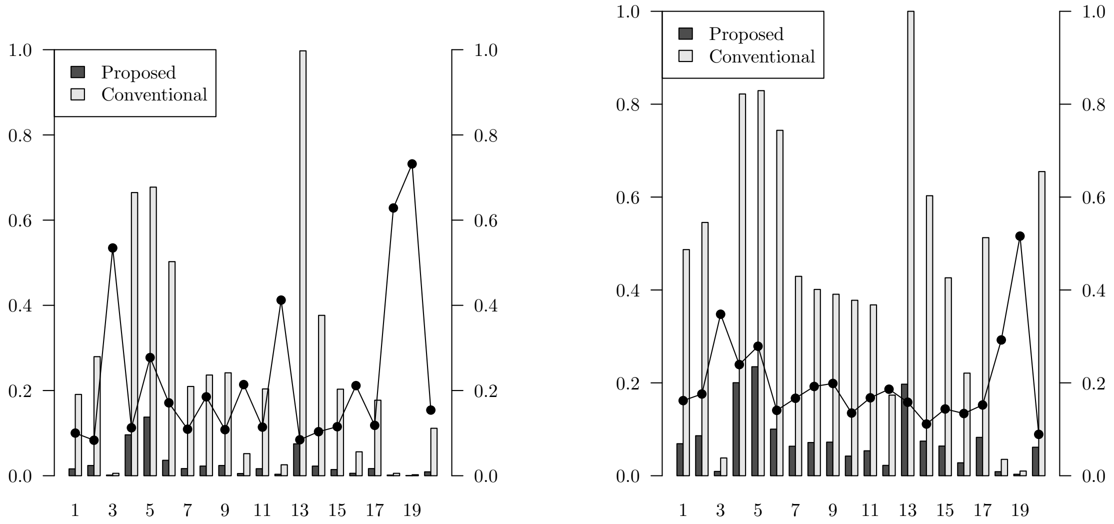

Figure 3: The first experiment for $n = 200$ (Left) and $n = 1000$ (Right): the axis $1 \leq i \leq 20$ indicates the variable $X = X^{(i)}$, and the two bars in each $1 \leq i \leq 20$ indicate the rates r for the proposed and existing procedures while the line graph indicates the ratio of the execution times.

4.1 IMPROVEMENTS IN THE QUOTIENT SCORE BASED ON JEFFREYS' PRIOR

The first experiment examines difference between the existing and proposed bounds for the quotient score based on Jeffreys' prior. As we have seen in Theorem 2, the proposed bound is no larger than the existing one, so that the former prunes unnecessary computation more often than the latter does. But, the actual difference depends on the data, and if we consider the overhead, the execution time for the latter might be small.

Figure 3 shows the two indexes for $n = 200$ (Left) and $n = 1000$ (Right). The axis that ranges over $1 \leq i \leq 20$ indicates the variable $X = X^{(i)}$, and each procedure finds the parent set $\pi(X, S)$ for all $S \subseteq V \backslash \{X\}$. The two bars in each $1 \leq i \leq 20$ indicate the rates r for the proposed and existing procedures, respectively, while the line graph indicates the ratio of the execution times (the proposed divided by the existing) for the 20 variables.

From Figure 3, we see that the numbers of subsets S for which actually $Q^n(X|S)$ were computed for the proposed procedure is less than one tenth and one fifth in many variables for $n = 200$ and $n = 1000$, respectively, and that the execution time ratio is one third and one fifth for $n = 200$ and $n = 1000$, respectively.

In general, how efficiently a pruning rule of branch and bound works depends on the database. In this sense, we cannot say any general statement of the efficiency. However, the significance is rather large, and it seems that the tendency holds for many databases.

4.2 BDeu VS THE QUOTIENT SCORE BASED on JEFFREYS' PRIOR

In this section, we compare the performances among the four criteria: the conditional and quotient scores based on Jeffreys', BDeu, and MDL.

Figure 4 shows the ratio r (Above) and the execution time (Below) for $n = 1000$. The axes that ranges over $1 \leq i \leq 20$ indicates the variable $X = X^{(i)}$, and each procedure finds the parent set $\pi(X, S)$ for all $S \subseteq V \backslash \{X\}$. The four bars in each $1 \leq i \leq 20$ show the performance for the four criteria.

From Figure 4, we have a couple of insights. First of all, the pruning rule for the BDeu is much less efficient compared with those for the other three criteria. It seems that this is due to the fact that the BDeu does not give regular BNSL while the other three criteria do: the BDeu does not stop computing $Q^n(X|T)$ for $T \supsetneq S$ even if the subset S fully divides the examples. Some claim (Silander, 2016) that the BDeu tends to select a complicated structure in particular for small n, which is consistent

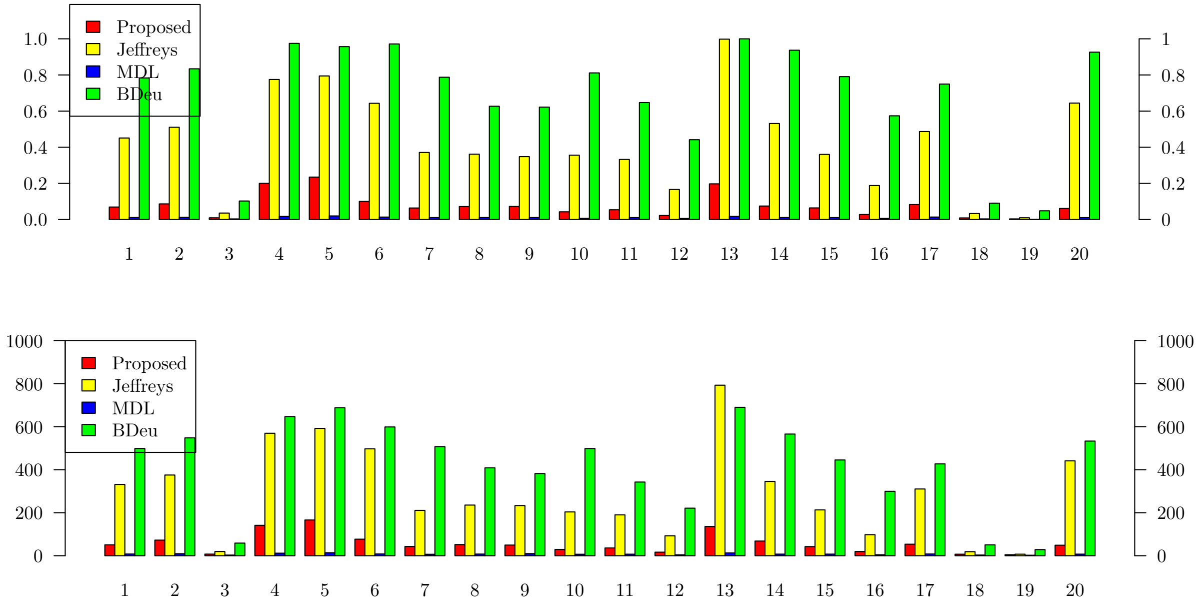

Figure 4: The ratio r (Above) and the execution time (Below) for $n = 1000$: the four bars in each $1 \leq i \leq 20$ shows the performance for the four criteria.

with irregularity of the BDeu.

Secondly, BNSL based on the quotient score based on Jeffreys' prior is much more efficient than the other two Bayesian criteria. It has been known that the pruning bound of the MDL is much more efficient than that of the BDeu (Campos and Ji, 2011). As expected from Theorem 4, the performance of the quotient score based on Jeffreys' prior is close to that of the MDL.

However, the performance is less efficient than the MDL, which seems to be due to the constant term $\log \frac{\Gamma(\sigma(S)/2)}{\Gamma(\alpha\sigma(S)/2)}$ in (14): if $\sigma(S)$ is too large for fixed n, then the bound (14) is too small, so that the pruning rule will be applied less often.

5 CONCLUDING REMARKS

In this paper, we proposed the pruning rule for the quotient score based on Jeffreys' prior. We found that the proposed bound is tighter than the existing one, and examined by the experiments that it runs much faster.

Also, we obtained a novel insight that regular BNSL has a tight pruning bound. We do not have any proof, but have illustrated how the phenomenon occurs, and have experimentally seen that for the three criteria (quotient and conditional Jeffreys' and MDL) that satisfy regularity, the pruning rules are applied more often than for the criterion (BDeu) that is not regularity.

Another significant contribution of this paper is that we obtained an efficient pruning rule for maximizing the posterior probability. Thus far, no efficient pruning rule has been obtained except for the MDL procedure. The result is related to the fact that the score and bound of the quotient score based on Jeffreys' prior behave similarly to those of the MDL procedure.

A problem to consider in the future would be to make the difference clear in correctness between the quotient and conditional scores. In general, each of us has his/her own prior probability over parameters, and a Bayesian solution depends on the belief, so that no one can reject any prior that other holds. However, even if we chose Jeffreys' prior, we see difference in correctness between the quotient and conditional scores. It is worth while considering whether we need to choose one of them.

Appendices A and B are in the Supplementary Material file. They will appear in arXiv math.

References

I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *The 2nd European Conference on Artificial Intelligence in Medicine*, pages 247–256, London, England, 1989. Springer-Verlag.

W. Buntine. Theory refinement on Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 52–60, Los Angels, CA, 1991.

C. P. Campos and Q. Ji. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research*, 12:663–689, 3 2011.

D. M. Chickering, C. Meek, and D. Heckerman. Large-sample learning of Bayesian networks is NP-hard. In *Uncertainty in Artificial Intelligence*, pages 124–133, Acapulco, Mexico, 2003. Morgan Kaufmann.

G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

J. Cussens and M. Bartlett. *GOBNILP 1.6.2 User/Developer Manual1*. University of York, 2015.

D. Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer-Verlag, 2013.

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.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 1939.

R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Information Theory*, IT-27(2):199–207, 1981.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Representation and Reasoning)*. Morgan Kaufmann, 2nd edition, 1988.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

T. Silander. Bayesian network structure learning with a quotient normalized maximum likelihood criterion. In *Proceedings of the Ninth Workshop on Information Theoretic Methods in Science and Engineering*, 2016.

T. Silander and P. Myllymaki. A simple approach for finding the globally optimal Bayesian network structure. In *Uncertainty in Artificial Intelligence*, pages 445–452, Arlington, Virginia, 2006. Morgan Kaufmann.

A. P. Singh and A. W. Moore. Finding optimal Bayesian networks by dynamic programming. Technical report, Carnegie Mellon University, 2005.

P. Spirtes, C. Glymour, and R. Scheines. *Causation,Prediction and Search*. Springer Verlag, Berlin, 1993.

J. Suzuki. Learning Bayesian belief networks based on the minimum description length principle: An efficient algorithm using the b & b technique. In *International Conference on Machine Learning*, pages 462–470, Bari, Italy, 1996. Morgan Kaufmann.

J. Suzuki. An efficient Bayesian network structure learning strategy. *Next Generation Computation Journal*, 1, 2016.

J. Suzuki. A theoretical analysis of the BDeu scores in Bayesian network structure learning. *Behaviormetrika*, 1:1–20, 2017.

J. Tian. A branch-and-bound algorithm for MDL learning Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 580–588, Stanford, CA, 2000. Morgan Kaufmann.

M. Ueno. Learning likelihood-equivalence Bayesian networks using an empirical Bayesian approach. *Behaviormetrika*, 35(2):115–135, 2008.