Constraint-Based Learning Bayesian Networks Using Bayes Factor

Kazuki Natori(✉), Masaki Uto, Yu Nishiyama,
Shuichi Kawano, and Maomi Ueno

Graduate School of Information Systems,
The University of Electro-Communications,
1-5-1, Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
{natori,uto,yu.nishiyama,skawano,ueno}@ai.is.uec.ac.jp

Abstract. A score-based learning Bayesian networks, which seeks the best structure with a score function, incurs heavy computational costs. However, a constraint-based (CB) approach relaxes this problem and extends the available learning network size. A severe problem of the CB approach is its lower accuracy of learning than that of a score-based approach. Recently, several CI tests with consistency have been proposed. The main proposal of this study is to apply the CI tests to CB learning Bayesian networks. This method allows learning larger Bayesian networks than the score based approach does. Based on Bayesian theory, this paper addresses a CI test with consistency using Bayes factor. The result shows that Bayes factor with Jeffreys' prior provides theoretically and empirically best performance.

Keywords: Bayesian networks · Conditional independence test · Jeffreys' prior · Learning Bayesian networks

1 Introduction

A Bayesian network is a probabilistic graphical model that represents relations of random variables using a directed acyclic graph (DAG) and a conditional probability table (Heckerman 1995; Pearl 1988). When a joint probability distribution has the DAG probabilistic structure, it can be decomposed exactly into a product of the conditional probabilities of variables given their parent variables. Therefore, a Bayesian network is guaranteed to provide a good approximation of the joint probability distribution. When we use a Bayesian network, it is necessary to estimate the structure of a Bayesian network from data because it is generally unknown. Estimating the structure is called "learning Bayesian network".

Two approaches can be used for learning Bayesian networks. First are score-based (SB) approaches (Chickering 2002; Cooper and Herskovits 1992; Heckerman 1995; Heckerman *et al.* 1995). The SB approach seeks the best structure with a score function that has consistency with the true DAG structure.

J. Suzuki and M. Ueno (Eds.): AMBN 2015, LNAI 9505, pp. 15–31, 2015.
DOI: 10.1007/978-3-319-28379-1_2

Therefore, this approach is called score-based learning. A popular Bayesian network learning score is the marginal likelihood (ML) score (using a Dirichlet prior over model parameters), which finds the maximum a posteriori (MAP) structure, as described by Buntine (1991) and Heckerman *et al.* (1995). In addition, the Dirichlet prior is known as a distribution which is only likelihood equivalent when certain conditions hold (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, BDeu requires an "equivalent sample size (ESS)", which is the value of a user-specified free parameter. Moreover, it has been demonstrated in recent studies that the ESS plays an important role in the resulting network structure estimate.

Several learning algorithms in this approach have been developed based on dynamic programming (Cowell 2009; Koivisto and Sood 2004; Silander and Myllymaki 2006), A* search (Yuan *et al.* 2011), branch and bound (Malone *et al.* 2011), and integer programming (Cussens 2011; Jaakkola *et al.* 2010). However, the Bayesian network score-based learning is adversely affected by exponential time and NP hard problems (Chickering 1996). Consequently, the SB approach makes it difficult to apply a large network.

Second is a constraint-based (CB) approach. Fundamentally, the solution of the CB approach sequentially checks conditional independence relations among all variables by statistical testing (CI), and directs edges of the structure from observed data. Actually, the CB approach can relax computational cost problems and can extend the available learning network size for learning. Recently, Yahezkel et al. (2009) proposed the recursive autonomy identification (RAI) algorithm. The RAI algorithm decomposes into autonomous sub-structures after the basic solution of CB approaches. This sequence is performed recursively for each sub-structure. The advantage of the RAI algorithm is to be able to minimize the number of parent nodes when using CI tests in the CB approach. The RAI algorithm is, therefore, the highest accuracy in CB approaches. However, the CB approach depends on the threshold of CI test. It has no consistency with the true DAG structure. Traditional CI tests use G^2 or χ^2 test, and mutual information (MI). Recently, several CI tests with a score function have been proposed for learning Bayesian networks. For example, de Campos (2006) proposed a new score function based on MI for CI tests (de Campos 2006). MI shows consistency for the conditional independence relations between two nodes, but it has not proved the strong consistency (van der Vaart 2000).

On the other hand, a Bayes factor is known to have a strong consistency (van der Vaart 2000). The Bayes factor indicates the ratio of the marginal likelihoods for two hypotheses. The marginal likelihood finds the maximum a posteriori (MAP) structure, as described by Buntine (1991) and Heckerman *et al.* (1995). Steck and Jaakkola (2002) proposed a CI test using a Bayes factor that set of BDeu as the marginal likelihood. The CI test does not address the orientation of edges between two variables. To detect the orientation correctly, BDeu adjusts

the number of parameters to be constant. However, this adjustment entails bias of the prior distribution (Ueno 2011).

In addition, Suzuki (2012) proposed a CI test that has strong consistent estimator of mutual information. As the result of the research, the proposed method corresponds to asymptotically a Bayes factor. But the method is only applied in the Chou–Liu algorithm and is not used in the learning Bayesian networks. Suzuki (2015) also proposed a CI test but he did not write how to use the test for learning Bayesian networks.

This study proposes constraint-based learning Bayesian networks using Bayes factor. A Bayes factor consists of the marginal likelihood for conditional joint probability distributions between two variables in Bayesian networks. This paper also shows that the Bayes factor using Jeffreys' prior is theoretically optimal for CI tests of Bayesian network. Clarke and Barron (1994) derived that the minimum risk value of the hyperparameter of Dirichlet prior is 1/2, which is Jeffreys' prior because it minimizes the entropy risk of prior. For a score-based learning Bayesian network, the Jeffreys' prior works worse than BDe(u) does because it does not satisfy the likelihood equivalence property. However, this study shows theoretically that Jeffreys' prior is the optimal for the proposed Bayes factor. In addition, some numerical experiments underscore the effectiveness of the proposed method. This study gives score-based learning for a large Bayesian network including more than 60 variables.

This paper is organized as follows. First, we introduce the learning Bayesian networks in Sect. 2. Section 3 shows traditional CI tests. Section 4 presents the CI test using the Bayes factor with consistency. Section 5 presents the theoretical analyses about the proposed method that is introduced into Sect. 4. Section 6 introduces the recursive autonomy identification algorithm, which is the state-of-the-art algorithm in the CB approach. Section 7 shows experimental evaluations using the RAI algorithm. In these experiments, we review the learning accuracy of the RAI algorithm according to comparison of each CI tests. Section 8 concludes the paper and suggests avenues of future work.

2 Learning Bayesian Networks

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

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

where G is the possible set of Bayesian network structures, and Π_i is the parent variable set of x_i.

Next, we introduce the problem of learning a Bayesian network. Let θ_{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 $\widehat{\Theta} = (\hat{\theta}_{ijk})$ $(i = 1, \cdots, N, j = 1, \cdots, q_i, k = 1, \cdots, r_i - 1)$:

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

Therein, n_{ijk} represents the number of samples of $x_i = k$ when $\Pi_i = j$, $n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$, α_{ijk} denotes the hyperparameters of the Dirichlet prior distributions (α_{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

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

Here, q_i signifies the number of instances of Π_i, where $q_i = \prod_{x_l \in \Pi_i} r_l$ and $\mathbf{X}$ 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 in the form of the following constraint related to hyperparameters of (3):

$$\alpha_{ijk} = \alpha p(x_i = k, \Pi_i = j \mid g^h). \tag{4}$$

Here, α is the user-determined equivalent sample size (ESS); 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, $\alpha_{ijk} = \alpha/(r_i q_i)$ is regarded as a special case of the BDe metric. Heckerman *et al.* (1995) called this special case "BDeu". Actually, $\alpha_{ijk} = \alpha/(r_i q_i)$ does not mean "uniform prior," but "is the same value of all hyperparameters for a variable".

These methods are called a "score based approach." Score-based learning Bayesian networks are hindered by heavy computational costs. However, a conditional independence (CI) based approach is known to relax this problem and to extend the available learning network size.

3 CI Tests

Common means of CI testing are by thresholding conditional mutual information (CMI) or a statistic that measures statistical independence between variables (in Pearson's chi-square or likelihood ratio G-test).

Mutual Information. Mutual Information (MI) between variables X and Y measures the amount of information shared between these variables, which is provided as

$$\mathrm{MI}(X; Y) = \sum_{x \in X, y \in Y} P(x, y) \log\{P(x, y)/(P(x)P(y))\}. \tag{5}$$

It also measures the degree to which uncertainty about Y decreases when X is observed (and vice versa) (Cover and Thomas 1991). Actually, MI is the Kullback–Leibler (KL) divergence between $P(x, y)$ and $P(x)P(y)$ (Cover and Thomas 1991), measuring how much the joint differs from the marginals' product, or how much the variables can be regarded as not independent.

The CMI between X and Y, given a conditioning set $\mathbf{Z}$, is given as

$$\mathrm{CMI}(X;Y \mid \mathbf{Z}) = \sum_{x\in X, y\in Y, z\in \mathbf{Z}} P(x,y,z)\log\{P(x,y \mid z)/(P(x \mid z)P(y \mid z))\}. \tag{6}$$

By definition, $\mathrm{MI}(X;Y)$ and $\mathrm{CMI}(X;Y \mid \mathbf{Z})$ are non-negative. $\mathrm{MI}(X;Y) = 0$ ($\mathrm{CMI}(X;Y \mid \mathbf{Z}) = 0$) if and only if X and Y are independent (given $\mathbf{Z}$). The true MI is unknown. The estimated $\widehat{\mathrm{MI}}$ is larger than MI (Treves and Panzeri 1995), and therefore for independent variables larger than 0. Practically, $\widehat{\mathrm{MI}}$ is compared to a small threshold, ε, to distinguish pairs of dependent and pairs of independent variables (Aliferis *et al.* 2010; Besson 2010; Cheng *et al.* 1999; 2002). If $\widehat{\mathrm{MI}}(X;Y) < \varepsilon$, X and Y are regarded as independent and the edge connecting them is removed. The test for CI using CMI is similar.

Pearson's chi-square and G^2 test Statistical tests compare the null hypothesis that two variables are independent of the alternative hypothesis. If the null is rejected (cannot be rejected), then the edge is learned (removed). A statistic that is asymptotically chi-square distributed is calculated and compared to a critical value. If it is greater (smaller) than the critical value, then the null is rejected (cannot be rejected) (Agresti 2002; Spirtes *et al.* 2000). In Pearson's chi-square test, the statistic X^2_{st} is

$$X^2_{st} = \sum_{x\in X, y\in Y} (O_{xy} - E_{xy})^2/E_{xy} \sim \chi^2_{\mathrm{d.f}=(|X|-1)(|Y|-1)}, \tag{7}$$

where $O_{xy}(E_{xy})$ is the number of records (expected to be if the null was correct) for which $X = x$, $Y = y$, and $|X|$ and $|Y|$ are the corresponding cardinalities. If the null is correct, $P(x, y) = P(x) \cdot P(y), \forall x \in X, y \in Y$. We expect that $E_{xy}/N = (E_x/N) \cdot (E_y/N), \forall x \in X, y \in Y$ and $E_{xy} = E_x \cdot E_y/N$ for E_x and E_y, which are the numbers of records in which $X = x$ and $Y = y$, respectively, and where N is the total number of records. If X^2_{st} is greater than a critical value for a significance value α, $X^2_{st} > X_{\mathrm{d.f}=(|X|-1)(|Y|-1),\alpha}$, then we reject the null hypothesis.

Instead, based on maximum likelihood, if the statistic

$$G^2_{st} = 2\sum_{x\in X, y\in Y} O_{xy}\log(O_{xy}/E_{xy}) \sim \chi^2_{\mathrm{d.f}=(|X|-1)(|Y|-1)} \tag{8}$$

is larger than the previous critical value $G^2_{st} > X^2_{\mathrm{d.f}=(|X|-1)(|Y|-1),\alpha}$, then we reject the null hypothesis.

However, the learning accuracy of the CB approach is less than that of score-based learning because these CI tests have no strong consistency (van der Vaart 2000).

4 Bayes Factor for CI Test

Traditional CI tests have used statistical tests without consistency. Therefore, the traditional CI tests are not guaranteed to obtain the correct structure even when the data size becomes large. In this paper, we propose a CI test with consistency using the Bayes factor to improve the traditional CI test.

The Bayes factor is the ratio of the marginal likelihood (ML) (Kass and Raftery 1995), which finds the maximum a posteriori (MAP) of the statistical model. Therefore, the Bayes factor has asymptotic consistency. For example, the Bayes factor is given as $p(\mathbf{X} \mid g_1)/p(\mathbf{X} \mid g_2)$, where g_1 and g_2 are the hypothetical structures from observed data $\mathbf{X}$. If the value is larger than 1.0, then g_1 is favored more than g_2, else g_2 is favored more than g_1.

Steck and Jaakkola (2002) proposed a CI test using the Bayes factor. In this method, $\mathbf{X}$ presents observed data for only two variables X_1 and X_2 given conditional variables as

$$\log \frac{p(\mathbf{X} \mid g_1)}{p(\mathbf{X} \mid g_2)}. \tag{9}$$

In the CI test, g_1 shows a dependent model in Fig. 1; g_2 shows an independent model in Fig. 2, where C is the conditional variables. When the log-Bayes factor takes a negative value, then the edge between x_1 and x_2 is deleted.

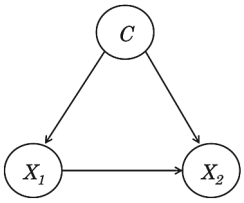

Fig. 1. g_1; dependent model.

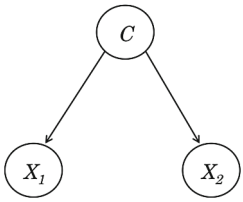

Fig. 2. g_2; independent model.

Steck and Jaakkola (2002) applied BDeu as the marginal likelihoods of the Bayes factor. However, Ueno (2010, 2011) pointed out that BDeu's prior is not non-informative. Especially, BDeu is not guaranteed to optimize CI tests because it was developed for score-based learning Bayesian network. The CI test does not address the orientation of edge between two variables. To detect the orientation correctly, BDeu adjusts the number of parameters to be constant. However, this adjustment causes the bias of the prior distribution (Ueno 2011).

To solve this problem, our approach uses a joint probability distribution of X_1 and X_2 because it is unnecessary to consider the orientation of edge between X_1 and X_2. Let $\theta_{jk_1k_2}$ represent $p(x_1 = k_1, x_2 = k_2 \mid \Pi_{(x_1,x_2)} = j, g_1)$, where $\Pi_{(x_1,x_2)}$ represents a set of common parents variables of x_1 and x_2. Here, $n_{jk_1k_2}$ denotes the number of samples of $x_1 = k_1$ and $x_2 = k_2$ when $\Pi_{(x_1,x_2)} = j$, $n_{k_1k_2} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} n_{jk_1k_2}$. It is noteworthy that $\theta_{jr_1r_2} = 1 - \sum_{k_1=1}^{r_1-1} \sum_{k_2=1}^{r_2-1} \theta_{jk_1k_2}$. Assuming a uniform prior $\alpha_{jk_1k_2} = \alpha$, the marginal likelihood is obtained as

$$p(\mathbf{X}|g_1) = \frac{\Gamma(r_1 r_2 \alpha)}{\Gamma(\alpha)} \prod_{j=1}^{q_i} \prod_{k_1=1}^{r_1} \prod_{k_2=1}^{r_2} \frac{\Gamma(\alpha + n_{jk_1k_2})}{\Gamma(r_1 r_2 \alpha + n_{k_1k_2})}, \tag{10}$$

$$p(\mathbf{X}|g_2) = \prod_{i=1,2} \frac{\Gamma(r_i \alpha)}{\Gamma(\alpha)} \prod_{j=1}^{q_i} \prod_{k_i=1}^{r_i} \frac{\Gamma(\alpha + n_{jk_i})}{\Gamma(r_i \alpha + n_{k_i})}. \tag{11}$$

The remaining problem is determination of the value of hyper-parameter α. Clarke and Barron (1994) described that the optimal minimum risk value of the hyperparameter of the Dirichlet prior is 1/2, which is Jeffreys' prior because it minimizes the entropy risk of prior. Ueno (2010, 2011) claimed that Jeffreys' prior is not efficient for score-based learning Bayesian network. However, this study specifically examines CI tests. The Jeffreys' prior is theoretically optimum for this problem.

Suzuki (2012) proposed a Bayes estimator of the mutual information for extending the Chow–Liu algorithm. The estimator is almost identical to the proposed Bayes factor in this paper. However, their purposes differ because Suzuki (2012) learned probabilistic tree structures to maximize the Bayes estimator.

Suzuki (2015) also proposed a CI test but he did not write how to use the test for learning Bayesian networks. The main proposal of this study is to apply the Bayes factor CI test to CB learning Bayesian networks.

5 Theoretical Analyses

In this section, we present results from some theoretical analyses of CI tests using the proposed method. From (3), the sum of hyperparameters α of BDeu is constant for the number of parents because $\alpha_{ijk} = \alpha/(r_i q_i)$, but that of the proposed method increases as the number of parents increases. For example, one might consider two binary variables with the empty set of C, as shown in Figs. 1 and 2. Then the proposed score for g_1 is calculable by

$$p(\mathbf{X} \mid g_1) = \frac{\Gamma(4\alpha)}{\Gamma(\alpha)} \prod_{k_1=1}^{2} \prod_{k_2=1}^{2} \frac{\Gamma(\alpha + n_{k_1k_2})}{\Gamma(4\alpha + n_{k_1k_2})}.$$

The proposed score for g_2 is obtained as

$$p(\mathbf{X} \mid g_2) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)} \prod_{k_1=1}^{2} \prod_{k_2=1}^{2} \frac{\Gamma(\alpha + n_{k_1k_2})}{\Gamma(2\alpha + n_{k_1k_2})}.$$

The proposed score for g_1 is equivalent to the BDeu score where ESS $= 4\alpha$, but the proposed score for g_2 is equivalent to the BDeu score where ESS $= 2\alpha$. Consequently, from the view of BDeu, the proposed score changes the ESS value according to the number of parameters. From this, the reader might suspect that the proposed method is affected by estimation bias.

To clarify the mechanisms of marginal likelihood of Bayesian network, Ueno (2010) analyzed the log-marginal likelihood asymptotically and derived the following theorem.

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

$$\log p(\mathbf{X} \mid g, \alpha) = \log p(\widehat{\Theta} \mid \mathbf{X}, g, \alpha) - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{r_i - 1}{r_i} \log \left(1 + \frac{n_{ijk}}{\alpha_{ijk}}\right) + const., \tag{12}$$

where

$$\log p(\widehat{\Theta} \mid \mathbf{X}, g, \alpha) = \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} (\alpha_{ijk} + n_{ijk}) \log \frac{(\alpha_{ijk} + n_{ijk})}{(\alpha_{ij} + n_{ij})},$$

and const. is the term that is independent of the number of parameters.

From (12), the log-marginal likelihood can be decomposed into two factors: (1) a log-posterior term $\log p(\widehat{\Theta} \mid \mathbf{X}, g, \alpha)$ and (2) a penalty term $\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{r_i - 1}{r_i} \cdot \log \left(1 + \frac{n_{ijk}}{\alpha_{ijk}}\right)$. $\sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{r_i - 1}{r_i}$ 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. This result suggests that a tradeoff exists between the role of α_{ijk} in the log-posterior (which helps to block extra arcs) and its role in the penalty term (which helps to add extra arcs).

From (12), the value of hyperparameter α_{ijk} should not be changed because the change of α_{ijk} strongly affects the penalty term of the score. The difference between BDeu and the proposed marginal likelihood is that the value of α_{ijk} in BDeu decreases as the number of parameters increases because $\alpha_{ijk} = \alpha/(r_i q_i)$ in BDeu, but that of the proposed method is constant for the different number of parameters. However, we use $\alpha_{ijk} = \alpha/(r_i q_i)$ only for correct orientation identification. Therefore, generally, the decrease of α_{ijk} leading to the increase the number of parameters in BDeu cannot be justified. Consequently, BDeu might show somewhat unstable performance in the CI test.

6 Recursive Autonomy Identification Algorithm

The remaining problem is which CB algorithm we employ to implement the Bayes factor CI test. In this study, we use the recursive autonomy identification (RAI) algorithm which is the state-of-art algorithm for the CB approach. In this section, we present the definition and procedure of the RAI algorithm.

Yehezkel and Lerner (2009) proposed the RAI algorithm to reduce unnecessary CI tests. They show that X and Y which are the variables of structure are independent conditioned on a set of conditional variables $\boldsymbol{S}$ using $X \perp Y \mid \boldsymbol{S}$, and make use of d-separation (Pearl 1988). Also, they define d-separation resolution as the purpose to evaluate d-separation for different the number of conditional variables, and an autonomous substructure.

D-Separation Resolution. The resolution of a d-separation relation between a pair of non-adjacent nodes in a graph is the size of the smallest condition set that d-separates the two nodes.

Exogenous Causes. A node Y in $g(\boldsymbol{V}, \boldsymbol{E})$ is an exogenous cause to $g'(\boldsymbol{V'}, \boldsymbol{E'})$, where $\boldsymbol{V'} \subset \boldsymbol{V}$ and $\boldsymbol{E'} \subset \boldsymbol{E}$, if $Y \notin \boldsymbol{V'}$ and $X \in \boldsymbol{V'}, Y \in \boldsymbol{P}a(X, g)$ or $Y \notin Adj(X, g)$ (Pearl 2000).

Autonomous Sub-structure. In DAG $g(\boldsymbol{V}, \boldsymbol{E})$, a sub-structure $g^A(\boldsymbol{V}^A, \boldsymbol{E}^A)$ such that $\boldsymbol{V}^A \subset \boldsymbol{V}$ and $\boldsymbol{E}^A \subset \boldsymbol{E}$ is said to be autonomous in g given a set $\boldsymbol{V}_{ex} \subset \boldsymbol{V}$ of exogenous causes to g^A if $\forall X \in \boldsymbol{V}^A$, $\boldsymbol{P}a(X, g) \subset \{\boldsymbol{V}^A \cup \boldsymbol{V}_{ex}\}$. If $\boldsymbol{V}_{ex}$ is empty, we say the sub-structure is (completely) autonomous.

They define sub-structure autonomy in the sense that the sub-structure holds the Markov property for its nodes. Given a structure g, any two non-adjacent nodes in an autonomous sub-structure g^A in g are d-separated given nodes either included in the sub-structure g^A or exogenous causes to g^A.

In this method, starting from a complete undirected graph and proceeding from low to high graph d-separation resolution, the RAI algorithm uncovers the correct pattern of a structure by performing the following sequence of operations.

First, all relations between nodes in the structure are checked using the CI test. Second, the edges are directed by orientation rules. Third, structure decomposes autonomous sub-structures. For each sub-structure, the RAI algorithm is applied recursively, while increasing the order of the CI tests. The important idea is that the entire structure decomposes autonomous sub-structures. By performing that procedure, decrease the high order of the CI tests. In the experimentally obtained results, the RAI algorithm was shown to be significant in comparison with other algorithms of the CB approach.

By the procedure, the RAI algorithm is able to realize the computational cost smaller than any other algorithm in the CB approach.

7 Numerical Experiments

This section presents some numerical experiments used to evaluate the effectiveness of our proposed method. For this purpose, we compare the learning accuracy of the proposed method with the other methods.

7.1 Experimental Design

We conducted some simulation experiments to evaluate the effectiveness of the proposed method. In the experiments, we compare the performances of Bayes factor with $\alpha_{ijk} = 1/2$, those with $\alpha_{ijk} = 1$, those with BDeu ($\alpha = 1$) (Steck and Jaakkola 2002), those of de Campos's method (2006), and those of the mutual information with the threshold of 0.003 which is derived as best value by Yehezkel and Lerner (2009). These methods are presented in Table 1.

In Sect. 7.2, we evaluate the performances of CI tests using three small network structures with binary variables. First structure shows a strongly skewed conditional probability distribution. Second has a skewed conditional probability distribution. Third has a uniform conditional probability distribution.

In Sects. 7.3 and 7.4, we present learning results obtained using large networks. We use the Alarm network in Sect. 7.3 and the win95pts network in Sect. 7.4. These benchmark networks were used from the *bnlearn repository* (Scutari 2010).

Table 1. Comparison of methods.

#	Methods
1	$\alpha_{ijk} = \frac{1}{2}$
2	$\alpha_{ijk} = 1$
3	BDeu ($\alpha = 1$)(Steck and Jaakkola 2002)
4	MI & χ^2 (de Campos 2006)
5	MI (Yehezkel and Lerner 2009)

7.2 Experimentation with Small Network

First, we evaluated the learning accuracy using a five-variable structure. Figure 3 has a strongly skewed conditional probability distribution. Figure 4 has a skewed conditional probability distribution. Figure 5 has a uniform conditional probability distribution.

The procedures of this experiment are described below.

1. We generated 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 samples from the three structures.
2. Using CI tests in Table 1, Bayesian network structures were estimated from 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 samples.
3. We repeated procedure 2 for 10 iterations for each number of samples.

We presented the average of the total learning errors for each CI test. The learning error shows the difference between the learned structure and the true structure, which is called the structure Hamming distance (SHD).

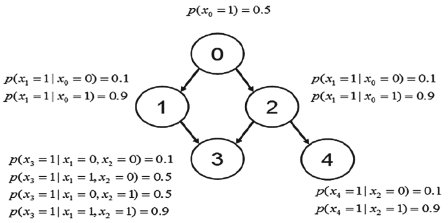

Fig. 3. Strongly skewed distribution.

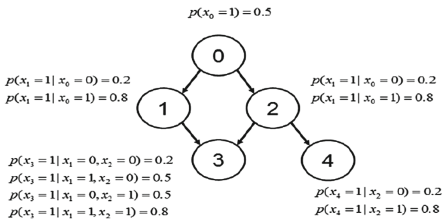

Fig. 4. Skewed distribution.

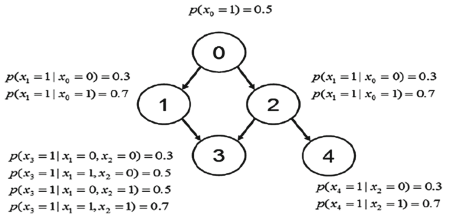

Fig. 5. Uniform distribution.

Tsamardinos *et al.* (2009) proposed the evaluation of the accuracy of the learning structure using the SHD, which is the most efficient metric between the learned and the true structure.

The results are depicted in Fig. 6. The results show that our proposed method (#1) produces the best performance. For a strongly skewed distribution (Fig. 3), our proposed method decreases the learning error faster than $\alpha_{ijk} = 1$ as the sample size becomes large. For a skewed distribution (Fig. 4), our proposed method decreases the learning error faster than $\alpha_{ijk} = 1$ as the sample size becomes large. For a uniform distribution (Fig. 5), all CI tests tend to be adversely affected, showing somewhat unstable behaviors. However, only the method with $\alpha_{ijk} = 1/2$ converges to zero error for a uniform distribution.

From Fig. 6, for a small network, performances with de Campos's method and MI are more adversely affected than those with the other methods because they have no strong consistency.

7.3 Experimentally Obtained Result with the Alarm Network

To evaluate a large network, we first used the Alarm network because it is widely known as a benchmark structure for the evaluation of learning Bayesian networks. The Alarm network includes 37 variables and 46 edges. The maximum in-degree is four. In this experiment, we determined the number of states of all variables as two.

To evaluate the CI test accuracy, we used learning errors of three types (Spirtes *et al.* 2000; Tsamardinos *et al.* 2006). An extra edge (EE) is a learned

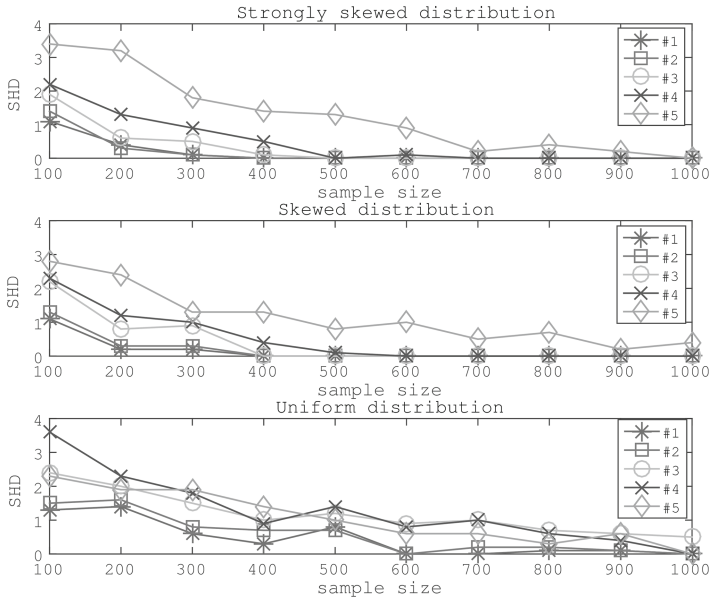

Fig. 6. Results of the learning small network.

edge, although it does not exist in the true graph. A missing edge (ME) is a missed edge from learning, although it exists in the true graph. Additionally, we used SHD.

For evaluation of learning of the Alarm network, we generated $N = 10,000$, 20,000, 50,000, 100,000, and 200,000 samples. Then we let the RAI algorithm with each CI test learn the structure using these samples. We repeated this procedure 10 times. We plot the MEs, EEs, and SHDs of the methods for each sample size to evaluate the learning accuracy in Figs. 7, 8, and 9. Additionally, we show the average of run-time in comparison with the method presented in Table 2.

Table 2. Comparison of the average run-time for each CI method in the Alarm network.

N	Average run-time results (s)				
	#1	#2	#3	#4	#5
10,000	80.9469	80.9974	80.2859	0.7680	0.5558
20,000	167.6280	168.2110	169.8730	1.1758	0.7945
50,000	423.5380	423.7020	424.3510	2.3933	1.6321
100,000	1.8034E+03	1.8283E+03	1.7869E+03	5.8928	4.3668
200,000	4.3404E+03	4.3984E+03	4.3753E+03	9.5591	7.1311

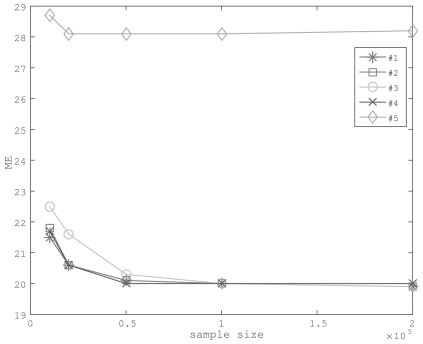

Fig. 7. Average numbers of MEs

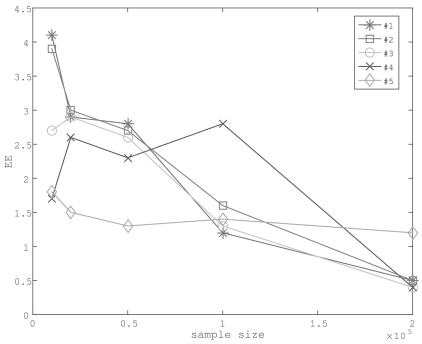

Fig. 8. Average numbers of EEs

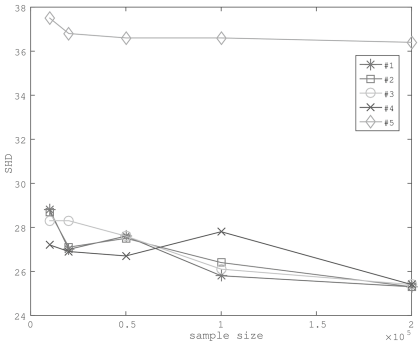

Fig. 9. Average numbers of SHDs

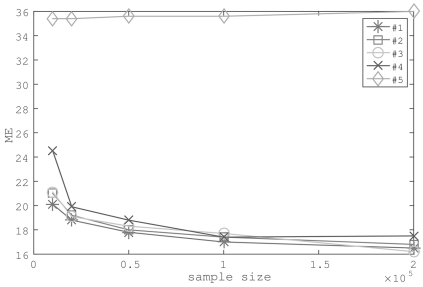

Fig. 10. Average numbers of MEs.

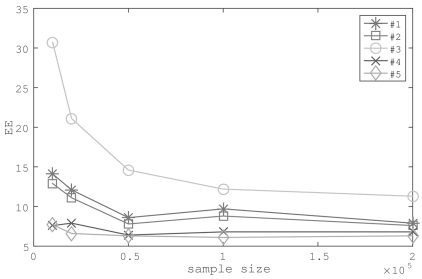

Fig. 11. Average numbers of EEs.

In Table 2, the proposed methods are shown to consume more run-time than the traditional MI methods do. In addition, the run-time of the proposed methods increases linearly as the sample size increases.

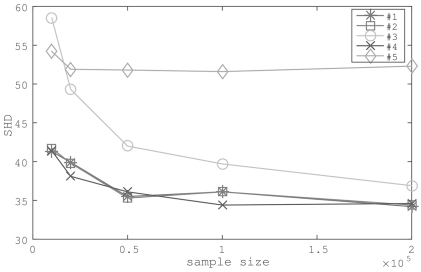

Fig. 12. Average numbers of SHDs.

From Figs. 7 and 9, Bayes factor with $\alpha_{ijk} = 1/2$ outperforms other methods in many cases. Our proposed method tends to be adversely affected more by extra edges for small sample sizes. As the sample size becomes larger than 100,000, the EEs of the proposed method show the best results.

7.4 Experimentally Obtained Results with the Win95pts Network

In the SB approach, Cussens (2011) proposed a learning algorithm using the integer programming and achieved the learning structure with 60 variables. To prove that our proposed method can learn a structure with more than 60 variables, we used the win95pts network. The network includes 76 variables and 112 edges. In addition, the maximum number of degrees is seven.

In this experiment, we also evaluated our proposed method using the same method as that used for learning the Alarm network. We compared the performances of the CI tests for $N = 10,000, 20,000, 50,000, 100,000$, and 200,000 samples. The procedure was repeated 10 times.

In Figs. 10, 11, and 12, we depict the experimentally obtained results from using MEs, EEs, and SHDs. Additionally, we show the average of run-time in comparison with the method presented in Table 3.

Table 3. Comparison of the average run-time for each CI method in the win95pts network.

N	Average run-time results (s)				
	#1	#2	#3	#4	#5
10,000	1.0222e+03	1.0642e+03	986.1200	7.7304	5.2507
20,000	2.1132e+03	1.9826e+03	2.0241e+03	13.5052	8.3541
50,000	4.9998e+03	5.1772e+03	4.7857e+03	21.9171	13.4316
100,000	1.5838e+04	1.5379e+04	1.4425e+04	39.9133	23.6153
200,000	3.3139e+04	3.2942e+04	3.2829e+04	66.8592	36.7520

From Fig. 10, our proposed method (#1) is shown to be the best. From Fig. 11, our proposed method (#1) tends to be adversely affected by extra edges. However, de Campos's method produces fewer extra edges. From Fig. 12, for a small sample size, the Bayes factor with $\alpha_{ijk} = 1$ exhibits superior performance. However, regarding the performance of the proposed method, the Bayes factor with $\alpha_{ijk} = 1$, and de Campos's method show almost identical performance when the sample size becomes large. Actually, de Campos's method without strong consistency provides the best performance because the sample size in this experiment is insufficiently large for this network.

From Table 3, the proposed methods consume more run-time than the traditional MI methods do. In addition, the run-time of the proposed methods increases linearly as the sample size increases. The run-time of the traditional MI methods increases rapidly as the network size increases. Consequently, the proposed method is expected to be applicable to extremely large networks.

8 Conclusion

As described herein, we proposed a new CI test using the Bayes factor with $\alpha_{ijk} = 1/2$ for learning Bayesian networks. Additionally, we provided some theoretical analyses of the proposed method. The results show that the prior distribution of BDeu for score-based learning is not non-informative, and it might cause biased and unstable estimations. The proposed CI test based on Jeffreys' prior minimizes the entropy risk of the prior and optimum the learning results. Using some experiments, we demonstrated that our proposed method improves learning accuracy compared with the other CI tests. Although the CI tests using Bayes Factor based on BDeu (Steck and Jaakkola 2002) have already been proposed, our proposed CI test worked better than the other CI tests did. However, for a large network, we were unable to find a significant difference from the other methods. For a large network, the proposed method requires a large sample size because it has asymptotic consistency.

On a different note, this work indicates that it begins taking a modest step towards improving the theory of the CB approach. A future work is to investigate the performance of the proposed method for larger networks and huge samples.

References

Agresti, A.: Categorical Data Analysis, Wiley Series in Probability and Statistics (2002)

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part I: algorithms and empirical evaluation. J. Mach. Learn. Res. **11**, 171–234 (2010)

van der Vaart, A.W.: Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics (2000)

Besson, P.: Bayesian networks and information theory for audio-visual perception modeling. Biol. Cybern. **103**, 2013–2226 (2010)

Buntine, W.L.: Theory refinement on Bayesian networks. In: Proceedings of the 7th International Conference on Uncertainty in Artificial Intelligence, pp. 52–60 (1991)
de Campos, L.M.: A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. J. Mach. Learn. Res. **7**, 2149–2187 (2006)
Cheng, J., Greiner, R.: Comparing Bayesian network classifiers. In: Proceedings of the 15th International Conference on Uncertainty in Artificial Intelligence, pp. 101–107 (1999)
Cheng, J., Greiner, R., Kelly, J., Bell, D., Liu, W.: Learning Bayesian networks from data: an information-theory based approach. Artif. Intell. **137**, 43–90 (2002)
Chickering, D.M.: Learning Bayesian networks is NP-complete. In: Learning from Data: Artificial Intelligence and Statistics V, pp. 121–130 (1996)
Chickering, D.M.: Optimal structure identification with greedy search. J. Mach. Learn. Res. **3**, 507–554 (2002)
Clarke, B.S., Barron, A.R.: Jeffreys' prior is asymptotically least favorable under entropy risk. J. Stat. Plann. Infer. **41**, 37–60 (1994)
Cover, T.M., Thomas, J.A.: Elements of Information Theory (2nd ed.), Wiley Series in Telecommunications and Signal Processing (1991)
Cooper, G.F., Herskovits, E.A.: Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**, 309–347 (1992)
Cowell, R.G.: Efficient maximum likelihood pedigree reconstruction. Theor. Popul. Biol. **76**(4), 285–291 (2009)
Cussens, J.: Bayesian network learning with cutting planes. In: Cozman, F.G., Pfeffer, A. (eds.), Proceedings of the 27th International Conference on Uncertainty in Artificial Intelligence, pp. 153–160. AUAI Press (2011)
Heckerman, D.: A tutorial on learning with Bayesian networks. In: Technical Report: TR-95-06, Microsoft Research (1995)
Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: The combination of knowledge and statistical data. Mach. Learn. **20**, 197–243 (1995)
Jaakkola, T., Sontag, D., Globerson, A., Meila, M.: Learning Bayesian network structure using LP relaxations. In: 13th International Conference on Artificial Intelligence and Statistics, vol. 9, pp. 358–365 (2010)
Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)
Koivisto, M., Sood, K.: Exact Bayesian structure discovery in Bayesian networks. J. Mach. Learn. Res. **5**, 549–573 (2004)
Malone, B., Yuan, C., Hansen, E., Bridges, S.: Improving the scalability of optimal Bayesian network learning with external-memory frontier breadth-first branch and bound search. In: Cozman, F.G., Pfeffer, A. (eds.), Proceedings of the 27th International Conference on Uncertainty in Artificial Intelligence, pp. 479–488. AUAI Press (2011)
Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan-Kaufmann, San Francisco (1988)
Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, New York (2000)
Scutari, M.: Learning Bayesian networks with the bnlearn R package. J. Stat. Softw. **35**(3), 1–22 (2010)
Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search, 2nd edn. MIT Press, Cambridge (2000)
Silander, T., Myllymaki, P.: A simple approach for finding the globally optimal Bayesian network structure. In: Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence, pp. 445–452 (2006)

Steck, H., Jaakkola, T.S.: On the Dirichlet prior and Bayesian regularization. In: Advances in Neural Information Processing Systems, pp. 697–704. MIT Press, Vancouve (2002)

Suzuki, J.: The Bayesian Chow-Liu algorithm. In: Sixth European Workshop on Probabilistic Graphical Models, pp. 315–322 (2012)

Suzuki, J.: Consistency of learning Bayesian network structures with continuous variables: an information theoretic approach. Entropy **17**(8), 5752–5770 (2015)

Treves, A., Panzeri, S.: The upward bias in measures of information derived from limited data samples. Neural Comput. **7**, 399–407 (1995)

Tsumardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Mach. Learn. **65**(1), 31–78 (2006)

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

Ueno, M.: Robust learning Bayesian networks for prior belief. In: Cozman, F.G., Pfeffer, A. (eds.), Proceedings of the 27th International Conference on Uncertainty in Artificial Intelligence, pp. 698–707. AUAI Press (2011)

Yehezkel, R., Lerner, B.: Bayesian network structure learning by recursive autonomy identification. J. Mach. Learn. Res. **10**, 1527–1570 (2009)

Yuan, C., Malone, B., Wu, X.: Learning optimal Bayesian networks using A* search. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (2011)