LEARNING LIKELIHOOD-EQUIVALENCE BAYESIAN NETWORKS USING AN EMPIRICAL BAYESIAN APPROACH

Article *in* Behaviormetrika · January 2007
DOI: 10.2333/bhmk.35.115

1 author:

Maomi Ueno
The University of Electro-Communications
121 PUBLICATIONS **372** CITATIONS

Behaviormetrika
Vol.35, No.2, 2008, 115–135

LEARNING LIKELIHOOD-EQUIVALENCE BAYESIAN NETWORKS USING AN EMPIRICAL BAYESIAN APPROACH

Maomi Ueno*

Many studies on learning Bayesian networks have used the Dirichlet prior score metric (DPSM). Although they assume different optimum hyper-parameter values for DPSM, few studies have focused on selection of optimum hyper-parameter values. Analyses of DPSM hyper-parameters for learning Bayesian networks are presented here along with the following results: 1. DPSM has a strong consistency for any hyper-parameter values. That is, the score metric DPSM, uniform prior score metric (UPSM), likelihood-equivalence Bayesian Dirichlet score metric (BDe), and minimum description length (MDL) asymptotically converge to the same results. 2. The optimal hyper-parameter values are affected by the true network structure and the number of data. 3. Contrary to Yang and Chang (2002)'s results, BDe based on likelihood equivalence is a theoretically and actually reasonable score metric, if the optimum hyper-parameter values can be found. Using these results, this paper proposes a new learning Bayesian network method based on BDeu that uses the empirical Bayesian approach. The unique features of this method are: 1. It is able to reflect a user's prior knowledge. 2. It has both the strong consistency and likelihood equivalence properties. 3. It finds the optimum hyper-parameter value of BDeu to maximize predictive efficiency, by adapting to domain and data size. In addition, this paper presents some numerical examples using the proposed method that demonstrate the effectiveness of the proposed method.

1. Introduction

Over the last few years, a method of reasoning using probabilities, variously called Bayesian networks, belief networks, or causal networks, has become popular within the AI probability and uncertainty community (see, for example Almond, 1995; Jensen, 2001; Korb and Nicholeson, 2004; Neapolitan, 1990; Pearl, 1988, and so on).

Learning Bayesian networks is one of the basic research topics in the field and concerns building a network structure based on data. Artificial intelligence researchers and statisticians have proposed a variety of scoring methods based on different assumptions to address this issue.

Cooper and Herskovits (1991), Cooper and Herskovits (1992), in the first significant attempt at learning Bayesian networks, assumed a uniform distribution and a general Dirichlet prior score of the Bayesian network model and derived two score metrics which will be referred to as uniform prior score metric (UPSM) and Dirichlet prior score metric (DPSM).

Buntine (1991) also assumed the Dirichlet prior and introduced a hyper-parameter that gives equivalent networks equivalent priors, meaning that marginal priors for individual variables are non-informative.

Spiegelhalter et al. (1993) proposed a score metric using the Bayes factors that use the

Key Words and Phrases: Artificial intelligence, Machine learning, Bayesian network, Bayesian Dirichlet equivalence uniform, likelihood equivalence

∗ Graduate School of Information Systems, The University of Electro-Communications

Dirichlet prior. They recommended the Dirichlet prior's hyper-parameter values, which are higher than 1.0 and propose a hyper-parameter determination method when the original structures are transformed into a junction tree as the basis for efficient computation.

Lam and Bacchus (1994) proposed a minimum description length (MDL) encoding of Bayesian networks. However, their encoding was not a function of the size of the data set, so the code length could not be efficient.

Suzuki (1993) proposed an alternative MDL code for Bayesian networks that is approximated from the DPSM with hyper-parameter $\frac{1}{2}$. This criterion has a strong consistency. Suzuki (1993) and Suzuki (1998) also proved that the DPSM converges to the MDL only when the hyper-parameter values are all $\frac{1}{2}$.

On the other hand, Bouckaert (1994a) and Bouckaert (1994b) also proved that the UPSM (DPSM when the prior distribution has a uniform distribution, or when the hyper-parameter values are all 1.0) converges to the MDL.

However, Suzuki (1998) claimed that Bouckaert (1994a)'s derivation was wrong and that DPSM converges to MDL only when the hyper-parameter values are all $\frac{1}{2}$.

Heckerman, Geiger, and Chickering (1995) proposed the likelihood equivalence assumption and showed that the Dirichlet prior with the constant sum of the hyper-parameters for a variable is a sufficient condition to satisfy the assumption. They pointed out that UPSM does not satisfy the likelihood equivalence assumption and called their new score metric the likelihood-equivalence Bayesian Dirichlet score metric (BDe metric). Buntine (1991)'s hyper-parameter can be interpreted as a special case of the BDe when the prior is assumed to be uniform. Heckerman, Geiger, and Chickering (1995) called Butine's metric the BDeu metric.

Kayaalp and Cooper (2002) also proposed a new score metric called the global uniform (GU) metric, which is a special instance of BDeu in which the constant sum of the hyper-parameters for a variable is equivalent to the sample size.

Yang and Chang (2002) compared the performances of the score metrics UPSM, DPSM, BDe, and MDL using some simulation experiments. Their results showed that the DPSM with hyper-parameter value 10 was best at identifying the true network structure. They also reported that BDe performed worst in their experiment, so assuming likelihood equivalence is unreasonable in learning Bayesian networks.

Thus, there are several different contradictory theories in the Bayesian network learning area that all derive different optimum hyper-parameter values or conditions for DPSM.

Steck and Jaakkola (2002) provided a study of foresight which focused on the hyper-parameter value of BDeu. They showed that asymptocally, as the hyper-parameter value goes to zero, the addition or deletion of an arc in a Bayesian network is infinitely favored or disfavored, and that the preference depends on effective degrees of freedom, a measure that is defined in terms of sufficient statistics, the sumple size of a variable i takes k-th value given a certain parents pattern, that equal zero. They also suggest that in the other extreme, when the hyper-parameter value approaches infinity, the number of arcs in the estimated structure probably increases. These results are significant, but since they are asymptotic they may not sound alarming enough.

Most recently, Silander, Kontakanen, and Myllymaki (2007) provided a series of con-

crete experiments to find the optimum parameter values of BDeu. The results did not give a definite answer and showed that the solution of the network structure is highly sensitive to the chosen hyper-parameter values. This study provided a significant hypothesises that the optimum hyper-parameter value of BDeu cahanges by the kinds of data.

Unfortunately, since their experiments employed several diffrent actual data, the results could not refer to the relationship between the original network structure and the optimum hyper parameter value. In addition, their study forcused on BDeu score and did not compare the performances of the BDeu chainging the hyper-parameter value with the other score metrics.

Therefore, this paper provides some simulation experiments by changing original network structures and the conditional probabilities parameters, and compares the performances with the other score metrics. The results are as follows:

1. DPSM has a strong consistency for any set of hyper-parameters. That is, the score metrics UPSM, DPSM, BDe, and MDL asymptotically converge to the same results.
2. The problem of setting optimal hyper-parameters is affected by the true network structure and the amount of data.
3. Contrary to Yang and Chang (2002)'s results, the BDe based on the likelihood equivalence is theoretically and actually a reasonable score metric.

These results unify and explain the apparently contradictory research on learning Bayesian networks.

This paper also proposes a new Bayesian network learning method based on the BDe from an empirical Bayesian approach. The unique features of this method are:

1. It can reflect a user's prior knowledge since this paper does not limit the prior distribution to the uniform prior as BDeu. If we have a prior structure, we can determine the equivalent sample size of BDe after a joint probability distribution of the prior knowledge is integrated to BDe metric.
2. It has both the strong consistency and likelihood equivalence properties.
3. It finds the optimum hyper-parameter value for DPSM, maximizing the predictive efficiency by adapting to domain and data size.
4. The cross varidation based empirical Bayesian approach can unify the different two problems of Bayesian networks, maximizing the probabilities inferences efficiency given certain evidence and approximating the true joint probabuility distribution, or finding the true network structure.

Finally, this paper presents some numerical examples using the proposed method and demonstrates the effectiveness of the method. It should be noted that Silander, Kontakanen, and Myllymaki (2007) also reffered to the need of an empirical Bayesian approach for learning Bayesian networks to solve the problem which they pointed out.

2. Bayesian networks

Let $U = \{x_1, x_2, \cdots, x_N\}$ be a set of N discrete variables; each can take values in

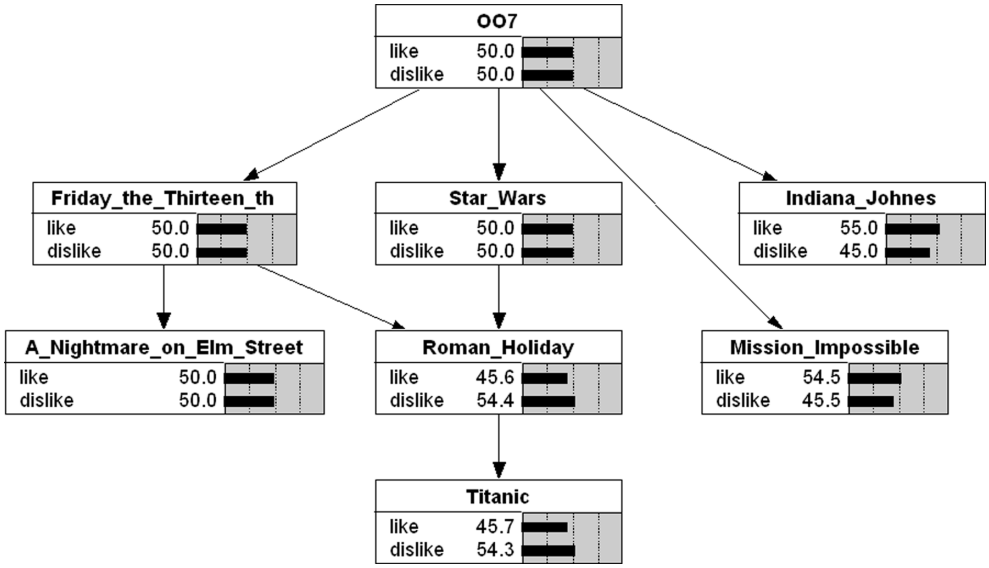

Figure 1: Prior probabilities of a Bayesian network

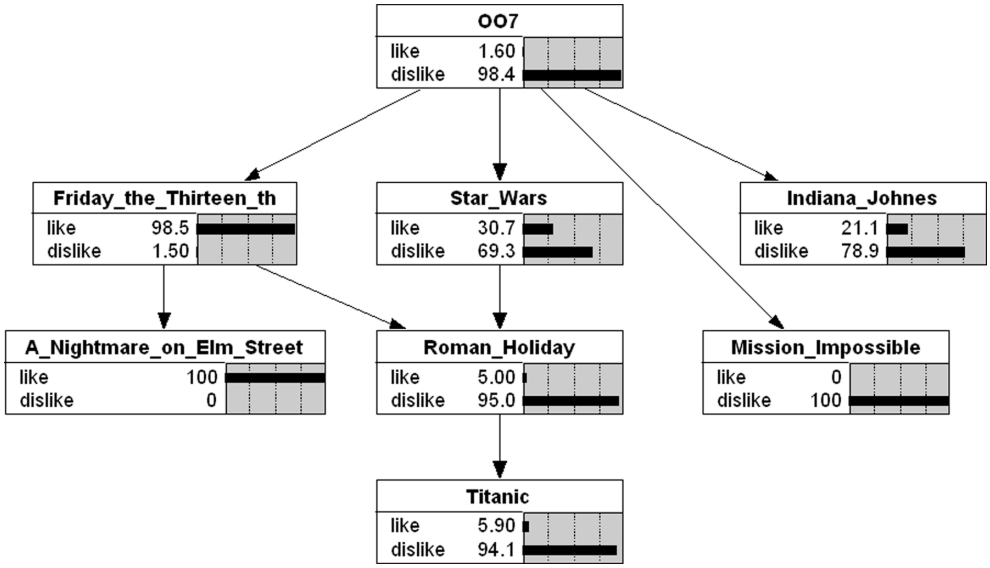

Figure 2: Posterior probabilities of a Bayesian network

the set $\{1, \cdots, r_i\}$. We write $x_i = k$ when we observe that variable x_i is state k. We use $p(x_i = k \mid x_j = k', \xi)$ to denote the probability of a person with background knowledge ξ for observation $x_i = k$ given observation $x_j = k'$. When we observe the state for every variable in set U, we call this set of observations an instance of U. We use $p(Y \mid Z, \xi)$ to denote the set of probabilities for all possible observations of Y given all possible observations of Z, where $Y \subset U$, $Z \subset U$, and $Y \cap Z = \phi$.

A Bayesian network represents a joint probability distribution over domain U by encoding assertions of conditional independence as well as a collection of probability distributions. From the chain rule of probability we know

$$p(x_1, x_2, \cdots, x_N \mid \xi) = \prod_{i=1}^{N} p(x_i \mid x_1, x_2, \cdots, x_{i-1}, \xi). \tag{1}$$

For each variable x_i, let $\Pi_i \subseteq \{x_1, \cdots, x_{i-1}\}$ be a set of variables called parent nodes that renders x_i and $\{x_1, x_2, \cdots, x_{i-1} \setminus \Pi_i\}$ conditionally independent. That is,

$$p(x_i \mid x_1, x_2, \cdots, x_{i-1}, \xi) = p(x_i \mid \Pi_i, \xi). \tag{2}$$

A Bayesian network is represented as a pair of a network structure B_S that encodes the assertions of conditional independence in this equation and a set of conditional probability parameters B_P, (B_S, B_P). Parameter B_S is a directed acyclic graph such that (1) each variable in U corresponds to a node in B_S, and (2) the parents of the node corresponding to x_i are the nodes corresponding to the variables in Π_i. Hereafter, this paper uses x_i to refer to both a variable and its corresponding node in a graph. Associated with node x_i in B_S are the probability distributions $p(x_i \mid \Pi_i, \xi)$. B_P is the union of these distributions. When (1) and (2) are combined, it can be seen that any network for U uniquely determines a joint probability distribution for U. That is,

$$p(x_1, x_2, \cdots, x_N \mid B_S) = \prod_{i=1}^{N} p(x_i \mid \Pi_i, B_S). \tag{3}$$

For example, Figure 1 shows a Bayesian network model for movie preference. Each node indicates a random variable of film preference, and the arcs to node i correspond to the conditional probabilities table $\{p(x_i|\Pi_i, B_S)\}$. The prior probabilities of root node *007* are decided, then the prior probabilities over the network are given as shown in Figure 1. If we know that a person likes *A Nightmare on Elm Street* and dislikes *Mission Impossible*, as shown in Figure 2, the probabilities over the networks are propagated by applying the Bayes theorem given the evidences for the nodes *A Nightmare on Elm Street* and *Mission Impossible*. The posterior probabilities given the evidence that a person likes *A Nightmare on Elm Street* and dislikes *Mission Impossible* are propagated as shown in Figure 2. We know that all prior probabilities are around 0.5 in Figure 1, but the propagated probabilities in Figure 2 show with high probability that the person prefers horror movies. The main problem, which this paper focuses on, is using score metrics methods to construct the network structures, as shown in Figure 1.

The following sections introduce a method of estimating network structure called learning Bayesian networks.

3. Dirichlet-multinomial model

This section introduces the Dirichlet-multinomial model (Cooper and Herskovits, 1991, 1992; Heckerman, Geiger, and Chickering, 1995) which is parameterized from (3).

Let θ_{ijk} be a conditional probability parameter of $x_i = k$ when jth instance of the parents of x_i is observed (We write $\Pi_i = j$), then the likelihood $L(\Theta_{B_S} \mid \mathbf{X})$ is given by

$$L(\Theta_{B_S} \mid \mathbf{X}, B_S) \propto \prod_{i=1}^{N} \prod_{j=1}^{q_i} \prod_{k=0}^{r_i-1} \theta_{ijk}^{n_{ijk}}, \tag{4}$$

where $\Theta_{B_S} = (\theta_{ijk})(i = 1, \cdots, N, j = 1, \cdots, q_i, k = 0, \cdots, r_i - 1)$, r_i is the number of states of x_i, q_i is the number of instances of $\prod_i$, $q_i = \prod_{x_l \in \Pi_i} r_l$, n_{ijk} is the number of samples of $x_i = k$ when $\Pi_i = j$, and $\mathbf{X}$ is a multinominal sample from Bayesian network $p(B_S, B_P)$.

If the resulting posterior distributions are in the same family, it is known that the prior distribution has a Dirichlet distribution as a conjugate prior, which is a class of likelihood functions of the multinomial distribution.

$$p(\Theta_{B_S}) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \prod_{k=0}^{r_i-1} \frac{\Gamma(\sum_{k=0}^{r_i-1} n'_{ijk})}{\prod_{k=0}^{r_i-1} \Gamma(n'_{ijk})} \prod_{k=0}^{r_i-1} \theta_{ijk}^{n'_{ijk}-1}, \tag{5}$$
$$n'_{ijk} > 0 (k = 0, \ldots, r_i - 1),$$

where Γ is the *Gamma function*, which satisfies $\Gamma(x + 1) = x\Gamma(x)$, and n'_{ijk} is the hyperparameter of the prior distribution corresponding to multinominal sample n_{ijk}.

Consequently, we obtain the posterior as follows:

$$p(\Theta_{B_S} \mid \mathbf{X}, B_S) \propto \prod_{i=1}^{N} \prod_{j=1}^{q_i} \prod_{k=0}^{r_i-1} \theta_{ijk}^{n'_{ijk}+n_{ijk}-1}. \tag{6}$$

Thus, if the prior distribution for Θ_{B_S} has a Dirichlet distribution, then so does the posterior distribution for Θ_{B_S}.

Given the Dirichlet distribution's properties, Cooper and Herskovits (1992), Heckerman, Geiger, and Chickering (1995) employed an unbiased estimator, the expectation of the parameter θ_{ijk} as the parameter estimator $\widehat{\theta_{ijk}}$. That, is

$$\widehat{\theta_{ijk}} = \frac{n'_{ijk} + n_{ijk}}{n'_{ij} + n_{ij}}, (k = 0, \cdots, r_i - 2), \tag{7}$$

where $n'_{ij} = \sum_{k=0}^{r_i-1} n'_{ijk}, n_{ij} = \sum_{k=0}^{r_i-1} n_{ijk}, \widehat{\theta_{ij(r_i-1)}} = 1 - \sum_{k=0}^{r_i-2} \widehat{\theta_{ijk}}$.

The predictive distribution is obtained as follows:

$$\begin{aligned} p(\mathbf{X} \mid B_S) &= \int_{\Theta_{B_S}} p(\mathbf{X} \mid \Theta_{B_S}, B_S) p(\Theta_{B_S}) d\Theta_{B_S} \\ &= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(n'_{ij})}{\Gamma(n'_{ij} + n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(n'_{ijk} + n_{ijk})}{\Gamma(n'_{ijk})} \end{aligned} \tag{8}$$

This criterion is called the Dirichlet prior score metric, (DPSM). The estimated structure of the Bayesian network can be obtained by maximizing the DPSM from data. To put it more precisely, it should be noticed that this criterion is not a predictive distribution,

because the estimator is not a well known Bayesian estiomator. However, almost all the score metrics of Bayesian networks assume this estimator, this paper also employs this estimator.

Furthermore, Cooper and Herskovits (1991), Cooper and Herskovits (1992) assumed that the prior distribution has a uniform $n'_{ijk} = 1, (i = 1, \cdots, N, j = 1, \cdots, q_i, k = 0, \cdots, r_i - 1)$ and then derived the following criterion,

$$p(\mathbf{X} \mid B_S) \propto \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(n_{ij} + r_i - 1)!} \prod_{k=0}^{r_i - 1} n_{ijk}!. \tag{9}$$

This criterion led to their famous causal discovery program "K2" and is called the uniform prior score metric (UPSM).

4. Minimum Description Length (MDL)

The minimum description length (MDL) inference was invented by Rissanen (1978), and its basic idea is to make a tradeoff between model simplicity and fit to the data by minimizing the length of a joint description of the model and the data, assuming the model is correct. The first MDL encoding of Bayesian networks was put forward by Lam and Bacchus (1994). However, their encoding is not a function of the size of the data set. This implies that the code length cannot be efficient. Suzuki (1993) proposes an alternative MDL code for Bayesian networks when the hyper parameter $n'_{ijk} = \frac{1}{2}, (i = 1, \cdots, N, j = 1, \cdots, q_i, k = 0, \cdots, r_i - 1)$:

$$I(B_S, \mathbf{X}) = \ln p(B_S) + \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=0}^{r_i - 1} \left[n_{ijk} \ln \frac{n_{ijk}}{n_{ij}} \right] - \frac{\sum_{i=1}^{N} q_i (r_i - 1)}{2} \ln n, \tag{10}$$

where $n = \sum_{j=1}^{q_i} n_{ij}$. This criterion has a strong consistency that guarantees that the estimated model converges to the true model as $n \to \infty$.

Suzuki (1993) and Suzuki (1998) proved that when the hyper parameter $n'_{ijk} = \frac{1}{2}(i = 1, \cdots, N, j = 1, \cdots, q_i, k = 0, \cdots, r_i - 1)$, the DPSM (8) converges to the MDL in (10).

However, Bouckaert (1994a) and Bouckaert (1994b) also proved that the UPSM (the DPSM when $n'_{ijk} = 1, (i = 1, \cdots, N, j = 1, \cdots, q_i, k = 0, \cdots, r_i - 1)$) converges to the MDL in (10). Suzuki (1998) claimed that Bouckaert (1994a)'s derivation was wrong and the DPSM converges to the MDL in (10) only when $n'_{ijk} = \frac{1}{2}, (i = 1, \cdots, N, j = 1, \cdots, q_i, k = 0, \cdots, r_i - 1)$.

This argument is very interesting, but the problem is that Suzuki (1993), Suzuki (1998), Bouckaert (1994a), and Bouckaert (1994b) assumed that the DPSM with only one prior condition produces the MDL. This paper shows below that the DPSMs with any hyper-parameter conditions derive a more general score metric. This means that both Suzuki (1993) and Bouckaert (1994a) showed the right results.

5. Likelihood equivalence assumption and BDe metric

Heckerman, Geiger, and Chickering (1995) introduced a likelihood equivalence assumption that states that if two structures are equivalent, their parameter joint probability density functions are identical. Theoretically, the likelihood equivalence assumption is written as follows:

Given network structures B_{S1} and B_{S2} such that $p(B_{S1} \mid \xi) > 0$ and $p(B_{S2} \mid \xi) > 0$, if B_{S1} and B_{S2} are equivalent, then $p(\Theta_{B_{S1}} \mid B_{S1}, \xi) = p(\Theta_{B_{S2}} \mid B_{S2}, \xi)$.

Heckerman, Geiger, and Chickering (1995) pointed out that formula (9) does not satisfy the likelihood equivalence assumption and used the likelihood equivalence assumption instead of the Dirichlet distribution assumption to derive the same formula as (8). They also presented a sufficient condition for satisfying the likelihood equivalence assumption as the following constraint about the hyper-parameters:

$$n'_{ijk} = n' p(x_i = k, \Pi_i = j \mid B_S^h, \xi), \tag{11}$$

where n' is the equivalent sample size determined by users and B_S^h is the hypothetical Bayesian network structure that reflects a user's prior knowledge. They called this metric the likelihood-equivalence Bayesian Dirichlet score metric (BDe).

Buntine (1991)'s uniform prior constraint $n'_{ijk} = n'/(r_i q_i)$ is considered a special case of the BDe metric, and Heckerman, Geiger, and Chickering (1995) called this special case BDeu ("u" stands for uniform joint distribution). Buntine noted that this metric satisfies the property of likelihood equivalence.

Kayaalp and Cooper (2002) also proposed a new score metric called the global uniform (GU) metric, which assumes the uniform prior and holds the likelihood-equivalence without any constant n' values.

This paper proves for the first time that the DPSM with any hyper-parameter values has strong consistency and converges to the MDL in any prior knowledge conditions.

6. Relationships among various score metrics

As mentioned above, Suzuki (1993), Suzuki (1998), Bouckaert (1994a), and Bouckaert (1994b) assumed that the DPSM with only one hyper-parameter value ($n'_{ijk} = 1$ or $n'_{ijk} = 1/2$) derives the MDL. However, they have not proved that the MDL cannot be derived from the DPSM, whose hyper-parameter values do not satisfy their setting conditions.

This section shows that DPSMs with any hyper-parameter values, including the BDe, the BDeu, and the GU, converge to the more general score metric, which has a strong consistency and is asymptotically equivalent to the MDL.

That is, the following theorem holds.

Theorem 1 *For* n'_{ijk}, $(i = 1, \cdots, N, j = 1, \cdots, q_i, k = 0, \cdots, r_i - 1)$,

$$\ln p(\boldsymbol{X} \mid B_S) \geq \ln p(\boldsymbol{X}, \hat{\Theta}_{B_S} \mid B_S) - \left(\frac{K}{2}\right) \ln \frac{n' + n}{2\pi} + const, (n \to \infty), \tag{12}$$

where K is the number of parameters in the Bayesian network model, n' is the equivalent sample size detemined by users, and n is the sample size.

[Proof]
From (8), the log-predictive distribution can be written by

$$\begin{aligned}\ln p(\mathbf{X} \mid B_S) &= \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\sum_{k=0}^{r_i-1}\ln\Gamma(n'_{ijk}+n_{ijk}) - \ln\Gamma\left[\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\right]\right)\\ &\quad +const\\ &= \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\sum_{k=0}^{r_i-1}\ln\Gamma(n'_{ijk}+n_{ijk}) - \ln\Gamma(n'_{ij}+n_{ij})\right)+const.\end{aligned}$$

Using the Stirling's series (Box and Tiao (1973), p.147. A.2.2.8)

$$\ln\Gamma(n) = \frac{1}{2}\ln(2\pi) + \left(n-\frac{1}{2}\right)\ln n - n + O\left(\frac{1}{n}\right),$$

we can expand $\ln p(\mathbf{X} \mid B_S)$ as follows;

$$\begin{aligned}\ln p(\mathbf{X} \mid B_S) &= \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\frac{r_i-1}{2}\ln(2\pi) + \sum_{k=0}^{r_i-1}\left(n'_{ijk}+n_{ijk}-\frac{1}{2}\right)\\ \ln(n'_{ijk}+n_{ijk})\\ -\left(n'_{ij}+n_{ij}-\frac{1}{2}\right)\ln(n'_{ij}+n_{ij})\end{array}\right)\\ &\quad +const, (n\to\infty)\\ &= \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\ln(n'_{ijk}+n_{ijk})\\ -(n'_{ij}+n_{ij})\ln(n'_{ij}+n_{ij}) + \frac{r_i-1}{2}\ln(2\pi)\\ -\frac{1}{2}\sum_{k=0}^{r_i-1}\ln(n'_{ijk}+n_{ijk}) + \frac{1}{2}\ln(n'_{ij}+n_{ij})\end{array}\right)\\ &\quad +const, (n\to\infty).\end{aligned}$$

Since $\ln(n'_{ij}+n_{ij}) \geq \ln(n'_{ijk}+n_{ijk})$, we get

$$\begin{aligned}\ln p(\mathbf{X} \mid B_S) &\geq \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\ln\frac{(n'_{ijk}+n_{ijk})}{(n'_{ij}+n_{ij})} + \frac{r_i-1}{2}\ln(2\pi)\\ -\frac{1}{2}\sum_{k=0}^{r_i-1}\ln(n'_{ij}+n_{ij}) + \frac{1}{2}\ln(n'_{ij}+n_{ij})\end{array}\right)\\ &\quad +const, (n\to\infty)\\ &= \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\ln\frac{(n'_{ijk}+n_{ijk})}{(n'_{ij}+n_{ij})} + \frac{r_i-1}{2}\ln(2\pi)\\ -\frac{r_i-1}{2}\ln(n'_{ij}+n_{ij})\end{array}\right)\\ &\quad +const, (n\to\infty)\end{aligned}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\begin{array}{l}\sum_{k=0}^{r_i-1}(n'_{ijk}+n_{ijk})\ln\frac{(n'_{ijk}+n_{ijk})}{(n'_{ij}+n_{ij})}\\ -\frac{r_i-1}{2}\ln\frac{(n'_{ij}+n_{ij})}{2\pi}\end{array}\right)$$
$$+const, (n\to\infty).$$

From $\ln(n'+n) \geq \ln(n'+n_{ij})$, we obtain

$$\ln p(\mathbf{X} \mid B_S) \geq \ln p(\mathbf{X}, \hat{\Theta}_{B_S} \mid B_S) - \left(\frac{K}{2}\right)\ln\frac{n'+n}{2\pi} + const, (n\to\infty).$$

This theorem shows that the various DPSMs, including BDe and BDeu with likelihood equivalence, converge to a strongly consistent score metric because the sufficient condition of strong consistency is that $\ln\ln n < c < n$ when the metric is described by $\ln p(\mathbf{X}, \hat{\Theta}_{B_S} \mid B_S) - c \times K$ (Nishi, 1988). That is, the summation of hyper-parameter n' must be satisfied for the score metric to have a strong consistency

$$\ln p(\mathbf{X}, \hat{\Theta}_{B_S} \mid B_S) \to \sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\left[n_{ijk}\ln\frac{n_{ijk}}{n_{ij}}\right], \text{as } n\to\infty$$
$$2\pi(\ln n)^2 - n < n' < 2\pi\exp(2n) - n.$$

In addition, we note that, as a score metric, formula (12) represents various score metrics by changing the values of the hyper-parameters. For example, when $n' = (2\pi - 1)n$, then this score metric converges to (10). The differences between formula (12) and the MDL (Bouckaert, 1994a; Suzuki, 1993) are that hyper-parameters remain in the log-posterior term and the penalty term of the metric and π remains in the penalty term of the formula (12). Therefore, asymptotically, these score metrics behave very similarly because they all have a strong consistency. However, term $\ln p(B_S) + \sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\left[n_{ijk}\ln\frac{n_{ijk}}{n_{ij}}\right]$ in MDL (10) is not a Bayesian estimator but a maximum likelihood estimator, so score metric (12) will be better when $n_{ij} = 0$ because of the small amount of data for the number of variables.

Moreover, score metric (12) converges to the following information criteria by changing its hyper-parameter values. When $n' = 200 * \pi - n$, the score metric (12) is equivalent to Akaike's information criterion (AIC) (Akaike, 1974). When $n' = 1$, score metric (12) is equivalent to ICOPMP (Bozdogan, 1990).

Thus, score metric (12) can unify the various score metrics for learning Bayesian networks. This means that the DPSM also provides the various score metrics by changing the hyper-parameter values.

7. Learning Bayesian networks from an empirical Bayesian approach

7.1 Previous work and problems

As mentioned first, Silander, Kontakanen, and Myllymaki (2007) provided a series of concrete experiments to find the optimum parameter values of BDeu and showed that the

solution of the network structure is highly sensitive to the chosen hyper-parameter values. However, their study forcused on BDeu score and did not compare the performances of the BDeu chainging the hyper-parameter value with the other score metrics.

Yang and Chang (2002) compared the performance of score metrics UPSM, DPSM, BDeu, and MDL using networks with three nodes and five nodes and the ALARM network structure. The experimental results show that when $n'_{ijk} = 10$, the DPSM is best at identifying the true network structure. They also reported that the BDeu performed worst in their experiments and that the likelihood equivalence assumption was unreasonable for actual learning Bayesian networks.

However, they investigated the performances of the DPSM only when $n'_{ijk} = 2$ and $n'_{ijk} = 10$, and the BDeu only when $n' = 4$, $n' = 16$, and $n' = 96$. There were not a sufficient amount of conditions in the experiments to find that $n'_{ijk} = 10$ is the optimum hyper-parameter and the BDe is not a good criterion.

It should be also noted that they implicitly assumed that there is a certain optimum hyper-parameter value for the learning Bayesian networks in the same way as Suzuki (1993), Suzuki (1998), Bouckaert (1994a), and Bouckaert (1994b). As theorem 1 showed, the score metrics UPSM, DPSM, BDeu, and MDL have the strong consistency, so they all asymptotically provide the same results. The selection of hyper-parameter values affects the learning Bayesian network results especially when there is only a small amount of data. If the prior distribution is close to the true distribution, then the probability of constructing the true network structure (or learning efficiency) increases even when there is only a small amount of data. Thus, the optimal hyper-parameter values are thought to depend on the true network structure and the amount of data. The next section presents some simulations that were used to analyze the hyper-parameters and proposes a learning Bayesian networks method from an empirical Bayesian approach.

7.2 Simulations

This section provides some simulations to test whether the optimal hyper-parameter values depend on the true network structures and the amount of data. Three Bayesian network structures are constructed. These experiments investigate the effects of different settings of the parameters on the prediction results by making the arcs "strong cousality" and "weak causality". Here, we define the causality strongness of $A \rightarrow B$ by the amount of mutual information $I(A, B)$ in accordance with Chow and Liu (1968). The mutual information between arcs in Figure 3(1) are $I(0, 1) = 0.037$, $I(0, 2) = 0.022$, $I(1\&2, 3) = 0.014$, and $I(2, 4) = 0.018$. The mutual information between arcs in Figure 3(2) are $I(0, 1) = 0.061$, $I(0, 2) = 0.067$, $I(1\&2, 3) = 0.063$, and $I(2, 4) = 0.057$. The mutual information between arcs in Figure 3(1) are $I(0, 1) = 0.0005$, $I(0, 2) = 0.004$, $I(1\&2, 3) = 0.018$, and $I(2, 4) = 0.0005$.

Bayesian network (1) in Figure 3 is a combination of strong and weak causality arcs. Here, a strong causality arc is one in which the conditional probabilities for a variable consist of very different values, such as the values around 1.0 and around 0.0. A weak causality arc is one in which the conditional probabilities of a variable take the almost

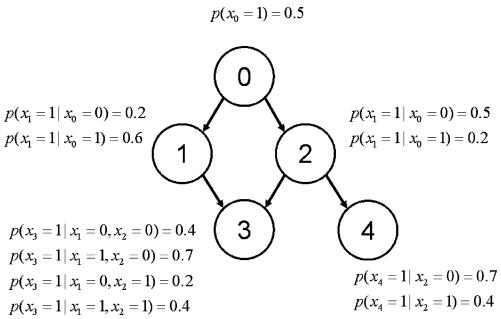

Figure 3: Bayesian network (1).

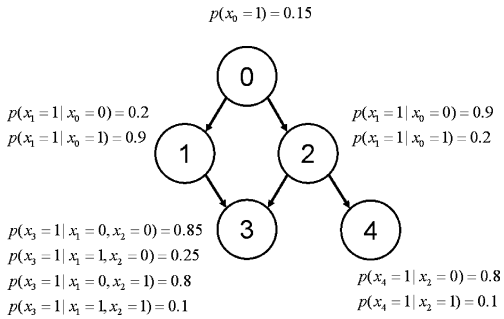

Figure 4: Bayesian network (2).

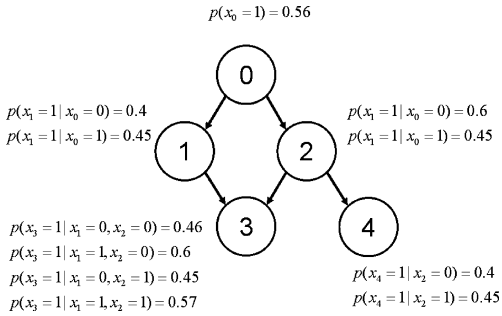

Figure 5: Bayesian network (3).

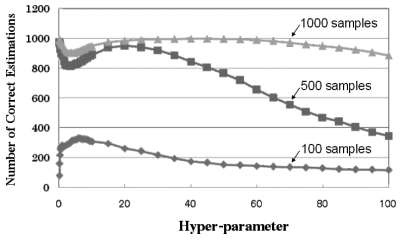

Figure 6: DPSM performance for Bayesian network (1).

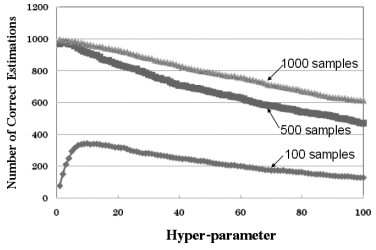

Figure 7: BDeu performance for Bayesian network (1).

same values. Bayesian network (2) in Figure 4 consists only of strong causality arcs. Bayesian network (3) in Figure 5 consists only of weak causality arcs. The procedures in the simulations using the three network structures are as follows:

1. There are three sets of 1000 samples generated from the three Bayesian network structures with the different conditional probabilities shown in Figures 3, 4, and 5.
2. Using the DPSM and the BDeu, Bayesian network structures are estimated based on 100, 500, and 1000 samples, respectively, from the datasets for the structures shown in Figures 3, 4, and 5, by changing the values of hyper-parameters from 1 to 100.
3. The number of times the estimated structure is the true structure is counted by repeating Procedure 2 1000 times.

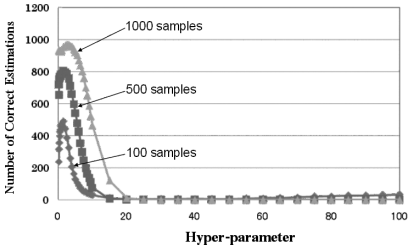

Figure 8: DPSM performance for Bayesian network (2).

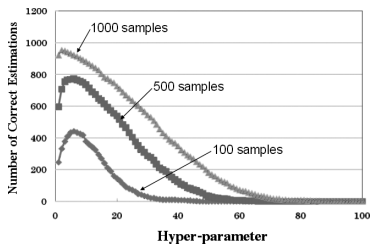

Figure 9: BDeu performance for Bayesian network (2).

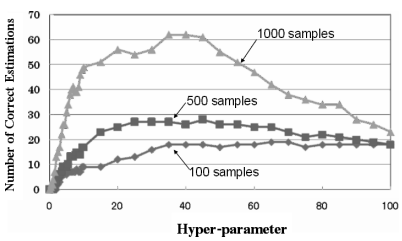

Figure 10: DPSM performance for Bayesian network (3).

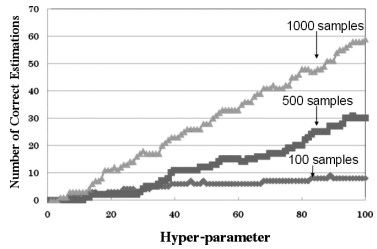

Figure 11: BDeu performance for Bayesian network (3).

Table 1: Comparisons of estimation performances among various score metrics

The structure	Sample sizes for	BDeu	DPSM	UPSM	AIC	MDL
Bayesian network 1	100	348 (n'=10)	330 (n'_{ijk}=6)	222	205	341
	500	979 (n'=2)	973 (n'_{ijk}=1)	870	480	972
	1000	999 (n'=1)	998 (n'_{ijk}=40)	897	424	989
Bayesian network 2	100	446 (n'=9)	486 (n'_{ijk}=2)	357	238	428
	500	773 (n'=9)	804 (n'_{ijk}=2)	676	561	726
	1000	956 (n'=2)	970 (n'_{ijk}=12)	896	496	928
Bayesian network 3	100	9 (n'=89)	19 (n'_{ijk}=65)	4	2	6
	500	31 (n'=93)	28 (n'_{ijk}=45)	26	7	26
	1000	59 (n'=100)	62 (n'_{ijk}=30)	47	6	51

Here, the greedy algorithm with no restriction on the number of parents is used as a search algorithm. The DPSM and the BDeu results obtained by changing the respective hyper-parameter values, n'_{ijk} and n', from 1 to 100 are shown in Figures 6, 7, 8, 9, 10, and 11. In each Figure, the horizontal axis indicates the value of the hyper-parameter, and the vertical axis indicates the number of correct estimations per 1000 estimations experiments using the learning method. From these figures, the optimum hyper-parameter values can be determined as one which shows the highest prediction performance (The highest number of correct estimation as the vertical axis).

It can be seen from these results that the optimum hyper-parameter values of the DPSM and the BDeu depend on the true structure and the number of samples. However, the traditional concerning researches have implicitly assumed that there is a certain optimum hyper-parameter value for learning Bayesian networks. The performances of the BDeu for

networks (1) and (2) show the peak when the hyper-parameter is around 0-10. When the true conditional probabilities of the network have values around 1.0 or 0.0, BDeu's performances show the peak when the hyper-parameter is small values. In contrast, when the true conditional probabilities of the network tend to have values around 0.5 like network (3), BDeu's performance increases in a monotonic curve because of the shrinkage effects (see, for example, Varian (1975)) of the hyper-parameter values.

The same simulations for the score metric with fixed hyper-parameter values, UPSM, and the score metrics with no hyper-parameters, AIC and MDL, are provided. (Note that the results of DPSM and BDeu show the results of the highest number of correct estimation by changing the hyper-parameter values. Table 1 shows the number of correct estimations for the DPSM, the BDeu, the UPSM, the AIC, and the MDL. As can be seen in the table, the BDeu performs best for network structure (1), which has both strong causality arcs and weak causality arcs, and the DPSM performs best for network (2) with strong causality arcs and (3) with weak causality arcs. It can be seen that BDeu and DPSM return similar results. The optimum hyper-parameters of DPSM and BDeu in Table 1 are different because DPSM's optimum hyper-parameter means $n'_{ijk} = const$, but BDeu's optimum hyper-parameter means the likelihood equivalent sample size $n' = \sum_{j=1}^{q_i} \sum_{k=0}^{r_i} n_{ijk}$. If the numbers of all variables' parameters $(r_i q_i)$ is constant at K, then DPSM's performance with n' are equivalent to BDeu's performance with $n' \times K$. This means that the best hyper-parameter for DPSM is smaller than the best hyper-parameter for BDeu when the numbers of all variables' parameters $(r_i q_i)$ is constant at K. The reasons why Table 1 shows some contrary results are that this experiment does not satisfy the constraint that the numbers of all variables' parameters $(r_i q_i)$ is constant at K and the search areas of the best hyper-parameter for the DPSM and the BDeu are different because the scales of the hyper-parameter are different between the DPSM and the BDeu.

In addition, it can be said that the prediction efficiency of BDeu and DPSM are almost the same from Table 1. However, BDeu, which satisfies the likelihood equivalence, is theoretically more sound than DPSM.

These results are quite different from Yang and Chang (2002)'s results because BDeu performed at its worst in their experiments. This was because their experiments used only $n' = 4$, $n' = 16$, and $n' = 96$ as hyper-parameters values. BDeu's performance changes a great deal when the value of hyper-parameter from Figures 7, 9, and 11 is changed, but Yang et al. did not use enough experimental conditions to compare the performances of the various score metrics.

In addition, although the results are not shown in Figures 6–11 and Table 1, the same simulations using 10,000 samples are performed. In those simulations, BDeu and DPMS both estimated correctly 100% of the time with any hyper-parameter value because they have the strong consistency for any hyper-parameters, as shown in Theorem 1.

This section showed that BDe is theoretically and practically the best score metric. The remaining problem is how to estimate the optimum hyper-parameter values of BDeu from data.

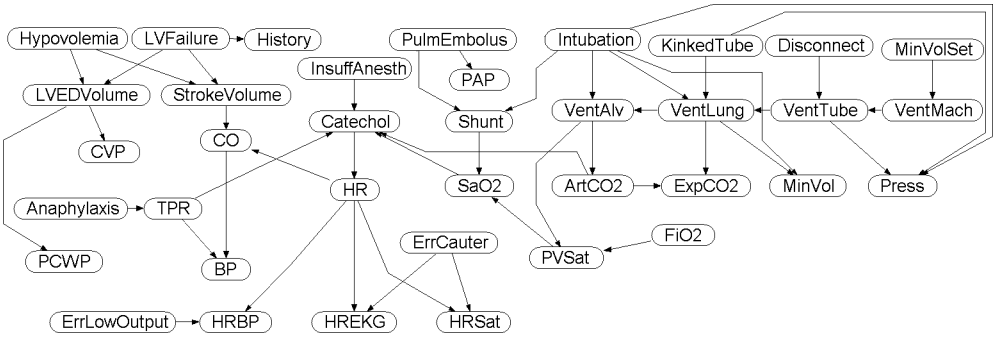

Figure 12: Alarm network structure

7.3 BDeu from an empirical Bayesian approach

The previous section showed that BDeu is a theorically and actually reasonable score metric for the various learning Bayesian networks methods. However, in the real world, it is difficult to evaluate the values of hyper-parameters because the true network structure is not known. This section proposes a hyper-parameter estimation method for BDeu based on an empirical Bayesian approach.

Given n data $\mathbf{X}$, we propose the following hyper-parameter values selection procedure using the cross-validation method:

1. Let the range of the hyper-parameter n' be $a \leq n' \leq b$.
2. Let the initial value of the hyper-parameter be $n' = a$.
3. Randomly sample as training data 50% of n data $\mathbf{X}$. The remaining data is called validation data.
4. Learn network structure using BDeu with n' from the training data.
5. Learn network structure using BDeu with n' from the validation data.
6. Count the errors of the estimated structures learned from the training data and the validation data. One error is counted if one parent node is missing or if there is one more parent node in the learned structure from the validation data than in the learned structure from the training data.
7. Propagate the value of the hyper-parameter $n' = n' + 1$ while $n' \leq b$.
8. Repeat procedures 2–7 100 times.
9. Calculate the average of the errors over 100 trials.
10. Select the value of the hyper-parameter to minimize the average of the errors.

To test the performance of the proposed method, this study generated random samples from the Alarm network in Figure 12, specifically three kinds of 2000-sample data, Bayesian networks A, B, and C, by changing the conditional probabilities of the Alarm network. The setting method of the conditional probabilities follows **7.2** experiment's method. Bayesian network A consists of a combnation of strong and weak causality arcs. Bayesian network B consists only of strong causality arcs. Bayesian network C consists only of weak causality arcs. As same as **7.2**, a strong causality arc is one in which the conditional probabilities for a variable consist of very different values and a weak causality

Table 2: Estimation performance of proposed method and other score metrics

The structure	Sample sizes	BDeu	UPSM	AIC	MDL
Bayesian network A	1000	7.86 (n'=19)	8.26	14.72	10.26
	1500	5.42 (n'=21)	7.58	13.61	8.94
	2000	3.71 (n'=18)	6.94	14.87	7.86
Bayesian network B	1000	6.94 (n'=5)	8.12	27.87	11.23
	1500	5.63 (n'=8)	7.23	28.65	8.27
	2000	3.12 (n'=18)	6.76	27.76	6.53
Bayesian network C	1000	7.27 (n'=11)	8.11	31.72	10.27
	1500	6.82 (n'=18)	7.38	31.87	8.96
	2000	3.93 (n'=23)	6.93	32.79	8.21

arc is one in which the conditional probabilities of a variable take the almost same value. However, the detail list of the conditional probabilities are omitted due to limitations of space.

Using these data sets, the proposed method with UPSM, MDL, and AIC using 1000, 1500, and 2000 samples were compared.

It is impossible to estimate the complete true structure in the case of the Alarm network because the network is large, therefore this experiment evaluates the average of the errors (the average number of arcs which the estimated structure missed) by each learning method. The results are shown in Table 2. It should be noted that Table 1 evaluated the number of correct estimations of the true structure but Table 2 evaluated the average of the errors (the average number of arcs which the estimated structure missed or wrongly added).

As can be seen in Table 2, the proposed method using BDeu performed best for various networks structures. BDeu's results were almost always better than UPSM's. In addition, MDL performance was somewhat unstable when there was not enough data, which could be anticipated based on its score metric expression in (10). Because $n_{ijk} \ln(n_{ijk}/n_{ij})$ is very sensitive to variations in n_{ijk}, and n_{ijk} may vary with database size, increasing the amount of data may not necessarily reduce the chance of making errors during structure induction.

The AIC performed worst because it does not have the strong consistency.

Thus, the proposed method performs better than any of the other learning Bayesian networks in the simulations. Speeding up the hyper-parameter values selection procedure is an urgent problem. The average number of errors for the structures learned from training data and validation data describe a concave curve for the hyper-parameter values. Using the concave curve property for the hyper-parameter values, we can use the high-speed searching algorithm of the hyper-parameter values selection.

8. Why is a cross varidation based empirical Bayesian approach employed?

In this research, we employ a cross varidation method instead of a well known typical empirical Bayesian method, analytically estimating the hyper-parameter values which

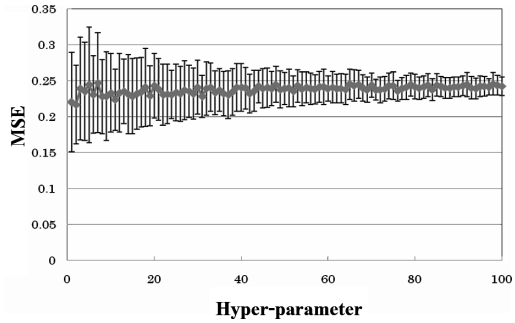

Figure 13: Evidence-based inference performance of network using 100 samples

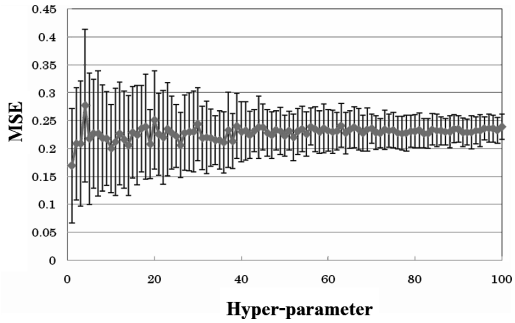

Figure 14: Evidence-based inference performance of network using 1000 samples

maximizes the predictive distribution of network structures. The main reason is that cross varidation based empirical Bayesian approach can also obtain the optimum hyper-parameter values for the probabilistic inferences.

There is a learning Bayesian network approach that is not based on scoring metrics that was not thoroughly discussed here. In this approach, Bayesian network structure encodes a group of conditional independence relationships among the nodes using the concept of *d-separation* (Pearl, 1988). This suggests that learning Bayesian networks are built by identifying the conditional independence relationships among the nodes. Using statistical tests such as Chi-squared test and mutual information test makes it possible to find the conditional independence relationships among the attributes and use these relationships as constraints to construct a Bayesian network. These algorithms are referred to as *CI*-based algorithms or constraint-based algorithms (Cheng, Bell, and Liu, 1997; Sprites Glymour, and Scheines, 1993).

Heckerman, Meek, and Cooper (1997) compared score metric approaches with *CI*-based approaches and showed that the score metric approaches often have certain advantages over the *CI*-based approaches, in terms of approximation of a joint probability distribution of the Bayesian network. However, Friedman, Geiger, and Goldszmidt (1997) showed theoretically that the general score metric approaches may produce poor classifiers because a good classifier maximizes a different function. Greiner, Grove, and Schuurmans (1997) reached the same conclusion through a different type of analysis. They also reported that score metric methods are often less efficient.

Cheng and Greiner (1999) referred to these results, proposing a modified *CI*-based method (Chow and Liu, 1968) and demonstrated that it performed better than the other learning Bayesian networks methods. As a Bayesian classifier, they used a sub-model of a Bayesian network, but their results do not mean that the score metric methods have no advantage over the *CI*-based methods. Rather, they mean that probabilistic inference, including classification by Bayesian network and approximation of joint probability distribution by a Bayesian network, are different problems.

For example, Dash and Cooper (2004) proposed a very interesting method that assumes a model that mixes several Bayesian networks. They proposed a score metric for learning

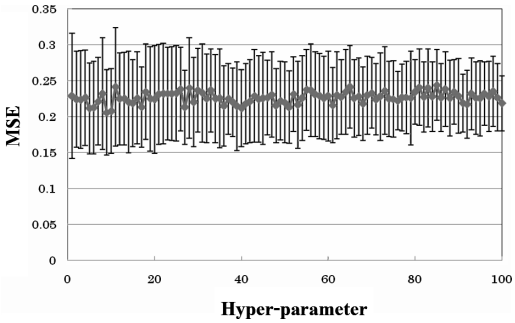

Figure 15: Evidence-based inference performance of network based on 100 samples

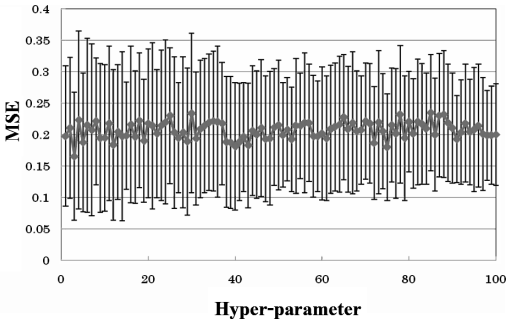

Figure 16: Evidence-based inference performance of the network based on 1000 samples

a mixture of Bayesian networks and showed that the constructed Bayesian network models provide accurate inferences and classifications. However, this method suffers from the serious problem that the search space of the network structure expands vastly, making it impossible to use for large variable data.

We argue here that a unified score metric for learning Bayesian networks can be represented by changing the hyper-parameter values for both approximation of the joint probability distribution and inference of a Bayesian network. To put it another way, we assume that there are optimal hyper-parameter values for approximating the joint probability distribution and the different optimal hyper-parameter values for inferring a Bayesian network. This means that we have to discriminate among the problems of approximating the joint probability distribution of Bayesian networks and inferring each variable value in a Bayesian network from the observed evidence.

For example, Figure 13 shows the evidence-based inferences performances of the Bayesian network learned based on 100 random samples from the true structure (1) in Figure 3. The horizontal axis indicates the value of hyper-parameter n' when the structure is learned using BDeu, and the vertical axis indicates the mean squared errors (MSE: drawn by heavy line) of the Bayesian network inferences for all variables over 100 simulation trials given one bit of evidence. The thin line in the figure means the 95% confidence interval of the MSE. Here, the evidence variable in each trial is randomly selected. Figure 14 shows the one evidence-based inference performance of the network learned based on 1000 random samples. Figure 15 shows the evidence-based inference performance of the network based on three bits of evidence learned based on 100 random samples. Figure 16 shows the three evidences based inferences performances of the network learned from 1000 random samples. The figures show that the optimum hyper-parameter values for the inferences given that some evidence is different from the optimum values for estimating the true network structure, or approximation of the joint probability distribution.

However, the cross varidation based empirical Bayesian approach can unify the two problems, maximizing the probabilities inferences efiiciency given certain evidence and approximating the true joint probabuility distribution, or finding the true network structure.

This is the main reason that we employ the cross varidation based empirical Bayesian approach.

9. Conclusion and discussions

This paper proposed a likelihood-equivalence Bayesian network learning method based on an empirical Bayesian approach. There have been several different apparently contradictory theories in the learning Bayesian network area that derive different optimum hyper-parameter values of DPSM. However, this paper has presented different results:

1. The DPSM has a strong consistency for any hyper-parameter set. That is, the score metrics UPSM, DPSM, BDe, and MDL asymptotically converge to the same results.
2. The problem of setting optimal hyper-parameters is affected by the true network structure and the amount of data.
3. BDe based on the likelihood equivalence is a theoretically and actually reasonable score metric.

To find optimum hyper-parameters, this paper also proposed a new method of learning Bayesian networks based on an empirical Bayesian approach. The unique features of this method are:

1. It is possible to reflect a user's prior knowledge.
2. It has both strong consistency and likelihood equivalence properties.
3. To maximize the predictive efficiency, adapting for the domain and data size,
4. The cross varidation based empirical Bayesian approach can unify the different two problems of Bayesian networks, maximizing the probabilities inferences efficiency given certain evidence and approximating the true joint probabuility distribution, or finding the true network structure.

Finally, using the proposed method, this paper presented some numerical examples that demonstrated the method's effectiveness.

Acknowledgements

The author would like to thank Dr. Chao-Lin Liu of National Chengchi University for his helpful comments to improve the presentation of this paper.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automatic Control*, **19**, 716–723.

Almond, R.G. (1995). *Graphical belief modeling*, Florida: CRC Press.

Bouckaert, R. (1994a). Probablistic network construction using the minimum description length principle, Technical Report RUU-CS-94-27, Utrecht University.

Bouckaert, R. (1994b). Properties of Bayesian network learning algorithm, *Proc. Uncertainty in Artificial Intelligence, California*, 102–109.

Box, G.E.P. & Tiao, G.C. (1973). *Bayesian inference in statistical analysis*, Addison-Wiely, Publisher.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models, *Communications in Statistics*, **A19**, No.1, 221–278.

Buntine, W. (1991). Theory refinement on Bayesian networks, *Proc. Uncertainty in Artificial Intelligence*, 52–60.

Buntine, W. (1996). A Guide to the Literature on Learning Probabilistic Networks from Data, *IEEE Transactions on Knowledge and Data Engineering*, **8**(2), 195–210.

Cheng, J., Bell, D.A., & Liu, W. (1997). An algorithm for Bayesian belief network construction from data. *Proc. AI & STAT'97*, 83–90.

Cheng, J. & Greiner, R. (1999). Comparing Bayesian Network Classifers, *Proc. Uncertainty in Artificial Intelligence*, 101–108.

Chow, C.K. & Liu, C.N. (1968). Approximating discrete probability distributions with dependence trees, *IEEE transactions on information theory*, **14**, 462–467.

Cooper, G.F. & Herskovits, E. (1991). A Bayesian Methods for the induction of probabilistic networks from data, Technical Report SMI-91-1, Section on Medical Informatics, Stanford University.

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

Dash, D. & Cooper, G.F. (2004). Model averaging for prediction with discrete bayesian networks, *Journal of Machine Learning Research*, **5**, 1177–1203.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifer, *Machine Learning*, **29**, 131–161.

Greiner, R., Grove, A., & Schuurmans, D. (1997). Learning Bayesian nets that perform well, *Proc. Uncertainty in Artificial Intelligence*, 198–206.

Heckerman, D. (1995). A Tutorial on Learning With Bayesian Networks, Technical Report MSR-TR-95-06, Microsoft Research.

Heckerman, D., Geiger, D., & Chickering, D.M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, **20**(3), 197–243.

Heckerman, D., Meek, C., & Cooper, G. (1997). A Bayesian approach to causal discovery, Technical Report MSR-TR-97-05. Microsoft Research.

Jensen, F. (2001). *Bayesian networks and decision graphs*, NY: Springer-Verlag Publisher.

Kayaalp, M. & Cooper, G.F. (2002). A Bayesian network scoring metric: That is based on Globally uniform parameter priors, *Proc. Uncertainty in Artificial Intelligence*, 251–258.

Korb, K.B. & Nicholeson, E.A. (2004). *Bayesian artificial intelligence*, Florida: Chapman & Hall/CRC.

Lam, W. & Bacchus, F. (1994). Learning Bayesian Belief Networks: An Approach Based on the MDL Principle, *Computational Intelligence*, **10**, 4, 269–293.

Neapolitan, R.E. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms*, NY: John Wiley & Sons.

Neapolitan, R.E. (2004). *Learning Bayesian networks*, NJ: Prentice Hall Pub.

Nishi, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified, *Journal of Multivariate Analysis*, **27**, 392–403.

Pearl, J. (1998). *Probabilistic reaoning in intelligent systems*, CA: Morgan Kaufman Publisher.

Rissanen, J. (1978). Modeling by shortest data description, *Autometrika*, **14**, 465–471.

Spiegelhalter, D., Dawid, A., Lauritzen, S., & Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, **8**, 219–282.

Sprites, C., Glymour, C., & Scheines, R. (1993). *Causation, Prediction and Search*, Springer-Verlag, New York.

Silander, T., Kontakanen, P., & Myllymaki, P. (2007). On sensitivity of the MAP Bayesian network structure to the equipment sample size parameter, Proc. the twenty-third conference of Uncertainty in Artificial Intteligence: 360–367.

Steck, H., & Jaakkola, T. (2002). On the dirichlet prior and Bayesian regularization. In Becker, S., Thrun, S., & Obermayer, K., editors, NIPS, pp.697–704, MIT Press.

Suzuki, J. (1993). A Construction of Bayesian networks from Databases on an MDL Principle, *Proc. Uncertainty in Artificial Intelligence*, 266–273.

Suzuki, J. (1998). Learning Bayesian belief networks based on the MDL principle:An efficient algorithm using the branch and bound technique, *IEICE Transction, Information and Systems*, Vol. **E81-D**, No.12.

Varian, H.R. (1975). A Bayesian Approach to Real Estate Assessment, in Fienberg, S.E. and Zellner, A., (eds.), Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage. Amsterdam: North-Holland, 195–208.

Yang, S. & Chang, K-C. (2002). Comparison of score metrics for Bayesian network learning, *IEEE Transaction on systems, Man and Cybernetics–PART A: Systems and Humans*, **32**(3), 419–428.

(Received November 20 2007, Revised April 16 2008)