Minimum Free Energy Principle for Constraint-Based Learning Bayesian Networks

Takashi Isozaki[1,2] and Maomi Ueno[1]

[1] Graduate School of Information Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
[2] Research and Technology Group, Fuji Xerox Co., Ltd. 3-1-1 Roppongi, Minato-ku, Tokyo 106-0032, Japan
{t-isozaki,ueno}@ai.is.uec.ac.jp

Abstract. Constraint-based search methods, which are a major approach to learning Bayesian networks, are expected to be effective in causal discovery tasks. However, such methods often suffer from impracticality of classical hypothesis testing for conditional independence when the sample size is insufficiently large. We propose a new conditional independence (CI) testing method that is effective for small samples. Our method uses the minimum free energy principle, which originates from thermodynamics, with the "Data Temperature" assumption recently proposed for relating probabilistic fluctuation to virtual thermal fluctuation. We define free energy using Kullback–Leibler divergence in a manner corresponding to an information-geometric perspective. This CI method incorporates the maximum entropy principle and converges to classical hypothesis tests in asymptotic regions. We provide a simulation study, the results of which show that our method improves the learning performance of the well known PC algorithm in some respects.

Keywords: Bayesian networks, Structure learning, Conditional independence test, Minimum free energy principle.

1 Introduction

Bayesian networks (BNs) [1] are graphical models and compact representations of joint probability distributions. This combination is suitable for modeling the uncertainty surrounding random variables. Actually, BNs are widely studied for various applications, such as expert systems, human modeling, autonomous agents, natural language processing, and computational biology. The network structure expresses conditional dependence–independence relations among random variables. A wide range of applications is considered. Therefore, structure discovery of BNs from observational data has become an attractive problem tackled by many researchers over the past decade.

The many proposed methods of structure learning can be categorized into two major approaches: score-and-search based methods (e.g. [2]) and constraint-based methods (e.g. [3]). Score-and-search based methods describe the fitness of

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 612–627, 2009.

each possible structure to the observed data while retaining appropriate complexity of models, for which scores such as AIC, BIC, MDL, and BDeu are typically used (e.g. [2]). Constraint-based methods are designed to estimate properties of conditional independence among the variables in the data. Both methods present distinct advantages and disadvantages. Constraint-based methods are computationally efficient and expected to find causal models and latent common causes under certain conditions [3].

However, in contrast to recent development of score-and-search based methods, disadvantages of constraint-based methods have increasingly achieved prominence [4]. Instances of conditional independence are often decided using classical hypothesis tests with χ^2 or G^2 test [3,4]. One disadvantage of the classical tests is their impracticality for use with small samples because of their use of asymptotic approximation of the statistic to the χ^2 distribution, which is justified for a sufficiently large sample size.

As described in this paper, we propose an alternative conditional independence testing method that presents an advantage of effectiveness for small samples and which has connectivity with the classical hypothesis tests. To realize this, we use the *minimum free energy principle*, which originated from thermodynamics (e.g. [5]). In fact, the minimum free energy (MFE) principle and similar ideas have recently attracted the attention of researchers in some domains of computer science such as clustering [6] and learning [7,8]. In thermal physics, free energy consists of an internal energy, an entropy, and temperature. All of these are expected to play important roles. Nevertheless, many studies that have applied free energy to statistical science have treated temperature as a fixed parameter or a free parameter, apparently because of its lack of clarity of the meaning in data science. Consequently, we consider that the potentials of free energies have not been well extracted.

We remain acutely aware that an advantage of using free energies in statistical science is their capability of expressing a tradeoff between the maximum likelihood (ML) [9] and the maximum entropy (ME) [10] principles, which are best used, respectively, for sufficiently large datasets and small datasets. As described in this paper, for solving this problem, we use a metaphor of the tradeoff between minimizing internal energies, which are dominant for low temperatures, and maximizing entropies, which are dominant for high temperatures in the MFE principle on thermodynamics. We can regard the temperature in free energies as a tradeoff parameter that determines the component fraction of the ML and ME concepts. Therefore, if we pursue the program, it is reasonable to relate temperature with the available data size. Consequently, we recently proposed the "Data Temperature" assumption by which the inverted temperature is a monotonic increasing function of the available data size. We demonstrated its effectiveness in parameter learning of BNs [11].

For this work, we adopt this assumption for learning structure BNs. However, a new manner of definition of free energies must be developed for constraint-based learning. These new definitions are related to a fact described in *information geometry* [12]. As a result, we obtain a new unified manner between

the constraint-based structure and parameter learning of BNs using the MFE principle with the "Data Temperature" assumption, in which we use only one hyperparameter introduced in our previous work.

Advantages of the proposed method are the following.

- Improving accuracy in some respects for constraint-based learning methods on BNs, especially for small samples, and having connectivity with the classical hypothesis tests for asymptotic regions by introducing the minimum free energy principle, which can treat the tradeoff of the maximum likelihood and maximum entropy principles explicitly.
- Developing a novel theoretical framework of unifying structure and parameter learning methods of BNs, which corresponds to an information-geometrical perspective.

Furthermore, we demonstrated the performance of our methods incorporated with the PC algorithm [3], which is a typical benchmark constraint-based algorithm. Our robust independence testing method turns out to be more effective on one level than the standard hypothesis tests for small or medium data sizes.

2 Background

2.1 Bayesian Networks

A Bayesian network (BN) is a set $\mathbb{B} = \langle \mathbb{G}, P \rangle$. Actually, $\mathbb{G} = \langle \mathrm{V}, \mathrm{E} \rangle$ denotes a directed acyclic graph (DAG) with nodes representing random variables V. In addition, P is a joint probability distribution on V. Furthermore, $\mathbb{G}$ and P must satisfy the Markov condition: all variables $X \in \mathrm{V}$ are independent of any subset of its non-descendant variables conditioned on the set of its parents [1]. The set of the parents of a variable X_i is in the graph $\mathbb{G}$ as Π_i. For a distribution P of n variables $\mathrm{V} = \{X_1, \ldots, X_n\}$, a BN $\mathbb{B}$ can be factorized as conditional probability distributions

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \,|\, \Pi_i) \ , \tag{1}$$

as proved easily using the Markov condition.

The graph of a BN presents some instances of entailed independence of the probability distribution. The d-separation [1] is a concept in DAGs, which characterizes entailed conditional independence in the graph. Two nodes X and Y are d-separated by $\boldsymbol{Z}$ in graph $\mathbb{G}$, denoted as $Dsep^{\mathbb{G}}(X; Y|\boldsymbol{Z})$. In addition, $Ind\,(X; Y|\boldsymbol{Z})$ is denoted as the conditional independence of X and Y given $\boldsymbol{Z}$ in P. It is known that $Dsep^{\mathbb{G}}(X; Y|\boldsymbol{Z}) \Rightarrow Ind\,(X; Y|\boldsymbol{Z})$ for a BN $\mathbb{B} = \langle \mathbb{G}, P \rangle$.

A BN $\mathbb{B} = \langle \mathbb{G}, P \rangle$ satisfies the *faithfulness condition* if the Markov condition entails all and only the instances of conditional independence in P [3]. In a faithful BN $\mathbb{B} = \langle \mathbb{G}, P \rangle$, $Dsep^{\mathbb{G}}(X; Y|\boldsymbol{Z}) \Leftrightarrow Ind\,(X; Y|\boldsymbol{Z})$ [1]. This equivalence relation enables us to infer structures of BNs from conditional independence relations. We assume for this discussion that BNs satisfy the faithfulness condition.

We describe some other assumptions used to conduct DAG structure inference in this paper, further to *i.i.d.*, as follows.

- Discrete Variables: each variable in V has a finite, discrete number of possible values.
- No Missing Values: let D be a database set with sample size N such that each sample has no missing values for all variables in set V.
- No Latent Variables: we consider DAGs without latent variables.

2.2 Constraint-Based Learning on BNs

Here we state a basic concept of the constraint-based structure learning methods using the preceding notation. For a BN that satisfies the faithfulness condition, the basic concepts are the following [13].

- Search for a set $\boldsymbol{Z}$ for each pair of variables X and Y in V such that $Ind(X;Y|\boldsymbol{Z})$ holds in P. Therefore, X and Y are conditionally independent given a set $\boldsymbol{Z}$ in P. Construct an undirected graph such that nodes X and Y are connected with an undirected edge if and only if no set $\boldsymbol{Z}$ can be found.
- For each pair of nonadjacent variables X and Y with a common neighbor W, check if $W \in \boldsymbol{Z}$. If not, then add arrowheads pointing at W (i.e., $X \rightarrow W \leftarrow Y$), the type of which is called a v-structure.
- Orient as many of the undirected edges as possible subject to two conditions: (i) the orientation should not create a new v-structure; and (ii) the orientation should not create a directed cycle graph.

For checking conditional independence, we describe classical hypothesis testing, which is frequently used within BN learning algorithm under a faithfulness assumption [3,4]; we also use it afterwards. Statistics such as χ^2 or G^2 are expressed here as S^2. If S^2 can be approximated to a χ^2 distribution with degrees of freedom df: χ^2_{df} and $S^2 < \chi^2_{\alpha,df}$, where $\chi^2_{\alpha,df}$ is a threshold value such that $P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$, in which α is a fixed confidence level, then we do not reject the null hypothesis of (conditional) independence between two selected variables given selected conditional sets; otherwise we reject it. The validity of approximation of statistics such as χ^2 or G^2 is proved in asymptotic regions [14]. However, for a small sample size, it is not justified. Spirtes et al. [3] have used, in their PC algorithm, a criterion for the validity: the algorithm does not perform an independence test if the sample size is less than 10 times the number of different possible joint patterns of the two variables and conditional sets, which means the variables are assumed to be conditionally dependent. This impracticality is a weak point of the constraint-based learning methods of BNs because learning BNs often process insufficient data.

3 Conditional Independence Testing Using the MFE Principle

In this section, we describe a new conditional independence testing method that is designed to be especially effective for small data size, and which is designed

to be connected asymptotically with classical hypothesis testing. We take an approach from the metaphor of thermodynamics, where entropy, energy, and temperature play important roles.

3.1 Why Is the MFE Principle Needed?

Many studies of learning BNs have often used mutual information (e.g., [15]) for measuring dependence, which often means *minimizing entropy*, as described below. For asymptotic regions, for which sufficiently large samples are available, the guiding principle in statistics is the maximum likelihood (ML) principle [9]. Friedman et al. [15] derived that, for a BN $\mathbb{G}$, given a dataset D with N data size for n random variables, maximizing the log likelihood $LL(\mathbb{G}|D)$ is equivalent to maximizing empirical mutual information between a node and its parent nodes (represented as Π_i for a node X_i): $LL(\mathbb{G}|D) = N(\sum_{i=1}^{n} \hat{I}(X_i; \Pi_i) - \sum_{i=1}^{n} \hat{H}(X_i))$, where $\hat{I}$ and $\hat{H}$ denotes empirical mutual information and Shannon entropy [10], and the second term of the right-hand-side of the equation has nothing to do with the learning structure. Therefore, from the definition of mutual information, it is readily derived that

$$LL(\mathbb{G}|D) = -N \sum_{i=1}^{n} \hat{H}(X_i \,|\, \Pi_i) = -N\, \hat{H}(X_1, \ldots, X_n) \ . \tag{2}$$

The last equation is derived from the definition of BNs described in (1). This equation means that maximizing the log likelihood is equivalent to minimizing the entropy of BNs. This equation also implies that maximizing the log likelihood for constructing the DAG structures engenders *complete DAG* because the following inequality is justified: $-N \sum_i \hat{H}(X_i \,|\, \Pi_i) \geq -N \sum_i \hat{H}(X_i)$, because $0 \leq H(X|Y) \leq H(X)$ [10].

In contrast, when we obtain insufficient data, it is reasonable to use the maximum entropy (ME) principle [10], which states that the most preferable probabilistic model should maximize its entropies under some constraint related to available data. Consequently, with no constraint, maximizing entropies of BNs engenders the DAG with *no edges*, which means that a BN is a collection of complete independent distributions: $P(\boldsymbol{X}) = \prod_i P(X_i)$.

A tradeoff exists between maximum likelihood and maximum entropy for obtaining the valid structures. In the asymptotic region, the ML principle is expected to be dominant; in an insufficient sample region, the ME principle is expected to be dominant. Therefore, we can set a problem of how to decide the tradeoff between the ML and ME principles according to an arbitrarily given sample size. We regard the situation as a metaphor of thermodynamics. The tradeoff between minimizing internal energy and maximizing entropy in thermodynamics seems to correspond to the tradeoff between maximizing likelihood and entropy in statistics, and temperature can be regarded as a parameter that brings harmony of the two amounts. These amounts can be treated in a unified manner as a *free energy* that is well known in thermal physics. Furthermore, this tradeoff can be determined using the *minimum free energy* (MFE) principle.

During recent decades, many researchers investigated Bayesian methods, which can avoid overfitting derived from using the ML with insufficient data. This can be regarded as the same problem setting described in this paper. For example, Dash et al. proposed a robust conditional independence testing procedure using Bayesian Dirichlet smoothing [16]. However, the Bayesian method presents the difficulty of deciding optimal hyperparameters simultaneously in both theoretical (e.g. [17]) and practical [18] perspectives, when no prior knowledge exists.

Similarly, temperature is an unknown parameter in the MFE principle. However, we recently presented a model of inverted temperature that is a monotonic increasing function of available data size, which we call the "Data Temperature" assumption [11]. Using this approach, we give the meaning of temperature of free energy in data/statistical science by regarding the probability fluctuation as a virtual thermal fluctuation. Furthermore, we showed that the approach is effective for parameter learning of BNs, and that the effect is not sensitive for selecting a hyperparameter. We consider that the approach can also be useful in structure learning BNs for estimating optimal entropies of the network structure. To realize this, we regard that remaining problems are how to define amounts corresponding to energies, entropies, and temperature.

3.2 Minimum Free Energy Principle

The (Helmholtz) free energies were introduced originally into the field of thermodynamics. The energies are defined such that a maximum thermodynamical work is the difference between values of free energies in two distinct states, [5] where the maximum work is obtained using an isothermal quasistatic operation from a closed system under the condition of a constant temperature. Therefore, the free energy can be regarded as an amount, in a constant temperature, corresponding to a potential energy in dynamics (e.g., gravitational and electro-magnetic potential energy). In this meaning, the free energy is viewed as an amount that is extracted freely from a thermodynamical system.

We regard the free energy in information systems as an amount that has a similar effect in thermodynamical systems: it is extracted freely from a data system under a given data size (corresponding to inverted temperature). This property is apparently preferred for various tasks such as inference, learning, and estimation under a finite available data size because we wish to obtain maximum effective information from limited exploitable data.

We use a principle of minimum free energy (MFE) for statistical testing of conditional independence. A free energy F is defined by an internal energy U, an entropy H, and inverted temperature β_0 ($= 1/temperature$) as

$$F := U - \frac{H}{\beta_0} \ , \tag{3}$$

where (inverted) temperature β_0, which is a parameter, balances the respective contributions to F of U and H. According to the principle of MFE, given some temperature β_0, the stable state of the system is realized to minimize F [5], where minimizing U and maximizing H are balanced.

3.3 Representation of Free Energy in Probabilistic Models

Different from usual application of the MFE principle in data science, we start with description of the free energy definitely as a function of internal energy, entropy, and temperature to recognize clearly important properties of temperature and use effectively free energies. Fortunately, entropy was introduced into information theory by Shannon. It has since become a fundamental concept in computer science and statistical science [10]. Therefore, we define the entropy of a random variable X as Shannon entropy. We hope that the entropy serves to avoid overfitting for small samples. We adopt the Kullback–Leibler (KL) divergence between two probabilistic distributions, which are an empirical distribution and a true distribution. Here, we follow the "Data Temperature" assumption [11], which makes the MFE principle express a harmony between the ML and ME principle according to available data size: *temperature is defined as a monotonic function of the available data size such that temperature* $\beta_0 \to \infty$ *if data size* $N \to \infty$*, and* $\beta_0 \to 0$ *if* $N \to 0$.

3.4 An MFE Representation of Hypothesis Testing on BNs

We represent the conditional independence tests using the MFE principle. To do so, as in the usual manner [3,4], we represent the null hypothesis as conditional independent relations, and the opposite hypothesis as conditional dependent relations between two variables X and Y given conditional sets $\boldsymbol{Z}$.

We define the internal energies for each hypothesis. First, we represent the internal energy U such that the relative entropy (KL divergence) between the graphs expressing the null hypothesis (expressed as $\mathbb{H}_1$, corresponding distributions as $\hat{P}_1$) and the true graphs (as $\mathbb{H}_0$, corresponding as P_0), where $\hat{P}_1$ of the null hypothesis is defined as a maximum likelihood distribution. Therefore, we can define an internal energy U_1 which expresses the null hypothesis such as

$$\begin{aligned} U_1(X, Y, \boldsymbol{Z}) :&= -D(\,\hat{P}_1(X, Y, \boldsymbol{Z}) \,||\, P_0(X, Y, \boldsymbol{Z})\,) \\ &= \sum_{x,y,\boldsymbol{z}} \hat{P}(x, y, \boldsymbol{z}) \log \frac{P(x, y \,|\, \boldsymbol{z})}{\hat{P}(x \,|\, \boldsymbol{z})\hat{P}(y \,|\, \boldsymbol{z})}\ , \end{aligned} \tag{4}$$

where $\hat{P}$ is a maximum likelihood distribution and P is a distribution that will be estimated using the MFE principle with a "Data Temperature" model. In turn, the internal energy U_2 expresses the opposite hypothesis, which expresses a dependent relation as

$$\begin{aligned} U_2(X, Y, \boldsymbol{Z}) :&= -D(\,\hat{P}_2(X, Y, \boldsymbol{Z}) \,||\, P_0(X, Y, \boldsymbol{Z})\,) \\ &= \sum_{x,y,\boldsymbol{z}} \hat{P}(x, y, \boldsymbol{z}) \log \frac{P(x, y \,|\, \boldsymbol{z})}{\hat{P}(x, y \,|\, \boldsymbol{z})}\ . \end{aligned} \tag{5}$$

In the next step, we define the entropy term with respect to each hypothesis. Probability distributions which constitute the entropy are estimated under given available samples. We describe the entropy of the null hypothesis as

$$H_1(X, Y, \boldsymbol{Z}) := -\sum_{x,y,\boldsymbol{z}} P(x, y, \boldsymbol{z}) \log(P(x \,|\, \boldsymbol{z}) P(y \,|\, \boldsymbol{z}) P(\boldsymbol{z})) \ . \tag{6}$$

The other entropy, that of the opposite hypothesis, is

$$H_2(X, Y, \boldsymbol{Z}) := -\sum_{x,y,\boldsymbol{z}} P(x, y, \boldsymbol{z}) \log(P(x, y \,|\, \boldsymbol{z}) P(\boldsymbol{z})) \ . \tag{7}$$

Now we are almost prepared to express the free energy of each hypothesis. We regard the temperature in each hypothesis (β_1 and β_2) as a *global temperature* over related variables. According to the "Data Temperature" assumption, we can consider that $\beta_1 = \beta_2 = \beta_0$, which means the same sample size. Therefore, the difference of the free energy of each hypothesis is

$$F_1(X, Y, \boldsymbol{Z}) - F_2(X, Y, \boldsymbol{Z}) = \hat{I}(X; Y|\boldsymbol{Z}) - \frac{1}{\beta_0} I(X; Y|\boldsymbol{Z}) \ , \tag{8}$$

where

$$\hat{I}(X; Y|\boldsymbol{Z}) = \sum_{x,y,\boldsymbol{z}} \hat{P}(x, y, \boldsymbol{z}) \log \frac{\hat{P}(x, y \,|\, \boldsymbol{z})}{\hat{P}(x \,|\, \boldsymbol{z}) \hat{P}(y \,|\, \boldsymbol{z})} \ , \tag{9}$$

and

$$I(X; Y|\boldsymbol{Z}) = \sum_{x,y,\boldsymbol{z}} P(x, y, \boldsymbol{z}) \log \frac{P(x, y \,|\, \boldsymbol{z})}{P(x \,|\, \boldsymbol{z}) P(y \,|\, \boldsymbol{z})} \ . \tag{10}$$

According to the notation used in our previous work, we define a new parameter β, which we call "Data Temperature" hereinafter as

$$\beta := \beta_0 / (\beta_0 + 1) \ , \tag{11}$$

where if $\beta_0 \to 0$, then $\beta \to 0$ (high temperature limit); if $\beta_0 \to \infty$, then $\beta \to 1$ (low temperature limit).

For estimating the non-empirical conditional mutual information $I(X, Y|\boldsymbol{Z})$, as described above, we follow our previous work associated with parameter learning method, for which a different definition of internal energies U is needed [11]. Let $P(\boldsymbol{X})$ and $\hat{P}(\boldsymbol{X})$ respectively represent probability distributions of joint random variables $\boldsymbol{X}$ to be estimated from the MFE principle and ML principle. Internal energies $U(\boldsymbol{X})$ are defined for parameter learning as

$$U(\boldsymbol{X}) = D(\, P(\boldsymbol{X}) \,||\, \hat{P}(\boldsymbol{X}) \,) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{\hat{P}(\boldsymbol{x})} \ . \tag{12}$$

Then, using Lagrange multipliers corresponding to minimizing the free energy with a constraint as $\sum_{\boldsymbol{X}=\boldsymbol{x}} P(\boldsymbol{X} = \boldsymbol{x}) = 1$, the estimated probability $P_\beta(\boldsymbol{x})$ is expressed in Boltzmann's formula, as shown below [11].

$$P_\beta(\boldsymbol{x}) = \frac{\exp(-\beta(-\log \hat{P}(\boldsymbol{x})))}{\sum_{\boldsymbol{x}'} \exp(-\beta(-\log \hat{P}(\boldsymbol{x}')))} = \frac{[\hat{P}(\boldsymbol{x})]^\beta}{\sum_{\boldsymbol{x}'} [\hat{P}(\boldsymbol{x}')]^\beta} \tag{13}$$

Therein, $\hat{P}$ is a relative frequency: the ML estimator.

Finally, we obtain the condition of conditional independence as

$$\hat{I}(X;Y|\boldsymbol{Z}) < \frac{1-\beta}{\beta} I_\beta(X;Y|\boldsymbol{Z}) \ , \tag{14}$$

where I_β is defined as

$$\begin{aligned} I_\beta(X;Y|\boldsymbol{Z}) &= \sum_{x,y,\boldsymbol{z}} P_\beta(x,y,\boldsymbol{z}) \log \frac{P_\beta(x,y\,|\,\boldsymbol{z})}{P_\beta(x\,|\,\boldsymbol{z})P_\beta(y\,|\,\boldsymbol{z})} \\ &= \sum_{x,y,\boldsymbol{z}} P_\beta(x,y,\boldsymbol{z}) \log \frac{P_\beta(x,y,\boldsymbol{z})P_\beta(\boldsymbol{z})}{P_\beta(x,\boldsymbol{z})P_\beta(y,\boldsymbol{z})} \ . \end{aligned} \tag{15}$$

Therein, β plays only the role of a symbolic index; it does not represent a sole parameter. In each estimator, β should be calculated using the explicit model of "Data Temperature," as described in the next subsection. Therefore, β in (15) represents *local temperature*. In (14), the left-hand-side corresponds to the likelihood term, which is dominant for a large data size (large β), and the right-hand-side corresponds to the entropy term, which is dominant for a small data size (small β). We designate g^2_β and represent the conditional independence (CI) condition with it as

$$g^2_\beta = \hat{I}(X;Y|\boldsymbol{Z}) - \frac{1-\beta}{\beta} I_\beta(X;Y|\boldsymbol{Z}) < 0 \ . \tag{16}$$

This is useful for combination with the classical hypothesis tests.

In these formulations, it might seem strange that two distinct definitions of internal energies exist between hypothesis tests and parameter estimation. We express internal energies in the hypothesis tests as $D(\,\hat{P}\,||\,Q)$, where $\hat{P}$ is an ML distribution. That differs from our formula in parameter learning, where we expressed the internal energies as $D(\,P\,||\,\hat{Q})$, where $\hat{Q}$ is an ML distribution. This difference is pointed out from the perspective of *information geometry* [12]. From an information theoretical viewpoint, hypothesis testing is related to the large deviation theorem via Sanov's theorem [10], and then, from an information geometrical perspective, it can be interpreted as a $\nabla^{(\mathrm{e})}$-projection, whereas the ML estimation can be interpreted as $\nabla^{(\mathrm{m})}$-projection [12]. In other words, the hypothesis testing and ML estimation are then different concepts in view of information theory. Consequently, in the definitions for learning structures and parameters, the difference is reasonable from this perspective.

3.5 "Data Temperature" Model

In searching for the values of β, we use our simple model of temperature, which is proposed as a function of data size N [11]. The model function of β is defined as

$$\begin{aligned} \beta &:= 1 - \exp\left(-\frac{N}{\gamma N_c}\right) \ , \\ \gamma &:= |\boldsymbol{X}| - 1 \ , \end{aligned} \tag{17}$$

where γ is defined as the degrees of freedom of related random variables $\boldsymbol{X}$, and where N_c is a decoupling constant, which can be regarded as a hyperparameter for β. We use N_c as only one *common* hyperparameter in learning of both parameter and structure. This explicit model shows good performance and robustness against selected hyperparameters N_c in classification tasks using Bayesian network classifiers with structure learning [11].

3.6 Asymptotic Theoretical Analysis

We hope that the proposed method has consistency with the classical hypothesis test for an asymptotic region because it is theoretically justified. However, our conditional independence conditions using the inequality (14) cannot be used straightforwardly for large data sizes because $g_\beta^2 \geq 0$ always for sufficiently large data size. That is true because $\hat{I}(X;Y|\boldsymbol{Z}) \geq 0$ and $[(1-\beta)/\beta]\, I_\beta(X;Y|\boldsymbol{Z}) \to 0$ as β goes to 1 (as N becomes sufficiently large), which means that our method would produce an overly dense graph for sufficiently large data size. In such regions, the effect of enlarging the entropy term has vanished and the likelihood term has become dominant. However, different from parameter learning, hypothesis testing for BNs means that extra edges should be removed even for a large sample size, based on *Occam's razor* [13]. This connecting problem is solved as described below.

For a large sample size region, we wish to use the G^2 statistic for conditional independence testing, which is often used [3,4]. The G^2 test is used to identify $Ind\,(X, Y|\boldsymbol{Z})$, by which the null hypothesis of conditional independence is represented. Let $N_{xy\boldsymbol{z}}$ represent the number of times in the data where $X = x, Y = y$ and $\boldsymbol{Z} = \boldsymbol{z}$. We define $N_{x\boldsymbol{z}}, N_{y\boldsymbol{z}}$, and $N_{\boldsymbol{z}}$ similarly. Consequently, the G^2 statistic is defined as follows:

$$G^2 = 2 \sum_{x,y,\boldsymbol{z}} N_{xy\boldsymbol{z}} \log \frac{N_{xy\boldsymbol{z}} N_{\boldsymbol{z}}}{N_{x\boldsymbol{z}} N_{y\boldsymbol{z}}} \ . \tag{18}$$

The degrees of freedom *df* are defined as

$$df = (|X| - 1)(|Y| - 1) \prod_{Z \in \boldsymbol{Z}} |Z| \ , \tag{19}$$

where we designate $|X|$ as the number of states in X. It is noteworthy that the G^2 statistics have a relation with the empirical mutual information with data size N [14] as

$$G^2 = 2N\, \hat{I}(X;Y|\boldsymbol{Z}) \ . \tag{20}$$

The statistic is proven to be approximated asymptotically to a χ^2 distribution with degrees of freedom *df* [14]. Therefore, in a large sample size region, we should set the condition in which the null hypothesis (i.e., conditional independence) is not rejected, as

$$G^2 < \chi^2_{\alpha,df} \ , \tag{21}$$

where α is a significance level such as 0.05, and where df are the degrees of freedom, as defined in (19).

Here, we intend to connect the classical condition with the MFE condition. We define a formal correspondence amount G^2_β to G^2, using (16) and (20) as

$$G^2_\beta := 2Ng^2_\beta = G^2 - 2N \frac{1-\beta}{\beta} I_\beta(X;Y|\boldsymbol{Z}) \ . \tag{22}$$

Using the explicit model of β expressed in (17), in an asymptotic region,

$$G^2_\beta = G^2 - 2N \frac{\exp(-N/\gamma N_c)}{1-\exp(-N/\gamma N_c)} I_\beta(X;Y|\boldsymbol{Z}) \to G^2 \ . \tag{23}$$

Then, we can treat the MFE and the classical condition uniformly because a condition $G^2_\beta < \chi^2_{\alpha,df}$ can include the CI condition (16) and the classical condition (21). Even when the data size is small and $G^2_\beta \geq 0$, we conduct the classical hypothesis tests because our method was shown to generate pseudo-samples similarly to the Bayesian methods [19]. Consequently, we can conduct conditional independence tests on variables X and Y given $\boldsymbol{Z}$ using the MFE principle and G^2 tests as described below.

- If $G^2_\beta < 0$, because of MFE principle, then we set $X \perp\!\!\!\perp Y \,|\, \boldsymbol{Z}$ (conditional independence),
- else if $0 \leq G^2_\beta < \chi^2_{\alpha,df}$, because of the classical test, then we set $X \perp\!\!\!\perp Y \,|\, \boldsymbol{Z}$,
- else, we set $X \not\perp\!\!\!\perp Y \,|\, \boldsymbol{Z}$ (conditional dependence).

We designate this conditional independence method as MFE-CI.

4 Experiments

We next demonstrate the performances of our approach compared with traditional statistical testing methods. Some experiments of learning BNs can be done using the PC algorithm [3], which is a well known benchmark algorithm of constraint-based methods, embedding our conditional independence tests or classical independence tests using χ^2 distributions with fixed significant level $\alpha = 0.05$ for each hypothesis test of conditional independence. We implemented the PC algorithm for embedding the MFE-CI method using C++ programming language. The PC algorithm, which constructs partial DAGs (PDAGs), is the following [20]:

The PC algorithm
1. Assume a non-negative integer $m = 0$.
2. Let $\mathbb{G}$ *be a complete undirected graph.*
3. Repeat:
(a) For all pairs of variables (X, Y)*, check* $Ind\,(X, Y|\boldsymbol{Z})$ *for all subsets* $\boldsymbol{Z}$ *such that* $|\boldsymbol{Z}| = m$ *and* $\boldsymbol{Z} \subset Adj(X)$ *or* $\boldsymbol{Z} \subset Adj(Y)$.
If there exists a $\boldsymbol{Z}$ *such as* $Ind\,(X, Y|\boldsymbol{Z})$,

then remove the edge $X - Y$ *from* $\mathbb{G}$*, and add* $\boldsymbol{Z}$ *to* $SepSet(XY)$*.*
(b) Set $m = m + 1$*.*
Until no variable has more than m *adjacencies,*
or a stopping condition is satisfied.
4. Orientation rules are performed.
5. Return the partially directed acyclic graph $\mathbb{G}$*.*

Therein, $|\boldsymbol{X}|$ denotes the size of members in $\boldsymbol{X}$; $Adj(X)$ is a set of adjacent nodes to X. The orientation rules [21], described in step 4 of the algorithm, are as follows:

4-1. If $U \notin SepSet(XY)$*, orient* $X - U - Y$ *as* $X \rightarrow U \leftarrow Y$ *(v-structure) for each uncoupled set of* X *and* Y *such as* $X - U - Y$*.*
4-2. Repeat this step while more edges can be oriented.
4-2-1. Orient $U - Y$ *as* $U \rightarrow Y$ *for each uncoupled set of* X *and* Y *such as* $X \rightarrow U - Y$*.*
4-2-2. Orient $X - Y$ *as* $X \rightarrow Y$ *for each set of* X *and* Y *such that a path exists from* X *to* Y*.*
4-2-3. Orient $U - W$ *as* $U \rightarrow W$ *for each uncoupled set of* X *and* Y *such as* $X - U - Y$*,* $X \rightarrow W$*,* $Y \rightarrow W$*, and* $U - W$*.*

The completeness of the rules was proved by Meek [22].

The PC algorithm is performed under the *faithfulness assumption* described in section 2. Consequently, the algorithm can infer correct graph structures by finding conditional independence for probability distributions. However, if the assumption is violated, even though the true graph means $Ind\,(X, Y|\boldsymbol{Z})$ for X and Y and a conditional set $\boldsymbol{Z}$, the algorithm might find another false conditional set $\boldsymbol{Z'}$ for the test between X and Y, and then add $\boldsymbol{Z'}$ to *SepSet(XY)*. This false detection has *no influence* on removing the edge between X and Y correctly. However, the algorithm decides the wrong direction of edges using the orientation rules described above. In this situation, finding correctly conditional sets has a large influence on the directionality of edges in BNs.

We conducted the simulation study with various quantities of variables: {10, 20, 40, 80}, where each variable has all four possible states, and with networks of two types, i.e. the sparser and denser graphs, where sparser cases have the same number of edges as variables; the denser cases have twice. For each such graph, a random structure network was constructed with conditional probability tables (CPTs) of five types that were set by random numbers. The available sample size is varied in a range of {500, 1000, 2500, 5000, 10000}. When the number of conditional sets $|\boldsymbol{Z}|$ is large, the number of CI tests is intractably large because of a combinatorial explosion. Therefore, we did not perform CI tests and assume conditional dependence when $|\boldsymbol{Z}| \geq 5$. We selected a value of the hyperparameter N_c for β in (17) as 2.0, which shows good performance in preliminary experiments.

We set the performance criterion as counting *added edges*, *removed edges*, and *reversed edges*. Counting added edges expresses the consequence where two variables X and Y are not adjacent in original BNs but where an edge exists between them in reconstructed BNs. On the other hand, counting removed edges mean the opposite. Counting reversed edges means that if $X \rightarrow Y$ in the original, then $Y \rightarrow X$ in the output. The results are presented in Table 1 for sample sizes of 500, 1000, and 2500, and in Table 2 for sample sizes of 5000 and 10000. The values in the tables are averaged values of simulations for five different randomly set CPTs. We designate the PC algorithm with a standard G^2 test as *Std-PC* or *Std*, and the PC embedded with the MFE-IC as *MFE-PC*. These tables show that the counted quantities of extra added edges were very small, even for a small sample size such as 500 and even for denser structures. In contrast, quantities of removed edges are very large in both Std-PC and MFE-PC. The MFE-PC removed true edges more than Std-PC. In reversed checks, many errors were found in Std-PC. These characteristics were noticeable in large and denser networks. We discuss these results later. A key is apparently the *faithfulness condition* for understanding the results.

The MFE-PC seemed to underscore the effectiveness for deciding the direction of edges. It might be unfair, however, to conclude that because the MFE-PC removed more edges than Std-PC. Therefore, we defined *reversed ratio* as (number of reversed edges)/((true number of edges) – (number of removed)). Results of *reversed ratio* for denser networks are portrayed in Fig. 1, where the horizontal axis expresses the true number of edges, the vertical axis expresses the *reversed ratio*, and G2 and MFE respectively signify Std-PC and MFE-PC. These figures show that the MFE-PC outperforms Std-PC in deciding the direction of edges, especially for denser networks, even using samples such as 5000, which are not small.

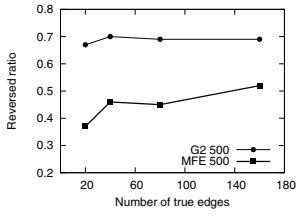
We discuss these comparative results. As designed using the MFE principle, MFE-PC is expected to remove edges more than PC does for two reasons. The first is that MFE-PC performs CI tests in more cases than Std-PC, which does the test only for sufficiently large data size. For example, even when the data

Table 1. Results for the simulation using data sizes of 500, 1000, and 2500

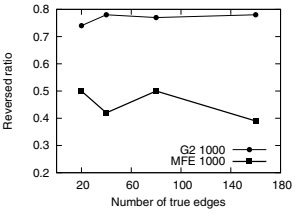
		500				1000				2500			
		Sparser		Denser		Sparser		Denser		Sparser		Denser	
Type	Nodes	Std	MFE	Std	MFE	Std	MFE	Std	MFE	Std	MFE	Std	MFE
Added	10	0	0	0.6	0.2	0.4	0	1.8	0	0	0	0	0
	20	0	0	0.4	0.4	0.2	0	1.2	0	0	0	0.2	0
	40	0.2	0	0.4	0	0.6	0.4	0.8	0	0	0	0	0
	80	0.6	0.4	1.4	0.8	0.4	0.2	2.2	0.2	0.2	0.2	0	0
Removed	10	3.0	3.4	9.0	14.0	2.4	2.8	3.6	11.6	1.8	1.8	4.4	8.6
	20	4.2	7.6	19.0	26.2	3.0	5.8	11.4	22.0	2.0	2.8	12.8	18.6
	40	11.2	15.8	41.6	53.6	6.4	12.0	25.0	46.6	6.0	7.2	26.0	37.2
	80	21.4	32.0	81.2	109	11.0	21.2	48.6	93.8	9.0	12.0	54.4	73.8
Reversed	10	1.8	1.4	7.4	2.2	0.8	0.8	12.2	4.2	0.6	0.6	9.2	5.2
	20	5.4	2.0	14.6	6.4	5.2	2.6	22.2	7.6	2.8	2.4	13.2	7.6
	40	10.6	5.2	26.6	11.8	10.2	4.0	42.6	16.8	6.6	6.4	31.8	19.4
	80	18.8	10.6	54.2	26.6	21.0	11.2	86.4	26.0	11.6	10.8	56.6	38.8

Table 2. Results for simulations using data sizes 5000 and 10000

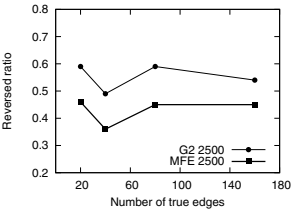
		5000				10000			
		Sparser		Denser		Sparser		Denser	
Type	Nodes	Std	MFE	Std	MFE	Std	MFE	Std	MFE
Added	10	0	0	1.2	0.2	0	0	0	0
	20	0	0	0.4	0.0	0	0	0	0
	40	0.4	0.4	0.0	0.0	0	0	0.2	0.2
	80	0.4	0.4	0.0	0.0	0.4	0.4	0	0
Removed	10	1.0	1.0	1.8	6.4	0.6	0.6	2.6	4.6
	20	0.4	1.4	6.2	14.6	0.6	0.6	8.6	10.4
	40	2.6	4.6	16.0	30.0	1.8	2.0	20.2	24.4
	80	4.6	7.2	24.2	59.2	4.4	4.6	40.0	48.0
Reversed	10	0.6	0.6	7.2	4.4	4.0	4.0	4.2	3.6
	20	2.6	2.4	20.8	9.6	6.2	6.2	12.0	11.4
	40	5.0	3.8	36.6	20.8	14.0	14.2	21.0	19.0
	80	8.4	7.8	78.0	38.0	24.0	24.2	48.5	44.3



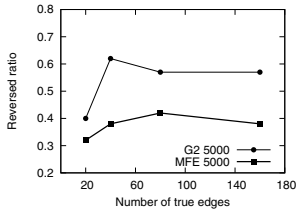
(a) Sample size = 500.



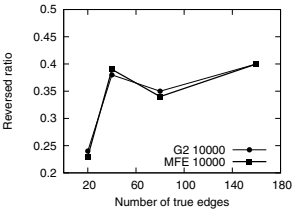
(b) Sample size = 1000.



(c) Sample size = 2500.



(d) Sample size = 5000.



(e) Sample size = 10000.

Fig. 1. Ratio of reversed edges in the resultant graphs with denser BNs from use of a standard PC and PC embedded with the MFE-IC method

size $N = 5000$, Std-PC was unable to perform CI tests for $|\boldsymbol{Z}| \geq 3$ in this simulation, which implies that Std-PC might sometimes correctly happen to maintain some existing edges. The second is that, when confidence of existing edges is small because the data size is insufficiently large, the null hypothesis is not rejected because of the effect of maximum entropy. Therefore, MFE-IC method seemingly tends to prefer sparser graphs. These mean that MFE-IC method does draw edges only when high confidence for dependence is obtained. Additionally, we can comment on the fact that Std-PC incorrectly decided the direction of edges more than MFE-PC. This fact means that Std-PC detected conditional independence for invalid conditional sets $\boldsymbol{Z}$. In this case, wrong

v-structures, divergence connections such as $X \leftarrow U \rightarrow Y$, and serial connections such as $X \rightarrow U \rightarrow Y$ were generated. This suggests that the simulation had difficulty realizing the faithfulness assumption. In other words, unfortunately, $Ind\,(X;Y|\boldsymbol{Z}) \Rightarrow Dsep^{\mathbb{G}}(X;Y|\boldsymbol{Z})$ was often violated in this simulation. This situation was also reported by Ramsey et al. for linear Gaussian models of DAGs [23]. However, we found the large difference of the reversed ratio, which shows that MFE-PC correctly found the conditional sets more than Std-PC. According to the discussion, we regard that MFE-IC method is especially preferable for causal discovery, where existing edges are expected to represent definite direct dependence between variables, and where direction has important meanings. In contrast, MFE-IC seems to be unsuitable for finding BNs in view of the predictive sense, for which the edges are allowed to be reversed for maintaining adequate parametric space size.

5 Conclusion

We proposed a method for improvement of conditional independence (CI) testing in small samples, which is a weak point of constraint-based learning Bayesian networks using the classical hypothesis tests. To do this, we introduced the minimum free energy principle with a "Data Temperature" assumption that relates probabilistic fluctuation to virtual thermal fluctuation. We defined a free energy using Kullback–Leibler divergence, which corresponds to an information-geometric view. This CI method incorporates the maximum entropy and maximum likelihood principles and converges to the classical hypothesis tests in asymptotic regions.

We also demonstrated the effectiveness of our method by embedding it in the well known PC algorithm. The results show that our method correctly identified the direction of the edges better than the standard tests did, which is expected to be effective for causal discovery where the orientation of edges is significant.

Acknowledgements

One author (T.I.) particularly thanks M. Kyojima, K. Shinozaki and T. Tanaka of Fuji Xerox Co., Ltd. for support and encouragement, and thanks T. Ogawa of the University of Electro-Communications for advice related to information geometry. The authors thank anonymous reviewers for fruitful comments, which help them improve the paper.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo (1988)
2. Heckerman, D., Geiger, D., Chickering, D.: Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20, 197–243 (1995)

3. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search, 2nd edn. MIT Press, Cambridge (2000)
4. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max–min hill-climbing Bayesian network structure learning algorithm. Machine Learning 65(1), 31–78 (2006)
5. Callen, B.H.: Thermodynamics and An Introduction to Thermostatistics, 2nd edn. John Wiley & Sons, Hoboken (1985)
6. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of English words. In: Proc. of Annual Meeting on Association for Computational Linguistics (ACL 1993), pp. 183–190 (1993)
7. Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Conference on Uncertainty in Artificial Intelligence (UAI 1999), pp. 289–296 (1999)
8. LeCun, Y., Huang, F.J.: Loss functions for discriminative training of energy-based models. In: Proc. of International Workshop on Artificial Intelligence and Statistics (AISTATS 2005), pp. 206–213 (2005)
9. Lehmann, L.E.: Testing Statistical Hypotheses, 2nd edn. John Wiley & Sons, New York (1986)
10. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. John Wiley & Sons, Hoboken (2006)
11. Isozaki, T., Kato, N., Ueno, M.: Minimum free energies with "data temperature" for parameter learning of Bayesian networks. In: Proc. of IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008), pp. 371–378 (2008)
12. Amari, S., Nagaoka, H.: Method of Information Geometry. Oxford University Press, New York (2000)
13. Pearl, J.: Causality, models, reasoning, and inference. Cambridge University Press, New York (2000)
14. Kullback, S.: Information Theory and Statistics. Dover Publications, Mineola (1968)
15. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning 29(2-3), 131–163 (1997)
16. Dash, D., Druzdzel, M.J.: Robust independence testing for constraint-based learning of causal structure. In: Proc. of Conference on Uncertainty in Artificial Intelligence (UAI 2003), pp. 167–174 (2003)
17. Clarke, B., Barron, A.: Jeffreys' prior is asymptotically least favorable under entropy risk. Journal of Statistical Planning and Inference 41, 37–60 (1994)
18. Silander, T., Kontkane, P., Myllymaki, P.: On sensitivity of the map Bayesian network structure to the equivalent sample size parameter. In: Proc. of Conference on Uncertainty in Artificial Intelligence (UAI 2007), pp. 360–367 (2007)
19. Isozaki, T., Kato, N., Ueno, M.: "Data temperature" in minimum free energies for parameter learning of Bayesian networks (to appear)
20. Neapolitan, R.E.: Learning Bayesian Networks. Prentice-Hall, Upper Saddle River (2004)
21. Verma, T., Pearl, J.: An algorithm for deciding if a set of observed independencies has a causal explanation. In: Proc. of Conference on Uncertainty in Artificial Intelligence (UAI 1992), pp. 323–330 (1992)
22. Meek, C.: Causal inference and causal explanation with background knowledge. In: Proc. of Conference on Uncertainty in Artificial Intelligence (UAI 1995), pp. 403–410 (1995)
23. Ramsey, J., Spirtes, P., Zhang, J.: Adjacency-faithfulness and conservative causal inference. In: Proc. of Conference on Uncertainty in Artificial Intelligence (UAI 2006), pp. 401–408 (2006)