# Minimum Free Energies with "Data Temperature" for Parameter Learning of Bayesian Networks

Takashi Isozaki[1,2], Noriji Kato[2], Maomi Ueno[1]
[1]Graduate School of Information Systems
The University of Electro-Communications, Japan
{t-isozaki, ueno}@ai.is.uec.ac.jp
[2]Research and Technology Group, Fuji Xerox Co., Ltd., Japan
noriji.kato@fujixerox.co.jp

## Abstract

*Maximum likelihood (ML) method for estimating parameters of Bayesian networks (BNs) is efficient and accurate for large samples. However, ML suffers from overfitting when the sample size is small. Bayesian methods, which are effective to avoid overfitting, have difficulties for determining optimal hyperparameters of prior distributions with good balance between theoretical and practical points of view when no prior knowledge is available.*

*In this paper, we propose an alternative estimation method of the parameters on BNs. The method uses a principle, with roots in statistical thermal physics, of minimizing free energy. We propose an explicit model of the temperature, which should be properly estimated. We designate the model "data temperature". In assessments of classification accuracy, we show that our method yields higher accuracy than that of the Bayesian method with normally recommended hyperparameters. Moreover, our method exhibits robustness for the choice of introduced hyperparameters.*

## 1. Introduction

Bayesian networks (BNs) [14], which are probabilistic models, are well suited for representing knowledge under uncertainty. They can often be expressed in compact forms and can be interpreted as causal models. They are therefore used in various fields and applications such as expert systems, user modeling, and computational biology. In fact, BNs are expressed as directed acyclic graphs (DAGs) in which random variables and their dependencies are respectively associated with nodes and directed edges. Qualitative relationships are expressed as their structures and quantitative relationships are expressed as their parameters. The parameters of BNs are conditional probabilities assigned to nodes, into which joint probability is decomposed. In this paper, we concentrate our efforts on parameter learning of BNs in cases where no prior knowledge is available.

The maximum likelihood (ML) method is often used for estimating conditional probabilities. However, when training data are few, the estimated parameters of BNs with ML are likely to fall into overfitting to the data. Bayesian methods, which involve prior distributions, are effective to avoid this problem. For expressing the prior distributions of discrete variables, the Dirichlet distribution function is usually used [6]. This prior has hyperparameters, which means prior imaginary instances (we designate it as $\alpha$ consistently in this paper). Some studies have used hyperparameters such as $\alpha = 1$ (meaning uniform prior distributions) or $\alpha = 0.5$ (meaning non-informative prior distributions) [3, 19] when no prior knowledge exists. However, it remains controversial to decide hyperparameters of prior distributions theoretically. Furthermore, from a practical perspective, Yang and Chang reported that $\alpha = 10$ is best for learning BNs [21]. Therefore, it seems difficult to find optimal $\alpha$ consistently from both perspectives.

Another approach to avoid overfitting to data is incorporation of proper entropy into estimators of the parameters. One effective idea for treating entropy is using the principle of minimum (Helmholtz's) free energy (MFE), which has its roots in statistical thermal physics. The free energy $F$, if described in the manner of physics, consists of (internal) energy $U$, entropy $H$, and (inverted) temperature $\beta$. In fact, $\beta$ balances the contributions of $U$ and $H$ to the free energy.

In recent years, the MFE principle and similar concepts have been used in wide areas of computer science. However, to our knowledge, the meaning of temperature has not been established yet. Consequently, $\beta$ is treated in various ways at this stage: annealing parameters ([15], [20]), fixed parameters ([1], [12], [22]), or optimizable parameters in

1082-3409/08 $25.00 © 2008 IEEE
DOI 10.1109/ICTAI.2008.56

each dataset ([10]). The pre-fixed method seems to have a poor foundation and the optimizing method using held-out data is not efficient in computational cost and not effective for very small data size. Consequently, universal or robust values of $\beta$ have not been reported to date.

Differently from the approach described above, we take *model-based approaches* for the $\beta$. For that purpose, a hyperparameter is introduced for $\beta$. Using this approach, we intend to explore robust estimation methods against the hyperparameter for BNs. In this paper, we propose a meaning of $\beta$ in the MFE principle by combining thermal fluctuation with probabilistic fluctuation, and an explicit model of $\beta$ as a result of our interpretation of the role of $\beta$. Then we assessed our method with respect to accuracies and robustness in relation to classification tasks using real world data in comparison to ML and the Bayesian method.

This paper is organized as follows: After a brief description of the background of BNs and their parameter learning, we provide an overview of free energy and a definition of it for estimating the parameters. Our concept of temperature in data science and an explicit model are introduced in Section 3. Some experiments are conducted for investigating the effectiveness of our method in Section 4. Finally, discussions of the results are presented in Section 5.

## 2. Background

### 2.1. Bayesian networks

In fact, BNs are directed acyclic graph (DAG) representations of joint probability distributions of a set of $n$ random variables $\{X_1, \ldots, X_n\}$, which is decomposed into products of conditional probabilities of $X_i$ given its parent set $Pa(X_i)$ as

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

A random variable is defined as a node $X_i$ in the graph and its parents are defined as nodes, from which edges extend to $X_i$. In BNs, each node is conditionally independent of its non-descendants given its parents. To use BNs effectively, we must be careful to decompose the joint probabilities correctly into products of conditional probabilities, and be careful to determine the conditional probabilities. The former means that it is necessary to construct a correct structure; the latter underscores the necessity of estimating correct conditional probabilities. Both are designed using expert knowledge or trained using algorithms. In the next subsection, we explain the method of parameter learning of BNs.

### 2.2. Parameter learning of BNs

For learning parameters of BNs, two methods are often used: maximum likelihood (ML) and Bayesian methods. We assume a complete dataset with no missing data values. Applying Lagrangian multipliers to the log likelihood function with multipliers to constrain the parameters to a normalized probability distribution, it is easy to show that ML estimators of conditional probabilities (we also express them as $\{\boldsymbol{\theta}\}$) are

$$\theta_{ijk} := P(x_i^k \,|\, \boldsymbol{\pi}_i^j) = \frac{N_{ijk}}{\sum_{k'=1}^{r_i} N_{ijk'}}, \tag{2}$$

where $i$ is an index of $X$, $j$ is an index of parent nodes' configurations, $r_i$ is the number of states of the $X_i$, $k$ and $k'$ are $X_i$'s state ($k, k' \leq r_i$), and $N_{ijk}$ is the number of cases in the dataset in which $X_i = x_i^k$, given the condition that $Pa(X_i) = \boldsymbol{\pi}_i^j$ [9].

When we use Bayesian statistics on discrete random variables, the estimators are often obtained using a posterior mean or by maximizing a posterior, where the Dirichlet distributions and their hyperparameters (smoothing parameters) $\{\boldsymbol{\alpha}\}$ are usually used. According to Bayesian statistics [6], a posterior probability density function $\rho(\theta|d)$ given data $d$ is expressed as follows from Bayes' theorem: $\rho(\theta|d) \propto \rho(\theta) f(d|\theta)$, where $f(d|\theta)$ is the likelihood function and $\rho(\theta)$ is a prior distribution. In discrete variables, the likelihood function is the multinomial distribution function. Prior and posterior functions are both written as Dirichlet distributions when $\rho(\theta)$ and $\rho(\theta|d)$ are both represented as *natural conjugate family distributions* of the likelihood function. Then we obtain the BN parameters by taking the posterior mean of the Dirichlet distributions as

$$\theta_{ijk} := P(x_i^k \,|\, \boldsymbol{\pi}_i^j) = \frac{\alpha_{ijk} + N_{ijk}}{\sum_{k'=1}^{r_i} (\alpha_{ijk'} + N_{ijk'})}. \tag{3}$$

As expressed in equation (3), it is clear that the hyperparameters, $\boldsymbol{\alpha}$, play a role in avoiding overfitting because $\boldsymbol{\alpha}$ can contribute to some extent when the data size is not large. However, because of the lack of intelligible meaning of $\alpha$ in a case without prior knowledge about the data, $\alpha$ is often assigned manually to various values including $\alpha = 1$ that means uniform prior distributions and $\alpha = 0.5$ that means non-informative prior distributions [3, 19].

## 3. The proposed method

We minimize the free energy with finite temperature for estimating proper parameters of BNs instead of maximizing likelihood or maximizing expected values of the Bayesian posterior distributions. For presenting the approach, we must properly define entropy, energy, and temperature. We particularly regard temperature as important: it decides the degree of contribution of entropy to the free energy.

### 3.1. The minimum free energy principle

In statistical physics, the Helmholtz free energy $F$ of a system is defined using internal energy $U$, entropy $H$, and the (inverted) temperature $\beta_0$ $(= 1/\text{temperature})$ as

$$F := U - \frac{H}{\beta_0}, \tag{4}$$

where (inverted) temperature $\beta_0$, which is a parameter, balances contributions of $U$ and $H$ to $F$. According to the principle of MFE, given some temperature $\beta_0$, the stable state of the system is realized to minimize $F$.

We use this principle for the parameter learning. We denote random variables as $X$, which are assumed to be discrete variables. For a definition of entropy terms, we adopt Shannon's entropy using probability distributions $P(X)$ as

$$H(X) := -\sum_x P(X = x) \log P(X = x). \tag{5}$$

We assume that the system variable is a probability distribution. We define $U$ as the Kullback-Leibler (KL) divergence, which represents the similarity or distortion between the hidden true distribution and the distribution estimated using the ML method because we incorporate the ML principle under a large sample limit. Hence, the internal energy is defined as

$$\begin{aligned} U(X) &:= D(P(X) \,||\, \hat{P}(X)) \\ &= \sum_x P(X = x) \log \frac{P(X = x)}{\hat{P}(X = x)}, \end{aligned} \tag{6}$$

where $\hat{P}(X)$ is the probability distribution estimated using the ML method and $P(X)$ is the true probability distribution. It is noteworthy that estimations obtained by minimizing the KL divergence in equation (6) are equivalent to ML estimations.

The probability distribution parameterized by $\beta_0$ is the solution of the minimization with a constraint as $\sum_{X=x} P(X = x) = 1$. Therefore, it is solved using Lagrangian multipliers. The Lagrangian $L$ is expressed as

$$\begin{aligned} L &= F + \lambda(\sum_x P(x) - 1) \\ &= \frac{1 + \beta_0}{\beta_0} \sum_x P(x) \log P(x) - \sum_x P(x) \log \hat{P}(x) \\ &\quad + \lambda(\sum_x P(x) - 1), \end{aligned} \tag{7}$$

where $\lambda$ is the Lagrange multiplier. For later convenience, we define a parameter $\beta$ transformed from the $\beta_0$, as

$$\beta := \frac{\beta_0}{\beta_0 + 1}. \tag{8}$$

In relation to that expression, if $\beta_0 \to 0$, then $\beta \to 0$ (high temperature limit); if $\beta_0 \to \infty$, then $\beta \to 1$ (low temperature limit). We designate the $\beta$ temperature later. Then the solution is derived from the partial derivative: $\partial L / \partial P(x) = 0$. Therefore, the estimated parameter $P_\beta(X)$ is expressed in the form of Boltzmann's law, which is well known in statistical physics, as

$$P_\beta(X = x) = \frac{\exp(-\beta(-\log \hat{P}(X = x)))}{\sum_{x'} \exp(-\beta(-\log \hat{P}(X = x')))}. \tag{9}$$

Practically, we use the equivalent form

$$P_\beta(X = x) = \frac{\hat{P}^\beta(X = x)}{\sum_{x'} \hat{P}^\beta(X = x')}, \tag{10}$$

where $\hat{P}$ is the relative frequency, i.e. the ML estimator. These formulas ((9) or (10)) have been reported elsewhere, including [10, 20]. However, the combination of the explicit definitions of $U$ such as eq. (6) and the transformation in eq. (8) have not been reported in the literature, which can more easily lead us to intuitive comprehension of the role of minimizing the free energy and temperature $\beta$ in data science. Therefore, we can proceed to modeling $\beta$. In closing of this subsection, it is also noteworthy that the principle of minimizing free energies can be regarded as an extension of minimizing KL divergences between true distributions and ML-estimated distributions by defining a *tempered KL divergence* denoted as tKL, which is defined as follows:

$$\text{tKL} := F = \sum_x P(x) \log \frac{P(x)^{1/\beta}}{\hat{P}(x)}. \tag{11}$$

### 3.2. Introducing "Data Temperature"

From the definitions of $U$, $H$, and $F$ given above, it is apparent that parameter estimators by MFE tend to be ML estimators at low temperature (large $\beta$) and tend to be dominated by the entropy at high temperature (small $\beta$). On the other hand, from the view of data science, we hope to realize ML-like estimators for large samples and to avoid overfitting for small samples. We therefore relate temperature to the number of samples as follows. *Large* sample size corresponds to *low* temperature, and *small* sample size corresponds to *high* temperature. In other words, probabilistic fluctuation, which is large for small data size, is regarded as thermal fluctuation, which is large for high temperature, and vice versa in our approach. We designate this concept "data temperature".

Based on the assumption of the relationship described between data size and $\beta$, we can express $\beta$ explicitly as some monotonic function of the number of samples, which enables us to leverage the "data temperature" concept effectively. Although the exact mode of measuring $\beta$ is left open, some clues for modeling $\beta$ exist. First, $\beta$ approaches $1$ such that estimators approach ML estimators when the data size is large, whereas $\beta$ approaches $0$ such that estimators are uniform for internal states when the data size is small. The larger the data size $N$ is, the smaller the difference coefficient of $\beta$ for $N$ seems to be. On the other hand, the smaller $N$ is, the larger the difference coefficient seems to be. Therefore, a reasonable model would be a convex upward function that fulfills the boundary conditions described above. Next, the necessary data size seems to be dependent on the degrees of freedom of the random variables $X$. In other words, the more degrees of freedom the random variables have, the larger the data size we would need to regard the estimators as near-ML estimators. Then, $\gamma$ and $N_c$ are introduced for separating effects of $X$'s degrees of freedom from the $\beta$. $\gamma$ is a function of the degrees of freedom, and $N_c$ is a decoupling constant, which is introduced as a hyperparameter for $\beta$, and is expected to play some role other than that related to $\gamma$.

Then, we create a model of $\beta$ as following a simple monotone function of data size $N$, $\gamma$, and $N_c$:

$$\beta := 1 - \exp\left(-\frac{N}{\gamma N_c}\right). \tag{12}$$

Three examples of the proposed function are shown in Fig. 1, which are the cases in which $\gamma$ is assumed to be $1$ for simplicity and $N_c = 1, 2, 5$.

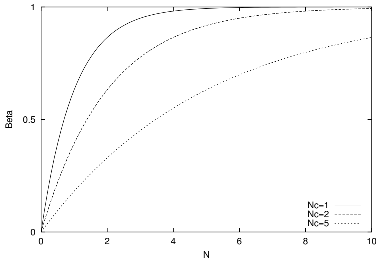

**Figure 1. Examples of the proposed exponential function.** $\gamma = 1$ **and** $N_c = 1, 2, 5$**.**

According to the description given above, the function $\gamma$ must necessarily be decided. The simplest form of $\gamma$ is one's own degrees of freedom,

$$\gamma := |X| - 1, \tag{13}$$

where $|X|$ is denoted as a number of states of a random variable $X$. We designate it as the "linear-state model". However, we consider that this model might be an approximate model under the limit of uniform distributions over the internal states. In practice, because data distributions have some bias, fewer data are necessary than in the uniform distributions. Therefore, we consider another model of $\gamma$ that is denoted as *effective* degrees of freedom, which is suppressed, because of the explanation given above, as the following:

$$\gamma := \log(|X|). \tag{14}$$

We consider that this form of $\gamma$ is an approximate expression of effective degrees of freedom. We denote the expression in equation (14) as a "log-state model". We designate these parameter learning methods as MFE with explicit $\beta$ (MFE-EB ) methods.

The relation between the temperature and data size can provide a perspective to unify the maximum likelihood and the maximum entropy principles under the minimum free energy principle with varying data size because the equation (9) is the same form of the ME principle; also, $\beta$ can be regarded as an associated constraint condition. Therefore, this estimation method bears some resemblance to the ME concept, where $\beta$ is a constraint condition.

Finally, it is straightforward to extend the above method to cases of multivariate systems by proper indexing for joint states. Therefore, Boltzmann's law, corresponding to eq. (9), becomes

$$P_\beta(\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(-\beta(-\log \hat{P}(\boldsymbol{X} = \boldsymbol{x})))}{\sum_{\boldsymbol{x}'} \exp(-\beta(-\log \hat{P}(\boldsymbol{X} = \boldsymbol{x}')))}, \tag{15}$$

where $\boldsymbol{X}$ is denoted as a multivariate set, and $\boldsymbol{x}$ is a joint state of $\boldsymbol{X}$.

### 3.3. Estimating conditional probabilities

In a BN that has discrete variables, conditional probability tables are often assumed to be independent in each conditioning event [17]. Using this local independent assumption, we naturally extend the form of $\beta$ to local forms, which we attach to each node and configuration of its parent set. Consequently, in BNs, the free energy is defined in each node and configuration. Therefore, more detailed control of entropy is possible in conditional probabilities than in multivariate joint probabilities.

In fact, $N_{ij}$ is defined as $N_{ij} = \sum_{k'} N_{ijk'}$ if the same indices $i, j, k$ and notation $N_{ijk}$ described in Section 2 are used. In addition, $\beta_{ij}$ is definable in an exponential function as

$$\beta_{ij} = 1 - \exp\left(-\frac{N_{ij}}{\gamma_i N_c}\right), \tag{16}$$

where we can adopt the "linear-state model", as

$$\gamma_i := |X_i| - 1, \tag{17}$$

or the “log-state model”, as

$$\gamma_i := \log(|X_i|). \tag{18}$$

Finally, the parameters of BNs, $\theta_{ijk}$, are expressed as the following.

$$\theta_{ijk} = \frac{\exp(-\beta_{ij}\ |MLL_{ijk}\ |)}{\sum_{k'} \exp(-\beta_{ij}\ |MLL_{ijk'}\ |)} = \frac{\hat{\theta}_{ijk}^{\beta}}{\sum_{k'} \hat{\theta}_{ijk'}^{\beta}} \tag{19}$$

Therein, $MLL_{ijk}$ is defined as an expression using ML estimators $\hat{\theta}_{ijk}$: $MLL_{ijk} = \log \hat{\theta}_{ijk} \leq 0$.

## 4. Experiments

In the experiments described in this section, we investigate whether the MFE-EB method can avoid overfitting in practice to the same degree that the Bayesian method can when the available data size is small. Furthermore, we compare robustness of our method against hyperparameters to that of the Bayesian method. For this purpose, we use UCI repository data [13] and the classification accuracy of Bayesian network classifiers (BNCs) under a pre-trained network structure to evaluate the parameter estimation accuracy because the classification accuracy depends on parameter estimation accuracy in such situations.

### 4.1. Bayesian network classifiers

In fact, BNCs are restricted models of BNs, which are often used in classification tasks. The BNCs have one class variable and other variables, and predict the class label with given evidence to the other attributes. Friedman et al. observed that, in many benchmark datasets, unrestricted BNs underperform Naive Bayes classifiers (NBs), which are strongly restricted BNs [5]. We use Naive Bayes type models for evaluation of our method. For this work, we adopted generally augmented Naive Bayes classifiers (GANs), which have an unrestricted network in attributes, except for class variables [2, 5]. This type of network is useful because it reportedly achieves the highest accuracies in many datasets among NBs, BNs, and GANs [2]. This type of network is shown in Fig. 2 as an example, where $X_c$ denotes a class node and $X_i$ $(i = 1, 2, \ldots, 10)$ are other attributes. In addition, an unrestricted structure is introduced among the attributes ($\{X_1, X_2, \ldots, X_{10}\}$).

In this paper, to execute structural learning of GANs, we used the PC algorithm [18], which is a representative algorithm of constraint-based search method for structure learning of BNs, and which usually uses $\chi^2$-tests or mutual information tests for identifying conditional independence relationships among variables. We modified the PC algorithm for application to BNCs according to the methodologies of Cheng and Greiner [2]: (1) replacing every mutual information test between attributes $X, Y \in \boldsymbol{X}$: $I(X;Y)$ with a conditional mutual information test $I(X;Y|X_c)$; (2) replacing every conditional mutual information test $I(X;Y|\boldsymbol{Z})$ with $I(X;Y|\boldsymbol{Z}, X_c)$, where $\boldsymbol{Z} \subset \boldsymbol{X} \setminus X_c$; (3) adding the class node $X_c$ as a parent of every other attribute, where the conditional independence of $X$ and $Y$ given subsets $\boldsymbol{Z}$ is measured using conditional mutual information in our experiments, as the following.

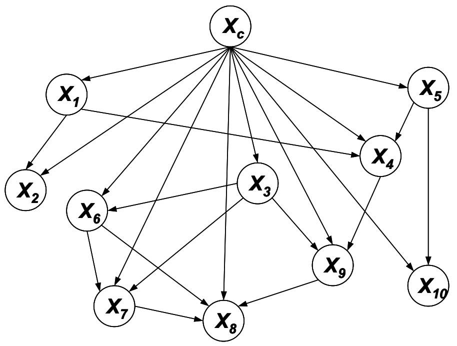

**Figure 2. Example of a generally augmented Naive Bayes classifier (GAN).**

$$I(X;Y|\boldsymbol{Z}) = \sum_{x,y,\boldsymbol{z}} P(x,y,\boldsymbol{z}) \log \frac{P(x,y \,|\, \boldsymbol{z})}{P(x \,|\, \boldsymbol{z}) P(y \,|\, \boldsymbol{z})} \tag{20}$$

Actually, $X$ and $Y$ are conditional independent given the condition set $\boldsymbol{Z}$ if $I(X;Y|\boldsymbol{Z})$ is smaller than a certain threshold value $\epsilon > 0$. The threshold values are decided respectively for each dataset by limiting each network to a maximum of five parents per variable. In addition, the graphs produced by the PC algorithm are partially directed acyclic graphs (PDAGs). Therefore, we oriented the undirected edges to avoid cyclic graphs and to reduce the number of parameters.

### 4.2. Evaluation using UCI data

We selected seven datasets from the UCI machine learning repository. These datasets were formatted and discretized by Greiner[1] using the Fayyad and Irani method [4] except for the Car and Nursery datasets, which were prediscretized. In choosing the datasets, we selected datasets with numerous cases to train the structure of GANs as precisely as possible because the effect of parameter estimation is expected to be separated from that of the correctness of structural inference. It can occur that the structures

[1] available from his site:
http://www.cs.ualberta.ca/ greiner/ELR/

**Table 1. Description of datasets used for the experiments.**

| Dataset | | Attributes | Classes | Instances | |
|---|---|---|---|---|---|
| | | | | Train | Test |
| 1 | Car | 6 | 4 | 1000 | 500 |
| 2 | Chess | 36 | 2 | 2130 | 1066 |
| 3 | Letter | 16 | 26 | 15000 | 5000 |
| 4 | Nursery | 8 | 5 | 8640 | 4320 |
| 5 | Satimage | 36 | 6 | 4435 | 2000 |
| 6 | Segment | 19 | 7 | 1540 | 770 |
| 7 | Shuttle-small | 9 | 7 | 3866 | 1934 |

are designed by prior knowledge but the parameters are not pre-estimated in real applications such as user modeling. A brief description of the datasets is presented in Table 1 [2].

First, structure learning was conducted using the PC algorithm. Next, parameter learning was conducted using the ML with eq. (2), Bayesian with eq. (3), and MFE-EB with eq. (16), eq. (19) and eq. (17) or eq. (18) processes. In contrast to structure learning, we used small samples that had been selected randomly from each dataset. Those sample sizes are selected to present large deviations of accuracy in some Bayesian hyperparameters from that in ML under the limit sample size 100 (Letter, 1000; Chess/Nursery/Satimage, 250; Car/Segment/Shuttle-small, 100). The Dirichlet hyperparameters are the same, $\alpha_{ijk} = \alpha$, because of the lack of prior knowledge about datasets. We compared their accuracy to that of ML to confirm the effectiveness of the Bayes and the MFE-EB. For avoiding zero probability, which generates contradiction when testing data have evidence that did not emerge in training data, we added a small positive number (0.0001) to all conditional frequencies.

We adopted $\alpha = 0.5, 1, 10$ because $\alpha = 1$ is the famous Laplace method, and $\alpha = 0.5$ and $\alpha = 10$ are recommended from theoretical [3, 19] and practical [21] perspectives. On the other hand, we examined some values for $N_c$ because no knowledge about it exists. We show accuracies of BNCs with parameters estimated using ML, the Bayesian, and maximum values of the MFE-EB ("linear-state" and "log-state"), as shown in Table 2. It is clear that the MFE-EB methods have effects of avoiding overfitting to small data size because of controlling entropy according to the available data size, as well as the Bayesian does. Moreover, regarding comparison with the Bayesian with recommended hyperparameters and with MFE-EB, the latter is superior with respect to maximum values of accuracy and variances. It seems to leverage likelihood and entropy more effectively than that of the Bayesian-Dirichlet method.

Next, we evaluate the robustness of the MFE-EB against various values of the hyperparameter $N_c$. Figures 3 and 4 show that the accuracies against the various values of $N_c$, which are both plotted in differences of accuracies from that by ML. Actually, MFE-EB seems to show good performance in common $N_c \sim 1$ for a linear model, and in common $N_c \sim 2$ for a log model in every dataset. For a more precise description, Table 3 shows the effective range over which the classification accuracy is greater than 95% of the maximum accuracy for each dataset, where "*" signifies minimum or maximum values of hyperparameters in the range of the experiments. In both the linear-state and log-state model, the range in which good performance is shown has some overlap among all datasets. It can be said that the hyperparameter $N_c$ has good common ranges of $(1 \leq N_c \leq 1.5)$ in linear-state MFE-EB, and ranges of $(2 \leq N_c \leq 4)$ in log-state MFE-EB. Therefore, MFE-EB can be said to be a robust estimation method of parameters in BNs. These results are expected to result from the adopted functional form of $\beta$, where $\beta$ approaches 1 rapidly with increasing the number of samples. In addition, the MFE-EB might be expected to have universal ranges of hyperparameters. Moreover, our MFE-EB method seems to be attractive in the sense that there is room for improvement of the function of $\beta$.

## 5. Discussion

The ML, the Bayesian, and the MFE-EB are all called generative methods, although discriminative methods have recently received significant attention in parameter learning of BNCs [7, 8, 11, 16]. Their approaches are aimed at improving the classification accuracies of BNCs given some restricted structures. They are not intended to estimate hidden true probabilistic distributions correctly. However, their studies suggest some insights about both structure and parameter learning of BNCs. Jing et al. found that, when the structure is incorrect, their discriminative methods outperform their generative counterparts [11]. Shen et al. showed that, the better the structures, the smaller the advantage of their discriminative method over the ML in classification tasks [16]. Their results imply that, in discriminative method, the parameters are trained to compensate incompleteness of structure in BNs (BNCs). Therefore, we consider that the hyperparameters in the generative methods can have equivalent effects. Table 2 shows that, for some datasets, the accuracies trained using the Bayesian method seem to be slightly superior to those using the MFE-EB methods. We consider that the results are attributable to the incompleteness of structure in BNCs. In such situations, the Bayes estimators can compensate for the incompleteness because $\alpha$ contributes to the parameters to some degree, even in situations with not a few data, whereas the

[2] The Car dataset was split into 1000 training samples and 500 test samples using random selection.

**Table 2. Accuracies [%] of each method.**

| Dataset | ML | Bayes ($\alpha = 0.5$) | Bayes ($\alpha = 1$) | Bayes ($\alpha = 10$) | MFE-EB (lin) | MFE-EB (log) |
|---|---|---|---|---|---|---|
| Car | 65.8 | 74.4 | 73.8 | 70.2 | 74.0 | 74.0 |
| Chess | 86.0 | 88.6 | 88.9 | 86.1 | 86.3 | 86.0 |
| Letter | 52.4 | 62.2 | 60.6 | 52.0 | 59.8 | 60.6 |
| Nursery | 66.2 | 75.3 | 76.5 | 77.5 | 76.7 | 76.7 |
| Satimage | 67.1 | 64.2 | 60.5 | 52.7 | 76.6 | 76.8 |
| Segment | 81.4 | 80.2 | 79.2 | 69.7 | 82.4 | 82.4 |
| Shuttle-small | 85.8 | 97.9 | 97.9 | 86.6 | 97.2 | 97.4 |
| Ave.$\pm$ $\sigma$ | 72.1$\pm$11.7 | 77.6$\pm$11.8 | 76.8$\pm$12.7 | 70.7$\pm$13.2 | 79.0$\pm$10.7 | 79.1$\pm$10.5 |

**Table 3. Effective ranges of hyperparameters.**

| Dataset | Linear-MFE ($N_c$) | | Log-MFE ($N_c$) | |
|---|---|---|---|---|
| | min | max | min | max |
| Car | 1.0 | 10* | 2.0 | 10* |
| Chess | 0.1* | 10* | 0.5* | 10* |
| Letter | 0.1* | 10* | 0.5* | 10* |
| Nursery | 1.0 | 10* | 2.0 | 10* |
| Satimage | 0.5 | 1.5 | 2.0 | 5.0 |
| Segment | 0.1* | 10* | 0.5* | 10* |
| Shuttle-small | 0.1* | 2.0 | 0.5* | 4.0 |

MFE-EB estimators cannot compensate because they are closer to ML estimators than the Bayes because of the functional form of $\beta$. Therefore, the MFE-EB method might be less effective in multivariate systems that are not properly inferred for their structures between variables, although the method can be extended theoretically to the case of discrete joint probability estimations.

It is worthwhile to discuss the possibility of improving the MFE-EB method. The change of accuracy over the hyperparameters presents similar behaviors by the two models of $\gamma$. For example, in the Nursery dataset, values of $N_c$ in both models, for which the accuracies are high, are larger than those in the other datasets. On the other hand, in the Satimage dataset, both are smaller than those in the other datasets. These results imply that the optimal ranges of values in hyperparameters depend on the dataset properties. Therefore, we consider that it is possible to improve MFE-EB by incorporating those properties of each dataset into the function of $\beta$.

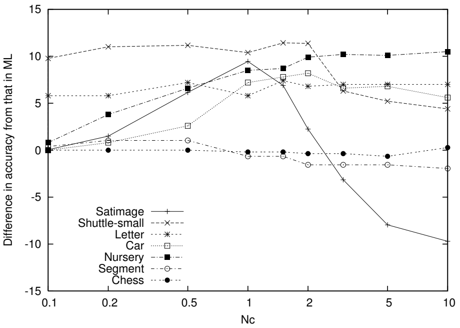

**Figure 3. MFE-EB (“linear-state”) estimation: differences in accuracy from ML [%].**

## 6. Conclusions

We explore a new parameter learning method of BNs. The method is robust against hyperparameter setting. Then we propose an alternative one based on the principle of minimum free energy (MFE), which is well known in statistical thermal physics. Our main conceptual contribution is “data temperature”, which is generated by combining thermal fluctuation with probabilistic fluctuation. Our explicit model of the “data temperature” is assumed to have monotonic functions according to the available data size. The approach enables treatment of the two major principles of maximum likelihood and maximum entropy in a unified manner in the MFE principle with varying data size. In other words, this approach is based on optimizing the contribution of entropy according to available data size, instead of maximizing likelihood or maximizing expected values of posterior probability distributions.

We showed that our method is robust in classification accuracy for choice of hyperparameters. Furthermore, it is superior to the Bayesian method with recommended Dirichlet hyperparameters, although our explicit model of temperature is not sophisticated. Our method provides an effective

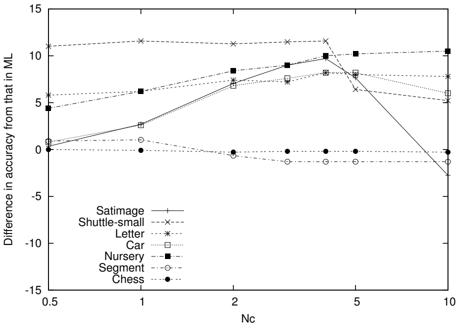

**Figure 4. MFE-EB ("log-state") estimation: differences in accuracy from ML [%].**

tool for use as a parameter estimation method, especially for small data size or sparse data.

In future work, it is necessary that this method be extended to treat prior knowledge. Moreover, it might be interesting to search for equivalent prior distributions generating our estimator, and to extend our approach to various parameter estimation methods.

## Acknowledgements

We thank T. Yamasaki and T. Yoshioka of Fuji Xerox Co., Ltd. for giving us the opportunity to conduct this study. One author (T.I.) particularly thanks H. Okamoto of Fuji Xerox Co., Ltd. for helpful comments and encouragement.

## References

[1] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85:549–559, 1998.

[2] J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 101–108, 1999.

[3] B. Clarke and A. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[4] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI-93)*, pages 1022–1027, 1993.

[5] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[6] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[7] R. Greiner and W. Zhou. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. In *National Conference on Artificial Intelligence (AAAI-02)*, pages 167–173, 2002.

[8] D. Grossman and P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proc. of International Conference on Machine Learning (ICML-04)*, pages 361–368, 2004.

[9] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995, revised June 1996.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 289–296, 1999.

[11] Y. Jing, V. Pavlović, and J. M. Rehg. Efficient discriminative learning Bayesian network classifier via boosted augmented naive Bayes. In *Proc. of International Conference on Machine Learning (ICML-05)*, pages 369–376, 2005.

[12] Y. LeCun and F. J. Huang. Loss functions for discriminative training of energy-based models. In *Proc. of International Workshop on Artificial Intelligence and Statistics (AISTATS-05)*, pages 206–213, 2005.

[13] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[15] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proc. of Annual Meeting on Association for Computational Linguistics (ACL-93)*, pages 183–190, 1993.

[16] B. Shen, X. Su, R. Greiner, P. Musilek, and C. Cheng. Discriminative parameter learning of general Bayesian network classifiers. In *Proc. of IEEE International Conference on Tools with Artificial Intelligence (ICTAI-03)*, pages 296–305, 2003.

[17] D. J. Spiegelhalter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

[18] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, Cambridge, MA, 2000.

[19] J. Suzuki. Learning Bayesian belief networks based on the minimum description length principle: An efficient algorithm using the B&B technique. In *Proc. of International Conference on Machine Learning (ICML-96)*, pages 462–470, 1996.

[20] N. Ueda and R. Nakano. Deterministic annealing variant of the EM algorithm. In *Advances in Neural Information Processing Systems 7 (NIPS 7)*, pages 545–552, 1995.

[21] S. Yang and K. Chang. Comparison of score metrics for Bayesian network learning. *IEEE Trans. on Systems, Man and Cybernetics Part A: Systems and Humans*, 32(3):419–428, 2002.

[22] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Information Theory*, 51(7):2282–2312, 2005.