entropy

MDPI

Article

Hidden Node Detection between Observable Nodes Based on Bayesian Clustering

Keisuke Yamazaki * and Yoichi Motomura

AI Research Center, National Institute of Advanced Industrial Science Technology, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan; y.motomura@aist.go.jp

* Correspondence: k.yamazaki@aist.go.jp; Tel.: +81-3-3599-8750

Received: 15 November 2018; Accepted: 3 January 2019; Published: 7 January 2019

Abstract: Structure learning is one of the main concerns in studies of Bayesian networks. In the present paper, we consider networks consisting of both observable and hidden nodes, and propose a method to investigate the existence of a hidden node between observable nodes, where all nodes are discrete. This corresponds to the model selection problem between the networks with and without the middle hidden node. When the network includes a hidden node, it has been known that there are singularities in the parameter space, and the Fisher information matrix is not positive definite. Then, the many conventional criteria for structure learning based on the Laplace approximation do not work. The proposed method is based on Bayesian clustering, and its asymptotic property justifies the result; the redundant labels are eliminated and the simplest structure is detected even if there are singularities.

Keywords: Bayesian clustering; structure learning in singular cases; model selection

1. Introduction

In learning Bayesian networks, one of the main concerns is structure learning. Many criteria to detect the network structure have been proposed such as the minimum description length (MDL) [1], the Bayesian information criterion (BIC) [2], the Akaike information criterion (AIC) [3], and the marginal likelihood [4]. Most of these criteria assume statistical regularity, which means that the network has identifiability on the parameter and then the nodes are observable.

The nodes of the network are not always observable in practical situations; there will be some underlying factors, which are difficult to observe and do not appear in the given data. In such cases, the criteria for the structure learning must be designed by taking account of the existence of the hidden nodes. However, the statistical regularity does not hold when the network contains hidden nodes [5,6].

The probabilistic models fall into two types: Regular and singular. If the parameter and the probability function expressed by the parameter have one-to-one mapping, the model has statistical regularity and is referred to as regular. Otherwise, there are singularities in the parameter space and the model is referred to as singular. Due to the singularities, the Fisher information matrix is not positive definite, which means that the conventional analysis based on the Laplace approximation or the asymptotic normality does not work in the singular models. Many probabilistic models such as mixture models, hidden Markov models, and neural networks are singular. To cope with the problem of the singularities, an analysis method based on algebraic geometry has been proposed [7], and asymptotic properties of the generalization performance and of the marginal likelihood have been investigated in mixture models [8], hidden Markov models [9], neural networks [7,10], etc.

It is known that the Bayesian network with hidden nodes is singular since the parametrization will change compared with the network without hidden nodes. Even in the simple structure such as the naive Bayesian network, the parameter space has singularities [5,11]. A method to select the optimal structure from some candidate networks has been proposed by using the algebraic geometrical method [5]. For general singular models, new criteria are developed; a widely applicable information criterion (WAIC) is based on the asymptotic form of the generalization error and a widely applicable Bayesian information criterion (WBIC) is derived from the asymptotic form of the marginal likelihood. BIC is also extended to the singular models [12].

The structure learning of the Bayesian network with hidden nodes is a very widely studied problem. Observable constraints from the Bayesian network with hidden nodes is considered in [13]. A model based on observable conditional independence constraints is proposed by [14]. For causal discovery, the related fast causal inference (FCI) algorithm has been developed, e.g., [15]. In the present paper, we consider a two-step method; the first step obtains the optimal structure with observable nodes and the second step detects the hidden nodes in each partial graph. Figure 1 shows the hidden-node detection. The left side of the figure describes the optimal structure with observable nodes only, based on some method of the structure learning. Then, as the second step, we focus on the connections between the observable nodes shown in the right side of the figure. In this example, the parent node x_4 has the domain $\{1,\dots,5\}$ and the child node x_6 has the one $\{1,\dots,4\}$. If the value of the child node is determined by only three factors, the middle node Z, which has the domain $\{1,2,3\}$, simplifies the conditional probability tables (CPTs). It has been known that the smaller dimension the parameter of the network is, the more accurate the parameter learning is. So, it is practically useful to find the simplest expression of the CPTs.

The issue comes down to detection of a hidden node between observable nodes. We compare two network structures, which are shown in Figure 2.

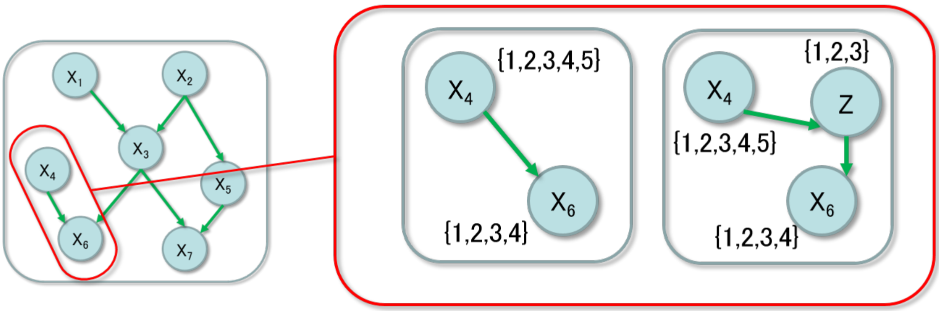


Figure 1. The two-step method: structure learning with observable nodes and hidden-node detection.

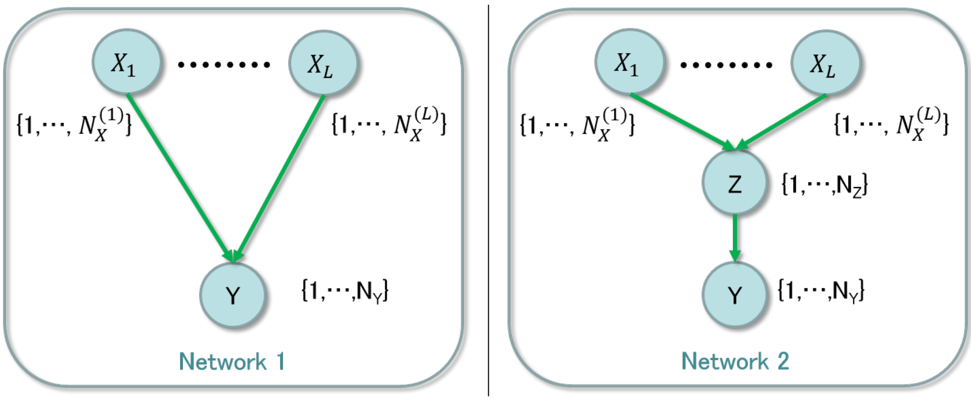


Figure 2. Two networks with and without a hidden node.

The left and the right panels are networks without and with the hidden node, respectively, where $X = (x_1,\dots,x_L)$ with the domain $x_l \in \{1,\dots,N_X^{(l)}\}$ and $Y \in \{1,\dots,N_Y\}$ are observable and $Z \in \{1,\dots,N_Z\}$ is hidden. Since the evidence data on X and Y are given and there is no information on Z, we need to consider whether the hidden node exists and its range N_Z. We propose a method to examine whether the middle hidden node should exist or not using Bayesian clustering. In order to obtain the simplest structure, there is a way to use the regularization technique [16], while it is not straightforward to prove the selected structure is theoretically optimal. Our method is justified based on a property of the entropy term in the asymptotic form of the marginal likelihood, which plays an essential role in the clustering. The result of clustering shows necessary labels to express the relation between the observable nodes X and Y. Counting the number of the used labels, we can determine the existence of the hidden node. Note that we do not consider whole possible structures of the network to reduce the computational complexity; in the present paper, we try to optimize the network from the limited structures, where for example there is no multiple inserted hidden nodes or connections between hidden nodes.

The remainder of this paper is organized as follows. Section 2 presents a formal definition of the network. Section 3 summarizes Bayesian clustering. Section 4 proposes the method to select the structure based on Bayesian clustering and derives its asymptotic behavior. Section 5 shows results of the numerical experiments validating the behavior. Finally, we present a discussion and our conclusions in Sections 6 and 7, respectively.

2. Model Settings

In this section, the network structure and its parameterization are formalized. The naive structure has been applied to classification and clustering tasks and its mathematical properties are studied [5] since it is expressed as a mixture model. As mentioned in the previous section, we consider the hidden node with both parent and child observable nodes. One of the simplest networks is shown in the right panel of Figure 2. Let the probabilities of $X = (X_1, \dots, X_L)$, Z, and Y be defined by

$$p(X_l = i^{(l)}) = a^{(l)}_{i^{(l)}}, \tag{1}$$

$$p(Z = j|X = i) = b_{ij}, \tag{2}$$

$$p(Y = k|Z = j) = c_{jk} \tag{3}$$

for $i \in I = \{(i^{(1)}, \dots, i^{(L)})\}$, $i^{(l)} \in \{1, \dots, N^{(l)}_X\}$, $j = 1, \dots, N_Z$, and $k = 1, \dots, N_Y$. Since they are probabilities, we assume that

$$a^{(l)}_i \geq 0,\ a^{(l)}_1 = 1 - \sum_{i=2}^{N^{(l)}_X} a^{(l)}_i, \tag{4}$$

$$b_{ij} \geq 0,\ b_{i1} = 1 - \sum_{j=2}^{N_Z} b_{ij}, \tag{5}$$

$$c_{jk} \geq 0,\ c_{j1} = 1 - \sum_{k=2}^{N_Y} c_{ij}. \tag{6}$$

It is easy to find that b_{ij} is the element of the CPT for Z and c_{jk} is that for Y. Let w be the parameter consisting of $a^{(l)}_i, b_{ij}, c_{jk}$, where the dimension is

$$\dim w = \sum_{l=1}^{L} (N^{(l)}_X - 1) + (N_Z - 1) \prod_{l=1}^{L} N^{(l)}_X + N_Z(N_Y - 1). \tag{7}$$

We also define the probabilities of the network shown in the left panel of Figure 2;

$$p(X^{(l)} = i^{(l)}) = d^{(l)}_{i^{(l)}}, \tag{8}$$

$$p(Y = j|X = i) = e_{ij}. \tag{9}$$

The parameter u consisting of d_i and e_{ij} has the dimension

$$\dim u = \sum_{l=1}^{L}(N_X^{(l)} - 1) + (N_Y - 1)\prod_{l=1}^{L} N_X^{(l)}. \tag{10}$$

If the relation between X and Y can be simplified, the degree of freedom $\dim u$ is not necessary and is reduced to $\dim w$ such as the case shown in Figure 1. This is similar to the dimension reduction of data with sandglass type neural networks or the non-negative matrix factorization, which have a smaller number of nodes in the middle layers than the one in the input and output layers. The relation between the necessary dimension of the parameter and the probability of the output is not always trivial [17]. The present paper focuses on the sufficient case in terms of the dimension reduction, where $\dim w < \dim u$ rewritten as

$$N_Y \prod_{l=1}^{L} N_X^{(l)} > N_Z \left(N_Y - 1 + \prod_{l=1}^{L} N_X^{(l)} \right). \tag{11}$$

Recall that X and Y are observable and Z is hidden, where N_X and N_Y are given and N_Z is unknown. When the minimum N_Z is detected from the given evidence pairs of X and Y, and is satisfied Equation (11), the network structure with the hidden node expresses the pairs with smaller dimension of the parameter. We use Bayesian clustering technique to detect the minimum N_Z.

3. Bayesian Clustering

In this section, let us formally introduce Bayesian clustering. Let the evidence be described by (x_i, y_i) and there are n pairs, which are denoted by $(X^n, Y^n) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Recall that $x_i = (x_i^{(1)}, \ldots, x_i^{(L)})$. The corresponding value of the hidden node is z_i and the set of n data is denoted by Z^n. We can estimate z_i based on the probability $p(Z^n|X^n, Y^n)$. In Bayesian clustering, it is defined by

$$p(Z^n|X^n, Y^n) = \frac{p(X^n, Z^n, Y^n)}{p(X^n, Y^n)}, \tag{12}$$

$$p(X^n, Z^n, Y^n) = \int \prod_{i=1}^{n} p(x_i, z_i, y_i|w)\varphi(w|\alpha)dw, \tag{13}$$

$$p(X^n, Y^n) = \sum_{Z^n} p(X^n, Z^n, Y^n), \tag{14}$$

where $\varphi(w|\alpha)$ is a prior distribution and α is the hyperparameter.

In the network with the hidden node,

$$p(x_i, z_i, y_i|w) = \prod_{l=1}^{L} \{a^{(l)}_{x_i^{(l)}}\} b_{x_i z_i} c_{z_i y_i}. \tag{15}$$

If the prior distribution is expressed as the Dirichlet distribution for $a^{(l)}_{i^{(l)}}$, b_{ij}, and c_{jk}, the numerator $p(X^n, Z^n, Y^n)$ is analytically computable. Based on the relation $p(Z^n|X^n, Y^n) \propto p(X^n, Z^n, Y^n)$, the Markov Chain Monte Carlo (MCMC) method provides the sampling of Z^n from $p(Z^n|X^n, Y^n)$. This is a common method to estimate hidden variables in machine learning; the underlying topics are estimated based on the Gibbs sampler in topic models such as the latent Dirichlet allocation [18].

4. Hidden Node Detection

In this section, the algorithm to detect the hidden node is introduced and its asymptotic property reducing the number of the used labels is revealed.

4.1. The Proposed Algorithm

When the size of the middle node is large such as

$$\prod_{l=1}^{L} N_X^{(l)} < N_Z, \tag{16}$$

there is no reason to have the node Z; the middle node should reduce the degree of freedom from X. If only $N_Z = 1$ satisfies Equation (11), the middle node is not necessary. Note that $N_Z = 1$ shows that there is no edge between X and Y, which is already excluded in structure learning.

Example 1. *When $L = 1$, $N_X^{(1)} = 3$ and $N_Y = 3$, only $N_Z = 1$ satisfies Equation (11), which shows that there is no hidden node between X and Y.*

The present paper proposes the following algorithm to determine the existence of Z;

Algorithm 2. *Assume that there is $N_Z > 1$ for given $N_X^{(l)}$ and N_Y, that is Equation (11) is satisfied. Apply the Bayesian clustering method to the given evidence (X^n, Y^n) and estimate Z^n based on the MCMC sampling. Let the number of used labels be denoted by $\hat{N}_Z$. If the following inequality holds, the hidden node $Z \in \{1, \ldots, \hat{N}_Z\}$ reduces the parameter,*

$$1 < \hat{N}_Z < \frac{N_Y \prod_{l=1}^{L} N_X^{(l)}}{N_Y - 1 + \prod_{l=1}^{L} N_X^{(l)}}. \tag{17}$$

4.2. Asymptotic Properties of the Algorithm

The MCMC method in Bayesian clustering is based on the probability $p(X^n, Z^n, Y^n)$ as shown in Section 3. Since the proposed method depends on this clustering method, let us consider the properties of $p(X^n, Z^n, Y^n)$. The negative logarithm of the probability is expressed as follows:

$$\begin{aligned}
F_\alpha(X^n, Z^n, Y^n) &= -\ln p(X^n, Z^n, Y^n) \\
&= -\ln \int \prod_{i=1}^{n} p(x_i, z_i, y_i|w) \varphi(w|\alpha) dw \\
&= \sum_{l=1}^{L} \left\{ \ln \Gamma(n + N_X \alpha_a) - \sum_{i=1}^{N_X} \ln \Gamma(n_i + \alpha_a) \right\} \\
&+ \sum_{i \in I} \left\{ \ln \Gamma\left(\sum_{j=1}^{N_Z} n_{ij} + N_Z \alpha_b \right) - \sum_{j=1}^{N_Z} \ln \Gamma(n_{ij} + \alpha_b) \right\} \\
&+ \sum_{j=1}^{N_Z} \left\{ \ln \Gamma\left(\sum_{k=1}^{N_Y} m_{jk} + N_Y \alpha_c \right) - \sum_{k=1}^{N_Z} \ln \Gamma(m_{jk} + \alpha_c) \right\} \\
&+ \sum_{l=1}^{L} \left\{ N_X^{(l)} \ln \Gamma(\alpha_a) - \ln \Gamma(N_X^{(l)} \alpha_a) \right\} \\
&+ \left\{ \prod_{l=1}^{L} N_X^{(l)} \right\} \left\{ N_Z \ln \Gamma(\alpha_b) - N_X \ln \Gamma(N_Z \alpha_b) \right\} \\
&+ N_Z \left\{ N_Y \ln \Gamma(\alpha_c) - \ln \Gamma(N_Y \alpha_c) \right\},
\end{aligned} \tag{18}$$

where n_i, n_{ij}, and m_{jk} are given as

$$n_i = \sum_{j=1}^{n} \prod_{l=1}^{L} \delta_{x_j^{(l)}, i^{(l)}}, \tag{19}$$

$$n_{ij} = \sum_{k=1}^{n} \delta_{z_k, j} \prod_{l=1}^{L} \delta_{x_k^{(l)}, i^{(l)}}, \tag{20}$$

$$m_{jk} = \sum_{l=1}^{n} \delta_{z_l,j}\delta_{y_l,k}, \tag{21}$$

respectively, and the prior distribution $\varphi(w|\alpha)$ consists of the Dirichlet distributions;

$$\varphi(w) = \prod_{l=1}^{L} \mathrm{Dir}(a^{(l)}|\alpha_a) \prod_{i\in I} \mathrm{Dir}(b_i|\alpha_b) \prod_{j=1}^{N_Z} \mathrm{Dir}(c_j|\alpha_c), \tag{22}$$

$$\mathrm{Dir}(a^{(l)}|\alpha_a) = \frac{\Gamma(N_X^{(l)}\alpha_a)}{\Gamma(\alpha_a)^{N_X^{(l)}}} \prod_{i=1}^{N_X^{(l)}} a_i^{(l)(\alpha_a-1)}, \tag{23}$$

$$\mathrm{Dir}(b_i|\alpha_b) = \frac{\Gamma(N_Z\alpha_b)}{\Gamma(\alpha_b)^{N_Z}} \prod_{j=1}^{N_Z} b_{ij}^{\alpha_b-1}, \tag{24}$$

$$\mathrm{Dir}(c_j|\alpha_c) = \frac{\Gamma(N_Y\alpha_c)}{\Gamma(\alpha_c)^{N_Y}} \prod_{k=1}^{N_Y} c_{jk}^{\alpha_c-1}. \tag{25}$$

The function δ_{ij} and $\Gamma(\cdot)$ are the Kronecker delta and the gamma function, respectively. The hyperparameter α consists of α_a, α_b, and α_c. The sampling result of Z^n is dominantly taken from the area, which makes $p(X^n, Z^n, Y^n)$ large. Then, we investigate which Z^n minimizes $F_\alpha(X^n, Z^n, Y^n)$ for given (X^n, Y^n).

Theorem 3. *When the number of the given data n is sufficiently large, $F(X^n, Z^n, Y^n)$ is written as*

$$F(X^n, Z^n, Y^n) = -nS + C\ln n + O_p(1), \tag{26}$$

$$\begin{aligned} S &= \textstyle\sum_{l=1}^{L} \sum_{i^{(l)}=1}^{N_X^{(l)}} \frac{n_{i(l)}^{(l)}}{n} \ln \frac{n_{i(l)}^{(l)}}{n} \\ &\quad + \textstyle\sum_{i\in I} \sum_{j=1}^{\tilde{N}_Z} \frac{\sum_{j'=1}^{\tilde{N}_Z} n_{ij'}}{n} \frac{n_{ij}}{\sum_{j'=1}^{\tilde{N}_Z} n_{ij'}} \ln \frac{n_{ij}}{\sum_{j'=1}^{\tilde{N}_Z} n_{ij'}} \\ &\quad + \textstyle\sum_{j=1}^{\tilde{N}_Z} \sum_{k=1}^{N_Y} \frac{m_j}{n} \frac{m_{jk}}{m_j} \ln \frac{m_{jk}}{m_j}, \end{aligned} \tag{27}$$

$$\begin{aligned} C = &\left\{ \textstyle\prod_{l=1}^{L} N_X^{(l)} \right\} (N_Z - \tilde{N}_Z)\alpha_b \\ &+ \tfrac{1}{2}\left\{ \textstyle\sum_{l=1}^{L} (N_X^{(l)} - 1) + \left(\prod_{l=1}^{L} N_X^{(l)} \right)(\tilde{N}_Z - 1) + \tilde{N}_Z(N_Y - 1) \right\}, \end{aligned} \tag{28}$$

$$m_j = \sum_{k=1}^{N_Y} m_{jk}, \tag{29}$$

where $\tilde{N}_Z$ is the number of m_j such that $m_j/n = O(1)$.

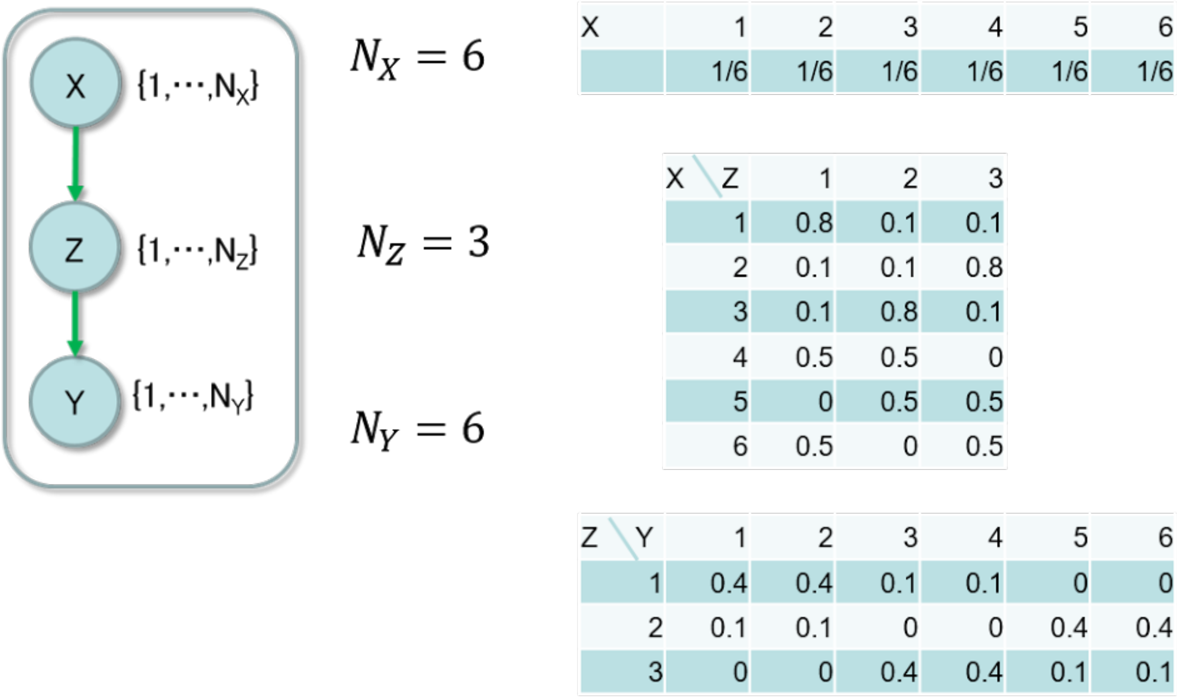
The proof will be shown in Appendix A. The first term $-nS$ is the dominant factor, and its coefficient S is maximized in the clustering. This coefficient determines $\tilde{N}_Z$, which is the number of used labels in the clustering result.

Assume that the true structure with the hidden node has the minimal expression, where the range of Z is $z = 1, \ldots, N_Z^*$, and that the estimated size is larger than the true one; $N_Z^* \leq \tilde{N}_Z$. We can easily confirm that Bayesian clustering chooses the minimum structure $\tilde{N}_Z = N_Z^*$ as follows. The three terms in the coefficient S correspond to the negative entropy functions of the parameter $a_i^{(l)}$, b_{ij}, and c_{jk}, respectively. Then, the minimum $\tilde{N}_Z$ obviously makes the coefficient S maximized since the number of elements of parameter should be minimized for the small entropy. When the hidden node has the redundant state, which means that two values of Z have completely same output distribution of Y, the second term of S is larger than the case of non-redundant situation $\hat{N}_Z = N_Z^*$. Based on the assumption that the true structure is minimal, the estimation therefore gets the minimum structure, $\tilde{N}_Z = N_Z^*$.

According to this property, the number of used label $\hat{N}_Z$ asymptotically goes to N_Z^*. The proposed algorithm compares the essential number of the values of Z and will be a criterion to select the proper structure when n is large. This property exists only in Bayesian clustering so far; the eliminating effect of the redundant labels has not been found in other method of the clustering such as the maximum-likelihood clustering based on the expectation-maximization algorithm.

5. Numerical Experiments

In this section, we validate the asymptotic property in numerical experiments. We set the data-generating model shown in Figure 3 and prepared ten evidence data sets.



X	1	2	3	4	5	6
	1/6	1/6	1/6	1/6	1/6	1/6

X \ Z	1	2	3
1	0.8	0.1	0.1
2	0.1	0.1	0.8
3	0.1	0.8	0.1
4	0.5	0.5	0
5	0	0.5	0.5
6	0.5	0	0.5

Z \ Y	1	2	3	4	5	6
1	0.4	0.4	0.1	0.1	0	0
2	0.1	0.1	0	0	0.4	0.4
3	0	0	0.4	0.4	0.1	0.1

Figure 3. The data-generating model.

There was a single parent node $L = 1$. The sizes of the nodes were $N_X^{(1)} = 6$, $N_Y = 6$ and $N_Z^* = 3$. The CPTs are described on the right-side of the figure, where the true parameter consists of these probabilities. There were 2000 pairs of (x, y) in each data set. Since the following condition is satisfied,

$$\frac{N_Y \prod_{l=1}^{L} N_X^{(l)}}{N_Y - 1 + \prod_{l=1}^{L} N_X^{(l)}} = \frac{6 \times 6}{6 - 1 + 6} = \frac{36}{11} > 3 = N_Z^*, \tag{30}$$

the structure of the data-generating model with the hidden node had smaller dimension of the parameter than the one without a hidden node.

We applied Bayesian clustering to each data set, where the model had the size of the hidden node $N_Z = 6$. According to the asymptotic property in Theorem 3, the MCMC method should take label assignment from the area, where the number of the used labels was reduced to three. The estimated model size was determined by the assignment, which minimized the function $F_\alpha(X^n, Z^n, Y^n)$. Since the sampling of the MCMC method depended on the initial assignment, we conducted ten trials for each data set and regarded the estimated size as the minimum one. The number of iterations in the MCMC method was 1000.

Table 1 shows the results of the experiments.

Table 1. The results of the estimated size.

Data-Set ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Estimated size	3	3	3	4	3	3	4	4	4	3

In all data sets, the size of the hidden node Z is reduced and the correct size is estimated in more than half sets, we confirmed the effect eliminating the redundant labels. Since the result of the MCMC method depends on the given data, the minimum size is not always found; the estimated size is four in some data sets instead of three. Even in such case, however, we could estimate the correct size after setting the initial size of the model as $N_Z = 4$. Repeating this procedure, we will be able to avoid the local optimal size and find the global one.

Figure 4 shows this estimation procedure in the practical cases. The initial model size starts from six. The left panel is the case, where the proper size is directly found and the estimated size does not change at size four. The right panel is the case, where the estimated size is first four and then the next result is three, which is the fixed point.

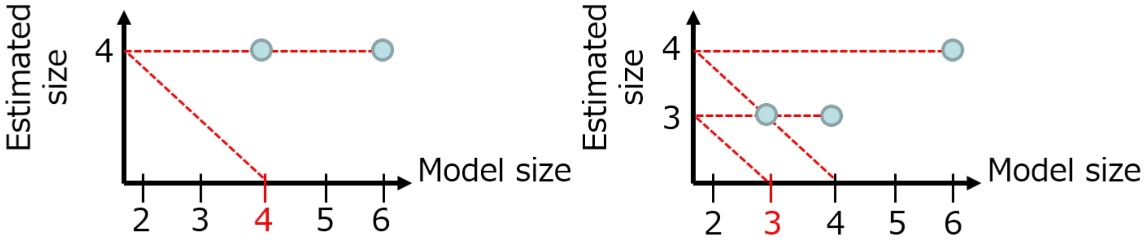
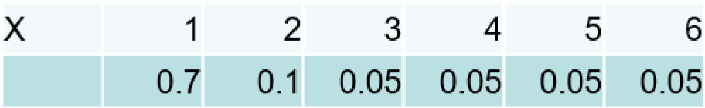


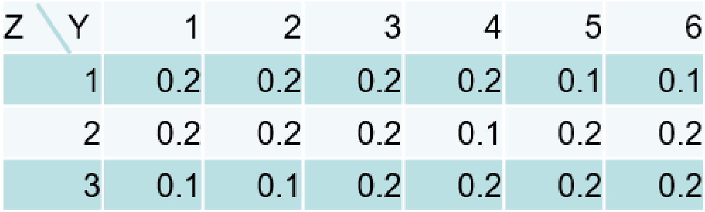
Figure 4. The estimation procedure in practical cases.

To investigate the properties of the estimated size, we tried some different numbers of pairs $n = 100, 500$ and a skewed distribution of the parent node (Figure 5), and nearly uniform distribution of the child node (Figure 6).



X	1	2	3	4	5	6
	0.7	0.1	0.05	0.05	0.05	0.05

Figure 5. The skewed distribution of the parent node X.



Z \ Y	1	2	3	4	5	6
1	0.2	0.2	0.2	0.2	0.1	0.1
2	0.2	0.2	0.2	0.1	0.2	0.2
3	0.1	0.1	0.2	0.2	0.2	0.2

Figure 6. The nearly uniform distribution of the parent node Y.

Table 2 shows the results of $n = 100, 500$.

Table 2. The results of the estimated size in $n = 100, 500$.

Data-Set ID	1	2	3	4	5	6	7	8	9	10
Estimated size (n = 100)	3	3	3	3	4	3	4	3	3	3
Estimated size (n = 500)	3	3	3	3	3	3	3	3	3	4

Since these CPTs of X, Z, Y are a straightforward case to distinguish the role of the hidden node, the smaller number of the pairs does not adversely affect the estimation. Table 3 shows the results of the different CPTs in the parent and the child nodes.

Table 3. The results of the estimated size in the different conditional probability tables (CPTs).

Data-Set ID	1	2	3	4	5	6	7	8	9	10
Estimated size (skewed parent node)	3	3	3	3	3	4	3	3	3	3
Estimated size (nearly-uniform child node)	1	1	1	1	2	1	1	2	1	1

The number of pairs was $n = 100$. Due to the CPT of Z, the skewed distribution of the parent node still keeps the sufficient variation of Z to estimate the size N_Z, which provides the same accuracy as the uniform distribution. On the other hand, the nearly uniform distribution of the child node makes the estimation difficult because each value of Z has the similar output distribution. The Dirichlet prior of Z has a strong effect to eliminate the redundancy, which means the estimated sizes tend to be smaller than the true one.

6. Discussion

In this section, we discuss the difference between the proposed method and other conventional criteria for the model selection. In the proposed method, the label assignment Z^n is obtained from the MCMC method, which takes the samples according to $p(X^n, Z^n, Y^n)$. The probability $p(X^n, Z^n, Y^n)$ is the marginal likelihood on the complete data (X^n, Z^n, Y^n); recall the definition,

$$p(X^n, Z^n, Y^n) = \int \prod_{i=1}^{n} p(x_i, z_i, y_i | w) \varphi(w|\alpha) dw. \tag{31}$$

This looks similar to the criteria based on the marginal likelihood such as BDu(e) [19,20] and its asymptotic form such as BIC [2], MDL [1]. Since it is assumed that the network has the statistical regularity or the nodes are all observable, many criteria do not work on the network with hidden nodes.

WBIC is proposed for the singular models. The main difference is that it is based on the marginal likelihood of the incomplete data X^n, Y^n;

$$\begin{aligned} p(X^n, Y^n) &= \textstyle\sum_{Z^n} p(X^n, Z^n, Y^n) \\ &= \textstyle\int \prod_{i=1}^{n} \sum_{z_i} p(x_i, z_i, y_i | w) \varphi(w|\alpha) dw. \end{aligned} \tag{32}$$

Due to the marginalization over Z^n, it requires the calculation of values for all candidate structures. For example, assume that we have candidate structures $N_Z = 1, 2, 3$ denoted by $p_1(X^n, Y^n)$, $p_2(X^n, Y^n)$, and $p_3(X^n, Y^n)$, respectively. In WBIC, we calculate all values and select the optimal structure;

$$\hat{N}_Z = \arg \min_{i=1,2,3} p_i(X^n, Y^n). \tag{33}$$

On the other hand, in the proposed method, we calculate the label assignment with the structure $N_Z = 3$ and obtain $\hat{N}_Z$, which shows the necessity of the node Z.

Another difference from the conventional criteria is the dominant order of the objective function, which determines the optimal structure. As shown in Corollary 6.1 of [6], the negative logarithm of the marginal likelihood of the incomplete data has the following asymptotic form;

$$\begin{aligned} F_\alpha(X^n, Y^n) &= -\ln p(X^n, Y^n) \\ &= -nS_{XY} + C_{XY} \ln n + o_p(\ln n), \end{aligned} \tag{34}$$

where the coefficient S_{XY} is the empirical entropy of the observation (X^n, Y^n) and C_{XY} depends on the data-generating distribution, the model, and the prior distribution. This form means that the optimal model is selected by $\ln n$ order term with the coefficient C_{XY}, while it is selected by n order term with the coefficient S of Theorem 3 in the proposed method. Since the largest terms are n order in both $F_\alpha(X^n, Y^n)$ and $F_\alpha(X^n, Z^n, Y^n)$, the proposed method will have stronger effect to distinguish the difference of the structures.

The asymptotic accuracy of Bayesian clustering has been studied [21], which considers the error function between the true distribution of the label assignment and the estimated one measured by the Kullback-Leibler divergence:

$$D(n) = E_{X^n,Y^n}\left[\sum_{Z^n} q(Z^n|X^n, Y^n) \ln \frac{q(Z^n|X^{,}Y^n)}{p(Z^n|X^n, Y^n)}\right], \tag{35}$$

where $E_{X^n,Y^n}[\cdot]$ is the expectation over all evidence data and

$$q(Z^n|X^n, Y^n) = \frac{q(X^n, Z^n, Y^n)}{\sum_{Z^n} q(X^n, Z^n, Y^n)}, \tag{36}$$

$$q(X^n, Z^n, Y^n) = \prod_{i=1}^{n} q(x_i, z_i, y_i). \tag{37}$$

The true network is denoted by $q(x, z, y)$. The proposed method minimizes this error function, which means that the label assignment Z^n is optimized in the sense of the density estimation. Even though the optimized function is not directly for the model selection, due to the asymptotic property of the Bayes clustering simplifying the label use, the proposed method is computationally efficient to determine the existence of the hidden node and the result asymptotically has coincident.

7. Conclusions

In this paper, we have proposed a method to detect a hidden node between observable nodes based on Bayesian clustering. The asymptotic behavior of the clustering has been revealed and it shows that the redundant labels are eliminated and the essential structure will be detected. Evaluation of the proposed method with numerical experiments is one of our future studies.

Author Contributions: Conceptualization, K.Y. and Y.M.; data curation, K.Y.; formal analysis, K.Y.; funding acquisition, Y.M. and K.Y.; investigation, K.Y. and Y.M.; project administration, Y.M.; software, K.Y.; writing, original draft, K.Y.; writing, review and editing, K.Y.

Funding: This research is based on results obtained from projects commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and KAKENHI 15K00299.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this section, we prove Theorem 3. Using the asymptotic relation $\ln \Gamma(x) = x \ln x - \frac{1}{2} \ln x - x + O(1)$ for sufficiently large x, we can obtain that

$$
\begin{aligned}
F_\alpha(X^n, Z^n, Y^n) \quad &= \textstyle\sum_{l=1}^{L} \Big\{ \big(n + N_X^{(l)}\alpha_a\big) \ln \big(n + N_X^{(l)}\alpha_a\big) \\
&- \tfrac{1}{2} \ln \big(n + N_X^{(l)}\alpha_a\big) - \big(n + N_X^{(l)}\alpha_a\big) \Big\} \\
&- \textstyle\sum_{l=1}^{L} \sum_{i^{(l)}=1}^{N_X^{(l)}} \Big\{ (n_{i^{(l)}}^{(l)} + \alpha_a) \ln(n_{i^{(l)}}^{(l)} + \alpha_a) \\
&- \tfrac{1}{2} \ln(n_{i^{(l)}}^{(l)} + \alpha_a) - (n_{i^{(l)}}^{(l)} + \alpha_a) \Big\} \\
&+ \textstyle\sum_{i \in I} \Big\{ \Big(\sum_{j=1}^{N_Z} n_{ij} + N_Z \alpha_b \Big) \ln \Big(\sum_{j=1}^{N_Z} n_{ij} + N_Z \alpha_b \Big) \\
&- \tfrac{1}{2} \ln \Big(\textstyle\sum_{j=1}^{N_Z} n_{ij} + N_Z \alpha_b \Big) - \Big(\sum_{j=1}^{N_Z} n_{ij} + N_Z \alpha_b \Big) \Big\} \qquad \text{(A1)} \\
&- \textstyle\sum_{i \in I} \sum_{j=1}^{N_Z} \Big\{ (n_{ij} + \alpha_b) \ln(n_{ij} + \alpha_b) - \tfrac{1}{2} \ln(n_{ij} + \alpha_b) - (n_{ij} + \alpha_b) \Big\} \\
&+ \textstyle\sum_{j=1}^{N_Z} \Big\{ \Big(\sum_{k=1}^{N_Y} m_{jk} + N_Y \alpha_c \Big) \ln \Big(\sum_{j=1}^{N_Y} m_{jk} + N_Y \alpha_c \Big) \\
&- \tfrac{1}{2} \ln \Big(\textstyle\sum_{k=1}^{N_Y} m_{jk} + N_Y \alpha_c \Big) - \Big(\sum_{k=1}^{N_Y} m_{jk} + N_Y \alpha_c \Big) \Big\} \\
&- \textstyle\sum_{j=1}^{N_Z} \sum_{k=1}^{N_Y} \Big\{ (m_{jk} + \alpha_c) \ln(m_{jk} + \alpha_c) - \tfrac{1}{2} \ln(m_{jk} + \alpha_c) - (m_{jk} + \alpha_c) \Big\} \\
&+ O_p(1).
\end{aligned}
$$

Collecting the constant terms and uniting them to $O_p(1)$, we rewrite $F_\alpha(X^n, Z^n, Y^n)$ as

$$
\begin{aligned}
F_\alpha(X^n, Z^n, Y^n) \quad &= \textstyle\sum_{l=1}^{L} \Big\{ n \ln n + N_X^{(l)} \alpha_a \ln n - \tfrac{1}{2} \ln n - n \Big\} \\
&- \textstyle\sum_{l=1}^{L} \sum_{i^{(l)}=1}^{N_X^{(l)}} \Big\{ n_{i^{(l)}}^{(l)} \ln n_{i^{(l)}}^{(l)} + \alpha_a \ln n_{i^{(l)}}^{(l)} - \tfrac{1}{2} \ln n_{i^{(l)}}^{(l)} - n_{i^{(l)}}^{(l)} \Big\} \\
&+ \textstyle\sum_{i \in I} \Big\{ \Big(\sum_{j=1}^{N_Y} n_{ij} \Big) \ln \sum_{j=1}^{N_Y} n_{ij} + N_Z \alpha_b \ln \sum_{j=1}^{N_Z} n_{ij} - \tfrac{1}{2} \ln \sum_{j=1}^{N_Z} n_{ij} - \sum_{j=1}^{N_Z} n_{ij} \Big\} \\
&- \textstyle\sum_{i \in I} \sum_{j=1}^{N_Z} \Big\{ n_{ij} \ln n_{ij} + \alpha_b \ln n_{ij} - \tfrac{1}{2} \ln n_{ij} - n_{ij} \Big\} \qquad \text{(A2)} \\
&+ \textstyle\sum_{j=1}^{N_Z} \Big\{ \sum_{k=1}^{N_Y} m_{jk} \ln \Big(\sum_{k=1}^{N_Y} m_{jk} \Big) + N_Y \alpha_b \ln \sum_{k=1}^{N_Y} m_{jk} \\
&- \tfrac{1}{2} \ln \textstyle\sum_{k=1}^{N_Y} m_{jk} - \sum_{k=1}^{N_Y} m_{jk} \Big\} \\
&- \textstyle\sum_{j=1}^{N_Z} \sum_{k=1}^{N_Y} \Big\{ m_{jk} \ln m_{jk} + \alpha_c \ln m_{jk} - \tfrac{1}{2} \ln m_{jk} - m_{jk} \Big\} + O_p(1).
\end{aligned}
$$

Using the following relations,

$$\sum_{i^{(l)}=1}^{N_X^{(l)}} n_{i^{(l)}}^{(l)} \ln n_{i^{(l)}}^{(l)} = n \sum_{i^{(l)}=1}^{N_X^{(l)}} \left(\frac{n_{i^{(l)}}^{(l)}}{n} \ln \frac{n_{i^{(l)}}^{(l)}}{n} \right) + n \ln n, \tag{A3}$$

$$\sum_{i \in I} \left(\sum_{j=1}^{N_Z} n_{ij} \right) \ln \sum_{j=1}^{N_Z} n_{ij} = n \sum_{i \in I} \left(\frac{n_i}{n} \ln \frac{n_i}{n} \right) + n \ln n, \tag{A4}$$

$$\sum_{i \in I} \sum_{j=1}^{N_Z} n_{ij} \ln n_{ij} = n \sum_{i \in I} \sum_{j=1}^{N_Z} \frac{n_{ij}}{n} \ln \frac{n_{ij}}{n} + n \ln n, \tag{A5}$$

$$\sum_{j=1}^{N_Z} \left(\sum_{k=1}^{N_Y} m_{jk} \right) \ln \sum_{k=1}^{N_Y} m_{jk} = n \sum_{j=1}^{N_Z} \frac{m_j}{n} \ln \frac{m_j}{n} + n \ln n, \tag{A6}$$

$$\sum_{j=1}^{N_Z} \sum_{k=1}^{N_Y} m_{jk} \ln m_{jk} = n \sum_{j=1}^{N_Z} \sum_{k=1}^{N_Y} \frac{m_{jk}}{n} \ln \frac{m_{jk}}{n} + n \ln n \tag{A7}$$

and focusing on the terms of order n and $\ln n$, we obtain the asymptotic form in the theorem.

References

1. Rissanen, J. Stochastic complexity and modeling. *Ann. Stat.* **1986**, *14*, 1080–1100. [CrossRef]
2. Schwarz, G.E. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
3. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
4. Good, I.J. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods, Research Monograph No. 30*; The MIT Press: Cambridge, MA, USA, 1965.
5. Rusakov, D.; Geiger, D. Asymptotic model selection for naive Bayesian networks. *J. Mach. Learn. Res.* **2005**, *6*, 1–35.
6. Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*; Cambridge University Press: New York, NY, USA, 2009.
7. Watanabe, S. Algebraic analysis for non-identifiable learning machines. *Neural Comput.* **2001**, *13*, 899–933. [CrossRef] [PubMed]
8. Yamazaki, K.; Watanabe, S. Singularities in mixture models and upper bounds of stochastic complexity. *Int. J. Neural Netw.* **2003**, *16*, 1029–1038. [CrossRef]
9. Yamazaki, K.; Watanabe, S. Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing* **2005**, *69*, 62–84. [CrossRef]
10. Aoyagi, M. Consideration on Singularities in Learning Theory and the Learning Coefficient. *Entropy* **2013**, *15*, 3714–3733. [CrossRef]
11. Geiger, D.; Heckerman, D.; Meek, C. Asymptotic Model Selection for Directed Networks with Hidden Variables. In Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence, Portland, OR, USA, 1–4 August 1996; pp. 283–290.
12. Drton, M.; Plummer, M. A Bayesian information criterion for singular models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2017**, *79*, 323–380. [CrossRef]
13. Verma, T.; Pearl, J. Equivalence and Synthesis of Causal Models. In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, USA, 27–29 July 1990; Elsevier Science Inc.: New York, NY, USA, 1991; pp. 255–270.
14. Richardson, T.; Spirtes, P. Ancestral Graph Markov Models. *Ann. Stat.* **2000**, *30*, 2002. [CrossRef]
15. Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **2008**, *172*, 1873–1896. [CrossRef]
16. Chaturvedi, I.; Ragusa, E.; Gastaldo, P.; Zunino, R.; Cambria, E. Bayesian network based extreme learning machine for subjectivity detection. *J. Frankl. Inst.* **2018**, *355*, 1780–1797. [CrossRef]
17. Allman, E.S.; Rhodes, J.A.; Sturmfels, B.; Zwiernik, P. Tensors of nonnegative rank two. *Linear Algebra Appl.* **2015**, *473*, 37–53. [CrossRef]

18. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
19. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **1995**, *20*, 197–243. [CrossRef]
20. Buntine, W. Theory Refinement on Bayesian Networks. In Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, Los Angeles, CA, USA, 13–15 July 1991; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1991; pp. 52–60.
21. Yamazaki, K. Asymptotic accuracy of Bayes estimation for latent variables with redundancy. *Mach. Learn.* **2016**, *102*, 1–28. [CrossRef]