# Hidden Node Detection between Two Observable Nodes Based on Bayesian Clustering

**Keisuke Yamazaki** K.YAMAZAKI@AIST.GO.JP
*AI Research Center*
*National Institute of Advanced Industrial Science Technology*
*Tokyo (Japan)*

**Yoichi Motomura** Y.MOTOMURA@AIST.GO.JP
*AI Research Center*
*National Institute of Advanced Industrial Science Technology*
*Tokyo (Japan)*

## Abstract

The structure learning is one of the main concerns in studies of the Bayesian networks. In the present paper, we consider the network consisting of both observable and hidden nodes, and propose a method to investigate the existence of a hidden node between two observable nodes, which is 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 the structure learning based on the Laplace approximation do not work. The proposed method is based on the 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

## 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 MDL (Rissanen, 1986), BIC (Schwarz, 1978), AIC (Akaike, 1974) and the marginal likelihood (Good, 1965). Most of these criteria assume the statistical regularity, which means that the network has identifiability on the parameter and then the nodes are observable.

The nodes in 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 case, 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 (Rusakov and Geiger, 2005; Watanabe, 2009).

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 the statistical regularity and is referred as regular. Otherwise, there are singularities in the parameter space and the model is 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 (Watanabe, 2001), and asymptotic properties of the generalization performance and of the marginal likelihood have been investigated in mixture models (Yamazaki and Watanabe, 2003), hidden Markov models (Yamazaki and Watanabe, 2005), neural networks (Watanabe, 2001; Aoyagi, 2013), etc.

If there exists a hidden node and its value is not observable, the Bayesian network is singular. Even in the simple structure such as the naive Bayesian network, the parameter space has singularities (Geiger et al., 1996; Rusakov and Geiger, 2005). A method to select the optimal structure from some candidate networks has been proposed by using the algebraic geometrical method (Rusakov and Geiger, 2005). 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 (Drton and Plummer, 2017).

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 Verma and Pearl (1991). A model based on observable conditional independence constraints are proposed by Richardson and Spirtes (2000). For causal discovery, the related FCI algorithm has been developed, e.g. (Zhang, 2008). In the present paper, we consider detection of a hidden node between two observable nodes;

$$X \to Z \to Y, \tag{1}$$

where $X$ and $Y$ are observable and $Z$ is hidden. When the evidence data on $X$ and $Y$ are given and there is no information on $Z$, we need to consider its existence and, if it exists, the range of the variable. This structure learning includes two tasks: the detection of the necessary values of $Z$, and the model selection between Eq. (1) and the network without the hidden node described by

$$X \to Y. \tag{2}$$

We propose a method to examine whether the middle hidden node should be exist or not using the Bayesian clustering. The 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 labels, we can determine the existence of the hidden node based.

The remainder of this paper is organized as follows. Section 2 presents a formal definition of the network. Section 3 summarizes the Bayesian clustering. Section 4 proposes the method to select the structure based on the Bayesian clustering and derives its asymptotic behavior. Finally, we present a discussion and our conclusions in Sections 5 and 6, 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 (Rusakov and Geiger, 2005) 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 structure is shown in Eq.(1). Let the probabilities of $X$, $Z$ and $Y$ be defined by

$$\begin{aligned} p(X=i) =& a_i, \\ p(Z=j|X=i) =& b_{ij}, \\ p(Y=k|Z=j) =& c_{jk} \end{aligned}$$

for $i=1,\dots,N_X$, $j=1,\dots,N_Z$, and $k=1,\dots,N_Y$, respectively. Since they are probabilities, we assume that

$$\begin{aligned} a_i \geq 0, & a_1 = 1 - \sum_{i=2}^{N_X} a_i, \\ b_{ij} \geq 0, & b_{i1} = 1 - \sum_{j=2}^{N_Z} b_{ij}, \\ c_{jk} \geq 0, & c_{j1} = 1 - \sum_{k=2}^{N_Y} c_{ij}. \end{aligned}$$

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

$$\dim w = N_X - 1 + N_X(N_Z - 1) + N_Z(N_Y - 1).$$

We also define the probabilities of the network shown in Eq.(2);

$$\begin{aligned} p(X=i) =& d_i, \\ p(Y=j|X=i) =& e_{ij}. \end{aligned}$$

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

$$\dim u = N_X - 1 + N_X(N_Y - 1).$$

If the relation between $X$ and $Y$ in Eq.(2) can be simplified, the degree of freedom $\dim u$ is not necessary and the network can be expressed as Eq.(1) with small $N_Z$. This is similar to the dimension reduction of data with sandglass type neural networks or the non-negative matrix factorization, which have the 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 sometimes trivial (Allman et al., 2015).

The present paper focuses on the sufficient case in terms of the dimension reduction, where $\dim w < \dim u$;

$$N_X N_Y > N_Z(N_X + N_Y - 1). \tag{3}$$

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 Eq.(3), the network structure Eq.(1) can express the pairs with smaller dimension of the parameter. 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 network structure. We use the Bayesian clustering technique to detect the minimum $N_Z$.

## 3. Bayesian Clustering

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

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

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

In the network Eq.(1),

$$p(x_i, z_i, y_i|w) = a_{x_i} b_{x_i z_i} c_{z_i y_i}.$$

If the prior distribution is expressed as the Dirichlet distribution for $a_i$, $b_{ij}$ 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 $p(Z^n|X^n, Y^n)$.

## 4. Hidden Node Detection

In this section, the algorithm to detect the hidden node is introduced and its asymptotic property is revealed.

### 4.1 The Proposed Algorithm

When the ranges of the observable nodes $X$ and $Y$ are $x = 1, \ldots, N_X$ and $y = 1, \ldots, N_Y$, respectively, there is no reason to have the middle hidden node $Z$ for $N_Z \geq N_X$; the node

$Z$ should reduce the degree of freedom from $X$ and the expression $p(Y|X) = \sum_Z p(Y|X,Z)$ with the network Eq.(1) should make the probability $p(Y|X)$ of Eq.(2) simplified. If there is no $N_Z > 1$ such that Eq.(3) is satisfied, we immediately find that the network of Eq.(2) does not have the middle node $Z$. Note that $N_Z = 1$ shows that there is no edge between $X$ and $Y$, which is already excluded due to the structure of Eq.(2).

**Example 1** *When $N_X = 3$ and $N_Y = 3$, only $N_Z = 1$ satisfies Eq.(3), 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 $N_Z > 1$ such that Eq.(3) is satisfied for given $N_X$ and $N_Y$. 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 network structure can be expressed as Eq.(1),*

$$1 < \hat{N}_Z < \frac{N_X N_Y}{N_X + N_Y - 1}.$$

### 4.2 Asymptotic Properties of the Algorithm

The MCMC method in the 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:

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

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

$$\begin{aligned} n_i =& \sum_{j=1}^{n} \delta_{x_j,i}, \\ n_{ij} =& \sum_{k=1}^{n} \delta_{x_k,i}\delta_{z_k,j}, \\ m_{jk} =& \sum_{l=1}^{n} \delta_{z_l,j}\delta_{y_l,k}, \end{aligned}$$

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

$$\begin{aligned} \varphi(w|\alpha) =& \mathrm{Dir}(a|\alpha_a) \prod_{i=1}^{N_X} \mathrm{Dir}(b_i|\alpha_b) \prod_{j=1}^{N_Z} \mathrm{Dir}(c_j|\alpha_c), \\ \mathrm{Dir}(a|\alpha_a) =& \frac{\Gamma(N_X\alpha_a)}{\Gamma(\alpha_a)^{N_X}} \prod_{i=1}^{N_X} a_i^{\alpha_a - 1}, \\ \mathrm{Dir}(b_i|\alpha_b) =& \frac{\Gamma(N_Z\alpha_b)}{\Gamma(\alpha_b)^{N_Z}} \prod_{j=1}^{N_Z} b_{ij}^{\alpha_b - 1}, \\ \mathrm{Dir}(c_j|\alpha_c) =& \frac{\Gamma(N_Y\alpha_c)}{\Gamma(\alpha_c)^{N_Y}} \prod_{k=1}^{N_Y} c_{jk}^{\alpha_c - 1}. \end{aligned}$$

The Kronecker delta and the gamma function are denoted by $\delta_{ij}$ and $\Gamma(\cdot)$, respectively. The hyperparameter $\alpha$ consists of $\alpha_a$, $\alpha_b$ and $\alpha_c$. 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 $(X^n, Y^n)$. When the number of the given data $n$ is sufficiently large, $F(X^n, Z^n, Y^n)$ is written as

$$\begin{aligned} F(X^n, Z^n, Y^n) =& -nS + C \ln n + O_p(1), \\ S =& \sum_{i=1}^{N_X} \sum_{j=1}^{\tilde{N}_Z} \frac{n_{ij}}{n} \ln \frac{n_{ij}}{n} - \sum_{j=1}^{\tilde{N}_Z} \frac{m_j}{n} \ln \frac{m_j}{n} + \sum_{j=1}^{\tilde{N}_Z} \sum_{k=1}^{N_Y} \frac{m_{jk}}{n} \ln \frac{m_{jk}}{n}, \\ C =& N_X(N_Z - \tilde{N}_Z)\alpha_b + \frac{1}{2}\{N_X - 1 + N_X(\tilde{N}_Z - 1) + \tilde{N}_Z(N_Y - 1)\}, \\ m_j =& \sum_{k=1}^{N_Y} m_{jk}, \end{aligned}$$

where $\tilde{N}_Z$ is the number of $m_j$ such that $m_j/n = O(1)$. The derivation of this expression will be shown in Appendix. 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 of Eq.(1) has the minimal expression, where the range of $Z$ is $z = 1, \dots, N_Z^*$ such that $N_Z^* \leq \tilde{N}_Z$. In the sense of the clustering, the true number of labels is $N_Z^*$ and the estimated one is $\tilde{N}_Z$. We can prove that the Bayesian clustering chooses the minimum structure $\tilde{N}_Z = N_Z^*$ as follows. First, the coefficient is rewritten as

$$S = \sum_{i=1}^{N_X} \frac{n_i}{n} \ln \frac{n_i}{n} + \sum_{i=1}^{N_X} \sum_{j=1}^{\tilde{N}_Z} \frac{n_i}{n} \frac{n_{ij}}{n_i} \ln \frac{n_{ij}}{n_i} + \sum_{j=1}^{\tilde{N}_Z} \sum_{k=1}^{N_Y} \frac{m_j}{n} \frac{m_{jk}}{m_j} \ln \frac{m_{jk}}{m_j}.$$

These three terms correspond to the negative entropy functions of the parameter $a_i$, $b_{ij}$ 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. 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 the 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. 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,

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

This looks similar to the criteria based on the marginal likelihood such as BDu(e) (Heckerman et al., 1995; Buntine, 1991) and its asymptotic form such as BIC (Schwarz, 1978), MDL (Rissanen, 1986). 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 $X^n, Y^n$;

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

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;

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

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 $Z$.

The asymptotic accuracy of the Bayesian clustering has been studied (Yamazaki, 2016), which considers the error function between the true distribution of the label assignment and the estimated one measured by the Kullback-Leibler divergence:

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

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

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

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.

## 6. Conclusions

In this paper, we have proposed a method to detect a hidden node between observable nodes based on the 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.

## Acknowledgments

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

## Appendix

In this section, we derive the asymptotic form of $F_\alpha(X^n, Z^n, Y^n)$. 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) =& (n + N_X\alpha_a)\ln(n + N_X\alpha_a) - \frac{1}{2}\ln(n + N_X\alpha_a) - (n + N_X\alpha_a) \\
& - \sum_{i=1}^{N_X}\left\{(n_i + \alpha_a)\ln(n_i + \alpha_a) - \frac{1}{2}\ln(n_i + \alpha_a) - (n_i + \alpha_a)\right\} \\
& + \sum_{i=1}^{N_X}\left\{\left(\sum_{j=1}^{N_Z} n_{ij} + N_Z\alpha_b\right)\ln\left(\sum_{j=1}^{N_Z} n_{ij} + N_Z\alpha_b\right)\right. \\
& \left. - \frac{1}{2}\ln\left(\sum_{j=1}^{N_Z} n_{ij} + N_Z\alpha_b\right) - \left(\sum_{j=1}^{N_Z} n_{ij} + N_Z\alpha_b\right)\right\} \\
& - \sum_{i=1}^{N_X}\sum_{j=1}^{N_Z}\left\{(n_{ij} + \alpha_b)\ln(n_{ij} + \alpha_b) - \frac{1}{2}\ln(n_{ij} + \alpha_b) - (n_{ij} + \alpha_b)\right\} \\
& + \sum_{j=1}^{N_Z}\left\{\left(\sum_{k=1}^{N_Y} m_{jk} + N_Y\alpha_c\right)\ln\left(\sum_{j=1}^{N_Y} m_{jk} + N_Y\alpha_c\right)\right. \\
& \left. - \frac{1}{2}\ln\left(\sum_{k=1}^{N_Y} m_{jk} + N_Y\alpha_c\right) - \left(\sum_{k=1}^{N_Y} m_{jk} + N_Y\alpha_c\right)\right\} \\
& - \sum_{j=1}^{N_Z}\sum_{k=1}^{N_Y}\left\{(m_{jk} + \alpha_c)\ln(m_{jk} + \alpha_c) - \frac{1}{2}\ln(m_{jk} + \alpha_c) - (m_{jk} + \alpha_c)\right\} \\
& + O_p(1), \\
=& n\ln n + N_X\alpha_a\ln n - \frac{1}{2}\ln n - n \\
& - \sum_{i=1}^{N_X}\left\{n_i\ln n_i + \alpha_a\ln n_i - \frac{1}{2}\ln n_i - n_i\right\} \\
& + \sum_{i=1}^{N_X}\left\{\left(\sum_{j=1}^{N_Y} n_{ij}\right)\ln\sum_{j=1}^{N_Y} n_{ij} + N_Z\alpha_b\ln\sum_{j=1}^{N_Z} n_{ij} - \frac{1}{2}\ln\sum_{j=1}^{N_Z} n_{ij} - \sum_{j=1}^{N_Z} n_{ij}\right\} \\
& - \sum_{i=1}^{N_X}\sum_{j=1}^{N_Z}\left\{n_{ij}\ln n_{ij} + \alpha_b\ln n_{ij} - \frac{1}{2}\ln n_{ij} - n_{ij}\right\} \\
& + \sum_{j=1}^{N_Z}\left\{\sum_{k=1}^{N_Y} m_{jk}\ln\left(\sum_{k=1}^{N_Y} m_{jk}\right) + N_Y\alpha_b\ln\sum_{k=1}^{N_Y} m_{jk}\right. \\
& \left. - \frac{1}{2}\ln\sum_{k=1}^{N_Y} m_{jk} - \sum_{k=1}^{N_Y} m_{jk}\right\} \\
& - \sum_{j=1}^{N_Z}\sum_{k=1}^{N_Y}\left\{m_{jk}\ln m_{jk} + \alpha_c\ln m_{jk} - \frac{1}{2}\ln m_{jk} - m_{jk}\right\} + O_p(1).
\end{aligned}$$

Using the following relations,

$$\begin{aligned}\sum_{i=1}^{N_X} n_i \ln n_i =& n\sum_{i=1}^{N_X}\left(\frac{n_i}{n}\ln\frac{n_i}{n}\right)+n\ln n,\\ \sum_{i=1}^{N_X}\left(\sum_{j=1}^{N_Z} n_{ij}\right)\ln\sum_{j=1}^{N_Z} n_{ij} =& n\sum_{i=1}^{N_X}\left(\frac{n_i}{n}\ln\frac{n_i}{n}\right)+n\ln n,\\ \sum_{i=1}^{N_X}\sum_{j=1}^{N_Z} n_{ij}\ln n_{ij} =& n\sum_{i=1}^{N_X}\sum_{j=1}^{N_Z}\frac{n_{ij}}{n}\ln\frac{n_{ij}}{n}+n\ln n,\\ \sum_{j=1}^{N_Z}\left(\sum_{k=1}^{N_Y} m_{jk}\right)\ln\sum_{k=1}^{N_Y} m_{jk} =& n\sum_{j=1}^{N_Z}\frac{m_j}{n}\ln\frac{m_j}{n}+n\ln n,\\ \sum_{j=1}^{N_Z}\sum_{k=1}^{N_Y} m_{jk}\ln m_{jk} =& n\sum_{j=1}^{N_Z}\sum_{k=1}^{N_Y}\frac{m_{jk}}{n}\ln\frac{m_{jk}}{n}+n\ln n\end{aligned}$$

and focusing on the terms of order $n$ and $\ln n$, we obtain the asymptotic form in Section 4.2.

## References

Hirotsugu Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19:716–723, 1974.

Elizabeth S. Allman, John A. Rhodes, Bernd Sturmfels, and Piotr Zwiernik. Tensors of nonnegative rank two. *Linear Algebra and its Applications*, 473:37 – 53, 2015. Special issue on Statistics.

Miki Aoyagi. Consideration on singularities in learning theory and the learning coefficient. *Entropy*, 15(9):3714–3733, 2013.

Wray Buntine. Theory refinement on bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'91, pages 52–60, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380, 2017. ISSN 1467-9868.

Dan Geiger, David Heckerman, and Christopher Meek. Asymptotic model selection for directed networks with hidden variables. In *UAI '96: Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, pages 283–290, 1996.

I. J. Good. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. Research Monograph No. 30, 1965.

David Heckerman, Dan Geiger, and David M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.*, 20(3):197–243, 1995.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30:2002, 2000.

Jorma Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

Dmitry Rusakov and Dan Geiger. Asymptotic model selection for naive bayesian networks. *Journal of Machine Learning Research*, 6:1–35, 2005.

Gideon E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 (2): 461–464, 1978.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc. ISBN 0-444-89264-8.

Sumio Watanabe. Algebraic analysis for non-identifiable learning machines. *Neural Computation*, 13 (4):899–933, 2001.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, New York, NY, USA, 2009.

Keisuke Yamazaki. Asymptotic accuracy of Bayes estimation for latent variables with redundancy. *Machine Learning*, 102(1):1–28, 2016.

Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks*, 16:1029–1038, 2003.

Keisuke Yamazaki and Sumio Watanabe. Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, 69(1-3):62–84, 2005.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873 – 1896, 2008.