Mach Learn (2009) 75: 199–215
DOI 10.1007/s10994-008-5099-x

Upper bound for variational free energy of Bayesian networks

Kazuho Watanabe · Motoki Shiga · Sumio Watanabe

Received: 20 February 2007 / Revised: 1 September 2008 / Accepted: 27 December 2008 /
Published online: 24 January 2009

Abstract In recent years, variational Bayesian learning has been used as an approximation of Bayesian learning. In spite of the computational tractability and good generalization in many applications, its statistical properties have yet to be clarified. In this paper, we focus on variational Bayesian learning of Bayesian networks which are widely used in information processing and uncertain artificial intelligence. We derive upper bounds for asymptotic variational free energy or stochastic complexities of bipartite Bayesian networks with discrete hidden variables. Our result theoretically supports the effectiveness of variational Bayesian learning as an approximation of Bayesian learning.

Keywords Bipartite Bayesian networks · Variational Bayes framework · Variational free energy

1 Introduction

Recently, Bayesian networks have been widely used in information processing and uncertain artificial intelligence (Jordan 1999; Jensen 2001). For example, they have been applied to bioinformatics, image analysis, and so on (Friedman 2004; Meltzer et al. 2005). In spite of

Editor: Zoubin Ghahramani.

K. Watanabe (✉)
Department of Complexity Science and Engineering, The University of Tokyo, Mail Box 409,
5-1-5 Kashiwanoha, Kashiwa, 277-8561 Japan
e-mail: kazuho@mns.k.u-tokyo.ac.jp

M. Shiga
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji,
Kyoto 611-0011, Japan
e-mail: shiga@kuicr.kyoto-u.ac.jp

S. Watanabe
P&I Lab., Tokyo Institute of Technology, Mail Box R2-5, 4259 Nagatsuda, Midori-ku, Yokohama,
226-8503 Japan
e-mail: swatanab@pi.titech.ac.jp

their wide range of application, statistical properties such as the generalization error have yet to be clarified.

The main reasons for this difficulty are due to their non-identifiability. If the mapping from the parameter w of the learning machine to the probabilistic model $p(x|w)$ is one-to-one, then the model is called identifiable, and otherwise, non-identifiable. An identifiable model is further called regular if the asymptotic normality of the maximum likelihood estimator holds. One of the difficulties in the analysis of the non-identifiable model is that we cannot apply the asymptotic theory of regular statistical models to a non-identifiable one. If the model attains the true distribution from which sample data are taken, the true parameter is not one point but an analytic set with singularities in the parameter space. This is why the mathematical properties of the non-identifiable models have been unknown.

In recent years, however, the asymptotic theory for Bayesian learning of non-identifiable models has been established with an algebraic-geometric method (Watanabe 2001). The method revealed the relation between model's singularities and its statistical properties. For Bayesian networks, the asymptotic form of the Bayesian stochastic complexity, namely the free energy, was derived (Yamazaki and Watanabe 2003b; Rusakov and Geiger 2005). The result shows that the stochastic complexity of Bayesian networks is much smaller than complexity of regular models.

On the other hand, performing Bayesian learning is computationally intractable for non-identifiable models. The variational Bayesian framework was proposed as an approximation method of Bayesian learning (Hinton and van Camp 1993) and extended to statistical models with hidden variables (Attias 1999; Beal 2003; Beal and Ghahramani 2003). This framework provides computationally tractable posterior distributions over the hidden variables and parameters with an iterative algorithm. Variational Bayesian learning has been applied to various learning machines and it has performed with good generalization at only a modest computational cost compared to Markov Chain Monte Carlo (MCMC) methods, which are the major schemes in Bayesian learning.

Several properties of the variational Bayesian approach have been studied recently. Wang and Titterington investigated the local convergence property of the variational Bayesian estimator (Wang and Titterington 2004, 2006) and examined the covariance of the estimator (Wang and Titterington 2005). Asymptotic forms of variational free energies of mixture models and hidden Markov models were derived (Watanabe and Watanabe 2005, 2006a, 2006b; Hosino et al. 2005).

In this paper, we analyze the variational free energy of Bayesian networks. We derive an upper bound for the variational free energy for bipartite structured graphical models with discrete hidden variables. The upper bound is obtained under the assumption that the true data generating distribution is included in the set of models. This assumption is essential for addressing model selection or hypothesis testing where we compare the variational free energies of the possible models. We decompose the variational free energy into the sum of the complexity-control term and the likelihood term. The likelihood term is evaluated by the empirical entropy up to an additive constant. The complexity-control term is given by the Kullback information from the approximate posterior distribution to the prior distribution. We give the upper bound on it in the form, $\nu \log n + C$, where n is the sample size, C is a constant independent of n, and the constant ν is identified by the structures of the Bayesian network and the true network.

The variational free energy, which is also called the variational stochastic complexity and corresponds to a lower bound for the Bayesian evidence, is a key quantity for model selection. Evaluating the variational free energy also contributes to the following two issues. One is the accuracy of variational Bayesian learning as an approximation method since the variational free energy shows the distance from the variational posterior distribution to the true

Bayesian posterior distribution in terms of Kullback information. Another is the influence of the hyperparameters on the learning process. Since the variational Bayesian algorithm minimizes the variational free energy function, the derived bounds indicate how the hyperparameters influence the learning process. The main results indicate how to determine the hyperparameter values before the learning process.

The paper is organized as follows. Section 2 introduces Bayesian networks. Section 3 reviews the framework of Bayesian learning and its asymptotic analysis. Section 4 describes the general framework of variational Bayesian learning and the variational Bayesian algorithm for the Bayesian network is also derived. Section 5 presents the main theorem which shows the upper bounds for the variational free energy of the Bayesian network. The proof of the main theorem is presented in Sect. 6. Section 7 presents the results of the numerical experiments demonstrating the tightness of the bounds.

In order to discuss the approximation accuracy of Variational Bayes, the derived bounds are compared to the Bayesian free energy obtained in previous works and the effect of the hyperparameters is also discussed in Sect. 8. Finally, Sect. 9 concludes the paper.

2 Bayesian networks

A graphical model expresses the relations among random variables by a graph. Bayesian networks are included in graphical models. Bayesian network is defined by a directed graph and conditional probabilities (Jensen 2001).

In this paper, we focus on a Bayesian network whose states of all hidden nodes influence those of all observation nodes, and assume that it has N observation nodes and K hidden nodes. The graphical structure of this Bayesian network is called bipartite and presented in Fig. 1.

The observation nodes are denoted by a vector $x = (x_1, x_2, \ldots, x_N)$, and the set of states of observation node x_j is $\{1, 2, \ldots, Y_j\}$. The hidden nodes are denoted by a vector $z = (z_1, z_2, \ldots, z_K)$, and the set of states of hidden node z_k is $\{1, 2, \ldots, T_k\}$.

The probability that the state of the hidden node z_k is i $(1 \leq i \leq T_k)$, is expressed as

$$a_{(k,i)} := P(z_k = i). \tag{1}$$

Because $a_k := \{a_{(k,i)}; i = 1, 2, \ldots, T_k\}$ is a probability distribution, $\sum_{i=1}^{T_k} a_{(k,i)} = 1$ holds for $k = 1, 2, \ldots, K$.

The conditional probability that the jth observation node x_j is l $(1 \leq l \leq Y_j)$, given the condition that the states of hidden nodes are $z = (z_1, z_2, \ldots, z_K)$, is denoted by

$$b_{(j,l|z)} := P(x_j = l|z). \tag{2}$$

Then $b_{(j,\cdot|z)} := \{b_{(j,l|z)}; l = 1, 2, \ldots, Y_j\}$ satisfies $\sum_{l=1}^{Y_j} b_{(j,l|z)} = 1$, for $j = 1, 2, \ldots, N$. Define $a := \{a_k; k = 1, 2, \ldots, K\}$ and $b := \{b_z; 1 \leq z_1 \leq T_1, \ldots, 1 \leq z_K \leq T_K\}$, where

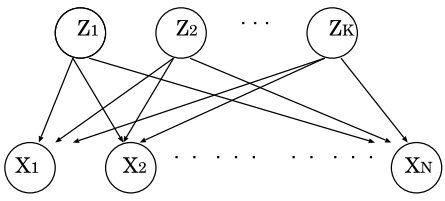

Fig. 1 Graphical structure of the Bayesian network

$b_z := \{b_{(j,\cdot|z)};\ j = 1, 2, \ldots, N\}$. Let $\omega = \{a, b\}$ be the set of all parameters. Then the joint probability that the states of observation nodes are $x = (x_1, x_2, \ldots, x_N)$ and the states of hidden nodes are $z = (z_1, z_2, \ldots, z_K)$ is

$$p(x, z|\omega) = c(x|b_z) \prod_{k=1}^{K} a_{(k,z_k)}, \tag{3}$$

where

$$c(x|b_z) = \prod_{j=1}^{N} b_{(j,x_j|z)}. \tag{4}$$

Therefore the marginal probability that the states of observation nodes are x is

$$\begin{aligned} p(x|\omega) &= \sum_{z} p(x, z|\omega) \\ &= \left\{ \prod_{k=1}^{K} \sum_{z_k=1}^{T_k} \right\} c(x|b_z) \prod_{k=1}^{K} a_{(k,z_k)}, \end{aligned} \tag{5}$$

where we use the notation $\sum_z = \{\prod_{k=1}^{K} \sum_{z_k=1}^{T_k}\} := \sum_{z_1=1}^{T_1} \sum_{z_2=1}^{T_2} \cdots \sum_{z_K=1}^{T_K}$ for the summation over all states of hidden nodes. Let

$$M = \sum_{j=1}^{N} (Y_j - 1),$$

which is the number of parameters to specify the conditional probability $c(x|b_z)$ of the states of all the observation nodes given the states of the hidden nodes. Then the number of the parameters of the model, d, is

$$d = M \prod_{k=1}^{K} T_k + \sum_{k=1}^{K} (T_k - 1). \tag{6}$$

A Bayesian network is non-identifiable if it has a hidden variable. Hence, as will be noted in Sect. 3, the Bayesian stochastic complexity can no longer be well approximated by the well-known model selection criteria such as the minimum description length (MDL) (Rissanen 1986). Indeed, it was pointed out that the MDL also known as the Bayesian information criterion (BIC) may not work well for Bayesian networks (Allen and Greiner 2000).

3 Bayesian learning

Suppose n training samples $X^n = \{X_1, X_2, \ldots, X_n\}$ are independently and identically taken from the true distribution $p_0(x)$. Let $\varphi(\omega)$ be the prior distribution of the parameters ω. Then the posterior distribution $p(\omega|X^n)$ is computed from the given dataset and the prior by

$$p(\omega|X^n) = \frac{1}{Z(X^n)} \exp(-nH_n(\omega))\varphi(\omega), \tag{7}$$

where $H_n(\omega) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|\omega)$, and $Z(X^n)$ is the normalization constant called the marginal likelihood or the evidence of the dataset X^n (Mackay 1992). The Bayesian predictive distribution $p(x|X^n)$ is given by averaging the model over the posterior distribution as follows,

$$p(x|X^n) = \int p(x|\omega)p(\omega|X^n)d\omega. \tag{8}$$

The Bayesian stochastic complexity $F(X^n)$ is defined by

$$F(X^n) = -\log Z(X^n), \tag{9}$$

which is also called the free energy and is important in most data modeling problems. In practice, it is used as a criterion by which the model is selected and the hyperparameters in the prior are optimized (Schwarz 1978).

Let $E_{X^n}[\cdot]$ be the expectation over all sets of training data and

$$S(X^n) = -\sum_{i=1}^{n} \log p_0(X_i) \tag{10}$$

be the empirical entropy. It was proved that the Bayesian stochastic complexity has the following asymptotic form (Watanabe 2001),

$$E_{X^n}\big[F(X^n) - S(X^n)\big] \simeq \lambda \log n - (m-1)\log\log n + O(1), \tag{11}$$

where λ and m are, respectively, the rational number and the natural number which are determined by the singularities of the true parameters. In regular models, 2λ is equal to the number of parameters and $m = 1$, while in non-identifiable models, 2λ is not larger than the number of parameters and $m \geq 1$ (Yamazaki and Watanabe 2003b; Rusakov and Geiger 2005).

In practice, Bayesian learning requires integration over the posterior distribution, which typically cannot be performed analytically. As an approximation, the variational Bayesian framework was introduced in neural networks (Hinton and van Camp 1993) and was extended to deal with statistical models containing hidden variables (Attias 1999).

4 Variational Bayesian learning for Bayesian networks

This section reviews the variational Bayesian framework and that for the Bayesian network model defined by (5).

4.1 Variational Bayesian learning

Let $\{X^n, Z^n\}$ be the complete data with the addition of the corresponding hidden variables $Z^n = \{Z_1, Z_2, \ldots, Z_n\}$. The variational Bayesian framework approximates the Bayesian posterior $p(Z^n, \omega|X^n)$ of the hidden variables and the parameters by the variational posterior $q(Z^n, \omega|X^n)$, which factorizes as

$$q(Z^n, \omega|X^n) = Q(Z^n|X^n)r(\omega|X^n), \tag{12}$$

where $Q(Z^n|X^n)$ and $r(\omega|X^n)$ are posteriors over the hidden variables and the parameters respectively. The variational posterior $q(Z^n, \omega|X^n)$ is chosen to minimize the functional $\overline{F}[q]$ defined by

$$\begin{aligned}\overline{F}[q] &= \sum_{Z^n} \int q(Z^n, \omega|X^n) \log \frac{q(Z^n, \omega|X^n)}{p(X^n, Z^n, \omega)} d\omega, \\ &= F(X^n) + K(q(Z^n, \omega|X^n)||p(Z^n, \omega|X^n)),\end{aligned} \tag{13}$$

where $K(q(Z^n, \omega|X^n)||p(Z^n, \omega|X^n))$ is the Kullback information between the true Bayesian posterior $p(Z^n, \omega|X^n)$ and the variational posterior $q(Z^n, \omega|X^n)$. The functional $\overline{F}[q]$ is called the variational free energy function and it measures the quality of the approximation. The above minimization problem leads to the following theorem. The proof is well known (Beal 2003).

Theorem 1 *If the functional $\overline{F}[q]$ is minimized under constraint (12) then the variational posteriors, $r(\omega|X^n)$ and $Q(Z^n|X^n)$, satisfy*

$$r(\omega|X^n) = \frac{1}{C_r} \varphi(\omega) \exp\langle \log p(X^n, Z^n|\omega) \rangle_Q, \tag{14}$$

$$Q(Z^n|X^n) = \frac{1}{C_Q} \exp\langle \log p(X^n, Z^n|\omega) \rangle_r, \tag{15}$$

where C_r and C_Q are the normalization constants.[1]

Note that (14) and (15) give only the necessary conditions for the functional $\overline{F}[q]$ to be minimized. The variational posteriors that satisfy (14) and (15) are computed by an iterative algorithm whose convergence is guaranteed.

We define the variational free energy $\overline{F}(X^n)$ by the minimum value of the functional $\overline{F}[q]$, that is,

$$\overline{F}(X^n) = \min_{r,Q} \overline{F}[q]. \tag{16}$$

From (13), the difference between $\overline{F}(X^n)$ and the Bayesian stochastic complexity $F(X^n)$ shows the accuracy of the variational Bayesian approach as an approximation of Bayesian learning.

4.2 Variational posterior for Bayesian networks

We assume that the prior distribution $\varphi(\omega)$ of the parameters $\omega = \{a, b\}$ is the conjugate prior distribution. Then $\varphi(\omega)$ is given by $\{\prod_{k=1}^K \varphi(a_k)\}\{\prod_z \prod_{j=1}^N \varphi(b_{(j,\cdot|z)})\}$, where

$$\varphi(a_k) = \frac{\Gamma(T_k \phi_0)}{\Gamma(\phi_0)^{T_k}} \prod_{z_k=1}^{T_k} a_{(k,z_k)}^{\phi_0 - 1}, \tag{17}$$

[1] Hereafter we use the notation $\langle\cdot\rangle_r$ and $\langle\cdot\rangle_Q$ for the expectation over $r(\omega|X^n)$ and $Q(Z^n|X^n)$ respectively.

and

$$\varphi(b_{(j,\cdot|z)}) = \frac{\Gamma(Y_j\xi_0)}{\Gamma(\xi_0)^{Y_j}} \prod_{x_j=1}^{Y_j} b_{(j,x_j|z)}^{\xi_0-1}, \tag{18}$$

are Dirichlet distributions with hyperparameters $\phi_0 > 0$ and $\xi_0 > 0$, and

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

is the gamma function. Let $\delta(n)$ be 1 when $n = 0$ and 0 otherwise, and define

$$\bar{n}_{(k,z_k)}^z := \sum_{i=1}^n \left\langle \delta(Z_i^{(k)} - z_k) \right\rangle_Q, \tag{19}$$

and

$$\bar{n}_{(j,x_j|z)}^x := \sum_{i=1}^n \delta(X_i^{(j)} - x_j) \left\langle \prod_{k=1}^K \delta(Z_i^{(k)} - z_k) \right\rangle_Q. \tag{20}$$

Here $X_i^{(j)}$ is the state of the jth observation node and $Z_i^{(k)}$ is the state of the kth hidden node when the ith training datum is observed. From (14), the variational posterior distribution of parameters $\omega = \{a, b\}$ is given by $r(\omega|X^n) = \{\prod_{k=1}^K r(a_k|X^n)\}\{\prod_z \prod_{j=1}^N r(b_{(j,\cdot|z)}|X^n)\}$, where

$$r(a_k|X^n) = \frac{\Gamma(n + T_k\phi_0)}{\prod_{z_k=1}^{T_k} \Gamma(\bar{n}_{(k,z_k)}^z + \phi_0)} \prod_{z_k=1}^{T_k} a_{(k,z_k)}^{\bar{n}_{(k,z_k)}^z + \phi_0 - 1}, \tag{21}$$

$$r(b_{(j,\cdot|z)}|X^n) = \frac{\Gamma(\bar{n}_z^x + Y_j\xi_0)}{\prod_{x_j=1}^{Y_j} \Gamma(\bar{n}_{(j,x_j|z)}^x + \xi_0)} \prod_{x_j=1}^{Y_j} b_{(j,x_j|z)}^{\bar{n}_{(j,x_j|z)}^x + \xi_0 - 1}, \tag{22}$$

and

$$\bar{n}_z^x := \sum_{i=1}^n \left\langle \prod_{k=1}^K \delta(Z_i^{(k)} - z_k) \right\rangle_Q.$$

Note that

$$\bar{n}_z^x = \sum_{x_j=1}^{Y_j} \bar{n}_{(j,x_j|z)}^x,$$

for $j = 1, \ldots, N$, and

$$\bar{n}_{(k,z_k)}^z = \sum_{z_{-k}} \bar{n}_z^x, \tag{23}$$

where $\sum_{z_{-k}}$ denotes the sum over z_i $(i \neq k)$.

It follows from (21) and (22) that

$$\langle \log a_{(k,z_k)} \rangle_r = \Psi(\bar{n}_{(k,z_k)}^z + \phi_0) - \Psi(n + T_k\phi_0),$$

for $k = 1, 2, \ldots, K$ and

$$\langle \log b_{(j,x_j|z)} \rangle_r = \Psi(\bar{n}^x_{(j,x_j|z)} + \xi_0) - \Psi(\bar{n}^x_z + Y_j\xi_0),$$

for $j = 1, 2, \ldots, N$ and given z, where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function. From (15), the variational posterior distribution of the hidden variables is given by $Q(Z^n|X^n) = \prod_{i=1}^n Q(Z_i|X^n)$, where

$$\begin{aligned} Q(Z_i = z|X^n) &= \sum_{Z_i} Q(Z_i|X^n)\delta(Z_i - z) \\ &= \left\langle \prod_{k=1}^K \delta(Z_i^{(k)} - z_k) \right\rangle_Q \\ &\propto \exp\left\{ \sum_{k=1}^K \{\Psi(\bar{n}^z_{(k,z_k)} + \phi_0) - \Psi(n + T_k\phi_0)\} \right. \\ &\quad \left. + \sum_{j=1}^N \{\Psi(\bar{n}^x_{(j,X_i^{(j)}|z)} + \xi_0) - \Psi(\bar{n}^x_z + Y_j\xi_0)\} \right\}. \end{aligned} \tag{24}$$

The variational Bayesian algorithm updates $\{\bar{n}^x_{(j,x_j|z)}\}$ using (20) and (24) iteratively.

5 Main results

We assume the following conditions.

(A1) The true distribution is defined by a Bayesian network with H hidden nodes, each with S_k states, where $H \le K$. Then the true distribution $p_0(x)$ is

$$p(x|\omega^*) = \left\{ \prod_{k=1}^H \sum_{z_k=1}^{S_k} \right\} c(x|b^*_z) \prod_{k=1}^H a^*_{(k,z_k)}, \tag{25}$$

where $c(x|b^*_z) = \prod_{j=1}^N b^*_{(j,x_j|z)}$ and the true parameters $\omega^* = \{a^*, b^*\}$ are given by

$$\begin{aligned} a^* &= \{a^*_k; k = 1, 2, \ldots, H\}, \\ b^* &= \{b^*_z; 1 \le z_1 \le S_1, 1 \le z_2 \le S_2, \ldots, 1 \le z_H \le S_H\}, \\ b^*_z &= \{b^*_{(j,\cdot|z)}; j = 1, 2, \ldots, N\}, \\ a^*_k &= \{a^*_{(k,z_k)}; 1 \le z_k \le S_k\} \quad (k = 1, 2, \ldots, H), \\ b^*_{(j,\cdot|z)} &= \{b^*_{(j,x_j|z)}; 1 \le x_j \le Y_j\} \quad (j = 1, 2, \ldots, N). \end{aligned} \tag{26}$$

For $k > H$, we define $S_k = 1$ and $a^*_{(k,z_k)} = \delta(z_k - 1)$.

The true distribution can be realized by the model, that is, the model given by (5) where $T_k \ge S_k$ holds for $k = 1, 2, \ldots, H$.

(A2) The prior distribution of parameters $\omega = (a, b)$ is the conjugate prior distribution, $\varphi(\omega) = \{\prod_{k=1}^K \varphi(a_k)\}\{\prod_z \prod_{j=1}^N \varphi(b_{(j,\cdot|z)})\}$, where $\varphi(a_k)$ and $\varphi(b_{(j,\cdot|(z_1,z_2))})$ are given by (17) and (18).

Under these conditions, we prove the following theorem. The proof will appear in the next section.

Theorem 2 (Main result) *Assume the conditions* (A1) *and* (A2). *Then for an arbitrary natural number n, the variational free energy satisfies*

$$\overline{F}(X^n) - S(X^n) \leq \nu \log n + C \tag{27}$$

with probability 1, *where C is a constant independent of n and*

$$\nu = \phi_0 \sum_{k=1}^{K} T_k - \frac{K}{2} + \min_{\{u_k\}} \left\{ \frac{M}{2} \prod_{k=1}^{K} u_k - \left(\phi_0 - \frac{1}{2}\right) \sum_{k=1}^{K} u_k \right\}. \tag{28}$$

The minimum is taken over the set of positive integers $\{u_k; S_k \leq u_k \leq T_k\}_{k=1}^{K}$.

If $K = 1$, this reduces to the case of the naive Bayesian networks whose free energy or stochastic complexity has been evaluated (Yamazaki and Watanabe 2003a; Rusakov and Geiger 2005). Bounds for their variational free energy have also been obtained (Watanabe and Watanabe 2005, 2006b).

The coefficient ν is given by the solution of the minimization problem. We present examples of the upper bound as corollaries below.

By taking $u_k = S_k$ for $1 \leq k \leq H$ and $u_k = 1$ for $H+1 \leq k \leq K$, we obtain the following upper bound for the variational free energy (Watanabe et al. 2006). This bound is minimal if $\phi_0 \leq (1 + M \min_{1 \leq k \leq K}\{S_k\})/2$.

Corollary 1 *For* $\phi_0 \leq (1 + M \min_{1 \leq k \leq K}\{S_k\})/2$,

$$\overline{F}(X^n) - S(X^n) \leq \nu \log n + C, \tag{29}$$

where C is a constant independent of n and

$$\nu = \phi_0 \sum_{k=1}^{K} T_k - \phi_0 K + \left(\phi_0 - \frac{1}{2}\right) H + \left(\frac{1}{2} - \phi_0\right) \sum_{k=1}^{H} S_k + \frac{M}{2} \prod_{k=1}^{H} S_k. \tag{30}$$

If $K = H = 2$, that is, the true network and the model both have 2 hidden nodes, solving the minimization problem gives the following corollary. Suppose $S_1 \geq S_2$ and $T_1 \geq T_2$.

Corollary 2 *If* $K = H = 2$,

$$\overline{F}(X^n) - S(X^n) \leq \nu \log n + C, \tag{31}$$

where C is a constant independent of n and

$$\nu = \begin{cases} (T_1 - S_1 + T_2 - S_2)\phi_0 + \frac{M}{2} S_1 S_2 + \frac{S_1+S_2}{2} - 1 & (0 < \phi_0 \leq \frac{1+S_2 M}{2}), \\ (T_2 - S_2)\phi_0 + \frac{M}{2} T_1 S_2 + \frac{T_1+S_2}{2} - 1 & (\frac{1+S_2 M}{2} < \phi_0 \leq \frac{1+T_1 M}{2}), \\ \frac{M}{2} T_1 T_2 + \frac{T_1+T_2}{2} - 1 & (\frac{1+T_1 M}{2} < \phi_0). \end{cases} \tag{32}$$

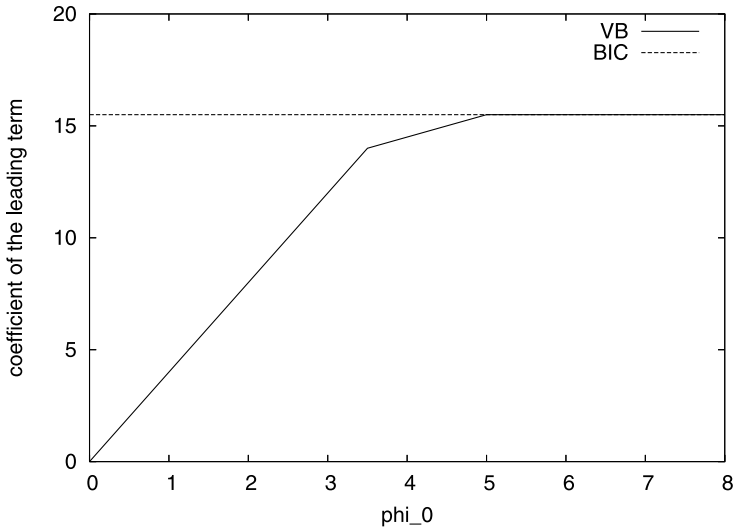

Fig. 2 The coefficient ν of the upper bound for the variational free energy (*solid line*) as a function of the hyperparameter ϕ_0 with $T_1 = T_2 = 3$, $S_1 = S_2 = 2$ and $M = 3$. The *dashed line* is the coefficient $d/2$ of the BIC where d is the number of the parameters

The penalty term in the BIC is given by $(d/2)\log n$ (Schwarz 1978) where d is the number of the parameters defined by (6). Corollary 2 claims that the coefficient ν of the leading term is smaller than $d/2$ when $\phi_0 \leq \frac{1+T_1 M}{2}$. The coefficient ν in the above corollary is demonstrated in Fig. 2 as a function of the hyperparameter ϕ_0 with $T_1 = T_2 = 3$, $S_1 = S_2 = 2$, and three binary observed nodes $Y_1 = Y_2 = Y_3 = 2$; that is, $M = 3$. The effect of the hyperparameter ϕ_0 is discussed in Sect. 8.3.

6 Proof of Theorem 2

This section proves the main theorem.

From (15), we can rewrite the variational free energy as follows,

$$\overline{F}(X^n) = \min_r \left[K\Big(r(\omega|X^n)||\varphi(\omega)\Big) - \log C_Q \right], \tag{33}$$

where

$$\log C_Q = \log \sum_{Z^n} \exp\Big\langle \log p(X^n, Z^n|\omega) \Big\rangle_r. \tag{34}$$

From (21), (22) and (24), we obtain $\log C_Q$ and $K(r(\omega|X^n)||\varphi(\omega))$ in (33) as follows,

$$\begin{aligned}
\log C_Q &= \sum_{i=1}^{n} \log \sum_{Z_i} \exp\langle \log p(X_i, Z_i|\omega)\rangle_r \\
&= \sum_{i=1}^{n} \log \left[\sum_{z} \exp \left\{ \sum_{k=1}^{K} \{\Psi(\bar{n}^z_{(k,z_k)} + \phi_0) - \Psi(n + T_k\phi_0)\} \right.\right. \\
&\quad \left.\left. + \sum_{j=1}^{N} \{\Psi(\bar{n}^x_{(j,X_i^{(j)}|z)} + \xi_0) - \Psi(\bar{n}^x_z + Y_j\xi_0)\} \right\} \right],
\end{aligned} \tag{35}$$

and

$$K\Big(r(\omega|X^n)||\varphi(\omega)\Big) = \sum_{k=1}^{K} K\left(r(a_k|X^n)||\varphi(a_k)\right) + \sum_{z}\sum_{j=1}^{N} K\left(r(b_{(j,\cdot|z)}|X^n)||\varphi(b_{(j,\cdot|z)})\right)$$

$$= \sum_{k=1}^{K}\Bigg[\sum_{z_k=1}^{T_k}\left\{\bar{n}^z_{(k,z_k)}\Psi(\bar{n}^z_{(k,z_k)}+\phi_0) - \log\Gamma(\bar{n}^z_{(k,z_k)}+\phi_0)\right\}$$

$$- n\Psi(n+T_k\phi_0) + \log\Gamma(n+T_k\phi_0) + \log\frac{\Gamma(\phi_0)^{T_k}}{\Gamma(T_k\phi_0)}\Bigg]$$

$$+ \sum_{z}\sum_{j=1}^{N}\Bigg[\sum_{x_j=1}^{Y_j}\left\{\bar{n}^x_{(j,x_j|z)}\Psi(\bar{n}^x_{(j,x_j|z)}+\xi_0) - \log\Gamma(\bar{n}^x_{(j,x_j|z)}+\xi_0)\right\}$$

$$- \bar{n}^x_z\Psi(\bar{n}^x_z+Y_j\xi_0) + \log\Gamma(\bar{n}^x_z+Y_j\xi_0) + \log\frac{\Gamma(\xi_0)^{Y_j}}{\Gamma(Y_j\xi_0)}\Bigg]. \tag{36}$$

Furthermore, by using the inequalities for the digamma function, for $x>0$

$$\frac{1}{2x} < \log x - \Psi(x) < \frac{1}{x} \quad (x>0)$$

and for the log-gamma function,

$$0 \le \log\Gamma(x) - \left(x-\frac{1}{2}\right)\log x + x - \frac{1}{2}\log 2\pi \le \frac{1}{12x} \quad (x>0)$$

we can bound $\log C_Q$:

$$\log C_Q \ge \sum_{i=1}^{n}\log\Bigg[\sum_{z}\exp\Bigg\{\sum_{k=1}^{K}\Bigg\{\log\frac{\bar{n}^z_{(k,z_k)}+\phi_0}{n+T_k\phi_0} - \frac{1}{\bar{n}^z_{(k,z_k)}+\phi_0} + \frac{1}{2(n+T_k\phi_0)}\Bigg\}$$

$$+ \sum_{j=1}^{N}\Bigg\{\log\frac{\bar{n}^x_{(j,X_i^{(j)}|z)}+\xi_0}{\bar{n}^x_z+Y_j\xi_0} - \frac{1}{\bar{n}^x_{(j,X_i^{(j)}|z)}+\xi_0} + \frac{1}{2(\bar{n}^x_z+Y_j\xi_0)}\Bigg\}\Bigg\}\Bigg]. \tag{37}$$

We can also bound $K(r(\omega|X^n)||\varphi(\omega))$ in (33):

$$K\Big(r(\omega|X^n)||\varphi(\omega)\Big) \le \sum_{k=1}^{K}\left\{\left(T_k\phi_0-\frac{1}{2}\right)\log(n+T_k\phi_0)\right\}$$

$$- \sum_{k=1}^{K}\sum_{z_k=1}^{T_k}\left\{\left(\phi_0-\frac{1}{2}\right)\log\left(\bar{n}^z_{(k,z_k)}+\phi_0\right)\right\}$$

$$+ \sum_{z}\sum_{j=1}^{N}\Bigg\{\left(Y_j\xi_0-\frac{1}{2}\right)\log\left(\bar{n}^x_z+Y_j\xi_0\right)$$

$$- \sum_{x_j=1}^{Y_j}\left(\xi_0-\frac{1}{2}\right)\log\left(\bar{n}^x_{(j,x_j|z)}+\xi_0\right)\Bigg\} + O(1). \tag{38}$$

From (33), since $\overline{F}(X^n)$ is given as the minimum value of the function of $\{\bar{n}^x_{(j,x_j|z)}\}$, we can obtain an upper bound for $\overline{F}(X^n)$ by substituting each $\bar{n}^x_{(j,x_j|z)}$ by any specific value. Therefore let u_k be a natural number such that $S_k \le u_k \le T_k$ for $k = 1, 2, \ldots, K$ and consider the following $\bar{n}^x_{(j,x_j|z)}$ for each j and x_j:

$$\bar{n}^x_{(j,x_j|z)} = nb^*_{(j,x_j|\overline{z})} \prod_{k=1}^{K} a'_{(k,z_k)}, \tag{39}$$

where $\overline{z} = (\min\{z_1, S_1\}, \min\{z_2, S_2\}, \ldots, \min\{z_H, S_H\})$ and

$$a'_{(k,z_k)} = \begin{cases} a^*_{(k,z_k)} & (1 \le z_k \le S_k - 1), \\ a^*_{(k,S_k)}/(u_k - S_k + 1) & (S_k \le z_k \le u_k), \\ 0 & (\text{otherwise}). \end{cases} \tag{40}$$

This corresponds to the case when u_k ($\ge S_k$) states of the kth hidden node are active for $k = 1, 2, \ldots, H$. Then we have $\bar{n}^x_z = n \prod_{k=1}^{K} a'_{(k,z_k)}$ and $\bar{n}^z_{(k,z_k)} = na'_{(k,z_k)}$. Substituting them into (38) yields

$$K\Big(r(\omega|X^n)||\varphi(\omega)\Big) \le \left\{ \phi_0 \sum_{k=1}^{K} T_k - \frac{K}{2} + \frac{M}{2} \prod_{k=1}^{K} u_k - \left(\phi_0 - \frac{1}{2}\right) \sum_{k=1}^{K} u_k \right\} \log n + O(1). \tag{41}$$

From (37), we obtain

$$\begin{aligned} \log C_Q &\ge \sum_{i=1}^{n} \log\left(p(X_i|\omega^*) \exp\left(\frac{C'}{n}\right)\right) \\ &= -S(X^n) + C', \end{aligned} \tag{42}$$

where C' is a constant. From (33), (42) and (41), we complete the proof. □

7 Experiments

We applied variational Bayesian learning to a Bayesian network with two hidden nodes and computed the variational free energy to compare it to the derived upper bound and to the BIC.

We generated data from the true network with $H = 1$ hidden node and $N = 4$ observed nodes. Each observed node has $Y_j = 4$ states and the hidden node is binary-valued, $S_1 = 2$. The parameters ω^* were set to fixed values. Specifically, $a^*_{(1,1)} = 1/3$, $a^*_{(1,2)} = 2/3$, $b^*_{(j,l|1)}$ for all j are 5/8, 1/8, 1/8, 1/8, and $b^*_{(j,l|2)}$ for all j are 1/8, 5/8, 1/8, 1/8 for $l = 1, 2, 3, 4$. We used as a learner the Bayesian network model with 2 binary-valued hidden nodes (one redundant node), that is, $K = 2$ and $T_1 = T_2 = 2$, to consider the non-identifiable parameter settings.

We compared the coefficient of the leading term of the free energy, which is of the order of $\log n$, instead of the energy itself since its constant order term (denoted by C in (27)) is not zero although it can be negligible in our asymptotic theoretical analysis. Two data sets with the size $n = 1000$ and $n = 500$ were generated from the true network. For these data sets, we ran the variational Bayesian algorithm and computed the variational free energies

subtracted by the empirical entropies $S(X^{1000})$ and $S(X^{500})$, respectively. Denoting them as $\overline{F}_0(X^{1000})$ and $\overline{F}_0(X^{500})$, we calculated

$$\hat{\nu} = (\overline{F}_0(X^{1000}) - \overline{F}_0(X^{500}))/\log 2$$

to estimate the coefficient of the leading term of the normalized variational free energy $\overline{F}(X^n) - S(X^n)$. Theorem 2 implies $\hat{\nu} \leq \nu$.

We also evaluated the corresponding value $\hat{\nu}_{BIC}$ for the BIC by replacing the variational free energy $\overline{F}(X^n)$ with the BIC,

$$F_{\mathrm{BIC}}(X^n) = \frac{d}{2}\log n - \sum_{i=1}^{n} \log p(X_i|\hat{\omega}_{\mathrm{MAP}}),$$

where $\hat{\omega}_{\mathrm{MAP}}$ is the maximum a posteriori (MAP) estimator of the parameters ω. The MAP estimator was obtained by slightly modifying the above variational Bayesian algorithm with the same initial conditions. More specifically, the terms $\Psi(\bar{n}^z_{(k,z_k)} + \phi_0) - \Psi(n + T_k\phi_0)$ and $\Psi(\bar{n}^x_{(j,X_i^{(j)}|z)} + \xi_0) - \Psi(\bar{n}^x_z + Y_j\xi_0)$ in (24) were replaced by $\log(\bar{n}^z_{(k,z_k)} + \phi_0) - \log(n + T_k\phi_0)$ and $\log(\bar{n}^x_{(j,X_i^{(j)}|z)} + \xi_0) - \log(\bar{n}^x_z + Y_j\xi_0)$, respectively. To take into account the difference of the likelihood terms between the variational free energy and the BIC, we also calculated

$$F_{\mathrm{VB-BIC}}(X^n) = \frac{d}{2}\log n - \log C_Q,$$

where $\log C_Q$ is defined by (34), and obtained $\hat{\nu}_{\mathrm{VB-BIC}}$ in the same way as we obtained $\hat{\nu}$ and $\hat{\nu}_{\mathrm{BIC}}$ from $\overline{F}(X^n)$ and $F_{\mathrm{BIC}}(X^n)$.

Figure 3 shows the results of $\hat{\nu}$, $\hat{\nu}_{\mathrm{BIC}}$ and $\hat{\nu}_{\mathrm{VB-BIC}}$ averaged over 100 draws of data sets for the different values of the hyperparameter ϕ_0. Another hyperparameter ξ_0 was set to $\xi_0 = 1$.

The results of $\hat{\nu}$ show a similar trend to ν and they are smaller than those of the BIC ($\hat{\nu}_{\mathrm{BIC}}$ and $\hat{\nu}_{\mathrm{VB-BIC}}$). This indicates the effect of the hyperparameter, which is suggested by the discussion in Sect. 8.3.

8 Discussion

In this paper, we obtained an asymptotic upper bound for the variational free energy of bipartite Bayesian networks with discrete hidden variables. In this section, we discuss some generalizations of the main results, the approximation accuracy of Variational Bayes and the effect of the hyperparameters.

8.1 Bounds for general Bayesian networks

In the previous sections, we considered any discrete random variables as observed nodes whose conditional probabilities are given by

$$c(x|b_z) = \prod_{j=1}^{N} b_{(j,x_j|z)},$$

with $M = \sum_{j=1}^{N}(Y_j - 1)$ parameters. Theorem 2 gives the upper bound for the variational free energy, which applies to any sets of training samples X^n of the discrete variable x. The

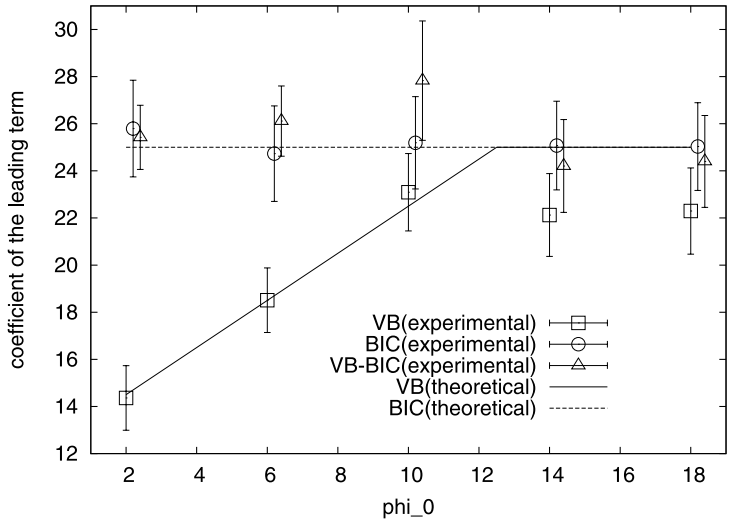

Fig. 3 The coefficients of the variational free energy for the hyperparameter ϕ_0. The open squares are the averages of $\hat{\nu}$ with the error bars showing 95% confidence intervals. The open circles and triangles are those of $\hat{\nu}_{\mathrm{BIC}}$ and $\hat{\nu}_{\mathrm{VB-BIC}}$ respectively. The *solid line* is the coefficient ν of the theoretical upper bound on the variational free energy, (32). The *dashed line* is the coefficient $d/2$ of the BIC where d is the number of parameters

above distribution of the observed variable x can also be generalized to the exponential-family distribution with M parameters by a similar argument as the one applied to the mixtures of exponential families (Watanabe and Watanabe 2005, 2006b). In this case, the observed variable x can also be continuous and the model (5) is given by

$$p(x|\omega) = \left\{ \prod_{k=1}^{K} \sum_{z_k=1}^{T_k} \right\} c(x|b_z) \prod_{k=1}^{K} a_{(k,z_k)},$$

where $c(x|b)$ is the probability density function of an exponential-family distribution with the parameters $b = (b^{(1)}, b^{(2)}, \ldots, b^{(M)})^T$,

$$c(x|b) = \exp(b \cdot h(x) + h_0(x) - g(b)).$$

Here $b \cdot h(x)$ is the inner product of the vector b and the vector-valued function $h(x) = (h_1(x), \ldots, h_M(x))^T$. The functions $h_0(x)$ and $g(b)$ are real-valued functions of the observed variable x and the parameters b, respectively.

8.2 Comparison to Bayes

Let us compare the variational free energy to the Bayesian free energy, namely the stochastic complexity of Bayesian networks and those of regular statistical models. The Bayesian stochastic complexities of regular models are also called the Bayesian information criterion (BIC) (Schwarz 1978).

For an arbitrary natural number n, the following inequality holds for the Bayesian stochastic complexity (Yamazaki and Watanabe 2003b; Rusakov and Geiger 2005),

$$E_{X^n}[F(X^n) - S(X^n)] \leq \mu \log n + O(1),$$

where

$$\mu = \frac{M}{2}\prod_{k=1}^{H} S_k - \frac{1}{2}\sum_{k=1}^{H} S_k + \frac{1}{2}H + \sum_{k=1}^{K} T_k - K. \tag{43}$$

These upper bounds are obtained under the conditions (A1), (A2), and $\phi_0 = 1$ in (17). Also the penalty term in the BIC is given by $\frac{d}{2}\log n$ where

$$d = \sum_{k=1}^{K}(T_k - 1) + M\prod_{k=1}^{K} T_k, \tag{44}$$

is the number of parameters. By putting $\phi_0 = 1$ in (30), from (43) and (44), we obtain

$$\nu = \mu < d/2.$$

This means the variational free energy is much smaller than the BIC, and is close to the Bayesian stochastic complexity. In other words, this implies the effectiveness of the variational Bayesian approach in terms of approximating the Bayesian posterior distributions and estimating the Bayesian stochastic complexities.

8.3 Effect of hyperparameters

Theorem 2 indicates how the hyperparameters influence the process of variational Bayesian learning. Consider the case discussed in Corollary 2 for example. The coefficient ν in (32) implies the states of hidden nodes that minimize the variational free energy function depending on the hyperparameter ϕ_0. More specifically, when $\phi_0 \leq \frac{1+S_2M}{2}$, all the redundant states of the first and second hidden nodes shrink and become inactive, that is, their probabilities approach 0. When $\frac{1+S_2M}{2} < \phi_0 \leq \frac{1+T_1M}{2}$, all the states of the first hidden node (with more possible states) become active while all the redundant states of the other hidden node are eliminated. When $\frac{1+T_1M}{2} < \phi_0$, all the states of both hidden nodes become active after the learning process.

In the general case considered in Theorem 2, the coefficient ν is identified by the minimum of a function of the numbers $u_1, u_2, \ldots, u_K$ of states. The minimum solution implies how many redundant states become active after the learning process according to the hyperparameter ϕ_0.

8.4 Utility of the bound and future work

Using the bound in the main theorem, one can experimentally investigate properties of the actual variational Bayesian algorithm which may converge to local minima of the variational free energy. Comparing the theoretical bound with experimental results, one can examine whether the algorithm converges to the optimal variational posterior.

Moreover, the theoretical bound would enable us to compare the accuracy of variational Bayesian approximation with other schemes such as the Laplace approximation or the MCMC method. In order to make such comparisons more accurately, one will need the

lower bound for the variational free energy as well as the upper bound. To obtain lower bounds, the identifiability of Bayesian networks should be taken into account (Whiley and Titterington 2002).

It is also important to assess the variational approximation in terms of the generalization error, or the accuracy of approximating the Bayesian predictive distributions in future studies.

9 Conclusion

In this paper, we obtained an upper bound for the variational free energy of Bayesian networks. The derived bound enabled us to assess the accuracy of the variational Bayesian approximation and the effect of the hyperparameters. It hence provides implications for the design of learning algorithms based on variational Bayesian approximation.

Acknowledgement This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for JSPS Fellows 0404637 and for Scientific Research 15500130.

References

Allen, T. V., & Greiner, R. (2000). Model selection criteria for learning belief nets: an empirical comparison. In *Proceedings of international conference on machine learning* (pp. 1047–1054).

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of uncertainty in artificial intelligence* (pp. 21–30). Stockholm, Sweden.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. Thesis, University College London.

Beal, M. J., & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian statistics* (Vol. 7). London: Oxford University Press.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, *303*, 799–805.

Hinton, G., & van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual ACM conference on computational learning theory* (pp. 5–13). New York: ACM Press.

Hosino, T., Watanabe, K., & Watanabe, S. (2005). Stochastic complexity of variational Bayesian hidden Markov models. In *Proceedings of international joint conference on neural networks* (Vol. 2, pp. 1114–1119).

Jensen, F. V. (2001). *Bayesian networks and decision graphs*. New York: Springer.

Jordan, M. I. (1999). *Learning in graphical models*. Cambridge: MIT Press.

Mackay, D. J. (1992). Bayesian interpolation. *Neural Computation*, *4*(2), 415–447.

Meltzer, T., Yanover, C., & Weiss, Y. (2005). Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *Proceedings of the tenth IEEE international conference on computer vision* (pp. 428–435).

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, *14*(3), 1080–1100.

Rusakov, D., & Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, *6*(1), 1–35.

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

Wang, B., & Titterington, D. M. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, *20*, 151–170.

Wang, B., & Titterington, D. M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Proceedings of the tenth international workshop on AISTATS* (pp. 373–380).

Wang, B., & Titterington, D. M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, *1*(3), 625–650.

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

Watanabe, K., & Watanabe, S. (2005). Stochastic complexity for mixture of exponential families in variational Bayes. In *Proceedings of international conference on algorithmic learning theory* (pp. 107–121). New York: Springer.

Watanabe, K., & Watanabe, S. (2006a). Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *Journal of Machine Learning Research*, *7*, 625–644.

Watanabe, K., & Watanabe, S. (2006b). Variational Bayesian stochastic complexity of mixture models. In *Advances in neural information processing systems* (Vol. 18, pp. 1465–1472). Cambridge: MIT Press.

Watanabe, K., Shiga, M., & Watanabe, S. (2006). Upper bounds for variational stochastic complexities of Bayesian networks. In *Proceedings of international conference on intelligent data engineering and automated learning* (pp. 139–146). New York: Springer.

Whiley, M., & Titterington, D. M. (2002). *Model identifiability in naive Bayesian networks* (Tech. Rep. 02-1). Department of Statistics, University of Glasgow.

Yamazaki, K., & Watanabe, S. (2003a). Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, *16*, 1023–1038.

Yamazaki, K., & Watanabe, S. (2003b). Stochastic complexity of Bayesian networks. In *Proceedings of uncertainty in artificial intelligence* (pp. 592–599). Acapulco, Mexico.