Journal of Machine Learning Research 7 (2006) 625–644 Submitted 1/05; Revised 9/05; Published 4/06

Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation

Kazuho Watanabe KAZUHO23@PI.TITECH.AC.JP
Department of Computational Intelligence and Systems Science
Tokyo Institute of Technology
MailBox R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan

Sumio Watanabe SWATANAB@PI.TITECH.AC.JP
P & I Lab.
Tokyo Institute of Technology
MailBox R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan

Editor: Tommi Jaakkola

Abstract

Bayesian learning has been widely used and proved to be effective in many data modeling problems. However, computations involved in it require huge costs and generally cannot be performed exactly. The variational Bayesian approach, proposed as an approximation of Bayesian learning, has provided computational tractability and good generalization performance in many applications.

The properties and capabilities of variational Bayesian learning itself have not been clarified yet. It is still unknown how good approximation the variational Bayesian approach can achieve. In this paper, we discuss variational Bayesian learning of Gaussian mixture models and derive upper and lower bounds of variational stochastic complexities. The variational stochastic complexity, which corresponds to the minimum variational free energy and a lower bound of the Bayesian evidence, not only becomes important in addressing the model selection problem, but also enables us to discuss the accuracy of the variational Bayesian approach as an approximation of true Bayesian learning.

Keywords: Gaussian mixture model, variational Bayesian learning, stochastic complexity

1. Introduction

A Gaussian mixture model is a learning machine which estimates the target probability density by the sum of normal distributions. This learning machine is widely used especially in statistical pattern recognition and data clustering. In spite of wide range of its applications, its properties have not yet been made clear enough. This is because the Gaussian mixture model is a non-regular statistical model. A statistical model is regular if and only if a set of conditions (referred to as "regularity conditions") that ensure the asymptotic normality of the maximum likelihood estimator is satisfied. The regularity conditions are not satisfied for mixture models because the parameters are not identifiable, in other words, the mapping from parameters to probability distributions is not one-to-one. Other than mixture models, statistical models with hidden variables such as hidden Markov models and Bayesian networks fall into the class of non-regular models.

Recently, a lot of attentions has been paid to the non-regular models. In Bayesian learning, mathematical foundation for analyzing non-regular models was established with an algebraic geometrical method (Watanabe, 2001). The Bayesian stochastic complexities or the marginal like-

©2006 Kazuho Watanabe and Sumio Watanabe.

lihoods of several non-regular models have been clarified in some recent studies (Yamazaki and Watanabe, 2003a,b). The Bayesian framework provides better generalization performance in non-regular models than the maximum likelihood (ML) method that tends to overfit the data.

In the Bayesian framework, rather than learning a single model, one computes the distribution over all possible parameter values and considers an ensemble with respect to the posterior distribution. However, computing the Bayesian posterior can seldom be performed exactly and requires some approximations. Well-known approximate methods include Markov chain Monte Carlo (MCMC) methods and the Laplace approximation. The former attempts to find the exact posterior distribution but typically requires huge computational resources. The latter approximates the posterior distribution by a Gaussian distribution, which can be insufficient for models containing hidden variables.

The variational Bayesian (VB) framework was proposed as another approximation for computations in the models with hidden variables (Attias, 1999; Ghahramani and Beal, 2000). This framework provides computationally tractable posterior distributions over the hidden variables and the parameters with an iterative algorithm. The variational Bayesian framework has been applied to various real-world data modeling problems and empirically proved to be both computational tractable and generalize well.

The properties of variational Bayesian learning remain unclear from a theoretical stand point. Although the variational Bayesian framework is an approximation, questions like how accurately it approximates the true distribution have yet to be answered.

In this paper, we focus on variational Bayesian learning of Gaussian mixture models. As the main contribution, we derive asymptotic upper and lower bounds on the variational stochastic complexity. It is shown that the variational stochastic complexity is smaller than in regular statistical models, so the advantage of Bayesian learning still remains in variational Bayesian learning. The variational stochastic complexity, which corresponds to the minimum variational free energy and a lower bound of the Bayesian evidence, is an important quantity for model selection. Giving the asymptotic bounds on it also contributes to the following two issues. One is the accuracy of variational Bayesian learning as an approximation method since the variational stochastic complexity 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, the derived bounds indicate how the hyperparameters influence the learning process. Our results indicate how to determine the hyperparameter values before the learning process.

We consider the case in which the true distribution is contained in the learned model, in other words, the model has redundant components to attain the true distribution. Analyzing the variational stochastic complexity in this case is most valuable for comparing variational Bayesian learning with true Bayesian learning. This is because the advantage of Bayesian learning is typical in this case (Watanabe, 2001). Furthermore, this analysis is necessary and essential for addressing the model selection and hypothesis testing problems.

This paper is organized as follows. In Section 2, the Gaussian mixture model is briefly introduced. In Section 3, we describe Bayesian learning. In Section 4, the variational Bayesian framework is outlined and the variational stochastic complexity is defined. In Section 5, we state the main theorem of this paper. The main theorem is proved in Section 6. In Section 7, we experimentally examine the quality of the bounds given in the main theorem. Discussion and conclusions follow in Section 8 and Section 9.

2. Gaussian Mixture Models

Denote by $g(x|\mu,\Sigma)$ a density function of an M-dimensional normal distribution whose mean is $\mu \in R^M$ and variance-covariance matrix is $\Sigma \in R^{M\times M}$. A Gaussian mixture model $p(x|\theta)$ of an M-dimensional input $x \in R^M$ with a parameter vector θ is defined by

$$p(x|\theta) = \sum_{k=1}^{K} a_k g(x|\mu_k, \Sigma_k),$$

where integer K is the number of components and $\{a_k | a_k \geq 0, \sum_{k=1}^{K} a_k = 1\}$ is the set of mixing proportions. The parameter θ of the model is $\theta = \{a_k, \mu_k, \Sigma_k\}_{k=1}^{K}$.

In some applications, the parameter is restricted to the means of each component and it is assumed that there is no correlation between each input dimension. In this case, the model is written by

$$p(x|\theta) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi\sigma_k^2}^M} \exp(-\frac{\|x-\mu_k\|^2}{2\sigma_k^2}), \tag{1}$$

where $\sigma_k > 0$ is a constant.

In this paper, we consider this type eq.(1) of Gaussian mixture models in the variational Bayesian framework and show upper and lower bounds of the variational stochastic complexity in Theorem 3.

The Gaussian mixture model can be rewritten as follows using a hidden variable $y = (y^1, \cdots, y^K) \in \{(1,0,\cdots,0),(0,1,\cdots,0),\cdots,(0,0,\cdots,1)\}$,

$$p(x,y|\theta) = \prod_{k=1}^{K} \Big[\frac{a_k}{\sqrt{2\pi\sigma_k^2}^M} \exp\{-\frac{\|x-\mu_k\|^2}{2\sigma_k^2}\} \Big]^{y^k}.$$

The hidden variable y is not observed and is representing the component from which the datum x is generated. If the datum x is from the kth component, then $y^k = 1$, if otherwise, $y^k = 0$. And

$$\sum_{y} p(x,y|\theta) = p(x|\theta)$$

holds where the sum over y ranges over all possible values of the hidden variable.

The Gaussian mixture model is a non-regular statistical model, since the parameters are non-identifiable. More specifically, if the true distribution can be realized by a model with the smaller number of components, the true parameter is not a point but an analytic set with singularities. If the parameters are non-identifiable, the usual asymptotic theory of regular statistical models cannot be applied. Some studies have revealed that Gaussian mixture models have quite different properties from those of regular statistical models. In particular, the Gaussian mixture model given by eq.(1) has been studied as a prototype of non-regular models in the case of the maximum likelihood estimation(Hartigan, 1985; Dacunha-Castelle and Gassiat, 1997).

3. Bayesian Learning

Suppose n training samples $X^n = \{x_1, \cdots, x_n\}$ are independently and identically taken from the true distribution $p_0(x)$. In Bayesian learning of a model $p(x|\theta)$ whose parameter is θ, first, the prior

distribution $\varphi(\theta)$ on the parameter θ is set. Then the posterior distribution $p(\theta|X^n)$ is computed from the given data set and the prior by

$$p(\theta|X^n) = \frac{1}{Z(X^n)}\varphi(\theta)\prod_{i=1}^{n} p(x_i|\theta),$$

where $Z(X^n)$ is the normalization constant that is also known as the marginal likelihood or the Bayesian evidence of the data set X^n (Mackay, 1992).

The Bayesian predictive distribution $p(x|X^n)$ is given by averaging the model over the posterior distribution as follows,

$$p(x|X^n) = \int p(x|\theta)p(\theta|X^n)d\theta,$$

and its generalization error can be measured by the Kullback information from the true distribution,[1]

$$K(p_0(x)||p(x|X^n)) = \int p_0(x)\log\frac{p_0(x)}{p(x|X^n)}dx.$$

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

$$F(X^n) = -\log Z(X^n), \tag{2}$$

which is also called the free energy and is important in most data modeling problems. Practically, it is used as a criterion by which the learning model is selected and the hyperparameters in the prior are optimized (Akaike, 1980; Schwarz, 1978).

The Bayesian posterior can be rewritten as

$$p(\theta|X^n) = \frac{1}{Z_0(X^n)}\exp(-nH_n(\theta))\varphi(\theta),$$

where $H_n(\theta)$ is the empirical Kullback information,

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p_0(x_i)}{p(x_i|\theta)}, \tag{3}$$

and $Z_0(X^n)$ is the normalization constant. Let

$$S(X^n) = -\sum_{i=1}^{n}\log p_0(x),$$

and define the normalized Bayesian stochastic complexity $F_0(X^n)$ by

$$\begin{aligned} F_0(X^n) &= -\log Z_0(X^n) \\ &= F(X^n) - S(X^n). \end{aligned} \tag{4}$$

1. Throughout this paper, we use the notation $K(q(x)||p(x))$ for the Kullback information from a distribution $q(x)$ to a distribution $p(x)$, that is,

$$K(q(x)||p(x)) = \int q(x)\log\frac{q(x)}{p(x)}dx.$$

It is noted that the empirical entropy $S(X^n)$ does not depend on the model $p(x|\theta)$. Therefore minimization of $F(X^n)$ is equivalent to that of $F_0(X^n)$.

Let $E_{X^n}[\cdot]$ denote the expectation over all sets of training samples. Then it follows from eq.(4) that

$$E_{X^n}[F(X^n) - F_0(X^n)] = nS,$$

where $S = -\int p_0(x)\log p_0(x)dx$ is the entropy. There is the following relationship between the average Bayesian stochastic complexity and the average generalization error(Levin et al., 1990),

$$\begin{aligned} E_{X^n}[K(p_0(x)||p(x|X^n))] &= E_{X^{n+1}}[F(X^{n+1})] - E_{X^n}[F(X^n)] - S \\ &= E_{X^{n+1}}[F_0(X^{n+1})] - E_{X^n}[F_0(X^n)]. \end{aligned} \tag{5}$$

Recently, in Bayesian learning, an advanced mathematical method for analyzing non-regular models was established(Watanabe, 2001), which enabled us to clarify the asymptotic behavior of the Bayesian stochastic complexity of non-regular models. More specifically, by using concepts in algebraic analysis, it was proved that the average normalized Bayesian stochastic complexity defined by $E_{X^n}[F_0(X^n)]$ has the following asymptotic form,

$$E_{X^n}[F_0(X^n)] \simeq \lambda\log n - (m-1)\log\log n + O(1), \tag{6}$$

where λ and m are the rational number and the natural number respectively which are determined by the singularities of the true parameter. In regular statistical models, 2λ is equal to the number of parameters and $m = 1$, whereas in non-regular models such as Gaussian mixture models, 2λ is not larger than the number of parameters and $m \geq 1$. This means non-regular models have an advantage in Bayesian learning. From eq.(5), if the asymptotic form of the average normalized Bayesian stochastic complexity is given by eq.(6), the average generalization error is given by

$$E_{X^n}[K(p_0(x)||p(x|X^n))] \simeq \frac{\lambda}{n} + o(\frac{1}{n}). \tag{7}$$

Since the coefficient λ is proportional to the average generalization error, Bayesian learning is more suitable for non-regular models than the maximum likelihood (ML) method.

However, in order to carry out Bayesian learning practically, one computes the Bayesian stochastic complexity or the predictive distribution by integrating over the posterior distribution, which typically cannot be performed analytically.

Hence, approximations must be made. The Laplace approximation is a well-known and simple method that approximates the posterior distribution by a Gaussian distribution. This approach gives reasonable approximation in the case of regular statistical models whose posteriors converge to normal distributions as the sample size n tends to infinity. In contrast, posterior distributions of non-regular models do not converge to normal distributions in general, even as n tends to infinity. Therefore, the Laplace approximation can be insufficient for non-regular models. Markov chain Monte Carlo (MCMC) method can provide a better approximation. It attempts to sample from the exact posterior distribution but typically requires vast computational resources.

As another approximation, the variational Bayesian framework was proposed (Attias, 1999; Beal, 2003; Ghahramani and Beal, 2000).

4. Variational Bayesian Learning

In this section, we outline the variational Bayesian framework and define the variational stochastic complexity.

4.1 The Variational Bayesian Framework

Using the complete likelihood of the data $\{X^n, Y^n\}$, with the corresponding hidden variables $Y^n = \{y_1, \cdots, y_n\}$, we can rewrite the Bayesian stochastic complexity eq.(2) as

$$\begin{aligned} F(X^n) &= -\log \int \sum_{Y^n} \varphi(\theta) \prod_{i=1}^{n} p(x_i, y_i|\theta) d\theta \\ &= -\log \int \sum_{Y^n} p(X^n, Y^n, \theta) d\theta, \end{aligned}$$

where the sum over Y^n ranges over all possible values of all hidden variables.

The variational Bayesian framework starts by upper bounding the Bayesian stochastic complexity. For an arbitrary conditional distribution $q(Y^n, \theta|X^n)$ on the hidden variables and the parameters, the Bayesian stochastic complexity can be upper bounded by applying Jensen's inequality,

$$\begin{aligned} F(X^n) &\leq \sum_{Y^n} \int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} d\theta \\ &\equiv \overline{F}[q]. \end{aligned}$$

This inequality becomes an equality if and only if $q(Y^n, \theta|X^n) = p(Y^n, \theta|X^n)$, that is, $q(Y^n, \theta|X^n)$ equals the Bayesian posterior distribution. This means that the smaller the functional $\overline{F}[q]$ is, the closer the distribution $q(Y^n, \theta|X^n)$ is to the true Bayesian posterior distribution. The functional $\overline{F}[q]$ is called the variational free energy.

The variational Bayesian approach makes an approximation to ensure a computationally tractable posterior. More specifically, assuming the parameters and the hidden variables are conditionally independent of each other, the variational Bayesian approach restricts the set of $q(Y^n, \theta|X^n)$ to distributions that have the form

$$q(Y^n, \theta|X^n) = Q(Y^n|X^n) r(\theta|X^n), \tag{8}$$

where $Q(Y^n|X^n)$ and $r(\theta|X^n)$ are probability distributions over the hidden variables and the parameters respectively. The distribution $q(Y^n, \theta|X^n)$ that minimizes the functional $\overline{F}[q]$ is termed the optimal variational posterior and generally differs from the true Bayesian posterior.

Minimization of the functional $\overline{F}[q]$ with respect to the distributions $Q(Y^n|X^n)$ and $r(\theta|X^n)$ can be performed by using variational methods. Solving the minimization problem under the constraints $\int r(\theta|X^n) d\theta = 1$ and $\sum_{Y^n} Q(Y^n|X^n) = 1$ gives the following theorem. The proof is well-known(Beal, 2003; Sato, 2001), thus it is omitted in this paper.

Theorem 1 *If the functional $\overline{F}[q]$ is minimized under the constraint eq.(8) then the variational posteriors, $r(\theta|X^n)$ and $Q(Y^n|X^n)$, satisfy*

$$r(\theta|X^n) = \frac{1}{C_r} \varphi(\theta) \exp \langle \log p(X^n, Y^n|\theta) \rangle_{Q(Y^n|X^n)}, \tag{9}$$

and

$$Q(Y^n|X^n) = \frac{1}{C_Q} \exp \langle \log p(X^n, Y^n|\theta) \rangle_{r(\theta|X^n)}, \tag{10}$$

where C_r and C_Q are the normalization constants.[2]

2. Hereafter for an arbitrary distribution $p(x)$, we use the notation $\langle \cdot \rangle_{p(x)}$ for the expected value over $p(x)$.

Note that eq.(9) and eq.(10) give only the necessary condition for $r(\theta|X^n)$ and $Q(Y^n|X^n)$ minimize the functional $\overline{F}[q]$. The variational posteriors that satisfy eq.(9) and eq.(10) are searched by an iterative algorithm. It is known that this algorithm is a natural gradient method when the model is in the general exponential family of models with hidden variables (Sato, 2001).

4.2 Stochastic Complexity of Variational Bayes

We define the variational stochastic complexity $\overline{F}(X^n)$ by the minimum value of the functional $\overline{F}[q]$ attained by the above optimal variational posteriors, that is ,

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

The variational stochastic complexity $\overline{F}(X^n)$ gives an estimate (upper bound) for the true Bayesian stochastic complexity $F(X^n)$, which is the minus log evidence. Therefore, $\overline{F}(X^n)$ is used for the model selection in variational Bayesian learning(Beal, 2003). Moreover, the difference between $\overline{F}(X^n)$ and the Bayesian stochastic complexity $F(X^n)$ is the Kullback information from the optimal variational posterior to the true posterior. That is

$$\overline{F}(X^n) - F(X^n) = \min_{r,Q} K(q(Y^n,\theta|X^n)||p(Y^n,\theta|X^n)).$$

Hence, comparison between $\overline{F}(X^n)$ and $F(X^n)$ shows the accuracy of the variational Bayesian approach as an approximation of true Bayesian learning.

We define the normalized variational stochastic complexity $\overline{F}_0(X^n)$ by

$$\overline{F}_0(X^n) = \overline{F}(X^n) - S(X^n). \tag{11}$$

From Theorem 1, the following lemma is obtained. The proof is given in Appendix.

Lemma 2

$$\overline{F}_0(X^n) = \min_{r(\theta|X^n)} \{K(r(\theta|X^n)||\varphi(\theta)) - (\log C_Q + S(X^n))\}, \tag{12}$$

where

$$C_Q = \sum_{Y^n} \exp \langle \log p(X^n,Y^n|\theta) \rangle_{r(\theta|X^n)}.$$

The variational posteriors $r(\theta|X^n)$ and $Q(Y^n|X^n)$ that satisfy eq.(9) and eq.(10) are parameterized by the variational parameter $\overline{\theta}$ defined by

$$\overline{\theta} = \langle \theta \rangle_{r(\theta|X^n)},$$

if the model $p(x,y|\theta)$ is included in the exponential family(Beal, 2003; Ghahramani and Beal, 2000). Then it is noted that C_Q in eq.(12) is also parameterized by $\overline{\theta}$. Therefore, henceforth we denote $r(\theta|X^n)$ and C_Q as $r(\theta|\overline{\theta})$ and $C_Q(\overline{\theta})$ when they are regarded as functions of the variational parameter $\overline{\theta}$.

We define the variational estimator $\overline{\theta}_{vb}$ of θ by the variational parameter $\overline{\theta}$ that attains the minimum value of the normalized variational stochastic complexity $\overline{F}_0(X^n)$. By this definition, Lemma 2 claims that

$$\overline{\theta}_{vb} = \operatorname*{argmin}_{\overline{\theta}} \{K(r(\theta|\overline{\theta})||\varphi(\theta)) - (\log C_Q(\overline{\theta}) + S(X^n))\}. \tag{13}$$

In variational Bayesian learning, the variational parameter $\overline{\theta}$ is updated iteratively to find the optimal solution $\overline{\theta}_{vb}$. Therefore, our aim is to evaluate the minimum value of the right hand side of eq.(13) as a function of the variational parameter $\overline{\theta}$.

5. Main Results

In this section, we describe two conditions and give the upper and lower bounds of the normalized variational stochastic complexity in Theorem 3.

We assume the following conditions.

(i) The true distribution $p_0(x)$ is an M-dimensional Gaussian mixture model $p(x|\theta_0)$ which has K_0 components and the parameter $\theta_0 = \{a_k^*, \mu_k^*\}_{k=1}^{K_0}$,

$$p(x|\theta_0) = \sum_{k=1}^{K_0} \frac{a_k^*}{\sqrt{2\pi}^M} \exp(-\frac{\|x-\mu_k^*\|^2}{2}),$$

where $x, \mu_k^* \in R^M$. And suppose that the true distribution can be realized by the model, that is, the model $p(x|\theta)$ has K components,

$$p(x|\theta) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi}^M} \exp(-\frac{\|x-\mu_k\|^2}{2}), \tag{14}$$

and $K \geq K_0$ holds.

(ii) The prior of the parameters is the product of the following two distributions on $\mathbf{a} = \{a_k\}_{k=1}^K$ and $\mu = \{\mu_k\}_{k=1}^K$

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^{K} a_k^{\phi_0 - 1}, \tag{15}$$

$$\varphi(\mu) = \prod_{k=1}^{K} \sqrt{\frac{\xi_0}{2\pi}}^M \exp(-\frac{\xi_0 \|\mu_k - \nu_0\|^2}{2}), \tag{16}$$

where $\xi_0 > 0$, $\nu_0 \in R^M$ and $\phi_0 > 0$ are constants called hyperparameters. These are Dirichlet and normal distributions respectively. They are the conjugate prior distributions and are often used in variational Bayesian learning of Gaussian mixture models.

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

Theorem 3 (Main Result) *Assume the conditions (i) and (ii). Then the normalized variational stochastic complexity $\overline{F}_0(X^n)$ defined by eq.(11) satisfies*

$$\underline{\lambda} \log n + nH_n(\overline{\theta}_{vb}) + C_1 \leq \overline{F}_0(X^n) \leq \overline{\lambda} \log n + C_2, \tag{17}$$

with probability 1 for an arbitrary natural number n where C_1, C_2 are constants independent of n and the coefficients $\underline{\lambda}$, $\overline{\lambda}$ are given by

$$\underline{\lambda} = \begin{cases} (K-1)\phi_0 + \frac{M}{2} & (\phi_0 \leq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}), \end{cases}$$

$$\overline{\lambda} = \begin{cases} (K-K_0)\phi_0 + \frac{MK_0+K_0-1}{2} & (\phi_0 \leq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}). \end{cases} \tag{18}$$

Taking expectation over all sets of training samples, we obtain the following corollary.

Corollary 4 *Assume the conditions (i) and (ii). Then the average of the normalized variational stochastic complexity $\overline{F}_0(X^n)$ satisfies*

$$\underline{\lambda}\log n + E_{X^n}[nH_n(\overline{\theta}_{vb})] + C_1 \leq E_{X^n}[\overline{F}_0(X^n)] \leq \overline{\lambda}\log n + C_2.$$

Remark. The following bounds for the variational stochastic complexity $\overline{F}(X^n) = \overline{F}_0(X^n) + S(X^n)$ are immediately obtained from Theorem 3 and Corollary 4,

$$S(X^n) + \underline{\lambda}\log n + nH_n(\overline{\theta}_{vb}) + C_1 \leq \overline{F}(X^n) \leq S(X^n) + \overline{\lambda}\log n + C_2,$$

and

$$nS + \underline{\lambda}\log n + E_{X^n}[nH_n(\overline{\theta}_{vb})] + C_1 \leq E_{X^n}[\overline{F}(X^n)] \leq nS + \overline{\lambda}\log n + C_2,$$

where $S(X^n) = -\sum_{i=1}^n \log p(x_i|\theta_0)$ is the empirical entropy and $S = -\int p(x|\theta_0)\log p(x|\theta_0)dx$ is the entropy.

Since the dimension of the parameter θ is $MK + K - 1$, the penalty term in the Bayesian information criterion (BIC) (Schwarz, 1978) is given by $\lambda_{\rm BIC}\log n$ where

$$\lambda_{\rm BIC} = \frac{MK + K - 1}{2}. \tag{19}$$

Note that, unlike for regular statistical models, the advantage of Bayesian learning for non-regular models is demonstrated by the asymptotic analysis as seen in eq.(6) and eq.(7). Theorem 3 claims that the coefficient $\overline{\lambda}$ of $\log n$ is smaller than $\lambda_{\rm BIC}$ when $\phi_0 \leq (M+1)/2$. This means the normalized variational stochastic complexity $\overline{F}_0(X^n)$ becomes smaller than the BIC and implies that the advantage of non-regular models in Bayesian learning still remains in variational Bayesian learning.

Theorem 3 also shows how the hyperparameters affect the learning process and implies that the hyperparameter ϕ_0 is the only hyperparameter that the leading term of the normalized variational stochastic complexity $\overline{F}_0(X^n)$ depends on. The effects of the hyperparameters are discussed in Section 8.

In the condition (i), we assume that the true distribution is contained in the learner model ($K_0 \leq K$). This assumption is necessary for assessing model selection or hypothesis testing methods and for developing a new method for these tasks. In real-world applications, the true distribution might not be represented by any model with finite components. Also if the model is complex enough to almost contain the true distribution with finite training samples, we need to consider the case when the model is redundant.

6. Proof of Theorem 3

In this section, we prove Theorem 3. First of all, we derive the variational posterior $r(\theta|X^n)$, $Q(Y^n|X^n)$ and the variational parameter $\overline{\theta}$ for the Gaussian mixture model given by eq.(14).

6.1 Variational Posterior for Gaussian Mixture Model

For the complete-data set $\{X^n, Y^n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, let

$$\overline{y}_i^k = \langle y_i^k \rangle_{Q(Y^n|X^n)}, \quad n_k = \sum_{i=1}^n \overline{y}_i^k \quad \text{and} \quad \nu_k = \frac{1}{n_k}\sum_{i=1}^n \overline{y}_i^k x_i,$$

where $y_i^k = 1$ if ith datum x_i is from the kth component, if otherwise, $y_i^k = 0$. The variable n_k is the expected number of times data come from the kth component and ν_k is the mean of them. Note that the variables n_k and ν_k satisfy the constraints $\sum_{k=1}^K n_k = n$ and $\sum_{k=1}^K n_k \nu_k = \sum_{i=1}^n x_i$. From eq.(9) and the respective prior eq.(15) and eq.(16), the variational posterior $r(\theta|X^n) = r(\mathbf{a}|X^n) r(\mu|X^n)$ is obtained as the product of the following two distributions,

$$r(\mathbf{a}|X^n) = \frac{\Gamma(n+K\phi_0)}{\prod_{k=1}^K \Gamma(\bar{a}_k(n+K\phi_0))} \prod_{k=1}^K a_k^{\bar{a}_k(n+K\phi_0)-1},$$

and

$$r(\mu|X^n) = \prod_{k=1}^K \frac{1}{\sqrt{2\pi\bar{\sigma}_k^2}^M} \exp\left(\frac{-\|\mu_k - \bar{\mu}_k\|^2}{2\bar{\sigma}_k^2}\right),$$

where

$$\bar{a}_k = \frac{n_k+\phi_0}{n+K\phi_0}, \quad \bar{\sigma}_k^2 = \frac{1}{n_k+\xi_0}, \quad \text{and} \quad \bar{\mu}_k = \frac{n_k\nu_k + \xi_0\nu_0}{n_k+\xi_0}.$$

From eq.(10), the variational posterior $Q(Y^n|X^n)$ is given by

$$Q(Y^n|X^n) = \frac{1}{C_Q} \prod_{i=1}^n \exp\left[y_i^k \{ \Psi(n_k+\phi_0) - \Psi(n+K\phi_0) - \frac{\|x_i - \bar{\mu}_k\|^2}{2} - \frac{M}{2}\left(\log 2\pi + \frac{1}{n_k+\xi_0}\right)\}\right],$$

where $\Psi(x) = \Gamma'(x)/\Gamma(x)$ is the di-gamma(psi) function and we used

$$\langle \log a_k \rangle_{r(\mathbf{a}|X^n)} = \Psi(n_k+\phi_0) - \Psi(n+K\phi_0).$$

The variational parameter $\bar{\theta}$ is given by $\bar{\theta} = \langle\theta\rangle_{r(\theta|X^n)} = \{\bar{a}_k, \bar{\mu}_k\}_{k=1}^K$. It is noted that $r(\theta|X^n)$ and $Q(Y^n|X^n)$ are parameterized by $\bar{\theta}$ since n_k can be replaced by using $\bar{a}_k = \frac{n_k+\phi_0}{n+K\phi_0}$. Henceforth, we denote $r(\theta|X^n)$ and C_Q as $r(\theta|\bar{\theta})$ and $C_Q(\bar{\theta})$.

6.2 Lemmas

Before proving Theorem 3, we show two lemmas where the two terms $K(r(\theta|\bar{\theta})||\varphi(\theta))$ and $(\log C_Q(\bar{\theta}) + S(X^n))$ in Lemma 2 are respectively evaluated. In the proofs (put in Appendix) of the two lemmas, we use inequalities on the di-gamma function $\Psi(x)$ and the log-gamma function $\log\Gamma(x)$, for $x > 0$ (Alzer, 1997),

$$\frac{1}{2x} < \log x - \Psi(x) < \frac{1}{x}, \tag{20}$$

and

$$0 \le \log\Gamma(x) - \{(x-\frac{1}{2})\log x - x + \frac{1}{2}\log 2\pi\} \le \frac{1}{12x}. \tag{21}$$

The inequalities (20) ensure that substituting $\log x$ for $\Psi(x)$ only contributes additive constant terms to the normalized variational stochastic complexity. The substitution for $\log\Gamma(x)$ is given by eq.(21) as well.

Lemma 5

$$\left| K(r(\theta|\bar{\theta})||\varphi(\theta)) - \{G(\bar{\mathbf{a}}) + \frac{\xi_0}{2}\sum_{k=1}^K \|\bar{\mu}_k - \nu_0\|^2\}\right| \le C,$$

holds where C is a constant and the function $G(\bar{\mathbf{a}})$ *of* $\bar{\mathbf{a}} = \{\bar{a}_k\}_{k=1}^K$ *is defined by*

$$G(\bar{\mathbf{a}}) = \frac{MK+K-1}{2}\log n + \{\frac{M}{2} - (\phi_0 - \frac{1}{2})\}\sum_{k=1}^{K}\log \bar{a}_k.$$

Lemma 6

$$\log C_Q(\bar{\theta}) = \sum_{i=1}^{n}\log\Big[\sum_{k=1}^{K}\frac{1}{\sqrt{2\pi}^M}\exp\{\Psi(n_k+\phi_0) - \Psi(n+K\phi_0) - \frac{\|x_i-\bar{\mu}_k\|^2}{2} - \frac{M}{2}\frac{1}{n_k+\xi_0}\}\Big], \quad (22)$$

and

$$nH_n(\bar{\theta}) - \frac{n}{n+K\phi_0} \le -(\log C_Q(\bar{\theta}) + S(X^n)) \le n\overline{H}_n(\bar{\theta}) - \frac{n}{2(n+K\phi_0)}, \quad (23)$$

where $H_n(\bar{\theta})$ *is given by eq.(3) and* $\overline{H}_n(\bar{\theta})$ *is defined by*

$$\overline{H}_n(\bar{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(x_i|\theta_0)}{\sum_{k=1}^{K}\frac{\bar{a}_k}{\sqrt{2\pi}^M}\exp\{-\frac{\|x_i-\bar{\mu}_k\|^2}{2} - \frac{M+2}{2(n_k+\min\{\phi_0,\xi_0\})}\}}.$$

6.3 Upper and Lower Bounds

Now from the above lemmas, we prove Theorem 3 by showing the upper bound and the lower bound respectively.

(Proof of Theorem 3)

Proof First we show the upper bound in eq.(17).

From Lemma 2, Lemma 5 and Lemma 6, it follows that

$$\overline{F}_0(X^n) \le \min_{\bar{\theta}} T_n(\bar{\theta}) + C, \quad (24)$$

where

$$T_n(\bar{\theta}) = G(\bar{\mathbf{a}}) + \frac{\xi_0}{2}\sum_{k=1}^{K}\|\bar{\mu}_k - \nu_0\|^2 + n\overline{H}_n(\bar{\theta}).$$

From eq.(24), it is noted that the function values of $T_n(\bar{\theta})$ at specific points of the variational parameter $\bar{\theta}$ give the upper bounds of the normalized variational stochastic complexity $\overline{F}_0(X^n)$. Hence, let us consider following two cases.

(I) :

$$\bar{a}_k = a_k^* \ \ (1 \le k \le K_0 - 1), \ \ \bar{a}_k = a_{K_0}^*/(K - K_0 + 1) \ \ (K_0 \le k \le K),$$

$$\bar{\mu}_k = \mu_k^* \ \ (1 \le k \le K_0 - 1), \ \ \bar{\mu}_k = \mu_{K_0}^* \ \ (K_0 \le k \le K),$$

then $n\overline{H}_n(\bar{\theta}) < \frac{K-K_0+1}{\min_k\{a_k^*\}}$ holds and

$$T_n(\bar{\theta}) < \frac{MK+K-1}{2}\log n + C' + O(\frac{1}{n}),$$

where C' is a constant.

(II) :

$$\overline{a}_k = a_k^* \frac{n+K_0\phi_0}{n+K\phi_0} \quad (1 \leq k \leq K_0), \quad \overline{a}_k = \frac{\phi_0}{n+K\phi_0} \quad (K_0+1 \leq k \leq K),$$

$$\overline{\mu}_k = \mu_k^* \quad (1 \leq k \leq K_0), \quad \overline{\mu}_k = \nu_0 \quad (K_0+1 \leq k \leq K),$$

then $n\overline{H}_n(\overline{\theta}) < (K-K_0)\phi_0 + \frac{1}{\min_k\{a_k^*\}} + O(1/n)$ holds and

$$T_n(\overline{\theta}) < \{(K-K_0)\phi_0 + \frac{MK_0+K_0-1}{2}\}\log n + C'' + O(\frac{1}{n}),$$

where C'' is a constant.

From eq.(24), we obtain the upper bound in eq.(17).

Next we show the lower bound in eq.(17). It follows from Lemma 2, Lemma 5 and Lemma 6,

$$\overline{F}_0(X^n) \geq \min_{\overline{\mathbf{a}}}\{G(\overline{\mathbf{a}})\} + nH_n(\overline{\theta}_{vb}) - C - 1. \tag{25}$$

If $\phi_0 > \frac{M+1}{2}$, then

$$G(\overline{\mathbf{a}}) \geq \frac{MK+K-1}{2}\log n - (\frac{M+1}{2} - \phi_0)K\log K, \tag{26}$$

since Jensen's inequality yields that $\sum_{k=1}^K \log \overline{a}_k \leq K\log(\frac{1}{K}\sum_{k=1}^K \overline{a}_k) = K\log(\frac{1}{K})$.

If $\phi_0 \leq \frac{M+1}{2}$, then

$$G(\overline{\mathbf{a}}) \geq \{(K-1)\phi_0 + \frac{M}{2}\}\log n + (\frac{M+1}{2} - \phi_0)(K-1)\log\phi_0 + O(\frac{1}{n}), \tag{27}$$

since $\overline{a}_k \geq \frac{\phi_0}{n+K\phi_0}$ holds for every k and the constraint $\sum_{k=1}^K \overline{a}_k = 1$ ensures that $|\log \overline{a}_k|$ is bounded by a constant independent of n for at least one index k. From eqs.(25),(26) and (27), we obtain the lower bound in eq.(17).

■

7. Experiments

In order to examine the quality of the theoretical bounds given in Theorem 3, we conducted experiments of variational Bayesian learning for Gaussian mixture models using $M = 1$ and $M = 10$ dimensional synthetic data. A set of models with different number of components ($K = 1,2,3,4,5$) was prepared. We applied the variational Bayesian algorithm to each model using the data set generated from the true distribution with $K_0 = 2$ components. The true distribution was set to a Gaussian mixture model with the parameter $a_1^* = a_2^* = 1/2$, $\mu_1^* = -2/\sqrt{M}\cdot\mathbf{1}$ and $\mu_2^* = 2/\sqrt{M}\cdot\mathbf{1}$ where $\mathbf{1}$ is the M-dimensional vector whose all entries are 1. The hyperparameters were set at $\phi_0 = 1.0$, $\nu_0 = 0$ and $\xi_0 = 1.0$. In order to achieve the minimum in eq.(13), the initial value of the variational parameter $\overline{\theta}$ was set around the true parameter, that is, around $\overline{a}_1 = \overline{a}_2 = 1/2$, $\overline{a}_k = 0$ $(k \geq 3)$, $\overline{\mu}_1 = \mu_1^*$, $\overline{\mu}_2 = \mu_2^*$ and $\overline{\mu}_k = 0$ $(k \geq 3)$. Two sample sets with the size $n = 1000$ and $n = 100$ were prepared. For each data set, the normalized variational stochastic complexity (the inside of the braces in eq.(13)) was

calculated when the variational Bayesian algorithm converged. Denoting the results for respective data sets by $\overline{F}_0(X^{1000})$ and $\overline{F}_0(X^{100})$, we calculated

$$\lambda_{\rm VB} = (\overline{F}_0(X^{1000}) - \overline{F}_0(X^{100}))/\log 10 \tag{28}$$

to estimate the coefficient of the leading term of the normalized variational stochastic complexity $\overline{F}_0(X^n)$. We averaged the values of $\lambda_{\rm VB}$ over 100 draws of sample sets. The results of the averages of $\lambda_{\rm VB}$ and the coefficient $\overline{\lambda}$ given by eq.(18) are presented in Figure 1 against the number K of components for the case of (a)$M = 1$ and (b)$M = 10$. In Figure 1, an upper bound of the coefficient of the Bayesian stochastic complexity and $\lambda_{\rm BIC}$ given by eq.(19) are also plotted for the comparison of variational Bayesian learning with true Bayesian learning in the next section. The variational Bayesian algorithm gave $\lambda_{\rm VB}$ that coincide with the coefficient $\overline{\lambda}$. This implies the upper bound in eq.(17) is tight.

We also calculated the generalization error defined by $K(p(x|\theta_0)||\langle p(x|\theta)\rangle_{r(\theta|X^n)})$, where $\langle p(x|\theta)\rangle_{r(\theta|X^n)}$ is the predictive distribution in variational Bayesian learning. In the case of the Gaussian mixture model, it is given by $\langle p(x|\theta)\rangle_{r(\theta|X^n)} = \sum_{k=1}^{K} \frac{\overline{a}_k}{\sqrt{2\pi(1+\overline{\sigma}_k^2)}^M} \exp(\frac{-\|x-\overline{\mu}_k\|^2}{2(1+\overline{\sigma}_k^2)})$. The generalization error, multiplied by n for scaling purposes, was approximated by

$$\lambda_{\rm G} = \frac{n}{n'} \sum_{i=1}^{n'} \log \frac{p(x_i'|\theta_0)}{\langle p(x_i'|\theta)\rangle_{r(\theta|X^n)}}, \tag{29}$$

with $n' = 10000$ test data $\{x_i'\}_{i=1}^{n'}$ generated from the true distribution. The results of the averages of $\lambda_{\rm G}$ over 100 draws of the data sets with the size $n = 1000$ are also plotted in Figure 1. The results of the averages of $\lambda_{\rm VB}$ and $\lambda_{\rm G}$ showed different behavior. More specifically, $\lambda_{\rm G}$ increased little while $\lambda_{\rm VB}$ grew proportionally to the number K of components. From eq.(6) and eq.(7), $\lambda_{\rm VB}$ and $\lambda_{\rm G}$ should have shown similar behavior if there were the same relation between the average normalized variational stochastic complexity and the average generalization error as in Bayesian learning. These results imply that in variational Bayesian learning, unlike in Bayesian learning, the coefficient of the average generalization error differs from that of the average variational stochastic complexity $E_{X^n}[\overline{F}(X^n)]$.

8. Discussion

In this paper, we showed upper and lower bounds of the variational stochastic complexity of the Gaussian mixture models. We discuss five topics.

8.1 Lower Bound

Let us discuss the lower bound. The lower bound in eq.(17) can be improved to give

$$\overline{F}_0(X^n) \geq \overline{\lambda}\log n + nH_n(\overline{\theta}_{vb}) + C_1, \tag{30}$$

if the consistency of the variational estimator $\overline{\theta}_{vb}$ is proven. Note that the coefficient $\overline{\lambda}$ is the same as that of the upper bound given in Theorem 3. The consistency means that the mixing coefficient $\overline{a}_k$ does not tend to zero for at least K_0 components and they are always used to learn the K_0 true components when the sample size n is sufficiently large. We conjecture that the variational estimator

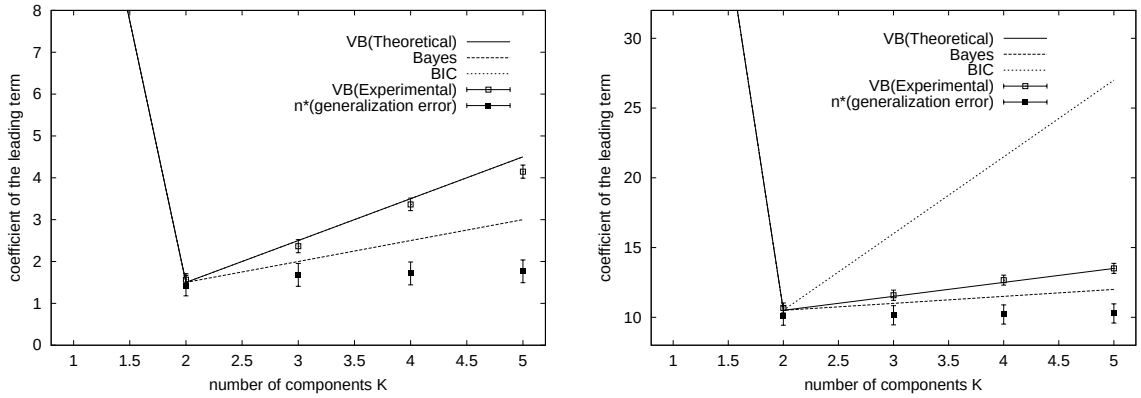

Figure 1: The coefficients of the stochastic complexities for the number K of components with $K_0 = 2$, $\phi_0 = 1$ and (a) $M = 1$, (b) $M = 10$. The solid line is $\overline{\lambda}$ of the variational Bayes eq.(18), the dashed line is the upper bound of λ in true Bayesian learning eq.(32) and the dotted line is $\lambda_{\rm BIC}$ of the BIC eq.(19). The open squares with error bars are the results of the averages of $\lambda_{\rm VB}$ eq.(28) and the full squares with error bars are the results of the averages of $\lambda_{\rm G}$ eq.(29). The error bars show 95% confidence intervals.

is consistent and the inequality (30) holds for the Gaussian mixture model. However, little has been known so far about the behavior of the variational estimator. Analyzing its behavior and investigating the consistency are important undertakings.

Furthermore, on the left hand side of eq.(17), $nH_n(\overline{\theta}_{vb})$ is a kind of training error. If the maximum likelihood estimator exists, it is lower bounded by

$$\min_{\theta} nH_n(\theta) = \min_{\theta} \sum_{i=1}^{n} \log \frac{p(x_i|\theta_0)}{p(x_i|\theta)},$$

which is the (maximum) likelihood ratio statistic with sign inversion. It is known that the likelihood ratio statistics of some non-regular models diverge to infinity as n grows and that the divergence of the likelihood ratio makes the generalization performance worse in the maximum likelihood estimation. In the case of the Gaussian mixture model, it is conjectured that the likelihood ratio diverges in the order of $\log\log n$ (Hartigan, 1985). Although this has not been proved, it suggests that the upper bound in eq.(17) is tight. More specifically, if eq.(30) holds and the order of divergence of the likelihood ratio is smaller than $\log n$, that is, $E_{X^n}[\min_\theta nH_n(\theta)] = o(\log n)$, then it immediately follows from Corollary 4 that

$$E_{X^n}[\overline{F}_0(X^n)]/\log n \to \overline{\lambda} \quad (n \to \infty). \tag{31}$$

This was suggested also by the experimental results presented in the previous section.

8.2 Comparison to Bayesian Learning

We compare the normalized variational stochastic complexity shown in Theorem 3 with the one in true Bayesian learning assuming eq.(31) holds. The Bayesian stochastic complexities of several

non-regular models have been clarified in some recent studies. For the Gaussian mixture model in particular, the following upper bound on the coefficient of the average normalized Bayesian stochastic complexity $E_{X^n}[F_0(X^n)]$ described as eq.(6) is known (Watanabe et al., 2004),

$$\lambda \leq (MK_0 + K - 1)/2, \tag{32}$$

under the same condition about the true distribution and the model as the condition (i) described in Section 5 and certain conditions about the prior distribution. Since these conditions about the prior are satisfied by putting $\phi_0 = 1$ in the condition (ii) of Theorem 3, we can compare the stochastic complexities in this case. Putting $\phi_0 = 1$ in eq.(18), we have

$$\overline{\lambda} = K - K_0 + (MK_0 + K_0 - 1)/2. \tag{33}$$

Let us compare this $\overline{\lambda}$ of variational Bayesian learning to λ in eq.(32) of true Bayesian learning.

For any M,

$$\overline{\lambda} - \lambda \geq (K - K_0)/2$$

holds. This implies that the more redundant components the model has, the more variational Bayesian learning differs from true Bayesian learning. However the difference $(K - K_0)/2$ is rather small since it is independent of the dimension M of the input space. This implies the variational posterior is close to the true Bayesian posterior. Moreover, it is noted that when $M = 1$, that is, the input is one-dimensional, $2\overline{\lambda}$ is equal to $2K - 1$ that is the number of the parameters of the model. Hence the Bayesian information criterion (BIC) (Schwarz, 1978) and the minimum description length (MDL) (Rissanen, 1986) correspond to $\overline{\lambda} \log n$ when $M = 1$.

Figure 1 shows the coefficients $\overline{\lambda}$, λ_{BIC} and the upper bound of the coefficient λ of the Bayesian stochastic complexity with respect to the number K of components for the case when $K_0 = 2$, $\phi_0 = 1$ and (a) $M = 1$ and (b) $M = 10$. In (a) of Figure 1, $\overline{\lambda}$ (solid line) and λ_{BIC} (dotted line) coincide. It is noted that $\overline{\lambda}$ of variational Bayesian learning eq.(33) relatively approaches the upper bound in Bayesian learning eq.(32) and becomes far smaller than that of BIC eq.(19) as the dimension M becomes larger.

8.3 Stochastic Complexity and Generalization

We have discussed how much the variational posterior differs from the true Bayesian posterior by comparing the stochastic complexities. In variational Bayesian learning, there is no apparent relationship between the average variational stochastic complexity and the average generalization error unlike in Bayesian learning where their leading terms are given by the same coefficient λ as in eq.(6) and eq.(7). This was also observed experimentally by the different behavior of λ_{VB} and λ_{G} in the previous section. Hence, assessing the generalization performance of the Gaussian mixture model in variational Bayesian learning is an important issue to be addressed. The term $(\log C_Q(\overline{\theta}) + S(X^n))$ in Lemma 6 may diverge to infinity as the likelihood ratio statistic in the maximum likelihood method as mentioned above. It would be important to clarify how this term affects the generalization performance in variational Bayesian learning.

8.4 Effect of Hyperparameters

Let us discuss the effects of the hyperparameters. From Theorem 3, only the hyperparameter ϕ_0 affects the leading term of the normalized variational stochastic complexity $\overline{F}_0(X^n)$ and the other

hyperparameters ξ_0 and ν_0 affect only the lower order terms. This is due to the influence of the hyperparameters on the prior probability density around the true parameters. Consider the case when $K_0 < K$. In this case, for a parameter that gives the true distribution, either of the followings holds, $a_k = 0$ for some k or $\mu_i = \mu_j$ for some pair (i, j). The prior distribution $\varphi(\mathbf{a})$ given by eq.(15) can drastically change the probability density around the points where $a_k = 0$ for some k by changing the hyperparameter ϕ_0 while the prior distribution $\varphi(\mu)$ given by eq.(16) always takes positive values for any values of the hyperparameters ξ_0 and ν_0.

We also point out that Theorem 3 shows how the hyperparameter ϕ_0 influence variational Bayesian learning. The coefficients $\underline{\lambda}$ and $\overline{\lambda}$ in eq.(18) are divided into two cases. These cases correspond to whether $\phi_0 \leq (M+1)/2$ holds, indicating that the influence of the hyperparameter ϕ_0 in the prior $\varphi(\mathbf{a})$ appears depending on the dimension M of the input space. More specifically, only when $\phi_0 \leq (M+1)/2$, the prior distribution reduces redundant components; otherwise it uses all the components.

8.5 Applications of the Bounds

Finally, let us give examples of how to use the theoretical bounds given in Theorem 3 and discuss issues to be addressed.

Comparing the theoretical bounds in eq.(17) with experimental results, one can investigate the properties of the actual iterative algorithm in variational Bayesian learning. Although the actual iterative algorithm gives the variational posterior that satisfies eq.(9) and eq.(10), it may converge to local minima of the functional $\overline{F}[q]$. Remember that eq.(9) and eq.(10) are just a necessary condition for $\overline{F}[q]$ to be minimized. One can examine experimentally whether the algorithm converges to the optimal variational posterior that minimizes the functional instead of local minima by comparing the experimental results with the theoretical bounds. Moreover, the theoretical bounds would enable us to compare the accuracy of variational Bayesian learning with that of the Laplace approximation or the MCMC method. However, in order to make such comparisons more accurately, one will need not only the leading term but also the lower order terms of the asymptotic form of the variational stochastic complexity. Giving the more accurate asymptotic form is important for such comparisons.

The Gaussian mixture model is included in general exponential family models with hidden variables (Sato, 2001) and furthermore, in general graphical models to which the variational Bayesian framework can be applied (Attias, 1999). Analyzing the variational stochastic complexities in the more general cases would be an important undertaking.

Furthermore, as mentioned in Section 4, the variational stochastic complexity $\overline{F}(X^n)$ is used as a criterion for model selection in variational Bayesian learning. Theorem 3 shows how accurately one can estimate the Bayesian stochastic complexity $F(X^n)$, the negative log of the Bayesian evidence, by its upper bound $\overline{F}(X^n)$. By the above comparison to Bayesian learning, it is expected that $\overline{F}(X^n)$ provides a rather good approximation to $F(X^n)$. This gives a theoretical justification for its use in model selection. Our result is important for developing effective model selection methods using $\overline{F}(X^n)$.

9. Conclusion

In this paper, we derived upper and lower bounds of the variational stochastic complexity of the Gaussian mixture models. Using the derived bounds, we discussed the influence of the hyperparameters and the accuracy of variational Bayesian learning as an approximation of true Bayesian

learning. These bounds can be used for evaluation and optimization of learning algorithms based on the variational Bayesian approximation.

Acknowledgments

The authors would like to thank Dr.Masa-aki Sato, Dr.Shin Ishii and Dr.Kenji Fukumizu for their helpful comments. This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for JSPS Fellows 16-4637 and for Scientific Research 15500130, 2005. An early version of this paper has been presented at IEEE conference on Cybernetics and Intelligent Systems (Watanabe and Watanabe, 2004).

Appendix A.

Proof of Lemma 2

Proof From the restriction of the variational Bayesian approximation eq.(8), $\overline{F}(X^n)$ can be divided into two terms,

$$\overline{F}(X^n) = \min_{r,Q}\Big[\langle \log \frac{r(\theta|X^n)}{\varphi(\theta)}\rangle_{r(\theta|X^n)} + \langle \log \frac{Q(Y^n|X^n)}{p(X^n,Y^n|\theta)}\rangle_{r(\theta|X^n)Q(Y^n|X^n)}\Big].$$

Since the optimal variational posteriors satisfy eq.(9) and eq.(10), if the variational posterior $Q(Y^n|X^n)$ is optimized, then

$$\langle \log \frac{Q(Y^n|X^n)}{p(X^n,Y^n|\theta)}\rangle_{r(\theta|X^n)Q(Y^n|X^n)} = -\log C_Q$$

holds. Thus we obtain eq.(12). ■

Proof of Lemma 5

Proof Calculating the Kullback information between the posterior and the prior, we obtain

$$K(r(\mathbf{a}|\bar{\mathbf{a}})||\varphi(\mathbf{a})) = \sum_{k=1}^{K} h(n_k) - n\Psi(n+K\phi_0) + \log\Gamma(n+K\phi_0) + \log\frac{\Gamma(\phi_0)^K}{\Gamma(K\phi_0)}, \tag{34}$$

where we use the notation $h(x) = x\Psi(x+\phi_0) - \log\Gamma(x+\phi_0)$. Similarly,

$$K(r(\mu|\bar{\mu})||\varphi(\mu)) = \sum_{k=1}^{K}\frac{M}{2}\log\frac{n_k+\xi_0}{\xi_0} - \frac{KM}{2} + \frac{1}{2}\xi_0\sum_{k=1}^{K}\{\frac{M}{n_k+\xi_0} + \|\bar{\mu}_k-\nu_0\|^2\}. \tag{35}$$

By using inequalities (20) and (21), we obtain

$$h(x) = -(\phi_0 - \frac{1}{2})\log(x+\phi_0) + x + O(1).$$

Thus we have, from eqs.(34),(35) and $K(r(\theta|\bar{\theta})||\varphi(\theta)) = K(r(\mathbf{a}|\bar{\mathbf{a}})||\varphi(\mathbf{a})) + K(r(\mu|\bar{\mu})||\varphi(\mu))$,

$$\Big|K(r(\theta|\bar{\theta})||\varphi(\theta)) - \Big\{G(\bar{\mathbf{a}}) + \frac{\xi_0}{2}\sum_{k=1}^{K}\|\bar{\mu}_k-\nu_0\|^2\Big\}\Big| \le C,$$

where C is a constant since $\frac{1}{n+\xi_0} < \frac{1}{n_k+\xi_0} < \frac{1}{\xi_0}$. ■

Proof of Lemma 6

Proof

$$\begin{aligned} C_Q(\bar{\theta}) &= \prod_{i=1}^{n} \sum_{y_i} \exp \langle \log p(x_i, y_i|\theta) \rangle_{r(\theta|\bar{\theta})} \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \frac{1}{\sqrt{2\pi}^M} \exp\{\Psi(n_k+\phi_0) - \Psi(n+K\phi_0) - \frac{\|x_i-\bar{\mu}_k\|^2}{2} - \frac{M}{2}\frac{1}{n_k+\xi_0}\}. \end{aligned} \tag{36}$$

Thus we have eq.(22).

Using again the inequalities (20), we obtain

$$-\log C_Q(\bar{\theta}) \leq -\sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} \frac{\bar{a}_k}{\sqrt{2\pi}^M} \exp\{-\frac{\|x_i-\bar{\mu}_k\|^2}{2} - \frac{M+2}{2(n_k+\min\{\phi_0,\xi_0\})}\}\Big] - \frac{n}{2(n+K\phi_0)},$$

and

$$-\log C_Q(\bar{\theta}) \geq -\sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} \frac{\bar{a}_k}{\sqrt{2\pi}^M} \exp\{-\frac{\|x_i-\bar{\mu}_k\|^2}{2}\}\Big] - \frac{n}{n+K\phi_0},$$

which give the upper and lower bounds in eq.(23) respectively. ■

References

H. Akaike. Likelihood and bayes procedure. In *Bayesian Statistics*, pages 143–166, Valencia, Spain, 1980. (Bernald J.M. eds.), University Press.

H. Alzer. On some inequalities for the Gamma and Psi functions. *Mathematics of computation*, 66 (217):373–389, 1997.

H. Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of Uncertainty in Artificial Intelligence(UAI'99)*, 1999.

M. J. Beal. *Variational Algorithms for approximate Bayesian inference*. PhD thesis, University College London, 2003.

D. Dacunha-Castelle and E. Gassiat. Testing in locally conic models, and application to mixture models. *Probability and Statistics*, 1:285–317, 1997.

Z. Ghahramani and M. J. Beal. Graphical models and variational methods. *Advanced Mean Field Methods – Theory and Practice, eds. D. Saad and M. Opper, MIT Press*, 2000.

J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer*, pages 807–810, 1985.

E. Levin, N. Tishby, and S. A. Solla. A statistical approaches to learning and generalization in layered neural networks. *Proc. of IEEE*, 78(10):1568–1674, 1990.

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

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

M. Sato. Online model selection based on the variational bayes. *Neural Computation*, 13(7):1649–1681, 2001.

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

K. Watanabe and S. Watanabe. Lower bounds of stochastic complexities in variational bayes learning of gaussian mixture models. In *Proceedings of IEEE conference on Cybernetics and Intelligent Systems (CIS04)*, pages 99–104, Singapore, 2004.

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

S. Watanabe, K. Yamazaki, and M. Aoyagi. Kullback information of normal mixture is not an analytic function. *Technical Report of IEICE (in Japanese)*, NC2004-50:41–46, 2004.

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

K. Yamazaki and S. Watanabe. Stochastic complexity of bayesian networks. In *Proceedings of Uncertainty in Artificial Intelligence(UAI'03)*, 2003b.