社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

変分ベイズ法による混合正規分布モデルの学習における確率的複雑さについて

渡辺　一帆†　　渡辺　澄夫††

† 東京工業大学大学院総合理工学研究科知能システム科学専攻　〒 226–8503 横浜市緑区長津田 4259
†† 東京工業大学精密工学研究所　〒 226–8503 横浜市緑区長津田 4259 MailBox:R2-5
E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

あらまし　変分ベイズ法はベイズ学習の一つの近似手法として提案され、少ない計算量とその有効性が多くの実問題を通じて検証されてきた。しかしながら、変分ベイズ法がどの程度精密な近似を与えるかはいまだに解明されてこなかった。本研究では、混合正規分布モデルの変分ベイズ学習について考察し、確率的複雑さの下界を与える。混合正規分布は特異モデルの 1 つであるが、特異モデルの学習理論の発展により、そのベイズ学習の性質についてもいくつかの結果が得られてきた。本研究で得られた確率的複雑さの値を真のベイズ法の場合と比較することで、変分ベイズ法の近似法としての精度について明らかにする。

キーワード　混合正規分布、変分ベイズ学習、確率的複雑さ、特異モデル

Stochastic Complexities in Learning of Normal Mixture Models by Variational Bayes Approach

Kazuho WATANABE† and Sumio WATANABE††

† Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology　4259 Nagatsuta, Midori-ku, Yokohama, 226–8503 Japan
†† P&I Lab, Tokyo Institute of Technology　4259 Nagatsuta, Midori-ku, Yokohama, 226–8503 Japan
E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

Abstract　The Variational Bayes approach, proposed as an approximation of the Baysian learning, has provided computational tractability and good generalization performance in many applications. In spite of these advantages, the properties and capabilities of the Variational Bayes learning itself have not been clarified yet. It is still unknown how good approximation the Variational Bayes approach can achieve. In this paper, we discuss the Variational Bayes learning of normal mixture models and derive the lower bounds of the stochastic complexities that show us the accuracy of Variational Bayes learning as an approximation.

Key words　Normal Mixture Model, Variational Bayes Learning, Stochastic Complexity, Singular Statistical Model

1. Introduction

A normal mixture model is a learning machine which estimates the target probability density by the sum of several normal distributions. This learning machine is widely used especially in statistical pattern recognition and data clustering. In spite of the wide range of its applications, the properties of its learning and generalization have not yet been made clear enough to design its optimal model structure. That is because statistical models with hidden variables are singular models along with many learning machines used in most data modelling problems, such as neural networks and hidden markov models.

The Bayesian learning is widely used to train learning machines including the normal mixture model. And it has been proven that the Bayesian framework provides better generalization performance in learning singular models than the maximum likelihood (ML) method that tends to produce a model overfitting the data. In the Bayesian framework, one computes the Bayesian posterior. However, The computation requires huge vosts and generally cannot exactly.

The Variational Bayes (VB) framework was proposed as

an approximation for computations in the Bayesian learning of models with hidden variables [1]. This framework has been applied to various real-world data modelling problems and empirically proved to be effective both in computational tractability and generalization performance.

In spite of its tractability, the properties of the Variational Bayes learning itself have not been clarified yet. Although the Variational Bayes framework is an approximation, questions like how accurately does it approximate the true Bayesian learning are yet to be answered.

On the other hand, a lot of attentions have been paid to the properties of singular statistical models. Especially in the Bayesian learning, mathematical foundation to analyze singular models was established based on algebraic geometrical method [5]. The Bayesian stochastic complexities and generalization errors of several singular models have been clarified in some recent studies [4]. These results enabled us to discuss the accuracy of several approximation schemes to realize the Bayesian learning.

In this paper, we focus on the learning of normal mixture models in the Variational Bayes framework and derive the lower bound of the stochastic complexity that enables us to see the difference between the Variational Bayes learning and the true one.

2. Normal 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 normal 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 $\{a_k | a_k \geqq 0, \sum_{k=1}^{K} a_k = 1\}$ is the set of coefficients. 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 supposed that there is no correlation between each input dimension. In this case, the model is written by

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

where $\sigma_k > 0$ is a constant.

In this paper, we consider learning this type of normal mixture models eq.(1) in the Variational Bayes framework, and show the lower bound of the stochastic complexity in Theorem 1.

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

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

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

A normal mixture model is a singular statistical model, since it has non-identifiability of the parameter. 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 a model parameter is non-identifiable, the usual asymptotic theory of a regular statistical model cannot be applied. Some studies have clarified that a normal mixture model has some different properties from those of regular statistical model [4].

3. The Bayesian Learning and the Variational Bayes

3.1 The 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 the 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 dataset 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 a normalization constant that is also known as the marginal likelihood or the evidence of the dataset X^n.

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

$$p(x|X^n) = \int p(x|\theta) p(\theta|X^n) d\theta. \quad (2)$$

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

$$F(X^n) = -\log Z(X^n), \quad (3)$$

which becomes 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.

The Bayesian posterior can be rewritten as

$$p(\theta|X^n) = \frac{1}{Z_0(X^n)} \exp(-nH_n(\theta)) \varphi(\theta), \quad (4)$$

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

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

and $Z_0(X^n)$ is a normalization constant. Putting $S(X^n) = -\sum_{i=1}^n \log p_0(x)$, we have

$$F(X^n) = -\log Z_0(X^n) + S(X^n). \tag{5}$$

It is noted that the empirical entropy $S(X^n)$ does not depend on the model $p(x|\theta)$ and its expectation value over all sets of training samples is equal to nS where $S = -\int p_0(x)\log p_0(x)dx$ is the entropy. Therefore minimization of $F(X^n)$ is equivalent to that of $-\log Z_0(X^n)$.

Evaluating the expectation value over all sets of training samples, we define the average normalized stochastic complexity $F(n)$ by

$$F(n) = \left\langle F(X^n)\right\rangle_{p_0(X^n)} - nS, \tag{6}$$

where $p_0(X^n) = \prod_{i=1}^n p_0(x_i)$(1).

On the Bayesian learning, mathematical foundation for general models including singular models was established [5], which enabled us to clarify the asymptotic behavior of the stochastic complexity of singular models. More specifically, by using a concept in algebraic analysis, it was proved that the average stochastic complexity $F(n)$ has the following asymptotic form,

$$F(n) \simeq \lambda \log n - (m-1)\log\log n + O(1), \tag{7}$$

where λ is a rational number determined by the pole of the zeta function of a learning machine and m is a natural number determined by the order of the pole. In regular statistical models, 2λ is equal to the number of parameters and $m = 1$, whereas in singular models such as normal mixture models, 2λ is not larger than the number of parameters and $m \geqq 1$. Since the increase of the stochastic complexity is equal to the generalization error, the Bayesian learning is more suitable for learning singular models than the maximum likelihood (ML) method.

However, in order to carry out the Bayesian learning practically, one computes the 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 the well-known and simplest method based on the assumption that all posterior distributions are normal. However, posterior distributions of singular models never converge to normal distributions in general, even as n tends to infinity. Therefore, the Laplace approximation is not sufficient for singular models. As a more exact approximation, Markov chain Monte Carlo (MCMC) method is well known. This approach attempts to sample from the exact posterior distribution but typically requires vast computational resources.

As an another approximation, the Variational Bayes framework was proposed. In the following, we explain the outline of it.

3.2 The Variational Bayes Framework

Using the likelihood on the whole data $\{X^n, Y^n\}$ added the corresponding hidden variables $Y^n = \{y_1, \cdots, y_n\}$, we can rewrite the stochastic complexity (3) 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 Bayes framework starts with upper bounding the stochastic complexity. For an arbitrary conditional distribution $q(Y^n, \theta|X^n)$ on the hidden variables and the parameters, the stochastic complexity can be upper bounded by applying Jensen's inequality.

$$\begin{aligned} F(X^n) &= -\log \sum_{Y^n} \int p(X^n, Y^n, \theta) d\theta \\ &\leqq \sum_{Y^n} \int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} d\theta \\ &\equiv \overline{F}(X^n). \end{aligned} \tag{8}$$

This inequality becomes an equality if and only if $q(Y^n, \theta|X^n) = \frac{1}{Z} p(Y^n, \theta|X^n)$, that is, $q(Y^n, \theta|X^n)$ equals to the Bayesian posterior distribution. This means that the smaller $\overline{F}(X^n)$, the closer the distribution $q(Y^n, \theta|X^n)$ to the true Bayesian posterior distribution.

The Variational Bayes approach makes approximations aiming at realizing a computationally tractable posterior. More specifically, assuming the parameters and the hidden variables are conditionally independent of each other, the VB approach restricts the set of $q(Y^n, \theta|X^n)$ to distributions that have the form

$$q(Y^n, \theta|X^n) = Q(Y^n|X^n) r(\theta|X^n), \tag{9}$$

where $Q(Y^n|X^n)$ and $r(\theta|X^n)$ are posteriors on the hidden variables and the parameters respectively. This posterior is termed the variational posterior and generally differ from the true Bayesian posterior.

Minimization of the functional $\overline{F}(X^n)$ 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 constraint $\int r(\theta|X^n)d\theta = 1$, we obtain the optimal variational posterior $r(\theta|X^n)$ over the parameters,

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

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

where C_r is a normalization constant. And similarly, we obtain the variational posterior over the hidden variables,

$$Q(Y^n|X^n) = \frac{1}{C_Q} \exp \left\langle \log p(X^n, Y^n|\theta) \right\rangle_{r(\theta|X^n)}, \qquad (11)$$

where C_Q is a normalization constant.

These optimal variational posteriors that satisfy eqs.(10) and (11) are computed by an iterative algorithm whose convergence is guaranteed.

Since $\overline{F}(X^n)$ in eq.(8), the stochastic complexity in the Variational Bayes learning, gives the upper bound of the true stochastic complexity $F(X^n)$, $\overline{F}(X^n)$ itself is an estimate of $F(X^n)$ and is used for the model selection in the VB learning. In the next section, we prove Theorem 1 that shows the lower bound of the average stochastic complexity in the VB learning in the situation that the model has redundant components to realize the true distribution. Investgating how the stochastic complexity increases in the situation becomes important when we consider the model selection and hypothesis testing. Moreover, the difference between the stochastic complexity in the VB learning and the true one shows us the accuracy of the VB approach as an approximation of the true Bayesian learning.

4. Main Theorem

The average normalized stochastic complexity $\overline{F}(n)$ in the Variational Bayes learning is defined by

$$\overline{F}(n) = \left\langle \overline{F}(X^n) \right\rangle_{p_0(X^n)} - nS. \qquad (12)$$

We assume the following conditions.

(i) The true distribution $p_0(x)$ is an M-dimensional normal 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 has K ($K \geqq K_0$) components,

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

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

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^k} \prod_{k=1}^{K} a_k^{\phi_0 - 1}, \qquad (13)$$

$$\varphi(\mu) = \prod_{k=1}^{K} \sqrt{\frac{\xi_0}{2\pi}}^M \exp\{-\frac{\xi_0\|\mu_k - \nu_0\|^2}{2}\}, \qquad (14)$$

where $\xi_0 > 0$, $\nu_0 \in R^M$ and $\phi_0 > 0$ are constants called hyperparemetes. These are Dirichlet and normal distributions respectively, which are commonly used in the Variational Bayes learning of normal mixture models.

On these conditions, we prove the following theorem.

[Theorem 1] Assume the conditions (i) and (ii). Then the average stochastic complexity (12) satisfies

$$\overline{F}(n)/\log n \to \overline{\lambda},$$

as n tends to infinity where

$$\overline{\lambda} \geqq \begin{cases} (K-K_0)\phi_0 + \frac{MK_0+K_0-1}{2} & (\phi_0 \leqq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}). \end{cases} \qquad (15)$$

(**Proof of Theorem 1**)

The difference between $\overline{F}(X^n)$ and the original stochastic complexity $F(X^n)$ is the Kullback information from the variational posterior to the true posterior[(2)]. That is

$$\overline{F}(X^n) - F(X^n) = \sum_{Y^n} \int q(Y^n, \theta) \log \frac{q(Y^n, \theta)}{P(Y^n, \theta|X^n)} d\theta$$

From the restriction of the VB approximation eq.(9), the above Kullback information can be divided into that of the hidden variables given the parameters and that of the parameters,

$$\begin{aligned} &\sum_{Y^n} \int q(Y^n, \theta) \log \frac{q(Y^n, \theta)}{P(Y^n|\theta, X^n)\varphi(\theta)} d\theta \\ &= \sum_{Y^n} \int Q(Y^n) r(\theta) \log \frac{Q(Y^n)}{P(Y^n|\theta, X^n)} d\theta \\ &\qquad + \int r(\theta) \log \frac{r(\theta)}{p(\theta|X^n)} d\theta. \end{aligned} \qquad (16)$$

Since the first term is the Kullback information, it is not negative. Hence $\overline{F}(X^n)$ is lower bounded as follows,

$$\begin{aligned} \overline{F}(X^n) &\geqq \int r(\theta) \log \frac{r(\theta)}{p(\theta|X^n)} d\theta + F(X^n) \\ &= \int r(\theta) \log \frac{r(\theta)}{\varphi(\theta)} d\theta + n \int H_n(\theta) r(\theta) d\theta + S(X^n), \end{aligned}$$

where we used eqs.(4) and (5). Denoting by $K(q(x)||p(x))$ the Kullback information between two distributions $q(x)$ and $p(x)$, we obtain

$$\overline{F}(X^n) - S(X^n) \geqq K(r(\theta)||\varphi(\theta)) + n\langle H_n(\theta)\rangle_{r(\theta)}. \qquad (17)$$

First, we evaluate the first term on the right-hand side of the above inequality by actually calculating the variational posterior $r(\theta)$ of the normal mixture model.

The log-likelihood of the whole data $\{X^n, Y^n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ is given by

$$\log p(X^n, Y^n|\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[y_i^k \{\log \frac{a_k}{\sqrt{2\pi}^M} - \|x_i - \mu_k\|^2/2\} \right],$$

(2) : In this proof, we omit the condition X^n of the variational posteriors, and abbreviate them to $q(Y^n, \theta)$, $Q(Y^n)$ and $r(\theta)$.

where $y_i^k = 1$ if ith datum x_i is from the kth component, if otherwise, $y_i^k = 0$. Computing the expectation w.r.t. the hidden variable posterior $Q(Y^n)$, we have

$$\langle \log p(X^n, Y^n|\theta)\rangle_{Q(Y^n)}$$
$$\propto \sum_{k=1}^{K}\{n_k(\log a_k - \frac{\|\mu_k - \nu_k\|^2}{2}) - \sum_{i=1}^{n}\frac{\overline{y}_i^k}{2}\|x_i - \nu_k\|^2\},$$

where

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

The variable n_k is the expected number of the data that are estimated to be from kth component, and ν_k is the mean of them. By multiplying the respective prior (13) and (14), the variational posterior $r(\theta) = r(\mathbf{a})r(\mu)$ is obtained as the product of the following two distributions,

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

and

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

where

$$\overline{\sigma}_k^2 = \frac{1}{n_k + \xi_0},\quad \overline{\mu}_k = \frac{n_k\nu_k + \xi_0\nu_0}{n_k + \xi_0}.$$

Calculating the Kullback informaion between the posterior and the prior, we have

$$K(r(\mathbf{a})||\varphi(\mathbf{a})) = \sum_{k=1}^{K}g(n_k) - n\Psi(n + K\phi_0)$$
$$+ \log\Gamma(n + K\phi_0) + \log\frac{\Gamma(\phi_0)^K}{\Gamma(K\phi_0)}, \qquad (18)$$

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

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

and the notation $g(x) = x\Psi(x+\phi_0) - \log\Gamma(x+\phi_0)$. Similarly,

$$K(r(\mu)||\varphi(\mu)) = \sum_{k=1}^{K}\frac{M}{2}\log\frac{n_k + \xi_0}{\xi_0} - \frac{KM}{2}$$
$$+\frac{1}{2}\xi_0\sum_{k=1}^{K}\{\frac{M}{n_k + \xi_0} + \big(\frac{n_k}{n_k + \xi_0}\big)^2\|\nu_k - \nu_0\|^2\}. \qquad (19)$$

From eqs,(18) and (19),

$$K(r(\theta)||\varphi(\theta)) \geqq G(n_1, \cdots, n_K) + O(1), \qquad (20)$$

where we defined the function $G(n_1, \cdots, n_K)$ by

$$G(n_1, \cdots, n_K) = \sum_{k=1}^{K}\{g(n_k) + \frac{M}{2}\log(n_k + \xi_0)\}$$
$$-n\Psi(n + K\phi_0) + \log\Gamma(n + K\phi_0). \qquad (21)$$

From the inequality (20), it is noted that we can evaluate $K(r(\theta)||\varphi(\theta))$ by analyzing the function G under the constraint that $\sum_{k=1}^{K}n_k = n$.

By using the asymptotic forms of the di-gamma function $\Psi(x)$ and the log-gamma function $\log\Gamma(x)$,

$$\Psi(x) \simeq \log x - \frac{1}{2x} + O(\frac{1}{x^2}),$$

and

$$\log\Gamma(x) \simeq \big(x - \frac{1}{2}\big)\log x - x + \frac{1}{2}\log 2\pi + O(\frac{1}{x}),$$

the function $g(x)$ is asymptotically expanded as follows

$$g(x) \simeq -\big(\phi_0 - \frac{1}{2}\big)\log(x + \phi_0) + x + O(1),$$

as x tends to infinity.

Provided each n_k is the α_k-th order of n , that is, $n_k = p_k n^{\alpha_k} + o(n^{\alpha_k}), (0 < p_k, 0 \leqq \alpha_k \leqq 1, k = 1, 2, \cdots, K)$ as $n \to \infty$, since

$$\log(n_k + \phi_0) = \log(n_k + \xi_0) = \alpha_k\log n + \log(p_k + o(1)),$$

we have

$$\frac{g(n_k) + \frac{M}{2}\log(n_k + \xi_0) - n_k}{\log n} \to \{-\big(\phi_0 - \frac{1}{2}\big) + \frac{M}{2}\}\alpha_k$$

and

$$\frac{G(n_1, \cdots, n_K)}{\log n} \to \sum_{k=1}^{K}\{-\big(\phi_0 - \frac{1}{2}\big) + \frac{M}{2}\}\alpha_k + \big(K\phi_0 - \frac{1}{2}\big), \qquad (22)$$

as $n \to \infty$. This means that the first term on the right-hand side of the inequality (17) asymptotically only depends on $\{\alpha_k\}$ that are the orders of $\{n_k\}$ w.r.t. n as $n \to \infty$.

Turning now to the second term on the right-hand side of the inequality (17). Using Jensen's inequality,

$$\langle H_n(\theta)\rangle_{r(\theta)} \geqq \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(x_i|\theta_0)}{\langle p(x_i|\theta)\rangle_{r(\theta)}} \equiv D_n(\theta'), \qquad (23)$$

where $\langle p(x|\theta)\rangle_{r(\theta)}$ is the predictive distribution and only depends on $\theta' = \{n_k, \nu_k\}_{k=1}^K$ given the dataset X^n. In the case of the normal mixture model, it is again given by a mixture of gaussians as follows,

$$\langle p(x|\theta)\rangle_{r(\theta)} = \sum_{k=1}^{K}\frac{a_k'}{\sqrt{2\pi(1 + \overline{\sigma}_k^2)}^M}\exp(\frac{-\|x - \overline{\mu}_k\|^2}{2(1 + \overline{\sigma}_k^2)}), \qquad (24)$$

where $a_k' = \frac{n_k + \phi_0}{n + K\phi_0}$. We define functions $D(\theta')$ and $\zeta(\theta')$ by

$$D(\theta') = K(p(x|\theta_0)||\langle p(x|\theta)\rangle_{r(\theta)})$$

and

$$\zeta(\theta') = \frac{\sqrt{n}[D_n(\theta') - D(\theta')]}{\sqrt{D(\theta')}}.$$

It follows that

$$D_n(\theta') = D(\theta') + \sqrt{\frac{D(\theta')}{n}}\zeta(\theta') \quad (25)$$

$$\geqq D(\theta') + \sqrt{\frac{D(\theta')}{n}} \min_{\theta'} \zeta(\theta')$$

$$\geqq -\frac{\{\sup_{\theta'} \zeta(\theta')\}^2}{4n}. \quad (26)$$

$\zeta(\theta')$ converges in law to a stochastic process whose mean is 0 and variance function is finite as $n \to \infty$. Since it can be proved that $E_{X^n}[\{\sup_{\theta'} \zeta(\theta')\}^2]$ is finite [5], from the inequalities (23) and (26), there exists a positive constant C_1 such that

$$E_{X^n}\left[\langle H_n(\theta)\rangle_{r(\theta)}\right] \geqq -\frac{C_1}{n}. \quad (27)$$

If the number of n_ks such that n_k/n converges to a positive constant is less than K_0, the number of components which the true distribution has, the predictive distribution (24) has less components than the true one as $n \to \infty$. Then the predictive distribution doesn't approach the true one and

$$\lim_{n\to\infty} D(\theta') > 0$$

holds in eq.(25). That means from the inequality (23), $n\langle H_n(\theta)\rangle_{r(\theta)}$ diverges in the order of n unless at least K_0 n_ks are in the order of n. Therefore, in order to minimize the right-hand side of the inequality (17), at least K_0 n_ks must be in the order of n, and if so, from the inequality (27), the average of the second term of (17) over all sets of samples can be lower bounded by some constant independent of n. Finally, minimizing the limiting value of eq.(22) with respect to $\{\alpha_k\}$ under the constraint that $0 \leqq \alpha_k \leqq 1$ and at least K_0 α_ks satisfy $\alpha_k = 1$, which completes the proof. **(Q.E.D)**

5. Discussion and Conclusion

In this paper, we showed the lower bound of the stochastic complexity of normal mixture models in the Variational Bayes learning. The stochastic complexities in the true Bayesian learning of several singular models have been clarified in some recent studies. On the normal mixture models in particular, the following upper bound on the coefficient of the stochastic complexity $F(n)$ described as (7) is known [6],

$$\lambda \leqq (MK_0 + K - 1)/2, \quad (28)$$

on the same condition (i) about the true distribution and the model described in Section 4 and certain conditions about the prior distribution. These conditions about the prior are satisfied by putting $\phi_0 = 1$ in the condition (ii) of Theorem 1. Putting $\phi_0 = 1$ in the inequality (15), we have

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

Let us compare this $\overline{\lambda}$ of the Variational Bayes learning to λ (28) of the true Bayesian learning. For any M,

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

holds. Since $K_0 \leqq K$, this means that the Variational Bayes approach can never achieve the true Bayesian learning if the model has redundant components, and the more redundant components the model has, the more the Variational Bayes learning differs from the true one. 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) [3] corresponds to $\overline{\lambda}\log n$ when $M = 1$.

The lower bound (15) in Theorem 1 is divided into two cases. These cases are divided by the condition $\phi_0 \leqq (M+1)/2$ or otherwise, implying that the influence of the hyperparameter ϕ_0 in the prior $\varphi(\mathbf{a})$ appears only when $\phi_0 \leqq (M+1)/2$. More specifically, only then, the prior distribution works to reduce the redundant components that the model has.

In this paper, we discussed learning normal mixture models in the Variational Bayes framework, and derived the lower bound of the stochastic complexity. We showed that the stochastic complexity increases more in the Variational Bayes learning than in the true one as the model becomes more redundant. Not only the lower bound, more exact evaluation of the stochastic complexity $\overline{F}(n)$ is required in the future work for more proper discussions about properties of the Variational Bayes learning.

Acknowledgment

This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research 15500130 and for JSPS Fellows 4637, 2004.

文　　献

[1] H.Attias, "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes," *Proc. of Uncertainty in Artificial Intelligence(UAI'99)*, 1999.

[2] Z.Ghahramani, M.J.Beal, "Graphical models and variational methods," *Advanced Mean Field Methods , MIT Press*, 2000.

[3] G.Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, Vol.6, No.2, pp.461-464, 1978.

[4] K.Yamazaki, S.Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," *International Journal of Neural Networks*, 16, pp.1029-1038, 2003 .

[5] S.Watanabe,"Algebraic analysis for non-identifiable learning machines," *Neural Computation*, Vol.13, No.4, pp.899-933, 2001.

[6] S.Watanabe, K.Yamazaki, M.Aoyagi, "Kullback information of normal mixture is not an analytic function,", Technical Report of IEICE (in Japanese), NC2004-50, pp.41-46.

[7] K.Watanabe, S.Watanabe, "Lower Bounds for Stochastic Complexities of Normal Mixture in Variational Bayes Learning," *Proc. of IBIS2004*, 2004, to appear.