Statistical Learning Theory of Variational Bayes

Department of Computational Intelligence and Systems Science
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology

Kazuho Watanabe

2006

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 has been unknown how good approximation the variational Bayesian approach can achieve. Consequently, the design of learning algorithms using variational Bayesian framework has rarely been dealt with theoretically.

This thesis establishes a theory for designing variational Bayesian algorithms in the principled manner. The theoretical properties of variational Bayesian learning are discussed for mixture models. Upper and lower bounds for variational stochastic complexities are derived as the main contribution. 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.

Preface

This work has been carried out at Watanabe laboratory, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology.

I wish to thank my supervisor Prof. Sumio Watanabe for his support and a lot of encouraging comments during my undergraduate, master and Ph.D. studies. I would also like to express my great respect and gratitude to him for giving me the opportunity to meet such fascinating studies.

I am grateful to Prof. Yoshiyuki Kabashima, Prof. Toshiaki Murofushi, Prof. Misako Takayasu, and Prof. Toru Aonishi for reviewing this thesis and providing comments to improve it.

I also wish to thank present and former members of Watanabe laboratory: Dr. Keisuke Yamazaki, Dr. Shinichi Nakajima, Dr. Miki Aoyagi, Kazumi Imbe, Tikara Hosino, Motoki Shiga, Katsuyuki Takahashi, Nobuhiro Nakano, Takayuki Higo, Satoshi Tsuji, Takeshi Matsuda, Kaori Fujiwara, Kenji Nagata, Shingo Takamatsu, Yu Nishiyama, and Ryosuke Iriguchi.

Finally, I wish to thank my family and friends who have given me so much support and encouragement.

Kazuho Watanabe

Contents

Chapter 1

Introduction

Nowadays, enormous amounts of data are available and waiting for being processed. These data can be DNA sequences, which have genes to be found, images in computer vision, which have objects to be recognized, documents which have topics to be categorized and so on. Learning systems can elicit such useful information from data by using them as examples. These systems include learning machines such as multi-layer perceptrons, mixture models, hidden Markov models, Bayesian networks and support vector machines. Statistical learning theory studies the properties and capabilities of learning systems. It ultimately aims at the better design of them to capture the probabilistic structure of the complex data more efficiently.

For example, the Gaussian mixture model is widely used especially in statistical pattern recognition and data clustering. This learning machine estimates the target probability density by the sum of normal distributions. 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 (Cramer, 1946). The regularity conditions are not satisfied in 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 have been paid to the non-regular models (Amari et al., 2006; Fukumizu et al., 2004; Watanabe, 2001a). In Bayesian

learning, mathematical foundation for analyzing non-regular models was established with an algebraic geometrical method (Watanabe, 2001a). The Bayesian stochastic complexities or the marginal likelihoods of several non-regular models have been clarified in some recent studies (Yamazaki and Watanabe, 2003a; Yamazaki and Watanabe, 2003b). 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 around the maximum a posteriori (MAP) estimator, which can be unrepresentative for models containing hidden variables.

The variational Bayesian (VB) framework was proposed as another approximation (Hinton and van Camp, 1993) and extended 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. Applications of variational Bayesian learning have been reported in various real-world data modeling problems in areas such as speech recognition (Watanabe et al., 2002), image signal processing (Cheng et al., 2005), bioinformatics(Beal et al., 2005). It has been empirically proved to be computationally 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. This has been the major setback in dealing with the design of learning systems with the variational Bayesian approach in a theoretically principled way.

The objective of this work is to establish a theory for the principled design of variational Bayesian learning algorithms for non-regular statistical models. The theoretical properties of variational Bayesian leaning are discussed by deriving the asymptotic form of the variational stochastic complexity or the minimum variational free energy for mixture models.

First, we focus on the Gaussian mixture model that is the most fundamental model in non-regular ones. As the main contribution, we derive asymptotic upper and lower bounds on the variational stochastic complexity that is defined by the minimum value of the objective function in variational Bayes, called the variational free energy. 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. Then, we consider generalizing the result to the wider class of mixture models. Consequently, the properties of variational Bayesian learning are discussed for the mixture of exponential families which include mixtures of distributions such as binomial, gamma and Gaussian.

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. The main results indicate how to determine the hyperparameter values before the learning process.

Throughout this thesis, it is assumed that 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 non-identifiability arises in the mixture models and the advantage of Bayesian learning is typical in this case (Watanabe, 2001a). Furthermore, this analysis is necessary and essential for addressing the model selection and hypothesis testing problems. These problems are addressed based on the statistical properties of the stochastic complexity, such as the asymptotic mean and asymptotic distribution when the model has redundant components.

Additionally, the theoretical results are compared to the ones of experiments where variational Bayesian learning is carried out for Gaussian mixture models using synthetic data. The experimental results demonstrate the properties of the practical variational Bayesian algorithm which involves iterative

updates and hence suffers from local minima in general.

This thesis is organized as follows.

- Chapter 2 introduces basic concepts and objectives in statistical learning as the preliminaries.
- Chapter 3 reviews the standard framework for Bayesian learning and defines the Bayesian stochastic complexity.
- Chapter 4 outlines the variational Bayesian framework. The variational stochastic complexity is defined.
- Chapter 5 introduces the Gaussian mixture model and shows Theorem 2 where its asymptotic variational stochastic complexity is derived.
- Chapter 6 introduces the mixture of exponential family model and shows Theorem 3 by extending the result in Theorem 2. Some examples of mixture models follow as applications of Theorem 3.
- Chapter 7 presents the results of experiments where the variational Bayesian algorithm is applied to the Gaussian mixture model.
- Chapter 8 gives discussions on the results which include the accuracy of the variational Bayesian approximation and the effect of the hyperparameters.
- Chapter 9 summarizes the results and concludes this thesis.

Chapter 2

Statistical Learning

Learning system consists of two components: a model and a learning algorithm. This thesis deals with the learning algorithm, variational Bayes, for non-regular statistical models. In order to make the discussions in the following chapters more meaningful, this chapter gives general framework of learning from data and basic concepts in statistical models and learning algorithms. Some illustrative examples of models and algorithms are also presented.

2.1 Learning from Examples

The main goal of statistical learning theory is to study the properties of learning systems in a statistical way. Learning systems are adapted by using a set of observed data, in other words, the system is expected to learn or estimate the true underlying data-generating process. The data set is also called the training data or samples and usually given as a set of real valued vectors, $X^n = \{x_1, x_2, \cdots, x_n\}$, where n is the number of the data.

A learning system consists of two elements: a model and a learning algorithm. Given the data, a model is prepared first. In statistical learning, the model is described by a probability density function or a probability function with parameters. The model is denoted as $p(x|\theta)$, where $\theta = (\theta^{(1)}, \cdots, \theta^{(d)})^T$ is a d-dimensional vector of parameters.

Specifying the model $p(x|\theta)$, then the learning task reduces to finding the values of the parameters so that the model becomes close to the underlying distribution of the data set X^n. In the Bayesian framework, what is obtained

is not a single point of the parameter vector θ, but a probability distribution over all possible values of θ. Learning algorithms are procedures to find the specific parameter vector θ or the distribution on it using the data set X^n. They are detailed in Section 2.3.

There are two types of learning: *supervised* or *unsupervised*. This thesis deals mainly with examples of unsupervised learning where the distribution of only the input variable is estimated by the model. Major unsupervised learning tasks include density estimation, where the distribution of the data is estimated by the model, clustering, where the data set is divided into disjoint sets of mutually similar elements, learning the support of a distribution and so on.

In contrast, supervised learning tasks aim at estimating the conditional distribution of the output given the input. This corresponds to learning the input-output relation given the training data $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ that are the examples of the pairs of the input and output vectors.

As mentioned above, the framework of statistical learning started with a set of the given data X^n. There is, however, another type of supervised learning where the input points of the training data X^n are automatically collected by the learning system itself. This framework is called *active learning* in the sense that the system actively asks questions and demands answers to construct the data set. It is used to improve the system's performance especially when collecting training data examples is laborsome.

Furthermore, the types of learning are also classified according to the way in which the data are generated or observed. There are mainly two ways: *batch learning* where all the data are given at the beginning of the learning process, and *on-line learning* where the learner is allowed to obtain only one example at a time, and supposed to make an estimate of the output, before obtaining the correct answer. An on-line learning algorithm updates the current model in response to each new example. Hence, on-line learning is indispensable and works well in situations where the real-time quick response of the system is required. These include adaptive filtering and control.

The subject of this thesis is unsupervised learning in the batch setting.

2.2 Statistical Models

A certain class of statistical models is called *regular* and has been studied intensively in statistics. There are well established sets of regularity condi-

tions under which the asymptotic normality and efficiency of the maximum likelihood (ML) estimator is ensured (Cramer, 1946; Akahira and Takeuchi, 1995; van der Vaart, 1998). In other words, the maximum likelihood estimator as well as the Bayesian posterior distribution, converges to the normal distribution as the sample size n tends to infinity. These can be contrasted with non-regular models which this thesis focuses on.

Let us consider a certain class of non-regular statistical models. One of the regularity conditions asserts that the Fisher information matrix, whose ijth entry is given by,

$$I_{ij}(\theta) = E_\theta\Big[\frac{\partial \log p(x|\theta)}{\partial \theta^{(i)}} \frac{\partial \log p(x|\theta)}{\partial \theta^{(j)}}\Big],$$

is finite and positive definite. Here, $E_\theta[\cdot]$ denotes the expectation with respect to $p(x|\theta)$. This condition is violated in many statistical models which are widely used in practical applications. These include neural networks, reduced rank regression models, mixture models, hidden Markov models, and Bayesian networks. This is because the parameters of such models are non-identifiable.

A statistical model $S = \{p(x|\theta)|\theta\}$ is identifiable if $\theta_1 \neq \theta_2$ implies $p(x|\theta_1) \neq p(x|\theta_2)$, that is, the mapping from the parameters to the probability distributions is one-to-one. Let us call the parameters that give the same probability distribution, singularities, if the set of them forms a continuous subset in the parameter space. If the model parameters are non-identifiable, then the Fisher information matrix degenerates on the singularities and hence the model is non-regular.

As a simple example, let us consider a Gaussian mixture model of the form,

$$p(x|a,b,c) = (1-a)g(x-b) + bg(x-c), \tag{2.1}$$

where $g(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ is the Gaussian density function with zero mean and unit variance. If $a = 0$ and $b = 0$, the model gives the standard normal distribution $g(x)$ for any c. Similarly, it becomes the same standard normal distribution, if $a = 1$ and $c = 0$ for any b or if $b = c = 0$ for any a. The parameter-settings in the set,

$$\Theta_0 = \{\theta|a = b = 0, c \in R\} \cup \{\theta|a = 1, c = 0, b \in R\} \cup \{\theta|b = c = 0, 0 \leq a \leq 1\},$$

give the same distribution and hence the model, eq.(2.1), is non-identifiable. In this model, Θ_0 is the set of singularities. Figure 2.1 illustrates the set Θ_0 in the parameter space.

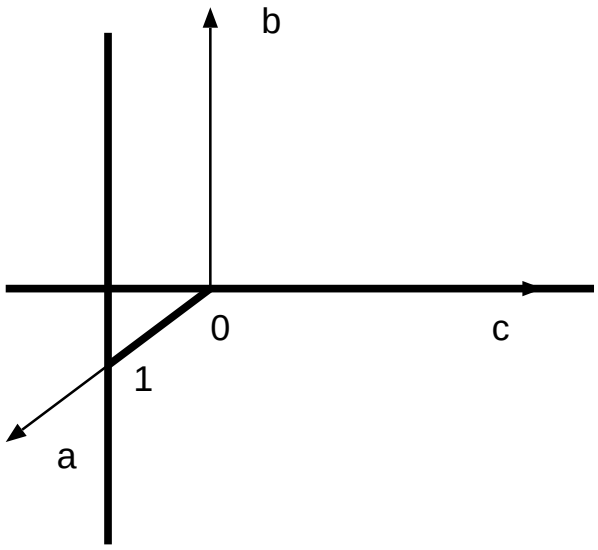


Figure 2.1: Singularities of the simple Gaussian mixture model, eq.(2.1).

2.3 Learning Algorithms

Once the statistical model $p(x|\theta)$ is specified, then a learning algorithm is applied to it. As most fundamental learning algorithms, this section introduces maximum likelihood (ML), maximum a posteriori (MAP), and Bayesian learning.

In ML learning, the parameter θ of the model is chosen to maximize the likelihood function to obtain the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$,

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i|\theta).$$

In MAP learning, the parameter θ is viewed as a random variable and some prior knowledge about the parameter values is reflected to the prior distribution $\varphi(\theta)$. Then the MAP estimator $\hat{\theta}_{\mathrm{MAP}}$ is obtained by maximizing the posterior probability density of the parameter θ,

$$\hat{\theta}_{\mathrm{MAP}} = \operatorname*{argmax}_{\theta} \varphi(\theta) \prod_{i=1}^{n} p(x_i|\theta).$$

In Bayesian learning, the parameter θ is again viewed as a random variable and the posterior distribution over all parameter values is computed as

$$p(\theta|X^n) \propto \varphi(\theta) \prod_{i=1}^{n} p(x_i|\theta).$$

After gaining the estimator or the posterior distribution of the parameter, the predictive distribution is obtained to estimate the true probability distribution that generates the data set X^n. In ML and MAP learning, the predictive distributions are given by $p(x|\hat{\theta}_{\mathrm{ML}})$ and $p(x|\hat{\theta}_{\mathrm{MAP}})$ respectively. These two methods uses only a one-point estimate of the parameter. In contrast, Bayesian learning uses the ensemble of models according to the posterior distribution $p(\theta|X^n)$ to obtain the predictive distribution,

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

The Bayesian framework will be revisited and elaborated in Chapter 3. The main subject of this thesis, variational Bayesian learning, is an approximation of Bayesian learning, where the Bayesian posterior distribution is approximated by another distribution as will be detailed in Chapter 4.

2.4 Model Selection and Hyperparameter Optimization

One of the main objectives in statistical learning theory is to evaluate the generalization performance of the learning system consisting of the model and the learning algorithm. More specifically, the quality of learning is assessed by how far the predictive distribution is from the true distribution as a probability distribution. Ultimately, this is to design better learning systems with the improved generalization performance. Model selection and hyperparameter optimization consider selecting a model or setting the prior distribution appropriately to improve the generalization performance.

Model selection itself is an inference problem where the complexity of the underlying distribution is estimated. The task is to select one appropriate model from a set of possible candidates of models $\{p_1(x|\theta_1), p_2(x|\theta_2), \cdots\}$, or to evaluate the posterior probability over the candidates. For example, the candidates may be chosen to be models with different number of components in the case of Gaussian mixture models.

Another approach is optimizing the prior distribution so that the quality of learning is improved. This can be carried out by parameterizing the prior distribution as $\varphi(\theta|\xi)$ and estimating the parameter ξ according to the data. The parameter ξ of the prior distribution is called the *hyperparameter* to distinguish it from the model parameter θ.

As mentioned in Section 2.2, the properties of regular statistical models have been studied in statistics. Consequently, some efficient methods for model selection have been constructed.

One of these is the Akaike's information criterion (AIC), which propose to select the model that minimizes the following quantity,

$$(\text{AIC}) = -\sum_{i=1}^{n} \log p(x_i|\hat{\theta}_{\text{ML}}) + d,$$

where d is the number of parameters of the model (Akaike, 1974).

Another is the Bayesian information criterion (BIC), which instead propose to minimize,

$$(\text{BIC}) = -\sum_{i=1}^{n} \log p(x_i|\hat{\theta}_{\text{MAP}}) + \frac{d}{2} \log n,$$

which corresponds to the minimum description length (MDL) criterion (Schwarz, 1978; Rissanen, 1986). Note that, however, all these methods were derived from the theory of regular statistical models and hence lose theoretical justification for non-regular models.

One of the hyperparameter optimization schemes chooses the hyperparameter ξ that minimizes the following quantity,

$$-\log \int \prod_{i=1}^{n} p(x_i|\theta)\varphi(\theta|\xi)d\theta.$$

This is called the *stochastic complexity* as will be defined again in Section 3.3. This approach to optimization of the hyperparameter is called the *empirical Bayes* (Good, 1965; Efron and Morris, 1973).

Chapter 3

Bayesian Learning

This chapter starts with the standard framework for Bayesian learning in Section 3.1 and continuing with the definition of the Kullback information in Section 3.2. The Bayesian stochastic complexity is defined in Section 3.3. The recent results of its asymptotic analysis are overviewed in Section 3.4. Some practical issues involved in Bayesian learning are reviewed in Section 3.5.

3.1 Bayesian Inference

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 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 the normalization constant that is also known as the marginal likelihood or the Bayesian 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,

$$p(x|X^n) = \int p(x|\theta)p(\theta|X^n)d\theta. \tag{3.1}$$

3.2 Kullback Information

The quantity of information related to a random variable x with a distribution $p(x)$ is often measured by the *entropy*, which is defined by

$$S = -\int p(x) \log p(x) dx.$$

If x is a discrete variable, the integral is replaced by the summation and the entropy is non-negative. If x is continuous, the entropy has no lower bound.

As the entropy measures the information content of a distribution, the *relative entropy* or the *Kullback information* $K(q(x)||p(x))$ is often used to measure the information for discriminating two distributions, $q(x)$ and $p(x)$. The Kullback information is defined by

$$K(q(x)||p(x)) = \int q(x) \log \frac{q(x)}{p(x)} dx, \tag{3.2}$$

if x is continuous, or by corresponding summation if x is discrete. The Kullback information is non-negative and equals to zero when $q(x) = p(x)$ almost everywhere as the proof is given in Appendix A.1. However, it is not symmetric as can be seen in eq.(3.2).

Hence, the generalization error of the predictive distribution, eq.(3.1), can be measured by the Kullback information from the true distribution,

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

3.3 Stochastic Complexity

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

$$F(X^n) = -\log Z(X^n), \tag{3.3}$$

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,

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_0(x_i)}{p(x_i|\theta)}, \tag{3.4}$$

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

$$\begin{aligned} F_0(X^n) &= -\log Z_0(X^n) \\ &= F(X^n) - S(X^n). \end{aligned} \tag{3.5}$$

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.(3.5) 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),

$$\begin{aligned} E_{X^n}[K(p_0(x)||p(x|X^n))] &= E_{X^{n+1}}[F(X^{n+1})] - E_{X^n}[F(X^n)] - S \\ &= E_{X^{n+1}}[F_0(X^{n+1})] - E_{X^n}[F_0(X^n)]. \end{aligned} \tag{3.6}$$

3.4 Asymptotic Analysis

Recently, in Bayesian learning, an advanced mathematical method for analyzing non-regular models was established (Watanabe, 2001a), 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,

$$E_{X^n}[F_0(X^n)] \simeq \lambda \log n - (m-1) \log \log n + O(1), \tag{3.7}$$

where λ and m are the rational number and the natural number respectively which are determined by the singularities of the true parameter. The derivation of eq.(3.7) is outlined in Appendix A.2. 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 because the Bayesian stochastic complexity corresponds to the cumulative loss of the Bayesian predictive distribution and the redundancy of the Bayesian method in coding (Clarke and Barron, 1990). From eq.(3.6), if the asymptotic form of the average normalized Bayesian stochastic complexity is given by eq.(3.7), the average generalization error is given by

$$E_{X^n}[K(p_0(x)||p(x|X^n))] \simeq \frac{\lambda}{n} + o(\frac{1}{n}). \tag{3.8}$$

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.

3.5 Practical Issues

The advantage of the Bayesian approach was mentioned in the previous section. There are, however, some practical issues involved in Bayesian learning. This section describes two issues, approximation methods and conjugate priors.

3.5.1 Approximation Schemes

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. The mean and covariance matrix of the approximating distribution are given by the MAP estimator,

$$\hat{\theta}_{\mathrm{MAP}} = \operatorname*{argmax}_{\theta} \varphi(\theta) \prod_{i=1}^{n} p(x_i|\theta),$$

and the inverse of the Fisher information matrix at the mean. This approach gives reasonable approximation in the case of regular statistical models whose Fisher information matrices are positive definite. In contrast, the Fisher information matrices of non-regular models degenerate around the true parameter. Also, the MAP solution can break down by overfitting the data. In other words, it tends to be the parameter with enormous posterior probability density, but the density is large over only a very small volume of the parameter space. This makes the MAP solution unrepresentative (Mackay, 2003, Ch.22). Therefore, the Laplace approximation can be insufficient for non-regular models.

In addition to the intractability of the integrals with respect to the posterior distribution, it also takes difficulties to sample from it directly. The *Markov chain Monte Carlo* (MCMC) methods are the most popular schemes that attempt to sample from the exact posterior distribution. The *Metropolis-Hastings algorithm* and the *Gibbs sampler* are the well-known prototypes of such methods(Metropolis et al., 1953). They have been improved based on the idea of the *extended ensemble* method and resulted in the methods named, *exchange Monte Carlo* and *multi-canonical sampling*(Iba, 2001). These methods were originally developed in the context of statistical physics. Another improved method, called the *Hybrid Monte Carlo* was in fact applied to neural network learning (Neal, 1996).

They generates the Markov chain that eventually converges in law to the true posterior distribution in principle but typically require vast computational resources. It is also hard to assess the convergence of the MCMC methods. One of the recent studies reported that the MCMC methods can be more erroneous for non-regular models than for regular ones (Nakano et al., 2005). It was also pointed out that the exchange Monte Carlo method has advantages in non-regular models though it is still computationally demanding (Nagata and Watanabe, 2006).

As another approximation, the variational Bayesian framework was proposed (Attias, 1999; Beal, 2003; Ghahramani and Beal, 2000) and it is elaborated in the next chapter.

3.5.2 Conjugate Prior

The conjugate prior distributions are often used in Bayesian learning or approximate Bayesian learning.

Given a model $p(x|\theta)$, the prior distribution $\varphi(\theta|\xi)$ parameterized by

the hyperparameter ξ is conjugate if the posterior distribution, $p(\theta|X^n) \propto p(X^n|\theta)\varphi(\theta|\xi)$, has the same functional form as $\varphi(\theta|\xi)$.

For example, consider the model $p(x|\mu)$ which is a 1-dimensional Gaussian distribution with the mean parameter μ,

$$p(x|\mu) = g(x|\mu, 1/\tau),$$

where $1/\tau$ is the variance assumed to be known and $g(x|\mu, \sigma^2)$ denotes the Gaussian density function with mean μ and variance σ^2. In this case, the conjugate prior distribution is Gaussian,

$$\varphi(\mu|\nu_0) = g(\mu|\nu_0, 1/\tau_0).$$

Given the training data $X^n = \{x_1, \cdots, x_n\}$, the posterior distribution also becomes Gaussian,

$$p(\mu|X^n) = g(\mu|\overline{\mu}, \overline{\sigma}^2),$$

where

$$\overline{\mu} = \frac{\tau \sum_{i=1}^n x_i + \tau_0}{n\tau + \tau_0}, \quad \text{and} \quad \overline{\sigma}^2 = \frac{1}{n\tau + \tau_0},$$

are the mean and the variance of the posterior distribution.

As demonstrated in this example, if the prior distribution is conjugate, the posterior distribution can easily be obtained by simply updating the hyperparameter of the conjugate prior. The conjugate prior makes the learning algorithm very simple in Bayesian learning.

Other examples of the conjugate prior distributions paired with the corresponding parameters are listed as follows: (Beta, parameter of binomial), (Dirichlet, parameter of multinomial), (Gamma, inverse variance parameter of 1-dimensional Gaussian) and (Wishart, inverse covariance matrix of multi-dimensional Gaussian). The beta distribution of the parameter $t \in [0 \;\; 1]$ of the binomial distribution defined in eq.(6.10) has the probability density function,

$$\varphi(t|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1},$$

where $\alpha > 0$ and $\beta > 0$ are the hyperparameters and

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

is the gamma function. The probability density functions of the Dirichlet and the gamma distributions are respectively given in eq.(5.3) and eq.(6.11).

The Wishart distribution is the joint distribution of the elements of $d \times d$ symmetric matrix $\mathbf{S}$ with the density function,

$$\varphi(\mathbf{S}|\eta, \mathbf{B}) = \frac{|\mathbf{B}|^{\eta/2}|\mathbf{S}|^{(\eta-d-1)/2}e^{-\frac{1}{2}\mathrm{tr}(\mathbf{SB})}}{2^{\eta d/2}\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma(\frac{\eta+1-i}{2})},$$

where η and the positive definite matrix $\mathbf{B}$ are the hyperparameters.

As discussed in Section 3.5.1, however, the integrals with respect to the posterior distribution are intractable for non-regular models. This means non-regular models do not have any appropriate conjugate prior distribution though this is not always the case.

In fact, the variational Bayesian framework takes advantage of the conjugacy by introducing hidden variables(Beal, 2003; Ghahramani and Beal, 2000).

Chapter 4

Variational Bayesian Learning

This chapter outlines the variational Bayesian framework in Section 4.1 and defines the variational stochastic complexity in Section 4.2. This framework, also known as mean field theory in statistical physics, originated from the general variational free energy approach. Other schemes to approximate the Bayesian posterior distribution are presented in Section 4.3.

4.1 Variational Bayesian Framework

The technique for approximating a complex distribution based on variational free energy minimization has been used in statistical physics (Feynman, 1972; Mackay, 2003, Ch.33). It was introduced to statistical data modelling in order to approximate the posterior distribution in learning of neural networks (Hinton and van Camp, 1993). The variational Bayesian framework described here is a generalization of such techniques and it can be applied to general graphical models with hidden variables (Attias, 1999).

By using the complete likelihood of the data $\{X^n, Y^n\}$, with the corresponding hidden variables $Y^n = \{y_1, \cdots, y_n\}$, the Bayesian stochastic complexity eq.(3.3) can be rewritten as follows,

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

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,

$$\begin{aligned} F(X^n) &\leq \sum_{Y^n} \int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} d\theta \\ &\equiv \overline{F}[q]. \end{aligned} \tag{4.1}$$

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 the objective function called the *variational free energy* and it measures the quality of the approximation.

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

$$q(Y^n, \theta|X^n) = Q(Y^n|X^n) r(\theta|X^n), \tag{4.2}$$

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.

Theorem 1 *If the functional $\overline{F}[q]$ is minimized under the constraint eq.(4.2) then the variational posteriors, $r(\theta|X^n)$ and $Q(Y^n|X^n)$, satisfy*

$$r(\theta|X^n) = \frac{1}{C_r} \varphi(\theta) \exp \langle \log p(X^n, Y^n|\theta) \rangle_{Q(Y^n|X^n)}, \tag{4.3}$$

and

$$Q(Y^n|X^n) = \frac{1}{C_Q} \exp \langle \log p(X^n, Y^n|\theta) \rangle_{r(\theta|X^n)}, \tag{4.4}$$

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

The proof is given in Appendix A.3.

Note that eq.(4.3) and eq.(4.4) give only the necessary conditions that $r(\theta|X^n)$ and $Q(Y^n|X^n)$ minimize the functional $\overline{F}[q]$. The variational posteriors that satisfy eq.(4.3) and eq.(4.4) are searched by an iterative algorithm. This algorithm is a natural gradient method (Amari, 1998) when the model is in the general exponential family of models with hidden variables and its on-line version has been derived (Sato, 2001).

4.2 Stochastic Complexity of Variational Bayes

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

$$\overline{F}(X^n) = \min_{r,Q} \overline{F}[q]. \tag{4.5}$$

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.

The above equation means that minimizing variational free energy gives the *e-projection* ($\alpha = 1$-projection) of the Bayesian posterior distribution to the set of the factorized distributions(Amari, 1985). Another interpretation of the variational stochastic complexity is obtained as the code length given by the coding scheme, named *bits-back coding* (Hinton and van Camp, 1993; Honkela and Valpola, 2004).

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

The normalized variational stochastic complexity $\overline{F}_0(X^n)$ is defined by

$$\overline{F}_0(X^n) = \overline{F}(X^n) - S(X^n). \tag{4.6}$$

From Theorem 1, the following lemma is obtained.

Lemma 1

$$\overline{F}_0(X^n) = \min_{r(\theta|X^n)} \{K(r(\theta|X^n)||\varphi(\theta)) - (\log C_Q + S(X^n))\}, \tag{4.7}$$

where

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

(Proof of Lemma 1)

From the restriction of the variational Bayesian approximation eq.(4.2), $\overline{F}(X^n)$ 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.(4.3) and eq.(4.4), 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.(4.7).**(Q.E.D)**

The variational posteriors $r(\theta|X^n)$ and $Q(Y^n|X^n)$ that satisfy eq.(4.3) and eq.(4.4) 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.(4.7) is also parameterized by $\overline{\theta}$. Therefore, henceforth $r(\theta|X^n)$ and C_Q are denoted as $r(\theta|\overline{\theta})$ and $C_Q(\overline{\theta})$ when they are regarded as functions of the variational parameter $\overline{\theta}$.

The variational estimator $\overline{\theta}_{vb}$ of θ is defined 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 1 claims that

$$\overline{\theta}_{vb} = \mathop{\rm argmin}_{\overline{\theta}} \{ K(r(\theta|\overline{\theta})||\varphi(\theta)) - (\log C_Q(\overline{\theta}) + S(X^n)) \}. \tag{4.8}$$

In variational Bayesian learning, the variational parameter $\overline{\theta}$ is updated iteratively to find the optimal solution $\overline{\theta}_{vb}$. Therefore, the asymptotic behavior of the variational stochastic complexities will be shown in Theorem 2 and Theorem 3 by evaluating the minimum value of the right hand side of eq.(4.8) as a function of the variational parameter $\overline{\theta}$.

4.3 Other Variational Approximations

In addition to the variational Bayesian approach, there are some other strategies for approximating a complex distribution.

One approach considers a lower bound $q(x)$ to the complex distribution $p(x)$ and optimizes the unnormalized density function $q(x)$. The lower bound $q(x)$ is chosen so that its normalization constant is maximized in order to gain the tightest fit to the distribution $p(x)$. This approach was introduced for approximating the posterior distributions of the logistic regression models (Jaakkola and Jordan, 2000).

Some other methods try to obtain the exact marginal distributions which usually become biased in the variational Bayesian approach. The properties of one of these methods, called *expectation propagation*, were investigated in several learning tasks (Minka, 2001).

Chapter 5

Variational Bayes for Gaussian Mixture Models

This chapter starts with the introduction to the Gaussian mixture models in Section 5.1. Then Theorem 2 is presented in Section 5.2, where the asymptotic form is obtained for the variational stochastic complexity of the Gaussian mixture models. The proof of Theorem 2 is given in Section 5.3.

5.1 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

$$p(x|\theta) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi\sigma_k^2}^M} \exp(-\frac{\|x-\mu_k\|^2}{2\sigma_k^2}), \tag{5.1}$$

where $\sigma_k > 0$ is a constant.

In this chapter, this type eq.(5.1) of the Gaussian mixture model is considered in the variational Bayesian framework and upper and lower bounds of the variational stochastic complexity will be shown in Theorem 2.

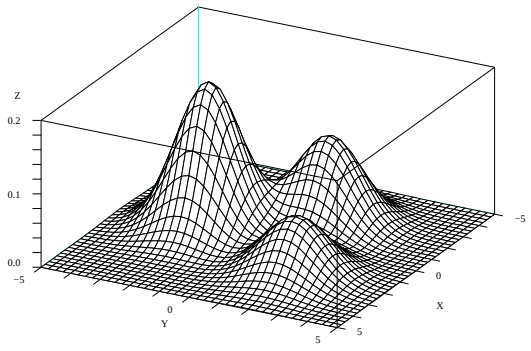


Figure 5.1: An example of the 2-dimensional Gaussian mixture model.

The Gaussian mixture model can be rewritten as follows by 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.

ML learning of the Gaussian mixture model is carried out by the expectation-maximization (EM) algorithm (Dempster et al., 1977), which corresponds to a clustering algorithm called the soft K-means (Mackay, 2003, Ch.22). The EM algorithm for ML and MAP learning is described in Appendix A.4.

The Gaussian mixture model is a non-regular statistical model, since the parameters are non-identifiable as demonstrated in Section 2.2. 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 (Akaho and Kappen, 2000; Yamazaki and Watanabe, 2003a). In particular, the Gaussian mixture model given by eq.(5.1) 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).

5.2 Variational Stochastic Complexity of Gaussian Mixture Models

In this section, we describe two conditions and give the upper and lower bounds of the normalized variational stochastic complexity in Theorem 2.

The following conditions are assumed.

(A5-1) 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,

$$p(x|\theta) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi}^M} \exp(-\frac{\|x-\mu_k\|^2}{2}), \tag{5.2}$$

and $K \geq K_0$ holds.

(A5-2) The prior of the parameters is the product of the following two distributions on $\mathbf{a} = \{a_k\}_{k=1}^K$ and $\mu = \{\mu_k\}_{k=1}^K$

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^K a_k^{\phi_0 - 1}, \tag{5.3}$$

$$\varphi(\mu) = \prod_{k=1}^K \sqrt{\frac{\xi_0}{2\pi}}^M \exp(-\frac{\xi_0 \|\mu_k - \nu_0\|^2}{2}), \tag{5.4}$$

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 2 *Assume the conditions (A5-1) and (A5-2). Then the normalized variational stochastic complexity $\overline{F}_0(X^n)$ defined by eq.(4.6) satisfies*

$$\underline{\lambda} \log n + nH_n(\overline{\theta}_{vb}) + C_1 \leq \overline{F}_0(X^n) \leq \overline{\lambda} \log n + C_2, \tag{5.5}$$

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

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

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

Remark. The constant C_2 in the above theorem is for example given by

$$C_2 = \begin{cases} C + C' & (\phi_0 > \frac{M+1}{2}), \\ C + C'' & (\phi_0 \leq \frac{M+1}{2}), \end{cases}$$

where the constants C, C' and C'' are indeed obtained in eqs.(5.10), (5.21) and (5.25) respectively.

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

Corollary 1 *Assume the conditions (A5-1) and (A5-2). 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 2 and Corollary 1,

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

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 average normalized stochastic complexity of regular statistical models, which coincides with the Bayesian information criterion (BIC) (Schwarz, 1978) and the minimum description length (MDL) (Rissanen, 1986), is given by $\lambda_{\rm BIC}\log n$ where

$$\lambda_{\rm BIC} = \frac{MK + K - 1}{2}. \tag{5.6}$$

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.(3.7) and eq.(3.8). Theorem 2 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 2 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 Chapter 8.

In the condition (A5-1), it is assumed 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, it is necessary to consider the case when the model is redundant.

5.3 Proof of Theorem 2

In this section, we prove Theorem 2. 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.(5.2).

5.3.1 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

$$\overline{y}_i^k = \langle y_i^k \rangle_{Q(Y^n|X^n)},$$

$$n_k = \sum_{i=1}^{n} \overline{y}_i^k$$

and

$$\nu_k = \frac{1}{n_k} \sum_{i=1}^{n} \overline{y}_i^k x_i,$$

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.(4.3) and the respective prior eq.(5.3) and eq.(5.4), 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,

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

and

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

where

$$\overline{a}_k = \frac{n_k + \phi_0}{n + K\phi_0},$$

$$\overline{\sigma}_k^2 = \frac{1}{n_k + \xi_0},$$

and

$$\overline{\mu}_k = \frac{n_k \nu_k + \xi_0 \nu_0}{n_k + \xi_0}.$$

From eq.(4.4), the variational posterior $Q(Y^n|X^n)$ is given by

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

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

5.3.2 Lemmas

Before proving Theorem 2, we show two lemmas where the two terms

$$K(r(\theta|\overline{\theta})||\varphi(\theta)) \quad \text{and} \quad (\log C_Q(\overline{\theta}) + S(X^n))$$

in Lemma 1 are respectively evaluated. In the proofs 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),

$$\frac{1}{2x} < \log x - \Psi(x) < \frac{1}{x}, \tag{5.7}$$

and

$$0 \leq \log \Gamma(x) - \{(x - \frac{1}{2}) \log x - x + \frac{1}{2} \log 2\pi\} \leq \frac{1}{12x}. \tag{5.8}$$

The inequalities (5.7) 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.(5.8) as well.

Lemma 2

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

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

$$G(\overline{\mathbf{a}}) = \frac{MK + K - 1}{2} \log n + \{\frac{M}{2} - (\phi_0 - \frac{1}{2})\} \sum_{k=1}^{K} \log \overline{a}_k. \tag{5.9}$$

Remark. The constant C in the above lemma is for example given by

$$\begin{aligned} C &= \frac{MK + K - 1}{2} K\phi_0 \\ &\quad + (K-1)\Big|\phi_0 - \frac{\log 2\pi}{2}\Big| + K \max\{1, \frac{1}{12\phi_0}\} + 1 + \frac{1}{12(1 + K\phi_0)} \\ &\quad + \left| \log \frac{\Gamma(\phi_0)^K}{\Gamma(K\phi_0)} \right| + \frac{MK}{2} \max\{|\log \phi_0|, |\log \xi_0|\} + \frac{MK}{2}. \end{aligned} \tag{5.10}$$

(Proof of Lemma 2)

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

$$K(r(\mathbf{a}|\overline{\mathbf{a}})||\varphi(\mathbf{a})) = \sum_{k=1}^{K} h(n_k) - n\Psi(n + K\phi_0) + \log \Gamma(n + K\phi_0) + \log \frac{\Gamma(\phi_0)^K}{\Gamma(K\phi_0)}, \tag{5.11}$$

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

$$K(r(\mu|\overline{\mu})||\varphi(\mu)) = \sum_{k=1}^{K} \frac{M}{2} \log \frac{n_k + \xi_0}{\xi_0} - \frac{KM}{2} + \frac{1}{2}\xi_0 \sum_{k=1}^{K} \{\frac{M}{n_k + \xi_0} + \|\overline{\mu}_k - \nu_0\|^2\}. \tag{5.12}$$

By using inequalities (5.7) and (5.8), we obtain

$$-1+\frac{12\phi_0-1}{12(x+\phi_0)} \leq h(x)+(\phi_0-\frac{1}{2})\log(x+\phi_0)-x-\phi_0+\frac{1}{2}\log 2\pi \leq 0. \quad (5.13)$$

Thus, from eqs.(5.11),(5.12),(5.13) and

$$K(r(\theta|\overline{\theta})||\varphi(\theta)) = K(r(\mathbf{a}|\overline{\mathbf{a}})||\varphi(\mathbf{a})) + K(r(\mu|\overline{\mu})||\varphi(\mu)),$$

it follows that

$$\begin{aligned}
&\Big|K(r(\theta|\overline{\theta})||\varphi(\theta)) - \Big\{G(\overline{\mathbf{a}}) + \frac{\xi_0}{2}\sum_{k=1}^{K}\|\overline{\mu}_k-\nu_0\|^2\Big\}\Big| \\
\leq\ & \frac{MK+K-1}{2}\log(1+\frac{K\phi_0}{n}) \\
&+(K-1)\Big|\phi_0-\frac{\log 2\pi}{2}\Big| + K + \sum_{k=1}^{K}\frac{|12\phi_0-1|}{12(n_k+\phi_0)} + \frac{12n+1}{12(n+K\phi_0)} \\
&+\left|\log\frac{\Gamma(\phi_0)^K}{\Gamma(K\phi_0)}\right| + \left|\sum_{k=1}^{K}\log(\frac{n_k+\xi_0}{n_k+\phi_0}) - \frac{MK}{2}(1+\log\xi_0) + \frac{\xi_0}{2}\sum_{k=1}^{K}\frac{M}{n_k+\xi_0}\right|.
\end{aligned}$$

The right hand side of the above inequality is bounded by a constant since

$$\frac{1}{n+\xi_0} < \frac{1}{n_k+\xi_0} < \frac{1}{\xi_0},$$

and

$$\frac{1}{n+\phi_0} < \frac{1}{n_k+\phi_0} < \frac{1}{\phi_0}.$$

(Q.E.D)

Lemma 3

$$\begin{aligned}
\log C_Q(\overline{\theta}) \ &= \ \sum_{i=1}^{n}\log\Big[\sum_{k=1}^{K}\frac{1}{\sqrt{2\pi}^M}\exp\{\Psi(n_k+\phi_0)-\Psi(n+K\phi_0) \\
&\quad -\frac{\|x_i-\overline{\mu}_k\|^2}{2} - \frac{M}{2}\frac{1}{n_k+\xi_0}\}\Big], \qquad (5.14)
\end{aligned}$$

and

$$nH_n(\overline{\theta})-\frac{n}{n+K\phi_0} \leq -(\log C_Q(\overline{\theta})+S(X^n)) \leq n\overline{H}_n(\overline{\theta})-\frac{n}{2(n+K\phi_0)}, \quad (5.15)$$

where $H_n(\overline{\theta})$ is given by eq.(3.4) and $\overline{H}_n(\overline{\theta})$ is defined by

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

(Proof of Lemma 3)

$$\begin{aligned} C_Q(\overline{\theta}) &= \prod_{i=1}^{n}\sum_{y_i} \exp\langle \log p(x_i,y_i|\theta)\rangle_{r(\theta|\overline{\theta})} \\ &= \prod_{i=1}^{n}\sum_{k=1}^{K} \frac{1}{\sqrt{2\pi}^M} \exp\{\Psi(n_k+\phi_0) - \Psi(n+K\phi_0) \\ &\qquad -\frac{\|x_i-\overline{\mu}_k\|^2}{2} - \frac{M}{2}\frac{1}{n_k+\xi_0}\}. \end{aligned} \tag{5.16}$$

Thus we have eq.(5.14).

Using again the inequalities (5.7), we obtain

$$\begin{aligned} -\log C_Q(\overline{\theta}) \le -\sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} \frac{\overline{a}_k}{\sqrt{2\pi}^M} \exp\{-\frac{\|x_i-\overline{\mu}_k\|^2}{2} - \frac{M+2}{2(n_k+\min\{\phi_0,\xi_0\})}\}\Big] \\ -\frac{n}{2(n+K\phi_0)}, \end{aligned} \tag{5.17}$$

and

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

which give the upper and lower bounds in eq.(5.15) respectively.**(Q.E.D)**

5.3.3 Upper and Lower Bounds

Now from the above lemmas, we prove Theorem 2 by showing the upper bound and the lower bound respectively.

(Proof of Theorem 2)

First we show the upper bound in eq.(5.5).

From Lemma 1, Lemma 2 and Lemma 3, it follows that

$$\overline{F}_0(X^n) \le \min_{\overline{\theta}} T_n(\overline{\theta}) + C, \tag{5.18}$$

where

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

From eq.(5.18), it is noted that the function values of $T_n(\overline{\theta})$ at specific points of the variational parameter $\overline{\theta}$ give the upper bounds of the normalized variational stochastic complexity $\overline{F}_0(X^n)$. Hence, let us consider following two cases.

(I) : Consider the case when all components including redundant ones are used to learn K_0 true components, that is,

$$\begin{aligned}
\overline{a}_k &= \frac{a_k^* n + \phi_0}{n + K\phi_0} \quad (1 \leq k \leq K_0 - 1), \\
\overline{a}_k &= \frac{a_{K_0}^* n/(K - K_0 + 1) + \phi_0}{n + K\phi_0} \quad (K_0 \leq k \leq K), \\
\overline{\mu}_k &= \mu_k^* \quad (1 \leq k \leq K_0 - 1), \\
\overline{\mu}_k &= \mu_{K_0}^* \quad (K_0 \leq k \leq K),
\end{aligned} \tag{5.19}$$

then

$$\begin{aligned}
& n\overline{H}_n(\overline{\theta}) \\
< & \sum_{i=1}^{n}\log p(x_i|\theta_0) - \sum_{i=1}^{n}\log\frac{n+\phi_0}{n+K\phi_0} \\
& -\sum_{i=1}^{n}\log\Big[\sum_{k=1}^{K_0-1}\frac{a_k^*}{\sqrt{2\pi}^M}\exp\{-\frac{\|x_i-\mu_k^*\|^2}{2} - \frac{M+2}{2(a_k^* n + \min\{\xi_0,\phi_0\})}\} \\
& + \frac{a_{K_0}^*}{\sqrt{2\pi}^M}\exp\{-\frac{\|x_i-\mu_k^*\|^2}{2} - \frac{M+2}{2(\frac{a_{K_0}^*}{K-K_0+1}n + \min\{\xi_0,\phi_0\})}\}\Big] \\
< & \sum_{i=1}^{n}\log\frac{\frac{n+K\phi_0}{n+\phi_0}p(x_i|\theta_0)}{p(x_i|\theta_0)\exp\{-\frac{(M+2)(K-K_0+1)}{2(\min_k\{a_k^*\}n+\min\{\xi_0,\phi_0\}(K-K_0+1))}\}} \\
< & \frac{(K-1)\phi_0 n}{n+\phi_0} + \frac{(M+2)(K-K_0+1)n}{2(\min_k\{a_k^*\}n + \min\{\xi_0,\phi_0\}(K-K_0+1))}, \\
\leq & (K-1)\phi_0 + \Big(\frac{M+2}{2}\Big)\frac{K-K_0+1}{\min_k\{a_k^*\}}
\end{aligned}$$

where the first inequality follows from $\frac{a_k^* n+\phi_0}{n+K\phi_0} > a_k^* \frac{n+\phi_0}{n+K\phi_0}$ and the third inequality follows from $\log(1+x) \leq x$ for $x > -1$.

It follows that

$$T_n(\overline{\theta}) < \frac{MK+K-1}{2}\log n + C', \tag{5.20}$$

where C' is a constant which is for example given by

$$\begin{aligned} C' &= \Big\{\frac{M}{2} - (\phi_0 - \frac{1}{2})\Big\}\Big\{\sum_{k=1}^{K_0-1} \log a_k^* + (K-K_0+1)\log\frac{a_{K_0}^*}{K-K_0+1} \\ &\quad + K\frac{(K-K_0+1)(1-\max_k\{a_k^*\}K)\phi_0}{\min_k\{a_k^*\}(1+K\phi_0)}\Big\} \\ &\quad + \frac{\xi_0}{2}\sum_{k=1}^{K_0-1}\|\mu_k^* - \nu_0\|^2 + \frac{\xi_0(K-K_0+1)}{2}\|\mu_{K_0}^* - \nu_0\|^2 \\ &\quad + (K-1)\phi_0 + \Big(\frac{M+2}{2}\Big)\frac{(K-K_0+1)}{\min_k\{a_k^*\}}. \end{aligned} \tag{5.21}$$

(II) : Consider the case when the redundant components are eliminated, that is,

$$\begin{aligned} \overline{a}_k &= \frac{a_k^* n+\phi_0}{n+K\phi_0} \quad (1 \leq k \leq K_0), \\ \overline{a}_k &= \frac{\phi_0}{n+K\phi_0} \quad (K_0+1 \leq k \leq K), \\ \overline{\mu}_k &= \mu_k^* \quad (1 \leq k \leq K_0), \\ \overline{\mu}_k &= \nu_0 \quad (K_0+1 \leq k \leq K), \end{aligned} \tag{5.22}$$

then

$$\begin{aligned} & n\overline{H}_n(\overline{\theta}) \\ < & \sum_{i=1}^n \log \frac{p(x_i|\theta_0)}{\frac{n+\phi_0}{n+K\phi_0}\sum_{k=1}^{K_0}\frac{a_k^*}{\sqrt{2\pi}^M}\exp\{-\frac{\|x_i-\mu_k^*\|^2}{2} - \frac{M+2}{2(a_k^* n+\min\{\xi_0,\phi_0\})}\}} \\ < & \frac{(K-1)\phi_0 n}{n+\phi_0} + \Big(\frac{M+2}{2}\Big)\frac{n}{\min_k\{a_k^*\}n + \min\{\xi_0,\phi_0\}} \\ \leq & (K-1)\phi_0 + \Big(\frac{M+2}{2}\Big)\frac{1}{\min_k\{a_k^*\}} \end{aligned} \tag{5.23}$$

holds. The first inequality follows from $\frac{a_k^* n+\phi_0}{n+K\phi_0} > a_k^* \frac{n+\phi_0}{n+K\phi_0}$ and

$$\sum_{k=K_0+1}^{K} \frac{\overline{a}_k}{\sqrt{2\pi}^M} \exp\{-\frac{\|x_i-\overline{\mu}_k\|^2}{2} - \frac{M+2}{2(n_k+\min\{\phi_0,\xi_0\})}\} > 0.$$

The second inequality follows from $\log(1+x) \leq x$ for $x > -1$.

It follows that

$$T_n(\overline{\theta}) < \{(K-K_0)\phi_0 + \frac{MK_0+K_0-1}{2}\}\log n + C'', \tag{5.24}$$

where C'' is a constant which is for example given by

$$\begin{aligned} C'' &= \Big\{\frac{M}{2} - (\phi_0 - \frac{1}{2})\Big\}\Big\{\sum_{k=1}^{K_0} \log a_k^* + (K-K_0)\log\phi_0 \\ &\quad + K_0 \frac{(1-\min_k\{a_k^*\}K)\phi_0}{\min_k\{a_k^*\}(1+K\phi_0)}\Big\} + \frac{\xi_0}{2}\sum_{k=1}^{K_0}\|\mu_k^* - \nu_0\|^2 \\ &\quad + (K-1)\phi_0 + \Big(\frac{M+2}{2}\Big)\frac{1}{\min_k\{a_k^*\}}. \end{aligned} \tag{5.25}$$

From eqs.(5.18), (5.20) and (5.24), we obtain the upper bound in eq.(5.5).

Next we show the lower bound in eq.(5.5). It follows from Lemma 1, Lemma 2 and Lemma 3,

$$\overline{F}_0(X^n) \geq \min_{\overline{\mathbf{a}}}\{G(\overline{\mathbf{a}})\} + nH_n(\overline{\theta}_{vb}) - C - 1. \tag{5.26}$$

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

$$G(\overline{\mathbf{a}}) \geq \frac{MK+K-1}{2}\log n - (\frac{M+1}{2} - \phi_0)K\log K, \tag{5.27}$$

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

$$\begin{aligned} G(\overline{\mathbf{a}}) &\geq \{(K-1)\phi_0 + \frac{M}{2}\}\log n + (\frac{M+1}{2} - \phi_0)(K-1)\log\frac{\phi_0 n}{n+K\phi_0} + C''' \\ &\geq \{(K-1)\phi_0 + \frac{M}{2}\}\log n + (\frac{M+1}{2} - \phi_0)(K-1)\log\frac{\phi_0}{1+K\phi_0} + C''', \end{aligned} \tag{5.28}$$

where C''' is a constant. The first inequality follows 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.(5.26),(5.27) and (5.28), we obtain the lower bound in eq.(5.5). **(Q.E.D)**

Chapter 6

Variational Bayes for Mixture of Exponential Families

In the previous chapter, the asymptotic bounds were obtained for the variational stochastic complexity of the Gaussian mixture models. In this chapter, we focus on variational Bayesian learning of more general mixture models, namely mixtures of exponential families that include mixtures of distributions such as Gaussian, binomial, and gamma.

This chapter starts with the introduction to the mixture of exponential families in Section 6.1. Then Theorem 3 is stated by generalizing Theorem 2 in the case of mixtures of exponential families in Section 6.2. Section 6.3 presents examples of mixture models where Theorem 3 applies. Theorem 3 is proven in the last section of this chapter.

6.1 Mixture of Exponential Family

Denote by $c(x|b)$ a probability density function of the input $x \in R^N$ given an M-dimensional parameter vector $b = (b^{(1)}, b^{(2)}, \cdots, b^{(M)})^T \in B$ where B is a subset of R^M. The general mixture model $p(x|\theta)$ with a parameter vector θ is defined by

$$p(x|\theta) = \sum_{k=1}^{K} a_k c(x|b_k),$$

where 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 $\theta = \{a_k, b_k\}_{k=1}^{K}$.

A model $p(x|\theta)$ is called a mixture of the exponential family (MEF) model or exponential family mixture model if the probability distribution $c(x|b)$ is given by

$$c(x|b) = \exp\{b \cdot f(x) + f_0(x) - g(b)\}, \tag{6.1}$$

where $b \in B$ is called the natural parameter, $b \cdot f(x)$ is its inner product with the vector $f(x) = (f_1(x), \cdots, f_M(x))^T$, $f_0(x)$, and $g(b)$ are real-valued functions of the input x and the parameter b, respectively (Brown, 1986). Suppose functions $f_1, \cdots, f_M$ and the constant function, 1, are linearly independent and the effective number of parameters in a single component distribution, $c(x|b)$, is M. Basic properties of the exponential family are summarized in Appendix A.5.

The conjugate prior distribution $\varphi(\theta)$ for the mixture of the exponential family model is defined by the product of the following two distributions on $\mathbf{a} = \{a_k\}_{k=1}^K$ and $\mathbf{b} = \{b_k\}_{k=1}^K$

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^K a_k^{\phi_0 - 1}, \tag{6.2}$$

$$\begin{aligned} \varphi(\mathbf{b}) &= \prod_{k=1}^K \varphi(b_k) \\ &= \prod_{k=1}^K \frac{1}{C(\xi_0, \nu_0)} \exp\{\xi_0(b_k \cdot \nu_0 - g(b_k))\}, \end{aligned} \tag{6.3}$$

where the function $C(\xi, \mu)$ of $\xi \in R$ and $\mu \in R^M$ is defined by

$$C(\xi, \mu) = \int \exp\{\xi(\mu \cdot b - g(b))\} db. \tag{6.4}$$

Constants $\xi_0 > 0$, $\nu_0 \in R^M$ and $\phi_0 > 0$ are the hyperparameters.

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

$$p(x, y|\theta) = \prod_{k=1}^K \Big[a_k c(x|b_k)\Big]^{y^k}.$$

Even though the exponential family distributions are regular statistical models, the mixtures of them are non-regular. It is obvious that the non-identifiability arises and the singularities occur in the mixture of exponential families as the simple example of the Gaussian mixture model presented in Section 2.2.

6.2 Variational Stochastic Complexity of Mixture of Exponential Families

The following conditions are assumed.

(A6-1) The true distribution $p_0(x)$ is represented by a mixture of the exponential family model $p(x|\theta_0)$, which has K_0 components and the parameter $\theta_0 = \{a_k^*, b_k^*\}_{k=1}^{K_0}$,

$$p(x|\theta_0) = \sum_{k=1}^{K_0} a_k^* \exp\{b_k^* \cdot f(x) + f_0(x) - g(b_k^*)\},$$

where $b_k^* \in R^M$ and $b_k^* \neq b_j^* (k \neq j)$. Also, assume that the true distribution can be achieved with the model, that is, the model $p(x|\theta)$ has K components,

$$p(x|\theta) = \sum_{k=1}^{K} a_k \exp\{b_k \cdot f(x) + f_0(x) - g(b_k)\},$$

and $K \geq K_0$ holds.

(A6-2) The prior distribution of the parameters is the conjugate prior $\varphi(\theta) = \varphi(\mathbf{a})\varphi(\mathbf{b})$, where $\varphi(\mathbf{a})$ and $\varphi(\mathbf{b})$ are given by eqs.(5.3) and (6.3) with hyperparameters ϕ_0, ξ_0 and ν_0. The prior distribution $\varphi(\mathbf{b})$ is bounded.

(A6-3) Regarding the distribution $c(x|b)$ of each component, the Fisher information matrix

$$I(b) = \frac{\partial^2 g(b)}{\partial b \partial b}$$

satisfies $0 < |I(b)| < +\infty$, for an arbitrary $b \in B^1$. The function $\mu \cdot b - g(b)$ has a stationary point at $\hat{b}$ in the interior of B for each $\mu \in \left\{ \frac{\partial g(b)}{\partial b} | b \in B \right\}$.

The following theorem will be proven under these conditions. The proof will appear in the next section.

[1]The notation $\frac{\partial^2 g(b)}{\partial b \partial b}$ is used for a matrix whose ijth entry is $\frac{\partial^2 g(b)}{\partial b^{(i)} \partial b^{(j)}}$ and $|\cdot|$ denotes the determinant of a matrix.

Theorem 3 *Assume the conditions (A6-1), (A6-2), and (A6-3). Then the normalized stochastic complexity* $\overline{F}_0(X^n)$ *defined by eq.(4.6) satisfies*

$$\underline{\lambda}\log n + nH_n(\overline{\theta}_{vb}) + C_1 \leq \overline{F}_0(X^n) \leq \overline{\lambda}\log n + C_2, \tag{6.5}$$

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*

$$\underline{\lambda} = \begin{cases} (K-1)\phi_0 + \frac{M}{2} & (\phi_0 \leq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}), \end{cases} \tag{6.6}$$

$$\overline{\lambda} = \begin{cases} (K-K_0)\phi_0 + \frac{MK_0+K_0-1}{2} & (\phi_0 \leq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}). \end{cases} \tag{6.7}$$

In this theorem, $nH_n(\overline{\theta}_{vb})$ is a training error. Let $\hat{\theta}$ be the maximum likelihood estimator, then it immediately follows from eq.(3.4) that

$$nH_n(\overline{\theta}_{vb}) \geq nH_n(\hat{\theta}), \tag{6.8}$$

where $nH_n(\hat{\theta}) = \min_\theta \sum_{i=1}^n \log \frac{p(x_i|\theta_0)}{p(x_i|\theta)}$ is the (maximum) likelihood ratio statistic with sign inversion. If the average of the likelihood ratio statistic is a bounded function of n, then it immediately follows from eq.(6.8) and Theorem 3 that

$$\underline{\lambda}\log n + O(1) \leq \overline{F}_0(n) \leq \overline{\lambda}\log n + O(1), \tag{6.9}$$

where $O(1)$ is a bounded function of n. It was proven that the likelihood ratio statistics of some non-regular models diverge to infinity as n increases. Some known facts about the divergence of the likelihood ratio are described in the examples below. Note, however, that even if $nH_n(\hat{\theta})$ diverges to minus infinity, eq.(6.8) does not necessarily mean $nH_n(\overline{\theta}_{vb})$ diverges in the same order.

6.3 Examples

The following are examples where Theorem 3 applies.

Example 1 (binomial). Consider a mixture of binomial component distributions. Each component has a one-dimensional parameter $t \in [0,\ 1]$,

$$c(x = k|t) = \binom{L}{k} t^k (1-t)^{L-k}, \tag{6.10}$$

where L is the number of Bernoulli trials and $k = 0, 1, 2, \cdots, L$. Hence, $M = 1$ and the natural parameter is given by $b = \log \frac{t}{1-t}$. Since the discrete variable x takes values in a finite set in this case, the (minus) likelihood ratio statistic is bounded below by some constant independent of n and is therefore $E_{X^n}[nH_n(\overline{\theta}_{vb})]$ from eq.(6.8). Then, eq.(6.9) holds where $\underline{\lambda}$ and $\overline{\lambda}$ are given by eqs.(6.6) and (6.7) with $M = 1$. Binomial mixture models are used for the gene analysis (Chernoff and Lander, 1995).

Example 2 (gamma). Consider the gamma component with shape parameter $\nu > 0$ and scale parameter $s > 0$,

$$c(x|\nu, s) = \frac{1}{s^\nu \Gamma(\nu)} x^{\nu-1} \exp(-\frac{x}{s}), \tag{6.11}$$

where $0 \leq x < \infty$. The natural parameter b is given by $b^{(1)} = 1/s$ and $b^{(2)} = \nu - 1$. Hence, eq.(6.5) holds where $\underline{\lambda}$ and $\overline{\lambda}$ are given by eqs.(6.6) and (6.7) with $M = 2$. When shape parameter ν is known, the likelihood ratio diverges in the order of $\log \log n$ (Liu et al., 2003). This implies that $E_{X^n}[nH_n(\overline{\theta}_{vb})] = O(\log \log n)$ from eq.(6.8). Gamma mixture models are applied to the image analysis (Agusta and Dowe, 2003).

Example 3 (Gaussian). Consider the N-dimensional Gaussian component with mean μ and covariance matrix Σ,

$$c(x|\mu, \Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}.$$

The natural parameter b is given by $\mu^T\Sigma^{-1}$ and Σ^{-1}. These are functions of the elements of μ and the upper half of Σ^{-1}. Hence, eq.(6.5) holds where $\underline{\lambda}$ and $\overline{\lambda}$ are given by eqs.(6.6) and (6.7) with $M = N + N(N+1)/2$. The covariance matrix Σ in some applications is assumed to be known and the parameter is restricted to mean μ. It is conjectured in this case that the likelihood ratio diverges in the order of $\log \log n$(Hartigan, 1985). This suggests that $E_{X^n}[nH_n(\overline{\theta}_{vb})] = O(\log \log n)$ from eq.(6.8). Applications of Gaussian mixture models are too numerous to mention.

Other than these examples, Theorem 3 applies to mixtures of distributions such as multinomial, Poisson and Weibull.

6.4 Proof of Theorem 3

In this section, Theorem 3 is proved by the similar way to Theorem 2.

6.4.1 Variational Posterior for Mixture of Exponential Family Model

In this subsection, we derive the variational posterior $r(\theta|X^n)$ for the mixture of the exponential family model using eq.(4.3) and then define the variational parameter and the variational estimator for this model.

Using the complete data $\{X^n, Y^n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, we put

$$\overline{y}_i^k = \langle y_i^k \rangle_{Q(Y^n)},$$

$$n_k = \sum_{i=1}^{n} \overline{y}_i^k,$$

and

$$\nu_k = \frac{1}{n_k} \sum_{i=1}^{n} \overline{y}_i^k f(x_i),$$

where $y_i^k = 1$ if the ith datum x_i is from the kth component, otherwise, $y_i^k = 0$. The variable n_k is the expected number of times data come from the kth component. Note that the variables n_k and ν_k satisfy the constraints $\sum_{k=1}^K n_k = n$ and $\sum_{k=1}^K n_k \nu_k = \sum_{i=1}^n f(x_i)$. From eq.(4.3) and the respective prior distributions, eqs.(5.3) and (6.3), the variational posterior $r(\theta|X^n) = r(\mathbf{a}|X^n) r(\mathbf{b}|X^n)$ is obtained as the product of the following two distributions,

$$r(\mathbf{a}|X^n) = \frac{\Gamma(n + K\phi_0)}{\prod_{k=1}^K \Gamma(n_k + \phi_0)} \prod_{k=1}^K a_k^{n_k + \phi_0 - 1}, \tag{6.12}$$

$$\begin{aligned} r(\mathbf{b}|X^n) &= \prod_{k=1}^K r(b_k|X^n) \\ &= \prod_{k=1}^K \frac{1}{C(\gamma_k, \overline{\mu}_k)} \exp\{\gamma_k(\overline{\mu}_k \cdot b_k - g(b_k))\}, \end{aligned} \tag{6.13}$$

where

$$\overline{\mu}_k = \frac{n_k \nu_k + \xi_0 \nu_0}{n_k + \xi_0},$$

and

$$\gamma_k = n_k + \xi_0.$$

Let

$$\overline{a}_k = \langle a_k \rangle_{r(\mathbf{a}|X^n)} = \frac{n_k + \phi_0}{n + K\phi_0}, \tag{6.14}$$

and

$$\overline{b}_k = \langle b_k \rangle_{r(b_k|X^n)} = \frac{1}{\gamma_k} \frac{\partial \log C(\gamma_k, \overline{\mu}_k)}{\partial \overline{\mu}_k}. \tag{6.15}$$

Note that

$$\frac{\phi_0}{n + K\phi_0} \leq \overline{a}_k \leq 1 - \frac{(K-1)\phi_0}{n + K\phi_0}$$

holds. Then the variational parameter $\overline{\theta}$ is defined by

$$\overline{\theta} = \langle \theta \rangle_{r(\theta|X^n)} = \{\overline{a}_k, \overline{b}_k\}_{k=1}^K. \tag{6.16}$$

It needs to be noted that $\overline{b}_k$ is the expectation parameter of b_k with the variational posterior $r(b_k|X^n)$. It also needs to be noted that the variational posterior $r(\theta|X^n)$ and C_Q in eq.(4.4) are parameterized by the variational parameter $\overline{\theta}$.

6.4.2 Lemmas

Since the variational posterior satisfies $r(\theta|\overline{\theta}) = r(\mathbf{a}|\overline{\mathbf{a}})r(\mathbf{b}|\overline{\mathbf{b}})$, we have

$$K(r(\theta|\overline{\theta})||\varphi(\theta)) = K(r(\mathbf{a}|\overline{\mathbf{a}})||\varphi(\mathbf{a})) + \sum_{k=1}^K K(r(b_k|\overline{b}_k)||\varphi(b_k)). \tag{6.17}$$

The following lemma is used for evaluating $K(r(b_k|\overline{b}_k)||\varphi(b_k))$ in the case of the mixture of exponential families.

Lemma 4

$$K(r(b_k|\overline{b}_k)||\varphi(b_k)) = \frac{M}{2}\log(n_k + \xi_0) - \log\varphi(\overline{b}_k) + O(1).$$

(Proof of Lemma 4)

Using the variational posterior, eq.(6.13), we obtain

$$K(r(b_k|\overline{b}_k)||\varphi(b_k)) = -\log\frac{C(\gamma_k, \overline{\mu}_k)}{C(\xi_0, \nu_0)} + n_k\{\nu_k\langle b_k\rangle_{r(b_k|\overline{b}_k)} - \langle g(b_k)\rangle_{r(b_k|\overline{b}_k)}\}, \tag{6.18}$$

where we used $\gamma_k = n_k + \xi_0$. Let us now evaluate the value of $C(\gamma_k, \overline{\mu}_k)$ when γ_k is sufficiently large. From Condition (A6-3), using the saddle point approximation, we obtain

$$C(\gamma_k, \overline{\mu}_k) = \exp\Big[\gamma_k\{\overline{\mu}_k \cdot \hat{b}_k - g(\hat{b}_k)\}\Big]\sqrt{\frac{2\pi}{\gamma_k}}^M \sqrt{|I(\hat{b}_k)|}^{-1}\Big\{1 + O(\frac{1}{\gamma_k})\Big\}, \quad (6.19)$$

where $\hat{b}_k$ is the maximizer of the function $\overline{\mu} \cdot b_k - g(b_k)$, that is,

$$\frac{\partial g(\hat{b}_k)}{\partial b_k} = \overline{\mu}_k.$$

Therefore, $-\log C(\gamma_k, \overline{\mu}_k)$ is evaluated as

$$-\log C(\gamma_k, \overline{\mu}_k) = \frac{M}{2}\log\frac{\gamma_k}{2\pi} + \frac{1}{2}\log|I(\hat{b}_k)| - \gamma_k(\overline{\mu}_k \cdot \hat{b}_k - g(\hat{b}_k)) + O(\frac{1}{\gamma_k}). \quad (6.20)$$

Applying the saddle point approximation to

$$b_k - \hat{b}_k = \frac{1}{C(\gamma_k, \overline{\mu}_k)}\int (b_k - \hat{b}_k)\exp\{\gamma_k(\overline{\mu}_k \cdot b_k - g(b_k))\}db,$$

we obtain

$$||\overline{b}_k - \hat{b_k}|| \leq \frac{A'}{\gamma_k} + O(\frac{1}{\gamma_k\sqrt{\gamma_k}}), \quad (6.21)$$

where A' is a constant. Since

$$g(b_k) - g(\hat{b}_k) = (b_k - \hat{b}_k)\overline{\mu}_k + \frac{1}{2}(b_k - \hat{b}_k)^T I(b_k^*)(b_k - \hat{b}_k), \quad (6.22)$$

for some point b_k^* on the line segment between b_k and $\hat{b}_k$, we have

$$g(\overline{b}_k) - g(\hat{b}_k) = (\overline{b}_k - \hat{b}_k)\overline{\mu}_k + O(\frac{1}{\gamma_k^2}), \quad (6.23)$$

and applying the saddle point approximation we obtain

$$\langle g(b_k)\rangle_{r(b_k|\overline{b}_k)} - g(\hat{b}_k) = (\overline{b}_k - \hat{b}_k)\overline{\mu}_k + \frac{M}{2\gamma_k} + O(\frac{1}{\gamma_k\sqrt{\gamma_k}}). \quad (6.24)$$

From eqs.(6.23) and (6.24)

$$\langle g(b_k)\rangle_{r(b_k|\overline{b}_k)} - g(\overline{b}_k) = \frac{M}{2\gamma_k} + O(\frac{1}{\gamma_k\sqrt{\gamma_k}}). \quad (6.25)$$

Thus, from eqs.(6.18), (6.23), (6.24), and (6.20), we obtain the lemma. **(Q.E.D)**

Lemma 2 and Lemma 3 are substituted by the following lemmas.

Lemma 5

$$\left| K(r(\theta|\overline{\theta})||\varphi(\theta)) - G(\overline{\mathbf{a}}) + \sum_{k=1}^{K} \log \varphi(\overline{b}_k) \right| \leq C \tag{6.26}$$

holds where C is a constant and the function $G(\overline{\mathbf{a}})$ is defined by eq.(5.9).

(Proof of Lemma 5)

From eqs.(5.11), (6.17), (5.13) and Lemma 4,

$$\left| K(r(\theta|\overline{\theta})||\varphi(\theta)) - G(\overline{\mathbf{a}}) + \sum_{k=1}^{K} \log \varphi(\overline{b}_k) \right|$$

is bounded above by a constant since

$$\frac{1}{n+\xi_0} < \frac{1}{n_k+\xi_0} < \frac{1}{\xi_0}.$$

(Q.E.D)

Lemma 6

$$nH_n(\overline{\theta}) + O(1) \leq -(\log C_Q(\overline{\theta}) + S(X^n)) \leq n\overline{H}_n(\overline{\theta}) + O(1) \tag{6.27}$$

holds where the function $H_n(\theta)$ is defined by eq.(3.4) and

$$\overline{H}_n(\overline{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta_0)}{\sum_{k=1}^{K} \overline{a}_k c(x_i|\overline{b}_k) \exp(-\frac{A}{n_k+\min\{\phi_0,\xi_0\}})},$$

where A is a constant.

(Proof of Lemma 6)

$$\begin{aligned} C_Q(\overline{\theta}) &= \prod_{i=1}^{n}\sum_{k=1}^{K} \exp\langle \log a_k c(x_i|b_k)\rangle_{r(\theta|\overline{\theta})} \\ &= \prod_{i=1}^{n}\sum_{k=1}^{K} \exp\{\Psi(n_k+\phi_0) - \Psi(n+K\phi_0) \\ &\qquad +\overline{b}_k \cdot f(x_i) - \langle g(b_k)\rangle_{r(\theta|\overline{\theta})} + f_0(x_i)\}. \end{aligned}$$

Again, using the inequalities in eqs.(5.7) and (6.25), we obtain

$$\log C_Q(\overline{\theta}) \geq \sum_{i=1}^{n} \log[\sum_{k=1}^{K} \overline{a}_k c(x_i|\overline{b}_k) \exp\{-\frac{M+2}{2(n_k + \min\{\phi_0, \xi_0\})} + O(n_k^{-\frac{3}{2}})\}] + O(1),$$

$$\log C_Q(\overline{\theta}) \leq \sum_{i=1}^{n} \log[\sum_{k=1}^{K} \overline{a}_k c(x_i|\overline{b}_k)] + O(1),$$

which give the upper and lower bounds in eq.(6.27), respectively. **(Q.E.D)**

6.4.3 Proof

(Proof of Theorem 3)

Since the prior distribution $\varphi(\mathbf{b})$ is bounded, from Lemma 5 and Lemma 6, we complete the proof in the same way as in Section 5.3.3. **(Q.E.D)**.

Chapter 7

Experiments

This chapter presents the results of the experiments where variational Bayesian learning is simulated for the Gaussian mixture model given by eq.(5.1). Variational Bayesian learning involves the iterative updates practically. Hence, the practical algorithm may converge to local minima. The properties of the practical variational Bayesian algorithm are investigated by comparing the experimental results with the theoretical ones presented in the previous chapters. The result of the first experiment is presented in Section 7.1 to discuss the dependency of the variational Bayesian algorithm on the number of components. This is followed by the result of the second experiment in Section 7.2 to discuss the effect of the hyperparameter.

7.1 Dependency on Model Size

In the first experiment, Gaussian mixture models with different number of components ($K = 1, 2, 3, 4, 5$) were trained by using $M = 1$ and $M = 10$ dimensional synthetic data. 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 the Gaussian mixture model,

$$p(x|\theta_0) = \frac{a_1^*}{\sqrt{2\pi}^M} \exp(-\frac{\|x - \mu_1^*\|^2}{2}) + \frac{a_2^*}{\sqrt{2\pi}^M} \exp(-\frac{\|x - \mu_2^*\|^2}{2}), \qquad (7.1)$$

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.(4.8), 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.(4.8)) 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

$$\lambda_{\rm VB} = (\overline{F}_0(X^{1000}) - \overline{F}_0(X^{100}))/\log 10 \tag{7.2}$$

to estimate the coefficient of the leading term of the normalized variational stochastic complexity $\overline{F}_0(X^n)$. The values of $\lambda_{\rm VB}$ were averaged over 100 draws of sample sets. The results of the averages of $\lambda_{\rm VB}$ and the coefficient $\overline{\lambda}$ given by eq.(6.7) are presented in Figure 7.1 against the number K of components for the case of (a) $M = 1$ and (b) $M = 10$ respectively. In these figures, an upper bound of the coefficient of the Bayesian stochastic complexity and $\lambda_{\rm BIC}$ given by eq.(5.6) are also plotted for the comparison of variational Bayesian learning with true Bayesian learning in the next chapter. The variational Bayesian algorithm gave $\lambda_{\rm VB}$ that coincide with the coefficient $\overline{\lambda}$. This implies the upper bound in eq.(6.5) 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

$$\lambda_{\rm G} = \frac{n}{n'} \sum_{i=1}^{n'} \log \frac{p(x_i'|\theta_0)}{\langle p(x_i'|\theta)\rangle_{r(\theta|X^n)}}, \tag{7.3}$$

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 7.1. 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.(3.7) and eq.(3.8), $\lambda_{\rm VB}$ 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 $E_{X^n}[\overline{F}(X^n)]$.

7.2 Dependency on Hyperparameter

In the second experiment, to investigate the effect of the hyperparameter ϕ_0, we calculated the average variational stochastic complexities ($\lambda_{\rm VB}$ in eq.(7.2)) of the $M = 2$ dimensional Gaussian mixture model with $K = 4$ components trained by the variational Bayes algorithm for various values of the hyperparameter ϕ_0. We used the training data sets generated by the true distribution defined by eq.(7.1) and calculated the values of $\lambda_{\rm VB}$ in the same way as the above. Figure 7.2 shows an example of the data sets.

The hyperparameters except for ϕ_0 were set at $\nu_0 = (0, 0)^T$ and $\xi_0 = 1.0$. The averages of $\lambda_{\rm VB}$ are presented in Figure 7.3 for two different types of the initial values of the variational parameter that are,

$$
\begin{aligned}
&(1): \quad \overline{a}_k = 1/K, \ \ \overline{\mu}_k = (0,0)^T, \quad (k = 1,2,3,4), \\
&(2): \quad \overline{a}_1 = \overline{a}_2 = 1/2, \overline{a}_3 = \overline{a}_4 = 0, \\
&\qquad \overline{\mu}_1 = (\sqrt{2}, \sqrt{2})^T, \overline{\mu}_2 = (-\sqrt{2}, -\sqrt{2})^T, \overline{\mu}_3 = \overline{\mu}_4 = (0,0)^T.
\end{aligned}
$$

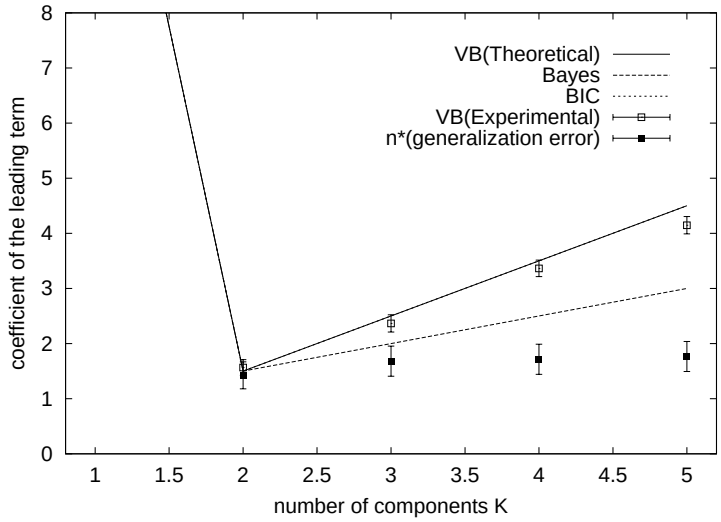
The generalization errors were also calculated and averages of them are presented in Figure 7.4.

As will be discussed in Chapter 8, Theorem 2 shows how the hyperparameter ϕ_0 influences the process of variational Bayesian learning. More specifically, only when $\phi_0 \leq (M + 1)/2$, the prior distribution eliminates the redundant components and otherwise it uses all the components. We can see in Figure 7.3, that when ϕ_0 is above $\frac{M+1}{2} = \frac{3}{2}$, the averages of $\lambda_{\rm VB}$ larger than the theoretical upper bound $\overline{\lambda}$ were obtained in the case of the initial value (2). This is due to the local minima since ϕ_0 larger than $\frac{M+1}{2}$ is not appropriate for the initial value (2) where two redundant components were reduced at the beginning of the learning process. However, if the initial value

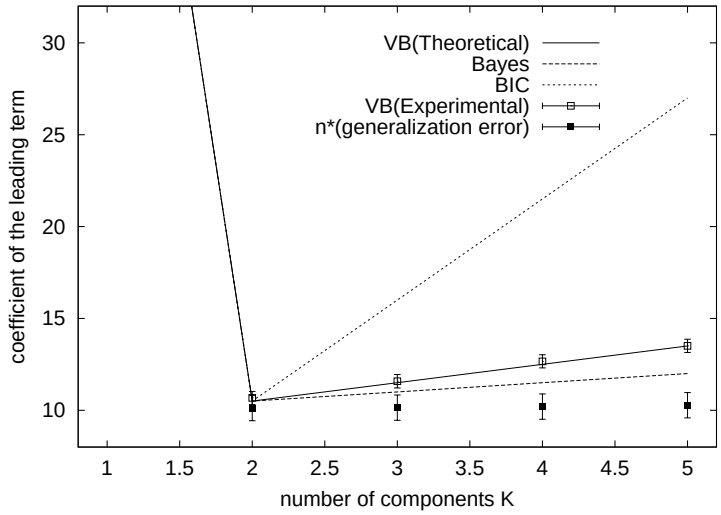
was set properly, the experimental results nearly coincide with the theoretical upper bound $\overline{\lambda}$. This implies the upper bound in eq.(5.5) is tight.

As can be seen in Figure 7.3 and Figure 7.4, λ_{VB} and λ_{G} showed different behavior, although the smaller the variational stochastic complexity (λ_{VB}), the smaller the generalization error (λ_{G}).

All these results imply that the appropriate initial value of the variational Bayesian algorithm can be set according to the hyperparameter ϕ_0.



(a) M=1



(b) M=10

Figure 7.1: 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.(6.7), the dashed line is the upper bound of λ in true Bayesian learning eq.(8.3) and the dotted line is $\lambda_{\rm BIC}$ of the BIC eq.(5.6). The open squares with error bars are the results of the averages of $\lambda_{\rm VB}$ eq.(7.2) and the full squares with error bars are the results of the averages of $\lambda_{\rm G}$ eq.(7.3). The error bars show 95% confidence intervals.

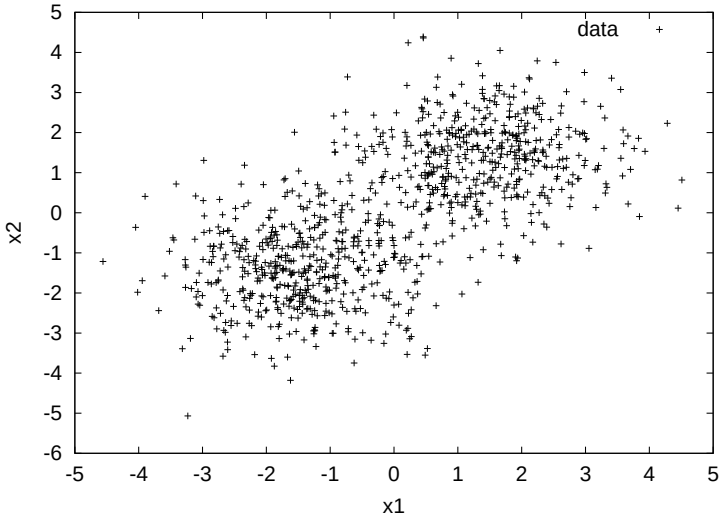


Figure 7.2: An example of the synthetic data sets, $M = 2$.

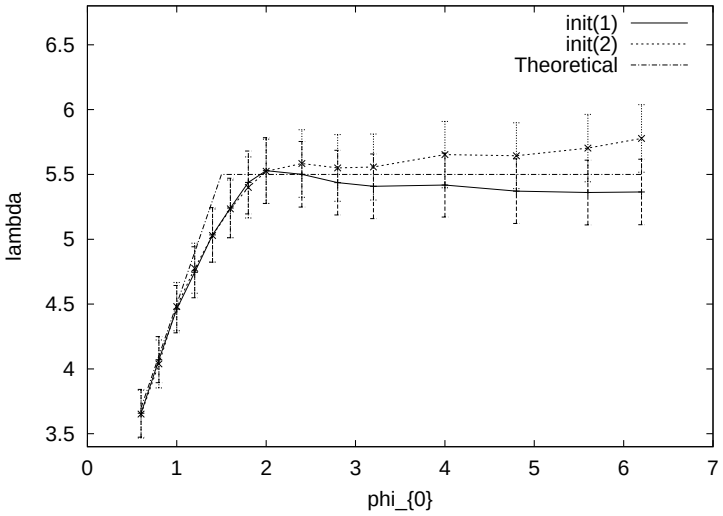


Figure 7.3: The average of λ_{VB} against the hyperparameter ϕ_0 for the two types (1) (solid line), (2) (dotted line) of initial values of the variational parameter and the theoretical upper bound $\overline{\lambda}$ (dashed line).

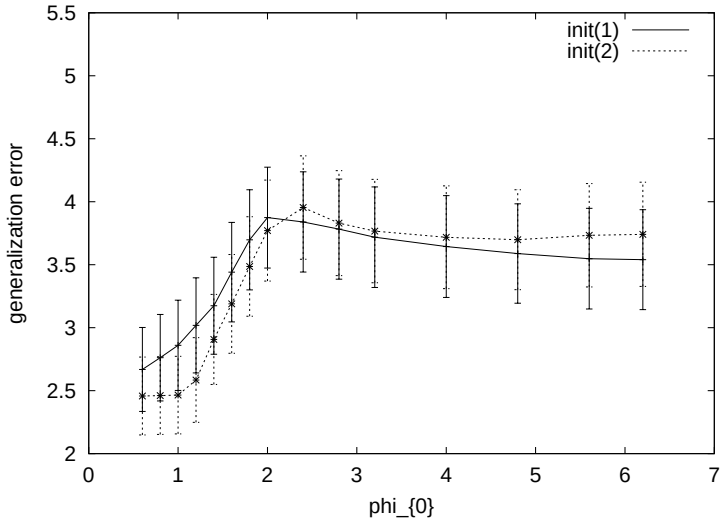


Figure 7.4: Generalization error against the hyperparameter ϕ_0 for the two types (1) (solid line), (2) (dotted line) of initial values of the variational parameter.

Chapter 8

Discussion

In this thesis, we showed upper and lower bounds of the variational stochastic complexity of the mixture models. Six topics are discussed in this chapter. The lower bound of the variational stochastic complexity is discussed in Section 8.1. The variational stochastic complexity is compared to the Bayesian stochastic complexity in Section 8.2 to examine the accuracy of the variational Bayesian approximation. The relation between the variational stochastic complexity and the generalization error is discussed in Section 8.3. Discussion on model selection is presented in Section 8.4. The effect of hyperparameters is pointed out in Section 8.5. Finally, some examples of the applications of the theoretical bounds in Theorem 2 and Theorem 3 are considered in Section 8.6.

8.1 Lower Bound

Let us discuss the lower bound. The lower bounds in eq.(5.5) and eq.(6.5) can be improved to give

$$\overline{F}_0(X^n) \geq \overline{\lambda} \log n + nH_n(\overline{\theta}_{vb}) + C_1, \tag{8.1}$$

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 2. The consistency means that the variational estimator converges to a parameter in the set of the true parameter, $\{\theta | p(x|\theta) = p(x|\theta_0)\}$, with probability 1 as the sample size n is sufficiently large. We conjecture that the variational estimator is consistent and the inequality (8.1) holds for some mixture

models. 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 sides of eqs.(5.5) and (6.5), $nH_n(\overline{\theta}_{vb})$ 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.(5.5) is tight. More specifically, if eq.(8.1) 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 1 that

$$E_{X^n}[\overline{F}_0(X^n)]/\log n \to \overline{\lambda} \quad (n \to \infty). \tag{8.2}$$

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

As discussed in Section 6.3, $E_{X^n}[\min_\theta nH_n(\theta)]$ is actually bounded or $o(\log n)$ for some mixture components. This implies the upper bound in eq.(6.5) is tight as well and eq.(8.2) holds for some mixture models.

8.2 Comparison to Bayesian Learning

Let us compare the normalized variational stochastic complexities shown in Theorem 2 and Theorem 3 with the one in true Bayesian learning, assuming eq.(8.2) holds. The Bayesian stochastic complexities of several non-regular models have been clarified in some recent studies. The following condition is assumed about the prior distribution.

(A8-1) The prior distribution $\varphi(\theta)$ is lower bounded by an infinitely differentiable function $\overline{\varphi}(\theta)$ whose support, $\{\theta|\overline{\varphi}(\theta) > 0\}$, is a compact set.

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.(3.7) is known (Watanabe et al., 2004),

$$\lambda \leq (MK_0 + K - 1)/2, \tag{8.3}$$

under the same condition about the true distribution and the model as the condition (A5-1) described in Chapter 5 and the condition (A8-1). Since the condition (A8-1) is satisfied by putting $\phi_0 = 1$ in the condition (A5-2) of Theorem 2, we can compare the stochastic complexities in this case. Putting $\phi_0 = 1$ in eq.(6.7), we have

$$\overline{\lambda} = K - K_0 + (MK_0 + K_0 - 1)/2. \tag{8.4}$$

Let us compare this $\overline{\lambda}$ of variational Bayesian learning to λ in eq.(8.3) 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 7.1 shows the coefficients $\overline{\lambda}$, $\lambda_{\rm BIC}$ 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 7.1, $\overline{\lambda}$ (solid line) and $\lambda_{\rm BIC}$ (dotted line) coincide. It is noted that $\overline{\lambda}$ of variational Bayesian learning eq.(8.4) relatively approaches the upper bound in Bayesian learning eq.(8.3) and becomes far smaller than that of BIC eq.(5.6) as the dimension M becomes larger.

Furthermore, for the general mixture model, the following upper bound on the coefficient λ is known(Yamazaki and Watanabe, 2003a; Yamazaki and Watanabe, 2003b),

$$\lambda \leq \begin{cases} (K + K_0 - 1)/2 & (M = 1), \\ (K - K_0) + (MK_0 + K_0 - 1)/2 & (M \geq 2), \end{cases} \tag{8.5}$$

It is noted from eq.(8.4) and eq.(8.5) that the upper bound of λ is equal to $\overline{\lambda}$ when $M \geq 2$. This implies that the variational posterior is close to the true Bayesian posterior when $M \geq 2$. More precise discussion about the accuracy of the approximation can be done for models on which tighter bounds or exact values of the coefficient λ in eq.(3.7) are given (Yamazaki and Watanabe, 2004).

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.(3.7) and eq.(3.8). This was also observed experimentally by the different behavior of λ_{VB} and λ_{G} in the previous chapter. Hence, assessing the generalization performance of the 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 1 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.

For linear neural networks, also known as reduced rank regression models, the generalization performance of the variational Bayesian approach was clarified (Nakajima, 2006). This is the first example where the generalization performance of variational Bayesian learning has been theoretically assessed. The explicit expression of the variational estimator will enable to apply the method used for linear neural networks to assessing generalization errors of the variational Bayesian approach for the mixture models and other models with hidden variables.

8.4 Model Selection

Let us consider the implications the main results have on the model selection problem. In the practical variational Bayesian framework, the posterior distribution $Q(K)$ over the number of components, K, are estimated using data

and the prior distribution $\varphi(K)$ (Attias, 1999). It is known in this procedure that the optimal posterior is given by

$$Q(K) \propto \varphi(K) \exp(-\overline{F}_K(X^n)),$$

(Ueda and Ghahramani, 2002). Here, we denote by $\overline{F}_K(X^n)$, the variational stochastic complexity defined by eq.(4.5), explicitly indicating the number of components, K. Theorem 2 and Theorem 3 suggest that this procedure is appropriate for selecting models. More specifically, if the true distribution is included in the possible models and $\varphi(K) > 0$ for all of them, the posterior probability $Q(K)$ has a strong peak at $K = K_0$, since the coefficients $\underline{\lambda}$ and $\overline{\lambda}$ in Theorem 2 and Theorem 3 are proportional to K.

A new method to estimate the number of mixture components has also been developed based on the fact that the stochastic complexity is a function of the true number of components (Yamazaki et al., 2005). Theorem 2 and Theorem 3 also suggest that this method can be applied to variational Bayesian learning.

8.5 Effect of Hyperparameters

Let us discuss the effects of the hyperparameters. From Theorem 2 and 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 $b_i = b_j$ for some pair (i, j). The prior distribution $\varphi(\mathbf{a})$ given by eq.(5.3) 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 $\varphi(\mathbf{b})$ given by eq.(6.3) 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 eqs.(6.6) and (6.7) 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 number M of parameters in each component. More specifically, only when $\phi_0 \leq (M+1)/2$, the prior

distribution reduces redundant components; otherwise it uses all the components.

8.6 Applications of the Bounds

Finally, let us give examples of how to use the theoretical bounds given in Theorem 2 and Theorem 3 and discuss issues to be addressed.

Comparing the theoretical bounds in eq.(6.5) 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.(4.3) and eq.(4.4), it may converge to local minima of the functional $\overline{F}[q]$. Remember that eq.(4.3) and eq.(4.4) 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. Indeed, such comparison was carried out in Chapter 7 and local minima were observed in Section 7.2. 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 mixture models are 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. In fact, the variational stochastic complexities of hidden Markov models (Hosino et al., 2006a), probabilistic context free grammars (Hosino et al., 2006b), and Bayesian networks (Watanabe et al., 2006) have been obtained in the similar way as used for proving Theorem 2 and Theorem 3 in this thesis. Consequently, the properties of variational Bayesian learning have been discussed for these learning machines.

As mentioned in Chapter 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. The main results of this thesis are important for developing effective model selection methods using $\overline{F}(X^n)$.

Furthermore, the main results will also contribute to deriving methods for optimizing the hyperparameters, now that their effects have been clarified as discussed in the previous section. The experimental results in Section 7.2 imply that the appropriate initial value of the practical variational Bayesian algorithm depends on the hyperparameters. In order to obtain the global minimum solution, the appropriate way to set the initial value should be derived based on the main results in this thesis.

Chapter 9

Conclusion

This thesis has established a theory for the principled design of variational Bayesian learning systems. Theoretical properties of variational Bayesian learning have been analyzed for mixture models.

The main contributions achieved in this thesis are summarized as follows,

1. For spherical Gaussian mixture models, the asymptotic upper and lower bounds were obtained on the variational stochastic complexity which was defined as an upper bound of the Bayesian stochastic complexity.

2. For mixtures of exponential families, the asymptotic upper and lower bounds were obtained on the variational stochastic complexity by generalizing the theorem for the Gaussian mixture model.

3. The theoretical results were compared to the ones of experiments using synthetic data. The properties of the practical variational Bayesian algorithm were discussed.

4. By comparing the variational stochastic complexity to the Bayesian stochastic complexity, the accuracy of the variational Bayesian approximation was discussed quantitatively. In other words, the asymptotic distance from the approximate posterior distribution to the true Bayesian posterior distribution was derived in terms of Kullback information. Additionally, the effect of hyperparameters was also clarified.

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

Appendix

The appendix gives supplements to some contents of this thesis. These include the proofs of some assertions. Section A.1 presents the proof of the property of the Kullback information defined in Section 3.2. Section A.2 presents the derivation of the asymptotic form of the Bayesian stochastic complexity. Section A.3 gives the derivation of the forms of the variational posteriors in Theorem 1 in Section 4.1. Section A.4 reviews the EM algorithm in ML and MAP learning. Section A.5 presents some basic properties of the exponential family distribution discussed in Section 6.1.

A.1 Property of Kullback Information

The Kullback information $K(q(x)||p(x))$ is almost positive definite, that is, $K(q(x)||p(x)) \geq 0$, with equality if and only if $q(x) = p(x)$ (Kullback, 1968). This section proves this assertion.

Let $g(x) = \frac{p(x)}{q(x)}$. Then

$$\begin{aligned} K(q(x)||p(x)) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\ &= \int q(x) \left\{ \frac{p(x)}{q(x)} - 1 - \log \frac{p(x)}{q(x)} \right\} dx \\ &= \int q(x) S(g(x)) dx, \end{aligned} \tag{A.1}$$

where $S(x) = x - 1 - \log x$.

Since $S(x) \geq 0$ with equality if and only if $x = 1$, it follows from eq.(A.1) that

$$K(q(x)||p(x)) \geq 0,$$

with equality if and only if $g(x) = \frac{p(x)}{q(x)} = 1$ almost everywhere.**(Q.E.D)**

A.2 Asymptotic Expansion of Bayesian Stochastic Complexity

This section outlines the derivation of the asymptotic form of the Bayesian stochastic complexity, eq.(3.7) (Watanabe, 2001a, 2001b).

Let

$$H(\theta) = K(p_0(x)||p(x|\theta)).$$

It can be proven that

$$\tilde{F}(n/2) - C \leq E_{X^n}[F_0(X^n)] \leq \tilde{F}(n), \tag{A.2}$$

where C is a constant independent of n and

$$\tilde{F}(n) = -\log \int \exp(-nH(\theta))\varphi(\theta)d\theta,$$

(Watanabe 2001a). Hence, we focus on the asymptotic expansion of $\tilde{F}(n)$ as $n \to \infty$, below.

The function $v(t)$ called the state density function is defined by

$$v(t) = \int \delta(t - H(\theta))\varphi(\theta)d\theta,$$

where $\delta(\cdot)$ is the Dirac delta function. It follows that

$$\tilde{F}(n) = -\log \int \exp(-t)v(\frac{t}{n})\frac{1}{n}dt, \tag{A.3}$$

which means $\tilde{F}(n)$ is given by the Laplace transform of the function $v(t)$.

On the other hand, the Mellin transform of the state density function $v(t)$ gives the zeta function

$$J(z) = \int t^z v(t)dt = \int H(\theta)^z \varphi(\theta)d\theta,$$

where z is a complex number.

The poles of the zeta function give the asymptotic form of $v(t)$ as $t \to +0$ and hence that of $\tilde{F}(n)$ as $n \to \infty$ from eq.(A.3). This is outlined below.

The poles of $J(z)$ are known to be rational negative numbers. Let $-\lambda_1 > -\lambda_2 > \cdots$ be the poles of the zeta function and $m_1, m_2, \cdots$ be

the corresponding orders of the poles. Then, $J(z)$ has the Laurent series representation around $z = -\lambda_k$,

$$J(z) = \sum_{m=1}^{m_k} \frac{c_{km}}{(z+\lambda_k)^m} + (\text{regular function}).$$

The inverse Mellin transform of the above form of $J(z)$ gives the asymptotic form of $v(t)$,

$$v(t) \simeq \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} \frac{c_{km}}{(m-1)!} t^{\lambda_k - 1} (-\log t)^{m-1}.$$

Substituting the above form into eq.(A.3), $\tilde{F}(n)$ is asymptotically expanded as

$$\tilde{F}(n) \simeq -\log \left\{ \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} \sum_{q=0}^{m-1} c'_{kmq} \frac{(\log n)^q}{n^{\lambda_k}} \int e^{-t} t^{\lambda_k - 1} (-\log t)^{m-1-q} dt \right\}.$$

By putting $\lambda = \lambda_1$ and $m = m_1$, the asymptotic form, eq.(3.7) is obtained from eq.(A.2).

A.3 Variational Posteriors (Proof of Theorem 1)

This section proves Theorem 1 in Section 4.1.

The variational free energy defined by eq.(4.1) can be manipulated into the following forms (neglecting additive constants) as functionals of $Q(Y^n|X^n)$ and $r(\theta|X^n)$ respectively.

$$\begin{aligned}
& \sum_{Y^n} \int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} d\theta \\
= & \sum_{Y^n} Q(Y^n|X^n) \left\langle \log \frac{Q(Y^n|X^n)}{p(X^n, Y^n|\theta)} \right\rangle_{r(\theta|X^n)} \\
= & \sum_{Y^n} Q(Y^n|X^n) \left\{ \log Q(Y^n|X^n) - \log(\frac{1}{C_Q} \exp\langle \log p(X^n, Y^n|\theta) \rangle_{r(\theta|X^n)}) \right\},
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{Y^n} \int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} d\theta \\
= &\int r(\theta|X^n) \left\langle \log \frac{r(\theta|X^n)}{p(X^n, Y^n, \theta)} \right\rangle_{Q(Y^n|X^n)} d\theta \\
= &\int r(\theta|X^n) \left\{ \log r(\theta|X^n) - \log(\frac{1}{C_r} \exp\langle \log p(X^n, Y^n|\theta)\rangle_{Q(Y^n|X^n)\varphi(\theta)}) \right\} d\theta.
\end{aligned}$$

Theorem 1 can be proven by applying the property of the Kullback information, described in Section A.1, to the above expressions.**(Q.E.D)**

A.4 EM algorithm for ML learning

This section reviews the derivation of the EM algorithm in ML and MAP learning for models with hidden variables (Dempster et al., 1977).

The maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$ is given by maximizing the log-likelihood,

$$L(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta).$$

The derivation of the EM algorithm starts with upper bounding the negative log-likelihood. For an arbitrary distribution $Q(Y^n)$ of the hidden variables $Y^n = \{y_1, \cdots, y_n\}$, the following inequality holds.

$$\begin{aligned}
-L(\theta) &= -\sum_{i=1}^{n} \log \sum_{y_i} p(x_i, y_i|\theta) \\
&\leq \sum_{i=1}^{n} Q(y_i) \log \frac{Q(y_i)}{p(x_i, y_i|\theta)} \qquad \text{(A.4)} \\
&\equiv \overline{L}(Q, \theta).
\end{aligned}$$

As can be seen in eq.(A.4), $\hat{\theta}_{\mathrm{ML}}$ is obtained by minimizing the functional $\overline{L}$ with respect to $Q(Y^n)$ and θ.

The EM algorithm to minimize $\overline{L}$ iterates the two steps, the E step, where the distribution $Q(Y^n)$ is optimized given a current parameter θ, and the M step, where $\overline{L}$ is minimized with respect to θ given $Q(Y^n)$.

In the E step, the optimal distribution $Q(Y^n)$ is given by the product of the distribution

$$Q(y_i) = p(y_i|x_i, \theta)$$

from eq.(A.4) and the property of the Kullback information described in Section A.1. In the M step, θ is updated to

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{y_i} Q(y_i) \log p(x_i, y_i|\theta),$$

since the rest of $\overline{L}$ does not depend on θ.

The EM algorithm for MAP learning is obtained similarly by only altering the update of θ in the M step to

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left\{ \log \varphi(\theta) + \sum_{i=1}^{n} \sum_{y_i} Q(y_i) \log p(x_i, y_i|\theta) \right\}.$$

In the case of the ML learning for the Gaussian mixture model discussed in Chapter 5, the E step gives

$$Q(y_i^k) \propto \hat{a}_k \exp(-\frac{1}{2}||x_i - \hat{\mu}_k||^2), \tag{A.5}$$

and the M step gives

$$\hat{a}_k = \frac{1}{n} \sum_{i=1}^{n} \langle y_i^k \rangle_{Q(y_i)}, \tag{A.6}$$

$$\hat{\mu}_k = \frac{1}{n\hat{a}_k} \sum_{i=1}^{n} \langle y_i^k \rangle_{Q(y_i)} x_i. \tag{A.7}$$

The maximum likelihood estimator for the Gaussian mixture model is given by iteratively updating $\{\hat{a}_k, \hat{\mu}_k\}$ and $Q(Y^n)$ by using eqs.(A.5),(A.6) and (A.7). These iterative updates are equivalent to those of the clustering algorithm called *soft K-means* (Mackay, 2003, Ch.22).

A.5 Statistical Exponential Families

This section presents some basic properties of the standard exponential family (Brown, 1986).

The family of probability distributions with densities of the form

$$c(x|b) = \exp\left[\sum_{k=1}^{M} b^{(k)} f_k(x) + f_0(x) - g(b) \right], \quad (b = (b^{(1)}, \cdots, b^{(M)})^T),$$

is called the *standard exponential family*. If $f_1(x), \cdots, f_M(x)$ are linearly independent, that is, $\alpha_1 f_1(x) + \cdots + \alpha_M f_M(x) = 0$ implies $\alpha_1 = \cdots = \alpha_M = 0$, then the number M is called the *order* of the exponential family.

Since $c(x|b)$ is a probability density function,

$$g(b) = \log \int \exp \left[\sum_{k=1}^{M} b^{(k)} f_k(x) + f_0(x) \right] dx$$

holds and $g(b)$ is called the *cumulant generating function*. The set B defined by

$$B = \left\{ (b^{(1)}, \cdots, b^{(M)}) | \int \exp \left[\sum_{k=1}^{M} b^{(k)} f_k(x) + f_0(x) \right] dx < \infty \right\}$$

is called the *natural parameter space*.

Here are important fundamental facts about the standard exponential families. Let $K(b) = e^{g(b)}$.

(1) B is a convex set in R^M, that is,

$$\lambda b_1 + (1 - \lambda) b_2 \in B$$

holds for any $b_1, b_2 \in B$ and any λ $(0 < \lambda < 1)$.

(2) $K(b)$ is analytic in the interior of B and

$$\begin{aligned} & \frac{\partial^{[p]} K(b)}{\partial (b^{(1)})^{p_1} \cdots \partial (b^{(M)})^{p_M}} \\ = & \int \frac{\partial^{[p]}}{\partial (b^{(1)})^{p_1} \cdots \partial (b^{(M)})^{p_M}} \exp \left[\sum_{k=1}^{M} b^{(k)} f_k(x) + f_0(x) \right] dx \end{aligned}$$

holds for any $p = (p_1, \cdots, p_M)^T$ where p_i are non-negative integers and $[p] = p_1 + \cdots + p_M$.

Bibliography

[1] Agusta, Y. and Dowe, D. L. (2003) Unsupervised learning of gamma mixture models using minimum message length. *Proceedings of 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 2003)*, Benalmadena, Spain, ACTA Press, Anaheim, CA, USA, 457-462.

[2] Akahira, M. and Takeuchi, K. (1995) *Non-regular Statistical Estimation*, Springer-Verlag, New York.

[3] Akaho, S. and Kappen, H. J. (2000) Nonmonotonic generalization bias of Gaussian mixture models. *Neural Computation*, 12 (4), 1411-1427.

[4] Akaike, H. (1974) A New look at statistical model identification. *IEEE Trans. Automatic Control*, 19 (6), 716-723.

[5] Akaike, H. (1980) Likelihood and Bayes procedure. *Bayesian Statistics*, (Bernald J.M. eds.) University Press, Valencia, Spain, 143-166, 1980.

[6] Alzer, H. (1997) On some inequalities for the Gamma and Psi functions. *Mathematics of computation*, 66 (217), 317-389.

[7] Amari, S. (1985) *Differential Geometrical Method in Statistics*, Springer-Verlag, New York.

[8] Amari, S. (1998) Natural gradient works efficiently in learning. *Neural Computation*, 10 (2), 251-276.

[9] Amari, S., Park, H., and Ozeki, T. (2006) Singularities affect dynamics of learning in neuromanifolds. *Neural Computation*, 18 (5), 1007-1065.

[10] Attias, H. (1999) Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, 21-30.

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

[12] Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C, and Wild, D. L. (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* , 21 (3), 349-356.

[13] Brown, L. D. (1986) *Fundamentals of statistical exponential families*. IMS Lecture Notes-Monograph Series 9.

[14] Cheng, L., Jiao, F., Schuurmans, D., and Wang, S. (2005) Variational Bayesian image modelling. *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 129-136.

[15] Chernoff, H., and Lander, E. (1995) Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference*, 43, 19-40.

[16] Clarke, B. S., and Barron, A. R. (1990) Information-theoretic asymptotics of Bayesian methods. *IEEE Trans. Information Theory*, 36 (3), 453-471 .

[17] Cramer, H. (1946) *Mathematical Methods of Statistics*, Princeton University Press.

[18] Dacunha-Castelle, D. and Gassiat, E. (1997) Testing in locally conic models, and application to mixture models. *Probability and Statistics*, 1, 285-317.

[19] Dempster, A. P., Laird, N. M., and Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39-B, 1-38.

[20] Efron, B., and Morris, C. (1973) Stein's estimation rule and its competitors-an empirical Bayes approach. *Journal of the American Statistical Association*, 68, 117-130.

[21] Feynman, R. P. (1972) *Statistical Mechanics: A Set of Lectures*, W.A.Benjamin, Inc.

[22] Fukumizu, K., Kuriki, S., Takeuchi, K., and Akahira, M. (2004) *Statistical Theory of Singular Models(in Japanese)*, Iwanami.

[23] Good, I. J. (1965) *The Estimation of Probabilities*, Cambridge, MA, MIT Press.

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

[25] Hartigan, J.A. (1985) A Failure of likelihood asymptotics for normal mixtures. *Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer*, 2, 807-810.

[26] Hinton, G. E., van Camp, D. (1993) Keeping the neural networks simple by minimizing the description length of the weights. *Proceedings of the sixth annual conference on Computational learning theory*, Santa Cruz, California, USA, 5-13.

[27] Honkela, A., and Valpola, H. (2004) Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Trans. Neural Networks*, 15 (4), 800-810.

[28] Hosino, T., Watanabe, K., and Watanabe, S. (2006a) Stochastic complexity of hidden Markov models on the variational Bayesian learning. *IEICE Trans.*, J89-D (6), 1279-1287.

[29] Hosino, T., Watanabe, K., and Watanabe, S. (2006b) Free energy of stochastic context free grammar on variational Bayes. *International Conference on Neural Information Processing (ICONIP2006)*, to appear.

[30] Iba, Y. (2001) Extended Ensemble Monte Carlo. *International Journal of Modern Physics*, C12, 623-656.

[31] Jaakkola, T. S., and Jordan, M. I. (2000) Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10, 25-37.

[32] Kiefer, J., and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27 (4), 887-906 .

[33] Kullback, S. (1968) *Information Theory and Statistics*, Dover Publications, New York.

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

[35] Liu, X., Pasarica, C., and Shao, Y. (2003) Testing homogeneity in gamma mixture models. *Scandinavian Journal of Statistics*, 30 (1), 227-239.

[36] Mackay, D. J. C.(1992) Bayesian interpolation. *Neural Computation*, 4 (2), 415-447.

[37] Mackay, D. J. C.(2003) *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

[38] McLachlan, G., and Peel, D. (2000) *Finite Mixture Models*, Wiley, New York.

[39] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.

[40] Minka, T. P. (2001) Expectation propagation for approximate Bayesian inference. *Proceedings of 17th Conference on Uncertainty in Artificial Intelligence*, Seattle, Washington, USA, 362-369.

[41] Nagata, K., and Watanabe, S. (2006) The Exchange Monte Carlo method for Bayesian learning in singular learning machines. *Proceedings of IEEE World Congress on Computational Intelligence*, 6383-6389.

[42] Nakajima, S. (2006) Asymptotic theory of empirical and variational Bayes learning. *Ph.D. Thesis*, Tokyo Institute of Technology.

[43] Nakano, N., Takahashi, K., and Watanabe, S. (2005) On the evaluation criterion of the MCMC method in singular learning machines(in Japanese). *IEICE Trans.*, J88-D2(10), 2011-2020.

[44] Neal, R. M. (1996) *Bayesian Learning for Neural Networks*, Springer.

[45] Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, 14, 1080-1100.

[46] Sato, M. (2001) Online model selection based on the variational Bayes. *Neural Computation*, 13 (7), 1649-1681.

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

[48] Ueda, N, and Ghahramani, Z. (2002) Bayesian model search for mixture models based on optimizing variational bounds. *International Journal of Neural Networks*, 15, 1223-1241.

[49] van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge University Press.

[50] Watanabe, S., Minami, Y., Nakamura, A, and Ueda, N. (2002) Application of variational Bayesian approach to speech recognition. Advances in Neural Information Processing Systems, 15, MIT Press, 1261-1268.

[51] Watanabe, K., and Watanabe, S. (2004) Lower bounds of stochastic complexities in variational Bayes learning of Gaussian mixture models. *Proceedings of IEEE conference on Cybernetics and Intelligent Systems (CIS04)*, Singapore, 99-104.

[52] Watanabe, K., and Watanabe, S. (2005a) Stochastic complexity for mixture of exponential families in variational Bayes. *Proceedings of the 16th International Conference on Algorithmic Learning Theory (ALT2005)*, Singapore, 107-121.

[53] Watanabe, K., and Watanabe, S. (2005b) On variational Bayes algorithms for exponential family mixtures. *Proceedings of International Symposium on Nonlinear Theory and its Applications (NOLTA2005)*, Bruges, Belgium, 393-396.

[54] Watanabe, K., and Watanabe, S. (2005c) Variational Bayesian algorithm and stochastic complexity for mixture models. *International Conference on Neural Information Processing (ICONIP2005)*, Taipei, Taiwan, 338-342.

[55] Watanabe, K., and Watanabe, S. (2006a) Variational Bayesian stochastic complexity of mixture models. *Advances in Neural Information Processing Systems*, 18, MIT Press, 1465-1472.

[56] Watanabe, K., and Watanabe, S. (2006b) Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *Journal of Machine Learning Research*, 7, 625–644.

[57] Watanabe, K., and Watanabe, S. (2006c) Stochastic complexities of general mixture models in variational Bayesian learning. *International Journal of Neural Networks*, to appear.

[58] Watanabe, K., Shiga, M., and Watanabe, S. (2006) Upper bounds for variational stochastic complexities of Bayesian networks. *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL2006)*, Burgos, Spain, 139-146.

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

[60] Watanabe, S. (2001b) *Learning Machines and Algorithms (in Japanese).* Kyoritsu.

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

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

[63] Yamazaki, K., and Watanabe, S. (2003b) Stochastic complexity of Bayesian networks, *Uncertainty in Artificial Intelligence*, Acapulco, Mexico, 592-599.

[64] Yamazaki, K., and Watanabe, S. (2004) Newton diagram and stochastic complexity in mixture of binomial distributions, *Algorithmic Learning Theory (ALT2004)*, 350-364.

[65] Yamazaki, K., Nagata, K., and Watanabe, S. (2005) A New method of model selection based on learning coefficient, *Proceedings of International Symposium on Nonlinear Theory and its Applications (NOLTA2005)*, Bruge, Belgium, 389-392.