T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	変分ベイズ学習における相転移現象および事前分布の最適設計
Title(English)	Phase transition in variational bayes learning and optimal design of prior distribution
著者(和文)	梶大介
Author(English)	Daisuke Kaji
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第8284号, 授与年月日:2011年3月26日, 学位の種別:課程博士, 審査員:渡邊 澄夫
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第8284号, Conferred date:2011/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Powered by T2R2 (Science Tokyo Research Repository)

Phase Transition in Variational Bayes Learning and Optimal Design of Prior Distribution

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

Daisuke Kaji

2011

Abstract

The variational Bayes method is widely applied to many practical problems. This method not only solves an expensive computational cost of the Bayes method, but also provides a good generalization performance. The variational posterior distribution is given by minimizing the variational Bayes free energy under the condition of the mean field approximation. However little is known about the theoretical properties on the influence which hyperparameters give to the variational Bayes free energy.

The purpose of this thesis is to clarify the relation between the variational Bayes free energy and the hyperparameters, and to give the guideline for design of the hyperparameters.

For this purpose, we firstly derive the asymptotic expansion of the variational Bayes free energy on the Bernoulli mixture. Then we show the existence of the phase transition phenomenon depending on hyperparameters of both mixing ratio and the Bernoulli distribution. In addition, we experimentally investigate the optimal setting of hyperparameters from the two viewpoints of prediction and clustering.

Finally, we discuss the significance which the phase transition phenomenon gives to the clustering problem and the learning theory.

Preface

This study is carried out at Watanabe laboratory, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology. Advices from the member of Watanabe laboratory make me notice many important things which I overlooked.

Assistant professor Kazuho Watanabe gave me fundamental and important advices about the variational Bayes learning and these advices were indispensable for this study.

Assistant professor Keisuke Yamazaki pointed out the essential problem from the viewpoint of Bayesian learning and it was greatly helpful to understand the phase transition phenomenon of the variational Bayes learning.

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

Finally, I am grateful to my supervisor Prof. Sumio Watanabe. I was taught a lot of important things about the theoretical approach method to the mathematical structure which practical problems intrinsically have.

Daisuke Kaji

Contents

Abstract **i**

Preface **iii**

1 Introduction **1**
1.1 Back Ground of Research 1
1.2 Purpose of Research 3
1.3 Outline of Thesis 4

2 Baysian Learning Theory **5**
2.1 Learning from Data 5
2.2 Bayesian Learning 6

3 Bernoulli Mixture Model **11**
3.1 Mixture Model 11
3.2 Bernoulli Mixture 13
3.3 Clustering using Bernoulli Mixture 14

4 Variational Bayes Learning **19**
4.1 Variational Bayes Algorithm 19
4.2 Model Selection by Variational Bayes Free Energy 22

5 Phase Transition in Variational Bayes Learning **25**
5.1 Asymptotic Expansion of Variational Bayes Free Energy . . . 25
5.2 Proof of Theorem 5.1.1 30
5.2.1 Calculation of Asymptotic Expansion 30

5.2.2 Derivation of Eq.(5.3) 35
5.2.3 Proof of Lemma 5.2.1 36
5.2.4 Proof of Lemma 5.2.2 39
5.3 Experiments 41
5.4 Discussion 42

6 Optimal Hyperparameter Design 43
6.1 Two Design Method of Hyperparameter 43
6.1.1 Hyperparameter for Generalized Learning 43
6.1.2 Hyperparameter for Knowledge Discovery 44
6.2 Experiments 45
6.2.1 Variational Bayes Free Energy and Variational Bayes Generalization Error 45
6.2.2 Knowledge Discovery 47
6.2.3 Application to Category Classification 48
6.3 Discussion 49

7 Discussion 51
7.1 Discussion from Learning Theory 51
7.2 Discussion from Optimal Design of Hyperparameter 54

8 Conclusion 57

Appendix 59
A.1 EM Algorithm of Bernoulli mixture 59
A.2 Conjugate Prior 61
A.3 Noninformative Prior and Jeffreys' Prior 62

Bibliography 65

Chapter 1

Introduction

In this Chapter, we first explain the background and purpose of our research, then an outline of this thesis is given.

1.1 Back Ground of Research

Technological advances of hardware and learning algorithm in recent years enable us to use more complicated statistical models such as multi-layer perceptron, graphical model and mixture model. Furthermore these advantages create the new flow of data analysis called data mining by tying up to large amounts of data. Mixture model, which is one of the generative models, has been extensively studied as the probabilistic model which makes it possible to analyze the latent structure of data. In particular, a multivariate Bernoulli mixture model is widely applied to image analysis, text classification, and so on, as a tool for analyzing binary data. In particular, a multivariate Bernoulli mixture model is widely applied to image analysis, text classification, and so on, as a tool for analyzing binary data[Bishop(2006), Juan & Vidal (2001, 2004)].

Among of its training algorithms, the variational Bayes learning is known as the algorithm which provides both computational tractability and good generalization performance for mixture models. In spite of wide range of the applications, its properties have not yet been made clear enough. This is because the mixture model is a non-regular model. A statistical model is

regular if and only if a set of conditions (referred to as“ regularity conditions ”) that ensure the asymptotic normality of the maximum likelihood estimator is satisfied. The regularity conditions are not satisfied for mixture models because the parameters are not identifiable, in other words, the mapping from parameters to probability distributions is not one-to-one.

In Bayesian learning, mathematical foundation for analyzing non-regular models was established using algebraic geometry, and the asymptotic behaviors of Bayes marginal likelihood and Bayes generalization error in singular learning machines were clarified [Watanabe (2001, 2009, 2010)]. In addition, the Bayesian stochastic complexities or the marginal likelihoods of several non-regular models have been clarified in recent studies [Yamazaki & Watanabe (2003a, b)]. In the Bayesian framework, the predictive distribution is derived as an ensemble with respect to the posterior distribution of parameters and provides better generalization performance in non-regular models than the maximum likelihood method that tends to overfit the data.

However, it is hard to compute the exact Bayesian posterior in general, therefore some approximations were proposed. Well-known approximate methods include Markov Chain Monte Carlo (MCMC) methods and the Laplace approximation. The former attempts to find the exact posterior distribution by using sampling methods but typically requires huge computational resources. The latter approximates the posterior distribution by a Gaussian distribution, which is insufficient for models containing hidden variables. The variational Bayesian framework was proposed as another approximation using the mean field approximation for computations in the models with hidden variables [Attias (1999), Ghahramani & Beal (2000), Smidl & Quinn (2006), Beal (2003)]. This framework provides computationally tractable posterior distributions over the hidden variables and the parameters with an iterative algorithm.

Although the properties of the variational Bayes learning, such as approximation accuracy for true distribution, remain unclear, recently the asymptotic variational free energy has been studied in case of mixture of the exponential families. It was reported that its upper and lower bounds are given by the non-smooth function of the hyperparameter of mixing ratio. This result indicates the existence of phase transition phenomenon [Watanabe &

Watanabe(2006a, b), (2007)]. On the other hand, it was also reported that the variational Bayes generalization error do not have a direct mathematical relationship with the variational free energy in three-layer linear neural networks [Nakajima & Watanabe (2007)].

1.2 Purpose of Research

The purpose of this thesis is to clarify the relation between the variational Bayes free energy and the hyperparameters and to give the guideline for the design method of the hyperparameters. For this purpose, we introduce the notion of "deterministic components" which generate fixed data. Then we theoretically derive the asymptotic expansion of the variational Bayes free energy using the deterministic components in Bernoulli mixture. The phase transition phenomenon of the variational Bayes posterior depending on the hyperparameters of mixing ratio and Bernoulli distributions is directly derived from this asymptotic expansion [Kaji et al. (2010)].
In the second half of this thesis, we experimentally investigate the optimal hyperparameter design method from the following two viewpoints:

- Prediction : inference of the probability of datum $\boldsymbol{x}$.
- Clustering : analysis of the data structure.

We first show that the behavior of the generalization error is similar to one of the variational Bayes free energy; hence the optimization of the hyperparameter by the minimization of the variational Bayes free energy is useful for the prediction. Next we apply the variational Bayes method to the clustering problem and show the hyperparameter design method for extraction of the "minority" and the "general tendency of the data". These experimental results demonstrate that the optimal hyperparameter for the different purposes are different from each other[Kaji & Watanabe (2009)].

Finally, we discuss the significance of the asymptotic expansion of the variational Bayes free energy on the learning theory. The asymptotic expansion gives the estimation for the asymptotic generalization error of the variational Bayes method in Bernoulli mixture. Furthermore we consider

the influence that the phase transition phenomenon gives to the clustering results.

1.3 Outline of Thesis

This thesis is organized as follows:

- Chapter 2 : We review the framework of statistical learning theory and the Bayes learning.
- Chapter 3 : We introduce the Bernoulli mixture model and its applications. The problem of clustering by maximum likelihood estimation is also discussed.
- Chapter 4 : We outline the variational Bayes learning. The variational Bayes free energy is defined.
- Chapter 5 : We derive the asymptotic expansion of the variational Bayes free energy. The phase diagram and the existence of the phase transition phenomenon are shown.
- Chapter 6 : We present the experimental results on the optimal hyperparameter setting using artificial data and practical data.
- Chapter 7 : We discuss the significance of our results from the both viewpoints of the learning theory and the optimal hyperparameter design.
- Chapter 8 : We summarize the results and concludes of this thesis.

Chapter 2

Baysian Learning Theory

Machine learning system is composed of a model and a learning algorithm. This thesis examine the variational Bayes learning introduced as an approximation method of Bayesian learning. In this chapter, we review Baysian learning theory. We start by explaining the framework of learning theory. Then we give the definition of Baysian learning and some recent results. The variational Bayes learning will be given in Chapter 4.

2.1 Learning from Data

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ be the data set. The learning model or the statistical model is defined by a conditional density function $p(\boldsymbol{x}|\boldsymbol{\theta})$ of $\boldsymbol{x} \in \mathbb{R}^d$, where $\boldsymbol{\theta} \in \Theta$ is a parameter and Θ is the set of all parameters. Then the statistical learning by the learning model is to estimate the parameter $\boldsymbol{\theta}$ from the data set $\boldsymbol{X}$, the following three learning algorithms are basic and representative:

- maximum likelihood estimator (MLE)
- maximum a posteriori estimator (MAP estimator)
- Bayes estimator

MLE $\boldsymbol{\theta}_{mle}$ and MAP estimator $\boldsymbol{\theta}_{map}$ are defined as follows:

MLE

$$\boldsymbol{\theta}_{mle} = \arg\max_{\boldsymbol{\theta}} \left\{ \frac{1}{N} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}) \right\},$$

where $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$ is called log likelihood of data set $\boldsymbol{X}$.

The MLE is determined by only the model and the data. In contrast, the prior distribution $\varphi(\boldsymbol{\theta})$ on parameter $\boldsymbol{\theta}$ is given in the MAP estimation. Then the MAP estimator is defined as follows:

MAP estimator

$$\boldsymbol{\theta}_{map} = \arg\max_{\boldsymbol{\theta}} \left\{ \frac{1}{N} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}) + \frac{1}{N} \log \varphi(\boldsymbol{\theta}) \right\}.$$

Some optimization methods, such as the steepest descent method, are applied to calculate the above estimators. We see from $\frac{1}{N} \log \varphi(\boldsymbol{\theta}) \to 0$ $(N \to \infty)$ that the MAP estimator gets closed to the MLE in $N \to \infty$. However, when the data size is small, the MAP estimator can avoid overfitting by the term $\frac{1}{N} \log \varphi(\boldsymbol{\theta})$ acting as the regularization term. Both algorithms determine the optimal parameter, hence the predictive distributions are given by $p(\boldsymbol{x}|\boldsymbol{\theta}_{mle})$ and $p(\boldsymbol{x}|\boldsymbol{\theta}_{map})$.
On the other hand, in the framework of Bayes estimation, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ is given by the data set $\boldsymbol{X}$. Then the predictive distribution is calculated by taking average on the parameter $\boldsymbol{\theta}$. We will describe the definition of the Bayes estimator and its properties in the next section.

2.2 Bayesian Learning

Let $p(\boldsymbol{x}|\boldsymbol{\theta})$ be a learning model which has parameter $\boldsymbol{\theta}$ and $\varphi(\boldsymbol{\theta})$ is a prior distribution of $\boldsymbol{\theta}$. In Bayesian learning, the following posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ is computed from the given data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$,

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z} \varphi(\boldsymbol{\theta}) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}),$$

where $Z = Z(\boldsymbol{X})$ is the normalization constant that is also known as the marginal likelihood or the Bayesian evidence of the data set $\boldsymbol{X}$. The Bayesian predictive distribution $p(\boldsymbol{x}|\boldsymbol{X})$ is given by averaging the model over the posterior distribution as follows:

$$p(\boldsymbol{x}|\boldsymbol{X}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}.$$

And in general the generalization error can be measured by the Kullback Leibler divergence

$$KL(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{X})) = \int p^*(\boldsymbol{x}) \log \frac{p^*(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{X})} d\boldsymbol{x},$$

where $p^*(\boldsymbol{x})$ is the true distribution. The free energy, the minus log likelihood or the Bayesian stochastic complexity $F(\boldsymbol{X})$ is defined by

$$F(\boldsymbol{X}) = -\log Z(\boldsymbol{X}),$$

which is important in data modeling problems, because it is equal to the likelihood of a statistical model. Therefore it is used as a criterion by which the learning model is selected and the hyperparameters in the prior are optimized [Akaike (1980)]. The Bayesian posterior can be rewritten as

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z_0(\boldsymbol{X})} \exp(-NH(\boldsymbol{\theta}))\varphi(\boldsymbol{\theta}),$$

where $H(\boldsymbol{\theta})$ is the empirical Kullback Leibler divergence,

$$H(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p^*(\boldsymbol{x}_n)}{p(\boldsymbol{x}_n|\boldsymbol{\theta})},$$

and $Z_0(\boldsymbol{X})$ is the normalization constant. By using the empirical entropy

$$S(\boldsymbol{X}) = -\frac{1}{N} \sum_{n=1}^{N} \log p^*(\boldsymbol{x}_n),$$

the normalized free energy $F_0(\boldsymbol{X})$ is defined by

$$F_0(\boldsymbol{X}) = -\log Z_0(\boldsymbol{X}) = F(\boldsymbol{X}) - NS(\boldsymbol{X}). \tag{2.1}$$

It is noted that the empirical entropy $S(\boldsymbol{X})$ does not depend on the model $p(\boldsymbol{x}|\boldsymbol{\theta})$ and the prior distribution $\varphi(\boldsymbol{\theta})$. Therefore minimization of $F(\boldsymbol{X})$ is equivalent to that of $F_0(\boldsymbol{X})$. Let $E_{\boldsymbol{X}}[\cdot]$ denote the expectation over all data sets. Then it follows from Eq.(2.1) that

$$E_{\boldsymbol{X}}[F(\boldsymbol{X}) - F_0(\boldsymbol{X})] = NS,$$

where $S = -\int p^*(\boldsymbol{x}) \log p^*(\boldsymbol{x}) d\boldsymbol{x}$ is the entropy. There is the following relationship between the average free energy and the average generalization error [Levin et al. (1990)],

$$\begin{aligned} E_{\boldsymbol{X}}[KL(p^*(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{X}))] &= E_{\boldsymbol{X}^{+1}}[F(\boldsymbol{X}^{+1})] - E_{\boldsymbol{X}}[F(\boldsymbol{X})] - S \\ &= E_{\boldsymbol{X}^{+1}}[F_0(\boldsymbol{X}^{+1})] - E_{\boldsymbol{X}}[F_0(\boldsymbol{X})], \end{aligned} \tag{2.2}$$

where $\boldsymbol{X}^{+1} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{x}_{N+1}\}$. Recently, in Bayesian learning, an advanced mathematical method for analyzing non-regular models was established [Watanabe (2001)], which enables us to clarify the asymptotic behavior of the free energy of non-regular models. More specifically, by using concepts in algebraic analysis, it was proved that the average normalized Bayesian stochastic complexity defined by $E_{\boldsymbol{X}}[F_0(\boldsymbol{X})]$ has the following asymptotic form

$$E_{\boldsymbol{X}}[F_0(\boldsymbol{X})] = \lambda \log N - (m-1) \log \log N + O(1), \tag{2.3}$$

where λ and m are the rational number and the natural number respectively which are determined by the singularities of the true parameter. In regular statistical models, 2λ is equal to the number of parameters and $m = 1$, whereas in non-regular models such as Gaussian mixture models, 2λ is not larger than the number of parameters and $m \geq 1$. This means non-regular models have an advantage in Bayesian learning. From Eq.(2.2), if the asymptotic form of the average normalized Bayesian stochastic complexity is given by Eq.(2.3), the average generalization error is given by

$$E_{\boldsymbol{X}}[KL(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{X}))] \simeq \frac{\lambda}{N} + o\left(\frac{1}{N}\right).$$

Since the coefficient λ is proportional to the average generalization error, Bayesian learning is more suitable for non-regular models than the maxi-

mum likelihood method. However the free energy and the predictive distribution can not be given analytically in general and it is also typically hard to compute them by integrating over the posterior distribution.

Chapter 3

Bernoulli Mixture Model

Bernoulli mixture model is widely applied in image processing and text mining as a clustering tool of binary data. In this chapter, we first give the definition of the mixture model and the Bernoulli mixture, then we discuss the problem of the clustering using the Bernoulli mixture by the MLE method.

3.1 Mixture Model

The mixture model, which is a linear combination of the probabilistic distributions, is given by following equation:

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\theta}_k),$$

where $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$ is a probabilistic distribution parameterized by $\boldsymbol{\theta}_k$ and π_k is a mixing ratio satisfying $\sum_{k=1}^{K} \pi_k = 1$ for $\pi_k \geq 0$ $(\forall k)$. Hereafter, we call $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$ k-th component. This linear combination of the probabilistic distribution makes it possible not only to express more complicated probabilistic distribution but to give the information of component by which the data are generated. That is, when $p(k)$ and $p(k|\boldsymbol{x})$ are given by

$$\pi_k = p(k), \quad p(\boldsymbol{x}|k) = p(\boldsymbol{x}|\boldsymbol{\theta}_k),$$

we can obtain the conditional distribution of k given $\boldsymbol{x}$ by the following equation:

$$\begin{aligned} p(k|\boldsymbol{x}) &= \frac{p(k)p(\boldsymbol{x}|k)}{\sum_l p(l)p(\boldsymbol{x}|l)} \\ &= \frac{\pi_k p(\boldsymbol{x}|\boldsymbol{\theta}_k)}{\sum_l \pi_l p(\boldsymbol{x}|\boldsymbol{\theta}_l)}. \end{aligned}$$

This notion leads to the expression of the mixture model using the hidden variables. Let $\boldsymbol{z} = (z_1, z_2, \cdots, z_K)$ be the K-dimensional competitive probabilistic variables over $\mathcal{C}^K = \{(\underbrace{1, 0, \cdots, 0}_{K}), (0, 1, 0, \cdots, 0), \cdots, (0, \cdots, 0, 1)\}$ and the probability of $z_k = 1$ is given by

$$p(z_k = 1) = \pi_k,$$

then $p(\boldsymbol{z})$ can be expressed by

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

Similarly, the conditional probability distribution $p(\boldsymbol{x}|\boldsymbol{z})$ is given by

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\theta}_k)^{z_k}.$$

Therefore we obtain the expression of the mixture model using the latent variable as follows:

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}),$$

where $\sum_{\boldsymbol{z}}$ means $\sum_{\boldsymbol{z} \in \mathcal{C}^K}$. As a result, we have a discussion using the joint probability $p(\boldsymbol{x}, \boldsymbol{z})$. This makes it possible to derive the EM(expectation-maximization) algorithm (See Appendix A.1) and the variational Bayes algorithm. In Chapter 4, we will derive the Variational Bayes algorithm.

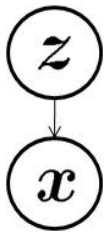

Figure 3.1: Graphical expression of mixture model

3.2 Bernoulli Mixture

In this section, we introduce a Bernoulli mixture and its prior distribution. Let $\boldsymbol{x} = (x_1, \cdots, x_M)^T$ be an M dimensional binary datum, $x_i = 0, 1$, and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_M)^T$ be a parameter which satisfies $0 \leq \mu_i \leq 1$. The (multivariate) Bernoulli distribution is defined by

$$B(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{M} \mu_i^{x_i}(1-\mu_i)^{1-x_i},$$

which shows the probability of $x_i = 0, 1$ is given by $\mu_i, 1-\mu_i$ respectively. The Bernoulli mixture is defined by

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \pi_k B(\boldsymbol{x}|\boldsymbol{\mu}_k), \tag{3.1}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)$, $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, ..., \mu_{kM})$ denote the mixing ratio of $\{B(\boldsymbol{x}|\boldsymbol{\mu}_k)\}$ and the parameter of the k-th Bernoulli distribution respectively and K is the number of components. Hence, using the hidden variables $\boldsymbol{z} = (z_1, z_2, ..., z_K)$, the simultaneous distribution of $(\boldsymbol{x}, \boldsymbol{z})$ is

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\mu}) = \prod_{k=1}^{K} \Big(\pi_k B(\boldsymbol{x}|\boldsymbol{\mu}_k)\Big)^{z_k}$$

which satisfies $\sum_z p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\mu}) = p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu})$. In variational Bayes learning, the conjugate prior distributions[1] are employed for the prior distributions.

[1]See Appendix A.2

In case of the Bernoulli mixture, the conjugate prior distributions are given by the Dirichlet and Beta distributions. They are defined by

$$\mathrm{Dir}(\boldsymbol{\pi}|\mathbf{a}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K}\Gamma(a_k)} \prod_{k=1}^{K} \pi_k^{a_k-1},$$
$$\mathrm{Beta}(\mu|\boldsymbol{b}) = \left(\frac{\Gamma(b_1+b_2)}{\Gamma(b_1)\Gamma(b_2)}\mu^{b_1-1}(1-\mu)^{b_2-1}\right),$$

where $\mathbf{a} = (a_1, a_2, ..., a_K)$ and $\mathbf{b} = (b_1, b_2)$ are sets of constants. In this paper, we study the case when the prior distributions of $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ are prepared respectively by

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|a, a, ..., a), \quad p(\boldsymbol{\mu}) = \prod_{k=1}^{K}\prod_{m=1}^{M} \mathrm{Beta}(\mu_{km}|b, b), \tag{3.2}$$

where $a > 0$ and $b > 0$ are hyperparameters. Let N be the number of data and $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ be the data set. The set of all corresponding hidden variables is denoted by $\boldsymbol{Z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$. Then the distribution of $\boldsymbol{X}$ and $\boldsymbol{Z}$ for given parameters $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\pi_k \prod_{m=1}^{M} \mu_{km}^{x_{nm}}(1-\mu_{km})^{1-x_{nm}}\right)^{z_{nk}}.$$

3.3 Clustering using Bernoulli Mixture

The Bernoulli mixture model is widely used in many applications as a clustering tool of binary data. For example, the binary image analysis is one of the important application field of the Bernoulli mixture [Juan & Vidal(2001),(2004)] Here we consider the mechanism of the clustering using the Bernoulli mixture by the MLE. The learning result is influenced by the property of non-regular model as described below. Suppose that $N\,(> 0)$ is multiples of 10 and the samples $\boldsymbol{x} = (x_1, x_2, x_3) = (1, 1, 0)$ and $(0, 1, 1)$ are given with size $0.5N$ respectively. When the learner $p(\boldsymbol{x}_n|\boldsymbol{\theta})$ has two

components, then the likelihood function

$$L_{mle}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

takes maximum value if and only if

$$(\pi_1, \pi_2) = (0.5, 0.5), (\mu_{11}, \mu_{12}, \mu_{13}) = (1, 1, 0), (\mu_{21}, \mu_{22}, \mu_{23}) = (0, 1, 1)$$

or

$$(\pi_1, \pi_2) = (0.5, 0.5), (\mu_{11}, \mu_{12}, \mu_{13}) = (0, 1, 1), (\mu_{21}, \mu_{22}, \mu_{23}) = (1, 1, 0).$$

In this case, above two mixture models give same clustering results.

On the other hand, let us suppose that the samples $(1, 1, 0), (0, 1, 1)$ and $(0, 1, 0)$ are given with size $0.4N, 0.4N, 0.2N$ respectively and the learner has three components. Then, for example, the following parameters give the maximum value of the likelihood function $L_{mle}(\boldsymbol{\theta})$ (See Figure.3.2),

$$\boldsymbol{\pi} = (0.4, 0.4, 0.2), \boldsymbol{\mu} = (1, 1, 0, 0, 1, 1, 0, 1, 0), \tag{3.3}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ and $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{31}, \mu_{32}, \mu_{33})$.

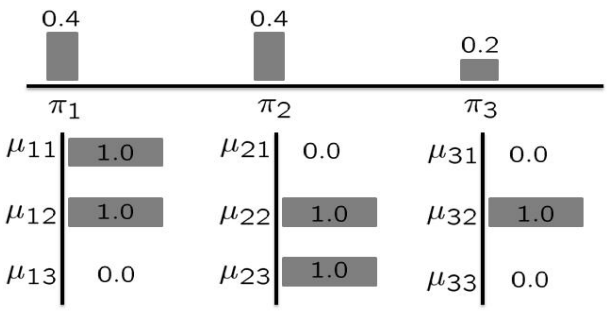

Figure 3.2: Clustering result 1

In contrast, the following parameters also give the maximum value of the likelihood function $L_{mle}(\boldsymbol{\theta})$ (See Figure.3.3),

$$\boldsymbol{\pi} = (0.6, 0.4, 0.0), \boldsymbol{\mu} = (2/3, 1, 0, 0, 1, 1, 0, 0, 0). \tag{3.4}$$

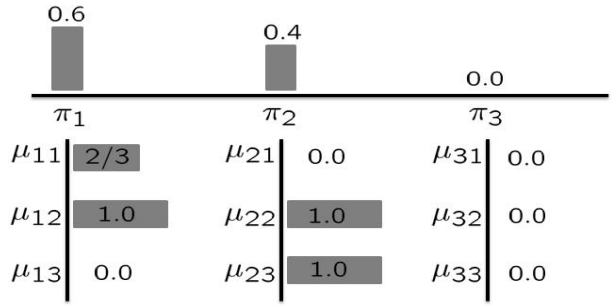

Figure 3.3: Clustering result 2

More generally,

$$\boldsymbol{\pi} = (0.5(1-t), 0.5(1-t), t), \boldsymbol{\mu} = \left(\frac{0.4}{0.5(1-t)}, 1, 0, 0, 1, \frac{0.4}{0.5(1-t)}, 0, 1, 0\right)$$

$(0 \leq t \leq 1)$ also give the maximum value. It follows from these results that the MLE does not have any mechanism to determine the model uniquely in the above case. However the results of (3.3) and (3.4) give different interpretations of the data in the context of clustering analysis. That is, (3.3) extracts following features:

- Data are separated into 3 types.
- The first component and the second component take place with same probability.

On the other hand, (3.4) gives the following features:

- Data are separated into 2 types.
- The first component takes place easier than the second component.

Next we see the case of MAP estimator. The likelihood function of MAP estimator is given by

$$L_{map}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log \varphi(\boldsymbol{\theta}) = L_{mle}(\boldsymbol{\theta}) + \log \varphi(\boldsymbol{\theta}),$$

where $\varphi(\boldsymbol{\theta})$ is the prior distribution. When both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ give the maximum value of L_{mle}, then the inequality relation of L_{map} depends on the value of $\log \varphi(\boldsymbol{\theta})$. As a result, the prior distribution determines the structure of the mixture model in this case.

How does the variational Bayes learning determine the clustering structure of mixture model? We discuss this problem in Chapter 5.

Chapter 4

Variational Bayes Learning

This chapter starts with an outline of the variational Bayes learning theory, then we give the concrete algorithm of the variational Bayes learning in the Bernoulli mixture. We also describe the model selection using the variational Bayes free energy in Section 4.2.

4.1 Variational Bayes Algorithm

In this section, we explain the well-known variational Bayes learning as the approximation of the Bayes Learning. Let $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu})$ be the set of parameters. For an arbitrary probability distribution $q(\boldsymbol{Z}, \boldsymbol{\theta})$, the functional is defined by

$$\bar{F}[q(\boldsymbol{Z}, \boldsymbol{\theta})] = \sum_{\boldsymbol{Z}} \int q(\boldsymbol{Z}, \boldsymbol{\theta}) \log \frac{q(\boldsymbol{Z}, \boldsymbol{\theta})}{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})} d\boldsymbol{\theta}, \tag{4.1}$$

where $p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}) = p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ is given by

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})$$

using $p(\boldsymbol{\pi})$ and $p(\boldsymbol{\mu})$ defined in (3.2). It follows that

$$\bar{F}[q(\boldsymbol{Z}, \boldsymbol{\theta})] = F(\boldsymbol{X}) + KL(q(\boldsymbol{Z}, \boldsymbol{\theta}) \| p(\boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{X})).$$

Since $F(\boldsymbol{X})$ does not depend on $q(\boldsymbol{Z}, \boldsymbol{\theta})$, the minimization of $\bar{F}[q(\boldsymbol{Z}, \boldsymbol{\theta})]$ is equivalent to that of $KL(q(\boldsymbol{Z}, \boldsymbol{\theta}) \| p(\boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{X}))$. In variational Bayes approximation, the method of the mean field approximation is applied to solving this

minimization problem. Namely, the trial distribution $q(\boldsymbol{Z}, \boldsymbol{\theta})$ is optimized by the minimization of $\bar{F}[q(\boldsymbol{Z}, \boldsymbol{\theta})]$ under the condition that

$$q(\boldsymbol{Z}, \boldsymbol{\theta}) = q_1(\boldsymbol{Z}) q_2(\boldsymbol{\theta}). \tag{4.2}$$

By using variational method, it is derived that the optimal distributions satisfy

$$\log q_1(\boldsymbol{Z}) = E_{q_2}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})] + C_1,$$
$$\log q_2(\boldsymbol{\theta}) = E_{q_1}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})] + C_2,$$

where C_1, C_2 are the normalization constants. Also, for the Bernoulli mixture model, it is shown that the optimal distribution $q_2(\boldsymbol{\theta})$ can be parameterized by $\boldsymbol{\alpha} = \{\alpha_k\}$ and $\{\eta_{km}, \eta'_{km}\}$,

$$q_2(\boldsymbol{\theta}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{k=1}^{K} \prod_{m=1}^{M} \mathrm{Beta}(\mu_{km}|\eta_{km}, \eta'_{km}),$$

and that $\boldsymbol{\alpha}$ and $\{\eta_{km}, \eta'_{km}\}$ are optimized by the following recursive procedures,

---

<u>*VB e-step*</u>

$$\begin{aligned} \log \rho_{nk} &= \psi(\alpha_k) - \psi\left(\sum_k^K \alpha_k\right) \\ &+ \sum_{m=1}^{M} \left(x_{nm}\psi(\eta_{km}) - x_{nm}\psi(\eta'_{km}) + \psi(\eta'_{km}) - \psi(\eta_{km} + \eta'_{km})\right), \end{aligned} \tag{4.3}$$

$$r_{nk} = \frac{\rho_{nk}}{\sum_{k=1}^{K} \rho_{nk}}. \tag{4.4}$$

<u>*VB m-step*</u>

$$N_k = \sum_{n=1}^{N} r_{nk}, \quad \nu_{km} = \sum_{n=1}^{N} r_{nk} x_{nm}, \quad \nu'_{km} = \sum_{n=1}^{N} r_{nk}(1 - x_{nm}), \tag{4.5}$$

$$\alpha_k = \alpha + N_k, \quad \eta_{km} = b + \nu_{km}, \quad \eta'_{km} = b + \nu'_{km}. \tag{4.6}$$

Here ψ denotes the digamma distribution $\psi(t) \equiv \Gamma'(t)/\Gamma(t)$. According to the information geometry, the mean field approximation can be regarded as the e-projection to the e-flat submanifold [Amari et al. (2001)]. However the interpretation of the variational Bayes method for mixture models using the information geometry has not yet been completely clarified and it is expected to explain the relation between the learning result and the hyperparameters from the geometrical perspective. At last, the minimized functional

$$\bar{F}(\boldsymbol{X}) \equiv \min_{q_1, q_2} \bar{F}[q(\boldsymbol{Z}, \boldsymbol{\theta})]$$

is called the variational Bayes free energy. It is a function of hyperparameters (a, b). The numerical value of the variational Bayes free energy can be obtained by the above recursive calculations, which is applied to the variational model evaluation. However, its theoretical behavior has been left unknown. We will describe the outline of the model evaluation by the variational Bayes free energy in next section.

4.2 Model Selection by Variational Bayes Free Energy

In this section, we explain the model selection using the variational Bayes free energy[Bishop (2006), Wang et al. (2003)]. Let $\boldsymbol{m}$ be the parameter which means the probabilistic model. The number of components and the hyperparameters are examples of $\boldsymbol{m}$. When $p(\boldsymbol{m})$ is a prior distribution of $\boldsymbol{m}$, then we can obtain the variational Bayes free energy by same discussion in Section 4.1 as follows:

$$
\begin{aligned}
\bar{F}[q(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{m})] &= \sum_{\boldsymbol{m}}\sum_{\boldsymbol{Z}}\int q(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{m})\log\frac{q(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{m})}{p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{m})}d\boldsymbol{\theta} \\
&= \sum_{\boldsymbol{m}} q(\boldsymbol{m})\Bigg\{\sum_{\boldsymbol{Z}}\int q(\boldsymbol{Z}|\boldsymbol{m})q(\boldsymbol{\theta}|\boldsymbol{m})\log\frac{q(\boldsymbol{Z}|\boldsymbol{m})}{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{m})}d\boldsymbol{\theta} \\
&\quad + \sum_{\boldsymbol{Z}}\int q(\boldsymbol{Z}|\boldsymbol{m})q(\boldsymbol{\theta}|\boldsymbol{m})\log\frac{q(\boldsymbol{\theta}|\boldsymbol{m})}{p(\boldsymbol{\theta}|\boldsymbol{m})}d\boldsymbol{\theta} \\
&\quad + \sum_{\boldsymbol{Z}}\int q(\boldsymbol{Z}|\boldsymbol{m})q(\boldsymbol{\theta}|\boldsymbol{m})\log\frac{q(\boldsymbol{m})}{p(\boldsymbol{m})}d\boldsymbol{\theta}\Bigg\} \\
&= \sum_{\boldsymbol{m}} q(\boldsymbol{m})\Bigg\{\sum_{\boldsymbol{Z}}\int q(\boldsymbol{Z}|\boldsymbol{m})q(\boldsymbol{\theta}|\boldsymbol{m})\log\frac{q(\boldsymbol{Z}|\boldsymbol{m})}{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{m})}d\boldsymbol{\theta} \\
&\quad + \int q(\boldsymbol{\theta}|\boldsymbol{m})\log\frac{q(\boldsymbol{\theta}|\boldsymbol{m})}{p(\boldsymbol{\theta}|\boldsymbol{m})}d\boldsymbol{\theta} + \log\frac{q(\boldsymbol{m})}{p(\boldsymbol{m})}\Bigg\}, \qquad (4.7)
\end{aligned}
$$

where $p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{m}) = p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{m})p(\boldsymbol{\theta})p(\boldsymbol{m})$. When the variational posterior distributions are denoted by $q_{vb}(\boldsymbol{Z}|\boldsymbol{m}), q_{vb}(\boldsymbol{\theta}|\boldsymbol{m})$, we obtain $\tilde{q}(\boldsymbol{m})$, which is the optimal distribution of $q(\boldsymbol{m})$, by substituting $q_{vb}(\boldsymbol{Z}|\boldsymbol{m})$ and $q_{vb}(\boldsymbol{\theta}|\boldsymbol{m})$ in (4.7) and maximizing $\bar{F}[q(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{m})]$ under the condition $\sum_{\boldsymbol{m}} q(\boldsymbol{m}) = 1$.

That is, $\tilde{q}(\boldsymbol{m})$ is given by

$$\begin{aligned}\tilde{q}(\boldsymbol{m}) &= -\frac{1}{C}p(\boldsymbol{m})\left\{\sum_{\boldsymbol{Z}}\int q_{vb}(\boldsymbol{Z}|\boldsymbol{m})q_{vb}(\boldsymbol{\theta}|\boldsymbol{m})\log\frac{q_{vb}(\boldsymbol{Z}|\boldsymbol{m})}{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{m})}d\boldsymbol{\theta}\right. \\ &\quad \left.+\int q_{vb}(\boldsymbol{\theta}|\boldsymbol{m})\log\frac{q_{vb}(\boldsymbol{\theta}|\boldsymbol{m})}{p(\boldsymbol{\theta}|\boldsymbol{m})}d\boldsymbol{\theta}\right\}, \qquad (4.8)\end{aligned}$$

where C is the normalization constant. Therefore the optimal model is obtained by $\boldsymbol{m}$ maximizing $\tilde{q}(\boldsymbol{m})$. If we give the uniform distribution as the prior distribution $p(\boldsymbol{m})$, the maximization of Eq.(4.8) is equivalent to the minimization of Eq.(4.1).

Chapter 5

Phase Transition in Variational Bayes Learning

In this chapter, we give the asymptotic expansion of the variational Bayes free energy for the Bernoulli mixture in Theorem 5.1.1. This theorem derives the phase diagram and the existence of the phase transition phenomenon (Theorem 5.1.2). We also confirm the phase transition phenomenon by the experiments using the artificial data.

5.1 Asymptotic Expansion of Variational Bayes Free Energy

In this section, we show the theoretical behavior of the variational Bayes free energy $\bar{F}(\boldsymbol{X})$ of the Bernoulli mixture.
In practical applications, the true distribution is unknown, hence the appropriate number of components is also unknown. If the true distribution can not be realized by the statistical model, then $\bar{F}(\boldsymbol{X})$ is very large. Therefore, in order to make a theoretical foundation for statistical model evaluation, we have to study the case when the true distribution is realizable by a statistical model. However, the Bernoulli mixture is a nonidentifiable and singular statistical model, resulting that its variational Bayes free energy has not been clarified. To describe the main theorem, we need the assumption and condition.

Assumption (A). Let $0 \leq K_1^* \leq K_0^* \leq K$. The true distribution from which training samples are taken is the model $p(\boldsymbol{x}|\boldsymbol{\theta}^*)$ represented by the minimum number K_0^* of components and the parameter is given by $\boldsymbol{\theta}^* = \{\boldsymbol{\pi}^*, \boldsymbol{\mu}_1^*, \cdots, \boldsymbol{\mu}_{K_0^*}^*\}$

$$p(\boldsymbol{x}|\boldsymbol{\theta}^*) = \sum_{k=1}^{K_0^*} \pi_k^* B(\boldsymbol{x}|\boldsymbol{\mu}_k^*), \tag{5.1}$$

where

$$\begin{aligned} 0 < \mu_{km}^* < 1 \qquad & (1 \leq k \leq K_1^*), \\ \mu_{km}^* = 0 \text{ or } 1 \qquad & (K_1^* + 1 \leq k \leq K_0^*). \end{aligned}$$

In the following discussion, we will use the notation $\Delta K^* = K_0^* - K_1^*$. Since the Bernoulli mixture is nonidentifiable, the mapping from the parameter to the probability distribution is not one-to-one. We assume the following condition that the estimated distribution converges to the true distribution when N tends to infinity.

Consistency Condition (C). Let Θ_0 be the set of true parameters

$$\Theta_0 = \{\boldsymbol{\theta} \; ; \; p(\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta}^*)\}.$$

The empirical probabilities are defined by using Eq.(4.5)

$$P_k \equiv \frac{N_k}{N}, \quad p_{km} \equiv \frac{\nu_{km}}{N}, \quad p'_{km} \equiv \frac{\nu'_{km}}{N}.$$

Then the distance between $\hat{\boldsymbol{\theta}} \equiv \{P_k, p_{km}, p'_{km}\}$ and Θ_0 converges to zero in probability.

In this paper, we adopt the assumption (A) and the condition (C), and prove the following theorems.

Theorem 5.1.1 *Assume (A) and (C). Let K_0 be the number of components which satisfy $P_k > 0$ and K_1 the number of components which satisfy $0 <$*

$p_{km}, p'_{km} < 1$. *Then*

$$\begin{aligned}
&\bar{F}(\boldsymbol{X}) - NS(\boldsymbol{X}) \\
&= \left(\left(\frac{M+1}{2} - a\right) K_1 + \left(\frac{1}{2} - a + Mb\right) \Delta K + Ka - \frac{1}{2}\right) \log N + O_p(1),
\end{aligned}$$

where $\Delta K = K_0 - K_1$ *and* $S(\boldsymbol{X}) = -\frac{1}{N} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta}^*)$.

In this theorem, K_0 and K_1 are random variables which are determined by the minimization of the variational Bayes free energy. The components of K_1 and ΔK are referred to as "nondeterministic (stochastic)" and "deterministic", respectively. We will give the proof of the above Theorem 5.1.1 in Section 5.2. The following theorem claims that they are essentially determined by the hyperparameters.

Theorem 5.1.2 *Assume (A) and (C). The random variables* K_0 *and* K_1 *are determined as follows:*

1. *If* $\frac{M+1}{2} - a > 0$, $\frac{1}{2} - a + Mb > 0$. *Then* $K_1 = K_1^*$, $\Delta K = \Delta K^*$.
2. *If* $\frac{M+1}{2} - a > 0$, $\frac{1}{2} - a + Mb < 0$. *Then* $K_1 = K_1^*$, $\Delta K = K - K_1^*$.
3. *If* $\frac{M+1}{2} - a < 0$, $\frac{1}{2} - a + Mb > 0$. *Then* $K_1 = K - \Delta K^*$, $\Delta K = \Delta K^*$.
4. *If* $\frac{M+1}{2} - a < 0$, $\frac{1}{2} - a + Mb < 0$. *Then,*
   (a) *If* $b > \frac{1}{2}$. *Then* $K_1 = K - \Delta K^*$, $\Delta K = \Delta K^*$.
   (b) *If* $b < \frac{1}{2}$. *Then* $K_1 = K_1^*$, $\Delta K = K - K_1^*$.

Proof. Based on Theorem 5.1.1, two random variables K_0 and K_1 are determined by the minimization of the variational Bayes free energy and vary depending on the hyperparameter (a, b).

Case 1 Both K_1 and ΔK become small, therefore $K_1 = K_1^*$ and $\Delta K = \Delta K^*$.

Case 2 K_1 becomes small and ΔK becomes large, therefore $K_1 = K_1^*$ and $\Delta K = K - K_1^*$.

Case 3 K_1 becomes large and ΔK becomes small, therefore $K_1 = K - \Delta K^*$ and $\Delta K = \Delta K^*$.

Case 4 The behavior of K_1 and ΔK depends on their coefficients.

Case (a) $\frac{M+1}{2} - a < \frac{1}{2} - a + Mb \Leftrightarrow b > \frac{1}{2}$: K_1 becomes large and ΔK becomes small, therefore $K_1 = K - \Delta K^*$ and $\Delta K = \Delta K^*$.

Case (b) $\frac{M+1}{2} - a > \frac{1}{2} - a + Mb \Leftrightarrow b < \frac{1}{2}$: K_1 becomes small and ΔK becomes large, therefore $K_1 = K_1^*$ and $\Delta K = K - K_1^*$.

This completes the proof of Theorem 5.1.2. (Q.E.D.)

As a result, we obtain the following phase transition diagram about the Bayes predictive distribution or the plug-in distribution of the mean parameter shown in Figure.5.1.

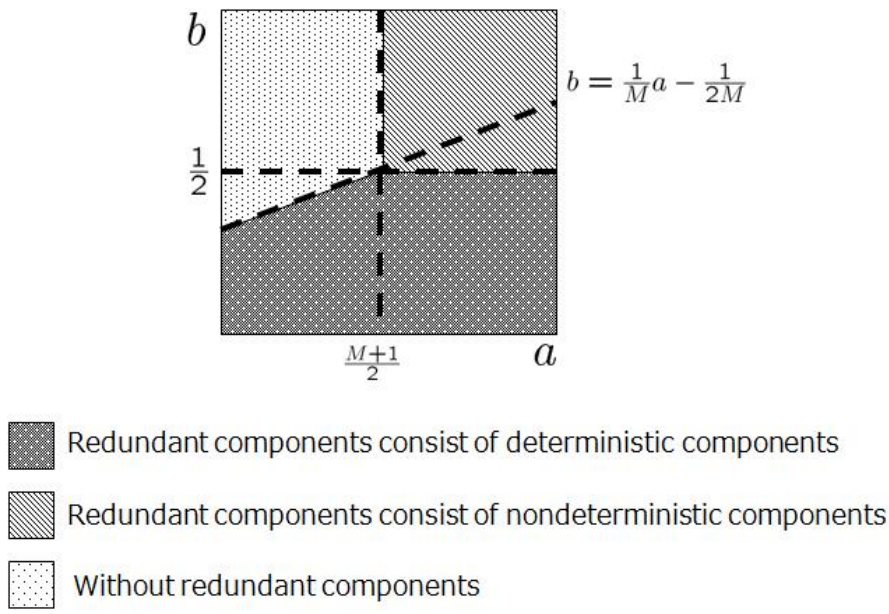

Figure 5.1: Phase transition diagram

When we apply the Bayes method or the variational Bayes method, we always confront the problem how to decide the prior distribution. Figure 5.1, which clarifies the asymptotic behavior of the variational Bayes method, is thought to give the direction of the hyperparameter design. We will discuss the meaning of the phase diagram on clustering analysis in Chapter 7.

By calculating the variational Bayes free energy using Eq.(5.2) for each case in Theorem 5.1.2, we obtain the following Corollary 5.1.1.

Corollary 5.1.1 *Assume (A) and (C), let $f_1(a)$ and $f_2(a,b)$ be given by $g_1(a) = \left(\frac{M+1}{2} - a\right)$ and $g_2(a,b) = \left(\frac{1}{2} - a + Mb\right)$ respectively. Then the variational Bayes free energy is divided into the following 3 cases for hyperparameter (a,b).*

Case 1 *If $a < \frac{M+1}{2}$ and $b > \frac{a}{M} - \frac{1}{2}$. Then,*

$$\begin{aligned}&\bar{F}(\boldsymbol{X}) - NS(\boldsymbol{X})\\ &= \left(g_1(a)K_1^* + g_2(a,b)\Delta K^* + Ka - \frac{1}{2}\right)\log N + O_p(1).\end{aligned}$$

Case 2 *If $a > \frac{M+1}{2}$ and $b > \frac{1}{2}$. Then,*

$$\begin{aligned}&\bar{F}(\boldsymbol{X}) - NS(\boldsymbol{X})\\ &= \left(g_1(a)(K - \Delta K^*) + g_2(a,b)\Delta K_1^* + Ka - \frac{1}{2}\right)\log N + O_p(1).\end{aligned}$$

Case 3 *If $[a < \frac{M+1}{2}$ and $b < \frac{a}{M} - \frac{1}{2}]$ or $[a > \frac{M+1}{2}$ and $b < \frac{1}{2}]$. Then,*

$$\begin{aligned}&\bar{F}(\boldsymbol{X}) - NS(\boldsymbol{X})\\ &= \left(g_1(a)K_1^* + g_2(a,b)(K - \Delta K^*) + Ka - \frac{1}{2}\right)\log N + O_p(1).\end{aligned}$$

Figure.5.2 shows the coefficient of $\log N$ term in the variational Bayes free energy for $K_1^* = 1, \Delta K^* = 1, K = 3$ and $M = 3$.

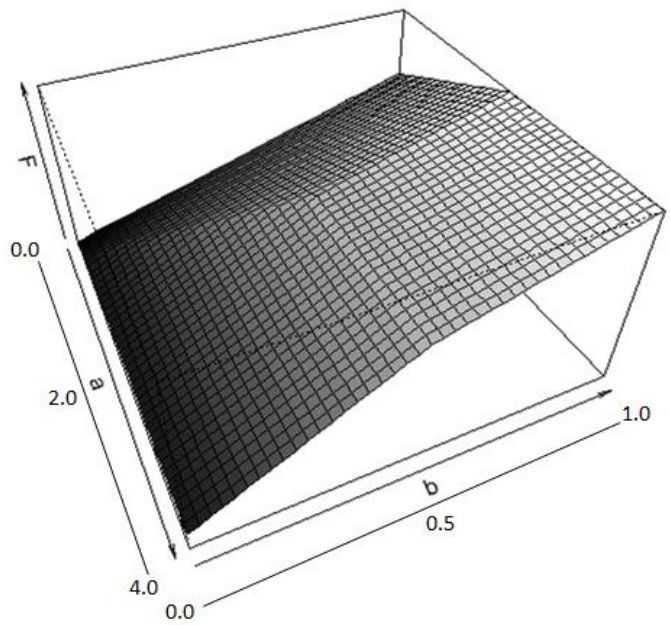

Figure 5.2: Variational Bayes free energy

5.2 Proof of Theorem 5.1.1

In this section, we give the proof of Theorem 5.1.1. The proof is divided into four parts. We first describe the outline of the proof, and then we give some supplements.

5.2.1 Calculation of Asymptotic Expansion

By using the assumption of the mean field approximation of the posterior distribution $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\boldsymbol{Z})q(\boldsymbol{\pi}, \boldsymbol{\theta})$, the variational Bayes free energy is

given by

$$\begin{aligned}
\bar{F}[q] &= \int q(\boldsymbol{Z}')\log\frac{q(\boldsymbol{Z}')}{p(\boldsymbol{X},\boldsymbol{Z}')}d\boldsymbol{Z}' \\
&= \sum_{\boldsymbol{Z}}\int\int q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})\log\frac{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}{p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}d\boldsymbol{\pi}d\boldsymbol{\mu} \\
&= -E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu})] - E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log p(\boldsymbol{Z}|\boldsymbol{\pi})] \\
&\quad -E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log p(\boldsymbol{\pi})] - E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log p(\boldsymbol{\mu})] \\
&\quad +E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log q(\boldsymbol{Z})] + E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log q(\boldsymbol{\pi})] \\
&\quad +E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log q(\boldsymbol{\mu})], \qquad (5.2)
\end{aligned}$$

where p is the probabilistic model and $\boldsymbol{Z}' = (\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})$ is the set of all unobservable variables introduced in Section 3.2. By rewriting the above equation using the Bernoulli mixture, we obtain the following equation (we give the details of the derivation in Section 5.2.2):

$$\bar{F}[q] = \sum_{i=1}^{5} F_i[q], \qquad (5.3)$$

where we define $F_i[q]$ $(i = 1, \cdots, 5)$ by the following equations:

$$\begin{aligned}
F_1[q] &= -\log\frac{\Gamma(Ka)}{\Gamma(a)^K}, \\
F_2[q] &= -\sum_{k=1}^{K}\sum_{m=1}^{M}\log\frac{\Gamma(2b)}{\Gamma(b)^2}, \\
F_3[q] &= \log\frac{\Gamma(\sum_k^K \alpha_k)}{\prod_k^K \Gamma(\alpha_k)}, \\
F_4[q] &= \sum_{k=1}^{K}\sum_{m=1}^{M}\log\frac{\Gamma(\eta_{km}+\eta'_{km})}{\Gamma(\eta_{km})\Gamma(\eta'_{km})}, \\
F_5[q] &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log r_{nk}.
\end{aligned}$$

Here a and b are the hyperparameters of mixing ratio and Bernoulli distributions in Eq.(3.2). By using the asymptotic forms of the log-gamma function

$\log\Gamma(x) = \left(x - \frac{1}{2}\right)\log x - x + \frac{1}{2}\log 2\pi + O\left(\frac{1}{x}\right)$, we have

$$F_3[q] = F_{31}[q] + F_{32}[q] + O_p(1), \quad F_4[q] = F_{41}[q] + F_{42}[q] + F_{43}[q] + O_p(1),$$

where

$$\begin{aligned}
F_{31}[q] &= \left(Ka + N - \frac{1}{2}\right)\log(Ka + N) \\
F_{32}[q] &= -\sum_{k=1}^{K}\left(a + N_k - \frac{1}{2}\right)\log(a + N_k), \\
F_{41}[q] &= \sum_{k=1}^{K}\sum_{m=1}^{M}\left(2b + N_k - \frac{1}{2}\right)\log(2b + N_k) \\
F_{42}[q] &= -\sum_{k=1}^{K}\sum_{m=1}^{M}\left(b + \nu_{km} - \frac{1}{2}\right)\log(b + \nu_{km}) \\
F_{43}[q] &= -\sum_{k=1}^{K}\sum_{m=1}^{M}\left(b + \nu'_{km} - \frac{1}{2}\right)\log(b + \nu'_{km}).
\end{aligned}$$

Without losing generality, we can assume $p_{km} \neq 0$ and $p'_{km} = 0$ $(1 \leq m \leq M, K_1 \leq k \leq K_0)$. Then we have the following equation:

$$\begin{aligned}
F_{32}[q] &= \sum_{k=1}^{K}\left(a + P_k N - \frac{1}{2}\right)\log\left\{N\left(\frac{a}{N} + P_k\right)\right\} + O_p(1) \\
&= \sum_{k=1}^{K_0}\left(a + P_k N - \frac{1}{2}\right)\left\{\log N + \log\left(P_k + O_p\left(\frac{1}{N}\right)\right)\right\} \\
&\qquad + \sum_{k=K_0+1}^{K}\left(a - \frac{1}{2}\right)\log(a) + O_p(1), \\
&= \sum_{k=1}^{K_0}\left\{P_k N\log N + P_k N\log P_k + \left(a - \frac{1}{2}\right)\log N\right\} + O_p(1).
\end{aligned} \tag{5.4}$$

Similarly,

$$F_{31}[q] = \left(Ka + N - \frac{1}{2}\right)\log N + O_p(1), \tag{5.5}$$

$$F_{41}[q] = \sum_{m=1}^{M}\sum_{k=1}^{K_0}\left\{P_k N \log N + P_k N \log P_k + \left(2b - \frac{1}{2}\right)\log N\right\} + O_p(1), \tag{5.6}$$

$$F_{42}[q] = \sum_{m=1}^{M}\sum_{k=1}^{K_0}\left\{p_{km} N \log N + p_{km} N \log p_{km} + \left(b - \frac{1}{2}\right)\log N\right\} + O_p(1), \tag{5.7}$$

$$F_{43}[q] = \sum_{m=1}^{M}\sum_{k=1}^{K_1}\left\{p'_{km} N \log N + p'_{km} N \log p'_{km} + \left(b - \frac{1}{2}\right)\log N\right\} + O_p(1). \tag{5.8}$$

Note that the range of sum of $F_{42}[q]$ differs from that of $F_{43}[q]$ due to $p'_{km} = 0$ $(1 \leq m \leq M, K_1 \leq k \leq K_0)$. Since $F_1[q]$ and $F_2[q]$ are constant for N,we have the following equation by summing up Eq.(5.4)-Eq.(5.7):

$$\begin{aligned}
&\bar{F}[q] - \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log r_{nk} \\
&= \sum_{m=1}^{M}\left(\sum_{k=1}^{K_0}(P_k \log P_k - p_{km}\log p_{km}) - \sum_{k=1}^{K_1}(p'_{km}\log p'_{km})\right) N \\
&\quad - \sum_{k=1}^{K_0} P_k N \log P_k + \left(Ka - \frac{1}{2}\right)\log N + \left(\frac{K_0}{2} - K_0 a\right)\log N \\
&\quad + \left((K_0 - K_1)Mb + \frac{MK_1}{2}\right)\log N + O_p(1),
\end{aligned} \tag{5.9}$$

From Eq.(4.4) and $\sum_{k=1}^{K} r_{nk} = 1$, we have

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log r_{nk} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log \rho_{nk} - \sum_{n=1}^{N}\log\sum_{k=1}^{K}\rho_{nk}. \tag{5.10}$$

Using Eq.(4.3) and $\psi(x) = \log x + o(1)$,

$$\begin{aligned}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\{\psi(\alpha_k)-\psi\left(\sum_{k}^{K}\alpha_k\right)\right\} &= \sum_{k=1}^{K} N_k\left\{\psi(a+N_k)-\psi\left(Ka+N\right)\right\}\\ &= \sum_{k=1}^{K_0} N_k(\log N+\log P_k) - \sum_{k=1}^{K_0} N_k \log N + o_p(1)\\ &= \sum_{k=1}^{K_0} P_k N \log P_k + o_p(1),\end{aligned}$$

$$\begin{aligned}&\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\{\sum_{m=1}^{M}\left(x_{nm}\psi(\eta_{km}) - x_{nm}\psi(\eta'_{km}) + \psi(\eta'_{km}) - \psi(\eta_{km}+\eta'_{km})\right)\right\}\\ &= \sum_{k=1}^{K}\sum_{m=1}^{M}\left\{\nu_{mk}\log(\eta_{km}) + \nu'_{mk}\log(\eta'_{km}) - N_k\log(\eta_{km}+\eta'_{km})\right\} + o_p(1)\\ &= \sum_{m=1}^{M}\sum_{k=1}^{K_0}\left\{-N_k(\log N+\log P_k) + \nu_{mk}(\log N + \log p_{km})\right\}\\ &\qquad + \sum_{k=1}^{K_1}\sum_{m=1}^{M}\left\{\nu'_{mk}(\log N + \log p'_{km})\right\} + o_p(1)\\ &= \sum_{m=1}^{M}\sum_{k=1}^{K_0}\left(-N_k\log P_k + \nu_{mk}\log p_{km}\right) + \sum_{m=1}^{M}\sum_{k=1}^{K_1}\nu'_{nm}\log p'_{km} + o_p(1).\end{aligned}$$

Hence, from Eq.(5.9), we have

$$\begin{aligned}\bar{F}[q] + \sum_{n=1}^{N}\log\sum_{k=1}^{K}\rho_{nk} &= \left(Ka - \frac{1}{2}\right)\log N + \left(\frac{K_0}{2} - K_0 a\right)\log N\\ &\quad + \left((K_0-K_1)Mb + \frac{MK_1}{2}\right)\log N + O_p(1).\end{aligned} \tag{5.11}$$

Finally, we apply the following lemmas to $\sum_{n=1}^{N}\log\sum_{k=1}^{K}\rho_{nk}$. We give the proof of these Lemmas in Section 5.2.3 and 5.2.4. Let the mean parameter be given by $\boldsymbol{\theta}_{vb} = \{\boldsymbol{\pi}_{vb}, \boldsymbol{\mu}_{vb}\} = \{E_{q(\boldsymbol{\pi})}[\boldsymbol{\pi}], E_{q(\boldsymbol{\mu})}[\boldsymbol{\mu}]\}$ ($q(\boldsymbol{\pi}), q(\boldsymbol{\mu})$ are the variational posterior distributions), then we have the following lemmas.

Lemma 5.2.1 *When $\boldsymbol{\theta}_{vb}$ is the mean parameter of the posterior distribution, then we have*

$$\sum_{n=1}^{N} \log \sum_{k=1}^{K} \rho_{nk} = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) + o_p(1). \tag{5.12}$$

Lemma 5.2.2 *When $\boldsymbol{\theta}_{vb}$ is the mean parameter of the posterior distribution, then we have*

$$-\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) = NS(\boldsymbol{X}) + O_p(1). \tag{5.13}$$

Consequently, we obtain the following equation:

$$\begin{aligned} \bar{F}(q) - NS(\boldsymbol{X}) &= \left(Ka - \frac{1}{2}\right)\log N + \left(\frac{K_0}{2} - K_0 a\right)\log N \\ &\quad + \left((K_0 - K_1)Mb + \frac{MK_1}{2}\right)\log N + O_p(1) \\ &= \left\{\left(\frac{M+1}{2} - a\right)K_1 + \left(\frac{1}{2} - a + Mb\right)\Delta K + Ka - \frac{1}{2}\right\}\log N + O_p(1). \end{aligned} \tag{5.14}$$

(Q.E.D.)

5.2.2 Derivation of Eq.(5.3)

Using

$$E_{q(\boldsymbol{\pi})}[\log \pi_k] = \psi(\alpha_k) - \psi\left(\sum_{k}^{K} \alpha_k\right),$$

$$E_{q(\boldsymbol{\mu})}[\log \mu_{km}] = \psi(\eta_{km}) - \psi(\eta_{km} + \eta'_{km})$$

and

$$E_{q(\boldsymbol{\mu})}[(1 - \log \mu_{km})] = \psi(\eta'_{km}) - \psi(\eta_{km} + \eta'_{km}),$$

each term of Eq.(5.2) is given by

$$\begin{aligned} -E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu})}[\log p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu})] &= -\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M}[x_{nm} r_{nk}\{\psi(\eta_{km}) - \psi(\eta_{km} + \eta'_{km})\} \\ &\quad + (1 - x_{nm}) r_{nk}\{\psi(\eta'_{km}) - \psi(\eta_{km} + \eta'_{km})\}], \end{aligned}$$

$$-E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta})}[\log p(\boldsymbol{Z}|\boldsymbol{\pi})] = -\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\psi(\alpha_k) - \psi\left(\sum_{k}^{K}\alpha_k\right)\right),$$

$$-E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta})}[\log p(\boldsymbol{\pi})] = -\log\frac{\Gamma(Ka)}{\Gamma(a)^K} - (a-1)\sum_{k=1}^{K}\left(\psi(\alpha_k) - \psi\left(\sum_{k}^{K}\alpha_k\right)\right),$$

$$\begin{aligned}-E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta})}[\log p(\boldsymbol{\theta})] = &-\sum_{k=1}^{M}\sum_{m=1}^{M}\log\frac{\Gamma(2b)}{\Gamma(b)^2} \\ &- \left((b-1)(\psi(\eta_{km}) - \psi(\eta_{km}+\eta'_{km})) + (b-1)(\psi(\eta'_{km}) - \psi(\eta_{km}+\eta'_{km}))\right),\end{aligned}$$

$$E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta})}[\log q(\boldsymbol{Z})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log r_{nk},$$

$$E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta})}[\log q(\boldsymbol{\pi})] = \log\frac{\Gamma(\sum_k^K \alpha_k)}{\prod_{k=1}^K\Gamma(\alpha_k)} + \sum_{k}^{K}(\alpha_k - 1)\left(\psi(\alpha_k) - \psi\left(\sum_{k}^{K}\alpha_k\right)\right),$$

$$\begin{aligned}E_{q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta})}[\log q(\boldsymbol{\theta})] = &\sum_{k=1}^{K}\sum_{m=1}^{M}\log\frac{\Gamma(\eta_{km}+\eta'_{km})}{\Gamma(\eta_{km})\Gamma(\eta'_{km})} \\ &+ \left((\eta_{km}-1)\left(\psi(\eta_{km}) - \psi(\eta_{km}+\eta'_{km})\right) + (\eta'_{km}-1)\left(\psi(\eta'_{km}) - \psi(\eta_{km}+\eta'_{km})\right)\right).\end{aligned}$$

By summing up above equations, we obtain Eq.(5.3).

5.2.3 Proof of Lemma 5.2.1

Lemma 5.2.1 *When $\boldsymbol{\theta}_{vb}$ is the mean parameter of the posterior distribution, then we have*

$$\sum_{n=1}^{N}\log\sum_{k=1}^{K}\rho_{nk} = \sum_{n=1}^{N}\log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) + o_p(1).$$

Proof. Using Eq.(4.3) and $\psi(x) < \log x - \frac{1}{2x}$,

$$
\begin{aligned}
&\sum_{n=1}^{N} \log \sum_{k=1}^{K} \rho_{nk} \\
&\quad < \sum_{n=1}^{N} \log \sum_{k=1}^{K} \left[\frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} \prod_{m=1}^{M} \left(\frac{\eta_{km}}{\eta_{km} + \eta'_{km}} \right)^{x_{nm}} \prod_{m=1}^{M} \left(\frac{\eta'_{km}}{\eta_{km} + \eta'_{km}} \right)^{1-x_{nm}} \right. \\
&\quad \left. \cdot \exp \frac{1}{2} \left(-\frac{1}{\alpha_k} + \frac{1}{\sum_{k=1}^{K} \alpha_k} - \sum_{m=1}^{M} \left(\frac{x_{nm}}{\eta_{km}} + \frac{1 - x_{nm}}{\eta'_{km}} - \frac{1}{\eta_{km} + \eta'_{km}} \right) \right) \right] \\
&\quad = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \left[\frac{a + P_k N}{Ka + N} \prod_{m=1}^{M} \left(\frac{b + p_{km} N}{2b + P_k N} \right)^{x_{nm}} \prod_{m=1}^{M} \left(\frac{b + p'_{km} N}{2b + P_k N} \right)^{1-x_{nm}} \right. \\
&\quad \left. \cdot \exp \frac{1}{2} \left(\frac{-1}{a + P_k N} + \frac{1}{Ka + N} - \sum_{m=1}^{M} \left(\frac{x_{nm}}{b + p_{km} N} + \frac{1 - x_{nm}}{b + p'_{km} N} - \frac{1}{2b + P_k N} \right) \right) \right].
\end{aligned}
$$

From the supposition $P_k \neq 0$ $(k = 1, \cdots, K_0)$ and $P_k = 0$ $(k = K_0 + 1, \cdots, K)$, we have

$$
\begin{aligned}
&\frac{a + P_k N}{Ka + N} \prod_{m=1}^{M} \left(\frac{b + p_{km} N}{2b + P_k N} \right)^{x_{nm}} \prod_{m=1}^{M} \left(\frac{b + p'_{km} N}{2b + P_k N} \right)^{1-x_{nm}} \\
&\qquad = \begin{cases} p(\boldsymbol{x}_n | \boldsymbol{z}_k, \boldsymbol{\theta}_{vb}), & (k \leq K_0) \\ \frac{a}{Ka+N} \left(\frac{1}{2} \right)^M, & (K_0 < k \leq K), \end{cases}
\end{aligned}
$$

where $\boldsymbol{z}_k = (0, \cdots, \overset{k}{1}, \cdots, 0)$. Hence,

$$
\begin{aligned}
\sum_{n=1}^{N} \log \sum_{k=1}^{K} \rho_{nk} \quad < \quad & \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K_0} p(\boldsymbol{x}_n | \boldsymbol{z}_k, \boldsymbol{\theta}_{vb}) \exp \frac{1}{2} \left(\frac{1}{Ka + N} + \frac{M}{2b + P_k N} \right) \right. \\
& \left. + \sum_{k=K_0+1}^{K} \frac{a}{Ka + N} \left(\frac{1}{2} \right)^M \exp \frac{1}{2} \left(\frac{1}{Ka + N} + \frac{M}{2b} \right) \right\}.
\end{aligned} \tag{5.15}
$$

Here we put $A = \sum_{k=1}^{K_0} p(\boldsymbol{x}_n|\boldsymbol{z}_k, \boldsymbol{\theta}_{vb}) \exp \frac{1}{2}\left(\frac{1}{Ka+N} + \frac{M}{2b+P_kN}\right)$, then,

$$\log \sum_{k=1}^{K} \rho_{nk} \leq \log A + \log \left\{ 1 + \frac{1}{A} \frac{a(K-K_0)}{Ka+N} \left(\frac{1}{2}\right)^M \exp \frac{M}{4b} \right\} + \frac{1}{2(Ka+N)}.$$

Since $\log(1+x) = x + o\left(\frac{1}{x^2}\right)$, we have $\log \left\{ 1 + \frac{1}{A} \frac{a(K-K_0)}{Ka+N} \left(\frac{1}{2}\right)^M \exp \frac{M}{4b} \right\} = o\left(\frac{1}{N}\right)$. Consequently, we have

$$\begin{aligned} \log \sum_{k=1}^{K} \rho_{nk} &< \log \sum_{k=1}^{K_0} p(\boldsymbol{x}_n|\boldsymbol{z}_k, \boldsymbol{\theta}_{vb}) + \frac{1}{2}\left(\frac{1}{Ka+N} + \frac{M}{2b+P_kN}\right) + o_p\left(\frac{1}{N}\right) \\ &= \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) + o_p\left(\frac{1}{N}\right). \end{aligned} \tag{5.16}$$

On the other hand, we can obtain the following equation by using $\psi(x) > \log x - \frac{1}{x}$,

$$\begin{aligned} &\sum_{n=1}^{N} \log \sum_{k=1}^{K} \rho_{nk} \\ &> \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K_0} p(\boldsymbol{x}_n|\boldsymbol{k}, \boldsymbol{\theta}_{vb}) \exp \left\{ \frac{-1}{a+P_kN} - \sum_{m=1}^{M} \left(\frac{x_{nm}}{b+p_{km}N} + \frac{1-x_{nm}}{b+p'_{km}N} \right) \right\} \right. \\ &+ \left. \sum_{k=K_0+1}^{K} \frac{a}{Ka+N} \left(\frac{1}{2}\right)^M \exp\left(\frac{-1}{a} - \frac{M}{b}\right) \right]. \end{aligned}$$

By applying the same argument of Eq.(5.15) for replacing $A = \sum_{k=1}^{K_0} p(\boldsymbol{x}_n|\boldsymbol{z}_k, \boldsymbol{\theta}_{vb}) \exp \left\{ \frac{-1}{a+P_kN} - \sum_{m=1}^{M} \left(\frac{x_{nm}}{b+p_{km}N} + \frac{1-x_{nm}}{b+p'_{km}N} \right) \right\}$, we have

$$\log \sum_{k=1}^{K} \rho_{nk} > \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) + o_p\left(\frac{1}{N}\right). \tag{5.17}$$

From Eq.(5.16) and Eq.(5.17), we obtain

$$\sum_{n=1}^{N} \log \sum_{k=1}^{K} \rho_{nk} = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) + o_p(1).$$

(Q.E.D.)

5.2.4 Proof of Lemma 5.2.2

Lemma 5.2.2 *When $\boldsymbol{\theta}_{vb}$ is the mean parameter of the posterior distribution, then we have*

$$-\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) = NS(\boldsymbol{X}) + O_p(1). \tag{5.18}$$

Proof. We first give the following discrete probability distribution

$$\tilde{p}(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}) = f(\boldsymbol{x}|\tilde{\theta}_1) * \cdots * f(\boldsymbol{x}|\tilde{\theta}_{2^M}),$$

where $f(\boldsymbol{x}|\tilde{\theta}_k) = \begin{cases} \tilde{\theta}_k, & \sum_{m=1}^{M} x_m 2^{m-1} + 1 = k \\ 1, & other \end{cases}$. Then the set of Bernoulli mixtures is included in the above discrete probability distribution models, hence

$$\begin{aligned} -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) - NS(\boldsymbol{X}) &\geq -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{mle}) - NS(\boldsymbol{X}) \\ &\geq -\sum_{n=1}^{N} \log \tilde{p}(\boldsymbol{x}_n|\tilde{\boldsymbol{\theta}}_{mle}) - NS(\boldsymbol{X}), \end{aligned}$$

where $\boldsymbol{\theta}_{mle}$ and $\tilde{\boldsymbol{\theta}}_{mle}$ are the maximum likelihood estimators. Using the property of the empirical Kullback-Leibler divergence $H(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \log \frac{p^*(\boldsymbol{x}_n)}{p(\boldsymbol{x}_n|\boldsymbol{\theta})}$ in non-singlar model,

$$H(\boldsymbol{\theta}_{vb}) \geq H(\tilde{\boldsymbol{\theta}}_{mle}) \geq \frac{C}{N} + o_p\left(\frac{1}{N}\right) \quad ,$$

where C is a random value depending on sample data, but not depending on the number of sample data N. Hence

$$-\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) - NS(\boldsymbol{X}) \geq O_p(1). \tag{5.19}$$

On the other hand, from Section 5.2 and Lemma 5.2.1, we have

$$\bar{F} = G(K_1, \Delta K) - \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{vb}) + O_p(1),$$

where

$$G(K_1, \Delta K) = \left\{ \left(\frac{M+1}{2} - a \right) K + \left(\frac{1}{2} - a + Mb \right) \Delta K + Ka - \frac{1}{2} \right\} \log N.$$

Here K_1 and ΔK are variables determined by minimizing the variational Bayes free energy. In what follows we will use $K_1^*, \Delta K^*$ and $\boldsymbol{\theta}^*$ as the parameters of the true distribution introduced in Assumption (A).

(i) Case of $K_1 \geq K_1^*$ and $\Delta K \geq \Delta K^*$.

The given probabilistic model can compose the true distribution. In addition, the variational Bayes free energy takes a minimum value at $\boldsymbol{\theta}_{vb}$. Therefore we have the following inequation:

$$\begin{aligned} \bar{F} &= G(K_1, \Delta K) - \sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta}_{vb}) + O_p(1) \\ &\leq G(K_1, \Delta K) - \sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta}^*) + O_p(1). \end{aligned}$$

Hence,

$$-\sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta}_{vb}) \leq NS(\boldsymbol{X}) + O_p(1), \tag{5.20}$$

using $-\sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta}^*) = NS(\boldsymbol{X})$.

(ii) Case that $K_1 \geq K_1^*$ and $\Delta K \geq \Delta K^*$ are not satisfied.

We have

$$\sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta}_{vb}) - \sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta}^*) = O_p(N)$$

from $KL(p(\boldsymbol{x}_n | \boldsymbol{\theta}^*) \| p(\boldsymbol{x}_n | \boldsymbol{\theta}_{vb})) > 0$. Hence $\bar{F}$ of Case (i) is smaller than that of this case. As a result, $K_1 \geq K_1^*$ and $\Delta K \geq \Delta K^*$ are always satisfied. Eq.(5.18) is derived from Eq.(5.19) and Eq.(5.20).

(Q.E.D.)

5.3 Experiments

To illustrate the influence of the phase transition phenomenon in actual data, we experimentally investigated the relation between the composition of the redundant components and hyperparameters. The true distribution was given by the left of Figure.5.3, where white means the high probability and the data dimension $M = 4$, the number of components $K^* = 2$. The number of components of learner was given by $K = 3$. The experimental process was as follows:

1. Execute the learning for 100 times with different initial conditions.
2. Extract the component which has the minimum mixing ratio as the redundant component.
3. Sort the redundant components in ascending order of $\bar{F}$.

The result is shown in the right of Figure.5.3.

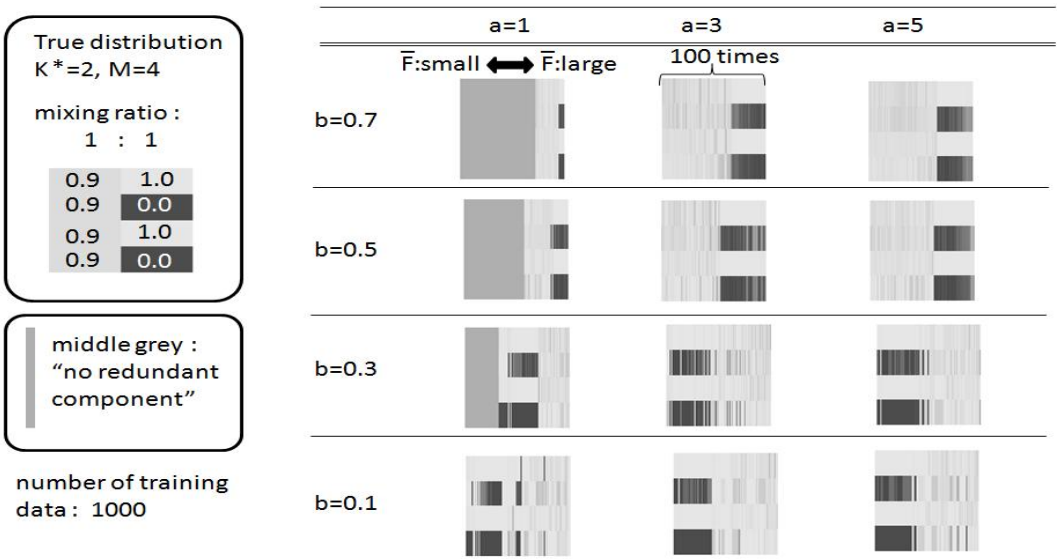

Figure 5.3: Relation between redundant component and hyperparameters

Although the learning results depended on the initial conditions, we can see the following trends derived from Theorem 5.1.2.

- When $a > \frac{M+1}{2} = 2.5$, the redundant components switched from "deterministic component" to "nondeterministic component" around $b = 0.5$.
- When $a < \frac{M+1}{2} = 2.5$, the redundant components switched from "deterministic component" to "no redundant component" at a point less than $b = 0.5$.

5.4 Discussion

In practical case, it is highly probable that the very small clusters are observed as the deterministic components, because the sample size is finite and the small cluster often has strong similarity among its elements. Actually, we will give the example of the minority clusters extracted as the deterministic component using a practical questionnaire in Section 6.2.

Chapter 6

Optimal Hyperparameter Design

In this chapter, we experimentally examine the optimal hyperparameter design method of the variational Bayes learning. The optimality is discussed from the two viewpoints: generalized learning and knowledge discovery.

6.1 Two Design Method of Hyperparameter

Prediction and clustering are major applications of the mixture model. We consider the hyperparameter design methods for both applications.

6.1.1 Hyperparameter for Generalized Learning

The variational predictive distribution is defined by

$$p(\boldsymbol{x}|\boldsymbol{X}) = \int p(\boldsymbol{x}|\boldsymbol{\theta}) q_2(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where $\boldsymbol{X}$ is the set of training data and $q_2(\boldsymbol{\theta})$ is the variational Bayes free energy defined by Eq.(4.2). The variational Bays generalization error G is defined by the Kullback-Leibler distance between the true distribution $p^*(\boldsymbol{x})$ and the variational predictive distribution,

$$G = E\left[\int p^*(\boldsymbol{x}) \log \frac{p^*(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{X})} d\boldsymbol{x}\right],$$

where $E[\cdot]$ denotes the expectation value over all sets of training data $\boldsymbol{X}$. It is thought that the variational Bays generalization error measures the prediction accuracy.

Both the variational Bayes free energy and the variational Bayes generalization error depend on the choice of the hyper parameters (a, b). In the real world problems, we do not know the true distribution, hence we can not directly estimate the variational Bayes generalization error. On the other hand, the variational free energy can be calculated using only the statistical model for a given hyperparameters. As described in Chapter 4.2, the variational Bayes free energy is used for determining the hyperparameter and it is sometimes proposed that the optimal hyperparameter for the minimum free energy is appropriate for the minimum generalization error. In the experiments, we study the effect of hyperparameter to the variational Bayes free energy and the variational Bayes generalization error.

6.1.2 Hyperparameter for Knowledge Discovery

A mixture model such as a Bernoulli mixture is used for the unsupervised clustering problem in application systems. In such cases, knowledge discovery or data mining is more important than the minimum generalization error.

Sometimes it is recommended that, if one has some knowledge about the object, then one had better use the such knowledge in the hyperparameter setting. In contrast, it is often referred that, if one does not have any the prior information, then a uniform distribution such as $a = 1, b = 1$ is appropriate. However, usually we have a purpose for using statistical models, even when we do not have any prior information.

For example, we have a requirement about the size of the clusters. In fact, when we classify the data, we sometimes consider not only main clusters but also small and minority clusters. Because small clusters which make the generalization error be large may contain an important information about data.

In a Bernoulli mixture, we propose that a small hyperparameter b is useful for such minority cluster extraction. If b is set as small, then the prior distribution generates 0 or 1 with high probability, resulting that the

predictive distribution becomes adapted to a small cluster that generates a number of specific terms. In addition, the Theorem 5.1.1 implies that the combination of parameters $a < \frac{M+1}{2}$ and small b enables us to extract the minority cluster without redundant components.

6.2 Experiments

In this section, we show the experimental results on the generalization error and the clustering properties of the variational Bayes learning.

6.2.1 Variational Bayes Free Energy and Variational Bayes Generalization Error

The true distribution was designed to have the parameter described in the left-hand of Figure.6.1. It consists of 3 mixture components and each component is a 5 dimensional Bernoulli distribution $M = 5$. A learning machine was made of 10 components, $K = 10$. The stop condition of the recursive procedure was set as "maximum variation of all parameters $< 10^{-3}$".

We first investigated the behaviors of the variational Bayes free energy and the variational Bayes generalization error. We used $1,000$ samples in one trial and calculated the experimental expectation values over 100 trials. The variational Bayes free energy is shown in Figure.6.1, where (a,b) is the set of hyperparameters of the mixing ratio and the Bernoulli distribution.
The variational Bayes generalization error appears in Fig.6.2. Peaks in the variational Bayes free energy appeared around $(a, b) = (3, 0.5)$. This phenomenon is thought to be related to the phase transition. In addition, Figure.6.2 shows a region of small a and the region around $b = 1$ was stable with respect to the variational Bayes generalization error. Therefore, in order to make the generalization error small, the hyperparameter $(a, b) = (\text{small}, 1)$ is recommended. The behaviors of the variational Bayes free energy was almost same as that of the variational Bayes generalization error, hence the minimum free energy was an appropriate criterion for the optimization of the hyperparameters.

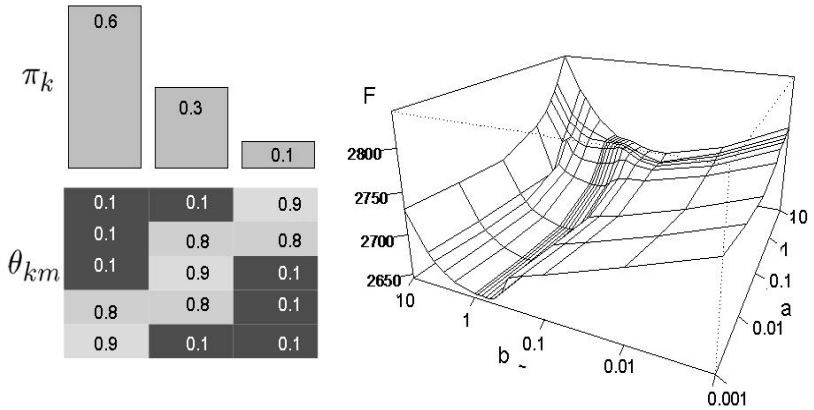

Figure 6.1: (left)True distribution and (right)variational Bayes free energy(a, b : log scale)

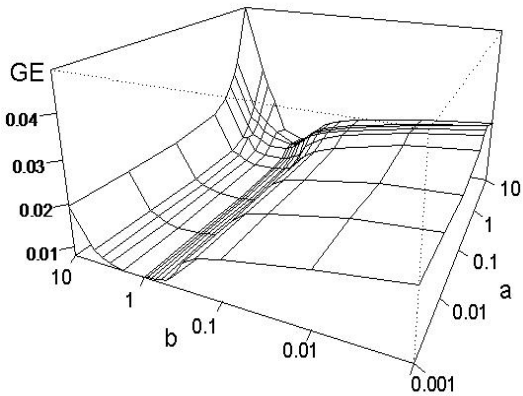

Figure 6.2: Variational Bayes generalization error (a, b : log scale)

6.2.2 Knowledge Discovery

Next we investigated the effect of the hyperparameter to the clustering analysis. We used the true distribution composed of three mixture components, in which one component, as a minority cluster, has a mixing ratio of 0.01. The predictive distribution by mean parameters is shown in Figure.6.3. When both a and b were set large, the learner could not find small clusters. In contrast to the combination of large a and small b, extracted several clusters. Therefore, it is thought that the cluster size can be controlled by adjusting the hyperparameter b. However this result contained the some redundant components. On the other hand, when small a and b were chosen, the learner simultaneously reduced the number of clusters and extracted a minority cluster. Consequently, by changing b with small a, we could obtain the desired cluster size without redundant components.

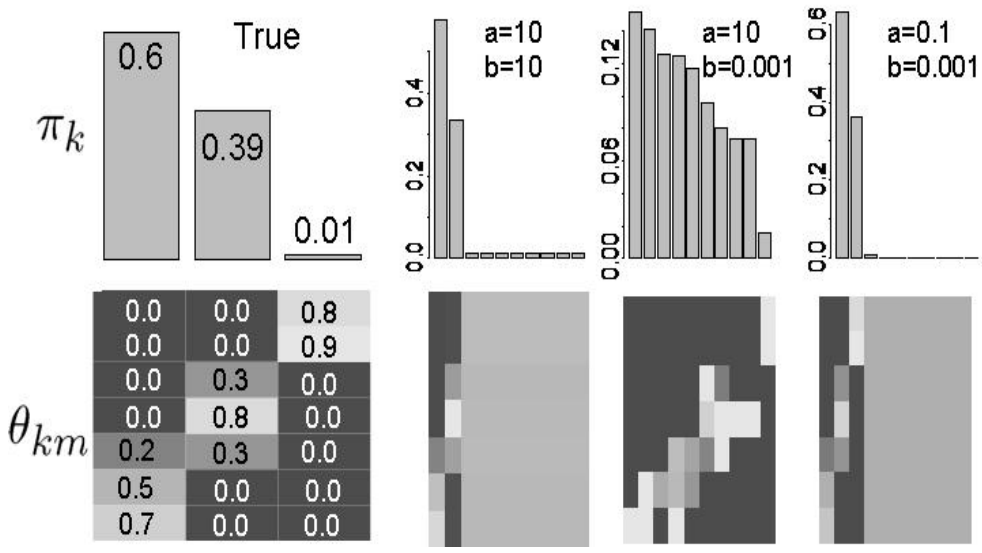

Figure 6.3: Effect of hyperparameter to clustering analysis

6.2.3 Application to Category Classification

Finally we applied the proposed design methods to the practical data obtained from the web site "http://kiwitobes.com/clusters/zebo.txt" [Segaran (2007)]. The data was given as a matrix composed of 83 items and 1750 users, in which elements are assigned a value of 1 if some users want (or own or love) the item, and other elements are assigned a value of 0.

The result is illustrated in Table.6.1 and Table.6.2, where the second line from the top is the mixing ratio of each category and the items are listed in the decreasing order of the probability for each category.
We could obtain these 13 categories when the both hyperparameters a and b were set small (Table.6.1). Here the category1 is a very large cluster which has the mixing ratio about 45%, The members of this category love general items like house, money, job, business. In contrast, the category11,12,13 were very small clusters, which have the mixing ratio of less than 1%. In this case, the probability of each category was expressed as either high probability or low probability, like 0.99997. These results suggest that these categories contain the minority clusters that have very similar interests.

On the other hand, when $(a, b) = (small, large)$ was set, two large cluster were extracted as the general tendency (Table.6.2). The members of category1 love conceptual items, like, money, love, friends, in contrast, the member of category2 loves concrete items, like, laptop, ipod, shoes, and cloths.

Table 6.1: Clustering result for practical data ($a = 0.0001, b = 0.0001$)

category1		category2		$\cdots$	category11		category12		category13	
0.45414		0.34500		$\cdots$	0.00171		0.00102		0.00057	
house	0.46400	laptop	0.19441		xbox 360	0.99997	plane	0.99994	cell phone	0.99990
money	0.19700	house	0.17127		ps3	0.99997	boat	0.99994	dog	0.99990
job	0.05883	ipod	0.15085		psp	0.99997	big house	0.55429	laptop	0.99990
business	0.05241	money	0.11379		ipod	0.99997	house	0.44571	cat	0.99990
clothes	0.05219	computer	0.08726	$\cdots$	mansion	0.33364	mansion	0.00006	big house	0.99990
shoes	0.05175	cell phone	0.07031		sports car	0.00003	sports car	0.00006	horse	0.99990
friends	0.04112	bike	0.04819		bike	0.00003	bike	0.00006	mp3 player	0.99990
big house	0.04047	friends	0.04379		clothes	0.00003	clothes	0.00006	mansion	0.00010

Table 6.2: Clustering result for practical data ($a = 0.0001, b = 1.0$)

category1		category2	
0.45414		0.34500	
house	0.39667	laptop	0.26457
money	0.24632	ipod	0.18966
love	0.07036	house	0.17169
friends	0.06517	shoes	0.14718
job	0.06448	clothes	0.11781
business	0.03925	computer	0.08738
mansion	0.03796	money	0.08401
big house	0.03671	cellphone	0.06589

6.3 Discussion

From the experimental results, we can propose two design methods for hyperparameter optimization. The former method is to minimize the generalization error and the latter for knowledge discovery. The behaviors of the variational Bayes free energy was almost same as that of the variational Bayes generalization error, hence it is thought that the hyperparameters which minimize the variational Bayes free energy also minimize the variational Bayes generalization error. Meanwhile it is possible to control the number of components and the cluster size by adjusting the hyperparameters of mixture ratio and Bernoulli distribution, respectively. We showed that these properties of hyperparameters enable us to extract the general tendency and minority using the practical data. Experimental results demonstrated that the optimal hyperparameters for the different purposes are different from each other.

Chapter 7

Discussion

In this chapter, we discuss the significance of our results from the two viewpoints: the learning theory and the optimal design method of the hyperparameter.

7.1 Discussion from Learning Theory

In statistical physics, the equilibrium state is described by the Boltzmann distribution

$$P_\beta(\boldsymbol{S}) = \frac{1}{Z(\beta)} e^{-\beta H(\boldsymbol{S})},$$

where H is a Hamiltonian of the state variable $\boldsymbol{S} = (S_1, \cdots, S_N)$, β is inverse of the temperature T and $Z(\beta)$ is normalization constant. Then the free energy is defined by

$$F(\beta) = -\log Z(\beta).$$

The free energy plays important roles in statistical physics because it is used to estimation of the macroscopic equilibrium state of the physical system [1] and the mean field approximation is applied to its calculation.

In statistical learning, as we mentioned in Section 2, the Hamiltonian

[1] The free energy for the Hamiltonian $\bar{H}(\boldsymbol{S}, \boldsymbol{h}) = H(\boldsymbol{S}) - \boldsymbol{h} \cdot \boldsymbol{S}$ is called the Helmholtz free energy. The derivation of the Helmholtz free energy gives the mean value for the Boltzmann distribution of the Hamiltonian $H(\boldsymbol{S})$.

$\bar{H}(\boldsymbol{\theta})$ is given by

$$\bar{H}(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) + \frac{1}{N}\log\varphi(\boldsymbol{\theta}),$$

where $H(\boldsymbol{\theta})$ is the empirical Kullback Leibler divergence,

$$H(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\log\frac{p^*(\boldsymbol{x}_n)}{p(\boldsymbol{x}_n|\boldsymbol{\theta})}.$$

Then the posterior distribution is given as the Boltzmann distribution of $\bar{H}(\boldsymbol{\theta})$, that is,

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z_0(\boldsymbol{X})}\exp(-N\bar{H}(\boldsymbol{\theta})).$$

The free energy for the normalization constant $Z_0(\boldsymbol{X})$ is called the normalized free energy and satisfies the following equation,

$$F_0(\boldsymbol{X}) = -\log Z_0(\boldsymbol{X}) = F(\boldsymbol{X}) - NS(\boldsymbol{X}).$$

This free energy also plays important roles in statistical learning, the following asymptotic forms are shown in [Watanabe (2001)],

$$\begin{aligned} E_{\boldsymbol{X}}[F_0(\boldsymbol{X})] &= \lambda\log N - (m-1)\log\log N + O(1), \\ E_{\boldsymbol{X}}[KL(p^*(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{X}))] &\simeq \frac{\lambda}{N} + o\left(\frac{1}{N}\right). \end{aligned} \tag{7.1}$$

On the other hand, the variational Bayes free energy is introduced as the approximation of the free energy using the mean field approximation. Therefore it is expected that the analysis of the variational Bayes free energy gives the information of the asymptotic behavior of the predictive distribution [2] and the generalization error in the variational Bayes learning. We showed the asymptotic behavior of the posterior distribution in Theorem 5.1.2. Here let us discuss the generalization error of the variational Bayes learning.

When we have no information about the data, the uniform distributions given by $(a, b) = (1, 1)$ are often used as the prior distribution. From the

[2] In statistical physics, the first-order phase transition is defined as the discontinuous of the first-order differential of the free energy and it can be explained from the degeneracy of the Fisher matrix by the thermodynamic limit. Whereas in this case the asymptotic behavior means limit for sample size.

Theorem 5.1.2, the predictive distribution which corresponds to $(a, b) = (1, 1)$ belongs to the region of "no redundant components" and its variational Bays free energy is given in Case 1 of Corollary 5.1.1,

$$\begin{aligned}
&\bar{F}_0(\boldsymbol{X}) \\
&= \left(\left(\frac{M+1}{2} - 1\right) K_1^* + \left(\frac{1}{2} - 1 + M\right) \Delta K^* + K - \frac{1}{2}\right) \log n + O_p(1),
\end{aligned}$$

where $\bar{F}_0(\boldsymbol{X})$ is the normalized variational Bayes free energy $\bar{F}_0(\boldsymbol{X}) = \bar{F}(\boldsymbol{X}) - NS(\boldsymbol{X})$. Hence λ_{VB} which is the coefficient of $\log N$ term is given by

$$\begin{aligned}
\lambda_{VB} &= \left(\left(\frac{M+1}{2} - 1\right) K_1^* + \left(\frac{1}{2} - 1 + M\right) \Delta K^* + K - \frac{1}{2}\right) \\
&= \frac{MK_1^*}{2} + K - M\Delta K^* - \frac{K_0^*}{2} - \frac{1}{2}.
\end{aligned}$$

On the other hand, the following upper bound $\bar{\lambda}_{Bays}$ on the coefficient λ in Eq.(7.1) is known for the general mixture model [Yamazaki et al.(2003a,b)],

$$\bar{\lambda}_{Bays} = (K - K_0^*) + \frac{MK_0^* + K_0^* - 1}{2} \quad (M \geq 2). \tag{7.2}$$

Therefore, using $\Delta K^* = K_0^* - K_1^*$, we can obtain

$$\bar{\lambda}_{Bayes} - \lambda_{VB} = \frac{3}{2} M \Delta K^*.$$

Accordingly, if the true distribution has no deterministic components (namely, $\Delta K^* = 0$), then λ_{VB} corresponds to the above upper bound. In addition, when the true distribution has deterministic components, we have $\bar{\lambda}_{Bayes} > \lambda_{VB}$. This implies that the variational posterior is not so different from the true Bayesian posterior.

Furthermore λ is equal to the number of parameters in regular statistical models. Hence if the regular statistic model has the same number of parameters as the Bernoulli mixture model, $\lambda_{regular} = KM + K - 1$. Then we have

$$\lambda_{regular} - \lambda_{VB} = \frac{M}{2}(K - K_1^*) + \frac{K_0^*}{2} + M\Delta K^* - \frac{K}{2} \geq 0.$$

From this result, we can see that the variational Bayes learning is much more effective than the regular statistic models in high-dimension data.

7.2 Discussion from Optimal Design of Hyperparameter

In this section, we would like to discuss the guideline of hyperparameter design of the variational Bayes learning in the Bernoulli mixture. The experimental results in Section 6.2.1 demonstrated that the minimization of the variational Bayes free energy is one of the effective methods of the hyperparameter design for the predictive distribution. Therefore, here we focus on the hyperparameter design for the clustering analysis. We obtained the following properties on the hyperparameters from the experiments of Section 6.2.2:

- The hyperparameter a adjusts the number of components.
- The hyperparameter b adjusts the size of cluster.

However these adjustments are not independent of each other. Therefore the hyperparameter a gives an effect to the size of cluster and the hyperparameter b also gives an effect to the number of components.
Consider next the implications of the phase diagram for the clustering analysis. In general, it is thought that the clustering result including the redundant components is not desirable for clustering analysis. On the other hand, the phase diagram suggests that :

- Redundant components tend to arise when the hyperparameter a is large.
- Redundant components tend to arise for small hyperparameter b even if the hyperparameter a is small.

In practical data analysis, the learning results depend strongly on the data. However, from the above discussion, we can propose the following guidelines of the hyperparameter design for the clustering analysis using the variational Bayes learning in the Bernoulli mixture.

Guideline of hyperparameter design for clustering

1. Adjust the size of the cluster by changing the hyperparameter b based on the following directions:
   - The hyperparameter b should be large to extract the general trend of data.
   - The hyperparameter b should be small to extract the information of minority.
2. Adjust the number of components by changing the hyperparameter a based on the following directions:
   - In order to avoid the clustering with the redundant components, the hyperparameter a less than $\frac{M+1}{2}$ (M : data dimension) is recommended.
   - In order to avoid the clustering with the redundant components, the hyperparameter a should be set smaller as the hyperparameter b becomes small.
3. Repeat the above steps as needed.

Chapter 8

Conclusion

In this thesis, we clarified the asymptotic behavior of the variational Bayes predictive distribution in the Bernoulli mixture model. In addition, we also gave the design guide of the prior distribution based on both the theoretical analysis of the variational Bayes free energy and the experimental results. The main results of this study are summarized as follows:

- Asymptotic expansion of the variational Bayes free energy in Bernoulli mixture was derived.
- Existence of the phase transition phenomenon depending on the hyperparameter of both the mixing ratio and the Bernoulli distribution was shown by using the notion of the deterministic components.
- It was experimentally shown that the optimization of the hyperparameter by minimizing the variational Bays free energy is useful in the hyperparameter design for the prediction.
- Design guide of the hyperparameters for clustering taking into account the purpose of data analysis was presented based on both the experimental results and the asymptotic theory of the variational Bayes free energy.

Appendix

In the appendix, we give supplements on some terminologies appeared in this thesis.

A.1 EM Algorithm of Bernoulli mixture

The EM(Expectation-Maximization) algorithm is well known as a calculation method of MLE for data including missing data [Dempster et al. (1977)]. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ denotes the data set and the probabilistic model with parameter $\boldsymbol{\theta}$ is given by

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}),$$

where $\boldsymbol{Z}$ is unobservable data such as the missing data or the hidden variables. Then we have the following decomposition on $\log p(\boldsymbol{X}|\boldsymbol{\theta})$,

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})),$$

where $\mathcal{L}(q, \boldsymbol{\theta})$ is given by

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}.$$

EM algorithm gives the parameter $\boldsymbol{\theta}$ which maximizes $\log p(\boldsymbol{X}|\boldsymbol{\theta})$ by the following steps:

1. E-step: Maximize $\mathcal{L}(q, \boldsymbol{\theta})$ by optimizing q under the fixed $\boldsymbol{\theta}$.
2. M-step: Maximize $\mathcal{L}(q, \boldsymbol{\theta})$ by optimizing $\boldsymbol{\theta}$ under the fixed q.

$q(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$ denotes $q(\boldsymbol{Z})$ which is optimized under the fixed parameter $\boldsymbol{\theta}^{old}$. Then we have the following expression of $\mathcal{L}(q, \boldsymbol{\theta})$.

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \log q(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \\
&= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) + const,
\end{aligned}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}).$$

Therefore $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ is used for calculation of M-step in general. In practical use, we need to calculate the above E-step and M-step for the concrete probabilistic model. For example, EM algorithm for the Bernoulli mixture Eq.(3.1) is given as follows:

<u>*E-step*</u>

$$r_{nk} = \frac{\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)}{\sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)},$$

<u>*M-step*</u>

$$N_k = \sum_{n=1}^{N} r_{nk}, \quad \bar{\boldsymbol{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n,$$

$$\boldsymbol{\mu}_k = \bar{\boldsymbol{x}}_k, \quad \pi_k = \frac{N_k}{N}.$$

A.2 Conjugate Prior

We introduced the Dirichlet distribution and the Beta distribution as the conjugate distributions in Chapter 3. In this section, we explain the general form of the conjugate prior [Bishop (2006)]. Let us start by giving the definition of the exponential family. The exponential family on $\boldsymbol{x}$ is defined by

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left[{}^t\boldsymbol{\eta}u(\boldsymbol{x})\right],$$

where $\boldsymbol{\eta}$ is a parameter called a natural parameter and u is an arbitrary function. This family includes a lot of important distributions such as Gaussian distribution, Bernoulli distribution, multinomial distribution, and so on.

Example A.2.1 *Gaussian distribution*

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Let η_1, η_2 be given by

$$\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = \frac{-1}{2\sigma^2},$$

then $p(x|\mu,\sigma^2)$ is rewritten as

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}}}\cdot(-2\eta_2)^{\frac{1}{2}}\exp\left(\frac{\eta_1^2}{4\eta_2}\right)\exp\left[(\eta_1,\eta_2)\begin{pmatrix} x \\ x^2 \end{pmatrix}\right].$$

This means that the Gaussian distribution belongs to the exponential family.

If the posterior distribution is in the same family as the prior distribution, the prior distribution is called the conjugate prior. This property is useful to derive the variational Bayes posterior and the predictive distribution. It is known that the conjugate prior of the exponential family is given by the following equation:

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\{\nu\left({}^t\boldsymbol{\eta}\boldsymbol{\chi}\right)\},$$

where $\boldsymbol{\chi}, \nu$ are parameters and $f(\boldsymbol{\chi},\nu)$ is normalization constant.

A.3 Noninformative Prior and Jeffreys' Prior

In this thesis, we gave the guidelines for the hyperparameter design method in accordance with the purpose. Aside from this, there is a notion referred to as noninformative prior which has less influence on the inference. The simplest example of the noninformative prior is the uniform distribution. However the integral of $p(\theta) = \text{const.}$ over θ diverges, hence it is impossible to normalize this distribution. This prior is called improper prior and sometimes used, when the posterior distribution becomes proper.
Next, let the prior of the parameter θ be given by

$$p(\theta) = c \ \ (c : \text{const.}), \ \ \theta \in [a, b].$$

When we give the parameter transform $\xi = \theta^2$, the probabilistic distribution $p(\xi)$ is given by

$$p(\xi) = p(\theta) \left| \frac{d\theta}{d\xi} \right| = c \cdot 2\xi$$

and this is not the uniform distribution. Therefore the condition of uniformity depends on the parameterization. However the uniformity of the distribution is preserved for the shift transform as follows.

Shift invariant prior
The condition that the probabilistic distribution $p(\theta)$ is invariant under the transform $\eta = \theta + c$ (c : const.) is given by

$$\int_A^B p(\theta) d\theta = \int_{A-c}^{B-c} p(\theta) d\theta = \int_A^B p(\theta - c) d\theta$$

for any A, B ($A < B$). Hence $p(\theta) = p(\theta - c)$ is satisfied for any c, $p(\theta) =$ const. is derived. This prior does not depend on the position of the origin.

By developing this notion further, we can obtain the Jeffreys' prior. The Jeffreys' prior is given by the same formula for any coordinate systems and defined by using the Fisher information matrix

$$I(\boldsymbol{\theta}) = I_{ij}(\boldsymbol{\theta}) \ \ = \ \ \int \frac{\partial L(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial L(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_j} p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x},$$

where $L(\boldsymbol{x}, \boldsymbol{\theta}) = \log p(\boldsymbol{x}|\boldsymbol{\theta})$.

Definition A.3.1 *Jeffreys' prior* *A probability density function $\varphi(\boldsymbol{\theta})$ on $\mathbb{R}^d$ is called Jeffreys' prior on Θ if*

$$\varphi(\theta) = \begin{cases} \frac{1}{Z}\sqrt{\det I(\boldsymbol{\theta})}, & (\boldsymbol{\theta} \in \Theta) \\ 0, & (\text{otherwise}), \end{cases}$$

where Z is normalization constant.

At a singularity, $\det I(\theta) = 0$, hence $\varphi(\theta) = 0$. Let $\boldsymbol{\xi} = g(\boldsymbol{\theta})$ be a diffeomorphism from an open set $\Theta \subset \mathbb{R}^d$ to an open set $\Xi \subset \mathbb{R}^d$. Then the Fisher information matrix of $p(\boldsymbol{x}|g(\boldsymbol{\theta}))$ is given by

$$\begin{aligned} I_{ij}(\boldsymbol{\xi}) &= \int \frac{\partial}{\partial \theta_i} L(\boldsymbol{x}, g(\boldsymbol{\theta})) \frac{\partial}{\partial \theta_j} L(\boldsymbol{x}, g(\boldsymbol{\theta})) p(\boldsymbol{x}|g(\boldsymbol{\theta})) d\boldsymbol{x} \\ &= \sum_k \sum_l \frac{\partial \xi_k}{\partial \theta_i} \frac{\partial \xi_l}{\partial \theta_j} \int \frac{\partial}{\partial \xi_k} L(\boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial}{\partial \xi_l} L(\boldsymbol{x}, \boldsymbol{\xi}) p(\boldsymbol{x}|\boldsymbol{\xi}) d\boldsymbol{x}. \end{aligned}$$

Therefore,

$$\det I(\boldsymbol{\theta}) = |g'(\boldsymbol{\theta})|^2 \det I(\boldsymbol{\xi}).$$

From the above, we have the following equation for the Jeffreys'priors $\frac{1}{Z}\sqrt{I(\boldsymbol{\xi})}$ and $\frac{1}{Z}\sqrt{I(\boldsymbol{\theta})}$ on respective spaces.

$$\begin{aligned} \frac{1}{Z}\sqrt{\det I(\boldsymbol{\xi})} d\boldsymbol{\xi} &= \frac{1}{Z}\sqrt{\det I(g(\boldsymbol{\theta}))} |g'(\boldsymbol{\theta})| d\boldsymbol{\theta} \\ &= \frac{1}{Z}\sqrt{\det I(\boldsymbol{\theta})} d\boldsymbol{\theta} \end{aligned}$$

This means that Jeffreys' prior is defined independently of coordinates. When the Jeffreys' prior is employed, it is known that the learning coefficient λ and m given by the coefficient of Eq.(2.3) satisfies the following (1) or (2)[Watanabe (2009)]:

(1) $\lambda = \frac{d}{2},\ m = 1$

(2) $\lambda > \frac{d}{2}$

This result shows that the Jeffreys' prior is not appropriate for the statistical estimation in general.

Meanwhile the following experimental results on the model selection was also shown using a 3-layer neural net [Nishiue & Watanabe (2001)].

The uniform distribution or the Jeffreys' prior was employed as the prior and the model was selected by minimization of the free energy in Bayesian learning.

- If the true distribution was included in models, Jeffreys' prior selected the true distribution with higher probability than the uniform distribution.
- If the true distribution is not included in models, the generalization error of the uniform distribution is smaller than that of the Jeffreys' prior.

Bibliography

[Akaike (1980)] Akaike, H. (1980). Likelihood and Bayes procedure. *In J. M. Bernald (Ed.),Bayesian statistics*, Valencia, Spain, University Press, 143–166.

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

[Amari et al. (2001)] Amari, S., Ikeda, S., & Shimokawa, H. (2001). Information geometry and mean field approximation: The $\alpha-$projection approach. *Advanced Mean Field Methods-Theory and Practice, eds.D. Saad and M. Opper*, MIT Press.

[Amari & Nagaoka (2000)] Amari, S., & Nagaoka, H. (2000). Methods of Information Geometry. *Translations of Mathematical Monographs*, 91, American Mathematical Society.

[Attias (1999)] Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *In Proceedings of Uncertainty in Artificial Intelligence(UAI '99)*, Stockholm, Sweden, 21–30.

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

[Baldi (2003)] Baldi, P., Frasconi, P., & and Smyth, P. (2003). Modeling the Internet and the Web. Wiley.

[Bishop (2006)] Bishop, C. M. (2006) Pattern Recognition and Machine Learning. Springer-Verlag.

[Dempster et al. (1977)] 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.

[Dubrovin et al. (1991)] Dubrovin, B. A., Fomenko, A. T., & Novikov, S. P. (1991). Modern Geometry - Methods and Applications: Part I: The Geometry of Surfaces, *Graduate Texts in Mathematics*, 93, Springer-Verlag.

[Duda & Hart (1973)] Duda, R. O., & Hart, P. E. (1973). Pattern Classification and Scene Analysis Wiley.

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

[Good (2003)] Good, I. J. (2003). The Estimation of Probabilities: An Essay on Modern Bayesian Methods. Cambridge, MA, MIT Press.

[Hastie et al. (2001)] T.Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of Statistical Learning. Springer-Verlag.

[Hertz et al.(1991)] Hertz, J., Krogh, A. & Palmer, R. G. (1991). Introduction to the Theory of Neural Computation. Addison-Wesley.

[Hinton & Camp (1993)] Hinton, G. E., & van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. *In Proc. of the Conference on Computational Learning Theory*, Santa Cruz, California, USA, 5–13.

[Izenman (2008)] Izenman, A. J. (2008). Modern Multivariate Statistical Techniques. Springer-Verlag.

[Juan & Vidal (2001)] Juan, A., & Vidal, E. (2001). On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35, 2705–2710.

[Juan & Vidal (2004)] Juan, A., & Vidal, E. (2004). Bernoulli mixture models for binary images. *In Proceedings of 17th International Conference on Pattern Recognition (ICPR04)*, 3, 367–370.

[Kaji & Watanabe (2009)] Kaji, D., & Watanabe, S. (2009). Optimal Hyperparameters for Generalized Learning and Knowledge Discovery in Variational Bayes. *Proceedings of 16th International Conference on Neural Information Processing (ICONIP 2009)*, Bangkok, Thailand, 476–483.

[Kaji et al. (2010)] Kaji, D., Watanabe, K., & Watanabe, S. (2010). Phase Transition of Variational Bayes Learning in Bernoulli Mixture. *Australian Journal of Intelligent Information Processing Systems*, 11(4), 35–39.

[Kass & Vos (1997)] Kass, R. E., & Vos, P. W. (1997). Geometrical Foundations of Asymptotic Inference. Wiley.

[Lazarsfeld & Henry (1968)] Lazarsfeld, P. F., Henry, N. W. (1968). Latent structure analysis. Houghton Mifflin, Boston.

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

[MacKay (1995)] MacKay, D. J. C. (1995). Developments in probabilistic modeling with neural networks ensemble learning. *In Proc. of the 3rd Ann. Symp. on Neural Networks*, Springer-Verlag, 191–198.

[Murray & Rice (1993)] Murray, M. K., & Rice, J. W. (1993). Differential Geometry and Statistics. Chapman & Hall.

[Nakajima & Watanabe (2007)] Nakajima, S., & Watanabe, S. (2007). Variational Bayes Solution of Linear Neural Networks and its Generalization Performance. *Neural Computation*, 19(4), 1112–1153.

[Nishiue & Watanabe (2001)] Nishiue, K., & Watanabe, S. (2001). On Inconsistency of Precise Prediction and Knowledge Discovery in Learning

and Estimation. *Workshop on Information-Based Induction. Sciences (IBIS).*

[Oyama & Watanabe (2009)] Oyama, S., & Watanabe, S. (2009). Phase Transition of Generalization Errors in Variational Bayes Learning. *Proc. of 2009 International Symposium on Nonlinear Theory and its Applications, Sapporo*, 133–136.

[Permuter et al.(2003)] Permuter, H., J. Francos, J., & Jermyn, I. H. (2003). Gaussian mixture models of texture and color for image database retrieval. *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, 1, 25–88.

[Robert et al.(2009)] Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys' Theory of Probability revisited. Stat Sci 2009, 24, 141–172

[Segaran (2007)] Segaran, T. (2007). Programming Collective Intelligence: Building Smart Web 2.0 Applications. O 'Reilly Media, Inc.

[Smidl & Quinn (2006)] Šmídl, V., & Quinn, A. (2006) . The Variational Bayes Method in Signal Processing. Springer-Verlag, New York.

[Wang et al. (2003)] Wang, J., Taguri, M., Tezuka, S., Kabashima., Y and Ueda, N. (2003). Computational statistics(in Japanese). Iwanami.

[Wang & Titterington (2004a)] Wang, B., & Titterington, T. M. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, Vol. 20, No. 3, Springer-Verlag, 151–170.

[Wang & Titterington (2004b)] Wang, B., & Titterington, T. M. (2004). Convergence and Asymptotic Normality of Variational Bayesian Approximations for Exponential Family Models with Missing Values. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 20, 577–584.

[Wang & Titterington (2005)] Wang, B., & Titterington, T. M. (2005). Variational Bayes estimation of mixing coefficients. *Deterministic and Statistical Methods in Machine Learning, Lecture Notes in Artificial Intelligence*, 3635, Springer-Verlag, 281–295.

[Warner (1983)] Warner, F. W. (1983). Foundations of Differentiable Manifolds and Lie Groups. *Graduate Texts in Mathematics*, 94, Springer-Verlag.

[Watanabe & Watanabe (2006a)] Watanabe, K., & Watanabe, S. (2006a). Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *Journal of Machine Learning Research*, 7(4), 625–644.

[Watanabe & Watanabe (2006b)] Watanabe, K., & Watanabe, S. (2006b). Variational Bayesian stochastic complexity of mixture models. *Advances in Neural Information Processing Systems*, 18, 1465–1472.

[Watanabe & Watanabe (2007)] Watanabe, K., & Watanabe, S. (2007). Stochastic complexities of general mixture models in Variational Bayesian Approximation. *Neural Computation*, 18(5), 1007–1065.

[Watanabe (2001)] Watanabe, S. (2001). Algebraic Analysis for Nonidentifiable Learning Machines. *Neural Computation*, 13(4), 899–933.

[Watanabe (2009)] Watanabe, S. (2009). Algebraic Geometry and Statistical Learning Theory. Cambridge University Press.

[Watanabe (2010)] Watanabe, S. (2010). Equations of states in singular statistical estimation. *Neural Networks*, 23(1), 20–34.

[Winkler (1995)] Winkler, G. (1995). Image Analysis, Random Fields and Dynamic Monte Carlo Methods: A Mathematical Introduction. Springer-Verlag.

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

[Yamazaki & Watanabe (2003b)] Yamazaki, K., & Watanabe, S. (2003b). Stochastic complexity of bayesian networks. *In Proceedings of Uncertainty in Artificial Intelligence(UAI '03)*, Acapulco, Mexico, 592–599.