論文 / 著書情報
Article / Book Information

題目(和文)	特異モデルにおけるベイズ仮説検定に関する理論的研究
Title(English)	Theoretical Study of Bayesian Hypothesis Testing for Singular Models
著者(和文)	仮屋夏樹
Author(English)	Natsuki Kariya
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11891号, 授与年月日:2021年3月26日, 学位の種別:課程博士, 審査員:渡邊 澄夫,三好 直人,金森 敬文,山下 真,中野 張
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第11891号, Conferred date:2021/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Thesis

Theoretical Study of Bayesian Hypothesis Testing for Singular Models

December 2020

Natsuki Kariya

Department of Mathematical and Computing Science,
Tokyo Institute of Technology

Abstract

In this thesis, we studied Bayesian hypothesis testing for singular models theoretically.

Hypothesis testing using mixture models has been widely used in various fields, but it is not easy to give the foundation of it theoretically, due to singularities in the models. Many statisticians have studied this problem because it is both appealing and challenging.

In recent years, the theory of statistical inference for singular models has been developed based on Bayesian statistics and algebraic geometry. Bayesian singular learning theory has been proved to be a very useful framework for the theoretical treatment of singular models.

We attempted a theoretical analysis of hypothesis testing for singular models from the perspective of Bayesian singular learning theory. Technically, the theory of hypothesis testing requires probability distribution of the test statistic, and it needs some theoretical extension that has not been sufficiently addressed, due to factors such as the need to include higher-order terms beyond the results obtained for statistical inference. Such problems have been hardly tackled despite their importance, as far as the author knows.

This thesis presents several new results obtained through the study. More specifically, we studied the test of homogeneity for normal mixture, because it is one of the most typical singular models.

Our first result is that the asymptotic distributions of the marginal likelihood ratio, which is the test statistics of the most powerful test in the Bayesian hypothesis testing, are derived theoretically in several cases.

Second, we applied the variational Bayes approximation for the first time, which is known useful for approximating the posterior distribution, to the problem of hypothesis testing for singular models. We could derive the asymptotic behavior of variational Bayesian free energy, which is the test statistics of our hypothesis testing, ad construct the hypothesis specifically.

Furthermore, we analyzed the type I and the type II errors in Bayesian hypothesis testing for singular models theoretically. The relationship between the errors in the hypothesis test and the real log canonical threshold, which is important

quantities that characterize the geometrical properties of singular models around their singularities are derived.

Acknowledgements

I would like to express my sincerest gratitude to Professor Sumio Watanabe for his patient and cheerful supervision throughout the course of my study. I greatly indebted him for generously accepting to supervise me in the doctoral course, although my major until my master's degree was physics. He taught me not only the rigor of the discipline of mathematics but also its universality and beauty. Thanks to him, I have had an irreplaceable experience in life, through the discussion with him. I would like to acknowledge Prof. Takafumi Kanamori, Prof. Naoto Miyoshi, Prof. Yumiharu Nakano, and Prof. Makoto Yamashita for illuminative comments and careful reading of the thesis. I would like to thank Mizuho Information & Research Institute, Inc., the company I work at, for allowing me to enter the doctoral course, and all my colleagues for their support. During my doctoral course, I have stimulated by the seminar in Watanabe Lab. I wish to thank all the previous and present members of Watanabe Lab., in particular, N. Hayashi, and K. Ohashi, for the discussion with them and constructive comments regarding my study. Last but not least, I wish to thank my family for their continuous support and encouragement for the completion of this study.

Contents

Chapter 1

Introduction

1.1 Interest and motivation

We are often faced with the need to analyze data that appear to have multiple peaks, but these are difficult to be described by a probability distribution with a single peak. Mixture models have been used in various fields including pattern recognition, clustering analysis, and anomaly detection to date as a suitable means of describing such models (McLachlan and Peel [2000]).

Let us show an example. Figure 1.1 is data from the results of seven studies on the age of onset of schizophrenia, which consist of 99 females and 152 males (Lewine [1981]). According to Lewine [1981], there are two types of schizophrenia: early-onset groups and late-onset groups. The former group is considered to be mainly male and the latter mainly female. This data set contains observed values of only male data. In Everitt et al. [2001], they fitted the male data using a two-component normal mixture model. The detail of their analysis and the results are beyond the scope of the present thesis, but it would suffice as an evidence that mixture models play an important role in the analysis of the actual data.

When using mixture models, to determine the number of components for describing the data is a very important problem. The hypothesis test is considered to be one of the useful tools for this purpose, and this type of the hypothesis tests is called testing homogeneity. The simplest but most important case is to decide which is the appropriate hypothesis, one component distribution, or two-component mixture. Testing homogeneity has been an important issue for various mixture models, especially for normal mixtures (Chauveau et al. [2019]).

Theoretically, mixture models often have singularity in their parameter space and this greatly affect the problem of the statistical inference and hypothesis testing.

Let us see the simplest case, two-component mixture model such as,

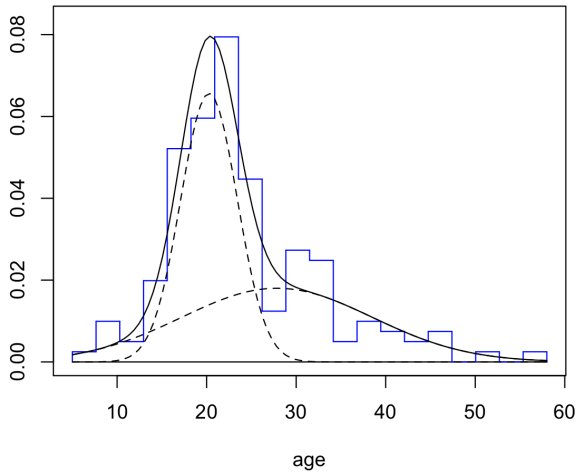

Fig. 1.1: The age of onset of schizophrenia of 152 males studied by Lewine [1981]. We fit the data by two-component normal mixture model. For fitting, we used R package MixtureInf.
http://cran.r-project.org/package=MixtureInf

$$p(X|a,b) = (1-a)f(X|b=b_0) + af(X|b) \tag{1.1}$$

where $p(X|a,b)$ is a probabilistic model we consider and a and b are parameter. b_0 is a fixed constant. $f(X|b)$ is a probabilistic density function.

We assume that the the data $\{X_1, X_2, \ldots X_n\}$ are generated from the true probabilistic model $q(x)$. What we should do is to determine the best value of a and b that approximate $q(x)$ sufficiently well.

This is a very common situation when the statistical inference and hypothesis testing are needed. It is natural to expect that the best (a,b) to be uniquely determined by the well-known method. (e.g. Maximum Likelihood Estimation)

However, This is not always the case. If $q(x)$ is $f(X|b_0)$, the best parameter sets $(a.b)$ is not uniquely determined. Actually, all parameter sets (a,b) which satisfies $a(b-b_0)=0$ are the same. (see Figure 1.2.)

In such a case, the Fisher information matrix becomes singular and there is no guarantee that conventional methods such as Maximum Likelihood Estimation work well, because such conventional methods assume the uniqueness of the solution.

In singularity, the Fisher information matrix becomes singular and this leads that the log likelihood ratio not converging to χ^2 distributions, unlike the case of the regular models. This is why the problem of testing homogeneity for mixture models is theoretically challenging.

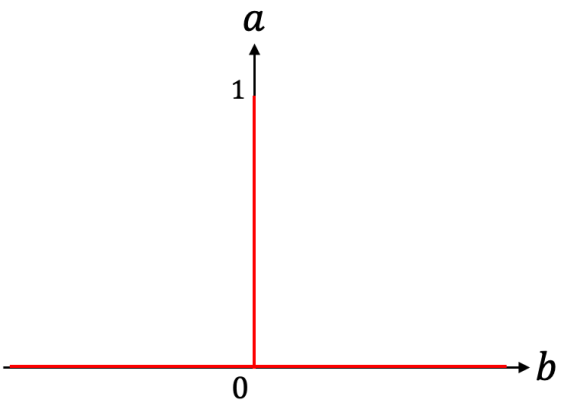

Fig. 1.2: The singularity of the model (1.1) in the parameter space a-b plane. The figure is the case that $b_0 = 0$. In such a case, $ab = 0$ (The red line in the figure) is the set of the singularity.

As an example, In Figure 1.3, we plot the log likelihood ratio of the two-component normal mixture, $(1-a)\mathcal{N}(0,1) + a\mathcal{N}(b,1)$ against the single component normal distribution $\mathcal{N}(0,1)$. The sample size is set as $n = 100$, and the sample is generated from the standard normal distribution $\mathcal{N}(0,1)$. It is clearly shown that the log likelihood ratio is localized around the singularity $ab = 0$, and the distribution of it is totally different from the normal distribution.

The problem has a long history. One of the achievement on the asymptote of the log likelihood ratio was developed in Ghosh and Sen [1985]. They developed the asymptotic theory of the test statics under some assumption of the identifiability, the mixed two distribution can be identified.

In Hartigan [1985], the likelihood ratio for testing homogeneity for normal mixture is studied. it is proved that the likelihood ratio test statistics goes to infinity when $n \to \infty$, if the range of the parameter is not bounded.

Actually, removing the condition assumed in Ghosh and Sen [1985] seems challenging. Many researchers have tackled this problem on their respective models, (Garel [2001] Liu and Shao [2003] Liu and Shao [2004]). For example, in Garel [2001], the asymptote of the test statistics, for testing homogeneity between the simple gaussian against the simple mixture, is represented by using the maximum of the gaussian process.

Also, various methods have been proposed to circumvent difficulties. For example, the modified likelihood ratio test, a method that adds a regularization term to the test statistics (Chen et al. [2001] Chen et al. [2004]), a D test Charnigo and Sun [2004], and applying an expectation-maximization (EM) algorithm for calculating the modified likelihood ratio (Chen and Li [2009] Chen et al. [2012]),

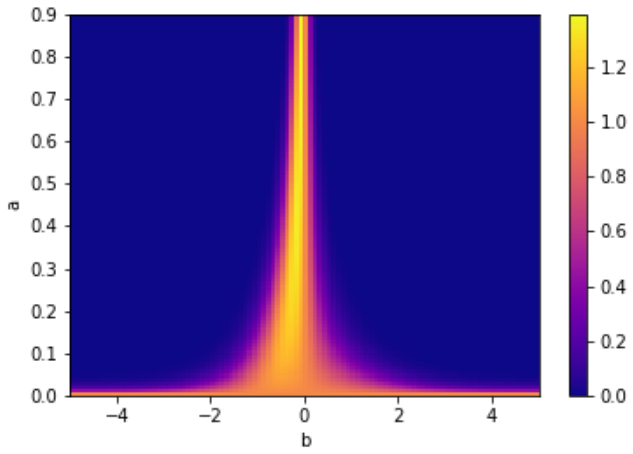

Fig. 1.3: An example of the log likelihood ratio of the two-component normal mixture, $(1-a)\mathcal{N}(0,1)+a\mathcal{N}(b,1)$ against the single component normal distribution $\mathcal{N}(0,1)$. The sample size is set as $n=100$, and the sample is generated from the standard normal distribution $\mathcal{N}(0,1)$, and the log likelihood ratio is plotted for each (a,b).

and so on. In the modified likelihood ratio test and EM test, the regularization term is added to the original likelihood ratio. These are known as very useful and practical method for testing homogeneity Li et al. [2016].

On the other hand, there are few studies based on a *Bayesian* approach. As we see the next subsection, the learning theory for statistical inference based on Bayesian formalism, to treat singular models have been developed recently and works well. Although the application of the Bayesian learning theory to the statistical hypothesis testing is very limited, it is natural to expect that this would be effective. This is one of the motivation of the studies in the thesis.

1.2 Bayesian singular learning theory

1.2.1 Summary of Bayesian inference

Let us briefly summarized the statistical inference based on the Bayesian singular learning theory (Watanabe [2018]).

We assume that sample $\{X^n\}$ is independent and identically distributed.

$$\{X_1, X_2, \dots X_n\} \in \mathbb{R}^N \overset{iid}{\sim} q(x). \tag{1.2}$$

To infer the true distribution $q(x)$, we choose the probabilistic model $(p(X|w)$ and the prior $\varphi(w)$. Here, $w \in W \subset \mathbb{R}^d$ means the parameter in the probabilistic model.

The posterior can be written as,

$$p(w|X^n) = \frac{1}{Z_n}\varphi(w)\prod_i p(X_i|w) \tag{1.3}$$

where the normalization constant Z_n is given as

$$Z_n \equiv \int_W dw\varphi(w)p(X^n|w).$$

Z_n is often referred to as the partition function, from the analogy with the statistical physics. It should be noted that Z_n is a probabilistic distribution, because

$$\int dX^n Z_n = 1$$

In the statistical inference, we are interested in the prediction. In the Bayesian statistics, this is performed through the predictive distribution. To obtain a good prediction, we should know the extent to which the predictive distribution approximates the true distribution.

$$p(x|X^n) \equiv \int_W dwp(x|w)p(w|X^n)$$

Although we can not know the true distribution itself, we can estimate how good/bad our model as an approximation of the true distribution. This is amazing feature of the theory of the statistical inference.

There are mainly two indicators for this purpose.

One is the log marginal likelihood,

$$F_n \equiv -\log Z_n. \tag{1.4}$$

F_n is also called as the free energy, from the analogy with the statistical physics.

If we consider the expectation value of F_n with respect to the true distribution, it becomes

$$\begin{aligned}\int q(X^n)F_n dX^n &= -\int q(X^n)\log q(X^n)dX^n + \int q(X^n)\log\frac{q(X^n)}{Z_n)}dX^n \\ &= nS + \int q(X^n)\log\frac{q(X^n)}{Z_n}dX^n\end{aligned}$$

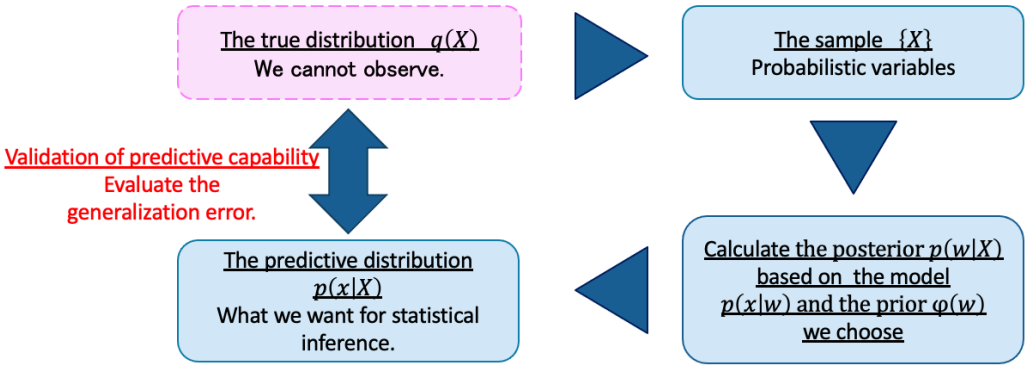

Fig. 1.4: The schematic picture of the sequence of the Bayesian inference.

where $S \equiv -\int q(x) \log q(x) dx$. The S is called as the entropy, from the analogy with the statistical physics.

S is determined solely from $q(x)$, and the second term of the equation (1.5) is minimized when $Z_n = q(X^n)$. Therefore, we can obtain the Z_n which approximates $q(X^n)$ best, by minimizing the expectation value of F_n.

Because the stochastic behavior of the Z_n is determined from the sample which we have already obtained, the minimization of the expectation of the F_n gives us the distribution which explained the obtained sample best, under the assumption of the probabilistic model $p(X^n|w)$.

The other is the generalization loss,

$$G_n \equiv -\int q(x) \log p(x|X^n) dx = S + \int q(x) \log \frac{q(x)}{p(x|X^n)} dx. \tag{1.5}$$

Similar to the case of the expectation value of F_n, we can easily seen that G_n is minimized by $p(x|X^n)$ which is the nearest to $q(x)$, in the sense of Kullback-Leibler divergence. Therefore, we can obtain $p(x|X^n)$ which gives the best prediction of the behavior of $q(x)$, by minimizing G_n.

but in the sense of the indicator of the goodness of the inference, F_n and G_n is a bit different. When we seek for the goodness of the prediction, we use G_n as an indicator, but when we seek for the explanation of the sample obtained, we should use F_n as an indicator.

The schematic picture of the sequence of the Bayesian inference is shown in Figure (1.4).

We should note that G_n and F_n is not independent, because

$$\int q(x) \left(F_{n+1} - F_n\right) dx = G(n).$$

As the main theme of this thesis is hypothesis testing, and the marginal likelihood ratio is often used as a good test statistics, we will see in the next section, we mainly discuss the behavior of F_n in the thesis.

1.2.2 Summary of the singular learning theory

In conventional statistics, the cases that the model is regular against the true distribution(The parameter in the model that minimizes the KL divergence between the true distribution and the model is determined uniquely) are mainly treated.

In this case, the Fisher information matrix is regular, that is, there exists the inverse matrix of the Fisher information matrix. In Bayesian inference, the posterior is well approximated by a Gaussian distribution in these cases.

However, in general, this is *not* holds, and in such cases, the Fisher information matrix becomes singular and the results obtained in the conventional statistics based on this are not justified. (see Figure (1.5)).

For example, the well-known maximum likelihood estimation may fail in singular cases. Also, the well-known χ^2 test is no longer valid in singular cases.

It should be noted that such circumstances are very common in today's data analysis. This is because the models we use become more complex and complex, and they tend to become singular because such complex models generally have high expressibility and redundancy. For example, mixture models such as normal mixture, Bernoulli mixture, multinomial mixture, etc., hidden Markov model, neural network, reduced rank regression, and so on.

Establishing a learning theory that can treat singular models has been a major challenge for many years. Watanabe shows that by performing proper coordinate transformations whose existence is guaranteed based on the Hironaka's resolution of singularities (Atiyah [1970]), it is possible to calculate the renormalized posterior distribution. He also derived the general asymptotic form of the free energy when the true distribution is in the singularity (Watanabe [2001]) such as,

$$F_n = nL_n(w_0) + \lambda \log n - (m-1) \log \log n + \mathcal{O}(1), \tag{1.6}$$

where $L_n(w_0)$ means the empirical log loss function,

$$L_n(w) \equiv -\frac{1}{n} \sum_i \log p(X_i|w),$$

at singularity $w = w_0$.

The constants in equation (1.6) reflects the geometrical property of the parameter space around the singularity. λ is known as a real canonical log threshold (RCLT) in algebraic geometry, and m is known as a multiplicity in algebraic geometry.

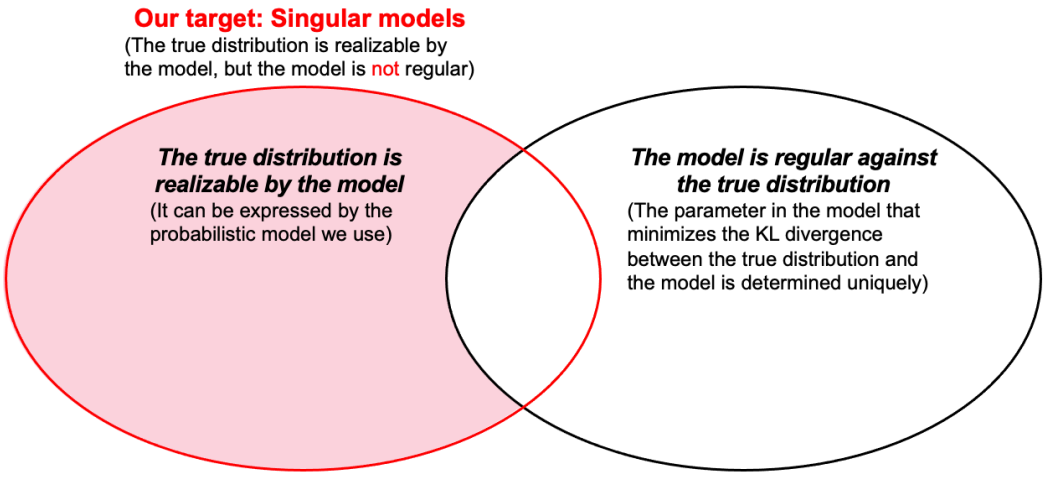

Fig. 1.5: The schematic picture of the venn diagram with respect to the relation between the true distribution and the probabilistic model.

This asymptote is clearly different from those obtained in the regular cases,

$$F_{regular} = nL_n(w^*) + \frac{d}{2}\log n + \mathcal{O}(1)$$

where w^* is the maximum likelihood estimator, and d means the dimension of the parameter space.

It can be readily shown that the generalization error G_n becomes

$$G_n = L(w_0) + \frac{\lambda}{n} + o\left(\frac{1}{n}\right) \tag{1.7}$$

where $L(w_0)$ means the log loss function,

$$L_n(w) \equiv -\int q(x)\log p(x|w)dx$$

at singularity $w = w_0$. When the true distribution is realizable, it is equal to the entropy S.

Therefore, the leading term of G_n that depends on n is $\mathcal{O}(\frac{1}{n})$, and this is determined by RCLT. In this sense, RCLT is a very important quantity for discussing the generalization performance of the model.

The results on the general theory derived above are a brilliant achievement, but the practical use of these results requires the determination of coefficients for each model. Studies have clarified that the RCLT for normal mixture (Yamazaki

and Watanabe [2003]), reduced rank regression (Aoyagi and Watanabe [2005]), non-negative matrix factorization (Hayashi and Watanabe [2017]), and so on.

It is often founded that the RCLT of singular models are smaller than $\frac{d}{2}$, and in this sense, Bayesian inference is superior to the conventional maximum likelihood estimation with respect to the generalization performance.

While RCLT for each singular models has been clarified, the stochastic behavior of the free energy for each models is not fully studied. One of the main difficulties for this is that the stochastic term in the asymptote of the free energy is $\mathcal{O}(1)$. Therefore, the analysis on higher-order terms is required to grasp the stochastic behavior of the free energy and relevant quantities.

As we will see in the next chapter, the stochastic behavior of the free energy needs to be clarified for constructing the hypothesis testing. This is one of the main themes of the present thesis.

1.3 Organization of the thesis

In this thesis, we study the hypothesis testing for singular models, mainly focusing on normal mixtures, one of the typical examples of the singular models. We apply the Bayesian singular learning theory that is a very strong framework to analyze the problem and already is known to be valid for analyzing the problem of the statistical inference.

However, as we commented in the previous section, to analyze the problem of the hypothesis testing requires the analysis of the high order term in the test statistics and it is not necessarily trivial. As far as the author knows, such kind of studies hardly exists.

In this thesis, the results of our study are presented as follows.

In Chapter 2, we briefly review the basics of hypothesis testing and the Bayesian version of hypothesis testing. We derive the conventional χ^2 test and point out the limitation of it when using singular models. Also, we briefly review the previous several studies on the hypothesis testing based on the Bayesian singular theory .

In Chapter 3, one of the main results of this thesis is presented (Kariya and Watanabe [2020a]). The asymptote of the marginal likelihood ratio is analytically derived for several cases, based on the technique of the singular learning theory.

In Chapter 4, another main result of this thesis is presented (Kariya and Watanabe [2020b]). Here, a new hypothesis testing scheme, VB test is offered which is based on our analytical result. We, for the first time, apply the variational approximation to the test statistics and derive the asymptote of the test statistics.

In Chapter 5, some analytical results on type I and type II error of the hypothesis testing for singular models are presented. This is not yet published, but it is in preparation.

In Chapter 6, we summarize the result obtained, and present the future perspective.

Chapter 2

Bayesian hypothesis testing

2.1 Hypothesis testing

In this section, we summarize the formulation of hypothesis testing. As we mentioned in the introduction, hypothesis testing is a very useful tool to decide which hypothesis is valid for an explanation of the data we obtained.

The typical situation is that we try to decide the better one between two hypotheses for an explanation of our data. This is the simplest of cases, but it is also the most frequent and important case we face.

We assume that the data $\{X_1, X_2 \cdots X_n\}$ is generated from the distribution $q(x)$ independently, and we treat the case the hypotheses we compare are on the distribution generating our data. We refer to the two hypotheses we compare as the null hypothesis (N. H.) and the alternative hypothesis(A. H.).

$$\begin{aligned} \text{N.H.} &: X_i \sim P_0, \\ \text{A.H.} &: X_i \sim P_1, . \end{aligned}$$

The hypothesis test is constructed by using a function $T(\{X^n\})$ and a constant η as,

$$\begin{aligned} T(\{X^n\}) \leq \eta &\Rightarrow \text{we adapt N.H.} \\ T(\{X^n\}) > \eta &\Rightarrow \text{we adapt A.H.} \end{aligned}$$

In other words, "to construct a hypothesis test" is nothing other than "'to properly choose $T(\{X^n\})$ and η".

There exists the possibility we choose wrong hypothesis as as result of testing, and we should construct a "*good*" test for preventing this.

The probabilities of such errors, which are often called as the Type I error α and the Type II error β, are defined as,

$$\begin{aligned} \alpha(T,\eta) &= \Pr(A.H.|N.H.), \\ \beta(T,\eta) &= \Pr(N.H.|A.H.). \end{aligned}$$

As it can be clearly seen, it is difficult to keep both the Type I error α and the Type II error β low enough at the same time. In other words, these two errors have a trade-off relation. Therefore, the "good" hypothesis test is characterized as one that gives the lowest β in the set of the hypothesis tests that gives the same α.

The similar notion on the error of the test, *Level* and *Power* have been used in the literature, such as,

$$\begin{aligned} Level(T,\eta) &= \Pr(A.H.|N.H.), \\ Power(T,\eta) &= \Pr(A.H.|A.H.). \end{aligned}$$

These are very well-known, and we also use them in our thesis.

Now we can characterize the goodness of the hypothesis test. Let us consider two hypothesis test T_1 and T_2. We assume that both tests have the same type I error α. When a hypothesis test T_1 shows larger power than another test T_2, T_1 is said to be more powerful than T_2.

It is known that the hypothesis test using $L(X^n)$ as a test statistics becomes the most powerful test, from Neyman-Pearson's lemma .

Theorem 1. (Neyman-Pearson Lemma)

Let $X_1, \cdots, X_n$ be generated from i.i.d. $q(x)$. Consider the decision problem between two hypothesis,

$$\begin{aligned} \text{N.H.} &: X_i \sim P_0, \\ \text{A.H.} &: X_i \sim P_1, . \end{aligned}$$

For a constant $\eta \geq 0$, we define a set

$$A(\{X_n\}, \eta) = \left\{ X_n : \frac{P_0(X_1, \cdots, X_n)}{P_1(X_1, \cdots, X_n)} > \eta \right\} \tag{2.1}$$

Let

$$\alpha^* = P_0(A^c(\{X_n\}, \eta)), \beta^* = P_1(A(\{X_n\}, \eta)) \tag{2.2}$$

be the error which corresponds to the decision region $A(\{X_n\}, \eta)$.

Let $B(\{X_n\}, \eta)$ be any other decision region of which error is α and β.

Then, if $\alpha \leq \alpha^*$, $\beta \geq \beta^*$ is satisfied.

proof. Let $B \in X^n$ be any other acceptance regions. Let us define ϕ_a and ϕ_b as indicator function of the region A and B.

For all $X_1, \cdots, X_n \in X^n$, the following is satisfied.

$$(\phi_a(\{X_n\}) - \phi_b(\{X_n\}))\,(P_0(\{X_n\}) - \eta P_1(\{X_n\})) \geq 0 \tag{2.3}$$

Then, by summing this over the all space, we obtain

$$\sum (\phi_a P_0 - \eta \phi_a P_1 - P_0 \phi_b + \eta P_1 \phi_b) \geq 0. \tag{2.4}$$

The left hand side becomes,

$$\begin{aligned}(l.h.s) &= \sum_A (P_0 - \eta P_1) - \sum_B (P_0 - \eta P_1) \\ &= (1 - \alpha^*) - \eta\beta^* - (1 - \alpha) + \eta\beta \\ &= \eta(\beta - \beta^*) - (\alpha^* - \alpha).\end{aligned}$$

As $\eta > 0$, the theorem is proved. □

We can readily conclude from Neyman-Pearson Lemma that the hypothesis test using the likelihood ratio as the test statistics becomes the most powerful test. Therefore, in this thesis, we mainly study the likelihood ratio test theoretically.

2.2 The framework of the Bayesian hypothesis testing

In this section, we briefly review the framework of a Bayesian hypothesis test.

Let $\{X^n = (X_1, X_2, ..., X_n) \in \mathbb{R}^1\}$ be a set of sample obtained. We assumed that these samples are generated independently and identically from a probabilistic model $p_0(x|w)$. where w means the set of the parameters in the model.

In the Bayesian framework, parameters w_0 is assumed to be generated from a prior $\varphi(w)$, which is described as

$$w_0 \sim \varphi(w), \quad X_i \sim p_0(x|w_0).$$

In the Bayseian hypothesis test, the null and alternative hypotheses we treat are set in terms of prior, such as

$$\begin{aligned}\text{N.H.} &: w_0 \sim \varphi_0(w), \quad X_i \sim p_0(x|w_0), \\ \text{A.H.} &: w_0 \sim \varphi_1(w), \quad X_i \sim p_0(x|w_0).\end{aligned}$$

As well as we saw in the previous section, we can compare the power of the two tests. Let us consider two hypothesis tests T and U. We will say that "T is a more powerful test than U" if and only if,

$$Level(T, \eta) = Level(U, \rho) \Rightarrow Power(T, \eta) \geq Power(U, \rho),$$

holds for an arbitrary set $(\eta, \rho) \in \mathbb{R}^2$.

A hypothesis test T is said to be "the most powerful test", if it is more powerful than any other test. The marginal likelihood ratio $L(X^n)$, which is the test statistics of the most powerful test, can be written as,

$$L(X^n) = \frac{\int \varphi_1(w) \prod_i p_0(X_i|w) dw}{\int \varphi_0(w) \prod_i p_0(X_i|w) dw}. \tag{2.5}$$

In the last of the this section, we should note that the Bayesian hypothesis test we consider in this thesis is different from the decision theory based on the Bayes Factor, value of the marginal likelihood ratio, not on the stochastic behavior of the marginal likelihood ratio.

The decision theory based on the value of Bayes factor is simple and clear, but we should keep in mind that the marginal likelihood ratio is a stochastic variable. The value of the marginal likelihood ratio can be variate greatly as a result of the stochastic behavior of the sample, and it affects the result of the decision made by our treatment. Such an example will be shown in the next chapter. Therefore, the stochastic behavior of the marginal likelihood ratio should be grasped for appropriate decision making.

2.3 Relation between the hypothesis testing and singular learning theory

In the previous subsection, we saw that the most powerful test is given as the well-known likelihood ratio test.

Therefore, it is important to know the distribution of the likelihood ratio, the test statistics which gives the most powerful test.

Let us see the simplest case (竹村彰通 [2020]). We consider the two hypotheses such as,

$$\begin{aligned} \text{N.H.} \quad &: \quad w \sim \delta(w - w_0), \\ \text{A.H.} \quad &: \quad w \sim \delta(w - w^*), w_0 \neq w^*, \end{aligned}$$

where w_0 is a fixed value and w^* means a value that is different from w_0 in the model $p(x|w)$.

The marginal likelihood ratio L becomes

$$L(X^n) = \frac{\prod_i p_0(X_i|w^*)dw}{\prod_i p_0(X_i|w_0)dw}. \tag{2.6}$$

The logarithm of $L(X^n)$ can be written as

$$l_n(w^*) \equiv \log L(X^n) = \sum_i \log(p(x|w^*)) - \log(p(x|w_0)) \tag{2.7}$$

Let us consider the case that N. H. is true. That is, the sample $\{X_n\}$ is generated from N.H..

The Taylor expansion of $l_n(w^*)$ around w_0 becomes,

$$\begin{aligned} l_n(w_0+\epsilon) &= l_n(w_0) + \sum_{j=1}^{d} \frac{\partial l_n(w_0)}{\partial w_j^*}\epsilon_j + \frac{1}{2}\sum_{j,k}^{d} \frac{\partial^2 l_n(w_0)}{\partial w_j^* \partial w_k^*}\epsilon_j\epsilon_k \\ &+ \mathcal{O}(\epsilon^3) \end{aligned}$$

where d means the dimension of the parameter space.

When we substitute $\frac{\gamma}{\sqrt{n}}$ for ϵ, the l_n becomes

$$\begin{aligned} l_n(w_0+\frac{\gamma}{\sqrt{n}}) &= l_n(w_0) + \frac{1}{\sqrt{n}}\sum_{j=1}^{d} \frac{\partial l_n(w_0)}{\partial w_j^*}\gamma_j + \frac{1}{2n}\sum_{j,k}^{d} \frac{\partial^2 l_n(w_0)}{\partial w_j^* \partial w_k^*}\gamma_j\gamma_k \\ &+ o(1) \end{aligned}$$

It is easily seen that

$$\int dx p(x|w)\frac{\partial \log p(x|w^*)}{\partial w_i^*} = \frac{\partial}{\partial w_i^*}\int dx p(x|w^*) = 0. \tag{2.8}$$

Also, the covariance matrix of $\frac{\partial \log p(x|w^*)}{\partial w_i^*}$ is nothing other than the Fisher information matrix of $p(x|w^*)$.

From the central limit theorem, there exists a stochastic variable $\{W_k\}$ that satisfies

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log p(X_i|w_0}{\partial w_j^*} &\rightarrow W_j \\ \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log p(X_i|w_0)}{\partial w_j^* \partial w_k^*} &\rightarrow -I_{jk}(w_0)\end{aligned}$$

Therefore, we obtain

$$l_n(w_0+\frac{\gamma}{\sqrt{n}}) = \sum_{j=1}^{d} W_j\gamma_j - \frac{1}{2}\sum_{j,k}^{d} I_{jk}\gamma_j\gamma_k + o(1) \tag{2.9}$$

Let us assume that all the eigenvalues of $I_{jk}(w_0)$ are positive. In that case, the $I_{jk}(w_0)$ can be expressed as,

$$I_{jk}(w_0) = \Sigma^2 \tag{2.10}$$

where Σ is a real-value symmetry matrix whose eigenvalue is positive.

By using this,

$$\begin{aligned}\sum_{j=1}^{d} W_j\gamma_j - \frac{1}{2}\sum_{j,k}^{d} I_{jk}\gamma_j\gamma_k &= Z^t\sum\gamma - \frac{1}{2}\gamma^t\Sigma^t\Sigma\gamma \\ &= Z^t\xi - \frac{1}{2}\xi^t\xi \\ &= -\frac{1}{2}\sum_{j=1}^{d}(\xi_j - Z_j)^2 + \frac{1}{2}\sum_{j=1}^{d} Z_i^2\end{aligned}$$

where Z is a vector whose components $\{Z_i\}$ satisfy $\{Z_i\} \sim \mathcal{N}(0,1)$.

Clearly, $l_n(w_0)$ is maximized at $\gamma = \Sigma^{-1}Z \sim \mathcal{N}(0, I_{jk}(w_0)^{-1})$. That is, the solution w^* maximizing the $l_n(w_0)$ is

$$w^* = \frac{1}{\sqrt{n}}\Sigma^{-1}Z \sim \mathcal{N}(0, \frac{1}{n}I_{jk}(w_0)^{-1})) \tag{2.11}$$

This is nothing other than the solution of the maximum likelihood estimation.

As a result of this, the following holds under $n \rightarrow \infty$.

$$2\max l_n(w^*) \sim \chi_d^2 \tag{2.12}$$

This shows that the maximum of the log likelihood ratio obeys the χ_d^2. It is the special version of the Wilks' Theorem.

$$2\log\frac{\max_{w\in H_1} L_n(w)}{\max_{w\in H_0} L_n(w)} \sim \chi^2(r_1 - r_0) \tag{2.13}$$

where r_1 means the dimension of the parameter w in H_1 and r_0 means the dimension of the parameter w in H_0.

This is a very general and useful result, but the assumption we use to prove this result, the Fisher information matrix is regular does not always holds. This breaks down especially for the more complex probabilistic models such as mixture models, neural networks, and so on. These models are known as *singular* models, which have the singularity where the Fisher information matrix becomes singular, in their parameter space.

These model becomes popular and popular recently, but actually, the theory on the hypothesis test for using these *singular* models are not developed. This is one of the motivations for the present study.

2.4 Previous studies on the Bayesian hypothesis testing from a singular learning theory perspective

In the previous subsection, we pointed out the importance of the study on the hypothesis test using singular models. As we mentioned in the introduction, there have been many studies on this topic, but they are mainly on the asymptotic distribution of the maximum likelihood.

We also mentioned in the Introduction that as for the statistical *inference* , there have been several study that shows the Bayesian singular learning theory.

The asymptote of the important quantities such as the free energy, the generalization error are obtained and the generalized information criteria WAIC and WBIC are derived (Watanabe [2018]).

However, few theoretical studies on *hypothesis testing* based on the Bayesian singular learning theory exists, except for on the problem of the change point detection using the marginal likelihood ratio as a test statistics (藤原香織 and 渡辺澄夫 [2008], 大橋耕也 and 渡辺澄夫 [2017]).

They consider the hypothesis testing such that N.H. that corresponds to the case that there is no change, and A.H. that corresponds to the case that some changes occur. When the N.H. corresponds to the singularity of the probabilistic model they use, the behavior of the marginal likelihood ratio becomes different from those well known χ^2. They examine the validity of their new methods numerically by using the artificial data and concluded that their new methods is valid.

While the effectiveness of the Bayesian singular learning theory is becoming

popular, there is no case that the Bayesian singular learning theory is applied to the problem of testing homogeneity, as far as the author knows, though the importance of the problem. This is one of the motivations for us to study this problem.

Chapter 3

Asymptotic analysis of the marginal likelihood ratio for testing homogeneity in normal mixtures

3.1 Introduction

In this chapter, we study the test of homogeneity of normal mixture based on the Bayesian framework for the first time.

To construct hypothesis test, we derive the asymptotic distribution of the test statistic, i.e., the marginal likelihood ratio. We consider here in three cases: (1) only the mixture ratio is a variable; (2) the mixture ratio and the mean of the mixed distribution in the A.H. are variables; (3) the mixture ratio, the mean, and the variance of the mixed distribution in the A. H. are variables.

In all cases, the marginal likelihood ratios does not converge to the well-known χ^2 distribution, but to more complex distributions. This results from an effect of the singularities in the model. These results suggest the need for the theoretical study for the hypothesis test for singular models. The present study can be considered to pave the way in this direction.

3.2 Asymptotic distribution of the marginal likelihood ratio

In the following, we derive the asymptotic probability distributions of $L(X^n)$. We treat three cases for A.H. such as,

1. $\varphi_1(a, b, c) = U_a(0, 1)\delta(b - \beta)\delta(c - 1)$,

2. $\varphi_1(a, b, c) = U_a(0, 1)U_b(0, B)\delta(c - 1)$,

3. $\varphi_1(a, b, c) = U_a(0, 1)U_{bc}(D)$,

where $U_a(0, 1)$ is the uniform distribution of a on the interval $(0, 1)$, $U_b(0, B)$ is the uniform distribution of b on $(0, B)$, and $U_{bc}(D)$ is the uniform distribution of a set D in $(b, 1/c)$ space.

Through the proofs of the theorems we show, the following notation is used. For a given sample X^n, two random variables ξ_n and η_n are defined as

$$\xi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i, \tag{3.1}$$

$$\eta_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} (X_i^2 - 1). \tag{3.2}$$

If $\{X^n\}$ is a sample generated from some i.i.d. given by the N.H., then both ξ_n and η_n converge to $\mathcal{N}(0, 1)$ in distribution and they are asymptotically independent.

3.2.1 Case 1 : the case only the mixture ratio is unknown

Here we consider the case 1, that is, only the mixture ratio is the variable.

This case can be regarded as a minimal model with which to study the effect of the singularity on the behavior of the marginal likelihood ratio. Hence we will study it as a first step towards more practical situations in the following sections.

Especially, We treat the case that the A.H. is near the N.H. in terms of the Kullback-Leibler divergence. In such a situation, it is not easy to discriminate the alternative hypothesis from the null one. This is a typical situation in which a hypothesis test is needed.

A similar situation is studied in the context of the Bayesian *inference*, where the true distribution which generates the sample is slightly deviates from the singularity of the model on the order of $O(n^{-1/2})$ (Watanabe and Amari [2003]). It was shown that the singularity greatly affects the behavior of the generalization error, even when the parameter set that represents the true model does not definitely match the singularity.

Although our problem is not an inference but a hypothesis test, we expected that a similar structure exists. We will see that this is true, and that the $n^{-1/2}$ scaling works as well. This is because the scaling is determined from the order of the Kullback-Leibler divergence between the A.H. and the singularity (N.H.).

Applying the scaling mentioned above, we can derive the asymptotic distribution of the marginal likelihood ratio as follows.

Theorem 2. Assume that the N.H. and A.H. are given as

$$\begin{aligned} \text{N.H.} \quad &: \quad \varphi_0(w) = \delta(a)\delta(b)\delta(c-1), \\ \text{A.H.} \quad &: \quad \varphi_1(w) = U_a(0,1)\delta(b-\beta)\delta(c-1), \end{aligned}$$

where $\beta = \beta_0 \times n^{-\frac{1}{2}}$ and β_0 is a nonzero constant. If $\{X_i\}$ is independently and identically generated from the N.H., the convergence in probability,

$$L(X^n) - L_\infty(\xi_n) \to 0$$

holds for $n \to \infty$, where

$$L_\infty(\xi_n) \quad = \quad \frac{\sqrt{2\pi}}{2\beta_0}\left[\mathrm{erf}\left(\frac{\beta_0 - \xi_n}{\sqrt{2}}\right) + \mathrm{erf}\left(\frac{\xi_n}{\sqrt{2}}\right)\right]\exp(\frac{\xi_n^2}{2}). \tag{3.3}$$

Here ξ_n is a random variable defined in eq.(3.1) and $\mathrm{erf}(x)$ is the error function,

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt.$$

Remark. Assume that ξ is a random variable whose probability distribution is $\mathcal{N}(0,1)$. By Theorem 2 and the convegence in distribution $\xi_n \to \xi$, the convergence in distribution $L(X^n) \to L_\infty(\xi)$ holds. Since $L_\infty(\xi)$ can be rewriten as

$$L_\infty(\xi) = \int_0^1 da \ \exp\left(-\frac{\beta_0^2 a^2}{2} + \beta_0 \xi a\right),$$

it is an increasing function of ξ. Therefore, we can determine the rejection region by using ξ.

proof. : The integral with respect to (b,c) is easily performed and the prior of a in the A.H. is a uniform distribution on $(0,1)$; it follows that

$$L(X^n) \quad = \quad \int_0^1 \exp(H(a))\ da, \tag{3.4}$$

where $H(a)$ is defined as,

$$H(a) \quad \equiv \quad \sum_{i=1}^n \log\frac{p(X_i|a,\beta,1)}{p(X_i|0,0,1)} \tag{3.5}$$

$$= \quad \sum_{i=1}^n \log\left\{(1-a) + a\exp\left(\beta X_i - \frac{\beta^2}{2}\right)\right\}. \tag{3.6}$$

Under the N.H., from a well-known result in extreme statistics, the order of the maximum of X_i is

$$X_M \equiv \max\{X_i\} = O_p\left(\sqrt{2\log n}\right).$$

This results in

$$\beta X_i - \frac{\beta^2}{2} \leq \beta X_M - \frac{\beta^2}{2} = O_p\left(\sqrt{\frac{\log n}{n}}\right). \tag{3.7}$$

Let α be a constant which satisfies $1 < \alpha$. Then

$$\beta X_i - \frac{\beta^2}{2} \sim o_p\left(\sqrt{\frac{(\log n)^\alpha}{n}}\right).$$

Hence

$$\begin{aligned}\exp\left(\beta X_i - \frac{\beta^2}{2}\right) &= 1 + \left(\beta X_i - \frac{\beta^2}{2}\right) + \frac{1}{2!}\left(\beta X_i - \frac{\beta^2}{2}\right)^2 \\ &\quad + \frac{1}{3!}\left(\beta X_i - \frac{\beta^2}{2}\right)^3 \times e^{C_0},\end{aligned}$$

where C_0 is a random variable that satisfies

$$|C_0| \leq \left|\beta X_i - \frac{\beta^2}{2}\right|.$$

Then,

$$\begin{cases} 0 \leq C_0 \leq \beta X_i - \frac{\beta^2}{2} & (\text{ if } \beta X_i - \frac{\beta^2}{2} \geq 0), \\ \beta X_i - \frac{\beta^2}{2} \leq C_0 \leq 0 & (\text{otherwise}). \end{cases}$$

Therefore,

$$\frac{1}{3!}\left(\beta X_i - \frac{\beta^2}{2}\right)^3 e^{C_0} \sim o_p\left(\frac{(\log n)^{3\alpha/2}}{n^{\frac{3}{2}}}\right).$$

It follows that

$$\begin{aligned} H(a) &= \sum_{i=1}^{n} \log\left[1 + a\left\{\left(\beta X_i - \frac{\beta^2}{2}\right) + \frac{1}{2}\left(\beta X_i - \frac{\beta^2}{2}\right)^2\right\} + o_p\left(\frac{1}{n}\right)\right] \\ &= \sum_{i=1}^{n} \log\left[1 + a\beta X_i - \frac{a\beta^2}{2} + \frac{a\beta^2 X_i^2}{2} + o_p\left(\frac{1}{n}\right)\right]. \end{aligned}$$

Then, by applying a Taylor expansion $\log(1+\epsilon) = \epsilon - \epsilon^2/2 + O(\epsilon^3)$ to this equation, we obtain

$$H(a) = \sum_{i=1}^{n} \left[a\beta X_i - \frac{1}{2}a\beta^2 + \frac{1}{2}a\beta^2 X_i^2 - \frac{1}{2}a^2\beta^2 X_i^2 \right] + o_p(1). \tag{3.8}$$

Let us use the following notations,

$$\begin{aligned} \gamma &\equiv \frac{\sum_i \beta X_i + \frac{1}{2}\sum_i (\beta X_i)^2 - \frac{1}{2}\beta^2}{\frac{1}{2}\sum_i (\beta X_i)^2}, \\ \delta &\equiv \frac{1}{2}\sum_i (\beta X_i)^2 . \end{aligned}$$

Accordingly, $H(a)$ can be written as

$$\begin{aligned} H(a) &= -\delta a^2 + \gamma\delta a \\ &= -\delta(a-\gamma/2)^2 + \delta\gamma^2/4. \end{aligned}$$

It follows that

$$\begin{aligned} L(X^n) &= \int_0^1 da \exp\left[-\delta(a-\frac{1}{2}\gamma)^2\right] \times \exp\left[\frac{1}{4}\times\gamma^2\delta\right] \\ &= \frac{\sqrt{\pi}}{2\sqrt{\delta}}\left[\mathrm{erf}\left(\frac{\gamma\sqrt{\delta}}{2}\right) + \mathrm{erf}\left(\sqrt{\delta}(1-\frac{\gamma}{2})\right)\right] \times \exp\left[\frac{1}{4}\times\gamma^2\delta\right], \end{aligned}$$

where $\mathrm{erf(x)}$ is the error function defined by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt.$$

As n goes to infinity, δ converges in probability as

$$\delta \to \frac{1}{2}\beta_0^2, \tag{3.9}$$

by which γ satisfies

$$\gamma = 2\xi_n + o_p(1), \tag{3.10}$$

which completes the theorem. □

To confirm the validity of the result obtained above, we numerically evaluated the value of $L(X^n)$ for finite samples. We used sets of samples generated from the N.H. distribution. Here, we set the sample size as $n = 50$ or $n = 100$, and

generated 10000 sets of samples from the null hypothesis. In each case, the level was calculated from them as the ratio of the number of sets X^n for which $L(X^n)$ falls into the critical region to the total number of sets.

The result is shown in the Table 1. We can see that the levels numerically calculated match those derived from the asymptote in both $n = 50$ and $n = 100$ cases. Therefore, it can be concluded that the asymptotic distribution we derived above is valid.

Table 3.1: Level calculated from the samples generated from the null hypothesis

parameters		level	
β	threshold of 5%	n = 50	n = 100
0.5	1.464	5.07%	4.76%
1	2.02	5.56%	5.15%
1.5	2.475	5.13%	4.92%
2	2.929	4.88%	5.03%

3.2.2 Case 2 : both the mixture ratio and the mean of the mixed distribution are unknown

In this section, we proceed towards a bit more general case. Here, the alternative hypothesis is a normal mixture whose mixture ratio and the mean are both variables. We assume that we have no prior knowledge about the A.H., and that the prior is a uniform distribution.

We will prove the following theorem on the asymptotic distribution of the marginal likelihood ratio L.

Theorem 3. Assume that N.H. and A.H. are given by

$$\begin{aligned} \text{N.H.} &: \varphi_0(w) = \delta(a)\delta(b)\delta(c-1), \\ \text{A.H.} &: \varphi_1(w) = U_a(0,1)U_b(0,B)\delta(c-1), \end{aligned}$$

respectively, where $B = B_0 \times n^{-\frac{1}{2}}$. Then, if $\{X^n\}$ is an i.i.d. sample generated from N.H., the convergence in probability

$$L(X^n) - L_\infty(\xi_n) \to 0$$

holds as $n \to \infty$, where

$$L_\infty(\xi_n) = \frac{1}{B_0}\int_0^{B_0^2} \frac{1}{2\sqrt{t}} \log\left(\frac{B_0^2}{t}\right) e^{-t/2} \cosh\left(\xi_n\sqrt{t}\right) dt, \qquad (3.11)$$

and ξ_n is a random variable defined in eq.(3.1).

Remark. Assume that ξ is a random variable whose probability distribution is $\mathcal{N}(0,1)$. By Theorem 3, the convergence in distribution

$$L(X^n) \to L_\infty(\xi)$$

holds. Hence the asymptotic rejection region of the most powerful test can be found by using $L_\infty(\xi)$.

proof. : The marginal likelihood ratio can be written as

$$L(X^n) = \int_0^1 da \int_0^B \frac{db}{B} \exp(H(a,b)),$$

where

$$H(a,b) = \sum_i \log\left\{(1-a) + a\exp\left(bX_i - \frac{1}{2}b^2\right)\right\}.$$

From the condition $b \in [0, B_0/\sqrt{n}]$, the $H(a,b)$ can be approximated in the same way as in the proof of Theorem 1 as,

$$H(a,b) = \sum_i \left[abX_i - \frac{1}{2}ab^2 + \frac{1}{2}ab^2X_i^2 - \frac{1}{2}a^2b^2X_i^2\right] + o_p(1). \tag{3.12}$$

Hence,

$$H(a,b) = -\frac{n}{2}a^2b^2 + \sum_i \left[abX_i + \frac{1}{2}ab^2\left(X_i^2-1\right) - \frac{1}{2}a^2b^2\left(X_i^2-1\right)\right] + o_p(1). \tag{3.13}$$

Under the N.H., by using the definitions, eqs.(3.1) and (3.2), we have

$$\begin{aligned} H(a,b) &= -\frac{n}{2}a^2b^2 + \sqrt{n}\left(ab\xi_n + \frac{1}{2}ab^2\eta_n - \frac{1}{2}a^2b^2\eta_n\right) + o_p(1) \\ &= -\frac{n}{2}a^2b^2 + \sqrt{n}ab\xi_n + o_p(1). \end{aligned}$$

Using the notation $L = L(X^n)$ for simplicity, it follows that

$$\begin{aligned} L &= \int_0^1 da \int_0^{B_0/\sqrt{n}} \frac{\sqrt{n}\,db}{B_0} \exp(-\frac{n}{2}a^2b^2 + \sqrt{n}ab\xi_n + o_p(1)) \\ &= \int_0^1 da \int_0^{B_0} \frac{db}{B_0} \exp(-\frac{1}{2}a^2b^2 + ab\xi_n + o_p(1)). \end{aligned}$$

Hence, the convergence in probability $L(X^n) - L_\infty(\xi_n) \to 0$ holds, where

$$L_\infty(\xi_n) = \int_0^1 da \int_0^{B_0} \frac{db}{B_0} \exp(-\frac{1}{2}a^2b^2 + ab\xi_n)$$

By using $b = t/a$, we have

$$\begin{aligned} L_\infty(\xi_n) &= \int_0^1 da \int_0^{aB_0} \frac{dt}{aB_0} \exp(-\frac{1}{2}t^2 + t\xi_n) \\ &= \int_0^{B_0} \frac{dt}{B_0} \int_{t/B_0}^1 \frac{da}{a} \exp(-\frac{1}{2}t^2 + t\xi) \\ &= \frac{1}{B_0} \int_0^{B_0} dt \, (\log(B_0) - \log t) \exp(-\frac{1}{2}t^2 + t\xi). \\ &= \frac{1}{2B_0} \int_0^{B_0} dt \, (\log(B_0) - \log t) \exp(-\frac{1}{2}t^2) \cosh(t\xi). \end{aligned}$$

Then eq.(3.11) is obtained by replacing the integration of t by $\sqrt{t}$. □

Similar to the previous example, we can see that the asymptotic behavior of the test statistics L does not explicitly depend on n.

We also note that the stochastic behavior of L is determined only by that of the random variable ξ. Clearly, L increases monotonously as the absolute value of ξ increases, and $\cosh\left(\xi\sqrt{t}\right)$ is an even function with respect to ξ, hence, we can determine the critical region in the same way as is done in a two-sided hypothesis test of ξ.

We numerically validated the effectiveness of the analytically derived distribution of L when the sample size is finite.

First, we prepared the 10000 sets of the n samples, where n means the sample size and we set n as $n = 50$ or $n = 100$. We calculated the L by substituting the ξ in the asymptote with $\frac{1}{\sqrt{n}}X_i$. Here, we fixed B_0 as 1. In each case, the level was calculated from them as the ratio of the number of sets X^n for which $L(X^n)$ falls into the critical region to the total number of sets. The levels were compared with those calculated from the level calculated from the asymptote of L.

Table 2 shows the result. It shows the asymptote we derived in the previous section works well even in the finite n cases.

Let us comment on the comparison of our results with those obtained by another well-known method of Bayesian hypothesis testing, i.e.,using the Bayes factor.

The log marginal likelihood ratio $F = -\log L$, which is also called the logarithm of the Bayes factor, can be used as a tools for hypothesis testing.

Table 3.2: The level calculated from the samples generated from the null hypothesis

level	10%	5%	1%
rejection region $L > r$	r=2.171	r=2.298	r=2.646
numerically calculated level(n=50)	9.75%	5.22%	1.04%
numerically calculated level(n=100)	9.91%	4.73%	0.97%
numerically calculated level(n=200)	9.74%	4.84%	0.99%

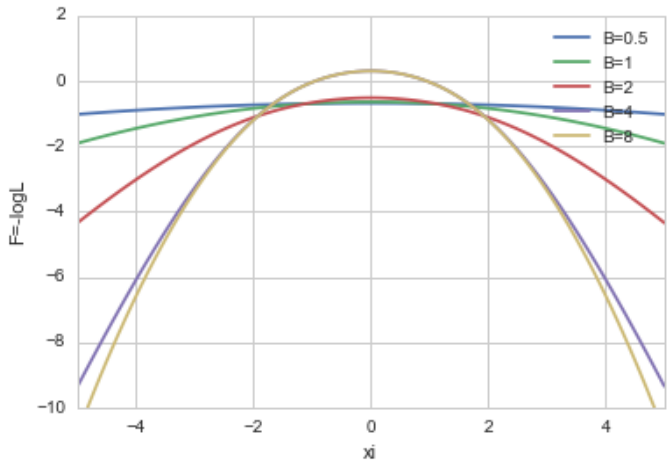

Fig. 3.1: The log marginal likelihood ratio F as a function of the random variable ξ for several values of B_0.

When the value of F calculated from the data becomes negative, we choose the alternative hypothesis, and if otherwise, we choose the null hypothesis otherwise.

For the present problem, we can consider two ways of hypothesis testing with the result we derived. One is based on the stochastic behavior of L, and the other is based on the F. Both use the same quantity L, but we will see below that the former may work more effective in the "delicate" situation.

Figure 3.1 shows the behavior of F as a function of ξ.

Interestingly, when B_0 is small, the value of F is always negative, regardless of any ξ, while in the large B_0 case, F becomes positive in the small ξ region. This can be understood as follows. When the two centers of the mixture distribution are so close as the distance between them is $\mathcal{O}(n^{-1/2})$, the overlap of the distribution of the null hypothesis and the distribution of the alternative hypothesis is large, and the sign of Bayes factor can become negative for any ξ.

In other words, when the two hypotheses are difficult to distinguish, the hypothesis test using the Bayes factor may choose the alternative hypothesis for any data, and it does not work well. On the other hand, the likelihood ratio test based on the stochastic behavior of L is expected to work in such delicate cases.

3.2.3 Case 3 : the case the mixture ratio, the mean of the distribution mixed, and the variance are unknown

Here, we discuss a more practical case in which the variance of the A.H. is also a variable. That is, we consider the following probabilistic model,

$$p(x|a,b,c) = (1-a)\,\mathcal{N}(0,1) + a\mathcal{N}(b,\frac{1}{c}). \tag{3.14}$$

We set the N.H. and the A.H. as

$$\begin{aligned} \text{N.H.} \quad &: \quad \varphi_0(a,b,c) = \delta(a)\delta(b)\delta(c-1), \\ \text{A.H.} \quad &: \quad \varphi_1(a,b,c) = U_a(0,1)U(b,c), \end{aligned}$$

where $U_a(0,1)$ is a uniform distribution on the interval $(0,1)$, and $U(b,c)$ is a uniform distribution on an ellipsoid in the (b,c) plane such as,

$$D = \left\{ (b,c)\ ;\ b^2 + \frac{(c-1)^2}{2} \leq \frac{R_0^2}{n} \right\},$$

where R_0 is a constant. The area of D is $\sqrt{2}\pi R_0^2/n$.

Theorem 4. When the sample size $n \to \infty$, convergence in probability $L(X^n) - L_\infty(\Xi_n) \to 0$ holds, where

$$L_\infty(\Xi_n) = \frac{1}{2R_0^2} \int_0^{R_0^2} \left(\frac{R_0}{\sqrt{t}} - 1 \right) \exp\left(-\frac{t}{2} \right) I_0(\sqrt{t}\Xi_n)\ dt.$$

Here, $I_0(t)$ is the modified Bessel function,

$$I_0(t) = \frac{1}{\pi} \int_0^\pi \cosh(t\sin\theta) d\theta,$$

which is monotone increasing in $t > 0$ and Ξ_n is a random variable defined by

$$\Xi_n = \sqrt{\xi_n^2 + \eta_n^2},$$

where ξ_n and η_n are defined in eq.(3.1) and (3.2), respectively.

Remark. Let Ξ be a random variable whose square is subject to a χ^2 distribution with freedom 2. In accordance with this theorem, $L(X^n)$ converges in distribution to $L_\infty(\Xi)$. Hence, the rejection region of the most poweful test can be asymptotically determined by $L_\infty(\Xi)$.

proof. The log density ratio function is given by

$$\begin{aligned} f(X_i, a, b, c) &= \log \frac{p(X_i|a, b, c)}{p(X_i|0, 0, 1)} \\ &= \log \left[1 - a + a\sqrt{c}\, e^{g(X_i,b,c)}\right], \end{aligned}$$

where $g(x, b, c)$ is a function defined by

$$g(X_i, b, c) = -\left(\frac{c}{2} - \frac{1}{2}\right) X_i^2 + bcX_i - \frac{b^2 c}{2}.$$

Hence the marginal likelihood ratio $L = L(X^n)$ is given by

$$L = \int_0^1 da \int_D dbdc\, \varphi_1(a, b, c)\, \exp\left(\sum_{i=1}^n f(X_i, a, b, c)\right),$$

where

$$\varphi_1(a, b, c) = \frac{n}{\sqrt{2}\pi R_0^2}.$$

Since the integrated region of (b, c) of this integral is D, $b = O_p(n^{-1/2})$, $c = O_p(n^{-1/2})$. It follows that

$$\begin{aligned} f(X_i, a, b, c) &= f|_{(b,c)=(0,1)} + \frac{\partial f}{\partial b}\Big|_{(b,c)=(0,1)} b + \frac{\partial f}{\partial c}\Big|_{(b,c)=(0,1)} (c-1) \\ &+ \frac{1}{2}\frac{\partial^2 f}{\partial b^2}\Big|_{(b,c)=(0,1)} b^2 + \frac{1}{2}\frac{\partial^2 f}{\partial c^2}\Big|_{(b,c)=(0,1)} (c-1)^2 \\ &+ \frac{\partial^2 f}{\partial b \partial c}\Big|_{(b,c)=(0,1)} b\,(c-1) + o_p(1/n), \end{aligned}$$

where

$$\begin{aligned} \frac{\partial f}{\partial b}\Big|_{(b,c)=(0,1)} &= aX_i \\ \frac{\partial f}{\partial c}\Big|_{(b,c)=(0,1)} &= \frac{a}{2}\left(X_i^2 - 1\right) \\ \frac{\partial^2 f}{\partial b^2}\Big|_{(b,c)=(0,1)} &= a\left(X_i^2 - 1\right) - a^2 X_i^2 \\ \frac{\partial^2 f}{\partial b \partial c}\Big|_{(b,c)=(0,1)} &= aX_i + \frac{a}{2}(1-a) X_i \left(1 - X_i^2\right) \\ \frac{\partial^2 f}{\partial c^2}\Big|_{(b,c)=(0,1)} &= -\frac{a}{4}\left(1 + X_i^2\right) - \frac{a}{4}\left(X_i^2 - X_i^4\right) - \frac{a^2}{4}\left(1 - X_i^2\right)^2. \end{aligned}$$

Hence

$$\begin{aligned} f(X_i,a,b,c) &= abX_i + \frac{a(c-1)}{2}\left(X_i^2-1\right) + \frac{1}{2}\left\{a\left(X_i^2-1\right) - a^2X_i^2\right\}b^2 \\ &\quad + \left\{aX_i + \frac{a}{2}(1-a)X_i\left(1-X_i^2\right)\right\}b\,(c-1) \\ &\quad + \frac{1}{2}\left[-\frac{a}{4}\left(1+X_i^2\right) - \frac{a}{4}\left(X_i^2-X_i^4\right) - \frac{a^2}{4}\left(1-X_i^2\right)^2\right](c-1)^2 \\ &\quad + o_p(1/n). \end{aligned} \tag{3.15}$$

Note that the order of the quadratic forms of $(b, c-1)$ is $1/n$ and

$$\begin{aligned} (1/n)\sum_i X_i &= o_p(1), \\ (1/n)\sum_i X_i^2 &= 1 + o_p(1), \\ (1/n)\sum_i X_i^3 &= o_p(1), \\ (1/n)\sum_i X_i^4 &= 3 + o_p(1). \end{aligned}$$

The log likelihood ratio function is given by

$$\begin{aligned} \sum_{i=1}^n f(X_i,a,b,c) &= ab\sum_{i=1}^n X_i + \frac{a(c-1)}{2}\sum_{i=1}^n\left(X_i^2-1\right) - \frac{n}{2}a^2b^2 \\ &\quad -\frac{n}{4}a^2(c-1)^2 + o_p(1). \end{aligned}$$

Let us define (r, θ) by

$$\begin{aligned} b &= r\cos\theta, \\ c &= 1 + \sqrt{2}r\sin\theta. \end{aligned}$$

Then by using eq.(3.1) and (3.2),

$$\begin{aligned} \sum_{i=1}^n f(X_i,a,b,c) &= -\frac{n}{2}a^2r^2 + \sqrt{n}ar(\xi_n\cos\theta + \eta_n\sin\theta) + o_p(1) \\ &= -\frac{n}{2}a^2r^2 + \sqrt{n}ar\sqrt{\xi_n^2+\eta_n^2}\sin(\theta+\theta_0) + o_p(1), \end{aligned}$$

where θ_0 is a random variable which satisfies $\tan\theta_0 = \xi_n/\eta_n$. By using the nota-

tion $R = R_0/\sqrt{n}$, the log marginal likelihood ratio can be written as

$$\begin{aligned} L &= \int_0^1 da \int_D dbdc \, \frac{n}{\sqrt{2}\pi R_0^2} \exp\left(\sum_i f_i(a,b,c) \right) \\ &= \int_0^1 da \int_0^R \frac{2r\, dr}{R^2} \int_0^{2\pi} \frac{d\theta}{2\pi} \exp\left(-\frac{n}{2}a^2 r^2 + ar\sqrt{\xi_n^2 + \eta_n^2}\sin(\theta + \theta_0) + o_p(1) \right) \\ &= \int_0^1 da \int_0^R \frac{2r\, dr}{R^2} \int_0^{2\pi} \frac{d\theta}{2\pi} \exp\left(-\frac{n}{2}a^2 r^2 + ar\Xi_n \sin(\theta) + o_p(1) \right). \end{aligned}$$

Then by replacing $r = \ell/\sqrt{n}$ with $dr = d\ell/\sqrt{n}$, it follows that

$$L = \int_0^1 da \int_0^{R_0} \frac{2\ell\, d\ell}{R_0^2} \int_0^{2\pi} \frac{d\theta}{2\pi} \exp\left(-\frac{1}{2}a^2\ell^2 + a\ell\Xi_n \sin(\theta) + o_p(1) \right).$$

We define $L_\infty(\Xi_n)$ by

$$L_\infty(\Xi_n) = \int_0^1 da \int_0^{R_0} \frac{2\ell\, d\ell}{R_0^2} \int_0^{2\pi} \frac{d\theta}{2\pi} \exp\left(-\frac{1}{2}a^2\ell^2 + a\ell\Xi_n \sin(\theta) \right).$$

Then, the convergence in probability $L(X^n) - L_\infty(\Xi_n) \to 0$ holds. $L_\infty(\Xi_n)$ can be rewritten as

$$L_\infty(\Xi_n) = \int_0^1 da \int_0^{R_0} \frac{2\ell\, d\ell}{R_0^2} \int_0^{\pi} \frac{d\theta}{2\pi} \exp\left(-\frac{1}{2}a^2\ell^2 \right) \cosh\left(a\ell\Xi_n \sin(\theta) \right).$$

By using

$$t = a^2\ell^2,$$

the random variable $L_\infty(\Xi_n)$ can be also rewritten as

$$\begin{aligned} L_\infty(\Xi_n) &= \int_0^1 da \int_0^{a^2R_0^2} \frac{dt}{a^2R_0^2} \int_0^{\pi} \frac{d\theta}{2\pi} \exp\left(-\frac{t}{2} \right) \cosh\left(\sqrt{t}\Xi_n \sin(\theta) \right), \\ &= \int_0^{R_0^2} dt \int_{\sqrt{t}/R_0}^1 \frac{da}{a^2R_0^2} \int_0^{\pi} \frac{d\theta}{2\pi} \exp\left(-\frac{t}{2} \right) \cosh\left(\sqrt{t}\Xi_n \sin(\theta) \right), \\ &= \frac{1}{2R_0^2} \int_0^{R_0^2} dt \left(R_0/\sqrt{t} - 1 \right) \exp\left(-\frac{t}{2} \right) I_0(\sqrt{t}\Xi_n), \end{aligned}$$

which completes the theorem. □

As well as the results we obtained in the previous sections, the asymptote of L does not explicitly depend on the sample size n. The reason for this behavior is the same as in the previous cases, the critical scaling $r \propto n^{-1/2}$.

Let us validate the asymptote we derived above.

We firstly prepared the 10000 sample sets, whose size is denoted by n, and conducted the validation for four different values of n, i.e., $n = 100, 200, 400, 800$. We calculated the L by using Ξ calculated from the finite sample. Here, we set R_0 as 1. Then, we estimated the level numerically in each case and compared them with the levels of case 2.

Table 3 shows the result. Compared with the previous cases that used simpler models, in the present case, the numerically calculated levels slightly deviate from the theoretical values derived from the asymptote. But we can see that as the n becomes larger, the numerically calculated levels approach the theoretical value, and we can conclude that they match well and the asymptote we derived in the previous section works well even in the case of finite n.

Table 3.3: Comparison of levels derived from asymptote and those numerically calculated levels

level	10%	5%	1%
rejection region $L > r$	r=0.550	r=0.581	r=0.659
numerically calculated level(n=100)	9.53%	4.81%	1.21%
numerically calculated level(n=200)	9.72%	4.64%	0.88%
numerically calculated level(n=400)	10.04%	4.91%	0.79%
numerically calculated level(n=800)	10.33%	5.09%	1.03%

3.3 Comments from the perspective of the singular learning theory

Let us mention the relation between the result obtained above and the general asymptotic form of the log marginal likelihood of the singular model, which is derived from the theory of algebraic geometry (Watanabe [2001]).

In Theorem 2, we derived the asymptotic form of L and saw that L did not depend on the sample size n as a result of the scaling law $B \propto n^{-1/2}$ that we applied.

We can consider another scaling $B \propto n^{-\alpha}$, where $\alpha > 0$ is a constant. As long as $\alpha \leq \frac{1}{2}$, we can calculate the asymptotic form of L in the same way as the derivation of Theorem 2. The result is as follows.

$$L = \frac{1}{B_0 n^{\frac{1}{2}-\alpha}} \int_0^{B_0^2 n^{1-2\alpha}} \frac{1}{\sqrt{t}} \log \left[\frac{B_0^2 n^{1-2\alpha}}{t} \right] e^{-t/2} \cosh \xi \sqrt{t} dt \tag{3.16}$$

We can immediately obtain the log marginal likelihood ratio $F = -\log L$.

$$F = \left(\frac{1}{2} - \alpha \right) \log n - (1 - 2\alpha) \log (\log n) + o_p(\log (\log n)) \tag{3.17}$$

According to Watanabe [2001], the general asymptotic form of log marginal likelihood becomes

$$F = \lambda \log n - (m - 1) \log (\log n) + o_p(\log (\log n)) \tag{3.18}$$

We can see that our result corresponds to $\lambda = \frac{1}{2} - \alpha$ and $m = (2 - 2\alpha)$. The sample size's dependency on the support of the prior affects the real canonical log threshold λ and the multiplicity m. In this paper, we treated $\alpha = \frac{1}{2}$ as a "critical" case, where the λ and m effectively vanish. In such a case, the main term of F becomes stochastic. This is why it can be difficult to apply conventional Bayes factor-based testing to such a case.

Let us comment more on the scaling $n^{-1/2}$. The Kullback-Leibler divergence K between the null hypothesis and the alternative hypothesis can be easily calculated.

$$K(a.b) = \int p(x|(0,0)) \log \frac{p(x|(0,0))}{p(x|w)} dx = -\frac{1}{2} a^2 b^2 \tag{3.19}$$

Here, $nK(a,b)$ is nothing other than the leading term of the $H(a.b)$.

In the proof of the Theorem 2, we mainly considered that $b \propto n^{-1/2}$, and $a \sim \mathcal{O}(1)$. The meaning of this setup is clear, the center of the mixed distribution deviates from the origin , as much as the variance of the distribution, and the null and alternative hypothesis are hard to discriminate.

As a result of this scaling, both na^2b^2 and $\sqrt{n}ab\xi$ becomes $\mathcal{O}(1)$, and this result in the n-independent asymptote of L.

However, as we can be easily seen, this "scaling" is not unique. So long as $ab \sim n^{-1/2}$ and $bX_i - \frac{b^2}{2}$ is small enough that the Taylor expansion of the exponential is valid, a proof similar to the one above can be constructed. For example, a scaling such as $a \sim n^{-1/4}$ and $b \sim n^{-1/4}$ will lead to the same results.

The important point here is that this can be understood as a Taylor expansion around the singularity $ab = 0$, and the deviation is described as a power of ab, not of b.

As we saw above, in this delicate situation, a hypothesis test based on the stochastic behavior of L works well, and to construct it, we need to find the singularity (in our setting, $ab = 0$) and an appropriate scaling (in our setting, $ab \sim n^{-1/2}$)is essentially important.

Therefore, to construct the hypothesis test using singular models, we should keep in mind the effect of the singularity, and consider whether the case under consideration is "delicate" or not, by computing the Kullback-Leibler divergence between the null hypothesis and the alternative hypothesis. The scaling is determined by the form of the Kullback-Leibler divergence that consists of a polynomial for each parameters. From the perspective of the singular learning theory, this is nothing other than the relation between the real log canonical threshold (RCLT) λ and the representation of the parameters in the model.

3.4 Discussion

In this chapter, we theoretically studied the test of homogeneity for normal mixtures in terms of the Bayesian framework, for the first time. By applying the mathematical technique developed for the analysis of singular models and by appropriately scaling from the singularity, we derived the asymptotic behavior of the marginal likelihood ratio for several forms of the prior.

The merits of our treatment are as follows.

First, the test statistics that we analyzed was the marginal likelihood ratio and as a result of this, the hypothesis test using it is guaranteed to be the most powerful test. Second, compared with other methods using the value of the (log) likelihood ratio, such as Bayes factor based ones, the hypothesis test based on the stochastic behavior of the marginal likelihood ratio is valid even when the null hypothesis and the alternative one are hard to discriminate, as we saw. The stochastic behavior of the test statistics we derived can be described as a function of the probability variables obeying well-known probability distributions. From the practical perspective, this gives us a clear and easy-to-use formalism.

To conclude our discussion, we should note that in the field of Bayesian learning theory, the study of hypothesis tests is not sufficient and there is much that remains to be studied. We believe that our method is very general, and that it can be applied to various singular models. This direction of study could be of practical value. We also believe that it is also important to study the methods of approximating the log marginal likelihood ratios with high efficiency. One candidate for this is variational Bayes, which is an efficient way to approximate the posterior distribution. However, the theory of hypothesis test based on variational Bayes is still insufficient. Therefore, we should study how to apply it to a Bayesian hypothesis test. This is the theme of the next section.

Chapter 4

Testing homogeneity for normal mixture models using variational Bayes

4.1 Introduction

As we saw in the previous section, the properties of hypothesis tests for singular models are completely different from those for regular models. It clearly shows the importance of the theory for the hypothesis test for singular models.

In the Bayesian hypothesis test, we have to calculate the marginal log likelihood, the test statistics which gives the most powerful test. However, this requires vast computational resources in general, and an efficient scheme for the task is needed.

Variational Bayes (also called variational inference) (Attias [2000] Blei et al. [2017]) is known as an effective approximation method for this purpose, and it is widely used in various situations. Although the power of the variational Bayes is well-known, no study has applied it to the approximation of the marginal log likelihood ratio (variational free energy) to construct a hypothesis test, as long as we know.

Although there have been not many theoretical studies on variational Bayes for normal mixtures, but some important facts on the variational free energy have been clarified through asymptotic analysis in Watanabe and Watanabe [2006] Watanabe and Watanabe [2007]. Through the asymptotic analysis of the variational free energy, the authors proved that the phase transition occurs when the hyperparameter exceeds some critical value. Here, the "hyperparameter" means the parameter contained in the prior of the mixture ratio in our model. More concretely, in Watanabe and Watanabe [2006] Watanabe and Watanabe [2007], Dirichlet dis-

tribution is chosen as the prior of the mixture ratio as $\varphi(\{a_k\}) \propto \prod_k \{a_k\}^{\phi-1}$, where φ is the prior, $\{a_k\}$ is the mixture ratio of each component, and ϕ is the hyperparameter. The authors of Watanabe and Watanabe [2006] Watanabe and Watanabe [2007] theoretically clarified when the ϕ exceeds some value, the phase transition occurs. When the phase transition occurs, the stochastic behavior of the test statistics is greatly changed. Therefore, to construct a hypothesis test properly, the properties of the phase transition should be grasped.

Another important result their asymptotic analysis has derived is that the leading order of the variational free energy is $\mathcal{O}(\log n)$, when the sample size n is large, while the order of the stochastic term in the variational free energy is $\mathcal{O}(1)$ (Watanabe and Watanabe [2006]). The stochastic behavior of the test statistics needs to be clarified for constructing a hypothesis test, but the specific expression of it has not been obtained.

In summary, we have to clarify the asymptotic behavior of the variational free energy, to construct a new way for testing homogeneity based on the variational Bayes framework. It requires the stochastic behavior of the variational free energy, which is given as $\mathcal{O}(1)$ term in the asymptote of the variational free energy. The previous studies only revealed the asymptote within the order of $\mathcal{O}(\log n)$, and we have to proceed further and clarify the higher order term for our purpose. Also, the phase transition induced by the hyperparameter greatly affects the behavior of the variational free energy, and we have to clarify the influence of it in our problem. These are the problem we solve in this chapter.

The results shown below are as follows. First, we show that our model has the phase transition induced by the hyperparameter under the variational Bayes approximation. We obtain the critical value of the phase transition. Second, we derive the asymptote of the variational free energy within the order of $\mathcal{O}(1)$, when the hyperparameter ϕ is larger than the critical value. Based on these results, we construct a new hypothesis test scheme for the first time and demonstrate its validity with numerical experiments.

4.2 Variational Bayes

In this section, we apply the variational approximation to the marginal log likelihood ratio. Before this, let us rewrite the marginal likelihood ratio by using latent variables, for convenience.

As well as the previous chapter, we consider two-component normal mixture model as the probabilistic model,

$$p_0(x|w) = (1-a)\mathcal{N}(0,1^2) + a\mathcal{N}(b,1^2). \tag{4.1}$$

In this chapter, we assume that the N.H and A.H are as follows,

$$\begin{aligned}\varphi_0(a,b) &= \delta(a)\delta(b), \\ \varphi_1(a,b) &= Dir(a|\phi) \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}b^2\right).\end{aligned}$$

where $Dir(a|\phi)$ is a Dirichlet distributions of a,

$$Dir(a|\phi) \equiv \frac{\Gamma\left(\sum_k \phi\right)}{\prod_k \Gamma(\phi)} \prod_k a_k^{\phi-1}.$$

Here we define $a_0 = (1-a)$ and $a_1 = a$, and $b_0 = 0$ and $b_1 = b$. Hereafter, the index k means the dummy variable for the index of the cluster, and k runs from 0 to 1.

We also note that we choose conjugate prior distribution as a A.H., for simplifying the calculation of the variational approximation.

To calculate the numerator of marginal likelihood ratio $L(X^n)$ is a difficult task. In the following sections, we will discuss how to approximate this quantity efficienty and construct the hypothesis test based on it.

For this purpose, we introduce here latent variables $\{y_{ik}\}$. The meaning of the variable $y_{ik} \in \{0,1\}$ is to which cluster in the model X_i belongs. We should note that $\{y_{ik}\}$ satisfy $\sum_k y_{ik} = 1$.

The posterior under the A. H. can be written in terms of the latent variables as,

$$p(w, \{y_{ik}\}|X^n) \equiv \frac{1}{Z_n} \prod_{i,k} \left\{ a_k \frac{1}{\sqrt{2\pi}} e^{-(X_i-b_k)^2/2} \right\}^{y_{ik}} \varphi_1(w) \tag{4.2}$$

Hereafter, we abbreviate the set of the parameter$\{a,b\}$ as w, and the summation of $\{y_{ik}\}$ is taken for all configurations, Z_n is

$$Z_n \equiv \int dw \sum_{\{y_{ik}\}} \prod_{i,k} \left\{ a_k \frac{1}{\sqrt{2\pi}} e^{-(X_i-b_k)^2/2} \right\}^{y_{ik}} \varphi_1(w).$$

Let us approximate $p(w, \{y_{ik}\}|X^n)$ under the variational approximation. We will find a variational function $[q(\{y_{ik}\})r(w)]$ that minimizes the Kullbuck-Leibler divergence,

$$D(qr||p) = \int dw \sum_{\{y_{ik}\}} q(\{y_{ik}\})r(w) \log \frac{q(\{y_{ik}\})r(w)}{p(w, \{y_{ik}\}|X^n)},$$

as an approximate of $p(w, \{y_{ik}\}|X^n)$.

The $q(\{y_{ik}\})$ and $r(w)$ should satisfy the following conditions, which are obtained as a result of the variational principle.

$$q(\{y_{ik}\}) \propto \exp\left[E_r\left\{\log p(w, \{y_{ik}\}\,|X^n)\right\}\right],$$

$$r(w) \propto \exp\left[E_q\left\{\log p(w, \{y_{ik}\}\,|X^n)\right\}\right].$$

where $E_r\{\cdot\}$ means the expectation value with respect to $r(w)$, and $E_q\{\cdot\}$ means the expectation value with respect to $q(\{y_{ik}\})$.

The logarithm of $p(w, \{y_{ik}\}, X^n)$ becomes

$$\begin{aligned}\log p(w, \{y_{ik}\}\,|X^n) &= \sum_i \sum_k y_{ik}\left[\log a_k - \frac{1}{2}\left(X_i - \delta_{k1}b\right)^2\right] \\ &- \frac{n}{2}\log(2\pi) + \log\varphi_1(a,b), \end{aligned} \tag{4.3}$$

It is linear with respect to y_{ik}, and $r(w)$ becomes,

$$\begin{aligned} r(w) &\propto \exp\left[E_q\left\{\log p(w, \{y_{ik}\}\,|X^n)\right\}\right] \\ &= \prod_k \prod_i a_k^{\hat{y_{ik}}} \frac{1}{\sqrt{2\pi}}\left\{\exp\left[-\frac{1}{2}\left(X_i - \delta_{k1}b\right)^2\right]\right\}^{\hat{y_{ik}}} \\ &\times \varphi_1(a,b), \end{aligned} \tag{4.4}$$

where $\hat{y_{ik}}$ means $E_q\{y_{ik}\}$.

As well as $r(w)$, $q(\{y_{ik}\})$ can be calculated as,

$$\begin{aligned} q(y_{ik}) &\propto \exp\left[\sum_i \sum_k y_{ik}\left\{\langle\log a_k\rangle - \frac{1}{2}\langle\left(X_i - \delta_{k1}b\right)^2\rangle\right\}\right] \\ &= \prod_i \prod_k \left\{\exp\left[\langle\log a_k\rangle - \frac{1}{2}\langle\left(X_i - \delta_{k1}b\right)^2\rangle\right]\right\}^{y_{ik}} \end{aligned} \tag{4.5}$$

where $\langle\{\cdot\}\rangle$ is an abbreviation of $E_r\{\cdot\}$.

The self-consistent equations that $\hat{y_{ik}}$ should satisfy become

$$\hat{y_{ik}} \propto \exp\left[\langle\log a_k\rangle - \frac{1}{2}\langle\left(X_i - \delta_{k1}b\right)^2\rangle\right] \tag{4.6}$$

The self-consistent equations a_k should satisfy are

$$\langle\log a_k\rangle = \psi\left(\sum_i \hat{y_{ik}} + \phi\right) - \psi\left(n + 2\phi\right), \tag{4.7}$$

where $\psi(z) \equiv \frac{\partial}{\partial z}\Gamma(z)$ is the digamma function. Here, $\Gamma(z)$ means Gamma function.

From these results, $r(w)$ can be written as,

$$\begin{aligned} r(w) &\propto \prod_k \prod_i a_k^{\hat{y_{ik}}} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2} + \sum_i \hat{y_{ik}}\right)\right] \\ &\times \exp\left(b - \frac{\sum_i X_i \hat{y_{ik}}}{\frac{1}{\sigma^2} + \sum_i \hat{y_{ik}}}\right)^2 \end{aligned} \tag{4.8}$$

and the equations which $\langle b \rangle$ and $\langle b^2 \rangle$ should satisfy become

$$\langle b \rangle = \frac{\sum_i X_i \hat{y_{ik}}}{\sum_i \hat{y_{ik}} + \frac{1}{\sigma^2}},$$

$$\langle b^2 \rangle = \langle b \rangle^2 + \frac{1}{\sum_i \hat{y_{ik}} + \frac{1}{\sigma^2}}.$$

As a result of them, the self-consistent equations for $\hat{y_{ik}}$ becomes as follows,

$$\hat{y_{i0}} \propto \exp\left[\psi\left(\sum_i \hat{y_{i0}} + \phi\right) - \psi(n + 2\phi) - \frac{1}{2}X_i^2\right] \tag{4.9}$$

$$\begin{aligned} \hat{y_{i1}} &\propto \exp\left[\psi\left(\sum_i \hat{y_{i1}} + \phi\right) - \psi(n + 2\phi) - \frac{1}{2}\langle (X_i - b)^2 \rangle\right] \\ &= \exp\left[\psi\left(\sum_i \hat{y_{i1}} + \phi\right) - \psi(n + 2\phi)\right] \\ &\times \exp\left[-\frac{1}{2}\left\{(X_i - \langle b \rangle)^2 + \frac{1}{\sum_i \hat{y_{i1}} + \frac{1}{\sigma^2}}\right\}\right] \end{aligned} \tag{4.10}$$

Using these results, the variational free energy becomes

$$\begin{aligned}
F &= \langle q(\{y_{ik}\})r(w)\log\frac{q(\{y_{ik}\})r(w)}{p(w,X^n)}\rangle \\
&= \sum_i\sum_k \hat{y_{ik}}\log\hat{y_{ik}} + \log\frac{\Gamma(\sum\langle a_k\rangle)}{\prod\Gamma(\langle a_k\rangle)} - \log\frac{\Gamma(2\phi)}{\prod_k\Gamma(\phi)} \\
&+ \frac{1}{2}\log\left(1+\sigma^2\sum_i\hat{y_{i1}}\right) + \frac{1}{2\sigma^2}\langle b\rangle^2 \\
&+ \frac{1}{2}\sum_i\sum_k \hat{y_{ik}}\left(X_i - \delta_{k1}\langle b_k\rangle\right)^2 + \frac{1}{2}\sum_i\sum_k \hat{y_{ik}}\log(2\pi) \\
&= \sum_i\sum_k \hat{y_{ik}}\log\hat{y_{ik}} + \log\frac{\Gamma(\sum_k\sum_i\hat{y_{ik}}+2\phi)}{\prod\Gamma(\sum_i\hat{y_{ik}}+\phi)} \\
&+ \frac{1}{2}\log\left(1+\sigma^2\sum_i\hat{y_{i1}}\right) + \frac{1}{2}\sum X_i^2 \\
&- \frac{1}{2}\frac{(\sum X_j\hat{y_{j1}})^2}{\sum_i\hat{y_{i1}}+\frac{1}{\sigma^2}} + \frac{n}{2}\log(2\pi) - \log\frac{\Gamma(2\phi)}{\prod_k\Gamma(\phi)}
\end{aligned} \tag{4.11}$$

The logarithm of the denominator of L is calculated as

$$\begin{aligned}
F_0 &= -\log\int\varphi_0(w)\prod p_0(X_i,w)dw \\
&= \frac{1}{2}\sum_i X_i^2 + \frac{n}{2}\log(2\pi)
\end{aligned}$$

Therefore, we obtain the logarithm of L, variational free energy, as

$$\begin{aligned}
F - F_0 &= \sum_i\sum_k \hat{y_{ik}}\log\hat{y_{ik}} + \log\frac{\Gamma(\sum_k\sum_i\hat{y_{ik}}+2\phi)}{\prod\Gamma(\sum_i\hat{y_{ik}}+\phi)} \\
&+ \frac{1}{2}\log\left(1+\sigma^2\sum_i\hat{y_{i1}}\right) - \frac{1}{2}\frac{(\sum X_j\hat{y_{j1}})^2}{\sum_i\hat{y_{i1}}+\frac{1}{\sigma^2}} \\
&- \log\frac{\Gamma(2\phi)}{\prod(\Gamma(\phi))}
\end{aligned} \tag{4.12}$$

To construct a hypothesis test, the stochastic behavior of $F - F_0$ is needed, and we should proceed further. It requires the stochastic behavior of $\hat{y_{i1}}$. In the following sections, we derive the the stochastic behavior of $\hat{y_{i1}}$ for this purpose.

We should also note that the variational free energy exhibits the phase transition when the hyperparameter ϕ changes. This affects the configuration and stochastic behavior of $\hat{y_{i1}}$, as we will see in the next section. This is the root of our motivation to investigate the phase transition of the model.

4.3 Phase transition induced by the hyperparameter

In our problem, the parameter sets that minimize the variational free energy should be those that correspond to the null hypothesis. Actually, such a parameter is not unique. More specifically, $\{y_{i1}\}$ that satisfies $\sum_{y_{i1}} = \mathcal{O}(1)$ and $\langle b \rangle = 0$ is one candidate, but also $\{y_{i1}\}$ that satisfies $\sum y_{i1} = 0$ is another one. In other words, these solutions seem degenerated.

However, the prior we consider has a hyperparameter ϕ, and it is natural to expect that the degeneration be solved when we tune this hyperparameter, and only one of the solution becomes a true solution of the variational Bayes. This is nothing other than the phase transition induced by the hyperparameter.

In a previous study Watanabe and Watanabe [2006], the upper and lower bounds of the asymptote of the variational free energy were studied, within $\mathcal{O}(\log(n))$. They showed that the phase transition occurs by tuning the hyperparameter. Therefore, it is clear that the phase transition drastically change the variational free energy even at the term of $\mathcal{O}(\log(n))$, which is *not* the stochastic term . It is natural for us to expect that the phase transition *also* greatly affect the $\mathcal{O}(1)$ term, which dominates the stochastic behavior of the variational free energy, the test statistics.

Therefore, from the perspective of our purpose, we consider that the effect of the phase transition should be studied and we should grasp what kind of solution is obtained as a function of the hyperparameter. This is the main purpose in this section. We firstly show the existence of the phase transition and derive the critical point $\phi_{\rm cr}$.

4.3.1 Asymptotic form of F when $\sum_i y_{i1}$ is $\mathcal{O}(n)$

Our main purpose here is to construct the hypothesis test using variational Bayes, and it is natural for us to focus on a situation when the hypothesis test plays an important role. That is the situation when the two hypotheses we treat are very similar and it is difficult to distinguish between them. Specifically, we consider the situation $\langle b \rangle$ is small and two gaussian distribution in our model are largely overlapped, $\langle b \rangle X_{\max} \sim o_p(1)$, under the null hypothesis.

Under this assumption above, the following theorem holds.

Theorem 5. When $\sum_i y_{i1}$ is $\mathcal{O}(n)$ and $\langle b \rangle \sim o(1/\sqrt{\log n})$, the asymptotic form of the variational free energy becomes

$$F - F_0 = \log n + o(\log n) \tag{4.13}$$

under the null hypothesis.

proof. Let us introduce $\bar{y}$ as $\sum_i \hat{y_{i1}} \equiv n_1$ and $n_1/n \equiv \alpha \sim \mathcal{O}(1)$ for brevity.

The self-consistent equation of $\{y_{i1}\}$becomes

$$\begin{aligned}
\hat{y_{i1}} &= \frac{n_1 e^{\langle b \rangle X_i - 1/2\langle b^2 \rangle}}{n - n_1 + n_1 e^{\langle b \rangle X_i - 1/2\langle b^2 \rangle}} \\
&= \frac{n_1}{n} + \frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)\langle b \rangle X_i \\
&+ \frac{1}{2}\frac{n_1}{n}\langle b \rangle^2 \left(1 - \frac{n_1}{n}\right)\left(X_i^2 - 1\right) \\
&- \left(\frac{n_1}{n}\right)^2 \langle b \rangle^2 X_i^2 + \left(\frac{n_1}{n}\right)^3 \langle b \rangle^2 X_i^2 + \mathcal{O}\left(\left(\langle b \rangle X_i\right)^3\right) \\
&= \alpha + \alpha(1-\alpha)\langle b \rangle X_i \\
&+ \frac{1}{2}\alpha \langle b \rangle^2 \left[X_i^2\left(1 - 3\alpha + 2\alpha^2\right) + \alpha - 1\right] + \mathcal{O}\left(\left(\langle b \rangle X_i\right)^3\right).
\end{aligned}$$

Using this, $\langle b \rangle$ becomes

$$\begin{aligned}
\langle b \rangle &= \frac{\sum_j \hat{y_{j1}} X_j}{\sum_j \hat{y_{j1}} + \frac{1}{\sigma^2}} \\
&= \frac{\sum_j X_j \left(\alpha + \alpha\left(1-\alpha\right)\langle b \rangle X_j + \mathcal{O}(\langle b \rangle^2)\right)}{n_1 + \frac{1}{\sigma^2}} \\
&= \frac{1}{n}\sum X_j + \langle b \rangle \left(1 - \alpha\right) + o(n^{-1/2}).
\end{aligned}$$

Therefore, we obtain

$$\langle b \rangle = \frac{1}{n_1}\sum_j X_j + o(n^{-1/2}). \tag{4.14}$$

From this result, $\hat{y_{i1}}$ becomes

$$\begin{aligned}
\hat{y_{i1}} &= \alpha + \frac{(1-\alpha)}{n}\sum_j X_j X_i \\
&+ \frac{1}{2\alpha n^2}\left(\sum_j X_j\right)^2 \left[\alpha - 1 + \left(1 - 3\alpha + 2\alpha^2\right) X_i^2\right] \\
&+ o\left(\frac{\log n}{n}\right).
\end{aligned} \tag{4.15}$$

Here we used the well known result that the maximum of the $\{X_i\}$ is at most in the order of $\mathcal{O}(\sqrt{\log n})$, when $X_i \sim \mathcal{N}(0, 1^2)$.

Now we can calculate the variational free energy explicitly from these results. Let us rewrite $\hat{y_{i1}}$ as a sum of the mean and the fluctuation term, for simplicity,

$$\hat{y_{i1}} = \alpha + \Delta y_i.$$

The entropy term becomes,

$$\begin{aligned}
& \sum_i \{\hat{y_{i1}} \log \hat{y_{i1}} + (1 - \hat{y_{i1}}) \log (1 - \hat{y_{i1}})\} \\
= & \sum_i (\alpha + \Delta y_i) \log (\alpha + \Delta y_i) \\
+ & [1 - (\alpha + \Delta y_i)] \log [1 - (\alpha + \Delta y_i)] \\
= & n [\alpha \log \alpha + (1 - \alpha) \log (1 - \alpha)] \\
+ & \sum_i \Delta y_i [\log \alpha - \log (1 - \alpha)] + \sum_i \frac{1}{2} (\Delta y_i)^2 \left[\frac{1}{\alpha} + \frac{1}{1 - \alpha}\right] \\
+ & \sum_i \mathcal{O}\left(\Delta y_i^3\right).
\end{aligned}$$

From the equation (16) , the sum of Δy_i becomes

$$\begin{aligned}
\sum_i \Delta y_i &= \frac{(1 - \alpha)}{n} \sum_{i,j} X_j X_i \\
&+ \frac{1}{2\alpha n^2} \left(\sum_j X_j\right)^2 \sum_i \left[\alpha - 1 + \left(1 - 3\alpha + 2\alpha^2\right) X_i^2\right] \\
&= \left[(1 - \alpha) + \frac{\alpha - 1}{2\alpha} + \frac{1 - 3\alpha + 2\alpha^2}{2\alpha}\right] \left(\frac{1}{\sqrt{n}} \sum_i X_i\right)^2 \\
&+ \frac{1}{2\alpha n} \left(1 - 3\alpha + 2\alpha^2\right) \sum_i \left(X_i^2 - 1\right) \left(\frac{1}{\sqrt{n}} \sum_i X_i\right)^2 \\
&= \frac{1}{2\alpha n} \left(1 - 3\alpha + 2\alpha^2\right) \sum_i \left(X_i^2 - 1\right) \left(\frac{1}{\sqrt{n}} \sum_i X_i\right)^2 \\
&= \mathcal{O}\left(n^{-1/2}\right).
\end{aligned}$$

and the sum of the square of Δy_i becomes

$$
\begin{aligned}
& \sum_i (\Delta y_i)^2 \\
= \; & \frac{(1-\alpha)^2}{n^2}\left(\sum_j X_j\right)^2 \sum_i X_i^2 - \frac{(1-\alpha)^2}{\alpha n^3}\left(\sum_j X_j\right)^4 \\
+ \; & \frac{(1-\alpha)}{\alpha n^3}(1-3\alpha+2\alpha^2)\left(\sum_j X_j\right)^3 \sum_i X_i^3 \\
+ \; & \frac{1}{4\alpha^2 n^4}\left(\sum_j X_j\right)^4 \sum_i \left\{\alpha - 1 + (1-3\alpha+2\alpha^2)X_i^2\right\}^2 \\
= \; & (1-\alpha)^2\left(\frac{1}{\sqrt{n}}\sum_i X_i\right)^2 + \mathcal{O}(n^{-1/2}).
\end{aligned}
$$

Therefore, the entropy term becomes

$$
\begin{aligned}
& \sum_i \left\{\hat{y_{i1}} \log \hat{y_{i1}} + (1-\hat{y_{i1}}) \log\left(1-\hat{y_{i1}}\right)\right\} \\
= \; & n\left[\alpha \log \alpha + (1-\alpha)\log(1-\alpha)\right] + \frac{1-\alpha}{2\alpha}\xi^2.
\end{aligned}
$$

The other terms can be calculated as follows:

$$
\begin{aligned}
& \log \frac{\Gamma(n+2\phi)}{\Gamma(n_1+\phi)\Gamma(n-n_1+\phi)} \\
= \; & \frac{1}{2}\log n - \left(n\alpha + \phi - \frac{1}{2}\right)\log \alpha \\
- \; & \left(n(1-\alpha) + \phi - \frac{1}{2}\right)\log(1-\alpha) - \frac{1}{2}\log 2\pi + o(1).
\end{aligned}
$$

$$
\begin{aligned}
& \frac{1}{2}\log\left(1+\sigma^2 \sum y_{i1}\right) \\
\simeq \; & \frac{1}{2}\left[\log n + \log\alpha + \log\sigma^2\right] + \frac{1}{2}\log\left(1+\frac{1}{n\alpha\sigma^2}\right) \\
= \; & \frac{1}{2}\left[\log n + \log\alpha + \log\sigma^2\right] + o(1).
\end{aligned}
$$

$$
\begin{aligned}
-\frac{1}{2}\frac{\left(\sum X_i \hat{y_{i1}}\right)^2}{\sum \hat{y_{i1}} + \frac{1}{\sigma^2}} &= -\frac{1}{2}\langle b\rangle^2\left(\sum \hat{y_{i1}} + \frac{1}{\sigma^2}\right) \\
&= -\frac{1}{2\alpha}\xi^2 + o\left(1\right).
\end{aligned}
$$

Here we used the asymptotic form of the gamma function,

$$\log\Gamma(n) = \left(n - \frac{1}{2}\right)\log n - n + \frac{1}{2}\log 2\pi + o(1).$$

By integrating them, we obtain the variational free energy

$$\begin{aligned} F - F_0 &= \log n + (1-\phi)\log\alpha - \left(\phi - \frac{1}{2}\right)\log(1-\alpha) + \frac{1}{2}\log\sigma^2 \\ &- \frac{1}{2}\xi^2 - \log\frac{\Gamma(2\phi)}{\prod(\Gamma(\phi))} - \frac{1}{2}\log 2\pi + o(1). \end{aligned} \tag{4.16}$$

From these results, we obtain

$$F - F_0 = \log n + o(\log n), \tag{4.17}$$

and the proof is completed. □

4.3.2 Asymptotic form of F when $\sum_i y_{i1}/n \to 0$

When $\sum_i y_{i1}/n \to 0$, the following theorem on the asymptotic form of F holds.

Theorem 6. Let us define the function $f(y_i)$ as,

$$\begin{aligned} f(y_i) &= \sum_i \{\hat{y_{i1}}\log\hat{y_{i1}} + (1-\hat{y_{i1}})\log(1-\hat{y_{i1}})\} \\ &- \frac{\left(\sum_i X_i\hat{y_{i1}}\right)^2}{2\left(n_1 + 1/\sigma^2\right)} \end{aligned}$$

When we fix $\sum_i \hat{y_{i1}} = n_1$, the minimum of $f(\hat{y_{i1}})$ satisfies

$$f(y_i) = -n_1\log n + n_1\log n_1 - n_1 + o(1) \tag{4.18}$$

and $F - F_0$ becomes

$$F - F_0 = \phi\log\frac{n}{n_1} + \log n_1 + \mathcal{O}_p(1) \tag{4.19}$$

proof. In this case, the logarithm of the ratio of the gamma function becomes,

$$\begin{aligned} &\frac{\log\Gamma(n+2\phi)}{\prod\log\Gamma(\sum_i \hat{y_{ik}}) + \phi)} = (n_1+\phi)\log n \\ - &\ (n_1 + \phi - \frac{1}{2})\log(n_1+\phi) - (n-n_1)\log\left(1 - \frac{n_1}{n}\right) + \mathcal{O}(1) \end{aligned}$$

where we define $\sum_i \hat{y_{i1}} \equiv n_1$. We should note that the leading order is different from the case we saw in the previous subsection.

Applying the method of Lagrange multipliers, we minimize the function such as,

$$\begin{aligned} f_1(\hat{y_{i1}}) &= \sum_i \{\hat{y_{i1}} \log \hat{y_{i1}} + (1-\hat{y_{i1}}) \log (1-\hat{y_{i1}})\} \\ &- \frac{(\sum_i X_i \hat{y_{i1}})^2}{2\,(n_1 + 1/\sigma^2)} + \lambda \left(\sum_i \hat{y_{i1}} - n_1 \right). \end{aligned}$$

The equation of the stationary condition is given as

$$\frac{\partial f_1}{\partial \hat{y_{i1}}} = \log \frac{\hat{y_{i1}}}{1-\hat{y_{i1}}} - \frac{\sum_j X_j \hat{y_{j1}} X_i}{(n_1 + 1/\sigma^2)} + \lambda = 0. \tag{4.20}$$

By solving it with y_i, we obtain

$$\hat{y_{i1}} = \frac{1}{1 + \exp(-A(X_i - B))}, \tag{4.21}$$

where

$$\begin{aligned} A &\equiv \frac{\sum_j X_j \hat{y_{j1}}}{(n_1 + 1/\sigma^2)} \\ B &\equiv \frac{\lambda}{A}. \end{aligned}$$

Let us assume that $A > 0$, and $X_1 \leq X_2 \leq ... \leq X_n$. This assumption does not lose the generality. Under this assumption, the following lemma on the asymptotic form of the trimmed sum of X_i holds.

Lemma 1. Let $X_1 \leq X_2 \leq ... \leq X_n$be an i.i.d sample generated from the standard normal distribution $\mathcal{N}(0, 1^2)$.

Let us consider the trimmed sum of the largest n_1th data from the sample. When $n_1 \to \infty$ and $n_1/n \to 0$, the asymptotic behavior of the sum is

$$S = \sum_{i=n-n_1+1}^{n} X_i \to \sqrt{2 \log \frac{n}{n_1}} + o_p(n_1) \tag{4.22}$$

proof of Lemma 1. As a normal distribution satisfies the von Mises conditions, the asymptote of the n_1th maximum values $x_{(n-n_1+1),n}$ satisfies

$$\left(X_{(n-n_1+1),n} - a_n\right)/b_n \to \mathcal{N}(0,1), \tag{4.23}$$

where $a_n \equiv F^{-1}\left(1-\frac{n_1}{n}\right)$ and $b_n \equiv \sqrt{n_1}/\left(nf(a_n)\right)$, here $F(x)$ means the cumulative distribution function of X_i, and $f(x)$ means the distribution function of X_i (see Theorem 8.3.4 and Theorem 8.5.3 in Barry C. Arnold and Nagaraja [2008]).

In our case, the asymptotic form of a_n becomes $a_n \to \sqrt{2\log\frac{n}{n_1} - \log\log\left(\frac{n}{n_1}\right)^2}$, $b_n \to \frac{\sqrt{n_1}}{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}a_n^2}$, and the leading term of the $X_{(n-n_1+1),n}$becomes

$$X_{(n-n_1+1),n} = a_n + o_p(\frac{n_1}{n}). \tag{4.24}$$

Let us proof the lemma using this result. First, we split the $X_{(n-n_1+1),n}, \ldots X_{(n),n}$ samples by T groups that satisfy $1 \ll T < n_1 \ll n$. Each group contains $[n_1/T]$ terms.
The maximum in the $t+1$th group, Y_{t+1} satisfies

$$\begin{aligned}
Y_{t+1} &\leq \sqrt{2\log\left(n/\left(n_1 * t/T\right)\right)} + o_p(\frac{n_1}{n}) \\
&= \sqrt{2\log\left(n/n_1\right) + 2\log\left(T/t\right)} + o_p(\frac{n_1}{n}) \\
&= \sqrt{2\log\left(n/n_1\right)} \times \sqrt{1+\log\left(T/t\right)/\log\left(n/n_1\right)} + o_p(\frac{n_1}{n}) \\
&\leq \sqrt{2\log\left(n/n_1\right)} \times \left(1+\log\left(T/t\right)/\log\left(n/n_1\right)\right) \\
&= \sqrt{2\log\left(n/n_1\right)} + \sqrt{2}\log\left(T/t\right)/\sqrt{\log\left(n/n_1\right)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i=n-n_1+1}^{n} x_{(i),n} &\leq \frac{n_1}{T}\sum_{t=1}^{T-1} Y_{t+1} + \frac{n_1}{T}\sqrt{2\log n} \\
(r.h.s) &= \frac{n_1}{T}\left(T-1\right)\sqrt{2\log\left(n/n_1\right)} \\
&+ \sqrt{2}\left(T-1\right)/\sqrt{\log\left(n/n_1\right)} + \frac{n_1}{T}\sqrt{2\log n} \\
&\leq \left(n_1 - \frac{n_1}{T}\right)\sqrt{2\log\left(n/n_1\right)} \\
&+ \frac{n_1}{T}\left(\sqrt{2\log n/n_1} + \sqrt{2}\frac{\log n_1}{\sqrt{\log(n/n_1)}}\right) \\
&= n_1\sqrt{2\log\left(n/n_1\right)} + \frac{n_1}{T}\frac{\sqrt{2}\log n_1}{\sqrt{\log\left(n/n_1\right)}} + o_p(n_1).
\end{aligned}$$

If we choose T that satisfies $1 \ll T < n_1 \ll n$ properly, e.g., $T = \sqrt{n_1}$, the second term becomes $o_p(n_1)$. and the lemma is proven.

Using this, we can obtain

$$0 < A \leq \sqrt{2 \log \frac{n}{n_1}}$$

As $\sum_i \hat{y_{i1}} = n_1$, and $\lim_{n\to\infty} \frac{n_1}{n} = 0$, the number of $\hat{y_{i1}}$ that satisfies $\hat{y_{i1}} > 1/2$ should not be $\mathcal{O}(n)$. Therefore, B should go to ∞ when $n \to \infty$.

Let Z be a constant that satisfies $Z \to \infty$ and $\frac{B}{Z} \to \infty$, and let us write the number of $\hat{y_{i1}}$ that satisfies $X_i < B/Z$ as αn_1, and $\beta \equiv 1 - \alpha$.

Under these assumption, $AB \to \infty$ when $n \to \infty$. This is because if AB does not diverge, $\hat{y_{i1}}$ becomes larger than 0, but this leads $n_1 = \sum_i \hat{y_{i1}}$ does not diverge. Therefore, when $n_1 \to \infty$, $AB \to \infty$ is needed.

As a result of this, the following holds from (4.21),

$$\begin{aligned} |X_i| \leq B/Z \Rightarrow \hat{y_{i1}} &= \frac{1}{1 + \exp(-AX_i + AB)} \\ &\sim \exp(-AB). \end{aligned} \tag{4.25}$$

Let us split $f(\hat{y_{i1}})$ into the three parts,

$$f(\hat{y_{i1}}) = f_1(\hat{y_{i1}}) + f_2(\hat{y_{i1}}) + f_3(\hat{y_{i1}}), \tag{4.26}$$

where

$$f_1 = \sum_{X_i < B/Z} \hat{y_{i1}} \log \hat{y_{i1}} + (1 - \hat{y_{i1}}) \log(1 - \hat{y_{i1}}), \tag{4.27}$$

$$f_2 = \sum_{X_i > B/Z} \hat{y_{i1}} \log \hat{y_{i1}} + (1 - \hat{y_{i1}}) \log(1 - \hat{y_{i1}}), \tag{4.28}$$

$$f_3 = -\frac{\left(\sum_i X_i \hat{y_{i1}}\right)^2}{2\left(n_1 + 1/\sigma^2\right)}. \tag{4.29}$$

From the convexity, f_1 satisfies the following inequality as,

$$\begin{aligned} f_1 &\geq n\left[\frac{\alpha n_1}{n} \log \frac{\alpha n_1}{n} + \left(1 - \frac{\alpha n_1}{n}\right) \log\left(1 - \frac{\alpha n_1}{n}\right)\right] \\ &= -\alpha n_1 \log n + \alpha n_1 \log \alpha n_1 - \alpha n_1 + \alpha^2 \frac{n_1^2}{n}. \end{aligned}$$

Also, f_2 satisfies

$$f_2 = \sum_{X_i > B/Z} \hat{y_{i1}} \log \hat{y_{i1}} + (1 - \hat{y_{i1}}) \log(1 - \hat{y_{i1}}) \geq -\beta n_1 \log 2, \tag{4.30}$$

because the minimum of the function $g(y) = y\log y + (1-y)\log(1-y)$ is $g(y=1/2) = -\log 2$.

As for $f_3(\hat{y_{i1}})$, the term $\sum_i X_i \hat{y_{i1}}$ in the numerator satisfies,

$$\begin{aligned}\sum_i X_i \hat{y_{i1}} &= \sum_{X_i > B/Z} X_i \hat{y_{i1}} + \sum_{X_i < B/Z} X_i \hat{y_{i1}} \\ &\leq \left(\sum_{X_i > B/Z} X_i \right) + \sum_{X_i < B/Z} X_i \hat{y_{i1}}\end{aligned}$$

In the last line, we use the result,

$$\alpha n_1 = n \exp\left(-AB\right). \tag{4.31}$$

Therefore,

$$\sum_i X_i \hat{y_{i1}} \leq \alpha \frac{n_1}{n} \sum_{X_i < B/Z} X_i + \beta n_1 \sqrt{2 \log \frac{n}{\beta n_1}}. \tag{4.32}$$

The condition in which equality is satisfied is $\alpha = 1, \beta = 0$, and

$$f_3(\hat{y_{i1}}) \geq -\frac{1}{2} \frac{\left(\alpha n_1 \frac{1}{n} \sum_{X_i < B/Z} X_i \right)^2}{(n_1 + 1/\sigma^2)}. \tag{4.33}$$

We can see that the α and β that gives the maximum of $f_1 + f_2$ under $\alpha + \beta = 1$ are also $\alpha = 1, \beta = 0$.
Therefore, we obtain

$$f(\hat{y_{i1}}) \geq -n_1 \log n + n_1 \log n_1 - n_1 + \frac{n_1^2}{n} + o(1). \tag{4.34}$$

By adding the log gamma term to it, we obtain the minimum of the variational free energy,

$$F - F_0 = \phi \log \frac{n}{n_1} + \log n_1 + \mathcal{O}_p(1). \tag{4.35}$$

□

From the results of Theorem 1 and Theom 2, we can obtain the asymptotic behavior of the variational free energy as a function of ϕ within the $\mathcal{O}(\log n)$ as

$$F - F_0 = \begin{cases} \phi \log n + o\left(\log n\right) & (\phi < 1) \\ \log n + o\left(\log n\right) & (otherwise) \end{cases}$$

This clearly shows that the phase transition exists in our model, and the critical value of the hyperparameter $\phi_{\rm cr}$ becomes $\phi_{\rm cr} = 1$. Note that these results are different from those obtained in the previous study (Watanabe and Watanabe [2006]), because our model and theirs have different parameter space.

We should also note that the configuration of $\{\hat{y_{i1}}\}$ of the solution is clearly different depending on the value of the hyperparameter. When the $\phi \geq 1$, the solution satisfies $\sum \hat{y_{i1}} \sim \mathcal{O}(1)$. This means that under the A. H., the data is described by a mixture of two clusters with very close averages and it is very difficult to distinguish between the two clusters.

In contrast, when $\phi < 1$, the $\sum \hat{y_{i1}}$ becomes small. This means that the vast majority of the sample can be described under the A. H. by a single normal distribution with a center at the origin, but there are a few data that can be regarded as belonging to a normal distribution with a center away from the origin. Under such a circumstance, the hypothesis test scheme based on this can be regarded as testing for the existence of outliers.

The lesson from our result is that we should choose an appropriate hyperparameter suitable for the purpose.

4.4 Asymptotic form of the variational free energy on the $\mathcal{O}(1)$

In this section, we consider again a situation in which it is difficult to distinguish whether or not a sample is generated from one cluster. From the discussion in the previous section, this corresponds to the case in which $\phi > 1$.

On the asymptotic form of the variational free energy, the following theorem holds under the above assumption.

Theorem 7. The variational free energy of the two component Gaussian mixture becomes

$$\begin{aligned} F - F_0 &= \log n - (\phi - 1)\log(\phi - 1) - \left(\phi - \frac{1}{2}\right)\log\left(\phi - \frac{1}{2}\right) \\ &+ \left(2\phi - \frac{3}{2}\right)\log\left(2\phi - \frac{3}{2}\right) + \frac{1}{2}\log\sigma^2 - \frac{1}{2}\xi^2 \\ &- \frac{1}{2}\log 2\pi - \log\frac{\Gamma(2\phi)}{\prod(\Gamma(\phi))} + o(1) \end{aligned} \tag{4.36}$$

when the hyperparameter satisfies $\phi > 1$.

Here, ξ is a probabilistic variable that obeys $\xi \sim \mathcal{N}(0, 1^2)$.

proof. As we proved in Theorem 5, the variational free energy becomes

$$\begin{aligned} F - F_0 &= \log n + (1-\phi)\log\alpha - \left(\phi - \frac{1}{2}\right)\log(1-\alpha) + \frac{1}{2}\log\sigma^2 \\ &- \frac{1}{2}\xi^2 - \log\frac{\Gamma(2\phi)}{\prod(\Gamma(\phi))} - \frac{1}{2}\log 2\pi + o(1). \end{aligned} \tag{4.37}$$

From the variational principle, α is determined as a solution of the following equation,

$$\alpha = \mathrm{argmin}\left[\mathrm{F}(\alpha)\right] \equiv \alpha_0. \tag{4.38}$$

α_0 is the solution of $\frac{dF}{d\alpha} = 0$, that is,

$$\alpha_0 = \frac{\phi - 1}{2\phi - \frac{3}{2}} = \frac{1}{2}\frac{4\,(\phi-1)}{1 + 4\,(\phi-1)} \tag{4.39}$$

By substituting this into F, we can obtain

$$\begin{aligned} F - F_0 &= \log n - (\phi-1)\log(\phi-1) - \left(\phi - \frac{1}{2}\right)\log\left(\phi - \frac{1}{2}\right) \\ &+ \left(2\phi - \frac{3}{2}\right)\log\left(2\phi - \frac{3}{2}\right) + \frac{1}{2}\log\sigma^2 - \frac{1}{2}\xi^2 \\ &- \frac{1}{2}\log 2\pi - \log\frac{\Gamma(2\phi)}{\prod(\Gamma(\phi))} + o(1). \end{aligned} \tag{4.40}$$

This is the result that we want to derive. □

In Figure 4.1, α_0 is plotted as a function of ϕ.

We can see that α_0 behaves as a function of ϕ, around the critical point $\phi_{\mathrm{cr}} = 1$ as,

$$\alpha_0 \sim (\phi - \phi_{\mathrm{cr}}) \tag{4.41}$$

The asymptotic form of F clearly shows that the stochastic behavior of the variational free energy is determined by the stochastic variable ξ.

Under the N. H., ξ follows a standard normal distribution and the distribution of the variational free energy can be described by a χ^2 distribution. We will examine the validity of the results in the next section.

4.5 Numerical experiment

In this section, we show the result of our numerical experiments to examine the validity of our theoretical results.

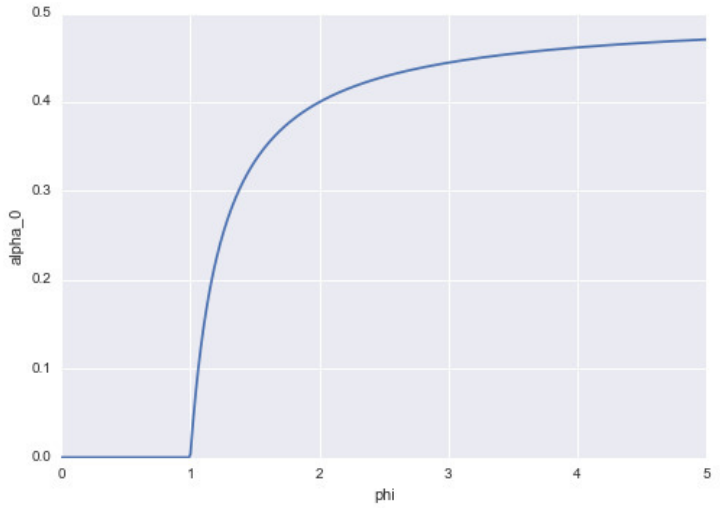

Fig. 4.1: Variational parameter α_0 that minimizes variational free energy F as function of the hyperparameter ϕ.

First, to see the validity of the asymptote for a finite sample size, we compared the asymptote derived in the previous section with one that is numerically calculated by an iterative algorithm, the variational Bayes-EM (VB-EM) algorithm. We set the hyperparameter as a sufficiently large value, $\phi = 20$, and calculated the asymptote in cases in which $n = 200, 400, 800, 1600, 3200, 6400$ cases. To see the variance, we calculated them for 100 different sample sets. The results are shown in figure 4.2. We can see that the asymptote theoretically derived and the numerically calculated result match well as a distribution.

We also calculated the level numerically for a finite sample with the VB-EM algorithm, and compared it with the threshold determined from the asymptote we derived. The procedure is as follows. First, we numerically calculated the variational free energy for many sample sets independently generated from the null hypothesis. After this, we determined the level as the ratio of the number of the sample set whose variational free energy becomes less than the threshold, to the total number of the sample sets.

Through the numerical experiments, the hyperparameter was set as $\phi = 20$ and we calculated the variational free energy for the 5000 sample sets generated from the null hypothesis.

The results are summarized in the table 4.1. The results show that the threshold derived from the asymptote functions correctly. Therefore, we can conclude that the asymptotic form of the variational free energy we derived is valid.

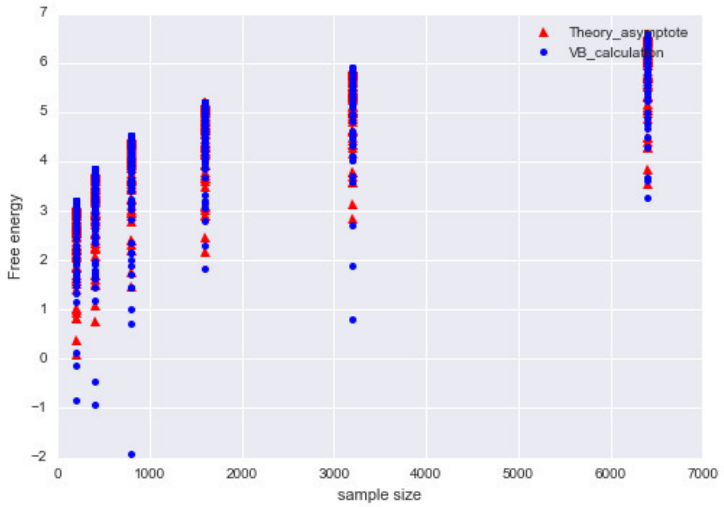

Fig. 4.2: Comparison of variational free energy numerically calculated by VB-EM algorithm with the asymptote theoretically derived, for different sample sizes. Red triangles correspond to the variational free energy calculated from theoretically derived asymptote, and blue circles correspond to the variational free energy numerically calculated.

From the results we have shown, now the hypothesis test of homogeneity based on variational Bayes, which we refer it to as the VB test, can be constructed as follows.

First, calculate the variational free energy from the sample numerically by VB-EM algorithm. In this procedure, the hyperparameter ϕ should be set as greater than one.

Second, test whether the variational free energy is below the threshold or not. The threshold is derived from the asymptote we derived in Section 5. The stochastic behavior of the asymptote is described by the square of the standard normal distribution, and it is easy to calculate the threshold for the rejection rates one needs, by combining the well-known behavior of the χ^2 distribution and the asymptote we derived.

4.6 Discussion

In this chapter, we apply the variational Bayes to the calculation of the marginal likelihood ratio. We construct a new hypothesis test for the homogeneity, the

Table 4.1: The rejection rate calculated numerically by variational Bayes. The threshold is calculated from the asymptote of the variational free energy theoretically derived.

sample size	rejection rates		
n	10%	5%	1%
100	7.1%	3.4%	0.5%
200	8.4%	4.1%	0.8%
400	9.5%	4.5%	0.7%
800	9.6%	5.1%	1.1%

VB test using variational Bayes based on the asymptotic form we theoretically derived. The variational free energy of the normal mixture model showed that the phase transition induced by the hyperparameter ϕ in the prior. As we saw, the phase transition affects the stochastic behavior of the variational free energy, and it is important to discuss the phase transition, when we try to construct a hypothesis test.

The application of variational Bayes for hypothesis tests is not limited to the problem we discussed in this chapter, although specific calculations are needed in each case. As future problems, it would also be interesting to construct hypothesis tests for other singular models, such as Poisson mixture model, Bernoulli mixture model, and so on. Although the distributions of the test statistics for these singular models are not trivial, these mixture models can be rewritten by using latent variables, and we can expect that variational approach also works.

Chapter 5

Asymptotic analysis on upper bounds of the type I and type II errors for singular models

5.1 Introduction

In the previous two chapters, we mainly discussed the asymptotic distribution of the test statistics and constructed the hypothesis test based on them. The performance of the test is validated through numerical experiments in the previous chapters, but to clarify the error in hypothesis testing theoretically is also a fascinating problem from both fundamental and practical perspective.

In conventional theoretical statistics which treats regular models, several celebrated results are obtained, such as Chernoff bound and so on (see e.g. Cover and Thomas [2006]). Therefore, it is natural for us to try to extend these results for treating singular models.

In this section, we theoretically derive the expression of the bound of type I and type II error of the hypothesis testing for singular models, based on Bayesian singular learning theory. The results are clearly different from the conventional result, and these reflect the geometrical properties of singularities.

5.2 Bayesian hypothesis test and the upper bound of the type I and type II errors

Let us briefly summarize the notation, because in this chapter, we mainly treat the theory for general hypothesis test, and the results are not limited to the case of the test of homogeneity for normal mixture.

Let $\{X^n = (X_1, X_2, ..., X_n) \in \mathbb{R}^n\}$ be sample, generated independently and identically.

We assume that the sample is obtained from a probabilistic model $p(x|w)$, where w means the set of the parameters in the model. In the Bayesian framework, parameters w is generated from a prior $\varphi(w)$.

This can be expressed as

$$w \sim \varphi(w), \quad X_i \sim p(x|w).$$

Generally, the null and alternative hypotheses for hypothesis tests are set as,

$$\begin{aligned} \text{N.H.} \ &: \ w \sim \varphi_0(w), \quad X_i \sim p(x|w), \\ \text{A.H.} \ &: \ w \sim \varphi_1(w), \quad X_i \sim p(x|w). \end{aligned}$$

Under these assumption, Bayesian marginal likelihood ratio is defined by

$$L(X^n) = \frac{\int \varphi_1(w) \prod_i p(X_i|w) dw}{\int \varphi_0(w) \prod_i p(X_i|w) dw}. \tag{5.1}$$

As we have already mentioned, in the Bayesian hypothesis test, $L(X^n)$ (or the logarithm of it) is the test statistics which gives the most powerful test. The error can be written as follows,

$$\begin{aligned} \alpha \equiv \Pr(A.H.|N.H.) &= \Pr(L(X^n) \geq \eta | N.H.) \\ \beta \equiv \Pr(N.H.|A.H.) &= \Pr(L(X^n) \leq \eta | A.H.) \end{aligned}$$

To construct the most powerful hypothesis test, we need to derive the stochastic behavior of $L(X^n)$ and clarify the η as a function of the error. Also, it is important to obtain the asymptotic behavior and the (upper) bound of the type I and type II errors as the function of the sample size n.

We studied the former problem, in previous chapters, for constructing the theory of testing homogeneity for normal mixture. In this chapter, we tackle the latter theoretically.

It is easily shown that when we choose a probabilistic variable U and a some constant η, for any $s \geq 0$, the lemma below holds.

Lemma 2. For any probabilistic variable U and any constant η, and any $s \geq 0$, the following holds,

$$\Pr(U \geq \eta) \leq e^{-s\eta} g(s) \tag{5.2}$$

where $g(s) \equiv \langle e^{sU} \rangle$ is the moment generating function of U.

proof.

$$\begin{aligned}
\Pr(U \geq \eta) &= \int_{\eta}^{\infty} p(U)dU \leq \int_{\eta}^{\infty} p(U)e^{s(U-\eta)}dU \\
&\leq \int_{-\infty}^{\infty} p(U)e^{s(U-\eta)}dU \\
&= e^{-s\eta}\int_{-\infty}^{\infty} p(U)e^{sU}dU \\
&= e^{-s\eta}g(s)
\end{aligned}$$

□

Therefore, we obtain the upper bound of $\Pr(U \geq \eta)$ as

$$\Pr(U \geq \eta) \leq \min_{s \geq 0} e^{-s\eta}g(s) \tag{5.3}$$

As well, the upper bound of $\Pr(U \leq \eta)$ can be obtained as,

$$\Pr(U \leq \eta) \leq \min_{s \leq 0} e^{-s\eta}g(s) \tag{5.4}$$

If we consider the case that U is the test statistics of a hypothesis test, the type I and type II errors become,

$$\begin{aligned}
\alpha &= \Pr(U \geq \eta | N.H.) \leq \min_{0 \leq s} \exp\left(-s\eta + \log\langle e^{sU}\rangle_{N.H.}\right) \\
\beta &= \Pr(U \leq \eta | A.H.) \leq \min_{s \leq 0} \exp\left(-s\eta + \log\langle e^{sU}\rangle_{A.H.}\right)
\end{aligned}$$

In the Bayesian likelihood ratio test, we can choose $\log L(X^n)$ as the test statistics, and the moment generating functions of the test statistics under N.H. and A.H. can be defined as,

$$\begin{aligned}
g_0(s) &\equiv \langle e^{sU}\rangle_{N.H.} = \langle (L(X^n))^s\rangle_{N.H.} \\
g_1(s) &\equiv \langle e^{sU}\rangle_{A.H.} = \langle (L(X^n))^s\rangle_{A.H.}
\end{aligned}$$

This gives,

$$\begin{aligned}
g_0(s) &= \int d\{X_i\} \int dw\varphi_0(w) \prod_i p(X_i|w) \left[\frac{\int dw\varphi_1(w)\prod_i p(X_i|w)}{\int dw\varphi_0(w)\prod_i p(X_i|w)}\right]^s \\
&= \int d\{X_i\} \left\{\int dw\varphi_0(w)\prod_i p(X_i|w)\right\}^{1-s} \left\{\int dw\varphi_1(w)\prod_i p(X_i|w)\right\}^s,
\end{aligned}$$

$$
\begin{aligned}
g_1(s) &= \int d\{X_i\} \int dw \varphi_1(w) \prod_i p(X_i|w) \left[\frac{\int dw \varphi_1(w) \prod_i p(X_i|w)}{\int dw \varphi_0(w) \prod_i p(X_i|w)} \right]^s \\
&= \int d\{X_i\} \left\{ \int dw \varphi_1(w) \prod_i p(X_i|w) \right\}^{1+s} \left\{ \int dw \varphi_0(w) \prod_i p(X_i|w) \right\}^{-s}.
\end{aligned}
$$

Therefore, for $-1 \leq s \leq 0$, $g_0(s)$ and $g_1(s)$ satisfies,

$$
g_1(s) = g_0(1 - |s|). \tag{5.5}
$$

This enables for us to rewrite bound for β as,

$$
\begin{aligned}
\beta &\leq \min_{-1 \leq r \leq 0} \exp\left[-r\eta + \log g_1(r) \right] \\
&= \min_{-1 \leq r \leq 0} \exp\left[-r\eta + \log g_0(1 - |r|) \right] \\
&= \min_{0 \leq s \leq 1} \exp\left[s\eta + \log g_0(1 - s) \right] \\
&= \min_{0 \leq q \leq 1} \exp\left[(1 - q)\eta + \log g_0(q) \right].
\end{aligned}
$$

This shows that it is sufficient to obtain the behavior of $g_0(s)$ for deriving the bound of the type I and type II errors.

5.3 The bound of the type I and type II errors for singular models

In the previous section, we review the well-known results on the upper bound of the type I and type II errors, and extend them for the Bayesian hypothesis test.

We saw that the bound is determined from the asymptotic behavior of the expectation value of the moment generating function for log marginal likelihood ratio, $\langle g_0(s) \rangle$. The leading term of $\langle g_0(s) \rangle$ is expected to become $\mathcal{O}(n)$ for the case the $N.H.$ and $A.H.$ which are well separated, and the asymptote of each error is often described as,

$$
\lim_{n \to \infty} \beta = -\frac{1}{n} \log \langle g_0(s) \rangle. \tag{5.6}
$$

The result above is well-known, but as we will see in this section, when the hypotheses we compare are described as probabilistic distributions and they are sufficiently close in the some region of their support in the parameter space, this

is not always the case. Such a situation is often seen in Bayesian hypothesis test, and it is important to clarify the bound of type I and II error in such cases.

Let us consider the case the null hypothesis is given as a point hypothesis in the parameter space, such as

$$\varphi_0(w) = \delta(w - w_0) \tag{5.7}$$

where w_0 is the singularity in the parameter space.

What we should is to calculate $\log g_0(s)$,

$$\begin{aligned} g_0(s) &= \int d\{X_i\} \prod_i p(X_i|w_0)^{1-s} \left\{ \int dw \varphi_1(w) \prod_i p(X_i|w) \right\}^s \\ &\equiv \left\langle \left\{ \frac{\int dw \varphi_1(w) \prod_i p(X_i|w)}{\prod_i p(X_i|w_0)} \right\}^s \right\rangle. \end{aligned}$$

It is known that when $n \to \infty$, the log likelihood ratio can be approximated as follows, from the large-deviation theory,

$$\lim_{n\to\infty} -\frac{1}{n} \log \left\{ \frac{\int dw \varphi_1(w) \prod_i p(X_i|w)}{\prod_i p(X_i|w_0)} \right\} = D(p(x|w^*)||p(x|w_0)), \tag{5.8}$$

where $D(\cdot||\cdot)$ means the Kullback-Leibler divergence,
and $w^* \equiv \arg\min_w D(p(x|w)||p(x|w_0))$.

This leads the following result.

Theorem 8. Let the null and alternative hypotheses for hypothesis tests are set as,

$$\begin{aligned} \text{N.H.} &: w \sim \varphi_0(w) = \delta(w - w_0), \\ \text{A.H.} &: w \sim \varphi_1(w). \end{aligned}$$

We assume that the $\varphi_1(w)$ has no support on the w_0 but under A.H., the $w^* \equiv \arg\min_w D(p(x|w)||p(x|w_0))$ is uniquely determined.

Then, for a fixed constant $\epsilon > 0$ and a threshold η for a hypothesis testing, the upper bounds of type I and type II errors become,

$$\begin{aligned} \alpha &\leq \exp\left[-(1-\epsilon)\eta - n(1-\epsilon)D(p(x|w^*)||p(x|w_0)) + o(n)\right], \\ \beta &\leq \exp\left[\epsilon\eta - n(1-\epsilon)D(p(x|w^*)||p(x|w_0)) + o(n)\right], \end{aligned}$$

proof. This result is readily led from (5.6) and (5.8). We should note that at $s = 1$, g_0 becomes definitely one. This is why we introduce a constant $\epsilon > 0$. □

However, this is not the case in general, when the null hypothesis $p(x|w_0)$ corresponds to the singularity in the parameter space of the probabilistic model we adopt $p(x|w)$ and the alternative hypothesis has its support on the singularity. Under such a circumstance, the parameter set w^* has no longer one-to-one correspondence to the w_0 (Watanabe [2018]).

This degeneracy causes singular behavior of the log marginal likelihood ratio, and we have to take this effect into consideration to clarify the statistical properties of the hypothesis test when using such singular models for hypothesis test.

We note that such a singular behavior is not seen only in special cases. Actually, it is common in various probabilistic models which are practically used in various fields, such as mixture models, the hidden Markov model, Bayesian network, neural network and so on.

The following holds.

Theorem 9. Let the null and alternative hypotheses for hypothesis tests are set as,

$$\begin{aligned} \text{N.H.} \quad &: \quad w \sim \varphi_0(w) = \delta(w - w_0), \\ \text{A.H.} \quad &: \quad w \sim \varphi_1(w). \end{aligned}$$

Here, we assume that the $\varphi_1(w)$ has support on the w which gives $p(x|w_0)$. However, we consider the case that the w which gives $p(x|w_0)$ is *not* uniquely determined.

Then, for a fixed constant $\epsilon > 0$ and a threshold η for a hypothesis testing, the upper bounds of type I and type II errors become,

$$\begin{aligned} \alpha &\leq \left\langle \exp\left[-(1-\epsilon)\eta - (1-\epsilon)(\lambda \log n - (m-1)\log\log n + \mathcal{O}(1))\right] \right\rangle \\ \beta &\leq \left\langle \exp\left[\epsilon\eta - (1-\epsilon)(\lambda \log n - (m-1)\log\log n + \mathcal{O}(1))\right] \right\rangle \end{aligned}$$

where λ is given as the coefficient of the leading $\mathcal{O}(\log n)$ term in the asymptote of the difference between the logarithm of the marginal likelihood ratio as,

$$\log\left\{\frac{\int dw\varphi_1(w)\prod_i p(X_i|w)}{\prod_i p(X_i|w_0)}\right\} \equiv F_n - nL_n(w_0)$$

where

$$F_n \equiv -\log\int dw\varphi_1(w)\prod_i p(X_i|w),$$

and

$$L_n(w_0) \equiv -\frac{1}{n}\sum_i p(X_i|w_0).$$

proof. We can rewrite g_0 in terms of these quantities as,

$$g_0(s) = \left\langle e^{-s(F_n - nL_n)} \right\rangle \tag{5.9}$$

From the result of Bayesian singular learning theory, the asymptotic form of F_n is obtained, which is valid even for the situation where the null hypothesis is in the singularity in the parameter space Watanabe [2018].

$$F_n = nL_n(w_0) + \lambda \log n - (m-1) \log \log n + \mathcal{O}(1), \tag{5.10}$$

where λ is a positive constant which is known as "real log canonical threshold" in Bayesian singular learning theory, and $m \geq 1$ is a some natural number which is known as "multiplicity" in Bayesian singular learning theory. It was shown that stochastic term in F_n is $\mathcal{O}(1)$.

The asymptote of g_0 becomes

$$g_0 = \left\langle e^{-s(\lambda \log n - (m-1) \log \log n + \mathcal{O}(1))} \right\rangle. \tag{5.11}$$

Therefore, we can construct the bound of the type I and type II errors as ,

$$\begin{aligned}
\alpha &\leq \min_{0 \leq s \leq 1} \left\langle \exp\left[-s\eta - s(\lambda \log n - (m-1) \log \log n + \mathcal{O}(1))\right] \right\rangle \\
\beta &\leq \min_{0 \leq s \leq 1} \left\langle \exp\left[(1-s)\eta - s(\lambda \log n - (m-1) \log \log n + \mathcal{O}(1))\right] \right\rangle
\end{aligned}$$

Under the $n \to \infty$, the minimum of (r.h.s)s are achieved when $s = 1$, and the asymptote of the these error becomes,

$$\begin{aligned}
\alpha &\leq \frac{1}{n^{\lambda}}, \\
\beta &\leq \frac{1}{n^{\lambda}},
\end{aligned}$$

and the theorem is proofed. □

5.4 Example - The bound of the type I and type II errors for normal mixture

To proceed further, let us consider a specific probabilistic model.

As well as other sections in this thesis, we choose the two-component normal mixture as a probabilistic model $p(x|w)$ and discuss the test of homogeneity for this model.

$$p(x|w) = \frac{1}{\sqrt{2\pi}} \left[(1-a)e^{-x^2/2} + ae^{-(x-b)^2/2} \right] \tag{5.12}$$

We choose N.H. and A.H. as follows,

$$\begin{aligned} \text{N.H.} &: w \sim \varphi_0(w) = \delta(a)\delta(b) \\ \text{A.H.} &: w \sim \varphi_1(w) = U_{a(0,1)} \times U_b(0,1) \end{aligned}$$

It can be seen that the distribution given by N.H. can be given is degenerated in the parameter space given by A. H.

In other words, in our normal mixture, the arbitrary parameter sets $(a.b)$ which satisfies $ab = 0$ gives the distribution given by N. H. .

As shown in section 3, the asymptotic form of the log marginal likelihood ratio becomes when the prior φ_1 contains the singularity,

$$F_n = \frac{1}{2}\log n - \log\log n - \log \int_0^\infty dt t^{-1/2} \log(1/t) \exp(-t/2) \cosh(\sqrt{t}\xi) \tag{5.13}$$

where $\xi \equiv \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_i X_i \sim \mathcal{N}(0,1)$.

It can be seen that the $\lambda = \frac{1}{2}$ in this model, and as a result,

$$\begin{aligned} \alpha &\leq \frac{1}{\sqrt{n}}, \\ \beta &\leq \frac{1}{\sqrt{n}}, \end{aligned}$$

5.5 Discussion

In this chapter, we discuss the bound of type I and type II error in the Bayesian hypothesis testing. We show that the bound can be given by using real log canonical threshold, the quantity that the geometrical properties of the singularities.

Let us remind the form of the regular models (see Introduction in the thesis). In the regular models, the coefficient of the logarithm term in the free energy is $\frac{d}{2}$. Therefore, in the singular models, the bound is changed from $\frac{d}{2}$ to λ. λ depends on each models, but in the present case, it is smaller than $\frac{d}{2}$.

When the N.H. locates on the singularity, if the support of the A.H. is separated from the singularity sufficiently, the leading term of the $\log L$ becomes $nK_n(w_{min})$, where w_{min} means the parameter that gives the minimum of Kullback-Leibler divergence K_{min} between $p(x|w)$ under A.H. and N.H..

As long as nK_n is sufficiently larger than $\lambda \log n$, the leading term of $\log L$ becomes nK_n. This can be regarded as a regular case. In this case, the bound of the error can be characterized by form $e^{-nK(n)}$.

However, nK_n is not sufficiently larger than $\lambda \log n$, this case is not necessarily regarded as a regular case, because the leading term is not clear. Therefore, the crossover is expected to occur as a result of the change of the leading term of $logL$ and the behavior of the bound of the error can be changed, from $e^{-nK(n)}$ to $\frac{1}{n^\lambda}$.

Chapter 6

Conclusion

6.1 Summary

In this thesis, we studied Bayesian hypothesis testing for singular models theoretically. The summary of the results we obtained is as follows.

- We theoretically studied the test of homogeneity for normal mixtures in terms of the Bayesian framework (Chapter 3).

  By applying the mathematical technique developed for the analysis of singular models, we derived the asymptotic behavior of the marginal likelihood ratio for several types of the alternative hypotheses. Our analysis is based on the scaling technique that is based on the Kullback-Leibler divergence between the null hypothesis and the alternative hypothesis, and it enables us to derive the asymptote systematically and easily. We also compare our method and Bayes factor-based testing and discussing the validity of them in the delicate setting.

- We apply the variational Bayes to the calculation of the marginal likelihood ratio for the first time (Chapter 4).

  We constructed a new hypothesis test for the homogeneity, the VB test which is based on variational Bayes. In VB test, the test statistics is the variational free energy. We derive the asymptotic form of it and construct the hypothesis test. We also clarified that the variational free energy shows the phase transition induced by the hyperparameter ϕ in the prior. As we saw, the phase transition affects the stochastic behavior of the variational free energy, and it is important to discuss the phase transition when we try to construct a hypothesis test.

- We discuss the bound of type I and type II error in the Bayesian hypothesis testing and analytically derived the bound in the singular case (Chapter 5) .

  We show that the bound can be given by using real log canonical threshold (RCLT), the quantity that the geometrical properties of the singularities. We also discuss the necessity for the bound of the singular case, that is different from the regular case, and the possibility of the crossover behavior of the bound.

6.2 Future problems

Through the series of the present study, we believe that the theoretical understanding of the Bayesian hypothesis testing for singular models is deepen to some extent. However, this research area is relatively new, as we mentioned in the Introduction, and there is still much remained that should be studied further. Let us show several examples.

- Constructing theory of the Bayesian hypothesis testing for another singular model

  As we mentioned in the Introduction, in the case of the statistical inference, the properties for each singular model are studied because the quantities e.g. RCLT and so on, which determine the generalization error, are peculiar to each model. Therefore, from the practical perspective, the theoretical understandings of the generalization error for each well-known singular model is required. The same is true for the problem of hypothesis testing. In this thesis, we mainly treated normal mixture, but we think other singular models should be studied in the future, e.g. various mixture models such as Poisson mixture, Bernoulli mixture, and so on. We should also note that although the distributions of the test statistics for these singular models are not trivial, these mixture models can be rewritten by using latent variables, and we can expect that the variational approach also works.

- Application of our theory for another research area such as (theoretical) physics

  In this thesis, we thoroughly treated the two-component normal mixture and discuss the theoretical properties of hypothesis testing. From practical perspective, it may seem too simple and may seem that these results have very limited applications. We admit that the setup we treated is relatively simple. However, we believe that the results shown here are meaningful from two perspectives.

First, these results can become a basis for the analysis of more complex models. Second, this is what we want to state here, a similar setup are studied extensively in another research area recently. More specifically, in theoretical physics (especially optics and quantum physics), the seek for high resolution of the signals from two optical sources is a very hot topic(see e.g. Tsang et al. [2016], Lu et al. [2018]). Especially, in recent years, the statistical treatment based on the calculation of the (quantum) Fisher information is paid attention. However, as we saw in the present thesis, these problems tend to become singular cases although the existing studies treat such problems as if they are regular, and the formalism based on the Fisher information does not work well. We hope that our treatment for the singular models may shed new light on such problems.

Bibliography

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in bayesian estimation. Neural Networks, 18(7):924–933, 2005. doi:https://doi.org/10.1016/j.neunet.2005.03.014.

Michael Francis Atiyah. Resolution of singularities and division of distributions. Communications on pure and applied mathematics, 23(2):145–150, 1970. doi:https://doi.org/10.1002/cpa.3160230202.

Hagai Attias. A variational baysian framework for graphical models. In Advances in neural information processing systems, volume 12, pages 209–215, 2000. URL https://proceedings.neurips.cc/paper/1999/file/74563ba21a90da13dacf2a73e3ddefa7-Paper.pdf.

N. Balakrishnan Barry C. Arnold and H. N. Nagaraja. A First Course in Order Statistics (Classics in Applied Mathematics). Society for Industrial and Applied Mathematics, 2008.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112 (518):859–877, 2017. doi:https://doi.org/10.1080/01621459.2017.1285773.

Richard Charnigo and Jiayang Sun. Testing homogeneity in a mixture distribution via the l2 distance between competing models. Journal of the American Statistical Association, 99(466):488–498, 2004. doi:https://doi.org/10.1198/016214504000000494.

D. Chauveau, B. Garel, and S. Mercier. Testing for univariate two-component Gaussian mixture in practice. Journal of the French Statistical Society, 160:86–113, 2019. URL http://journal-sfds.fr/article/view/729/775.

Hanfeng Chen, Jiahua Chen, and John D. Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. Journal of the Royal

Statistical Society. Series B (Statistical Methodology), 63(1):19–29, 2001. doi:https://doi.org/10.1111/1467-9868.00273.

Hanfeng Chen, Jiahua Chen, and John D. Kalbfleisch. Testing for a finite mixture model with two components. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(1):95–115, 2004. doi:10.1111/j.1467-9868.2004.00434.x.

Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The EM approach. The Annals of Statistics, 37(5A):2523–2542, 2009. doi:https://doi.org/10.1214/08-AOS651.

Jiahua Chen, Pengfei Li, and Yuejiao Fu. Inference on the order of a normal mixture. Journal of the American Statistical Association, 107(499):1096–1105, 2012. doi:10.1080/01621459.2012.695668.

Thomas M. Cover and Joy A. Thomas. Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, July 2006. ISBN 0471241954.

B. S. Everitt, S. Landau, and M. Leese. Cluster Analysis. 4th Edition. Arnold, London., 2001. ISBN 978-0-340-76119-9.

Bernard Garel. Likelihood ratio test for univariate gaussian mixture. Journal of Statistical Planning and Inference, 96(2):325 – 350, 2001. doi:https://doi.org/10.1016/S0378-3758(00)00216-0.

J. Ghosh and P. Sen. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 1985, 2:789–806, 1985. URL http://www.lib.ncsu.edu/resolver/1840.4/3493.

J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 1985, 2:807–810, 1985.

Naoki Hayashi and Sumio Watanabe. Upper bound of bayesian generalization error in non-negative matrix factorization. Neurocomputing, 266:21–28, 2017. doi:https://doi.org/10.1016/j.neucom.2017.04.068.

Natsuki Kariya and Sumio Watanabe. Asymptotic analysis of singular likelihood ratio of normal mixture by bayesian learning theory for testing homogeneity. Communications in Statistics - Theory and Methods, 49, 2020a. doi:https://doi.org/10.1080/03610926.2020.1849721.

Natsuki Kariya and Sumio Watanabe. Testing homogeneity for normal mixture models: Variational bayes approach. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 103 (11):1274–1282, 2020b. doi:https://doi.org/10.1587/transfun.2019EAP1172.

Richard R Lewine. Sex differences in schizophrenia: Timing or subtypes? Psychological Bulletin, 90(3):432, 1981. doi:https://doi.org/10.1037/0033-2909.90.3.432.

S. Li, J. Chen, and P. Li. Mixtureinf: Inference for finite mixture models., 2016. URL https://cran.r-project.org/web/packages/MixtureInf/index.html.

Xin Liu and Yongzhao Shao. Asymptotics for likelihood ratio tests under loss of identifiability. The Annals of Statistics, 31(3):807–832, 06 2003. doi:10.1214/aos/1056562463.

Xin Liu and Yongzhao Shao. Asymptotics for the likelihood ratio test in a two-component normal mixture model. Journal of Statistical Planning and Inference, 123(1):61 – 81, 2004. doi:https://doi.org/10.1016/S0378-3758(03)00138-1.

Xiao-Ming Lu, Hari Krovi, Ranjith Nair, Saikat Guha, and Jeffrey H Shapiro. Quantum-optimal detection of one-versus-two incoherent optical sources with arbitrary separation. npj Quantum Information, 4(1):1–8, 2018. doi:https://doi.org/10.1038/s41534-018-0114-y.

G. J. McLachlan and D. Peel. Finite mixture models. Wiley Series in Probability and Statistics, New York, 2000. ISBN 978-0471721185.

Mankei Tsang, Ranjith Nair, and Xiao-Ming Lu. Quantum theory of superresolution for two incoherent optical point sources. Physical Review X, 6(3):031033, 2016. doi:https://doi.org/10.1103/PhysRevX.6.031033.

Kazuho Watanabe and Sumio Watanabe. Stochastic complexities of gaussian mixtures in variational bayesian approximation. Journal of Machine Learning Research, 7(Apr):625–644, 2006. URL http://jmlr.org/papers/v7/watanabe06a.html.

Kazuho Watanabe and Sumio Watanabe. Stochastic complexities of general mixture models in variational bayesian learning. Neural Networks, 20(2):210–219, 2007. doi:https://doi.org/10.1016/j.neunet.2006.05.030.

Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. Neural Computation, 13(4):899–933, 2001. doi:10.1162/089976601300014402.

Sumio Watanabe. Mathematical Theory of Bayesian Statistics. Chapman and Hall/CRC, New York, 2018. ISBN 9781482238068.

Sumio Watanabe and Shun-ichi Amari. Learning coefficients of layered models when the true distribution mismatches the singularities. Neural Computation, 15(5):1013–1033, 2003. doi:10.1162/089976603765202640.

Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. Neural networks, 16(7):1029–1038, 2003. doi:https://doi.org/10.1016/S0893-6080(03)00005-4.

大橋耕也 and 渡辺澄夫. 変化点検出問題におけるベイズ検定統計量の導出と検出力の実験的考察. 研究報告数理モデル化と問題解決 (MPS), 2017(11): 1–6, 2017.

竹村彰通. 新装改訂版　現代数理統計学. 学術図書出版社, 2020. ISBN 78-4-7806-0860-1.

藤原香織 and 渡辺澄夫. 特異モデルにおけるベイズ検定と時系列解析への応用. 電子情報通信学会論文誌. D, 情報・システム = The IEICE transactions on information and systems (Japanese edition), 91(4):889–896, apr 2008. URL https://ci.nii.ac.jp/naid/110007381036/.