Mach Learn (2016) 102:1–28
DOI 10.1007/s10994-015-5482-3

Asymptotic accuracy of Bayes estimation for latent variables with redundancy

Keisuke Yamazaki

Received: 8 May 2013 / Accepted: 19 January 2015 / Published online: 11 February 2015

Abstract Hierarchical parametric models consisting of observable and latent variables are widely used for unsupervised learning tasks. For example, a mixture model is a representative hierarchical model for clustering. From the statistical point of view, the models can be regular or singular due to the distribution of data. In the regular case, the models have the identifiability; there is one-to-one relation between a probability density function for the model expression and the parameter. The Fisher information matrix is positive definite, and the estimation accuracy of both observable and latent variables has been studied. In the singular case, on the other hand, the models are not identifiable and the Fisher matrix is not positive definite. Conventional statistical analysis based on the inverse Fisher matrix is not applicable. Recently, an algebraic geometrical analysis has been developed and is used to elucidate the Bayes estimation of observable variables. The present paper applies this analysis to latent-variable estimation and determines its theoretical performance. Our results clarify behavior of the convergence of the posterior distribution. It is found that the posterior of the observable-variable estimation can be different from the one in the latent-variable estimation. Because of the difference, the Markov chain Monte Carlo method based on the parameter and the latent variable cannot construct the desired posterior distribution.

Keywords Unsupervised learning · Hierarchical parametric models · Bayes statistics · Algebraic geometry · Singularities

Editor: Peter Flach.

This research was partially supported by the Kayamori Foundation of Informational Science Advancement and KAKENHI 23500172 and 24700139.

K. Yamazaki (✉)
Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology,
G5-19 4259 Nagatsuta, Midori-ku, Yokohama, Japan
e-mail: k-yam@math.dis.titech.ac.jp

1 Introduction

Hierarchical parametric models are employed for unsupervised learning in many data-mining and machine-learning applications. Statistical analysis of the models plays an important role for not only revealing the theoretical properties but also the practical applications. For example, the asymptotic forms of the generalization error and the marginal likelihood are used for model selection in the maximum-likelihood and Bayes methods, respectively (Akaike 1974; Schwarz 1978; Rissanen 1986).

Parametric models generally fall into two cases: regular and singular. The present paper focuses on the models, the function of which are continuous and sufficiently smooth with respect to the parameter. In regular cases, the Fisher information matrix is positive definite, and there is a one-to-one relation between the parameter and the expression of the model as a probability density function. Otherwise, the model is singular, and the parameter space includes singularities. Due to these singularities, the Fisher information matrix is not positive definite, and so the conventional analysis methods that rely on its inverse matrix are not applicable. In this case, an algebraic geometrical approach can be used to analyze the Bayes method (Watanabe 2001, 2009).

Hierarchical models have both observable and latent variables. The latent variables represent the underlying structure of the model, while the observable ones correspond to the given data. For example, unobservable labels in clustering are expressed as the latent variables in mixture models, and the system dynamics of time-series data is a sequence of the variables in hidden Markov models. Hierarchical models thus have two estimation targets: observable and latent variables. The well-known generalization error measures the performance of the prediction of a future observable variable. Combining the two model cases and the two estimation targets, there are four estimation cases, which are summarized in Table 1. We will use the abbreviations shown in the table to specify the target variable and the model case; for example, Reg-OV estimation stands for estimation of the observable variable in the regular case.

In the present paper, we will investigate the asymptotic performance of the Sing-LV estimation. One of the main concerns in unsupervised learning is the estimation of unobservable parts and in practical situations, the ranges of the latent variables are unknown, which corresponds to the singular case. The other estimation cases have already been studied; the accuracy of the Reg-OV estimation has been clarified on the basis of the conventional analysis method, and the results have been used for model selection criteria, such as AIC (Akaike 1974). The primary purpose for using the algebraic geometrical method is to analyze the Sing-OV estimation, and the asymptotic generalization error of the Bayes method has been derived for many models (Aoyagi and Watanabe 2005; Aoyagi 2010; Rusakov and Geiger 2005; Yamazaki and Watanabe 2003a, b, 2005a, b; Zwiernik 2011). Recently, an error function for the latent-variable estimation was formalized in a distribution-based manner, and its asymptotic form was determined for the Reg-LV estimation of both the maximum likelihood and Bayes methods (Yamazaki 2014). Hereinafter, the estimation method will be assumed to be the Bayes method unless it is explicitly stated otherwise.

Table 1 Estimation classification according to the target variable and the model case

Estimation target \model case	Regular case	Singular case
Observable variable	Reg-OV estimation	Sing-OV estimation
Latent variable	Reg-LV estimation	Sing-LV estimation

In the Bayes estimation, parameter sampling from the posterior distribution is an important process for practical applications. The behavior of posterior distributions has been studied in the statistical literature. The convergence rate of the posterior distribution has been analyzed (e.g., Ghosal et al. 2000; Le Cam 1973; Ibragimov and Has' Minskii 1981). Specifically, the rate based on the Wasserstein metrics is elucidated in finite and infinite mixture models (Nguyen 2013). To avoid singularities, conditions for the identifiability guaranteeing the positive Fisher matrix are necessary. Allman et al. (2009) use algebraic techniques to clarify the identifiability in some hierarchical models. In the regular case, the posterior distribution has the asymptotic normality, which means that it converges to a Gaussian distribution. Because the variance of the distribution goes to zero when the number of data is sufficiently large, the limit distribution is the delta distribution. Then, the sample sequence from the posterior distribution converges to a point. On the other hand, in the singular case, the posterior distribution does not have the asymptotic normality and the sequence converges to some area of the parameter space (Watanabe 2001). Studies on the Sing-OV estimation such as Yamazaki and Kaji (2013) have shown that the convergence area of the limit distribution depends on a prior distribution. The behavior of the posterior distribution has not been clarified in the Sing-LV estimation. The analysis of the present paper enables us to elucidate the relation between the prior and the limit posterior distributions.

The main contributions of the present paper are summarized as follows:

1. The algebraic geometrical method for the Sing-OV estimation is applicable to the analysis of the Sing-LV estimation.
2. The asymptotic form of the error function is obtained, and its dominant order is larger than that of the Reg-LV estimation.
3. There is a case, where the limit posterior distribution in the Sing-LV estimation is different from that in the Sing-OV estimation.

The third result is important for practical applications: in some priors, parameter-sampling methods based on latent variables, such as Gibbs sampling in the Markov chain Monte Carlo (MCMC) method, cannot construct the proper posterior distribution because the sample sequence of the MCMC method follows the posterior of the Sing-LV estimation, which has a different convergence area from the desired one in the Sing-OV estimation.

The rest of this paper is organized as follows. The next section formalizes the hierarchical model and the singular case, and introduces the performance of the Reg-OV and the Sing-OV estimations. Section 3 explains the asymptotic analysis of the free energy function and the convergence of the posterior distribution based on the results of the Sing-OV estimation. In Sect. 4, the latent-variable estimation and its evaluation function are formulated in a distribution-based manner. Section 5 shows the main results: the asymptotic error function of general hierarchical models, and the detailed error properties in mixture models. In Sect. 6, we discuss the limit distribution of the posterior in the Sing-LV estimation and differences from the Sing-OV estimation. Finally, Sect. 7 presents conclusions.

2 The singular case and accuracy of the observable-variable estimation

In this section, we introduce the singular case and formalize the Bayes method for the observable-variable estimation. This section is a brief summary of the results on the Reg-OV and the Sing-OV estimations.

2.1 Hierarchical models and singularities

Let a learning model be defined by

$$p(x|w) = \sum_{y=1}^{K} p(x, y|w) = \sum_{y=1}^{K} p(y|w)p(x|y, w),$$

where $x \in R^M$ is an observable variable, $y \in \{1, \dots, K\}$ is a latent one, and $w \in W \subset R^d$ is a parameter. For the discrete x such that $x \in \{1, 2, \dots, M\}$, all results hold by replacing $\int dx$ with $\sum_{x=1}^{M}$.

Example 1 A mixture of distributions is described by

$$p(x|w) = \sum_{k=1}^{K} a_k f(x|b_k), \tag{1}$$

where f is the density function associated with a mixture component, which is identifiable for any $b_k \in W_b \subset R^{d_c}$. The mixing ratios have constraints $a_k \geq 0$ and $\sum_{k=1}^{K} a_k = 1$. We regard a_1 as a function of the parameters $a_1 = 1 - \sum_{k=2}^{K} a_k$. The parameter w consists of $\{a_2, \dots, a_K\}$ and $\{b_1, \dots, b_K\}$, where $w \in \{[0, 1]^{K-1}, W_b^{Kd_c}\}$. The latent variable y is the component label.

Assume that the number of data is n and the observable data $X^n = \{x_1, \dots, x_n\}$ are independent and identically distributed from the true model, which is expressed as

$$q(x) = \sum_{y=1}^{K^*} q(y)q(x|y).$$

Note that the value range of the latent variable y described as $[1, \dots, K^*]$ is generally unknown and can be different from the one in the learning model. In the example of the mixture model, the true model is expressed as

$$q(x) = \sum_{k=1}^{K^*} a_k^* f(x|b_k^*). \tag{2}$$

We also assume that the true model satisfies the minimality condition:

$$k \neq j \in \{1, \dots, K^*\} \Rightarrow q(x|y = k) \neq q(x|y = j).$$

For example, consider a three-component model such that $q(x|y = 1) \neq q(x|y = 2) = q(x|y = 3)$. This model does not satisfy the minimality condition. Defining a new label, we obtain the following two-component expression, which satisfies the condition;

$$\begin{aligned} q(x) &= q(y = 1)q(x|y = 1) + \{q(y = 2) + q(y = 3)\}q(x|y = 2) \\ &= q(y = 1)q(x|y = 1) + q(y = \bar{2})q(x|y = \bar{2}), \end{aligned}$$

where $y \in \{1, \bar{2}\}$ and $\bar{2} = \{2, 3\}$.

The present paper focuses on the case in which the true model is in the class of the learning model. More formally, there is a set of parameters expressing the true model such that

$$W_X^t = \{w^*; p(x|w^*) = q(x)\} \neq \emptyset,$$

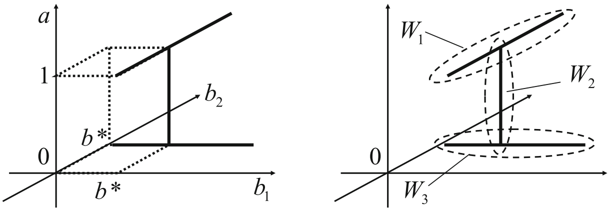

Fig. 1 The true parameter set W_X^t (the *left panel*), and the parameter areas W_1, W_2, and W_3 (the *right panel*)

which is referred to as the true parameter set for x. This means that the latent variable range satisfies $K = K^*$ or $K > K^*$. The former relation corresponds to the regular case and the latter one to the singular case. The true parameter set W_X^t includes $K!$ isolated points in the regular case due to the symmetry of the parameter space. On the other hand, it consists of an analytic set in the singular case. We explain this structure using the following model settings.

Example 2 Assume that $K = 2$ and $K^* = 1$ in the mixture model. For illustrative purposes, let the learning and the true models be defined by

$$\begin{aligned} p(x|w) &= af(x|b_1) + (1-a)f(x|b_2), \\ q(x) &= f(x|b^*), \end{aligned}$$

respectively, where $x \in R^1$ and $w = \{a, b_1, b_2\}$ such that $a \in [0, 1]$ and $b_1, b_2 \in W_b \subset R^1$. We can confirm that the true parameter set consists of the following analytic set:

$$\begin{aligned} W_X^t &= W_1^t \cup W_2^t \cup W_3^t, \\ W_1^t &= \{a = 1, b_1 = b^*\}, \\ W_2^t &= \{b_1 = b_2 = b^*\}, \\ W_3^t &= \{a = 0, b_2 = b^*\}. \end{aligned}$$

As shown in Fig. 1, let W_1, W_2, and W_3 be the neighborhood of W_1^t, W_2^t, and W_3^t, respectively. The Fisher information matrix is not positive definite in W_X^t. Moreover, the intersections of W_1^t, W_2^t and W_3^t are singularities.

When $K = K^*$, W_X^t is a set of points, which corresponds to the regular case;

Example 3 If both the learning and the true models have two components,

$$\begin{aligned} p(x|w) &= af(x|b_1) + (1-a)f(x|b_2), \\ q(x) &= a^*f(x|b_1^*) + (1-a^*)f(x|b_2^*) \end{aligned}$$

for $a^* \neq 0, 1$ and $b_1^* \neq b_2^*$, the estimation will be in the regular case. Due to $K! = 2! = 2$, the set consists of two isolated points;

$$W_X^t = \{(a = a^*, b_1 = b_1^*, b_2 = b_2^*), (a = 1 - a^*, b_1 = b_2^*, b_2 = b_1^*)\},$$

where the Fisher information matrix is positive definite.

2.2 The observable-variable estimation and its performance

In Bayesian statistics, estimation of the observable variables is defined by

$$p(x|X^n) = \int p(x|w)p(w|X^n)dw,$$

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

where $\varphi(w;\eta)$ is a prior distribution with the hyperparameter η, $p(w|X^n)$ is the posterior distribution of the parameter, and its normalizing factor is given by

$$Z(X^n) = \int \prod_{i=1}^{n} p(x_i|w)\varphi(w;\eta)dw.$$

This formulation is available for both the Reg-OV and Sing-OV estimations. In the mixture model, the Dirichlet distribution is often used for the prior distribution of the mixing ratio;

$$\varphi(w;\eta) = \varphi(a;\eta_1)\varphi(b;\eta_2), \tag{3}$$

$$\varphi(a;\eta_1) = \frac{\Gamma(K\eta_1)}{\Gamma(\eta_1)^K}\prod_{i=k}^{K} a_k^{\eta_1-1}, \tag{4}$$

where $a = \{a_1,\ldots,a_K\}$, $b = \{b_1,\ldots,b_K\}$, $\eta = \{\eta_1,\eta_2\} \in R^2_{>0}$, and Γ is the gamma function. Since a_k has the same exponential part for all k, $\varphi(a;\eta_1)$ is referred to as a symmetric Dirichlet distribution.

The estimation accuracy is measured by the average Kullback–Leibler divergence:

$$G(n) = E_X\left[\int q(x)\ln\frac{q(x)}{p(x|X^n)}dx\right],$$

where the expectation is

$$E_X\left[f(X^n)\right] = \int f(X^n)q(X^n)dX^n.$$

Let us define the free energy as

$$F(X^n) = -\ln Z(X^n),$$

which plays an important role in Bayes statistics as a criterion for selecting the optimal model. In the Reg-OV estimation, the Bayesian information criterion (BIC; Schwarz 1978) and the minimum-description-length principle (MDL; Rissanen 1986) are both based on the asymptotic form of $F(X^n)$. Theoretical studies often analyze the average free energy given by

$$F_X(n) = -nS_X + E_X[F(X^n)],$$

where the entropy function is defined by

$$S_X = -\int q(x)\ln q(x)dx.$$

The model that minimizes $F(X^n)$ is then selected as optimal from among the candidate models. The energy function $F_X(n)$ allows us to investigate the average behavior of the selection. Note that the entropy term does not affect the selection result because it is independent of the candidate models. According to the definitions, the average free energy and the generalization error have the relation

$$\begin{aligned}
G(n) &= E_{X^n}\left[\int q(x_{n+1}) \ln \frac{q(x_{n+1})}{p(x_{n+1}|X^n)} dx_{n+1}\right] \\
&= E_{X^n, x_{n+1}}\left[\ln \frac{q(x_{n+1})}{p(x_{n+1}|X^n)}\right] \\
&= E_{X^n, x_{n+1}}\left[\ln \frac{\prod_{i=1}^{n+1} q(x_i)}{\int \prod_{i=1}^{n+1} p(x_i|w)\varphi(w;\eta)dw}\right] \\
&\quad - E_{X^n}\left[\ln \frac{\prod_{i=1}^{n} q(x_i)}{\int \prod_{i=1}^{n} p(x_i|w)\varphi(w;\eta)dw}\right] \\
&= F_X(n+1) - F_X(n), \qquad (5)
\end{aligned}$$

which implies that the asymptotic form of $F(n)$ also relates to that of $G(n)$. The rest of the paper discusses the case $W_X^t \neq \emptyset$, although it is also important to consider the case $W_X^t = \emptyset$, where the learning model cannot attain the true model.

The algebraic geometrical analysis (Watanabe 2001, 2009) is applicable to both the regular and singular cases for deriving the asymptotic form of $F_X(n)$. Its result shows that the form is expressed as

$$F_X(n) = \lambda_X \ln n - (m_X - 1) \ln \ln n + O(1),$$

where the coefficients λ_X and m_X are positive rational and natural, respectively. The reason why the free energy has this form will be explained in the next section. According to the relation Eq. (5), the asymptotic form of the generalization error is given by

$$G(n) = \frac{\lambda_X}{n} - \frac{m_X - 1}{n \ln n} + o\left(\frac{1}{n \ln n}\right). \qquad (6)$$

Since the learning model can attain the true model, we can confirm that the generalization error converges to zero for $n \to \infty$. The coefficients are $\lambda_X = d/2$ and $m_X = 1$ in the regular case. It is proved that $\lambda_X < d/2$ in the singular case (Section 7, Watanabe 2009).

3 Asymptotic analysis of the free energy and posterior convergence

This section introduces the asymptotic analysis of $F_X(n)$ based on algebraic geometry and explains how the prior distribution affects convergence of the posterior distribution. The topics in this section have already been elucidated in the studies on the Sing-OV estimation (e.g., Watanabe 2009).

3.1 Relation between the free energy and the zeta function

Let us define another Kullback–Leibler divergence,

$$H_X(w) = \int q(x) \ln \frac{q(x)}{p(x|w)} dx,$$

which is assumed to be analytic (Fundamental Condition I, Watanabe 2009). We consider the prior distribution $\varphi(w;\eta) = \psi_1(w;\eta)\psi_2(w;\eta)$, where $\psi_1(w;\eta)$ is a positive function of class C^∞ and $\psi_2(w;\eta)$ is a nonnegative analytic function (Fundamental Condition II, Watanabe 2009). Let the zeta function of a parametric model be given by

$$\zeta_X(z) = \int H_X(w)^z \varphi(w;\eta)dw,$$

where z is a complex variable. From algebraic analysis, we know that its poles are real, negative, and rational (Atiyah 1970). Let the largest pole and its order be $z = -\lambda_X$ and m_X, respectively. The zeta function includes the term

$$\zeta_X(z) = \frac{f_c(z)}{(z+\lambda_X)^{m_X}} + \cdots,$$

where $f_c(z)$ is a holomorphic function. We define the state density function of $t > 0$ as

$$v(t) = \int \delta(t - H_X(w))\varphi(w;\eta)dw.$$

The zeta function is its Mellin transform:

$$\zeta_X(z) = \mathcal{M}[v(t)] = \int_0^\infty v(t)t^z dt.$$

Moreover, it is known that the inverse Laplace transform of $v(t)$ has the same asymptotic form as $F_X(n)$;

$$\begin{aligned}\mathcal{L}^{-1}[v(t)] &= \int v(t)e^{nt}dt\\ &= \int e^{nH_X(w)}\varphi(w;\eta)dw = F_X(n).\end{aligned}$$

Then, there is the following relation,

$$F_X(n) \overset{\mathcal{L}}{\Longleftrightarrow} v(t) \overset{\mathcal{M}}{\Longleftrightarrow} \zeta_X(z).$$

Based on the Laplace and the Mellin transforms, the asymptotic forms of all functions are available if one of them is given. Following the transforms from $\zeta_X(z)$ to $F_X(n)$ through $v(t)$, we obtain the asymptotic form

$$F_X(n) = \lambda_X \ln n - (m_X - 1)\ln\ln n + O(1).$$

Let us define the effective area of the parameter space, which plays an important role in the convergence analysis of the posterior distribution. According to the results on the Sing-OV estimation, it has been found that the largest pole exists in a restricted parameter space. In Example 2, the parameter space is divided into W_1, W_2, W_3 and the rest of the support of $\varphi(w;\eta)$. The first three sets are neighborhoods of the analytic sets W_1^t, W_2^t and W_3^t constructing W_X^t, respectively. Assume that a pole $z = -\lambda_e$ of the zeta function

$$\zeta_e(z) = \int_{W_e} H_X(w)^z \varphi(w;\eta)dw$$

is equal to the largest pole $z = -\lambda_X$, where $W_e = W_1 \cap W_2$. In the present paper, we refer to W_e as *the effective area*. Let the effective area be denoted by the minimum set $W_1 \cap W_2$. In other words, we do not call W_1 the effect area even though W_1 includes W_e. If the largest pole of $\int_{W_1 \setminus W_1 \cap W_2} H_X(w)^z \varphi(w;\eta)dw$ is also equal to $z = -\lambda_X$, the effective area is W_1 since $W_1 \cap W_2$ can not cover the area.

3.2 Phase transition

A switch in the underlying function of the free energy is generally referred to as a phase transition. When the prior of the mixing ratio parameters is the Dirichlet distribution. the phase transition is observed in $F_X(n)$. Combining the results of Yamazaki et al. (2010) and Yamazaki and Kaji (2013), we obtain the following lemma;

Lemma 1 *Suppose that* $K = 2$, $K^* = 1$ *in the mixture model, where the true and the learning models are given by*

$$q(x) = f(x|b^*),$$
$$p(x|w) = af(x|b_1) + (1-a)f(x|b_2),$$

respectively. Let the component be expressed as

$$f(x = m|b_k) = \binom{M}{m} b_k^m (1-b_k)^{M-m},$$

where $x \in \{1, \ldots, M\}$, M *is an integer such that* $K < M$, *and* $(M\ m)^\top$ *is the binomial coefficient. We consider the case* $0 < b^* < 1$. *Let the prior distribution for the mixing ratio be the symmetric Dirichlet distribution, and the one for* b_k *be analytic and positive. Then the largest pole of the zeta function* $\zeta_X(z)$ *is*

$$\lambda_X = \begin{cases} \frac{1+\eta_1}{2} & \eta_1 \le 1/2, \\ \frac{3}{4} & \eta_1 > 1/2, \end{cases}$$

$$m_X = \begin{cases} 2 & \eta_1 = 1/2, \\ 1 & otherwise. \end{cases}$$

Moreover, the effective area W_e *is given by*

$$W_e = \begin{cases} W_1 \cup W_3 & \eta_1 < 1/2, \\ (W_1 \cap W_2) \cup (W_3 \cap W_2) & \eta_1 = 1/2, \\ W_2 & \eta_1 > 1/2. \end{cases}$$

The proof is in Appendix 3. Lemma 1 indicates that the free energy has the phase transition at $\eta_1 = 1/2$.

3.3 Convergence area of the posterior distribution

The asymptotic form of the free energy determines the limit structure of the posterior distribution. In this subsection, we will show that the convergence area is the effective parameter area.

The free energy $F(X^n)$ has an asymptotic form similar to the average energy $F_X(n)$ (Watanabe 2009, Main Formula II),

$$F(X^n) = nS(X^n) + \lambda_X \ln n - (m_X - 1)\ln\ln n + O_p(1), \tag{7}$$

where $S(X^n) = \frac{1}{n}\sum_{i=1}^n \ln q(x_i)$. According to $Z(X^n) = \exp(-F(X^n))$, the posterior distribution has the expression,

$$p\left(w|X^n\right) = \frac{\prod_{i=1}^n p(x_i|w)\varphi(w;\eta)}{\exp\left\{-nS(X^n) - \lambda_X \ln n + o_p(\ln n)\right\}}.$$

Let us divide the neighborhood of W_X^t into $W_e \cup W_o$, where W_e is the effective area. Then, there is a pole $z = -\mu_X$ such that $\mu_X > \lambda_X$ in the other area W_o, and the posterior value of W_o is described by

$$\begin{aligned} p(W_o|X^n) &= \int_{W_o} p(w|X^n)dw \\ &= \frac{\int_{W_o} \prod_{i=1}^n p(x_i|w)\varphi(w;\eta)dw}{\exp\{-nS(X^n) - \lambda_X \ln n + o_p(\ln n)\}} \\ &= \frac{\exp\{-nS(X^n) - \mu_X \ln n + o_p(\ln n)\}}{\exp\{-nS(X^n) - \lambda_X \ln n + o_p(\ln n)\}} \\ &= n^{-\mu_X + \lambda_X} + o_p(n^{-\mu_X + \lambda_X}). \end{aligned}$$

The posterior asymptotically has zero value in W_o, which means that it converges to the effective area.

According to Lemma 1, the effective area depends on the hyperparameter. Therefore, the convergence area changes at the phase transition point $\eta_1 = 1/2$. It also shows how the learning model realizes the true one. In $W_1 \cup W_3$, the true model is expressed by one-component model, which means that the redundant component is eliminated. On the other hand, all components of the learning model are used in W_2.

The phase transition is observed in general mixture models;

Theorem 1 *Let a learning model and the true one be expressed as Eqs.* (1) *and* (2)*, respectively. When the prior of the mixing ratio is the Dirichlet distribution of Eq.* (4)*, the average free energy $F_X(n)$ has at least two phases: the phase that eliminates all redundant components when η_1 is small, and the one that uses them when η_1 is sufficiently large.*

The proof is in Appendix 3.

4 Formal definition of the latent-variable estimation and its accuracy

This section formulates the Bayes latent-variable estimation and an error function that measures its accuracy.

We first consider a detailed definition of a latent variable. Let $Y^n = \{y_1, \ldots, y_n\}$ be unobservable data, which correspond to the latent parts of the observable X^n. Then, the complete form of the data is (x_i, y_i), and (X^n, Y^n) and X^n are referred to as complete and incomplete data, respectively. The true model generates the complete data (X^n, Y^n), where the range of the latent variables is $y_i \in \{1, \ldots, K^*\}$. The learning model, on the other hand, has the range $y_i \in \{1, \ldots, K\}$. For a unified description, we define that the true model has probabilities $q(y) = 0$ and $q(x, y) = 0$ for $y > K^*$.

We define the true parameter set for (x, y) as

$$W_{XY}^t = \{w^*; p(x, y|w^*) = q(x, y)\},$$

which is a proper subset of W_X^t. In Example 2,

$$W_{XY}^t = \{a = 1, b_1 = b^*\} = W_1^t \subset W_X^t.$$

The subsets $W_2 = \{b_1 = b_2 = b^*\}$ and $W_3 = \{a = 0, b_2 = b^*\}$ in W_X^t are excluded since W_{XY}^t takes account of the representation with respect to not only x but also y. Due to the assumption $W_X^t \neq \emptyset$, W_{XY}^t is not empty. The set W_{XY}^t again consists of an analytic set in the singular case, and it is a unique point in the regular case.

While latent-variable estimation falls into various types according to the target of the estimation, the present paper focuses on the Type-I estimation of Yamazaki (2014): the joint probability of $(y_1, \ldots, y_n)$ is the target and is written as $p(Y^n|X^n)$. The Bayes estimation has two equivalent definitions:

$$p(Y^n|X^n) = \int \prod_{i=1}^{n} \frac{p(x_i, y_i|w)}{p(x_i|w)} p(w|X^n)dw \tag{8}$$

$$= \frac{Z(X^n, Y^n)}{Z(X^n)}, \tag{9}$$

where the marginal likelihood for the complete data is given by

$$Z(X^n, Y^n) = \int \prod_{i=1}^{n} p(x_i, y_i|w)\varphi(w; \eta)dw.$$

It is easily confirmed that $Z(X^n) = \sum_{Y^n} Z(X^n, Y^n)$.

The true probability of Y^n is uniquely given by

$$q(Y^n|X^n) = \frac{q(X^n, Y^n)}{q(X^n)} = \prod_{i=1}^{n} \frac{q(x_i, y_i)}{q(x_i)}. \tag{10}$$

The accuracy of the estimation is measured by the difference between $q(Y^n|X^n)$ and $p(Y^n|X^n)$. Thus, we define the error function as the average Kullback–Leibler divergence,

$$D(n) = \frac{1}{n} E_{XY}\left[\ln \frac{q(Y^n|X^n)}{p(Y^n|X^n)} \right], \tag{11}$$

where the expectation is defined as

$$E_{XY}\left[f(X^n, Y^n)\right] = \int \sum_{y_1=1}^{K} \cdots \sum_{y_n=1}^{K} f(X^n, Y^n)q(X^n, Y^n)dX^n.$$

5 Asymptotic analysis of the error function

In this section, we show that the algebraic geometrical analysis is applicable to the Sing-LV estimation, and present the asymptotic form of the error function $D(n)$.

5.1 Conditions for the analysis

Before showing the asymptotic form of the error function, we state necessary conditions.

Let us define the zeta function on the complete data (x, y) as

$$\zeta_{XY}(z) = \int H_{XY}(w)^z \varphi(w; \eta)dw,$$

where the Kullback–Leibler divergence $H_{XY}(w)$ is given by

$$H_{XY}(w) = \sum_{y=1}^{K} \int q(x, y) \ln \frac{q(x, y)}{p(x, y|w)} dx.$$

Let the largest pole of $\zeta_{XY}(z)$ be $z = -\lambda_{XY}$, and let its order be m_{XY}.

We consider the following conditions:

(A1) The divergence functions $H_{XY}(w)$ and $H_X(w)$ are analytic.

(A2) The prior distribution has the compact support, which includes W_X^t, and has the expression $\varphi(w; \eta) = \psi_1(w; \eta)\psi_2(w; \eta)$, where $\psi_1(w; \eta) > 0$ is a function of class C^∞ and $\psi_2(w; \eta) \geq 0$ is analytic on the support of $\varphi(w; \eta)$.

They correspond to the Fundamental Conditions I and II in Watanabe (2009), respectively. It is known that models with discrete x such as the binomial mixture satisfy (A1) (Yamazaki et al. 2010). On the other hand, if x is continuous, there are some models, of which $H_X(w)$ is not analytic;

Example 4 (**Example 7.3 in** Watanabe 2009) In the Gaussian mixture, $H_X(w)$ is not analytic, which means that the mixture model does not satisfies (A1). Let us consider a simple case; $K = 2$ and $K^* = 1$, where the true model and a learning model are given by

$$q(x) = f(x|0),$$
$$p(x|a) = af(x|2) + (1-a)f(x|0),$$

respectively, where $x \in R^1$,

$$f(x|b) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-b)^2}{2}\right\},$$

and $b \in W^1 = R^1$. Then,

$$\begin{aligned} H_X(a) &= \int q(x) \ln \frac{q(x)}{p(x|a)} dx \\ &= -\int q(x) \ln\{1 + a(\exp(2x-2) - 1)\}dx \\ &= \int \sum_{j=1}^{\infty} \frac{a^j}{j}(1 - \exp(2x-2))^j q(x)dx, \end{aligned}$$

where the last expression is a formal expansion. Since its convergence radius is zero at $a = 0$, $H_X(a)$ is not analytic. Based on the similar way, we can find that $H_X(w)$ is not analytic in a general Gaussian mixture .

The following example shows a prior distribution for the mixture model satisfying (A2).

Example 5 The symmetric Dirichlet distribution satisfies the condition (A2) because Eq. (4) is obviously analytic and non negative in its support. Choosing an analytic distribution for $\varphi(b; \eta_2)$, we obtain the prior $\varphi(w; \eta)$ satisfying the condition (A2).

5.2 Asymptotic form of the error function

Now, we show the main theorem on the asymptotic form of the error function:

Theorem 2 *Let the true distribution of the latent variables and the estimated distribution be defined by Eqs.* (9) *and* (10)*, respectively. By assuming the conditions (A1) and (A2), the asymptotic form of $D(n)$ is expressed as*

$$D(n) = (\lambda_{XY} - \lambda_X)\frac{\ln n}{n} - (m_{XY} - m_X)\frac{\ln \ln n}{n} + o\left(\frac{\ln \ln n}{n}\right).$$

The proof is in Appendix 1. The theorem indicates that the algebraic geometrical method plays an essential role for the analysis of the Sing-LV estimation because the coefficients consist of the information of the zeta functions such as λ_{XY}, λ_X, m_{XY} and m_X. The order $\ln n/n$ has not ever appeared in the Reg-LV estimation. In the Reg-LV estimation such that $K = K^*$, the asymptotic error function has the following form (Yamazaki 2014);

$$D(n) = \frac{1}{n}\mathrm{Tr}\left[I_{XY}(w^*)I_X(w^*)^{-1}\right] + o\left(\frac{1}{n}\right),$$

$$\{I_{XY}(w)\}_{ij} = \sum_{y=1}^{K}\int \frac{\partial \ln p(x,y|w)}{\partial w_i}\frac{\partial \ln p(x,y|w)}{\partial w_j}p(x,y|w)dx,$$

$$\{I_X(w)\}_{ij} = \int \frac{\partial \ln p(x|w)}{\partial w_i}\frac{\partial \ln p(x|w)}{\partial w_j}p(x|w)dx,$$

where w^* is the unique point consisting of W^t_{XY}. The dominant order is $1/n$, and the coefficient is determined by the Fisher information matrices on $p(x,y|w)$ and $p(x|w)$. Theorem 2 implies that the largest possible order is $\ln n/n$ in the Sing-LV estimation. This order change is adverse for the performance because the error converges more slowly to zero. In singular cases, the probability $p(Y^n|X^n)$ is constructed over the space $Y^n \in K^n$ while the true probability $q(Y^n|X^n)$ is over $Y^n \in K^{*n}$. The size of the redundant space $K^n - K^{*n}$ grows exponentially with the amount of training data. For realizing $p(x,y|w^*)$, where $w^* \in W^t_{XY}$, we must assign zero to the probabilities on the vast redundant space. The increased order reflects the cost of assigning these values.

Let us compare the dominant order of $D(n)$ with that of the generalization error. We find that both Reg-OV and Sing-OV estimations have the same dominant order $1/n$ as shown in Eq. (6) while the redundancy and the hyperparameter affect the coefficients. Thus, changing the order is a unique phenomenon of the latent-variable estimation.

5.3 Asymptotic error in the mixture model

In Theorem 2, the possible dominant order was calculated as $\ln n/n$. However, there is no guarantee that this is the actual maximum order; the order can decrease to $1/n$ if the coefficients are zero, where the zeta functions $\zeta_{XY}(z)$ and $\zeta_X(z)$ have their largest poles in the same position and their multiple orders are also the same. The result of the following theorem clearly shows that the dominant order is $\ln n/n$ in the mixture models.

Theorem 3 *Let the learning and the true models be mixtures defined by Eqs.* (1) *and* (2)*, respectively. Assume the conditions (A1) and (A2). The Bayes estimation for the latent variables, Eq.* (9)*, with the prior represented by Eqs.* (3) *and* (4) *has the following bound for the asymptotic error:*

$$D(n) \geq \frac{(K-K^*)\eta_1}{2}\frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right).$$

The proof is in Appendix 1. Due to the definition of the Dirichlet distribution, η_1 is positive. Combining this with the assumption $K^* < K$, we obtain that the coefficient of $(\ln n)/n$ is positive, which indicates that it is the dominant order.

The Dirichlet prior distribution for the mixing ratio is qualitatively known to have a function controlling the number of available components, the so-called automatic relevance determination (ARD); a small hyperparameter tends to have a result with few components due to the shape of the distribution. Theorem 3 quantitatively shows an effect of the Dirichlet prior.

The lower bound in the theorem mathematically supports the ARD effect; the redundancy $K - K^*$ and the hyperparameter η_1 have a linear influence on the accuracy.

Theorem 3 holds in a wider class of the mixture models since the error is evaluated as the lower bound. The following corollary shows that the Gaussian mixture has the same bound for the error even though it does not satisfy (A1) as shown in Example 4.

Corollary 1 *Assume that in a mixture model, $H_{XY}(w)$ is analytic, and the prior distribution for the mixing ratio is the symmetric Dirichlet distribution. If there is a positive constant C_1 such that*

$$H_X(w) \le C_1 \int \left(\frac{p(x|w)}{q(x)} - 1 \right)^2 dx,$$

the error function has the same lower bound as Theorem 3. In the Gaussian mixture, components of which are defined by

$$f(x|b) = \frac{1}{\sqrt{2\pi}^M} \exp\left\{ -\frac{||x-b||^2}{2} \right\},$$

where $x \in R^M$ and $b \in W^M = R^M$, $H_{XY}(w)$ is analytic and the inequality holds.

The proof is in Appendix 1.

6 Discussion

Theorem 2 shows that the asymptotic error has the coefficient $\lambda_{XY} - \lambda_X$, which is the difference of the largest poles in the zeta functions. Based on the free energy of the complete data defined as $F(X^n, Y^n) = -\ln Z(X^n, Y^n)$, we find that the error is determined by the different properties between $F(X^n, Y^n)$ and $F(X^n)$ since their asymptotic forms are expressed as

$$F(X^n, Y^n) = nS(X^n, Y^n) + \lambda_{XY} \ln n - (m_{XY} - 1) \ln \ln n + O_p(1),$$
$$F(X^n) = nS(X^n) + \lambda_X \ln n - (m_X - 1) \ln \ln n + O_p(1),$$

where $S(X^n, Y^n) = -\frac{1}{n} \sum_{i=1}^n \ln q(x_i, y_i)$.

In this section, we examine the properties of $F(X^n, Y^n)$ and indicate that the difference from those of $F(X^n)$ affects the behavior of the Sing-LV estimation and the parameter sampling from the posterior distribution.

6.1 Effect to eliminate redundant labels

According to Eq. (9), the MCMC sampling of the Y^n's following $p(Y^n|X^n)$ is essential for the Bayes estimation. The following relation indicates that we do not need to calculate $Z(X^n)$ and that the value of $Z(X^n, Y^n)$ determines the properties of the estimation:

$$p(X^n, Y^n) = Z(X^n, Y^n) \propto p(Y^n|X^n) = \frac{Z(X^n, Y^n)}{Z(X^n)}. \quad (12)$$

The expression of $p(X^n, Y^n)$ can be tractable with a conjugate prior, which marginalizes out the parameter integral (Dawid and Lauritzen 1993; Heckerman 1999).

We determine where the estimated distribution $p(Y^n|X^n)$ has its peak. Obviously, the label assignment Y^n minimizing $F(X^n, Y^n)$ provides the peak due to the definition $F(X^n, Y^n) = -\ln Z(X^n, Y^n)$ and Eq. (12). Let this assignment be described as $\bar{Y}^n$;

$$\bar{Y}^n = \arg\max_{Y^n} p(Y^n|X^n) = \arg\min_{Y^n} F(X^n, Y^n).$$

The following discussion shows that $\bar{Y}^n$ does not include the redundant labels.

We have to consider the symmetry of the latent variable in order to discuss the peak. In latent-variable models, both the latent variable and the parameter are symmetric. In Example 2, the component $f(x|b^*)$ of the true model can be attained by the first component $a_1 f(x|b_1)$ or the second one $(1-a_1)f(x|b_2)$ of the learning model. Because the true label $y = 1$, which the true model provides, is unobservable, there are two proper estimation results $Y^n = \{1, \ldots, 1\}$ and $Y^n = \{2, \ldots, 2\}$ to indicate that the true model consists of one component. This is the symmetry of the latent variable. In the parameter space, it corresponds to the symmetric structure of W_1 and W_3 shown in Fig. 1. The symmetry makes it difficult to interpret the estimation results, which is known as the label-switching problem.

For the purpose of the theoretical evaluation, the definition of the error function $D(n)$ selects the true assignment of the latent variable. In the above example, only $Y^n = \{1, \ldots, 1\}$ is accepted as the proper result. However, there is no selection of the true assignment in the estimation process; other symmetric assignments such as $Y^n = \{2, \ldots, 2\}$ will be the peak of $p(Y^n|X^n)$. Then, the true parameter area W^t_{XY} is not sufficient to describe the peak. Taking account of the symmetry, we define another analytic set of the parameter as

$$W^p_{XY} = \cup_{\sigma \in \Sigma} \left\{ w; a_{\sigma(k)} = a^*_k, b_{\sigma(k)} = b^*_k \quad \text{for} \quad 1 \le k \le K^* \right\},$$

Σ is the set of injective functions from $\{1, \ldots, K^*\}$ to $\{1, \ldots, K\}$. It is easy to confirm that $W^t_{XY} \subset W^p_{XY}$. In Example 2, $W^t_{XY} = W^t_1 \subset W^t_1 \cup W^t_3 = W^p_{XY}$. Note that the redundant components are eliminated in $p(x|w^*)$, where $w^* \in W^p_{XY}$.

Let us analyze the location of the peak. Define that

$$S'(X^n, Y^n) = -\frac{1}{n} \sum_{i=1}^{n} \ln p(x_i, y_i|w^*),$$

where $w^* \in W^p_{XY}$. Switching the label based on the symmetry, we can easily prove that $\max_{w^*, Y^n} S'(X^n, Y^n) = \max_{Y^n} S(X^n, Y^n)$. Moreover, $-\frac{1}{n}\sum_{i=1}^{n} \ln p(x_i, y_i|w)$ with $w \in W^t_X \setminus W^p_{XY}$, such as $w \in W^t_2$ in Example 2, cannot realize $S(X^n, Y^n)$ according to a simple calculation as shown in the next paragraph. Because the leading term of the asymptotic $F(X^n, Y^n)$ is $nS(X^n, Y^n)$ and $nS'(X^n, \bar{Y}^n)$ realizes it, the true assignment $\bar{Y}^n$ follows the parameter $w^* \in W^p_{XY}$. Recalling that the redundant components are eliminated when $w \in W^p_{XY}$, we can conclude that the redundant labels are eliminated in $\bar{Y}^n$. This elimination occurs in any prior distribution if its support includes W^p_{XY}.

Let us confirm the elimination in Example 2. We consider three parameters; $w^*_1 \in W^t_1 = \{a = 1, b_1 = b^*\}$, $w^*_2 \in W^t_2 = \{b_1 = b_2 = b^*\}$ and $w^*_3 \in W^t_3 = \{a = 0, b_2 = b^*\}$. The leading term of the asymptotic $F(X^n, Y^n)$ is expressed as

$$nS'_j(X^n, Y^n) = -\sum_{i=1}^{n} \ln p(x_i, y_i|w^*_j)$$

for $j = 1, 2, 3$. This is rewritten as

$$\begin{aligned} nS'_j(X^n, Y^n) = &-\sum_{i=1}^{n} \delta_{y_i,1} \ln a - \sum_{i=1}^{n} \delta_{y_i,2} \ln(1-a) \\ &- \sum_{i=1}^{n} \delta_{y_i,1} \ln f(x_i|b_1) - \sum_{i=1}^{n} \delta_{y_i,2} \ln f(x_i|b_2), \end{aligned} \tag{13}$$

where $\delta_{i,j}$ is the Kronecker delta. The assignment $\bar{Y}^n$ depends on w_j^*. For example, $\bar{Y}^n = \{1, \ldots, 1\}$ for w_1^* and $\bar{Y}^n = \{2, \ldots, 2\}$ for w_3^*. Then, we obtain that

$$nS'_j(X^n, Y^n) = \begin{cases} -\sum_{i=1}^{n} \ln f(x_i|b^*) & j = 1 \\ -N_1 \ln a - N_2 \ln(1-a) - \sum_{i=1}^{n} \ln f(x_i|b^*) & j = 2 \\ -\sum_{i=1}^{n} \ln f(x_i|b^*) & j = 3, \end{cases}$$

where $N_1 = \sum_{i=1}^{n} \delta_{y_i,1}$ and $N_2 = \sum_{i=1}^{n} \delta_{y_i,2}$. The cases $j = 1$ and $j = 3$ have the same value and the case $j = 2$ is smaller than the others due to the first two terms in Eq. (13), which holds for any value of $0 < a < 1$ in W_2^t. This means that W_2^t cannot make $p(Y^n|X^n)$ maximum. In other words, the assignment Y^n using both labels 1 and 2 is not the peak.

6.2 Two approaches to calculate $p(Y^n|X^n)$ and their difference

It is necessary to emphasize that the calculation of $p(Y^n|X^n)$ based on sampling from $p(w|X^n)$ following Eq. (8) can be inaccurate. According to Theorem 1 and Eq. (7), we confirm that $F(X^n)$ has a phase transition in mixture models due to the hyperparameter of the Dirichlet prior. This means that, when the hyperparameter η_1 is large, the Monte Carlo sampling are from the area, in which all the components are used such as W_2^t. In the numerical computation, the integrand of Eq. (8) will be close to $\prod_{i=1}^{n} p(x_i, y_i|w_2^*)/p(x_i|w_2^*)$, where $w_2^* \in W_2^t$. Because $w_2^* \in W_2^t \subset W_X^t$,

$$\begin{aligned} \prod_{i=1}^{n} \frac{p(x_i, y_i|w_2^*)}{p(x_i|w_2^*)} &= \exp\left\{ \sum_{i=1}^{n} \ln p(x_i, y_i|w_2^*) - \sum_{i=1}^{n} \ln p(x_i|w_2^*) \right\} \\ &= \exp\big\{ -nS'_2(X^n, Y^n) + nS(X^n) \big\}. \end{aligned}$$

On the other hand, based on Eq. (9), the desired value of $p(Y^n|X^n)$ is calculated as

$$\begin{aligned} \frac{Z(X^n, Y^n)}{Z(X^n)} &= \exp\left\{ F(X^n) - F(X^n, Y^n) \right\} \\ &= \exp\{-nS(X^n, Y^n) + nS(X^n)\} + o(\exp(-n)). \end{aligned}$$

Since $S'_2(X^n, Y^n) > S(X^n, Y^n)$, the value of Eq. (8) is much smaller than that of Eq. (9). Therefore, the result of the numerical integration in Eq. (8) is almost zero. The parameter area providing non-zero value of integrand in Eq. (8) is located in the tail of the posterior distribution when $p(w|X^n)$ converges to $W_X^t \setminus W_{XY}^p$.

6.3 Failure of parameter sampling from the posterior distribution

In the previous subsection, parameter sampling from the posterior distribution can make an adverse effect on the calculation of the distribution of the latent variable. Here, in the other way, we show that latent-variable sampling can construct an undesired posterior distribution.

There are methods to sample a sequence of $\{w, Y^n\}$ from $p(w, Y^n|X^n)$. Ignoring Y^n, we obtain the sequence $\{w\}$. The Gibbs sampling in the MCMC method (Robert and Casella 2005) is one of the representative techniques.

[Gibbs Sampling for a Model with a Latent Variable]

1. Initialize the parameter;
2. Sample Y^n based on $p(Y^n|w, X^n)$;
3. Sample w based on $p(w|Y^n, X^n)$;
4. Iterate by alternately updating Step 2 and Step 3.

The sequence of $\{w, Y^n\}$ obtained by this algorithm follows $p(w, Y^n|X^n)$. The extracted parameter sequence $\{w\}$ is assumed to be samples from the posterior because $p_G(w|X^n) = \sum_{Y^n} p(w, Y^n|X^n)$ is theoretically equal to $p(w|X^n)$. However, in the mixture models, the practical value of $p_G(w|X^n)$ based on the Monte Carlo method can be different from that of the original posterior $p(w|X^n)$ when the hyperparameter for the mixing ratio η_1 is large.

Let us consider the expression

$$\begin{aligned} -\ln p(X^n, Y^n, w) &= -\ln \prod_{i=1}^{n} \prod_{k=1}^{K} a_k^{\delta_{y_i k}} f(x_i|b_k)^{\delta_{y_i k}} - \ln \varphi(w; \eta) \\ &= -\sum_{k=1}^{K} \delta_{y_i k} \ln a_k - \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{y_i k} \ln f(x_i|b_k) - \ln \varphi(w; \eta). \end{aligned}$$

We determine a location of a pair $(\bar{w}, \bar{Y}^n)$ that minimizes this expression in the asymptotic case $n \to \infty$ because the relation $p(X^n, Y^n, w) \propto p(w, Y^n|X^n)$ indicates that the sequence $\{w, Y^n\}$ is mainly taken from the neighborhood of the pair. The third term of the last expression does not have any asymptotic effect because it has the constant order on n. The first two terms have the same expression as Eq. (13). Based on the calculation of $S'_j(X^n, Y^n)$, $\bar{w} \in W^p_{XY}$ and $\bar{Y}^n = \arg\max_{Y^n} p(X^n, Y^n, \bar{w})$. Therefore, the practical value of $p_G(w|X^n)$ is calculated by the sequence $\{w\}$ around W^p_{XY} for any η_1 while the convergence area of the original $p(w|X^n)$ depends on the phase of $F(X^n)$ controlled by η_1.

In Example 2, the posterior $p(w|X^n)$ converges to W^t_2 when η_1 is large. On the other hand, the sampled sequence based on $p(X^n, Y^n, w)$ are mainly from $W_1 \cup W_3$ since $S'_2(X^n, \bar{Y}^n_2) > S'_1(X^n, \bar{Y}^n_1) = S'_3(X^n, \bar{Y}^n_3)$, where $\bar{Y}^n_j$ stands for the assignment minimizing $S'_j(X^n, Y^n)$. In order to construct the sequence $\{w\}$ following $p(w|X^n)$, we need samples $(w, Y^n) \in W_2 \times \bar{Y}^n_2$, which are located in the tail of $p(w, Y^n|X^n)$. In theory, the sequence $\{w\}$ from $p(w, Y^n|X^n)$ realizes the one from $p(w|X^n)$. However, in practice, it is not straightforward to obtain $\{w, Y^n\}$ from the tail of $p(w, Y^n|X^n)$. This property of the Gibbs sampling has been reported in a Gaussian mixture model (Nagata and Watanabe 2009). The experimental results show that the obtained sequence of $\{w\}$ is localized in the area corresponding to W^p_{XY}. Note that there is no failure of the MCMC method when η_1 is sufficiently small, where the peaks of $p(w|X^n)$ and $p(w, Y^n|X^n)$ are in the same area. Thus, to judge the reliability of the MCMC sampling, we have to know the phase transition point such as $\eta_1 = 1/2$ in Lemma 1.

7 Conclusions

The present paper clarifies the asymptotic accuracy of the Bayes latent-variable estimation. The dominant order is at most $\ln n/n$, and its coefficient is determined by a positional relation between the largest poles of the zeta functions. According to the mixture-model case, it is suggested that the order is dominant and the coefficient is affected by the redundancy of the learning model and the hyperparameters. The accuracy of prediction can be approximated by methods such as the cross-validation and bootstrap methods. On the other hand, there is no approximation for the accuracy of latent-variable estimation, which indicates that the theoretical result plays a central role in evaluating the model and the estimation method.

Appendix 1

Here, we prove Theorems 2 and 3, and Corollary 1.

Proof of Theorem 2

Proof Let us define another average free energy as

$$F_{XY}(n) = -nS_{XY} + E_{XY}\Big[-\ln Z(X^n, Y^n)\Big],$$

where the entropy function is given by

$$S_{XY} = -\sum_{y=1}^{K} \int q(x, y) \ln q(x, y) dx.$$

According to the definitions of the error function $D(n)$ and the Bayes estimation method Eq. (9), it holds that

$$\begin{aligned} nD(n) &= E_{XY}\Big[\ln \frac{q(X^n, Y^n)}{Z(X^n, Y^n)}\Big] - E_X\Big[\ln \frac{q(X^n)}{Z(X^n)}\Big] \\ &= -nS_{XY} - E_{XY}\Big[\ln Z(X^n, Y^n)\Big] + nS_X + E_X\Big[\ln Z(X^n)\Big] \\ &= F_{XY}(n) - F_X(n). \end{aligned}$$

Based on (A1), (A2), and algebraic geometrical analysis, we obtain the asymptotic forms of $F_{XY}(n)$ and $F_X(n)$:

$$\begin{aligned} F_{XY}(n) &= \lambda_{XY} \ln n - (m_{XY} - 1) \ln \ln n + O(1), \\ F_X(n) &= \lambda_X \ln n - (m_X - 1) \ln \ln n + O(1), \end{aligned}$$

which proves the theorem. □

Outline of the calculation of a pole of the zeta function

We will show the outline of calculation to find a pole. Let us introduce some useful lemmas for the zeta function. The proofs are omitted because they are almost obvious due to the relation between the free energy and the zeta function.

Lemma 2 *Let the largest poles of the zeta functions $\int H_1(w)^z \varphi(w)dw$ and $\int H_2(w)^z \varphi(w)dw$ be $z = -\lambda_1$ and $z = -\lambda_2$, respectively. It holds that $\lambda_1 \leq \lambda_2$ when $H_1(w) \leq H_2(w)$ on the support of $\varphi(w)$.*

Lemma 3 *Under the same conditions as Lemma 2, it holds that* $\lambda_1 = \lambda_2$ *if there exist positive constants* C_1 *and* C_2 *such that* $C_1 H_2(w) \leq H_1(w) \leq C_2 H_2(w)$.

We define an equivalence relation $H_1(w) \equiv H_2(w)$ due to $\lambda_1 = \lambda_2$ in Lemma 3.

Let us now calculate a general zeta function $\int H(w)^z \varphi(w) dw$. First, we focus on the restricted area W_{res}, which is the neighborhood of $\{w : H(w) = 0\}$ in the parameter space because poles of the zeta function do not depend on other areas (Watanabe 2001). Next, we need a function $H^{\text{alg}}(w)$, which is a polynomial of w and satisfies $H(w) \equiv H^{\text{alg}}(w)$. Based on Lemma 3, the largest pole of the zeta function $\int_{W_{res}} H^{\text{alg}}(w)^z \varphi(w) dw$ is the same as that of the original zeta function. According to the resolution of singularities (Hironaka 1964), there is a mapping $u = \Phi(w)$ such that

$$H^{\text{alg}}(\Phi(w)) = a(u) u_1^{2\alpha_1} u_2^{2\alpha_2} \ldots u_d^{2\alpha_d}, \tag{14}$$

where $a(u)$ is a non-zero analytic function in $\{u{:}H(\Phi(w)) = 0\}$, and $\alpha_1, \ldots, \alpha_d$ are integers. Let $|\Phi| = |u_1|^{\beta_1} \ldots |u_d|^{\beta_d}$ be the Jacobian, and the prior distribution is described as $\varphi(\Phi(w)) = u_1^{\gamma_1} \ldots u_d^{\gamma_d}$, where β_i and γ_i are integers. Then, it holds that

$$\begin{aligned}\int_{W_{res}} H^{\text{alg}}(w)^z \varphi(w) dw &= \int_{\Phi(W_{res})} H^{\text{alg}}(\Phi(w))^z \varphi(\Phi(w)) |\Phi| du \\ &= \int_{\Phi(W_{res})} a(u) \prod_{i=1}^{d} u_i^{2\alpha_i z + \gamma_i} |u_i|^{\beta_i} du.\end{aligned}$$

Calculating the integral over u_i in the last expression, we find that the zeta function has factors $(2\alpha_i z + \beta_i + \gamma_i + 1)^{-1}$. This means that there are poles $z = -(\beta_i + \gamma_i + 1)/(2\alpha_i)$.

When it is not straightforward to find the multiple form such as Eq. (14), we can consider a partially-multiple form;

$$H^{\text{alg}}(\Phi(w)) = a(u) u_1^{2\alpha_1} g(u \setminus u_1),$$

where the function $g(u \setminus u_1)$ can be a polynomial of $u \setminus u_1$. The zeta function is written as

$$\int_{\Phi(w)} a(u) g(u \setminus u_1) u_1^{2\alpha_1 z + \gamma_1} |u_1|^{\beta_1} du.$$

Calculating the integral over u_1, we obtain a pole $z = -(\beta_1 + \gamma_1 + 1)/(2\alpha_1)$.

Assume that we obtain a partially-multiple form as the upper bounds such that

$$H^{\text{alg}}(w) \leq a(u) u_1^{2\alpha_1} g(u \setminus u_1),$$

where the Jacobian and the prior include factors $|u_1|^{\beta_1}$ and $u_1^{\gamma_1}$, respectively. Due to Lemma 2, a pole of the zeta function with respect to the right-hand side provides the upper bounds $\lambda \leq (\beta_1 + \gamma_1 + 1)/(2\alpha_1)$.

Proof of Theorem 3

The following lemma shows the calculation of λ_{XY}.

Lemma 4 *The largest pole of the zeta function* $\zeta_{XY}(z)$ *is*

$$\begin{aligned}\lambda_{XY} &= \frac{K^* - 1 + K^* d_c}{2} + (K - K^*)\eta_1, \\ m_{XY} &= 1.\end{aligned}$$

Proof (Lemma 4) We consider a restricted parameter space W_1, which is a neighborhood of W^t_{XY} given by

$$\begin{aligned} a_k &= a_k^* \quad (2 \leq k \leq K^*), \\ a_k &= 0 \quad (k > K^*), \\ b_k &= b_k^* \quad (1 \leq k \leq K^*). \end{aligned}$$

This is a generalization of W_1 in Example 2. The Kullback–Leibler divergence has the expression

$$\begin{aligned} H_{XY}(w) &= \sum_{k=1}^{K^*} a_k^* \left\{ \ln \frac{a_k^*}{a_k} + \int f(x|b_k^*) \ln \frac{f(x|b_k^*)}{f(x|b_k)} dx \right\} \\ &\equiv \left(1 - \sum_{k=2}^{K^*} a_k^*\right) \ln \frac{1 - \sum_{k=2}^{K^*} a_k^*}{1 - \sum_{k=2}^{K} a_k} + \int f(x|b_1^*) \ln \frac{f(x|b_1^*)}{f(x|b_1)} dx \\ &\quad + \sum_{k=2}^{K^*} \left\{ a_k^* \ln \frac{a_k^*}{a_k} + \int f(x|b_k^*) \ln \frac{f(x|b_k^*)}{f(x|b_k)} dx \right\}. \end{aligned}$$

Based on the shift transformation $\Phi_1(w)$, such that

$$\begin{aligned} \bar{a}_k &= a_k - a_k^* \quad (2 \leq k \leq K^*), \\ \bar{a}_k &= a_k \quad (k > K^*), \\ \bar{b}_{km} &= b_{km} - b_{km}^* \quad (1 \leq k \leq K^*, 1 \leq m \leq d_c), \\ \bar{b}_{km} &= b_{km} \quad (k > K^*, 1 \leq m \leq d_c), \end{aligned}$$

we can find an equivalent polynomial described as

$$H_{XY}(\Phi_1(w)) \equiv \sum_{k=2}^{K^*} \bar{a}_k^2 + \sum_{k=K^*+1}^{K} \bar{a}_k + \sum_{k=1}^{K^*} \bar{b}_k^2., \tag{15}$$

where the detailed derivation is in Appendix 2. Let the right-hand side of Eq. (15) be $H^{\text{alg}}_{XY}(\Phi_1(w))$, and consider a zeta function given by

$$\zeta_1(z) = \int H^{\text{alg}}_{XY}(\Phi_1(w))^z \varphi(\Phi_1(w); \eta) d\Phi_1(w).$$

According to Lemma 3, the positions of the poles of $\zeta_1(z)$ are the same as those of $\zeta_{XY}(z)$. By using a blow-up Φ_2 defined by

$$\begin{aligned} u_2 &= \bar{a}_2, \\ u_2 u_k &= \bar{a}_k \quad (2 < k \leq K^*), \\ u_2^2 u_k &= \bar{a}_k \quad (k > K^*), \\ u_2 v_{km} &= \bar{b}_{km} \quad (1 \leq k \leq K^*, 1 \leq m \leq d_c), \\ v_{km} &= \bar{b}_{km} \quad (k > K^*, 1 \leq m \leq d_c), \end{aligned}$$

we obtain the following expression in the restricted area,

$$\zeta_1(z) = \int_{\Phi_2\Phi_1(W_1)} f_1(\Phi_2\Phi_1(w)) u_2^{2z} \varphi(\Phi_2\Phi_1(w); \eta) |u_2|^{K^*-2+K^*d_c+2(K-K^*)} d\Phi_2\Phi_1(w),$$

where f_1 is a function consisting of the parameters except for u_2, and a factor on $|u_2|$ is derived from the Jacobian of Φ_2. Note that there is not u_1 as a parameter in $\Phi_2\Phi_1(w)$ since w_1 is already omitted on the basis of the relation $a_1 = 1 - \sum_{k=2}^{K} a_k$. The symmetric Dirichlet prior has a factor $\prod_{k=2} a_k^{\eta_1 - 1}$ in the original parameter space. According to $\Phi_2\Phi_1(a_k) = u_2^2 u_k$ for $k > K^*$, it has a factor $u_2^{2(K-K^*)(\eta_1-1)}$ in the space of $\Phi_2\Phi_1(w)$, which indicates that $\zeta_1(z)$ has a pole at $z = -(K^* - 1 + K^* d_c)/2 - (K - K^*)\eta_1$. Considering the symmetry of the parameters in $H_{XY}^{\text{alg}}(w)$, we determine that this pole is the largest and that its order is $m_{XY} = 1$, which proves Lemma 4. □

The result for λ_X is shown in the following lemma.

Lemma 5 *The largest pole of the zeta function $\zeta_X(z)$ has the bound*

$$\lambda_X \le \mu = \frac{K^* - 1 + K^* d_c}{2} + \frac{(K - K^*)\eta_1}{2}.$$

Proof (Lemma 5) It is known [cf. Yamazaki et al. (2010) ; Section 7.8 of Watanabe (2009)] that, in the restricted area W_1, there are positive constants C_1 and C_1' such that

$$\begin{aligned} H_X(w) \le & C_1 \int \left\{ \frac{p(x|w)}{q(x)} - 1 \right\}^2 dx \\ \equiv & C_1' \int \left\{ p(x|w) - q(x) \right\}^2 dx. \end{aligned}$$

Using $\Phi_1(w)$, we obtain

$$\begin{aligned} H_X(\Phi_1(w)) \le & C_1' \int \Bigg\{ \sum_{k=2}^{K^*} \bar{a}_k \big(f(x|\bar{b}_k + b_k^*) - f(x|\bar{b}_1 + b_1^*)\big) \\ & + \sum_{k=2}^{K^*} a_k^* \big(f(x|\bar{b}_k + b_k^*) - f(x|b_k^*)\big) \\ & + \left(1 - \sum_{k=2}^{K^*} a_k^*\right) \big(f(x|\bar{b}_1 + b_1^*) - f(x|b_1^*)\big) \\ & + \sum_{k>K^*}^{K} \bar{a}_k \big(f(x|\bar{b}_k) - f(x|\bar{b}_1 + b_1^*)\big) \Bigg\}^2 dx. \end{aligned}$$

Because $f(x|b_k)$ is a regular model, there is a positive constant C_2 such that

$$H_X(\Phi_1(w)) \le C_2 \left\{ \sum_{k=2}^{K^*} \bar{a}_k^2 + \sum_{k=2}^{K^*} \bar{b}_k^2 + \bar{b}_1^2 + \sum_{k>K^*}^{K} \bar{a}_k^2 \right\} \tag{16}$$

in W_1, where the detailed derivation is in Appendix 2. Let the right-hand side be $H_X^{\text{alg}}(\Phi_1(w))$, and consider a zeta function given by

$$\zeta_2(z) = \int_{\Phi_1(W_1)} H_X^{\text{alg}}(\Phi_1(w))^z \varphi(\Phi_1(w); \eta) d\Phi_1(w).$$

According to Lemma 2, a pole $z=-\mu$ of the zeta function $\zeta_2(z)$ provides bounds for the largest pole of $\zeta_X(z)$, such that $z=-\lambda_X\geq-\mu$. By using a blow-up Φ_3 defined by

$$\begin{aligned}
u_2 &= \bar{a}_2,\\
u_2u_k &= \bar{a}_k \quad (2<k\leq K^*),\\
u_2u_k &= \bar{a}_k \quad (k>K^*),\\
u_2v_{km} &= \bar{b}_{km} \quad (1\leq k\leq K^*, 1\leq m\leq d_c),\\
v_{km} &= \bar{b}_{km} \quad (k>K^*, 1\leq m\leq d_c),
\end{aligned}$$

we obtain

$$\zeta_2(z)=\int_{\Phi_3\Phi_1(W_1)} f_2(\Phi_3\Phi_1(w))u_2^{2z}\varphi(\Phi_3\Phi_1(w);\eta)|u_2|^{K^*-2+K^*d_c+(K-K^*)}d\Phi_3\Phi_1(w),$$

where f_2 is a function of the parameters except for u_2, and the factor on $|u_2|$ is derived from the Jacobian of Φ_3. It is easy to confirm that the Dirichlet prior has a factor $u_2^{(K-K^*)(\eta_1-1)}$. Therefore, $\zeta_2(z)$ has a pole at $z=-\mu=-(K^*-1+K^*d_c)/2-(K-K^*)\eta_1/2$, which proves Lemma 5. □

We are now prepared to prove Theorem 3.

Proof (Theorem 3) According to Theorem 2, it holds that

$$D(n)=(\lambda_{XY}-\lambda_X)\frac{\ln n}{n}+o\left(\frac{\ln n}{n}\right).$$

Combining Lemmas 4 and 5, we obtain

$$\begin{aligned}
D(n)\geq&\left\{\frac{K^*-1+K^*d_c}{2}+(K-K^*)\eta_1\right.\\
&\left.-\frac{K^*-1+K^*d_c}{2}-\frac{(K-K^*)\eta_1}{2}\right\}\frac{\ln n}{n}+o\left(\frac{\ln n}{n}\right)\\
=&\frac{(K-K^*)\eta_1}{2}\frac{\ln n}{n}+o\left(\frac{\ln n}{n}\right),
\end{aligned}$$

which completes the proof. □

Proof of Corollary 1

Proof Since $H_X(w)$ has the bound,

$$H_X(w)\leq C_1\int\left(\frac{p(x|w)}{q(x)}-1\right)^2dx, \qquad (17)$$

Lemma 5 immediately holds. Due to the analytic divergence $H_{XY}(w)$, Lemma 4 also holds. Combining these lemmas, we obtain the same lower bound as Theorem 3. In the Gaussian mixture,

$$\begin{aligned}
H_{XY}(w)&=\sum_{y=1}^{K^*}\int q(x,y)\ln\frac{a_y^*f(x|b_y^*)}{a_yf(x|b_y)}dx\\
&=\sum_{y=1}^{K^*}a_y^*\ln\frac{a_y^*}{a_y}+\sum_{y=1}^{K^*}a_y^*f(x|b_y^*)\ln\frac{f(x|b_y^*)}{f(x|b_y)}dx.
\end{aligned}$$

Because $f(x|b)$ is identifiable, $H_{XY}(w)$ is analytic. Section 7.8 in (Watanabe 2009) shows that $H_X(w)$ has the upper bound expressed as Eq. (17) in the Gaussian mixture, which proves the corollary. □

Appendix 2

This section shows supplementary proofs for some equations in the proof of Theorem 3.

According to the analysis with the Newton diagram (Yamazaki et al. 2010), the following relations hold;

$$w_1 + \left\{h_0(w \setminus w_1) + w_1 h_1(w)\right\}^2 \equiv w_1 + h_0(w \setminus w_1)^2, \tag{18}$$

$$w_1^2 + \left\{h_0(w \setminus w_1) + w_1 h_1(w)\right\}^2 \equiv w_1^2 + h_0(w \setminus w_1)^2, \tag{19}$$

$$w_1 + w_1 h_1(w) \equiv w_1, \tag{20}$$

$$w_1^2 + w_1^2 h_1(w) \equiv w_1^2, \tag{21}$$

where $w = \{w_1, w_2, \ldots, w_d\}$, and h_0 and h_1 are polynomial. Using these relations, we prove Eqs. (15) and (16).

Proof of Equation (15)

Recall that the Kullback–Leibler divergence has the following equivalent expression;

$$\begin{aligned} H_{XY}(w) &= \sum_{k=1}^{K^*} a_k^* \left\{ \ln \frac{a_k^*}{a_k} + \int f(x|b_k^*) \ln \frac{f(x|b_k^*)}{f(x|b_k)} dx \right\} \\ &\equiv \left(1 - \sum_{k=2}^{K^*} a_k^*\right) \ln \frac{1 - \sum_{k=2}^{K^*} a_k^*}{1 - \sum_{k=2}^{K} a_k} + \int f(x|b_1^*) \ln \frac{f(x|b_1^*)}{f(x|b_1)} dx \\ &\quad + \sum_{k=2}^{K^*} \left\{ a_k^* \ln \frac{a_k^*}{a_k} + \int f(x|b_k^*) \ln \frac{f(x|b_k^*)}{f(x|b_k)} dx \right\}. \end{aligned}$$

Based on the transformation $\Phi_1(w)$ and the Taylor expansion of $\ln(1 + \Delta x)$ around $|\Delta x| = 0$, we obtain

$$\begin{aligned} H_{XY}(\Phi_1(w)) &\equiv -\left(1 - \sum_{k=2}^{K^*} a_k^*\right) \ln \left(1 - \sum_{k=2}^{K} \frac{\bar{a}_k}{1 - \sum_{k=2}^{K^*} a_k^*}\right) \\ &\quad + \int f(x|b_1^*) \ln \frac{f(x|b_1^*)}{f(x|b_1^* + \bar{b}_1)} dx \\ &\quad + \sum_{k=2}^{K^*} \left\{ - a_k^* \ln \left(1 + \frac{\bar{a}_k}{a_k^*}\right) + \int f(x|b_k^*) \ln \frac{f(x|b_k^*)}{f(x|b_k^* + \bar{b}_k)} dx \right\} \\ &\equiv \sum_{k=2}^{K} \bar{a}_k - \sum_{k=2}^{K^*} \bar{a}_k \\ &\quad - \frac{1}{2}\left(1 - \sum_{k=2}^{K^*} a_k^*\right)^{-1} \left(\sum_{k=2}^{K} \bar{a}_k\right)^2 + \frac{1}{2} \sum_{k=1}^{K^*} a_k^{*-1} \bar{a}_k^2 + h_r(w) \end{aligned}$$

$$+\sum_{k=1}^{K^*}\int f(x|b_k^*)\ln\frac{f(x|b_k^*)}{f(x|b_k^*+\bar{b}_k)}dx,$$

where $h_r(w)$ includes the higher order terms on $\bar{a}_k$. By applying Eq. (19) to $\bar{a}_k^2$, it holds that

$$\begin{aligned}H_{XY}(\Phi_1(w)) \equiv& \sum_{k=2}^{K}\bar{a}_k-\sum_{k=2}^{K^*}\bar{a}_k\\&+\frac{1}{2}\sum_{k=1}^{K^*}a_k^{*-1}\bar{a}_k^2+h_r(w)\\&+\sum_{k=1}^{K^*}\int f(x|b_k^*)\ln\frac{f(x|b_k^*)}{f(x|b_k^*+\bar{b}_k)}dx\\=&\sum_{k=K^*+1}^{K}\bar{a}_k+\frac{1}{2}\sum_{k=1}^{K^*}a_k^{*-1}\bar{a}_k^2+h_r(w)\\&+\sum_{k=1}^{K^*}\int f(x|b_k^*)\ln\frac{f(x|b_k^*)}{f(x|b_k^*+\bar{b}_k)}dx.\end{aligned}$$

Due to Eqs. (20) and (21), $h_r(w)$ is excluded;

$$H_{XY}(\Phi_1(w)) \equiv \sum_{k=2}^{K^*}\bar{a}_k^2+\sum_{k=K^*+1}^{K}\bar{a}_k+\sum_{k=1}^{K^*}\int f(x|b_k^*)\ln\frac{f(x|b_k^*)}{f(x|b_k^*+\bar{b}_k)}dx.$$

Because $f(x|b_k)$ is regular, it is known that

$$\int f(x|b_k^*)\ln\frac{f(x|b_k^*)}{f(x|b_k^*+\bar{b}_k)}dx \equiv \bar{b}_k^2,$$

which proves that

$$H_{XY}(\Phi_1(w)) \equiv \sum_{k=2}^{K^*}\bar{a}_k^2+\sum_{k=K^*+1}^{K}\bar{a}_k+\sum_{k=1}^{K^*}\bar{b}_k^2.$$

□

Proof of Equation (16)

Recall that the Kullback–Leibler divergence $H_X(\Phi_1(w))$ has the following bound;

$$\begin{aligned}H_X(\Phi_1(w)) \le C_1'\int\Bigg\{&\sum_{k=2}^{K^*}\bar{a}_k\big(f(x|\bar{b}_k+b_k^*)-f(x|\bar{b}_1+b_1^*)\big)\\&+\sum_{k=2}^{K^*}a_k^*\big(f(x|\bar{b}_k+b_k^*)-f(x|b_k^*)\big)\\&+\big(1-\sum_{k=2}^{K^*}a_k^*\big)\big(f(x|\bar{b}_1+b_1^*)-f(x|b_1^*)\big)\end{aligned}$$

$$+\sum_{k>K^*}^{K}\bar{a}_k\big(f(x|\bar{b}_k)-f(x|\bar{b}_1+b_1^*)\big)\Big\}^2dx.$$

In the area $\Phi_1(W_1)$, there is a positive constant C_1'' such that

$$H_X(\Phi_1(w))\le C_1''\Bigg\{\sum_{k=1}^{K^*}\bar{a}_k^2\int\big(f(x|\bar{b}_k+b_k^*)-f(x|\bar{b}_1+b_1^*)\big)^2dx$$
$$+\sum_{k=1}^{K^*}\int\big(f(x|\bar{b}_k+b_k^*)-f(x|b_k^*)\big)^2dx+\sum_{k>K^*}^{K}\bar{a}_k^2\Bigg\}. \qquad (22)$$

The Taylor expansion at $\bar{b}_k$ yields

$$f(x|\bar{b}_k+b_k^*)=f(x|b_k^*)+\bar{b}_k^\top\frac{\partial}{\partial\bar{b}_k}f(x|b_k^*)+\cdots.$$

The second term of the right-hand side in Eq. (22) has the following bound,

$$\sum_{k=1}^{K^*}\int\Big(f(x|\bar{b}_k+b_k^*)-f(x|b_k^*)\Big)^2dx\le C_b\sum_{k=1}^{K^*}\Big\{\bar{b}_k^2+\bar{b}_k^2h_r(\bar{b}_k)\Big\},$$

where C_b is a positive constant and $\bar{b}_k^2h_r(\bar{b}_k)$ stands for the rest of the terms. Based on Eq. (21), the bound has the equivalent form,

$$\sum_{k=1}^{K^*}\Big\{\bar{b}_k^2+\bar{b}_k^2h_r(\bar{b}_k)\Big\}\equiv\sum_{k=1}^{K^*}\bar{b}_k^2,$$

which changes the first term of Eq. (22) into

$$\bar{a}_k^2\int\big(f(x|\bar{b}_k+b_k^*)-f(x|\bar{b}_1+b_1^*)\big)^2dx\equiv\bar{a}_k^2$$

due to Eq. (19). Then, there is a positive constant C_2 such that

$$H_X(\Phi_1(w))\le C_2\Bigg\{\sum_{k=2}^{K^*}\bar{a}_k^2+\sum_{k=2}^{K^*}\bar{b}_k^2+\bar{b}_1^2+\sum_{k>K^*}^{K}\bar{a}_k^2\Bigg\}.$$

□

Appendix 3

Proof of Lemma 1

Proof The calculation is based on the way of the proof of Theorem 3. Define the shift transformation Φ_4 given by

$$\bar{a}=1-a,$$
$$\bar{b}_1=(1-\bar{a})(b_1-b^*)+\bar{a}\bar{b}_2,$$
$$\bar{b}_2=b_2-b^*.$$

This corresponds to focusing on the area $W_1\cup W_2$. Following the calculation of Yamazaki et al. (2010), we obtain

$$H_X(\Phi_4(w)) \equiv \bar{b}_1^2 + \bar{a}^2\bar{b}_2^4.$$

Let the right-hand side be $H_{X2}^{\mathrm{alg}}(w)$, and consider a zeta function given by

$$\zeta_3(z) = \int_{W_1 \cup W_2} H_{X2}^{\mathrm{alg}}(\Phi_4(w))^z \varphi(\Phi_4(w); \eta) d\Phi_4(w).$$

By using a blow-up Φ_5 defined by

$$\begin{aligned} \bar{a} &= v_1 v_2, \\ \bar{b}_1 &= u_1^2 v_1, \\ \bar{b}_2 &= u_1, \end{aligned}$$

we obtain the following expression,

$$\zeta_3(z) = \int_{\Phi_5\Phi_4(W_1 \cup W_2)} f_3(\Phi_5\Phi_4(w)) u_1^{4z} v_1^{2z} \varphi(\Phi_5\Phi_4(w); \eta)|u_1|^2|v_1| d\Phi_5\Phi_4(w),$$

where f_3 is a function of the parameter v_2. The prior has a factor $v_1^{\eta_1-1}$. Therefore, $\zeta_3(z)$ has poles at $z = -3/4$ and $z = -(1+\eta_1)/2$, which are calculated from the factors u_1 and v_1, respectively. Considering the cases $u_1 = 0$ and $v_1 = 0$, we find that the effective area of the pole $z = -3/4$ is W_2 and that of $z = -(1+\eta_1)/2$ is W_1. Due to the symmetry, the area $W_2 \cup W_3$ has the same poles. Then, the largest pole changes at $\eta_1 = 1/2$, where the order of the pole is $m_X = 2$. This completes the proof. □

Proof of Theorem 1

First, we introduce tighter upper bounds on λ_X.

Lemma 6 *Under the same condition as in Theorem 3, it holds that*

$$\lambda_X \le \begin{cases} \frac{K^*-1+K^*d_c}{2} + \frac{K-K^*}{2}\eta_1 & \eta_1 \le d_c, \\ \frac{K^*-1+K^*d_c}{2} + \frac{(K-K^*)d_c}{2} & \eta_1 > d_c. \end{cases}$$

Proof Consider the area W_2, which is the neighborhood of

$$\begin{aligned} a_k &= a_k^* \quad (2 \le k \le K^*) \\ b_{km} &= b_{km}^* \quad (1 \le k \le K^*, 1 \le m \le d_c) \\ b_{km} &= b_{1m}^* \quad (k > K^*, 1 \le m \le d_c). \end{aligned}$$

Let us define the shift transformation Φ_5 given by

$$\begin{aligned} \bar{a}_k &= a_k - a_k^* \quad (2 \le k \le K^*) \\ \bar{b}_{km} &= b_{km} - b_{km}^* \quad (1 \le k \le K^*, 1 \le m \le d_c) \\ \bar{b}_{km} &= b_{km} - b_{1m}^* \quad (k > K^*, 1 \le m \le d_c). \end{aligned}$$

Based on the Taylor expansion of $f(x|\bar{b}_k + b_k^*)$, there is a positive constant C_3 such that

$$H_X(\Phi_5(w)) \le C_3 \left\{ \sum_{k=2}^{K^*} \bar{a}_k^2 + \sum_{k=1}^{K} \bar{b}_k^2 \right\}.$$

Let the right-hand side be $H_{X3}^{\mathrm{alg}}(w)$, and consider a zeta function given by

$$\zeta_4(z) = \int_{\Phi_6(W_{21})} H_{X3}^{\mathrm{alg}}(\Phi_6(w))^z \varphi(\Phi_6(w);\eta) d\Phi_6(w).$$

By using a blow-up Φ_7 defined by

$$\begin{aligned} u_2 &= \bar{a}_2, \\ u_2 u_k &= \bar{a}_k \quad (2 \le k \le K^*), \\ u_k &= \bar{a}_k \quad (k > K^*), \\ u_2 v_{km} &= \bar{b}_{km} \quad (1 \le k \le K, 1 \le m \le d_c), \end{aligned}$$

we obtain the following expression:

$$\zeta_4(z) = \int_{\Phi_7\Phi_6(W_{21})} f_4(\Phi_7\Phi_6(w)) u_2^{2z} \varphi(\Phi_7\Phi_6(w);\eta)|u_2|^{K^*-2+Kd_c} d\Phi_7\Phi_6(w),$$

where f_4 is a function consisting of the parameters except for u_2. Therefore, $\zeta_4(z)$ has a pole at $z = -(K^* - 1 + Kd_c)/2$, which shows that

$$\lambda_X \le \frac{K^* - 1 + K^* d_c}{2} + \frac{(K - K^*)d_c}{2}.$$

Compared to the result of Lemma 5, we find that the bounds are tighter when $\eta_1 > d_c$, which proves the lemma. □

Second, the following lemma shows the lower bound of λ_X;

Lemma 7 *Under the same condition as in Theorem 3, it holds that*

$$\lambda_X > \frac{K^* - 1 + K^* d_c}{2}.$$

Proof We can immediately obtain the inequality based on the minimality condition of $q(x)$ and $d > K^* - 1 + K^* d_c$. □

Last, using these lemmas, we prove Theorem 1. As shown in the proofs of Lemmas 5 and 6, λ_X is a linear function of η_1 due to the factor $a_k^{\eta_1 - 1}$ in the Dirichlet prior. The upper and lower bounds imply that, for η_1 close to zero, there exists a constant α such that

$$\lambda_X = \alpha\eta_1 + \beta,$$

where $\beta = (K^* - 1 + K^* d_c)/2$. Eliminated components appear in $\alpha\eta_1$ since their mixing ratio parameters converge to zero in the effective area, and the prior factor $a_k^{\eta_1 - 1}$ works on the calculation of the pole of $\zeta_X(z)$. The phase in the upper bounds eliminates all redundant components, and the constant term β in the above expression is the same value as that of the bounds. This means that the redundant components are all eliminated in this phase. On the other hand, the upper bounds also indicate that λ_X must be a constant function for a sufficiently large η_1. When there is no linear factor of η_1 in λ_X, all mixing ratio parameters converge to nonzero values; all components are used in this phase. Therefore, we have found the two phases, as desired. □

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, *19*, 716–723.

Allman, E., Matias, C., & Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, *37*, 3099–3132.

Aoyagi, M. (2010). Stochastic complexity and generalization error of a restricted boltzmann machine in Bayesian estimation. *Journal of Machine Learning Research*, *11*, 1243–1272.

Aoyagi, M., & Watanabe, S. (2005). Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, *18*, 924–933.

Atiyah, M. F. (1970). Resolution of singularities and division of distributions. *Communications on Pure and Applied Mathematics*, *23*, 145–150.

Dawid, A. P., & Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, *21*(3), 1272–1317.

Ghosal, S., Ghosh, J. K., & Vaart, A. W. V. D. (2000). Convergence rates of posterior distributions. *Annals of Statistics, 28*, 500–531.

Heckerman, D. (1999). Learning in graphical models. In M. I. Jordan (Ed.), *A tutorial on learning with Bayesian networks* (pp. 301–354). Cambridge, MA, USA: MIT Press.

Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero I. *Annals of Mathematics*, *79*(1), 109–203.

Ibragimov, I. A., & Has' Minskii, R. Z. (1981). *Statistical estimation-asymptotic theory* (Vol. 16). Berlin: Springer.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics, 1*, 38–53.

Nagata, K., & Watanabe, S. (2009). Design of exchange monte carlo method for Bayesian learning in normal mixture models. In: *Proceedings of the 15th international conference on advances in neuro-information processing—Volume Part I* (pp. 696–706). Berlin, Heidelberg: Springer, ICONIP'08.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics, 41*, 370–400.

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

Robert, C. P., & Casella, G. (2005). *Monte Carlo statistical methods (Springer texts in statistics)*. Secaucus, NJ: Springer New York Inc.

Rusakov, D., & Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, *6*, 1–35.

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

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

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. New York, NY: Cambridge University Press.

Yamazaki K (2014) Asymptotic accuracy of distribution-based estimation for latent variables. *Journal of Machine Learning Research*, 13, 3541–3562.

Yamazaki, K., & Kaji, D. (2013). Comparing two Bayes methods based on the free energy functions in Bernoulli mixtures. *Neural Networks*, *44C*, 36–43.

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

Yamazaki, K., & Watanabe, S. (2003b). Stochastic complexity of Bayesian networks. In *Proceedings of UAI*, pp. 592–599.

Yamazaki, K., & Watanabe, S. (2005a). Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, *69*(1–3), 62–84.

Yamazaki, K., & Watanabe, S. (2005b). Singularities in complete bipartite graph-type Boltzmann machines and upper bounds of stochastic complexities. *IEEE Transactions on Neural Networks*, *16*(2), 312–324.

Yamazaki, K., Aoyagi, M., & Watanabe, S. (2010). Asymptotic analysis of Bayesian generalization error with Newton diagram. *Neural Networks*, *23*, 35–43.

Zwiernik, P. (2011). An asymptotic behaviour of the marginal likelihood for general Markov models. *Journal of Machine Learning Research*, *999888*, 3283–3310.