Asymptotic Accuracy of Bayesian Estimation for a Single Latent Variable

Keisuke Yamazaki
k-yam@math.dis.titech.ac.jp
Department of Computational Intelligence and Systems Science,
Tokyo Institute of Technology
G5-19 4259 Nagatsuta, Midori-ku, Yokohama, Japan

Abstract

In data science and machine learning, hierarchical parametric models, such as mixture models, are often used. They contain two kinds of variables: observable variables, which represent the parts of the data that can be directly measured, and latent variables, which represent the underlying processes that generate the data. Although there has been an increase in research on the estimation accuracy for observable variables, the theoretical analysis of estimating latent variables has not been thoroughly investigated. In a previous study, we determined the accuracy of a Bayes estimation for the joint probability of the latent variables in a dataset, and we proved that the Bayes method is asymptotically more accurate than the maximum-likelihood method. However, the accuracy of the Bayes estimation for a single latent variable remains unknown. In the present paper, we derive the asymptotic expansions of the error functions, which are defined by the Kullback-Leibler divergence, for two types of single-variable estimations when the statistical regularity is satisfied. Our results indicate that the accuracies of the Bayes and maximum-likelihood methods are asymptotically equivalent and clarify that the Bayes method is only advantageous for multivariable estimations.

Keywords: unsupervised learning, hierarchical parametric models, latent variable, Bayes estimation

1 Introduction

In machine learning and data science, hierarchical parametric models, such as mixture models, are often used. These models contain two kinds of variables: observable and latent. The observable variables represent the observable, measurable data, while the latent variables express the underlying processes that generate the data. For example, a common hierarchical model is a mixture of Gaussian distributions defined by

$$p(x|w) = \sum_{k=1}^{K} a_k \mathcal{N}(x|\mu_k, \Sigma),$$

where $x \in R^M$ is the observable position, w is the parameter containing a_k and μ_k, $a_k \geq 0$ is the mixing ratio, and $\mathcal{N}(x|\mu, \Sigma)$ is a Gaussian distribution with mean μ and variance-covariance matrix Σ. Let us consider cluster analysis, which is a typical task of unsupervised learning. The observable variable is the data position x, and the latent variable is the ungiven cluster label $k \in \{1, \ldots, K\}$, which indicates to which component/cluster the data belong.

Since the parameter is unknown, in practice, it is often necessary to deal with both the parameter and the observable or the latent variable. The parameter is usually estimated in one of two ways: the maximum-likelihood method or the Bayes method. The maximum-likelihood method estimates the parameter that maximizes the likelihood function, while the Bayes method determines the optimal (posterior) distribution for the parameter.

It has been noted that the hierarchical models include singularities in the parameter space (Amari and Ozeki, 2001; Watanabe, 2001b). At a singular point, the relation between the parameter w and the probability $p(x|w)$ is not one to one, and the Fisher information matrix is not positive definite. Let the K^* component Gaussian mixture be the data-generating distribution, and let the K component mixture be a learning model. The case $K > K^*$ corresponds to a singular case: there are redundant components and their parameters contain singularities. On the other hand, the well-specified case $K = K^*$ does not have singularities, and in the present paper, we call it a regular case.

The estimation of an unseen observable variable is referred to as a prediction. Let a set of the given data be $X^n = \{x_1, \ldots, x_n\}$. The task is to

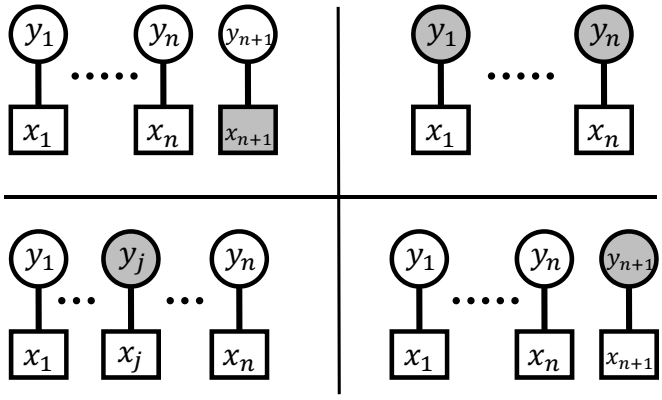

Figure 1: Predictions of observable variables and estimations of latent variables. The observable data are $\{x_1, \ldots, x_n\}$. Rectangles and circles represent the observable and unobservable variables, respectively. Gray nodes are the estimation targets.

predict the next data position based on the given data; this is formulated as the estimation of the probability $p(x_{n+1}|X^n)$. In order to measure the accuracy of the task, we define the error function to be the Kullback-Leibler divergence,

$$E_{X^n}\left[\int q(x_{n+1}) \ln \frac{q(x_{n+1})}{p(x_{n+1}|X^n)} dx_{n+1}\right],$$

where $q(x)$ is the data-generating distribution and $E_{X^n}[\cdot]$ is the expectation over all of the given data. In the example of the Gaussian mixture, the prediction task is to estimate the next unseen data positions.

The estimation of the latent variables is not the same as the prediction task. The target variable of the estimation is unobservable, and in many practical situations, its true value is not given; this makes it difficult to evaluate the result. In a previous study (Yamazaki, 2014), we formulated the accuracy of the latent-variable estimation in a distribution-based manner. The estimation of latent variables is divided into three classes. Let a set of latent variables be $Y^n = \{y_1, \ldots, y_n\}$, where y_i is the corresponding variable to x_i. Figure 1 shows the prediction of observable variables and the three types of estimations of latent variables. Rectangles and circles indicate the observable and latent variables, respectively. The gray nodes are the targets of the estimations. The top left panel shows the prediction, which is expressed

as the estimation of $p(x_{n+1}|X^n)$. The top right panel shows the estimation of the joint probability $p(Y^n|X^n)$, in which all of the latent variables are targets; we will refer to this as Type I. The bottom left panel shows the estimation of the probability of a specific latent variable $p(y_j|X^n)$; we will refer to this as Type II. The bottom right panel shows the estimation of the probability of a latent variable in the unseen data $p(y_{n+1}|X^n)$; we will refer to this as Type III. In the example of a Gaussian mixture, these three types of latent-variable estimation correspond to the cluster analysis process of assigning labels to data.

When the number of data points n is sufficiently large, the form of the error function is referred to as the asymptotic expansion, and the calculation of this form has been exhaustively studied for the prediction process. In the maximum-likelihood method, the asymptotic error is well known, and it has been used as a criterion for selecting models (Akaike, 1974; Takeuchi, 1976; White, 1982). In the Bayes method, the estimation depends on the posterior distribution, and the theoretical properties of its convergence have been studied (Le Cam, 1973; Ghosal *et al.*, 2000; Nguyen, 2013). The normalizing factor of the posterior distribution is the marginal likelihood, and this has a direct relation with the error function (Levin *et al.*, 1990). Since the asymptotic expansion of the marginal likelihood has been derived for the regular case (Schwarz, 1978; Clarke and Barron, 1990), this relation allows us to calculate the asymptotic error. In the singular case, algebraic geometry plays an effective role; in particular, the resolution of singularities (Hironaka, 1964) can be used to clarify the asymptotic marginal likelihood and the asymptotic error (Watanabe, 2001a; Aoyagi and Watanabe, 2004; Yamazaki and Watanabe, 2003; Rusakov and Geiger, 2005; Watanabe, 2009; Zwiernik, 2011; Naito and Yamazaki, 2014).

These studies on the predictive error have focused on the estimation of a single observable variable. Based on their definitions, in the maximum-likelihood method, the error function for the joint probability of multiple variables is equivalent to that for the probability of a single variable. The form of the error of the Bayes method depends on the number of variables. For the regular case, an information criterion that uses the asymptotic error of the joint probability was devised for use with the selection of a Bayesian model (Ando, 2007).

Although there are a number of studies that consider the estimation of observable variables and the convergence of the parameters, the theory of estimating latent variables has not been thoroughly analyzed. The error

functions of Types I, II, and III are defined as the Kullback-Leibler divergence from the data-generating distribution to the estimated one, and its theoretical behavior has been analyzed. The error function of Type III with the maximum-likelihood method has been derived, and a model-selection criterion has been proposed for the regular case (Shimodaira, 1993). The asymptotic expansions of Type I in the Bayes method and of the rest of the types in the maximum-likelihood method have been calculated for the regular case, and we found that with the maximum-likelihood method, their asymptotic errors are equivalent and that for Type I, the Bayes method is more accurate than the maximum-likelihood method. The singular case has been considered, and its error has been derived only for Type I (Yamazaki, 2015).

The asymptotic errors of Types II and III with the Bayes method are as yet unknown in both the regular and the singular cases. Since the asymptotic analysis for these estimations of a single variable requires the calculation of higher-order terms of the marginal likelihood, deriving the asymptotic expansions is not straightforward. In the present paper, we reveal one of the higher-order terms and show the asymptotic errors of Types II and III for the regular case. Comparing the results of this to those of the maximum-likelihood method, we determined that the Bayes method is advantageous only for multivariable estimations, such as those for Type I.

The remainder of this paper is organized as follows: The three types of estimations and their error functions are formally defined in Section 2. The results from our previous study are introduced in Section 3. Section 4 presents the main results on the accuracy of estimations of Types II and III. The advantage of the Bayes estimation is discussed in Section 5.

2 Three Types of Estimations of Latent Variables

In this section, we formulate the three types of estimations of latent variables.

2.1 Formulation of a Hierarchical Probabilistic Model

Let $x \in R^M$ and $y \in \{1, \ldots, K\}$ be observable and latent variables, respectively. The model is represented by

$$p(x, y|w) = p(y|w)p(x|y, w),$$

where the parameter is expressed as $w \in W \subset R^d$. The probabilistic density function of x is then expressed as

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

In the data-generating process of the rightmost expression, we assume that y is selected based on $p(y|w)$, and then x is determined by $p(x|y, w)$. In machine learning, this mixture-type form is used to express many hierarchical models, such as Bayesian networks.

Let $\{X^n, Y^n\} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the i.i.d. data set generated by the true distribution $q(x, y)$. We assume that the true distribution is expressed as

$$q(x, y) = p(x, y|w^*),$$

where w^* is referred to as the true parameter.

2.2 The Three Estimations and their Error Functions

First, we introduce the maximum-likelihood estimator and the posterior distribution, which play important roles in the maximum-likelihood and Bayes methods, respectively. The likelihood function is defined by

$$L(w) = \prod_{i=1}^{n} p(x_i|w).$$

The maximum-likelihood estimator is given by

$$\hat{w} = \arg\max_{w} L(w).$$

In the maximum-likelihood method, this estimator is regarded as the optimal parameter. For example, the prediction of unseen data x is given by

$$p(x|X^n) = p(x|\hat{w}).$$

On the other hand, the Bayes method is defined based on the posterior distribution $p(w|X^n)$. Using a prior distribution $\varphi(w)$, we define the posterior distribution as

$$p(w|X^n) = \frac{1}{Z(X^n)} L(w)\varphi(w),$$

where $Z(X^n)$ is a normalizing factor given by

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

The prediction $p(x|X^n)$ is given by

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

We assume that the true parameter w^* is included in the support of the prior distribution.

Next, for both the maximum-likelihood and Bayes methods, we define the estimated probabilities of the latent variable for each of the three types. For Type I, the given data are $X^n = \{x_1, \ldots, x_n\}$, and the estimation target is $Y^n = \{y_1, \ldots, y_n\}$. The estimated probability of the maximum-likelihood estimation is defined by

$$p(Y^n|X^n) = \prod_{i=1}^{n} p(y_i|x_i, \hat{w}) = \prod_{i=1}^{n} \frac{p(x_i, y_i|\hat{w})}{p(x_i|\hat{w})}.$$

In the Bayes estimation, it is defined by

$$\begin{aligned} p(Y^n|X^n) &= \int \prod_{i=1}^{n} p(y_i|x_i, w)p(w|X^n)dw \\ &= \int \prod_{i=1}^{n} \frac{p(x_i, y_i|w)}{p(x_i|w)} p(w|X^n)dw \\ &= \frac{\int \prod_{i=1}^{n} p(x_i, y_i|w)\varphi(w)dw}{\int \prod_{i=1}^{n} p(x_i|w)\varphi(w)dw}. \end{aligned}$$

For Type II, the given data are X^n, and the estimation target is one of the elements in Y^n. Let the target be $y_j \in Y^n$. The estimated probability of the

maximum-likelihood estimation is defined by

$$p(y_j|X^n) = p(y_j|x_j, \hat{w}) = \frac{p(x_j, y_j|\hat{w})}{p(x_j|\hat{w})}.$$

In the Bayes estimation, it is defined by

$$\begin{aligned} p(y_j|X^n) &= \int p(y_j|x_j, w)p(w|X^n)dw \\ &= \int \frac{p(x_j, y_j|w)}{p(x_j|w)} p(w|X^n)dw \\ &= \frac{\int p(x_j, y_j|w) \prod_{i \neq j} p(x_i|w)\varphi(w)dw}{\int \prod_{i=1}^n p(x_i|w)\varphi(w)dw}. \end{aligned}$$

For Type III, the given data are $X^{n+1} = \{X^n, x_{n+1}\}$, and the estimation target is y_{n+1}. The estimated probability of the maximum-likelihood estimation is defined by

$$p(y_{n+1}|X^{n+1}) = p(y_{n+1}|x_{n+1}, \hat{w}) = \frac{p(x_{n+1}, y_{n+1}|\hat{w})}{p(x_{n+1}|\hat{w})}.$$

In the Bayes estimation, it is defined by

$$\begin{aligned} p(y_{n+1}|X^{n+1}) &= \int p(y_{n+1}|x_{n+1}, w)p(w|X^n)dw \\ &= \frac{\int p(y_{n+1}|x_{n+1}, w) \prod_{i=1}^n p(x_i|w)\varphi(w)dw}{\int \prod_{i=1}^n p(x_i|w)\varphi(w)dw}. \end{aligned}$$

Finally, we define the error functions that measure the accuracy of these estimations, and these are based on the average Kullback-Leibler divergence. In Type I, the true probability of Y^n is expressed by

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

The error function is given by

$$D_{\mathrm{I}}(n) = \frac{1}{n} E_n \left[\ln \frac{q(Y^n|X^n)}{p(Y^n|X^n)} \right],$$

where the expectation is described as

$$E_n[f(X^n, Y^n)] = \int \sum_{y_1=1}^{K} \cdots \sum_{y_n=1}^{K} q(X^n, Y^n) f(X^n, Y^n) dx_1 \dots dx_n.$$

In Types II and III, the error functions are given by

$$\begin{aligned} D_{\mathrm{II}}(n) =& \frac{1}{n} \sum_{j=1}^{n} E_n \left[\ln \frac{q(y_j|x_j)}{p(y_j|X^n)} \right], \\ D_{\mathrm{III}}(n) =& E_{n+1} \left[\ln \frac{q(y_{n+1}|x_{n+1})}{p(y_{n+1}|X^{n+1})} \right], \end{aligned}$$

respectively.

3 Previous Results on Asymptotic Error Functions

This section presents results that we published previously (Yamazaki, 2014). We obtained the asymptotic expansions of $D_{\mathrm{I}}(n)$ for both estimation methods, and the asymptotic expansions of $D_{\mathrm{II}}(n)$ and $D_{\mathrm{III}}(n)$ for the maximum-likelihood estimation. The Fisher information matrices of $p(x, y|w)$ and $p(x|w)$ are defined as

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

respectively. Let $I_{Y|X}(w)$ be their difference:

$$I_{Y|X}(w) = I_{XY}(w) - I_X(w).$$

In the present paper, we assume that these Fisher information matrices exist and that the maximum-likelihood estimator converges almost surely to w^* (Wald, 1949). In other words, the models $p(x, y|w)$ and $p(x|w)$ are regular around w^*, and the estimator is consistent (van der Vaart, 1998).

Because the latent variable is not observable, there is a set of symmetric points W_X^* such that $q(x) = p(x|w_X^*)$ for $w_X^* \in W_X^*$. Note that the true parameter w^* is one of the elements of W_X^*. For example, let us consider the two-component Gaussian mixture given by

$$p(x|w) = a\mathcal{N}(x|\mu_1, \Sigma) + (1-a)\mathcal{N}(x|\mu_2, \Sigma),$$

and let us assume that the true distribution is defined by

$$\begin{aligned} q(x, y=1) =& a^*\mathcal{N}(x|\mu_1^*, \Sigma), \\ q(x, y=2) =& (1-a^*)\mathcal{N}(x|\mu_2^*, \Sigma), \end{aligned}$$

where a^*, μ_1^*, and μ_2^* are constants. This means that the true parameter w^* is described by

$$w^* = (a^*, \mu_1^*, \mu_2^*)^\top.$$

The parameter w_s^* given by

$$w_s^* = (1-a^*, \mu_2^*, \mu_1^*)^\top$$

also satisfies $q(x) = p(x|w_s^*)$, where the labels $y = 1, 2$ are switched. Then, $W_X^* = \{w^*, w_s^*\}$, and we refer to these points as symmetric, since they provide symmetric label assignments. This symmetry appears when $K \geq 2$.

Thus, the maximum-likelihood estimator does not always converge to w^*, and in cluster analysis, this is known as the label-switching problem. To avoid this problem and to theoretically analyze the error function, we consider the case $\hat{w} \to w^*$.

Under the above assumptions, the following theorem has been proven.

Theorem 1 *The error functions have the following asymptotic expansion:*

$$D(n) = \frac{c}{n} + o\left(\frac{1}{n}\right),$$

where $D(n)$ is a general notation for $D_{\mathrm{I}}(n)$, $D_{\mathrm{II}}(n)$, and $D_{\mathrm{III}}(n)$, and the coefficient c for each case is shown in Table 1. The rows indicate the maximum-likelihood (ML) and Bayes methods, respectively. The matrices $I_{XY}(w^*)$, $I_X(w^*)$, and $I_{Y|X}(w^*)$ are abbreviated in a form that does not include the true parameter, i.e., I_{XY}, I_X, or $I_{Y|X}$, respectively.

Table 1: Coefficients of the dominant order $1/n$ in the error functions

	Type I	Type II	Type III
ML	$\mathrm{Tr}[I_{Y\|X} I_X^{-1}]/2$	$\mathrm{Tr}[I_{Y\|X} I_X^{-1}]/2$	$\mathrm{Tr}[I_{Y\|X} I_X^{-1}]/2$
Bayes	$\ln\det[I_{XY} I_X^{-1}]/2$	unknown	unknown

The following corollary compares the two estimation methods with Type I, and shows the advantages of the Bayes estimation.

Corollary 2 *Let the error functions for the maximum-likelihood and Bayes methods be denoted by* $D_{\mathrm{I}}^{\mathrm{ML}}(n)$ *and* $D_{\mathrm{I}}^{\mathrm{Bayes}}(n)$*, respectively. For any true parameter* w^**, there exists a positive constant* c_d *such that*

$$D_{\mathrm{I}}^{\mathrm{ML}}(n) - D_{\mathrm{I}}^{\mathrm{Bayes}}(n) \geq \frac{c_d}{n} + o\left(\frac{1}{n}\right).$$

Corollary 2 indicates that, based on the leading term in the error function, $D_{\mathrm{I}}^{\mathrm{ML}}(n) > D_{\mathrm{I}}^{\mathrm{Bayes}}(n)$ in the asymptotic case of large n.

4 Main Results

This section presents the asymptotic expansions of the error functions for Types II and III.

4.1 Asymptotic Errors of Types II & III in the Bayes Method

Due to the assumptions about the Fisher information matrices and the convergence of the maximum-likelihood estimator, we can determine that

$$|\hat{w} - w^*| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{1}$$

Let $\hat{w}_{n-1}(j)$ be the maximum-likelihood estimator based on the dataset $X^n \setminus x_j$:

$$\hat{w}_{n-1}(j) = \arg\max_w \prod_{i \neq j}^{n} p(x_i|w).$$

In order to simplify the notation, we will use $\hat{w}_{n-1}$ for $\hat{w}_{n-1}(j)$. This estimator also converges to the true parameter, and

$$|\hat{w}_{n-1} - w^*| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{2}$$

Now, we consider the asymptotic expansions of the error functions. In the Bayes method, the error functions $D_{\rm II}(n)$ and $D_{\rm III}(n)$ are written as

$$D_{\rm II}(n) = \frac{1}{n}\sum_{j=1}^{n}\left\{E_n\left[\ln q(x_j, y_j) - \ln q(x_j)\right] + F_1(n) - F_2(n)\right\}, \tag{3}$$

$$D_{\rm III}(n) = E_{n+1}\left[\ln q(y_{n+1}|x_{n+1})\right] + F_3(n) - F_2(n), \tag{4}$$

respectively, where

$$F_1(n) = E_n\left[-\ln \int p(x_j, y_j|w)\prod_{i\neq j}^{n} p(x_i|w)\varphi(w)dw\right],$$

$$F_2(n) = E_n\left[-\ln \int \prod_{i=1}^{n} p(x_i|w)\varphi(w)dw\right],$$

$$F_3(n) = E_{n+1}\left[-\ln \int p(y_{n+1}|x_{n+1}, w)\prod_{i=1}^{n} p(x_i|w)\varphi(w)dw\right].$$

Then, the asymptotic expansions of $F_1(n)$, $F_2(n)$, and $F_3(n)$ are necessary. Let us define the following negative log marginal likelihood:

$$F_\xi(n) = E_z E_n\left[-\ln \int \xi(z|w)\prod_{i=1}^{n} p(x_i|w)\varphi(w)dw\right].$$

This is a unified expression for $F_1(n)$, $F_2(n)$, and $F_3(n)$, in which the function $\xi(z|w)$ will be replaced with the parametric models $p(x_j, y_j|w)$, $p(x_i|w)$, and $p(y_{n+1}|x_{n+1}, w)$, respectively. Thus, we assume that z is independent of x_i. The expectation $E_z[\cdot]$ is based on $\xi(z|w^*)$. In the case of $\xi(z|w) = p(x, y|w)$, the expectation is defined as

$$E_z[f(z)] = \int \sum_{y=1}^{K} f(x, y)p(x, y|w^*)dx,$$

and the function $F_1(n)$ is given by

$$F_1(n) = F_\xi(n-1).$$

The following lemma plays an essential role in the asymptotic analysis of the error functions.

Lemma 3 *The function $F_\xi(n)$ is expressed as*

$$\begin{aligned} F_\xi(n) = &- E_n\left[\sum_{i=1}^{n} \ln p(x_i|\hat{w})\right] + \frac{d}{2}\ln n \\ &- E_z E_n\left[\ln \xi(z|\hat{w})\varphi(\hat{w})\right] \\ &- \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln\det I_X(w^*) \\ &- \frac{1}{2n}\mathrm{Tr}E_n\left[\frac{1}{\varphi(\hat{w})}\frac{\partial^2\varphi(\hat{w})}{\partial w^2}\right] I_X(w^*)^{-1} + o\left(\frac{1}{n}\right). \end{aligned} \tag{5}$$

This form contains terms of the order $1/n$, and this order is higher than the constant terms derived in Clarke and Barron (1990). Note that $\xi(z|w)$ does not affect the $1/n$-order terms. Since the error functions are the differences between the $F_\xi(n)$, as shown in Eqs. 3 and 4, it is easy to see that the error functions depend on only the first and the third terms of Eq. 5, which include the maximum-likelihood estimator $\hat{w}$. This implies a connection between the Bayes method and the maximum-likelihood method.

Based on this lemma, we can prove the following two theorems.

Theorem 4 *Let the error functions for the maximum-likelihood and Bayes methods be denoted by $D_{\mathrm{II}}^{\mathrm{ML}}(n)$ and $D_{\mathrm{II}}^{\mathrm{Bayes}}(n)$, respectively. Asymptotically, they have the following relation:*

$$\begin{aligned} D_{\mathrm{II}}^{\mathrm{Bayes}}(n) =& D_{\mathrm{II}}^{\mathrm{ML}}(n) + o\left(\frac{1}{n}\right) \\ =& \frac{\mathrm{Tr}I_{Y|X}(w^*)I_X(w^*)^{-1}}{2n} + o\left(\frac{1}{n}\right). \end{aligned}$$

Theorem 5 *Let the error functions for the maximum-likelihood and Bayes methods be denoted by $D_{\mathrm{III}}^{\mathrm{ML}}(n)$ and $D_{\mathrm{III}}^{\mathrm{Bayes}}(n)$, respectively. Asymptotically,*

they have the following relation:

$$D_{\mathrm{III}}^{\mathrm{Bayes}}(n) = D_{\mathrm{III}}^{\mathrm{ML}}(n) + o\left(\frac{1}{n}\right) = \frac{\mathrm{Tr} I_{Y|X}(w^*) I_X(w^*)^{-1}}{2n} + o\left(\frac{1}{n}\right).$$

In Types II and III, the asymptotic errors of the Bayes estimation are equivalent to those of the maximum-likelihood estimation. Since Table 1 shows $D_{\mathrm{II}}^{\mathrm{ML}}(n) = D_{\mathrm{III}}^{\mathrm{ML}}(n)$, the errors of Types II and III are also asymptotically the same as those for the Bayes method.

The following corollary summarizes the relative magnitudes of the error functions.

Corollary 6 *Based on the leading terms of the error functions, the relative magnitudes are as follows:*

$$D_{\mathrm{I}}^{\mathrm{Bayes}}(n) < D_{\mathrm{I}}^{\mathrm{ML}}(n) = D_{\mathrm{II}}^{\mathrm{Bayes}}(n) = D_{\mathrm{II}}^{\mathrm{ML}}(n) = D_{\mathrm{III}}^{\mathrm{Bayes}}(n) = D_{\mathrm{III}}^{\mathrm{ML}}(n).$$

Considering these results, in Section 5, we will discuss why the Bayes estimation is more accurate than the maximum-likelihood estimation for the Type I estimation.

4.2 Proof of Lemma 3

Based on a saddle-point approximation, we have

$$F_\xi(n) = E_z E_n \Bigg[-\sum_{i=1}^{n} \ln p(x_i|\hat{w}) - \frac{1}{2} \ln 2\pi \det\{n I_X(w^*)\}^{-1} - \ln \int g(w) \mathcal{N}(w|\hat{w}_n, \{n I_X(w^*)\}^{-1}) dw \Bigg], \tag{6}$$

where $g(w) = \xi(z|w)\varphi(w)e^{r(w)}$ and $r(w) = O_p((w-\hat{w})^3)$. According to Eq. 1 and the asymptotic distribution of $\hat{w}$,

$$
\begin{aligned}
&-\ln \int g(w)\mathcal{N}(w|\hat{w}, \{nI_X(w^*)\}^{-1})dw \\
&= -\ln g(\hat{w}) - \ln \int \left\{1 + \frac{1}{2g(\hat{w})}(w-\hat{w})^\top \frac{\partial^2 g(\hat{w})}{\partial w^2}(w-\hat{w}) + \dots \right\} \\
&\quad \times \mathcal{N}(w|\hat{w}, \{nI_X(w^*)\}^{-1})dw \\
&= -\ln g(\hat{w}) - \ln \left\{1 + \frac{1}{2g(\hat{w})}\frac{1}{n}\mathrm{Tr}\frac{\partial^2 g(\hat{w})}{\partial w^2} I_X(w^*)^{-1} + o_p\left(\frac{1}{n}\right)\right\} \\
&= -\ln g(\hat{w}) - \frac{1}{2g(\hat{w})}\frac{1}{n}\mathrm{Tr}\frac{\partial^2 g(\hat{w})}{\partial w^2} I_X(w^*)^{-1} + o_p\left(\frac{1}{n}\right). \qquad (7)
\end{aligned}
$$

Using the Taylor expansion at w^*, we obtain

$$
\frac{1}{\xi(z|\hat{w})} = \frac{1}{\xi(z|w^*)} + o_p\left(\frac{1}{n}\right).
$$

Considering this equation and the order of $r(\hat{w})$ with Eq. 1, we average Eq. 7:

$$
\begin{aligned}
&-E_z E_n\left[\ln \int g(w)\mathcal{N}(w|\hat{w}, \{nI_X(w^*)\}^{-1})dw\right] \\
&= -E_z E_n\big[\ln \xi(z|\hat{w})\varphi(\hat{w})\big] \\
&\quad - \frac{1}{2n}E_n\left[\frac{1}{\varphi(\hat{w})e^{r(\hat{w})}}\mathrm{Tr}\frac{\partial^2}{\partial w^2}\left\{E_z\left[\frac{\xi(z|\hat{w})}{\xi(z|w^*)}\right]\varphi(\hat{w})e^{r(\hat{w})}\right\} I_X(w^*)^{-1}\right] + o\left(\frac{1}{n}\right) \\
&= -E_z E_n\big[\ln \xi(z|\hat{w})\varphi(\hat{w})\big] - \frac{1}{2n}E_n\left[\frac{1}{\varphi(\hat{w})}\mathrm{Tr}\frac{\partial^2 \varphi(\hat{w})}{\partial w^2} I_X(w^*)^{-1}\right] + o\left(\frac{1}{n}\right),
\end{aligned}
$$

where the last expression is based on $E_z[\frac{\xi(z|w)}{\xi(z|w^*)}] = 1$ for any w. Replacing this expression with the last term of Eq. 6, we obtain Eq. 5.

4.3 Proof of Theorem 4

In the case of $\xi(z|w) = p(x_j, y_j|w)$,

$$
\begin{aligned}
F_1(n) =& E_z E_{n-1}\left[-\ln \int \xi(z|w)\prod_{i\neq j}^{n} p(x_i|w)\varphi(w)dw\right] \\
=& F_\xi(n-1).
\end{aligned}
$$

In the case of $\xi(z|w) = p(x_j|w)$,

$$\begin{aligned} F_2(n) =& E_z E_{n-1}\left[-\ln \int \xi(z|w) \prod_{i\neq j}^{n} p(x_i|w)\varphi(w)dw \right] \\ =& F_\xi(n-1). \end{aligned}$$

Based on Lemma 3, $F_1(n)$ and $F_2(n)$ can be written as

$$\begin{aligned} F_1(n) =& -E_n\left[\sum_{i\neq j}^{n} \ln p(x_i|\hat{w}_{n-1}) \right] + \frac{1}{2}\ln\frac{1}{2\pi}\det\{(n-1)I_X(w^*)\} \\ & - E_n[\ln p(x_j, y_j|\hat{w}_{n-1})\varphi(\hat{w}_{n-1})] \\ & - \frac{1}{2(n-1)}\mathrm{Tr}E_n\left[\frac{1}{\varphi(\hat{w}_{n-1})}\frac{\partial^2\varphi(\hat{w}_{n-1})}{\partial w^2} \right] I_X(w^*)^{-1} + o\left(\frac{1}{n}\right), \\ F_2(n) =& -E_n\left[\sum_{i\neq j}^{n} \ln p(x_i|\hat{w}_{n-1}) \right] + \frac{1}{2}\ln\frac{1}{2\pi}\det\{(n-1)I_X(w^*)\} \\ & - E_n[\ln p(x_j|\hat{w}_{n-1})\varphi(\hat{w}_{n-1})] \\ & - \frac{1}{2(n-1)}\mathrm{Tr}E_n\left[\frac{1}{\varphi(\hat{w}_{n-1})}\frac{\partial^2\varphi(\hat{w}_{n-1})}{\partial w^2} \right] I_X(w^*)^{-1} + o\left(\frac{1}{n}\right), \end{aligned}$$

where we used the asymptotic behavior of $\hat{w}_{n-1}$ and the order shown in Eq. 2.

Using these forms and the relation of Eq. 3, we obtain

$$D_{\mathrm{II}}^{\mathrm{Bayes}}(n) = \frac{1}{n}\sum_{j=1}^{n} E_n\left[\ln\frac{q(y_j|x_j)}{p(y_j|x_j, \hat{w}_{n-1})} \right] + o\left(\frac{1}{n}\right).$$

According to the definition of the error for Type III,

$$E_n\left[\ln\frac{q(y_j|x_j)}{p(y_j|x_j, \hat{w}_{n-1})} \right] = D_{\mathrm{III}}^{\mathrm{ML}}(n-1) + o\left(\frac{1}{n}\right),$$

which is independent of j. Thus,

$$D_{\mathrm{II}}^{\mathrm{Bayes}}(n) = D_{\mathrm{III}}^{\mathrm{ML}}(n-1) + o\left(\frac{1}{n}\right).$$

Using Theorem 1, we can derive the following form:

$$D_{\mathrm{II}}^{\mathrm{Bayes}}(n) = \frac{\mathrm{Tr}I_{Y|X}(w^*)I_X(w^*)^{-1}}{2(n-1)} + o\left(\frac{1}{n}\right),$$

which proves Theorem 4.

4.4 Proof of Theorem 5

In the case of $\xi(z|w) = p(y_{n+1}|x_{n+1}, w)$,

$$\begin{aligned} F_3(n) =& E_z E_n \Big[-\ln \int \xi(z|w) \prod_{i=1}^{n} p(x_i|w)\varphi(w) dw \Big] \\ =& F_\xi(n). \end{aligned}$$

Based on Lemma 3, $F_3(n)$ can be rewritten as

$$\begin{aligned} F_3(n) =& E_{n+1}\Big[-\sum_{i=1}^{n} \ln p(x_i|\hat{w}) - \frac{1}{2}\ln 2\pi \det\{nI_X(w^*)\}^{-1} \\ & - \ln p(y_{n+1}|x_{n+1}, \hat{w})\varphi(\hat{w}) - \frac{1}{2n}\mathrm{Tr}\frac{1}{\varphi(\hat{w})}\frac{\partial^2 \varphi(\hat{w})}{\partial w^2} I_X(w^*)^{-1}\Big] + o\Big(\frac{1}{n}\Big). \end{aligned}$$

Let us assume that the function ξ is a constant $\xi(z|w) = 1$. Then, $F_2(n)$ has another expression;

$$\begin{aligned} F_2(n) =& E_z E_n \Big[-\ln \int 1 \cdot \prod_{i=1}^{n} p(x_i|w)\varphi(w) dw \Big] \\ =& F_\xi(n). \end{aligned}$$

It is easily confirmed that Lemma 3 holds in this case, and $F_2(n)$ has the following form;

$$\begin{aligned} F_2(n) =& E_n\Big[-\sum_{i=1}^{n} \ln p(x_i|\hat{w}) - \frac{1}{2}\ln 2\pi \det\{nI_X(w^*)\}^{-1} \\ & - \ln \varphi(\hat{w}) - \frac{1}{2n}\mathrm{Tr}\frac{1}{\varphi(\hat{w})}\frac{\partial^2 \varphi(\hat{w})}{\partial w^2} I_X(w^*)^{-1}\Big] + o\Big(\frac{1}{n}\Big). \end{aligned}$$

Using these forms and the relation of Eq. 4, we obtain

$$\begin{aligned} D_{\mathrm{III}}^{\mathrm{Bayes}}(n) =& E_{n+1}\Big[\ln \frac{q(y_{n+1}|x_{n+1})}{p(y_{n+1}|x_{n+1}, \hat{w})}\Big] + o\Big(\frac{1}{n}\Big) \\ =& D_{\mathrm{III}}^{\mathrm{ML}}(n) + o\Big(\frac{1}{n}\Big). \end{aligned}$$

According to Theorem 1,

$$D_{\mathrm{III}}^{\mathrm{Bayes}}(n) = \frac{\mathrm{Tr} I_{Y|X}(w^*) I_X(w^*)^{-1}}{2n} + o\Big(\frac{1}{n}\Big),$$

which proves Theorem 5.

5 Discussion

In the previous section, we found that the accuracy of the Bayes estimation was asymptotically equivalent to that of the maximum-likelihood estimation for Types II and III. In this section, we investigate the mathematical reason why the Bayes estimation is advantageous for Type I.

In Section 5.1, Types II and III are extended to multivariable estimations, and their asymptotic errors are introduced. The results indicate that the Bayes method is again more accurate. In Section 5.2, we compare single-variable and multivariable predictions, and we find that the Bayes estimation is advantageous not only when estimating latent variables but also when estimating observable variables. In Section 5.3, we formally decompose the error functions of the multivariable estimations and elucidate the difference between the Bayes and maximum-likelihood methods.

5.1 Other Estimations of Multiple Latent Variables

Let us consider the variants of Types II and III, in which there are multiple estimation targets. We consider a positive constant α, where αn is an integer. We will use the following notation for the data:

$$\begin{aligned} X_1 =& \{x_1, \dots, x_{\alpha n}\}, \\ Y_1 =& \{y_1, \dots, y_{\alpha n}\}, \\ X_2 =& \{x_{n+1}, \dots, x_{n+\alpha n}\}, \\ Y_2 =& \{y_{n+1}, \dots, y_{n+\alpha n}\}. \end{aligned}$$

Definition 7 (Type II′) *Assume that $0 < \alpha < 1$. Let X^n be the observable data, and let Y_1 be the estimation targets. The maximum-likelihood estimation is given by*

$$p(Y_1|X^n) = \prod_{i=1}^{\alpha n} p(y_i|x_i, \hat{w}) = \prod_{i=1}^{\alpha n} \frac{p(x_i, y_i|\hat{w})}{p(x_i|\hat{w})},$$

and the Bayes estimation is given by

$$p(Y_1|X^n) = \frac{\int \prod_{j=1}^{\alpha n} p(x_j, y_j|w) \prod_{i=\alpha n+1}^{n} p(x_i|w)\varphi(w;\eta)dw}{\int \prod_{i=1}^{n} p(x_i|w)\varphi(w;\eta)dw}.$$

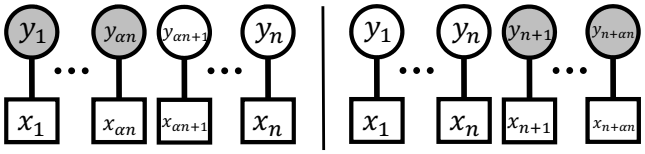

Figure 2: Variants of Types II and III.

In Type II′, the estimation is on the joint probability of Y_1, where $Y^n \setminus Y_1$ is marginalized out. Note that $X^n \setminus X_1 = \{x_{\alpha n+1}, \ldots, x_n\}$ is used for both estimations, since $\hat{w}$ is based on X^n in the maximum-likelihood method and the numerator and the denominator include the factor $\prod_{i=\alpha n+1}^{n} p(x_i|w)$ in the Bayes method. Type II′ lies between Types I and II; it is equivalent to Type I for $\alpha = 1$ and formally converges to Type II as $\alpha \to 1/n$.

Definition 8 (Type III′) *Let X^n and X_2 be the observable data, and let Y_2 be the estimation targets. The maximum-likelihood estimation is given by*

$$p(Y_2|X^n, X_2) = \prod_{i=n+1}^{n+\alpha n} p(y_i|x_i, \hat{w}) = \prod_{i=n+1}^{n+\alpha n} \frac{p(x_i, y_i|\hat{w})}{p(x_i|\hat{w})},$$

and the Bayes estimation is given by

$$p(Y_2|X^n, X_2) = \int \prod_{i=n+1}^{n+\alpha n} \frac{p(x_i, y_i|w)}{p(x_i|w)} p(w|X^n) dw.$$

In Type III′, the estimation is on the joint probability of Y_2. Type III′ formally converges to Type III as $\alpha \to 1/n$.

Figure 2 shows these types; the left panel shows Type II′, which is the multitarget estimation of Type II, and the right panel shows Type III′, which is the multitarget estimation of Type III.

The error functions of Type II′ and III′ are defined by

$$D_{\mathrm{II}'}(n) = \frac{1}{\alpha n} E_{X^n}\left[\sum_{Y_1} q(Y_1|X^n) \ln \frac{q(Y_1|X^n)}{p(Y_1|X^n)}\right],$$

$$D_{\mathrm{III}'}(n) = \frac{1}{\alpha n} E_{X^n, X_2}\left[\sum_{Y_2} q(Y_2|X_2) \ln \frac{q(Y_2|X_2)}{p(Y_2|X_2, X^n)}\right],$$

respectively. In the maximum-likelihood estimation, according to the definitions, $D_{\mathrm{II}'}(n) = D_{\mathrm{II}}(n)$ and $D_{\mathrm{III}'}(n) = D_{\mathrm{III}}(n)$. By comparing $D_{\mathrm{II}'}(n)$ and $D_{\mathrm{III}'}(n)$ with $D_{\mathrm{II}}(n)$ and $D_{\mathrm{III}}(n)$, respectively, we can clarify whether the Bayes method is advantageous for multivariable estimations.

Let us define a mixture of the Fisher information matrices:

$$K_{XY}(w) = \alpha I_{XY}(w) + (1-\alpha) I_X(w).$$

In a previous study (Yamazaki, 2014), we proved the following lemmas.

Lemma 9 *In the Bayes estimation for Type II′, the error function has the following asymptotic expansion:*

$$D_{\mathrm{II}'}^{\mathrm{Bayes}}(n) = \frac{1}{2\alpha n} \ln \det[K_{XY}(w^*) I_X(w^*)^{-1}] + o\left(\frac{1}{n}\right).$$

Lemma 10 *In the Bayes estimation for Type III′, the error function has the following asymptotic expansion:*

$$D_{\mathrm{III}'}^{\mathrm{Bayes}}(n) = \frac{1}{2\alpha n} \ln \det[K_{XY}(w^*) I_X(w^*)^{-1}] + o\left(\frac{1}{n}\right).$$

These lemmas show the following relations, based on the leading terms:

$$D_{\mathrm{II}'}^{\mathrm{Bayes}}(n) < D_{\mathrm{II}'}^{\mathrm{ML}}(n) = D_{\mathrm{II}}^{\mathrm{ML}}(n),$$
$$D_{\mathrm{III}'}^{\mathrm{Bayes}}(n) < D_{\mathrm{III}'}^{\mathrm{ML}}(n) = D_{\mathrm{III}}^{\mathrm{ML}}(n).$$

By comparing these relations with Corollary 6, we see that the Bayes method is advantageous when there are multiple estimation targets:

$$D_{\mathrm{II}'}^{\mathrm{Bayes}}(n) < D_{\mathrm{II}}^{\mathrm{Bayes}}(n),$$
$$D_{\mathrm{III}'}^{\mathrm{Bayes}}(n) < D_{\mathrm{III}}^{\mathrm{Bayes}}(n).$$

5.2 Estimation of Multiple Observable Variables

In the previous subsection, it was proved that the Bayes method was advantageous for all multivariable estimations of latent variables. Let us consider the following two cases for estimating observable variables.

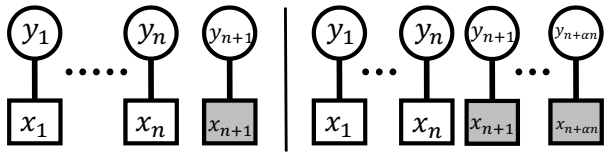

Figure 3: Predictions of a single observable variable (the left panel) and of multiple variables (the right panel).

Definition 11 (Single-target prediction) *Let X^n be the observable data, and let x_{n+1} be the estimation target. The maximum-likelihood estimation is given by*

$$p(x_{n+1}|X^n) = p(x_{n+1}|\hat{w}),$$

and the Bayes estimation is given by

$$p(x_{n+1}|X^n) = \int p(x_{n+1}|w)p(w|X^n)dw.$$

Definition 12 (Multiple-target prediction) *Let X^n be the observable data, and let X_2 be the estimation target. The maximum-likelihood estimation is given by*

$$p(X_2|X^n) = \prod_{i=n+1}^{n+\alpha n} p(x_i|\hat{w}),$$

and the Bayes estimation is given by

$$p(X_2|X^n) = \int \prod_{i=n+1}^{n+\alpha n} p(x_i|w)p(w|X^n)dw.$$

Figure 3 shows these predictions; the left and right panels show the predictions for a single target and for multiple targets, respectively.

The error functions for the single-target prediction (STP) and multiple-target prediction (MTP) are defined by

$$D_{\mathrm{STP}}(n) = E_{n+1}\left[\ln \frac{q(x_{n+1})}{p(x_{n+1}|X^n)}\right],$$
$$D_{\mathrm{MTP}}(n) = \frac{1}{\alpha n} E_{n+\alpha n}\left[\ln \frac{q(X_2)}{p(X_2|X^n)}\right],$$

respectively.

The following lemma shows that we obtain a smaller error with the Bayes method not only when estimating the latent variables but also when estimating the observable variables.

Lemma 13 *The error functions in the predictions have the following asymptotic expansions:*

$$\begin{aligned}
D^{\mathrm{ML}}_{\mathrm{STP}}(n) =& \frac{d}{2n} + o\left(\frac{1}{n}\right), \\
D^{\mathrm{ML}}_{\mathrm{MTP}}(n) =& D^{\mathrm{ML}}_{\mathrm{STP}}(n), \\
D^{\mathrm{Bayes}}_{\mathrm{STP}}(n) =& D^{\mathrm{ML}}_{\mathrm{STP}}(n) + o\left(\frac{1}{n}\right), \\
D^{\mathrm{Bayes}}_{\mathrm{MTP}}(n) =& \frac{\ln(1+\alpha)}{\alpha}\frac{d}{2n} + o\left(\frac{1}{n}\right),
\end{aligned}$$

where d is the dimension of the parameter.

The proofs are given in the Appendix. We can now obtain the following relations, based on the leading terms:

$$D^{\mathrm{Bayes}}_{\mathrm{MTP}}(n) < D^{\mathrm{ML}}_{\mathrm{MTP}}(n) = D^{\mathrm{ML}}_{\mathrm{STP}}(n) = D^{\mathrm{Bayes}}_{\mathrm{STP}}(n).$$

Again, we see that the Bayes estimation is more accurate in the multiple-target case, and its accuracy is equivalent to that of the maximum-likelihood estimation in the single-target case.

5.3 Analysis of the Advantage in Multivariable Estimations

In the Bayesian multivariable estimations, the estimated distributions are defined by the integrals of the parameter, where all target variables are also used for their estimation. For example, the Bayes estimation of the multiple-target prediction is given by

$$\begin{aligned}
p(X_2|X^n) =& \int \prod_{i=n+1}^{n+\alpha n} p(x_i|w)p(w|X^n)dw \\
=& \frac{\int \prod_{i=1}^{n+\alpha n} p(x_i|w)\varphi(w;\eta)dw}{\int \prod_{i=1}^{n} p(x_i|w)\varphi(w;\eta)dw},
\end{aligned}$$

where the numerator includes the likelihood of $X^n \cup X_2 = \{x_1, \ldots, x_{n+\alpha n}\}$. On the other hand, the maximum-likelihood estimation is based on the likelihood of X^n. This implies that the dependent way of the Bayes estimation will be more accurate than repetition of the single-variable estimation.

We can mathematically explain this advantage as follows. In the prediction problem, the MTP error is formally expressed as

$$\begin{aligned}
D_{\mathrm{MTP}}(n) &= \frac{1}{\alpha n} E_{n+\alpha n}\Big[\ln q(X_2) - \ln p(X_2|X^n) \Big] \\
&= \frac{1}{\alpha n} E_{n+\alpha n}\Big[\ln q(X_2) - \ln p(x_{n+1}|X_2 \setminus x_{n+1}, X^n) \\
&\qquad\qquad - \ln p(X_2 \setminus x_{n+1}|X^n) \Big] \\
&= \frac{1}{\alpha n} E_{n+\alpha n}\Big[\ln q(X_2) - \ln p(x_{n+1}|X_2 \setminus x_{n+1}, X^n) \\
&\qquad\qquad - \ln p(x_{n+2}|X_2 \setminus \{x_{n+1}, x_{n+2}\}, X^n) \\
&\qquad\qquad - \ln p(X_2 \setminus \{x_{n+1}, x_{n+2}\}|X^n) \Big] \\
&= \frac{1}{\alpha n} E_{n+\alpha n}\Big[\sum_{i=1}^{\alpha n} \ln q(x_{n+i}) - \ln p(x_{n+1}|X_2 \setminus x_{n+1}, X^n) \\
&\qquad\qquad - \ln p(x_{n+2}|X_2 \setminus \{x_{n+1}, x_{n+2}\}, X^n) \\
&\qquad\qquad \cdots - \ln p(x_{n+\alpha n-1}|x_{n+\alpha n}, X^n) \\
&\qquad\qquad - \ln p(x_{n+\alpha n}|X^n) \Big].
\end{aligned}$$

Then,

$$\begin{aligned}
D_{\mathrm{MTP}}(n) =& \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} D_{\mathrm{MTP},i}(n), \\
D_{\mathrm{MTP},i}(n) =& E_{n+\alpha n}\left[\ln \frac{q(x_{n+i})}{p(x_{n+i}|X_2 \setminus \{x_{n+1}, \ldots, x_{n+i}\}, X^n)} \right].
\end{aligned}$$

Because the maximum-likelihood estimation determines $\hat{w}$ from X^n,

$$D^{\mathrm{ML}}_{\mathrm{MTP},i}(n) = E_{n+\alpha n}\left[\ln \frac{q(x_{n+i})}{p(x_{n+i}|\hat{w})} \right] = D^{\mathrm{ML}}_{\mathrm{STP}}(n),$$

which means that

$$D^{\mathrm{ML}}_{\mathrm{MTP}}(n) = \frac{1}{\alpha n}\sum_{i=1}^{\alpha n} D^{\mathrm{ML}}_{\mathrm{STP}}(n) = D^{\mathrm{ML}}_{\mathrm{STP}}(n).$$

Comparing this with the maximum-likelihood estimation $p(x_{n+i}|X^n) = p(x_{n+i}|\hat{w})$, we find that the Bayes estimation $p(x_{n+i}|X_2 \setminus \{x_{n+1}, \ldots, x_{n+i}\}, X^n)$ uses the additional data set $X_2 \setminus \{x_{n+1}, \ldots, x_{n+i}\}$, which results in a more accurate prediction.

Now, we consider the estimation of latent variables. Let us define the following notation:

$$\begin{aligned}
Y_i^n =& Y^n \setminus \{y_i, \ldots, y_n\} = \{y_1, \ldots, y_{i-1}\},\\
Y_{1,i} =& Y_1 \setminus \{y_i, \ldots, y_{\alpha n}\} = \{y_1, \ldots, y_{i-1}\},\\
Y_{2,i} =& Y_2 \setminus \{y_{n+i}, \ldots, y_{n+\alpha n}\} = \{y_{n+1}, \ldots, y_{n+i-1}\}.
\end{aligned}$$

For example, the estimated probability of Type I can be written as

$$\begin{aligned}
p(Y^n|X^n) =& p(y_n|Y_n^n, X^n)p(Y_n^n|X^n)\\
=& p(y_n|Y_n^n, X^n)p(y_{n-1}|Y_{n-1}^n, X^n)p(Y_{n-1}^n|X^n)\\
=& \prod_{i=1}^{n} p(y_i|Y_i^n, X^n).
\end{aligned}$$

In the same way,

$$\begin{aligned}
p(Y_1|X^n) =& \prod_{i=1}^{\alpha n} p(y_i|Y_{1,i}, X^n),\\
p(Y_2|X^n, X_2) =& \prod_{i=1}^{\alpha n} p(y_{n+i}|Y_{2,i}, X_2, X^n),
\end{aligned}$$

for Type II′ and III′, respectively. Then, the error functions can be rewritten as

$$\begin{aligned}
D_{\mathrm{I}}(n) =& \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{I},i}(n),\\
D_{\mathrm{II}'}(n) =& \frac{1}{\alpha n}\sum_{i=1}^{\alpha n} D_{\mathrm{II}',i}(n),\\
D_{\mathrm{III}'}(n) =& \frac{1}{\alpha n}\sum_{i=1}^{\alpha n} D_{\mathrm{III}',i}(n),
\end{aligned}$$

where

$$
\begin{aligned}
D_{\mathrm{I},i}(n) =& E_n\left[\ln\frac{q(y_i|x_i)}{p(y_i|Y_i^n,X^n)}\right],\\
D_{\mathrm{II}',i}(n) =& E_n\left[\ln\frac{q(y_i|x_i)}{p(y_i|Y_{1,i},X^n)}\right],\\
D_{\mathrm{III}',i}(n) =& E_{n+\alpha n}\left[\ln\frac{q(y_{n+i}|x_{n+i})}{p(y_{n+i}|Y_{2,i},X_2,X^n)}\right],
\end{aligned}
$$

respectively. Note that, in these formal product forms, a target y_i is estimated based on the results of other targets; for example, Type I has the probability $p(y_i|Y_i^n,X^n)$, where y_i depends on the results of $Y_i^n=\{y_1,\ldots,y_{i-1}\}$. However, in the maximum-likelihood method, the estimated probabilities are expressed as

$$
\begin{aligned}
p(y_i|Y_i^n,X^n) =& p(y_i|\hat{w}) = p(y_i|X^n),\\
p(y_i|Y_{1,i},X^n) =& p(y_i|\hat{w}) = p(y_i|X^n),\\
p(y_{n+i}|Y_{2,i},X_2,X^n) =& p(y_{n+i}|x_{n+i},\hat{w}) = p(y_{n+i}|x_{n+i},X^n),
\end{aligned}
$$

respectively, where additional data, such as Y_i^n, $Y_{1,i}$, and $Y_{2,i}$, is ignored. It can easily be found that for $i\geq 2$,

$$
\begin{aligned}
D_{\mathrm{I},i}^{\mathrm{ML}}(n) =& E_n\left[\ln\frac{q(y_i|x_i)}{p(y_i|X^n)}\right]\\
>& E_n\left[\ln\frac{q(y_i|x_i)}{p(y_i|Y_i^n,X^n)}\right] = D_{\mathrm{I},i}^{\mathrm{Bayes}}(n),\\
D_{\mathrm{II}',i}^{\mathrm{ML}}(n) =& E_n\left[\ln\frac{q(y_i|x_i)}{p(y_i|X^n)}\right]\\
>& E_n\left[\ln\frac{q(y_i|x_i)}{p(y_i|Y_{1,i},X^n)}\right] = D_{\mathrm{II}',i}^{\mathrm{Bayes}}(n),\\
D_{\mathrm{III}',i}^{\mathrm{ML}}(n) =& E_{n+\alpha n}\left[\ln\frac{q(y_{n+i}|x_{n+i})}{p(y_{n+i}|X_2,X^n)}\right]\\
>& E_{n+\alpha n}\left[\ln\frac{q(y_{n+i}|x_{n+i})}{p(y_{n+i}|Y_{2,i},X_2,X^n)}\right] = D_{\mathrm{III}',i}^{\mathrm{Bayes}}(n),
\end{aligned}
$$

which shows that the error of the maximum-likelihood method is larger than that of the Bayes method.

Let us consider the single-variable estimations from the perspective of this additional data. In the multivariable estimation, the Bayes method has an advantage, because the error functions defined by the Kullback-Leibler divergence are decomposed into terms such as $D_{\mathrm{I},i}(n)$, $D_{\mathrm{II}',i}(n)$, and $D_{\mathrm{III}',i}(n)$, which express the error on each y_i. Thus, the use of additional data, such as Y_i^n, $Y_{1,i}$, and $Y_{2,i}$, improves the accuracy. Note that these data points are also the estimation targets in other terms. On the other hand, the single-variable estimations do not have any other targets, and thus the error function does not decompose and the Bayes method does not have an advantage. Using Theorems 4 and 5, we quantitatively confirmed that the asymptotic accuracies of the Bayes and maximum-likelihood methods were equal.

6 Conclusion

The present paper derived the asymptotic accuracy of the Bayes latent-variable estimation for Types II and III, which are both single-variable estimations. The results indicate that the accuracy of the Bayes method is equivalent to that of the maximum-likelihood method. This clarifies that the Bayes method is only advantageous for multivariable estimations, such as Types I, II′, and III′.

Acknowledgments

This research was partially supported by the CASIO Science Promotion Foundation and KAKENHI 23500172.

Appendix

Proof of Lemma 13

Since the first equation is a well-known result and is shown in the literature (Akaike, 1974; Watanabe, 2009), we omit the proof.

The second equation is derived from the definitions of the error functions:

$$\begin{aligned}
D_{\mathrm{MTP}}^{\mathrm{ML}}(n) =& \frac{1}{\alpha n} E_{n+\alpha n}\left[\sum_{i=n+1}^{n+\alpha n} \ln \frac{q(x_i)}{p(x_i|\hat{w})}\right] \\
=& \frac{1}{\alpha n} \sum_{i=n+1}^{n+\alpha n} E_{n+\alpha n}\left[\ln \frac{q(x_i)}{p(x_i|\hat{w})}\right] \\
=& \frac{1}{\alpha n} \sum_{i=n+1}^{n+\alpha n} D_{\mathrm{STP}}^{\mathrm{ML}}(n) \\
=& D_{\mathrm{STP}}^{\mathrm{ML}}(n).
\end{aligned}$$

Using the form of $F_2(n)$ shown in Section 4.4, we obtain

$$\begin{aligned}
D_{\mathrm{STP}}^{\mathrm{Bayes}}(n) =& E_{n+1}\big[\ln q(x_{n+1}) + F_2(n+1) - F_2(n)\big] \\
=& E_{n+1}\left[\sum_{i=1}^{n+1} \ln q(x_i) + F_2(n+1)\right] - E_n\left[\sum_{i=1}^{n} \ln q(x_i) + F_2(n)\right] \\
=& \frac{d}{2}\ln(n+1) - \frac{d}{2}\ln n + o\left(\frac{1}{n}\right) \\
=& \frac{d}{2n} + o\left(\frac{1}{n}\right),
\end{aligned}$$

which proves the third equation.

Based on the same form of $F_2(n)$, the last equation is derived as follows:

$$\begin{aligned}
D_{\mathrm{MTP}}^{\mathrm{Bayes}}(n) =& \frac{1}{\alpha n}\left\{E_{n+\alpha n}\left[\sum_{i=1}^{n+\alpha n} \ln q(x_i) + F_2(n+\alpha n)\right]\right. \\
& \left. - E_n\left[\sum_{i=1}^{n} \ln q(x_i) + F_2(n)\right]\right\} \\
=& \frac{1}{\alpha n}\left\{\frac{d}{2}\ln(n+\alpha n) - \frac{d}{2}\ln n\right\} + o\left(\frac{1}{n}\right) \\
=& \frac{\ln(1+\alpha)}{\alpha}\frac{d}{2n} + o\left(\frac{1}{n}\right).
\end{aligned}$$

References

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

Amari, S. and Ozeki, T. (2001). Differential and algebraic geometry of multilayer perceptrons. *IEICE Trans*, **E84-A 1**, 31–38.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, **94**(2), 443–458.

Aoyagi, M. and Watanabe, S. (2004). The generalization error of reduced rank regression in bayesian estimation. In *Proc. of ISITA*, pages 1068–1073.

Clarke, B. and Barron, A. R. (1990). Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, **36**, 453–471.

Ghosal, S., Ghosh, J. K., and Vaart, A. W. V. D. (2000). Convergence rates of posterior distributions. *Ann. Statist*, pages 500–531.

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

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

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

Naito, T. and Yamazaki, K. (2014). Asymptotic marginal likelihood on linear dynamical systems. *IEICE Transactions*, **97-D**(4), 884–892.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist*, pages 370–400.

Rusakov, D. and 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.

Shimodaira, H. (1993). A new criterion for selecting models from partially observed data. In *Oldford, Eds., Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics 89*, pages 381–386. Springer-Verlag.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Science*, **153**, 12–18. in Japanese.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, **20**(4), 595–601.

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

Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, **14**(8), 1049–1060.

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, New York, NY, USA.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**(1), 1–25.

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

Yamazaki, K. (2015). Asymptotic accuracy of Bayes estimation for latent variables with redundancy. *Machine Learning*. to appear.

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

Zwiernik, P. (2011). An asymptotic behaviour of the marginal likelihood for general markov models. *Journal of Machine Learning Research*, **12**, 3283–3310.