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 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 forms of the error functions, which are defined by the Kullback-Leibler divergence, for two types of single-variable estimations. Our results indicate that the accuracies of the Bayes and maximum-likelihood methods are asymptotically equivalent and 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 and hidden Markov models, are often used. These models contain two kinds of variables: observable and latent. The observable variables literally 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

\[
p(x|\{a_k, \mu_k, \Sigma_k\}) = \sum_{k=1}^{K} a_k \mathcal{N}(x|\mu_k, \Sigma_k),
\]

where \(x \in \mathbb{R}^M\) is the observable position, \(a_k \geq 0\) is the mixing ratio, and \(\mathcal{N}(x|\mu, \Sigma)\) is a Gaussian distribution with mean \(\mu\) and variance-covariance matrix \(\Sigma\). In the data-generating process for this model, we assume that the cluster \(y\) is selected based on \(\{a_k\}\), and then the data position \(x\) is determined by \(\mathcal{N}(x|\mu_y, \Sigma_y)\).

The estimation of an unseen observable variable is referred to as a prediction and its statistical properties have been exhaustively studied. Let a set of the given data be \(X^n = \{x_1, \ldots, x_n\}\). The task is to predict the next data position based on the given data, which 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 \(E_{X^n}\left[ \cdot \right]\) is the expectation over all of the given data. When the number of data points \(n\) is sufficiently large, the asymptotic form of this function is revealed in the maximum-likelihood estimation and the Bayes estimation. The maximum-likelihood method determines the optimal parameter, which is called the maximum-likelihood estimator. The asymptotic distribution of the estimator plays an essential role in the prediction accuracy (Akaike, 1974; 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, 1999).
Since the marginal likelihood has a direct connection to the error function, the asymptotic form has been clarified in various hierarchical models (Schwarz, 1978; Watanabe, 2001; Aoyagi and Watanabe, 2004; Yamazaki and Watanabe, 2003; Rusakov and Geiger, 2005; Watanabe, 2009; Zwiernik, 2011; Naito and Yamazaki, 2014).

Although there are a number of studies on estimating observable variables and determining the convergence of the parameters, the theory for estimating latent variables has not been thoroughly analyzed. In a previous study (Yamazaki, 2014), we divided the estimation of latent variables into three classes. Let a set of latent variables be $Y^n = \{y_1, \ldots, y_n\}$, where $(x_i, y_i)$ is a complete pair of data. 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 $p(Y^n|X^n)$, which we refer to as Type I, in which all of the latent variables are targets. The bottom-left panel shows the estimation of the probability of a specific latent variable $p(y_j|X^n)$, which we refer to as Type II. The bottom-right panel shows the estimation of the probability of a latent variable in the unseen data $p(y_{n+1}|X^n)$, which we refer to as Type III. The error function of each of these is defined as
the Kullback-Leibler divergence from the true distribution to the estimated one, and its theoretical behavior has been analyzed (Yamazaki, 2014). The asymptotic forms of Type I in the Bayes method and of all types in the maximum-likelihood method have been calculated, 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 asymptotic errors of Types II and III with the Bayes method are as yet unknown. These types have single estimation targets, and this requires the marginalization of the rest of the latent variables. In the present paper, we consider the estimations for single targets and derive their asymptotic forms.

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 Hierarchical Probabilistic Model

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

$$p(x, y \mid w) = p(y \mid w_m)p(x \mid y, w_c),$$

where the parameter is expressed as $w = \{w_m, w_c\} \in W \subset \mathbb{R}^d$ such that $w_m \cap w_c = \emptyset$. The probabilistic density function of $x$ is then expressed as

$$p(x \mid w) = \sum_{y=1}^{K} p(x, y \mid w) = \sum_{y=1}^{K} p(y \mid w_m)p(x \mid y, w_c).$$
In machine learning, many hierarchical models are expressed in this mixture-type form such as hidden Markov models and Bayesian networks. For example, a Gaussian mixture is defined by

\[ p(x|\omega) = \sum_{y=1}^{K} a_{y}N(x|\mu_{y}, \Sigma_{y}), \]

where the parameter is expressed as \( \omega = \{a_{y}, \mu_{y}, \Sigma_{y}\} \). In this model, \( p(y|w_{m}) = a_{y} \) and \( p(x|y, w_{c}) = N(x|\mu_{y}, \Sigma_{y}) \), where \( w_{m} = \{a_{y}\} \) and \( w_{c} = \{\mu_{y}, \Sigma_{y}\} \).

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

\[ L(\omega) = \prod_{i=1}^{n} p(x_{i}|\omega). \]

The maximum-likelihood estimator is given by

\[ \hat{\omega} = \arg \max_{\omega} L(\omega). \]

Using a prior distribution \( \varphi(\omega) \), we define the posterior distribution of \( \omega \) as

\[ p(\omega|X^{n}) = \frac{1}{Z(X^{n})} L(\omega) \varphi(\omega), \]

where \( Z(X^{n}) \) is a normalizing factor given by

\[ Z(X^{n}) = \int L(\omega) \varphi(\omega) d\omega = \int \prod_{i=1}^{n} p(x_{i}|\omega) \varphi(\omega) d\omega. \]
Assume that the true parameter $w^*$ is included in the support of the posterior distribution.

Next, for both the maximum-likelihood and Bayesian methods, we define the estimated probabilities of the latent variable for each of the three types. For Type I, the given data is $X^n = \{x_1, \ldots, x_n\}$, 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

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

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

$$= \frac{\int \prod_{i=1}^{n} p(x_i, y_i|w)\varphi(w)dw}{\int \prod_{i=1}^{n} p(x_i|w)\varphi(w)dw}.$$  

For Type II, the given data is $X^n$, 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

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

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

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

For Type III, the given data is $X^{n+1} = \{X^n, x_{n+1}\}$, 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
\[
p(y_{n+1}|X^{n+1}) = \int p(y_{n+1}|x_{n+1}, w)p(w|X^n)dw
= \int p(y_{n+1}|x_{n+1}, w)\prod_{i=1}^{n} p(x_i|w)\varphi(w)dw
\]
\[
\int \prod_{i=1}^{n} p(x_i|w)\varphi(w)dw.
\]

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
\[
D_1(n) = \frac{1}{n} \mathbb{E}_n \left[ \ln \frac{q(Y^n|X^n)}{p(Y^n|X^n)} \right],
\]
where the expectation is described as
\[
\mathbb{E}_n[f(X^n, Y^n)] = \int \sum_{y_1=1}^{K} \cdots \sum_{y_n=1}^{K} q(X^n, Y^n) f(X^n, Y^n) dx_1 \cdots dx_n.
\]

In Types II and III, the error functions are given by
\[
D_{II}(n) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_n \left[ \ln \frac{q(y_j|x_j)}{p(y_j|X^n)} \right],
\]
\[
D_{III}(n) = \mathbb{E}_{n+1} \left[ \ln \frac{q(y_{n+1}|x_{n+1})}{p(y_{n+1}|X^{n+1})} \right],
\]
respectively.
3 Previous Results on Asymptotic Error Functions

This section presents results we published previously (Yamazaki, 2014). We obtained the asymptotic forms of $D_1(n)$ for the both estimation methods, and the asymptotic forms of $D_{II}(n)$ and $D_{III}(n)$ for the maximum-likelihood estimation. The Fisher information matrices of $p(x, y|w)$ and $p(x|w)$ are defined as

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

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

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, 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 $q(x|w_X^*) = p(x|w_X^*)$ for $w_X^* \in W_X^*$. Note that the true parameter $w^*$ is one of the elements of $W_X^*$. 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 $\hat{w} \to w_X^* = w^*$.

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

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

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

where $D(n)$ is a general notation for $D_1(n)$, $D_{II}(n)$, and $D_{III}(n)$, 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    | $\text{Tr}[I_{Y|X}I_X^{-1}]/2$ | $\text{Tr}[I_{Y|X}I_X^{-1}]/2$ | $\text{Tr}[I_{Y|X}I_X^{-1}]/2$ |
| Bayes | $\ln \text{det}[I_{XY}I_X^{-1}]/2$ | unknown | unknown |

The following corollary compares the two estimation methods in 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 $D_{\text{ML}}^{I}(n)$ and $D_{\text{Bayes}}^{I}(n)$, respectively. For any true parameter $w^*$, there exists a positive constant $c_d$ such that

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

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

4 Main Results

This section presents the asymptotic forms 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

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

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

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

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

The following two theorems show the asymptotic forms of the error functions.

**Theorem 3** Let the error functions for the maximum-likelihood and Bayes methods be denoted by $D_{\text{II}}^{\text{ML}}(n)$ and $D_{\text{II}}^{\text{Bayes}}(n)$, respectively. Asymptotically, they have the following relation:

$$D_{\text{II}}^{\text{Bayes}}(n) = D_{\text{II}}^{\text{ML}}(n) + o\left(\frac{1}{n}\right)$$

$$= \frac{\text{Tr} I_Y(w^*)I_X(w^*)^{-1}}{2n} + o\left(\frac{1}{n}\right).$$

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

$$D_{\text{III}}^{\text{Bayes}}(n) = D_{\text{III}}^{\text{ML}}(n) + o\left(\frac{1}{n}\right)$$

$$= \frac{\text{Tr} I_Y(w^*)I_X(w^*)^{-1}}{2n} + o\left(\frac{1}{n}\right).$$

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 $D_{\text{II}}^{\text{ML}}(n) = D_{\text{III}}^{\text{ML}}(n)$, 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 5** Based on the leading terms of the error functions, the relative magnitudes are as follows:

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

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 Theorem

The error function can be rewritten as

\[ D_{\text{Bayes}}(n) = \frac{1}{n} \sum_{j=1}^{n} \{ E_n[\ln q(x_j, y_j) - \ln q(x_j)] + F_1(n) - F_2(n) \}, \]

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

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

Based on a saddle-point approximation, we have

\[ F_1(n) = E_n \left[ - \sum_{i \neq j} \ln p(x_i | \hat{w}_{n-1}) - \frac{1}{2} \ln 2\pi \det \{(n-1)I_X(w^*)\}^{-1} \right. \]

\[ - \ln \int p(x_j, y_j | w) \varphi(w) e^{r_1(w)} N(w | \hat{w}_{n-1}, \{(n-1)I_X(w^*)\}^{-1}) dw \],

\[ F_2(n) = E_n \left[ - \sum_{i=1}^{n} \ln p(x_i | \hat{w}) - \frac{1}{2} \ln 2\pi \det \{nI_X(w^*)\}^{-1} \right. \]

\[ - \ln \int \varphi(w) e^{r_2(w)} N(w | \hat{w}, \{nI_X(w^*)\}^{-1}) dw \],

where \( r_1(w) = O_p((w - \hat{w}_{n-1})^3) \), \( r_2(w) = O_p((w - \hat{w})^3) \). Using the Taylor expansion at \( \hat{w}_{n-1} \) and \( \hat{w} \), we obtain

\[ F_1(n) = E_n \left[ - \sum_{i \neq j} \ln p(x_i | \hat{w}_{n-1}) - \frac{1}{2} \ln 2\pi \det \{(n-1)I_X(w^*)\}^{-1} \right. \]

\[ - \ln \int p(x_j, y_j | w) \varphi(w) e^{r_1(w)} N(w | \hat{w}_{n-1}, \{(n-1)I_X(w^*)\}^{-1}) dw \],

\[ F_2(n) = E_n \left[ - \sum_{i=1}^{n} \ln p(x_i | \hat{w}) - \frac{1}{2} \ln 2\pi \det \{nI_X(w^*)\}^{-1} \right. \]

\[ - \ln \int \varphi(w) e^{r_2(w)} N(w | \hat{w}, \{nI_X(w^*)\}^{-1}) dw \].
respectively. Based on Eq. \ref{eq:asymptotic} and the asymptotic distribution of $\hat{w}_{n-1}$, the last term of $F_1(n)$ can be expressed as

$$- \ln f_1 - \ln \int \left\{ 1 + \frac{1}{2f_1} (w - \hat{w}_{n-1})^T \frac{\partial^2 f_1}{\partial w^2} (w - \hat{w}_{n-1}) + \ldots \right\} \times N(w|\hat{w}_{n-1}, \{(n-1)I_X(w^*)\}^{-1})dw$$

$$= - \ln f_1 - \ln \left\{ 1 + \frac{1}{2f_1} \frac{1}{n-1} \text{Tr} \frac{\partial^2 f_1}{\partial w^2} I_X(w^*)^{-1} + o_p \left( \frac{1}{n} \right) \right\}$$

$$= - \ln f_1 - \frac{1}{2(n-1)} \text{Tr} E_n \left[ \frac{1}{\varphi(\hat{w}_{n-1})} \frac{\partial^2 \varphi(\hat{w}_{n-1})}{\partial w^2} \right] I_X(w^*)^{-1} + o_p \left( \frac{1}{n} \right),$$

where $f_1 = p(x_j, y_j|\hat{w}_{n-1})\varphi(\hat{w}_{n-1})e^{\tau_1(\hat{w}_{n-1})}$. Using the Taylor expansion at $\hat{w}_{n-1} = w^*$, we obtain

$$\frac{1}{p(x_j, y_j|\hat{w}_{n-1})} = \frac{1}{p(x_j, y_j|w^*)} + o_p(1).$$

The average of the last term can be written as

$$- E_n \left[ \ln \int p(x_j, y_j|w)\varphi(w)e^{\tau_1(w)}N(w|\hat{w}_{n-1}, \{(n-1)I_X(w^*)\}^{-1})dw \right]$$

$$= - E_n [\ln f_1] - \frac{1}{2(n-1)} \text{Tr} E_n \left[ \frac{1}{\varphi(\hat{w}_{n-1})} \frac{\partial^2 \varphi(\hat{w}_{n-1})}{\partial w^2} \right] I_X(w^*)^{-1} + o \left( \frac{1}{n} \right).$$

Therefore, the asymptotic form of $F_1(n)$ can be written as

$$F_1(n) = - E_n \left[ \sum_{i \neq j}^n \ln p(x_i|\hat{w}_{n-1}) \right] + \frac{1}{2} \ln 2\pi \det \{(n-1)I_X(w^*)\} \right.$$  

$$- E_n \left[ \ln p(x_j, y_j|\hat{w}_{n-1})\varphi(\hat{w}_{n-1}) \right]$$

$$- \frac{1}{2(n-1)} \text{Tr} E_n \left[ \frac{1}{\varphi(\hat{w}_{n-1})} \frac{\partial^2 \varphi(\hat{w}_{n-1})}{\partial w^2} \right] I_X(w^*)^{-1} + o \left( \frac{1}{n} \right).$$

Now, we consider the asymptotic form of $F_2(n)$. Due to Eq. \ref{eq:asymptotic} and the asymptotic distribution of $\hat{w}$, the last term of $F_2(n)$ can be written as

$$- E_n \left[ \ln \int \varphi(w)e^{\tau_2(w)}N(w|\hat{\omega}, \{nI_X(w^*)\}^{-1})dw \right]$$

$$= - E_n [\ln f_2] - \frac{1}{2n} \text{Tr} E_n \left[ \frac{1}{\varphi(\hat{\omega})} \frac{\partial^2 \varphi(\hat{\omega})}{\partial w^2} \right] I_X(w^*)^{-1} + o \left( \frac{1}{n} \right),$$
where \( f_2 = \varphi(\hat{w})e^{r_2(\hat{w})} \). Then, \( F_2(n) \) can be written as

\[
F_2(n) = -E_n \left[ \sum_{i=1}^{n} \ln p(x_i|\hat{w}) \right] + \frac{1}{2} \ln 2\pi \det \{ nI_X(w^*) \} \\
- E_n[\ln \varphi(\hat{w})] - \frac{1}{2n} \text{Tr} E_n \left[ \frac{1}{\varphi(\hat{w})} \frac{\partial^2 \varphi(\hat{w})}{\partial \hat{w}^2} \right] I_X(w^*)^{-1} + o\left( \frac{1}{n} \right).
\]

Using these asymptotic forms, we obtain

\[
D_{\text{Bayes}}^\Pi(n) = \frac{1}{n} \sum_{j=1}^{n} E_n \left[ \sum_{i \neq j} \ln q(x_i) + \ln q(x_j, y_j) - F_1(n) \right] \\
- E_n \left[ \sum_{i=1}^{n} \ln q(x_i) - F_2(n) \right] \\
= \frac{1}{n} \sum_{j=1}^{n} E_n \left[ \sum_{i \neq j} \ln \frac{q(x_i)}{p(x_i|\hat{w}_{n-1})} + \ln \frac{q(x_j, y_j)}{p(x_j, y_j|\hat{w})} \right] \\
- E_n \left[ \sum_{i=1}^{n} \ln \frac{q(x_i)}{p(x_i|\hat{w})} \right] + o\left( \frac{1}{n} \right).
\]

Based on the asymptotic property of the training error in the maximum-likelihood method [Akaike, 1974], there is a constant \( c_{\text{ML}} \) such that

\[
E_n \left[ \sum_{i \neq j} \ln \frac{q(x_i)}{p(x_i|\hat{w}_{n-1})} \right] = -d + \frac{c_{\text{ML}}}{n - 1} + o\left( \frac{1}{n} \right),
\]

\[
E_n \left[ \sum_{i=1}^{n} \ln \frac{q(x_i)}{p(x_i|\hat{w})} \right] = -d + \frac{c_{\text{ML}}}{n} + o\left( \frac{1}{n} \right).
\]

Then, we can rewrite the error as

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

According to the definition of the error in Type III,

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

13
which is independent of \( j \). Thus,
\[
D_{II}^{\text{Bayes}}(n) = D_{III}^{\text{ML}}(n - 1) + o\left(\frac{1}{n}\right).
\]
Using Theorem \[ \] we can derive the following form:
\[
D_{II}^{\text{Bayes}}(n) = \frac{\text{Tr} I_Y X(w^*) I_X(w^*)^{-1}}{2(n - 1)} + o\left(\frac{1}{n}\right),
\]
which proves Theorem \[ \]

4.3 Proof of Theorem \[ \]

The error function can be rewritten as
\[
D_{III}^{\text{Bayes}}(n) = E_{n+1}\left[ \ln q(y_{n+1}|x_{n+1}) + F_3(n) - F_2(n) \right],
\]
\[
F_3(n) = E_n \left[ - \ln \int p(y_{n+1}|x_{n+1}, w) \prod_{i=1}^n p(x_i|w) \varphi(w) dw \right].
\]
Based on a saddle-point approximation,
\[
F_3(n) = E_{n+1} \left[ - \sum_{i=1}^n \ln p(x_i|\hat{w}) - \frac{1}{2} \ln 2\pi \det \{nI_X(w^*)\}^{-1}
\]
\[
- \ln \int p(y_{n+1}|x_{n+1}, w) \varphi(w) e^{r_3(w)} N(w|\hat{w}, \{nI_X(w^*)\}^{-1}) dw \right],
\]
where \( r_3(w) = O_p((w - \hat{w}_{n-1})^3) \). In the same way as we did in the proof of Theorem \[ \] we obtain
\[
F_3(n) = E_{n+1} \left[ - \sum_{i=1}^n \ln p(x_i|\hat{w}) - \frac{1}{2} \ln 2\pi \det \{nI_X(w^*)\}^{-1}
\]
\[
- \ln f_3 - \frac{1}{2n} \text{Tr} \frac{1}{\varphi(\hat{w})} \partial^2 \phi(\hat{w}) \partial w^2 I_X(w^*)^{-1} \right] + o\left(\frac{1}{n}\right),
\]
where \( f_3 = p(y_{n+1}|x_{n+1}, \hat{w}) \varphi(\hat{w}) e^{r_3(\hat{w})} \). Replacing \( F_2(n) \) and \( F_3(n) \) with their asymptotic forms, we can rewrite the error as
\[
D_{III}^{\text{Bayes}}(n) = E_{n+1} \left[ \ln \frac{q(y_{n+1}|x_{n+1})}{p(y_{n+1}|x_{n+1}, w)} \right] + o\left(\frac{1}{n}\right)
\]
\[
= D_{III}^{\text{ML}}(n) + o\left(\frac{1}{n}\right).
\]
According to Theorem \[ \text{III} \]
\[
D_{\text{Bayes}}^{\text{III}}(n) = \frac{\text{Tr} I_Y w^* I_X w^*}{2n} + o\left(\frac{1}{n}\right),
\]
which proves Theorem \[ \text{IV} \]

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. Assume that \( an \) is a positive integer, where \( 0 < \alpha < 1 \).

We will use the following notation for the data:

\[
\begin{align*}
X_1 &= \{x_1, \ldots, x_{an}\}, \\
Y_1 &= \{y_1, \ldots, y_{an}\}, \\
X_2 &= \{x_{n+1}, \ldots, x_{n+an}\}, \\
Y_2 &= \{y_{n+1}, \ldots, y_{n+an}\}.
\end{align*}
\]

**Definition 6 (Type II')** Let \( X^n \) be the observable data, and let \( Y_1 \) be the estimation targets. The maximum-likelihood estimation is given by

\[
p(Y_1|X^n) = \prod_{i=1}^{an} p(y_i|x_i, \hat{w}) = \prod_{i=1}^{an} \frac{p(x_i, y_i|\hat{w})}{p(x_i|\hat{w})}.
\]
and the Bayes estimation is given by
\[ p(Y_1|X^n) = \frac{\int \prod_{j=1}^{\alpha n} p(x_j, y_j|w) \prod_{i=\alpha n+1}^{n} p(x_i|w) \varphi(w; \eta)dw}{\int \prod_{i=1}^{n} p(x_i|w) \varphi(w; \eta)dw}. \]

In Type II', the estimation is on the joint probability of \( Y_1 \), where \( Y \setminus Y_1 \) is marginalized out.

**Definition 7 (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
\[
p(Y_2|X^n, X^2) = \prod_{i=\alpha n+1}^{n+\alpha n} p(y_i|x_i, \hat{w}) = \prod_{i=\alpha n+1}^{n+\alpha n} \frac{p(x_i, y_i|\hat{w})}{p(x_i|\hat{w})},
\]
and the Bayes estimation is given by
\[
p(Y_2|X^n, X^2) = \int \prod_{i=\alpha n+1}^{n+\alpha n} p(x_i, y_i|w) p(w|X^n)dw.
\]

In Type III', the estimation target is extended to \( Y_2 \).

Figure 2 shows these types; the left panel shows Type II', which is the multi-target estimation of Type II, and the right panel shows Type III', which is the multi-target estimation of Type III.

The error functions of Type II' and III' are defined by
\[
D_{II'}(n) = \frac{1}{\alpha n} E_{X^n} \left[ \sum_{Y_1} q(Y_1|X^n) \ln \frac{q(Y_1|X^n)}{p(Y_1|X^n)} \right],
\]
\[
D_{III'}(n) = \frac{1}{\alpha n} E_{X^n, X^2} \left[ \sum_{Y_2} q(Y_2|X_2) \ln \frac{q(Y_2|X_2)}{p(Y_2|X_2, X^n)} \right],
\]
respectively.
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 8** In the Bayes estimation for Type II, the error function has the following asymptotic form:

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

**Lemma 9** In the Bayes estimation for Type III, the error function has the following asymptotic form:

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

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

\[ D_{\text{Bayes}}^{\text{II}}(n) < D_{\text{ML}}^{\text{II}}(n), \]
\[ D_{\text{Bayes}}^{\text{III}}(n) < D_{\text{ML}}^{\text{III}}(n). \]

By comparing these relations with Corollary 5, we see that the Bayes method is advantageous when there are multiple estimation targets.

### 5.2 Estimation of Multiple Observable Variables

The Bayes method is advantageous for all multivariable estimations, both of latent variables and observable variables. Let us consider the following two cases for estimating observable variables.

**Definition 10 (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 11 (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

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

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

respectively.

Lemma 12 The error functions in the predictions have the following asymp-
totic forms,

\[ D_{\text{ML STP}}(n) = \frac{d}{2n} + o\left(\frac{1}{n}\right), \]
\[ D_{\text{ML MTP}}(n) = D_{\text{ML STP}}(n), \]
\[ D_{\text{Bayes STP}}(n) = D_{\text{ML STP}}(n) + o\left(\frac{1}{n}\right), \]
\[ D_{\text{Bayes MTP}}(n) = \frac{\ln(1 + \alpha)}{\alpha} \frac{d}{2n} + o\left(\frac{1}{n}\right), \]

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:

\[ D_{\text{Bayes MTP}}(n) < D_{\text{ML MTP}}(n) = D_{\text{ML STP}}(n) = D_{\text{Bayes STP}}(n). \]

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

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

\[ D_{\text{MTP}}(n) = \frac{1}{\alpha n} E_{n+\alpha n} \left[ \ln q(X_2) - \ln p(X_2|X^n) \right] \]

\[ = \frac{1}{\alpha n} E_{n+\alpha n} \left[ \ln q(X_2) - \ln p(x_{n+1}|X_2 \setminus x_{n+1}, X^n) \right. \]
\[ - \ln p(X_2 \setminus x_{n+1}|X^n) \]
\[ = \frac{1}{\alpha n} E_{n+\alpha n} \left[ \ln q(X_2) - \ln p(x_{n+1}|X_2 \setminus x_{n+1}, X^n) \right. \]
\[ - \ln p(x_{n+2}|X_2 \setminus \{x_{n+1}, x_{n+2}\}, X^n) \]
\[ - \ln p(X_2 \setminus \{x_{n+1}, x_{n+2}\}|X^n) \]
\[ = \frac{1}{\alpha n} E_{n+\alpha n} \left[ \sum_{i=1}^{\alpha n} \ln q(x_{n+i}) - \ln p(x_{n+1}|X_2 \setminus x_{n+1}, X^n) \right. \]
\[ - \ln p(x_{n+2}|X_2 \setminus \{x_{n+1}, x_{n+2}\}, X^n) \]
\[ \cdots - \ln p(x_{n+\alpha n-1}|x_{n+\alpha n}, X^n) \]
\[ - \ln p(x_{n+\alpha n}|X^n) \].

Then,

\[ D_{\text{MTP}}(n) = \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} D_{\text{MTP},i}(n), \]

\[ D_{\text{MTP},i}(n) = E_{n+\alpha n} \left[ \ln \frac{q(x_{n+i})}{p(x_{n+i}|X_2 \setminus \{x_{n+1}, \ldots, x_{n+i}\}, X^n)} \right]. \]

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

\[ D_{\text{MTP},i}^{\text{ML}}(n) = E_{n+\alpha n} \left[ \ln \frac{q(x_{n+i})}{p(x_{n+i} | \hat{w})} \right] = D_{\text{LTR},i}^{\text{ML}}(n), \]

20
which means that
\[ D_{\text{ML}}^{\text{MTP}}(n) = \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} D_{\text{ML}}^{\text{STP}}(n) = D_{\text{STP}}^{\text{ML}}(n). \]

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:
\[
Y^n = Y^n \setminus \{y_i, \ldots, y_n\} = \{y_1, \ldots, y_{i-1}\},
Y_1 = Y_1 \setminus \{y_i, \ldots, y_{\alpha n}\} = \{y_1, \ldots, y_{i-1}\},
Y_2 = Y_2 \setminus \{y_{n+i}, \ldots, y_{n+\alpha n}\} = \{y_{n+1}, \ldots, y_{n+i-1}\}.
\]

For example, the estimated probability of Type I can be written as
\[
p(Y^n|X^n) = p(y_n|Y^n_n, X^n)p(Y^n_n|X^n)
= p(y_n|Y^n_n, X^n)p(y_{n-1}|Y^n_{n-1}, X^n)p(Y^n_{n-1}, X^n)
= \prod_{i=1}^{n} p(y_i|Y^n_i, X^n).
\]

In the same way,
\[
p(Y_1|X^n) = \prod_{i=1}^{\alpha n} p(y_i|Y^n_{i,1}, X^n),
p(Y_2|X^n, X_2) = \prod_{i=1}^{\alpha n} p(y_{n+i}|Y^n_{2,i}, X_2, X^n),
\]
for Type II’ and III’, respectively. Then, the error functions can be rewritten as
\[
D_I(n) = \frac{1}{n} \sum_{i=1}^{n} D_{I,i}(n),
D_{II'}(n) = \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} D_{II',i}(n),
D_{III'}(n) = \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} D_{III',i}(n),
\]
where

\[ D_{I,i}(n) = E_n \left[ \ln \frac{q(y_i|x_i)}{p(y_i|Y^n_i, X^n)} \right], \]

\[ D_{IV,i}(n) = E_n \left[ \ln \frac{q(y_i|x_i)}{p(y_i|Y^n_{1,i}, X^n)} \right], \]

\[ D_{III,i}(n) = E_{n+\alpha} \left[ \ln \frac{q(y_{n+i}|x_{n+i})}{p(y_{n+i}|Y^n_{2,i}, X^n)} \right], \]

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 \( p(y_i|Y^n_i, X^n) \), where \( y_i \) depends on the results of \( Y^n_i = \{y_1, \ldots, y_{i-1}\} \).

However, in the maximum-likelihood method, the estimated probabilities are expressed as

\[ p(y_i|Y^n_i, X^n) = p(y_i|\hat{w}) = p(y_i|X^n), \]

\[ p(y_i|Y^n_{1,i}, X^n) = p(y_i|\hat{w}) = p(y_i|X^n), \]

\[ p(y_{n+i}|Y^n_{2,i}, X^n) = p(y_{n+i}|\hat{w}, X^n) = p(y_{n+i}|x_{n+i}, X^n), \]

respectively, where additional data, such as \( Y^n_i, Y^n_{1,i}, Y^n_{2,i} \), is ignored. It can be easily found that

\[ D_{I,i}^{ML}(n) = E_n \left[ \ln \frac{q(y_i|x_i)}{p(y_i|X^n)} \right] > E_n \left[ \ln \frac{q(y_i|x_i)}{p(y_i|Y^n_i, X^n)} \right] = D_{I,i}^{Bayes}(n), \]

\[ D_{IV,i}^{ML}(n) = E_n \left[ \ln \frac{q(y_i|x_i)}{p(y_i|X^n)} \right] > E_n \left[ \ln \frac{q(y_i|x_i)}{p(y_i|Y^n_{1,i}, X^n)} \right] = D_{IV,i}^{Bayes}(n), \]

\[ D_{III,i}^{ML}(n) = E_{n+\alpha} \left[ \ln \frac{q(y_{n+i}|x_{n+i})}{p(y_{n+i}|X^n)} \right] > E_{n+\alpha} \left[ \ln \frac{q(y_{n+i}|x_{n+i})}{p(y_{n+i}|Y^n_{2,i}, X^n)} \right] = D_{III,i}^{Bayes}(n), \]

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 $D_{I,i}(n)$, $D_{II,i}(n)$, and $D_{III,i}(n)$, which express the error on each $y_i$. Thus, the use of additional data, such as $Y_i, Y_{1,i}$, 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. Theorems 3 and 4 confirm that the asymptotic accuracies of the Bayes and maximum-likelihood methods are equal.

6 Conclusion

The present paper derived the asymptotic forms of the 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'.

Acknowledgment

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

Appendix

Proof of Lemma 12

Since the first equation is a well-known result, which is shown in (Akaike, 1974; Watanabe, 2009), we omit the proof.
The second equation is derived from the definitions of the error functions:

\[
D_{\text{ML}}(n) = \frac{1}{\alpha n} E_{n+\alpha n} \left[ \sum_{i=n+1}^{n+\alpha n} \ln \frac{q(x_i)}{p(x_i|\hat{w})} \right] \\
= \frac{1}{\alpha n} \sum_{i=n+1}^{n+\alpha n} E_{n+\alpha n} \left[ \ln \frac{q(x_i)}{p(x_i|\hat{w})} \right] \\
= \frac{1}{\alpha n} \sum_{i=n+1}^{n+\alpha n} D_{\text{ML}}^n(n) \\
= D_{\text{ML}}(n).
\]

Using \( F_2(n) \), we obtain

\[
D_{\text{Bayes}}^n = E_{n+1} \left[ \ln q(x_{n+1}) + F_2(n+1) - F_2(n) \right] \\
= E_{n+1} \left[ \sum_{i=1}^{n+1} \ln q(x_i) + F_2(n+1) \right] - E_n \left[ \sum_{i=1}^{n} \ln q(x_i) + F_2(n) \right] \\
= - \frac{d}{2(n+1)} + \frac{d}{2} \ln(n+1) + \frac{d}{2n} - \frac{d}{2} \ln n + o \left( \frac{1}{n} \right) \\
= \frac{d}{2n} + o \left( \frac{1}{n} \right),
\]

which proves the third equation.

Based on \( F_2(n) \), the last equation is derived as follows:

\[
D_{\text{Bayes}}^n = \frac{1}{\alpha n} \left\{ E_{n+\alpha n} \left[ \sum_{i=1}^{n+\alpha n} \ln q(x_i) + F_2(n+\alpha n) \right] \\
- E_n \left[ \sum_{i=1}^{n} \ln q(x_i) + F_2(n) \right] \right\} \\
= \frac{1}{\alpha n} \left\{ - \frac{d}{2(1+\alpha)n} + \frac{d}{2} \ln(n+\alpha n) + \frac{d}{2n} - \frac{d}{2} \ln n \right\} + o \left( \frac{1}{n} \right) \\
= \ln(1+\alpha) \frac{d}{\alpha} \frac{1}{2n} + o \left( \frac{1}{n} \right).
\]
References

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

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

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

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

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.

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

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*, 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. *J. Mach. Learn. Res.*, 999888, 3283–3310.