Accuracy of Latent-Variable Estimation in Bayesian Semi-Supervised Learning *

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
Hierarchical probabilistic models, such as Gaussian mixture models and hidden Markov models, are widely used for unsupervised learning tasks. These models consist of observable and latent variables, which represent the observable data and the underlying data-generation process, respectively. Unsupervised learning tasks, such as cluster analysis, are regarded as estimations of latent variables based on the observable ones. The estimation of latent variables in semi-supervised learning, where some labels are observed, will be more precise than that in unsupervised, and one of the concerns is to clarify the effect of the labeled data. However, there has not been sufficient theoretical analysis of the accuracy of the estimation of latent variables. In a previous study, a distribution-based error function was formulated, and its asymptotic form was calculated for unsupervised learning with generative models. It has been shown that, for the estimation of latent variables, the Bayes method is more accurate than the maximum-likelihood method. The present paper reveals the asymptotic forms of the error function in Bayesian semi-supervised learning for both discriminative and generative models. The results show that the generative model, which uses all of the given data, performs better when the model is well specified. If the model is misspecified, the discriminative model is preferred. Based on the asymptotic form of the error function for the discriminative model, we propose an information criterion for selecting the optimal model structure.

keywords: Latent-variable estimation, Generative and discriminative models, Bayes statistics, Model selection

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

Hierarchical statistical models, such as the Gaussian mixture model, are widely employed in unsupervised learning. They consist of observable and latent variables, which express the given data and the underlying data-generation process, respectively. A typical task of unsupervised learning is clustering, in which observable data is used to estimate labels that indicate which cluster the data are from. The Gaussian mixture model is often used for cluster analysis, and in this model, a Gaussian component represents a cluster. If the parameter is known, the probabilities of the labels for each data point are easily computed. However, in many practical situations, the parameter is unknown, and both it and the latent variable must be estimated. Some learning algorithms, such as the expectation–maximization (EM) algorithm and the variational Bayes method, have two explicit estimation steps (the E step and the M step). The present paper focuses on the Bayesian approach, which uses the posterior distribution to marginalize out the parameter and calculates the probabilities of the latent variables.

There are two ways to use hierarchical models: to predict unseen observable data or to estimate the latent variables. The accuracy of prediction has been analyzed theoretically. The generalization error measures the accuracy, and, in many cases, its asymptotic behavior is well known. On the other hand, latent-variable (LV) estimation has not been analyzed sufficiently. Recently, a distribution-based error function was formulated to determine the accuracy of the LV estimation, and its asymptotic form was calculated (Yamazaki, 2012b). The results showed the different properties from those of the estimation of observable variables (OVs).

There are discriminative and generative approaches to defining the model; the discriminative approach results in a model that expresses the causal effect of the observable data on the latent variables, while the generative approach results in a model that explains the data-generation process. Our previous study mainly analyzed the generative model (Yamazaki, 2012b). The present paper compares these approaches in terms of the estimation of LVs.

In the estimation of LVs, partially observed labels will improve accuracy because they have information about the targets of the estimation. Learning from a mixture of labeled and unlabeled data is referred to as semi-supervised learning (Zhu, 2007). Its main concerns include clarifying how the supplemental information affects the accuracy of the estimation and developing a learning algorithm that achieves better results based on this advantage. The present paper analyzes the statistical properties of semi-supervised learning, and the main contributions are as follows:

1. The asymptotic forms of the error function are derived for both the discriminative and generative models.

2. The generative model performs better in the well-specified case, while the
discriminative one is more promising for the misspecified case.

3. In order to design the optimal structure of the model, we propose an information criterion that is based on the asymptotic form of the discriminative model.

Asymptotic analysis generally assumes that the amount of training data is sufficiently large. If the number of labels that are given is not large, there may be no effect on the estimation, or it may be very weak. To magnify the effect of the observed labels, we assume that the number of labeled data points is \( \alpha n \), where \( n \) is the number of the training data points and \( \alpha \) is the ratio of the labels where \( 0 < \alpha < 1 \) (see Fig. 1). Asymptotic analysis has provided criteria for model selection; these include the Akaike information criterion (AIC; [Akaike, 1974]), the Bayesian information criterion (BIC; [Schwarz, 1978]), and the minimum description length (MDL) principle ([Rissanen, 1986]). Our criterion is similar to the AIC in that the error function is minimized. The clear difference between them is that ours is for the LV estimation, while the AIC is for the OV estimation.

The rest of this paper is organized as follows. The next section formalizes the data structure and the model expression. In Section 3, we introduce the Bayesian LV estimation and an error function to measure its accuracy. The discriminative and the generative approaches are explained. Section 4 shows the asymptotic analysis of the error function and compares the approaches. In Section 5, we consider the error function of the misspecified case, and we present an asymptotic analysis of the discriminative model. Based on these results, we propose a new model-selection criterion and evaluate it by performing some numerical experiments. In Section 6, the Bayes method is compared with the maximum-likelihood method. Section 7 discusses extensions of the results for the discriminative model to the covariate-shift case and to overspecified modeling. Finally, Section 8 presents our conclusions.

2 Data Structure and Model Expression

2.1 Formal Expressions of Data and Model

This subsection formulates the settings of the data and the model.

Fig. 1 shows the structure of the data; the observable and latent variable are \( x \in \mathbb{R}^M \) and \( y \in \{1, \ldots, K\} \), respectively. The data points \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) are independent and identically distributed, and \( \alpha n \) data points \( \{(x_1, y_1), \ldots, (x_{\alpha n}, y_{\alpha n})\} \) are labeled, where \( 0 < \alpha < 1 \) and \( \alpha n \) is an integer. We define the following data
Figure 1: Data structure: pairs $(x_i, y_i)$ are generated. The solid circles represent observable data, which is used for training, and the dashed circles represent hidden data, which are the target of the estimation.

sets:

$$X_1 = \{x_1, \ldots, x_{\alpha n}\},$$
$$X_2 = \{x_{\alpha n+1}, \ldots, x_n\},$$
$$Y_1 = \{y_1, \ldots, y_{\alpha n}\},$$
$$Y_2 = \{y_{\alpha n+1}, \ldots, y_n\},$$
$$X^n = \{X_1, X_2\},$$
$$Y^n = \{Y_1, Y_2\},$$
$$D = \{X_1, Y_1, X_2\} = \{X^n, Y_1\},$$

where $X_1$ is the set of $\alpha n$ observable data points, $Y_1$ is the set of the corresponding labels, and the target of the LV estimation is $Y_2$. The set $D$ contains the available data for the estimation. The number of the labels grows linearly with the total number of data points $n$.

The generative model represents the underlying process of data generation. In the present paper, we assume that the observable variables are caused by the latent variables. The mathematical expression for this is $p(x, y|w) = p(x|y, w)p(y|w)$, where $w$ is the parameter. On the other hand, the discriminative model expresses
the probability of the latent variable based on the observable variable; the classifier of the learning model is defined by \( p(y|x, w) \). If the discriminative model is defined on the basis of the generative one, the model expression is given by \( p(y|x, w) = p(x, y|w)/p(x|w) \), where \( p(x|w) = \sum_y p(x, y|w) \). According to Fig. 4, the target of the LV estimation is \( p(Y_2|D) \). There are various ways to define \( p(Y_2|D) \), as will be shown in the next section. Let \( q(y|x) \) and \( q(x) \) be the true classifier and the true density of \( x \), respectively; the data \( (X^n, Y^n) \) are generated from \( q(x, y) = q(y|x)q(x) \).

### 2.2 An Example of the Model

For illustrative purposes, we present the following data and model:

**Example 1 (Data distribution)** Define \( x \) and \( y \) so that \( x \in R^2 \) and \( y \in \{1, 2, 3\} \). The sample data follow the following distribution:

\[
q(x, y) = a_j^* \mathcal{N}(x|b_j^*, \sigma_j^*),
\]

where \( a_j^* \) is the mixing ratio and \( \mathcal{N}(x|\mu, \sigma) \) is the two-dimensional Gaussian distribution with mean \( \mu \in R^2 \) and variance vector \( \sigma_y = (\sigma_{y1}, \sigma_{y2})^\top \in R_{>0}^2 \). The element \( \sigma_{yi} \) is the variance of the \( i \)th dimension in the \( y \)th component, and the covariance is zero.

Note that the density of \( x \) is also based on \( q(x, y) \):

\[
q(x) = \sum_{y=1}^{3} q(x, y).
\]

Fig. 2 shows a sample set of the data from \( q(x) \), where \( a_1^* = 0.4 \), \( a_2^* = 0.4 \), \( a_3^* = 0.2 \), \( b_1^* = (0, 1)^\top \), \( b_2^* = (0, -1)^\top \), \( b_3^* = (6, 1)^\top \), and \( \sigma_1^* = \sigma_2^* = \sigma_3^* = (2, 1)^\top \).

**Example 2 (Learning model)** The three-component Gaussian mixture learning model is defined by

\[
p(x|w) = \sum_{k=1}^{3} a_k \mathcal{N}(x|b_k, \sigma_k),
\]

where \( 0 \leq a_k \leq 1 \), \( a_2 = 1 - a_1 - a_3 \), \( b_k \in R^2 \), and \( \sigma_k = (\sigma_{k1}, \sigma_{k2})^\top \). For simplicity, suppose that \( \sigma_k \) is a constant vector; this means that the parameter consists of \( w = \{a_1, b_1, b_2, a_3, b_3\} \). The discriminative model is based on the following mixture:

\[
p(y = k|x, w) = \frac{p(x, y = k, w)}{\sum_{j=1}^{3} p(x, y = j, w)} = \frac{a_k \mathcal{N}(x|b_k, \sigma_k)}{\sum_{j=1}^{3} a_j \mathcal{N}(x|b_j, \sigma_k)}.
\]

If the components have a common variance vector, such that \( \sigma_k = \sigma_k^* \), then there exists a true parameter \( w^* \) such that \( q(y|x) = p(y|x, w^*) \).
2.3 Redundant Parameters in the Discriminative Model

The classifier $p(y|x, w)$ based on $p(x, y|w)$ does not provide a one-to-one relation between the model expression and the parameter. For example, suppose that $a_k^* \neq 0$. Even though there is no redundancy in the number of clusters, there is redundancy
in the parameter:

\[
p(y = 1|x, w) = \frac{a_1 \exp\{- (x_1 - b_{11})^2/(2\sigma_{11}^2) - (x_2 - b_{12})^2/(2\sigma_{12}^2)\}}{Z_m} = \frac{1}{1 + f_1(x, w) + f_2(x, w)},
\]

\[
Z_m = \sum_{y=1}^{3} \frac{a_y}{\sigma_{y1}\sigma_{y2}} \exp\{- (x_1 - b_{y1})^2/(2\sigma_{y1}^2) - (x_2 - b_{y2})^2/(2\sigma_{y2}^2)\},
\]

\[
\ln f_1(x, w) = -\frac{1}{2} \left( \frac{1}{\sigma_{21}^2} - \frac{1}{\sigma_{11}^2} \right) x_1^2 - \frac{1}{2} \left( \frac{1}{\sigma_{22}^2} - \frac{1}{\sigma_{12}^2} \right) x_2^2 + \frac{b_{21}}{\sigma_{21}^2 - \sigma_{11}^2} x_1 + \frac{b_{22}}{\sigma_{22}^2 - \sigma_{12}^2} x_2 - \frac{1}{2} \left( \frac{b_{21}^2}{\sigma_{21}^2 - \sigma_{11}^2} - \frac{b_{11}^2}{\sigma_{11}^2} \right) x_1 + \frac{b_{22}^2}{\sigma_{22}^2 - \sigma_{12}^2} x_2
\]

\[
\ln f_2(x, w) = -\frac{1}{2} \left( \frac{1}{\sigma_{31}^2} - \frac{1}{\sigma_{11}^2} \right) x_1^2 - \frac{1}{2} \left( \frac{1}{\sigma_{32}^2} - \frac{1}{\sigma_{12}^2} \right) x_2^2 + \frac{b_{31}}{\sigma_{31}^2 - \sigma_{11}^2} x_1 + \frac{b_{32}}{\sigma_{32}^2 - \sigma_{12}^2} x_2 - \frac{1}{2} \left( \frac{b_{31}^2}{\sigma_{31}^2 - \sigma_{11}^2} - \frac{b_{11}^2}{\sigma_{11}^2} \right) x_1 + \frac{b_{32}^2}{\sigma_{32}^2 - \sigma_{12}^2} x_2
\]

where the coefficients of \(x_1\) and \(x_2\), and the constant terms contain more elements of the parameter than are needed to express the function. Let us reparameterize \(f_1(x, w)\) and \(f_2(x, w)\) as

\[
\ln f_1(x, \bar{w}) = -\frac{1}{2} \left( \frac{1}{\sigma_{21}^2} - \frac{1}{\sigma_{11}^2} \right) x_1^2 - \frac{1}{2} \left( \frac{1}{\sigma_{22}^2} - \frac{1}{\sigma_{12}^2} \right) x_2^2 + c_{11} x_1 + c_{12} x_2 + c_{13},
\]

\[
\ln f_2(x, \bar{w}) = -\frac{1}{2} \left( \frac{1}{\sigma_{31}^2} - \frac{1}{\sigma_{11}^2} \right) x_1^2 - \frac{1}{2} \left( \frac{1}{\sigma_{32}^2} - \frac{1}{\sigma_{12}^2} \right) x_2^2 + c_{21} x_1 + c_{22} x_2 + c_{23},
\]

where \(\bar{w} = \{c_{11}, \ldots, c_{23}\}\). Considering the cases \(y = 2, 3\), we can easily confirm that the parameter \(\bar{w}\) is sufficient to express \(p(y = 2|x, w)\) and \(p(y = 3|x, w)\). This means that the essential dimension of the parameter is \(\dim \bar{w} = 6\) instead of \(\dim w = 8\). To eliminate the redundancy, we regard \(a_k\) as a positive constant and let the reduced parameter be \(\bar{w} = w \setminus \{a_k\}\). For the general dimension of data \(M\), we can calculate that \(\dim \bar{w} = \dim w - M\).
3 The Bayes LV Estimation and Error Function

This section introduces the Bayes LV estimation and an error function to measure its accuracy. Let \( L(w, X^n, Y^n) \) be a likelihood function on \( \{X^n, Y^n\} \), and let \( L(w, D) = \sum_{Y^n} L(w, X^n, Y^n) \) be one on \( D \). In the Bayesian model, the LV estimation, which corresponds to constructing \( p(Y_2|D) \), is written as

\[
p(Y_2|D) = \frac{\int L(w, X^n, Y^n) \varphi(w|\eta) dw}{\int L(w, D) \varphi(w|\eta) dw},
\]

where \( \varphi(w|\eta) \) is a prior distribution and \( \eta \) is the hyperparameter. We will consider here the following three likelihood functions:

(Model 1)

\[
L_1(w, X^n, Y^n) = \prod_{i=1}^{n} p(y_i|x_i, w),
\]

\[
L_1(w, D) = \prod_{i=1}^{\alpha n} p(y_i|x_i, w);
\]

(Model 2)

\[
L_2(w, X^n, Y^n) = \prod_{i=1}^{\alpha n} p(y_i|x_i, w) \prod_{i=\alpha n+1}^{n} p(x_i, y_i|w),
\]

\[
L_2(w, D) = \prod_{i=1}^{\alpha n} p(y_i|x_i, w) \prod_{i=\alpha n+1}^{n} p(x_i|w);
\]

(Model 3)

\[
L_3(w, X^n, Y^n) = \prod_{i=1}^{n} p(x_i, y_i|w),
\]

\[
L_3(w, D) = \prod_{i=1}^{\alpha n} p(x_i, y_i|w) \prod_{i=\alpha n+1}^{n} p(x_i|w).
\]

The first and third models correspond to the discriminative and generative models, respectively. The second one is a hybrid model; the labeled data are used in the discriminative expression, and the unlabeled data are used in the generative expression. Note that Model 2 cannot use the reduced parameterization \( \bar{w} \) in the discriminative expression, since it must use \( w \) in the generative part.
The formulation defined by Eq. 3 has the following equivalent expressions:

\[ p(Y_2|D) = \int \prod_{i=\alpha n+1}^{n} p(y_i|x_i, w)p(w|D)dw, \quad (4) \]

\[ p(w|D) = \frac{L(w, D)\varphi(w|\eta)}{\int L(w, D)\varphi(w|\eta)dw}, \quad (5) \]

where \( p(w|D) \) is the posterior distribution. The first definition of \( p(Y_2|D) \), Eq. 3, is used for theoretical calculations in the asymptotic analysis, and the second one, Eq. 4, is useful for numerical computations.

Since the data are i.i.d., the true joint probability of \( Y_2 \) is given by

\[ q(Y_2|D) = q(Y_2|X_2) = \prod_{i=\alpha n+1}^{n} q(y_i|x_i). \quad (6) \]

The error function is defined as the difference between the true and the estimated probabilities of \( Y_2 \). Following previous work \cite{Yamazaki2012}, the error is based on the Kullback–Leibler divergence:

\[ D(n) = \frac{1}{(1 - \alpha)n} E_D \left[ \sum_{Y_2} q(Y_2|D) \ln \frac{q(Y_2|D)}{p(Y_2|D)} \right], \]

where the expectation means

\[ E_D[f(X_1, Y_1, X_2)] = \int \sum_{Y_1} f(X_1, Y_1, X_2)q(Y_1|X_1)q(X_1)q(X_2)dX_1dX_2. \]

Since the number of elements in \( Y_2 \) is \( (1 - \alpha)n \), the error function is the average divergence for one latent variable.

**4 Asymptotic Analysis of the Error Function**

This section shows one of the main results of this paper: the asymptotic forms of the error functions of the three models and a comparison between them.

We assume the following condition:

\[(A0) \text{ There exists a true parameter } w^* \text{ such that } q(y|x) = p(y|x, w^*) \text{ and } q(x, y) = p(x, y|w^*) \text{ in the support of } \varphi(w|\eta), \text{ and the following Fisher information ma-}\]
trices in the neighborhood of $w^*$ exist and are positive definite:

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

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

$$
\{I_x(w)\}_{ij} = \int \sum y \frac{\partial \ln p(x|w)}{\partial w_i} \frac{\partial \ln p(x|w)}{\partial w_j} p(x|w) dx.
$$

This condition indicates the ideal situation for the estimation, where the estimated probability $p(Y_2|D)$ in all models converges to the true one and the model is identifiable. The true parameter $w^*$ is a unique point, although there exist some points for the parameter $w^*_x$ such that $p(x|w^*_x) = q(x)$; this is due to symmetry in the latent-variable and parameter spaces.

When the discriminative model $p(y|x, w)$ is based on the generative expression, such as in Example 2, then $I(w^*) = I_{xy}(w^*) - I_x(w^*)$. Note that here we use the common parameter setting and not the reduced one $\tilde{w}$; this means that, due to the redundancy of the parameters, the rank of $I(w^*)$ is not more than $\dim w - M$. The parameter redundancy in the discriminative model must be eliminated for the condition (A0). Thus, we use the notation $\tilde{w}$ to indicate the parameter setting satisfying (A0), and we use the notation $p(y|x, \tilde{w})$ even if the model is not based on the generative expression.

The following theorem shows the asymptotic behavior of the error function.

**Theorem 3** Let $D_1(n)$, $D_2(n)$, and $D_3(n)$ be the error functions of Models 1, 2, and 3, respectively. Under the condition (A0), it holds that

$$
D_1(n) = \frac{\dim \tilde{w} \ln 1/\alpha}{2} \frac{1}{1 - \alpha} \frac{1}{n} + o\left(\frac{1}{n}\right),
$$

$$
D_2(n) = \frac{1}{2} \ln \det K_2(w^*) \frac{1}{1 - \alpha} \frac{1}{n} + o\left(\frac{1}{n}\right),
$$

$$
D_3(n) = \frac{1}{2} \ln \det K_3(w^*) \frac{1}{1 - \alpha} \frac{1}{n} + o\left(\frac{1}{n}\right),
$$

where

$$
K_2(w) = (I_{xy}(w) - \alpha I_x(w)) (\alpha I_{xy}(w) + (1 - 2\alpha) I_x(w))^{-1},
$$

$$
K_3(w) = I_{xy}(w) (\alpha I_{xy}(w) + (1 - \alpha) I_x(w))^{-1}.
$$

The proof is in the appendix.
The theorem shows that, in all models, the dominant order is $1/n$, which is the speed at which the error converges to zero. The accuracy of Model 1 depends on the dimension of $\bar{w}$ instead of the position of $w^*$; note that, in the other models, the coefficients of the dominant order are functions of $w^*$.

Let us compare these error values. We assume that Model 1 is based on the generative expression and uses the reduced parameter $\bar{w}$. It is known that the dimension of the parameter has a strong influence on its accuracy; a small value of the dimension makes the convergence of the posterior distribution faster (Schwarz, 1978), and the coefficient of the dominant term is a half of the parameter dimension in the asymptotic generalization error of the OV estimation (Akaike, 1974). Since $\dim \bar{w} < \dim w$, Model 1 has an advantage in the parameter dimension. On the other hand, the actual amount of data used for the estimation is larger in Models 2 and 3; according to the second definition of $p(Y_2|D)$, Eq. 4, the posterior distribution of Model 1 is constructed by only $\{X_1, Y_1\}$, while those of Models 2 and 3 require $D = \{X_1, Y_1, X_2\}$. Obviously, a larger amount of training data increases the accuracy of the estimation. Models 2 and 3 have an advantage based on the amount of data that is used. Thus, in the accuracy, we have a trade-off between the parameter dimension and the amount of data.

The following theorem shows that, in order to improve accuracy, increasing the amount of data is more effective than reducing the dimension of the parameter. Model 1 is thus at a disadvantage.

**Theorem 4** If the nonzero eigenvalues of $I(w^*)I_x(w^*)^{-1}$ are all positive, the following inequality holds asymptotically:

$$D_3(n) < D_2(n) < D_1(n).$$

The proof is in the appendix.

5 Designing the Optimal Structure of the Discriminative Model

This section introduces a criterion for model selection. First, we experimentally confirm that the discriminative model is a good candidate for model selection; in practical situations, Models 2 and 3 do not satisfy (A0). The asymptotic form of $D_1(n)$ is then shown in a more general case. Using this form, we propose a criterion.

Usually, model selection means selecting the optimal number of clusters in the mixture models. However, under a semi-supervised learning situation, we already know how many clusters there are, based on the labeled data. Thus, in this section, we will select the optimal structure for the components.
In the previous section, we analyzed the error function under the ideal condition (A0), and we found that the generative model had the best performance. In this section, we examine the condition and determine how strict it is. When considering a method for designing a model, we need to consider the misspecified case, in which the model cannot attain the true distribution. Let us define three maximum-likelihood estimators:

\[ \hat{w}_i = \arg \max_w L_i(w, D) \]

for \( i = 1, 2, 3 \), which correspond to Models 1, 2, and 3, respectively. The maximum-likelihood estimator indicates the asymptotic location of the posterior distribution; this is because \( L_i(w, D) \) is the main factor in the numerator of the distribution.

Fig. 3 shows representative estimation results \( \hat{Y}_2 = \arg \max_{Y_2} p(Y_2|X_2, \hat{w}_i) \). We used the data shown in Fig. 2. The model \( p(x, y|w) \) was the same as that in Example 2. The variances had a common value, \( \sigma_k = (1, 1) \), for all \( k \). The left, the middle, and the right panels correspond to the results of \( \hat{w}_1 \), \( \hat{w}_2 \), and \( \hat{w}_3 \), respectively. The discriminative model with labeled data estimates the clusters accurately. Obviously, neither \( \hat{w}_2 \) nor \( \hat{w}_3 \) converges to the desired point. The probability \( p(x|w) \) does not attain \( q(x) \) if the variance \( \sigma_k \) is different from \( \sigma^*_k \), and the likelihood functions including \( p(x|w) \) tend to cover all the given data with isotropic Gaussian distributions. This is why the unlabeled data \( X_2 \) adversely affect the results.

Even in such a simple setting, the condition (A0) is not satisfied in Models 2 and 3, which shows that these models are not appropriate for model selection. We will thus focus on Model 1 and analyze the asymptotic behavior of \( D_1(n) \).

### 5.2 The Asymptotic Form of \( D_1(n) \) in a Generalized Case

We assume that the following conditions are satisfied:
If there exists a parameter $w^*$ such that $q(y|x) = p(y|x, w^*)$ in the support of $\varphi(w|\eta)$, then the following Fisher matrix $I(w)$ and its variant $J(w)$ in the neighborhood of $w^*$ exist and are positive definite:

$$\{J(w)\}_{ij} = -\int \sum_y \frac{\partial^2 \ln p(y|x, w)}{\partial w_i \partial w_j} p(y|x, w) q(x) dx.$$

If $q(y|x) \neq p(y|x, w)$ for any parameter, then $w_0$ minimizing the Kullback–Leibler divergence from $q(y|x)$ to $p(y|x, w)$ uniquely exists in the support of $\varphi(w|\eta)$, and $I(w_0)$ is positive definite. The parameter $w_0$ is formally defined as

$$w_0 = \arg \min_w \left\{ \int \sum_y q(y|x) \ln \frac{q(y|x)}{p(y|x, w)} q(x) dx \right\}.$$

The observable data in both sets $X_1$ and $X_2$ are distributed according to $q(x)$, and the number of clusters $K$ is known.

Condition (A3) indicates that the labeled data are known to be properly sampled. The following theorem then holds.

**Theorem 5** If the conditions (A1)–(A3) are satisfied, then the error function has the asymptotic form

$$D_1(n) = E \left[ \ln \frac{q(y|x)}{p(y|x, w_0)} \right] + \frac{\dim w \ln 1/\alpha}{2} \frac{1}{1 - \alpha} \frac{1}{n} + o\left(\frac{1}{n}\right),$$

where the expectation is

$$E[f(x, y)] = \int \sum_y f(x, y) q(y|x) q(x) dx.$$

If the parameter $w^*$ exists, then $w_0 = w^*$, which indicates that the first term in Eq. (7) is zero.

The proof is in the appendix.

This theorem includes the results for $D_1(n)$ from Theorem 3. In the misspecified case, there is a bias term; the error converges to a positive value instead of to zero. A model with a broad expression can be close to the true classifier, and this makes the bias term small; in this case, the parameter generally has a large dimension, which makes the second term large. This is why an expressive model is not always good, and model selection is necessary.
5.3 A Model Selection Criterion

There are criteria for selecting the optimal model that are based on the asymptotic forms of the evaluation functions; these include the generalization error and the free-energy function (Akaike, 1974; Schwarz, 1978; Watanabe, 2010; Yamazaki et al., 2006). In this subsection, we propose a criterion that selects a model by minimizing the error function $D_1(n)$.

Theorem 5 shows the asymptotic form of the error function. The first term depends on the model and its optimal parameter $w_0$; this means that we cannot calculate the term or compare the error values among some candidate models. Another asymptotic expression of the error is derived in the following theorem.

**Theorem 6** Let $\hat{w}_\alpha$ be the maximum a posteriori probability estimator from the labeled data,

$$\hat{w}_\alpha = \arg \max_w \prod_{i=1}^n p(y_i|x_i, w)\varphi(w|\eta).$$

By using the estimator, the asymptotic form of the error function is described as

$$D_1(n) = E[\ln q(y|x)] - E_{\alpha n} \left[ \frac{1}{\alpha n} \sum_{i=1}^n \ln p(y_i|x_i, \hat{w}_\alpha) \right] + \frac{1}{2\alpha n} \text{Tr}\{I(w_0)J(w_0)^{-1}\} + \frac{\dim w \ln 1/\alpha}{2} \frac{1}{1 - \alpha} n + o\left(\frac{1}{n}\right)$$

under the conditions (A1)-(A3).

The proof is in the appendix.

In the model selection, we can ignore the first term, which depends on the true classifier, because it is a constant term with respect to the model $p(y|x, w)$. The third term requires the optimal parameter $w_0$. Replacing $w_0$ with $\hat{w}_\alpha$, and the matrices $I(w)$ and $J(w)$ with

$$\{I_n(w)\}_{ij} = \frac{1}{n} \sum_{l=1}^n \frac{\partial \ln p(y_l|x_l, w)}{\partial w_i} \frac{\partial \ln p(y_l|x_l, w)}{\partial w_j},$$

$$\{J_n(w)\}_{ij} = -\frac{1}{n} \sum_{l=1}^n \frac{\partial^2 \ln p(y_l|x_l, w)}{\partial w_i \partial w_j},$$

respectively, we obtain that the model minimizing the value

$$\text{IC} = -\frac{2}{\alpha} \sum_{i=1}^{\alpha n} \ln p(y_i|x_i, \hat{w}_\alpha) + \frac{1}{\alpha} \text{Tr}\{I_{\alpha n}(\hat{w}_\alpha)J_{\alpha n}(\hat{w}_\alpha)^{-1}\} + \frac{\ln 1/\alpha}{1 - \alpha} \dim w$$

is optimal on average, in the sense of minimizing the error, when the amount of data $n$ is sufficiently large.
5.4 Experiments with Model Selection

This subsection briefly evaluates the model-selection criterion, using the settings of Examples 1 and 2: $a_1^* = 0.4$, $a_2^* = 0.4$, $a_3^* = 0.2$, $b_1^* = (1, 3)^\top$, $b_2^* = (0, 0)^\top$, $b_3^* = (3, -1)^\top$. The variances are $\sigma_1^* = (2, 2)^\top$, $\sigma_2^* = (1, 1)^\top$, and $\sigma_3^* = (3, 3)^\top$. The generated data are shown in Fig. 4. If $\sigma_k = \sigma_k^*$ for $k = 1, 2, 3$, the learning model attains the true classifier $q(y|x)$. In other words, the model is not appropriate if $\sigma_k$ includes a different value. The number of data points was $n = 5,000$, and the ratio of labeled data was $\alpha = 0.1$. Table 1 shows the models with different variances; the left and right columns are the variances and the IC values, respectively. The model in the first row attains the true distribution. For the model in the last row, the variances $\sigma_k$ such that $\sigma_{k1} = \sigma_{k2}$ for $k = 1, 2, 3$ are a part of the parameter. This model can also attain the true distribution when the variances are correctly estimated. There are three additional dimensions for $\sigma_k$ in $w$. Note that there are two degrees of freedom; this is due to the expressions of $f_1$ and $f_2$ in Example 2. Thus, $\dim w = 8$ for the last model, and $\dim \hat{w} = 6$ for the other models; this adversely affects the last model, because the larger dimension increases the last term of the IC. As expected, the first model has the smallest value for the IC. Comparing the second and the third models, the variance $\sigma_1$ has a greater effect on the accuracy than does $\sigma_3$. For this data set, the second model is more appropriate than the last one. Since the second model has a value of $\sigma_3$ that is different from that of the true
Table 1: Experimental results of model selection; the true distribution has variance \( \sigma_1^* = (2, 2)^\top, \sigma_2^* = (1, 1)^\top, \) or \( \sigma_3^* = (3, 3)^\top. \)

| \((\sigma_{11}, \sigma_{12}), (\sigma_{21}, \sigma_{22}), (\sigma_{31}, \sigma_{32})\) | Value of IC |
|-------------------------------------------------|-------------|
| \((2,2),(1,1),(3,3)\)                          | 4495.22     |
| \((2,2),(1,1),(4,4)\)                          | 4522.18     |
| \((1.5,2),(1,1),(3,3)\)                        | 4651.79     |
| \((2,2),(4,4),(3,3)\)                          | 6346.61     |
| estimated as the parameter                      | 4560.04     |

distribution, the last model will be better, on average, for the asymptotic case.

6 Comparison with the Maximum-Likelihood Estimation

This section introduces the maximum-likelihood estimation and compares its asymptotic error to that of the Bayes estimation. The estimated probability of \( Y_2 \) is defined by

\[
p(Y_2|D) = p(Y_2|X_2, \hat{w}) = \prod_{i=\alpha n+1}^{n} p(y_i|x_i, \hat{w}),
\]

where \( \hat{w} \) is the estimator. Once \( \hat{w} \) is obtained, \( y_i \) for \( i > \alpha n \) is independently estimated on the basis of each respective \( x_i \), which results in the final expression.

We will analyze the error function of \( \hat{w}_1 \):

\[
D_{ML}(n) = \frac{1}{(1 - \alpha)n} E_D \left[ \sum_{Y_2} q(Y_2|D) \ln \frac{q(Y_2|D)}{p(Y_2|X_2, \hat{w}_1)} \right] = E_D E \left[ \ln \frac{q(y|x)}{p(y|x, \hat{w}_1)} \right].
\]

Since the estimation corresponds to the prediction in the regression problem, the error function is equivalent to the generalization error. The following result is immediately obtained.

Lemma 7 Assume that the conditions (A1)–(A3) are satisfied. The asymptotic form of the generalization error is expressed as

\[
D_{ML}(n) = E \left[ \frac{q(y|x)}{p(y|x, w_0)} \right] + \frac{1}{2\alpha n} \text{Tr}[I(w_0)J(w_0)^{-1}] + o\left(\frac{1}{n}\right).
\]
The proof is omitted.

Comparing Lemma 7 with Theorem 5, we find that the magnitude relation of the error functions is determined by the value of $\text{Tr}\{I(w_0)J(w_0)^{-1}\}$. We obtain the following corollary.

**Corollary 8** A sufficient condition for the asymptotic advantage of the Bayes estimation is as follows:

If $\text{Tr}\{I(w_0)J(w_0)^{-1}\} \geq \text{dim } w$, then $D_1(n) - D_{ML}(n) \leq o(1/n)$ for any $\alpha$.

**Proof of Corollary** We consider the dominant terms of $D_1(n)$ and $D_{ML}(n)$.

$$D_1(n) - D_{ML}(n) = \frac{\text{dim } w \ln 1/\alpha}{2} \frac{1}{1 - \alpha} \frac{1}{n} - \frac{1}{2\alpha n} \text{Tr}\{I(w_0)J(w_0)^{-1}\} + o\left(\frac{1}{n}\right)$$

$$\leq \frac{\text{dim } w}{2} \left(\ln 1/\alpha - \frac{1}{1 - \alpha}\right) \frac{1}{n} + o\left(\frac{1}{n}\right)$$

$$\leq \frac{\text{dim } w}{2} \left(\frac{1}{1 - \alpha} - \frac{1}{\alpha}\right) \frac{1}{n} + o\left(\frac{1}{n}\right) = o\left(\frac{1}{n}\right),$$

which completes the proof. (End of Proof)

## 7 Extensions of Theorem 5

In this section, we extend the theorem to the case in which (A3) does not hold. More precisely, when the density $q(x)$ of $X_2$ is different from that of $X_1$ (Section 7.1), and when the number of clusters $K$ is unknown (Section 7.2). For simplicity, the existence of $w^*$ is assumed.

### 7.1 Covariate-Shift Cases

Let $q_1(x)$ and $q_2(x)$ be the density of $X_1$ and $X_2$, respectively. The regression problem refers to the case $q_2(x) \neq q_1(x)$ as the covariate shift (Shimodaira, 2000). In many practical situations, the labels are manually assigned, which means that obtaining them is expensive. It frequently happens that the density of the observable data with labels is different from the density of the unlabeled target data.

To analyze the asymptotic error, we assume that the learning model $p(y|x, w)$ can realize the true classifier $q(y|x)$. Then, both estimators given by

$$\hat{w}_1 = \arg \max_w \prod_{i=1}^{an} p(y_i|x_i, w),$$

$$\hat{w}_2 = \arg \max_w \prod_{i=an+1}^{n} p(y_i|x_i, w)$$

are consistent.
converge to \( w^* \) for \( n \to \infty \). Let the Fisher information matrices with these density functions be defined by

\[
\{I_1(w)\}_{ij} = \int \sum_y \frac{\partial \ln p(y|x, w)}{\partial w_i} \frac{\partial \ln p(y|x, w)}{\partial w_j} p(y|x, w) q_1(x) dx,
\]

\[
\{I_2(w)\}_{ij} = \int \sum_y \frac{\partial \ln p(y|x, w)}{\partial w_i} \frac{\partial \ln p(y|x, w)}{\partial w_j} p(y|x, w) q_2(x) dx.
\]

We assume the following conditions:

(A1-1) There exists a parameter \( w^* \) such that \( q(y|x) = p(y|x, w^*) \) in the support of \( \varphi(w|\eta) \), and \( \hat{w}_1 \) and \( \hat{w}_2 \) converge to \( w^* \).

(A3-1) The number of clusters \( K \) is known.

The following theorem then holds.

**Theorem 9** Let the estimated and true probabilities of \( Y_2 \) be defined by Eqs. 3 and 6, respectively. If the conditions (A1-1) and (A3-1) are satisfied, the error function has the asymptotic form,

\[
D(n) = \frac{1}{2} \ln \det \left( \frac{1}{1 - \alpha} \left( \alpha I_1(w^*) + (1 - \alpha) I_2(w^*) \right) I_1(w^*)^{-1} \right) \frac{1}{n} + \frac{\dim w \ln 1/\alpha}{2} \frac{1}{1 - \alpha} n + o \left( \frac{1}{n} \right).
\]

The proof is in the appendix. The effect of the covariate shift appears in the first term of the asymptotic form; this term is zero in Theorem 5.

### 7.2 Overspecified Cases

When it is not clear that the observable labels cover all possibilities, a sufficiently large number of the labels will be assigned. Let \( K^* \) and \( K \) be the numbers of clusters in the true and the estimated probabilities, respectively. The above-mentioned situation is formulated as \( K > K^* \). In order to unify the notation, we define \( q(y = k|x) = 0 \) for all \( x \) when \( k > K^* \). Then, the parameter \( w^* \) is not represented by a unique point. For example, let us consider \( K^* = 2 \) and \( K = 3 \), where \( a_3^* = 0 \) in Example 1 and the model \( p(y|x, w) \) is given by Example 2. Then, \( w^* \) is a set of parameters:

\[
W_t = \{w^*: q(y|x) = p(y|x, w^*)\} = \{a_1 = a_1^*, b_1 = b_1^*, b_2 = b_2^*, a_3 = 0\}.
\]

Since \( a_3 = 0 \), \( b_3 \) can take an arbitrary value. The Fisher information matrix \( I(w^*) \) is not positive definite for the set \( W_t \).
Even in the case of a non-negative Fisher matrix, the algebraic geometrical method (Watanabe, 2001) is applicable to the asymptotic analysis. The previous studies with this method revealed the prediction accuracy of the observable variable and the free-energy function in the generative models (Aoyagi and Watanabe, 2005; Aoyagi, 2010; Rusakov and Geiger, 2005; Zwiernik, 2011). In recent years, it has been shown that the method is also applicable to analysis of the accuracy of the LV estimation (Yamazaki, 2012a).

In this subsection, we apply the algebraic geometrical method to Bayesian semi-supervised learning in a discriminative model. In the same way as in Examples 1 and 2, let the joint probabilities be defined by

\[
q(x, y = k) = a^*_k \mathcal{N}(x | b^*_k, \sigma^*),
\]

\[
p(x, y = k | w) = a_k \mathcal{N}(x | b_k, \sigma^*),
\]

where \( x \in \mathbb{R}^M \), \( y = \{1, \ldots, K\} \), \( 0 < a^*_k < 1 \) for \( k \leq K^* \), and the variance vectors have the same value \( \sigma^* \) for all components of \( q(x, y) \) and \( p(x, y | w) \). The classifiers are given by

\[
q(y | x) = \frac{q(x, y)}{\sum_{y=1}^{K} q(x, y)},
\]

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

The parameter \( w \) consists of \( a_k \) and \( b_k \), where \( a_{K^*} = 1 - \sum_{k \neq K^*} a_k \). The dimension of the parameter is \( \text{dim } w = K - 1 + KM \), but not all elements of \( w \) are necessary to express \( p(y | x, w) \).

The prior distribution is defined as

\[
\varphi(w | \eta) = \varphi(a | \eta_1) \varphi(b | \eta_2),
\]

\[
\varphi(a | \eta_1) = \frac{\Gamma(K \eta_1)}{\Gamma(\eta_1)^K} \prod_{k=1}^{K} a^{\eta_1 - 1}_k,
\]

\[
\varphi(b | \eta_2) = \prod_{k=1}^{K} \varphi(b_k | \eta_2),
\]

where \( \Gamma \) is the Gamma function, and \( \varphi(b_k | \eta_2) \) is a \( C^\infty \) function that has compact support. The hyperparameter is \( \eta = \{\eta_1, \eta_2\} \). The prior distribution for the mixing ratios \( a_k \) is the symmetric Dirichlet distribution.

The following theorem shows the asymptotic behavior of the error.
**Theorem 10** Let the estimated and true probabilities of $Y_2$ be defined by Eqs. 3 and 6, respectively. If the models and the prior distribution are given by the above definitions, the error function has the asymptotic form

$$D(n) = \left\{ \frac{K^* M}{2} + (K - K^*) \eta_1 \right\} \frac{\ln 1/\alpha}{1 - \alpha n} + o\left(\frac{1}{n}\right).$$

The proof is in the appendix. It is known that the Dirichlet prior for the mixing ratio has the automatic relevance determination (ARD) effect, which automatically eliminates unnecessary components. The ARD effect can be recognized by the shape of the prior, which has abrupt peaks at $a_k = 0$ and $a_k = 1$ for $0 < \eta_1 < 1$; this implies that the necessary components tend to have a nonzero value of $a_k$, due to the latter peak, and that the unnecessary ones will be removed, due to the former peak. The ARD effect quantitatively appears in Theorem 10: the value of $\eta_1$ linearly affects the $K - K^*$ redundant components and, when it is small, it eliminates the redundancy.

## 8 Conclusions

The generative and discriminative approaches were compared in terms of the LV estimation. The results showed that the generative approach performed better in the well-specified case. However, the situations to which this is applicable are quite limited. Considering the misspecified case in Fig. 3, we must conclude that model selection in the case of unsupervised learning is difficult, if not impossible. The distribution shape of the components strongly affects the estimation result, which causes an inappropriate convergence of the estimation. In the misspecified unsupervised case, we cannot obtain the proper result; indeed, we do not even have a way to check if the result is proper. In semi-supervised learning, some of the latent variables are observable; this allows us to establish a criterion for model selection. This shows that labeling information is crucial for the estimation of latent variables in real-world problems.

## Appendix

This section shows the proofs of the theorems.

### Proof of Theorem 3

This subsection derives the asymptotic forms of $D_2(n)$ and $D_3(n)$. The asymptotic form of $D_1(n)$ is shown as a special case in the proof of Theorem 5.
Define the free-energy functions as

\[
F_{xy}(n) = E_n \left[ \ln \prod_{i=1}^{n} q(x_i, y_i) - \ln \int \prod_{i=1}^{n} p(x_i, y_i | w) \varphi(w | \eta) dw \right],
\]

\[
F_{xy,x}(n) = E_n \left[ \ln \prod_{i=1}^{\alpha n} q(x_i, y_i) \prod_{i=\alpha n + 1}^{n} q(x_i)
- \ln \int \prod_{i=1}^{\alpha n} p(x_i, y_i | w) \prod_{i=\alpha n + 1}^{n} p(x_i | w) \varphi(w | \eta) dw \right],
\]

where the expectation is

\[
E_n[f(X^n, Y^n)] = \int \sum_{Y^n} f(X^n, Y^n) q(Y^n | X^n) q(X^n) dX^n.
\]

The error function \(D_3(n)\) is rewritten as

\[
(1 - \alpha) n D_3(n) = E_D \left[ \sum_{Y_2} q(Y_2 | D) \left\{ \ln \frac{\prod_{i=1}^{n} q(x_i, y_i)}{\int \prod_{i=1}^{n} p(x_i, y_i | w) \varphi(w | \eta) dw} \right. \\
- \ln \frac{\prod_{i=1}^{\alpha n} q(x_i, y_i) \prod_{i=\alpha n + 1}^{n} q(x_i)}{\int \prod_{i=1}^{\alpha n} p(x_i, y_i | w) \prod_{i=\alpha n + 1}^{n} p(x_i | w) \varphi(w | \eta) dw} \right\} \right]
\]

\[
= F_{xy}(n) - F_{xy,x}(n).
\]  

(8)

The maximum-likelihood estimator is defined by

\[
\hat{w}_{xy} = \arg \max_w \prod_{i=1}^{n} p(x_i, y_i | w).
\]

Due to (A0), the estimators \(\hat{w}_{xy}\) and \(\hat{w}_3\) converge to \(w^*\), which means that the essential parameter area for the integration is the neighborhood of \(w^*\), \(\hat{w}_{xy}\), and \(\hat{w}_3\).
According to the Taylor expansion at $w = \hat{w}_{xy}$,

$$
E_n \left[ \ln \int \prod_{i=1}^{n} p(x_i, y_i|w) \varphi(w|\eta) dw \right] \\
= E_n \left[ \ln \int \exp \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln p(x_i, y_i|w) \right\} \varphi(w|\eta) dw \right] \\
= E_n \left[ \ln \int \exp \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln p(x_i, y_i|\hat{w}_{xy}) \\
+ \frac{1}{n} (w - \hat{w}_{xy})^\top \frac{\partial}{\partial w} \sum_{i=1}^{n} \ln p(x_i, y_i|\hat{w}_{xy}) \\
+ \frac{1}{2} (w - \hat{w}_{xy})^\top \frac{\partial^2}{\partial w^2} \sum_{i=1}^{n} \ln p(x_i, y_i|\hat{w}_{xy}) (w - \hat{w}_{xy}) + r_1(w) \right\} \varphi(w|\eta) dw \right],
$$

where $r_1(w)$ is the remainder term. Based on the saddle-point approximation,

$$
E_n \left[ \ln \int \prod_{i=1}^{n} p(x_i, y_i|w) \varphi(w|\eta) dw \right] \\
= E_n \left[ \sum_{i=1}^{n} \ln p(x_i, y_i|\hat{w}_{xy}) \right] \\
+ E_n \left[ \ln \int \exp(nr_1(w)) \varphi(w|\eta) \mathcal{N}(\hat{w}_{xy}, (nI_{xy}(w^*))^{-1}) dw \right] \\
- \left\{ \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det I_{xy}(w^*) \right\} + o(1)
$$

$$
= E_n \left[ \sum_{i=1}^{n} \ln p(x_i, y_i|\hat{w}_{xy}) \right] - \frac{\dim w}{2} \ln n \\
+ \frac{\dim w}{2} \ln 2\pi - \frac{1}{2} \ln \det I_{xy}(w^*) + \ln \varphi(w^*) + o(1),
$$

where $\mathcal{N}(\mu, \Sigma)$ is a $\dim w$-dimensional Gaussian distribution with mean $\mu \in \mathbb{R}^{\dim w}$ and variance-covariance matrix $\Sigma$. Then, we obtain

$$
F_{xy}(n) = E_n \left[ \sum_{i=1}^{n} \ln \frac{p(x_i, y_i|w^*)}{p(x_i, y_i|\hat{w}_{xy})} \right] \\
+ \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det I_{xy}(w^*) - \ln \varphi(w^*|\eta) + o(1).
$$
According to the Taylor expansion at \( \hat{w}_{xy} \),

\[
E_n \left[ \sum_{i=1}^{n} \ln p(x_i, y_i | w^*) \right] \\
= E_n \left[ \sum_{i=1}^{n} \ln p(x_i, y_i | \hat{w}_{xy}) \right] \\
+ (w^* - \hat{w}_{xy})^\top \frac{\partial}{\partial w} \sum_{i=1}^{n} \ln p(x_i, y_i | \hat{w}_{xy}) \\
+ \frac{1}{2} (w^* - \hat{w}_{xy})^\top \frac{\partial^2}{\partial w^2} \sum_{i=1}^{n} \ln p(x_i, y_i | \hat{w}_{xy}) (w^* - \hat{w}_{xy}) + r_2(w) \\
= E_n \left[ \sum_{i=1}^{n} \ln p(x_i, y_i | \hat{w}_{xy}) \right] - n E_n \left[ (w^* - \hat{w}_{xy})^\top I_{xy}(w^*) (w^* - \hat{w}_{xy}) \right] + o(1),
\]

where \( r_2(w) \) is the remainder term. The estimator \( \hat{w}_{xy} \) has asymptotic normality, and it converges to the Gaussian distribution with mean \( w^* \) and variance–covariance matrix \( (nI_{xy}(w^*))^{-1} \). It holds that

\[
E_n \left[ \sum_{i=1}^{n} \ln p(x_i, y_i | w^*) \right] = E_n \left[ \sum_{i=1}^{n} \ln p(x_i, y_i | \hat{w}_{xy}) \right] - \frac{\dim w}{2} + o(1).
\]

The free-energy function has the asymptotic form

\[
F_{xy}(n) = - \frac{\dim w}{2} + \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi \\
+ \frac{1}{2} \ln \det I_{xy}(w^*) - \ln \varphi(w^* | \eta) + o(1).
\]

In the same way, we obtain that

\[
F_{xy,x}(n) = - \frac{\dim w}{2} + \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi \\
+ \frac{1}{2} \ln \det \{\alpha I_{xy}(w^*) + (1 - \alpha) I_z(w^*)\} - \ln \varphi(w^* | \eta) + o(1).
\]

Based on the relation in Eq. 8, we obtain that

\[
(1 - \alpha) n D_3(n) = \frac{1}{2} \ln \det \left\{ I_{xy}(w^*) (\alpha I_{xy}(w^*) + (1 - \alpha) I_z(w^*))^{-1} \right\} + o(1),
\]

which is the asymptotic form of \( D_3(n) \).
Define the free-energy functions by

\[ F_{y|x,xy}(n) = \mathbb{E}_n \left[ \ln \prod_{i=1}^{\alpha n} q(y_i|x_i) \prod_{i=\alpha n+1}^{n} q(x_i, y_i) \right. \]
\[ - \ln \int \prod_{i=1}^{\alpha n} p(y_i|x_i, w) \prod_{i=\alpha n+1}^{n} p(x_i, y_i|w) \varphi(w|\eta) dw \right], \]
\[ F_{y|x,x}(n) = \mathbb{E}_n \left[ \ln \prod_{i=1}^{\alpha n} q(y_i|x_i) \prod_{i=\alpha n+1}^{n} q(x_i) \right. \]
\[ - \ln \int \prod_{i=1}^{\alpha n} p(y_i|x_i, w) \prod_{i=\alpha n+1}^{n} p(x_i|w) \varphi(w|\eta) dw \right]. \]

The error function can be rewritten as

\[ (1 - \alpha) n D_2(n) = F_{y|x,xy}(n) - F_{y|x,x}(n). \] (9)

The maximum-likelihood estimator is defined by

\[ \hat{w}_{y|x,xy} = \arg \max_w \prod_{i=1}^{\alpha n} p(y_i|x_i, w) \prod_{i=\alpha n+1}^{n} p(x_i, y_i|w). \]

Due to (A0), the estimators \( \hat{w}_{y|x,xy} \) and \( \hat{w}_2 \) converge to \( w^* \), which means that the essential parameter area for the integration is the neighborhood of \( w^* \), \( \hat{w}_{y|x,xy} \), and \( \hat{w}_2 \). According to the Taylor expansion and the saddle-point approximation, the free-energy functions have the following asymptotic forms:

\[ F_{y|x,xy}(n) = - \frac{\dim w}{2} + \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi \]
\[ + \frac{1}{2} \ln \det \{I_{xy}(w^*) - \alpha I_x(w^*)\} - \ln \varphi(w^*|\eta) + o(1), \]
\[ F_{y|x,xy}(n) = - \frac{\dim w}{2} + \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi \]
\[ + \frac{1}{2} \ln \det \{\alpha I_{xy}(w^*) + (1 - 2\alpha) I_x(w^*)\} - \ln \varphi(w^*|\eta) + o(1). \]

Based on the relation in Eq. [9] we obtain that

\[ (1 - \alpha) n D_2(n) = \frac{1}{2} \ln \det \left\{ K_{21}(w^*) K_{22}(w^*)^{-1} \right\} + o(1), \]
\[ K_{21}(w) = I_{xy}(w) - \alpha I_x(w), \]
\[ K_{22}(w) = \alpha I_{xy}(w) + (1 - 2\alpha) I_x(w), \]

which is the asymptotic form of \( D_3(n) \). (End of Proof)
Proof of Theorem 4

According to the condition, let the eigenvalues of $I(w^*)I_x(w^*)$ be
\[ \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\tilde{d}} > 0, \]
\[ \sigma_{\tilde{d}+1} = \cdots = \sigma_{d} = 0, \]
where $\tilde{d} = \dim \tilde{w}$. First, we compare $D_1(n)$ and $D_3(n)$. Focusing on the factor $\ln \det K_3(w^*)$ of the dominant term in $D_3(n)$, we obtain
\[
K_3(w^*) = \ln \det I_{xy}(w^*) - \ln \det \{\alpha I_{xy}(w^*) + (1 - \alpha)I_x(w^*)\}
= \ln \det \{I(w^*) + I_x(w^*)\} - \ln \det \{\alpha I(w^*) + I_x(w^*)\}
= \ln \det \{I(w^*)I_x(w^*)^{-1} + E\} - \ln \det \{\alpha I(w^*)I_x(w^*)^{-1} + E\}
= \sum_{i=1}^{d} \ln (\sigma_i + 1) - \sum_{i=1}^{d} \ln (\alpha \sigma_i + 1)
= \sum_{i=1}^{\tilde{d}} \ln \frac{\sigma_i + 1}{\alpha \sigma_i + 1}
= \tilde{d} \ln \frac{1}{\alpha} + \sum_{i=1}^{\tilde{d}} \ln \left(1 + \frac{1 - 1/\alpha}{\sigma_i + 1/\alpha}\right),
\]
where $E$ is the $d \times d$ unit matrix and the relation $I(w) = I_{xy}(w) - I_x(w)$ was applied. Because $\frac{1 - 1/\alpha}{\sigma_i + 1/\alpha} < 0$, the second term in the last expression is less than zero. Thus,
\[
K_3(w^*) < \tilde{d} \ln 1/\alpha,
\]
which shows that
\[
D_3(n) - D_1(n) = \frac{1}{2} \{K_2(w^*) - \tilde{d} \ln 1/\alpha\} \frac{1}{(1 - \alpha)n} + o\left(\frac{1}{n}\right) < 0.
\]
Next, we compare $D_1(n)$ and $D_2(n)$.

\[
K_2(w) = \ln \det \left\{ I_{xy}(w^*) - \alpha I_x(w^*) \right\} - \ln \det \left\{ \alpha I_{xy}(w^*) + (1 - 2\alpha) I_x(w^*) \right\}
\]
\[
= \ln \det \left\{ I(w^*) + (1 - \alpha) I_x(w^*) \right\} - \ln \det \left\{ \alpha I(w^*) + (1 - \alpha) I_x(w^*) \right\}
\]
\[
= \ln \det \left\{ I(w^*) I_x(w^*)^{-1} + (1 - \alpha) E \right\} - \ln \det \left\{ \alpha I(w^*) I_x(w^*)^{-1} + (1 - \alpha) E \right\}
\]
\[
= \sum_{i=1}^{d} \ln(\sigma_i + (1 - \alpha)) - \sum_{i=1}^{d} \ln(\alpha \sigma_i + (1 - \alpha))
\]
\[
= \tilde{d} \ln \frac{1}{\alpha} + \sum_{i=1}^{d} \ln \frac{\sigma_i + 1 - \alpha}{\sigma_i + (1 - \alpha)/\alpha}
\]
\[
= \tilde{d} \ln \frac{1}{\alpha} + \sum_{i=1}^{d} \ln \left( 1 + \frac{(1 - \alpha)(1 - 1/\alpha)}{\sigma_i + (1 - \alpha)/\alpha} \right)
\]
\[
= \tilde{d} \ln \frac{1}{\alpha} + \sum_{i=1}^{d} \ln \left( 1 + \frac{1 - 1/\alpha}{\sigma_i/(1 - \alpha) + 1/\alpha} \right).
\]

Because $\frac{1 - 1/\alpha}{\sigma_i/(1 - \alpha) + 1/\alpha} < 0$, the second term in the last expression is less than zero. Thus,

\[
K_2(w^*) < \tilde{d} \ln 1/\alpha,
\]

which shows that

\[
D_2(n) - D_1(n) = \frac{1}{2} \left\{ K_3(w) - \tilde{d} \ln 1/\alpha \right\} \frac{1}{(1 - \alpha)n} + o \left( \frac{1}{n} \right) < 0.
\]

Comparing $K_3(w^*)$ and $K_2(w^*)$, we find that

\[
\sum_{i=1}^{d} \ln \left( 1 + \frac{1 - 1/\alpha}{\sigma_i + 1/\alpha} \right) < \sum_{i=1}^{d} \ln \left( 1 + \frac{1 - 1/\alpha}{\sigma_i/(1 - \alpha) + 1/\alpha} \right).
\]

Therefore, $K_3(w^*) < K_2(w^*)$, which shows that $D_3(n) < D_2(n)$. \textbf{(End of Proof)}

\section*{Proof of Theorem 5}

Define the free-energy function by

\[
F_{y|x}(n) = E_n \left[ \ln \prod_{i=1}^{n} q(y_i|x_i) - \ln \int \prod_{i=1}^{n} p(y_i|x_i, w) \psi(w|\eta) dw \right].
\]
The error function can be rewritten as

\[ (1 - \alpha)nD_1(n) = E_D \left[ \sum_{y_2} q(Y_2|D) \left\{ \ln \int \prod_{i=1}^{n} p(y_i|x_i, w) \varphi(w|\eta) dw \right. \right. \]
\[ - \ln \left. \left. \int \prod_{i=1}^{\alpha n} p(y_i|x_i) \varphi(w|\eta) dw \right\} \right] \]
\[ = F_{y|x}(n) - F_{y|x}(\alpha n). \tag{10} \]

It is sufficient to calculate the asymptotic form of \( F_{y|x}(n) \). The maximum-likelihood estimator is defined by

\[ \hat{w}_n = \arg \max_w \prod_{i=1}^{n} p(y_i|x_i, w). \]

Due to (A2), the estimator converges to \( w_0 \), which means that the essential parameter area for the integration is the neighborhood of \( w_0 \) and \( \hat{w}_n \). According to the Taylor expansion at \( w = \hat{w}_n \),

\[ E_n \left[ \ln \int \prod_{i=1}^{n} p(y_i|x_i, w) \varphi(w|\eta) dw \right] \]
\[ = E_n \left[ \ln \int \exp \left\{ n \frac{1}{n} \sum_{i=1}^{n} \ln p(y_i|x_i, w) \right\} \varphi(w|\eta) dw \right] \]
\[ = E_n \left[ \ln \int \exp \left\{ n \frac{1}{n} \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right. \right. \]
\[ + \frac{1}{n}(w - \hat{w}_n)^\top \frac{\partial}{\partial w} \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \]
\[ + \frac{1}{2}(w - \hat{w}_n)^\top \frac{\partial^2}{\partial w^2} \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n)(w - \hat{w}_n) + r_1(w) \}
\[ \varphi(w|\eta) dw \right], \]
where $r_1(w)$ is the remainder term. Based on the saddle-point approximation,

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

$$= E_n \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right]$$

$$+ E_n \left[ \ln \int \exp(nr_1(w)) \varphi(w|\eta) \mathcal{N}(\hat{w}_n, (nJ(w_0))^{-1}) dw \right]$$

$$- \left\{ \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det J(w_0) \right\} + o(1)$$

$$= E_n \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right] \frac{\dim w}{2} \ln n$$

$$+ \frac{\dim w}{2} \ln 2\pi - \frac{1}{2} \ln \det J(w_0) + \ln \varphi(w_0) + o(1).$$

Thus, we obtain

$$F_{y|x}(n) = E_n \left[ \sum_{i=1}^{n} \ln \frac{q(y_i|x_i)}{p(y_i|x_i, w_0)} \right] + E_n \left[ \sum_{i=1}^{n} \ln \frac{p(y_i|x_i, w_0)}{p(y_i|x_i, \hat{w}_n)} \right]$$

$$+ \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det J(w_0) - \ln \varphi(w_0|\eta) + o(1).$$

According to the Taylor expansion at $\hat{w}_n$,

$$E_n \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, w_0) \right]$$

$$= E_n \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right]$$

$$+ (w_0 - \hat{w}_n)^\top \frac{\partial}{\partial w} \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n)$$

$$+ \frac{1}{2} (w_0 - \hat{w}_n)^\top \frac{\partial^2}{\partial w^2} \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) (w_0 - \hat{w}_n) + o(1)$$

$$= E_n \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right] - nE_n \left[ (w_0 - \hat{w}_n)^\top J(w_0)(w_0 - \hat{w}_n) \right] + o(1),$$

where $r_2(w)$ is the remainder term. The estimator $\hat{w}_n$ has asymptotic normality, and it converges to the Gaussian distribution with mean $w_0$ and variance–covariance.
matrix \((nJ(w_0)I(w_0)^{-1}J(w_0))^{-1}\). It holds that
\[
E_n \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, w_0) \right] = E_n \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right] - \frac{1}{2} \text{Tr} \{ I(w_0)J(w_0)^{-1} \} + o(1).
\]

The free-energy function has the asymptotic form
\[
F_{y|x}(n) = nE \left[ \ln \frac{q(y|x)}{p(y|x, w_0)} \right] - \frac{1}{2} \text{Tr} \{ I(w_0)J(w_0)^{-1} \}
+ \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det J(w_0) - \ln \varphi(w_0|\eta) + o(1).
\]

Based on the relation in Eq. 10, we obtain
\[
(1 - \alpha) nD_1(n) = (1 - \alpha) nE \left[ \ln \frac{q(y|x)}{p(y|x, w_0)} \right] - \frac{\dim w}{2} \ln \alpha + o(1),
\]
which completes the proof. (End of Proof)

**Proof of Theorem 6**

The first term of the asymptotic form in Theorem 5 can be rewritten as
\[
E \left[ \ln \frac{q(y|x)}{p(y|x, w_0)} \right] = E[\ln q(y|x)] - E_{\alpha n} \left[ \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} \ln p(y_i|x_i, w_0) \right].
\]

According to (A2), the estimator \(\hat{w}_\alpha\) also converges to \(w_0\). The Taylor expansion at \(\hat{w}_\alpha\) changes the second term into
\[
- E_{\alpha n} \left[ \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} \ln p(y_i|x_i, \hat{w}_\alpha) \right]
- E_{\alpha n} \left[ \frac{1}{\alpha n} (w_0 - \hat{w}_\alpha) \right] \sum_{i=1}^{\alpha n} \frac{\partial}{\partial w} \ln p(y_i|x_i, \hat{w}_\alpha)
- \frac{1}{2} E_{\alpha n} \left[ \frac{1}{\alpha n} (w_0 - \hat{w}_\alpha) \right] \sum_{i=1}^{\alpha n} \frac{\partial^2}{\partial w^2} \ln p(y_i|x_i, \hat{w}_\alpha)(w_0 - \hat{w}_\alpha) + o\left(\frac{1}{n}\right)
= - E_{\alpha n} \left[ \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} \ln p(y_i|x_i, \hat{w}_\alpha) \right]
+ \frac{1}{2} E_{\alpha n} \left[ (w_0 - \hat{w}_\alpha) J(w_0)(w_0 - \hat{w}_\alpha)^\top \right] + o\left(\frac{1}{n}\right).
\]
The estimator \( \hat{w}_\alpha \) has asymptotic normality, and it converges to the Gaussian distribution with mean \( w_0 \) and variance–covariance matrix \((\alpha n J(w_0)I(w_0)^{-1}J(w_0))^{-1}\).

The second term of the last expression can be described as
\[
\frac{1}{2} E_{\alpha n} \left[ (w_0 - \hat{w}_\alpha)^T J(w_0)(w_0 - \hat{w}_\alpha) \right] = \frac{1}{2\alpha n} \text{Tr}\{I(w_0)J(w_0)^{-1}\} + o\left(\frac{1}{n}\right).
\]

Then, we obtain
\[
E \left[ \ln \frac{q(y|x)}{p(y|x, w_0)} \right] = E[\ln q(y|x)] - E_{\alpha n} \left[ \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} \ln p(y_i|x_i, \hat{w}_\alpha) \right] + \frac{1}{2\alpha n} \text{Tr}\{I(w_0)J(w_0)^{-1}\} + o\left(\frac{1}{n}\right),
\]
which completes the proof. (End of Proof)

**Proof of Theorem 9**

Due to (A1-1), the matrices given by
\[
J_1(w) = \int \sum_y \frac{\partial^2 \ln p(y|x, w)}{\partial w_i \partial w_j} p(y|x, w) q_1(x) dx,
\]
\[
J_2(w) = \int \sum_y \frac{\partial^2 \ln p(y|x, w)}{\partial w_i \partial w_j} p(y|x, w) q_2(x) dx
\]
have the relations \( I_1(w) = J_1(w) \) and \( I_2(w) = J_2(w) \), respectively.

Similar to Eq. 10, the error \( D(n) \) has the following expression:
\[
(1 - \alpha) n D(n) = F_{12}(n) - F_1(\alpha n),
\]
\[
F_{12}(n) = E_{12} \left[ \sum_{i=1}^{n} \ln q(y_i|x_i) - \ln \prod_{i=1}^{n} p(y_i|x_i, w) \varphi(w|\eta) dw \right],
\]
\[
F_1(\alpha n) = E_{\alpha n} \left[ \sum_{i=1}^{\alpha n} \ln q(y_i|x_i) - \ln \prod_{i=1}^{\alpha n} p(y_i|x_i, w) \varphi(w|\eta) dw \right],
\]
where the expectation is
\[
E_{12}[f(X^n, Y^n)] = \int \sum_{Y^n} f(X^n, Y^n) q(Y^n|X^n) q_1(X_1) q_2(X_2) dX^n.
\]
Because $F_1(\alpha n) = F_{y|x}(\alpha n)$ and $w_0 = w^*$, the following asymptotic form of $F_1(\alpha n)$ is immediately obtained from the proof of Theorem 5:

$$F_1(\alpha n) = -\frac{\dim w}{2} + \frac{\dim w}{2} \ln \alpha n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det I_1(w^*) - \ln \varphi(w^*|\eta) + o(1).$$

Due to (A1-1) and the saddle-point approximation, it holds that

$$F_{12}(n) = E_{12} \left[ \sum_{i=1}^{n} \frac{p(y_i|x_i, w^*)}{p(y_i|x_i, \hat{w}_n)} \right] + \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det \{\alpha J_1(w^*) + (1 - \alpha) J_2(w^*)\} - \ln \varphi(w^*|\eta) + o(1),$$

where the estimator $\hat{w}_n$ converges to $w^*$. In the same way as in the proof of Theorem 5, the Taylor expansion at $\hat{w}_n$ enables us to obtain

$$E_{12} \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right] = E_{12} \left[ \sum_{i=1}^{n} \ln p(y_i|x_i, \hat{w}_n) \right] - \text{Tr}\{\alpha I_1(w^*) + (1 - \alpha) I_2(w^*)\} \ln \varphi(w^*|\eta) + o(1).$$

By using the equivalence between $I_i(w)$ and $J_i(w)$ for $i = 1, 2$, the asymptotic form of $F_{12}(n)$ can be written as

$$F_{12}(n) = -\frac{\dim w}{2} + \frac{\dim w}{2} \ln n - \frac{\dim w}{2} \ln 2\pi + \frac{1}{2} \ln \det \{\alpha I_1(w^*) + (1 - \alpha) I_2(w^*)\} - \ln \varphi(w^*|\eta) + o(1).$$

Calculating the difference $F_{12}(n) - F_{y|x}(\alpha n)$, we obtain the asymptotic form in Theorem 6 (End of Proof).

**Proof of Theorem 10**

Assume that the free-energy function has the asymptotic form

$$F_{y|x}(n) = \lambda \ln n + O(1).$$

Note that, because of the existence of $w^*$, there is no $n$th-order term. According to the conditions of Theorem 10, Eq. 10 still holds. This indicates that the error function can be described as

$$D(n) = \lambda \ln 1/\alpha \frac{1}{1 - \alpha n} + o \left( \frac{1}{n} \right).$$
Even if the Fisher information matrix is not positive definite, the algebraic geometrical method provides a method for calculating $F_{y|x}(n)$ \cite{Watanabe2001}. More precisely, it has been proved that the energy function has the asymptotic form

$$F_{y|x}(n) = \lambda \ln - (m - 1) \ln \ln n + O(1).$$

The coefficients $\lambda$ and $m$ are obtained as follows. Let a Kullback–Leibler divergence be defined by

$$H(w) = \int \sum_y p(y|x, w^*) q(x) \ln \frac{p(y|x, w^*)}{p(y|x, w)} dx,$$

and assume that $H(w)$ is analytic. It is known that the poles of the zeta function

$$\zeta(z) = \int H(w)^z \varphi(w|\eta) dw$$

are real, negative, and rational. The largest pole, $z = -\lambda$ with multiplicity $m$, determines the coefficients of $F_{y|x}(n)$.

Before calculating the pole, we state without proof some useful lemmas for the zeta function.

**Lemma 11** 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 12** Under the same conditions as in Lemma 11, 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 can define an equivalence relation $H_1(w) \equiv H_2(w)$ when $\lambda_1 = \lambda_2$ in Lemma 12. Considering the Newton diagram \cite{Yamazaki2010}, a polynomial of $w$ has a simpler equivalent expression.

**Lemma 13** Let $H_1(w)$ be a polynomial of $w = \{w_1, \ldots, w_d\}$, and let $H_2(w \setminus w_1)$ be the same as $H_1(w)$ except for the terms with the factor $w_i^i$ for $i \geq 1$. It holds that

$$cw_1 + H_1(w) \equiv cw_1 + H_2(w \setminus w_1),$$

where $c$ is a constant.
In the discriminant model of Theorem 10, the function $H(w)$ can be written as follows:

$$H(w) = E \left[ \ln \frac{a_y \exp \left\{ \sum_{i=1}^{K^*} a_i \exp \left\{ \sum_{m=1}^{M} \frac{(x_m - b_{im})^2}{2\sigma^*} \right\} \right\}}{\sum_{i=1}^{K^*} a_i \exp \left\{ \sum_{m=1}^{M} \frac{(x_m - b_{im})^2}{2\sigma^*} \right\}} \right]$$

$$= E \left[ \ln \left\{ 1 + \sum_{i=1}^{K^*} \exp \left\{ - \sum_{m=1}^{M} \frac{(x_m - b_{im})^2}{2\sigma^*} + \sum_{m=1}^{M} \frac{(x_m - b_{im})^2}{2\sigma^*} - \ln a_i \right\} \right\} \right]$$

Using a reparameterization similar to that in Example 2, we obtain

$$H(w) = E \left[ \ln \frac{1 + \sum_{i=1}^{K^*} \exp \left\{ \sum_{m=1}^{M} (c_{im} x_m + d_{im}) + \sum_{i=1}^{K^*} a_i f_i(w) \right\}}{1 + \sum_{i=1}^{K^*} \exp \left\{ \sum_{m=1}^{M} (c_{im}^* + d_{im}^*) \right\}} \right],$$

where the function $f_i$ stands for the exponential part. The pole of the zeta function is affected by the neighborhood of $W_i$ (Watanabe [2001]). Then, $c_{im}$ and $d_{im}$ are close to $c_{im}^*$ and $d_{im}^*$, respectively. Let the $\Phi_1(w)$ be defined by

$$u_{im} = c_{im} - c_{im}^*, \quad v_{im} = d_{im} - d_{im}^*,$$

for $1 \leq i \leq K^*, 1 \leq m \leq M$. We obtain the following expression:

$$H(\Phi_1(w)) = E \left[ \ln \frac{1 + \sum_{i=1}^{K^*} \sum_{m=1}^{M} g_{im}(\Phi_1(w)) + \sum_{i=1}^{K^*} a_i f_i(\Phi_1(w))}{1 + \sum_{i=1}^{K^*} \exp \left\{ \sum_{m=1}^{M} (c_{im}^* + d_{im}^*) \right\}} \right],$$

where $g_{im}$ is given by

$$g_i(\Phi_1(w)) = \exp(c_{im}^* x_m + d_{im}^*) \{1 - \exp(u_{im} x_m + v_{im})\}.$$

In this neighborhood, the numerator is close to zero because $u_{im}, v_{im}$ for $(1 \leq i \leq K^*)$ and $a_i$ for $(i > K^*)$ are all zero in $W_i$. Based on the Taylor expansion, we find that

$$H(\Phi(w)) = E \left[ \frac{\sum_{i=1}^{K^*} \sum_{m=1}^{M} g_{im}(\Phi_1(w)) + \sum_{i=1}^{K^*} a_i f_i(\Phi_1(w))}{1 + \sum_{i=1}^{K^*} \exp \left\{ \sum_{m=1}^{M} (c_{im}^* + d_{im}^*) \right\}} + \ldots \right].$$

33
Since $f_i$ is the exponential function, there exists a positive constant $\beta_i$ such that $f_i(\Phi_1(w)) = \beta_i + \ldots$. Based on Lemma 13, it holds that

$$H(\Phi_1(w)) \equiv \sum_{i=K^*+1}^K \beta_i a_i + E \left[ \frac{\sum_{i \neq y}^{K^*} \sum_{i=1}^M g_{im}(\Phi_1(w))}{1 + \sum_{i \neq y}^{K^*} \exp\left(\sum_{m=1}^M (c^*_i + d^*_i)\right)} \right] + \ldots,$$

where the terms with the factor $a_i^j$ for $i \geq K^* + 1$ and $j \geq 1$ are excluded, and the rest of the terms are the same as in $H(\Phi_1(w))$. This indicates that

$$H(\Phi_1(w)) \equiv \sum_{i=K^*+1}^K a_i + E \left[ \ln \left(1 + \frac{\sum_{i \neq y}^{K^*} \sum_{i=1}^M g_{im}(\Phi_1(w))}{1 + \sum_{i \neq y}^{K^*} \exp\left(\sum_{m=1}^M (c^*_i + d^*_i)\right)} \right] \right].$$

The second term corresponds to the case $K = K^*$. When the Fisher information matrix is positive definite, $H(w)$ is equivalent to the quadratic function of the necessary parameter for $q(y|x)$ based on the proper reparameterization. Let $\bar{w}$ be the necessary parameter, where $\dim \bar{w} = K^*M$. It holds that

$$H(\Phi_1(w)) \equiv \sum_{i=K^*+1}^K a_i + \sum_{i=1}^{K^*M} \bar{w}_i^2.$$

According to Lemma 12 the following zeta function has the same pole as the original one:

$$\zeta_{alg}(z) = \int \left( \sum_{i=1}^{K^*M} \bar{w}_i^2 + \sum_{i=K^*+1}^K a_i \right)^z \varphi(\Phi_2(w)|\eta) |\Phi_2| d\Phi_2(w),$$

where $\Phi_2(w)$ is the total transformation of the parameter, and $|\Phi_2|$ is its Jacobian. Consider the mapping $\Phi_3$:

$$a_i = \gamma_i \omega_1^2 (K^* + 1 \leq i \leq K),$$
$$\bar{w}_1 = \omega_1,$$
$$\bar{w}_i = \omega_i \omega_1 (1 \leq i \leq K^*M);$$

in algebraic geometry, this is referred to as blowing up. Using this mapping, the zeta function is described as

$$\zeta_{alg}(z) = \int \omega_1^{2z} \left(1 + \sum_{i=2}^{K^*M} \omega_i^2 + \sum_{i=K^*+1}^K \gamma_i \right)^z \varphi(\Phi_3(\Phi_2(w))) |\Phi_3| \Phi_2 d\Phi_3\Phi_2(w).$$
We focus on the factors with $\omega_1$. The Jacobian $|\Phi_3\Phi_2|$ has $\omega_1^{K^*M-1+2(K-K^*)}$, and the prior distribution $\varphi(\Phi_3\Phi_2(w))$ has $\omega_1^{2(K-K^*)(m-1)}$. Integration over $\omega_1$ shows a pole,

$$\zeta_{\text{alg}}(z) = \int \omega_1^{2z+K^*M-1+2(K-K^*)+2(K-K^*)(m-1)} d\omega_1 h(z)$$

$$= \frac{C}{2z + K^*M + 2(K - K^*)\eta_1} h(z),$$

where $C$ is a positive constant, and $h(z)$ is the remaining part of the function. The last expression clearly reveals that there is a pole

$$z = -\frac{K^*M}{2} - (K - K^*)\eta_1$$

with multiplicity $m = 1$. By considering blowing up with different parameters, we can easily see that this pole is the largest one. (End of Proof)

References

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

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. and Watanabe, S. (2005). Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 18, 924–933.

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

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. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 227–244.

Watanabe, S. (2001). Algebraic analysis for non-identifiable learning machines. *Neural Computation*, 13 (4), 899–933.
Watanabe, S. (2010). Equations of states in singular statistical estimation. *Neural Networks, 23*(1), 20–34.

Yamazaki, K. (2012a). Asymptotic accuracy of bayes estimation for latent variables with redundancy. arXiv:1205.3234.

Yamazaki, K. (2012b). Asymptotic accuracy of distribution-based estimation for latent variables. arXiv:1204.2069.

Yamazaki, K., Nagata, K., Watanabe, S., and Müller, K.-R. (2006). A model selection method based on bound of learning coefficient. In *LNCS*, volume 4132, pages 371–380. Springer.

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

Zhu, X. (2007). Semi-supervised learning literature survey. Technical Report TR1530, Computer Science, University of Wisconsin Madison.

Zwiernik, P. (2011). An asymptotic behaviour of the marginal likelihood for general markov models. *J. Mach. Learn. Res., 999888*, 3283–3310.