Neural Networks 69 (2015) 11–19

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Asymptotic accuracy of Bayesian estimation for a single latent variable

Keisuke Yamazaki

Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, G5-19 4259 Nagatsuta, Midori-ku, Yokohama, Japan

ARTICLE INFO

Article history:
Received 7 October 2014
Received in revised form 20 April 2015
Accepted 30 April 2015
Available online 8 May 2015

Keywords:
Unsupervised learning
Hierarchical parametric models
Latent variable
Bayes estimation

ABSTRACT

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

1. Introduction

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

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

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

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

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

The estimation of an unseen observable variable is referred to as a prediction. Let a set of the given data be $X^n = \{x_1, \ldots, x_n\}$. The task is to predict the next data position based on the given data; this is formulated as the estimation of the probability $p(x_{n+1}|X^n)$. In order to measure the accuracy of the task, we define the error function to be the Kullback–Leibler divergence,

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

E-mail address: k-yam@math.dis.titech.ac.jp.

http://dx.doi.org/10.1016/j.neunet.2015.04.014
0893-6080/

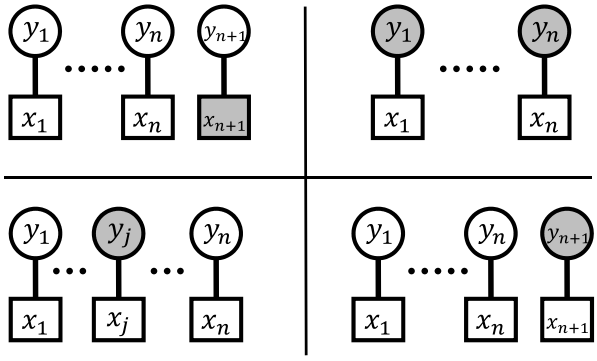

Fig. 1. Predictions of observable variables and estimations of latent variables. The observable data are $\{x_1, \ldots, x_n\}$. Rectangles and circles represent the observable and unobservable variables, respectively. Gray nodes are the estimation targets.

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

The estimation of the latent variables is not the same as the prediction task. The target variable of the estimation is unobservable, and in many practical situations, its true value is not given; this makes it difficult to evaluate the result. In a previous study (Yamazaki, 2014), we formulated the accuracy of the latent-variable estimation in a distribution-based manner. The estimation of latent variables is divided into three classes. Let a set of latent variables be $Y^n = \{y_1, \ldots, y_n\}$, where y_i is the corresponding variable to x_i. Fig. 1 shows the prediction of observable variables and the three types of estimations of latent variables. Rectangles and circles indicate the observable and latent variables, respectively. The gray nodes are the targets of the estimations. The top left panel shows the prediction, which is expressed as the estimation of $p(x_{n+1}|X^n)$. The top right panel shows the estimation of the joint probability $p(Y^n|X^n)$, in which all of the latent variables are targets; we will refer to this as Type I. The bottom left panel shows the estimation of the probability of a specific latent variable $p(y_j|X^n)$; we will refer to this as Type II. The bottom right panel shows the estimation of the probability of a latent variable in the unseen data $p(y_{n+1}|X^n)$; we will refer to this as Type III. In the example of a Gaussian mixture, these three types of latent-variable estimation correspond to the cluster analysis process of assigning labels to data.

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

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

Although there are a number of studies that consider the estimation of observable variables and the convergence of the parameters, the theory of estimating latent variables has not been thoroughly analyzed. The error functions of Types I, II, and III are defined as the Kullback–Leibler divergence from the data-generating distribution to the estimated one, and its theoretical behavior has been analyzed. The error function of Type III with the maximum-likelihood method has been derived, and a model-selection criterion has been proposed for the regular case (Shimodaira, 1993). The asymptotic expansions of Type I in the Bayes method and of the rest of the types in the maximum-likelihood method have been calculated for the regular case, and we found that with the maximum-likelihood method, their asymptotic errors are equivalent and that for Type I, the Bayes method is more accurate than the maximum-likelihood method. The singular case has been considered, and its error has been derived only for Type I (Yamazaki, in press).

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

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

2. Three types of estimations of latent variables

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

2.1. Formulation of a hierarchical probabilistic model

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

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

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

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

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

Let $\{X^n, Y^n\} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the i.i.d. dataset 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.