# DISTRIBUTION-BASED ESTIMATION OF THE LATENT VARIABLES AND ITS ACCURACY

*Keisuke Yamazaki*

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

## ABSTRACT

Hierarchical probabilistic models such as a mixture of distributions and a hidden Markov model are widely used for unsupervised learning. They consist of the observable and the latent variables, which represent the observed data and the underlying data-generating process, respectively. There are two type of use due to the estimated variable; the prediction of future/unseen data is the observable-variable (OV) estimation, and the analysis how the given data were generated is the latent-variable (LV) estimation. The asymptotic accuracy of the OV estimation has been elucidated in many models. On the other hand, the LV estimation has not sufficiently been studied. In this talk, the error function to measure the accuracy of the LV estimation is formulated, and its asymptotic form is derived for the maximum-likelihood (ML) and the Bayes methods. The results provide a distribution-based evaluation of the unsupervised learning, and show that the Bayes method has the better accuracy than the ML method.

## 1. INTRODUCTION

Hierarchical probabilistic models such as a mixture of distributions and a hidden Markov model are widely used for unsupervised learning tasks. One of representative models is a mixture of Gaussian distributions in the cluster analysis; the task is to find clusters in data and each cluster is approximated by a Gaussian distribution. The hierarchical model consists of two types of variables: the observable and the latent variables. The observable variable represents the observable parts of the given data, and the latent one does the hidden parts. In the example of the cluster analysis, the observable variable expresses the data location and the latent one corresponds to the label showing which cluster the data belong to. The unsupervised learning task is interpreted as estimation of the latent variable based on the observable one.

The use of the models falls into two ways: prediction of unseen observable variable and estimation of the latent variables for the given data, which we refer to as the OV and the LV estimations, respectively. From the theoretical point of view, there have been many studies on analysis of the accuracy of the OV estimation. The error function measuring the accuracy is the generalization error and its asymptotic form has been derived in many models. Moreover, the form is used for selecting optimal structure of the model [1].

However, the accuracy of the LV estimation has not been clarified yet. This talk introduces the maximum-likelihood (ML) and the Bayes methods for the LV estimation. The error function is formulated and its asymptotic form is derived. Comparing these method, the asymptotic result shows the advantage of the Bayes method.

## 2. FORMAL DEFINITIONS OF THE LV ESTIMATION

This section introduces the definitions of the LV estimation and its accuracy in the ML and the Bayes methods, which is the main part of the talk and a brief summary of [2].

Let the observable and hidden parts of data be $x \in R^M$ and $y \in \{1, \ldots, K\}$, respectively. The hierarchical model is expressed as $p(x, y|w) = p(x|y, w)p(y|w)$, where $w$ is the parameter. A Gaussian mixture model has the form

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

where $w = \{a_k, b_k\}$ and $\mathcal{N}(x|b)$ is the Gaussian distribution with the parameter $b$. The mixing ratio $a_k$ satisfies that $a_k \geq 0$ and $\sum_{k=1}^{K} a_k = 1$. We can easily confirm that $p(x, y|w) = a_y \mathcal{N}(x|b_y)$, $p(y|x) = a_y$, and $p(x|y, w) = \mathcal{N}(x|b_y)$.

Assume that there is the true distribution $q(x, y) = q(x|y)q(y)$ generating i.i.d. data

$$\{X^n, Y^n\} = (x_1, y_1), \ldots, (x_n, y_n),$$

where $x_i$ is the observable part of data and $y_i$ is the corresponding hidden part. The task in the unsupervised learning is to estimate $Y^n = \{y_1, \ldots, y_n\}$ on the basis of $X^n = \{x_1, \ldots, x_n\}$. The result of the LV estimation is expressed as the distribution $p(Y^n|X^n)$.

The ML and the Bayes methods are representative in the OV estimation. The target distribution of the OV estimation is $p(x|X^n)$, where $x$ represents unseen/future ob-

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

servable data. The maximum likelihood estimator is defined by

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

The ML method is given by

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

In the Bayes method, the posterior distribution is written as

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

where $\varphi(w)$ is a prior distribution and $Z(X^n)$ is the normalization factor. The predictive distribution is then defined by

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

The ML method uses the optimal value of the parameter while the Bayes method does the distribution of the parameter for expectation.

Now, we introduce these two methods of the LV estimation. The ML method is given by

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

and the Bayes method is defined by

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

which has the equivalent definition,

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

We can find that the ML method is plug-in and the Bayes method is the expectation.

Let us evaluate the estimation result. In the OV estimation, the error function to measure the accuracy is the generalization error. The KL divergence is often used for the error function,

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

where the expectation means that

$$E[f(X^n)] = \int f(X^n)q(X^n)dX^n.$$

The asymptotic form of the error has been studied well. Specifically, in the Bayes method, the relation to algebraic geometry has been found [3], and the form is calculated in many types of the hierarchical models [4, 5, 6, 7, 8, 9, 10, 11]. In the similar way, the error function of the LV estimation is defined by

$$D(n) = \frac{1}{n} E\left[\sum_{Y^n} q(Y^n|X^n) \ln \frac{q(Y^n|X^n)}{p(Y^n|X^n)}\right].$$

In the Bayes method, the error function has a connection to the marginal likelihood, of which the asymptotic form [12] is the essential factor of the analysis. See [2] for the asymptotic forms of the ML and Bayes methods and a comparison of the accuracy.

## 3. DISCUSSION

This section considers the normalized maximum likelihood approach [13] for the LV estimation. The distribution of latent variables are formally written as

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

The normalized likelihood distributions are given by

$$p_{\mathrm{NML}}(X^n, Y^n) = \frac{\max_{w\in W_c} p(X^n, Y^n|w)}{\int \sum_{Y^n} \max_{w\in W_c} p(X^n, Y^n|w)dX^n},$$
$$p_{\mathrm{NML}}(X^n) = \frac{\max_{w\in W_c} p(X^n|w)}{\int \max_{w\in W_c} p(X^n|w)dX^n},$$

where $W_c$ is the compact set and the conditions i)-v) in [13] are satisfied. However, the ratio $p_{\mathrm{NML}}(X^n, Y^n)/p_{\mathrm{NML}}(X^n)$ cannot be the estimation of the normalized ML method because

$$p_{\mathrm{NML}}(X^n) \neq \sum_{Y^n} p_{\mathrm{NML}}(X^n, Y^n),$$

i.e. the ratio is not the probability of $Y^n$. For the proper definition, there are two possible ways; the estimator with respect to $X^n$ defines

$$p(Y^n|X^n) = \frac{\max_{w\in W_c} p(X^n, Y^n|w(X^n))}{\sum_{Y^n} \max_{w\in W_c} p(X^n, Y^n|w(X^n))},$$

and the estimator with respect to $X^n$ and $Y^n$ does

$$p(Y^n|X^n) = \frac{\max_{w\in W_c} p(X^n, Y^n|w(X^n, Y^n))}{\sum_{Y^n} \max_{w\in W_c} p(X^n, Y^n|w(X^n, Y^n))}.$$

The former definition is the same as the one of the ML method. Thus, let the latter one be the normalized ML (NML) method. Using the maximum likelihood of the joint probability

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

we give the equivalent expression of the NML method as

$$p(Y^n|X^n) = \frac{p(X^n, Y^n|\hat{w}_{XY})}{\sum_{Y^n} p(X^n, Y^n|\hat{w}_{XY})}.$$

Let us analyze the asymptotic form of the error function. Based on the definition, the error function is rewritten as

$$nD(n) = E\left[\ln \frac{q(X^n, Y^n)}{q(X^n)}\right] - E\left[\ln \frac{p(X^n, Y^n|\hat{w}_{XY})}{\sum_{Y^n} p(X^n, Y^n|\hat{w}_{XY})}\right],$$

where the expectation means that

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

Define the following distribution

$$p^{(y)}_{\rm NML}(X^n) = \frac{\sum_{Y^n} p(X^n, Y^n|\hat{w}_{XY})}{\int \sum_{Y^n} p(X^n, Y^n|\hat{w}_{XY}) dX^n}.$$

Then, it holds that

$$nD(n) = E\left[\ln \frac{q(X^n, Y^n)}{q(X^n)}\right] - E[\ln p_{\rm NML}(X^n, Y^n)] + E[\ln p_{\rm NML}(X^n)] + E\left[\ln \frac{p^{(y)}_{\rm NML}(X^n)}{p_{\rm NML}(X^n)}\right].$$

For simplicity, we assume that there is $w^* \in W_c$ such that $p(x, y|w^*) = q(x, y)$. Based on the asymptotic form of the normalized maximum likelihood distribution [13], we easily obtain the following bounds;
if $E[-\ln p_{\rm NML}(X^n)] \leq E[-\ln p^{(y)}_{\rm NML}(X^n)]$,

$$D(n) \leq \frac{\ln(c_{XY}/c_X)}{n} + o(1),$$

otherwise,

$$D(n) > \frac{\ln(c_{XY}/c_X)}{n} + o(1),$$

where

$$c_{XY} = \int_{W_c} \sqrt{\det I_{XY}(w)} dw,$$
$$c_X = \int_{W_c} \sqrt{\det I_X(w)} dw,$$
$$\{I_{XY}(w)\}_{ij} = E_{xy}\left[\frac{\partial \ln p(x, y|w)}{\partial w_i} \frac{\partial \ln p(x, y|w)}{\partial w_j}\right],$$
$$\{I_X(w)\}_{ij} = E_{xy}\left[\frac{\partial \ln p(x|w)}{\partial w_i} \frac{\partial \ln p(x|w)}{\partial w_j}\right].$$

The expectation means that

$$E_{xy}[f(x, y)] = \int \sum_{y=1}^{K} f(x, y) p(x, y|w) dx.$$

It is an important future study to derive the asymptotic expression of $E[-\ln p^{(y)}_{\rm NML}(X^n)]$ for the exact form of $D(n)$.

## 4. REFERENCES


[1] Hirotsugu Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, pp. 716–723, 1974.

[2] Keisuke Yamazaki, "Asymptotic accuracy of distribution-based estimation for latent variables," arXiv:1204.2069, 2012.

[3] Sumio Watanabe, "Algebraic analysis for non-identifiable learning machines," *Neural Computation*, vol. 13 (4), pp. 899–933, 2001.

[4] Keisuke Yamazaki and Sumio Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," *International Journal of Neural Networks*, vol. 16, pp. 1029–1038, 2003.

[5] Keisuke Yamazaki and Sumio Watanabe, "Stochastic complexity of bayesian networks," in *Proc. of UAI*, 2003, pp. 592–599.

[6] Keisuke Yamazaki and Sumio Watanabe, "Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities," *IEEE Trans. on Neural Networks*, vol. 16, no. 2, pp. 312–324, 2005.

[7] Keisuke Yamazaki and Sumio Watanabe, "Generalization errors in estimation of stochastic context-free grammar," in *The IASTED International Conference on ASC*, 2005, pp. 183–188.

[8] Miki Aoyagi and Sumio Watanabe, "Stochastic complexities of reduced rank regression in bayesian estimation," *Neural Networks*, vol. 18, pp. 924–933, 2005.

[9] Miki Aoyagi, "Stochastic complexity and generalization error of a restricted boltzmann machine in bayesian estimation," *Journal of Machine Learning Research*, vol. 11, pp. 1243–1272, 2010.

[10] Dmitry Rusakov and Dan Geiger, "Asymptotic model selection for naive bayesian networks," *Journal of Machine Learning Research*, vol. 6, pp. 1–35, 2005.

[11] Piotr Zwiernik, "An asymptotic behaviour of the marginal likelihood for general markov models," *J. Mach. Learn. Res.*, vol. 999888, pp. 3283–3310, Nov. 2011.

[12] Bertrand Clarke and Andrew R. Barron, "Information-theoretic asymptotics of bayes methods," *IEEE Transactions on Information Theory*, vol. 36, pp. 453–471, 1990.

[13] Jorma J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, 1996.