2005 International Symposium on Nonlinear
Theory and its Applications (NOLTA2005)
Bruges, Belgium, October 18-21, 2005

On Variational Bayes Algorithms for Exponential Family Mixtures

Kazuho Watanabe[†] and Sumio Watanabe[‡]

†Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology
‡P& I Lab., Tokyo Institute of Technology
Mail Box:R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan
Email: kazuho23@pi.titech.ac.jp, swatanab@pi.titech.ac.jp

Abstract—In this paper, we empirically analyze the behaviors of the Variational Bayes algorithm for the mixture model. While the Variational Bayesian learning has provided computational tractability and good generalization performance in many applications, little has been done to investigate its properties. Recently, the stochastic complexity of mixture models in the Variational Bayesian learning was clarified. By comparing the experimental results with the theoretical ones, we discuss the properties of the practical Variational Bayes algorithm.

1. Introduction

Mixture models are widely used especially in statistical pattern recognition or data clustering and closely related to several neural network models[2]. The Variational Bayesian (VB) framework was proposed as an approximation of the Bayesian learning for the models with hidden variables including mixture models[1][4]. The VB learning has been applied to various learning machines and it has performed good generalization with only modest computational costs compared to Markov chain Monte Carlo methods that are the major schemes of the Bayesian learning. However, little has been done to investigate the properties of the VB learning itself.

Recently, as an initial discussion of the theoretical properties of the VB learning, the asymptotic form was obtained for the stochastic complexities in the VB learning of mixtures of exponential-family distributions[5]. This enabled us to investigate the properties of the practical VB learning involving an iterative algorithm and suffering from the problems such as local minima.

In this paper, we experimentally analyze the behaviors of the VB algorithm for the mixture model and discuss the properties of it in terms of the redundancy of the model and the hyperparameter in the prior distribution. The VB algorithm is a procedure of minimizing the functional that finally gives the stochastic complexity. We experimentally examine whether the algorithm converges to the optimal solution instead of local minima by a comparison of the experimental results with the theoretical upper bound of the stochastic complexity.

In Section 2, the mixture of exponential family model is introduced. In Section 3, the VB learning is outlined, then the VB algorithm and the theoretical upper bound of the stochastic complexity for the mixture models are described. We present the experimental results in Section 4 and discuss them in Section 5. Finally, conclusion follows in Section 6.

2. Mixture of Exponential Family

Denote by $c(x|b)$ a probability density function of the input $x \in R^N$ given an M-dimensional parameter vector $b = (b^{(1)}, b^{(2)}, \cdots, b^{(M)})^T \in B$ where B is a subset of R^M. The general mixture model $p(x|\theta)$ with a parameter vector θ is defined by

$$p(x|\theta) = \sum_{k=1}^{K} a_k c(x|b_k),$$

where K is the number of components and $\{a_k | a_k \geq 0, \sum_{k=1}^{K} a_k = 1\}$ is the set of mixing proportions. The model parameter θ is $\{a_k, b_k\}_{k=1}^{K}$.

A model $p(x|\theta)$ is called a mixture of exponential family (MEF) model if $c(x|b)$ is given by the form,

$$c(x|b) = \exp\{b \cdot f(x) + f_0(x) - g(b)\}, \tag{1}$$

where $b \in B$ is called the natural parameter, $b \cdot f(x)$ is the inner product with the vector $f(x) = (f_1(x), \cdots, f_M(x))^T$, $f_0(x)$ and $g(b)$ are real-valued functions of the input x and the parameter b, respectively. Suppose functions $f_1, \cdots, f_M$ and a constant function are linearly independent.

The conjugate prior distribution $\varphi(\theta)$ for the MEF model is given by the product of the following two distributions on $\mathbf{a} = \{a_k\}_{k=1}^{K}$ and $\mathbf{b} = \{b_k\}_{k=1}^{K}$

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^k} \prod_{k=1}^{K} a_k^{\phi_0 - 1}, \tag{2}$$

$$\varphi(\mathbf{b}) = \prod_{k=1}^{K} \frac{\exp\{\xi_0(b_k \cdot \nu_0 - g(b_k))\}}{C(\xi_0, \nu_0)}. \tag{3}$$

This work was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for JSPS Fellows 4637 and for Scientific Research 15500130, 2005.

Constants $\xi_0 > 0$, $\nu_0 \in R^M$ and $\phi_0 > 0$ are called hyperparameters and $C(\xi, \mu) = \int \exp\{\xi(\mu \cdot b - g(b))\}db$ is a function of $\xi \in R$ and $\mu \in R^M$.

The mixture model can be rewritten by using a hidden variable $y = (y^1, \cdots, y^K) \in \{(1, 0, \cdots, 0), \cdots, (0, 0, \cdots, 1)\}$, as

$$p(x, y|\theta) = \prod_{k=1}^{K} \Big[a_k c(x|b_k)\Big]^{y^k}.$$

If and only if the datum x is generated from the kth component, $y^k = 1$.

3. Variational Bayesian Learning

Suppose n training samples $X^n = \{x_1, \cdots, x_n\}$ are independently and identically taken from the true distribution $p_0(x)$. The Bayesian posterior distribution is defined by

$$p(\theta|X^n) = \frac{1}{Z(X^n)} \exp(-nH_n(\theta))\varphi(\theta), \qquad (4)$$

where $\varphi(\theta)$ is the prior distribution, $H_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log \frac{p_0(x_i)}{p(x_i|\theta)}$ and $Z(X^n)$ is the normalization constant. The stochastic complexity is defined by

$$F(X^n) = -\log Z(X^n), \qquad (5)$$

which is also called the free energy and is important in most data modelling problems. However, the Bayesian posterior distribution and the stochastic complexity typically cannot be obtained analytically. Let $\{X^n, Y^n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ be the complete data. In the VB framework, the Bayesian posterior $p(Y^n, \theta|X^n)$ of the hidden variables and the parameters is approximated by the variational posterior $q(Y^n, \theta|X^n)$, which factorizes as

$$q(Y^n, \theta|X^n) = Q(Y^n|X^n)r(\theta|X^n), \qquad (6)$$

where $Q(Y^n|X^n)$ and $r(\theta|X^n)$ are posteriors on the hidden variables and the parameters respectively. The variational posterior $q(Y^n, \theta|X^n)$ is chosen to minimize the functional $\overline{F}[q]$ defined by

$$\begin{aligned}\overline{F}[q] &= \sum_{Y^n} \int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} d\theta, \qquad (7)\\ &= F(X^n) + K(q(Y^n, \theta|X^n)||p(Y^n, \theta|X^n)), \qquad (8)\end{aligned}$$

where $K(q(Y^n, \theta|X^n)||p(Y^n, \theta|X^n))$ is the Kullback information between the true Bayesian posterior $p(Y^n, \theta|X^n)$ and the variational posterior $q(Y^n, \theta|X^n)$ [1]. This leads to the following theorem. The proof is well known[3][4].

Theorem 1 *If the functional $\overline{F}[q]$ is minimized under the constraint (6) then the variational posteriors, $r(\theta|X^n)$ and $Q(Y^n|X^n)$, satisfy*

$$r(\theta|X^n) = \frac{1}{C_r}\varphi(\theta) \exp \left\langle \log p(X^n, Y^n|\theta) \right\rangle_{Q(Y^n|X^n)}, \qquad (9)$$

$$Q(Y^n|X^n) = \frac{1}{C_Q} \exp \left\langle \log p(X^n, Y^n|\theta) \right\rangle_{r(\theta|X^n)}, \qquad (10)$$

where C_r and C_Q are the normalization constants[2].

Hereafter, we omit the condition X^n of the variational posteriors and abbreviate them to $q(Y^n, \theta)$, $Q(Y^n)$ and $r(\theta)$. We define the stochastic complexity in the VB learning $\overline{F}(X^n)$ by the minimum of the functional $\overline{F}[q]$, that is,

$$\overline{F}(X^n) = \min_{r,Q} \overline{F}[q].$$

Note that eqs.(9) and (10) give only a necessary condition that $r(\theta)$ and $Q(Y^n)$ minimize $\overline{F}[q]$.

3.1. VB algorithm for MEF Model

Let $\overline{y}_i^k = \langle y_i^k \rangle_{Q(Y^n)}$, $n_k = \sum_{i=1}^n \overline{y}_i^k$ and $\nu_k = \frac{1}{n_k}\sum_{i=1}^n \overline{y}_i^k f(x_i)$, where $y_i^k = 1$ iff the ith datum x_i is from the kth component. From (9) and the respective prior (2) and (3), in the case of the MEF model, the variational posterior $r(\theta)$ is obtained as the product of the following two distributions.

$$r(\mathbf{a}) = \frac{\Gamma(n + K\phi_0)}{\prod_{k=1}^K \Gamma(n_k + \phi_0)} \prod_{k=1}^K a_k^{n_k + \phi_0 - 1}, \qquad (11)$$

$$r(\mathbf{b}) = \prod_{k=1}^K \frac{1}{C(\gamma_k, \overline{\mu}_k)} \exp\{\gamma_k(\overline{\mu}_k \cdot b_k - g(b_k))\}, \qquad (12)$$

where $\overline{\mu}_k = \frac{n_k\nu_k + \xi_0\nu_0}{n_k + \xi_0}$ and $\gamma_k = n_k + \xi_0$. Let

$$\overline{a}_k = \langle a_k \rangle_{r(\mathbf{a})} = \frac{n_k + \phi_0}{n + K\phi_0}, \qquad (13)$$

$$\overline{b}_k = \langle b_k \rangle_{r(b_k)} = \frac{1}{\gamma_k} \frac{\partial \log C(\gamma_k, \overline{\mu}_k)}{\partial \overline{\mu}_k}, \qquad (14)$$

and define the variational parameter by $\overline{\theta} = \langle \theta \rangle_{r(\theta)} = \{\overline{a}_k, \overline{b}_k\}_{k=1}^K$.Then, putting (10) into (7), we obtain

$$\overline{F}(X^n) = \min_{\overline{\theta}} \{K(r(\theta)||\varphi(\theta)) - (\log C_Q + S(X^n))\}, \qquad (15)$$

where $S(X^n) = -\sum_{i=1}^n \log p_0(x)$. Hence, the VB algorithm is an update rule for the variational parameter $\overline{\theta}$ to attain the minimum in eq.(15) although it may converges to local minima[4].

Recently, the following theorem was shown ([5]) on the average stochastic complexity defined by

$$\overline{F}(n) = E_{X^n}[\overline{F}(X^n)], \qquad (16)$$

where $E_{X^n}[\cdot]$ denotes the expectation value over all sets of training samples.

[1] $K(q(x)||p(x))$ denotes the Kullback information from a distribution $q(x)$ to a distribution $p(x)$, that is,

$$K(q(x)||p(x)) = \int q(x) \log \frac{q(x)}{p(x)} dx.$$

[2] $\langle \cdot \rangle_{p(x)}$ denotes the expectation over $p(x)$.

Theorem 2 *Assume that the true distribution is a mixture of exponential family model with K_0 components. Then the average stochastic complexity $\overline{F}(n)$ satisfies*

$$\overline{F}(n) \leq \overline{\lambda} \log n + C, \tag{17}$$

for an arbitrary natural number n, where C is a constant independent of n and

$$\overline{\lambda} = \begin{cases} (K - K_0)\phi_0 + \frac{MK_0+K_0-1}{2} & (\phi_0 \leq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}). \end{cases} \tag{18}$$

4. Experiment

In this section, we present the results of experiments where the VB learning is simulated for the mixture model with the 2-dimensional gaussian component $c(x|b) = \frac{1}{2\pi}\exp(-\frac{||x-b||^2}{2})$. This means $M = 2$.

In the first experiment, we trained the gaussian mixture models with $K = 2, 3, 4, 5$ components by the VB algorithm using the data generated by the true distribution with $K_0 = 2$ components. The true distribution was set to

$$p(x|\theta_0) = \frac{1}{2}c(x|(2,2)^T) + \frac{1}{2}c(x|(-2,-2)^T). \tag{19}$$

The hyperparameters were set at $\phi_0 = 1.0$, $\nu_0 = (0,0)^T$ and $\xi_0 = 1.0$. We prepared two sample sets with the sample size $n = 1000$ and $n = 100$. The value of $K(r(\theta)||\varphi(\theta))$ in eq.(15) was calculated when the VB algorithm for each data set converged since it gives the leading term of the stochastic complexity $\overline{F}(X^n)$. The difference of them was divided by $\log 10$ so that the average of it gives the coefficient λ if the average of $K(r(\theta)||\varphi(\theta))$ is asymptotically expanded as

$$E_{X^n}[K(r(\theta)||\varphi(\theta))] \simeq \lambda \log n + O(1). \tag{20}$$

Then from Theorem 2, $\lambda \leq \overline{\lambda}$ should hold unless the VB algorithm converges to local minima.

We averaged the values of λ over 100 draws of the sample sets. The results of the averages of λ are presented in Figure 1 against the number K of components for two different types of the initial values of the variational parameter that are (1): $\overline{a}_k = 1/K, \overline{b}_k = (0,0)^T (k = 1, 2, \cdots, K)$ and (2): $\overline{a}_1 = \overline{a}_2 = 1/2, \overline{a}_k = 0 (k \geq 3), \overline{b}_1 = (2,2)^T, \overline{b}_2 = (-2,-2)^T, \overline{b}_k = (0,0)^T (k \geq 3)$.

We also calculated the training error $T(X^n) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i|\theta_0)}{<p(x_i|\theta)>_{r(\theta)}}$ and the generalization error $G(X^n) = K(p(x|\theta_0)||\langle p(x|\theta)\rangle_{r(\theta)})$ where $\langle p(x|\theta)\rangle_{r(\theta)}$ is the predictive distribution in the VB learning. The generalization error was approximated by $\frac{1}{n'}\sum_{i=1}^{n'} \log \frac{p(x'_i|\theta_0)}{<p(x'_i|\theta)>_{r(\theta)}}$ with test data $\{x'_i\}_{i=1}^{n'=10000}$ generated from the true distribution eq.(19).

Figure 2 shows the averages of the training errors and the generalization errors for the data set with the size $n = 1000$. All these results are multiplied by $n = 1000$ for scaling purposes.

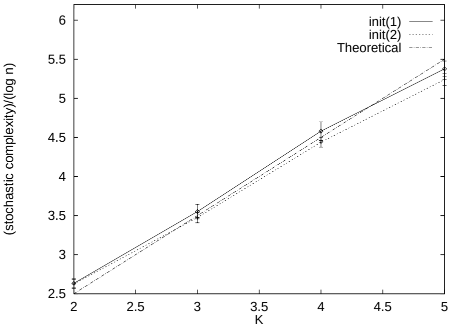

Figure 1: The stochastic complexity against the number K of components for the two types (1) (solid line), (2) (dotted line) of initial values of the variational parameter and the theoretical bound $\overline{\lambda}$ (dashed line).

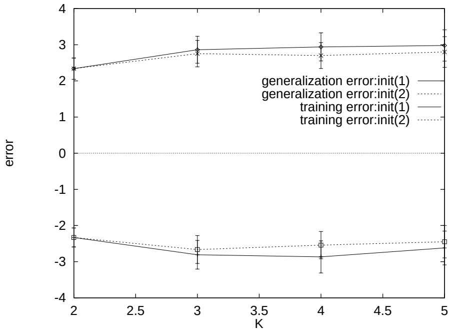

Figure 2: Generalization errors (upper two lines) and training errors (lower two lines) against the number K of components for the two types (1) (solid lines), (2) (dotted lines) of initial values of the variational parameter.

In the second experiment, to investigate the effect of the hyperparameter ϕ_0, we calculated the average stochastic complexities (λ in eq.(20)) of the gaussian mixture model with $K = 4$ components trained by the VB algorithm for various values of the hyperparameter ϕ_0. We used the same training sets generated by the true distribution eq.(19) and calculated the values of λ in the same way as the above. The hyperparameters except for ϕ_0 were set at $\nu_0 = 0$ and $\xi_0 = 1.0$. The averages of λ are presented in Figure 3 for the above two types of the initial values of the variational parameter.

The training and generalization errors are also calculated and averages of them are presented in Figure 4 and Figure 5.

5. Discussion

In this section, we discuss the experimental results.

We point out that Theorem 2 shows how the hyperparameter ϕ_0 influences the process of the VB learning. More specifically, only when $\phi_0 \leq (M+1)/2$, the prior distribution works to eliminate the redun-

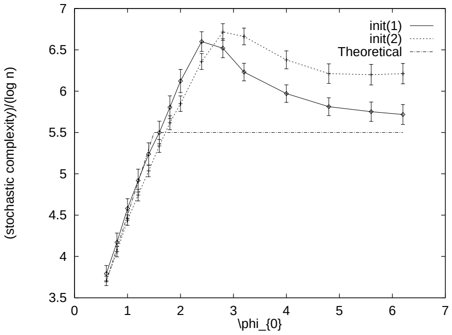

Figure 3: The stochastic complexity against the hyperparameter ϕ_0 for the two types (1) (solid line), (2) (dotted line) of initial values of the variational parameter and the theoretical upper bound $\overline{\lambda}$ (dashed line).

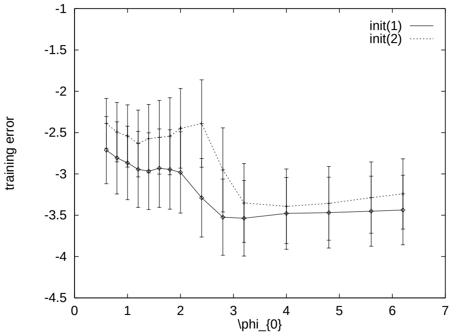

Figure 4: Training error against the hyperparameter ϕ_0 for the two types (1) (solid line), (2) (dotted line) of initial values of the variational parameter.

dant components that the model has and otherwise it works to use all the components. First, we discuss this effect of the hyperparameter ϕ_0 on the actual iterative algorithm of the VB learning. We see in Figure 1 that when $\phi_0 = 1$, although there is slight difference between the two types of the initial values, the experimental results nearly coincide with the theoretical upper bound $\overline{\lambda}$ in eq.(17). This is true for the results in Figure 3 when $\phi_0 \leq \frac{M+1}{2} = \frac{3}{2}$. However, when ϕ_0 is above $\frac{M+1}{2} = \frac{3}{2}$, the results of λ values are larger than the theoretical upper bound $\overline{\lambda}$. This means that for ϕ_0 just above $\frac{M+1}{2}$, the VB algorithm tends to converge to local minima at least for the two types (1) and (2) of the initial values.

Next, we discuss the relationship between the stochastic complexity and the training or generalization errors. Although the theoretical behaviors of the average generalization error and the average training error are still unknown, we observe the following tendencies about the relationship between them. As we can see in Figure 1 and Figure 2, the smaller the stochastic complexity (the value of λ), the smaller the generalization error. However, the generalization error increases little while the stochastic complexity grows proportionally to the number K of the components, that is, the redundancy of the model. As can be seen also in Figure 3 and Figure 5, the smaller the stochastic complexity, the smaller the generalization error. However, this is not true for the training error in Figure 2 and Figure 4. The smaller training error does not mean the better generalization. These results suggest that the stochastic complexity $\overline{F}(X^n)$ is more appropriate than the training error as a criterion to select the model whose generalization is good.

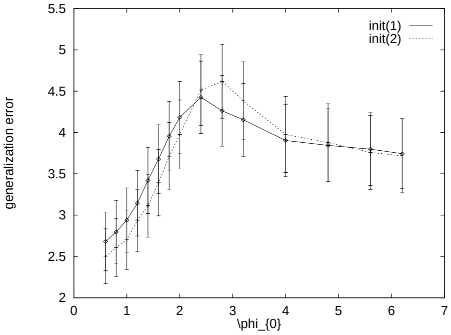

Figure 5: Generalization error against the hyperparameter ϕ_0 for the two types (1) (solid line), (2) (dotted line) of initial values of the variational parameter.

6. Conclusion

In this paper, we presented the experimental results of the VB learning of the gaussian mixture model. Comparing them with the theoretical results, we investigated the properties of the practical VB algorithm.

We conclude with the three observations on the VB algorithm. (i). For $\phi_0 \leq \frac{M+1}{2}$, the VB algorithm often finds the minimum of the stochastic complexity. However, for ϕ_0 just above $\frac{M+1}{2}$, it tends to converge to local minima. (ii). The model with the smaller training error does not have the smaller generalization error. (iii). The model with the smaller stochastic complexity often has the smaller generalization error.

References

[1] H.Attias, "Inferring parameters and structure of latent variable models by variational bayes," *Proc. of UAI*, 1999.

[2] G.McLachlan, D.Peel,"Finite mixture models," Wiley, 2000.

[3] M.Sato, “Online model selection based on the variational bayes,” *Neural Computation*, Vol.13, No.7, pp.1649-1681, 2004.

[4] U.Ueda, Z.Ghahramani, “Bayesian model search for mixture models based on optimizing variational bounds,” *Neural Networks*, 15, pp.1223-1241, 2002.

[5] K.Watanabe, S.Watanabe, "Stochastic complexity for mixture of exponential families in variational bayes" *Proc. of ALT05*, to appear, 2005.