LETTER Communicated by Hagai Attias

Online Model Selection Based on the Variational Bayes

Masa-aki Sato
Information Sciences Division, ATR International, and CREST, Japan Science and Techonology Corporation, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

The Bayesian framework provides a principled way of model selection. This framework estimates a probability distribution over an ensemble of models, and the prediction is done by averaging over the ensemble of models. Accordingly, the uncertainty of the models is taken into account, and complex models with more degrees of freedom are penalized. However, integration over model parameters is often intractable, and some approximation scheme is needed.

Recently, a powerful approximation scheme, called the variational bayes (VB) method, has been proposed. This approach defines the free energy for a trial probability distribution, which approximates a joint posterior probability distribution over model parameters and hidden variables. The exact maximization of the free energy gives the true posterior distribution. The VB method uses factorized trial distributions. The integration over model parameters can be done analytically, and an iterative expectation-maximization-like algorithm, whose convergence is guaranteed, is derived.

In this article, we derive an online version of the VB algorithm and prove its convergence by showing that it is a stochastic approximation for finding the maximum of the free energy. By combining sequential model selection procedures, the online VB method provides a fully online learning method with a model selection mechanism. In preliminary experiments using synthetic data, the online VB method was able to adapt the model structure to dynamic environments.

1 Introduction

The learning of model parameters from observed data can be accomplished by using the maximum likelihood (ML) method for probabilistic models (Bishop, 1995). The expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) provides a general framework for calculating the ML estimator for models with hidden variables. The fundamental problems of the ML method are overfitting and the inability to account for the model complexity, so it is unable to determine the model structure.

The Bayesian framework overcomes these problems in principle (Bishop, 1995; Cooper & Herskovitz, 1992; Gelman, Carlin, Stern, & Rubin, 1995;

Heckerman, Geiger, & Chickering, 1995; Mackay, 1992a, 1992b). The Bayesian method estimates a probability distribution over an ensemble of models, and the prediction is done by averaging over the ensemble of models. Accordingly, the uncertainty of the models is taken into account, and complex models with more degrees of freedom are penalized. The evidence, which is the marginal posterior probability given the data, gives a criterion for the model selection (Mackay, 1992a, 1992b). However, an integration over model parameters is often intractable, and some approximation scheme is needed (Chickering & Heckerman, 1997; Mackay, 1999; Neal, 1996; Richardson & Green, 1997; Roberts, Husmeier, Rezek, & Penny, 1998). Markov chain Monte Carlo (MCMC) methods and the Laplace approximation method have been developed to date. MCMC methods can, in principle, find exact results, but they require a huge amount of time for computation. In addition, it is difficult to determine when these algorithms converge. The Laplace approximation method makes a local gaussian approximation around a maximum a posteriori parameter estimate. This approximation is valid only for a large sample limit. Unfortunately, it is not suited to parameters with constraints such as mixing proportions of mixture models.

Recently, an alternative approach, variational Bayes (VB), has been proposed (Waterhouse, Mackay, & Robinson, 1996; Attias, 1999, 2000; Ghahramani & Beal, 2000; Bishop, 1999). This approach defines the free energy for a trial probability distribution, which approximates a joint posterior probability distribution over model parameters and hidden variables. The maximum of the free energy gives the log evidence for an observed data set. Therefore, the exact maximization of the free energy gives the true posterior distribution over the parameters and the hidden variables. The VB method uses trial distributions in a restricted space where the parameters are assumed to be conditionally independent of the hidden variables. Once this approximation is made, the remaining calculations are all done exactly. As a result, an iterative EM-like algorithm, whose convergence is guaranteed, is derived. The predictive distribution is also calculated analytically.

The VB method has several attractive features. The method requires only a modest amount of computational time, comparable to that of the EM algorithm. The Bayesian information criterion (BIC) (Schwartz, 1978) and the minimum description length (MDL) criterion (Rissanen, 1987) for the model selection are obtained from the VB method in a large sample limit (Attias, 1999). In this limit, the VB algorithm becomes equivalent to the ordinary EM algorithm. The VB method can be easily extended to the hierarchical Bayes method. Sequential model selection procedures (Ghahramani & Beal, 2000; Ueda, 1999) have also been proposed by combining the VB method and the split-and-merge algorithm (Ueda, Nakano, Ghahramani, & Hinton, 1999).

In this article, we derive an on-line version of the VB algorithm and prove its convergence by showing that it is a stochastic approximation for finding the maximum of the free energy. We also prove that the VB algorithm is a gradient method with the inverse of the Fisher information matrix for

the posterior parameter distribution as a coefficient matrix. Namely, the VB method is a type of natural gradient method (Amari, 1998). By combining sequential model selection procedures, the online VB algorithm provides a fully online learning method with a model selection mechanism. It can be applied to real-time applications. In preliminary experiments using synthetic data, the online VB method was able to adapt the model structure to dynamic environments. We also found that the introduction of a discount factor was crucial for a fast convergence of the online VB method.

We study the VB method for general exponential family models with hidden variables (EFH models) (Amari, 1985), although the VB method can be applied to more general graphical models (Attias, 1999). The use of the EFH models makes the calculations transparent. Moreover, the EFH models include a lot of interesting models such as normalized gaussian networks (Sato & Ishii, 2000), hidden Markov models (Rabiner, 1989), mixture of gaussian models (Roberts, Husmeier, Rezek, & Penny, 1998), mixture of factor analyzers (Ghahramani & Beal, 2000), mixture of probabilistic principal component analyzers (Tipping & Bishop, 1999), and others (Roweis & Ghahramani, 1999; Titterington, Smith, & Makov, 1985).

2 Variational Bayes Method

In this section, we review the VB method (Waterhouse et al., 1996; Attias, 1999, 2000; Ghahramani & Beal, 2000; Bishop, 1999) for EFH models (Amari, 1985).

2.1 Bayesian Method. In the maximum likelihood (ML) approach, the objective is to find the ML estimator that maximizes the likelihood for a given data set. The ML approach, however, suffers from overfitting and the inability to determine the best model structure.

In the Bayesian method (Bishop, 1995; Mackay, 1992a, 1992b), a set of models, $\{\mathcal{M}_m | m = 1, \ldots, \mathcal{N}\}$, where $\mathcal{M}_m$ denotes a model with a fixed structure, is considered. A probability distribution for a model $\mathcal{M}_m$ with a model parameter $\boldsymbol{\theta}_m$ is denoted by $P(\mathbf{x}|\boldsymbol{\theta}_m, \mathcal{M}_m)$, where $\mathbf{x}$ represents an observed stochastic variable. For a set of observed data, $\mathbf{X}\{T\} = \{\mathbf{x}(t)|t = 1, \ldots, T\}$, and a prior probability distribution $P(\boldsymbol{\theta}_m|\mathcal{M}_m)$, the Bayesian method calculates the posterior probability over the parameter,

$$P(\boldsymbol{\theta}_m|\mathbf{X}\{T\}, \mathcal{M}_m) = \frac{P(\mathbf{X}\{T\}|\boldsymbol{\theta}_m, \mathcal{M}_m)P(\boldsymbol{\theta}_m|\mathcal{M}_m)}{P(\mathbf{X}\{T\}|\mathcal{M}_m)}.$$

Here, the data likelihood is defined by

$$P(\mathbf{X}\{T\}|\boldsymbol{\theta}_m, \mathcal{M}_m) = \prod_{t=1}^{T} P(\mathbf{x}(t)|\boldsymbol{\theta}_m, \mathcal{M}_m),$$

and the evidence for a model $\mathcal{M}_m$ is defined by

$$P(\mathbf{X}\{T\}|\mathcal{M}_m) = \int d\mu(\boldsymbol{\theta}_m) P(\mathbf{X}\{T\}|\boldsymbol{\theta}_m, \mathcal{M}_m) P(\boldsymbol{\theta}_m|\mathcal{M}_m).$$

The posterior probability over model structure is given by

$$P(\mathcal{M}_m|\mathbf{X}\{T\}) = \frac{P(\mathbf{X}\{T\}|\mathcal{M}_m)P(\mathcal{M}_m)}{\sum_{m'} P(\mathbf{X}\{T\}|\mathcal{M}_{m'})P(\mathcal{M}_{m'})}.$$

If we use a noninformative prior for the model structure, that is, $P(\mathcal{M}_m) = 1/\mathcal{N}$, the posterior for a given model $P(\mathcal{M}_m|\mathbf{X}\{T\})$ is proportional to the evidence of the model $P(\mathbf{X}\{T\}|\mathcal{M}_m)$. Therefore, the model selection can be done according to the evidence. In the following, we will calculate the evidence for a given model by using the VB method, and the explicit model structure dependence is omitted.

2.2 Exponential Family Model with Hidden Variables. An EFH model for an N-dimensional vector variable $\mathbf{x} = (x_1, \ldots, x_N)^T$ is defined by a probability distribution,

$$\begin{aligned} P(\mathbf{x}|\boldsymbol{\theta}) &= \int d\mu(\mathbf{z}) P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}), \\ P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) &= \exp\left[\mathbf{r}(\mathbf{x}, \mathbf{z}) \cdot \boldsymbol{\theta} + r_0(\mathbf{x}, \mathbf{z}) - \psi(\boldsymbol{\theta})\right], \end{aligned} \tag{2.1}$$

where $\mathbf{z} = (z_1, \ldots, z_M)^T$ denotes an M-dimensional vector hidden variable and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ denotes a set of model parameters called the natural parameter.[1] A set of sufficient statistics is denoted by $\mathbf{r}(\mathbf{x}, \mathbf{z}) = (r_1(\mathbf{x}, \mathbf{z}), \ldots, r_K(\mathbf{x}, \mathbf{z}))^T$. An inner product of two vectors $\mathbf{r}$ and $\boldsymbol{\theta}$ is denoted by $\mathbf{r} \cdot \boldsymbol{\theta} = \sum_{k=1}^K r_k \theta_k$. Measures on the observed and the hidden variable spaces are denoted by $d\mu(\mathbf{x})$ and $d\mu(\mathbf{z})$, respectively. The normalization factor $\psi(\boldsymbol{\theta})$ is determined by

$$\exp\left[\psi(\boldsymbol{\theta})\right] = \int d\mu(\mathbf{x}) d\mu(\mathbf{z}) \exp\left[\mathbf{r}(\mathbf{x}, \mathbf{z}) \cdot \boldsymbol{\theta} + r_0(\mathbf{x}, \mathbf{z})\right], \tag{2.2}$$

which is derived from the probability condition $\int d\mu(\mathbf{x}) P(\mathbf{x}|\boldsymbol{\theta}) = 1$. $P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ represents the probability distribution for a complete event $(\mathbf{x}, \mathbf{z})$. If the hidden variable is a discrete variable, the integration $\int d\mu(\mathbf{z})$ is replaced by the summation $\sum_{\mathbf{z}}$.

The expectation parameter for the EFH model, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)^T$, is defined by

$$\begin{aligned} \boldsymbol{\phi} &= \partial\psi(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = (\partial\psi/\partial\theta_1, \ldots, \partial\psi/\partial\theta_K)^T \\ &= E\left[\mathbf{r}(\mathbf{x}, \mathbf{z})|\boldsymbol{\theta}\right] = \int d\mu(\mathbf{x}) d\mu(\mathbf{z}) \mathbf{r}(\mathbf{x}, \mathbf{z}) P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}). \end{aligned} \tag{2.3}$$

[1] In general, the natural parameter is a function of another model parameter φ: $\theta = \theta(\varphi)$. The following discussion can also be applied in this case.

2.3 Evidence and Free Energy. The evidence for a data set $\mathbf{X}\{T\}$ is defined by

$$P(\mathbf{X}\{T\}) = \int d\mu(\boldsymbol{\theta}) P(\mathbf{X}\{T\}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta}), \tag{2.4}$$

where $d\mu(\boldsymbol{\theta})$ denotes a measure on the model parameter space and $P_0(\boldsymbol{\theta})$ denotes a prior distribution for the model parameters. The integration over model parameters in equation 2.4 penalizes complex models with more degrees of freedom (Bishop, 1995). This integration, however, is often difficult to perform. In order to evaluate this integration, the VB method introduces a trial probability distribution $Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})$, which approximates the posterior probability distribution over the model parameter $\boldsymbol{\theta}$ and the hidden variables $\mathbf{Z}\{T\} = \{\mathbf{z}(t)|t = 1, \ldots, T\}$:

$$P(\boldsymbol{\theta}, \mathbf{Z}\{T\}|\mathbf{X}\{T\}) = \frac{P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta})}{P(\mathbf{X}\{T\})}, \tag{2.5}$$

where the probability distribution for a complete data set $(\mathbf{X}\{T\}, \mathbf{Z}\{T\})$ is given by

$$P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) = \prod_{t=1}^{T} P(\mathbf{x}(t), \mathbf{z}(t)|\boldsymbol{\theta}). \tag{2.6}$$

The Kullback-Leibler (KL) divergence between the trial distribution $Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})$ and the true posterior distribution $P(\boldsymbol{\theta}, \mathbf{Z}\{T\}|\mathbf{X}\{T\})$ is given by

$$\begin{aligned} KL(Q||P) &= \int d\mu(\boldsymbol{\theta}) d\mu(\mathbf{Z}\{T\})\, Q(\boldsymbol{\theta}, \mathbf{Z}\{T\}) \\ &\quad \times \log\left(\frac{Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})}{P(\boldsymbol{\theta}, \mathbf{Z}\{T\}|\mathbf{X}\{T\})}\right) \\ &= \log P(\mathbf{X}\{T\}) - F(\mathbf{X}\{T\}, Q), \end{aligned} \tag{2.7}$$

where the free energy $F(\mathbf{X}\{T\}, Q)$ is defined by

$$\begin{aligned} F(\mathbf{X}\{T\}, Q) &= \int d\mu(\boldsymbol{\theta}) d\mu(\mathbf{Z}\{T\})\, Q(\boldsymbol{\theta}, \mathbf{Z}\{T\}) \\ &\quad \times \log\left(\frac{P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta})}{Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})}]\right). \end{aligned} \tag{2.8}$$

Therefore, the true posterior distribution is obtained by maximizing the free energy with respect to the trial distribution $Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})$. The maximum of the free energy is equal to the log evidence:

$$\log P(\mathbf{X}\{T\}) = \max_{Q} F(\mathbf{X}\{T\}, Q) \geq F(\mathbf{X}\{T\}, Q), \tag{2.9}$$

Equation 2.9 implies that the lower bound for the log evidence can be evaluated by using some trial posterior distributions $Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})$.

2.4 Variational Bayes Algorithm. In the VB method, trial posterior distributions are assumed to be factorized as

$$Q(\boldsymbol{\theta}, \mathbf{Z}\{T\}) = Q_\theta(\boldsymbol{\theta}) Q_z(\mathbf{Z}\{T\}). \tag{2.10}$$

The hidden variables are assumed to be conditionally independent with the model parameters given the data. We also assume that the prior distribution $P_0(\boldsymbol{\theta})$ is given by the conjugate prior distribution[2] for the EFH model, equation 2.1,

$$P_0(\boldsymbol{\theta}) = \exp\left[\gamma_0(\boldsymbol{\alpha}_0 \cdot \boldsymbol{\theta} - \psi(\boldsymbol{\theta})) - \Phi(\boldsymbol{\alpha}_0, \gamma_0)\right], \tag{2.11}$$

where $(\boldsymbol{\alpha}_0, \gamma_0)$ are prior hyperparameters. The normalization factor $\Phi(\boldsymbol{\alpha}_0, \gamma_0)$ is determined by

$$\exp\left[\Phi(\boldsymbol{\alpha}_0, \gamma_0)\right] = \int d\mu(\boldsymbol{\theta}) \exp\left[\gamma_0(\boldsymbol{\alpha}_0 \cdot \boldsymbol{\theta} - \psi(\boldsymbol{\theta}))\right]. \tag{2.12}$$

Equations 2.10 and 2.11 are the only assumptions in this method. Under these assumptions, we try to maximize the free energy $F(\mathbf{X}\{T\}, Q)$. The maximum free energy with respect to factorized Q, 2.10, gives an estimate (lower bound) for the log evidence $\log(P(\mathbf{X}\{T\}))$.

The free energy can be maximized by alternately maximizing the free energy with respect to Q_θ and Q_z. This process closely resembles the free energy formulation for the EM algorithm (Neal & Hinton, 1998) for finding the ML estimator. In the VB expectation step (E-step), the free energy is maximized with respect to $Q_z(\mathbf{Z}\{T\})$ under the condition $\int d\mu(\mathbf{Z}\{T\}) Q_z(\mathbf{Z}\{T\}) = 1$, while $Q_\theta(\boldsymbol{\theta})$ is fixed. The maximum solution is given by the posterior distribution for the hidden variables with the ensemble average of model parameters (see appendix A):

$$Q_z(\mathbf{Z}\{T\}) = \prod_{t=1}^{T} Q_z(\mathbf{z}(t)), \tag{2.13}$$

$$Q_z(\mathbf{z}(t)) = P(\mathbf{z}(t)|\mathbf{x}(t), \bar{\boldsymbol{\theta}}) = P(\mathbf{x}(t), \mathbf{z}(t)|\bar{\boldsymbol{\theta}}) / P(\mathbf{x}(t)|\bar{\boldsymbol{\theta}}), \tag{2.14}$$

$$\bar{\boldsymbol{\theta}} = \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \boldsymbol{\theta}. \tag{2.15}$$

In the VB maximization step (M-step), the free energy is maximized with respect to $Q_\theta(\boldsymbol{\theta})$ under the condition $\int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) = 1$, while $Q_z(\mathbf{Z}\{T\})$ obtained in the VB E-step is fixed. The maximum solution is given by the

[2] It is also possible to use noninformative priors (Attias, 1999).

conjugate distribution for the EFH model with posterior hyperparameters $(\boldsymbol{\alpha}, \gamma)$ (see appendix A):

$$Q_\theta(\boldsymbol{\theta}) = P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma) = \exp\left[\gamma\left(\boldsymbol{\alpha}\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta})\right) - \Phi(\boldsymbol{\alpha}, \gamma)\right], \tag{2.16}$$

$$\gamma = T + \gamma_0, \tag{2.17}$$

$$\boldsymbol{\alpha} = \frac{1}{\gamma}\left[T\langle \mathbf{r}(\mathbf{x}, \mathbf{z})\rangle_{\bar{\boldsymbol{\theta}}} + \boldsymbol{\alpha}_0\cdot\gamma_0\right], \tag{2.18}$$

$$\langle \mathbf{r}(\mathbf{x}, \mathbf{z})\rangle_{\bar{\boldsymbol{\theta}}} = \frac{1}{T}\sum_{t=1}^{T}\int d\mu(\mathbf{z}(t))P(\mathbf{z}(t)|\mathbf{x}(t), \bar{\boldsymbol{\theta}})\mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)). \tag{2.19}$$

The effective amount of data $\gamma = (T + \gamma_0)$ represents the reliability (or uncertainty) of the estimation. As the amount of data T increases, the reliability of the estimation increases. The prior hyperparameter γ_0 represents the reliability of the prior belief on the prior hyperparameter $\boldsymbol{\alpha}_0$. The posterior hyperparameter $\boldsymbol{\alpha}$ is determined by the expectation value of the sufficient statistics. The prior hyperparameter $\boldsymbol{\alpha}_0$ gives the initial value for $\boldsymbol{\alpha}$.

Since the posterior parameter distribution $Q_\theta(\boldsymbol{\theta})$ is given by the conjugate distribution $P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma)$, which is also an exponential family model, the integration over the parameter $\boldsymbol{\theta}$ in equation 2.15 can be explicitly calculated as

$$\bar{\boldsymbol{\theta}} = \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}, \tag{2.20}$$

$$\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}} = \int d\mu(\boldsymbol{\theta})P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma)\boldsymbol{\theta} = \frac{1}{\gamma}\frac{\partial\Phi}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \gamma). \tag{2.21}$$

The natural parameter of the conjugate distribution is given by $(\gamma\,\boldsymbol{\alpha}, \gamma)$. The corresponding expectation parameters are given by the ensemble averages of the model parameters: $\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}$ defined in equation 2.21 and

$$\begin{aligned}\langle\psi(\boldsymbol{\theta})\rangle_{\boldsymbol{\alpha}} &= \int d\mu(\boldsymbol{\theta})P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma)\psi(\boldsymbol{\theta}) \\ &= \frac{1}{\gamma}\frac{\partial\Phi}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \gamma)\cdot\boldsymbol{\alpha} - \frac{\partial\Phi}{\partial\gamma}(\boldsymbol{\alpha}, \gamma).\end{aligned} \tag{2.22}$$

2.5 Parameterized Free Energy Function. Since the optimal solution simultaneously satisfies equations 2.14 and 2.16, the trial posterior distributions, $Q_\theta(\boldsymbol{\theta})$ and $Q_z(\mathbf{Z}\{T\})$, can be parameterized as

$$Q_\theta(\boldsymbol{\theta}) = P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma) = \exp\left[\gamma\left(\boldsymbol{\alpha}\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta})\right) - \Phi(\boldsymbol{\alpha}, \gamma)\right], \tag{2.23}$$

$$Q_z(\mathbf{Z}\{T\}) = \prod_{t=1}^{T} Q_z(\mathbf{z}(t)), \tag{2.24}$$

$$Q_z(\mathbf{z}(t)) = P(\mathbf{z}(t)|\mathbf{x}(t), \bar{\boldsymbol{\theta}}), \tag{2.25}$$

where γ , $\boldsymbol{\alpha}$, and $\bar{\boldsymbol{\theta}}$ are arbitrary variational parameters. By substituting this parameterized form into the definition of the free energy, equation 2.8, one can get the parameterized free energy function:

$$\begin{aligned} F(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \gamma) = & \sum_{t=1}^{T} \log P(\mathbf{x}(t)|\bar{\boldsymbol{\theta}}) + (\gamma_0 \boldsymbol{\alpha}_0 - \gamma\, \boldsymbol{\alpha}) \cdot \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} \\ & + T \langle \mathbf{r}(\mathbf{x}, \mathbf{z}) \rangle_{\bar{\boldsymbol{\theta}}} \cdot (\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} - \bar{\boldsymbol{\theta}}) - (T + \gamma_0 - \gamma) \langle \psi(\boldsymbol{\theta}) \rangle_{\boldsymbol{\alpha}} \\ & + T\psi(\bar{\boldsymbol{\theta}}) + \Phi(\boldsymbol{\alpha}, \gamma) - \Phi(\boldsymbol{\alpha}_0, \gamma_0), \end{aligned} \tag{2.26}$$

where the ensemble averages of the parameters $\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}}$ and $\langle \psi(\boldsymbol{\theta}) \rangle_{\boldsymbol{\alpha}}$ are given by equations 2.21 and 2.22, respectively.

The VB E-step equation, 2.20, can be derived from the free energy maximization condition with respect to $\bar{\boldsymbol{\theta}}$:

$$\partial F(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \gamma) / \partial \bar{\boldsymbol{\theta}} = 0. \tag{2.27}$$

The derivative of the free energy with respect to $\bar{\boldsymbol{\theta}}$ is given by (see appendix B)

$$\begin{aligned} \partial F / \partial \bar{\boldsymbol{\theta}} &= U(\bar{\boldsymbol{\theta}}) \cdot (\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} - \bar{\boldsymbol{\theta}}), \\ U(\bar{\boldsymbol{\theta}}) &= \sum_{t=1}^{T} \Big\langle \Big(\mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)) - \langle \mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)) \rangle_{\bar{\boldsymbol{\theta}}} \Big) \\ &\qquad \times \Big(\mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)) - \langle \mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)) \rangle_{\bar{\boldsymbol{\theta}}} \Big)^{T} \Big\rangle_{\bar{\boldsymbol{\theta}}}, \\ \langle \mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)) \rangle_{\bar{\boldsymbol{\theta}}} &= \int d\mu(\mathbf{z}(t)) P(\mathbf{z}(t)|\mathbf{x}(t), \bar{\boldsymbol{\theta}}) \mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)). \end{aligned} \tag{2.28}$$

Since the coefficient matrix U is positive definite, the maximization condition, equation 2.27, leads to the VB E-step equation, 2.20. The Hessian of the free energy with respect to $\bar{\boldsymbol{\theta}}$ at the VB E-step solution is given by $(-U)$. This shows that the VB E-step solution is actually a maximum of the free energy with respect to $\bar{\boldsymbol{\theta}}$.

The VB M-step equations, 2.17 and 2.18, can be derived from the free energy maximization condition with respect to $(\boldsymbol{\alpha}, \gamma)$:

$$\begin{aligned} \partial F(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \gamma) / \partial \gamma &= 0, \\ \partial F(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \gamma) / \partial \boldsymbol{\alpha} &= 0. \end{aligned} \tag{2.29}$$

The derivative of the free energy with respect to $(\boldsymbol{\alpha}, \gamma)$ is given by (see

appendix B)

$$\begin{pmatrix} \frac{1}{\gamma}(\partial F/\partial \boldsymbol{\alpha}) \\ (\partial F/\partial \gamma) \end{pmatrix} = \begin{pmatrix} V_{\boldsymbol{\alpha},\boldsymbol{\alpha}} & , & V_{\boldsymbol{\alpha},\gamma} \\ V^T_{\boldsymbol{\alpha},\gamma} & , & V_{\gamma,\gamma} \end{pmatrix} \times \begin{pmatrix} T\langle \mathbf{r}(\mathbf{x},\mathbf{z})\rangle_{\bar{\boldsymbol{\theta}}} + \gamma_0 \boldsymbol{\alpha}_0 - (T+\gamma_0)\boldsymbol{\alpha} \\ T + \gamma_0 - \gamma \end{pmatrix}, \tag{2.30}$$

where the Fisher information matrix V for the posterior parameter distribution $P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha},\gamma)$ is given by

$$\begin{aligned} V_{\boldsymbol{\alpha},\boldsymbol{\alpha}} &= \frac{1}{\gamma^2}\left\langle \left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}^T}\right)\right\rangle_{\boldsymbol{\alpha}} \\ &= \left\langle (\boldsymbol{\theta} - \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}})(\boldsymbol{\theta} - \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}})^T \right\rangle_{\boldsymbol{\alpha}}, \\ V_{\boldsymbol{\alpha},\gamma} &= \frac{1}{\gamma}\left\langle \left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial \log P_\alpha}{\partial \gamma}\right)\right\rangle_{\boldsymbol{\alpha}} \\ &= \left\langle (\boldsymbol{\theta} - \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}})(g(\boldsymbol{\theta}) - \langle g(\boldsymbol{\theta})\rangle_{\boldsymbol{\alpha}})\right\rangle_{\boldsymbol{\alpha}}, \\ V_{\gamma,\gamma} &= \left\langle \left(\frac{\partial \log P_\alpha}{\partial \gamma}\right)\left(\frac{\partial \log P_\alpha}{\partial \gamma}\right)\right\rangle_{\boldsymbol{\alpha}} \\ &= \left\langle (g(\boldsymbol{\theta}) - \langle g(\boldsymbol{\theta})\rangle_{\boldsymbol{\alpha}})(g(\boldsymbol{\theta}) - \langle g(\boldsymbol{\theta})\rangle_{\boldsymbol{\alpha}})\right\rangle_{\boldsymbol{\alpha}}, \\ g(\boldsymbol{\theta}) &= \boldsymbol{\alpha}\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta}). \end{aligned} \tag{2.31}$$

Since the Fisher information matrix V is positive definite, the free energy maximization condition, 2.29, leads to the VB M-step equations, 2.17 and 2.18. From equation 2.30, it is shown that the VB M-step solution is a maximum of the free energy with respect to $(\boldsymbol{\alpha},\gamma)$, as in the VB-E step.

The VB algorithm is summarized as follows. First, γ is set to $(T+\gamma_0)$. In the VB E-step, the ensemble average of the parameter $\bar{\boldsymbol{\theta}}$ is calculated by using equation 2.20. Subsequently, the expectation value of the sufficient statistics $\langle \mathbf{r}(\mathbf{x},\mathbf{z})\rangle_{\bar{\boldsymbol{\theta}}}$, 2.19, is calculated by using the posterior distribution for the hidden variable $P(\mathbf{z}(t)|\mathbf{x}(t),\bar{\boldsymbol{\theta}})$, equation 2.14. In the VB M-step, the posterior hyperparameter $\boldsymbol{\alpha}$ is updated by using equation 2.18. Repeating this process, the free energy function, equation 2.26, increases monotonically. This process continues until the free energy function converges.

Using equations 2.28 and 2.30, VB equations 2.20 and 2.18 can be expressed as the gradient method:

$$\begin{aligned} \Delta\bar{\boldsymbol{\theta}} &= \bar{\boldsymbol{\theta}}_{new} - \bar{\boldsymbol{\theta}} = \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}} - \bar{\boldsymbol{\theta}} \\ &= U^{-1}(\bar{\boldsymbol{\theta}})\frac{\partial F}{\partial \bar{\boldsymbol{\theta}}}(\mathbf{X}\{T\},\bar{\boldsymbol{\theta}},\boldsymbol{\alpha},\gamma), \end{aligned} \tag{2.32}$$

$$\Delta\boldsymbol{\alpha} = \boldsymbol{\alpha}_{new} - \boldsymbol{\alpha} = \frac{1}{\gamma}(T\langle \mathbf{r}(\mathbf{x},\mathbf{z})\rangle_{\bar{\boldsymbol{\theta}}} + \gamma_0\boldsymbol{\alpha}_0 - \gamma\,\boldsymbol{\alpha})$$

$$= \frac{1}{\gamma^2} V^{-1}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma)\frac{\partial F}{\partial \boldsymbol{\alpha}}(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \gamma), \tag{2.33}$$

together with $\gamma = T + \gamma_0$. Substituting the VB E-step equation, 2.20, into the free energy equation, 2.26, the VB algorithm is further rewritten as

$$\Delta\boldsymbol{\alpha} = \frac{1}{\gamma^2} V^{-1}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma)\frac{\partial F}{\partial \boldsymbol{\alpha}}(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}} = \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \gamma). \tag{2.34}$$

This shows that the VB algorithm is the gradient method with the inverse of the Fisher information matrix as a coefficient matrix. Namely, it is a type of natural gradient method (Amari, 1998), which gives the optimal asymptotic convergence. This fact is proved for the first time in this article. The natural gradient gives the steepest direction in the hyperparameter space, which has the Riemannian structure according to the information geometry. It should be noted that the learning rate in the VB algorithm, 2.34, is automatically determined by the inverse of the Fisher information matrix.

When the VB algorithm converges, the free energy equation, 2.26, can be written in a simple form:

$$F(\mathbf{X}\{T\}) = \log(P(\mathbf{X}\{T\}|\bar{\boldsymbol{\theta}})P_0(\bar{\boldsymbol{\theta}})) - \int d\mu(\boldsymbol{\theta})\; Q_\theta(\boldsymbol{\theta})\log(Q_\theta(\boldsymbol{\theta}))$$

$$+ \gamma\,[\Psi(\bar{\boldsymbol{\theta}}) - \langle\Psi(\boldsymbol{\theta})\rangle_{\boldsymbol{\alpha}}]. \tag{2.35}$$

The first term on the right-hand side is the log likelihood together with the prior. It is estimated at the ensemble average of the parameters. The second term is the entropy of the posterior parameter distribution. It penalizes the complex models and overfitting. The third term represents the deviation from the ensemble average of the parameters and becomes negligible in the large sample limit.

2.6 Predictive Distribution. If the posterior parameter distribution is obtained by using the VB algorithm, one can calculate the predictive distribution for the observed variable $\mathbf{x}$. The predictive distribution for $\mathbf{x}$ is given by

$$P(\mathbf{x}|\mathbf{X}\{T\}) = \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) P(\mathbf{x}|\boldsymbol{\theta})$$

$$= \int d\mu(\boldsymbol{\theta}) \int d\mu(\mathbf{z}) \exp\left[(\mathbf{r}(\mathbf{x},\mathbf{z}) + \gamma\,\boldsymbol{\alpha})\cdot\boldsymbol{\theta} + r_0(\mathbf{x},\mathbf{z})\right.$$

$$\left. - (1+\gamma)\psi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\alpha},\gamma)\right]. \tag{2.36}$$

By interchanging the integration with respect to $\boldsymbol{\theta}$ and $\mathbf{z}$, one can get

$$P(\mathbf{x}|\mathbf{X}\{T\}) = \int d\mu(\mathbf{z}) \times \exp\left[r_0(\mathbf{x},\mathbf{z}) + \Phi(\hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}),\gamma+1) - \Phi(\boldsymbol{\alpha},\gamma)\right], \quad (2.37)$$

$$\hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}) = (\gamma\,\boldsymbol{\alpha} + \mathbf{r}(\mathbf{x},\mathbf{z}))/(1+\gamma).$$

For a finite T, this predictive distribution has a different functional form from the model distribution $P(\mathbf{x}|\boldsymbol{\theta})$, equation 2.1.

2.7 Large Sample Limit. When the amount of observed data becomes large ($T \gg 1 : \gamma \gg 1$), the solution of the VB algorithm becomes the ML estimator (Attias, 1999). In this limit, the integration over the parameters with respect to the posterior parameter distribution can be approximated by using a stationary point approximation:

$$\exp[\Phi(\boldsymbol{\alpha},\gamma)] = \int d\mu(\boldsymbol{\theta}) \exp[\gamma(\alpha\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta}))] \sim \exp\left[\gamma(\alpha\cdot\hat{\boldsymbol{\theta}} - \psi(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\log\left|\gamma\,\frac{\partial^2\psi}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\right| + O(1/\gamma)\right], \quad (2.38)$$

where $\hat{\boldsymbol{\theta}}$ is the maximum of the exponent $\left(\boldsymbol{\alpha}\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta})\right)$, that is,

$$\frac{\partial\psi}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\alpha}. \quad (2.39)$$

Therefore, Φ can be approximated as

$$\Phi(\boldsymbol{\alpha},\gamma) \sim \gamma(\alpha\cdot\hat{\boldsymbol{\theta}} - \psi(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\log\left|\gamma\,\frac{\partial^2\psi}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\right| + O(1/\gamma). \quad (2.40)$$

Consequently, the ensemble average of the parameter $\bar{\boldsymbol{\theta}}$ can be approximated as

$$\bar{\boldsymbol{\theta}} = \frac{1}{\gamma}\frac{\partial\Phi}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma) \sim \frac{1}{\gamma}\frac{\partial}{\partial\boldsymbol{\alpha}}(\gamma(\boldsymbol{\alpha}\cdot\hat{\boldsymbol{\theta}} - \psi(\hat{\boldsymbol{\theta}}))) = \hat{\boldsymbol{\theta}}. \quad (2.41)$$

The relations 2.39 and 2.41 imply that the posterior hyperparameter $\boldsymbol{\alpha}$ is equal to the expectation parameter of the EFH model, $\boldsymbol{\phi}$ (see equation 2.3) in this limit. Furthermore, equations 2.18, 2.39, and 2.41 are equivalent to the

ordinary EM algorithm for the EFH model. In a large sample limit, the data term is dominant over the model complexity term. Consequently, the free energy maximization becomes equivalent to the likelihood maximization. Using equations 2.40 and 2.41, the free energy becomes

$$F \sim \sum_{t=1}^{T} \log P(\mathbf{x}|\bar{\boldsymbol{\theta}}) - \frac{K}{2}\log\gamma \tag{2.42}$$

$$- \frac{1}{2}\log\left|\frac{\partial^2\psi}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}})\right| + \gamma_0(\alpha_0\cdot\bar{\boldsymbol{\theta}} - \psi(\bar{\boldsymbol{\theta}})) - \boldsymbol{\Phi}(\boldsymbol{\alpha}_0, \gamma_0) + O(1/\gamma).$$

This expression coincides with the BIC/MDL criteria (Rissanen, 1987; Schwartz, 1978).

The predictive distribution $P(\mathbf{x}|\mathbf{X}\{T\})$ in this limit coincides with the model distribution using the ML estimator $P(\mathbf{x}|\bar{\boldsymbol{\theta}})$. This can be shown by using the following relations:

$$\hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}) \sim \boldsymbol{\alpha} + \frac{1}{\gamma}(\mathbf{r}(\mathbf{x},\mathbf{z}) - \boldsymbol{\alpha}) + O(1/\gamma^2),$$

$$\boldsymbol{\Phi}(\hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}), r+1) \sim \boldsymbol{\Phi}(\boldsymbol{\alpha},\gamma) + \frac{1}{\gamma}(\mathbf{r}(\mathbf{x},\mathbf{z}) - \boldsymbol{\alpha})\frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma) + \frac{\partial\boldsymbol{\Phi}}{\partial\gamma}(\boldsymbol{\alpha},\gamma)$$

$$\sim \boldsymbol{\Phi}(\boldsymbol{\alpha},\gamma) + \mathbf{r}(\mathbf{x},\mathbf{z})\cdot\bar{\boldsymbol{\theta}} - \psi(\bar{\boldsymbol{\theta}}).$$

3 Online Variational Bayes Method

3.1 Expectation Value of the Free Energy. In this section, we derive an online version of the VB algorithm. The amount of data increases over time in the online learning. Therefore, it is desirable to calculate the free energy corresponding to a fixed amount of data. For this purpose, let us define an expectation value of the log evidence for a finite amount of data:

$$E\left[\log P(\mathbf{X}\{T\})\right]_\rho = \int d\mu(\mathbf{X}\{T\})\rho(\mathbf{X}\{T\})$$

$$\times \log\left(\int d\mu(\boldsymbol{\theta})P(\mathbf{X}\{T\}|\boldsymbol{\theta})P_0(\boldsymbol{\theta})\right), \tag{3.1}$$

where ρ represents an unknown probability distribution for observed data. The corresponding VB free energy is given by

$$E\left[F(\mathbf{X}\{T\}, Q_\theta, Q_z)\right]_\rho = T\int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})$$

$$\times E\left[\int d\mu(\mathbf{z})Q_z(\mathbf{z})\log\left(P(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})/Q_z(\mathbf{z})\right)\right]_\rho$$

$$+ \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \log \left(P_0(\boldsymbol{\theta}) / Q_\theta(\boldsymbol{\theta}) \right). \tag{3.2}$$

The ratio (γ_0/T) determines the relative reliability between the observed data and the prior belief for the parameter distribution. The expected free energy, equation 3.2, can be estimated by

$$\begin{aligned} F(\mathbf{X}\{\tau\}, Q_z\{\tau\}, Q_\theta, T) = \left(\frac{T}{\tau}\right) \sum_{t=1}^{\tau} \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \int d\mu(\mathbf{z}(t)) Q_z(\mathbf{z}(t)) \\ \times \log \left(P(\mathbf{x}(t), \mathbf{z}(t)|\boldsymbol{\theta}) / Q_z(\mathbf{z}(t)) \right) \\ + \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \log \left(P_0(\boldsymbol{\theta}) / Q_\theta(\boldsymbol{\theta}) \right), \end{aligned} \tag{3.3}$$

where $Q_z\{\tau\} = \{Q_z(\mathbf{z}(t))|t = 1, \ldots, \tau\}$. Note that τ represents the actual amount of observed data, and it increases over time while T is fixed. The estimation of the posterior distribution $Q_z(\mathbf{z}(t))$ is inaccurate in the early stage of the online learning and gradually becomes accurate as learning proceeds. However, the early inaccurate estimations and the later accurate estimations contribute to the free energy (see equation 3.3) in equal weight. This might cause slow convergence of the learning process. Therefore, we introduce a time-dependent discount factor $\lambda(t)$ $(0 \leq \lambda(t) \leq 1, t = 2, 3, \ldots)$ for forgetting the earlier inaccurate estimation effects. Accordingly, a discounted free energy is defined by

$$\begin{aligned} F^\lambda(\mathbf{X}\{\tau\}, Q_z\{\tau\}, Q_\theta, T) = T\eta(\tau) \sum_{t=1}^{\tau} \left(\prod_{s=t+1}^{\tau} \lambda(s) \right) \\ \times \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \int d\mu(\mathbf{z}(t)) Q_z(\mathbf{z}(t)) \\ \times \log \left(P(\mathbf{x}(t), \mathbf{z}(t)|\boldsymbol{\theta}) / Q_z(\mathbf{z}(t)) \right) \\ + \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \log \left(P_0(\boldsymbol{\theta}) / Q_\theta(\boldsymbol{\theta}) \right), \end{aligned} \tag{3.4}$$

where $\eta(\tau)$ represents a normalization constant:

$$\eta(\tau) = \left[\sum_{t=1}^{\tau} \left(\prod_{s=t+1}^{\tau} \lambda(s) \right) \right]^{-1}. \tag{3.5}$$

3.2 Online Variational Bayes Algorithm. The online VB algorithm can be derived from the successive maximization of the discounted free energy (see equation 3.4). Let us assume that $Q_z\{\tau - 1\} = \{Q_z(\mathbf{z}(t))|t = 1, \ldots, \tau - 1\}$ and $Q_\theta^{(\tau-1)}(\boldsymbol{\theta})$ have been determined for an observed data set $\mathbf{X}\{\tau - 1\} =$

$\{\mathbf{x}(t)|t = 1, \ldots, \tau - 1\}$. With new observed data $\mathbf{x}(\tau)$, the discounted free energy $F^{\lambda}(\mathbf{X}\{\tau\}, Q_z\{\tau\}, Q_\theta, T)$ is maximized with respect to $Q_z(\mathbf{z}(\tau))$, while Q_θ is set to $Q_\theta^{(\tau-1)}(\boldsymbol{\theta})$. The solution is given by

$$\begin{aligned} Q_z(\mathbf{z}(\tau)) &= P(\mathbf{z}(\tau)|\mathbf{x}(\tau), \bar{\boldsymbol{\theta}}(\tau)) \\ \bar{\boldsymbol{\theta}}(\tau) &= \int d\mu(\boldsymbol{\theta}) Q_\theta^{(\tau-1)}(\boldsymbol{\theta})\boldsymbol{\theta}. \end{aligned} \tag{3.6}$$

In the next step, the discounted free energy $F^{\lambda}(\mathbf{X}\{\tau\}, Q_z\{\tau\}, Q_\theta, T)$ is maximized with respect to $Q_\theta(\boldsymbol{\theta})$, while $Q_z\{\tau\}$ is fixed. The solution is given by

$$\begin{aligned} Q_\theta^{(\tau)}(\boldsymbol{\theta}) &= \exp\left[\gamma\left(\boldsymbol{\alpha}(\tau)\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta})\right) - \Phi(\boldsymbol{\alpha}(\tau), \gamma)\right], \\ \gamma &= T + \gamma_0, \\ \gamma\,\boldsymbol{\alpha}(\tau) &= T\langle\!\langle \mathbf{r}(\mathbf{x}, \mathbf{z})\rangle\!\rangle(\tau) + \gamma_0\boldsymbol{\alpha}_0, \end{aligned} \tag{3.7}$$

where the discounted average $\langle\!\langle\cdot\rangle\!\rangle(\tau)$ is defined by

$$\begin{aligned} \langle\!\langle \mathbf{r}(\mathbf{x}, \mathbf{z})\rangle\!\rangle(\tau) &= \eta(\tau)\sum_{t=1}^{\tau}\left(\prod_{s=t+1}^{\tau}\lambda(s)\right) \\ &\quad\times\int d\mu(\mathbf{z}(t))Q_z(\mathbf{z}(t))\mathbf{r}(\mathbf{x}(t), \mathbf{z}(t)). \end{aligned} \tag{3.8}$$

The discounted average can be calculated by using a step-wise equation:

$$\begin{aligned} \langle\!\langle \mathbf{r}(\mathbf{x}, \mathbf{z})\rangle\!\rangle(\tau) &= (1 - \eta(\tau))\langle\!\langle \mathbf{r}(\mathbf{x}, \mathbf{z})\rangle\!\rangle(\tau - 1) \\ &\quad + \eta(\tau)E_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x}(\tau), \mathbf{z}(\tau))|\bar{\boldsymbol{\theta}}(\tau)\right], \\ \eta(\tau) &= \left(1 + \lambda(\tau)/\eta(\tau - 1)\right)^{-1}. \end{aligned} \tag{3.9}$$

By using equation 3.6, the expectation value of the sufficient statistics for the current data $E_{\mathbf{z}}[\mathbf{r}(\mathbf{x}(\tau), \mathbf{z}(\tau))|\bar{\boldsymbol{\theta}}(\tau)]$ is given by

$$\begin{aligned} E_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x}(\tau), \mathbf{z}(\tau))|\bar{\boldsymbol{\theta}}(\tau)\right] &= \int d\mu(\mathbf{z}(\tau)) \\ &\quad\times P(\mathbf{z}(\tau)|\mathbf{x}(\tau), \bar{\boldsymbol{\theta}}(\tau))\mathbf{r}(\mathbf{x}(\tau), \mathbf{z}(\tau)). \end{aligned} \tag{3.10}$$

The recursive formula for $\boldsymbol{\alpha}(\tau)$ is derived from the above equations:

$$\Delta\boldsymbol{\alpha}(\tau) = \boldsymbol{\alpha}(\tau) - \boldsymbol{\alpha}(\tau - 1)$$

$$= \frac{1}{\gamma}\eta(\tau)\Big(TE_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x}(\tau), \mathbf{z}(\tau))|\bar{\boldsymbol{\theta}}(\tau)\right] + \gamma_0\boldsymbol{\alpha}_0 - \gamma\,\boldsymbol{\alpha}(\tau-1)\Big). \tag{3.11}$$

By using equation 3.7, the ensemble average of the parameter $\bar{\boldsymbol{\theta}}(\tau)$ defined in equation 3.6 can be calculated as

$$\bar{\boldsymbol{\theta}}(\tau) = \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}(\tau-1)} = \frac{1}{\gamma}\frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha}(\tau-1), \gamma). \tag{3.12}$$

The online VB algorithm is summarized as follows. In the VB E-step, the ensemble average of the parameter $\bar{\boldsymbol{\theta}}(\tau)$ is determined by equation 3.12. Using this value, one calculates the expectation value of the sufficient statistics (see equation 3.10) for the current data. The posterior hyperparameter $\boldsymbol{\alpha}(\tau)$ is updated by equation 3.11 in the VB M-step. This process is repeated when new data are observed. By combining the VB E-step (3.12) and VB M-step (3.11) equations, one can get the recursive update equation for $\boldsymbol{\alpha}(\tau)$:

$$\Delta\boldsymbol{\alpha}(\tau) = \frac{1}{\gamma}\eta(\tau)\Big(TE_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x}(\tau), \mathbf{z}(\tau))|\langle\theta\rangle_{\boldsymbol{\alpha}(\tau-1)}\right] + \gamma_0\boldsymbol{\alpha}_0 - \gamma\,\boldsymbol{\alpha}(\tau-1)\Big). \tag{3.13}$$

3.3 Stochastic Approximation. Unlike the VB algorithm, the discounted free energy in the online VB algorithm does not always increase, because a new contribution is added to the discounted free energy at each time instance. In the following, we prove that the online VB algorithm can be considered as a stochastic approximation (Kushner & Yin, 1997) for finding the maximum of the expected free energy defined in equation 3.2, which gives a lower bound for the expected log evidence defined in equation 3.1. The expected free energy (see equation 3.2), in which the maximization with respect to Q_z has been performed, can be written as

$$\max_{Q_z} E\left[F(\mathbf{X}\{T\}, Q_\theta, Q_z)\right]_\rho = E\left[F_M(\mathbf{x}, \boldsymbol{\alpha}, T)\right]_\rho, \tag{3.14}$$

where

$$F_M(\mathbf{x}, \boldsymbol{\alpha}, T) = T\int d\mu(\mathbf{z})P(\mathbf{z}|\mathbf{x}, \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}) \times \int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})\log\left(P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})/P(\mathbf{z}|\mathbf{x}, \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}})\right)$$

$$+ \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \log \left(P_0(\boldsymbol{\theta}) / Q_\theta(\boldsymbol{\theta}) \right)$$

$$= T \log P(\mathbf{x}|\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}}) + \boldsymbol{\Phi}(\boldsymbol{\alpha}, \gamma) - \boldsymbol{\Phi}(\boldsymbol{\alpha}_0, \gamma_0)$$
$$- [(\gamma \boldsymbol{\alpha} - \gamma_0 \boldsymbol{\alpha}_0) \cdot \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} - T \boldsymbol{\psi}(\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}})] \tag{3.15}$$

The gradient of F_M is calculated as

$$\frac{\partial F_M}{\partial \boldsymbol{\alpha}}(\mathbf{x}, \boldsymbol{\alpha}, T) = \gamma V_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \gamma)$$
$$\times \left(T E_{\mathbf{z}} [\mathbf{r}(\mathbf{x}, \mathbf{z}) | \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}}] + \gamma_0 \boldsymbol{\alpha}_0 - \gamma \boldsymbol{\alpha} \right), \tag{3.16}$$

where the Fisher information matrix $V_{\boldsymbol{\alpha},\boldsymbol{\alpha}}$ for the posterior parameter distribution, $Q_\theta(\boldsymbol{\theta}) = P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma)$, is defined in equation 2.31. Consequently, the online VB algorithm, 3.13, can be written as

$$\Delta \boldsymbol{\alpha}(\tau) = \frac{1}{\gamma^2} \eta(\tau) V^{-1}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\alpha}(\tau - 1), \gamma) \cdot \frac{\partial F_M}{\partial \boldsymbol{\alpha}}(\mathbf{x}(\tau), \boldsymbol{\alpha}(\tau - 1), T). \tag{3.17}$$

If the effective learning rate $\eta(\tau)(\geq 0)$ satisfies the condition (Kushner & Yin, 1997)

$$\sum_{t=1}^{\infty} \eta(\tau) = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta^2(\tau) < \infty, \tag{3.18}$$

the online VB algorithm, equation 3.13, defines the stochastic approximation for finding the maximum of the expected free energy (see equation 3.2).

When there is no discount factor, that is, $\lambda(\tau) = 1$, the effective learning rate $\eta(\tau)$ is given by

$$\eta(\tau) = 1/\tau. \tag{3.19}$$

This satisfies the stochastic approximation condition, equation 3.18. However, the learning speed becomes very slow if this schedule is adopted (see section 5). The reason for this slow convergence is that earlier inaccurate hyperparameter estimations affect the hyperparameter estimations in later learning stages because there is no discount factor in the sufficient statistics average (see equation 3.8). The introduction of the discount factor is crucial for fast convergence.

As in the online EM algorithm we proposed previously (Sato, 2000), we employ the following discount schedule,

$$1 - \lambda(\tau) = \frac{1}{(\tau - 2)\kappa + \tau_0}, \tag{3.20}$$

which can be calculated recursively:

$$\lambda(\tau) = 1 - \frac{1 - \lambda(\tau - 1)}{1 + \kappa(1 - \lambda(\tau - 1))}. \tag{3.21}$$

The corresponding effective learning rate $\eta(\tau)$ satisfies

$$\eta(\tau) \underset{\tau \to \infty}{\longrightarrow} \left(\frac{\kappa + 1}{\kappa}\right) \frac{1}{\tau}, \tag{3.22}$$

so that the stochastic approximation condition (see equation 3.18) is satisfied. The constants appearing in equation 3.20 have clear physical meanings. τ_0 represents how many samples contribute to the discounted average for the sufficient statistics (see equation 3.8) in the early stage of learning. κ controls the asymptotic decreasing ratio for the effective learning constant $\eta(\tau)$, as in equation 3.22. The values of τ_0 and κ control the learning speeds in the early and later stages of learning, respectively.

3.4 Discussion. There are some comments on our online VB method. The objective function of our online VB method is the expected log evidence (see equation 3.1) for a fixed amount of data. Unlike the usual Bayesian method, our online VB method estimates the same quantity even if the amount of the observed data is increased. The observed data are used for improving the estimation quality of the expected free energy (see equation 3.2) which approximates the expected log evidence (see equation 3.1). Therefore, our online VB method does not converge to the EM algorithm in the large sample limit ($\tau \to \infty$ and T : fixed). Nevertheless, the method can perform both learning and prediction concurrently in an online fashion.

The batch VB method can be derived from our online VB method if the same data set is repeatedly presented. Let suppose that the posterior hyperparameters are updated only at the end of each epoch, in which all of the data are presented once, by using equation 3.7, while the discounted averages are calculated every time by using equation 3.9. If the discount factor is given by $\lambda(\tau) = 0$ at the beginning of each epoch, or $\lambda(\tau) = 1$ otherwise, then this online VB method is equivalent to the batch VB method.

It is also possible to use the log evidence for the observed data as in the usual Bayesian method. The corresponding online VB algorithm can be derived by equating the amount of data τ to T in equation 3.3. In this case, the objective function changes over time as the amount of data increases. The target posterior parameter distribution also changes over time. The variance of the posterior parameter distribution is proportional to $1/\tau$. This gives too much confidence for the posterior parameter distribution because the free energy is not fully maximized for each datum; that is, the VB-E and VB-M steps are performed once for each datum. Consequently, this type of online VB method soon converges to the EM algorithm: the posterior parameter

distribution becomes sharply peaked near the maximum likelihood estimator. In order to avoid this difficulty, the VB-E and VB-M steps should be iterated until convergence for each datum. However, the above algorithm is not an online algorithm in usual sense.

4 Online Model Selection

In the usual Bayesian procedure for model selection, one prepares a set of models with different structures and calculates the evidence for each model. Then the best model that gives the highest evidence is selected or the average over models with different structures is taken. An alternative approach is to change model structure sequentially. Sequential model selection procedures have been proposed based on either the batch VB method (Ghahramani & Beal, 2000; Ueda, 1999) or the MCMC method (Richardson & Green, 1997).

In this article, we adopt sequential model selection procedures based on the online VB method. We can consider several sequential model selection strategies. In the first method, we start from an initial model with a given structure. The VB learning process for this model is continued by monitoring the free energy value. When the free energy converges, the model structure is changed according to some criterion, and the initial model is saved as the base model. The VB learning process for the current model is continued until the free energy converges. If the free energy of the current model is greater than that of the base model, the current model is saved as the base model. Otherwise the base model is not changed. One always keeps the base model as the best model to date. A new trial model is selected based on the base model. This process continues until further attempts do not improve the base model.

The above procedure is a deterministic process. We can consider a stochastic model selection process based on the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). In this case, the base model is different from the best model to date. If the free energy of the current model is greater than that of the base model, the current model is saved as the base model. Otherwise the base model is changed to the current model with the probability $\exp(\beta(F_{current} - F_{base}))$. This stochastic process can be applied to model selection in dynamic environments.

We also propose an online model selection procedure using a hierarchical mixture of models with different structures. We train them concurrently by using the online VB method for each model. They are trained independently with each other. (It is also possible to train them competitively by using the online VB method for the hierarchical mixture of models.) When the free energies converge, these models are compared according to their free energy values. Then the current best model structure is selected. Structures of the other models are changed based on the current best model structure. This sequential model selection procedure is suitable for dynamic environment.

This method is used in a model selection task for dynamic environments in section 5.

In the next section, we study the model selection problem for mixture of gaussian models. As a mechanism for structural change, we adopt the split-and-merge method proposed by Ueda et al. (1999) (see also Richardson & Green, 1997; Ghahramani & Beal, 2000; Ueda, 1999). For mixture models, the split-and-merge method provides a simple procedure for structural changes. We choose either to split a unit into two or to merge two units into one in the sequential model selection process. In the current implementation, the same process is applied if the previous attempt was successful. Otherwise the other process is applied.

A criterion for splitting a unit is given by the unit's free energy, which is assigned to each unit (see appendix C). The split is applied to the unit with the lowest free energy among unattempted units. A criterion for merging units is given by the correlation between the two units' activities, which are represented by the posterior probability that the units will be selected for given data. The unit pair with the highest correlation among unattempted unit pairs is selected for merging. The deletion of units is also performed for units with very small activities, which indicate that the units have not been selected at all.

We adopted the above model selection procedure because of its simplicity. Other model selection procedures using the split-and-merge algorithm have also been proposed (Ghahramani & Beal, 2000; Ueda, 1999).

By combining the sequential model selection procedure with the online VB learning method, a fully online learning method with a model selection mechanism is obtained, and it can be applied to real-time applications.

5 Experiments

As a preliminary study on the performance of the online VB method, we considered model selection problems for two-dimensional mixture of gaussian (MG) models (see appendix C). We borrowed two tasks from Roberts et al. (1998). Data set A, consisting of 200 points, was generated from a mixture of four gaussians with the centers $(0, 0), (2, \sqrt{12}), (4, 0)$, and $(-2, -\sqrt{12})$ (see Figure 1A). The gaussians had the same isotropic variance $\sigma^2 = (1.2)^2$. In addition, data set B, consisting of 1000 points, was generated from a mixture of four gaussians (see Figure 1B). In this case, they were paired such that each pair had a common center—$\mathbf{m}_1 = \mathbf{m}_2 = (2, \sqrt{12})$ and $\mathbf{m}_3 = \mathbf{m}_4 = (-2, -\sqrt{12})$—but different variances—$\sigma_1^2 = \sigma_3^2 = (1.0)^2$ and $\sigma_2^2 = \sigma_4^2 = (5.0)^2$. Although these models were simple, the model selection tasks for them were rather difficult because of the overlap between the gaussians (Roberts et al., 1998).

In the first experiment, we examined the usual Bayes model selection procedure. A set of models consisting of different numbers of units was

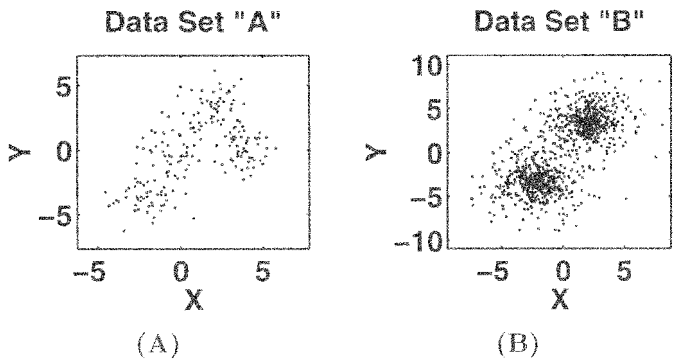

Figure 1

Figure 1: (A) 200 points in data set A generated from mixture of four gaussians with different centers. (B) 1000 points in data set B generated from mixture of four gaussians. Pairs of gaussians have the same centers but different variances.

prepared. The VB method was applied to each model, and the maximum free energy was calculated. Three VB methods were compared: the batch VB, online VB without the discount factor, and online VB with the discount factor whose schedule is given by equation 3.20 with $\tau_0 = 100$ and $\kappa = 0.01$. We used a nearly noninformative prior for all cases: $\gamma_0 = 0.01$.

The learning speed was measured according to epoch numbers. In one epoch, all training data were supplied to each VB method once. The online VB method updated the ensemble average of parameters for each datum, while the batch VB method updated them once according to the average of the sufficient statistics over all of the training data.

The results are summarized in Figure 2. Both the batch VB method and the online VB method gave the highest free energy for the true model consisting of four units. The online VB method with the discount factor showed faster and better performance than the batch VB method, especially for large amounts of data (see Figure 2). The reason for this performance difference can be considered as follows. In the online VB method, the posterior probability for hidden variables is calculated by using the newly calculated ensemble average of the parameters improved at each observation. The batch VB method, in contrast, uses the ensemble average of the parameters calculated in the previous epoch for all data. Therefore, the estimation quality of the posterior probability for the hidden variables improves rather slowly. This becomes more prominent for larger amounts of data. In this case, the online VB method can find the optimal solution within one epoch, as shown in Figure 2B.

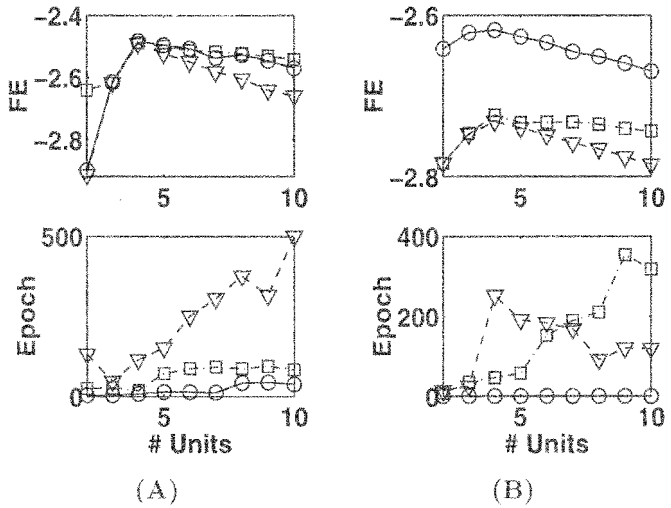

Figure 2

Figure 2: Maximum free energies (FE) obtained by three learning methods and their convergence times measured by epoch numbers are plotted for various models. Three methods are batch VB (dash-dotted line with square), online VB with discount factor (solid line with circles), and online VB without discount factor (dashed line with triangles). The abscissa denotes the number of gaussian units in trained models. (A) Results for data set A. (B) Results for data set B.

The online VB method without the discount factor showed poor performance and slow convergence for all cases. This result implies that the introduction of the discount factor is crucial for good performance of the online VB method, as pointed out in section 3. If there is no discount factor, the early inaccurate estimations contribute to the sufficient statistics average even in the later stages of the learning process and degrade the quality of estimations.

In the second experiment, the sequential model selection procedure using a trial model (see section 4) was tested. When the free energy converged, the structure of the trial model was changed based on the base model, which was the best model to date. We tested two initial model configurations consisting of 2 units and 10 units. When the model structure was changed, the discount factor and the effective learning constant in the online VB method were reset as $(1 - \lambda(\tau)) = 0.01$ and $\eta(\tau) = 0.01$. The online VB method was able to find the best model in all cases (see Figures 3 and 4). It should be noted that the VB method sometimes increased the free energy while decreasing the data likelihood (see Figures 3–6). This was achieved as a result of the decrease

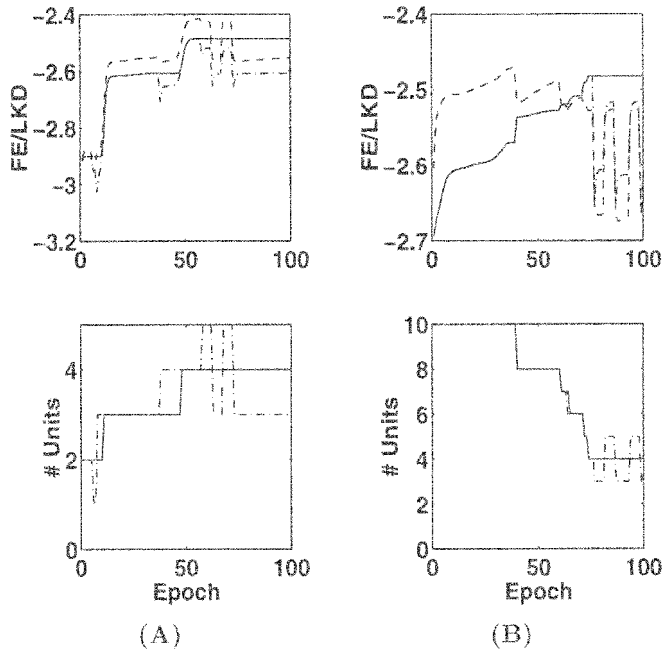

Figure 3

Figure 3: Online model selection processes using the online VB method for data set A. Free energies (FE) and number of units for the base model (solid line), which is the best model to date, and the current trial model (dash-dotted line) are shown. Log likelihood (LKD) for the current trial model (dashed line) is also shown. (A) The initial model consists of 2 units. (B) The initial model consists of 10 units.

in the model complexity. The batch VB method also found the best model (see Figures 5 and 6) except for one case, in which the batch VB method got stuck in a local maximum (see Figure 6A).

We also examined the online model selection procedure for a dynamic environment using a hierarchical mixture of MG models described in section 4. The hierarchical mixture in this experiment consisted of two MG models. Both models started from the same model structure with 10 units. When the free energies converged, the structure of a MG model with lower free energy was changed. At the moment, the discount factor and the effective learning constant were reset as $(1 - \lambda(\tau)) = 0.01$ and $\eta(\tau) = 0.01$.

In this experiment, the data generation model was changed at the fifty-first epoch. In the first 50 epochs, the number of gaussians was four. After the

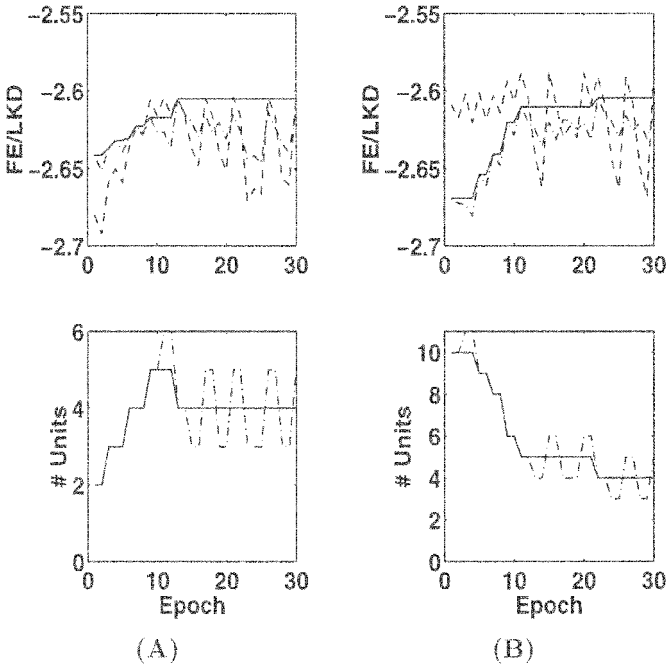

Figure 4

Figure 4: Online model selection processes using the online VB method for data set B. Free energies (FE) and number of units for the base model (solid line), which is the best model to date, and the current trial model (dash-dotted line) are shown. Log likelihood (LKD) for the current trial model (dashed line) is also shown. (A) The initial model consists of 2 units. (B) The initial model consists of 10 units.

fifty-first epoch, the number of gaussians was six. In the learning process, each model observed 1000 data in one epoch. Figure 7 shows the learning process. After the twentieth epoch, the optimal model structure with four units was found. When the environment was changed at the fifty-first epoch, the free energies of both models dropped dramatically. The trained models tried to search for the new optimal structure. After the seventy-fifth epoch, the optimal model structure with six units was found. The result showed that the online VB method was able to adjust the model structure to dynamic environment.

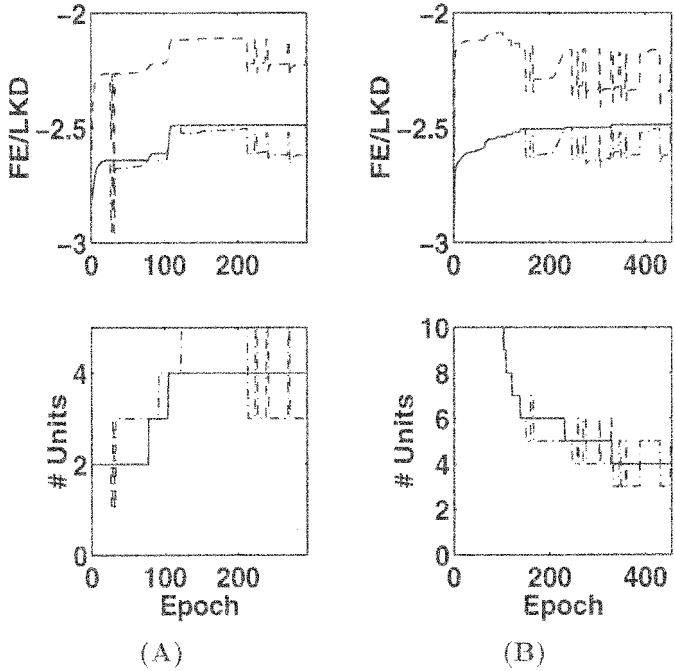

Figure 5

Figure 5: Sequential model selection processes using the batch VB method for data set A. Free energies (FE) and number of units for the base model (solid line), which is the best model to date, and the current trial model (dash-dotted line) are shown. Log likelihood (LKD) for the current trial model (dashed line) is also shown. (A) The initial model consists of 2 units. (B) The initial model consists of 10 units.

6 Conclusion

We derived an online version of the VB algorithm and proved its convergence by showing that it is a stochastic approximation for finding the maximum of the free energy. A fully online learning method with a model selection mechanism was also proposed based on the online VB method, together with a sequential model selection procedure. This method can be applied to real-time applications.

We considered the Bayes model without hierarchy. The current method can be easily extended to the hierarchical Bayes model (see appendix D).

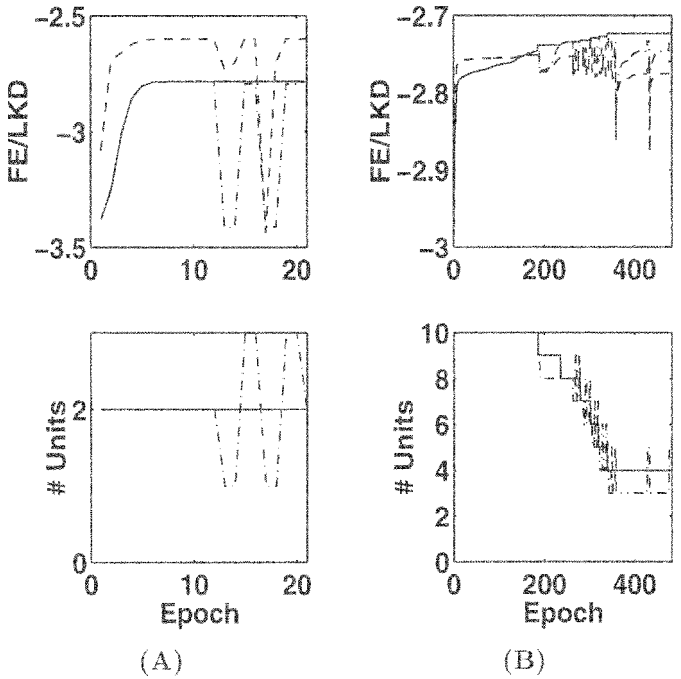

Figure 6

Figure 6: Sequential model selection processes using the batch VB method for data set B. Free energies (FE) and number of units for the base model (solid line), which is the best model to date, and the current trial model (dash-dotted line) are shown. Log likelihood (LKD) for the current trial model (dashed line) is also shown. (A) The initial model consists of 2 units. (B) The initial model consists of 10 units.

In preliminary experiments using synthetic data, the online VB method was able to adapt the model structure to dynamic environments. A detailed study on the performance of the online VB method will be published in a forthcoming article. It also remains for future study to find better sequential model selection procedure.

Appendix A

Free energy maximization can be done by using the following theorem:

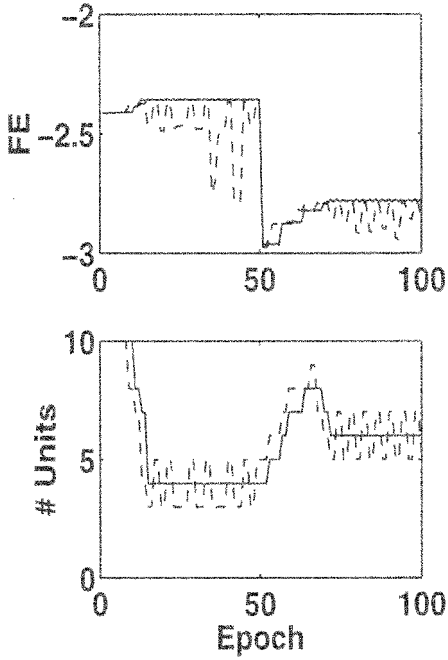

Figure 7

Figure 7: Online model selection processes for dynamic environment using hierarchical mixture of MG models. The number of units in the data generation model is four in the first 50 epochs and six after the fifty-first epoch. Free energies (FE) and number of units for best model (solid line), which is selected at the previous competition, and trial model (dashed line) are shown.

Theorem 1. *The maximum of* $(\int d\mu(\mathbf{y})Q(\mathbf{y})(f(\mathbf{y}) - \log(Q(\mathbf{y}))))$ *under the condition,* $\int d\mu(\mathbf{y})Q(\mathbf{y}) = 1$, *is given by*

$$Q(\mathbf{y}) = \frac{\exp[f(\mathbf{y})]}{\int d\mu(\mathbf{y}')\exp[f(\mathbf{y}')]}.$$

The theorem can be proved with the help of the Lagrange multiplier method. The VB equations, 2.13 through 2.19, can be proved by using this theorem and the following relations:

$$F = \int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})$$
$$\times \left[T\left(\langle \mathbf{r}(\mathbf{x},\mathbf{z})\rangle_{Q_z}\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta})\right) + \log P_0(\boldsymbol{\theta}) - \log Q_\theta(\boldsymbol{\theta})\right]$$

$$
\begin{aligned}
&+ Q_\theta(\boldsymbol{\theta})\text{-independent terms} \\
&= \int d\mu(\mathbf{Z}\{T\}) Q_z(\mathbf{Z}\{T\}) \\
&\quad \times \left[\sum_{t=1}^{T} (\mathbf{r}(\mathbf{x}(t), \mathbf{Z}(t)) \cdot \langle \boldsymbol{\theta} \rangle_{Q_\theta} + \mathbf{r}_0(\mathbf{x}(t), \mathbf{Z}(t))) - \log Q_z(\mathbf{Z}\{T\}) \right] \\
&\quad + Q_z(\mathbf{Z}\{T\}) \text{ - independent terms,}
\end{aligned}
$$

where $\langle \cdot \rangle_{Q_\theta}$ and $\langle \cdot \rangle_{Q_z}$ denote the expectation value with respect to $Q_\theta(\boldsymbol{\theta})$ and $Q_z(\mathbf{Z}\{T\})$, respectively.

Appendix B

The calculation of the derivative of the parameterized free energy (see equation 2.26) is lengthy but straightforward. The outline of the calculation is shown below. The derivative with respect to $\bar{\boldsymbol{\theta}}$ can be calculated as

$$
\partial F / \partial \bar{\boldsymbol{\theta}} = T \left(\frac{\partial}{\partial \bar{\boldsymbol{\theta}}} \langle \mathbf{r}(\mathbf{x}, \mathbf{z}) \rangle_{\bar{\boldsymbol{\theta}}} \right) (\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} - \bar{\boldsymbol{\theta}}),
$$

by using the relation

$$
\frac{\partial}{\partial \bar{\boldsymbol{\theta}}} \left(\sum_{t=1}^{T} \log P(\mathbf{x}(t) | \bar{\boldsymbol{\theta}}) \right) = T \left(\langle \mathbf{r}(\mathbf{x}, \mathbf{z}) \rangle_{\bar{\boldsymbol{\theta}}} - \frac{\partial \psi}{\partial \bar{\boldsymbol{\theta}}} \right).
$$

The coefficient matrix $T(\partial \langle \mathbf{r} \rangle_{\bar{\boldsymbol{\theta}}} / \partial \bar{\boldsymbol{\theta}})$ turns out to be $U(\boldsymbol{\theta})$ defined in equation 2.28.

The derivatives with respect to $(\boldsymbol{\alpha}, \gamma)$ are given by

$$
\begin{aligned}
\frac{1}{\gamma} \frac{\partial F}{\partial \boldsymbol{\alpha}} &= \left(\frac{1}{\gamma^2} \frac{\partial^2 \Phi}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \right) \cdot \left(T \langle \mathbf{r}(\mathbf{x}, \mathbf{z}) \rangle_{\bar{\boldsymbol{\theta}}} + \gamma_0 \boldsymbol{\alpha}_0 - (T + \gamma_0) \boldsymbol{\alpha} \right) \\
&\quad + (T + \gamma_0 - \gamma) \left(\frac{1}{\gamma} \frac{\partial^2 \Phi}{\partial \boldsymbol{\alpha} \partial \gamma} - \frac{1}{\gamma^2} \frac{\partial \Phi}{\partial \boldsymbol{\alpha}} \right), \\
\frac{\partial F}{\partial \gamma} &= \left(\frac{1}{\gamma} \frac{\partial^2 \Phi}{\partial \boldsymbol{\alpha} \partial \gamma} - \frac{1}{\gamma^2} \frac{\partial \Phi}{\partial \boldsymbol{\alpha}} \right) \cdot \left(T \langle \mathbf{r}(\mathbf{x}, \mathbf{z}) \rangle_{\bar{\boldsymbol{\theta}}} + \gamma_0 \boldsymbol{\alpha}_0 - (T + \gamma_0) \boldsymbol{\alpha} \right) \\
&\quad + (T + \gamma_0 - \gamma) \left(\frac{\partial^2 \Phi}{\partial \gamma \, \partial \gamma} \right).
\end{aligned}
$$

Equations 2.30 and 2.31 can be derived by using the above and the following equations:

$$
\frac{1}{\gamma^2} \frac{\partial^2 \Phi}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} = \frac{1}{\gamma^2} \left\langle \left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}} \right) \left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}^T} \right) \right\rangle_{\boldsymbol{\alpha}},
$$

$$\frac{1}{\gamma}\frac{\partial^2\Phi}{\partial\boldsymbol{\alpha}\partial\gamma} - \frac{1}{\gamma^2}\frac{\partial\Phi}{\partial\boldsymbol{\alpha}} = \frac{1}{\gamma}\left\langle\left(\frac{\partial\log P_\alpha}{\partial\boldsymbol{\alpha}}\right)\left(\frac{\partial\log P_\alpha}{\partial\gamma}\right)\right\rangle_{\boldsymbol{\alpha}},$$

$$\frac{\partial^2\Phi}{\partial\gamma\,\partial\gamma} = \left\langle\left(\frac{\partial\log P_\alpha}{\partial\gamma}\right)\left(\frac{\partial\log P_\alpha}{\partial\gamma}\right)\right\rangle_{\boldsymbol{\alpha}}.$$

Appendix C

The VB algorithm for the mixture of Gaussian model is briefly explained in this appendix. (Notations in this appendix are slightly different from those in the text.) We first explain more general mixture models: the mixture of exponential family (MEF) models. The probability distribution for the ith unit in the MEF model is defined by

$$P(\mathbf{x}|\boldsymbol{\theta}_i, i) = \exp[\mathbf{r}_i(\mathbf{x})\cdot\boldsymbol{\theta}_i + r_{i,0}(\mathbf{x}) - \Psi_i(\boldsymbol{\theta}_i)]. \tag{C.1}$$

The conjugate distribution for equation C.1 is given by

$$P_\alpha(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i, \gamma_i, i) = \exp[\gamma_i(\boldsymbol{\alpha}_i\cdot\boldsymbol{\theta}_i - \Psi_i(\boldsymbol{\theta}_i)) - \Phi_i(\boldsymbol{\alpha}_i, \gamma_i)]. \tag{C.2}$$

The probability distribution for the MEF model is then defined by

$$\begin{aligned} P(\mathbf{x}|\mathbf{g}, \boldsymbol{\theta}) &= \sum_{i=1}^{M} g_i P(\mathbf{x}|\boldsymbol{\theta}_i, i) \\ &= \sum_{\{\mathbf{z}\}}\exp\left[\sum_{i=1}^{M}\Big(z_i(\log g_i - \Psi_i(\boldsymbol{\theta}_i)) \right. \\ &\qquad \left. + z_i\mathbf{r}_i(\mathbf{x})\cdot\boldsymbol{\theta}_i + z_i r_{i,0}(\mathbf{x})\Big)\right], \end{aligned} \tag{C.3}$$

where the hidden variable $\mathbf{z} = \{z_i | i = 1, \ldots, M\}$ is an indicator variable: $z_i = 0$ or 1, and $\sum_{i=1}^{M} z_i = 1$. $\sum_{\{\mathbf{z}\}}$ denotes the summation over M possible configurations of $\mathbf{z}$. The mixing proportion $\mathbf{g} = \{g_i | i = 1, \ldots, M\}$ satisfies the constraint $\sum_{i=1}^{M} g_i = 1$, which is automatically satisfied by the expression $g_i = e^{\varphi_i}/(\sum_{j=1}^{M} e^{\varphi_i})$.

The set of model parameters $\{\mathbf{g}, \boldsymbol{\theta}\} = \{g_i, \boldsymbol{\theta}_i | i = 1, \ldots, M\}$ is not the natural parameter of the MEF model (see equation C.3). The natural parameter is given by $\{\boldsymbol{\omega}, \boldsymbol{\theta}\} = \{\omega_i = \varphi_i - \Psi_i(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i | i = 1, \ldots, M\}$. The corresponding sufficient statistics is given by $\{z_i, z_i\mathbf{r}_i(\mathbf{x}) | i = 1, \ldots, M\}$. Accordingly, the MEF model can be written as the EFH model:

$$P(\mathbf{x}|\boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{\{\mathbf{z}\}}\exp\left[\sum_{i=1}^{M}\left(z_i\omega_i + z_i\mathbf{r}_i(\mathbf{x})\cdot\boldsymbol{\theta}_i\right) - \Psi_\theta(\boldsymbol{\omega}, \boldsymbol{\theta})\right], \tag{C.4}$$

$$\Psi_\theta(\boldsymbol{\omega}, \boldsymbol{\theta}) = \log\left[\sum_{i=1}^{M} \exp(\omega_i + \Psi_i(\boldsymbol{\theta}_i))\right]. \tag{C.5}$$

The conjugate distribution for the MEF model, equation C.3, is given by the product of the Dirichlet distribution and the conjugate distribution for each unit:

$$\begin{aligned} P_\alpha(\mathbf{g}, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\nu}, \gamma) &= \exp\left[\gamma \sum_{i=1}^{M} \nu_i \left(\log g_i - \Psi_i(\boldsymbol{\theta}_i) + \boldsymbol{\alpha}_i \cdot \boldsymbol{\theta}_i\right) - \Phi_\alpha(\boldsymbol{\alpha}, \boldsymbol{\nu}, \gamma)\right] \\ &= \exp\left[\gamma \sum_{i=1}^{M} (\nu_i \omega_i + \nu_i \boldsymbol{\alpha}_i \cdot \boldsymbol{\theta}_i)\right. \\ &\quad \left. -\gamma\, \Psi_\theta(\boldsymbol{\omega}, \boldsymbol{\theta}) - \Phi_\alpha(\boldsymbol{\alpha}, \boldsymbol{\nu}, \gamma)\right], \end{aligned} \tag{C.6}$$

$$\begin{aligned} \Phi_\alpha(\boldsymbol{\alpha}, \boldsymbol{\nu}, \gamma) &= \sum_{i=1}^{M} \log\Gamma(\gamma\, \nu_i + 1) \\ &\quad - \log\Gamma(\gamma + M) + \sum_{i=1}^{M} \Phi_i(\boldsymbol{\alpha}_i, \gamma\, \nu_i), \end{aligned} \tag{C.7}$$

where ν_i satisfies $\sum_{i=1}^{M} \nu_i = 1$, and $\Gamma(\gamma)$ is the gamma function, that is, $\Gamma(\gamma) = \int_0^\infty ds e^{-s} s^{\gamma - 1}$. The VB algorithm for the MEF model can be derived by using Φ_α as described in section 2. The VB E-step equation is given by

$$\bar{\boldsymbol{\theta}}_i = \langle \boldsymbol{\theta}_i \rangle_{\boldsymbol{\alpha}} = \frac{1}{\gamma\, \nu_i} \frac{\partial \Phi_\alpha}{\partial \boldsymbol{\alpha}_i}, \tag{C.8}$$

$$\bar{\omega}_i = \langle \omega_i \rangle_{\boldsymbol{\alpha}} = \frac{1}{\gamma} \frac{\partial \Phi_\alpha}{\partial \nu_i} - \boldsymbol{\alpha}_i \cdot \langle \boldsymbol{\theta}_i \rangle_{\boldsymbol{\alpha}}. \tag{C.9}$$

The VB M-step equation is given by

$$\gamma = T + \gamma(0), \tag{C.10}$$

$$\nu_i = \frac{1}{\gamma}\left(T \langle z_i \rangle_{\bar{\boldsymbol{\theta}}} + \gamma(0) \nu_i(0)\right), \tag{C.11}$$

$$\boldsymbol{\alpha}_i = \frac{1}{\gamma\, \nu_i}\left(T \langle z_i \mathbf{r}_i(\mathbf{x}) \rangle_{\bar{\boldsymbol{\theta}}} + \gamma(0) \nu_i(0) \boldsymbol{\alpha}_i(0)\right), \tag{C.12}$$

where $\gamma(0)$ and $\{\nu_i(0), \boldsymbol{\alpha}_i(0) | i = 1, \ldots, M\}$ are the prior hyperparameters of the prior parameter distribution. $\langle \cdot \rangle_{\bar{\boldsymbol{\theta}}}$ denotes the expectation value (see equation 2.19) with respect to $P(\mathbf{z}|\mathbf{x}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{\theta}})$.

The free energy of the MEF model after the VB M-step is expressed as

$$F = \sum_{i=1}^{M} \Big[T\langle z_i \log P(\mathbf{x}|\bar{\omega}, \bar{\theta})\rangle_{\bar{\theta}} - T\langle z_i \log P(\mathbf{x}, \mathbf{z}|\bar{\omega}, \bar{\theta})\rangle_{\bar{\theta}} + \log\Gamma(\gamma\,\nu_i + 1) - \nu_i \log\Gamma(\gamma\, + M) - \log\Gamma(\gamma\,(0)\nu_i(0) + 1) + \nu_i(0)\log\Gamma(\gamma\,(0) + M) + \Phi_i(\alpha_i, \gamma\,\nu_i) - \Phi_i(\alpha_i(0), \gamma\,(0)\nu_i(0)) \Big]. \tag{C.13}$$

The mixture of Gaussian (MG) model is obtained when the component distribution $P(\mathbf{x}|\boldsymbol{\theta}_i, i)$ is the normal distribution:

$$P(\mathbf{x}|\mathbf{m}_i, \Sigma_i, i) = (2\pi)^{-N/2}|\Sigma_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \Sigma_i (\mathbf{x} - \mathbf{m}_i)\right]$$

$$= \exp\left[-\frac{1}{2}\mathbf{x}^T\Sigma_i\mathbf{x} + \mathbf{x}^T\Sigma_i\mathbf{m}_i - \Psi_i(\mathbf{m}_i, \Sigma_i)\right], \tag{C.14}$$

$$\Psi_i(\mathbf{m}_i, \Sigma_i) = \frac{1}{2}\mathbf{m}_i^T\Sigma_i\mathbf{m}_i + \frac{1}{2}\log|\Sigma_i| - \frac{N}{2}\log(2\pi), \tag{C.15}$$

where $\mathbf{m}_i$ and Σ_i denote the center and the inverse covariance matrix of the ith gaussian. The natural parameter of the normal distribution is given by $\boldsymbol{\theta}_i = (\Sigma_i, \Sigma_i\mathbf{m}_i)$. The conjugate distribution for the normal distribution (see equation C.14) is given by the normal Wishart distribution (Gelman et al., 1995),

$$P_\alpha(\mathbf{m}_i, \Sigma_i|\mathbf{c}_i, \Delta_i, \gamma_i) = \exp\Bigg[-\frac{1}{2}\gamma_i(\mathbf{m}_i - \mathbf{c}_i)^T\Sigma_i(\mathbf{m}_i - \mathbf{c}_i) - \frac{1}{2}\gamma_i \mathrm{Tr}(\Sigma_i\Delta_i^{-1}) + \frac{1}{2}(\gamma_i - N)\log|\Sigma_i| - \Phi_i(\Delta_i, \gamma_i)\Bigg], \tag{C.16}$$

$$\Phi_i(\Delta_i, \gamma_i) = \frac{1}{2}\gamma_i \log|\Delta_i^{-1}| + \sum_{n=1}^{N}\log\Gamma\left(\frac{\gamma_i + 1 - n}{2}\right) - \frac{1}{2}\gamma_i N\log\left(\frac{\gamma_i}{2}\right) - \frac{N}{2}\log\left(\frac{\gamma_i}{2\pi}\right) + \frac{1}{4}N(N-1)\log\pi. \tag{C.17}$$

The natural parameter of the conjugate distribution, equation C.16, is given by

$$(\gamma_i\boldsymbol{\alpha}_i, \gamma_i) = (\gamma_i(\Delta_i^{-1} + \mathbf{c}_i\mathbf{c}_i^T), \gamma_i\mathbf{c}_i, \gamma_i). \tag{C.18}$$

The VB algorithm for the MG model can be derived by using the above equations.

Appendix D

The VB method can be easily extended to the hierarchical Bayes model. Let us consider the EFH model (see equation 2.1) with the prior distribution $P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}_0, \gamma_0)$. The evidence for the hierarchical Bayes model is given by the marginal likelihood with respect to the model parameter $\boldsymbol{\theta}$ and the prior hyperparameter $\boldsymbol{\alpha}_0$,

$$P(\mathbf{X}\{T\}) = \int d\mu(\theta) d\mu(\boldsymbol{\alpha}_0) P(\mathbf{X}\{T\}|\boldsymbol{\theta}) P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}_0, \gamma_0) P_0(\boldsymbol{\alpha}_0), \tag{D.1}$$

where $P_0(\boldsymbol{\alpha}_0)$ is the prior distribution for the prior hyperparameter $\boldsymbol{\alpha}_0$. The free energy is defined by

$$F(\mathbf{X}\{T\}, Q) = \int d\mu(\boldsymbol{\theta}) d\mu(\boldsymbol{\alpha}_0) d\mu(\mathbf{Z}\{T\}) Q(\boldsymbol{\theta}, \boldsymbol{\alpha}_0, \mathbf{Z}\{T\}) \times \log\left(\frac{P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}_0, \gamma_0) P_0(\boldsymbol{\alpha}_0)}{Q(\boldsymbol{\theta}, \boldsymbol{\alpha}_0, \mathbf{Z}\{T\})}\right). \tag{D.2}$$

The hierarchical VB method can be obtained assuming the conjugate prior for $P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}_0, \gamma_0)$,

$$P_0(\boldsymbol{\alpha}_0) = \exp\left[b_0\left(\mathbf{a}_0 \boldsymbol{\alpha}_0 \gamma_0 - \Phi_\alpha(\boldsymbol{\alpha}_0, \gamma_0)\right) - \Phi_a(\mathbf{a}_0, b_0)\right], \tag{D.3}$$

and the factorization for the trial posterior distribution,

$$Q(\boldsymbol{\theta}, \boldsymbol{\alpha}_0, \mathbf{Z}\{T\}) = Q_\theta(\boldsymbol{\theta}) Q_\alpha(\boldsymbol{\alpha}_0) Q_z(\mathbf{Z}\{T\}). \tag{D.4}$$

The remaining calculations can be done by the same way as in the VB method. The VB algorithm in this case consists of three steps. The posterior probability for the hidden variable $P(\mathbf{z}(t)|\mathbf{x}(t), \bar{\boldsymbol{\theta}})$ is calculated in the VB E-step by using the ensemble average of the parameters

$$\bar{\boldsymbol{\theta}} = \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}}. \tag{D.5}$$

The posterior hyperparameter $\boldsymbol{\alpha}$ is calculated in the VB M-step,

$$\gamma\, \boldsymbol{\alpha} = T \langle \mathbf{r}(\mathbf{x}, \mathbf{z}) \rangle_{\bar{\boldsymbol{\theta}}} + \gamma_0 \langle \boldsymbol{\alpha}_0 \rangle_{\mathbf{a}}, \tag{D.6}$$

$$\gamma_0 \langle \boldsymbol{\alpha}_0 \rangle_{\mathbf{a}} = \gamma_0 \int d\mu(\boldsymbol{\alpha}_0) Q_\alpha(\boldsymbol{\alpha}_0) \boldsymbol{\alpha}_0 = \frac{1}{b} \frac{\partial \Phi_a}{\partial \mathbf{a}}(\mathbf{a}, b), \tag{D.7}$$

together with $\gamma = T + \gamma_0$. The posterior hyper-hyperparameter $(\mathbf{a}, b)$ is then calculated:

$$\mathbf{a} = \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} + \mathbf{a}_0, \tag{D.8}$$

$$b = b_0 + 1, \tag{D.9}$$

$$\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} = \frac{1}{\gamma} \frac{\partial \Phi_\alpha}{\partial \boldsymbol{\alpha}} (\boldsymbol{\alpha}, \gamma). \tag{D.10}$$

Repeating the above three steps, the free energy monotonically increases. The online VB algorithm can be similarly derived.

References

Amari, S. (1985). *Differential geometrical method in statistics*. New York: Springer-Verlag.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation, 10*, 251–276.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence* (pp. 21–30).

Attias, H. (2000). A variational Bayesian framework for graphical models. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12*. Cambridge, MA: MIT Press (pp. 206–212).

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Bishop, C. M. (1999). Variational principal components. In *IEE Conference Publication on Artificial Neural Networks ICANN99* (pp. 509–514).

Chickering, D. M., & Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning, 29*, 181–212.

Cooper, G., & Herskovitz, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–347.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B, 39*, 1–22.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Ghahramani, Z., & Beal, M. J. (2000). Variational inference for Bayesian mixture of factor analysers. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12*. Cambridge, MA: MIT Press (pp. 449–455).

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*, 197–243.

Kushner, H. J., & Yin, G. G. (1997). *Stochastic approximation algorithms and applications*. New York: Springer-Verlag.

Mackay, D. J. C. (1992a). Bayesian interpolation. *Neural Computation, 4*, 405–447.

Mackay, D. J. C. (1992b). A practical Bayesian framework for backpropagation networks. *Neural Computation, 4*, 448–472.

Mackay, D. J. C. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation, 11,* 1035–1068.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21,* 1087–1092.

Neal, R. M. (1996). *Bayesian learning for neural networks.* New York: Springer-Verlag.

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 355–368). Norwell, MA: Kluwer.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77,* 257–286.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B, 59,* 731–792.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society B, 49,* 223–239, 253–265.

Roberts, S. J., Husmeier, D., Rezek, I., & Penny, W. (1998). Bayesian approaches to gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1133–1142.

Roweis, S. T., & Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Computation, 11,* 305–345.

Sato, M. (2000). Convergence of on-line EM algorithm. In *Proc. of 7th International Conference on Neural Information Processing ICONIP-2000* Vol. 1 (pp. 476–481).

Sato, M., & Ishii, S. (2000). On-line EM algorithm for the normalized gaussian network. *Neural Computation, 12,* 407–432.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Computation, 11,* 443–482.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley.

Ueda, N. (1999). *Variational Bayesian learning with split and merge operations* (Tech. Rep. No. PRMU99-174, 67-74). Institute of Electronics, Information, and Communications Engineers. (in Japanese)

Ueda, N., Nakano, R., Ghahramani, Z., & Hinton, G. E. (1999). SMEM algorithm for mixture models. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems, 11* (pp. 599–605). Cambridge, MA: MIT Press.

Waterhouse, S., Mackay, D., & Robinson, T. (1996). Bayesian methods for mixture of experts. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems, 8* (pp. 351–357). Cambridge, MA: MIT Press.

Received December 28, 1999; accepted September 27, 2000.