TR-H-282

On-Line Model Selection Based on The Variational Bayes.

Masa-aki SATO (ATR-HIP/CREST)

2000.1.13

ATR人間情報通信研究所
〒619-0288　京都府相楽郡精華町光台2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Telephone: +81-774-95-1011
Fax　　　: +81-774-95-1008

On-line Model Selection Based on the Variational Bayes

Masa-aki Sato †‡
†ATR Human Information Processing Research Laboratories
‡CREST, Japan Science and Techonology Corporation
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
TEL: (+81) 774-95-1039 FAX: (+81) 774-95-1008
E-mail: masaaki@hip.atr.co.jp

Abstract

The Bayesian framework provides a principled way of the 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 by Attias (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 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 EM-like algorithm, whose convergence is guaranteed, is derived.

In this paper, 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. By combining the split and merge algorithm proposed by Ueda et al. (1999), the on-line VB algorithm provides a fully on-line learning method with a model selection mechanism. In preliminary experiments using synthetic data, the on-line VB method showed a faster and better performance than the batch VB method.

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 et al. 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 et al. 1995; Heckerman et al. 1995; Mackay 1992a; Mackay 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; Mackay 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 et al. 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 only valid 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 called Variational Bayes (VB) has been proposed by Attias (1999), (see also Neal & Hinton 1998; Waterhouse et al. 1996). 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 only requires a modest amount of computational time comparable with the EM algorithm. The BIC/MDL model selection criteria (Rissanen 1987; Schwartz 1978) 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 1999; Ueda 1999) have also been proposed by combining the VB method and the split and merge algorithm (Ueda et al. 1999).

In this paper, 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 on-line VB algorithm provides a fully on-line learning method with a model selection mechanism. It can be applied to real-time applications. In preliminary experiments using synthetic data, the on-line VB method showed a faster and better performance than the batch VB method. We also found that the introduction of a discount factor is crucial for a fast convergence of the on-line 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 1999), hidden Markov models (Rabiner 1989), mixture of Gaussian models (Roberts et al. 1998), mixture of factor analyzers (Ghahramani & Beal 1999), mixture of probabilistic principal component analyzers (Tipping & Bishop 1999), and others (Roweis & Ghahramani 1999; Titterington et al. 1985).

2 Variational Bayes Method

2.1 Exponential family model with hidden variables

In this section, we review the Variational Bayes (VB) method (Attias 1999) for the general Exponential Family models with Hidden variables (EFH models) (Amari 1985). An EFH model for an N-dimensional vector variable $\mathbf{x} = (x_1, ..., x_N)^T$ is defined by a probability distribution,

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

where $\mathbf{z} = (z_1, ..., z_M)^T$ denotes an M-dimensional vector hidden variable and $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^T$ denotes a set of model parameters called the natural parameter.[1] A set of sufficient statistics is denoted by $\mathbf{r}(\mathbf{x}, \mathbf{z}) = (r_1(\mathbf{x}, \mathbf{z}), ..., r_K(\mathbf{x}, \mathbf{z}))^T$. 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 $\Psi_\theta(\boldsymbol{\theta})$ is determined by

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

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 $(\mathbf{x}, \mathbf{z})$.

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

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

2.2 Evidence and free energy

The likelihood for a set of observed events $\mathbf{X}\{T\} = \{\mathbf{x}(t)|t = 1, ..., T\}$, is defined by

$$P(\mathbf{X}\{T\}|\boldsymbol{\theta}) = \prod_{t=1}^{T} P(\mathbf{x}(t)|\boldsymbol{\theta}). \tag{2.4}$$

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

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

inability to determine the best model structure. Bayesian approaches overcome these difficulties by averaging over the ensemble of models. The evidence for a data set $\mathbf{X}\{T\}$ is defined by

$$P(\mathbf{X}\{T\}) = \int d\mu(\boldsymbol{\theta}) P(\mathbf{X}\{T\}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta}), \tag{2.5}$$

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 (2.5) 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 (Attias 1999) introduces a trial posterior distribution for model parameters $\boldsymbol{\theta}$ and hidden variables $\mathbf{Z}\{T\} = \{\mathbf{z}(t)|t = 1, ..., T\}$, $Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})$. The free energy for a data set $\mathbf{X}\{T\}$ is defined by

$$F(\mathbf{X}\{T\}, Q) = \int d\mu(\boldsymbol{\theta}) d\mu(\mathbf{Z}\{T\}) Q(\boldsymbol{\theta}, \mathbf{Z}\{T\}) \log(P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta}) / Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})), \tag{2.6}$$

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

$$P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) = \prod_{t=1}^{T} P(\mathbf{x}(t), \mathbf{z}(t)|\boldsymbol{\theta}). \tag{2.7}$$

The log evidence is given by the maximum of the free energy (see Appendix A),

$$\log P(\mathbf{X}\{T\}) = \max_{Q} F(\mathbf{X}\{T\}, Q) \geq F(\mathbf{X}\{T\}, Q), \tag{2.8}$$

under the probability condition $\int d\mu(\boldsymbol{\theta}) d\mu(\mathbf{Z}\{T\}) Q(\boldsymbol{\theta}, \mathbf{Z}\{T\}) = 1$. The maximum solution is given by

$$Q(\boldsymbol{\theta}, \mathbf{Z}\{T\}) = P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta}) / P(\mathbf{X}\{T\}) = P(\boldsymbol{\theta}, \mathbf{Z}\{T\}|\mathbf{X}\{T\}), \tag{2.9}$$

which is the true posterior distribution for the model parameters and the hidden variables. The equation (2.8) implies that the lower bound for the log evidence can be evaluated by using some trial posterior distributions $Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})$.

2.3 Variational Bayes algorithm

In the VB method, trial posterior distributions are assumed to be factorized as

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

We also assume that the prior distribution $P_0(\boldsymbol{\theta})$ is given by the conjugate prior distribution[2] for the EFH model (2.1) :

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

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

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

The 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 $\log(P(\mathbf{X}\{T\}))$.

[2] It is also possible to use non-informative priors (Attias 1999).

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 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):

$$
\begin{aligned}
Q_z(\mathbf{Z}\{T\}) &= \prod_{t=1}^{T} Q_z(\mathbf{z}(t)), && (2.13)\\
Q_z(\mathbf{z}(t)) &= P(\mathbf{z}(t)|\mathbf{x}(t),\bar{\boldsymbol{\theta}}) = P(\mathbf{x}(t),\mathbf{z}(t)|\bar{\boldsymbol{\theta}})/P(\mathbf{x}(t)|\bar{\boldsymbol{\theta}}), && (2.14)\\
\bar{\boldsymbol{\theta}} &= \int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})\boldsymbol{\theta}. && (2.15)
\end{aligned}
$$

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

$$
\begin{aligned}
Q_\theta(\boldsymbol{\theta}) &= P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha},\gamma) = \exp\left[\gamma(\boldsymbol{\alpha}\cdot\boldsymbol{\theta} - \Psi_\theta(\boldsymbol{\theta})) - \Phi_\alpha(\boldsymbol{\alpha},\gamma)\right], && (2.16)\\
\gamma &= T + \gamma_0, && (2.17)\\
\boldsymbol{\alpha} &= \frac{1}{\gamma}\left[T\langle\mathbf{r}(\mathbf{x},\mathbf{z})\rangle_{\bar{\boldsymbol{\theta}}} + \boldsymbol{\alpha}_0\cdot\gamma_0\right], && (2.18)\\
\langle\mathbf{r}(\mathbf{x},\mathbf{z})\rangle_{\bar{\boldsymbol{\theta}}} &= \frac{1}{T}\sum_{t=1}^{T}\int d\mu(\mathbf{z}(t))P(\mathbf{z}(t)|\mathbf{x}(t),\bar{\boldsymbol{\theta}})\mathbf{r}(\mathbf{x}(t),\mathbf{z}(t)). && (2.19)
\end{aligned}
$$

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 (2.15) can be explicitly calculated as

$$
\begin{aligned}
\bar{\boldsymbol{\theta}} &= \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}, && (2.20)\\
\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}} &= \int d\mu(\boldsymbol{\theta})P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha},\gamma)\boldsymbol{\theta} = \frac{1}{\gamma}\frac{\partial\Phi_\alpha}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma). && (2.21)
\end{aligned}
$$

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 (2.21) and

$$
\langle\Psi_\theta(\boldsymbol{\theta})\rangle_{\boldsymbol{\alpha}} = \int d\boldsymbol{\theta}P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha},\gamma)\Psi_\theta(\boldsymbol{\theta}) = \frac{1}{\gamma}\frac{\partial\Phi_\alpha}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma)\cdot\boldsymbol{\alpha} - \frac{\partial\Phi_\alpha}{\partial\gamma}(\boldsymbol{\alpha},\gamma). \qquad (2.22)
$$

2.4 Parameterized free energy function

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

$$
Q_\theta(\boldsymbol{\theta}) = P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha},\gamma) = \exp\left[\gamma(\boldsymbol{\alpha}\cdot\boldsymbol{\theta} - \Psi_\theta(\boldsymbol{\theta})) - \Phi_\alpha(\boldsymbol{\alpha},\gamma)\right], \qquad (2.23)
$$

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

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

where γ, $\boldsymbol{\alpha}$, and $\bar{\boldsymbol{\theta}}$ are arbitrary variational parameters. By substituting this parameterized form into the definition of the free energy (2.6), one can get the parameterized free energy function :

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

where the ensemble averages of the parameters $\langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}}$ and $\langle \Psi_\theta(\boldsymbol{\theta}) \rangle_{\boldsymbol{\alpha}}$ are given by (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)

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

Since the coefficient matrix U is positive definite, the maximization condition (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)

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

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

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

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 the 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 (2.20). Subsequently, the expectation value of the sufficient statistics $\langle \mathbf{r}(\mathrm{x}, \mathrm{z})\rangle_{\bar{\boldsymbol{\theta}}}$ (2.19) is calculated by using the posterior distribution for the hidden variable $P(\mathrm{z}(t)|\mathrm{x}(t), \bar{\boldsymbol{\theta}})$ (2.14). In the VB M-step, the posterior hyperparameter $\boldsymbol{\alpha}$ is updated by using (2.18). Repeating this process, the free energy function (2.26) increases monotonically. This process continues until the free energy function converges.

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

$$\begin{aligned}
\Delta\bar{\boldsymbol{\theta}} &= \bar{\boldsymbol{\theta}}_{new} - \bar{\boldsymbol{\theta}} = \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}} - \bar{\boldsymbol{\theta}} \\
&= U^{-1}(\bar{\boldsymbol{\theta}})\frac{\partial F}{\partial\bar{\boldsymbol{\theta}}}(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \gamma), \qquad (2.32)\\
\Delta\boldsymbol{\alpha} &= \boldsymbol{\alpha}_{new} - \boldsymbol{\alpha} = \frac{1}{\gamma}(T\langle\mathbf{r}(\mathrm{x},\mathrm{z})\rangle_{\bar{\boldsymbol{\theta}}} + \gamma_0\boldsymbol{\alpha}_0 - \gamma\boldsymbol{\alpha}) \\
&= \frac{1}{\gamma^2}V^{-1}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma)\frac{\partial F}{\partial\boldsymbol{\alpha}}(\mathbf{X}\{T\}, \bar{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \gamma), \qquad (2.33)
\end{aligned}$$

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

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

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). This fact is proved for the first time in this paper.

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

$$\begin{aligned}
F(\mathbf{X}\{T\}, \boldsymbol{\alpha}) &= \sum_{t=1}^{T}\log P(\mathrm{x}|\bar{\boldsymbol{\theta}}) + T\Psi_\theta(\bar{\boldsymbol{\theta}}) \\
&\quad + (\Phi_\alpha(\boldsymbol{\alpha},\gamma) - \gamma\boldsymbol{\alpha}\cdot\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}) - (\Phi_\alpha(\boldsymbol{\alpha}_0,\gamma_0) - \gamma_0\boldsymbol{\alpha}_0\cdot\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}). \qquad (2.35)
\end{aligned}$$

The first term on the r.h.s. of (2.35) is the log-likelihood with the ensemble average of the parameters. The remaining terms represent the penalty due to the model complexity. This will become clear in a large sample limit as will be shown later.

2.5 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

$$\begin{aligned}
P(\mathrm{x}|\mathbf{X}\{T\}) &= \int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})P(\mathrm{x}|\boldsymbol{\theta}) \\
&= \int d\mu(\boldsymbol{\theta})\int d\mu(\mathrm{z})\exp\left[(\mathbf{r}(\mathrm{x},\mathrm{z}) + \gamma\boldsymbol{\alpha})\cdot\boldsymbol{\theta} + r_0(\mathrm{x},\mathrm{z})\right. \\
&\quad \left. -(1+\gamma)\Psi_\theta(\boldsymbol{\theta}) - \Phi_\alpha(\boldsymbol{\alpha},\gamma)\right]. \qquad (2.36)
\end{aligned}$$

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

$$\begin{aligned} P(\mathbf{x}|\mathbf{X}\{T\}) &= \int d\mu(\mathbf{z}) \exp\left[r_0(\mathbf{x},\mathbf{z}) + \Phi_\alpha(\hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}), \gamma+1) - \Phi_\alpha(\boldsymbol{\alpha},\gamma)\right], \\ \hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}) &= (\gamma\boldsymbol{\alpha} + \mathbf{r}(\mathbf{x},\mathbf{z}))/(1+\gamma). \end{aligned} \qquad (2.37)$$

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

2.6 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:

$$\begin{aligned} \exp\left[\Phi_\alpha(\boldsymbol{\alpha},\gamma)\right] &= \int d\mu(\theta) \exp\left[\gamma(\alpha \cdot \boldsymbol{\theta} - \Psi_\theta(\boldsymbol{\theta}))\right] \\ &\sim \exp\left[\gamma(\alpha \cdot \hat{\boldsymbol{\theta}} - \Psi_\theta(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\log\left|\gamma \frac{\partial^2 \Psi_\theta}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\right| + O(1/\gamma)\right], \end{aligned} \qquad (2.38)$$

where $\hat{\boldsymbol{\theta}}$ is the maximum of the exponent $(\boldsymbol{\alpha} \cdot \boldsymbol{\theta} - \Psi_\theta(\boldsymbol{\theta}))$, i.e.,

$$\frac{\partial \Psi_\theta}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\alpha}. \qquad (2.39)$$

Therefore, Φ_α can be approximated as

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

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

$$\begin{aligned} \bar{\boldsymbol{\theta}} &= \frac{1}{\gamma}\frac{\partial \Phi_\alpha}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma) \\ &\sim \frac{1}{\gamma}\frac{\partial}{\partial \boldsymbol{\alpha}}(\gamma(\boldsymbol{\alpha} \cdot \hat{\boldsymbol{\theta}} - \Psi_\theta(\hat{\boldsymbol{\theta}}))) = \hat{\boldsymbol{\theta}}. \end{aligned} \qquad (2.41)$$

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 Eq.(2.3)) in this limit. Furthermore, the 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 (2.40) and (2.41), the free energy becomes

$$\begin{aligned} F \sim\ & \sum_{t=1}^{T} \log P(\mathbf{x}|\bar{\boldsymbol{\theta}}) - \frac{K}{2}\log\gamma \\ & -\frac{1}{2}\log\left|\frac{\partial^2 \Psi_\theta}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}})\right| + \gamma_0(\boldsymbol{\alpha}_0 \cdot \bar{\boldsymbol{\theta}} - \Psi_\theta(\bar{\boldsymbol{\theta}})) - \Phi_\alpha(\boldsymbol{\alpha}_0,\gamma_0) + O(1/\gamma). \end{aligned} \qquad (2.42)$$

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:

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}) &\sim \boldsymbol{\alpha} + \frac{1}{\gamma}(\mathbf{r}(\mathbf{x},\mathbf{z}) - \boldsymbol{\alpha}) + O(1/\gamma^2), \\
\Phi_\alpha(\hat{\boldsymbol{\alpha}}(\mathbf{x},\mathbf{z}), r+1) &\sim \Phi_\alpha(\boldsymbol{\alpha},\gamma) + \frac{1}{\gamma}(\mathbf{r}(\mathbf{x},\mathbf{z}) - \boldsymbol{\alpha})\frac{\partial \Phi_\alpha}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma) + \frac{\partial \Phi_\alpha}{\partial \gamma}(\boldsymbol{\alpha},\gamma) \\
&\sim \Phi_\alpha(\boldsymbol{\alpha},\gamma) + \mathbf{r}(\mathbf{x},\mathbf{z}) \cdot \bar{\boldsymbol{\theta}} - \Psi_\theta(\bar{\boldsymbol{\theta}}).
\end{aligned}$$

3 On-line Variational Bayes method

3.1 Expectation value of the free energy

In this section, we derive an on-line version of the VB algorithm. The amount of data increases over time in the on-line learning. Therefore, it is desirable to calculate a 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 :

$$E\left[\log P(\mathbf{X}\{T\})\right]_\rho = \int d\mu(\mathbf{X}\{T\})\rho(\mathbf{X}\{T\}) \log\left(\int d\mu(\boldsymbol{\theta}) P(\mathbf{X}\{T\}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta})\right), \tag{3.1}$$

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

$$\begin{aligned}
E\left[F(\mathbf{X}\{T\}, Q_\theta, Q_z)\right]_\rho &= T\int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) E\left[\int d\mu(\mathbf{z}) Q_z(\mathbf{z}) \log\left(P(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})/Q_z(\mathbf{z})\right)\right]_\rho \\
&\quad + \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \log\left(P_0(\boldsymbol{\theta})/Q_\theta(\boldsymbol{\theta})\right).
\end{aligned} \tag{3.2}$$

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 (3.2) can be estimated by

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

Note that τ represents the actual amount of observed data, and it increases over time while T is fixed. In on-line learning, parameters are updated each time new data is observed. Therefore, $Q_z(\mathbf{z}(t))$ is parameterized with different parameter values for different data:

$$Q_z(\mathbf{z}(t)) = P(\mathbf{z}(t)|\mathbf{x}(t), \bar{\boldsymbol{\theta}}(t)). \tag{3.4}$$

A time dependent discount factor $\lambda(t)$ $(0 \leq \lambda(t) \leq 1,\ t = 2, 3, ...)$ is introduced, and a discounted free energy is defined by

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

$$
\begin{aligned}
& +\int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})\log\left(P_0(\boldsymbol{\theta})/Q_\theta(\boldsymbol{\theta})\right) \\
= \; & T\eta(\tau)\sum_{t=1}^{\tau}\left(\prod_{s=t+1}^{\tau}\lambda(s)\right)\Big[\log P(\mathbf{x}(t)|\bar{\boldsymbol{\theta}}(t)) \\
& +\int d\mu(\mathbf{z}(t))P(\mathbf{z}(t)|\mathbf{x}(t),\bar{\boldsymbol{\theta}}(t))\mathbf{r}(\mathbf{x}(t),\mathbf{z}(t))\cdot(\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}-\bar{\boldsymbol{\theta}}(t))+\Psi_\theta(\bar{\boldsymbol{\theta}}(t))\Big] \\
& +(\gamma_0\boldsymbol{\alpha}_0-\gamma\boldsymbol{\alpha})\cdot\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}+\Phi_\alpha(\boldsymbol{\alpha},\gamma)-\Phi_\alpha(\boldsymbol{\alpha}_0,\gamma_0), \qquad (3.5)
\end{aligned}
$$

where $\gamma = (T + \gamma_0)$ is used and $\eta(\tau)$ represents a normalization constant:

$$\eta(\tau) = \left[\sum_{t=1}^{\tau}\left(\prod_{s=t+1}^{\tau}\lambda(s)\right)\right]^{-1}. \qquad (3.6)$$

3.2 On-line variational Bayes algorithm

The on-line VB algorithm can be derived from the successive maximization of the discounted free energy (3.5). The calculations can be done in the same way as in Sec. 2.4. Let us assume that $\bar{\boldsymbol{\theta}}\{\tau-1\} = \{\bar{\boldsymbol{\theta}}(t)|t = 1,...,\tau-1\}$ and $\boldsymbol{\alpha}(\tau-1)$ have been determined for an observed data set $\mathbf{X}\{\tau-1\} = \{\mathbf{x}(t)|t = 1,...,\tau-1\}$. With new observed data $\mathbf{x}(\tau)$, the discounted free energy (3.5) is maximized with respect to $\bar{\boldsymbol{\theta}}(\tau)$:

$$\partial F^\lambda(\mathbf{X}\{\tau\},\bar{\boldsymbol{\theta}}\{\tau\},\boldsymbol{\alpha}(\tau-1),T)/\partial\bar{\boldsymbol{\theta}}(\tau) = 0. \qquad (3.7)$$

The solution is given by

$$\bar{\boldsymbol{\theta}}(\tau) = \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}(\tau-1)} = \frac{1}{\gamma}\frac{\partial\Phi_\alpha}{\partial\boldsymbol{\alpha}}(\boldsymbol{\alpha}(\tau-1),\gamma). \qquad (3.8)$$

In the next step, the discounted free energy is maximized with respect to $\boldsymbol{\alpha}$, while $\bar{\boldsymbol{\theta}}\{\tau\}$ is fixed:

$$\partial F^\lambda(\mathbf{X}\{\tau\},\bar{\boldsymbol{\theta}}\{\tau\},\boldsymbol{\alpha},T)/\partial\boldsymbol{\alpha} = 0. \qquad (3.9)$$

The solution is given by

$$\gamma\boldsymbol{\alpha}(\tau) = T\langle\!\langle\mathbf{r}(\mathbf{x},\mathbf{z})\rangle\!\rangle(\tau) + \gamma_0\boldsymbol{\alpha}_0, \qquad (3.10)$$

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

$$\langle\!\langle\mathbf{r}(\mathbf{x},\mathbf{z})\rangle\!\rangle(\tau) = \eta(\tau)\sum_{t=1}^{\tau}\left(\prod_{s=t+1}^{\tau}\lambda(s)\right)\int d\mu(\mathbf{z}(t))P(\mathbf{z}(t)|\mathbf{x}(t),\bar{\boldsymbol{\theta}}(t))\mathbf{r}(\mathbf{x}(t),\mathbf{z}(t)). \qquad (3.11)$$

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

$$
\begin{aligned}
\langle\!\langle\mathbf{r}(\mathbf{x},\mathbf{z})\rangle\!\rangle(\tau) &= (1-\eta(\tau))\langle\!\langle\mathbf{r}(\mathbf{x},\mathbf{z})\rangle\!\rangle(\tau-1) \\
&\quad +\eta(\tau)\int d\mu(\mathbf{z}(\tau))P(\mathbf{z}(\tau)|\mathbf{x}(\tau),\bar{\boldsymbol{\theta}}(\tau))\mathbf{r}(\mathbf{x}(\tau),\mathbf{z}(\tau)), \qquad (3.12) \\
\eta(\tau) &= (1+\lambda(\tau)/\eta(\tau-1))^{-1}. \qquad (3.13)
\end{aligned}
$$

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

$$
\begin{aligned}
\Delta\boldsymbol{\alpha}(\tau) &= \boldsymbol{\alpha}(\tau)-\boldsymbol{\alpha}(\tau-1) \\
&= \frac{1}{\gamma}\eta(\tau)\left[T\int d\mu(\mathbf{z}(\tau))P(\mathbf{z}(\tau)|\mathbf{x}(\tau),\bar{\boldsymbol{\theta}}(\tau))\mathbf{r}(\mathbf{x}(\tau),\mathbf{z}(\tau))+\gamma_0\boldsymbol{\alpha}_0-\gamma\boldsymbol{\alpha}(\tau-1)\right]. \qquad (3.14)
\end{aligned}
$$

The on-line VB algorithm is summarized as follows. In the VB E-step, the ensemble average of parameter $\bar{\boldsymbol{\theta}}(\tau)$ is determined by equation (3.8). Using this value, one calculates the expectation value for the sufficient statistics,

$$E_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x}(\tau),\mathbf{z}(\tau))|\bar{\boldsymbol{\theta}}(\tau)\right] = \int d\mu(\mathbf{z}(\tau))P(\mathbf{z}(\tau)|\mathbf{x}(\tau),\bar{\boldsymbol{\theta}}(\tau))\mathbf{r}(\mathbf{x}(\tau),\mathbf{z}(\tau)). \tag{3.15}$$

The posterior hyperparameter is updated by (3.14) in the VB M-step. By combining the VB E-step (3.8) and VB M-step equations (3.14), one can get the recursive update equation for $\boldsymbol{\alpha}(\tau)$:

$$\Delta\boldsymbol{\alpha}(\tau) = \frac{1}{\gamma}\eta(\tau)\left(TE_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x}(\tau),\mathbf{z}(\tau))|\langle\theta\rangle_{\boldsymbol{\alpha}(\tau-1)}\right] + \gamma_0\boldsymbol{\alpha}_0 - \gamma\boldsymbol{\alpha}(\tau-1)\right). \tag{3.16}$$

3.3 Stochastic approximation

Unlike the VB algorithm, the discounted free energy in the on-line 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 on-line VB algorithm can be considered as a stochastic approximation (Kushner & Yin 1997) for finding the maximum of the expected free energy defined in (3.2), which gives a lower bound for the expected log evidence defined in (3.1). The expected free energy (3.2), in which the maximization with respect to Q_z has been performed, can be written as

$$\max_{Q_z} E\left[F(\mathbf{X}\{T\},Q_\theta,Q_z)\right]_\rho = E\left[F_M(\mathbf{x},\boldsymbol{\alpha},T)\right]_\rho, \tag{3.17}$$

where

$$\begin{aligned} F_M(\mathbf{x},\boldsymbol{\alpha},T) &= T\int d\mu(\mathbf{z})P(\mathbf{z}|\mathbf{x},\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}})\int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})\log\left(P(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})/P(\mathbf{z}|\mathbf{x},\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}})\right) \\ &\quad + \int d\mu(\boldsymbol{\theta})Q_\theta(\boldsymbol{\theta})\log\left(P_0(\boldsymbol{\theta})/Q_\theta(\boldsymbol{\theta})\right) \\ &= T\log P(\mathbf{x}|\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}) - \left[(\gamma\boldsymbol{\alpha}-\gamma_0\boldsymbol{\alpha}_0)\cdot\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}} - T\Psi_\theta(\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}})\right] \\ &\quad + \Phi_\alpha(\boldsymbol{\alpha},\gamma) - \Phi_\alpha(\boldsymbol{\alpha}_0,\gamma_0). \end{aligned} \tag{3.18}$$

The gradient of F_M is calculated as

$$\frac{\partial F_M}{\partial\boldsymbol{\alpha}}(\mathbf{x},\boldsymbol{\alpha},T) = \gamma V_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\alpha},\gamma)\cdot\left[TE_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x},\mathbf{z})|\langle\boldsymbol{\theta}\rangle_{\boldsymbol{\alpha}}\right] + \gamma_0\boldsymbol{\alpha}_0 - \gamma\boldsymbol{\alpha}\right], \tag{3.19}$$

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 (2.31). Consequently, the on-line VB algorithm (3.16) can be written as

$$\Delta\boldsymbol{\alpha}(\tau) = \frac{1}{\gamma^2}\eta(\tau)V^{-1}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\alpha}(\tau-1),\gamma)\cdot\frac{\partial F_M}{\partial\boldsymbol{\alpha}}(\mathbf{x}(\tau),\boldsymbol{\alpha}(\tau-1),T). \tag{3.20}$$

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

$$\sum_{t=1}^{\infty}\eta(\tau) = \infty \quad \text{and} \quad \sum_{t=1}^{\infty}\eta^2(\tau) < \infty, \tag{3.21}$$

the on-line VB algorithm (3.16) defines the stochastic approximation for finding the maximum of the expected free energy (3.2).

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

$$\eta(\tau) = 1/\tau. \tag{3.22}$$

This satisfies the stochastic approximation condition (3.21). However, the learning speed becomes very slow if this schedule is adopted (see Sec. 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 (3.11). The introduction of the discount factor is crucial for fast convergence.

As in the on-line EM algorithm proposed in our previous paper (Sato 1999), we employ the following discount schedule

$$1 - \lambda(\tau) = \frac{1}{(\tau - 2)\kappa + \tau_0}, \tag{3.23}$$

which can be calculated recursively:

$$\lambda(\tau) = 1 - \frac{1 - \lambda(\tau - 1)}{1 + \kappa(1 - \lambda(\tau - 1))}. \tag{3.24}$$

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

$$\eta(\tau) \underset{\tau \to \infty}{\longrightarrow} \left(\frac{\kappa + 1}{\kappa} \right) \frac{1}{\tau}, \tag{3.25}$$

so that the stochastic approximation condition (3.21) is satisfied. The constants appearing in (3.23) have clear physical meanings. τ_0 represents how many samples contribute to the discounted average for the sufficient statistics (3.11) in the early stage of learning. κ controls the asymptotic decreasing ratio for the effective learning constant $\eta(\tau)$ as in (3.25). The values of τ_0 and κ control the learning speeds in the early and later stages of learning, respectively.

4 On-line 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.

In this paper, we adopt sequential model selection procedures (Ghahramani & Beal 1999; Ueda 1999). 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 et al. 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.

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 1999; 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. 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 1999; Ueda 1999).

By combining the sequential model selection procedure with the on-line VB learning method, a fully on-line 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 on-line VB method, we considered model selection problems for two-dimensional Mixture of Gaussian (MG) models (see Appendix C). We borrowed two tasks from a paper by Roberts et al. (1998). A 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 $(-2,-\sqrt{12})$ (Fig. 1A). The Gaussians had the same isotropic variance $\sigma^2=(1.2)^2$. In addition, a data set 'B' consisting of 1000 points was generated from a mixture of four Gaussians (Fig. 1B). In this case, they were paired such that each pair had a common center, i.e., $\mathbf{m}_1=\mathbf{m}_2=(2,\sqrt{12})$ and $\mathbf{m}_3=\mathbf{m}_4=(-2,-\sqrt{12})$, but had different variances, i.e., $\sigma_1^2=\sigma_3^2=(1.0)^2$ 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 prepared. The VB method was applied to each model and the maximum free energy was calculated. We used a nearly non-informative prior for all cases, i.e., $\gamma_0=0.01$. The on-line VB method used the discount schedule (3.23) with $\tau_0=100$ and $\kappa=0.01$ for all cases.

The learning speed was measured according to epoch numbers. In one epoch, all training data were supplied to each VB method once. The on-line 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 Fig. 2. Both the batch VB method and the on-line VB method gave the highest free energy for the true model consisting of four units. The on-line VB method showed a faster and better performance than the batch VB method, especially for large amounts of data (Fig. 2). The reason for this performance difference can be considered as follows. In the on-line 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 on-line VB method can find the optimal solution within one epoch, as shown in Fig. 2D.

The on-line VB method without the discount factor showed a poor performance and slow convergence for all cases. This result implies that the introduction of the discount factor is crucial for a good performance of the on-line VB method, as pointed out in Sec. 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 experiments, the sequential model selection procedure described in Sec. 4 was tested using two initial model configurations consisting of two units and ten units. When the model structure was changed, the discount factor and the effective learning constant in the on-line VB method were reset as $(1-\lambda(\tau)) = 0.01$ and $\eta(\tau) = 0.01$. The on-line VB method was able to find the best model in all cases (Fig. 3). It should be noted that the VB method sometimes increased the free energy while decreasing the data likelihood (Figs. 3 and 4). This was achieved as a result of the decrease in the model complexity. The batch VB method also found the best model except for one case, in which the batch VB method got stuck in a local maximum (Fig. 4C).

In summary, the on-line VB method showed a better and faster performance than the batch VB method in all cases.

6 Conclusion

In this paper, we derived an on-line version of the Variational Bayes (VB) algorithm and proved its convergence by showing that it is a stochastic approximation for finding the maximum of the free energy. A fully on-line learning method with a model selection mechanism was also proposed based on the on-line VB method together with a sequential model selection procedure. This method can be applied to the model selection problem in dynamic environments.

In this paper, we considered the Bayes model without hierarchy. The current method can be easily extended to the hierarchical Bayes model (see Appendix D).

In preliminary experiments using synthetic data, the on-line VB method showed a faster and better performance than the batch VB method. A detailed study on the performance of the on-line VB method will be published in a forthcoming paper. It is also remained for future study to find better sequential model selection procedure.

References

[1] Amari, S. (1985),*Differential geometrical method in statistics*, Springer Lecture Notes in Statistics,**28**, Springer.

[2] Amari, S. 1998, Natural Gradient Works Efficiently in Learning, *Neural Computation*, **10**, 251-276.

[3] Attias, H. 1999. Inferring parameters and structure of latent variable models by variational Bayes. *Proc. 15th Conference on Uncertainty in Artificial Intelligence.*

[4] Bishop, C.M. 1995. *Neural Networks for Pattern Recognition*, Oxford University Press. New York.

[5] Chickering, D.M. and Heckerman, D. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, **29**, 181-212.

[6] Cooper, G. and Herskovitz, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309-347.

[7] 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.

[8] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 1995. *Bayesian Data Analysis*, Chapman & Hall.

[9] Ghahramani, Z. and Beal, M.J. 1999. Variational inference for Bayesian mixture of factor analysers. To appear in *Advances in Neural Information Processing Systems 12.*

[10] Heckerman, D., Geiger, D., and Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, **20**, 197-243.

[11] Kushner, H. J., & Yin, G. G. 1997. *Stochastic Approximation Algorithms and Applications*, New York: Springer-Verlag.

[12] Mackay, D.J.C. 1992a. Bayesian interpolation. *Neural Computation*, **4**, 405-447.

[13] Mackay, D.J.C. 1992b. A practical Bayesian framework for backpropagation networks. *Neural Computation*, **4**, 448-472.

[14] Mackay, D.J.C. 1999. Comparison of Approximate Methods for Handling Hyperparameters *Neural Computation*, **11**, 1035-1068.

[15] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equations of state calculations by fast computing mathines. *Journal of Chemical Physics*, **21**, 1087-1092.

[16] Neal, R. M. 1996. *Bayesian learning for neural networks*, Springer-Verlag.

[17] 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 Academic Press.

[18] Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected apprications in speech recognition. In *Proceedings of the IEEE*, **77**, 257-286.

[19] Richardson, S. and 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.

[20] Rissanen, J. 1987. Stochastic complexity. *Journal of the Royal Statistical Society B*, **49**, 223-239 and 253-265.

[21] Roberts, S.J., Husmeier, D., Rezek, I., and Penny, W. 1998. Bayesian approaches to Gaussian mixture modeling. *IEEE PAMI*, **20**, 1133-1142.

[22] Roweis, S. T. and Ghahramani, Z. 1999. A unifying review of linear Gaussian models. *Neural Computation*, **11**, 305-345.

[23] Sato, M. 1999. Fast Learning of On-line EM Algorithm.
Technical Report (http://www.hip.atr.co.jp/masaaki/).

[24] Sato, M., & Ishii, S. 1999. On-line EM Algorithm for the Normalized Gaussian Network.
In press, *Neural Computation*, **12**.

[25] Schwartz, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

[26] Titterington, D.M., Smith, A.F.M., and Makov, U.E. 1985. *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester.

[27] Tipping, M. E. and Bishop,C. M. 1999. Mixtures of probabilistic principal component analyzers. *Neural Computation*, **11**, 443-482.

[28] Ueda, N., Nakano, R., Ghahramani, Z., and Hinton, G. E., 1999. SMEM Algorithm for Mixture Models. In M. S. Kearns, S. A. Solla, and D. A. Cohn, (Eds.), *Advances in Neural Information Processing Systems 11*, pp. 599-605, Cambridge, MA: MIT Press.

[29] Ueda, N. 1999. Variational Bayesian learning with split and merge operations. *Technical Report of IEICE*, **PRMU99-174**, 67-74, (in Japanese).

[30] Waterhouse, S., Mackay, D., and Robinson, T. 1996. Bayesian methods for mixture of experts. In Touretzky, D. S., Mozer, M. C. and Hasselmo, M. E., (Eds.), *Advances in Neural Information Processing Systems 8*, pp. 351-357. Cambridge, MA: MIT Press.

Appendix A

The free energy (2.6) can be maximized with respect to $Q(\boldsymbol{\theta}, \mathbf{Z}\{T\})$ by using the following theorem:

- 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

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

The theorem can be proven with the help of the Lagrange multiplier method. The VB equations, (2.13)-(2.19), can also be proven by using this theorem and the following relations:

$$\begin{aligned}
F &= \int d\mu(\boldsymbol{\theta}) Q_\theta(\boldsymbol{\theta}) \left[T(\langle \mathbf{r}(\mathbf{x},\mathbf{z})\rangle_{Q_z} \cdot \boldsymbol{\theta} - \Psi_\theta(\boldsymbol{\theta})) + \log P_0(\boldsymbol{\theta}) - \log Q_\theta(\boldsymbol{\theta})\right] \\
&+ \quad Q_\theta(\boldsymbol{\theta})\text{-independent terms} \\
&= \sum_{t=1}^{T} \int d\mu(\mathbf{z}(t)) Q_z(\mathbf{z}(t)) \left[\mathbf{r}(\mathbf{x}(t),\mathbf{z}(t)) \cdot \langle \boldsymbol{\theta}\rangle_{Q_\theta} + r_0(\mathbf{x}(t),\mathbf{z}(t)) - \log Q_z(\mathbf{z}(t))\right] \\
&+ \quad Q_z(\mathbf{z}(t))\text{-independent terms,}
\end{aligned}$$

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 on the derivative of the parameterized free energy (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,

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

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

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

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

The equations (2.30) and (2.31) can be derived by using the above and the following equations.

$$\begin{aligned}
\frac{1}{\gamma^2}\frac{\partial^2 \Phi_\alpha}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} &= \frac{1}{\gamma^2}\left\langle \left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}^T}\right)\right\rangle_{\boldsymbol{\alpha}}, \\
\frac{1}{\gamma}\frac{\partial^2 \Phi_\alpha}{\partial \boldsymbol{\alpha}\partial\gamma} - \frac{1}{\gamma^2}\frac{\partial \Phi_\alpha}{\partial \boldsymbol{\alpha}} &= \frac{1}{\gamma}\left\langle \left(\frac{\partial \log P_\alpha}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial \log P_\alpha}{\partial \gamma}\right)\right\rangle_{\boldsymbol{\alpha}}, \\
\frac{\partial^2 \Phi_\alpha}{\partial\gamma\partial\gamma} &= \left\langle \left(\frac{\partial \log P_\alpha}{\partial \gamma}\right)\left(\frac{\partial \log P_\alpha}{\partial \gamma}\right)\right\rangle_{\boldsymbol{\alpha}}.
\end{aligned}$$

Appendix C

The VB algorithm for the Mixture of Gaussian model is briefly explained in this Appendix.[3] We first explain more general mixture models, i.e., the Mixture of Exponential Family (MEF) models. The probability distribution for the i-th unit in the MEF model is defined by

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

The conjugate distribution for (A.1) is given by

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

The probability distribution for the MEF model is, then, defined by

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

where the hidden variable $\mathbf{z} = \{z_i | i = 1, \cdots, M\}$ is an indicator variable, i.e., $z_i = 0$ 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 $\mathbf{g} = \{g_i | i = 1, \cdots, M\}$ satisfies the constraint $\sum_{i=1}^{M} g_i = 1$, which is automatically satisfied by the expression, $g_i = e^{\varphi_i} / (\sum_{j=1}^{M} e^{\varphi_j})$.

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

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

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

The conjugate distribution for the MEF model (A.3) is given by the product of the Dirichlet distribution and the conjugate distribution for each unit:

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

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

[3] Notation in this Appendix are slightly different from those in the text.

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

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

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

The VB M-step equation is given by

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

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

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

where $\gamma(0)$ and $\{\nu_i(0), \boldsymbol{\alpha}_i(0) | i = 1, \cdots, M\}$ are the prior hyperparameters of the prior parameter distribution. $\langle \cdot \rangle_{\bar{\boldsymbol{\theta}}}$ denotes the expectation value (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

$$\begin{aligned} F = & \sum_{i=1}^{M} \Big[T \langle z_i \log P(\mathbf{x}|\bar{\boldsymbol{\omega}}, \bar{\boldsymbol{\theta}}) \rangle_{\bar{\boldsymbol{\theta}}} - T \langle z_i \log P(\mathbf{x}, \mathbf{z}|\bar{\boldsymbol{\omega}}, \bar{\boldsymbol{\theta}}) \rangle_{\bar{\boldsymbol{\theta}}} \\ & + \log \Gamma(\gamma \nu_i + 1) - \nu_i \log \Gamma(\gamma + M) + \Phi_i(\boldsymbol{\alpha}_i, \gamma \nu_i) - \Phi_i(\boldsymbol{\alpha}_i(0), \gamma(0)/M) \Big], \end{aligned} \tag{A.13}$$

where we assumed $\nu_i(0) = 1/M$.

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

$$\begin{aligned} P(\mathbf{x}|\mathbf{m}_i, \Sigma_i, i) &= (2\pi)^{-N/2} |\Sigma_i|^{1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^T \Sigma_i (\mathbf{x} - \mathbf{m}_i) \right] \\ &= \exp\left[-\frac{1}{2} \mathbf{x}^T \Sigma_i \mathbf{x} + \mathbf{x}^T \Sigma_i \mathbf{m}_i - \Psi_i(\mathbf{m}_i, \Sigma_i) \right], \end{aligned} \tag{A.14}$$

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

where $\mathbf{m}_i$ and Σ_i denote the center and the inverse covariance matrix of the i-th 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 (A.14) is given by the normal-Wishart distribution (Gelman et al. 1995),

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

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

The natural parameter of the conjugate distribution (A.16) is given by

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

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 (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 $\boldsymbol{\alpha}_0$:

$$P(\mathbf{X}\{T\}) = \int d\mu(\theta) d\mu(\boldsymbol{\alpha}_0) P(\mathbf{X}\{T\}|\boldsymbol{\theta}) P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}_0, \gamma_0) P_0(\boldsymbol{\alpha}_0), \tag{A.19}$$

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

$$\begin{aligned} F(\mathbf{X}\{T\}, Q) \quad = \quad & \int d\mu(\boldsymbol{\theta}) d\mu(\boldsymbol{\alpha}_0) d\mu(\mathbf{Z}\{T\}) Q(\boldsymbol{\theta}, \boldsymbol{\alpha}_0, \mathbf{Z}\{T\}) \\ & \times \log \left(P(\mathbf{X}\{T\}, \mathbf{Z}\{T\}|\boldsymbol{\theta}) P_\alpha(\boldsymbol{\theta}|\boldsymbol{\alpha}_0, \gamma_0) P_0(\boldsymbol{\alpha}_0) / Q(\boldsymbol{\theta}, \boldsymbol{\alpha}_0, \mathbf{Z}\{T\}) \right). \end{aligned} \tag{A.20}$$

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

$$P_0(\boldsymbol{\alpha}_0) = \exp\left[b_0(\mathbf{a}_0 \boldsymbol{\alpha}_0 \gamma_0 - \Phi_\alpha(\boldsymbol{\alpha}_0, \gamma_0)) - \Phi_a(\mathbf{a}_0, b_0) \right] \tag{A.21}$$

and the factorization for the trial posterior distribution,

$$Q(\boldsymbol{\theta}, \boldsymbol{\alpha}_0, \mathbf{Z}\{T\}) = Q_\theta(\boldsymbol{\theta}) Q_\alpha(\boldsymbol{\alpha}_0) Q_z(\mathbf{Z}\{T\}). \tag{A.22}$$

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

$$\bar{\boldsymbol{\theta}} = \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}}. \tag{A.23}$$

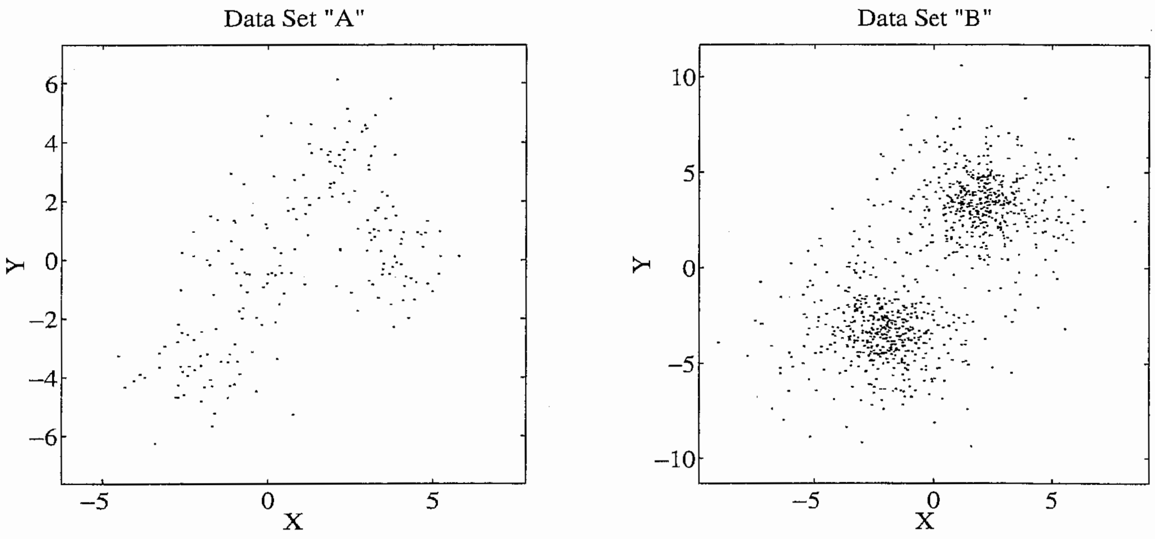
The posterior hyperparameter $\boldsymbol{\alpha}$ is calculated in the VB M-step;

$$\begin{aligned} \gamma \boldsymbol{\alpha} \quad &= \quad T \langle \mathbf{r}(\mathbf{x}, \mathbf{z}) \rangle_{\bar{\boldsymbol{\theta}}} + \gamma_0 \langle \boldsymbol{\alpha}_0 \rangle_{\mathbf{a}}, & \text{(A.24)} \\ \gamma_0 \langle \boldsymbol{\alpha}_0 \rangle_{\mathbf{a}} \quad &= \quad \gamma_0 \int d\mu(\boldsymbol{\alpha}_0) Q_\alpha(\boldsymbol{\alpha}_0) \boldsymbol{\alpha}_0 = \frac{1}{b} \frac{\partial \Phi_a}{\partial \mathbf{a}}(\mathbf{a}, b), & \text{(A.25)} \end{aligned}$$

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

$$\begin{aligned} \mathbf{a} \quad &= \quad \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} + \mathbf{a}_0, & \text{(A.26)} \\ b \quad &= \quad b_0 + 1, & \text{(A.27)} \\ \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\alpha}} \quad &= \quad \frac{1}{\gamma} \frac{\partial \Phi_\alpha}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}, \gamma). & \text{(A.28)} \end{aligned}$$

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



(A)

(B)

Figure 1

Fig.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.

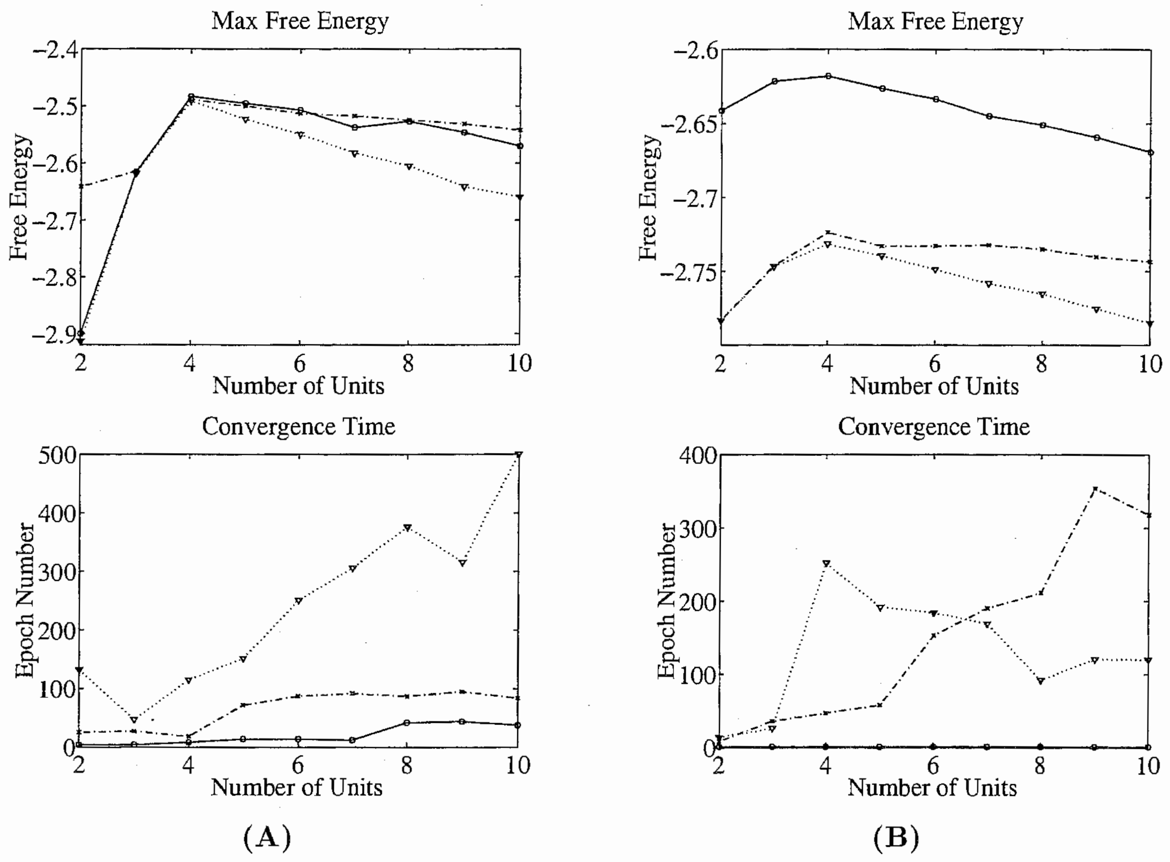


Figure 2

Fig.2 Maximum free energies 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 crosses), on-line VB (solid line with circles), and on-line VB without discount factor (dotted line with triangles). Abscissa denotes number of Gaussian units in trained models. (A): Results for data set 'A'. (B): Results for data set 'B'.

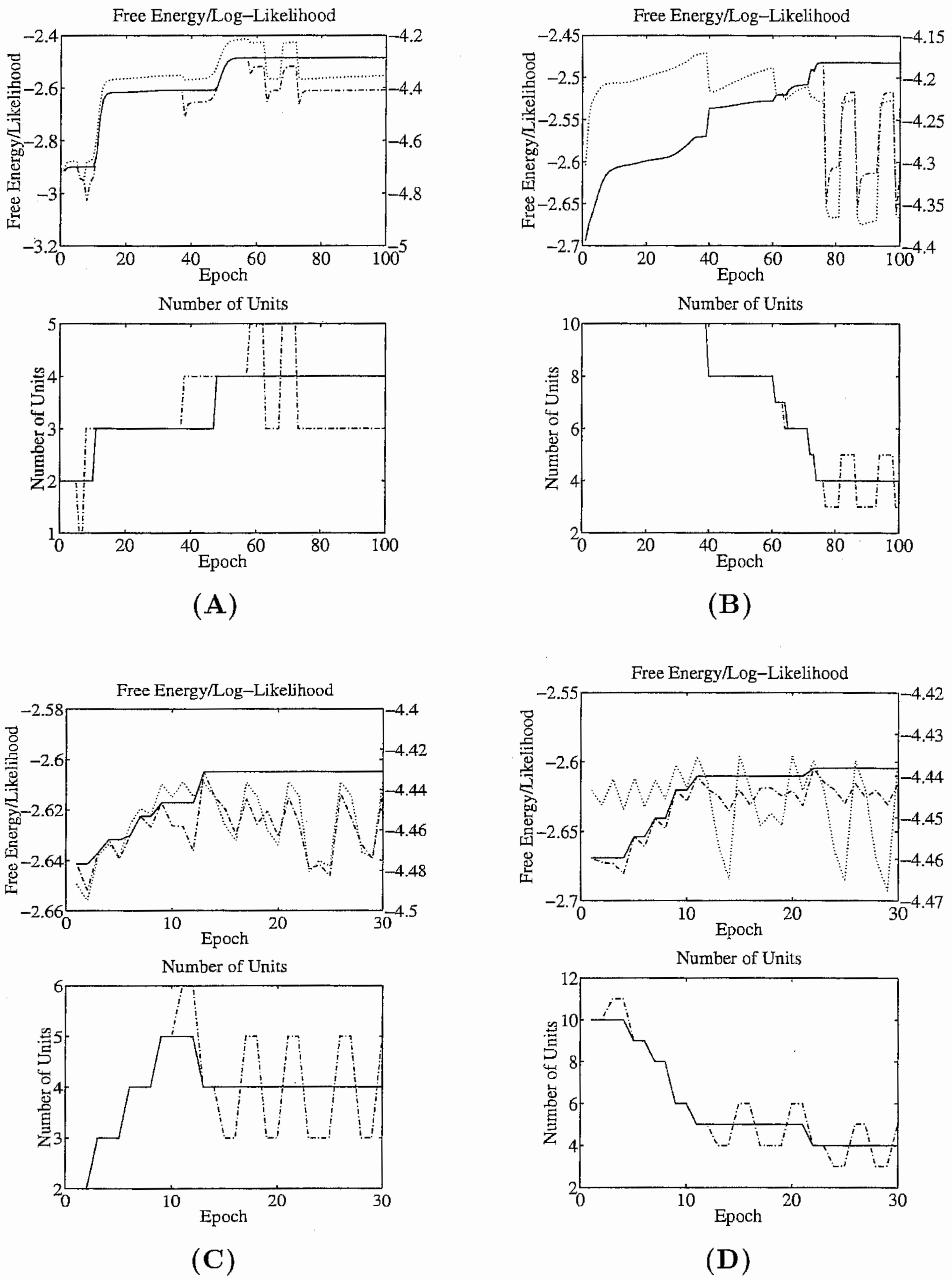


Figure 3

Fig. 3 On-line model selection processes using on-line VB method. Free energies and number of units for best model (solid line) and current model (dash-dotted line) are shown. Log-likelihood for current model (dotted line) is also shown. (A) and (B): Results for data set 'A'. (C) and (D): Results for data set 'B'.

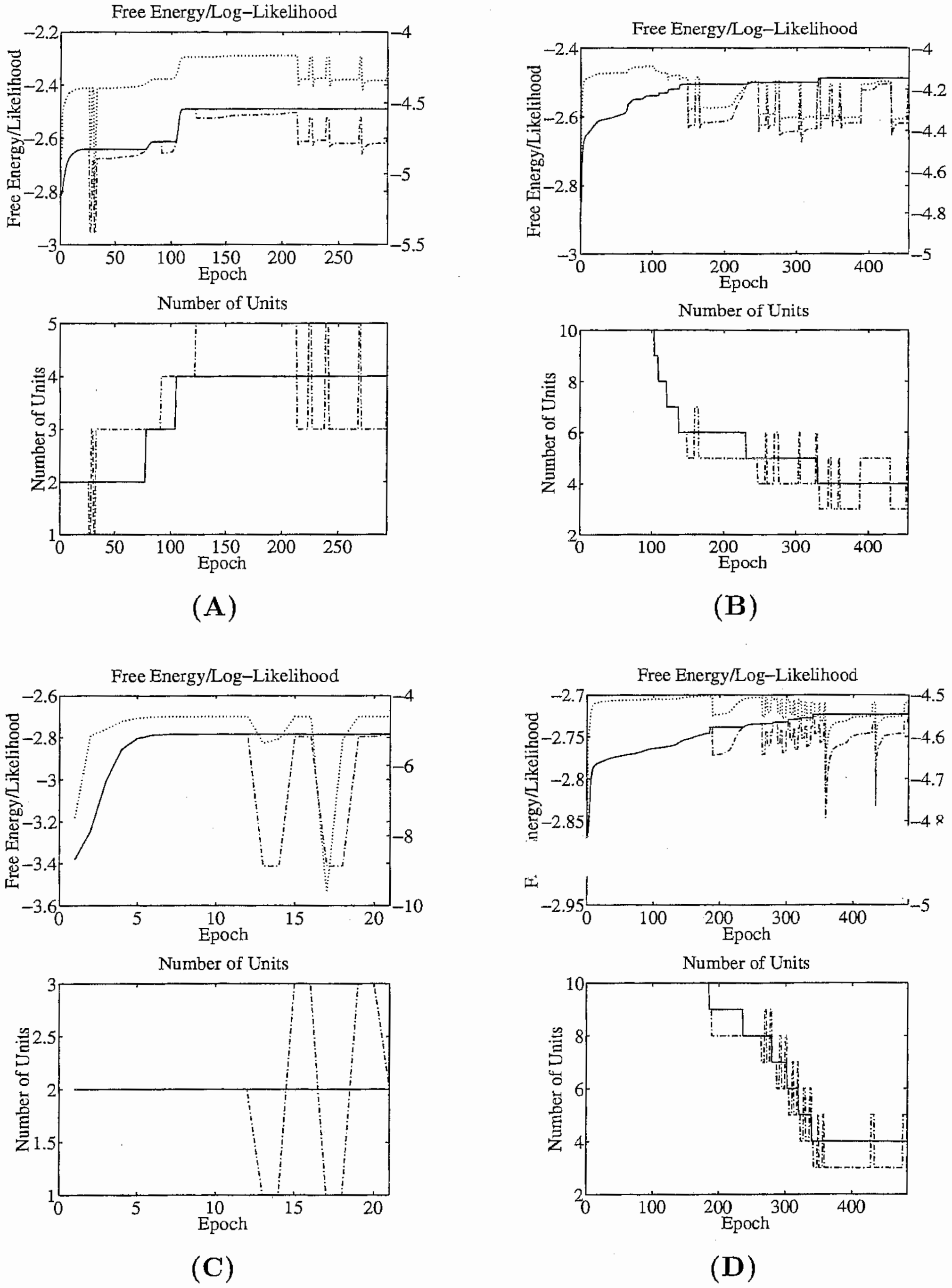


Figure 4

Fig. 4 Sequential model selection processes using the batch VB method. Free energies and number of units for best model (solid line) and current model (dash-dotted line) are shown. Log-likelihood for current model (dotted line) is also shown. (A) and (B): Results for data set 'A'. (C) and (D): Results for data set 'B'.