
PERGAMON

Neural Networks 15 (2002) 1223–1241

Neural Networks

www.elsevier.com/locate/neunet

Bayesian model search for mixture models based on optimizing variational bounds

Naonori Ueda[a,*], Zoubin Ghahramani[b]

[a]NTT Communication Science Laboratories, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
[b]Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London WC1N 3AR, UK

Received 31 October 2001; revised 15 April 2002; accepted 15 April 2002

Abstract

When learning a mixture model, we suffer from the local optima and model structure determination problems. In this paper, we present a method for simultaneously solving these problems based on the variational Bayesian (VB) framework. First, in the VB framework, we derive an objective function that can simultaneously optimize both model parameter distributions and model structure. Next, focusing on mixture models, we present a deterministic algorithm to approximately optimize the objective function by using the idea of the split and merge operations which we previously proposed within the maximum likelihood framework. Then, we apply the method to mixture of expers (MoE) models to experimentally show that the proposed method can find the optimal number of experts of a MoE while avoiding local maxima.

Keywords: Variational Bayes; Bayesian model search; Mixture models; Mixture of experts models; EM algorithm

1. Introduction

The aim of statistical learning is to estimate a generative model behind observed data. Recently, there has been an emphasis on using mixture models to analyze complex phenomena. However, when learning a mixture model, we are confronted by two difficulties in practice. The first is the local optima problem. That is, a learning algorithm can be trapped in poor local optima near an initial parameter value. The second is the problem of determining an appropriate model structure. If the model structure is too complicated, then learning results tends to overfit the noisy training data. Solving these problems is, therefore, of considerable importance for obtaining accurate predictions for unknown data.

As for the first problem mentioned above, we recently proposed the split and merge Expectation Maximization (SMEM) algorithm for mixture models within the maximum likelihood framework by simultaneously splitting and merging model components (Ueda, Nakano, Ghaharamani, & Hinton, 1999, 2000). The model structure (i.e. the number of mixture components), however, was fixed there since the ML framework suffers from the fact that the likelihood in general increases as the model structure becomes complex. The SMEM algorithm, therefore, cannot find the optimal model structure since the likelihood function is used as its objective function.

Within the ML framework, for linear models, we can utilize well known information criteria such as AIC (Akaike, 1974) and TIC (Takeuchi, 1983) to determine the model structure. These criteria are based on asymptotic normality assumption. Therefore, when the number of training data is small, these criteria are not valid due to the failure of the assumption. Computationally heavy cross-validation procedures (Stone, 1974) also becomes unreliable in the small sample case.

Bayesian approach, on the other hand, can theoretically determine the model structure through a posterior distribution over the model structure, conditional on the training data. Moreover, the Bayesian approach yields a posterior distribution over the model parameters and provides not a single prediction as in the ML approach, but a predictive distribution. The Bayesian approach, therefore, can mitigate the over-fitting problem.

In the Bayesian approach, however, we have to compute expectations which include difficult integrals. Recently, Waterhouse, MacKay, and Robinson (1995) proposed the Variational Bayesian (VB) method of avoiding overfitting by incorporating the *variational approximation* technique

* Corresponding author. Tel.: +81-774-93-5130; fax: +81-774-93-5155.
E-mail address: ueda@cslab.kecl.ntt.co.jp (N. Ueda).

0893-6080/02/$ - see front matter
PII: S0893-6080(02)00040-0

(Jaakkola, 1997) into Bayesian inference. The VB method can be more accurate than the Laplace approximation (MacKay, 1992a,b) in that it does not assume Gaussian distribution of the posterior. Moreover, it is much more efficient than Markov chain Monte Carlo (MCMC) methods (Gamerman, 1997) in that it results in a *deterministic* learning algorithm.

The VB method presented in Waterhouse et al. (1995), however, does not optimize the model structure. Moreover, as in the ML approach, it often suffers from the local maxima problem in practice. Attias (1999) have extended the VB to perform model selection by introducing a posterior over model structures to the VB formulation. The local optimum problem, however, has not been solved yet. Recently, Ghahramani and Beal (2001) have successfully applied the VB to state space models.

One of the authors have already presented a basic idea to solve the local optima problem for mixture of factor analyzers (Ghahramani & Beal, 2000) within the VB framework. However, an explicit objective function for finding the optimal model structure has not been shown there. In contrast, in this paper, for general nonlinear models, we formally derive an objective function that can optimize a model over parameter distributions and model structure *simultaneously* within the VB framework. Then, focusing on mixture models, we devise a Bayesian SMEM algorithm to efficiently optimize the objective function. We also apply the proposed method to the learning of a mixture of experts model and show that unlike the conventional methods, it can automatically find the optimal number of experts without being trapped in poor local optima. One of the authors has already published a short paper (Ueda, 2000) related to the same topic as the present paper. In the present paper, however, we formally derive VB formulae and moreover give complete derivations of VB algorithm for mixture of experts model, which is quite informative to readers. That is, this paper can be regarded as an extended version of the previous short paper.

The organization of the rest of the paper is as follows. In Section 2 we will first review the VB approach. In Section 3, we derive an objective function for *simultaneous* optimization over the distributions and model structure. Then we apply the method to the training of a mixture of experts model in Section 4 and show some experimental results to demonstrate the proposed algorithm in Section 5. Final remarks are presented in Section 6.

2. Variational Bayesian framework

2.1. Bayesian approach

Let d be a random variable in some statistical model to be considered. d can be a scalar, vector, or matrix. Let $\mathcal{H}_{\mathcal{M}} = \{p(d|\theta, \mathcal{M})\}$ denote a class of probability distributions with parameter θ under a fixed model structure $\mathcal{M}$. The model structure is the complexity of a model. More specifically, in the case of mixture model, it corresponds to the number of mixture components. Note that although the parameter θ depends on the model structure, we just write θ for notational simplicity. Then, the ML approach estimates the optimal hypothesis that maximizes the log−likelihood function $\log p(\mathcal{D}|\theta, \mathcal{M})$ using given training data $\mathcal{D}$.That is, in the ML approach, the best hypothesis is $p(d|\hat{\theta}, \mathcal{M})$, where $\hat{\theta}$ represents the ML estimate. However, as pointed out by many researchers, the ML approach often overfits the training data, which decreases generalization ability.

In contrast, the Bayesian approach tries not to estimate the parameter value like in the ML approach, but to estimate *posterior predictive distribution* $p(d^*|\mathcal{D}, \mathcal{M})$ for a new observation d^* defined by

$$p(d^*|\mathcal{D}, \mathcal{M}) = \int p(d^*|\theta, \mathcal{M}) p(\theta|\mathcal{D}, \mathcal{M}) \mathrm{d}\theta. \quad (1)$$

The RHS of Eq. (1) represents an average weighted by a posterior distribution of θ, say, $p(\theta|\mathcal{D}, \mathcal{M})$. Thus, the Bayesian approach can mitigate overfitting since the parameters are integrated out. In this sense, the Bayesian approach can provide a more reliable prediction than the ML approach. Moreover, in the Bayesian approach, by regarding $\mathcal{M}$ as a random variable, we can introduce a model posterior distribution $P(\mathcal{M}|\mathcal{D})$. The best model structure $\mathcal{M}^*$ can be identified by

$$\mathcal{M}^* = \arg \max_{\mathcal{M}} P(\mathcal{M}|\mathcal{D}).$$

Also, we can consider a model structure averaging

$$p(\cdot|\mathcal{D}) = \sum_{\mathcal{M}} p(\cdot|\mathcal{D}, \mathcal{M}) P(\mathcal{M}|\mathcal{D}). \quad (2)$$

The Bayesian approach, however, requires integrals that are in general hard to compute. There have been two kinds of approaches that approximate the integral: the Laplace approximation methods and Markov chain Monte Carlo (MCMC) methods. Recently, a new approach called *variational Bayesian* (VB) approach, which have been proposed and successfully applied to several inference problems (Attias, 1999; Ghahramani & Beal, 2000; Waterhouse et al., 1995).

In Section 2.2, we will describe the VB approach for general nonlinear models including mixture models.

2.2. Basic principle of the VB method

Now consider a Directed Acyclic Graph (DAG) shown in Fig. 1 for a nonlinear model including mixture models. Circles denote the unknowns (random variables) and double square box represents observed data. As shown in Fig. 1, Z, θ, ϑ and $\mathcal{M}$ are treated as random variables. DAG graphically shows the conditional independence between two random variables. For example, in Fig. 1, observed data $\mathcal{D}$ is independent of ϑ given θ. Namely, once θ has been

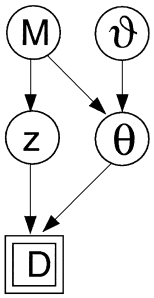

Fig. 1. Graphical model (directed acyclic graph) for a general model. Circles denote the unknowns and double square box represents observed data.

known, $\mathscr{D}$ does not depend on ϑ, but depends on θ. Z is a set of *latent* (unobserved) variables. θ denotes a set of parameters with prior distributions, and ϑ a set of hyperpriors (i.e. prior of prior) with hyperprior distributions. Of course, top level random variables, $\mathscr{M}$ and ϑ usually have hyperparameters (usually predefined as some constants). However, for notational simplicity, we omit the hyperparamters in this section.[1]

Then, the complete data likelihood of the nonlinear model paramterized by θ with a fixed model structure $\mathscr{M}$ is represented by $p(\mathscr{D},Z|\theta,\mathscr{M})$. In the VB approach, we consider the log marginal likelihood in which all random quantities are marginalized:

$$\begin{aligned}\mathscr{L}(\mathscr{D}) &= \log p(\mathscr{D}) \\ &= \log \sum_{\mathscr{M}} \sum_{Z} \int\int p(\mathscr{D}, Z|\theta, \mathscr{M}) p(\theta|\vartheta, \mathscr{M}) p(\vartheta|\mathscr{M}) \\ &\quad \times P(\mathscr{M}) \mathrm{d}\theta \, \mathrm{d}\vartheta.\end{aligned}$$

Here $p(\theta|\vartheta,\mathscr{M})$ and $P(\mathscr{M})$ are priors for θ and $\mathscr{M}$, and $p(\vartheta|\mathscr{M})$ is a hyperprior.

Next, by introducing a new distribution $Q(Z, \theta, \vartheta, \mathscr{M})$ for all random quantities and using Jensen's inequality for the convex function $\log(\cdot)$, $\mathscr{L}$ can be bounded as

$$\begin{aligned}\mathscr{L}(\mathscr{D}) &\geq \sum_{\mathscr{M}} \sum_{Z} \int\int Q(Z, \theta, \vartheta, \mathscr{M}) \\ &\quad \times \log \frac{(\mathscr{D}, Z|\theta, \mathscr{M}) p(\theta|\vartheta, \mathscr{M}) p(\vartheta|\mathscr{M}) P(\mathscr{M})}{Q(Z, \theta, \vartheta, \mathscr{M})} \mathrm{d}\theta \, \mathrm{d}\vartheta \\ &\equiv \mathscr{F}[Q], \end{aligned} \tag{3}$$

where Q is an approximation of the true posterior, $p(Z, \theta, \vartheta, \mathscr{M}|\mathscr{D})$, and is termed the *variational posterior*.[2]

The quantity $\mathscr{F}[Q]$ provides a rigorous lower bound on the log marginal likelihood and it can be shown that the following important relationship between $\mathscr{L}$ and $\mathscr{F}$ holds:

$$\begin{aligned}\mathscr{L}(\mathscr{D}) &= \mathscr{F}[Q(Z, \theta, \vartheta, \mathscr{M})] \\ &\quad + \mathrm{KL}[Q(Z, \theta, \vartheta, \mathscr{M}) \| p(Z, \theta, \vartheta, \mathscr{M}|\mathscr{D})].\end{aligned}$$

Here $\mathrm{KL}[\cdot\|\cdot]$ denotes the Kullback–Leibler divergence defined by

$$\begin{aligned}&\mathrm{KL}[Q(Z, \theta, \vartheta, \mathscr{M}) \| p(Z, \theta, \vartheta, \mathscr{M}|\mathscr{D})] \\ &= \sum_{\mathscr{M}} \sum_{Z} \int\int Q(Z, \theta, \vartheta, \mathscr{M}) \log \frac{Q(Z, \theta, \vartheta, \mathscr{M})}{p(Z, \theta, \vartheta, \mathscr{M}|\mathscr{D})} \mathrm{d}\theta \, \mathrm{d}\vartheta.\end{aligned}$$

That is, since $\mathscr{L}$ is a constant under a fixed $\mathscr{D}$, maximizing $\mathscr{F}[Q]$ w.r.t. Q is equivalent to minimizing the Kullback–Leibler divergence between Q and the true posterior distribution. In other words, the optimal Q that maximizes $\mathscr{F}$ is the best approximation of the true posterior under whatever constraints are imposed on Q. Unlike the Laplace approximation, we do not assume any particular functional form for Q, but we only assume the factorizing form as $Q = Q(Z|\mathscr{M})Q(\theta|\mathscr{M})Q(\vartheta|\mathscr{M})Q(\mathscr{M})$ as a practical requirement. In addition, we further assume that the model priors and hyperpriors factorize:

$$\begin{aligned}p(\theta|\vartheta, \mathscr{M}) &= \prod_{i=1}^{I} p(\theta_i|\vartheta_i, \mathscr{M}) \qquad \text{and} \\ p(\vartheta|\mathscr{M}) &= \prod_{i=1}^{I} p(\vartheta_i|\mathscr{M}).\end{aligned} \tag{4}$$

Note that $\theta = \{\theta_i\}_{i=1}^{I}$ and $\vartheta = \{\vartheta_i\}_{i=1}^{I}$, where I is the number of independent parameters. Correspondingly, we assume that $Q(\theta|\mathscr{M})$ and $Q(\vartheta|\mathscr{M})$ can be factorized as $Q(\theta|\mathscr{M}) = \prod_{i=1}^{I} Q(\theta_i|\mathscr{M})$ and $Q(\vartheta|\mathscr{M}) = \prod_{i=1}^{I} Q(\vartheta_i|\mathscr{M})$. Here, $Q(\theta|\mathscr{M})$, $Q(\vartheta|\mathscr{M})$, $Q(Z|\mathscr{M})$ and $Q(\mathscr{M})$ are approximations of the true posterior distributioins $p(\theta|\mathscr{D},\mathscr{M})$, $p(\vartheta|\mathscr{D},\mathscr{M})$, $P(Z|\mathscr{D},\mathscr{M})$, and $P(\mathscr{M}|\mathscr{D})$, respectively. They are called *variational posteriors*.

From these assumtions, $\mathscr{F}[Q]$ is rewritten as

$$\begin{aligned}\mathscr{F}[Q] &= \sum_{\mathscr{M}} Q(\mathscr{M}) \left\{ \left\langle \log \frac{p(\mathscr{D}, Z|\theta, \mathscr{M})}{Q(Z|\mathscr{M})} \right\rangle_{Q(Z|\mathscr{M}), \prod_{i=1}^{I} Q(\theta_i|\mathscr{M})} \right. \\ &\quad + \sum_{i=1}^{I} \left\langle \log \frac{p(\theta_i|\vartheta_i,)}{Q(\theta_i|\mathscr{M})} \right\rangle_{Q(\theta_i|\mathscr{M}), Q(\vartheta_i|\mathscr{M})} \\ &\quad \left. + \sum_{i=1}^{I} \left\langle \log \frac{p(\vartheta_i|\mathscr{M})}{Q(\vartheta_i|\mathscr{M})} \right\rangle_{Q(\vartheta_i|\mathscr{M})} + \log \frac{P(\mathscr{M})}{Q(\mathscr{M})} \right\}.\end{aligned} \tag{5}$$

Here the notation $\langle f(x) \rangle_{p(x)}$ represents the expectation of $f(x)$ w.r.t. the distribution $p(x)$:

$$\langle f(x) \rangle_{p(x)} = \int f(x) p(x) \mathrm{d}x$$

[1] In an application to mixture of experts model in Section 4, we explicitly describe the hyperparameters.

[2] Note that we should write $Q(\cdot|\mathscr{D})$, but to make the notation simple, the dependence of the variational posterior on the data $\mathscr{D}$ is hereafter omitted.

In the case of a discrete random variable z, the notation $\langle f(z)\rangle_{P(z)}$ represents $\langle f(z)\rangle_{P(z)} = \sum_z f(z)$.

2.3. Inference of optimal variational posteriors and model structure

According to the method of the *variational calculus* by differentiating $\mathcal{F}[Q]$ w.r.t. each of Q distributions and setting them to zero, an EM-like procedure presented below for estimating the optimal variational posteriors can be obtained. We call them *Variational Bayesian (VB) EM steps*. That is, at the $(t+1)$th iteration.

VB E-step computes:

$$Q(Z|\mathcal{M})^{(t+1)} \propto \exp\{\langle \log p(\mathcal{D}, Z|\theta, \mathcal{M})\rangle_{Q(\theta|\mathcal{M})^{(t)}}\}. \tag{6}$$

VB M-step updates: For $i = 1, \ldots, I$,

$$Q(\theta_i|\mathcal{M})^{(t+1)} \propto \exp\{\langle \log p(\mathcal{D}, Z|\theta, \mathcal{M})\rangle_{Q(Z|\mathcal{M})^{(t+1)}, Q(\theta_{-i}|\mathcal{M})^{(t)}} + \langle \log p(\theta_i|\vartheta_i, \mathcal{M})\rangle_{Q(\vartheta_i|\mathcal{M})^{(t)}}\}, \tag{7}$$

$$Q(\vartheta_i|\mathcal{M})^{(t+1)} \propto p(\vartheta_i|\mathcal{M}) \exp\{\langle \log p(\theta_i|\vartheta_i, \mathcal{M})\rangle_{Q(\theta_i|\mathcal{M})^{(t+1)}}\}. \tag{8}$$

The symbol θ_{-i} in Eq. (7) denotes all parametes in θ other than θ_i. By alternately and repeatedly performing the VB EM steps above until the convergence, we can obtain the local optimum estimates of $Q(Z|\mathcal{M})$, $Q(\theta_i|\mathcal{M})$ and $Q(\vartheta_i|\mathcal{M})$.

Note that if we were remove $Q(\theta)$ and $Q(\vartheta)$ from Eq. (6), we would get

$$Q(Z|\mathcal{M})^{(t+1)} \propto p(\mathcal{D}, Z|\theta^{(t)}, \mathcal{M}),$$

which is the same as the posterior distribution computed at the E-step in usual EM algorithm based (Dempster, Laird, & Rubin, 1977) on the ML approach. This means that the VB EM steps above include the ML-based EM algorithm as a special case where θ and ϑ are not random variables, but mathematical variables. This is a reason why we call the above algorithm the VB EM algorithm.

Let $Q(Z|\mathcal{M})^*$, $Q(\theta_i|\mathcal{M})^*$, and $Q(\vartheta_i|\mathcal{M})^*$ denote the estimated posteriors. Then, according to Attias (1999), using these estimated posteriors, the optimal posterior over the model structure can be obtained in a *closed form* as

$$Q(\mathcal{M})^* \propto \exp\left\{\left\langle \log \frac{p(\mathcal{D}, Z|\theta, \mathcal{M})}{Q(Z|\mathcal{M})^*}\right\rangle_{Q(Z|\mathcal{M})^*, Q(\theta|\mathcal{M})^*} + \sum_{i=1}^{I}\left\langle \log \frac{p(\theta_i|\vartheta_i\mathcal{M})}{Q(\theta_i|\mathcal{M})^*}\right\rangle_{Q(\theta_i|\mathcal{M})^*, Q(\vartheta_i|\mathcal{M})^*} + \sum_{i=1}^{I}\left\langle \log \frac{p(\vartheta_i|\mathcal{M})}{Q(\vartheta_i|\mathcal{M})^*}\right\rangle_{Q(\vartheta_i|\mathcal{M})^*} + \log P(\mathcal{M})\right\}. \tag{9}$$

The optimal model structure in the MAP sense can be found as:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} Q(\mathcal{M})^*.$$

However, since $Q(\theta_i|\mathcal{M})$, $Q(\vartheta_i|\mathcal{M})$, and $Q(Z|\mathcal{M})$ are iteratively optimized by using the steepest ascent procedure given by the VB EM steps presented above, we suffer from the local maxima problem as in the ML approach. That is, if these posteriors converge to poor local maxima, we no longer find the appropriate model structure since $Q(\mathcal{M})^*$ value depends on these converged values. In other words, we cannot find the optimal model structure without solving this local optimum problem in the VB learning.

3. Optimal model search

3.1. An objective function

Let $\mathcal{F}_{\mathcal{M}}$ denote all terms independent of $Q(\mathcal{M})$ of $\mathcal{F}[Q]$. That is

$$\mathcal{F}_{\mathcal{M}} = \left\langle \log \frac{p(\mathcal{D}, Z|\theta, \mathcal{M})p(\theta|\vartheta\mathcal{M})p(\vartheta|\mathcal{M})}{Q(Z|\mathcal{M})Q(\theta|\mathcal{M})Q(\vartheta|\mathcal{M})}\right\rangle_{Q(Z|\mathcal{M}), Q(\theta|\mathcal{M}), Q(\vartheta|\mathcal{M})}. \tag{10}$$

Then, using $\mathcal{F}_{\mathcal{M}}$, the lower bound can be rewritten as:

$$\mathcal{F}[Q(Z, \theta, \vartheta, \mathcal{M})] = \langle \mathcal{F}_{\mathcal{M}}\rangle_{Q(\mathcal{M})} - \mathrm{KL}[Q(\mathcal{M})\|P(\mathcal{M})]. \tag{11}$$

Since $\mathcal{F}_{\mathcal{M}}$ does not depend on $Q(\mathcal{M})$[3] and the KL term depends only on $Q(\mathcal{M})$, the conventional VB learning mentioned in Section 2.3 is equivalent to the following steps:

[Conventional VB learning algorithm]

Step 1. For each $\mathcal{M}$, setting $Q(Z|\mathcal{M})^{(0)}$, $Q(\theta|\mathcal{M})^{(0)}$ and $t \leftarrow 0$, perform below until convergence.

$$Q(Z|\mathcal{M})^{(t+1)} = \arg\max_{Q(Z|\mathcal{M})} \mathcal{F}_{\mathcal{M}}[Q(Z|\mathcal{M}), Q(\theta|\mathcal{M})^{(t)}, Q(\vartheta|\mathcal{M})^{(t)}], \tag{12}$$

$$Q(\theta_i|\mathcal{M})^{(t+1)} = \arg\max_{Q(\theta_i|\mathcal{M})} \mathcal{F}_{\mathcal{M}}[Q(Z|\mathcal{M})^{(t+1)}, Q(\theta|\mathcal{M}), Q(\vartheta|\mathcal{M})^{(t)}], \quad i = 1, \ldots, I \tag{13}$$

$$Q(\vartheta_i|\mathcal{M})^{(t+1)} = \arg\max_{Q(\vartheta_i|\mathcal{M})} \mathcal{F}_{\mathcal{M}}[Q(Z|\mathcal{M})^{(t+1)}, Q(\theta|\mathcal{M})^{(t+1)}, Q(\vartheta|\mathcal{M})], \quad i = 1, \ldots, I, \ t \leftarrow t+1 \tag{14}$$

Step 2. For each $\mathcal{M}$, maximize $\mathcal{F}_{\mathcal{M}}$ w.r.t. $\mathcal{M}$.

[3] Note that $\mathcal{F}_{\mathcal{M}}$ does not depend on $Q(\mathcal{M})$, but depends on $\mathcal{M}$.

Note that the re-estimate equations for $Q(Z|\mathcal{M})$, $Q(\theta_i|\mathcal{M})$, and $Q(\vartheta_i|\mathcal{M})$ derived at Step 1 are equivalent to Eqs. (6)–(8).

Let $\mathcal{F}^*_{\mathcal{M}}$ represent the optimal value of $\mathcal{F}_{\mathcal{M}}$ obtained by Step 1 above. Then, from Eq. (6), the optimal model posterior, denoted by $Q(\mathcal{M})^*$, obtained at Step 2 is given by

$$Q(\mathcal{M})^* \propto P(M)\exp\{\mathcal{F}^*_{\mathcal{M}}\}. \tag{15}$$

Note that Eq. (15) is equivalent to Eq. (9). The important point to note that from Eq. (15), the optimal model structure $\mathcal{M}^*$ that maximizes $Q(\mathcal{M})$ is equivalent to the one that maximizes $\mathcal{F}^*_{\mathcal{M}} + \log P(\mathcal{M})$. Since $P(\mathcal{M})$ is assumed to be uniform, the optimal model structure can be found by $\mathcal{F}^*_{\mathcal{M}}$ without computation of $Q(\mathcal{M})^*$.

That is, letting

$$\mathcal{F}^{(t)}_{\mathcal{M}} = \left\langle \log \frac{p(\mathcal{D}, Z|\theta, \mathcal{M})p(\theta|\vartheta, \mathcal{M})p(\vartheta|\mathcal{M})}{Q(Z|\mathcal{M})^{(t)}Q(\theta|\mathcal{M})^{(t)}Q(\vartheta|\mathcal{M})^{(t)}} \right\rangle_{Q(Z|\mathcal{M})^{(t)},Q(\theta|\mathcal{M})^{(t)},Q(\vartheta|\mathcal{M})^{(t)}}, \tag{16}$$

and assuming that $P(\mathcal{M})$ is uniform, we can show the following monotonicity property:

$$\text{If } \mathcal{F}^{(t)}_{\mathcal{M}'} \geq \mathcal{F}^{(t)}_{\mathcal{M}}, \qquad \text{then } Q(\mathcal{M}')^{(t)} \geq Q(\mathcal{M})^{(t)} \text{ holds}$$

This indicates that by maximizing $\mathcal{F}_{\mathcal{M}}$ with respect to not only $Q(Z|\mathcal{M})$ and $Q(\Theta|\mathcal{M})$, but also $\mathcal{M}$, we can obtain the optimal parameter distribution and model structure *simultaneously*, in the sense of the MAP estimate, by using $\mathcal{F}_{\mathcal{M}}$ instead of $\mathcal{F}[Q]$.

3.2. Bayesian SMEM algorithm for mixture models

As mentioned before, the VB learning algorithm often can get caught by local maxima. In the case of mixture models, the local maxima often involve having too many components in one part of the space and too few in another. To escape from such configurations, we employ the idea of the SMEM algorithm (Ueda et al., 2000) that we previously developed within the ML framework.

The application of the idea of the SMEM algorithm to the VB is straightforward. In the case of mixture models, $\mathcal{M}$ corresponds to the number of mixture components. Let m denote the number of mixture components. Then, the objective function $\mathcal{F}_m$ can be represented in the form of a direct sum

$$\mathcal{F}_m = \sum_{j=1}^{m} \mathcal{F}_{(j)},$$

where $\mathcal{F}_{(j)}$ is the objective funtion corresponding to the jth component model of a mixture model with m components. After the conventional VB learning algorithm has converged, the objective function can be rewritten as

$$\mathcal{F}^*_m = \mathcal{F}^*_{(i)} + \mathcal{F}^*_{(j)} + \mathcal{F}^*_{(k)} + \sum_{u,u\neq i,j,k} \mathcal{F}^*_{(u)}. \tag{17}$$

Here, $\mathcal{F}^*_{(i)}$ denote the objective function value of $\mathcal{F}_{(i)}$ after the convergence. We then try to increase the first three terms of the RHS of Eq. (17) by merging model components i and j to produce a model component i' and by splitting the model component k into two model components j' and k'.

In the SMEM algorithm, since the likelihood value monotonically increases as the number of models increases, we repeatedly and simultaneously perform split and merge operations during the learning processes to fix the number of components. On the other hand, in VB since $\mathcal{F}_m$ can be optimized w.r.t. both m and $Q(\theta,Z|m)$, here we employ either the split or merge operation alone in addition to the simultaneous split and merge operation. Clearly, since the split (merge) operation alone means $m \leftarrow m+1 (m \leftarrow m-1)$, these three kinds of operations (split, merge, split and merge) play a role not only to avoid the local maxima but also to search for the optimal number of models.

The variational posteriors of these newly generated models are initialized and reestimated with the other models. If the $\mathcal{F}_m$ value is improved, then we accept the new estimate. Otherwise, we reject it and try another candidate. This procedure is repeatedly performed until the objective function value, $\mathcal{F}_m$, is no longer improved.

By iteratively maximizing $\mathcal{F}_m$ w.r.t. $Q(Z|m)$, $Q(\theta|m)$ and m, we can expect to simultaneously solve both the local maxima and the optimal model structure selection problems. The criteria for choosing split and merge candidates and the process of initialization for newly generated models can be straight forwardly defined as those in the SMEM algorithm by using the MAP estimates. Since these points are not essential in this paper, we will omit them here. See Ueda et al. (2000).

[Variational Bayesian SMEM Algorithm]

Step 1. Perform the conventional VB updates presented in Eqs. (12)–(14). Let $Q(Z|m)^*$, $Q(\theta|m)^*$ and $Q(\vartheta|m)^*$ denote the estimated variational posteriors. Let $F \leftarrow \mathcal{F}^*_m$ and $m^* \leftarrow m$.

Step 2. Sort the split and merge candidates by computing split and merge criteria based on the MAP estimates of the posteriors obtained at Step 1.

Step 3. Perform the following steps independently:

(3-1)

Merge option. For each of the $C_{\max}$ merge candidates, perform merge operation and reestimate the posteriors in order. If a candidate that improves F^* has been found, then set the objective function value to F_1^{**} and ignore the other candidates. If all candidates could not improve F^*, then set $F_1^{**} \leftarrow F^*$.

(3-2)
Split and merge option. For each of the $C_{\max}$ merge candidates, perform split and merge operations and reestimate the posteriors in order. If a candidate that improves F^* as been found, then set the objective function value to F_2^{**} and ignore the other candidates. If all candidates could not improve F^*, then set $F_2^{**} \leftarrow F^*$.
(3-3)
Split option. For each of the $C_{\max}$ merge candidates, perform split and merge operations and reestimate the posteriors in order. If a candidate that improves F^* has been found, then set the objective function value to F_3^{**} and ignore the other candidates. If all candidates could not improve F^*, then set $F_3^{**} \leftarrow F^*$.

Step 4. If there is no candidate that improves F^*, then halt with the current posteriors and m^* as the final solution. Otherwise, let $F^* \leftarrow \max\{F_1,F_2,F_3\}$. If $F^* = F_1^{**}$, then accept the result of (3-1), set $m^* \leftarrow m^* - 1$ and go to Step 2. If $F^* = F_2^{**}$, then accept the result of (3-2) and go to Step 2. If $F^* = F_3^{**}$, then accept the result of (3-3), set $m^* \leftarrow m^* + 1$ and go to Step 2.

Note that in each of steps (3-1), (3-2) and (3-3), when a certain candidate which improves the objective function is found, the other successive candidates are excluded. There is no guarantee therefore that the chosen candidates will give the largest possible improvement in the objective function. This is not a major problem, because the split and merge operations are performed repeatedly.

Clearly, since Step 3 is a *greedy search*, the algorithm above tries to find a better local maximum of $\mathcal{F}_m$. In this sense, we cannot theoretically guarantee the global optimality of the algorithm. However, since the objective function value is monotonically improved, we can efficiently obtain a better local maximum. Each of steps (3-1), (3-2) and (3-3) corresponds to the search of better posterior under a fixed m^* and step 4 selects the best model structure m^*. By repeatedly performing these steps, we can find a better model parameter distribution and model structure, simultaneously.

4. Application to mixture of experts

4.1. Probability model

Our probability formulation of a MoE is based on the *random-regressor* (RR) model in which both input and output are treated as *random variables*. In other words, the input variable is also measured with error (see e.g. Seber & Wild, 1989). Using this model enables us to derive all variational posteriors without approximation. On the other hand, in Waterhouse et al. (1995), Gaussian approximation is forced to be used to derive a variational posterior on the gating network parameter since their formulation is based on the *fixed-regressor* model where only output value is treated as a random variable. Moreover, in our formulation, unlike the formulation by Waterhouse et al. (1995), the model structure (i.e. the number of experts) is also treated as a random variable and therefore it can be automatically optimized. Due to the lack of space, we only describe our Bayesian formulation for a MoE below.

Let $\mathbf{x} \in \mathscr{R}^d$ be a d-dimensional input and $f_i(\mathbf{x},\mathbf{w}_i) \in \mathscr{R}$ be the corresponding output[4] of expert i. Accordingly, the output value of a MoE for an input $\mathbf{x}$ is given by

$$y = \sum_{i=1}^{m} G_i(\mathbf{x}|\mathbf{\Phi}) f_i(\mathbf{x}, \mathbf{w}_i). \tag{18}$$

We restrict each expert to a *linear* funtion, $f_i(\mathbf{x}, \mathbf{w}_i) = \mathbf{w}_i^{\mathrm{T}}\bar{\mathbf{x}}$, where $\mathbf{w}_i \in \mathscr{R}^{d+1}$ is an unknown parameter. Here, $\bar{\mathbf{x}} = (\mathbf{x}^{\mathrm{T}}1)^{\mathrm{T}} \in \mathscr{R}^{d+1}$. As a gating network, we use a *normalized Gaussian* function (Xu, Jordan, & Hinton, 1994) defined by

$$G_i(\mathbf{x}|\mathbf{\Phi}) = \frac{\varphi_i \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{S}_i^{-1}\right)}{\sum_{j=1}^{m} \varphi_j \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_j, \mathbf{S}_j^{-1}\right)}. \tag{19}$$

Here, $\mathbf{\Phi} = \{\varphi_i, \boldsymbol{\mu}_i, \mathbf{S}_i, i = 1, \ldots, m\}$ is a set of unknown parameters, and φ_i is a mixing proportion satisfying $\varphi_i \geq 0$ and $\sum_{i=1}^{m} \varphi_i = 1$. The notation $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{S}^{-1})$ denotes the d-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{S}^{-1}$. Note that $\mathbf{S}$ is called a *precision matrix*.

The probability model for a MoE is given by

$$p(y|\mathbf{x}, \mathbf{\Phi}, \Theta, m) = \sum_{i=1}^{m} P(i|\mathbf{x}, \mathbf{\Phi}) p(y|\mathbf{x}, i, \theta_i), \tag{20}$$

where $P(i|\mathbf{x},\mathbf{\Phi})$ is the conditional probability of selecting expert i given input $\mathbf{x}$, that is, $P(i|\mathbf{x}, \mathbf{\Phi}) = G_i(\mathbf{x}|\mathbf{\Phi})$. $p(y|\mathbf{x},i,\theta_i)$ is the generative model of the ith expert and is usually assumed to be Gaussian with mean $f_i(\mathbf{x}, \mathrm{w}_i) = \mathbf{w}_i^{\mathrm{T}}\bar{\mathbf{x}}$ and variance β_i^{-1} (i.e. precision β_i):

$$p(y|\mathbf{x}, i, \theta_i) = (2\pi)^{-1/2} \beta_i^{1/2} \exp\left\{ -\frac{\beta_i}{2}\left(y - \mathbf{w}_i^{\mathrm{T}}\bar{\mathbf{x}}\right)^2 \right\}. \tag{21}$$

Here, $\theta_i = (\mathbf{w}_i, \beta_i)$. Let $\mathbf{w}_i = (w_i^{\mathrm{T}} w_{i0})^{\mathrm{T}}$, where $w_i \in \mathscr{R}^d$ and $w_{i0} \in \mathscr{R}$. Then, as derived in Appendix A, the joint density of the MoE based on the random regressor

[4] In this paper we focus on the scalar output, but the results can be extended to multivariate utput in a straight-forward way.

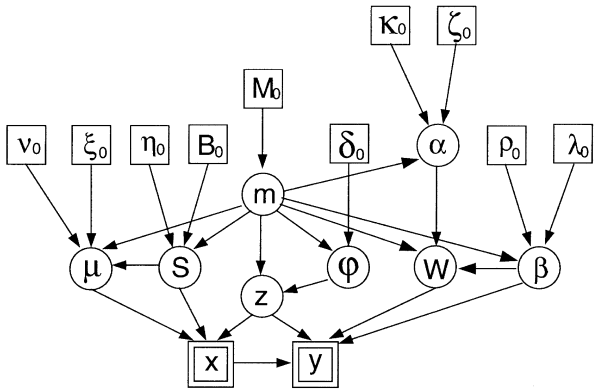

Fig. 2. Graphical model (directed acyclic graph) for the MoE. Circles denote the unknown, square boxed represents fixed quantities, and double square boxes represent observed data.

model becomes a mixture of Gaussians given by

$$p\left(\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix}\middle|\boldsymbol{\Phi},\Theta\right)=\sum_{i=1}^{m}\varphi_i\mathcal{N}\left(\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix}\middle|\begin{pmatrix}\boldsymbol{\mu}_i\\ \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\mu}_i+w_{i0}\end{pmatrix}\right.$$
$$\left.\times,\begin{pmatrix}\mathbf{S}_i+\beta_i\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} & -\beta_i\mathbf{w}_i\\ -\beta_i\mathbf{w}_i^{\mathrm{T}} & \beta_i\end{pmatrix}^{-1}\right). \quad (22)$$

From these, we can compute the following complete data likelihood:

$$p(\mathscr{D},Z|\boldsymbol{\Phi},\Theta,m)=P(Z,X|\boldsymbol{\Phi},m)p(Y|X,Z,\Theta,m)$$
$$=\prod_{i=1}^{m}\prod_{n=1}^{N}\left\{\varphi_i\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_i,\mathbf{S}_i^{-1}\right)\mathcal{N}\left(y_n|f_i(\mathbf{x}_n|\mathbf{w}_i),\beta_i^{-1}\right)\right\}^{z_i^n}. \quad (23)$$

Here, $\Theta=\{\theta_i\}_{i=1}^m=\{\mathbf{w}_i,\beta_i,i=1,\ldots,m\}$. $Z=\{z_i^n\}_{i=1,n=1}^{m,N}$ denotes a set of latent *allocation* variables (latent variables). N is the number of training data. That is, if $(\mathbf{x}_n,y_n)$ was generated from the ith model, $z_i^n=1$; otherwise $z_i^n=0$. Unlike the fixed regressor model (Waterhouse et al., 1995), since Eq. (23) is factorizable with respect to i, as shown later, all variational posteriors can be derived analytically.

It follows that

$$\langle\log p(\mathscr{D},Z|\boldsymbol{\Phi},\Theta,m)\rangle_{Q(Z|m)}$$
$$\propto\sum_{i=1}^{m}\sum_{n=1}^{N}\bar{z}_i^n\left[\log\varphi_i+\log|\mathbf{S}_i|^{1/2}\right.$$
$$-\frac{1}{2}\mathrm{Tr}\left\{\mathbf{S}_i(\mathbf{x}_n-\boldsymbol{\mu}_i)(\mathbf{x}_n-\boldsymbol{\mu}_i)^{\mathrm{T}}\right\}$$
$$\left.+\log\beta_i^{1/2}-\frac{\beta_i}{2}\left(y_n-\tilde{\mathbf{x}}_n^{\mathrm{T}}\mathbf{w}_i\right)^2\right], \quad (24)$$

$$\langle\log p(\mathscr{D},Z|\boldsymbol{\Phi},\Theta,m)\rangle_{Q(Z|m)}$$
$$=\sum_{i=1}^{m}\left[\bar{N}_i\left(\log\varphi_i+\log|\mathbf{S}_i|^{1/2}+\log\beta_i^{1/2}\right)\right.$$
$$-\frac{1}{2}\mathrm{Tr}\left\{\mathbf{S}_i\left(\bar{N}_i(\boldsymbol{\mu}_i-\bar{\mathbf{x}}_i)(\boldsymbol{\mu}_i-\bar{\mathbf{x}}_i)^{\mathrm{T}}+\bar{\mathbf{C}}_i\right)\right\}$$
$$\left.-\frac{\beta_i}{2}(Y-X\mathbf{w}_i)^{\mathrm{T}}\bar{\mathbf{V}}_i(Y-X\mathbf{w}_i)\right] \quad (25)$$

where

$$\bar{z}_i^n=\langle z_i^n\rangle_{Q(z_i^n|m)},\qquad \bar{N}_i=\sum_{n=1}^{N}\bar{z}_i^n,\qquad \bar{\mathbf{x}}_i=\frac{1}{\bar{N}_i}\sum_{n=1}^{N}\bar{z}_i^n\mathbf{x}_n,$$

$$\bar{\mathbf{C}}_i=\sum_{n=1}^{N}\bar{z}_i^n(\mathbf{x}_n-\bar{\mathbf{x}}_i)(\mathbf{x}_n-\bar{\mathbf{x}}_i)^{\mathrm{T}}\in\mathscr{R}^{d\times d}, \quad (26)$$

$$\bar{\mathbf{V}}_i=\mathrm{diag}(\bar{z}_i^1,\ldots,\bar{z}_i^N)\in\mathscr{R}^{N\times N},$$

$$X=(\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_N)^{\mathrm{T}}\in\mathscr{R}^{N(d+1)},\qquad \text{and}$$
$$Y=(y_1,\ldots,y_N)^{\mathrm{T}}\in\mathscr{R}^{N}.$$

Here, $\mathrm{diag}(\bar{z}_i^1,\ldots,\bar{z}_i^N)$ denotes a diagonal matrix with N diagonal elements $\bar{z}_i^1,\ldots,\bar{z}_i^N$.

4.2. Priors

We assume a probabilistic structure for priors shown in Fig. 2 and explain each of the priors as follows. Eq. (19) indicates that an input variable $\mathbf{x}$ is assumed to be from a mixture of Gaussians. It is well known that the *natural conjugate* prior of a single multivariate normal density is a *normal-Wishart* distribution (e.g. Bernardo & Smith, 1994). Accordingly, we employ this distribution on the parameters $\{\boldsymbol{\mu}_i,\mathbf{S}_i\}$ as follows:

$$p(\{\boldsymbol{\mu}_i,\mathbf{S}_i\}|m)=\prod_{i=1}^{m}\mathcal{N}\left(\boldsymbol{\mu}_i|\boldsymbol{\nu}_0,(\xi_0\mathbf{S}_i)^{-1}\right)\mathscr{W}(\mathbf{S}_i|\eta_0,\mathbf{B}_0). \quad (27)$$

The Wishart distribution is defined by

$$\mathscr{W}(\mathbf{S}_i|\eta_0,\mathbf{B}_0)=c|\mathbf{S}_i|^{(1/2)(\eta_0-d-1)}\exp\left\{-\frac{1}{2}\mathrm{Tr}\{\mathbf{B}_0\mathbf{S}_i\}\right\},$$

where c is a normalization constant given by

$$c=\frac{|\mathbf{B}_0|^{\eta_0/2}}{2^{\eta_0 d/2}\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\left(\frac{\eta_0+1-i}{2}\right)}.$$

Here, $\Gamma(\,)$ denotes the gamma function. The mean matrix is $\mathbf{E}\{\mathbf{S}_i\}=\langle\mathbf{S}_i\rangle_{Q(\mathbf{S}_i|m)}=\eta_0\mathbf{B}_0^{-1}$. Note that variables with '0' ($\boldsymbol{\nu}_0$, ξ_0, etc.) are parameters of the prior distribution, called *hyperparameters*. In this paper, we set them constants (scalar, vector or matrix) which are predetermined in some heuristic manner.

The prior on a set of mixture proportions $\boldsymbol{\varphi}$ will always

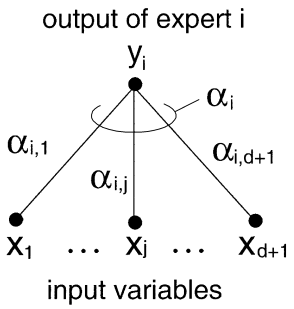

Fig. 3. $\alpha_{i,j}$ corresponds to the variance of $w_{i,j}$ and therefore $\alpha_{i,j}^{-1}=0$ indicates that $\mathbf{x}_j$ is relevant to form the distribution of y_i.

be taken as the *Dirichlet* distribution:

$$p(\boldsymbol{\varphi}|m)=\mathscr{D}(\{\varphi_i\}_{i=1}^m|\delta_0)=c'\prod_{i=1}^m\varphi_i^{\delta_0-1}, \tag{28}$$

where the normalization constant is given by

$$c'=\Gamma\left(\sum_{j=1}^m\varphi_j\right)\Big/\prod_{i=1}^m\Gamma(\varphi_i).$$

For the prior on $(\mathbf{w}_i,\beta_i)$, considering that β_i is the inverse of variance and takes a positive real value, we assume normal–Gamma distribution:

$$p(\{\mathbf{w}_i,\beta_i\}|\{\boldsymbol{\alpha}_i\},m)=\prod_{i=1}^m\mathscr{N}\left(\mathbf{w}_i|\mathbf{0},(\beta_i\Lambda_i)^{-1}\right)\mathscr{G}(\beta_i|\rho_0,\lambda_0). \tag{29}$$

Here, $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_m)$, $\boldsymbol{\alpha}_i=(\alpha_{i,1},\ldots,\alpha_{i,d+1})$ and $\Lambda_i=\mathrm{diag}(\alpha_{i,1},\ldots,\alpha_{i,d+1})$. $\alpha_{i,j}$ is the hyperprior on which $\mathbf{w}_{i,j}$ depends, and corresponds to the inverse of the variance of $w_{i,j}$. $\boldsymbol{\alpha}_i$ controls the magnitude of the weight on connection between the output of the ith expert and the input variable $x_j\in\mathscr{R}$ in $\bar{\mathbf{x}}$. More specifically, $\alpha_{i,j}^{-1}=0$ indicates that x_j is irrelevant to form the distribution of the ith expert's output value y_i as shown in Fig. 3. Using this hyperprior, relevant input variables are automatically selected for each exprt and therfore more flexible predictions would be expected. This kind of hyperprior for input variable selection, called the Automatic Relevance Determination (ARD), is proposed by (MacKay, 1994; Neal, 1996) and has been successfully used in several models (Bishop, 1999; Ghahramani & Beal, 2000; Tipping, 2000). We assume that the distribution of $\alpha_{i,j}$ is a Gamma, $p(\alpha_{i,j})=\mathscr{G}(\alpha_{i,j}|\kappa_0,\zeta_0)$. The Gamma distribution is defined by

$$\mathscr{G}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{a-1}\exp\{-bx\}.$$

Moreover, a prior on m is assumed to be uniform (noninformative prior), $P(m)=1/M_0$.

4.3. Optimal variational posteriors

As for the variational posteriors, we assume the following factorizing form:

$$\begin{aligned}Q&=Q(m)Q(Z|m)Q(\boldsymbol{\Phi}|m)Q(\Theta|m)\\&=Q(m)Q(Z|m)Q(\boldsymbol{\varphi}|m)Q(\boldsymbol{\mu},\mathbf{S}|m)Q(\mathbf{W},\boldsymbol{\beta}|m)Q(\boldsymbol{\alpha}|m),\end{aligned} \tag{30}$$

where $\boldsymbol{\varphi}=\{\varphi_i\}_{i=1}^m$, $\boldsymbol{\mu}=\{\boldsymbol{\mu}_i\}_{i=1}^m$, $\mathbf{S}=\{\mathbf{S}_i\}_{i=1}^m$, $\mathbf{W}=\{w_i\}_{i=1}^m$, $\boldsymbol{\beta}=\{\beta_i\}_{i=1}^m$, and $\boldsymbol{\alpha}=\{\alpha_{ij}\}_{i=1,j=1}^{m,d+1}$. Thus, the objective function shown in Eq. (10) is now specified as follows:

$$\begin{aligned}\mathscr{F}_m=&\left\langle\log\frac{p(\mathscr{D},Z|\boldsymbol{\Phi},\Theta,m)}{Q(Z|m)}\right\rangle_{Q(Z,\boldsymbol{\varphi},\boldsymbol{\mu},\mathbf{S},\mathbf{W},\boldsymbol{\beta}|m)}\\&+\left\langle\log\frac{p(\boldsymbol{\varphi}|m)}{Q(\boldsymbol{\varphi}|m)}\right\rangle_{Q(\boldsymbol{\varphi}|m)}+\left\langle\log\frac{p(\boldsymbol{\mu},\mathbf{S}|m)}{Q(\boldsymbol{\mu},\mathbf{S}|m)}\right\rangle_{Q(\boldsymbol{\mu},\mathbf{S}|m)}\\&+\left\langle\log\frac{p(\mathbf{W},\boldsymbol{\beta}|\boldsymbol{\alpha},m)}{Q(\mathbf{W},\boldsymbol{\beta}|m)}\right\rangle_{Q(\mathbf{W},\boldsymbol{\beta},\boldsymbol{\alpha}|m)}+\left\langle\log\frac{p(\boldsymbol{\alpha}|m)}{Q(\boldsymbol{\alpha}|m)}\right\rangle_{Q(\boldsymbol{\alpha}|m)}\end{aligned} \tag{31}$$

Although the optimal variational posterior distributions can be obtained by setting the functional derivative of $\mathscr{F}_m$ w.r.t. each of Q to zero, we can just use the results shown in Eqs. (6)–(8). The derived variational posterior distributions are summarized below. The detailed derivations are provided in Appendix B.

Results. $\{\varphi_i\}_{i=1}^m$ follows a Dirichlet distribution:

$$Q(\{\varphi_i\}_{i=1}^m|m)=\mathscr{D}(\{\varphi_i\}_{i=1}^m|\{\delta_0+\bar{N}_i\}_{i=1}^m). \tag{32}$$

$\boldsymbol{\mu}_i$ follows a multivariate Student's-T distribution:

$$Q(\boldsymbol{\mu}_i|m)=\mathscr{T}(\mu|\bar{\boldsymbol{\mu}}_i,\Sigma_{\boldsymbol{\mu}_i},f_{\boldsymbol{\mu}_i}), \tag{33}$$

where

$$\bar{\boldsymbol{\mu}}_i=\frac{\bar{N}_i\bar{\mathbf{x}}_i+\xi_0\boldsymbol{\nu}_0}{\bar{N}_i+\xi 0},\qquad \Sigma_{\boldsymbol{\mu}_i}=\frac{1}{(\bar{N}_i+\xi_0)f_{\boldsymbol{\mu}_i}}\mathbf{B}_i,$$

$$f_{\boldsymbol{\mu}_i}=\bar{N}_i+\eta_0+1-d, \tag{34}$$

$$\mathbf{B}_i=\mathbf{B}_0+\bar{\mathbf{C}}_i+\frac{\bar{N}_i\xi_0}{\bar{N}_i+\xi_0}(\bar{\mathbf{x}}_i-\boldsymbol{\nu}_0)(\bar{\mathbf{x}}_i-\boldsymbol{\nu}_0)^{\mathrm{T}}.$$

Here, $\mathscr{T}(\,)$ denotes a d-dimensional Student's-T distribution defined by

$$\mathscr{T}(\mathbf{x}|\boldsymbol{\mu},\Sigma,\nu)\propto\left\{1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}^{-(\nu+d)/2}, \tag{35}$$

with $\nu(>0)$ degrees of freedom, mode $\boldsymbol{\mu}$, and scale matrix Σ (a symmetric, positive-definite matrix[5]). Note that the mean and covariance matrix of $\mathbf{x}\in\mathscr{R}^d$ following the Student's-T distribution are $\mathbf{E}\{\mathbf{x}\}=\boldsymbol{\mu}$ and $\mathrm{Var}\{\mathbf{x}\}=\nu/(\nu-2)\Sigma$, respectively.

[5] When $d=1$, the matrix reduces to a positive scalar value σ.

$\mathbf{S}_i$ follows a Wishart distribution:

$$Q(\mathbf{S}_i|m) = \mathcal{W}(\mathbf{S}_i|\eta_0 + \bar{N}_i, \mathbf{B}_i). \tag{36}$$

β_i follows a gamma distribution:

$$Q(\beta_i|m) = \mathcal{G}\left(\beta_i|\rho_0 + \frac{\bar{N}_i}{2}, \lambda 0 + \frac{R_i}{2}\right), \tag{37}$$

where

$$R_i = (Y - X\bar{\mathbf{w}}_i)^{\mathrm{T}}\bar{\mathbf{V}}_i(Y - X\bar{\mathbf{w}}_i) + \bar{\mathbf{w}}_i^{\mathrm{T}}X^{\mathrm{T}}\left(X^{\mathrm{T}}\bar{\mathbf{V}}_iX + \langle\Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)^{-1}\langle\Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\bar{\mathbf{w}}_i. \tag{38}$$

Moreover

$$\langle\Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)} = \mathrm{diag}(\langle\alpha_{i,1}\rangle_{Q(\alpha_{,1}|m)}, \ldots, \langle\alpha_{i,d+1}\rangle_{Q(\alpha_{i,d+1}|m)}) = \mathrm{diag}\left(\frac{2\zeta_{i,1}}{2\kappa_0 + 1}, \ldots, \frac{2\zeta_{i,d+1}}{2\kappa_0 + 1}\right). \tag{39}$$

Here, $\zeta_{i,j}$ will be defined later.

$\mathbf{w}_i$ follows a multivariate Student's-T distribution:

$$Q(\mathbf{w}_i|m) = \mathcal{T}(\mathbf{w}_i|\bar{\mathbf{w}}_i, \Sigma_{\mathbf{w}_i}, f_{\mathbf{w}_i}), \tag{40}$$

where the parameters are given by

$$\bar{\mathbf{w}}_i = \left(X^{\mathrm{T}}\bar{\mathbf{V}}_iX + \langle\Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)^{-1}X^{\mathrm{T}}\bar{\mathbf{V}}_iY,$$
$$\Sigma_{\mathbf{w}_i} = \left(\frac{2\lambda_0 + R_i}{2\rho_0 + \bar{N}_i}\right)\left(X^{\mathrm{T}}\bar{\mathbf{V}}_iX + \langle\Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)^{-1}, \tag{41}$$
$$f_{\mathbf{w}_i} = 2\rho_0 + \bar{N}_i.$$

$\alpha_{i,j}$ follows a gamma distribution:

$$Q(\alpha_{ij}|m) = \mathcal{G}\left(\alpha_{ij}|\kappa_0 + \frac{1}{2}, \zeta_{i,j}\right), \tag{42}$$

where

$$\zeta_{i,j} = \zeta_0 + \frac{1}{2}\left(\frac{2\rho_0 + \bar{N}_i}{2\lambda_0 + \bar{N}_i}\right)\left(\frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i} - 2}(\Sigma_{\mathbf{w}_i})_{j,j} + \bar{w}_{i,j}^2\right). \tag{43}$$

Finally,

$$\bar{z}_i^n = Q\left(z_i^n = 1|m\right) = \frac{\exp\{\gamma_i^n\}}{\sum_{j=1}^{m}\exp\{\gamma_j^n\}}, \tag{44}$$

where

$$\gamma_i^n = \Psi(\delta_0 + \bar{N}_i) - \Psi\left(m\delta_0 + \sum_{i=1}^{m}\bar{N}_i\right) + \frac{1}{2}\sum_{j=1}^{d}\Psi\left(\frac{\eta_0 + \bar{N}_i + 1 - j}{2}\right) - \frac{1}{2}\log|\mathbf{B}_i| - \frac{1}{2}\mathrm{Tr}\left\{(\eta_0 + \bar{N}_i)\mathbf{B}_i^{-1}\left(\frac{f_{\boldsymbol{\mu}_i}}{f_{\boldsymbol{\mu}_i} - 2}\Sigma_{\boldsymbol{\mu}_i} + (\mathbf{x}_n - \bar{\boldsymbol{\mu}}_i)(\mathbf{x}_n - \bar{\boldsymbol{\mu}}_i)^{\mathrm{T}}\right)\right\} + \frac{1}{2}\left(\Psi\left(\rho_0 + \frac{\bar{N}_i}{2}\right) - \log\left(\lambda_0 + \frac{R_i}{2}\right)\right) - \frac{1}{2}\left(\frac{2\rho_0 + \bar{N}_i}{2\lambda_0 + R_i}\right)\left\{\left(y_n - \tilde{\mathbf{x}}_n^{\mathrm{T}}\bar{\mathbf{w}}_i\right)^2 + \frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i} - 2}\tilde{\mathbf{x}}_n^{\mathrm{T}}\Sigma_{\mathbf{w}_i}\tilde{\mathbf{x}}_n\right\}. \tag{45}$$

Here, $\Psi(x)$ denotes the digamma function defined by

$$\Psi(x) = \frac{\partial\log\Gamma(x)}{\partial x}$$

Note that the above equations are analytically derived, but each distribution cannot be obtained analytically since these posterior distributions are mutually dependent. Instead, as shown in Eqs. (6) and (7), these distributions are iteratively estimated.

Since the joint distribution of x and y, as shown in Eq. (22), becomes a mixture of Gaussians, we apply the split and merge cirteria used in density estimation problems (Ueda et al., 2000a) to the joint distribution. From these settings, we can perform the VB SMEM algorithm for optimal model search of a MoE.

4.4. Prediction

Once we have the optimal variational posteriors, now our goal is to estimate *predictive posterior distribution* of y_{N+1} corresponding to an unknown input $\mathbf{x}_{N+1}$. In the case of the random regressor models, the predictive distribution is found by computing $p(y, \mathbf{x}|\mathcal{D})$, then substituting $\mathbf{x} = \mathbf{x}_{N+1}$ into the distribution and rearranging it w.r.t. y_{N+1}.

First, using Q as an approximation to the posterior over parameters, the joint posterior distribution is given by

$$p(y, \mathbf{x}|\mathcal{D}) = \int\sum_{i=1}^{m} = G_i(\mathbf{x}|\boldsymbol{\Phi})\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{S}_i^{-1}\right)\mathcal{N}\left(y|\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}, \beta_i^{-1}\right) \times Q(\boldsymbol{\Phi}|m)Q(\boldsymbol{\mu}_i, \mathbf{S}_i|m)Q(\mathbf{w}_i, \beta_i|m) \times \mathrm{d}\boldsymbol{\Phi}\,\mathrm{d}\boldsymbol{\mu}_i\,\mathrm{d}\mathbf{S}_i\,\mathrm{d}\mathbf{w}_i\,\mathrm{d}\beta_i, \tag{46}$$

where m denotes the optimal number of experts. Due to the

nonlinearlity of the function G_i, we approximate the integration w.r.t. $\mathbf{\Phi}$ by the MAP estimate. That is

$$p(y, \mathbf{x}|\mathscr{D}) \simeq \sum_{i=1}^{m} G_i(\mathbf{x}|\mathbf{\Phi}_{\mathrm{MAP}}) \times \int \mathscr{N}\left(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{S}_i^{-1}\right) Q(\boldsymbol{\mu}_i, \mathbf{S}_i|m) \mathrm{d}\boldsymbol{\mu}_i \, \mathrm{d}\mathbf{S}_i \times \int \mathscr{N}\left(y|\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}, \beta_i^{-1}\right) Q(\mathbf{w}_i, \beta_i|m) \mathrm{d}\mathbf{w}_i \, \mathrm{d}\beta_i. \quad (47)$$

Moreover, given $\mathbf{x}_{N+1}$, the first integral does not depend on y_{N+1} and can be regarded as a constant. Therefore, the predictive posterior distribution of y_{N+1} is

$$p(y_{N+1}|\mathbf{x}_{N+1}, \mathscr{D}) \simeq \sum_{i=1}^{m} G_i(\mathbf{x}_{N+1}|\mathbf{\Phi}_{\mathrm{MAP}}) \times \int \mathscr{N}\left(y|\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}, \beta_i^{-1}\right) Q(\mathbf{w}_i, \beta_i|m) \mathrm{d}\mathbf{w}_i \, \mathrm{d}\beta_i. \quad (48)$$

The integration of the R.H.S. of Eq. (48) can be analytically computed and we can see that the distribution is a mixture of univariate Student-T distributions:

$$p(y_{+1+1}|\mathbf{x}_{N+1}, \mathscr{D}) \simeq \sum_{i=1}^{m} G_i(\mathbf{x}_{N+1}|\mathbf{\Phi}_{\mathrm{MAP}}) \mathscr{T}(y_{N+1}|\bar{\mathbf{w}}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}, \sigma_i, 2\rho_0 + \bar{N}_i), \quad (49)$$

where the scale parameter σ_i is given by

$$\sigma_i = \frac{2\lambda_0 + R_i}{(2\rho_0 + \bar{N}_i)\left\{1 - \tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}(X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle + \tilde{\mathbf{x}}_{N+1}\tilde{\mathbf{x}}_{N+1}^{\mathrm{T}})^{-1}\tilde{\mathbf{x}}_{N+1}\right\}}. \quad (50)$$

The definition of the Student-T distribution is presented in Eq. (35). Note that the mean and variance of y_{N+1} are given by

$$E\{y_{N+1}\} = \sum_{i=1}^{m} G_i(\mathbf{x}_{N+1}|\mathbf{\Phi}_{\mathrm{MAP}}) \bar{\mathbf{w}}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}, \quad (51)$$

$$\mathrm{Var}\{y_{N+1}\} = \sum_{i=1}^{m} G_i(\mathbf{x}_{N+1}|\mathbf{\Phi}_{\mathrm{MAP}}) \frac{(2\rho_0 + \bar{N}_i)\sigma_i}{2\rho_0 + \bar{N}_i - 2}. \quad (52)$$

The detailed derivations of these results are provided in Appendix C.

5. Experiments

5.1. Synthetic data

To visually demonstrate the behavior of the proposed algorithm, we first show the result of one-dimensional input and output synthetic data. Fig. 4(a) shows the true function and synthetically generated data (300 points) with small noise. Clearly, the MoE with six linear experts is optimum. We initialized a MoE with $m = 6$ as shown in Fig. 4(b) and performed the conventional VB learning. Clearly, it converged to a poor local maxima shown in Fig. 4(c). Note that each straight line, thick curve, and dotted line correspond to each expert ($\bar{\mathbf{w}}_i^{\mathrm{T}}\tilde{\mathbf{x}}$), the expected prediction value, and the standard deviation interval, respectively.

On the other hand, initializing an MoE model with $m = 3$ (Fig. 4(d), we successfully found the optimal model structure (Fig. 4(h)). In this case, the split operation was only accepted three times. That is, m changed 4, 5, and 6 monotonically. Note that the number of steps t does not include rejected learning steps.

Fig. 4(i) shows the trajectories of the objective function value, $\mathscr{F}_m$, and the mean squared error (MSE) for independent test data (500 points) during the learning process from Fig. 4(d)–(h). $\mathscr{F}_m$ values corresponding to Fig. 4(e)–(h) were -41.5, -15.8, -1.6 and 1.2. One can see that the MSE values decrease as the $\mathscr{F}_m$ value increases. The $\mathscr{F}_m$ value corresponding to Fig. 4(c) was -14.6, which is smaller than that of Fig. 4(g). This indicates that it is possible to develop a situation where it is hard to find the optimum model structure by the conventional VB learning algorithm due to the local maxima problem.

5.2. Realistic data

We also applied the proposed algorithm to 'kin-8nm' data in the DELVE database (Rasmussen et al., 1996) in which the local optimum problem is more crucial. The dataset is synthetically generated from a realistic simulation of the forward kinematics of an eight-link all-revolute robot arm. It consists of eight inputs and one ouput with medium noise and highly nonlinearity. The number of training (test) data was 256 (256). Table 1 except the last column shows maximum and minimum values of $\mathscr{F}_m$ and MSE obtained by performing the conventional VB training with fixed m (i.e. without model search) over 10 trials with different initialization. One sees that due to the local optima, $\mathscr{F}_m$ values for each m were unstable and therefore, the batch type model selection based on the $\mathscr{F}_m$ value was unreliable. Starting with $m = 5, \ldots, 10$, we performed the proposed model search algorithm independently. For each starting value m, we performed the proposed model search algorithm just one trial. For all $m = 5, \ldots, 10$, the algorithm converged to the same $m = 8$ and maximum and minimum $\mathscr{F}_m$ and MSE values are shown in the last column (marked by '$*$') in Table 1, which was very stable. Our model search method found the best

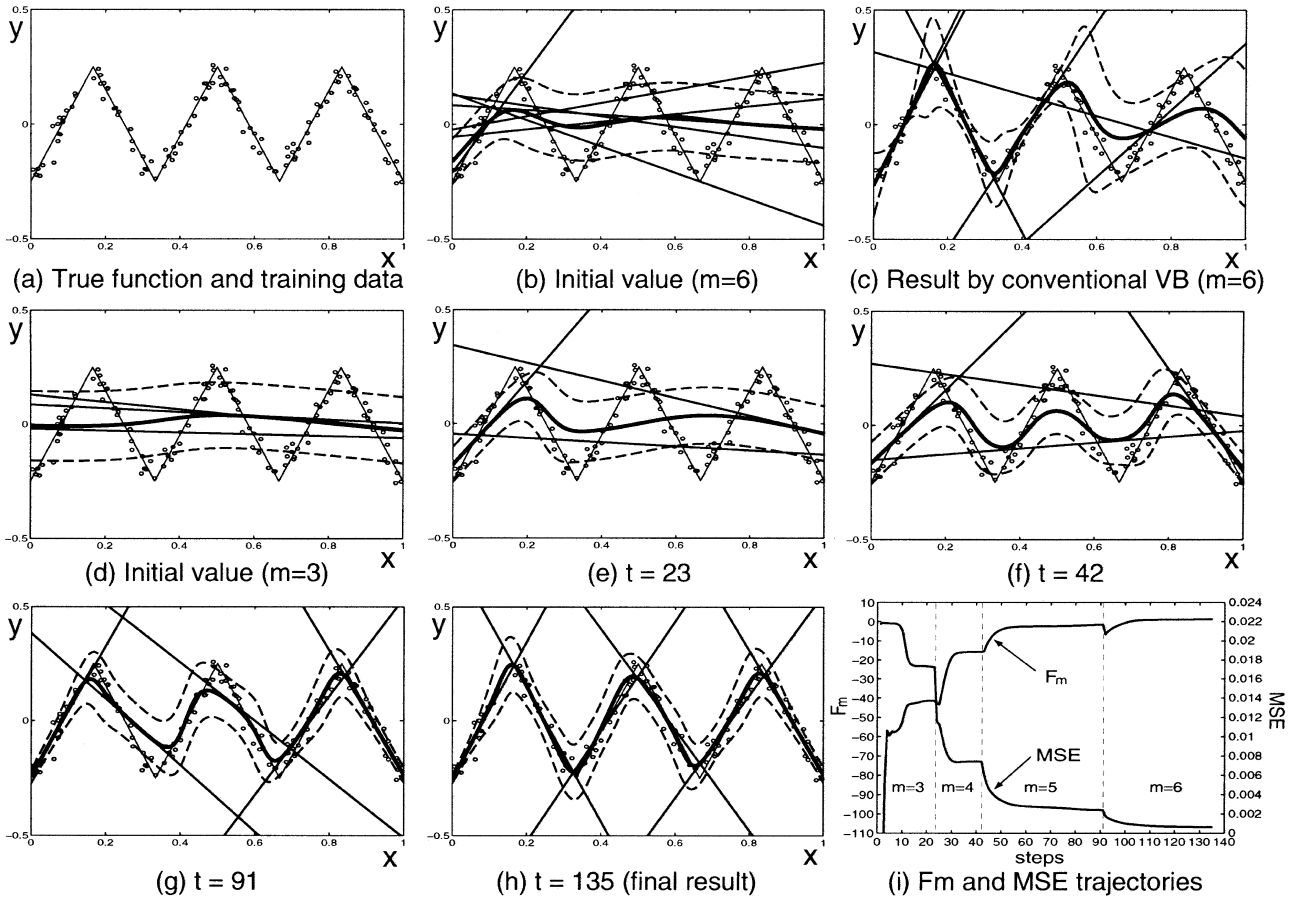

Fig. 4. Results for synthetic data.

results, although we performed just one trial at each initial m.

6. Conclusions

We have proposed a novel method for simultaneously solving the local optima and model structure optimization problems for mixture models based on the variational Bayesian framework. We have applied the proposed method to the mixture of linear expert models and demonstrated the usefulness of the method through experimental results using synthetic and realistic data.

In this paper, the formulation is based on the *joint density models*. In the case of the *fixed regressor models* in which only output value is random variable, we should use conditional density models instead of the joint density models. Recently, within the ML framework, the conditional EM (CEM) algorithm has been proposed by (Jebra & Pentland, 2000, 2002) to conditionally estimate the density. Extensions of this technique to conditional variational Bayesian methods should be used for the fixed regressor models.

Table 1
$\mathcal{F}_m$ and MSE values for each m and by the proposed model search ($*$)

m		5	6	7	8	9	10	$*$
$\mathcal{F}_m$	Min	−3002	−2985	−2911	−2821	−2969	−2927	−2401
	Max	−2671	−2590	−2587	−2514	−2567	−2715	−2381
MSE	Min	0.481	0.498	0.476	0.465	0.475	0.480	0.457
	Max	0.502	0.497	0.481	0.489	0.502	0.531	0.465

Appendix A. Derivation of Eq. (22)

From our probability setting of a MoE, we have

$$p\left(\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix}\middle|\boldsymbol{\Phi},\Theta\right)=\sum_{i=1}^{m}\varphi_i\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_i,\mathbf{S}_i^{-1}\right)N\left(y|\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}\beta_i^{-1}\right)=\sum_{i=1}^{m}(2\pi)^{((d+1)/2)}|\mathbf{S}_i|^{1/2}\beta_i^{1/2}\times\exp\left\{-\frac{1}{2}\underbrace{\left((\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{S}_i(\mathbf{x}-\boldsymbol{\mu}_i)+\beta_i\left(y-\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}\right)^2\right)}_{J}\right\}. \tag{A1}$$

Here, J can be rewritten as

$$J=\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ y-\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\mathbf{S}_i & \mathbf{0}_d\\ \mathbf{0}_d^{\mathrm{T}} & \beta_i\end{pmatrix}\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ y-\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}\end{pmatrix}, \tag{A2}$$

where $\mathbf{0}_d$ denotes d-dimensional zero vector. Moreover

$$\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ y-\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}\end{pmatrix}=\begin{pmatrix}\mathbf{I}_d & \mathbf{0}_d\\ -\mathbf{w}_i^{\mathrm{T}} & 1\end{pmatrix}\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ y-(\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\mu}_i+w_{i0})\end{pmatrix}, \tag{A3}$$

where $\mathbf{I}_d$ is the d-dimensional identity matrix and $\mathbf{w}_i=(w_i^{\mathrm{T}}w_{i0})^{\mathrm{T}}$. Substituting Eq. (A3) into (A2)

$$J=\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ y-(\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\mu}_i+w_{i0})\end{pmatrix}^{\mathrm{T}}\underbrace{\begin{pmatrix}\mathbf{I}_d & -\mathbf{w}_i\\ \mathbf{0}_d^{\mathrm{T}} & 1\end{pmatrix}\begin{pmatrix}\mathbf{S}_i & \mathbf{0}_d\\ \mathbf{0}_d^{\mathrm{T}} & \beta_i\end{pmatrix}\begin{pmatrix}\mathbf{I}_d & \mathbf{0}_d\\ -\mathbf{w}_i^{\mathrm{T}} & 1\end{pmatrix}}_{\begin{pmatrix}\mathbf{S}_i+\beta_i\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} & -\beta_i\mathbf{w}_i\\ -\beta_i\mathbf{w}_i^{\mathrm{T}} & \beta_i\end{pmatrix}}\begin{pmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ y-(\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\mu}_i+w_{i0})\end{pmatrix}.$$

Noting that

$$\begin{vmatrix}\mathbf{S}_i+\beta_i\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} & -\beta_i\mathbf{w}_i\\ -\beta_i\mathbf{w}_i^{\mathrm{T}} & \beta_i\end{vmatrix}=|\mathbf{S}_i|\beta_i, \tag{A4}$$

we arrive at

$$\begin{aligned}p\left(\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix}|\boldsymbol{\Phi},\Theta\right)&=\sum_{i=1}^{m}\varphi_i(2\pi)^{((d+1)/2)}|\Sigma_i|^{1/2}\exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{\mu}_i\\ \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\mu}_i+w_{i0}\end{pmatrix}^{\mathrm{T}}\Sigma_i\begin{pmatrix}\boldsymbol{\mu}_i\\ \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\mu}_i+w_{i0}\end{pmatrix}\right\}\\ &=\sum_{i=1}^{m}\varphi_i\mathcal{N}\left(\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix}\middle|\begin{pmatrix}\boldsymbol{\mu}_i\\ \mathbf{w}_i^{\mathrm{T}}\boldsymbol{\mu}_i+w_{i0}\end{pmatrix},\Sigma_i^{-1}\right),\end{aligned} \tag{A5}$$

where

$$\Sigma_i=\begin{pmatrix}\mathbf{S}_i+\beta_i\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} & -\beta_i\mathbf{w}_i\\ -\beta_i\mathbf{w}_i^{\mathrm{T}} & \beta_i\end{pmatrix}.$$

Thus, we have found that the joint distribution is a mixture of Gaussian shown in Eq. (22).

Appendix B. Derivations of the optimal variational posteriors

B.1. $Q(\{\varphi_i\}_{i1}^m|m)$

By replacing θ_i in Eq. (7) for φ, we get

$$Q(\boldsymbol{\varphi}|m)\propto p(\boldsymbol{\varphi}|m)\exp\{\langle\log(\mathcal{D},Z|\boldsymbol{\varphi},\boldsymbol{\mu},\mathbf{S},\mathbf{W},\boldsymbol{\beta},m)\rangle_{Q(Z|m),Q(\boldsymbol{\mu},\mathbf{S},\mathbf{W},\boldsymbol{\beta}|m)}\}. \tag{B1}$$

By ignoring no $\boldsymbol{\varphi}$-dependence terms in Eq. (25) as constants and using Eq. (32)

$$Q(\boldsymbol{\varphi}|m) \propto \prod_{i=1}^{m} \exp\left\{ \sum_{n=1}^{m} \bar{z}_i^n \log \varphi_i + \log \varphi_i^{\delta_0 - 1} \right\} = \prod_{i=1}^{m} \varphi_i^{\delta_0 + \bar{N}_i - 1} = \mathscr{D}(\{\varphi_i\}_{i=1}^m | \{\delta_0 + \bar{N}_i\}_{i=1}^m). \tag{B2}$$

Eq. (B2) indicates that $Q(\boldsymbol{\varphi}|m)$ is a Dirichlet distribution given by Eq. (32).

B.2. $Q(\boldsymbol{\mu}_i|m)$ and $Q(\boldsymbol{S}_i|m)$

Similarly, replacing θ_i in Eq. (7) for $\{\boldsymbol{\mu},\mathbf{S}\}$, we obtain

$$\begin{aligned}
Q(\boldsymbol{\mu}, \mathbf{S}) &\propto p(\boldsymbol{\mu}, \mathbf{S}) \exp\{\langle \log p(\mathscr{D}, Z | \boldsymbol{\varphi}, \boldsymbol{\mu}, \mathbf{S}, \mathbf{W}, \boldsymbol{\beta}, m) \rangle_{Q(z|m), Q(\boldsymbol{\varphi}, \mathbf{W}, \boldsymbol{\beta}|m)}\} \\
&\propto \prod_{i=1}^{m} \exp\Bigg\{ \bar{N}_i \log |\mathbf{S}_i|^{1/2} - \frac{1}{2} \mathrm{Tr}\Big\{ \mathbf{S}_i \Big(\bar{N}_i (\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^{\mathrm{T}} + \bar{\mathbf{C}}_i \Big) \Big\} + \log |\mathbf{S}_i|^{1/2} - \frac{1}{2} \mathrm{Tr}\Big\{ \xi_0 \mathbf{S}_i (\boldsymbol{\mu}_i - \boldsymbol{\nu}_0)(\boldsymbol{\mu}_i - \boldsymbol{\nu}_0)^{\mathrm{T}} \Big\} \\
&\quad + \log |\mathbf{S}_i|^{(1/2)(\eta_0 - d + 1)} - \frac{1}{2} \mathrm{Tr}\{\mathbf{B}_0 \mathbf{S}_i\} \Bigg\} \\
&\propto \prod_{i=1}^{m} |\mathbf{S}_i|^{1/2} \exp\Bigg\{ -\frac{1}{2} \mathrm{Tr}\Bigg\{ \mathbf{S}_i (\bar{N}_i + \xi_0) \Bigg(\boldsymbol{\mu}_i - \frac{\bar{N}_i \bar{\mathbf{x}}_i + \xi_0 \boldsymbol{\nu}_0}{\bar{N}_i + \xi_0} \Bigg) \Bigg(\boldsymbol{\mu}_i - \frac{\bar{N}_i \bar{\mathbf{x}}_i + \xi_0 \boldsymbol{\nu}_0}{\bar{N}_i + \xi_0} \Bigg)^{\mathrm{T}} \Bigg\} |\mathbf{S}_i|^{(-1/2)(\eta_0 + \bar{N}_i - d - 1)} \\
&\quad \times \exp\Bigg\{ -\frac{1}{2} \mathrm{Tr}\Bigg\{ \mathbf{S}_i \Bigg(\mathbf{B}_0 + \bar{\mathbf{C}}_i + \frac{\bar{N}_i \xi_0}{\bar{N}_i + \xi_0} (\bar{\mathbf{x}}_i - \boldsymbol{\nu}_0)(\bar{\mathbf{x}}_i - \boldsymbol{\nu}_0)^{\mathrm{T}} \Bigg) \Bigg\} \Bigg\} \Bigg\} \Bigg\} \\
&= \prod_{i=1}^{m} \underbrace{N\Big(\boldsymbol{\mu}_i | \bar{\boldsymbol{\mu}}_i (\bar{N}_i + \xi_0)^{-1} \mathbf{S}_i^{-1} \Big)}_{Q(\boldsymbol{\mu}_i | \mathbf{S}_{i,m})} \underbrace{W(\mathbf{S}_i | \eta_0 + \bar{N}_i, \mathbf{B}_i)}_{Q(\mathbf{S}_i | m)},
\end{aligned} \tag{B3}$$

where

$$\boldsymbol{\mu}_i = \frac{\bar{N}_i \bar{\mathbf{x}}_i + \xi_0 \boldsymbol{\nu}_0}{\bar{N}_i + \xi_0} \qquad \text{and} \qquad \mathbf{B}_i = \mathbf{B}_0 + \bar{\mathbf{C}}_i + \frac{\bar{N}_i \xi_0}{\bar{N}_i + \xi_0} (\bar{\mathbf{x}}_i - \boldsymbol{\nu}_0)(\bar{\mathbf{x}}_i - \boldsymbol{\nu}_0)^{\mathrm{T}}.$$

It follows that $Q(\mathbf{S}_i|m)$ is the Wishart distribution given by Eq. (36).

Next, $Q(\boldsymbol{\mu}_i|m)$ can be obtained by marginalizing $Q(\boldsymbol{\mu}_i, \mathbf{S}_i|m)$ w.r.t. $\mathbf{S}_i$. Namely

$$Q(\boldsymbol{\mu}_i|m) = \int Q(\boldsymbol{\mu}_i, \mathbf{S}_i|m) \mathrm{d}\mathbf{S}_i.$$

Here, $\int \cdot \mathrm{d}\mathbf{S}_i$ is understood to be w.r.t. the $d(d+1)/2$ distinct elements of the matrix $\mathbf{S}_i$. Note that the integral range is defined by over all possible values of positive-definite $d \times d$ matrix $\mathbf{S}_i$

$$\begin{aligned}
Q(\boldsymbol{\mu}_i|m) &\propto \int |\mathbf{S}_i|^{(1/2)(\eta_0 + \bar{N}_i - d)} \exp\Big\{ -\frac{1}{2} \Big\{ \mathbf{S}_i \Big((\bar{N}_i + \xi_0)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)^{\mathrm{T}} + \mathbf{B}_i \Big) \Big\} \Big\} \mathrm{d}\mathbf{S}_i \propto |(\bar{N}_i + \xi_0)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)^{\mathrm{T}} \\
&\quad + \mathbf{B}_i|^{(1/2)(\eta_0 + \bar{N}_i + 1)} \underbrace{\int \mathscr{W}\Big(\mathbf{S}_i | \eta_0 + \bar{N}_i + 1, (\bar{N}_i + \xi_0)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)^{\mathrm{T}} + \mathbf{B}_i \Big) \mathrm{d}\mathbf{S}_i}_{=1} \\
&\propto \Big| (\bar{N}_i + \xi_0)^{-1} \mathbf{B}_i + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)^{\mathrm{T}} \Big|^{(-1/2)(\eta_0 + \bar{N}_i + 1)} \\
&= \Big\{ 1 + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i)^{\mathrm{T}} (\bar{N}_i + \xi_0) \mathbf{B}_i^{-1} (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_i) \Big\}^{(-1/2)(\eta_0 + \bar{N}_i + 1)}.
\end{aligned} \tag{B4}$$

Note that the last line in Eq. (B4) is obtained via the formula that $|\mathbf{A} + \mathbf{a}\mathbf{a}^{\mathrm{T}}| = |\mathbf{A}|(1 + \mathbf{a}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{a})$ for $\mathbf{A}$ nonsingular. Thus, Eq. (B4) indicates that $Q(\boldsymbol{\mu}_i|m)$ becomes a d-dimensional Student's-T distribution:

$$Q(\boldsymbol{\mu}_i|m) = \mathscr{T}(\mu_i | \bar{\boldsymbol{\mu}}_i, \Sigma_{\boldsymbol{\mu}_i}, f_{\boldsymbol{\mu}_i}), \tag{B5}$$

where

$$f_{\boldsymbol{\mu}_i} = \eta_0 + \bar{N}_i + 1 - d \text{ and } \boldsymbol{\Sigma}_{\boldsymbol{\mu}_i} = \frac{\mathbf{B}_i}{(\bar{N}_i + \xi_0) f_{\boldsymbol{\mu}_i}}.$$

B.3. $Q(\boldsymbol{W}|m)$ and $Q(\boldsymbol{\beta}|m)$

In a similar way, using Eqs. (25) and (29)

$$\begin{aligned} Q(\mathbf{W}, \boldsymbol{\beta}|m) &\propto \exp\Big\{\langle \log p(\mathscr{D}, Z|\boldsymbol{\varphi}, \boldsymbol{\mu}, \mathbf{S}, \mathbf{W}, \boldsymbol{\beta}, m\rangle_{Q(Z|m), Q(\boldsymbol{\varphi}, \boldsymbol{\mu}, \mathbf{S}|m)} + \langle \log p(\mathbf{W}, \boldsymbol{\beta}|\boldsymbol{\alpha})\rangle_{Q(\boldsymbol{\alpha}|m)}\Big\} \\ &\propto \prod_{i=1}^{m} \exp\left\{\bar{N}_i \log \beta_i^{1/2} - \frac{\beta_i}{2}(Y - X\mathbf{w}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\mathbf{w}_i) + \log \beta_i^{((d+1)/2)} - \frac{\beta_i}{2} \mathbf{w}_i^{\mathrm{T}} \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)} \mathbf{w}_i + \log \beta_i^{\rho_0^{-1}} - \lambda_0 \beta_i \right\} \\ &= \prod_{i=1}^{m} \exp\left\{\log \beta_i^{(\rho_0 + (1/2)(\bar{N}_i + d + 1) - 1)} - \frac{\beta_i}{2}\Big\{(Y - X\mathbf{w}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\mathbf{w}_i) + \mathbf{w}_i^{\mathrm{T}} \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)} \mathbf{w}_i\Big\}\right\}. \end{aligned} \tag{B6}$$

where $\langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}$ will be specified in Section B.4.

Next, let $\hat{\mathbf{w}}_i$ be the generalized least-squares estimator minimizing $(Y - X\mathbf{w}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\mathbf{w}_i)$. That is

$$\hat{\mathbf{w}}_i = (X^{\mathrm{T}} \bar{\mathbf{V}}_i X)^{-1} X^{\mathrm{T}} \bar{\mathbf{V}}_i Y. \tag{B7}$$

By adding and subtracting $X\hat{\mathbf{w}}_i$, we get

$$\begin{aligned} (Y - X\mathbf{w}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\mathbf{w}_i) &= (Y - X\hat{\mathbf{w}}_i) - (X\mathbf{w}_i - X\hat{\mathbf{w}}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i ((Y - X\hat{\mathbf{w}}_i) - (X\mathbf{w}_i - X\hat{\mathbf{X}}_i)) \\ &= (Y - X\hat{\mathbf{w}}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\hat{\mathbf{w}}_i) + (\mathbf{w}_i - \hat{\mathbf{w}}_i)^{\mathrm{T}} X^{\mathrm{T}} \bar{\mathbf{V}}_i X (\mathbf{w}_i - \hat{\mathbf{w}}_i). \end{aligned} \tag{B8}$$

Note that using the orthogonality property of least-square estimator, say $X^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\hat{\mathbf{w}}_i) = \mathbf{0}$, the cross-product terms vanish.

Consequently, we get

$$\begin{aligned} (Y - X\mathbf{w}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\mathbf{w}_i) + \mathbf{w}_i^{\mathrm{T}} \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)} \boldsymbol{w}_i &= (Y - X\hat{\mathbf{w}}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\hat{\mathbf{w}}_i) + (\mathbf{w}_i - \hat{\mathbf{w}}_i)^{\mathrm{T}} X^{\mathrm{T}} \bar{\mathbf{V}}_i X (\mathbf{w}_i - \hat{\mathbf{w}}_i) + \mathbf{w}_i^{\mathrm{T}} \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)} \boldsymbol{w}_i \\ &= R_i + (\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}} (X^{\mathrm{T}} \bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}) (\mathbf{w}_i - \bar{\mathbf{w}}_i), \end{aligned} \tag{B9}$$

where

$$R_i = (Y - X\hat{\mathbf{w}}_i)^{\mathrm{T}} \bar{\mathbf{V}}_i (Y - X\hat{\mathbf{w}}_i) + \hat{\mathbf{w}}_i^{\mathrm{T}} X^{\mathrm{T}} \bar{\mathbf{V}}_i X \Big(X^{\mathrm{T}} \bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}\Big)^{-1} \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)} \hat{\mathbf{w}}_i,$$

and

$$\bar{\mathbf{w}}_i = \Big(X^{\mathrm{T}} \bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}\Big)^{-1} X^{\mathrm{T}} \bar{\mathbf{V}}_i Y.$$

Substituting Eq. (B9) into (B6), we have

$$\begin{aligned} Q(\mathbf{W}, \boldsymbol{\beta}|m) \propto \prod_{i=1}^{m} \Bigg[\beta_i^{(d+1)/2} \exp\Bigg\{ &- \frac{1}{2} \mathrm{Tr}\Big\{\beta_i \Big(X^{\mathrm{T}} \bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}\Big)(\mathbf{w}_i - \bar{\mathbf{w}}_i)(\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}}\Big\}\Bigg\} \beta_i^{(\rho_0 + (\bar{N}_i/2) - 1)} \\ &\times \exp\Bigg\{ - \beta_i \Big(\lambda_0 + \frac{R_i}{2}\Big)\Bigg\}\Bigg], \end{aligned} \tag{B10}$$

$$Q(\mathbf{W}, \boldsymbol{\beta}|m) = \prod_{i=1}^{m} \underbrace{\mathscr{N}\Big(\mathbf{w}_i|\bar{\mathbf{w}}_i, \beta_i^{-1}\Big(X^{\mathrm{T}} \bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}\Big)^{-1}\Big)}_{Q(\mathbf{w}_i|\beta_i, m)} \underbrace{\mathscr{G}\Big(\beta_i|\rho_0 + \tfrac{\bar{N}_i}{2}, \lambda_0 + \tfrac{R_i}{2}\Big)}_{Q(\beta_i|m)}. \tag{B11}$$

Hence, $Q(\beta_i|m)$ is the gamma distribution:

$$Q(\beta_i|m) = \mathcal{G}\left(\beta_i|\rho_0 + \frac{\bar{N}_i}{2}, \lambda_0 + \frac{R_i}{2}\right). \tag{B12}$$

Next, $Q(\mathbf{w}_i|m)$ can be computed by marginalizing $Q(\mathbf{w}_i, \beta_i|m)$ w.r.t. $\mathbf{w}_i$

$$\begin{aligned} Q(\mathbf{w}_i|m) &= \int Q(\mathbf{w}_i, \beta_i|m)\mathrm{d}\beta_i \propto \left\{\lambda_0 + \frac{R_i}{2} + \frac{1}{2}(\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}}\left(X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)(\mathbf{w}_i - \bar{\mathbf{w}}_i)\right\}^{(1/2)(2\rho_0+\bar{N}_i+d+1)} \\ &\times \int \mathcal{G}\left(\beta_i|\rho_0 + \frac{1}{2}(\bar{N}_i + d + 1), \lambda_0 + \frac{R_i}{2} + \frac{1}{2}(\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}}\left(X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)^{-1}(\mathbf{w}_i - \bar{\mathbf{w}}_i)\right)\mathrm{d}\beta_i \\ &\propto \left\{1 + (\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}}\frac{\left(X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)}{(2\lambda_0 + R_i)}(\mathbf{w}_i - \bar{\mathbf{w}}_i)\right\}^{-(1/2)(2\rho_0+\bar{N}_i+d+1)}. \end{aligned} \tag{B13}$$

Eq. (B13) indicates that as shown in Eq. (40), $Q(\mathbf{w}_i|m)$ is a multivariate Student's-T distribution with $f_{\mathbf{w}_i} = 2\rho_0 + \bar{N}_i$ degrees of freedom, mode

$$\bar{\mathbf{w}}_i = \left(X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)^{-1} X^{\mathrm{T}}\bar{\mathbf{V}}_i Y, \tag{B14}$$

and scale parameter

$$\Sigma_{\mathbf{w}_i} = \left(\frac{2\lambda_0 + R_i}{2\rho_0\bar{N}_i}\right)\left(X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i\rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)^{-1}. \tag{B15}$$

B.4. $Q(\boldsymbol{\alpha}|m)$

Applying Eq. (8) to $\boldsymbol{\alpha}$, we get

$$Q(\boldsymbol{\alpha}|m) \propto p(\boldsymbol{\alpha}|m)\exp\left\{\langle \log p(\mathbf{W}, \boldsymbol{\beta}|\boldsymbol{\alpha}, m)\rangle_{Q(\mathbf{W}|m),Q(\boldsymbol{\beta}|m)}\right\}. \tag{B16}$$

Here, since $p(\boldsymbol{\alpha}_i|m)$ is a gamma distribution

$$p(\boldsymbol{\alpha}|m) \propto \prod_{i=1}^{m}\prod_{i=1}^{d+1}\mathcal{G}(\alpha_{,j}|\kappa_0, \zeta_0) \propto \prod_{i=1}^{m}\prod_{j=1}^{d+1}\alpha_{i,j}^{\kappa_0 - 1}\exp\{-\zeta_0\alpha_{1,j}\}. \tag{B17}$$

On the other hand, using the results of $Q(\mathbf{w}_i|m)$ and $Q(\beta_i|m)$

$$\langle \log p(\mathbf{W}, \boldsymbol{\beta}|\boldsymbol{\alpha}, m)\rangle_{Q(\mathbf{W}|m),Q(\boldsymbol{\beta}|m)} \propto \sum_{i=1}^{m}\langle \log p(\mathbf{w}_i|\beta_i, \boldsymbol{\alpha}_i)\rangle_{Q(\mathbf{w}_i|m),Q(\beta_i|m)} \propto \sum_{i=1}^{m}\left\{\log\prod_{j=1}^{d+1}\alpha_{1,j}^{1/2} - \frac{1}{2}\langle \beta_i\rangle_{Q(\beta_i|m)}\left\langle \mathbf{w}_i^{\mathrm{T}}\Lambda_i\mathrm{w}_i\right\rangle_{Q(\mathbf{w}_i|m)}\right\}. \tag{B18}$$

Here, since $\langle \beta_i\rangle_{Q(\beta_i|m)}$ is the mean of the gamma distribution given by Eq. (B12), we easily have

$$\langle \beta_i\rangle_{Q(\beta_i|m)} = \frac{2\rho_0 + \bar{N}_i}{2\lambda_0 + \bar{N}_i}. \tag{B19}$$

Moreover, adding and subtracting the mean vector $\bar{\mathbf{w}}_i$ of the distribution $Q(\mathbf{w}_i|m)$

$$\begin{aligned} \left\langle \mathbf{w}_i^{\mathrm{T}}\Lambda_i\mathbf{w}_i\right\rangle_{Q(\mathbf{w}_i|m)} &= \mathrm{Tr}\left\{\Lambda_i\left\langle \mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}\right\rangle_{Q(\mathbf{w}_i|m)}\right\} = \mathrm{Tr}\left\{\Lambda_i\left(\left\langle (\mathbf{w}_i - \bar{\mathbf{w}}_i)(\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}}\right\rangle_{Q(\mathbf{w}_i|m)} + \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^{\mathrm{T}}\right)\right\} = \mathrm{Tr}\left\{\Lambda_i\left(\frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i} - 2}\Sigma_{\mathbf{w}_i} + \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^{\mathrm{T}}\right)\right\} \\ &= \sum_{j=1}^{d+1}\alpha_{i,j}\left\{\frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i} - 2}(\Sigma_{\mathbf{w}_i})_{j,j} + \bar{\mathbf{w}}_{i,j}^2\right\}. \end{aligned} \tag{B20}$$

Note that since $\bar{\mathbf{w}}_i$ is the mean vector of the random vector $\mathbf{w}_i$ following the Student's-T distribution given by Eqs. (40) and (41),

the term $\langle(\mathbf{w}_i - \bar{\mathbf{w}}_i)(\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}}\rangle_{Q(\mathbf{w}_i|m)}$ corresponds to the variance of $\mathbf{w}_i$. The notation $(\Sigma_{\mathbf{w}_i})_{j,j}$ represents the jth diagonal element of $\Sigma_{\mathbf{w}_i}$.

Substituting Eqs. (B17)–(B20) into Eq. (B16)

$$\begin{aligned} Q(\boldsymbol{\alpha}|m) &= \prod_{i=1}^{m} Q(\boldsymbol{\alpha}_i|m) \\ &= \prod_{i=1}^{m}\prod_{i=1}^{d+1} \alpha_{i,j}^{\kappa_0+(1/2)-1} \exp\Bigg\{ -\alpha_{i,j} \\ &\quad \times\left(\zeta_0 + \frac{1}{2}\left(\frac{2\rho_0+\bar{N}_i}{2\lambda_0+\bar{N}_i}\right)\left(\frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i}-2}(\Sigma_{\mathbf{w}_i})_{j,j} + \bar{w}_{i,j}^2\right)\right)\Bigg\} \\ &= \prod_{i=1}^{m}\prod_{j=1}^{d+1} Q(\alpha_{i,j}|m). \end{aligned} \tag{B21}$$

It follows that

$$Q(\alpha_{i,j}|m) = \mathcal{G}\left(\alpha_{i,j}|\kappa_0 + \frac{1}{2}, \zeta_{i,j}\right), \tag{B22}$$

where

$$\zeta_{i,j} = \zeta_0 + \frac{1}{2}\left(\frac{2\rho_0+\bar{N}_i}{2\lambda_0+\bar{N}_i}\right)\left(\frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i}-2}(\Sigma_{\mathbf{w}_i})_{j,j} + \bar{w}_{i,j}^2\right).$$

Since we have found that $\alpha_{i,j}$ follows a gamma distribution, its mean is easily computed as

$$\begin{aligned} \langle\Lambda_i\rangle_{Q(\alpha_{i,j}|m)} &= \mathrm{diag}(\langle\alpha_{i,1}\rangle_{Q(\alpha_{i,1}|m)}, \ldots, \langle\alpha_{i,d+1}\rangle_{Q(\alpha_{i,d+1}|m)}) \\ &= \mathrm{diag}\left(\frac{2\zeta_{i,1}}{2\kappa_0+1}\ldots, \frac{2\zeta_{i,d+1}}{2\kappa_0+1}\right). \end{aligned} \tag{B23}$$

B.5. $Q(Z|m)$

Using Eqs. (6) and (24)

$$\begin{aligned} Q(Z|m) &\propto \exp\{\langle\log p(\mathscr{D}, Z|\boldsymbol{\varphi}, \boldsymbol{\mu}, \mathbf{S}, \mathbf{W}, \boldsymbol{\beta}, m)\rangle_{Q(\boldsymbol{\mu},\mathbf{S},\mathbf{W},\boldsymbol{\beta}|m)}\} \\ &= \prod_{i=1}^{m}\prod_{n=1}^{N} \exp\Bigg\{ z_i^n \Bigg(\langle\log\varphi_i\rangle_{Q(\varphi_i|m)} + \frac{1}{2}\langle\log|\mathbf{S}_i|\rangle_{Q(\mathbf{S}_i|m)} \\ &\quad -\frac{1}{2}\mathrm{Tr}\{\langle\mathbf{S}_i\rangle_{Q(\mathbf{S}_i|m)}\langle(\mathbf{x}_n-\boldsymbol{\mu}_i)(\mathbf{x}_n-\boldsymbol{\mu}_i)^{\mathrm{T}}\rangle_{Q(\boldsymbol{\mu}_i|m)}\} \\ &\quad +\frac{1}{2}\langle\log\beta_i\rangle_{Q(\beta_i|m)} - \frac{1}{2}\langle\beta_i\rangle_{Q(\beta_i|m)} \\ &\quad \times\left\langle\left(y_n - \mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}_n\right)^2\right\rangle_{Q(\mathbf{w}_i|m)}\Bigg)\Bigg\}. \end{aligned} \tag{B24}$$

Here, since $\mathbf{S}_i$ and β_i follow the Wishart and gamma distributions, respectively, we easily compute their mean values as follows:

$$\langle\mathbf{S}_i\rangle_{Q(\mathbf{S}_i|m)} = (\eta_0 + \bar{N}_i)\mathbf{B}_i^{-1}, \tag{B25}$$

$$\langle\beta_i\rangle_{Q(\beta_i|m)} = \frac{2\rho_0+\bar{N}_i}{2\lambda_0+R_i}. \tag{B26}$$

Moreover, $\langle(\mathbf{x}_n-\boldsymbol{\mu}_i)(\mathbf{x}_n-\boldsymbol{\mu}_i)^{\mathrm{T}}\rangle_{Q(\boldsymbol{\mu}_i|m)}$ and $\langle(y_n - \mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}_n)^2\rangle_{Q(\mathbf{w}_i|m)}$ are computed below

$$\begin{aligned} \left\langle(\mathbf{x}_n-\boldsymbol{\mu}_i)(\mathbf{x}_n-\boldsymbol{\mu}_i)^{\mathrm{T}}\right\rangle_{Q(\boldsymbol{\mu}_i|m)} &= \underbrace{\left\langle(\boldsymbol{\mu}_i-\bar{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i-\bar{\boldsymbol{\mu}}_i)^{\mathrm{T}}\right\rangle_{Q(\boldsymbol{\mu}_i|m)}}_{\mathrm{Var}\{\boldsymbol{\mu}_i\}} \\ &\quad + (\mathbf{x}_n-\bar{\boldsymbol{\mu}}_i)(\mathbf{x}_n-\bar{\boldsymbol{\mu}}_i)^{\mathrm{T}} \\ &= \frac{f_{\boldsymbol{\mu}_i}}{f_{\boldsymbol{\mu}_i}-2}\Sigma_{\boldsymbol{\mu}_i} + (\mathbf{x}_n-\bar{\boldsymbol{\mu}}_i)(\mathbf{x}_n-\bar{\boldsymbol{\mu}}_i)^{\mathrm{T}}, \end{aligned} \tag{B27}$$

$$\begin{aligned} &\left\langle\left(y_n - \mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}_n\right)^2\right\rangle_{Q(\mathbf{w}_i|m)} \\ &= \left\langle\left(y_n - \tilde{\mathbf{x}}_n^{\mathrm{T}}\mathbf{w}_i\right) - \tilde{\mathbf{x}}_n^{\mathrm{T}}(\mathbf{w}_i-\bar{\mathbf{w}}_i)^2\right\rangle_{Q(\mathbf{w}_i|m)} \\ &= \left(y_n - \tilde{\mathbf{x}}^{\mathrm{T}}\bar{\mathbf{w}}_i\right)^2 + \tilde{\mathbf{x}}_n^{\mathrm{T}}\underbrace{\left\langle(\mathbf{w}_i-\bar{\mathbf{w}}_i) - (\mathbf{w}_i-\bar{\mathbf{w}}_i)^{\mathrm{T}}\right\rangle_{Q(\mathbf{w}_i|m)}}_{\mathrm{Var}\{\mathbf{w}_i\}} \\ &= \left(y_n - \tilde{\mathbf{x}}^{\mathrm{T}}\bar{\mathbf{w}}_i\right)^2 + \frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i}-2}\tilde{\mathbf{x}}^{\mathrm{T}}\Sigma_{\mathbf{w}_i}\tilde{\mathbf{x}}. \end{aligned} \tag{B28}$$

The other special expetations such as $\langle\log\varphi_i\rangle_{Q(\varphi_i|m)}$ and $\langle\log|\mathbf{S}_i|\rangle_{Q(\mathbf{S}_i|m)}$ can be computed as follows:

$$\begin{aligned} &\langle\log\varphi_i\rangle_{Q(\{\varphi_i\}_{i=1}^m|m)} \\ &= \frac{\Gamma\left(\sum_{j=1}^{m}(\delta_0+\bar{N}_j)\right)}{\prod_{j=1}^{m}\Gamma(\delta_0+\bar{N}_j)}\int_{\varphi_1}\cdots\int_{\varphi_m}(\log\varphi_i)\prod_{j=1}^{m}\varphi_j^{\delta_0-1}\mathrm{d}\varphi_1\cdots\mathrm{d}\varphi_m. \end{aligned} \tag{B29}$$

Note that the integral in Eq. (B28) is performed so as to satisfy $\sum_{i=1}^m \varphi_i = 1$. On the other hand, since

$$\int_{\varphi_1}\cdots\int_{\varphi_m}\mathscr{D}(\{\varphi_i\}_{i=1}^m|\{\delta_0+\bar{N}_i\}_{i=1}^m)\mathrm{d}\varphi_1\cdots\mathrm{d}\varphi_m = 1,$$

we get

$$\int_{\varphi_1}\cdots\int_{\varphi_m}\prod_{j=1}^{m}\varphi_j^{\delta_0-1}\mathrm{d}\varphi_1\cdots\mathrm{d}\varphi_m = \frac{\prod_{j=1}^{m}\Gamma(\delta_0+\bar{N}_i)}{\Gamma\left(\sum_{j=1}^{m}(\delta_0+\bar{N}_j)\right)}. \tag{B30}$$

To make the form of the R.H.S. of Eq. (B29), we differentiate both sides of Eq. (B30) w.r.t. $\bar{N}_i$. It follows

that

$$\int_{\varphi_1}\cdots\int_{\varphi_m}(\log\varphi_i)\prod_{j=1}^{m}\varphi_j^{\delta_0+\bar{N}_j-1}\mathrm{d}\varphi_1\cdots\mathrm{d}\varphi_m = \frac{\prod_{j=1}^{m}\Gamma(\delta_0+\bar{N}_j)}{\Gamma\left(\sum_{j=1}^{m}(\delta_0+\bar{N}_j)\right)}\left(\Psi(\delta_0+\bar{N}_i)-\Psi\left(m\delta_0+\sum_{j=1}^{m}\bar{N}_j\right)\right), \tag{B31}$$

where $\Psi(\cdot)$ is the digamma function defined by

$$\Psi(\mathbf{x}) = \frac{\partial\log\Gamma(\mathbf{x})}{\partial\mathbf{x}}$$

Substituting Eq. (B31) into Eq. (B29)

$$\langle\log\varphi_i\rangle_{Q(\{\varphi_i\}_{i=1}^m|m)} = \Psi(\delta_0+\bar{N}_i)-\Psi\left(m\delta_0+\sum_{j=1}^{m}\bar{N}_j\right). \tag{B32}$$

In a similar manner, since

$$\int_0^\infty \mathcal{G}(\beta_i|\rho_i,\lambda_i)\mathrm{d}\beta_i = 1,$$

where $\rho_i = \rho_0 + (\bar{N}_i/2)$ and $\lambda_i = \lambda_0 + (R_i/2)$, we have

$$\frac{\partial}{\partial\rho_i}\int_0^\infty \beta_i^{\rho_i-1}\,\mathrm{e}^{-\lambda_i\beta_i}\mathrm{d}\beta_i = \frac{\partial}{\partial\rho_i}\left(\frac{\Gamma(\rho_i)}{\lambda_i^{\rho_i}}\right). \tag{B33}$$

It follows that

$$\int(\log\beta_i)\beta_i^{\rho_i-1}\mathrm{e}^{-\lambda_i\beta_i}\mathrm{d}\beta_i = \frac{1}{\lambda_i^{\rho_i}}\left(\frac{\partial\Gamma(\rho_i)}{\partial\rho_i}-\Gamma(\rho_i)\log\lambda_i\right). \tag{B34}$$

Using Eq. (B34), we get

$$\langle\log\beta_i\rangle_{Q(\beta_i|m)} = \frac{\lambda_i}{\Gamma(\rho_i)}\int(\log\beta_i)\beta_i^{\rho_i-1}\mathrm{e}^{-\lambda_i\beta_i}\mathrm{d}\beta_i = \Psi(\rho_i)-\log\lambda_i. \tag{B35}$$

Similarly, since

$$\int\mathcal{W}(\mathbf{S}_i|\eta_i,\mathbf{B}_i)\mathrm{d}S_i = 1,$$

where $\eta_i = \eta_0 + \bar{N}_i$, we have

$$\frac{\partial}{\partial\eta_i}\int|\mathbf{S}_i|^{(1/2)(\eta_i-d+1)}\exp\{-\mathrm{Tr}\{\mathbf{B}_i\mathbf{S}_i\}\}\mathrm{d}\mathbf{S}_i = \frac{\partial}{\partial\eta_i}\left(\frac{2^{\eta_i d/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{\eta_i+1-j}{2}\right)}{|\mathbf{B}_i|^{\eta_i/2}}\right). \tag{B36}$$

It follows that

$$\int(\log|\mathbf{S}_i|)|\mathbf{S}_i|^{(1/2)(\eta_i-d-1)}\exp\left\{-\frac{1}{2}\mathrm{Tr}\{\mathbf{B}_i\mathbf{S}_i\}\right\}\mathrm{d}\mathbf{S}_i = \frac{2^{\eta_i d/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{\eta_i+1-j}{2}\right)}{|\mathbf{B}_i|^{\eta_i/2}}\times\left\{d\log 2-\log|\mathbf{B}_i|+\sum_{j=1}^{d}\Psi\left(\frac{\eta_i+1-j}{2}\right)\right\}. \tag{B37}$$

Thus,

$$\langle\log|\mathbf{S}_i|\rangle_{Q(\mathbf{S}_i|m)} = d\log 2-\log|\mathbf{B}_i|+\sum_{j=1}^{d}\Psi\left(\frac{\eta_i+1-j}{2}\right). \tag{B38}$$

Finally, substituting Eqs. (B25)–(B28), (B32), (B35) and (B38) into (B24), we get

$$Q(Z|m)\propto\prod_{i=1}^{m}\prod_{n=1}^{N}\exp\{z_i^n\gamma_i^n\}, \tag{B39}$$

where

$$\begin{aligned}\gamma_i^n = {}&\Psi(\delta_0+\bar{N}_i)-\Psi\left(m\delta_0+\sum_{i=1}^{m}\bar{N}_i\right)\\ &+\frac{1}{2}\sum_{j=1}^{d}\Psi\left(\frac{\eta_0+\bar{N}_i+1-j}{2}\right)-\frac{1}{2}\log|\mathbf{B}_i|\\ &-\frac{1}{2}\mathrm{Tr}\left\{\left(\eta_0+\bar{N}_i\right)\mathbf{B}_i^{-1}\right.\\ &\left.\times\left(\frac{f_{\boldsymbol{\mu}_i}}{f_{\boldsymbol{\mu}_i}-2}\boldsymbol{\Sigma}_{\boldsymbol{\mu}_i}+(\mathbf{x}_n-\bar{\boldsymbol{\mu}}_i)(\mathbf{x}_n-\bar{\boldsymbol{\mu}}_i)^{\mathrm{T}}\right)\right\}\\ &+\frac{1}{2}\left(\Psi\left(\rho_0+\frac{\bar{N}_i}{2}\right)-\log\left(\lambda_0+\frac{R_i}{2}\right)\right)\\ &-\frac{1}{2}\left(\frac{2\rho_0+\bar{N}_i}{2\lambda_0+R_i}\right)\left\{\left(y_n-\tilde{\mathbf{x}}_n^{\mathrm{T}}\bar{\mathbf{w}}_i\right)^2+\frac{f_{\mathbf{w}_i}}{f_{\mathbf{w}_i}-2}\tilde{\mathbf{x}}_n^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{w}_i}\tilde{\mathbf{x}}_n\right\}.\end{aligned} \tag{B40}$$

Thus,

$$\bar{z}_i^n = Q(z_i^n=1|m) = \frac{\exp\{\gamma_i^n\}}{\sum_{j=1}^{m}\exp\{\gamma_j^n\}}.$$

Appendix C. Derivation of the predictive posterior distribution

Setting the integration of the R.H.S. of Eq. (49) to I and using the result that $Q(\mathbf{w}_i,\beta_i|m)$ is the normal–gamma

distribution (Eq. (B10)), we get

$$I = \int \mathcal{N}\left(y|\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}, \beta_i^{-1}\right) Q(\mathbf{w}_i, \beta_i|m)\mathrm{d}\mathbf{w}_i\,\mathrm{d}\beta_i$$
$$\propto \int \beta_i^{(1/2)(2\rho_0+\bar{N}_i+d)} \exp\left\{ -\frac{\beta_i}{2}\left[2\lambda_0 + R_i + (\mathbf{w}-\bar{\mathbf{w}}_i)^{\mathrm{T}} \times \mathbf{K}_i(\mathbf{w}-\bar{\mathbf{w}}_i) + \left(y_{N+1} - \mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}\right)^2\right]\right\}\mathrm{d}\mathbf{w}_i\,\mathrm{d}\beta_i, \quad \text{(C1)}$$

where $\mathbf{K}_i = X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}$. On the other hand

$$(\mathbf{w}_i - \bar{\mathbf{w}}_i)^{\mathrm{T}}\mathbf{K}_i(\mathbf{w}_i - \bar{\mathbf{w}}_i) + \left(y_{N+1} - \mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}\right)^2$$
$$= \mathbf{w}_i^{\mathrm{T}}\left(\mathbf{K}_i + \tilde{\mathbf{x}}_{N+1}\tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\right)\mathbf{w}_i - 2\mathbf{w}_i^{\mathrm{T}}(\mathbf{K}_i\bar{\mathbf{w}}_i + \tilde{\mathbf{x}}_{N+1}y_{N+1}) + y_{N+1}^2 + \bar{\mathbf{w}}_i^{\mathrm{T}}\mathbf{K}_i\bar{\mathbf{w}}_i$$
$$= (\mathbf{w}_i - \tilde{\mathbf{w}}_i)^{\mathrm{T}}\mathbf{M}_i(\mathbf{w}_i - \tilde{\mathbf{w}}_i) + U_i(y_{N+1}), \quad \text{(C2)}$$

where

$$\mathbf{M}_i = \mathbf{K}_i + \tilde{\mathbf{x}}_{N+1}\tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}, \quad \text{(C3)}$$

$$\bar{\mathbf{w}}_i = \mathbf{M}_i^{-1}(\mathbf{K}_i\bar{\mathbf{w}}_i + \tilde{\mathbf{x}}_{N+1}y_{N+1}) = \mathbf{M}_i^{-1}\left(X^{\mathrm{T}}\bar{\mathbf{V}}_i Y + \tilde{\mathbf{x}}_{N+1}y_{N+1}\right), \quad \text{(C4)}$$

$$U_i(y_{N+1}) = y_{N+1}^2 + \bar{\mathbf{w}}_i^{\mathrm{T}}\mathbf{K}_i\bar{\mathbf{w}}_i - \tilde{\mathbf{w}}_i^{\mathrm{T}}\mathbf{M}_i\tilde{\mathbf{w}}_i. \quad \text{(C5)}$$

Substituting Eq. (C2) into (C1) and rearranging $\mathbf{w}_i$ and β_i terms, we have

$$I \propto \underbrace{\int \mathcal{N}\left(\mathbf{w}_i|\bar{\mathbf{w}}_i, \beta_i^{-1}\mathbf{M}_i^{-1}\right)\mathrm{d}\mathbf{w}_i}_{=1} \int \beta_i^{(1/2)(2\rho_0+\bar{N}_i+1)} \times \exp\left\{ -\frac{\beta_i}{2}(2\lambda_0 + R_i + U_i(y_{N+1}))\right\}\mathrm{d}\beta_i$$
$$= \underbrace{\int \mathcal{G}\left(\beta_i|\frac{1}{2}(2\rho_0 + \bar{N}_i + 1), \frac{1}{2}(2\lambda_0 + R_i + U_i(y_{N+1}))\right)\mathrm{d}\beta_i}_{=1} \times \left\{\frac{1}{2}(2\lambda_0 + R_i + U_i(y_{N+1}))\right\}^{-(1/2)(2\rho_0+\bar{N}_i+1)}$$
$$\propto (2\lambda_0 + R_i + U_i(y_{N+1}))^{-(1/2)(2\rho_0+\bar{N}_i+1)}. \quad \text{(C6)}$$

Next, to obtain the distribution of y_{N+1}, arranging $U_i(y_{N+1})$ w.r.t. y_{N+1}

$$U_i(y_{N+1}) = y_{N+1}^2 + \bar{\mathbf{w}}_i^{\mathrm{T}}\mathbf{K}_i\bar{\mathbf{w}}_i - \left(X^{\mathrm{T}}\bar{\mathbf{V}}_i Y + \tilde{\mathbf{x}}_{N+1}y_{N+1}\right)^{\mathrm{T}} \times \mathbf{M}_i^{-1}\left(X^{\mathrm{T}}\bar{\mathbf{V}}_i Y + \tilde{\mathbf{x}}_{N+1}y_{N+1}\right)$$
$$= \left(1 - \tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\mathbf{M}_i^{-1}\tilde{\mathbf{x}}_{N+1}\right)(y_{N+1} - \bar{y}_i)^2, \quad \text{(C7)}$$

where

$$\bar{y}_i = \left(1 - \tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\mathbf{M}_i^{-1}\tilde{\mathbf{x}}_{N+1}\right)^{-1}\tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\mathbf{M}_i^{-1}X^{\mathrm{T}}\bar{\mathbf{V}}_i Y$$
$$= \tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\underbrace{\left(X^{\mathrm{T}}\bar{\mathbf{V}}_i X + \langle \Lambda_i \rangle_{Q(\boldsymbol{\alpha}_i|m)}\right)^{-1}X^{\mathrm{T}}\bar{\mathbf{V}}_i Y}_{\bar{\mathbf{w}}_i} \equiv \bar{\mathbf{w}}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}. \quad \text{(C8)}$$

Thus

$$I \propto \left\{2\lambda_0 + R_i\left(1 - \tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\mathbf{M}_i^{-1}\tilde{\mathbf{x}}_{N+1}\right) \times \left(y_{N+1} - \bar{\mathbf{w}}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}\right)^2\right\}^{-(1/2)(2\rho_0+\bar{N}_i+1)}$$
$$\propto \left\{1 + \frac{1 - \tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\mathbf{M}_i^{-1}\tilde{\mathbf{x}}_{N+1}}{2\lambda_0 + R_i} \times \left(y_{N+1} - \bar{\mathbf{w}}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}\right)^2\right\}^{-(1/2)(2\rho_0+\bar{N}_i+1)}$$
$$= \mathcal{T}\left(y_{N+1}| - \bar{\mathbf{w}}_i^{\mathrm{T}}\tilde{\mathbf{x}}_{N+1}, \sigma_i, 2\rho_0 + \bar{N}_i\right), \quad \text{(C9)}$$

where the scale paramter σ_i is given by

$$\sigma_i = \frac{2\lambda_0 + R_i}{(2\rho_0 + \bar{N}_i)\left(1 - \tilde{\mathbf{x}}_{N+1}^{\mathrm{T}}\mathbf{M}_i^{-1}\tilde{\mathbf{x}}_{N+1}\right)}.$$

From these results, we obtain Eq. (49).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on AC, 19-6*, 716–723.

Attias, H. (1999). Learning parameters and structure of latent variable models by variational Bayes. *Proceedings of Uncertainty in Artificial Intelligence (UAI)*.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Bishop, C. M., & Bayesian, P. C. A. (1999). Advances in neural information processing systems. *MIT Press*, *NIPS11*, 382–388.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Gamerman, D. (1997). *Markov chain Monte Carlo*. London: Chapman & Hall.

Ghahramani, Z., & Beal, M. J. (2000). Variational inference for Bayesian mixture of factor analyzers. *Advances in Neural Information Processing Systems*, *NIPS12*, 449–455. MIT Press.

Ghahramani, Z., & Beal, M. J. (2001). Propagation algorithm for

variational Bayesian learning. *Advances in Neural Information Processing Systems*, *NIPS13*, 507–513. MIT Press.

Jaakkola, T. (1997). *Variational methods for inference and learning in graphical models*. PhD thesis, MIT Press.

Jebara, T., & Pentland, A. (2000). Maximum conditional likelihood via bound maximization and the CEM algorithm. *Advances in Neural Information Processing Systems*, *NIPS11*, 494–500.

Jebara, T., & Pentland, A. (2002). On reversing Jensen's inequality. *Advances in Neural Information Processing Systems*, *NIPS13* in press.

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

MacKay, D. J. C. (1992b). A practical Bayesian framework for back-propagation networks. *Neural Computation*, *4*, 448–472.

MacKay, D. J. C. (1994). Bayesian non-linear modeling for the prediction competition. *ASHRAE Transactions*, *100*, 1053–1062.

Neal, R. M. (1996). *Bayesian learning for neural networks* (*Vol. 118*). *Lecture Notes in Statistics*, Berlin: Springer.

Rasmussen, C. E., Neal, R. M., Hinton, G. E., Camp, D. van, Revow, M., Ghahramani, Z., Kustra, R., & Tibshirani, R. (1996). The DELVE Manual. http://www.cs.utoronto.ca/~delve/.

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. New York: Wiley.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, *39*, 44–47.

Takeuchi, K. (1983). On the selection of statistical models by AIC. *Journal of the Society of Instrument and Control Engineering*, *22*(5), 445–453. in Japanese.

Tipping, M. E. (2000). The relevance vector machine. *Advances in Neural Information Processing Systems*, *NIPS12*, 652–658.

Ueda, N. (2000). Variational Bayesian learning for optimal model search. *Journal of Japanese Society for Artificial Intelligence*, *16*(2), SP-F in Japanese.

Ueda, N., Nakano, R., Ghahramani, Z., & Hinton, G. E. (1999). SMEM algorithm for mixture models. *Advances in Neural Information Processing Systems*, *NIPS11*, 599–605.

Ueda, N., Nakano, R., Ghahramani, Z., & Hinton, G. E. (2000). SMEM algorithm for mixture models. *Neural Computation*, *12*(9), 2109–2128.

Waterhouse, S. R., MacKay, D., & Robinson, A. J. (1995). Bayesian methods for mixture of experts. *Advances in Neural Information Processing Systems*, *NIPS8*, 351–357.

Xu, L., Jordan, M. I., & Hinton, G. E. (1994). An alternative model for mixtures of experts. *Advances in Neural Information Processing Systems*, *NIPS7*, 633–640.