

# Variational Bayes for predictive distributions

Edward Snelson

Zoubin Ghahramani

Gatsby Computational Neuroscience Unit  
University College London  
17 Queen Square, London WC1N 3AR, UK  
{snelson,zoubin}@gatsby.ucl.ac.uk

February 17, 2005

## 1 Introduction

Variational Bayes (VB) [Beal, 2003] is a method for lower-bounding the evidence, or marginal likelihood, of a Bayesian latent variable model, which is then useful for model comparison. An approximation to the posterior distribution over model parameters is also obtained, and this can be used to obtain an approximation to the predictive distribution. However the variational approximation is optimized with respect to the evidence, rather than the predictive distribution. In this note we show how an alternative variational procedure can be derived that is tailored to the predictive distribution itself.

## 2 Variational Bayesian EM

Suppose we have a probabilistic model with an observed i.i.d. data set  $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ , corresponding latent variables  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , and parameters  $\theta$ . The variational Bayesian EM algorithm [Beal, 2003] forms a lower bound to the log marginal likelihood  $\log p(Y)$  by assuming a factorized variational distribution on the latent variables and parameters:  $q(X, \theta) = q_X(X)q_\theta(\theta)$ . The resulting EM-like updates take the following form:

$$\text{VBE step} \quad q_{\mathbf{x}_i}(\mathbf{x}_i) \propto \exp \left\langle \log p(\mathbf{x}_i, \mathbf{y}_i | \theta) \right\rangle_{q_\theta} \quad (1)$$

$$\text{VBM step} \quad q_\theta(\theta) \propto p(\theta) \exp \left\langle \log p(X, Y | \theta) \right\rangle_{q_X} . \quad (2)$$

Here the factorization  $q_X(X) = \prod_i q_{\mathbf{x}_i}(\mathbf{x}_i)$  arises from the i.i.d. property, not from any additional approximation.

## 2.1 Conjugate-exponential models

For a broad class of models, called the conjugate-exponential (CE) family, the VBEM updates are tractable and take on simple general forms. Conjugate-exponential models satisfy two conditions. First the complete data likelihood is in the exponential family:

$$p(\mathbf{x}_i, \mathbf{y}_i | \theta) = g(\theta) f(\mathbf{x}_i, \mathbf{y}_i) \exp [\phi(\theta)^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)] , \quad (3)$$

where  $\phi$  is a vector of natural parameters,  $\mathbf{u}$  is a vector of sufficient statistics, and  $g$  is a normalisation constant. Second the parameter prior is conjugate:

$$p(\theta | \eta, \nu) = h(\eta, \nu) g(\theta)^\eta \exp [\phi(\theta)^\top \nu] , \quad (4)$$

where  $\eta$  and  $\nu$  are hyperparameters and  $h$  is a normalisation constant.

The VBEM updates for CE models take the following form:

$$\text{VBE step} \quad q_{\mathbf{x}_i}(\mathbf{x}_i) = p(\mathbf{x}_i | \mathbf{y}_i, \tilde{\theta}) \quad (5)$$

$$\text{VBM step} \quad q_\theta(\theta) = p(\theta | \tilde{\eta}, \tilde{\nu}) , \quad (6)$$

where

$$\tilde{\theta} = \phi^{-1} \langle \phi(\theta) \rangle_{q_\theta} \quad (7)$$

$$\tilde{\eta} = \eta + n \quad (8)$$

$$\tilde{\nu} = \nu + \sum_{i=1}^n \langle \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i) \rangle_{q_{\mathbf{x}_i}} . \quad (9)$$

We see that  $q_{\mathbf{x}_i}$  remains in the exponential family and  $q_\theta$  remains conjugate.

## 3 Predictive variational Bayesian EM

The predictive distribution for a new observation  $\mathbf{y}'$  is:

$$p(\mathbf{y}' | Y) = \int d\theta p(\mathbf{y}' | \theta) p(\theta | Y) . \quad (10)$$

The problem with the standard variational scheme as described above is that even if we approximate the posterior  $p(\theta | Y)$  by  $q(\theta)$ , the integral is still intractable. In fact it is exactly like the marginal likelihood calculation because the approximate posterior is in the same conjugate family as the prior. Clearly we can lower bound this integral in a similar way to which we lower bounded the marginal likelihood. However we would obtain a separate lower bound for each value of  $\mathbf{y}'$ , and it is not clear that an accurate predictive distribution could be obtained from these bounds. An alternative, which is often used, is to simply predict from the model with the mean variational parameter:

$$p(\mathbf{y}' | Y) \approx p(\mathbf{y}' | \langle \theta \rangle) = \int d\mathbf{x}' p(\mathbf{x}', \mathbf{y}' | \langle \theta \rangle) . \quad (11)$$

We describe a new variational method that focuses its approximation on the predictive distribution, rather than on lower-bounding the marginal likelihood. We start by trying to minimize the KL divergence between an approximating predictive distribution  $q_{\mathbf{y}'}$  and the true distribution:

$$\begin{aligned}
& \min_{q_{\mathbf{y}'}} \text{KL}[q(\mathbf{y}')||p(\mathbf{y}'|Y)] \\
&= \max_{q_{\mathbf{y}'}} \int d\mathbf{y}' q(\mathbf{y}') \log \frac{\int d\theta p(\mathbf{y}'|\theta)p(\theta|Y)}{q(\mathbf{y}')} \quad (12) \\
&= \max_{q_{\mathbf{y}'}} \int d\mathbf{y}' q(\mathbf{y}') \log \frac{\int d\theta d\mathbf{x}' \prod_i d\mathbf{x}_i p(\theta)p(\mathbf{x}', \mathbf{y}'|\theta) \prod_i p(\mathbf{x}_i, \mathbf{y}_i|\theta)}{q(\mathbf{y}')} .
\end{aligned}$$

To deal with the integral inside the log we now introduce a factorized variational distribution over the corresponding variables:

$$q(\theta, \mathbf{x}', X|\mathbf{y}') = q(\theta)q(\mathbf{x}'|\mathbf{y}')q(X) , \quad (13)$$

and use Jensen's inequality to lower bound the above expression. Maximizing this lower bound with respect to all the variational distributions then gives the following updates:

$$\text{PVBE step} \quad q_{\mathbf{x}_i}(\mathbf{x}_i) \propto \exp \left\langle \log p(\mathbf{x}_i, \mathbf{y}_i|\theta) \right\rangle_{q_\theta} \quad (14)$$

$$q_{\mathbf{x}'\mathbf{y}'}(\mathbf{x}', \mathbf{y}') \propto \exp \left\langle \log p(\mathbf{x}', \mathbf{y}'|\theta) \right\rangle_{q_\theta} \quad (15)$$

$$\text{PVBM step} \quad q_\theta(\theta) \propto p(\theta) \exp \left[ \left\langle \log p(X, Y|\theta) \right\rangle_{q_X} + \left\langle \log p(\mathbf{x}', \mathbf{y}'|\theta) \right\rangle_{q_{\mathbf{x}'\mathbf{y}'}} \right] \quad (16)$$

Notice the similarity between these updates and the standard VBEM updates. To obtain the final predictive distribution we simply marginalize the joint variational distribution  $q_{\mathbf{x}'\mathbf{y}'}$ .

### 3.1 PVBEM for CE models

As for VBEM, the particular case of CE models leads to simple update rules:

$$\text{VBE step} \quad q_{\mathbf{x}_i}(\mathbf{x}_i) = p(\mathbf{x}_i|\mathbf{y}_i, \tilde{\theta}) \quad (17)$$

$$q_{\mathbf{x}'\mathbf{y}'}(\mathbf{x}', \mathbf{y}') = p(\mathbf{x}', \mathbf{y}'|\tilde{\theta}) \quad (18)$$

$$\text{VBM step} \quad q_\theta(\theta) = p(\theta|\tilde{\eta}, \tilde{\nu}) , \quad (19)$$

where

$$\tilde{\theta} = \phi^{-1} \left\langle \phi(\theta) \right\rangle_{q_\theta} \quad (20)$$

$$\tilde{\eta} = \eta + n + 1 \quad (21)$$

$$\tilde{\nu} = \nu + \sum_{i=1}^n \left\langle \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i) \right\rangle_{q_{\mathbf{x}_i}} + \left\langle \mathbf{u}(\mathbf{x}', \mathbf{y}') \right\rangle_{q_{\mathbf{x}'\mathbf{y}'}} . \quad (22)$$

Again, the final predictive distribution is formed by marginalising:

$$q(\mathbf{y}') = \int d\mathbf{x}' p(\mathbf{x}', \mathbf{y}' | \tilde{\theta}) . \quad (23)$$

We see that this distribution has exactly the same form as (11). However the PVBEM update rules tell us that we should do our parameter averaging in *natural* parameter space, and the parameter distribution  $q_\theta$  includes a contribution from the predictive point itself.

## 4 Conclusions

For some simple models in the CE family, we tried comparing this predictive distribution to the one obtained from standard VB. We found very little difference between the two approaches. This is probably because with a reasonable quantity of data the effect of the predictive point in the approximation is negligible. However the framework is still interesting from a theoretical point of view, and the approximation may prove more accurate in certain cases, particularly when only a very small number of data points are available.

## References

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.