Variational Bayes for 1-dimensional Mixture Models

W.D. Penny and S.J. Roberts,
Technical Report PARG-2000-01,
Department of Engineering Science,
Oxford University.

November 3, 2000

Abstract

We describe a Variational Bayesian (VB) learning algorithm for 1-dimensional mixture models. VB model order selection is compared to the Minimum Description Length (MDL) criterion and the algorithm is demonstrated on a number of synthetic data sets.

1 Introduction

Attias has recently [2] described a 'Variational Bayes framework for Graphical Models'. The key property of a Graphical Model in this context is that it has hidden variables. In this report we show how to apply the VB framework to the simplest possible such Graphical Model; a one-dimensional Gaussian Mixture [3] ie. where the data is one-dimensional (a single variable). The hidden variables are indicators which indicate which Gaussian generated which data point.

2 Mixture model

Below is a graph showing the probabilistic relationships between variables in a 1-dimensional Gaussian Mixture model with m-components. Component s has mixing coefficient π_s, mean μ_s and precision β_s. We can write the parameters collectively as the vectors $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_m]$, $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_m]$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_m]$ and can concatenate these into the overall parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}\}$. We are given a set of data points $D = \{y_n\}$ with $n = 1..N$. The likelihood of a data point y_n is given by the mixture model

$$p(y_n|\boldsymbol{\theta}) = \sum_{s=1}^{m} p(s_n = s|\boldsymbol{\pi}) p(y_n|\beta_s, \mu_s) \tag{1}$$

where s_n is an indicator variable indicating which component is selected for which data point. These are chosen probabilistically according to

$$p(s_n = s|\boldsymbol{\pi}) = \pi_s \tag{2}$$

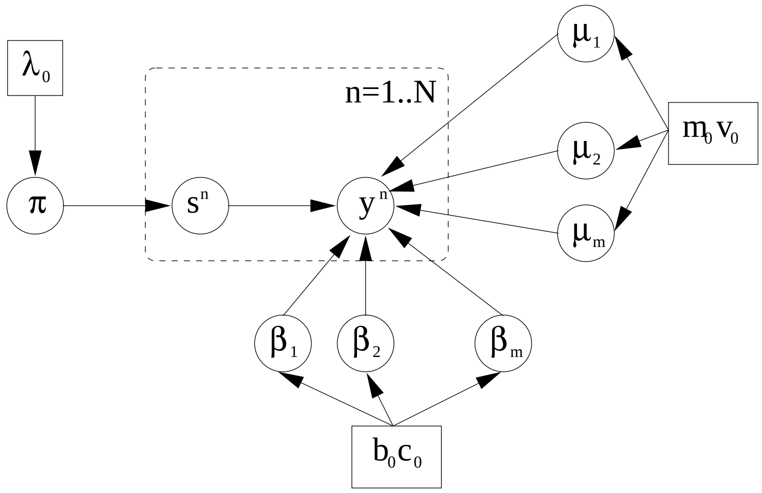

Figure 1: *A 1-dimensional Mixture model*

Each component is a Gaussian with

$$p(y_n|\beta_s, \mu_s) = (2\pi)^{-1/2} \beta_s^{1/2} \exp\left(\frac{-\beta_s}{2}(y_n - \mu_s)^2\right) \tag{3}$$

The prior on the model parameters is

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_s p(\beta_s) \prod_s p(\mu_s) \tag{4}$$

where the mixing prior is a symmetric Dirichlet

$$p(\boldsymbol{\pi}) = \frac{\Gamma(m\lambda_0)}{\Gamma(\lambda_0)^m} \prod_{s=1}^{m} \pi_s^{\lambda_0 - 1} \tag{5}$$

the prior over the precisions is a Gamma

$$p(\beta_s) = \Gamma(\beta_s; b_0, c_0) \tag{6}$$

and the prior over the means is a Normal

$$p(\mu_s) = N(\mu_s; m_0, v_0) \tag{7}$$

The joint likelihood of a data point and indicator variable is

$$p(y_n, s_n|\theta) = p(s_n = s|\boldsymbol{\pi}) p(y_n|\beta_s, \mu_s) \tag{8}$$

which, given that the data are IID, gives

$$p(\boldsymbol{Y}, \boldsymbol{S}|\theta) = \prod_{n=1}^{N} p(s_n = s|\boldsymbol{\pi}) p(y_n|\beta_s, \mu_s) \tag{9}$$

over the whole date set, where $\boldsymbol{Y} = [y_1, y_2, ..., y_N]$ and $\boldsymbol{S} = [s_1, s_2, ..., s_N]$.

3 Variational Bayes

We approximate the posterior distribution over parameters with the factorised density

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\pi})q(\boldsymbol{\beta})q(\boldsymbol{\mu}) \tag{10}$$

and the posterior distribution over hidden variables by $q(\boldsymbol{S})$. The Variational Bayes (VB) framework for models with hidden variables maximises the negative free energy

$$F_m =< \log \frac{p(\boldsymbol{Y}, \boldsymbol{S}|\theta)}{q(\boldsymbol{S})} >_{\boldsymbol{S},\boldsymbol{\theta}} -KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \tag{11}$$

where the first term corresponds to an average likelihood and the second term is the KL distance between the priors and the posteriors. This objective function can be maximised via an EM-like strategy (see [2]). In the E-step, $q(\boldsymbol{S})$ is updated according to

$$q(\boldsymbol{S}) \propto \exp[I(\boldsymbol{S})] \tag{12}$$

where

$$I(\boldsymbol{S}) =< \log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) >_{\boldsymbol{\theta}} \tag{13}$$

and the expectation is taken wrt. $q(\boldsymbol{\theta})$. In the M-step $q(\boldsymbol{\theta})$ is updated according to

$$q(\boldsymbol{\theta}) \propto \exp[I(\boldsymbol{\theta})]p(\boldsymbol{\theta}) \tag{14}$$

where

$$I(\boldsymbol{\theta}) =< \log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) >_{\boldsymbol{S}} \tag{15}$$

and the expectation is taken wrt. $q(\boldsymbol{S})$.

For the mixture model it turns out that the posteriors factorize into the following parametric forms. For the precisions we have $q(\boldsymbol{\beta}) = \prod_s q(\beta_s)$ and Gamma densities

$$q(\beta_s) = \Gamma(\beta_s; b_s, c_s) \tag{16}$$

For the means, we have $q(\boldsymbol{\mu}) = \prod_s q(\mu_s)$ and Normal densities

$$q(\mu_s) = N(\mu_s; m_s, v_s) \tag{17}$$

and for the mixing coefficients we have a Dirichlet

$$q(\pi) = \Gamma(\sum_{s'} \lambda_{s'}) \prod_{s=1}^{m} \frac{\pi_s^{\lambda_s - 1}}{\Gamma(\lambda_s)} \tag{18}$$

For the indicator posteriors we have $q(\boldsymbol{S}) = \prod_n q(s_n)$ and we write $\gamma_s^n \equiv q(s_n)$.

3.1 E-step

The E-step consists of updating the indicator posterior according to

$$\tilde{\gamma}_s^n = \tilde{\pi}_s \tilde{\beta}_s^{1/2} \exp\left[-\frac{1}{2}\bar{\beta}_s(y_n^2 + m_s^2 + v_s - 2m_s y_n)\right] \tag{19}$$

where

$$\begin{aligned} \log \tilde{\pi}_s &= \Psi(\lambda_s) - \Psi(\sum_{s'} \lambda_{s'}) \\ \log \tilde{\beta}_s &= \Psi(c_s) + \log b_s \\ \bar{\beta}_s &= b_s c_s \end{aligned} \tag{20}$$

where $\Psi()$ is the digamma function. The RHS of the equations consist of hyperparameters from the parameter posteriors (ie. λ_s, m_s, v_s, c_s and b_s). We then normalise to give

$$\gamma_s^n = \frac{\tilde{\gamma}_s^n}{\sum_{s'} \tilde{\gamma}_{s'}^n} \tag{21}$$

which is the probability that component s is responsible for data point y_n.

3.2 M-step

First of all, we define the following variables

$$\begin{aligned} \bar{\pi}_s &= \frac{1}{N}\sum_{n=1}^{N} \gamma_s^n \\ \bar{N}_s &= N\bar{\pi}_s \\ \bar{y}_s &= \frac{1}{N}\sum_{n=1}^{N} \gamma_s^n y_n \\ \tilde{y}_s^2 &= \frac{1}{N}\sum_{n=1}^{N} \gamma_s^n y_n^2 \end{aligned} \tag{22}$$

where $\bar{\pi}_s$ is the proportion of data in component s and $\bar{N}_s$ is the number of data points in component s. The quantities $\bar{y}_s$ and $\tilde{y}_s^2$ are the weighted data values and weighted squared data values respectively.

We then update the hyperparameters as follows. The mixing hyperparameters are updated by adding the data counts to the prior counts

$$\lambda_s = \bar{N}_s + \lambda_0 \tag{23}$$

If we define the average variance of component s as

$$\tilde{\sigma}_s^2 = \tilde{y}_s^2 + \bar{\pi}_s(m_s^2 + v_s) - 2m_s\bar{y}_s \tag{24}$$

then the hyperparameters for the precisions are updated as

$$\begin{aligned} \frac{1}{b_s} &= \frac{N}{2}\tilde{\sigma}_s^2 + \frac{1}{b_0} \\ c_s &= \frac{\bar{N}_s}{2} + c_0 \end{aligned} \tag{25}$$

We can understand these equations by looking at the corresponding mean variance (the inverse of the mean precision) which is given by $1/(b_s c_s)$. If we ignore terms involving the prior this comes out to be $\tilde{\sigma}_s^2/\bar{\pi}_s$ which is the average variance of that component re-weighted according to the number of examples that the component is responsible for.

If we first define the means and precisions of the component means *as estimated from the data* as

$$\begin{aligned} m_{data}(s) &= \bar{y}_s/\bar{\pi}_s \\ \tau_{data}(s) &= \bar{N}_s\bar{\beta}_s \end{aligned} \tag{26}$$

This last equation is reminiscent of the formula for the standard error (the variance of the estimate for the mean = the sample variance / the number of samples). The means and precisions of the posteriors are then given by

$$\begin{aligned} \tau_s &= \tau_0 + \tau_{data}(s) \\ m_s &= \frac{\tau_0}{\tau_s} m_0 + \frac{\tau_{data}(s)}{\tau_s} m_{data}(s) \end{aligned} \tag{27}$$

where $\tau_0 = 1/v_0$ is the prior precision and $\tau_s = 1/v_s$ is the posterior precision.

3.3 Setting the Priors

The means μ_s, have a Normal prior distribution with mean m_0 and variance v_0. We set m_0 to the mean of the data and set $v_0 = (\frac{x_{max} - x_{min}}{3})^2$.

The precisions β_s have a Gamma prior distribution. We use an uninformative prior (see eg. [7]) by setting $b = 10^3$ and $c = 10^{-3}$. The mixing coefficients $\boldsymbol{\pi}$ have a Dirichlet distribution parameterised by λ_0 which we set to 5.

3.4 Initialisation

We first apply the k-means clustering algorithm [3] which returns estimates for π_s, μ_s and β_s. The posterior for the means is a Gaussian with mean m_s and variance v_s which we initialise to $m_s = \mu_s$ and $v_s = (1/\beta_s)/(N\pi_s)$ where the RHS quantities are the k-means estimates. The initialisation of the variance is reminiscent of the formula for the standard error.

The posterior for the precisions is a Gamma whose mean is initialised to β_s and variance to $Var(\beta_s)$. The b_s and c_s parameters are set accordingly.

The posterior mixing distribution is parameterised by the coefficients λ_s which we initially set to $\lambda_s = 100\pi_s$ where π_s is the k-means estimate.

The VB equations can then be applied iteratively until a consistent solution is reached. Convergence can be measured by evaluating the negative free energy F_m which we now show how to compute.

3.5 Convergence

The negative free energy is composed of an average likelihood term L_{av} minus the KL divergence between the estimated posteriors and the priors and an entropy term. For the MOG model we have

$$F_m = L_{av} - KL(q(\boldsymbol{\pi})||p(\boldsymbol{\pi})) - KL(q(\beta)||p(\beta)) - KL(q(\mu)||p(\mu)) \tag{28}$$

where

$$L_{av} = H(q(\boldsymbol{S})) + \sum_{s=1}^{m} \bar{N}_s(\log \tilde{\pi}_s + 0.5 \log \tilde{\beta}_s) - 0.5N \sum_{s=1}^{m} \bar{\beta}_s \tilde{\sigma}_s^2 \tag{29}$$

The KL terms can be calculated from the relations given in the Appendix and the entropy over the hidden variables is

$$H(q(\boldsymbol{S})) = -\sum_{n=1}^{N} \sum_{s=1}^{m} \gamma_s^n \log \gamma_s^n \tag{30}$$

We calculate F_m at each iteration and terminate optimization if the proportionate increase from one iteration to the next is less than 0.01 per cent.

3.6 Co-mean algorithm

A simplified form of the Gaussian mixture model is one where all the means are tied to zero. This is useful, for example, for approximating unimodal non-Gaussian distributions. This simplifies a number of update equations. Equation 19 becomes

$$\tilde{\gamma}_s^n = \tilde{\pi}_s \tilde{\beta}_s^{1/2} \exp\left[-\frac{1}{2}\bar{\beta}_s y_n^2\right] \tag{31}$$

and equation 24 simplifies to

$$\tilde{\sigma}_s^2 = \tilde{y}_s^2 \tag{32}$$

and equations 26 and 27 drop out.

4 Model order selection

If we have a prior distribution over model order $p(m)$ and a posterior distribution $q(m)$ then we can define a more general objective function

$$F = \sum_m q(m)\left(F_m + \log\frac{p(m)}{q(m)}\right) \tag{33}$$

The optimal value of $q(m)$ is therefore

$$q(m) = \frac{\exp(F_m)p(m)}{\sum_{m'}\exp(F_{m'})p(m')} \tag{34}$$

For a uniform prior $p(m) = 1/M$ we get

$$q(m) = \frac{\exp(F_m)}{\sum_{m'}\exp(F_{m'})} \tag{35}$$

4.1 Comparison with Minimum Description Length

The Variational Bayes (VB) framework maximises the negative free energy

$$F_m = < \log p(D|\boldsymbol{\theta}) > -KL(Q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \tag{36}$$

where the first term is the average likelihood of the data and the second term is the KL-divergence between the approximating posteriors and the priors. This acts as a penalty term which penalizes more complex models.

As the number of samples increases the parameter posterior becomes sharply peaked about the most probable values (which are also the maximum likelihood values), $\hat{\boldsymbol{\theta}}$. It can then be shown that, in the large sample limit $N \to \infty$, F_m becomes equivalent to the Bayesian Information Criterion [2, 4]

$$BIC(m) = \log p(D|\hat{\boldsymbol{\theta}}) - \frac{p}{2}\log N \tag{37}$$

where p is the number of parameters in the model. BIC is itself equal to the negative of the Minimum Description Length (MDL) ie. $BIC(m) = -MDL(m)$. These popular model order selection criteria can therefore be seen as a limiting case of the VB framework. For the univariate Gaussian Mixture model we have

$$BIC(m) = \sum_{n=1}^{N}\log\left[\sum_{s=1}^{m}\pi_s N(y_n;\mu_s,1/\beta_s)\right] - \frac{3m}{2}\log N \tag{38}$$

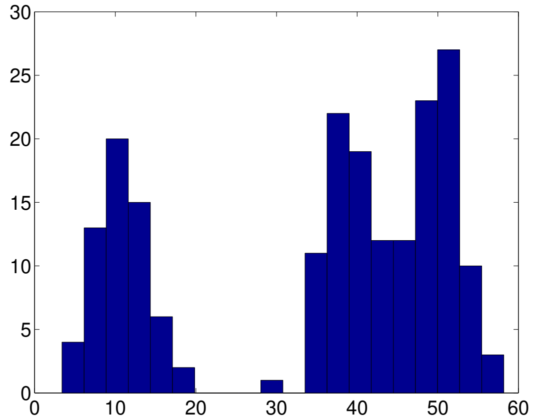

Figure 2: A histogram of 200 data points of 3 cluster data.

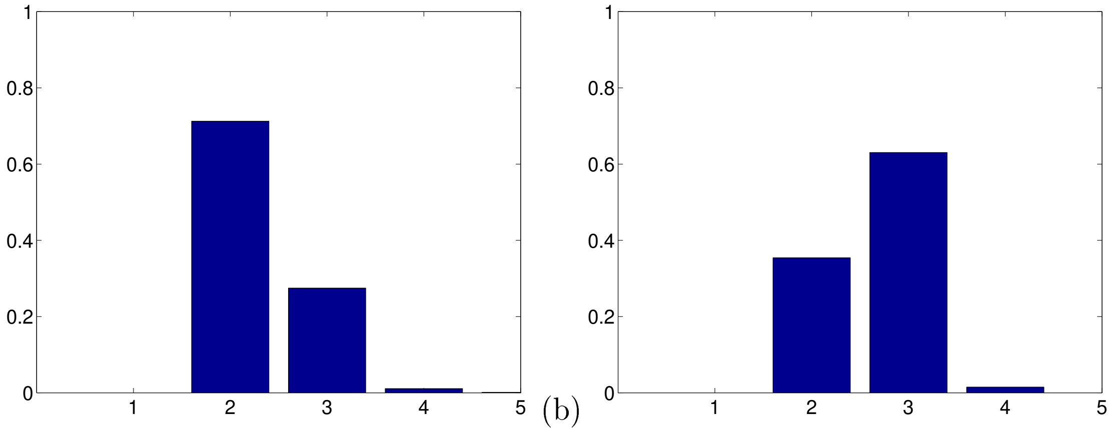

Figure 3: Model Order Selection for 3 cluster data with 50 data points (a) BIC/MDL and (b) VB.

5 Results

5.1 Overlapping data

We generated data from a 3-component Gaussian Mixture model with $\boldsymbol{\pi} = [0.3, 0.35, 0.35]$, $\boldsymbol{\mu} = [10, 40, 50]$, $\boldsymbol{Var} = [10, 10, 5]$. Figure 2 shows a histogram from a sample of 200 data points. This is a difficult data set as the second and third Gaussians have a strong overlap. We compare the model order selection ability of VB with that of BIC/MDL in Figures 3 and 4. Each plot is the average of 10 runs (where each different run is on a different randomly generated data set and we have averaged probabilities not free energies/likelihoods). VB is seen to be superior. On less difficult data sets (eg. non-overlapping clusters) and/or with more examples, VB and MDL are nearly always in agreement.

5.2 Co-mean data

We generated data from a mixture of two Gaussians with equal mixing priors, zero means and variances of 1 and 10. Figure 5 shows VB model order selection on data sets containing 100 data points (the histograms show the average of

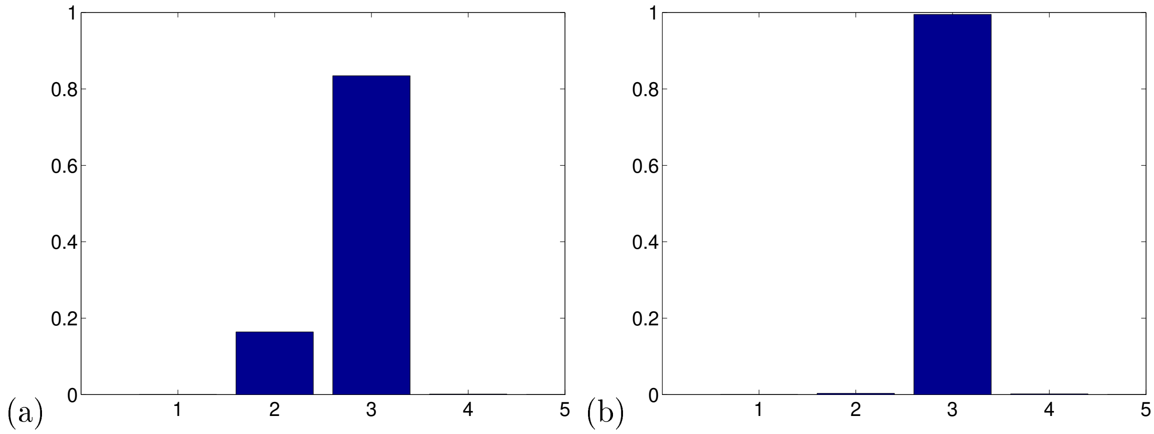

Figure 4: Model Order Selection for 3 cluster data with 200 data points (a) BIC/MDL and (b) VB.

ten runs, as before). Using k-means initialisation VB basically gets it wrong (so does MDL). Using a 'co-mean' initialisation scheme [1], however, VB gets it right. The difference in results is entirely attributable to the initialisation method which is important in GMM models as, even with VB-learning, there are many local maxima. We note that the separation of 1-dimensional co-mean data is much more difficult than the separation of multi-dimensional co-mean data. This is because, in multiple dimensions, the directions of the covariances (ie. the principal components) help enormously in the identification of the components; the equivalent problem in higher dimensions is therefore that of a mixture of co-mean isotropic (equal variance in all directions) Gaussians.

Of course, instead of using the full Gaussian mixture model with co-mean initialisation we could use the co-mean Gaussian mixture model (see section 2.6), and indeed, this is the preferred approach if the data is suspected to be unimodal (but non-Gaussian).

6 Discussion

Attias has derived a Variational Bayes algorithm for the general Gaussian mixture model [1, 2], ie. where the dimension of the data is d. The algorithm described here is therefore just a special case of Attias' work with $d = 1$.

References

[1] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[1] The co-mean initialisation scheme is as follows. We calculate the new variable $z = |y - \bar{y}|$ and apply a GMM(p) model to z, ie. a cluster analysis of absolute deviations. This results in mixing coefficients λ_z and means m_z. The posterior variances are set to m_z^2 and the posterior mixing priors to $100\lambda_z$. The posterior means are, of course, initialised to the same value $\bar{y}$. The variance of the posterior mean is initialised to the 10^3.

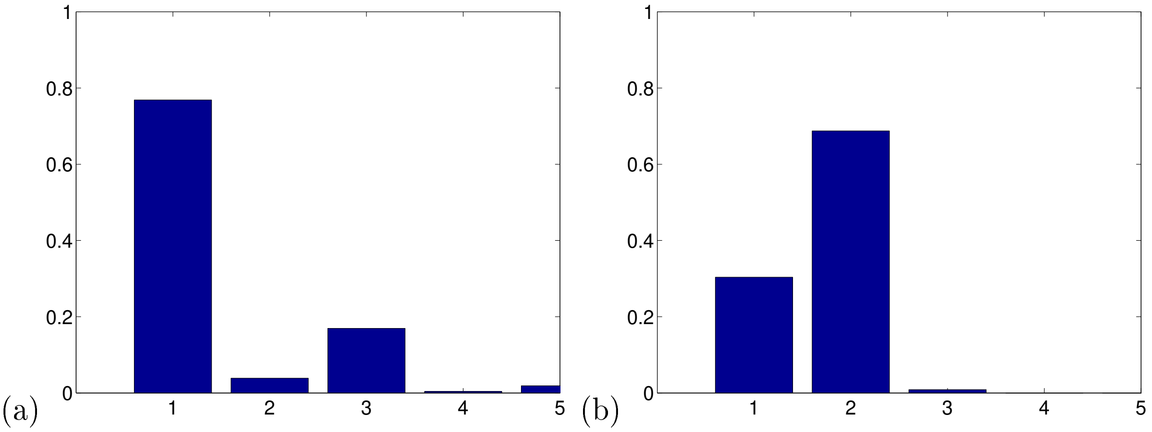

Figure 5: VB model order selection on two-component co-mean data using (a) k-means initialisation and (b) co-mean initialisation.

[2] H. Attias. A Variational Bayesian Framework for Graphical Models. In T. Leen et al, editor, *NIPS 12*, Cambridge, MA, 2000. MIT Press.

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

[4] D.M. Chickering and D. Heckerman. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. Technical Report MSR-TR-96-08, Microsoft Research, 1996.

[5] N. Lawrence. A Variational Bayesian Committee of Neural Networks. Technical report, Computer laboratory, Cambridge University, 2000.

[6] P. M. Lee. *Bayesian Statistics: An Introduction.* Arnold, 2 edition, 1997.

[7] J.J.K. O'Ruanaiaidh and W.J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* Springer, 1996.

A Derivation of update rules

A.1 E-Step

In the E-step, $q(\boldsymbol{S})$ is updated according to

$$q(\boldsymbol{S}) \propto \exp[I(\boldsymbol{S})] \tag{39}$$

where

$$I(\boldsymbol{S}) =< \log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) >_{\boldsymbol{\theta}} \tag{40}$$

and the expectation is taken wrt. $q(\boldsymbol{\theta})$. We have

$$\log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) = \sum_{n=1}^{N} \log \pi_s + 0.5 \log \beta_s - 0.5\beta_s (y_n - \mu_s)^2 \tag{41}$$

Taking expections wrt. the distributions $q(\boldsymbol{\pi})$, $q(\boldsymbol{\beta})$ and $q(\boldsymbol{\mu})$ gives

$$I(\boldsymbol{S}) = \sum_{n=1}^{N} I(s_n) \tag{42}$$

where

$$I(s_n) = \log \tilde{\pi}_s + 0.5 \log \tilde{\beta}_s - 0.5\bar{\beta}_s (y_n^2 + m_s^2 + v_s - 2m_s y_n) \tag{43}$$

and

$$\begin{aligned}
\log \tilde{\pi}_s &= \int q(\pi_s) \log \pi_s d\pi_s \qquad (44)\\
&= \Psi(\lambda_s) - \Psi(\sum_{s'} \lambda_{s'})\\
\log \tilde{\beta}_s &= \int q(\beta_s) \log \beta_s d\beta_s\\
&= \Psi(c_s) + \log b_s\\
\bar{\beta}_s &= \int q(\beta_s) \beta_s d\beta_s\\
&= b_s c_s
\end{aligned}$$

where $\Psi()$ is the digamma function (the above integrals are solved by integrating by parts as shown in Lawrence et al. [5]). The posterior distribution over hidden variables then factorises

$$q(\boldsymbol{S}) = \prod_{n=1}^{N} q(s_n) \tag{45}$$

If we define $\tilde{\gamma}_s^n \equiv q(s_n)$ then

$$\tilde{\gamma}_s^n \propto \exp[I(s_n)] \tag{46}$$

Hence,

$$\tilde{\gamma}_s^n = \tilde{\pi}_s \tilde{\beta}_s^{1/2} \exp\left[-\frac{1}{2}\bar{\beta}_s (y_n^2 + m_s^2 + v_s - 2m_s y_n)\right] \tag{47}$$

A.2 M-Step

In the M-step $q(\boldsymbol{\theta})$ is updated according to

$$q(\boldsymbol{\theta}) \propto \exp[I(\boldsymbol{\theta})]p(\boldsymbol{\theta}) \tag{48}$$

where

$$I(\boldsymbol{\theta}) =< \log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) >_{\boldsymbol{S}} \tag{49}$$

and the expectation is taken wrt. $q(\boldsymbol{S})$. We have

$$\log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) = \sum_{n=1}^{N} \log \pi_s + 0.5 \log \beta_s - 0.5\beta_s(y_n - \mu_s)^2 \tag{50}$$

A.2.1 Mixing coefficients

$$q(\boldsymbol{\pi}) \propto \exp[I(\boldsymbol{\pi})]p(\boldsymbol{\pi}) \tag{51}$$

where

$$I(\boldsymbol{\pi}) =< \log p(\boldsymbol{Y}, \boldsymbol{S}|\pi) >_{\boldsymbol{S}} \tag{52}$$

and the expectation is taken wrt. $q(\boldsymbol{S})$. Now,

$$\log p(\boldsymbol{Y}, \boldsymbol{S}|\pi) = \sum_{n=1}^{N} \log \pi_s \tag{53}$$

Hence, averaging over $q(\boldsymbol{S})$ gives

$$I(\boldsymbol{\pi}) = \sum_{s=1}^{m} I(\pi_s) \tag{54}$$

where

$$I(\pi_s) = \sum_{n=1}^{N} \gamma_s^n \log \pi_s \tag{55}$$

If we let

$$\bar{\pi}_s = \frac{1}{N} \sum_{n=1}^{N} \gamma_s^n \tag{56}$$

$$\bar{N}_s = N\bar{\pi}_s \tag{57}$$

then

$$I(\pi_s) = \bar{N}_s \log \pi_s \tag{58}$$

The above summation means that

$$q(\boldsymbol{\pi}) = \prod_{s=1}^{m} q(\pi_s) \tag{59}$$

where

$$q(\pi_s) \propto \exp[I(\pi_s)]p(\pi_s) \tag{60}$$

Now,

$$\log q(\pi_s) = \bar{N}_s \log \pi_s + (\lambda_0 - 1) \log \pi_s \tag{61}$$

Hence, the optimal form for $q(\pi_s)$ is a Dirichlet with parameters

$$\lambda_s = \bar{N}_s + \lambda_0 \tag{62}$$

A.2.2 Means

$$q(\boldsymbol{\mu}) \propto \exp[I(\boldsymbol{\mu})]p(\boldsymbol{\mu}) \tag{63}$$

where

$$I(\boldsymbol{\mu}) =< \log p(\boldsymbol{Y}, \boldsymbol{S}|\mu) >_{\boldsymbol{S}} \tag{64}$$

and the expectation is taken wrt. $q(\boldsymbol{S})$. Now,

$$\log p(\boldsymbol{Y}, \boldsymbol{S}|\mu) = 0.5\beta_s \sum_{n=1}^{N} (y_n - \mu_s)^2 \tag{65}$$

Hence, averaging over $q(\boldsymbol{S})$ gives

$$I(\boldsymbol{\mu}) = \sum_{s=1}^{m} I(\mu_s) \tag{66}$$

where

$$I(\mu_s) = -0.5\bar{\beta}_s \sum_{n=1}^{N} \gamma_s^n (y_n - \mu_s)^2 \tag{67}$$

where we have also averaged over β_s to eliminate the dependence on the precisions. Because $I(\boldsymbol{\mu})$ splits into a sum over s we have

$$q(\boldsymbol{\mu}) = \prod_{s=1}^{m} q(\mu_s) \tag{68}$$

where

$$q(\mu_s) \propto \exp[I(\mu_s)]p(\mu_s) \tag{69}$$

Now

$$\begin{aligned} \exp(I(\mu_s)) &\propto \exp[-0.5\bar{\beta}_s \sum_{n=1}^{N} \gamma_s^n (y_n - \mu_s)^2] \\ &\propto \exp[-0.5\bar{\beta}_s (\bar{N}_s \mu_s^2 - 2\mu_s N \bar{y}_s + N \tilde{y}_s^2)] \\ &\propto \exp\left[-0.5\bar{\beta}_s \bar{N}_s \left(\mu_s - \frac{\bar{y}_s}{\bar{\pi}_s}\right)^2\right] \end{aligned} \tag{70}$$

where in the last step we have discarded terms not involving μ_s. This gives a normal distribution with mean $\frac{\bar{y}_s}{\bar{\pi}_s}$ and precision $\bar{\beta}_s \bar{N}_s$. The combination of the two Gaussians results in a single Gaussian whose precision is the sum of the two precisions and whose mean is the sum of the two means, each weighted according to their relative precisions [6]. Hence

$$q(\mu_s) = N(\mu_s; m_s, 1/\tau_s) \tag{71}$$

where

$$\begin{aligned} \tau_s &= \tau_0 + \tau_{data}(s) \\ m_s &= \frac{\tau_0}{\tau_s} m_0 + \frac{\tau_{data}(s)}{\tau_s} m_{data}(s) \end{aligned} \tag{72}$$

where $\tau_0 = 1/v_0$ is the prior precision and $\tau_s = 1/v_s$ is the posterior precision and

$$\begin{aligned} m_{data}(s) &= \frac{\bar{y}_s}{\bar{\pi}_s} \\ \tau_{data}(s) &= \bar{\beta}_s \bar{N}_s \end{aligned} \tag{73}$$

A.2.3 Precisions

$$q(\boldsymbol{\beta}) \propto \exp[I(\boldsymbol{\beta})]p(\boldsymbol{\beta}) \tag{74}$$

where

$$I(\boldsymbol{\beta}) =< \log p(\boldsymbol{Y}, \boldsymbol{S}|\beta) >_{\boldsymbol{S}} \tag{75}$$

and the expectation is taken wrt. $q(\boldsymbol{S})$. Now,

$$\log p(\boldsymbol{Y}, \boldsymbol{S}|\beta) = 0.5 \sum_{n=1}^{N} \log \beta_s - 0.5\beta_s \sum_{n=1}^{N} (y_n - \mu_s)^2 \tag{76}$$

Hence, averaging over $q(\boldsymbol{S})$ gives

$$I(\boldsymbol{\beta}) = \sum_{s=1}^{m} I(\beta_s) \tag{77}$$

where

$$I(\beta_s) = 0.5 \sum_{n=1}^{N} \gamma_s^n \log \beta_s - 0.5\beta_s \sum_{n=1}^{N} \gamma_s^n (y_n - \mu_s)^2 \tag{78}$$

If we integrate the second summation over $q(\mu_s)$ (and normalise)

$$\tilde{\sigma}_s^2 = \frac{1}{N} \int d\mu_s q(\mu_s) \sum_{n=1}^{N} \gamma_s^n (y_n - \mu_s)^2 \tag{79}$$

then

$$\tilde{\sigma}_s^2 = \tilde{y}_s^2 + \bar{\pi}_s(m_s^2 + v_s) - 2m_s\bar{y}_s \tag{80}$$

where

$$\begin{aligned} \bar{y}_s &= \frac{1}{N} \sum_{n=1}^{N} \gamma_s^n y_n \\ \tilde{y}_s^2 &= \frac{1}{N} \sum_{n=1}^{N} \gamma_s^n y_n^2 \end{aligned} \tag{81}$$

Hence

$$I(\beta_s) = 0.5\bar{N}_s \log \beta_s - 0.5N\tilde{\sigma}_s^2\beta_s \tag{82}$$

Because $I(\boldsymbol{\beta})$ splits into a sum over s we have

$$q(\boldsymbol{\beta}) = \prod_{s=1}^{m} q(\beta_s) \tag{83}$$

where

$$q(\beta_s) \propto \exp[I(\beta_s)]p(\beta_s) \tag{84}$$

Now combining the exponential term with the Gamma prior gives

$$\log q(\beta_s) = (0.5\bar{N}_s + c_0 - 1) \log \beta_s - (0.5N\tilde{\sigma}_s^2 + 1/b_0)\beta_s \tag{85}$$

Hence, the optimal form for $q(\beta_s)$ is a Gamma with parameters

$$\begin{aligned} \frac{1}{b_s} &= \frac{N}{2}\tilde{\sigma}_s^2 + \frac{1}{b_0} \\ c_s &= \frac{\bar{N}_s}{2} + c_0 \end{aligned} \tag{86}$$

B Derivation of Negative Free Energy

The average likelihood term is given by

$$L_{av} = < \log \frac{p(\boldsymbol{Y}, \boldsymbol{S}|\theta)}{q(\boldsymbol{S})} >_{\boldsymbol{S},\boldsymbol{\theta}} \tag{87}$$

Splitting this up gives

$$L_{av} = < \log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) >_{\boldsymbol{S},\boldsymbol{\theta}} + H(q(\boldsymbol{S})) \tag{88}$$

where $H(q(\boldsymbol{S}))$ is the entropy over the hidden variables. The first term is the average of

$$\log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) = \log \pi_s + 0.5 \log \beta_s - 0.5\beta_s (y_n - \mu_s)^2 \tag{89}$$

which is

$$\begin{aligned} < \log p(\boldsymbol{Y}, \boldsymbol{S}|\theta) > &= \sum_{s=1}^{m} \sum_{n=1}^{N} \gamma_s^n [\int d\pi_s q(\pi_s) \log \pi_s \\ &+ 0.5 \int d\beta_s q(\beta_s) \log \beta_s - 0.5\bar{\beta}_s \int d\mu_s q(\mu_s)(y_n - \mu_s)^2] \\ &= \sum_{s=1}^{m} \bar{N}_s (\log \tilde{\pi}_s + 0.5 \log \tilde{\beta}_s) - 0.5N \sum_{s=1}^{m} \bar{\beta}_s \tilde{\sigma}_s^2 \end{aligned} \tag{90}$$

C Single Gaussian

For the special case of a single Gaussian (ie. $m = 1$) having mean μ and precision β the prior distributions are $p(\mu) = N(\mu; m_0, v_0)$ and $p(\beta) = \Gamma(\beta, b_0, c_0)$. The posteriors are $q(\mu) = N(\mu; m, v)$ and $q(\beta) = \Gamma(\beta, b, c)$. We first define

$$\begin{aligned} s_1 &= \sum_{n=1}^{N} y_n \\ s_2 &= \sum_{n=1}^{N} y_n^2 \\ s_v &= s_2 + N(m^2 + v) - 2ms_1 \end{aligned} \tag{91}$$

The update rule for the precision is then

$$\begin{aligned} \frac{1}{b} &= \frac{s_v}{2} + \frac{1}{b_0} \\ c &= \frac{N}{2} + c_0 \end{aligned} \tag{92}$$

and for the mean

$$\begin{aligned} \beta_m &= N\bar{\beta} + \beta_0 \\ v &= \frac{1}{\beta_m} \\ m &= \frac{\beta_0}{\beta_m} m_0 + \frac{N\bar{\beta}}{\beta_m} \frac{s_1}{N} \end{aligned} \tag{93}$$

The negative free energy is

$$F_1 = L_{av} - KL(q(\mu)||p(\mu)) - KL(q(\beta)||p(\beta)) \tag{94}$$

where the average likelihood is

$$L_{av} = 0.5N \log \tilde{\beta} - 0.5\bar{\beta} s_v \tag{95}$$

and $\bar{\beta} = bc$ and

$$\log \tilde{\beta} = \Psi(c) + \log b \tag{96}$$

D Densities and Divergences

D.1 Gamma densities

The Gamma density is defined as

$$\Gamma(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \tag{97}$$

The mean of a Gamma density is given by bc and the variance by $b^2 c$. Logs of gamma densities can be written as

$$\log \Gamma(x; b, c) = \frac{-x}{b} + (c-1)\log x + K \tag{98}$$

where K is a quantity which does not depend on x; the log of a gamma density comprises a term in x and a term in $\log x$.

D.2 KL-Divergence for Dirichlet densities

For the Dirichlet density with $q(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}; \boldsymbol{\lambda}_q)$ and $p(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}; \boldsymbol{\lambda}_p)$ where the number of states is m and $\lambda_{qt} = \sum_{s=1}^{m} \lambda_q(s)$, $\lambda_{pt} = \sum_{s=1}^{m} \lambda_p(s)$ the KL-divergence is

$$\begin{aligned} KL(q||p) \quad = \quad & \log \frac{\Gamma(\lambda_{qt})}{\Gamma(\lambda_{pt})} + \sum_{s=1}^{m} (\lambda_q(s) - \lambda_p(s))(\Psi(\lambda_q(s)) - \Psi(\lambda_{qt}) \\ & + \sum_{s=1}^{m} \log \frac{\Gamma(\lambda_p(s))}{\Gamma(\lambda_q(s))} \end{aligned} \tag{99}$$

and $\Psi()$ is the digamma function.

D.3 KL-Divergence for Univariate Normal densities

For Normal densities $q(x) = N(x; \mu_q, \sigma_q^2)$ and $p(x) = N(x; \mu_p, \sigma_p^2)$ the KL-divergence is

$$KL(q||p) = 0.5 \log \frac{\sigma_p^2}{\sigma_q^2} + \frac{\mu_q^2 + \mu_p^2 + \sigma_q^2 - 2\mu_q \mu_p}{2\sigma_p^2} - 0.5 \tag{100}$$

D.4 KL-Divergence for Gamma densities

For Gamma densities $q(x) = \Gamma(\boldsymbol{x}; b_q, c_q)$ and $p(x) = \Gamma(\boldsymbol{x}; b_p, c_p)$ the KL-divergence is

$$\begin{aligned} KL(q||p) &= (c_q - 1)\Psi(c_q) - \log b_q - c_q - \log\Gamma(c_q) \\ &+ \log\Gamma(c_p) + c_p \log b_p - (c_p - 1)(\Psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p} \end{aligned} \tag{101}$$

where $\Psi()$ is the digamma function.