Variational Bayes for d-dimensional Gaussian Mixture Models

W.D. Penny

Wellcome Department of Cognitive Neurology.
University College London.

July 4, 2001

Abstract

We derive the model order selection criterion used in the Variational Bayesian (VB) framework for d-dimensional Gaussian mixture models.

1 Introduction

Attias has recently developed a fully Bayesian approach to Gaussian Mixture modelling [1]. In this report we show how the model order selection criterion is derived.

2 Bayesian Gaussian Mixture model

Component s has mixing coefficient π_s, mean μ_s and precision Γ_s. We can write the parameters collectively as the vectors $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_m]$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_m]$ and $\boldsymbol{\Gamma} = [\Gamma_1, \Gamma_2, ..., \Gamma_m]$ and can concatenate these into the overall parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Gamma}\}$. We are given a set of data points $D = \{\boldsymbol{y}_n\}$ with $n = 1..N$. The likelihood of a d-dimensional data point $\boldsymbol{y}_n$ is given by the mixture model

$$p(\boldsymbol{y}_n|\boldsymbol{\theta}) = \sum_{s=1}^{m} p(s_n = s|\boldsymbol{\pi}) p(\boldsymbol{y}_n|\Gamma_s, \boldsymbol{\mu}_s) \quad (1)$$

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

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

Each component is a Gaussian with

$$p(\boldsymbol{y}_n|\boldsymbol{\Gamma}_s, \boldsymbol{\mu}_s) = (2\pi)^{-d/2}|\boldsymbol{\Gamma}_s|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{\mu}_s)^T \boldsymbol{\Gamma}_s (\boldsymbol{y}_n - \boldsymbol{\mu}_s)\right) \quad (3)$$

The prior on the model parameters is

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_s p(\boldsymbol{\Gamma}_s) p(\boldsymbol{\mu}_s|\boldsymbol{\Gamma}_s) \quad (4)$$

where the mixing prior is a symmetric Dirichlet

$$p(\boldsymbol{\pi}) = \mathsf{Dir}(\lambda_0 \boldsymbol{I}_m) \tag{5}$$

the prior over each precision matrix is a Wishart

$$p(\boldsymbol{\Gamma}_s) = \mathsf{W}(a_0, B_0) \tag{6}$$

and the prior over the means conditioned on the precisions is a Normal

$$p(\boldsymbol{\mu}_s|\boldsymbol{\Gamma_s}) = N(\boldsymbol{m}_0, \beta_0 \boldsymbol{\Gamma_s}) \tag{7}$$

Together, the prior $p(\boldsymbol{\mu}_s, \boldsymbol{\Gamma_s})$ is a Normal-Wishart density. The joint likelihood of a data point and indicator variable is

$$p(\boldsymbol{y}_n, s_n|\theta) = p(s_n = s|\boldsymbol{\pi}) p(\boldsymbol{y}_n|\Gamma_s, \boldsymbol{\mu}_s) \tag{8}$$

which, given that the data are IID, gives

$$p(\boldsymbol{Y}, \boldsymbol{S}|\theta) = \prod_{n=1}^{N} p(s_n = s|\boldsymbol{\pi}) p(\boldsymbol{y}_n|\Gamma_s, \boldsymbol{\mu}_s) \tag{9}$$

over the whole data set, where $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n]$ and $\boldsymbol{S} = [s_1, s_2, ..., s_N]$.

A variational approximation to the posterior [1] leads to the following components. For the precisions we have $q(\boldsymbol{\Gamma}) = \prod_s q(\boldsymbol{\Gamma}_s)$ and Wishart densities

$$q(\Gamma_s) = \mathsf{W}(a_s, B_s) \tag{10}$$

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

$$q(\boldsymbol{\mu}_s|\boldsymbol{\Gamma}_s) = N(\boldsymbol{m}_s, \beta_s \boldsymbol{\Gamma_s}) \tag{11}$$

and for the mixing coefficients we have a Dirichlet

$$q(\boldsymbol{\pi}) = \mathsf{Dir}(\boldsymbol{\lambda}) \tag{12}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ..., \lambda_m]$. For the indicator posteriors we have $q(\boldsymbol{S}) = \prod_n q(s_n)$ and we write as $\gamma_s^n \equiv q(s_n)$.

2.1 E-step

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

$$\tilde{\gamma}_s^n = \tilde{\pi}_s \tilde{\boldsymbol{\Gamma}}_s^{1/2} \exp\left[-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{m}_s)^T \bar{\boldsymbol{\Gamma}}_s (\boldsymbol{y}_n - \boldsymbol{m}_s)\right] \exp\left[\frac{-d}{2\beta_s}\right] \tag{13}$$

where

$$\begin{aligned} \log \tilde{\pi}_s &= \Psi(\lambda_s) - \Psi(\sum_{s'} \lambda_{s'}) & (14) \\ \log \tilde{\boldsymbol{\Gamma}}_s &= \sum_{i=1}^{d} \Psi((a_s + 1 - i)/2) - \log|\boldsymbol{B}_s| + d \log 2 \\ \bar{\boldsymbol{\Gamma}}_s &= a_s \boldsymbol{B}_s^{-1} \end{aligned}$$

where $\Psi()$ is the digamma function. The first two lines in the above equation correspond to equations 34 and 134 in Minka [3]. 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

$$\gamma_s^n = \frac{\tilde{\gamma}_s^n}{\sum_{s'} \tilde{\gamma}_{s'}^n} \tag{15}$$

which is the probability that component s is responsible for data point $\boldsymbol{y}_n$.

2.2 M-step

First of all, we define the following variables

$$
\begin{aligned}
\bar{\pi}_s &= \frac{1}{N}\sum_{n=1}^{N}\gamma_s^n && (16)\\
\bar{N}_s &= N\bar{\pi}_s\\
\bar{\mu}_s &= \frac{1}{\bar{N}_s}\sum_{n=1}^{N}\gamma_s^n \boldsymbol{y}_n\\
\bar{\boldsymbol{\Sigma}}_s &= \frac{1}{\bar{N}_s}\sum_{n=1}^{N}\gamma_s^n(\boldsymbol{y}_n-\bar{\boldsymbol{\mu}}_s)(\boldsymbol{y}_n-\bar{\boldsymbol{\mu}}_s)^T
\end{aligned}
\tag{17}
$$

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. We then update the hyperparameters as follows. For the mixing coefficients we have a Dirichlet

$$q(\boldsymbol{\pi}) = \mathsf{Dir}(\boldsymbol{\lambda}) \tag{18}$$

where the hyperparameters are updated by adding the data counts to the prior counts

$$\lambda_s = \bar{N}_s + \lambda_0 \tag{19}$$

For the means, we have

$$q(\boldsymbol{\mu}_s|\boldsymbol{\Gamma}_s) = N(\boldsymbol{m}_s, \beta_s\boldsymbol{\Gamma}_{\boldsymbol{s}}) \tag{20}$$

where

$$
\begin{aligned}
\boldsymbol{m}_s &= \frac{\bar{N}_s\bar{\boldsymbol{\mu}}_s + \beta_0\boldsymbol{m}_0}{\bar{N}_s+\beta_0} && (21)\\
\beta_s &= \bar{N}_s + \beta_0
\end{aligned}
$$

The above two equations are identical to eq.s 90 and 91 in Minka [3]. For the noise precision matrix we have a Wishart density

$$q(\boldsymbol{\Gamma}_s) = \mathsf{W}(a_s, \boldsymbol{B}_s) \tag{22}$$

where

$$
\begin{aligned}
\boldsymbol{B}_s &= \bar{N}_s\bar{\Sigma}_s + \frac{\bar{N}_s\beta_0(\bar{\boldsymbol{\mu}}_s-\boldsymbol{m}_0)(\bar{\boldsymbol{\mu}}_s-\boldsymbol{m}_0)^T}{\bar{N}_s+\beta_0} + \boldsymbol{B}_0 && (23)\\
a_s &= \bar{N}_s + a_0
\end{aligned}
$$

The above two equations are identical to Eq.s 89 and 91 in Minka [3].

2.3 Model order selection

The negative free energy is given by

$$
\begin{aligned}
F &= \int q(\pi)\log\frac{p(\pi)}{q(\pi)}d\pi && (24)\\
&+ \sum_{s=1}^{m}\int q(\Gamma_s)\log\frac{p(\Gamma_s)}{q(\Gamma_s)}d\Gamma_s
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{s=1}^{m} \int q(\mu_s|\Gamma_s) \log \frac{p(\mu_s|\Gamma_s)}{q(\mu_s|\Gamma_s)} d\mu_s \\
&+ \sum_{s=1}^{m} \sum_{n=1}^{N} q(s_n) \int q(\pi) \log \frac{p(s_n|\pi)}{q(s_n)} d\pi \\
&+ \sum_{s=1}^{m} \sum_{n=1}^{N} q(s_n) \int \int q(\Gamma_s) q(\mu_s|\Gamma_s) \log p(y_n|\Gamma_s, \mu_s, s_n) d\Gamma_s d\mu_s
\end{aligned}
$$

The first three terms are KL-divergences and the last two constitute the average likelihood. This can be evaluated as

$$
\begin{aligned}
F &= -KL_{Dir}(\boldsymbol{\lambda}; \boldsymbol{\lambda}_0) \\
&- \sum_{s=1}^{m} KL_W(a_s, B_s; a_0, B_0) \\
&- \sum_{s=1}^{m} KL_N(m_s, B_s/(\beta_s a_s); m_0, B_s/(\beta_0 a_s))) \\
&+ \sum_{s=1}^{m} L_{av}(s)
\end{aligned} \tag{25}
$$

where

$$
\begin{aligned}
L_{av}(s) &= \bar{N}_s \log \tilde{\pi}_s - \sum_{n=1}^{N} \gamma_s^n \log \gamma_s^n \\
&+ \frac{\bar{N}_s}{2} \left(-d \log 2\pi + \log \tilde{\Gamma}_s - Tr(\bar{\Gamma}_s \bar{\Sigma}_s) - \frac{d}{\beta_s} \right)
\end{aligned} \tag{26}
$$

These terms can also be simplified (esp. term 3) and there are a number of elements which cancel when the terms are added (eg. $\bar{N}_s \log \tilde{\pi}_s$, $(\bar{N}_s/2) \log \tilde{\Gamma}_s$), although, in this report we use the model order criterion based on the KL divergences as in equation 25. These KL-divergences are given in [4].

2.4 Minimum Description Length

The Bayesian Information Criterion is

$$
BIC(m) = \sum_{n=1}^{N} \log p(\boldsymbol{y}_n|\hat{\boldsymbol{\theta}}) - \frac{N_m}{2} \log N \tag{27}
$$

where N_m is the number of parameters in a model of size m. For a Gaussian Mixture Model we have

$$
N_m = m \left(1 + d + \frac{d(d+1)}{2} \right) \tag{28}
$$

where the first term is for the mixing coefficients, the second for the means and the third for the covariances. We have $MDL(m) = -BIC(m)$.

2.5 Practicalities

For priors we use

$$
\begin{aligned}
\lambda_0 &= 1 && (29) \\
a_0 &= d \\
B_0 &= 0.01d\boldsymbol{I} \\
\boldsymbol{m}_0 &= \bar{\boldsymbol{y}} \\
\beta_0 &= 1 \\
& && (30)
\end{aligned}
$$

The prior covariance, of each component, is therefore $B_0/a_0 = 0.01\boldsymbol{I}$. The algorithm is initialised using k-means [2].

For the special case of a single Gaussian we simply substitute $\gamma_s^n = 1$ for all n into the M-Step equations. The update equations are then solved in a single iteration, and an expression is given for F.

3 Results

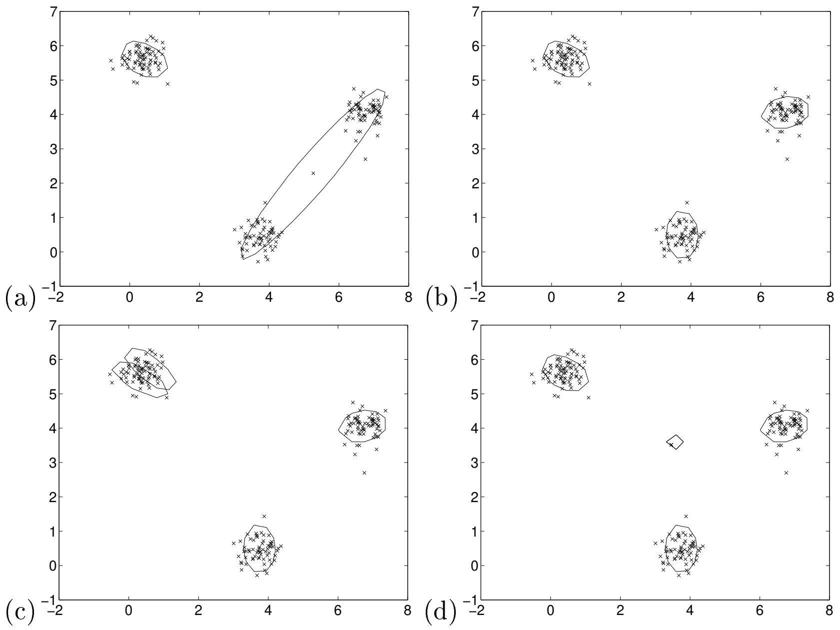

Figure 1: *A 2D data set containing 3 clusters modelled by 2, 3, 4 and 5 cluster GMM models - plots a to d, respectively. These models have negative free energies of* $F(m) = -581, -462, -491, -470$, *showing that the 3-cluster model is preferred.*

References

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

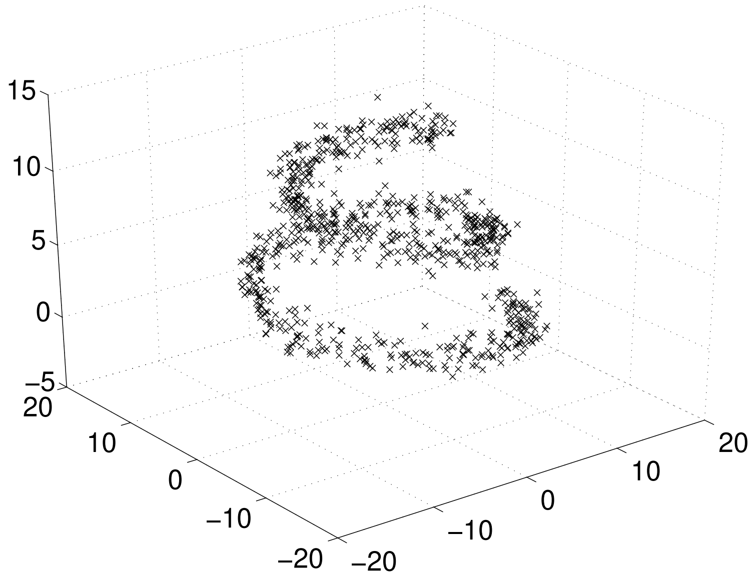

Figure 2: *Spirals data set*

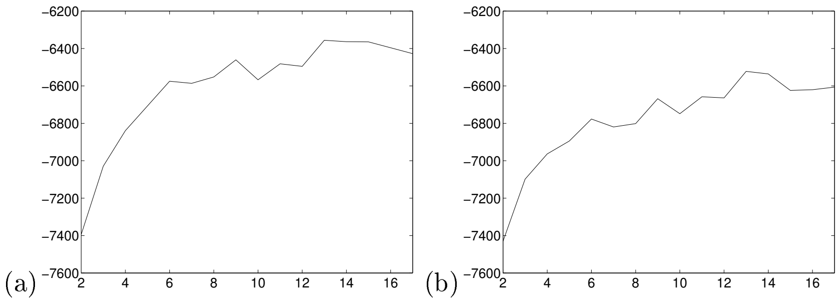

Figure 3: *Model order selection for the spirals data set with (a) BIC and (b) Negative Free Energy, $F(m)$. Both criteria prefer 13 clusters.*

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

[3] T.P. Minka. Using lower bounds to approximate integrals. Technical report, MIT Media Laboratory, 2001. http://vismod.www.media.mit.edu/ tpminka.

[4] W.D. Penny. Kullback-Liebler Divergences of Normal, Gamma, Dirichlet and Wishart densities. Technical report, Wellcome Department of Cognitive Neurology, 2001.