Variational Bayesian Model Selection for Mixture Distributions

Adrian Corduneanu[*] adrianc@mit.edu
Christopher M. Bishop cmbishop@microsoft.com
Microsoft Research
7 J. J. Thomson Avenue, Cambridge, CB3 0FB, U.K.
http://research.microsoft.com/∼cmbishop

In *Artificial Intelligence and Statistics 2001*, T. Jaakkola and T. Richardson (Eds), 27–34, Morgan Kaufmann

Abstract

Mixture models, in which a probability distribution is represented as a linear superposition of component distributions, are widely used in statistical modelling and pattern recognition. One of the key tasks in the application of mixture models is the determination of a suitable number of components. Conventional approaches based on cross-validation are computationally expensive, are wasteful of data, and give noisy estimates for the optimal number of components. A fully Bayesian treatment, based on Markov chain Monte Carlo methods for instance, will return a posterior distribution over the number of components. However, in practical applications it is generally convenient, or even computationally essential, to select a single, most appropriate model. Recently it has been shown, in the context of linear latent variable models, that the use of hierarchical priors governed by continuous hyper-parameters whose values are set by type-II maximum likelihood, can be used to optimize model complexity. In this paper we extend this framework to mixture distributions by considering the classical task of density estimation using mixtures of Gaussians. We show that, by setting the mixing coefficients to maximize the marginal log-likelihood, unwanted components can be suppressed, and the appropriate number of components for the mixture can be determined in a single training run without recourse to cross-validation. Our approach uses a variational treatment based on a factorized approximation to the posterior distribution.

1 Introduction

Mixture models are widely used as computationally convenient representations for modeling complex probability distributions, and are based on a linear combination of some number M of simpler, component distributions. In this paper we shall focus on the case in which the components of the mixture are multivariate normal distributions $\mathcal{N}(x|\mu, T)$ where x is a continuous multidimensional variable, and μ and T are the mean and inverse covariance parameters respectively. The mixture distribution for M components is then

$$P(x|\pi, \mu, T) = \sum_{i=1}^{M} \pi_i \mathcal{N}(x|\mu_i, T_i) \qquad (1)$$

where π_i are called mixing coefficients, and satisfy $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{M} \pi_i = 1$. Note that we use π to denote the set $\{\pi_i\}_{i=1}^{M}$, and similarly for $\mu \equiv \{\mu_i\}$ and $T \equiv \{T_i\}$.

Consider an observed data set D comprising N observations x_n, where $n = 1, \ldots, N$, which are assumed to be drawn independently from the mixture distribution (1). The probability of the observed data set, given the mixing coefficients and the parameters of the components is then given by

$$P(D|\pi, \mu, T) = \prod_{n=1}^{N} \left[\sum_{i=1}^{M} \pi_i \mathcal{N}(x_n|\mu_i, T_i) \right]. \qquad (2)$$

Viewed as a function of (π, μ, T), this is called the *likelihood* function.

The maximum likelihood framework chooses specific values for the model parameters which correspond to a (local) maximum of the likelihood function. A powerful approach to finding maximum likelihood solutions is based on the iterative EM (expectation-maximization) algorithm in which the mixture distribution is re-interpreted as a latent variable

[*] Present address: MIT AI Lab
545 Technology Square, Cambridge, MA 02139

model (as discussed in the next Section). This interpretation also forms the basis of our Bayesian variational treatment of the mixture model.

One difficulty with this procedure is that maximum likelihood is strictly not well defined due to the presence of singularities in the likelihood function in which one (or more) of the component densities collapses onto a specific data point (its mean μ_i becomes equal to the data vector x_n and the corresponding covariance goes to zero thus assigning infinite density at the location of the data point). In practice we must seek a good local maximum of the likelihood function, often with the use of heuristics to avoid encountering the singularities of the likelihood function. It should also be noted that, in addition to the singularities, the likelihood function is typically characterized by multiple local maxima, and that good initialization heuristics for EM (for example based on the K-means algorithm) can be important in order to consistently find good solutions.

A further limitation of maximum likelihood is that it does not provide any guidance on the choice of the model order M. Larger values of M allow the model to achieve better fits to the training data and hence to assign larger values of the likelihood function for the observed data set. However, the generalization capability of the model, i.e. its ability to assign a high probability to a data set drawn independently from the same distribution as the training set, is best for some specific value of M, with larger (as well as smaller) values having poorer generalization. Determination of the optimum model order for a specific problem is a central goal of the research presented in this paper.

Many traditional approaches to such model selection problems are based on cross-validation in which a range of candidate models are optimized to a training set and their predictive performance subsequently compared on an independent validation set. This approach is both computationally expensive and wasteful of valuable data that could otherwise be used for training. Furthermore, it is only applicable if there are one, or perhaps two, discrete model complexity parameters to be optimized, since an exhaustive search over the combinatorially large space of several such parameters would be computationally prohibitive.

A fully Bayesian treatment of the mixture modeling problem involves the introduction of prior distributions over the mixing coefficients and the parameters of the component distributions, as well as over the number of components in the mixture. Conditioning on the observed data leads to a posterior distribution over the number of components, where it is hoped that the most probable number corresponds to the model with the best generalization. Effectively this approach must consider all possible values of the number of components M up to some maximum value. In more complex models where there may be several such discrete parameters, such an approach can become intractable (although sampling methods could in principle sample preferentially from the regions of high posterior probability).

In the neural networks literature MacKay and Neal have advocated the use of continuous hyper-parameters as a mechanism for avoiding discrete model search [10]. They call this procedure ‘automatic relevance determination’ (ARD). Values of the hyper-parameters are determined using ‘type 2’ maximum likelihood in which the values of the hyper-parameters are chosen to optimize the marginal likelihood of the observed data in which the model parameters have been integrated out. Although this approach has met with limited success in the context of neural networks (probably due to the complexity of the likelihood function) it has subsequently proved to be very successful in the Bayesian treatment of principal component analysis (PCA) [2]. Here a separate hyper-parameter, representing the inverse variance, was introduced for each potential principal component. In the posterior distribution, hyper-parameters with large means represent components which are suppressed with high probability. Thus only a single model is considered (the one with the largest number of principal components) and the mean of the posterior distribution captures the most probable model complexity. This approach has been extended to a mixture of (a fixed number of) Bayesian PCA models [2, 4] in which each model can independently determine its own effective dimensionality, something which would be computationally prohibitive to tackle using cross-validation.

In this paper we extend the continuous hyper-parameter framework to address the problem of choosing the number of components in a Gaussian mixture model. Conventionally this problem may be solved by exhaustive cross-validation in the number of components up to some maximum value. Alternatively statistical tests may be used, for example Polymenis and Titterington [11] describe a recent approach, and also provide a survey of the history of this area. Techniques based on complexity criteria are discussed by Figueiredo and Jain [5] and references therein. The problem has also been approached from a Bayesian perspective using reversible jump Markov chain Monte Carlo [7] and using variational methods [1, 6, 4]. Both approaches return a posterior distribution (or samples from the posterior) over the number of components (up to some maximum), and therefore these methods effectively consider all possible intermediate models explicitly.

Our approach involves the use of a mixture model having a fixed number of potential components (corresponding to the maximum number considered above), in which the mixing coefficients are optimized using type 2 maximum likelihood. This causes the mixing coefficients corresponding to unwanted components to go to zero. The means and variances of the Gaussian components, as well as the discrete latent variables, are marginalized out using variational techniques. Our results indicate that this approach is able to recover the appropriate number of components in synthetic data problems, and that it also provides a useful and practical approach to density estimation in real world data sets.

2 Bayesian Mixture Model

We begin our treatment of Gaussian mixtures by setting out the probabilistic specification of our model in Section 2.1. This specifies the joint distribution $p(D, \mu, T, s|\pi)$ over the data set D, the component means μ, the inverse covariances T and the discrete latent variables s, conditioned on the mixing coefficients π. Our goal is to optimize the values of the mixing coefficients π by maximizing the marginal likelihood of the data given by $p(D|\pi)$ which requires that we marginalize over μ, T and s. Since this marginalization is intractable, we resort to a variational approximation scheme based on maximization of a lower bound on $\ln P(D|\pi)$, discussed in Section 2.2. The evaluation of the lower bound itself is discussed in Section 2.3, and the procedure for optimizing the mixing coefficients is discussed in Section 2.4.

2.1 Model Specification

It is convenient to re-interpret the mixture distribution as a latent variable model, in which we introduce, for each data point x_n, a set of binary latent variables $s_{in} \in \{0, 1\}$ where $i = 1, \ldots, M$, and $\sum_{i=1}^{M} s_{in} = 1$. From a generative perspective these latent variables describe which component in the mixture gave rise to each of the data points, so that if a given data point x_n is generated from component j then $s_{in} = 1$ if $i = j$ and $s_{in} = 0$ if $i \neq j$. Conditional on $s = \{s_{in}\}$, the data points are assumed to be independently drawn from a Gaussian distribution with mean μ_i and inverse covariance T_i so that

$$P(D|\mu, T, s) = \prod_{n=1}^{N} \prod_{i=1}^{M} \mathcal{N}(x_n|\mu_i, T_i)^{s_{in}}. \qquad (3)$$

The latent variables $\{s_{in}\}$ are given discrete distributions governed by the mixing coefficients π_i

$$P(s|\pi) = \prod_{i=1}^{M} \prod_{n=1}^{N} \pi_i^{s_{in}}. \qquad (4)$$

If we marginalize (3) over the s_{in}, weighted by the prior (4), then we recover the expression (2) for the marginal distribution of the observed data, conditioned on the mixing coefficients and the component means and covariances.

The model specification is completed by introducing conjugate priors over the means and inverse covariances

$$P(\mu) = \prod_{i=1}^{M} \mathcal{N}(\mu_i|0, \beta I) \qquad (5)$$

$$P(T) = \prod_{i=1}^{M} \mathcal{W}(T_i|\nu, V) \qquad (6)$$

where β is a fixed parameter with a small value corresponding to a broad prior over μ, I is the unit matrix, $\mathcal{W}$ denotes the Wishart distribution, and ν and V are the degrees of freedom and scale matrix again chosen to give a broad prior for T.

Thus the joint distribution of all of the random variables, conditioned on the mixing coefficients, is given by

$$P(D, \mu, T, s|\pi) = P(D|\mu, T, s)P(s|\pi)P(\mu)P(T). \qquad (7)$$

This model can be expressed as a directed graph, as illustrated in Figure 1.

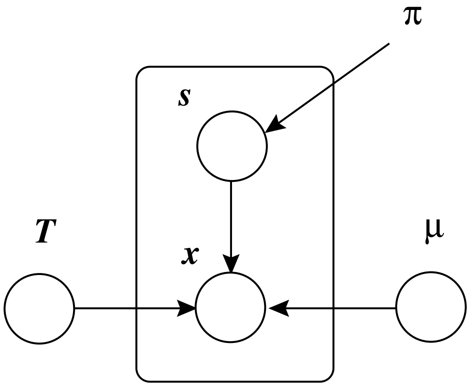

Figure 1: Representation of the Gaussian mixture model as a directed acyclic graph. The observed variable x is shown by the shaded node, while the box denotes a 'plate' representing the N independent samples from the data set.

2.2 Variational Approximation

In order to evaluate $P(D|\pi)$ we must marginalize (7) with respect to s, μ and T which is analytically intractable. We therefore use variational methods [9, 3] to find a tractable lower bound on $P(D|\pi)$. To simplify the notation we use θ to denote the $\{\mu, T, s\}$. Then the marginal likelihood which we wish to evaluate is given by

$$P(D|\pi) = \int P(D, \theta|\pi)\, d\theta.$$

Note that we use an integral to denote the joint integration over $\{\mu, T\}$ and summation over s. Variational methods involve the introduction of a distribution $Q(\theta)$ which, as we shall see shortly, provides an approximation to the true posterior distribution. Consider the following transformation applied to the log marginal likelihood

$$\begin{aligned}\ln P(D|\pi) &= \ln \int P(D, \theta|\pi)\, d\theta \\ &= \ln \int Q(\theta) \frac{P(D, \theta|\pi)}{Q(\theta)}\, d\theta\end{aligned}$$

$$\begin{aligned} &\geq \int Q(\theta) \ln \frac{P(D,\theta|\pi)}{Q(\theta)}\, d\theta \\ &= \mathcal{L}(Q) \end{aligned} \tag{8}$$

where we have applied Jensen's inequality. We see that the function $\mathcal{L}(Q)$ forms a rigorous lower bound on the true log marginal likelihood. The significance of this transformation is that, through a suitable choice for the Q distribution, the quantity $\mathcal{L}(Q)$ may be tractable to compute, even though the original log-likelihood function is not. From (8) it is easy to see that the difference between the true log marginal likelihood $\ln P(D|\pi)$ and the bound $\mathcal{L}(Q)$ is given by

$$\mathrm{KL}(Q\|P) = -\int Q(\theta) \ln \frac{P(\theta|D,\pi)}{Q(\theta)}\, d\theta \tag{9}$$

which is the Kullback-Leibler (KL) divergence between the approximating distribution $Q(\theta)$ and the true posterior $P(\theta|D,\pi)$. The relationship between the various quantities is shown in Figure 2.

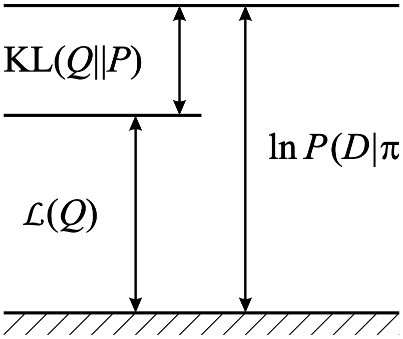

Figure 2: The quantity $\mathcal{L}(Q)$ provides a rigourous lower bound on the true log marginal likelihood $\ln P(D|\pi)$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$ between the approximating distribution $Q(\theta)$ and the true posterior $P(\theta|D,\pi)$.

The goal in a variational approach is to choose a suitable form for $Q(\theta)$ which is sufficiently simple that the lower bound $\mathcal{L}(Q)$ can readily be evaluated and yet which is sufficiently flexible that the bound is reasonably tight. We generally choose some family of Q distributions and then seek the best approximation within this family by maximizing the lower bound. Since the true log likelihood is independent of Q we see that this is equivalent to minimizing the Kullback-Leibler divergence.

Suppose we consider a completely free-form optimization over Q, allowing for all possible Q distributions. Using the well-known result that the KL divergence between two distributions $Q(\theta)$ and $P(\theta)$ is minimized by $Q(\theta) = P(\theta)$ we see that the optimal Q distribution is given by the true posterior, in which case the KL divergence is zero and the bound becomes exact. However, this will not lead to any simplification of the problem. In order to make progress it is necessary to consider a more restricted range of Q distributions.

Here we consider a constrained family of variational distributions by assuming that $Q(\theta)$ factorizes over subsets $\{\theta_i\}$ of the variables in θ, so that

$$Q(\theta) = \prod_i Q_i(\theta_i). \tag{10}$$

The KL divergence can then be minimized over all possible factorial distributions by performing a free-form minimization over the Q_i, leading to the following result

$$Q_i(\theta_i) = \frac{\exp \left\langle \ln P(D,\theta) \right\rangle_{k \neq i}}{\int \exp \left\langle \ln P(D,\theta) \right\rangle_{k \neq i} d\theta_i} \tag{11}$$

where $\langle \cdot \rangle_{k\neq i}$ denotes an expectation with respect to the distributions $Q_k(\theta_k)$ for all $k \neq i$. Note that this approach makes no assumptions about the form of the posterior distribution beyond the factorization implied by (10).

It is easily seen that, for conjugate hierarchical models, the expressions on the right hand side of (11) will have the same functional forms as in the priors. The sufficient statistics of each distribution Q_i will depend on moments of the other distributions $Q_{k\neq i}$ and so the expressions (11) represent an implicit solution for the variational posterior. These coupled equations can be solved by choosing some initialization for the sufficient statistics of the factors, and then iteratively updating them by taking each factor in turn and replacing its sufficient statistics by revised estimates given by the above equations. At each step of this re-estimation process, the lower bound will increase unless the variational posterior is already at a maximum of the bound.

We can apply this approach to the Gaussian mixture model by introducing a variational posterior distribution of the form

$$Q(\mu, T, s) = Q_\mu(\mu) Q_T(T) Q_s(s). \tag{12}$$

Application of (11) then gives the following solutions for the factors of the variational posterior

$$Q_s(s) = \prod_{n=1}^{N} \prod_{i=1}^{M} p_{in}^{s_{in}} \tag{13}$$

$$Q_\mu(\mu) = \prod_{i=1}^{M} \mathcal{N}(\mu_i | m_\mu^{(i)}, T_\mu^{(i)}) \tag{14}$$

$$Q_T(T) = \prod_{i=1}^{M} \mathcal{W}(T_i | \nu_T^{(i)}, V_T^{(i)}) \tag{15}$$

where we have defined

$$p_{in} = \frac{\widetilde{p}_{in}}{\sum_{j=1}^{M} \widetilde{p}_{jn}} \tag{16}$$

$$
\begin{aligned}
\widetilde{p}_{in} &= \exp(\langle \ln |T_i| \rangle /2 + \ln \pi_i \\
&\quad - \frac{1}{2} \mathrm{Tr}\{ \langle \mathrm{T_i} \rangle (\mathrm{x_n x_n^T} - \langle \mu_\mathrm{i} \rangle \mathrm{x_n^T} \\
&\quad - x_n \langle \mu_i \rangle^\mathrm{T} + \langle \mu_i \mu_i^\mathrm{T} \rangle)\}) \qquad (17)
\end{aligned}
$$

$$T_\mu^{(i)} = \beta I + \langle T_i \rangle \sum_{n=1}^{N} \langle s_{in} \rangle \qquad (18)$$

$$m_\mu^{(i)} = T_\mu^{(i)^{-1}} \langle T_i \rangle \sum_{n=1}^{N} x_n \langle s_{in} \rangle \qquad (19)$$

$$\nu_T^{(i)} = \nu + \sum_{n=1}^{N} \langle s_{in} \rangle \qquad (20)$$

$$
\begin{aligned}
V_T^{(i)} &= V + \sum_{n=1}^{N} x_n x_n^\mathrm{T} \langle s_{in} \rangle - \sum_{n=1}^{N} x_n \langle s_{in} \rangle \langle \mu_i^\mathrm{T} \rangle \\
&\quad - \langle \mu_i \rangle \sum_{n=1}^{N} x_n^\mathrm{T} \langle s_{in} \rangle + \langle \mu_i \mu_i^\mathrm{T} \rangle \sum_{n=1}^{N} \langle s_{in} \rangle . \qquad (21)
\end{aligned}
$$

The expected values in the above formulas are given by

$$\langle s_{in} \rangle = p_{in} \qquad (22)$$

$$\langle \mu_i \rangle = m_\mu^{(i)} \qquad (23)$$

$$\langle \mu_i \mu_i^\mathrm{T} \rangle = T_\mu^{(i)^{-1}} + m_\mu^{(i)} m_\mu^{(i)^\mathrm{T}} \qquad (24)$$

$$\langle T_i \rangle = \nu_T^{(i)} V_T^{(i)^{-1}} \qquad (25)$$

$$
\begin{aligned}
\langle \ln |T_i| \rangle &= \sum_{s=1}^{d} \psi((\nu_T^{(i)} + 1 - s)/2) \\
&\quad + d \ln 2 - \ln |V_T^{(i)}| . \qquad (26)
\end{aligned}
$$

Thus we see that the solutions for the variational factors Q_μ, Q_T and Q_s, given by (13), (14) and (15) respectively, are mutually coupled through their dependence on moments of the other factors. These can be solved iteratively as discussed above. Note that there will typically be multiple maxima in the variational bound and so in principle it may be beneficial to run the optimization a number of times using different initializations in order to find a good maximum. In practice we have found that, for the applications considered in this paper, a single initialization is sufficient to give good results.

2.3 Lower Bound on Marginal Likelihood

Given the functional forms for the variational factors Q_μ, Q_T and Q_s it is straightforward to evaluate the lower bound (8) to give

$$
\begin{aligned}
\mathcal{L} &= \langle \ln P(D|\mu, T, s) \rangle + \langle \ln P(s) \rangle \\
&\quad + \langle \ln P(\mu) \rangle + \langle \ln P(T) \rangle - \langle \ln Q_s(s) \rangle \\
&\quad - \langle \ln Q_\mu(\mu) \rangle - \langle \ln Q_T(T) \rangle \qquad (27)
\end{aligned}
$$

where

$$
\begin{aligned}
\langle \ln P(D|\mu, T, s) \rangle &= \sum_{i=1}^{M} \sum_{n=1}^{N} \langle s_{in} \rangle \left\{ \frac{1}{2} \langle \ln |T_i| \rangle \right. \\
&\quad - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \mathrm{Tr}(\langle \mathrm{T_i} \rangle (\mathrm{x_n x_n^T} - \\
&\quad \left. - x_n \langle \mu_i^\mathrm{T} \rangle - \langle \mu_i \rangle x_n^\mathrm{T} + \langle \mu_i \mu_i^\mathrm{T} \rangle))) \right\} \qquad (28)
\end{aligned}
$$

$$\langle \ln P(s) \rangle = \sum_{i=1}^{M} \sum_{n=1}^{N} \langle s_{in} \rangle \ln \pi_i \qquad (29)$$

$$\langle \ln P(\mu) \rangle = 2 \frac{Md}{2} \ln \left(\frac{\beta}{2\pi} \right) - \frac{\beta}{2} \sum_{i=1}^{M} \langle \mu_i^\mathrm{T} \mu_i \rangle \qquad (30)$$

$$
\begin{aligned}
\langle \ln P(T) \rangle &= M \left\{ - \frac{\nu d}{2} \ln 2 - \frac{d(d-1)}{4} \ln \pi \right. \\
&\quad \left. - \sum_{s=1}^{d} \ln \Gamma \left(\frac{\nu + 1 - s}{2} \right) + \frac{\nu}{2} \ln |V| \right\} + \\
&\quad + \frac{\nu - d - 1}{2} \sum_{i=1}^{M} \langle \ln |T_i| \rangle \\
&\quad - \frac{1}{2} \mathrm{Tr} \left(\mathrm{V} \sum_{\mathrm{i=1}}^{\mathrm{M}} \langle \mathrm{T_i} \rangle \right) \qquad (31)
\end{aligned}
$$

$$\langle \ln Q_s(s) \rangle = \sum_{i=1}^{M} \sum_{n=1}^{N} \langle s_{in} \rangle \ln \langle s_{in} \rangle \qquad (32)$$

$$\langle \ln Q_\mu(\mu) \rangle = \sum_{i=1}^{M} \left\{ - \frac{d}{2} (1 + \ln(2\pi)) + \frac{1}{2} \ln |T_\mu^{(i)}| \right\} \qquad (33)$$

$$
\begin{aligned}
\langle \ln Q_T(T) \rangle &= \sum_{i=1}^{M} \left\{ - \frac{\nu_T^{(i)} d}{2} \ln 2 - \frac{d(d-1)}{4} \ln \pi \right. \\
&\quad - \sum_{s=1}^{d} \ln \Gamma \left(\frac{\nu_T^{(i)} + 1 - s}{2} \right) + \frac{\nu_T^{(i)}}{2} \ln |V_T^{(i)}| \\
&\quad + \frac{\nu_T^{(i)} - d - 1}{2} \langle \ln |T_i| \rangle \\
&\quad \left. - \frac{1}{2} \mathrm{Tr} \left(\mathrm{V_T^{(i)}} \langle \mathrm{T_i} \rangle \right) \right\} \qquad (34)
\end{aligned}
$$

2.4 Optimizing the Mixing Coefficients

We have now obtained a variational lower bound $\mathcal{L}(Q)$ which approximates the true marginal log-likelihood

$\ln P(D|\pi)$. By maximizing this bound with respect to π we obtain our required estimates for the mixing coefficients. However, the solutions for the factors of the variational posterior, and hence the value of the lower bound, will depend on the values of π. We therefore adopt an EM procedure in which we alternately maximized $\mathcal{L}(Q)$ with respect to π (maximization step) and then optimize Q by iterative updating of the variational solutions for Q_μ, Q_T and Q_s (expectation step). For computational efficiency we perform just one re-estimation of each of the variational factors in each E-step, before re-estimating π in the M-step.

The re-estimation M-step equations for updating the π are obtained by setting the derivative of the lower bound with respect to π to zero, giving

$$\pi_i = \frac{1}{N} \sum_{n=1}^{N} p_{in}. \tag{35}$$

Numerical evaluation of the lower bound $\mathcal{L}$ after each M-step and after each update within each E-step provides a useful check on the software implementation, any such update should not lead to a decrease of $\mathcal{L}$. The improvement in $\mathcal{L}$ during the EM optimization can be used to monitor convergence and to set a suitable stopping criterion.

3 Results

We verify through experiments both that the maximum variational bound is a good score for model selection, and that the proposed iterative algorithm achieves maximum bound. We test the algorithm on synthetic and real data sets: 600 data points from a mixture of five Gaussians (means $[0,0]$, $[3,-3]$, $[3,3]$, $[-3,3]$, $[-3,-3]$ and covariances $[1,0;0,1]$, $[1,0.5;0.5,1]$, $[1,-0.5;-0.5,1]$, $[1,0.5;0.5,1]$, $[1,-0.5;-0.5,1]$); 900 data points from three mixtures of means $[0,-2]$,$[0,0]$, $[0,2]$ and same covariance $[2,0;0,0.2]$ from [5]; 400 data points from three mixtures of same mean $[0,0]$ and covariances $[1,0;0,0.2]$, $[0.02,-0.08;-0.08,1.5]$, $[0.5,0.4;0.05,0.05]$; "Enzime", "Acidity" and "Galaxy" univariate data sets from [12], and "Old Faithful" bivariate data from [8].

As a first check, we ran the variational optimization on a fixed number of components without updating the mixing coefficients, with an initial value selected through plain EM. Under this setting the variational likelihood bound becomes a model selection score. We compared the best model given by this score with the one selected through EM and cross-validation, and found that the variational score was invariantly maximum at the correct number of components on all of 100 instances of the 5-mixture synthetic data set (Figure 3). The same qualitative results were achieved on all considered data sets, indicating that finding the global maximum of the variational likelihood does select the correct model.

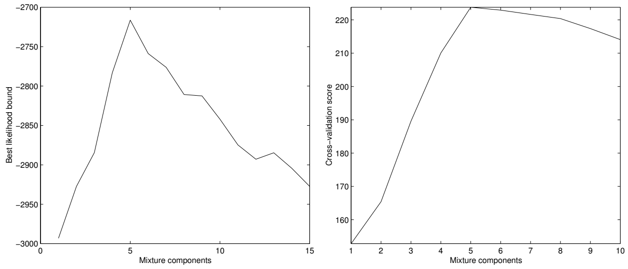

Figure 3: Best variational likelihood bound and plain EM cross-validation score as a function of the fixed assumed number of mixture components on 600 points drawn from a mixture of 5 bivariate Gaussians.

However, the power of our method consists in its ability to optimize mixing coefficients and variational score jointly. This could be a very difficult task because of the complexity of the likelihood manifold; for instance, in the case of plain EM it is very common to converge to a non-global maximum, especially if the number of components is large. Moreover, the variational likelihood manifold is very complex because of its symmetry with respect to mixture permutations, which replicates both global and local maxima. Our hope is that integrating out the parameters of the Gaussians to derive the variational likelihood smoothed the likelihood manifold removing most local maxima.

The possibility of local extrema makes the initial choice of mixture parameters important. If the initial means are equal or too close, it becomes hard to differentiate between components during the optimization. Under these circumstances the variational approximation converges very slowly, and as a result too many mixture components could get removed. This may happen because we update mixing coefficients after each variational iteration before full optimization, and a component that is out of place could be eventually removed if it does not find its place quickly enough. We approach this issue by initializing the means through K-means clustering. In order to avoid a strong initial bias from the assignment of components to K-means clusters, we choose large initial covariance matrices; otherwise in the beginning each Gaussian would be confined to its local cluster, and could remain in a local minimum. We found that K-means with large covariance initialization is enough to avoid local maxima.

We initialize the algorithm with a mixture of many components (15 in this paper) with equal mixing coefficients. The optimization has the property that Gaussians with similar parameters fitting the same cluster become unbalanced until one dominates and the others get removed. In addition, when a mixing coefficient is close to 0, it converges to 0 faster and faster; therefore we can remove components with very small mixing coefficients ($< 10^{-5}$). During the optimization, the variational bound increases with each iteration, by small amounts when all mixing coefficients are

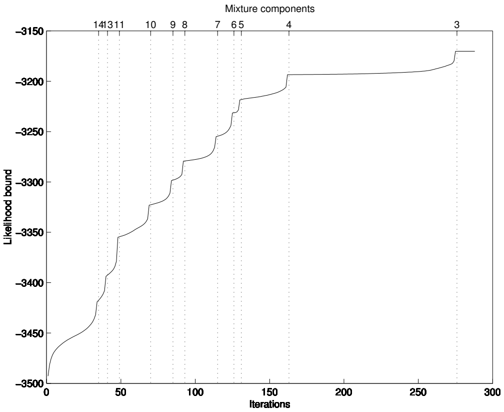

Figure 4: Variational likelihood bound over the model optimization of 900 data points drawn from a mixture of 3 same-covariance bivariate Gaussians. Initially the model had 15 mixtures. Vertical lines indicate cancellation of components.

large, and very fast when one coefficient is close to and drops to 0 (Figure 4). Although the convergence of the variational approximation can be quite slow, it was never slower than EM on the same mixing and Gaussian parameters. If necessary, optimization can be made faster by speculating that small and decreasing mixing coefficients converge to 0, and verifying directly if there is an increase in variational bound with such a change. Nevertheless, such heuristics of combining discrete search with continuous optimization were not employed in the results presented here.

We found that the model optimization automatically recovered the number of generating mixtures in all proposed synthetic data sets (Figure 5). In the case of the data set from [5], we recover the generating mixture even when given only 200 samples. "Old Faithful" data was fitted with three components (mixing 0.63, 0.33, 0.04). Also, the univariate data sets have been fitted with similar results as in [12] (Figure 6).

One advantage of the variational approximation is that it not only provides a likelihood bound but it also gives explicit values for means and covariances from the variational parameters. We test how close the variational parameters are to the maximum likelihood ones by comparing the log-likelihood of the data: under the best model found by the variational iteration; after running EM starting from the best variational parameters and keeping mixing coefficients fixed; and after running full EM initialized to the best variational parameters. Table 1 shows that EM does not significantly increase the likelihood, except for the case of the three mixtures with equal means. This is probably due to an inaccurate variational approximation for highly overlapping Gaussians. Nevertheless, the number of mixture components quickly converges to the correct one also for this data set.

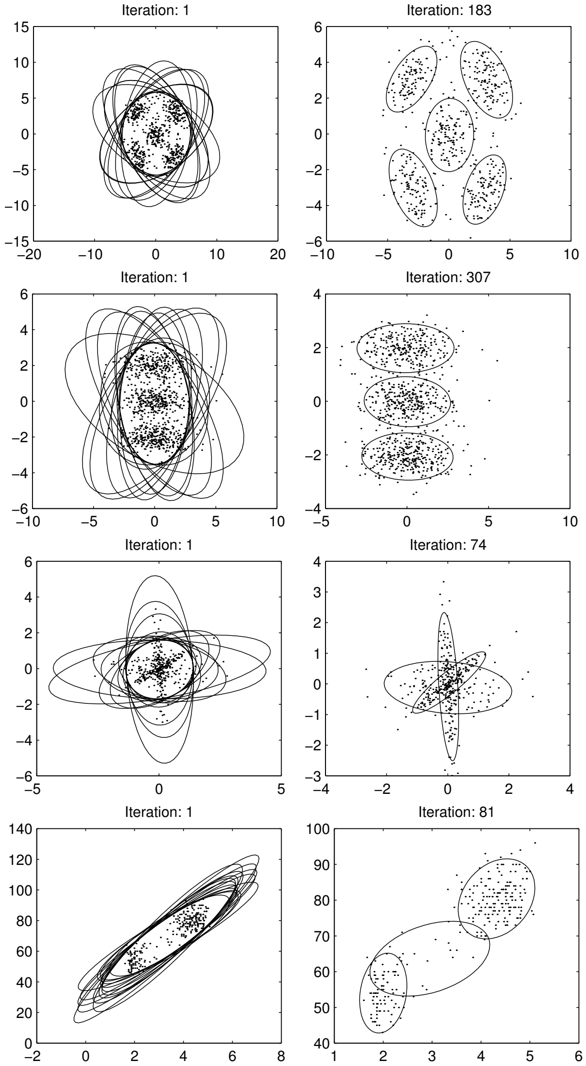

Figure 5: Initial and final configurations for the model optimization of three generated and one real data set: 5 Gaussians (600 points), 3 same-covariance Gaussians (900 points), 3 same-mean Gaussians (400 data points) and the "Old Faithful" data set.

As a last remark, we found that the number of mixture components selected by the algorithm was not sensitive to large variations in the non-informative priors for the means and covariance matrices.

4 Conclusions

In this paper we have shown how a discrete search over the number of components in a mixture distribution can be avoided through the introduction of continuous hyperparameters whose values are chosen to maximize the marginal likelihood. The framework has been developed for the case of mixtures of multivariate normal distributions. It is easily generalized to mixtures of models

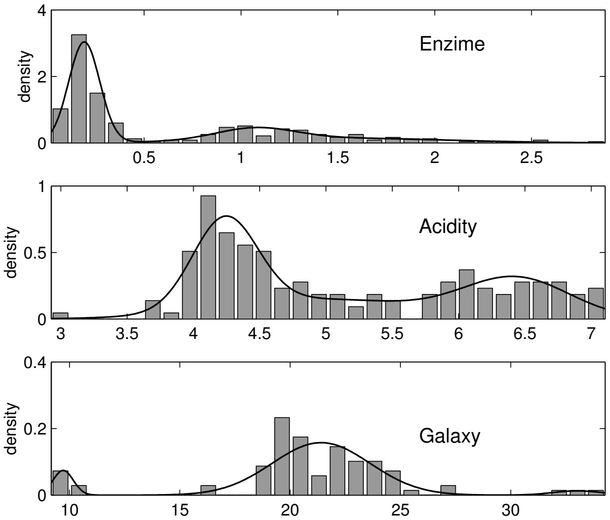

Figure 6: Model optimization of the "Enzime", "Acidity", and "Galaxy" data sets.

Data set	Variational	EM (fixed)	EM (full)
5 mixtures	-2577.51	-2577.46	-2577.46
3 mixtures (1^{st})	-3080.72	-3080.65	-3080.65
3 mixtures (2^{nd})	-712.213	-691.924	-689.619
Old Faithful	-1122.44	-1119.49	-1119.64
Enzime	-47.8791	-47.8504	-47.8268
Acidity	-178.917	-178.869	-178.754
Galaxy	-203.634	-203.482	-203.482

Table 1: The log-likelihood of the data sets under a mixture of Gaussians specified by: the best variational parameters; EM initialized to the best variational parameters without changing mixing coefficients; and same EM without the mixing constraint.

which can be specified as directed acyclic graphs of linear-Gaussian units with Wishart priors. Examples include mixtures of Kalman filters, and mixtures of Bayesian principal component analysis or factor analysis models.

Finally, we note that, due to the marginalization over component parameters, our approach does not suffer from the problem of singularities that plagues conventional maximum likelihood.

Acknowledgements

We would like to thank John Winn and Matthew Beal for helpful discussions.

References

[1] H. Attias. Learning parameters and structure of latent variables by variational Bayes, 1999. Proceedings Fifthteenth Conference on Uncertainty in Artificial Intelligence.

[2] C. M. Bishop. Bayesian PCA. In S. A. Solla M. S. Kearns and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 382–388. MIT Press, 1999.

[3] C. M. Bishop. Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, volume 1, pages 509–514. IEE, 1999.

[4] C. M. Bishop and J. Winn. Non-linear Bayesian image modelling. In *Proceedings Sixth European Conference on Computer Vision, Dublin*, volume 1, pages 4–17. Springer-Verlag, 2000.

[5] M. A. T. Figueiredo and A. K. Jain. Unsupervised selection and estimation of finite mixture models. In *International Conference on Pattern Recognition, Barcelona*, pages 87–90. IEEE, 2000.

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

[7] P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[8] W. Härdle. *Smoothing techniques with implementation in S.* Springer, New York, 1991.

[9] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. Kluwer, 1998.

[10] D. J. C. MacKay. Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

[11] A. Polymenis and D. M. Titterington. On the determination of the number of components in a mixture. *Statistics and Probability Letters*, 38:295–298, 1998.

[12] S. Richardson and P. Green. On bayesian analysis of mixtures with unknown number of compo nents. *Journal of the Royal Statistical Society B*, 59:731–792, 1997.