Blind Source Separation using Variational Expectation-Maximization Algorithm

Nikolaos Nasios and Adrian G. Bors
Dept. of Computer Science, University of York, York YO10 5DD, United Kingdom
E-mail: {n.n.adrian.bors}@cs.york.ac.uk

Abstract. In this paper we suggest a new variational Bayesian approach. Variational Expectation-Maximization (VEM) algorithm is proposed in order to estimate a set of hyperparameters modeling distributions of parameters characterizing mixtures of Gaussians. We consider maximum log-likelihood (ML) estimation for the initialization of the hyperparameters. The ML estimation is employed on distributions of parameters obtained from successive runs of the EM algorithm on the same data set. The proposed algorithm is used for unsupervised detection of quadrature amplitude and phase-shift-key modulated signals.

1 Introduction

A large variety of algorithms have been employed for data modeling. Expectation-maximization (EM) algorithm has been used successfully in many applications requiring the maximization of the log-likelihood of data [1]. More recently, Bayesian approaches consider the integration over distributions of parameters in order to achieve a better data modelling and generalization capability [2]. Various algorithms including stochastic approaches such as Monte Carlo Markov Chains (MCMC) and variational approximations have been used for Bayesian learning of graphical models [3, 4]. Variational Bayes (VB) algorithm is an EM like algorithm which is used in order to estimate hyperparameters characterizing distributions of parameters [5-7]. The performance of both EM and VB algorithms depends on a suitable initialization. In this paper we employ a maximum log-likelihood estimation for the initialization of the hyperparameters. We use distributions of parameters resulted from successive runs of the EM algorithm for the maximum log-likelihood estimation. We consider the graphical model of a mixture of Gaussians and we apply the proposed algorithm for unsupervised detection of modulated signals.

In Section 2 we introduce the variational Bayesian methodology for mixtures of Gaussians. Section 3 provides the maximum log-likelihood estimation for initializing the VEM algorithm while Section 4 describes the variational Bayes algorithm. Section 5 presents experimental results when applying the proposed algorithm in blind source separation of modulated signals and Section 6 provides the conclusions of the present study.
2 Bayesian methodology for mixtures of Gaussians

Due to their good approximation properties, mixtures of Gaussians have been used in various applications [5, 8]. A mixture of Gaussians can model any continuous probability function:

\[ p(x) = \sum_{i=1}^{N} \frac{\alpha_i}{\sqrt{(2\pi)^d|\Sigma_i|}} \exp \left[ -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) \right] \]  

(1)

where \( d \) is the dimension, \( \{\alpha, \Sigma, \mu\} \) represent parameters and \( N \) is the number of components. Furthermore, we consider that the sum of mixture probabilities is \( \sum_{i=1}^{N} \alpha_i = 1 \). In the classical estimation we estimate the parameters of the model. In Bayesian approaches we take into account the uncertainty in parameter estimation. Instead of parameters we estimate the hyperparameters modelling distributions of parameters. The \textit{a posteriori} probability is calculated as the integration over the space of parameters replacing the posterior with an approximation. The aim of variational Bayesian learning is to maximize the lower bound of the data log-likelihood probability approximation and therefore make the approximate posterior as close as possible to the true posterior distribution.

The parameters modelling a probability density function have their probabilities modelled as the conjugate priors [5, 6]. In the case of the parameters used for a mixture of Gaussians we have the mean, covariance matrix and mixing probability for each component. The conjugate prior distribution for the means is Gaussian, \( N(\mu|m, \beta S) \) where \( \beta \) is a scaling factor:

\[ N(\mu|m, \beta S) \sim \frac{1}{\sqrt{(2\pi)^d|\beta S|}} \exp \left[ -\frac{1}{2}(\mu - m)^T (\beta S)^{-1}(\mu - m) \right] \]  

(2)

A Wishart distribution \( W(\Sigma|\nu, S) \) is the conjugate prior for the inverse covariance matrix:

\[ W(\Sigma|\nu, S) \sim \frac{|S|^{-\nu/2} |\Sigma|^{(\nu-d-1)/2}}{2^{d/2} \pi^{d(d-1)/4} \prod_{k=1}^{d} \Gamma(\nu \frac{d-1}{2})} \exp \left[ -\frac{Tr(S^{-1} \Sigma)}{2} \right] \]  

(3)

where \( \nu \) are the degrees of freedom, \( Tr \) denotes the trace of the resulting matrix (the sum of the diagonal elements) and \( \Gamma(x) \) represents the Gamma function:

\[ \Gamma(x) = \int_{0}^{\infty} \tau^{x-1} \exp(-\tau) d\tau \]  

(4)

For the mixture probabilities we consider a Dirichlet distribution \( D(\alpha|\lambda_1, \ldots, \lambda_N) \):

\[ D(\alpha|\lambda_1, \ldots, \lambda_N) = \frac{\Gamma(\sum_{i=1}^{N} \lambda_i)}{\prod_{j=1}^{N} \Gamma(\lambda_j)} \prod_{i=1}^{N} \alpha_i^{\lambda_i - 1} \]  

(5)

The variational learning is expected to provide better data modelling by taking into account the uncertainty in the parameter estimation. On the other hand it provides better generalization while maintaining the good localization and modelling capabilities.
3 Maximum log-likelihood hyperparameter initialization

For certain datasets EM algorithm may not converge due to an unsuitable initialization. If we increase the number of parameters used in the modelling we are facing an even more challenging problem in choosing their initial values. Usually, random initialization is employed for EM or Variational Bayes algorithms. In this paper we adopt a hierarchical approach to the hyperparameter estimation. In the first stage we employ the EM algorithm using a set of random initializations. After several runs of the EM algorithm on the same data set, we form distributions of the parameters provided by the EM algorithm. Afterwards, maximum log-likelihood estimation is employed onto these distributions of parameters in order to initialize the hyperparameters for the Variational Expectation-Maximization algorithm.

The EM algorithm for Mixture of Gaussians model is applied on the given data in the first stage. In the E-step the \textit{a posteriori} probabilities are estimated:

\[
\hat{P}_{EM}(i|x_j) = \frac{\hat{\alpha}_i (|\Sigma_i|)^{-1/2} \exp \left[ -\frac{1}{2}(x_j - \hat{\mu}_i)^T \Sigma_i^{-1} (x_j - \hat{\mu}_i) \right]}{\sum_{k=1}^N \hat{\alpha}_k (|\Sigma_k|)^{-1/2} \exp \left[ -\frac{1}{2}(x_j - \hat{\mu}_k)^T \Sigma_k^{-1} (x_j - \hat{\mu}_k) \right]} \tag{6}
\]

In the M-step we update the parameters of the Gaussian mixture model:

\[
\hat{\alpha}_i = \frac{\sum_{j=1}^M \hat{P}_{EM}(i|x_j)}{M} \tag{7}
\]

\[
\hat{\mu}_{i,EM} = \frac{\sum_{j=1}^M x_j \hat{P}_{EM}(i|x_j)}{\sum_{j=1}^M \hat{P}_{EM}(i|x_j)} \tag{8}
\]

\[
\hat{\Sigma}_i = \frac{\sum_{j=1}^M \hat{P}_{EM}(i|x_j)(x_j - \hat{\mu}_{j,EM})(x_j - \hat{\mu}_{j,EM})^T}{\sum_{j=1}^M \hat{P}_{EM}(i|x_j)} \tag{9}
\]

We run the EM algorithm \( L \) times considering random initializations. All the parameters estimated in each of the runs are stored individually, forming parameter distributions. We assume that these distributions can be characterized parametrically by a set of hyperparameters. The parametric description of these probabilities is given by (2) for means \( \mu \), by (3) for covariance matrices \( \Sigma \), and by (5) for mixing probabilities \( \alpha \).

The next step consists in estimating the hyperparameters characterizing the distributions formed in the previous step. This estimation would correspond to a second level of embedding, characterizing the initial estimation of the hyperparameters. The distributions of the means resulting from the EM algorithm (the outputs from (8)) can be modelled as a mixture of Gaussians. We apply a second EM algorithm onto the distributions of parameters. The updating equations of the second EM are similar with (6), (7), (8) and (9). In the second EM we consider the given data samples \( x_j, j = 1, \ldots, M \) as the initial starting points for the centers of the mean distributions. The hypermeans \( \hat{\mathbf{m}}(0) \) are calculated as the average of the means resulting from several runs of the second EM. The corresponding covariance matrices for the Gaussian distribution of means \( \mathbf{S} \), are stored as well. The parameter \( \beta \) represents a scaling factor of the covariance
matrices corresponding to distribution $\hat{\Sigma}$, resulting from the given data set according to (9), and that of the mean distribution, $\bar{\Sigma}$, respectively. This parameter is initialized as the average of the eigenvalues of the matrix $\Sigma \Sigma^{-1}$, which can be calculated using the trace:

$$\beta_i(0) = \frac{\sum_{i=1}^{L} Tr(\Sigma_{ik} S_{ik}^{-1})}{Ld}$$

(10)

where $L$ is the number of runs for the EM algorithm.

The Wishart distribution $W(\Sigma|\nu, \bar{S})$ characterizes the covariance matrix. We initialize the degrees of freedom $\nu_i(0) = d$, while for the initialization of $S$ we consider the distribution of $\Sigma$ resulting from (9). We apply a Cholesky factorization onto the matrices $\Sigma_i$, $k = 1, \ldots, L$ resulted from successive runs of the EM algorithm. The Cholesky factorization results into an upper triangular matrix $R_k$ and a lower triangular matrix $R_k$ such that:

$$\Sigma_{ik}^{-1} = R_k R_k^T$$

(11)

We generate $d$ independent samples from a normal distribution of variance 1 $N(0, 1)$, and form a vector denoted as $\mathbf{N}$. The matrix $S_i$ will be initialized as [2]:

$$S_i(0) = \sum_{k=1}^{L} R_k \mathbf{N} (\mathbf{N} R_k)^T$$

(12)

For the Dirichlet parameters we use the maximum log-likelihood estimation for (5). After applying the logarithm on (5) and differentiating the resulting expression with respect to the parameters $\lambda_i$, $i = 1, \ldots, N$ we obtain the following equation which is applied iteratively:

$$\psi(\lambda_{i,t}) = \psi(\sum_{k=1}^{N} \lambda_{i,t}) + \log E[\hat{\lambda}_i]$$

(13)

where $t$ is the iteration number, $\log E[\hat{\lambda}_i]$ is the expectation of the mixing probability $\hat{\lambda}_i$, which is derived from the distributions obtained from successive runs of equation (7), and $\psi(x)$ is the digamma function (the logarithmic derivative of the Gamma function):

$$\psi(\lambda_i) = \frac{\Gamma'(\lambda_i)}{\Gamma(\lambda_i)}$$

(14)

where $\Gamma(x)$ function is provided in (4). The mean of the mixing probability distribution is considered as an appropriate estimate for $E[\hat{\lambda}_i]$. The parameter $\lambda_i$ is initialized by inverting the digamma function, $\lambda_{i,0} = \psi^{-1}(\log E[\hat{\lambda}_i])$. The iterative algorithm uses Newton’s method for updating $\lambda_i$ follows:

$$\lambda_{i,t} = \lambda_{i,t-1} - \frac{\psi(\lambda_{i,t}) - \psi(\lambda_{i,t-1})}{\psi'(\lambda_{i,t})}$$

(15)

Just a few iterations of (13) and (15) are necessary, and the result achieved at convergence provides the Dirichlet parameters $\lambda_i(0)$, $i = 1, \ldots, N$

4 Variational Bayes algorithm

Integrating over the entire parameter space would amount to a very heavy computational task, involving multidimensional integrals. Variational Bayes algorithm has been derived in order to estimate the hyperparameters of a mixture
model [5, 6]. In our approach we use the initialization provided by the maximum log-likelihood for the initialization of the Bayesian estimation algorithm. The variational Bayes is an iterative algorithm which consists of two steps at each iteration: variational expectation (VB-E) and variational maximization (VB-M).

In the first step we compute the a posteriori probabilities, given the hidden variable distributions and their hyperparameters. In the VB-M step we find the hyperparameters that maximize the log-likelihood, given the observed data and their a posteriori probabilities.

In the VB-E step for a mixture of Gaussians model we calculate the a posteriori probabilities for each data sample $x_j$, depending on the hyperparameters:

$$
\hat{P}(i|x_j) = \exp \left[ -\frac{1}{2} \log |S_i| + \frac{1}{2} d \log 2 + \frac{1}{2} \sum_{k=1}^{d} \psi \left( \frac{\nu_j + 1}{2} \right) + \psi(a_k) - \psi \left( \sum_{k=1}^{N} a_k \right) - \frac{\nu_j}{2} (x_j - m_i)^T S_i^{-1} (x_j - m_i) - \frac{d}{2} \beta_i \right]
$$

where $i = 1, \ldots, N$ is the mixture component, $d$ is number of dimensions, $j = 1, \ldots, M$ denotes the data index, and $\psi(x)$ is the digamma function from (14).

In the VB-M step we perform an intermediary calculation of the mean parameter as in the EM algorithm, but considering the a posteriori probabilities from (16):

$$
\hat{\mu}_{i, VEM} = \frac{\sum_{j=1}^{M} x_j \hat{P}(i|x_j)}{\sum_{j=1}^{M} \hat{P}(i|x_j)}
$$

The hyperparameters of the distribution are updated as follows:

$$
m_i = \frac{\beta_i(0)m_i(0) + \sum_{j=1}^{M} \hat{P}(i|x_j)x_j}{\beta_i(0) + \sum_{j=1}^{M} \hat{P}(i|x_j)}
$$

$$
S_i = S_i(0) + \sum_{j=1}^{M} \hat{P}(i|x_j)(x_j - \hat{\mu}_{i, VEM})(x_j - \hat{\mu}_{i, VEM})^T + \frac{\beta_i(0)(\hat{\mu}_{i, VEM} - m_i(0))(\hat{\mu}_{i, VEM} - m_i(0))^T \sum_{j=1}^{M} \hat{P}(i|x_j)}{\beta_i(0) + \sum_{j=1}^{M} \hat{P}(i|x_j)}
$$

while the additional hyperparameters for Wishart and Dirichlet distributions are updated as:

$$
\beta_i = \beta_i(0) + \sum_{j=1}^{M} \hat{P}(i|x_j); \quad \nu_i = \nu_i(0) + \sum_{j=1}^{M} \hat{P}(i|x_j); \quad \lambda_i = \lambda_i(0) + \sum_{j=1}^{M} \hat{P}(i|x_j)
$$

The effectiveness of the modeling is shown by the increase in the log-likelihood with each iteration. The convergence is achieved when we obtain a small variation in the log-likelihood for the given set of a posteriori probabilities:

$$
| \sum_{j=1}^{M} \log \hat{p}(x_j|\Phi(t)) - \sum_{j=1}^{M} \log \hat{p}(x_j|\Phi(t-1)) | < \varepsilon
$$
where \( p(x_j|\Phi(t)) \) is the probability function for the given set of hyperparameters \( \Phi \) at iteration \( t \), and \( \varepsilon \) is a small quantity. The number of mixture components is found using minimum description length (MDL) criterion.

## 5 Experimental results

We have applied the proposed algorithm in blind signal detection problems. We consider two cases of modulated signals: quadrature amplitude modulated signals (QAM) and phase-shifting-key (PSK) modulated signals. The perturbation channel equations considered in the case of 8-PSK signals are provided by [8]:

\[
x_I(t) = I(t) + 0.2I(t-1) - 0.2Q(t-1) + 0.04Q(t - 1) + \mathcal{N}(0, 0.11)
\]

\[
x_Q(t) = Q(t) + 0.2Q(t-1) + 0.2I(t) + 0.04I(t - 1) + \mathcal{N}(0, 0.11)
\]

where \((x_I(t), x_Q(t))\) makes up the in-phase and in-quadrature signal components at time \( t \) on the communication line, and \( I(t) \) and \( Q(t) \) correspond to the signal symbols (there are eight signal symbols in 8-PSK, equi-distantly located on a circle). The noise considered in this case is Gaussian and corresponds to SNR = 22 dB. We consider all possible symbol combinations for \((I, Q)\) and we generate a total of 64 signals which can be grouped in 8 signal constellations corresponding to the distorted signals [8]. We have generated 960 signals, by assuming equal probabilities for all intersymbol combinations. The signal constellations are represented in Figure 1. For 4-QAM signals we assume only additive noise, with SNR of 8 dB.

![Blind detection of 8-PSK modulated signals using VEM algorithm](image)

**Fig. 1.** Blind detection of 8-PSK modulated signals using VEM algorithm.
The blind detection problem is treated as an unsupervised classification task in which we want to model the superclusters (each formed from 8 clusters). The VEM algorithm properly initialized with the maximum likelihood estimation from distributions of parameters is applied on the given data. The covariance matrices characterizing the Wishart distribution $\mathbf{S}$, and the distribution of the means $\beta \mathbf{S}$, as well as the initial location and the ideal location for the hypermeans are marked in Figure 1. In Figure 2 the convergence of the mean distributions measured by the Kullback-Leibler (KL) divergence [2] is shown, while the global convergence for the proposed algorithm and for the variational Bayes (VB) algorithm considering various random initializations [5,6] is displayed in Figure 3. The favourable initialization for the VEM algorithm, achieved by applying maximum log-likelihood techniques can be easily identified on the curve marked with circles from Figure 3. The MDL criterion found four components for 4-QAM and eight components for 8-PSK. Table 1 shows comparative results when applying VEM, VB and EM algorithms on 4-QAM and 8-PSK modulated signals. The errors are measured in terms of global estimation by using KL divergence for the posterior distributions and misclassification errors on both training and testing set, and locally for the estimation of the individual parameters, respectively for the hypermean bias and for mixing probability bias. We have considered eight different random initializations for the VB and EM algorithms. We can observe from all these results that VEM algorithm provides a better estimation of the model parameters and achieves better source separation while the number of its necessary iterations is lower than in the other algorithms.

6 Conclusions

We propose a new Bayesian estimation algorithm applied to mixtures of Gaussians models. The proposed algorithm has two stages. In the first stage we model distributions of parameters resulting from repetitive runs of the EM algorithm on the same data set. In the second stage we apply maximum log-likelihood estimation in order to obtain initial estimators for the proposed variational expectation-maximization algorithm. We have considered appropriate hyperparameter initial
estimates for the parameter distributions under consideration: normal for the means, Wishart for the covariance matrix, and Dirichlet for the mixing probabilities. The proposed algorithm is compared with variational Bayes, which considers a similar updating algorithm, but random initialization, and with EM algorithm using random initialization. The algorithms have been tested on data sets representing 8-PSK and 4-QAM modulated signals, under inter-symbol and co-channel interference, when additive Gaussian noise is assumed. The experimental results show that the proposed VEM algorithm eliminates the dependence on the initialization which characterizes EM and VB algorithms and provides better estimation for the individual model parameters.

### References