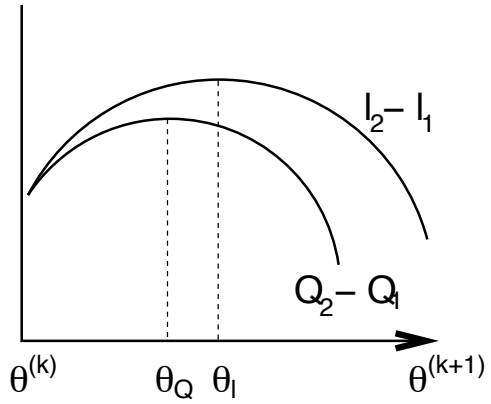
University of Cambridge Engineering Part IIB

Module 4F10: Statistical Pattern Processing

Handout 4: Expectation-Maximisation



Phil Woodland
pcw@eng.cam.ac.uk
Lent 2007

Introduction

In the last lecture we looked at Gaussian mixture models and found that an iterative procedure could be used to estimate the parameters of the Gaussian mixture model.

The iterative procedure for Gaussian Mixtures was a specific instance of the Expectation-Maximisation (EM) Algorithm which can be applied in many cases when direct maximum likelihood parameter estimation is not possible without knowledge of the values of hidden or latent variables. In the case of the Gaussian mixture model the latent variable determines which of the Gaussian mixture components is associated with each vector in the training set for the model.

In this lecture we will examine the

- mathematical basis of E-M for Gaussian mixtures
- auxiliary functions
- an alternative general formulation of E-M
- application of E-M to continuous and discrete latent variables

Deriving the E-M Mixture Updates

First consider a mixture distribution in which the parameter values (means, covariances, component priors) are changed from $\boldsymbol{\theta}^{(k)}$ on the k^{th} iteration to $\boldsymbol{\theta}^{(k+1)}$ on the $k+1^{\text{th}}$ iteration, with changes in PDF from $p(\mathbf{x}|\boldsymbol{\theta}^{(\mathbf{k})})$ to $p(\mathbf{x}|\boldsymbol{\theta}^{(\mathbf{k+1})})$. The increase in log likelihood is

$$l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} \log\left(\frac{p(\mathbf{x}_i|\boldsymbol{\theta}^{(k+1)})}{p(\mathbf{x}_i|\boldsymbol{\theta}^{(k)})}\right)$$

For a mixture distribution, denoting the m^{th} mixture component as ω_m,

$$\begin{aligned} & l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)}) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{p(\mathbf{x}_i|\boldsymbol{\theta}^{(k)})} \sum_{m=1}^{M} \left(p(\mathbf{x}_i, \omega_m|\boldsymbol{\theta}^{(k+1)})\right)\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{p(\mathbf{x}_i|\boldsymbol{\theta}^{(k)})} \sum_{m=1}^{M} \left(\frac{p(\mathbf{x}_i, \omega_m|\boldsymbol{\theta}^{(k+1)}) P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})}{P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})}\right)\right) \end{aligned}$$

Since $\log()$ is strictly concave we can use **Jensen's Inequality** which states that for $\lambda_m \geq 0$ and $\sum_m \lambda_m = 1$

$$\log\left(\sum_{m=1}^{M} \lambda_m x_m\right) \geq \sum_{m=1}^{M} \lambda_m \log\left(x_m\right)$$

Now using the numerator $P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})$ as λ_m gives

$$\begin{aligned} & l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)}) \geq \\ & \sum_{i=1}^{n} \sum_{m=1}^{M} P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) \log\left(\frac{p(\mathbf{x}_i, \omega_m|\boldsymbol{\theta}^{(k+1)})}{p(\mathbf{x}_i|\boldsymbol{\theta}^{(k)}) P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})}\right) \end{aligned}$$

which can be written as

$$l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$$

where

$$Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \sum_{i=1}^{n} \sum_{m=1}^{M} P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) \log\left(p(\mathbf{x}_i, \omega_m|\boldsymbol{\theta}^{(k+1)})\right)$$

which is known as the auxiliary function (more on this later).

So in other words, the difference

$$Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$$

gives a lower bound on the increase in the log likelihood. Given that $Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$ depends only on the old parameters, then if we maximise the value of $Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)})$ the value of the log likelihood lower bound will also be maximised.

To maximise, find the derivatives of $Q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)})$ with respect to the new parameters and equate to zero, noting that for the case of the component priors (mixture weights) again a Lagrange multiplier solution is needed. It can also be shown that the maximum that is found here is a global maximum of the auxiliary function.

This leads to the update equations for the mixture parameters presented earlier

Jensen's Inequality

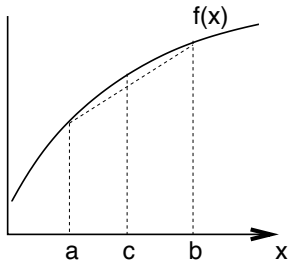
One useful inequality, commonly used in the derivation of the update formulae for mixture models, is *Jensen's inequality*. It states that

$$f\left(\sum_{m=1}^{M}\lambda_m x_m\right) \geq \sum_{m=1}^{M}\lambda_m f(x_m)$$

where $f()$ is any *concave function* and

$$\sum_{m=1}^{M}\lambda_m = 1, \quad \lambda_m \geq 0 \;\; m = 1,\ldots,M$$

As shown above, this can be used in the derivation of the EM algorithm for Gaussian mixture distributions.



A simple example is given above. Let $c = (1-\lambda)a + \lambda b$. From the diagram

$$f(c) = f((1-\lambda)a + \lambda b) \geq (1-\lambda)f(a) + \lambda f(b)$$

Kullback-Leibler Distance

A related derivation uses properties of the Kullback-Leibler distance between two PDFs. Consider two PDFs, $p(\mathbf{x})$ and $q(\mathbf{x})$. Looking at the relative entropy, or *Kullback-Leibler* distance, $\mathcal{D}(p(\mathbf{x}), q(\mathbf{x}))$,

$$\begin{aligned}\mathcal{D}(p(\mathbf{x}), q(\mathbf{x})) &= \int p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x} \\ &= -\int p(\mathbf{x}) \log\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}\end{aligned}$$

Using $\log(y) \leq y - 1$, we can write

$$\begin{aligned}\int p(\mathbf{x}) \log\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} &\leq \int p(\mathbf{x}) \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} - 1\right) d\mathbf{x} \\ &= \int (q(\mathbf{x}) - p(\mathbf{x}))\, d\mathbf{x} \\ &= 0\end{aligned}$$

This gives the following inequality

$$\int p(\mathbf{x}) \log\left(p(\mathbf{x})\right) d\mathbf{x} \geq \int p(\mathbf{x}) \log\left(q(\mathbf{x})\right) d\mathbf{x}$$

Similarly for the discrete version

$$\sum_{\forall \mathbf{x}} P(\mathbf{x}) \log\left(P(\mathbf{x})\right) \geq \sum_{\forall \mathbf{x}} P(\mathbf{x}) \log\left(Q(\mathbf{x})\right) d\mathbf{x}$$

where $Q(\mathbf{x})$ and $P(\mathbf{x})$ are valid PMFs. It directly follows from these inequalities that

$$\mathcal{D}(p(\mathbf{x}), q(\mathbf{x})) \geq 0$$

KL Distance for Gaussians

For the case of two Gaussian distributions the KL distance has a simple closed form solution. Consider

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ q(\mathbf{x}) &= \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \end{aligned}$$

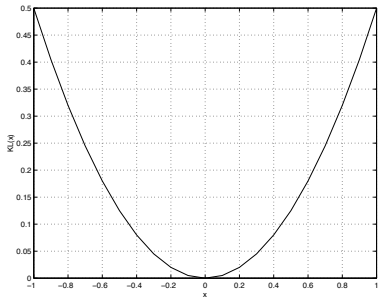
Then the KL distance between the two is given by

$$\begin{aligned} \mathcal{D}(p(\mathbf{x}), q(\mathbf{x})) = \frac{1}{2} \Big(& \text{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1 - \boldsymbol{I}) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ & + \log \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \right) \Big) \end{aligned}$$

For a simple example where

$$\begin{aligned} p(x) &= \mathcal{N}(x; 0, 1) \\ q(x) &= \mathcal{N}(x; \mu, 1) \end{aligned}$$

Then the plot as we vary μ is given by



Expectation Maximisation

EM is a general iterative optimisation technique. We would like a new estimate so that for the parameters at the $k+1^{th}$ iteration

$$l(\boldsymbol{\theta}^{(k+1)}) \geq l(\boldsymbol{\theta}^{(k)})$$

Alternatively we aim to ensure that

$$l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)}) \geq 0$$

We introduce a new set of discrete random variables $\mathbf{Z}$ which are dependent on the observations $\mathbf{X}$ and model parameters $\boldsymbol{\theta}^{(k)}$. From the definition of a PMF we can write

$$\begin{aligned} & \log(p(\mathbf{X}|\boldsymbol{\theta}^{(k+1)})) - \log(p(\mathbf{X}|\boldsymbol{\theta}^{(k)})) = \\ & \quad \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \left(\log(p(\mathbf{X}|\boldsymbol{\theta}^{(k+1)})) - \log(p(\mathbf{X}|\boldsymbol{\theta}^{(k)})) \right) \end{aligned}$$

since

$$\begin{aligned} \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log(p(\mathbf{X}|\boldsymbol{\theta}^{(k+1)})) &= \log(p(\mathbf{X}|\boldsymbol{\theta}^{(k+1)})) \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \\ &= \log(p(\mathbf{X}|\boldsymbol{\theta}^{(k+1)})) \end{aligned}$$

From the definition of conditional probability

$$p(\mathbf{X}|\boldsymbol{\theta}^{(k+1)}) = \frac{p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta}^{(k+1)})}{P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k+1)})}$$

so

$$\sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log(p(\mathbf{X}|\boldsymbol{\theta}^{(k+1)})) = \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)})}{P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k+1)})} \right)$$

and similarly for the second term.

EM (cont)

We can now write

$$\begin{aligned} l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)}) &= \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right) \\ &- \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k+1)}) \right) \\ &- \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k)}) \right) \\ &+ \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \right) \end{aligned}$$

From the previous inequality

$$\sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \right) \geq \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k+1)} \right.$$

So it follows that

$$\begin{aligned} l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)}) &\geq \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right) \\ &- \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k)}) \right) \end{aligned}$$

If we can ensure that the right-hand size is positive then the left-hand side must also be positive. So EM states that if

$$\sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right) \geq \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k)}) \right)$$

then

$$l(\boldsymbol{\theta}^{(k+1)}) \geq l(\boldsymbol{\theta}^{(k)})$$

EM (cont)

It is common to define the *auxiliary function* as

$$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right)$$

and for the continuous version

$$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right) d\mathbf{Z}$$

Thus the auxiliary function is the *expected value of the log likelihood of the joint distribution* of $\mathbf{Z}$ and $\mathbf{X}$.

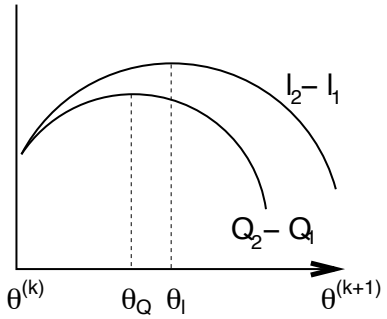
Note that if the auxiliary function increases then the likelihood is guaranteed increase, i.e. if

$$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$$

then

$$l(\boldsymbol{\theta}^{(k+1)}) \geq l(\boldsymbol{\theta}^{(k)})$$

EM (cont)



The diagram above illustrates the optimisation. The graph shows two lines,

$$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$$

and

$$l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)})$$

The maxima of the two lines occur at $\boldsymbol{\theta}_Q$ and $\boldsymbol{\theta}_l$

Using the value at $\boldsymbol{\theta}_Q$ does yield an increase in the log-likelihood, but has not hit the maximum value. It is necessary to iterate to find a local maximum of the likelihood. In common with gradient descent schemes EM is only guaranteed to find a local, not global, maximum of the likelihood function.

Hidden Variables

The set of variables $\mathbf{Z}$ are called hidden or latent variables. They may be discrete variable (for example in mixture models), or continuous (for example in *Factor Analysis*).

The set of data $\{\mathbf{Z}, \mathbf{X}\}$ is sometimes referred to as the *complete dataset*. It consists of the *observed* data $\mathbf{X}$ (the feature vectors) and *unobserved* data $\mathbf{Z}$ (the hidden variables).

The nature of the latent variable is highly important. It must be selected so that:

- given the complete dataset $\{\mathbf{Z}, \mathbf{X}\}$ it is simple to optimise $\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)})$ with respect to $\boldsymbol{\theta}^{(k+1)}$;
- the difference between the likelihoods and auxiliary functions is small. The difference is given by

$$\Big(l(\boldsymbol{\theta}^{(k+1)}) - l(\boldsymbol{\theta}^{(k)})\Big) - \Big(\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})\Big) = \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(\frac{P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)})}{P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k+1)})} \right)$$

  As the increase in the auxiliary function is a *lower* bound on the increase in the log-likelihood, the tighter the bound the better.

In practise the ability to optimise the auxiliary function is more important. The second consideration affects the rate of convergence of the algorithm.

EM Optimisation

We have seen that simply maximising the auxiliary function *does not* (in general) take us to the ML solution we need to iterate. From the definition of the *auxiliary* function

$$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right)$$

EM can be seen to have two stages:

1. **Expectation**: given the current set of parameters $\boldsymbol{\theta}^{(k)}$ calculate the posterior PMF of the latent variable, $P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)})$. Given this distribution calculate the expected value of log-likelihood of the complete dataset in terms of the new model parameters, $\boldsymbol{\theta}^{(k+1)}$,

   $$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \mathcal{E}\left\{ \log \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right) |\mathbf{X}, \boldsymbol{\theta}^{(k)} \right\}$$

   where the expectation is over the distribution of the latent variables given the current model parameters. The auxiliary function is only a function of the new parameters $\boldsymbol{\theta}^{(k+1)}$.

2. **Maximisation**: maximise the value of the auxiliary function, $\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)})$, with respect to $\boldsymbol{\theta}^{(k+1)}$.

One major issue is that some initial set of model parameters $\boldsymbol{\theta}^{(0)}$ are required. If there are many local maxima then EM will only find a local, not global, maximum. Which maxima is obtained depends on the choice of the initial parameters.

Mixture Models & E-M

Mixture models of a particular family of distributions are very well suited for estimation using EM (e.g. Gaussian, Poisson etc). For mixture models the hidden variable is which component of the mixture should be associated with each training vector.

We will use a discrete hidden variable to *indicate* which of the components of the mixture model generated an observation:

$$z_{ij} = \begin{cases} 1 & \text{observation } \mathbf{x}_i \text{ was generated by component } \omega_j \\ 0 & \text{otherwise} \end{cases}$$

If we look at a single point $\mathbf{x}_i$ and know that it was generated by component ω_j, then we can write

$$\begin{aligned} p(\mathbf{z}_i, \mathbf{x}_i|\boldsymbol{\theta}) &= p(\mathbf{x}_i|\omega_j, \boldsymbol{\theta}_j) P(\omega_j) \\ &= \prod_{m=1}^{M} \left[p(\mathbf{x}_i|\omega_m, \boldsymbol{\theta}_m) P(\omega_m) \right]^{z_{im}} \end{aligned}$$

As all the data points are independent then the hidden variables associated with the data points will also be independent of one another. So

$$p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{z}_i, \mathbf{x}_i|\boldsymbol{\theta})$$

Taking a $\log()$ we can write

$$\log(p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})) = \sum_{i=1}^{n} \log(p(\mathbf{z}_i, \mathbf{x}_i|\boldsymbol{\theta}))$$

This is the basis of estimating mixture model parameters.

Expectation

As mentioned in the *expectation* stage we need to compute $P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)})$. As all the observations are independent we need only consider $P(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(k)})$, where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix}$$

and

$$\mathbf{z}_i = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{iM} \end{bmatrix}$$

Recall that we will need the probability that the observation $\mathbf{x}_i$ was generated by component ω_j, which we saw before may be simply written as

$$P(\omega_j|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) = \frac{p(\mathbf{x}_i|\omega_j, \boldsymbol{\theta}_j^{(k)})P^{(k)}(\omega_j)}{\sum_{m=1}^{M} p(\mathbf{x}_i|\omega_m, \boldsymbol{\theta}_m^{(k)})P^{(k)}(\omega_m)}$$

This will use the fact that

$$\sum_{i=1}^{n}\sum_{\forall \mathbf{z}_i} P(\mathbf{z}_i|\mathbf{x}_i) \sum_{m=1}^{M} z_{im} \log(p(\mathbf{x}_i|\omega_m)) = \sum_{m=1}^{M}\left[\sum_{i=1}^{n} P(\omega_m|\mathbf{x}_i) \log(p(\mathbf{x}_i|\omega_m))\right]$$

Maximisation

Now we need to maximise the auxiliary function, $\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)})$. This may be written as

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) &= \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(k)}) \log\left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)})\right) \\
&= \sum_{i=1}^{n}\sum_{\forall \mathbf{z}_i} P(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) \log\left(p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}^{(k+1)})\right) \\
&= \sum_{i=1}^{n}\sum_{\forall \mathbf{z}_i} P(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) \sum_{m=1}^{M} z_{im} \log\left(p(\mathbf{x}_i|\omega_m, \boldsymbol{\theta}_m^{(k+1)})\right) \\
&\quad + \sum_{i=1}^{n}\sum_{\forall \mathbf{z}_i} P(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) \sum_{m=1}^{M} z_{im} \log\left(P^{(k+1)}(\omega_m)\right) \\
&= \sum_{m=1}^{M}\left[\sum_{i=1}^{n} P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) \log\left(p(\mathbf{x}_i|\omega_m, \boldsymbol{\theta}_m^{(k+1)})\right)\right] \\
&\quad + \sum_{m=1}^{M}\left[\sum_{i=1}^{n} P(\omega_m|\mathbf{x}_i, \boldsymbol{\theta}^{(k)}) \log\left(P^{(k+1)}(\omega_m)\right)\right]
\end{aligned}$$

Compare this to the ML estimation of the parameters of a single Gaussian pdf

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(p(\mathbf{x}_i|\boldsymbol{\theta})\right)$$

So, as we saw before, in EM we simply weight each of the observations log-likelihoods according to the hidden variable PMF.

Gaussian Mixture Models Revisited

For Gaussian Mixture Models (or mixtures of Gaussians), the log likelihood for component ω_m (d-dimensional data) is

$$\log\left(p(\mathbf{x};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)\right) = -\frac{1}{2}\left(\log((2\pi)^d|\boldsymbol{\Sigma}_m|) + (\mathbf{x}-\boldsymbol{\mu}_m)'\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right)$$

The auxiliary function may be written as

$$\begin{aligned}
&\mathcal{Q}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^{(k+1)}) \\
&= \sum_{m=1}^{M}\left[\sum_{i=1}^{n} P(\omega_m|\mathbf{x}_i,\boldsymbol{\theta}^{(k)})\left(-\frac{1}{2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_m)'\hat{\boldsymbol{\Sigma}}_m^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_m)\right)\right] \\
&\quad + \sum_{m=1}^{M}\left[\sum_{i=1}^{n} P(\omega_m|\mathbf{x}_i,\boldsymbol{\theta}^{(k)})\left(-\frac{1}{2}\log((2\pi)^d|\hat{\boldsymbol{\Sigma}}_m|)\right)\right] \\
&\quad + \sum_{m=1}^{M}\left[\sum_{i=1}^{n} P(\omega_m|\mathbf{x}_i,\boldsymbol{\theta}^{(k)})\log\left(P^{(k+1)}(\omega_m)\right)\right]
\end{aligned}$$

where $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$ are the mean and covariance matrix of component ω_m at iteration $k+1$.
This yields the re-estimation formulae for the mean and covariance matrix of component ω_j

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{n} P(\omega_j|\mathbf{x}_i,\boldsymbol{\theta}^{(k)})\mathbf{x}_i}{\sum_{i=1}^{n} P(\omega_j|\mathbf{x}_i,\boldsymbol{\theta}^{(k)})}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{i=1}^{n} P(\omega_j|\mathbf{x}_i,\boldsymbol{\theta}^{(k)})(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_j)'}{\sum_{i=1}^{n} P(\omega_j|\mathbf{x}_i,\boldsymbol{\theta}^{(k)})}$$

Simple Continuous E-M Example

Given n noisy measurements $x_1,\ldots,x_n$, with the noise known to be zero mean and unit variance, and that the "true" data is Gaussian distributed with variance σ^2. What is the mean, μ of the true data? From the question we know that

$$x_i = t_i + z, \quad z \sim \mathcal{N}(0,1)$$

t_i is the true data at i. We therefore know that

$$p(x_i|\theta) = \mathcal{N}(x_i;\mu,\sigma^2+1)$$

Could directly find the ML estimate for the parameters, but what about using EM? Let the "new" estimate of the parameters be $\hat{\theta}$ and the old estimate θ. Let the hidden variable be the noise value for a particular observation, z_i. So

$$p(x_i|z_i,\theta) = \mathcal{N}(x_i;\mu+z_i,\sigma^2)$$

We first need to compute the posterior $p(z_i|x_i,\theta)$

$$\begin{aligned}
p(z_i|x_i,\theta) &= \frac{p(x_i|z_i,\theta)p(z_i)}{p(x_i|\theta)} \\
&= \mathcal{N}\left(z_i;\frac{(x_i-\mu)}{(1+\sigma^2)},\frac{\sigma^2}{(1+\sigma^2)}\right)
\end{aligned}$$

So writing down the auxiliary function

$$\begin{aligned}
\mathcal{Q}(\theta,\hat{\theta}) &= \sum_{i=1}^{n}\int\left(p(z_i|x_i,\theta)\log(p(x_i,z_i|\hat{\theta})\right)dz_i \\
&= \sum_{i=1}^{n}\int\left(p(z_i|x_i,\theta)\log(p(x_i|z_i,\hat{\theta})\right)dz_i \\
&\quad + \sum_{i=1}^{n}\int\left(p(z_i|x_i,\theta)\log(p(z_i)\right)dz_i
\end{aligned}$$

The second term is not dependent on the new model parameters, the distribution of z_i is known. This leaves the first term. From the previous definitions

$$\begin{aligned}\tilde{\mathcal{Q}}(\theta,\hat{\theta}) &= \sum_{i=1}^{n}\int p(z_i|x_i,\theta)\log(p(x_i|z_i,\hat{\theta}))dz_i \\ &= \sum_{i=1}^{n}\int p(z_i|x_i,\theta)\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)-\frac{(x_i-z_i-\hat{\mu})^2}{2\sigma^2}\right]dz_i \\ &= \sum_{i=1}^{n}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)-\frac{(x_i-\hat{\mu})^2-2(x_i-\hat{\mu})\mathcal{E}\{z_i|\theta,x_i\}+\mathcal{E}\{z_i^2|\theta,x_i\}}{2\sigma^2}\right]\end{aligned}$$

We know that

$$\begin{aligned}\mathcal{E}\{z_i|\theta,x_i\} &= \frac{(x_i-\mu)}{(1+\sigma^2)} \\ \mathcal{E}\{z_i^2|\theta,x_i\} &= \frac{\sigma^2}{(1+\sigma^2)}+\left(\frac{(x_i-\mu)}{(1+\sigma^2)}\right)^2\end{aligned}$$

Differentiating with respect to $\hat{\mu}$ gives

$$\frac{\partial\tilde{\mathcal{Q}}(\theta,\hat{\theta})}{\partial\hat{\mu}}=\sum_{i=1}^{n}\frac{1}{\sigma^2}\left(x_i-\hat{\mu}-\mathcal{E}\{z_i|\theta,x_i\}\right)$$

so

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\frac{(x_i-\mu)}{(1+\sigma^2)}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{(\sigma^2x_i+\mu)}{(1+\sigma^2)}$$

In this case the standard ML estimation for this problem is trivial, but the above should illustrate the use of EM.

Factor Analysis

E-M can also be used to generate the parameters of a factor analysis model. In factor analysis, d-dimensional data, $\mathbf{x}$, is modelled using a p-dimensional vector of factors $\mathbf{z}$ and observations $\mathbf{x}$ are generated by

$$\mathbf{x}=\mathbf{A}\mathbf{z}+\mathbf{v}$$

where $\mathbf{A}$ is the $d\times p$ factor loading matrix ($d>p$). The factors are Gaussian distributed with zero mean and identity covariance matrix. $\mathbf{v}$ has a diagonal covariance matrix, $\mathbf{\Sigma}$.The data is zero mean.

According to this model, $\mathbf{x}$ is Gaussian distributed with zero mean and covariance $\mathbf{A}\mathbf{A}^{'}+\mathbf{\Sigma}$, and the goal is to find $\mathbf{A}$ and $\mathbf{\Sigma}$ using E-M, that best models the covariance structure of $\mathbf{x}$.

The hidden variables for factor analysis are the values of $\mathbf{z}$ associated with each training sample.

Use of E-M involves setting up the auxiliary function and the solution requires finding the expectations $\mathcal{E}(\mathbf{z}|\mathbf{x}_i)$ and $\mathcal{E}(\mathbf{z}\mathbf{z}^{'}|\mathbf{x}_i)$.