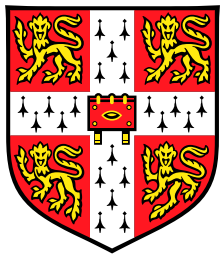
Engineering Part IIB: Module 4F10
Statistical Pattern Processing
Lecture 4: Expectation-Maximisation

Phil Woodland: pcw@eng.cam.ac.uk

Michaelmas 2012



Cambridge University Engineering Department

Introduction

Previously looked at GMMs & found an iterative procedure could be used to estimate parameters.

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

In this lecture we will examine the

- mathematical basis of EM for Gaussian mixtures
- auxiliary functions
- application of EM to continuous and discrete latent variables

The training data (for one class) will be $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

Expectation-Maximisation

- In EM, rather than directly optimising the log likelihood $\mathcal{L}(\boldsymbol{\theta})$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{x}_i|\boldsymbol{\theta})$$

  use an iterative approach so that on iteration k

$$\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) \geq 0$$

  where $\mathcal{Q}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$ is a lower-bound on $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)})$

- If $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ can be simply optimised wrt $\boldsymbol{\theta}$, then iterate until convergence
  - Since $\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$ depends only on iteration k parameters maximising the auxiliary function maximises the log likelihood lower bound.
- Need to select an appropriate form for auxiliary function.

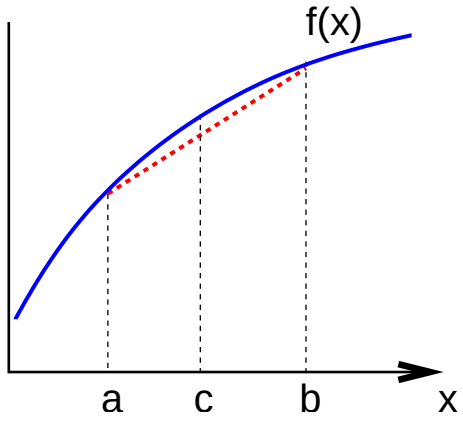
Jensen's Inequality

Jensen's inequality can be used to derive EM updates for GMMs. It states

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

where $f()$ is any concave function and

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



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

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

Lower Bound for Mixture Models

Consider the change in log likelihood function

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

Expand mixture model & multiply numerator/denominator by $P(\omega_m|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})$

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

Treating numerator $P(\omega_m|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)})$ as λ_m in Jensen's inequality ($\log()$ concave) gives

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

Definition of Auxiliary Function

- Recall the desired change

$$\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) \geq 0$$

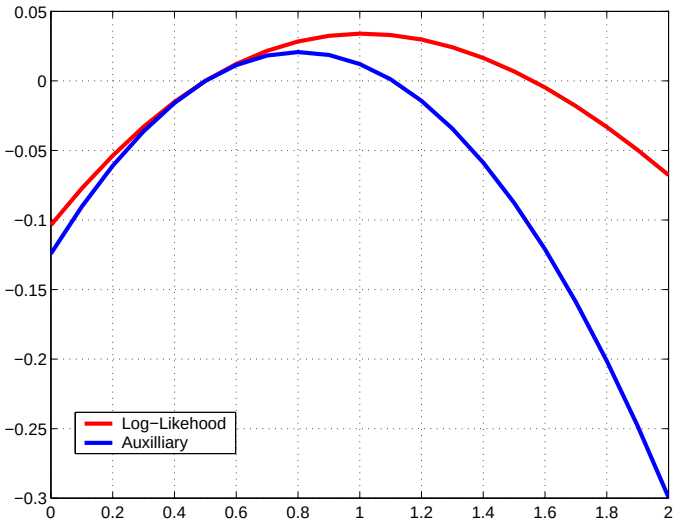
  Compare with the definition from Jensen's inequality

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

- To ensure that the log likelihood change is ≥ 0 require

$$\mathcal{Q}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) \geq 0$$

- Auxiliary function value can be maximised & unique global maximum found (differentiate and equate to zero, use Lagrange multipliers for component prior updates). This leads to the earlier update formulae for mixture models.



EM Example

Data generated from the following GMM:

$$x \sim 0.4 \times \mathcal{N}(1,1) + 0.6 \times \mathcal{N}(-1,1)$$

Initial estimate of the model parameters is

$$x^{(0)} \sim 0.4 \times \mathcal{N}(0.5,1) + 0.6 \times \mathcal{N}(-1,1)$$

Plot shows the variation of the log-likelihood difference and auxiliary function difference as the estimate of the mean of component 1 changes

- auxiliary function difference always a lower-bound
- peak of auxiliary function about 0.8
- peak of log-likelihood function 1.0
- gradient at current value (0.5) same for both
- need to iterate to get to max of log likelihood function

Expectation Maximisation

EM is a general iterative optimisation technique. Require $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \geq 0$.

Consider the situation where the likelihood of the observations can be expressed in terms of a set of latent variables $\mathbf{Z}$. Thus

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

For the GMM case the latent variable is the component ω_k.

The form of auxiliary function for this general case is

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

The continuous latent variable case version is

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

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 $\{\mathbf{Z}, \boldsymbol{X}\}$ is sometimes referred to as the complete dataset. It consists of the observed data $\boldsymbol{X}$ (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 $\{\mathbf{Z}, \boldsymbol{X}\}$ 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 function increases is small (a tight bound).

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

EM can be seen to have two stages (on each iteration):

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

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

   where the expectation is over the distribution of the latent variables, $\mathbf{Z}$, given the current model parameters.

2. Maximisation: maximise $\mathcal{Q}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)})$, 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

Discrete hidden variable to indicate which component generated an observation:

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

For a point $\boldsymbol{x}_i$ generated by component ω_j can write

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

As all data points are independent then hidden variables associated with data points also independent. The auxiliary function now becomes

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

Gaussian Mixture Models Revisited

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

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

The auxiliary function may be written as

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

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

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

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

Simple Continuous Latent Variable

Given n noisy measurements $x_1, \ldots, x_n$, with noise 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? If t_i is the true data at i, then

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

As the noise is independent of the observation, and the sum of two Gaussian distributed variables is Gaussian distributed, we therefore know that

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

Could directly find ML estimate for parameters, but what about using EM?

First the choice of the latent variable need to be made. In this case the value of the noise at each time instance can be used. Let the hidden variable be the noise value for a particular observation, z_i. So

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

Auxiliary Function

$$\mathcal{Q}(\theta^{(k+1)}, \theta^{(k)}) = \sum_{i=1}^{n} \int p(z_i|x_i, \theta^{(k)}) \log\left(p(x_i, z_i|\theta^{(k+1)})\right) dz_i$$

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

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

So writing down the auxiliary function

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

The 2nd term doesn't depend on new parameters, distribution of z_i is known.

Maximisation

Only the first term is needed

$$\tilde{\mathcal{Q}}(\theta^{(k+1)}, \theta^{(k)}) = \sum_{i=1}^{n} \int p(z_i|x_i, \theta^{(k)}) \log(p(x_i|z_i, \theta^{(k+1)}))dz_i$$

$$= \sum_{i=1}^{n} \int p(z_i|x_i, \theta^{(k)}) \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i - z_i - \mu^{(k+1)})^2}{2\sigma^2} \right] dz_i$$

$$= \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i - \mu^{(k+1)})^2 - 2(x_i - \mu^{(k+1)})\mathcal{E}\{z_i|\theta^{(k)}, x_i\} + \mathcal{E}\{z_i^2|\theta^{(k)}, x_i\}}{2\sigma^2} \right]$$

We know that

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

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

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

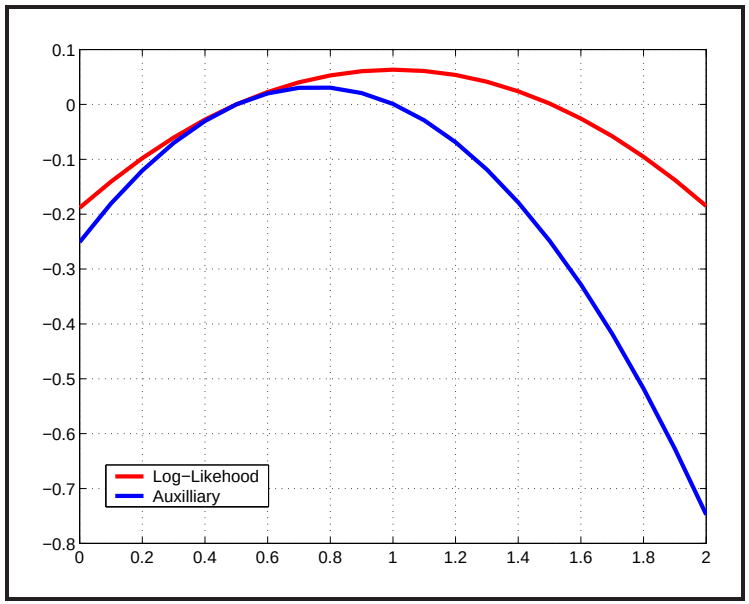
so

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

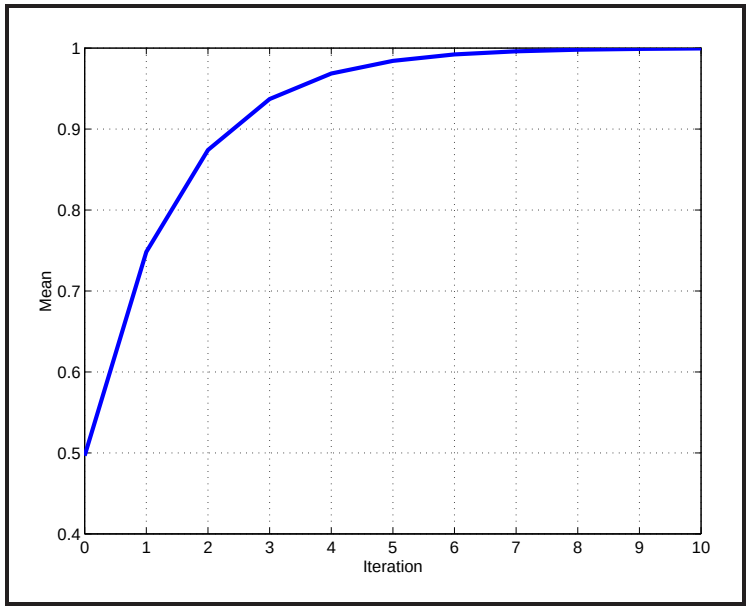
In this case the standard ML estimation for this problem is trivial, but the above should illustrate the use of EM.

Optimisation Example

True data has a mean of 1 and a variance of 1. Initial estimate of the mean is 0.5



The above diagram shows the difference in the log-likelihood and auxiliary function at this first iteration.



The above diagram shows the change in estimate of the mean.

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 $\mathcal{E}(\mathbf{z}\mathbf{z}^{'}|\mathbf{x}_i)$.