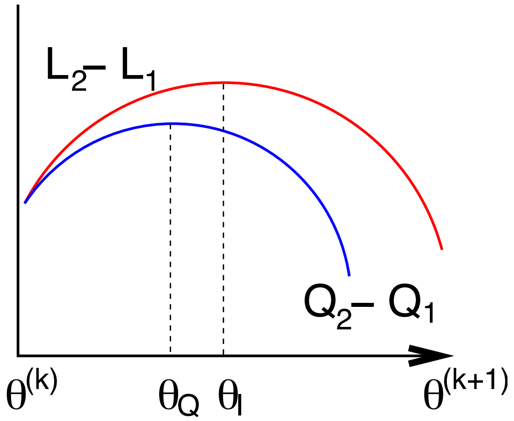
University of Cambridge
Engineering Part IIB

Module 4F10: Statistical Pattern Processing

Handout 4: Expectation-Maximisation



Mark Gales
mjfg@eng.cam.ac.uk
Michaelmas 2013

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.

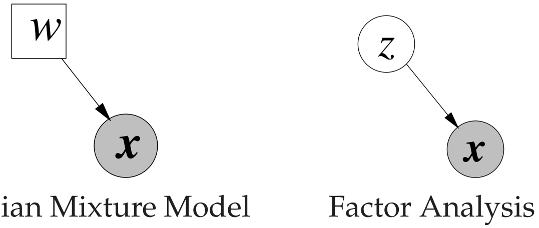
In this lecture we will examine the

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

The training data (for one class) will be

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$$

GMMs and FA



- Gaussian mixture models: these have the form

$$p(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})$$

  this may be though as selecting a component from the PMF (formed of the component priors). Given the selected component w the observation is generated from the specified Gaussian component.

- Factor analysis: this is best described in terms of a generative model

$$\begin{aligned} \boldsymbol{z} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}) \\ \boldsymbol{x} &= \mathbf{C}\boldsymbol{z} + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{(w)}) \\ p(\boldsymbol{x}) &= \int p(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \end{aligned}$$

  Here $\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ means distributed according to a multivariate Gaussian distribution of mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The overall covariance matrix is given by

$$\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}' + \boldsymbol{\Sigma}^{(w)}$$

Jensen's Inequality

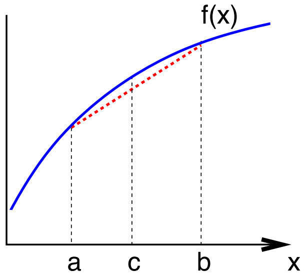
One useful inequality, commonly used in the derivation of the update formulae for mixture models, is Jensen's inequality. It states that

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

where $f()$ is any concave function and

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

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

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

Deriving the EM 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 $p(\boldsymbol{x}|\boldsymbol{\theta}^{(k)})$ to $p(\boldsymbol{x}|\boldsymbol{\theta}^{(k+1)})$. The increase in log likelihood is

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

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

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

Since $\log()$ is strictly concave we can use Jensen's Inequality which states that if $\lambda_m \geq 0$ 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 $P(\omega_m|\boldsymbol{x}_i,\boldsymbol{\theta}^{(k)})$ as λ_m gives

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

Auxiliary Functions

This inequality can be written as

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

where

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

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

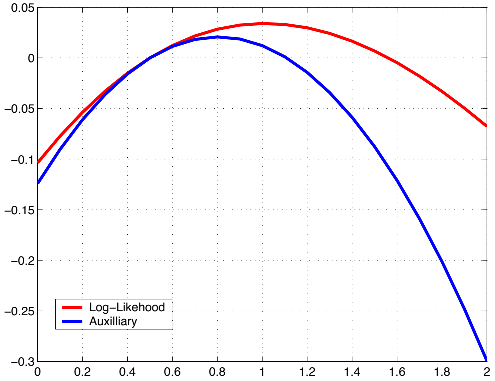
EM Example

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

- 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

How tight is the lower bound?

Kullback-Leibler Divergence

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

$$\begin{aligned} \mathcal{KL}(p(\boldsymbol{x})||q(\boldsymbol{x})) &= \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x} \\ &= -\int p(\boldsymbol{x}) \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) d\boldsymbol{x} \end{aligned}$$

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

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

This gives the following inequality

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

Similarly for the discrete version

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

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

$$\mathcal{KL}(p(\boldsymbol{x})||q(\boldsymbol{x})) \geq 0$$

KL Divergence for Gaussians

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

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

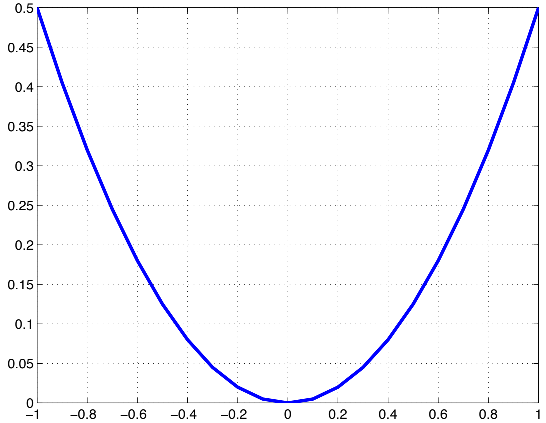
Then the KL divergence between the two is given by

$$\begin{aligned} \mathcal{KL}(p(\boldsymbol{x})||q(\boldsymbol{x})) &= \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1 - \boldsymbol{I}) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right. \\ &\quad \left. + \log\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right)\right) \end{aligned}$$

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

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

Alternatively we aim to ensure that

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

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

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

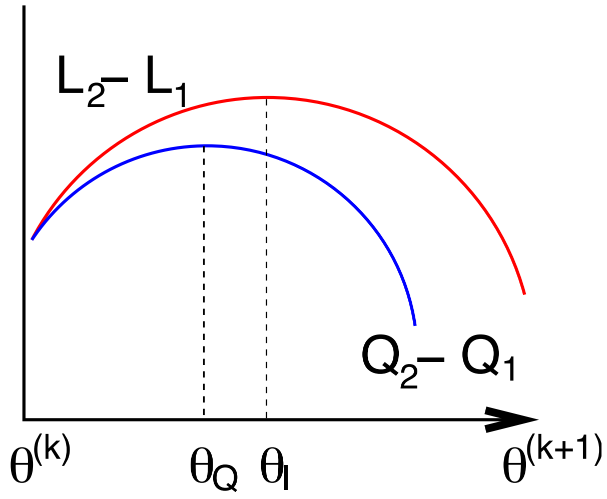
The continuous latent variable case version is

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

With these functions it is possible to show (notes at back)

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

Optimisation Example



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

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

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.

How Tight is the Bound?

An interesting question is how tight is the bound -

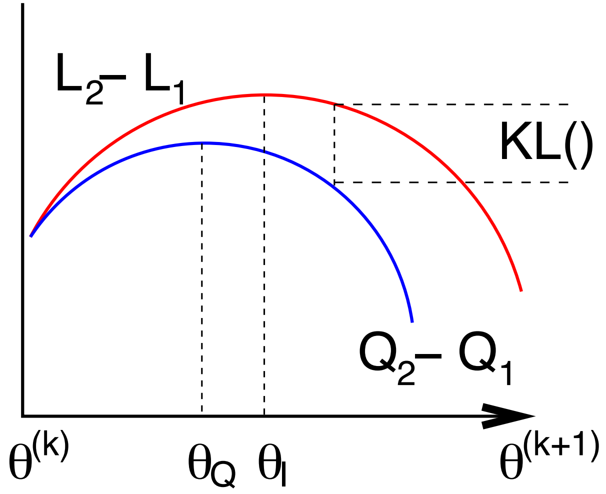
$$\left(\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)})\right) - \left(\mathcal{Q}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^{(k+1)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^{(k)})\right)$$

The tighter the bound the better!

It can be shown (see end of slides) that

$$\left(\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)})\right) = \left(\mathcal{Q}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^{(k+1)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)},\boldsymbol{\theta}^{(k)})\right) \\ + \mathcal{KL}\left(P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)})||P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k+1)})\right)$$

The difference (how tight the bound) is the KL-divergence.



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). The difference is given by
$$\mathcal{KL}\left(P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)})||P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k+1)})\right)$$
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
$$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)}) \log\left(p(\boldsymbol{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)})\right)$$
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, $P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)})$. Given this distribution calculate the expected value of log-likelihood of the complete dataset in terms of the new model parameters, $\boldsymbol{\theta}^{(k+1)}$,
$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) &= \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)}) \log\left(p(\boldsymbol{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)})\right) \\ &= \mathcal{E}\left\{\log\left(p(\boldsymbol{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)})\right)|\boldsymbol{X}, \boldsymbol{\theta}^{(k)}\right\}\end{aligned}$$
where the expectation is over the distribution of the latent variables, $\mathbf{Z}$, given the current model parameters. 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

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:

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

If we look at a single point $\boldsymbol{x}_i$ and know that it was generated by component ω_j, then we 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 the data points are independent then the hidden variables associated with the data points will also be independent of one another. The auxiliary function now becomes

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

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

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

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 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.

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 the ML estimate for the 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

The form of the auxiliary function is required

$$\begin{aligned}\mathcal{Q}(\theta^{(k)}, \theta^{(k+1)}) &= \int p(\mathbf{Z}|\boldsymbol{X}, \theta^{(k)}) \log\left(p(\boldsymbol{X}, \mathbf{Z}|\theta^{(k+1)})\right) d\mathbf{Z} \\ &= \sum_{i=1}^{n} \int p(z_i|x_i, \theta^{(k)}) \log\left(p(x_i, z_i|\theta^{(k+1)})\right) dz_i\end{aligned}$$

where the new estimate of the parameters is $\theta^{(k+1)}$ and the old estimate $\theta^{(k)}$.

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

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

So writing down the auxiliary function

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

The second term is not dependent on the new model parameters, the distribution of z_i is known.

Maximisation

Only the first term is needed

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

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

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

We know that

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

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

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

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

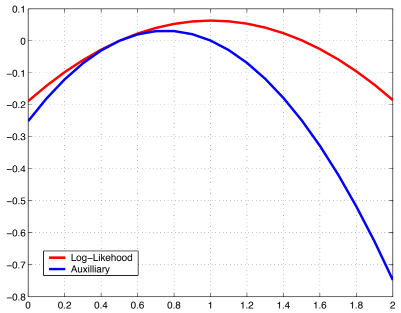
so

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

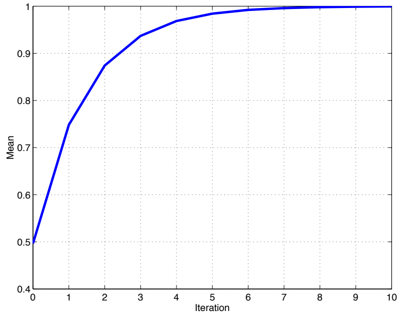
In this case the standard ML estimation for this problem is trivial, but the above should illustrate the use of EM.

Optimisation

Take the example where the true data has a mean of 1 and a variance of 1. The 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.

Variational Approaches

So far assumed that it is possible to compute $P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)})$.

sometimes intractable!

An alternative way of expressing the likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = \int P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta})\log\left(\frac{p(\mathbf{Z},\boldsymbol{X}|\boldsymbol{\theta})}{P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta})}\right)d\mathbf{Z}$$

Simple to show using KL-divergence that

$$\mathcal{L}(\boldsymbol{\theta}) \geq \int P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)})\log\left(\frac{p(\mathbf{Z},\boldsymbol{X}|\boldsymbol{\theta})}{P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)})}\right)d\mathbf{Z}$$

Standard auxiliary function.

Introduce a general function of the latent variable $q(\mathbf{Z},\boldsymbol{\theta}^{(k)})$.
Again from KL-divergence

$$\mathcal{L}(\boldsymbol{\theta}) \geq \int q(\mathbf{Z},\boldsymbol{\theta}^{(k)})\log\left(\frac{p(\mathbf{Z},\boldsymbol{X}|\boldsymbol{\theta})}{q(\mathbf{Z},\boldsymbol{\theta}^{(k)})}\right)d\mathbf{Z}$$

Variational auxiliary function

- no need to compute $P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)})$
- but not guaranteed to increase likelihood

EM Proof/Bound Tightness (slide 9)

Interested in ensuring that

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

From the definition of a PMF we can write

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

since

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

From the definition of conditional probability

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

so

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

and similarly for the second term.

EM Proof (cont)

We can now write

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

From the discussion about the KL-divergence

$$\begin{aligned}&\mathcal{KL}(P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)})||P(\mathbf{Z}|\boldsymbol{x},\boldsymbol{\theta}^{(k+1)})) = \\ &\sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)}) \log\left(\frac{P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k)})}{P(\mathbf{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(k+1)})}\right) \\ &\geq 0\end{aligned}$$

So it follows that

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

where the difference between the left and right-hand sides is the KL divergence given above.

EM Proof (cont)

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

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

then

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

It is common to define the auxiliary function as

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

and for the continuous version

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

Thus the auxiliary function is the expected value of the log likelihood of the joint distribution of **Z** and **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

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

Mixture Model Expectation (slide 14)

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

$$\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}, \qquad \mathbf{z}_i = \left[\begin{array}{c} z_{i1} \\ \vdots \\ z_{iM} \end{array}\right]$$

As the observations are independent

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

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

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

This will use the fact that

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

Mixture Model Maximisation

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

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

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

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

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

Maximisation Details

Here is a more detailed derivation for the previous slide. Using the fact that the observations and latent variables for each training example are independent of one another, so

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

Then note that summing over all $\mathbf{Z}$ that have a specific value for $\mathbf{z}_i$

$$\begin{aligned}
\sum_{\mathbf{Z}:\mathbf{z}_i \in \mathbf{Z}} P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)}) f(\mathbf{z}_i) &= \sum_{\mathbf{Z}:\mathbf{z}_i \in \mathbf{Z}} f(\mathbf{z}_i) \prod_{j=1}^{n} P(\mathbf{z}_j|\boldsymbol{x}_j, \boldsymbol{\theta}^{(k)}) \\
&= P(\mathbf{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) f(\mathbf{z}_i) \prod_{j \neq i}^{n} \left(\sum_{\forall \mathbf{z}_j} P(\mathbf{z}_j|\boldsymbol{x}_j, \boldsymbol{\theta}^{(k)}) \right) \\
&= P(\mathbf{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) f(\mathbf{z}_i)
\end{aligned}$$

The following set of equalities can be written

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) &= \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\boldsymbol{X}, \mathbf{Z}|\boldsymbol{\theta}^{(k+1)}) \right) \\
&= \sum_{\forall \mathbf{Z}} P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)}) \sum_{i=1}^{n} \log \left(p(\boldsymbol{x}_i, \mathbf{z}_i|\boldsymbol{\theta}^{(k+1)}) \right) \\
&= \sum_{i=1}^{n} \sum_{\forall \mathbf{z}_i} \sum_{\mathbf{Z}:\mathbf{z}_i \in \mathbf{Z}} P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(k)}) \log \left(p(\boldsymbol{x}_i, \mathbf{z}_i|\boldsymbol{\theta}^{(k+1)}) \right) \\
&= \sum_{i=1}^{n} \sum_{\forall \mathbf{z}_i} P(\mathbf{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) \log \left(p(\boldsymbol{x}_i, \mathbf{z}_i|\boldsymbol{\theta}^{(k+1)}) \right)
\end{aligned}$$