Advanced Data Analytics
Lecture week 3: Mixture Models

Ian T. Nabney

University of BRISTOL

Overview

- Understand Gaussian mixture models in greater generality
- Able to explain how the EM algorithm works
- Understand a high-level view of the variational analysis of the EM algorithm from multiple points of view

Based on Sections 9.2–9.4 of Bishop. A reminder that the PDF can be downloaded from `https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/`

University of BRISTOL

Mixtures of Gaussians

- You have already seen the Gaussian mixture model (GMM) and the EM training algorithm briefly in week 15 of the unit Introduction to AI.
- We are going to cover this in more detail here for a number of reasons:
  - There is a greater variety of covariance structures than the simple spherical covariance you have seen alread.
  - We will formulate the GMM in terms of latent variables which will underpin many other models we consider.
  - You need to understand the EM algorithm to greater depth so that we can explain the dimensionality reduction technique of the Generative Topographic Mapping.
  - This is also the first step towards understanding variational inference that will be one of the ways in which we can apply Bayesian modelling in practice.

Latent variables

- The Gaussian mixture distribution can be written as a linear superposition of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{1}$$

- We introduce a K-dimensional binary random variable $\mathbf{z}$ with a 1-of-K representation and define the joint distribution $p(\mathbf{x}, \mathbf{z})$ in terms of a marginal distribution $p(\mathbf{z})$ and a conditional distribution $p(\mathbf{x}|\mathbf{z})$.
- The marginal distribution over $\mathbf{z}$ is specified in terms of the mixing coefficients π_k, such that $p(z_k = 1) = \pi_k$ where the parameters $\{\pi_k\}$ must satisfy

$$0 \leqslant \pi_k \leqslant 1 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1 \tag{2}$$

  in order to be valid probabilities.
- Because $\mathbf{z}$ uses a 1-of-K representation, we can also write this distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{3}$$

University of BRISTOL

Marginal and joint distributions

- Similarly, the conditional distribution of $\mathbf{x}$ given a particular value for $\mathbf{z}$ is a Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

  which can also be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}. \tag{4}$$

- The joint distribution is given by $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$, and the marginal distribution of $\mathbf{x}$ is obtained by summing the joint distribution over all possible states of $\mathbf{z}$ to give

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5}$$

- If we have several observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, then, because we have represented the marginal distribution in the form $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$, it follows that for every observed data point $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$.

Responsibilities

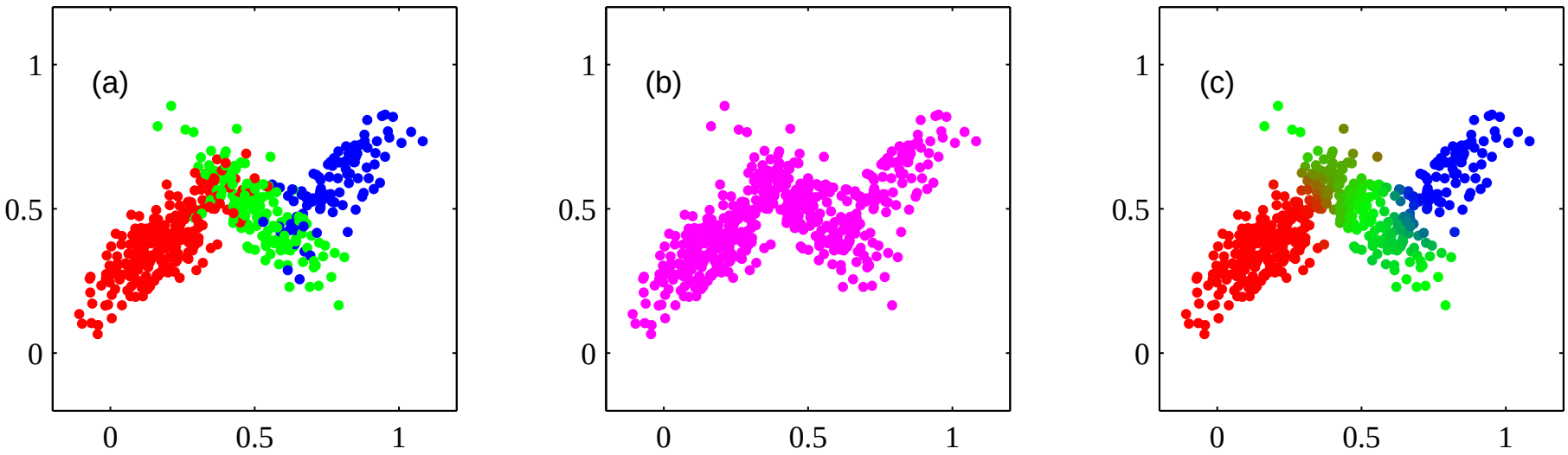
- Another important quantity is the conditional probability of $\mathbf{z}$ given $\mathbf{x}$. We shall use $\gamma(z_k)$ to denote $p(z_k = 1|\mathbf{x})$, whose value can be found using Bayes' theorem

$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \end{aligned} \tag{6}$$

- We shall view π_k as the prior probability of $z_k = 1$, and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed $\mathbf{x}$. $\gamma(z_k)$ can also be viewed as the responsibility that component k takes for 'explaining' the observation $\mathbf{x}$.

University of BRISTOL

Sampling from the distributions



Example of 500 points drawn from the mixture of 3 Gaussians. (a) Samples from the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ in which the three states of $\mathbf{z}$, corresponding to the three components of the mixture, are depicted in red, green, and blue, and (b) the corresponding samples from the marginal distribution $p(\mathbf{x})$, which is obtained by simply ignoring the values of $\mathbf{z}$ and just plotting the $\mathbf{x}$ values. (c) The same samples in which the colours represent the value of the responsibilities $\gamma(z_{nk})$ associated with data point $\mathbf{x}_n$, obtained by plotting the corresponding point using proportions of red, blue, and green ink given by $\gamma(z_{nk})$ for $k = 1, 2, 3$, respectively.

Maximising likelihood

- The log of the likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{7}$$

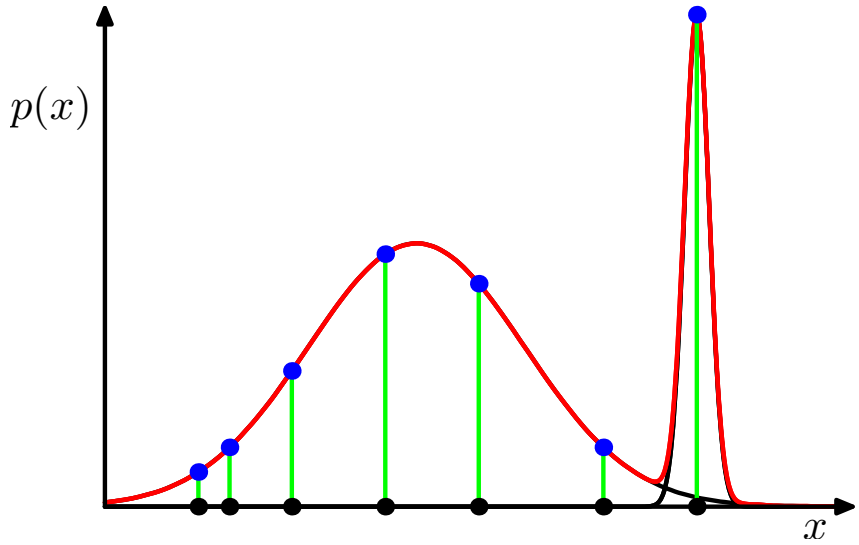
- There is a significant problem associated with the maximum likelihood framework applied to Gaussian mixture models, due to the presence of singularities.
- For simplicity, considerthe case of spherical covariance matrices given by $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$. Suppose that the jth component of the mixture model has its mean $\boldsymbol{\mu}_j$ exactly equal to one of the data points so that $\boldsymbol{\mu}_j = \mathbf{x}_n$ for some value of n.
- This data point will then contribute a term in the likelihood function of the form

$$\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}. \tag{8}$$

  If we consider the limit $\sigma_j \to 0$, then we see that this term goes to infinity and so the log likelihood function will also go to infinity.

Singularities

- These singularities provide another example of the severe over-fitting that can occur in a maximum likelihood approach. We shall see that this difficulty does not occur if we adopt a Bayesian approach.
- In applying maximum likelihood to GMMs we must avoid finding such pathological solutions and instead seek local maxima of the likelihood function that are well behaved.



We use heuristics, e.g. by detecting when a Gaussian component is collapsing and resetting its mean to a randomly chosen value while also resetting its covariance to some large value, and then continuing with the optimization.

M step: means

- Setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the means $\boldsymbol{\mu}_k$ of the Gaussian components to zero, we obtain

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{9}$$

- Multiplying by $\boldsymbol{\Sigma}_k^{-1}$ (which we assume to be nonsingular) and rearranging we obtain

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{10}$$

where we have defined

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{11}$$

We can interpret N_k as the effective number of points assigned to cluster k.

M step: covariances

- If we set the derivative of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}_k$ to zero, making use of the result for the maximum likelihood solution for the covariance matrix of a single Gaussian, we obtain

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \tag{12}$$

  which has the same form as the corresponding result for a single Gaussian fitted to the data set, but again with each data point weighted by the corresponding posterior probability and with the denominator given by the effective number of points associated with the corresponding component.

M step: mixing coefficients

- Here we must take account of the constraint which requires the mixing coefficients to sum to one. This can be achieved using a Lagrange multiplier and maximizing

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{13}$$

  which gives

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \tag{14}$$

  where again we see the appearance of the responsibilities.

- If we now multiply both sides by π_k and sum over k making use of the constraint, we find $\lambda = -N$. Using this to eliminate λ and rearranging we obtain

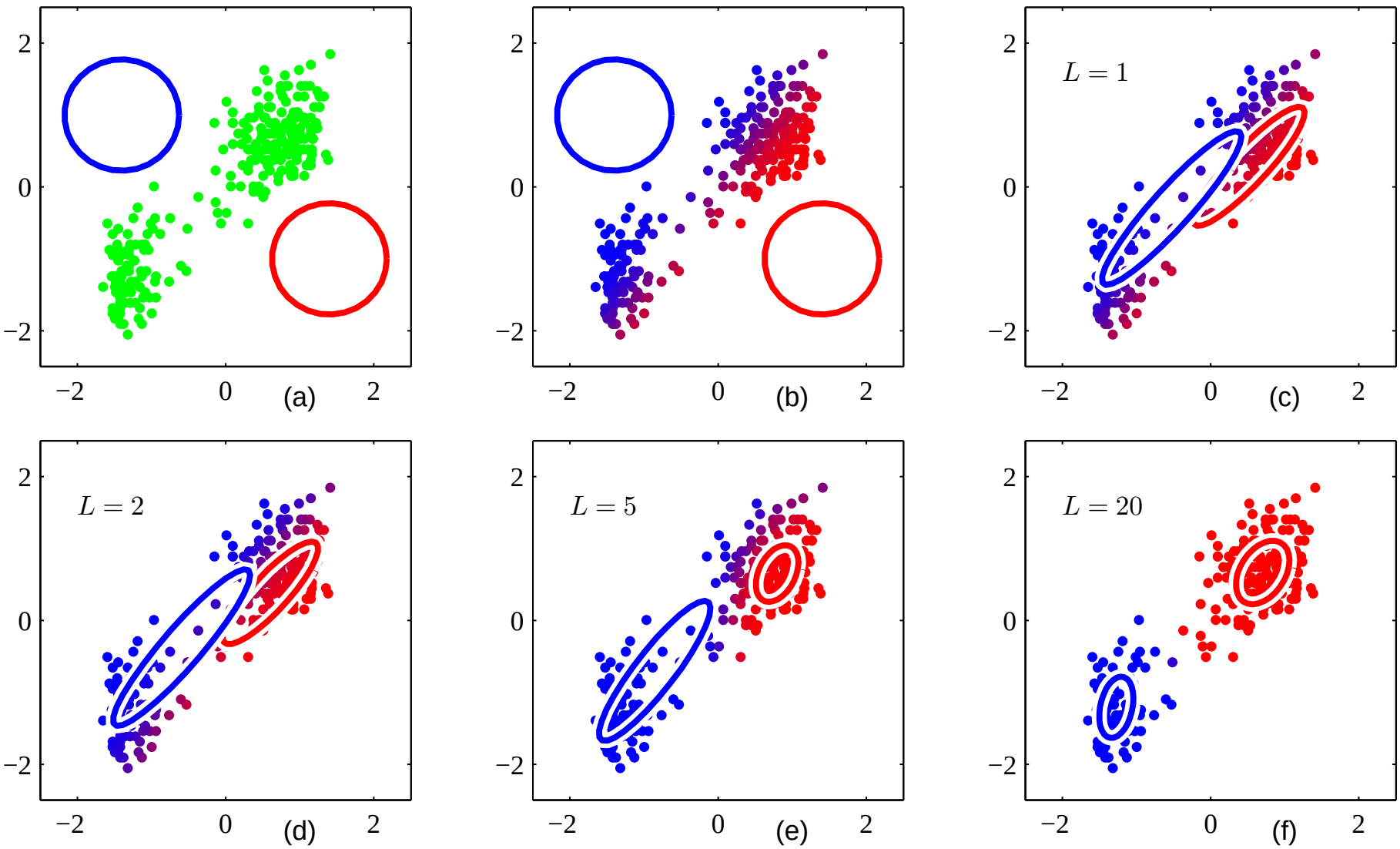
$$\pi_k = \frac{N_k}{N} \tag{15}$$

  so that the mixing coefficient for the k^{th} component is given by the average responsibility which that component takes for explaining the data points.

EM algorithm

- These equations do not constitute a closed-form solution for the parameters of the mixture model because the responsibilities $\gamma(z_{nk})$ depend on those parameters in a complex way.
- We first choose some initial values for the means, covariances, and mixing coefficients.
- Then we alternate between the following two updates that we call the E step and the M step.
  - In the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities, or responsibilities.
  - We then use these probabilities in the maximization step, or M step, to re-estimate the means, covariances, and mixing coefficients. Note that in so doing we first evaluate the new means and then use these new values to find the covariances.
- The algorithm is deemed to have converged when the change in the log likelihood function, or in the parameters, falls below some threshold.

EM algorithm: Old Faithful dataset



Algorithm implementation

- The EM algorithm takes many more iterations to reach (approximate) convergence than the K-means algorithm, and each cycle requires significantly more computation.
- It is therefore common to run the K-means algorithm in order to find a suitable initialization for the means of a Gaussian mixture model that is subsequently adapted using EM.
- The covariance matrices can be initialized to the sample covariances of the clusters found by the K-means algorithm (though in practice we usually inflate these a bit), and the mixing coefficients can be set to the fractions of data points assigned to the respective clusters.
- There are generally multiple local maxima of the log likelihood function, and EM is not guaranteed to find the largest of these maxima.

University of BRISTOL

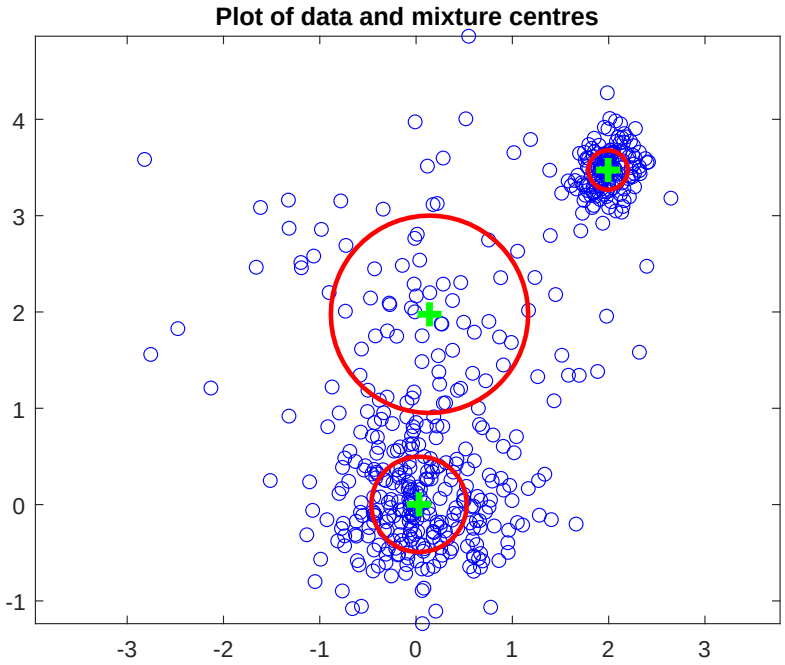
Covariance matrix structure: spherical

- The covariance matrix is a scalar multiple of the identity matrix, $\mathbf{\Sigma}_j = \sigma_j^2 \mathbf{I}$ so that

$$p(\mathbf{x}|j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left\{ -\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2\sigma_j^2} \right\}. \tag{16}$$

- M-step

$$(\sigma_j^{(m+1)})^2 = \frac{1}{d} \frac{\sum_{n=1}^N P^{(m)}(j|\mathbf{x}^n) \|\mathbf{x}^n - \boldsymbol{\mu}_j^{(m+1)}\|^2}{\sum_{n=1}^N P^{(m)}(j|\mathbf{x}^n)}. \tag{17}$$



Covariance matrix structure: diagonal

- The covariance matrix is diagonal

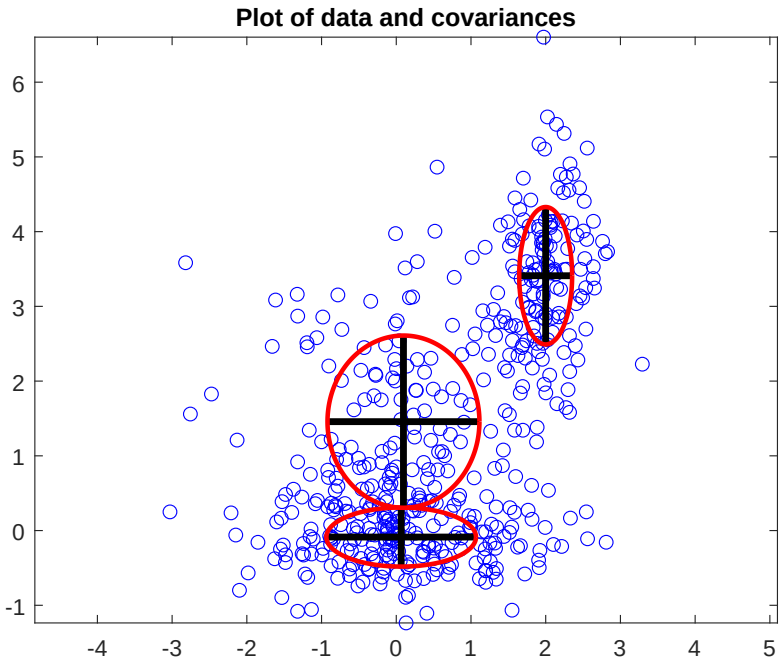
$$\mathbf{\Sigma}_j = \text{diag}(\sigma_{j,1}^2, \ldots, \sigma_{j,d}^2) \qquad (18)$$

and the density function is

$$p(\mathbf{x}|j) = \frac{1}{(2\pi \prod_{i=1}^{d} \sigma_{j,i}^2)^{d/2}} \exp\left\{ -\sum_{i=1}^{d} \frac{(x_i - mu_{j,i})^2}{2\sigma_{j,i}^2} \right\}.$$

- M-step

$$(\sigma_{i,j}^{(m+1)})^2 = \frac{\sum_{n=1}^{N} P^{(m)}(j|\mathbf{x}^n)(x_i^n - \mu_{i,j}^{(m+1)})^2}{\sum_{n=1}^{N} P^{(m)}(j|\mathbf{x}^n)}. \qquad (19)$$



Another view of EM

- We are going to review the EM algorithm focusing more on the latent variable point of view. This will allow us to generalise the algorithm to other models.
- We denote the set of all observed data by $\mathbf{X}$ and the set of all latent variables by $\mathbf{Z}$. The set of all model parameters is denoted by $\boldsymbol{\theta}$, and so the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}. \tag{20}$$

- The analysis applies equally well to continuous latent variables simply by replacing the sum over $\mathbf{Z}$ with an integral.
- The summation over the latent variables appears inside the logarithm. Even if the joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ belongs to the exponential family, the marginal distribution $p(\mathbf{X}|\boldsymbol{\theta})$ typically does not as a result of this summation.

- Now suppose that, for each observation in $\mathbf{X}$, we were told the corresponding value of the latent variable $\mathbf{Z}$. Then $\{\mathbf{X}, \mathbf{Z}\}$ is the complete data set, and the observed data $\mathbf{X}$ is incomplete.
- The likelihood function for the complete data set has the form $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$, and we shall suppose that maximization of this complete-data log likelihood function is straightforward.
- We can only compute the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$. Because we cannot use the complete-data log likelihood, we consider instead its expected value under the posterior distribution of the latent variable, which corresponds to the E step of the EM algorithm.
- In the subsequent M step, we maximize this expectation.

E and M step definitions

- In the E step, we use the current parameter values $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$. We then use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value $\boldsymbol{\theta}$.
- This expectation, denoted $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{21}$$

- In the M step, we determine the revised parameter estimate $\boldsymbol{\theta}^{\text{new}}$ by maximizing this function

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}). \tag{22}$$

Note that in the definition of $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, the logarithm acts directly on the joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$, and so the corresponding M-step maximization will, by supposition, be tractable.

University of BRISTOL

Extending the EM algorithm

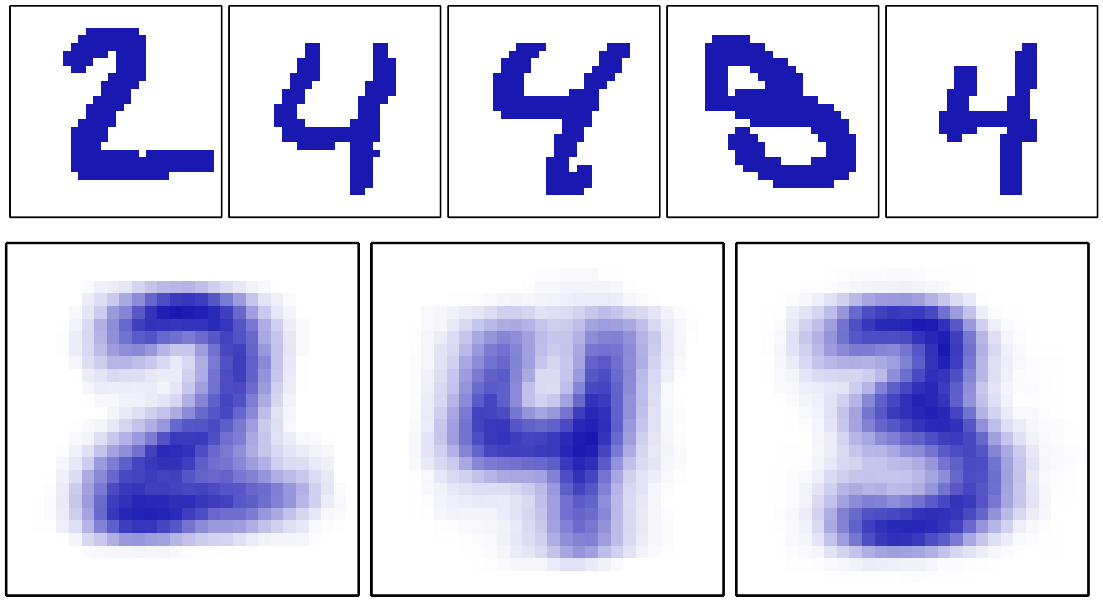
- The EM algorithm can also be used to find MAP (maximum posterior) solutions for models in which a prior $p(\boldsymbol{\theta})$ is defined over the parameters.
- In this case the E step remains the same as in the maximum likelihood case, whereas in the M step the quantity to be maximized is given by $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$.
- Suitable choices for the prior will remove the singularities such as variance collapse.
- EM can also be applied when the unobserved variables correspond to missing values in the data set. The distribution of the observed values is obtained by taking the joint distribution of all the variables and then marginalizing over the missing ones. EM can then be used to maximize the corresponding likelihood function.

Application to Bernoulli mixtures

- Form of Bernoulli (binary)

$$\mu_i^{x_i}(1-\mu_i)^{(1-x_i)}$$

- Also known as latent class analysis.
- Binary image data: lower images show three component means.



The EM algorithm in general

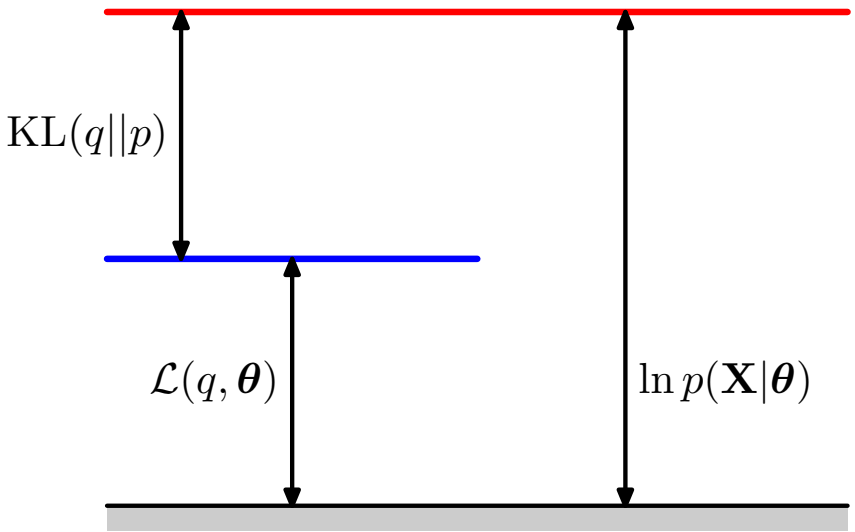
- We introduce a distribution $q(\mathbf{Z})$ defined over the latent variables. Note that the following decomposition holds

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p) \tag{23}$$

where we have defined

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \tag{24}$$

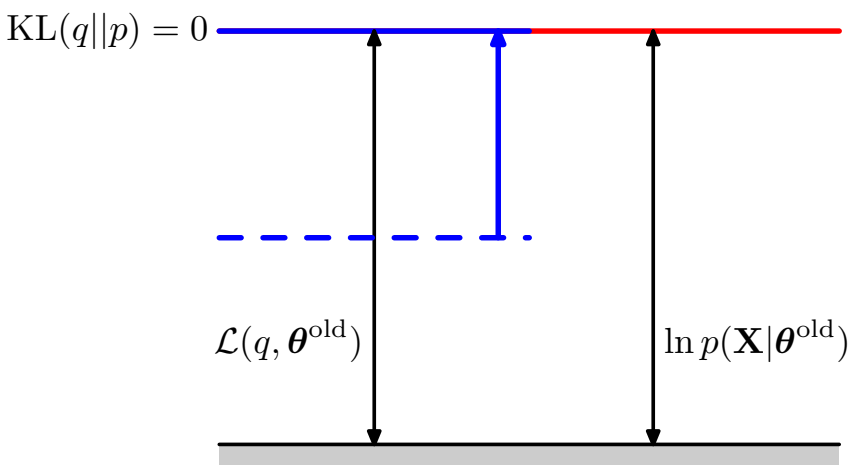
$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}. \tag{25}$$



Because the Kullback-Leibler divergence satisfies $\mathrm{KL}(q\|p) \geqslant 0$, we see that the quantity $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

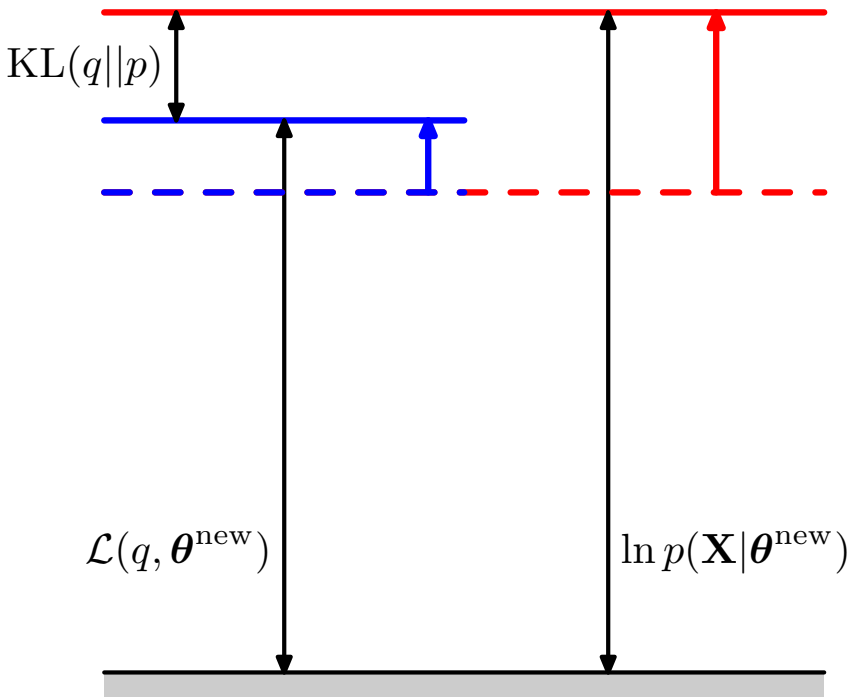
E step definition

- Suppose that the current value of the parameter vector is $\boldsymbol{\theta}^{\text{old}}$. In the E step, the lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ is maximized with respect to $q(\mathbf{Z})$ while holding $\boldsymbol{\theta}^{\text{old}}$ fixed.
- Note that the value of $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$ does not depend on $q(\mathbf{Z})$ and so the largest value of $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ will occur when the Kullback-Leibler divergence vanishes, in other words when $q(\mathbf{Z})$ is equal to the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- In this case, the lower bound will equal the log likelihood



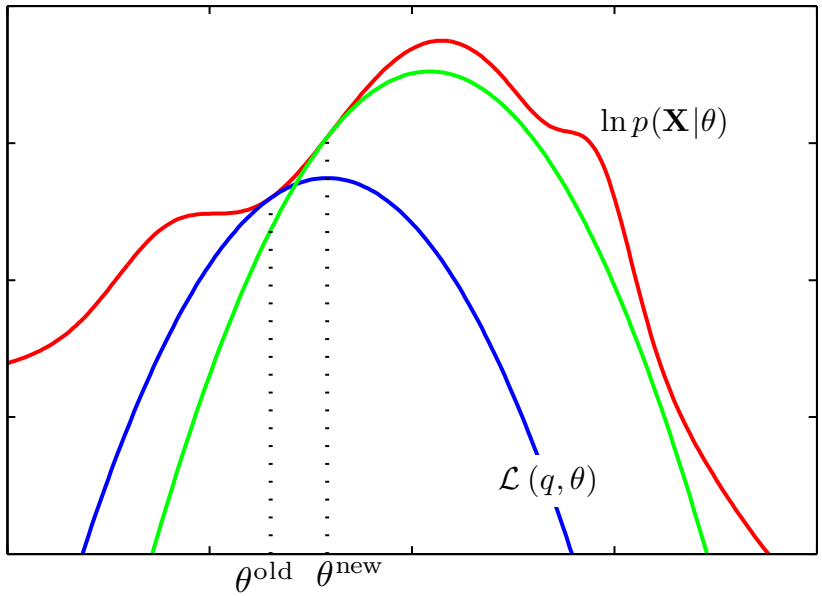
M step definition

- The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$.
- This causes the lower bound $\mathcal{L}$ to increase (unless it is already at a maximum), which will necessarily cause the corresponding log likelihood function to increase.
- Because the distribution q is determined using the old parameter values rather than the new values and is held fixed during the M step, it will not equal the new posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})$, and hence there will be a non-zero KL divergence.
- The increase in the log likelihood function is therefore greater than the increase in the lower bound



EM algorithm in parameter space

- The red curve depicts the (incomplete data) log likelihood function whose value we wish to maximize.
- We start with some initial parameter value $\boldsymbol{\theta}^{\mathrm{old}}$, and in the first E step we evaluate the posterior distribution over latent variables, which gives rise to a lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\mathrm{old})})$ whose value equals the log likelihood at $\boldsymbol{\theta}^{(\mathrm{old})}$, as shown by the blue curve.
- Note that the bound makes a tangential contact with the log likelihood at $\boldsymbol{\theta}^{(\mathrm{old})}$, so that both curves have the same gradient. This bound is a convex function having a unique maximum (for mixture components from the exponential family).
- In the M step, the bound is maximized giving the value $\boldsymbol{\theta}^{(\mathrm{new})}$, which gives a larger value of log likelihood than $\boldsymbol{\theta}^{(\mathrm{old})}$.
- The subsequent E step then constructs a bound that is tangential at $\boldsymbol{\theta}^{(\mathrm{new})}$ as shown by the green curve.



Overview

- Understand Gaussian mixture models in greater generality
- Able to explain how the EM algorithm works
- Understand a high-level view of the variational analysis of the EM algorithm from multiple points of view

Based on Sections 9.2–9.4 of Bishop. A reminder that the PDF can be downloaded from `https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/`

University of BRISTOL