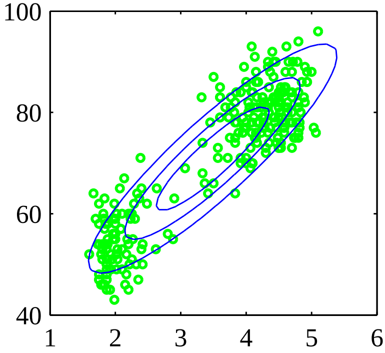
Elements of Statistical Learning 2010/11

Prof Peter Sollich

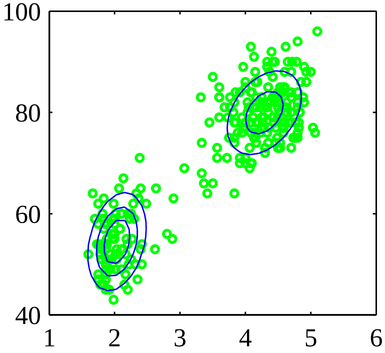
Gaussian Mixtures & Expectation-Maximization (EM)

Mixtures of Gaussians(1)

Old Faithful data set



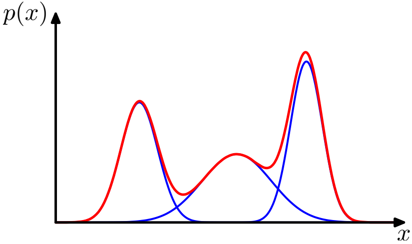
Single Gaussian



Mixture of two Gaussians

Mixtures of Gaussians(2)

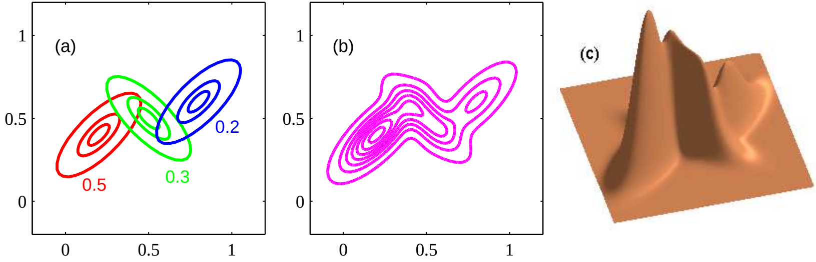
- Get a complex model from a combination of simple models
  $p(\boldsymbol{X}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{X} \,|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$
  e.g., for $K = 3$



- π_k are the Mixing coefficients
- and $\mathcal{N}\left(\boldsymbol{X} \,|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ are the components
- Note that $\forall k : \pi_k \geq 0$ and from normalization $\sum_{k=1}^{K} \pi_k = 1$.

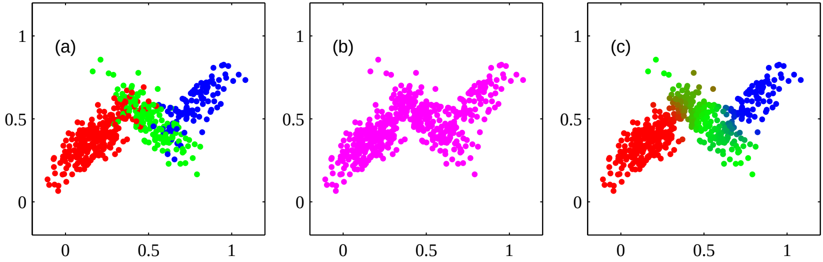
Mixtures of Gaussians(3)

An illustration of a mixture of 3 Gaussians in two-dimensional space



Gaussian Mixtures & Expectation-Maximization (EM)

A demonstration for the *responsibilities*



Example of 500 points drawn from the mixture of 3 Gaussians. (a) Samples from the joint distribution $p(z)p(x|z)$ in which the three states of 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 $p(x)$, which is obtained by simply ignoring the values of z and just plotting the x values. The data set in (a) is said to be complete, whereas that in (b) is incomplete. (c) The same samples in which the colours represent the value of the responsibilities $\gamma(z_{nk})$ associated with data point 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.

Gaussian Mixtures & Expectation-Maximization (EM)

The Maximum Likelihood (M-L) approach

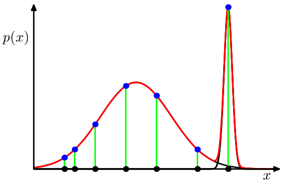
One can estimate the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by maximizing the log-likelihood

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

with respect to $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Difficulties:

- (a) Because of the sum inside the logarithm there is no closed form solution.
- (b) Over-fitting.



- (c) ...

Over-fitting - singularities in the M-L approach

Look at the log-likelihood function

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

and suppose (for simplicity) that for one of the components k

- $\Sigma_k = \sigma_k^2 \boldsymbol{I}$.
- μ_k is exactly equal to one of the data, i.e. $\mu_k = x_n$ for some n.

In that case we obtain $\mathcal{N}\left(x_n | x_n, \sigma_k^2 \boldsymbol{I}\right) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_k}$.

If we consider $\sigma_k \to 0$ then we see that this component goes to infinity, and so the log-likelihood function diverges. In other words, the log-likelihood function is not bounded, which renders the problem of finding its maximum ill-posed!

This cannot happen for a single Gaussian!

Further problems with M-L - identifiability

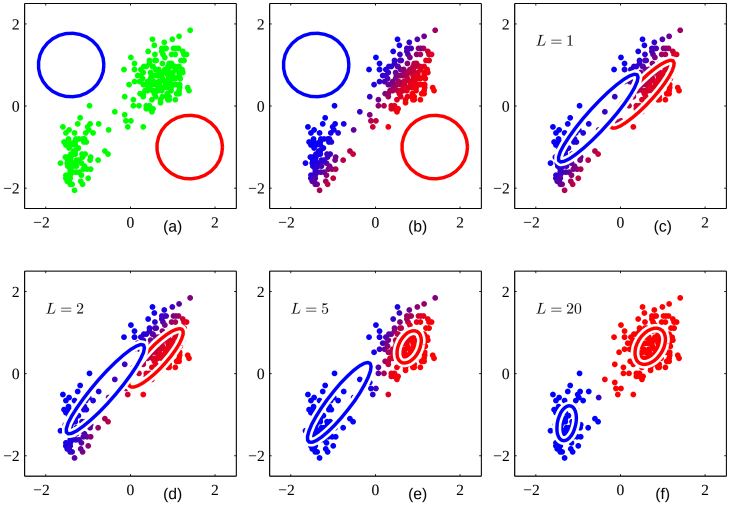
- (a) Because of the sum inside the logarithm there is no closed form solution.
- (b) Over-fitting / Singularities in the log-likelihood function.
- (c) Identifiability: For any given maximum-likelihood solution, a K-component mixture will have $K!$ equivalent solutions corresponding to the $K!$ ways of assigning K sets of parameters to K components.

Gaussian Mixtures & Expectation-Maximization (EM)

Summary of EM for Gaussian Mixtures

- 1. Initialize the means μ_k, covariances Σ_k and mixing coefficients π_k, and evaluate the initial value of the log-likelihood.
- 2. **E step**. Evaluate the responsibilities using the current parameter values $\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$.
- 3. **M step**. Re-estimate the parameters using the current responsibilities
  $\mu_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\boldsymbol{x}_n$,
  $\Sigma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})^T$,
  $\pi_k^{new} = \frac{N_k}{N}$,
  with $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.
- 4. Evaluate the log-likelihood $\ln p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})$ and check for convergence. If not converged, return to step 2.

Illustration of the EM algorithm using the Old Faithful set



Gaussian Mixtures & Expectation-Maximization (EM)

Summary of the General EM Algorithm

- 1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{old}$.
- 2. **E step**. Evaluate $p\left(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}\right)$.
- 3. **M step**. Evaluate $\boldsymbol{\theta}^{new}$ given by

$$\boldsymbol{\theta}^{new} = \max_{\boldsymbol{\theta}} \arg \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{new}),$$

  where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{new}) = \sum_{\boldsymbol{Z}} p\left(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}\right) \ln p\left(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}\right).$$

- 4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let $\boldsymbol{\theta}^{new} \rightarrow \boldsymbol{\theta}^{old}$, and return to step 2.

Gaussian Mixtures & Expectation-Maximization (EM)

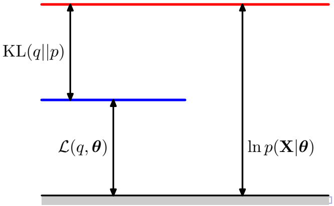
Decomposition of the log-likelihood function

- Within the framework of the latent variables, the likelihood is $p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$
- Given a distribution $q(\boldsymbol{Z})$ over the hidden variable $\boldsymbol{Z}$, one can always decompose:

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

where

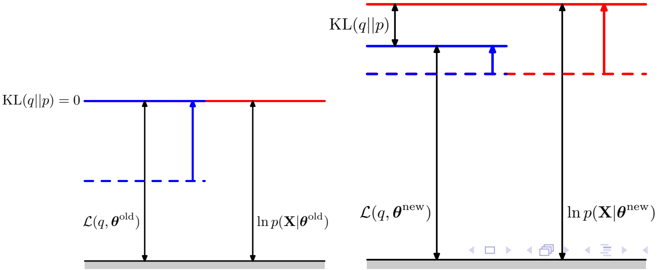
- $\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$
- $KL(q||p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$ (Kullback-Leibler divergence)



Increasing likelihood with the EM iterations

From this point of view the EM algorithm can be seen as a two-stage iterative optimization technique for finding M-L solutions:

- 1. **E step**. The lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ is maximized w.r.t. $q(\boldsymbol{Z})$ while keeping $\boldsymbol{\theta}^{old}$ fixed. This is achieved by fixing $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \quad \Leftrightarrow \quad KL(q||p) = 0$.
- 2. **M step**. $q(\boldsymbol{Z})$ is held fixed., and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized w.r.t. $\boldsymbol{\theta}$, to give $\boldsymbol{\theta}^{new}$. $\mathcal{L}(q, \boldsymbol{\theta})$ will never decrease in this step + typically the new $KL(q||p) > 0$.



Exercises you could try

- 9.7 /8 (Derivation of the EM equations)
- 9.25 (Properties of the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$)