Lecture 14. Expectation Maximisation Algorithm

COMP90051 Statistical Machine Learning

Semester 2, 2017
Lecturer: Andrey Kan

THE UNIVERSITY OF MELBOURNE

This lecture

- Expectation Maximisation (EM) algorithm
  - Introduction in general form
  - Jensen's inequality
  - EM as a coordinate descent approach
- EM applied to Gaussian Mixture Model
  - An iterative approach for parameter estimation
  - K-means as a limiting case of EM for GMM

Expectation Maximisation Algorithm

For a moment, let's put our GMM problem aside. In this section, we'll be talking about generic EM. Then in the next section, we'll apply it to the GMM

Motivation of EM

- Consider a parametric probabilistic model $p(\boldsymbol{X}|\boldsymbol{\theta})$, where $\boldsymbol{X}$ denotes data and $\boldsymbol{\theta}$ denotes a vector of parameters
- According to MLE, we need to maximise $p(\boldsymbol{X}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$
  * equivalently maximise $\log p(\boldsymbol{X}|\boldsymbol{\theta})$
- There can be a couple of issues with this task

1. Sometimes we don't observe some of the variables needed to compute the log likelihood
   * Example: GMM cluster membership is not known in advance
2. Sometimes the form of the log likelihood is inconvenient to work with
   * Example: taking a derivative of GMM log likelihood results in a cumbersome equation

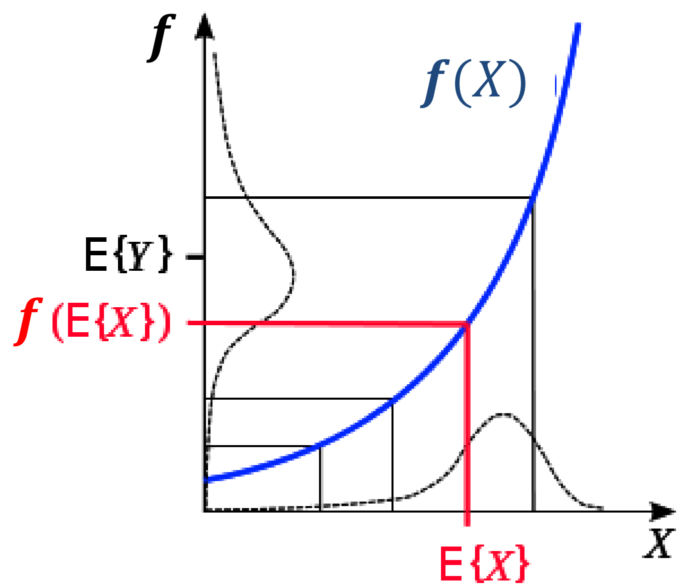
art: Ebrahim at Wikimedia Commons (CC4)

Key idea: Introduce latent variables

- Assume that the data consists of observed variables $\boldsymbol{X}$ and unobserved (aka *latent*) variables collectively denoted as $\boldsymbol{Z}$
- Such an approach directly models the situation where some variables are indeed unobserved
- Introducing additional variables might seem redundant
- However, a smart choice of latent variables can make calculations easier
  * Example: in GMM, if we let z_i denote true cluster membership for each point $\boldsymbol{x}_i$, computing the likelihood with known values $\boldsymbol{z}$ is simplified (see next section)

Side note: Jensen's inequality

- Compares effect of averaging before and after applying a convex function: $f\left(Average(\boldsymbol{x})\right) \leq Average\left(f(\boldsymbol{x})\right)$
- Example:
  - Let f be some convex function, such as $f(x) = x^2$
  - Consider $\boldsymbol{x} = [1,2,3,4,5]'$, then $f(\boldsymbol{x}) = [1,4,9,16,25]'$
  - Average of input $Average(\boldsymbol{x}) = 3$
  - $f\left(Average(\boldsymbol{x})\right) = 9$
  - Average of output $Average\left(f(\boldsymbol{x})\right) = 12.4$
- Proof follows from the definition of convexity
  - Proof by induction
- General statement:
  - If $\boldsymbol{X}$ random variable, f is a convex function
  - $f(\mathbb{E}[\boldsymbol{X}]) \leq \mathbb{E}[f(\boldsymbol{X})]$



plot: MHz'as at Wikimedia Commons (public domain)

Putting the latent variables in use

- We want to maximise $\log p(\boldsymbol{X}|\boldsymbol{\theta})$. We don't know $\boldsymbol{Z}$, but consider an arbitrary non-zero distribution $p(\boldsymbol{Z})$
- $\boxed{\log p(\boldsymbol{X}|\boldsymbol{\theta})} = \log \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ ← Rule of marginal distribution (here $\sum_{\boldsymbol{Z}} \dots$ iterates over all possible values of $\boldsymbol{Z}$)
- $= \log \sum_{\boldsymbol{Z}} \left(p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \frac{p(\boldsymbol{Z})}{p(\boldsymbol{Z})} \right)$
- $= \log \sum_{\boldsymbol{Z}} \left(p(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z})} \right)$
- $= \log \mathbb{E}_{\boldsymbol{Z}} \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z})} \right]$
- $\geq \mathbb{E}_{\boldsymbol{Z}} \left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z})} \right]$ ← Jensen's inequality holds since $\log(\dots)$ is a concave function
- $= \boxed{\mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z})]}$

Maximising the lower bound (1/2)

- $\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z})]$
- The right hand side (RHS) is a lower bound on the original log likelihood
  - This holds for any $\boldsymbol{\theta}$ and any non zero $p(\boldsymbol{Z})$
- Intuitively, we want to push the lower bound up
- This lower bound is a function of two “variables” $\boldsymbol{\theta}$ and $p(\boldsymbol{Z})$. We want to maximise the RHS as a function of these “variables”
- It is hard to optimise with respect to both at the same time, so EM resorts to an iterative procedure

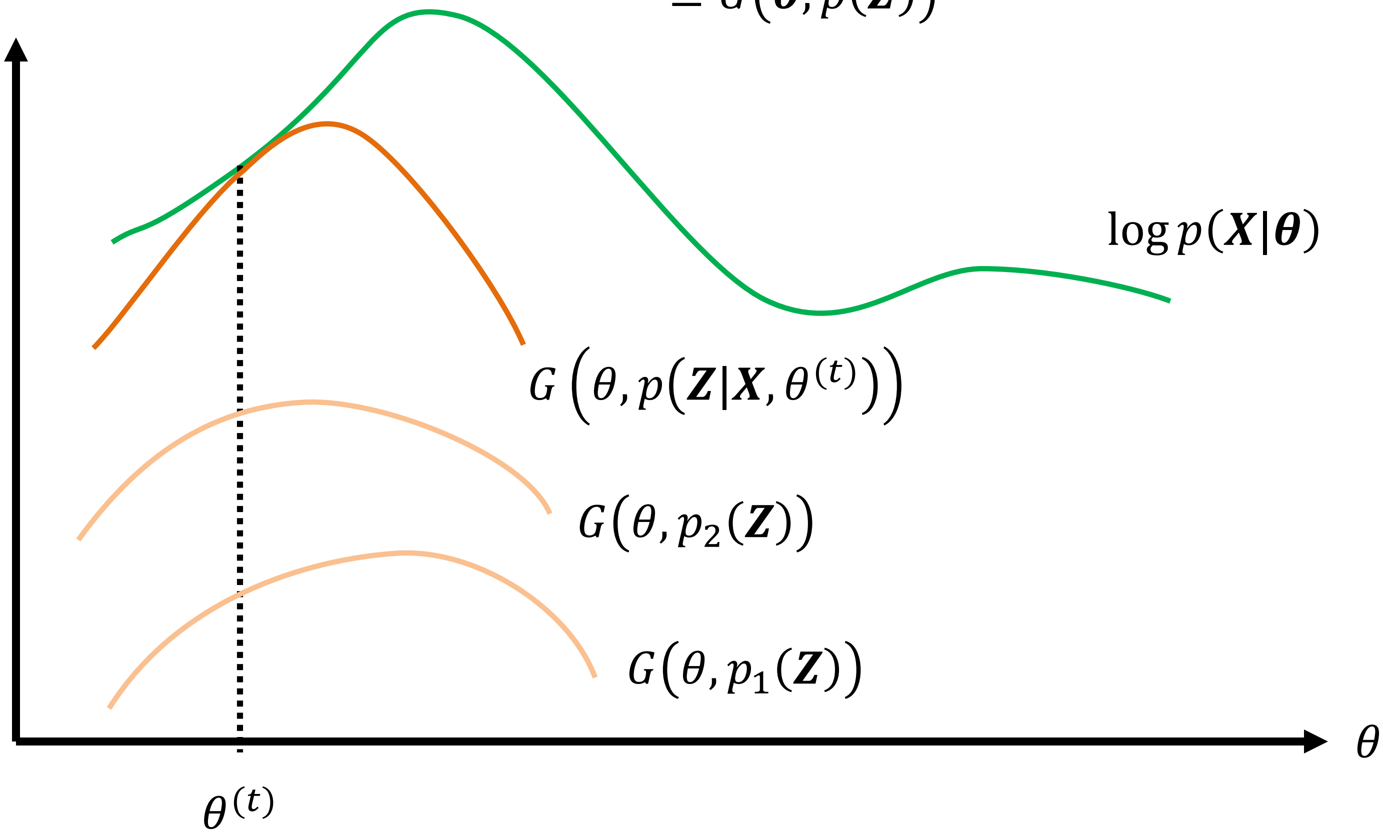
Maximising the lower bound (2/2)

- $\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z})]$
- EM is essentially coordinate descent:
  * Fix $\boldsymbol{\theta}$ and optimise the lower bound for $p(\boldsymbol{Z})$
  * Fix $p(\boldsymbol{Z})$ and optimise for $\boldsymbol{\theta}$
- The convenience of EM follows from the following
- For any point $\boldsymbol{\theta}^*$, it can be shown that setting $p(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^*)$ makes the lower bound tight
- For any $p(\boldsymbol{Z})$, the second term does not depend on $\boldsymbol{\theta}$
- When $p(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^*)$, the first term can usually be maximised as a function of $\boldsymbol{\theta}$ in a closed-form
  * If not, then probably don't use EM

we will prove this shortly

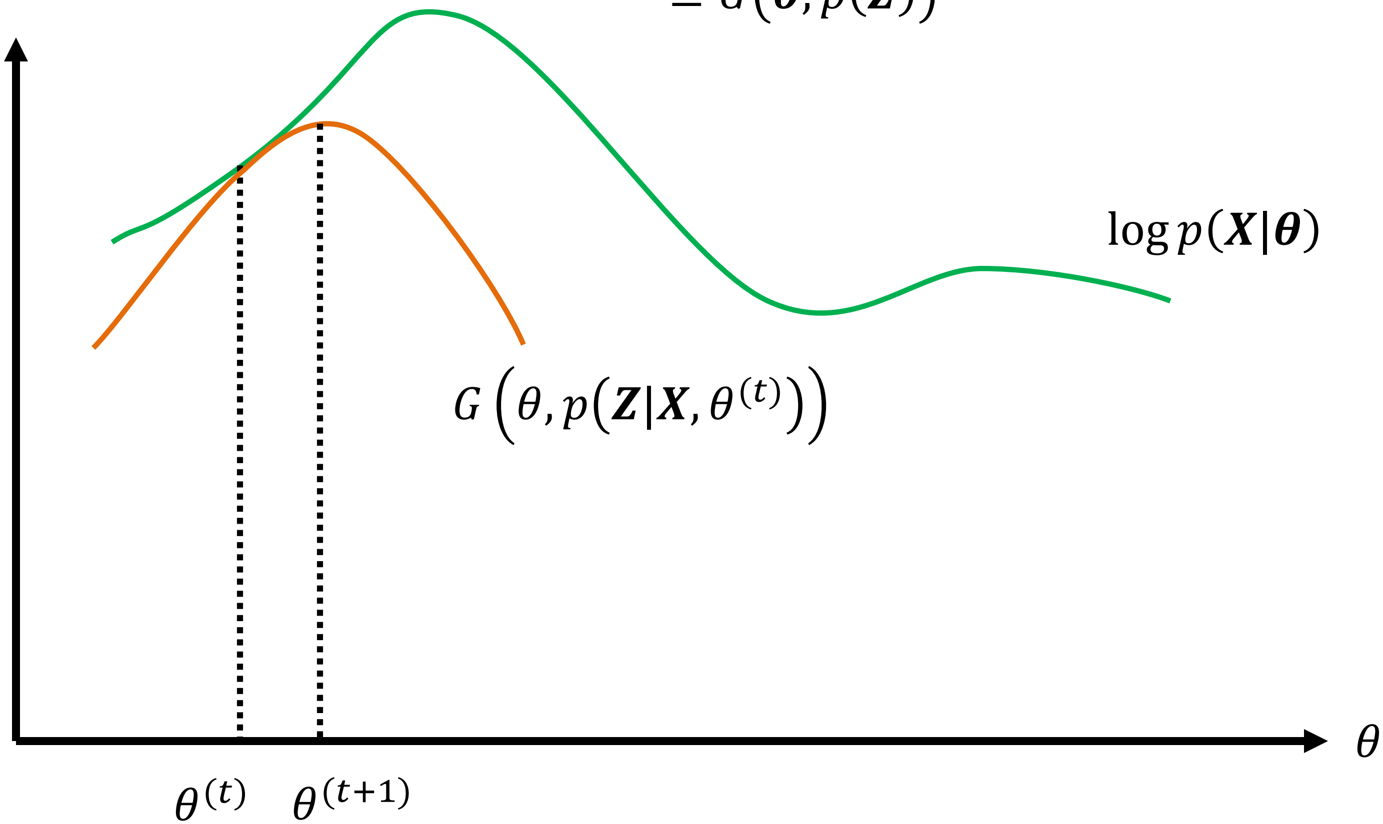
Example (1/3)

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z})]}_{\equiv G(\boldsymbol{\theta}, p(\boldsymbol{Z}))}$$



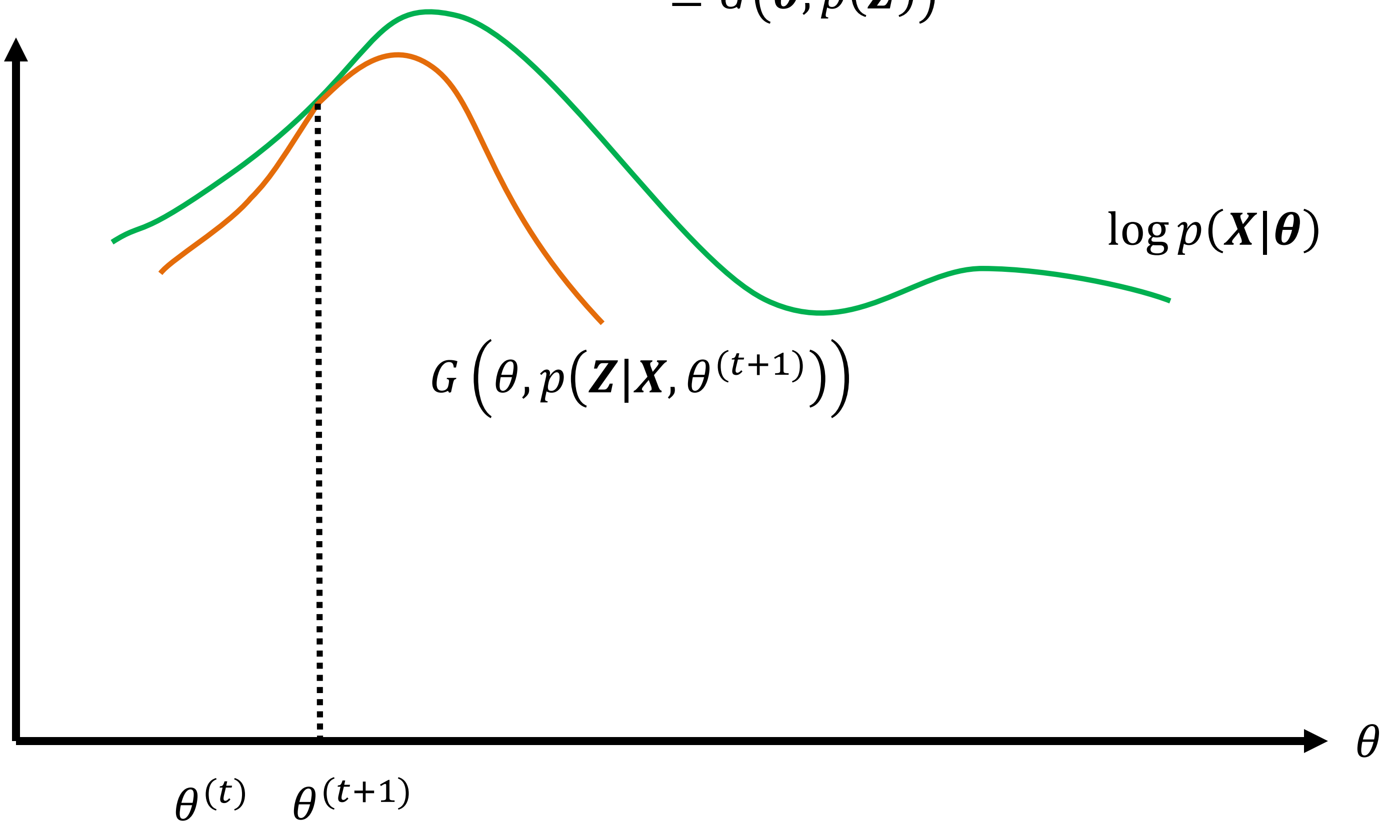
Example (2/3)

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z})]}_{\equiv G(\boldsymbol{\theta}, p(\boldsymbol{Z}))}$$



Example (3/3)

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z})]}_{\equiv G(\boldsymbol{\theta}, p(\boldsymbol{Z}))}$$



EM as iterative optimisation

1. Initialisation: choose initial values of $\boldsymbol{\theta}^{(1)}$
2. Update:
   * E-step: compute $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}\right) \equiv \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})]$
   * M-step: $\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}\right)$
3. Termination: if no change then stop
4. Go to Step 2

This algorithm will eventually stop (converge), but the resulting estimate can be only a local maximum

Maximising the lower bound (2/2)

- $\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \mathbb{E}_Z[\log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_Z[\log p(\boldsymbol{Z})]$
- EM is essentially coordinate descent:
  * Fix $\boldsymbol{\theta}$ and optimise the lower bound for $p(\boldsymbol{Z})$
  * Fix $p(\boldsymbol{Z})$ and optimise for $\boldsymbol{\theta}$
- The convenience of EM follows from the following
- For any point $\boldsymbol{\theta}^*$, it can be shown that setting $p(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^*)$ makes the lower bound tight
- For any $p(\boldsymbol{Z})$, the second term does not depend on $\boldsymbol{\theta}$
- When $p(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^*)$, the first term can usually be maximised as a function of $\boldsymbol{\theta}$ in a closed-form
  * If not, then probably don't use EM

we will prove this now

Putting the latent variables in use

- We want to maximise $\log p(\boldsymbol{X}|\boldsymbol{\theta})$. We don't know $\boldsymbol{Z}$, but consider an arbitrary non-zero distribution $p(\boldsymbol{Z})$
- $\boxed{\log p(\boldsymbol{X}|\boldsymbol{\theta})} = \log \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ ← Rule of marginal distribution (here $\sum_{\boldsymbol{Z}} \dots$ iterates over all possible values of $\boldsymbol{Z}$)
- $= \log \sum_{\boldsymbol{Z}} \left(p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \frac{p(\boldsymbol{Z})}{p(\boldsymbol{Z})} \right)$
- $= \log \sum_{\boldsymbol{Z}} \left(p(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z})} \right)$
- $= \log \mathbb{E}_{\boldsymbol{Z}} \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z})} \right]$
- $\boxed{\geq \mathbb{E}_{\boldsymbol{Z}} \left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z})} \right]}$ ← Jensen's inequality holds since $\log(\dots)$ is a concave function
- $= \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z})]$

Setting a tight lower bound (1/2)

- $\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{Z}}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z})}\right]$
    - $= \mathbb{E}_{\boldsymbol{Z}}\left[\log \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta})}{p(\boldsymbol{Z})}\right]$
    - $= \mathbb{E}_{\boldsymbol{Z}}\left[\log \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{Z})} + \log p(\boldsymbol{X}|\boldsymbol{\theta})\right]$
    - $= \mathbb{E}_{\boldsymbol{Z}}\left[\log \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{Z})}\right] + \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}|\boldsymbol{\theta})]$
    - $= \mathbb{E}_{\boldsymbol{Z}}\left[\log \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{Z})}\right] + \log p(\boldsymbol{X}|\boldsymbol{\theta})$
- $\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{Z}}\left[\log \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{Z})}\right] + \log p(\boldsymbol{X}|\boldsymbol{\theta})$

Setting a tight lower bound (2/2)

Ultimate aim: maximise this

Lower bound of what we want to maximise

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{Z}}\left[\log \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{Z})}\right] + \log p(\boldsymbol{X}|\boldsymbol{\theta})$$

First, note that this term* ≤ 0

Second, note that if $p(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$, then

$$\mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\log \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\right] = \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[\log 1] = 0$$

For any $\boldsymbol{\theta}^*$, setting $p(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^*)$ maximises the lower bound on $\log p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ and makes it tight

*Negative Kullback-Leibler divergence between $p(\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$

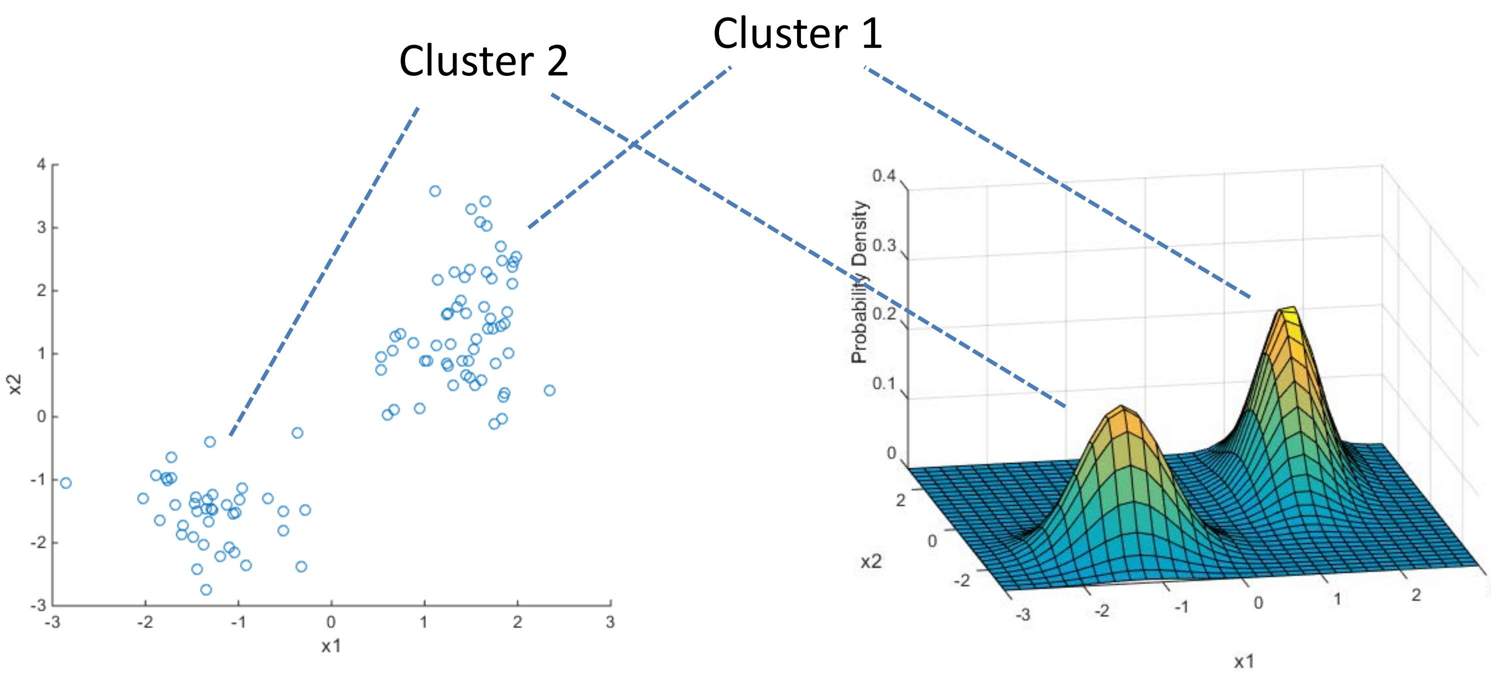
Estimating Parameters of Gaussian Mixture Model

A classical application of the
Expectation Maximisation algorithm

Clustering: Probabilistic interpretation

Clustering can be viewed as identification of components of a probability density function that generated the data

Identifying cluster centroids can be viewed as finding modes of distributions

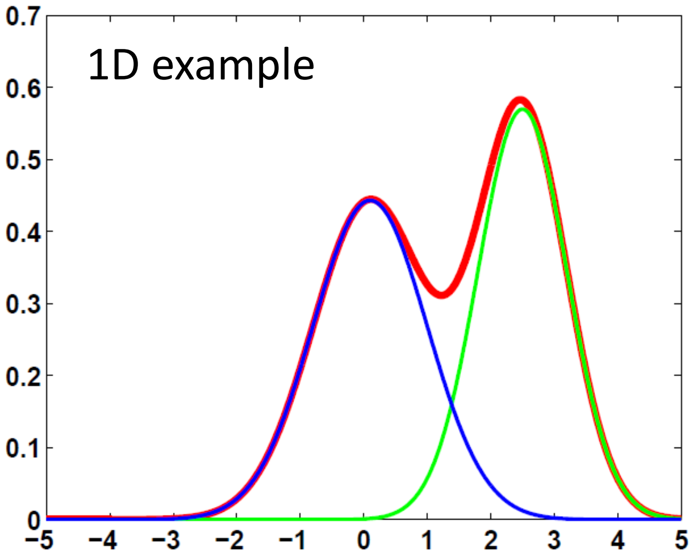


Gaussian mixture model (GMM)

- Gaussian mixture distribution (for one data point):

$$p(\boldsymbol{x}) \equiv \sum_{c=1}^{k} w_c \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

- Here $w_c \geq 0$ and $\sum_{c=1}^{k} w_c = 1$
- That is, $w_1, \dots, w_k$ is a probability distribution over components
- Parameters of the model are $w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, c = 1, \dots, k$



Mixture and individual component densities are re-scaled for visualisation purposes
Figure: Bishop

Fitting a GMM model to data

- Our aim is to find $w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, c = 1, \dots, k$ that maximise

$$\log p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \sum_{i=1}^{n} \log\left(\sum_{c=1}^{k} w_c \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right)$$

- Taking the derivative of this expression is challenging because the log cannot be pushed inside the sum
- Let's see how EM ideas can help

Latent variables of GMM

- Let $z_1, \ldots, z_n$ denote true origins of the corresponding points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Each z_i is a discrete variable that takes values in $1, \ldots, k$, where k is a number of clusters
- Now compare the original log likelihood

$$\log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sum_{i=1}^{n} \log\left(\sum_{c=1}^{k} w_c \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right)$$

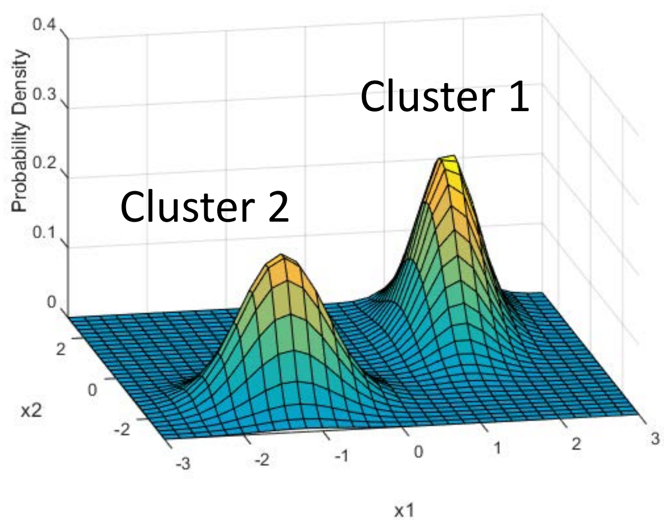
- With *complete data log likelihood* (if we knew $\boldsymbol{z}$)

$$\log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z}) = \sum_{i=1}^{n} \log\left(w_{z_i} \mathcal{N}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)\right)$$

- Recall that taking a log of a normal density function results in a tractable expression

Handling uncertainty about $\boldsymbol{z}$

- We cannot compute complete log likelihood because we don't know $\boldsymbol{z}$
- EM algorithm handles this uncertainty replacing $\log p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})$ with expectation $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}[\log p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})]$
- This in turn requires the distribution of $p\left(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right)$ given current parameter estimates
- Assuming that z_i are pairwise independent, we need to define $P\left(z_i = c|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}\right)$
- E.g., suppose $\boldsymbol{x}_i = (-2, -2)$. What is the probability that this point originated from Cluster 1



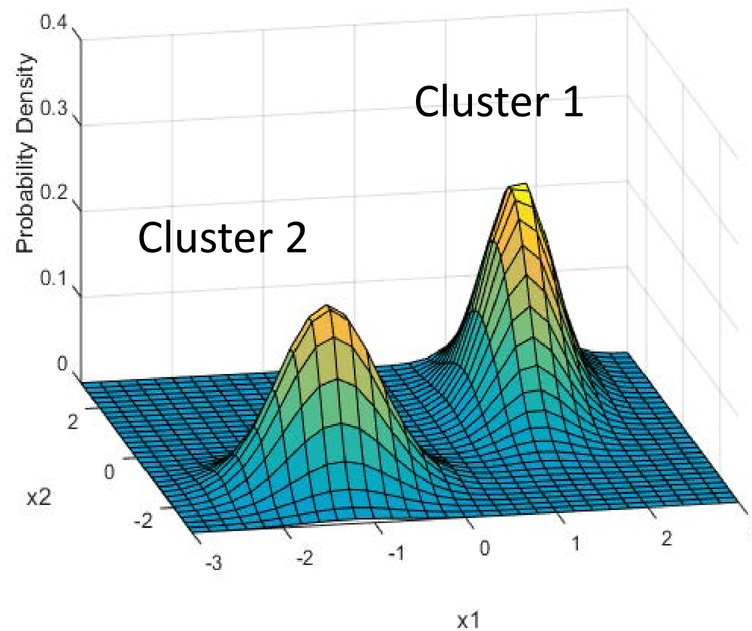
Defining cluster responsibilities

- It is reasonable to use

$$P\left(z_i = c|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}\right) = \frac{w_c \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{l=1}^{k} w_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

- This probability is called *responsibility* that cluster c takes for data point i

$$r_{ic} \equiv P\left(z_i = c|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}\right)$$



Expectation step for GMM

- To simplify notation, we denote $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as $\boldsymbol{X}$, and omit superscript t
- $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}\right) \equiv \mathbb{E}_{\boldsymbol{z}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}[\log p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})]$
- $= \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right) \log p(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\theta})$
- $= \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right) \sum_{i=1}^{n} \log w_{z_i} \mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)$
- $= \sum_{i=1}^{n} \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right) \log w_{z_i} \mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)$
- $= \sum_{i=1}^{n} \sum_{c=1}^{k} r_{ic} \log w_{z_i} \mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)$
- $= \sum_{i=1}^{n} \sum_{c=1}^{k} r_{ic} \log w_{z_i}$
- $+ \sum_{i=1}^{n} \sum_{c=1}^{k} r_{ic} \log \mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)$

Maximisation step for GMM

- In the maximisation step, take partial derivatives of $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}\right)$ with respect to each of the parameters and set the derivatives to zero to obtain new parameter estimates

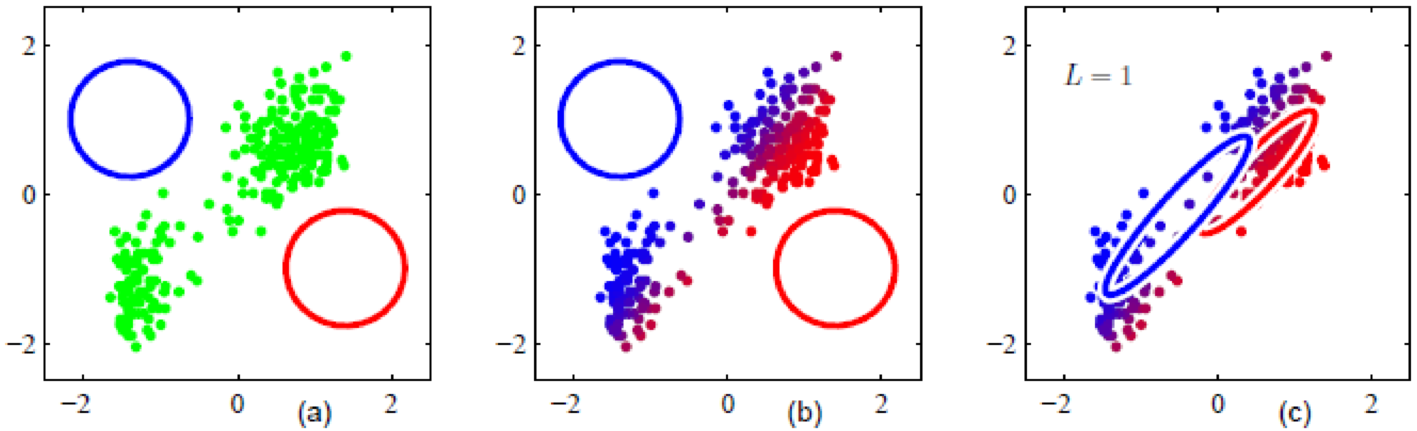
- $w_c^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ic}$

- $\boldsymbol{\mu}_c^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ic}\boldsymbol{x}_i}{r_c}$
  * Here $r_c \equiv \sum_{i=1}^{n} r_{ic}$

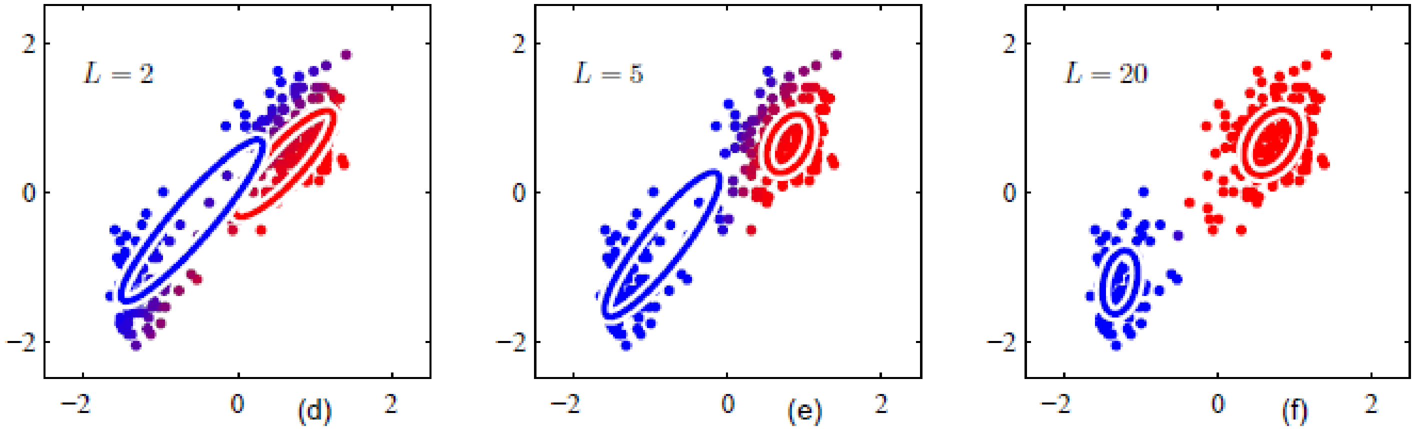
- $\boldsymbol{\Sigma}_c^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ik}\boldsymbol{x}_i\boldsymbol{x}_i'}{r_k} - \boldsymbol{\mu}_c^{(t)}\left(\boldsymbol{\mu}_c^{(t)}\right)'$

- Note that these are the estimates for step $(t+1)$

Example of fitting Gaussian Mixture model



(a) Initial (b) E-step (c) M-step



(d) 2 cycles (e) 5-cyclces (f) 20-cycles

K-means as a EM for a restricted GMM

- Consider a GMM model in which all components have the same fixed probability $w_c = 1/k$, and each Gaussian has the same fixed covariance matrix $\boldsymbol{\Sigma}_c = \sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix
- In such a model, only component centroids $\boldsymbol{\mu}_c$ need to be estimated
- Next approximate a probabilistic cluster responsibility $r_{ic} = P\left(z_i = c | \boldsymbol{x}_i, \boldsymbol{\mu}_c^{(t)}\right)$ with a deterministic assignment $r_{ic} = 1$ if centroid $\boldsymbol{\mu}_c^{(t)}$ is closest to point $\boldsymbol{x}_i$, and $r_{ic} = 0$ otherwise
- Such a formulation results in a E-step where $\boldsymbol{\mu}_c$ should be set as a centroid of points assigned to cluster c
- In other words, k-means algorithm is a EM algorithm for the restricted GMM model described above

This lecture

- Expectation Maximisation (EM) algorithm
  - Introduction in general form
  - Jensen's inequality
  - EM as a coordinate descent approach
- EM applied to Gaussian Mixture Model
  - An iterative approach for parameter estimation
  - K-means as a limiting case of EM for GMM