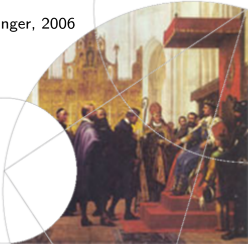
Mixture Modeling and EM Algorithm

Christian Igel
igel@diku.dk

Department of Computer Science
University of Copenhagen

pictures from C. M. Bishop: *Pattern Recognition and Machine Learning*, Springer, 2006

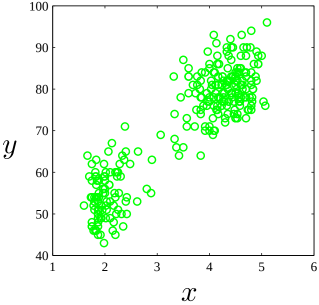


Outline

Outline

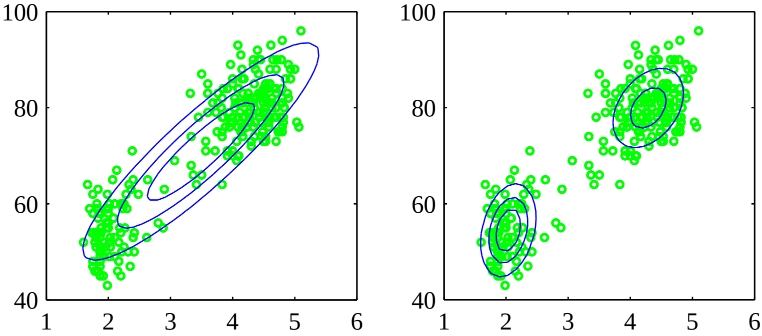
Example: Old Faithful

- Hydrothermal geyser in Yellowstone National Park, Wyoming, USA



- x-axis duration of eruption in minutes
- y-axis time to next eruption in minutes

Density estimation

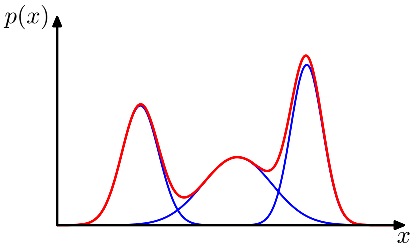


Outline

Mixture modeling

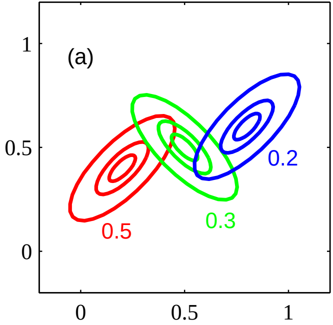
Mixture modeling: Describing complex distributions by convex combinations of simpler distributions:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x}|k) \quad , \quad \sum_{k=1}^{K} p(k) = 1$$



Red line: $p(x)$

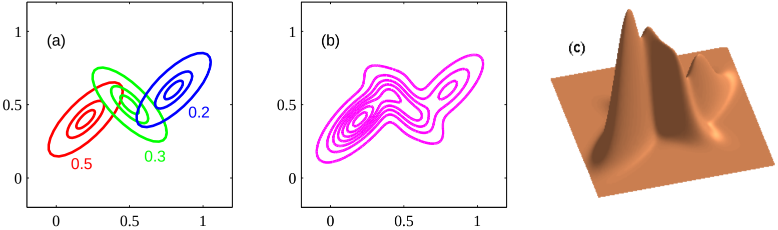
Blue lines: $p(k)p(x|k), k = 1, 2, 3$



Gaussians Mixture Models (GMMs)

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad , \quad \sum_{k=1}^{K} \pi_k = 1 \quad , \quad \boldsymbol{x} \in \mathbb{R}^D$$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^D \det|\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\mathsf{T} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right)$$



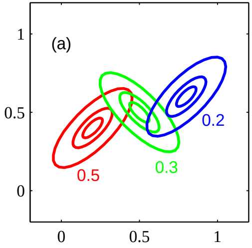
Generative process

To sample GMM distribution

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

we take a generative view:

1. First draw a component number k with relative probabilities π_k,
2. then draw a random vector $\boldsymbol{x}$ from the given component with density $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.



Outline

Maximum likelihood learning

- Training set is $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ..., \boldsymbol{x}_N\}$
- Likelihood function (assuming i.i.d. data) is given by

$$p(S \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

- Parameters $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$
- Cost function is the negative logarithmic likelihood (notice sum inside log):

$$\begin{aligned} E(\boldsymbol{\theta}) &= -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log \sum_{k=1}^{K} p(\boldsymbol{x}_n|\boldsymbol{\theta}_k)\pi_k \\ &= -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

GMM maximum likelihood

- Necessary condition for an extremum is that the partial derivatives w.r.t. the parameters $\boldsymbol{\theta} = \{\pi_k, \ldots, \boldsymbol{\mu}_k, \ldots, \boldsymbol{\Sigma}_k, \ldots\}$ of

$$E(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

  are zero.

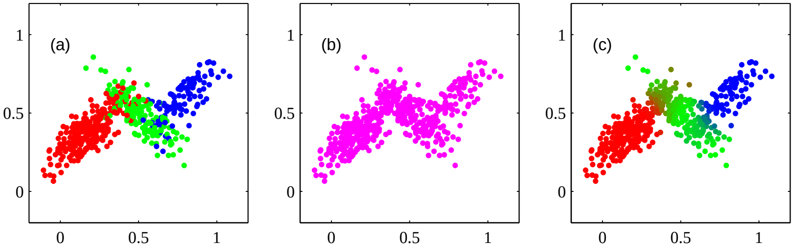
- We define the *responsibility* ($\gamma(z_{nk})$ in Bishop's textbook)

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \in [0, 1] \ .$$

- Responsibilities depend on GMM parameters; let us assume them to be fixed for the moment.

Responsibility

$$
\begin{aligned}
\gamma_{nk} &= \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \\
&= \frac{p(k) p(\boldsymbol{x}_n | k)}{\sum_{k'} p(k') p(\boldsymbol{x}_n | k')} = \frac{p(k) p(\boldsymbol{x}_n | k)}{p(\boldsymbol{x}_n)} \\
&= p(\text{pattern } n \text{ generated by component } k \,|\, \boldsymbol{x}_n) \in [0, 1]
\end{aligned}
$$



Maximizing likelihood I

Using (recall $\frac{\partial}{\partial \boldsymbol{z}} \boldsymbol{z}^{\mathsf{T}} \boldsymbol{B} \boldsymbol{z} = (\boldsymbol{B} + \boldsymbol{B}^{\mathsf{T}})\boldsymbol{z}$)

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

setting

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} E(\boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\mu}_k} \left\{ \sum_{n=1}^{N} \log \sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}) \right\}$$

to zero yields

$$0 = \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \ .$$

Maximizing likelihood II

- From $0 = \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)$ we conclude:

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \text{ with } N_k = \sum_{n=1}^{N} \gamma_{nk}$$

- The same can be done for the other parameters, but is more difficult (and an excellent exercise). We get:

$$\boldsymbol{\Sigma}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \text{ and } \pi_k \leftarrow \frac{N_k}{N}$$

- This is not the gradient of $E(\boldsymbol{\theta})$, because responsibilities depend on $\boldsymbol{\theta}$.
- We do not do std. gradient descent on negative logarithmic likelihood $E(\boldsymbol{\theta})$, but apply an iterative, two-step optimization scheme.

Expectation Maximization (EM) for GMM

EM Algorithm for GMMs

init parameters

while *termination criterion not met* **do**

```
/* E step                                                    */
```

for $k = 1, \ldots, K$ **do**

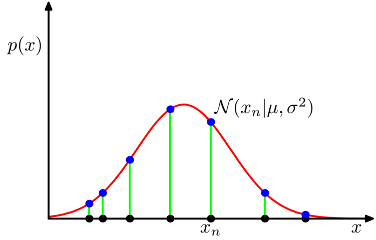
$$\gamma_{nk} \leftarrow \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

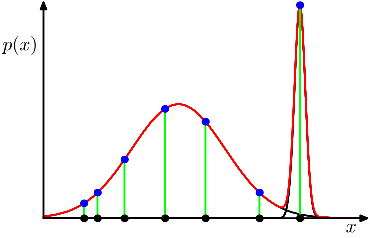
```
/* M step                                                    */
```

$$N_k \leftarrow \sum_{n=1}^{N} \gamma_{nk} \quad , \quad \pi_k \leftarrow \frac{N_k}{N}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \quad , \quad \boldsymbol{\Sigma}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T$$

Nature of the maximum likelihood solution

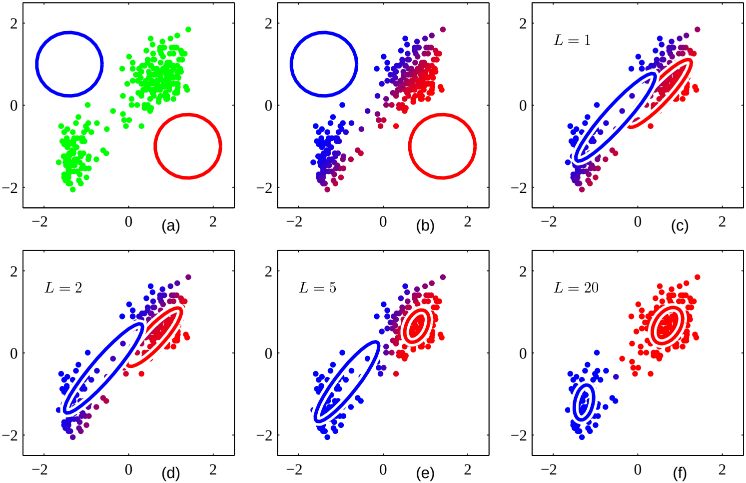




$$E(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Consider cost when $\boldsymbol{\mu}_k = \boldsymbol{x}_n$, $\pi_k > 0$ and $\boldsymbol{\Sigma}_k \to 0$

GMM for Old Faithful



Outline

Problem statement

- Given a set of observed $\boldsymbol{X}$ and hidden (latent) $\boldsymbol{Z}$ random variables as well as a probabilistic model p with parameters $\boldsymbol{\theta}$.
- We want to compute and/or maximize the likelihood

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

  or its logarithm

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \ .$$

- This is difficult, especially because of the sum which prevents the logarithm to act directly on the joint distribution.
- In contrast, $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ may be easy to compute if p belongs to the exponential family of distributions.

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence between two distribution p and q over $\boldsymbol{Z}$

$$\mathrm{KL}(q\|p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})} = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}$$

(sum turns to integral for continuous random variables) is

- a (non-symmetric) measure of difference between distributions,
- always positive, zero iff the distributions are the same.

Variational lower bound

Proposition

Given some distribution $q(\boldsymbol{Z})$ over the hidden variables, it holds

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{F}(q, \boldsymbol{\theta}) + \mathrm{KL}(q \| p_{\boldsymbol{Z}|\boldsymbol{X}})$$

with Kullback-Leibler divergence

$$\mathrm{KL}(q \| p_{\boldsymbol{Z}|\boldsymbol{X}}) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})}$$

and variational lower bound ("free energy")

$$\mathcal{F}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \quad .$$

Proof

We substitute

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) + \ln p(\boldsymbol{X}|\boldsymbol{\theta})$$

into the definition of the variational lower bound

$$\mathcal{F}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \quad .$$

and get

$$\begin{aligned}\mathcal{F}(q, \boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \left(\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) + \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - \ln q(\boldsymbol{Z})\right) \\ &= -\mathrm{KL}(q \| p_{\boldsymbol{Z}|\boldsymbol{X}}) + \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \quad .\end{aligned}$$

Substituting this into the decomposition shows its correctness.

Kullback-Leibler divergence and free energy

- Free energy $\mathcal{F}$ differs from KL divergence as it
  - has the reverse sign and
  - contains the joint instead of the conditional distribution.
- The properties of the KL divergence imply the lower bound property:
$$\mathcal{F}(q(\boldsymbol{Z}), \boldsymbol{\theta}) \leq \ln p(\boldsymbol{X}|\boldsymbol{\theta})$$
- $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ is also called logarithmic *evidence*, and the variational lower bound is referred to as *evidence lower bound (ELBO)*

General EM algorithm

- The EM algorithm is an iterative method for increasing $p(\boldsymbol{X}|\boldsymbol{\theta})$ by adapting $\boldsymbol{\theta}$.
- In each iteration n, an E step (expectation step) and an M step (maximization step) is performed.

EM Algorithm

init $\boldsymbol{\theta}^{(0)}$, $n \leftarrow 0$
while *termination criterion not met* **do**
  $q^{(n+1)} \leftarrow \operatorname{argmax}_q \mathcal{F}(q, \boldsymbol{\theta}^{(n)})$ (E step)
  $\boldsymbol{\theta}^{(n+1)} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{F}(q^{(n+1)}, \boldsymbol{\theta})$ (M step)
  $n \leftarrow n + 1$

E step

- We have

$$\begin{aligned}\mathcal{F}(q,\boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\ &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\ &= -\operatorname{KL}(q \| p_{\boldsymbol{Z}|\boldsymbol{X}}) + \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \quad .\end{aligned}$$

- This expression is maximized w.r.t. q if the Kullback-Leibler divergence vanishes, that is, if $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$.
- That is, after an optimal E step, we have:

$$\mathcal{F}(q^{(n+1)},\boldsymbol{\theta}^{(n)}) = \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{(n)})$$

M step

- Plugging $q^{(n+1)}(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(n)})$ into

$$\mathcal{F}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln q(\boldsymbol{Z})$$

gives

$$\mathcal{F}(q^{(n+1)}, \boldsymbol{\theta}) = \underbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(n)}) \ln p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\theta})}_{Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})} - \underbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(n)}) \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(n)})}_{-H(q^{(n+1)})}$$

- As $-H(q^{(n+1)})$ is independent of $\boldsymbol{\theta}$, optimizing $\mathcal{F}(q^{(n+1)}, \boldsymbol{\theta})$ w.r.t. to $\boldsymbol{\theta}$ in the M step corresponds to

$$\boldsymbol{\theta}^{(n+1)} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) \ .$$

EM increases likelihood

- Assume the algorithm has not converged and $\boldsymbol{\theta}^{(n+1)} \neq \boldsymbol{\theta}^{(n)}$.
- The M step gave $\mathcal{F}(q^{(n+1)}, \boldsymbol{\theta}^{(n+1)}) > \mathcal{F}(q^{(n+1)}, \boldsymbol{\theta}^{(n)}) = \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{(n)})$.
- Further, $q^{(n+1)}(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(n)})$ and $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(n+1)})$ have a positive Kullback-Leibler divergence.
- Thus, we have

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{(n+1)}) = \underbrace{\mathcal{F}(q^{(n+1)}, \boldsymbol{\theta}^{(n+1)})}_{> \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{(n)})} + \underbrace{\mathrm{KL}(q^{(n+1)} \| p^{(n+1)}_{\boldsymbol{Z}|\boldsymbol{X}})}_{>0} > \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{(n)})$$

showing the increase of both the log-likelihood $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ as well as its lower bound $\mathcal{F}(q, \boldsymbol{\theta})$.

GMM learning and general EM

- Hidden/latent discrete random variables $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$
- 1-hot encoding: $\boldsymbol{z}_i \in \mathbb{R}^K$, $[\boldsymbol{z}_i]_k = 1$ and $[\boldsymbol{z}_i]_j = 0$ for $j \neq k$ if $\boldsymbol{x}_i$ was generated by component k
- Distribution $q(\boldsymbol{Z})$ can be described by $N \times K$ values q_{nk} giving the probability that pattern n was generated by component k.
- $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{(i)})$ is given by $q_{nk} = \gamma_{nk}^{(i)}$ maximizing $\mathcal{F}$ in the E step
- M step directly corresponds to maximizing the likelihood for fixed responsibilities.

Concluding remarks

- EM algorithm is a general technique for maximum likelihood estimation.
- It can be applied to adapt Gaussian mixture models; k-means clustering can be interpreted in the EM framework.
- The general EM algorithm is the basis for *variational methods*.

Note

$$\begin{aligned}
\mathcal{F}(q,\boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})p(\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathbb{E}_{q(\boldsymbol{Z})}(\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta}) + \ln p(\boldsymbol{Z}|\boldsymbol{\theta}) - \ln q(\boldsymbol{Z})) \\
&= \mathbb{E}_{q(\boldsymbol{Z})}(\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})) + \mathbb{E}_{q(\boldsymbol{Z})}(\ln p(\boldsymbol{Z}|\boldsymbol{\theta}) - \ln q(\boldsymbol{Z})) \\
&= \mathbb{E}_{q(\boldsymbol{Z})}(\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})) - \mathrm{KL}(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{\theta}))
\end{aligned}$$

All considerations hold if q depends on $\boldsymbol{X}$. Thus, we can replace $q(\boldsymbol{Z})$ by $q(\boldsymbol{Z}|\boldsymbol{X})$. If we additionally have a prior $p(\boldsymbol{Z}|\boldsymbol{\theta})$ that does not depend on $\boldsymbol{\theta}$, we get the ELBO:

$$\mathbb{E}_{q(\boldsymbol{Z}|\boldsymbol{X})}(\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})) - \mathrm{KL}(q(\boldsymbol{Z}|\boldsymbol{X})\|p(\boldsymbol{Z}))$$

This resembles the common form of the ELBO for variational autoencoders (VAEs).