

# Hidden Variable Models 1: Mixture Models

Chris Williams

School of Informatics, University of Edinburgh

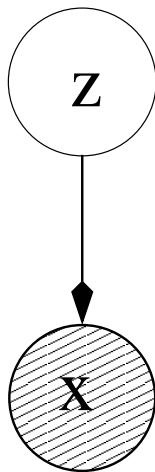
October 2008

# Overview

- Hidden variable models
- Mixture models
- Mixtures of Gaussians
- Aside: Kullback-Leibler divergence
- The EM algorithm
- Bishop §9.2, 9.3, 9.4

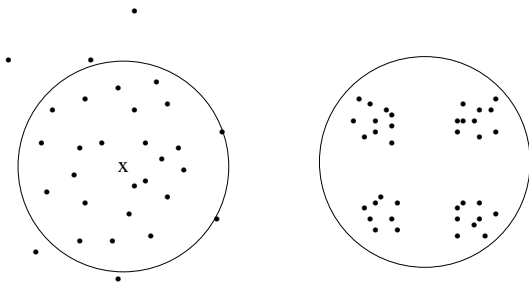
# Hidden Variable Models

- Simplest form is 2 layer structure
- $\mathbf{z}$  hidden (latent) ,  $\mathbf{x}$  visible (manifest)
- Example 1:  $\mathbf{z}$  is discrete → mixture model
- Example 2:  $\mathbf{z}$  is continuous → factor analysis



# Mixture Models

- A single Gaussian might be a poor fit



- Need mixture models for a *multimodal* density

- Let  $\mathbf{z}$  be a 1-of- $k$  indicator variable, with  $\sum_j z_j = 1$ .
- $p(z_j = 1) = \pi_j$  is the probability of that the  $j$ th component is active
- $0 \leq \pi_j \leq 1$  for all  $j$ , and  $\sum_{j=1}^k \pi_j = 1$
- The  $\pi_j$ 's are called the *mixing proportions*

$$p(\mathbf{x}) = \sum_{j=1}^k p(z_j = 1)p(\mathbf{x}|z_j = 1) = \sum_{j=1}^k \pi_j p(\mathbf{x}|\theta_j)$$

- The  $p(\mathbf{x}|\theta_j)$ 's are called the mixture components

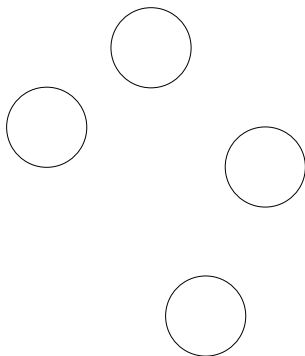
## Generating data from a mixture distribution

for each datapoint

    Choose a component with probability  $\pi_j$

    Generate a sample from the chosen component density

end for



# Responsibilities

$$\begin{aligned}\gamma(z_j) \equiv p(z_j = 1 | \mathbf{x}) &= \frac{p(z_j = 1) p(\mathbf{x} | z_j = 1)}{\sum_{\ell} p(z_{\ell} = 1) p(\mathbf{x} | z_{\ell} = 1)} \\ &= \frac{\pi_j p(\mathbf{x} | z_j = 1)}{\sum_{\ell} \pi_{\ell} p(\mathbf{x} | z_{\ell} = 1)}\end{aligned}$$

- $\gamma(z_j)$  is the posterior probability (or responsibility) for component  $j$  to have generated datapoint  $\mathbf{x}$

## Maximum likelihood estimation for mixture models

$$L(\theta) = \sum_{i=1}^n \ln \left\{ \sum_{j=1}^k \pi_j p(\mathbf{x}_i | \theta_j) \right\}$$

$$\frac{\partial L}{\partial \theta_j} = \sum_i \frac{\pi_j}{\sum_{\ell} \pi_{\ell} p(\mathbf{x}_i | \theta_{\ell})} \frac{\partial p(\mathbf{x}_i | \theta_j)}{\partial \theta_j}$$

now use

$$\frac{\partial p(\mathbf{x}_i | \theta_j)}{\partial \theta_j} = p(\mathbf{x}_i | \theta_j) \frac{\partial \ln p(\mathbf{x}_i | \theta_j)}{\partial \theta_j}$$

and therefore

$$\frac{\partial L}{\partial \theta_j} = \sum_i \gamma(z_{ij}) \frac{\partial \ln p(\mathbf{x}_i | \theta_j)}{\partial \theta_j}$$

## Example: 1-d Gaussian mixture

$$p(x|\theta_j) = \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp - \left\{ \frac{(x - \mu_j)^2}{2\sigma_j^2} \right\}$$

$$\frac{\partial L}{\partial \mu_j} = \sum_i \gamma(z_{ij}) \frac{(x_i - \mu_j)}{\sigma_j^2}$$

$$\frac{\partial L}{\partial \sigma_j^2} = \frac{1}{2} \sum_i \gamma(z_{ij}) \left[ \frac{(x_i - \mu_j)^2}{\sigma_j^4} - \frac{1}{\sigma_j^2} \right]$$

At a maximum, set derivatives = 0

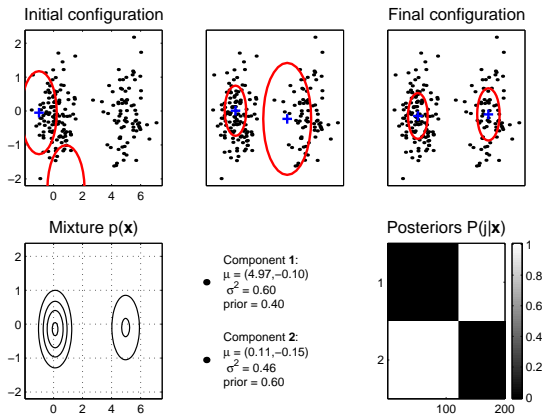
$$\hat{\mu}_j = \frac{\sum_{i=1}^n \gamma(z_{ij})x_i}{\sum_{i=1}^n \gamma(z_{ij})}$$
$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \gamma(z_{ij})(x_i - \hat{\mu}_j)^2}{\sum_{i=1}^n \gamma(z_{ij})}$$
$$\hat{\pi}_j = \frac{1}{n} \sum_i \gamma(z_{ij}).$$

Generalize to multivariate case

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^n \gamma(z_{ij}) \mathbf{x}_i}{\sum_{i=1}^n \gamma(z_{ij})}$$
$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{i=1}^n \gamma(z_{ij}) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)^T}{\sum_{i=1}^n \gamma(z_{ij})}$$
$$\hat{\pi}_j = \frac{1}{n} \sum_i \gamma(z_{ij}).$$

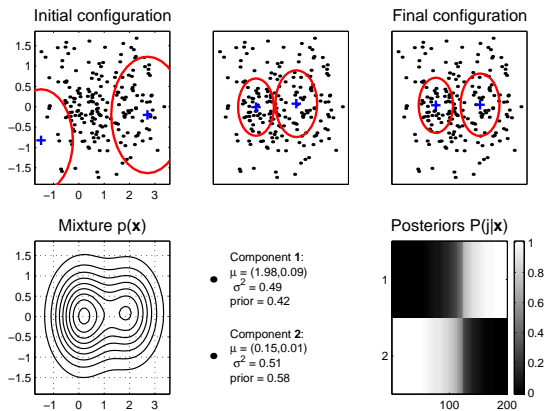
- What happens if a component becomes responsible for a single data point?

# Example



(Tipping, 1999)

# Example 2



## Kullback-Leibler divergence

- Measuring the “distance” between two probability densities  $P(x)$  and  $Q(x)$ .

$$KL(P||Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

- Also called the relative entropy
- Using  $\log z \leq z - 1$ , can show that  $KL(P||Q) \geq 0$  with equality when  $P = Q$ .
- Note that  $KL(P||Q) \neq KL(Q||P)$

# The EM algorithm

- Q: How do we estimate parameters of a Gaussian mixture distribution?
- A: Use the re-estimation equations

$$\hat{\mu}_j \leftarrow \frac{\sum_{i=1}^n \gamma(z_{ij}) x_i}{\sum_{i=1}^n \gamma(z_{ij})}$$
$$\hat{\sigma}_j^2 \leftarrow \frac{\sum_{i=1}^n \gamma(z_{ij}) (x_i - \hat{\mu}_j)^2}{\sum_{i=1}^n \gamma(z_{ij})}$$
$$\hat{\pi}_j \leftarrow \frac{1}{n} \sum_i \gamma(z_{ij}).$$

- This is intuitively reasonable, but the EM algorithm shows that these updates will converge to a local maximum of the likelihood

# The EM algorithm

EM = Expectation-Maximization

- Applies where there is incomplete (or *missing*) data
- If this data were known a maximum likelihood solution would be relatively easy
- In a mixture model, the missing knowledge is which component generated a given data point
- Although EM can have slow convergence to the local maximum, it is usually relatively simple and easy to implement. For Gaussian mixtures it is the method of choice.

## The nitty-gritty

$$L(\theta) = \sum_{i=1}^n \ln p(\mathbf{x}_i|\theta)$$

Consider for just one  $\mathbf{x}_i$  first

$$\log p(\mathbf{x}_i|\theta) = \log p(\mathbf{x}_i, \mathbf{z}_i|\theta) - \log p(\mathbf{z}_i|\mathbf{x}_i, \theta).$$

Now introduce  $q(\mathbf{z}_i)$  and take expectations

$$\begin{aligned} \log p(\mathbf{x}_i|\theta) &= \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \log p(\mathbf{x}_i, \mathbf{z}_i|\theta) - \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \log p(\mathbf{z}_i|\mathbf{x}_i, \theta) \\ &= \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \log \frac{p(\mathbf{x}_i, \mathbf{z}_i|\theta)}{q(\mathbf{z}_i)} - \sum_{\mathbf{z}_i} q(\mathbf{z}_i) \log \frac{p(\mathbf{z}_i|\mathbf{x}_i, \theta)}{q(\mathbf{z}_i)} \\ &\stackrel{\text{def}}{=} \mathcal{L}_i(q_i, \theta) + KL(q_i||p_i) \end{aligned}$$

From the non-negativity of the KL divergence, note that

$$\mathcal{L}_i(q_i, \theta) \leq \log p(\mathbf{x}_i|\theta)$$

i.e.  $\mathcal{L}_i(q_i, \theta)$  is a *lower bound* on the log likelihood

We now set  $q(\mathbf{z}_i) = p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old})$  [E step]

$$\begin{aligned}\mathcal{L}_i(q_i, \theta) &= \sum_{z_i} p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old}) \log p(\mathbf{x}_i, \mathbf{z}_i|\theta) - \sum_{z_i} p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old}) \log p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old}) \\ &\stackrel{def}{=} Q_i(\theta|\theta^{old}) + H(q_i)\end{aligned}$$

Notice that  $H(q_i)$  is independent of  $\theta$  (as opposed to  $\theta^{old}$ )

Now sum over cases  $i = 1, \dots, n$

$$\mathcal{L}(\mathbf{q}, \theta) = \sum_{i=1}^n \mathcal{L}_i(\mathbf{q}_i, \theta) \leq \sum_{i=1}^n \log p(\mathbf{x}_i | \theta)$$

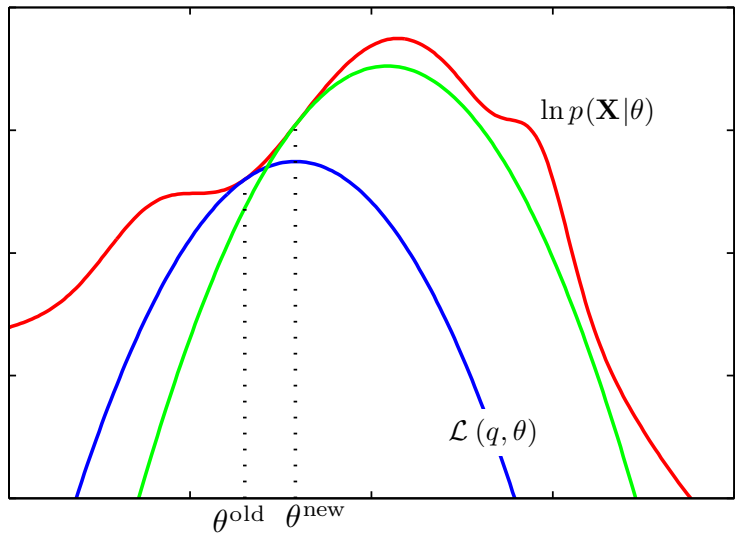
and

$$\begin{aligned} \mathcal{L}(\mathbf{q}, \theta) &= \sum_{i=1}^n Q_i(\theta | \theta^{old}) + \sum_{i=1}^n H(\mathbf{q}_i) \\ &\stackrel{def}{=} Q(\theta | \theta^{old}) + \sum_{i=1}^n H(\mathbf{q}_i) \end{aligned}$$

where  $Q$  is called the expected complete-data log likelihood.  
Thus to increase  $\mathcal{L}(\mathbf{q}, \theta)$  wrt  $\theta$  we need only increase  $Q(\theta | \theta^{old})$

Best to choose [M step]

$$\theta = \operatorname{argmax}_{\theta} Q(\theta | \theta^{old})$$



# EM algorithm: Summary

**E-step** Calculate  $Q(\theta|\theta^{old})$  using the responsibilities  $p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old})$

**M-step** Maximize  $Q(\theta|\theta^{old})$  wrt  $\theta$

EM algorithm for mixtures of Gaussians

$$\begin{aligned}\mu_j^{new} &\leftarrow \frac{\sum_{i=1}^n p(j|\mathbf{x}_i, \theta^{old}) \mathbf{x}_i}{\sum_{i=1}^n p(j|\mathbf{x}_i, \theta^{old})} \\ (\sigma_j^2)^{new} &\leftarrow \frac{\sum_{i=1}^n p(j|\mathbf{x}_i, \theta^{old}) (\mathbf{x}_i - \mu_j^{new})^2}{\sum_{i=1}^n p(j|\mathbf{x}_i, \theta^{old})} \\ \pi_j^{new} &\leftarrow \frac{1}{n} \sum_{i=1}^n p(j|\mathbf{x}_i, \theta^{old}).\end{aligned}$$

[Do mixture of Gaussians demo here]

## *k*-means clustering

```
initialize centres  $\mu_1, \dots, \mu_k$ 
while (not terminated)
  for  $i = 1, \dots, n$ 
    calculate  $|\mathbf{x}_i - \mu_j|^2$  for all centres
    assign datapoint  $i$  to the closest centre
  end for
  recompute each  $\mu_j$  as the mean of the
  datapoints assigned to it
end while
```

*k*-means algorithm is equivalent to the EM algorithm for spherical covariances  $\sigma_j^2 I$  in the limit  $\sigma_j^2 \rightarrow 0$  for all  $j$