Parametric Unsupervised Learning
Expectation Maximization (EM)

Lecture 20.a

Some slides are due to Christopher Bishop

BCS Summer School, Exeter, 2003

Limitations of K-means

- Hard assignments of data points to clusters – small shift of a data point can flip it to a different cluster
- Not clear how to choose the value of K
- Solution: replace ‘hard’ clustering of K-means with ‘soft’ probabilistic assignments
- Represents the probability distribution of the data as a *Gaussian mixture model*

The Gaussian Distribution

- Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

mean

covariance

- Define precision to be the inverse of the covariance

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$$

- In 1-dimension

$$\tau = \frac{1}{\sigma^2}$$

Gaussian Mixtures

- Linear super-position of Gaussians

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

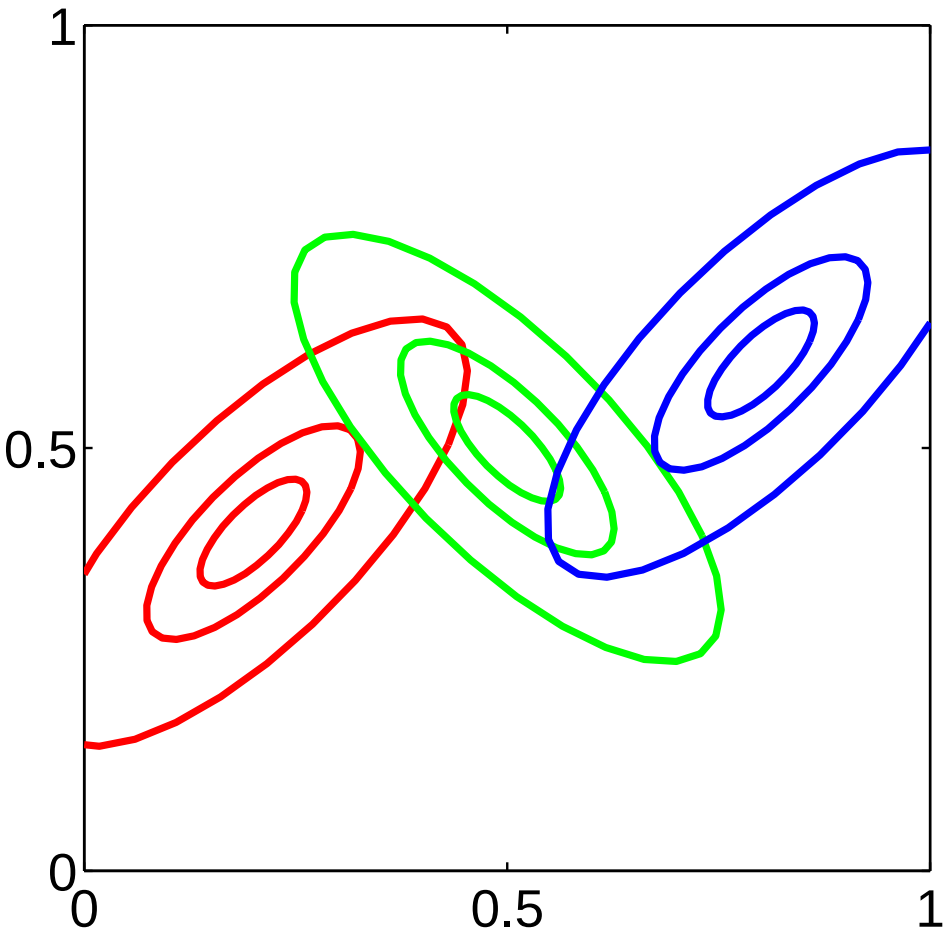
- Normalization and positivity require

$$\sum_{k=1}^{K} \pi_k = 1 \qquad 0 \leqslant \pi_k \leqslant 1$$

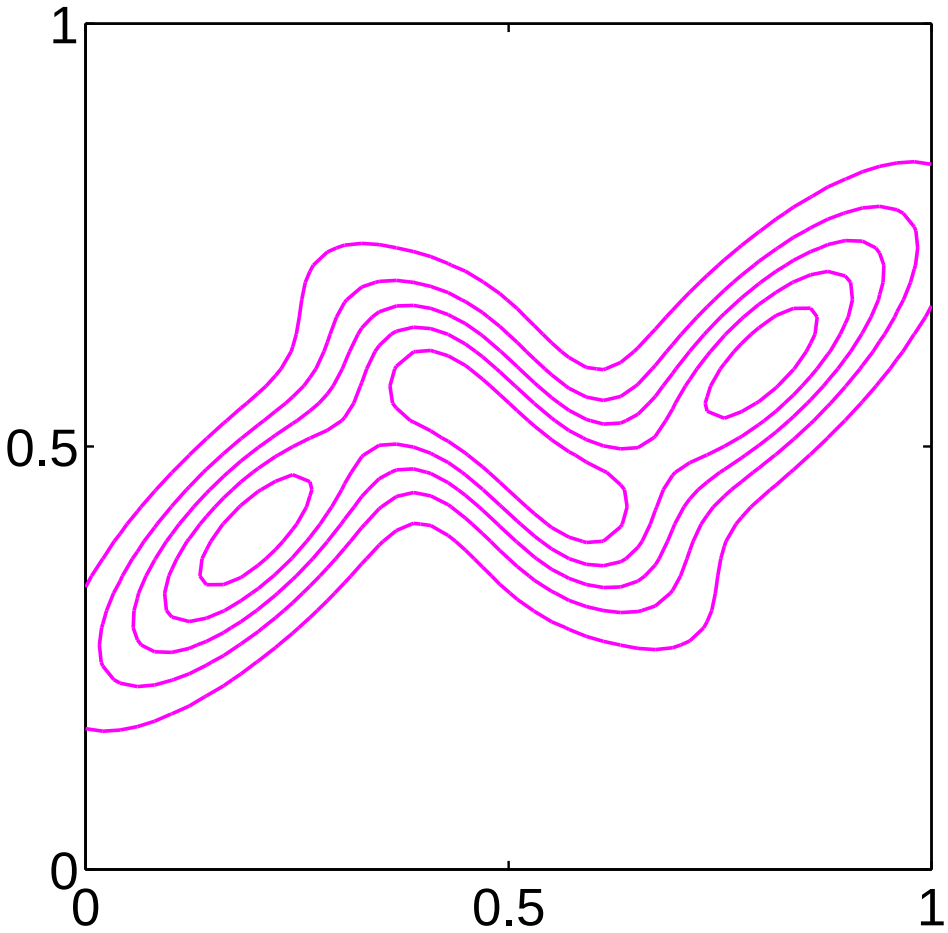
- Can interpret the mixing coefficients as prior probabilities

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

Example: Mixture of 3 Gaussians



Contours of Probability Distribution



Sampling from the Gaussian

- To generate a data point:
  - first pick one of the components with probability π_k
  - then draw a sample $\mathbf{x}_n$ from that component
- Repeat these two steps for each new data point

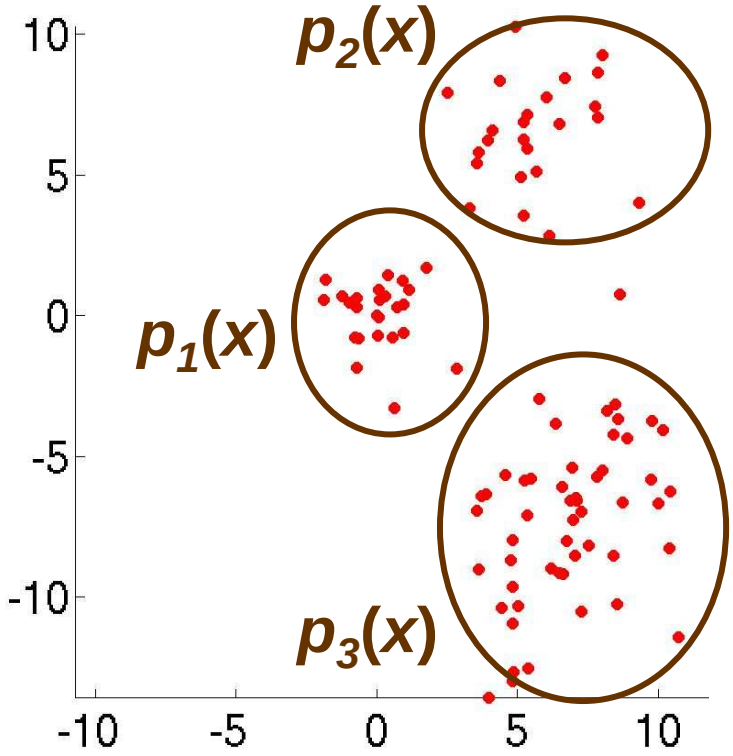
Example: Gaussian Mixture Density

- Mixture of 3 Gaussians

$$p_1(x) \cong N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$$

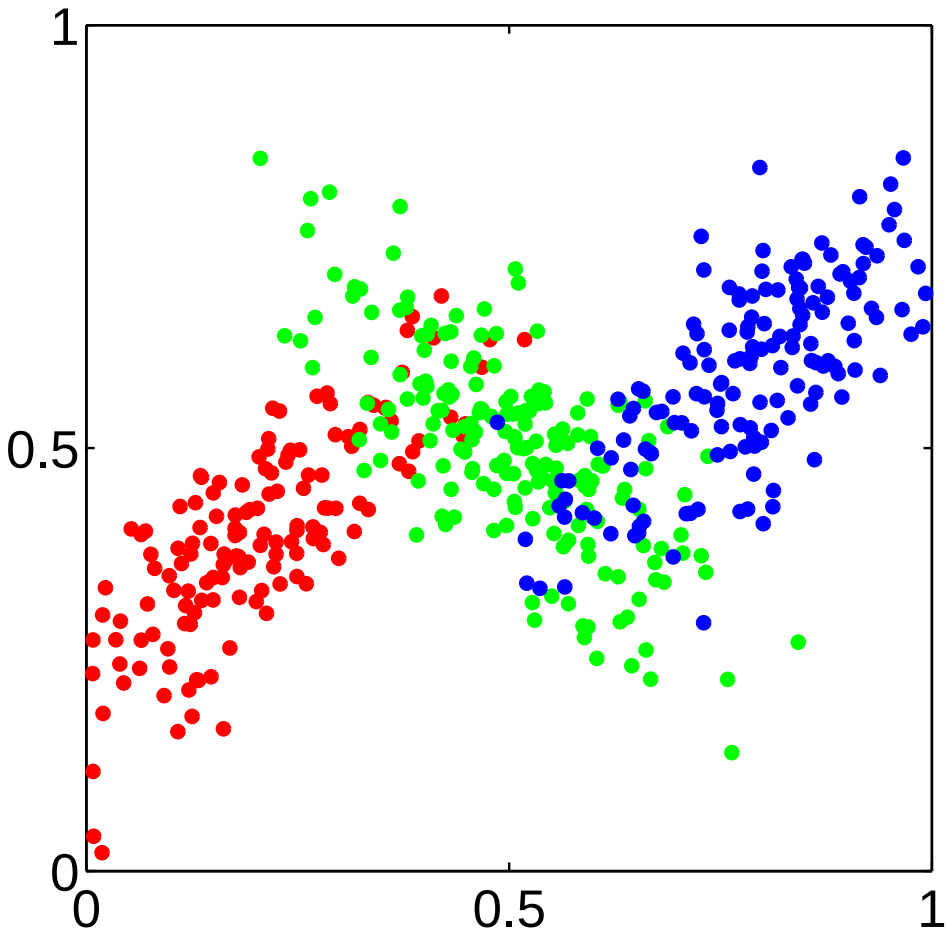
$$p_2(x) \cong N\left([6,6], \begin{bmatrix}4 & 0\\0 & 4\end{bmatrix}\right)$$

$$p_3(x) \cong N\left([7,-7], \begin{bmatrix}6 & 0\\0 & 6\end{bmatrix}\right)$$



$$p(x) = 0.2p_1(x) + 0.3p_2(x) + 0.5p_3(x)$$

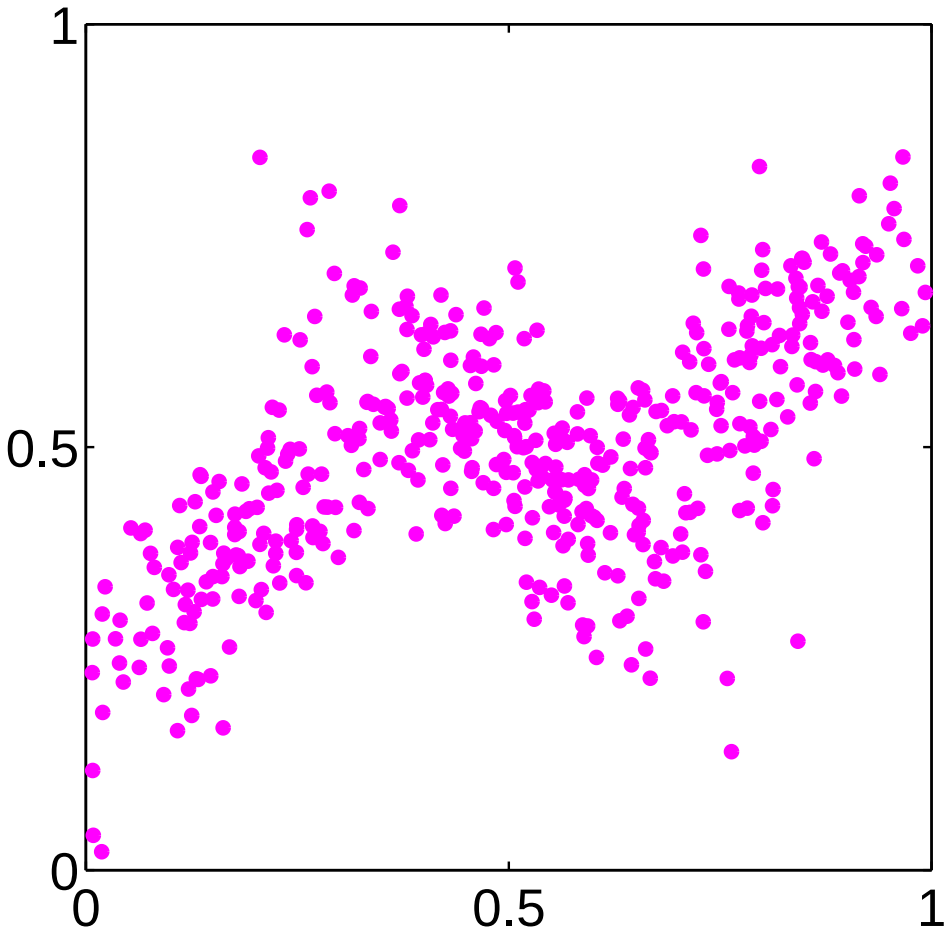
Synthetic Data Set



Fitting the Gaussian Mixture

- We wish to invert this process – given the data set, find the corresponding parameters:
  - mixing coefficients
  - means
  - covariances
- If we knew which component generated each data point, the maximum likelihood solution would involve fitting each component to the corresponding cluster
- Problem: the data set is unlabelled
- We shall refer to the labels as *latent* (= hidden) variables

Synthetic Data Set Without Labels



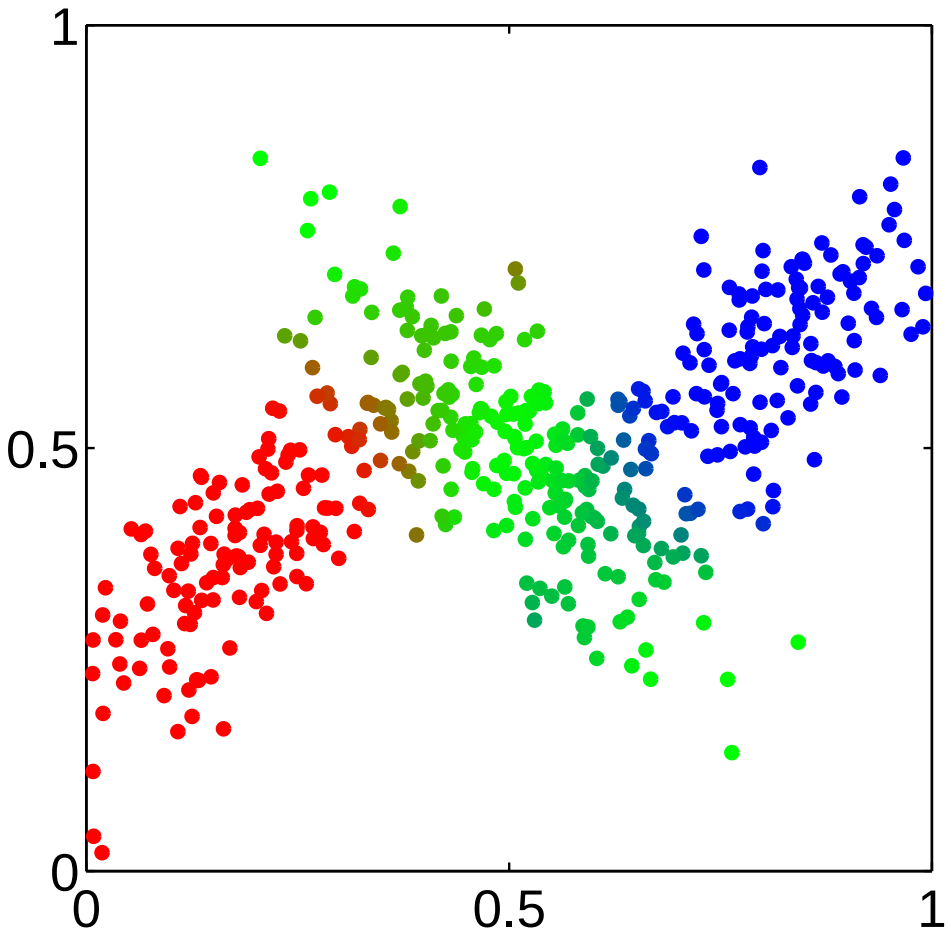
Posterior Probabilities

- We can think of the mixing coefficients as prior probabilities for the components
- For a given value of $\mathbf{x}$ we can evaluate the corresponding posterior probabilities, called *responsibilities*
- These are given from Bayes' theorem by

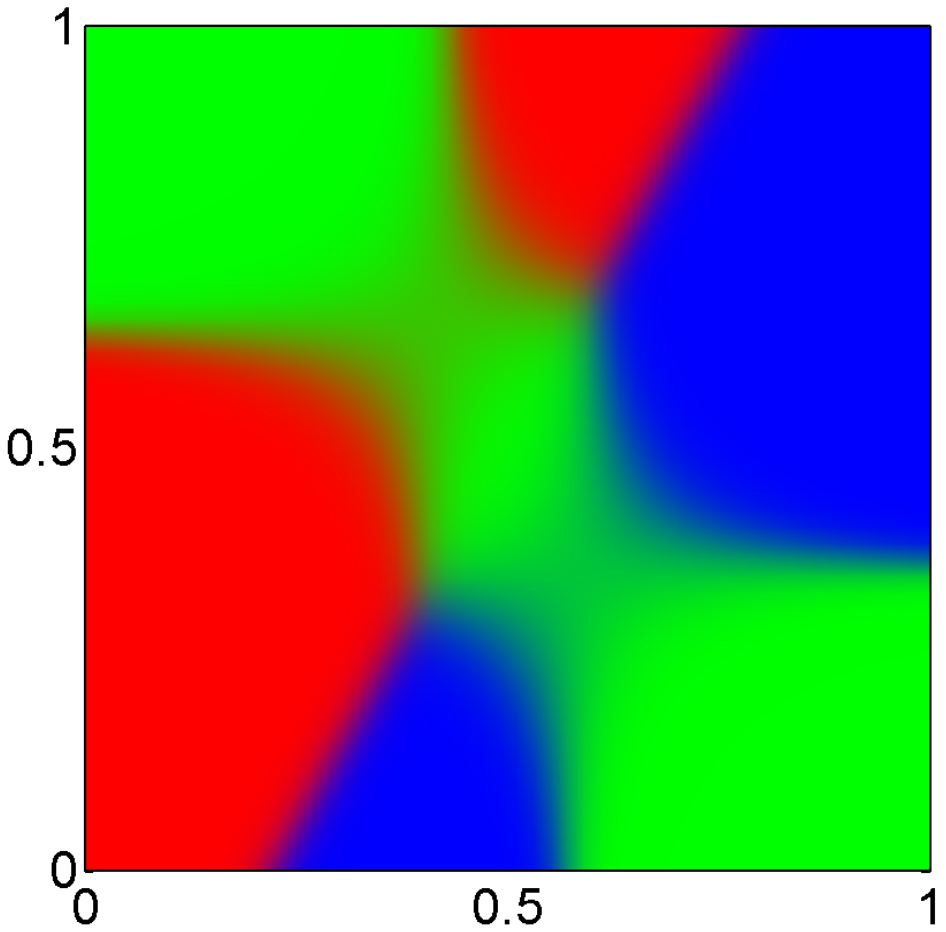
$$\gamma_k(\mathbf{x}) \equiv p(k|\mathbf{x}) = \frac{p(k)p(\mathbf{x}|k)}{p(\mathbf{x})}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Posterior Probabilities (colour coded)



Posterior Probability Map



Maximum Likelihood for the GMM

- The log likelihood function takes the form

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

- Note: sum over components appears *inside* the log
- There is no closed form solution for maximum likelihood
- How to maximize the log likelihood
  - solved by expectation-maximization (EM) algorithm

EM Algorithm – Informal Derivation

- Let us proceed by simply differentiating the log likelihood
- Setting derivative with respect to $\boldsymbol{\mu}_j$ equal to zero gives

$$-\sum_{n=1}^{N} \underbrace{\frac{\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}}_{\gamma_j(\mathbf{x}_n)} \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_j) = 0$$

giving

$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)}$$

which is simply the weighted mean of the data

EM Algorithm – Informal Derivation

- Similarly for the covariances

$$\boldsymbol{\Sigma}_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_j)(\mathbf{x}_n - \boldsymbol{\mu}_j)^{\mathsf{T}}}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)}$$

- For mixing coefficients use a Lagrange multiplier to give

Fraction of points assigned to component j

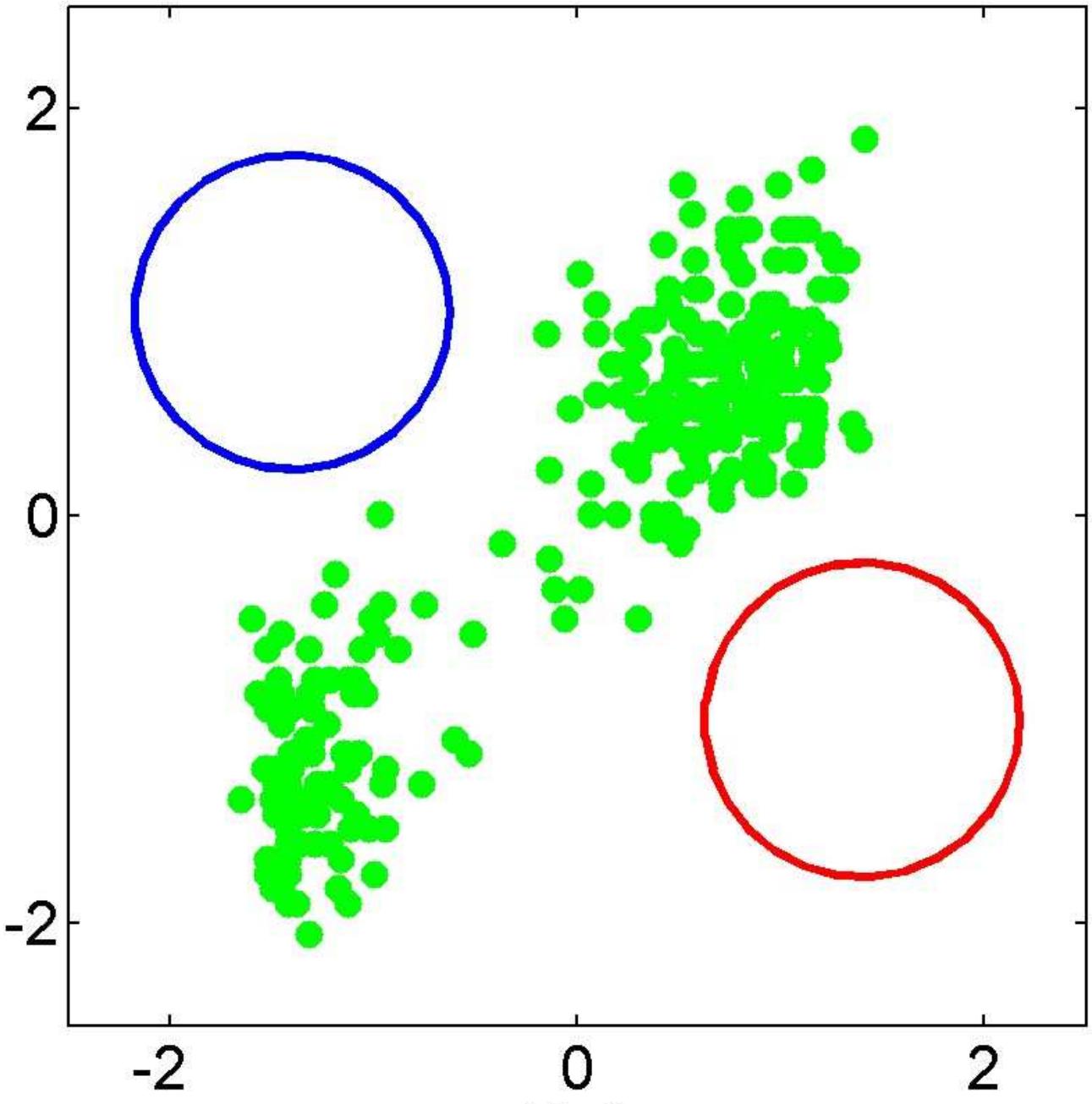
$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)$$

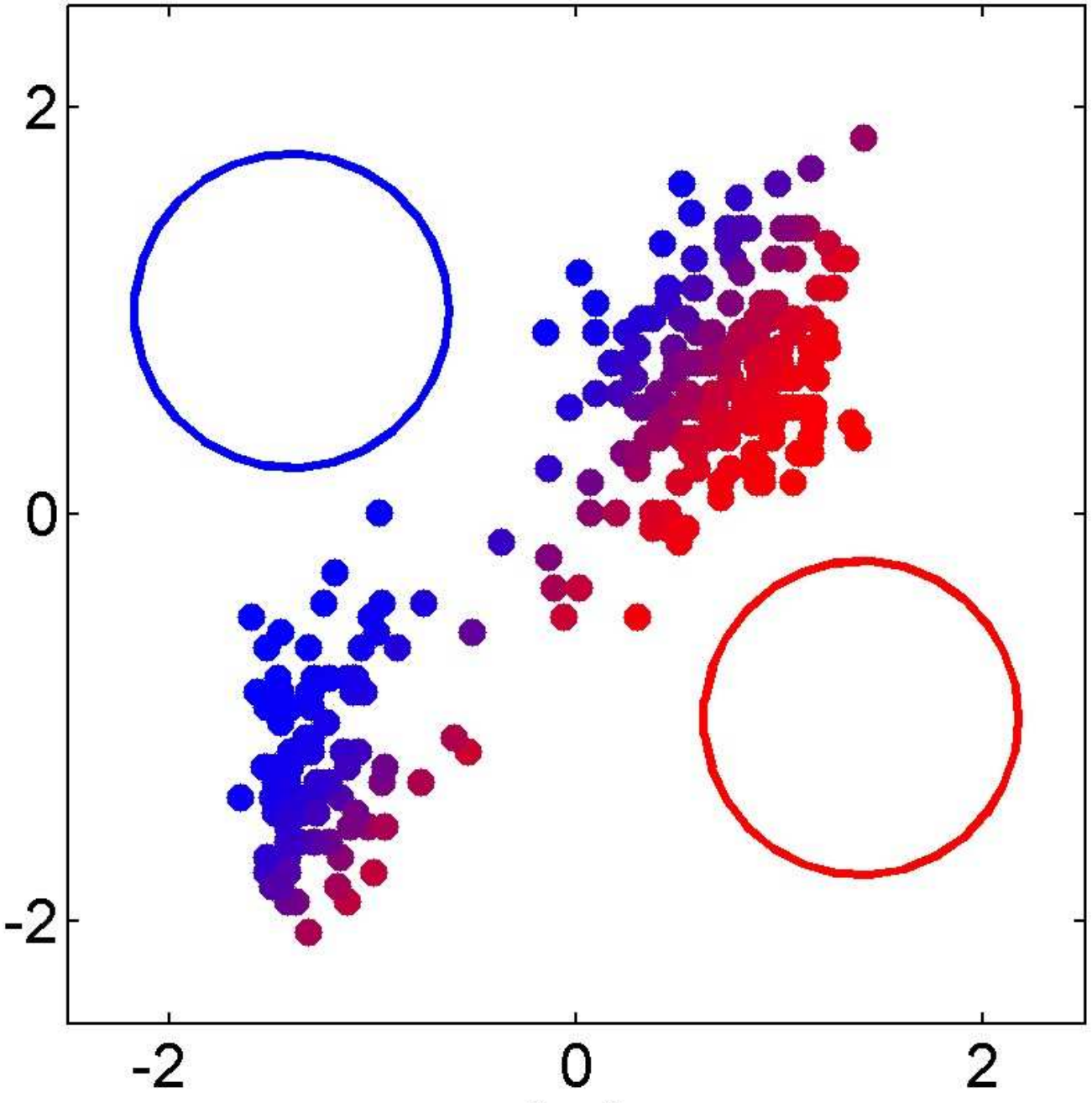
effective number of points assigned to cluster j.

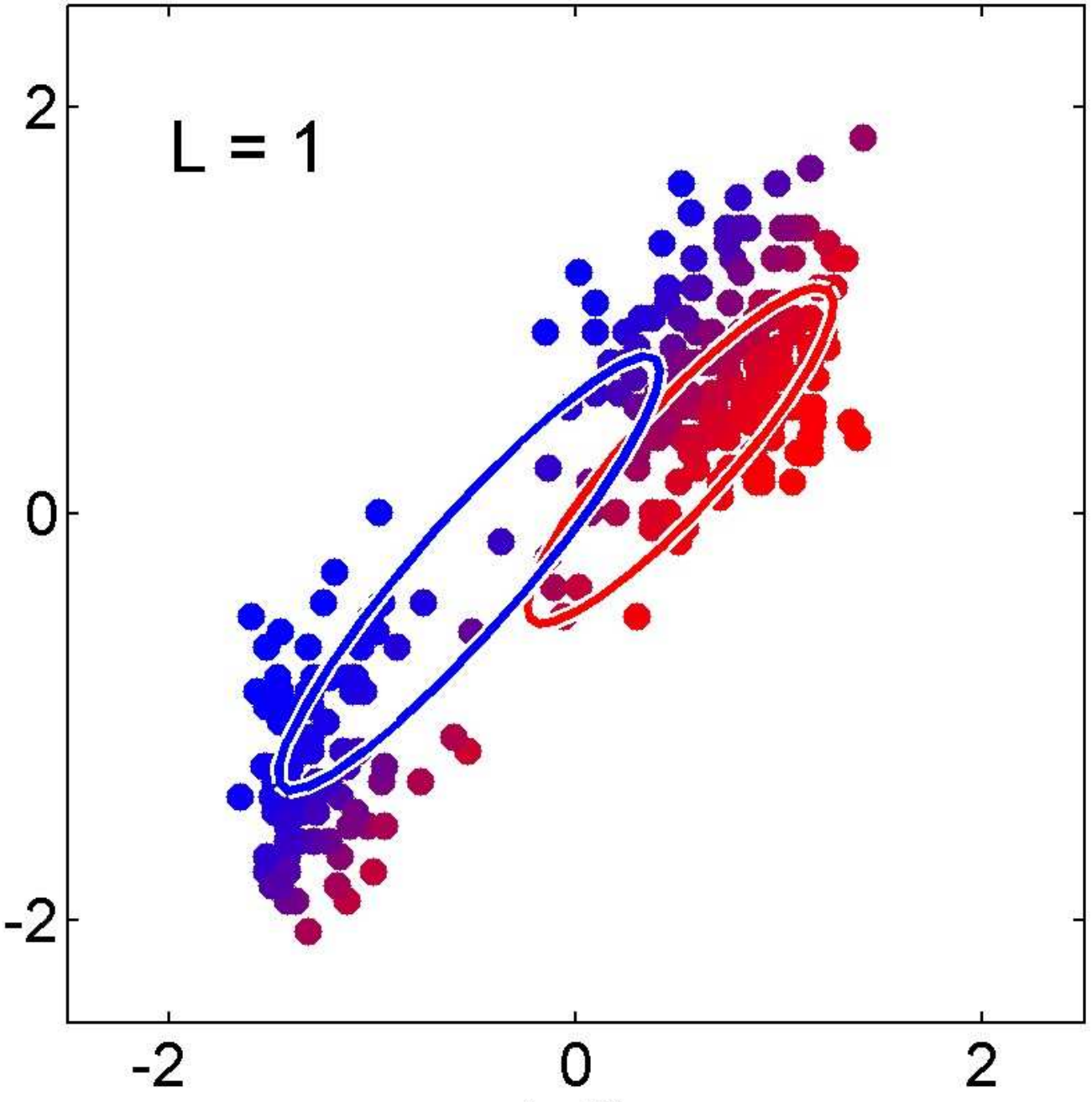
Average responsibility which component j takes for explaining the data points.

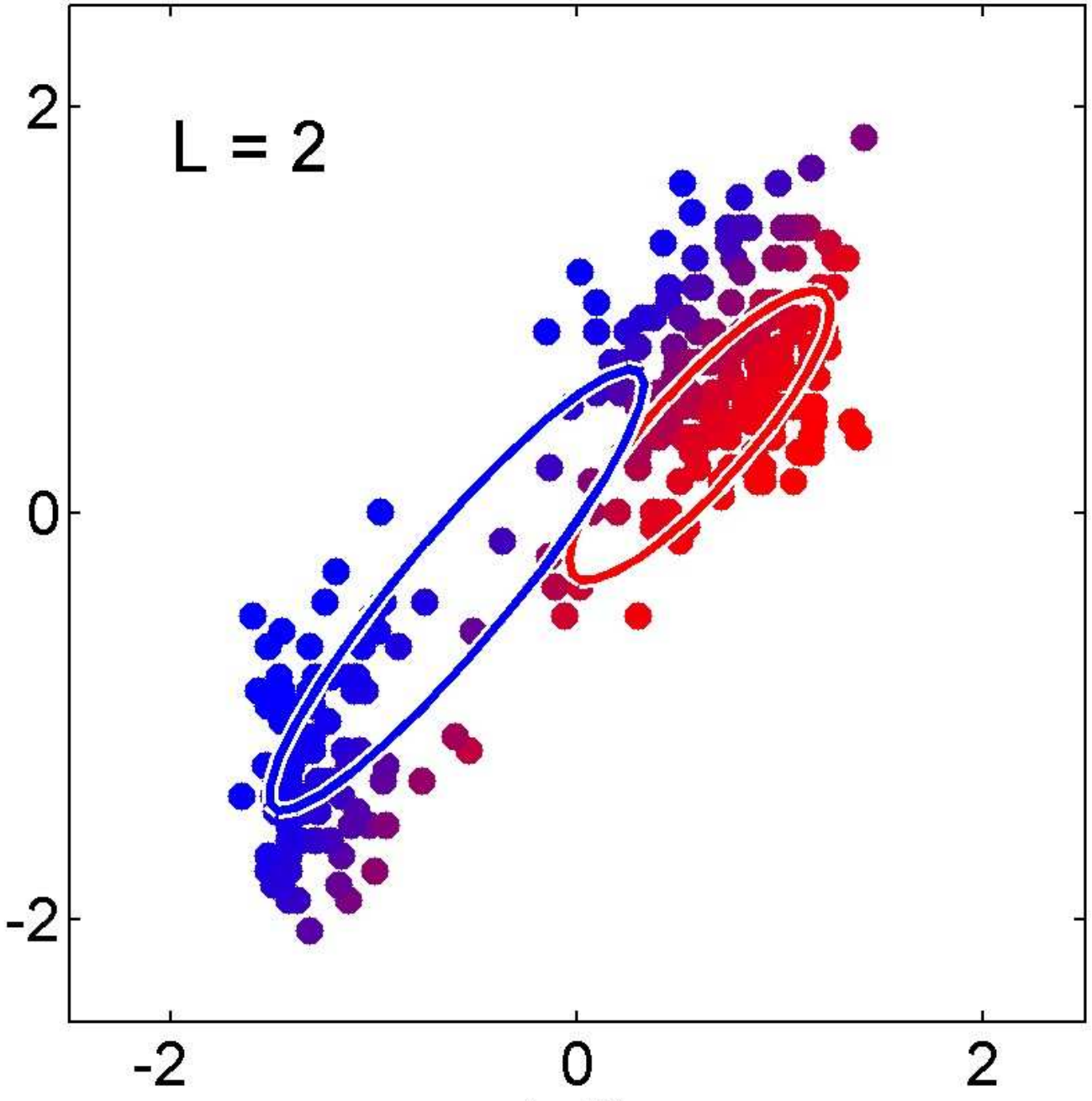
EM Algorithm – Informal Derivation

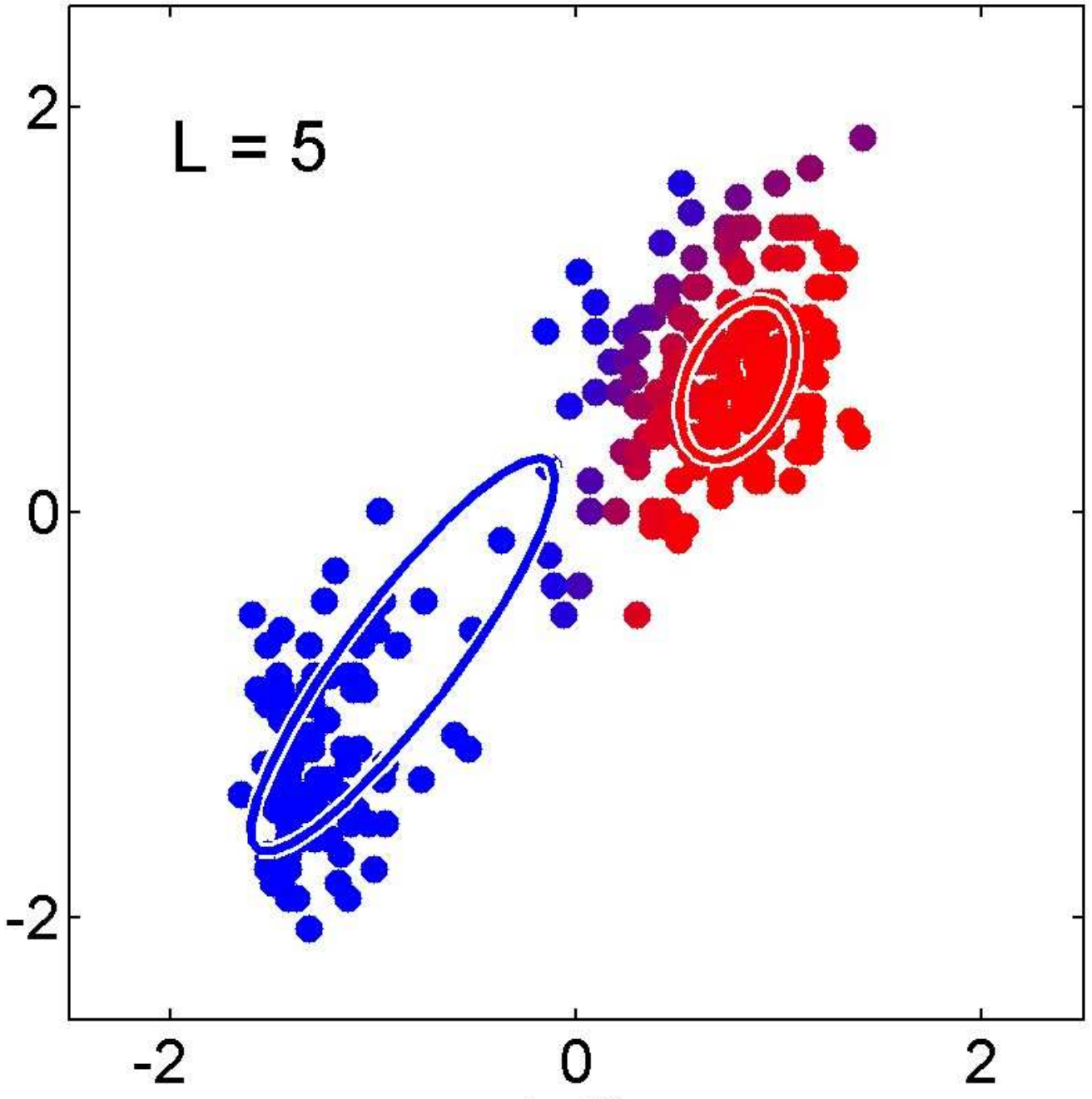
- The solutions are not closed form since they are coupled
- Suggests an iterative scheme for solving them:
  - Make initial guesses for the parameters
  - Alternate between the following two stages:
    1. E-step: evaluate responsibilities
    2. M-step: update parameters using ML results

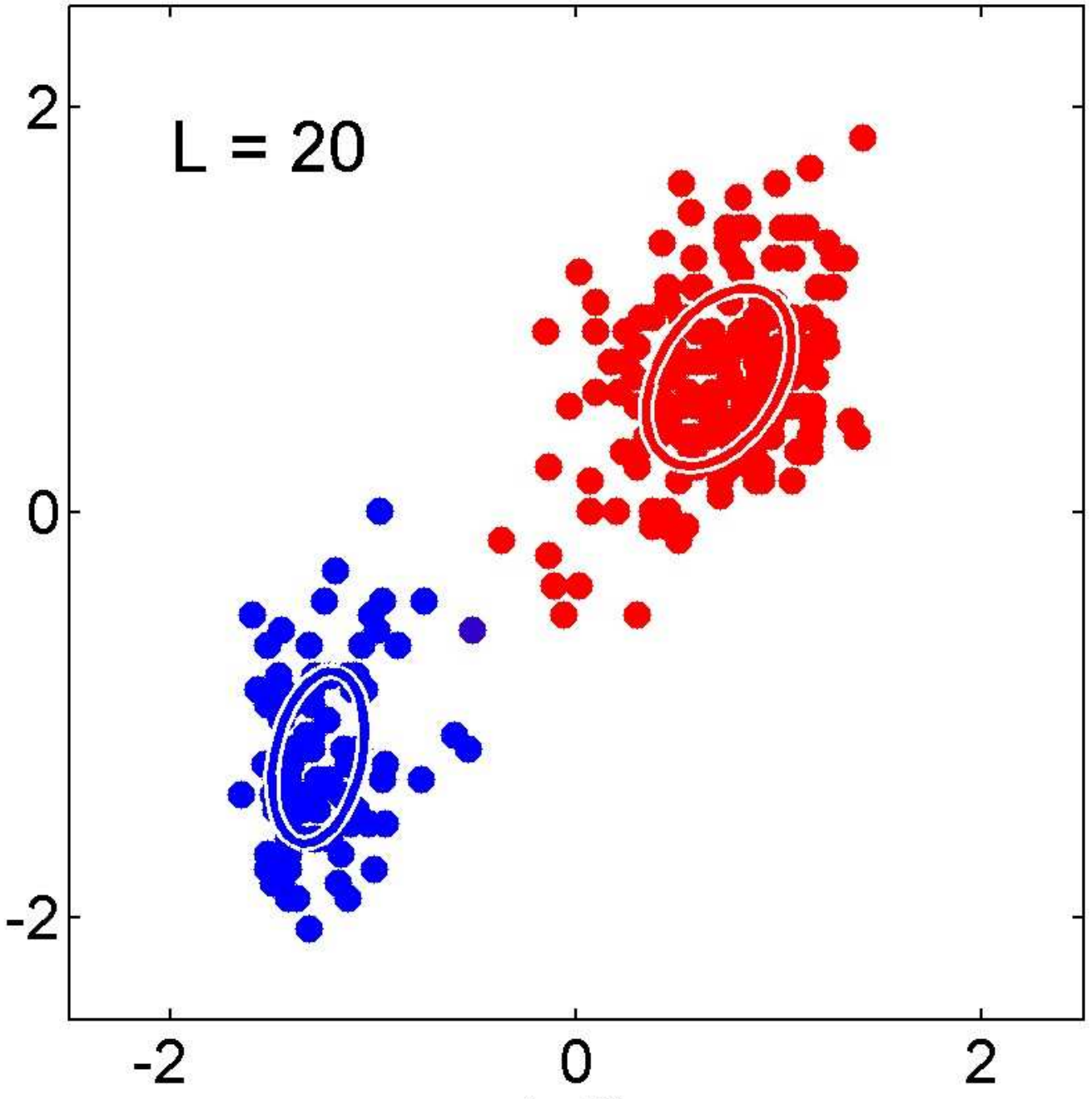

2
0
-2
-2
0
2


2
0
-2
-2
0
2


L = 1
2
0
-2
-2
0
2


L = 2
2
0
-2
-2
0
2


L = 5
2
0
-2
-2
0
2


L = 20
2
0
-2
-2
0
2

EM – Latent Variable Viewpoint

- Binary latent variables $\mathbf{z} = \{z_{kn}\}$ describing which component generated each data point
- Conditional distribution of observed variable

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

- Prior distribution of latent variables

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

- Marginalizing over the latent variables we obtain

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

Expected Value of Latent Variable

- From Bayes' theorem the posterior distribution:

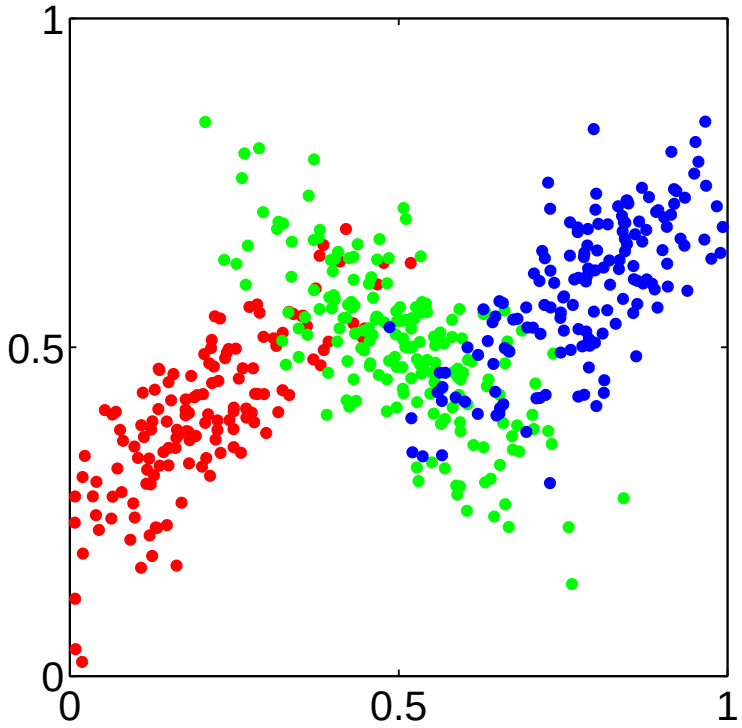
$$p(\mathrm{Z} \mid \mathrm{X}, \mu, \Sigma, \pi) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathrm{x}_n \mid \mu_k, \Sigma_k) \right]^{z_{nk}}$$

- The expectation of z_{nk} under this posterior distribution

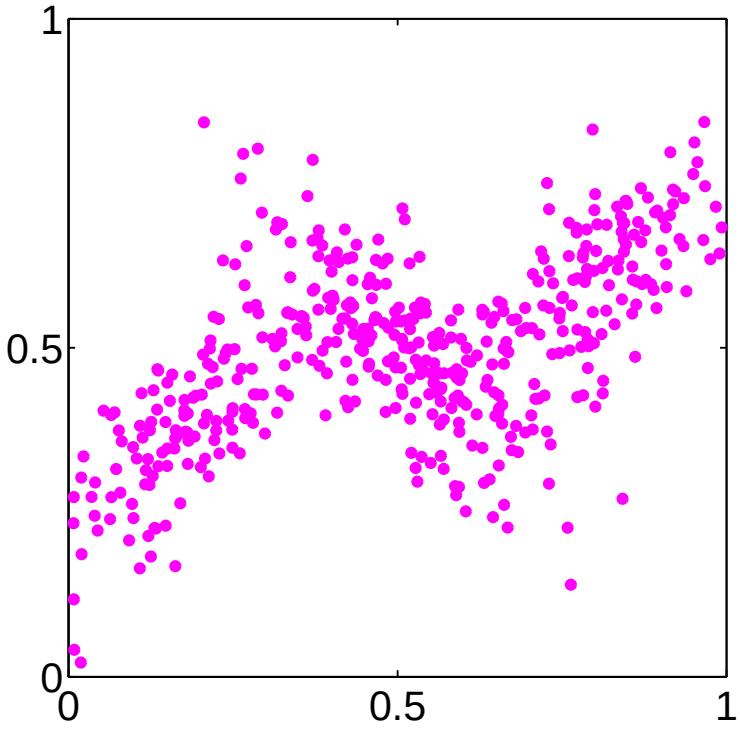
$$E[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k \mathcal{N}(\mathrm{x}_n \mid \mu_k, \Sigma_k) \right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j \mathcal{N}(\mathrm{x}_n \mid \mu_j, \Sigma_j) \right]^{z_{nj}}}$$

$$= \frac{\pi_k \mathcal{N}(\mathrm{x}_n \mid \mu_k, \Sigma_k)}{\sum_{z_{nj}} \pi_j \mathcal{N}(\mathrm{x}_n \mid \mu_j, \Sigma_j)} = \gamma(\mathrm{x}_n)$$

Complete and Incomplete Data



complete



incomplete

Latent Variable View of EM

- If we knew the values for the latent variables, we would maximize the complete-data log likelihood

$$\ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

  which gives a trivial closed-form solution (fit each component to the corresponding set of data points)
- We don't know the values of the latent variables
- However, for given parameter values we can compute the expected values of the latent variables

Expected Complete-Data Log Likelihood

- Suppose we make a guess $\boldsymbol{\theta}_{\text{old}}$ for the parameter values (means, covariances and mixing coefficients)
- Use these to evaluate the responsibilities
- Consider expected complete-data log likelihood

$$\mathrm{E}_{\mathbf{z}}[\ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})] = \sum_{n=1}^{N} \sum_{i=1}^{K} \gamma_i(\mathbf{x}_n) \{\ln \pi_i + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}$$

  where responsibilities are computed using $\boldsymbol{\theta}_{\text{old}}$
- We are implicitly 'filling in' latent variables with best guess
- Keeping the responsibilities fixed and maximizing with respect to the parameters give the previous results

EM in General

Given p(X,Z|θ) over observed variables X and latent variables Z, the goal is to maximize p(X| θ) with respect to θ

1. Choose an initial setting for parameters θ^{old}.
2. **E step:** Evaluate p(Z|X, θ^{old}).
3. **M step:** Evaluate θ^{new} given by

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

where $Q(\theta, \theta^{old}) = \sum_{Z} p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta).$

4. Check for convergence of either log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\theta^{old} \leftarrow \theta^{new}$$

and return to step 2.

K-means Algorithm

- Goal: represent a data set in terms of K clusters each of which is summarized by a prototype $\boldsymbol{\mu}_k$
- Initialize prototypes, then iterate between two phases:
  - E-step: assign each data point to nearest prototype
  - M-step: update prototypes to be the cluster means

Responsibilities

- *Responsibilities* assign data points to clusters

$$r_{nk} \in \{0, 1\}$$

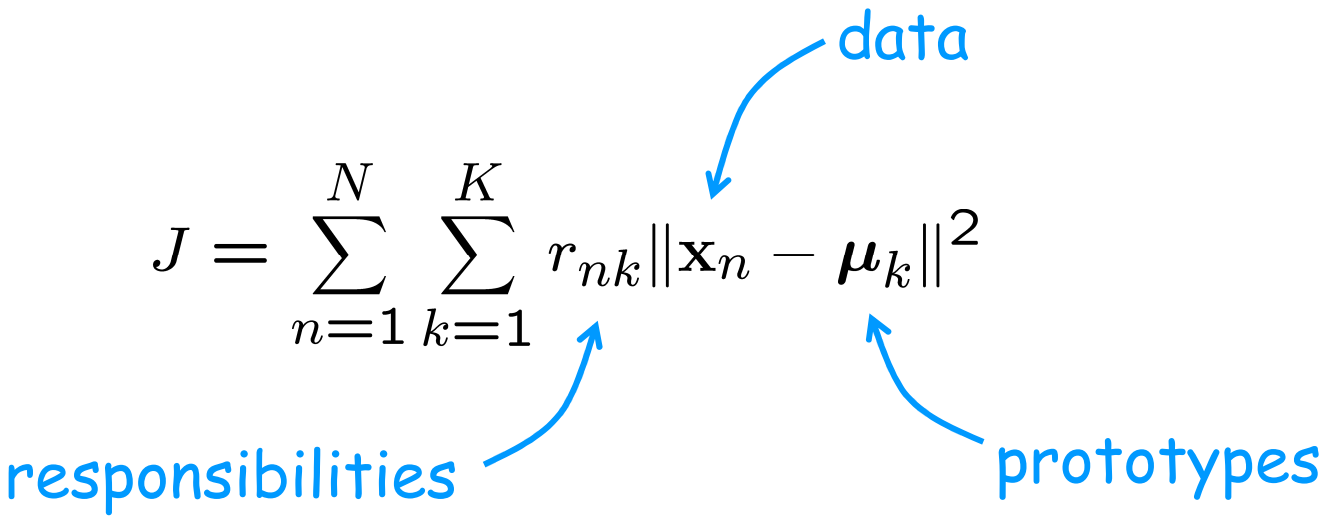
such that

$$\sum_k r_{nk} = 1$$

- Example: 5 data points and 3 clusters

$$(r_{nk}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

K-means Cost Function



Minimizing the Cost Function

- E-step: minimize J w.r.t. r_{nk}
  - assigns each data point to nearest prototype
- M-step: minimize J w.r.t $\boldsymbol{\mu}_k$
  - gives

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{kn}\mathbf{x}_n}{\sum_n r_{kn}}$$

  - each prototype set to the mean of points in that cluster
- Convergence guaranteed since there is a finite number of possible settings for the responsibilities

K-means Revisited

- Consider GMM with common covariances $\boldsymbol{\Sigma}_k = \epsilon \mathbf{I}$
- Take limit $\epsilon \rightarrow 0$
- Responsibilities become binary

$$\gamma_i(\mathbf{x}_n) = \frac{\pi_i \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_i\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}} \rightarrow r_{ni} \in \{0, 1\}$$

- Expected complete-data log likelihood becomes

$$\mathrm{E}_{\mathbf{Z}}[L_{\mathrm{C}}] \rightarrow -\frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{K} r_{ni} \|\mathbf{x}_n - \boldsymbol{\mu}_i\|^2 + \text{const.}$$