EE496 : COMPUTATIONAL INTELLIGENCE
PA02: EXPECTATION MAXIMIZATION

UGUR HALICI
METU: Department of Electrical and Electronics Engineering (EEE)
METU-Hacettepe U: Neuroscience and Neurotechnology (NSNT)

Preliminaries

- We assume that the dataset *X* has been generated by a *parametric* distribution *p(X)*.
- Estimation of the parameters of *p* is known as *density estimation*.
- We consider Gaussian distribution.

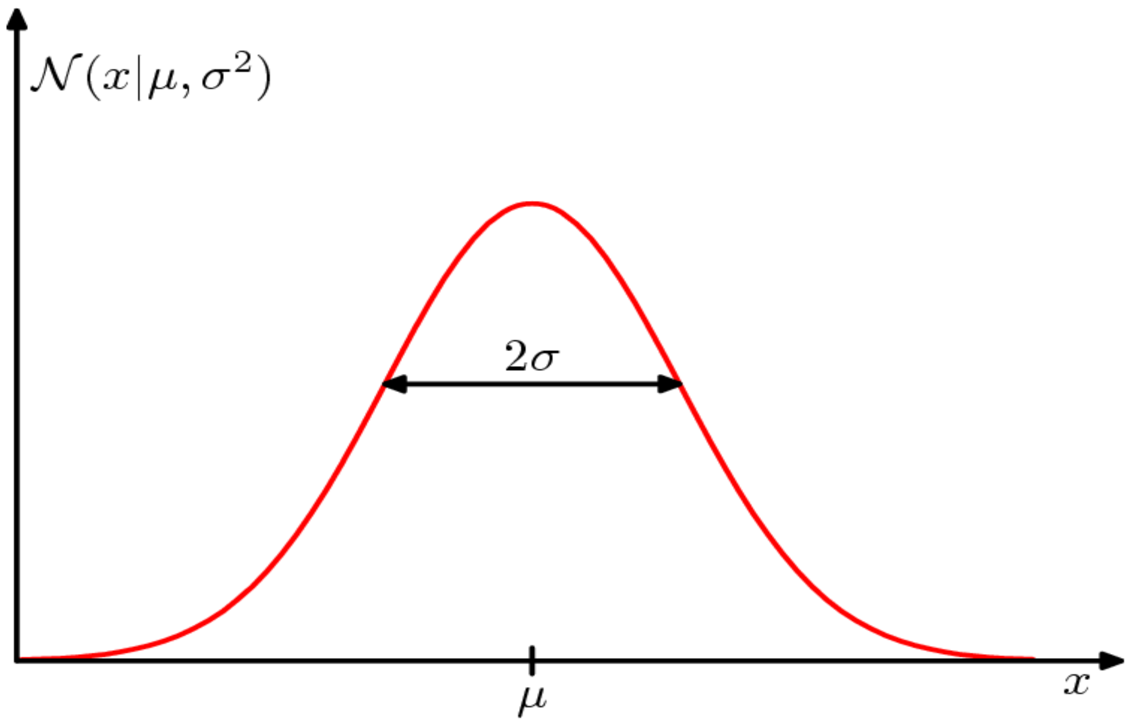
Figures taken from:
http://research.microsoft.com/~cmbishop/PRML/

Typical parameters

- *Mean* (μ): average value of $p(X)$, also called expectation.
- *Variance* (σ): provides a measure of variability in $p(X)$ around the mean.
- *Covariance*: measures how much two variables vary together.
- *Covariance matrix*: collection of covariances between all dimensions.
  - Diagonal of the covariance matrix contains the variances of each attribute.

One-dimensional Gaussian

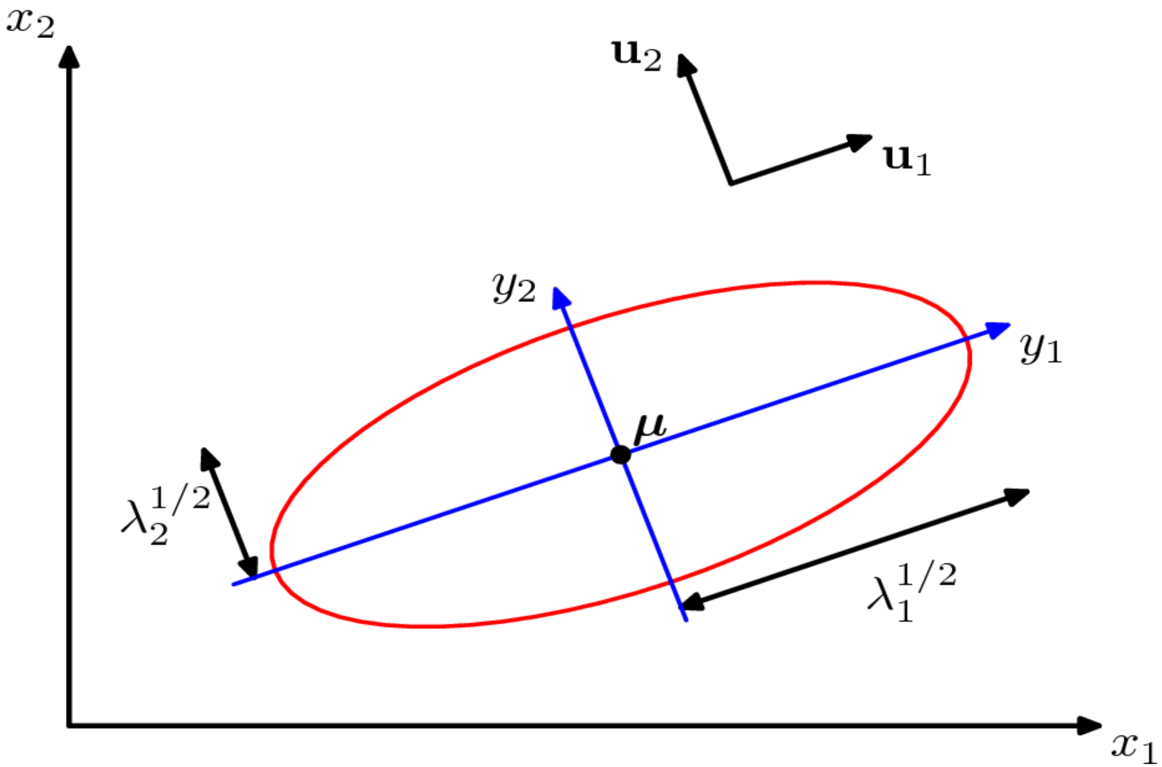
$$\text{Normal}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (x-\mu)^2 \right\}$$



- Parameters to be estimated are the mean (μ) and variance (σ)

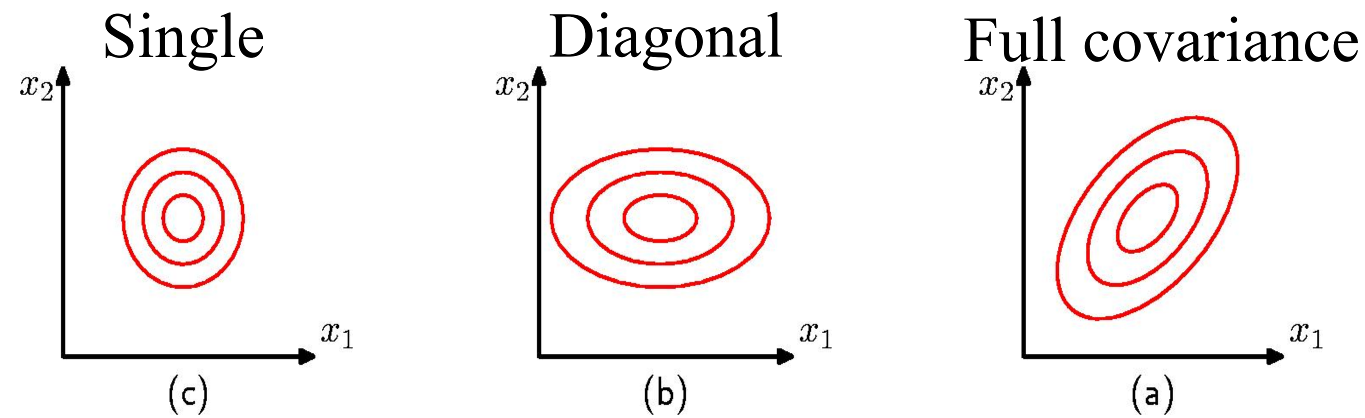
Multivariate Gaussian (1)

$$\text{Normal}(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^2} \frac{1}{\det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu\right)^T \Sigma \left(\mathbf{x}-\mu\right)\right\}$$



- In multivariate case we have covariance matrix instead of variance

Multivariate Gaussian (2)



$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1{}^2 & 0 \\ 0 & \sigma_2{}^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_{11}{}^2 & \sigma_{12}{}^2 \\ \sigma_{12}{}^2 & \sigma_{22}{}^2 \end{pmatrix}$$

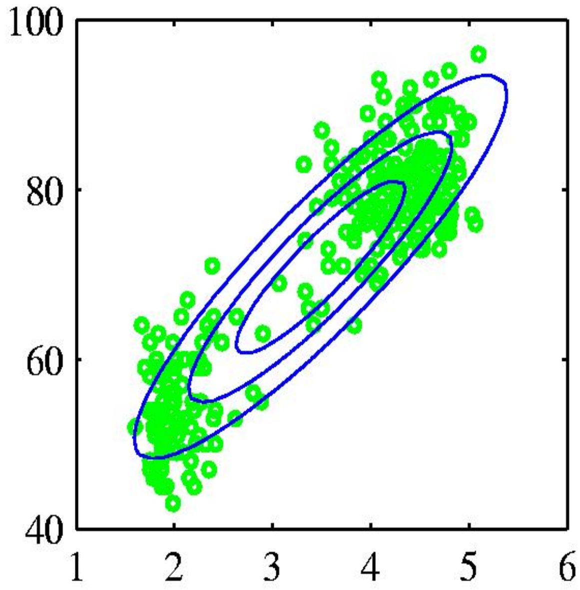
Complete data log likelihood:

$$\ln p(X) = \ln \prod_{n=1}^{N} \mathrm{Normal}(\mathbf{x}_n \mid \mu, \Sigma)$$

Maximum Likelihood (ML) parameter estimation

- Maximize the log likelihood formulation
- Setting the gradient of the complete data log likelihood to zero we can find the closed form solution.
  - Which in the case of mean, is the sample average.

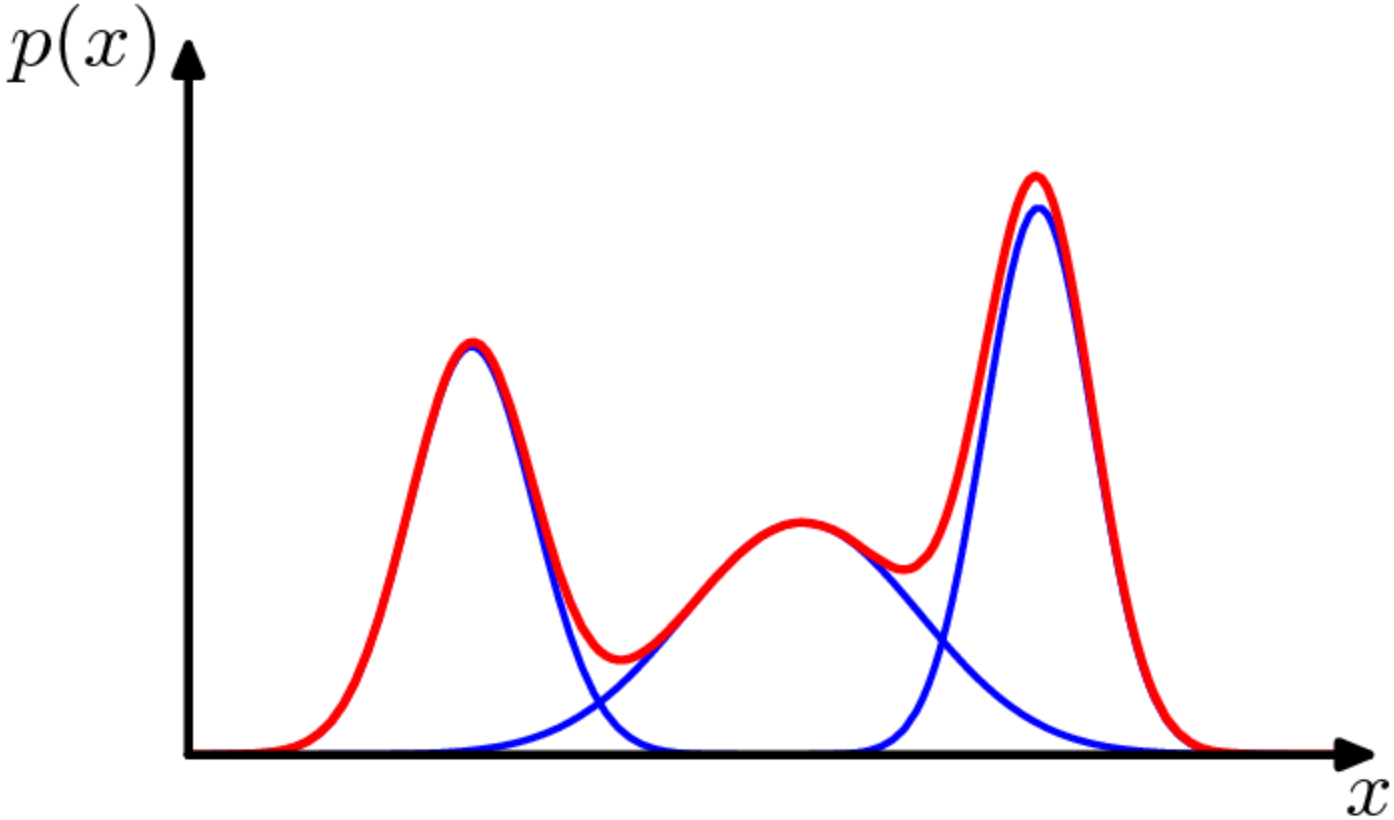
When one Gaussian is not enough



- Real world datasets are rarely unimodal!

Mixtures of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^{M} \pi_k \mathrm{Normal}(\mathbf{x} \mid \mu_k, \Sigma_k)$$



Mixtures of Gaussians

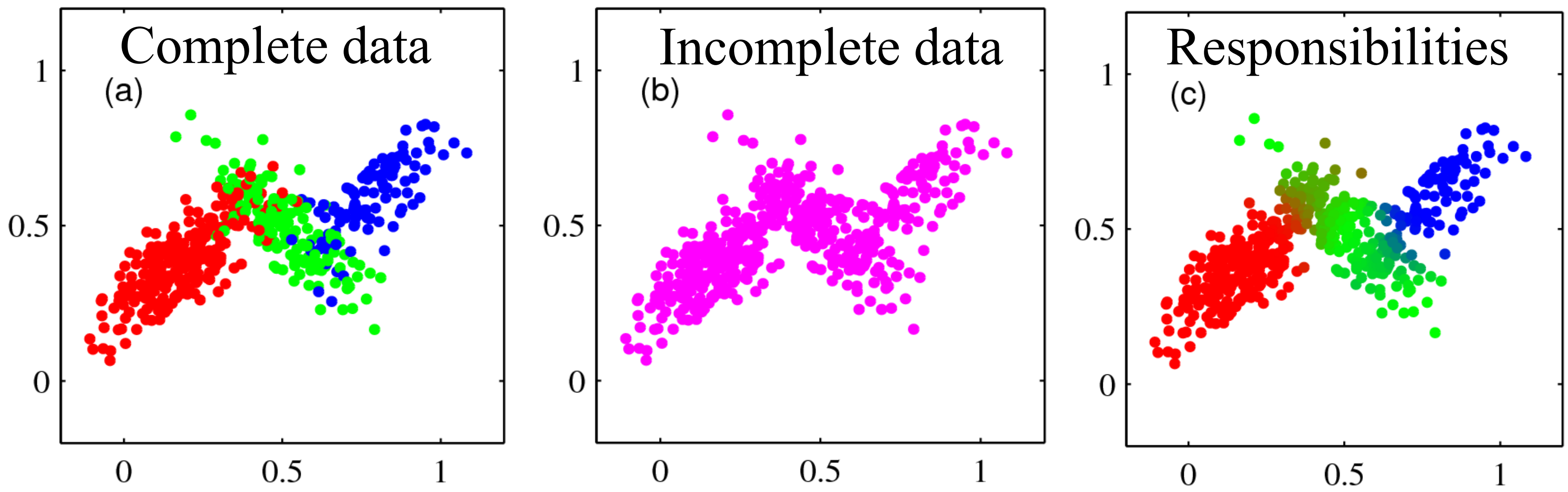
- In addition to mean and covariance parameters (now *M* times), we have mixing coefficients π_k.

Following properties hold for the mixing coefficients:

$$\sum_{k=1}^{M} \pi_k = 1 \qquad 0 \leq \pi_k \leq 1$$

It can be seen as the prior probability of the component *k*

Responsibilities (1)



- Component labels (red, green and blue) cannot be observed.
- We have to calculate approximations (responsibilities).

Responsibilities (2)

- Responsibility describes, how probably observation vector $\boldsymbol{x}$ is from component k.
- In clustering, responsibilities take values 0 and 1, and thus, it defines the hard partitioning.

Responsibilities (3)

We can express the marginal density *p*(*x*) as:

$$p(\mathbf{x}) = \sum_{k=1}^{M} p(k)\, p(\mathbf{x} \mid k)$$

From this, we can find the responsibility of the k[th] component of *x* using Bayesian theorem:

$$\gamma_k(\mathbf{x}) = p(k \mid \mathbf{x})$$

$$= \frac{p(\mathbf{x})\, p(\mathbf{x} \mid k)}{\sum_l p(l)\, p(\mathbf{x} \mid l)}$$

$$= \frac{\pi_k \mathrm{Normal}(\mathbf{x} \mid \mu_k, \Sigma_k)}{\sum_l \pi_l \mathrm{Normal}(\mathbf{x} \mid \mu_l, \Sigma_l)}$$

Expectation Maximization (EM)

- Goal: Maximize the log likelihood of the whole data

$$\ln p(\mathbf{X} \mid \Pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{M} \pi_k \mathrm{Normal}(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right\}$$

- When responsibilities are calculated, we can maximize individually for the means, covariances and the mixing coefficients!

Exact update equations

New mean estimates:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)\mathbf{x}_n \qquad N_k = \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)$$

Covariance estimates

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T$$

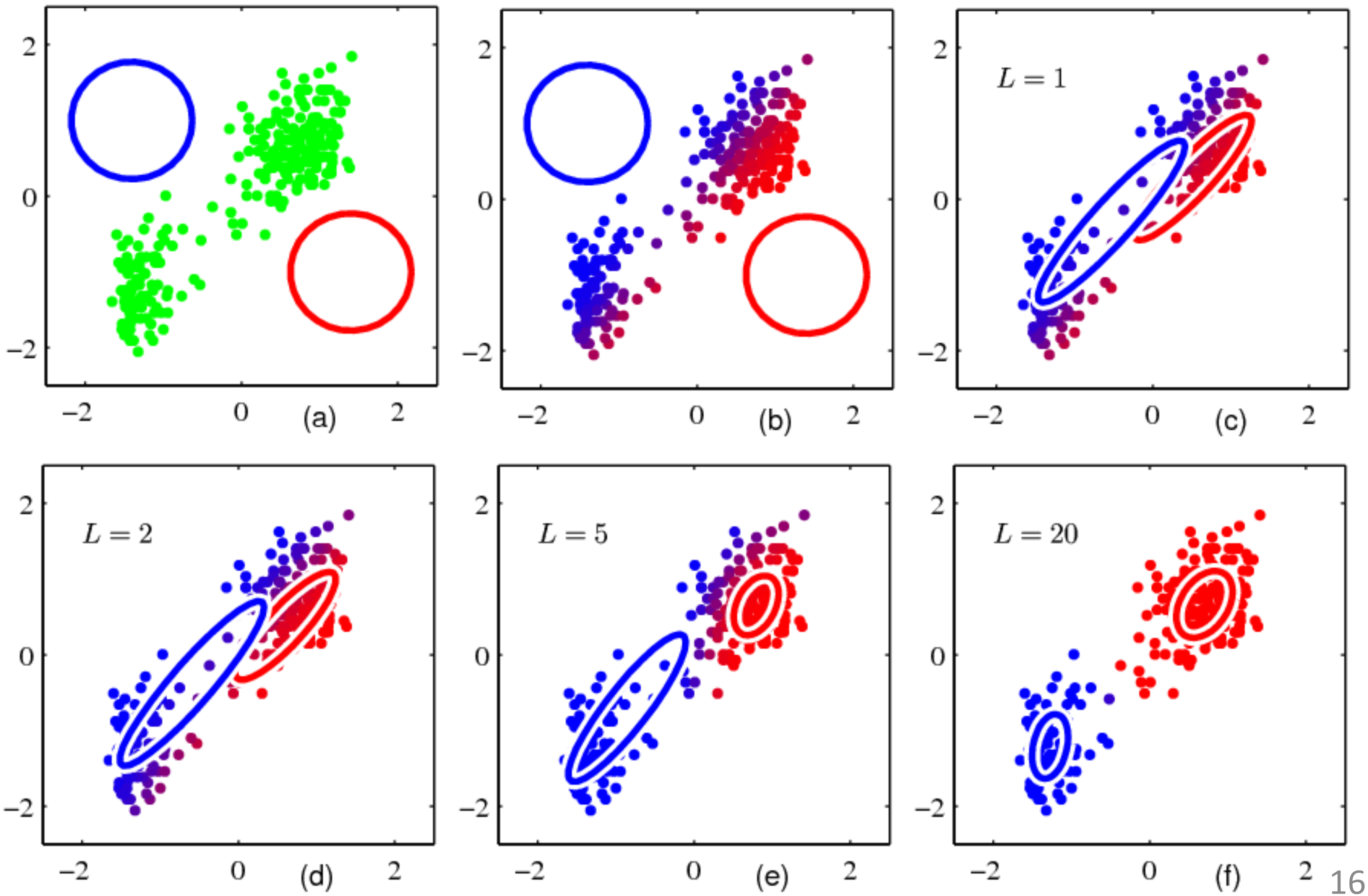
Mixing coefficient estimates

$$\pi_k = \frac{N_k}{N}$$

EM Algorithm

- Initialize parameters
- while not converged
  - **E step:** Calculate responsibilities.
  - **M step:** Estimate new parameters
  - Calculate log likelihood of the new parameters

Example of EM



Computational complexity

- Hard clustering with MSE criterion is NP-complete.
- Can we find optimal GMM in polynomial time?
- Finding optimal GMM is in class NP

Some insights

- In GMM we need to estimate the parameters, which all are real numbers
  - Number of parameters:

    M+M(D) + M(D(D-1)/2)
- Hard clustering has no parameters, just set partitioning (remember optimality criteria!)

Some further insights

- Both optimization functions are mathematically rigorous!
- Solutions minimizing MSE are always meaningful
- Maximization of log likelihood might lead to singularity!

Example of singularity