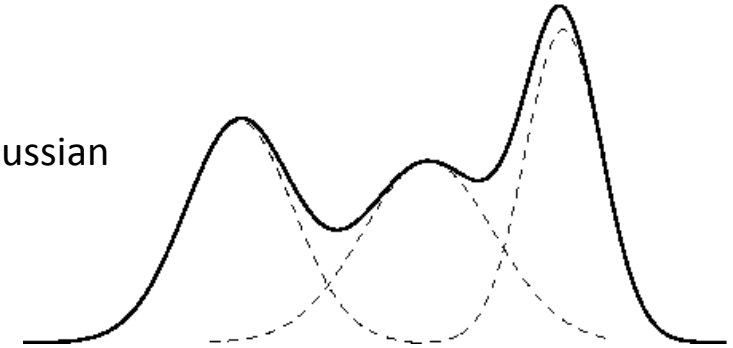
Statistics background:
Maximum Likelihood and Expectation Maximization

Mixture of Gaussians

- A simple linear superposition of Gaussian components
- Provides a richer class of density models than the single Gaussian
- GMM are formulated in terms of discrete latent variables
  - provide a deeper insight
  - motivate the EM algorithm

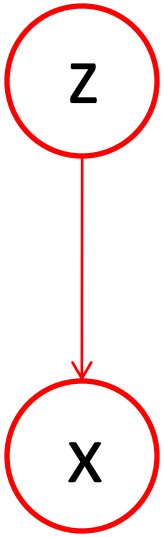


$$p(x) = \sum_k \pi_k N(x \mid \mu_k, \Sigma_k)$$

$$\sum_k \pi_k = 1 \text{ and } 0 \le \pi_k \le 1$$

- Let's introduce variable z with a 1-of-K representation.
- The joint distribution of x and z is:

$$p(x, z) = p(x \mid z)\, p(z)$$



Conditional and marginal distributions

The conditional distribution p(x | z) is:

$$p(x \mid z_k = 1) = N(x \mid \mu_k, \Sigma_k)$$

Which can be written int the more general form:

$$p(x \mid z) = \prod_k N(x \mid \mu_k, \Sigma_k)^{z_k}$$

Since

$$p(z) = \prod_k \pi_k^{z_k}$$

the marginal distribution of x p(x) can be obtained by summing over all values of z:

$$p(x) = \sum_z p(x \mid z)\, p(z) = \sum_k \pi_k N(x \mid \mu_k, \Sigma_k)$$

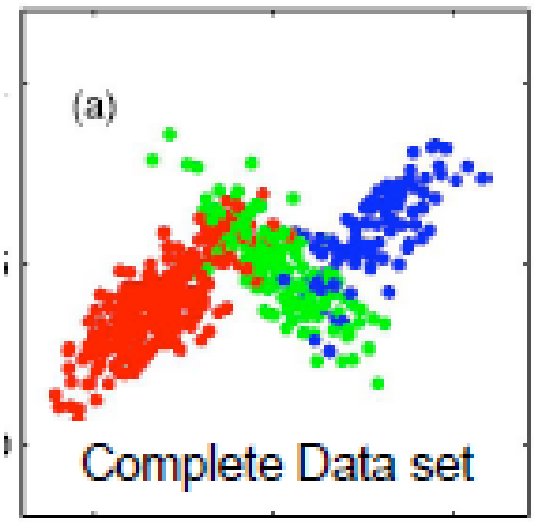
Responsibilities (another conditional distribution)

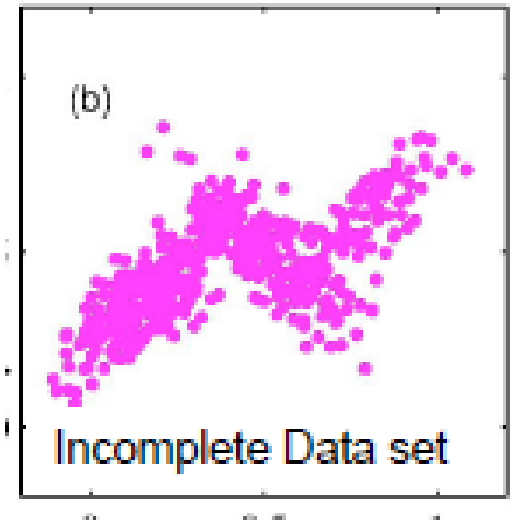
The conditional distribution p(z |x) can be obtained by Bayes rule:

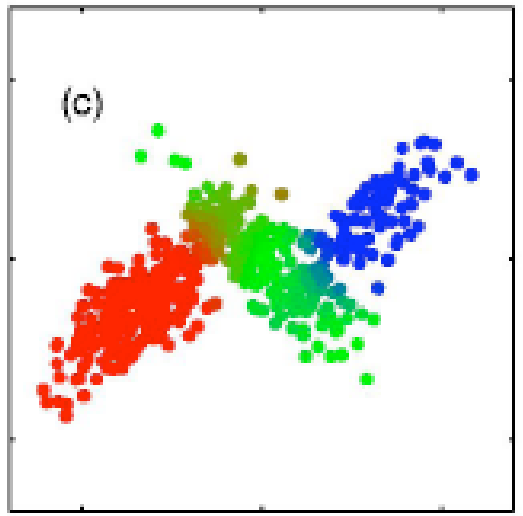
$$p(z_k=1|\,x)=\frac{p(z_k=1)\,p(x\,|\,z_k=1)}{\sum_j p(z_j=1)\,p(x\,|\,z_j=1)}=\frac{\pi_k N(x\,|\,\mu_k,\Sigma_k)}{\sum_j \pi_j N(x\,|\,\mu_j,\Sigma_j)}=\gamma(z_k)$$

These are often called responsibilities, i.e. how responsible is component k of the mixture for the generation of sample x.

Mixture of 3 Gaussians: a) x and z known, b) x known, c) responsibilities plotted.







MLE for Mixture of Gaussians

The mixture density function is:

$$p(x) = \sum_k \pi_k N(x \mid \mu_k, \Sigma_k)$$

The likelihood of data **X** is therfore:

$$p(X \mid \pi, \mu, \Sigma) = \prod_n \sum_k \pi_k N(x_n \mid \mu_k, \Sigma_k)$$

Taking the log-likelihood:

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_n \ln \left\{ \sum_k \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}$$

which we wish to maximize. Optimization is difficult because of the summation in the log term.

EM for mixture of Gaussians

Begin with the computation of the log likelihood (suppose to have mixture parameters).

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_n \ln \left\{ \sum_k \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}$$

Then take partial derivatives w.r.t μ, Σ, π.

$$0 = \frac{\partial L}{\partial \mu} = \sum_n \frac{\pi_k N(x \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(x \mid \mu_j, \Sigma_j)} \Sigma_k^{-1} (x_n - \mu_k)$$

$$\mu_k = \frac{1}{N_k} \sum_n \gamma_{nk} x_n, N_k = \sum_n \gamma_{nk}$$

Where N_k is the number of point in cluster k.

EM for mixture of Gaussians

Now differentiate :

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_n \ln \left\{ \sum_k \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}$$

w.r.t Σ and set it to zero:

$$\Sigma_k = \frac{1}{N_k} \sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)'$$

EM for mixture of Gaussians

Finally optimize w.r.t mixing coefficients using Lagrange multipliers (mix. coeff. < 1)

$$\ln p(X \mid \pi, \mu, \Sigma) + \lambda\left(\sum_k \pi_k - 1\right)$$

Since the sum of mix. coeff is 1 we find λ = -N and therefore:

$$0 = \sum_n \frac{N(x \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(x \mid \mu_j, \Sigma_j)} + \lambda$$

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

EM algorithm:

Initialize μ,Σ,π

E-step:

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n \mid \mu_j, \Sigma_j)}$$

M-step:

$$\mu_k^{new} = \frac{1}{N_k} \sum_n \gamma_{nk} x_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_n \gamma_{nk} (x_n - \mu_k^{new})(x_n - \mu_k^{new})'$$

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

EM algorithm convergence

When do we stop?

After each iteration *it* of E and M steps recompute the log likelihood:

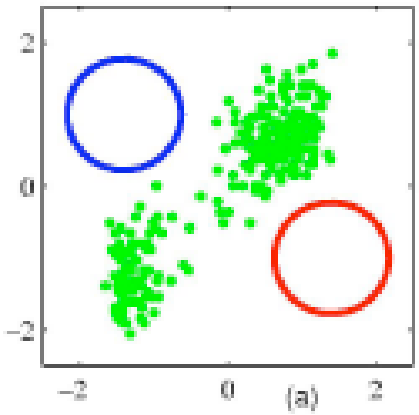
$$\ln p(X \mid \pi^{it}, \mu^{it}, \Sigma^{it}) = \sum_n \ln \left\{ \sum_k \pi_k^{it} N(x_n \mid \mu_k^{it}, \Sigma_k^{it}) \right\}$$

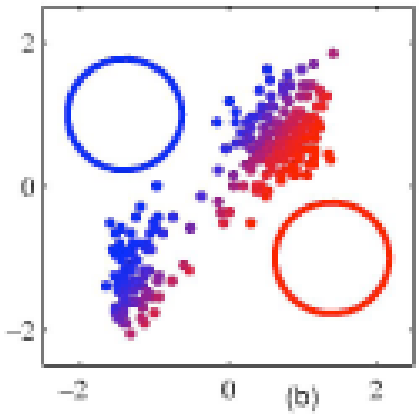
A simple convergence test:

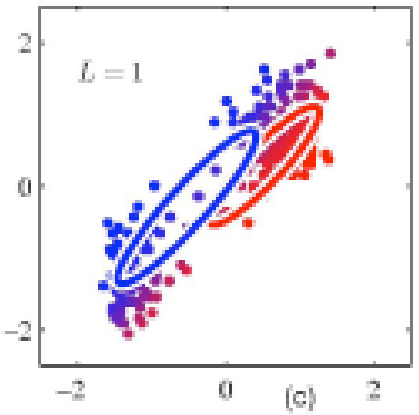
$$\left| \ln p(X \mid \pi^{it}, \mu^{it}, \Sigma^{it}) - \ln p(X \mid \pi^{it-1}, \mu^{it-1}, \Sigma^{it-1}) \right| < \varepsilon$$

Stop when log likelihood is not increasing anymore (e.g. ε = 1e-5).

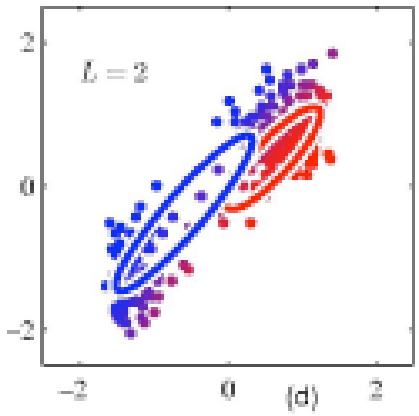
EM on the Old Faithful geyser dataset (k=2), x=[eruption duration, time to eruption]



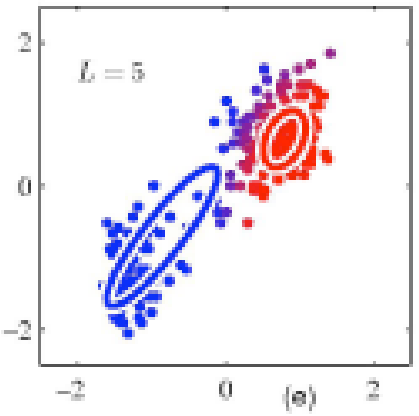




After 2 cycles



After 5 cycles



After 20 cycles