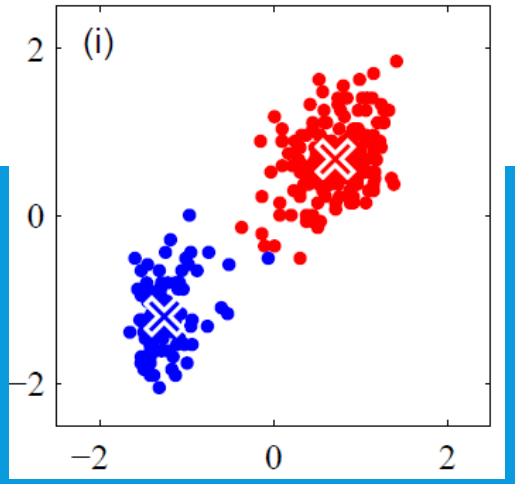
MIXTURE MODELS AND EM

Kyu Cho

MIXTURE MODELS

Hard Clustering : clusters do not overlap

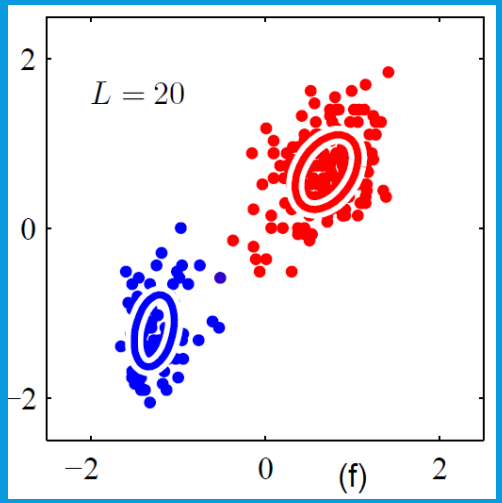
ex) cluster 1 or 2



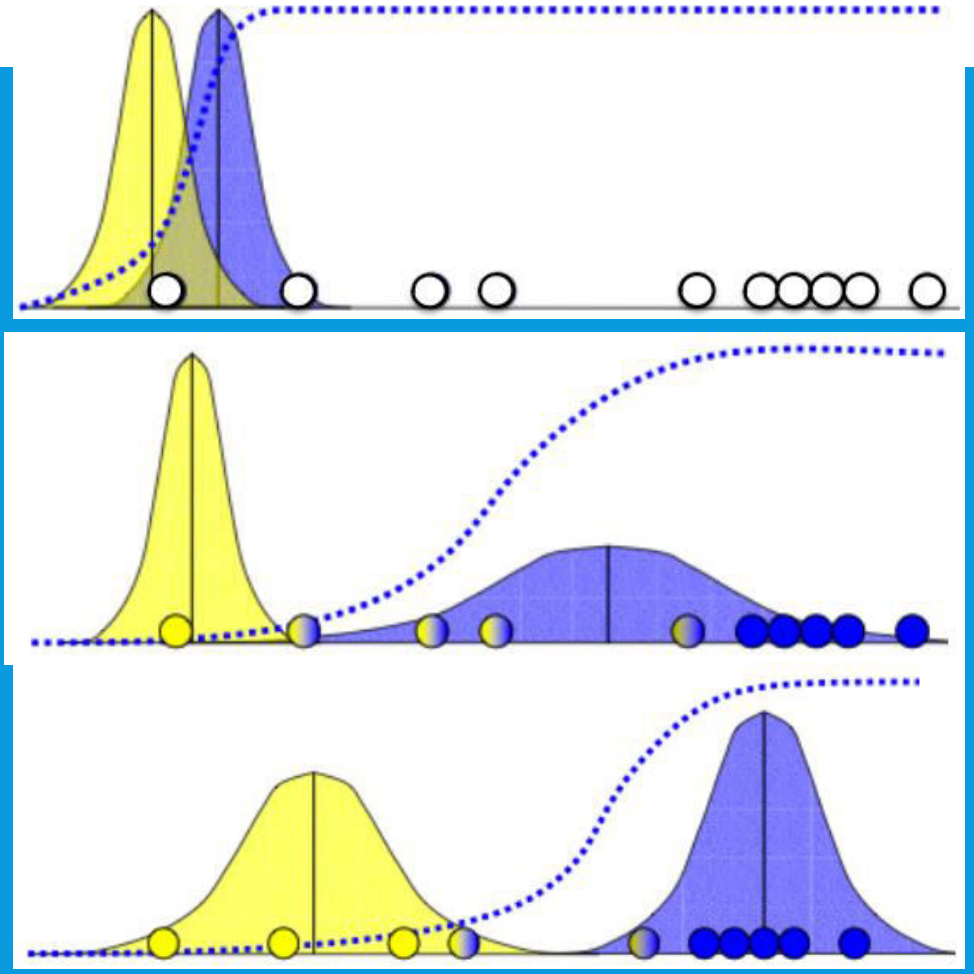
Soft Clustering : clusters may overlap

Each cluster is presented as probability distribution.

ex) 60% of cluster 1, 40% of cluster 2



MIXTURE MODEL IN 1-D



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \ldots + b_n x_{n_b}}{b_1 + b_2 + \ldots + b_n}$$

$$\sigma_b^2 = \frac{b_1 (x_1 - \mu_1)^2 + \ldots + b_n (x_n - \mu_n)^2}{b_1 + b_2 + \ldots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \ldots + a_n x_{n_b}}{a_1 + a_2 + \ldots + a_n}$$

$$\sigma_a^2 = \frac{a_1 (x_1 - \mu_1)^2 + \ldots + a_n (x_n - \mu_n)^2}{a_1 + a_2 + \ldots + a_n}$$

HOW TO PICK K?

Probabilistic model

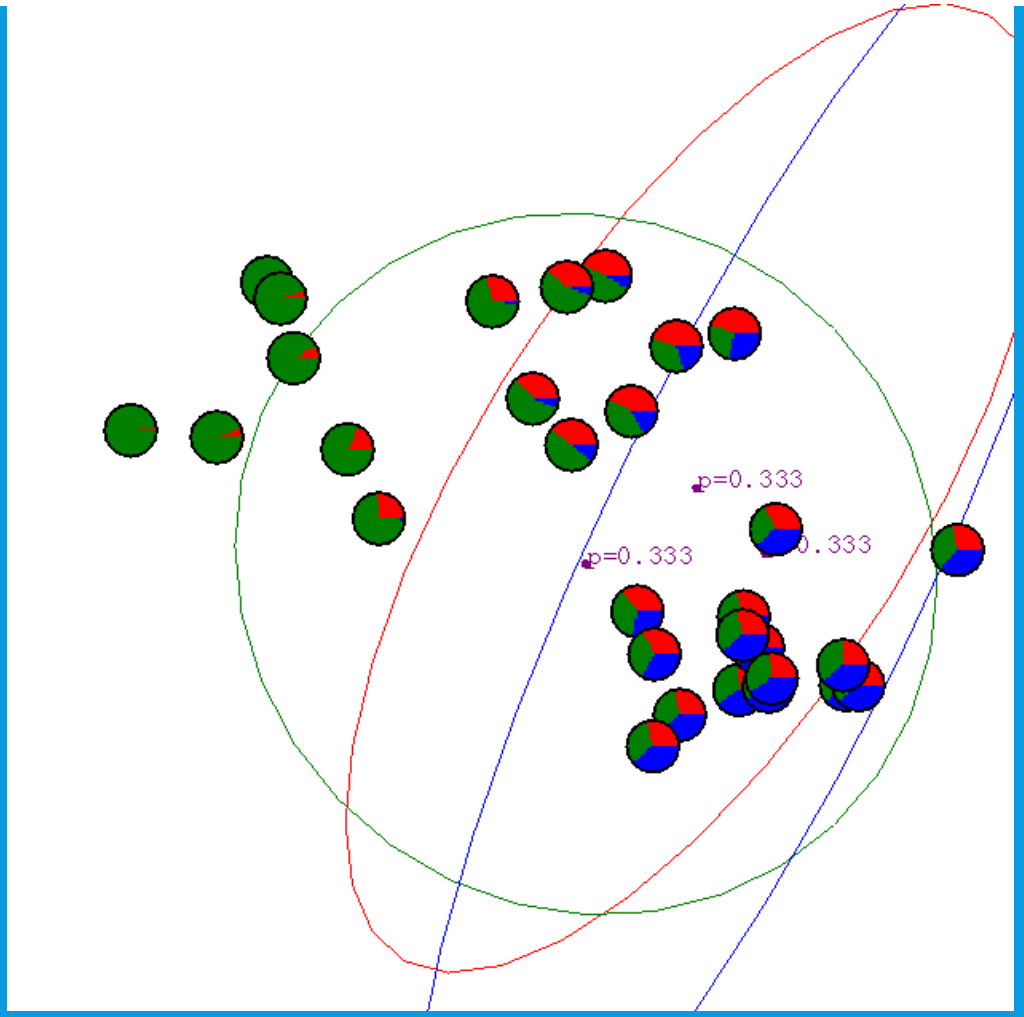
$$L = \log P(x_1 \ldots x_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(x_i \mid k) P(k)$$

- tries to “fit” the data (maximize likelihood)

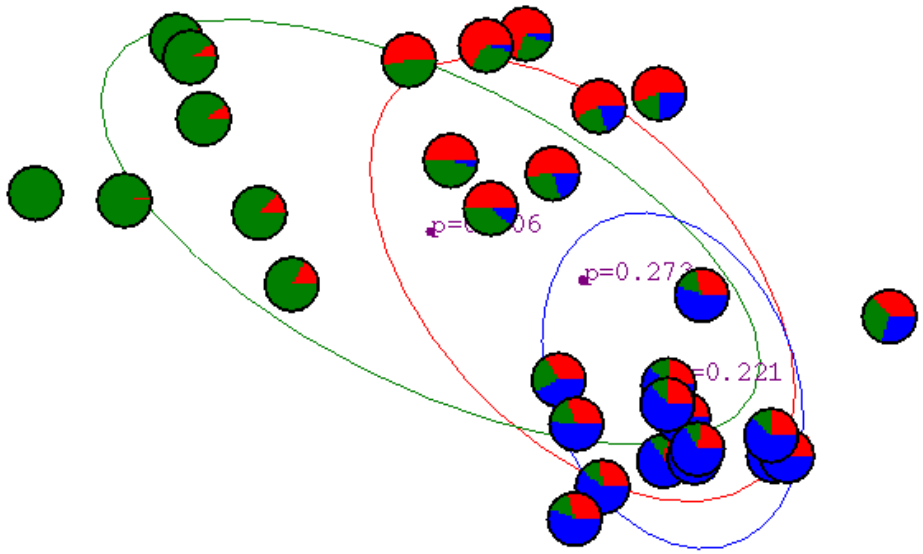
Pick K that makes L as large as possible

Set threshold to stop the iteration

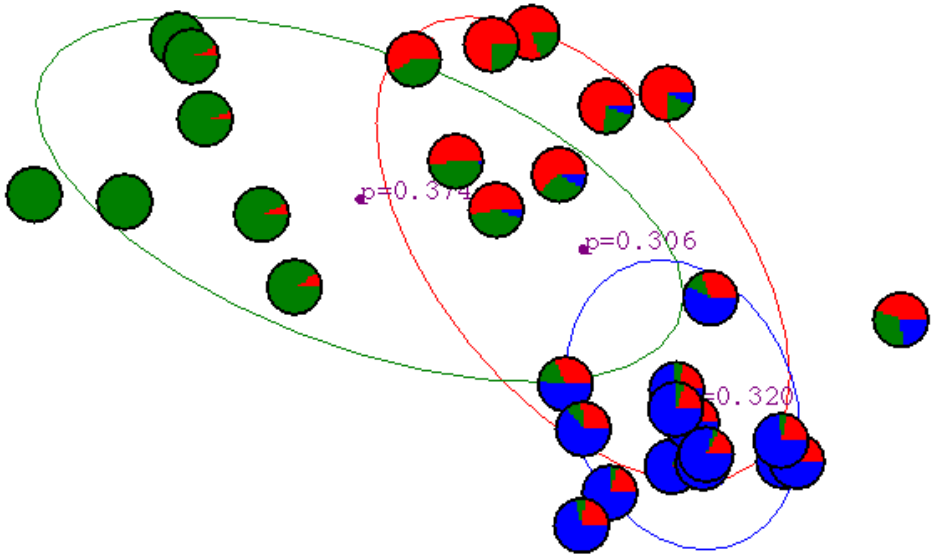
GAUSSIAN MIXTURE EXAMPLE: START



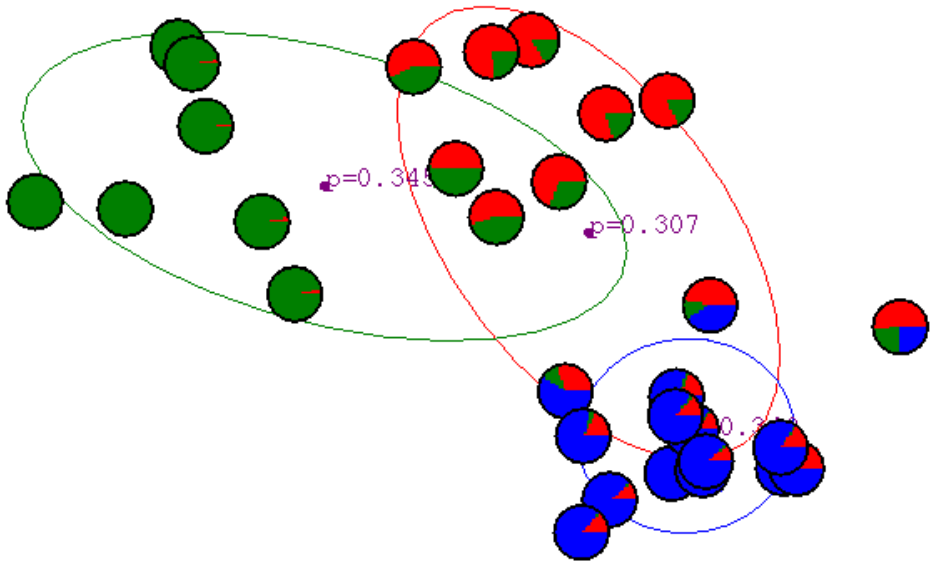
1ST ITERATION



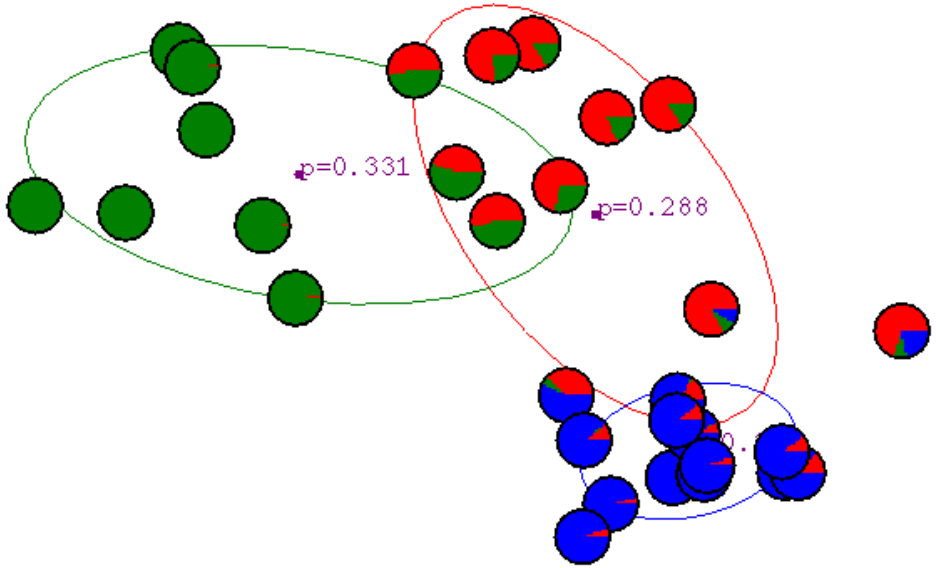
2ND ITERATION



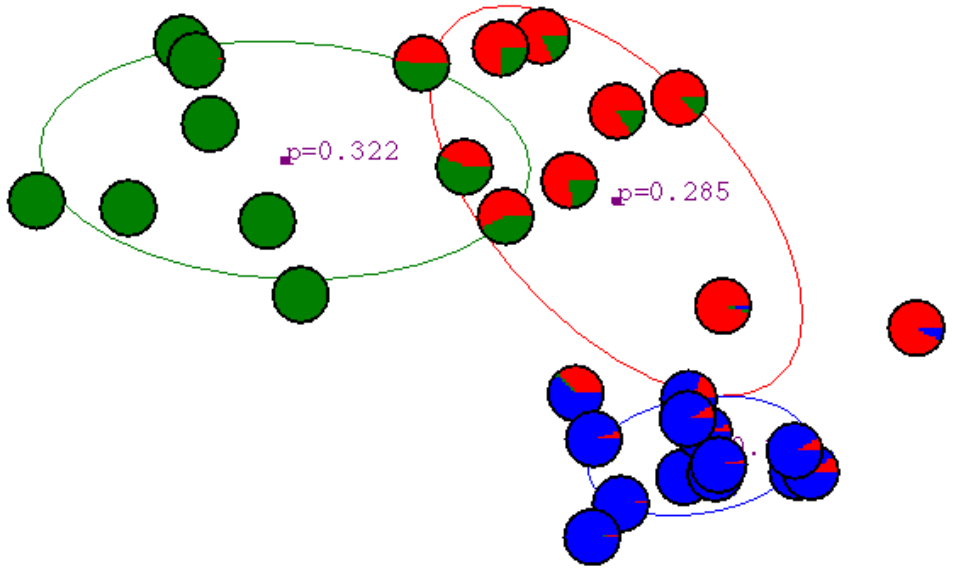
3RD ITERATION



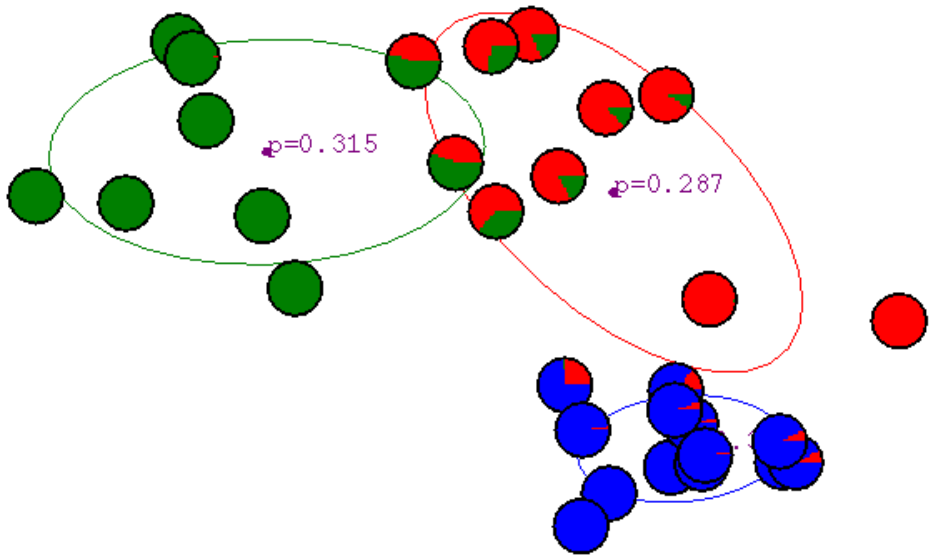
4TH ITERATION



5TH ITERATION



6TH ITERATION



20TH ITERATION