Gaussian mixture models

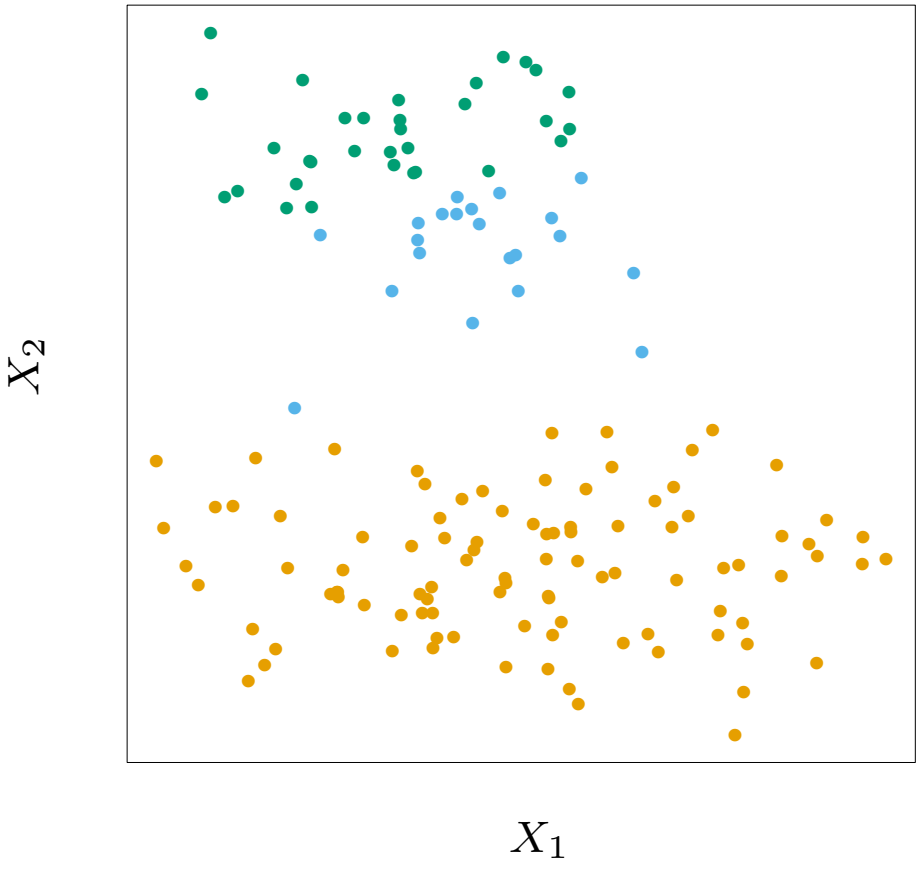
These are like kernel density estimates, but with a small number of components (rather than one component per data point)

Outline

- k-means clustering
- a soft version of k-means: EM algorithm for Gaussian mixture model
- EM algorithm for general missing data problems

K-means clustering

See  pp 461.



Simulated data in the plane, clustered into three classes (represented by red, blue and green) by the K-means clustering algorithm

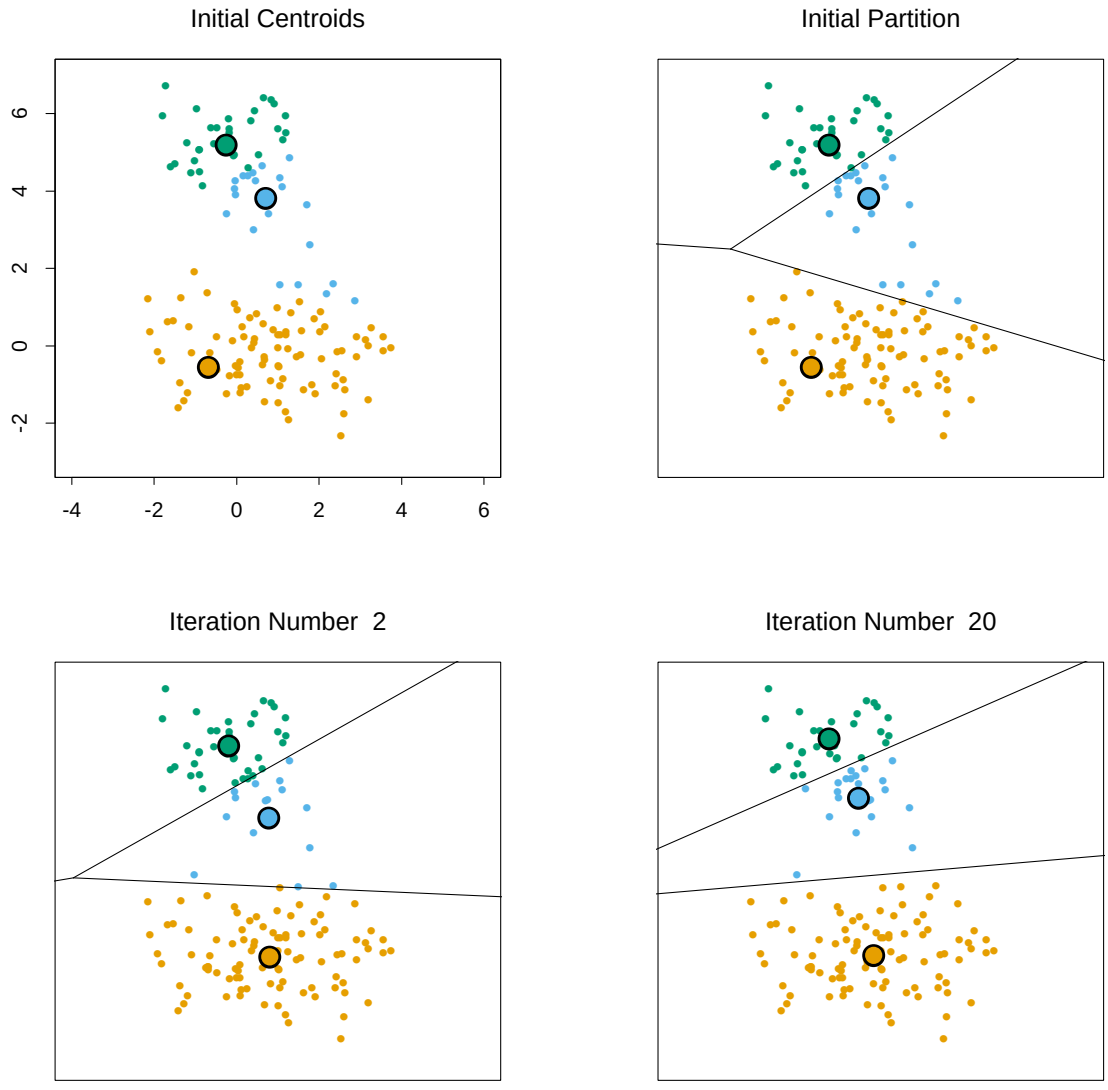
K-means algorithm

(1) For each data point, the closest cluster center (in Euclidean distance) is identified;

(2) Each cluster center is replaced by the coordinate-wise average of all data points that are closest to it.

- Steps 1 and 2 are alternated until convergence. Algorithm converges to a local minimum of the within-cluster sum of squares.
- Typically one uses multiple runs from random starting guesses, and chooses the solution with lowest within cluster sum of squares.

Kmeans in action

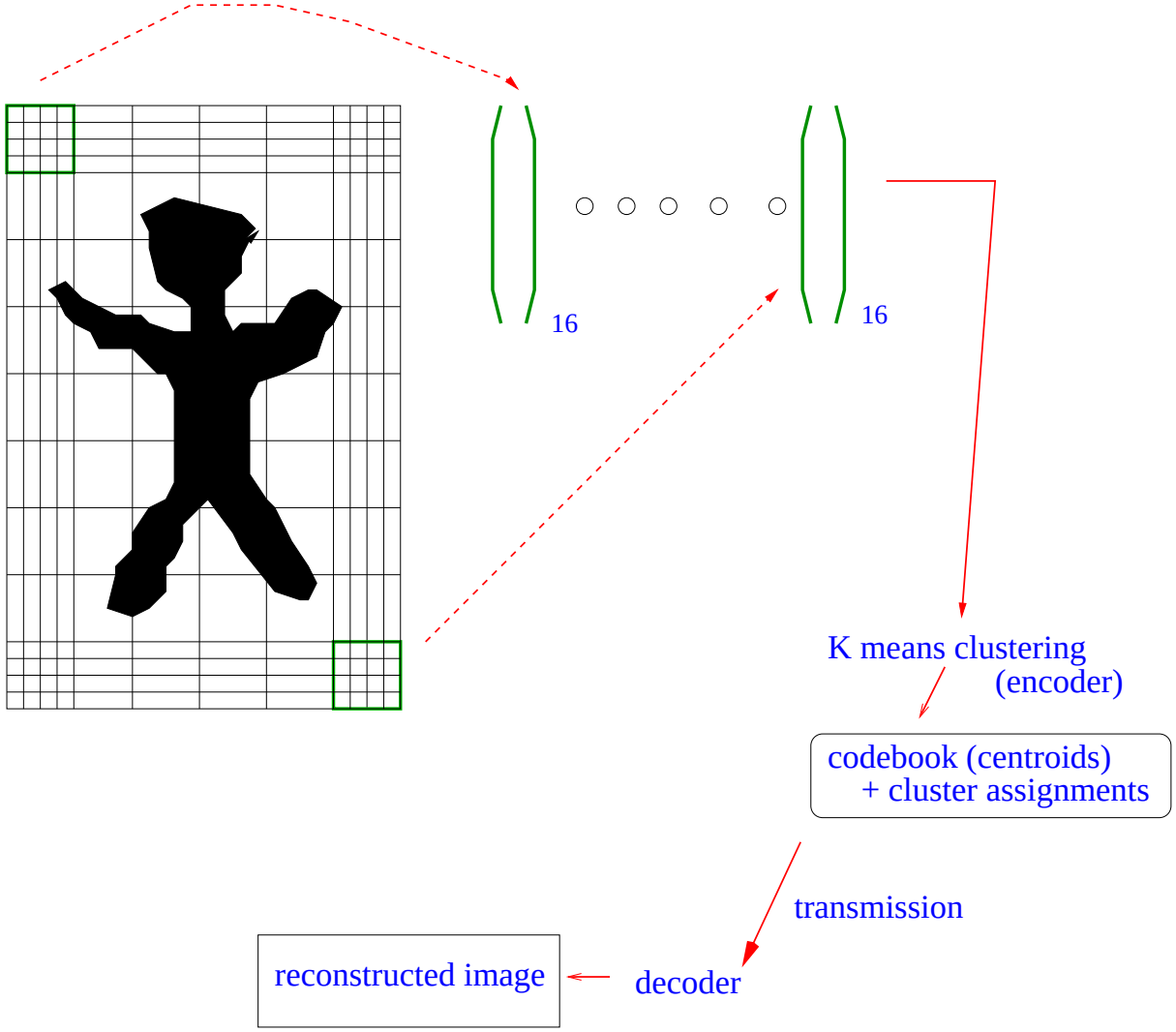


Successive iterations of the K-means clustering algorithm for the simulated data.

Vector Quantization

See  pp 466.

- VQ is k-means clustering, applied to vectors arising from the blocks of an image



Real application

Sir Ronald A. Fisher (1890-1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 greyscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel

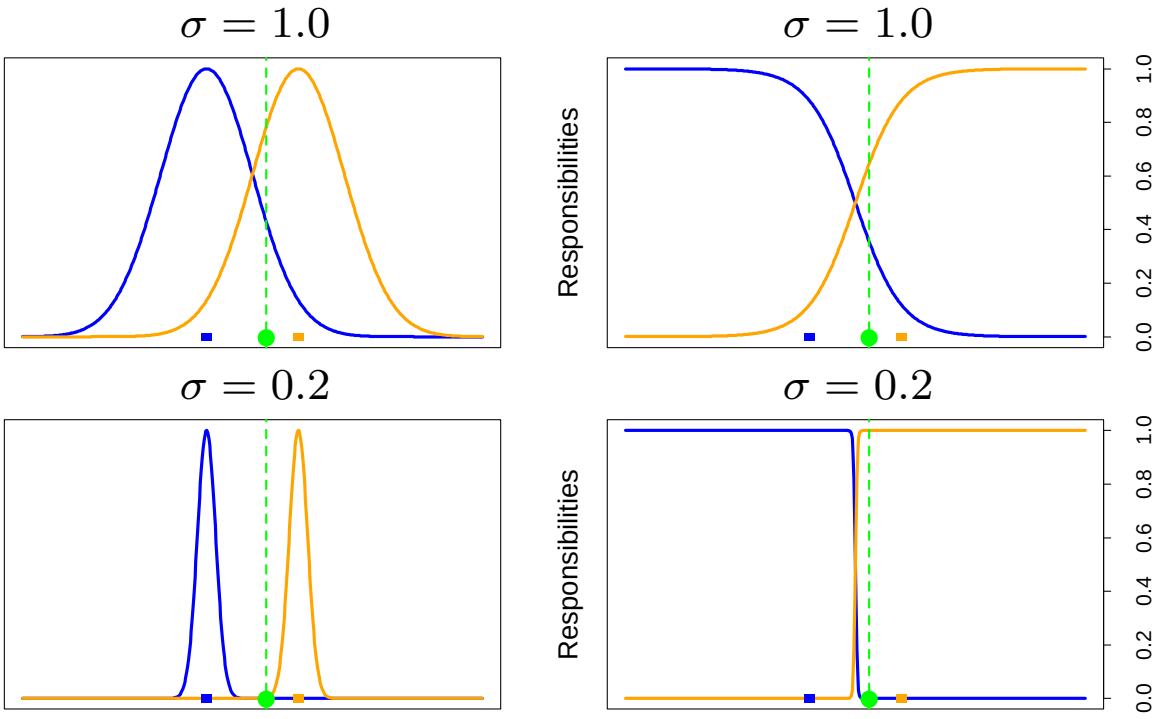
Gaussian mixtures and EM

Soft k-means clustering

See  pp 463.

Mixture Model: $f(x) = (1-\pi)g_1(x) + \pi g_2(x)$

Gaussian mixture: $g_j(x) = \phi_{\theta_j}(x)$, $\theta_j = (\mu_j, \sigma_j^2)$



Details of figure

- Left panels: two Gaussian densities $g_1(x)$ and $g_2(x)$ (blue and orange) on the real line, and a single data point (green dot) at $x = 0.5$. The colored squares are plotted at $x = -1.0$ and $x = 1.0$, the means of each density.
- Right panels: the relative densities $g_1(x)/(g_1(x) + g_2(x))$ and $g_2(x)/(g_1(x) + g_2(x))$, called the "responsibilities" of each cluster, for this data point. In the top panels, the Gaussian standard deviation $\sigma = 1.0$; in the bottom panels $\sigma = 0.2$.
- The EM algorithm uses these responsibilities to make a "soft" assignment of each data point to each of the two clusters. When σ is fairly large, the responsibilities can be near 0.5 (they are 0.36 and 0.64 in the top right panel).
- As $\sigma \to 0$, the responsibilities $\to 1$, for the cluster center closest to the target point, and 0 for all other clusters. This "hard" assignment is seen in the bottom right panel.

The EM Algorithm:

Two-Component Mixture Model

The left panel of Figure 1 shows a histogram of the 20 fictitious data points in Table 1.

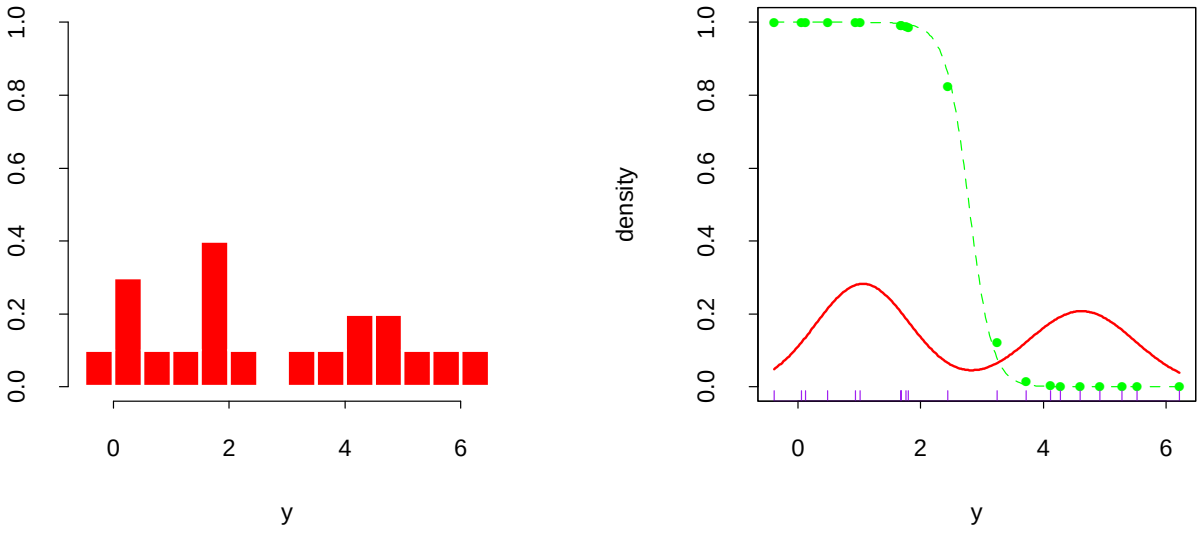


Figure 1: *Mixture example. Left panel: histogram of data. Right panel: maximum likelihood fit of Gaussian densities (solid red) and responsibility (dotted green) of the left component density for observation y, as a function of y.*

Table 1: *20 fictitious data points used in the two-component mixture example in Figure 1.*

-0.39	0.12	0.94	1.67	1.76	2.44	3.72	4.28	4.92	5.53
0.06	0.48	1.01	1.68	1.80	3.25	4.12	4.60	5.28	6.22

$$\begin{aligned} Y_1 &\sim N(\mu_1, \sigma_1^2), \\ Y_2 &\sim N(\mu_2, \sigma_2^2), \\ Y &= (1-\Delta)\cdot Y_1 + \Delta \cdot Y_2, \end{aligned}$$

where $\Delta \in \{0, 1\}$ with $\Pr(\Delta = 1) = \pi$.

Let $\phi_\theta(x)$ denote the normal density with parameters $\theta = (\mu, \sigma^2)$. Then the density of Y is

$$g_Y(y) = (1-\pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y).$$

The log-likelihood based on the N training cases is

$$\ell(\theta; \mathbf{z}) = \sum_{i=1}^{N} \log[(1-\pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)]. \quad (1)$$

Direct maximization of $\ell(\theta; \mathbf{z})$ is quite difficult numerically, because of the sum of terms inside the logarithm. There is, however, a simpler approach. We consider unobserved latent variables Δ_i taking values 0 or 1: if $\Delta_i = 1$ then Y_i comes from model 2, otherwise it comes from model 1. Suppose we knew the values of the Δ_i's. Then the log-likelihood would be

$$\begin{aligned}\ell_0(\theta; \mathbf{z}, \mathbf{\Delta}) &= \sum_{i=1}^{N} [(1-\Delta_i) \log \phi_{\theta_1}(y_i) + \Delta_i \log \phi_{\theta_2}(y_i)] \\ &\quad + \sum_{i=1}^{N} [(1-\Delta_i) \log \pi + \Delta_i \log(1-\pi)]\end{aligned}$$

Since the values of the Δ_i's are actually unknown, we proceed in an iterative fashion, substituting for each Δ_i its expected value

$$\gamma_i(\theta) = \mathrm{E}\,(\Delta_i|\theta, \mathbf{z}) = \Pr(\Delta_i = 1|\theta, \mathbf{z}), \qquad (2)$$

also called the responsibility of model 2 for observation i. We use a procedure called the EM algorithm.

EM algorithm for two-component Gaussian mixture.

- Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ (see text).
- *Expectation Step*: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}.(y_i)}, \; i = 1, 2, \ldots, N. \quad (3)$$

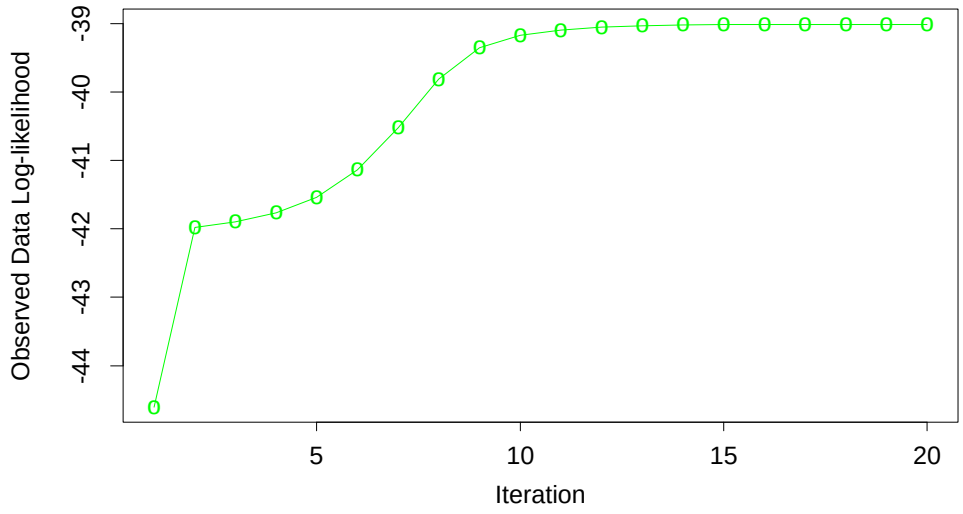
- *Maximization Step*: compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

- Iterate these steps until convergence.



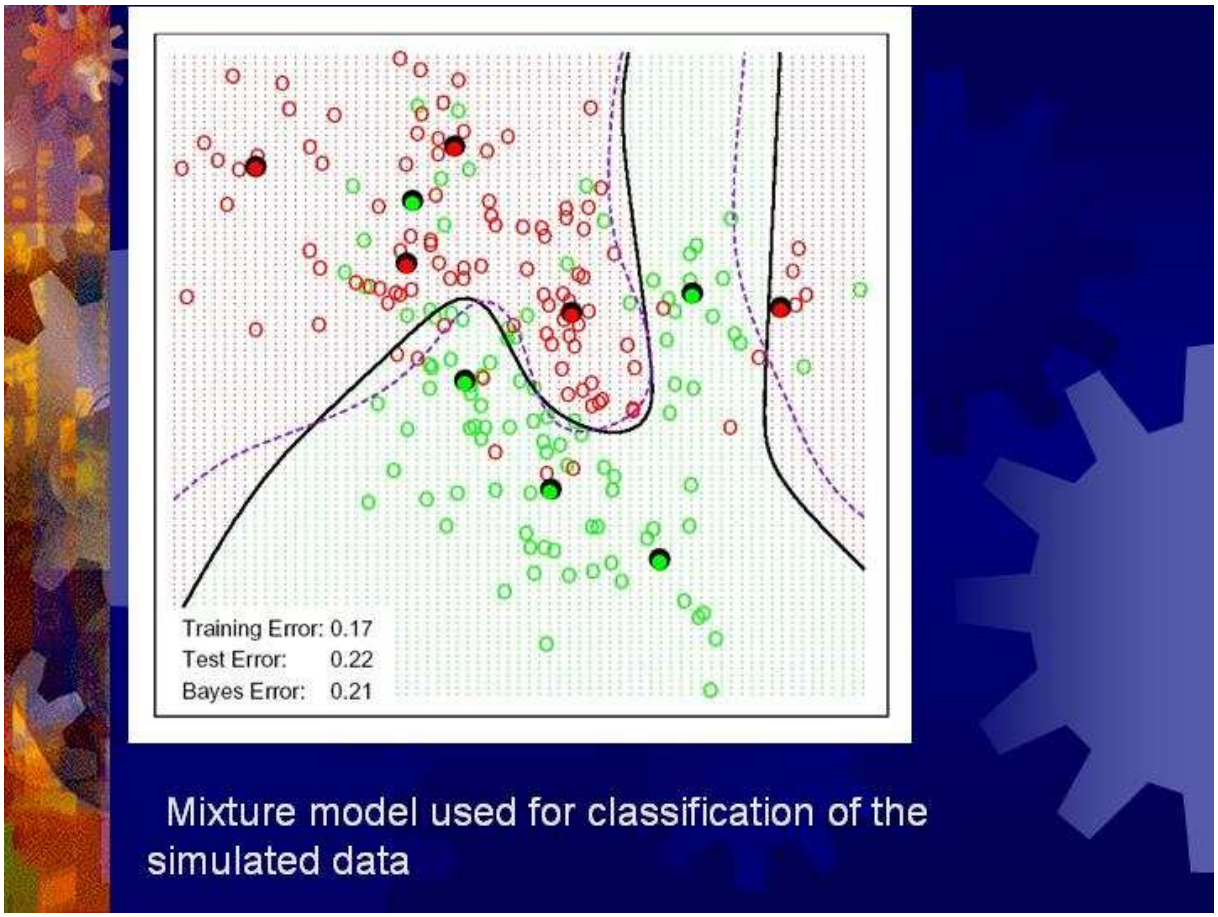
EM algorithm: observed data log-likelihood as a function of the iteration number.

Table 2: *Selected iterations of the EM algorithm for mixture example.*

Iteration	$\hat{\pi}$
1	0.485
5	0.493
10	0.523
15	0.544
20	0.546

The final maximum likelihood estimates are

$$\hat{\mu}_1 = 4.62, \qquad \hat{\sigma}_1^2 = 0.87,$$
$$\hat{\mu}_2 = 1.06, \qquad \hat{\sigma}_2^2 = 0.77,$$
$$\hat{\pi} = 0.546.$$



Mixture model used for classification of the simulated data

EM for general missing data problems

- Our observed data is $\mathbf{z}$, having log-likelihood $\ell(\theta; \mathbf{z})$ depending on parameters θ.
- The latent or missing data is $\mathbf{z}^m$, so that the complete data is $\mathbf{t} = (\mathbf{z}, \mathbf{z}^m)$ with log-likelihood $\ell_0(\theta; \mathbf{t})$, ℓ_0 based on the complete density.
- In the mixture problem $(\mathbf{z}, \mathbf{z}^m) = (\mathbf{y}, \Delta)$.
- EM paper in 1977 has interesting discussion- many including Hartley and Baum said that they had already done this work!

The EM algorithm.

1. Start with initial guesses for the parameters $\hat{\theta}^{(0)}$.

2. *Expectation Step*: at the jth step, compute

$$Q(\theta', \hat{\theta}^{(j)}) = \mathrm{E}\left(\ell_0(\theta'; \mathbf{t})|\mathbf{z}, \hat{\theta}^{(j)}\right) \tag{4}$$

   as a function of the dummy argument θ'.

3. *Maximization Step*: determine the new estimate $\hat{\theta}^{(j+1)}$ as the maximizer of $Q(\theta', \hat{\theta}^{(j)})$ over θ'.

4. Iterate steps 2 and 3 until convergence.

Proof that EM works

Since

$$\Pr(\mathbf{z}^m|\mathbf{z},\theta') = \frac{\Pr(\mathbf{z}^m,\mathbf{z}|\theta')}{\Pr(\mathbf{z}|\theta')}, \tag{5}$$

we can write

$$\Pr(\mathbf{z}|\theta') = \frac{\Pr(\mathbf{t}|\theta')}{\Pr(\mathbf{z}^m|\mathbf{z},\theta')}. \tag{6}$$

In terms of log-likelihoods, we have $\ell(\theta';\mathbf{z}) = \ell_0(\theta';\mathbf{t}) - \ell_1(\theta';\mathbf{z}^m|\mathbf{z})$, where ℓ_1 is based on the conditional density $\Pr(\mathbf{z}^m|\mathbf{z},\theta')$. Taking conditional expectations with respect to the distribution of $\mathbf{t}|\mathbf{z}$ governed by parameter θ gives

$$\begin{aligned} \ell(\theta';\mathbf{z}) &= \mathrm{E}\left[\ell_0(\theta';\mathbf{t})|\mathbf{z},\theta\right] - \mathrm{E}\left[\ell_1(\theta';\mathbf{z}^m|\mathbf{z})|\mathbf{z},\theta\right] \\ &\equiv Q(\theta',\theta) - R(\theta',\theta). \end{aligned} \tag{7}$$

In the M step, the EM algorithm maximizes $Q(\theta', \theta)$ over θ', rather than the actual objective function $\ell(\theta'; \mathbf{z})$.

Why does it succeed in maximizing $\ell(\theta'; \mathbf{z})$? Note that $R(\theta^*, \theta)$ is the expectation of a log-likelihood of a density (indexed by θ^*), with respect to the same density indexed by θ, and hence (by Jensen's inequality) is maximized as a function of θ^*, when $\theta^* = \theta$ (see Exercise 8.1). So if θ' maximizes $Q(\theta', \theta)$, we see that

$$\begin{aligned} \ell(\theta'; \mathbf{z}) - \ell(\theta; \mathbf{z}) &= [Q(\theta', \theta) - Q(\theta, \theta)] - [R(\theta', \theta) - R(\theta, \theta)] \\ &\geq 0. \end{aligned} \tag{8}$$

Hence the M step never decreases the log-likelihood.

A Different view

EM as a Maximization–Maximization Procedure

- Consider the function

$$F(\theta', \mathbf{P}) = \mathrm{E}_{\mathbf{P}}[\ell_0(\theta'; \mathbf{t})] - \mathrm{E}_{\mathbf{P}}[\log \mathbf{P}(\mathbf{z}^m)]. \quad (9)$$

- Here $\mathbf{P}(\mathbf{z}^m)$ is any distribution over the latent data $\mathbf{z}^m$. In the mixture example, $\mathbf{P}(\mathbf{z}^m)$ comprises the set of probabilities $\gamma_i = \Pr(\Delta_i = 1|\theta, \mathbf{z})$.

- Note that F evaluated at $\mathbf{P}(\mathbf{z}^m) = \Pr(\mathbf{z}^m|\mathbf{z}, \theta')$, is the log-likelihood of the observed data.

- The EM algorithm can be viewed as a joint maximization method for F over θ' and $\mathbf{P}(\mathbf{z}^m)$, by fixing one argument and maximizing over the other. The maximizer over $\mathbf{P}(\mathbf{z}^m)$ for fixed θ' can be shown to be

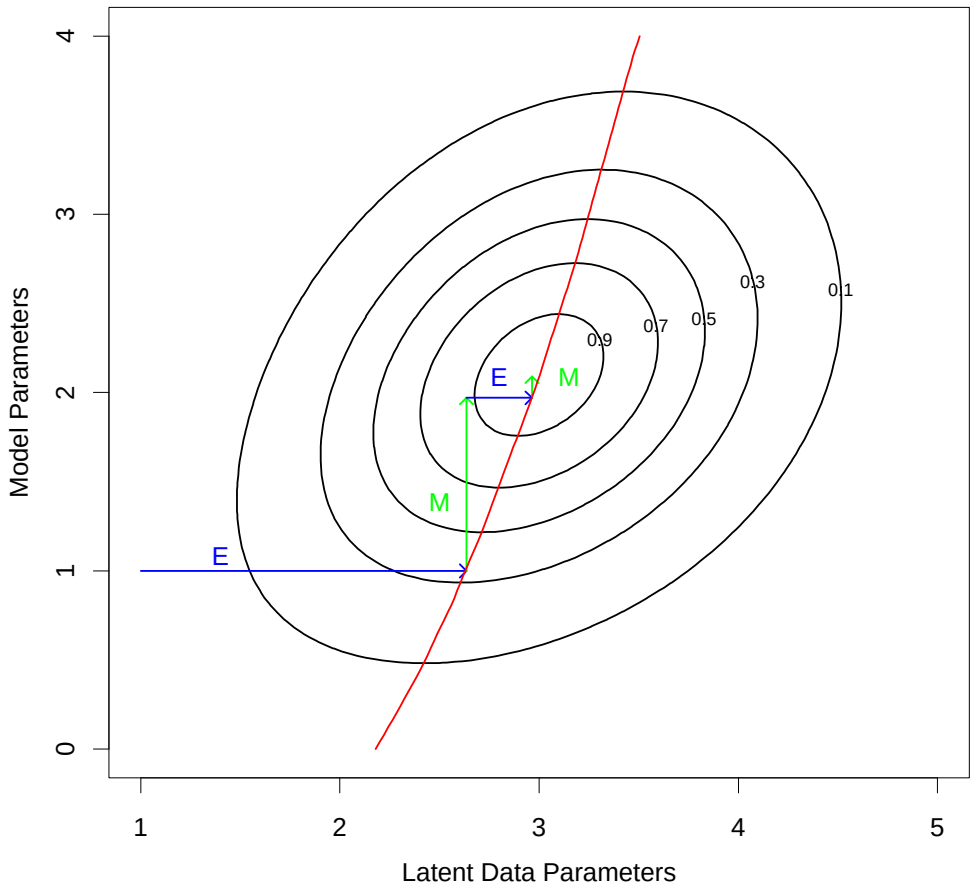
$$\mathbf{P}(\mathbf{z}^m) = \Pr(\mathbf{z}^m|\mathbf{z}, \theta') \tag{10}$$

  (Exercise 8.3).

  This is the distribution computed by the E step.

- In the M step, we maximize $F(\theta', \mathbf{P})$ over θ' with $\mathbf{P}$ fixed: this is the same as maximizing the first term $\mathrm{E}_{\mathbf{P}}[\ell_0(\theta'; \mathbf{t})|\mathbf{z}, \theta]$ since the second term does not involve θ'.

- Finally, since $F(\theta', \mathbf{P})$ and the observed data log-likelihood agree when $\mathbf{P}(\mathbf{z}^m) = \Pr(\mathbf{z}^m|\mathbf{z}, \theta')$, maximization of the former accomplishes maximization of the latter.



Maximization–maximization view of the EM algorithm. Shown are the contours of the (augmented) observed data log-likelihood $F(\theta', \tilde{P})$. The E step is equivalent to maximizing the log-likelihood over the parameters of the latent data distribution. The M step maximizes it over the parameters of the log-likelihood. The red curve corresponds to the observed data log-likelihood, a profile obtained by maximizing $F(\theta', \tilde{P})$ for each value of θ'.