Expectation-Maximization (EM) and Gaussian Mixture Models

Reference: The Elements of Statistical Learning,
by T. Hastie, R. Tibshirani, J. Friedman, Springer

The E.M. Algorithm

- But now we'll look at an even simpler case with hidden information.
- The EM algorithm
  - Can do trivial things, such as the contents of the next few slides.
  - An excellent way of doing our unsupervised learning problem, as we'll see.
  - Many, many other uses, including inference of Hidden Markov Models

Silly Example

Let events be "grades in a class"

w_1 = Gets an A $\quad P(A) = \frac{1}{2}$

w_2 = Gets a B $\quad P(B) = \mu$

w_3 = Gets a C $\quad P(C) = 2\mu$

w_4 = Gets a D $\quad P(D) = \frac{1}{2}-3\mu$

(Note $0 \le \mu \le 1/6$)

Assume we want to estimate μ from data. In a given class there were

a A's
b B's
c C's
d D's

What's the maximum likelihood estimate of μ given a,b,c,d ?

Trivial Statistics

$P(A) = \frac{1}{2}$ $P(B) = \mu$ $P(C) = 2\mu$ $P(D) = \frac{1}{2} - 3\mu$

$P(a,b,c,d \mid \mu) = K(\frac{1}{2})^a(\mu)^b(2\mu)^c(\frac{1}{2}-3\mu)^d$

$\log P(a,b,c,d \mid \mu) = \log K + a\log \frac{1}{2} + b\log \mu + c\log 2\mu + d\log (\frac{1}{2}-3\mu)$

$$\text{FOR MAX LIKE } \mu, \text{ SET } \frac{\partial \text{LogP}}{\partial \mu} = 0$$

$$\frac{\partial \text{LogP}}{\partial \mu} = \frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{1/2 - 3\mu} = 0$$

$$\text{Gives max like } \mu = \frac{b+c}{6(b+c+d)}$$

So if class got

A	B	C	D
14	6	9	10

$$\text{Max like } \mu = \frac{1}{10}$$

Same Problem with Hidden Information

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

What is the max. like estimate of μ now?

REMEMBER

$P(A) = \frac{1}{2}$

$P(B) = \mu$

$P(C) = 2\mu$

$P(D) = \frac{1}{2} - 3\mu$

Same Problem with Hidden Information

REMEMBER
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

What is the max. like estimate of μ now?

We can answer this question circularly:

EXPECTATION

If we know the value of μ we could compute the expected value of a and b

Since the ratio a:b should be the same as the ratio ½ : μ

$$a = \frac{\frac{1}{2}}{\frac{1}{2}+\mu}h \qquad b = \frac{\mu}{\frac{1}{2}+\mu}h$$

MAXIMIZATION

If we know the expected values of a and b we could compute the maximum likelihood value of μ

$$\mu = \frac{b+c}{6(b+c+d)}$$

E.M. for our Trivial Problem

REMEMBER
P(A) = ½
P(B) = μ
P(C) = 2μ
P(D) = ½-3μ

We begin with a guess for μ

We iterate between EXPECTATION and MAXIMALIZATION to improve our estimates of μ and *a* and *b*.

Define μ(t) the estimate of μ on the t'th iteration

b(t) the estimate of *b* on t'th iteration

$$\mu(0) = \text{initial guess}$$

$$b(t) = \frac{\mu(t)h}{\frac{1}{2} + \mu(t)} = \mathrm{E}\left[b \mid \mu(t)\right]$$

E-step

$$\mu(t+1) = \frac{b(t)+c}{6\left(b(t)+c+d\right)}$$

$$= \text{max like est of } \mu \text{ given } b(t)$$

M-step

Continue iterating until converged.

E.M. Convergence

- Convergence proof based on fact that Prob(data | μ) must increase or remain same between each iteration [NOT OBVIOUS]
- But it can never exceed 1 [OBVIOUS]

So it must therefore converge [OBVIOUS]

In our example, suppose we had

$h = 20$

$c = 10$

$d = 10$

$\mu(0) = 0$

Convergence is generally <u>linear</u>: error decreases by a constant factor each time step.

t	μ(t)	b(t)
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

Unsupervised Learning
Motivation

- Unsupervised learning aims at finding some patterns or characteristics of the data.
- It does not need the *class* attribute.

Consider the following data set:

-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

- Model the density of the data points
- A simple and common way: single Gaussian model

Unsupervised Learning

Motivation

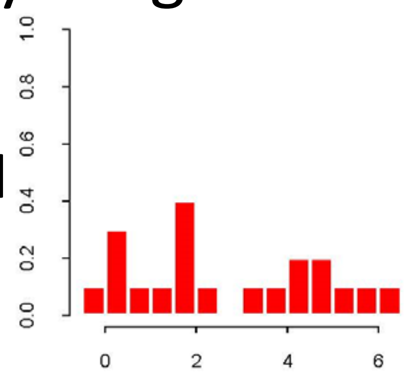
- Unsupervised learning aims at finding some patterns or characteristics of the data.
- It does not need the *class* attribute.

Consider the following data set:

-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

- Model the density of the data points
- A simple and common way: single Gaussian model

From histogram of the data points, single Gaussian model is poor



Mixture Model

Basic Framework

- Also called clustering
- Relate to grouping or segmenting a collection of objects into subsets or “clusters”
- Within each cluster are more closely related to one another than objects assigned to different clusters
- Form descriptive statistics to ascertain whether or not the data consists of a set distinct subgroups

Mixture Model

Basic Framework

- The mixture model is a probabilistic clustering paradigm.
- It is a useful tool for density estimation.
- It can be viewed as a kind of kernel method.
- Gaussian mixture model:

$$f(x) = \sum_{m=1}^{M} \alpha_m \phi(x; \mu_m, \boldsymbol{\Sigma}_m)$$

Mixture Model

Basic Framework

- Gaussian mixture model:

$$f(x) = \sum_{m=1}^{M} \alpha_m \phi(x; \mu_m, \boldsymbol{\Sigma}_m)$$

- α_m are mixing proportions, and $\sum_m \alpha_m = 1$
- Each Gaussian density has a mean μ_m and covariance matrix $\boldsymbol{\Sigma}_m$
- Can use any component densities in place of the Gaussian
- The Gaussian mixture model is by far the most popular

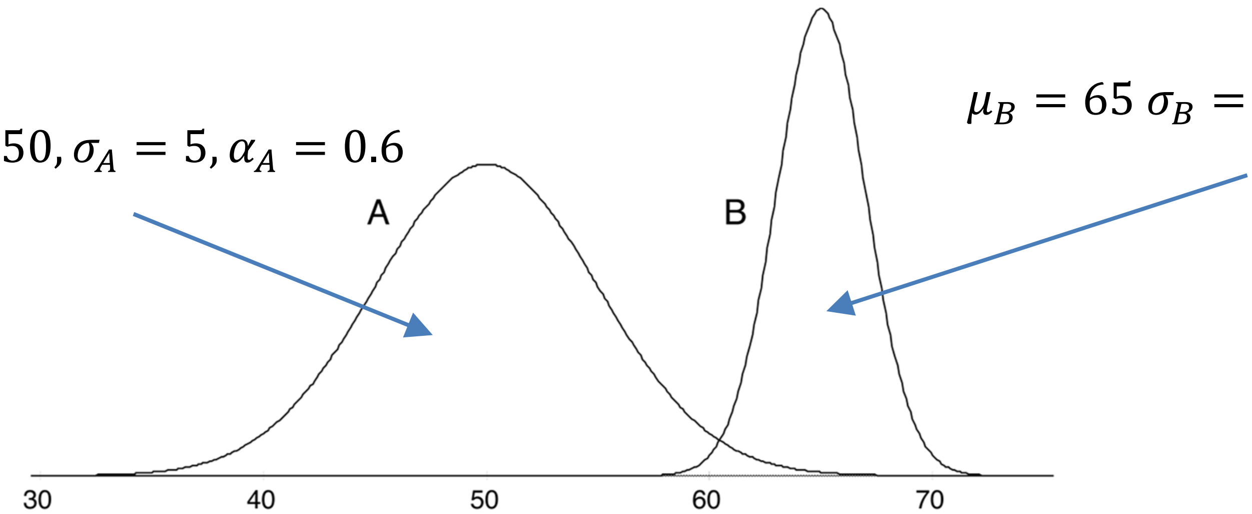
Mixture Model

Example

An example of Gaussian mixture model with 2 components.

$\mu_A = 50, \sigma_A = 5, \alpha_A = 0.6$

$\mu_B = 65\ \sigma_B = 2, \alpha_B = 0.4$



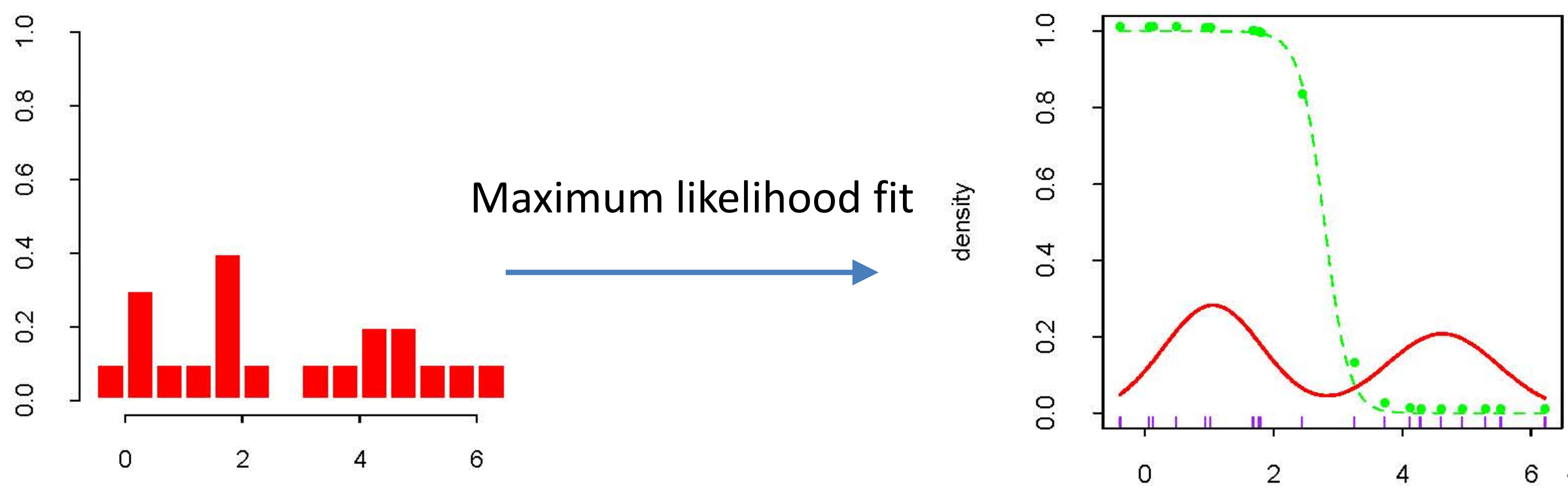
Sample data points generated from the model

A	51	B	62	B	64	A	48	A	39	A	51
A	43	A	47	A	51	B	64	B	62	A	48
B	62	A	52	A	52	A	51	B	64	B	64
B	64	B	64	B	62	B	63	A	52	A	42
A	45	A	51	A	49	A	43	B	63	A	48
A	42	B	65	A	48	B	65	B	64	A	41
A	46	A	48	B	62	B	66	A	48		
A	45	A	49	A	43	B	65	B	64		
A	45	A	46	A	40	A	46	A	48		

Mixture Model Learning

Sample Result

- Due to the apparent bi-modality
  → Single Gaussian distribution would not be appropriate
- A simple mixture model for density estimation
- Associated EM algorithm for carrying out maximum likelihood estimation



Mixture Model Learning

Two-Component Model

- Two separate underlying regimes
  → instead model Y as mixture of two normal distributions:
  $$Y_1 \sim N(\mu_1, \sigma_1^2)$$
  $$Y_2 \sim N(\mu_2, \sigma_2^2)$$
  $$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2$$
  where $\Delta \in \{0, 1\}$ with $\Pr(\Delta = 1) = \pi$
- Generative representation is explicit: generate a $\Delta \in \{0, 1\}$ with probability π
- Depending on outcome, deliver Y_1 or Y_2

Mixture Model Learning

Two-Component Model

- Let $\phi_\theta(x)$ denote the normal density with parameters $\theta = (\mu, \sigma^2)$
- Density of Y:

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

Mixture Model Learning

Two-Component Model

- Denote the training data by $\boldsymbol{Z} = \{\boldsymbol{y_1}, \cdots, \boldsymbol{y_N}\}$
- Fit the model to the data by maximum likelihood, the parameters:

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

- Log-likelihood based on the N training cases:

$$\ell(\theta; \boldsymbol{Z}) = \sum_{i=1}^{N} \log\left[(1-\pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)\right]$$

Mixture Model Learning

Two-Component Model

- Direct maximization of $\ell(\theta; \boldsymbol{Z})$ is quite difficult numerically, because of the sum of terms inside the logarithm
- Consider unobserved latent variables Δ_i taking values 0 or 1
  - if $\Delta_i = 1$ → Y_i comes from model 2
  - otherwise, comes from model 1

Mixture Model Learning

Two-Component Model

- Suppose knew the values of the Δ_i's
  → the log-likelihood:
  $$\ell_0(\theta; \boldsymbol{Z}, \Delta)$$
  $$= \sum_{i=1}^{N} \left[(1 - \Delta_i) \log \phi_{\theta_1}(y_i) + \Delta_i \log \phi_{\theta_2}(y_i)\right]$$
  $$+ \sum_{i=1}^{N} \left[(1 - \Delta_i) \log(1 - \pi) + \Delta_i \log \pi\right]$$

Mixture Model Learning

Two-Component Model

- Maximum likelihood estimates:
  μ_1 and σ_1^2 - sample mean and variance for those data with $\Delta_i = 0$
  μ_2 and σ_2^2 - sample mean and variance for those data with $\Delta_i = 1$
- Estimate of π would be the proportion of $\Delta_i = 1$
- Δ_i is unknown → iterative fashion, substituting for each Δ_i in its expected value

$$\gamma_i(\theta) = E(\Delta_i|\theta, \boldsymbol{Z}) = \Pr(\Delta_i = 1|\theta, \boldsymbol{Z})$$

- γ_i is also called *responsibility* of model 2 for observation i

Two-Component Mixture Model
EM Algorithm

EM algorithm for two-component Gaussian mixtures:

1. Take initial guesses for the parameters

$$\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$$

2. Expectation Step: compute the responsibilities

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

Two-Component Mixture Model

EM Algorithm

3. Maximization Step:

Compute the weighted means and variances

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

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

and the mixing probability

$$\hat{\pi} = \sum_{i=1}^{N}\hat{\gamma}_i/N$$

4. Iterate steps 2 and 3 until convergence

Two-Component Mixture Model

EM Algorithm

- In expectation step – do soft assignment of each observation to each model:
  - Current estimates of the parameters are used to assign responsibilities according to the relative density of the training points under each model
- In maximization step – weighted maximum-likelihood fits to update the estimates of the parameters

Two-Component Mixture Model

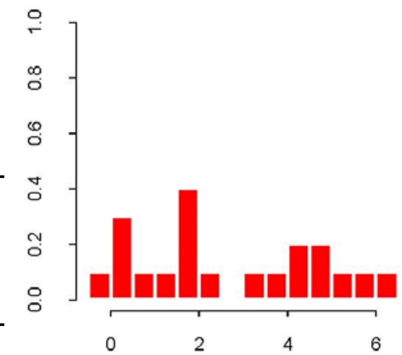
EM Algorithm

- Construct initial guesses for $\hat{\mu}_1$ and $\hat{\mu}_2$: choose two of the y_i at random
- Both $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ set equal to the overall sample variance $\sum_{i=1}^{N}(y_i - \bar{y})^2/N$
- Mixing proportion $\hat{\pi}$ can be started at the value 0.5

Two-Component Mixture Model

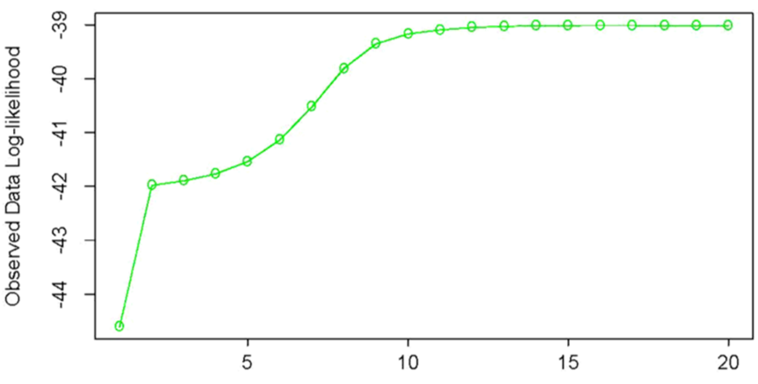
Example of Running EM

- Returning to the previous data set



-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

- The progress of the EM algorithm in maximizing the log-likelihood



- $\hat{\pi} = \sum_i \hat{\gamma}_i / N$
  the maximum likelihood estimate of the proportion of observations in class 2, at selected iterations of the EM procedure

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

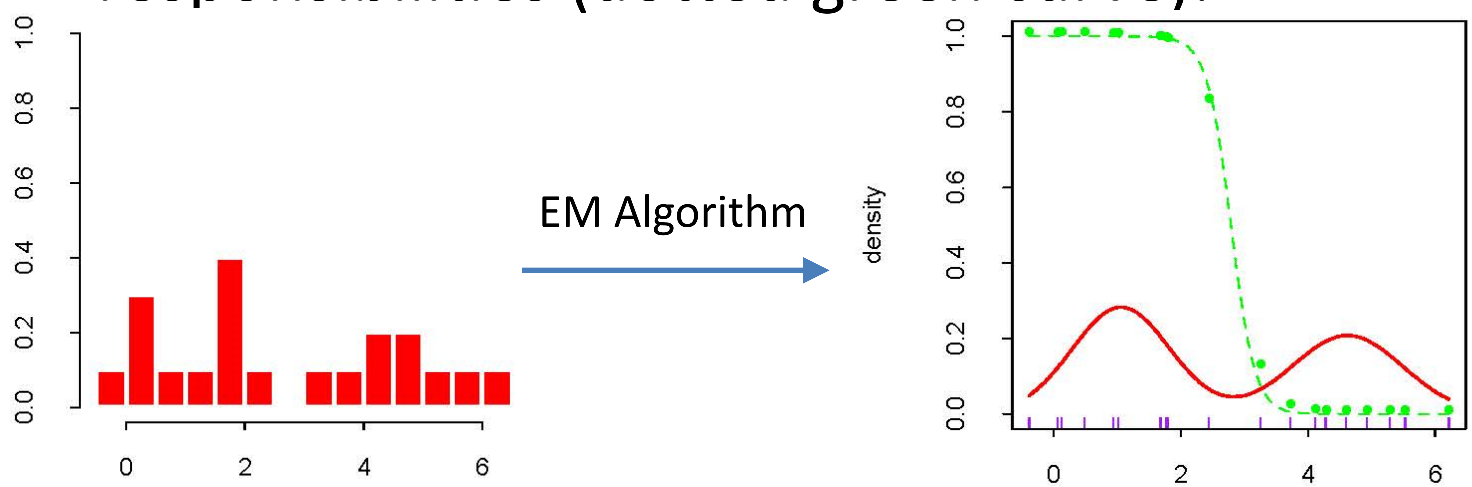
Two-Component Mixture Model

Example of Running EM

- The final maximum likelihood estimates:

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

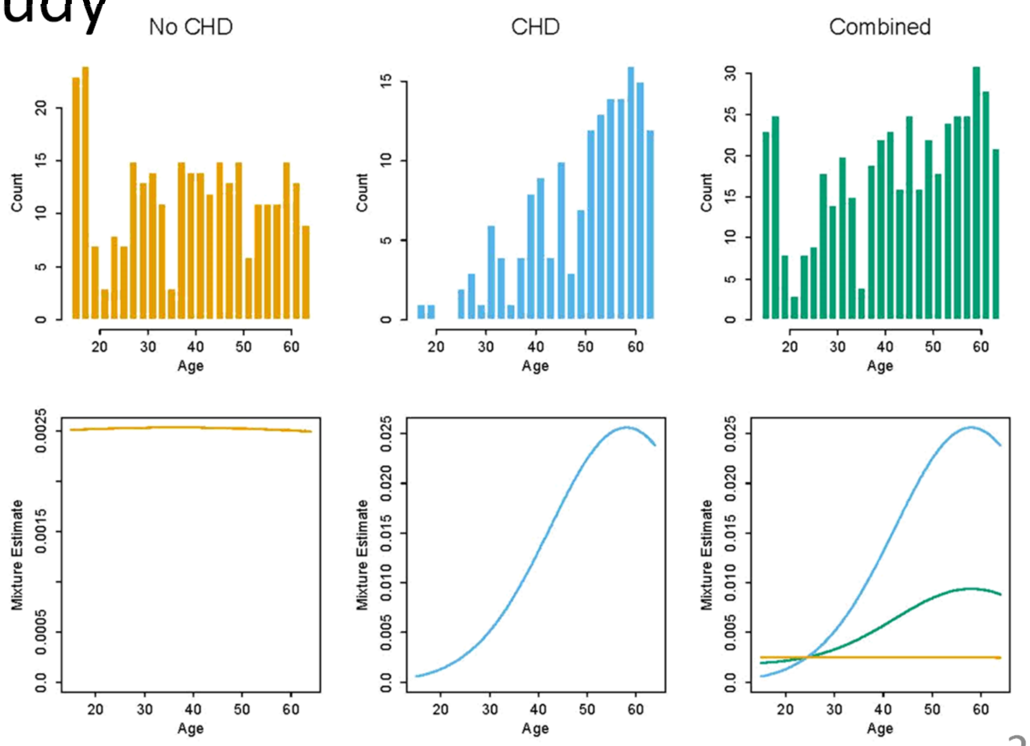
- The estimated Gaussian mixture density from this procedure (solid red curve), along with the responsibilities (dotted green curve):



Mixture Models

Heart Disease Risk Data Set

- Using Bayes' theorem, separate mixture densities in each class lead to flexible models for $\Pr(G|X)$
- An application of mixtures to the heart disease risk factor (CHD) study



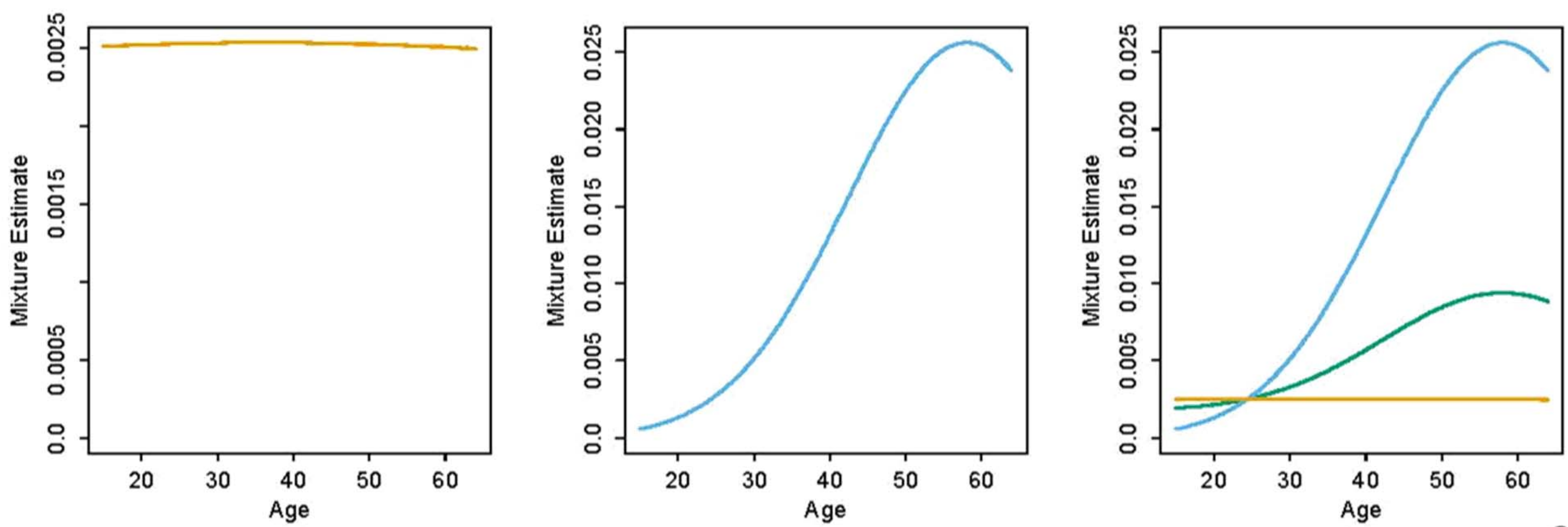
Mixture Models
Heart Disease Risk Data Set

- Using the combined data
  → fit a two-component mixture of the form with the (scalars) $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ not constrained to be equal
- Fitting via the EM algorithm: procedure does not knowledge of the CHD labels
- Resulting estimates:
  $\hat{\mu}_1 = 36.4, \quad \widehat{\boldsymbol{\Sigma}}_1 = 157.7 \quad \hat{\alpha}_1 = 0.7$
  $\hat{\mu}_2 = 58.0, \quad \widehat{\boldsymbol{\Sigma}}_2 = 15.6 \quad \hat{\alpha}_2 = 0.3$

Mixture Models

Heart Disease Risk Data Set

- Lower-left and middle panels:
  Component densities $\phi(\hat{\mu}_1, \widehat{\boldsymbol{\Sigma}}_1)$ and $\phi(\hat{\mu}_2, \widehat{\boldsymbol{\Sigma}}_2)$
- Lower-right panel:
  Component densities (orange and blue) along with the estimated mixture density (green)



Mixture Models

Heart Disease Risk Data Set

- Mixture model provides an estimate of the probability – observation i belongs to component m:

$$\hat{r}_{im} = \frac{\hat{\alpha}_m \phi\left(x_i; \hat{\mu}_m, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{m}}\right)}{\sum_{k=1}^{M} \hat{\alpha}_k \phi\left(x_i; \hat{\mu}_k, \widehat{\boldsymbol{\Sigma}}_k\right)}$$

  where x_i is Age in the example
- Suppose threshold each value $\hat{r}_{i2}$
  → define $\hat{\delta}_i = I(\hat{r}_{i2} > 0.5)$

Mixture Models
Heart Disease Risk Data Set

- Compare the classification of each observation by CHD and the mixture model:

		Mixture model	
		$\hat{\delta} = 0$	$\hat{\delta} = 1$
CHD	No	232	70
	Yes	76	84

- Although did not use the CHD labels, can discover the two CHD subpopulations
- Error rate: $\frac{76+70}{462} = 32\%$

Mixture Models

Heart Disease Risk Data Set

- Linear logistic regression, using CHD as a response:
  same error rate (32%) when fit to these data using maximum-likelihood