Gaussian Mixture Models Clustering

-Applied Multivariate Analysis & Statistical Learning-

MMA 15.4.2 and these notes

Lecturer: Darren Homrighausen, PhD

Preamble:

- Review marginal and conditional distributions
- Define density based clustering
- Give a quick overview of one implementation based on mixtures of multivariate normal distributions

Reminder: Marginalization and conditional probability

Suppose we have the joint density of two random variables Z, W. If we don't the values of W, we just look at the marginal density of Z as:

$$f(z) = \int f(z, w) dw$$

Also, we can always write

$$f(z, w) = f(z|w) f(w)$$

Therefore,

$$f(z) = \int f(z|w) f(w) dw$$

Density-based clustering approaches

Suppose that X has some label Y, but we don't observe it

Clustering can be thought of attempting to estimate this label, but having access to no labeled observations

Therefore, we can write

$$p(X) = \sum_{j=1}^{K} p(X, Y = j) = \sum_{j=1}^{K} p(X|Y = j)p(Y = j)$$

Now, we have a distribution on X alone. This gives us a likelihood

$$\prod_{i=1}^{n} \sum_{j=1}^{K} p(X_i|Y_i = j)p(Y = j)$$

Now, if we can just maximize this likelihood...

(Using expectation maximization (EM))

Gaussian mixture models (GMM)

Now, there are many choices for $p(X|Y=j)$:

- $N(\mu_j, \sigma^2 I)$ (corresponds nearly to K-means)
  (In fact, as $\sigma^2 \to 0$, GMM converges to K-means)
- $N(\mu_j, \sigma_j^2 I)$
- $N(\mu_j, \Sigma)$
- $N(\mu_j, \Sigma_j)$

Note: these are in increasing order of complexity.

We can use BIC to choose the number of parameters for maximum likelihood in this situation!

ESL Chapter 14.7.3

Gaussian mixture models (GMM) in R

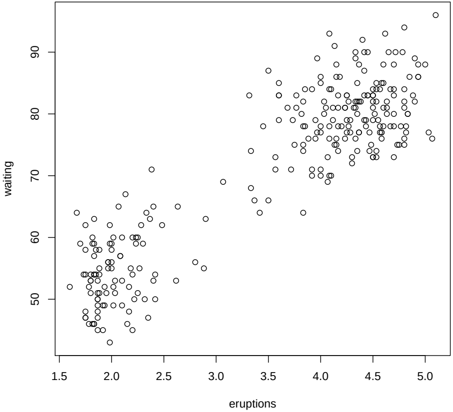


Figure: Data from 'Old Faithful' geyser in Yellowstone. Eruptions is length of the eruption and waiting is time until next eruption.

Gaussian mixture models (GMM) in R

identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
E		•	•	(univariate)	equal		
V		•	•	(univariate)	variable		
EII	λI	•	•	Spherical	equal	equal	NA
VII	λ_k I	•	•	Spherical	variable	equal	NA
EEI	λA		•	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_k A$		•	Diagonal	variable	equal	coordinate axes
EVI	λA_k		•	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_k A_k$		•	Diagonal	variable	variable	coordinate axes
EEE	$\lambda D A D^T$	•	•	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_k A D_k^T$		•	Ellipsoidal	equal	equal	variable
VEV	$\lambda_k D_k A D_k^T$		•	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k D_k A_k D_k^T$	•	•	Ellipsoidal	variable	variable	variable

Figure: Acronyms for `Mclust` in `R`.

Gaussian mixture models (GMM) in R

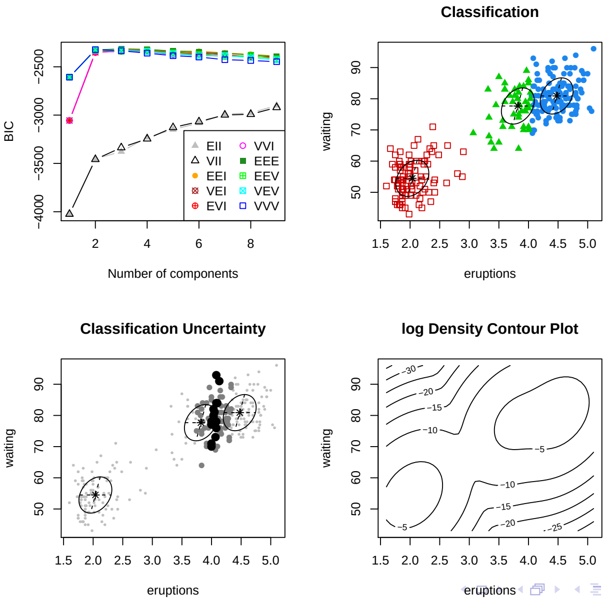
```
library(mclust)
faithGMM = Mclust(faithful)
> summary(faithGMM)
----------------------------------------------------
Gaussian finite mixture model fitted by
EM algorithm
----------------------------------------------------

Mclust EEE (elliposidal, equal volume,
shape and orientation) model with 3 components:

 log.likelihood   n df       BIC
      -1126.361 272 11 -2314.386

Clustering table:
  1   2   3
130  97  45
```

Gaussian mixture models (GMM) in R



Postamble:

- Review marginal and conditional distributions

  (Main result: we can write the density (and hence likelihood of X) as a function of X, treating the label Y as a nuisance parameter)

- Define density based clustering

  (Density-based clustering assigns clusters based on maximizing the probability. It is commonly referred to as 'soft' clustering as we really estimate probabilities of being in a cluster)

- Give a quick overview of one implementation based on mixtures of multivariate normal distributions

  (The R packages mclust provides a nice implementation for doing density-based clustering using a mixture of Gaussians)