Model-based clustering

- One disadvantage of hierarchical clustering algorithms, k-means algorithms and others is that they are largely heuristic and not based on formal models. Formal inference is not possible.

- Not necessarily a disadvantage since clustering is largely exploratory.

- Model-based clustering is an alternative. Banfield and Raftery (1993, *Biometrics*) is the classic reference. A more comprehensive and up-to-date reference is Melnykov and Maitra (2010, *Statistics Surveys*) also available on Professor Maitra's "Manuscripts Online" link.

- SAS will not implement model-based clustering algorithms.

- With R, you need to load a package called `mclust` and accept the terms of the (free) license. `mclust` is a very good package, but it can have issues with initialization.

Basic idea behind Model-based Clustering

- Sample observations arise from a distribution that is a mixture of two or more components.

- Each component is described by a density function and has an associated probability or "weight" in the mixture.

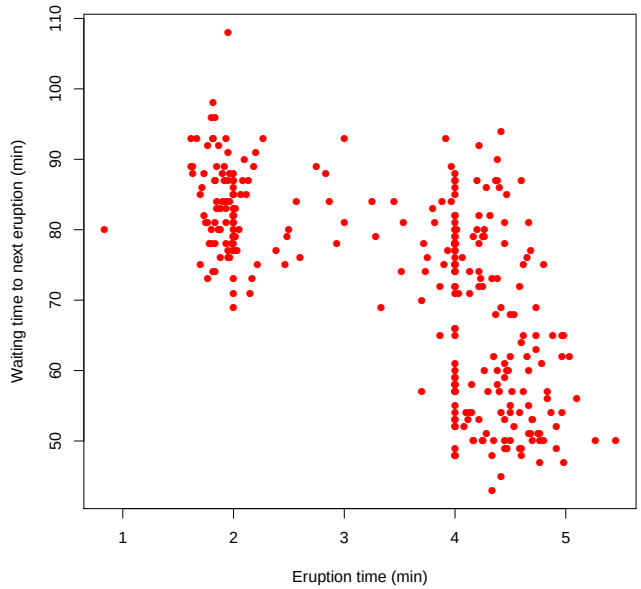
- In principle, we can adopt any probability model for the components, but typically we will assume that components are p-variate normal distributions. (This does not necessarily mean things are easy: inference in tractable, however.)

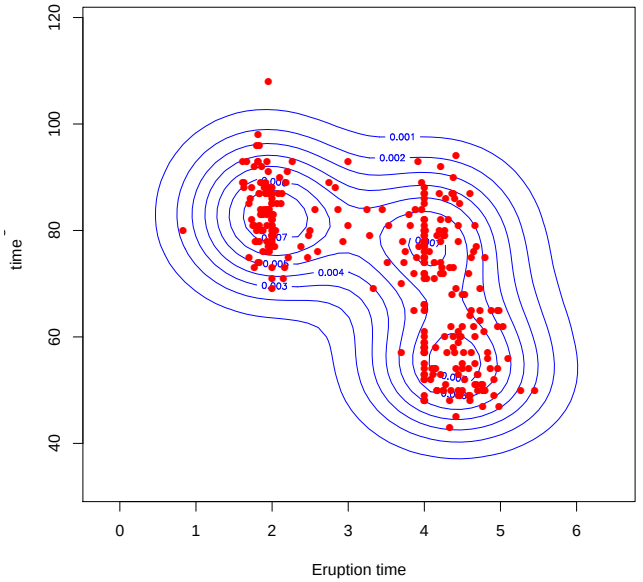
- Thus, the probability model for clustering will often be a mixture of multivariate normal distributions.

- Each component in the mixture is what we call a cluster.

Example: Old Faithful eruptions

Data: 272 observations of the waiting time between eruptions and the duration of the eruptions of Old Faithful (`geyser` in `MASS`)






```
library(KernSmooth)
est <- bkde2D(geyser[, 2:1], bandwidth=c(0.7, 7))
contour(est$x1, est$x2, est$fhat, col = "blue", ylab = "Waiting
time", xlab = "Eruption time")
points(geyser[,2:1], col = "red", pch = 16)
```

Old Faithful Geyser Data (cont'd)

- There seem to be at least three components (perhaps four) in the mixture.

- Contours suggest a mixture of three approximately bivariate normal distributions.

- Since the clouds of points form ellipses that appear to be "similar" in terms of volume, shape and orientation, we might anticipate that the three components (or four?) of this mixture might have homogeneous covariance matrices.

Model-based clustering (continued)

- Set-up: n p-dimensional observations $x_1, \cdots, x_n$. We assume that the joint distribution is a mixture of G components, each of which is multivariate normal with density $f_k(x|\mu_k, \Sigma_k), k = 1, ..., G$.

- The mixture model is then

$$
\begin{aligned}
f(x|\pi, \mu, \Sigma) &= \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k f_k(x_i|\mu_k, \Sigma_k) \\
&= \text{const.} \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k |\Sigma_k|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \mu_k)' \Sigma_k^{-1} (x_i - \mu_k)\},
\end{aligned}
$$

  where π_k = probability that x_i belongs to the kth component ($0 < \pi_k < 1$, $\sum_k \pi_k = 1$).

- Difficult to find MLEs of these parameters directly: use EM.

Parameter Estimation

- Suppose we observed the random variables which indicates which component each $\boldsymbol{x}_i$ belongs to.

- Let $\zeta_{ik} = 1$ if ith observation belongs to the kth group, zero otherwise.

- With complete data, obtaining the MLE of the component parameters would be trivial:

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i \in E_k} \boldsymbol{x}_i}{n_k}, \quad \hat{\pi}_k = \frac{n_k}{n}, \quad \hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i \in E_k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)'}{n_k},$$

  where $E_k = \{i : \gamma_i = k\}$ and n_k is the number of elements in E_k.

- Unfortunately, we do not know the ζ_{ik}s, so we treat these as missing observations.

Parameter Estimation via EM

- Note: If the group identifiers were known, estimation would be easy.
- Then the complete loglikelihood and corresponding Q-function are:

$$\ell(\boldsymbol{\vartheta}; \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\zeta_{ik}\left\{\log|\boldsymbol{\Sigma}_k| + (\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}$$
$$+\sum_{i=1}^{n}\sum_{k=1}^{K}\zeta_{ik}\log\pi_k - \frac{pn}{2}\log 2\pi.$$

$$Q(\boldsymbol{\vartheta}; \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\left\{\log|\boldsymbol{\Sigma}_k| + (\boldsymbol{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}$$
$$+\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\log\pi_k - \frac{pn}{2}\log 2\pi.$$

where $\pi_{ik} = E(\zeta_{ik}|\boldsymbol{x}_i)$ is calculated at the E-step.

The E- and M-steps in Parameter Estimation

- The E-step at the sth iteration:

$$\pi_{ik}^{(s)} = \frac{\pi_k^{(s-1)} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(s-1)}, \boldsymbol{\Sigma}_k^{(s-1)})}{\sum_{k'=1}^{K} \pi_{k'}^{(s-1)} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_{k'}^{(s-1)}, \boldsymbol{\Sigma}_{k'}^{(s-1)})}.$$

- M-step at the sth iteration:

$$\pi_k^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \pi_{ik}^{(s)}, \qquad \boldsymbol{\mu}_k^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}^{(s)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \pi_{ik}^{(s)}},$$

$$\text{and } \boldsymbol{\Sigma}_k^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}^{(s)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(s)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(s)})'}{\sum_{i=1}^{n} \pi_{ik}^{(s)}}.$$

- iterate from initial values (**need to choose carefully**) till convergence.

Modeling Σ_k

- The components or clusters in the mixture model have elipsoidal shape and are centered at μ_k.
- However, the G groups need not have homogenous covariance matrices.

- Since attributes of Σ_k determine the geometric features of the kth component, we can parameterize each Σ_k in the mixture model as

$$\Sigma_k = \lambda_k D_k A_k D_k',$$

  where D_k is the orthogonal matrix of eigenvectors of Σ_k, A_k is a diagonal matrix with elements proportional to the eigenvalues of Σ_k and λ_k is a scalar.
- D_k determines the orientation of the principal components of Σ_k, A_k determines the shape of the density countours, and λ_k determines the volume of the ellipsoid, which is proportional to $\lambda_k^p |A_k|$.

Mclust to implement algorithm

- The model-based clustering algorithm can be implemented using mclust package (`Mclust` function) in `R`.
- Mclust uses an identifier for each possible parametrization of the covariance matrix that has three letters: `E` for "equal", `V` for "variable" and `I` for "coordinate axes".
- The first identifier refers to volume, the second to shape and the third to orientation. For example:
  - `EEE` means that the G clusters have the same volume, shape and orientation in $p-$dimensional space.
  - `VEI` means variable volume, same shape and orientation equal to coordinate axes.
  - `EIV` means same volume, spherical shape and variable orientation.
- There are a total of 10 combinations of volume, shape and orientation included in the package.

Choosing the best model

- Given G, how do we choose a model?
- Any model selection criterion (AIC, likelihood ratio, BIC) can be used to select the best fitting model given G.
- Mclust uses the Bayesian Information Criterion (BIC, or Schwarz Information Criterion) to choose the best model given G.

$$BIC = -2\log(L) + m\log(n),$$

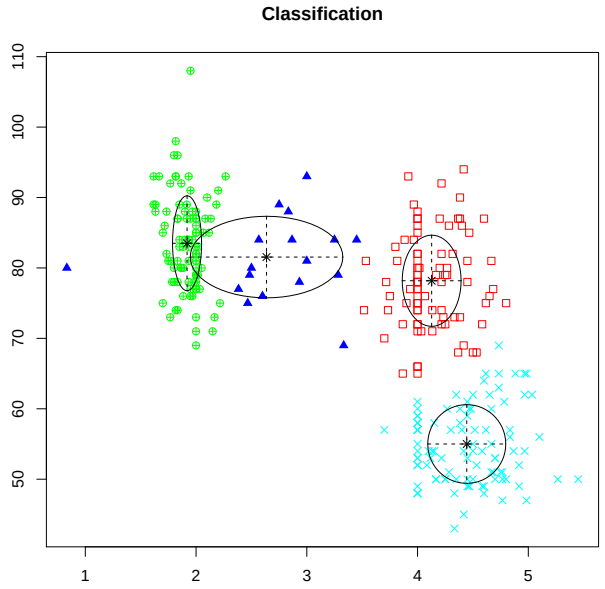
  where L is the likelihood function and m is the number of free parameters to be estimated. A model with low BIC fits the data better than one with high BIC.

- BIC penalizes large models (large m) more than AIC for which the penalty is $2m$.

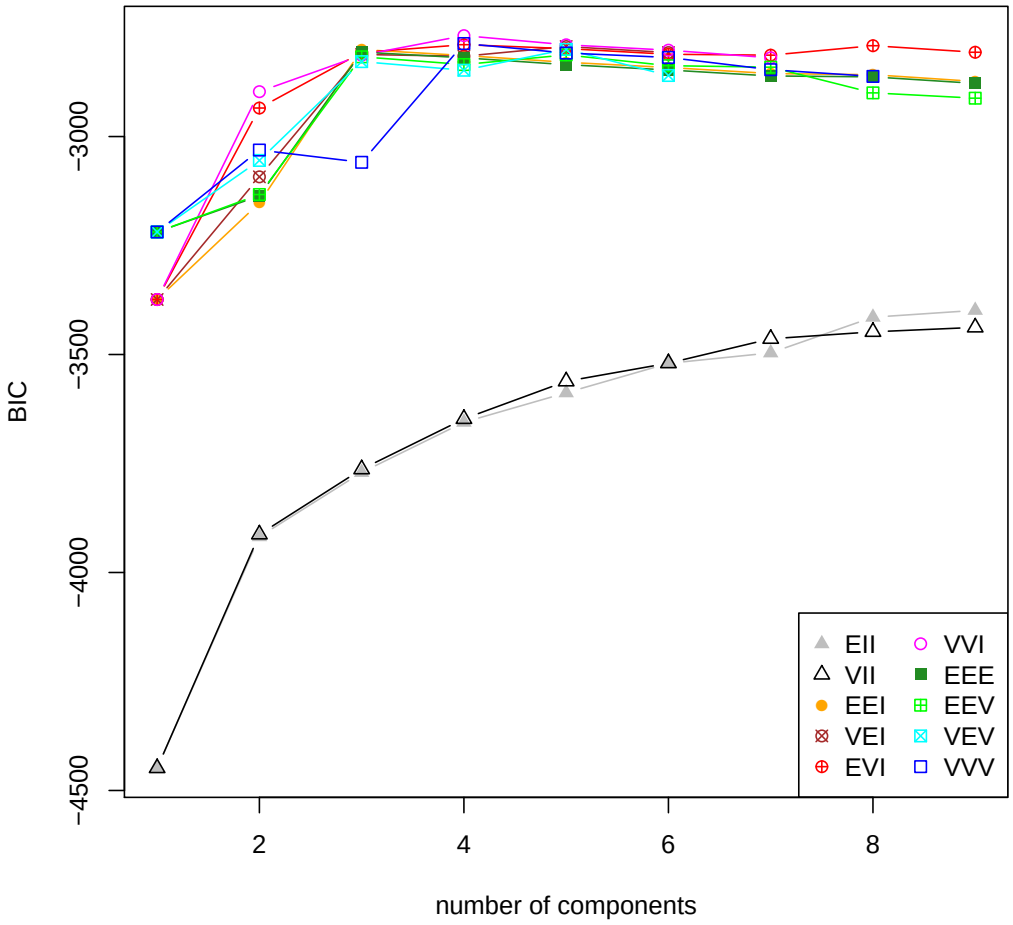
Back to Old Faithful

- We used `Mclust` to fit a mixture model to the Old Faithful data.

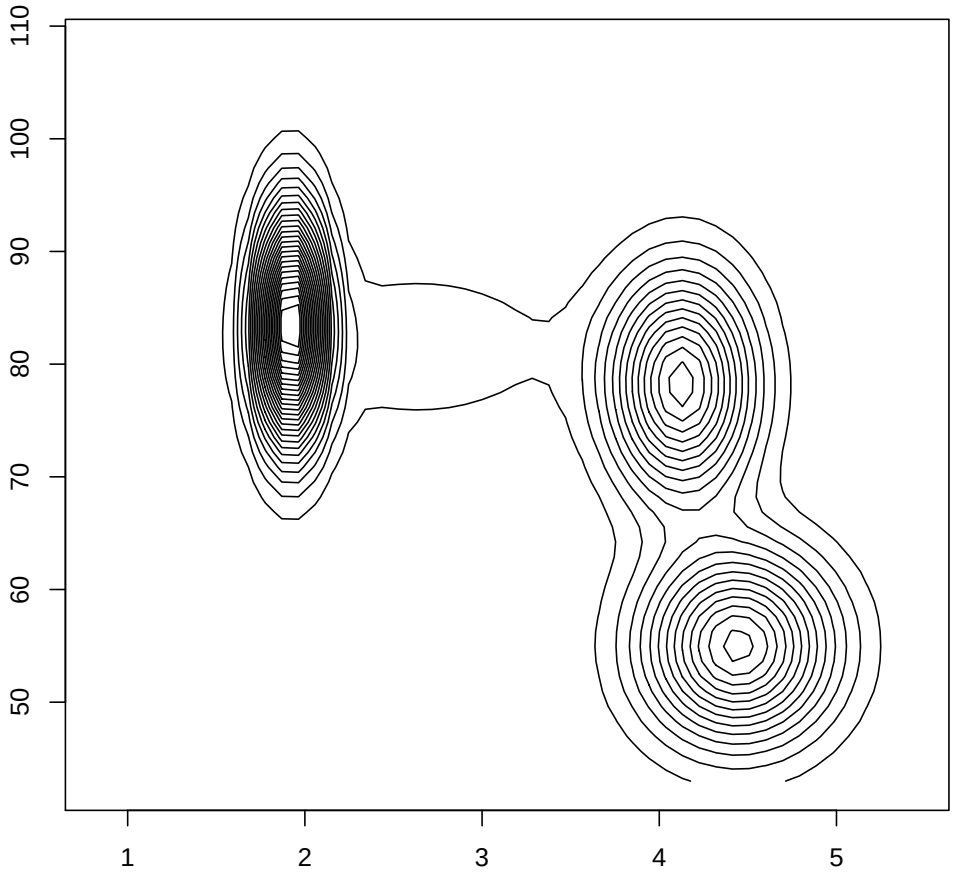
- The 4-component `VVI` model provided the best fit (although the 3-component model was also good). A `VVI` model is one where components have variable volume and shape but orientation in the direction of the main coordinate axes.



BIC values


BIC
−3000
−3500
−4000
−4500
2
4
6
8
number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EEV
VEV
VVV

Countours of estimated normal components



Uncertainty about component membership