Model-Based Clustering

- Hierarchical clustering algorithms, k-means algorithms and others are exploratory methods and not based on formal models.

- Not necessarily a disadvantage since clustering is largely exploratory.

- Model-based clustering is an alternative. Banfield and Raftery (1993, *Biometrics*) is the classic reference.

- With R, you can install a package called `Mclust` to do clustering based on normal distributions.

Why Model-based Clustering?

- The basic idea is to use a mixture of two or more component distributions
  - each component distribution corresponds to a “cluster”.
- It provides a statistical tool for making probabilistic inference (e.g., estimation, prediction).
- It comes with probabilistic errors (i.e., uncertainty quantification).
- It can deal with heterogeneity in the data.

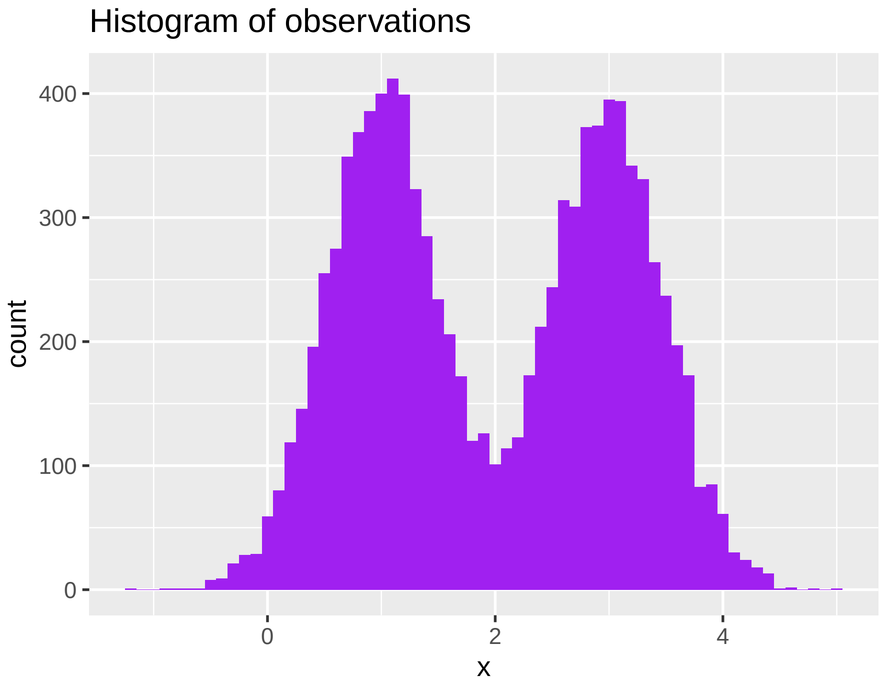
What is a finite mixture distribution?

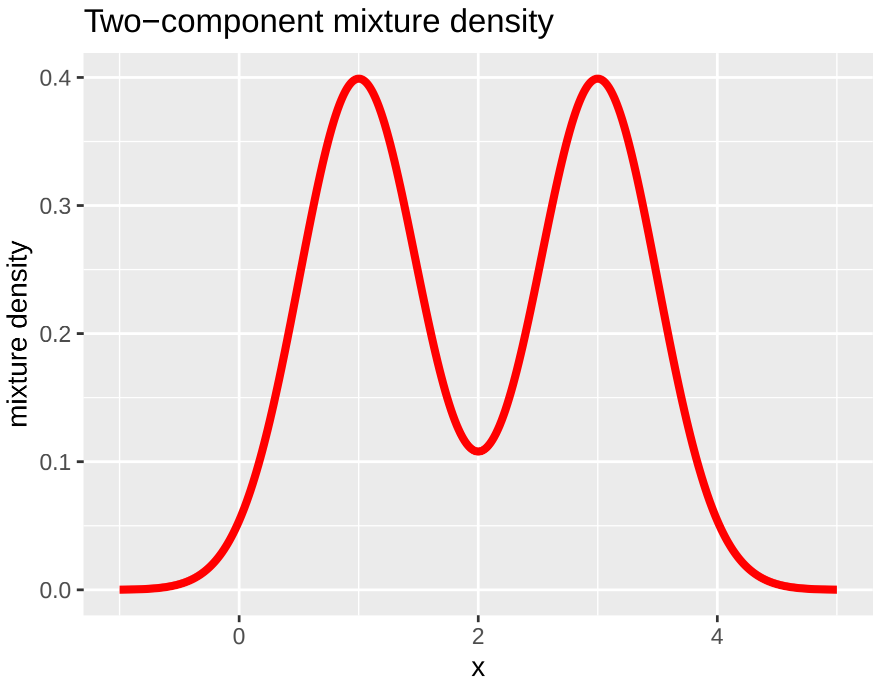
A random variable X is said to have a mixture distribution f with K components:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x).$$

- π_k is a mixing weight with $\sum_{k=1}^{K} \pi_k = 1$ and $0 < \pi_k < 1$.
  - The probability of selecting component k is π_k.
- f_k is a component density f_k (e.g., normal, Poisson).

What is a finite mixture distribution?

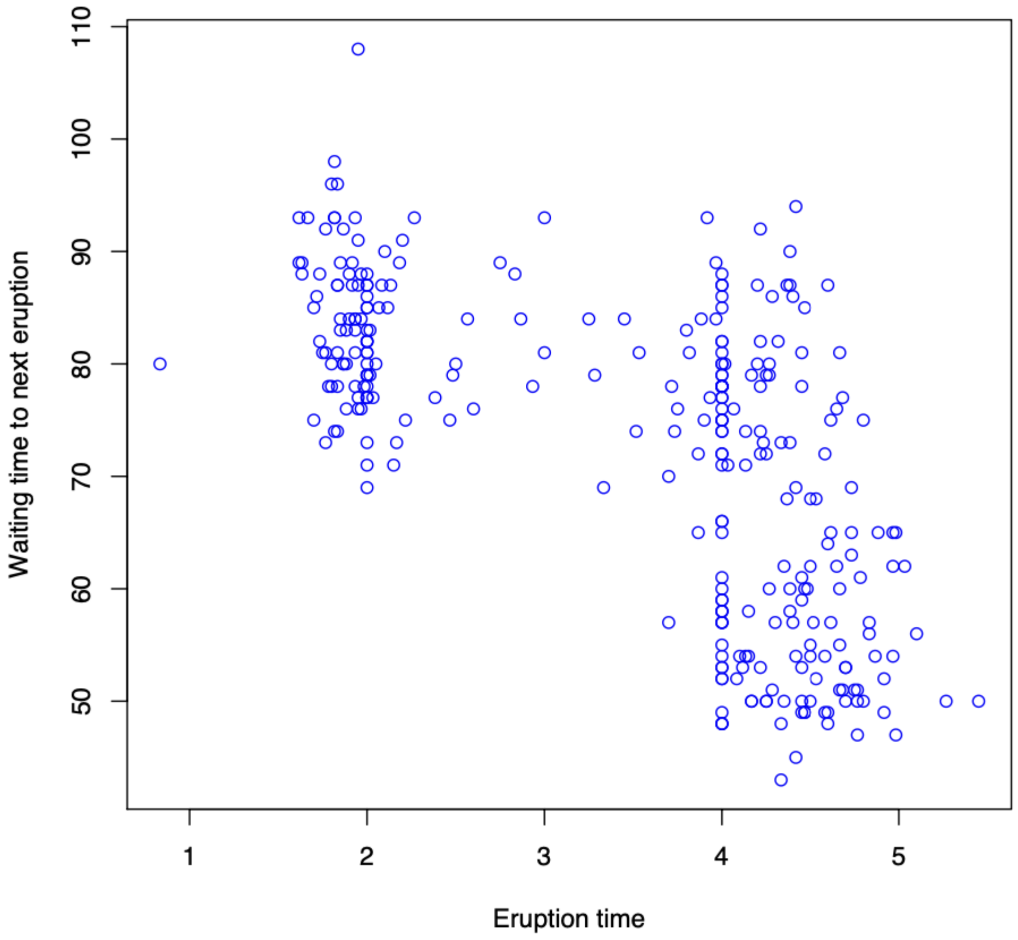

Histogram of observations


Two−component mixture density

Equivalent Representation of GMM

- Let z_i denote the latent component indicator or latent class/cluster for data point $\mathbf{x}_i$ with $z_i \overset{iid}{\sim} \mathsf{Cat}(\boldsymbol{\pi})$.
- $z_i = k$ means that the data point $\mathbf{x}_i$ falls into cluster k.
- An equivalent representation of GMM is that $x_i|z_i = k \sim f_k$.

Old Faithful Eruptions



Gaussian mixture models (GMM)

- Data: $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n **continuous** observations in p-dimensional space.

- Assume a Gaussian mixture model for the data:

$$\mathbf{x}_i \overset{ind}{\sim} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i | \mu_k, \Sigma_k), i = 1, \ldots, n,$$

  where $f_k(\mathbf{x}_i | \mu_k, \Sigma_k)$ is a p-dim multivariate normal distribution.

- Parameter estimation are usually based on the likelihood function (e.g., EM algorithm).

- What is the mixture model likelihood function?

More on GMM

- The components or clusters in the mixture model have ellipsoidal shapes and are centered at μ_k.

- The K component distributions need not have homogeneous covariance matrices.

- Σ_k is usually parameterized as $\Sigma_k = \nu_k D_k \Delta_k D_k'$, where
  - D_k is the orthogonal matrix of eigenvectors of Σ_k that describes the orientation of axes.
  - Δ_k is a diagonal matrix with elements proportional to the eigenvalues of Σ_k that determines the shape of the density contours
  - ν_k is a scalar that determines the volume of the ellipsoid (proportional to $\nu_k^p|\Delta_k|$).

Fit GMMs with the Mclust Algorithm

- The model-based clustering can be implemented using the 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 K clusters have the same volume, shape and orientation in $p-$dimensional space.
  - `VEI` means variable volume, same shape and orientation parallel to the coordinate axes.
  - `EIV` means same volume, spherical shape and variable orientation.

The Mclust Algorithm

There are a total of 10 combinations of volume, shape and orientation of covariance matrices available in the package:

EEE Ellipsoidal, equal volume, equal shape, equal orientation

EEV Ellipsoidal, equal volume, equal shape, varying orientation

VEV Ellipsoidal, equal shape, varying size and orientation

VVV Ellipsoidal, varying volume, shape and orientation

EEI Ellipsoidal, equal volume and shape, axes parallel to coordinate axes

. VEI Ellipsoidal, varying volume, equal shape, axes parallel to coord. axes

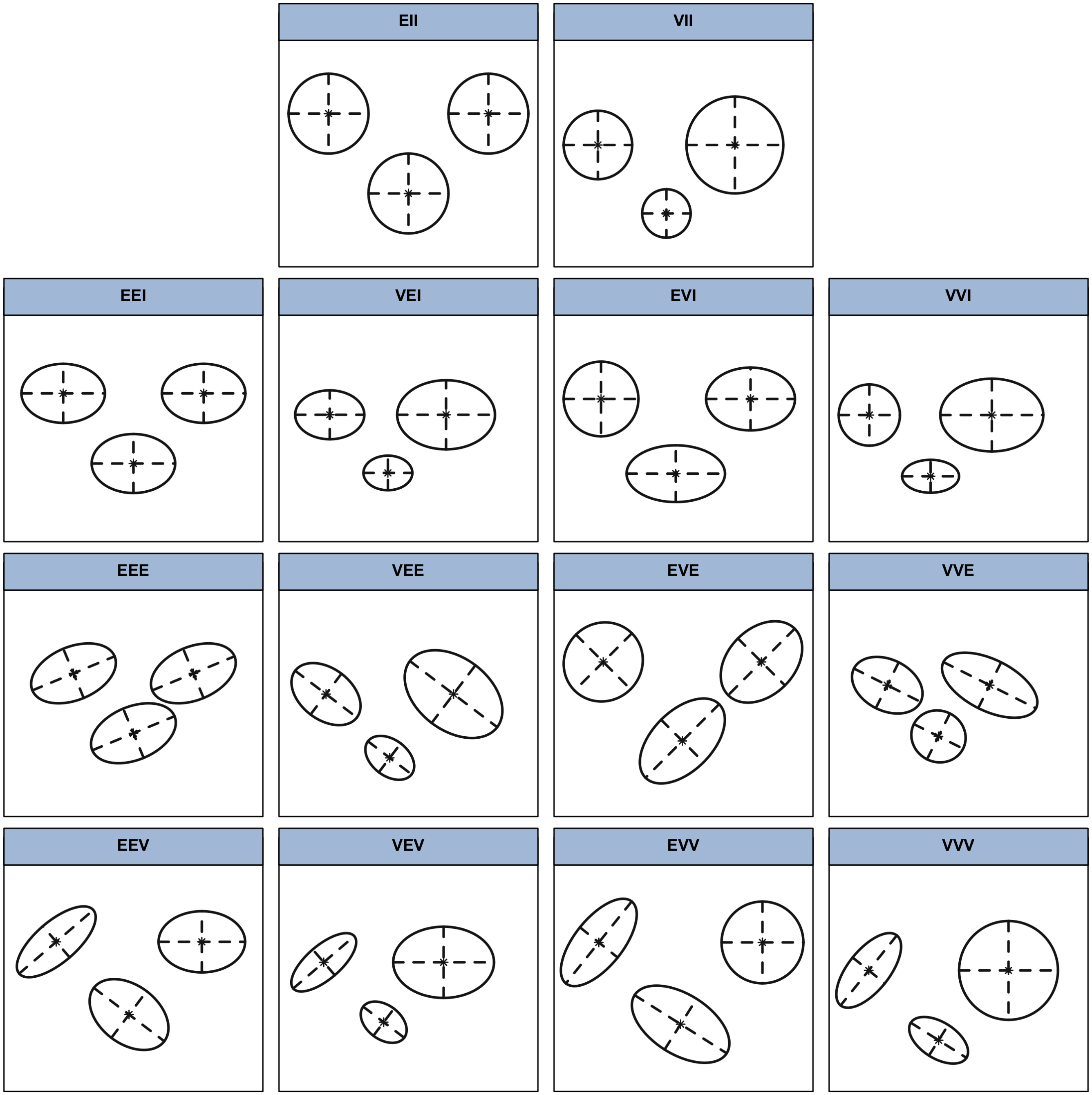
EVI Ellipsoidal, equal volume, varying shape, axes parallel to coord. axes

VVI Ellipsoidal, varying volume and shape, axes parallel to coord. axes

EII Spherical, equal volume

VII Spherical, varying volume

These place restrictions on possible sizes, shapes and orientation of contours of constant density.


EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV

The Mclust Algorithm

- Given the number of cluster, K, how do we choose a model?
- Any model selection criterion (AIC, BIC, ...) can be used to select the best model for a specified number of cluster K.
- Mclust uses the Bayesian Information Criterion (BIC, or Schwarz Information Criterion) to choose the best model for a given K.

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

  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 $2m$. BIC is large for the best cluster models.

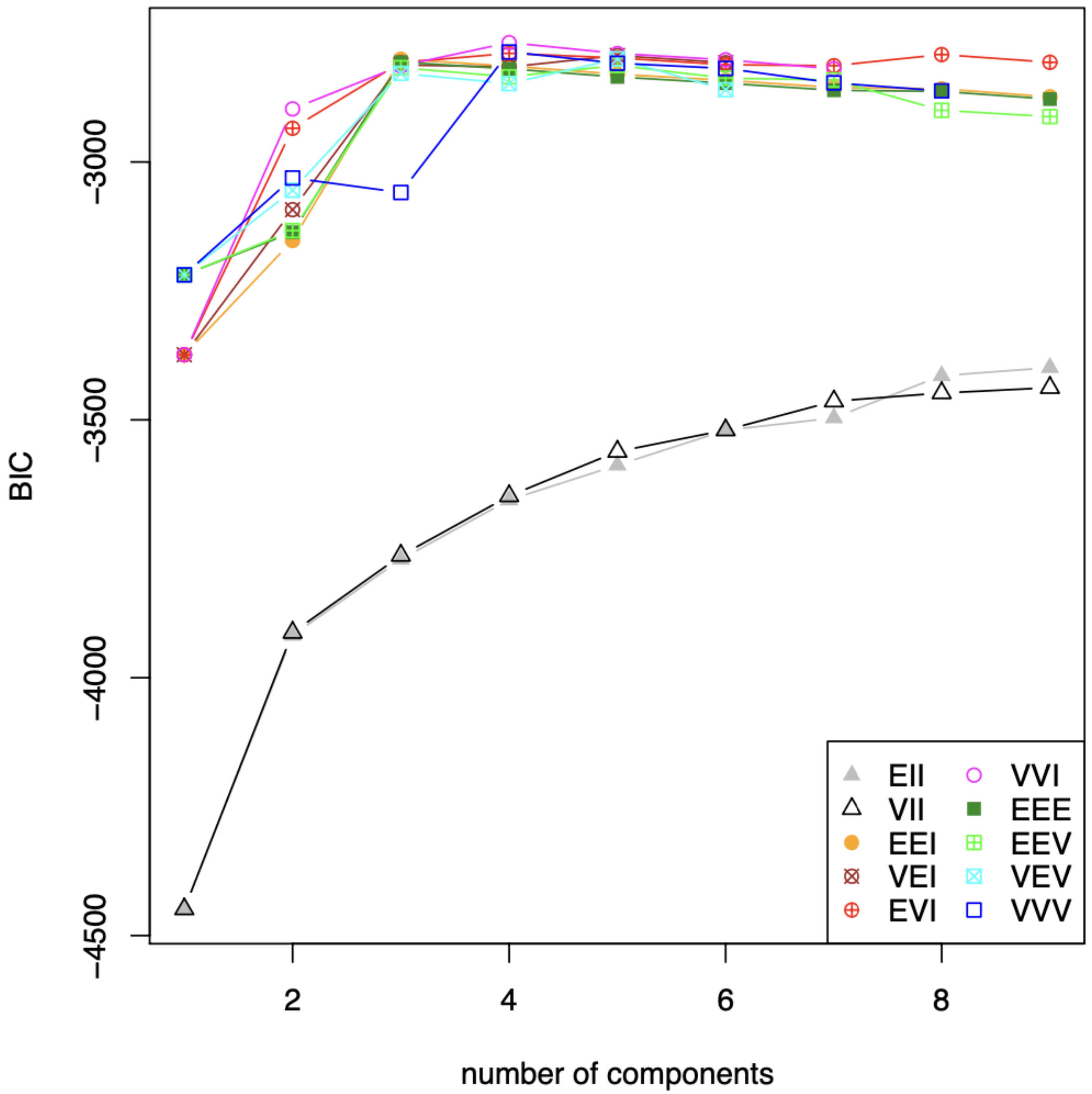
Back to Old Faithful

- We will use Mclust to fit a mixture model to the eruption data.

- The best fitting model had four components (although the model with three components was also good) and was `VVI`, meaning components had variable volume and shape but they are orientated in directions parallel to of the coordinate axes.

- The BIC plot is displayed on the next page, and a table of BIC values for all possible models is shown after that.

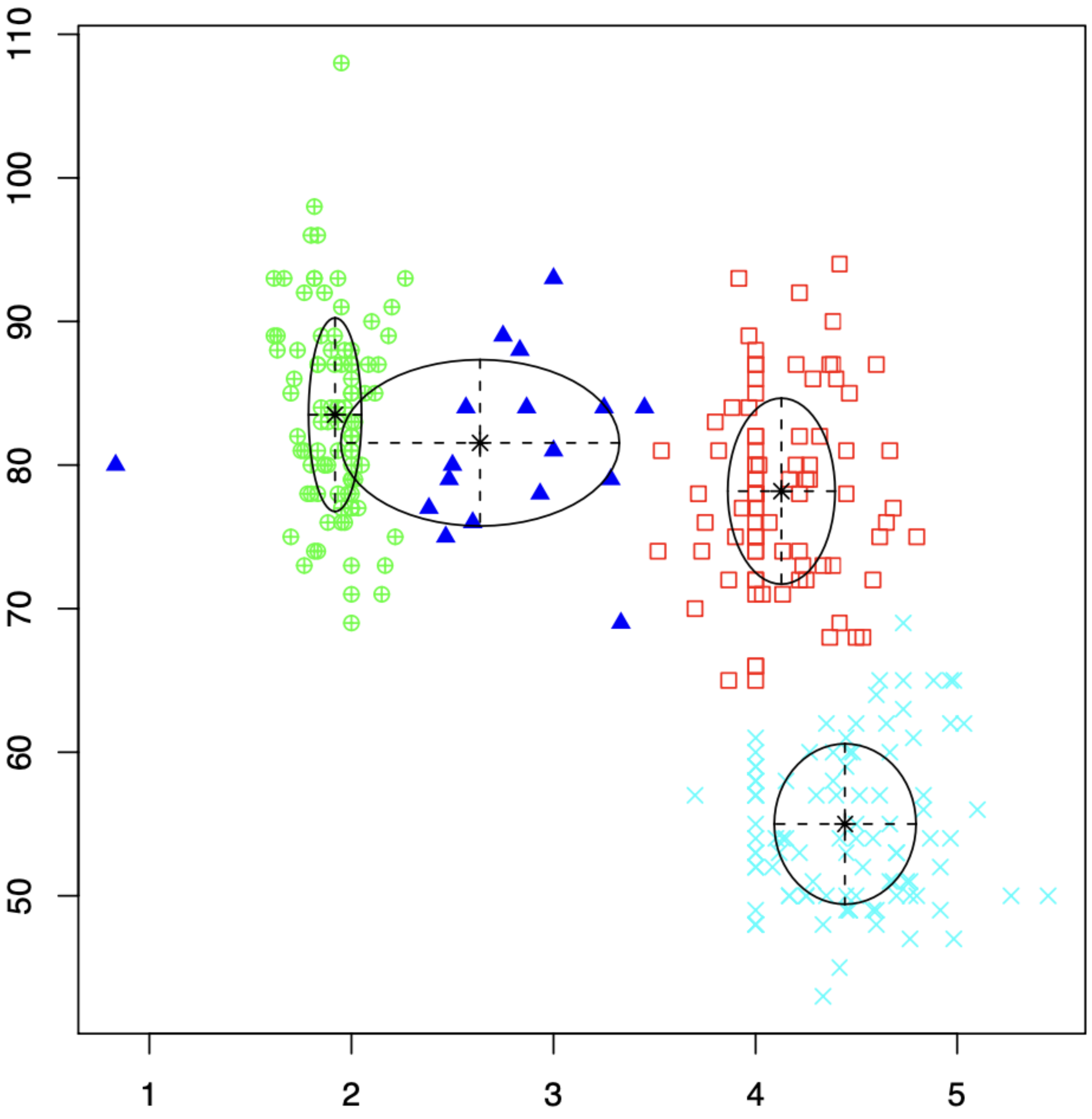
BIC Plot


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

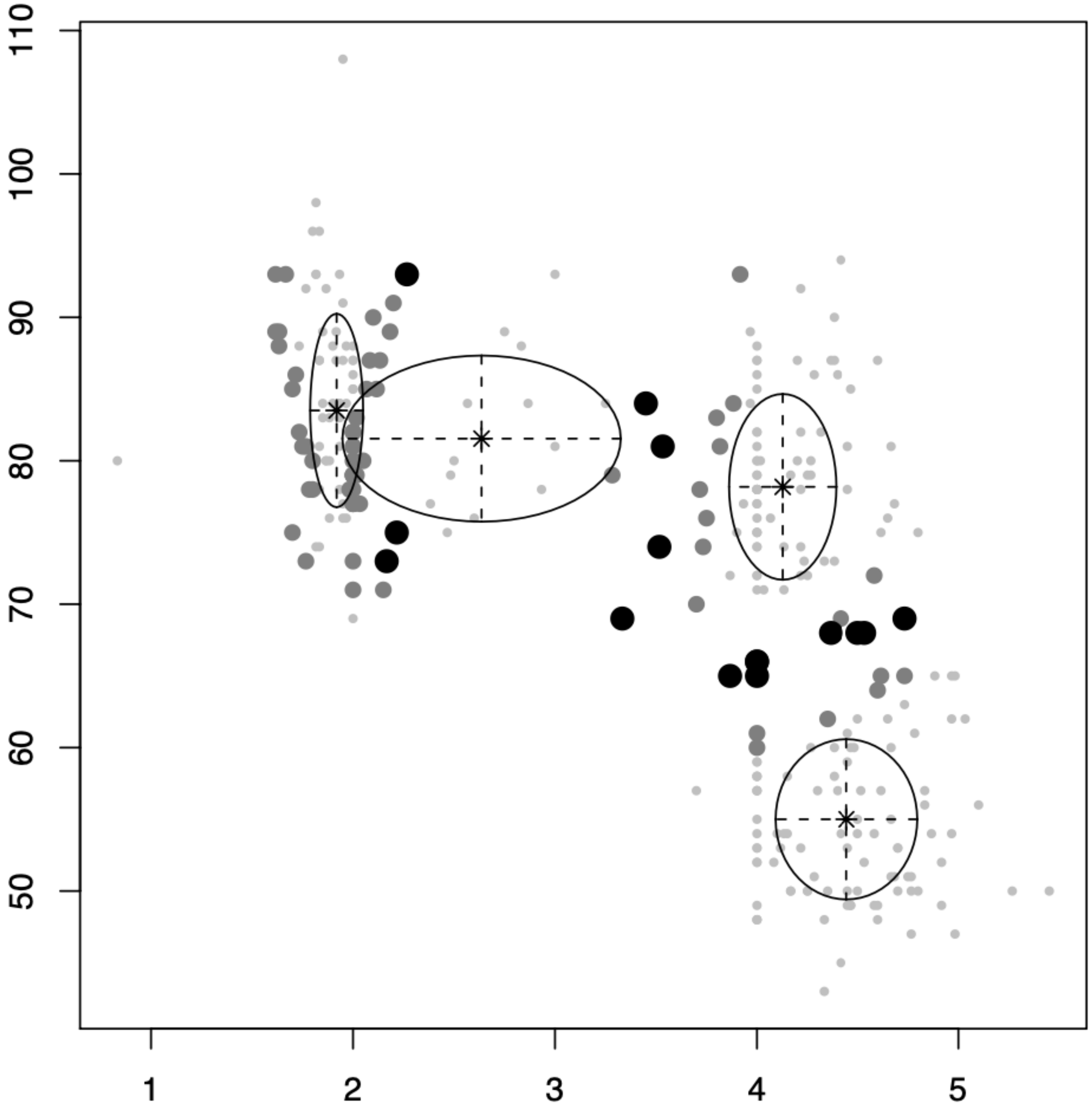
BIC Table

```
     "EII" "VII" "EEI" "VEI" "EVI" "VVI" "EEE" "EEV" "VEV" "VVV"
"1" -4449 -4449 -3374 -3374 -3374 -3374 -3219 -3219 -3219 -3219
"2" -3918 -3913 -3152 -3092 -2935 -2897 -3136 -3133 -3055 -3031
"3" -3771 -3763 -2801 -2812 -2807 -2814 -2806 -2817 -2828 -3059
"4" -3655 -3648 -2814 -2816 -2789 -2769 -2819 -2835 -2848 -2787
"5" -3588 -3561 -2830 -2794 -2798 -2790 -2835 -2813 -2800 -2808
"6" -3520 -3520 -2842 -2807 -2811 -2802 -2847 -2837 -2860 -2819
```

Normal Contour Plot



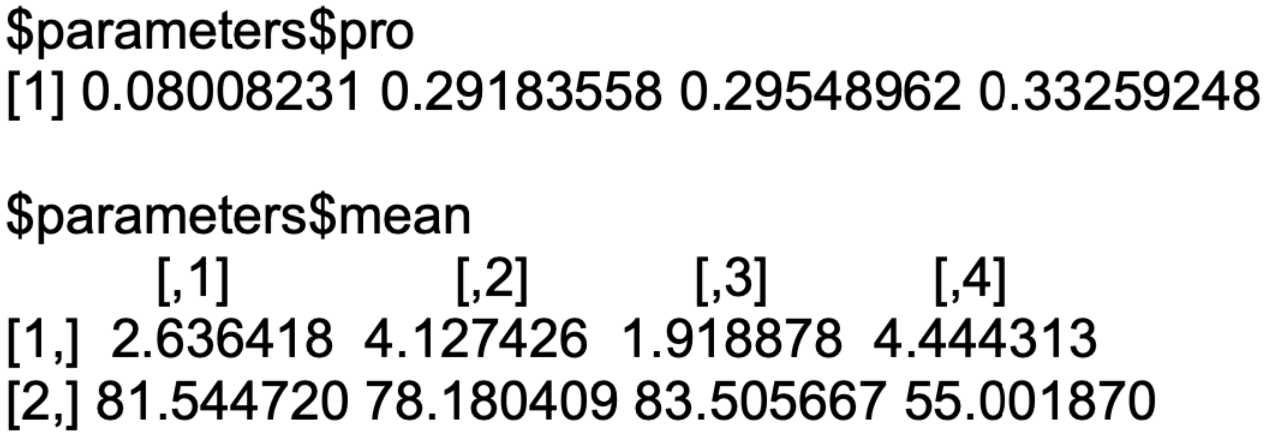
Uncertainty Plot



Estimated Parameters

- Mclust provides estimates of the $\pi_k, \mu_k, \nu_k, \Delta_k, D_k$ for each component. To get an estimate of Σ_k we have to reconstruct it from its components.

- For Old Faithful:



```
$parameters$pro
[1] 0.08008231 0.29183558 0.29548962 0.33259248

$parameters$mean
          [,1]      [,2]      [,3]      [,4]
[1,]  2.636418  4.127426  1.918878  4.444313
[2,] 81.544720 78.180409 83.505667 55.001870
```

Clustering with GMMs

- Given $\pi_k, \mu_k, \Sigma_k, k = 1, \ldots, K$, a GMM defines a joint distribution over $\mathbf{x}_i, z_i$.
- Using Bayes' rule, the conditional distribution of $z_i = k$ given the data $\mathbf{x}_i$ is

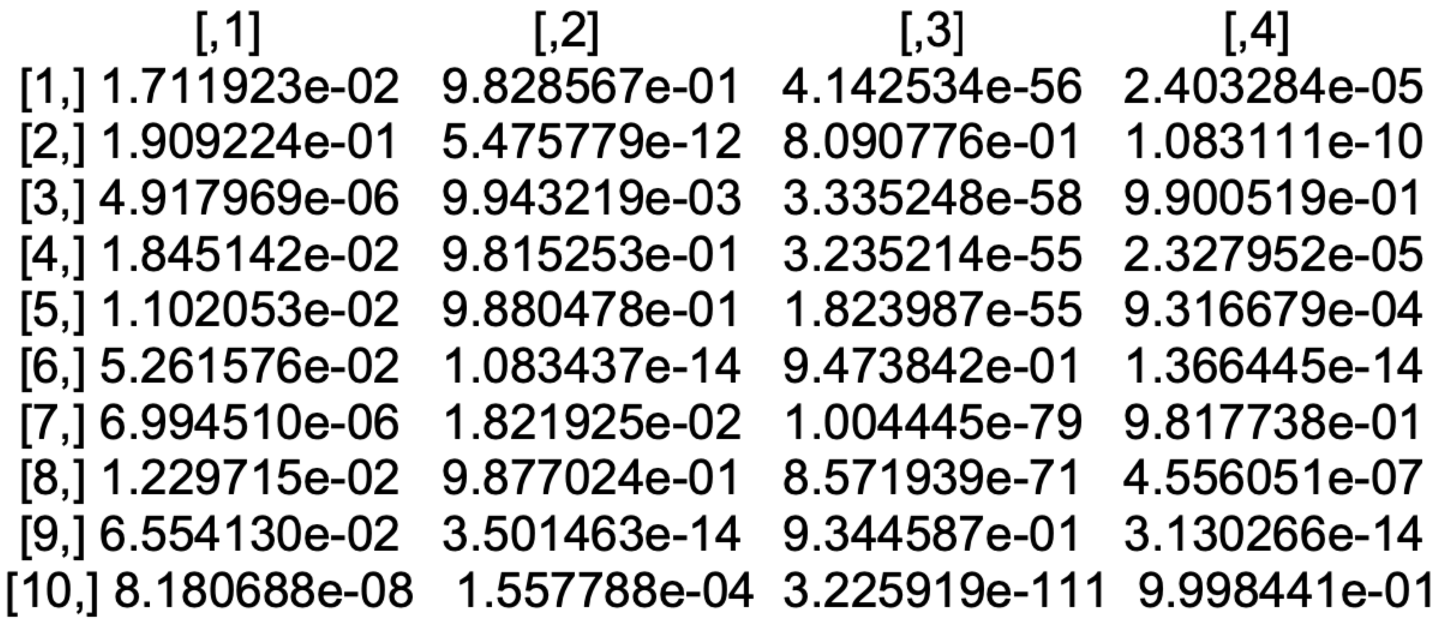
$$P(z_i = k|\mathbf{x}_i) = \frac{P(z_i = k)p(\mathbf{x}_i|z_i = k)}{p(\mathbf{x}_i)} = \frac{\pi_k f_k(\mathbf{x}_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i|\mu_k, \Sigma_k)}.$$

- This conditional distribution assigns some probability to $\mathbf{x}_i$ belonging to each cluster, also known as *soft clustering*.

- To obtain a *hard clustering* (an assignment of $\mathbf{x}_i$ to a single cluster), one typically selects a mode of the conditional distribution argmax$_k P(z_i = k|\mathbf{x}_i)$.

Estimated Cluster Assignments

- For each observation $\mathbf{x}_i$, we can also get $p(z_i = k|\mathbf{x}), k = 1, ..., K$. For the first 10 observations:



```
               [,1]         [,2]          [,3]         [,4]
 [1,] 1.711923e-02 9.828567e-01  4.142534e-56 2.403284e-05
 [2,] 1.909224e-01 5.475779e-12  8.090776e-01 1.083111e-10
 [3,] 4.917969e-06 9.943219e-03  3.335248e-58 9.900519e-01
 [4,] 1.845142e-02 9.815253e-01  3.235214e-55 2.327952e-05
 [5,] 1.102053e-02 9.880478e-01  1.823987e-55 9.316679e-04
 [6,] 5.261576e-02 1.083437e-14  9.473842e-01 1.366445e-14
 [7,] 6.994510e-06 1.821925e-02  1.004445e-79 9.817738e-01
 [8,] 1.229715e-02 9.877024e-01  8.571939e-71 4.556051e-07
 [9,] 6.554130e-02 3.501463e-14  9.344587e-01 3.130266e-14
[10,] 8.180688e-08 1.557788e-04 3.225919e-111 9.998441e-01
```

Final Comments

- mclust is a package written by Fraley and Raftery. A description of the package can be found at `http://www.stat.washington.edu/mclust`.
- See the online introduction of mclust: `https://mclust-org.github.io/mclust-book/`
- For categorial data, one can use multinomial distributions as mixing distributions ⇒ Multinomial Mixture Model (MMM).