Model-based clustering
Gaussian mixture models

Erik-Jan van Kesteren & Daniel L. Oberski

Last week

- Hierarchical clustering
- K-means clustering
- Assessing cluster solutions
  - Stability
  - Internal metrics
  - External validation

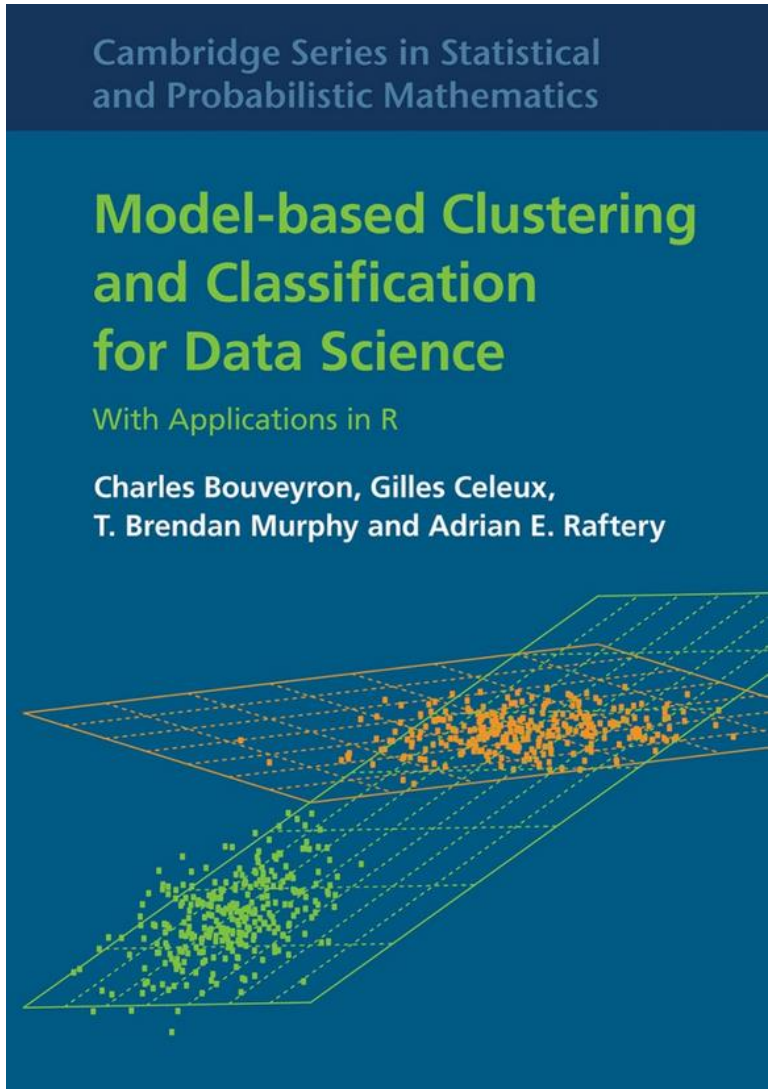
Today

- Model-based clustering
- Maximum likelihood estimation
- EM algorithm
- Multivariate model-based clustering
- Assumptions & restrictions

- **Goal: understand, apply, and assess model-based clustering methods**

Reading materials

- Mixture models: latent profile and latent class analysis (Oberski, 2016) http://daob.nl/wp-content/papercite-data/pdf/oberski2016mixturemodels.pdf
- MBCC sections 2.1 and 2.2



Model-based clustering

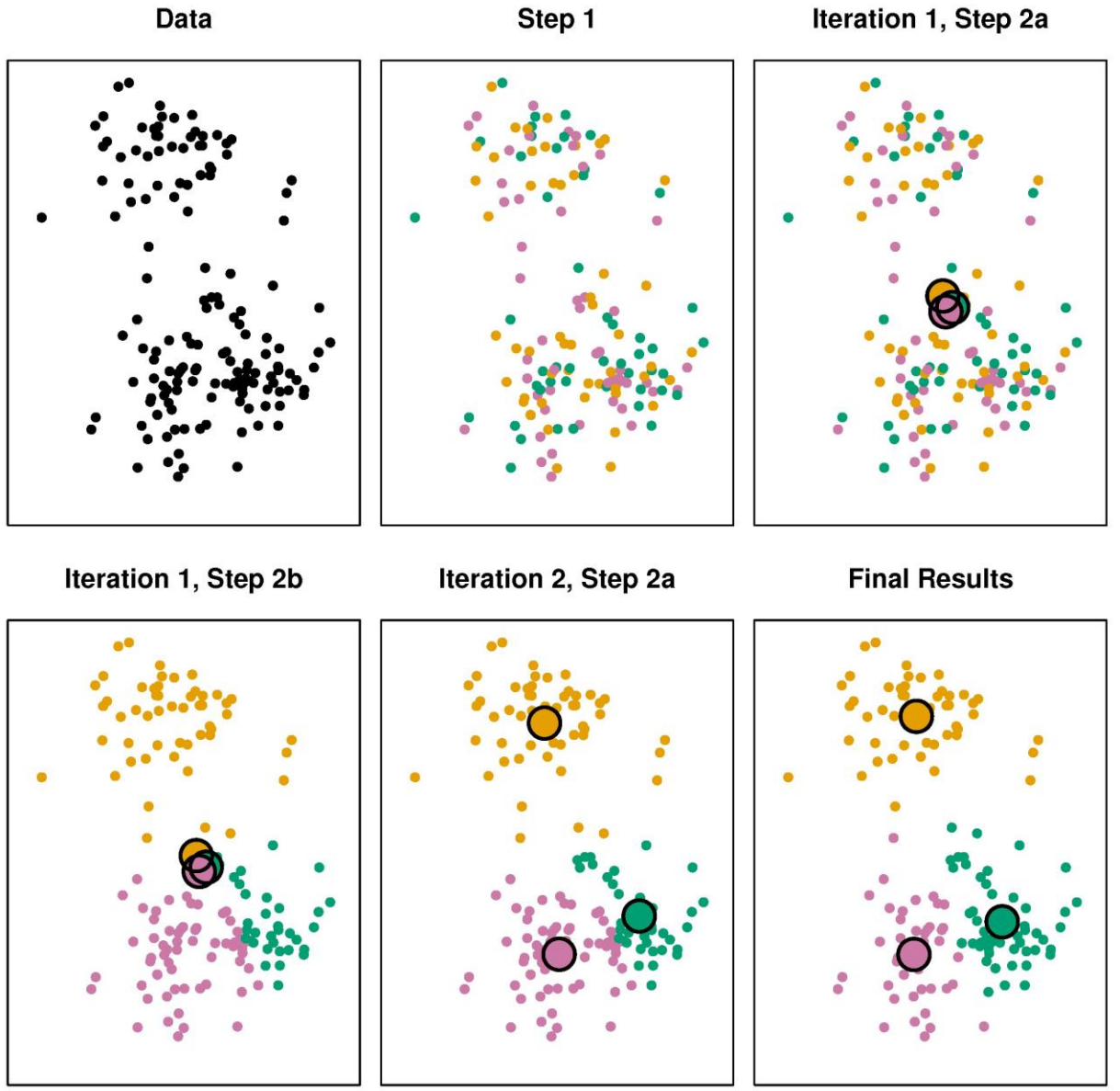
K-means again

1. Assign examples to K clusters

2. a. Calculate K cluster centroids;

   b. Assign examples to cluster with closest centroid;

3. If assignments changed, back to step 2a; else stop.



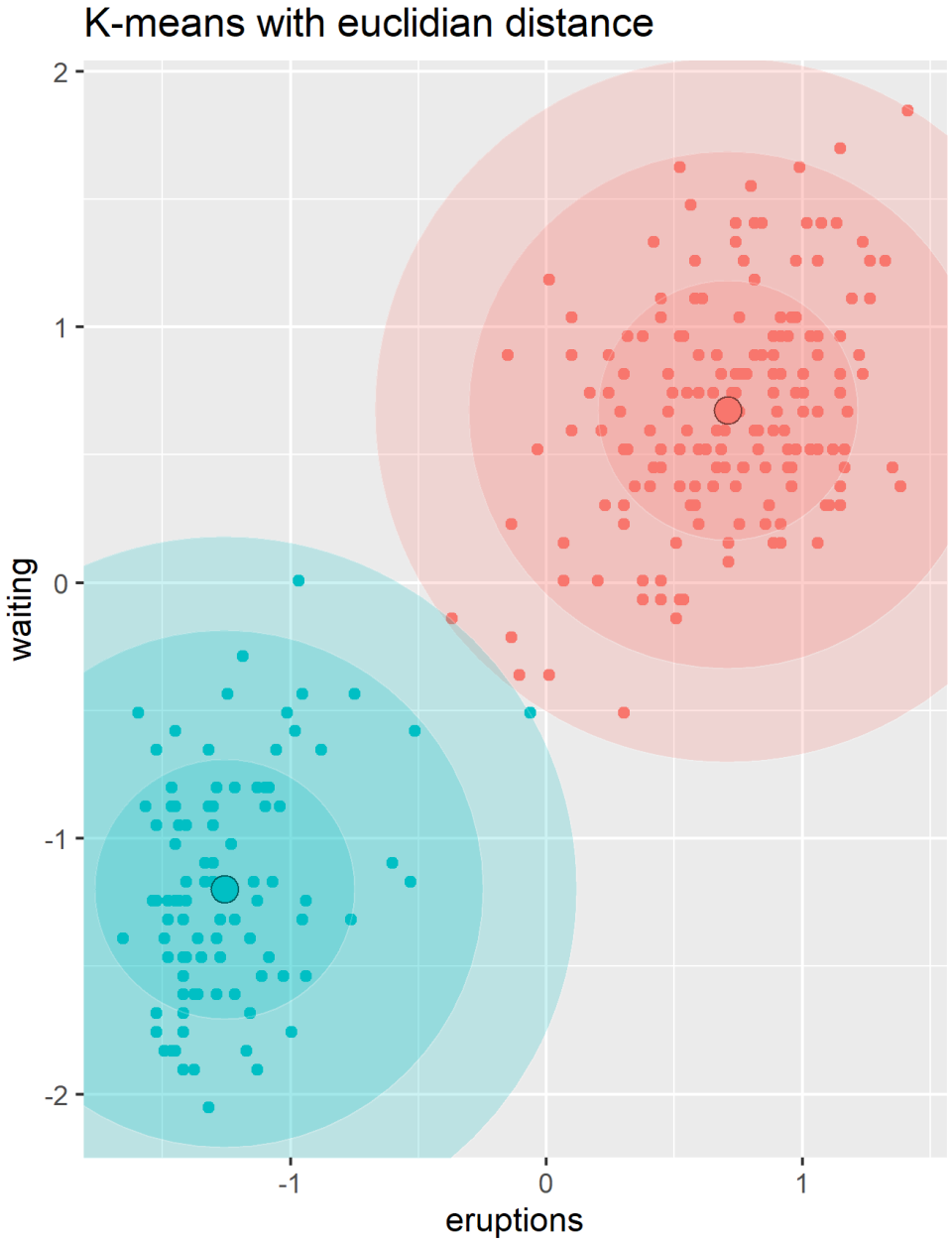
K-means again

- K-means is based on a **rule**
- Why this rule and not some other rule?
- What kind of data does the rule work well for?
- In what situations would the rule fail?
- What happens if we want to change the rule?

All **difficult to answer by staring at the algorithm.**

K-means again

- k-means algorithm makes clusters which are **circular in the space of the data**.
- Is this reasonable?
- Maybe x and y covary within the clusters, in the same way or even differently?
- Maybe we need ellipses?



Model-based clustering

Steps:

1. Pretend we believe in some *statistical model* that describes data as belonging to unobserved ("latent") groups;
2. Estimate ("train") this model using the data.

The rule follows from the model!

- Instead of worrying about *algorithm*, we worry about model.
- Earlier mentioned questions are easier to answer.

Model-based clustering

- Assumptions about the clusters are explicit, not implicit.
- We will look at the most used family of models:

Gaussian mixture models (GMMs)

- Data within each cluster *(multivariate) normally distributed*.
- Parameters can be either the same or different across groups:
  - Volume (size of the clusters in data space);
  - Shape (circle or ellipse);
  - Orientation (the angle of the ellipse).

Model-based clustering

Another major advantage

- For each observation, get a posterior probability of belonging to each cluster
- Reflects that cluster membership is uncertain
- Cluster assignment can be done based on the highest probability cluster for each observation

Model-based clustering

Remember silhouette?

- a_i = avg. distance to fellow cluster members (cohesion)
- b_i = min. distance to member from different cluster (separation)

$$s_i = \frac{(b_i - a_i)}{\max(a_i,\ b_i)}$$

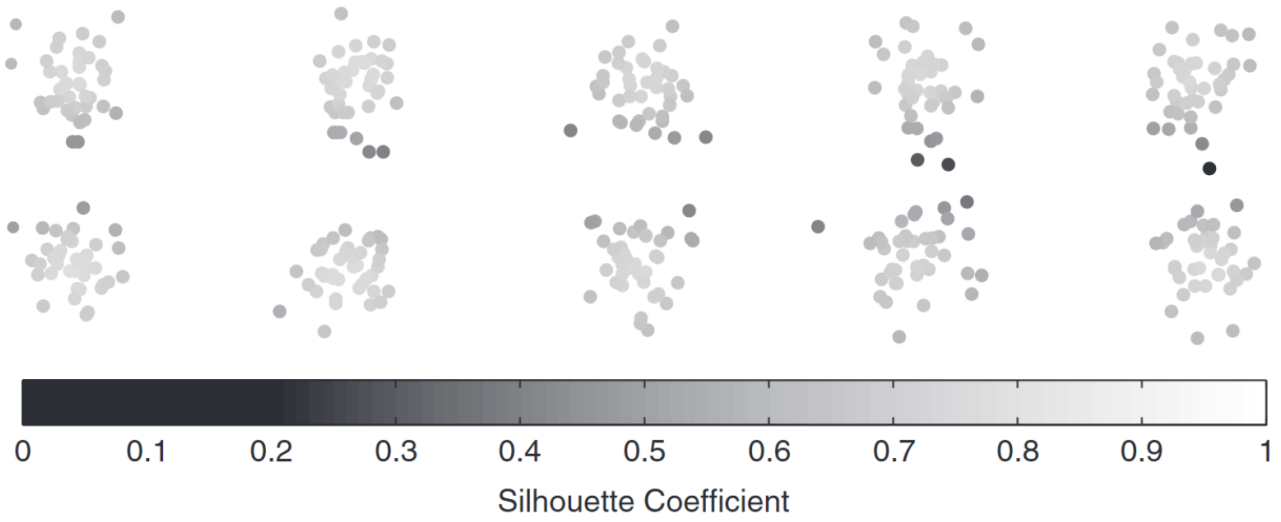


Figure 7.29. Silhouette coefficients for points in ten clusters.

Model-based clustering

Specific examples of model-based clustering:

- Gaussian mixture models
- Latent profile analysis
- Latent class analysis (categorical observations)
- Latent Dirichlet allocation

Gaussian mixture modelling

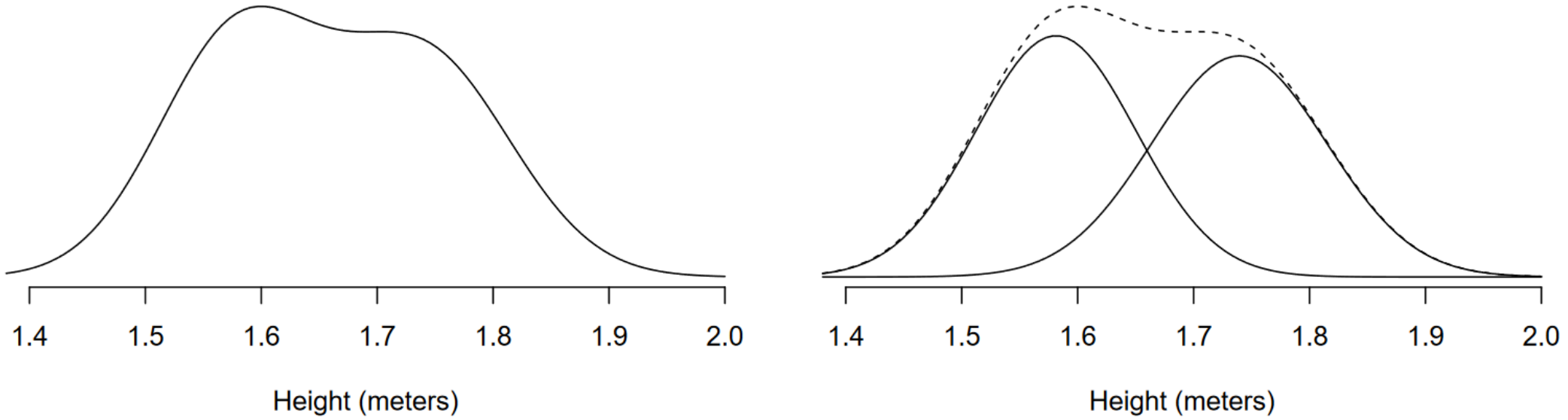


Fig. 1 Peoples' height. Left: observed distribution. Right: men and women separate, with the total shown as a dotted line.

Model-based clustering

- Statistical model + assumptions defines a **likelihood**:

$$p(data \mid parameters) = p(y \mid \theta)$$

- **Maximum likelihood estimation**: find the parameters θ for which it is most likely to observe this data
- This is how models can be estimated / fit / trained

- NB: the model and its assumptions are debatable!

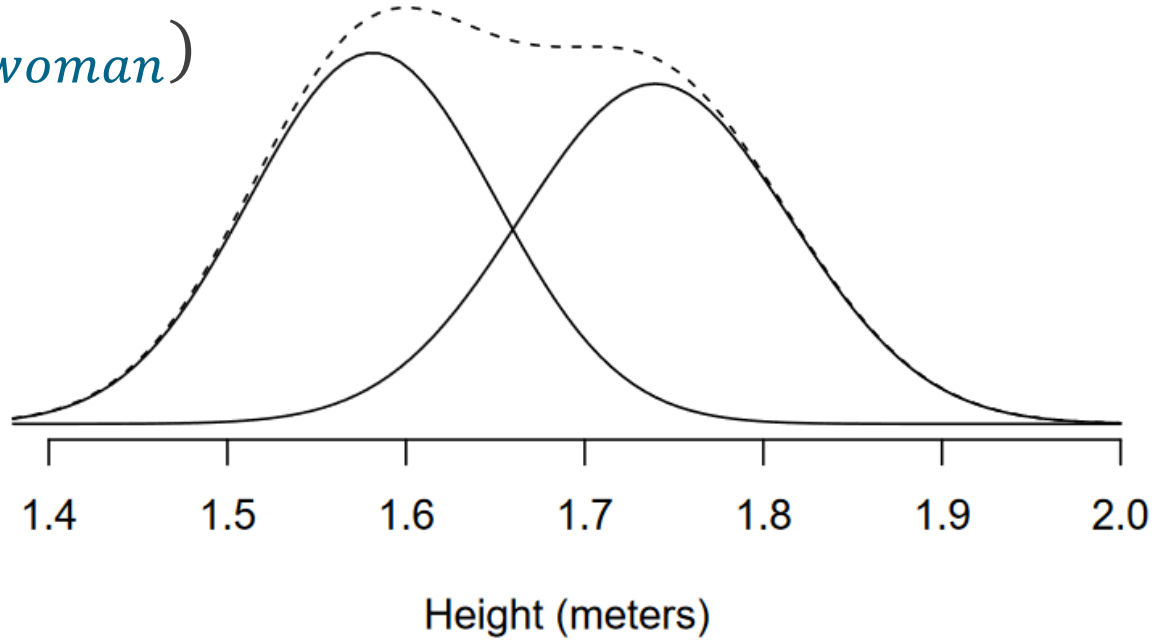
Model-based clustering

Likelihood (density) for height data:

$$p(height \mid \theta) = Pr(man)Normal(\mu_{man}, \sigma_{man}) + Pr(woman)Normal(\mu_{woman}, \sigma_{woman})$$

Or, in clearer notation:

$$p(height \mid \theta) = \pi_1^X Normal(\mu_1, \sigma_1) + (1 - \pi_1^X)Normal(\mu_2, \sigma_2)$$



Model-based clustering

Gaussian mixture parameters:

- π_1^X determines the relative cluster sizes
  - Proportion of observations to be expected in each cluster
- μ_1 and μ_2 determine the locations of the clusters
  - Like centroids in k-means clustering
- σ_1 and σ_2 determine the volume of the clusters
  - how large / spread out the clusters are in data space

Together, these 5 unknown parameters describe our model of how the data is generated.

Estimation: the EM algorithm

If we know who is a man and who is a woman, it's easy to find the maximum likelihood estimates for μ and σ:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N_1} height_i}{N_1}, \qquad \hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{N_1} (height_i - \hat{\mu}_1)^2}{N_1 - 1}}$$

(and same for $\hat{\mu}_2$ and $\hat{\sigma}_2$)

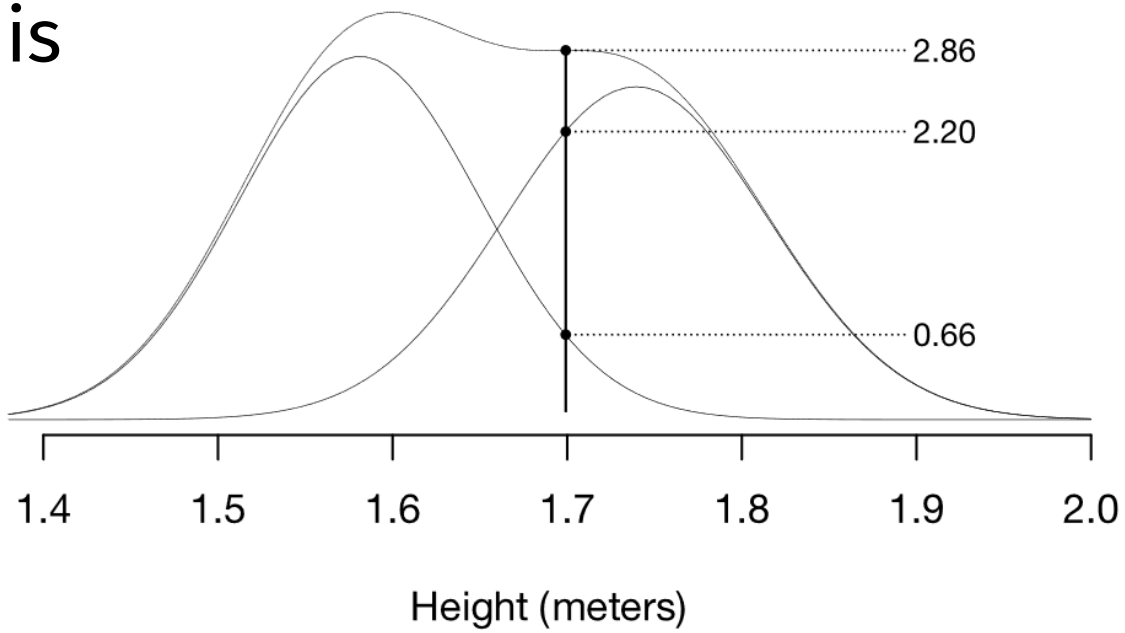
But we don't know this!

-> Assignments need to be estimated too.

Estimation: the EM algorithm

- Solution: Figure out the posterior probability of being a man/woman, given the current estimates of the means and sds
- If we know cluster locations and shapes, how likely is it that a 1.7m person is a man or a woman?

$$\pi_{man}^{X} = \frac{2.20}{2.86} \approx 0.77$$

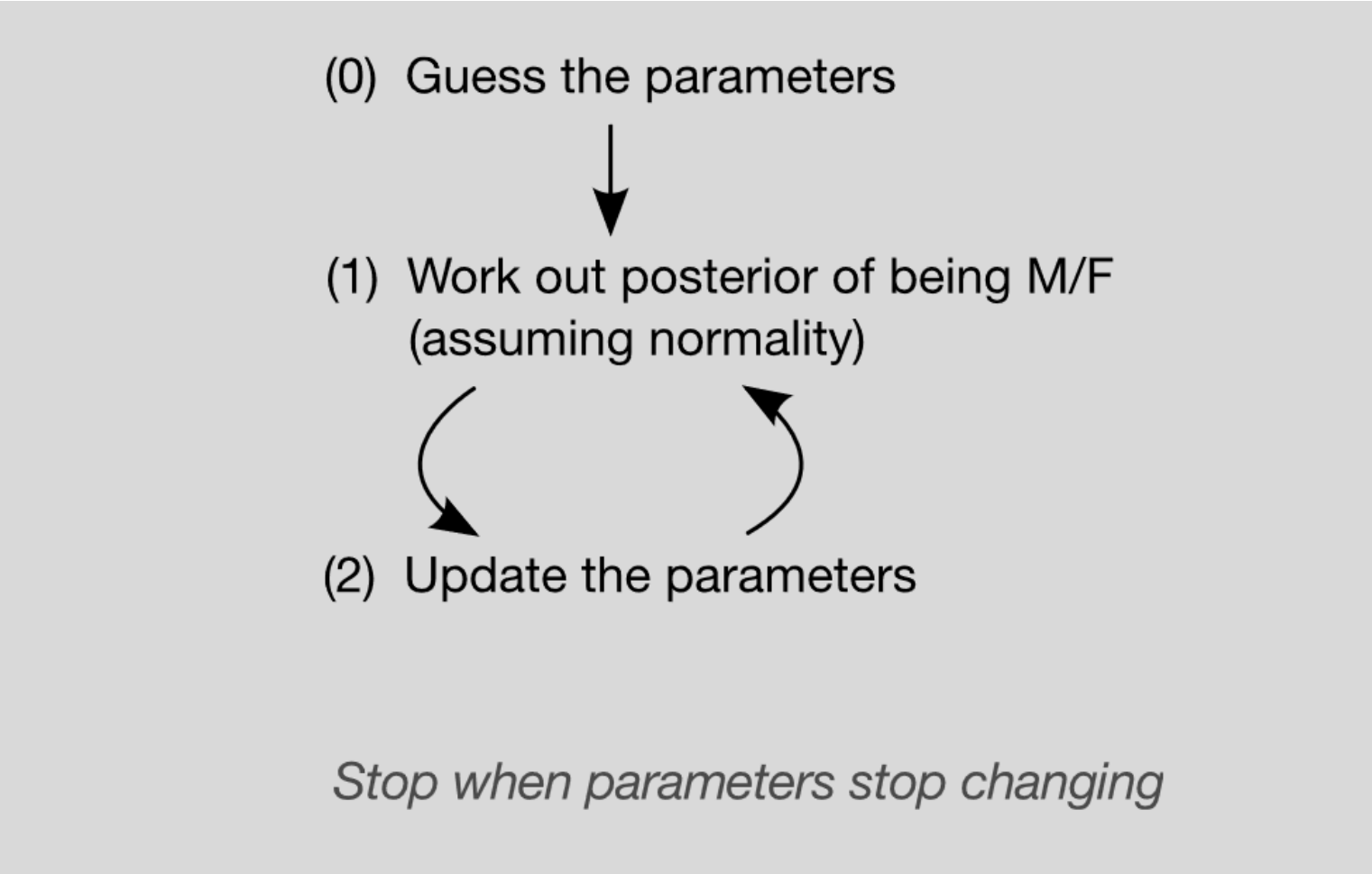


Estimation: the EM algorithm

- Now we have some class assignments (probabilities);
- So we can go back to the parameters and update them using our easy rule (M-step)
- Then, we can compute new posterior probabilities (E-step)

Does it remind you of something...?

Estimation: the EM algorithm



Live coding EM

Break

Multivariate model-based clustering

Multivariate model-based clustering

- With 2 observed features:
  - mean becomes a vector of 2 means
  - standard deviation turns into a 2x2 variance-covariance matrix determining the shape of the cluster
- So we have multiple within-cluster parameters:
  - Two means
  - Two variances, one for each observed variable
  - A single covariance among the features
- Together, the **11 parameters** define the likelihood in bivariate space, which from the top looks like ellipses

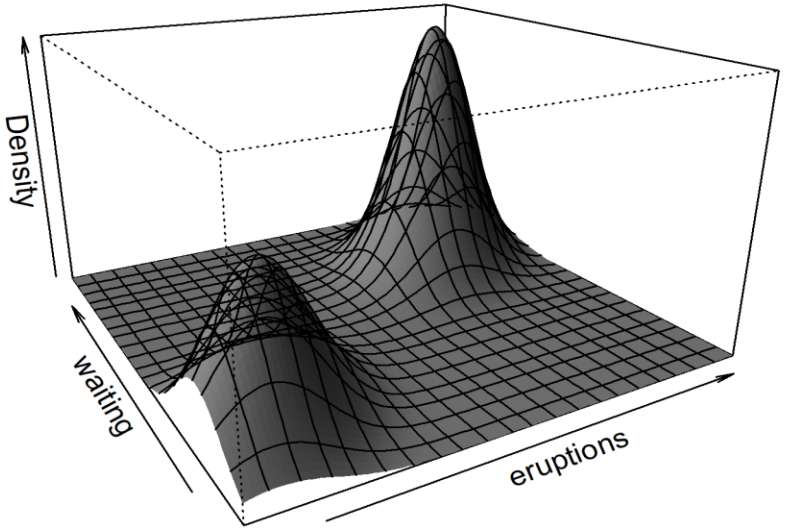
Multivariate normal distribution

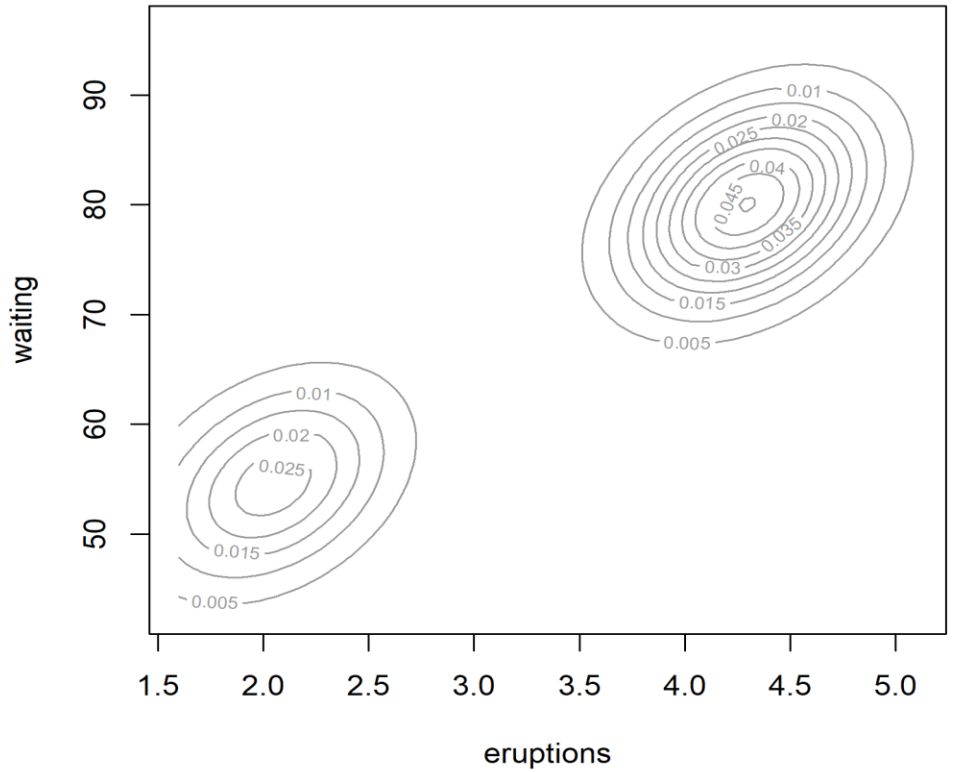
$$Normal(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$MVN(x;\mu,\sigma) = \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

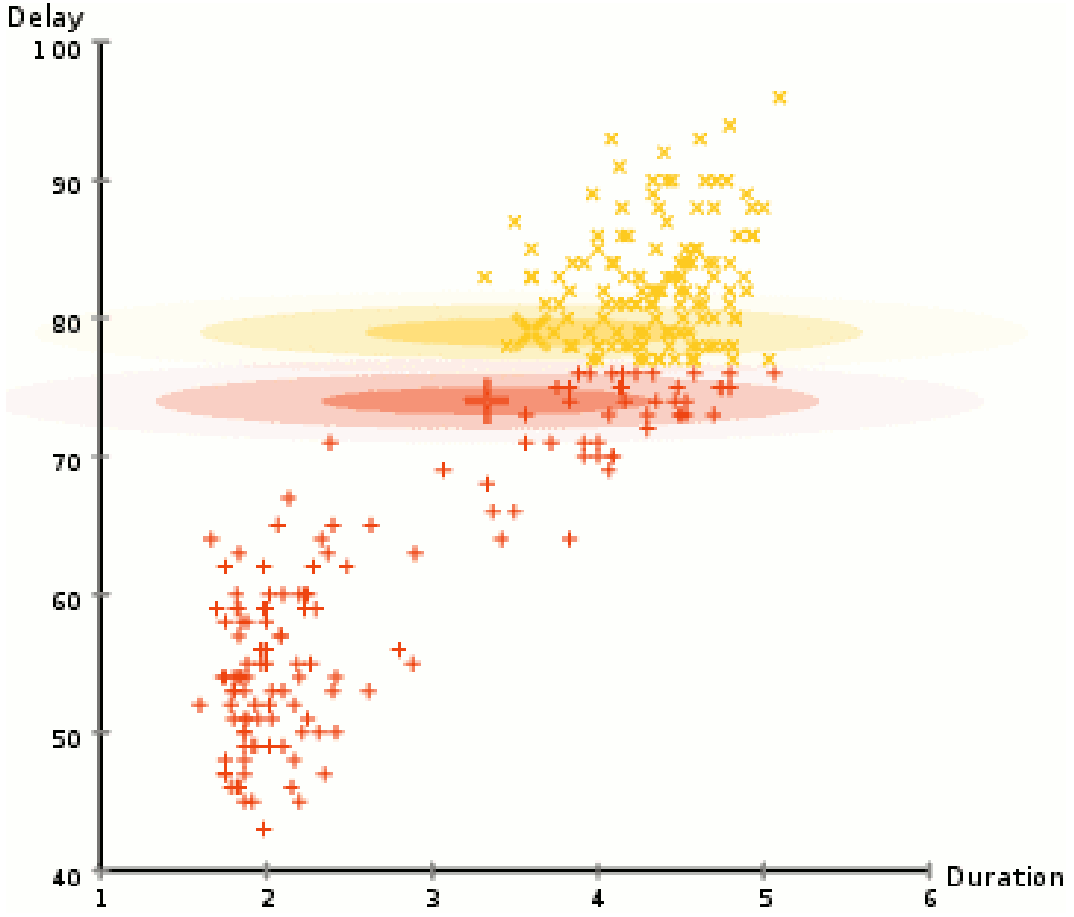
Multivariate model-based clustering

$$p(\boldsymbol{y}|\theta) = \pi_1^X MVN(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}) + \left(1 - \pi_1^X\right) MVN(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$$





Estimation: the EM algorithm



Multivariate model-based clustering

- Cluster shape parameters (the variance-covariance matrix) can be constrained to be equal across clusters
  - Same as k-means
- Can also be different across clusters
  - not possible in k-means
- More flexible, complex model
  - Think about the bias-variance tradeoff!

TOP SECRET SLIDE

- K-means clustering is a GMM with the following model:
  - All prior class proportions are 1/K
  - EII model: equal volume, only circles
  - All posterior probabilities are either 0 or 1

TOP SECRET SLIDE 2

- **G**MM has trouble with clusters that are not ellipses
- Secret weapon: **merging**

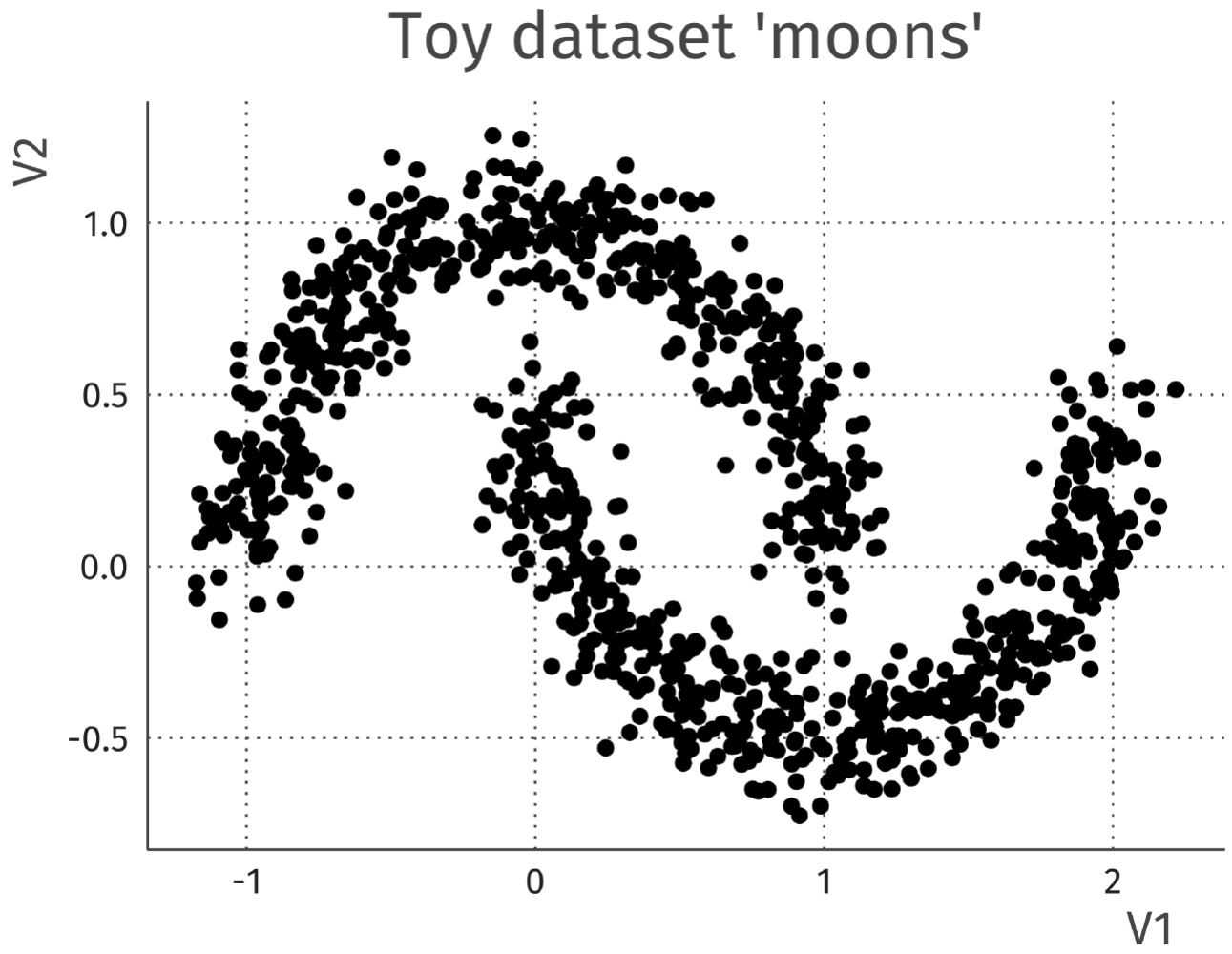
Powerful idea:

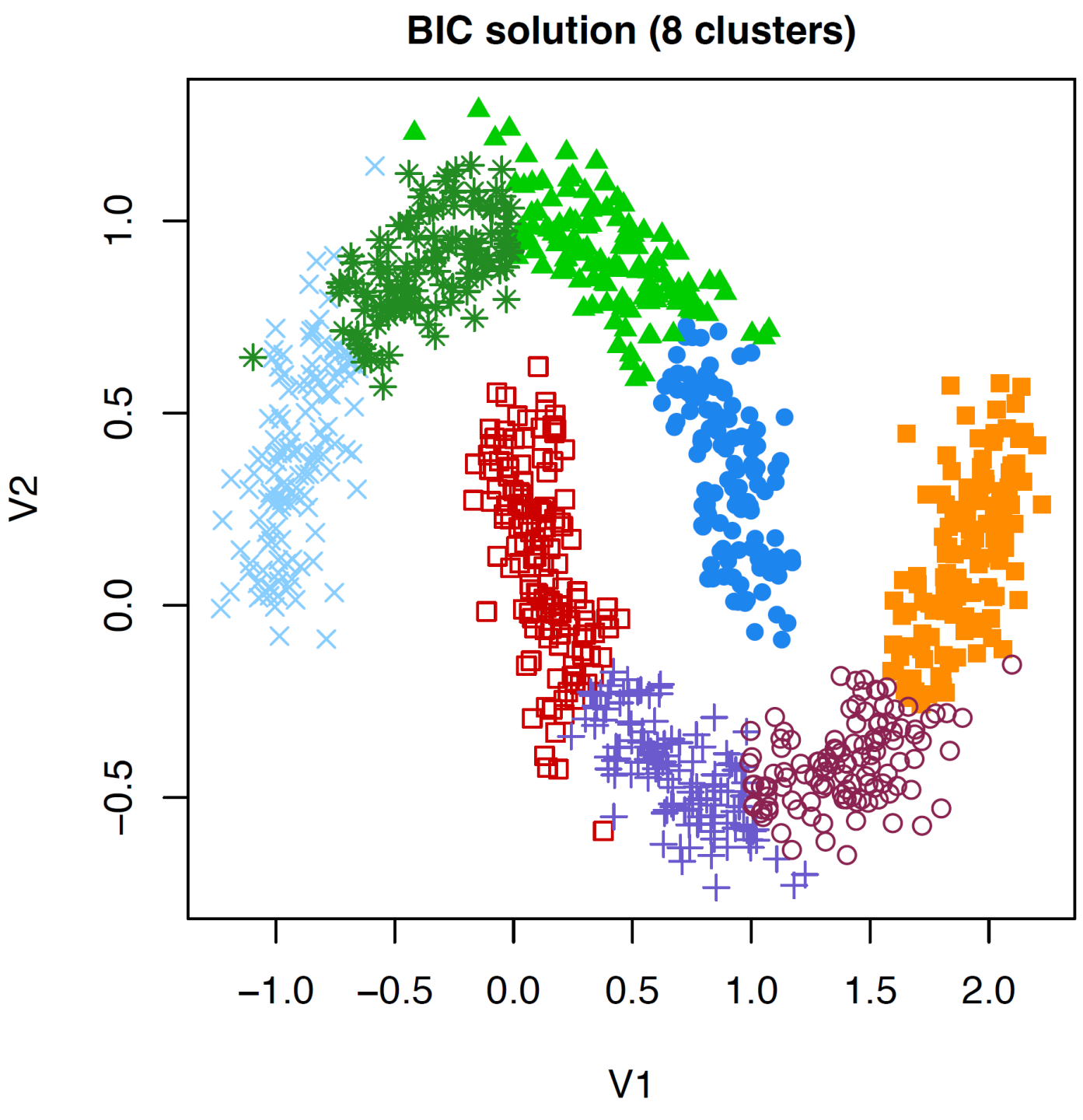
- Start with Gaussian mixture solution
- **Merge** “similar” components to create non-Gaussian clusters

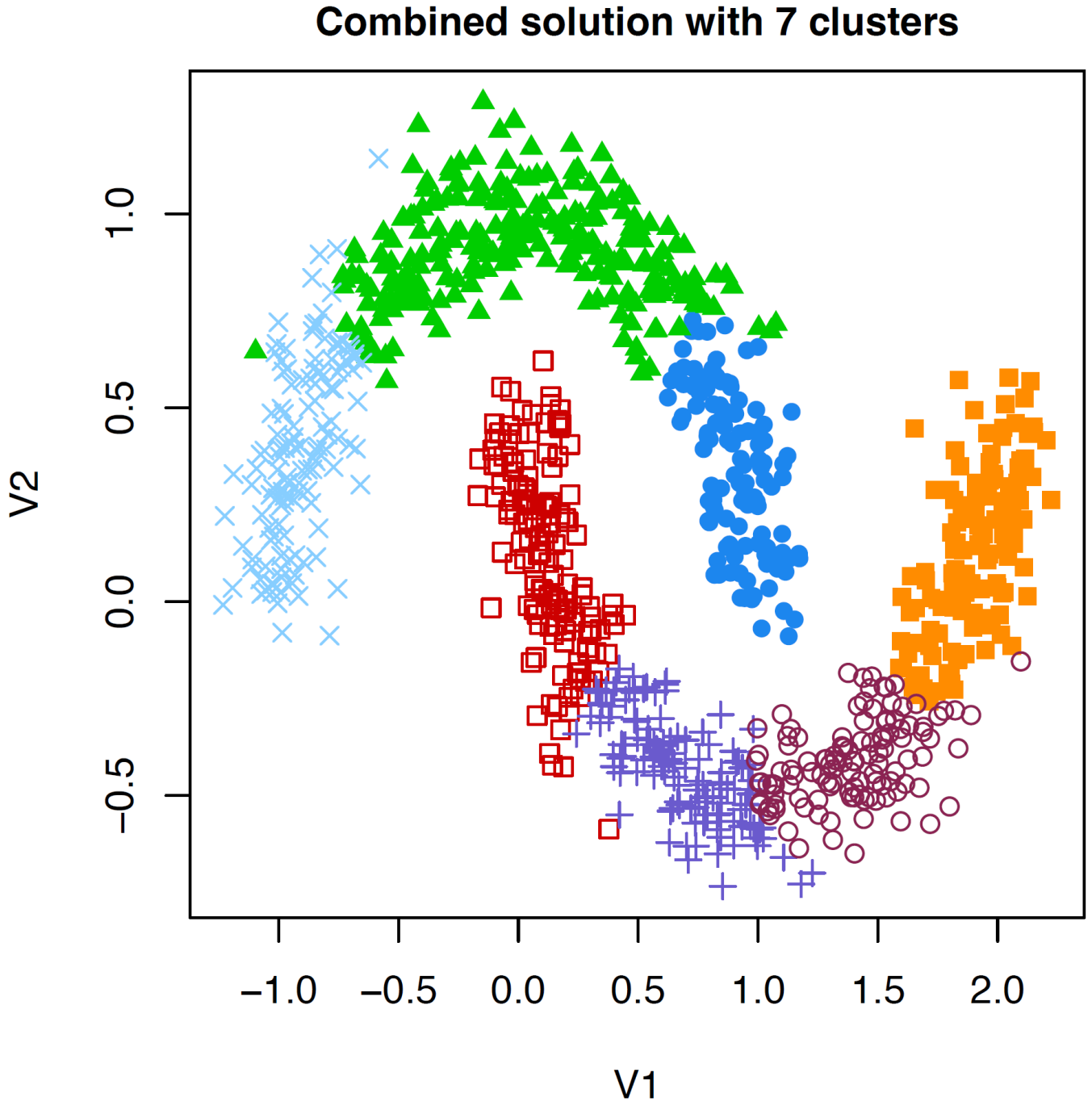
NB: we’re distinguishing “components” from “clusters” now

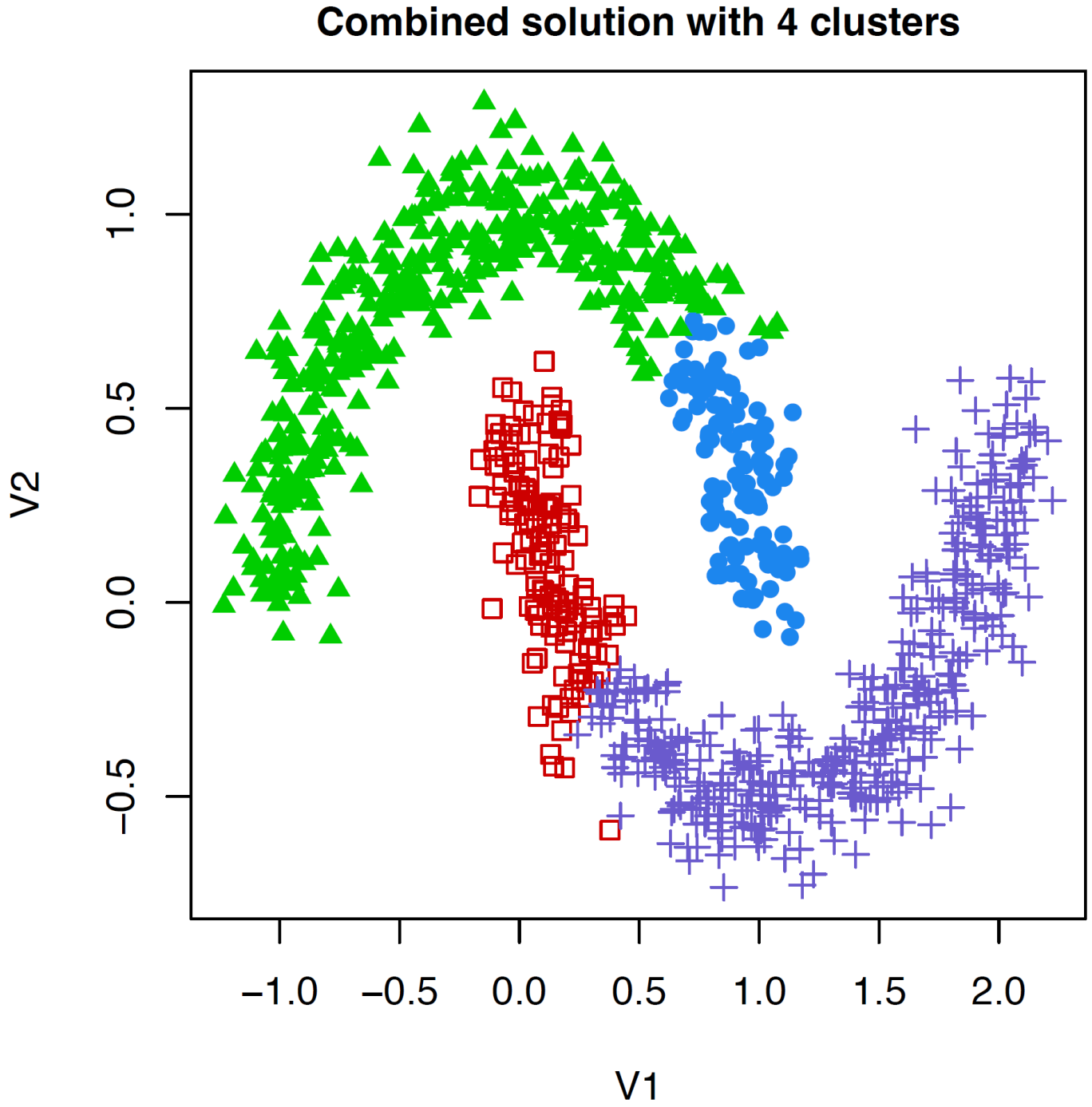
Merging

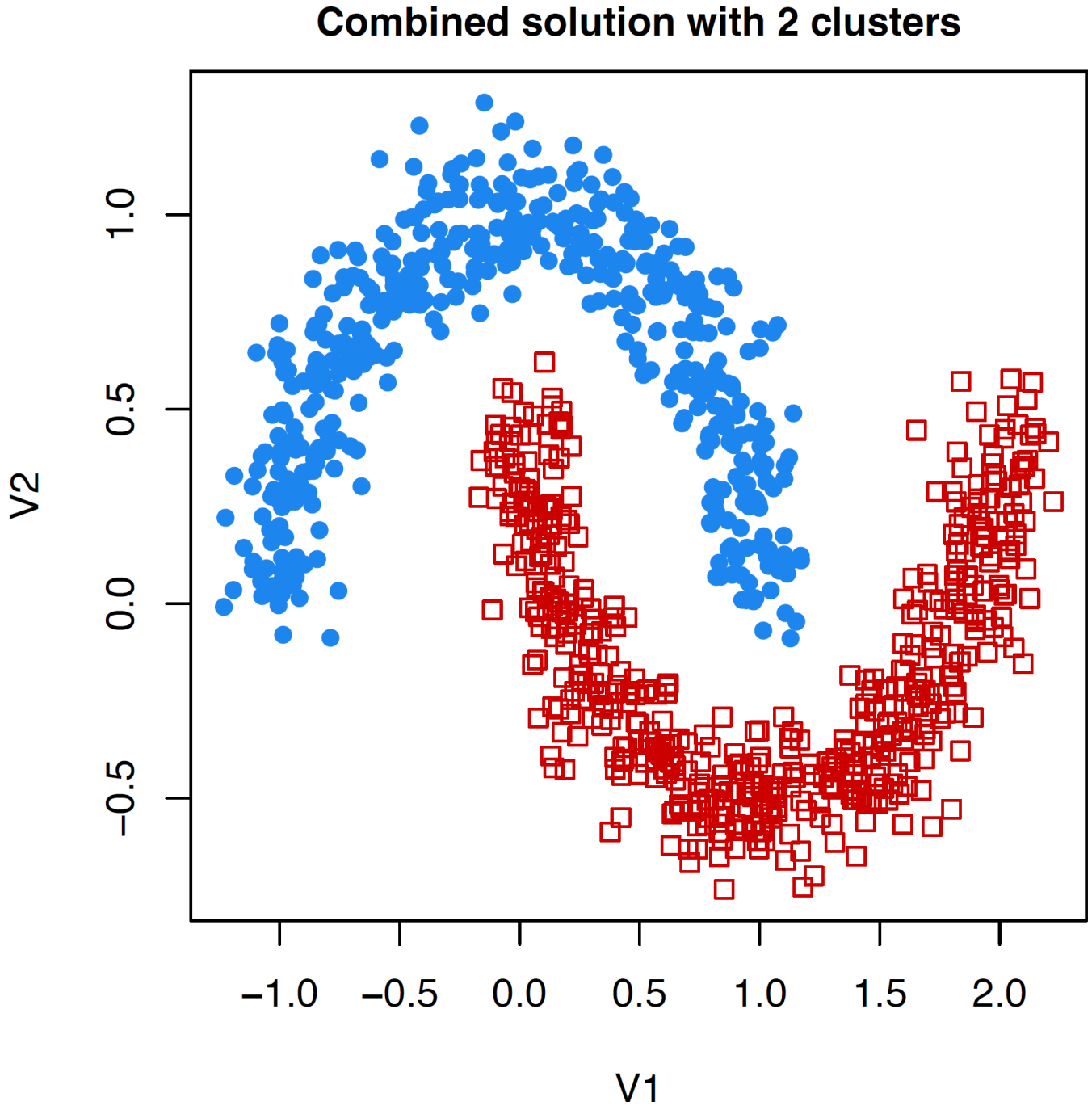
```
library(mclust)
out <- Mclust(x)
com <-
  clustCombi(out)

plot(com)
```




BIC solution (8 clusters)
V2
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
V1


Combined solution with 7 clusters
V2
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
V1


Combined solution with 4 clusters
V2
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
V1


Combined solution with 2 clusters
V2
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
V1

Assessing clustering results

Methods to assess whether the obtained clusters are “good”:

- Stability (previous lecture)
- External validity (previous lecture)
- Model fit

Model fit

How well does the model fit to the data?

Log-likelihood

$$\ell(\theta) = \log p(y|\theta) = \log \prod_{n=1}^{N} p(y_n|\theta) = \sum_{n=1}^{N} \log p(y_n|\theta)$$

The higher the log-likelihood, the more likely the data (if we assume this model is correct)

Deviance

$-2 \cdot \ell(\theta)$ (lower deviance is better)

Information criteria

Deviance forms the basis of **information criteria**, which balance **fit** and **complexity**

Akaike information criterion

$AIC = -2\ell(\theta) + 2k$

(where k is the number of parameters)

Bayesian information criterion

$BIC = -2\ell(\theta) + k \log n$

(where n is the number of rows in your data)

Information criteria

Think: bias and variance tradeoff!

- Variance also has to do with stability

Better fit & lower complexity = better cluster solution

(other assessment methods also available for model-based clustering)

High-dim!

How to do GMM in high dimensions?

- Same solution as we are used to by now!
  - Perform clustering on dimension reduction version of original data
  - Integrate regularization / dimension reduction in your GMM optimization method

- Bouveyron et al. (2007) High-dimensional data clustering; *Computational Statistics & Data Analysis 52*, 502 – 519
  - The second solution
  - Akin to “mixtures of probabilistic PCA”

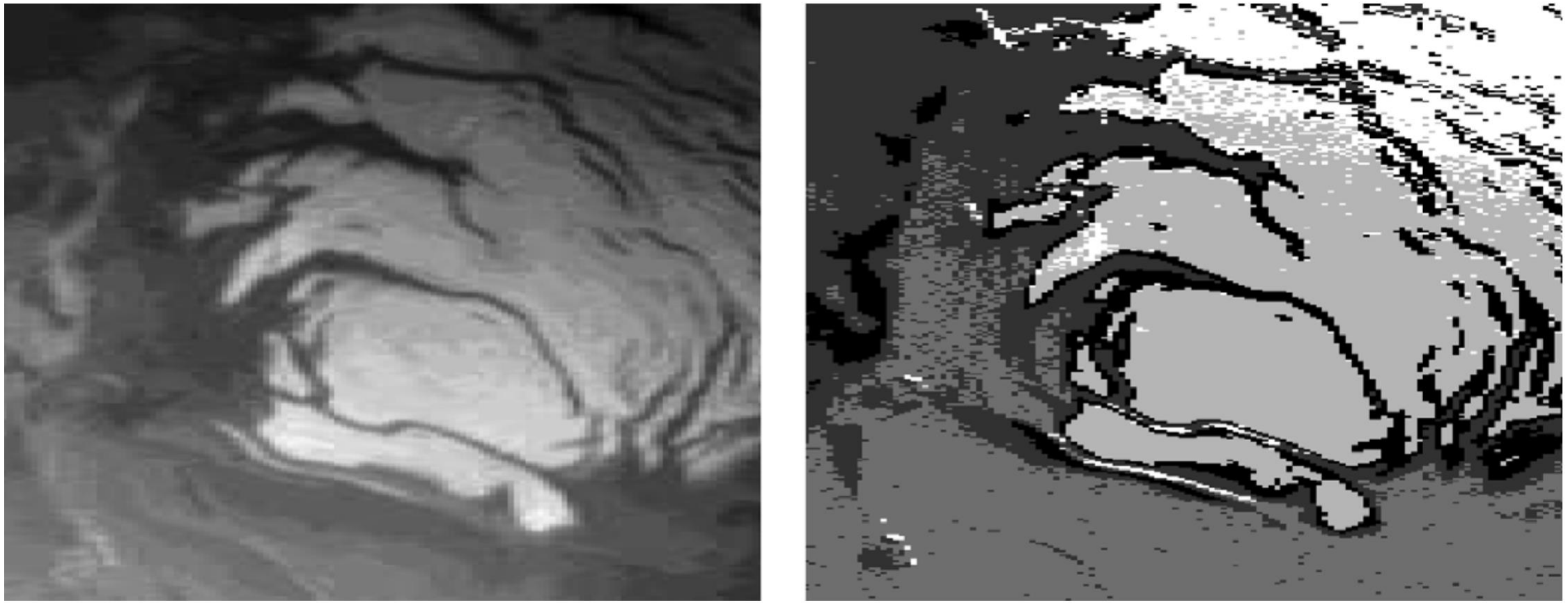


Fig. 8. Characterization of the Martian surface composition using HDDC: on the left, image of the studied zone and, on the right, segmentation using HDDC on the 256-dimensional spectral data associated to the image.

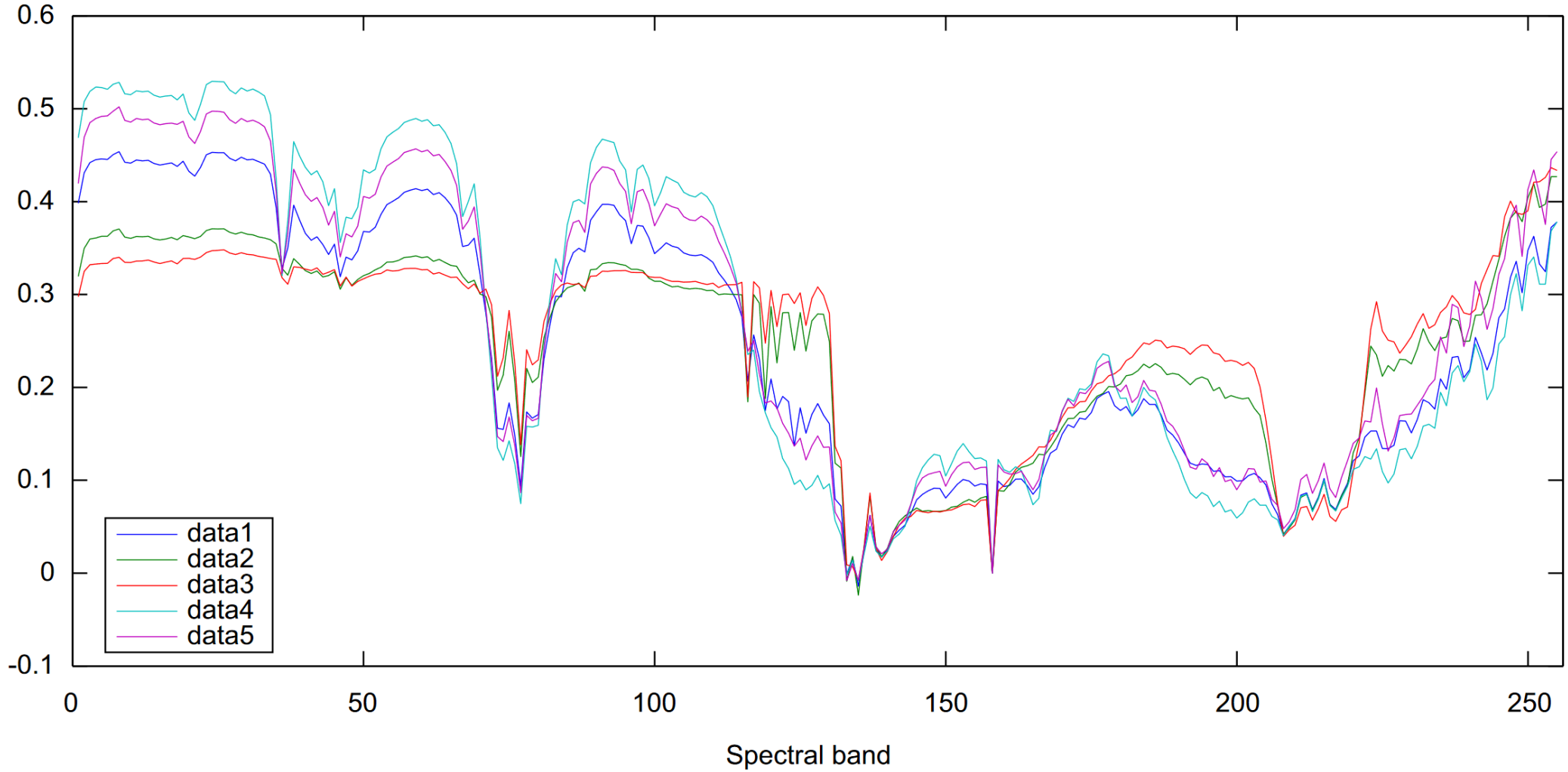


Fig. 9. Spectral means of the five mineralogical classes found using HDDC.

Model-based clustering in R

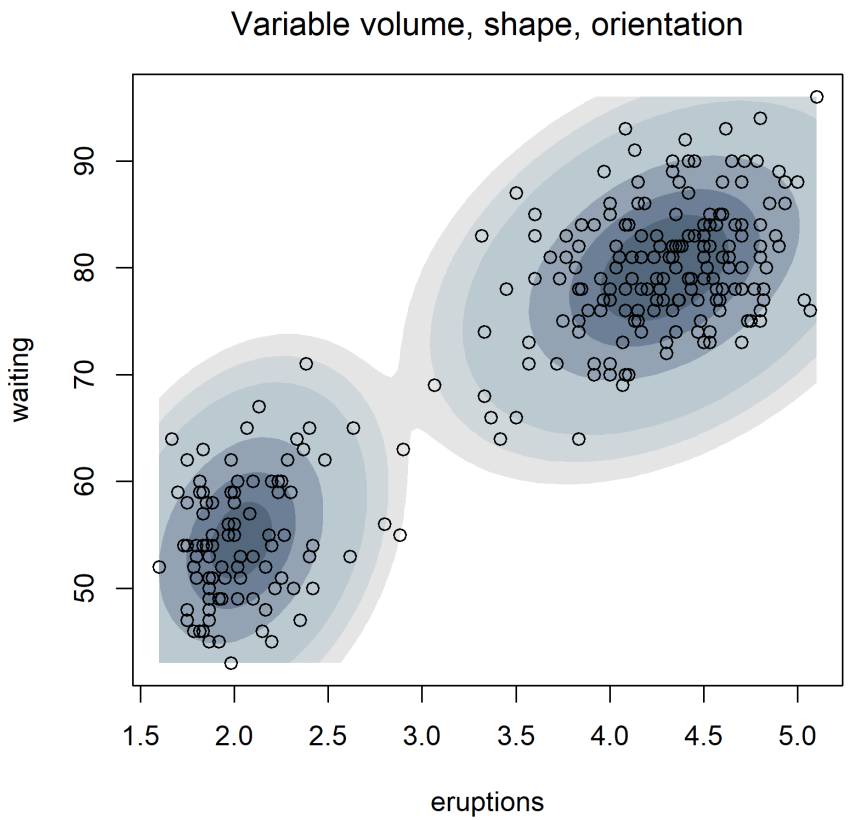
Model-based clustering in R

- **Mclust** implements multivariate model-based clustering
- Provides an easy interface to fit several parameterizations
- Model comparison with BIC
- Plotting functionality

Model-based clustering in R

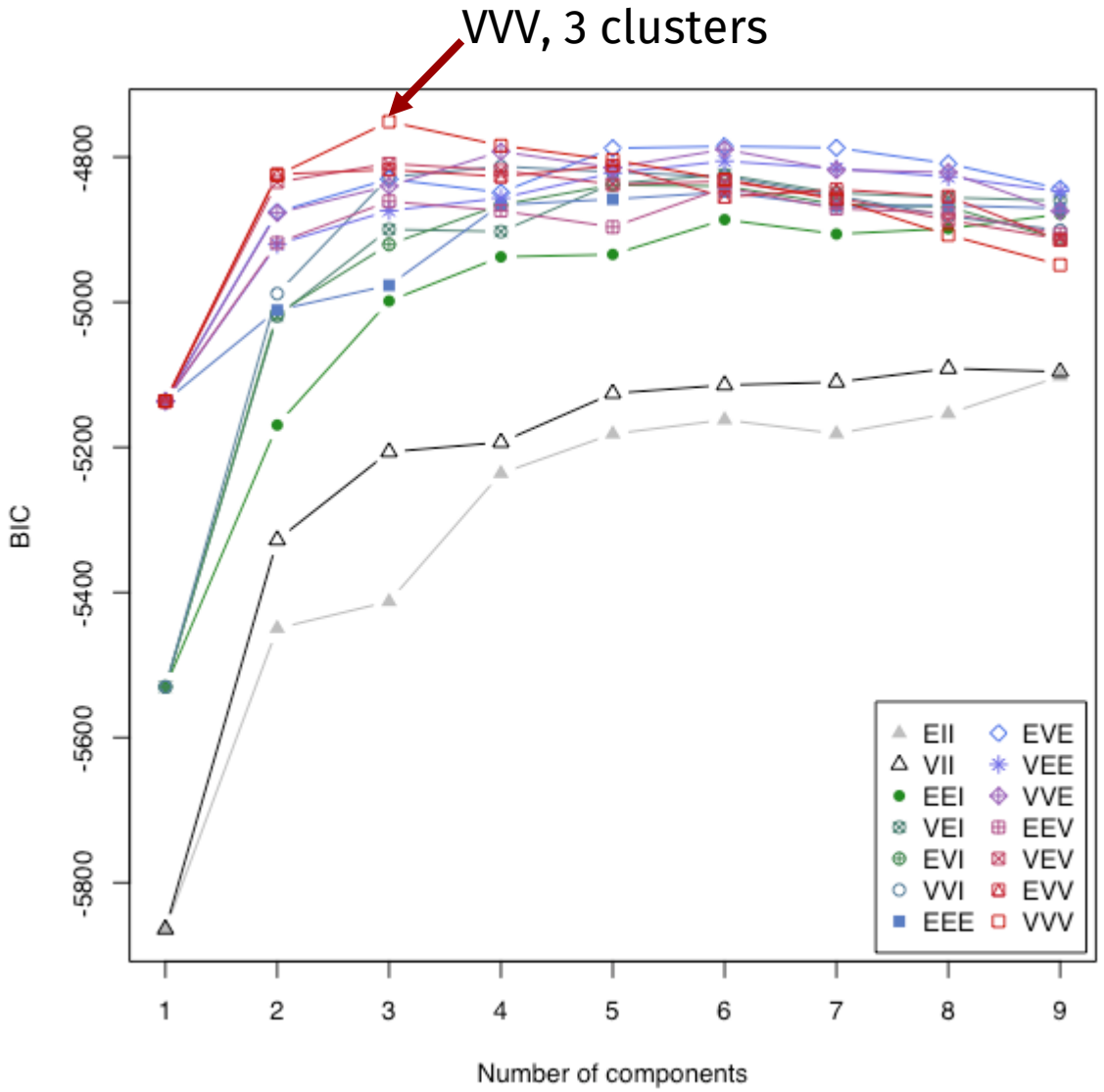
- Mclust uses an identifier for each possible parametrization of the cluster shape: E for equal, V for variable in:
  - **Volume** (size of the clusters in data space)
  - **Shape** (circle or ellipse)
  - **Orientation** (the angle of the ellipse)
- So an EEE model has equal volume, shape and orientation
- A VVV model has variable volume, shape, and orientation
- A VVE model has variable volume and shape but equal orientation

Model-based clustering in R



Model-based clustering in R

- How Mclust optimizes hyperparameters:
  - Fit all the models with up to 9 clusters (or more, your choice!)
  - Compute the BIC of each model
  - Choose the model with the lowest BIC



Practical: perform model-based clustering

Take-home exercises: 1-11

Questions?