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

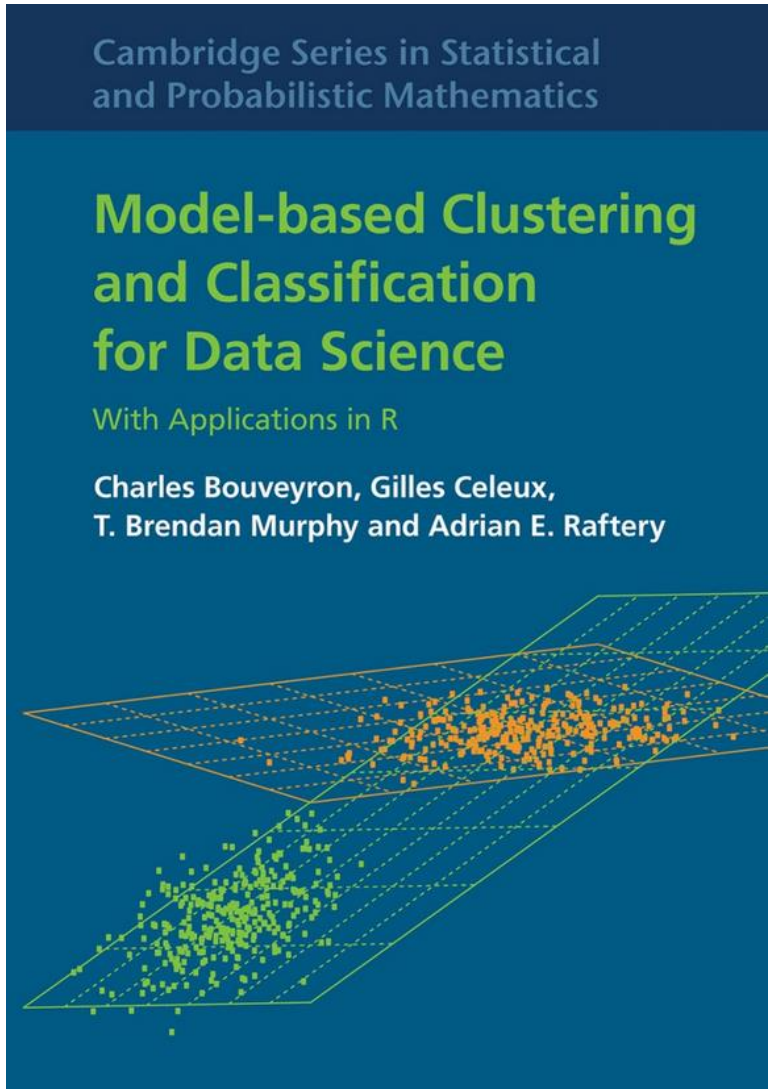
Today

- 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

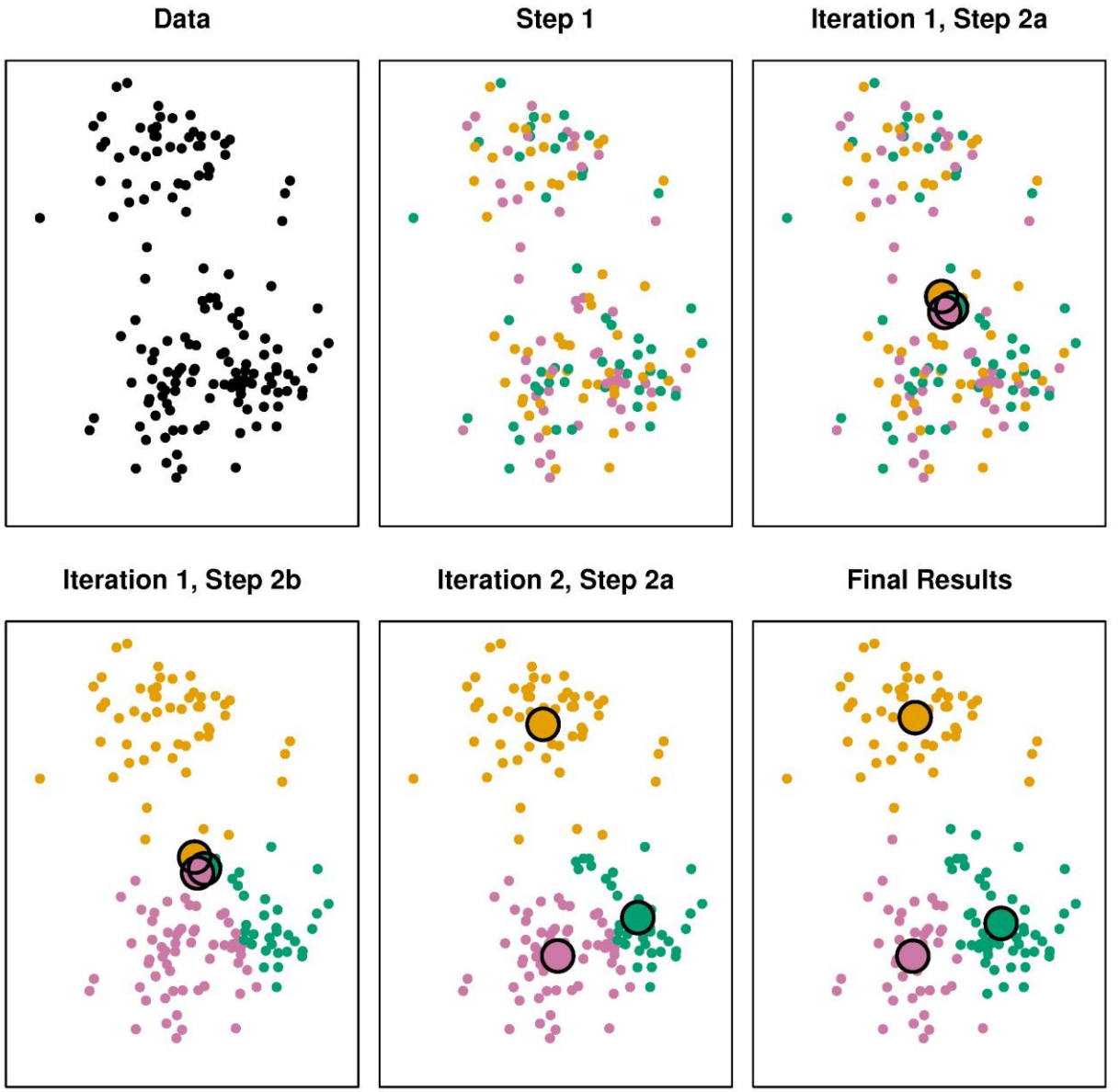
K-means again

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.**

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.
- Questions are easier to answer.

Model-based clustering

- Assumptions about the clusters are explicit, not implicit.
- We will look at the most commonly 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

Specific examples of model-based clustering:

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

Gaussian mixture modelling

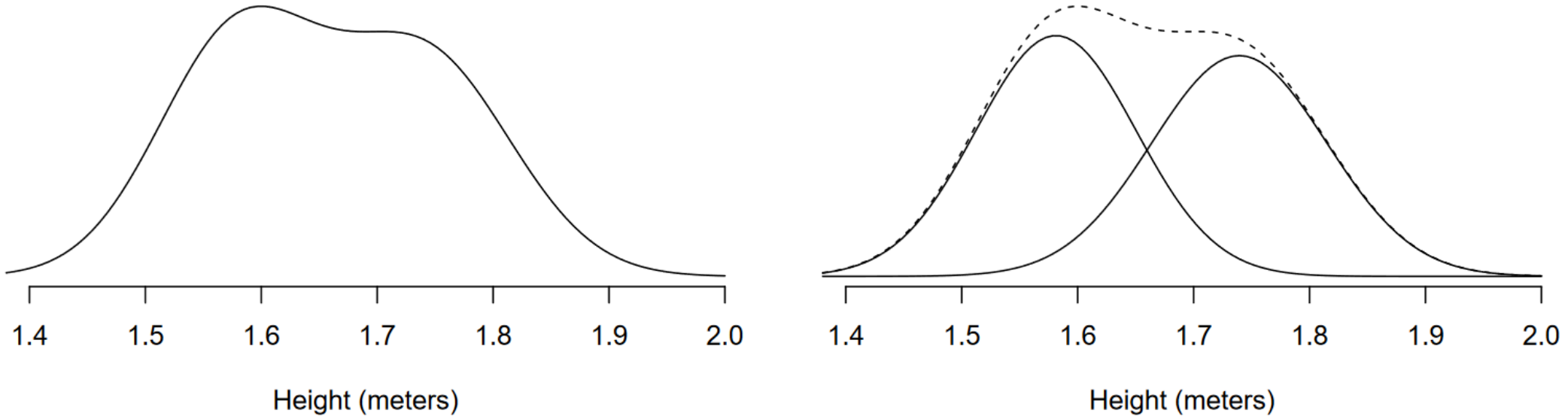


Fig. 1 Peoples' height. Left: observed distribution. Right: men and women separate, with the total shown as a dotted line.

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!

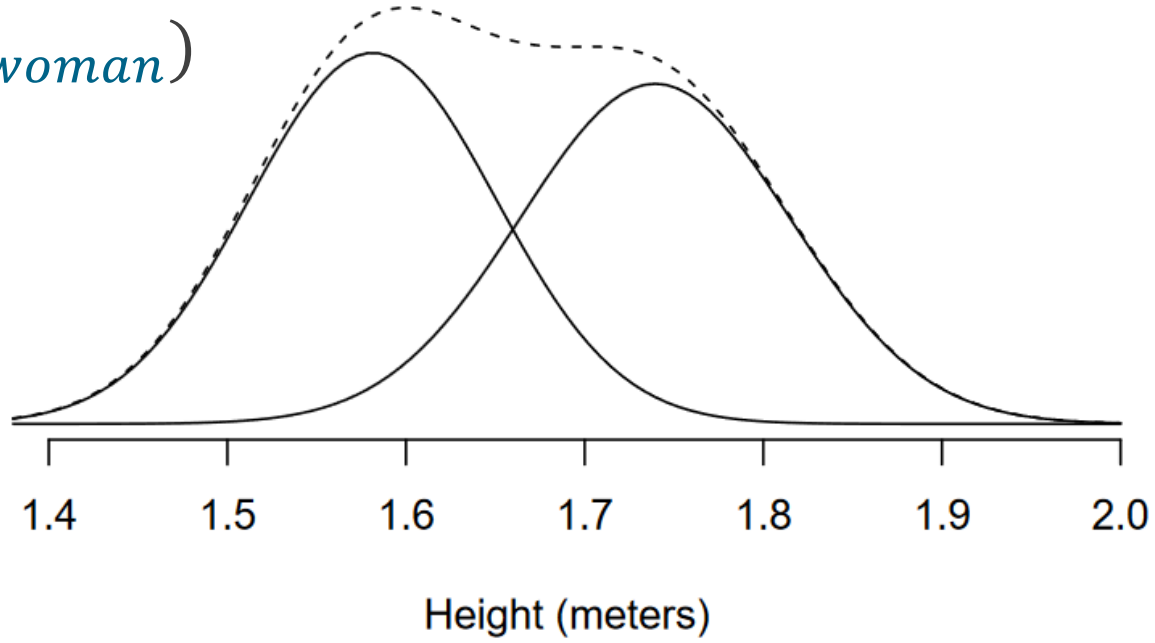
Model-based clustering

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 $\hat{\sigma}_2$)

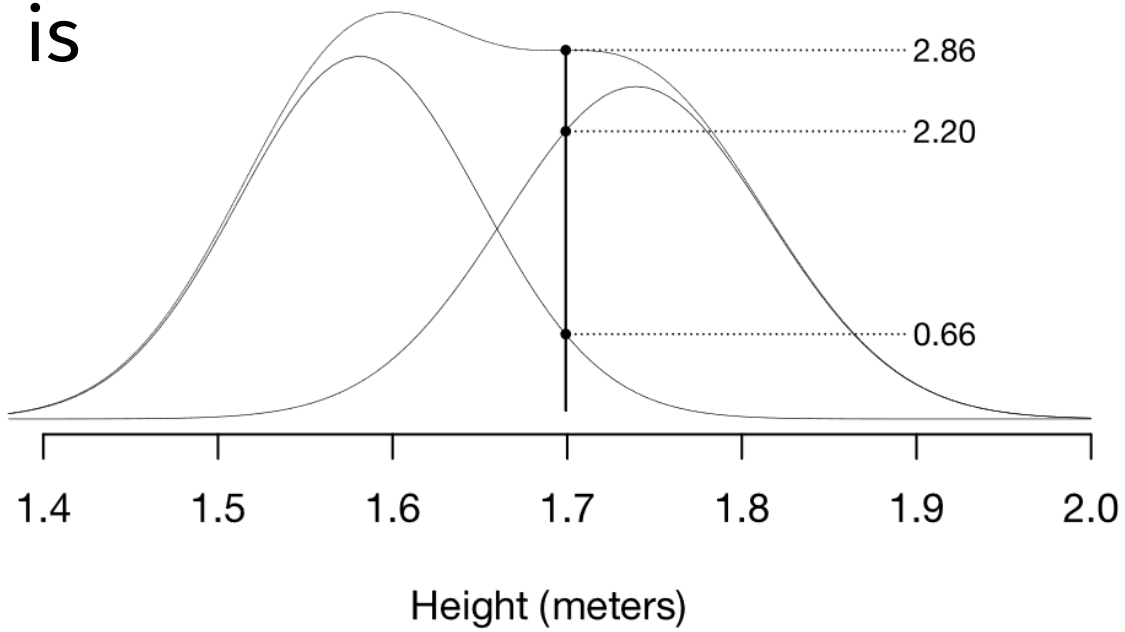
But we don't know this!

-> 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?

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

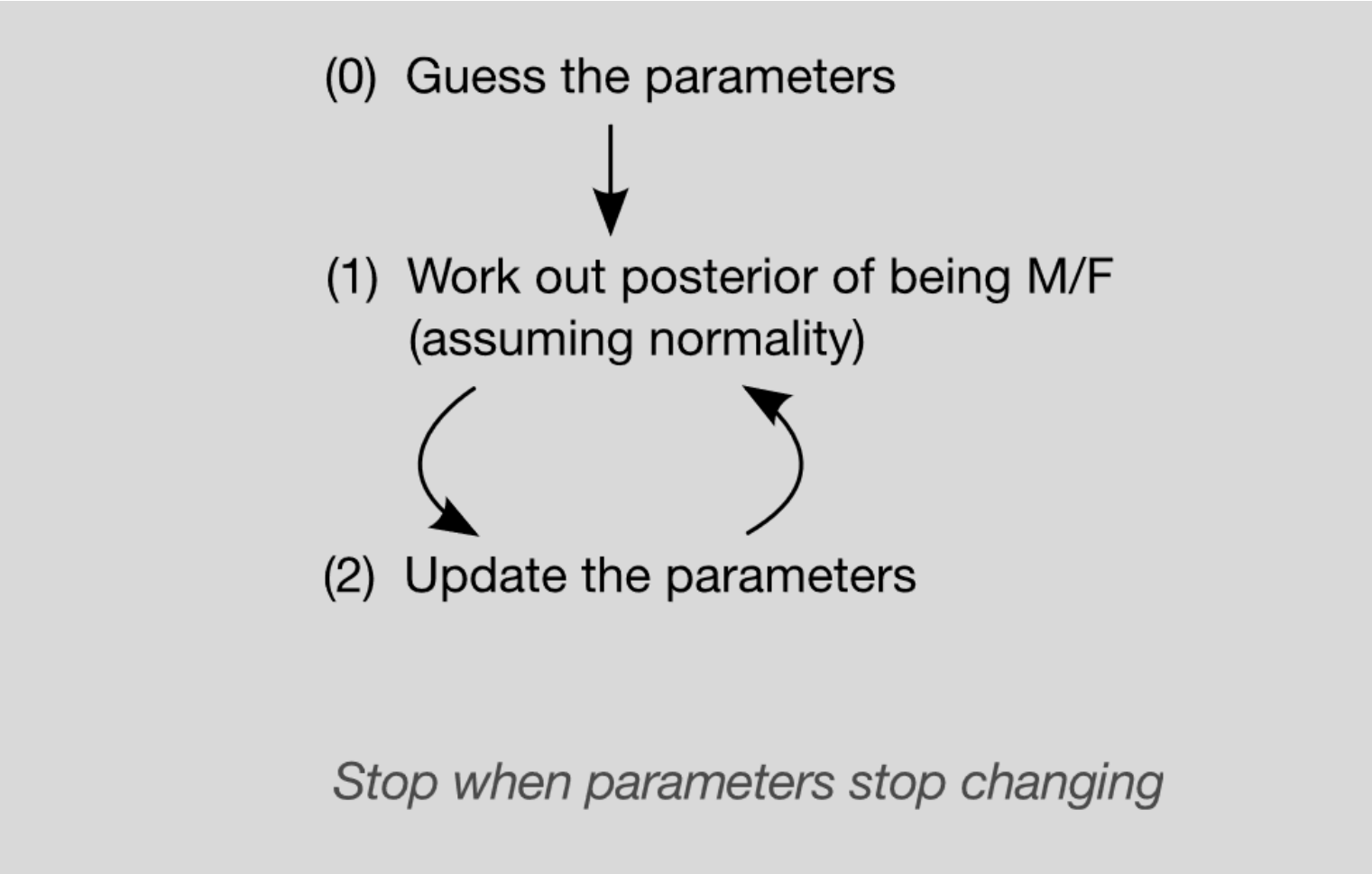


Estimation: the EM algorithm

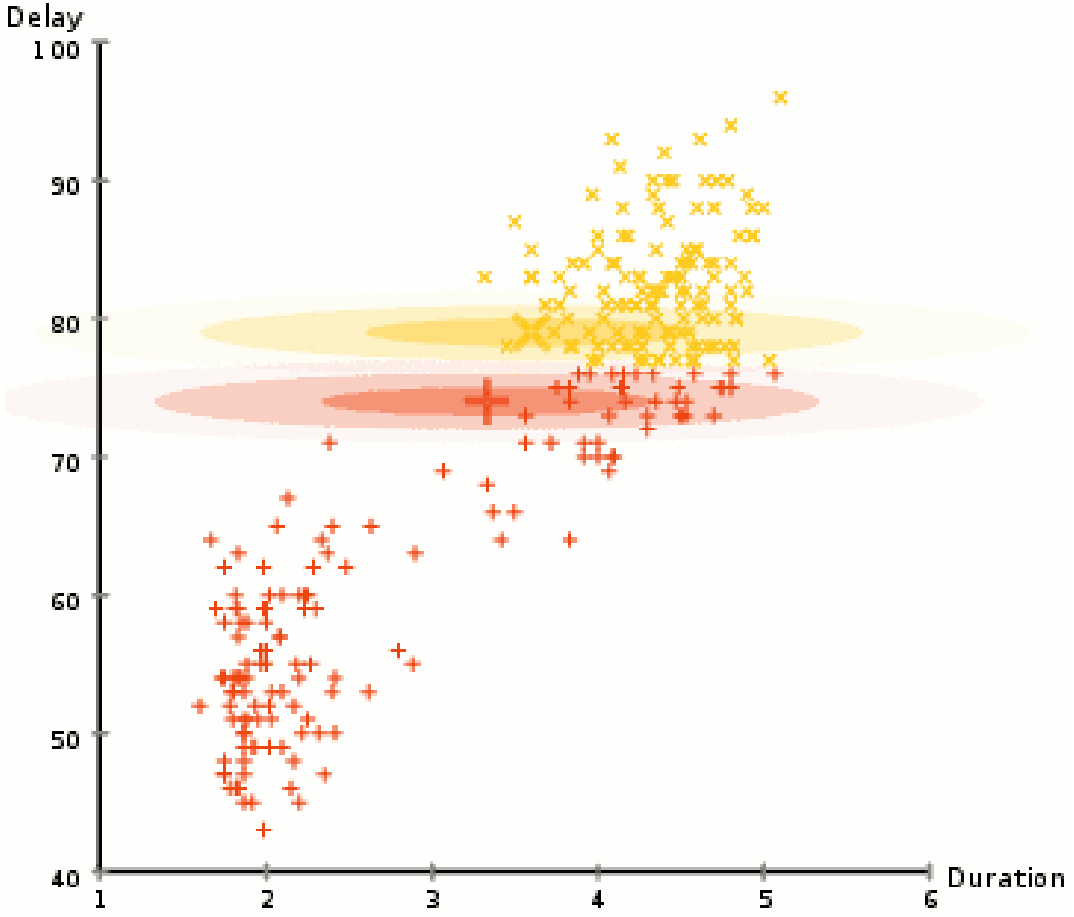
- Now we have some class assignments (probabilities);
- So we can go back to the parameters and update them using our easy rule (M-step)
- Then, we can compute new posterior probabilities (E-step)

Does it remind you of something...?

Estimation: the EM algorithm



Estimation: the EM algorithm



Questions?