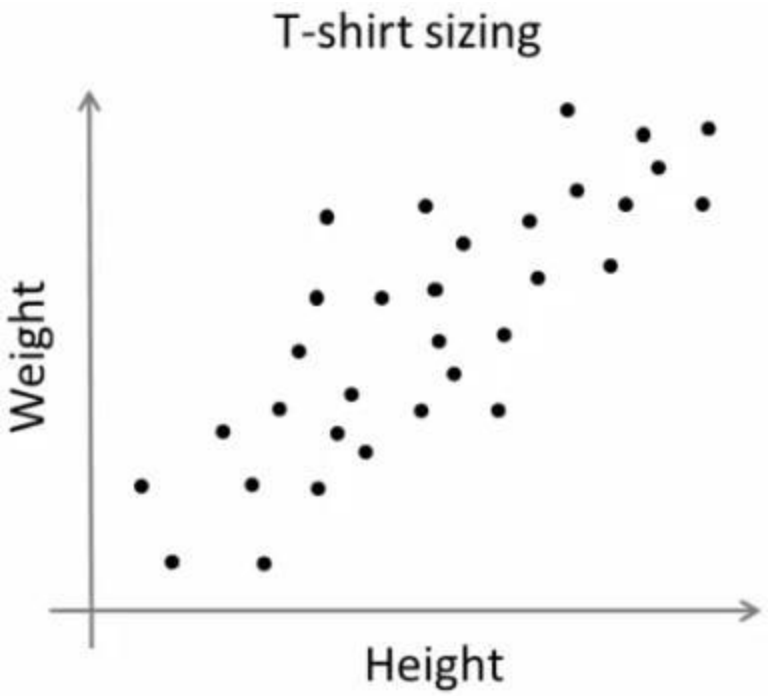
Introduction to Data Science

Clustering

Gordon Anderson

Clustering Data

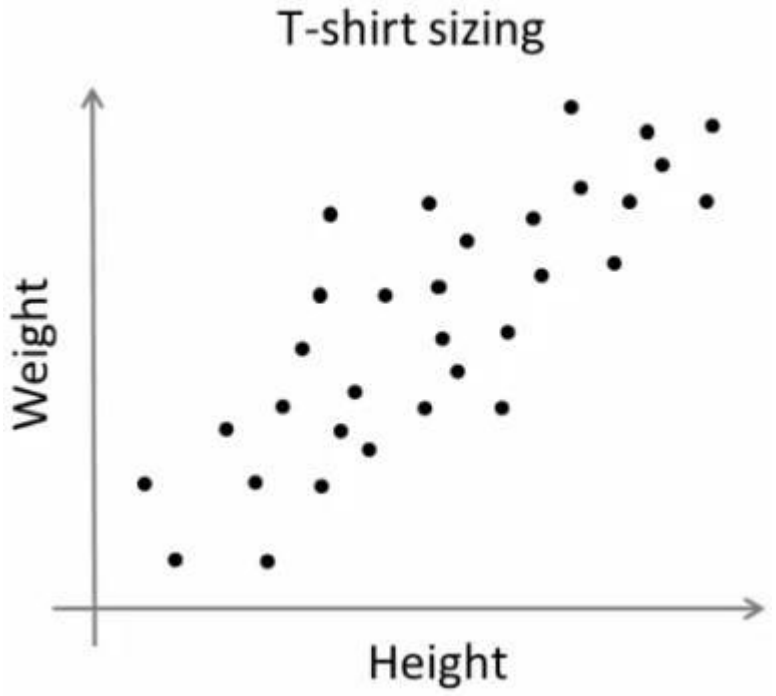


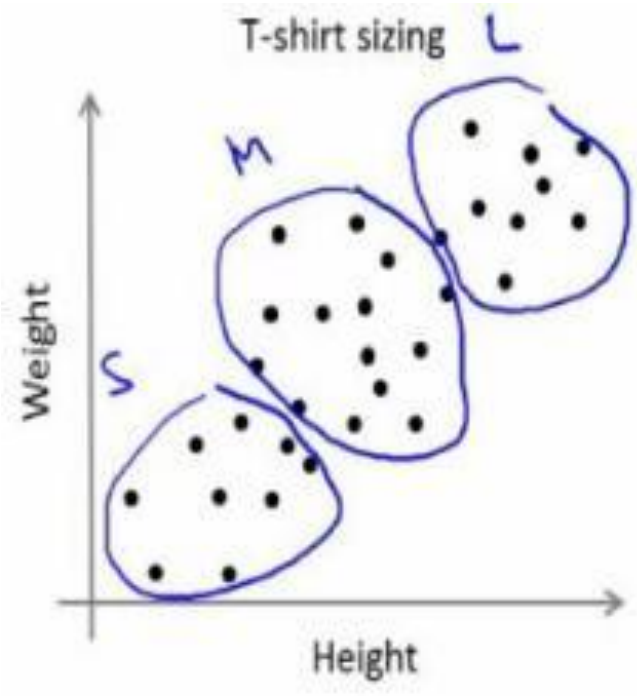
Are there “natural” groups in these data?

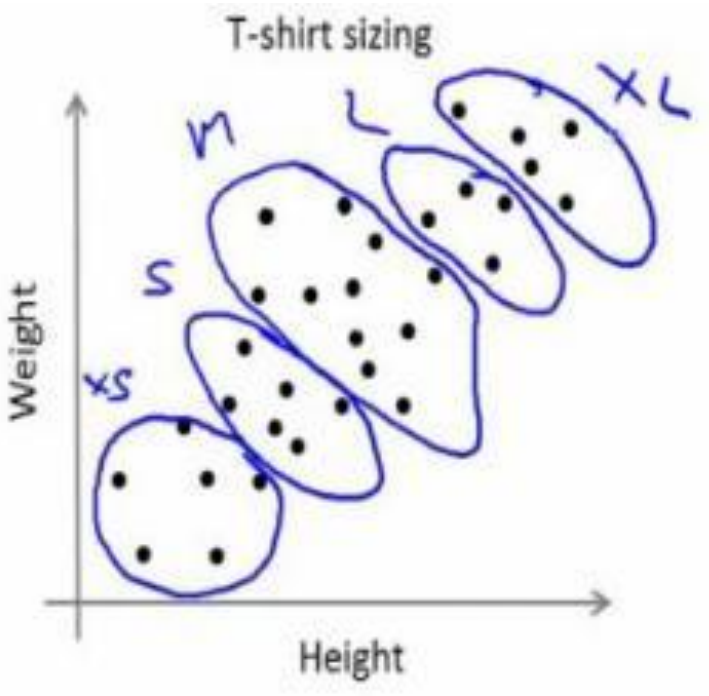
If so, what can we say about members of a group?

Clustering is often arbitrary

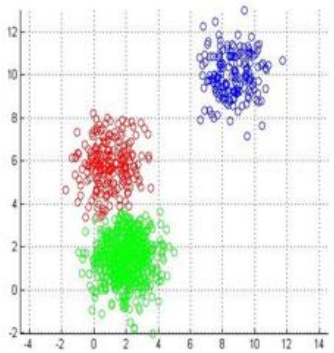
Often used to create reasonable “segments” of a population.







Clustering



- Task: find “natural” groupings or clusters in data.
- The clusters must have a “cohesiveness”.
- Mostly based on a distance metric: within-cluster distances less than between cluster distances.
- Also referred to as “segmentation”.
- Done by Unsupervised ML- there are no training examples.
- Use some variables to determine clusters, use other to “profile” the characteristics of each cluster discovered.

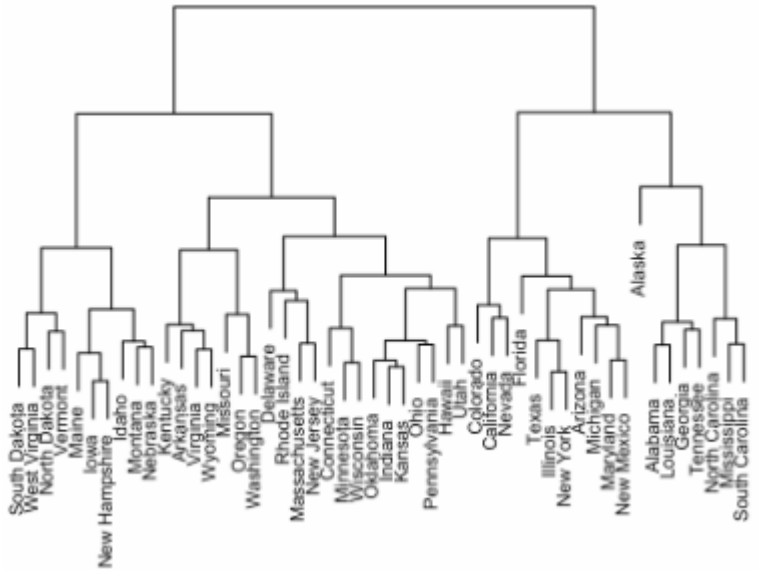
Clustering applications

- Genetics: find groups of genes based on aspects of their functionality.
- Information retrieval: group search returns in meaningful way.
- Business: find groupings of customers based on click behavior
- Epidemiology: identify progression of a disease geographically.
- City-planning: identify groups of houses according to their type, value and location.
- And many, many more…

Clustering- Overview of Methods

1. Hierarchical: agglomerative/divisive
2. Partitioning: k-means, PAM
3. Model-based: Gaussian mixture model

Hierarchical Clustering



- Produces a tree-like representation.
- Number of clusters depends on the level of the tree.
- Uses a distance metric.
- Two main types:
    1. Agglomerative: bottom-up. Good at identifying small clusters. Uses several "linkage" methods- how to compute dissimilarity (distance).
    2. Divisive: top-down. Good at identifying large clusters.

    (We won't be covering this type of clustering).

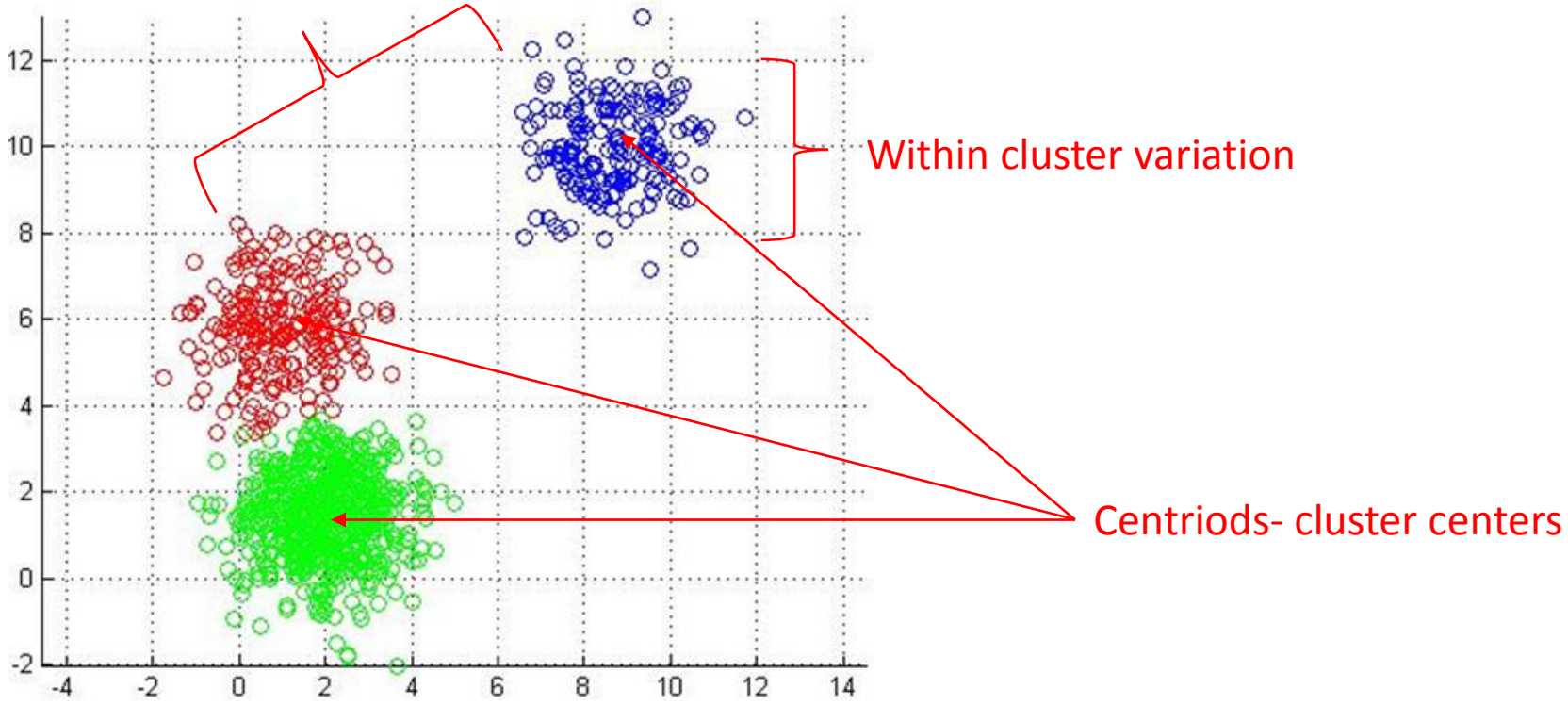
Partitioning: K-means

- Uses a distance metric.
- Defines centroids as cluster centers.
- Centroids are geometric locations in an n-dimensional space.
- Goal: minimize the within-cluster distances to centroids.
- The algorithm requires that the suspected number of clusters, k, be specified as well as the distance metric.
- Input to algorithm: data in a form that can be used by the distance metric- usually real-valued numbers.
- Output: cluster memberships of each data record. Usually the centroid coordinates are available as well.

Partitioning: K-means

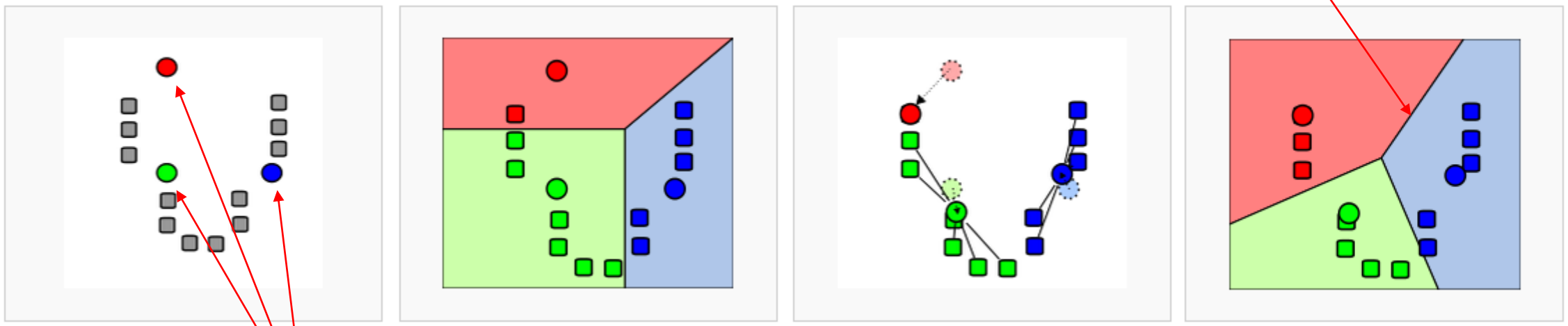
Goal: minimize the within-cluster variation.

Between cluster variation



K-means Algorithm

Have to specify the number of clusters K, distance metric beforehand.



1- generate K centers in the data space at *random* locations.
2- assign data points to closest center (requires a similarity metric).
3- calculate a new center for each cluster
4- repeat from step 2 until no new reassignments.

K-means Cluster Evaluation

Since you have to specify the number of clusters before you run the algorithm, How do you know what the best number is?

1- You have a hypothesis about groups in the data and you want to investigate it.

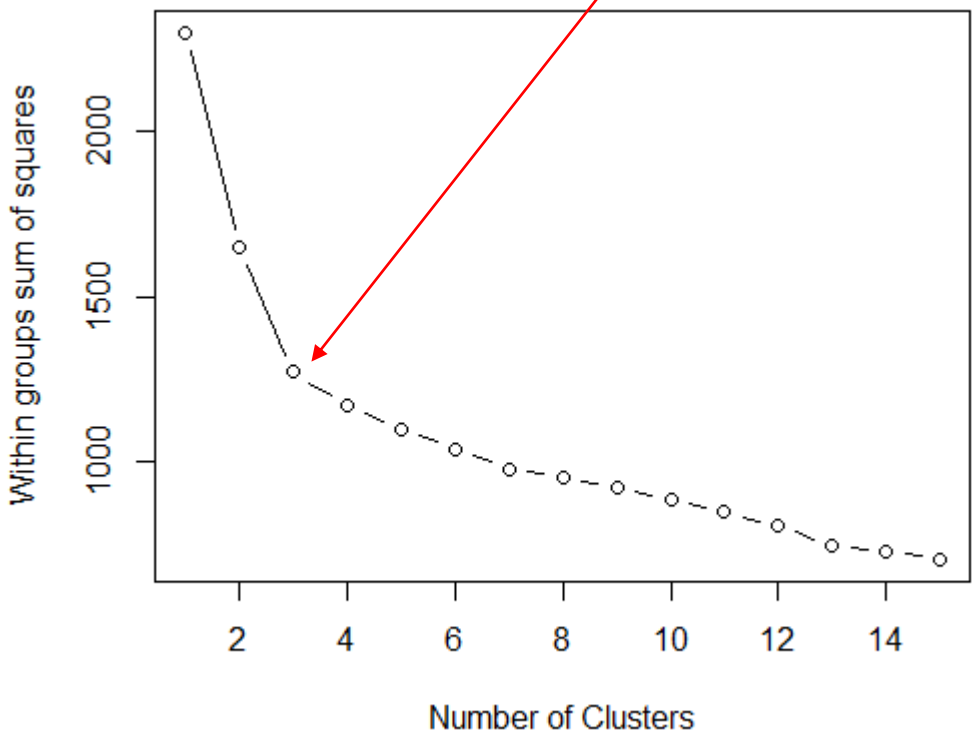
2- You run K-means on a lot of different values for K and select the best run.

"Elbow"- where the error increases sharply with fewer clusters. This is a likely best number of clusters.

Sum of squares = sum of squared errors, or residual sum of squares, RSS.

$$RSS = \sum_{k} \sum_{i \in S} \sum_{j=1}^{n} \left(x_{ij} - \bar{x}_{kj}\right)^2$$

The sum of the distances between all data points in a cluster summed over all clusters k.



Partitioning: K-medioids, or PAM

- A medioid is an actual data point- a centroid is a point in feature space.
- PAM works like K-means except that cluster centers are data points:

1. Choose k data points as medioids (cluster centers) at random.
2. Assign each data point to the cluster center it is closest to.
3. As long as the "cost*" is decreasing, do the following (this is a "while" loop):
4. Move the cluster center to the average point for all of its member data points.
5. Repeat steps 2 and 3 until there are no cluster reassignments or until some stopping criterion is met.

* The cost is computed by some function. It could be the sum of the distances (dissimilarities) of points to their medoid (cluster center).

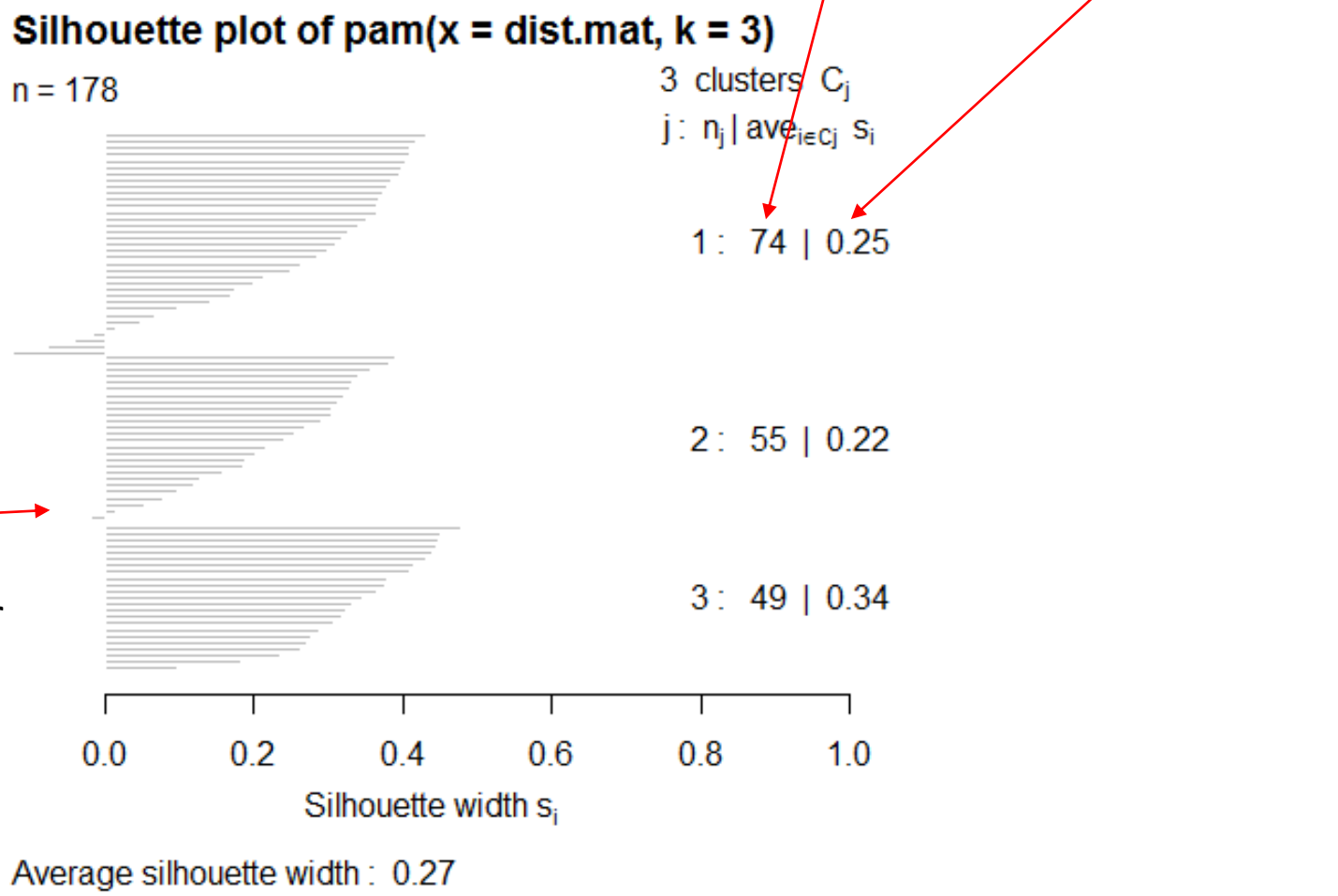
Cluster Evaluation- Silhouette

This plot is interpreted as follows:

Each cluster member is represented as a grey horizontal line. The length of the line is its "silhouette width", a measure of how well it fits with its cluster.

Large, positive values are good.

Negative lines indicate poorly fitting members as they could probably just as well go in another cluster, or they may not really fit well into any cluster.



PAM- Advantages

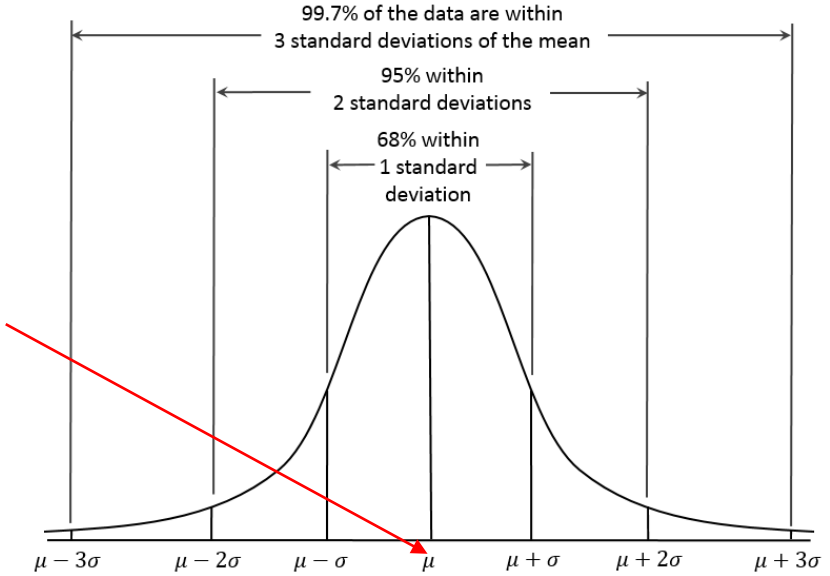
- K-means sensitive to outliers- PAM more robust with respect to outliers.
- Using actual data points can be more meaningful for interpreting clusters- provides an exemplar for each cluster.

Model-based: Gaussian Mixture Model

- Gaussian (Normal) probability distribution models the dispersal of each cluster.
- Each cluster has its own distribution.

A generic image of the Gaussian distribution

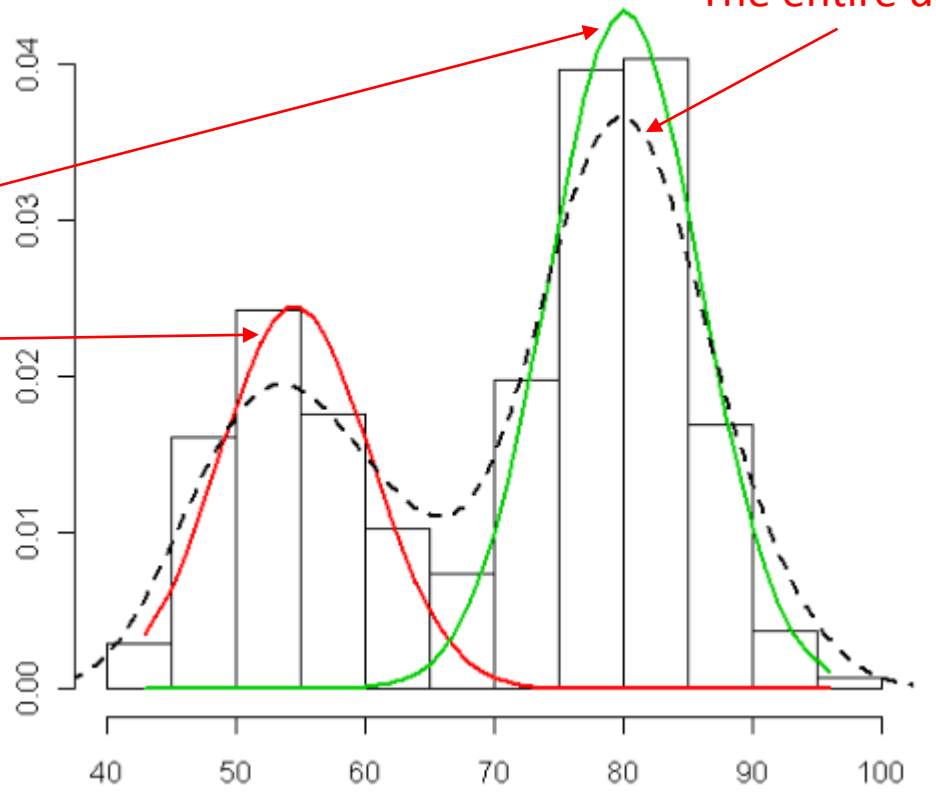
The mean will represent cluster centers



Gaussian Mixture Model

Each cluster is represented as a Gaussian distribution, with its mean as the cluster center.

The entire feature space is, therefore, a mixture of Gaussian distributions.



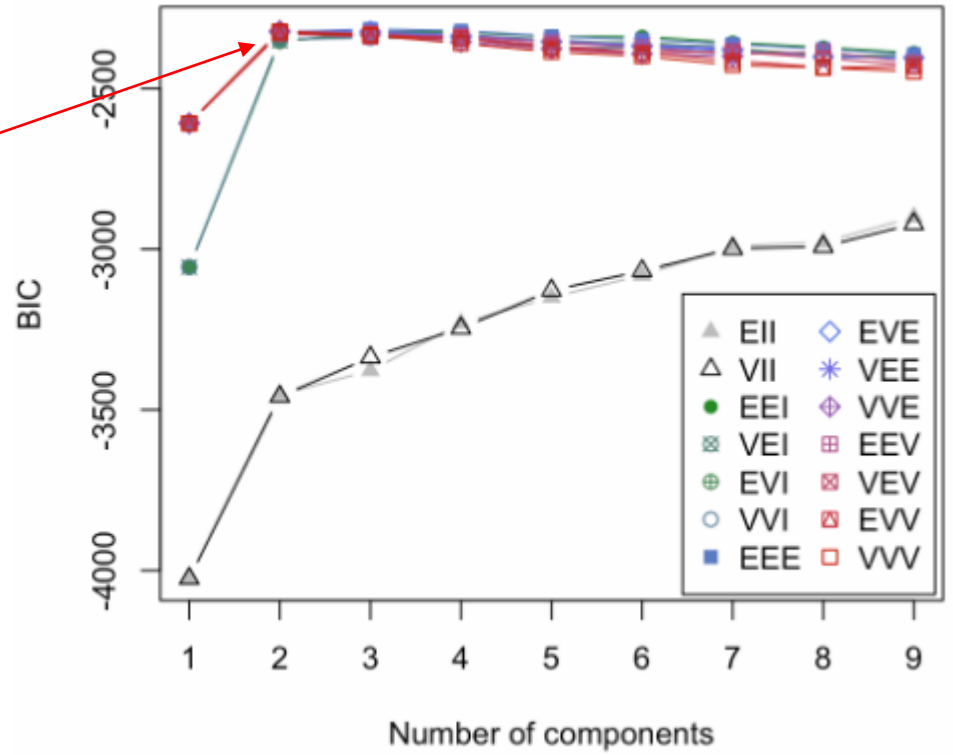
Gaussian Cluster Modeling

- Instead of a distance metric, each data point has a probability of belonging to a cluster.
- Gaussian distributions can vary by shape, volume and orientation in the feature space.
- Typically, a maximum likelihood algorithm is used to fit all these models for a range of k components, or clusters.
- The “best model” is selected using the Bayesian Information Criterion or BIC- a penalty for more complex models. A *larger* BIC score indicates stronger evidence for the corresponding model.

Example Clustering Evaluation

- Model options, in mclust package, are represented by identifiers including: EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV and VVV.
- The first letter refers to volume, the second to shape and the third to orientation. E stands for “equal”, V for “variable” and I for “coordinate axes”.

The “best” number of clusters is 2 or 3. Several models are equally good at fitting the data.

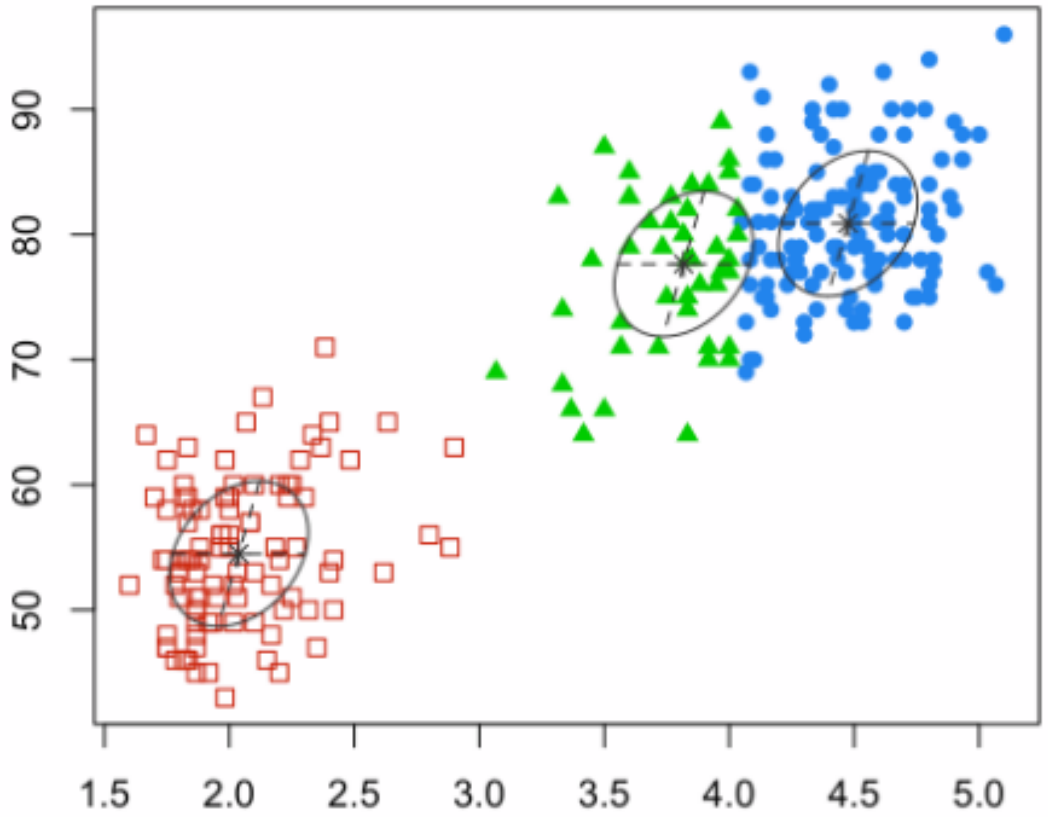


Example Clustering Output

The output is cluster centers, the means of the three distributions, as well as the uncertainty of this particular clustering.

Plot of the optimal model:
EEE: ellipsoidal shape, equal volume, orientation.

Advantages of Model-Based approach is that it determines the “best” number of clusters and the best probabilistic model for the clusters. Doesn’t rely on distance in the way that partitioning algorithms do.



Clustering Summary

- We focused on partitioning and model-based clustering.
- Generally take numeric data that may be scaled or normalized.
- There are many clustering algorithms for all kinds of data.
- Try several different types of clustering in your analysis.
- Vary the model settings- i.e. try many values of k and perhaps different distance metrics in K-means and PAM.
- The clusters must “make sense” in terms of the context of your study.
- Don’t blindly accept the results of an algorithm as the “truth”.
- That’s thinking like a data scientist!