Machine Learning pt. 1: Unsupervised

ACE 592 SAE

Unsupervised Machine Learning

Most of us know what supervised machine learning is, but what is “unsupervised learning”?

Essentially, it is inference for data that is ***not labeled***.

This amounts to attempting to learn latent patterns or variables in the data rather than predict explicit labels.

Two Algorithms We Will Go Over

1. *Principal Component Analysis*
    - For learning the implicit “components” of a data set.

2. *K-Means Clustering*
    - For discovering groupings in the variable space.

Principal Component Analysis (PCA)

- Essentially an orthogonal, linear transformation that translates the data into a new coordinate system.

- In this coordinate system, the **first component** explains the most variance, the **second component** explains the second most amount of variance, etc.

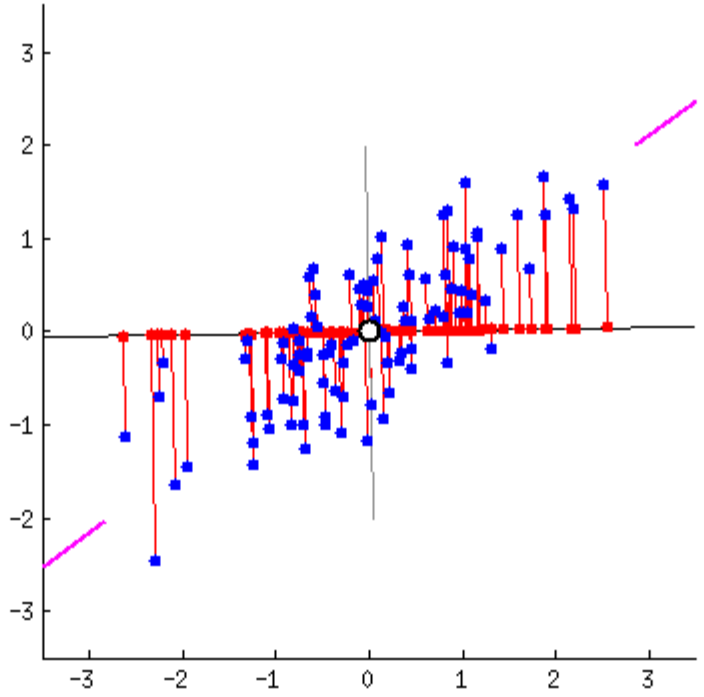
- The new dimensions of the system are linear combinations of existing variables.

How PCA chooses the components

For a given set of data, it tries to find a line/plane where:

1. Errors are minimized (like OLS).
2. Variation is maximized (red dots on the line are as spread out as possible).

So PCA is **also trying to maximize the variation the component explains**!



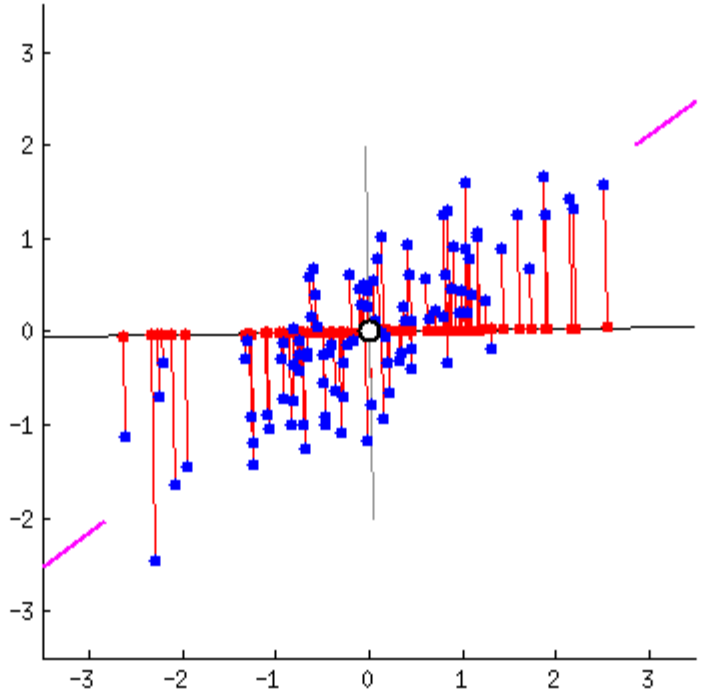
Graph taken from [this very cool explanation from Stack Overflow]

PCA and Linear Algebra

Using spectral decomposition, we can always make a *new orthogonal coordinate* system for any square symmetric matrix. The result is given by its ***eigenvectors.***

When we do this to the covariance matrix of the data, **the eigenvectors are equivalent to the principal components.**

Notice that at the magenta lines the new "frame" would make the points uncorrelated.



Graph taken from [this very cool explanation from Stack Overflow]

What this means

- PCA components can be found by Singular Value Decomposition or other methods of finding eigenvectors.

- **PCA components will be orthogonal** to one another (a property of eigenvectors of symmetric matrices).

- PCA components will be linear combinations of all the variables in the data.

Why do we care?

- With PCA, we can express large numbers of features concisely in a way that maximizes explained variance.
- We can use this to reduce the dimensionality of a dataset.
- We can construct “indices” that represent several different variables on aggregate.

The Downside

- PCA components do not have clear “units.” They can be included as control variables, but seldom have an interpretation in a regression model.

- The eigenvectors may be related to variables in the data in any way, and not necessarily in a way that makes them interpretable.

- Because of the above, PCA is a great first step for a prediction model or clustering model. Whether it actually tells you useful things about the data is variable.

Example: Photo Classification

- In a given photo, its likely that not every pixel matters.
- Doing PCA on a photo will summarize several pixels into a reduced number of components.
- Classifying on these components can be just as effective and much more efficient!

Economics Example: Asset Indices

- In Demography and Public Health, it is popular to use PCA with a data set of assets to produce an “asset index.”
- The result will be some components which are correlated with variables in the data, and may or may not be interpretable.
- This can be used to “rank” households based on the amount of assets they have.

PCA Hyperparameters

The principal hyperparameter in PCA is **the number of components**.

Recall that the first will explain the most variance, and the second the second most. The components will explain less and less as they increase.

How should we choose?

PCA Hyperparameters

The principal hyperparameter in PCA is **the number of components**.

Recall that the first will explain the most variance, and the second the second most. The components will explain less and less as they increase.

How should we choose?

The number of components that explains “enough of the data.”

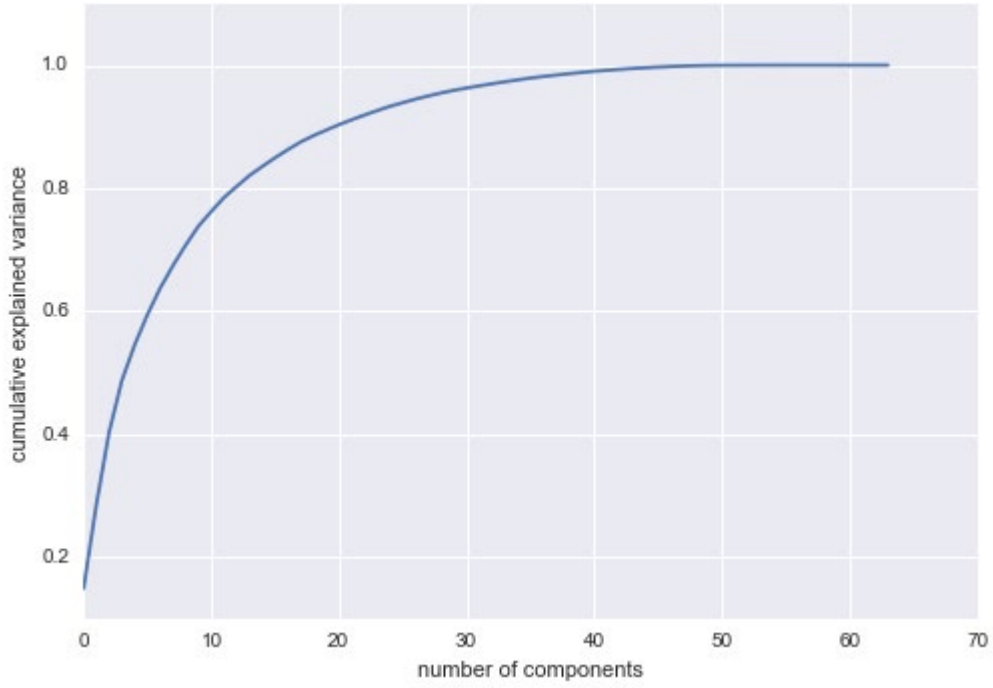
PCA Hyperparameters

Heuristics for Selection:

1. A threshold of explained variance.

Ex: *keep the number of components that explains 70% of the data.*

Good for when you are putting PCA components into a model, for example.



The “elbow rule”: choose where the gains start to level out.

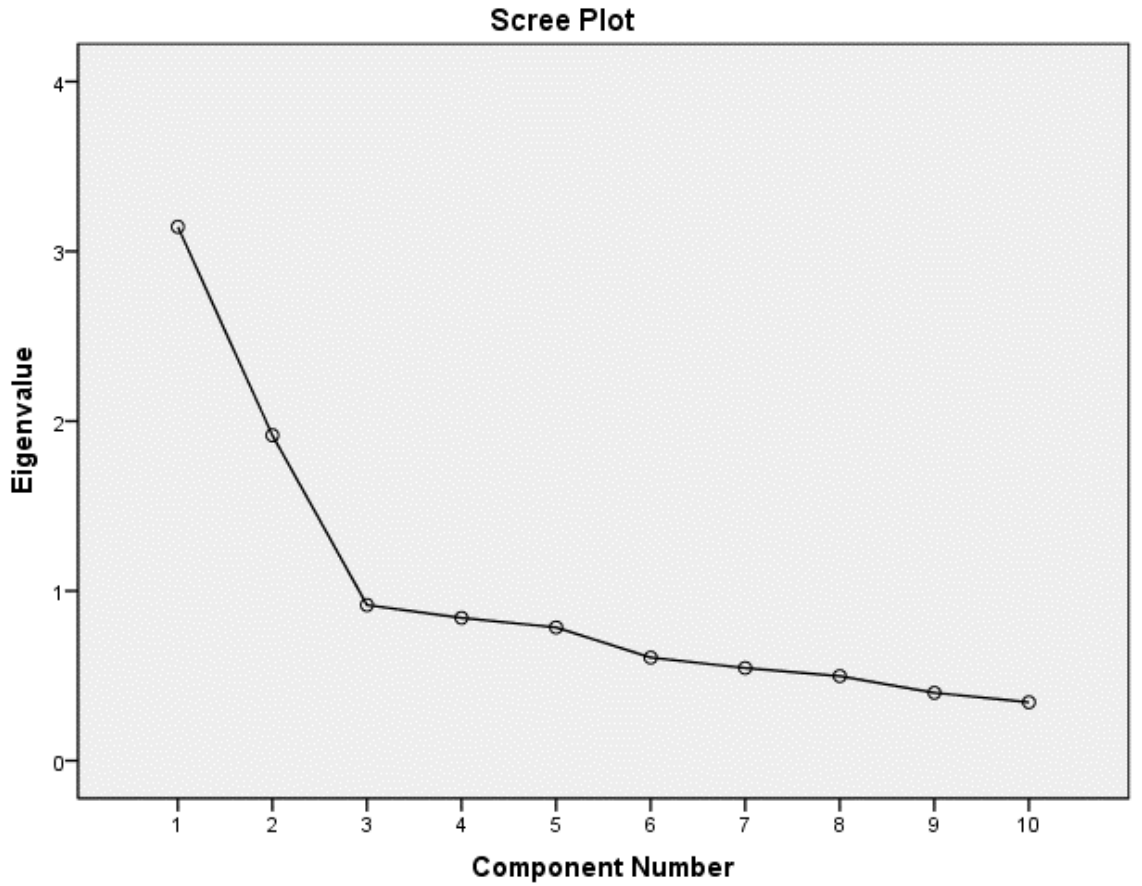
PCA Hyperparameters

Heuristics for Selection:

2. A threshold of eigenvalues.

Ex: *Keep components that have eigenvalues more than 1, or the "Kaiser Criterion."*

This rule is more for keeping the components that "matter the most."



The "elbow rule": but this time upside down!

Application of PCA

K-Means Clustering

- An algorithm to determine implicit groupings in data based on correlations in features.

- Determines K clusters from N observations where each observation belongs to the cluster with the ***nearest mean, or cluster centroid***.

- Minimizes within-cluster ***squared Euclidean distances.***

How it Basically Works (Lloyd's algorithm)

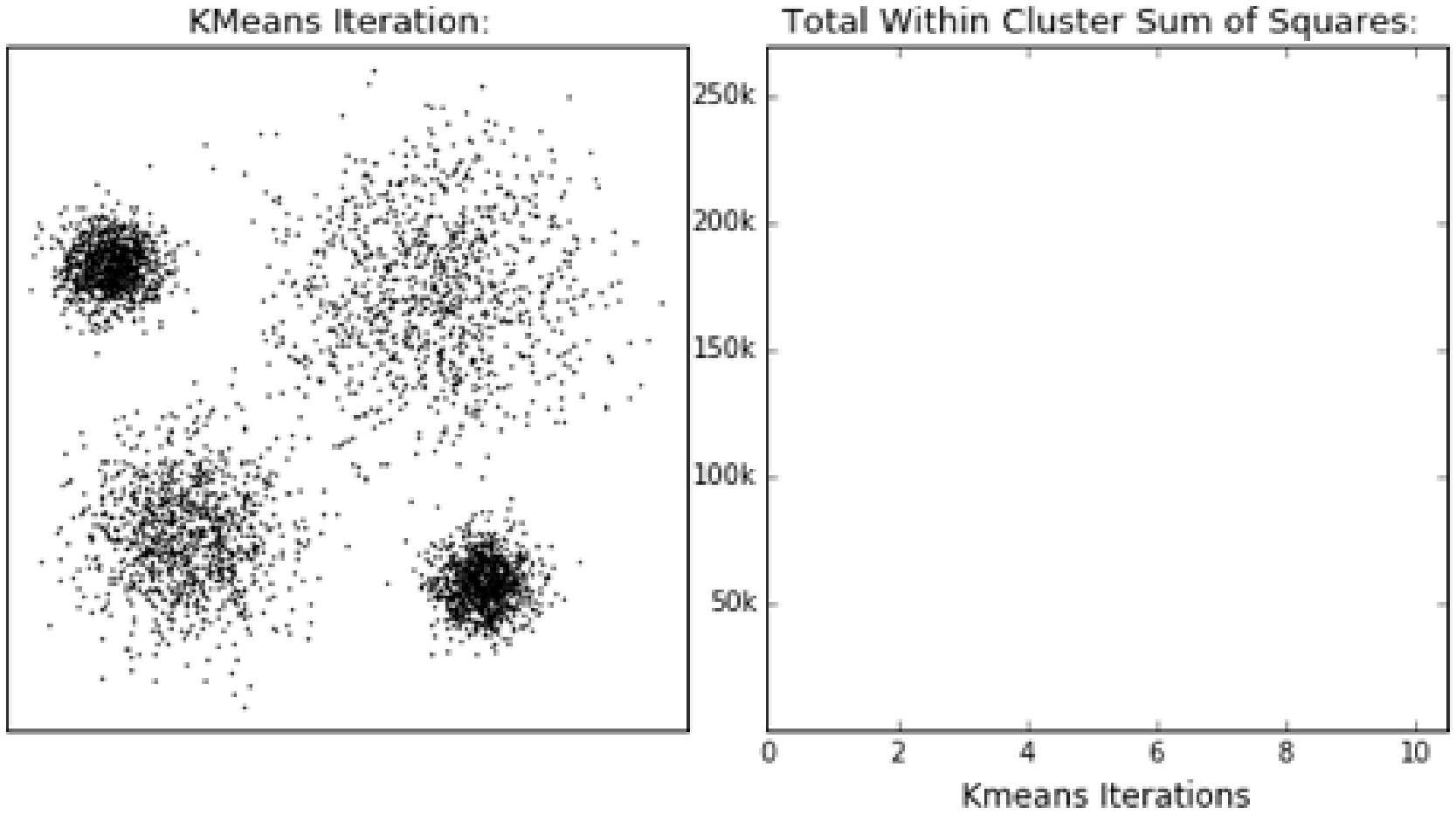
1. Assign K initial centroid points.

2. Assign each observation to the centroid which is the closest (minimum Euclidean distance).

3. Recalculate the means for each cluster.

4. Repeat 2 and 3 until the means stop moving.

How it Basically Works (Lloyd's algorithm)



How it Basically Works (Lloyd's algorithm)

- This method is **not guaranteed to reach the minimum.**

- There are several variations of this method:
  - K-medians clustering (uses medians).
  - K-mediods clustering (uses mediod points that are actually in the data).
  - Gaussian mixture models (generalized version of K-means).

Example: Dimension Reduction

- An Econometrica paper by Bonhomme and Manresa applies a clustering algorithm to reduce the number of fixed effects necessary in a model.

- They refer to this as a “group fixed effects estimator.”

- They benchmark this method against Mixture Distributions, a more widely used method in economics.

Example: Data Exploration with PCA

- K-means clustering algorithm is fairly useful for trying to understand the implicit patterns in our data.
- PCA is often used as a first step to reduce the dimensions of the data. The clustering algorithm is then applied to find groups.
- The groups may or may not be economically meaningful, but it can be a good starting point.

K-Means Hyperparameter

The main hyperparameter that must be chosen is the number of clusters.

How do we know which number is the best?

K-Means Hyperparameter

The main hyperparameter that must be chosen is the number of clusters.

How do we know which number is the best?

2 is likely to not be helpful, whereas K=N is equivalent to what we already have.

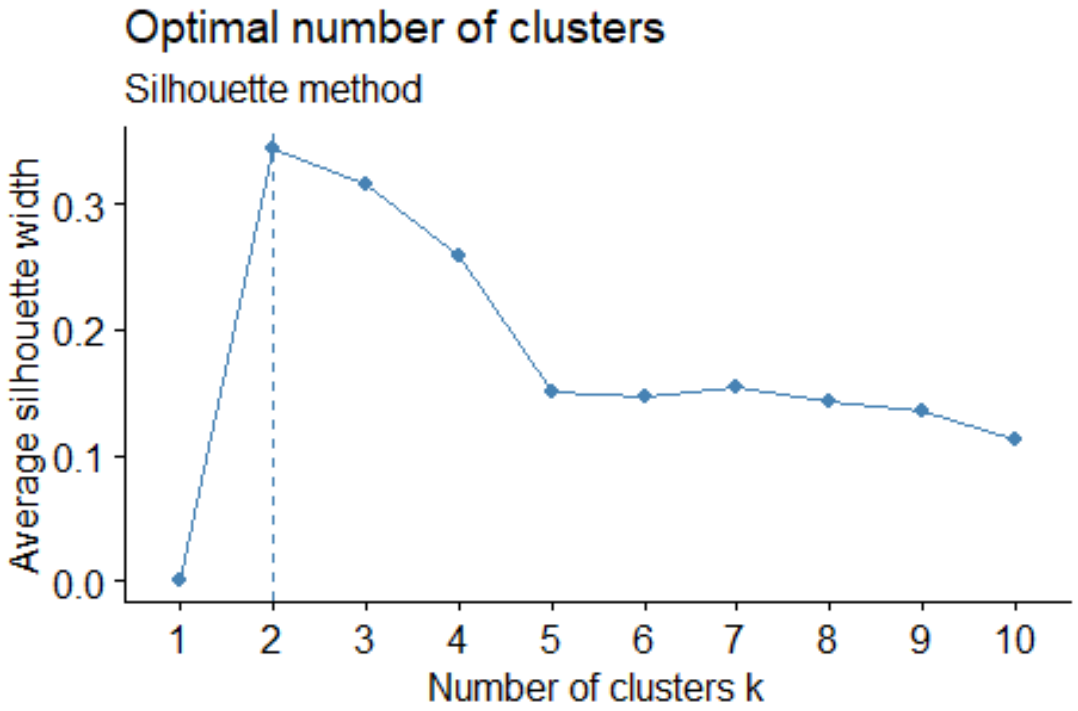
We need just enough clusters to see groupings.

Heuristics for Selection

Heuristic #1: Choose the K at the elbow of the average silhouette score.

Silhoutte score measures the distance between points and their cluster centers.

Obs have positive values if they are closest to their assigned cluster, negative if closer to another cluster.

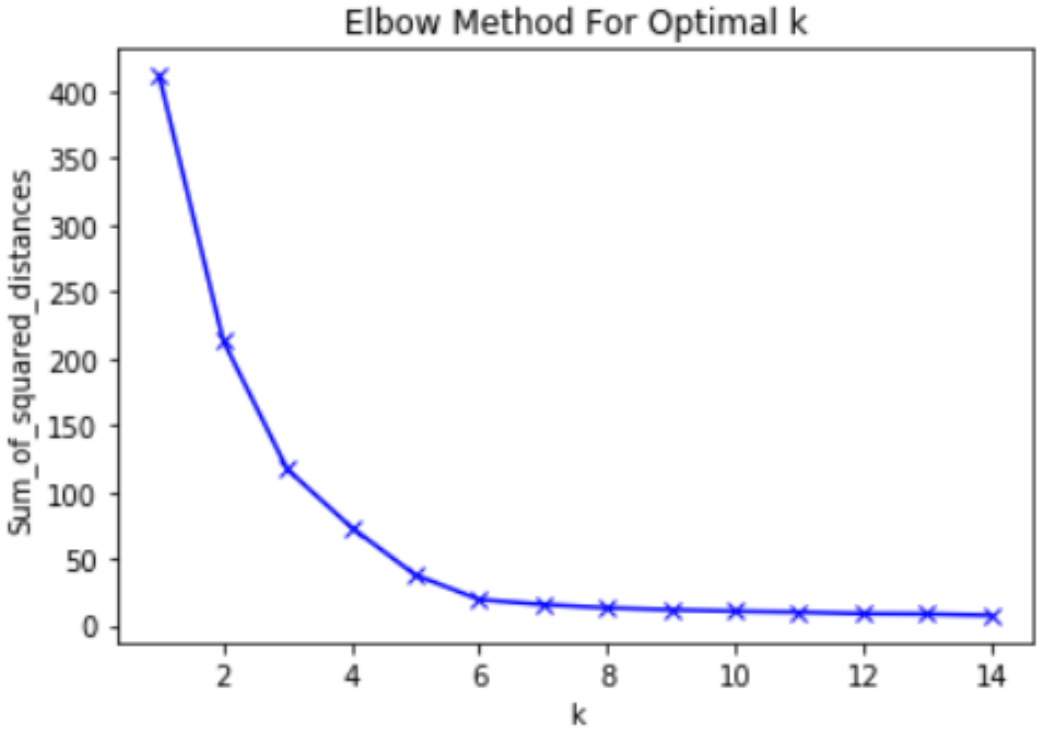


Heuristics for Selection

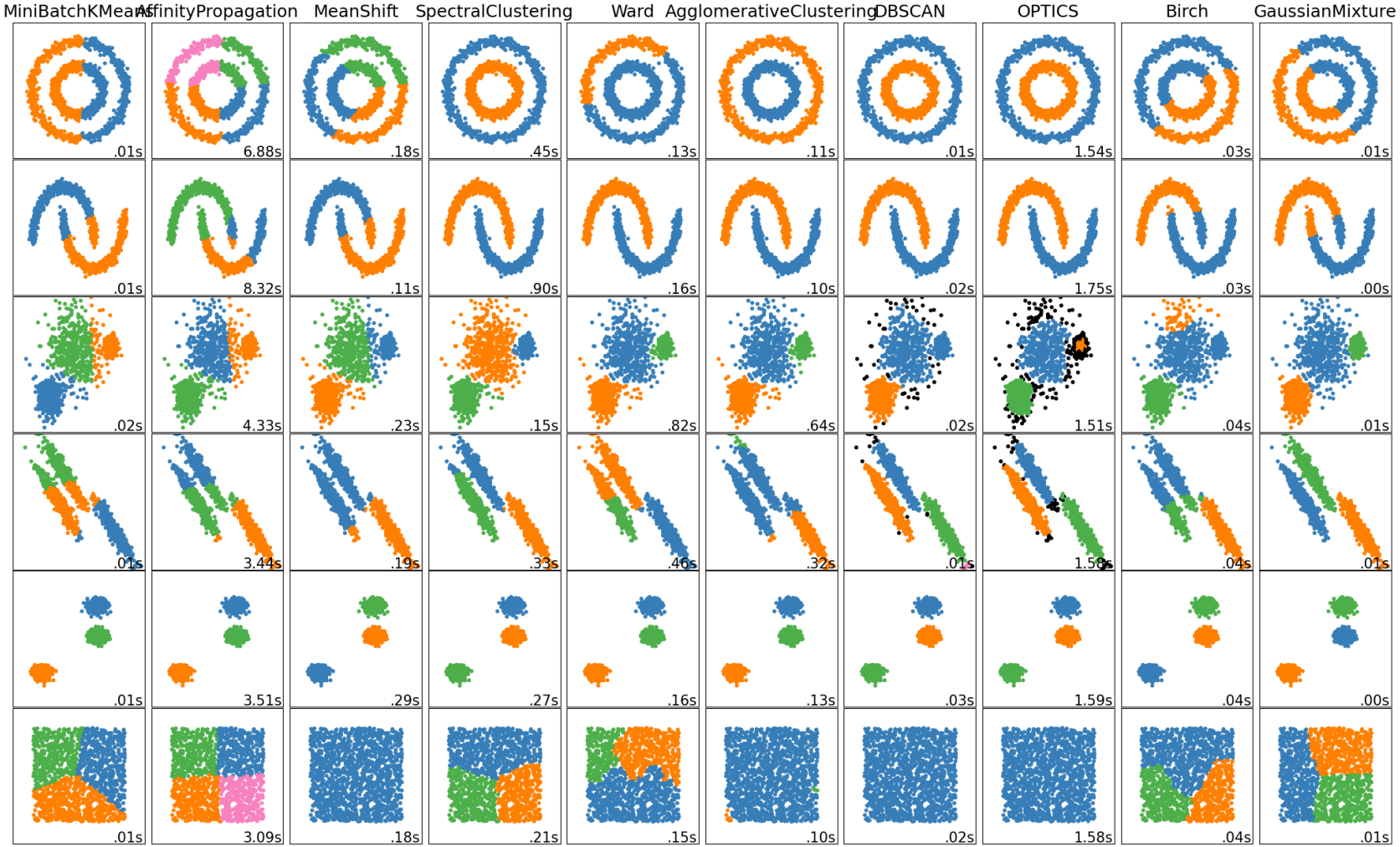
Heuristic #2: Choose the K at the elbow of the within-sum of squared distances (called "inertia").

Inertia is essentially the function being minimized. In theory, this should be smallest for the optimal number.

Because K-means doesn't always find the minimum***, it won't always decrease monotonically.***



Alternative Clustering Methods

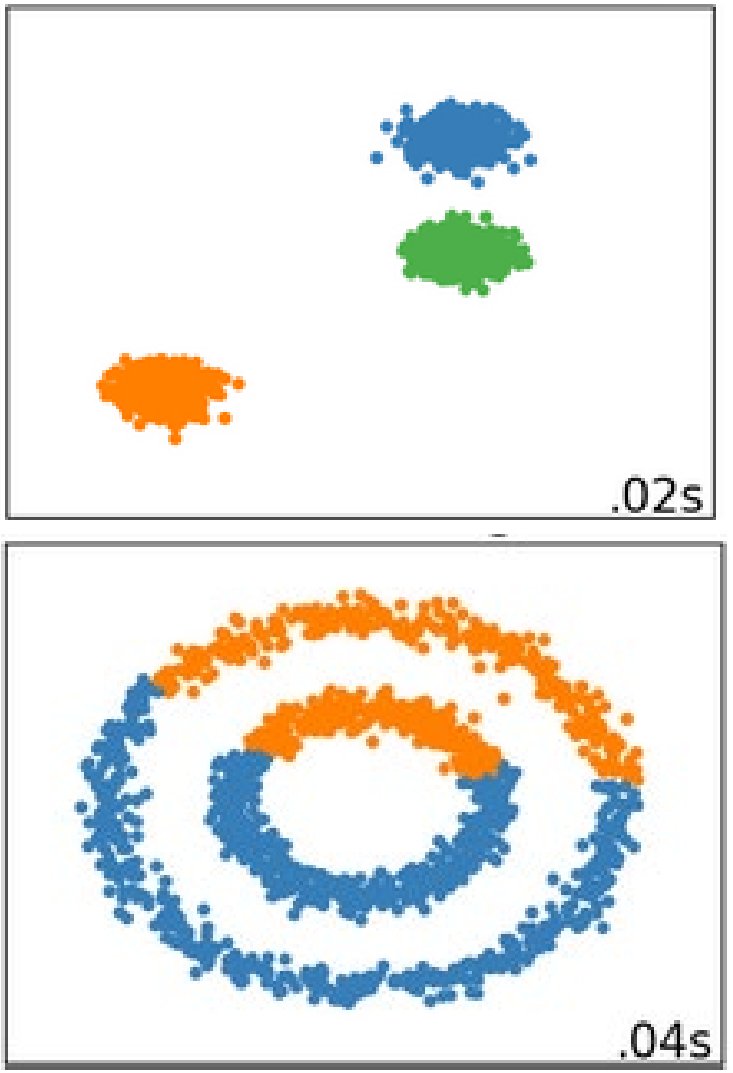


Weaknesses of K-means

- It does well for ***compactness***, but not well for ***connectivity***.

- For cases in which there is high connectivity, ***Spectral Clustering*** or ***Hierarchical Clustering*** is a better option.

Good for K-means



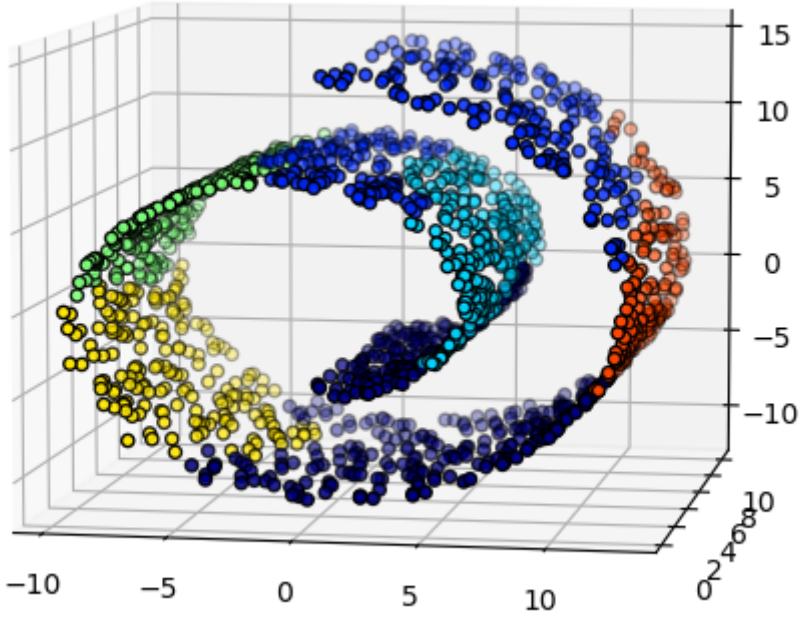
Bad for K-means

Imposing Constraints

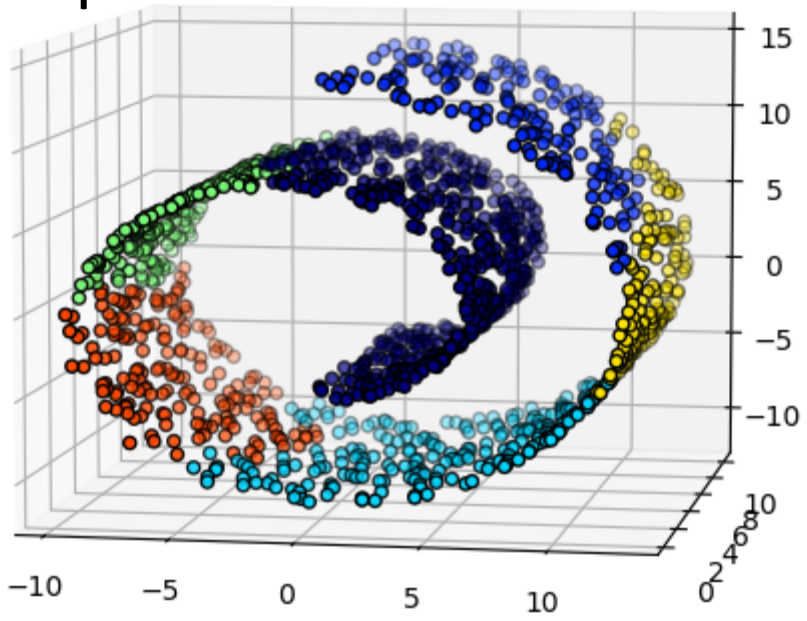
- Some methods can allow you to impose constraints such as ***Hierarchical Clustering***.

- This helps if you have a priori information about the shape.

No constraints



Shape constraints



Application of K-means