3. What is the elbow method in clustering? And which flaws does it pose to assess quality?

The elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.

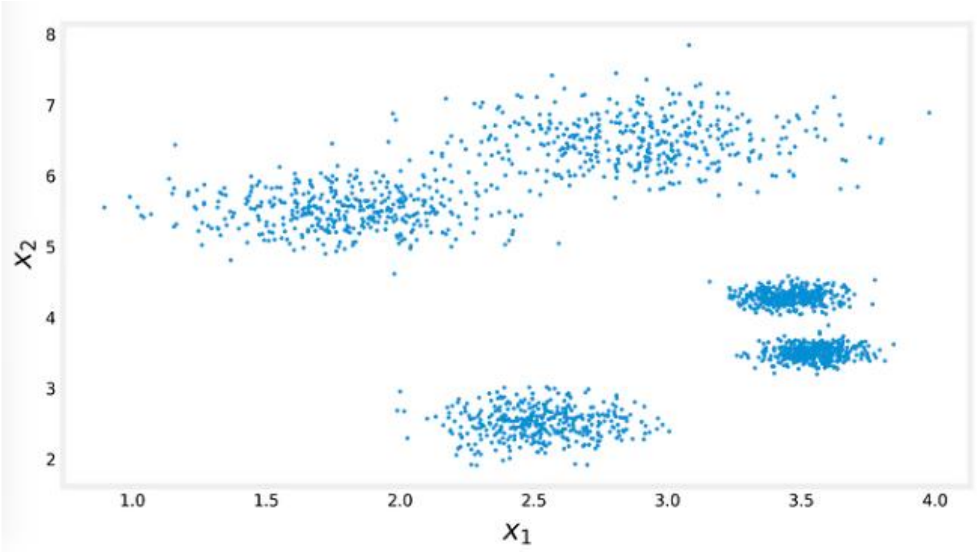
The intuition is that increasing the number of clusters will naturally improve the fit (explain more of the variation), since there are more parameters (more clusters) to use, but that at some point this is over-fitting, and the elbow reflects this. For example, given data that actually consist of k labeled groups – for example, k points sampled with noise – clustering with more than k clusters will "explain" more of the variation (since it can use smaller, tighter clusters), but this is over-fitting, since it is subdividing the labeled groups into multiple clusters. The idea is that the first clusters will add much information (explain a lot of variation), since the data actually consist of that many groups (so these clusters are necessary), but once the number of clusters exceeds the actual number of groups in the data, the added information will drop sharply, because it is just subdividing the actual groups. Assuming this happens, there will be a sharp elbow in the graph of explained variation versus clusters: increasing rapidly up to k (under-fitting region), and then increasing slowly after k (over-fitting region).

There are various measures of "explained variation" used in the elbow method. Most commonly, variation is quantified by variance, and the ratio used is the ratio of between-group variance to the total variance. Alternatively, one uses the ratio of between-group variance to within-group variance, which is the one-way ANOVA F-test statistic.

Flaws:

- The elbow method is considered both subjective and unreliable. In many practical applications, the choice of an "elbow" is highly ambiguous as the plot does not contain a sharp elbow. This can even hold in cases where all other methods for determining the number of clusters in a data set agree on the number of clusters.

- The K-Means algorithm aims to minimize the mean squared distance between each instance and its closest centroid, defined as inertia. However, from this definition, arises a problem. As long as we keep increasing the number of clusters k, the inertia will always decrease because the points will be closer to their centroids. Therefore, when choosing the right number of clusters for K-Means, we are looking for the minimum number of clusters that gives us reasonable inertia. That's exactly what we try to achieve using the Elbow Method. Suppose we have the following data:



For humans, it may be clear that the data come from 5 different clusters, but when dealing with high-dimensional data, we are unable to visualize it with ease.

Using the Elbow Method, we would probably choose k = 4, as indicated on the left plot. Note that, since two of the clusters are relatively close to one another, the Elbow Method leads us to think that those clusters are just one because if we place a centroid in-between both clusters, the relative distance from the data points to it will be short. Thus, we need a more precise, rigorous, and reliable approach to define the optimal number of clusters for our clustering task.