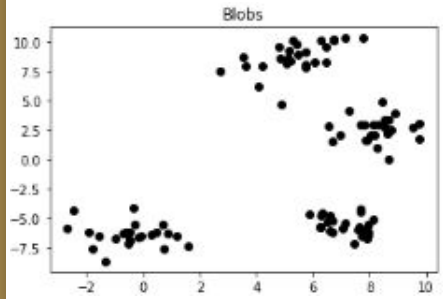
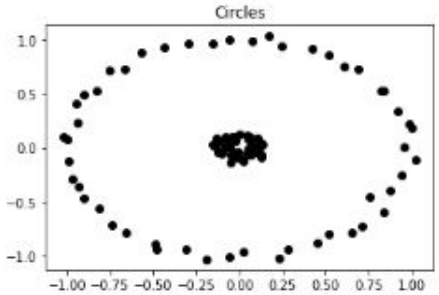
Impact of linkage on hierarchical clustering and covariance type on GMM clustering

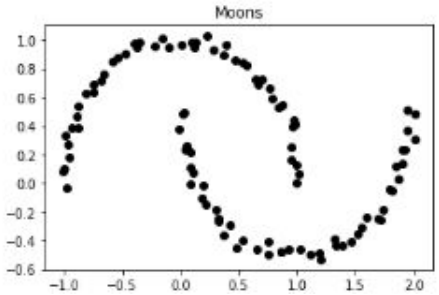
Agaton Fransson

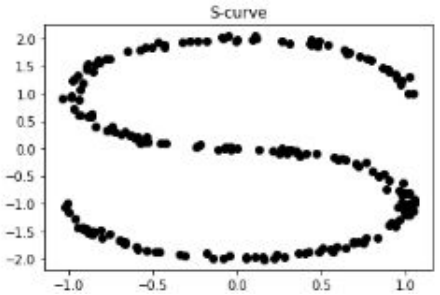
Different types of linkage data sets

- Used python's sklearn.dataset to generate different shapes
- We see four "blobs", one circle within a circle, two moons and one s-curve



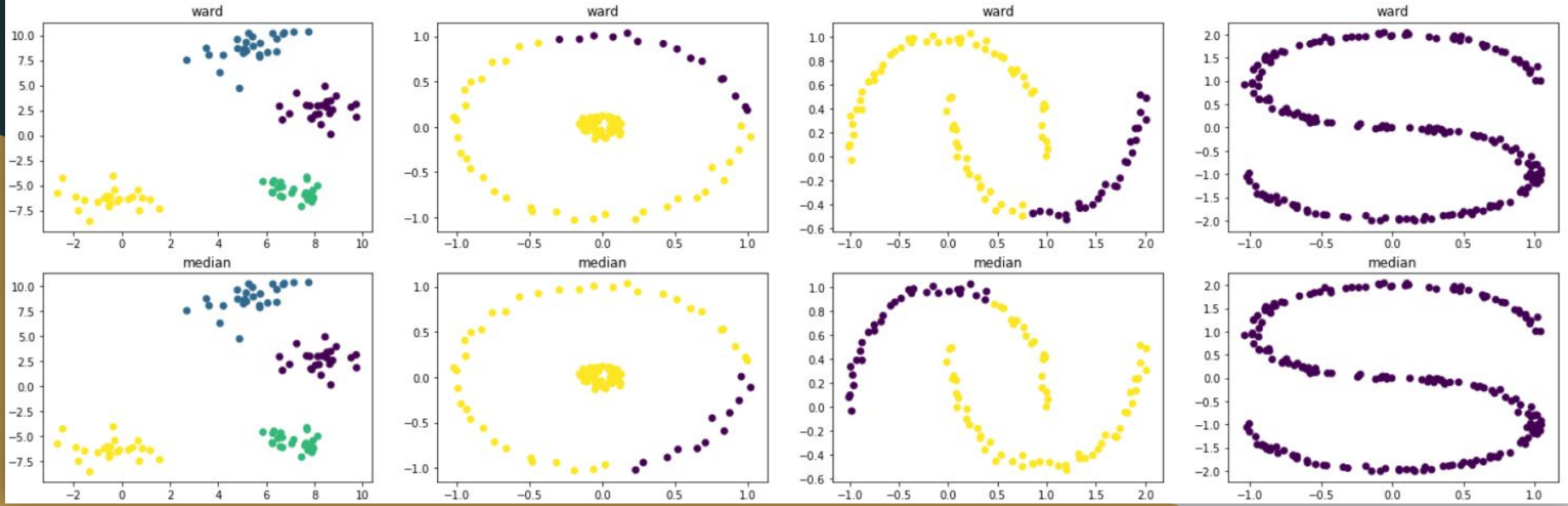






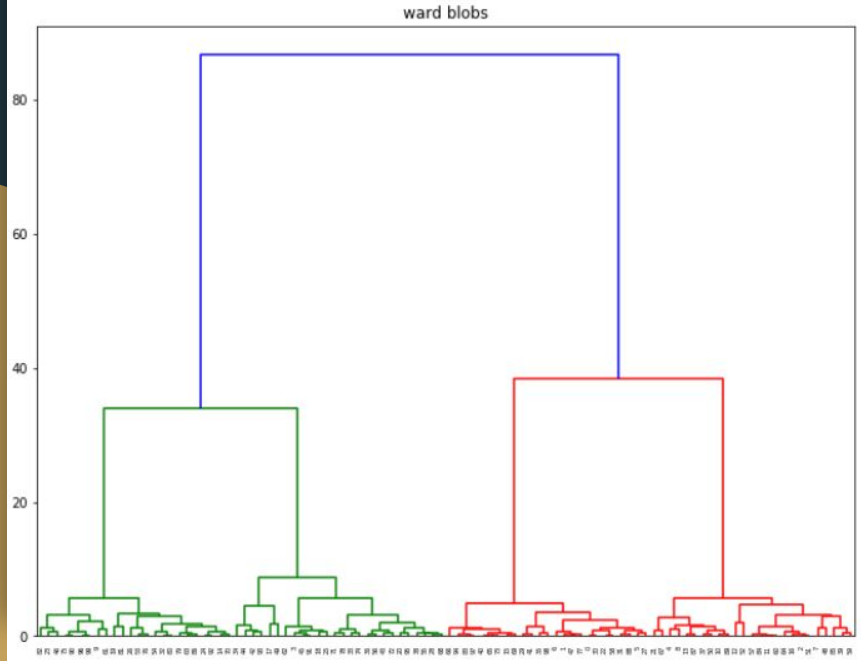
Linkage types

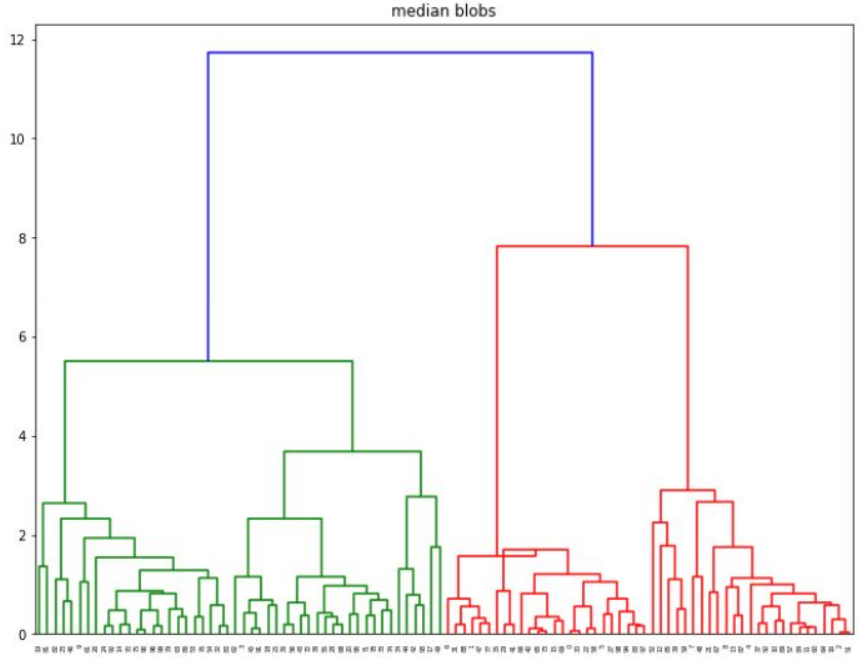
- Used the “ward”- and “median”-methods to calculate the distance between the newly formed clusters.
- To find clusters the “maximum clusters”-criteria was used and set to the number of clusters for each shape.
- We see that both methods captured the “blobs” and the “S”, but the “S” only has one cluster so they couldn’t fail to find it. The amount of success is seen in the dendrograms.



Dendrograms of success

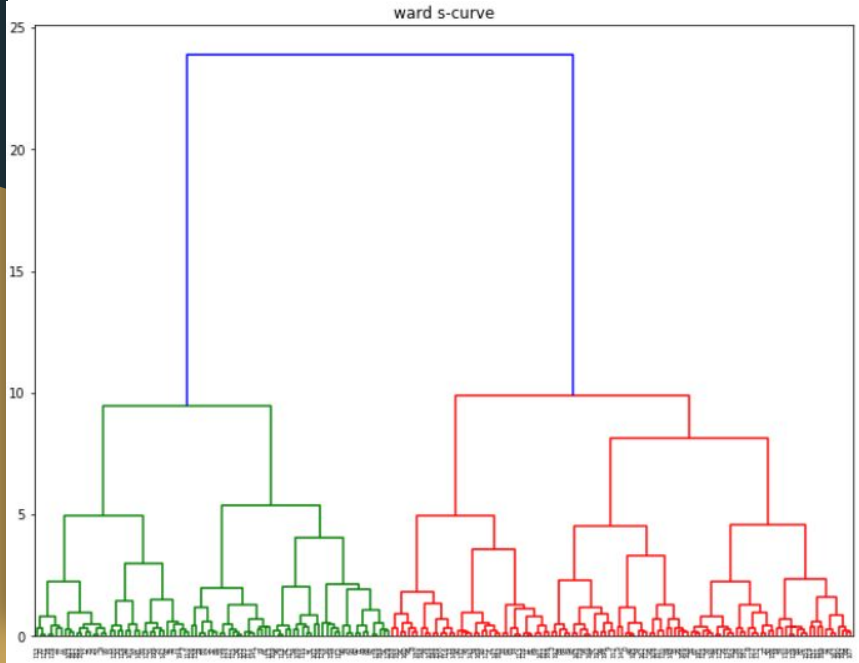
- We see that the ward-method has a very big jump from 4 to 3 clusters which suggests that it was sure of there being 4 clusters.
- For the median-method the percentual increase in distance between the clustering is pretty constant, which suggests that it is a worse method for this purpose, although it succeeded in finding the correct clusters.

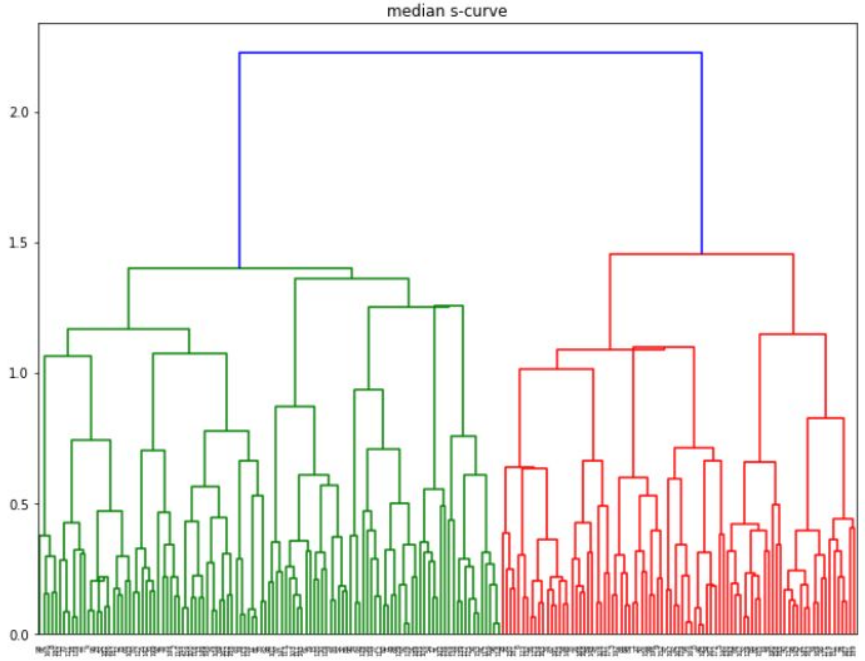




Dendrograms of failure

- These dendrograms suggests that both methods wanted there to be two clusters due to the big distance.
- The consistent increase in distance suggests that neither method was secure in finding the cluster, the dendrograms for the other failed clusters looked very similar.



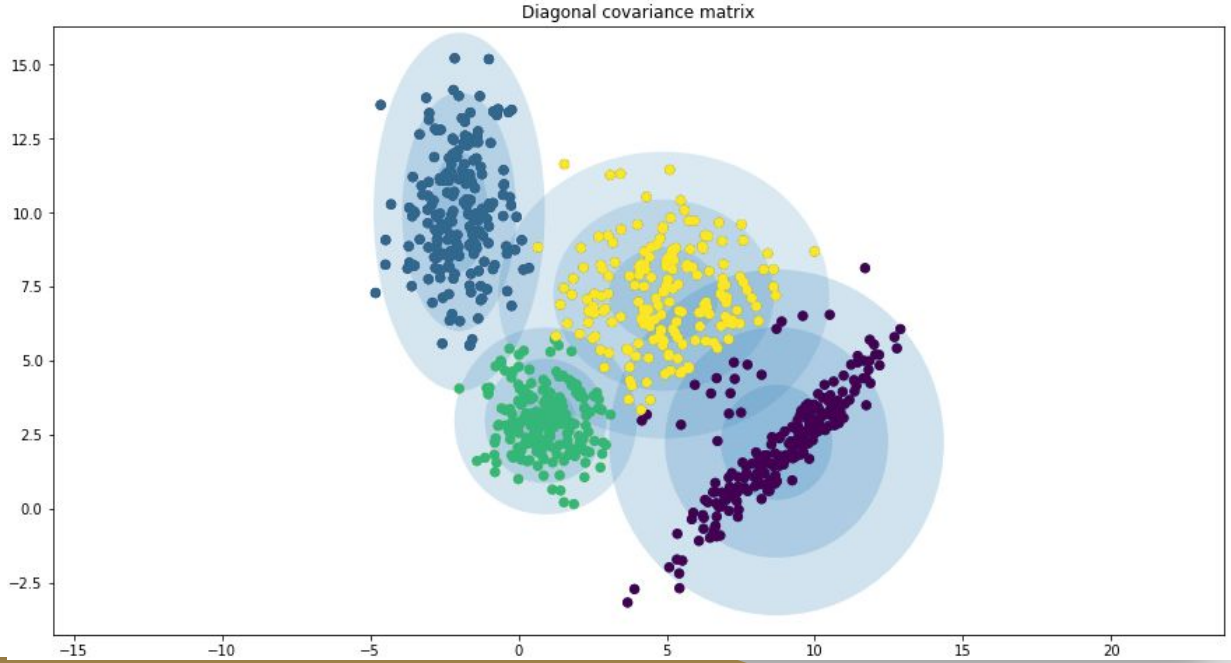


Conclusions

- Both the median- and ward-method were good at finding clusters when they are separated and shaped like circles
- Both were bad at finding clusters where the density of the data increased closer to the centre of the cluster

GMM clustering

- 4 multivariate normally distributed data sets with different means and covariance matrices, some diagonal (blue, green, yellow) and some not (purple).
- GMM set to assume data has diagonal covariance matrices
- See that the blue, green and yellow clusters are captured well
- The purple is not captured well and steals some data points from yellow



Conclusions

- The simplification of assuming that the covariance matrix is diagonal for a cluster works well when the covariance matrix of the cluster is diagonal, and bad otherwise, since a diagonal covariance matrix leads to a data distribution that has biggest and smallest variation along the x- and y-axis