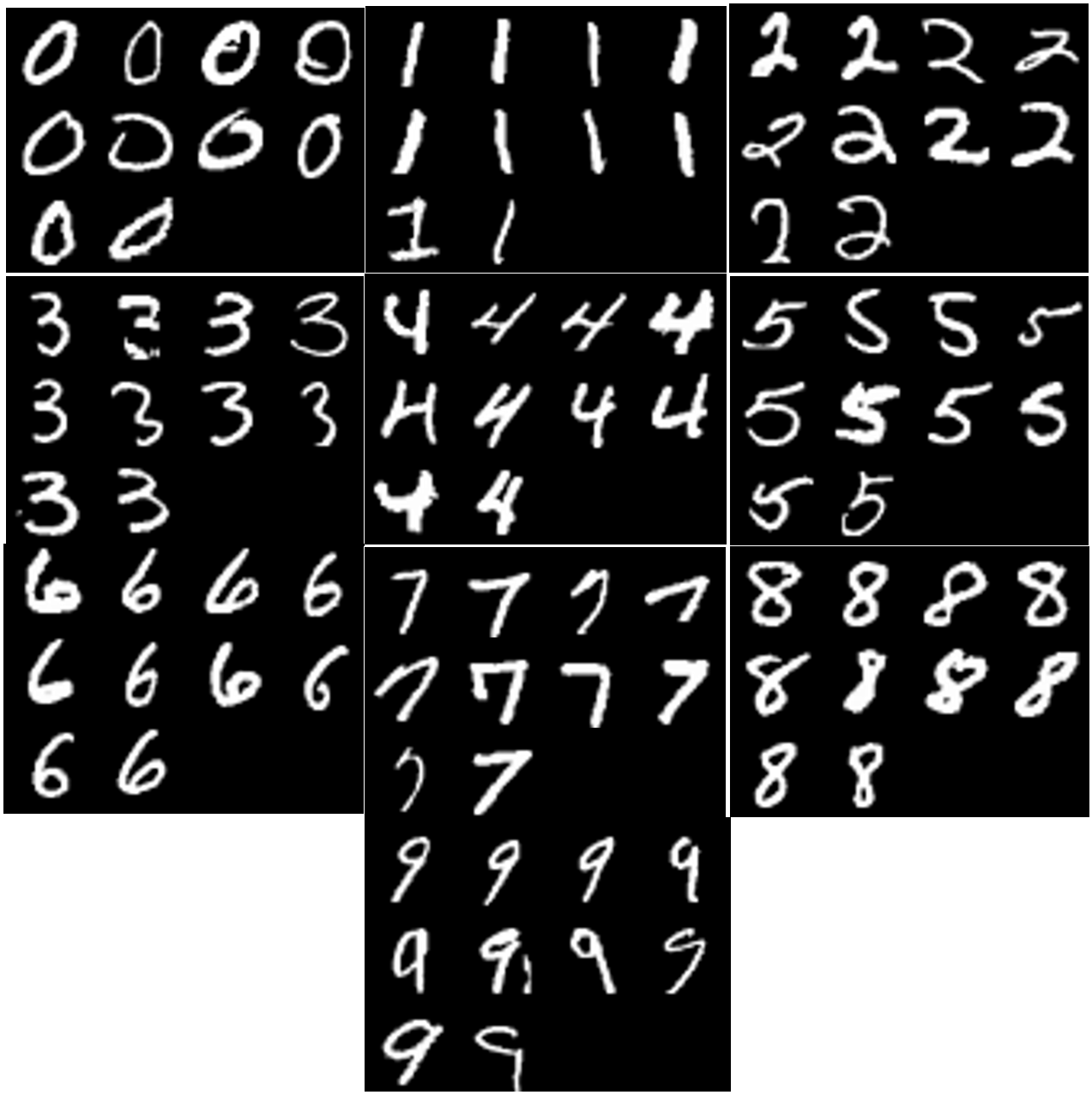
Barnabas Ujvari
S1758100

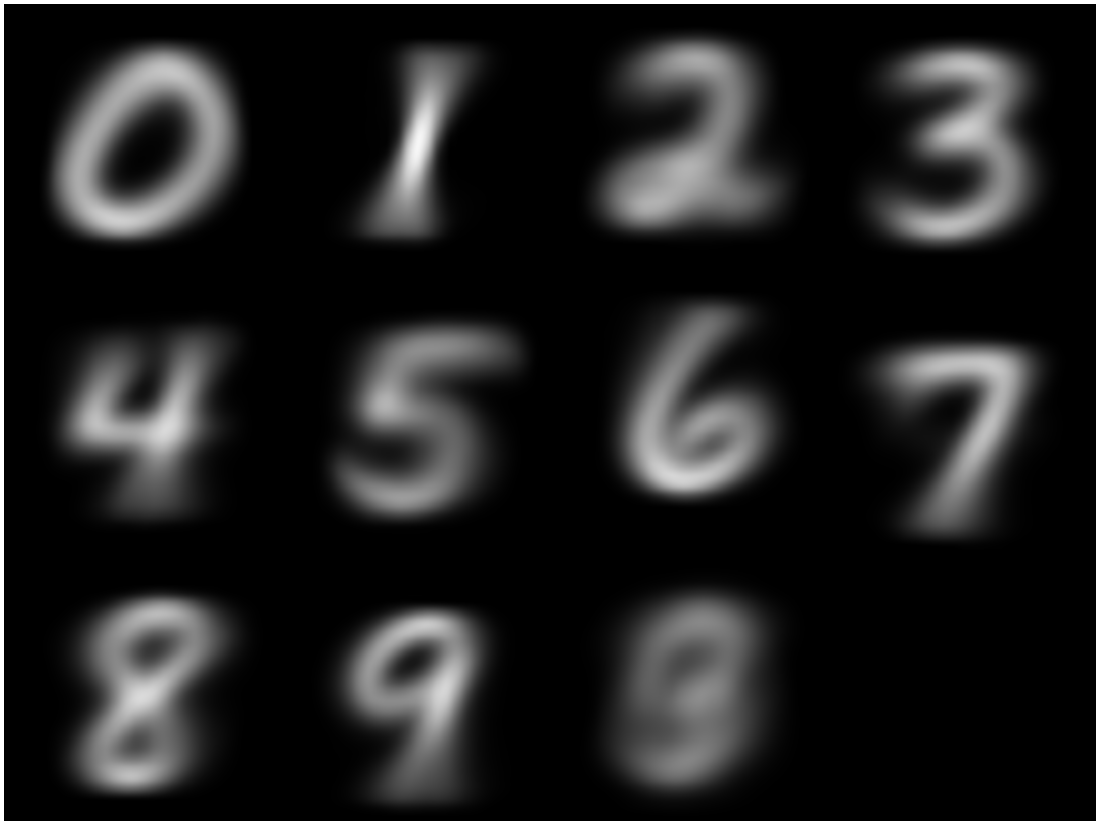
Inf-2B
Coursework 2

Task 1 -PCA and Clustering

Task 1.1

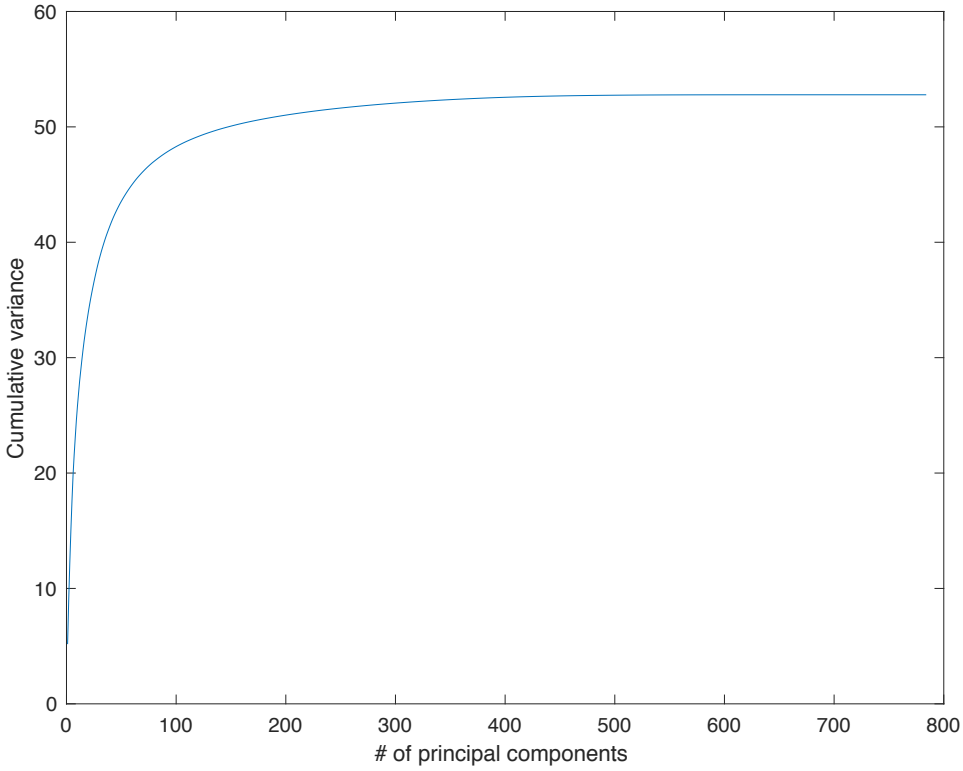


Task 1.2

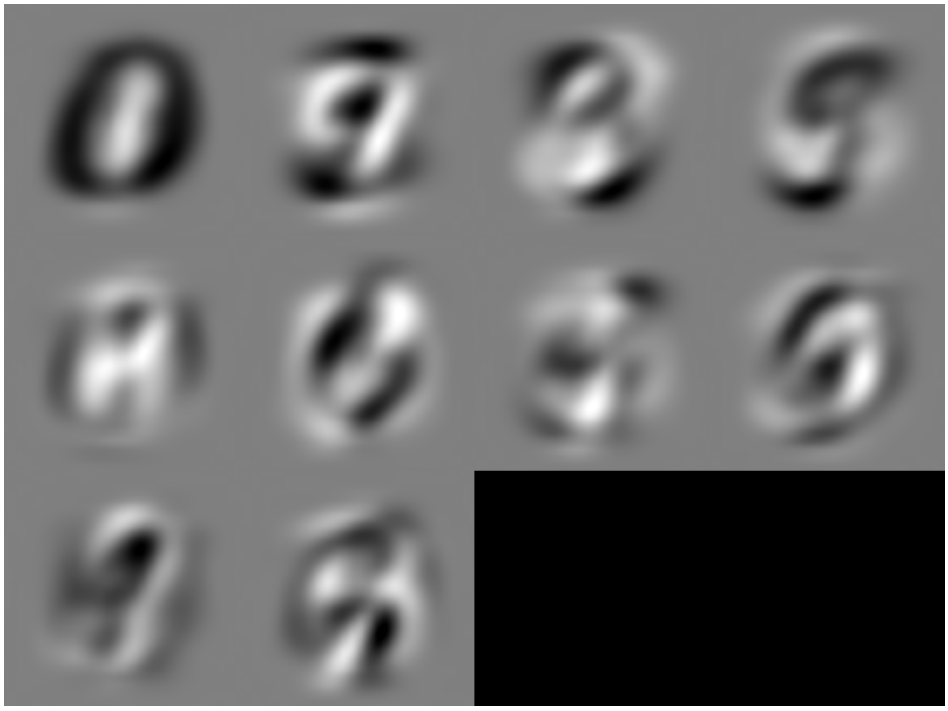


Task 1.3

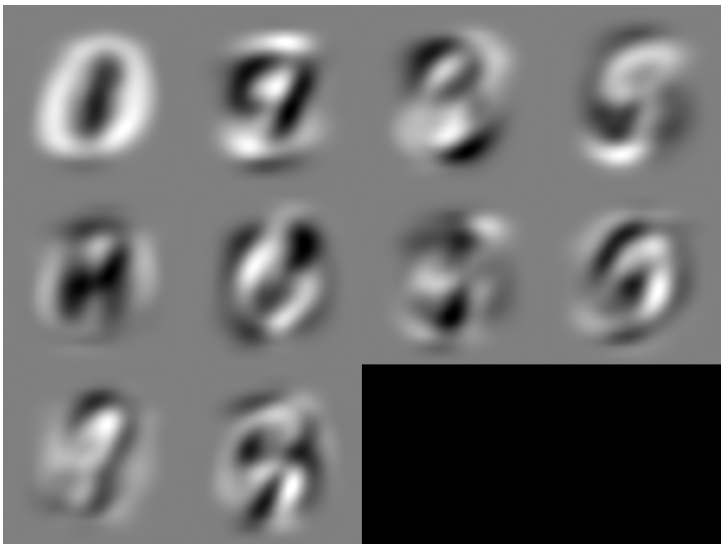
Accuracy aim	Minimum Dimension
70%	26
80%	44
90%	87
95%	154



Task 1.4

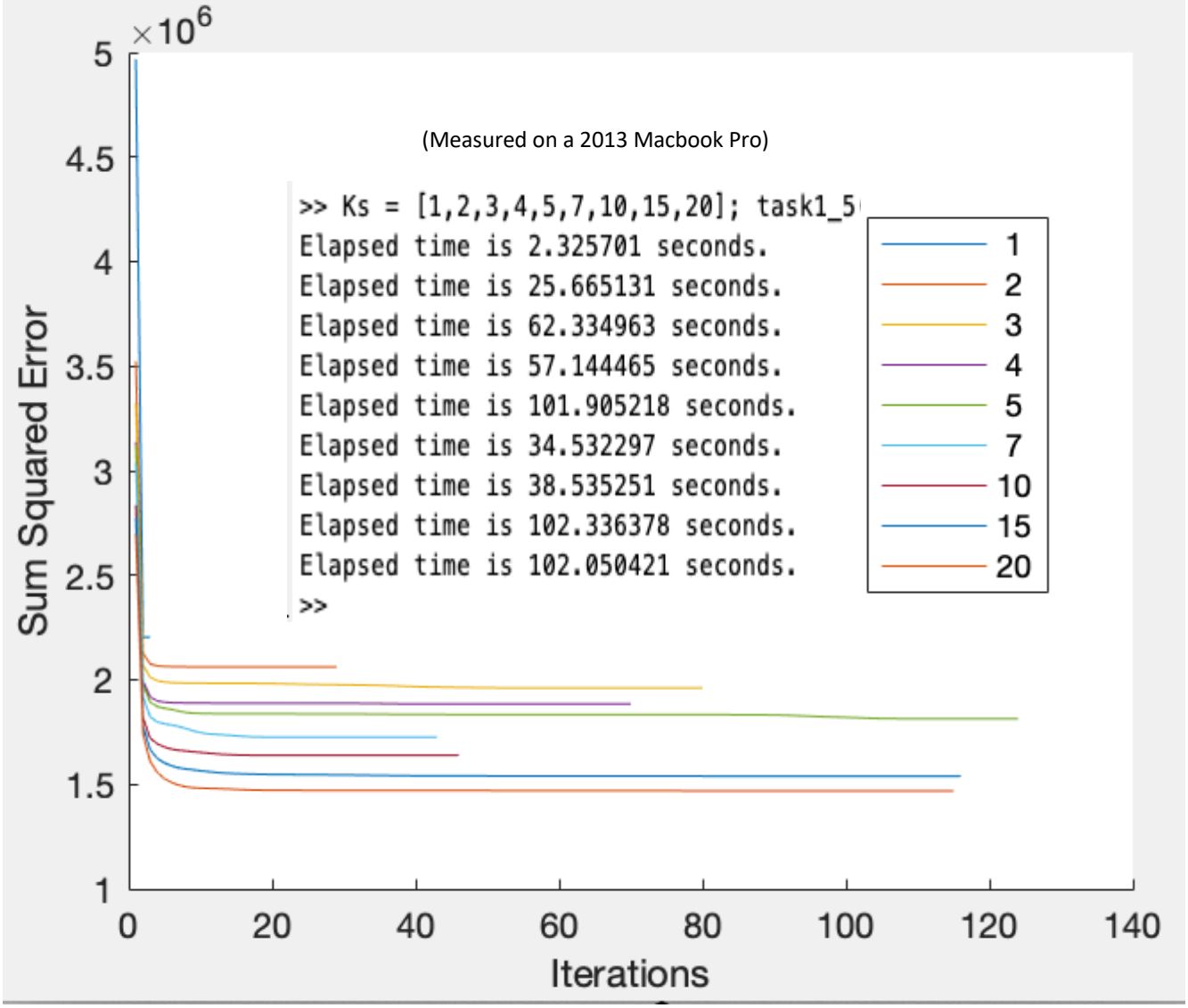


Solution strictly according to the specifications



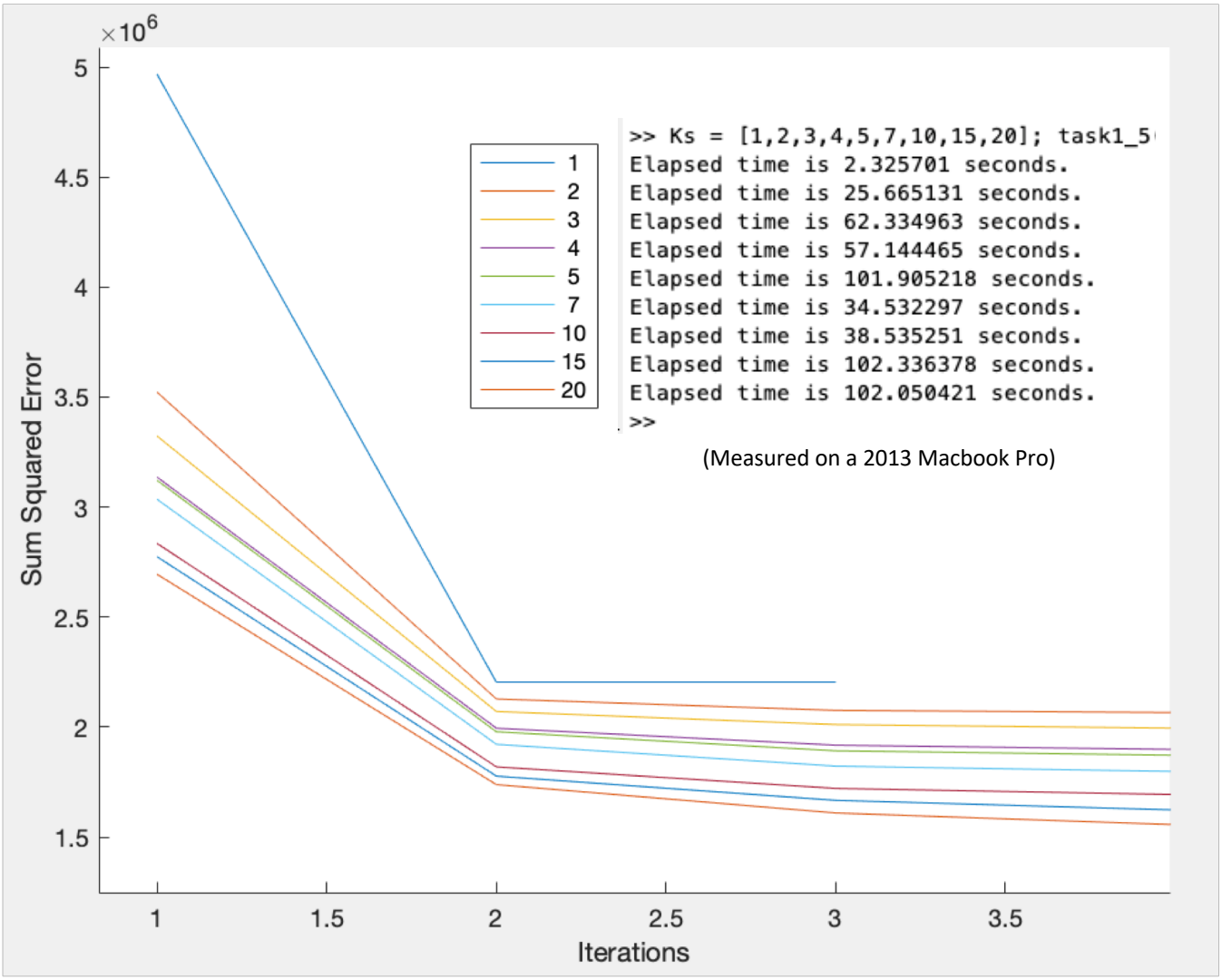
Solution with specific inversion to match the example

Task 1.5



Graph of Sum Squared Error

Task 1.5 (continued)

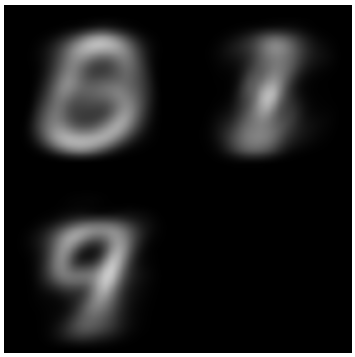


Graph of Sum Squared Error
(close-up on the starting few iterations)

Task 1.6

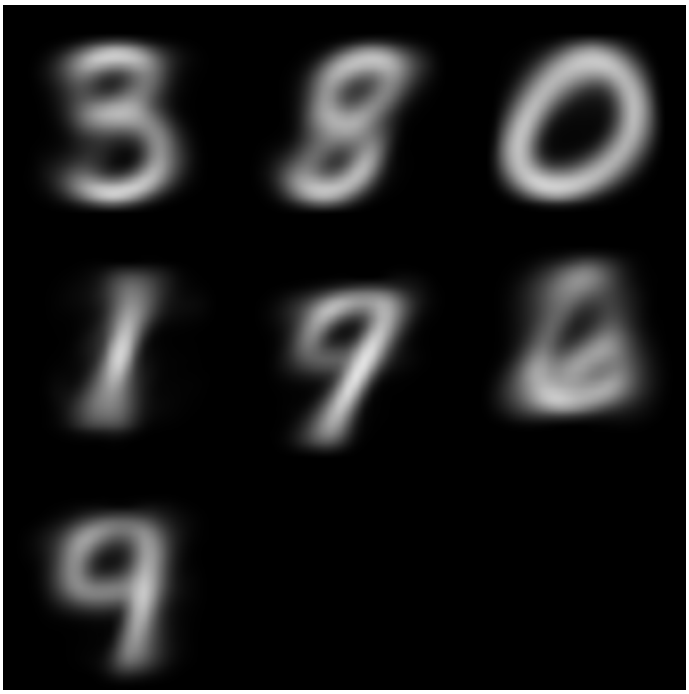
1 Cluster

2 Clusters

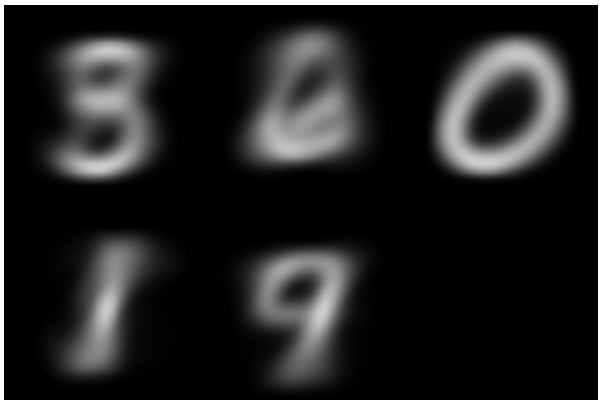


3 Clusters

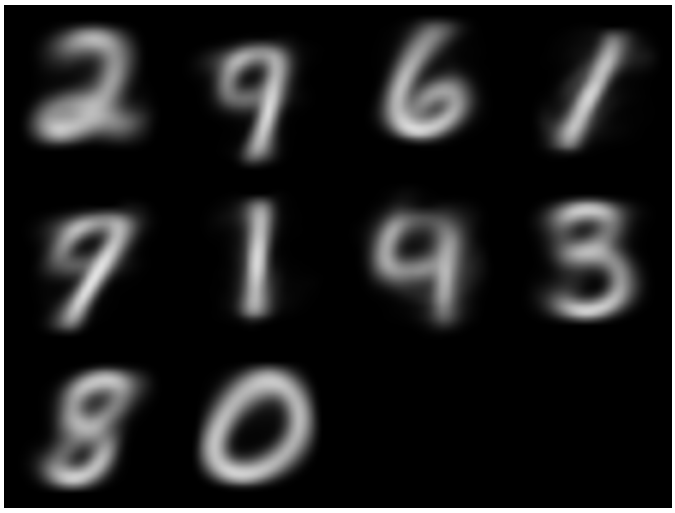
4 Clusters



7 Clusters

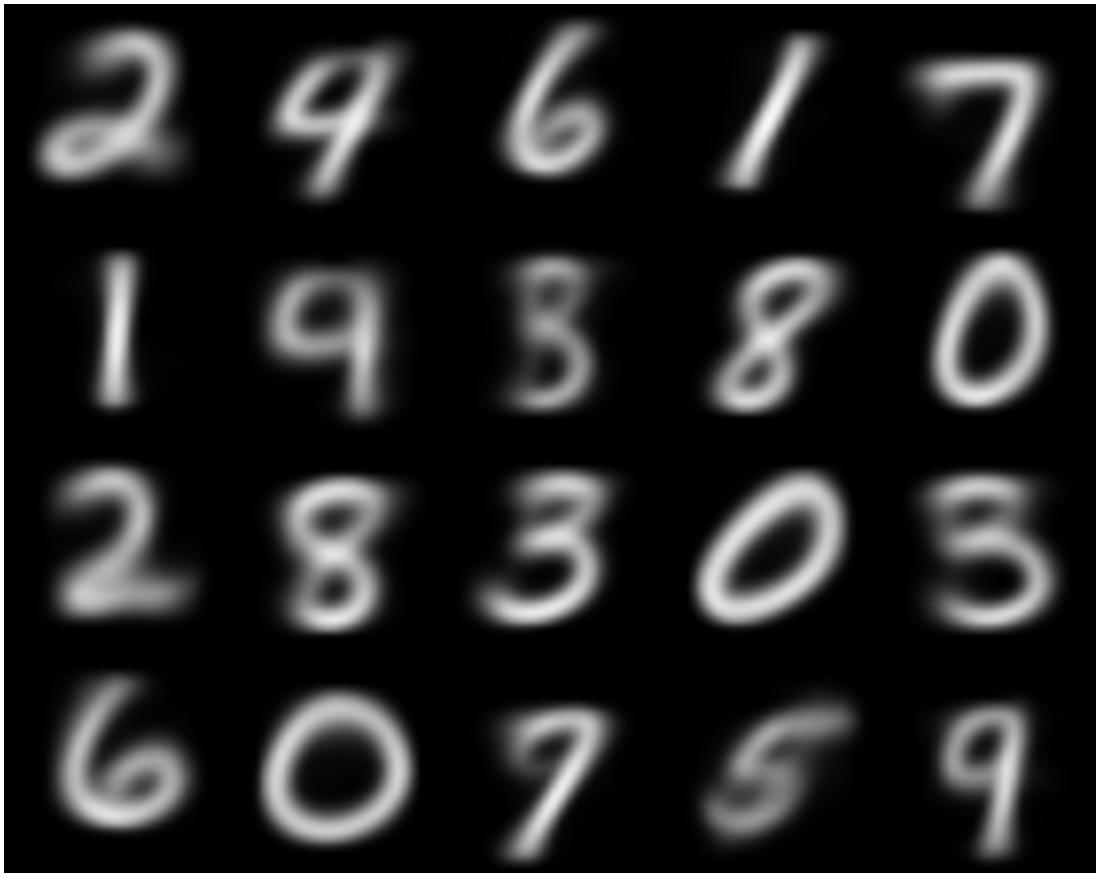


5 Clusters



10 Clusters

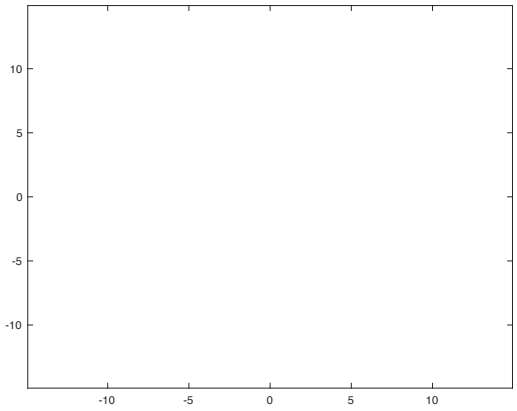
15 Clusters



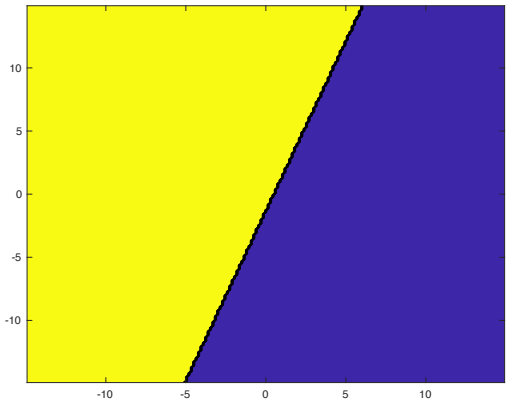
20 Clusters

Task 1.7

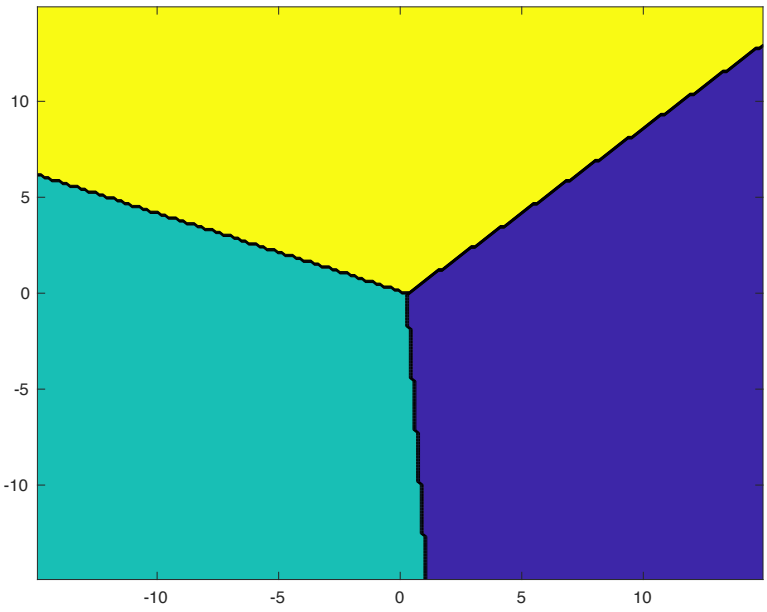
Visualising method: I made a grid of values between the given (m+-5o) and applied the k-means algorithm on each one 2D point in it. This was then passed on to the contour() function to colour the graph, giving the same colour to the same values. These can be seen on the picture below, one for each element in k = 1, 2, 3, 5, 10, representing the number of clusters.



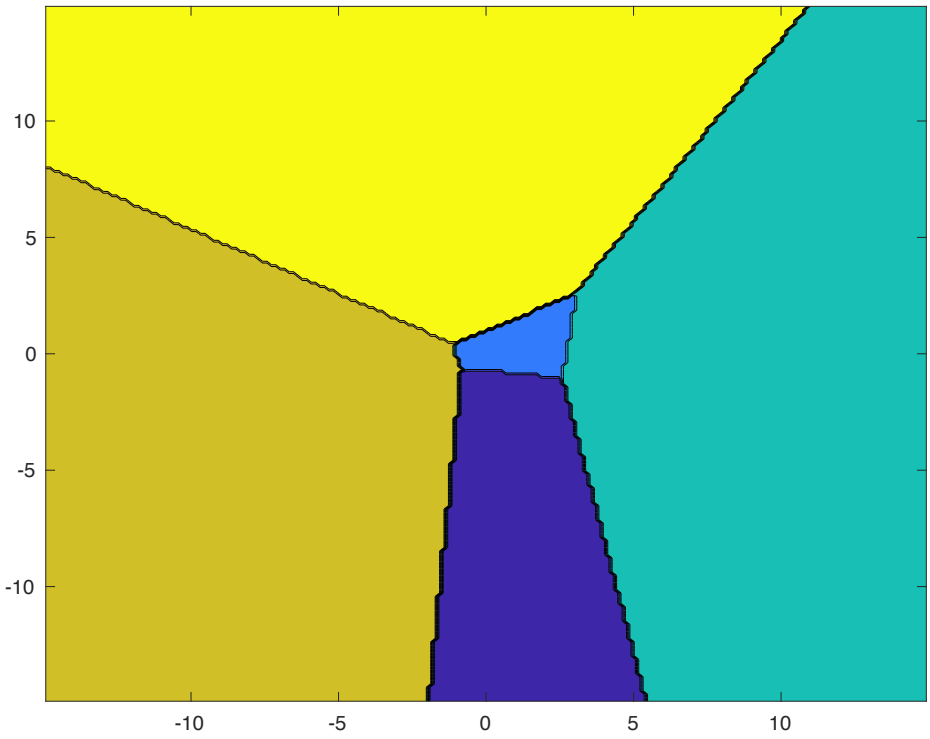
Solid white, as all dots have the same cluster



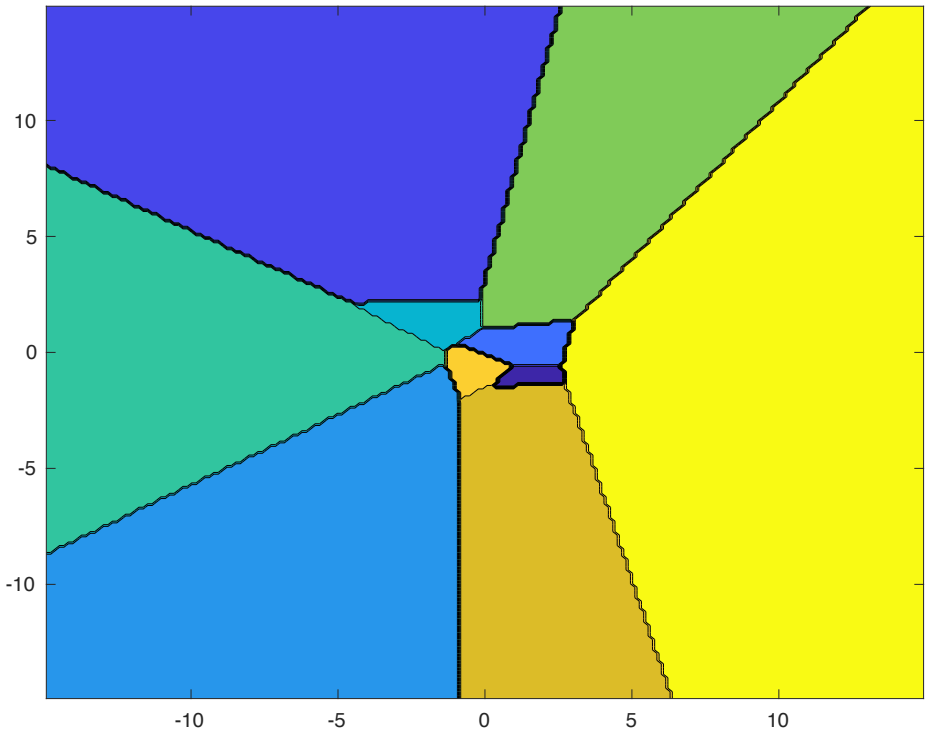
2 Cluster centres



3 Cluster centres



5 Cluster centres



10 Cluster centres

Task 1.8 -Research

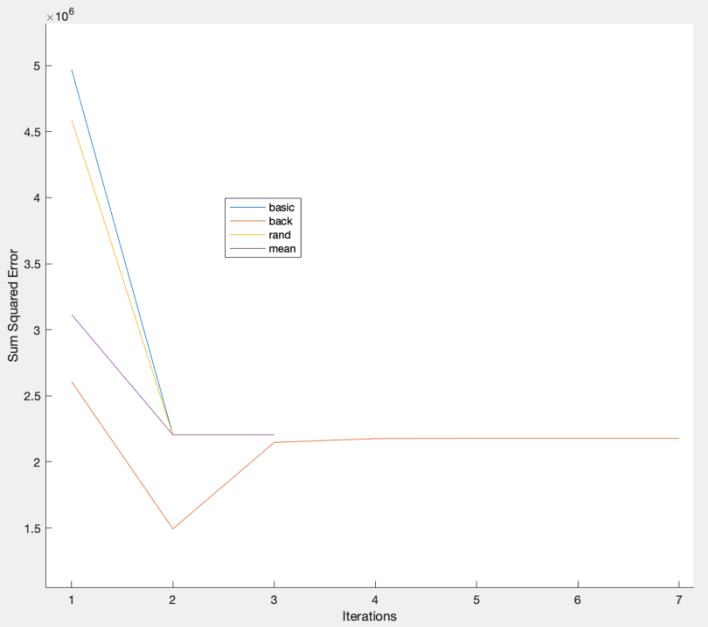
Hypothesis: `Choosing the cluster centers have no significant effect on the final SSH, we always get very similar results.`

My research method:

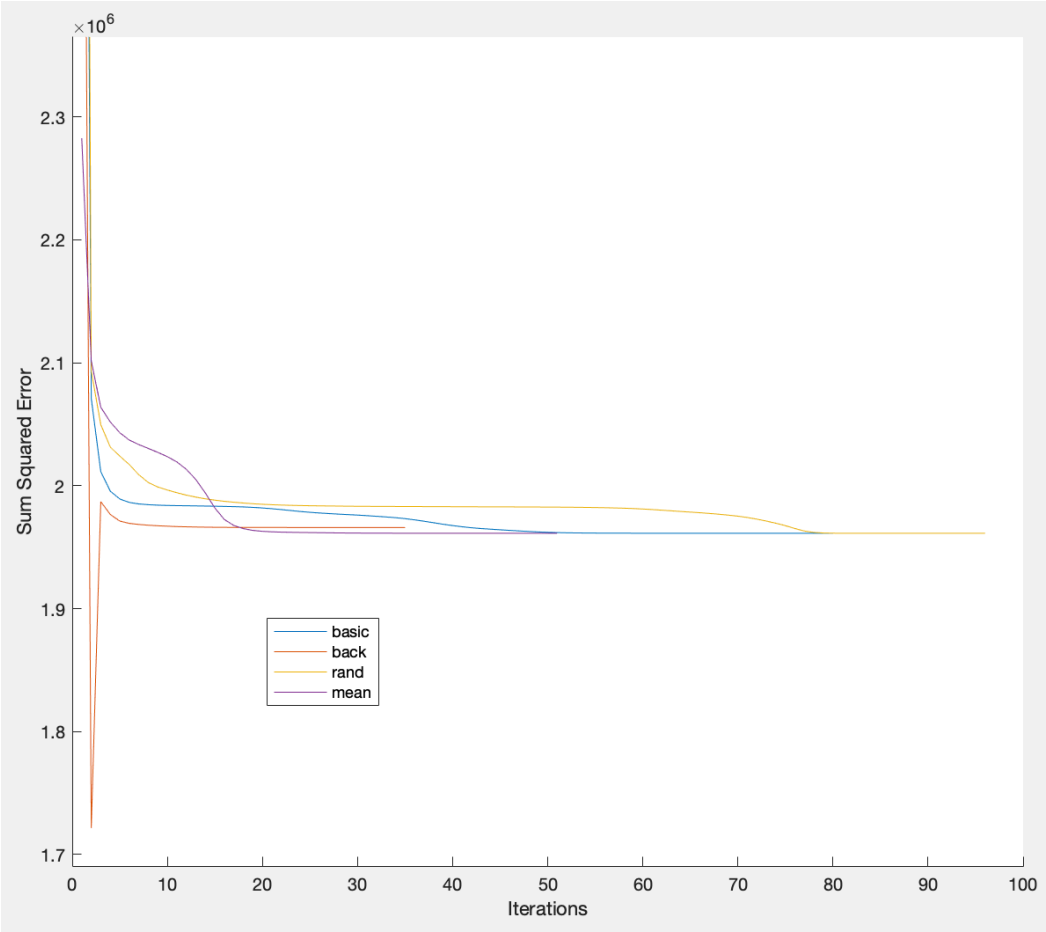
Calculate and plot 4 different cluster choosing method's results for 6 different number of clusters. I intend to back my hypothesis by the graphs of these calculations and brief notes.

4 methods for choosing:

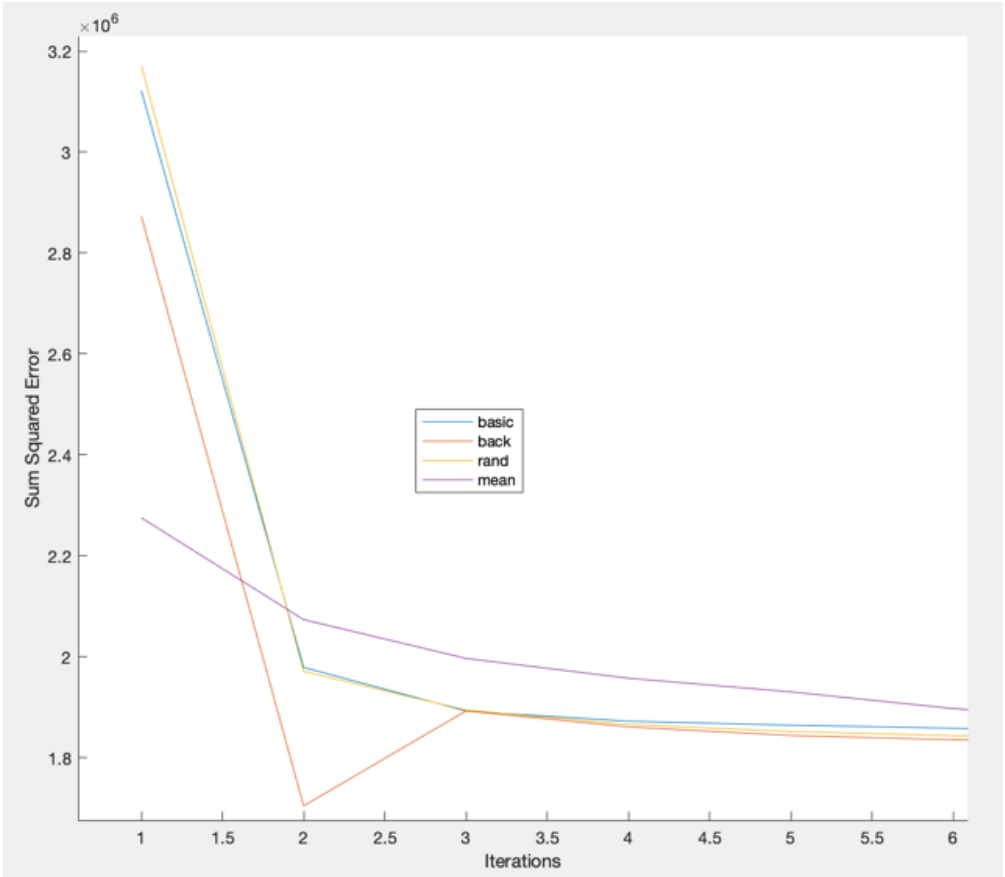
1. **The basic method**: using the first k element of the data set
2. **Last k elements** of the data set: a different group of similarly consecutive values
    a. should make no notable difference(may with coincidence)
3. **Random**ly choosing the k elements from the set:
    a. should be the most consistent result on different datasets, but we didn't have a similar dataset to try it on.
4. Choosing from **Mean** of the classes : k random value, each corresponding to a mean value forms the initial centers.
    a. Expected to have a significantly lower initial SSH with higher k(e.g. k >= 10).

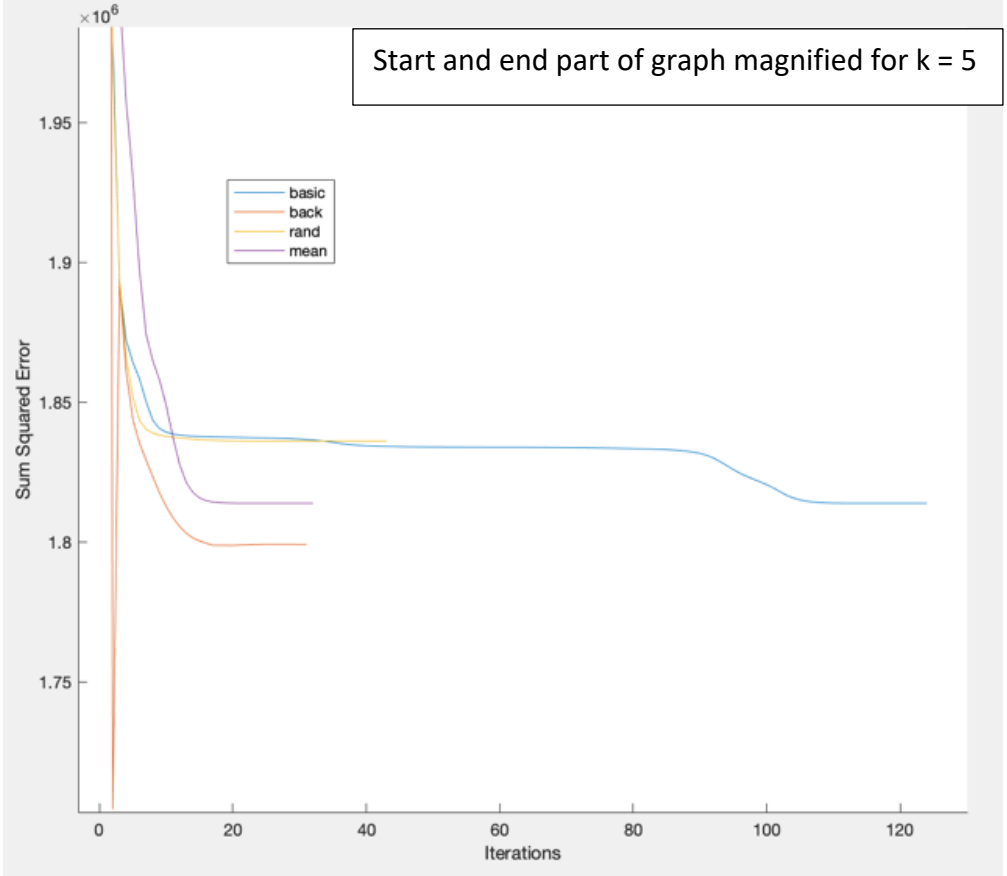


Back had a lucky start but eventually all converges to the same value for **k = 1**
Note that all other methods finish before Back even gets close to the final solution



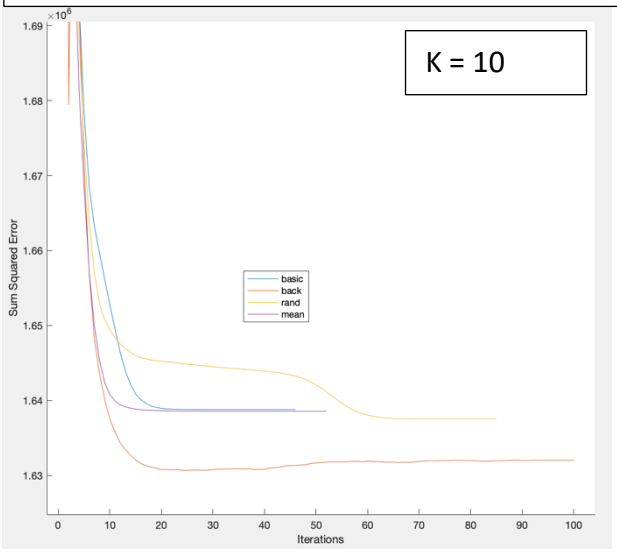
For **k= 3** and **k = 5** (next page)we observe the same phenomena with back, but these times it finishes earlier than the others.
The hypothesis still stands, as besides occasional faster runtime, the accuracy doesn't improve.

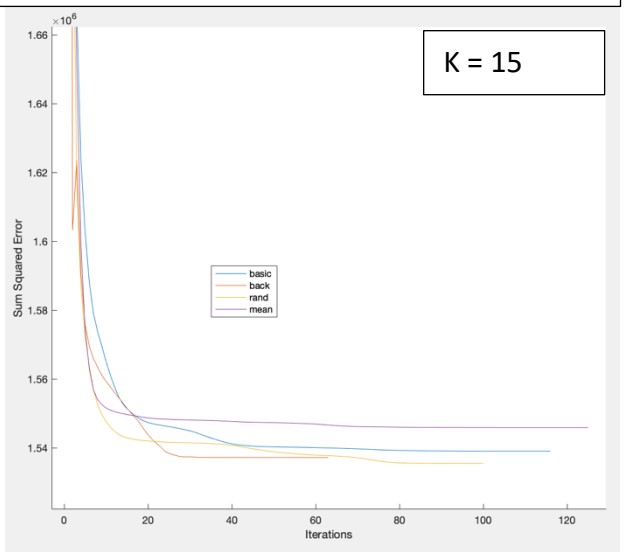




Start and end part of graph magnified for k = 5

The numbers on the Y axis for all 3 graph solidifies our hypothesis. Even the magnified endings fail to show any major differences in the final SSE.





As a conclusion, the evidence suggests that the hypothesis is true in most cases, but no hard proof was presented that would give a range that the difference wont exceed. For majority of purposes the graphs will be enough basis to make an informed decision and make the choice of initial centers a non-priority task. To be able to state more, one must expand the research