**Problem 1 Clustering synthetic data**

(a) Implement the three clustering algorithms. You will reuse these algorithms in the next problem,

so try to make them as general as possible.

(b) Run the algorithms on the three synthetic datasets. Qualitatively, how does each clustering

algorithm perform on each dataset? Comment on the advantages and limitations of each

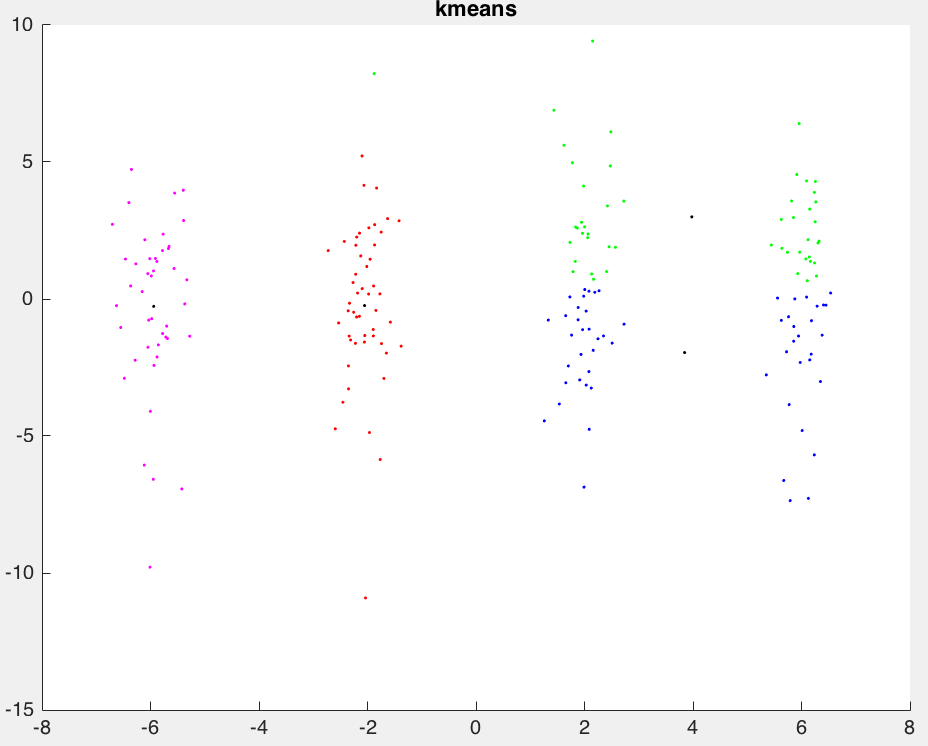
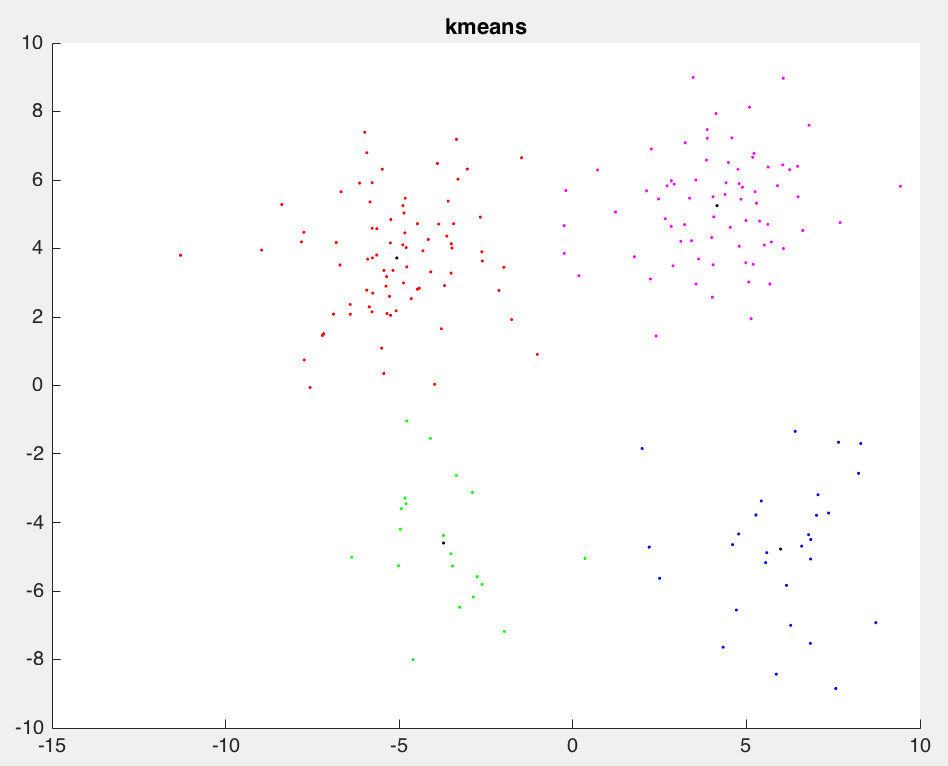
algorithm, in terms of the configuration of the data.

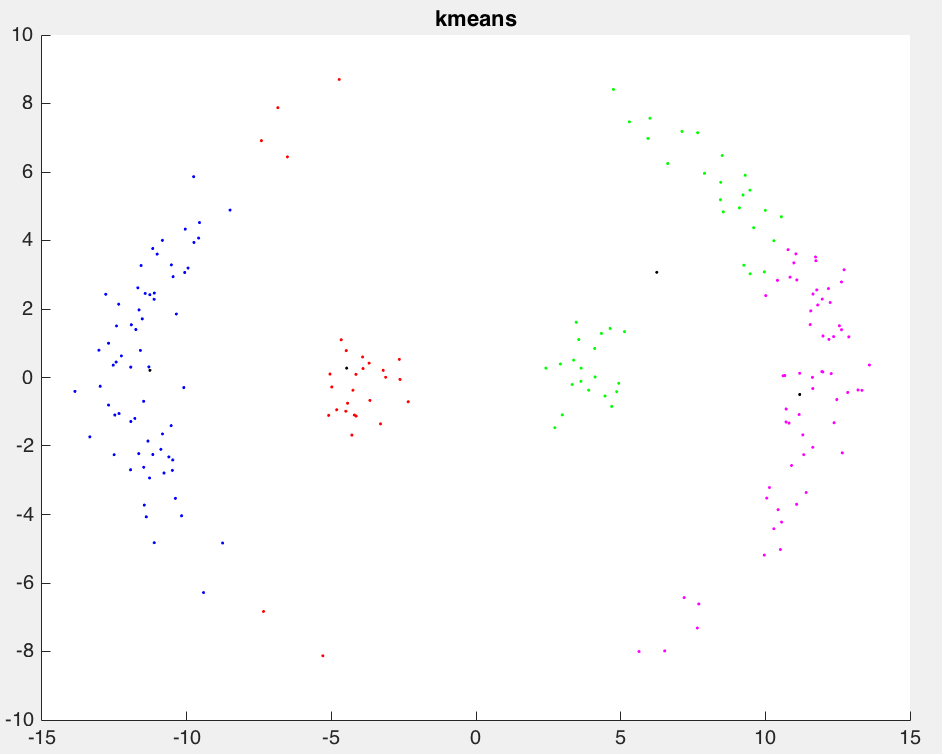
Answer:

K-means calculate conventional Euclidean distance, while EM calculates weighted distance. So EM algorithm would be more accurate and robust, but K-means spends less time on same data set than others. It is hard to initiate k-means’ center at first which will influence cluster result. Mean-shift has best performance above those three, but Mean-shift is sensitive to the bandwidth parameter h.

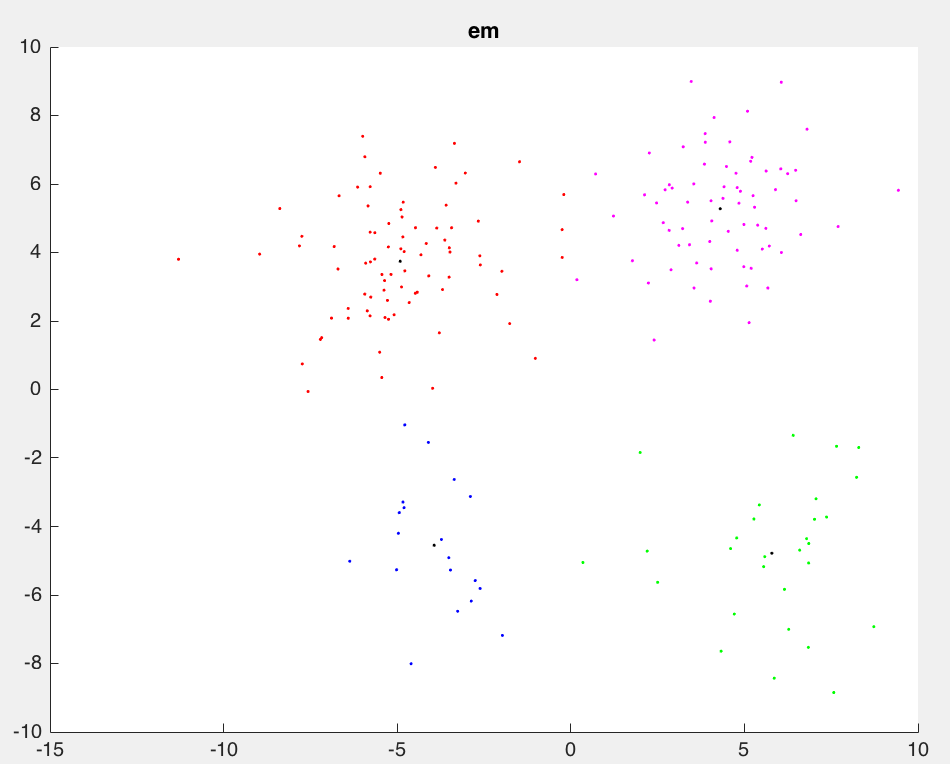
Three algorithm with different dataset (data\_A, data\_B, data\_C):

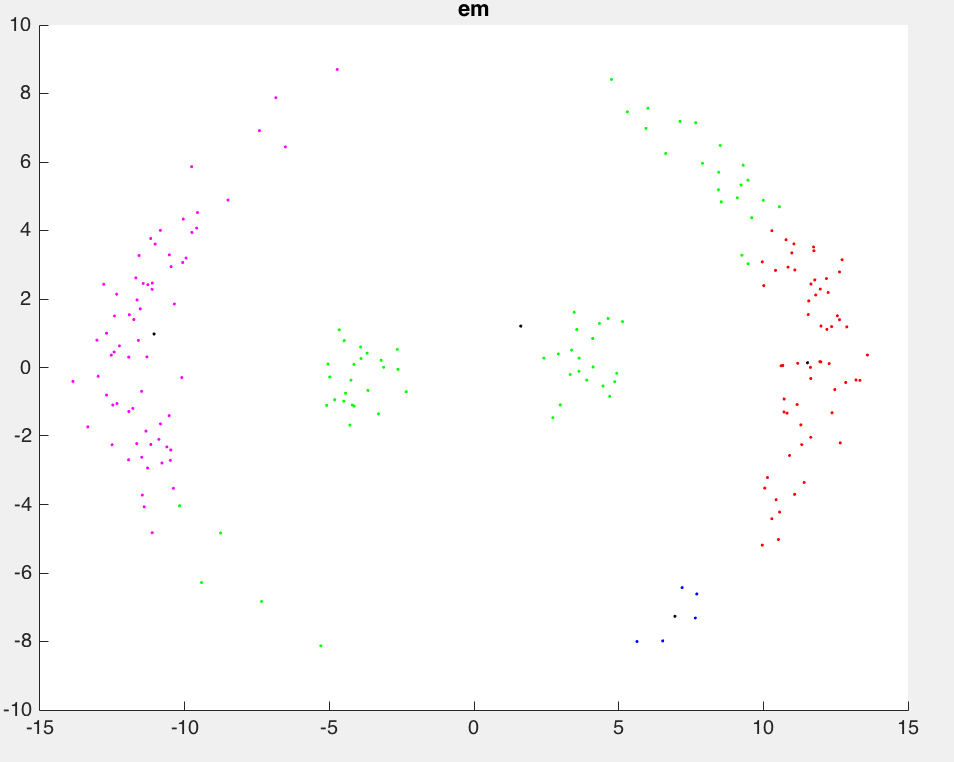
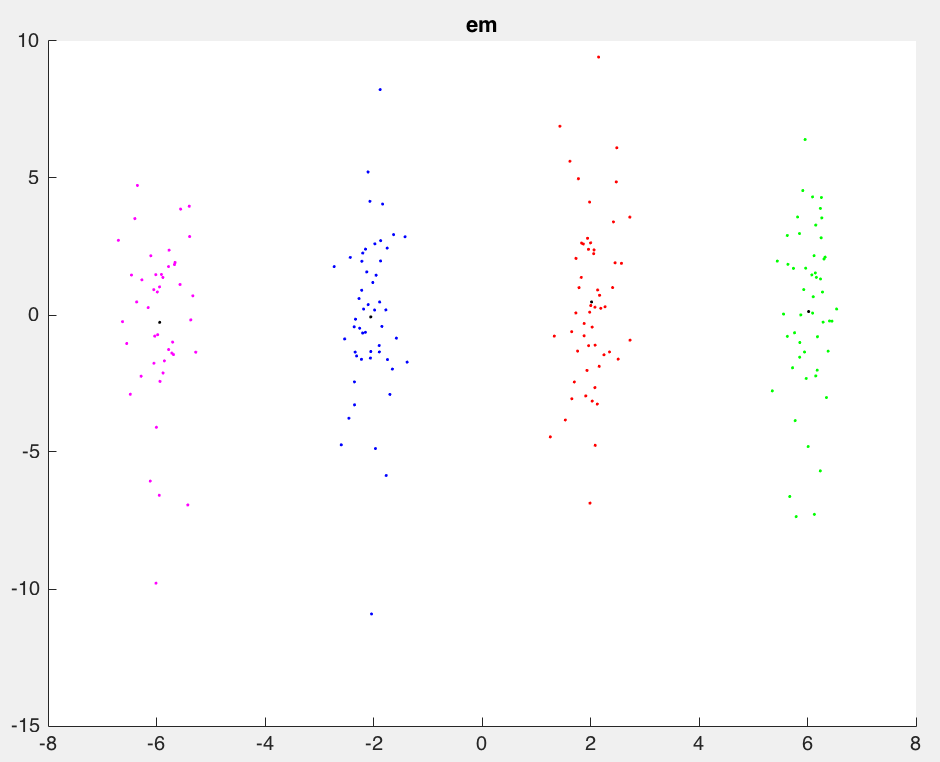
K-means algorithm:



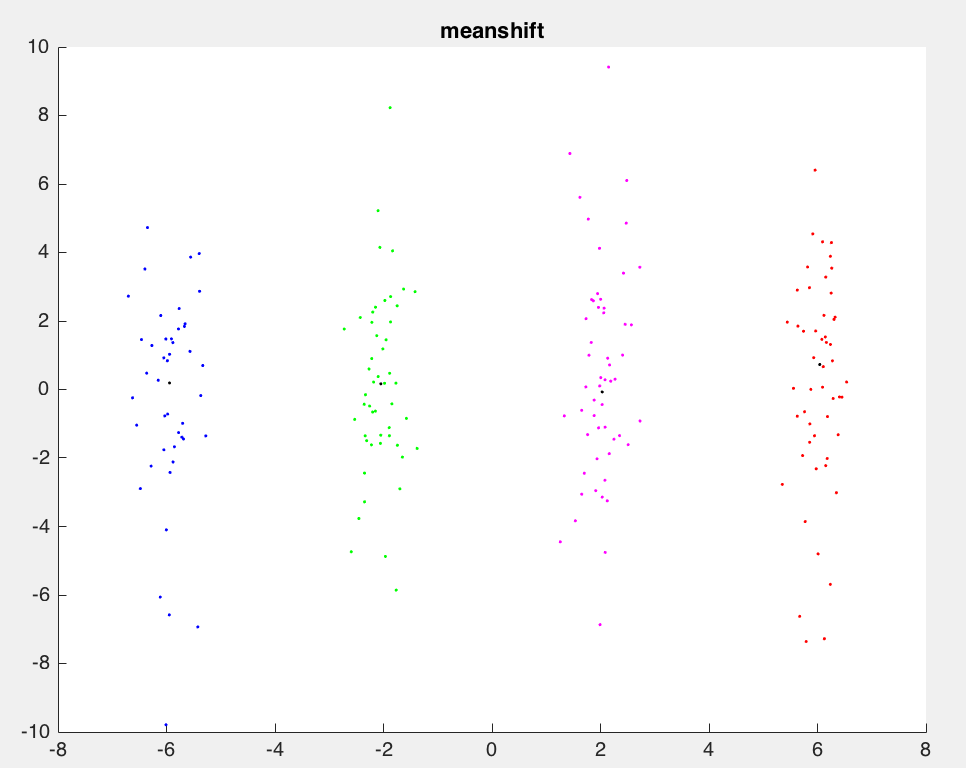
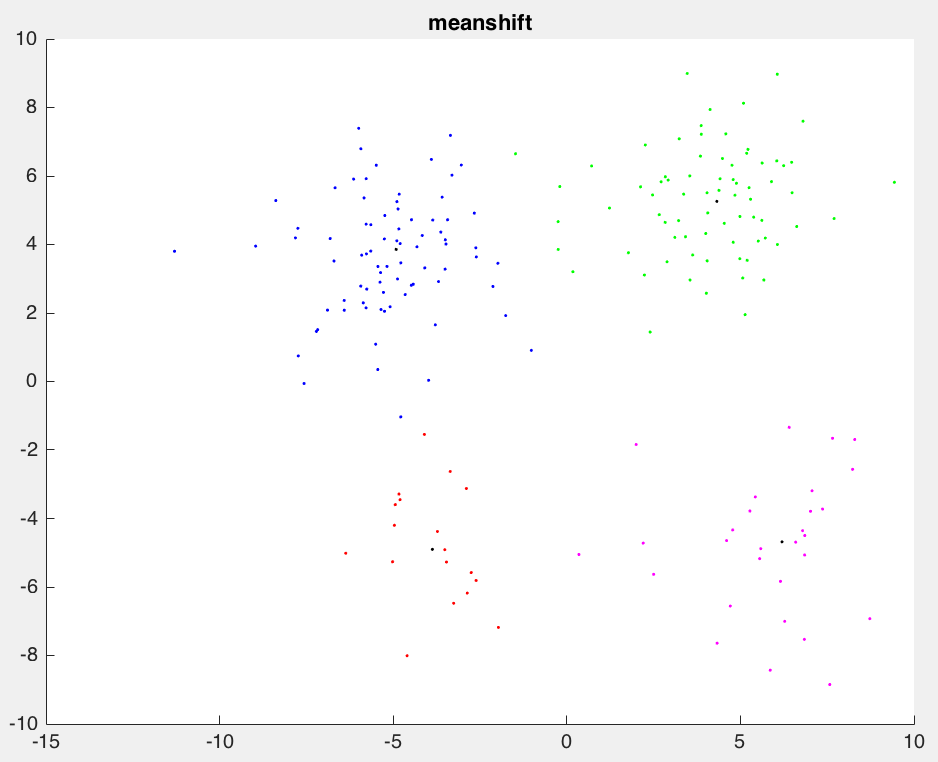


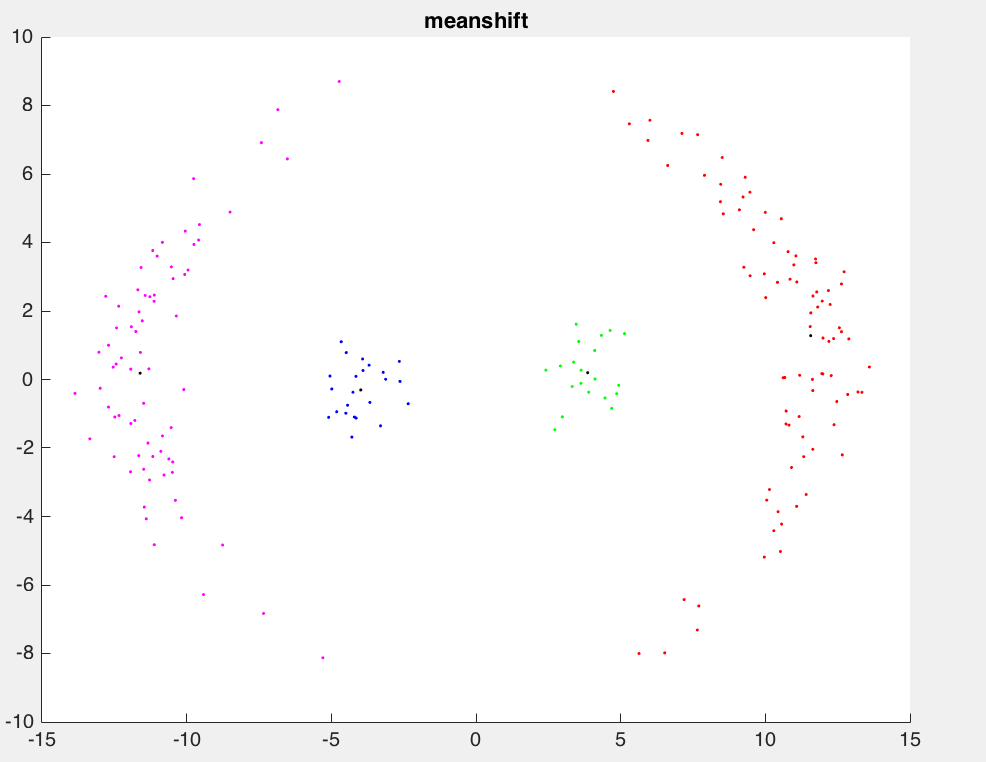
EM algorithm for Gaussian mixture models:





Mean-shift algorithm:





(c) How sensitive is mean-shift to the bandwidth parameter h?

Mean-shift is sensitive to the selection of bandwidth. A small h can slow down the convergence. A large h can speed up convergence but might merge two modes.

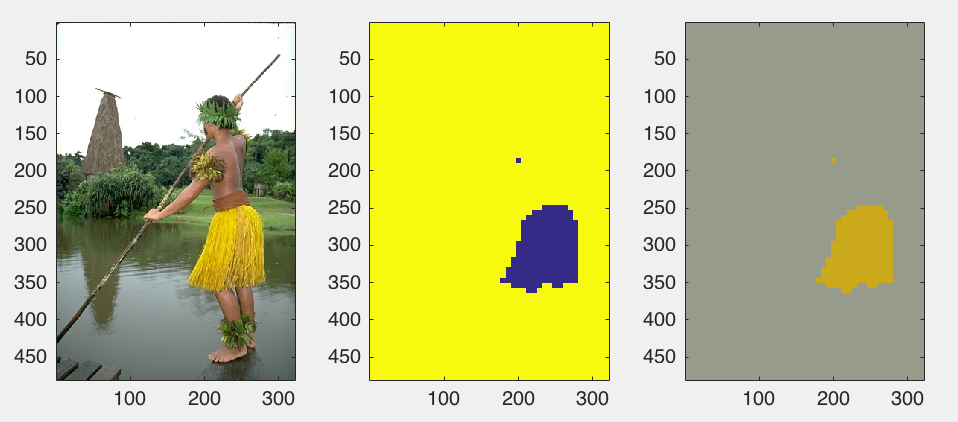
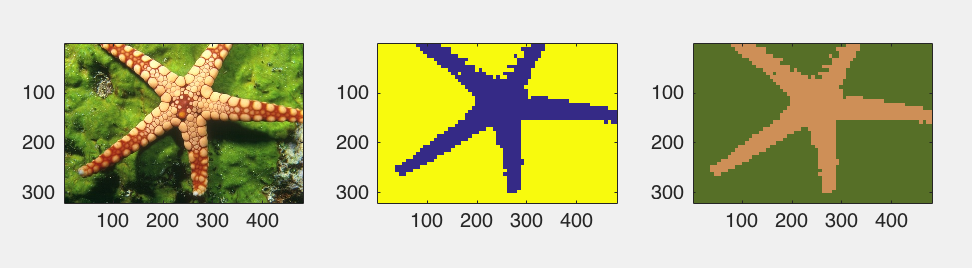
**Problem 2 A real world clustering problem - image segmentation**

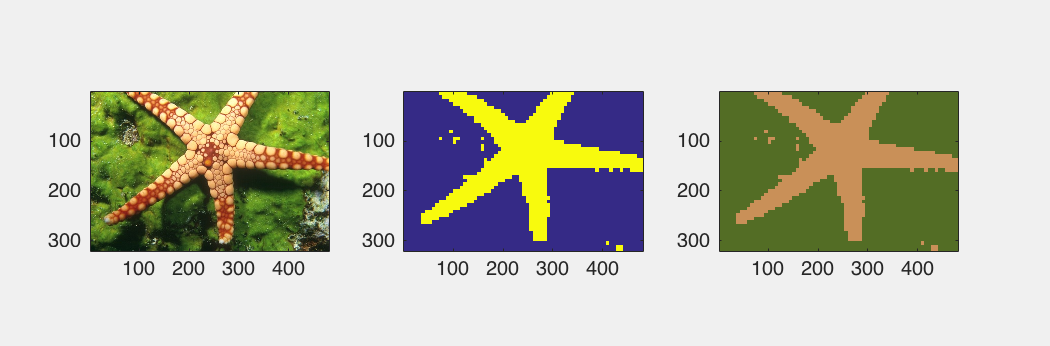
(a) Use the three clustering algorithms to segment a few of the provided images. Qualitatively,

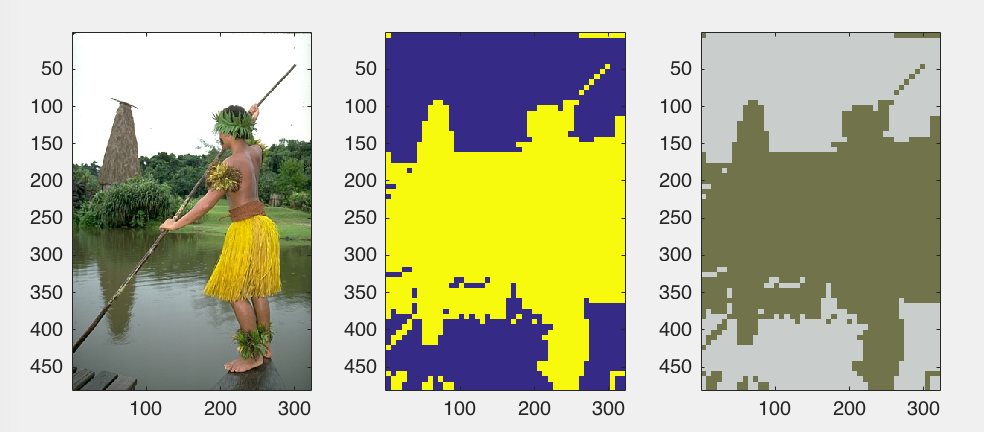
which algorithm gives better results? How do the results change with different K and h?

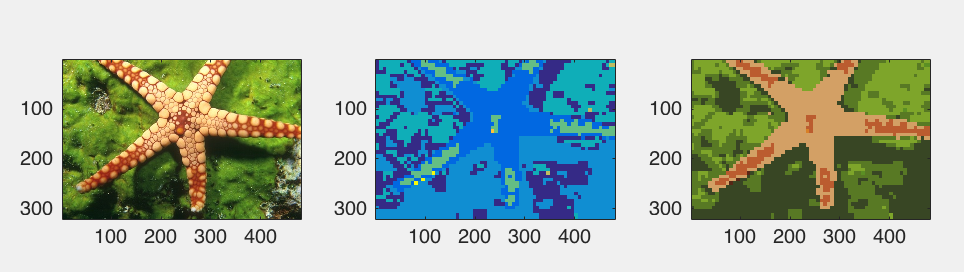
Which is less sensitive to changes in the parameters? Comment on any interesting properties

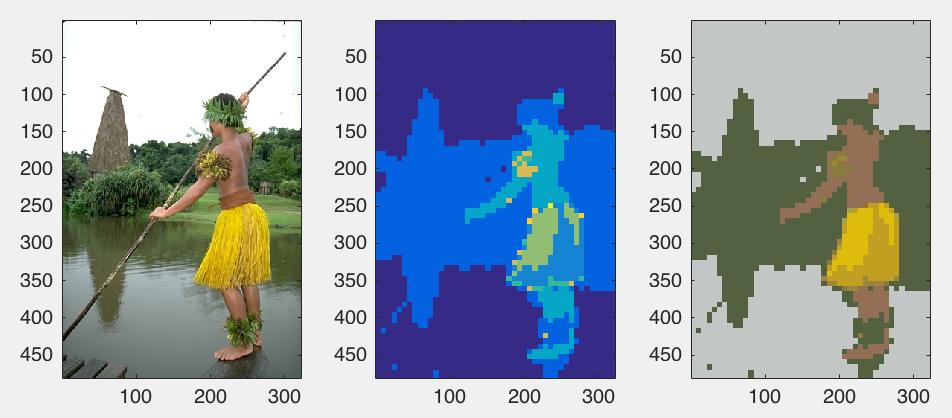
or limitations observed about the clustering algorithms.











Answer:

Mean-shift algorithm gives better results. When K increase, algorithm will get more different cluster labels. On the contrary, when h increase, Mean-shift algorithm will get less cluster label. K-means algorithm is less sensitive to change in parameters. For pictures with

(b) Modify your K-means and mean-shift implementations to allow different feature scaling. Hint:

changing the distance in (7) or kernel in (8) is equivalent to scaling each dimension in the feature

vector x by an appropriate amount. Rerun the segmentation experiment. Do the segmentation

results improve?

Yes, the segmentation results improve.