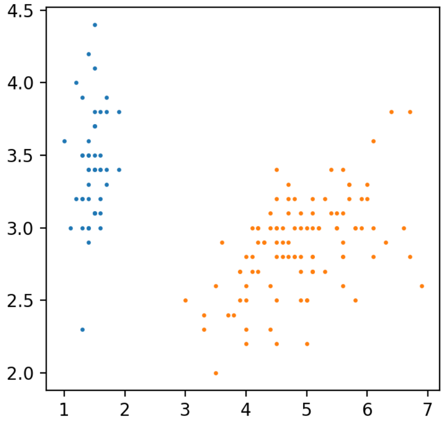
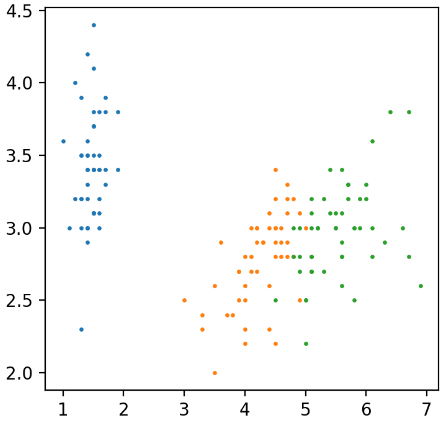
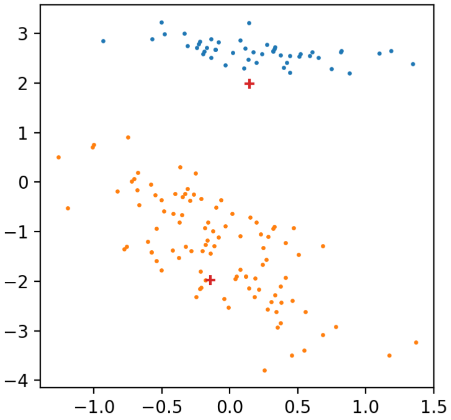
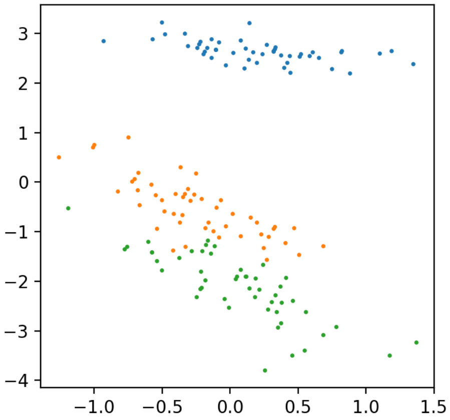
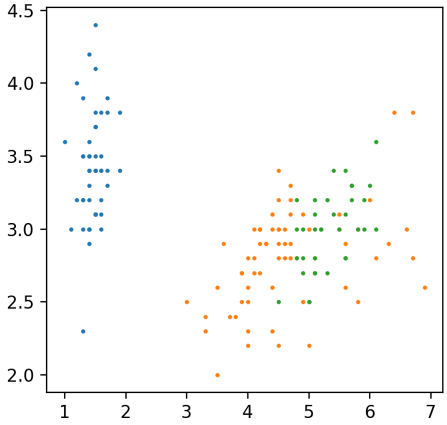
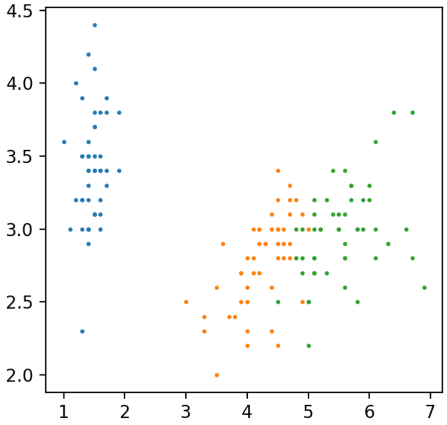
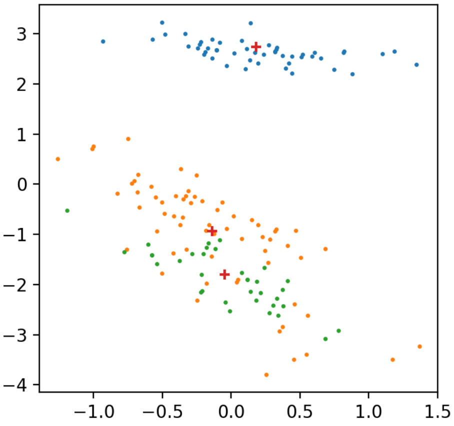
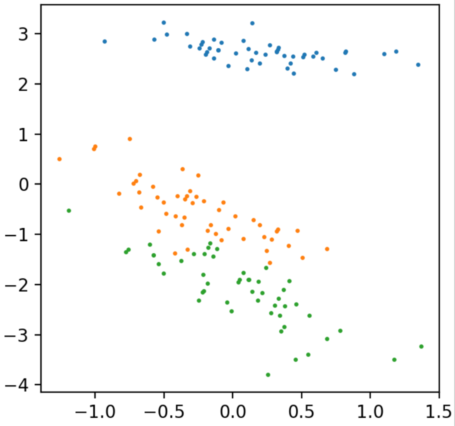
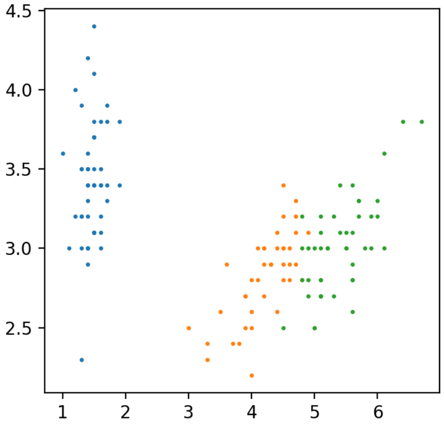
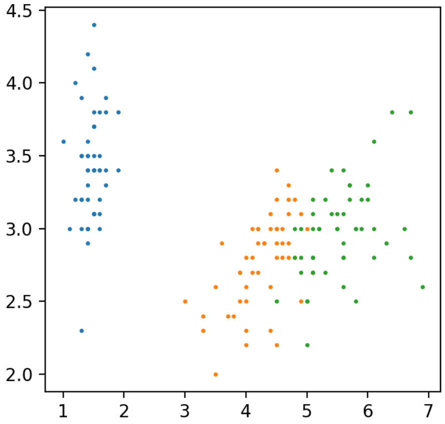
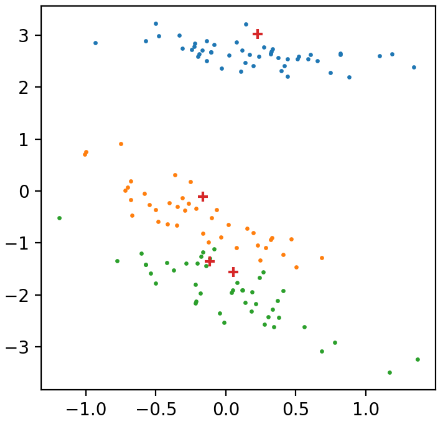
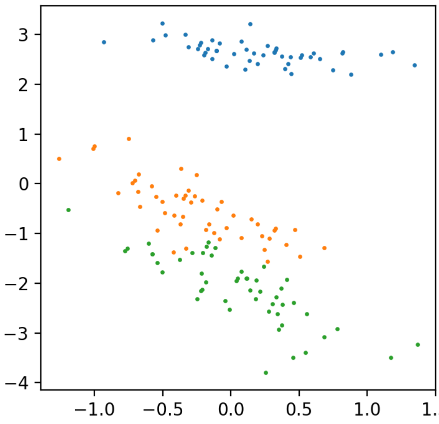
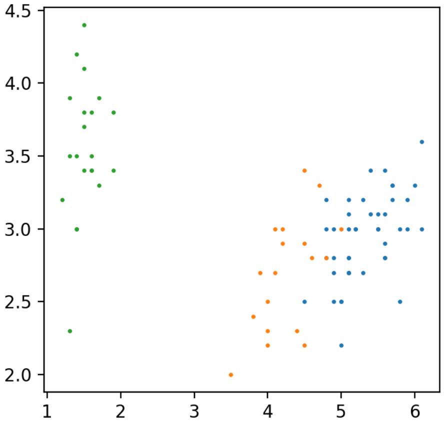
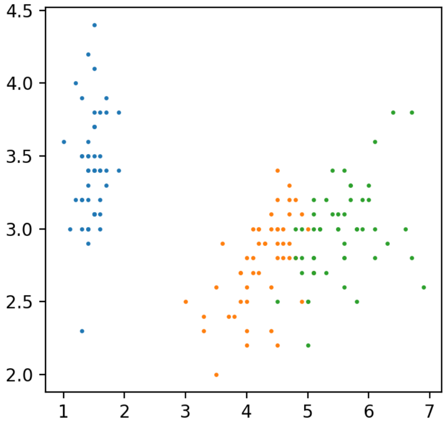
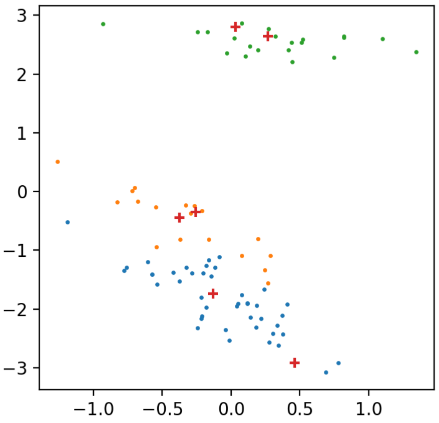
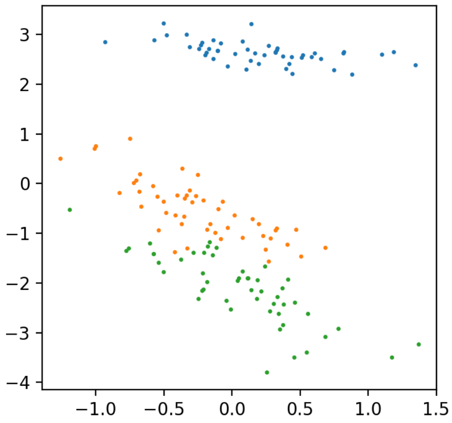
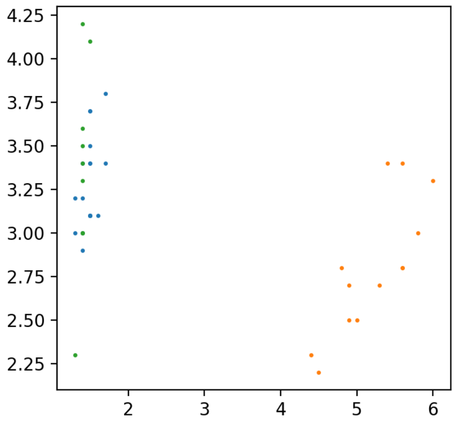
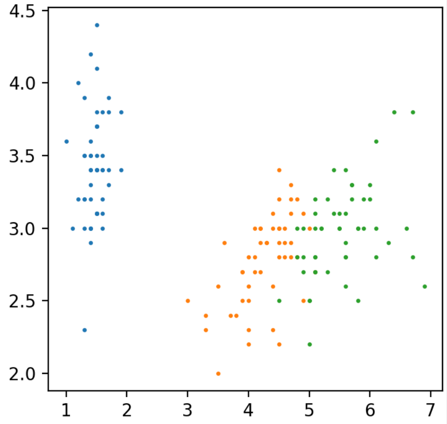
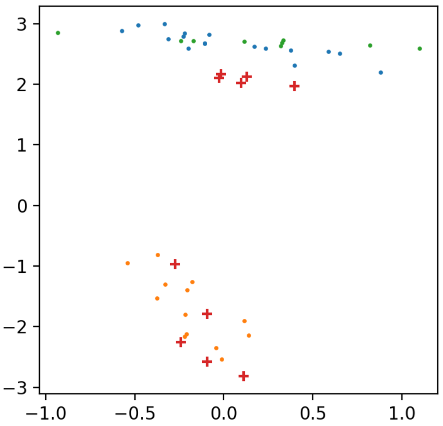
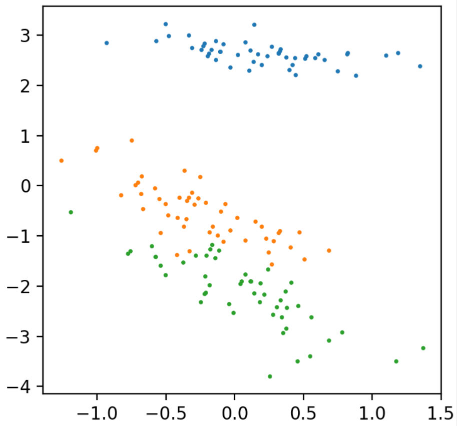
HW6 — Clustering with Gaussian Mixture Model

1. EM Algorithm

The UCI Iris dataset containing features of 150 plants from 3 species was clustered using a fully vectorized EM algorithm with no loops in the implementation. The algorithm only took about one second to complete 4000 iterations, and the final data classification for different numbers of clusters is shown in the graphs below

# of Clusters	Visualization	Predicted Clusters	True Clusters
2	Slice		
	PCA		
3	Slice		

# of Clusters	Visualization	Predicted Clusters	True Clusters
	PCA		
4	**Slice**		
	PCA		
6	**Slice**		

# of Clusters	Visualization	Predicted Clusters	True Clusters
	PCA		
10	**Slice**		
	PCA		

2. Conclusions

The EM algorithm works quite effectively when the number of clusters set is similar to the number of clusters there are in the true dataset. For the UCI dataset with 3 species of plants we can see that the clusterings we obtain from the EM algorithm when we set 3 clusters is relatively accurate. Although not all points are correctly classified in the same clusters, a majority of them seem to be correct. Similarly for 4 clusters the EM algorithm also seems to produce relatively accurate clusterings, although there is an extraneous cluster. If we set the number of clusters to be only 2 then we see that one of the predicted clusters is a conglomeration of two of the closely distributed true clusters. Conversely if we set the number of clusters to be 10 then we can see that a lot of the data is clustered incorrectly due to the extraneous clusters.

Another large factor in results of the EM algorithm is how the clusters are initialized. I have run the algorithm multiple times and each time the clustering slightly vary depending on the initialization of the clusters. This variance is exacerbated when the number of iterations run is quite low, and if the number of iterations is high this variance decreases.