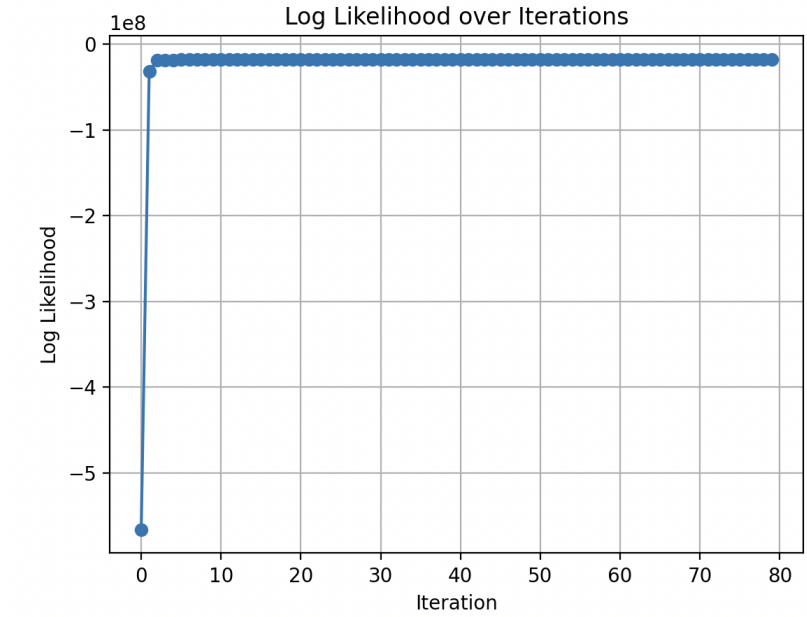
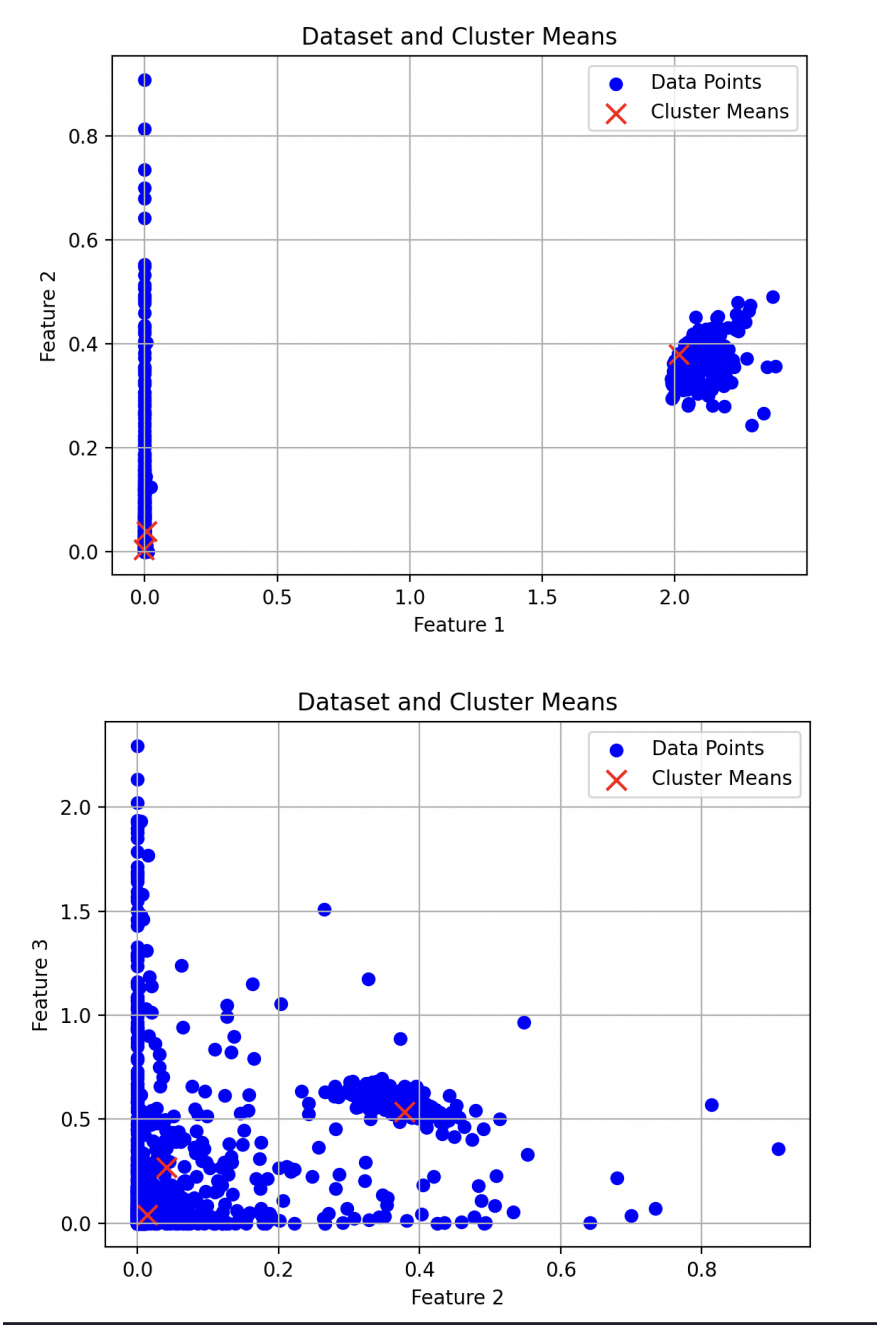
Rachel Springer (rs4127)
November 5, 2023

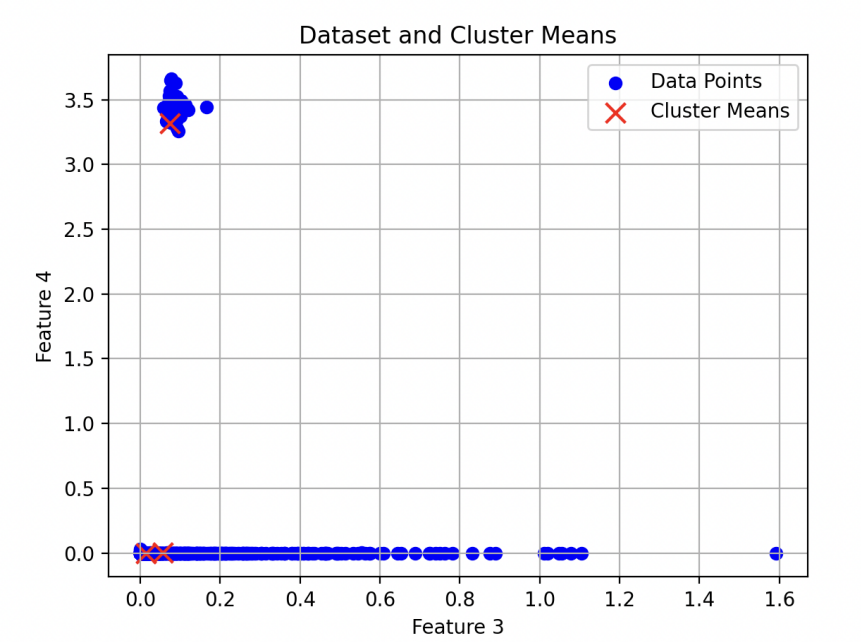
I implemented Gibbs sampling on a Gaussian mixture model. I applied the model to a dataset from Orcasound, which contained 5537 audio samples with human-confirmed orca vocalizations in a training dataset and 828 samples in a test dataset. To extract features from the audio, I parsed the tsv file of annotations to extract relevant segments of the wav files, and fed it through YAMNet (a pretrained audio-classification neural network) to get 1024-dimensional embeddings of each file.

Here we can see the log of the joint over iterations- it converged quite quickly, and visual confirmation on some of dimensions suggests that this number of clusters is logical. I ran the data on the held out log likelihood and found that it improved up to 3 clusters, so I chose to analyze it at K=3.




These means seemed logical from a visual analysis of some of the dimensions as well, though in many dimensions there was a more clear 2-cluster pattern.


Dataset and Cluster Means
Data Points
Cluster Means
Feature 2
Feature 1
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
Dataset and Cluster Means
Data Points
Cluster Means
Feature 3
Feature 2
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8




After running the clustering algorithm, I listened to some of the audio files form each of the clusters. The three seemed to be split into groups of high pitched squeals, more chirpy varied squeals, and whistly noises. (for example, the wav files corresponding to rows 2, 4, and 4024 in the first group; rows 4167, 4954, 5017, in group 2; 48 1776, and 5256 in group three, as found in the dataset at https://github.com/orcasound/orcadata/wiki/Orca-training-data). Under two clusters, the few "chirpier" squeals were combined with the other squeals, which seems audibly logical. It seemed to do this relatively effectively despite different background noise levels in the different recording locations. Scientifically, it might be more interesting to cluster the audio further to try to isolate orca-linguistic patterns within each of the vocalization styles, though this might require a different form of embedding, as YAMNet was originally designed to classify between audio categories with a larger variation in between categories (ex cats vs dogs vs drums). The clustering of the datapoints, also, doesn't always seem best suited to a Gaussian mixture, as we see above. This could be an area for improvement or experimentation in future versions as well.