Python Lab 5 SSY316

Matthew Newson, Erik Norlin

December 2023

Fitting GMMs using the EM algorithm

Activity 1

Gaussian mixture models (GMM) were implemented to find the optimal number of clusters K of the Iris data set considering all four variables *sepal width (cm), petal width (cm), sepal length (cm), petal length (cm).* Ranging K from 1 to 10, evaluating every GMM with the BIC criterion it was found that the optimal K was consistently 2 with the lowest obtained BIC value of 574. The obtained optimal mean, μ_{opt}, and (full) covariance, C_{opt}, were

$$\mu_{opt} \approx \begin{bmatrix} (3.43 & 0.25 & 5.01 & 1.46) \\ (2.87 & 1.68 & 6.26 & 4.91) \end{bmatrix}$$

and

$$C_{opt} \approx \begin{bmatrix} \begin{pmatrix} 0.14 & 0.01 & 0.1 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 \\ 0.1 & 0.01 & 0.12 & 0.02 \\ 0.01 & 0.01 & 0.02 & 0.03 \end{pmatrix} \\ \\ \begin{pmatrix} 0.11 & 0.08 & 0.12 & 0.14 \\ 0.08 & 0.18 & 0.17 & 0.29 \\ 0.12 & 0.17 & 0.43 & 0.45 \\ 0.14 & 0.29 & 0.45 & 0.67 \end{pmatrix} \end{bmatrix}$$

Activity 2

Repeating the same analysis as above but only considering the two variables *sepal width (cm), petal width (cm)* we obtained different optimal K for every run without any consistent pattern. For one of the runs the optimal K was 4 with a BIC value of 158. The obtained optimal mean and covariance (diag) were

$$\mu_{opt} \approx \begin{bmatrix} (3.38 & 0.20) \\ (3.02 & 1.98) \\ (2.69 & 1.30) \\ (3.50 & 0.31) \end{bmatrix}$$

$$C_{opt} \approx \begin{bmatrix} \begin{pmatrix} 0.09 & 0.00 \\ 0.00 & 0.00 \end{pmatrix} \\ \\ \begin{pmatrix} 0.08 & 0.00 \\ 0.00 & 0.09 \end{pmatrix} \\ \\ \begin{pmatrix} 0.08 & 0.00 \\ 0.00 & 0.04 \end{pmatrix} \\ \\ \begin{pmatrix} 0.2 & 0.00 \\ 0.00 & 0.02 \end{pmatrix} \end{bmatrix}$$

We can conclude that we do not get the same result as considering all four variables. It might be that including all four variables gives a more robust estimation of the clustering since we are considering more essential information.

Activity 3

Here we wanted to see if the optimal K was the same as the true number of clusters. Figure 1 shows sampled data from the three spherical multivariate Gaussians

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right), \mathcal{N}\left(\begin{bmatrix} -6 \\ 3 \end{bmatrix}, \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix}\right), \text{ and } \mathcal{N}\left(\begin{bmatrix} 3 \\ -4 \end{bmatrix}, \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix}\right)$$

Fitting GMMs to the data for different numbers of K, we obtained that the optimal K was consistently three and the optimal covariance matrices were spherical every run. This is aligned with the true distributions that we defined.

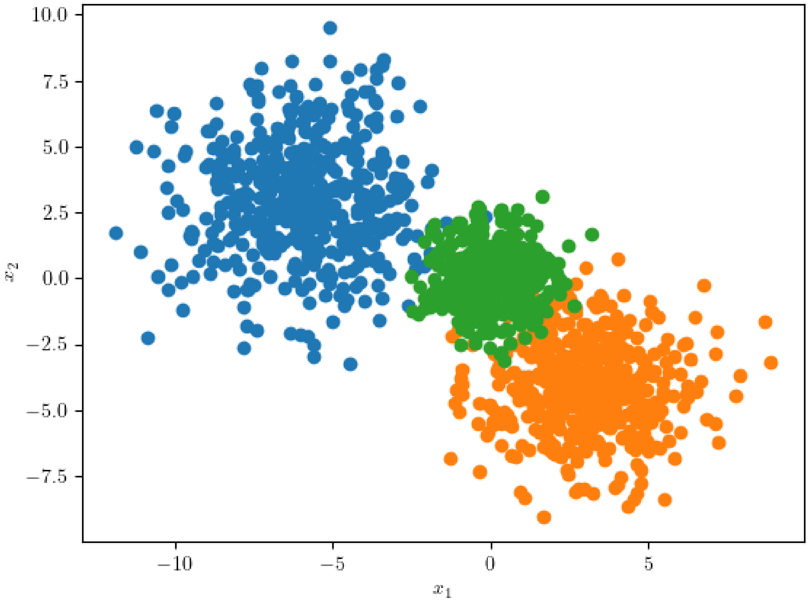

Figure 1: Sampled data from the three different multivariate Gaussian distributions.

Activity 4

Again we wanted to see if the optimal K was the same as the true number of clusters. This time, Figure 2 shows sampled data from the four multivariate Gaussians with full covariances

$$\mathcal{N}\left(\begin{bmatrix}0\\10\end{bmatrix},\begin{bmatrix}10&35\\35&1\end{bmatrix}\right),\mathcal{N}\left(\begin{bmatrix}-6\\3\end{bmatrix},\begin{bmatrix}16&48\\48&32\end{bmatrix}\right),\mathcal{N}\left(\begin{bmatrix}-5\\-4\end{bmatrix},\begin{bmatrix}18&27\\27&9\end{bmatrix}\right),\text{ and }\mathcal{N}\left(\begin{bmatrix}-1\\2\end{bmatrix},\begin{bmatrix}9&18\\18&9\end{bmatrix}\right)$$

Fitting GMMs to the data for different numbers of K, we obtained that the optimal K was consistently four and the optimal covariance matrices were full every run. This is aligned with the true distributions that we defined.

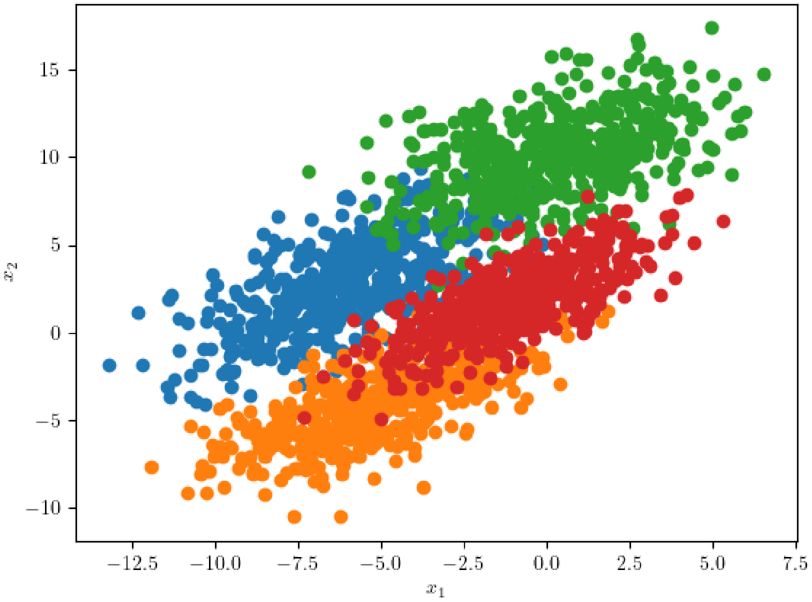

Figure 2: Sampled data from the four different multivariate Gaussian distributions.

Overfitted (Bayesian) Gaussian Mixtures

Activity 5

Here we wanted to see how a Bayesian Gaussian mixture model (BGMM) deal with overfitting compared to a GMM. A GMM and a BGMM were initialized with $K = 10$ and fitted to the Iris data set. We can see in Table 1 that the GMM distributed the data to all ten cluster. In Table 2 we can see that the BGMM distributed the data (on average) to four clusters. This could indicate that BGMM is less prone to overfitting than GMM.

Table 1: Predicted number of individuals by the GMM of the Iris data set.

Cluster	No. individuals
1	7
2	10
3	13
4	5
5	12
6	9
7	25
8	16
9	19
10	34

Table 2: Predicted number of individuals by the BGMM of the Iris data set.

Cluster	No. individuals
1	50
2	72
3	15
4	0
5	15
6	0
7	0
8	0
9	0
10	0

K-Means Clustering and PCA of Human Activity Recognition

Activity 6

Considering the high dimensional Human Activity Recognition data set, Figure 3 shows how the within-cluster sum of squares (WCSS) changes with varying K. We can see that the error is continuously decreasing because the WCSS will tend to zero as K approaches the number of data points. This is because every data point can then be assigned to its own cluster. However, having K be equal to the number of data points is not meaningful. To determine the optimal K we use the *silhouette score*. The clustered data with the largest silhouette score corresponds to the optimal K. In this case, the largest silhouette score corresponded to $K = 2$.

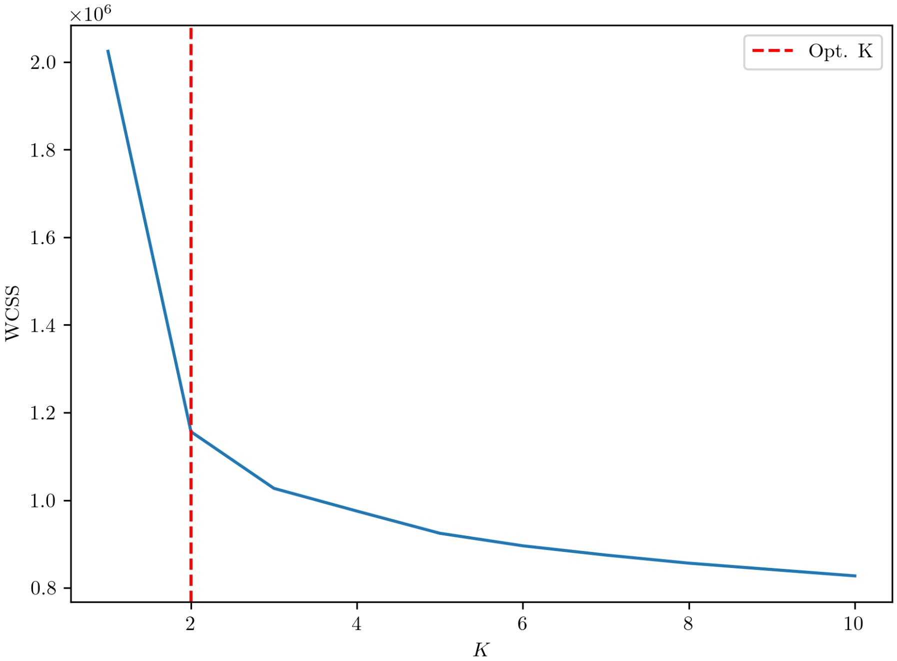

Figure 3: How the WCSS changes with varying number of clusters K. The largest silhouette score corresponds to $K = 2$ which determines the optimal K.

Activity 7

Figure 4 shows the explainable variance ratio and the cumulative sum ratio of λ_i where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$. We can verify in the left subfigure that the first principle direction accounts for 51% of the variance of the data. Following by the second principle direction with a huge drop to 6%, and the third 3% etc. In the right subfigure, we can verify that the first 103 principle directions account for 95% of the variance of the data which means that if we decide to be satisfied with 95% of the explainable variance of the data we can throw away over 80% of all principle directions.

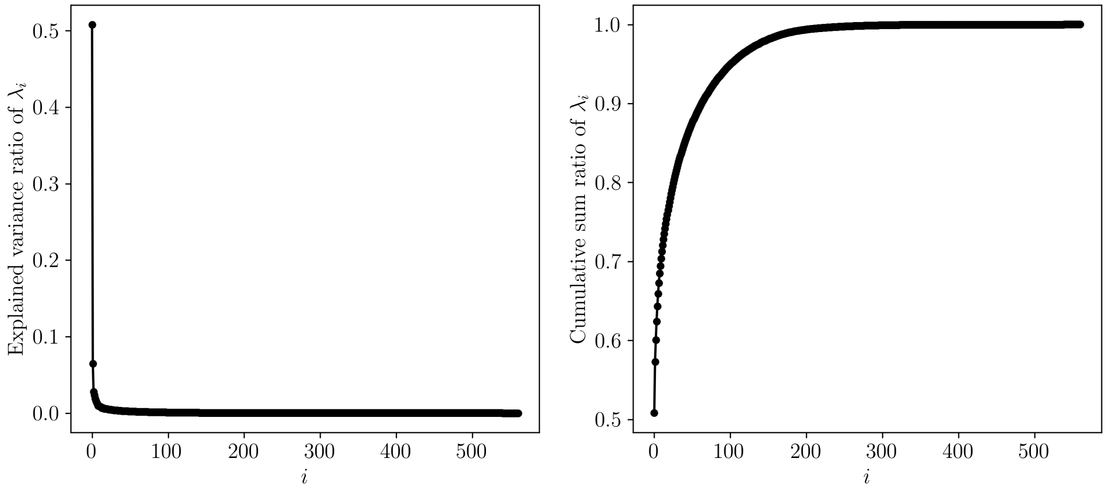

Figure 4: Left shows the explainable variance ratio of λ_i. Right shows the cumulative sum ratio of λ_i.

Activity 8

For illustration purposes we project the high dimensional data down to three dimensions using the first three principle directions accounting for 60% of the explainable variance of the data. Clustering the resulting principal components, Figure 5 shows us how the WCSS change with varying K, and the silhouette score tells us that the optimal K is two here as well.

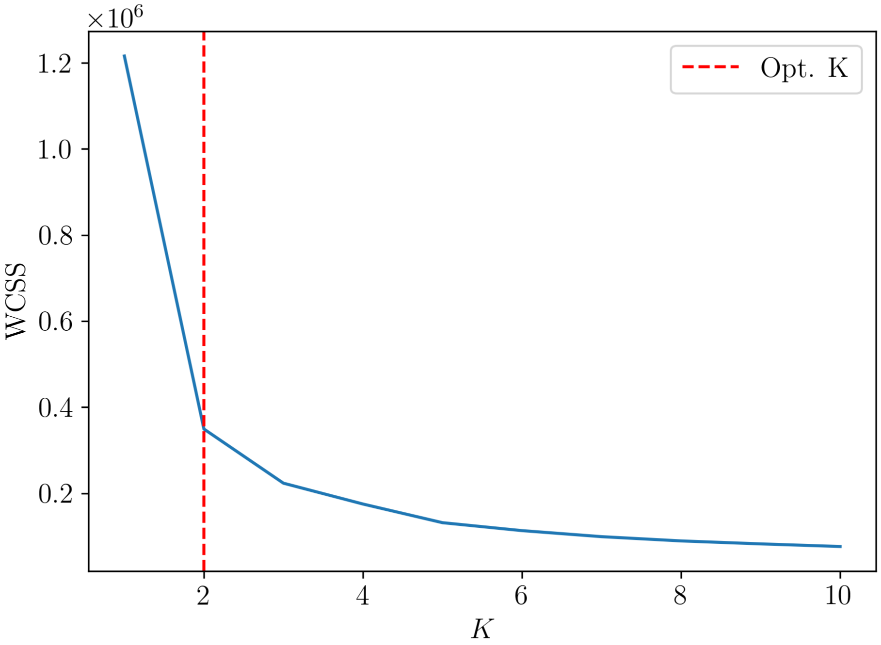

Figure 5: How the WCSS change with varying K after projecting the data to three dimensions. According to the silhouette score, the optimal K is two.

Figure 6 shows the clustering of the principal components in 3D. From inspection, we can see that the principal components are clustered into two groups that makes sense. Because we are visualizing this in 3D, the centers of the clusters are unfortunately not visible because they are completely surrounded by the clustering points. We can conclude that even though we discarded 99% of all principle components we were still able to retrieve a humanly interpretable clustering of the original data. However, it is important to note that since we are projecting the data into a subspace of lower dimension we indeed lose some information on the way.

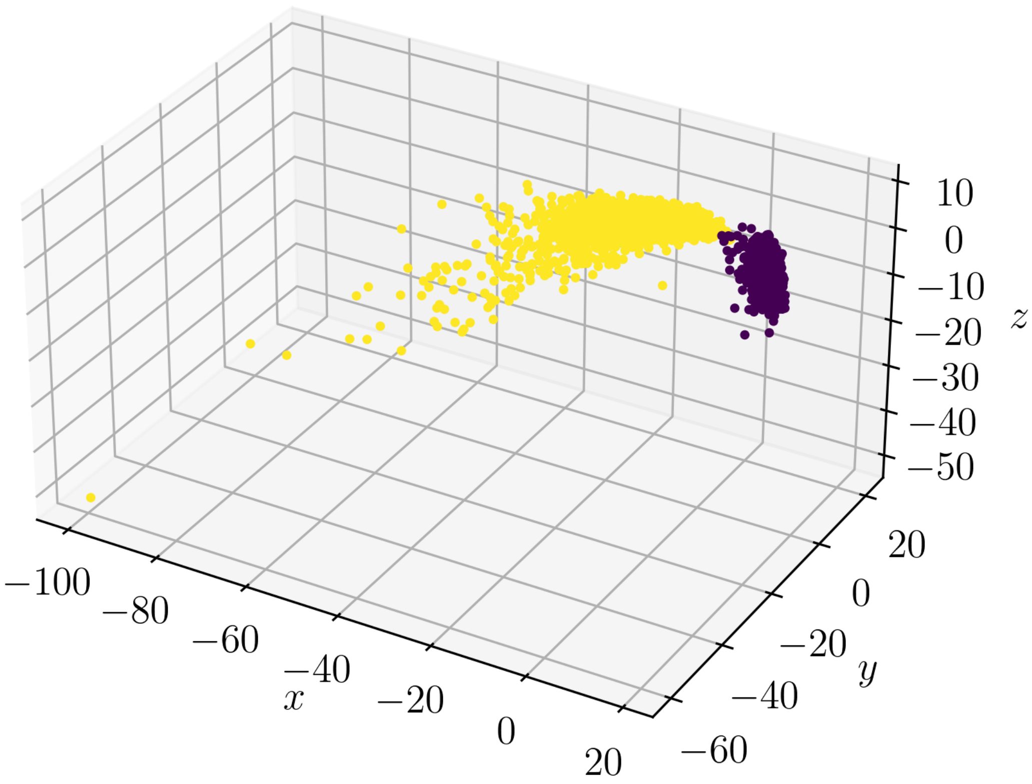

Figure 6: The first three principal components clustered into two groups.