Statistical Learning Project 1

Jonathan Cheung 01063446

February 18, 2020

Analysing Spotify music data with distance based methods

Introduction

https://www.kaggle.com/leonardopena/top-50-spotify-songs-by-each-country/data. This is a dataset containing the top 50 songs on Spotify for 20 countries during Christmas time in 2019. For each song we will use 9 variables: beats per minute, energy, danceability, loudness, liveness, valence, duration, accousticness, and speechiness. We can apply multidimensional scaling to the top 50 songs of a given country to visualise the similarities between these songs. MDS is useful here because music can be described with many variables, and we want to reduce these dimensions and see if there are key variables that can help find patterns or clusters of songs by the country that they are popular in.

World top 50 songs

First we apply MDS with euclidean distance onto the top 50 songs worldwide. From fig.1 we see 3 significantly large eigenvalues. Although not visible, there are some negative eigenvalues of order 10^{-11}, so the non Euclidean error is almost non existent and that the data can be represented with 3 dimensions. A cumulative sum plot (fig.2) confirms this as we can see that 3 eigenvalues represents over 75 percent of the information. An alternative metric would be the Canberra distance since the variables have different scales and Canberra measures relative difference. However, in fig.3 we see more significant negative eigenvalues, so we will stick to Euclidean distance.

Fig.4 is our plot visualising the 50 songs in 2d. The number in the plot corresponds to the songs ranking in the top 50 songs worldwide. The songs are reasonably well spread across the two dimensions and there are no clear clusters. Using isoMDS (fig.5) on this world data does not help much either. The scatter plot looks very similar and stress is reduced from 20.154 to 20.146. Thus, there are no groups of songs that are particularly similar in the top 50. Though this could be due to a lack of variables; a musician could argue that cadences are more important than our variables. If we had this information, it would interesting to see if pop songs, which have similar chord progressions, would cluster together.

Other interesting countries

For the rest of this report we will discuss countries with interesting results, and avoid those with no clear clusters. The top 50 African songs (rows 51-100) are shown in fig.6 There are two clear clusters. The cluster on the right represents the 39-50th ranked songs. If we check what these songs are, we find that they are in fact African genres like afro house and kwaito. These songs also all have 112-125 BPM and low acousticness. In the left cluster, the different genres are distributed fairly evenly. However, songs with large y value are from more upbeat artists like Ariane Grande and The Jackson 5. Whereas the songs with small y value are songs like “Blue Christmas” and “It’s beginning to look a lot like Christmas”.

If we look at songs 6,19,10,21,30, they have 147-202 BPM and large y value. Songs 14,17,31 have 75-93 bpm and small y axis value. The cluster of afro house songs fits between these two ends with bpm around 115. From this we conclude that the Y axis could represents BPM. Looking back to the afro house songs we see they have a longer duration than the remaining songs; they have an average length of 385 compared to 184 for the other 38 songs. So we conclude that the x axis could represent duration of the songs. Although we were hoping for something more interesting, it is not entirely surprising that bpm is a key dimension to the data since a lot of songs today are defined by their beat.

Spain (rows 551-600) has two main clusters of songs separated with vertical space y (fig.7). We can confirm this with k-means clustering (fig.8). This means that these top songs can be grouped into those with a high BPM and those with a low BPM. Spain has a strong dance culture and perhaps songs with an extremely high or low BPM are more enjoyable to dance to. Chile (rows 501-550) displays similar behaviour (fig.9). This could be explained with either the fact that Chile also speaks Spanish or that it has strong dance culture, like most Latin American countries. Lastly, India seems to have all their songs with a similar BPM (fig.10) since their favourite music genre, desipop, always has a similar beat.

Conclusion

With our given dataset, BPM and duration seem to be the most fundamental dimensions. We can not always group songs by genre with these two variables, but some genres like afro house, desipop, and latin pop seem to be more consistent with these variables. Worldwide the variety of music is big, but a music artist should pay attention to the country that they are releasing their music in since some countries have more specific tastes which can be identified with the variables in our dataset.

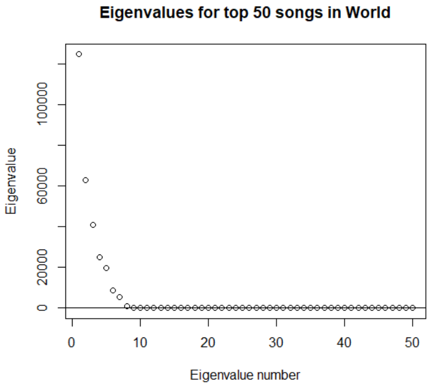

Figure 1: Eigenvalue plot using Euclidean distance

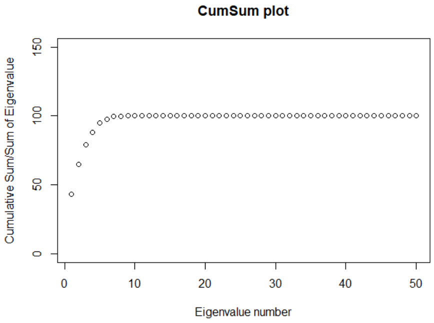

Figure 2: CumSum plot for Euclidean distance

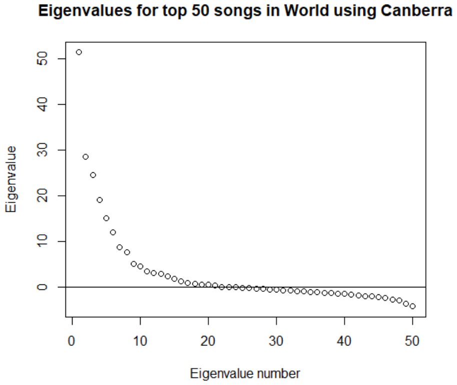

Figure 3: Eigenvalue plot using Canberra distance

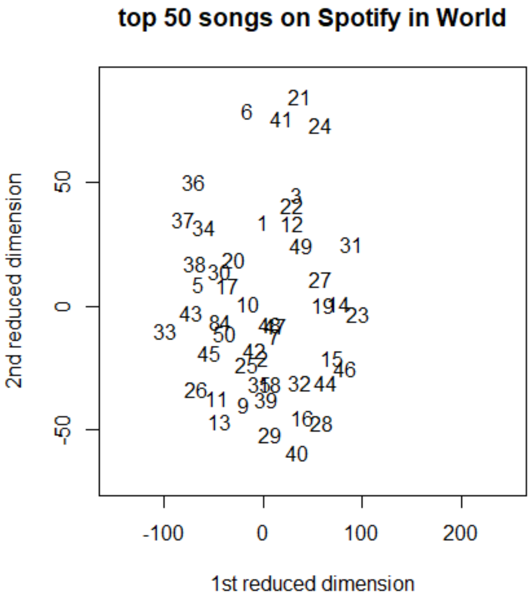

Figure 4: MDS scatter plot

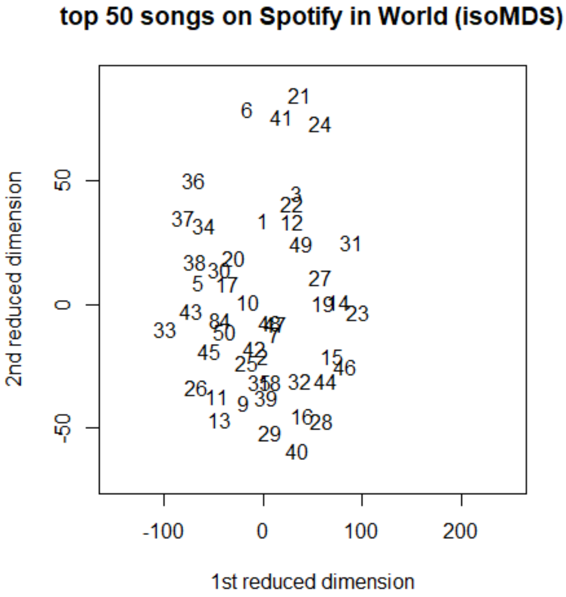

Figure 5: isoMDS scatter plot

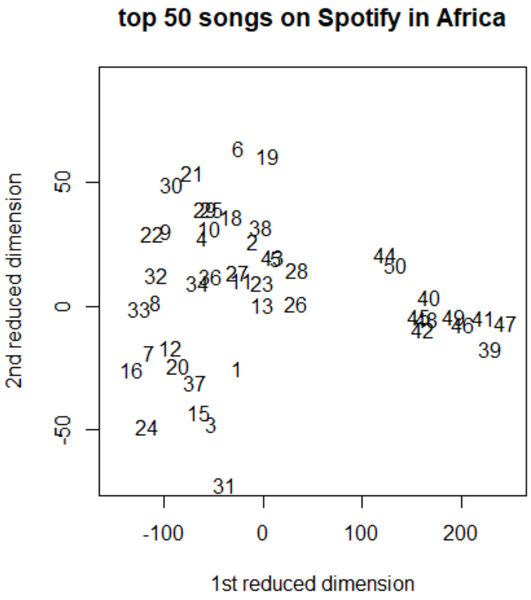

Figure 6: MDS scatter plot

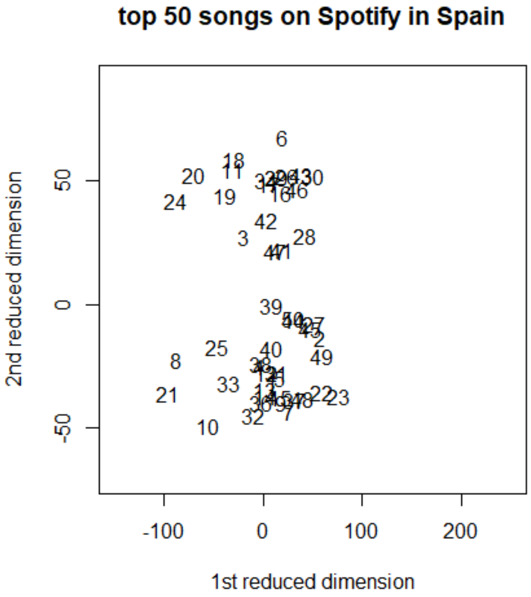

Figure 7: MDS scatter plot

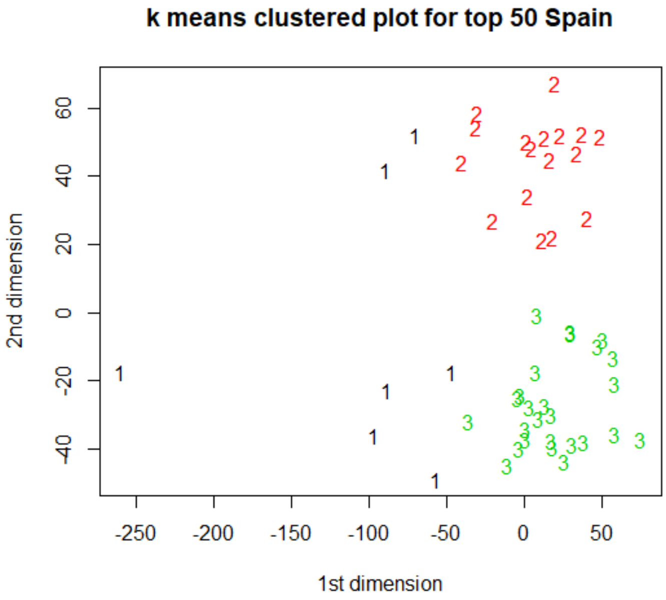

Figure 8: k means clustering

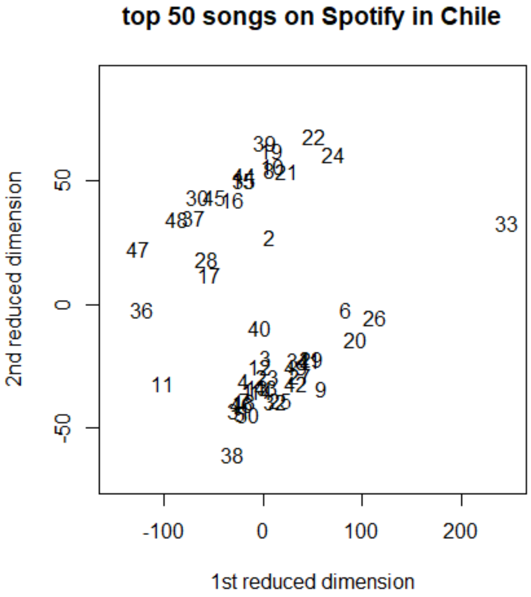

Figure 9: MDS scatter plot

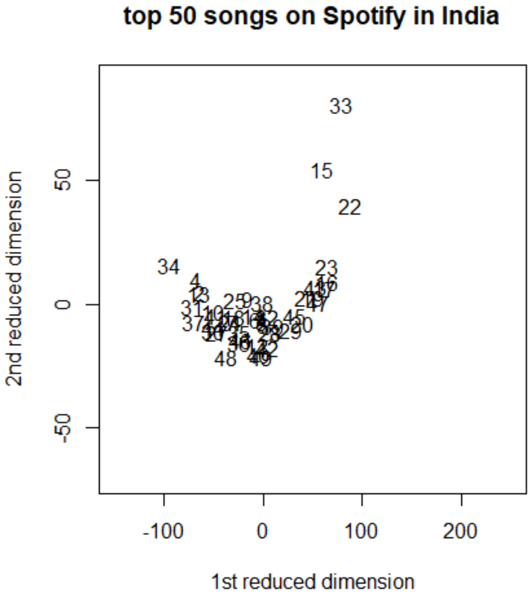

Figure 10: MDS scatter plot