MS-E2112
Multivariate statistical analysis

Project assignment
Eeli Friman
903628

1 Introduction

1.1 Problem formulation

Many famous bands have their style of music change during the years. Usually this happens due to new trends in the music industry, or simply due the artist wanting to create or try something new. Some music creators create the similar type of music for decades with a very noticeable and distinct tone and style.

One of my personal favourite bands is a British band called Coldplay. I've listened to the band for well over a decade, and I will finally see them live next summer. My personal interest towards the band sparked the idea for this project, which will take a closer look to their albums from the past 25 years.

The aim of this project is to try to determine the similarity, or difference, between the albums of Coldplay. The songs are analyzed through uni- and bivariate measures, as well as a multivariate analysis method: Fisher's linear discriminant. The methods were chosen to be suitable in achieving the goal.

The songs from their respective albums are characterized by different musical and auditive properties. The attributes are obtained from Spotify API, which is a diagnostic tool by a music streaming company, Spotify. The dataset collected from the API data is obtained from Kaggle, and is available from https://www.kaggle.com/datasets/faizalkarim/coldplay-albums-and-live-shows?resource=download. [1]

1.2 Univariate analysis

The original dataset consists of 16 columns and 233 rows. The 16 variables of the original dataset are introduced in table 1.

Table 1. The variables and their descriptions from the original data set [1].

Variable	Description
Name	name of the song
Duration	length of the song in seconds
Release date	album release date
Album name	name of the album
Explicit	indicates whether the song contains explicit lyrics, boolean value
Popularity	popularity of the song, measured, value between 0 and 100
Acousticness	confidence measure about song's acousticity, value between 0 and 1.0
Danceability	the song's suitability for dancing, value between 0 and 1.0
Energy	measure of intensity and activity of the song, value between 0 and 1.0
Instrumentalness	confidence that song is considered instrumental, value between 0 and 1.0
Liveness	probability of an audience being present, value between 0 and 1.0
Loudness	loudness of the track, measured -60 to 0 decibels (dB)
Speechiness	presence of spoken words, probability value between 0 and 1.0
Tempo	tempo of the track, pace, estimated in beats per minute (BPM)
Time signature	indicates the amount of beats in a bar, value from 3/4 to 7/4
Valence	measure of musical positiveness, value between 0 and 1.0

From the original dataset, variables "name", "release date", "liveness", "explicit", and "time signature" were removed. The variables "name" and "release date" were removed, since they strictly associated with a certain album. The rows of the data set corresponding to live albums were removed, since live collection albums are usually fused from numerous studio albums. Hence, the variable "liveness", which for non-live tracks is irrelevant, was also removed. The variable "explicit" was removed, since the measures used in uni- and bivariate analyses aren't sensible with boolean values. "Time signature" was removed due to how it is measured by Spotify API. Song's with a duration less than 2 minutes were removed from the final data set, since they most often correspond to intros or outros of the album, which can't be considered proper songs.

Intros and outros cause outliers on variables such as “speechiness”

From the 9 studio albums, only the most popular 6 were chosen for the final data set. Some of the studio albums were removed, since visualizations and categorizing of 9 variables gets too cluttered. The final albums are across the whole timeline of Coldplay’s activity, and are in releasing order: Parachutes, X&Y, Viva La Vida or Death to All His Friends, Mylo Xyloto, A Head Full of Dreams and Music of the Spheres. The final data set contained a total of 65 tracks.

The variable “album name” can be considered a label of this project, since the other variables are used to predict and group the albums, which are classified by their name. The rest of the variables can be treated as the features. The prediction using the multivariate analysis method is further examined in section 2. The feature variables and their mean, median and standard deviation values are represented in table 2. The means of the variables of each album are plotted in figure 1.

Table 2. Statistical measures of the 10 variables of the final data set.

Variable	Mean	Standard deviation	Median
Duration	267	76.5	256
Popularity	59.2	18.3	61.0
Acousticness	0.188	0.273	0.0267
Danceability	0.423	0.122	0.432
Energy	0.614	0.200	0.661
Instrumentalness	0.120	0.263	0.00449
Loudness	-7.89	2.14	-7.23
Speechiness	0.0374	0.00932	0.035
Tempo	126	24.3	131
Valence	0.242	0.132	0.227

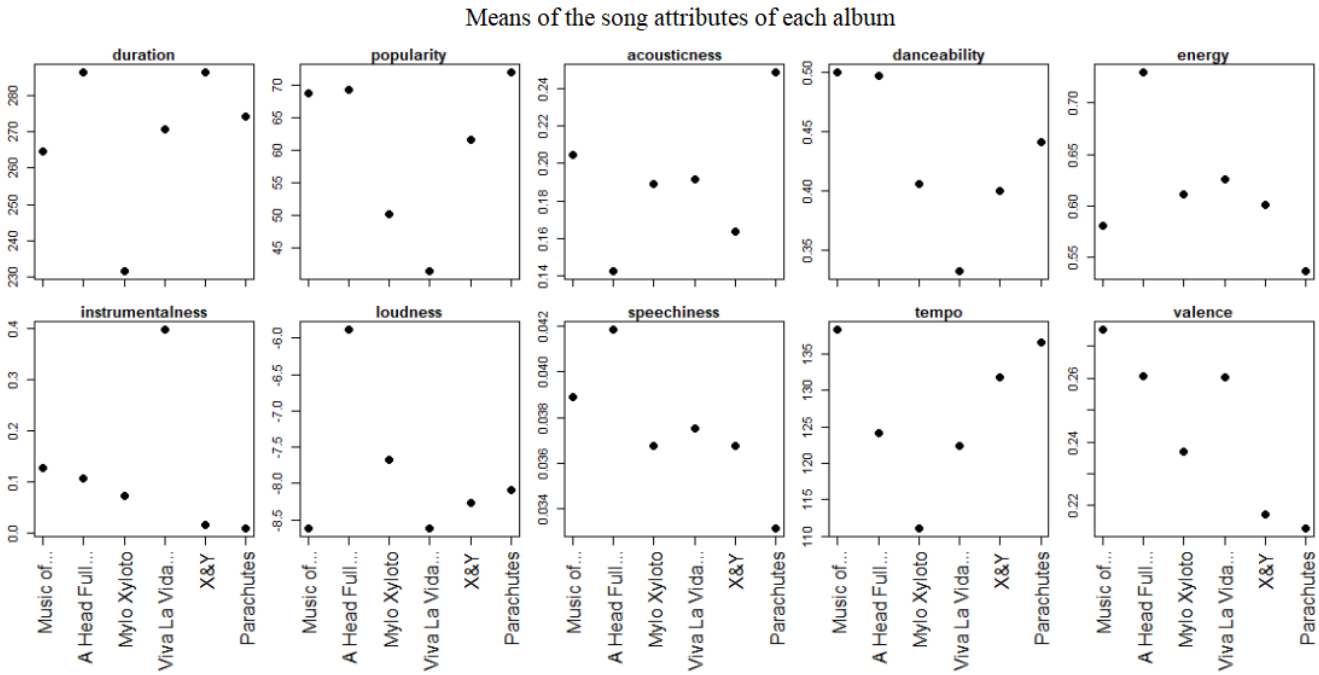

Figure 1. Mean values of the song attributes, grouped by albums.

From the plots in figure 1, we can see some differences and similarities across the albums. Duration-wise, the album Mylo Xyloto has average song length of almost 40 seconds less than any other album. From a

popularity standpoint, Viva La Vida and Mylo Xyloto performed significantly worse to the other albums. From the instrumentalness and danceability plots, it can be observed that Viva La Vida has the most instrumental aspects, and the lowest danceability, which could be potential factors in the lower popularity. The album Parachutes, with the highest popularity, is the most acoustic album of them all, with a notably high tempo.

1.3 Bivariate analysis

The bivariate analysis was conducted to study dependencies between the pairs of variables. Figure 2 represents a scatterplot of the 10 features represented in table 2. The lower diagonal represents the pairwise scatterplot of two variables. The upper diagonal represents the Pearson's correlation value between the pair of variables, which measures linear correlation between them.

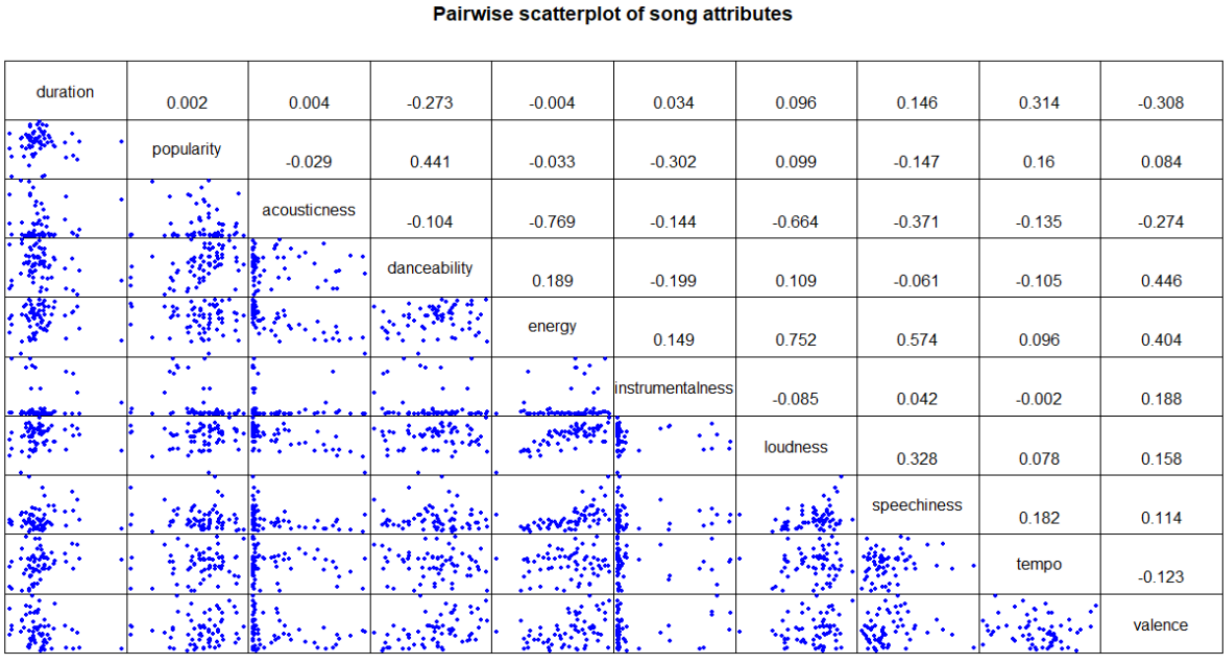

Figure 2. Pairwise scatterplots and Pearson's correlations of the variables. The axes are removed to keep the plot clear. In the lower diagonal, the column variable is plotted on the x-axis, and the row variable is plotted on the y-axis

Intuitively, variable acousticness has a high negative correlation with energy and loudness. Similar observation can be made with valence and energy, which both relate to the positivity. The most interesting pairs of the plot are pairs that contain "popularity" as the second variable. What song attributes affect the song's popularity positively or negatively? It is interesting to see whether a song's properties such as duration, energy, or tempo relate to the success of the track.

The highest correlation with popularity is danceability, 0.441, and lowest with instrumentalness, -0.302. A careful conclusion can be drawn, that the songs with high danceability have a higher rate of success. Correlation value of 0.002 between duration and popularity indicates that the song's length doesn't effect the success. However, from figure 1 in section 1.2, it can be observed that the average duration remains almost constant across the albums, where as popularity has more cross-album variance. Therefore, the effect and correlation of a song's duration to popularity could be simply nullified by the near-constant song duration across all of the albums. Therefore, it can't be confirmed that very long or very short tracks have as high success rate as song's with average duration.

2 Multivariate analysis

2.1 Method

The multivariate analysis method applied in this project was Fisher's linear discriminant analysis. The Fisher's method was chosen, since it can be applied to find a rule, or a way to separate classes out of a data set of measurements. The method can be used to further predict and group new measurements into one of the grouped classes, and works in dimension reduction. The method is therefore well suitable in achieving and studying the goal of this project, which is to group the albums according to the song attributes. The choice was also made due to personal interests, as the method appeared intriguing.

The Fisher's discriminant is a linear function, that classifies groups with a linear separator. The separator is constructed by maximizing the ratio of between groups dispersion, and within groups dispersions. The separator is constructed as a linear combination of variables, that provide the largest separation between groups. A new observation can then be classified to one of the groups by comparing it with the groups made by the linear function.

To study the effectiveness of the linear separator function, the missclassification rate can be calculated. Missclassification rate is defined as the rate of wrong classifications of groups out of total classes. In the context of the project, the missclassification rate is the wrongly predicted album of the song divided by the amount of tracks. The rate can be used to determine the efficiency of Fisher's method, by observing the method's prediction accuracy from the rate. The smaller the missclassification rate, the better the linear function classifies the groups according to the measurements.

The missclassification rate was calculated using the leave-one-out cross validation (LOOCV). In LOOCV, every element of the sample is left out as the test set, and the rest of the sample is used to train the linear function. The test set is finally classified using the Fisher's discriminant. The procedure is repeated for every element of the sample, and the missclassification rate is obtained by dividing the wrong classifications by the total number of elements in the data set. The total number of elements correspond to the true group sizes.

2.2 Implementation

All of the implementation of the analysis was conducted in R. After pruning of the data set with respect to section 1.2, the Fisher's linear discriminant analysis was applied to the data set. The method was applied with the function lda from package MASS.

When applying the lda function on a data set, a linear discriminator is constructed as a linear combination of the feature variables. The lda function constructs the separator for groups to have the highest distance between them. The function can be further applied on new measurements to classify them into one of the separated groups.

The function lda was also used to conduct the LOOCV to obtain the missclassification rate. The LOOCV can be directly performed using the lda function. When applying cross validation with lda, the function automatically classifies, groups and performs LOOCV on every row of the data set. The missclassification rate can be calculated from the results, by comparing the predicted album to the real album of the song. The missclassification rate was also calculated by predicting the group of the test set consisting of the left out element. The prediction was done in a for-loop with an lda model trained with the training set.

2.3 Results

The true and predicted albums of the songs using the lda function are presented in table 3. The values under "True predictions" represent the songs that were correctly classified. Values under "False predictions" represent the songs that were falsely classified to wrong album by the linear discriminant function.

Table 3. Total songs and True/False predictions by the model

Album	Total songs/truth	True predictions	False predictions
Parachutes	9	0	9
X&Y	13	3	10
Viva La Vida or Death to All His	11	5	6
Mylo Xyloto	14	2	12
A Head Full of Dreams	9	4	5
Music of the Spheres	9	1	8

The obtained missclassification rate was 0.77, which implies that only around every 5th song was classfied to the correct album. Figure 3 represents a scatterplot similar to figure 2. In figure 3, all of the albums visualized with different colors and plot characters, to visualize albums and detect the albums through their differences.

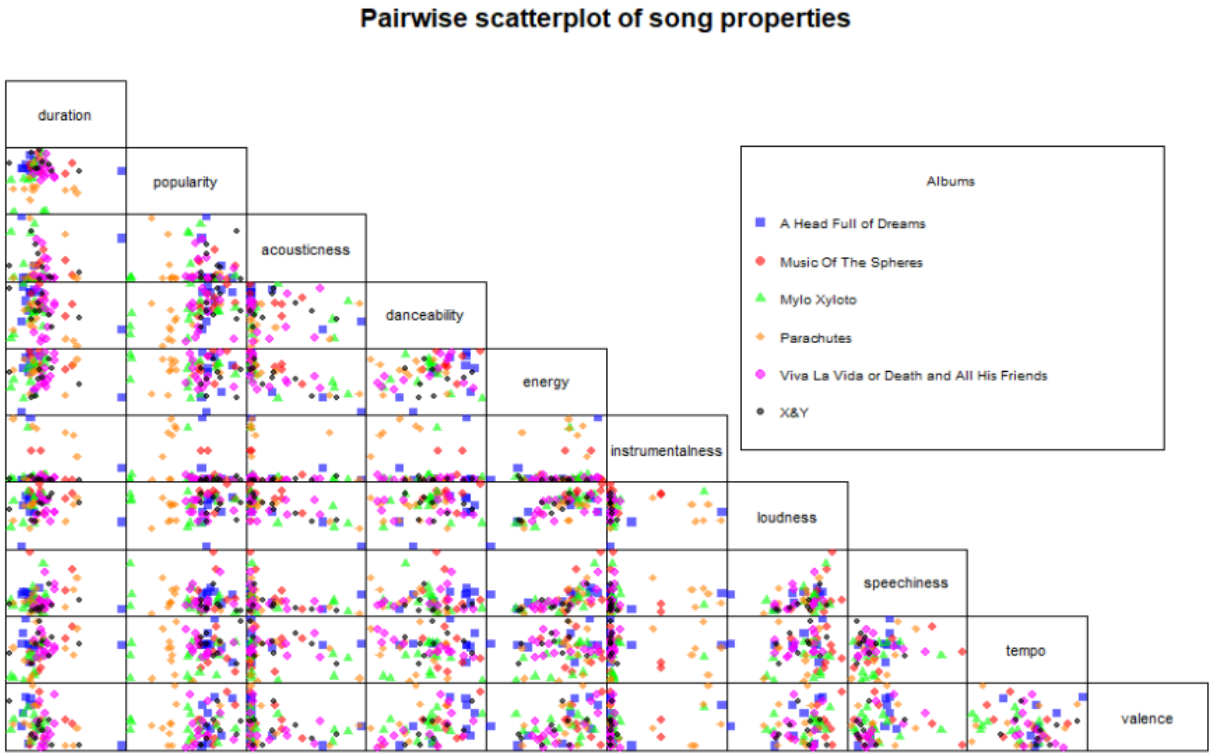

Figure 3. Pairwise scatterplot of song properties with different albums grouped.

From the figure 3, it can be observed that only the albums Parachutes and Mylo Xyloto appear as groups in some of the scatterplots. In the pairwise plots including popularity, the album Parachutes appears as a group, while the other albums aren't as disctinct.

3 Conclusion

The bivariate analysis was conducted using a Pearson's correlation measure, which measures linear correlation. By only studying the linear correlation, other degree correlations are left unobserved. Therefore, some of the conclusions drawn in section 1.2 can be faulty, and the key to a popular song might have not been solved. Many of the results of bivariate analysis may have also been affected by factors outside the songs, such as the music listened during that era and evolution of technology from 2000 onwards.

From the missclassification rate, obtained in section 2.3, it is clear that the Fisher's linear discriminant failed to group the albums. Missclassification rate of 0.77 indicates that the method performed very poorly, which could be due to numerous factors. The choice of variables could have been more carefully selected, and some of them discraded. The song attributes obtained through Spotify API can simply be too similar for a single band, which causes the loss of distinct groups. The albums also consist of only 9 – 14 songs, which may lack the trends required for the discriminant analysis. Some other classification method could have performed better on the data. However, I assume that the songs are just too similar, and the groups hold too many entries to be distinctly grouped.

Even though the multivariate analysis failed in the classification task, some interesting aspects were still found out about the data through other measures. Application of the multivariate analysis was also very interesting. Whether the method worked precisely or not, I am still very excited to finally see the band perform next summer.

References

1. Md Faizal, K., *Coldplay - Albums & Live shows*. 2023, Kaggle.