Vrije Universiteit Amsterdam

Multivariate statistics

Group 6

Academic year 2024-2025

PCA Analysis: Musical Landscape

Authors:

Pooya Mers - (2734368) - (p.zare.shah.mers@student.vu.nl)

March 19, 2025

1 Introduction

This report encompasses the findings of the Principal Component Analysis(PCA) of the pop-music landscape, using the 'popmusic.csv' dataset.

In the dataset, there are 833 songs in total, from 17 different artists. The scope of this report is the following ten numerical variables: (minor, chordnum, danceability, energy, loudness acousticness, liveness, valence, tempo, duration_sec), the detailed description of which can be found in the appendix.

Since ten variables are too many to visualize insightfully, they will be summarized into four new summary variables, with the aim of retaining as much variation as possible. The reduced number of variables(i.e. dimensions) then enables easier interpretation of the variation in the dataset.

1.1 Principle Component Analysis (PCA)

To start with PCA, an initial hurdle must first be tackled: the variables in the dataset X have substantially different scales. This is an issue, since PCA is a variance-maximization process, and if left untreated, PCA will load on the variables with the largest variances, which would (generally) be the ones with the largest scales. Therefore, we define $D = diag(s)$, where s is a column array of the variances of the variables.

Additionally, since the means of the variables are different, they would be difficult to interpret visually. We can use the "demeaning matrix" $H = I_n - n^{-1}\iota_n\iota_n^T$, where ι_n is an array of ones.

Lastly, ensuring that the variables are centered around zero would make visual interpretation easier, and we can apply this by diving the data by a factor of $\sqrt{n}$. Combining these transformations, the new normalized dataset can be defined as follows:

$X^* = \frac{1}{\sqrt{n}} H X D^{-1/2}$

We can then perform PCA in the following steps:

1. Compute each eigenvalue λ and the corresponding eigenvector of ${X^*}^T X^*$.
2. Sort the eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4$

3. Construct the loadings $\gamma_i, 1 \leq i \leq 4$, where γ_i is the eigenvector corresponding to λ_i.

4. Construct the factors $F_i = X^* \gamma_i$

Then the *compressed* dataset is then defined as $X^{PCA} = [F_1, F_2, F_3, F_4]$

To have a better idea of what the principal components represent, we can inspect the loadings γ_i:

	PC_1	PC_2	PC_3	PC_4
minor	-0.009	0.296	0.633	0.168
chordnum	0.129	-0.331	0.342	-0.332
danceability	-0.329	0.499	0.104	0.012
energy	-0.484	-0.258	0.023	0.138
loudness	-0.459	-0.237	-0.02	0.151
acousticness	0.464	0.158	-0.01	-0.214
liveness	-0.049	-0.254	0.632	-0.284
valence	-0.396	0.186	0.052	-0.206
tempo	-0.135	-0.344	-0.232	-0.514
duration_sec	0.19	-0.436	0.12	0.622

Table 1: Table of $\gamma_{i=1,2,3,4}$: Principal components vs the original variables

Using the results in table 1, we can form an interpretation of how each principal component is associated with the original variables. To identify the most important variables, we pick the largest ones in magnitude. Selecting a boundary for what is considered a large number is subjective, but here, we pick 0.3 to be the minimum, because the maximum possible number in the eigenvectors is around 0.6, and a bound of half this value ensures that there is at least some significance. Therefore, we can make the following interpretations for each principal component:

PC_1 (Principal Component 1)

Highest loadings: Energy (-0.484), Loudness (-0.459), Acousticness (0.464), Valence (-0.396), Danceability (-0.329)

Interpretation: PC_1 appears to represent "musical energy and acoustic balance," as it is dominated by energy, loudness, acousticness, and to a lower extent, danceability. Negative loadings for energy and loudness suggest that this component separates songs that are intense and loud from those that are softer and more acoustic.

PC_2 (Principal Component 2)

Highest loadings: Danceability (0.499)Duration_sec (-0.436), Tempo (-0.344), Chordnum (-0.331)

Interpretation: PC_2 seems to represent "rhythmic and tonal characteristics," as it combines (lower) musical complexity, danceability, minor key presence, and tempo/duration traits. Positive loadings for danceability suggest this component captures songs with rhythmic appeal.

PC_3 (Principal Component 3)

Highest loadings: Minor (0.633), Liveness (0.632), Chordnum (0.342)

Interpretation: PC_3 appears to capture "musical mood and live performance characteristics," as it is influenced by whether a song is in a minor key, its liveness, and chord complexity.

PC_4 (Principal Component 4)

Highest loadings: Duration_sec (0.622), Tempo (-0.514), Chordnum (-0.332)

Interpretation: PC_4 seems to represent "tempo and length characteristics," focusing on how long or fast-paced a song is, with negative tempo loadings suggesting slower songs. The songs also have lower chord numbers.

We can also use the squared correlation between the original variables and loadings to determine the % variance of each variable explained by each Principal Component: Since the data is normalized, we can use the following formula: $Corr^2(X_j, \gamma_k^T) = (\lambda_k^{1/2}\Gamma_{jk})^2$ to construct a squared correlation table 2.

Some of the original variables such as energy and loudness retain much of their variance through the four PCs, while others such as chordnum retain much less of their variance explained through the PCs. Therefore, the interpretations we make based on the loadings may not be as accurate for some variables, because the factors do not explain a large enough portion of the total variation of that variable. We can further explore how much of the total variation of the *dataset* is retained through the principal components.

	PC_1	PC_2	PC_3	PC_4
minor	0	0.12	0.43	0.03
chordnum	0.05	0.15	0.12	0.11
danceability	0.35	0.33	0.01	0
energy	0.77	0.09	0	0.02
loudness	0.69	0.08	0	0.02
acousticness	0.71	0.03	0	0.04
liveness	0.01	0.09	0.42	0.08
valence	0.51	0.05	0	0.04
tempo	0.06	0.16	0.06	0.25
duration_sec	0.12	0.25	0.02	0.37

Table 2: % Variance of each original variable explained by each PC

1.2 How much of the variation is retained?

Throughout the analysis, we took a pre-determined choice of selecting four PCs from the original ten variables. However, it's important to investigate whether this is a good selection, and whether more or less PCs should be selected.

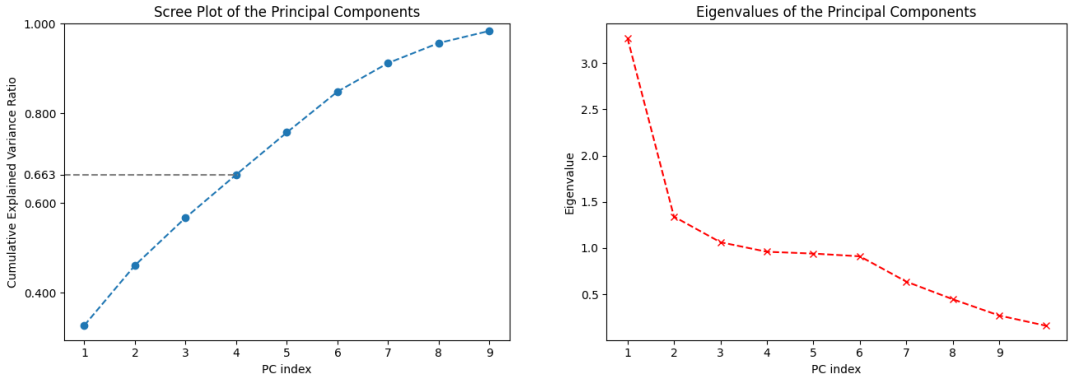
We can compute the variation explained by each principle component(F_i), using the fraction of the associated eigenvalue(λ_i) to the total sum of all eigenvalues:

$$\psi_r = \frac{\sum_{i=1}^{r^*} \lambda_i}{\sum_{i=1}^{r} \lambda_i} \tag{1}$$

In particular, after reducing the number of of variables from $r = 10$ to $r^* = 4$ we have:

$$\psi_4 = \frac{\sum_{j=1}^{4} \lambda_i}{\sum_{j=1}^{10} \lambda_i} \approx \frac{6.635}{10.000} = 66.35\% \tag{2}$$

Furthermore, we can use equation (1) to construct a scree plot of the principal components:



(a) Scree Plot

(b) Eigenvalues

Figure 1: On the left: Scree Plot of the PCs. The blue dots show the cumulative variance retained by the . On the right: Eigenvalue plot of the PCs. The red dots show the respective eigenvalue of a PC.

Scree plots can be used as a tool to select the number of principal components, and they are most useful when there is a notable change in marginal variance retained after selecting a certain number of PCs. In the scree plot we can see that after selecting additional principal components F_5, F_6, there would be notable marginal increase in variance retained, but this seems to start to flatten from F_7. Therefore, it would also be justifiable to take two additional PCs, totaling six. However, more principal components would reduce interpret-ability, because the additional noise would distort the underlying correlation between the variables that we wish to summarize.

1.3 Analyzing the songs along the two main Principle Components

We can scatter plot the compressed data of X^* along the two dimensions with the highest explained variation: PC_1 and PC_2.

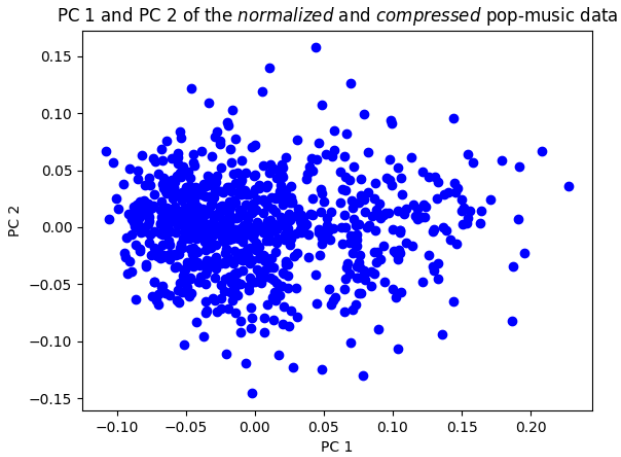


Figure 2: Plot of the compressed data X^{PCA} along PC_1, PC_2; the two dimensions with the highest explained variation.

In general, the songs seem to be distributed in a main cluster on the left(low PC_1) and center(near zero PC_2), and a second smaller cluster towards the right (high PC_1). There are scattered points in extremities of PC_2 and the higher end of PC_1.

We saw before that higher values of PC_1 represented sadder songs, while the lower values represented happier, more energetic songs. Higher values of PC_2 on the other hand, represented songs that were shorter, more danceable, less musically complex, and with more minor chords. The main cluster of songs is centered at a low PC_1 value, from (-0.10) to around (0.05) which indicates that most songs are happy and energetic. The largest PC_1 value is more than (0.20), far higher than the average. There also seems to be a second smaller cluster of songs to the right of $PC_1 = 0.05$, ranging until around (0.15). This second cluster with higher PC_1 values may represent more melancholic and acoustic pop songs.

With regards to PC_2, most of the songs are scattered around $PC_2 = 0$, ranging from around (-0.7) to (0.05). The number of positive points seems to be higher than negative ones, suggesting that pop songs have high danceability and low duration. Here, there are very high anomalies near (0.15), but also very low anomalies near (-0.15). This suggests that pop songs can have very high or very low danceability and duration.

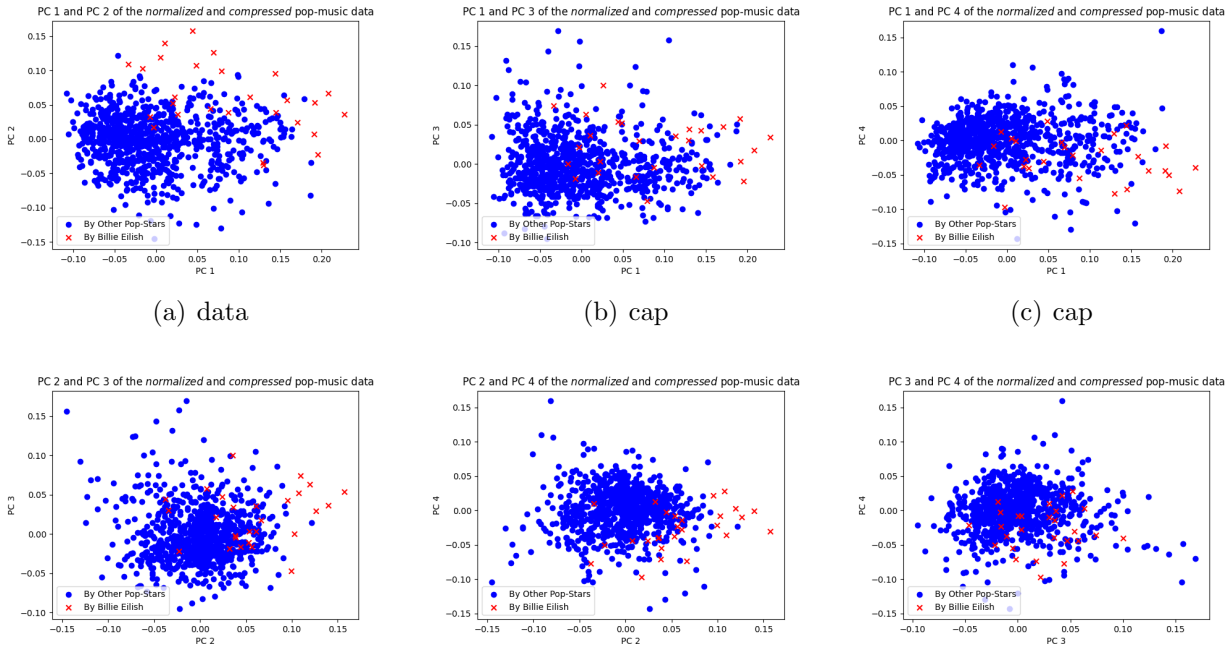
Some people claim that "all pop songs sound the same", but the findings supports this claim partially. We see that indeed, most pop songs are clustered together with respect to PC_1 and

PC_2 but there are many pop songs that differ in terms of the variables that form these principal components.

1.4 Billie Eilish's songs

In the recent years, Billie Eilish has risen above the popularity ranks of pop-stars, winning many accolades such as The Grammy and American Music awards (1). Her song "bad guy" has been streamed more than two billion times on Spotify as of March 2025 (2). It is a good question to ask; how are her songs different in comparison to other pop-stars?

We can plot her songs in a different color (red) compared to the songs from other artists (blue) to see if they are placed differently.



(a) data (b) cap (c) cap

(d) cap (e) cap (f) cap

Figure 3: Scatter plots of the compressed data $X^{PCA}PC_i$, PC_j for $i, j = 1, 2, 3, 4$ and $i \neq j$. The red crosses show the songs that were made by Billie Eilish, and the blue dots show the songs by other pop-artists.

Figure 3 shows interesting results. Billie Eilish's songs generally have a wide range of values of PC_1, but they seem mostly skewed towards the right. Only two of her songs are placed in the main cluster, and the others place outside this range. She has the most songs with the highset

PC_1 values out of all other artists. In regards to PC_2, her songs are mostly placed above the main cluster, and she has again the most songs with the highest PC_2 values.

Her songs tend to have, on average, slightly higher PC_3 and lower PC_4 values compared to songs by other artists, but the difference is not as significant of those of PC_1 and PC_2. In fact, many of Billie's songs fall into the main cluster with respect to PC_3 and PC_4. Therefore, it is not possible to claim that Billie's songs are in general unique in terms of these two principle components.

Looking back at the interpretations of the principal components, we can make some general claims about Billie's songs, comparatively:

1. Her songs are comparatively sad and melancholy, less energetic, and quiet. Additionally, her songs possess more acousticness.

2. Her songs are somewhat similar to those of other artists in terms of duration, tempo, and liveness. The danceability of her songs may also be different, but since this is a variable with a negative coefficient for PC_1 and a positive coefficient for PC_2 in table 1, there is not enough evidence to make a claim about it.

Billie's songs are unique in terms of being melancholy and low energy, but at the same time danceable. It seems that one of the reasons behind Billie's success has been coming up with a style that it different in these characteristics, which shows the reward of creativity in the musical landscape.

References

[1] Website, *Award list of Billie Eilish, March 2025* (https://www.grammy.com/artists/billie-eilish/251741).

[2] Website, *Spotify stream count for 'bad guy' by Billie Eilish, March 2025* (https://open.spotify.com/track/2Fxmhks0bxGSBdJ92vM42m)

2 Appendix

Variable definitions

In both the songlist and popmusic dataset, variables are defined as follows:

- acousticness: A measure from 0.0 to 1.0 (0-100 in the songlist dataset) indicating to what extent the track has an acoustic feel. Acoustic songs use little amplification (no loud electric guitars, etc) and have an intimate and cosy feel. 1.0 represents fully unamplified, acoustic songs.
- danceability: describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. (rescaled to 0-100 in the songlist dataset).
- duration_sec: duration of the track in seconds.
- energy: a measure from 0.0 to 1.0 (rescaled to 0-100 in the songlist dataset) which represents a measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Features contributing to this attribute include dynamic range, perceived loudness, timbre, etc.
- liveness: indicator between 0.0-1.0 (rescaled to 0-100 in the songlist dataset) for how live (i.e. not recorded in a studio) the performance feels. Higher liveness values represent an increased feeling that the song was performed live or that it features cheering, applause, several people singing, audience noises, etc.
- loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track. Values typically range between -60 and 0 db.
- tempo: The overall estimated tempo of a track in beats per minute (BPM). Musical tempo is the speed or pace of a given song and derives directly from the average beat duration.
- valence: A measure from 0.0 to 1.0 (rescaled to 0-100 in the songlist dataset) describing the musical positiveness conveyed by a track. Tracks with high "valence" sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- chordnum: integer indicating how many distinct chords appear in total in the song. The number of chords is an indication of the musical complexity of the song. If a song features a high number of distinct chords, it tends to be more musically complicated.
- minor: a value between 0.0 to 1.0 indicating the presence of many minor chords in the song. Minor chords are generally sad sounding chords. If the value is low, then major chords (which are happy sounding chords) are more frequent.
- popul: integer between 0 and 100 indicating the popularity of the song on Spotify, as indicated by the number of streams. A value of 100 indicates it is among the most streamed songs on Spotify.

Figure 4: Definitions of the variables in the dataset