Principal Component Analysis (PCA) for Dimensionality Reduction of Spotify Music Data

Author: Matias Mu

Tool Used: IBM SPSS Statistics

Technique: Principal Component Analysis (PCA)

Objective: Reduce multicollinear numerical variables into uncorrelated principal components

Applying Principal Component Analysis to Digital Music Data

1. Introduction

In this project we will analyse the Spotify 1921-2020 dataset containing music data for more than 170,654 songs. The main objective of this project is to reduce the variables of the dataset by grouping them based on their correlation using Principal Component Analysis (PCA) in order to facilitate further analysis of these variables without having to exclude relevant data.

By applying PCA, we also aim to discover how many components are needed to explain the variability of the data, also to identify which musical attributes are most influential in creating the principal components and finally to interpret the principal components meaning in a practical and realistic way.

2. Background

To understand the purpose and application of Principal Component Analysis it is important to have a clear definition of it. In the book Principal Component Analysis (2nd ed., Springer, 2002), PCA is defined as “The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables”. In the context of a project, the use of Principal Component Analysis is essential in order to reduce the size of variables by grouping them into principal components, therefore containing as much variation as possible without having to exclude variables.

It is important to mention that in order to reduce the number of variables, multicollinearity is necessary. According to paper 1404-2017. Multicollinearity: What Is It, Why Should We Care, and How Can It Be Controlled? (Schreiber-Gregory, 2017), Multicollinearity is defined as “the statistical phenomenon wherein there exists a perfect or exact relationship between predictor variables. From a conventional standpoint, this occurs in regression when several predictors are highly correlated. Another way to think of collinearity is ‘co-dependence’ of variables”. That said, if there is no multicollinearity, they cannot be grouped into principal components as they are not correlated with each other.

On the other hand, the Spotify 1921-2020 dataset, created by Yamac Eren Ay and published in Kaggle, has been selected for this project. This dataset contains more than 170,654 songs extracted through the Spotify Web API (Ay, 2021). The size of the dataset is significant in order to perform PCA. According to Statistics Laerd, it is recommended that “a minimum of 150 cases, or 5 to 10 cases per variable, it’s recommended as a minimum sample size”. (Laerd Statistics, n.d.).

3. Data Description

Spotify 1921-2020 dataset contains 170,654 observations and 19 columns, including different types of variables such as categorical, numeric, binary and ordinal variables. This dataset has been extracted using the official Spotify Web API. In order to perform the PCA, only numerical variables were selected, these are the follows:

Name	SPSS Type	Range	Example	Description
valence	Scale	0-1	0.428	Tracks with high valence sound more positive, while tracks with low valence sound more negative.
acousticness	Scale	0-1	0.00242	How acoustic a track is
danceability	Nominal	-	0.585	How suitable a track is for dancing. A value of 0.0 is least danceable and 1.0 is most danceable.
duration_ms	Nominal	-	237040	The duration of the track in milliseconds.
energy	Nominal	-	0.842	Intensity and activity of a track. (e.g. feel fast, loud, and noisy).
instrumentalness	Nominal	-	0.00686	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context.
liveness	Nominal	-	0.0866	The presence of an audience in the recording.
loudness	Nominal	-	-5.883	The overall loudness of a track in decibels (dB).
speechiness	Nominal	-	0.00556	Presence of spoken words in a track.
tempo	Nominal	-	118.211	The speed or pace of a given piece and derives directly from the average beat duration.
popularity	Scale	0-100	5	Popularity of track in range 0 to 100
year	Nominal	-	1921	The year of release

Table 1: Data description table. Sources: Kaggle and Spotify for Developers.

Categorical variables such as artist, name, id and also binary variables such as explicit, key and mode were excluded because these variables were not relevant for the purpose of the PCA. On the other hand, the variables used in the analysis were not standardised, as the correlation matrix in SPSS was applied, which is the equivalent of standardising the variables. (National College of Ireland, 2024. Horn, 2024).

4. Results

KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.699
Bartlett's Test of Sphericity	Approx. Chi-Square	884863.141
	df	66
	Sig.	.000

Image 1: KMO and Bartlett's Results. Source: SPSS.

The KMO should exceed 0.5. (NCI, 2024. Horn, 2024, slide 376). In this case the value of KMO is (0.699), so the dataset is suitable for the PCA. On the other hand, for the analysis to be valid, the null hypothesis must be rejected, so the p-value must be (< 0.05), (NCI, 2024. Horn, 2024 slide 376). In this case it is (0.000), so we can reject the null hypothesis and confirm that the data are adequate to perform the PCA.

Correlation Matrix

Correlation Matrix													
		valence	year	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	popularity	speechiness	tempo
Correlation	valence	1.000	-.028	-.184	.559	-.192	.354	-.199	.004	.314	.014	.046	.172
	year	-.028	1.000	-.614	.189	.080	.530	-.272	-.057	.488	.862	-.168	.141
	acousticness	-.184	-.614	1.000	-.267	-.076	-.749	.330	-.024	-.562	-.573	-.044	-.207
	danceability	.559	.189	-.267	1.000	-.140	.222	-.278	-.100	.285	.200	.235	.002
	duration_ms	-.192	.080	-.076	-.140	1.000	.042	.085	.047	-.003	.060	-.085	-.025
	energy	.354	.530	-.749	.222	.042	1.000	-.281	.126	.782	.485	-.071	.251
	instrumentalness	-.199	-.272	.330	-.278	.085	-.281	1.000	-.047	-.409	-.297	-.122	-.105
	liveness	.004	-.057	-.024	-.100	.047	.126	-.047	1.000	.056	-.076	.135	.008
	loudness	.314	.488	-.562	.285	-.003	.782	-.409	.056	1.000	.457	-.139	.210
	popularity	.014	.862	-.573	.200	.060	.485	-.297	-.076	.457	1.000	-.172	.133
	speechiness	.046	-.168	-.044	.235	-.085	-.071	-.122	.135	-.139	-.172	1.000	-.012
	tempo	.172	.141	-.207	.002	-.025	.251	-.105	.008	.210	.133	-.012	1.000

Image 2: Correlation Matrix Results. Source: SPSS.

Positive correlations: Year and Popularity (0.862), Energy and Loudness (0.782), Danceability and Valence (0.559), Energy and Year (0.530). There are also negative correlations such as Acousticness and Energy (-0.749), Acousticness and Year (-0.614) and Popularity and Acousticness (-0.573). Based on these results we can infer there are significant correlations and therefore, it is necessary to carry out the PCA.

4.3. Communalities

Communalities

	Initial	Extraction
valence	1.000	.753
year	1.000	.841
acousticness	1.000	.715
danceability	1.000	.720
duration_ms	1.000	.313
energy	1.000	.810
instrumentalness	1.000	.406
liveness	1.000	.715
loudness	1.000	.706
popularity	1.000	.809
speechiness	1.000	.692
tempo	1.000	.461

Extraction Method: Principal Component Analysis.

Image 3: Communalities Results. Source: SPSS.

The most explained variance of the variables are: Year (0.841), Energy (0.810), Valence (0.753) and Popularity (0.809), which indicates that its representation is well captured. Variables with lower scores are duration_ms (0.313), instrumentalness (0.406) and tempo (0.461), which have a lower representation. This result shows that most of the variables are sufficiently representative to perform the PCA.

4.4. Total Variance Explained

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.928	32.733	32.733	3.928	32.733	32.733	3.528	29.397	29.397
2	1.753	14.612	47.345	1.753	14.612	47.345	1.786	14.881	44.278
3	1.182	9.847	57.192	1.182	9.847	57.192	1.374	11.446	55.724
4	1.077	8.979	66.171	1.077	8.979	66.171	1.254	10.447	66.171
5	.915	7.622	73.793						
6	.874	7.283	81.076						
7	.733	6.112	87.188						
8	.620	5.164	92.352						
9	.360	3.002	95.353						
10	.304	2.533	97.886						
11	.133	1.104	98.991						
12	.121	1.009	100.000						

Extraction Method: Principal Component Analysis.

Image 4: Total Variance Explained Results. Source: SPSS.

Four principal components have been created, which explain 66.17% of the total variance. In this case we are excluding the other 33.83%, but doing so is still more representative than eliminating variables. According to Initial Eigenvalues % of Variance, PC1 explains 32.73%, PC2 explains 14.61%, PC3 explains 9.84% and finally PC4 explains 8.98%. Creating these Principal Components we are avoiding to fail into redundancy of data without excluding information.

4.5. Scree Plot

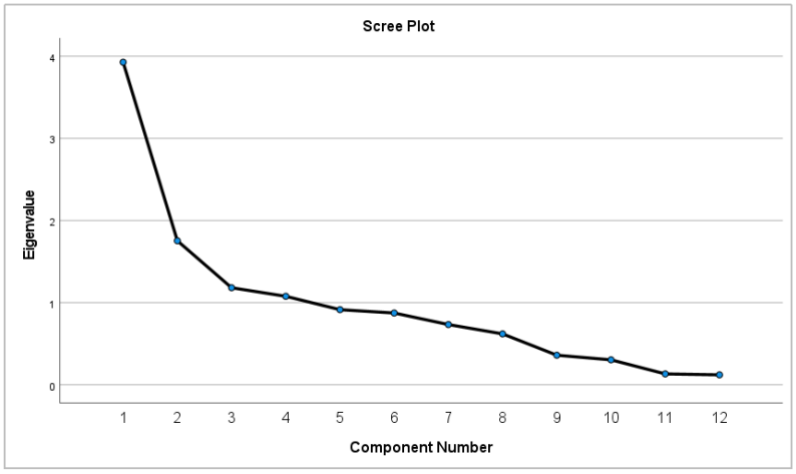

Image 5: Scree Plot. Source: SPSS.

The Scree Plot shows a steep drop from PC1 to PC3, where it is more difficult to see the drop between PC3 and PC4. So in this case it is better to consider the results from the Total Variance Explained Table.

4.6 Rotated Component Matrix

Rotated Component Matrix[a]

	Component			
	1	2	3	4
valence		.756		
year	.890			
acousticness	-.806			
danceability		.775		
duration_ms		-.526		
energy	.714		.515	
instrumentalness				
liveness				.756
loudness	.675			
popularity	.869			
speechiness				.677
tempo			.671	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 9 iterations.

Image 6: Rotated Component Matrix Results. Source: SPSS.

We can observe that the main loads of PC1 were Year (0.890), Popularity (0.869), Energy (0.714) and Loudness (0.675). For PC2, the main loads were Danceability (0.775), Valence (0.756) and Duration_ms (-0.526). It is important to note that the negative variable loaded represents the opposite of a positive value, in this case the songs are the opposite of a high duration.In PC3 the loads were Tempo (0.671) and Energy (0.515). Finally, in PC4, the main loads were Liveness (0.756) and speechiness (0.677).

** It is important to mention that the principal components were kept in SPSS for future analysis.*

FAC1_1	FAC2_1	FAC3_1	FAC4_1
-1.49266	-3.63885	.03234	2.01156
-.98010	2.29240	-1.28009	1.29572
-1.59811	-1.86529	.03039	-.49206
-1.27310	-.99525	.38399	.83167
-1.33949	-.16069	.00185	.19690
-.83705	-.14942	-.17052	.24507
-1.22125	.53825	-.80364	-.13368
-1.81850	-1.12754	-.51033	.38035
-1.77089	.83333	-.86727	-.39064
-1.52745	1.10727	-.23579	2.36783
-1.67778	.84754	.08871	.29259
-2.29122	-.02407	-.64530	-.89355
-1.72582	-.22916	1.83954	.42805
-1.50590	-.43943	-.22039	-.36493
-.89922	-.08990	2.26794	.84111
-1.97253	-.61899	-.02633	-.76573
-1.40789	-.69025	-.23507	.25694
-1.99272	-.07268	.98892	.32738

Image 6: Principal Components Variables. Source: SPSS.

5. Discussion and Analysis of Results

Having obtained the results, we can then translate them into practical terms for interpretation and application. Each principal component will be analysed and explained below:

PC1: Modernity

This PC is loaded by 4 variables, the ones with the highest loads are Year (0.890) and popularity (0.869), these variables can be interpreted for example as latest hits. The remaining variables are energy (0.714) and loudness (0.675), these variables can be interpreted as more vitality and noisy. These variables together suggest a musical pattern that is mostly characteristic of modern songs. That said, we could infer that energetic and noisy new songs may have a strong correlation with this component. This type of song may perform well at young people's events, shopping malls or clothing shops.

PC2: Danceability

This PC is loaded by 3 variables which are Danceability (0.775), Valence (0.756) and Duration_ms (-0.526). This set of correlations shows that this component is related to happy songs, which can be danceable and do not have a long duration. It is important to note that the duration_ms variable, being a negative correlation, suggests that these songs have a shorter duration than the others, which is congruent with danceable songs. That said, if a song has a high correlation with PC2, it means that these songs are very suitable for parties, discos or dancing events.

PC3: Rhythm

PC3 is loaded by 2 variables which are Tempo (0.671) and Energy (0.515). This shows that this component has a relationship with fast songs, with marked rhythm and energy. It could be inferred that songs that have a strong correlation with PC3 are suitable for contexts that demand energy or physical activity such as a gym, training or sporting events.

PC4: Live Music

PC4 is loaded by 2 variables. The first variable Liveness (0.756) reveals whether the song was recorded live or has speech interaction rather than musical sounds. Speechiness (0.677) refers to whether the song contains spoken rather than singing voices. That said, this PC is different from others, as depending on the level of correlation with a variable we will be able to identify whether the song is Live Music or not. Therefore, it is not possible to infer a specific place to use this type of songs.

6. Conclusion

The objective of this project was to reduce the dimensionality of the Spotify dataset using PCA to facilitate future statistical analysis without having to exclude relevant data. The feasibility of applying the PCA was validated by the KMO (0.699) and Bartlett's Test with (0.00). Each component was interpreted based on its loadings and translated into practical terms.

After conducting the PCA, it was possible to reduce from 12 variables to 4 variables explaining 67.17% of the total variance of the data set. Based on the results, we have been able to discover that the new principal components generated have a unique meaning for future analysis without having multicollinearity between them. Within the principal components we can interpret PC1. Modernity can be used to identify recent, energetic and loud song music, on the other hand, P2. Danceability can be used to identify music related to cheerful and short duration. Also PC3. Rhythm can be used to identify high tempo and energy songs and finally P4. Live Music, that can be used to identify songs that are live and spoken.

The result obtained not only helped us to understand that there was similarity between the variables of the dataset, but also in terms of the generation of the principal components will allow us to have clarity in terms of interpretation and implementation of future statistical analysis without having to eliminate variables and still having a high representativeness of the data.

7. References

- Ay, Y. E. (2021). *Spotify Dataset 1921-2020, 160k+ Tracks*. Kaggle. https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks
- Horn, C. *Courtesy slide. National College of Ireland*. (2024). Dimension reduction. Page 377.https://moodle2024.ncirl.ie/pluginfile.php/78636/mod_resource/content/1/Stats2_10.pdf
- Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer. http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf
- Laerd Statistics. (n.d.). *Principal components analysis (PCA) using SPSS Statistics*. Statistics Laerd. https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php
- Schreiber-Gregory, D. N. (2017). Multicollinearity: What Is It, Why Should We Care, and How Can It Be Controlled? Henry M Jackson Foundation / National University.https://support.sas.com/resources/papers/proceedings17/1404-2017.pdf
- Spotify Developer. (n.d.). *Get audio features*. Spotify. https://developer.spotify.com/documentation/web-api/reference/get-audio-features