PCA implementation on Spotify Recommender Systems

Professor: Upendra Persaud

Student: Jasmin Morsy

Class: M544 – Numerical Linear Algebra for Big Data

Date: 8/17/2024

Introduction/Objectives

Principal Component Analysis (PCA) is a powerful dimensionality reduction technique that can be effectively applied to Spotify music data, which typically contains a vast array of features such as tempo, liveness, danceability. Those features significantly influence a song's popularity by aligning with listener preferences and cultural trends. Songs with appealing and engaging features are more likely to resonate with a broader audience, leading to higher streaming numbers

By reducing the number of dimensions, PCA helps in uncovering underlying patterns in the music data, making it easier to analyze and visualize. This condensed data can then be utilized to implement a music recommendation system, which suggests tracks based on the principal components that capture the essential characteristics of the songs. Such a system not only enhances user experience by offering personalized music choices but also improves computational efficiency by focusing on the most relevant data features.

Data Sourcing, Description, and Preparation

(Appendix A)

Data Collection:

https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset/input

Data Shape: (170653, 19)

Data Cleaning: (Appendix B)

- Multi-Label Binarizer for Only the Top 100 Artists
- Standardizing the numerical variables
- Dropping unnecessary columns: 'id', 'artists', 'release_date'
- **Missing values**: 'release_year' missing values (119798), imputed with the median year
- **Infinite values:** (119798) Imputed with K-nearest neighbors (KNN)
- **Duplicate rows**: 597 rows were dropped
- **Outliers:** were capped with nearest acceptable value within the threshold

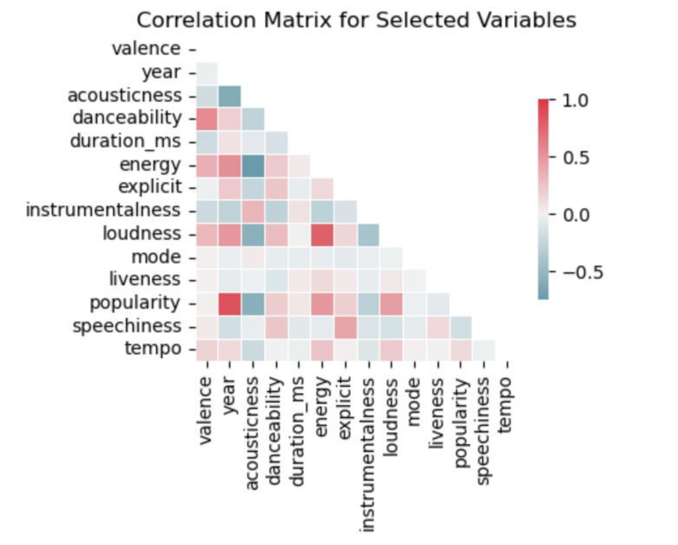
Data Shape after Cleaning (169929, 116)

Data Exploration, Summary Statistics And Data Visualization

Summary Statistics (Appendix C): count, mean, std, min, Q1, Q2, Q3, max

Distribution of Key Variables (Appendix D): Density Plot

Correlation Matrix for Variables of Interest (Appendix E):



Checking for Multicollinearity:

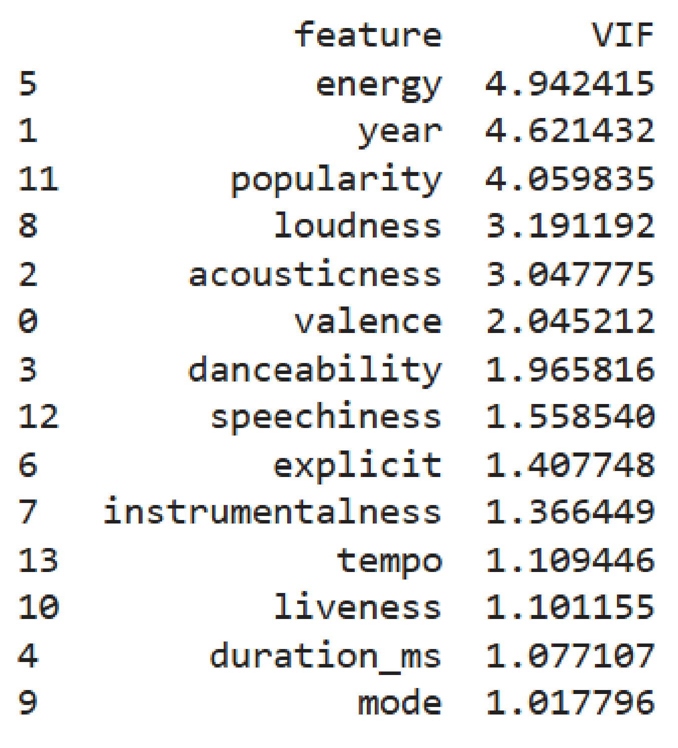
Threshold for Variance Inflation Factor:

VIF = 1: No multicollinearity

1 < VIF < 5 :moderate multicollinearity

VIF >= 5: high multicollinearity

VIF > 10: very high multicollinearity



```
             feature       VIF
5             energy  4.942415
1               year  4.621432
11        popularity  4.059835
8           loudness  3.191192
2       acousticness  3.047775
0            valence  2.045212
3       danceability  1.965816
12       speechiness  1.558540
6           explicit  1.407748
7   instrumentalness  1.366449
13             tempo  1.109446
10          liveness  1.101155
4        duration_ms  1.077107
9               mode  1.017796
```

Data Visualization (Appendix F):

K-mean clustering + Pairplot for variables of interest

PCA Implementation

1) Covariance Matrix: (Appendix G)

Key Observations:

High positive covariance

Valence and Danceability (0.557):
Year and Popularity (0.859)
Energy and Loudness (0.779)
Explicit and Speechness (0.415)

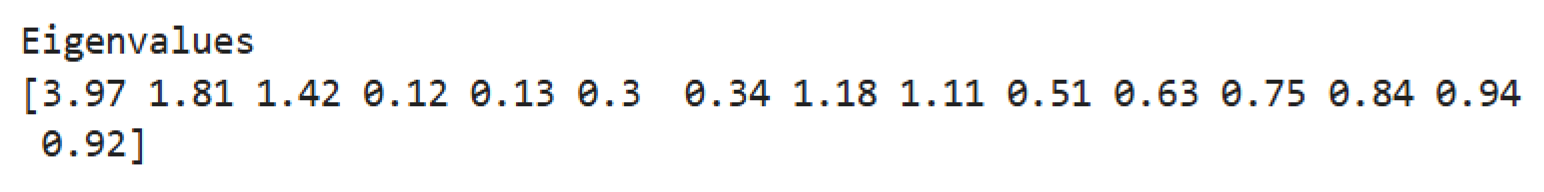
High Negative Covariance

Acousticness and Energy (-0.748)
Acousticness and Loudness (-0.558)
Acousticness and Year (-0.61)
Acousticness and Popularity (-0.569)

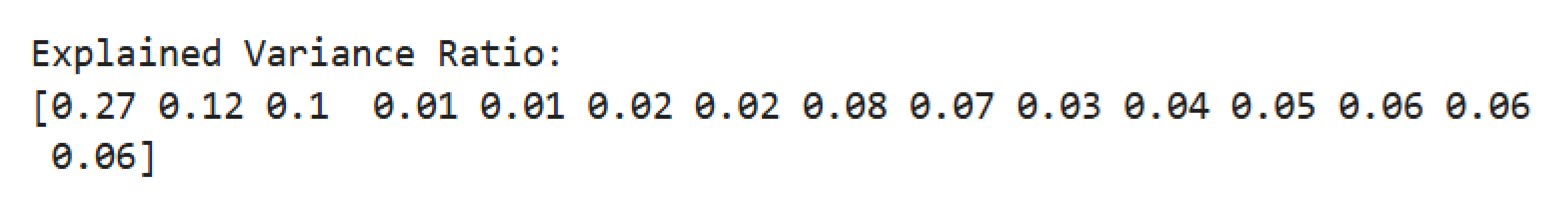
Covariance Matrix Conclusion

- **Trends in Music Production**: The strong positive covariance between year and popularity suggests that newer music is more popular. Additionally, newer tracks seem to be less acoustic and have higher energy, reflecting potential trends in modern music production.
- **Relationships between Musical Features**: The covariance between valence, energy, danceability, and loudness suggests that tracks with higher positivity and loudness also tend to be more danceable and energetic.
- **Listener Preferences**: The negative covariance between acousticness and both popularity and energy suggests that listeners may prefer less acoustic and more energetic tracks.

2) Eigenvalues Decomposition (Appendix H)



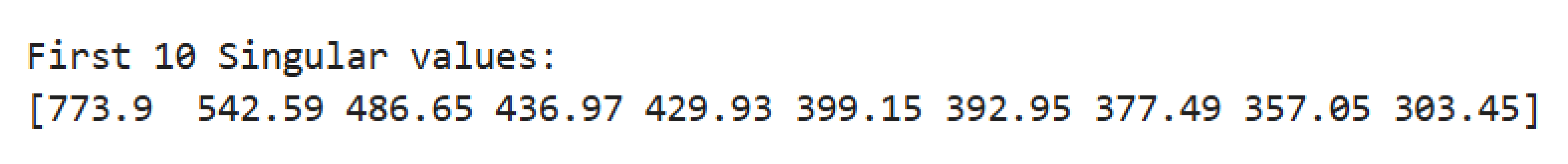
```
Eigenvalues
[3.97 1.81 1.42 0.12 0.13 0.3  0.34 1.18 1.11 0.51 0.63 0.75 0.84 0.94
 0.92]
```



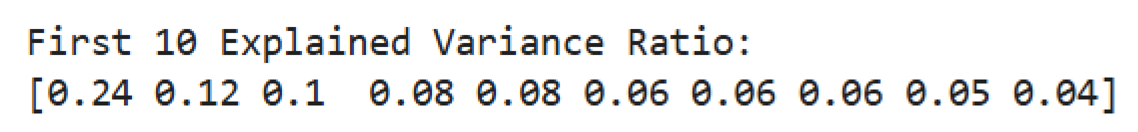
```
Explained Variance Ratio:
[0.27 0.12 0.1  0.01 0.01 0.02 0.02 0.08 0.07 0.03 0.04 0.05 0.06 0.06
 0.06]
```

Typically, the first few principal components (those with the highest eigenvalues) explain the majority of the variance in the dataset. In our dataset, the first 3 components capture only 49% of the most important patterns in the data. And other principle components that are associated with low eigenvalues contribute very little to explaining the variance in the dataset. We then conclude that Eigenvalues decomposition might not be the best approach to reduce the dimensionality of our dataset.

3) SVD Based Approach (Appendix I)



```
First 10 Singular values:
[773.9  542.59 486.65 436.97 429.93 399.15 392.95 377.49 357.05 303.45]
```

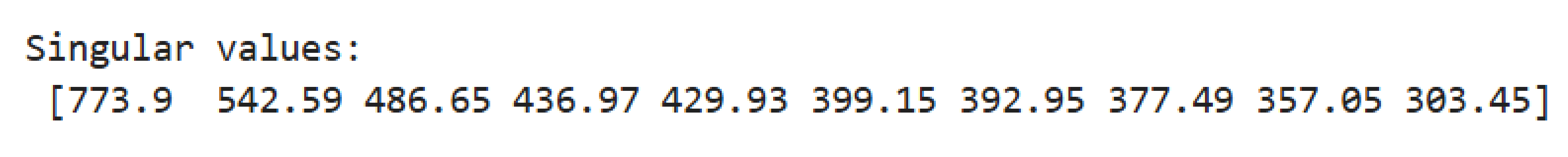


```
First 10 Explained Variance Ratio:
[0.24 0.12 0.1  0.08 0.08 0.06 0.06 0.06 0.05 0.04]
```

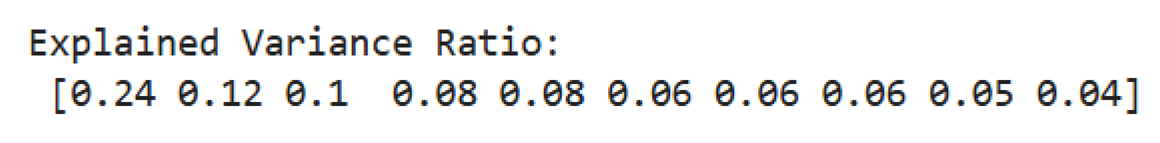
Based on the Singular Value Decomposition (SVD) results, the first principal component explains 26% of the variance in the dataset, indicating it captures the most significant variation among the features. The first two components together explain 38% of the variance, and by the time you include the **first three components**, **47%** of the variance is captured. The cumulative explained variance reaches **88%** after considering the **top ten components**, suggesting that these ten components collectively represent most of the important information in the dataset. This implies that dimensionality reduction to these ten components would retain a substantial portion of the dataset's variability, making it a reasonable choice for simplifying the data without losing much information.

.

4) Truncated SVD (Appendix J)



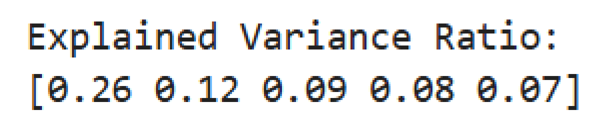
```
Singular values:
 [773.9  542.59 486.65 436.97 429.93 399.15 392.95 377.49 357.05 303.45]
```



```
Explained Variance Ratio:
 [0.24 0.12 0.1  0.08 0.08 0.06 0.06 0.06 0.05 0.04]
```

The truncated SVD approach reveals that the first few components capture the majority of the variance in the dataset. With just the first three components, nearly half (46%) of the variance is explained, and with the first six components, more than two-thirds (68%) of the variance is captured. This suggests that dimensionality reduction using SVD is effective in preserving much of the essential information from the original data, allowing for a more compact representation while retaining most of the variability.

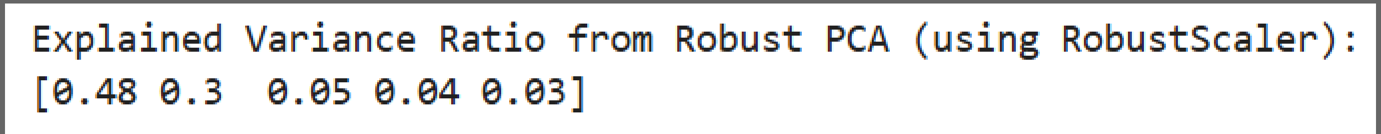
4) PCA (Appendix K)



```
Explained Variance Ratio:
[0.26 0.12 0.09 0.08 0.07]
```

The explained variance ratio passed on PCA suggests that the first five components collectively account for 62% of the total variance, suggesting that more than half of the dataset's variability can be represented by these five components alone. This reduction in dimensionality can simplify the dataset while retaining most of the important information.

3) Robust PCA (Appendix L)



```
Explained Variance Ratio from Robust PCA (using RobustScaler):
[0.48 0.3  0.05 0.04 0.03]
```

The explained variance ratio from Robust PCA using RobustScaler indicates that the first principal component (PC1) captures 48% of the total variance in the dataset, making it the most significant component. The second principal component (PC2) explains 30% of the variance, so together, the first two components account for 78% of the total variance. This means that these two components effectively summarize the majority of the dataset's variability.

Heat Map of the First Three RPCA Components

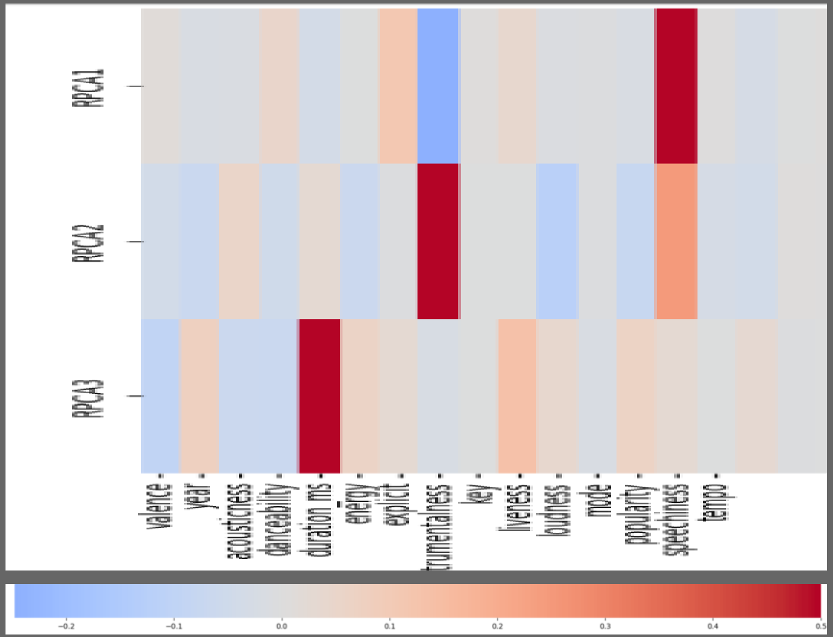
This Heat MAP visualizes the first 3 RPCA components and their relationship with different features.

Dominant Features:

Features show strong positive or negative correlations, indicated by red and blue colors. There’s a strong positive correlation in RPCA2with “speechiness: and a stong negative correlation with “intrumentalness”. Similarly, RPCA2 captures a strong positive correlation and a moderately negative correlation with “Instrumentalness” and “loudness” respectively. In addition, RPCA3 shows a stong positive correlation with “duration_ms”

Feature Contribution:

Features like “valence”, “dutation_ms”, "loudness“, “speechiness“, “instrumentalness” and “liveliness” appear to have higher contributions to RPCA components as indicated by the clear (not pale) red and blue colors.



Building a Recommendation System Based on RPCA

Robust Principal Component Analysis (RPCA) can be a valuable tool in building recommendation systems, particularly when dealing with noisy or incomplete data. RPCA effectively separates a data matrix into a low-rank component, which captures the underlying structure of the data, and a sparse component, which represents noise or outliers.

This decomposition is particularly useful in recommendation systems where user preferences or item features may be corrupted by noise or include missing values. By focusing on the low-rank structure, RPCA enhances the accuracy of recommendations, ensuring that they are based on the essential patterns in the data while minimizing the influence of anomalies. This approach leads to more reliable and robust recommendations, especially in scenarios with large, complex datasets.