Aalto University

MS-E2112 Project Work

Hierarchical cluster analysis of Billboard 200 number-one albums

Kalle Alaluusua (kalle.alaluusua@aalto.fi)

January 26, 2021

1 Introduction

This study examines acoustic features of the Billboard 200 number-one albums' content. The Billboard 200 is a record chart ranking the 200 most popular music albums and EPs in the United States. It is published weekly by Billboard magazine. This study focuses on the Billboard 200 number-one albums from 5.1.1963 to 19.1.2019. In the nearly 60-year time span the music industry has undergone dramatic changes. In this study, we explore the the number-one albums' content's acoustic features in order to discover common and differentiating factors. We hope to gain insight on the makings of best selling albums through the years.

2 Description of the data

The research data consists of 14607 rows containing acoustic data for tracks from Billboard 200 number-one albums from 5.1.1963 to 19.1.2019 [1]. Each row contains track name, album name, artist name, values for Spotify EchoNest acoustic data (acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, key, time signature, and valence), duration in milliseconds, and release date of the album.

We choose to ignore the week of the chart and the release date of the album in the analysis phase instead of treating the data as time series. Consequently, each album that has reached the number-one rank more than once is treated as single observation.

3 Preliminary analysis

Before further analysis, the variables in the acoustic features table, referred to as $\mathbf{X}$ from now on, are centered and scaled. As a result, each standardized variable has a mean of 0 and unit variance, which renders the data comparable. The standardized data are presented as a whole in Figure 1. To visualize both the shapes of the distributions and the specific point locations, a combination of a density plot and a strip chart with jitter was selected. The univariate variables are plotted on the vertical axis, and their values on the horizontal axis. A small random vertical displacement is added to each observation to reduce the overlap.

Three distributions stand out the most from the rest of distributions displayed in Figure 1 due to their discrete/categorical nature. The value corresponding to the time signature 4 (4/4) dominates other time signature values. The mode variable represents the modality of a song. Most of the songs are in a Major key as Major

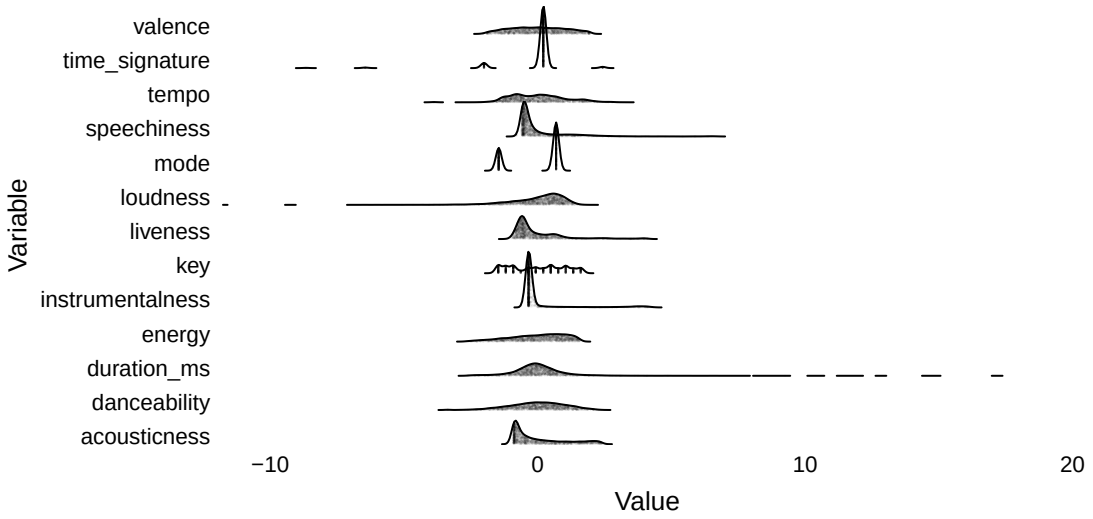

Figure 1: Density plots of the standardized acoustic features, where the line is displayed when its height is 0.1% or more relative to the overall maximum.

is represented by 1 and Minor by 0 in the original data. Finally, the key variable is approximately uniformly distributed.

The variables valence, tempo and danceability appear normally distributed. The rest of the variables depict varying levels of asymmetry. The duration_ms variable is significantly left skewed. However, the tail of the distribution is light, as the median falls near the the distribution's peak. We observe some univariate outliers. However, we choose to use classical estimators instead of their robust counterparts as the outliers lie within a reasonable range from the corresponding distribution's tails and the volume of regular observations is large.

Preliminary data analysis consisted of examining the pairwise scatter plots, univariate histograms and measures of correlation of the acoustic features data set $\mathbf{X}$. As a measure of correlation, Spearman rank correlation coefficient $\rho(X_i, X_j)$ was used. The estimator is defined as the Pearson correlation coefficient between the ranked variables and it is appropriate for both continuous and discrete ordinal variables, which suits the data at hand. Spearman rank correlation coefficient of the variables are depicted in Figure 2. The figure depicts low to medium pairwise correlations among the variables apart from $\rho(\text{loudness}, \text{energy}) = 0.74$ and $\rho(\text{acousticness}, \text{energy}) = -0.64$.

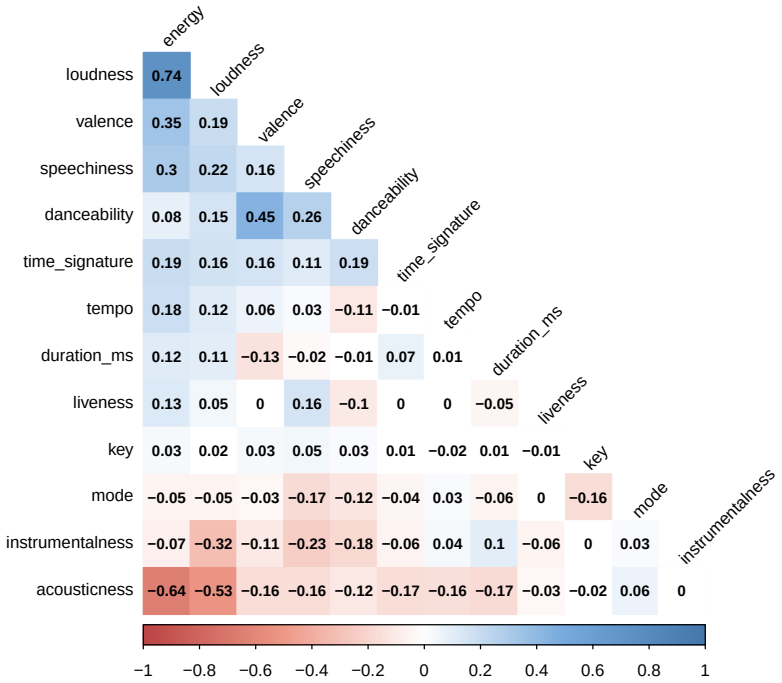

Figure 2: Spearman rank correlation coefficients $\rho(X_i, X_j)$. Correlations with p-value > 0.01 are considered as insignificant. In this case the corresponding cells are left white.

4 Hierarchical cluster analysis

The preliminary analysis revealed low to medium pairwise correlations among the majority of the variables. This suggests that PCA-based methods may be inappropriate for further analysis of the data variability. Regardless, we visualize the component scores of $\mathbf{X}$ corresponding to the first and second principal components to gain a better understanding of the multivariate nature of the data. We choose to ignore the categorical nature of the variables time signature, mode and key for the same purpose. The scores are depicted in Figure 3 and the loadings in Table 1.

The first and second component explain 0.23% and 0.11% of the variance in the data set $\mathbf{X}$, respectively. Table 1 depicts that the highly correlated variables loudness, acousticness and energy dominate the first PC, while the variables danceability and speechiness constitute most to the second PC. We observe no apparent cluster structures in the data.

To find common and differentiating factors within the data, we turn to hierarchical

Table 1: First and second principal component loadings of $\mathbf{X}$.

PC	aco.	dan.	dur.	ene.	ins.	key	liv.	lou.	mod.	spe.	tem.	tim.	val.
1st	0.46	-0.28	-0.11	-0.5	0.18	-0.04	-0.04	-0.48	0.08	-0.09	-0.13	-0.24	-0.31
2nd	0.19	0.5	-0.36	-0.19	-0.1	0.15	0.02	-0.19	-0.22	0.46	-0.29	0.06	0.36

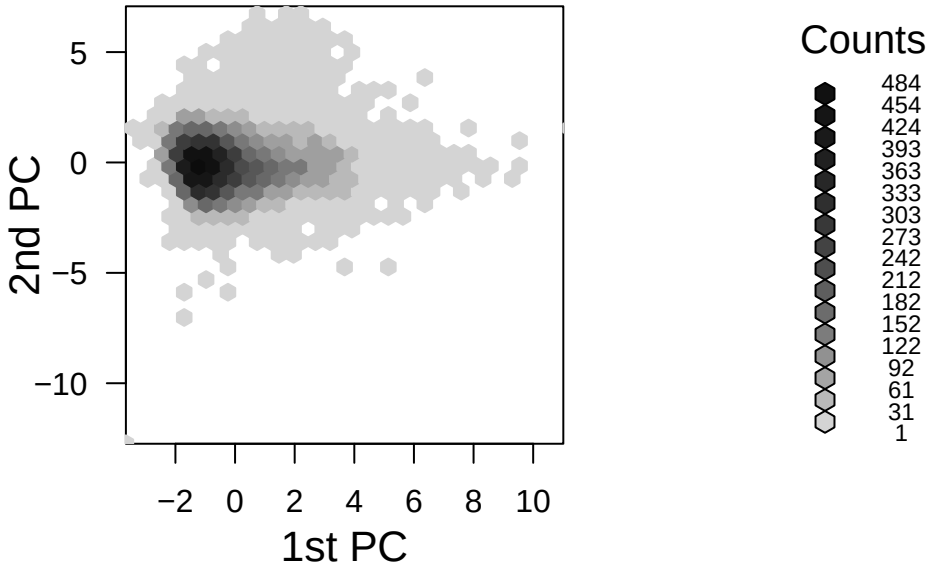

Figure 3: Hexbin scatter plot of the first and second principal component scores of the data set $\mathbf{X}$.

cluster analysis. Based on PCA, the data appears poorly separable. Regardless, it is in our interests to separate the data into regions of maximal similarity.

We apply Ward error sum of squares hierarchical clustering method, which finds clusters with minimum within-cluster variance, and maximum between-cluster variance. Ward's method is the only one among the agglomerative clustering methods that is based on a classical sum-of-squares criterion. This is also the criterion used in the widely used k-means clustering method, which unlike Ward's method requires knowing the number of clusters k.

For ease of interpretation, we limit the maximum number of clusters k to 20. To select the optimal k, we examine two validation measures for clustering results: Dunn index and average silhouette width. Both average silhouette width and Dunn Index combine measures of compactness and separation of the clusters. Dunn index is the ratio between the minimum inter-cluster distance to the maximum intra-cluster diameter. Dunn index ranges from zero to infinity, and in order to have well separated and compact clusters we aim to maximize Dunn index. Average silhouette width is the average of each observation's Silhouette value. Silhouette value measures the similarity of a point to its own cluster compared to other clusters. Silhouette value ranges from -1 to 1, and a value near 1 means that a point is similar to its own cluster and dissimilar to other clusters and vice versa.

Dunn indices and average silhouette widths for the number of clusters $k = 2, ..., 20$ are depicted in Figure 4. The average silhouette width is small for each k, which is expected as the data seems poorly separable. The optimal number of clusters appears to be between 10 and 12. Again, for ease of interpretation we select the smaller $k = 10$. Dendrogram of the data set $\mathbf{X}$ clustered into 10 classes using Ward's method is depicted in Figure 5. We denote the 10 clusters enumerated in Figure 5 by $\mathbf{C}_i,\ i = 1, ..., 10$. The cluster sizes and the sizes relative to the whole population size are depicted in Table 2. The clusters $\mathbf{C}_1$ and $\mathbf{C}_5$ together contain over 50% of the data.

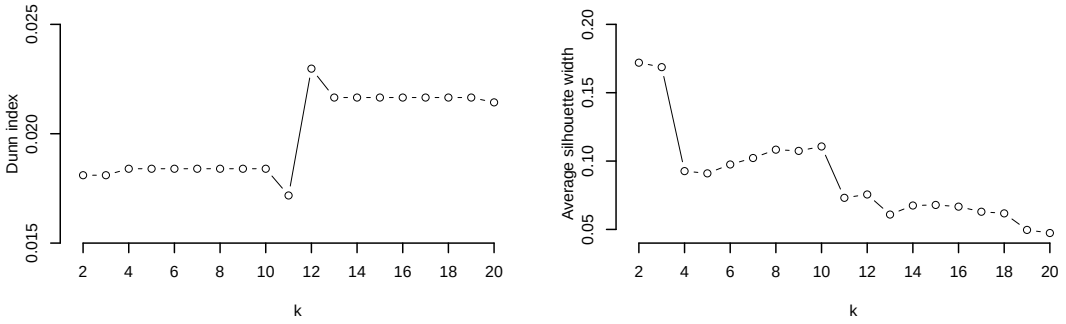

Figure 4: Dunn index and average silhouette width for the data set $\mathbf{X}$ clustered using Ward's method for the number of clusters $k = 2, ..., 20$.

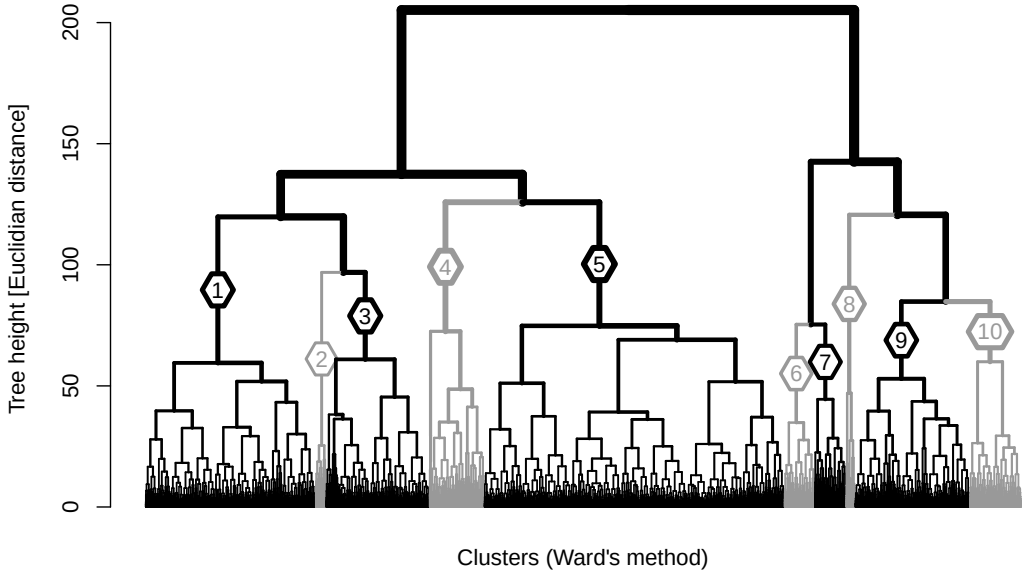

Figure 5: Dendrogram of the data set $\mathbf{X}$ clustered into 10 classes using Ward's method.

Table 2: Population sizes and relative sizes of clusters $\mathbf{C}_i,\ i = 1, ..., 10$.

Cluster	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_3$	$\mathbf{C}_4$	$\mathbf{C}_5$	$\mathbf{C}_6$	$\mathbf{C}_7$	$\mathbf{C}_8$	$\mathbf{C}_9$	$\mathbf{C}_{10}$
Size	2838	175	1725	915	5006	506	623	152	1910	857
Relative size	0.19	0.01	0.12	0.06	0.34	0.03	0.04	0.01	0.13	0.06

4.1 Interpretation of cluster analysis

We interpret the denrdrogram in Figure 5 as a decision tree, in order to describe which acoustic features separate and connect the number-one albums' content. Specifically, we start the analysis from the root of the dendrogram and compare the centroids of the two sets of clusters each node separates. This is justified, since the clustering is based on a classical sum-of-squares criterion. Let $\mathbf{C}_{\{i,j\}}$ denote the union of the clusters $\mathbf{C}_i$ and $\mathbf{C}_j$. Furtermore, we denote the L2-normalized difference of the cluster centroids of $\mathbf{C}_i$ and $\mathbf{C}_j$ by $\Delta(\mathbf{C}_i, \mathbf{C}_j)$ such that

$$\Delta(\mathbf{C}_i, \mathbf{C}_j) = \frac{\mu(\mathbf{C}_i) - \mu(\mathbf{C}_j)}{|\mu(\mathbf{C}_i) - \mu(\mathbf{C}_j)|_2},$$

where $\mu(\mathbf{C}_j)$ is the centroid of cluster $\mathbf{C}_j$. We interpret each branch of the dendrogram by inspecting the Δ-value at each node along the branch. For the purpose of demonstration, the normalized differences Δ of the dendrogram branch leading to clusters $\mathbf{C}_4$ and $\mathbf{C}_5$ are depicted in Table 3.

Table 3: The L2-normalized differences of the cluster centroids Δ of the dendrogram branch leading to clusters $\mathbf{C}_4$ and $\mathbf{C}_5$.

$\Delta(\mathbf{C}_{\{1,2,3,4,5\}}, \mathbf{C}_{\{6,7,8,9,10\}})$

aco.	dan.	dur.	ene.	ins.	key	liv.	lou.	mod.	spe.	tem.	tim.	val.
-0.25	0.16	-0.06	0.26	-0.32	0.01	-0.13	0.75	-0.09	0.14	0.09	0.09	-0.34

$\Delta(\mathbf{C}_{\{1,2,3\}}, \mathbf{C}_{\{4,5\}})$

aco.	dan.	dur.	ene.	ins.	key	liv.	lou.	mod.	spe.	tem.	tim.	val.
-0.07	0.11	-0.1	-0.21	-0.19	-0.13	0.34	-0.01	-0.22	0.29	0.11	0	-0.78

$\Delta(\mathbf{C}_4, \mathbf{C}_5)$

aco.	dan.	dur.	ene.	ins.	key	liv.	lou.	mod.	spe.	tem.	tim.	val.
0.24	-0.5	-0.09	-0.34	0.05	0.01	0.06	-0.54	0.21	-0.47	0.06	-0.05	-0.04

Table 3 shows that the cluster $\mathbf{C}_{\{1,2,3,4,5\}}$ is associated with high loudness value compared to the cluster $\mathbf{C}_{\{6,7,8,9,10\}}$. The "loud" cluster $\mathbf{C}_{\{1,2,3,4,5\}}$ further partitions into the clusters $\mathbf{C}_{\{1,2,3\}}$ and $\mathbf{C}_{\{4,5\}}$ associated with lower and higher valence, respectively. We refer to these clusters as "loud and negative" and "loud and positive", as tracks with low valence sound more negative (e.g. sad, depressed, angry), while tracks with high valence sound more positive (e.g. happy, cheerful, euphoric).

The "loud and posivitive" cluster $\mathbf{C}_{\{4,5\}}$ contains 40% of the of the original data. The majority of this content ($\mathbf{C}_5$) is associated with higher danceability, energy, loudness and speechiness values, while some of the songs ($\mathbf{C}_4$) are associated with higher acousticness. Speechiness detects the presence of spoken words in a track, so tracks that may contain both music and speech, either in sections or layered, include such cases as rap music.

Using the same approach, we interpret rest of Figure 5 dendrogram branches. The "loud and negative" cluster $\mathbf{C}_{\{1,2,3\}}$ contains 32% of the original data. Majority of this content ($\mathbf{C}_1$) is associated with higher danceability and speechiness, while the rest is associated with higher energy, ($\mathbf{C}_{\{2,3\}}$) and even liveness ($\mathbf{C}_2$).

The "quiet" cluster $\mathbf{C}_{\{6,7,8,9,10\}}$ partitions into "quiet and speechy" cluster $\mathbf{C}_{\{6,7\}}$, which contains both songs with higher danceability, energy, loudness and speechiness ($\mathbf{C}_6$), and more acoustic content ($\mathbf{C}_7$). The "quiet and non-speech-like" cluster $\mathbf{C}_{\{8,9,10\}}$ partitions into cluster $\mathbf{C}_8$ with shorter songs, the clusters $\mathbf{C}_9$ with longer, more danceable and energetic content, and $\mathbf{C}_{10}$ with longer, more acoustic and instrumentall content.

The interpretations of the cluster are depicted in Table 4 along with quartiles of the song release dates within each clusters. The three largest clusters are associated with high danceability and together contain songs mainly from '80s, '00s and '10s. We observe that an average best-selling song has become louder and shorter since the '60s and '70s. The 1990s mark a raise in increasing speechiness values, which is likely explained by the increase in popularity of rap music. The 1990s comes across as a decade of experimentality, since the hit songs of that era partition into many smaller clusters, some of which are short lived in therms of song release dates. Over the past decade, negativity has become more relevant than before.

Table 4: Interpretation of the clusters sorted by the decreasing cluster size, and quartiles of the song release dates within each clusters.

Cluster	Interpretation	Q_1	Q_2	Q_3
$\mathbf{C}_5$	Loud, positive, danceable, speechy	2004	2007	2010
$\mathbf{C}_1$	Loud, negative, danceable, speechy	2008	2016	2018
$\mathbf{C}_9$	Quiet, long, danceable	1979	1988	1992
$\mathbf{C}_3$	Loud, negative, energetic	2005	2014	2015
$\mathbf{C}_4$	Loud, positive, acoustic	1988	2012	2013
$\mathbf{C}_{10}$	Quiet, long, acoustic	1971	1972	1974
$\mathbf{C}_7$	Quiet, speechy, acoustic	1996	1996	1997
$\mathbf{C}_6$	Quiet, speechy, danceable	1998	1999	2014
$\mathbf{C}_2$	Loud, negative, energetic, live	1987	2010	2015
$\mathbf{C}_8$	Quiet, short	1995	1996	1996

5 Summary and conclusions

This study examined acoustic features of Billboard 200 number-one albums from 5.1.1963 to 19.1.2019 to discover common and differentiating factors. We applied Ward error sum of squares hierarchical clustering method and used Dunn index and average silhouette width to determine an optimal number of clusters. For ease of interpretation, we limited the maximum number of clusters to 20. The method produced a dendrogram partitioning the data into 12 cluster, which we interpreted by inspecting normalized differences of cluster centroids.

The results suggest that a large majority of the number-one albums' content since 1980s is more danceable compared to earlier hit albums characterized as acoustic. We observed that an average hit song has become louder and shorter since the 1970s. The number-one album variability appears the largest during the 1990s. Since that era, majority of number-one albums feature spoken words, presumable due to increasing popularity of rap music. While the majority of both 2000s and 2010s best selling albums are characterized as loud, danceable and speechy, the positivity (e.g. happiness, cheerfulness, euphoria) of 2000s, has been widely replaced by negativity (e.g. sadness, depression, anger) during the 2010s.

During the analysis, we chose to ignore the presence of outliers and treat categorical data variables as continuous. The outliers may have distorted the characteristics of certain smaller clusters, while the effect of the categorical variables on the cluster structure may have been dampened. Since the data appeared poorly separable, the eventual cluster structure may be highly sensitive to the choice of linkage criterion and the measure of distance. This could be verified for example by examining a

measure of association, such as Baker's Gamma Index, between two dendrograms produced using different methods.

References

[1] Andrew Thompson. Acoustic and meta features of albums and songs on the billboard 200. `https://components.one/datasets/billboard-200/`, apr 2019.