Assigment on dimension reducion

Rafał Rysiejko

12/23/2019

Introduction

The recorded music industry was worth $19.1 billion in 2018, which was almost a double-digit gain (9.7%) from the year prior[1]. In this entertainment market, an increasingly large number of song products are introduced each year. However, only a small group achieves mainstream success, and among these, the distribution of market success is becoming increasingly skewed. By analyzing the top songs from the popular music streaming service provider, we can investigate if some underlying patterns make a song truly hit.

This assignment will try to determine whether there are some distinctive features of song that make it successful and if so, are there any patterns among those features. For this, a set of statistical methods will be used, including dimensionality reduction techniques. Dataset[2] used for this assignment contains information about the top 50 most listened songs in the world on music streaming platform Spotify in 2019. Each song has a set of 13 variables:

Table 1: Variable description

Variable	Description
Track.Name	Song Title
Artist.Name	Artist performing the song
Genre	The genre of the track
Beats.Per.Minute	Variable describing the tempo of the song.
Energy	The energy of a song - the higher the value, the more energtic song.
Danceability	The higher the value, the easier it is to dance to this song.
Loudness..dB.	The higher the value, the louder the song.
Liveness	The higher the value, the more likely the song is a live recording.
Valence	The higher the value, the more positive mood for the song.
Length	The duration of the song.
Acousticness	The higher the value the more acoustic the song is.
Speechiness	The higher the value the more spoken word the song contains.
Popularity	The higher the value the more popular the song is.

Exploratory data analysis

Installing and running the libraries:

```
requiredPackages = c("tidyverse","factoextra","stats","clustertend","flexclust","ggforce"
,"fpc","cluster","ClusterR","knitr","kableExtra","DataExplorer","caret",
"reshape2","corrplot","labdsv","smacof","clusterSim","pastecs","psych","pca3d","pls")
for(i in requiredPackages){if(!require(i,character.only = TRUE)) install.packages(i)}
for(i in requiredPackages){library(i,character.only = TRUE) }
```

Loading the data:

[1] https://www.forbes.com/sites/hughmcintyre/2019/04/02/the-global-music-industry-hits-19-billion-in-sales-in-2018-jumping-by-almost-10/#52c7da3e18a9

[2] Dataset available at: https://www.kaggle.com/leonardopena/top50spotify2019/data

```
data_full <- read.csv("top50.csv",stringsAsFactors = F)
data_full <- data_full[,2:14]
data_init <- data_full
```

Table 2: Summary statistics of the dataset

	min	max	range	median	mean	SE.mean	CI.mean.0.95	var	std.dev
Beats.Per.Minute	85	190	105	104.5	120.06	4.37	8.78	954.71	30.90
Energy	32	88	56	66.5	64.06	2.01	4.04	202.55	14.23
Danceability	29	90	61	73.5	71.38	1.69	3.39	142.32	11.93
Loudness..dB..	-11	-2	9	-6.0	-5.66	0.29	0.58	4.23	2.06
Liveness	5	58	53	11.0	14.66	1.57	3.16	123.62	11.12
Valence.	10	95	85	55.5	54.60	3.16	6.35	498.90	22.34
Length.	115	309	194	198.0	200.96	5.54	11.12	1532.24	39.14
Acousticness..	1	75	74	15.0	22.16	2.69	5.40	360.83	19.00
Speechiness.	3	46	43	7.0	12.48	1.58	3.17	124.58	11.16
Popularity	70	95	25	88.0	87.50	0.64	1.28	20.17	4.49

Songs in the data set are grouped into 21 genres.

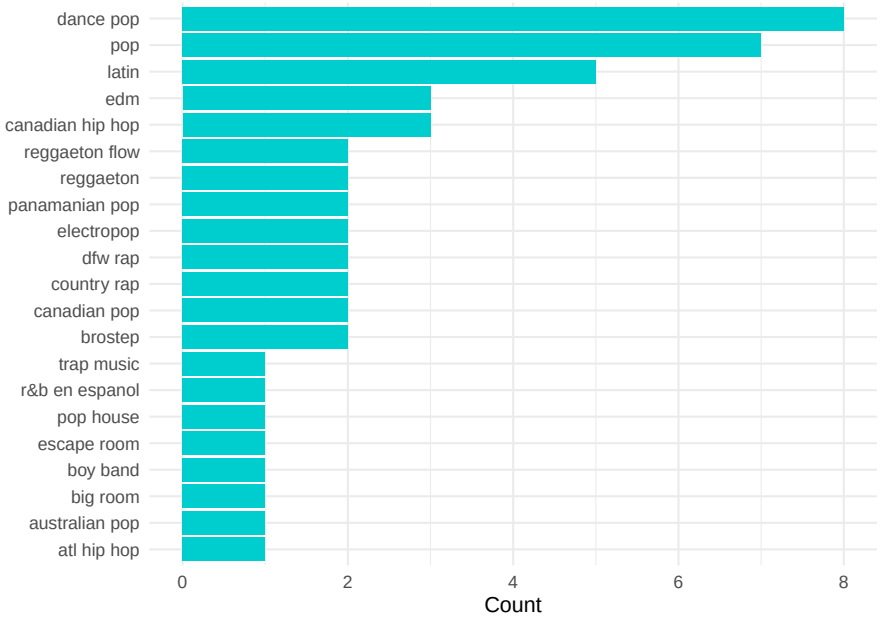

Figure 1: Genre distribution.

As there a lot of subgenres with a low number of instances, a new variable `aggregate.genres` is created.

```
data_init$aggregate.genre <- NA
data_init$aggregate.genre <- ifelse(data_init$Genre  %in% c("canadian pop","pop","dance pop",
                       "electropop","panamanian pop","pop house","australian pop"),"pop",
ifelse(data_init$Genre  %in% c("dfw rap","country rap","canadian hip hop","atl hip hop"),"hip hop",
ifelse(data_init$Genre  %in% c("r&b en espanol","latin"),"latin",
ifelse(data_init$Genre  %in% c("reggaeton flow","reggeaeton"),"reggaeton",
ifelse(data_init$Genre  %in% c("edm","trap music","brostep"),"electronic",
ifelse(data_init$Genre  %in% c("escape room","boy band","big room"),"other",data_init$Genre))))))
```

```
data_init$aggregate.genre <- as.factor(data_init$aggregate.genre)
```

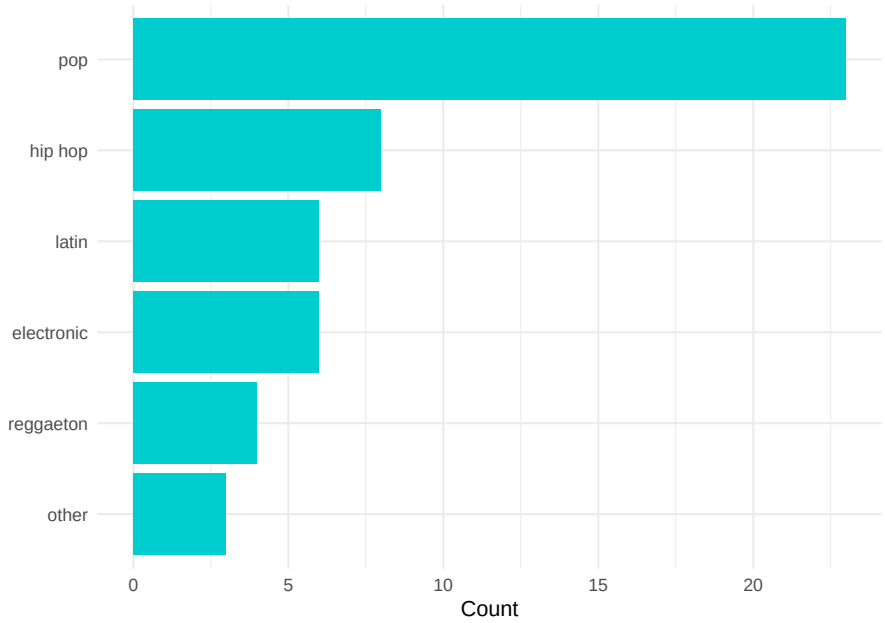

Figure 2: Aggregated genre distribution.

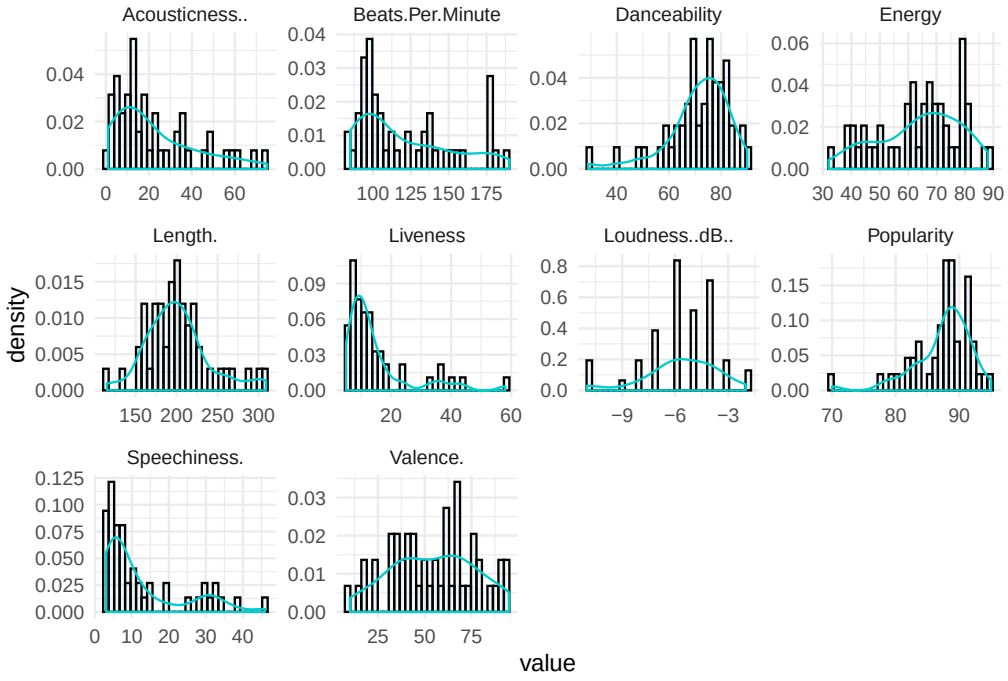

Figure 3: Histograms of numerical variables.

Variables `Liveness` and `Speechiness.` are highly positively skewed, suggesting the existence of outliers. A log transformation is applied to them as the PCA algorithm is sensitive to them.

```
transformed_var <- c("Liveness", "Speechiness.")
data_init <- data_init %>% mutate_at(vars(transformed_var), log)
```

Distribution of variables `Liveness` and `Speechiness.` after log transformation

To further investigate the relationship between variables, a correlation analysis was carried using a Pearson correlation coefficient.

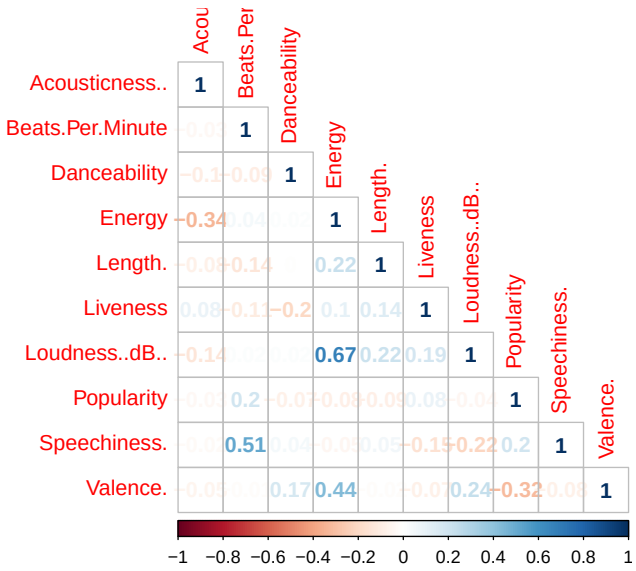

Figure 4: Correlation matrix.

From Figure 4, we can see a high positive correlation between variables `Speechiness` and `Beats.Per.Minute`, typical for rap songs. Another relationship is visible between variables `Energy` and `Loudness.dB.` suggesting that highly energetic songs tend to be louder. Also, a more energetic song, on average less acoustic.

Principal Component Analysis.

The data set is first split into train and test subsets for following PCA regression purposes with an 80% to 20% ratio.

```
smp_size <- floor(0.80 * nrow(data_init))

set.seed(123)
train_ind <- sample(seq_len(nrow(data_init)), size = smp_size)

data <- data_init[train_ind, ]
data_test <- data_init[-train_ind, ]
```

As different variables in the data set have different units of measurement, a normalization is required.

```
data_pca <- data_init %>%  dplyr::select(-Popularity) %>% select_if(is.numeric)
data_pca_stand<-data.Normalization(data_pca, type="n1", normalization="column")
```

Performing PCA

```
pca1<-prcomp(data_pca_stand, center=TRUE,scale. = TRUE)
x2 <- round(summary(pca1)$importance,2)
```

Summary of PCA

Table 3: Summary statistics of the PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.47	1.26	1.14	1.02	0.95	0.85	0.80	0.63	0.48
Proportion of Variance	0.24	0.18	0.15	0.12	0.10	0.08	0.07	0.04	0.03
Cumulative Proportion	0.24	0.42	0.56	0.68	0.78	0.86	0.93	0.97	1.00

Based on the summary in table 3, we see that 56 percent of the variance is contributed by the first three principal components. We can also see that there is no significant decrease in explained variance with each variable, suggesting that each of them contributes significantly to the explained variance.

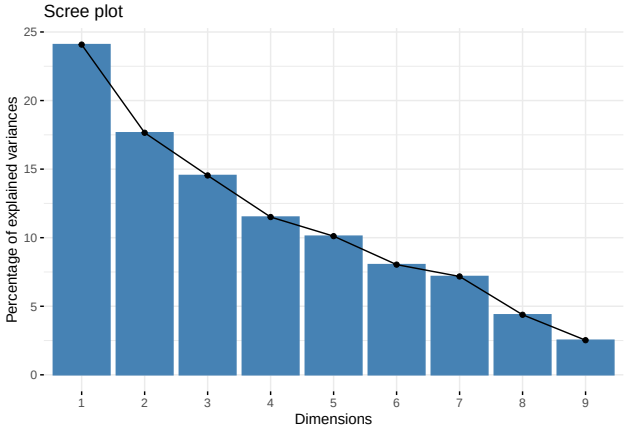

Figure 5: Scree Plot.

Table 4: Contribution of a variable to a given principal component

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Beats.Per.Minute	1.03	35.37	12.56	6.19	0.00	1.93	3.37	36.99	2.55
Energy	35.60	3.47	0.60	0.74	0.25	0.32	0.04	11.39	47.59
Danceability	0.43	3.05	35.24	8.72	11.58	35.83	3.21	0.74	1.21
Loudness..dB..	30.74	0.00	2.30	1.72	0.14	0.65	31.54	0.84	32.05
Liveness	2.41	13.35	23.20	0.53	1.02	43.49	15.82	0.18	0.00
Valence.	12.86	3.02	11.06	17.02	8.99	3.89	32.94	3.93	6.29
Length.	6.36	0.56	7.02	45.76	17.21	11.83	0.43	10.79	0.04
Acousticness..	6.75	6.25	1.00	15.95	54.49	1.56	8.16	1.81	4.03
Speechiness.	3.82	34.94	7.01	3.37	6.32	0.49	4.48	33.34	6.23

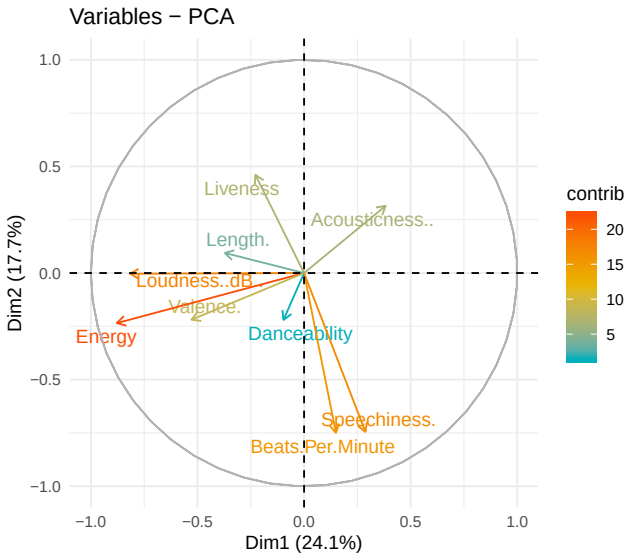

Figure 6: Variable contribution to Principal Component

Figure 6 represents the contribution of variables to the first two Principal Components. Positively correlated variables point to the same side of the plot. Negatively correlated variables point to opposite sides of the graph.

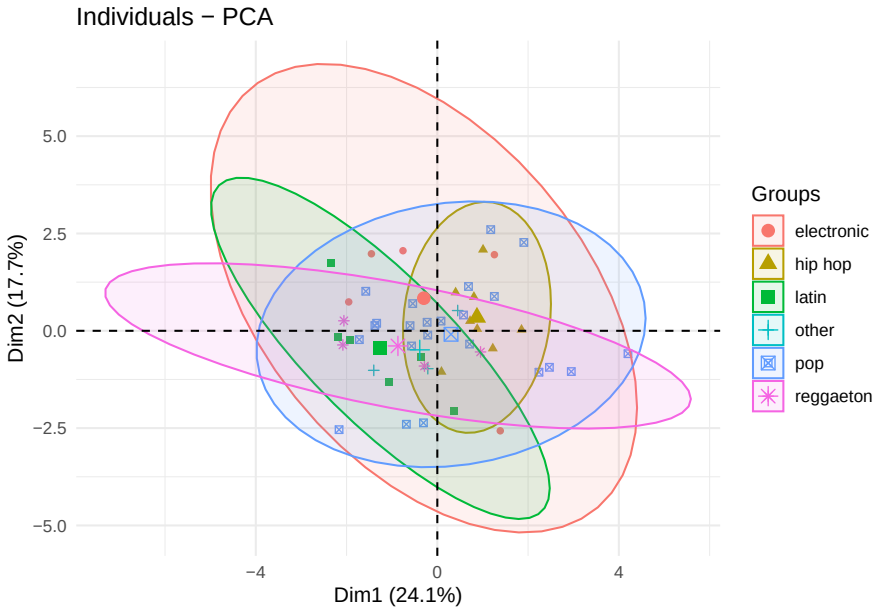

Figure 7: Graph of individual songs by aggregated genre.

Figure 7 represents the position of individual songs on the space of principal components, where songs with a similar profile are grouped together. We can see that the songs from the same genre tend to be placed near each other, except for `pop`, suggesting that this genre has high variability in analyzed features.

Rotated PCA using varimax rotation.

```
n = 3
pca_rot <- principal(data_pca_stand, nfactors=n, rotate="varimax")
x3 <- pca_rot
```

Table 5: Output 3

	RC1	RC2	RC3
Beats.Per.Minute	0.06	0.86	0.03
Energy	0.90	0.02	0.13
Danceability	0.19	-0.14	-0.68
Loudness..dB..	0.77	-0.12	0.29
Liveness	0.06	-0.21	0.72
Valence.	0.58	-0.11	-0.34
Length.	0.31	-0.04	0.37
Acousticness..	-0.46	-0.13	0.17
Speechiness.	-0.07	0.85	-0.09

When analyzing the rotated components, we can distinguish the following composite features:

- First rotated component RC1 – Loudness and Energy -> dance club readiness
- Second rotated component RC2 – Beats per minute and Speechiness -> lyrical rhythm
- Third rotated component RC3 – danceability and liveness -> calm and concert ready

Table 6: Quality measures of th PCA

	Complexity	Uniqueness
Beats.Per.Minute	1.01	0.25
Energy	1.04	0.17
Danceability	1.24	0.48
Loudness..dB..	1.33	0.30
Liveness	1.18	0.43
Valence.	1.70	0.53
Length.	1.97	0.76
Acousticness..	1.42	0.74
Speechiness.	1.03	0.27

Table 8 represents two quality measures *Complexity*, which indicates how many variables constitute a single factor. High complexity is an undesirable feature because it implies a more difficult interpretation of factors. *Uniqueness*, on the other hand, explains the proportion of variance that is not shared with other variables, meaning that variable with lower value carries less additional information in relation to other variables in the model.

Principal components regression

```
data_pca_regression <- data.Normalization(data_init %>% select_if(is.numeric),
type="n1", normalization="column")

reg.pcr.train<- pcr(Popularity~., data=data_pca_regression, scale=TRUE,
                    center = TRUE,validation="CV",ncomp = 3)

data_test_regression <- data.Normalization(data_test
```

```
      %>% select_if(is.numeric), type="n1", normalization="column")

reg.pcr.pred<-predict(reg.pcr.train, data_test_regression, ncomp = 3)
```

```
summary(reg.pcr.train)
```

```
## Data:    X dimension: 50 9
##  Y dimension: 50 1
## Fit method: svdpc
## Number of components considered: 3
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps
## CV            1.01    1.011    1.006    1.012
## adjCV         1.01    1.008    1.007    1.007
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps
## X            24.080   41.731    56.27
## Popularity    2.873    4.049     9.67
```

The summary provides the percentage of variance explained in the predictors and the outcome (Popularity) using different numbers of components For example, 56.27%% of the variation (or information) contained in the predictors is captured by three principal components (ncomp = 3). Additionally, setting ncomp = 3 captures only 9.67% of the information in the outcome variable (Popularity), which is low.

Model performance metrics

```
d1 <- data.frame(
  RMSE = caret::RMSE(reg.pcr.pred, data_test_regression$Popularity),
  Rsquare = caret::R2(reg.pcr.pred, data_test_regression$Popularity))
```

Table 7: Quality measures of the prediction using PCA

RMSE	R - Squared
0.93	0.04

Performance metrics contained in Table 7indicates that used independent variables are not explaining much in the variation of the dependent variable.

Conclusion

From the performed analysis, we concluded that over 50% of the variation in the data, and as that there is no significant drop in explained variance with each additional variable. Performing a Principal Component Regression on this problem revealed that the Popularity of songs is not strongly dependent on used independent variables. Further research on this topic could benefit from including more samples and from adding data from different periods to assess the importance of varying loading to principal components in different periods.