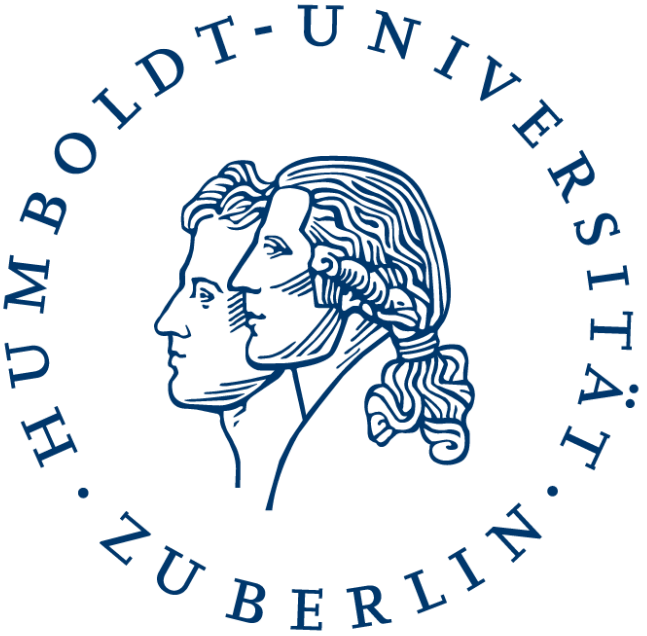
Humboldt-University Berlin

School of Business and Economics
Module: Datenanalyse II
Examiner: Dr. Sigbert Klinke
Winter term 2019/20



Assignment: Datenanalyse II

A multivariate analysis of music streaming

Gerome Wolf
Student-ID: 577552
E-Mail: wolfgerome@gmail.com
Programme: Economics (Master)
Date: 26th March 2020

Contents

List of Figures

List of Tables

Chapter 1

Problem statement and data collection

What is music? According to the Oxford Dictionary of English (Stevenson, 2010) it is "vocal or instrumental sounds (or both) combined in such a way as to produce beauty of form, harmony, and expression of emotion". Apparently, music is a medium that is characterized by some degree of structure that allows to capture and communicate emotions and can be utilized to regulate internal and external tensions within an individual or a social group. By definition, "emotional" and "rational" are opposing concepts, although music has the property of integrating aspects of mathematics (music theory), physics (acoustics), biology (neurology) and experiences of the composer all together (Fisher, 1929). Therefore, this makes it a fascinating object to analyze in detail and to uncover patterns that can generally describe some of the aforementioned characteristics.

Economic implications of music streaming

Recently, with the adoption of multifunctional devices such as smartphones and an increasing number of users who are connected to the internet, music streaming platforms have become major suppliers of music whereas traditional, local mediums such as vinyl, CD and MP3 players have grown out of fashion. Also, as users reveal their preferences through choosing their utility-maximizing bundles recommendations have become a core selling proposition and revenue generating service of music streaming platforms. The more accurate the recommendations are the more music is being consumed and the less

likely the customer is to cancel the subscription.

Since music consumption reflects the consumers' preferences and (partially) their emotional states, say being rather optimistic or pessimistic, the Bank of England has recently started to incorporate music consumption behaviour (based on lyrics) in their sentiment analyses from which, according to Haldane (2018), former Chief Economist, the resulting index of sentiment does at least as well in tracking consumer spending as the Michigan survey of consumer confidence.

The compensation schemes for music streams vary significantly across platforms as do their operating costs, target groups and enterprise structures (e.g. Youtube Music vs. Spotify) but according to a recent study out of seven music streaming services an artist needed on average about 373 streams to generate one Euro (Beat, 2020). As profit maximization is also an objective of the music industry and music labels as well as artists are assumed to be risk averse there is evidence (Thompson and Daniels, 2018) that particular hit songs (or one-hit-wonders) have converged on a style, meaning to exhibit an overall narrower range of sounds. "The storied, solitary figure working out musical problems at a piano while filling up an ashtray has been replaced by teams of digital production specialists and subspecialists, each assigned to a snare track, a bass track, and so on, mixed and matched and stuck together like Legos." (Thompson and Daniels, 2018).

In 2018 global recorded music revenues reached 19.1 billion USD with revenues from subscription based audio streams accounting for about half of total revenues (37% paid vs. 10% ad-supported), comprising 255 million users of paid subscription streaming accounts (International Federation of the Phonographic Industry, 2019).

A sample of unique songs was obtained from scraping the Spotify Charts website for Germany for a period spanning the previous 12 months on a daily basis which provided track title, artist name, number of total streams per day, its rank among the absolute top 200 on each day, the tracks' 22-character ID as used for further data augmentation and implicitly a label that this track appeared at least once among the most streamed tracks. 50 songs from the most popular playlists on Spotify of the genres Techno and classical music each were added such that the sample consists of 50 tracks per genre from (Pop, Hip Hop, Techno and classical music). Pop and Hip Hop tracks were sufficiently covered in the Charts sample already. Next, using the Python library for the Spotify Web API Spotipy and constraining all following methods to availability on the German

market track specific variables were obtained. In total, there were four genres left (the overwhelming majority of observations from the Charts sample was comprised of Hip Hop and Hop, cumulatively about 75%).
Most important for this study are the eight so-called audio features developed by Jehan (2005) that describe the psychoacoustical characteristics of any audio recording in quantitative terms (see table 1.1) and were obtained via the *audio features* method.
To add data from outside the Spotify API and to assess a (free) substitute platform for music listening metrics from the Youtube Data API were obtained.

The dataset ("DA2_Dataset.csv") can be found along with an R markdown file (Notebook_Submission.Rmd) in the GitHub repository: https://github.com/gerwolf/DA2_Homework

A rendered version of the Markdown notebook can be found on https://rpubs.com/gerwolf/da2_assignment_spotify

Table (1.1) Audio features

	Description
`acousticness`	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
`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.
`energy`	A measure from 0.0 to 1.0 and represents a perceptual 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. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
`instrumentalness`	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
`liveness`	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
`loudness`	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
`speechiness`	Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
`valence`	A measure from 0.0 to 1.0 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).

Chapter 2

Descriptive and bivariate statistics

2.1 Descriptive statistics

This section is to describe the statistical moments and properties of the variables of interest and is just provided for completeness and a quick overview of the data set.

Base data

The raw data as such consists mainly of continuous variables which are represented in histograms and kernel density plots in figure 9.1 and in Box plots in figure 9.2 in the appendix. Only `Artist_Popularity` and `Track_Popularity` seem to be approximately normally distributed which is important as the further analysis will revolve around these two variables as proxies for stream count.
A formal test on normality, the Kolmogorov-Smirnov test (KS-Test), finds that only `Artist_Popularity` is approximately normally distributed (with a test statistic $D = 0.0579 < \frac{1.3581}{\sqrt{196}} = D_{n=196,\alpha=0.05}$.

Standardized data

Standardizing the data gives the following histograms and Box plots, figures 9.3 and 9.4, respectively.
Again, testing for normality with the KS-Test gives that `Artist_Popularity` is approx-

imately normally distributed.

Box-Cox transformation

Selecting only those continuous variables which are strictly positive leaves 17 variables to transform and test for. Using the optimized values for λ and conducting the KS-Test once more gives that now `Artist_Follower`, `Artist_Popularity`, `Artist_Singles_Number`, `Artist_Singles_Tracks_Number`, `loudness`, `speechiness`, `tempo`, `valence`, `likeCount` and `viewsCount` are approximately normally distributed. The corresponding histograms and Box plots are in shown in figures 9.5 and 9.6, respectively.

2.2 Correlation

Only those coefficients which passed Steiger's Z test, also despite violations of jointly normal distribution, are displayed with a coloured field in figures 9.7 (Pearson) and 9.8 (outlier-robust measure, Spearman). As can be seen from comparing both correlograms using the Spearman rank correlation coefficient produces more statistically significant pairs than the Bravais-Pearson correlation coefficient. Variables describing artist output (the block from albums, singles, appearances and compilations) appear to be alltogether positively correlated as well as this block is negatively correlated with `danceability`, `energy` and `loudness`. `acousticness` and `energy` are negatively correlated, indicating opposing components of music compositions. The variables from the Youtube block (`viewsCount`, `dislikeCount`, `likeCount` and `commentCount`) are alltogether positively correlated as this block is with `Artist_Popularity` and `Track_Popularity`.

2.2.1 Steiger's Z test

When conducting Steiger's Z test the assumption for the test statistic $T = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}}$ (where R_{xy} is a given correlation coefficient between to RV x and y) under the Null hypothesis to be t distributed with $n-2$ degrees of freedom is that the two variables are jointly normally distributed.
We're particularly interested in the magnitude and relevance of the relationship be-

tween `Artist_Popularity` and `Artist_Follower`. Figure 2.1 below shows the effects of transforming the variables against the untransformed variables.

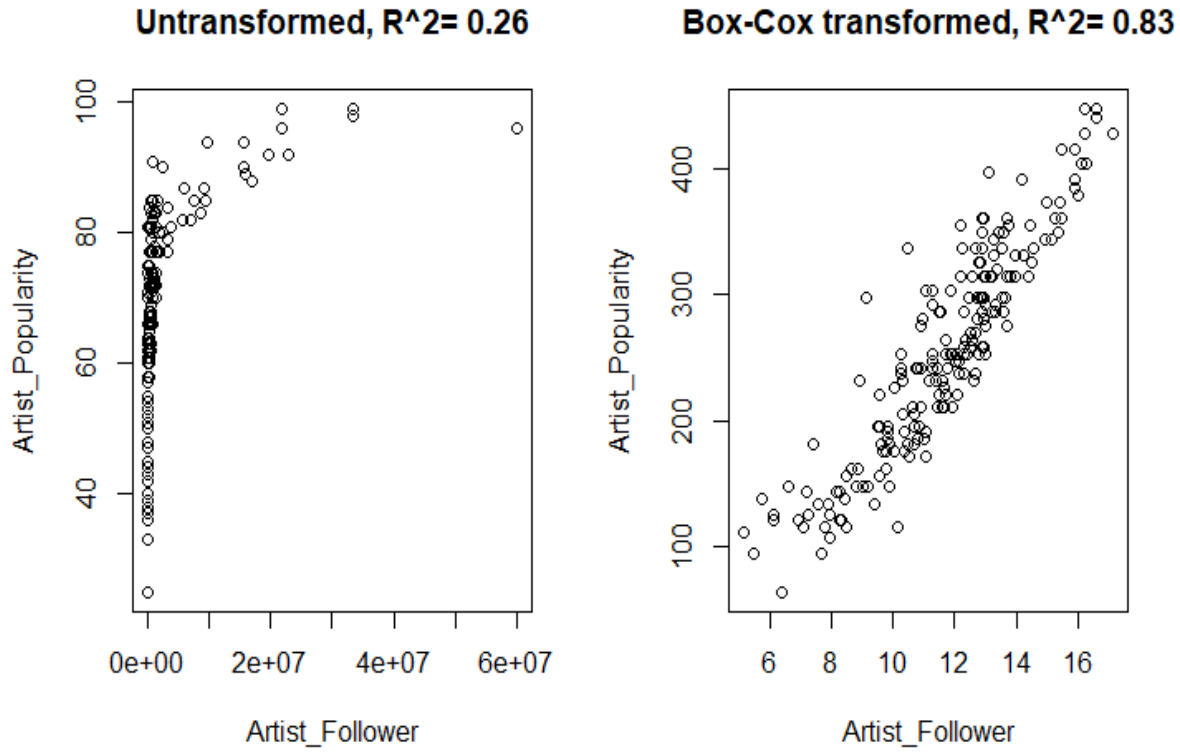


Figure (2.1) Scatter plots (base variables LHS, Box-Cox transformed variables RHS)

Recall that in the base specification only `Artist_Popularity` is approximately normally distributed, `Artist_Follower` is not. Testing for joint normality using the *MVN* package from `R` gives that the two variables are (approx.) jointly normally distributed according to the Energy statistic (Székely and Rizzo, 2017). Since we have verified that in the base specification `Artist_Popularity` and `Artist_Follower` are jointly normally distributed we can conduct Steiger's Z test and accept the test decision as being accurate. Looking at the entry for the two variables in figure 9.7 shows a correlation coefficient of 0.51 which did also pass Steiger's Z test. Looking at the entry for the two variables in figure 9.8 the rank correlation coefficient increases to 0.91 and did also pass Steiger's Z test.

2.3 Visual data analysis

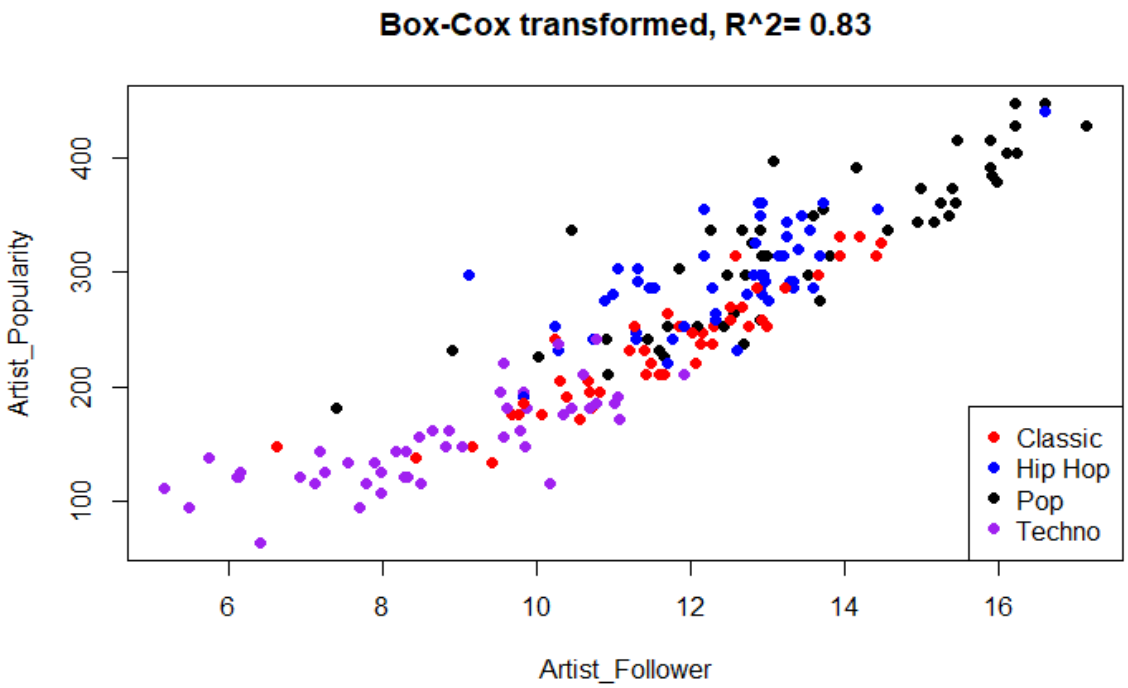
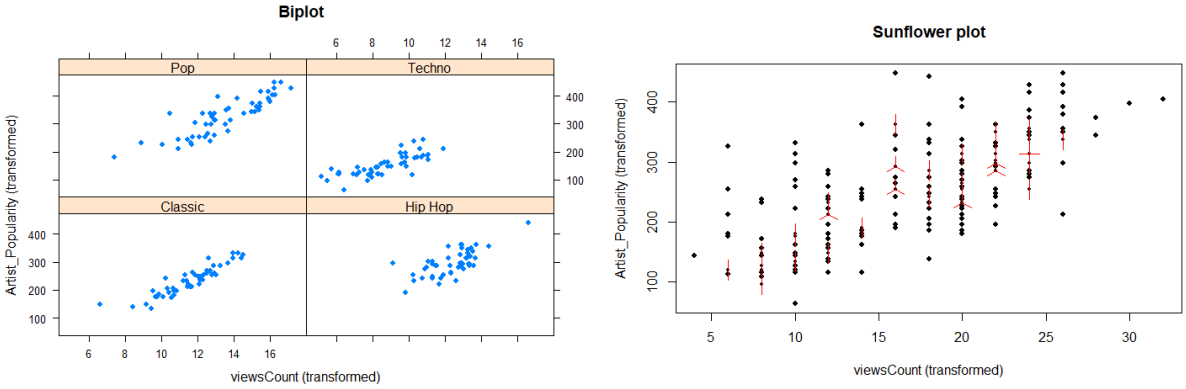


Figure (2.2) Scatter plot (colours = genres)

From figure 2.2 above it is clear that the four music genres share the same relationship between `Artist_Follower` and `Artist_Popularity` in terms of direction but differ in terms of magnitude. The right panel of figure 2.3 shows a sunflower plot which is particularly useful if overplotting points hide a significant density around spots. Different shapes indicate different degrees of concentration at these affected spots.

Figure (2.3) Bivariate graphics



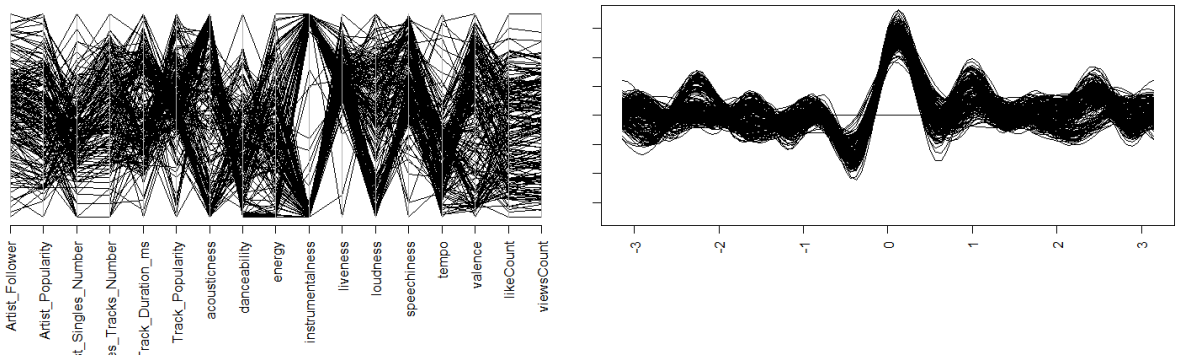
(a) Biplot

(b) Sunflower plot

A pairs plot for the Box-Cox transformed variables can be found in figure 9.9 in the appendix.

One possible way is to express every observation as a line resulting in a parallel coordinates plot (figure 2.5, left panel). A modification of parallel coordinates are Andrews curves which allows to identify outliers and clusters in the data (figure 2.5, right panel).

Figure (2.5) Nonorthogonal projections (all continuous variables)

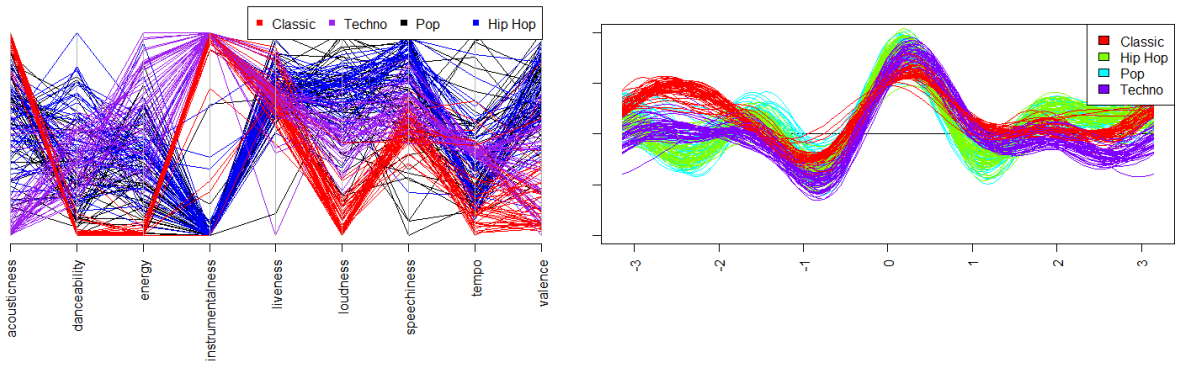


(a) Parallel coordinates plot

(b) Andrews curves

Assuming structure in the audio features it is sensible to analyze these variables in isolation and color them according to their genres (figure 2.7 below).

Figure (2.7) Nonorthogonal projections (audio features)



(a) Parallel coordinates plot

(b) Andrews curves

Clearly, both plots indicate clusters in the data that can be meaningfully attributed to the four different genre populations the samples were drawn from.
Scagnostics returns a group of outlying plots together with a group of exemplars, i.e. those plots that exhibit a high degree of similarity. In short, all variables from the group "artist output / supply" seem to be outliers in the sense that they are closely related (e.g. `Artist_Compilations_Tracks_Number` and `Artist_Compilations_Number` have a $\rho = 0.94$ meaning that per compilation an artist appeared at most once in it which makes sense considering the desired diversity on these compilations/playlists). Also, the relationship between `Artist_Follower` and `Artist_Popularity` highlighted in figure 2.2 but this time in the base specification has been detected as an outlying one. Interesting relationships between (`Track_Duration_ms`, `instrumentalness`) and (`instrumentalness`, `valence`) have been found, indicating that shorter tracks are less instrumental and very instrumental tracks have lower valence, meaning they sound more negative/sad/depressed/angry (see figure 2.9 below).

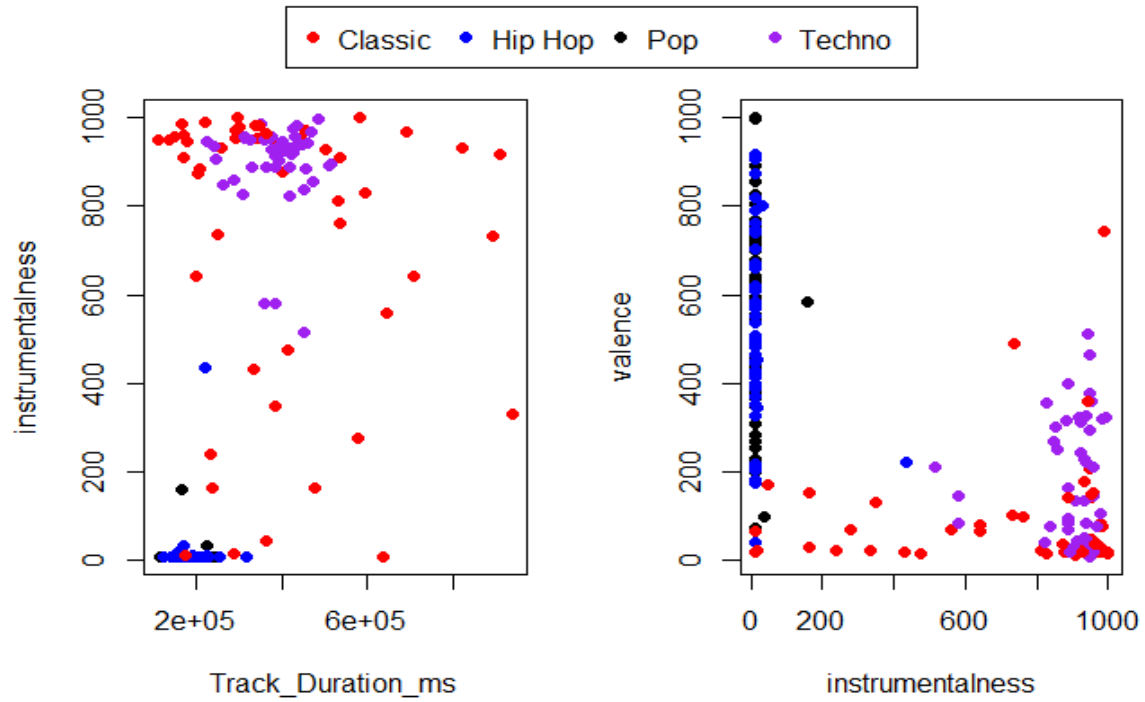


Figure (2.9) Scagnostics (outlier)

There were four pairs found to be exemplary, that is (`Artist_Appearances_Tracks_Number`, `Track_Popularity`), (`Artist_Albums_Tracks_Number`, `energy`), (`Artist_Appearances_Number`, `likeCount`) and (`tempo`, `viewsCount`) with (`x`, `y`), respectively.

2.4 Discrete variables

As there are almost no categorical variables in the data set they were generated indicating the 1^{st} (bottom), 2^{nd}, 3^{rd} and 4^{th} (top) quantiles of `Artist_Popularity`.

2.4.1 χ^2-Test of independence

	1st	2nd	3rd	4th	Total
Classic	9	21	13	6	49
Hip Hop	0	8	25	17	50
Pop	1	8	10	31	50
Techno	37	10	0	0	47
Total	47	47	48	54	196

Table (2.1) Contingency table for quantiles of `Artist_Popularity` by `Genre`

A formal χ^2-Test of independence rejects the Null hypothesis of independence ($\chi^2 = 156.23 > \chi^2_{df=9,\alpha=0.05} = 16.92$) and we can conclude that there exists a relationship. Cramer's $V = 0.52$ indicates that the relationship is strong.

2.4.2 Kendall's Tau

		Quantiles of `Artist_Popularity`				
		1st	2nd	3rd	4th	Total
Quantiles of `viewsCount`	1st	36	6	4	3	49
	2nd	10	19	12	8	49
	3rd	1	18	19	11	49
	4th	0	4	13	32	49
	Total	47	47	48	54	196

Table (2.2) Contingency table for quantiles

Again, a formal χ^2-Test of independence rejects the Null hypothesis of independence ($\chi^2 = 133.34 > \chi^2_{df=9,\alpha=0.05} = 16.92$) and we can conclude that there exists a relationship between the placement of an artist's music video on Youtube within the distribution and the placement of its corresponding artist within Spotify's popularity distribution. Cramer's $V = 0.48$ indicates that the relationship is semi-strong. The rank correlation coefficient by Kendall's Tau (here for a quadratic table) is around 0.59 and the test yields that this coefficient is significant, hence we can conclude that there exists a positive

relationship between the variables, meaning a higher rank of the song's music video in Youtube views is associated with a higher rank inside of Spotify's artists.

2.4.3 Proportional Reduction of Error (PRE) measures

From table 2.1 we know that the mode of of `Artist_Popularity` is in the top quantile (54), so the best prediction for a new artist without further knowledge is that she will fall in the top quantile of the popularity distribution. Without any knowledge about an additional feature other than artist popularity quantile itself and using the mode across classes as best predictor for class assignment for a new observation, the number of falsely predicted cases would be 142 or 72.45%.
Now having knowledge about an association between quantile of `Artist_Popularity` and its `Genre` and using class-internal modes for the additional feature, the number of falsely predicted cases would be 82 or 41.84% which is already much lower compared to the error rate from predicting without any knowledge about an association. On the other hand, 114 or 58.16% would have been predicted correctly (compared to 27.55% if there was no knowledge about genre).
Goodman and Kruskals λ is 0.42 which indicates that the improvement in predictability is substantial, resulting in an increased accuracy of predictions by more than 30%.
Turning to the second case the corresponding contingency table has been shown in table 2.2. Goodman and Kruskals λ is 0.37. Comparing both possible predictors for quantile of `Artist_Popularity` we can see that the PRE measures are both substantial but `Genre` seems to be a better predictor for quantile of `Artist_Popularity` than quantile of `viewsCount` on their own.

Chapter 3

Principal component analysis

Principal Component Analysis (PCA) aims to reduce the number of dimensions (or variables) in a dataset by expressing the original data points as linear combination(s) of the variables, using some coefficients to project the observed data points. Contrary to OLS one can express the criterion of the sum of squared differences between the projected data points *with reference to the origin* to be *maximized*, although both criteria are equivalent. Maximizing this L2 norm is equivalent to maximizing the variance of the projections as the data have been mean-centered. Now if an eigenvalue is equal to 1 that means it explains as much variance as a single variable (or the construct of multiple variables) since all variables have been scaled to mean 0 and unit-variance and it should be discarded.

3.1 Single value decomposition

From the fundamental property of eigenvectors

$$Av = \lambda v \tag{3.1}$$

where A is a nxn square matrix, v the (right) eigenvector and λ the eigenvalue corresponding to v. In words, any linear transformation applied to A doesn't change the direction of an eigenvector v, it only alters its magnitude in terms of the corresponding eigenvalue λ by a scalar. Generalizing for a set of eigenvectors, Q, which is a square nxn

matrix whose ith column is the eigenvector $v_i = q_i$ of A and Λ is a diagonal nxn matrix containing the ith eigenvalue λ on the diagonal element Λ_{ii} we get

$$AQ = Q\Lambda \tag{3.2}$$

$$\Leftrightarrow \quad A = Q\Lambda Q^{-1} \tag{3.3}$$

$$\Leftrightarrow \quad A = Q\Lambda Q' \tag{3.4}$$

with the property of an orthogonal matrix $Q'Q = I \Leftrightarrow Q^{-1} = Q'$ and such factorization holds for matrix A if it is diagonalizable which is the case if Q is invertible (or alternatively if $Q^{-1}AQ$ is a diagonal matrix) and this in turn is only possible iff (n) eigenvectors of A are linearly independent. This is always the case for eigenvectors corresponding to distinct eigenvalues of a symmetric matrix, also said that a set of eigenvectors forms a basis of the domain A (eigenbasis).

3.2 Scores and model fitting

The objective of this section is to find a model which projects the multivariate data points of audio features ($k = 9$) to a lower dimensional space capable of retaining a sufficient amount of variation from the original, full-feature space.

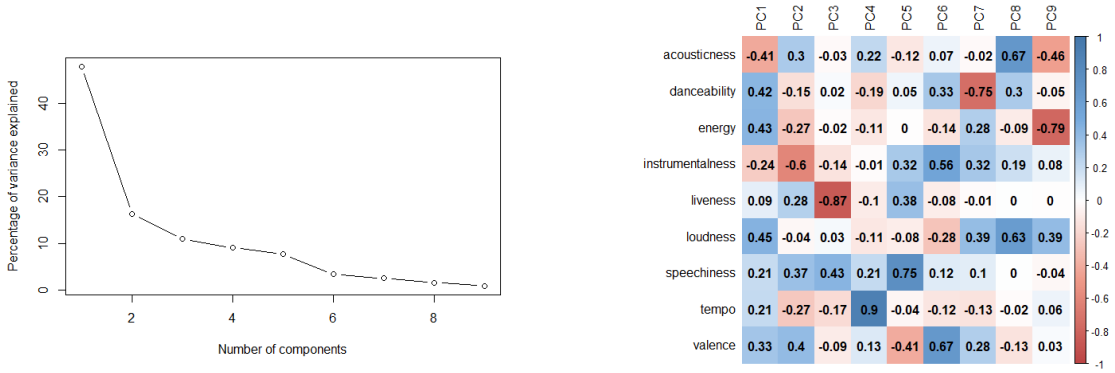
PC	Eigen value	Variance explained (%)	Cumulative variance explained (%)
1	4.31	48	48
2	1.46	16	64
3	0.98	11	75
4	0.81	9	84
5	0.69	8	92
6	0.30	3	95
7	0.22	2	98
8	0.14	2	99
9	0.08	1	100

Table (3.1) Eigenvalues

The largest eigenvalue is 4.31, meaning that the associated component explains more of

the variance than four of the original variables and there are only two eigenvalues which are greater than 1, thus two components should be chosen according to the Kaiser criterion. From the loadings plot in the right panel from figure 3.1 below we can see that the variables `acousticness` and `instrumentalness`, two distinct characteristics of classical music load highly negatively on the first component whereas `loudness`, `energy` and `danceability` load highly positively on the first component.
Within the second component `valence` and `speechiness`, distinct shared characteristics between Pop music and Hip Hop music, load highly positively and `instrumentalness` and `tempo` load both highly negatively. As we will see, the first component can clearly differentiate between classical music and Techno + Pop + Hip Hop music together. The second component can differentiate between Techno and Pop + Hip Hop music together.

Figure (3.1) PCA



(a) Scree plot (percentage of variance explained) (b) Loadings / normalized eigenvectors plot

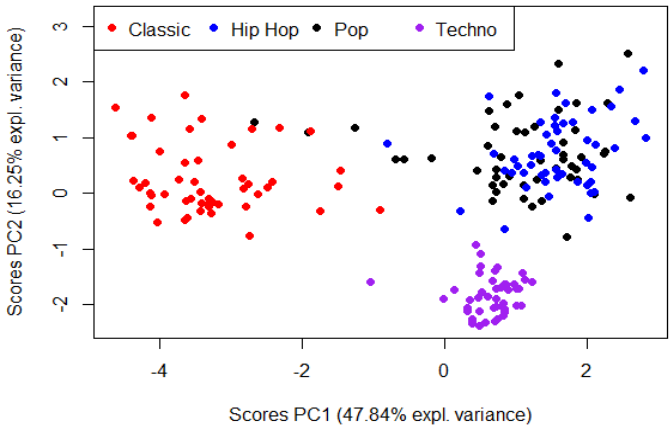


Figure (3.3) Score plot

Figure 9.10 shows the RMSE and R^2 for different numbers of components.

3.3 Q-residuals / Squared prediction errors

In order to assess the goodness of fit for a given number of components k we analyze the residuals since they represent the magnitude of variation that remains in each sample after projection through the model. In contrast to linear regression where these residuals are scalars that are to be minimized by the OLS criterion the PCA residuals are vectors for each observation. The Q-residual for observation i is computed in the following way:

$$Q_i = e_i e_i' = z_i(I - Q_k Q_k')z_i' \tag{3.5}$$

where z_i is the ith row of the standardized data matrix Z and Q_k is the subset of the projection matrix containing the k first eigenvectors.
Box (1954) showed that the squared prediction errors (SPEs) are well approximated by a weighted chi-squared distribution ($g\chi_h^2$) where the weight (g) and the degrees of freedom (h) are both functions of the eigenvalues of A (Nomikos and MacGregor, 1995). This approximation was adopted and was used as well as matching moments to obtain the weight (g) in order to compute $1 - \alpha$ control limits for the SPEs. The critical value for a sample of SPEs is given by

$$SPE_\alpha = \left(\frac{v}{2m}\right)\chi^2_{df=\frac{2m^2}{v},\alpha} \tag{3.6}$$

where v is the sample variance of the SPEs and m is the sample mean of the SPEs.

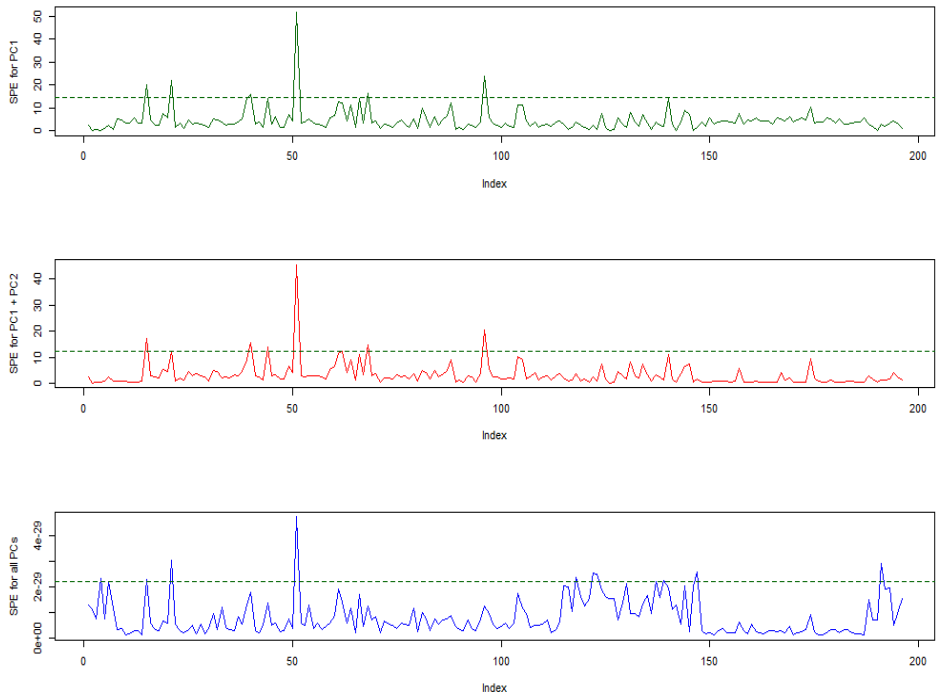


Figure (3.4) Squared Prediction Errors and control limits for different numbers of components

The SPE plots in figure 3.4 above show several things:

1) The more components are added to the model the smaller the SPEs become. The bottom panel shows the SPEs obtained when fitting the data to all available components, thus the displayed errors are marginal and are equal to 0.

2) When increasing the number of components for fitting from one to two (top and middle panels) the errors remain almost constant up to a scalar factor, highlighting certain observations that are not consistent with the model.

3) Observations that are not consistent with the model, i.e. those cases that cannot be sufficiently explained by the chosen number of components, lie above the 95% limit (the dashed lines). Naturally, we would expect 95% of the SPEs to fall below the control limit and only 5% above it. For the model with number of components $k = 1$ this is the case for only six observations (or 3%), for $k = 2$ seven observations (or 3.6%) and for the full "model" where $k = 9$ ten observations (or 5%) but for

this case the errors are meaningless.

To summarize, the residual analysis suggests that two components are indeed an appropriate reduction of dimensionality as there are no significant violations of model consistency.

3.4 Hotelling's T^2

In contrast to the Q-residuals Hotelling's T^2 values represent a measure of the variation in each sample within the model and are conceptually related to leverage which will be discussed in section 6.3 on outlier detection in linear regression. Hotelling's T^2 for observation i is computed according to

$$T_i^2 = t_i \lambda^{-1} t_i' = z_i Q_k \lambda^{-1} Q_k' z_i' \tag{3.7}$$

where t_i corresponds to the scores columns of observation i up to k and λ^{-1} is the inverted diagonal matrix with eigenvalues up to k. The expression is also known as Mahalanobis distance (Mahalanobis, 1936) for $k > 1$.
The test statistic (Hotelling's T^2) is distributed according to the F-distribution with k degrees of freedom in the numerator and $N - k$ degrees of freedom in the denominator. The control limit is computed according to

$$T_{k,\alpha}^2 = \frac{(N-1)(N+1)k}{N(N-k)} F_\alpha(k, N-k) \tag{3.8}$$

which is the equation for an ellipsis when $k = 2$ (Dunn, 2015). The Hotelling's T^2 values can be visualised in a sequence plot which is particulary useful if the row order in the dataset has a meaning (e.g. time-series).

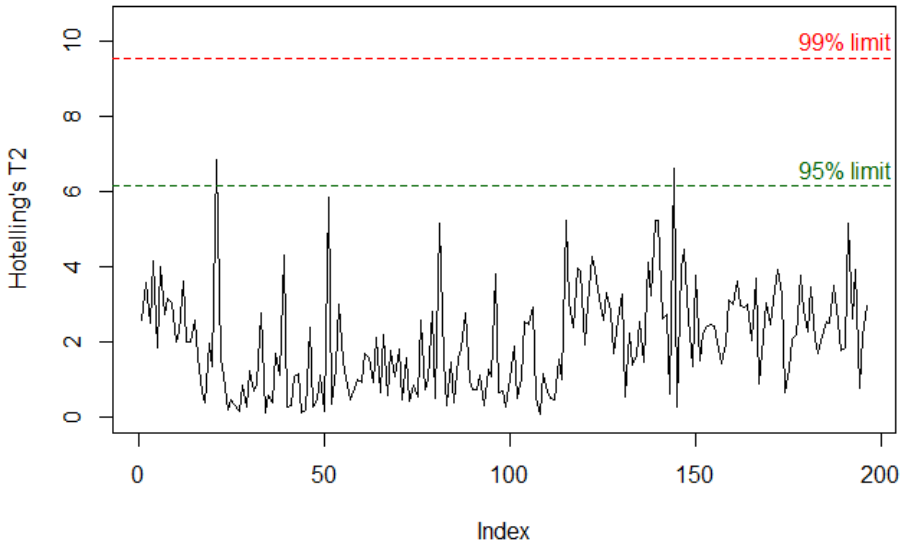


Figure (3.5) Sequence plot of Hotelling's T^2 and control limits

Only two observations fall above the 95% confidence limit, indicating that these observations are far away from the center (scores = 0) of the model but since we expect naturally about 5% of the observations to have a Hotelling's T^2 value above the 95% control limit there are no severe outliers present that could point at a model inconsistency. It is useful to plot the first two scores along with the confidence limits (95% and 99%) corresponding to each observation's Hotelling's T^2 value to identify those observations that are potentially outliers and could be influential on the model fit (figure 3.6).

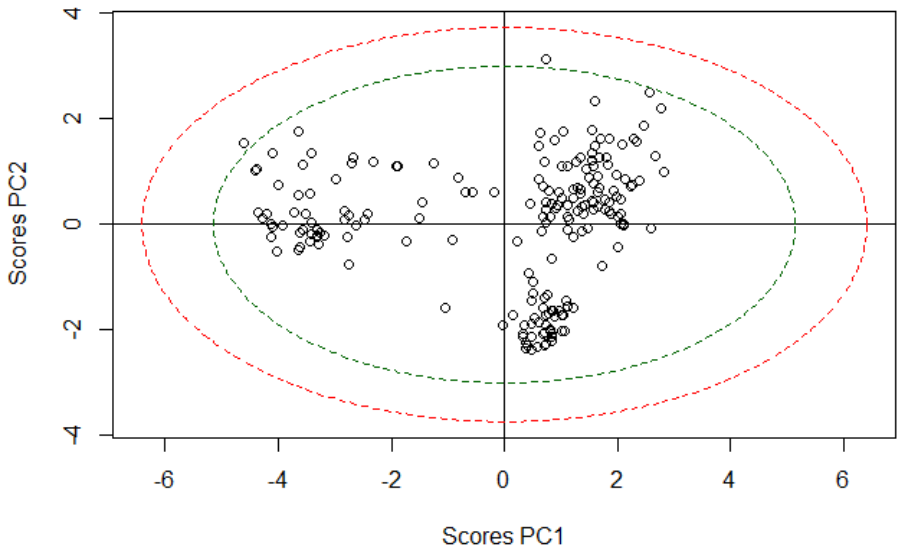


Figure (3.6) Score plot with Hotelling's T^2 control limits

By visual inspection one can clearly identify the same two observations as in figure 3.5 that seem to be furthest away from the center of the scores as well as from the other observations. Out of curiosity they must be music compositions that contain extreme realisations of one or more music theoretical properties. A closer look at these observations (21 and 146) confirms that they have extreme values among the audio features: observation 21 has a z-value of 4.30 in `speechiness` which represents the maximum value of this variable. Observation 146 has a z-value of 1.26 in `instrumentalness` which is close to the maximum value in this variable and a z-value of 2.13 in `liveness`, pointing at a potential inconsistency between these two features that cannot be captured by the model.

	#21	#146
Track_Title	Gibt es Dich?	Recomposed By Max Richter: Vivaldi, The Four Seasons: Spring 1
Track_Artist	Shirin David	Max Richter
Genre	Pop	Classic
acousticness	0.88	0.35
danceability	0.21	0.49
energy	-0.51	-0.79
instrumentalness	-0.90	1.26
liveness	-0.23	2.13
loudness	-0.04	-0.20
speechiness	4.30	-0.61
tempo	-0.95	-0.41
valence	1.08	-0.68

Table (3.2) Outlying observations from PCA ($k = 2$)

Finally, one can combine the Q-residuals and Hotelling's T^2 values and their respective confidence limits to compactly describe the variation within and outside the model (figure 3.7 below). The dashed green lines indicate the 95% confidence limits for Q-residuals (horizontal line) and for Hotelling's T^2 (vertical line). In terms of score outliers there are two observations above the limit for the model where $k = 2$ with the corresponding components explaining 64.09% of the total variance. 35.91%, on the other hand, remains unexplained by the model and six observations are above the control limit.

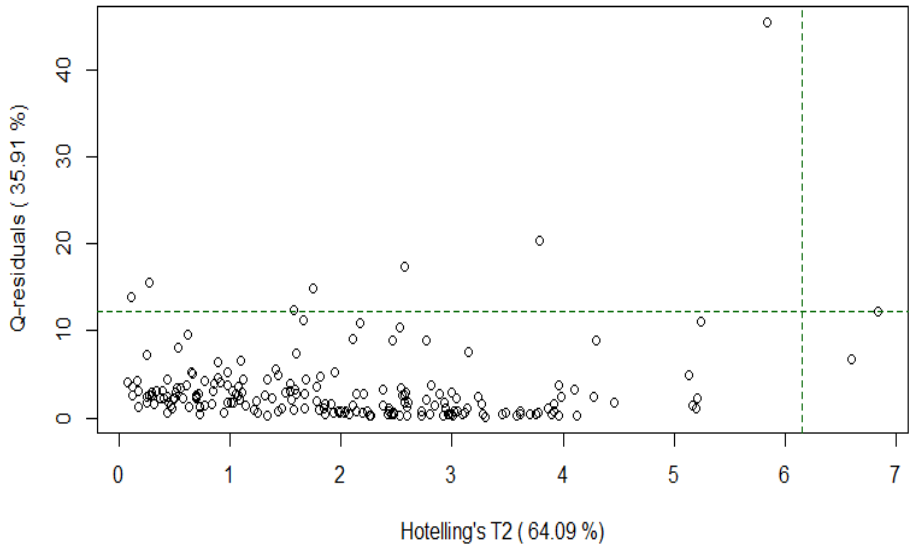


Figure (3.7) Q-residuals vs. Hotelling's T2 (95% control limits)

The adjusted multivariate outlyingness confirms there are seven observations (#21 is one of them) classified as outliers which is mainly due to inconsistent realisations in `liveness` and `energy` relative to their centers but overall there is no evidence for influential observations and severe model inconsistencies.
A parallel analysis of Horn finds two adjusted eigenvalues > 1, indicating there are two dimensions to retain.

Chapter 4

Factor analysis

For this chapter *audio features* as well as any direct indicator for artist popularity and track popularity are excluded in order to guarantee absence of severe multicollinearity and a maximum degree of exogeneous variation when predicting `Artist_Popularity` in section 6.

4.1 Suitability of factor analysis

The inverse correlation matrix (left panel of figure 9.11 in the appendix) suggests that the model holds. The right panel (partial correlations) shows that for many variables the correlation of the residuals is high (i.e. negative partial correlation is small). According to the Kaiser-Meyer-Olkin criterion the overall measure of sampling adequacy (MSA) is 0.74, indicating a suitability of factor analysis for the given dataset in the "middling" range with every individual MSA value at least as high as 0.53 (min: 0.53 [`Artist_Appearances_Tracks_Number`], max: 0.85 [`commentCount`, `viewsCount`]). The Bartlett test of sphericity confirms that the correlation matrix is not the identity matrix, so we can proceed with exploratory FA. Parallel analysis by Horn suggests that there are four components to retain.

4.2 Rotations and loadings

	ML2	ML3	ML1	ML4
Artist_Albums_Number	-0.04	0.75	0.50	0.02
Artist_Albums_Tracks_Number	-0.09	0.64	0.23	0.59
Artist_Appearances_Number	-0.04	0.07	0.21	0.87
Artist_Appearances_Tracks_Number	0.06	-0.07	0.23	0.55
Artist_Compilations_Number	-0.04	0.88	0.36	-0.08
Artist_Compilations_Tracks_Number	-0.05	0.90	0.41	-0.05
Artist_Follower	0.34	-0.05	0.05	-0.05
Artist_Singles_Number	0.07	0.36	0.69	-0.21
Artist_Singles_Tracks_Number	0.09	-0.11	0.99	-0.00
Track_Duration_ms	-0.10	0.06	0.27	0.37
commentCount	0.97	0.03	-0.14	0.00
dislikeCount	0.97	0.04	-0.16	-0.00
likeCount	0.98	0.04	-0.14	0.00
viewsCount	0.94	0.05	-0.11	0.04
days_release	-0.09	0.38	0.11	0.39

Table (4.1) Unrotated factor loadings (maximum likelihood)

From table 4.1 above one can see a simple structure (Thurstone, 1944), that is each row of loadings matrix contains one zero at least, so each item is described by a maximum of $(q = 4) - 1 = 3$ factors in this case. $|\text{Loadings}| > 0.4$ are highlighted. Each column of loadings contains $q = 4$ (near) zero-loadings and there are only two items which load on more than one factor each. Three items (`Artist_Follower`, `Track_Duration_ms` and `days_release`, i.e. the number of days the track has been available to the world since publication) don't load on any factor. At least four absolute loadings are greater than 0.6 (11 items in total), so using a relatively small sample of $n = 196$ shouldn't affect the results (Guadagnoli and Velicer, 1988).

After (*varimax*) rotation `Track_Duration_ms` and `days_release` both load on factor 4, `Artist_Follower` still doesn't load on any factor. Some loadings blocks are substantially high, e.g. 95% of variance in `viewsCount` can be explained by the common "Youtube" factor. A promax rotation of the model's factor loadings shows that two factors, namely ML3 and ML4 are correlated ($\rho = 0.37$) and an orthogonal rotation such as *varimax* is not appropriate. A factor congruence test, however, shows that the factors between an orthogonally rotated factor model (here *varimax*) and an obliquely rotated factor model (here *promax*), both Kaiser normalized, are almost identical with entries of 1 on the

	ML2	ML3	ML4	ML1
Artist_Albums_Number	-0.04	0.86	0.16	0.24
Artist_Albums_Tracks_Number	-0.06	0.62	0.64	-0.11
Artist_Appearances_Number	-0.03	0.04	0.90	-0.02
Artist_Appearances_Tracks_Number	0.05	-0.05	0.59	0.12
Artist_Compilations_Number	-0.03	0.95	0.03	0.09
Artist_Compilations_Tracks_Number	-0.04	0.98	0.08	0.12
Artist_Follower	0.33	-0.04	-0.04	0.11
Artist_Singles_Number	0.02	0.56	-0.03	0.58
Artist_Singles_Tracks_Number	0.00	0.18	0.23	0.95
Track_Duration_ms	-0.11	0.10	0.43	0.14
commentCount	0.97	-0.03	-0.04	-0.05
dislikeCount	0.98	-0.02	-0.05	-0.07
likeCount	0.99	-0.03	-0.04	-0.05
viewsCount	0.95	-0.01	0.01	-0.03
days_release	-0.07	0.36	0.42	-0.11

Table (4.2) Rotated loadings (varimax, ML)

diagonal. The FA solution of the *promax* rotation is adopted due to the correlatedness of two factors.

For orthogonal rotations it holds that the loadings ("pattern") matrix is equal to the structure matrix which both contain the correlation coefficients between the item and the factor since the factor intercorrelation matrix Φ is the identity. For oblique rotations $\Phi \neq I$ and the structure matrix which is the product of the patterns and Φ should be used for interpretation of factors. From table 4.3 we can clearly identify common items loading on either one of the four factors:

- ML2: "Youtube popularity"
- ML3: "Artist's content supply" (pos. correlated with ML4)
- ML4: "Artist long-term activity" (pos. correlated with ML3)
- ML1: "One-hit-wonder"

The almost uncorrelatedness of factors ML2 and ML3 is clear from figure 4.1 below:

	Pattern				Structure			
	ML2	ML3	ML4	ML1	ML2	ML3	ML4	ML1
Artist_Albums_Number	0.01	0.86	0.01	0.17	-0.09	0.89	0.32	0.31
Artist_Albums_Tracks_Number	0.03	0.55	0.55	-0.16	-0.13	0.72	0.74	-0.07
Artist_Appearances_Number	0.03	-0.12	0.94	-0.02	-0.08	0.21	0.89	-0.05
Artist_Appearances_Tracks_Number	0.08	-0.18	0.64	0.13	0.02	0.07	0.57	0.10
Artist_Compilations_Number	0.02	1.00	-0.16	0.01	-0.08	0.94	0.20	0.18
Artist_Compilations_Tracks_Number	0.02	1.02	-0.12	0.04	-0.09	0.98	0.25	0.21
Artist_Follower	0.32	-0.03	-0.02	0.11	0.33	-0.06	-0.07	0.13
Artist_Singles_Number	0.01	0.54	-0.11	0.54	0.00	0.58	0.08	0.63
Artist_Singles_Tracks_Number	-0.03	0.04	0.26	0.96	0.00	0.29	0.27	0.96
Track_Duration_ms	-0.08	0.01	0.44	0.13	-0.14	0.20	0.45	0.13
commentCount	0.98	0.03	0.01	-0.07	0.97	-0.10	-0.11	0.01
dislikeCount	0.99	0.04	0.00	-0.08	0.98	-0.10	-0.12	-0.01
likeCount	1.00	0.03	0.01	-0.06	0.99	-0.10	-0.12	0.01
viewsCount	0.96	0.04	0.06	-0.05	0.95	-0.07	-0.07	0.03
days_release	-0.01	0.32	0.36	-0.14	-0.11	0.43	0.48	-0.09

Table (4.3) Rotated factors (promax, ML)

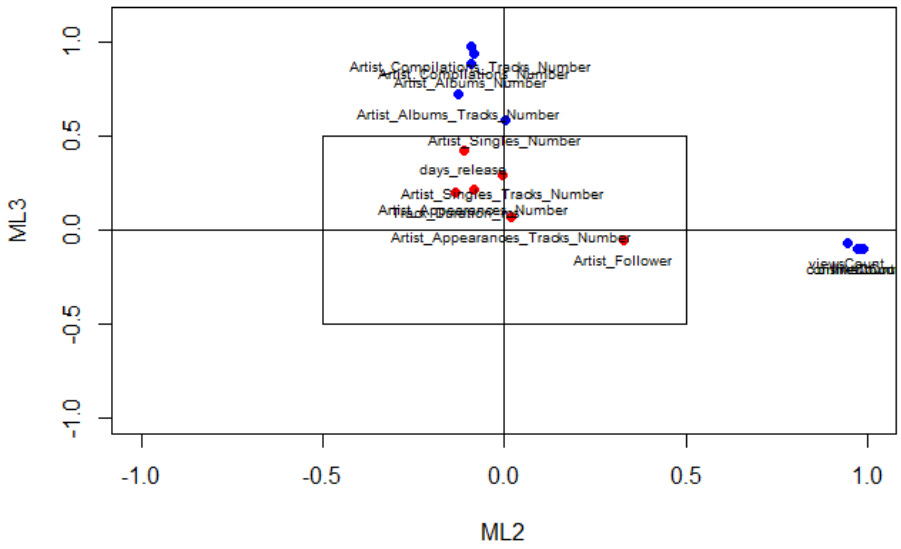


Figure (4.1) Structures plot (promax, ML |threshold| > 0.5)

For the sake of interpretability scores were computed using `psych::keys` and the `psych::scoreItems` functions omitting structure matrix entries below 0.5 (which didn't account for ambiguity).

4.3 Factor reliability

Cronbach's α

	ML2	ML3	ML4	ML1
Std. α	0.99	0.91	0.74	0.78

Table (4.4) Cronbach's α

For the first construct (ML2), the standardized α's decrease if a single item is dropped or remain almost unchanged, indicating a strong internal consistency for this construct. For ML3 the standardized α's increase when `Artist_Albums_Tracks_Number` or `Artist_Singles_Number` are dropped, reflecting the ambiguity of the first item and the low magnitude of the second item's loading shown in table 4.3. For ML4 the deleted Cronbach's α drops significantly if `Artist_Appearances_Number` is dropped, underlining its major correlation with the underlying factor. For ML1 Cronbach's α drops as well when dropping either one of the two items, relative to the overall value of standardized α. Overall, for all constructs Cronbach's α suggests that internal consistency is between "good" and "excellent".

Tukey's test on additivity

For constructs ML2 and ML3 the Null hypothesis that the sumscore appropriately describes the input variables are rejected. For constructs ML4 and ML1 it can't be rejected, however the number of items in those is rather small and a single item could be a perfect predictor for itself, leaving the test ambiguous about true reliability.

Chapter 5

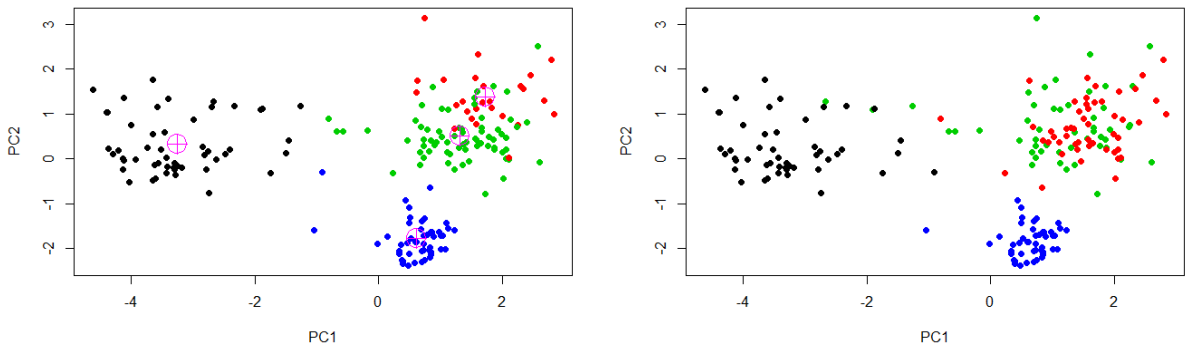
Cluster analysis

This chapter is to empirically assign memberships of music tracks based on the *audio features* which could already be visually recognized in figure 3.3.

5.1 k-Clustering

We know that there are four clusters in the dataset, `Genre`, and figure 9.13 confirms this as an appropriate number for k.

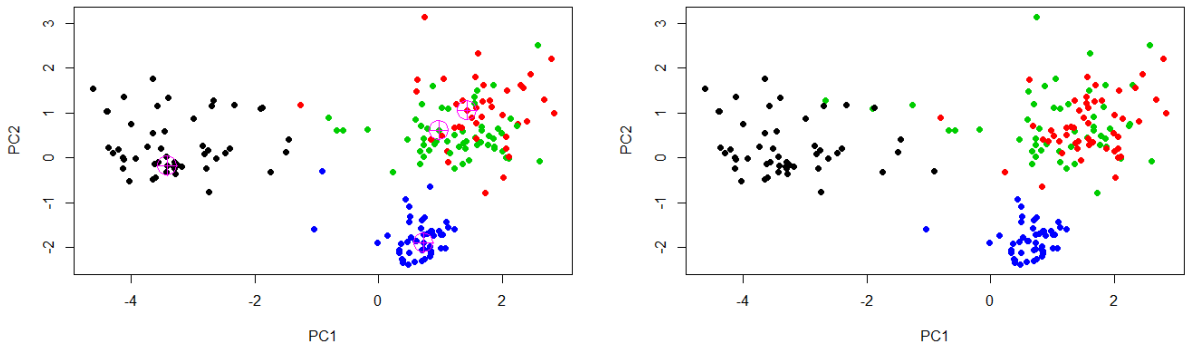
Figure (5.1) k-means clustering



(a) (k=4) solution with theoretical centroids

(b) True classes

Figure (5.3) k-medoids clustering



(a) (k=4) solution with medoids

(b) True classes

After reassigning memberships to correspond to their actual classes tables 5.1 and 5.2 below show the percentage proportions of observations by `Genre` in each cluster C.

	Classic	Hip Hop	Pop	Techno
$C1$	98	0	6	0
$C2$	0	36	16	0
$C3$	0	62	78	0
$C4$	2	2	0	100

Table (5.1) Class memberships k-means (k=4, accuracy=0.78)

	Classic	Hip Hop	Pop	Techno
$C1$	98	0	4	0
$C2$	0	56	26	0
$C3$	0	42	70	0
$C4$	2	2	0	100

Table (5.2) Class memberships k-medoids (k=4, accuracy=0.81)

For *k-means* it seems that Classic music and Techno music can be accurately distinguished but between Hip Hop and Pop there are many false negatives, e.g. 62% of the tracks were predicted to be of the genre Pop although the truth was that they were of class Hip Hop, underlining the ambiguity between those genres in terms of music theoretical characteristics nowadays. Vice versa, 16% of the tracks predicted to be Hip Hop music were actually Pop music. The medoids found by the cluster algorithm are given by

#	Track_Title	Track_Artist	Genre
142	Khachaturian: Spartacus Suite No. 2: I. Adagio of Spartacus and Phrygia	Aram Khachaturian	Classic
149	DT64	Moguai	Techno
80	Lambo Diablo GT (feat. Nimo & Juju) - Remix	Capo	Hip Hop
94	CASINO ROYAL	Kianush	Hip Hop

Table (5.3) Medoid observations

These observations should represent the most centric cases in terms of musictheoretical charateristics w.r.t. their class. Interestingly, a Hip Hop track (#80) has been assigned as the cluster center for the genre Pop, reflecting the apparent similarities between these genres.
Instead of finding clusters over the entire feature space one can first run a factor analysis and then k-clustering over the obtained scores. For number of factors = 2 the result was worse compared to the full feature space as the majority of observations from the locus of Pop and Hip Hop music were found to belong to one single cluster and a new, non-existing cluster was found to be between the classic music and Pop/Hip Hop music clusters.
Soft-clustering is an alternative to assigning class membership discretely and coming up with probabilities of association for each observation and cluster. For the problem at hand it was difficult to distinguish between the Pop/Hip Hop clusters.

5.2 Other clustering methods

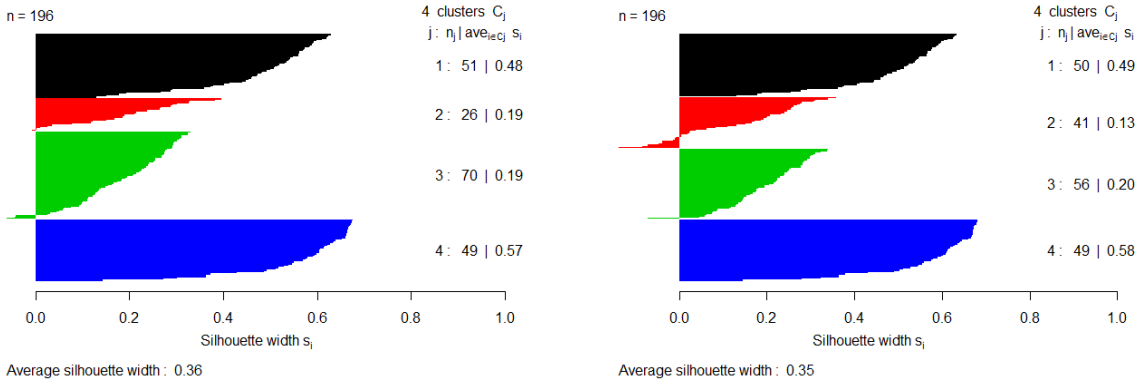
For the given data using hierarchical clustering a 4-cluster solution was not optimal and many observations from Pop/Hip Hop were assigned to Techno (a two or three cluster solution was suggested according to the dendrogram). Turning distances to similarities (figure 9.14 in the appendix) supports this proposition. Using Ward's minimum variance distance returns more equally sized clusters but a two or three cluster solution seems more appropriate than the true four cluster solution (figure 9.15 in the appendix). The biggest "jumps" occur at heights of 38.3 (1 cluster) to 21.9 (2 cluster) and 21.9 (2 cluster) to 13.3 (3 cluster).

If each cluster follows a multivariate normal distribution then a parametric model can be used to maximize the probability that an observation i belongs to a cluster k in an iterative fashion (Bayes-Theorem and EM-Clustering). For the given problem this assumption is clearly not satisfied and a six cluster solution was proposed.
DBSCAN is a density-based algorithm which distinguishes between core observations and noise observations that aren't close enough (closeness determined by a free parameter ϵ) to any cluster. For varying values of ϵ the results differ strongly with $\epsilon = 1$ and `MinPts` $= 5$ being the only parametrization able to identify a three cluster solution with a major amount of noise observations.
Mixed clustering determines in a first step clusters according to hierarchical clustering (Ward's distance) and supplies the cluster centers (means of items) to *k-means* as starting guesses to improve the initial solution from the dendrogram in the first step. Also here the method fails to distinguish between Pop and Hip Hop music.

5.3 Silhouette method

The Calinski-Harabasz index proposed a two cluster solution. An ensemble of 24 indices found a three cluster solution according to a majority vote. Since we know there are four clusters in the data the Silhouette scores for the k-means and k-medoids solutions were computed when $k = 4$.

Figure (5.5) Silhouette plots



(a) Silhouette scores k-means

(b) Silhouette scores k-medoids

The cluster solution of k-means has a slightly higher average silhouette score than the k-medoids solution. Both clustering methods seem to capture clusters 1 and 4 well with most intra-cluster sample silhouette scores greater than the total average score. A few observations seem to have been misassigned (cluster 3, left panel figure 5.5), whereas k-medoids seems to have misassigned additionally observations from cluster 2 (right panel of figure 5.5). Naturally, the k-means solution should be accepted according to the Silhouette criterion but since true labels are known and the k-medoids solution has a higher accuracy this solution was adapted for further modelling steps.

Chapter 6

Regression analysis

In this chapter a linear model that is able to explain the variation in `Artist_Popularity` as a proxy for actual streamed tracks of this artist based on the components and factors scores extracted in chapters 3 and 4 and other measures that might be correlated with `Artist_Popularity` such as `Artist_Follower`, the duration this artist's track has been available to the world `days_release` and the popularity of the artist's track `Track_Popularity` is developed. All following models' results are presented in table 6.3 at the end of section 6.3.

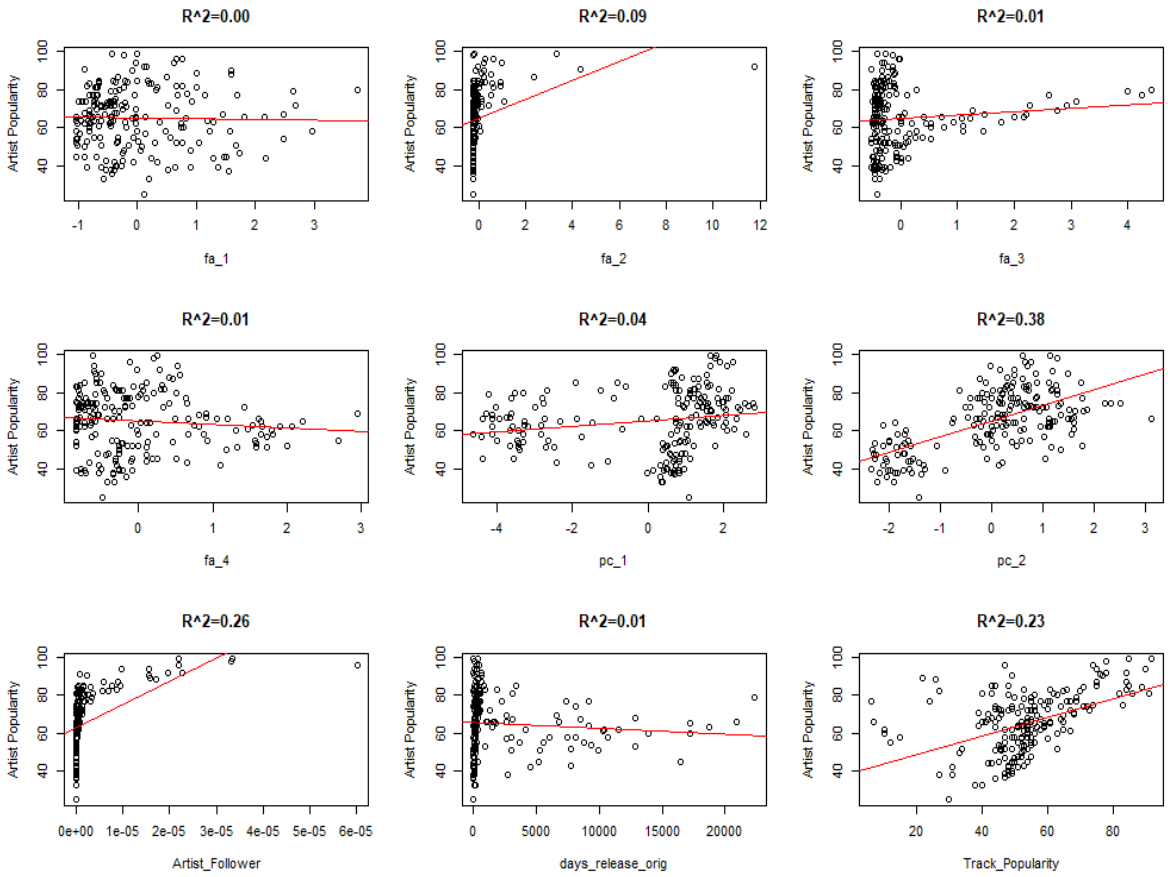


Figure (6.1) Partial regression plots

The initial model (`model1`) specification in base variables is

$$Artist\_Popularity_i = \beta_0 + \beta_1 fa\_1_i + \beta_2 fa\_2_i + \beta_3 fa\_3_i + \beta_4 fa\_4_i + \beta_5 pc\_1_i + \beta_6 pc\_2_i + \beta_7 Artist\_Follower_i + \beta_8 days\_release_i + \epsilon_i \tag{6.1}$$

All coefficients but for `fa_1` and `fa_4` are statistically significant and the signs are as expected. Computed variance inflating factors show that `fa_1`, `fa_3`, `fa_4`, `pc_1` impose mild collinearity (>2) to the model, potentially invalidating statistical inference. The condition index for the variable with the smallest eigenvalue (3.91), however, suggests that multicollinearity is neither severe nor mild. The R^2 is 0.64.

Removing `fa_1` and `fa_4` due to insignificance gives the next model specification (`model2`)

$$Artist\_Popularity_i = \beta_0 + \beta_1 fa\_2_i + \beta_2 fa\_3_i + \beta_3 pc\_1_i + \beta_4 pc\_2_i + \beta_5 Artist\_Follower_i + \beta_6 days\_release_i + \epsilon_i \quad (6.2)$$

Standardizing the dependent and independent variables gives the coefficients the interpretation of proportion of the dependent variable's variance (which is 1) when squaring them. In this case, the squared standardized regression coefficients of `model2` sum up to 60% with `pc_2` explaining about 32% and `Artist_Follower` 11%.
Visual inspection of figure 6.1 suggests nonlinearity in the regression function. The partial regression plots (figure 6.2 below) without and with accounting for a quadratic term in `fa_2`, `pc_2` + `Artist_Follower` and the respective R^2 values add support to this model specification.

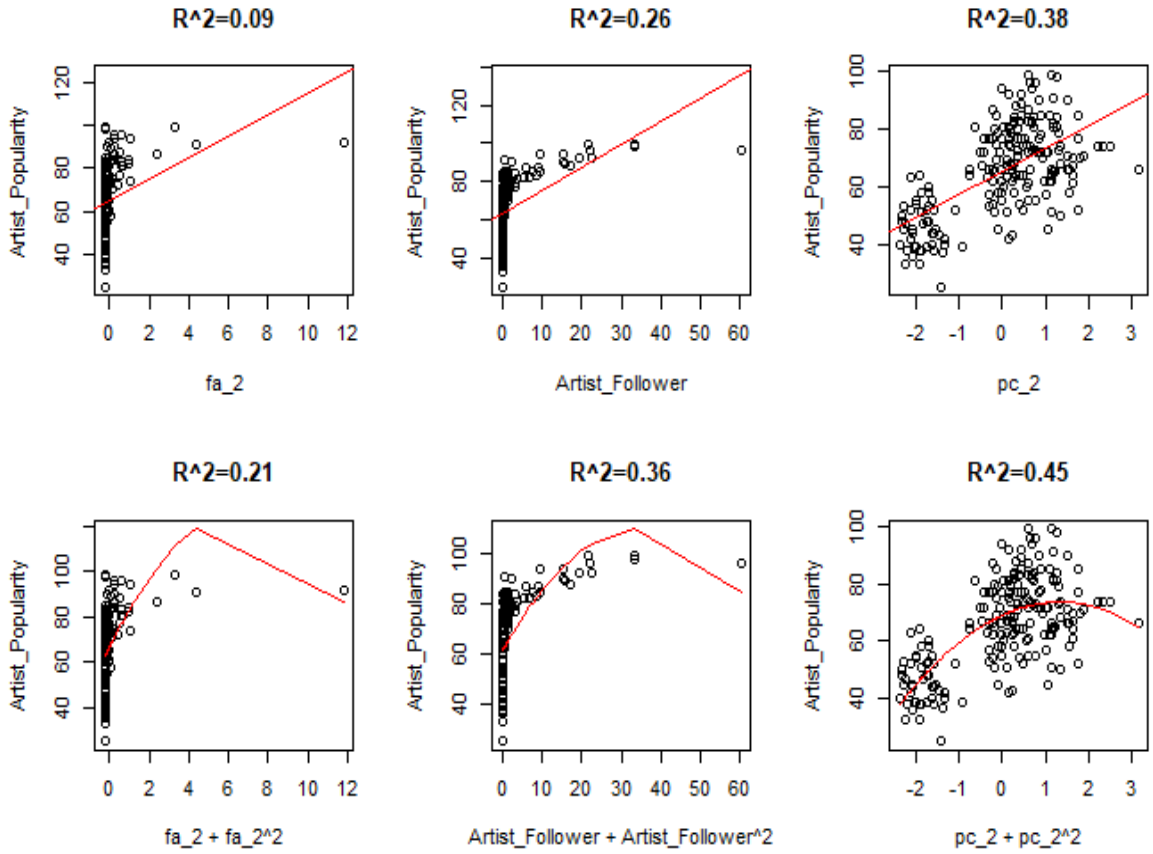


Figure (6.2) Partial regression plots w/o squared terms

6.1 Residual diagnostics

Normality

One assumption of linear regression required for asymptotic inference on the coefficients is that the error term is normally distributed. The KS-Test for the residuals from `model1` and `model2` suggests that the normality assumption is (approximately) satisfied.

Nonlinearity

Nonlinearity in the regression function as mentioned above can also be identified by residual diagnostics. Figure 9.16 in the appendix confirms this suspect in the variables `fa_2`, `Artist_Follower` and `pc_2`. Figure 9.17 hints at nonlinearity in the fitted values that can explain the residuals.

Variance

The linear regression model assumes that the variance-covariance matrix of the error term is spherical, i.e. the variance for each observation's error term is constant and identical across all observations and that the error terms are uncorrelated between the observations. In short, it must hold that $\mathbb{V}(\epsilon) = \sigma^2 I_n$ (or just *"iid"*).
Testing the constancy of the error one can apply the Breusch-Pagan (BP) test by constructing an auxiliary regression to regress the squared residuals from the original regression (i.e. from `model2`) on the same explanatory variables (the White test introduces polynomial and interactive terms). In this case, the test statistic is $0.1067 \cdot 196 = 20.91$ and $p = 6$. $\chi^2_{df=p,\alpha=0.05} = 12.59$ and the probability that the test statistic is greater than (since the χ^2-test is one-sided) the critical value 12.59 is $prob$ = 1- quantile of $\chi^2(20.91, df = 6) = 0.0019$ (also known as p-value). As this p-value is less than the conventional level of 0.05, meaning the probability of computing a test statistic as extreme as 20.91 *conditional* on that the Null-hypothesis was true (i.e. there was NO heteroskedasticity $\iff$ constancy of the error satisfied), the observed test statistic cannot be believed to be reproducable given there was truly homoskedasticity and hence the Null hypothesis is rejected. Consequently, asymptotic inference of the regression coefficients would be invalidated since inflated variance of the regression coefficients would lead to a situation where the Null hypothesis is failed to be rejected longer than neces-

sary.
On the correlational structure of the estimated error terms (= residuals) one can conduct the Durbin-Watson (DW) test and, in this case, the Null hypothesis of no autocorrelation cannot be rejected ($p = 0.2323$).

6.2 Model adjustment

As indicated by figure 6.2 and the residuals diagnostics (figures 9.16 and 9.17) a modification to account for nonlinearity in the regression function seems appropriate. However, not every seemingly existing curvilinear relationship carries a truly quadratic component if the predictor is e.g. highly skewed. Recall that before any transformation `Artist_Follower` was extremely right skewed and after a power transformation the relationship between `Artist_Follower` and `Artist_Popularity` was (almost perfectly) linear. Power transformations for `fa_2` and `pc_2`, however, do not lead to such a linear relationship with `Artist_Popularity` and the linear terms being not orthogonal to their quadratic counterparts. Additionally, the vertices for all three variables lie inside their respective variables' range, an indication that there's truly a quadratic component in the relationships. Therefore, `Artist_Follower` is also treated as the untransformed variable in the following and contains a quadratic component.

Linear term	quadratic term	Vertex	Range	s.e.	z	lower $CI_{95\%}$	upper $CI_{95\%}$	p-$value_{df=186}$
`fa_2`	`fa_2`^{2}	5.5283	`True`	0.6272	8.8146	4.2990	6.7575	4.440892e−16
`pc_2`	`pc_2`^{2}	1.3638	`True`	0.3006	4.5361	0.7745	1.9530	5.125243e−06
`Artist_Follower`	`Artist_Follower`^{2}	35.5193	`True`	3.6797	9.6527	28.3071	42.7314	0.000

Table (6.1) Nonlinear combinations

Table 6.1 above shows that e.g. the vertex of `Artist_Follower` is at 35.52 millions followers, after that point `Artist_Popularity` decreases with any additional follower overally. The test statistic z for testing significance of a nonlinear combination of each variable and its quadratic term is the coefficient (here the vertex) divided by the standard error which has been computed with the delta method. All tests reject the Null hypothesis (i.e. the vertices $\neq 0$) and add further support to account for quadratic relationships. (The conducted routine is equivalent to the `nlcom` function in `Stata`.)

The new model specification is (`model3`)

$$\begin{aligned} Artist\_Popularity_i = \beta_0 + \beta_1 fa\_2_i + \beta_2 fa\_2_i^2 + \beta_3 fa\_3_i \\ + \beta_4 pc\_1_i + \beta_5 pc\_2_i + \beta_6 pc\_2_i^2 + \beta_7 Artist\_Follower_i \\ + \beta_8 Artist\_Follower_i^2 + \beta_9 days\_release_i + \epsilon_i \end{aligned} \tag{6.3}$$

Residual diagnostics

The residuals are still (approximately) normally distributed (KS-Test: $D = 0.0368 < 0.0970$).

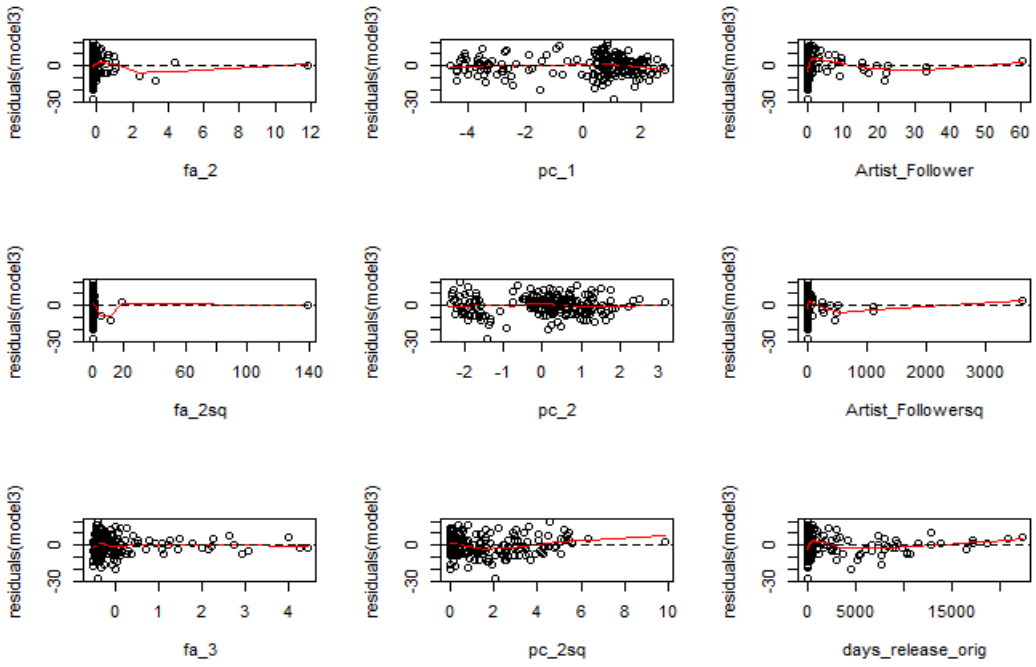


Figure (6.3) Residuals vs. predictors (`model3`)

Figure 6.3 indicates that the nonlinearity in the residuals has disappeared. The residuals against fitted values from `model3` show no obvious structure anymore.

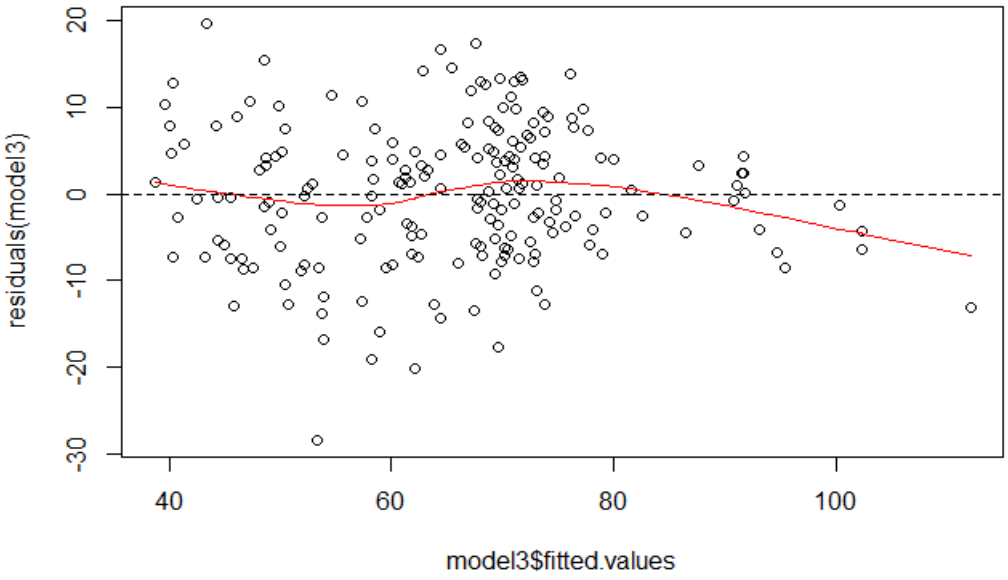
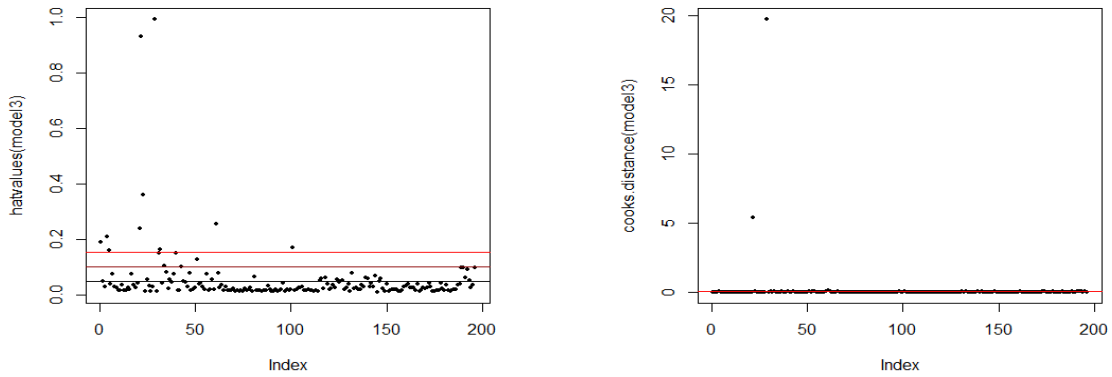


Figure (6.4) Residuals vs. fitted values (`model2`)

The test statistic for the BP-Test is $R^2 \cdot n = 0.0676 \cdot 196 = 13.25$ and the p-value is 0.1518, hence we fail to reject the Null hypothesis of homoskedasticity and note that after accounting for diminishing marginal effects the variance of the residuals is constant. The DW-Test yields that the errors are still uncorrelated. The correlational structure of the residuals from `model3` is consistent with the assumption of spherical errors from the Gauss-Markov theorem.

6.3 Influential observations and outlier detection

Figure (6.5) Outlier detection in `model3`



(a) Leverage

(b) Cook's distance

From figure 6.5 (left panel) we can see ten observations i $h_i = x_i'(X'X)^{-1}x_i$ exceeding $\frac{3(p+1)}{n}$, indicating those observations are relatively far away from the center of data and constitute unusual realizations ("leverage"). Assessing an observation i's influence on the overall fit of `model3` ("Cook's distance") shows that two observations change the model's coefficients substantially ($C_i > \frac{4}{196}$) compared to the other data points (figure 6.5, right panel). Standardized differences of coefficients ("SDBETA") show the influence of each observation on each coefficient by the change of including and excluding an observation. Therefore, coefficient-specific robustness can be tested.

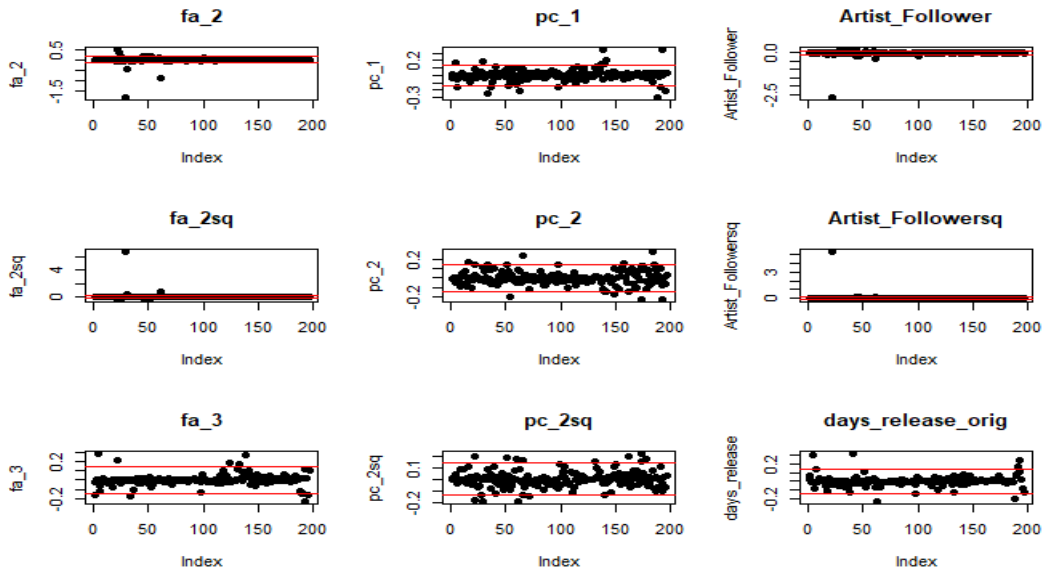


Figure (6.7) Regression deletion β's

Figure 6.7 shows that especially coefficients for `fa_2` and `Artist_Follower` (and their quadratic effects) are significantly altered by a few observations. Another measure of regression deletion on overall fit is conceptually related to Cook's distance and can be found in figure 9.18 in the appendix. SDFITS identified three observations, two of them are those identified by Cook's distance and all of them are also covered by those cases whose leverage is exceeding the highest threshold of $\frac{3(p+1)}{n}$. These cases are:

#	Track_Title	Track_Artist	Genre	Artist_ Popularity	Artist_ Follower	viewsCount
1	The Well-Tempered Clavier: Book 1, BWV 846-869: 1. Prelude in C Major, BWV 846	Johann Sebastian Bach	Classic	80	2,492,413	8,580
4	Clarinet Concerto in A K622 (1990 Digital Remaster): II. Adagio	Wolfgang Amadeus Mozart	Classic	79	3,321,232	509
5	Symphony No. 5 in C Minor, Op. 67: 1. Allegro con brio	Ludwig van Beethoven	Classic	77	3,104,946	3,845
21	Gibt es Dich?	Shirin David	Pop	66	370,642	201,567
22	Take Me Back to London (feat. Stormzy)	Ed Sheeran	Pop	96	60,136,077	60,116,012
23	Dance Monkey	Tones and I	Pop	91	748,052	748,367,421
29	Girls Like You (feat. Cardi B)	Maroon 5	Pop	92	22,742,618	2,598,870,760
32	Intentions	Justin Bieber	Pop	99	33,340,962	281,241
61	all the good girls go to hell	Billie Eilish	Pop	99	21,832,072	117,405,850
101	In Too Deep	Eminem	Hip Hop	98	33,281,605	1,865,363

Table (6.2) Outlying observations (`model3`)

Since the factor scores of `fa_2` can be interpreted as "Youtube popularity" these observations outlyingness is primarily driven by extreme realizations in this factor which is comprised by `viewsCount`, among others. The identified observations have unusually many clicks, even so extreme that one of them (# 29) qualified for the Global Top 30 most-viewed YouTube videos[1] and it is sensible to disregard those observations for the rest of this section. The final linear model `model4` was fitted according to the same model specification as in equation 6.3 without the ten aforementioned observations from table 6.2.

[1] Wikipedia: List of most-viewed YouTube videos, accessed on 03/19/2020

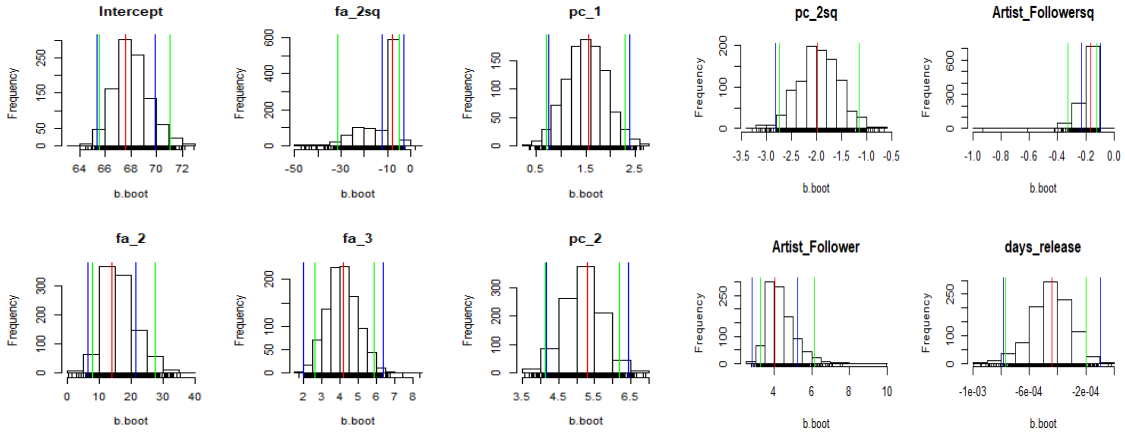
Table (6.3) Regression results

	Dependent variable:			
	Artist_Popularity			
	(model1)	(model2)	(model3)	(model4)
fa_1	−0.082			
	(1.139)			
fa_2	1.848**	1.857**	6.086***	14.029***
	(0.757)	(0.755)	(1.728)	(3.777)
fa_2sq			−0.550***	−7.703***
			(0.162)	(2.381)
fa_3	5.409***	5.431***	4.497***	4.181***
	(1.392)	(1.058)	(0.901)	(1.114)
fa_4	−1.299			
	(1.283)			
pc_1	1.402**	1.745***	1.813***	1.566***
	(0.591)	(0.478)	(0.409)	(0.420)
pc_2	7.454***	7.450***	5.951***	5.299***
	(0.657)	(0.600)	(0.537)	(0.567)
pc_2sq			−2.182***	−1.983***
			(0.389)	(0.424)
Artist_Follower	0.809***	0.785***	1.661***	4.031***
	(0.120)	(0.115)	(0.229)	(0.623)
Artist_Followersq			−0.023***	−0.163***
			(0.005)	(0.035)
days_release	−0.0005**	−0.001**	−0.0005***	−0.0004**
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Constant	64.514***	64.650***	67.561***	67.578***
	(0.869)	(0.856)	(0.988)	(1.156)
Observations	196	196	196	186
R^2	0.636	0.634	0.745	0.745
Adjusted R^2	0.621	0.623	0.732	0.732
Residual Std. Error	9.765 (df = 187)	9.740 (df = 189)	8.206 (df = 186)	7.855 (df = 176)
F Statistic	40.905*** (df = 8; 187)	54.645*** (df = 6; 189)	60.249*** (df = 9; 186)	57.058*** (df = 9; 176)
Note:				*p<0.1; **p<0.05; ***p<0.01

6.4 Inference

When bootstrapping regression coefficients strong deviations from the asymptotic limits indicate high uncertainty around the point estimates and test results are suspect.

Figure (6.8) Bootstrapped confidence limits



Clearly, all bootstrapped coefficients for the variables except those associated with `fa_2` and `Artist_Follower` are centered around their one-off estimation (red line) with the bootstrapped lower and upper 95% quantiles (green lines) in line with the asymptotics-based limits (blue lines). The uncertainty around `fa_2` is substantial (the true parameter lying somewhere between 7 and 30 cp. 14.029 from the regression) and its squared term as low as -37 which seems implausibly low for a diminishing marginal effect. Overall, the final specification and sample yields robust estimates and a high goodness of fit.

6.5 Forward, backward and stepwise regression

Running a forward stepwise regression with AIC as decision criterion in order to maximize predictiveness and providing additional variables for selection such as `Track_Popularity` as well as those excluded from `model1` (`fa_1` and `fa_4`) confirms that `pc_2` is the variable with the largest absolute correlation with `Artist_Popularity` and only inclusion of `Track_Popularity` could improve the model's goodness of it (though this coefficient is negligible in magnitude). The resulting R^2 is 0.7897. Using a backward regression yields the exact same specification with `days_release` being the only variable whose inclusion not adding significantly to the goodness of fit. A major drawback of stepwise models is their tendency to overfit by minimizing internal bias but not being able to produce out-of-sample robust estimates (i.e. high variance).

6.6 Regularization methods and cross validation

Regularization models such as Ridge or the Least Absolute Shrinkage Selection Operator (LASSO) regression impose a bias (L2 and L1 norm, respectively) to the least squares minimization problem which essentially penalizes the size of the coefficients if they don't contribute to the explanation of the dependent variable by shrinking them towards zero (but only LASSO can set them exactly to zero). The "rigor" (or sensitivity) of penalization is controlled by the hyperparameter λ and can be determined by cross-validation. In this case, λ corresponding to the most regularized model such that the error is within one standard error of the minimum is 0.3063. For both parametrizations of `lambda.min` and `lambda.1se` the model coefficient signs are consistent, however those when setting $\lambda =$ `lambda.1se` are smaller in size which was expected under a more rigorous penalization. The model parameter selection, however, is identical for both values for λ and also to the variables included by backward stepwise modelling (`fa_1` and `fa_4` are shrunk to 0/excluded). In terms of variable importance (not mentioning the intercept) `fa_2`, `pc_2` have the largest coefficients, `days_release` and `Track_Duration_ms` the smallest.
Elasticnet combines L1 and L2 penalties of LASSO and Ridge regression, respectively, linearly such that minimum MSE (which is the mean-variance trade-off) is achieved.
It is good practice to check model performance on separate sample parts, the training and testing sets (e.g. train: 75%, test: 25% and assign observations randomly to either one), to see how well the model is able to generalize over unseen data. Parametertuning is done on the training dataset with cross-validation on internal validation sets. By comparing unregularized (low bias, high variance) with regularized models (high bias, low variance) we expect to achieve better model performance on unseen data (the test set). This is indeed the case for all regularization methods which yield lower RMSEs/higher R^2 values on the test set than the full linear model which can be seen from figure 9.21 in the appendix (figures 9.19 and 9.20 in the appendix show higher goodness of fit on seen data).

6.7 Non-parametric and semi-parametric regression

Non-parametric and partial linear models are regression methods where the predictors are not taken as predetermined form but are additionally estimated itself by kernel den-

sities. As these models are generally difficult to interpret and to trace some prominent techniques and the results are only briefly presented for the sake of completeness.
A univariate kernel density estimator for `Track_Duration_ms` suggests a bimodal distribution (see figure 9.1).

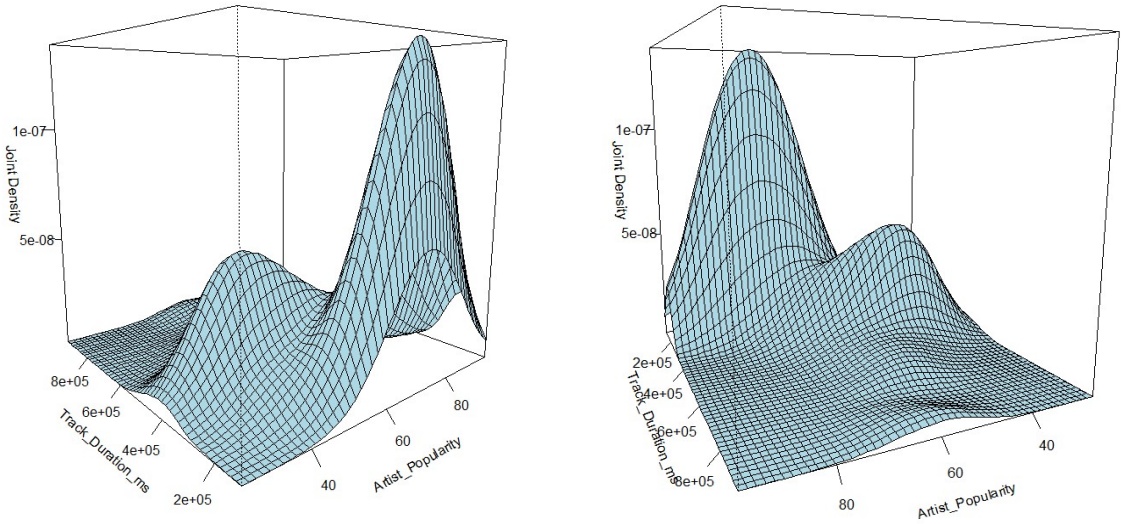


Figure (6.10) Bivariate kernel density plots

A bivariate kernel density plot (figure 6.10, left panel) above indicates one mode at `Track_Duration_ms` 500 seconds (about eight and a half minutes) and `Artist_Popularity` 50 (meaning longer tracks' artists are less popular). The other mode is where `Track_Duration_ms` 200 seconds (about three and a half minutes) and `Artist_Popularity` 80.
An entirely additive model (regressing the dependent variable on only smoothed functions of the predictors) using the same full regression function as in section 6.6 indicates nonlinearity in some variables, however the linear terms which have been accounted for with a quadratic term such as `pc_2` show an almost perfect linear relationship with their respective smoothed functions. The residuals indicate no nonlinearity at all. The R^2 is 0.9503.
A fully specified single index model yields an R^2 of 0.87 and no nonlinearity can be identified from the residuals against fitted values. A projection pursuit regression with five terms yields an R^2 of 0.9590 (which is the highest measured across all non-cross-validated models) and some structure in the residuals.

Chapter 7

Decision trees and neural networks

Decision trees are a structural representation of sequences of rules which divide the dataset into subgroups such that some impurity measure (e.g. Gini index reduction, entropy for classification and variance reduction for regression) is minimized.

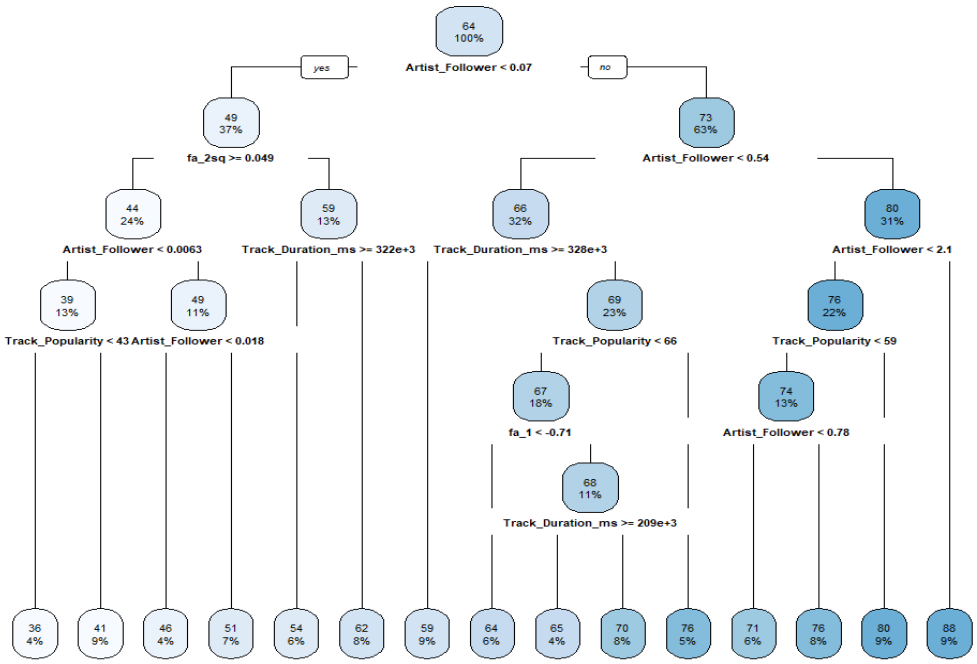


Figure (7.1) Simple decision tree (regression, full sample)

The decision tree in figure 7.1 above shows that only five variables are sufficient to ex-

plain 91.32% in the variation of `Artist_Popularity` of which `Artist_Follower` and its square term have the highest variable importance (the sum of gains in the node's purity measure of all nodes and surrogates splits). Pruning aims to reduce overfitting by reducing the tree's complexity which results in a marginally lower R^2 of 0.9062. A 5-fold cross validation yields an R^2 of 0.75 on the test set.

An ensemble of several trees ("weak" base learners) comprised of randomly selected features, a *random forest* algorithm, produces an almost perfect fit ($R^2 = 0.98$) on the full sample and generalizes well for unseen data with an $R^2 = 0.89$, the highest achieved among all regression models. The cross-validated set of hyperparameters was found to be `mtry` $= 5$ which corresponds to the recommendation of $\sqrt{p}$ or $\frac{p}{3}$ (Breiman, 2003) and determines the number of available features to randomly choose when constructing a single tree, `sampsize` $= 70$ (bootstrap sample size) and `ntree` $= 50$, the total number of base learners to form the final ensemble and conduct a voting about majority agreement over.

Neural networks emulate the functioning of the human brain by arranging neurons ("face-off spots", units) in a sequential way and processing inputs (features) through the entire network such that some target metric w.r.t. the truth (response or class) is optimized. In this case, `size` $= 2$ (number of units per hidden layer and there is only one hidden layer which satisfies the universal approximation property) and `decay` $= 0.001$, a penalization parameter on the weights.

Both approaches can also be used for classification. For instance, we can use the features used for prediction of `Artist_Popularity` to classify whether the artist/track appeared in the German Spotify Charts at least once in the past 12 months.

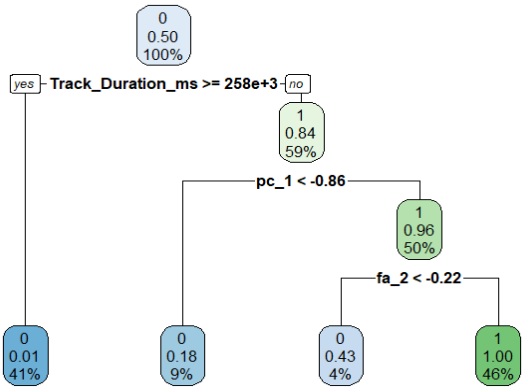


Figure (7.2) Simple decision tree (classification, full sample)

Figure 7.2 above shows that the best predictors to differentiate between those tracks/artists which appeared in the charts (1) and those which did not (0) are `Track_Duration_ms`, `pc_1` and `fa_2`. The rules found are such that the probability of being in the charts is 50% without any split (as there are 93 observations which did appear in the charts and 93 which did not) and tracks longer than approx. four minutes have a probability of 1% of appearing in the charts. The next surrogate split for shorter tracks distinguishes observations differing in their degrees of `danceability`, `energy`, `loudness` (higher values) and `acousticness` and `instrumentalness` (smaller values) such that more danceable, energetic and loud tracks are more likely to appear in the charts (96%). In the terminal node the degree in factor 2 `fa_2` attributed with "YouTube popularity" separates the remaining 55 observations that have lower values in `fa_2` than -0.22 and did not appear in the charts (4%) from those in the node that appeared in the charts with certainty. The overall accuracy for the full, unpruned tree is 0.9624. The best complexity parameter found during cross validation is 0.0108 with an accuracy of 0.9569 on the full sample. The confusion matrices from the cross-validated models for the train/test sets are given by with a complexity parameter for the simple decision tree `cp` = 0.7647.

Table (7.1) Simple decision tree (confusion matrices)

		Truth	
		Not in charts (0)	In charts (1)
Prediction	Not in charts (0)	71	3
	In charts (1)	4	64

Train set (accuracy=0.95)

	0	1
0	19	0
1	3	22

Test set (accuracy=0.93)

Table (7.2) Random forest (confusion matrices)

	0	1
0	74	0
1	1	67

Train set (accuracy=0.99)

	0	1
0	19	0
1	0	25

Test set (accuracy=1.00)

Table (7.3) Neural network (confusion matrices)

	0	1
0	74	0
1	0	68

Train set (accuracy=1.00)

	0	1
0	19	0
1	0	25

Test set (accuracy=1.00)

Chapter 8

Concluding remarks and future considerations

This paper has comprehensively reviewed and applied data analysis methods using a small data set compiled from the Spotify Web API and the YouTube Data API. Music-theoretical metrics were shown to be descriptors able to successfully distinguish between music genres but for the most popular genres, Pop and Hip Hop music, these differences were marginal in terms of psychoacoustics and other characteristics. Latent factors were extracted and interpreted according to popularity on another music consumption platform, an artists' content supply and long-term / short-term effort and success. Clustering techniques confirmed the presence of at least three clusters in the data set (there were four clusters or genres) but turned out to be insufficient to clearly separate Pop music from Hip Hop music. A linear regression model was developed which, after thoroughly testing robustness, was able to explain about 75% in the variation of the artists' popularity, a proxy for royalties received from streaming with emphasis on the "true" model parameters rather than low variance. Lastly, predictiveness was prioritized over interpretability and cross validation was applied for more advanced regression and classification tasks. The non-cross validated performance metrics on the full sample are presented along with the cross validated ones on the training and test data sets in table 8.1 below.

	Full sample		train set		test set	
Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
Linear model (full)	6.91	0.79	6.91	0.80	7.14	0.75
Ridge regression	7.26	0.77	7.39	0.77	8.15	0.68
Lasso regression	7.14	0.78	7.06	0.79	7.57	0.72
Elasticnet	7.18	0.78	7.09	0.79	7.63	0.72
GAM	3.37	0.95	2.92	0.96	6.94	0.80
SIM	5.51	0.87	5.91	0.86	5.87	0.84
PPR	3.06	0.96	3.46	0.95	7.17	0.76
Decision tree	4.46	0.91	4.47	0.92	7.22	0.75
Random forest	2.11	0.98	3.51	0.95	4.91	0.89
Neural network	-	-	4.21	0.93	6.64	0.80

Table (8.1) Performance metrics (regression)

Based on out-of-sample performance (lowest RMSE $\Longleftrightarrow$ highest R^2) the random forest model should be chosen. For classification the random forest model and the neural network yielded the maximum accuracy on the test set.

More complex models might be able to incorporate a broader diversity of genres. [1] Future research might be able to model `Track_Popularity` rather than `Artist_Popularity` and obtain actual streaming counts to identify determinants of "hit songs" based on a broader set of predictors than audio features and one substitute platform for music consumption. Eventually, a long-term perspective on converging music styles as suggested by various studies might shed light on music consumption patterns and its relationships with overall sentiment and economic activity, both from the perspective of music firms as well as consumers.

[1] For example, Spotify currently uses 2,156 distinct genres for their recommendation system, see Every Noise at Once

Chapter 9

Appendix

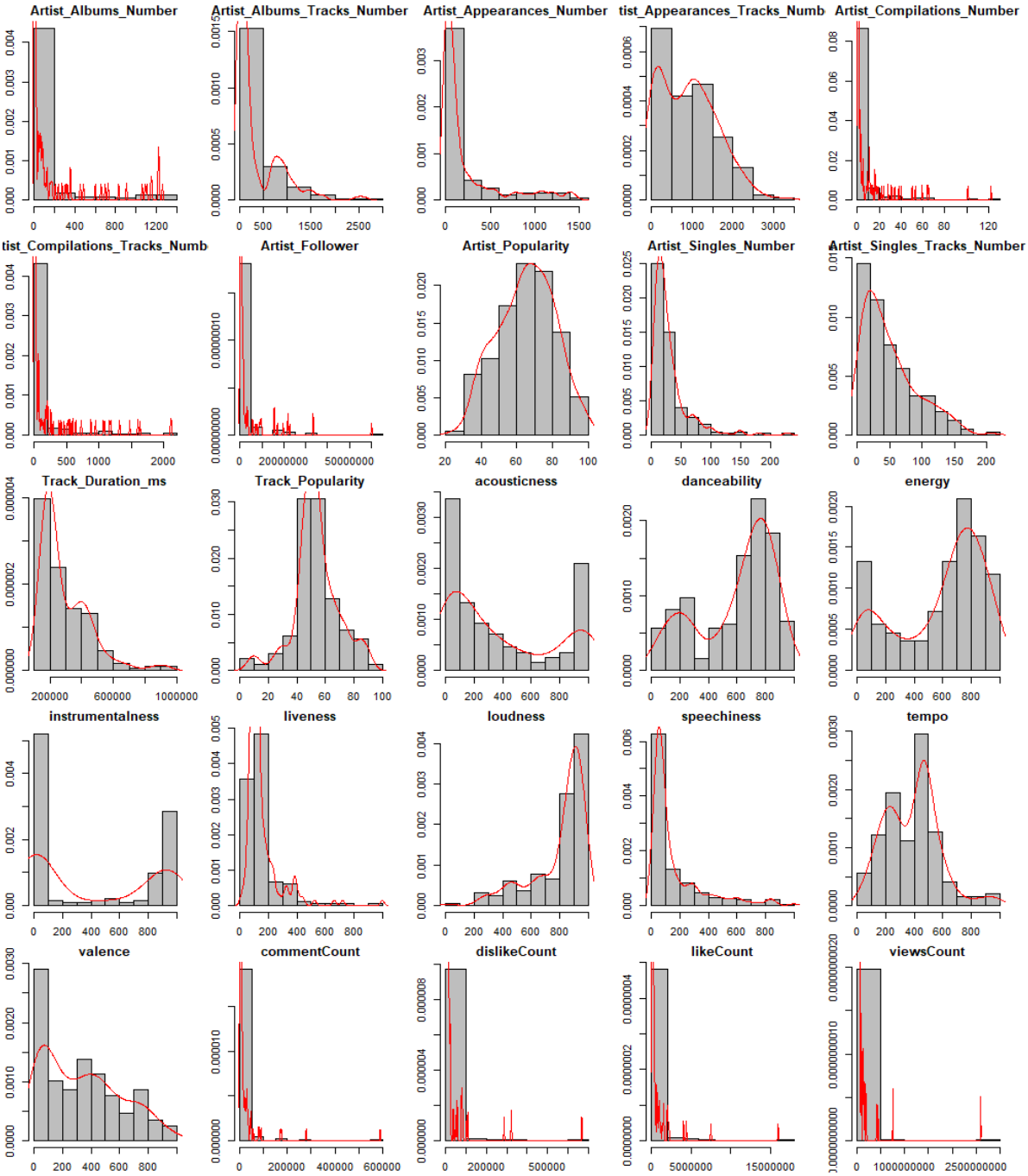


Figure (9.1) Histograms and kernel density plots of continuous variables

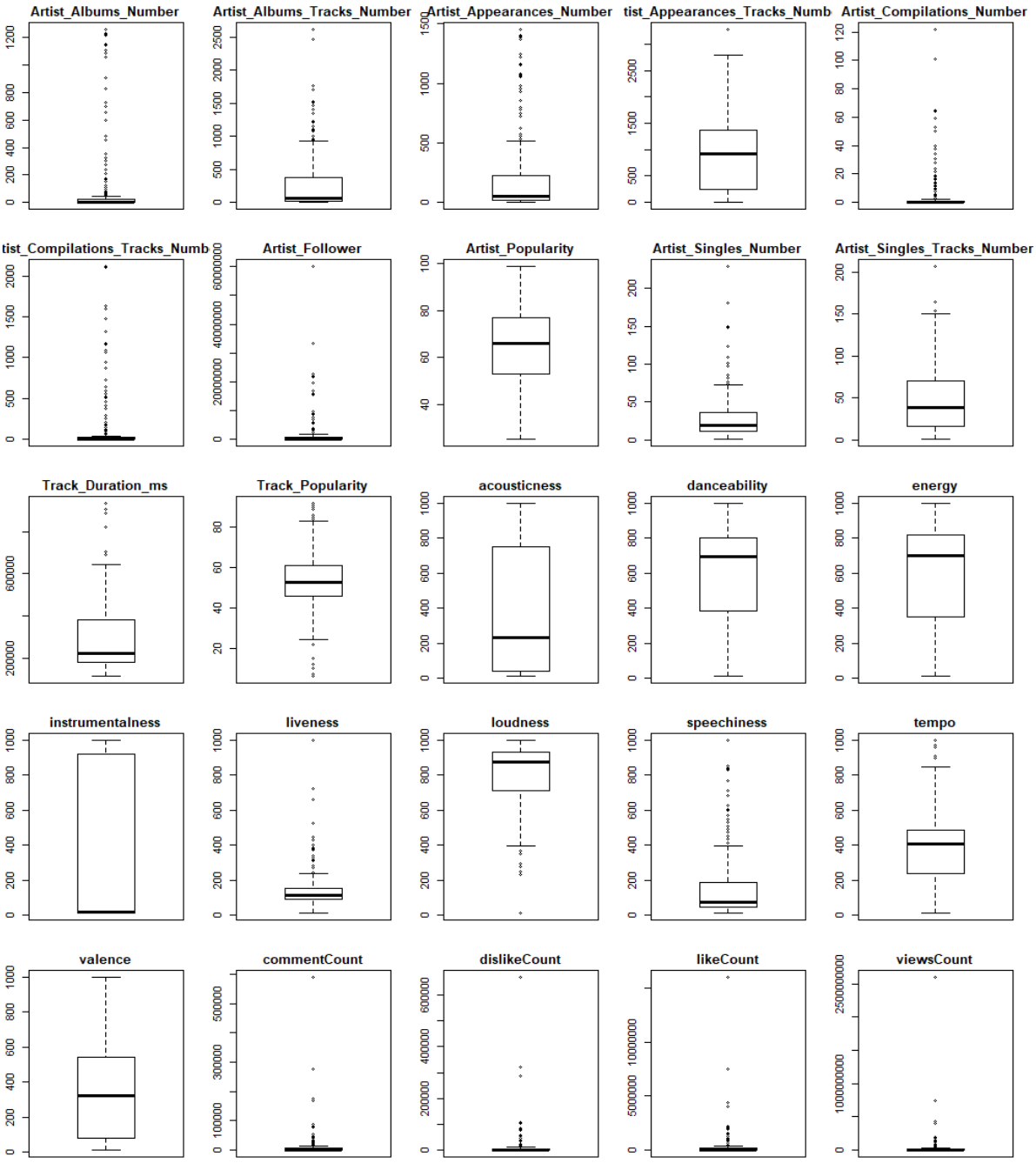


Figure (9.2) Box plots of continuous variables

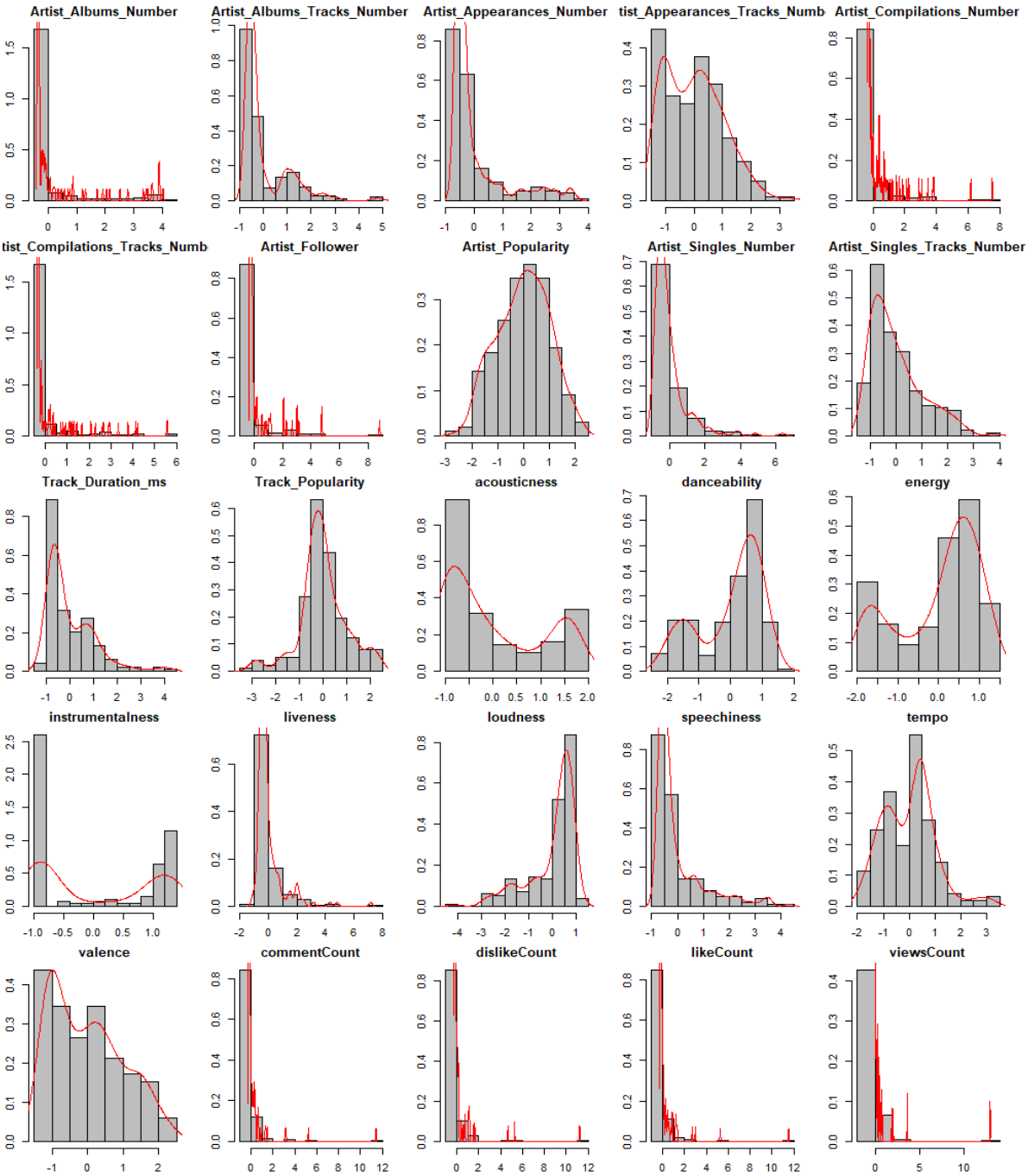


Figure (9.3) Histograms and kernel density plots of continuous variables (standardized)

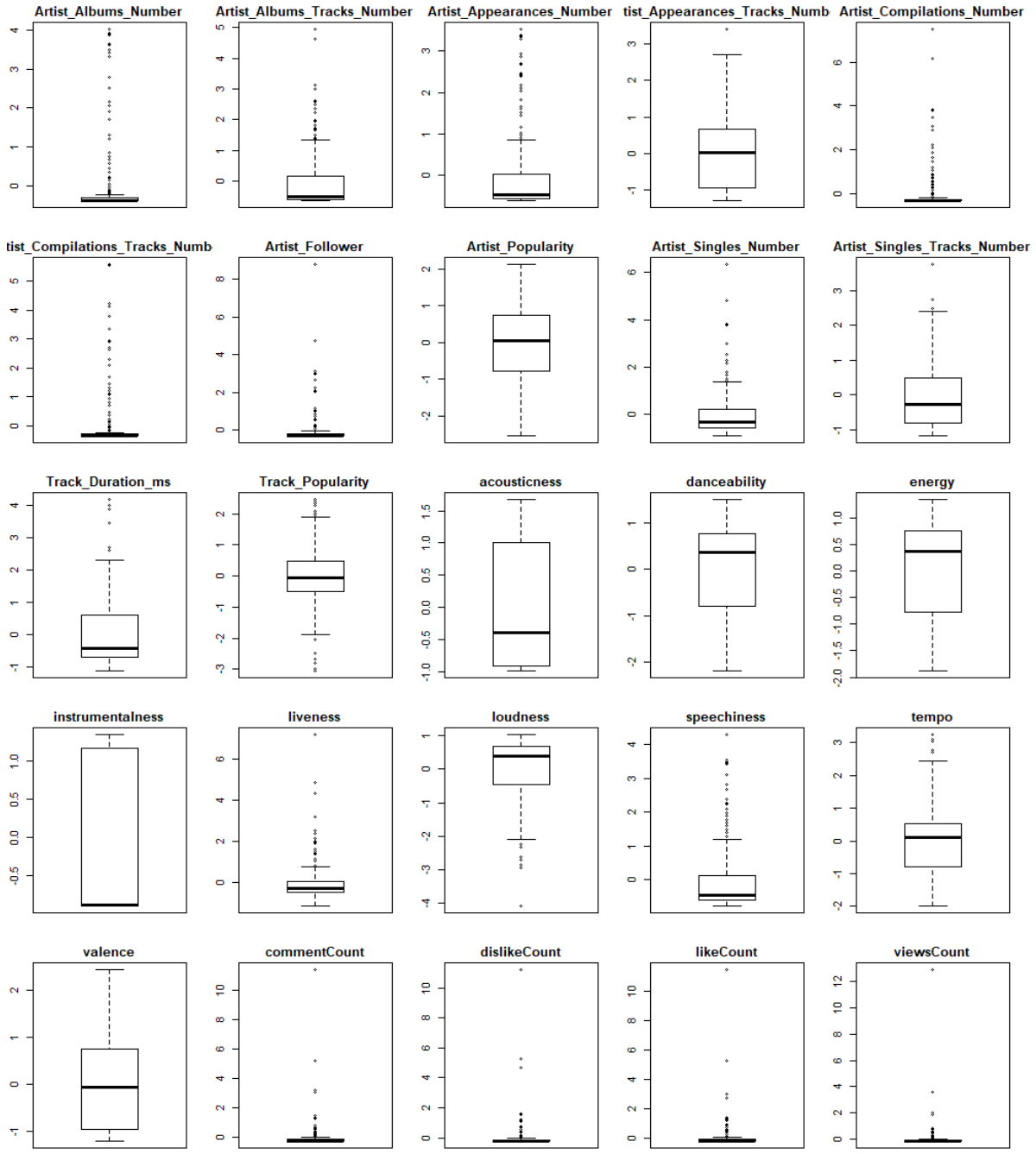


Figure (9.4) Box plots of continuous variables (standardized)

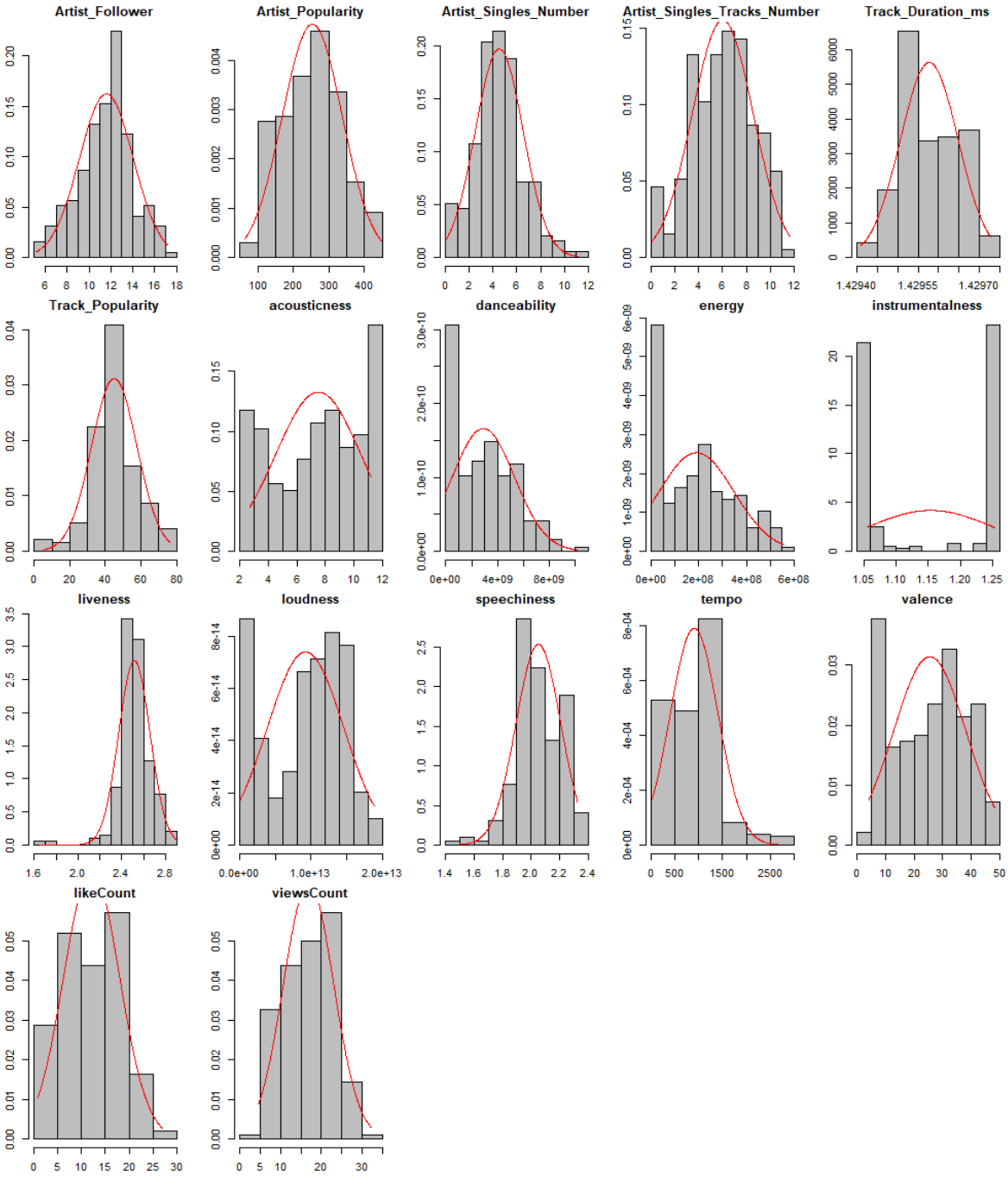


Figure (9.5) Histograms and kernel density plots of continuous variables (Box-Cox transformed)

Figure (9.6) Box plots of continuous variables (Box-Cox transformed)

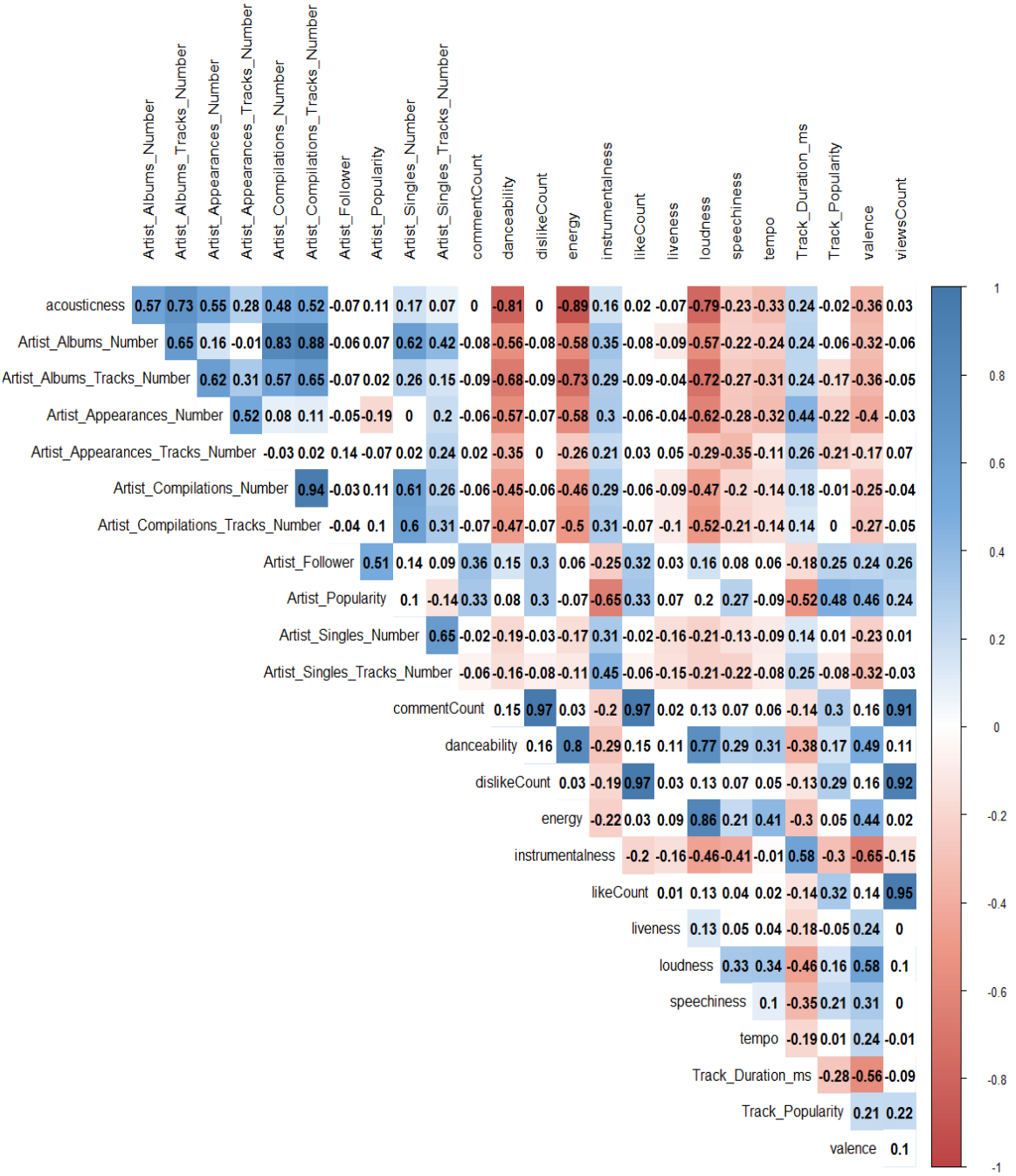


Figure (9.7) Correlogram for Bravais-Pearson correlation coefficient

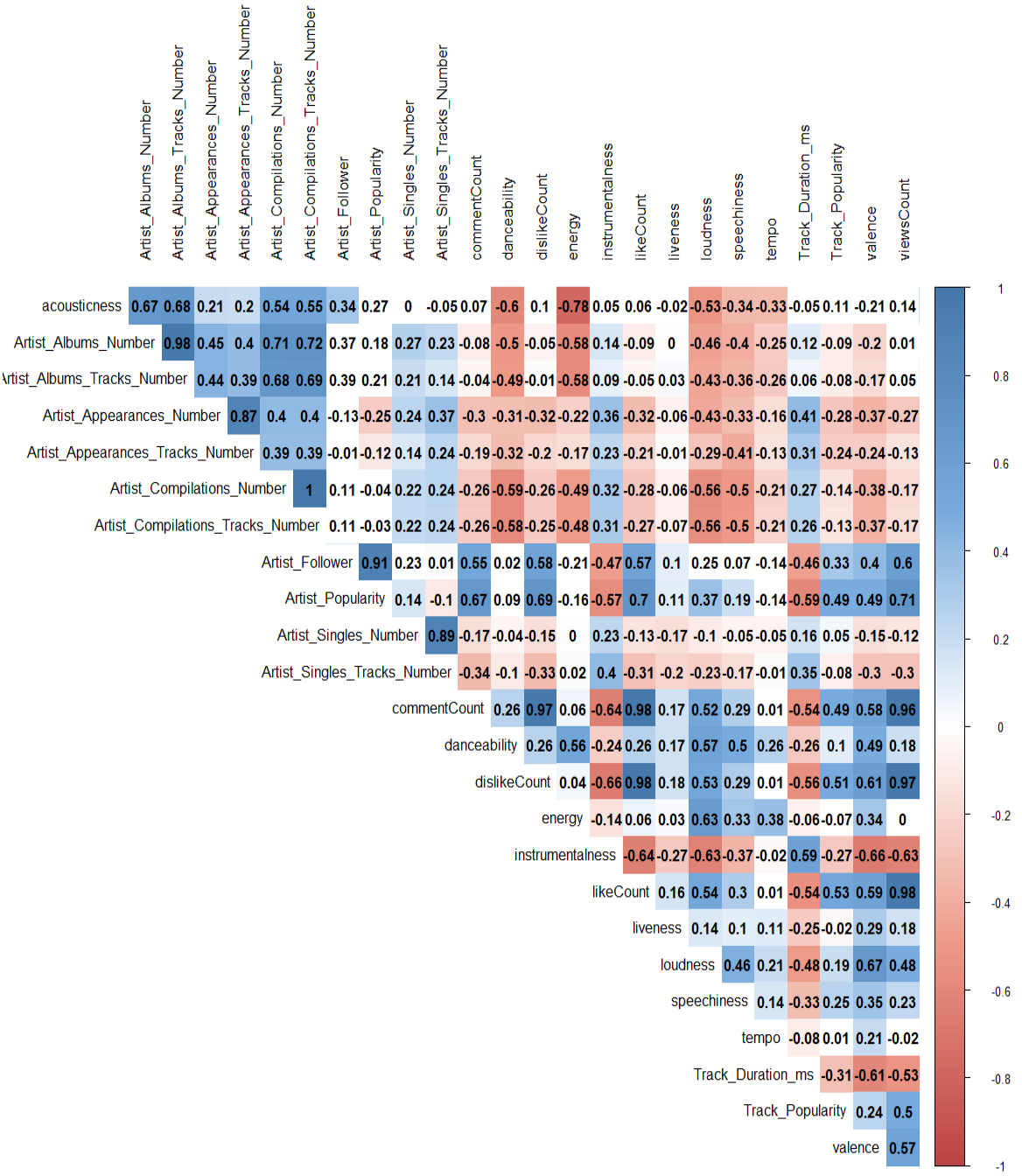


Figure (9.8) Correlogram for Spearman rank correlation coefficient

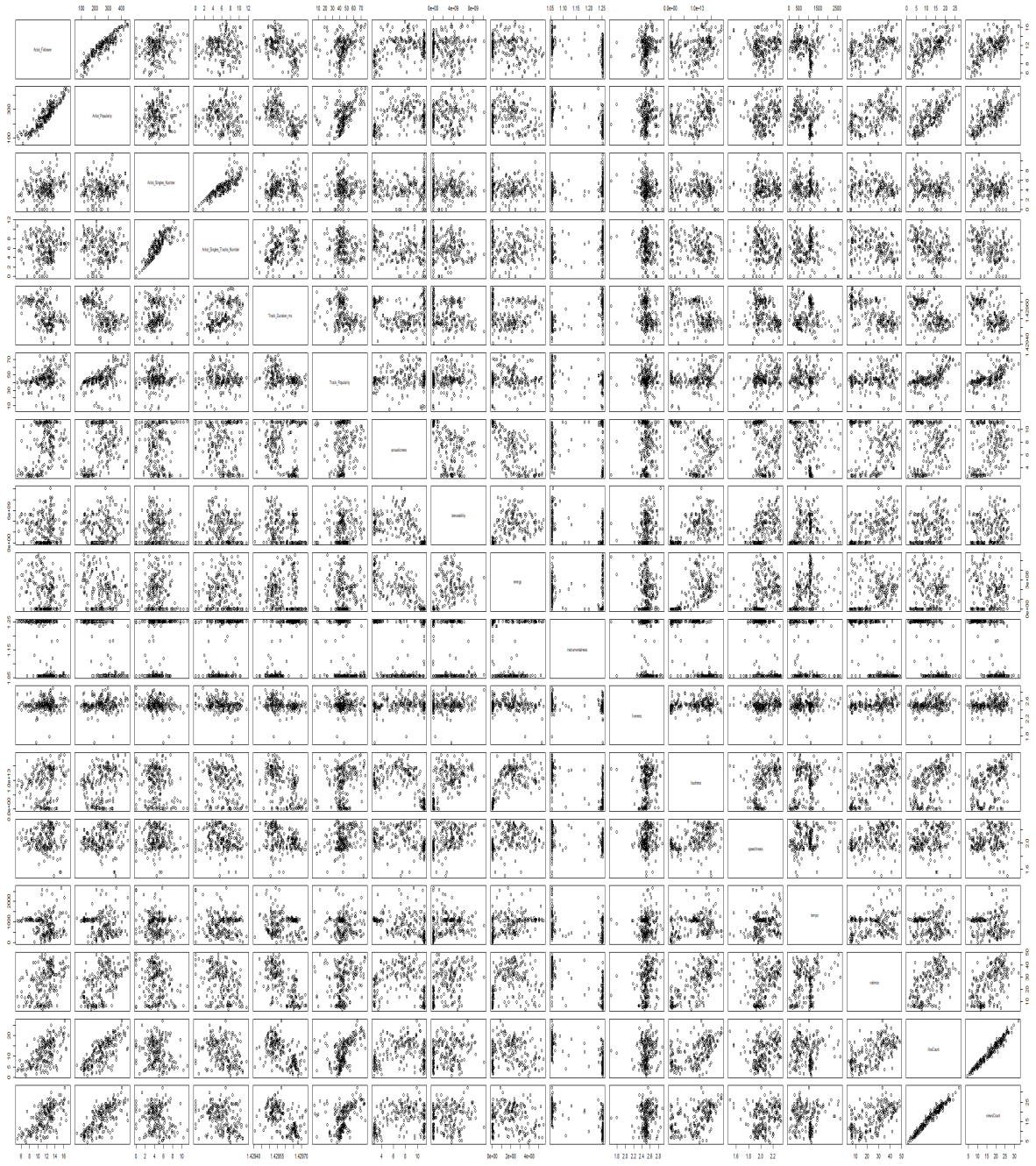


Figure (9.9) Pairs plot of Box-Cox transformed variables

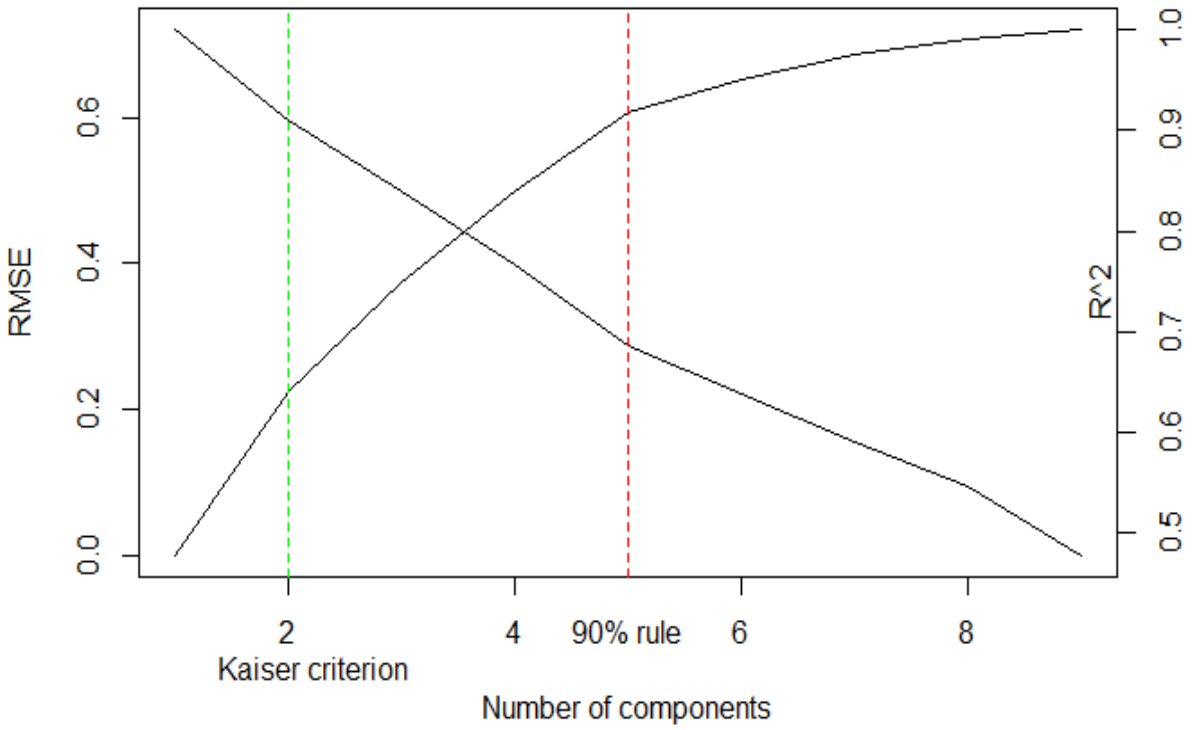
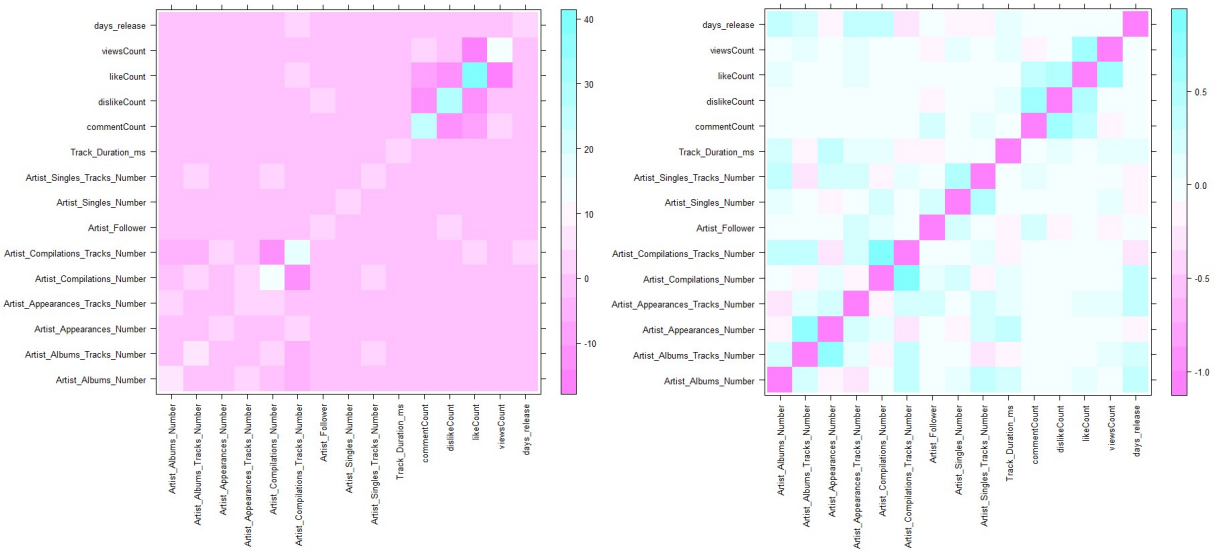


Figure (9.10) RMSE and R^2 for different number of components

Figure (9.11) FA suitability diagnostics



(a) Inverse correlation matrix

(b) Partial correlations

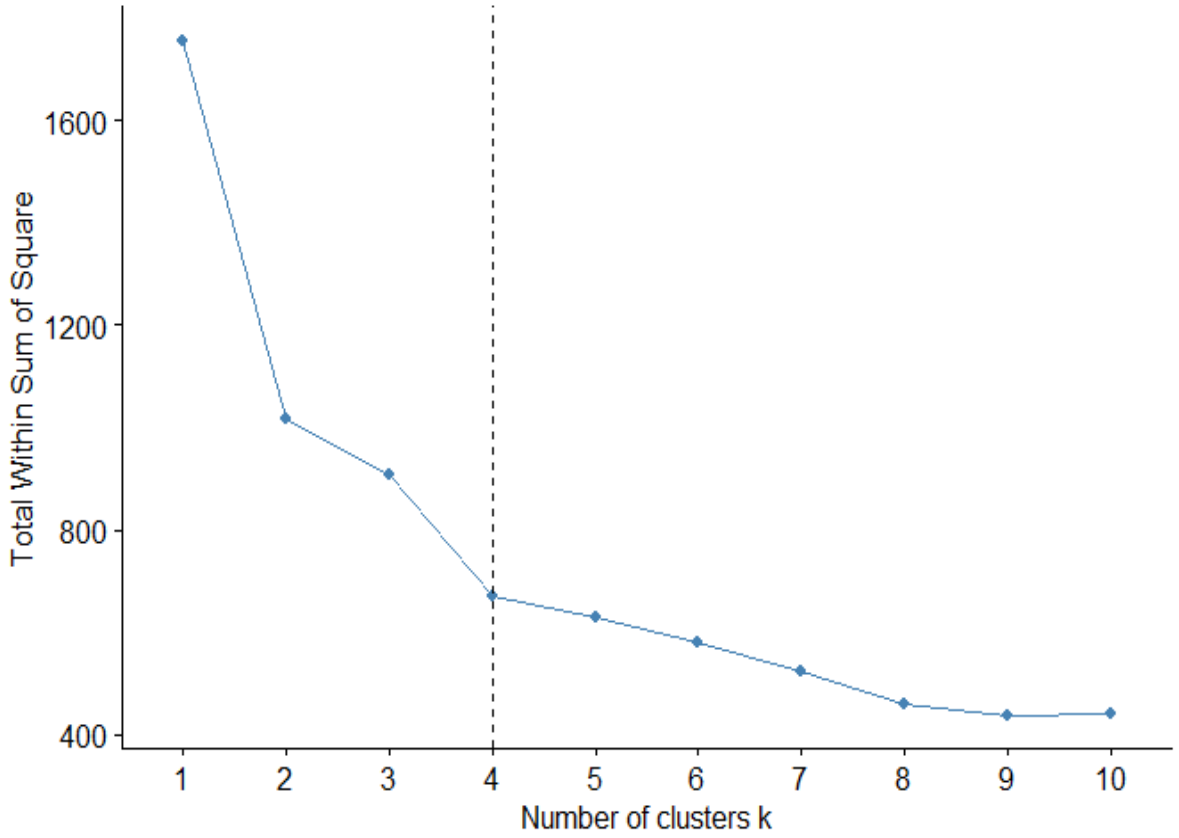


Figure (9.13) Scree plot k-means

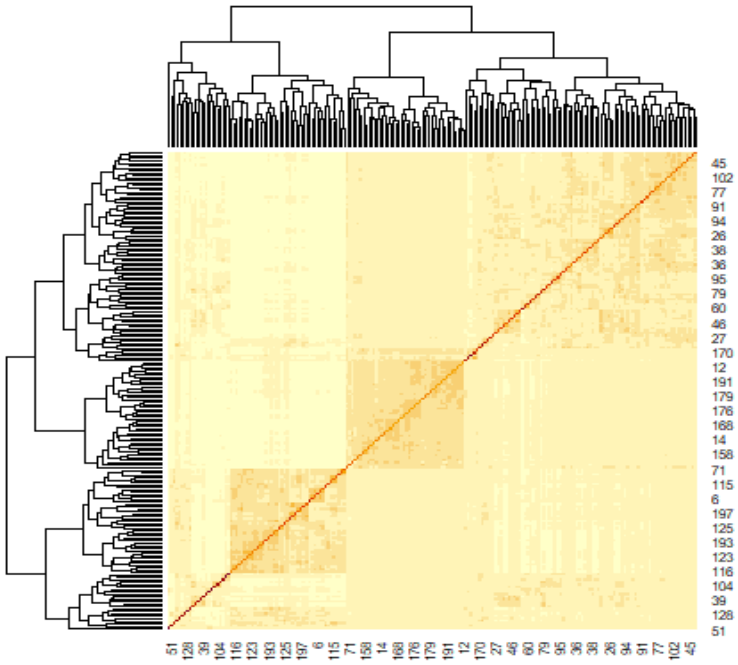


Figure (9.14) Similarity heatmap

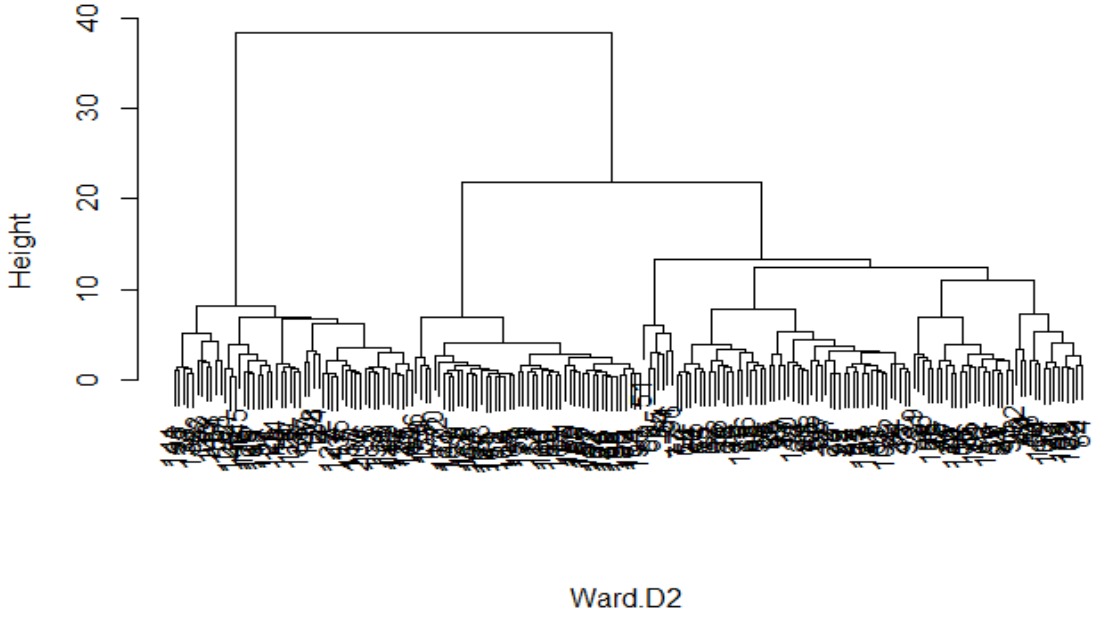


Figure (9.15) Dendrogram (Ward.D2)

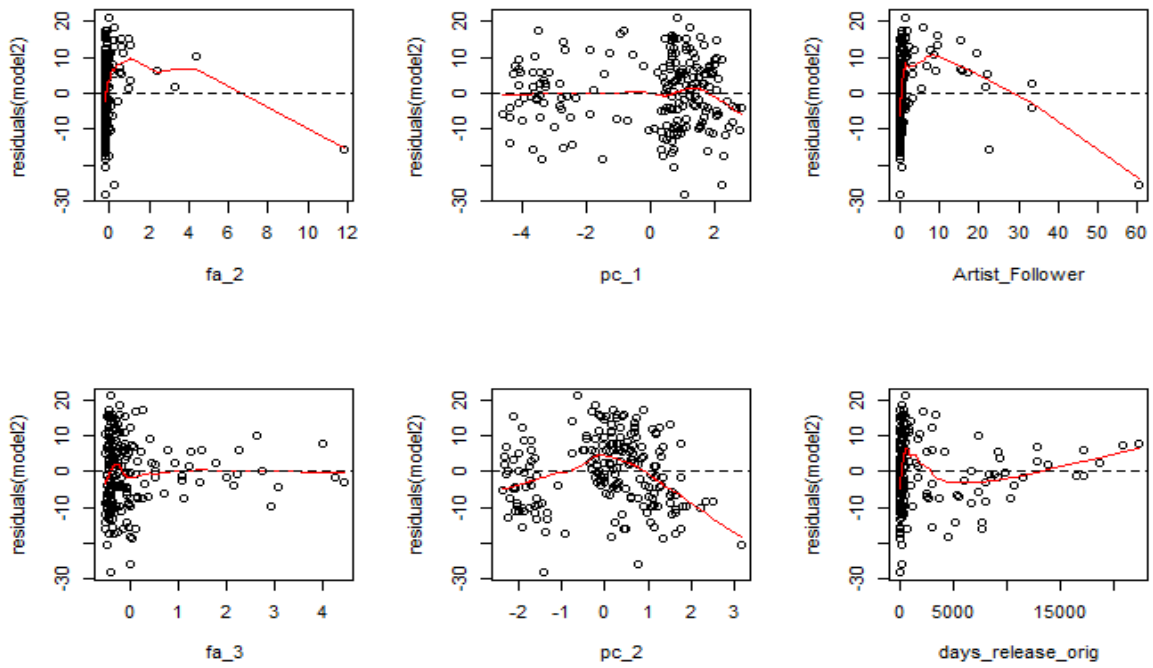


Figure (9.16) Residuals vs. predictors (`model2`)

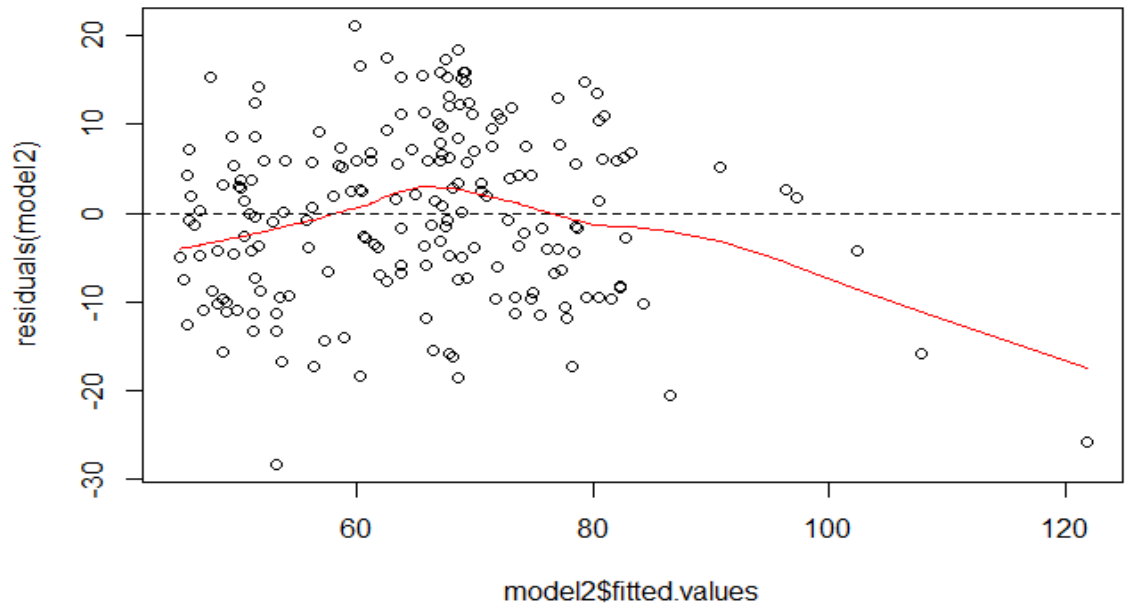


Figure (9.17) Residuals vs. fitted values (`model2`)

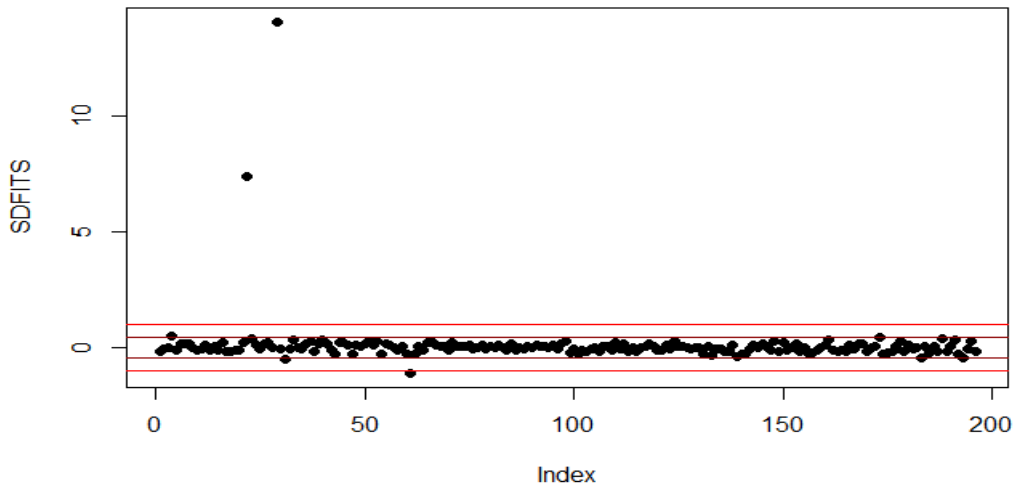


Figure (9.18) Regression deletion fit

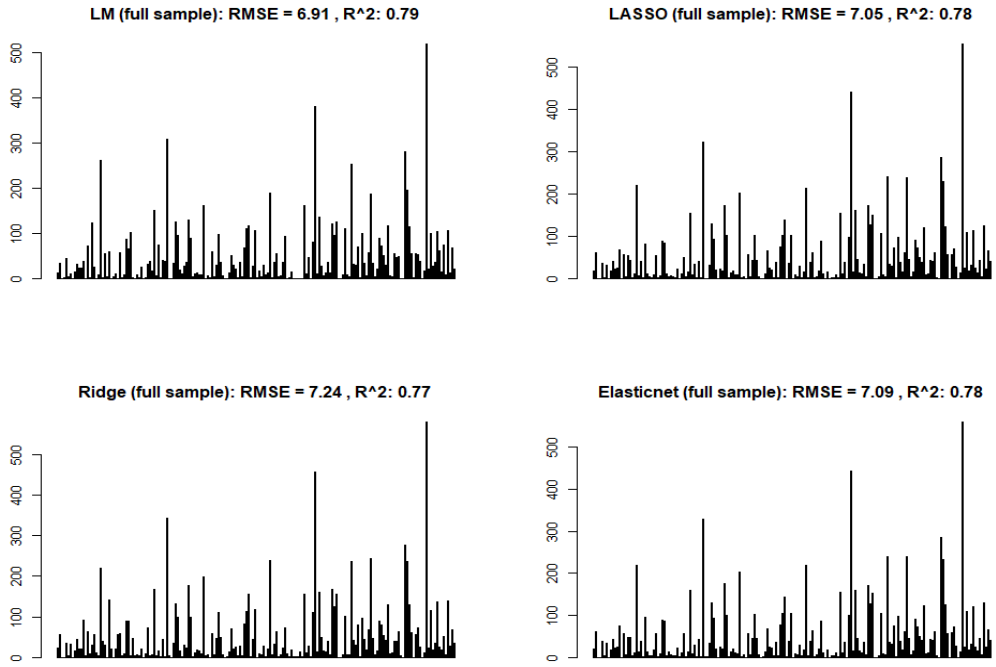


Figure (9.19) Residuals (full sample)

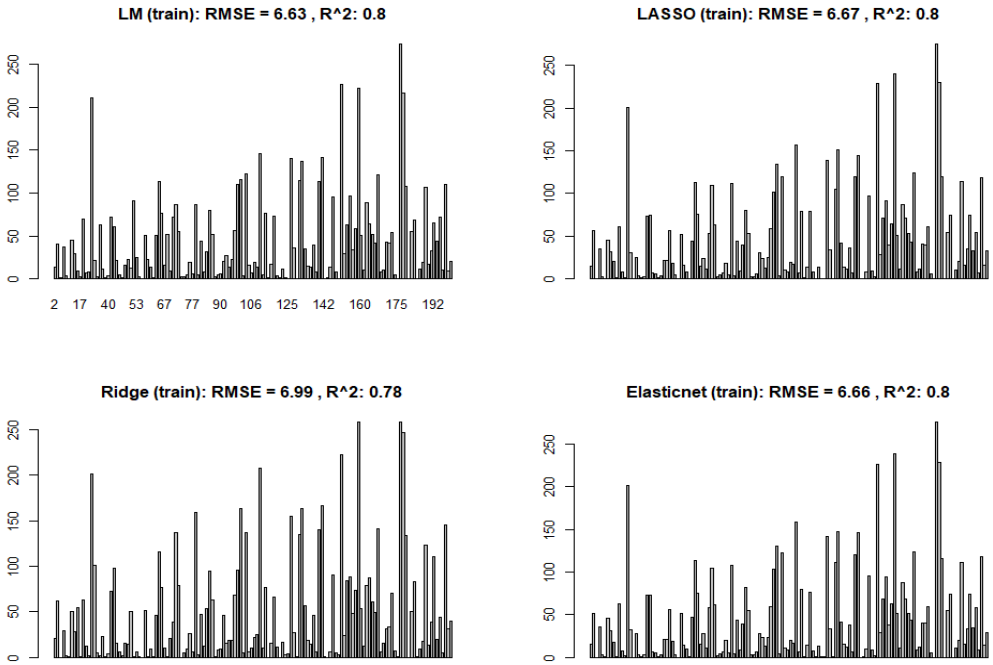


Figure (9.20) Residuals (train sample)

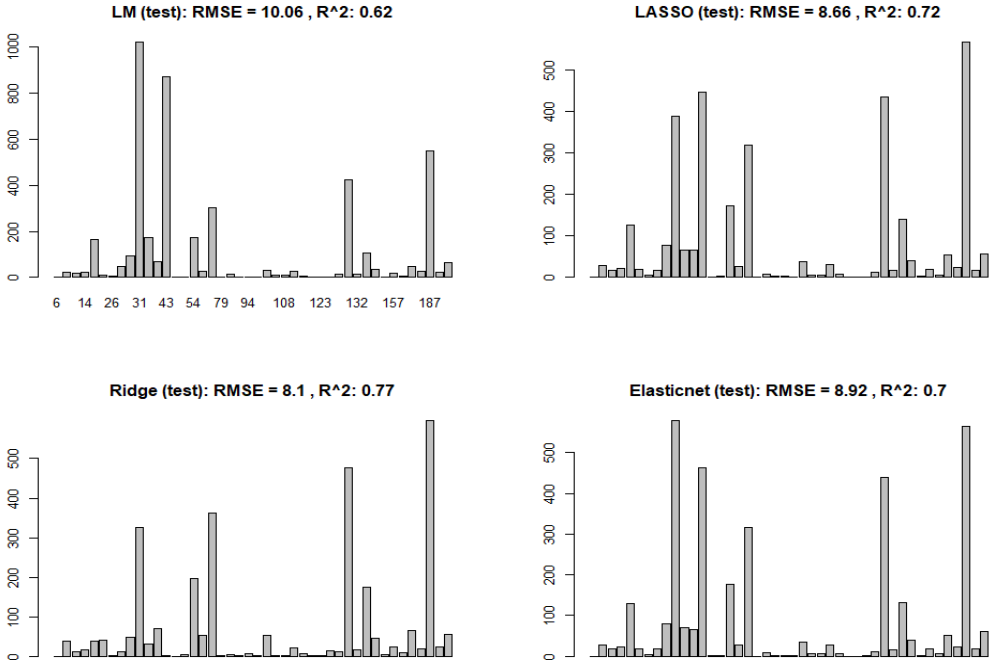


Figure (9.21) Residuals (test sample)

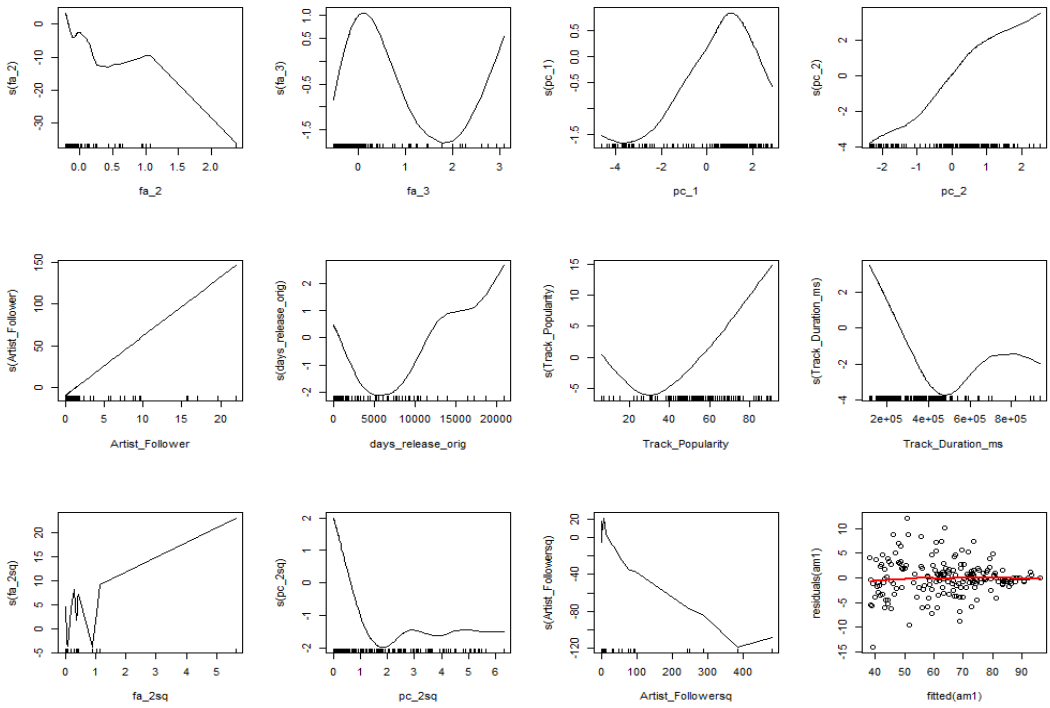


Figure (9.22) GAM: smoothed against predictors

Bibliography

Beat (2020). Spotify, napster und co: So viele streams brauchen künstler für einen euro - manager magazin. https://www.manager-magazin.de/unternehmen/artikel/spotify-napster-und-co-so-viele-streams-brauchen-kuenstler-fuer-einen-euro-a-1305050.html. (Accessed on 03/15/2020).

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *Ann. Math. Statist.*, 25(2):290–302.

Breiman, L. (2003). *Setting Up, Using, And Understanding Random Forests*. University of California, Berkeley. (Accessed on 03/15/2020).

Dunn, K. (2015). Process improvement using data. https://learnche.org/pid/contents. (Accessed on 03/15/2020).

Fisher, W. A. (1929). What is music? *The Musical Quarterly*, 15(3):360–370.

Guadagnoli, E. and Velicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological bulletin*, 103:265–75.

Haldane, A. (2018). Will big data keep its promise? Data Analytics for Finance and Macro Research Centre, King's Business School (Accessed on 03/15/2020).

International Federation of the Phonographic Industry (2019). Ifpi global music report 2019. https://www.ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2019. (Accessed on 03/15/2020).

Jehan, T. (2005). *Creating Music by Listening*. PhD thesis.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

Nomikos, P. and MacGregor, J. F. (1995). Multivariate spc charts for monitoring batch processes. *Technometrics*, 37(1):41–59.

Stevenson, A. (2010). *Oxford Dictionary of English*. Oxford University Press.

Székely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, 4(1):447–479.

Thompson, A. and Daniels, M. (2018). Are hit songs becoming less musically diverse? https://pudding.cool/2018/05/similarity/. (Accessed on 03/15/2020).

Thurstone, L. L. (1944). Second-order factors. *Psychometrika*, 9(2):71–100.