

A Bayesian Analysis of Spotify Data

Nathaniel Maxwell, Jessie Bierschenk

01 May, 2021

Introduction

Music-making is often thought of as an artform—a subjective expression that falls into a specific “genre” according to its musical attributes. Beginning in the 1960s, pop music had been dominated by the verse-chorus form where “the verse sets the scene, the pre-chorus builds tension, and the chorus reaches a climax,” with the cycle predictably repeating itself [5]. This musical formula dominated the industry; in fact, “music theorist Jay Summach has found that by the end of the 1960s, 42 percent of hit songs used verse-chorus form. By the end of the 1980s, that figure had doubled to 84 percent” [5].

With the advent of the 21st Century, however, the digitization of music production paired with the introduction of streaming platforms has warped the fundamental structure of songs. On popular media platforms, only snapshots of songs reach the ears of the public: five-second memes, 15-second TikToks, or 30-second ads. The limitless access to songs on streaming platforms has changed the landscape of song-making—“the gist of it: songwriters now get to the good stuff sooner” [2]. This phenomenon exists as increasing accessibility of songs results in decreasing revenue for artists. “Artists are paid per play—provided the listener stays tuned for at least 30 seconds. Each stream earns a tiny fraction of a cent. And just 13% of that goes to the songwriter, says David Israelite of the National Music Publishers Association” [Economist]. In turn, for an artist to make a decent living, their songs need millions of plays.

For many musicians, the art of composing/performing/marketing a new song is an arduous process. Even after all the work has been completed and a song is ready to be played to the public, the biggest uncertainty still awaits: How will the song be received? Will it become a hit? Will it be a song that everyone skips over, or never becomes popular? The purpose of this analysis is to investigate which characteristics of a song (such as tempo, duration, mode, acousticalness, etc.) would make it more “likeable,” less likely to be skipped, or more popular. Of course, music taste is a very subjective matter, and thus, there will be quite a bit of uncertainty around any variables that are deemed important/unimportant. What one person likes; another person may dislike. Therefore, looking at such musical characteristics through a Bayesian lens will help to quantify the uncertainty surrounding any of our findings. Through this analysis we hope to provide some conclusions that an aspiring musician (or even a well-established musician) can have at their disposal when creating new music.

These findings beg the question: so what are the features that make a song popular or appealing to a listener? These answers would be valuable for any musician seeking success in today’s music industry.

Pre-Analysis

Data

Data set 1: In our initial data set, we take a sample of 83,939 tracks from 167,881 released by Spotify that document musical attributes of the track, as well as if it was skipped by a listener on Spotify or not. The set of 167,881 observations was a sample of a full set of over 30 million observations. The original purpose of

the full data set was to analyze track attributes in order to predict whether a track would be skipped by a listener in the future. The tracks within the data set were not confined to any prerequisite of genre or form; perhaps some were not even songs, but rather audiobooks or podcasts. We wanted to attempt to narrow the track selection to represent only songs, and see if any variables could impact the “popularity” of a track. The explanation of the narrowing process is explained in detail in the variable explanation section.

Because this data set was not confined to an individual “taste” nor genre, limitations of this data were that song features may not be determined as significant due to the broadness of the data collected. Furthermore, there was no genre assigned to each data point, so it was impossible to narrow the data to be more specific to a particular sub-industry. Because of this, we recognized that in order to create a more meaningful report, we would have to narrow our focus and our data.

Data set 2: With the limitation of the first data set, we decided to narrow our focus and analyze data that pertained to one specific individual that recorded his tastes for 2,017 songs. For this second dataset, the size was not as big because it presented the opinions of a single individual. George McIntire assembled this data when exploring the explanation behind his varying taste in music. He created two playlists—each with about 1,000 songs—one with songs he liked and the other with songs he did not like [4]. In order to minimize bias, he had to make the “BAD” playlist equally as diverse as the “GOOD” playlist by putting in songs of different form rather than an entire album of a single artist he did not like.

Before beginning our analysis of the data, we removed the name and the artist of each song to avoid an unnecessary amount of dummy variables. Furthermore, as a new musician cannot attempt to be a different artist, that information is not important to our analysis. However, as a side note to keep in mind when interpreting, in the original data it was clear that popular artists such as Drake and Young Thug frequented list. We also decided to exclude the variable “key”, since only about 1 in 10,000 people have perfect pitch, and thus only .01% of the population would be able to identify the key [3].

Limitations of this data can be attributed to the source being a single individual as well as the possibility that “taste” in music can not be fully associated with the variables in the dataset.

Data set 1 Importation Method: The initial dataset came in two files. The first file contained the track id and whether or not the track was skipped (167,881 observations), and the second file contained the attributes of the tracks (of which there were 50,704 distinct tracks). Given that Rstudio only takes file sizes of less than 5MB, we extracted every other observation to cut the first file size in half to 83,939. For the second file, we discarded variables that would have no interpretation, and given that the file was still above 5MB, we split it into half, and imported both of them, along with the first file, into Rstudio. Finally, we combined the three files all into 1 file to obtain 83,939 observations containing track attributes and whether or not that track was skipped.

Data set 2 Importation Method: The second data set used came from an individual who was passionate about exploring his personal tastes in music. He had assembled a data set of 2,000 songs: half of which he “liked” (“target”= 1), and the other half which he “did not like” (“target”=0) The data needed very little manipulation besides scaling the numeric, continuous values and factoring the variables “mode”, “time_signature”, and the response variable “target”.

Data set Descriptors: Here are the descriptors of variables for the two data sets. The variables in the second data set without descriptions had the same descriptions as the duplicate variables in the first data set.

1. The first dataset consists of the response variable “skipped” (1 if skipped and 0 if not), and is dependent upon the variables containing musical attributes, which will be described below. Each track has the following characteristics:
 - (a) Release Year (Year the song was released)
 - (b) Duration (length of song in seconds)
 - (c) US Popularity Estimate (A popularity rating of song, on a scale 1-100)

- (d) Acousticness (A confidence measure from 0-1 on whether the track is acoustic, where values near 1 represent high confidence that the track is acoustic)
- (e) Beat Strength (The strength of the beat from 0-1, where 1 represents a very strong sense of beat)
- (f) Bounciness (A rating of the bounciness from 0-1, where 1 represents a strong sense of bounciness)
- (g) Danceability (A rating from 0-1 of how suitable the track is for dancing, where values near 1 represent high suitability)
- (h) Energy (A rating from 0-1 representing a perceptual measure of intensity and activity, where values near 1 represent high energy)
- (i) Instrumentalness (A rating from 0-1 that predicts whether a track has no vocals, where values close to 1 represent high confidence that there are no vocals)
- (j) Mode (Predicts whether or not a song is major or minor)
- (k) Speechiness (A rating from 0-1 that detects the presence of spoken words in a track, with values near 1 representing an exclusively speech-like track)
- (l) Tempo (The estimated tempo of the track in Beats Per Minute (BPM))
- (m) Valence (A rating from 0-1 that represents the positivity of the song, with 1 representing high positivity)
- (n) Skipped (Denotes whether or not that particular track was skipped or played the entire way through)

Note: In order to try to obtain tracks most representative of new music, only the following tracks were kept (total of 65,417 observations kept):

- (a) Tracks from 2010-present
 - (b) Tracks with a speechiness value ≤ 0.4 (filters out tracks that are mostly spoken, such as podcasts and ebooks)
 - (c) Tracks with an instrumentalness value ≤ 0.6 (filters out tracks that contain no vocals)
 - (d) Tracks with a duration ≤ 360 seconds (given that the average new song is 3-5 minutes, a cutoff of 6 minutes seemed appropriate) Furthermore, the variables year and track id were not utilized in the analysis, as their information provides no benefit to a songwriter.
2. The second dataset consisted of 2017 songs compiled by a single person, where a portion of the songs are songs that he likes, and the other portion are songs that he dislikes. This dataset includes similar variables as the first dataset, including:
- (a) Acousticness
 - (b) Danceability
 - (c) Duration
 - (d) Energy
 - (e) Instrumentalness
 - (f) Liveness (rating from 0-1 of whether the track was performed live, with 1 representing high confidence the track was performed live)
 - (g) Loudness (Overall loudness of the track in decibels (dB))
 - (h) Mode
 - (i) Speechiness
 - (j) Tempo
 - (k) Time Signature (The way in which beats of the song are organized)
 - (l) Valence

Data Set 1

Exploratory Data Analysis

When observing the variables of the dataset, nearly all of the histograms of values were approximately normal, aside from a few that had a very heavy left tail or right tail skew. Furthermore, a correlation plot can help give us an idea for potential interaction terms, depending on if the single variables are significant by themselves. This plot can be seen in the appendix. Beat Strength, Bounciness, and Danceability all have very high correlation values. If those variables end up being significant, it will be interesting to see if/how they interact.

Model Selection

For the first dataset, our response variable, \mathbf{y} , will be modeled by a *Bernoulli*(θ) distribution, where 1 means the track was skipped, and 0 means the track was not skipped. To obtain the variable θ , we will use the logit link, where $\text{logit}(\theta) = \eta$, and $\eta = \mathbf{x}^T \boldsymbol{\beta}$, where $\mathbf{x} \in \mathbf{X}$ is the covariate space for \mathbf{Y} . In other words, we are creating a logistic regression model where θ is obtained from a linear model. We can write our sampling distribution for \mathbf{y} as $[\mathbf{y}|\theta]$, and since θ is dependent upon $\boldsymbol{\beta}$ and \mathbf{x} , we can write this as $[\mathbf{y}|\boldsymbol{\beta}, \mathbf{x}]$.

What we really want is to estimate the values of the coefficients $\boldsymbol{\beta}$ for each of the variables to find out how they impact whether or not a track is skipped. In order to form a posterior estimate for $\boldsymbol{\beta}$, we must specify a prior. We assume little knowledge about each variable's effect, so we propose a weakly informative prior for $[\boldsymbol{\beta}]$: Using recommendations from Gelman, Jakulin, Pittau, and Su [7], we use a Cauchy(0,2.5) prior for each variable. Furthermore, we centered all the variables and then scaled them to have the same standard deviation, so that no variable could have a disproportionate effect on the outcome, and thus have better interpretability. Now that we have a prior model $[\boldsymbol{\beta}]$ and sampling model $[\mathbf{y}|\theta]$, we can use all $\mathbf{y}_i \in \mathbf{Y}$ to calculate the posterior $[\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}]$ using Baye's Theorem and proportionality. This will not be done by hand. Instead, using the rstanarm package, Rstudio will compute the posterior and draw MCMC samples from the $[\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}]$. In this way we will obtain estimates for the values of $\beta_0, \beta_1, \dots, \beta_k \in \boldsymbol{\beta}$.

Note: In order to allow efficient posterior sampling, a random sample of 5000 tracks were used.

Posterior Check and Estimate

After running the rstanarm function and including all of the variables, we see that there is only variable whose 95% confidence interval does not include 0. That variable is duration, and when calculating the 'leave-one-out' cross-validation information criterion (looic), we see that this model barely has a slightly lower value (6366) than the looic of a baseline model (6373) with no predictors. To attempt to find the best model, we will drop all variables that were not deemed significant at a 95% confidence interval (included 0 in their posterior interval), and rerun the model. In this case, 'duration' is the only variable remaining.

The model with only "duration" as a variable proved to be better, with a looic value of 6358. To ensure that the MCMC samples created from the stan_glm function converged and mixed well, a traceplot and ACF plot confirmed that they indeed did converge and mix well (See appendix for graphs). To calculate our posterior predictive accuracy, we used the following method. If the posterior probability of a track being skipped is greater or equal to 0.5, then we would predict that observation to have a value of 1 (and similarly for less than 0.5). For each observation, we can compare the posterior prediction to the actual observed value. The proportion of times we correctly predict an individual (i.e. [prediction = 0 and observation = 0] or [prediction = 1 and observation = 1]) is our classification accuracy. In our case, the posterior classification accuracy is 0.67. While we would really want to also calculate the estimated accuracy on "unseen" data, or data that doesn't actually affect this model, because our number of observations is so large (even using a sample), we would expect the value to be the same when using a LOOCV, and this is indeed true: the value is still 0.67.

Results

Given the 12 variables in this data set, only 1 of them turned out to be significant. Hence, we will provide an interpretation of the variable “duration” so that it can be utilized when musicians are creating new music.

The posterior estimate for $\hat{\beta}_{duration}$ is 0.126, and the 95% confidence interval for this variable is (0.066, 0.185). We will not provide a literal interpretation for the variable, for two reasons: 1, the data is scaled, making interpretation much harder, and 2; a musician is ultimately looking to create an art form for people to enjoy, and given the subjectiveness of art appreciation, a literal interpretation of the variable wouldn’t hold much weight in the real world. Instead, it is sufficient to know that as the duration of a song increases, the likelihood that it will be skipped increases as well. In other words, the longer a song is, the less likely a listener will stay engaged with it. This makes sense on an intuitive level, and supporting evidence for this result will be presented in the conclusion.

Conclusion

Our analysis of this dataset turned out to be rather inconclusive. It only returned one variable as significant, and that was duration. In effect, it stated that the longer the track is, the more likely it is to be skipped. Other than that, there were no other findings. This lack of conclusive analysis could be explained in a variety of ways. One, we do not have any guarantee that the tracks analyzed were actually songs. While we applied filters to attempt to filter out any tracks that were not songs, we still do not actually know the content of the track. So the uncertainty surrounding the track content is one factor that is certainly a model limitation. A second reason behind inconclusiveness is the fact that music taste is a very subjective field, and what one person likes, another person may not. Given that this data is pulled from a multitude of Spotify users, there is going to be a wide range of musical preferences. Therefore, it would be much harder to find any nuances in music style that might affect the broader population’s tendency to skip or not skip a song. An interesting note is that the duration of a song, our only significant variable, has been decreasing when looking specifically at the #1 song by year in the last few years, according to the New York Times [5]. This overall trend for all songs is backed by an article from the Economist that echoes that same sentiment [2]. It is clear, then, that this trend is evident in a Spotify user’s likelihood to skip or not skip a song.

Data Set 2

Exploratory Data Analysis

We have also included a correlation plot for the second data set. There were not any apparent correlations between variables aside from a possible correlation between loudness and energy. And a possible negative correlation between energy and acousticness.

Model Selection

We will use the exact same procedure that was used for the first data set to obtain posterior estimates for β , as well as create a model of best fit.

Before our model selection, we saw that the distribution of the “time_signature” factors in the posterior mean graphs were centered around zero with disproportionately high degrees of deviation relative to the other variables. Therefore, we omitted “time_signature” from our full model before beginning our model selection process. In addition, this decision was fueled by the need to save computational costs, as running LOOIC values for many models is very time expensive. To begin our modeling selection process, we used a backwards selection process where we systematically omitted one of the 11 variables from the full model. We then compared the full model (with all variables included) and the other models (each with one variable removed) by using their LOOIC (leave one out information criterion) values. Using the LOOIC produces

values that can be used to compare models to find which has the highest strength of explanation for the given response variable. Whichever model produced the lowest LOOIC compared to the other models and the full model then became the new full model. Using this process, we first omitted the variable “energy.” Now using this model (with only energy omitted) as the full model, we repeated the steps outlined above and found that this new full model had the lowest LOO value. Once we had our full model, we wanted to test for possible interaction terms. The interaction terms in question were “valence x mode”, “duration x danceability” and “acousticness x liveness”. We tested the valence and mode interaction because a minor mode is often associated with a “sad” sound, so we wanted to explore if there is an association with minor key and a lower perceived value of positivity (lower valence). We tested the duration and danceability interaction because we wanted to explore if more “danceable” songs had a tendency to be longer or shorter songs. We tested the acousticness and liveness interaction because we believed that the more likely the song was performed live, the less likely a song would be electronic-based (it would probably tend to be more acoustic). We repeated the steps outlined above, comparing the LOO value of the full model that included the interaction terms with the value of the full model without the interaction terms. We found that the full model that included the interaction terms had a lower LOO value so we adopted it as our new full-model. After looking at the posterior mean distributions of the variables for this model, however, we found that the interaction term between mode and valence seemed to be distributed around zero. We proceeded further and used the LOO comparison method and compared the full model with all 3 interaction terms and the full model with all interaction terms excluding mode x valence. After this comparison, we found that the full model with all interaction terms excluding mode*valence had the lower LOO value, so we adopted this as our final model.

Posterior Check and Estimate

To ensure that the MCMC samples created from the `stan_glm` function converged and mixed well for each variable, traceplots and ACF plots confirmed that they indeed did converge and mix well (See appendix for graphs). To calculate our posterior predictive accuracy, we used the same method used for data set 1 [see above]. If the posterior probability of the song being liked for an particular song is greater or equal to 0.5, then we would predict that that song to have a value of 1 (and similarly for less than 0.5). For each observation, we can compare the posterior prediction to the actual observed value. The proportion of times we correctly predict an individual (i.e. [prediction = 0 and observation = 0] or [prediction = 1 and observation = 1]) is our classification accuracy. In our case, the posterior classification accuracy is 0.688. We would really want to calculate the estimated accuracy on “unseen” data, or data that doesn’t actually affect this model, but using LOOCV, this value is nearly the same: the value is 0.683.

Results

Our resulting model found from the LOOIC model selection is as follows:

$$\begin{aligned} \text{logit}(\theta) = & \beta_0 + \text{acousticness}\beta_1 + \text{danceability}\beta_2 + \text{duration}\beta_3 + \\ & \text{instrumentalness}\beta_4 + \text{liveness}\beta_5 + \text{loudness}\beta_6 + \text{mode1}\beta_7 + \text{speechiness}\beta_8 + \text{tempo}\beta_9 \\ & + \text{valence}\beta_{10} + \text{danceability} \cdot \text{duration}\beta_{11} + \text{acousticness} \cdot \text{liveness}\beta_{12} \end{aligned}$$

$\beta_1 = -0.348$ with a confidence interval of $[-0.473, -0.223]$,

$\beta_2 = 0.328$ with a confidence interval of $[0.218, 0.448]$,

$\beta_3 = 0.353$ with a confidence interval of $[0.218, 0.489]$,

$\beta_4 = 0.365$ with a confidence interval of $[0.249, 0.483]$,

$\beta_5 = 0.136$ with a confidence interval of $[0.038, 0.238]$,

$\beta_6 = -0.409$ with a confidence interval of $[-0.547, -0.280]$,

$\beta_7 = -0.167$ with a confidence interval of $[-0.370, 0.031]$,

$\beta_8 = 0.364$ with a confidence interval of $[0.260, 0.474]$,

$\beta_9 = 0.088$ with a confidence interval of $[-0.012, 0.187]$,

$\beta_{10} = 0.189$ with a confidence interval of $[0.080, 0.299]$,

$\beta_{11} = 0.374$ with a confidence interval of $[0.265, 0.492]$,

$\beta_{12} = 0.218$ with a confidence interval of $[0.099, 0.340]$

Using similar logic as the first data set, we will not provide literal descriptions of the variables but explain their significance and general effect on the response variable “target.” To begin, all variables included in our best fit model proved to be significant using the 95% confidence interval except for “tempo” and “mode.” The variables that had a positive effect on the outcome (a higher probability of a song being “liked”) within our best-fit model were danceability, duration, instrumentality, liveness, speechiness, valence, danceability x duration, and acousticness x liveness. In essence, this means that if a song is longer, is more “danceable,” has less vocals, is performed in a live setting, has more spoken words, is more “positive” in nature, then that song has a higher probability of being “liked” by this individual according to our model. In addition, the interaction effects indicate that a higher danceability value corresponds to a higher effect of duration on the probability of a given song being “liked” (and vice versa), and, likewise, a higher level of acousticness corresponds to a higher effect of liveness on the probability of a given song being “liked” (and vice versa). Furthermore, the variables that had a negative effect on the outcome (a lower probability of a song being “liked”) within our best-fit model were acousticness and loudness. In essence, this means that if a given song is more acoustic or louder, then that song has a lower probability of being “liked” by this individual according to our model.

Conclusions

After an inconclusive result with the first analysis, the narrowness of the second analysis allowed for a higher level of insight. The systematic process of comparing the LOO values of different models that omitted different variables allowed us to eliminate insignificant variables such as “energy”. Our final model suggested particular characteristics of a given song that would increase or decrease the likelihood of this specific individual “liking” the song. While the predictive power (about 68%), was not as high as we would’ve ideally wanted, the value is still higher than 50%, so it is more informative than simply guessing whether this individual will like a song or not. Ultimately this model could be used by this individual to predict whether he will like a given song (given its characteristics), or it could purely be informative for this individual when reflecting on his personal tastes. Furthermore, given that this individual stated that his taste in music seems to span across multiple genres and seems “inconsistent”, our model could be used by artists wanting to entice listeners outside of their usual following. An individual like the one featured in our analysis proves to be a prime example of someone that could be interested in listening to new types of music. By keeping in mind the effects of the variables in the model, a musician/musicians can tailor his/her/their songs to increase their chances of gaining popularity in the musical industry.

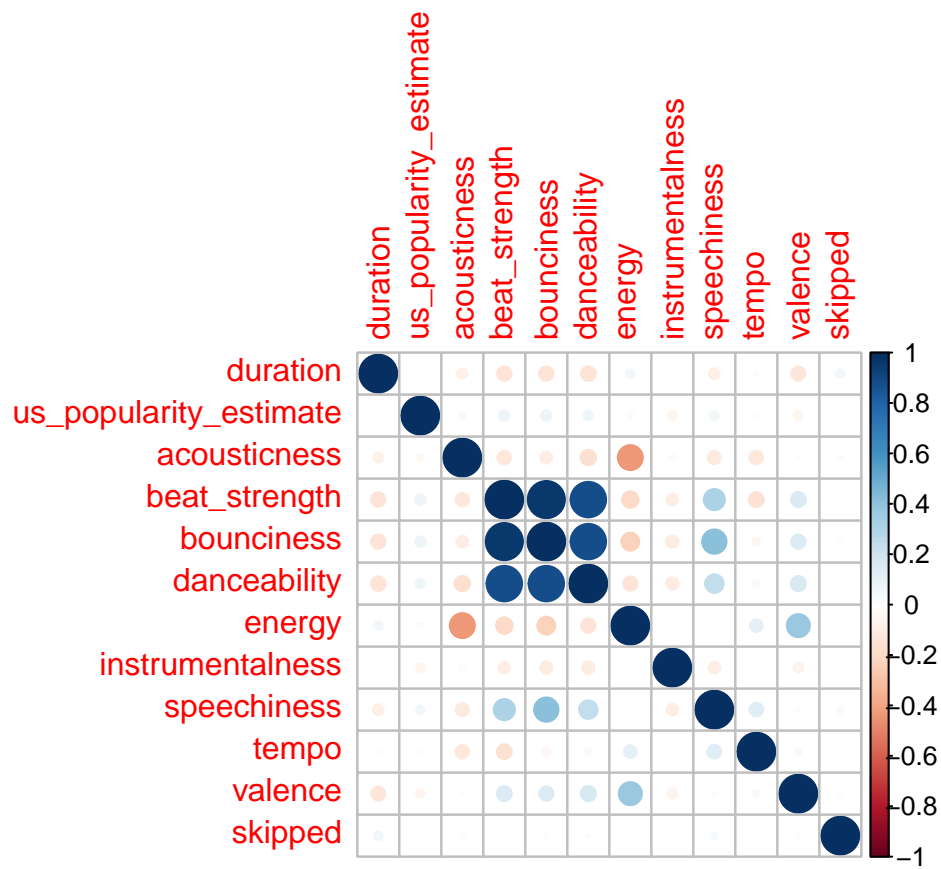
Sources

- [1] Betancourt, Michael. How the Shape of a Weakly Informative Prior Affects Inferences, mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html.
- [2] “The Economics of Streaming Is Changing Pop Songs.” The Economist, The Economist Newspaper, 5 Oct. 2019, www.economist.com/finance-and-economics/2019/10/05/the-economics-of-streaming-is-changing-pop-songs.
- [3] Gander, Kashmira. “Perfect Pitch: Why Some People Might Have Rare Musical Skill Possessed by Bach and Mozart.” Newsweek, Newsweek, 22 Feb. 2019, www.newsweek.com/perfect-pitch-why-rare-musical-skill-bach-mozart-1326380.

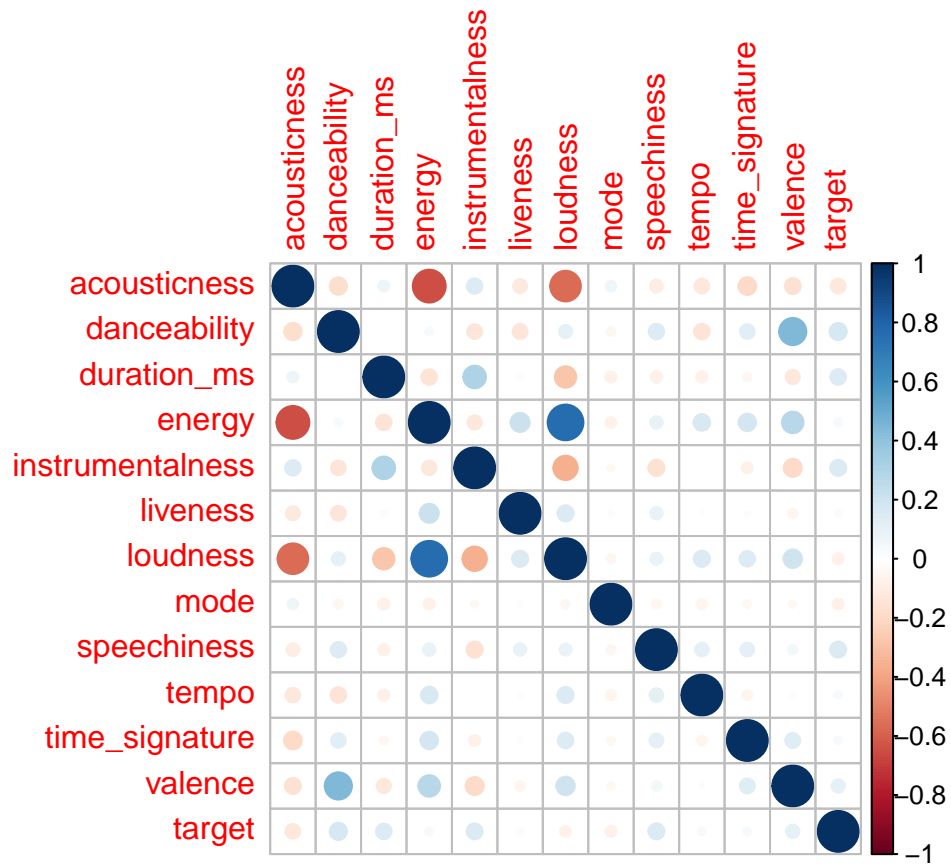
- [4] McIntire, George. “A Machine Learning Deep Dive into My Spotify Data.” Open Data Science - Your News Source for AI, Machine Learning & More, 5 Apr. 2018, opendatascience.com/a-machine-learning-deep-dive-into-my-spotify-data/.
- [5] Sloan, Nate, and Charlie Harding. “The Culture Warped Pop, for Good.” The New York Times, The New York Times, 14 Mar. 2021, www.nytimes.com/interactive/2021/03/14/opinion/pop-music-songwriting.html?auth=login-google1tap&login=google1tap.
- [6] Vehtari, Aki, et al. “Bayesian Logistic Regression with Rstanarm.” Github, 4 Dec. 2019, avehtari.github.io/modelselection/diabetes.html.
- [7] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su. “A weakly informative default prior distribution for logistic and other regression models.” The Annals of Applied Statistics, 2(4) 1360-1383 December 2008, <https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-4/A-weakly-informative-default-prior-distribution/10.1214/08-AOAS191.full?tab=ArticleLink>

Appendix

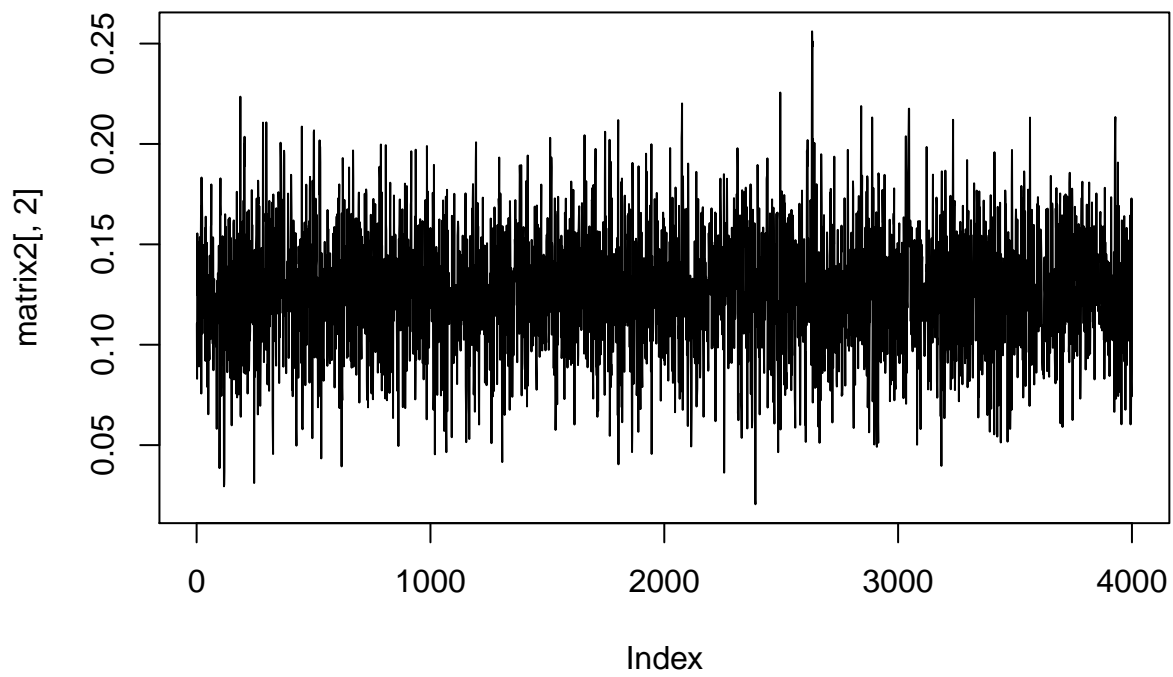
Correlation plot for data set 1.



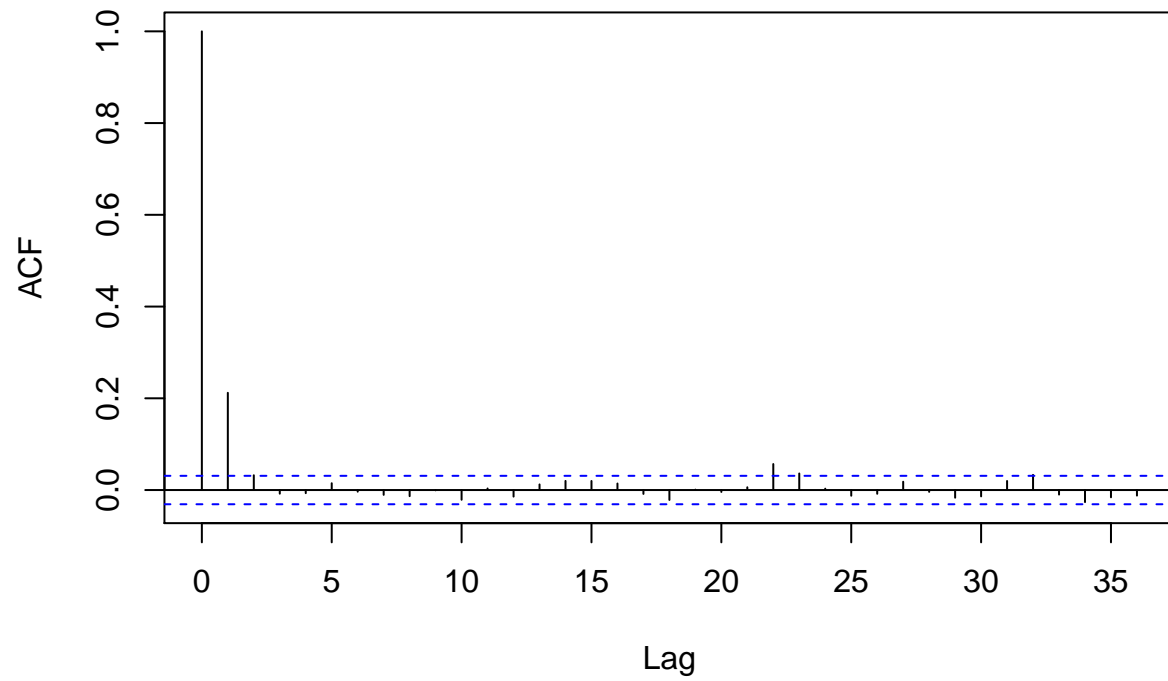
Correlation plot for data set 2.



ACF and Traceplots for the duration variable in data set 1

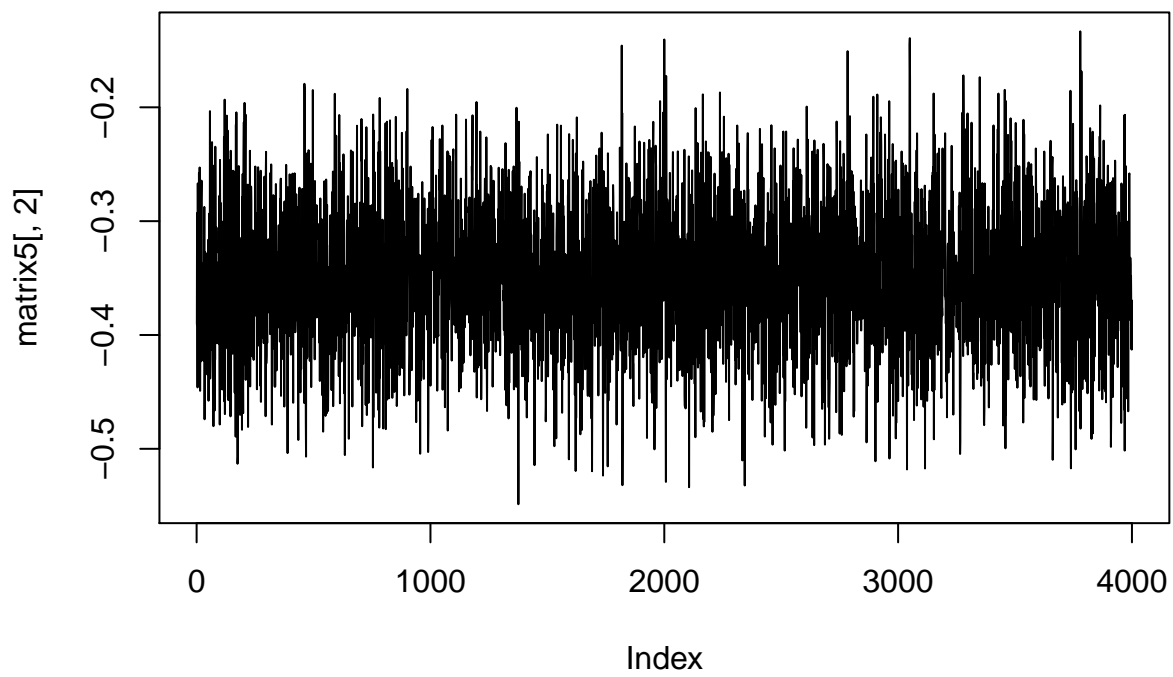


Series matrix2[, 2]

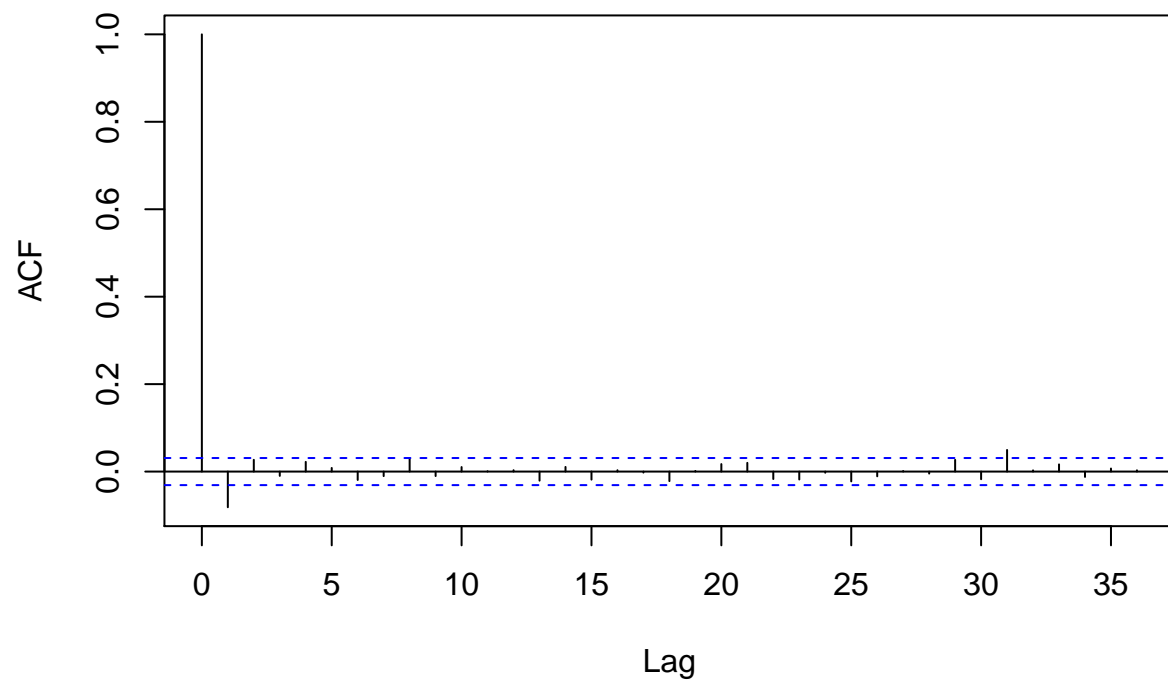


ACF and Traceplots for the all variables in final model in data set 2

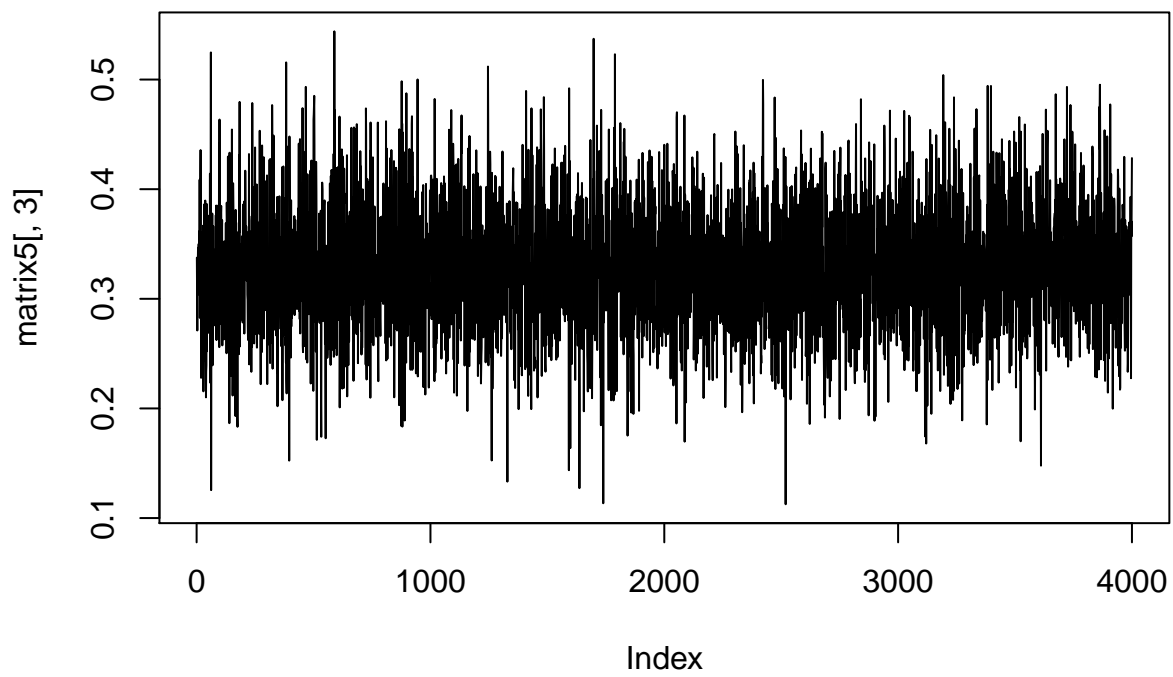
```
## [1] "acousticness"
```



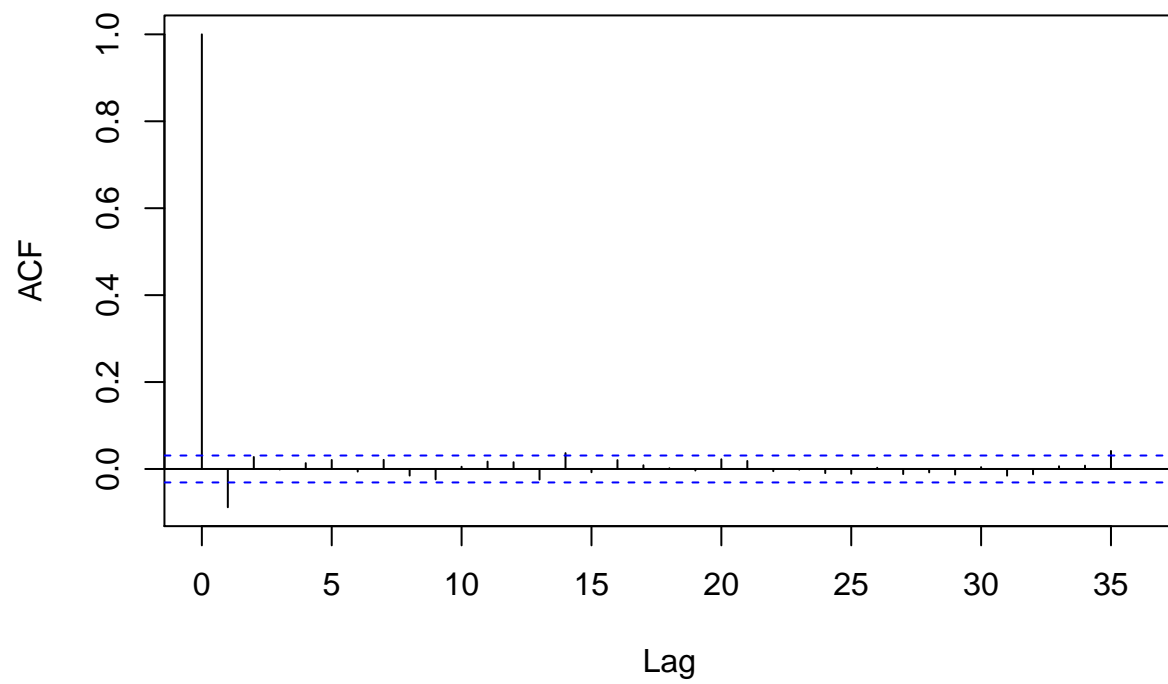
Series matrix5[, 2]



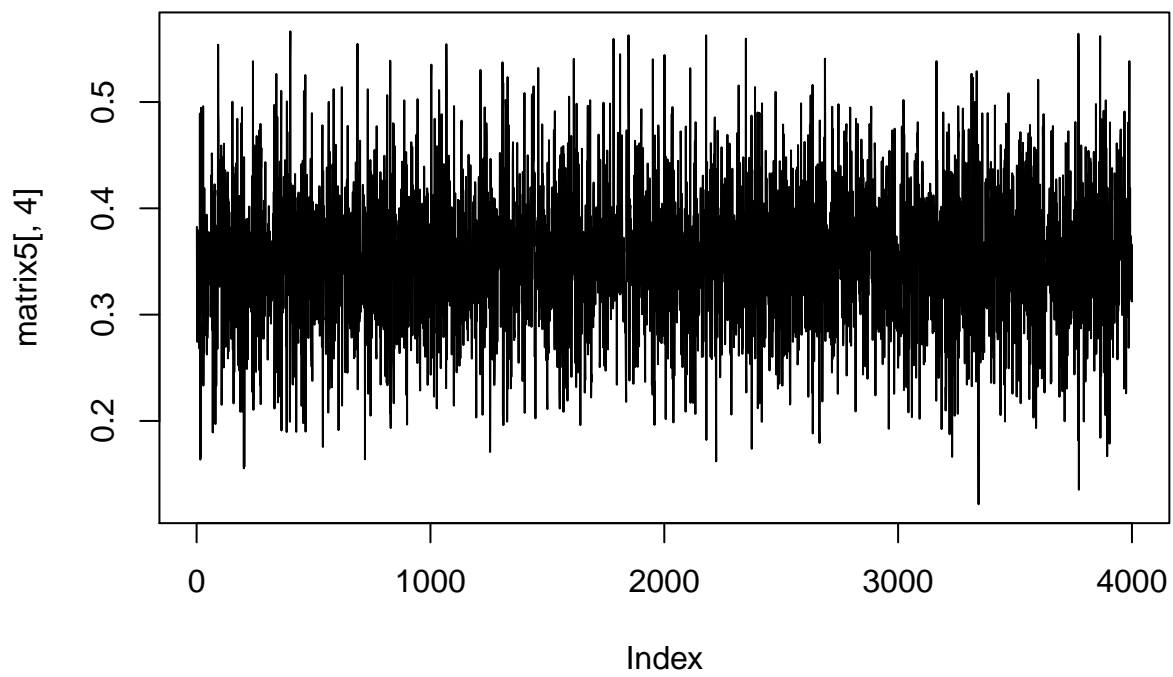
```
## [1] "danceability"
```



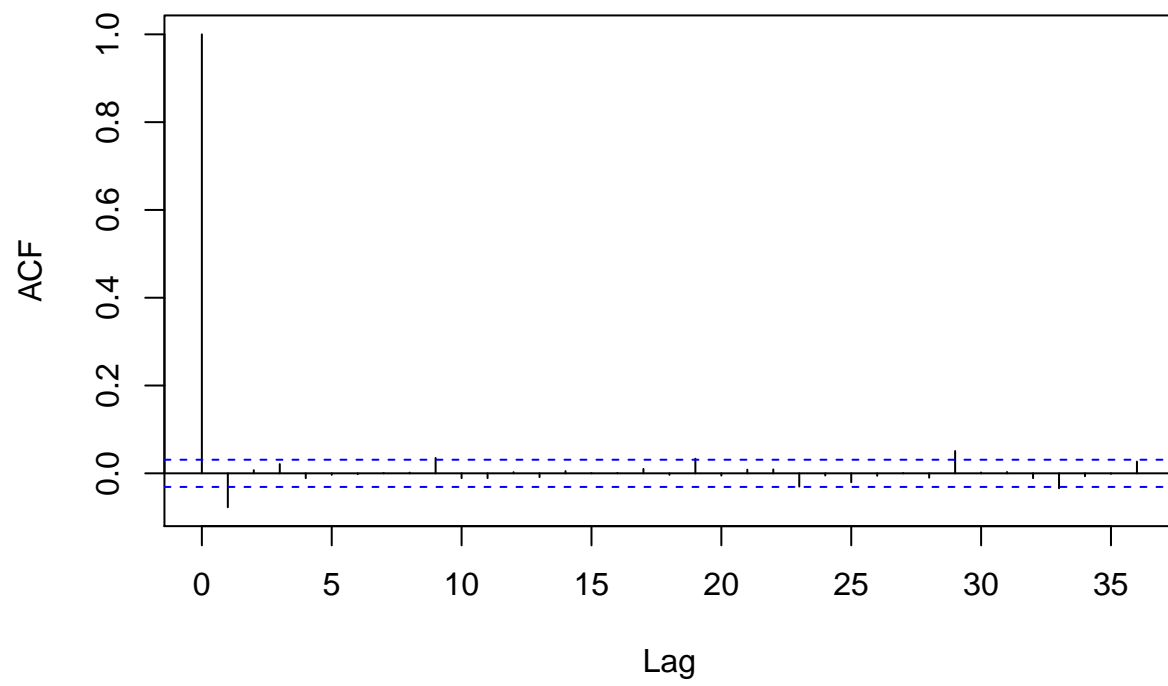
Series matrix5[, 3]



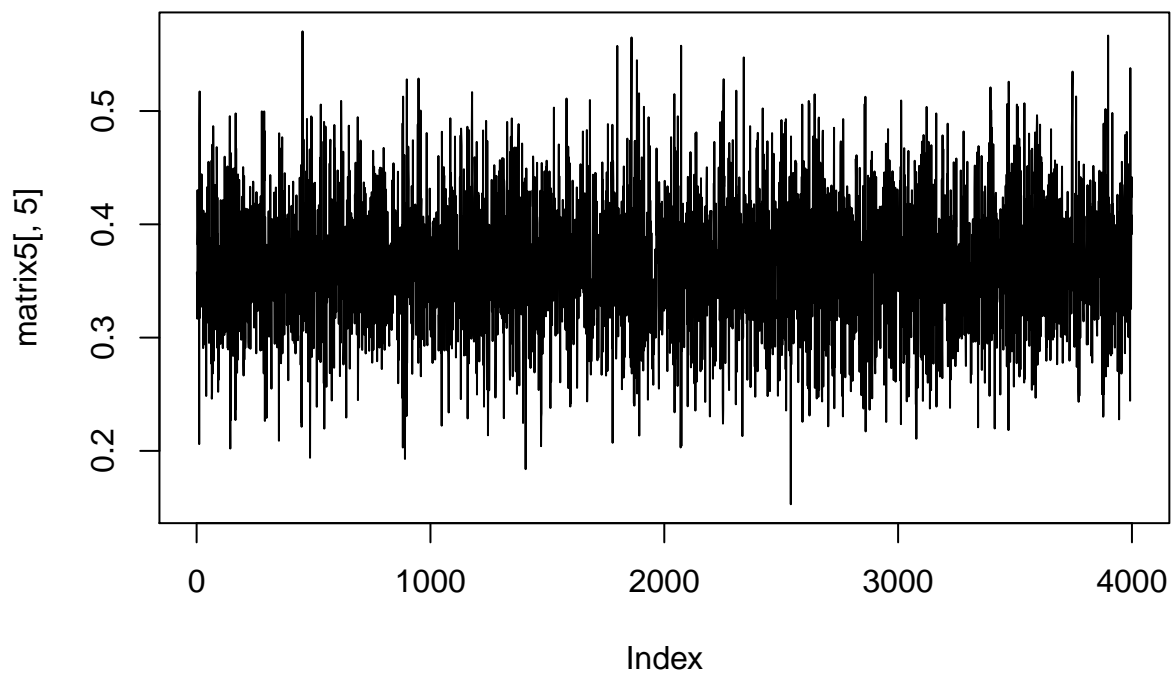
```
## [1] "duration"
```

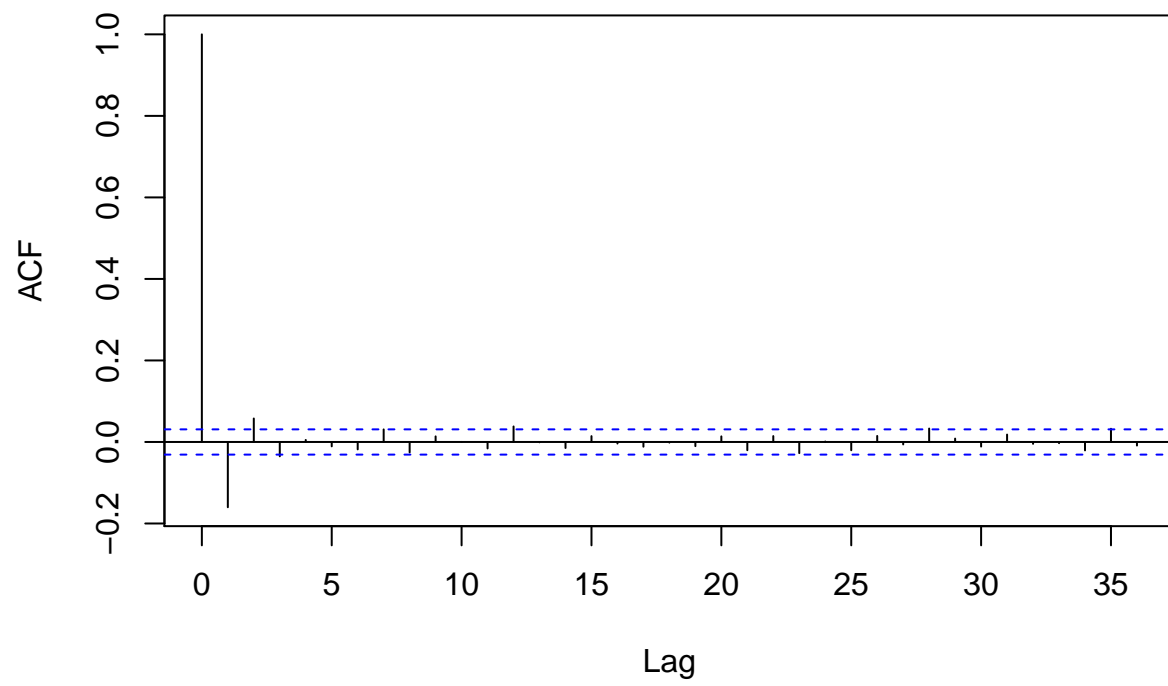
Series matrix5[, 4]



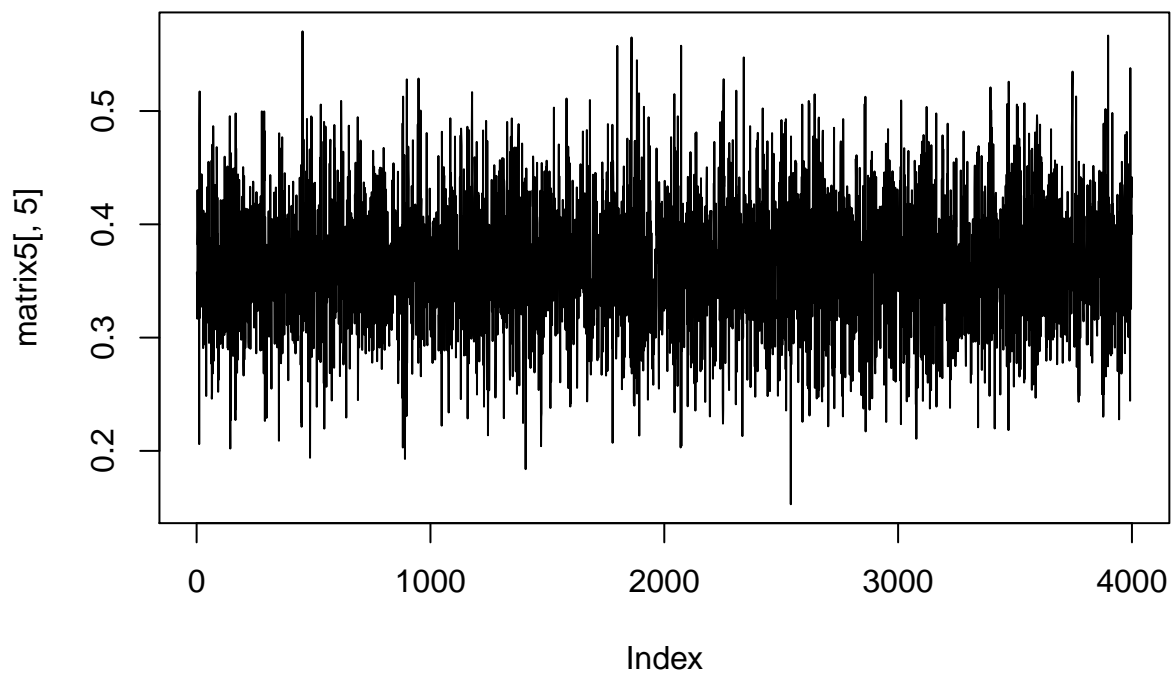
```
## [1] "instrumentalness"
```



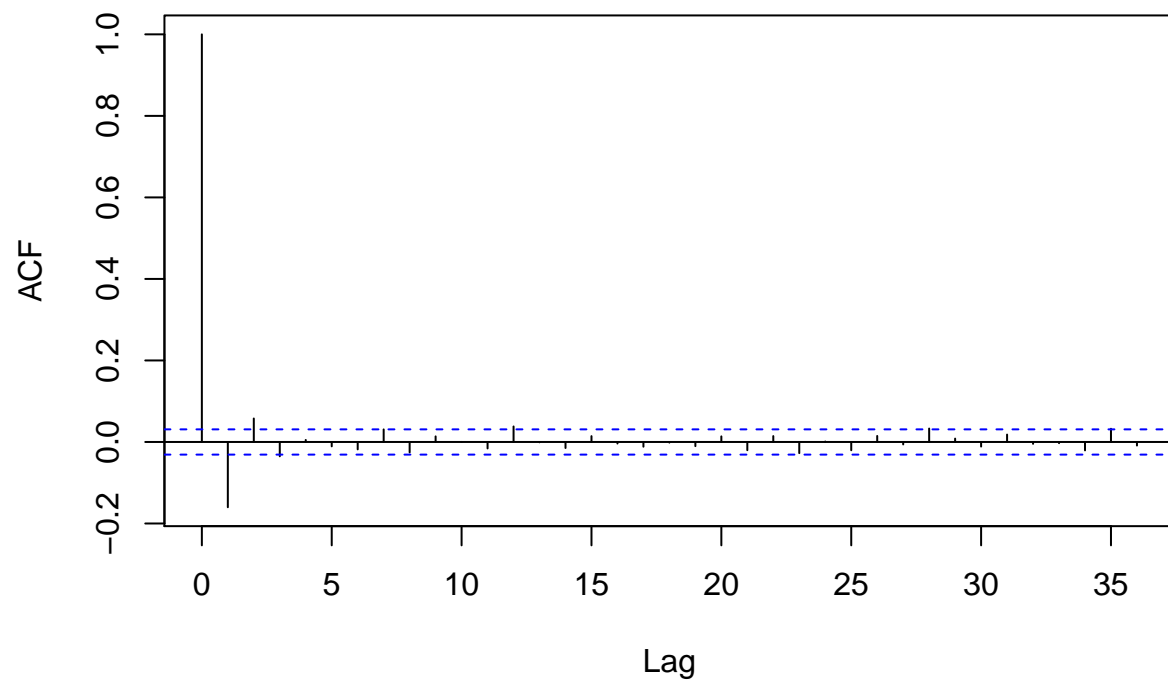
Series matrix5[, 5]



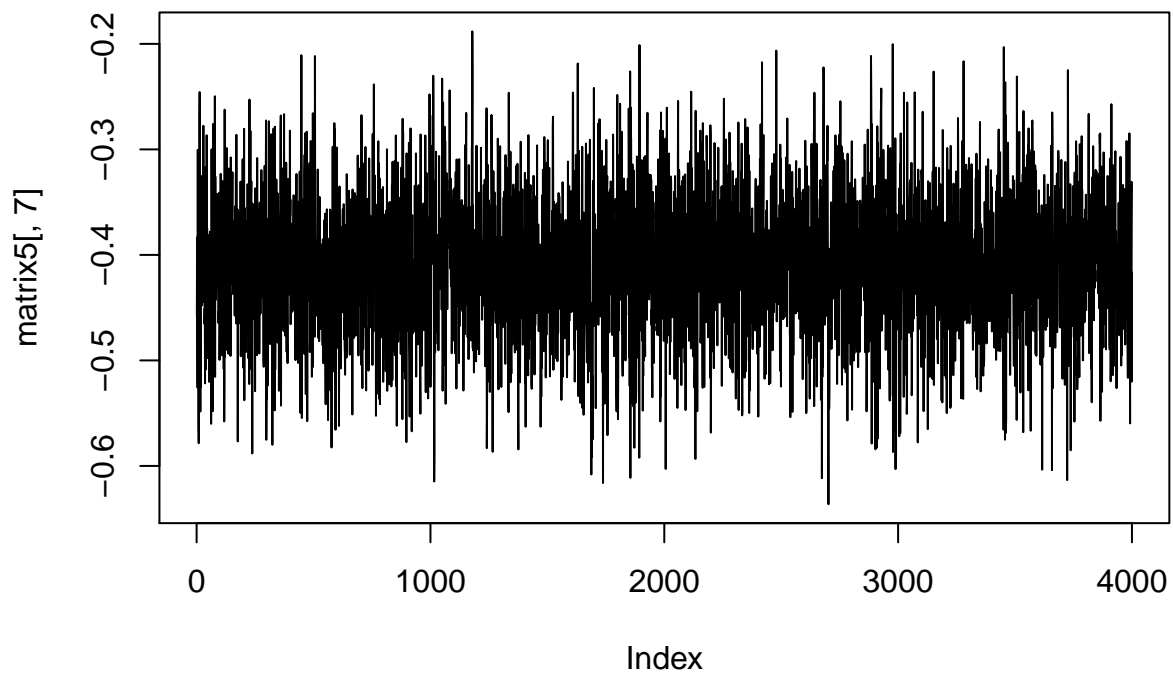
```
## [1] "liveness"
```



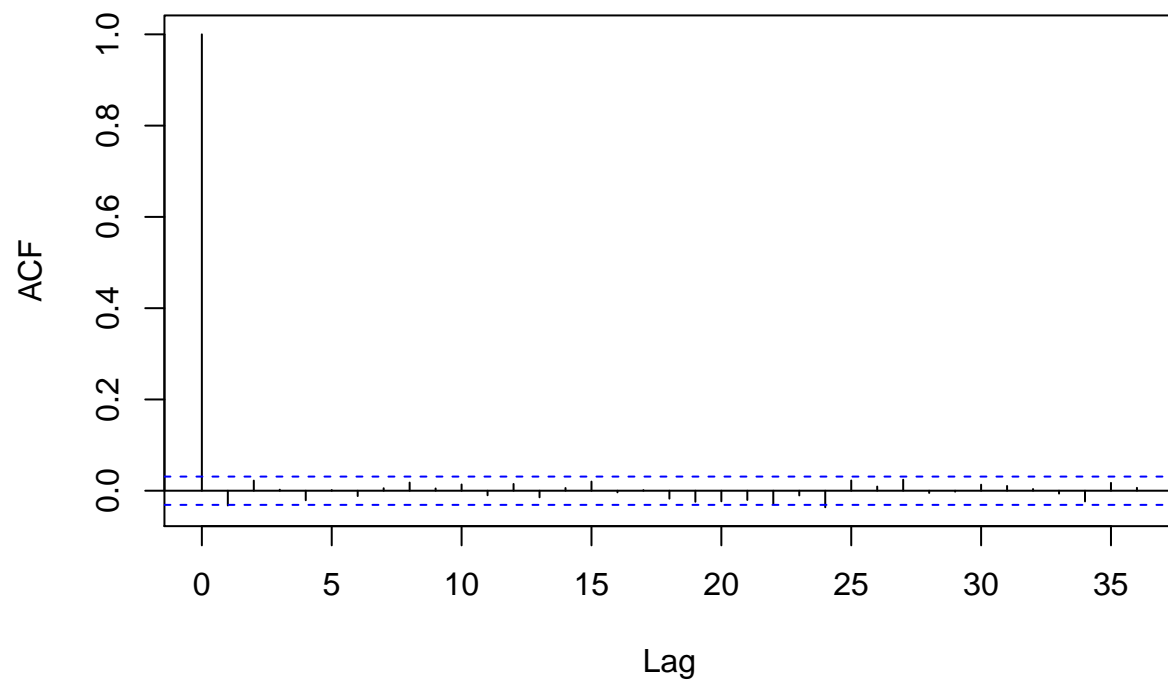
Series matrix5[, 5]



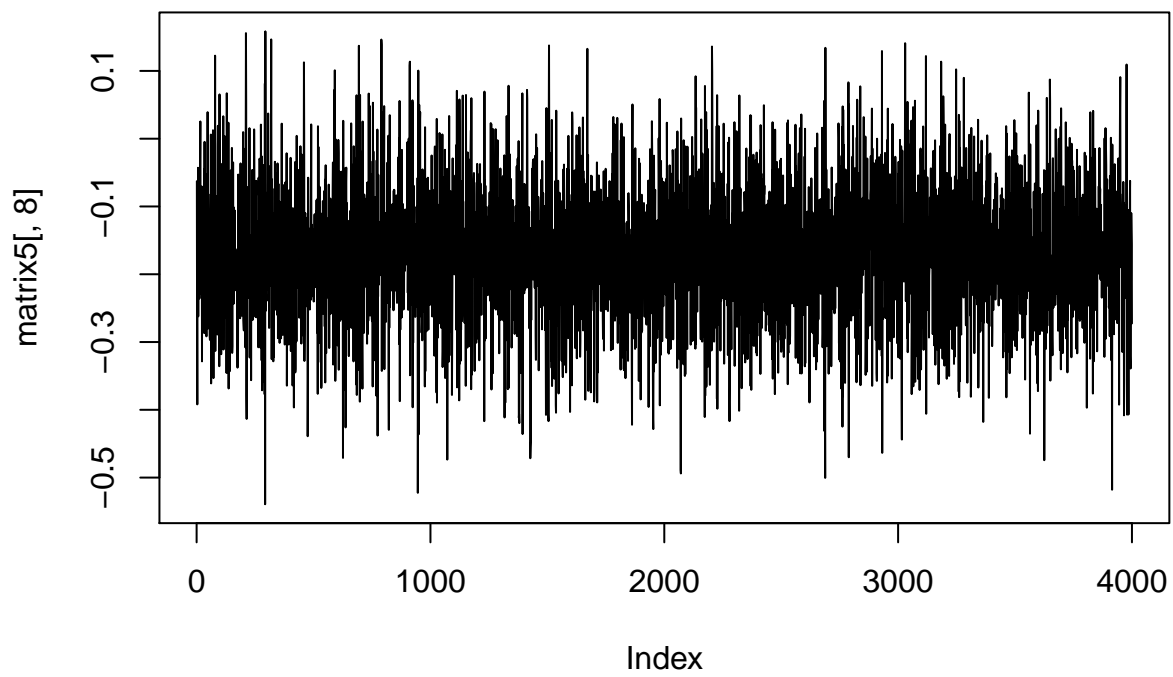
```
## [1] "loudness"
```



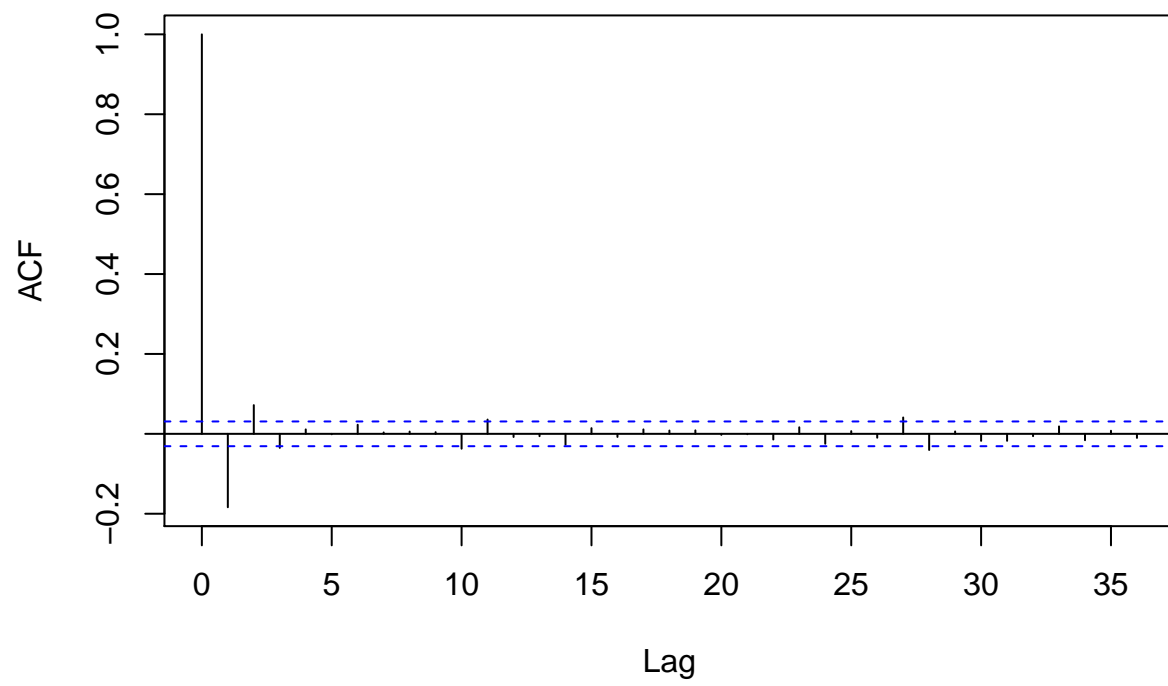
Series matrix5[, 7]



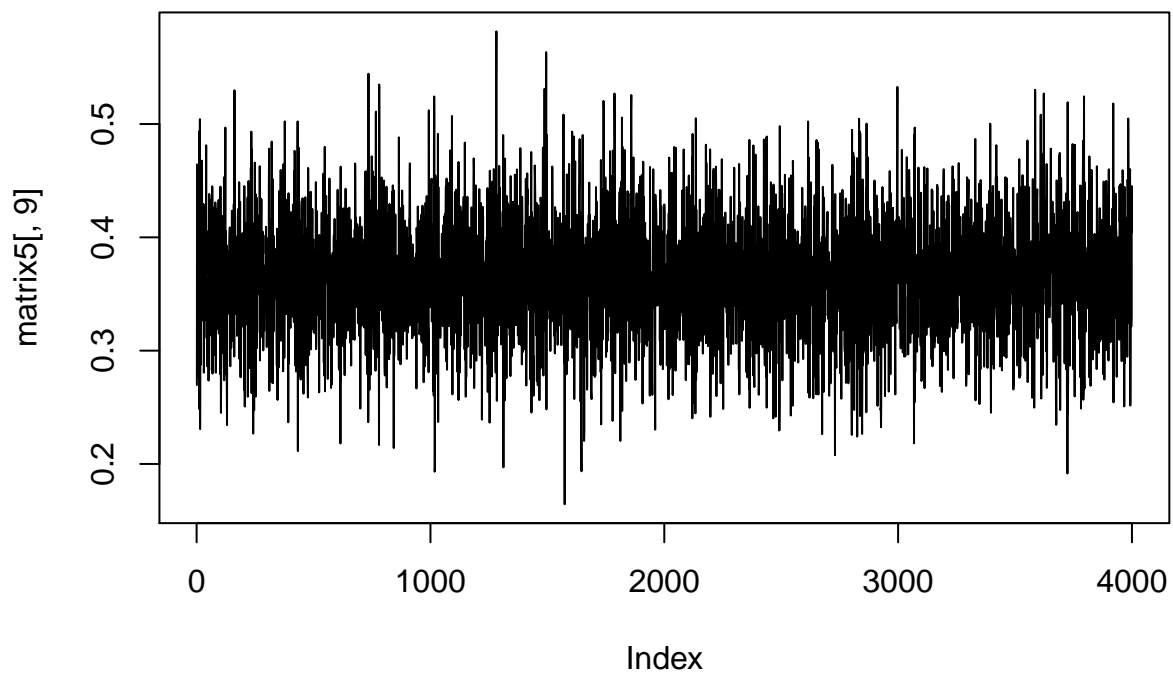
```
## [1] "mode"
```

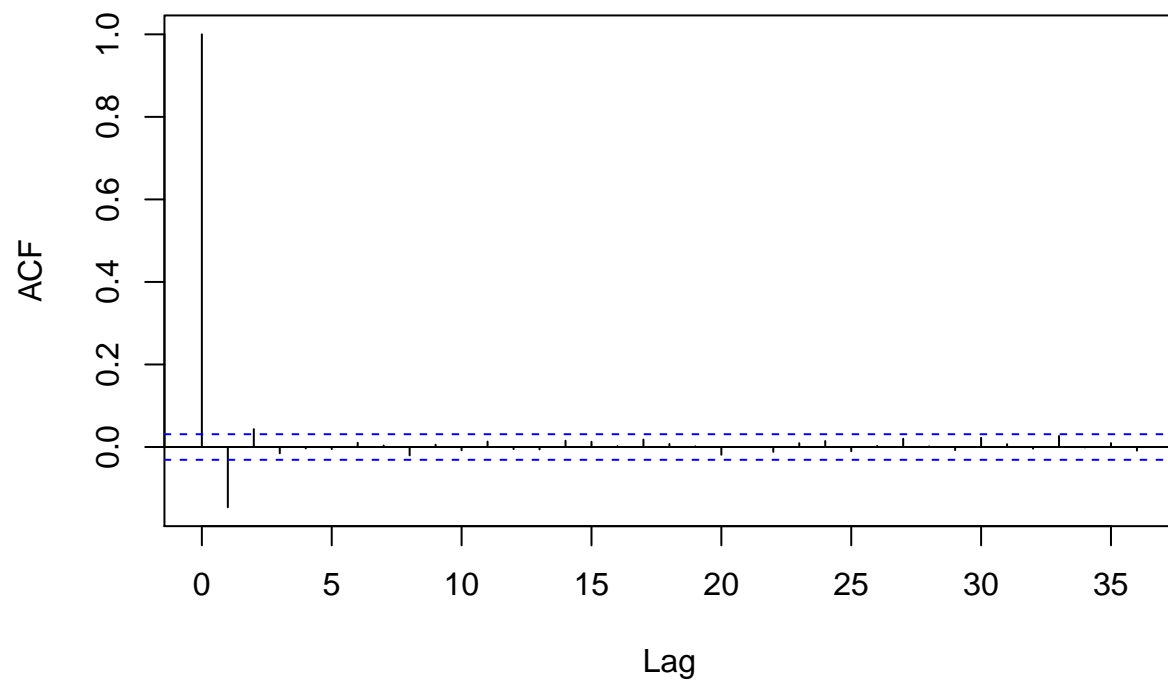
Series matrix5[, 8]



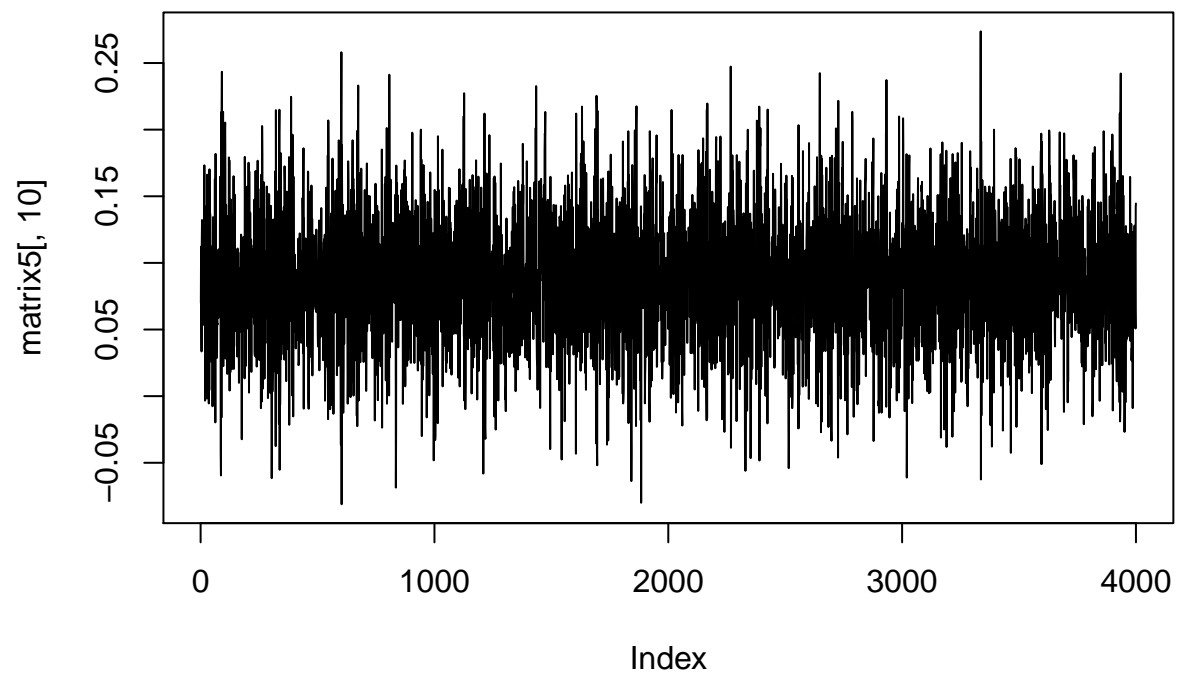
```
## [1] "speechness"
```



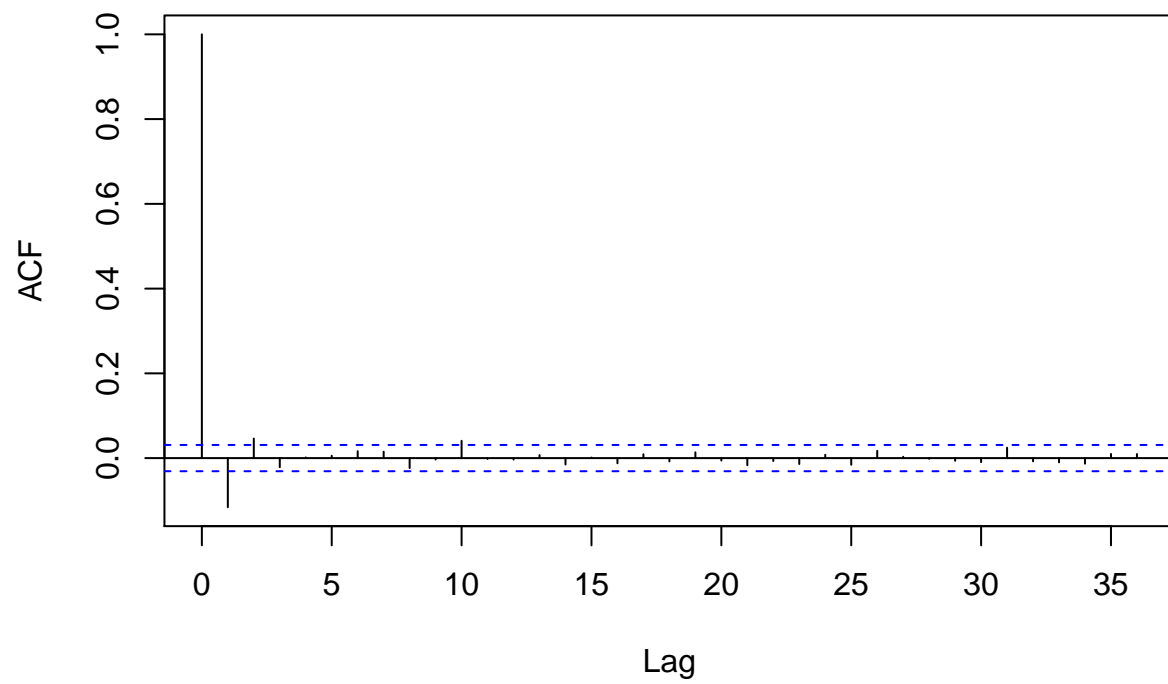
Series matrix5[, 9]



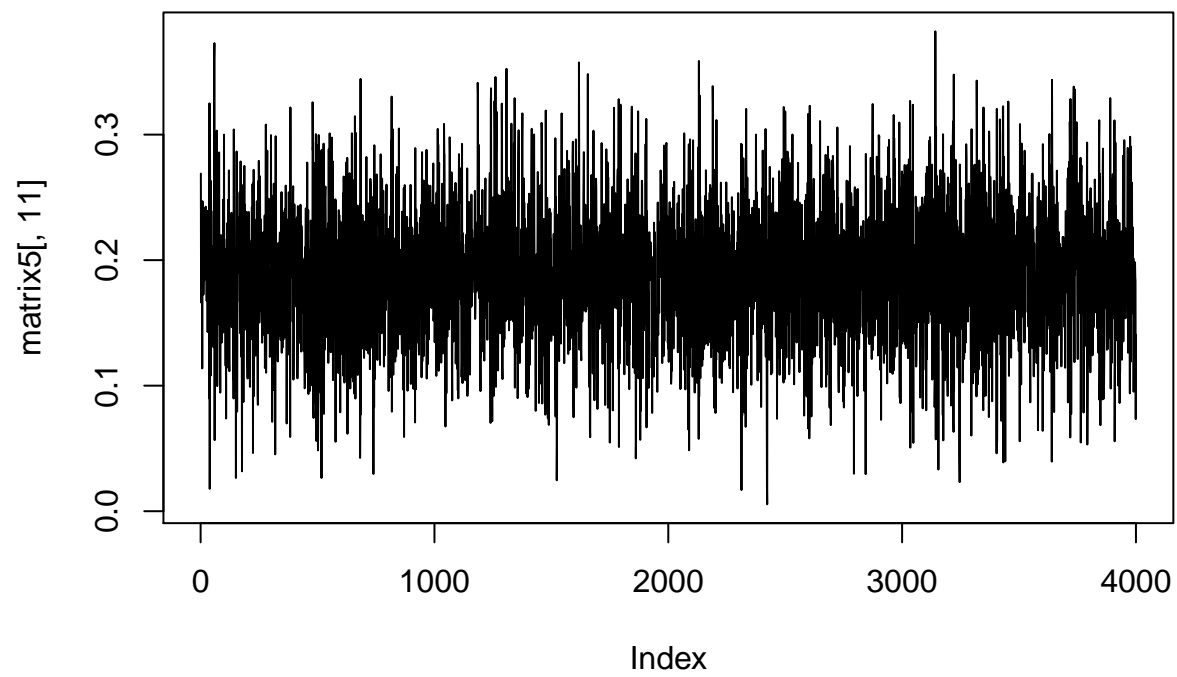
```
## [1] "tempo"
```



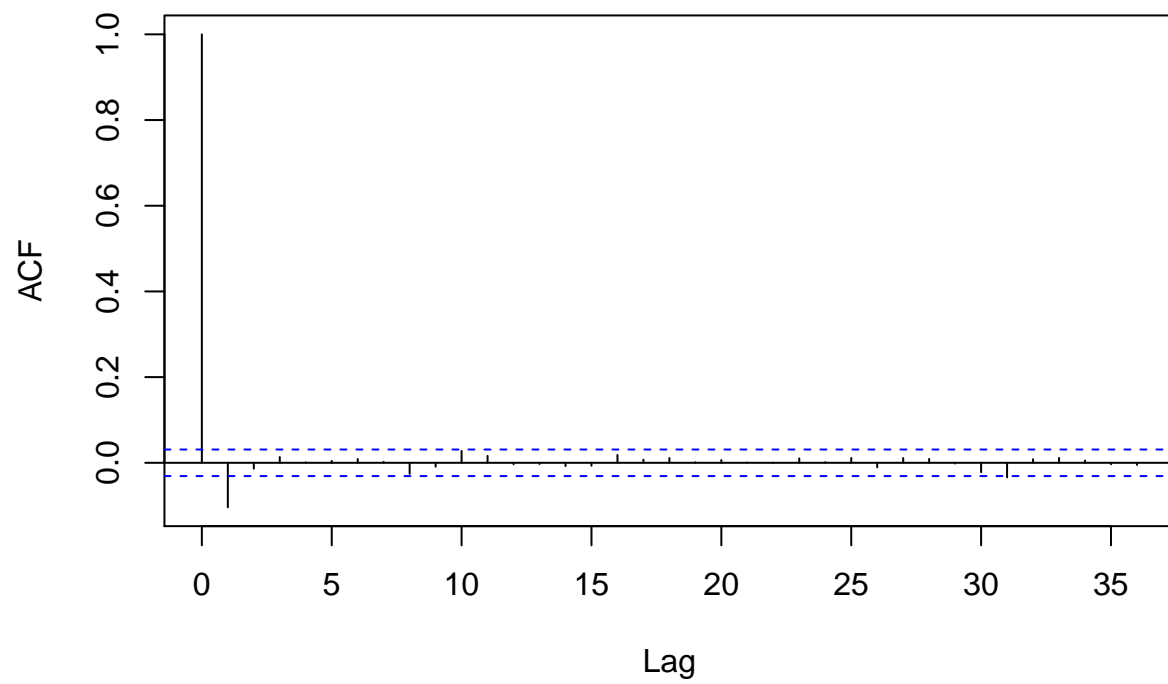
Series matrix5[, 10]



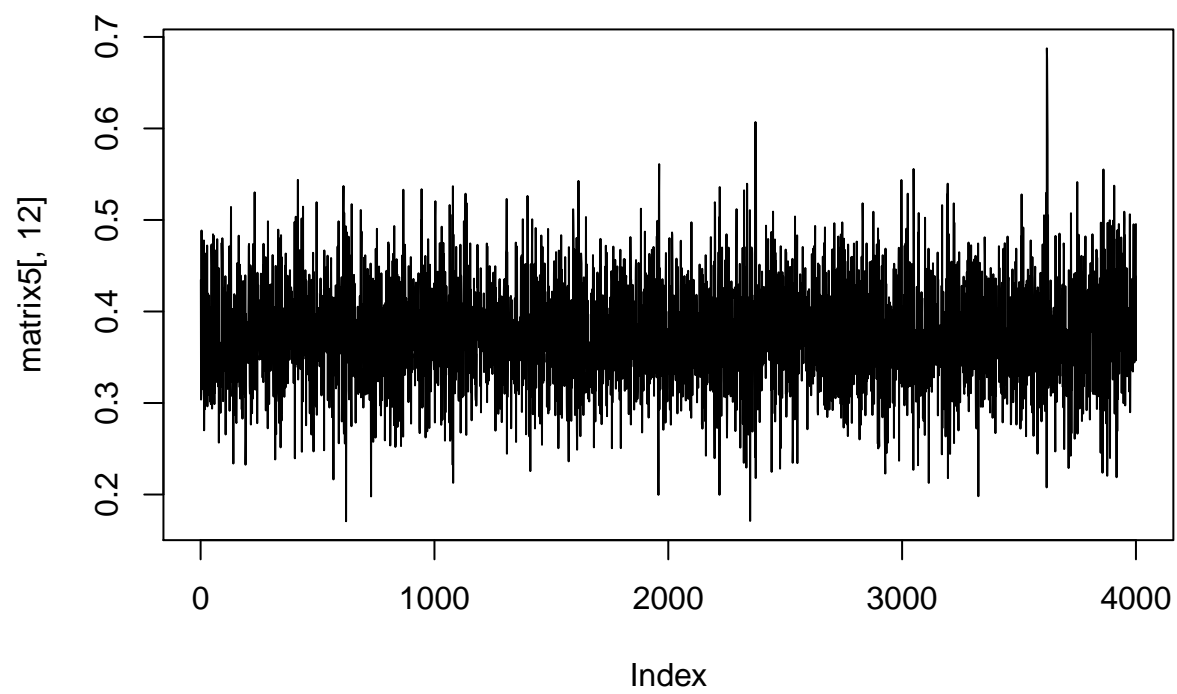
```
## [1] "valence"
```



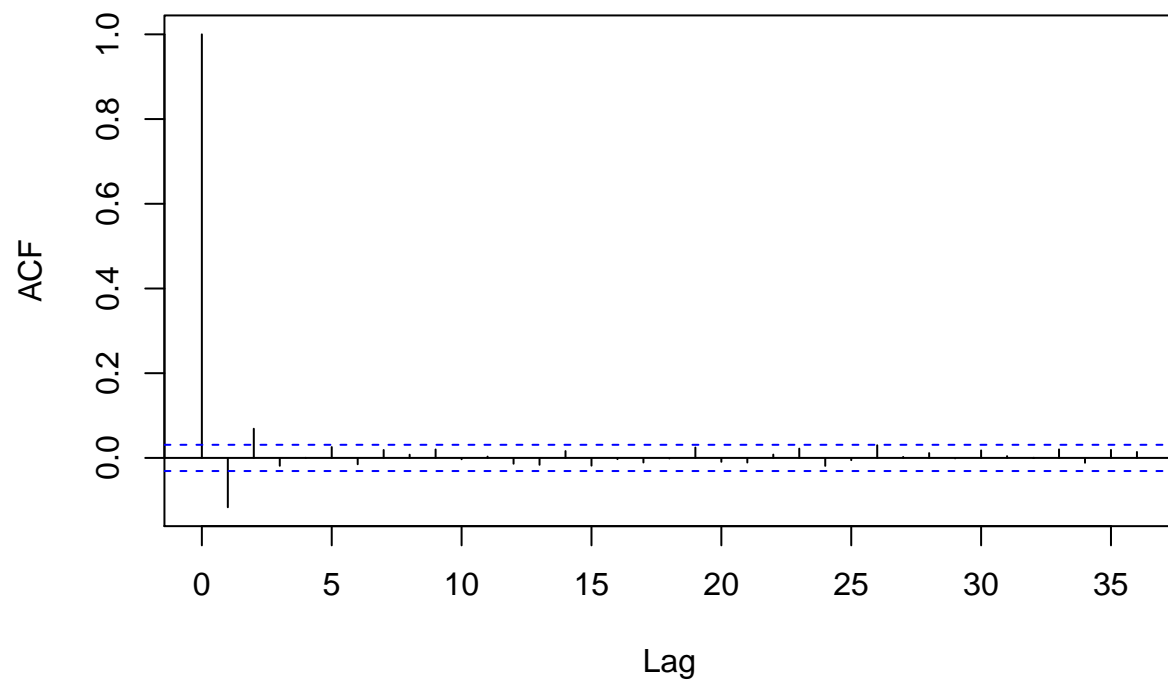
Series matrix5[, 11]



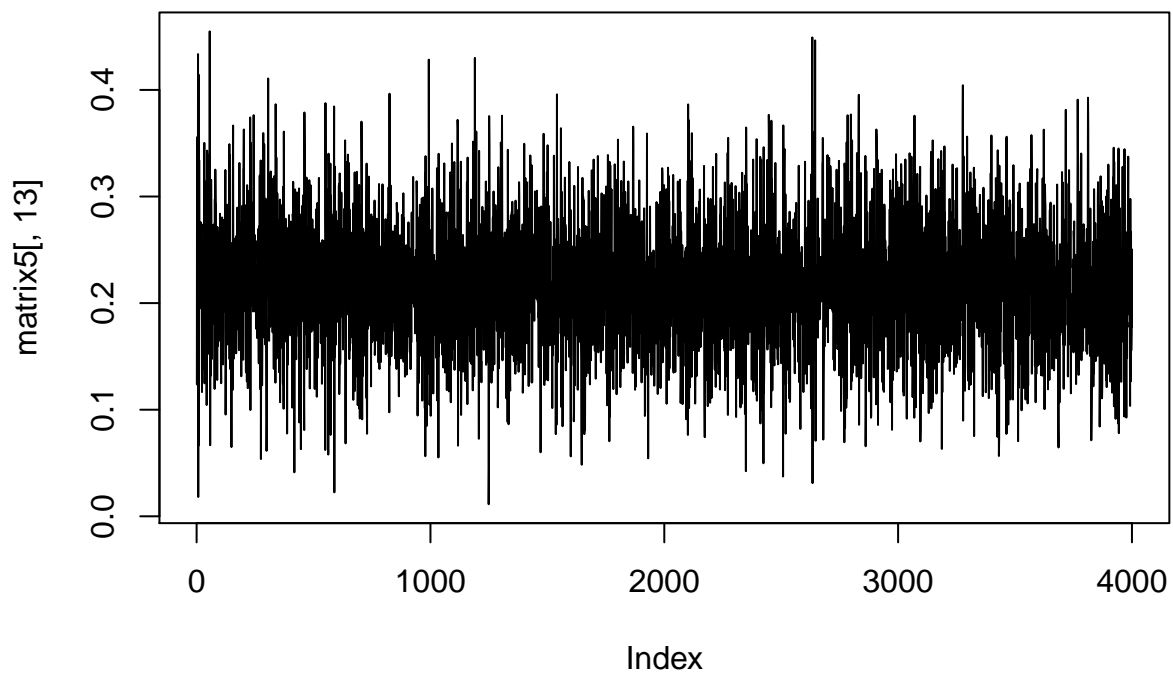
```
## [1] "interaction"
```

Series matrix5[, 12]



```
## [1] "interaction2"
```



Series matrix5[, 13]

