Multivariate Statistics Research Portfolio

Sharnya Govindaraj

Created as part of the course
Advanced Multivariate Statistics
taught by Dr. Hervé Abdi and Luke Moraglia

Principal Component Analysis

Method: Principal component analysis

Principal component analysis (PCA) is used to analyze one table of quantitative data. PCA mixes the input variables to give new variables, called principal components. The first principal component is the line of best fit. It is the line that maximizes the inertia (similar to variance) of the cloud of data points. Subsequent components are defined as orthogonal to previous components, and maximize the remaining inertia.

PCA gives one map for the rows (called *factor scores*), and one map for the columns (called *loadings*). These 2 maps are related, because they both are described by the same components. However, these 2 maps project different kinds of information onto the components, and so they are *interpreted differently*. Factor scores are the coordinates of the row observations. They are interpreted by the distances between them, and their distance from the origin. Loadings describe the column variables. Loadings are interpreted by the angle between them, and their distance from the origin.

The distance from the origin is important in both maps, because squared distance from the mean is inertia (variance, information; see sum of squares as in ANOVA/regression). Because of the Pythagorean Theorem, the total information contributed by a data point (its squared distance to the origin) is also equal to the sum of its squared factor scores.

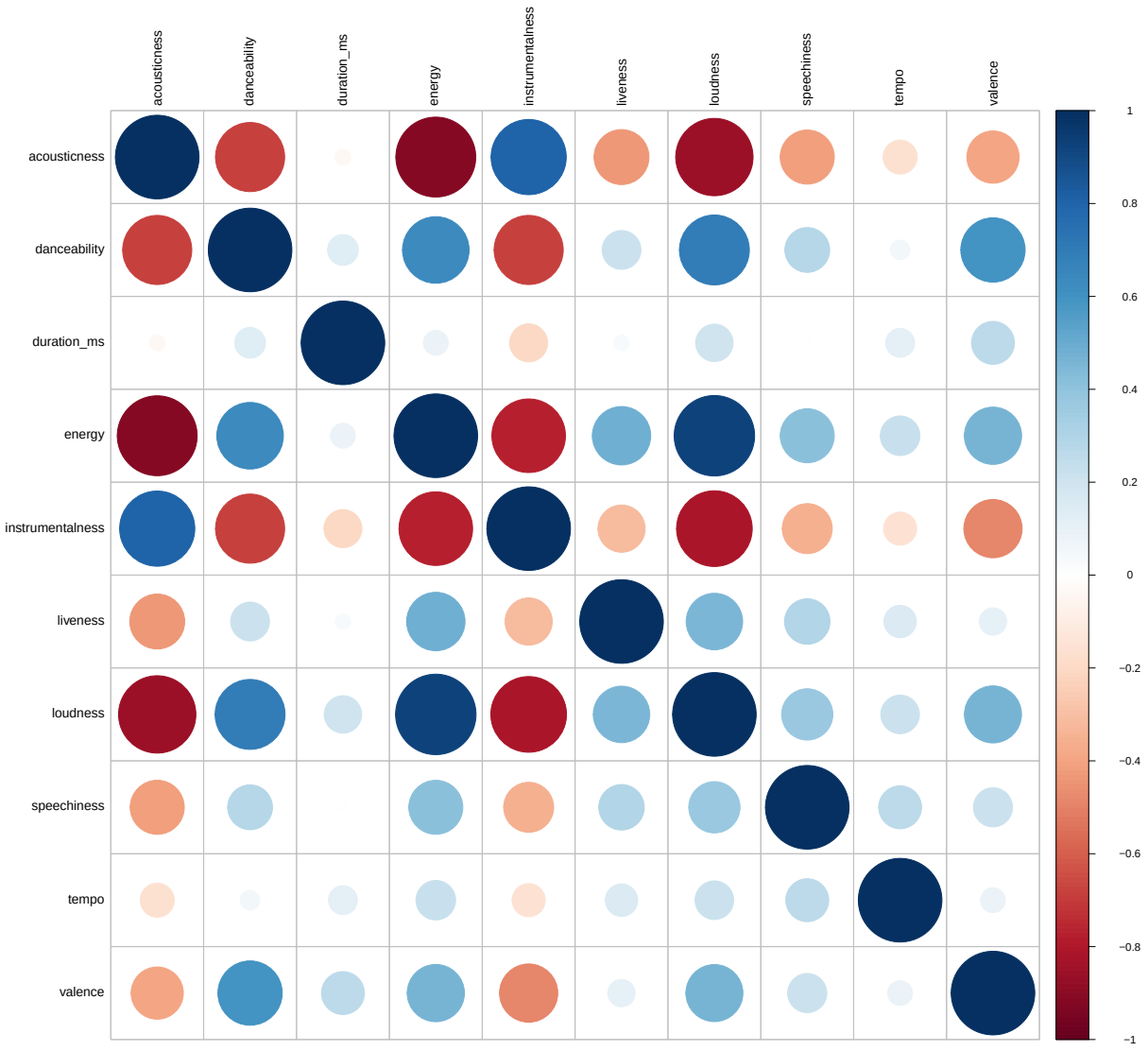
Data set: Audio features

This is a dataset which describes audio features of songs in Spotify playlists. Specifically, the music.track dataset measures 165 songs on 16 variables, of which 11 are quantitative. Some of the audio features described are acousticness, danceability, and energy.

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo
16	0.845	0.515	313560	0.519	6.96e-01	0.1020	-9.631	0.0391	128.346
48	0.843	0.656	216453	0.217	4.30e-06	0.2960	-13.725	0.0538	90.065
2	0.873	0.571	290293	0.346	5.19e-01	0.0980	-12.569	0.0409	93.885
46	0.369	0.567	342067	0.500	5.50e-05	0.5300	-9.294	0.0330	80.626
45	0.050	0.707	272000	0.508	0.00e+00	0.2550	-9.629	0.0344	88.989
47	0.315	0.589	250867	0.437	1.27e-03	0.0944	-12.581	0.0487	83.263

The correlation plot

In the correlation plot, we see some interesting, strong correlations. For instance, energy and acousticness are highly negatively correlated, just as loudness and acousticness. As expected, energy and loudness are strongly positively correlated, as are instrumentalness and acousticness.



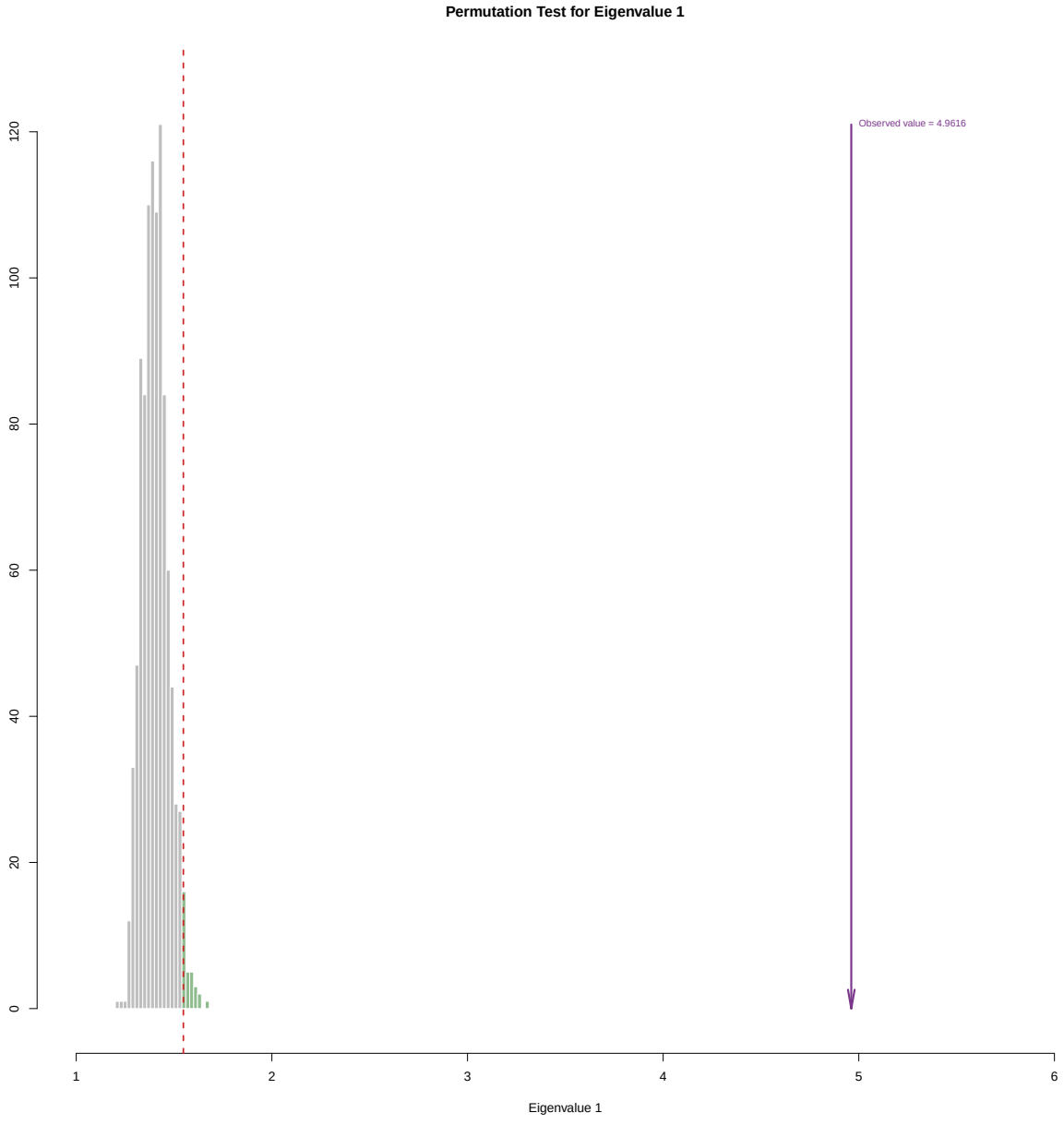

```
res_pca <- epPCA(df, center = TRUE, scale = "SS1", DESIGN = hwdata$genre, graphs = FALSE)
```

Analysis - Inference PCA

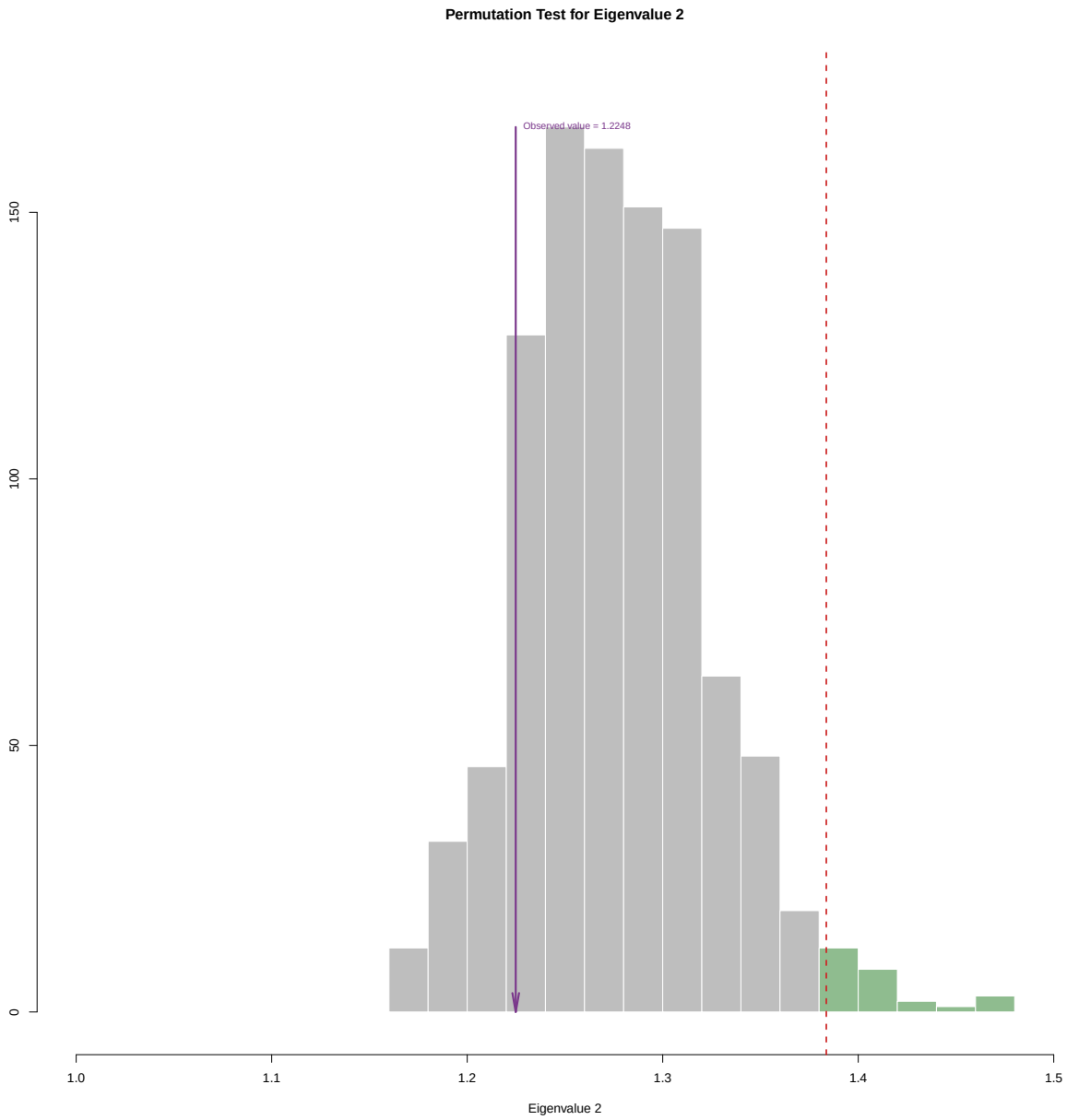
```
res_pcaInf <- epPCA.inference.battery(df, center = TRUE, scale = "SS1",
                                      DESIGN = music.track$genre, graphs = FALSE,
                                      test.iters = 999)
```

Permutation test plots

The plot for Eigenvalue 1 shows that the observed value is very far from the 5% line (red dotted), hence implying that there is a possibility of the null being true less than 5% of the time.

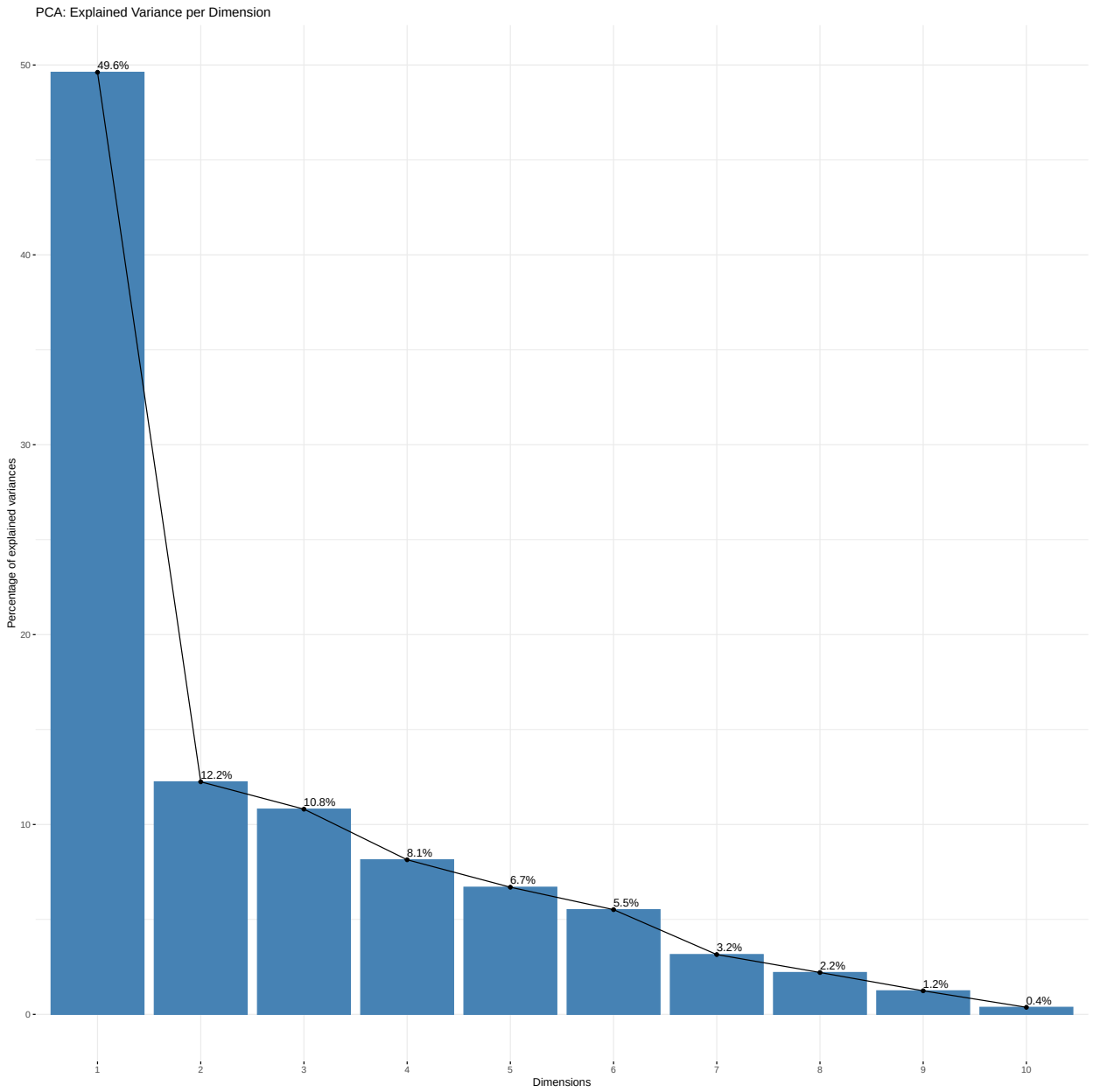


The permutation test for eigenvalue 2 does not show statistical significance



Scree Plot

The scree plot indicates that one component represents most of the variance in the dataset. The second and third components are quite close to each other, which warrants further investigation (but not at the cost of the first component).



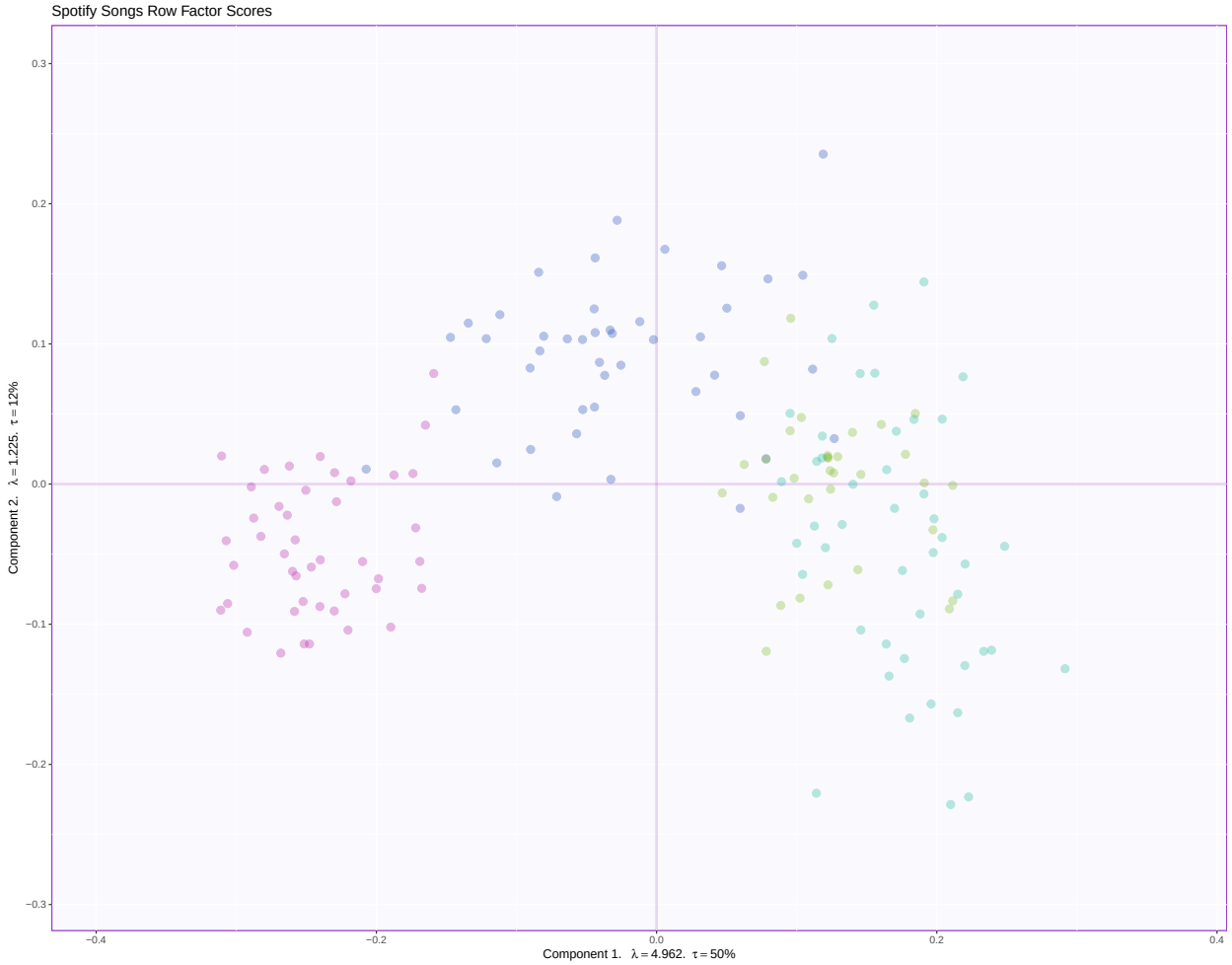
Row factor scores - inference PCA

The row factor scores plot shows the distribution of data, in this case, colored by different genres of music (sleep, dinner, party, and workout).

Legend:

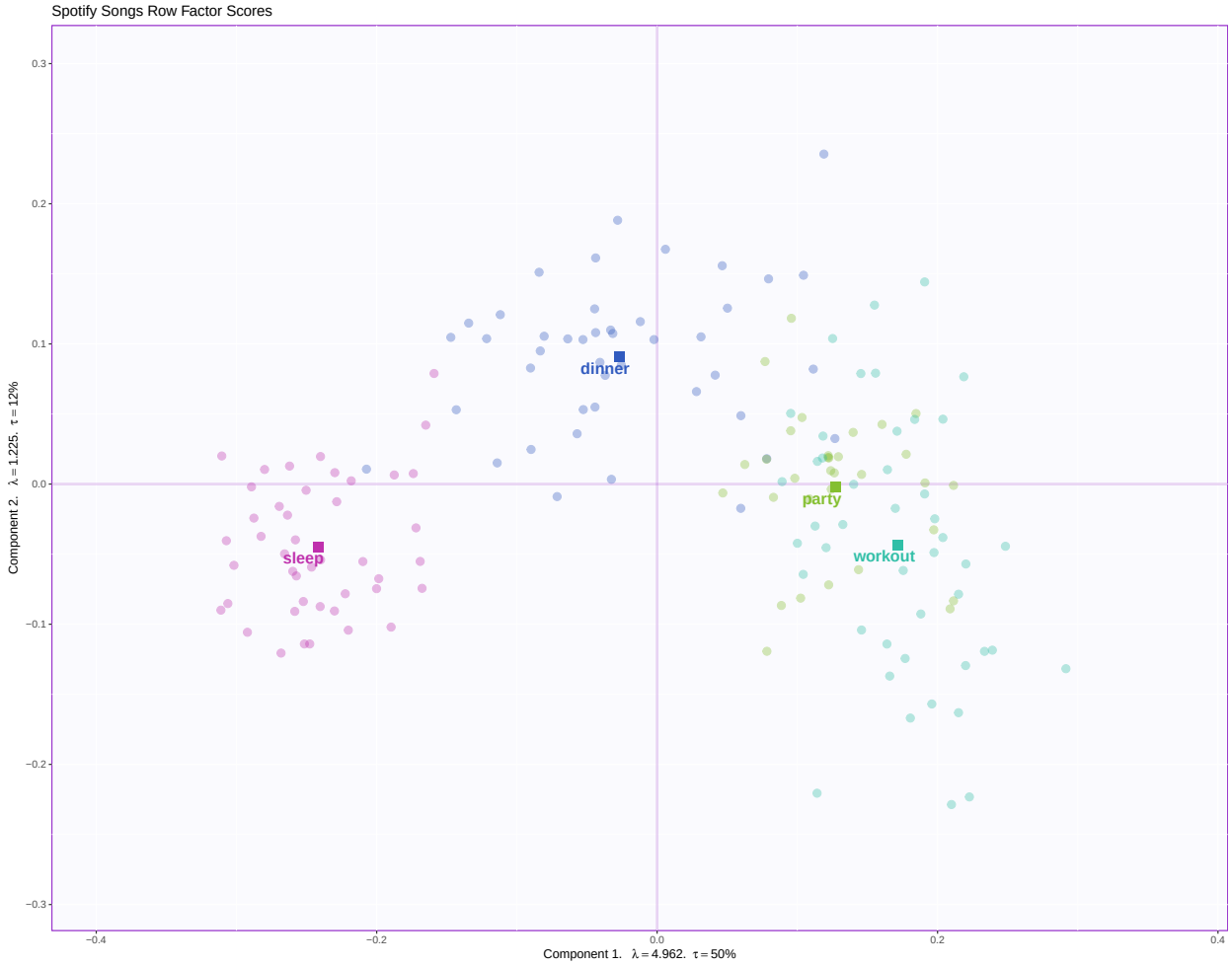
Pink = Sleep Blue = Dinner Green = Party Cyan = Workout

The scores are distributed nicely and the components make sense (particularly component 1).



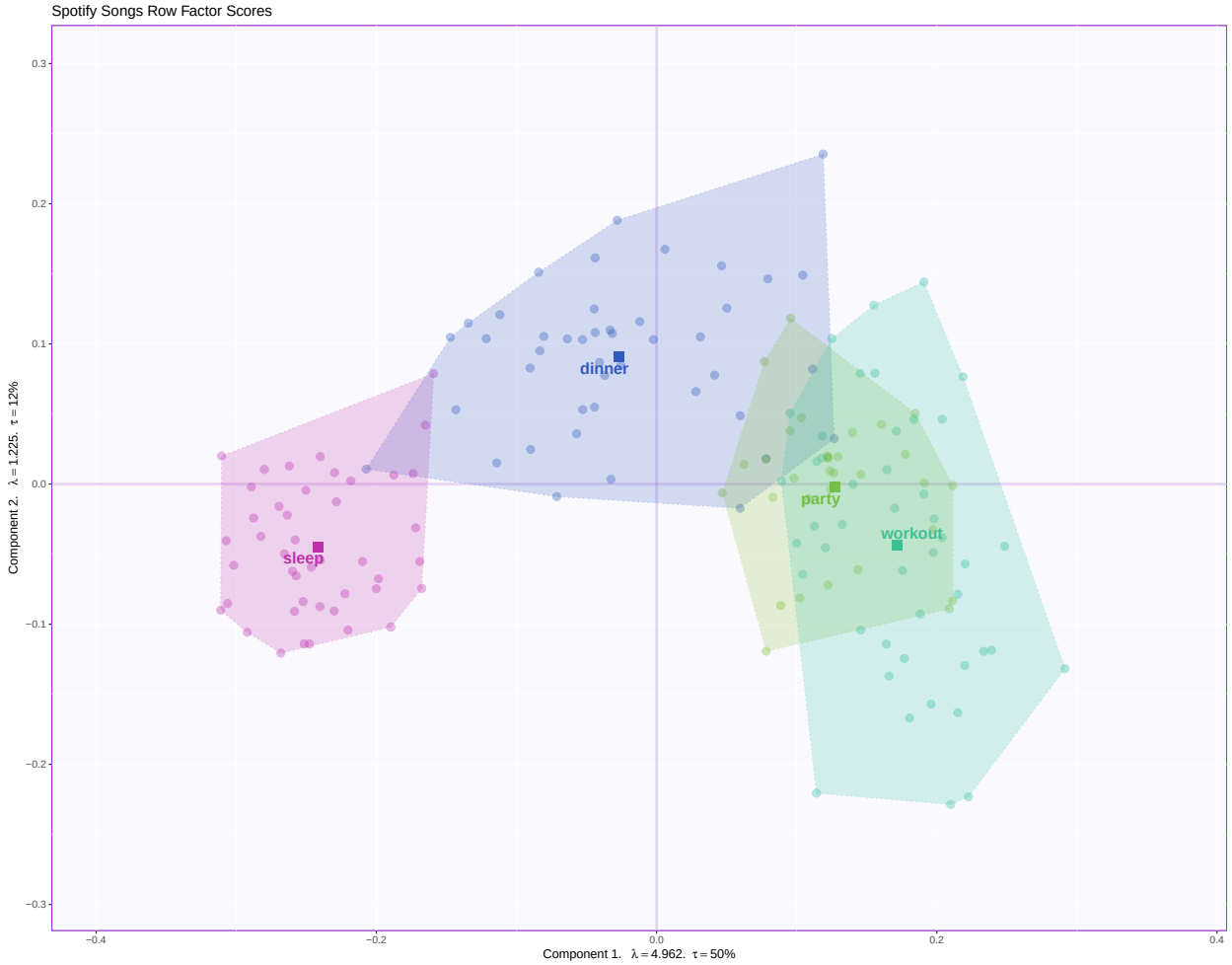
Plotting the means

Adding means of the genre to the row factor scores plot helps in getting a clearer idea about where the centre of each group lies. Further, the similarity/difference between the genres can be understood by constructing confidence intervals around the means.



Tolerance interval

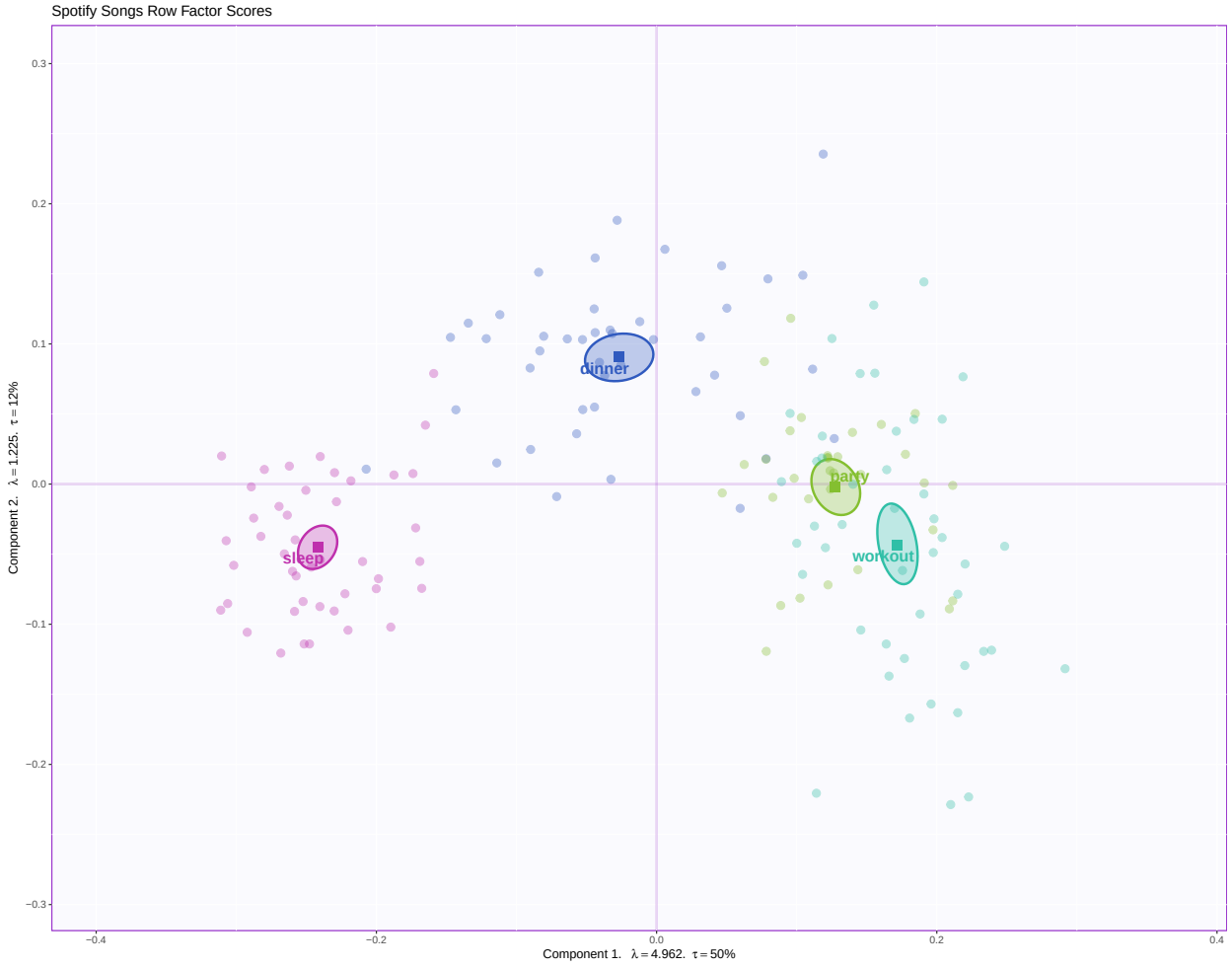
We can plot the tolerance interval for each genre.



Bootstrap interval

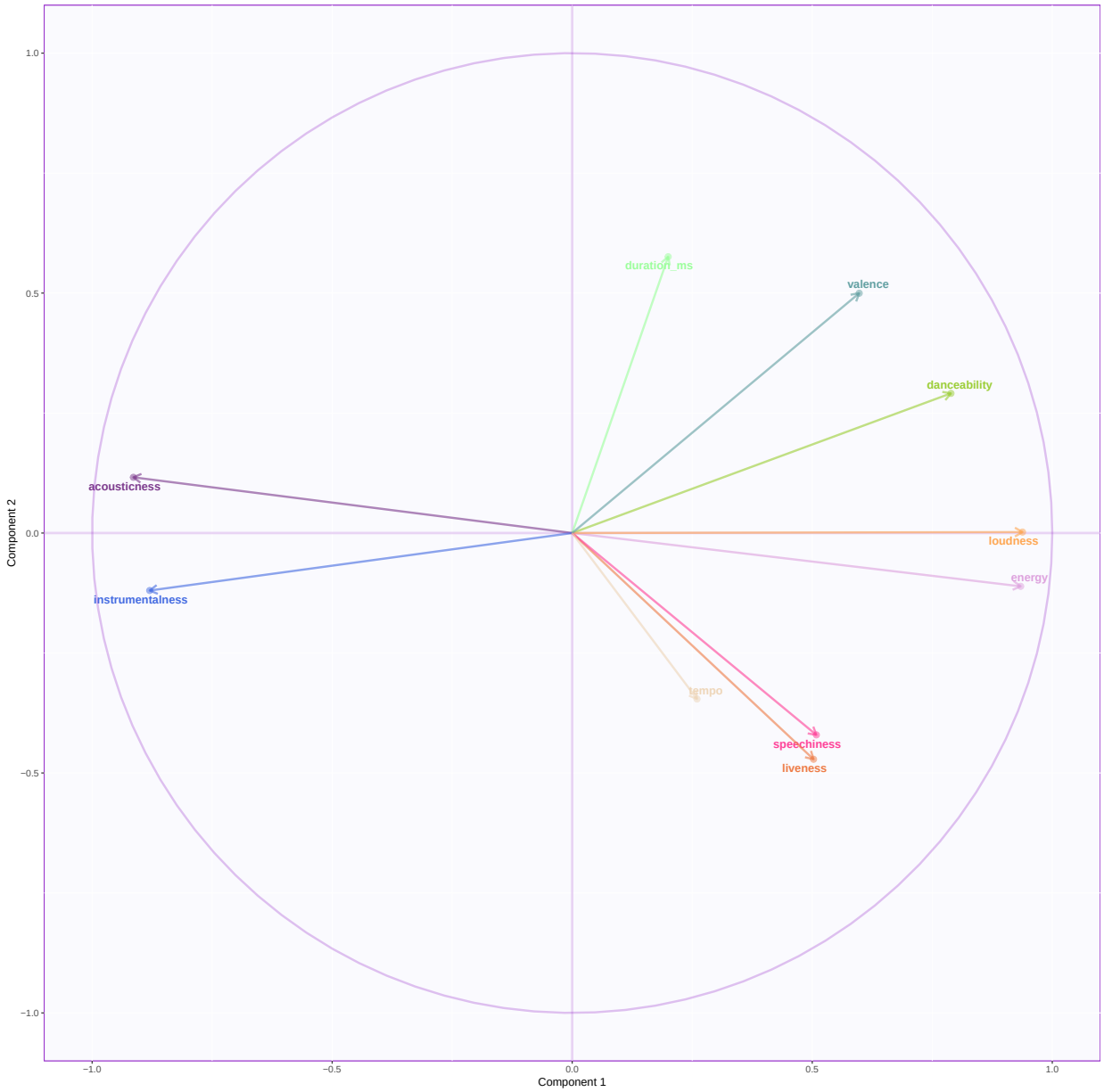
We can also add the bootstrap interval for the group means to see if these group means are significantly different.

The ellipses around the group means indicates the bootstrap intervals. The smaller their radii, the more confidence we can assume of our group mean estimate.



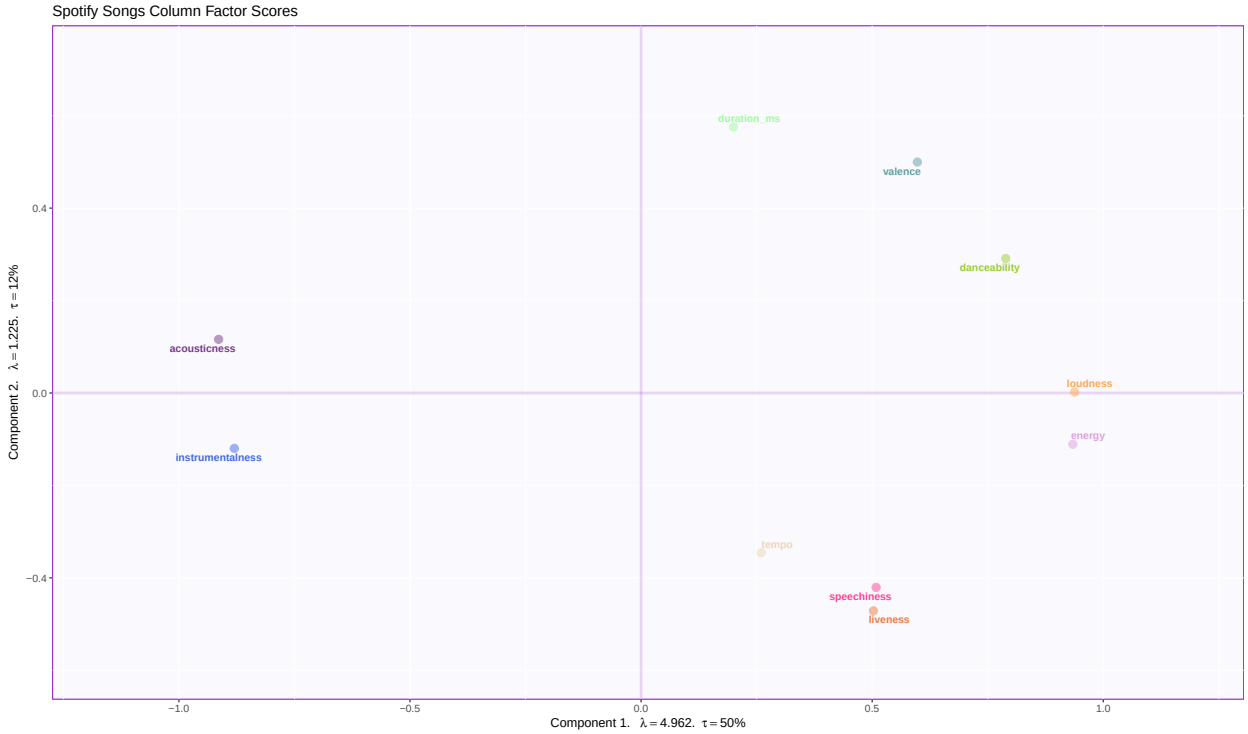
Loadings

The circle of loadings shows the nature of correlations between the columns (variables).The interpretation is similar to that of plain PCA, with the angular distances indicative of the strnegth of correlation between two variables and the closer they are to the circumference, the higher their contribution to the dataset.

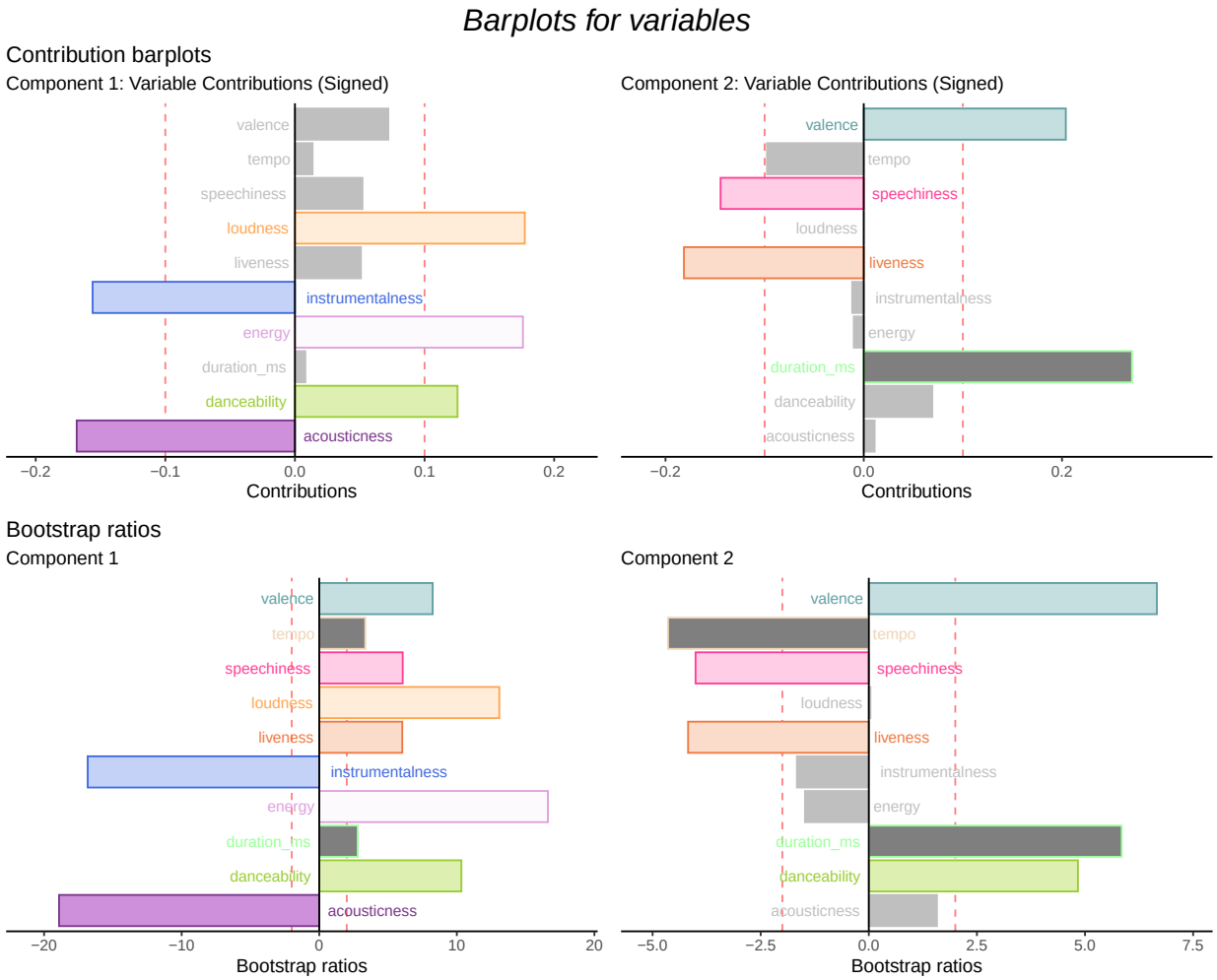


Column factor scores

This plot shows the distribution of the variables in the PCA space. Once again looking at component 1, it is clear that party and workout songs are closely related to variables like danceability, valence, loudness, energy etc. On the other hand, dinner and sleep songs are correlated with instrumentalness and acousticness.


Spotify Songs Column Factor Scores
duration_ms
valence
danceability
acousticness
loudness
energy
instrumentalness
tempo
speechiness
liveness
0.4
0.0
−0.4
−1.0
−0.5
0.0
0.5
1.0
Component 2. λ = 1.225. τ = 12%
Component 1. λ = 4.962. τ = 50%

Contribution barplots and Bootstrap Ratio of columns



Summary

When we interpret the factor scores and loadings together, the PCA revealed:

- Component 1: Sleep and dinner songs are different from party and workout songs. They differ on the characteristics of music such as danceability, loudness, valence, instrumentalness etc.

Certain variables contribute more to the variance in the dataset than the others (acousticness, instrumentalness, valence, energy, danceability, loudness).

Sleep and dinner songs are closely associated with acousticness and instrumentalness, while party and workout songs are explained more by valence, energy, danceability, and loudness.

Correspondence Analysis

Method: Correspondence analysis

Correspondence analysis is generalized principal component analysis for qualitative data. It can be used to analyze one data table by transforming it into two sets of factor scores. The factor scores represent the similarity structure between the rows and columns. Elements have *masses* and *weights*. The mass of each row represents its importance in the table, i.e., the mass of row i is the proportion of i with respect to the entire table. On the other hand, columns have weights, which reflect the information that a column provides to the identification of a given row.

There are two kinds of plots - *symmetric* and *asymmetric*. Asymmetric plots can be interpreted in a straightforward manner - the distance between the row and column factor scores is directly interpretable. In a symmetric plot, one of the factor scores is normed/standardized before plotting.

Data set: Powder soft drinks and emotions

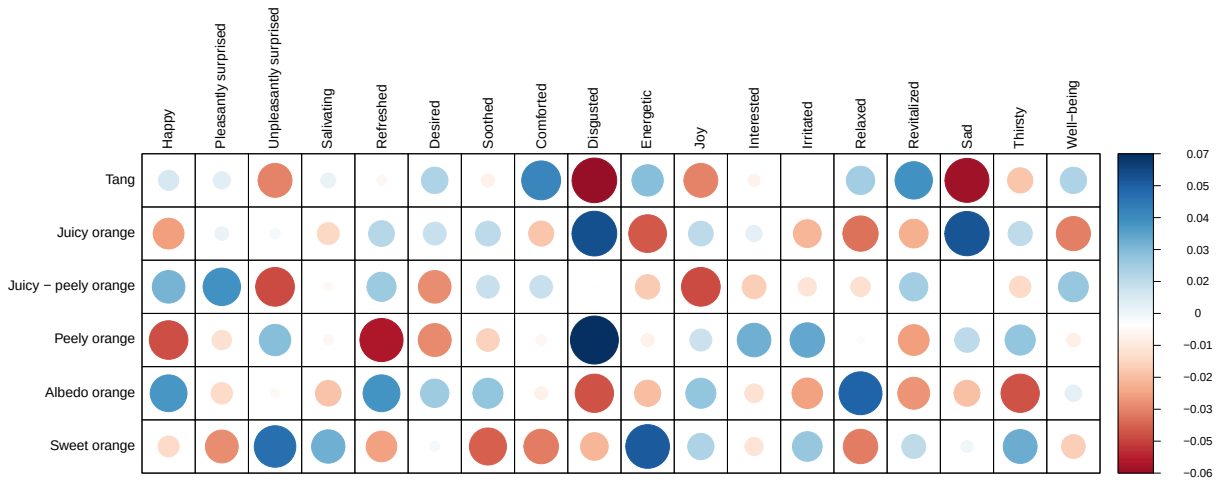
Researchers evaluated whether different orange flavor profiles would be better associated with the Tang brand, with respect to emotional response. All 73 respondents were women from Argentina, with 50% aged between 25 and 35 years and the rest between 36 and 49 years. Only one question was asked - how do you feel when you taste the product? Participants were presented with the stimuli (orange juice) in a sequential monadic way and they could choose an emotion, all related to the product, from a list.

	Happy	Pleasantly surprised	Unpleasantly surprised	Salivating	Refreshed	Desired	Sooth
Tang	10	9	4	7	14	8	
Juicy orange	6	8	6	5	15	7	
Juicy - peely orange	11	11	3	6	16	4	
Peely orange	5	7	8	6	8	4	
Albedo orange	12	7	6	5	18	8	
Sweet orange	7	5	9	8	10	6	

Chi-square matrix

Unlike PCA, correspondence analysis takes into account chi-square values, so we generate a chi-square matrix instead of the standard correlation plot.

In contrast to the correlation plot which tells us how two variables “move” together/in opposite directions, the chi-square matrix shows us the distance between the two variables. This is more interpretable when the data is qualitative (in this case, the choice of emotions associated with orange juice flavors).

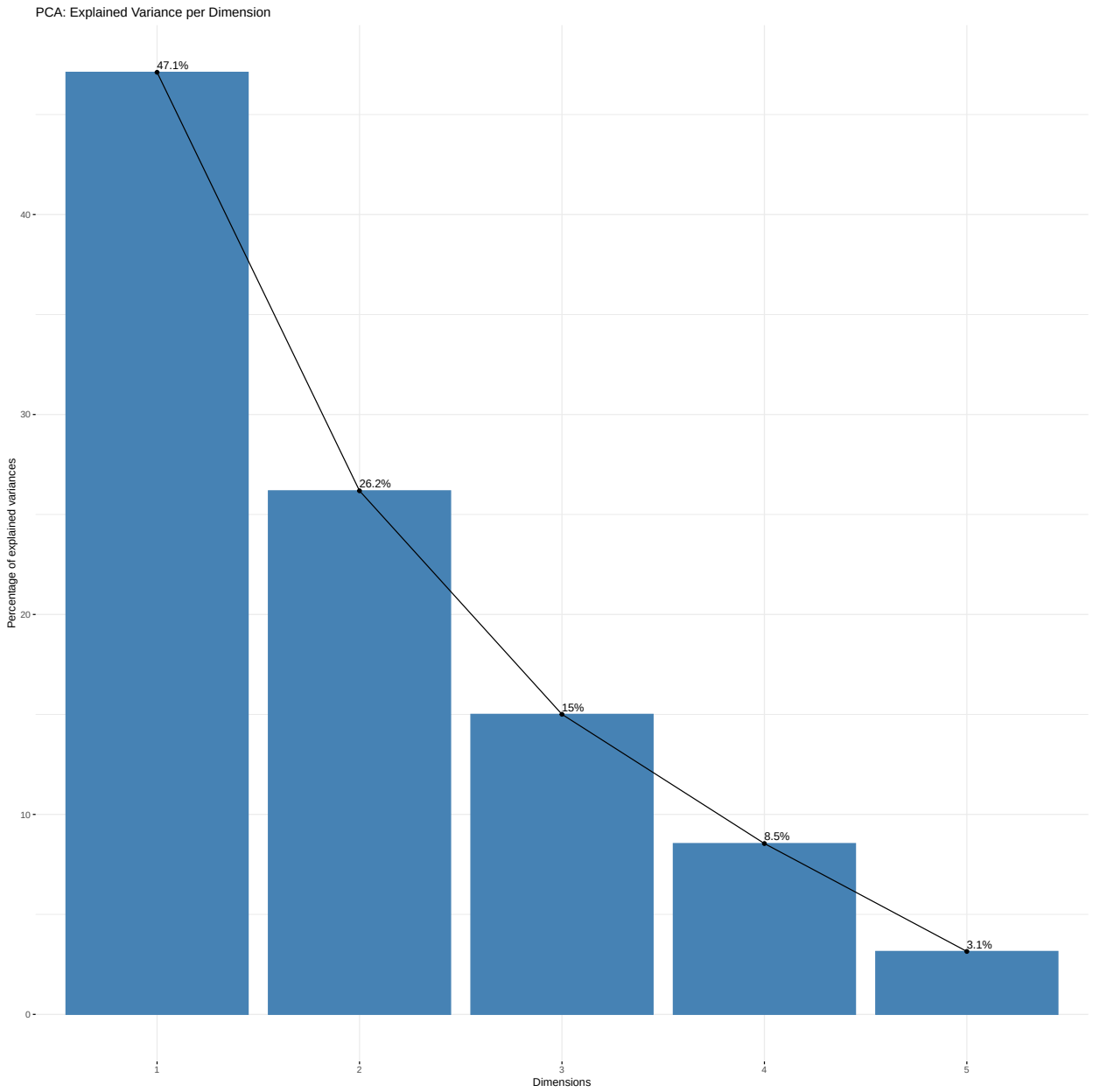


Analysis

```
# run CA
resCA.sym  <- epCA(X, symmetric = TRUE, graphs = FALSE)
resCAinf.sym4bootJ  <- epCA.inference.battery(X, symmetric = TRUE, graphs = FALSE, test.iters = 25)
resCAinf.sym4bootI  <- epCA.inference.battery(t(X), symmetric = TRUE, graphs = FALSE, test.iters = 25)
```

Scree Plot

The scree plot shows the dimensions extracted from the correspondence analysis and how much variance they contribute to the dataset. In this case, dimension 1 contributes about 45% of the variance and dimension 2 contributes about 25%.

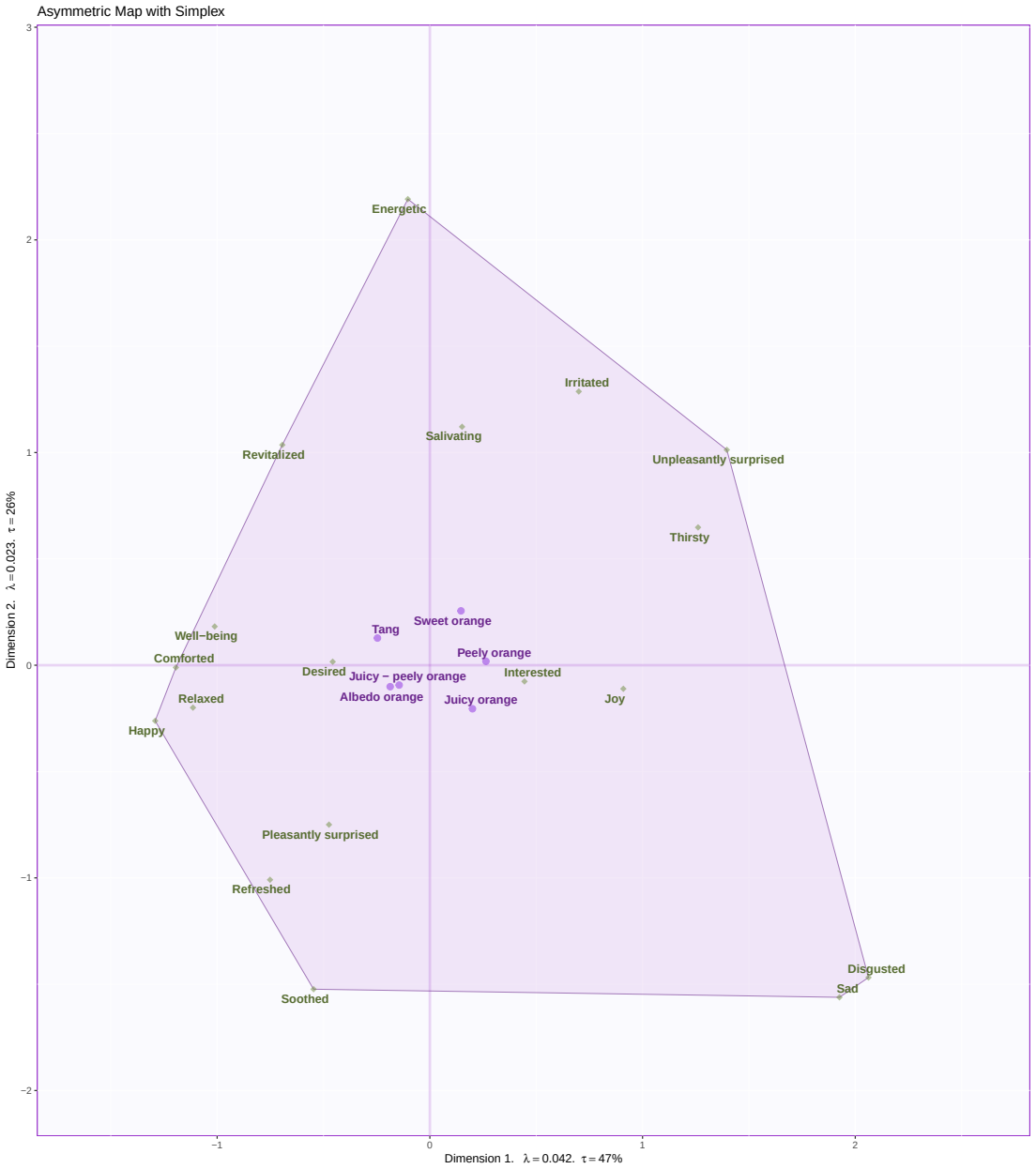


Asymmetric plot

Interpretation:

- Dispersion of the data points in the simplex - The unpleasant emotions are grouped together, while the happy emotions are grouped together (as we would expect intuitively). The distinctness is clear along dimension 1.
- Eigenvalue - In both dimensions, the eigenvalues are relatively small, but dimension 1 explains almost half (47%) the inertia in the dataset.

The distances between the flavors of orange juice and the emotional outcome are directly interpretable in the asymmetric plot. Tang is closely associated with "Desired", while Juicy Orange is associated with "Interested" for example.



Symmetric plot

Unlike the asymmetric plot, the distance between the row and column scores is not directly interpretable since one of them is scaled/normed. The symmetric plot is more useful for understanding the intra row/column distances and relationships.

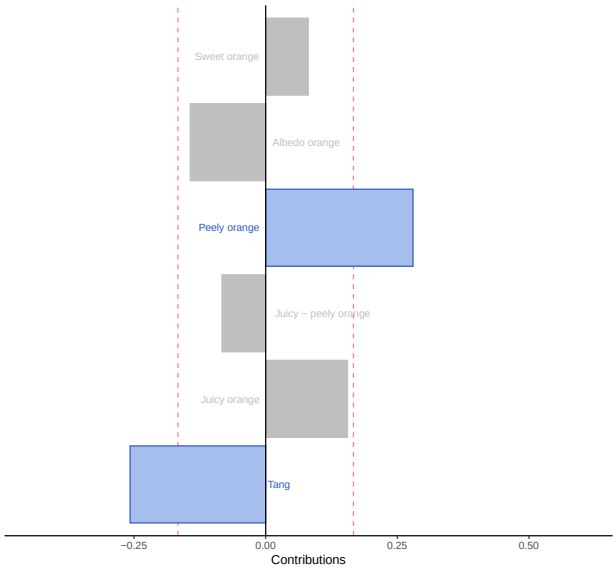
Contributions and bootstrap ratios barplots

Contribution barplots indicate which variables contribute the most significantly to the dataset. The bootstrap ratio barplots show which variables cross the threshold when our data is resampled several times. Basically, whether the same variables hold good (in terms of significant contributions) even after the same experiment is repeated several times.

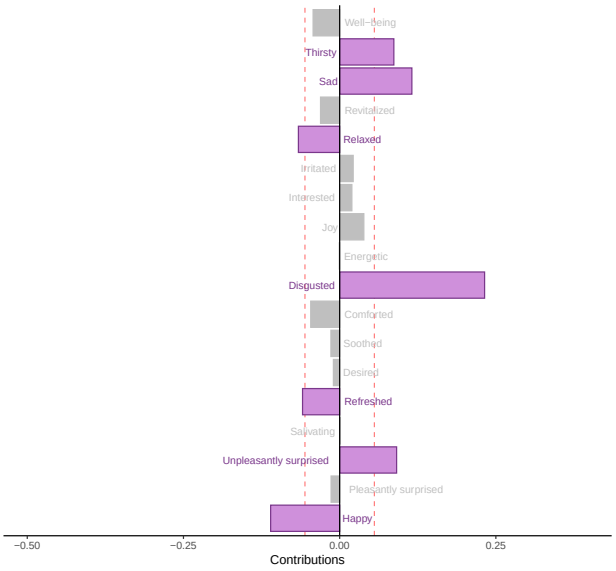
Contributions

Component 1

rows

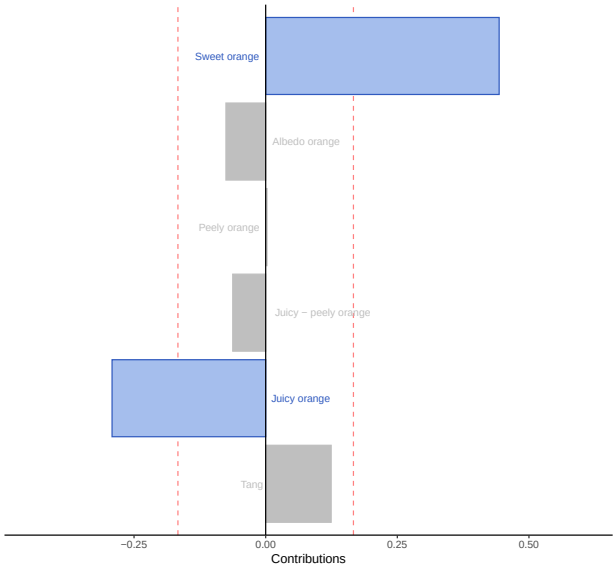


columns

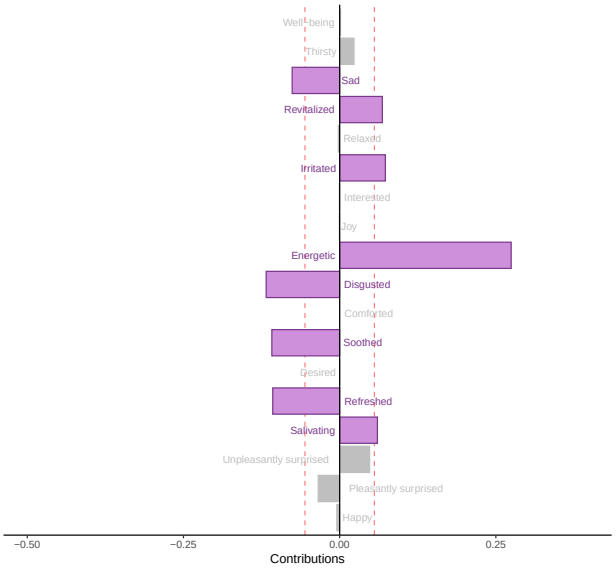


Component 2

rows

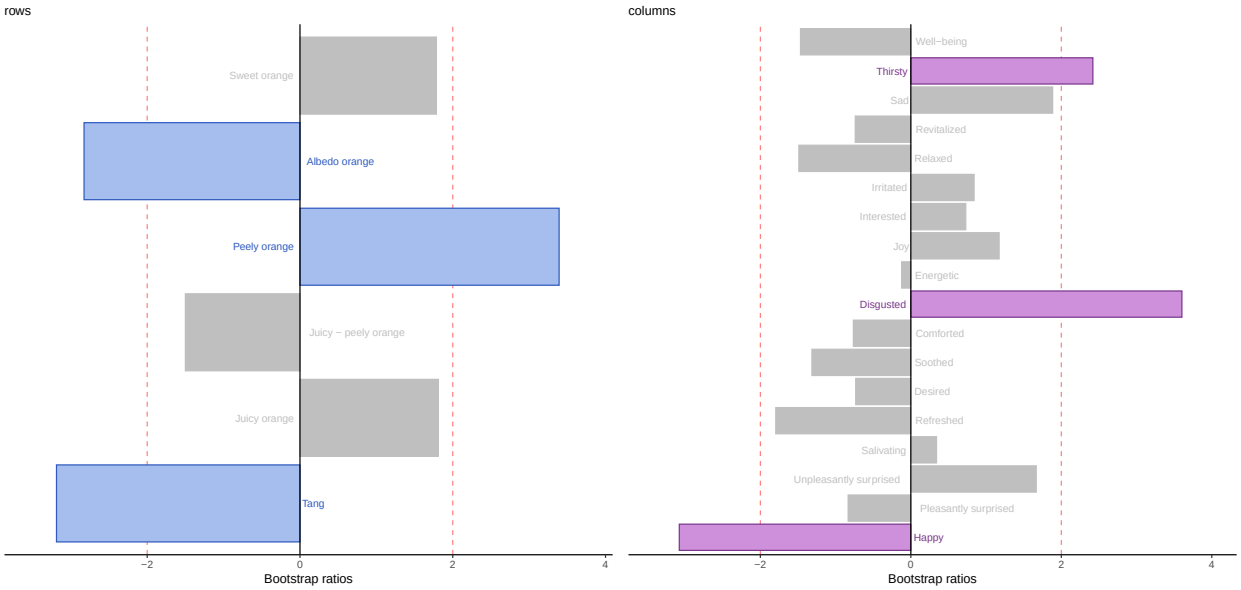


columns

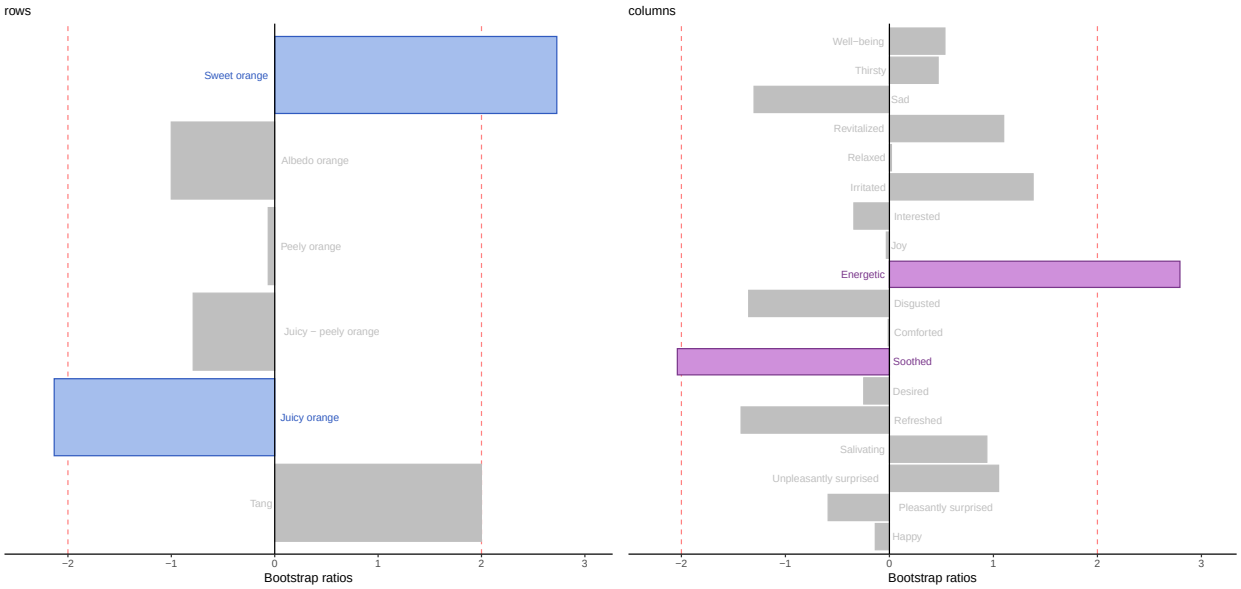


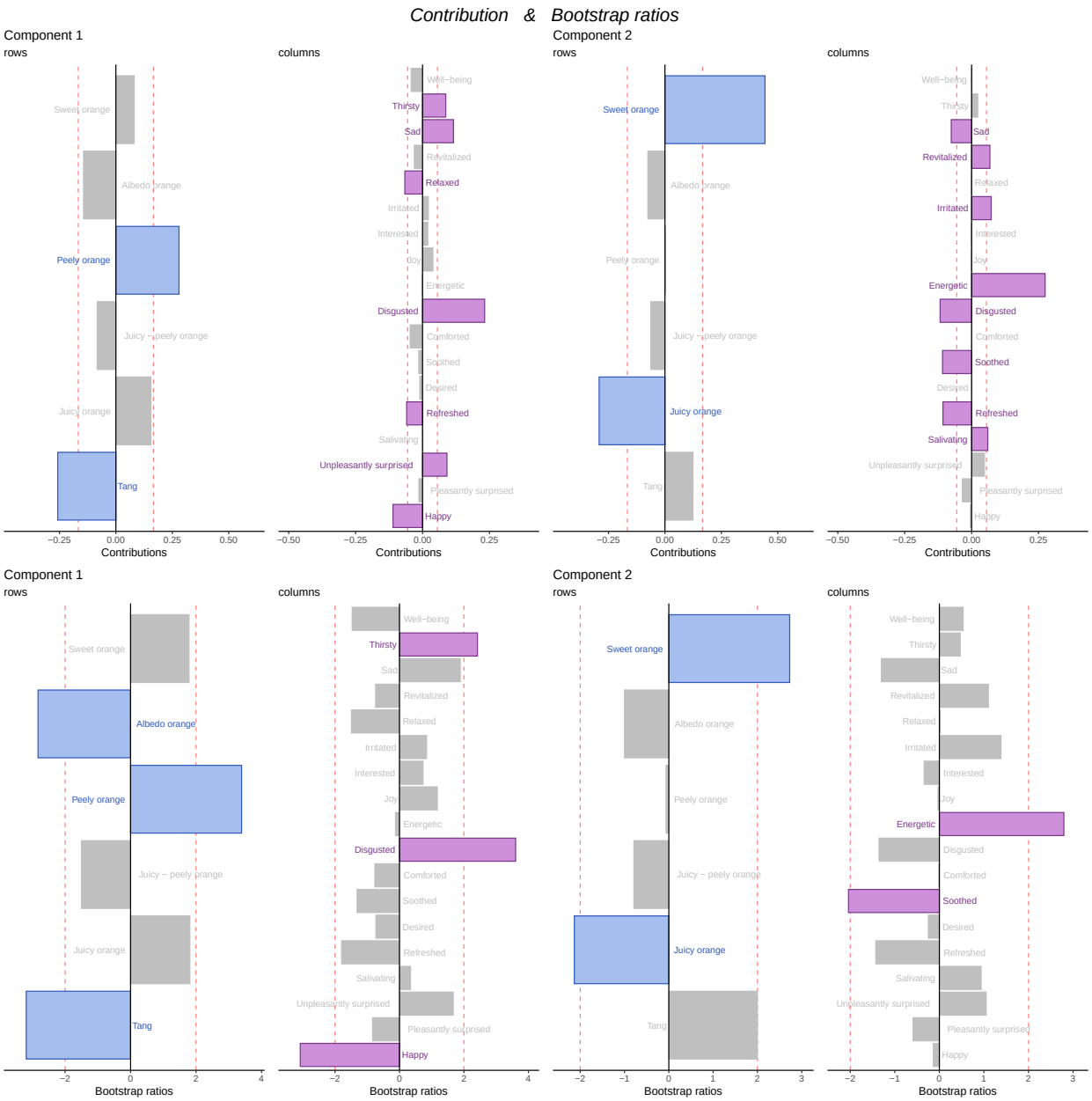
Bootstrap ratios

Component 1



Component 2





Summary

When we interpret the factor scores and loadings together, the CA revealed:

- Preferred plot - Asymmetric (distance between row and column factor scores is directly interpretable)
- Dimension 1: The Tang brand was associated with feelings of comfort and familiarity, while Peely orange was associated with unpleasant feelings.
- Dimension 2: Sweet orange and Juicy-peely orange were associated with positive emotions such as "Energetic".

Multiple Correspondence Analysis

Method: Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is an extension of correspondence analysis (CA) which explores the pattern of relationships of categorical dependent variables. CA is technically principal component analysis when the data table is qualitative instead of quantitative.

In MCA, the dataset to be supplied is a set of nominal variables. Each level of the nominal variable is coded as 0 or 1. For example, for gender, the code for a male respondent will be 1 0. That way, the whole data table will consist of columns of 0s and 1s. Per nominal variable only 1 column can have 1. Rest should be 0.

MCA can also be run on quantitative data (such as this example). An extra step is to convert them into categorical data by binning (see sample below).

The idea behind coding is that each row has a total of 1, which in CA implies that each row has the same mass

Source: Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage. pp. 651-657

Data set: Audio features

This is a dataset which describes audio features of songs in Spotify playlists. Specifically, the music.track dataset measures 165 songs on 16 variables, of which 11 are quantitative. Some of the audio features described are acousticness, danceability, and energy.

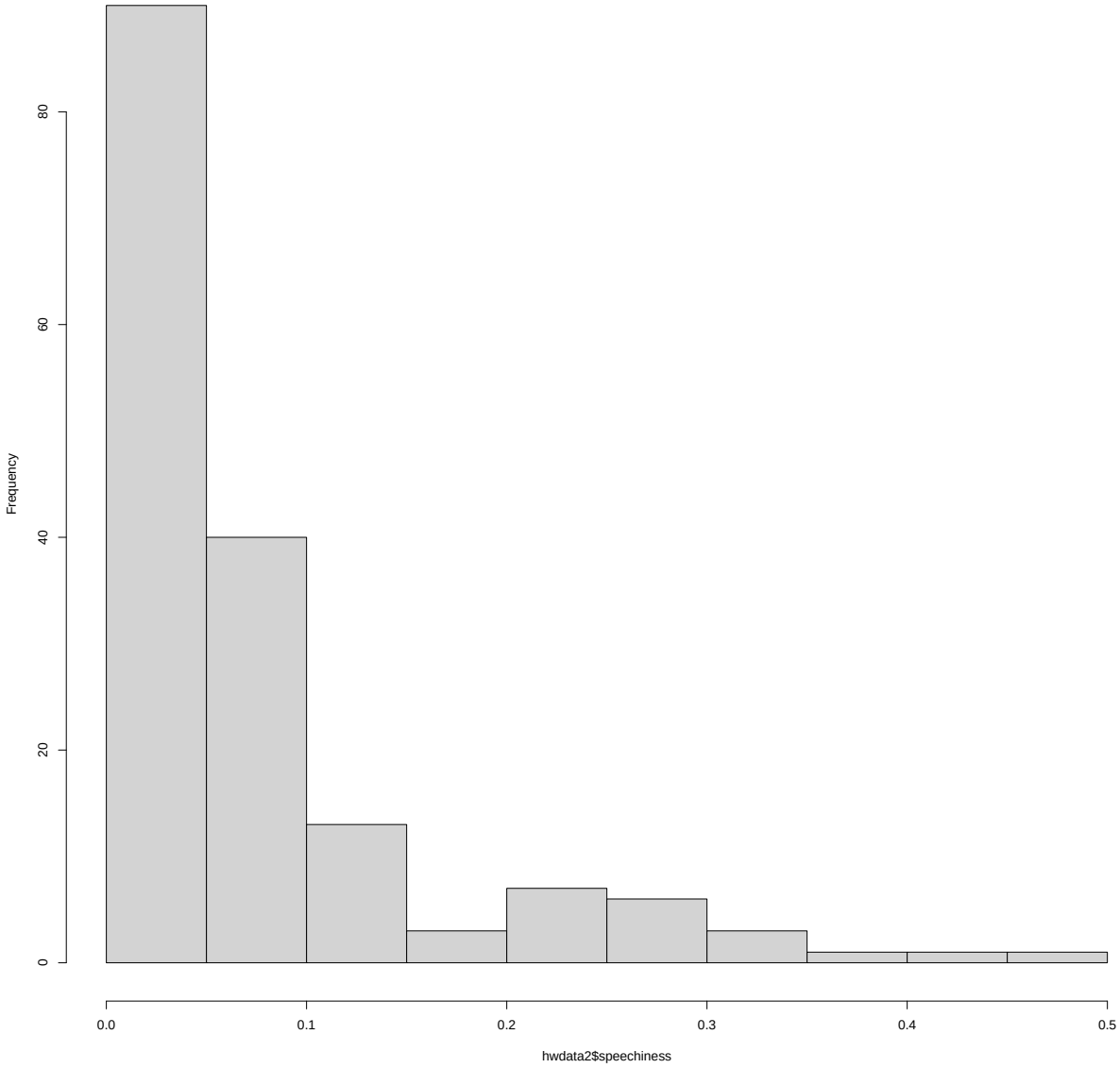
To analyze using MCA, we have to convert the data table into a categorical dataset. So, we bin the variables depending on what the histograms for each of the variables look like.

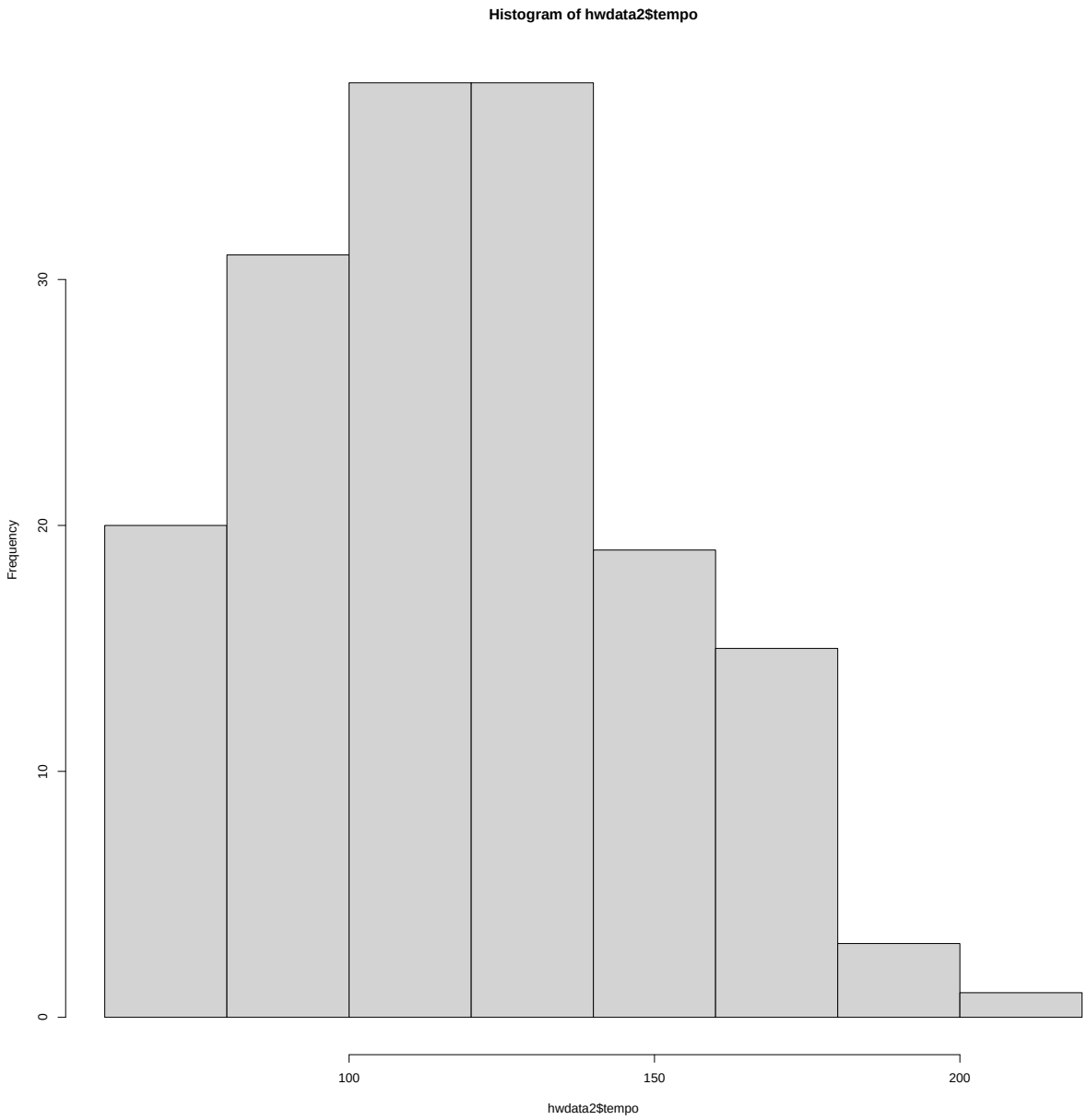
The table below shows the dataset along with the categorical variables.

	tempo	valence	Acous_ch	Dance_ch	Dur_ch	Ener_ch	Ins_ch	Live_ch	Loud_ch	Speech_ch	T
16	128.346	0.343	highAc	medDance	highDur	medEner	medIns	lowLiv	highLoud	lowSpeech	m
48	90.065	0.217	highAc	medDance	medDur	lowEner	lowIns	medLiv	medLoud	lowSpeech	lo
2	93.885	0.687	highAc	medDance	medDur	medEner	medIns	lowLiv	medLoud	lowSpeech	lo
46	80.626	0.400	medAc	medDance	highDur	medEner	lowIns	medLiv	highLoud	lowSpeech	lo
45	88.989	0.580	lowAc	highDance	medDur	medEner	lowIns	medLiv	highLoud	lowSpeech	lo

Sample binning of data

Here are two examples of how the data was binned. Based on the histogram, the speechiness variable was divided into two groups - low and high speechiness. On the other hand, the tempo histogram had more scope to be finely divided, hence three categories - low, medium, and high tempos.


Histogram of hwdata2$speechiness
Frequency
0
20
40
60
80
0.0
0.1
0.2
0.3
0.4
0.5
hwdata2$speechiness



Analysis

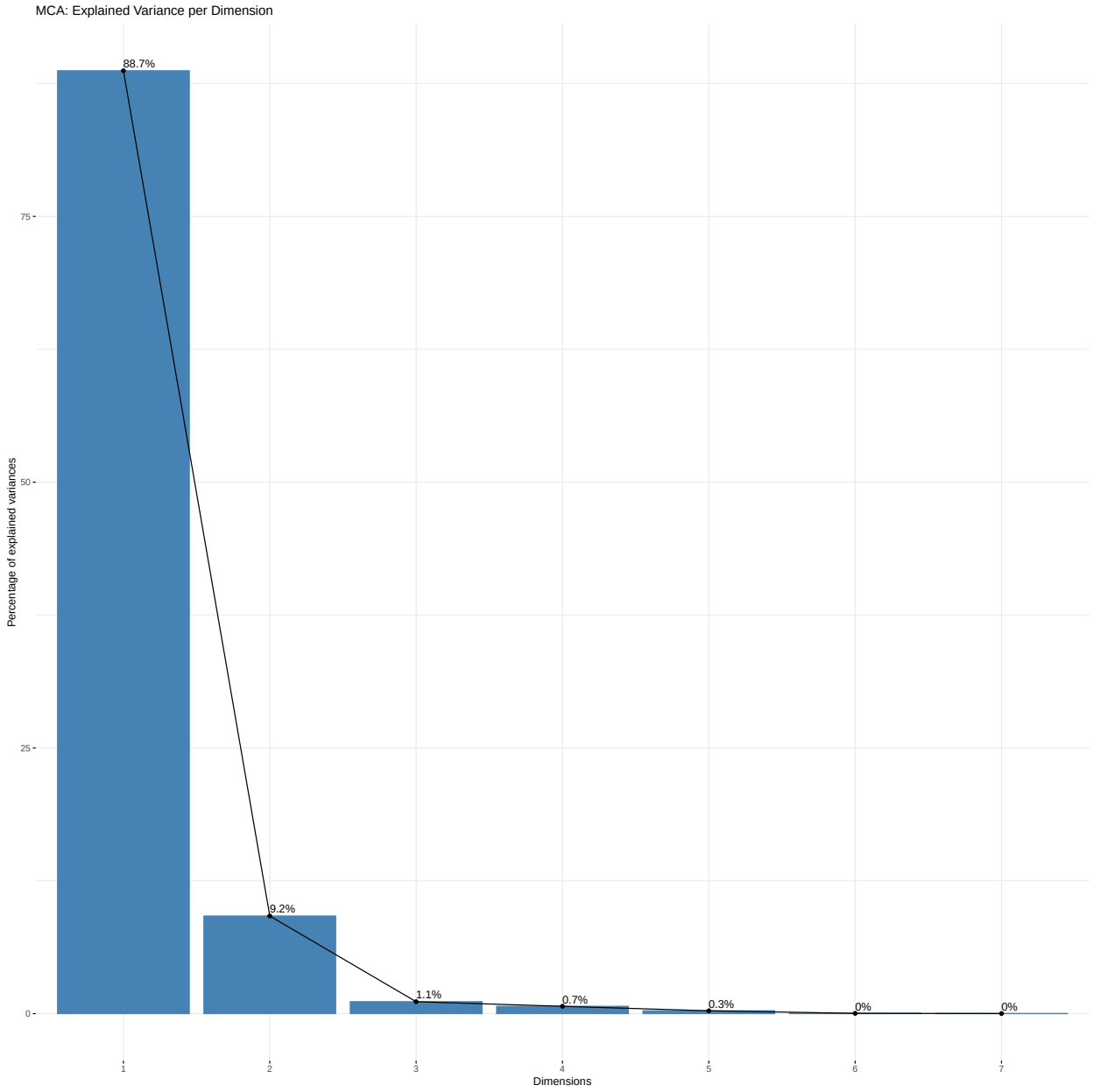
```
resMCA <- epMCA(hwdata, graphs = FALSE)

resMCA.inf <- epMCA.inference.battery(hwdata,
                                      graphs = FALSE)
```

Scree plot

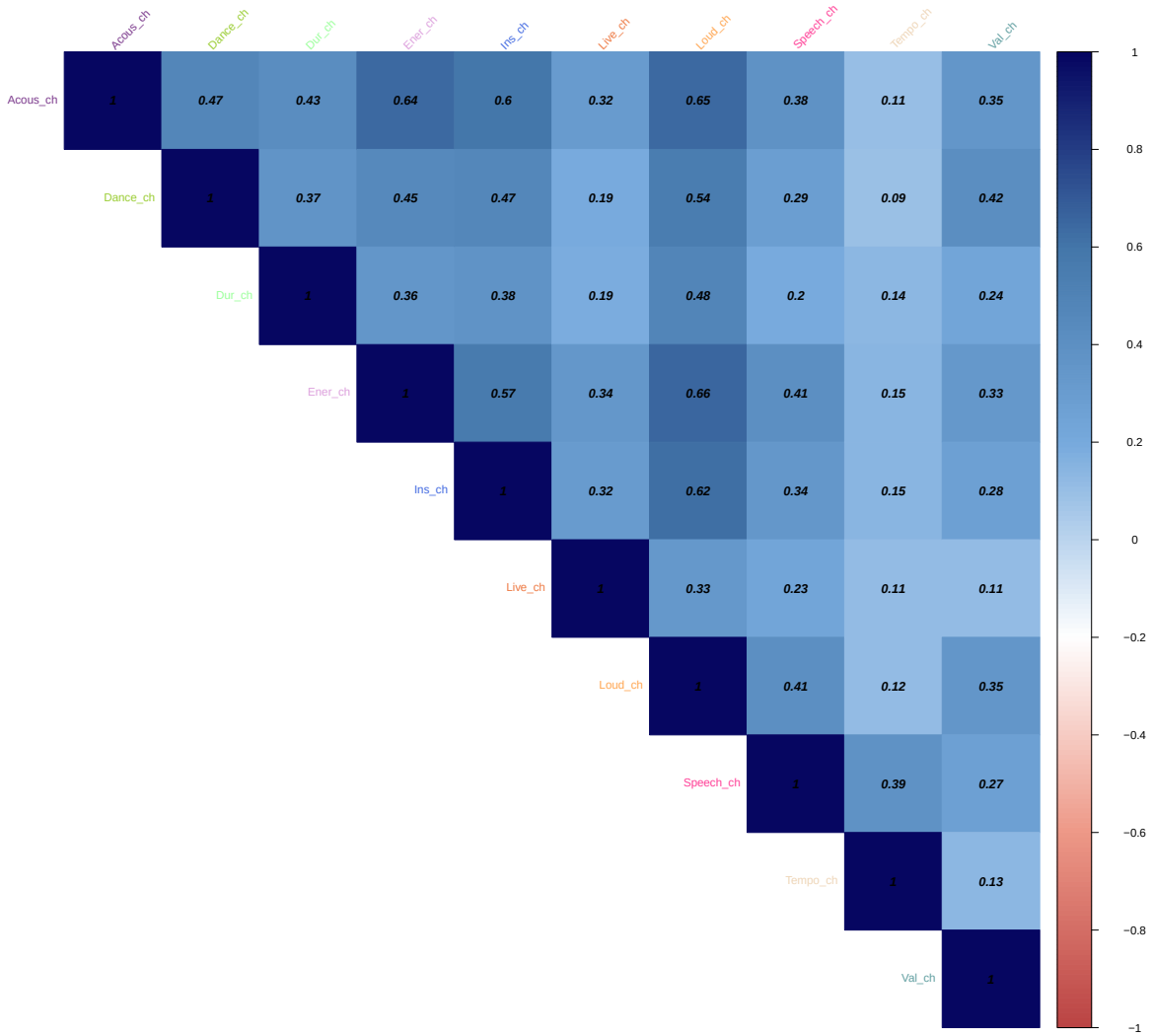
The scree plot is used to determine how many of the dimensions could be interpreted based on the amount of variance that is contributed to the dataset. As per this plot, Component 1 is explaining about 85% of

the variance, while dimension 2 explains about 10%. The purple dots are the significant dimensions worth taking a look at based on inferential MCA.



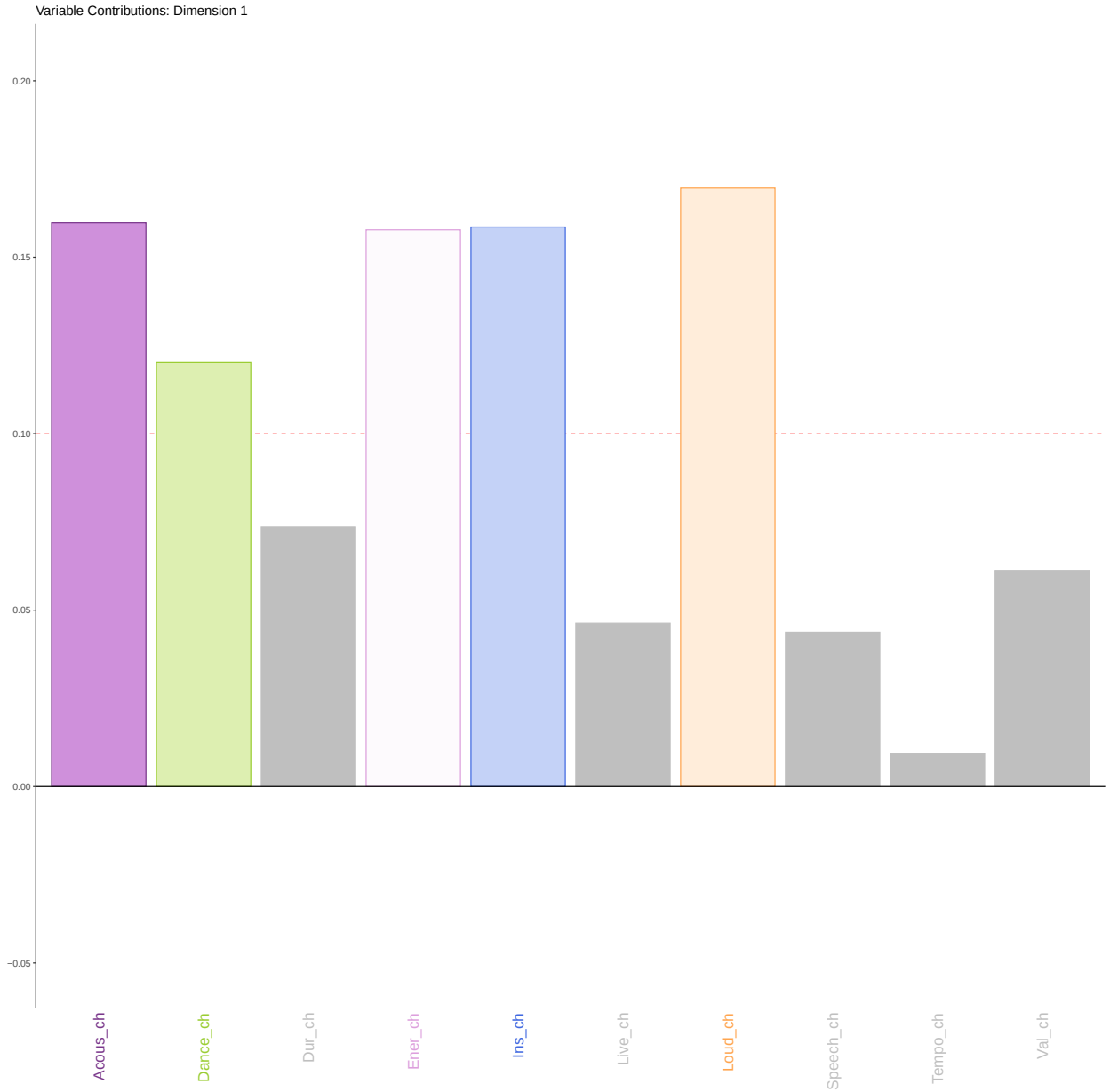
The pseudo-correlation heatmap

We can use this representation of the data to compare with plots generated while conducting a Principal Component Analysis. It is called a "pseudo" correlation because the numbers are not directly calculated but have to be derived from the contingency tables.

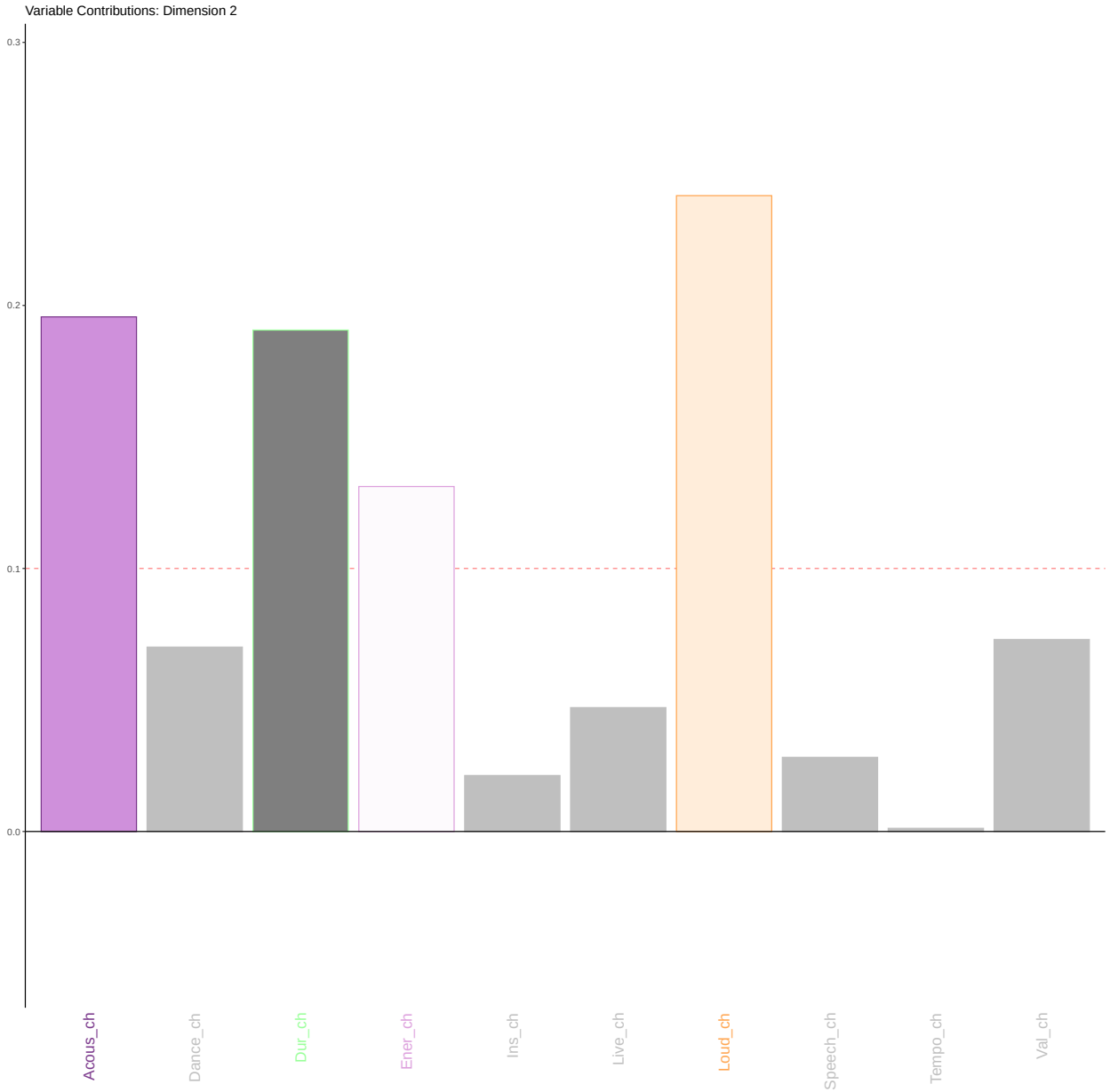


Variable contributions

In Dimension 1, the variables contribute to the inertia as outlined in the barplot below.

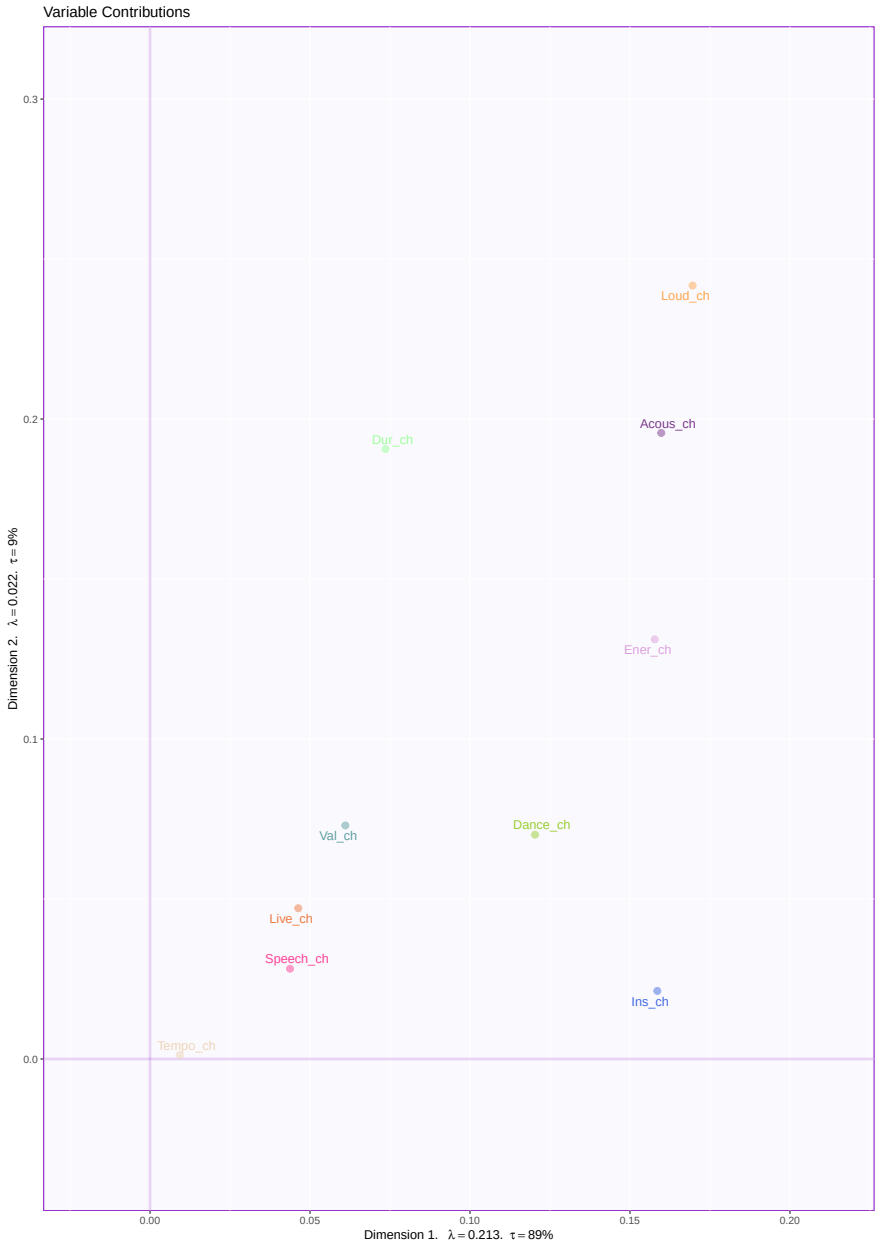


In Dimension 2, the variables contribute to the inertia as outlined in the barplot below.

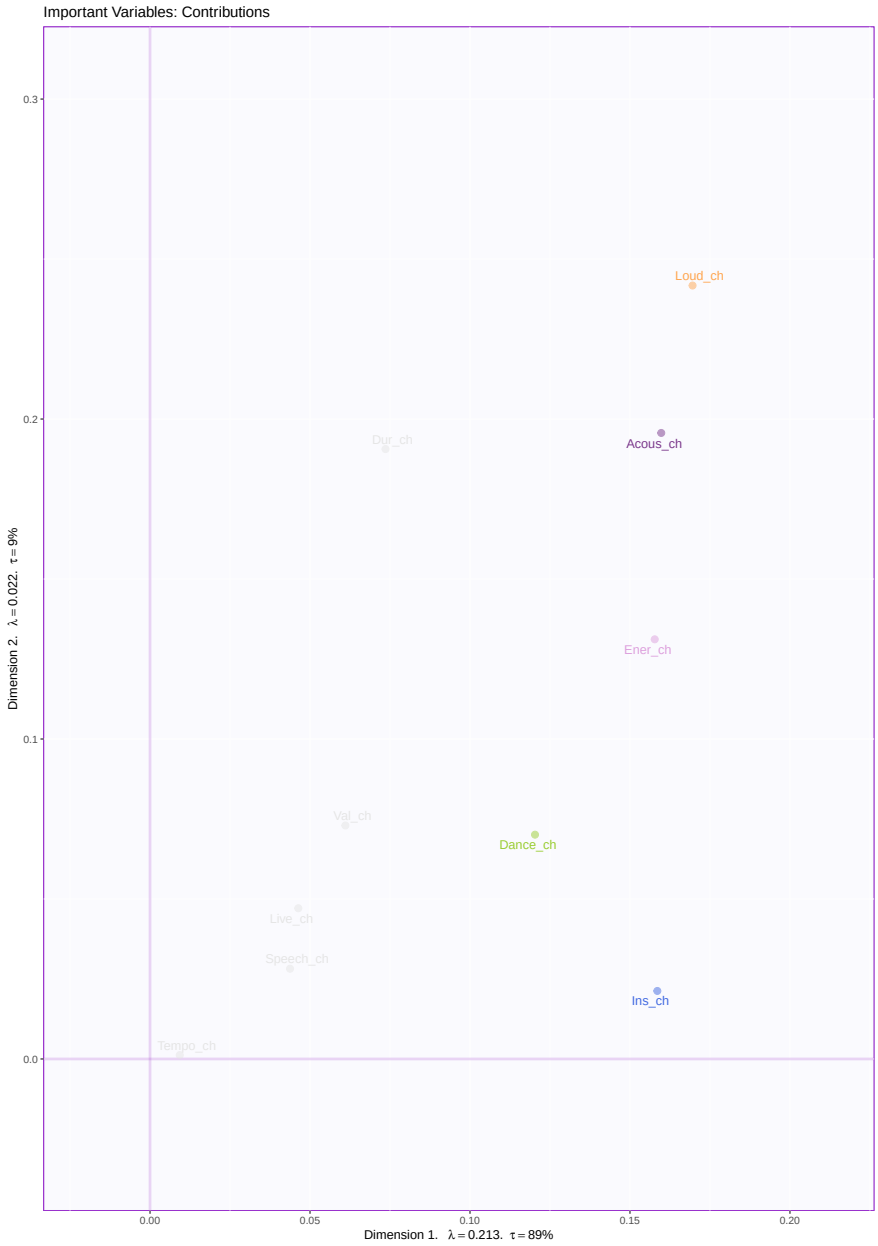


Factor scores

The variable factor scores indicate how closely related the columns are to each other. Further, we can also see which variables contribute relatively more to the inertia in a particular dimension.



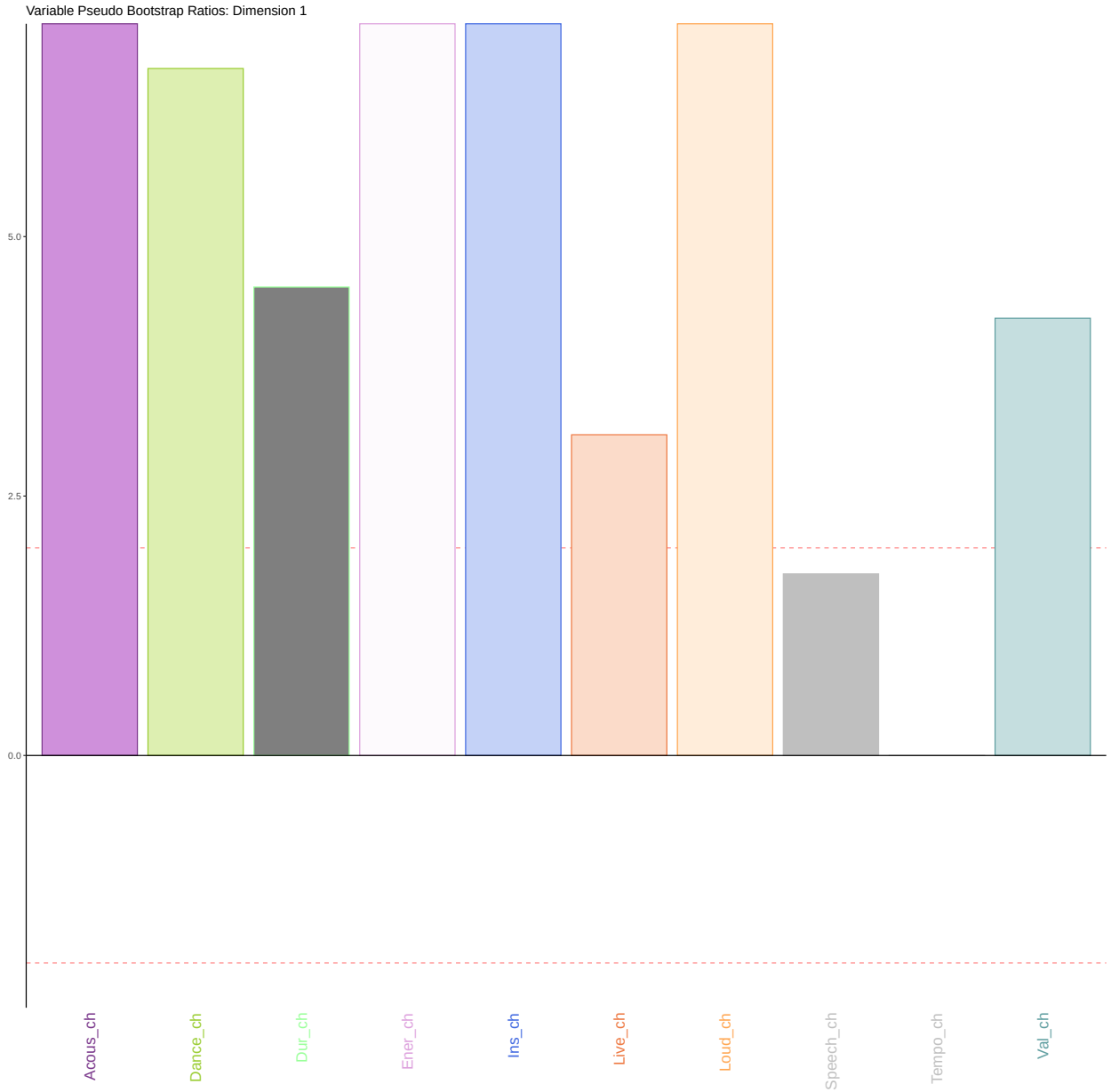
Among all the factor scores, the ones that are important are colored and the others are grayed out in the following plot.



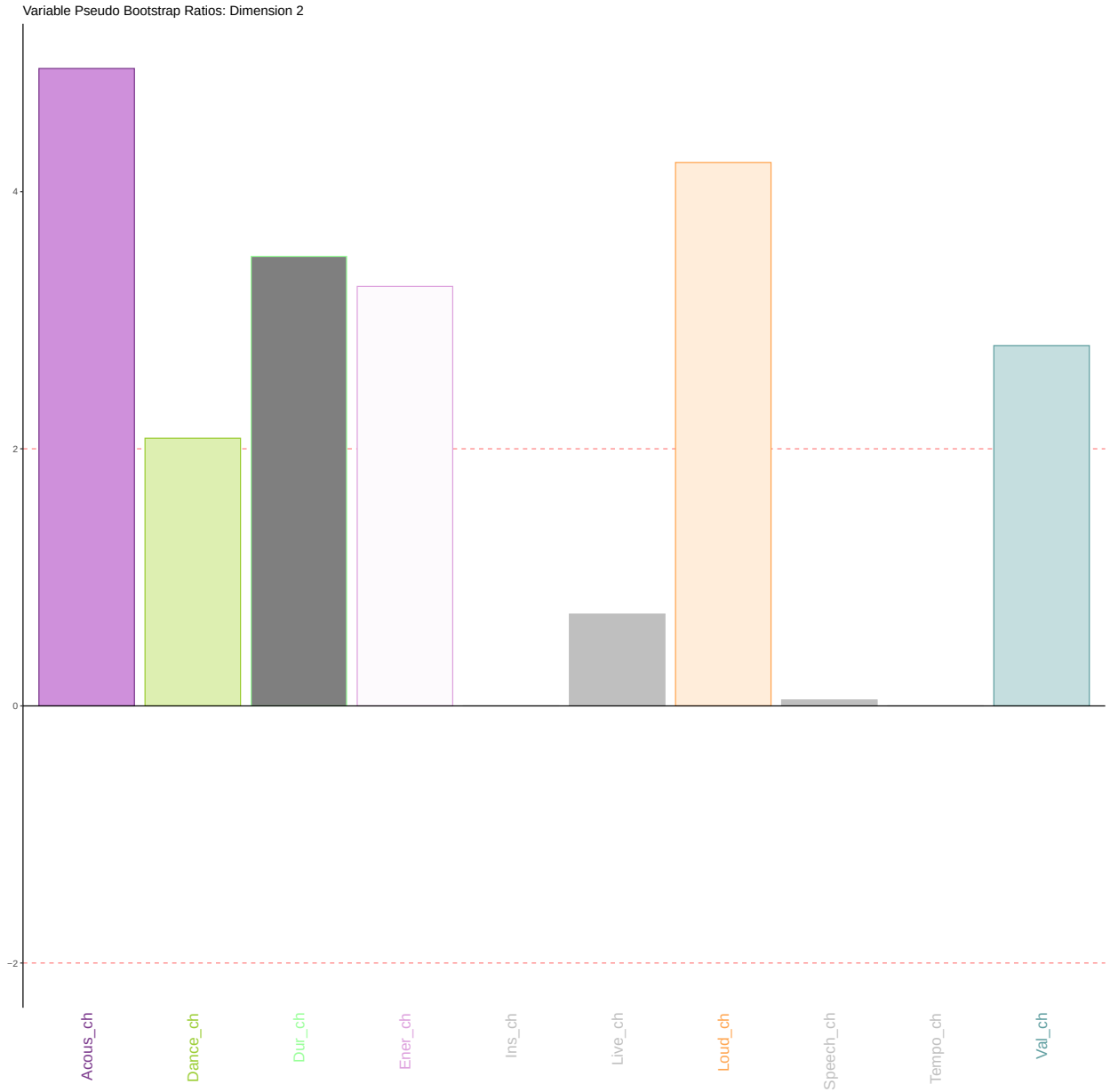
Pseudo Bootstrap ratios

The following barplots show pseudo bootstrap ratios for all the variables.

Dimension 1:

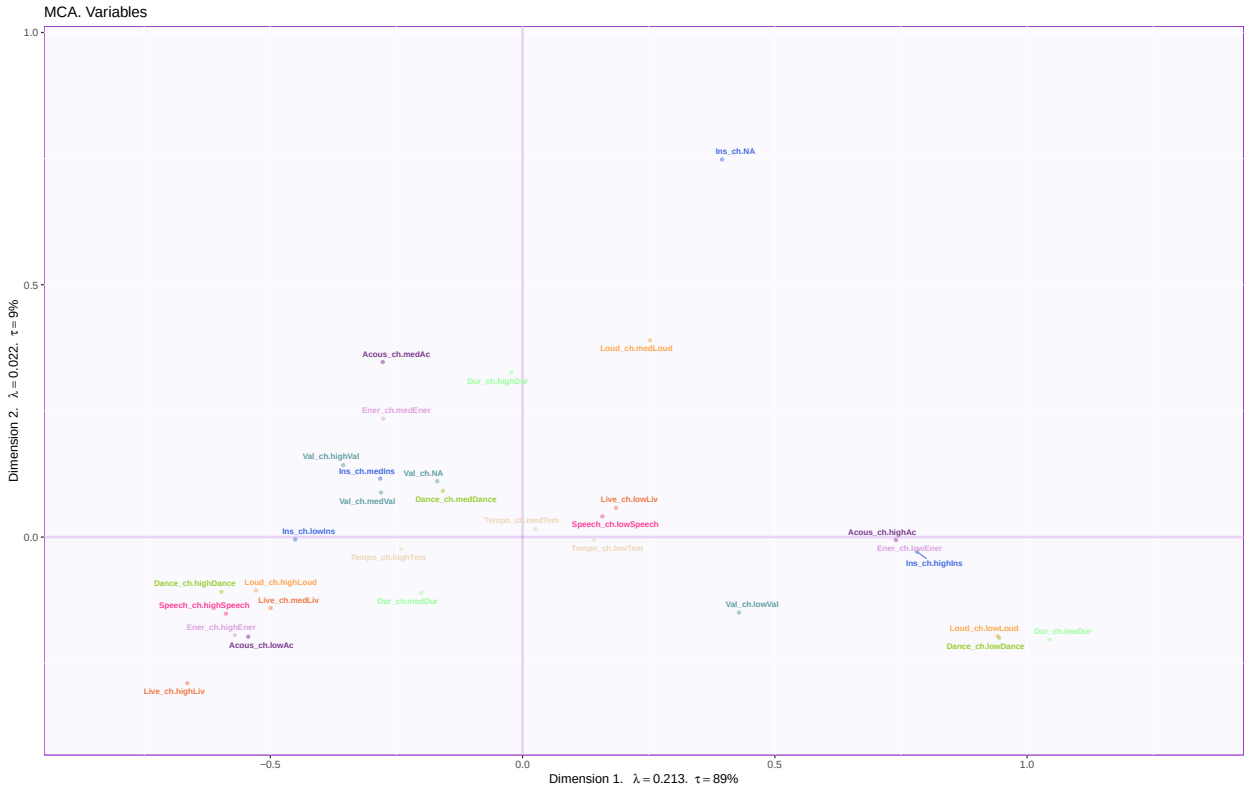

Variable Pseudo Bootstrap Ratios: Dimension 1
5.0
2.5
0.0
Acous_ch
Dance_ch
Dur_ch
Ener_ch
Ins_ch
Live_ch
Loud_ch
Speech_ch
Tempo_ch
Val_ch

Dimension 2:



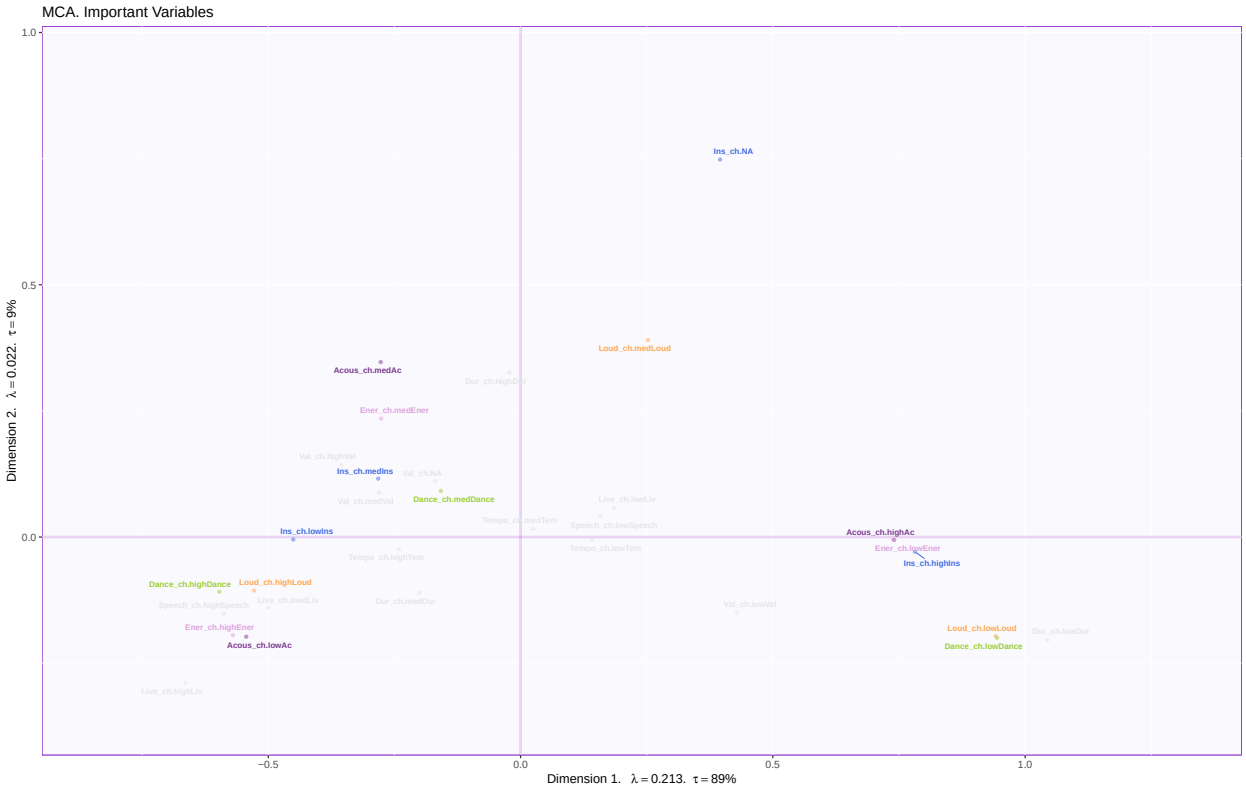
Important variables and their contribution to the dataset

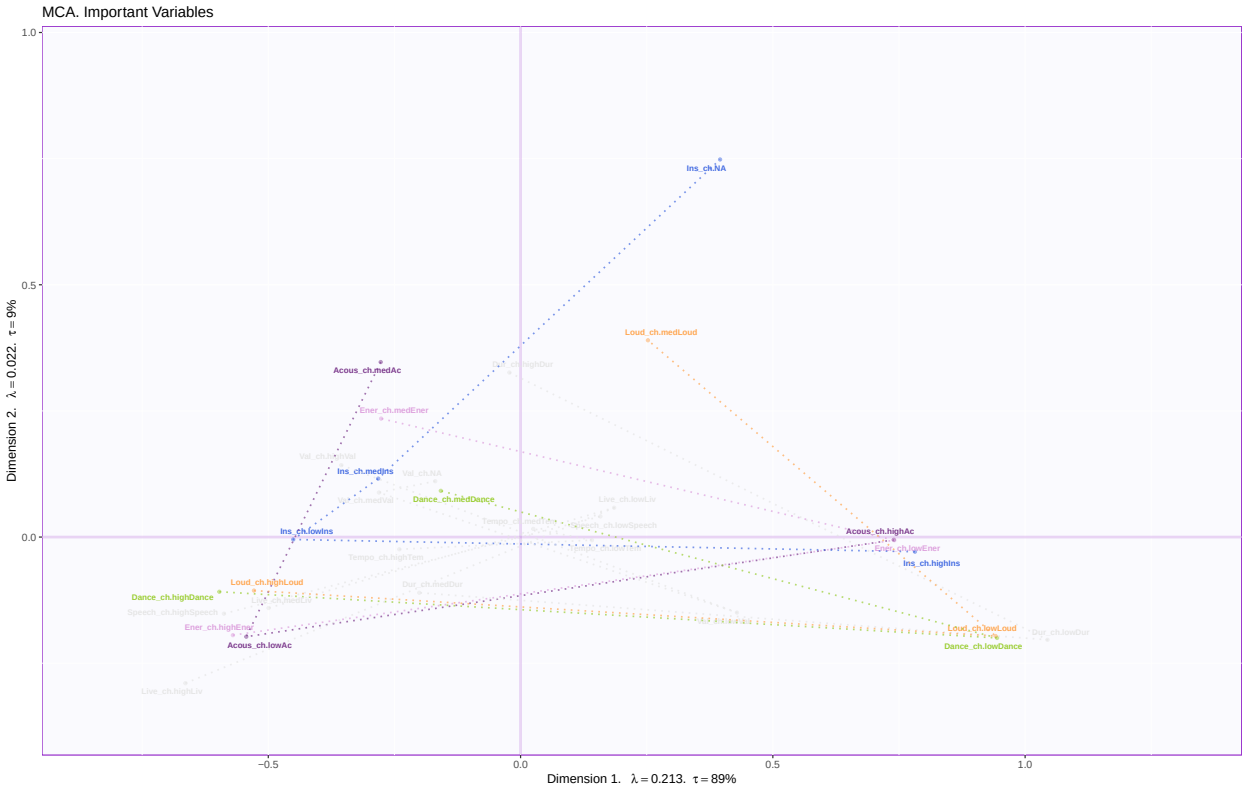
Going another step, we can plot the levels of the variables along the dimensions to understand more about what exactly in the variables are important. There is no single trend for how the variables vary from lower to higher levels along the dimensions.

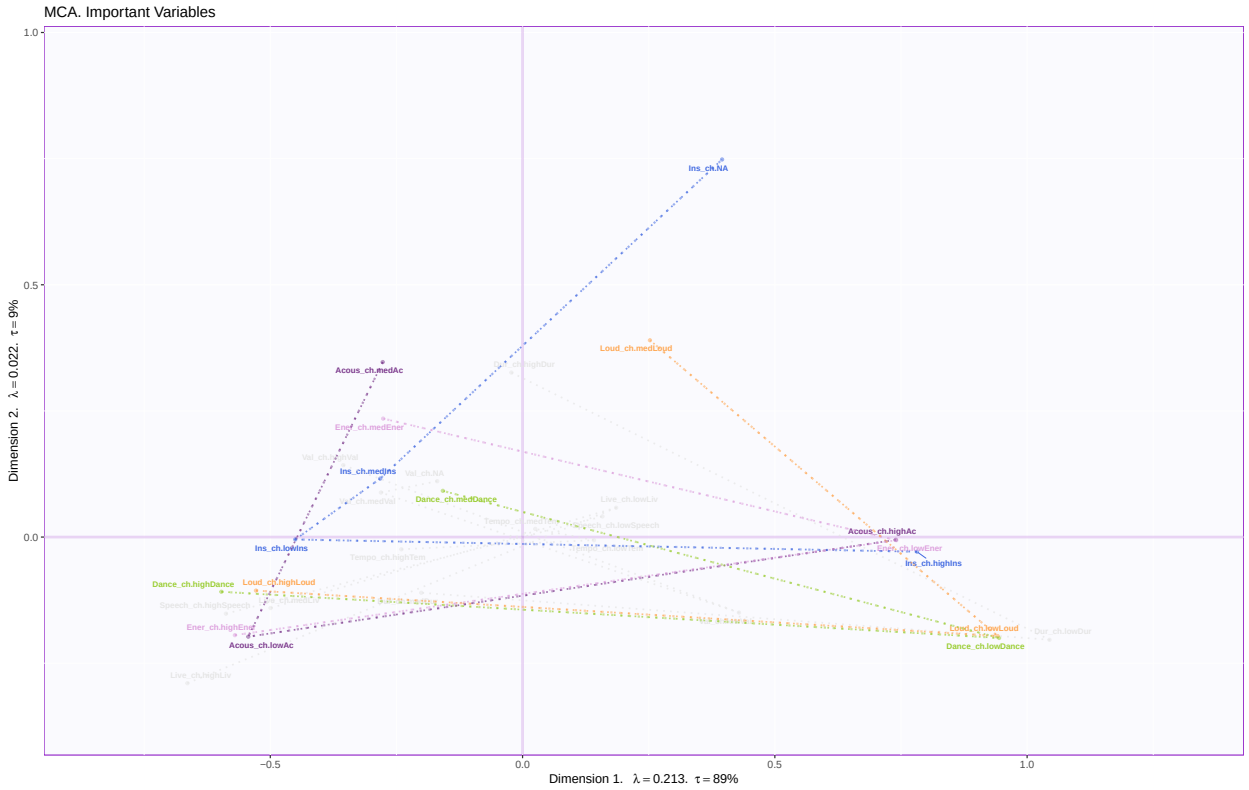


Important variable levels (connected by lines):

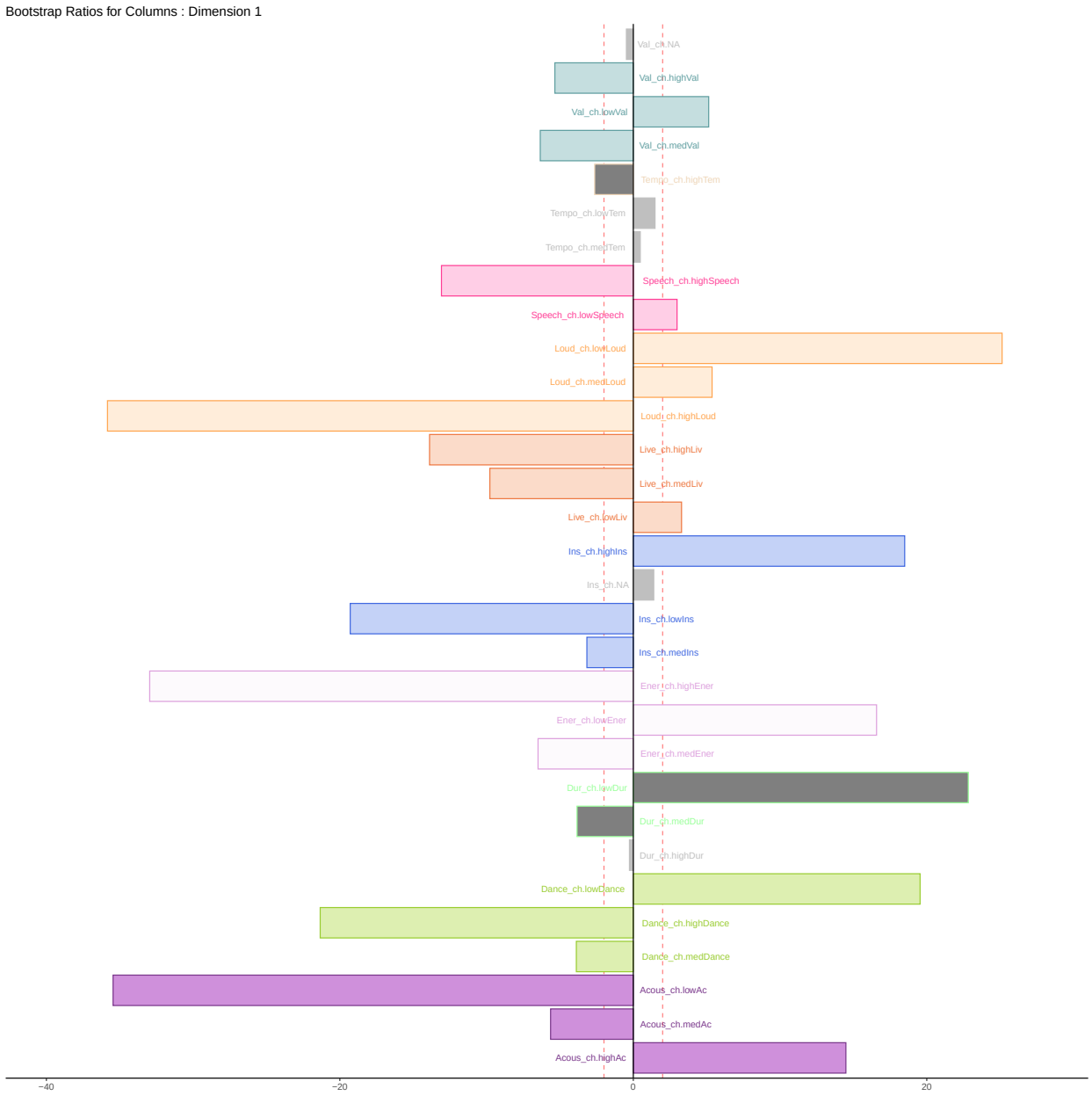
The other variables are in grayscale


MCA. Important Variables
Dimension 2. λ = 0.022. τ = 9%
Dimension 1. λ = 0.213. τ = 89%
1.0
0.5
0.0
−0.5
0.0
0.5
1.0
Ins_ch.NA
Loud_ch.medLoud
Acous_ch.medAc
Dur_ch.highDur
Ener_ch.medEner
Val_ch.highVal
Ins_ch.medIns
Val_ch.NA
Dance_ch.medDance
Live_ch.lowLiv
Tempo_ch.medTem
Speech_ch.lowSpeech
Ins_ch.lowIns
Acous_ch.highAc
Ener_ch.lowEner
Tempo_ch.lowTem
Ins_ch.highIns
Tempo_ch.highTem
Dance_ch.highDance
Loud_ch.highLoud
Live_ch.medLiv
Dur_ch.medDur
Speech_ch.highSpeech
Val_ch.lowVal
Ener_ch.highEner
Loud_ch.lowLoud
Dur_ch.lowDur
Acous_ch.lowAc
Dance_ch.lowDance
Live_ch.highLiv


MCA. Important Variables
Dimension 2. λ = 0.022. τ = 9%
Dimension 1. λ = 0.213. τ = 89%
1.0
0.5
0.0
−0.5
0.0
0.5
1.0
Ins_ch.NA
Loud_ch.medLoud
Acous_ch.medAc
Ener_ch.medEner
Ins_ch.medIns
Dance_ch.medDance
Ins_ch.lowIns
Acous_ch.highAc
Ener_ch.highEner
Ins_ch.highIns
Loud_ch.highLoud
Dance_ch.highDance
Ener_ch.highEnv
Acous_ch.lowAc
Loud_ch.lowLoud
Dance_ch.lowDance

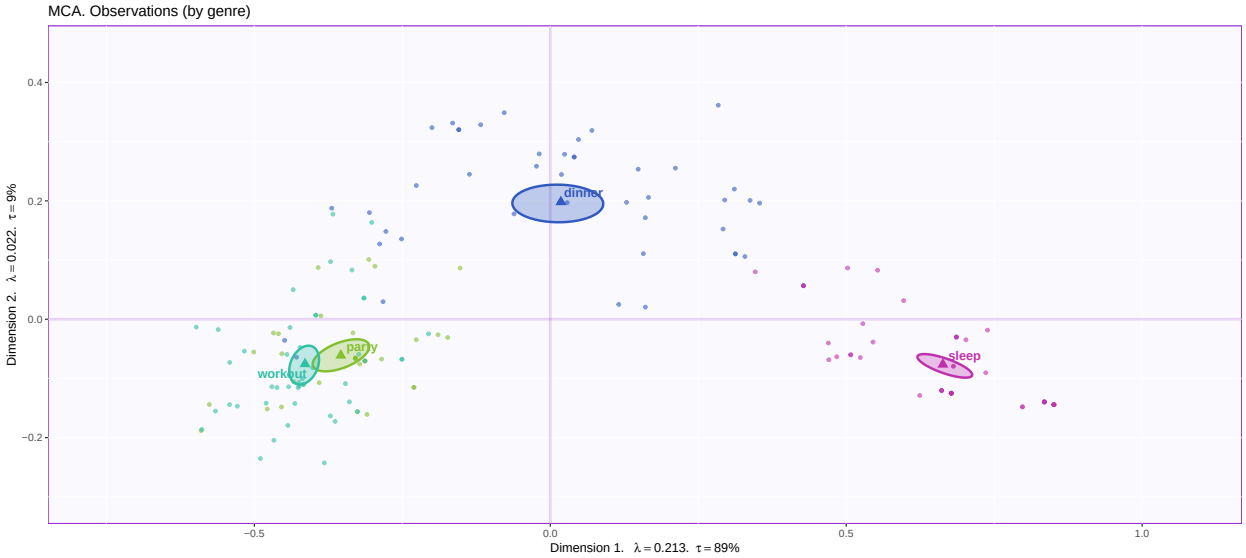

MCA. Important Variables
Dimension 2. λ = 0.022. τ = 9%
Dimension 1. λ = 0.213. τ = 89%
1.0
0.5
0.0
−0.5
0.0
0.5
1.0
Ins_ch.NA
Loud_ch.medLoud
Acous_ch.medAc
Ener_ch.medEner
Ins_ch.medIns
Dance_ch.medDance
Acous_ch.highAc
Ener_ch.lowEner
Ins_ch.highIns
Ins_ch.lowIns
Dance_ch.highDance
Loud_ch.highLoud
Loud_ch.lowLoud
Dance_ch.lowDance
Ener_ch.highEner
Acous_ch.lowAc

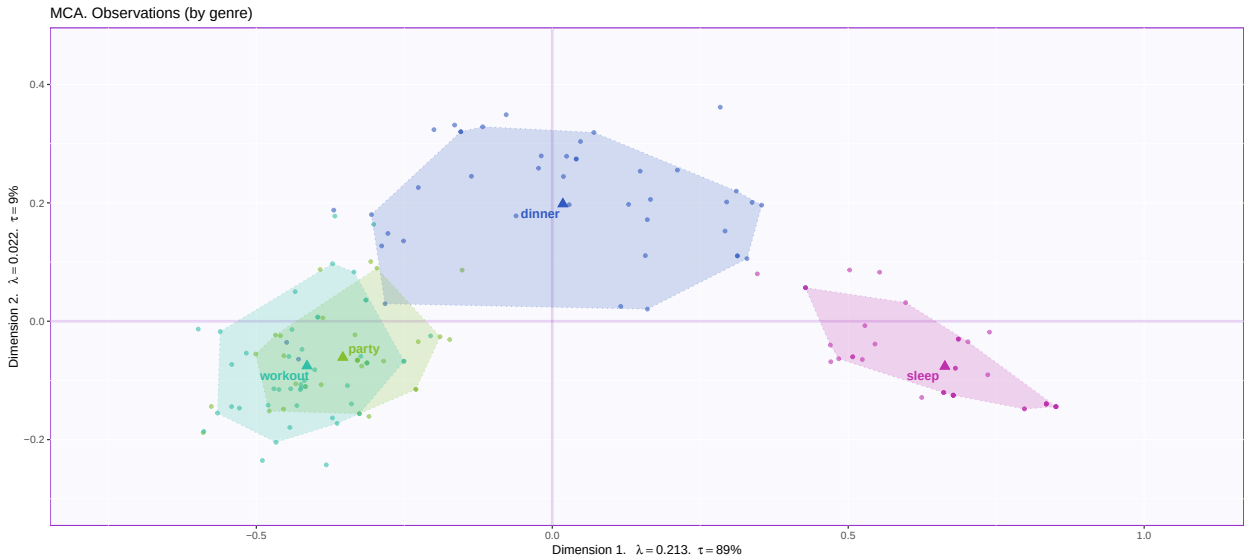
Bootstrap ratios for Dimension 1



Plot of the observations

In this plot, the observations are grouped by genre (sleep, dinner, party, workout). Confidence intervals generated by bootstrapping are marked as ellipsis around the group means. The tolerance intervals are marked around the means. Along dimension 1, we see that party and workout almost overlap entirely with each other, while being separated from dinner and sleep. Similar to PCA, dinner and sleep songs are closer to each other and farther apart from party and workout songs.


MCA. Observations (by genre)
Dimension 2. λ = 0.022. τ = 9%
Dimension 1. λ = 0.213. τ = 89%
dinner
party
workout
sleep
0.4
0.2
0.0
−0.2
−0.5
0.0
0.5
1.0



Summary

- Dimension 1: Loudness, Acousticness, Energy, Danceability, and Instrumentalness contribute the most to dimension 1. These closely correspond to Party and Workout songs.
- Dimension 2: Dinner songs are quite different from party and workout songs, most explained by Duration and Loudness.

Barycentric Discriminant Analysis

Method: Barycentric Discriminant Analysis (BADA)

Discriminant analysis is focused on grouping observations into categories. Barycentric Discriminant Analysis is employed to analyze multiple variables describing a set of observations in which each observation belongs to one and only one category. The categories are predefined.

BADA creates new combinations of the variables that best separate the groups in the data table. Using these new variables, a fresh observation measured using the same variables can placed in a category. The accuracy of placing the new observation into a category correctly is also calculated using cross validation techniques.

Source - https://bit.ly/3k37ceO

Data set: Audio features

This is a dataset which describes audio features of songs in Spotify playlists. Specifically, the music.track dataset measures 165 songs on 16 variables, of which 11 are quantitative. Some of the audio features described are acousticness, danceability, and energy.

```
##    acousticness danceability duration_ms energy instrumentalness key liveness
## 16        0.845        0.515      313560  0.519         6.96e-01   7    0.102
## 48        0.843        0.656      216453  0.217         4.30e-06   5    0.296
## 2         0.873        0.571      290293  0.346         5.19e-01   0    0.098
## 46        0.369        0.567      342067  0.500         5.50e-05  11    0.530
## 45        0.050        0.707      272000  0.508         0.00e+00   8    0.255
##    loudness mode speechiness
## 16   -9.631    1      0.0391
## 48  -13.725    1      0.0538
## 2   -12.569    0      0.0409
## 46   -9.294    1      0.0330
## 45   -9.629    0      0.0344
```

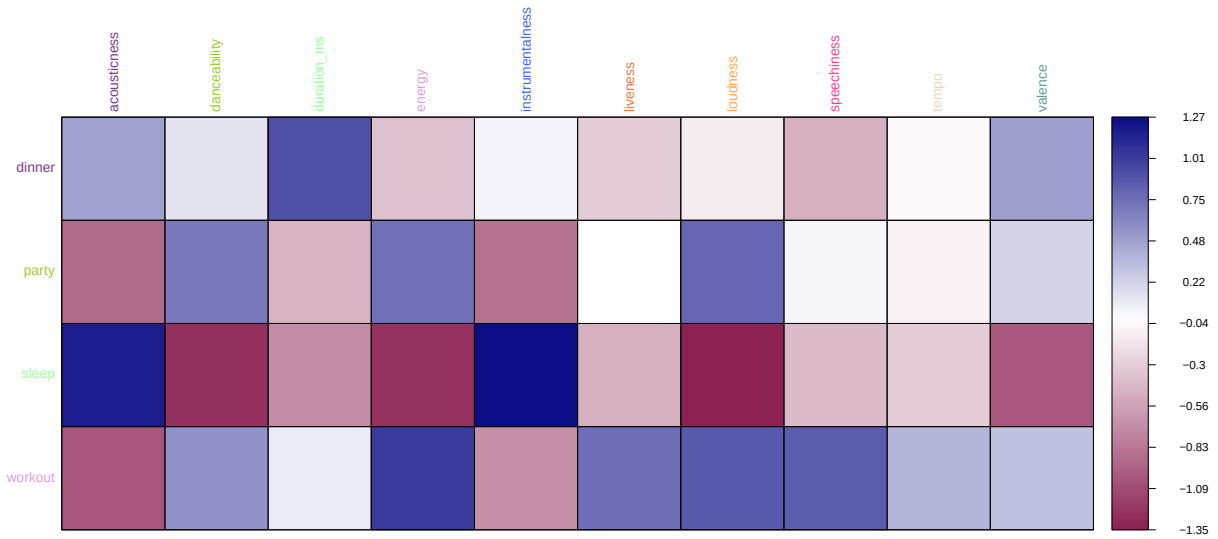
Analysis

```
# Run BADA  ----
resBADA <- tepBADA(XYmat, DESIGN = rawData$genre,
                   graphs = FALSE)
XYmat <- na.omit(XYmat)
```

```
nIter = 1000
resBADA.inf <- tepBADA.inference.battery(XYmat,
                  DESIGN = rawData$genre,
                  test.iters = nIter,
                  graphs = FALSE)
```

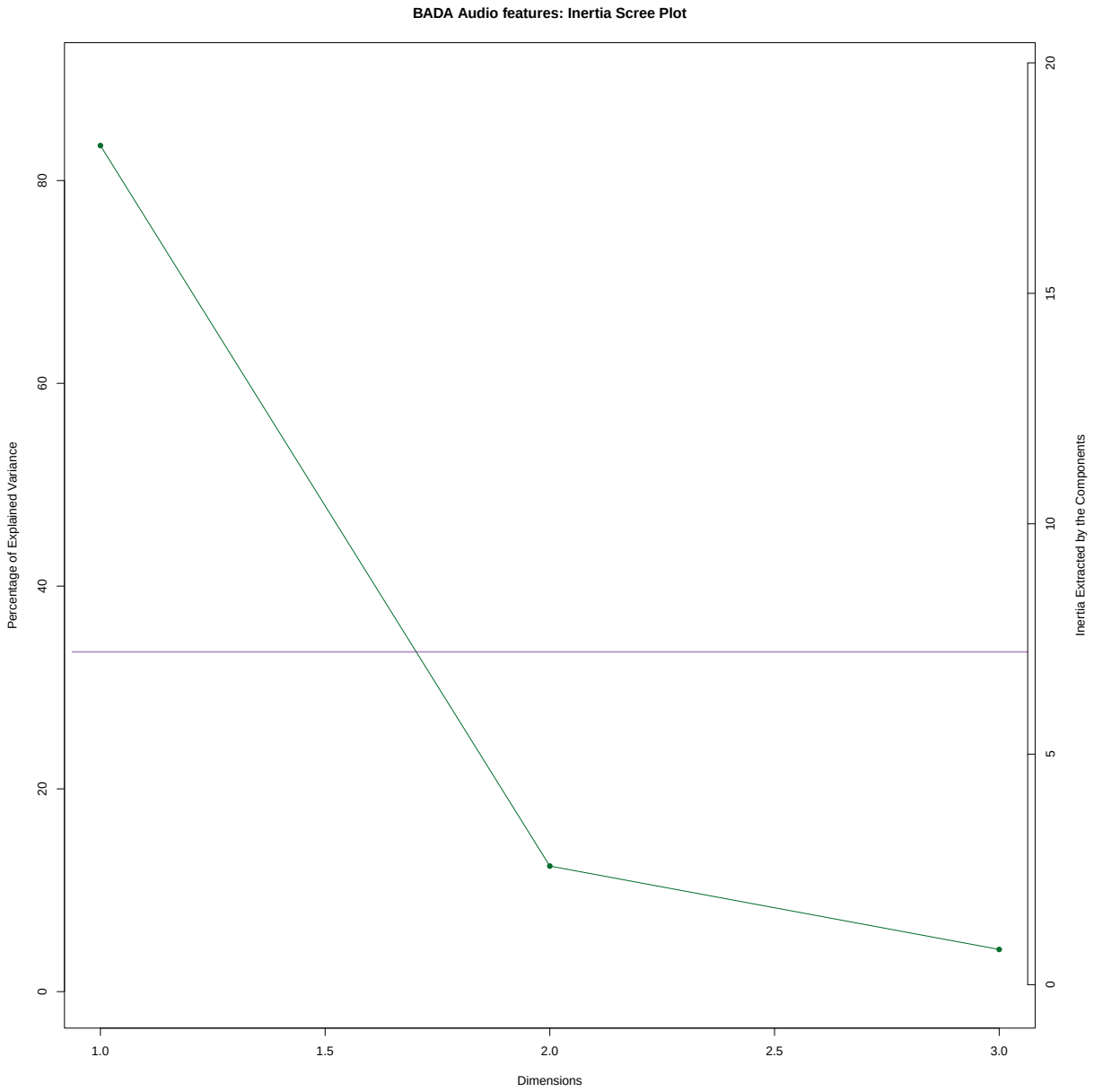
Heatmap

The heatmap is simply a visual representation of the relationship between the variables and the genres of songs. As we can see, the most distinct relationship is between sleep songs and audio features of acousticness and instrumentalness (positive). On the other hand, sleep songs and energy have a negative relationship.



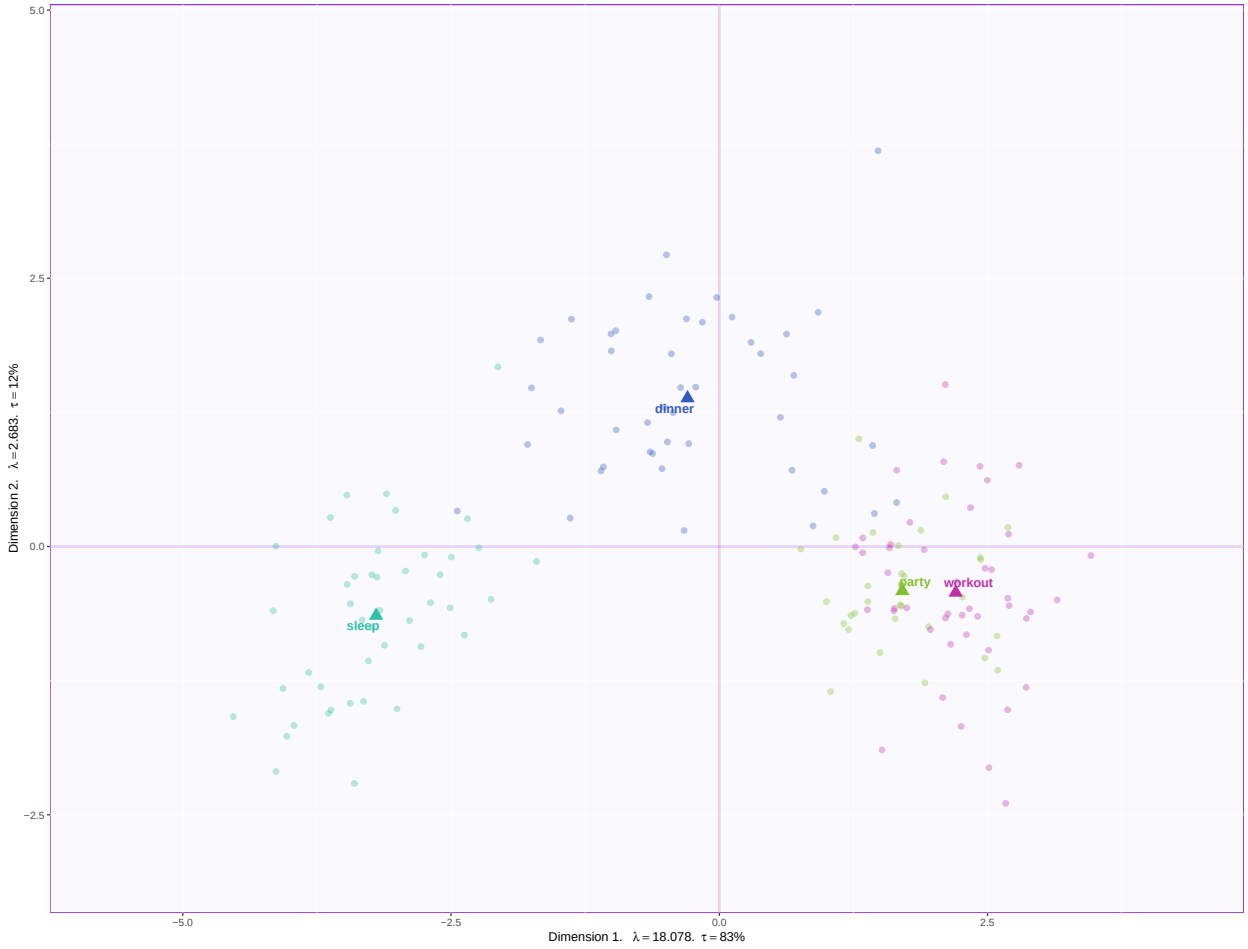
The Scree plot

The scree plot shows us how many dimensions contribute to the variance in the data. In this plot, Dim 1 contributes more than 80% or the variance. Hence, it would be a good place to start.



Map of row factors

This plot shows us the distribution of all elements in the data table. They are colored as per the genre, and we also see the group means marked. So pretty!

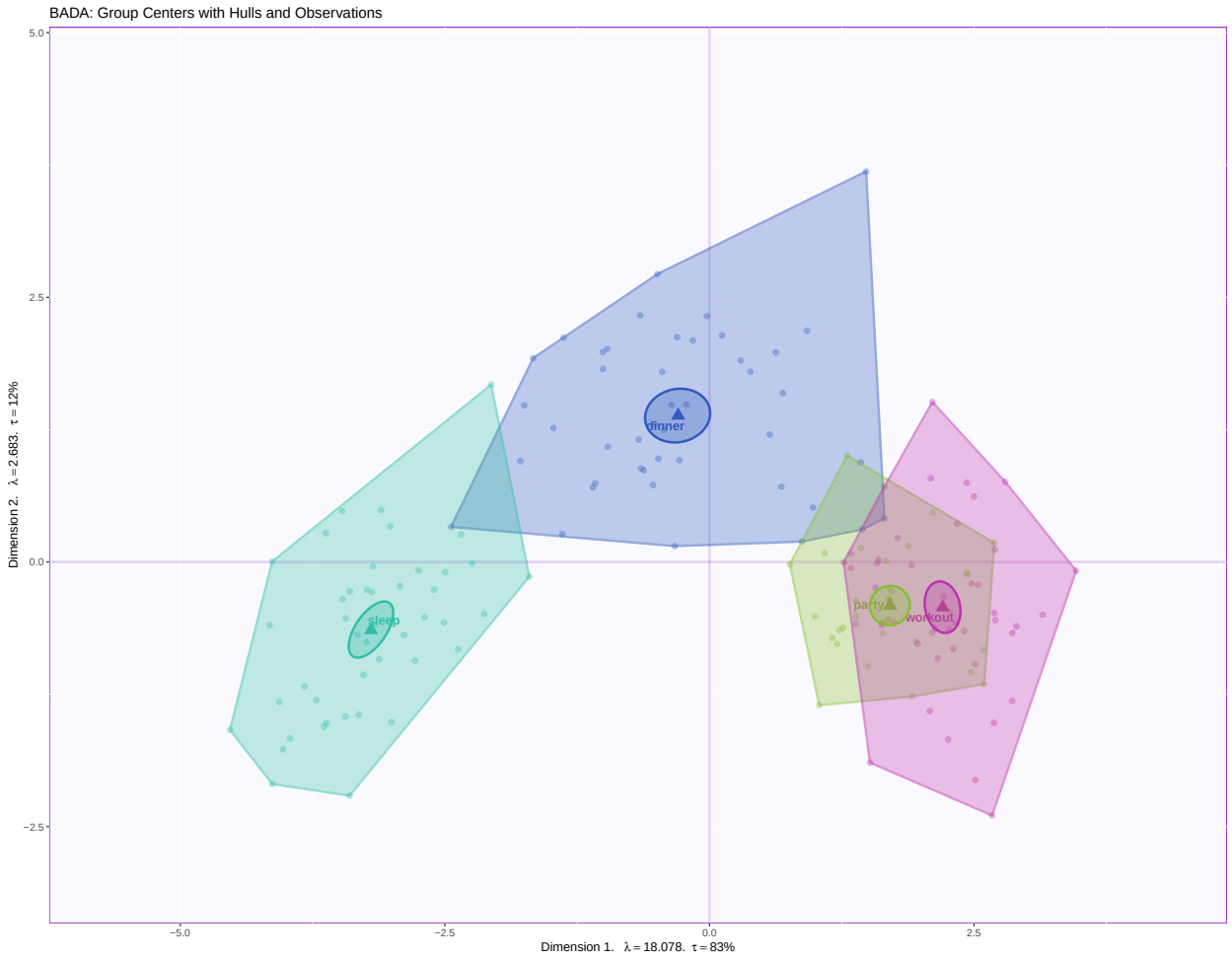


Confidence intervals generated by bootstrapping

The 95% confidence intervals around the group means are generated. They show the range of values the mean can assume 95% of the times the experiment is conducted, and the smaller the ellipsis, the higher confidence in the mean.

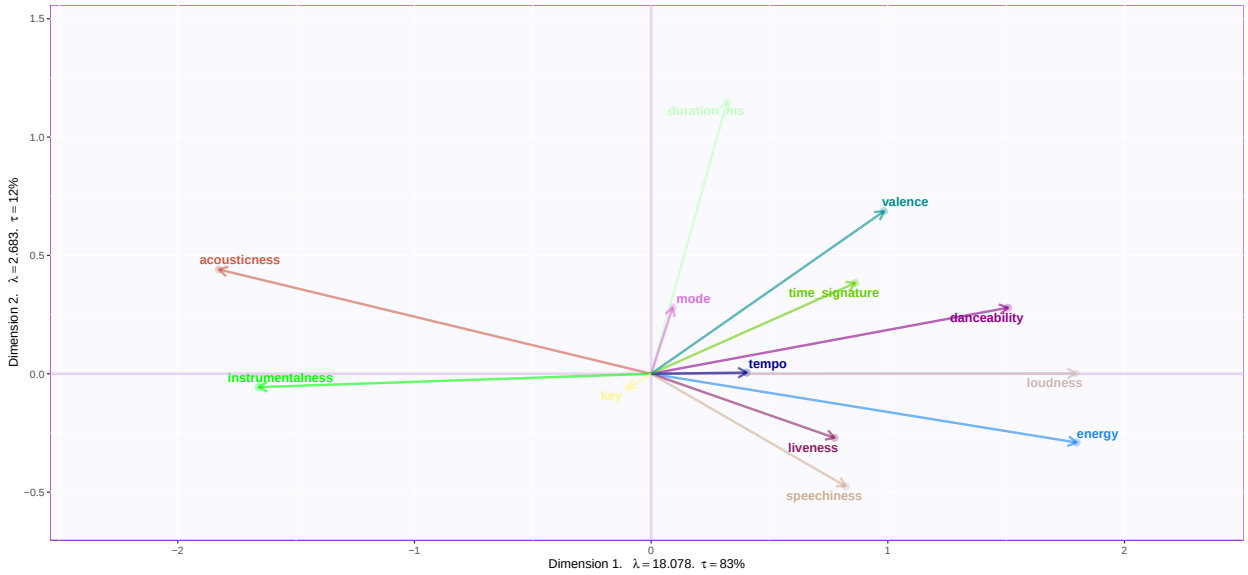
Row factors with hull

Hulls are drawn around each group.



Column factor scores

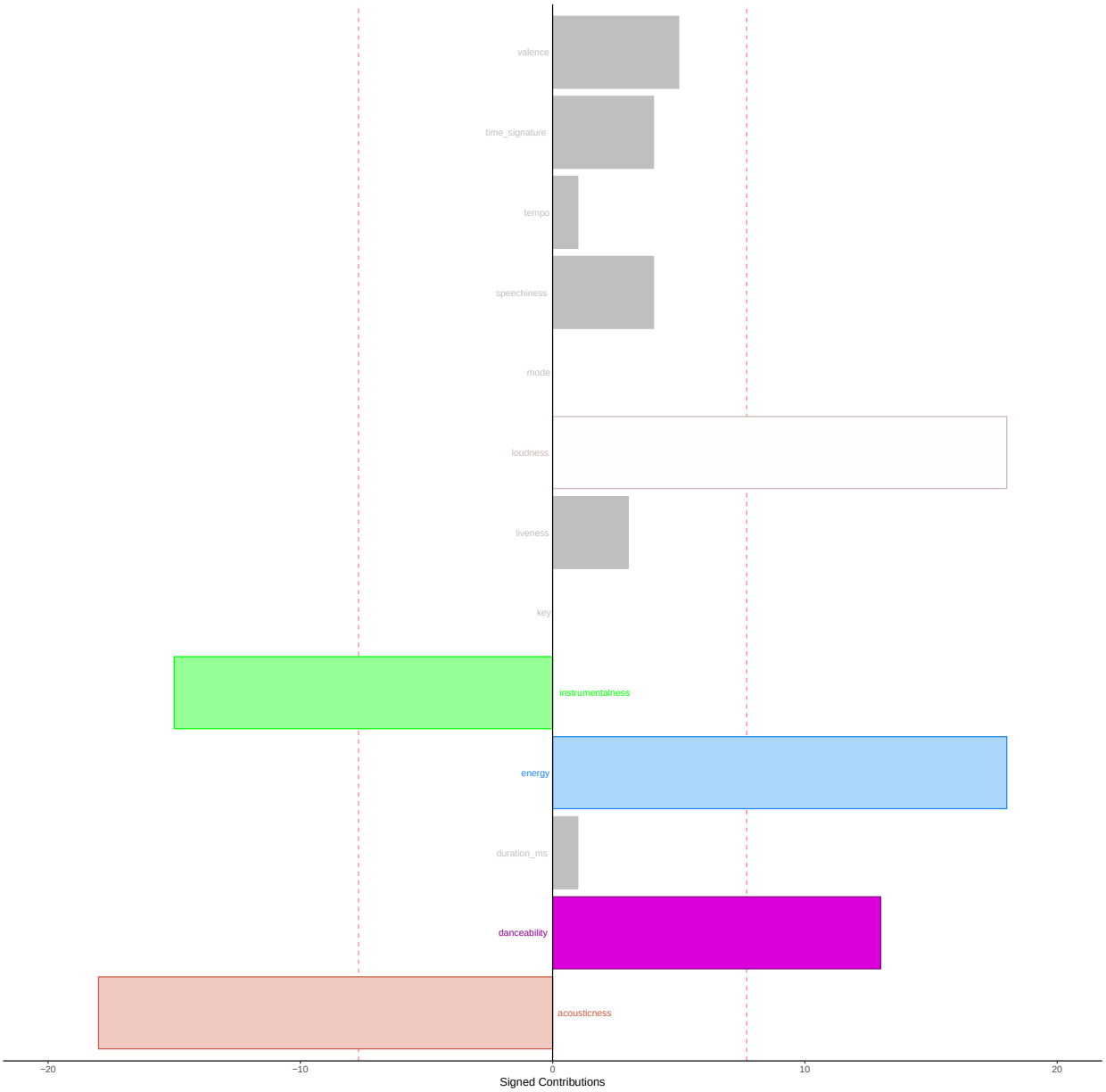
This plot shows us the correlations between the different variables - the angle between 2 arrows is an indication of the strength of relationship. The direction of relationship can be determined by looking at which quadrants the arrows lie in.



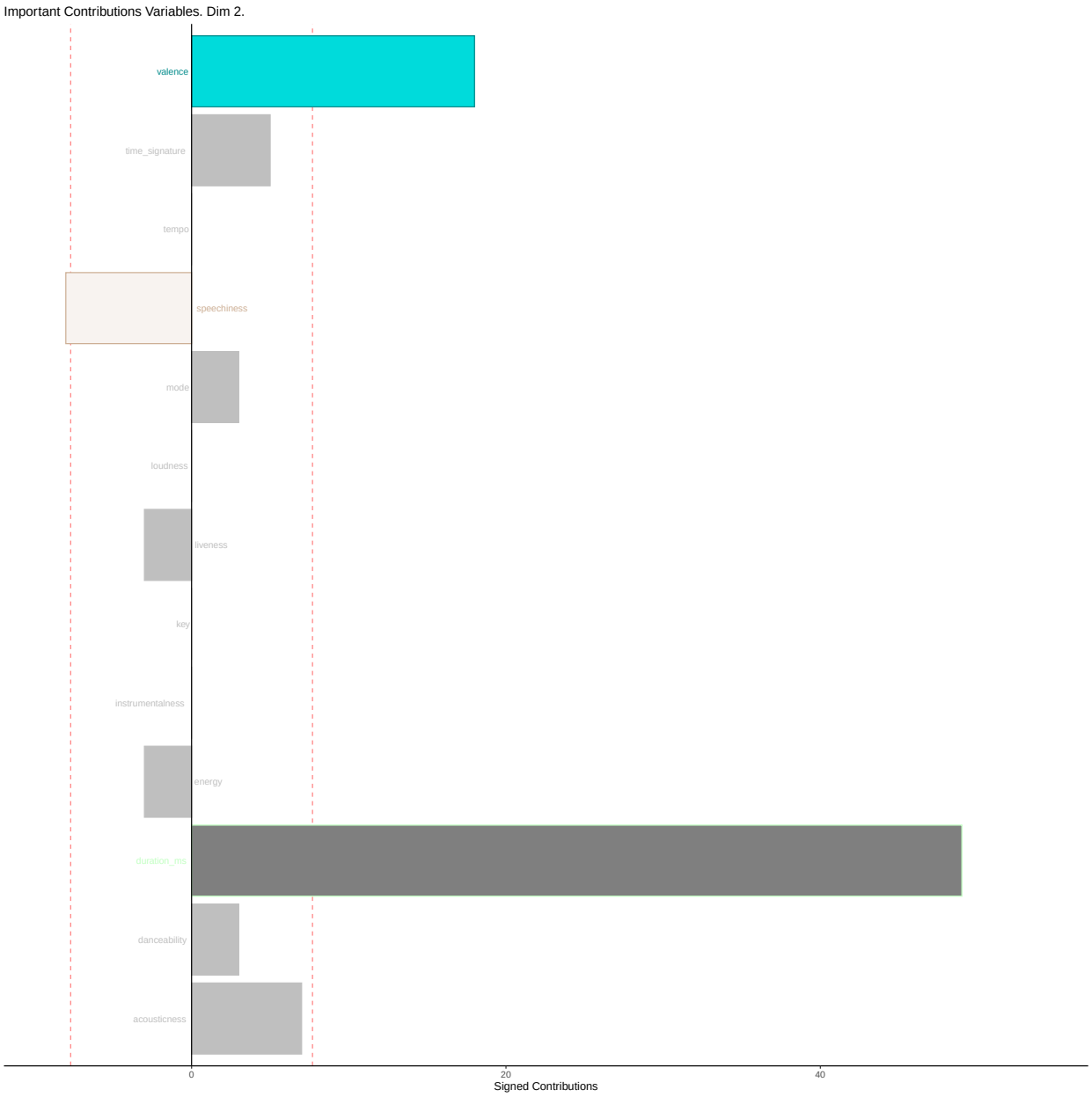
Contribution bar plots for dimension 1

The contribution bar plots show how much each variable is contributing to a certain dimension. Bootstrap ratios that follow are derived from the inference battery.

Important Contributions Variables. Dim 1.

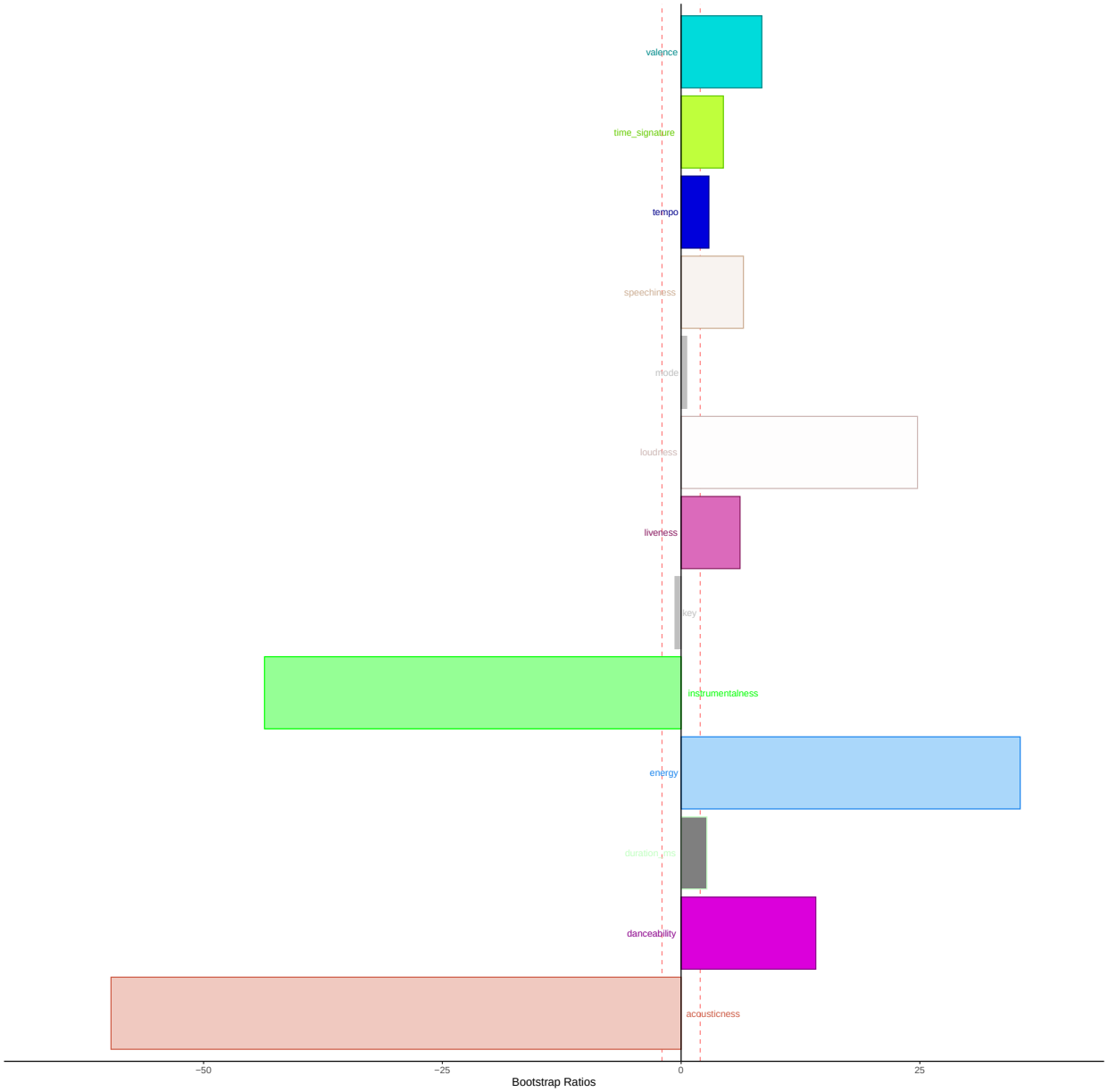


Contribution bar plots for dimension 2

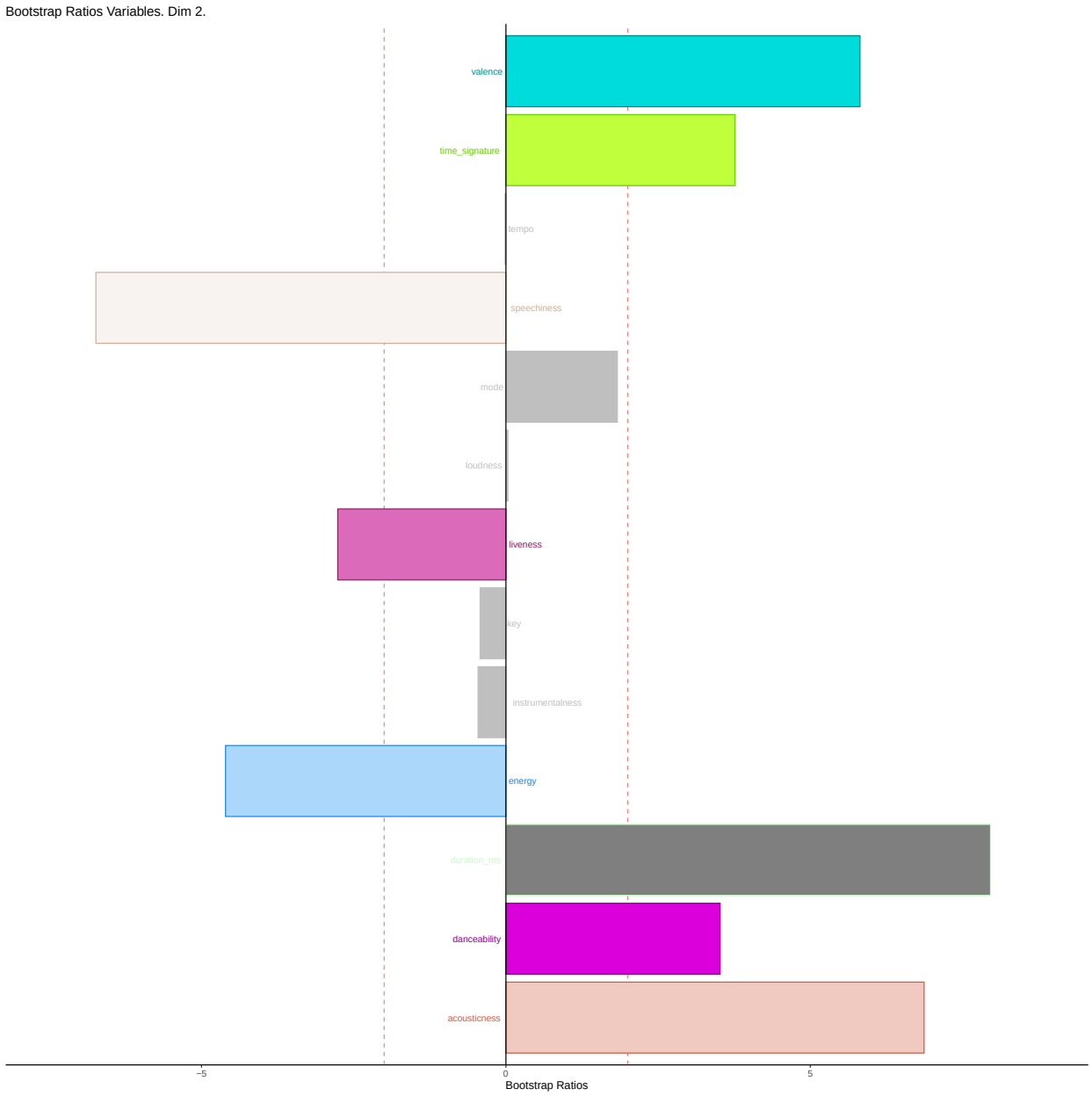


Bootstrap ratio - Dimension 1

Bootstrap Ratios Variables. Dim 1.



Bootstrap ratio - Dimension 2



Confusion matrix - fixed effects

The fixed effects confusion matrix gives us an idea of how good our predictions were to create the categories in the existing data. The accuracy value quantifies the same. On the other hand, random effects confusion matrix outlines the quality of prediction/categorizing accuracy of *new* observations in the dataset.

Jackknife is a technique used to train models for prediction by separating the dataset of interest into training and testing data. Testing data will contain only one row/subject/participant/observation, which will be predicted using the attributes of the rest of the data.

```
## [1] 0.830303
```

	.dinner	.party	.sleep	.workout
.dinner	37	0	1	0
.party	5	25	0	13
.sleep	1	0	43	0
.workout	1	7	0	32

Confusion matrix - random effects

```
## [1] 0.8121212
```

	.dinner.actual	.party.actual	.sleep.actual	.workout.actual
.dinner.predicted	37	0	1	0
.party.predicted	5	25	0	16
.sleep.predicted	1	0	43	0
.workout.predicted	1	7	0	29

Summary:

Dimension 1 Rows: Dinner and Sleep VS Party and Workout Cols: Danceability, loudness, energy are highly positively correlated, acousticness and energy strongly negatively correlated

Dimension 2 Sleep and Party/Workout songs are distinct from each other. Duration contributes to the dimension more than others. Seems to have an impact on dinner songs.

Interpretation: Dinner and sleep music that are more acoustical/instrumental are distinct from other genres

Discriminant Correspondence Analysis

Method: Discriminant Correspondence Analysis (DiCA)

BADA for qualitative data! Just like PCA and CA.

Source - https://bit.ly/300ioBM

Data set: Audio features

This is a dataset which describes audio features of songs in Spotify playlists. Specifically, the music.track dataset measures 165 songs on 16 variables, of which 11 are quantitative. Some of the audio features described are acousticness, danceability, and energy.

The data is binned in the same way as MCA.

```
##    acousticness danceability duration_ms energy instrumentalness key liveness
## 16        0.845        0.515      313560  0.519         6.96e-01   7    0.102
## 48        0.843        0.656      216453  0.217         4.30e-06   5    0.296
## 2         0.873        0.571      290293  0.346         5.19e-01   0    0.098
## 46        0.369        0.567      342067  0.500         5.50e-05  11    0.530
## 45        0.050        0.707      272000  0.508         0.00e+00   8    0.255
##    loudness
## 16   -9.631
## 48  -13.725
## 2   -12.569
## 46   -9.294
## 45   -9.629
```

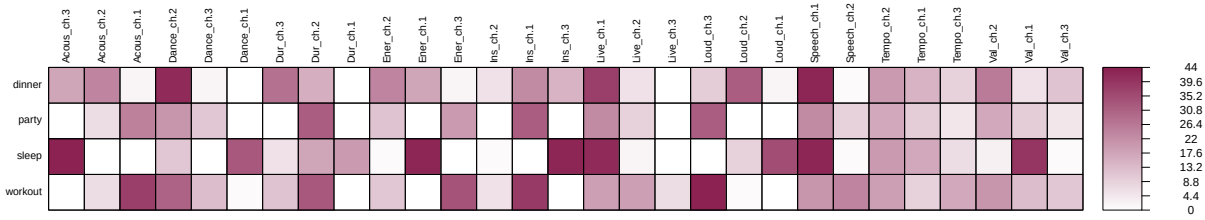
Analysis

```
## Run DiCA  ----
resDiCA <- tepDICA(XYmat,
                   make_data_nominal = TRUE,
                   DESIGN = hwdata$genre,
                   graphs = FALSE)
##  Inferences ----
set.seed(70301) # set the seed
# to ensure same results.
```

```
nIter <- 100
resDiCA.inf <- tepDICA.inference.battery(as.matrix(XYmat),
                                         make_data_nominal = TRUE,
                 DESIGN = hwdata$genre,
                 test.iters = nIter,
                 graphs = FALSE)
```

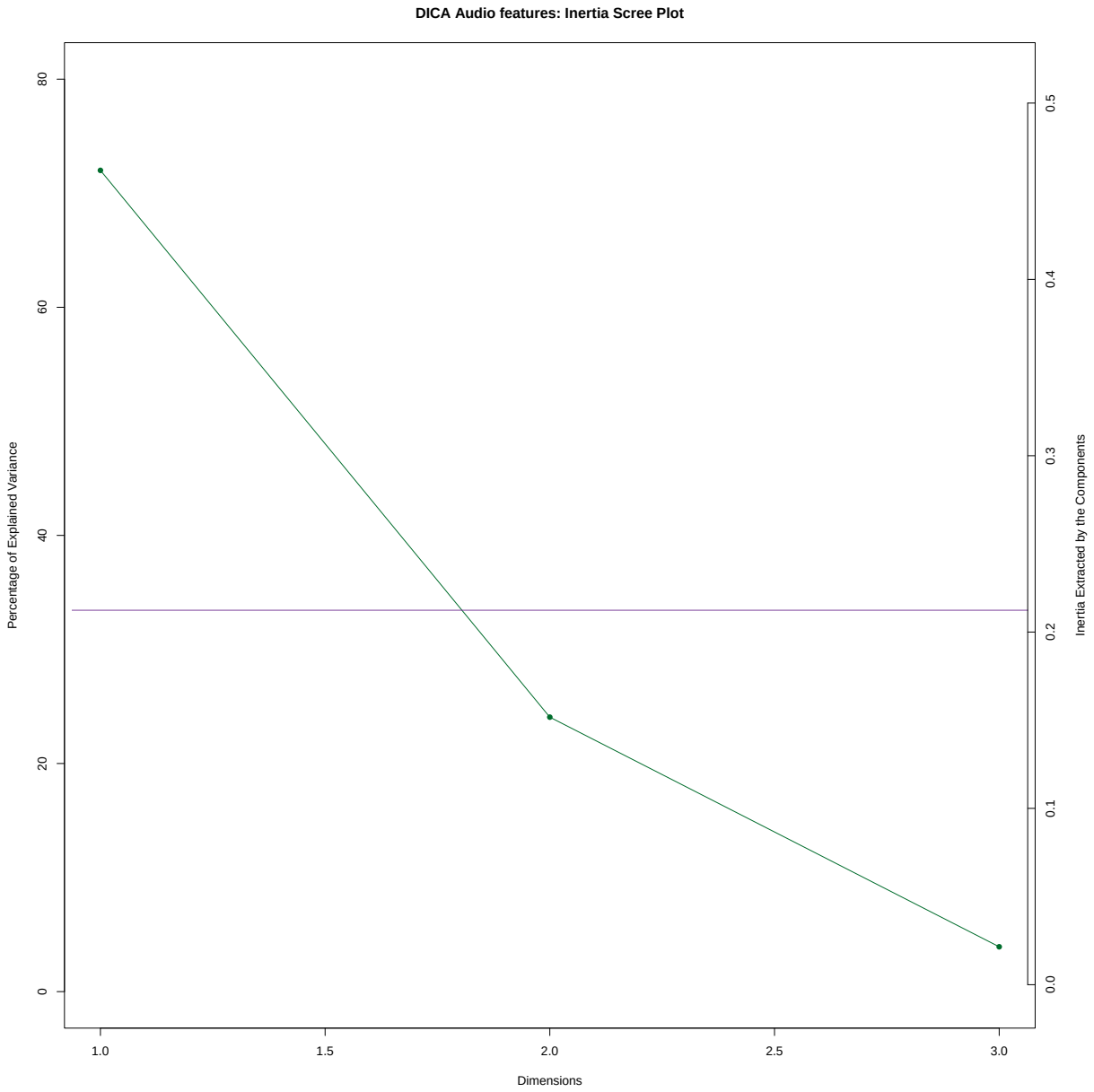
Heatmap

The heatmap is simply a visual representation of the relationship between the variables and the genres of songs. As we can see, the most distinct relationship is between sleep songs and audio features of acousticness and instrumentalness (positive). On the other hand, sleep songs and energy have a negative relationship.



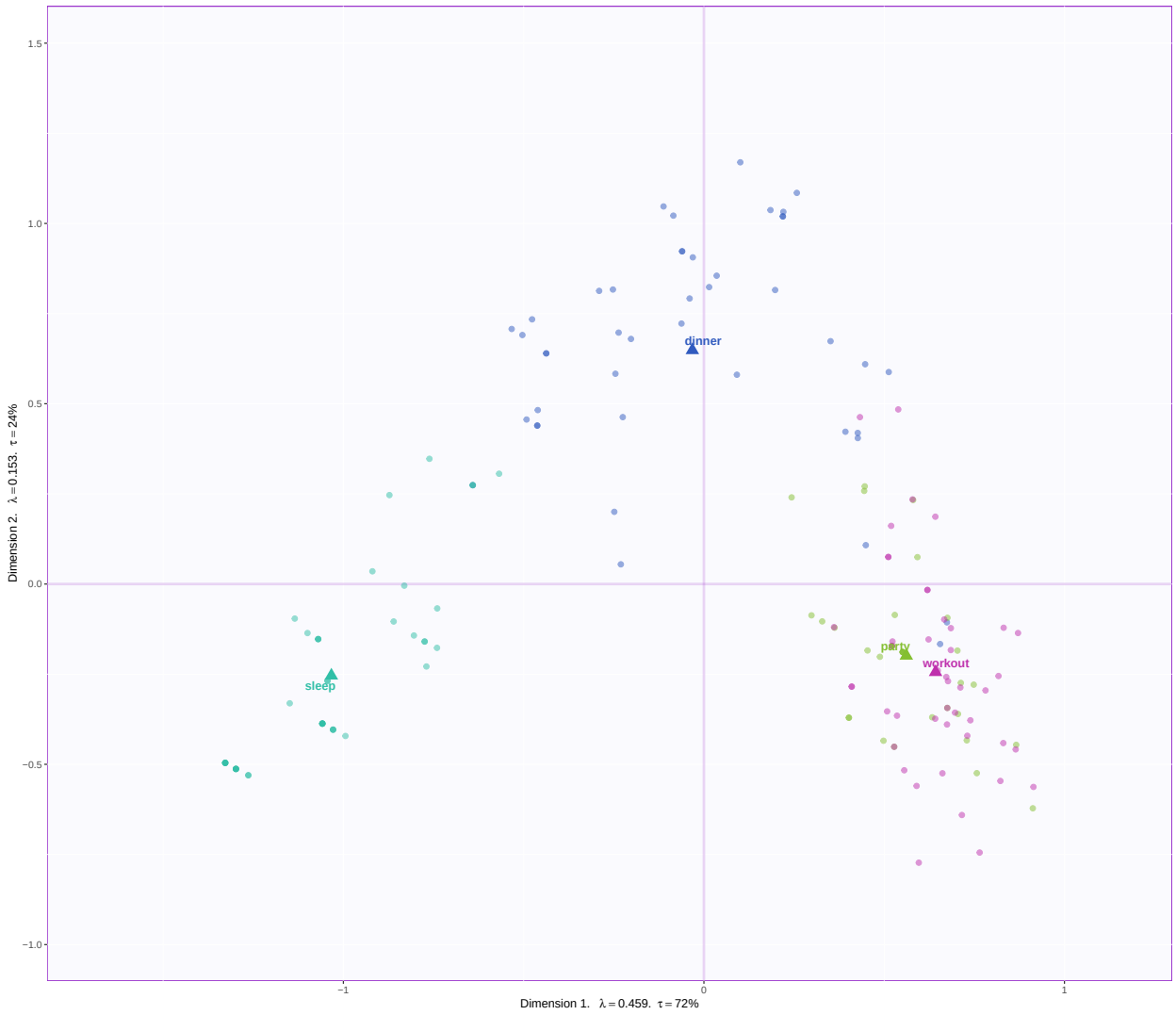
The Scree plot

The scree plot shows us how many dimensions contribute to the variance in the data. In this plot, Dim 1 contributes more than 80% or the variance. Hence, it would be a good place to start.



Map of row factors

This plot shows us the distribution of all elements in the data table. They are colored as per the genre, and we also see the group means marked. So pretty!

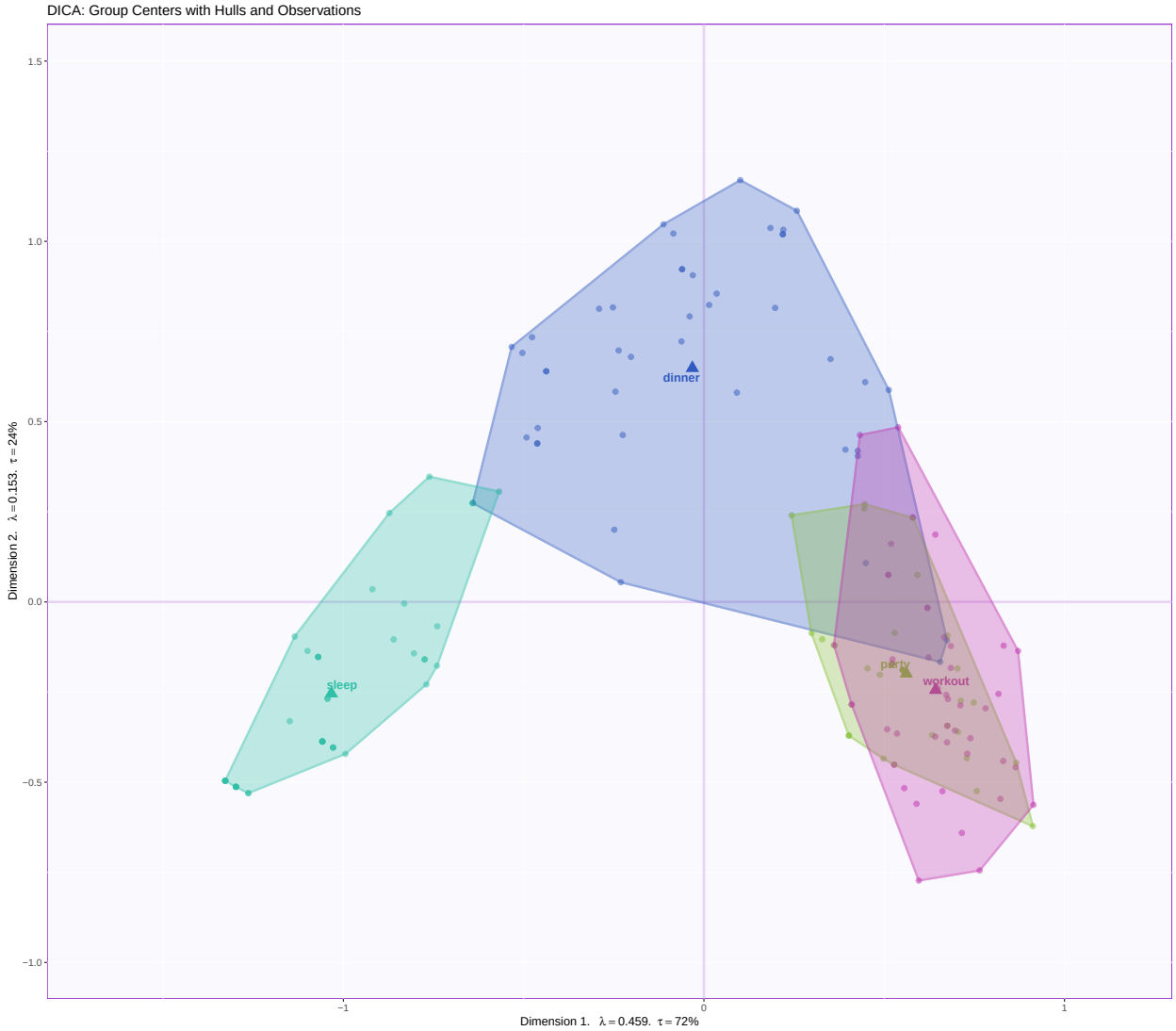


Confidence intervals generated by bootstrapping

The 95% confidence intervals around the group means are generated. The smaller their radii, the more confidence we can assume of our group mean estimate.

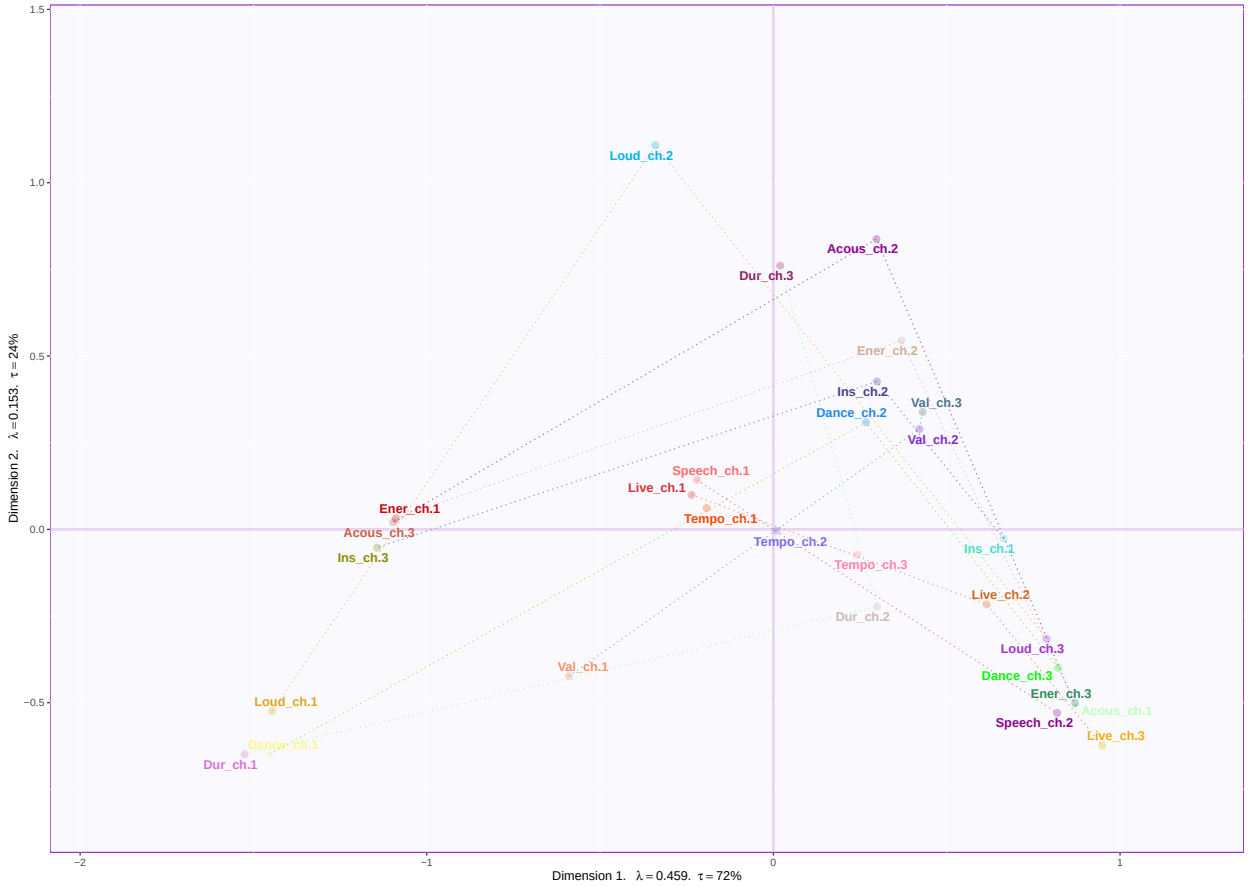
Row factors with hull

Hulls are drawn around each group.



Column factor scores

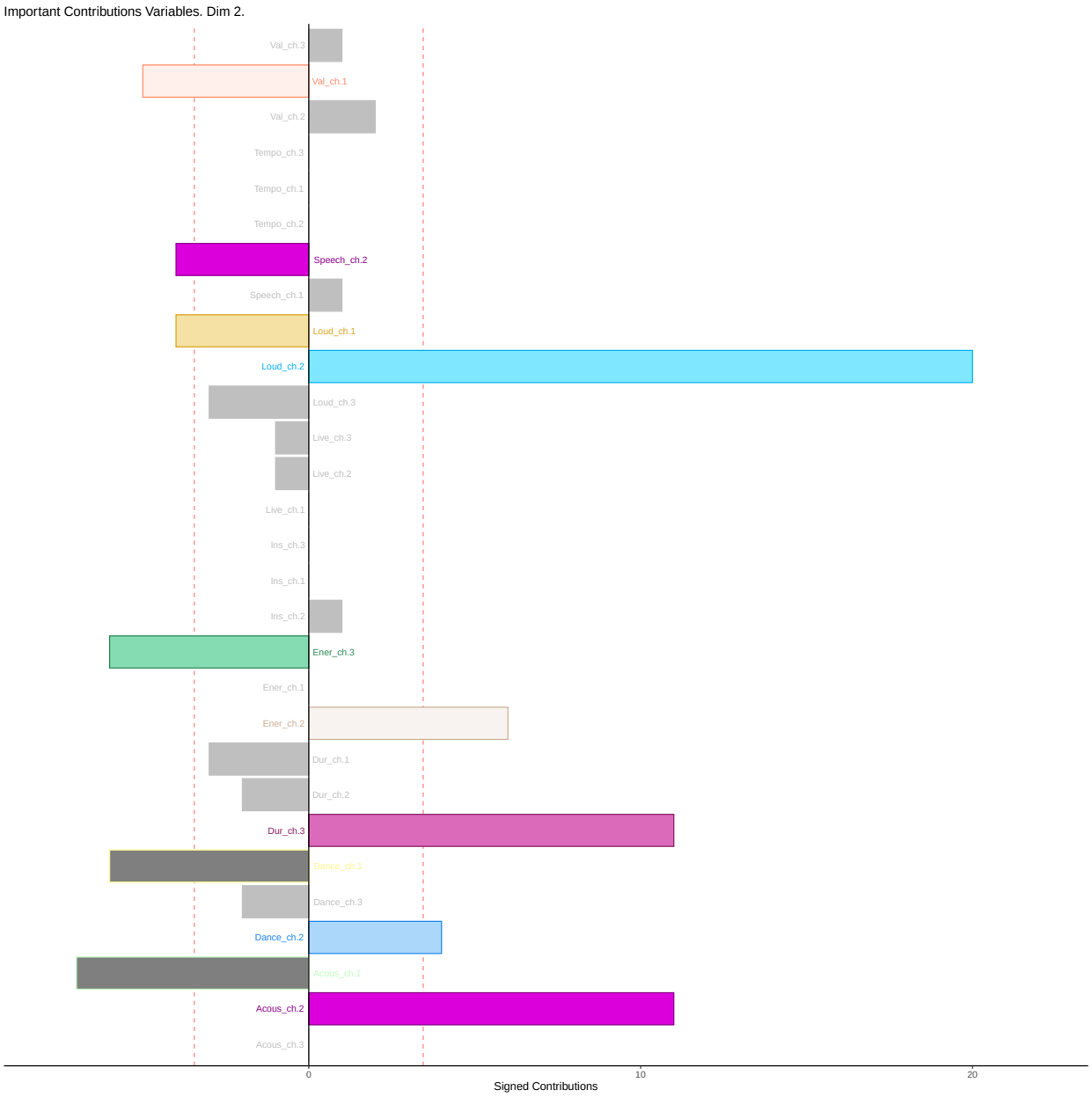
This plot shows us the correlations between the different variables - the angle between 2 arrows is an indication of the strength of relationship. The direction of relationship can be determined by looking at which quadrants the arrows lie in.



Contribution bar plots for dimension 1

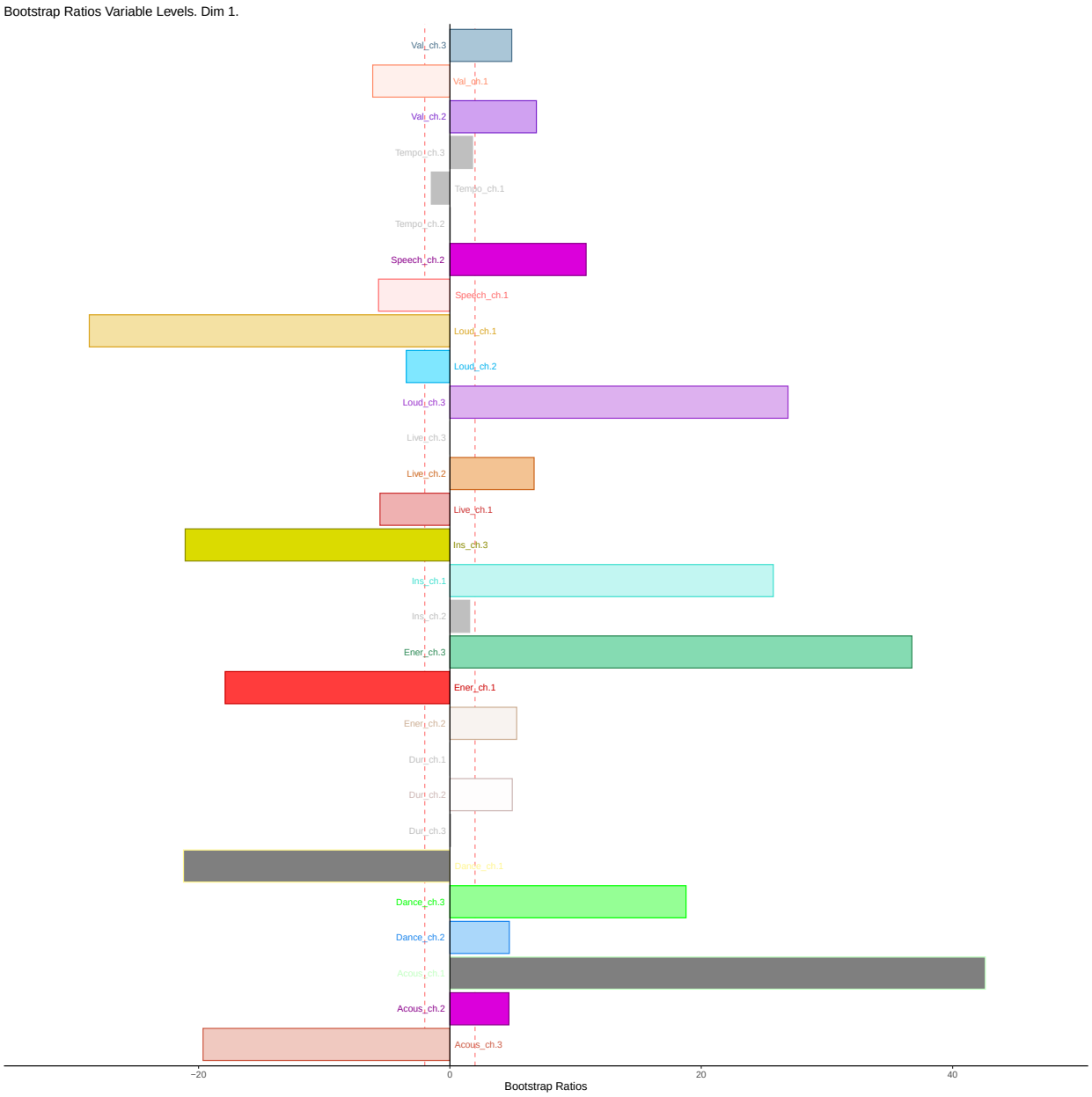
The contribution bar plots show how much each variable is contributing to a certain dimension. Bootstrap ratios that follow are derived from the inference battery.

Contribution bar plots for dimension 2

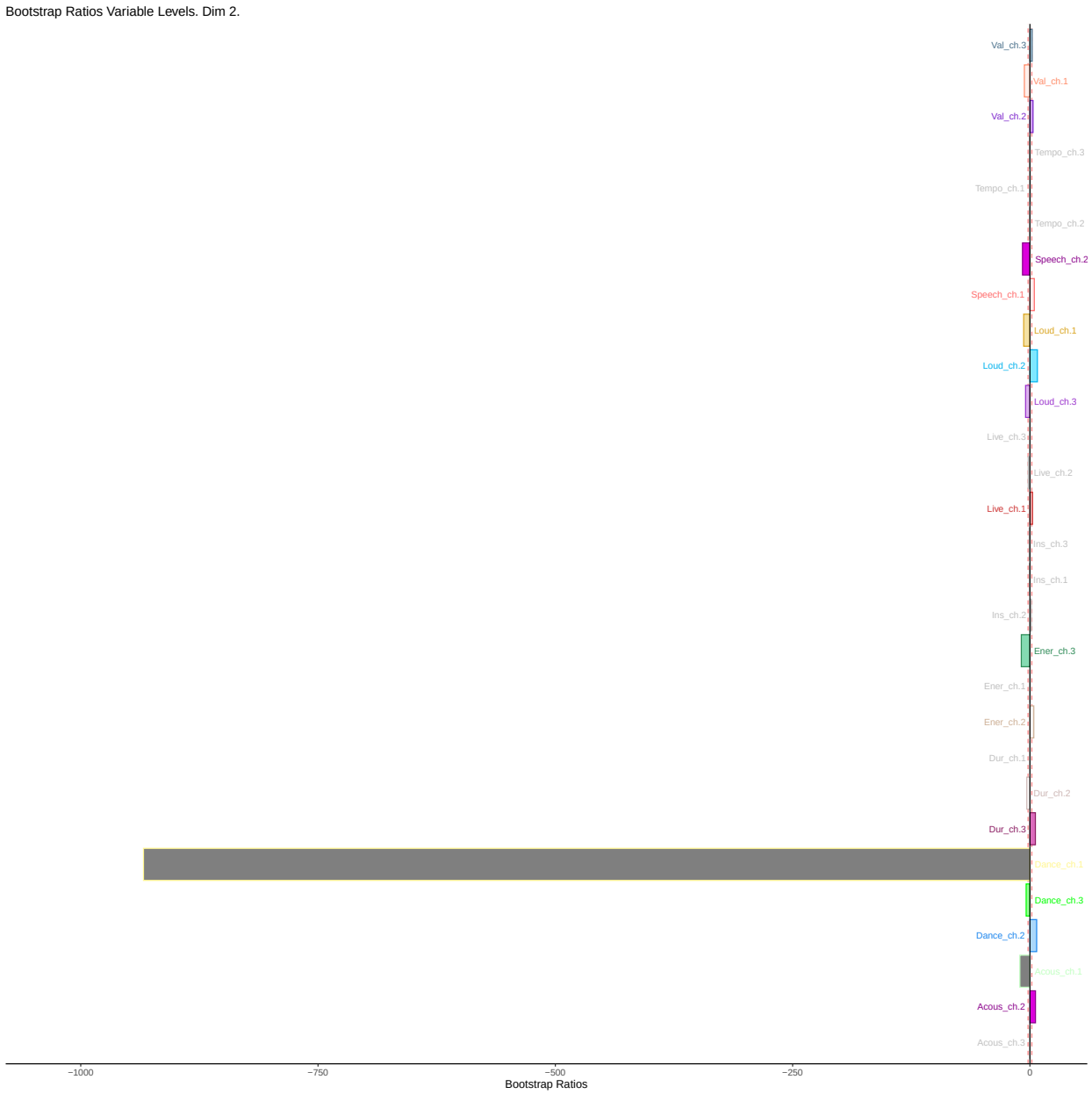


Contribution maps

Bootstrap ratio - Dimension 1



Bootstrap ratio - Dimension 2



Confusion matrix - Fixed effects

The fixed effects confusion matrix gives us an idea of how good our predictions were to create the categories in the existing data. The accuracy value quantifies the same. On the other hand, random effects confusion matrix outlines the quality of prediction/categorizing accuracy of *new* observations in the dataset.

Jackknife is a technique used to train models for prediction by separating the dataset of interest into training and testing data. Testing data will contain only one row/subject/participant/observation, which will be predicted using the attributes of the rest of the data.

```
## [1] 0.8363636
```

	.dinner	.party	.sleep	.workout
.dinner	39	0	1	2
.party	4	29	0	16
.sleep	1	0	43	0
.workout	0	3	0	27

Confusion matrix - random effects

```
## [1] 0.8
```

	.dinner.actual	.party.actual	.sleep.actual	.workout.actual
.dinner.predicted	38	0	1	2
.party.predicted	5	24	0	16
.sleep.predicted	1	0	43	0
.workout.predicted	0	8	0	27

Summary:

Dimension 1 Rows: Dinner and Sleep VS Party and Workout Cols: Danceability, loudness, energy are highly positively correlated, acousticness and energy strongly negatively correlated

Dimension 2 Sleep and Party/Workout songs are distinct from each other

Interpretation: Dinner and sleep music that are more acoustical/instrumental are distinct from other genres.

Partial Least Squares Correlation

Method: Partial Least Squares Correlation (PLSC)

Partial Least Squares Correlation is a technique used to analyze two data tables with different variables measuring the same observations. The data is condensed into latent variables which are linear combinations of original variables. The idea is to find how much information is shared between the two tables. The same technique can be applied when one data table is used to predit the other one (PLS Regression).

PLSC can handle large data sets because it essentially operates on reducing dimensionality in the data. The latent variables are supposed to explain the maximum covariance shared between the two tables. In other words (or PCA contemporary), latent variables are factor scores. On the other hand, the contributions of the variables to the shared covariance ("loadings") are called saliences.

Source - https://bit.ly/3H88Wxc

Data set: Audio features

This is a dataset which describes audio features of songs in Spotify playlists.

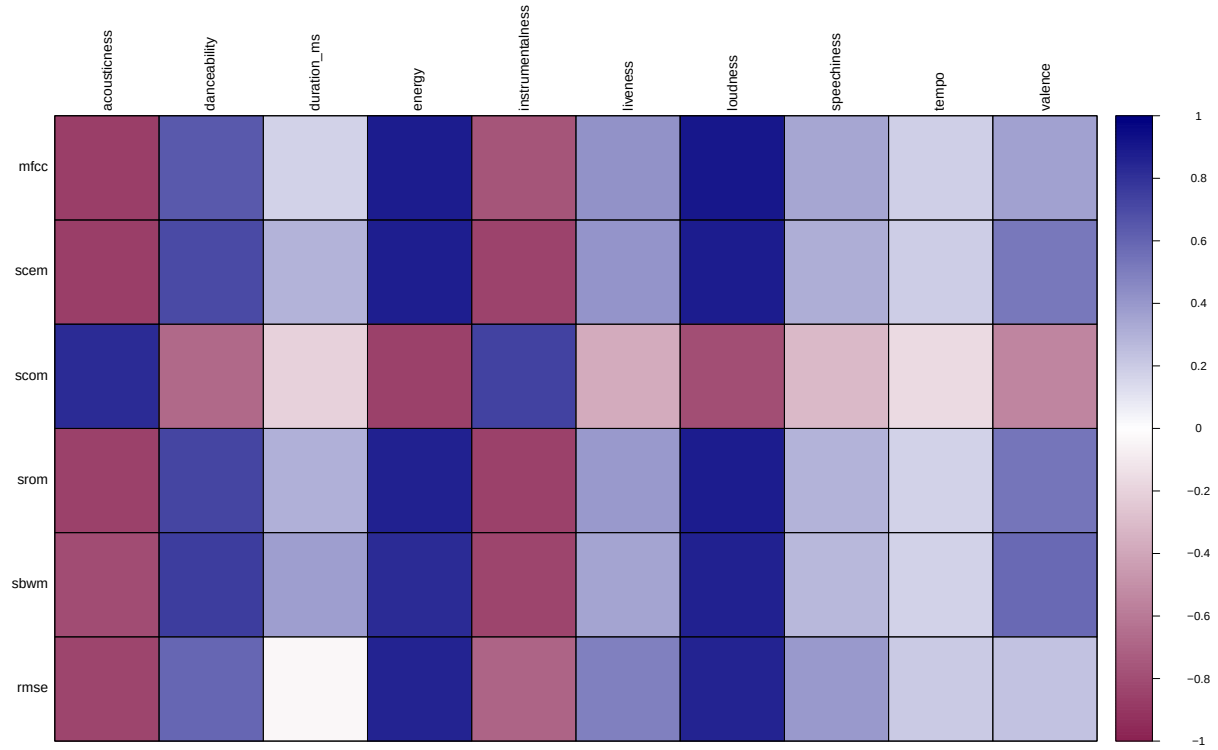
Specifically, the music.track dataset measures 165 songs on 16 variables, of which 11 are quantitative. Some of the audio features described are acousticness, danceability, and energy. Additionally, music.audio contains 7 quantitative variables that describe features of the audio signals; mel-frequency cepstral coefficients (MFCC) –> timbre, spectral centroid (SCEM) –> brightness of sound, spectral contrast (SCOM) –> harmonic/non-harmonic music, spectral roll-off (SROM) & bandwidth (SBWM) –> timbre, root mean square energy (RMSE) –> energy per frame

```
##         mfcc     scem     scom     srom     sbwm
## 1  2.1820905 2008.227 22.16629 4271.333 2362.907
## 2 -0.4209786 1810.502 22.81968 3800.144 2217.649
## 3  0.4319727 1920.575 20.06893 4231.883 2553.859
## 4  5.9844995 1777.566 21.45459 3898.110 2247.093
## 5  1.4752268 2629.842 21.64794 5873.931 2734.930
```

```
##    acousticness danceability duration_ms energy instrumentalness
## 16        0.845        0.515      313560  0.519         6.96e-01
## 48        0.843        0.656      216453  0.217         4.30e-06
## 2         0.873        0.571      290293  0.346         5.19e-01
## 46        0.369        0.567      342067  0.500         5.50e-05
## 45        0.050        0.707      272000  0.508         0.00e+00
```

Correlation plot

The correlation plot shows the magnitude and direction of the relationship between variables. As we can see, some are strongly negatively correlated (acousticness and MFCC, SCEM), and few others are strongly positively correlated (energy/loudness and SROM, SBWM, RMSE).

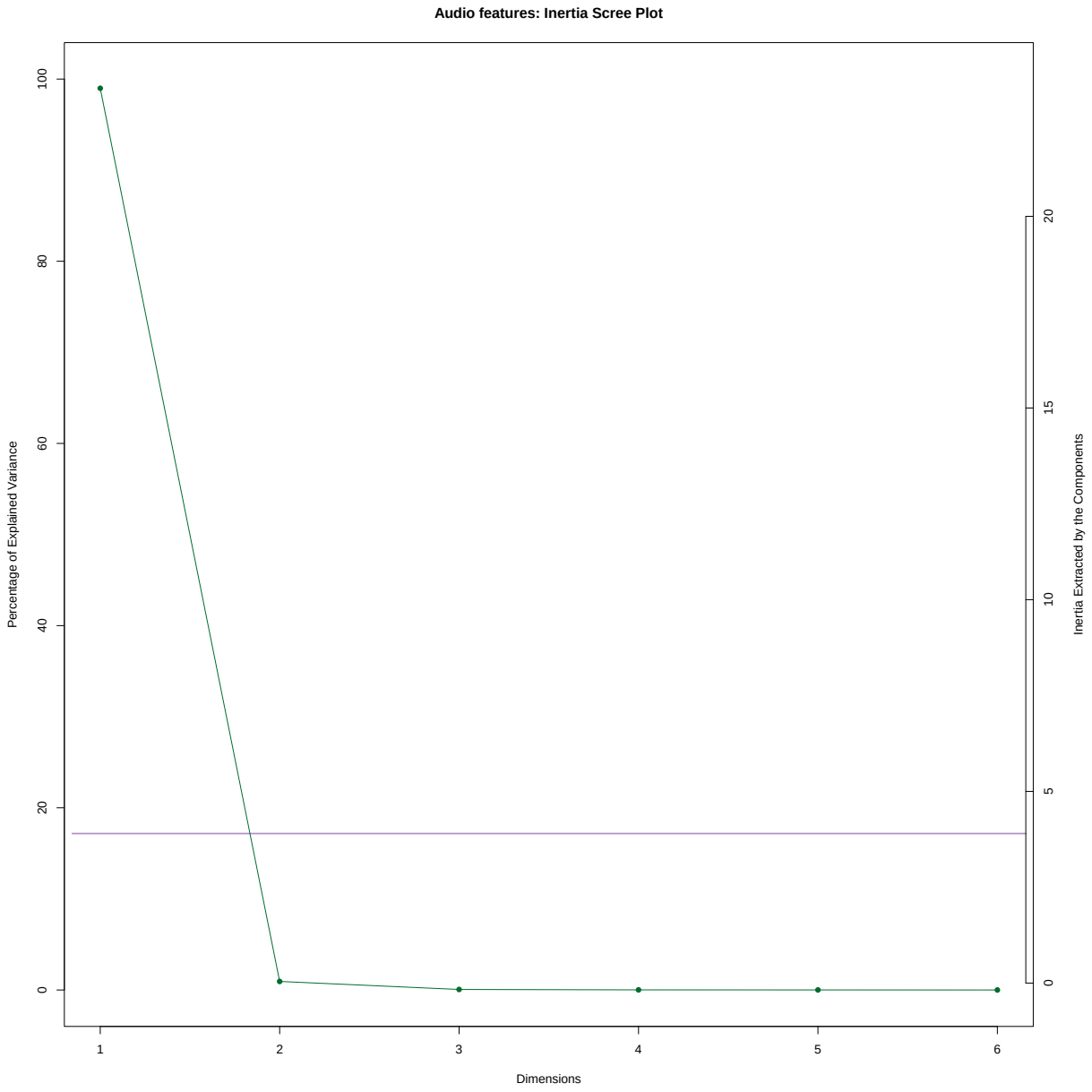


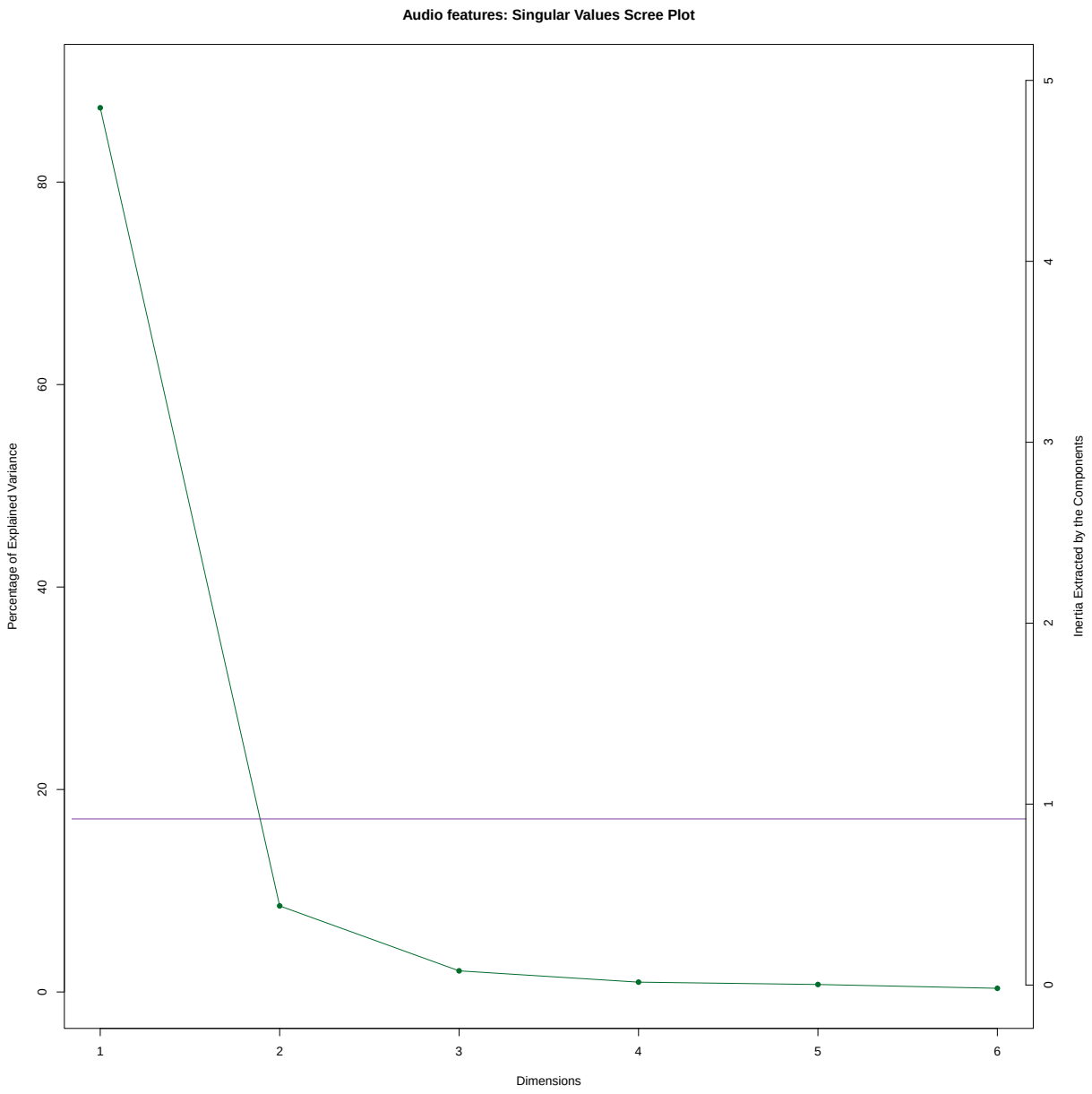
Analysis

```
# Run PLSC  ----
resPLSC <- tepPLS(Xmat,
                  Ymat,
                  DESIGN = rawData$genre,
                  graphs = FALSE)
```

The Scree plot

The scree plot shows us how many dimensions contribute to the variance in the data. In this plot, Dim 1 contributes more than 80% or the variance. Hence, it would be a good place to start.


Audio features: Inertia Scree Plot
Percentage of Explained Variance
Inertia Extracted by the Components
Dimensions
0
20
40
60
80
100
0
5
10
15
20
1
2
3
4
5
6



Latent variable - 1

In PLSC, linear combination of variables are projected as latent variables. When there are 2 data tables and both are considered the dependent variables, the goal of the analysis is to explore how much variance they share. Taking this into consideration, the "dimension" containing the first set of latent variables ideally holding most of the covariance/shared variance between the two data tables.

Below is the first set of latent variables, contribution barplots and bootstrap plots.

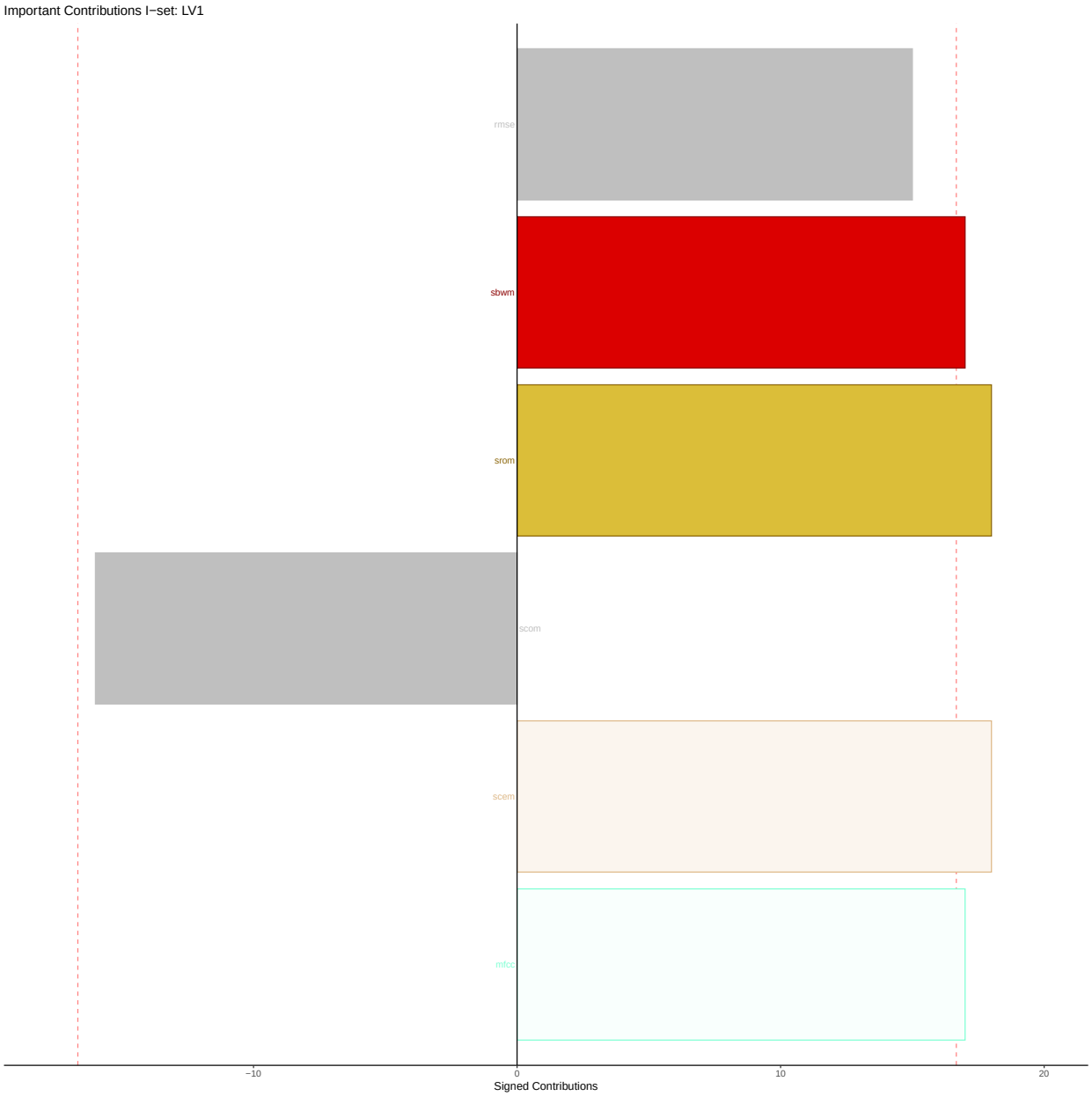
Also, just like PCA and all other multivariate techniques that I have dealt with so far, dinner and sleep music are tied together and workout and party are almost overlapping here.

Confidence intervals generated by bootstrapping

The 95% confidence intervals constructed around the means indicate the likelihood of the population mean falling within the range 95% of the times the experiment is conducted.

Contributions - music.audio

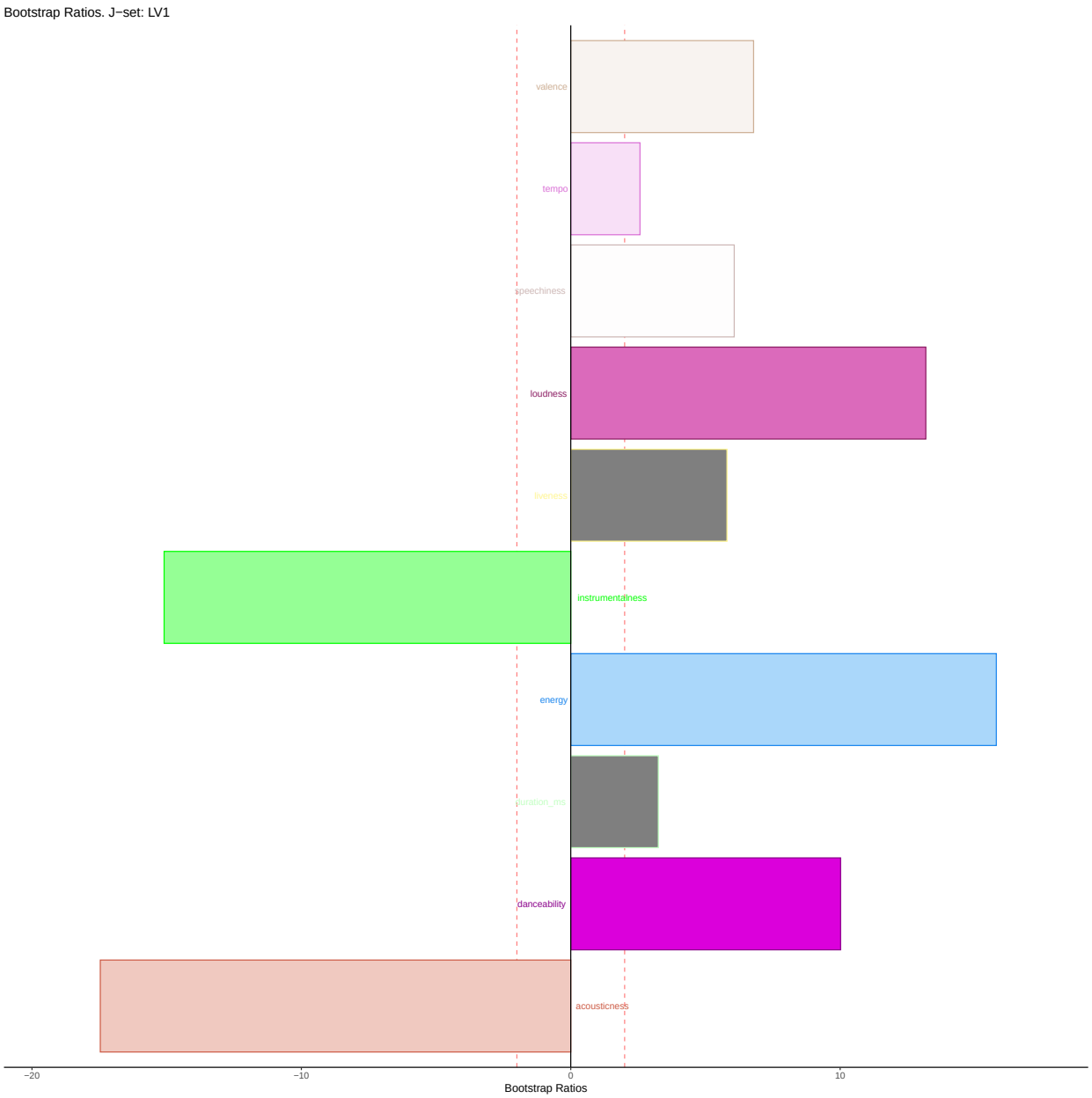
All 7 variables contribute significantly to the variance.



Contributions - music.track

In the second data table, acousticness, loudness, energy, danceability, and instrumentalness contribute to most of the shared variance.


Bootstrap Ratios. I−set: LV1
rmse
sbwm
srom
scom
scem
mfcc
−10
0
10
Bootstrap Ratios



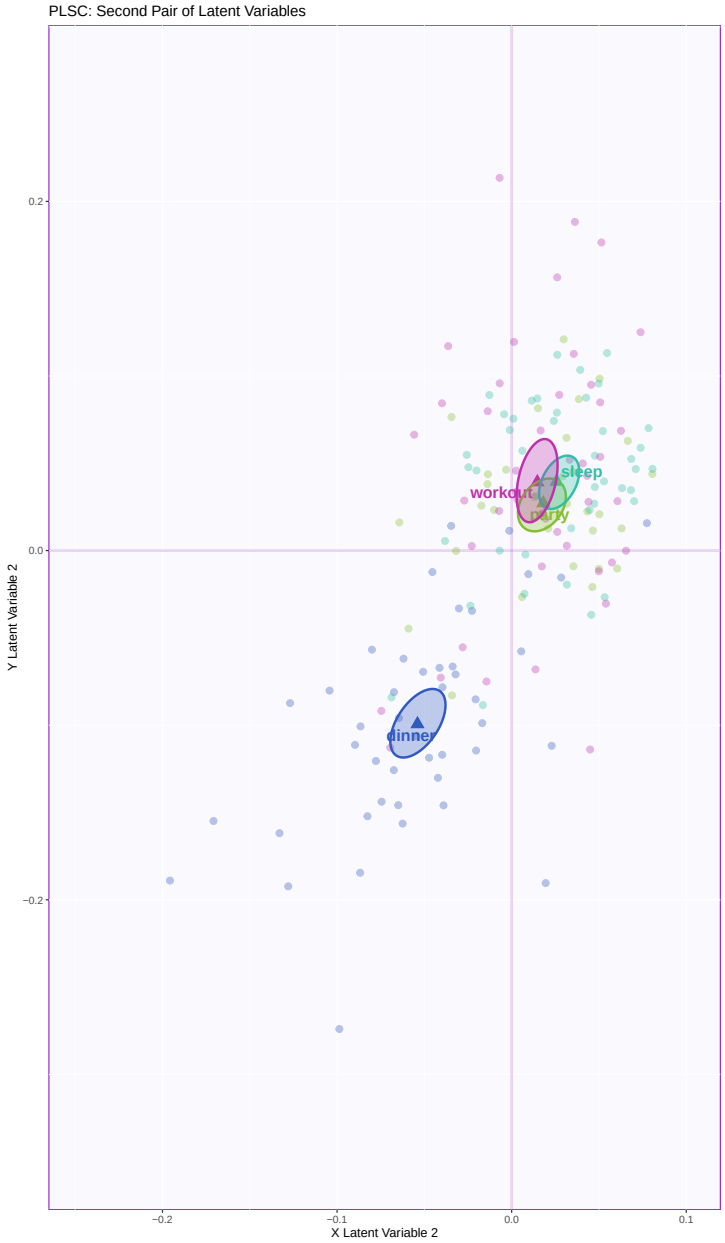
Bootstrap ratio for dimension 2

Latent variable - 2

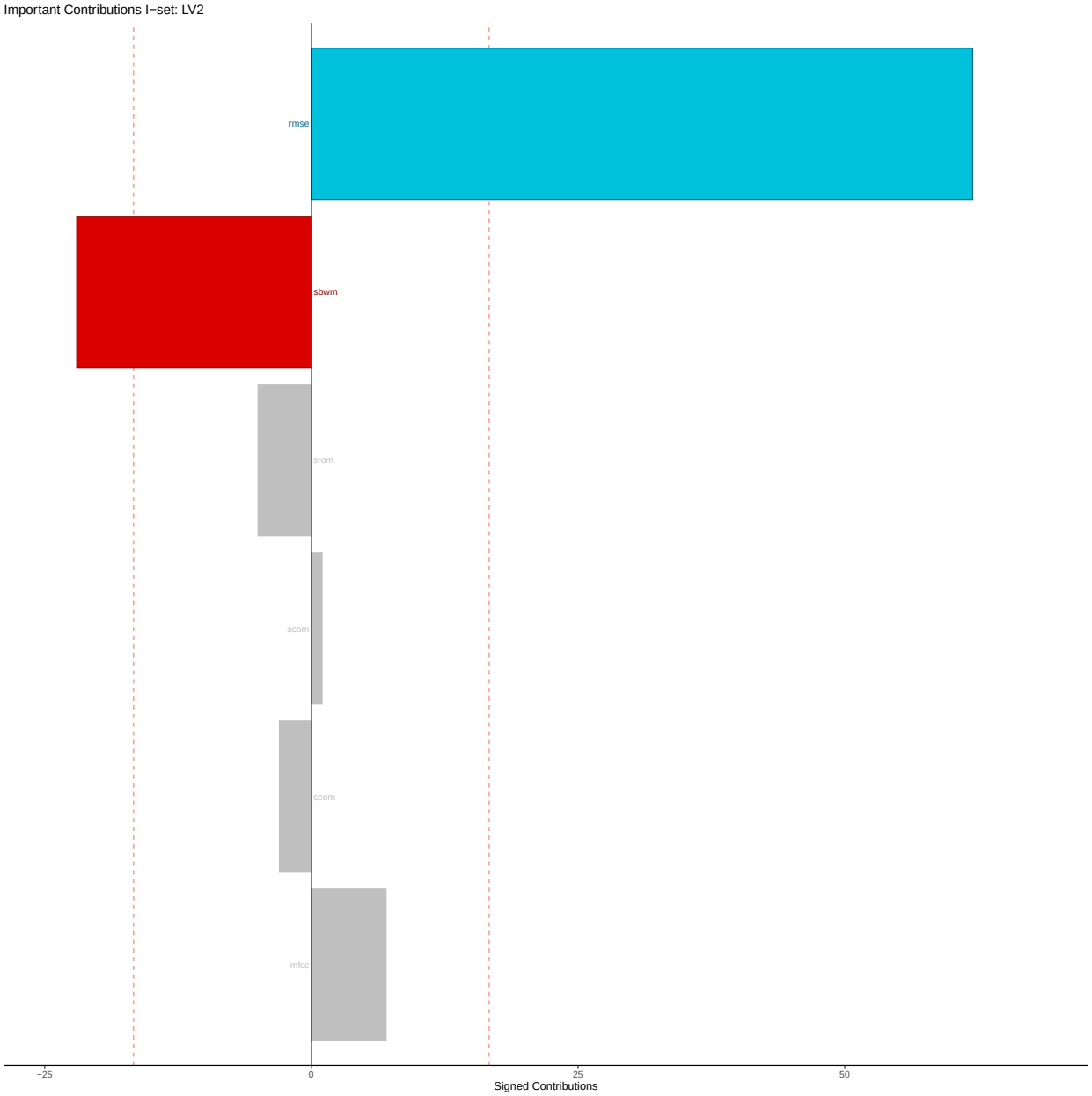
Just to check if the latent variables constructed using the second dimension say a different story.


PLSC: Second Pair of Latent Variables
0.2
0.0
−0.2
Y Latent Variable 2
sleep
workout
party
dinner
−0.2
−0.1
0.0
0.1
X Latent Variable 2

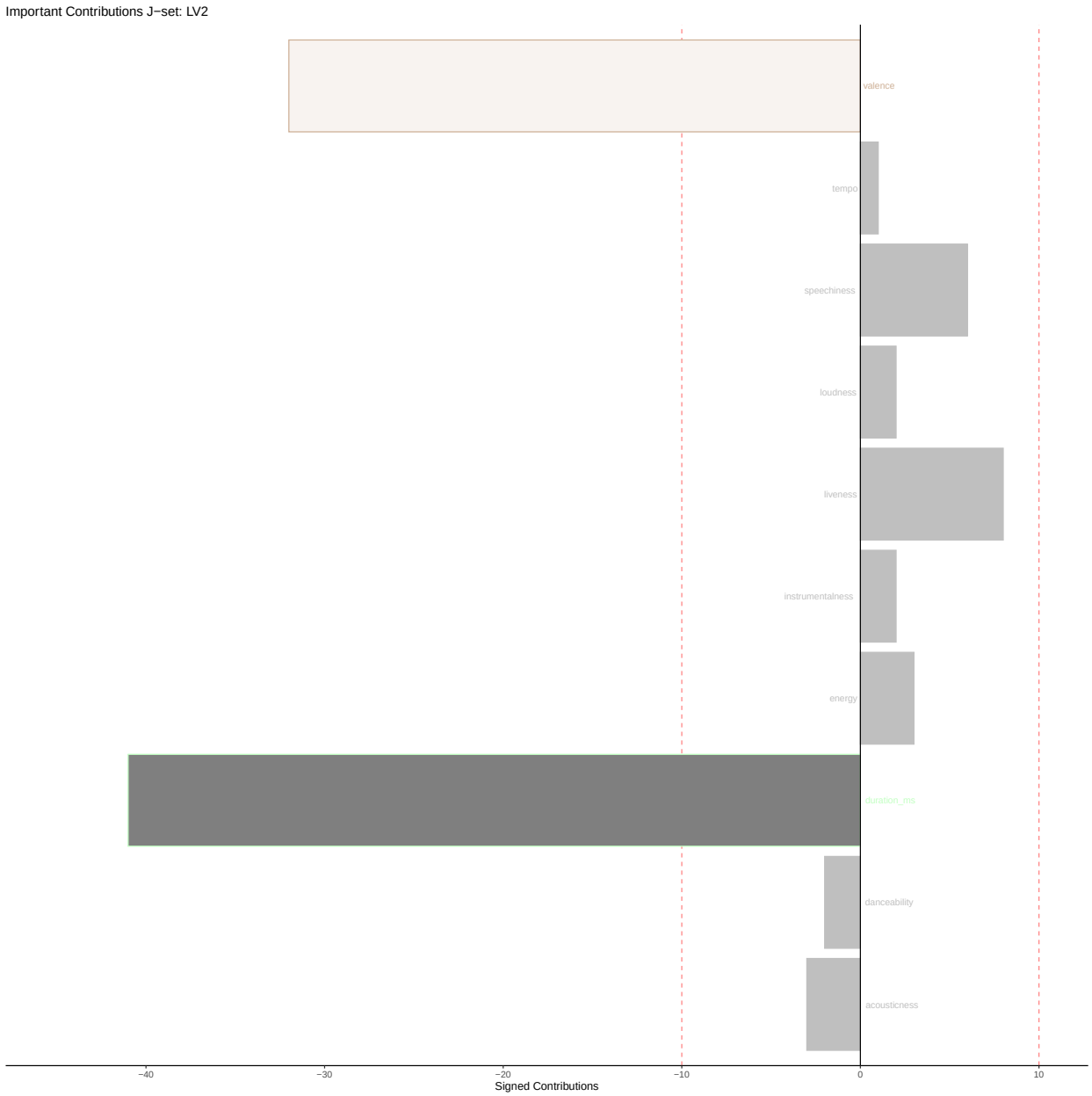
Confidence intervals generated by bootstrapping - latent variable 2



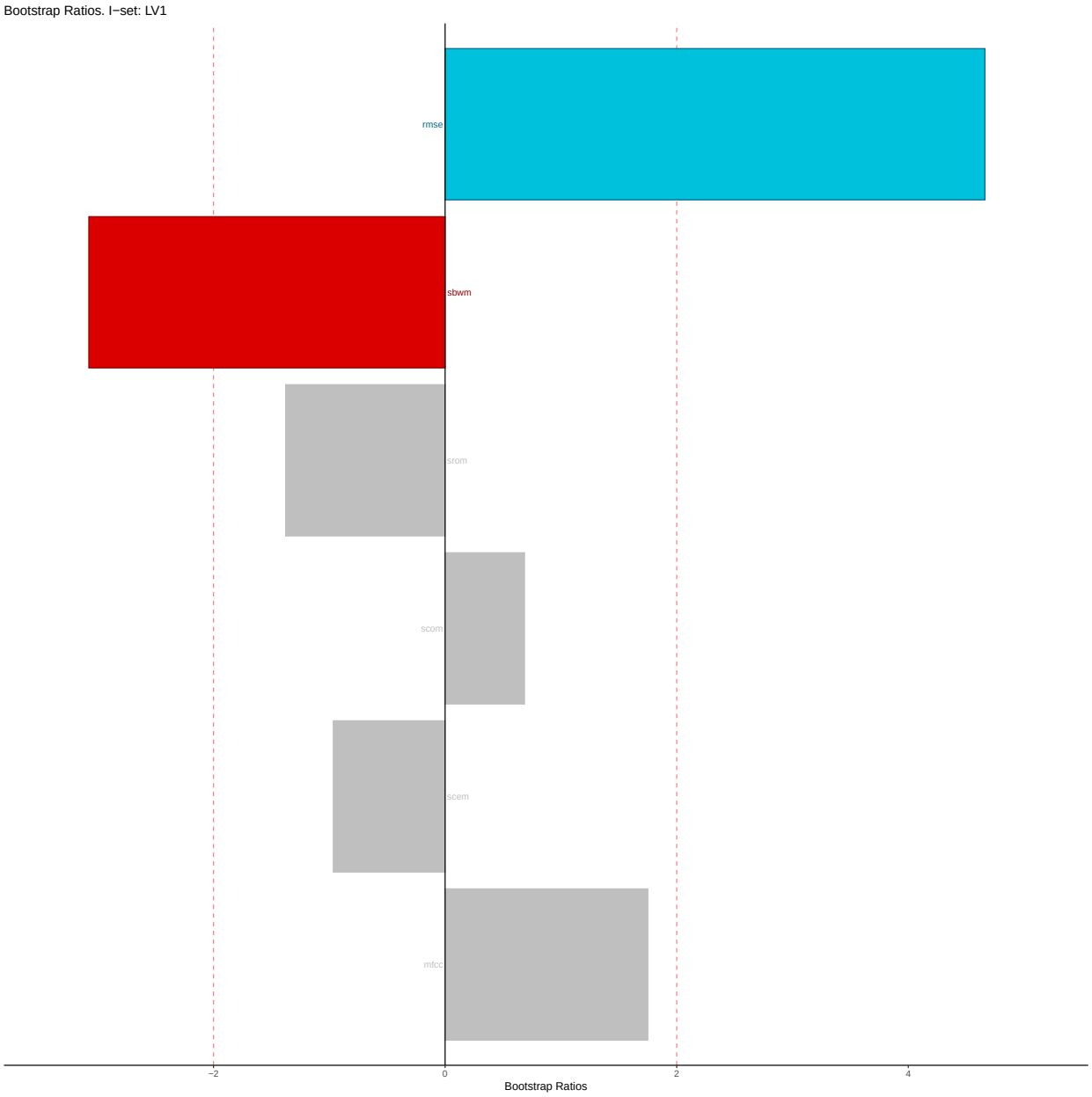
Row factor scores - latent variable 2

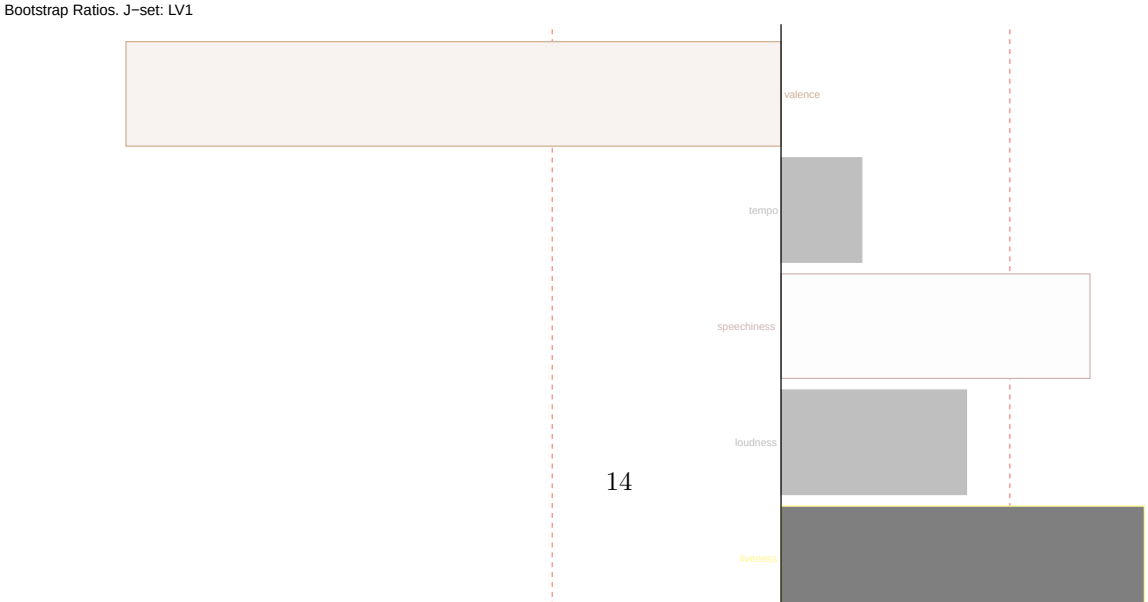


Column factor scores - latent variable 2

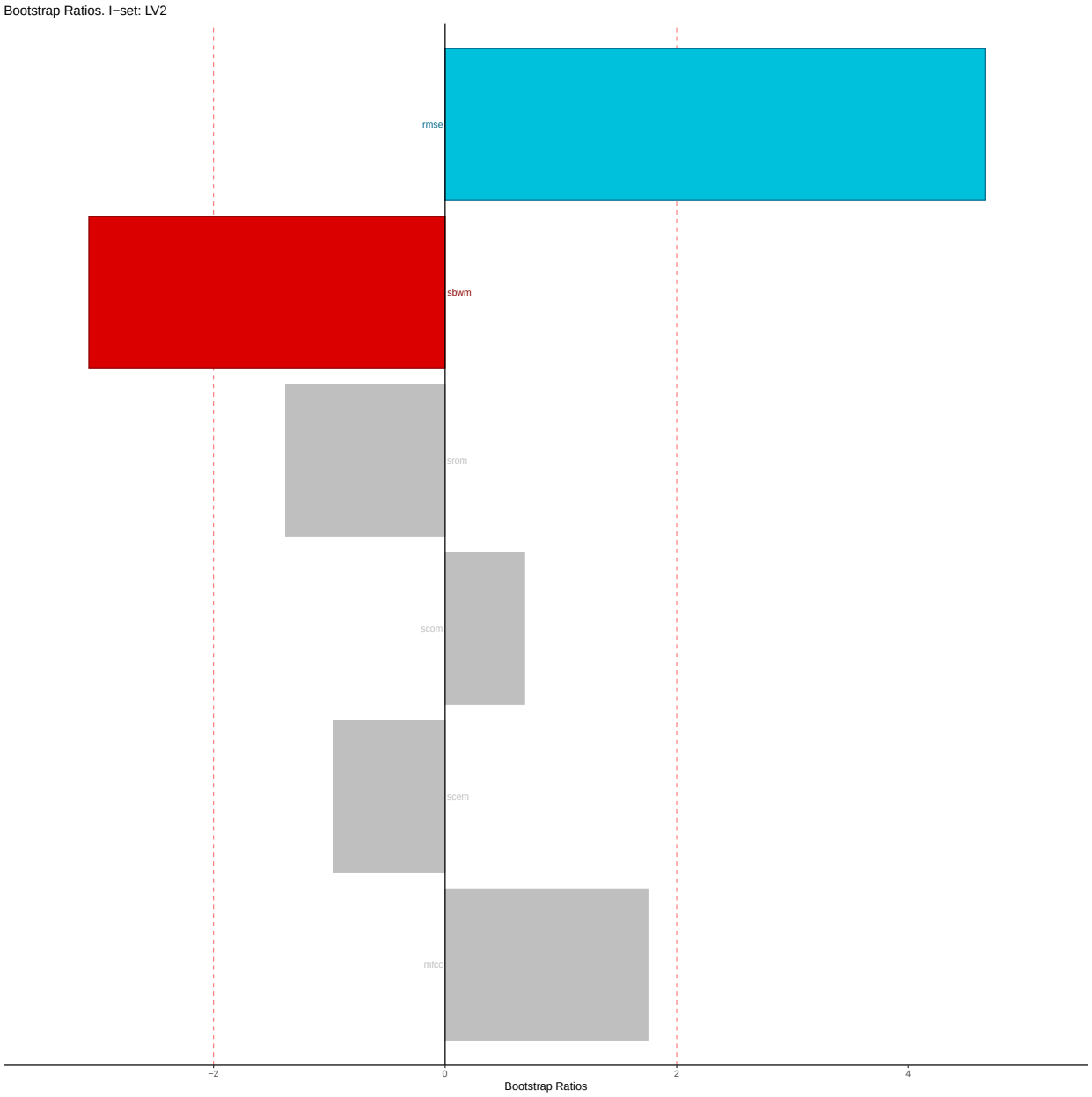


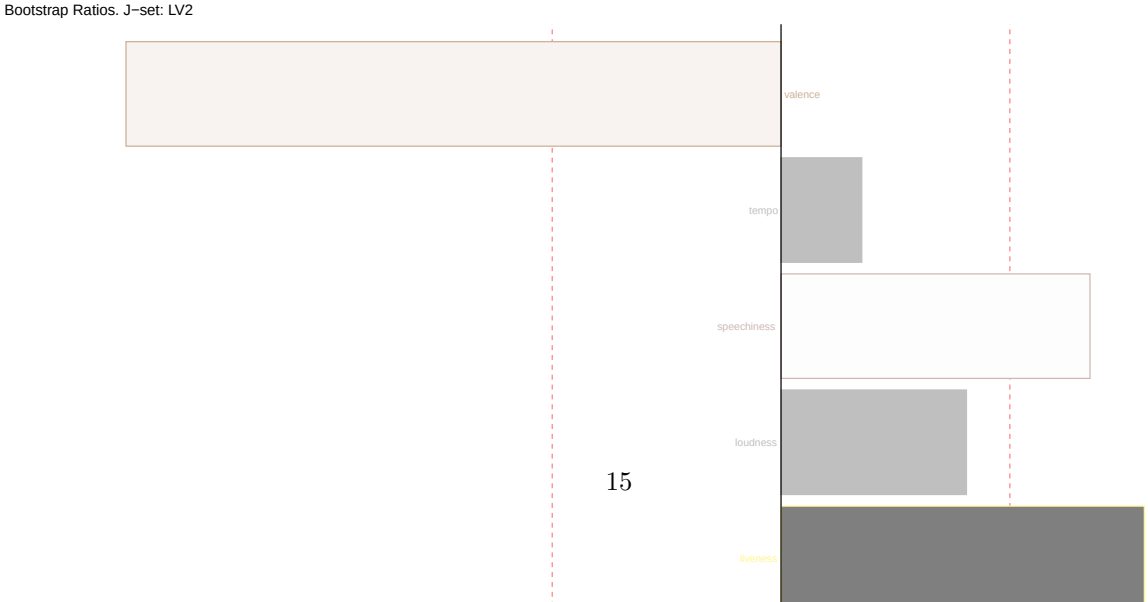
Bootstrap ratio for dimension 1 - latent variable 2





Bootstrap ratio for dimension 2 - latent variable 2





Summary:

The latent variable explaining dimension 1 is being interpreted in this summary. The maximum contributions towards musical features are from acousticness, instrumentalness, danceability, loudness, and energy. Similarly, from music.audio which describes the audio signal features, all variables except tempo contribute significantly.

Tying the two together, it is quite an intuitive match - acousticness and instrumentalness can be described by the timbre which is indicated by MFCC, spectral roll-off/bandwidth, while danceability, loudness and energy are indicated by spectral centroids and root mean square energy.

DiSTATIS

Method: DiSTATIS

Derived from the work of Escouffier, STATIS is a technique used to handle multiple data tables containing different variables measuring the same observations. It is an extension of principal component analysis.

In another variant of STATIS (called dual-STATIS), same variables measuring different observations in multiple tables are analyzed.

Either way, the first step is to analyze the similarity in data between tables and compute an optimal set of weights that are used to compute "compromises", which are similar to components in a PCA. Components are linear combinations of data tables. The second step is to conduct a PCA on the compromise, which gives an optimal map of the observations.

One of the "types" of STATIS is DiSTATIS. It is used when there are K distance matrices on the same set of observations. The distance matrices are converted into cross-product matrices and then STATIS is applied.

Source - https://bit.ly/3DbS0Tb

Data set: 19 judges sort 18 wines

This is a dataset which describes how 19 judges sort 18 wines into groups. Gender acts as the descriptor for the variables.

```
## # A tibble: 18 x 19
##       J1    J2    J3    J4    J5    J6    J7    J8    J9   J10   J11   J12   J13
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     6     1     6     3     3     3     2     1     2     5     5     3     1
##  2     3     4     2     1     2     1     3     2     2     6     5     2     1
##  3     4     1     1     5     3     3     2     1     1     4     5     2     4
##  4     2     2     3     1     4     2     3     2     4     7     4     2     6
##  5     2     3     6     5     1     4     1     1     3     3     2     3     9
##  6     1     1     4     5     1     4     2     1     2     5     5     1     2
##  7     4     2     5     1     3     4     3     2     3     4     3     1     3
##  8     4     1     4     2     1     3     2     1     1     7     1     3     4
##  9     3     5     5     4     3     2     3     2     2     4     3     2     8
## 10     2     1     1     2     2     1     1     1     5     2     2     3     4
## 11     7     1     6     3     3     3     1     1     1     7     5     3     4
## 12     5     1     6     3     1     3     2     1     1     4     1     3     2
## 13     1     1     5     3     1     4     1     1     1     1     5     3     2
## 14     4     3     3     1     4     1     3     2     5     2     3     1     7
## 15     5     1     6     3     2     3     2     1     1     5     5     3     1
## 16     5     1     2     4     1     2     2     1     1     7     5     3     3
## 17     3     4     2     1     4     1     3     2     3     6     2     2     5
## 18     8     1     1     2     1     1     2     1     1     1     5     3     4
## # ... with 6 more variables: J14 <dbl>, J15 <dbl>, J16 <dbl>, J17 <dbl>,
## #   J18 <dbl>, J19 <dbl>
```

Distance matrix

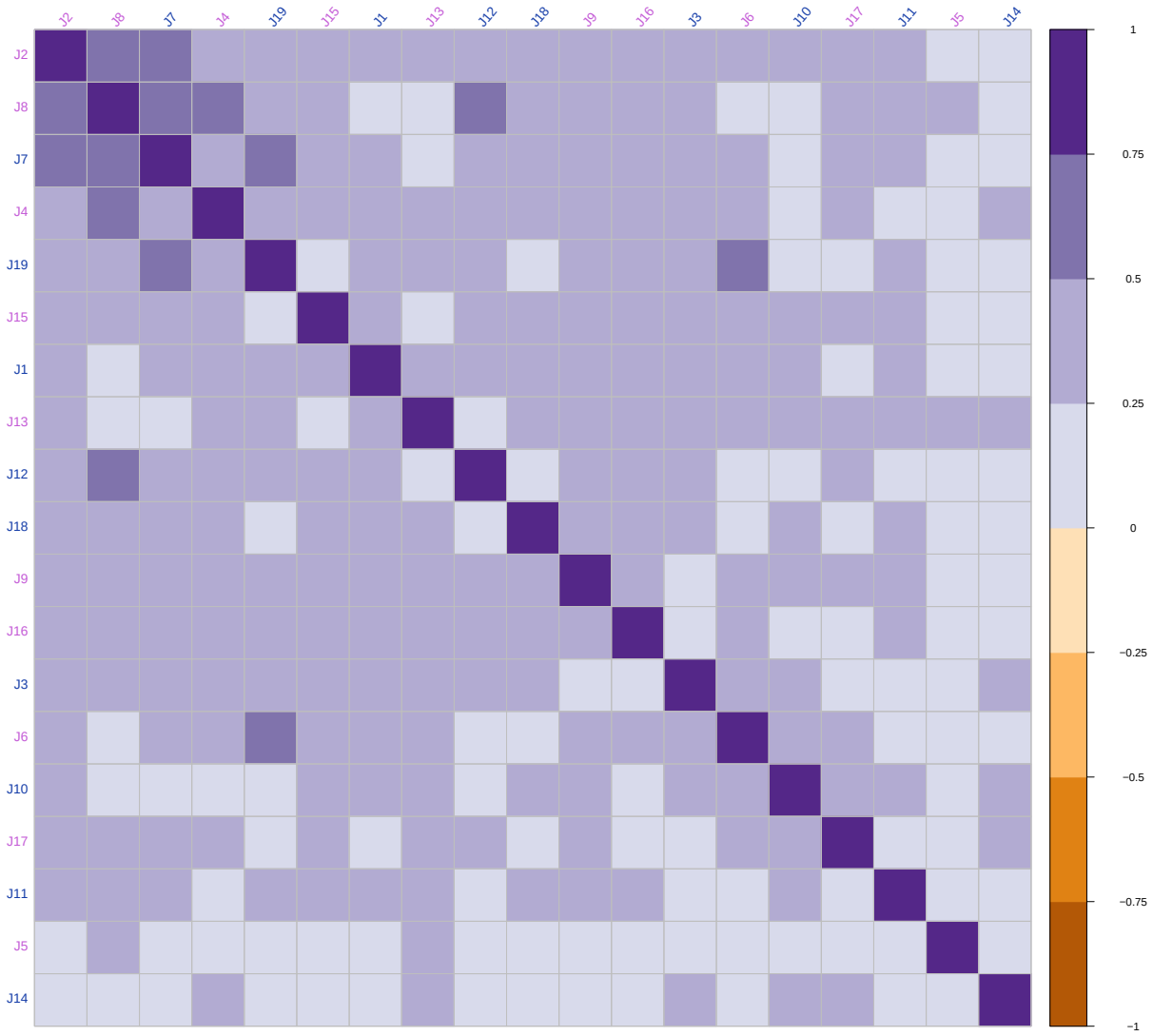
The main object used in DiSTATIS is the distance matrix or how far the judges are from each other on each observation.

```
## ----getCube-----------------------------------
DistanceCube <- DistanceFromSort(multiSort)
DistanceCube <- na.omit(DistanceCube)
```

Analysis

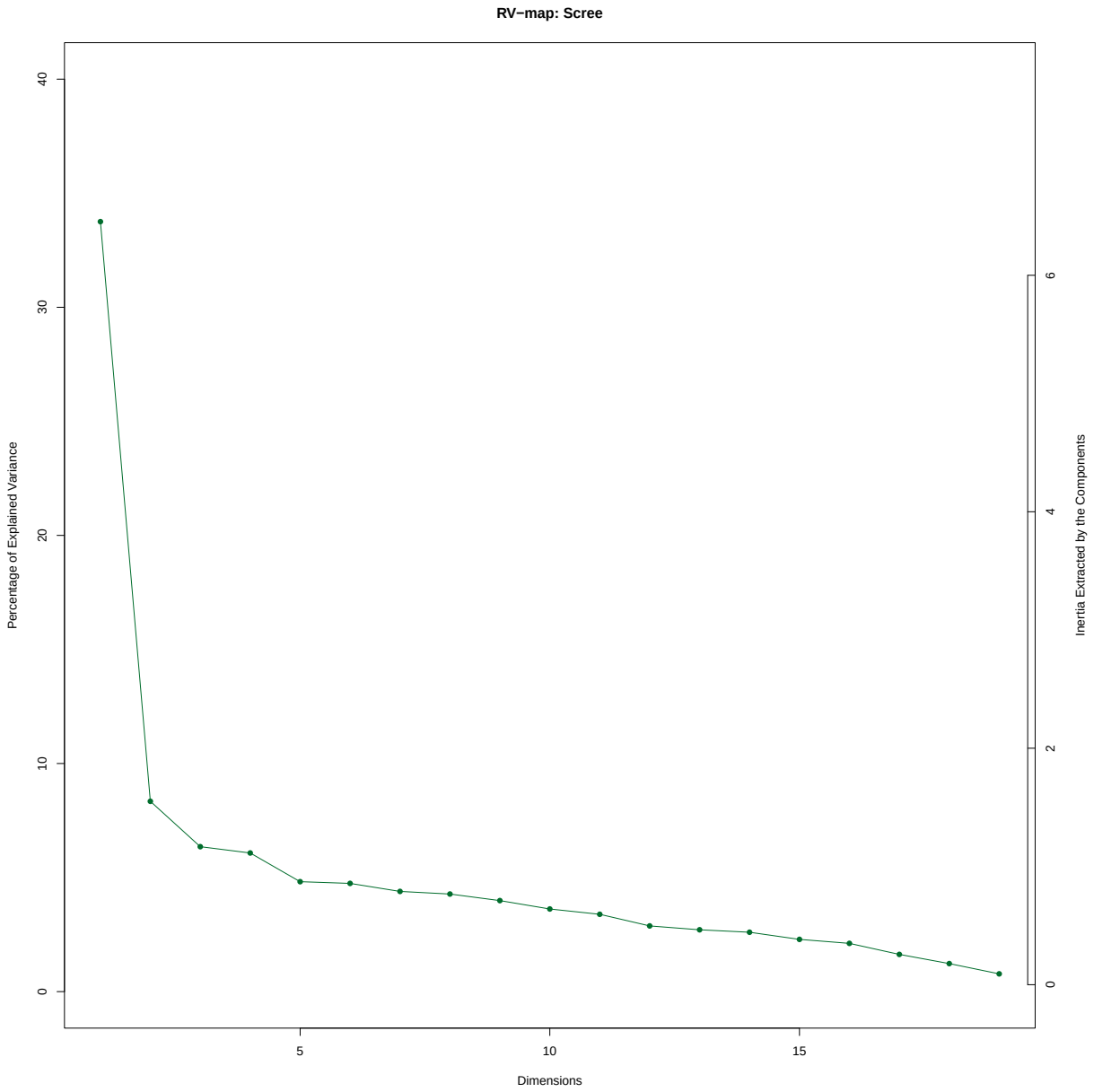
```
## runDistatis-----------------------------------
resDistatis <- distatis(DistanceCube,
                        nfact2keep = 3)
n.active <- dim(DistanceCube)[2]
```

Rv heat map



The Scree plot

The scree plot shows us how many dimensions contribute to the variance in the data and how much. In this plot, Dim 1 contributes about 47% or the variance. Hence, it would be a good place to start.

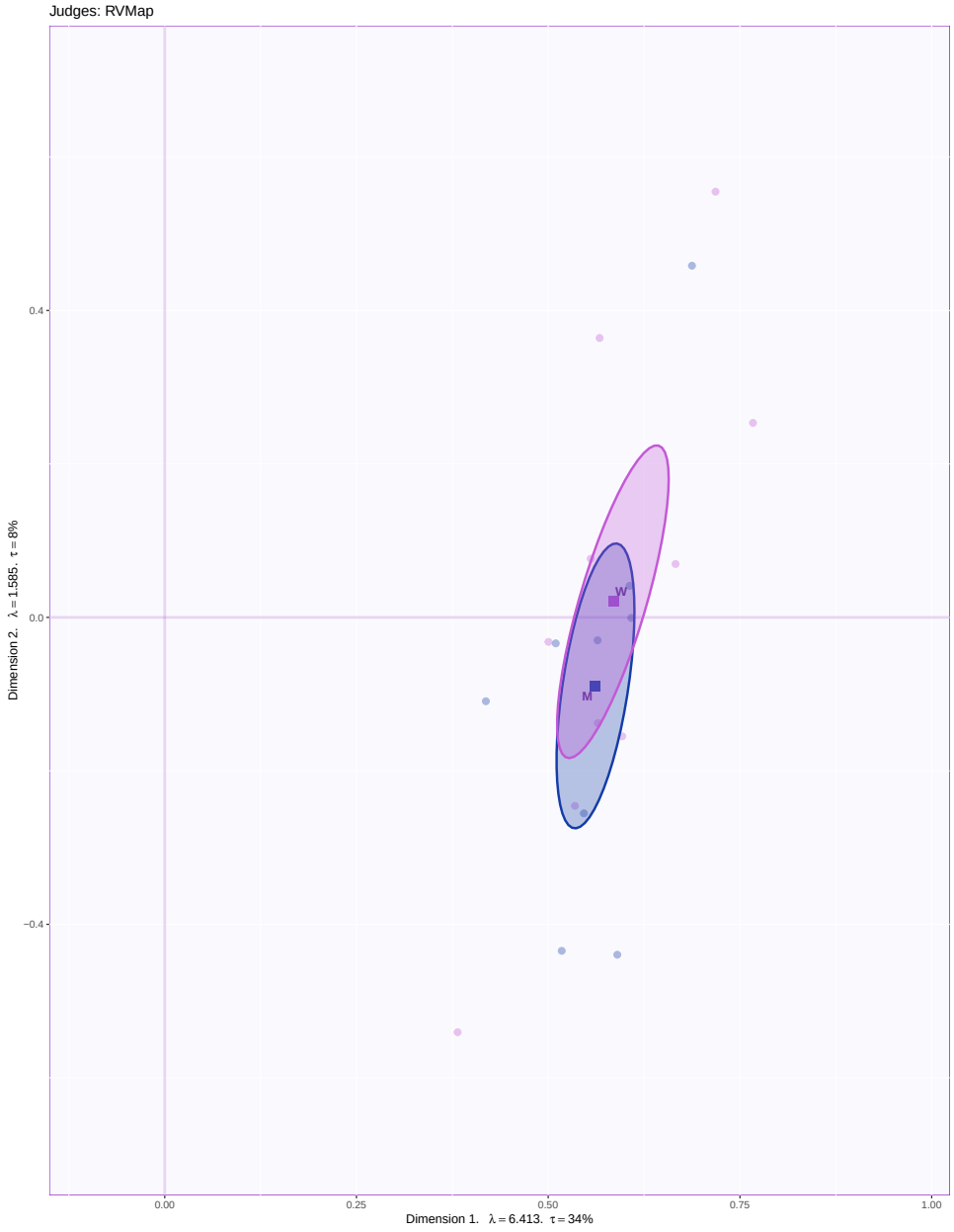


Rv Plot

As a first step, a PCA is run on the data table. The Rv coefficients are plotted on the components. The factor scores are obtained from the eigen-decomposition of the between distance matrices cosine matrix (often a matrix of Rv coefficients). The points are colored by the gender of the judge. Blue is for males and pink is for females. The 95% confidence intervals around the mean are also generated based on bootstrapping estimates.

```
## ------------------------------------------------------------------------------
## Basic Factor Maps (with ggplot2)
## ------------------------------------------------------------------------------
## $zeMap              A standard map with background, points, and Labels
## $zeMap_background   The background map
```

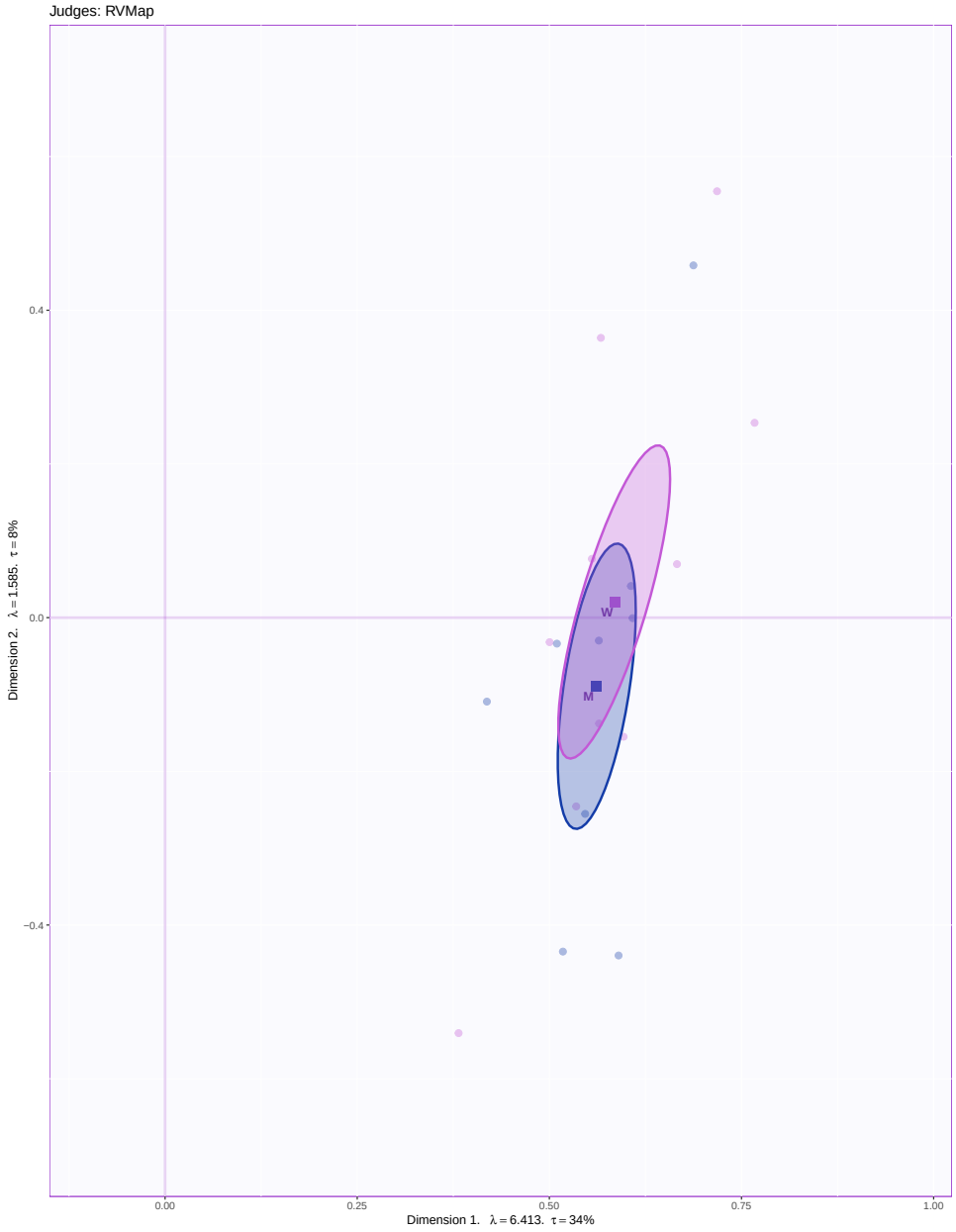
```
## $zeMap_dots         The points map
## $zeMap_text         The labels map
## $factorScores       The factor scores (coordinates)
## $constraints        map constraints (a list with minx miny maxx maxy)
## ------------------------------------------------------------------------
```

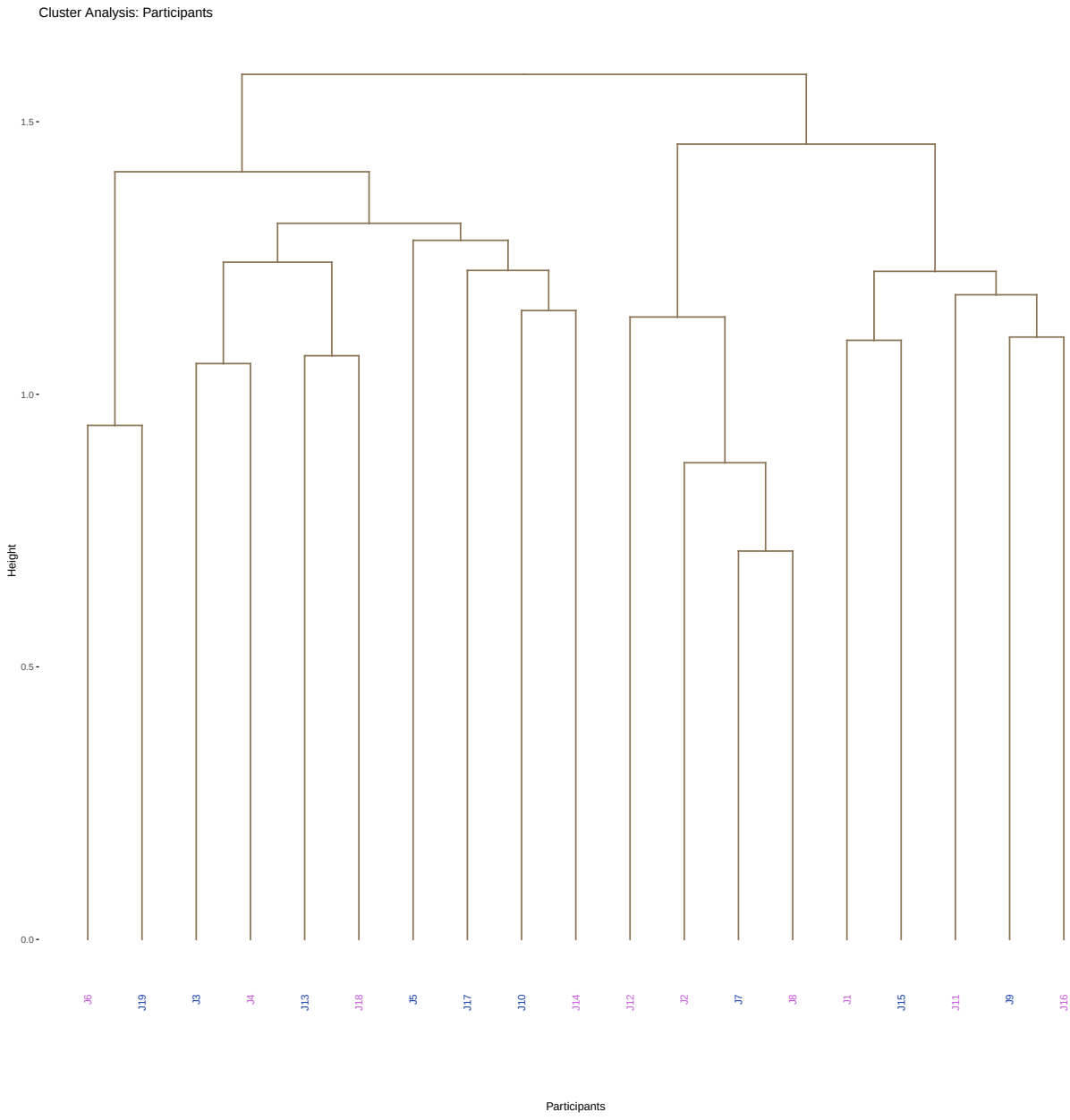


Hierarchical clustering analysis

In this step, the judges are mapped to different clusters, all placed in a hierarchical structure. The judges clustered closer to each other rated the wines in a similar fashion. (Since the gender variable is made up, we may not be seeing clear trends here).

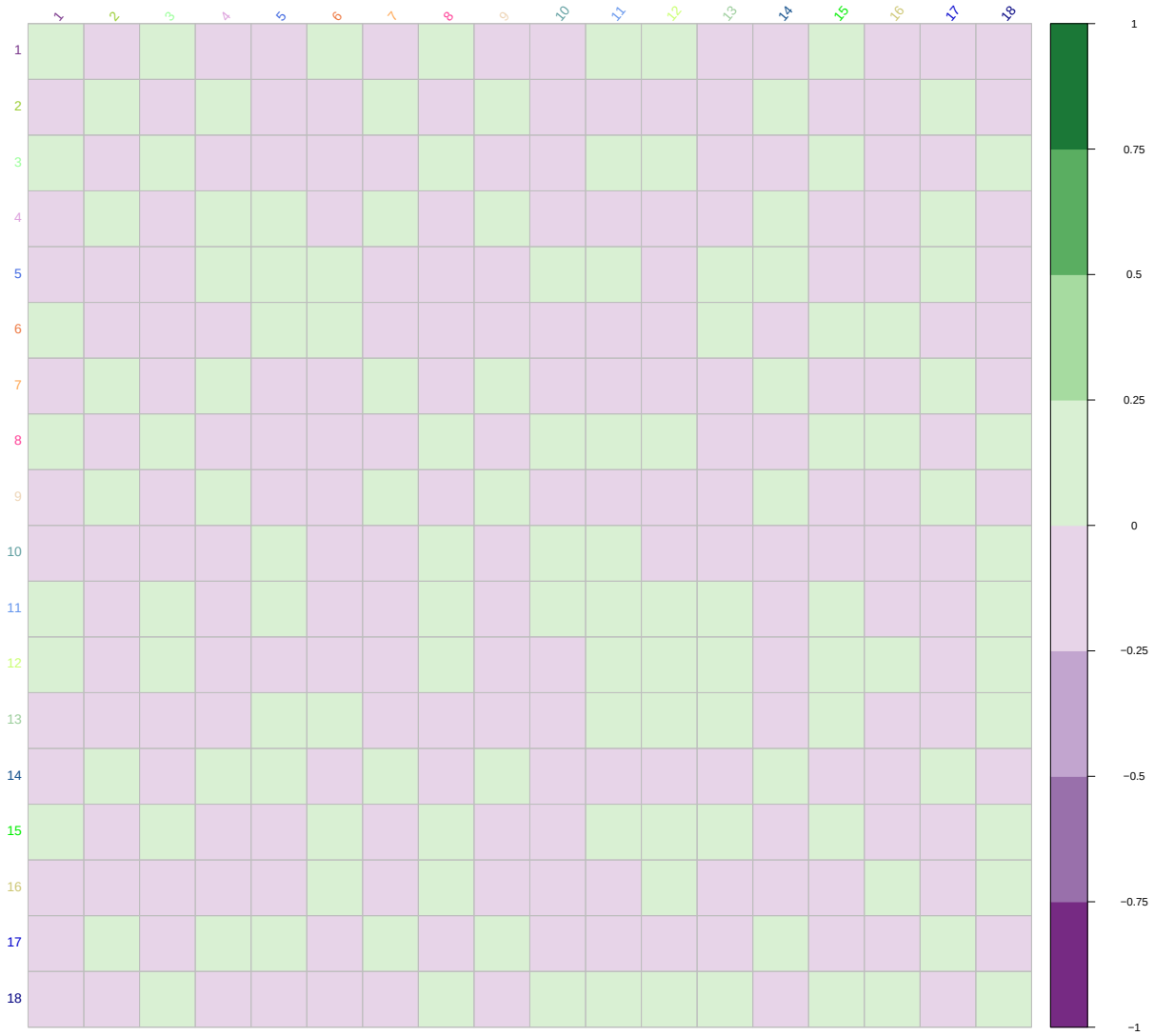
	dim 1	dim 2	dim 3
M	0.5610070	-0.0893042	0.0312385
W	0.5852095	0.0207661	-0.0227888

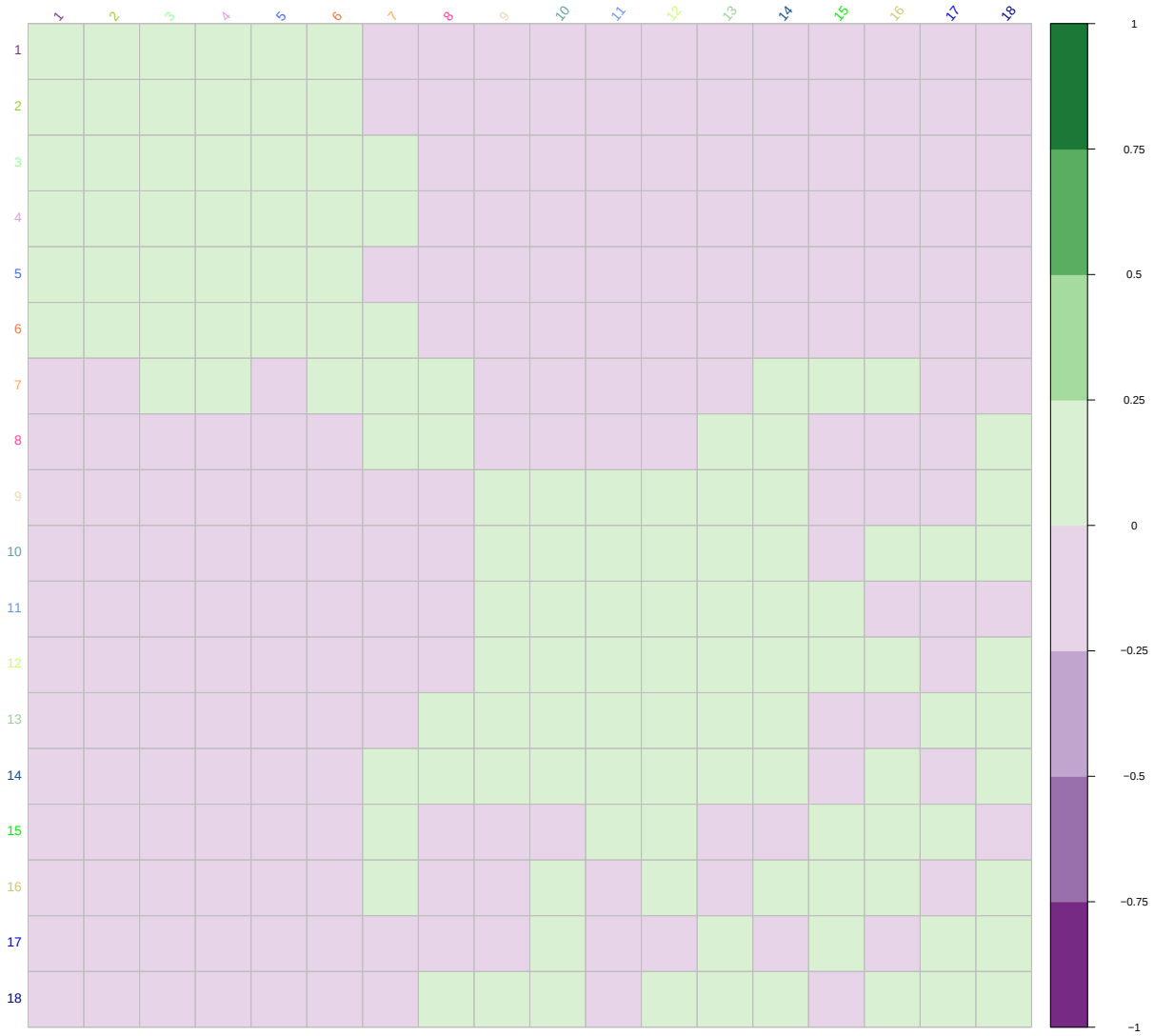




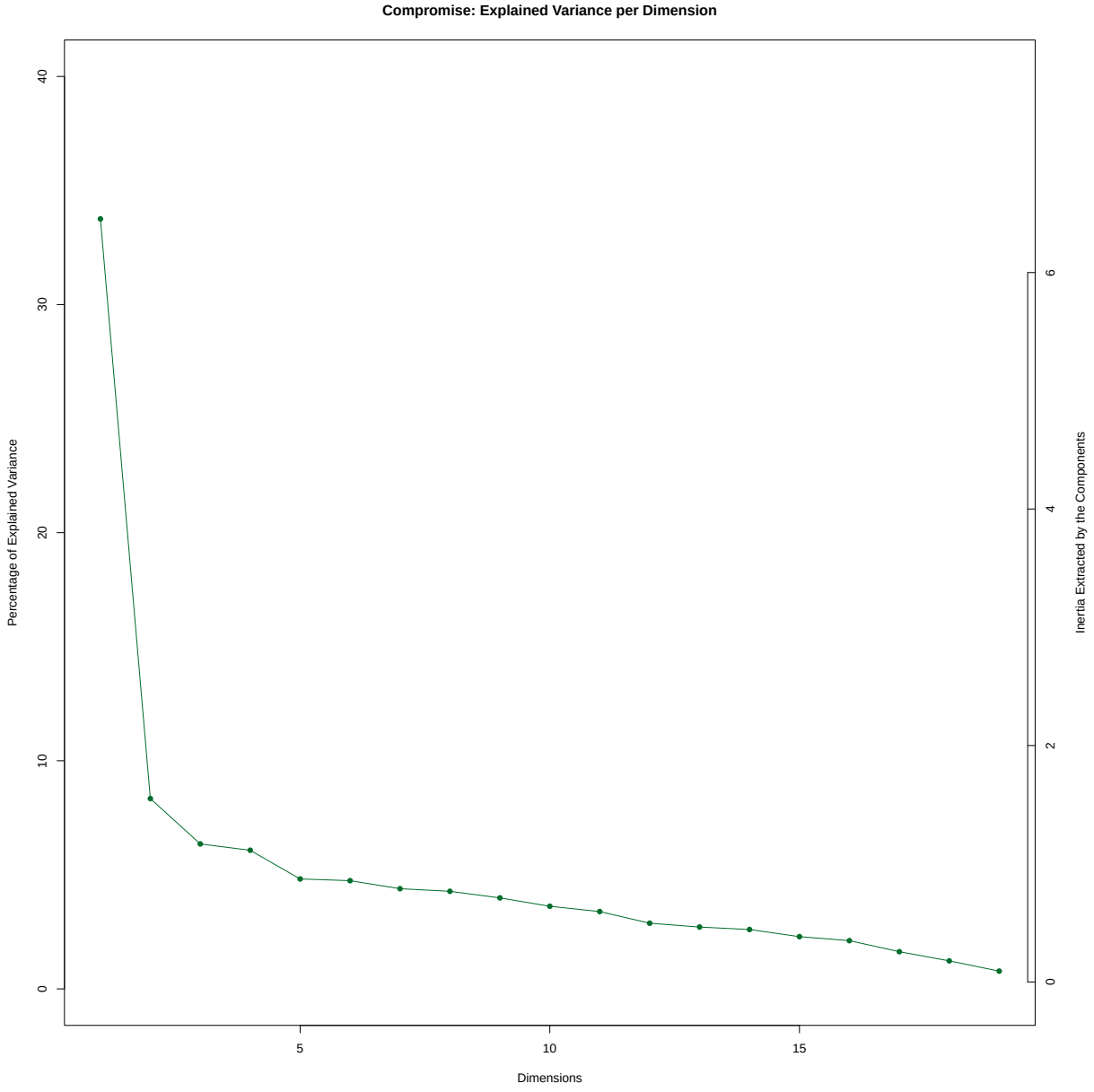
Heat map after HCA

After conducting the cluster analysis, this is what the heat map looks like.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
1
0.75
0.5
0.25
0
−0.25
−0.5
−0.75
−1


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
1
0.75
0.5
0.25
0
−0.25
−0.5
−0.75
−1

Scree plot - Compromise/Global data table



Summary

DiSTATIS was used to study a dataset which contained 18 wines judged by 19 judges. The global factor scores map shows that reds and whites were judged distinctly.

Multiple factor analysis

Method: Multiple Factor Analysis (MFA)

Much like the other techniques, Multiple factor analysis (aka multiple factorial analysis) is an extension of PCA specifically proficient in handling multiple data tables containing different variables measuring the same observations, or vice versa in dual-MFA (same variables measuring different observations).

Similar to DiSTATIS, MFA proceeds in 2 steps. The first step is to do a PCA of each data table and 'normalize' the data tables by using the first singular value obtained in the PCA. The next step is to aggregate all the normalized data tables into a grand data table and run a PCA again on that. This PCA gives factor scores for the observations and loadings for the variables on a "Global" map.

Additionally, MFA provides a "partial factor scores" plot for each data table which reflects the specific view-point of the data table.

Source - https://bit.ly/3H88Wxc

Data set: Audio features

This is a dataset which describes audio features of songs in Spotify playlists.

Specifically, the music.track dataset measures 165 songs on 16 variables, of which 11 are quantitative. Some of the audio features described are acousticness, danceability, and energy. Additionally, music.audio contains 7 quantitative variables that describe features of the audio signals; mel-frequency cepstral coefficients (MFCC) –> timbre, spectral centroid (SCEM) –> brightness of sound, spectral contrast (SCOM) –> harmonic/non-harmonic music, spectral roll-off (SROM) & bandwidth (SBWM) –> timbre, tempo –> estimated tempo of the track, root mean square energy (RMSE) –> energy per frame

The music.track table has been further divided into 2 tables - one containing only duration and tempo, with the other containing the rest of the variables. This is done because MFA needs at least 3 tables for analysis.

Dataset 1 –> Audio signal features Dataset 2 –> Musical features (perceived) Dataset 2 –> Technical features (Duration, Tempo in BPM)

```
##         mfcc     scem     scom     srom     sbwm   tempo.x     rmse
## 1  2.1820905 2008.227 22.16629 4271.333 2362.907 129.19922 4.348369
## 2 -0.4209786 1810.502 22.81968 3800.144 2217.649  89.10291 3.749902
## 3  0.4319727 1920.575 20.06893 4231.883 2553.859  95.70312 4.450774
## 4  5.9844995 1777.566 21.45459 3898.110 2247.093 161.49902 6.248868
## 5  1.4752268 2629.842 21.64794 5873.931 2734.930  89.10291 3.963258
##   acousticness
## 1        0.845
## 2        0.843
## 3        0.873
## 4        0.369
## 5        0.050
```

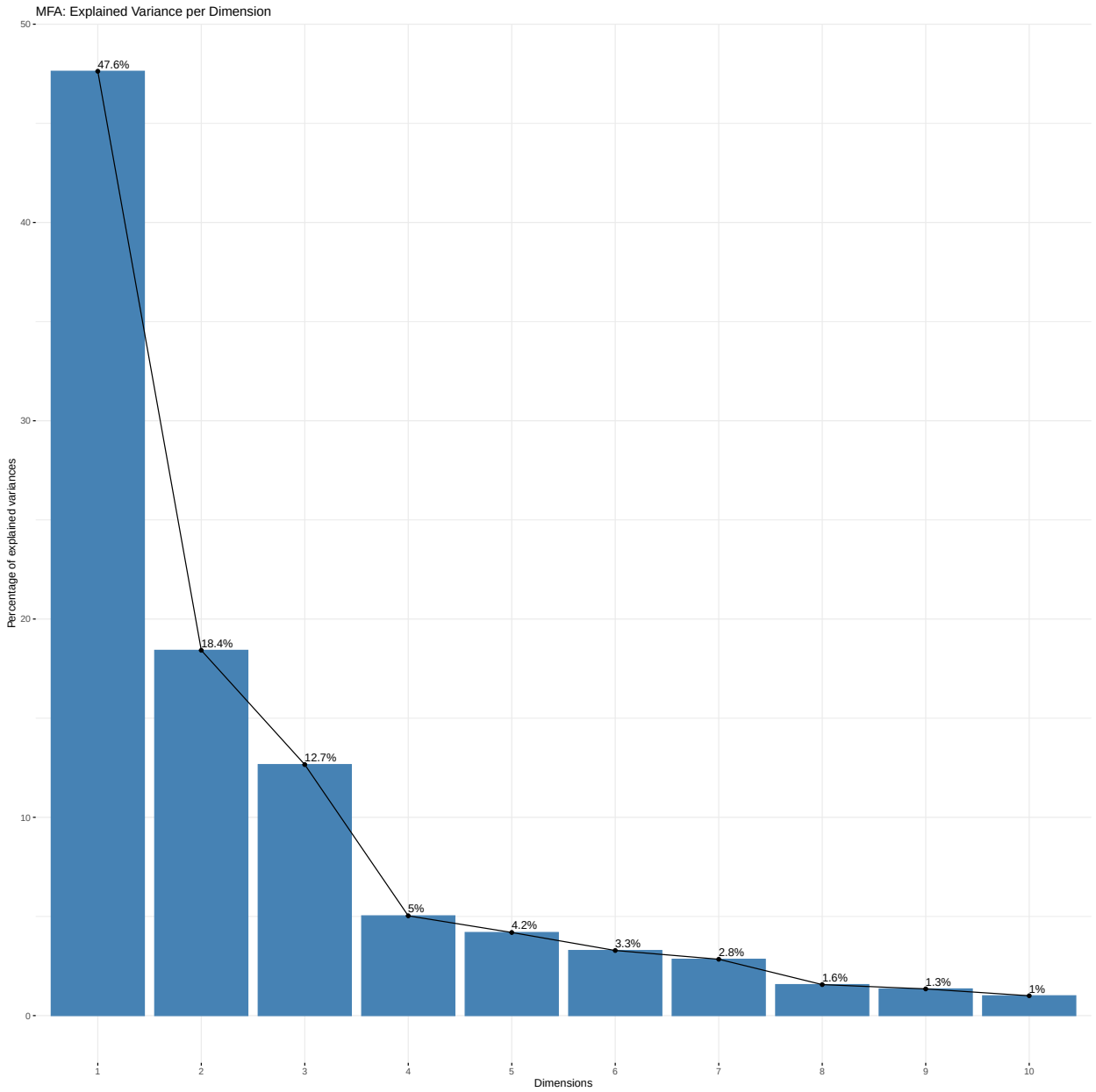
Analysis

```
## call  MFA ----
resMFA <- FactoMineR::MFA(finaldataset,
     group = c(7, 8, 2),
     type = c("s", "s", "s"),
```

```
    name.group = c("audiosignal", "musicalfeature", "technicalfeature"),
    graph = FALSE
)
```

The Scree plot

The scree plot shows us how many dimensions contribute to the variance in the data and how much. In this plot, Dim 1 contributes about 47% or the variance. Hence, it would be a good place to start.

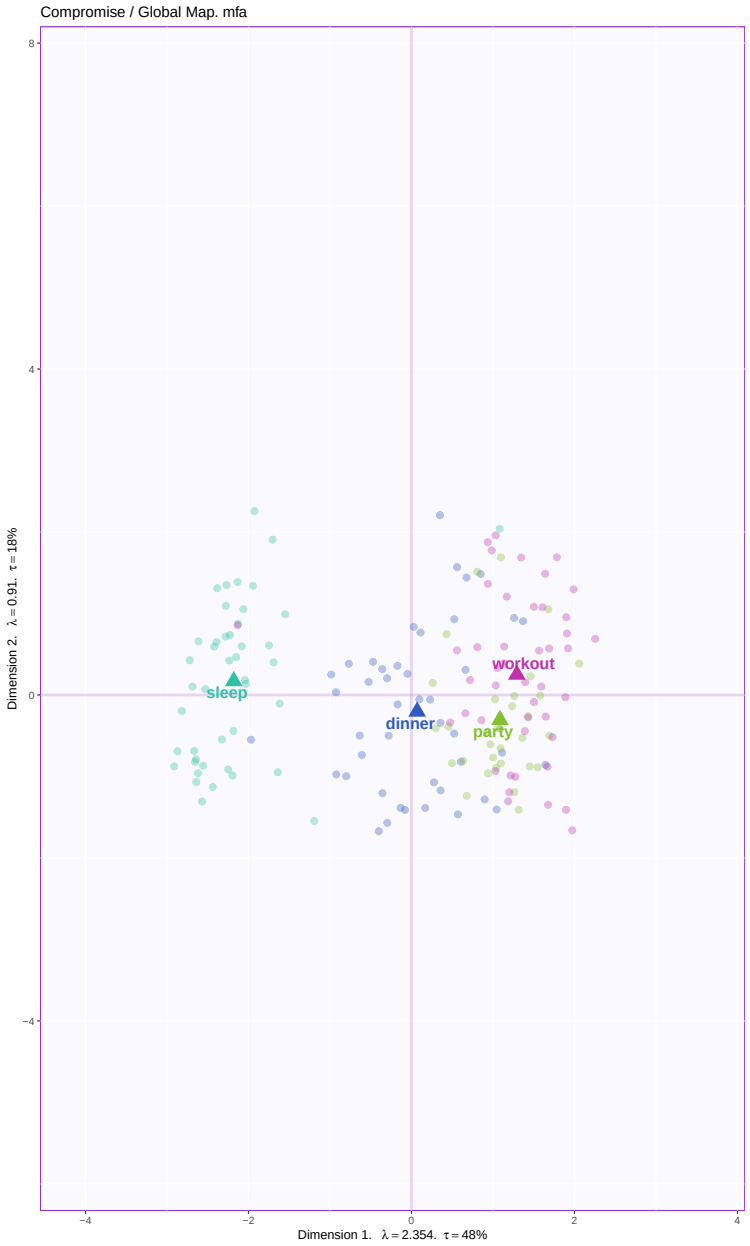


Rv Plot

As a first step, a PCA is run on the data tables as if they are 3 variables (so to say). The Rv coefficients are plotted on the components. The Rv coefficient can be interpreted as a non-centered squared coefficient of correlation between two matrices. We see in the plot that the audio signals and musical features overlap almost entirely, which is the trend observed in previous multivariate techniques as well. Duration and Tempo always stood out in Dimension 2.

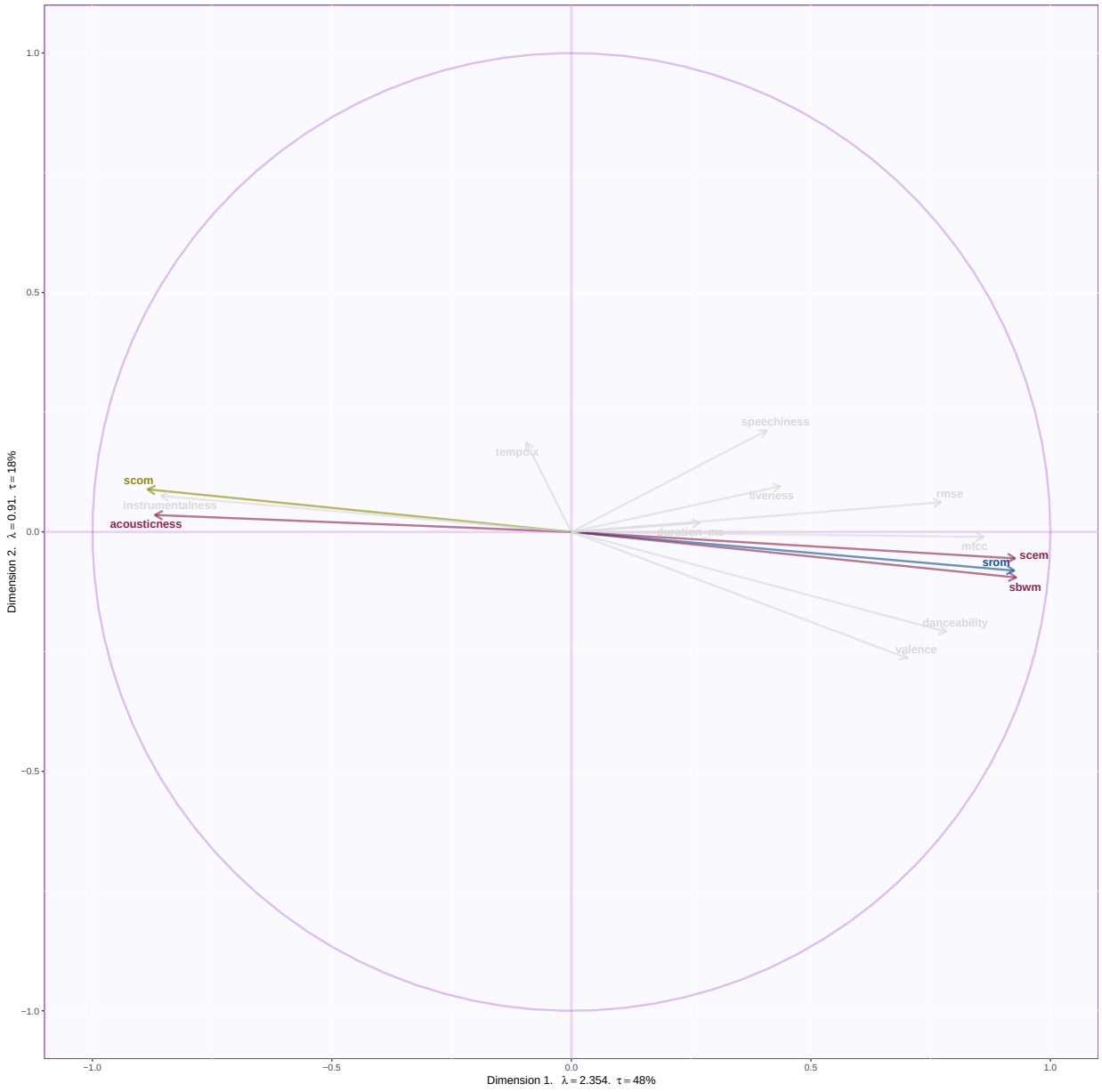
Global factor scores

This is a plot of the factor scores from the grand data table. The dots are colored by genre. Dinner is closer to party and workout songs compared to its distance from sleep songs.



Correlation between variables and factors

Back to PCA stuff! The circle of correlations can be interpreted using the angle between 2 variables (correlation magnitude and direction) and the distance between the circumference and arrowhead (the ones closer to the circumference are more important). Another method of visualizing would be to grey out the unimportant ones, like in this plot. SCOM and acousticness and strongly negatively correlated with SCEM, SBWM, SROM (same trend seen earlier in PLSC).



Summary:

From the three data tables, we infer that audio signal features and musical perception features are closer to each other and duration and tempo are farther apart from the former two. Also, acousticness and spectral contrast (harmonic/non-harmonic music) are strongly negatively correlated with spectral centroid, bandwidth and rolloff, all of which measure the timbre of an instrument.