# Country Environmental Factors correlated with Happiness

*Ritesh Malaiya*

*2018-12-10*

# Contents

# Part I

# Country-Level Environmental Measurements Dataset

# Chapter 1

# Introduction

## 1.1 Context

Assessing country-level social and economic statistics are often limited to socio-economic data. Not any more! This dataset will be maintained and updated with miscellaneous environmental data for countries across the globe.

## 1.2 Content

This data is all acquired through Google Earth Engine (https://earthengine.google.com/) where publicly available remote sensing datasets have been uploaded to the cloud to be manipulated by the average Joe like you and I. Most of the data is derived by calculating the mean for each country at a reduction scale of about 10km.

## 1.3 Inspiration

Can you use environmental statistics to predict social and economic data? Are people more happy in sunny countries? How do economies in forested countries compare with those dominated by grassland/desert?

## 1.4 Research Question (scope of this book)

Which of the varables correlate most with Happiness?

# Chapter 2

# Dataset

- Data: Measurements of environment conditions in Countries
- Rows: There are 137 observations, 1 for each country.
- Columns: Total 29 variables
- Qualitative: Country (nominal), Happiness (Ordinal).
- Quantitative: Aspect, Slope Crop Land, Tree Canopy Wind Cloud & Multiple variables for Temp & Rain

## 2.1 Structure of Data

```
## 'data.frame':    137 obs. of  29 variables:
##  $ Country                : Factor w/ 137 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Happiness_Rank         : Ord.factor w/ 3 levels "VH"<"H"<"U": 3 2 2 3 1 3 1 1 2 2 ...
##  $ accessibility_to_cities: num  317.7 73.8 1212.8 378.2 209.2 ...
##  $ elevation              : num  1832 652 557 1061 683 ...
##  $ aspect                 : num  201 192 185 174 145 ...
##  $ slope                  : num  1.516 1.89 0.171 0.193 0.624 ...
##  $ cropland_cover         : num  9.51 23.35 3.69 2.79 21.96 ...
##  $ tree_canopy_cover      : num  0.375 12.805 0.177 19.87 8.834 ...
##  $ isothermality          : num  35.9 33.2 40.3 64.3 49.9 ...
##  $ rain_coldestQuart      : num  128.72 392.51 25.29 8.05 79.09 ...
##  $ rain_driestMonth       : num  1.722 40.088 0.935 0.26 17.183 ...
##  $ rain_driestQuart       : num  8.3 138.15 6.09 4.43 60.49 ...
##  $ rain_mean_annual       : num  311.3 1151.1 79.5 1023.4 539.9 ...
##  $ rain_seasonailty       : num  91.6 38.5 67.1 91.5 48.3 ...
##  $ rain_warmestQuart      : num  12.69 138.33 9.51 318.54 183.14 ...
##  $ rain_wettestMonth      : num  67.8 159 13.4 202.2 79.2 ...
##  $ rain_wettestQuart      : num  175.8 435.9 33.3 524.3 211.7 ...
##  $ temp_annual_range      : num  40.3 27.1 36.5 21.5 26.8 ...
##  $ temp_coldestQuart      : num  -0.261 3.58 13.152 18.794 8.024 ...
##  $ temp_diurnal_range     : num  14.72 9.11 14.87 13.85 13.46 ...
##  $ temp_driestQuart       : num  21.1 19.6 26.9 18.9 11.1 ...
##  $ temp_max_warmestMonth  : num  32 26.3 41.5 31 28.2 ...
##  $ temp_mean_annual       : num  11.5 11.5 23 21.6 14.2 ...
##  $ temp_min_coldestMonth  : num  -8.312 -0.806 5.058 9.549 1.443 ...
##  $ temp_seasonality       : num  88.2 62.7 75.1 18.5 47.6 ...
##  $ temp_warmestQuart      : num  22.7 19.6 32.5 23.3 20.2 ...
```

Table 2.1: Environment variables for 137 countries

| Country | Happiness_Rank | accessibility_to_cities | elevation | aspect | slope |
|---|---|---|---|---|---|
| Afghanistan | U | 317.71575 | 1831.7444 | 201.4298 | 1.5156001 |
| Albania | H | 73.83086 | 651.8155 | 192.1303 | 1.8900753 |
| Algeria | H | 1212.79982 | 556.7583 | 184.9747 | 0.1708615 |
| Angola | U | 378.20239 | 1061.4790 | 174.2569 | 0.1926286 |
| Argentina | VH | 209.21958 | 682.7993 | 145.0314 | 0.6238553 |
| Armenia | U | 97.29452 | 1850.4830 | 183.5375 | 2.3188956 |

```
##  $ temp_wettestQuart      : num  3.95 5.27 20.81 22.76 16.48 ...
##  $ wind                   : num  3.43 2.47 4.03 2.16 4.27 ...
##  $ cloudiness             : num  114.2 181.1 90.7 187.5 159 ...
```

## 2.2 Datatable

# Part II

# Single Table for Country Environment Dataset

# Chapter 3

# Principal Component Analysis

## 3.1 Description

Principal component analysis (PCA), part of descriptive analytics, is used to analyze one table of quantitative data, specifically useful for *high dimensional data* and comparitively lesser data rows. 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 and Loadings describe the column variables. Both can be interpreted through their distance from origin. However, Factor scores are also interpreted by the distances between them and Loadings interpreted by the angle between them.

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.

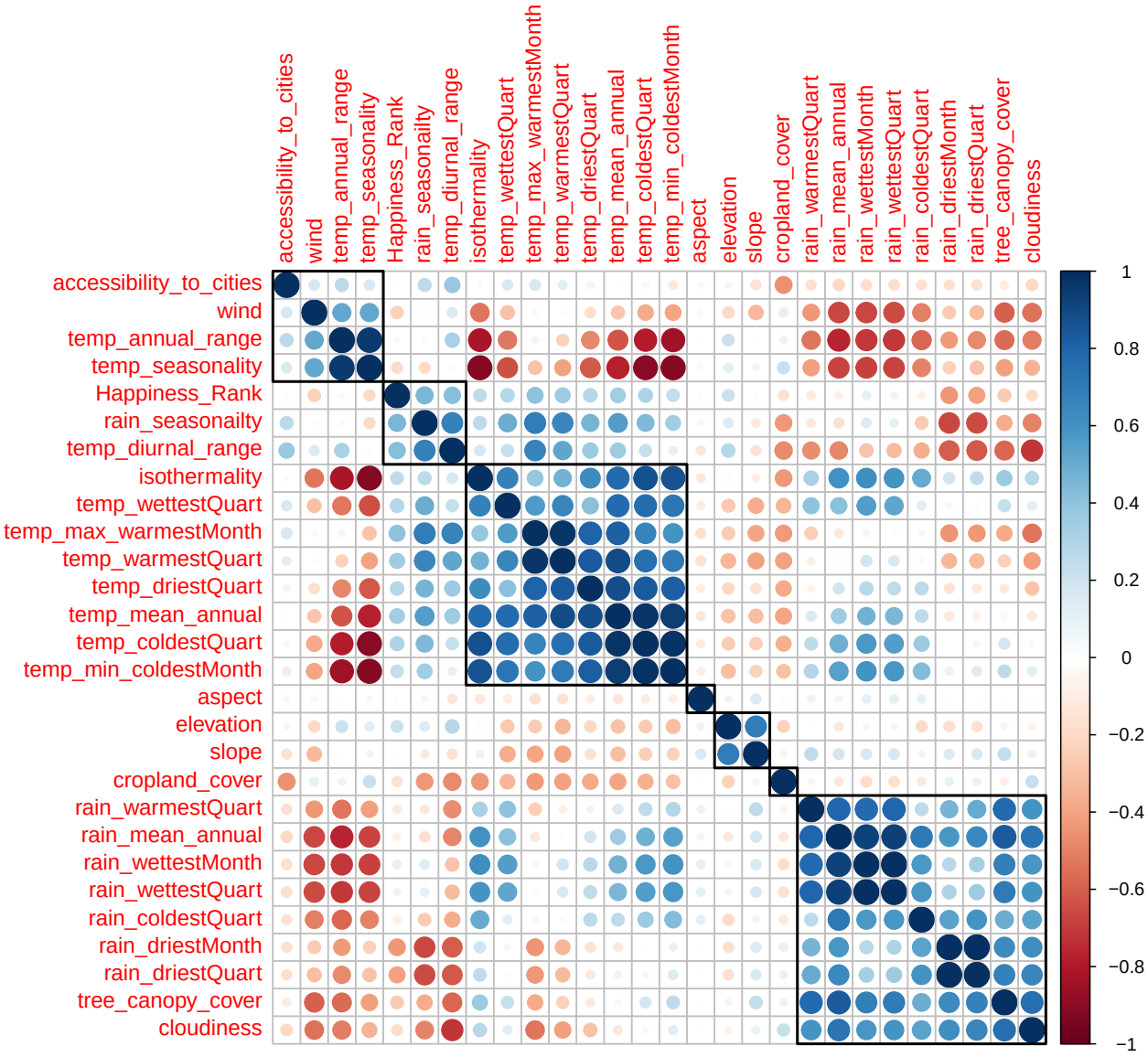
With both Factor and Loadings maps combined we can interpret which grouping criteria of rows of data is most impacted by which columns. This can interpreted visually by observing which a factors and loadings on a particular component and the distance on this component.

PCA also helps in *dimensionality reduction.* Using SVD, we get eigen values arranged in descending order in the diagonal matrix. We can simply ignore the lower eigen values to reduce dimensions. We can also take help of SCREE plot to visually analyze importance of eigen values.

There are multiple variables representing rain and Temp. Hence, for analysis purposes, lets choose annual mean for Rain and Temp.

## 3.2 Correlation Plot

Visually analyze multicollinearity in the system.

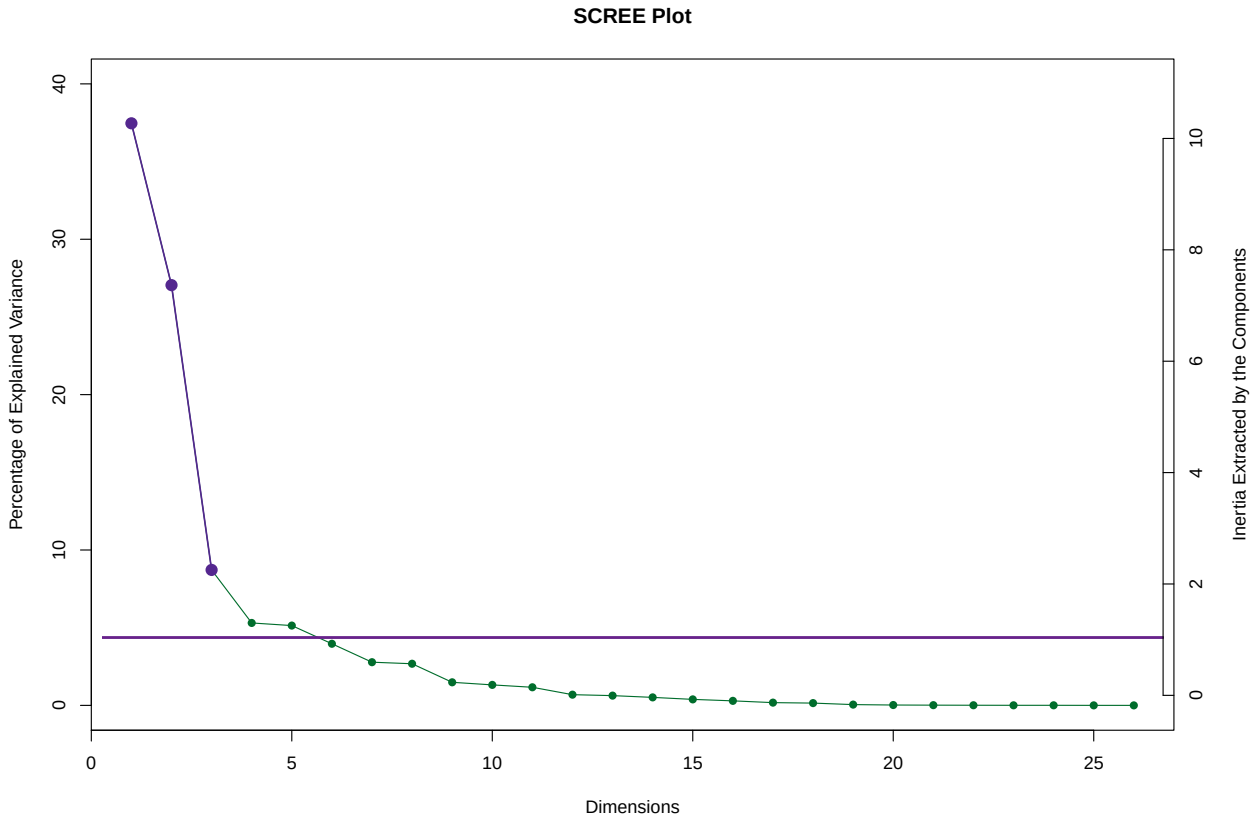



Now we have Factor scores and Loadings. * Factor Scores are the new Data points w.r.t. new Components achieved with help of SVD. * Loadings represent correlation between variables w.r.t the choosen Components. Can be interpreted in 3 ways + As slices of inertia of the contribution data table w.r.t. the choosen Components + As correlation between columns (features) of Original Data and Factor scores of each Components (latent features). + As coefficients of optimal linear combination i.e. Right Sigular Vectors (Q matrix of SVD)

## 3.3 Scree Plot

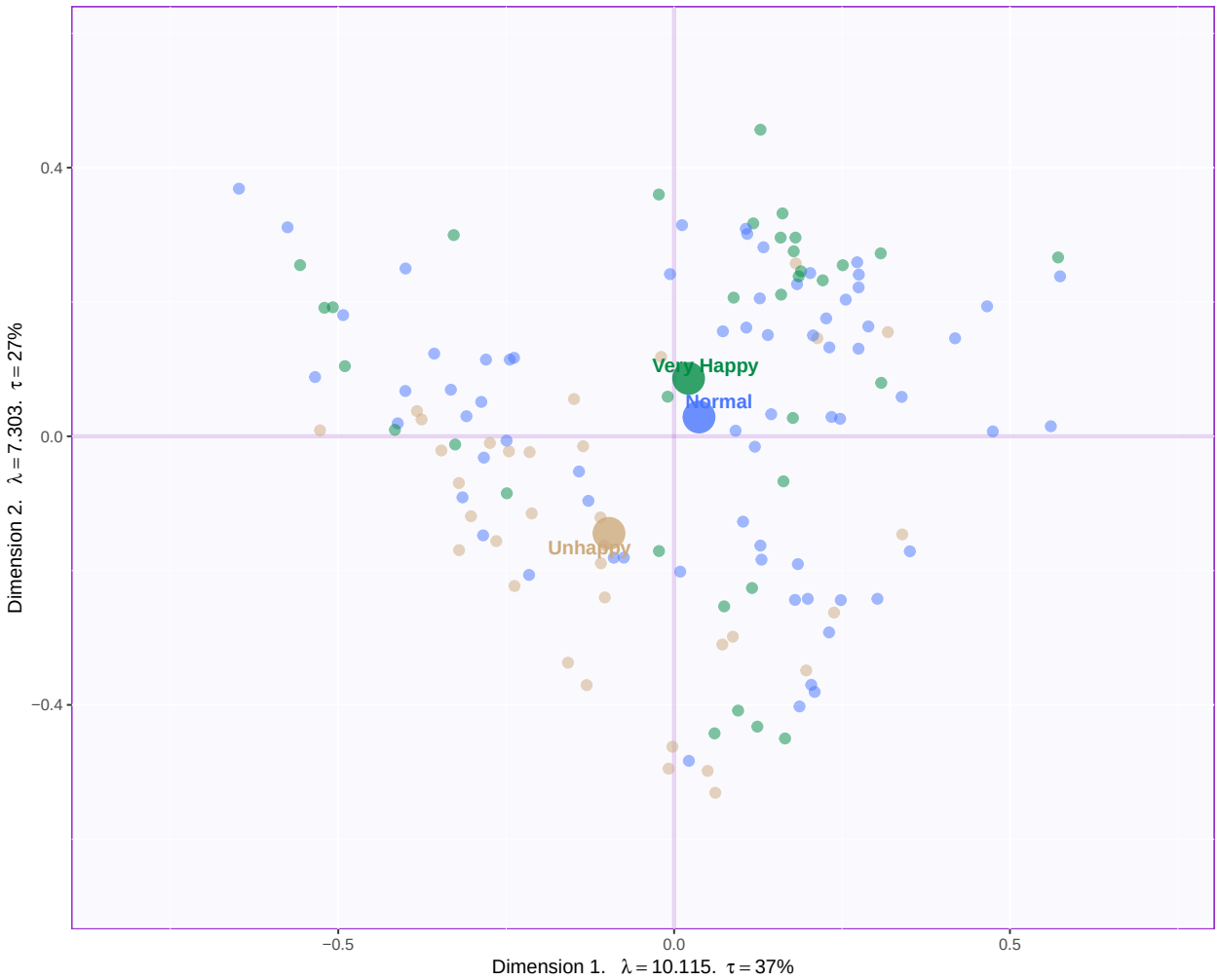
Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

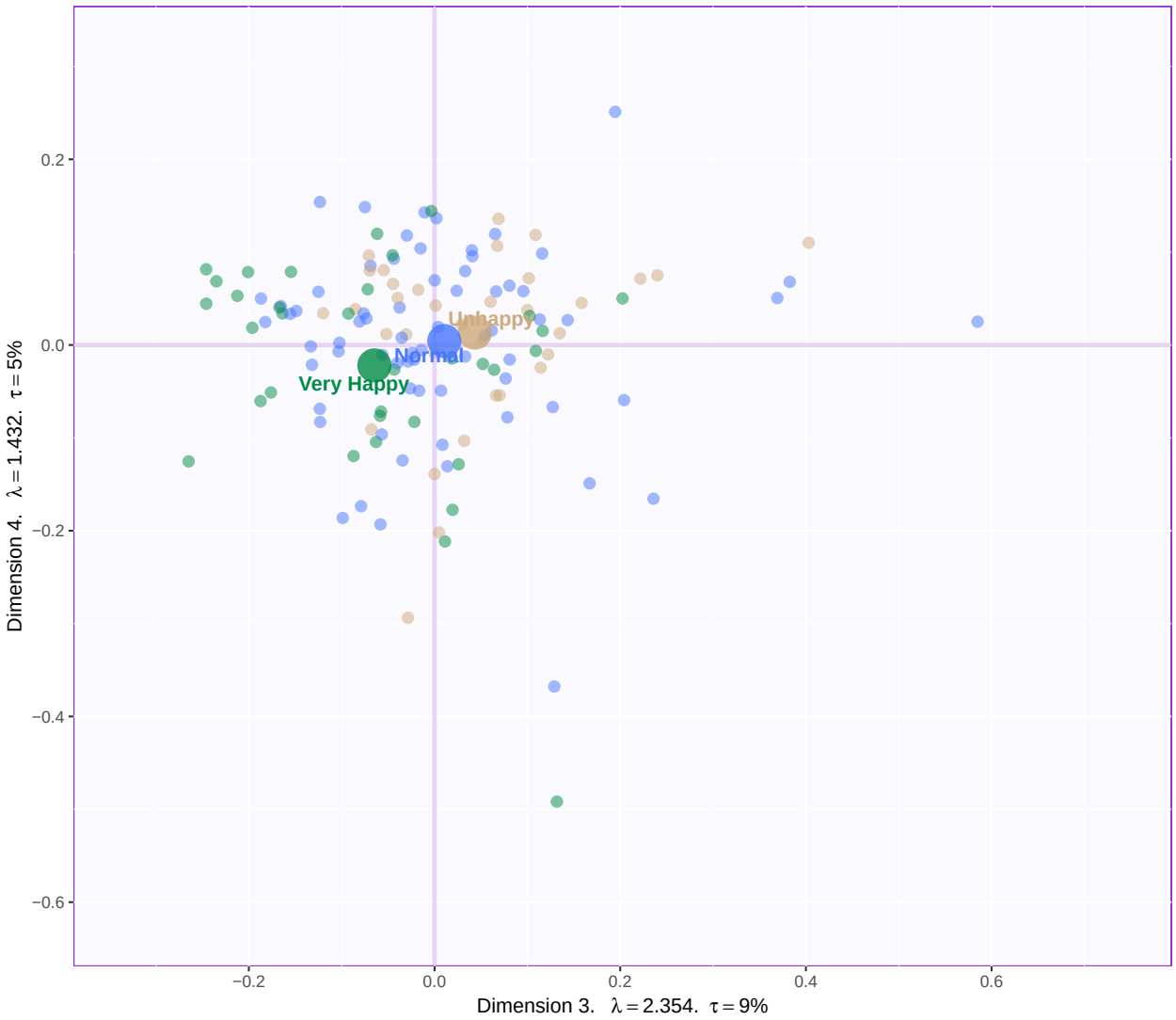
P.S. The most contribution component may not always be most useful for a given research question.

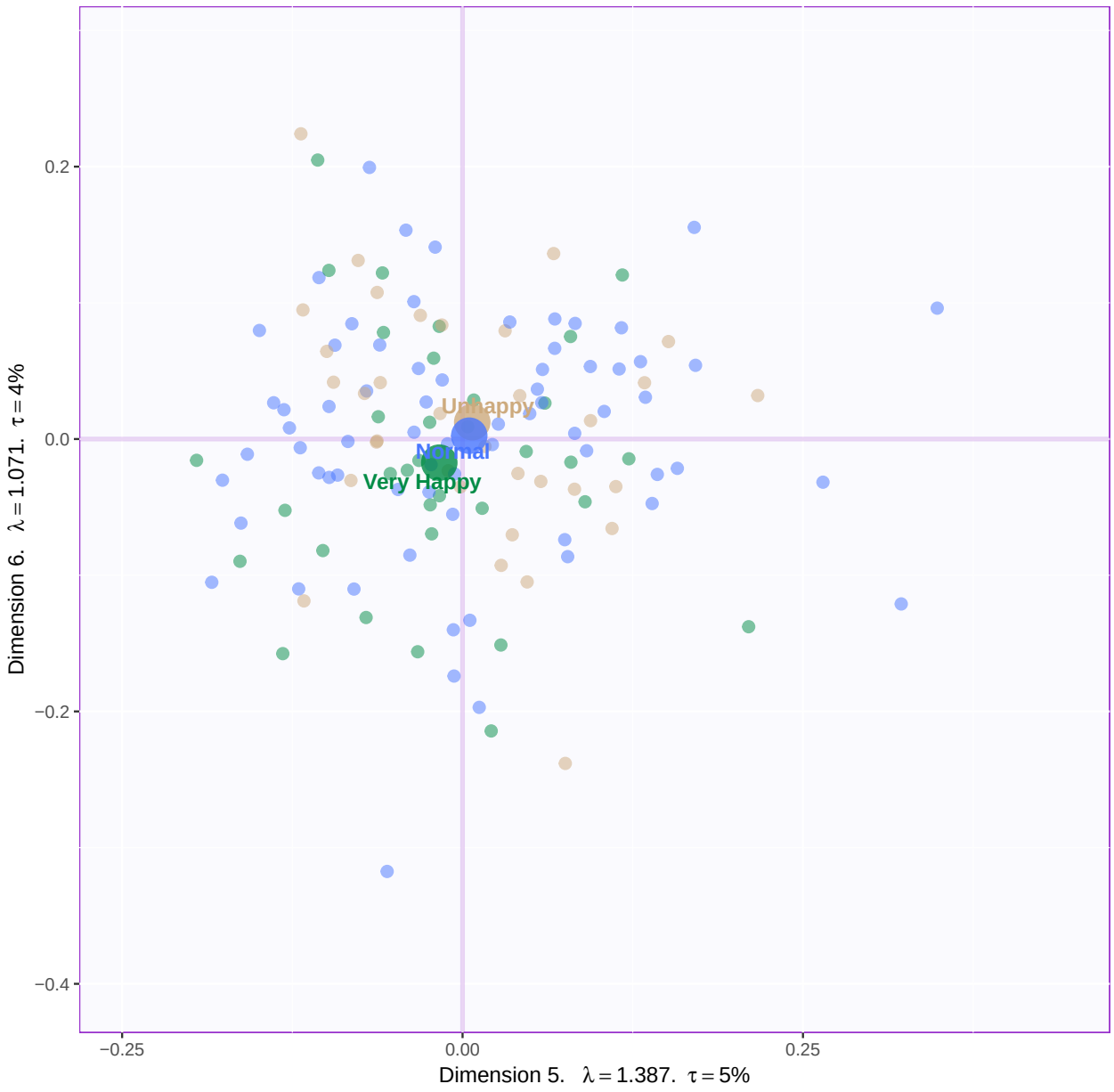


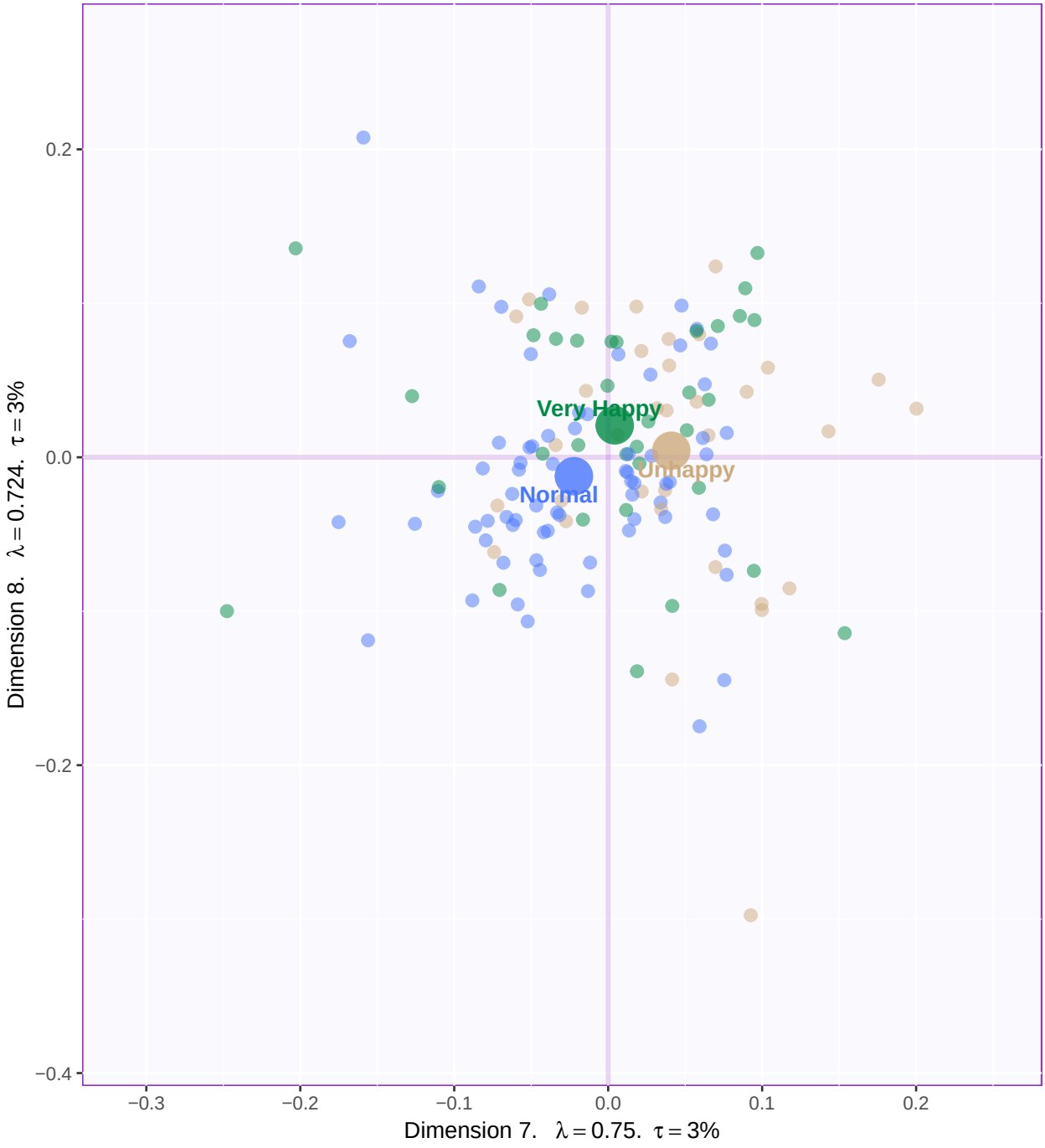

## 3.4 Factor Scores

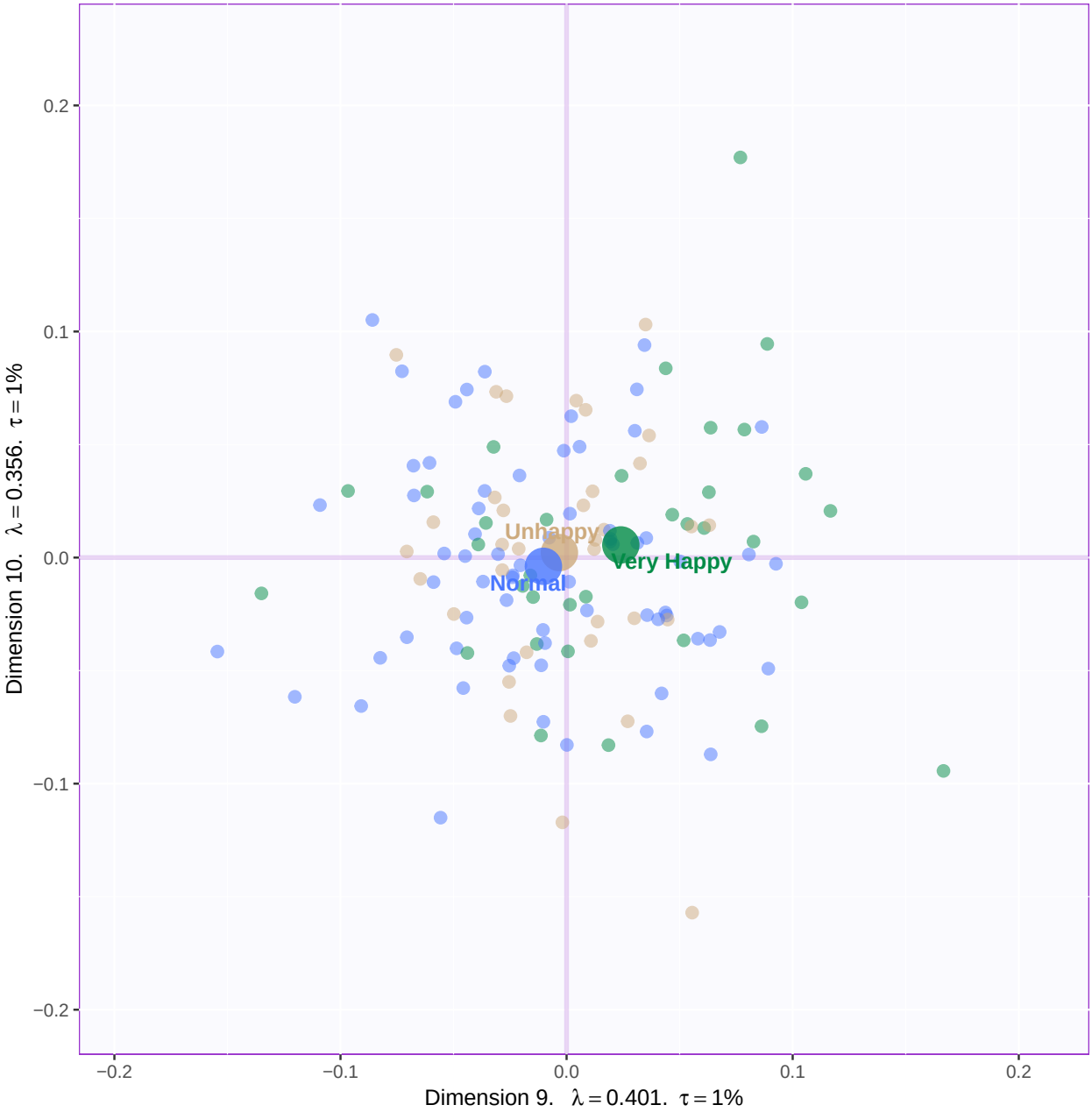
Lets visualize happiness categories for components 1-10, to make a decision (visually) on the most important components.


Very Happy
Normal
Unhappy
0.4
0.0
−0.4
−0.5
0.0
0.5
Dimension 2. λ = 7.303. τ = 27%
Dimension 1. λ = 10.115. τ = 37%


Unhappy
Normal
Very Happy
0.2
0.0
−0.2
−0.4
−0.6
−0.2
0.0
0.2
0.4
0.6
Dimension 4. λ = 1.432. τ = 5%
Dimension 3. λ = 2.354. τ = 9%


Unhappy
Normal
Very Happy
0.2
0.0
−0.2
−0.4
−0.25
0.00
0.25
Dimension 6. λ = 1.071. τ = 4%
Dimension 5. λ = 1.387. τ = 5%


Very Happy
Unhappy
Normal
0.2
0.0
−0.2
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
Dimension 8. λ = 0.724. τ = 3%
Dimension 7. λ = 0.75. τ = 3%




Since, it's not very straightforward to decide which components may be best suited for the research question at hand, let's represent, in a tabular format, which component helps to differentiate between which design variable values (Unhappy, Normal, Very Happy)

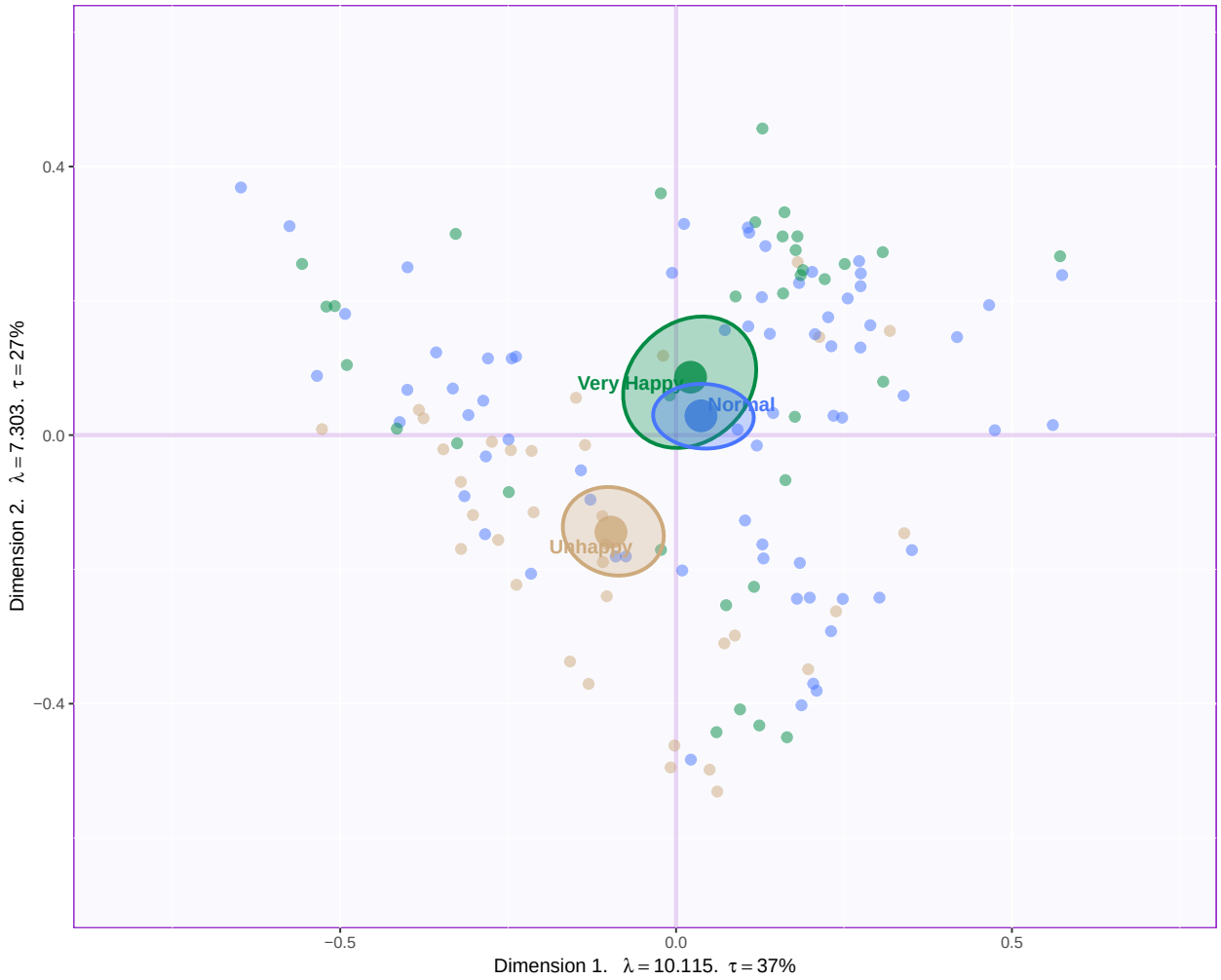
P.S. here -1 represents -ve quadrant of the component and +1 represent +ve quadrant. 0 represents that component was not decisive enough to clearly seperate happiness levels.

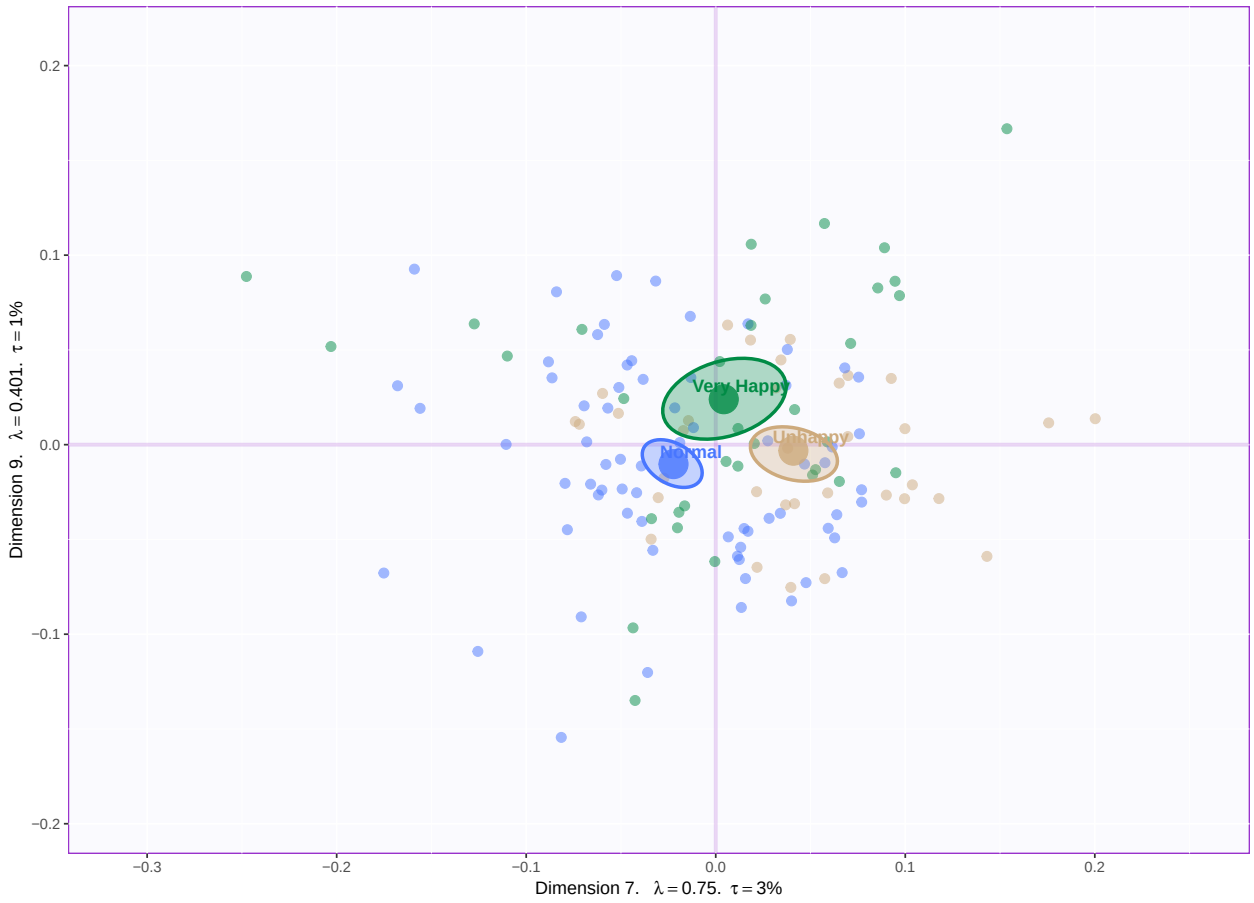
Looking at the table, it seems component 1, 2, 7, 9 may be able to best represent all 3 happiness levels. Although, SCREE Plot suggests that $3^{rd}$ and $4^{th}$ components might be useful, from our above analysis we know otherwise. Also, SCREE plot suggests that component $6^{th}$ and onwards might not be useful which is contradicting our findings above. Hence, let's plot components 1 vs 2 and 7 vs 9. Similarily, we will also plot Loading plots for these componenets.

- With Confidence Interval

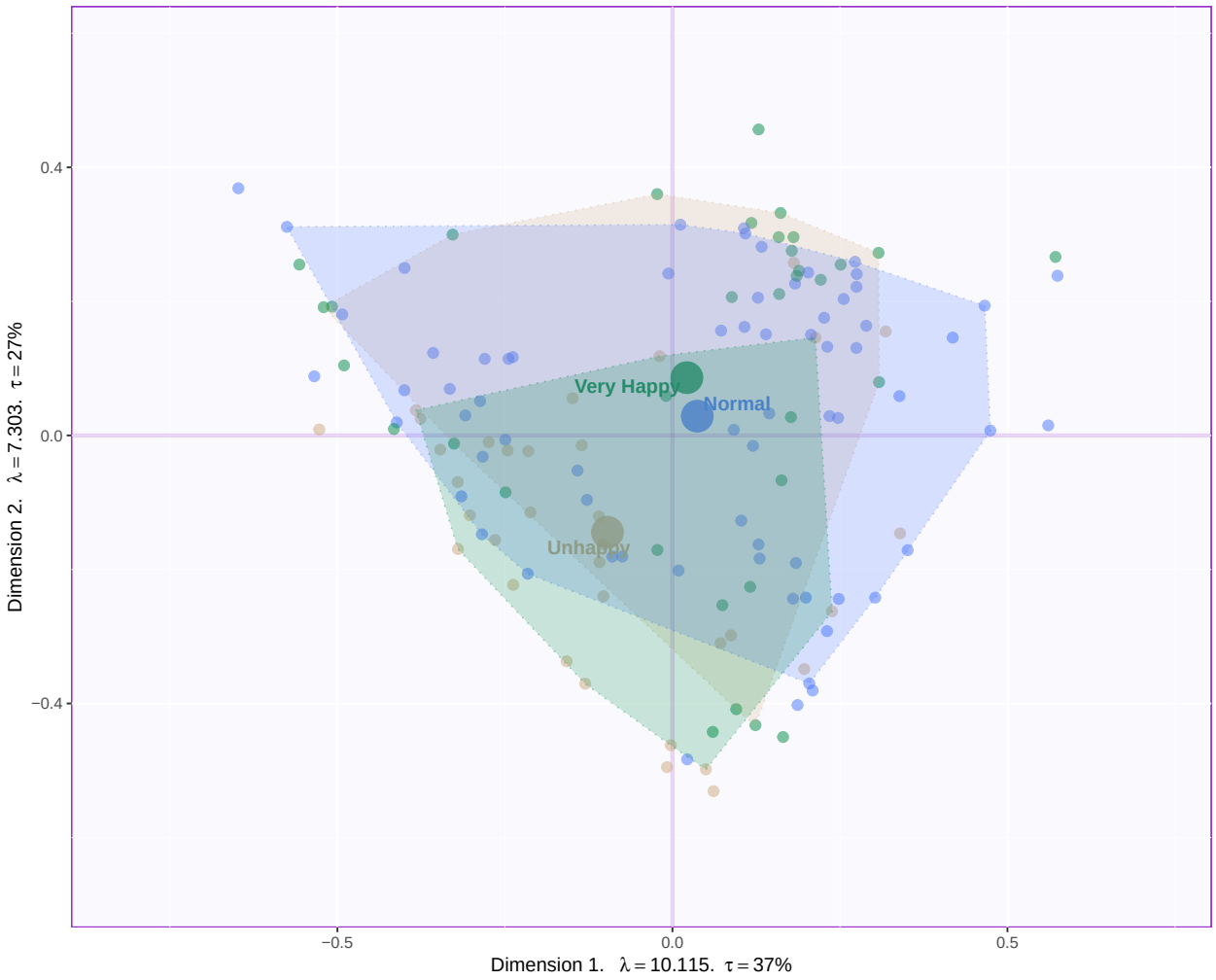
Table 3.1: Identify Components best describing happiness levels

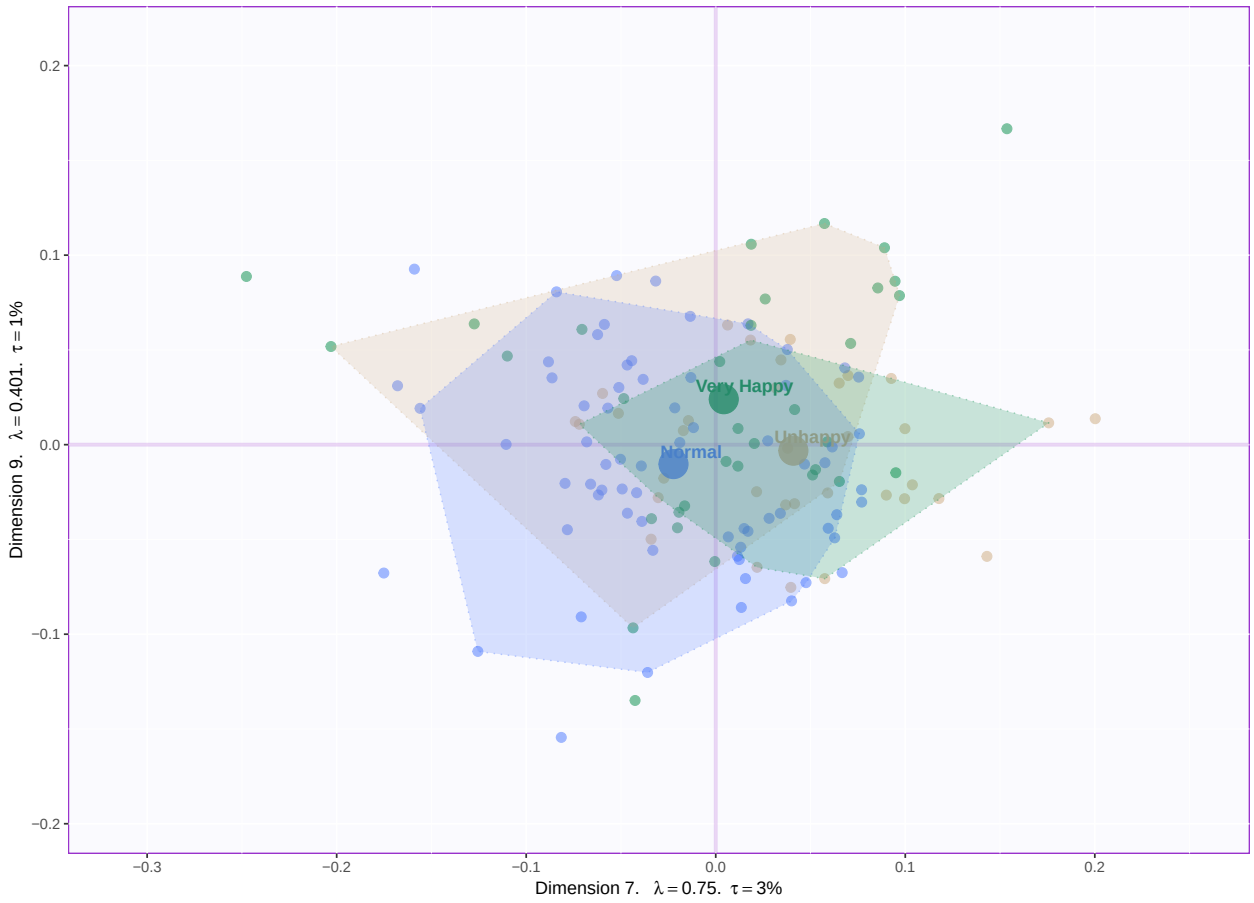
| | Unhappy | Normal | VeryHappy |
|---|---|---|---|
| Component 1 | -1 | 1 | 0 |
| Component 2 | -1 | 0 | 1 |
| Component 3 | 1 | 0 | -1 |
| Component 4 | 0 | 0 | 0 |
| Component 5 | 0 | 0 | 0 |
| Component 6 | 0 | 0 | 0 |
| Component 7 | 1 | -1 | 0 |
| Component 8 | 0 | 0 | 0 |
| Component 9 | 0 | -1 | 1 |
| Component 10 | 0 | 0 | 0 |



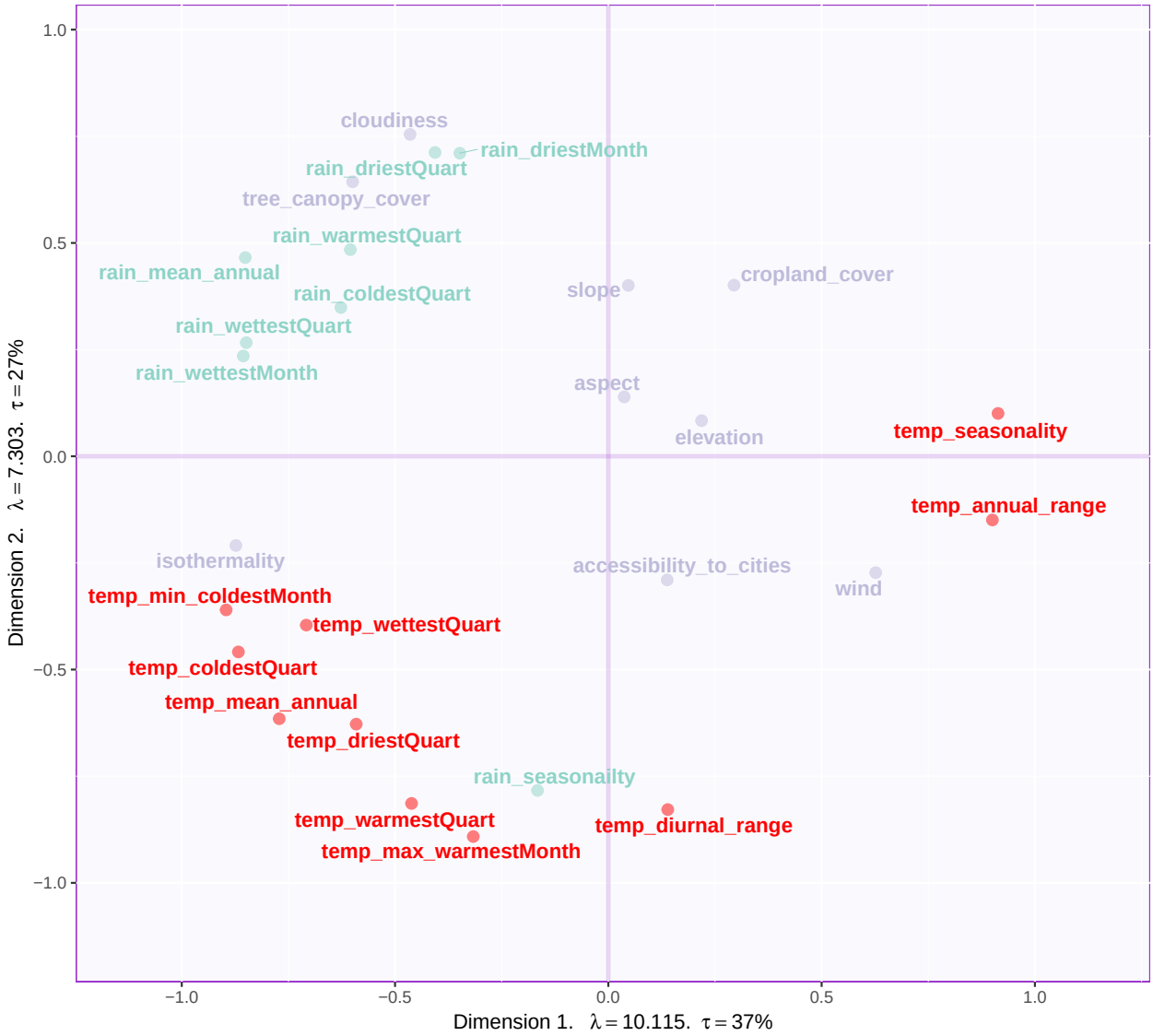


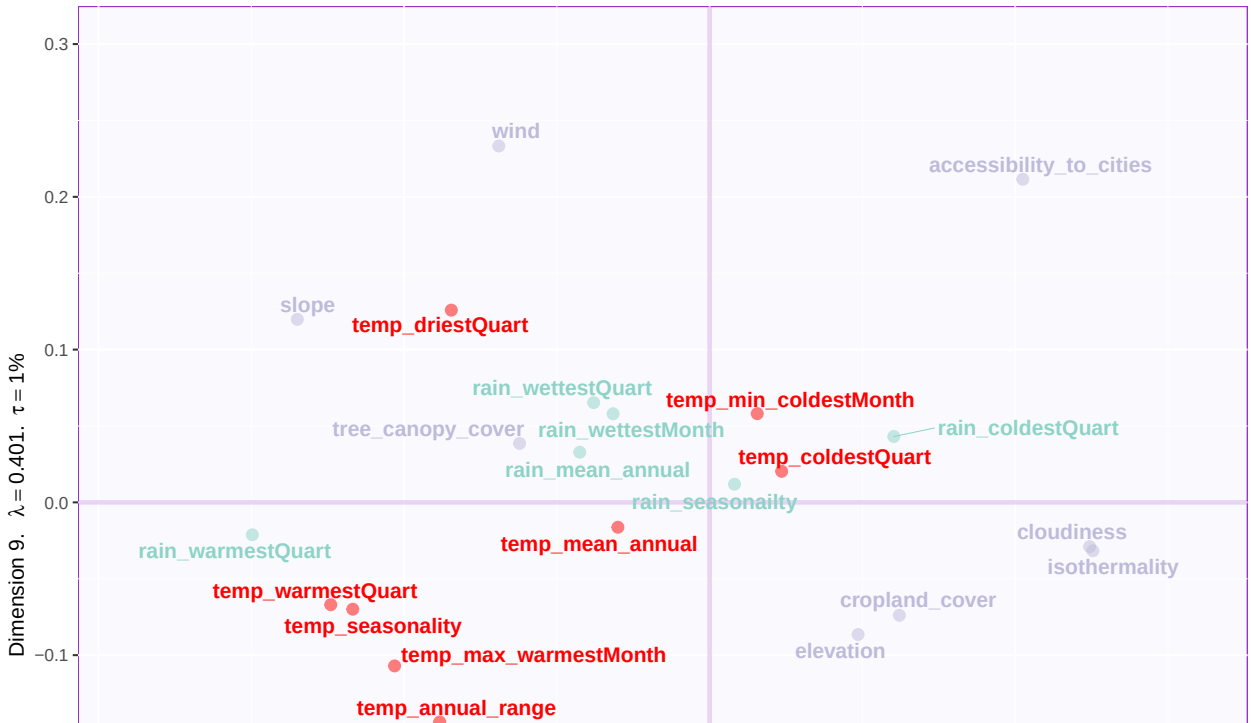

- With Tolerance Interval


Very Happy
Normal
Unhappy
0.4
0.0
−0.4
−0.5
0.0
0.5
Dimension 2. λ = 7.303. τ = 27%
Dimension 1. λ = 10.115. τ = 37%

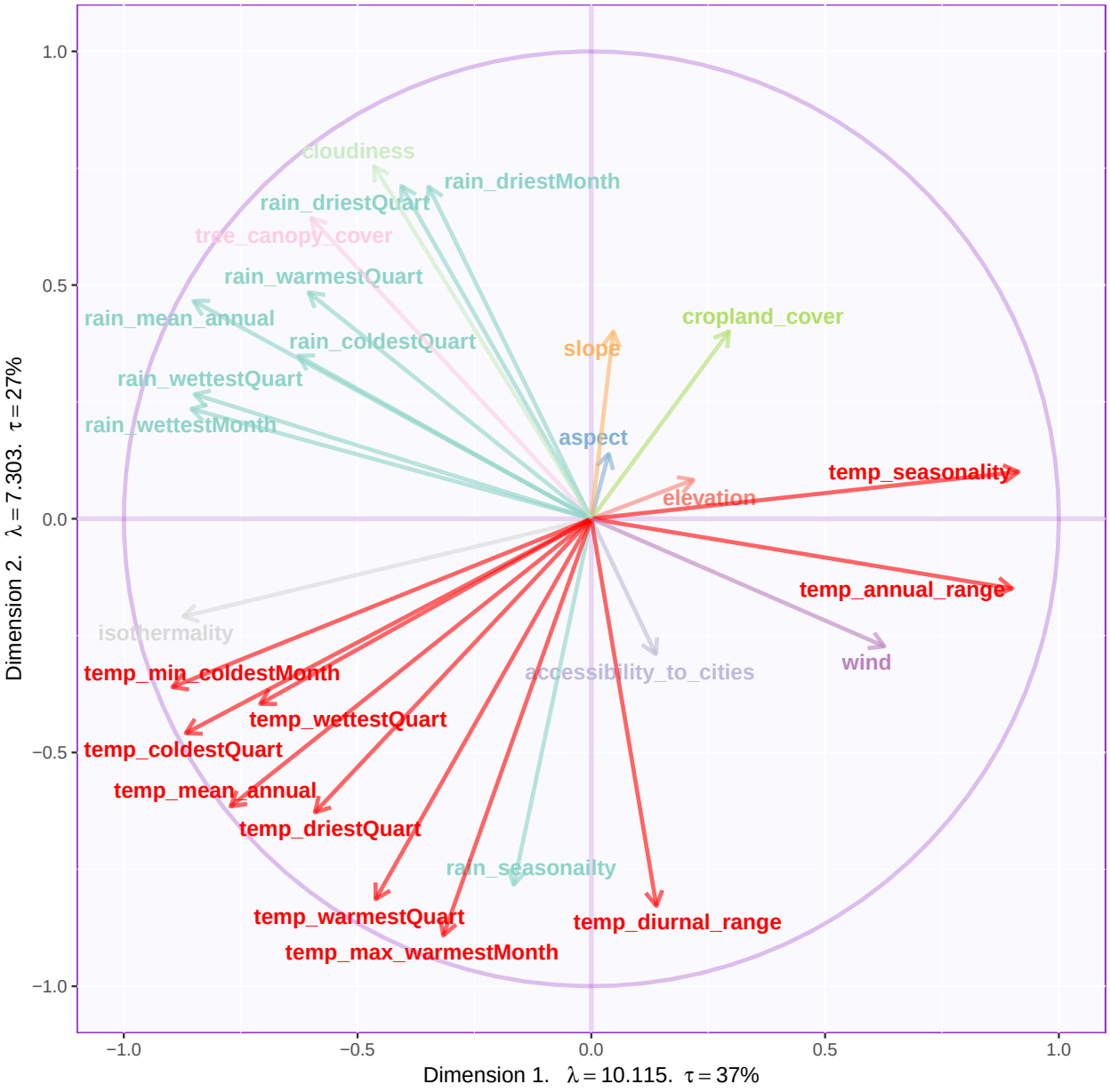

Very Happy
Unhappy
Normal
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
Dimension 9. λ = 0.401. τ = 1%
Dimension 7. λ = 0.75. τ = 3%

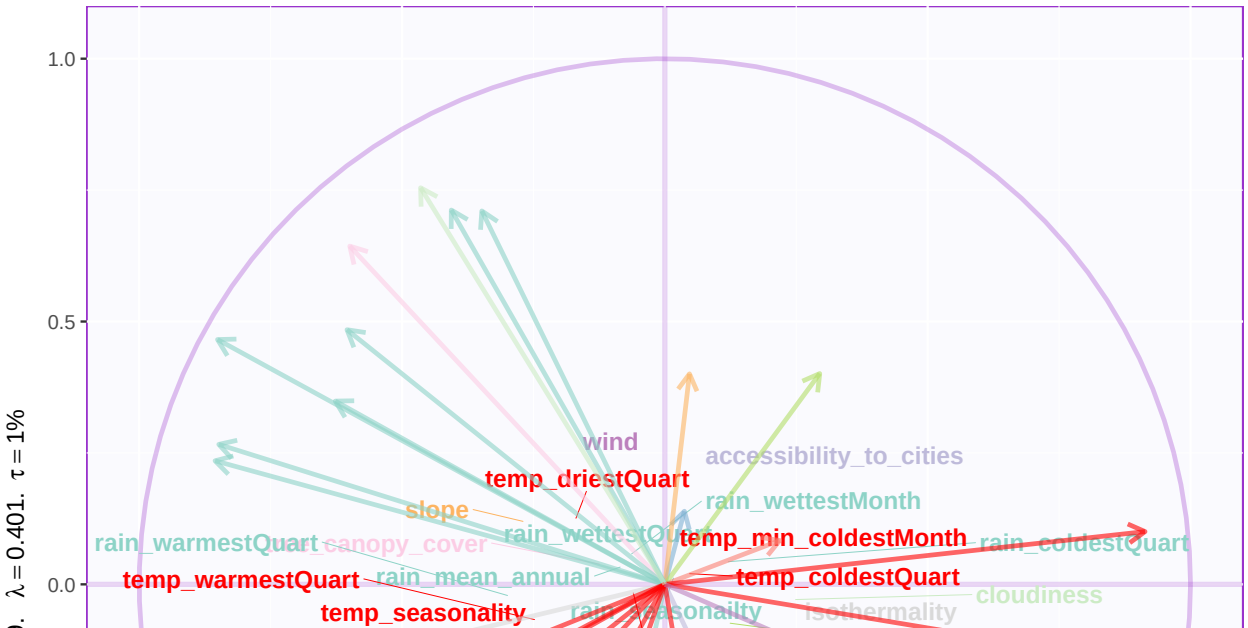
## 3.5 Loadings


cloudiness
rain_driestMonth
rain_driestQuart
tree_canopy_cover
rain_warmestQuart
rain_mean_annual
rain_coldestQuart
slope
cropland_cover
rain_wettestQuart
rain_wettestMonth
aspect
elevation
temp_seasonality
temp_annual_range
isothermality
accessibility_to_cities
wind
temp_min_coldestMonth
temp_wettestQuart
temp_coldestQuart
temp_mean_annual
temp_driestQuart
rain_seasonality
temp_warmestQuart
temp_diurnal_range
temp_max_warmestMonth
Dimension 2. λ = 7.303. τ = 27%
Dimension 1. λ = 10.115. τ = 37%


wind
accessibility_to_cities
slope
temp_driestQuart
rain_wettestQuart
temp_min_coldestMonth
tree_canopy_cover
rain_wettestMonth
rain_coldestQuart
temp_coldestQuart
rain_mean_annual
rain_seasonality
cloudiness
rain_warmestQuart
temp_mean_annual
isothermality
temp_warmestQuart
cropland_cover
temp_seasonality
elevation
temp_max_warmestMonth
temp_annual_range
Dimension 9. λ = 0.401. τ = 1%

## 3.6 Correlation Circle

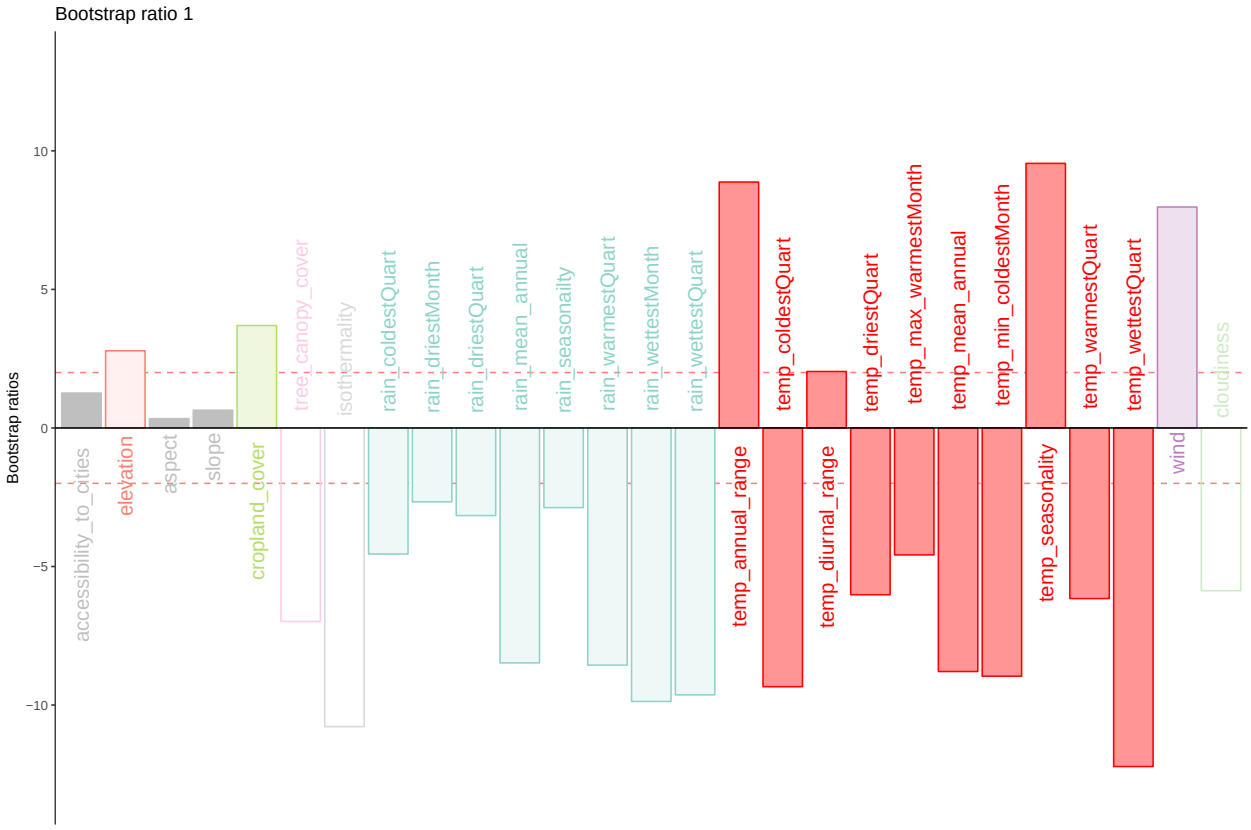


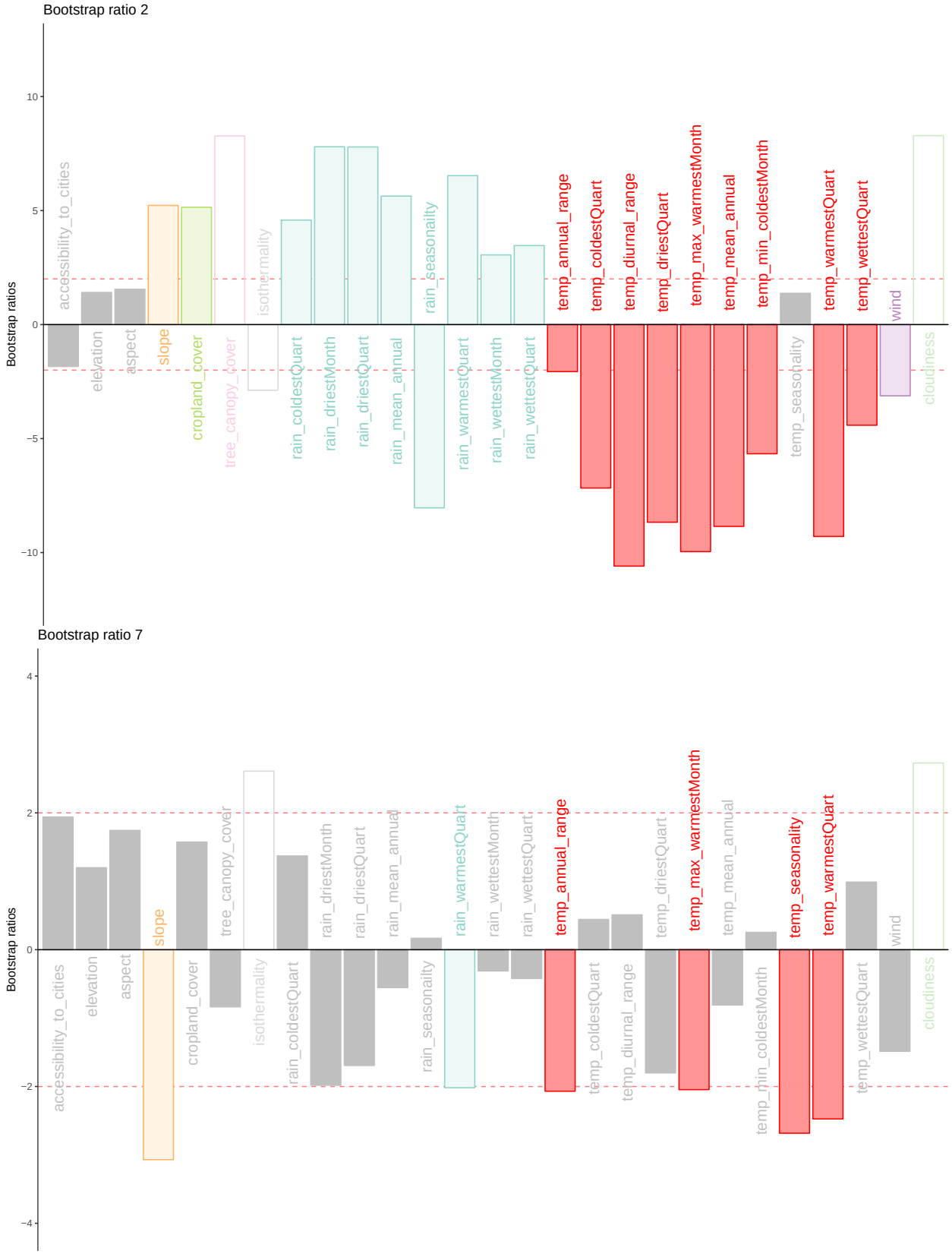



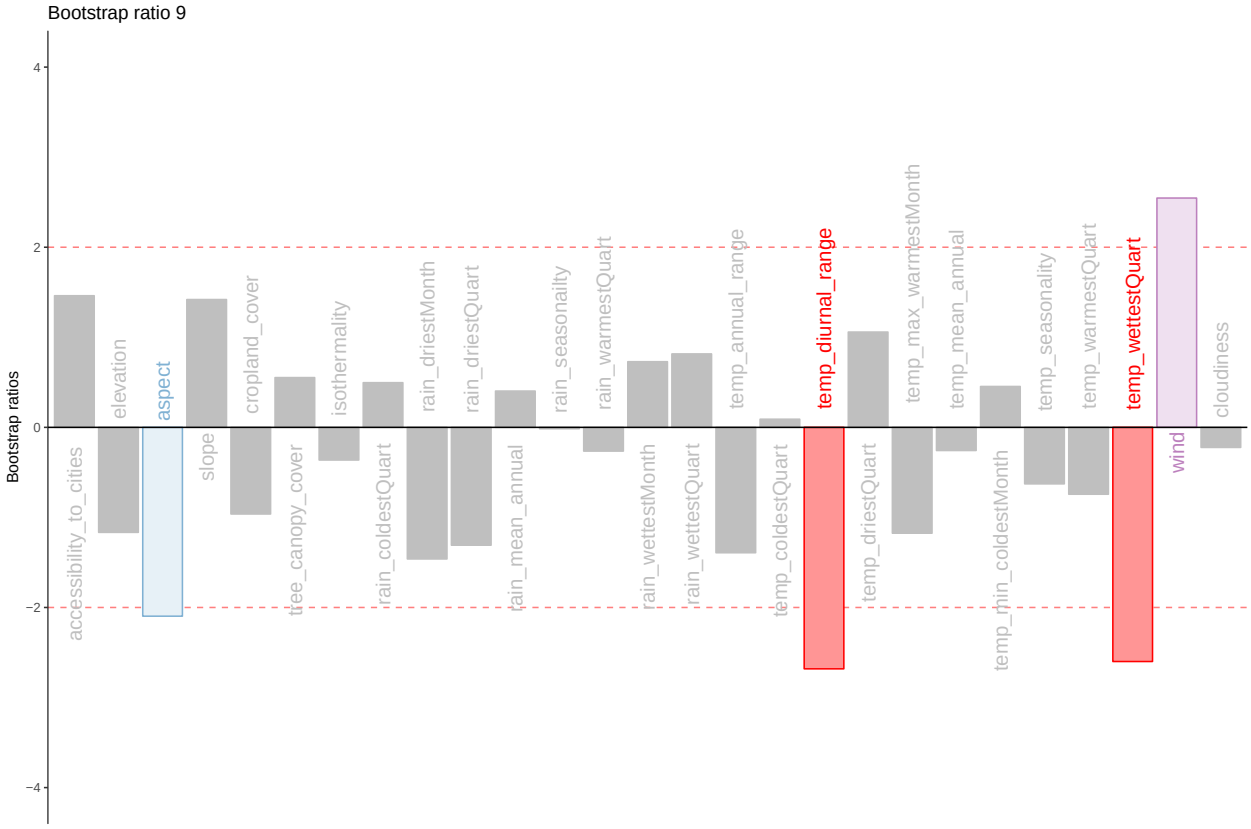
## 3.7 Most Contributing Variables

Let's plot variable contributions against each chosen components i.e. 1, 7, 9.

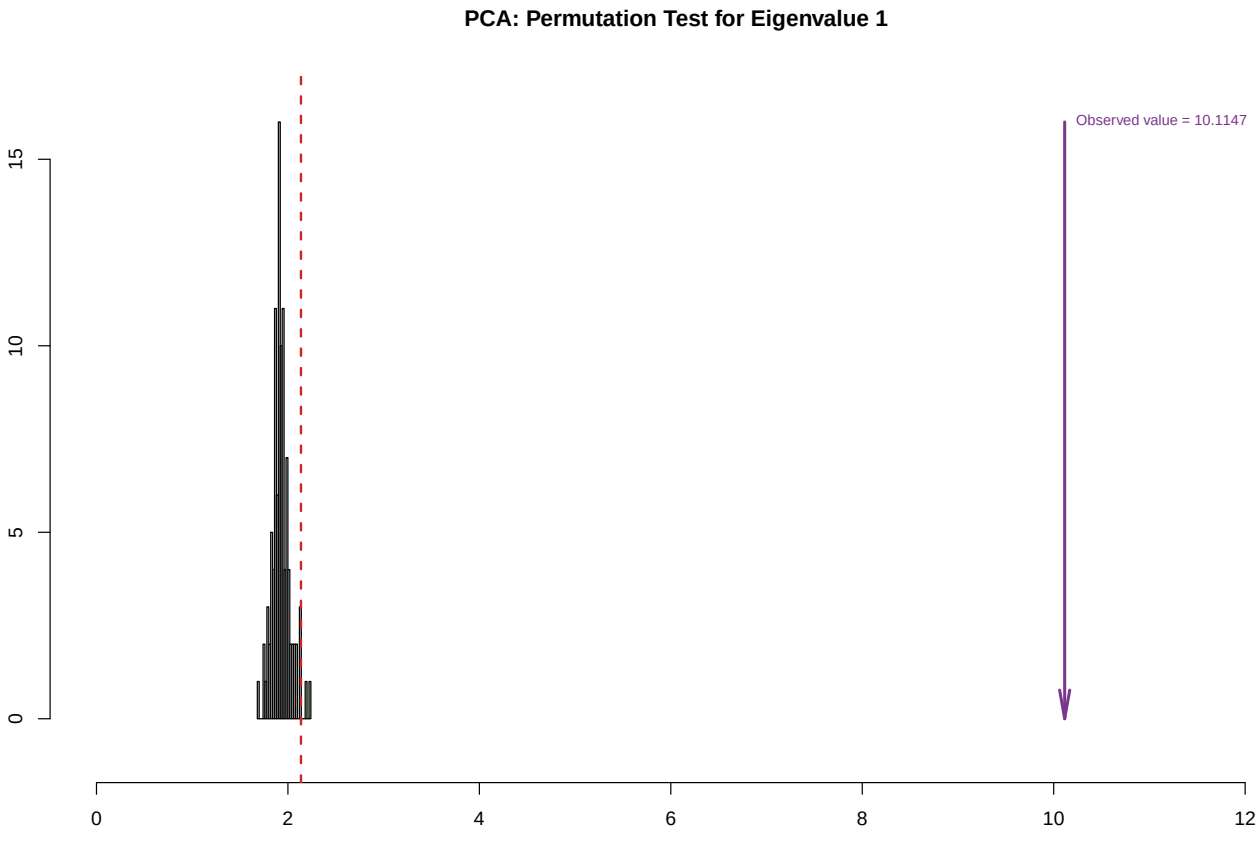
- With Bootstrap Ratio

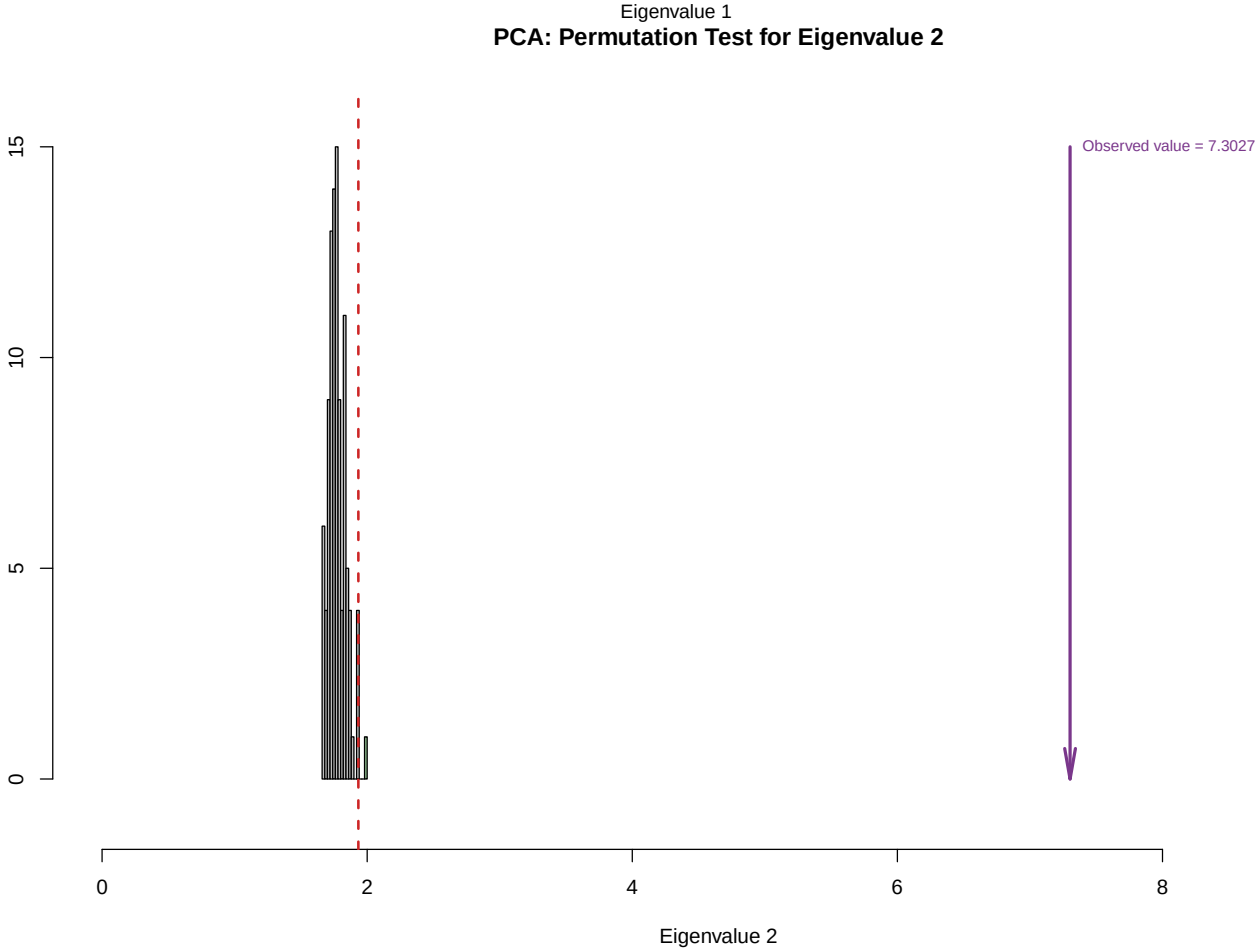


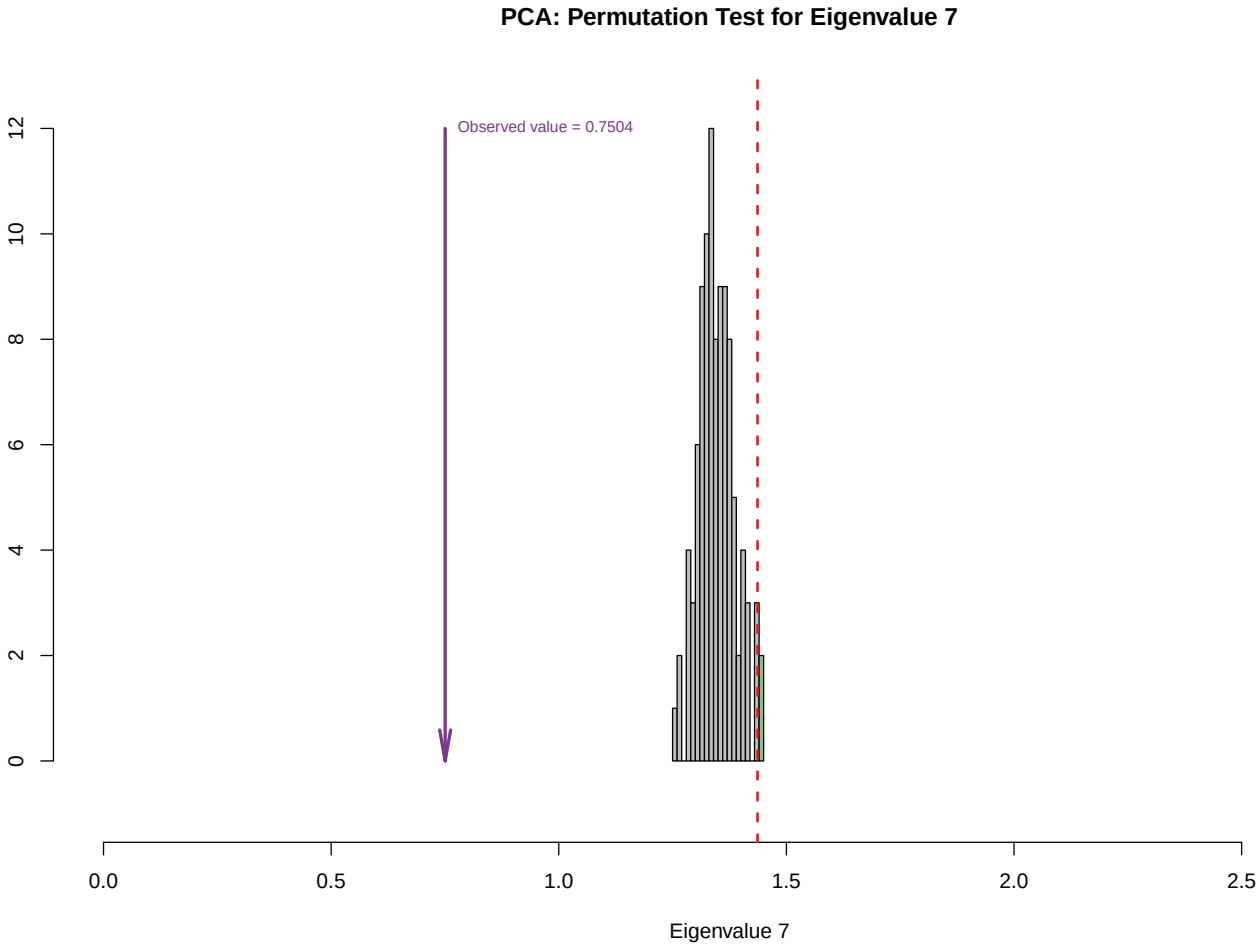

Bootstrap ratio 2
Bootstrap ratios
10
5
0
−5
−10
accessibility_to_cities
elevation
aspect
slope
cropland_cover
tree_canopy_cover
isothermality
rain_coldestQuart
rain_driestMonth
rain_driestQuart
rain_mean_annual
rain_seasonality
rain_warmestQuart
rain_wettestMonth
rain_wettestQuart
temp_annual_range
temp_coldestQuart
temp_diurnal_range
temp_driestQuart
temp_max_warmestMonth
temp_mean_annual
temp_min_coldestMonth
temp_seasonality
temp_warmestQuart
temp_wettestQuart
wind
cloudiness
Bootstrap ratio 7
Bootstrap ratios
4
2
0
−2
−4
accessibility_to_cities
elevation
aspect
slope
cropland_cover
tree_canopy_cover
isothermality
rain_coldestQuart
rain_driestMonth
rain_driestQuart
rain_mean_annual
rain_seasonality
rain_warmestQuart
rain_wettestMonth
rain_wettestQuart
temp_annual_range
temp_coldestQuart
temp_diurnal_range
temp_driestQuart
temp_max_warmestMonth
temp_mean_annual
temp_min_coldestMonth
temp_seasonality
temp_warmestQuart
temp_wettestQuart
wind
cloudiness

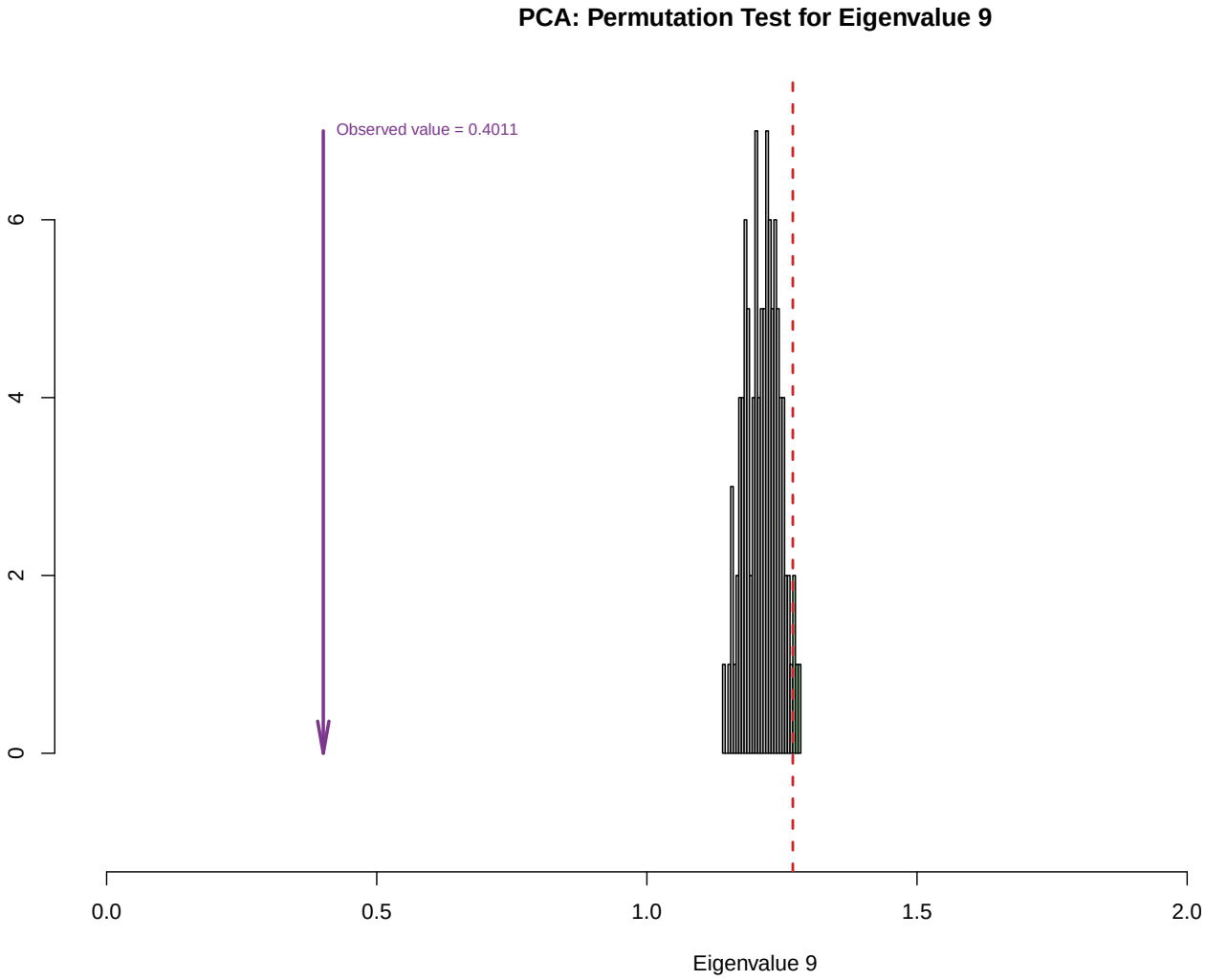

Bootstrap ratio 9
Bootstrap ratios
4
2
0
−2
−4
accessibility_to_cities
elevation
aspect
slope
cropland_cover
tree_canopy_cover
isothermality
rain_coldestQuart
rain_driestMonth
rain_driestQuart
rain_mean_annual
rain_seasonality
rain_warmestQuart
rain_wettestMonth
rain_wettestQuart
temp_annual_range
temp_coldestQuart
temp_diurnal_range
temp_driestQuart
temp_max_warmestMonth
temp_mean_annual
temp_min_coldestMonth
temp_seasonality
temp_warmestQuart
temp_wettestQuart
wind
cloudiness

## 3.8 Permutation Test

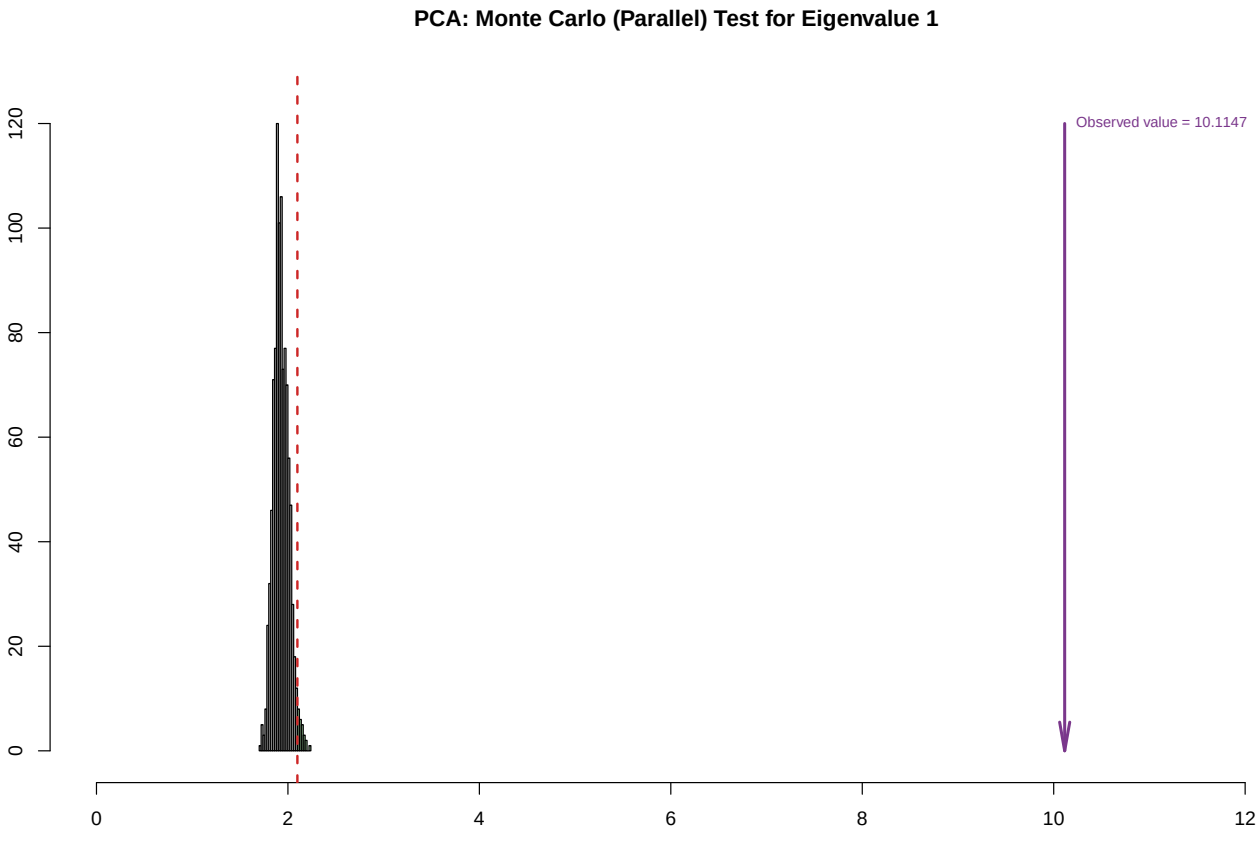


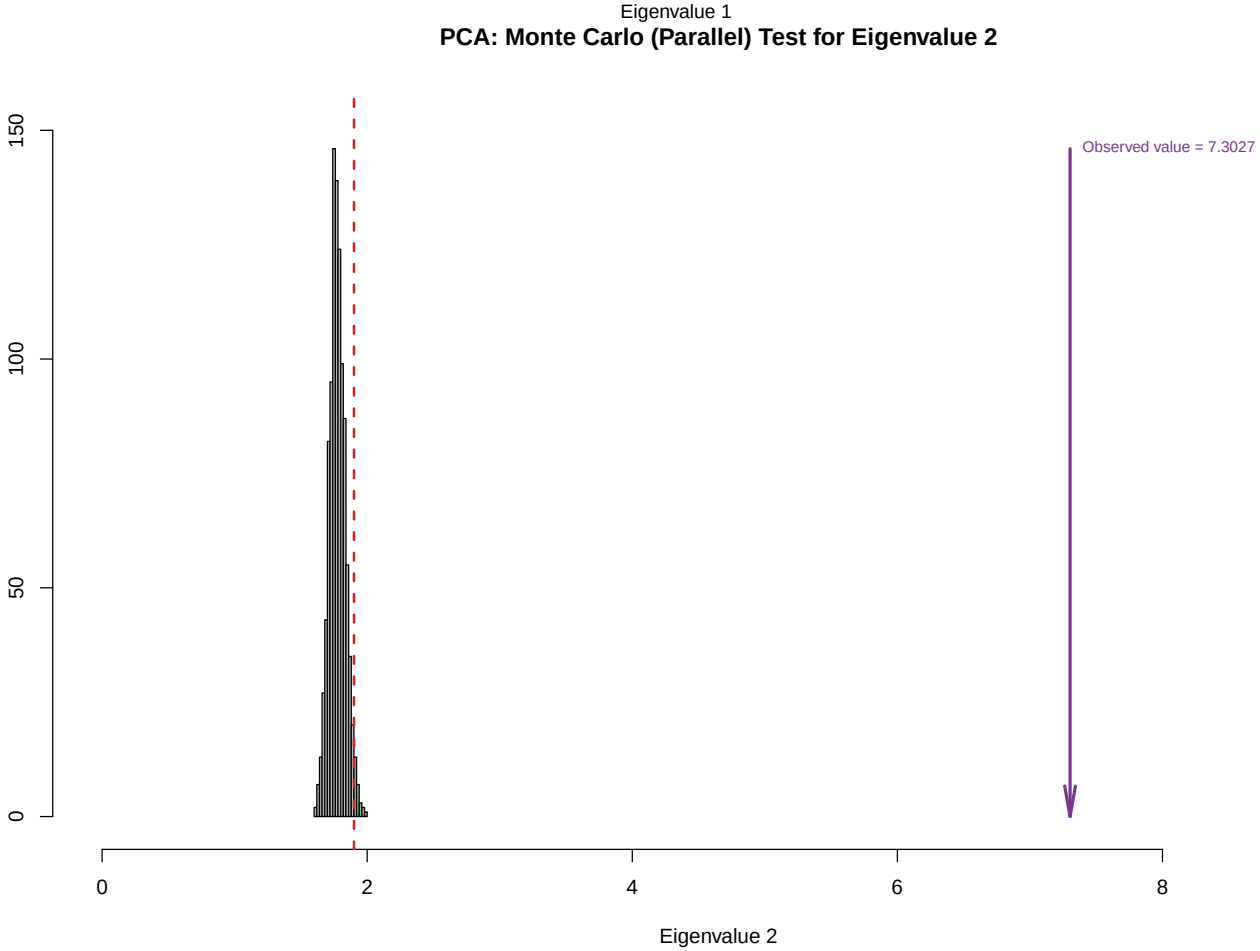



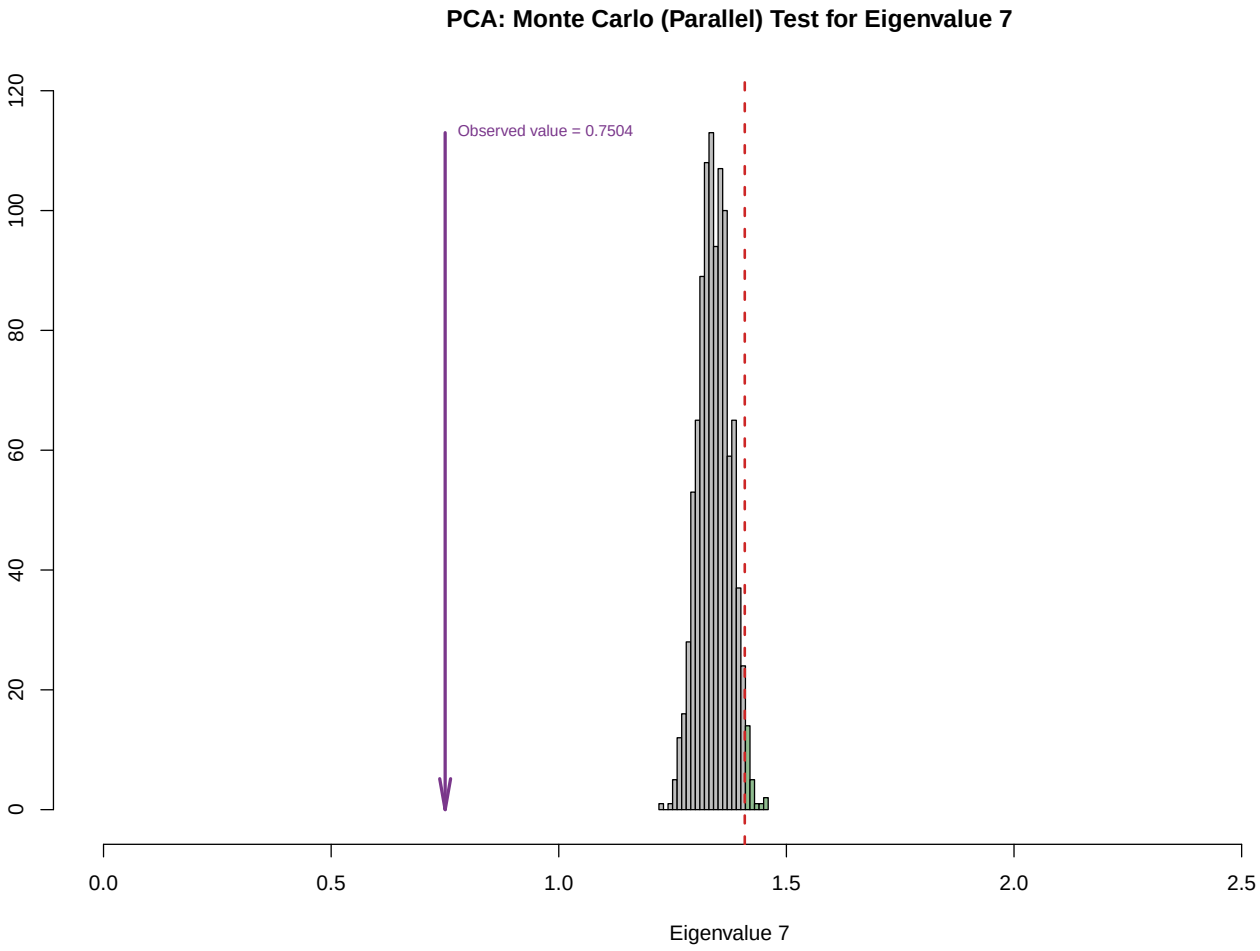

PCA: Permutation Test for Eigenvalue 7
Observed value = 0.7504
0
2
4
6
8
10
12
0.0
0.5
1.0
1.5
2.0
2.5
Eigenvalue 7

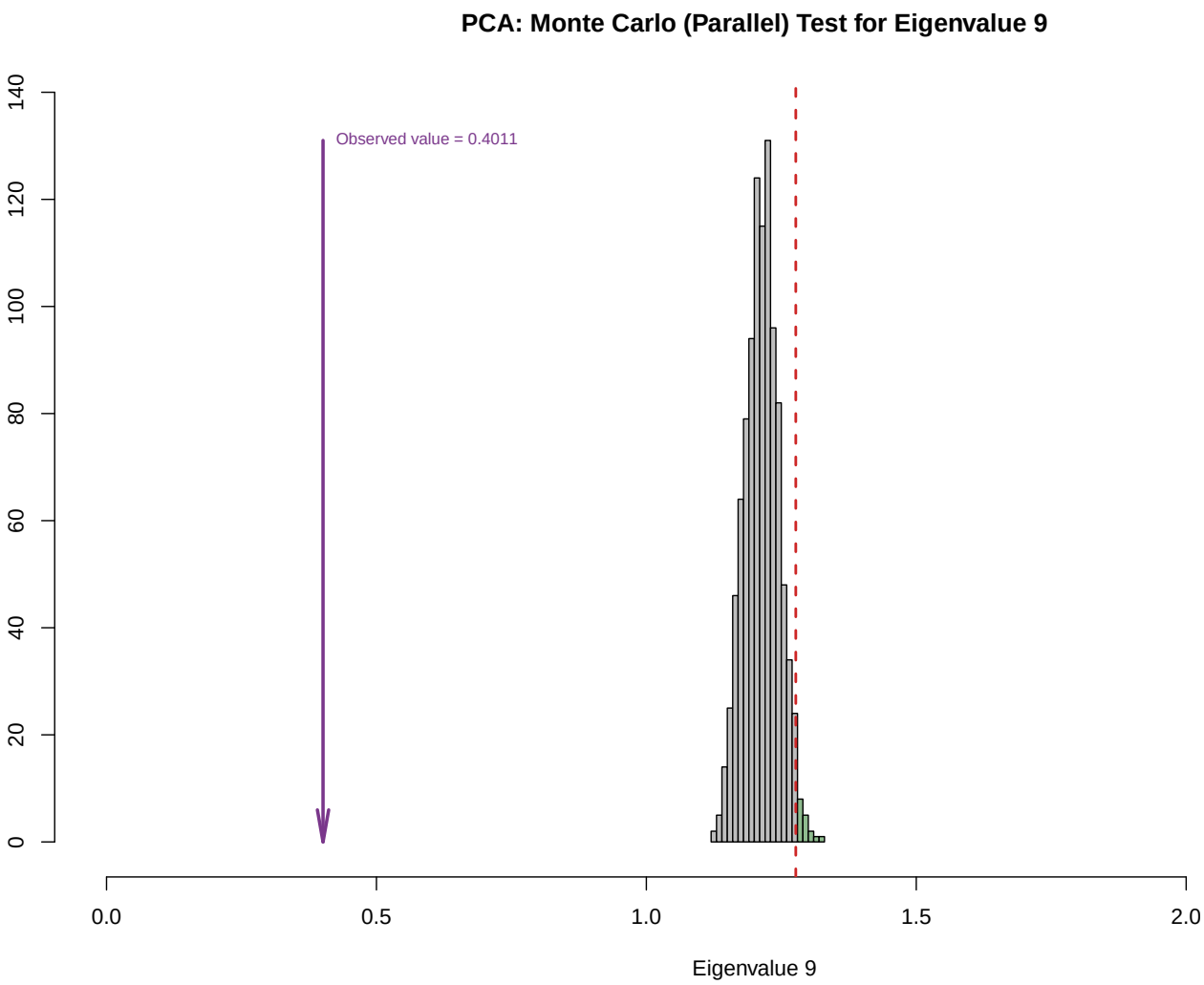

PCA: Permutation Test for Eigenvalue 9
Observed value = 0.4011
0
2
4
6
0.0
0.5
1.0
1.5
2.0
Eigenvalue 9

## 3.9 Parallet Test

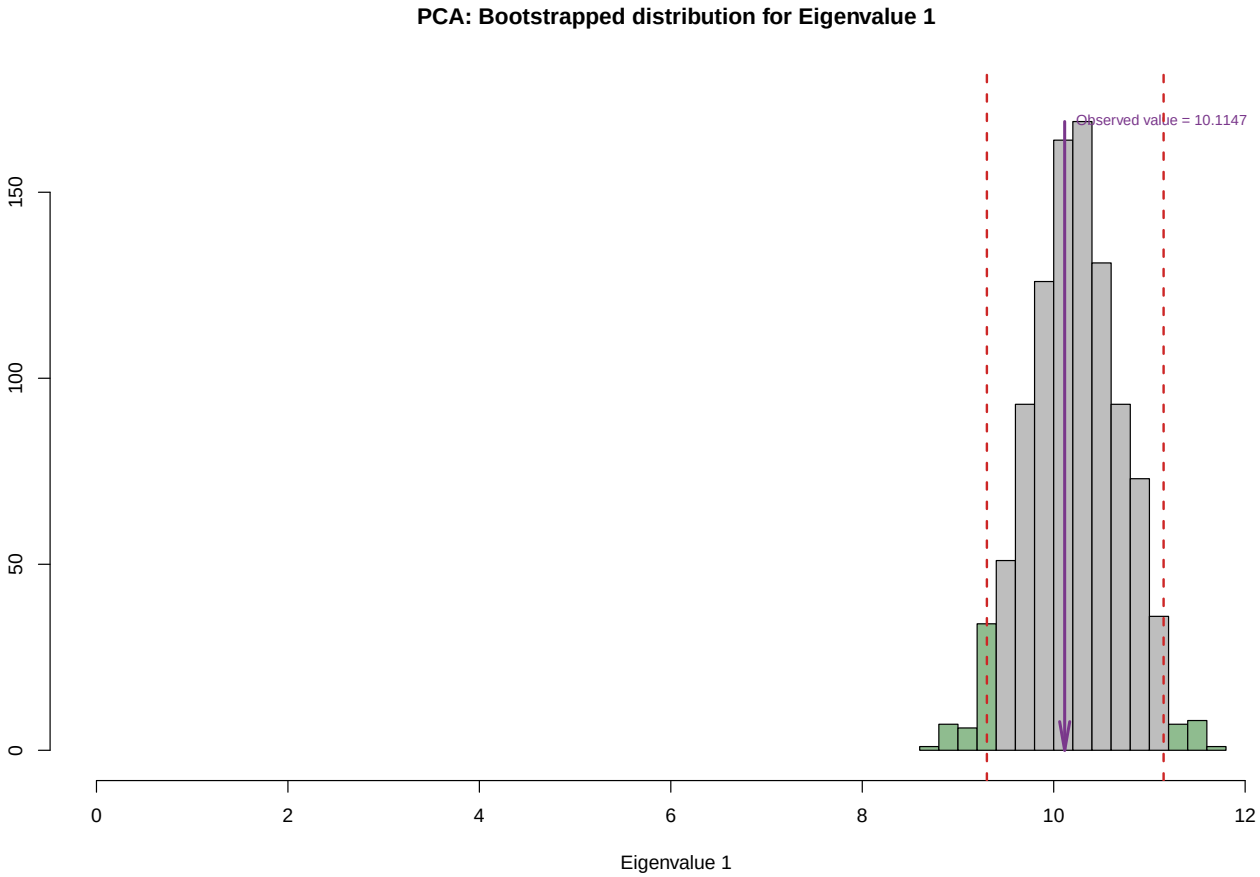


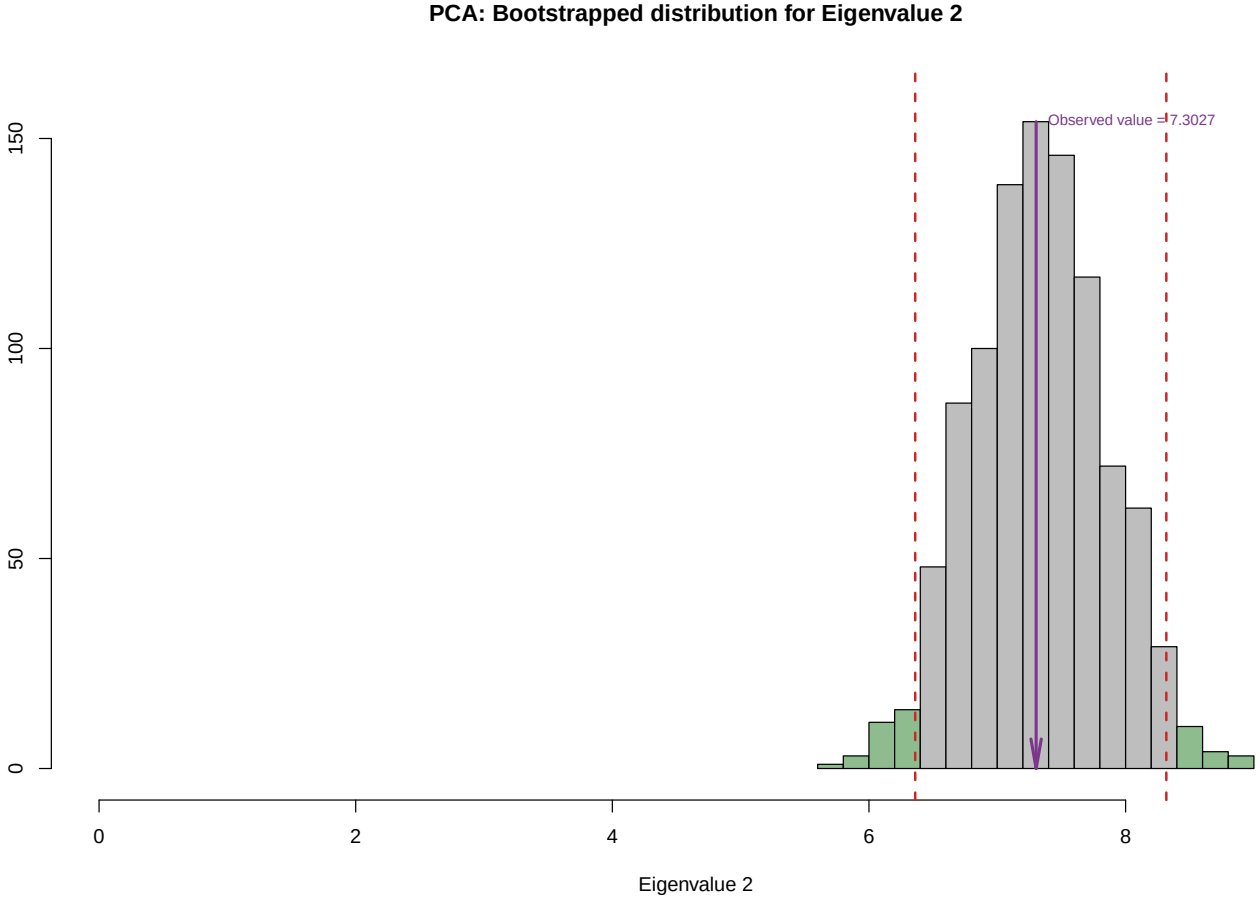



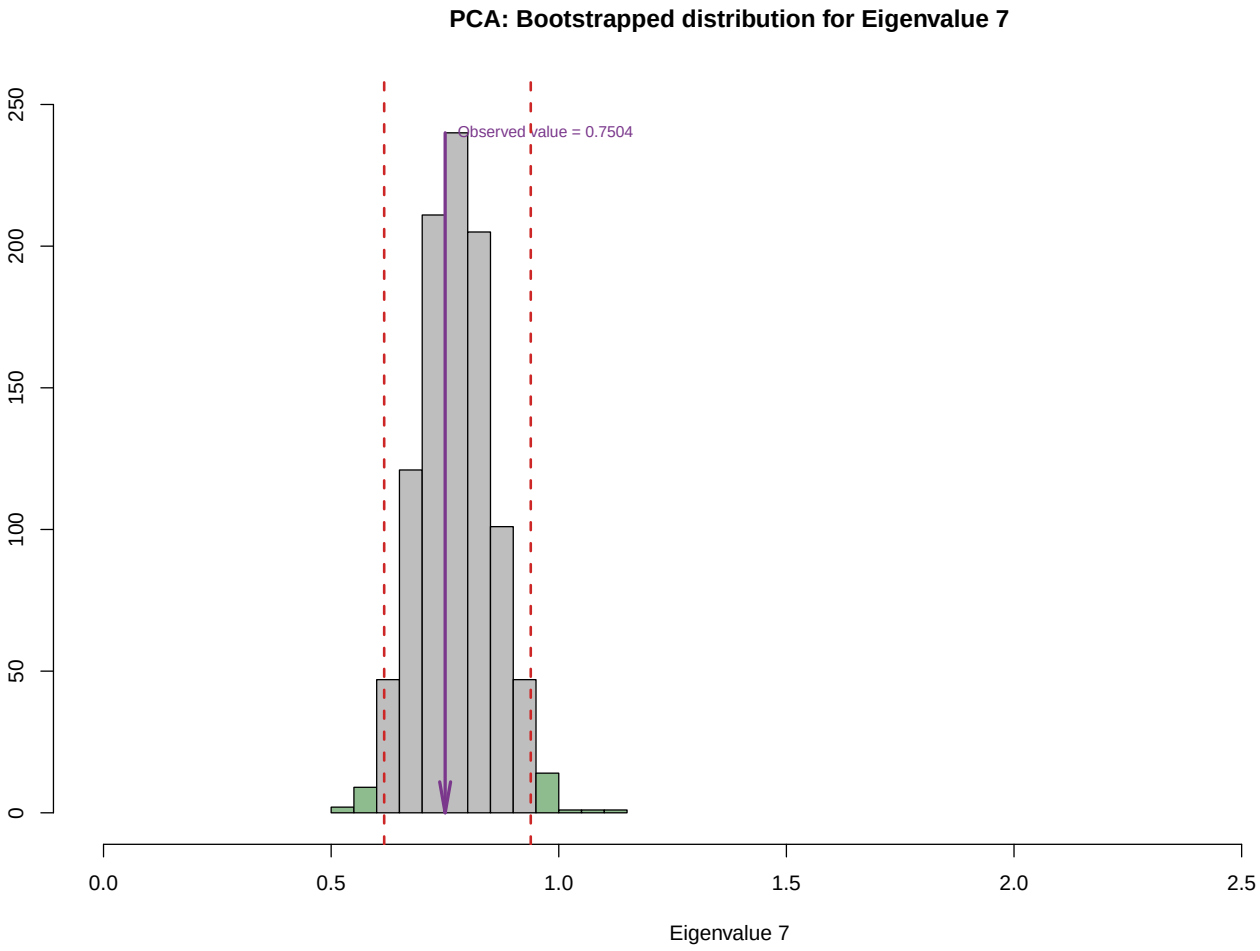

PCA: Monte Carlo (Parallel) Test for Eigenvalue 7
Observed value = 0.7504
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
Eigenvalue 7

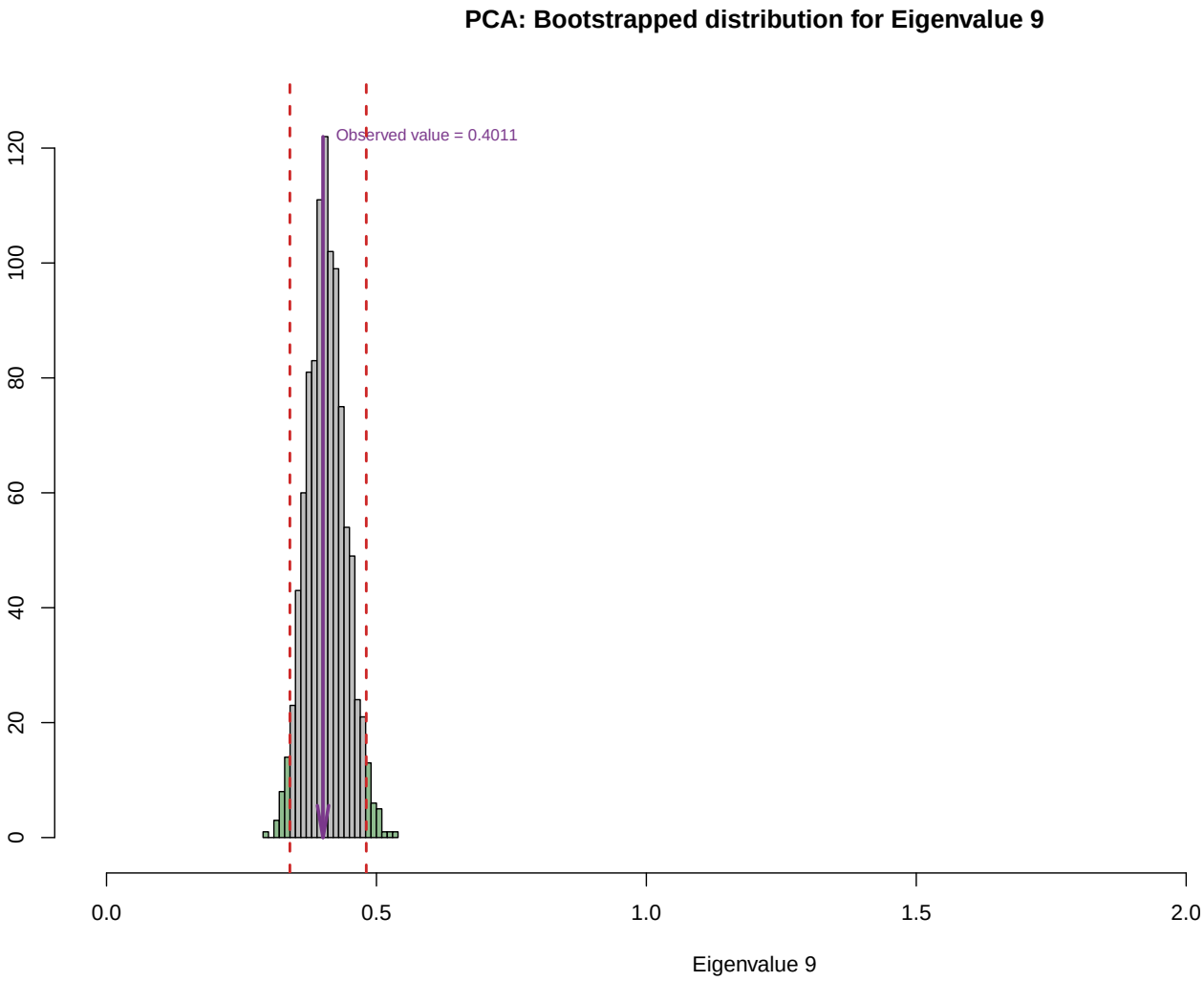

PCA: Monte Carlo (Parallel) Test for Eigenvalue 9
Observed value = 0.4011
0
20
40
60
80
100
120
140
0.0
0.5
1.0
1.5
2.0
Eigenvalue 9

## 3.10 Bootstrap Test




PCA: Bootstrapped distribution for Eigenvalue 7
Observed value = 0.7504
0
50
100
150
200
250
0.0
0.5
1.0
1.5
2.0
2.5
Eigenvalue 7




## 3.11 Conclusion

- Component 1:
  - Rows: Normal & Happy
  - Columns: Cloudiness & Rain vs Cropland, Aspect, Elevation
  - Interpret: People in countries with more Cloudiness, Trees and Rain tends to be happier.
- Component 7:
  - Rows: Happy & Unhappy
  - Columns: Temp and Rain vs Accessibility and Cropland
  - Interpret: Rain and Temp seems to be main reason for unhappiness and Cropland is important for Happiness.
- Component 9:
  - Rows: Happy & Very Happy
  - Columns: Temp vs Rain
  - Interpret: Rain and Temp seems to be main reason for Happiness. *This contradicts with Component 7 and 1.*

# Chapter 4

# Multiple Component Analysis

## 4.1 Description

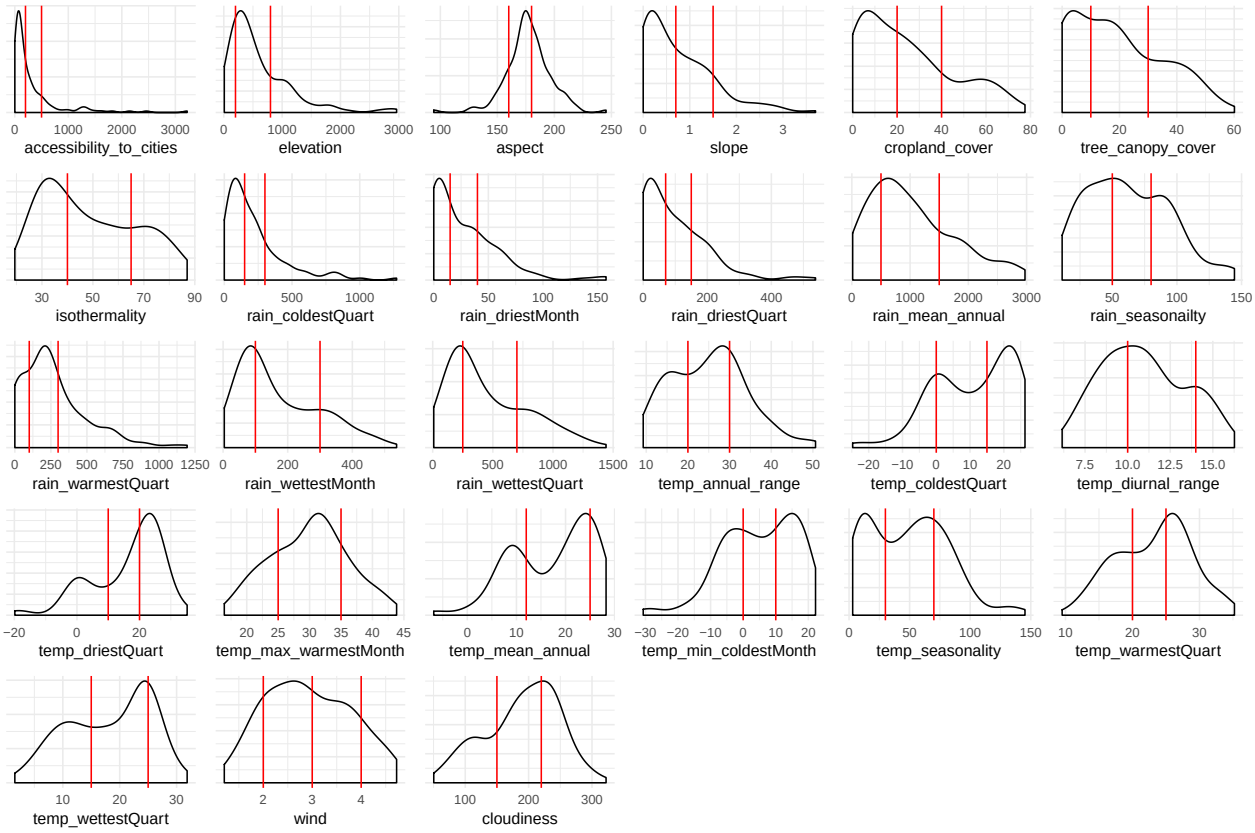
Multiple correspondence analysis (MCA) is an extension of correspondence analysis(CA) which allows one to analyze the pattern of relationships of several categorical dependent variables. As such, it can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative. Because MCA has been (re)discovered many times, equivalent methods are known under several different names such as optimal scaling, optimal or appropriate scoring, dual scaling, homogeneity analysis,scalogram analysis, and quantification method.

**Interpreting MCA** Multiple correspondence analysis locates all the categories in a Euclidean space.

- The first two dimensions of this space are plotted to examine the associations among the categories.
- The top-right quadrant of the plot shows the categories.
- The bottom-left quadrant shows the association.
- This interpretation is based on points found in approximately the same direction from the origin and in approximately the same region of the space. Distances between points do not have a straightforward interpretation.

## 4.2 Density Plot

Let's observe the distribution of each variables to get an intuition of how we can bin these variables. It's important to have nearly equal number of observations in the each bin and to try to cut the variables in a way to so that each new binned distribution is nearly Gaussian. We can also verify that our binning is appropiate by calculating Spearman Correlation for each of original variable and binned variable, the correlation coefficient should be close to 1.




## 4.3 Binning

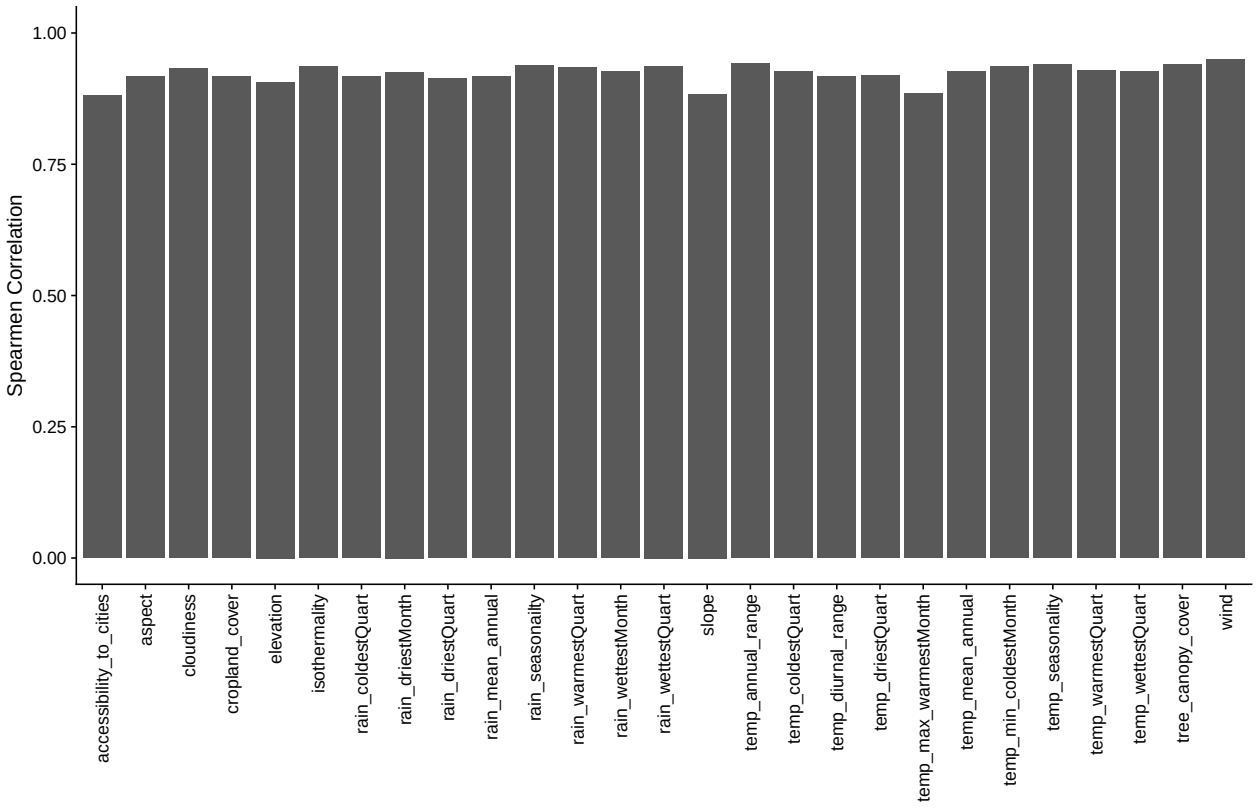
Structure of Data after binning based on above observation.

```
## 'data.frame':    137 obs. of  27 variables:
##  $ accessibility_to_cities: Factor w/ 3 levels "1","2","3": 2 1 3 2 2 1 3 1 1 1 ...
##  $ elevation              : Factor w/ 3 levels "1","2","3": 3 2 2 3 2 3 2 3 2 1 ...
##  $ aspect                 : Factor w/ 3 levels "1","2","3": 3 3 3 2 1 3 3 2 1 2 ...
##  $ slope                  : Factor w/ 3 levels "1","2","3": 3 3 1 1 1 3 1 2 2 1 ...
##  $ cropland_cover         : Factor w/ 3 levels "1","2","3": 1 2 1 1 2 2 1 2 2 3 ...
##  $ tree_canopy_cover      : Factor w/ 3 levels "1","2","3": 1 2 1 2 1 1 1 3 1 2 ...
##  $ isothermality          : Factor w/ 3 levels "1","2","3": 1 1 2 2 2 1 2 1 1 2 ...
##  $ rain_coldestQuart      : Factor w/ 3 levels "1","2","3": 1 3 1 1 1 1 1 2 1 1 ...
##  $ rain_driestMonth       : Factor w/ 3 levels "1","2","3": 1 3 1 1 2 2 1 3 2 1 ...
##  $ rain_driestQuart       : Factor w/ 3 levels "1","2","3": 1 2 1 1 1 1 1 3 1 1 ...
##  $ rain_mean_annual       : Factor w/ 3 levels "1","2","3": 1 2 1 2 2 2 1 2 1 3 ...
##  $ rain_seasonailty       : Factor w/ 3 levels "1","2","3": 3 1 2 3 1 1 2 1 1 3 ...
##  $ rain_warmestQuart      : Factor w/ 3 levels "1","2","3": 1 2 1 3 2 2 2 3 1 3 ...
##  $ rain_wettestMonth      : Factor w/ 3 levels "1","2","3": 1 2 1 2 1 1 1 2 1 3 ...
##  $ rain_wettestQuart      : Factor w/ 3 levels "1","2","3": 1 2 1 2 1 1 1 2 1 3 ...
##  $ temp_annual_range      : Factor w/ 3 levels "1","2","3": 3 2 3 2 2 3 2 2 3 2 ...
##  $ temp_coldestQuart      : Factor w/ 3 levels "1","2","3": 1 2 2 3 2 1 2 1 2 3 ...
##  $ temp_diurnal_range     : Factor w/ 3 levels "1","2","3": 3 1 3 2 2 2 2 1 1 1 ...
##  $ temp_driestQuart       : Factor w/ 3 levels "1","2","3": 3 2 3 2 2 1 2 1 2 2 ...
##  $ temp_max_warmestMonth  : Factor w/ 3 levels "1","2","3": 2 2 3 2 2 1 3 1 2 2 ...
##  $ temp_mean_annual       : Factor w/ 3 levels "1","2","3": 1 1 2 2 2 1 2 1 1 3 ...
```

```
##  $ temp_min_coldestMonth  : Factor w/ 3 levels "1","2","3": 1 1 2 2 2 1 2 1 1 3 ...
##  $ temp_seasonality       : Factor w/ 3 levels "1","2","3": 3 2 3 1 2 3 2 2 3 2 ...
##  $ temp_warmestQuart      : Factor w/ 3 levels "1","2","3": 2 1 3 2 2 1 3 1 2 3 ...
##  $ temp_wettestQuart      : Factor w/ 3 levels "1","2","3": 1 1 2 2 2 1 2 1 1 3 ...
##  $ wind                   : Factor w/ 4 levels "1","2","3","4": 3 2 4 2 4 1 4 2 2 2 ...
##  $ cloudiness             : Factor w/ 3 levels "1","2","3": 1 2 1 2 2 2 1 3 2 2 ...
```

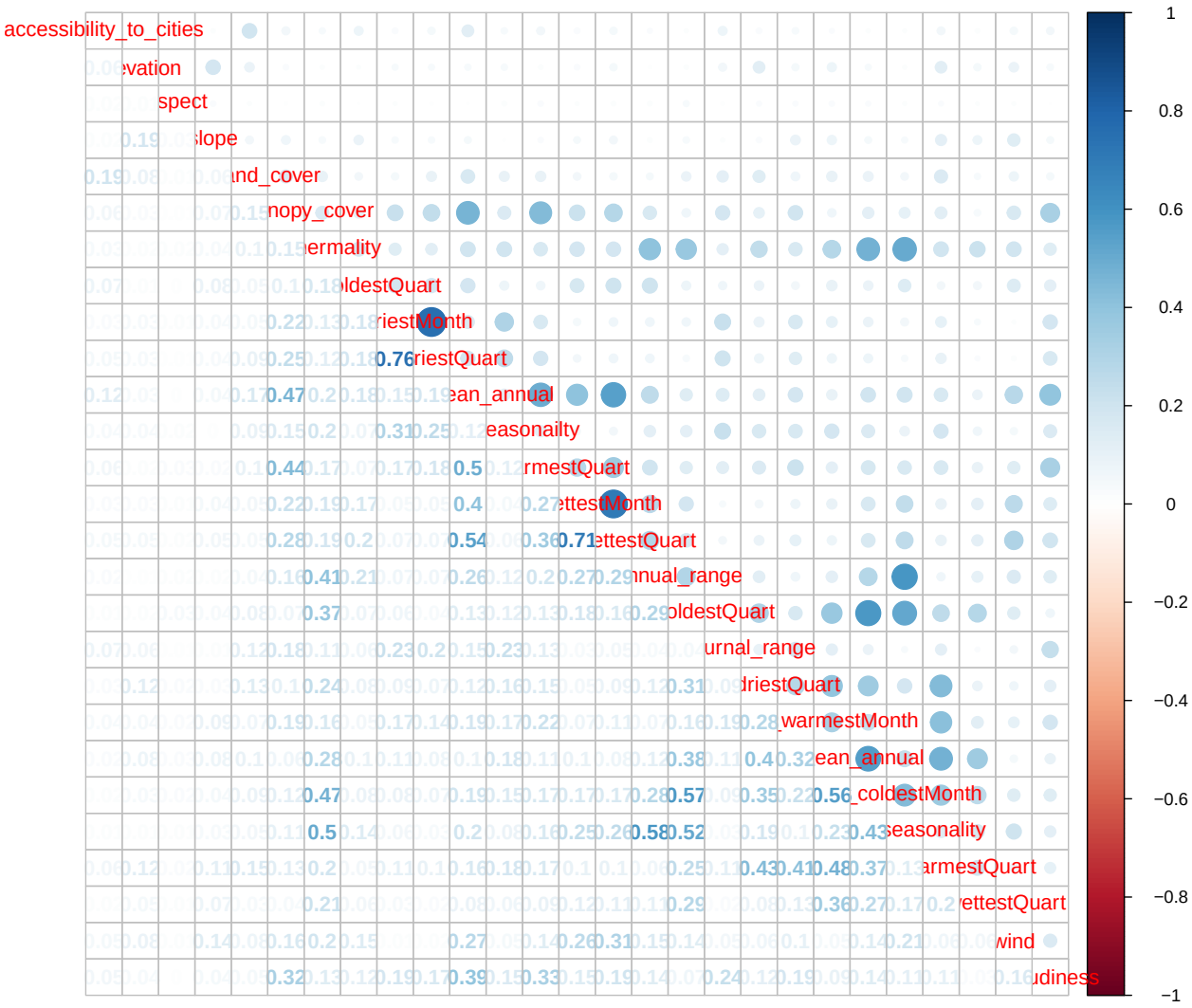
## 4.4 Spearman Correlation

Let's observe correlation between original data and binned data to make sure that neither the correlation ceofficient is too low or perfect.




## 4.5 Heatmap

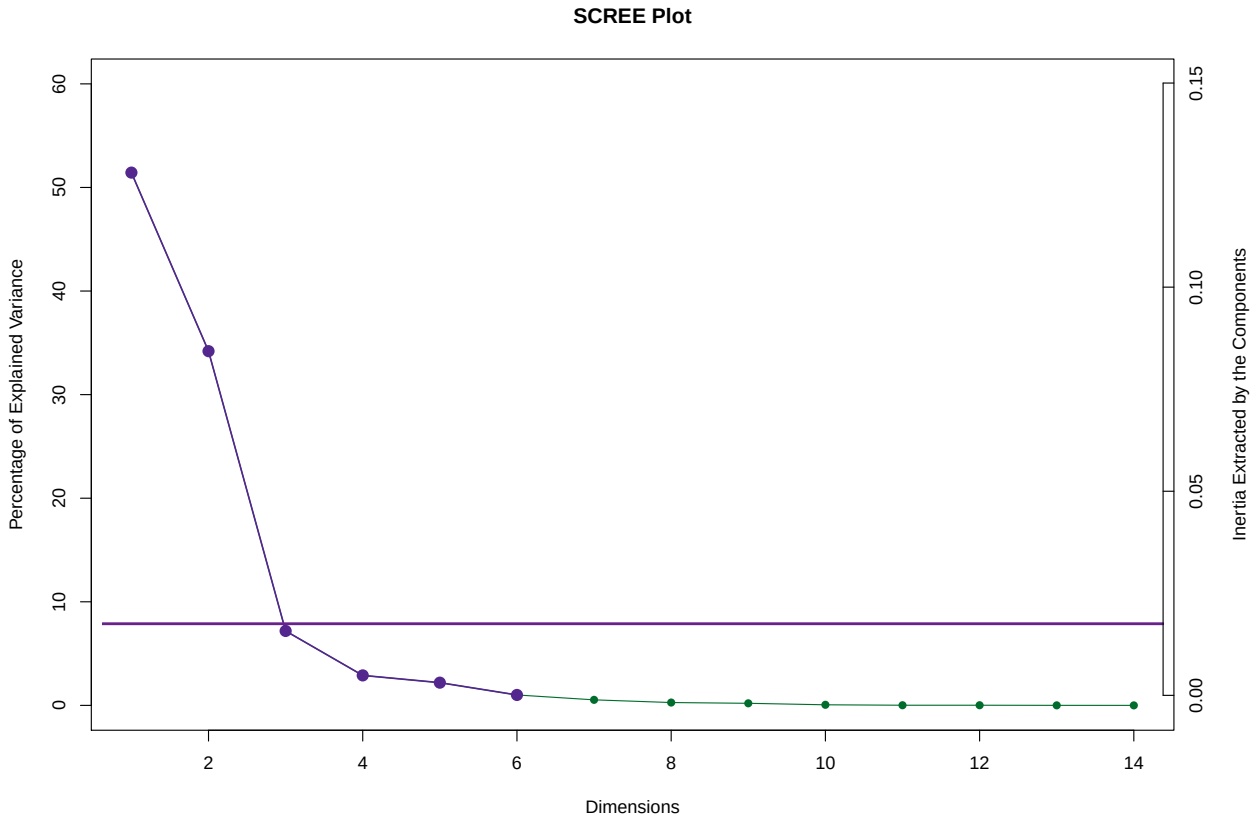
- For binned data




## 4.6 Scree Plot

Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

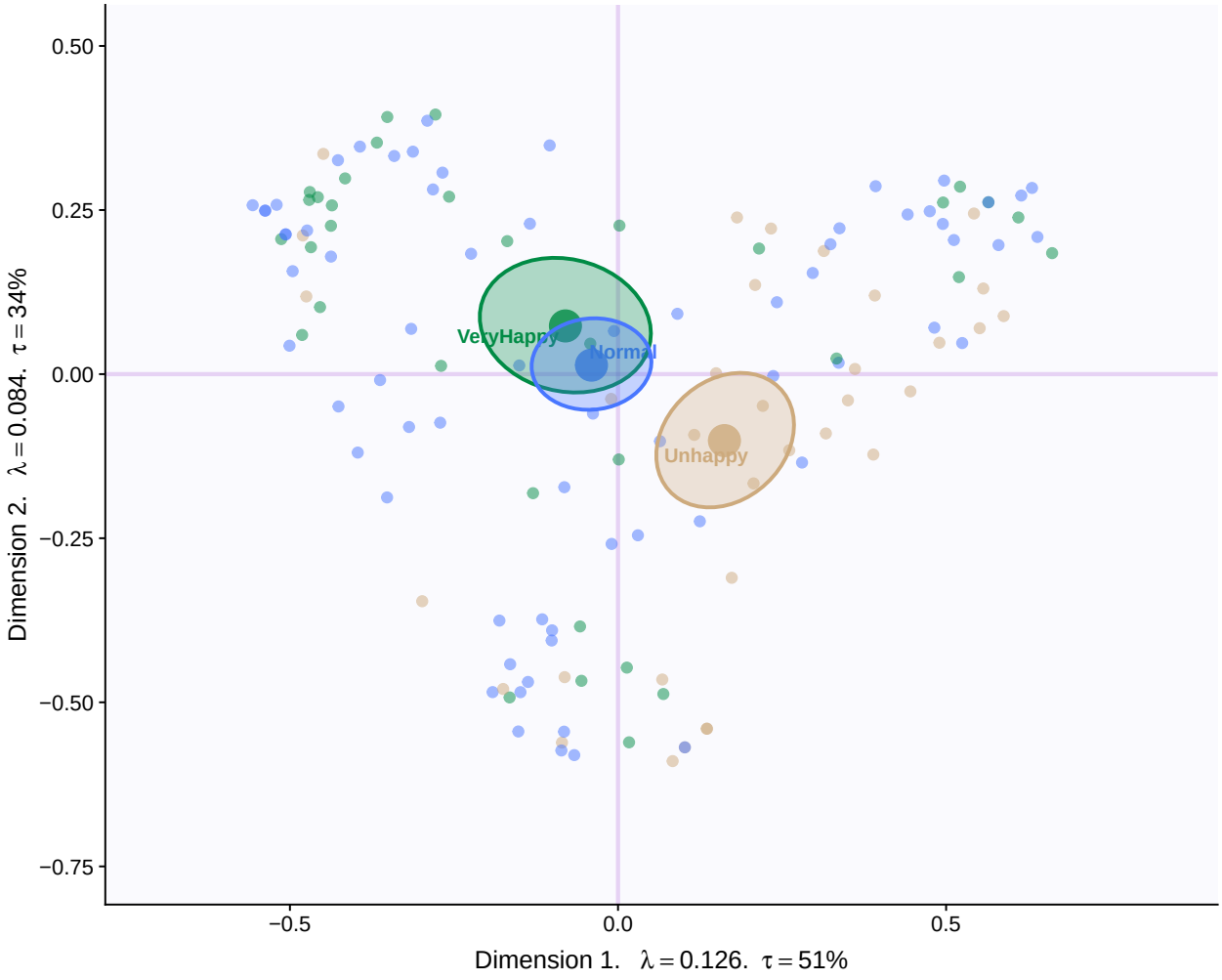
P.S. The most contribution component may not always be most useful for a given research question.

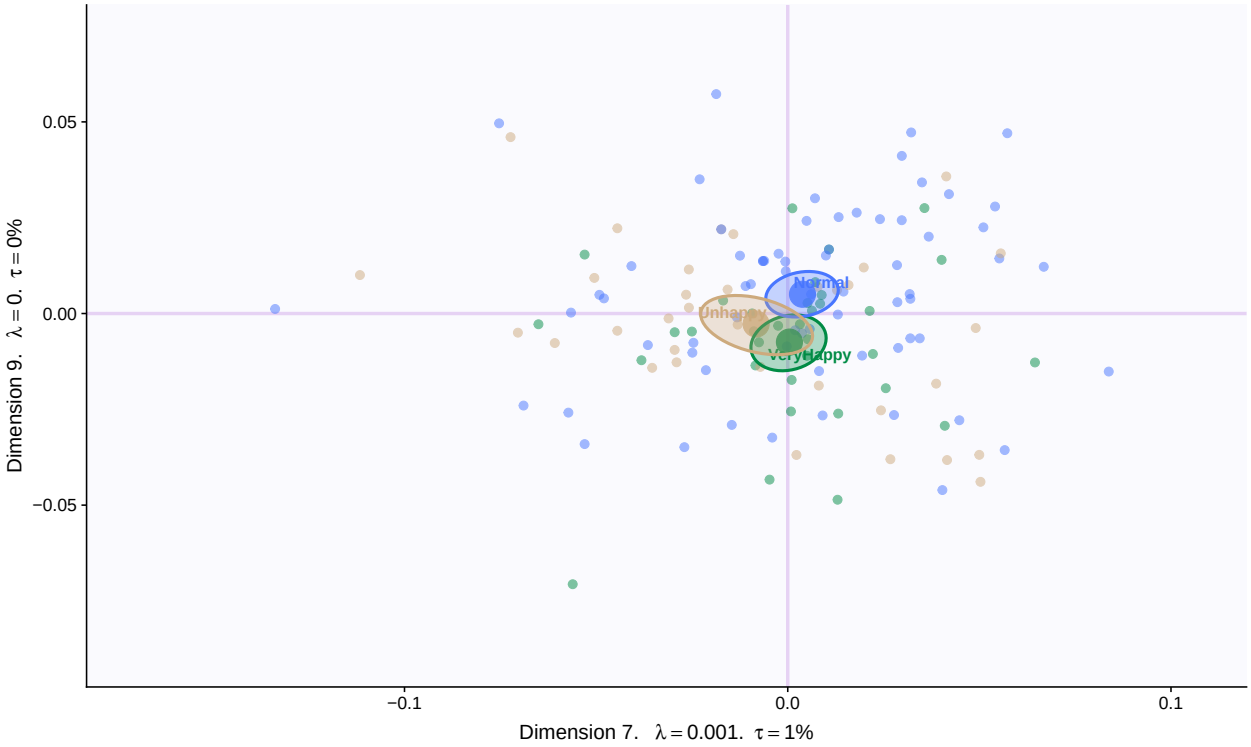



## 4.7 Factor Scores

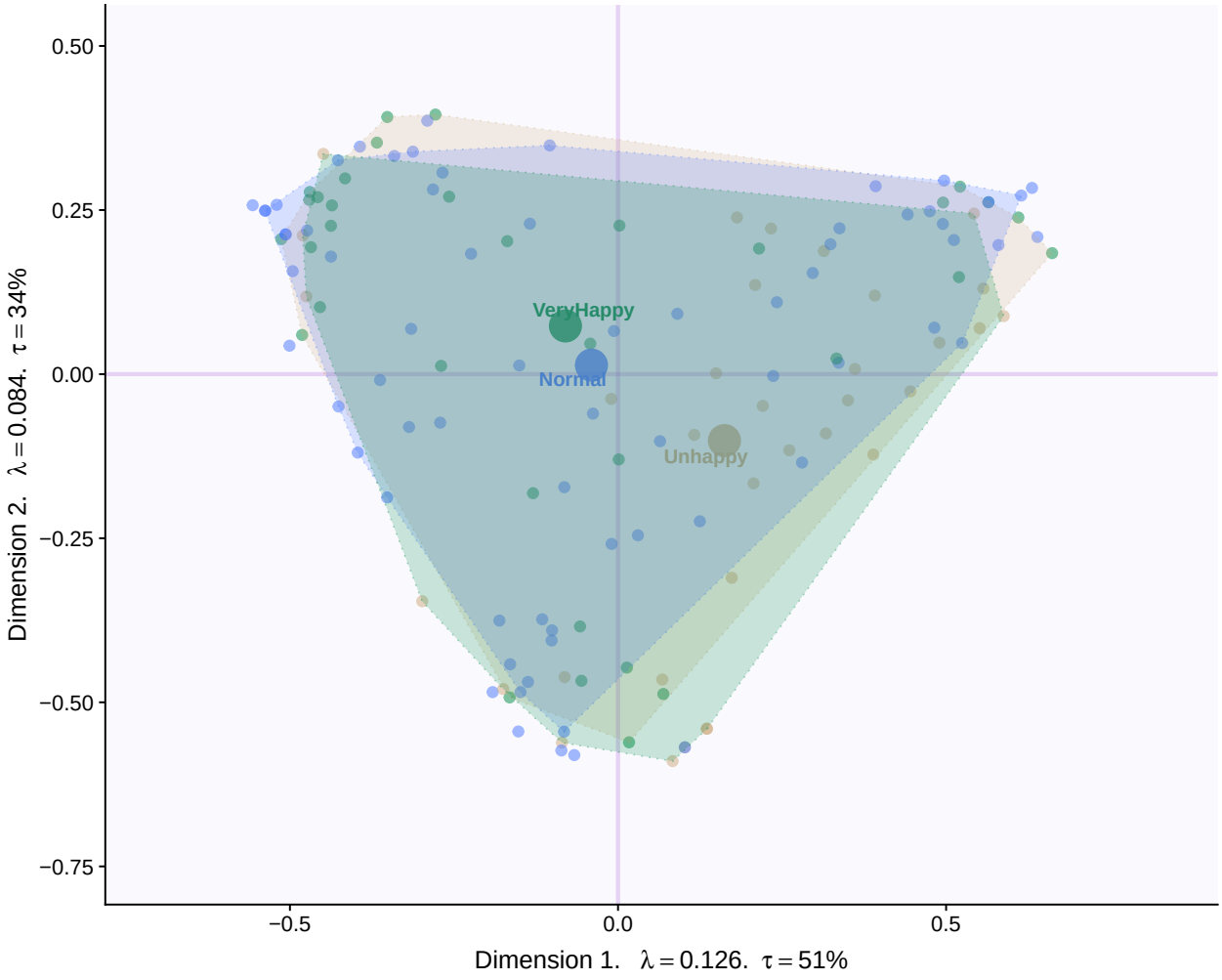
Lets visualize happiness categories for components 1-10, to make a decision (visually) on the most important components.

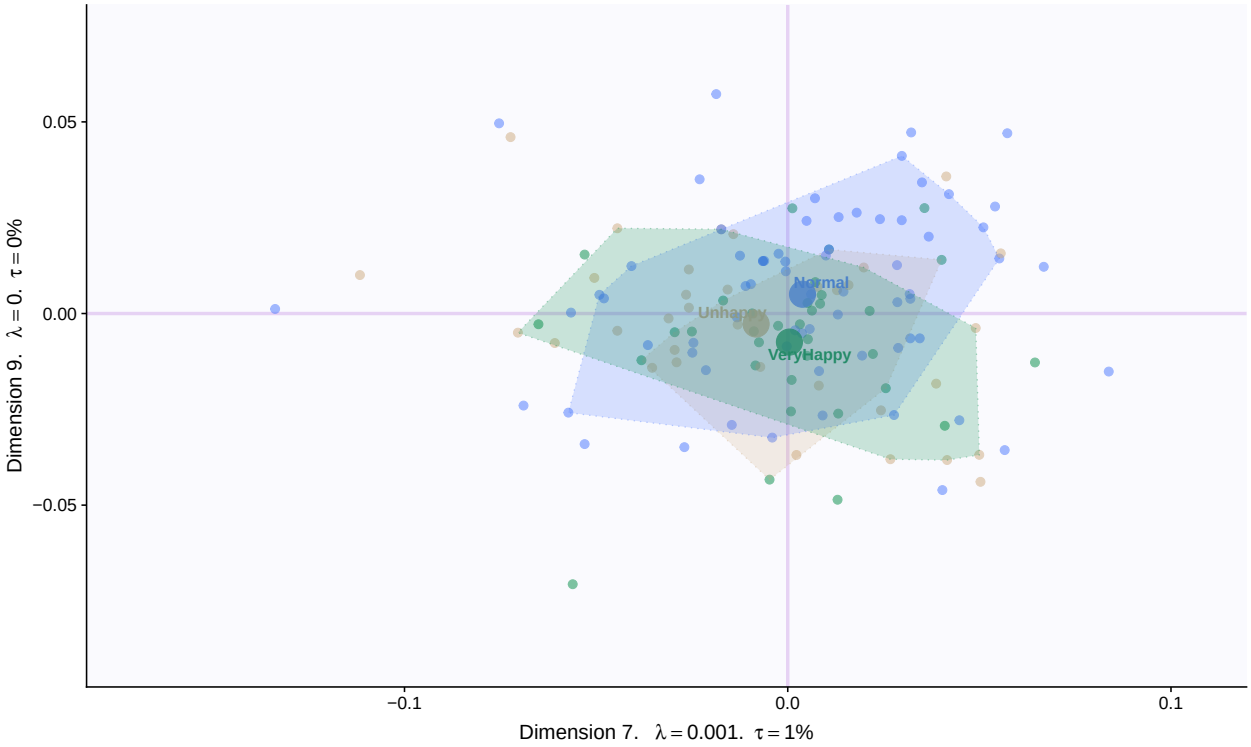
**With Confidence Interval**


0.50
0.25
0.00
−0.25
−0.50
−0.75
Dimension 2. λ = 0.084. τ = 34%
VeryHappy
Normal
Unhappy
−0.5
0.0
0.5
Dimension 1. λ = 0.126. τ = 51%

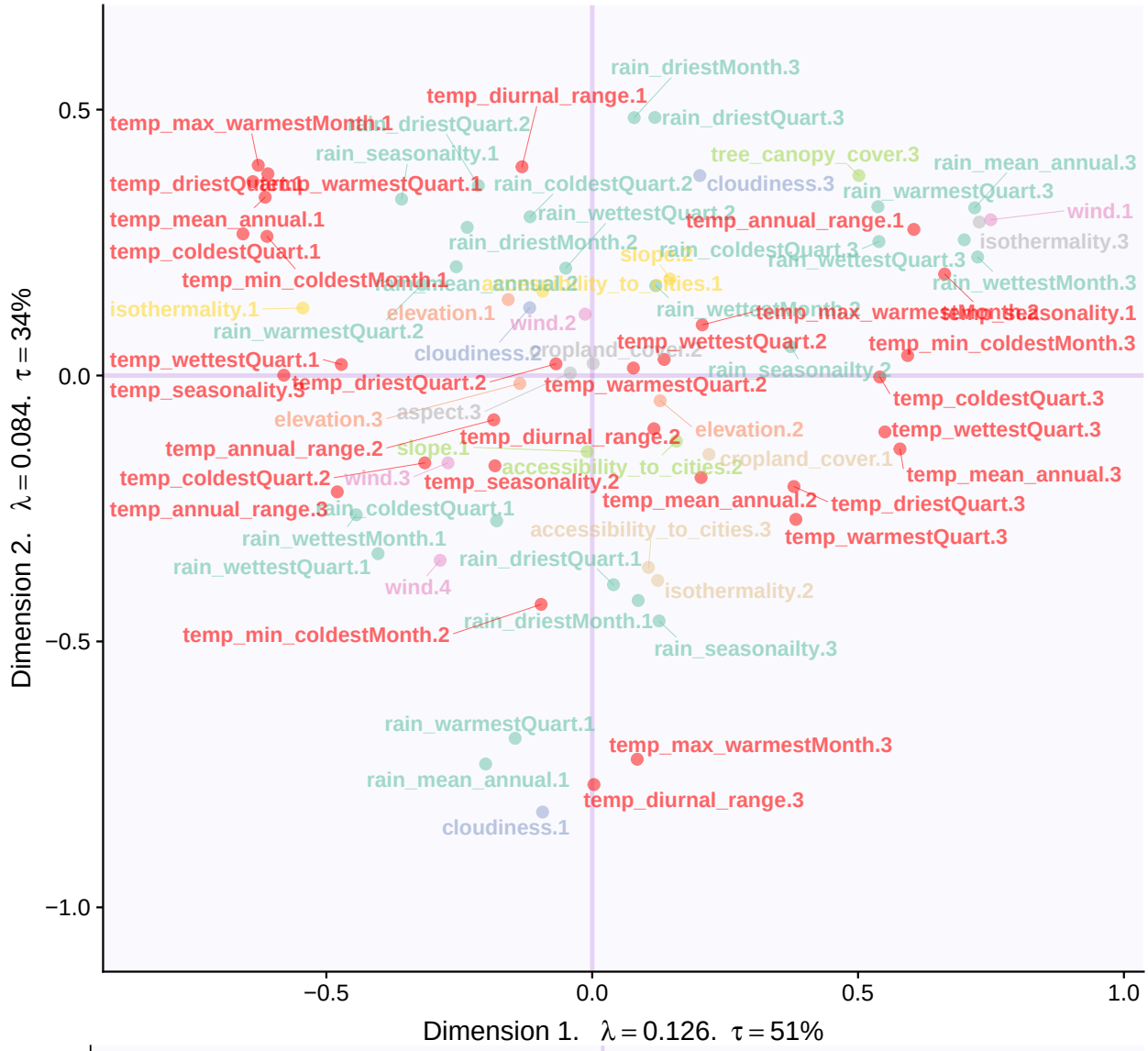


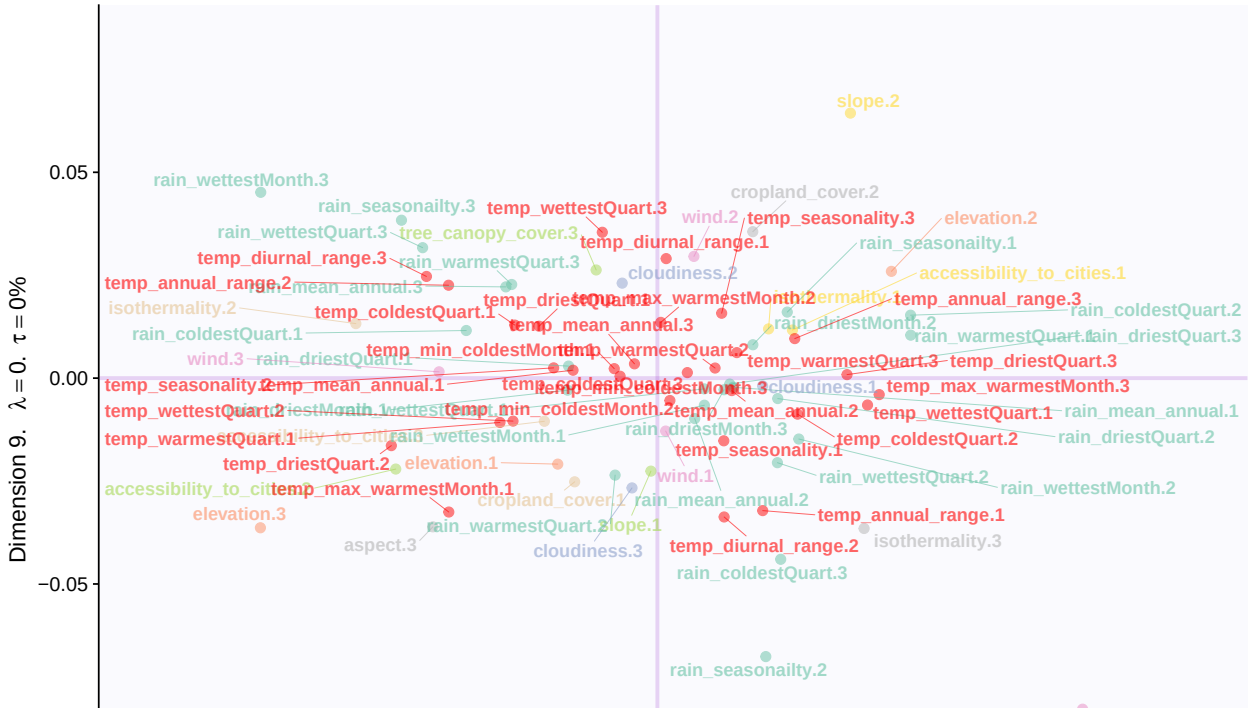

**With Tolerance Interval**


0.50
0.25
0.00
−0.25
−0.50
−0.75
−0.5
0.0
0.5
VeryHappy
Normal
Unhappy
Dimension 2. λ = 0.084. τ = 34%
Dimension 1. λ = 0.126. τ = 51%

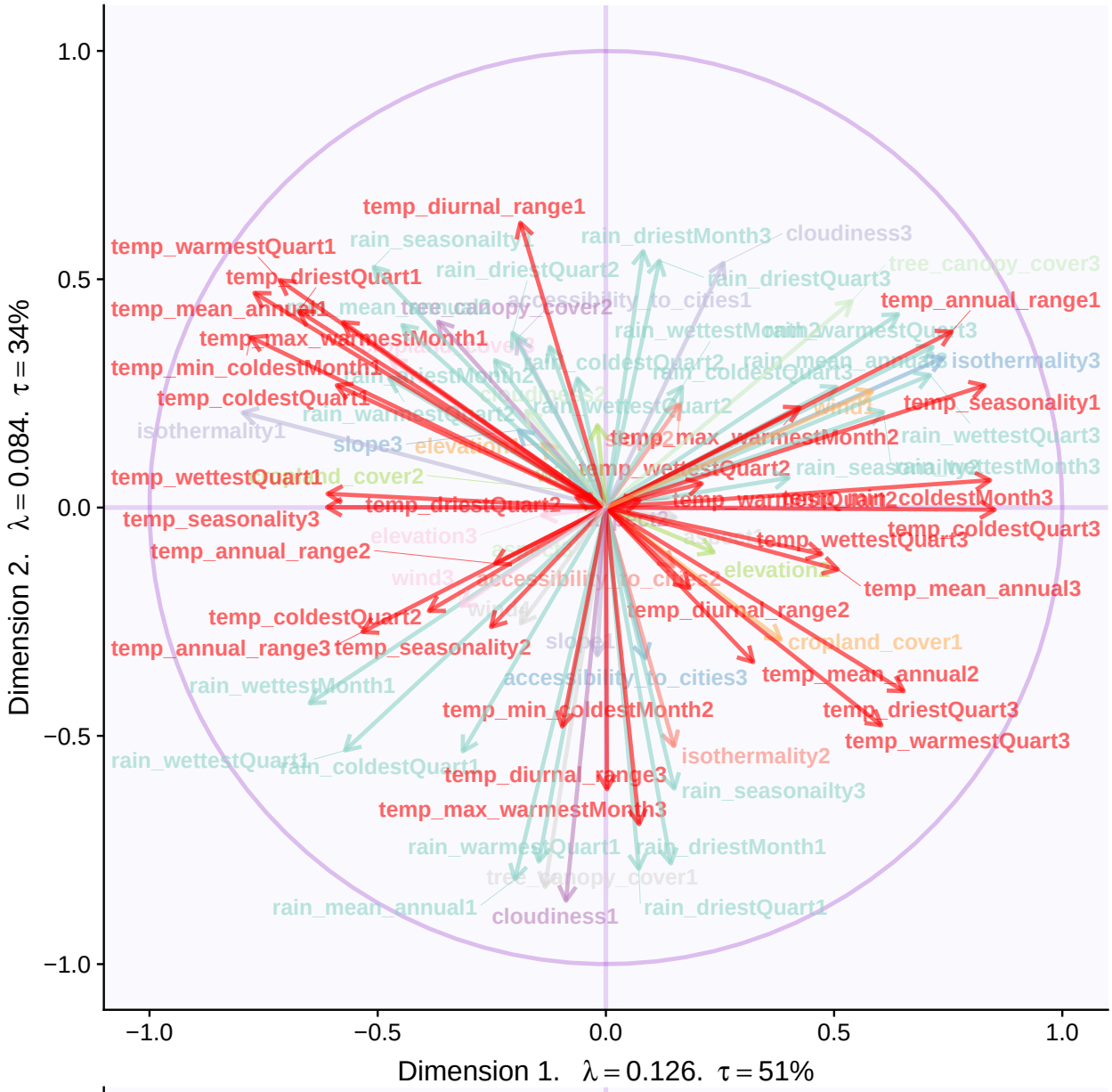

0.05
0.00
−0.05
Dimension 9. λ = 0. τ = 0%
Normal
Unhappy
VeryHappy
−0.1
0.0
0.1
Dimension 7. λ = 0.001. τ = 1%

## 4.8 Loadings


rain_driestMonth.3
temp_diurnal_range.1
temp_max_warmestMonth.1
rain_driestQuart.3
rain_driestQuart.2
rain_seasonailty.1
tree_canopy_cover.3
rain_mean_annual.3
temp_driestQuart.1
temp_warmestQuart.1
rain_coldestQuart.2
cloudiness.3
rain_warmestQuart.3
temp_mean_annual.1
rain_wettestQuart.2
temp_annual_range.1
wind.1
temp_coldestQuart.1
rain_driestMonth.2
isothermality.3
slope.2
rain_coldestQuart.3
rain_wettestQuart.3
temp_min_coldestMonth.1
rain_mean_annual.2
accessibility_to_cities.1
rain_wettestMonth.3
isothermality.1
elevation.1
wind.2
rain_wettestMonth.2
temp_max_warmestMonth.2
temp_seasonality.1
rain_warmestQuart.2
cloudiness.2
cropland_cover.2
temp_wettestQuart.2
temp_min_coldestMonth.3
temp_wettestQuart.1
rain_seasonailty.2
temp_driestQuart.2
temp_warmestQuart.2
temp_seasonality.3
temp_coldestQuart.3
elevation.3
aspect.3
elevation.2
temp_wettestQuart.3
temp_annual_range.2
temp_diurnal_range.2
slope.1
cropland_cover.1
temp_coldestQuart.2
wind.3
temp_seasonality.2
accessibility_to_cities.2
temp_mean_annual.3
temp_annual_range.3
rain_coldestQuart.1
temp_mean_annual.2
temp_driestQuart.3
rain_wettestMonth.1
accessibility_to_cities.3
temp_warmestQuart.3
rain_wettestQuart.1
rain_driestQuart.1
wind.4
isothermality.2
rain_driestMonth.1
temp_min_coldestMonth.2
rain_seasonailty.3
rain_warmestQuart.1
temp_max_warmestMonth.3
rain_mean_annual.1
temp_diurnal_range.3
cloudiness.1
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
Dimension 2. λ = 0.084. τ = 34%
Dimension 1. λ = 0.126. τ = 51%


slope.2
rain_wettestMonth.3
cropland_cover.2
rain_seasonailty.3
temp_wettestQuart.3
wind.2
temp_seasonality.3
elevation.2
rain_wettestQuart.3
tree_canopy_cover.3
temp_diurnal_range.1
rain_seasonailty.1
temp_diurnal_range.3
rain_warmestQuart.3
cloudiness.2
accessibility_to_cities.1
temp_annual_range.2
rain_mean_annual.3
isothermality.2
temp_coldestQuart.1
temp_driestQuart.1
temp_max_warmestMonth.2
temp_annual_range.3
rain_coldestQuart.2
rain_coldestQuart.1
temp_mean_annual.3
rain_driestMonth.2
rain_warmestQuart.1
rain_driestQuart.3
wind.3
rain_driestQuart.1
temp_min_coldestMonth.1
temp_warmestQuart.2
temp_warmestQuart.3
temp_driestQuart.3
temp_seasonality.2
temp_mean_annual.1
temp_coldestQuart.3
temp_min_coldestMonth.3
cloudiness.1
temp_max_warmestMonth.3
temp_wettestQuart.2
temp_min_coldestMonth.2
temp_mean_annual.2
temp_wettestQuart.1
rain_mean_annual.1
temp_warmestQuart.1
rain_wettestMonth.1
rain_driestMonth.3
rain_driestQuart.2
temp_coldestQuart.2
temp_driestQuart.2
elevation.1
temp_seasonality.1
wind.1
rain_wettestQuart.2
accessibility_to_cities.2
temp_max_warmestMonth.1
rain_wettestMonth.2
cropland_cover.1
rain_mean_annual.2
elevation.3
temp_annual_range.1
rain_warmestQuart.2
slope.1
aspect.3
temp_diurnal_range.2
isothermality.3
cloudiness.3
rain_coldestQuart.3
rain_seasonailty.2
0.05
0.00
−0.05
Dimension 9. λ = 0. τ = 0%

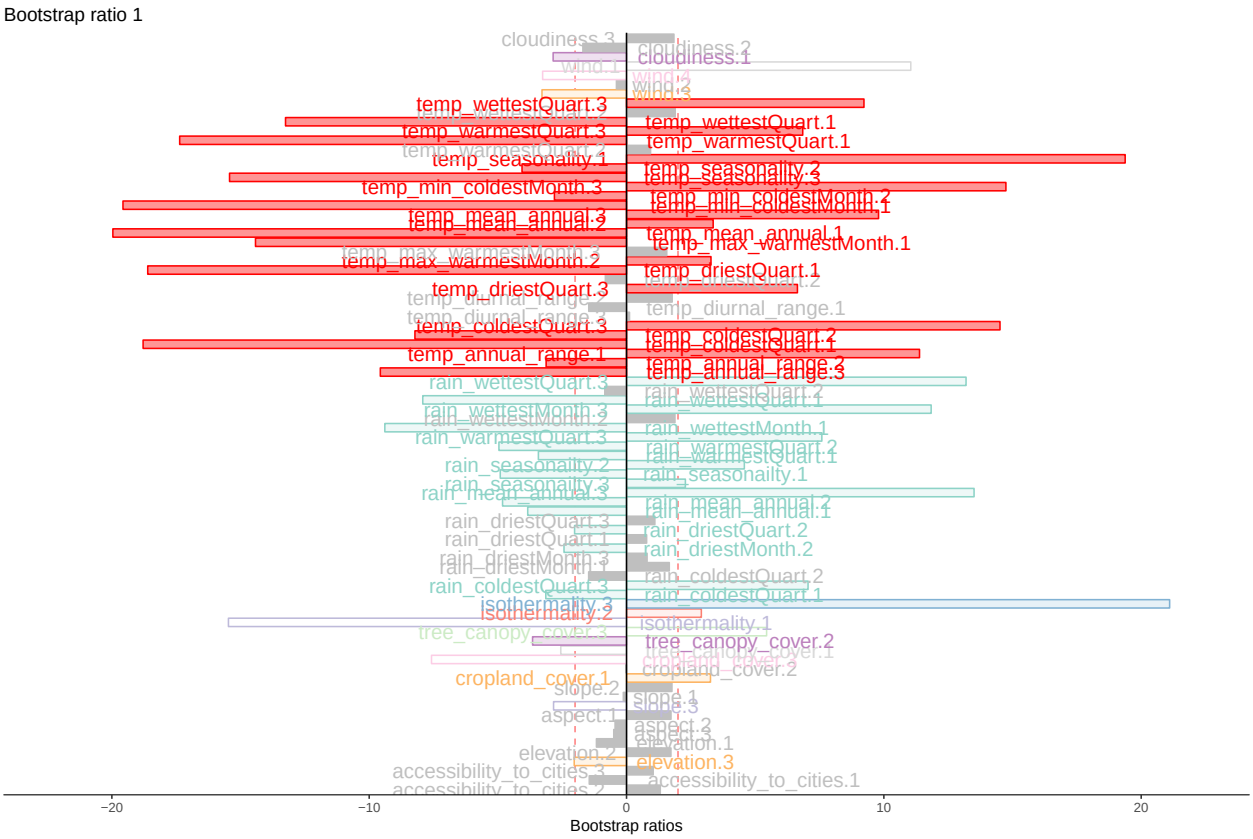
## 4.9 Loadings (correlation plot)



## 4.10 Most Contributing Variables (Inference)

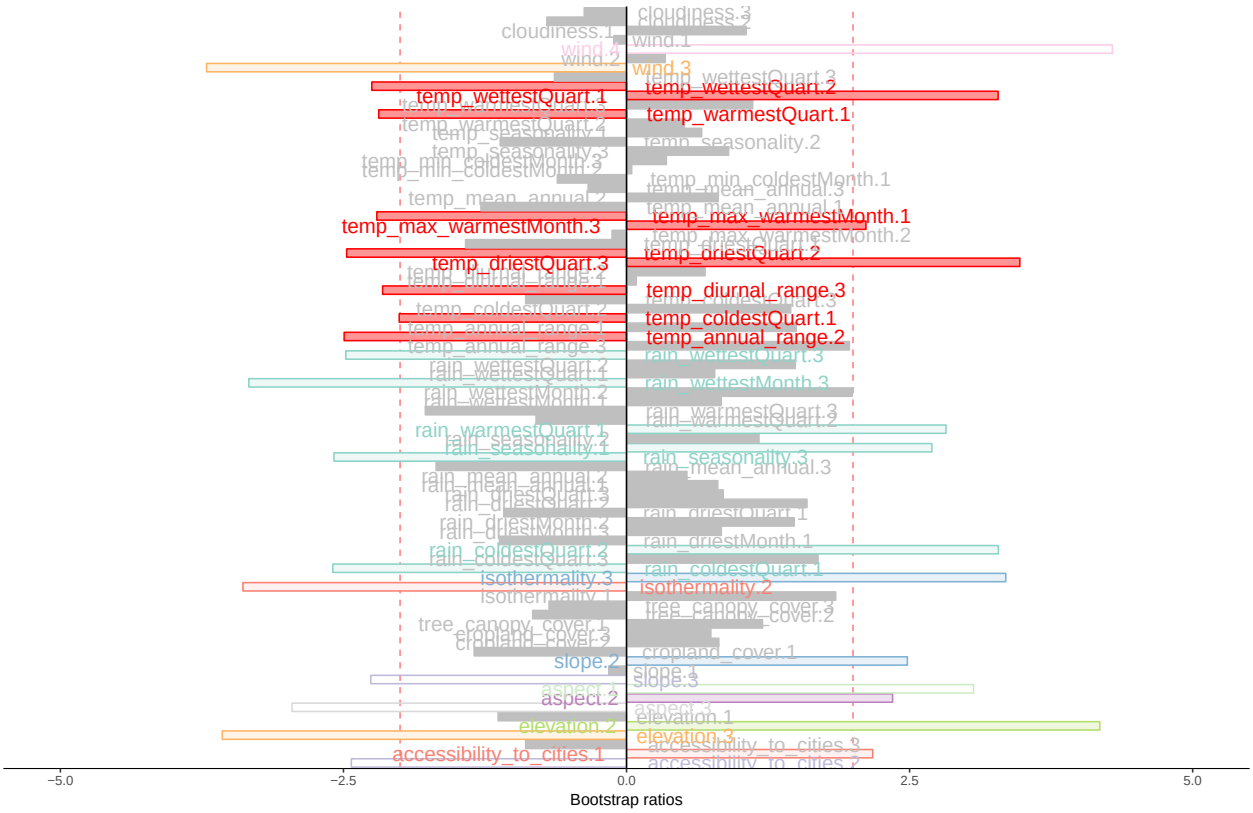
Let's plot variable contributions against each chosen components i.e. 1, 2, 7, 9.

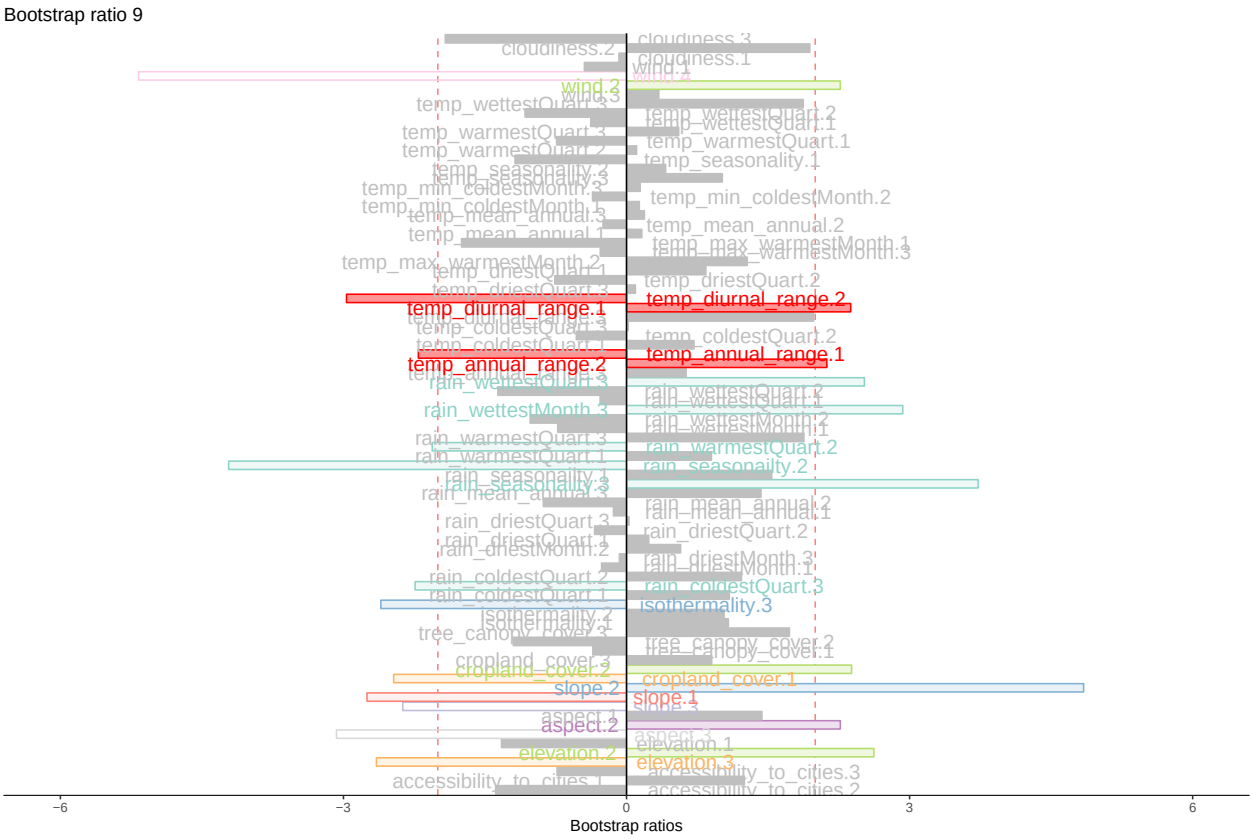
### With Bootstrap Ratio



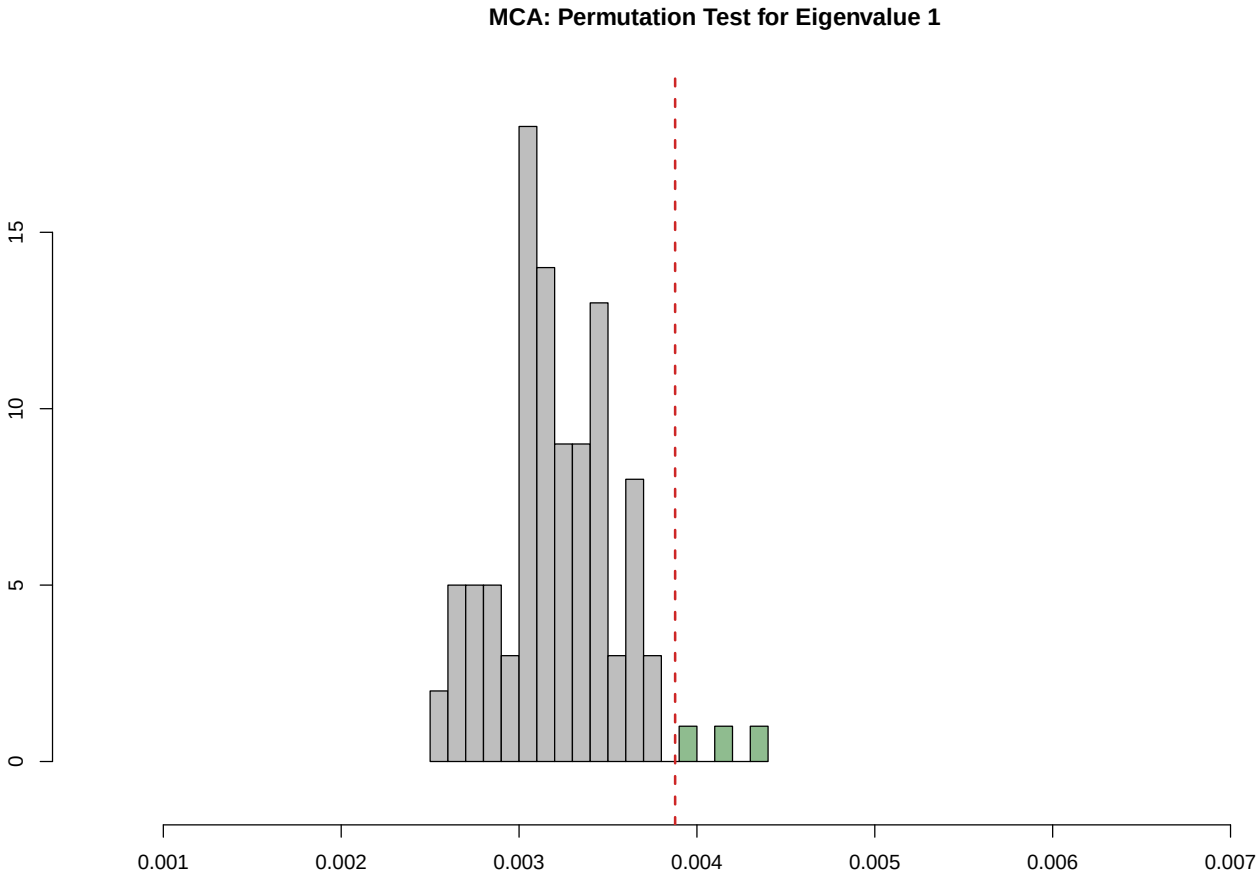
Bootstrap ratio 2

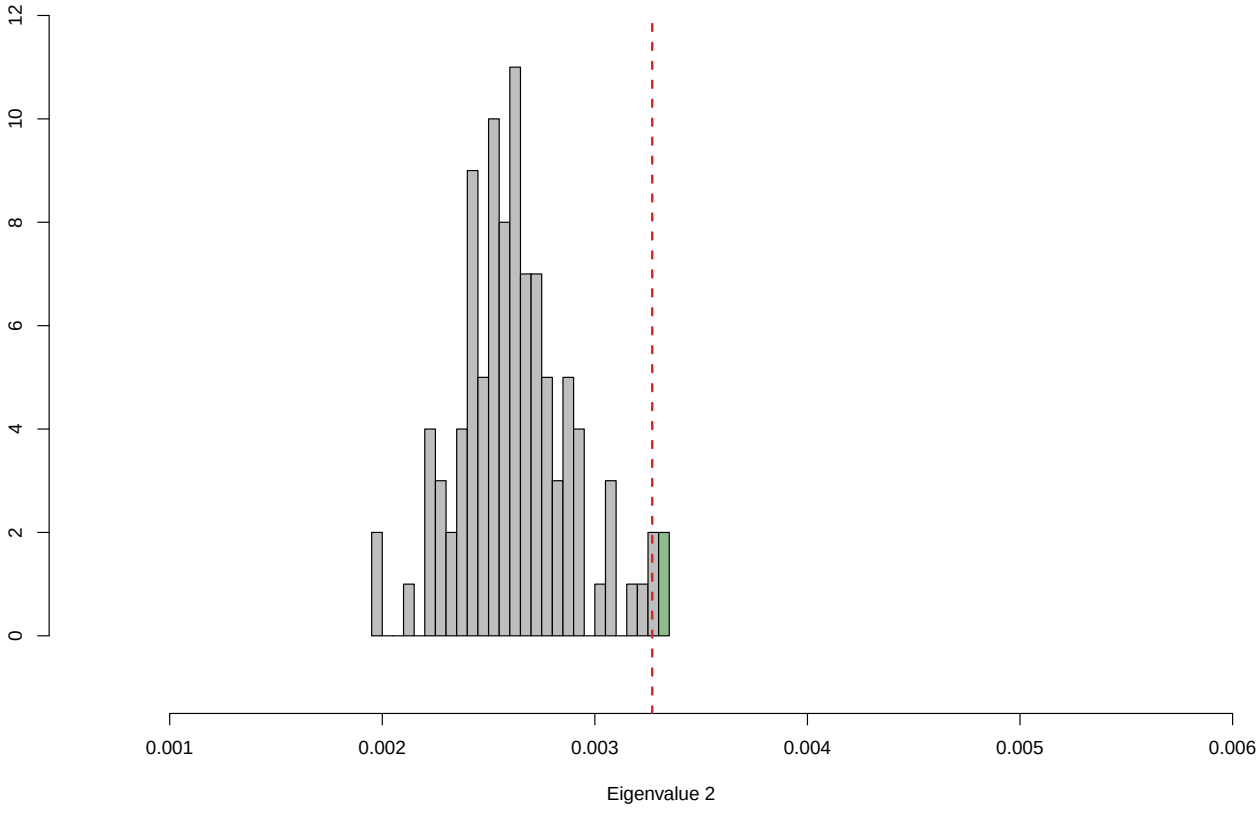

Bootstrap ratio 7

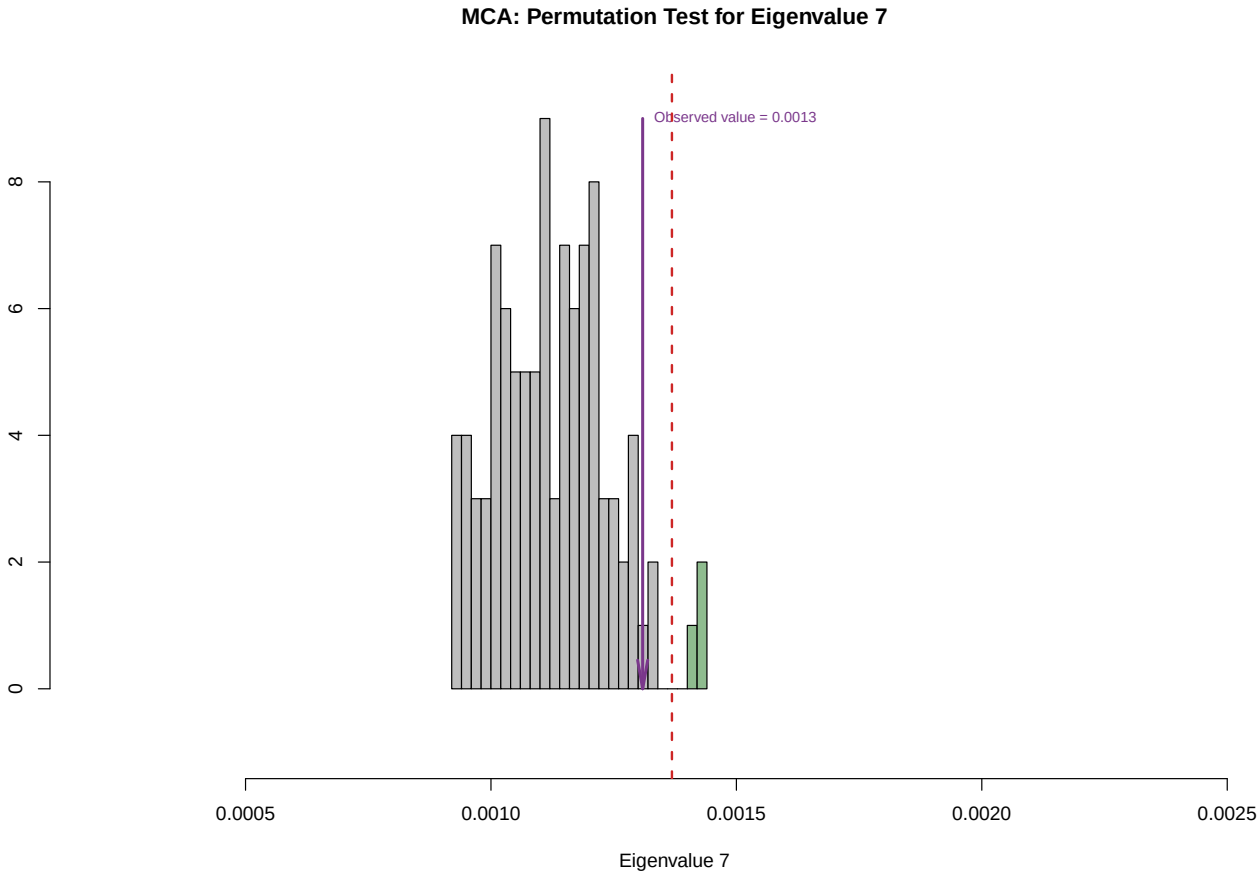


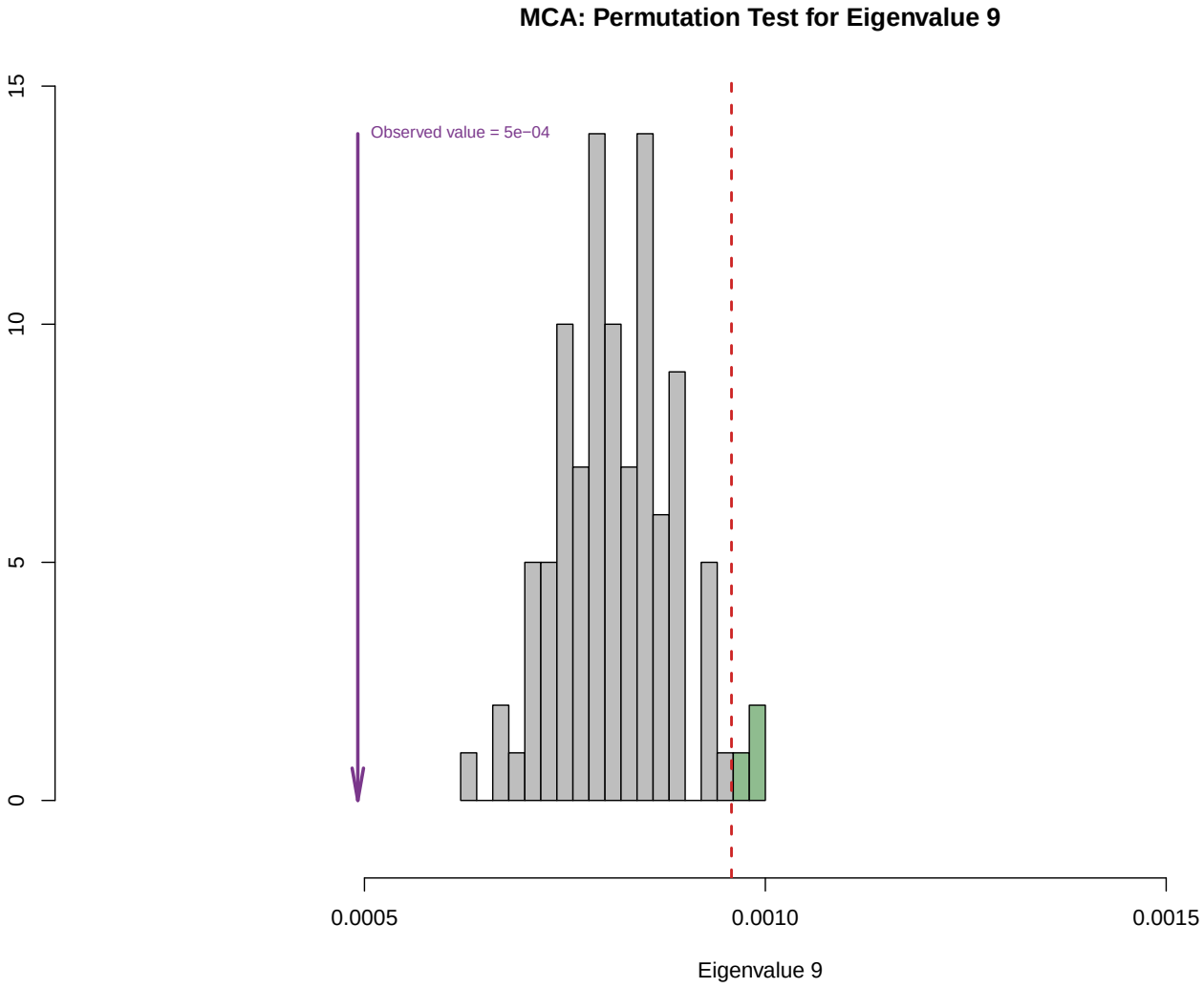

Bootstrap ratio 9
cloudiness.3
cloudiness.2
cloudiness.1
wind.1
wind.2
wind.3
temp_wettestQuart.3
temp_wettestQuart.2
temp_wettestQuart.1
temp_warmestQuart.3
temp_warmestQuart.2
temp_warmestQuart.1
temp_seasonality.3
temp_seasonality.2
temp_seasonality.1
temp_min_coldestMonth.3
temp_min_coldestMonth.2
temp_min_coldestMonth.1
temp_mean_annual.3
temp_mean_annual.2
temp_mean_annual.1
temp_max_warmestMonth.1
temp_max_warmestMonth.3
temp_max_warmestMonth.2
temp_driestQuart.1
temp_driestQuart.2
temp_driestQuart.3
temp_diurnal_range.1
temp_diurnal_range.2
temp_coldestQuart.3
temp_coldestQuart.2
temp_coldestQuart.1
temp_annual_range.2
temp_annual_range.1
rain_wettestQuart.3
rain_wettestQuart.2
rain_wettestQuart.1
rain_wettestMonth.3
rain_wettestMonth.2
rain_wettestMonth.1
rain_warmestQuart.3
rain_warmestQuart.1
rain_warmestQuart.2
rain_seasonality.1
rain_seasonality.2
rain_seasonality.3
rain_mean_annual.3
rain_mean_annual.2
rain_mean_annual.1
rain_driestQuart.3
rain_driestQuart.2
rain_driestQuart.1
rain_driestMonth.2
rain_driestMonth.3
rain_driestMonth.1
rain_coldestQuart.2
rain_coldestQuart.3
rain_coldestQuart.1
isothermality.3
isothermality.2
isothermality.1
tree_canopy_cover.3
tree_canopy_cover.2
tree_canopy_cover.1
cropland_cover.3
cropland_cover.2
cropland_cover.1
slope.2
slope.1
slope.3
aspect.1
aspect.2
aspect.3
elevation.1
elevation.2
elevation.3
accessibility_to_cities.3
accessibility_to_cities.1
accessibility_to_cities.2
−6
−3
0
3
6
Bootstrap ratios

## 4.11 Permutation Test

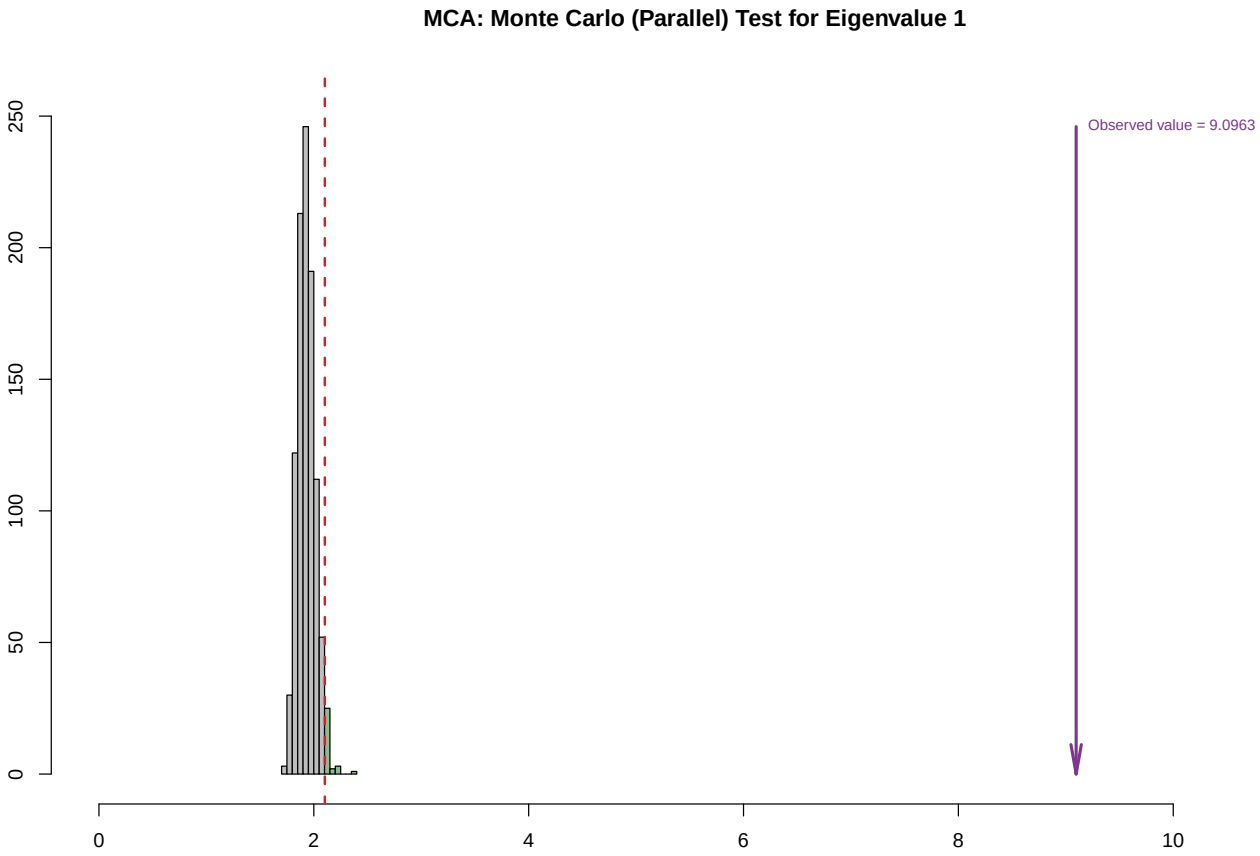


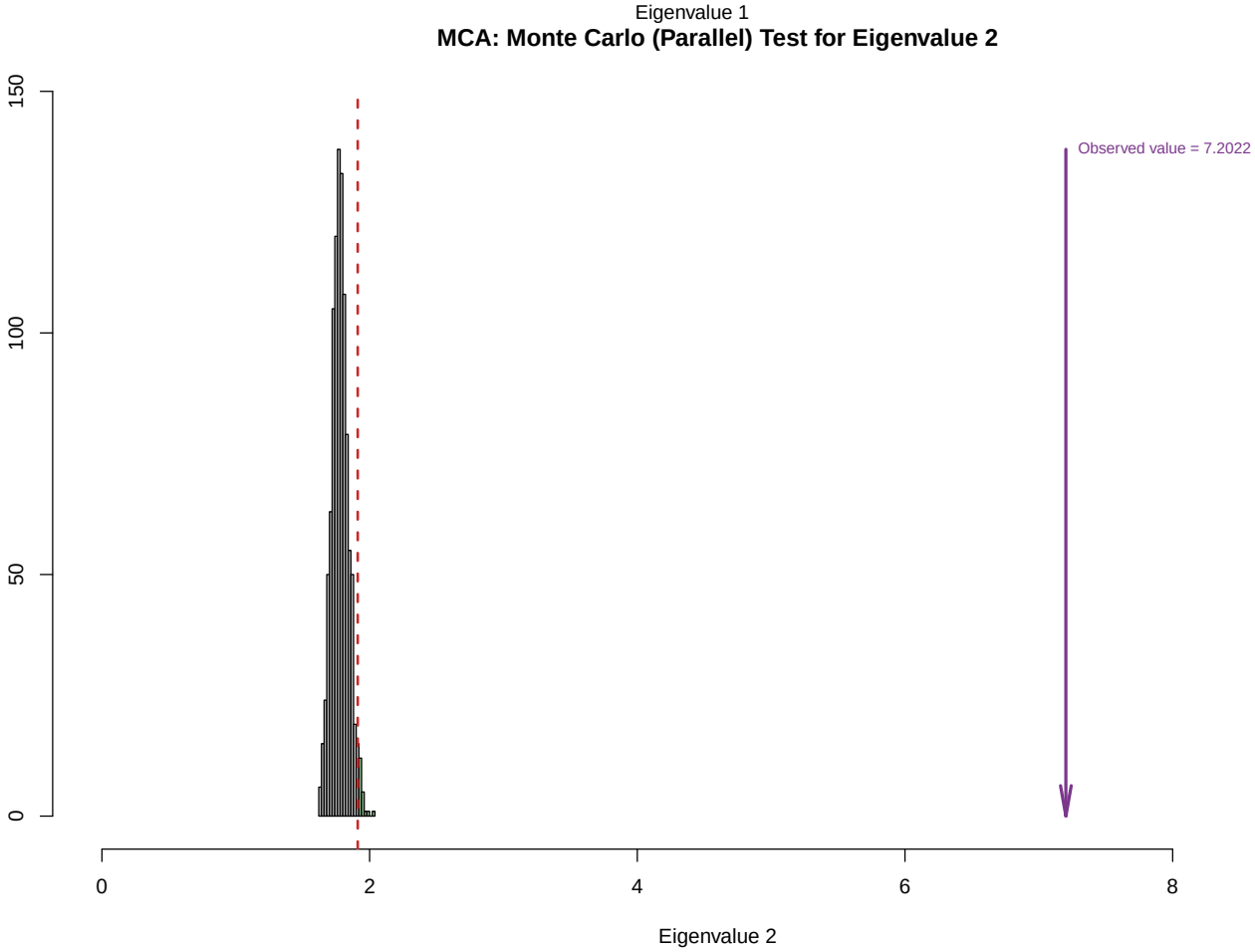



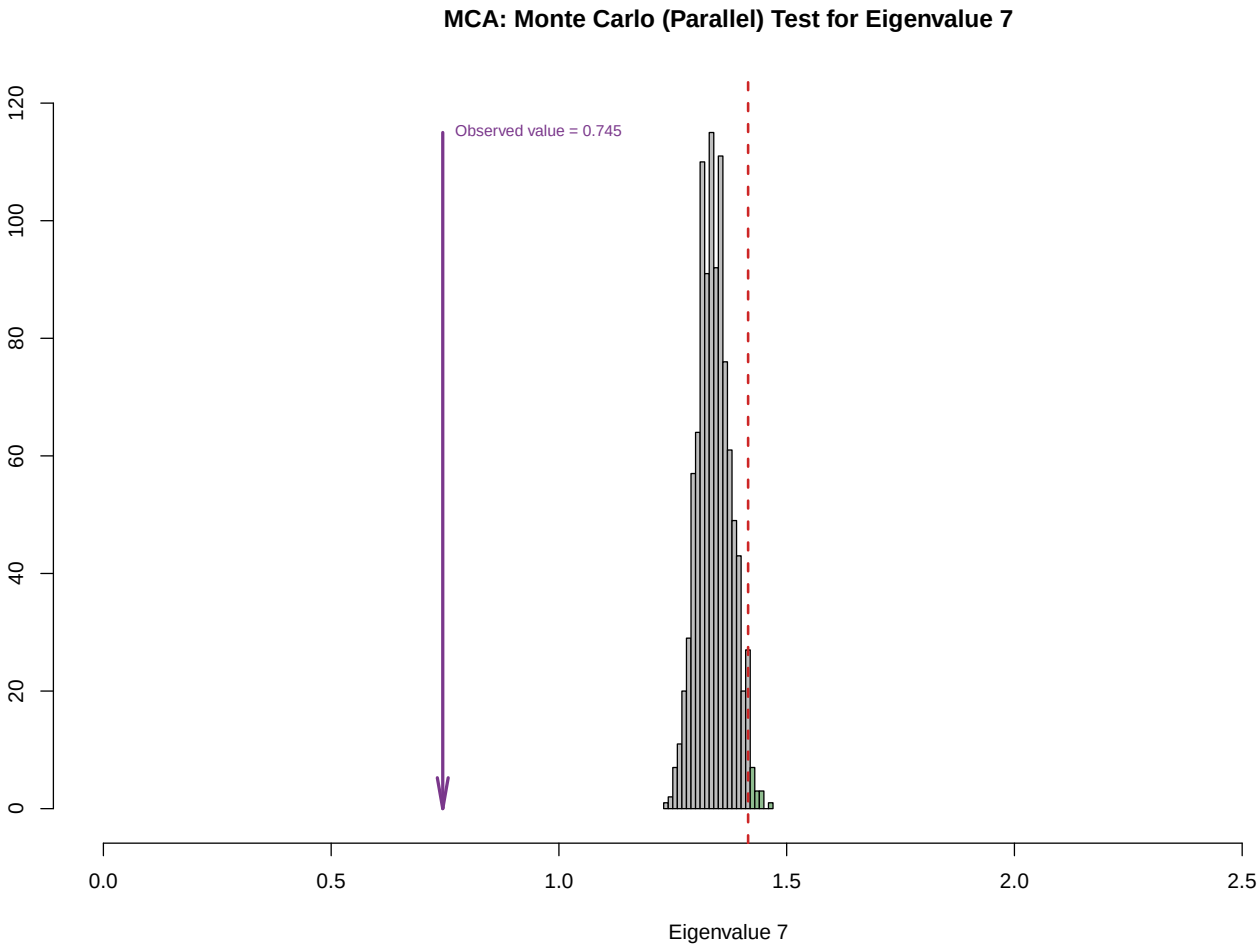

MCA: Permutation Test for Eigenvalue 7
Observed value = 0.0013
0
2
4
6
8
0.0005
0.0010
0.0015
0.0020
0.0025
Eigenvalue 7

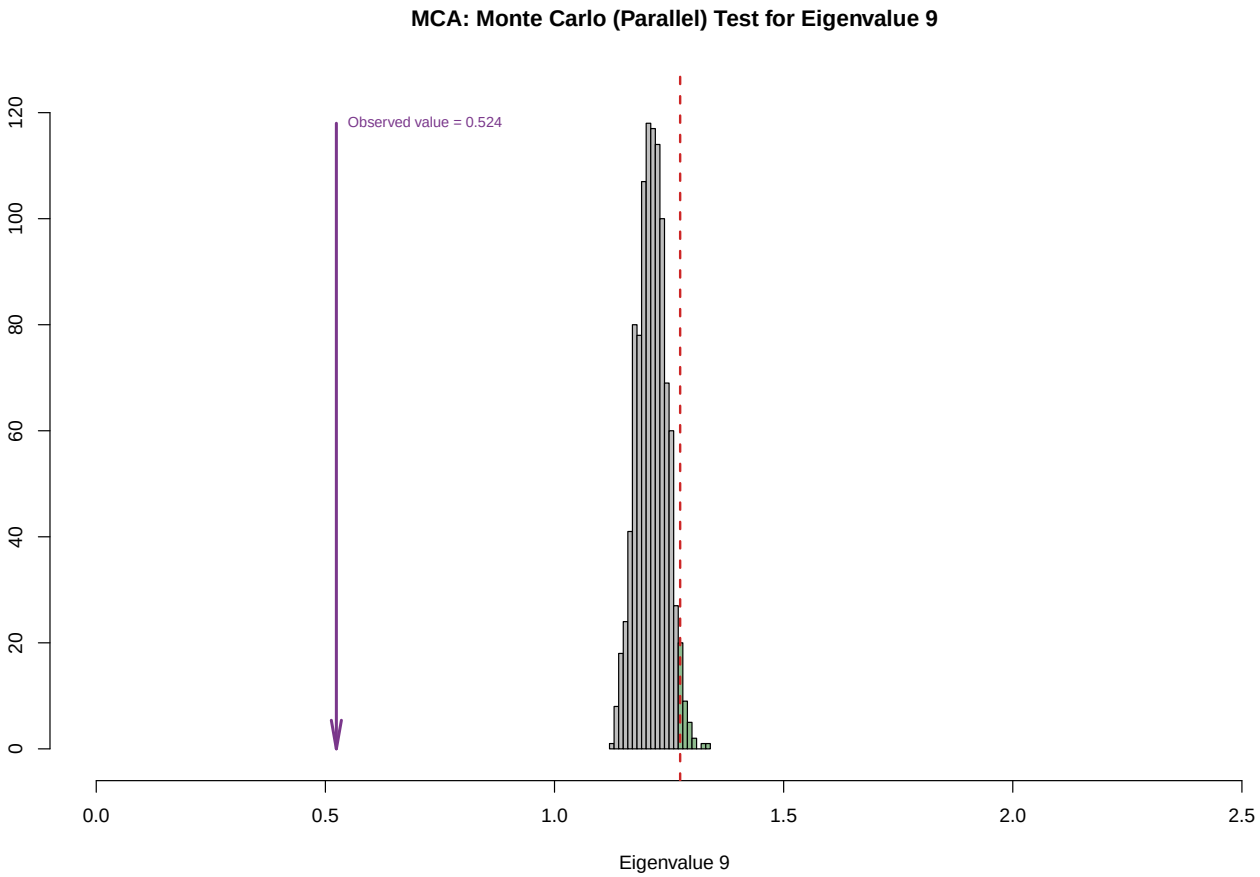

MCA: Permutation Test for Eigenvalue 9
Observed value = 5e−04
0
5
10
15
0.0005
0.0010
0.0015
Eigenvalue 9

## 4.12 Parallet Test

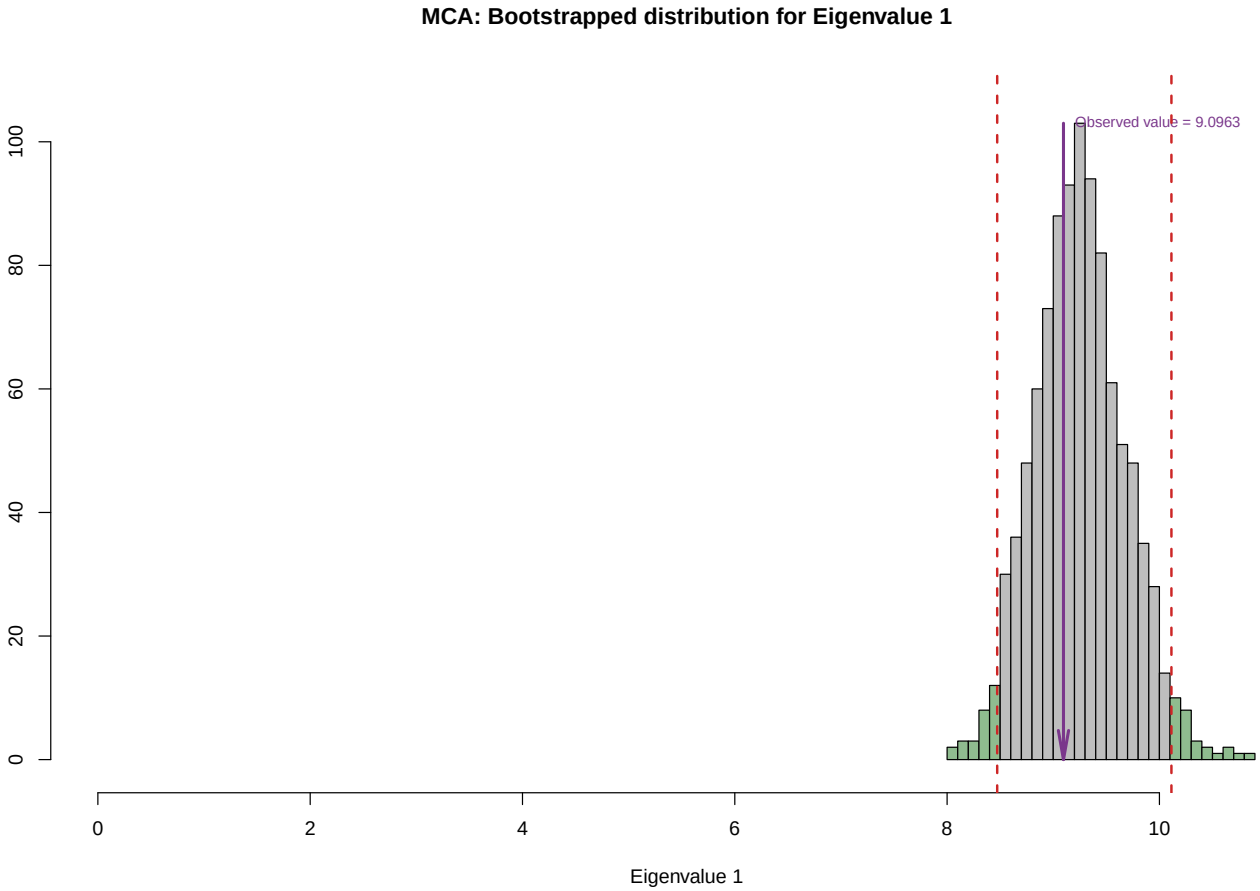


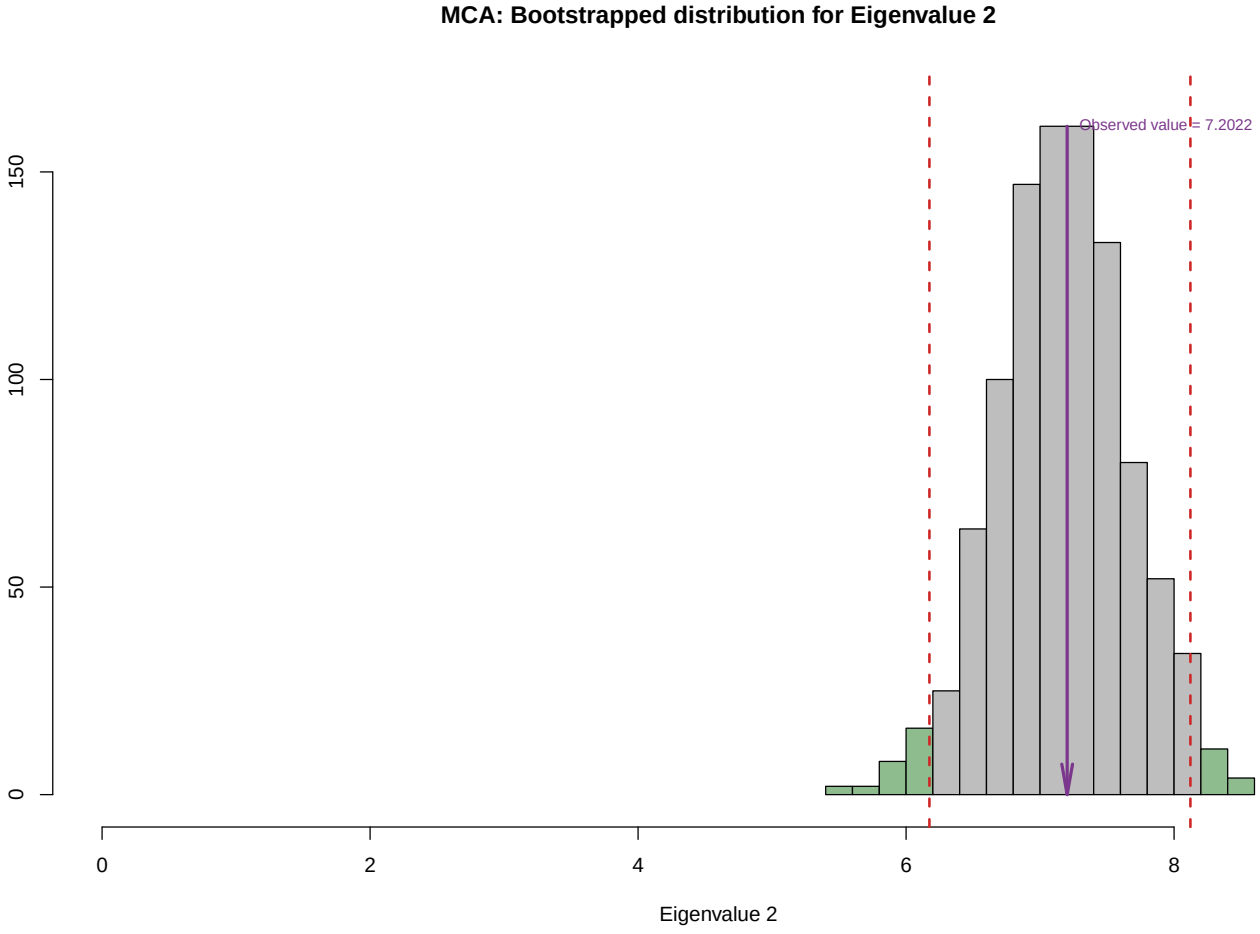



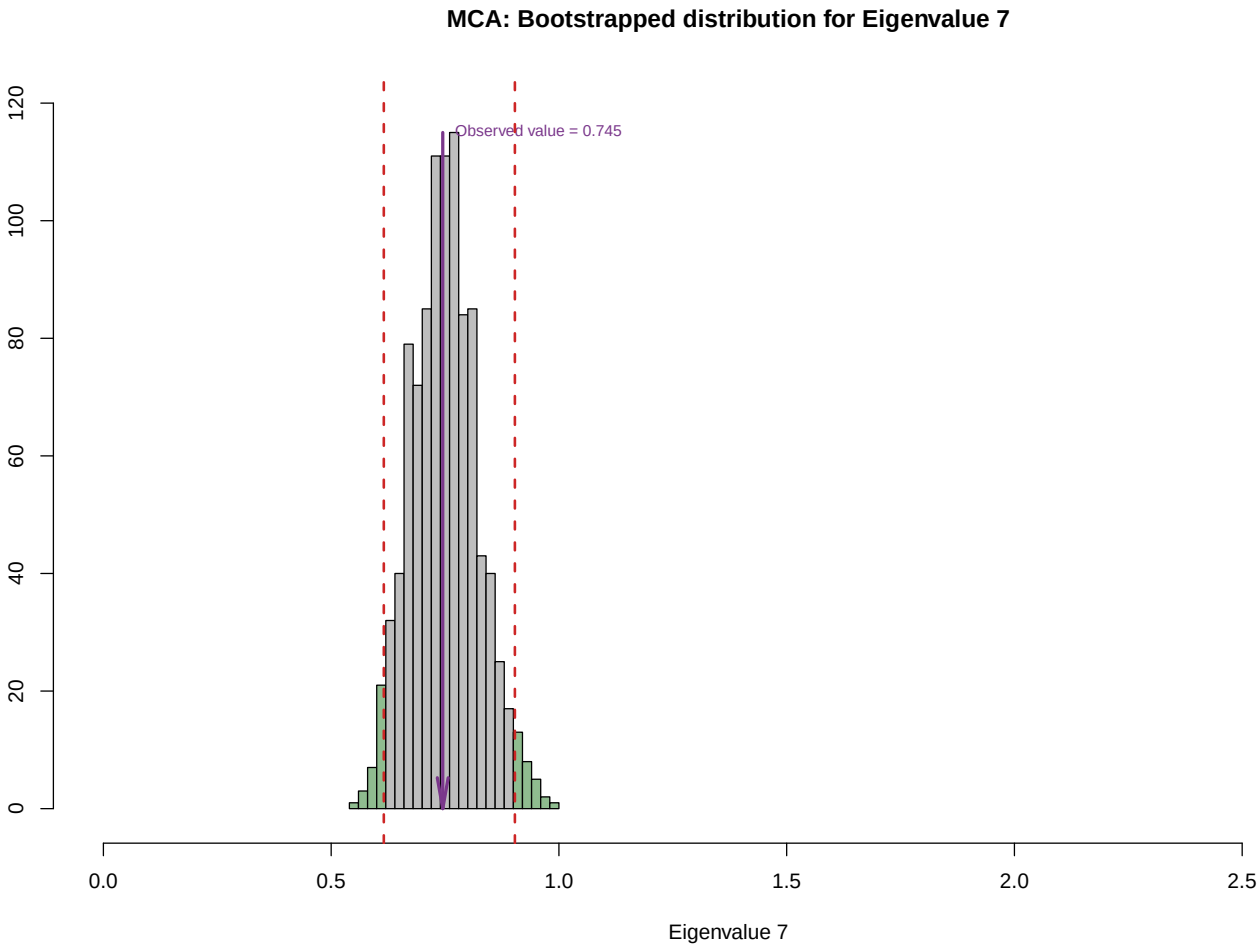

MCA: Monte Carlo (Parallel) Test for Eigenvalue 7
Observed value = 0.745
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
Eigenvalue 7

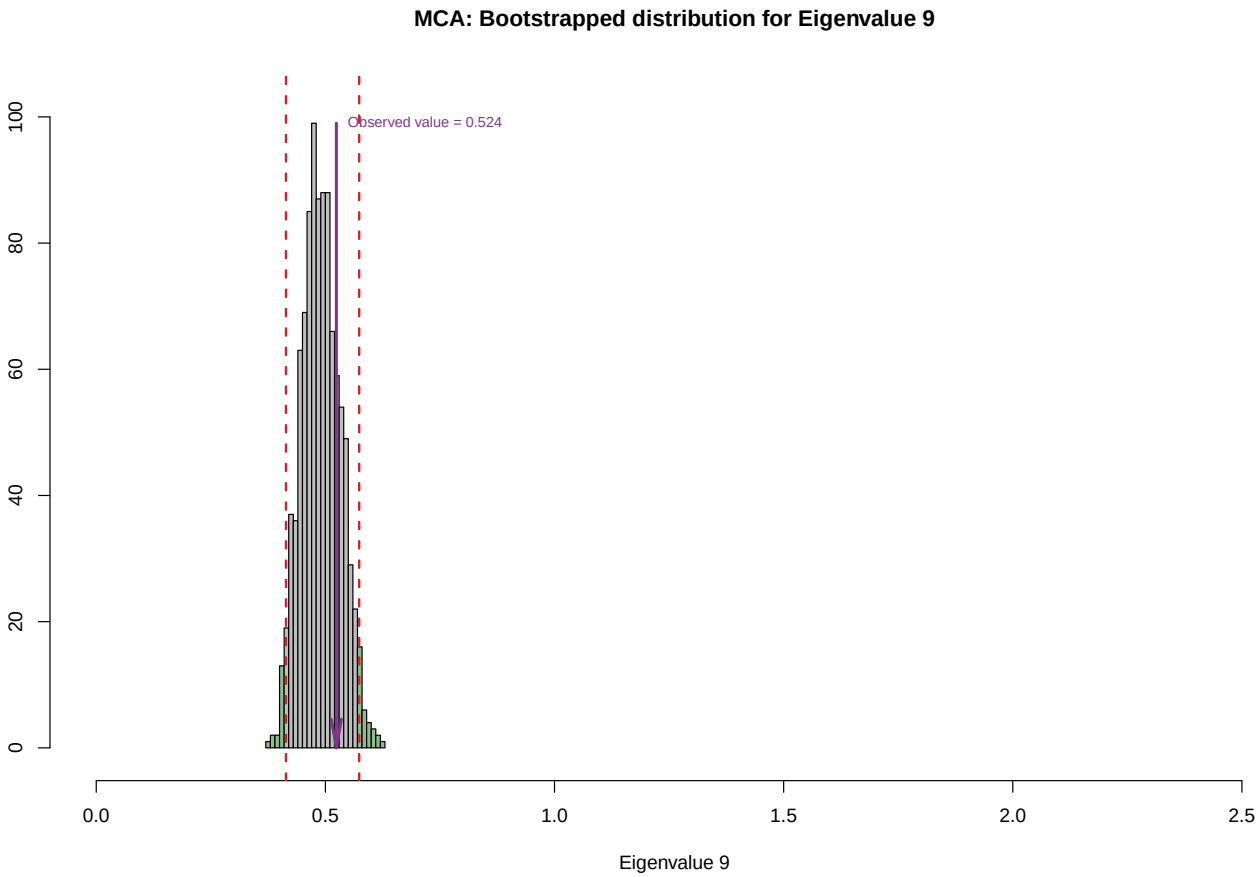

MCA: Monte Carlo (Parallel) Test for Eigenvalue 9
Observed value = 0.524
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
Eigenvalue 9

## 4.13 Bootstrap Test




MCA: Bootstrapped distribution for Eigenvalue 7
Observed value = 0.745
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
Eigenvalue 7




## 4.14 Conclusion

| Methods | Unhappy | Normal | Very Happy | Reliability |
|---|---|---|---|---|
| **MCA** | warm summers, cold winters, high rain | N/A | Warm winter, cold summer, low rain | Components have significant contribution but convex hull has overlapping areas |

# Part III

# Two / Multi Table Analysis for Country Environment Dataset

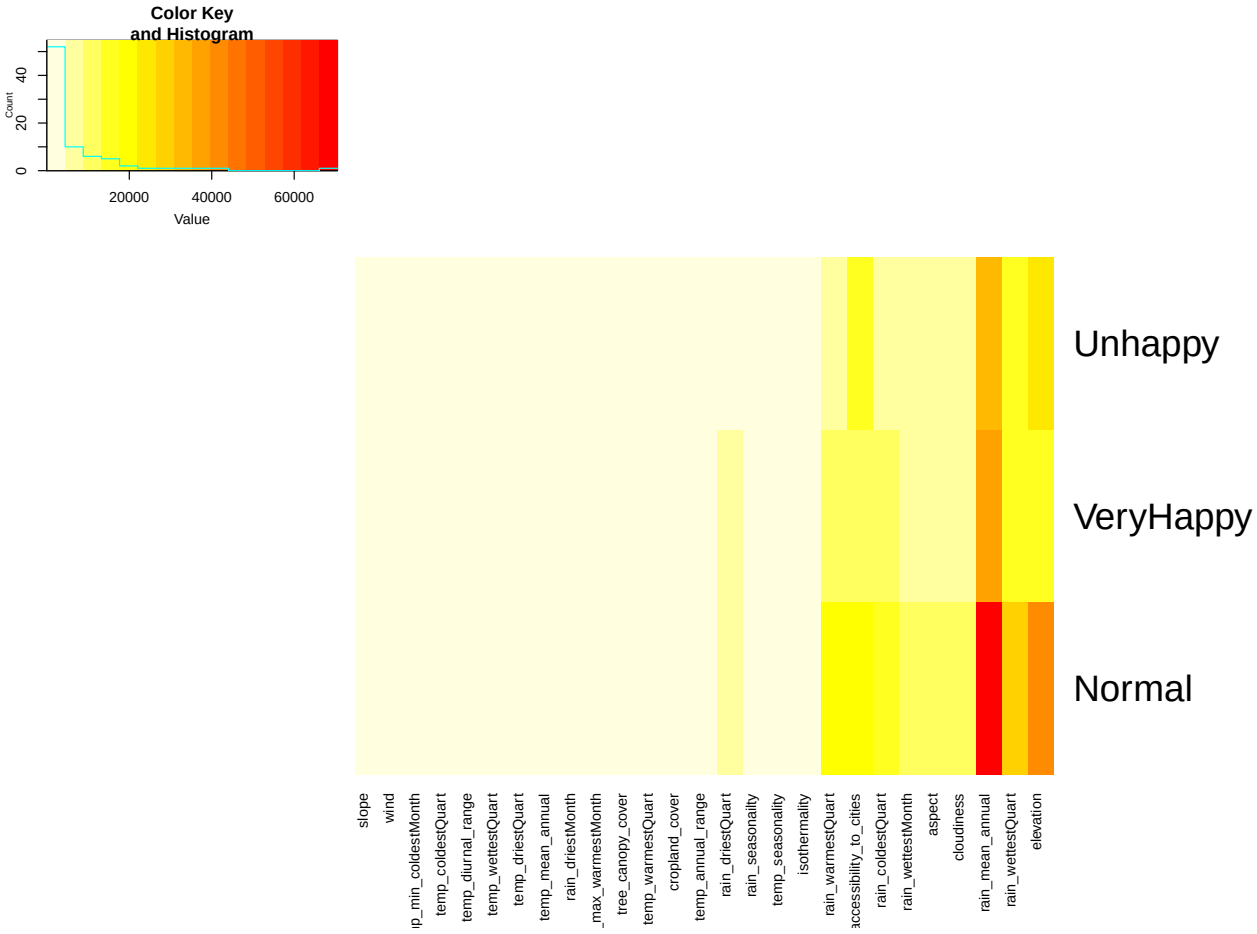
# Chapter 5

# Barycentric Discriminant Analysis

## 5.1 Description

Barycentric discriminant analysis(BADA) is a robust version of discriminant analysis that is used to assign, top re-defined groups(also called categories), observations described by multiple variables. By contrast with traditional discriminant analysis, BADA can be used even when the number of observations is smaller than the number of variables. This makes BADA particularly suited for the analysis of Big Data.

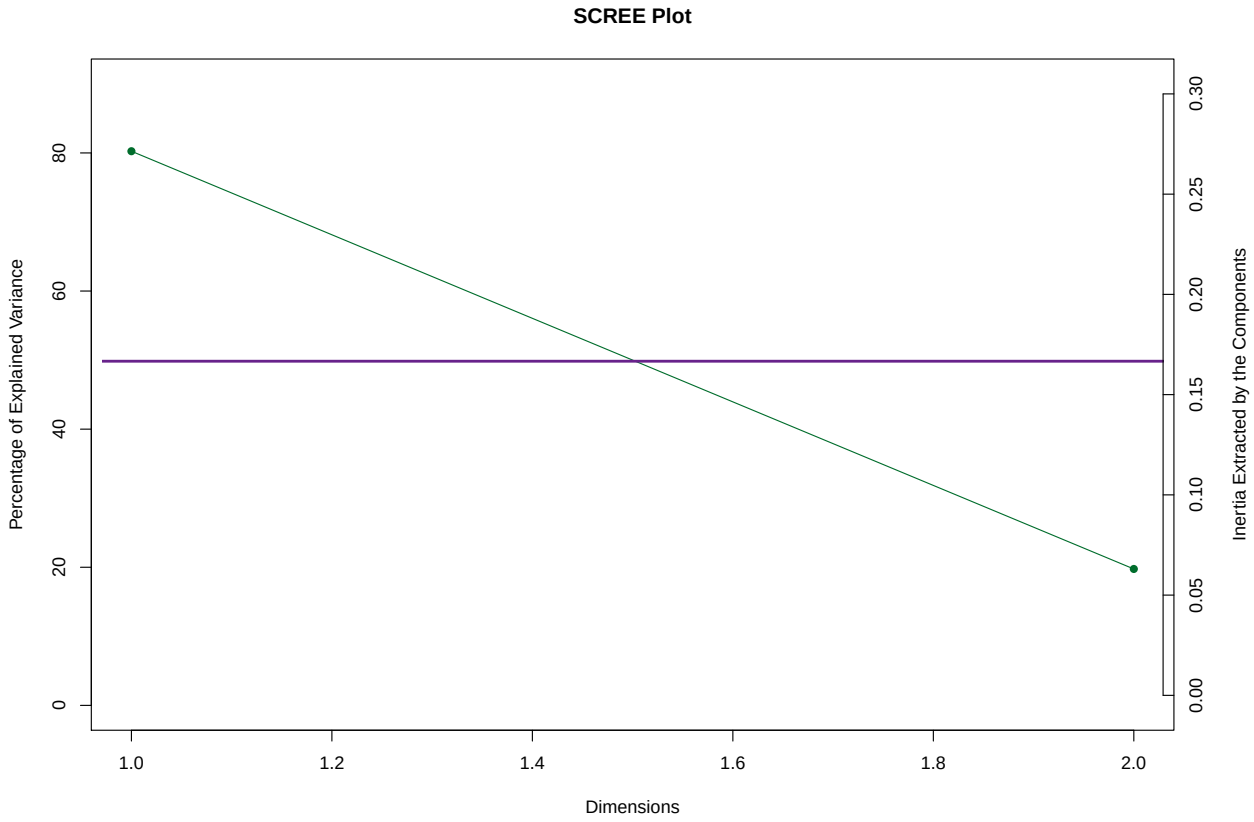
## 5.2 Heatmap



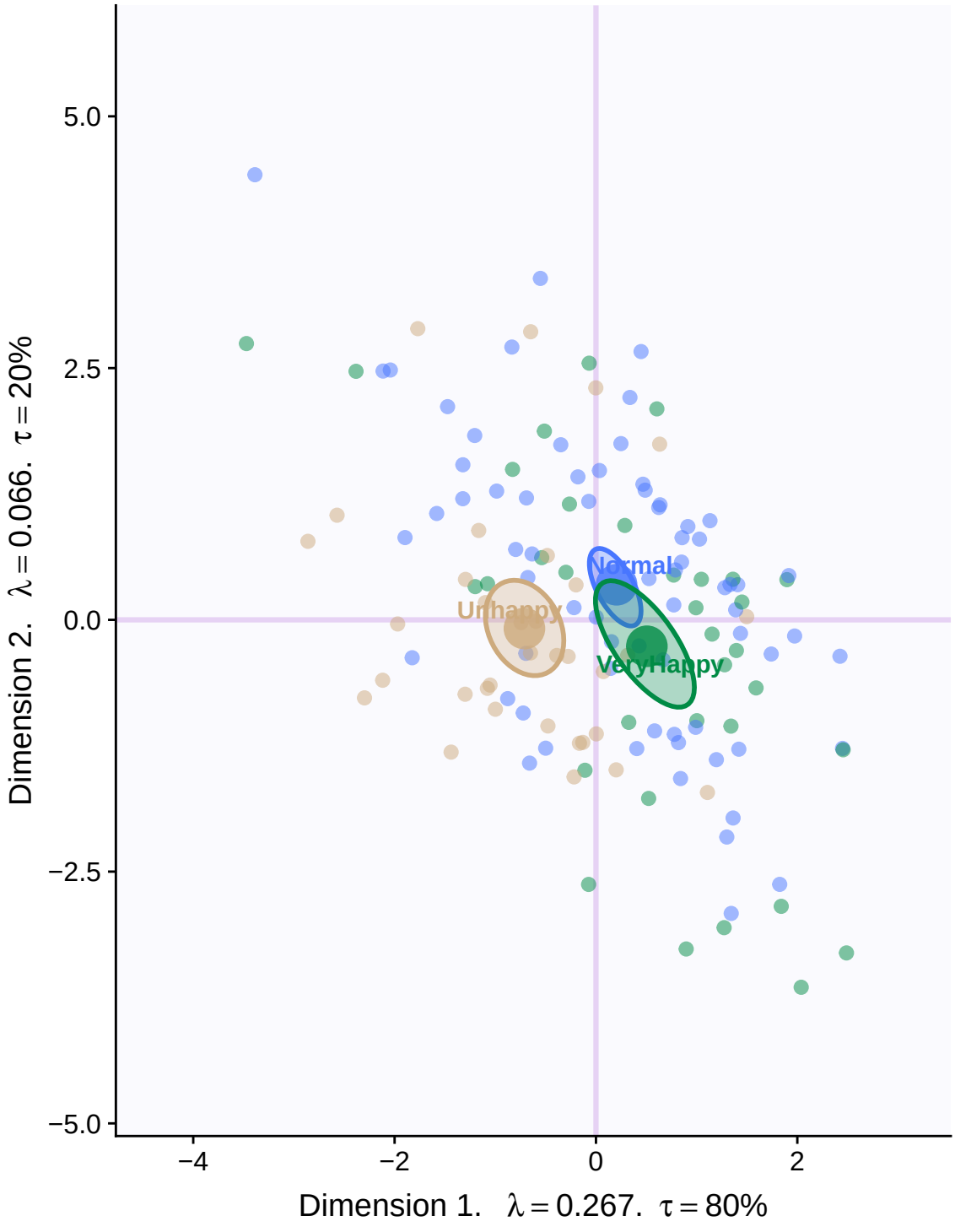

## 5.3 Scree Plot

Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

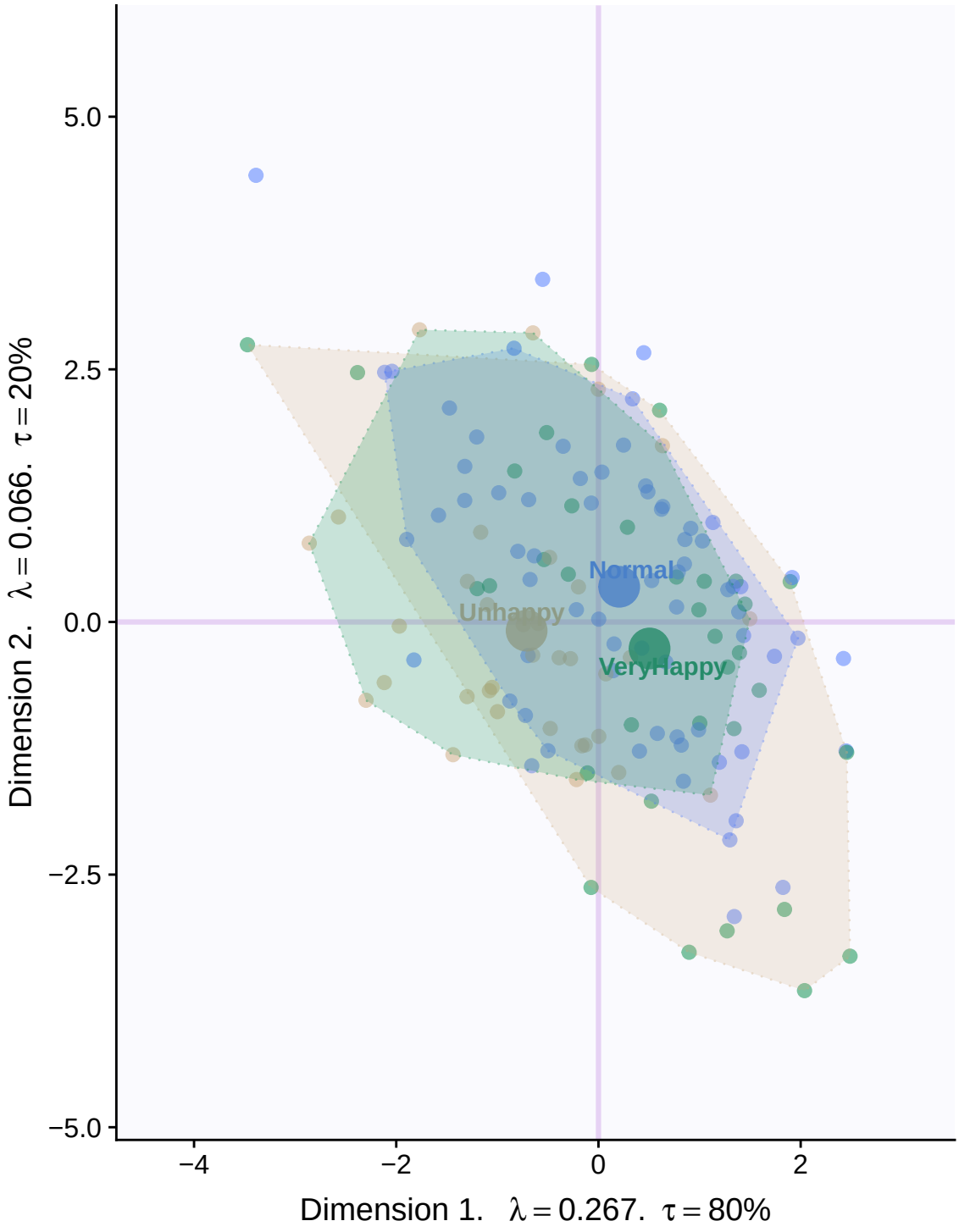
P.S. The most contribution component may not always be most useful for a given research question.


SCREE Plot
Percentage of Explained Variance
0
20
40
60
80
Inertia Extracted by the Components
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1.0
1.2
1.4
1.6
1.8
2.0
Dimensions

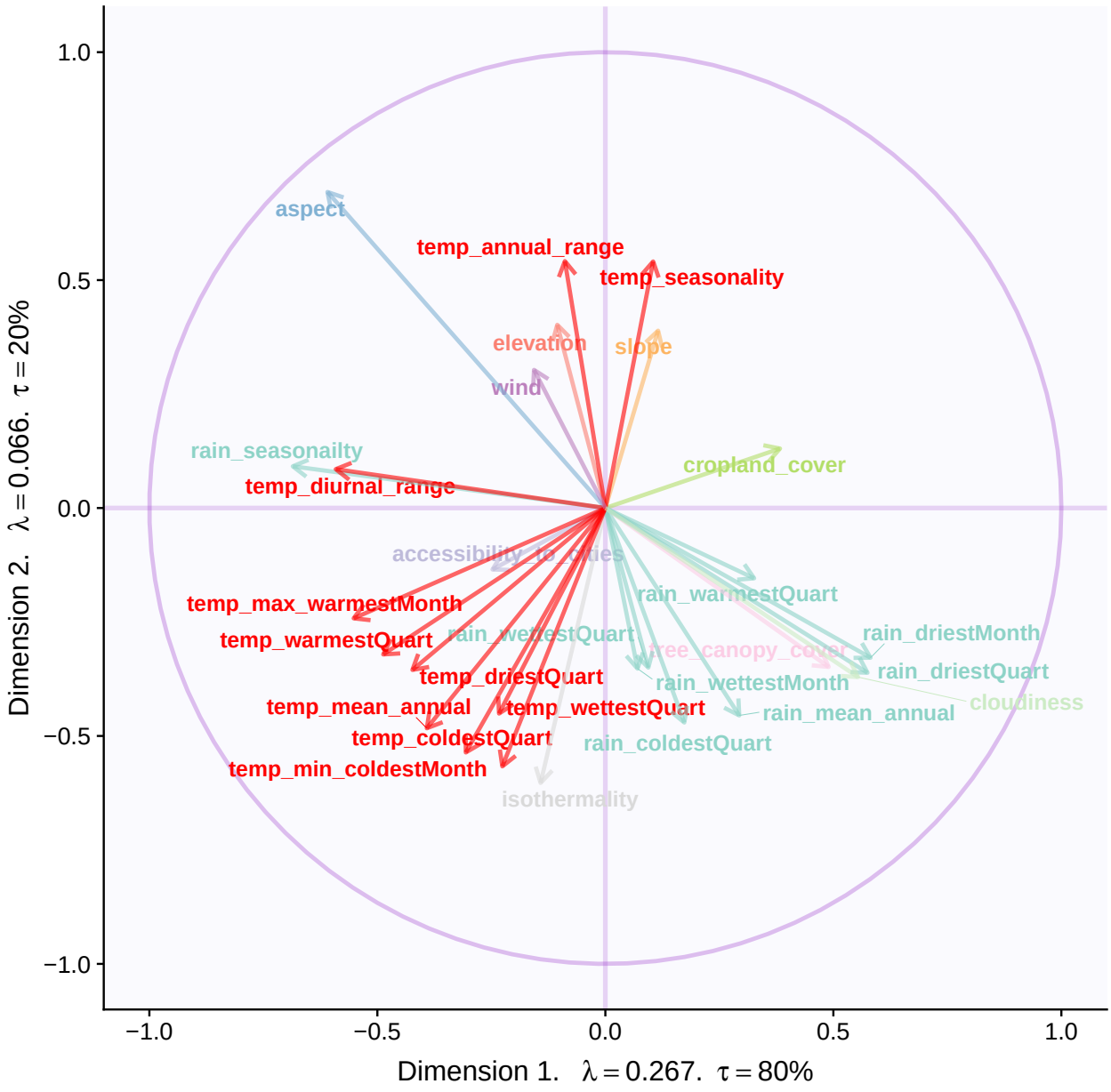
## 5.4 Factor Scores




- With Tolerance Interval

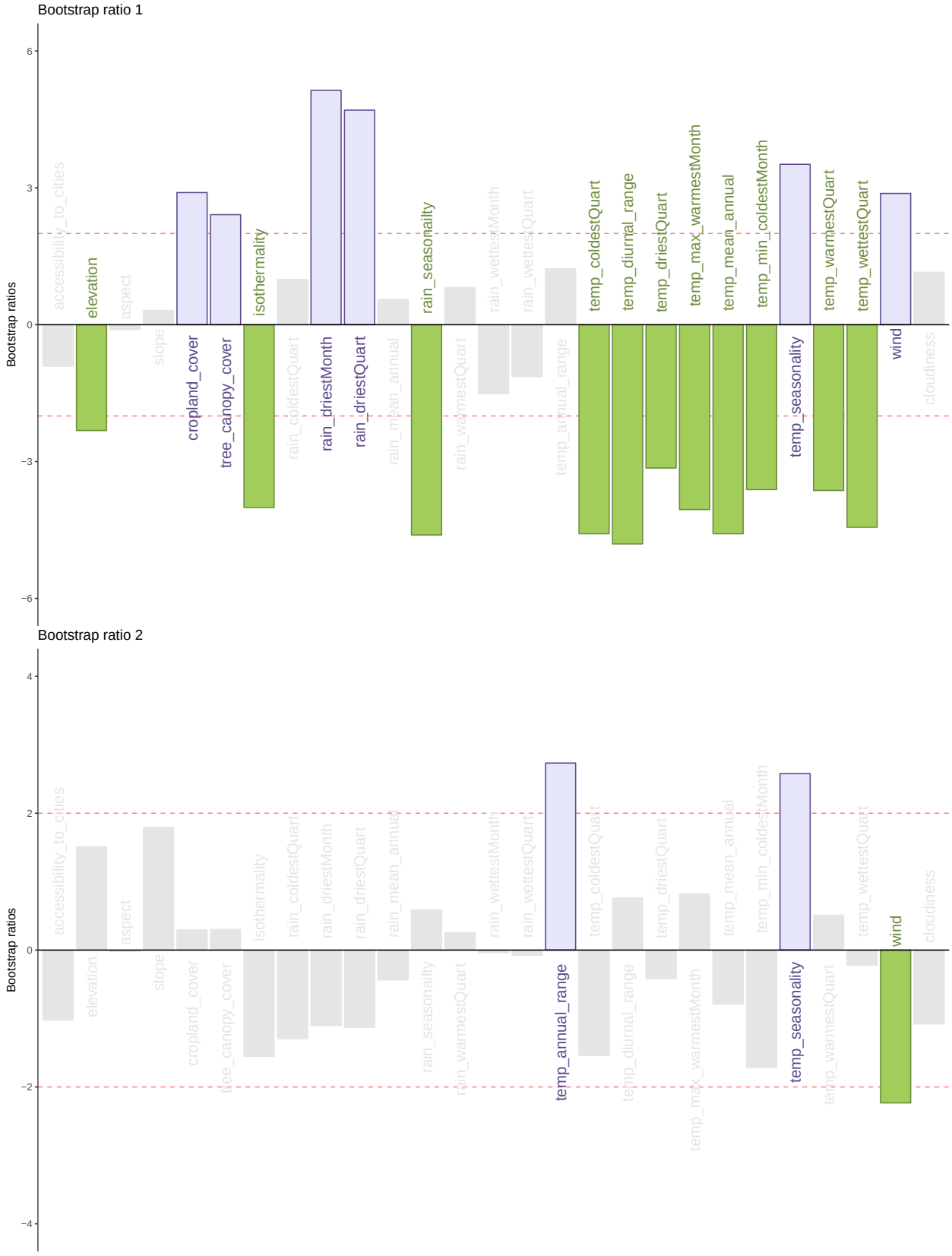

5.0
2.5
0.0
−2.5
−5.0
−4
−2
0
2
Dimension 2. λ = 0.066. τ = 20%
Dimension 1. λ = 0.267. τ = 80%
Normal
Unhappy
VeryHappy

## 5.5 Loadings

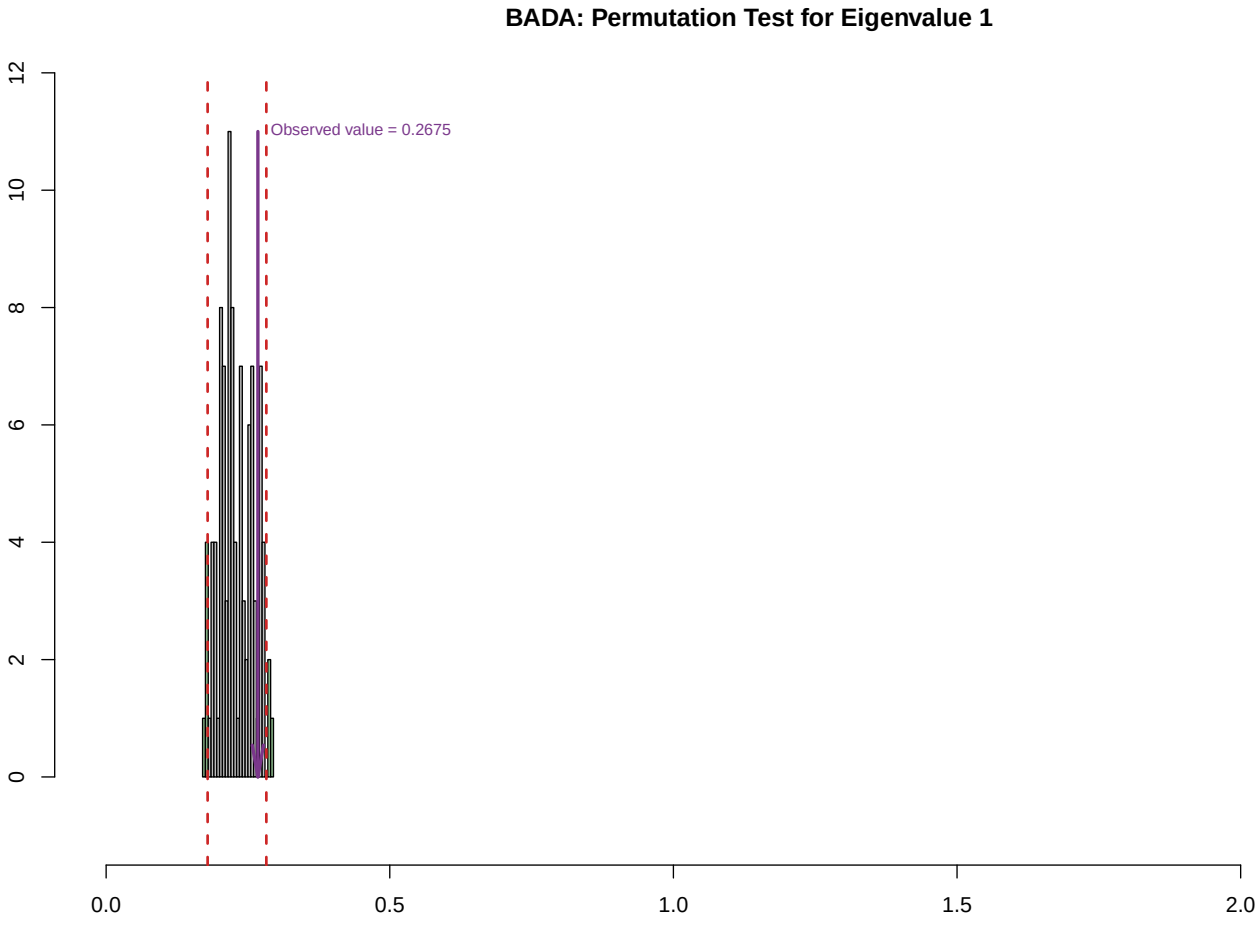


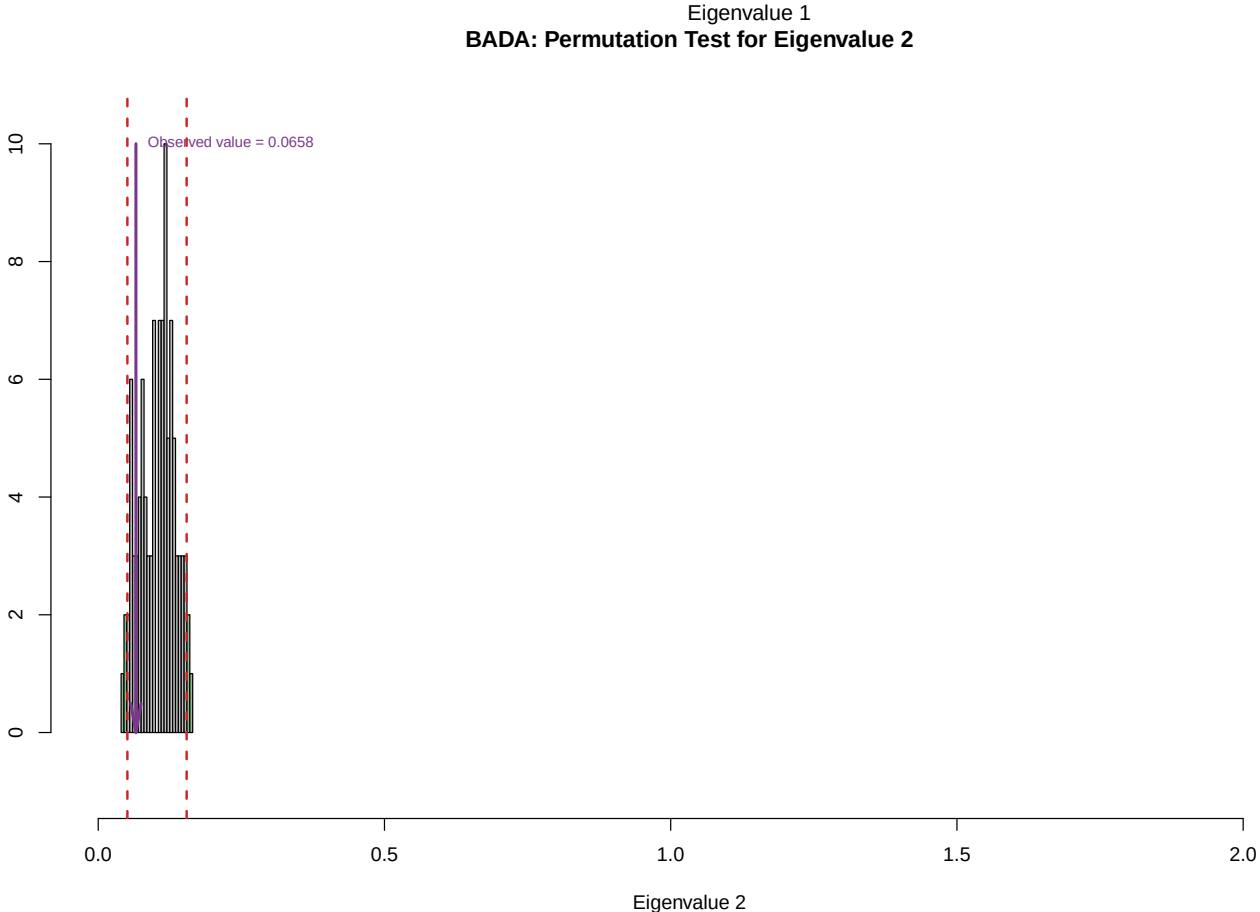

## 5.6 Most Contributing Variables

- With Bootstrap Ratio

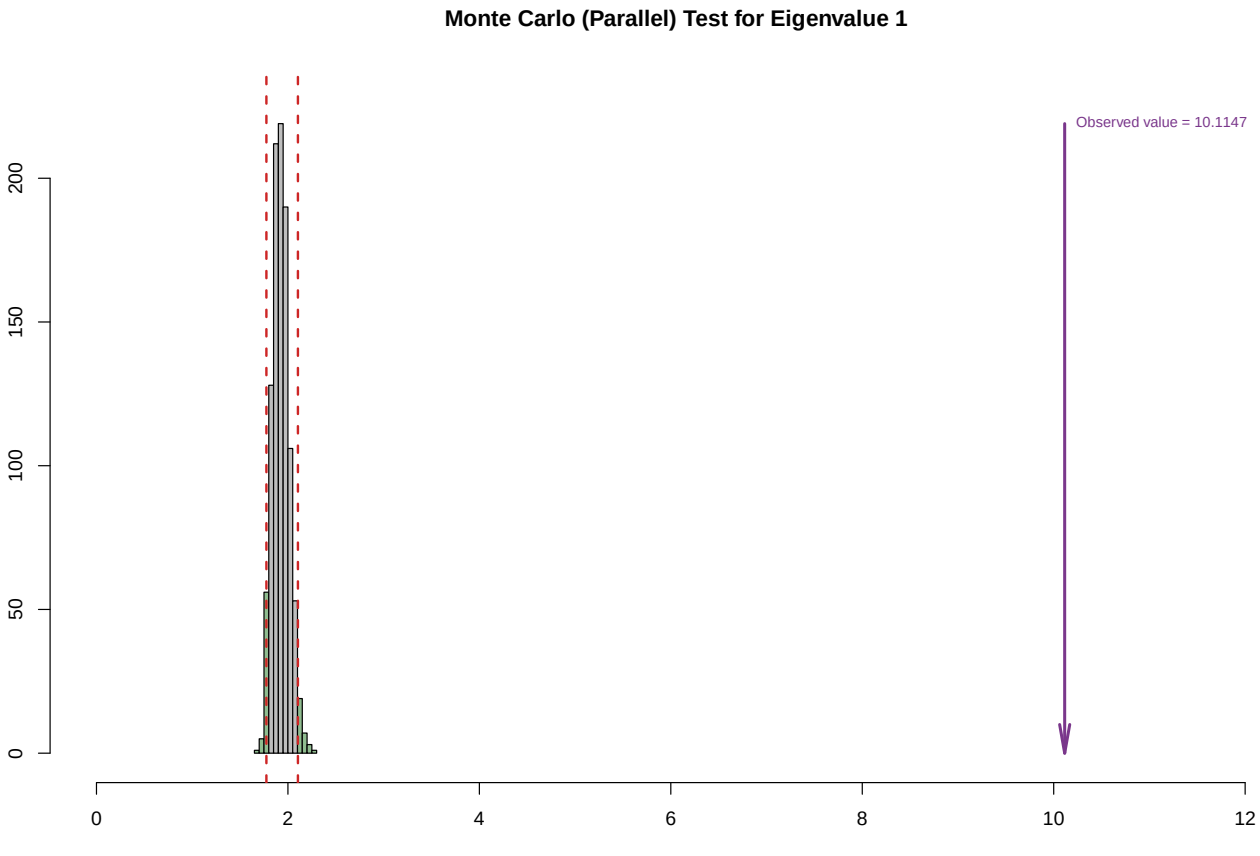

Bootstrap ratio 1
Bootstrap ratios
6
3
0
−3
−6
accessibility_to_cities
elevation
aspect
slope
cropland_cover
tree_canopy_cover
isothermality
rain_coldestQuart
rain_driestMonth
rain_driestQuart
rain_mean_annual
rain_seasonality
rain_warmestQuart
rain_wettestMonth
rain_wettestQuart
temp_annual_range
temp_coldestQuart
temp_diurnal_range
temp_driestQuart
temp_max_warmestMonth
temp_mean_annual
temp_min_coldestMonth
temp_seasonality
temp_warmestQuart
temp_wettestQuart
wind
cloudiness
Bootstrap ratio 2
Bootstrap ratios
4
2
0
−2
−4
accessibility_to_cities
elevation
aspect
slope
cropland_cover
tree_canopy_cover
isothermality
rain_coldestQuart
rain_driestMonth
rain_driestQuart
rain_mean_annual
rain_seasonality
rain_warmestQuart
rain_wettestMonth
rain_wettestQuart
temp_annual_range
temp_coldestQuart
temp_diurnal_range
temp_driestQuart
temp_max_warmestMonth
temp_mean_annual
temp_min_coldestMonth
temp_seasonality
temp_warmestQuart
temp_wettestQuart
wind
cloudiness

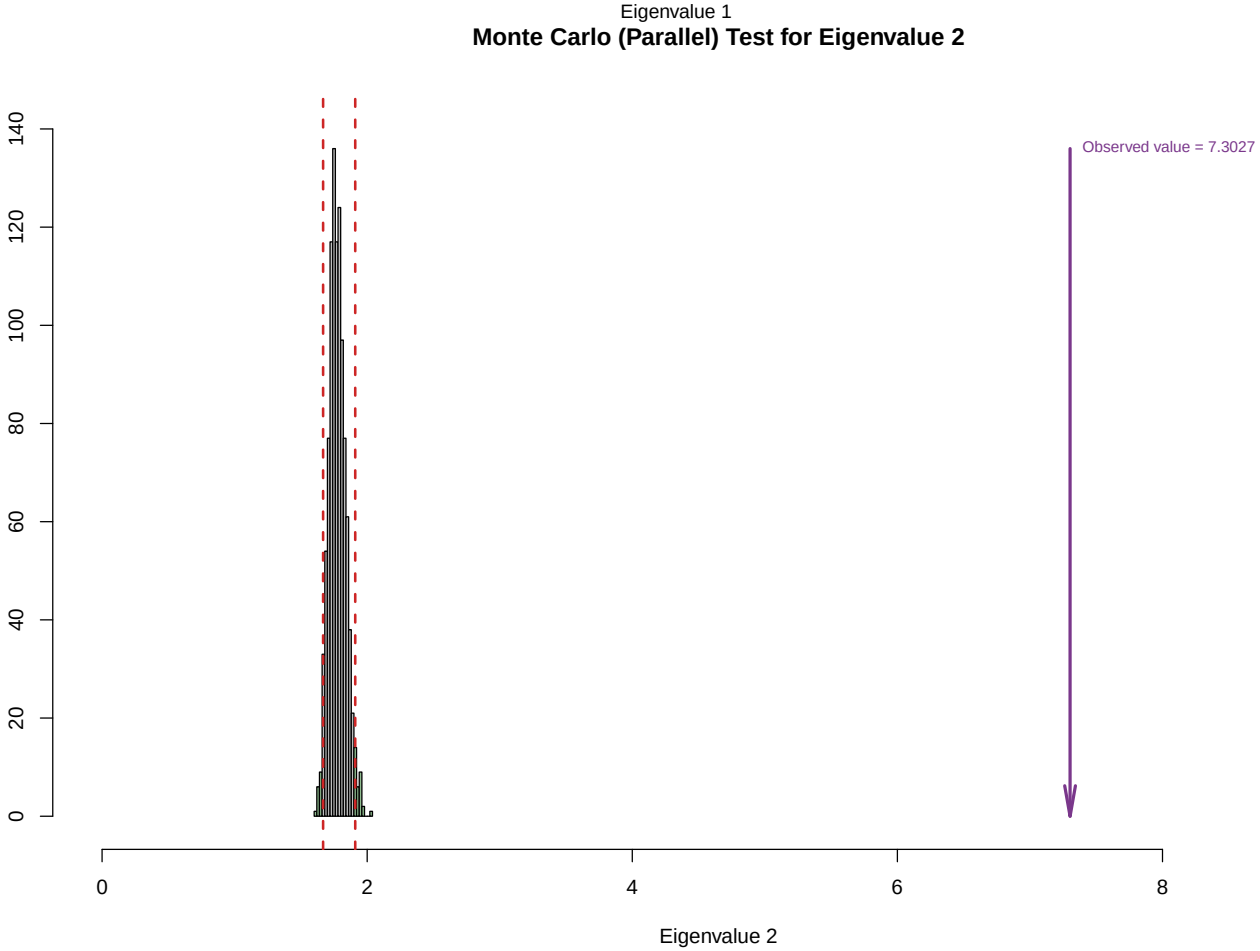
## 5.7 Permutation Test



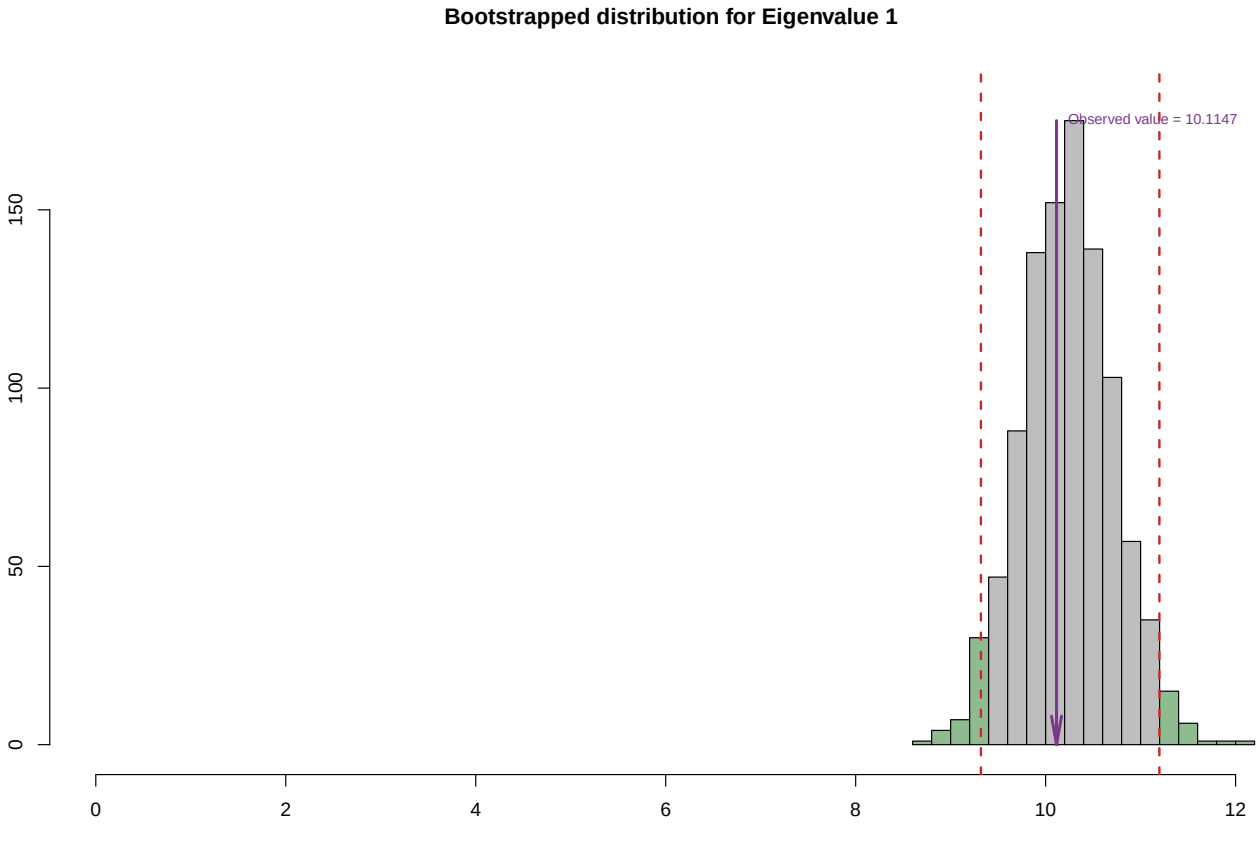



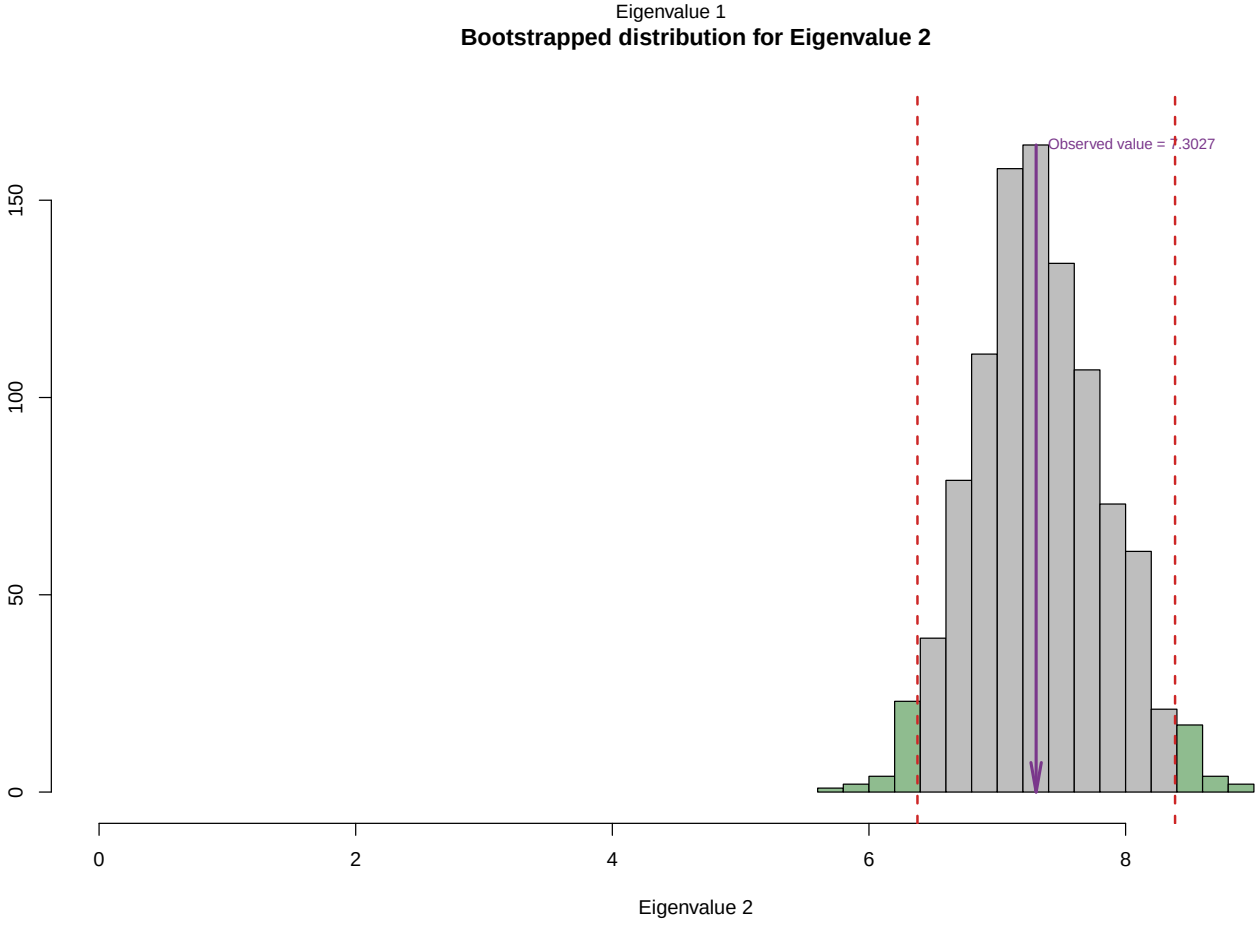
## 5.8 Parallet Test



## 5.9 Bootstrap Test



## 5.10 Conclusion

| Methods | Unhappy | Normal | Very Happy | Reliability |
|---|---|---|---|---|
| **BADA** | Temp | Rain | Rain | Components have significant contribution but convex hull has overlapping areas |

# Chapter 6

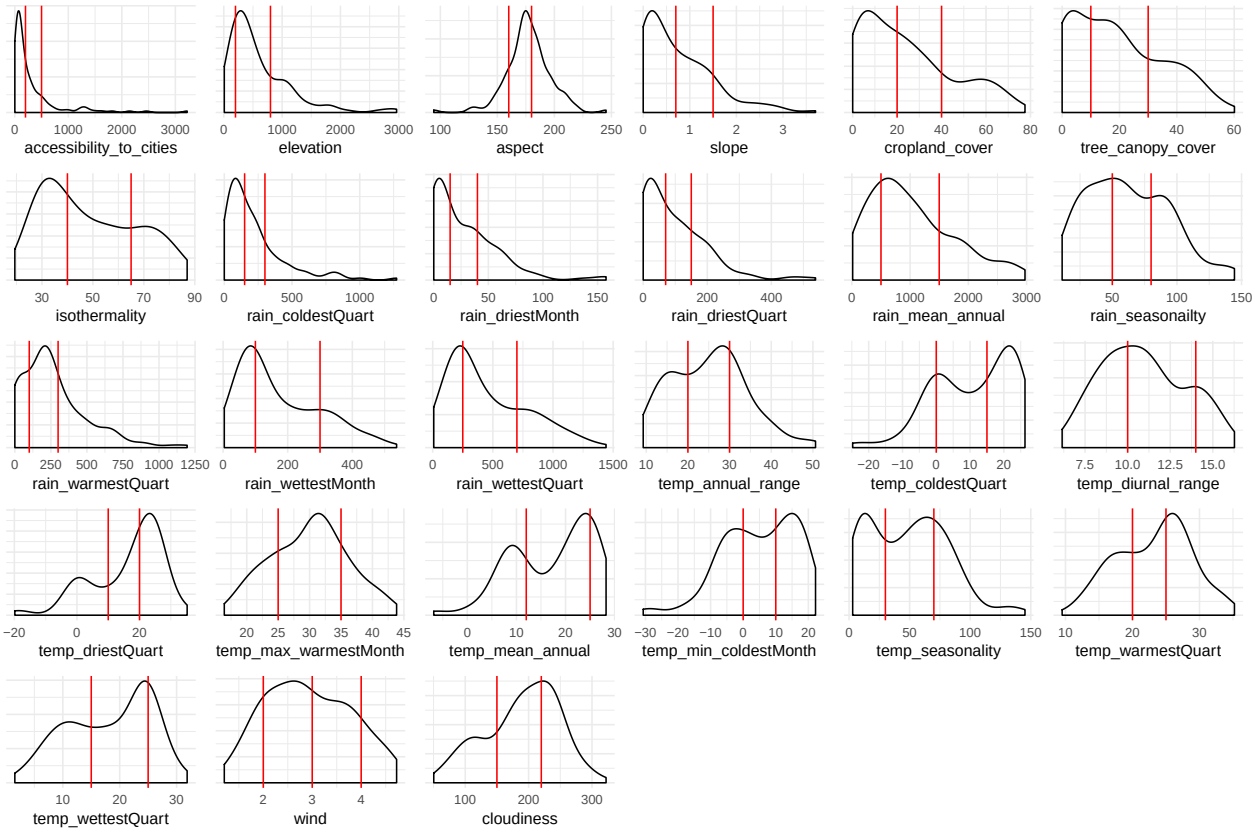
# Discriminant Correspondence Analysis

## 6.1 Description

As the name indicates, discriminant correspondence analysis(DCA) is an extension of discriminant analysis (DA) and correspondence analysis (CA). Like discriminant analysis, the goal of DCA is to categorize observations in pre-defined groups, and like correspondence analysis,it is used with nominal variables. The main idea behind DCA is to represent each group by the sum of its observations and to perform a simple CA on the groups by variables matrix. The original observations are then projected as supplementary elements and each observation is assigned to the closest group. The comparison between the a priori and the a posteriori classifications can be used to assess the quality of the discrimination. A similar procedure can be used to assign new observations to categories. The stability of the analysis can be evaluated using cross-validation techniques such as jack knifing or bootstrapping.

## 6.2 Density plot

Let's observe the distribution of each variables to get an intuition of how we can bin these variables. It's important to have nearly equal number of observations in the each bin and to try to cut the variables in a way to so that each new binned distribution is nearly Gaussian. We can also verify that our binning is appropiate by calculating Spearman Correlation for each of original variable and binned variable, the correlation coefficient should be close to 1.




## 6.3 Binning

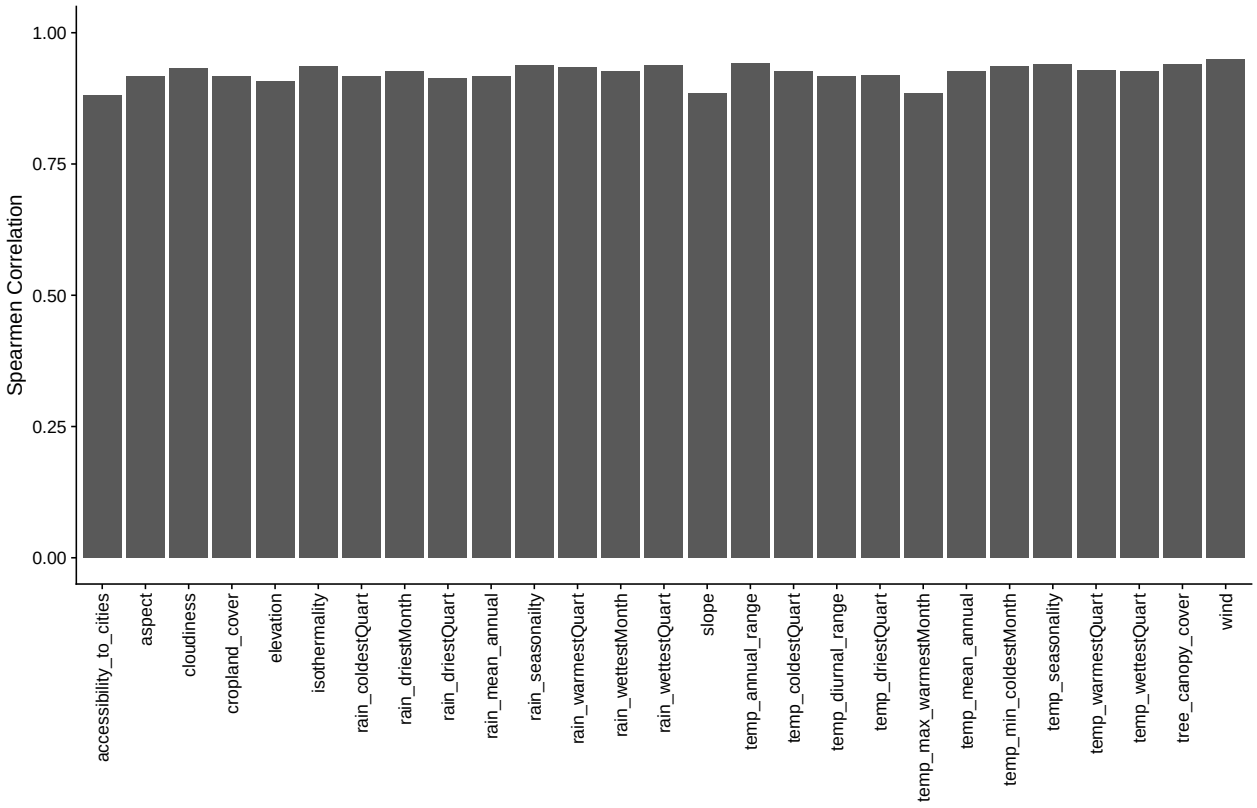
Structure of Data after binning based on above observation.

```
## 'data.frame':    137 obs. of  27 variables:
##  $ accessibility_to_cities: int  2 1 3 2 2 1 3 1 1 1 ...
##  $ elevation              : int  3 2 2 3 2 3 2 3 2 1 ...
##  $ aspect                 : int  3 3 3 2 1 3 3 2 1 2 ...
##  $ slope                  : int  3 3 1 1 1 3 1 2 2 1 ...
##  $ cropland_cover         : int  1 2 1 1 2 2 1 2 2 3 ...
##  $ tree_canopy_cover      : int  1 2 1 2 1 1 1 3 1 2 ...
##  $ isothermality          : int  1 1 2 2 2 1 2 1 1 2 ...
##  $ rain_coldestQuart      : int  1 3 1 1 1 1 1 2 1 1 ...
##  $ rain_driestMonth       : int  1 3 1 1 2 2 1 3 2 1 ...
##  $ rain_driestQuart       : int  1 2 1 1 1 1 1 3 1 1 ...
##  $ rain_mean_annual       : int  1 2 1 2 2 2 1 2 1 3 ...
##  $ rain_seasonailty       : int  3 1 2 3 1 1 2 1 1 3 ...
##  $ rain_warmestQuart      : int  1 2 1 3 2 2 2 3 1 3 ...
##  $ rain_wettestMonth      : int  1 2 1 2 1 1 1 2 1 3 ...
##  $ rain_wettestQuart      : int  1 2 1 2 1 1 1 2 1 3 ...
##  $ temp_annual_range      : int  3 2 3 2 2 3 2 2 3 2 ...
##  $ temp_coldestQuart      : int  1 2 2 3 2 1 2 1 2 3 ...
##  $ temp_diurnal_range     : int  3 1 3 2 2 2 2 1 1 1 ...
##  $ temp_driestQuart       : int  3 2 3 2 2 1 2 1 2 2 ...
##  $ temp_max_warmestMonth  : int  2 2 3 2 2 1 3 1 2 2 ...
##  $ temp_mean_annual       : int  1 1 2 2 2 1 2 1 1 3 ...
```

```
##  $ temp_min_coldestMonth  : int  1 1 2 2 2 1 2 1 1 3 ...
##  $ temp_seasonality       : int  3 2 3 1 2 3 2 2 3 2 ...
##  $ temp_warmestQuart      : int  2 1 3 2 2 1 3 1 2 3 ...
##  $ temp_wettestQuart      : int  1 1 2 2 2 1 2 1 1 3 ...
##  $ wind                   : int  3 2 4 2 4 1 4 2 2 2 ...
##  $ cloudiness             : int  1 2 1 2 2 2 1 3 2 2 ...
```

## 6.4 Spearman Correlation

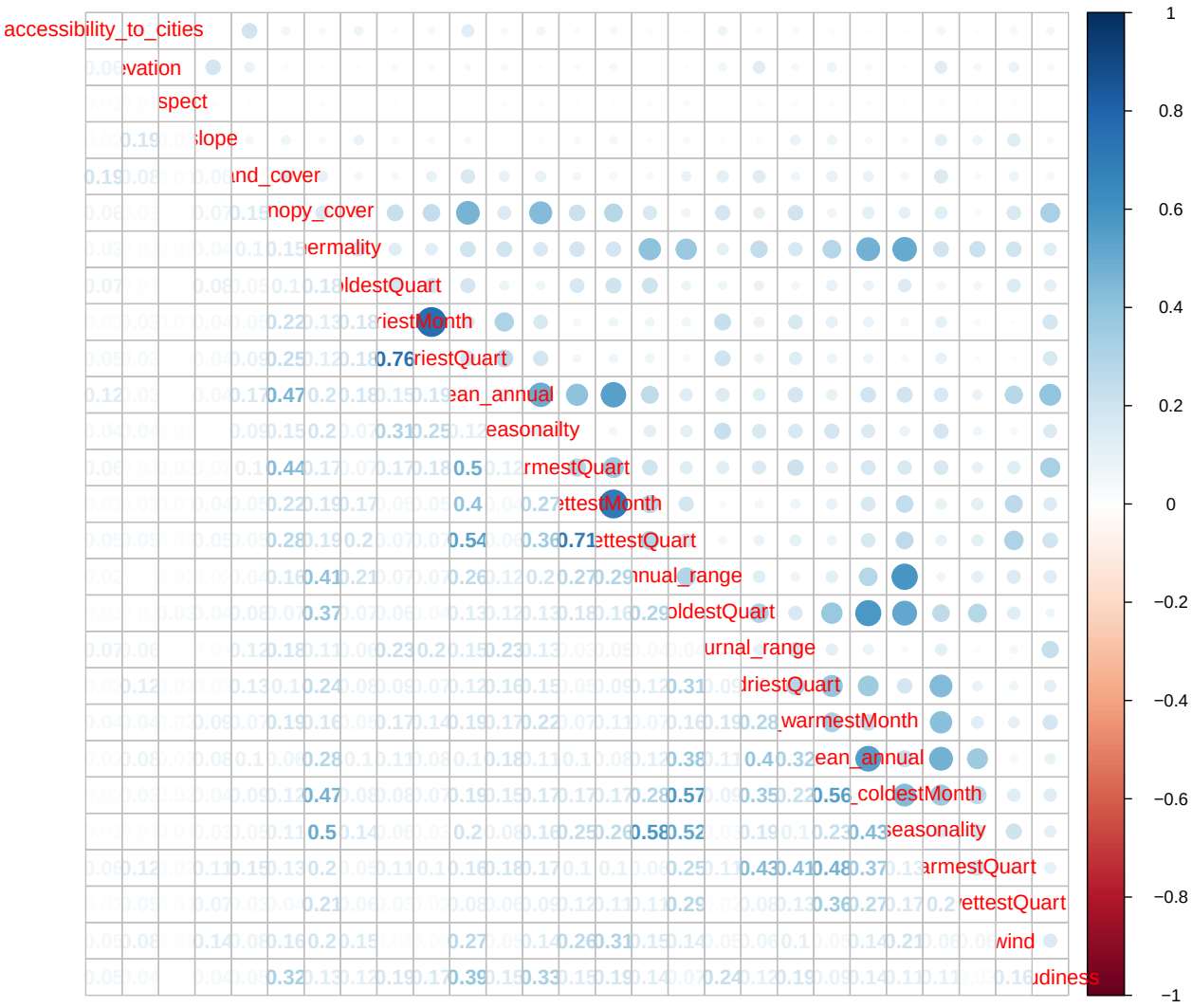
Let's observe correlation between original data and binned data to make sure that neither the correlation ceofficient is too low or perfect.




## 6.5 Heatmap

- For binned data

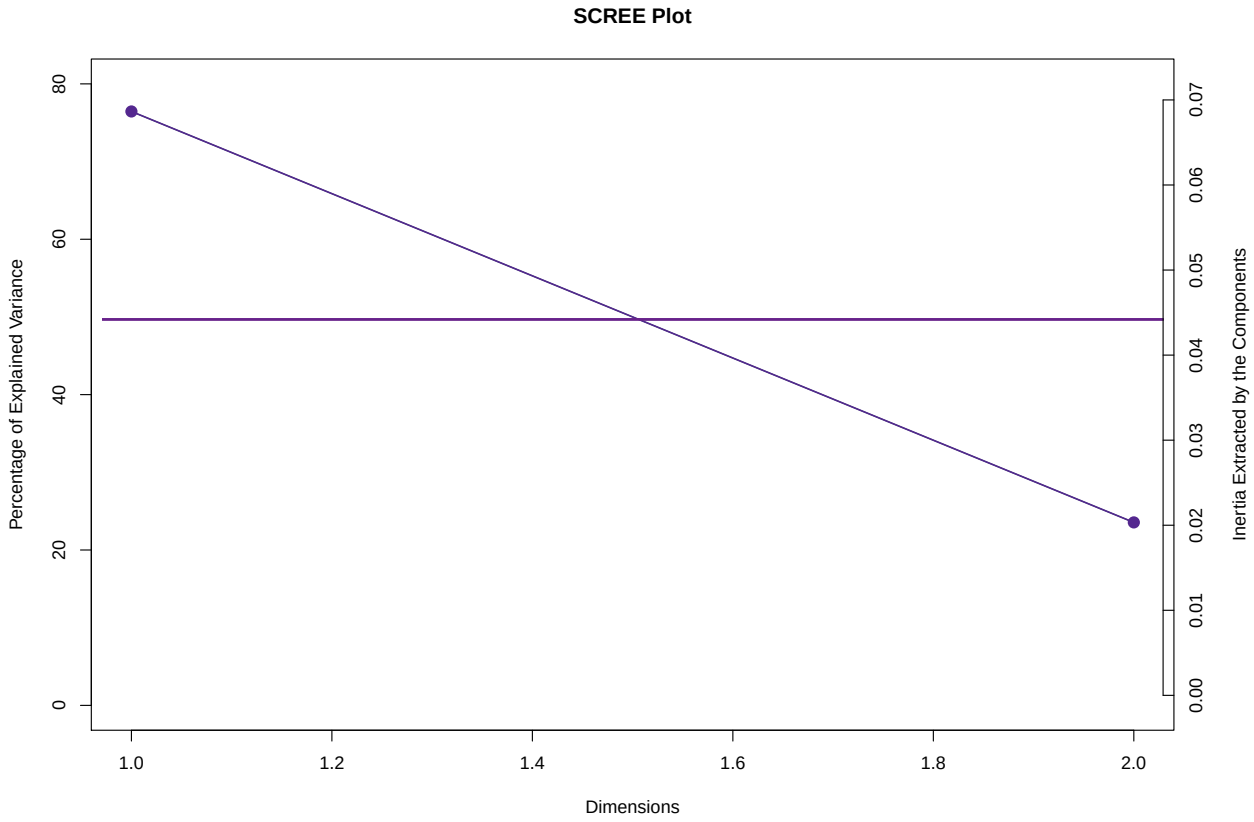
Visually analyze multicollinearity in the system of the original data



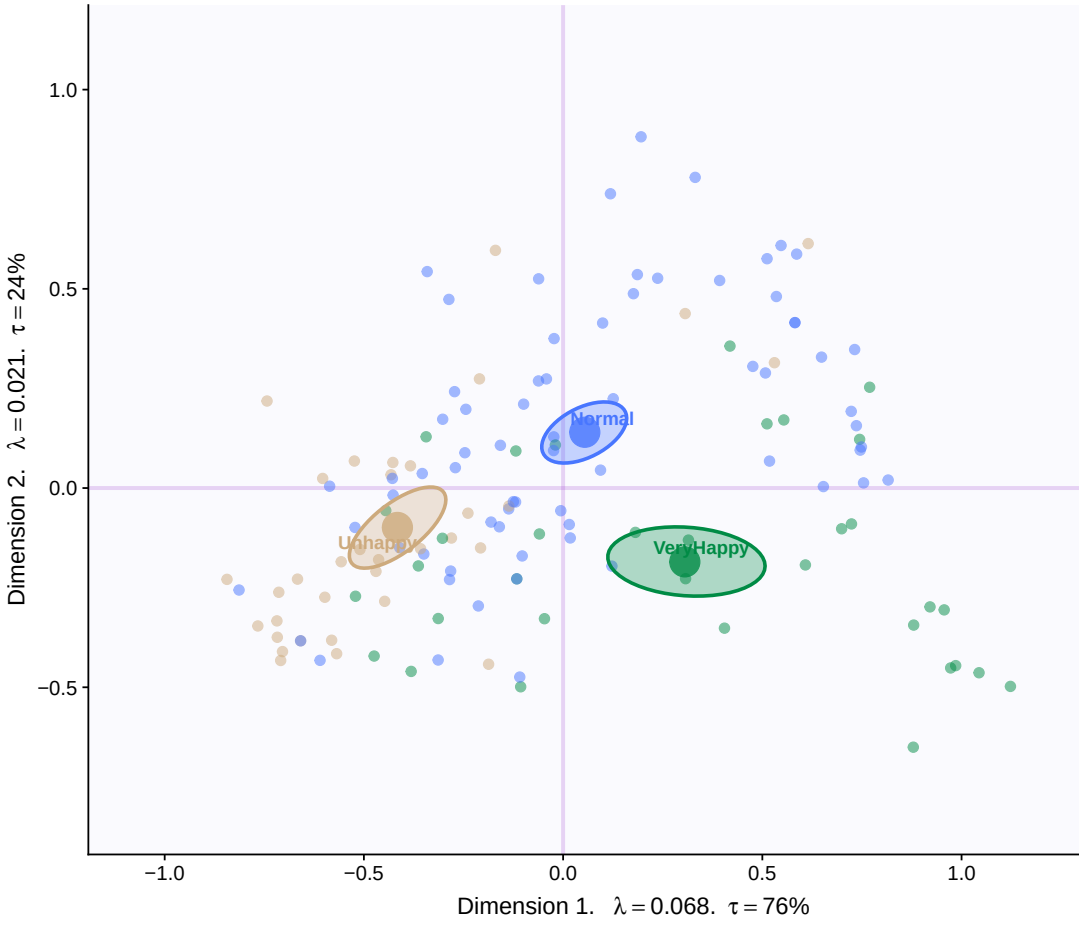

## 6.6 Scree Plot

Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

P.S. The most contribution component may not always be most useful for a given research question.

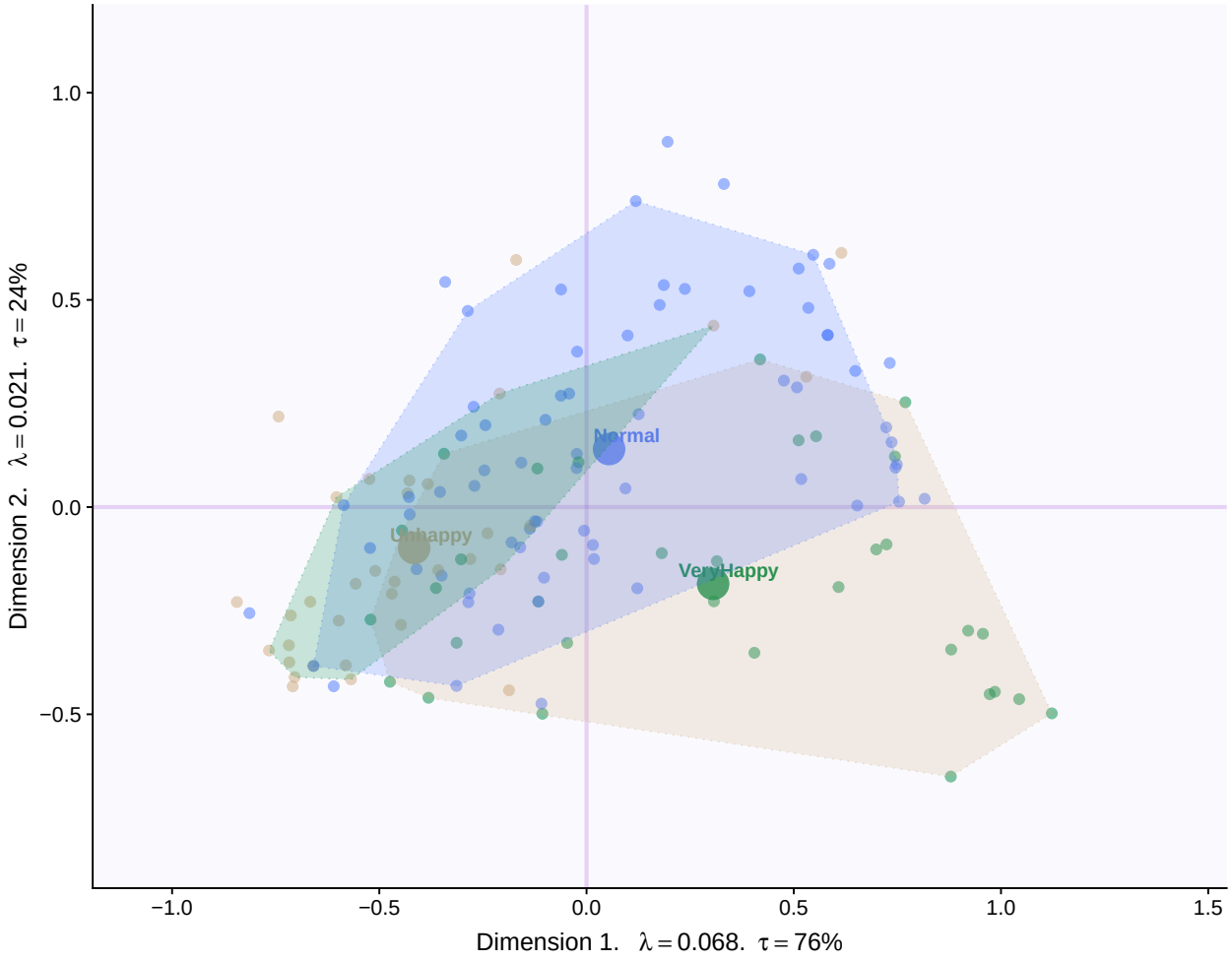

SCREE Plot
Percentage of Explained Variance
0
20
40
60
80
Inertia Extracted by the Components
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
1.0
1.2
1.4
1.6
1.8
2.0
Dimensions

## 6.7 Factor Scores

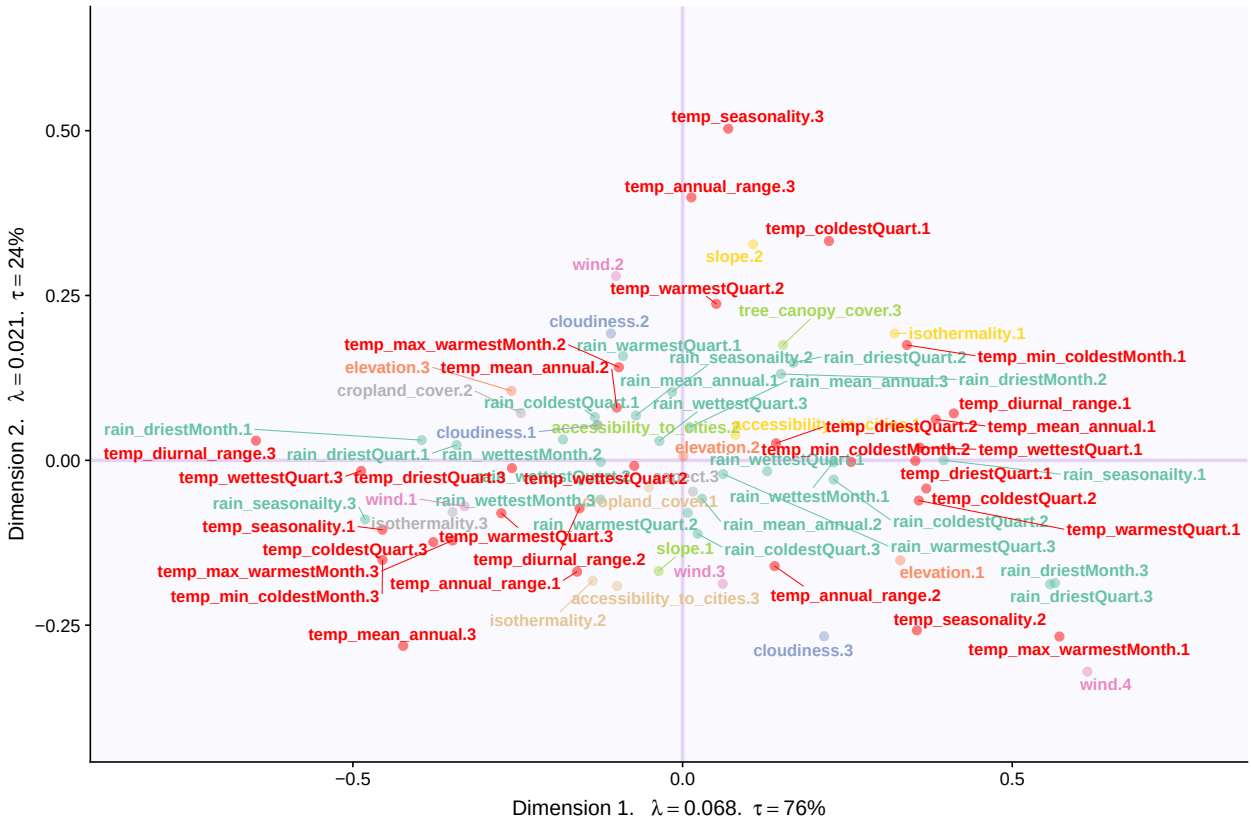



- With Confidence Interval

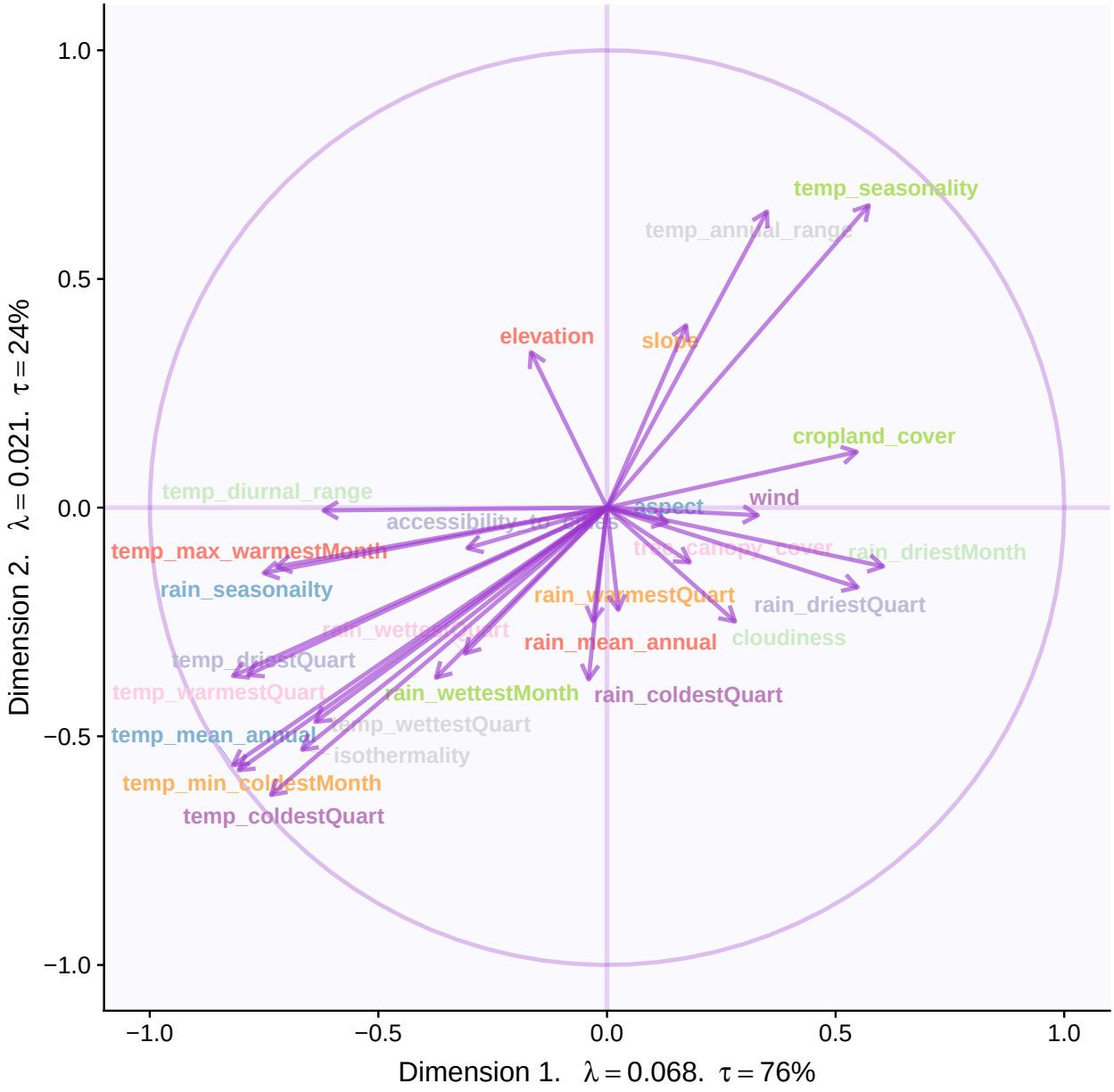
- With Tolerance Interval


Normal
Unhappy
VeryHappy
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Dimension 2. λ = 0.021. τ = 24%
Dimension 1. λ = 0.068. τ = 76%

## 6.8 Loadings


temp_seasonality.3
temp_annual_range.3
temp_coldestQuart.1
slope.2
wind.2
temp_warmestQuart.2
tree_canopy_cover.3
cloudiness.2
isothermality.1
temp_max_warmestMonth.2
rain_warmestQuart.1
rain_seasonailty.2
rain_driestQuart.2
temp_min_coldestMonth.1
elevation.3
temp_mean_annual.2
rain_mean_annual.1
rain_mean_annual.3
rain_driestMonth.2
cropland_cover.2
rain_coldestQuart.1
rain_wettestQuart.3
temp_diurnal_range.1
rain_driestMonth.1
cloudiness.1
accessibility_to_cities.2
temp_driestQuart.2
temp_mean_annual.1
temp_diurnal_range.3
rain_driestQuart.1
rain_wettestMonth.2
elevation.2
temp_min_coldestMonth.2
temp_wettestQuart.1
rain_wettestQuart.1
temp_wettestQuart.3
temp_driestQuart.3
temp_wettestQuart.2
temp_driestQuart.1
rain_seasonailty.1
rain_seasonailty.3
wind.1
rain_wettestMonth.3
cropland_cover.1
rain_wettestMonth.1
temp_coldestQuart.2
temp_seasonality.1
isothermality.3
rain_warmestQuart.2
rain_mean_annual.2
rain_coldestQuart.2
temp_warmestQuart.1
temp_warmestQuart.3
temp_coldestQuart.3
slope.1
rain_coldestQuart.3
rain_warmestQuart.3
temp_diurnal_range.2
temp_max_warmestMonth.3
wind.3
elevation.1
rain_driestMonth.3
temp_annual_range.1
temp_min_coldestMonth.3
accessibility_to_cities.3
temp_annual_range.2
rain_driestQuart.3
isothermality.2
temp_seasonality.2
temp_mean_annual.3
cloudiness.3
temp_max_warmestMonth.1
wind.4
0.50
0.25
0.00
−0.25
−0.5
0.0
0.5
Dimension 2. λ = 0.021. τ = 24%
Dimension 1. λ = 0.068. τ = 76%

## 6.9 Loadings (correlation plot)

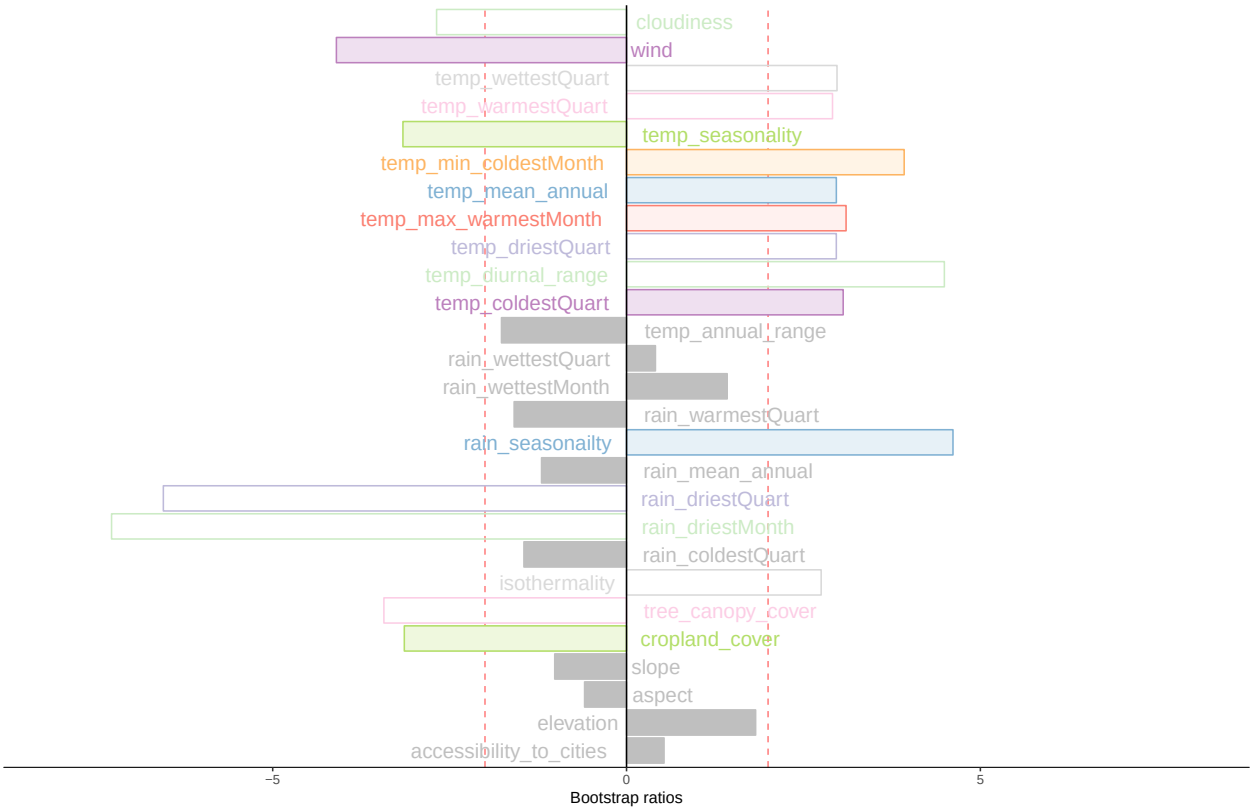



## 6.10 Most Contributing Variables (Inference)

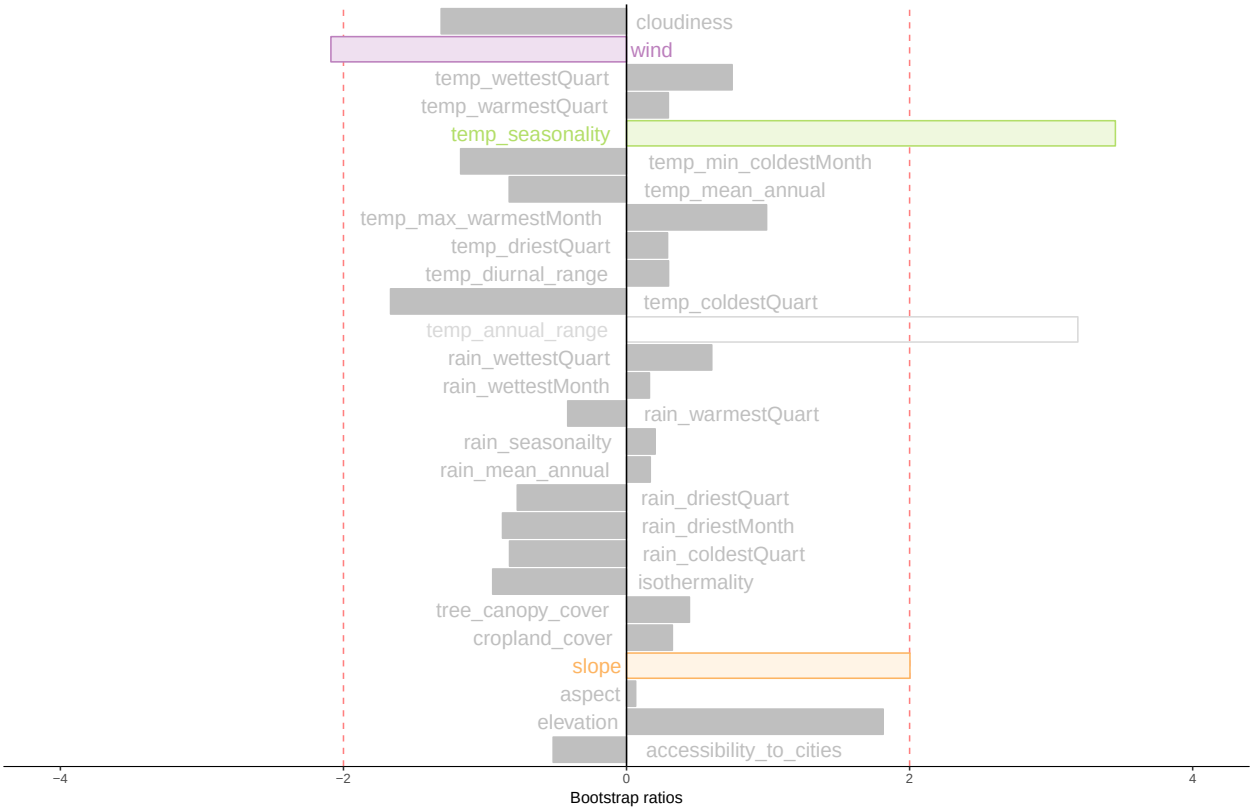
Let's plot variable contributions against each chosen components i.e. 1, 2.

- With Bootstrap Ratio

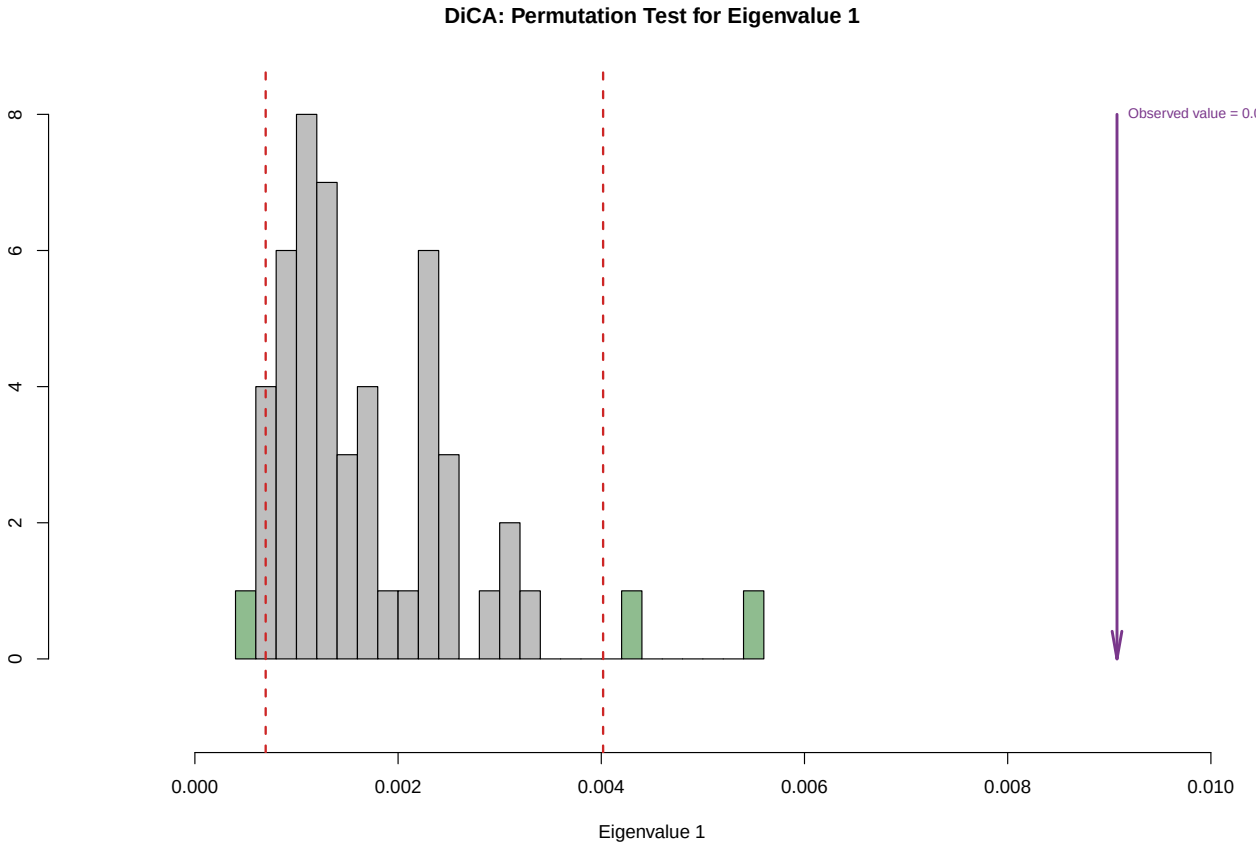
DiCA: Bootstrap ratio 1

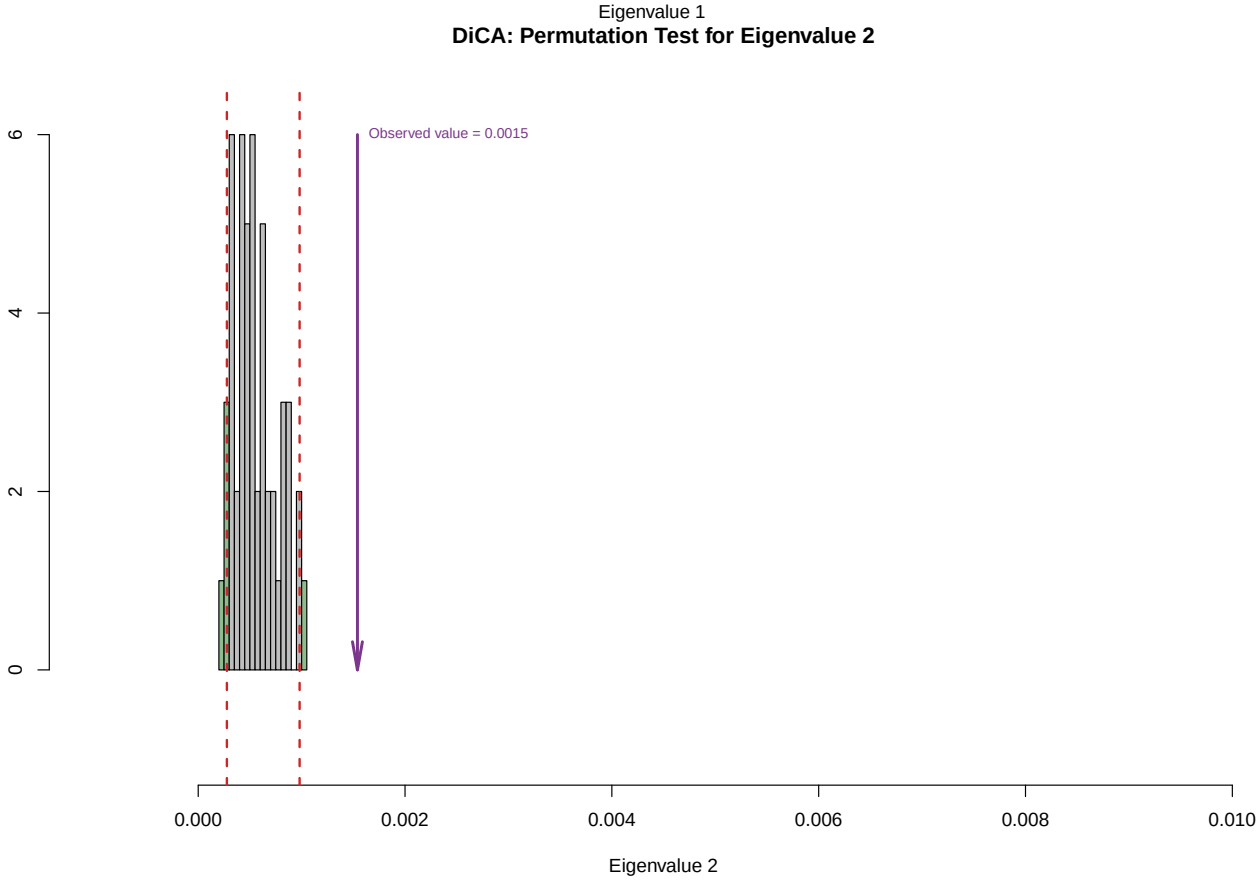



DiCA: Bootstrap ratio 2

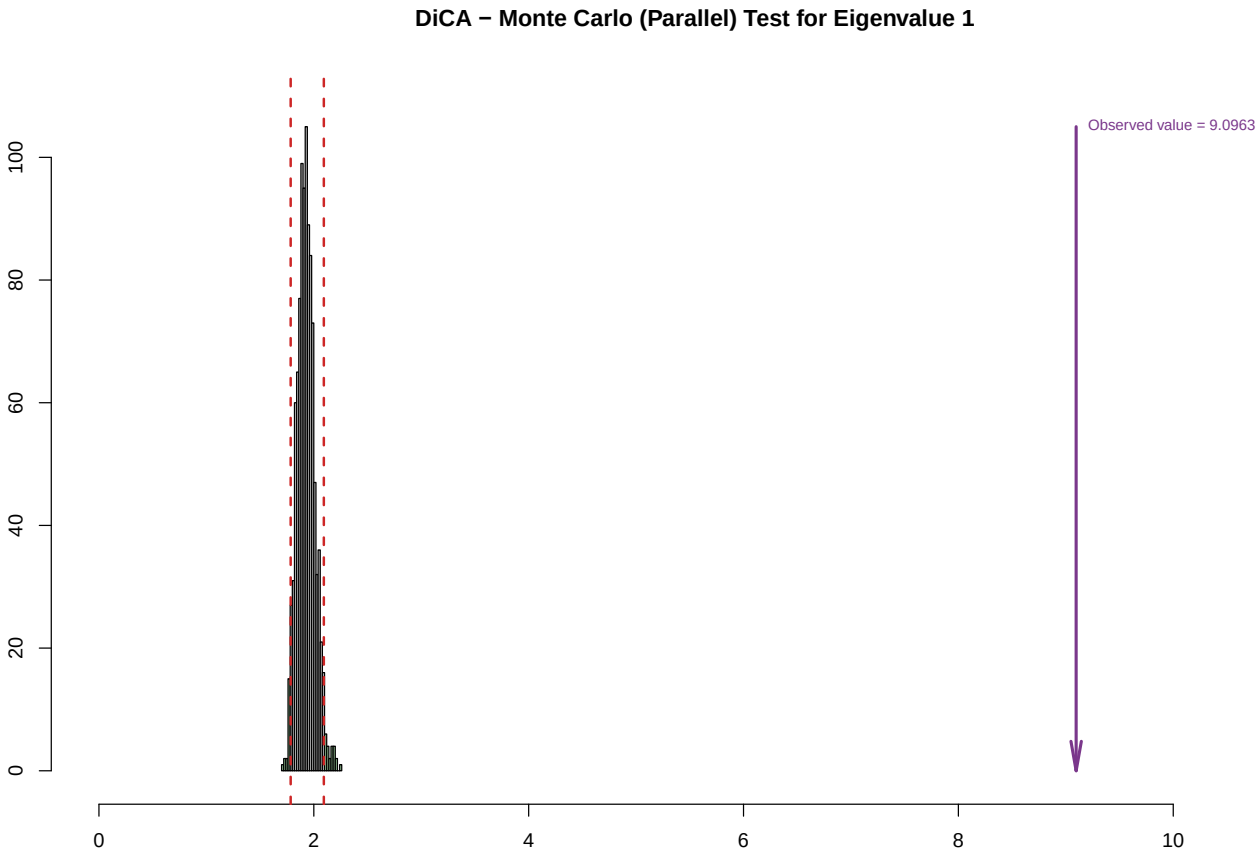


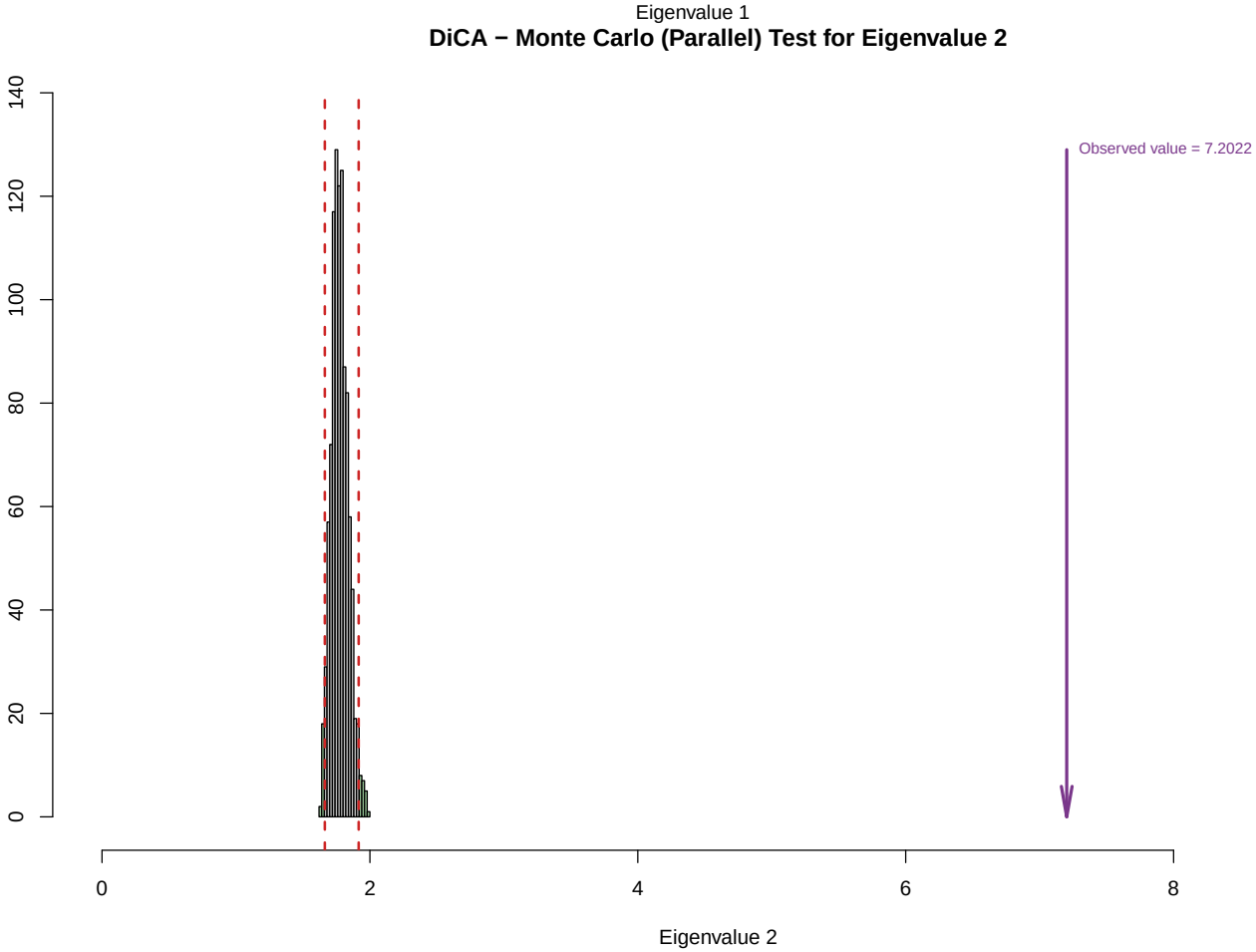
## 6.11 Permutation Test



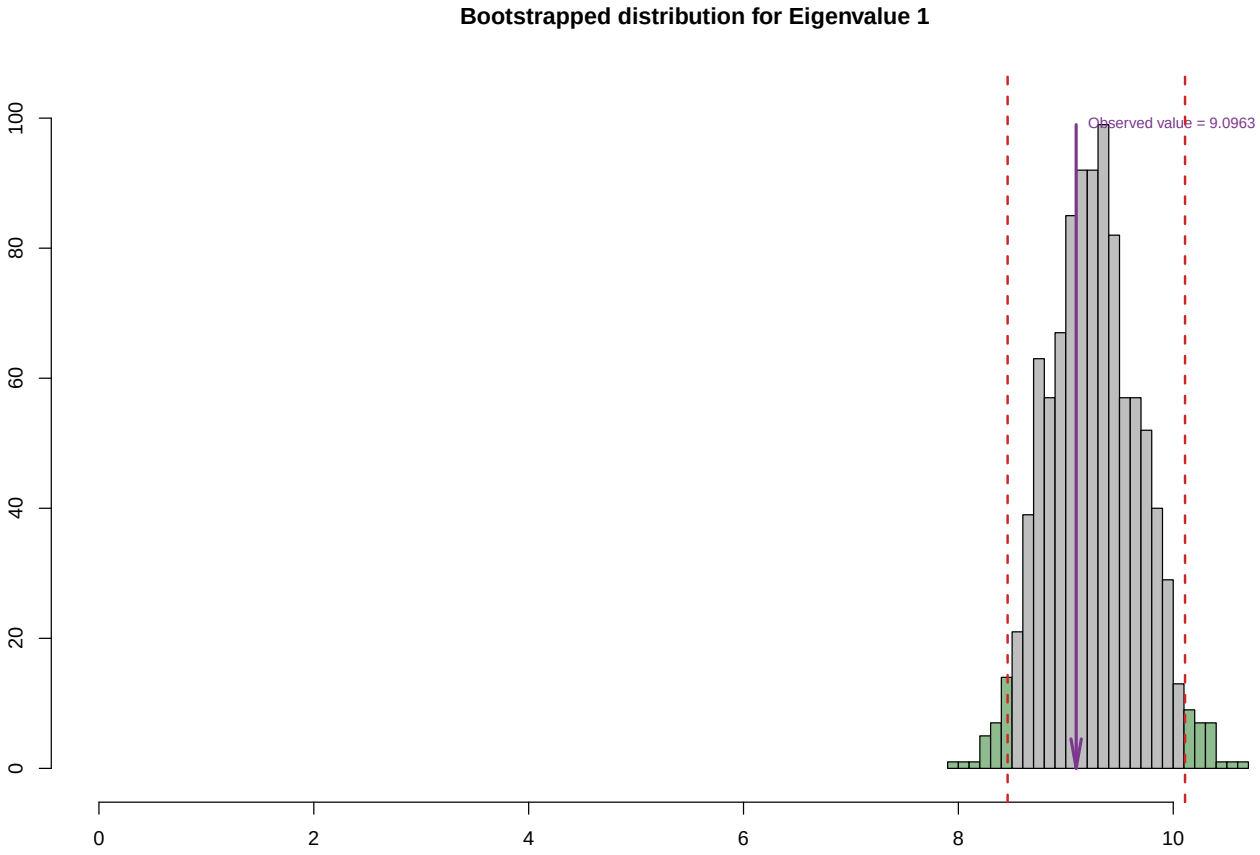



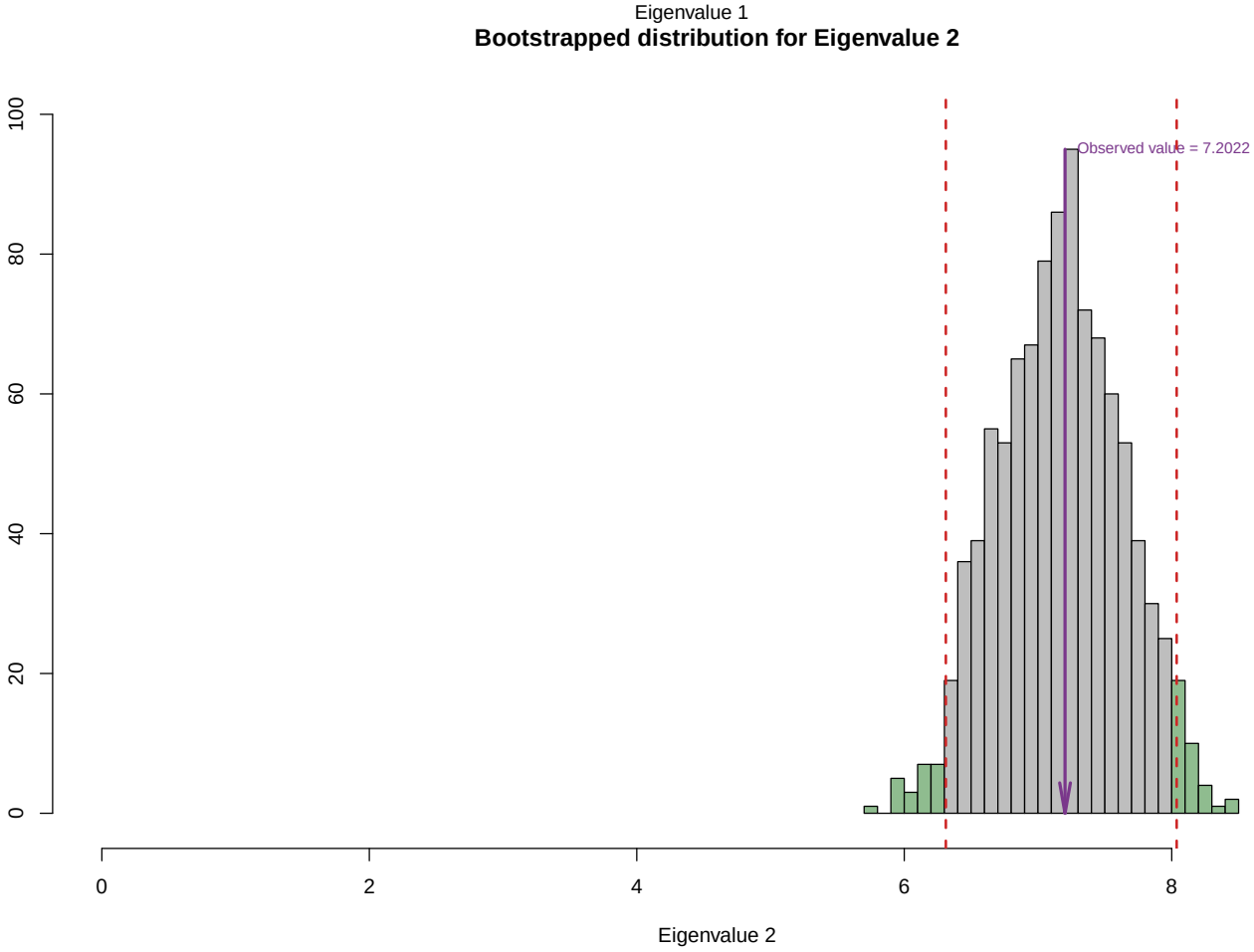
## 6.12 Parallet Test



## 6.13 Bootstrap Test



## 6.14 Conclusion

| Methods | Unhappy | Normal | Very Happy | Reliability |
|---|---|---|---|---|
| **DiCA** | warm summers, cold winters, high rain | Higher variation in temperature is correlated with lower happiness | Warm winter, cold summer, low rain, windy | Convex hulls are separeted but second component only has temp variables as significant |

# Chapter 7

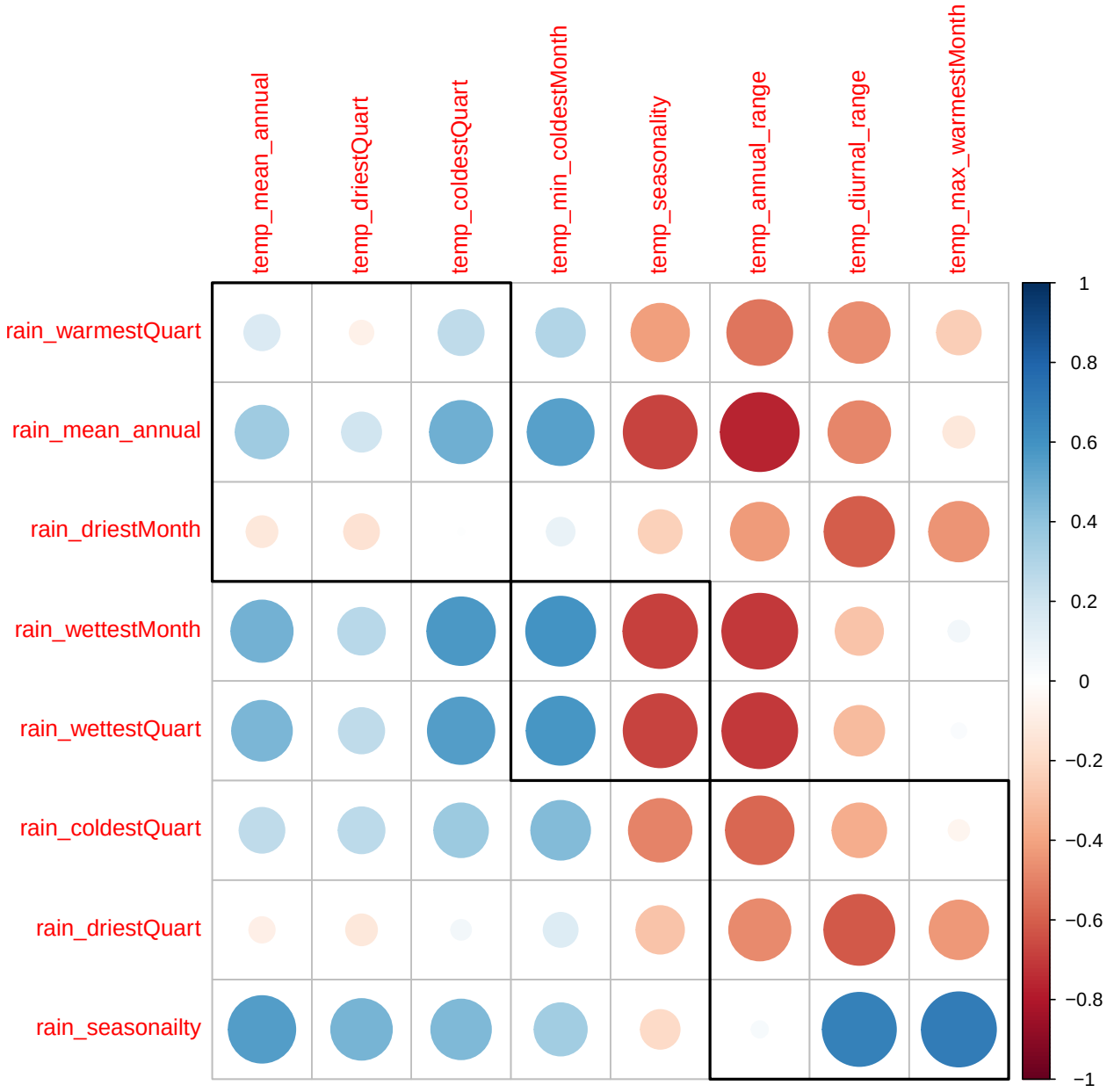
# Partial Least Squares - Correlation

## 7.1 Description

PLS is used to find the fundamental relations between two matrices (X and Y), i.e. a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among X values. PLS bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y data are projected to new spaces, the PLS family of methods are known as bilinear factor models.

## 7.2 Correlation Plot

Visually analyze multicollinearity between all varaibles in Rain and Temperature tables.




## 7.3 PLS-C

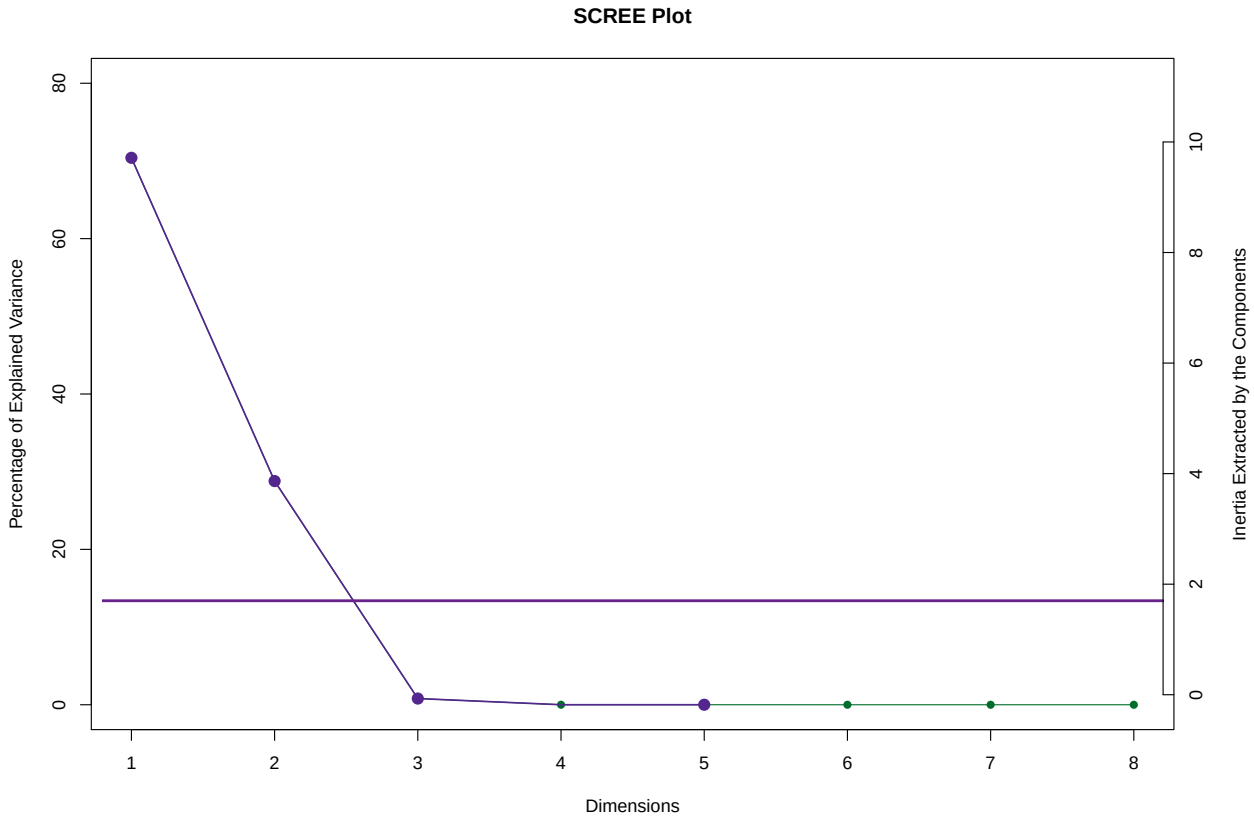
```
## ------------------------------------------------------------------------
##  Results of Permutation Test for PLSC of X'*Y = R
##  for Omnibus Inertia and Eigenvalues
## ------------------------------------------------------------------------
## $ fixedInertia      the Inertia of Matrix X
## $ fixedEigenvalues  an L*1 vector of the eigenvalues of X
## $ pOmnibus          the probablity associated to the Inertia
## $ pEigenvalues      an L* 1 matrix of p for the eigenvalues of X
## $ permInertia       vector of the permuted Inertia of X
## $ permEigenvalues   matrix of the permuted eigenvalues of X
## ------------------------------------------------------------------------
```

Now we have Latent Variables and Saliences. * Latent Variables are the new Data points w.r.t. correlation between both the tables. Latent Variables exists for each table. * Saliences represent correlation between variables of each table.

## 7.4 Scree Plot

Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

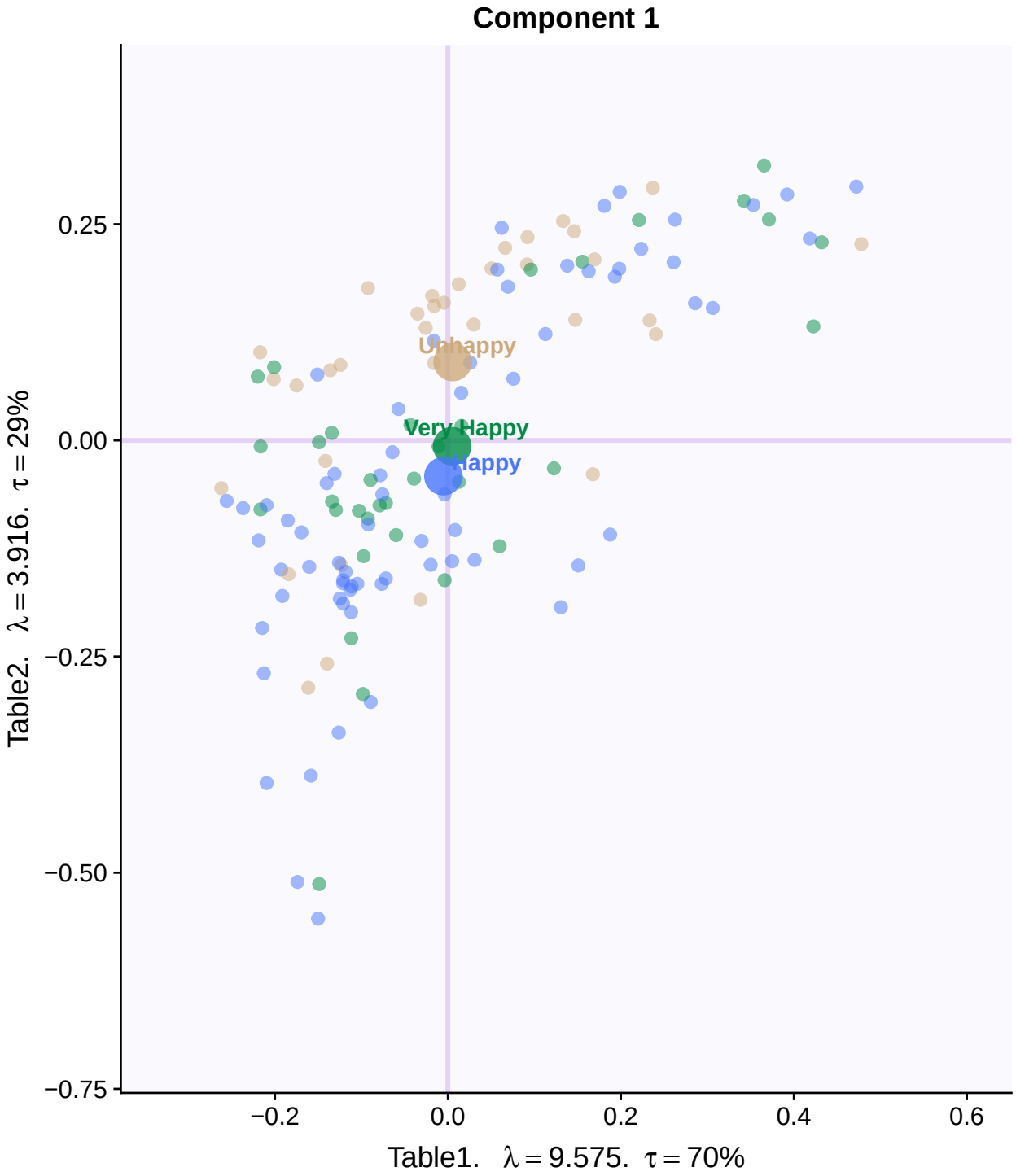
P.S. The most contribution component may not always be most useful for a given research question.



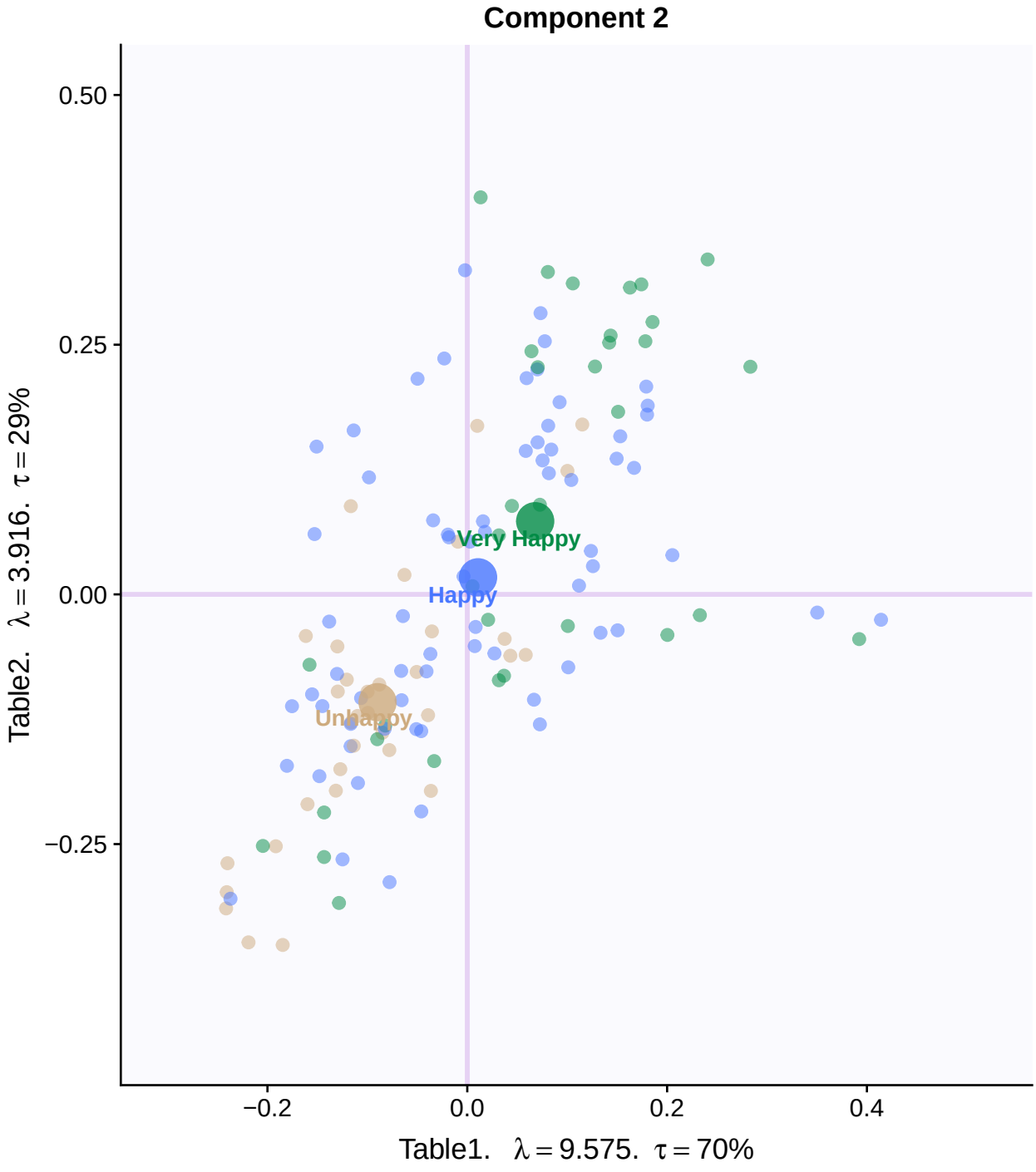

## 7.5 Latent Variables

Lets visualize happiness categories for Components 1 of each table

### 7.5.1 Component 1 for both Tables: Rain and Temperature

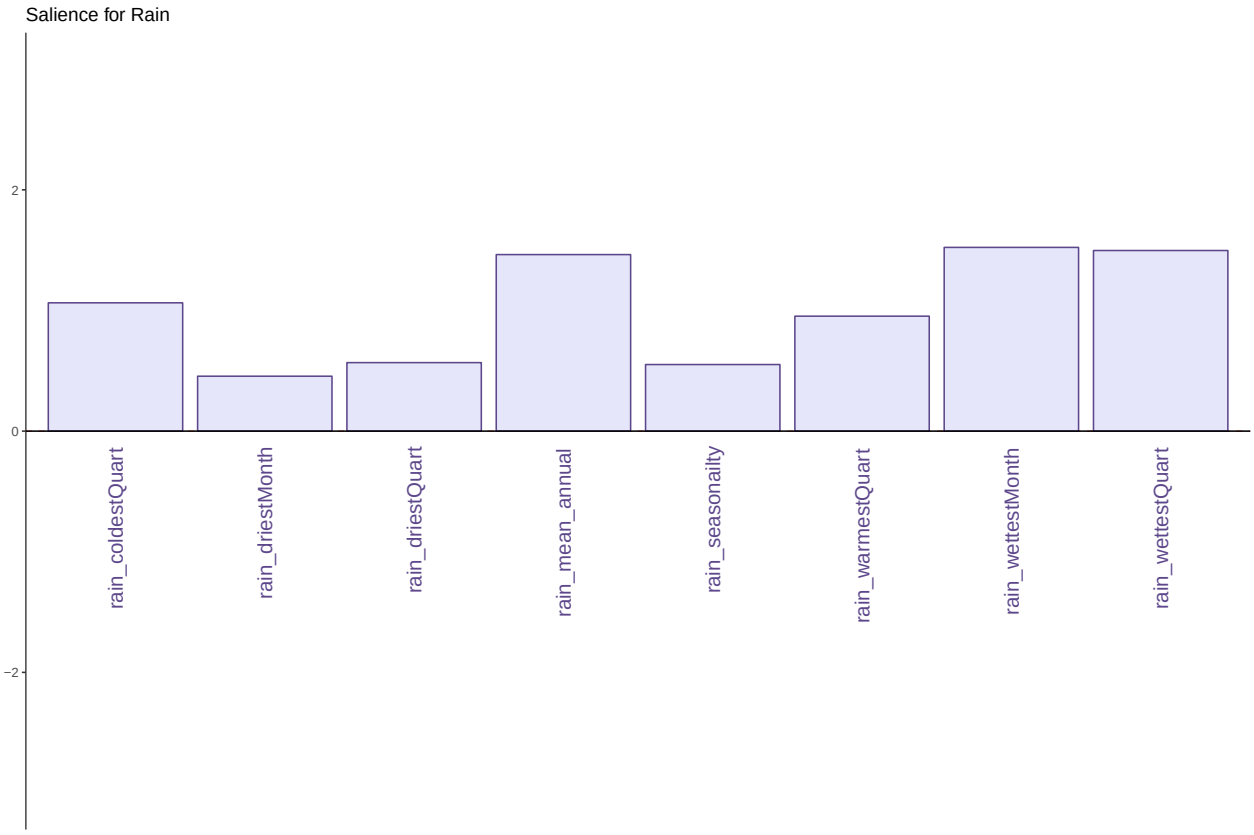


### 7.5.2 Component 2 for both Tables: Rain and Temperature

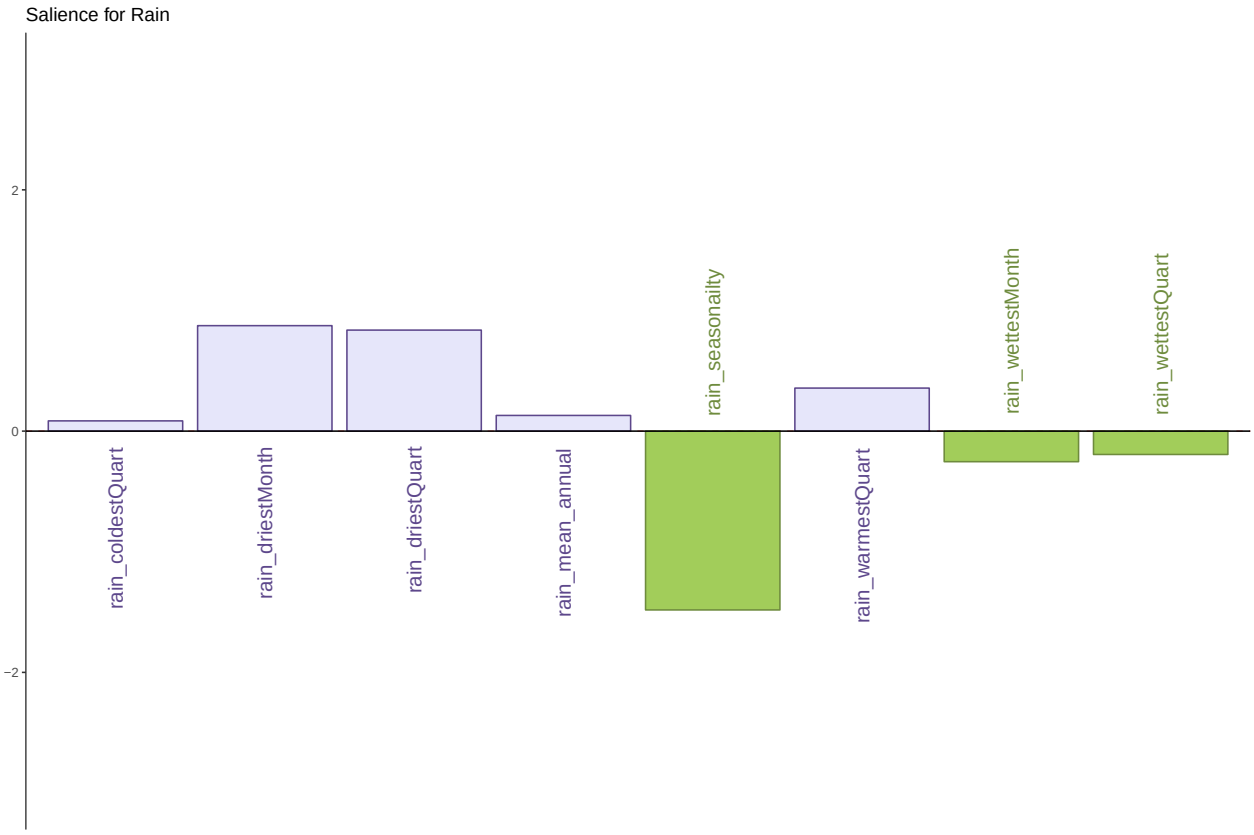


## 7.6 Salience for Rain

### 7.6.1 Components 1

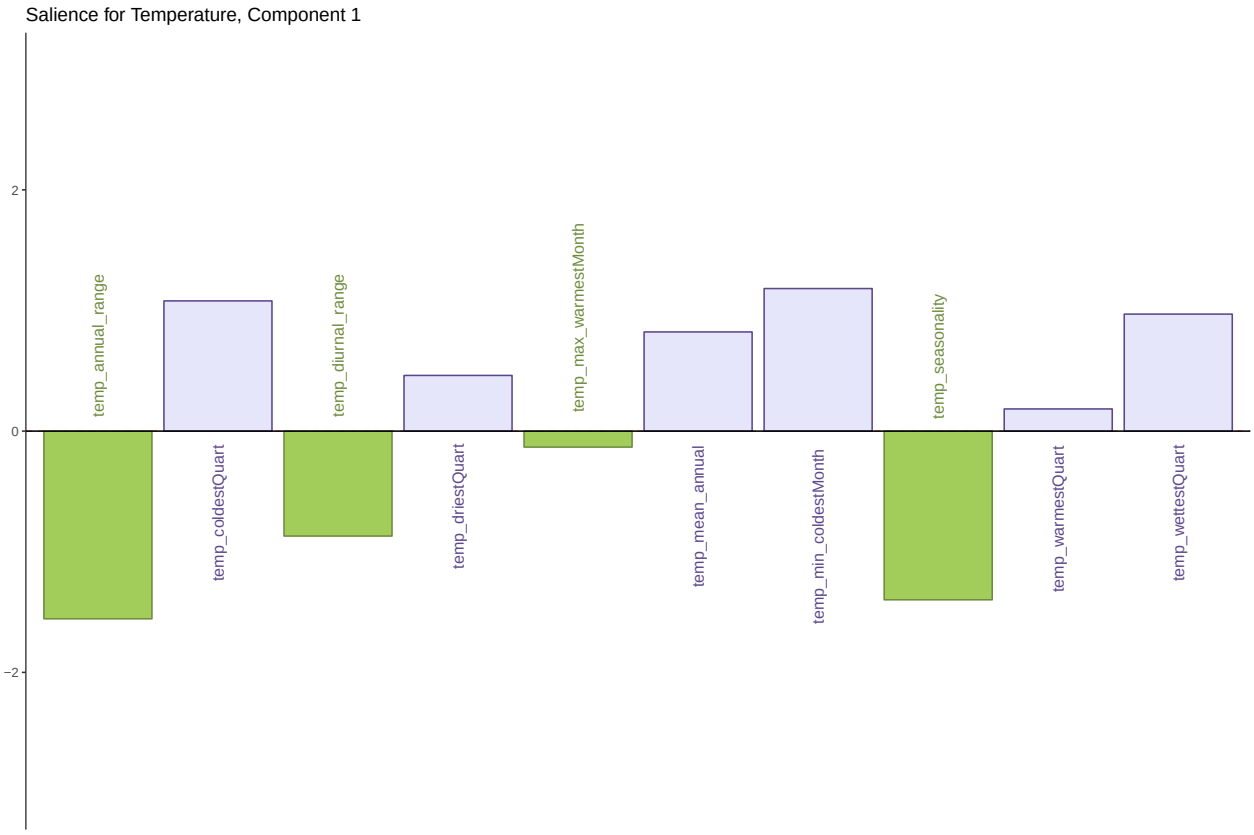


### 7.6.2 Component 2

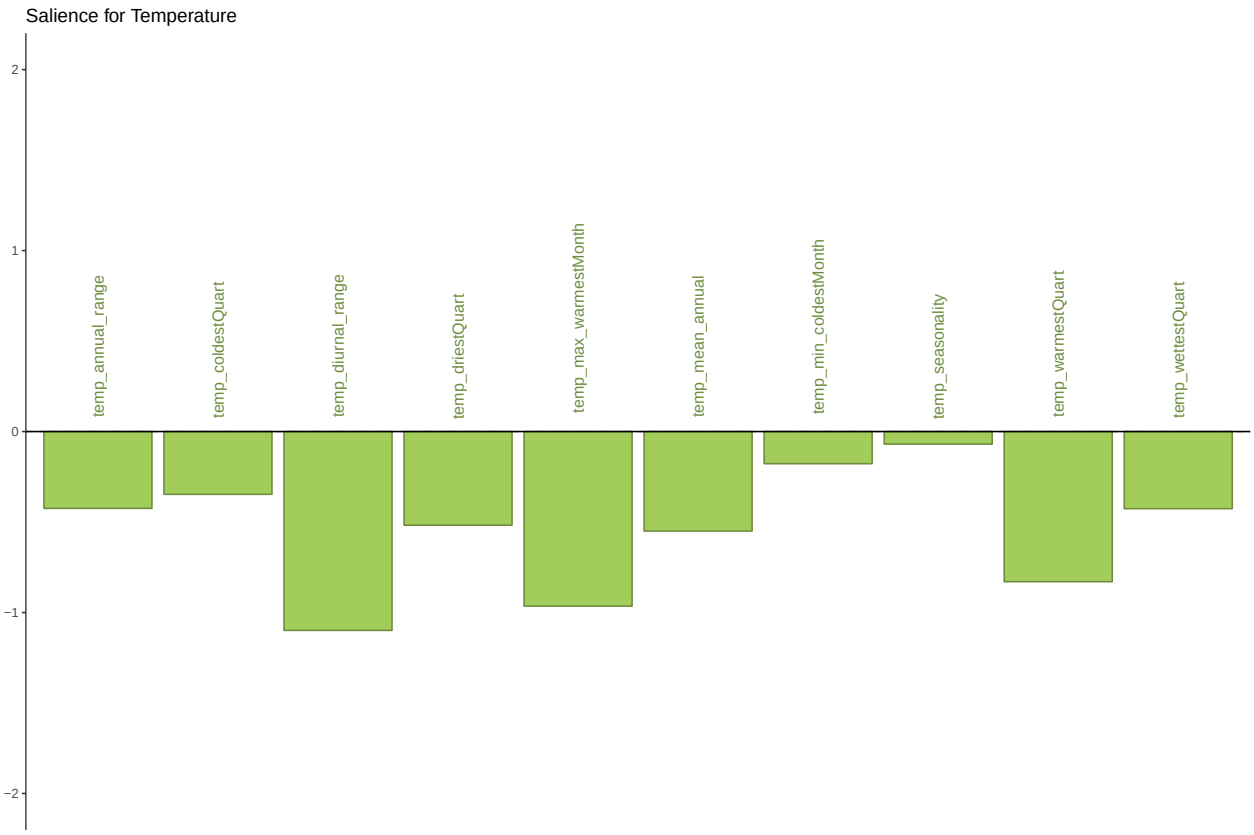


## 7.7 Salience for Temperature

### 7.7.1 Component 1



### 7.7.2 Component 2

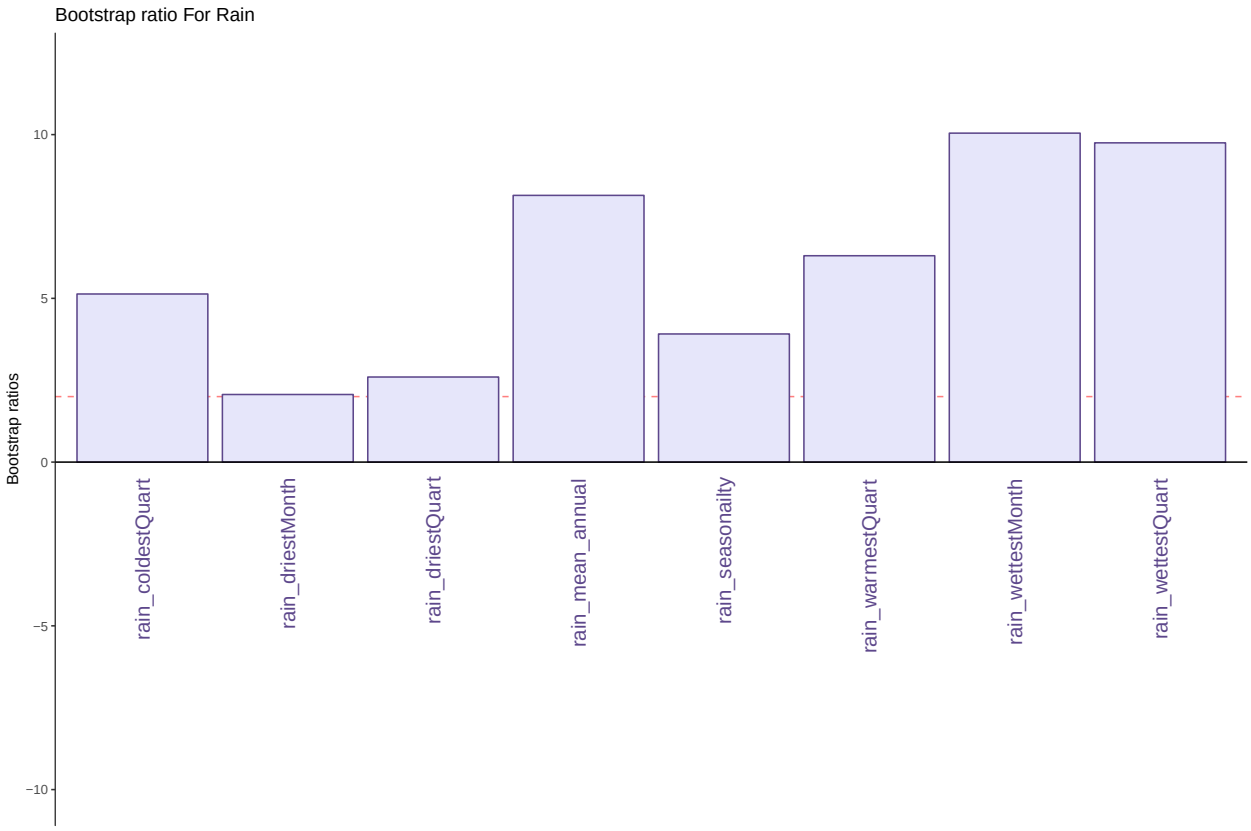



## 7.8 Most Contributing Variables - PLS-C (with Inference)

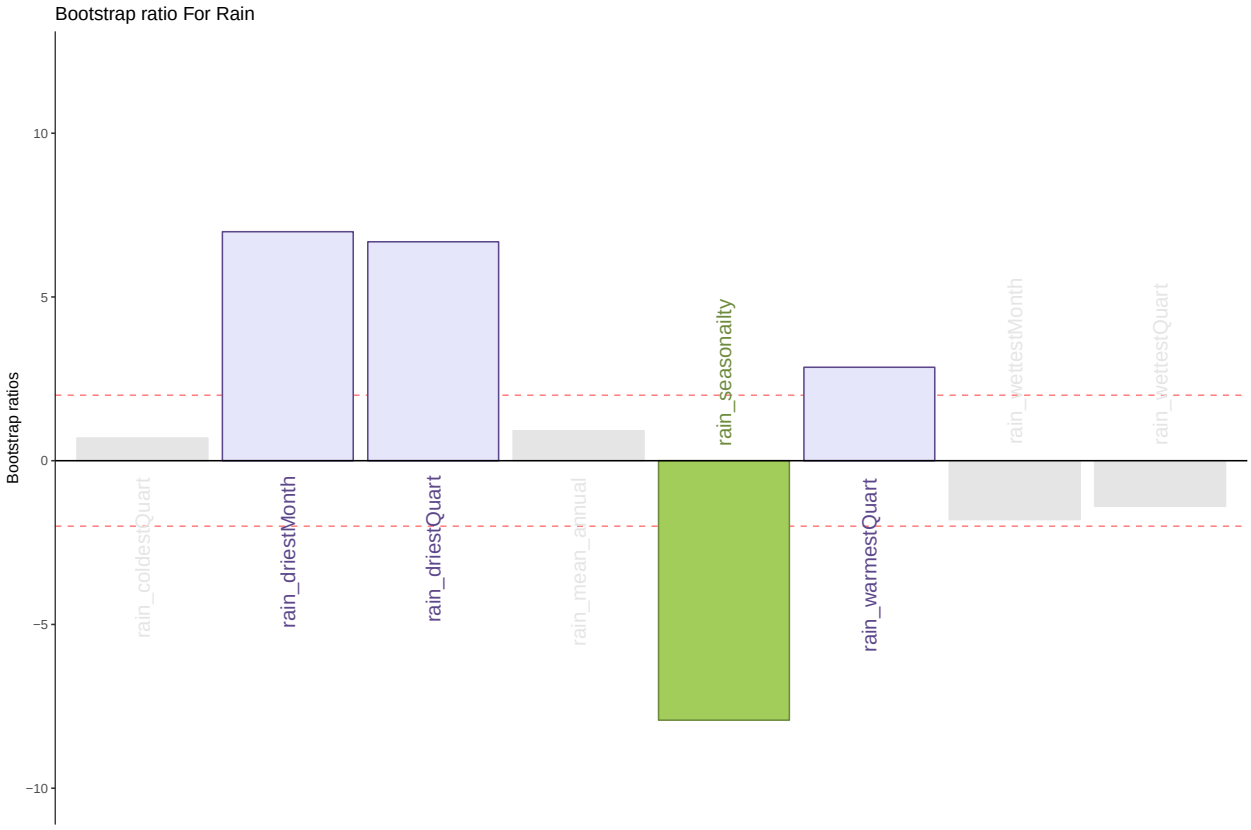
```
## ------------------------------------------------------------------------
##  Bootstraped Factor Scores (BFS) and Bootstrap Ratios  (BR)
##  for the I and J-sets of a PLSC (obtained from multinomial resampling of X & Y)
## ------------------------------------------------------------------------
## $ bootstrapBrick.i          an I*L*nIter Brick of BFSs  for the I-Set
## $ bootRatios.i              an I*L matrix of BRs for the I-Set
## $ bootRatiosSignificant.i   an I*L logical matrix for significance of the I-Set
## $ bootstrapBrick.j          a  J*L*nIter Brick of BFSs  for the J-Set
## $ bootRatios.j              a  J*L matrix of BRs for the J-Set
## $ bootRatiosSignificant.j   a  J*L logical matrix for significance of the J-Set
## ------------------------------------------------------------------------
```

### 7.8.1 Bootstrap Test

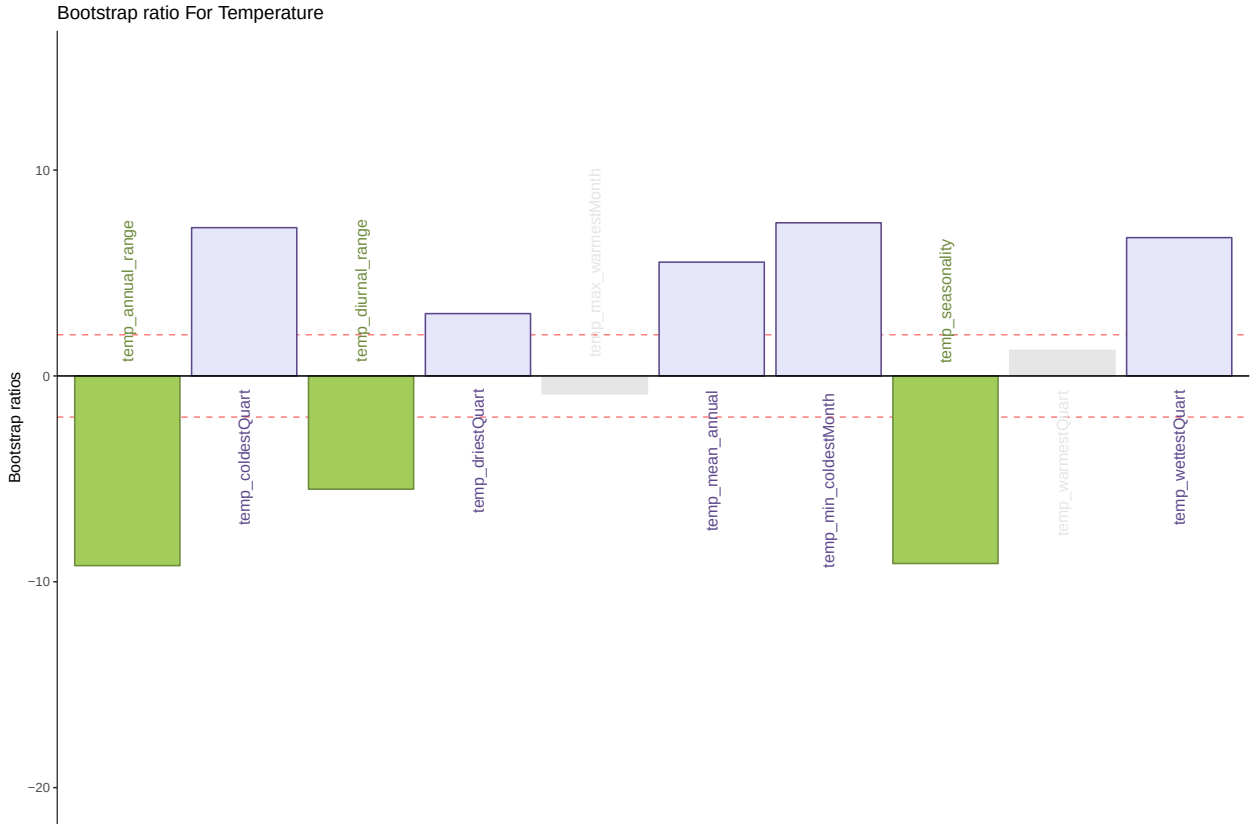
- Rain - Component 1



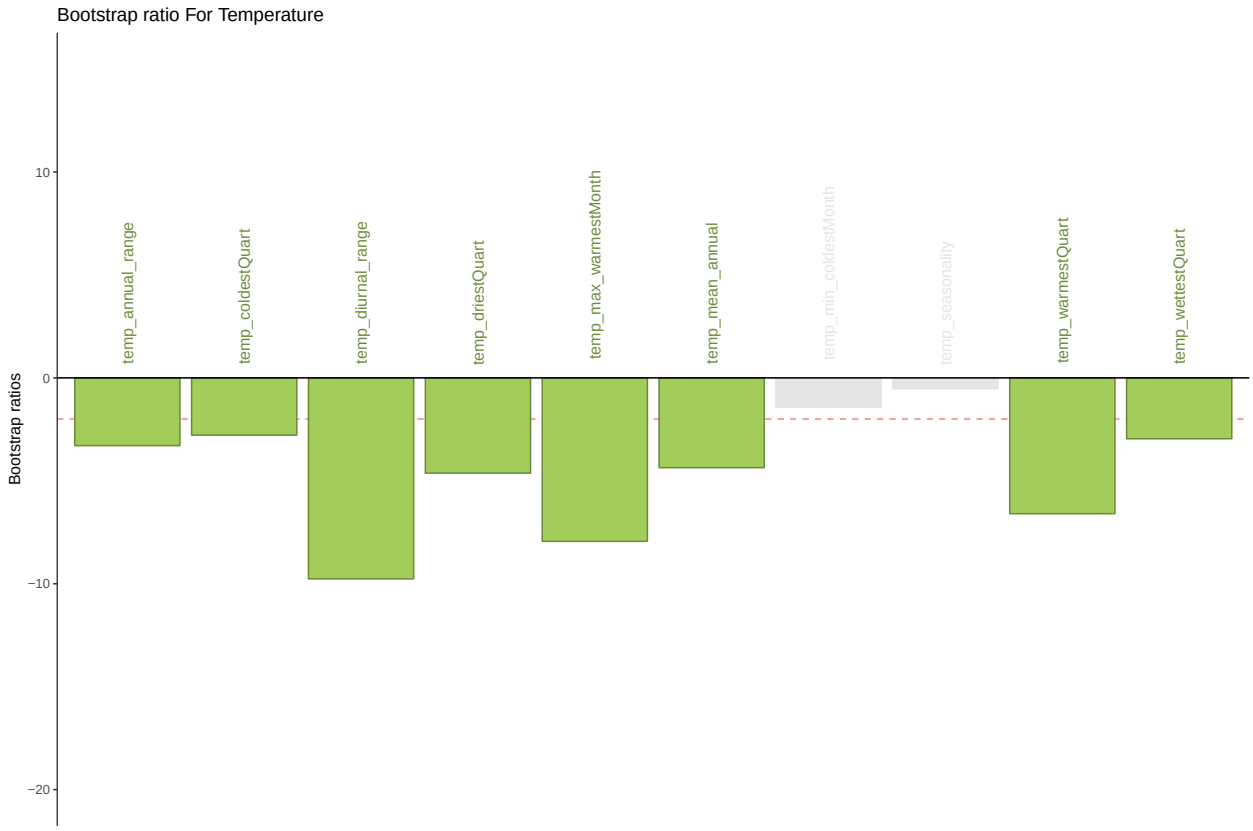

- Rain - Component 2



- Temperature - Component 1




- Temperature - Component 2




## 7.9 Conclusion

Here Component 2 seems to best seperate Happiness levels. Let's compare Component 2 for both tables.

- Table 1 & 2 Component 2
  - Latent Variables: Very Happy vs Unhappy (for Rain and Temperature both)
  - Salience:
    - Rain: It seems dryness and wetness at a montly scale have more effect than coldness or yearly patterns.
    - Temperature: All temperature variations at a monthly and yearly scale seems to impact happiness.

# Chapter 8

# Multiple Factor Analyis

## 8.1 Description

Multiple factor analysis is an extension of PCA tailored to handle multiple datatables that measure sets of variables collected on the same observations. MFA proceeds in two steps 1. It computes a PCA of each data table and 'normalizes' each data table by dividing all its elements by the first singular value obtained from its PCA. 2. All the normalized data tables are aggregated into a grand data table that is analyzed via a (non-normalized) PCA that gives a set of factor scores for the observations and loadings for the variables.

In addition, MFA provides for each data table a set of partial factor scores for the observations that reflects the specific 'view-point' of this data table. Interestingly, the common factor scores could be obtained by replacing the original normalized data tables by the normalized factor scores obtained from the PCA of each of these tables.

## 8.2 MFA

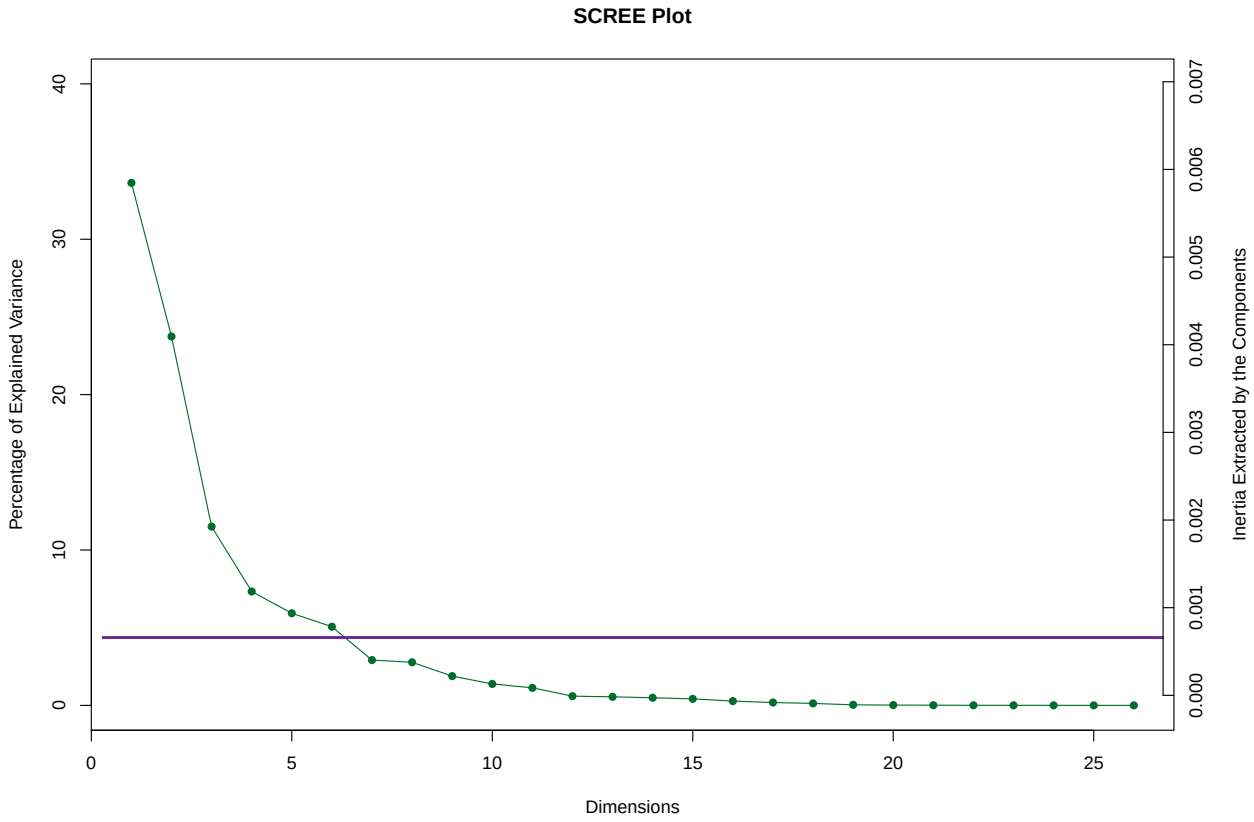
We have divided the data into 3 tables, separate tables for rain and temperature related columns and 3rd table for rest of the columns.

```
## [1] "Preprocessed the Rows of the data matrix using:  None"
## [1] "Preprocessed the Columns of the data matrix using:  Center_1Norm"
## [1] "Preprocessed the Tables of the data matrix using:  MFA_Normalization"
## [1] "Preprocessing Completed"
## [1] "Optimizing using:  None"
## [1] "Processing Complete"
```

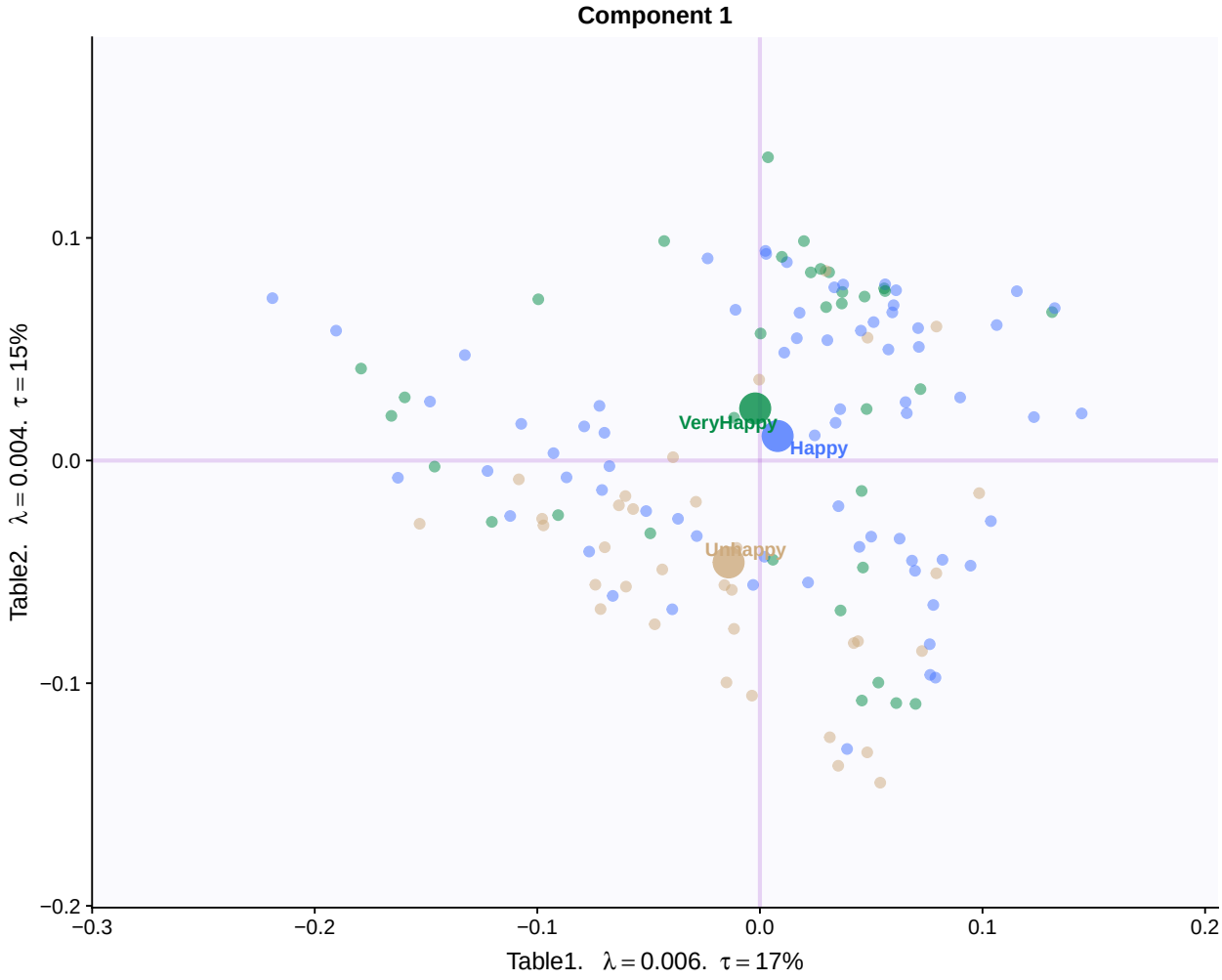
## 8.3 Scree Plot

Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

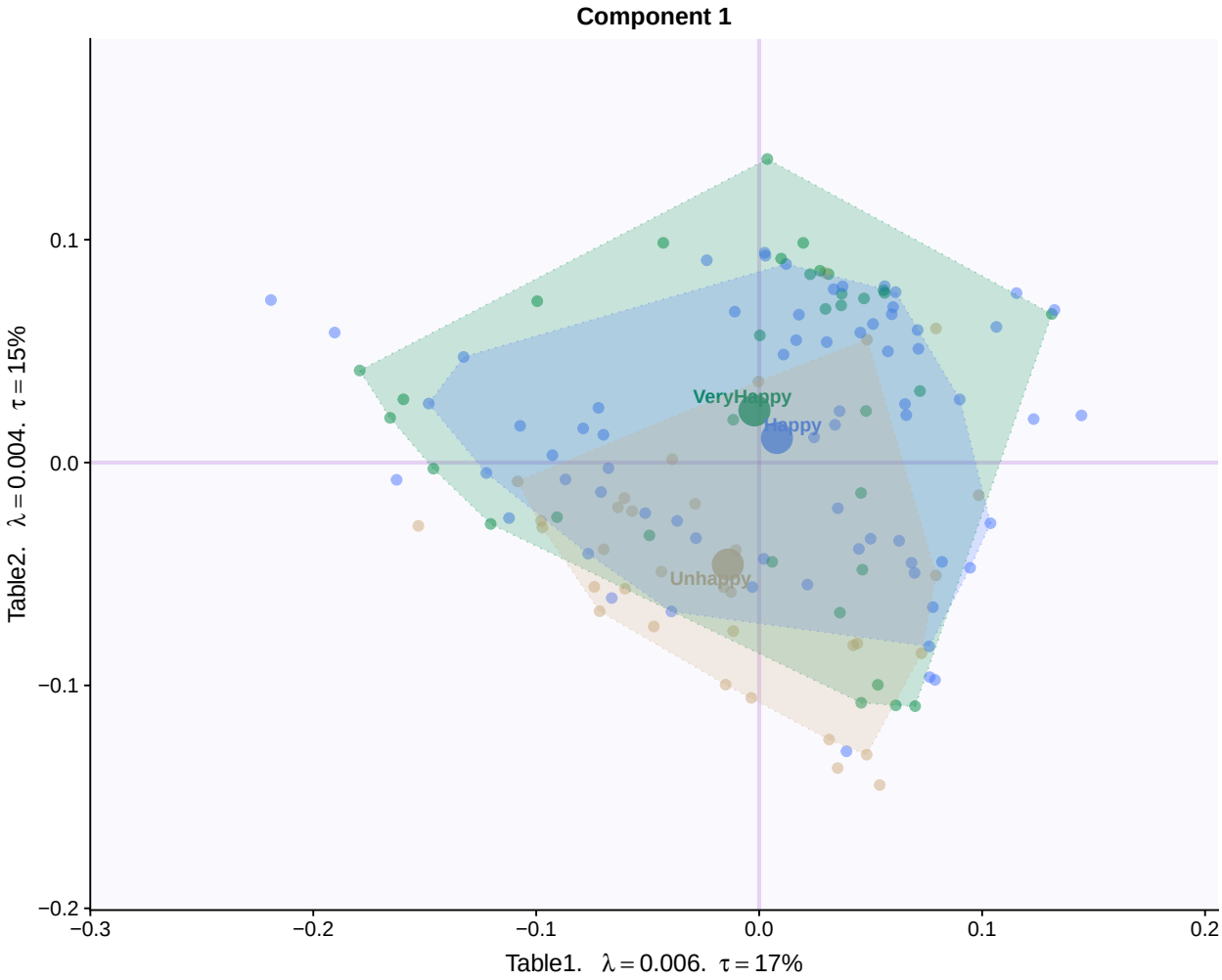
P.S. The most contribution component may not always be most useful for a given research question.


SCREE Plot
Percentage of Explained Variance
Inertia Extracted by the Components
Dimensions
0
10
20
30
40
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0
5
10
15
20
25

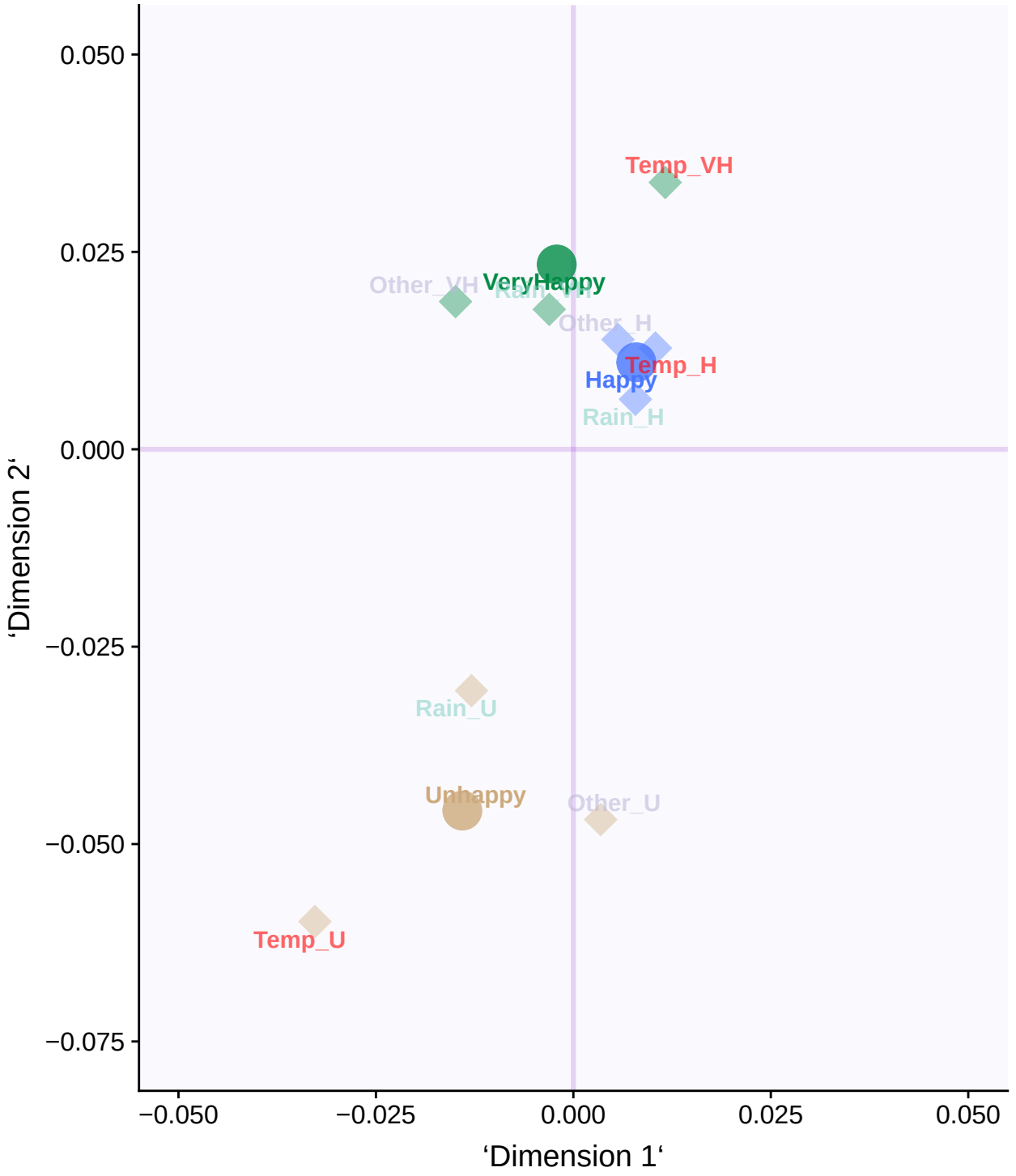
## 8.4 Factor Scores



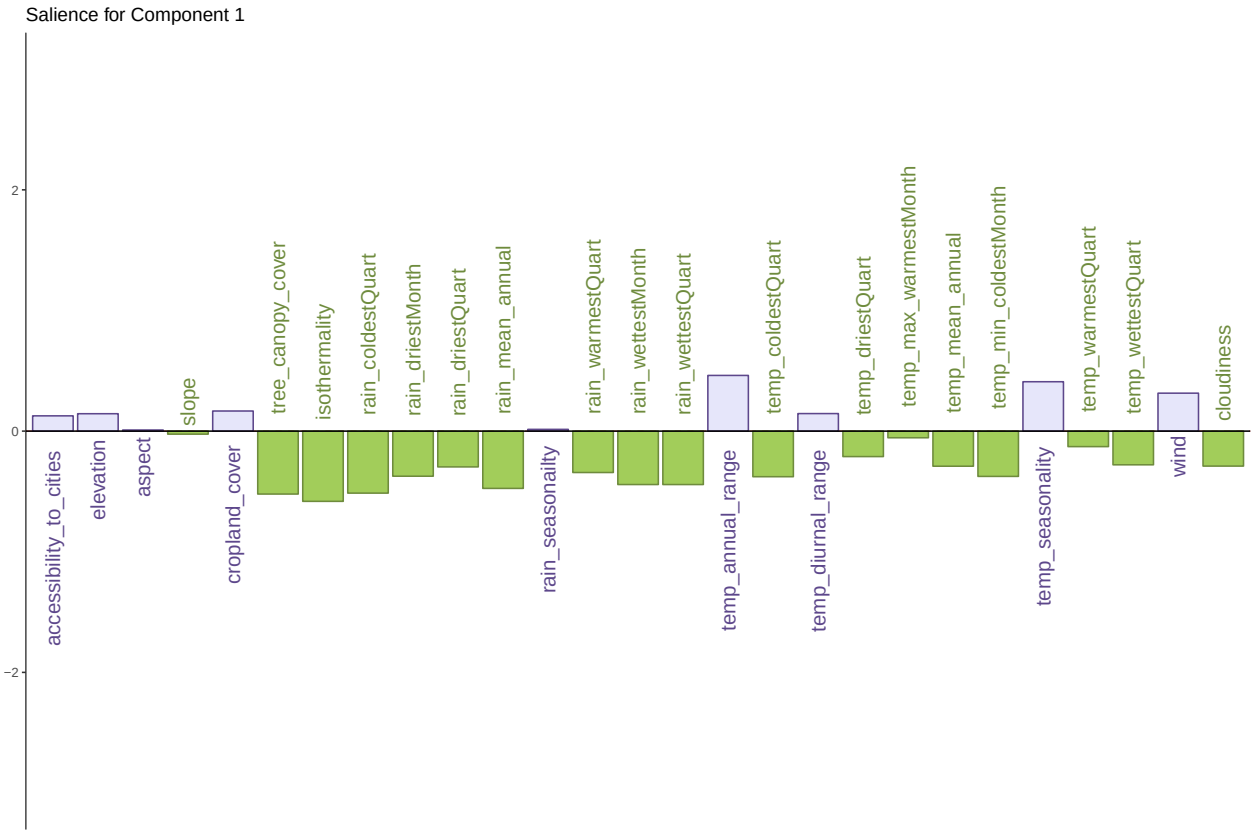

- With Tolerance Interval

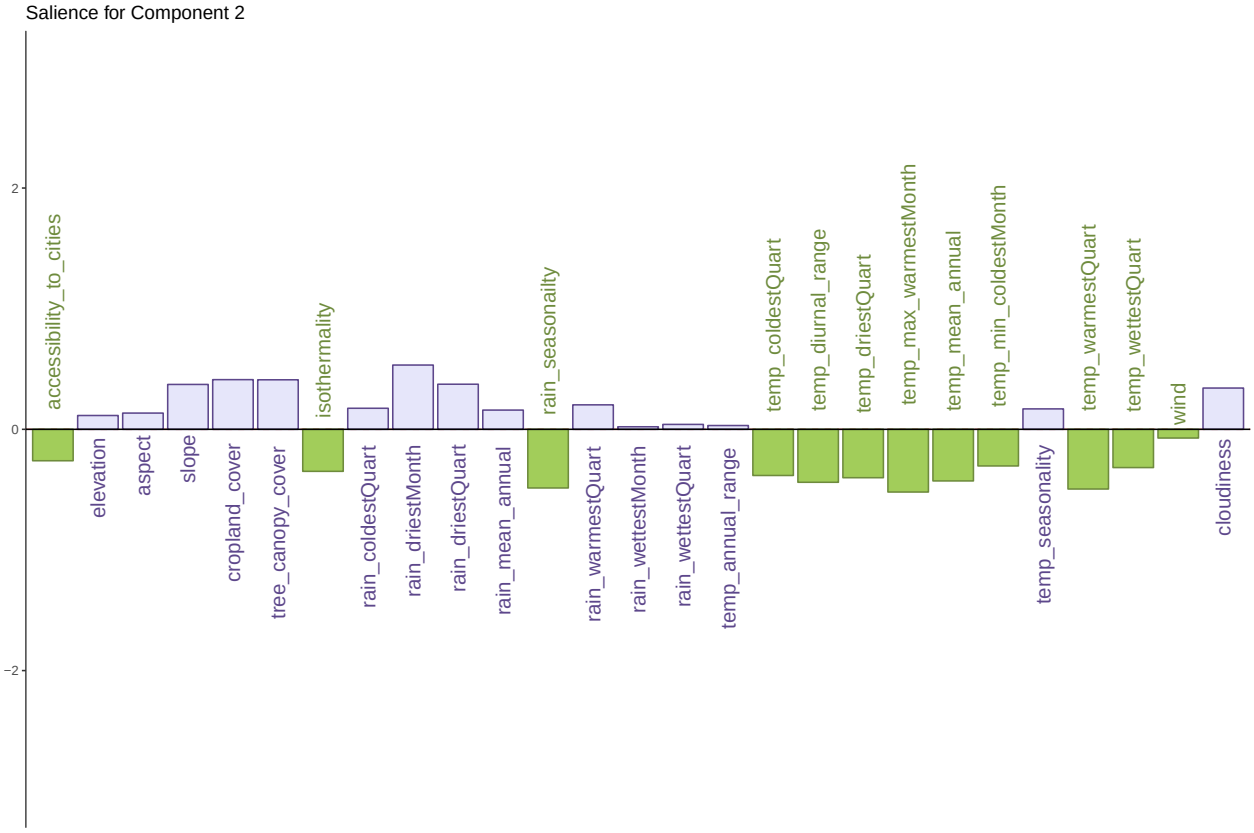



- With Partial Factor Scores

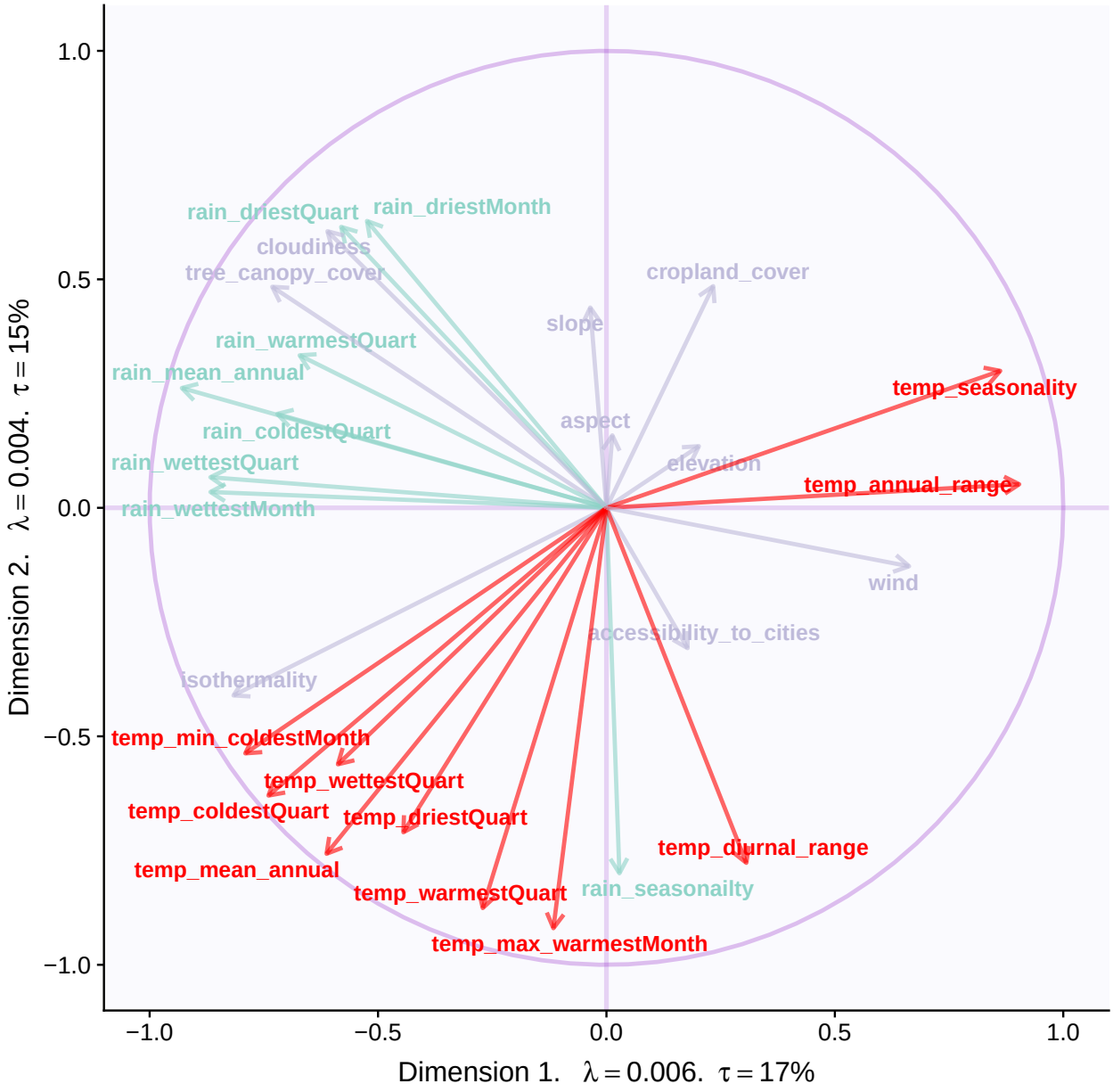

0.050
0.025
0.000
−0.025
−0.050
−0.075
'Dimension 2'
−0.050
−0.025
0.000
0.025
0.050
'Dimension 1'
Temp_VH
VeryHappy
Other_VH
Rain_VH
Other_H
Temp_H
Happy
Rain_H
Rain_U
Unhappy
Other_U
Temp_U

## 8.5 Loadings


Salience for Component 1
2
0
−2
accessibility_to_cities
elevation
aspect
slope
cropland_cover
tree_canopy_cover
isothermality
rain_coldestQuart
rain_driestMonth
rain_driestQuart
rain_mean_annual
rain_seasonailty
rain_warmestQuart
rain_wettestMonth
rain_wettestQuart
temp_annual_range
temp_coldestQuart
temp_diurnal_range
temp_driestQuart
temp_max_warmestMonth
temp_mean_annual
temp_min_coldestMonth
temp_seasonality
temp_warmestQuart
temp_wettestQuart
wind
cloudiness


Salience for Component 2
2
0
−2
accessibility_to_cities
elevation
aspect
slope
cropland_cover
tree_canopy_cover
isothermality
rain_coldestQuart
rain_driestMonth
rain_driestQuart
rain_mean_annual
rain_seasonailty
rain_warmestQuart
rain_wettestMonth
rain_wettestQuart
temp_annual_range
temp_coldestQuart
temp_diurnal_range
temp_driestQuart
temp_max_warmestMonth
temp_mean_annual
temp_min_coldestMonth
temp_seasonality
temp_warmestQuart
temp_wettestQuart
wind
cloudiness

## 8.6 Correlation Circle




## 8.7 Conclusion

| Methods | Unhappy | Normal | Very Happy | Reliability |
|---|---|---|---|---|
| **MFA** | Partial factors dominated by Temp, then rain and other variables | Neither of partial factors seems to have sufficient effect | Partial factors dominated by Temp and other variables, lesser effect of rain | Convex hull has overlapping areas |

# Chapter 9

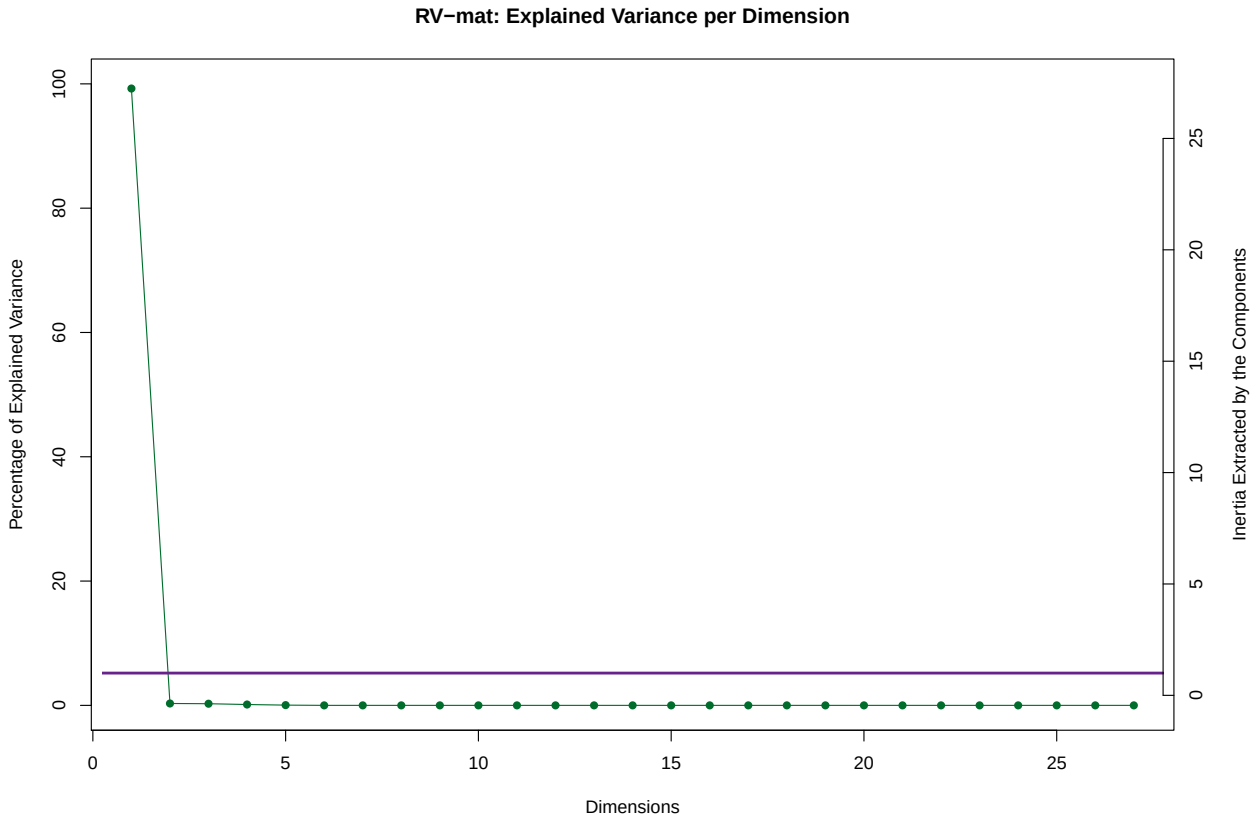
# DiSTATIS

## 9.1 Description

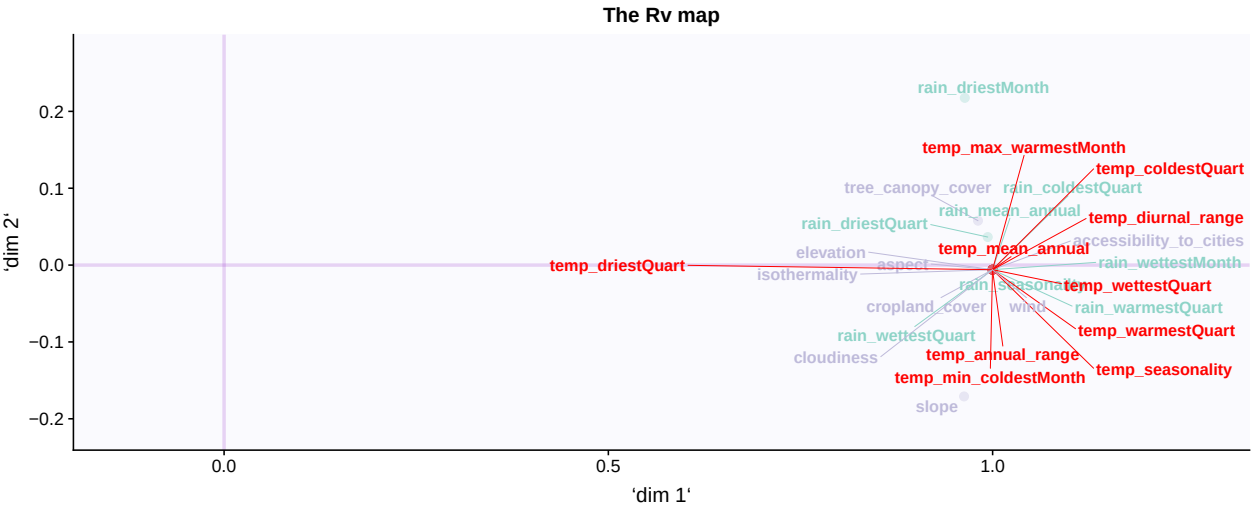
DISTATIS is a new method that can be used to compare algorithms when their outputs consist of distance matrices computed on the same set of objects. The method first evaluates the similarity between algorithms using a coefficient called the RV coefficient. From this analysis, a compromise matrix is computed which represents the best aggregate of the original matrices. In order to evaluate the differences between algorithms, the original distance matrices are then projected onto the compromise. The goal of DISTATIS is to analyze a set of distance matrices. In order to compare distance matrices, DISTATIS combines them into a common structure called a compromise and then projects the original distance matrices onto this compromise.

```
## [1] Bootstrap On Factor Scores. Iterations #:
## [2] 1000
```

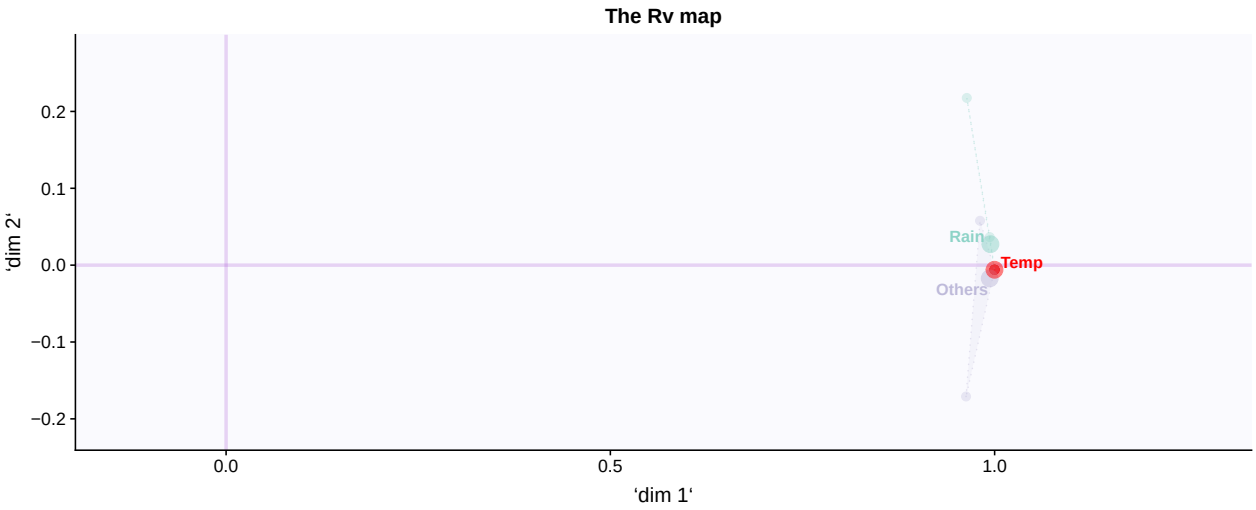
## 9.2 SCREE Plot - RV-MAT



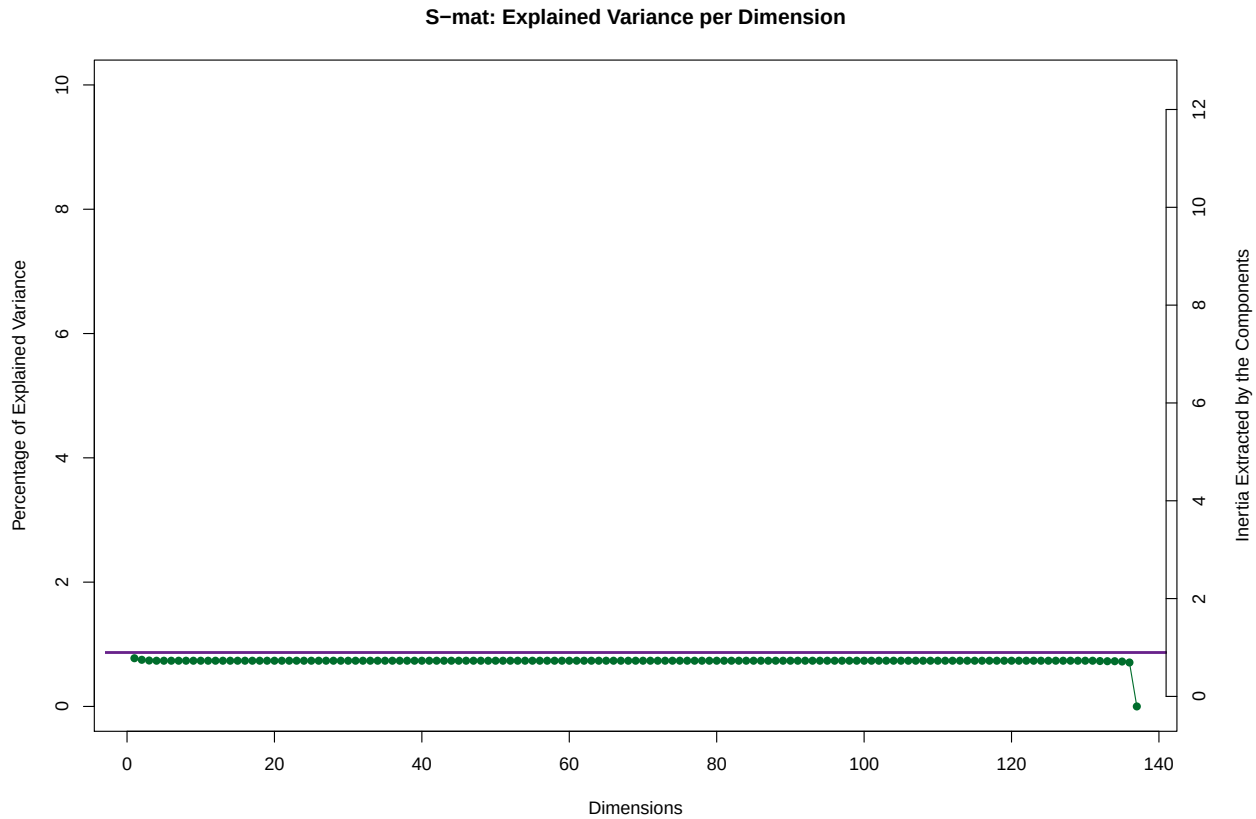

## 9.3 Plotting Assessor Matrix



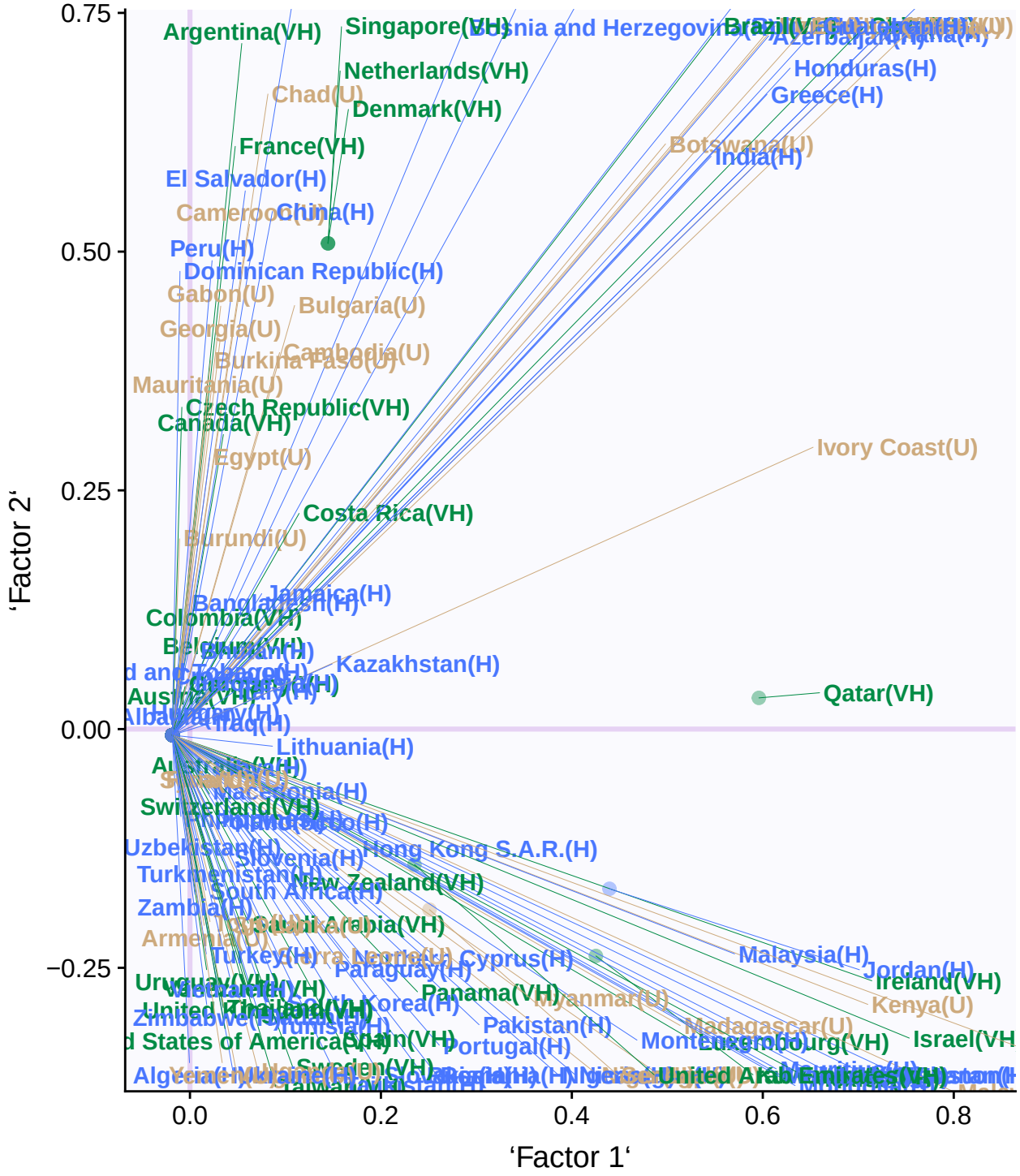
### 9.3.1 ConvexHull

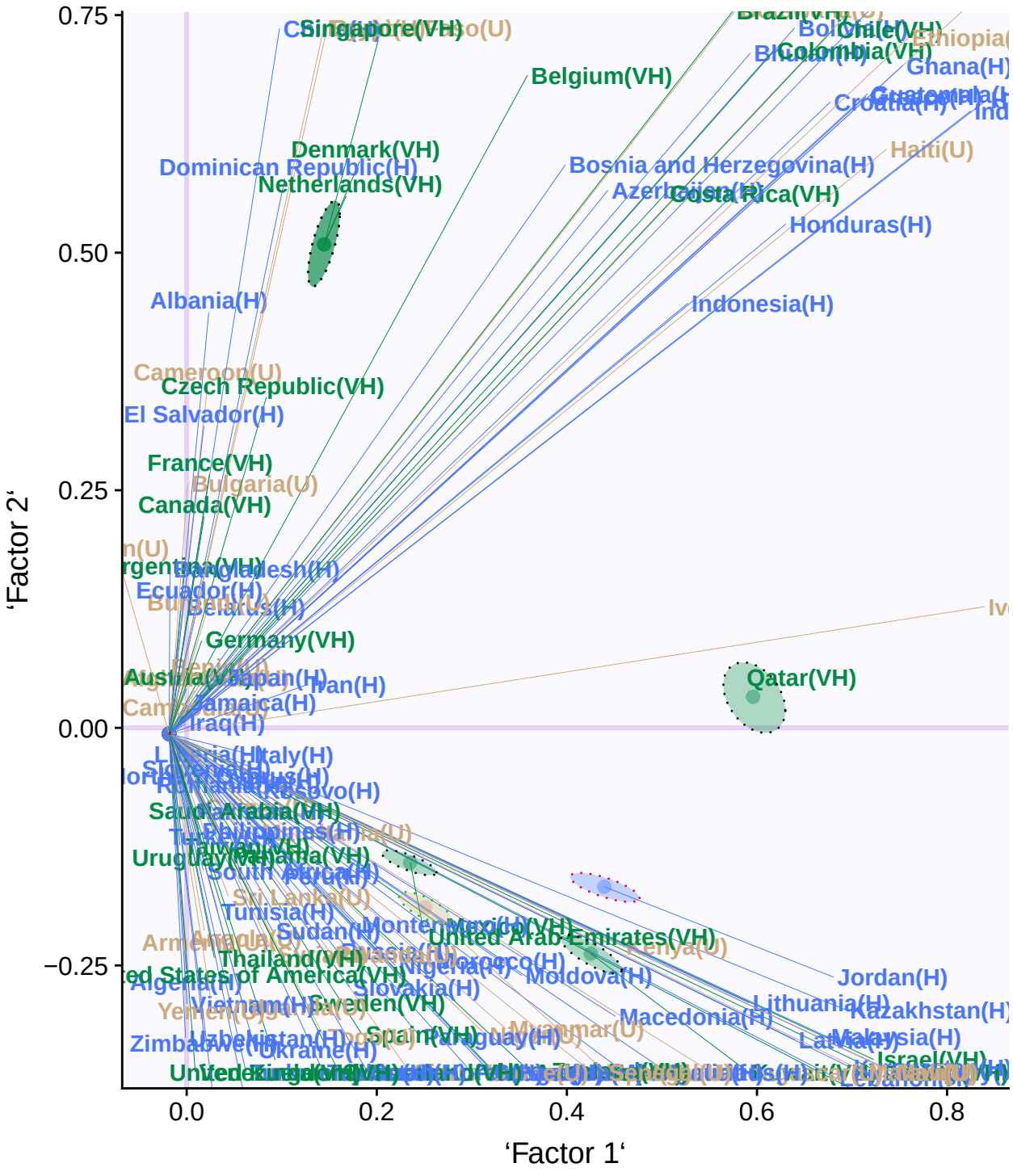



## 9.4 SCREE Plot - SV-MAT



## 9.5 I Set

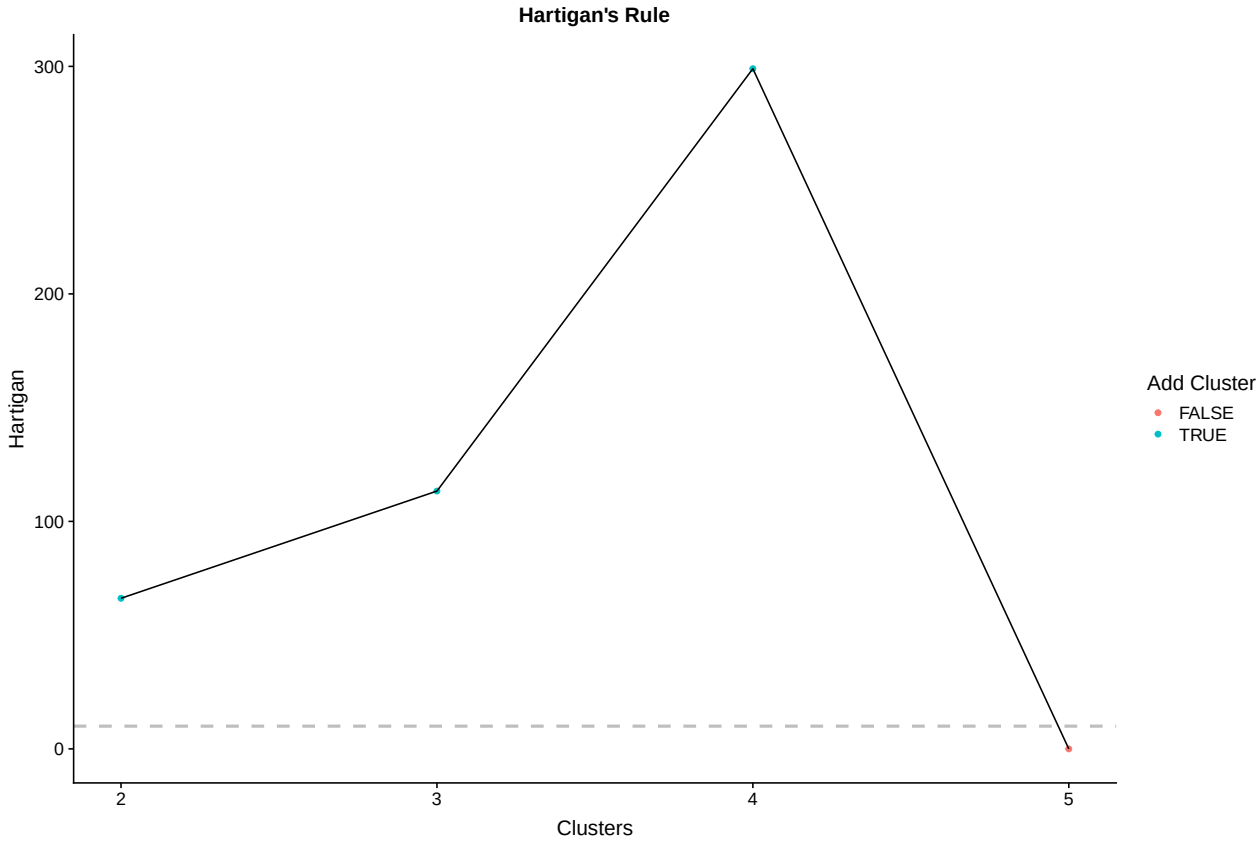

0.75
0.50
0.25
0.00
−0.25
0.0
0.2
0.4
0.6
0.8
'Factor 2'
'Factor 1'
Argentina(VH)
Singapore(VH)
Netherlands(VH)
Chad(U)
Denmark(VH)
France(VH)
El Salvador(H)
China(H)
Peru(H)
Dominican Republic(H)
Gabon(U)
Bulgaria(U)
Georgia(U)
Cambodia(U)
Mauritania(U)
Czech Republic(VH)
Canada(VH)
Egypt(U)
Costa Rica(VH)
Burundi(U)
Jamaica(H)
Colombia(VH)
Kazakhstan(H)
Qatar(VH)
Lithuania(H)
Switzerland(VH)
Hong Kong S.A.R.(H)
New Zealand(VH)
South Africa(H)
Zambia(H)
Armenia(U)
Turkey(H)
Cyprus(H)
Paraguay(H)
Panama(VH)
Pakistan(H)
Portugal(H)
Sweden(VH)
Malaysia(H)
Jordan(H)
Ireland(VH)
Kenya(U)
Madagascar(U)
Israel(VH
Honduras(H)
Greece(H)
Botswana(U)
India(H)
Ivory Coast(U)




## 9.6 Cluster Analysis (Hartigan's Rule)

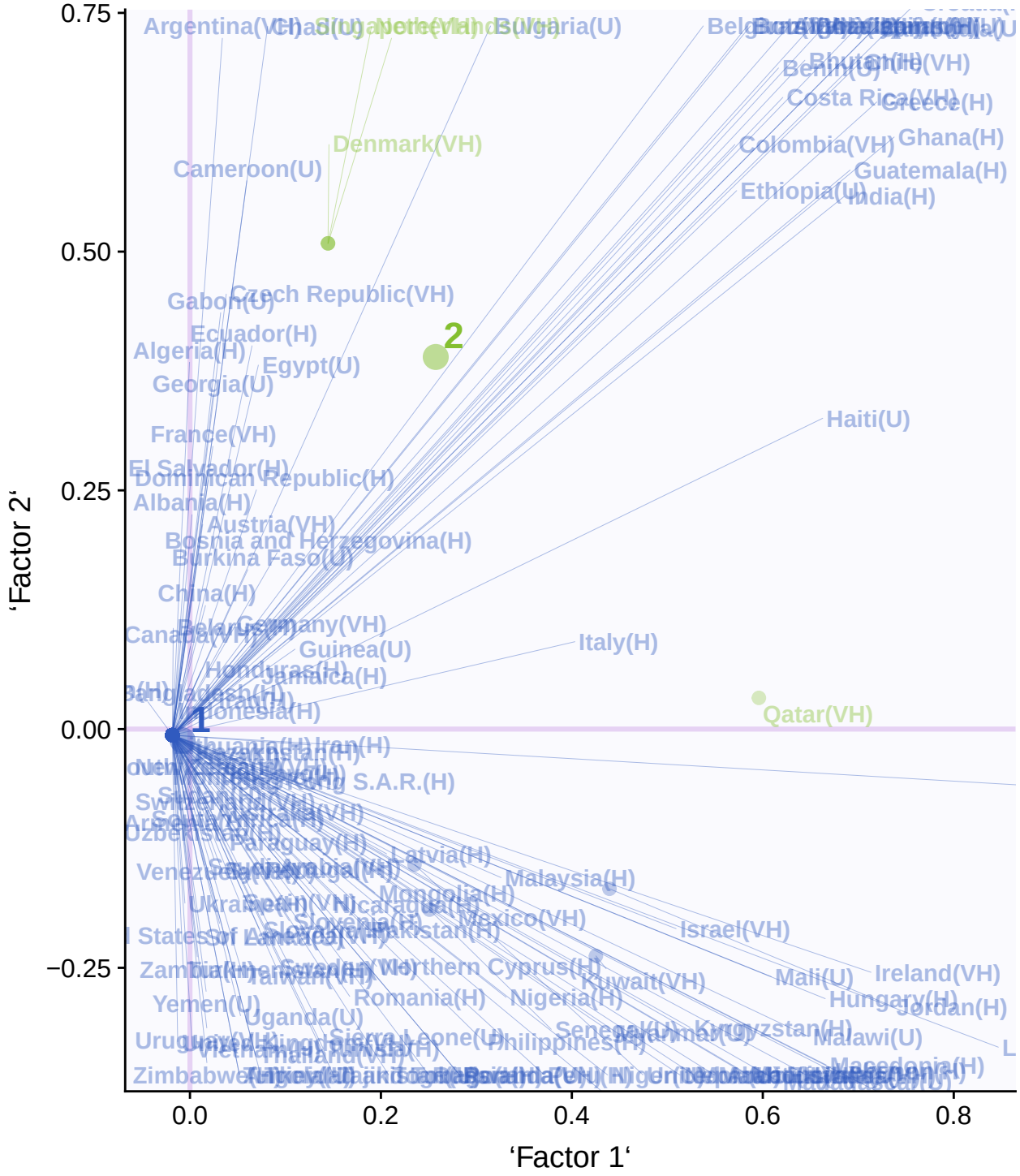
```
##   Clusters  Hartigan AddCluster
## 1        2  66.20137       TRUE
## 2        3 113.33226       TRUE
## 3        4 298.98109       TRUE
```

```
## 4         5   0.00000      FALSE
```

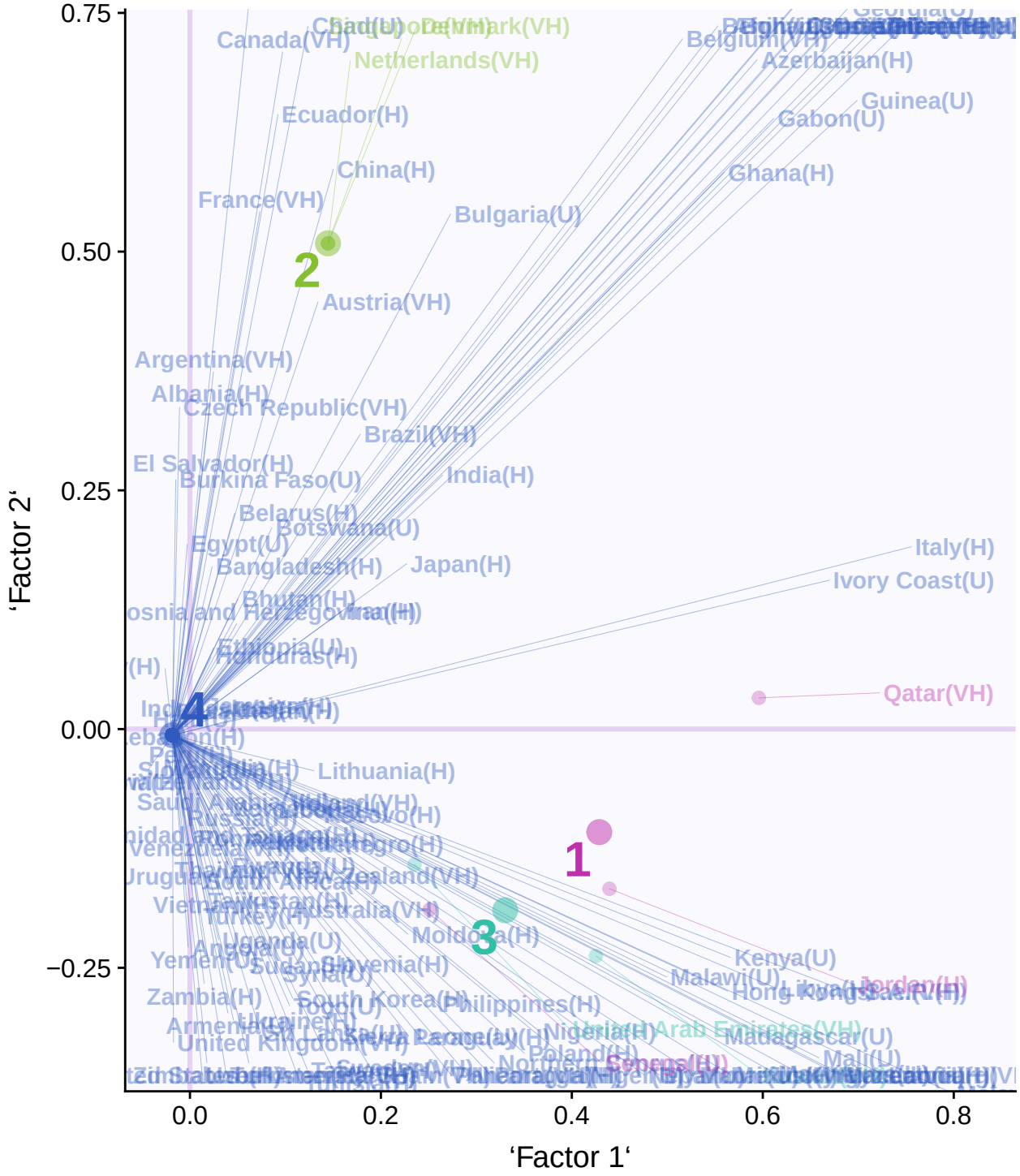


## 9.7 Cluster Analysis (K-Means)

### 9.7.1 Two centers

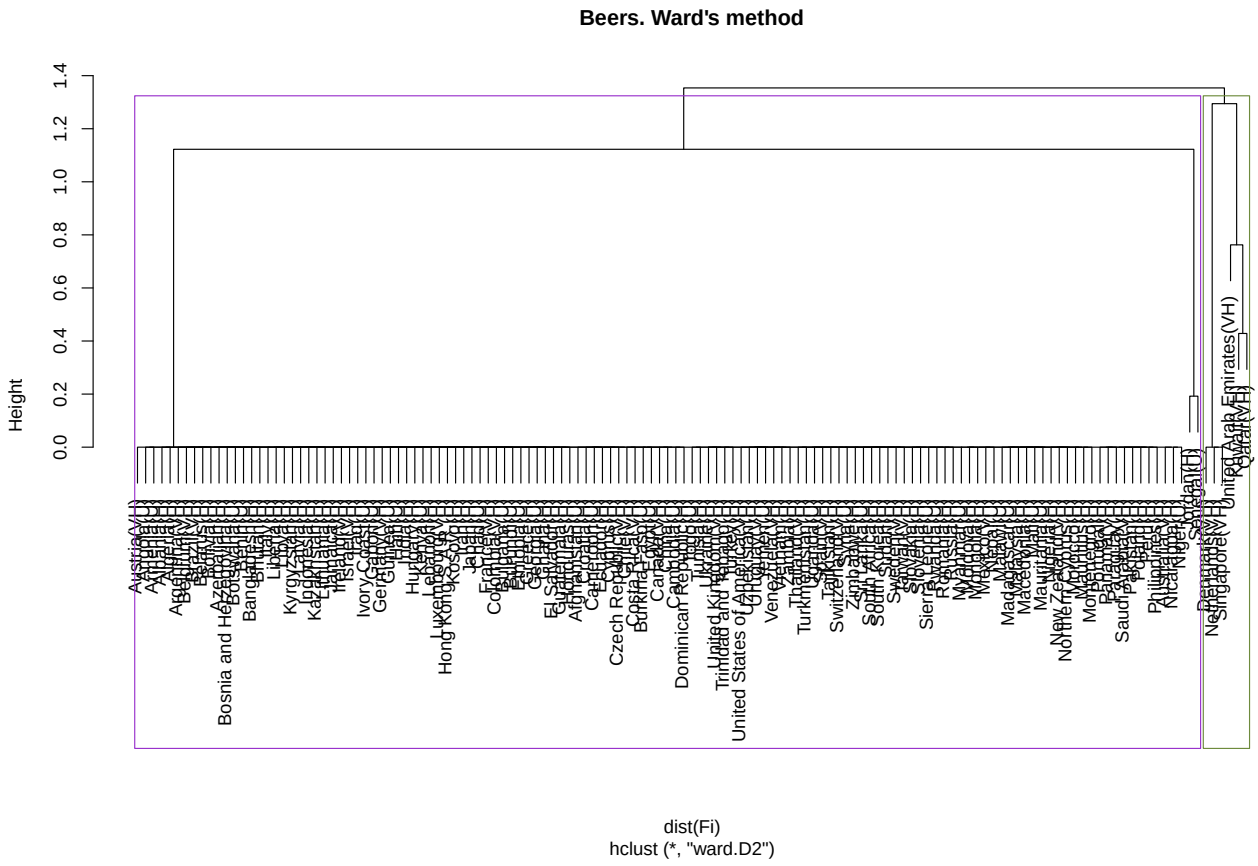


### 9.7.2 Four centers

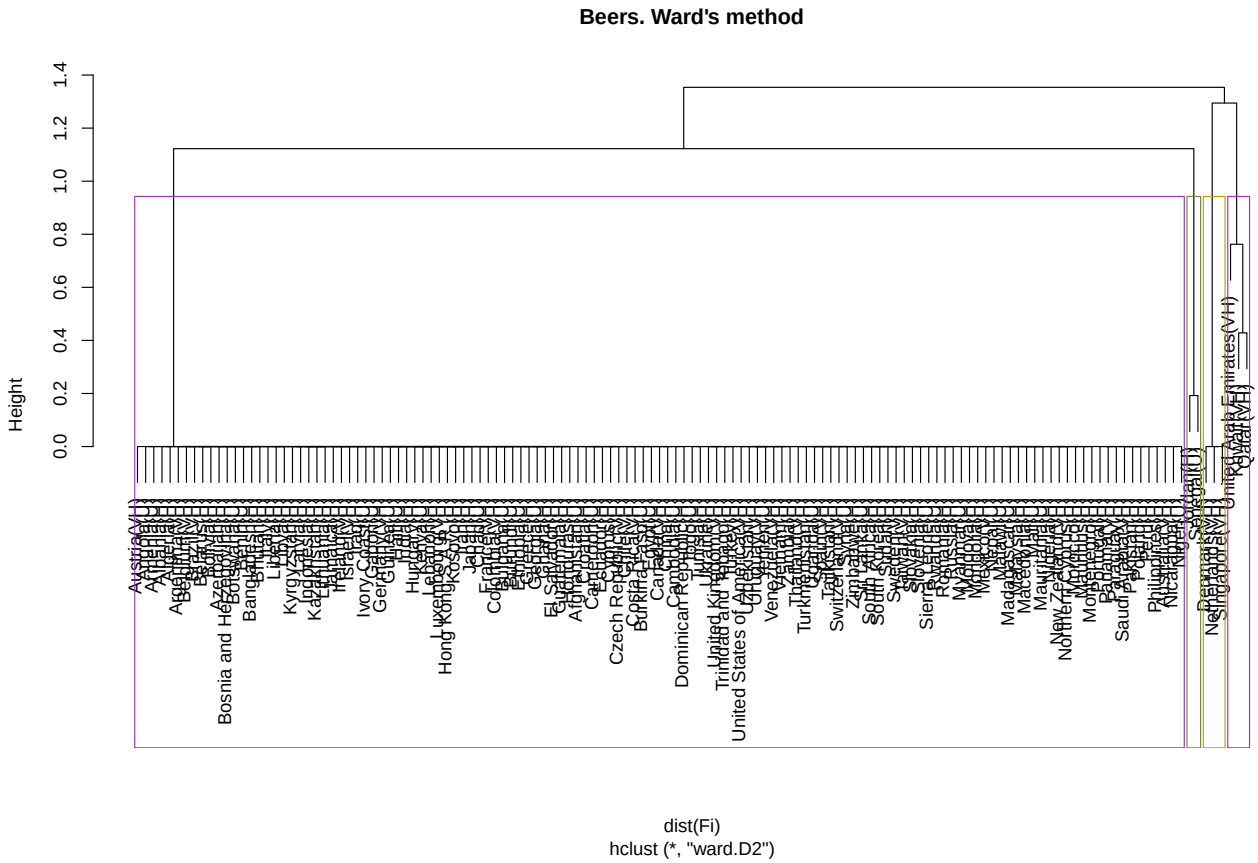


## 9.8 Cluster Analysis (hclust)

### 9.8.1 Two Clusters



### 9.8.2 Fours Clusters



# Chapter 10

# Conclusion for all Methods

| Methods | Unhappy | Normal | Very Happy | Reliability |
|---|---|---|---|---|
| **PCA** | Others | Temp & Rain | N/A | Components have significant contribution but convex hull has overlapping areas and Component 2 & 7 contradicts |
| **MCA** | warm summers, cold winters, high rain | N/A | Warm winter, cold summer, low rain | Components have significant contribution but convex hull has overlapping areas |
| **BADA** | Temp | Rain | Rain | Components have significant contribution but convex hull has overlapping areas |

| Methods | Unhappy | Normal | Very Happy | Reliability |
|---|---|---|---|---|
| **DiCA** | warm summers, cold winters, high rain | Higher variation in temperature is correlated with lower happiness | Warm winter, cold summer, low rain, windy | Convex hulls are separeted but second component only has temp variables as significant |
| **PLS-C** | Rain | Temp | Temp | Second component has more rain variables as significant than temp variables |
| **MFA** | Partial factors dominated by Temp, then rain and other variables | Neither of partial factors seems to have sufficient effect | Partial factors dominated by Temp and other variables, lesser effect of rain | Convex hull has overlapping areas |

- **MCA** and **DiCA** agrees:
  - Warmer winter, colder summer, low rain, windy cities makes people *happy*
  - Colder Winter, warmer summers, high rain, less windy makes people *unhappy*

However, even though **MCA** shows that most the variables has high contribution for the strongest signal in the data - **DiCA** shows that temp, rain and wind variables contributes significantly.

*Hence*,

- Happiness doesn't seem to be highly correlated to environmental conditions
- Temperature, rain and wind seem to be slightly correlated with happiness.
- Cluster Analysis doesn't seem to show any patterns in the data.

# Part IV

# Other Datasets

# Chapter 11

# Correspondence Analysis

## 11.1 Description

Correspondence Analysis (CA) is a multivariate graphical technique designed to explore relationships among categorical variables. The outcome from correspondence analysis is a graphical display of the rows and columns of a contingency table that is designed to permit visualization of the salient relationships among the variable responses in a low-dimensional space. Such a representation reveals a more global picture of the relationships among row-column pairs which would otherwise not be detected through a pairwise analysis.

**Calculate CA:**

- Step 1: Compute row and column averages
- Step 2: Compute the expected values
- Step 3: Compute the residuals
- Step 4: Plotting labels with similar residuals close together
- Step 5: Interpreting the relationship between row and column labels

**How to Interpret Correspondence Analysis Plots**

Correspondence analysis does not show us which rows have the highest numbers, nor which columns have the highest numbers. It instead shows us the relativities.

- The further things are from the origin, the more discriminating they are.
- Look at the length of the line connecting the row label to the origin. Longer lines indicate that the row label is highly associated with some of the column labels (i.e., it has at least one high residual).
- Look at the length of the label connecting the column label to the origin. Longer lines again indicate a high association between the column label and one or more row labels.
- Look at the angle formed between these two lines. Really small angles indicate association. 90 degree angles indicate no relationship. Angles near 180 degrees indicate negative associations.

## 11.2 Dataset - Weekly earnings by Race

- Data: Measurements of Weekly Earnings per Race
- Rows: There are 6 observations representing Asian/White/Black, Men/Woman.
- Columns: Total 6 variables grouping people based on Decile and Quartile ranges of their weekly income.

```
##                          White.men White.women Black.men Black.Women
## 1st decile                     412         374       361         331
## 1st quartile                   594         506       483         423
```

Table 11.1: Measurements of Weekly Earnings per Race

| | 1stQ | 2ndQ | 3rdQ |
|---|---|---|---|
| White.men | 594 | 326 | 547 |
| White.women | 506 | 237 | 397 |
| Black.men | 483 | 197 | 366 |
| Black.Women | 423 | 192 | 320 |
| Asian.Men | 648 | 481 | 731 |
| Asian.Women | 551 | 326 | 534 |

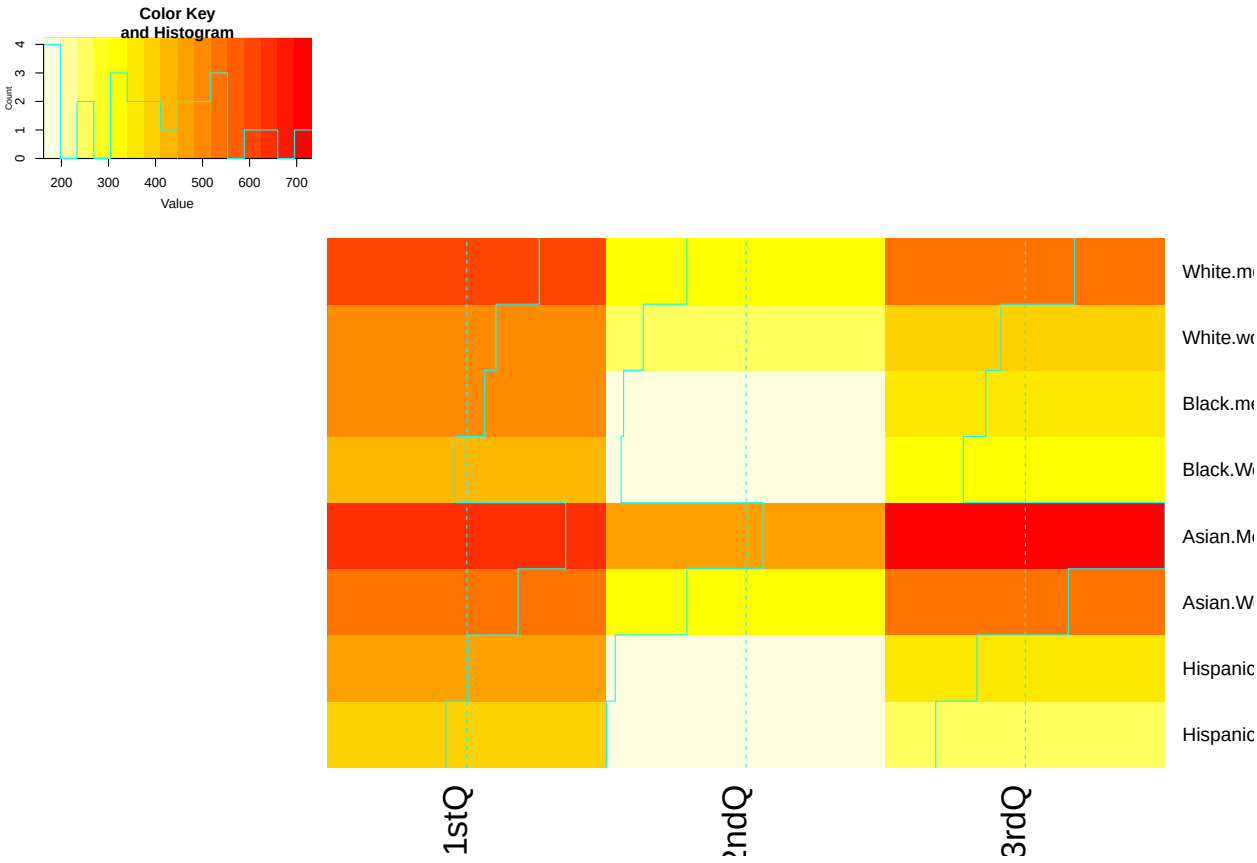
```
## 2nd quartile                      920         743       680         615
## 3rd quartile                     1467        1140      1046         935
## 9th decile                       2278        1726      1551        1453
## Total people (in thousands)     48746       36698      6445        7142
##                              Asian.Men Asian.Women Hispanic.Men
## 1st decile                         420         385          358
## 1st quartile                       648         551          451
## 2nd quartile                      1129         877          631
## 3rd quartile                      1860        1411          979
## 9th decile                        2699        2024         1498
## Total people (in thousands)       3684        2954        11142
##                              Hispanic.Women
## 1st decile                              320
## 1st quartile                            404
## 2nd quartile                            566
## 3rd quartile                            830
## 9th decile                             1266
## Total people (in thousands)            7168
```

However, here we can see that it may not be advisable to include Quartile and Decile intervals in the same analysis. Hence, we go ahead with Quartile Ranges only.

- Research Question

  - Does total earning of different races differ.
  - Which race get less than median salary (2nd Quartile)

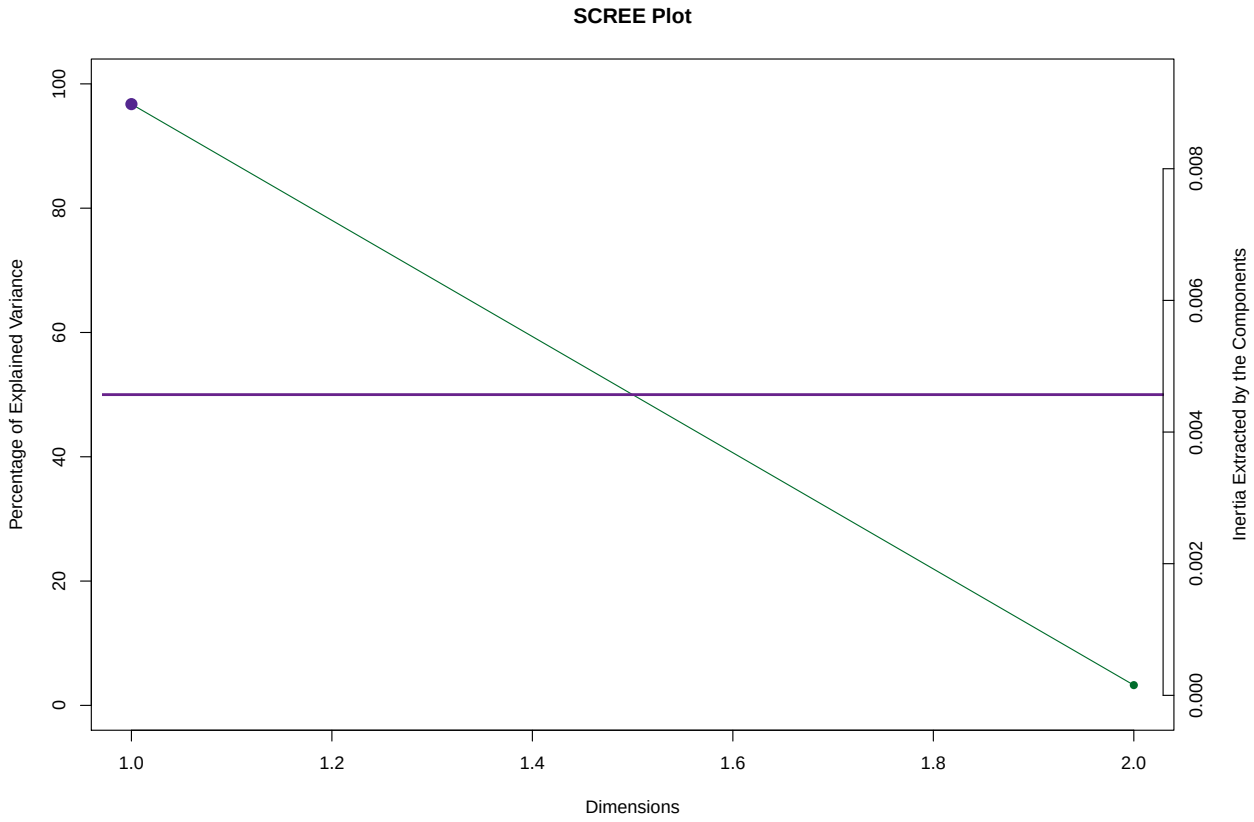
## 11.3 Heatmap




## 11.4 Scree Plot

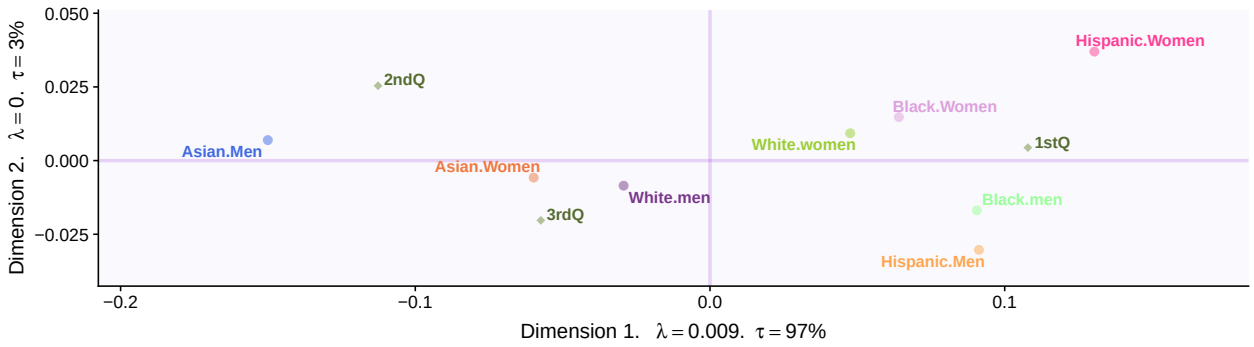
Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

P.S. The most contribution component may not always be most useful for a given research question.

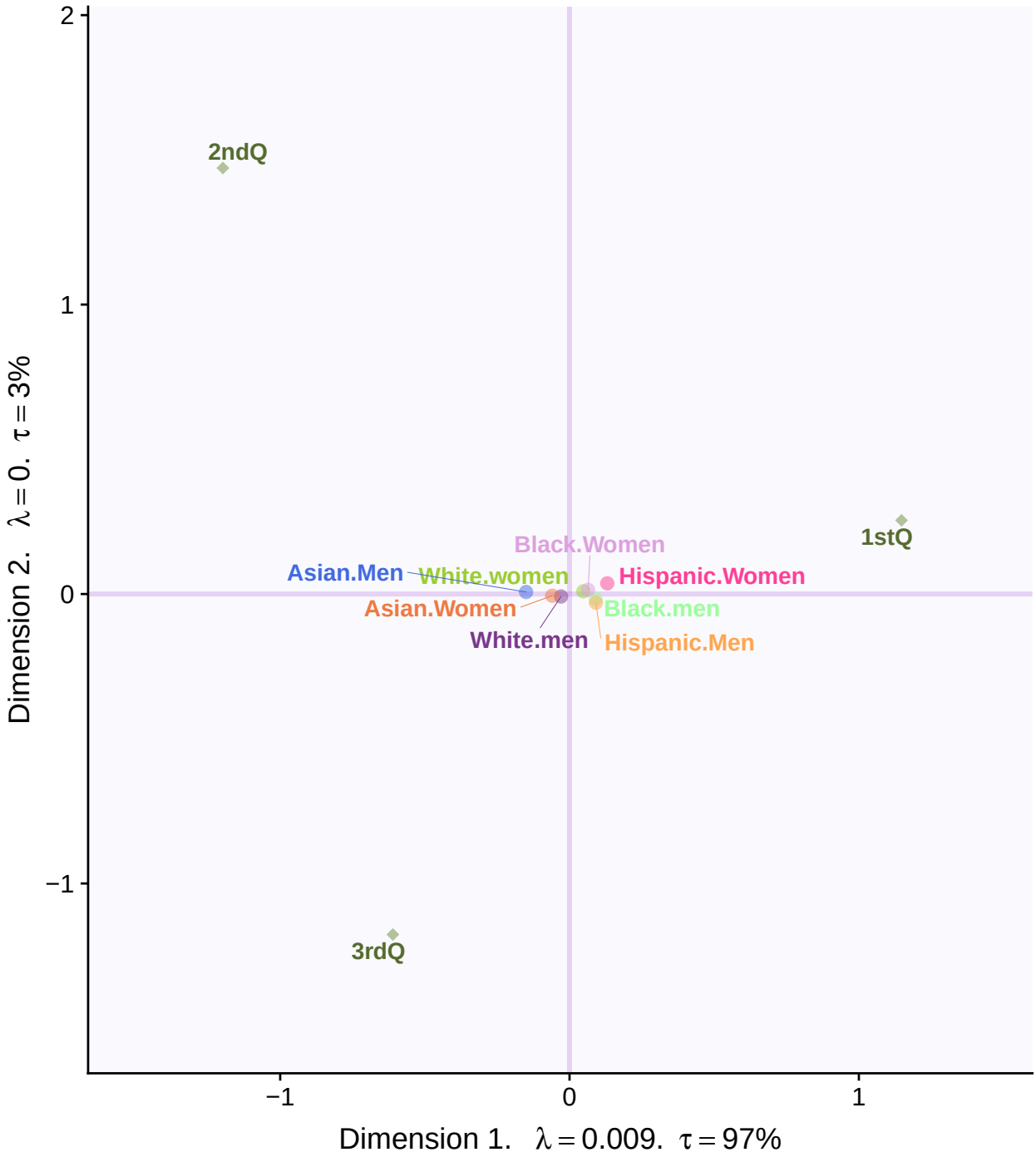



## 11.5 Factor Scores

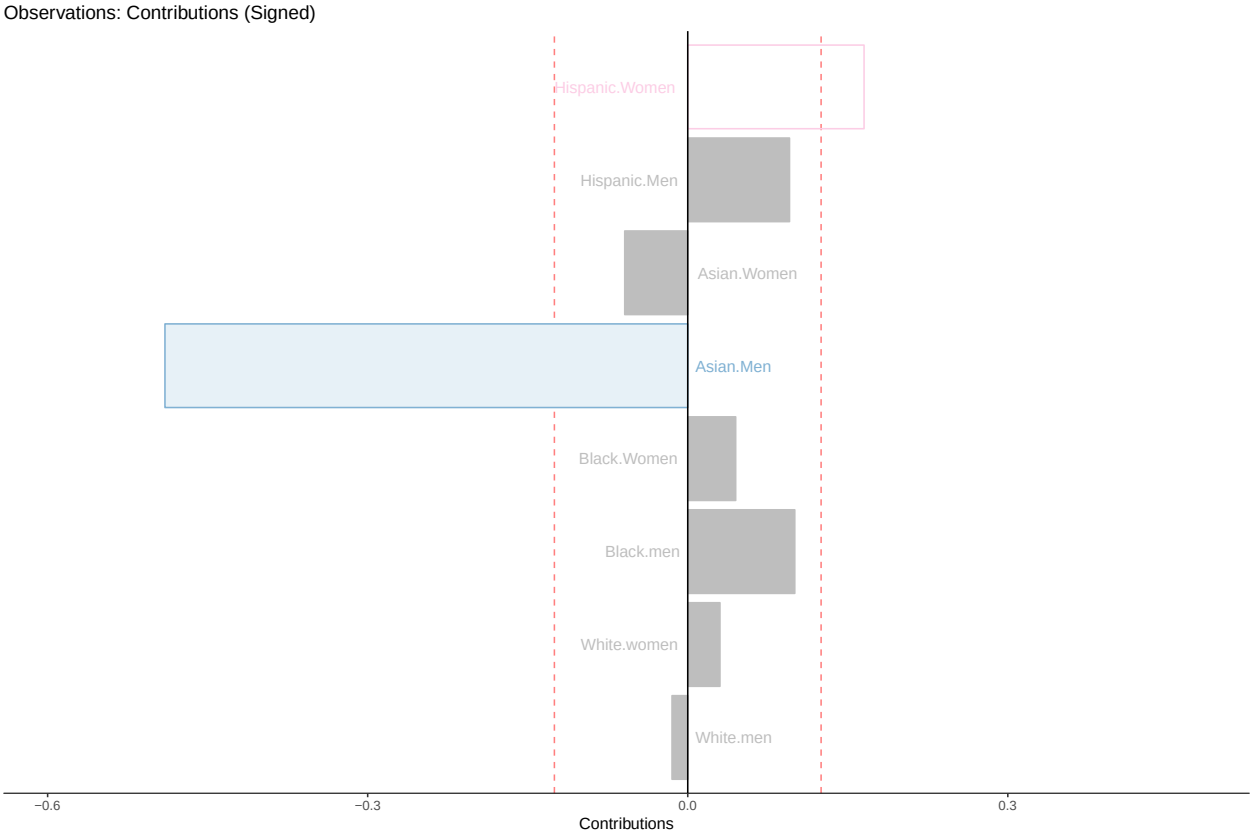
### 11.5.1 Symmetric Plot

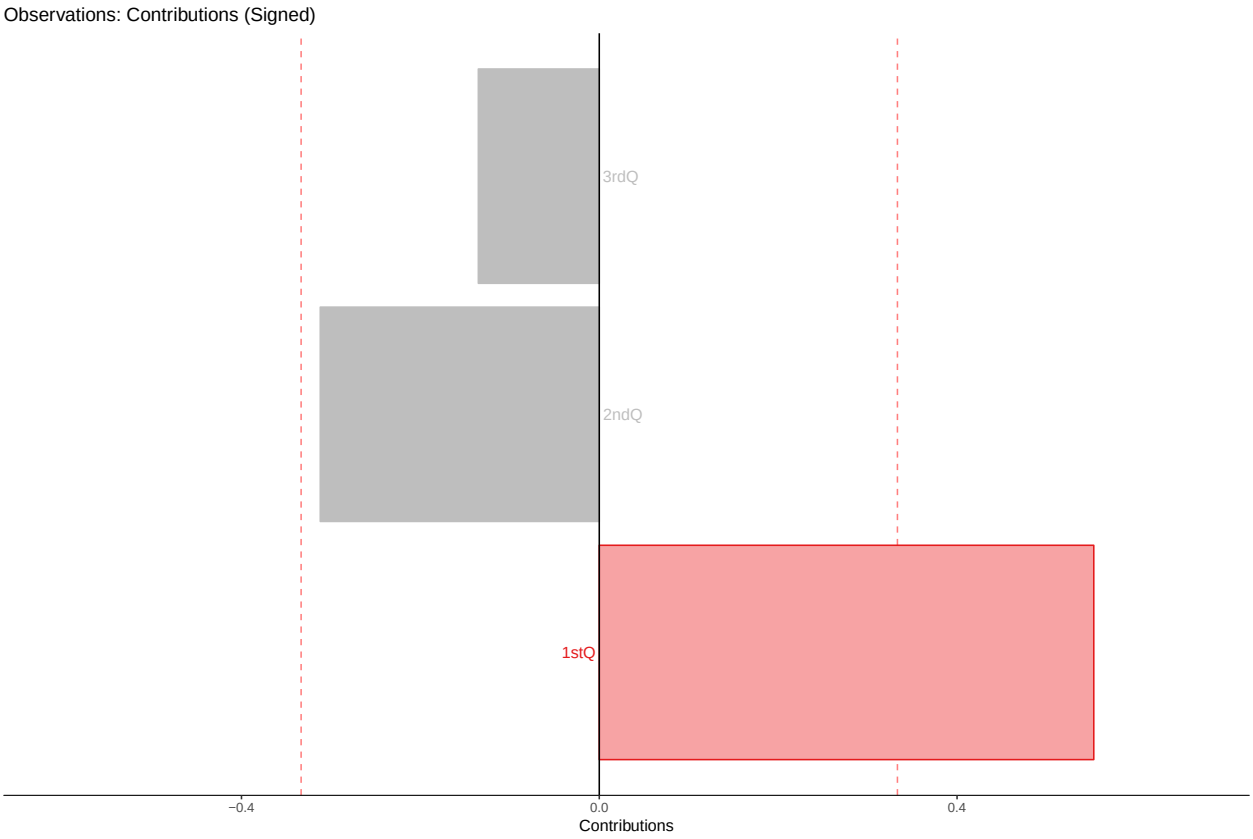


### 11.5.2 Asymmetric Plot

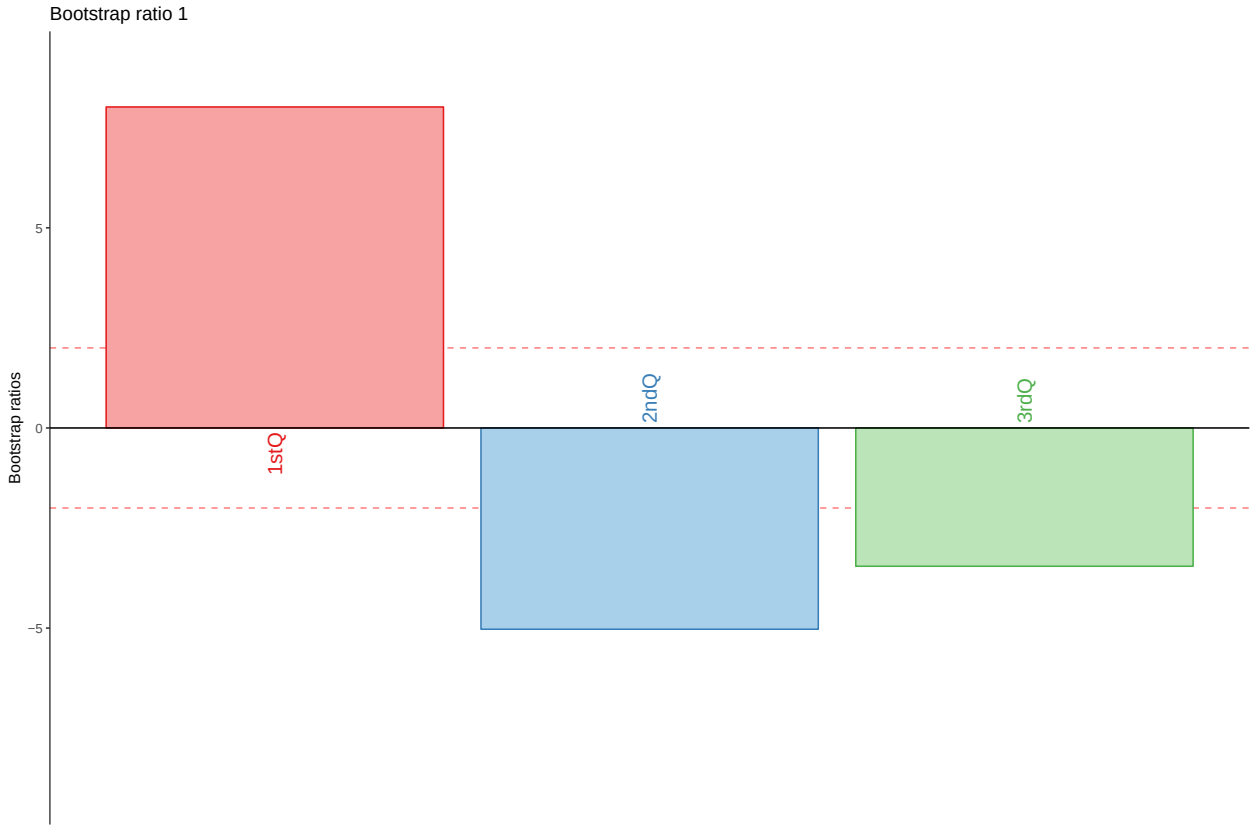


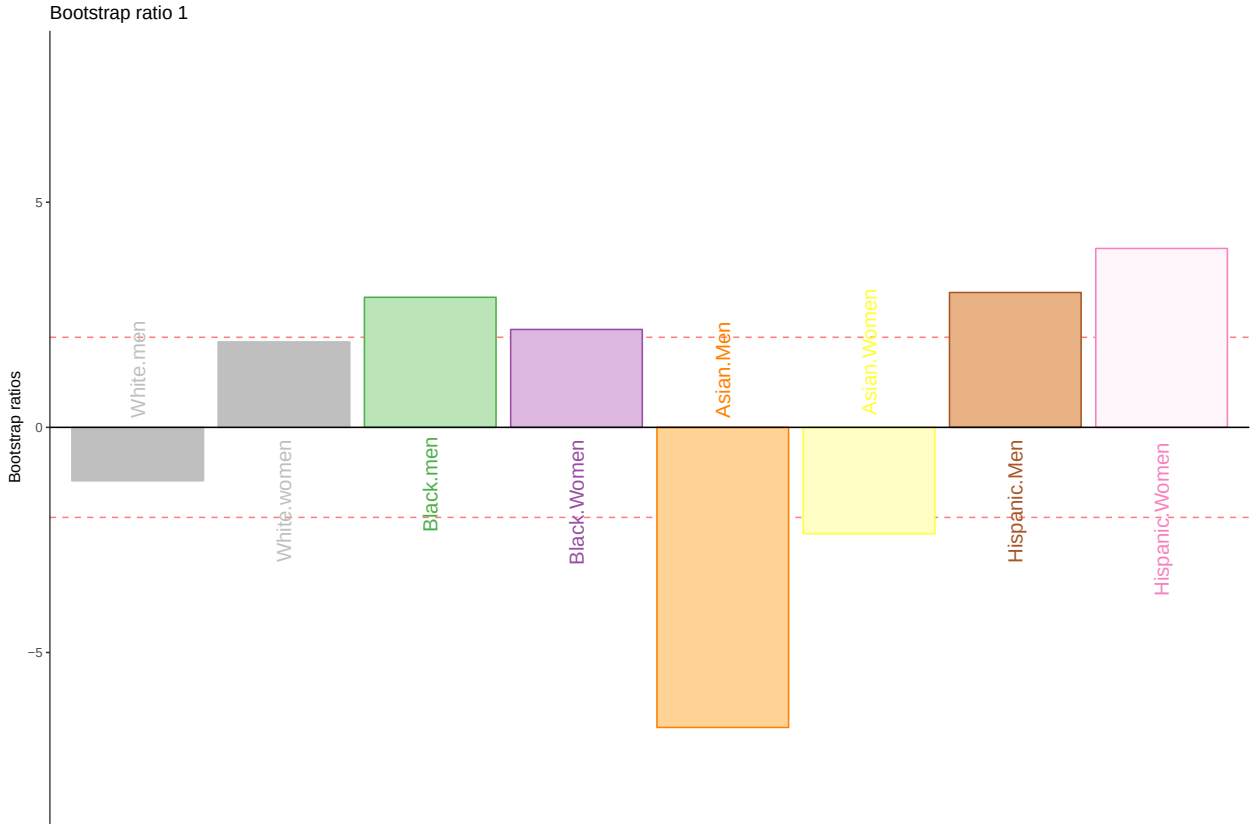
## 11.6 Most Contributing Variables




Observations: Contributions (Signed)
3rdQ
2ndQ
1stQ
−0.4
0.0
0.4
Contributions

## 11.7 Inference CA




Bootstrap ratio 1
Bootstrap ratios
5
0
−5
White.men
White.women
Black.men
Black.Women
Asian.Men
Asian.Women
Hispanic.Men
Hispanic.Women

# Chapter 12

# DiSTATIS

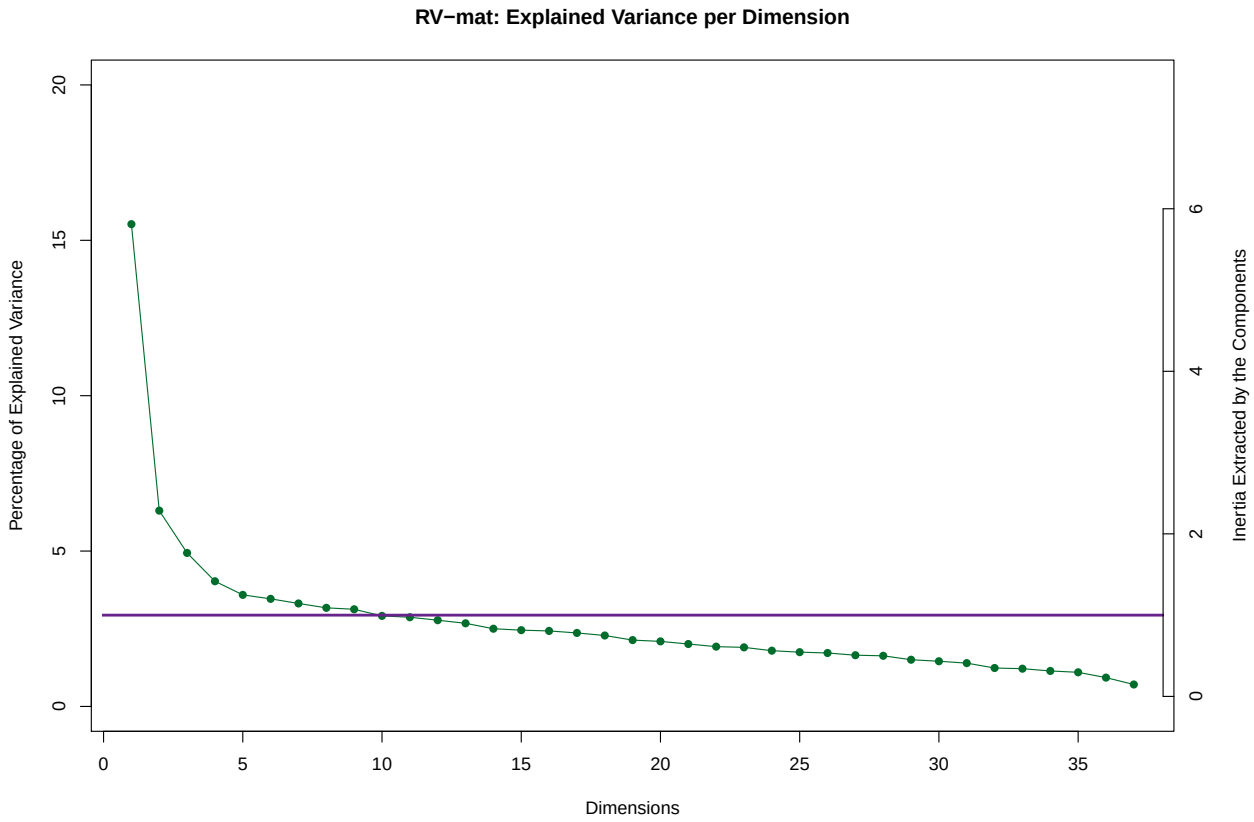
## 12.1 Description

DISTATIS is a new method that can be used to compare algorithms when their outputs consist of distance matrices computed on the same set of objects. The method first evaluates the similarity between algorithms using a coefficient called the RV coefficient. From this analysis, a compromise matrix is computed which represents the best aggregate of the original matrices. In order to evaluate the differences between algorithms, the original distance matrices are then projected onto the compromise. The goal of DISTATIS is to analyze a set of distance matrices. In order to compare distance matrices, DISTATIS combines them into a common structure called a compromise and then projects the original distance matrices onto this compromise.

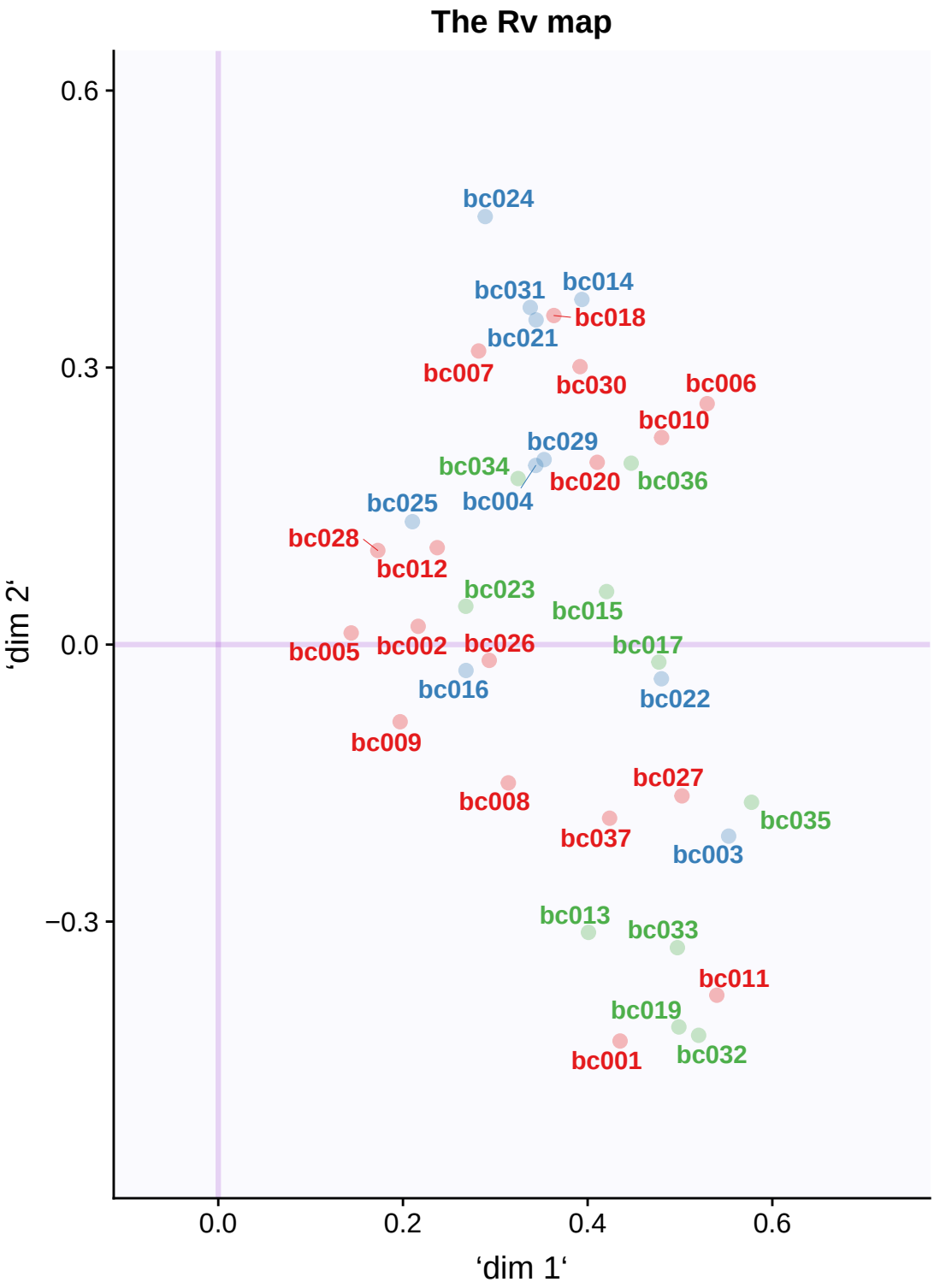
## 12.2 Dataset - Pianists for Composers

```
## [1] Bootstrap On Factor Scores. Iterations #:
## [2] 1000
```

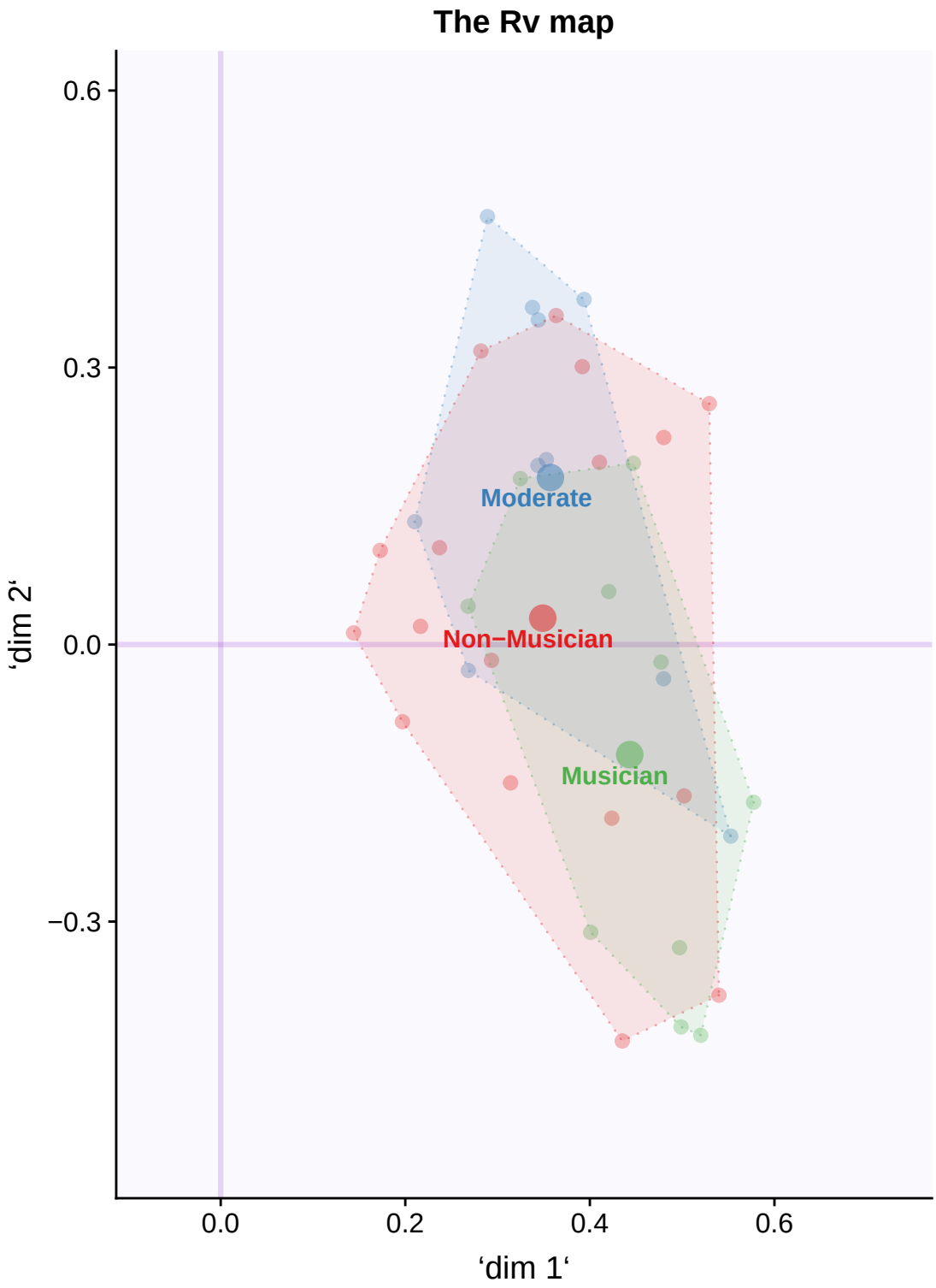
## 12.3 SCREE Plot - RV-MAT



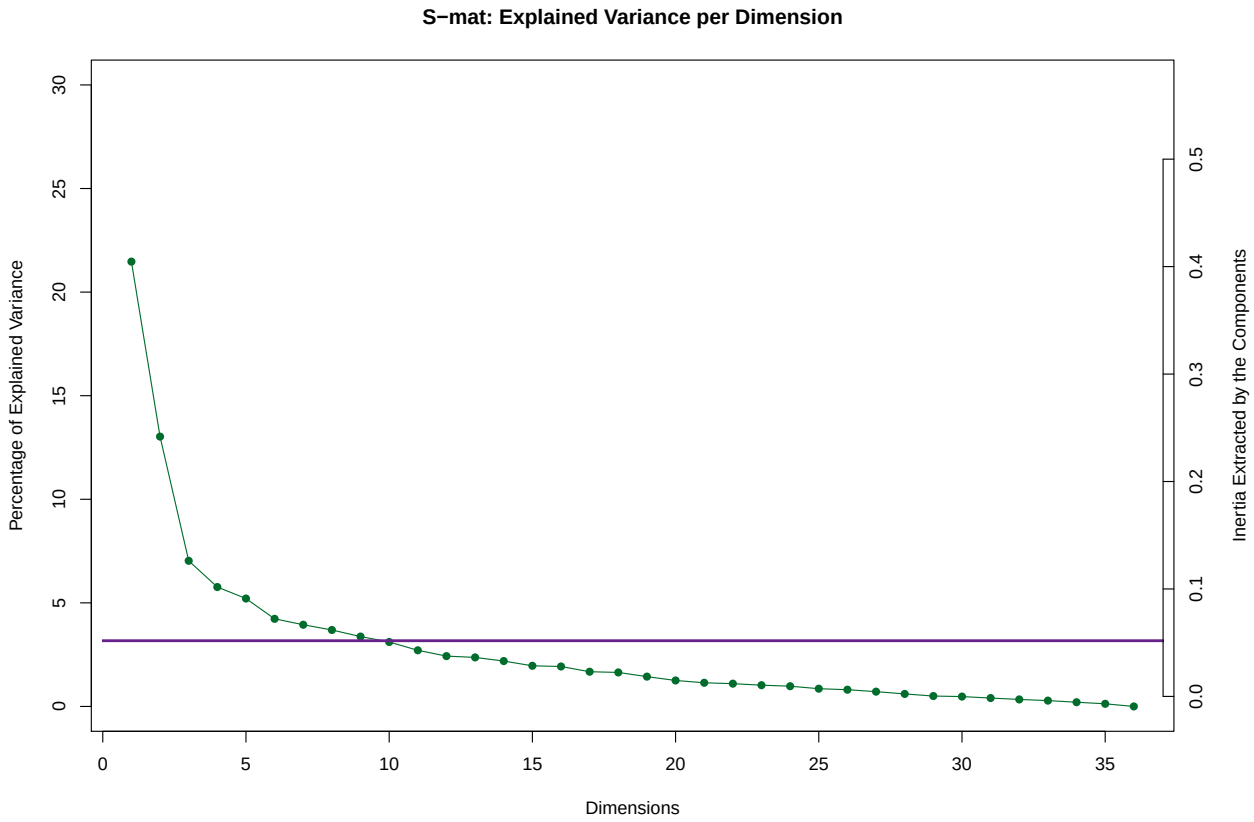
## 12.4 Plotting Assessor Matrix



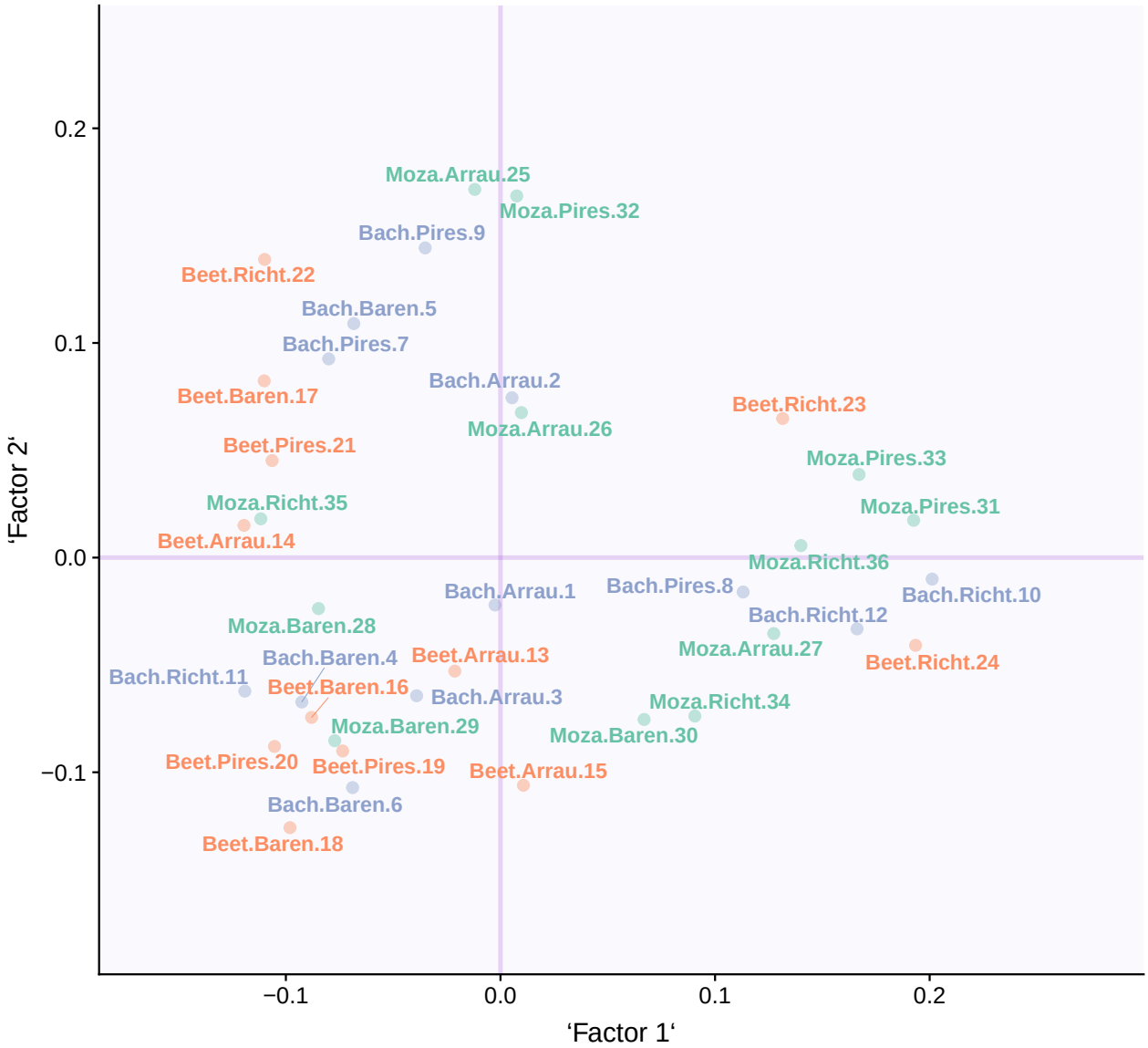
### 12.4.1 ConvexHull

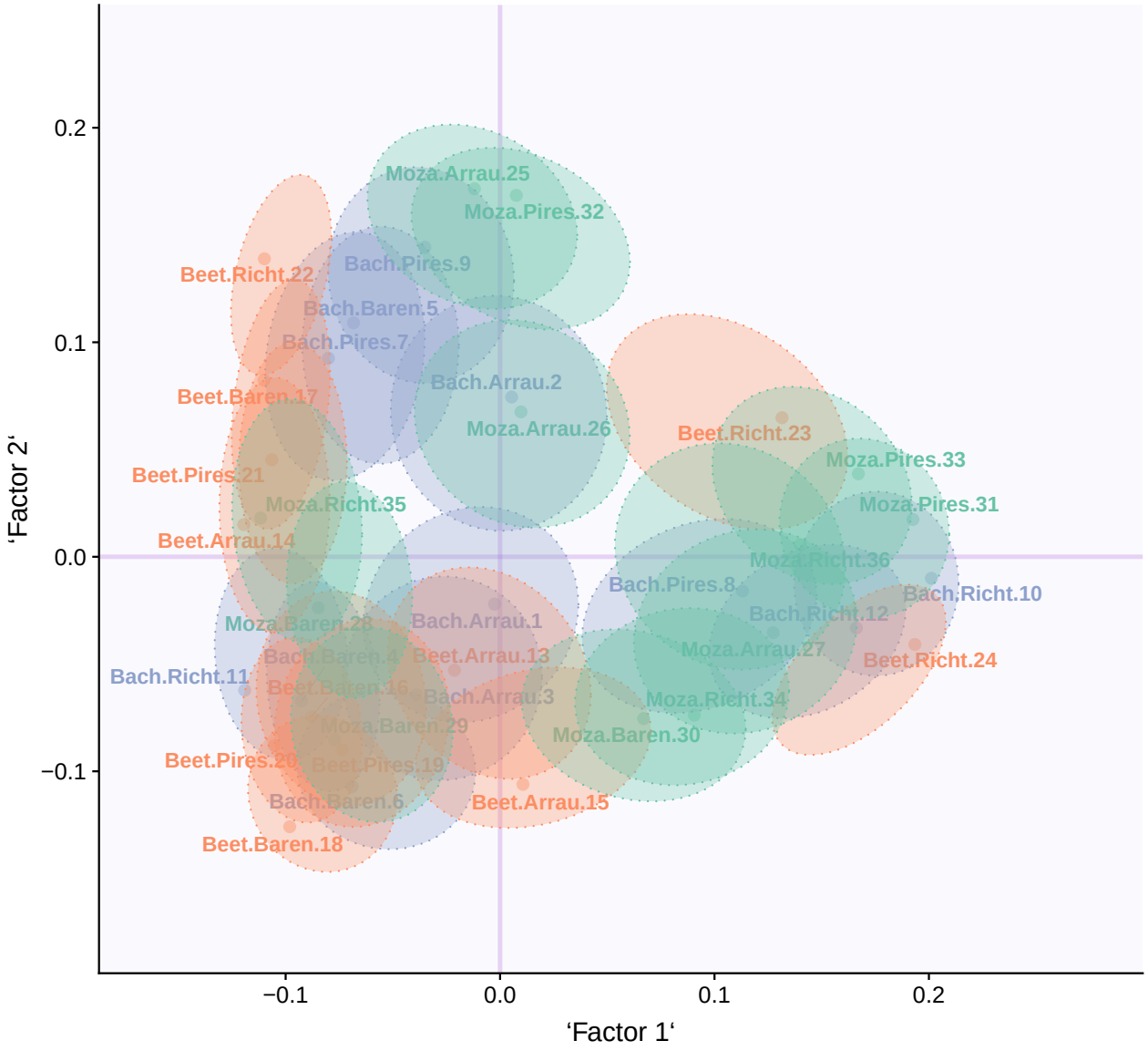


## 12.5 SCREE Plot - SV-MAT

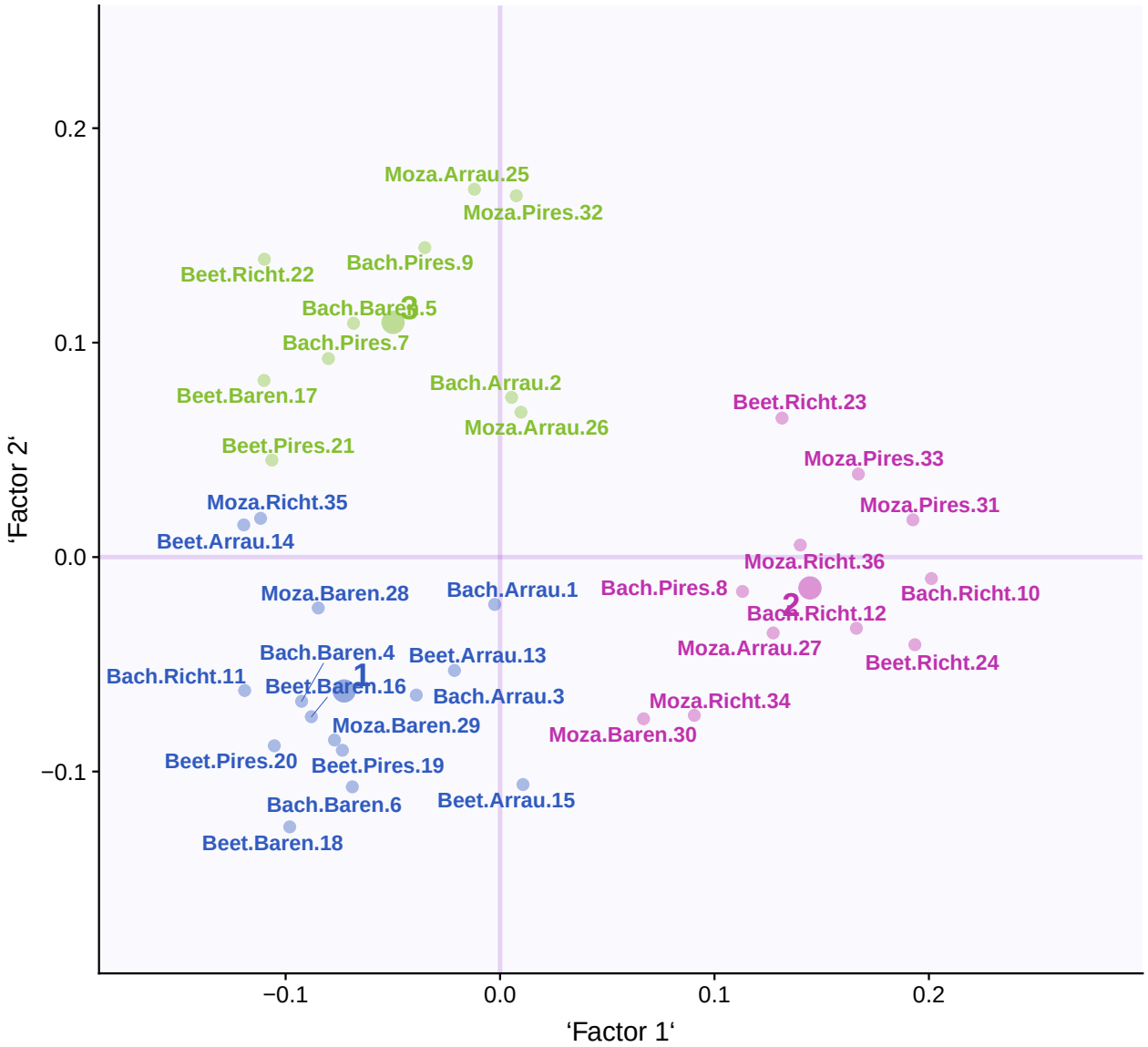


## 12.6 I Set

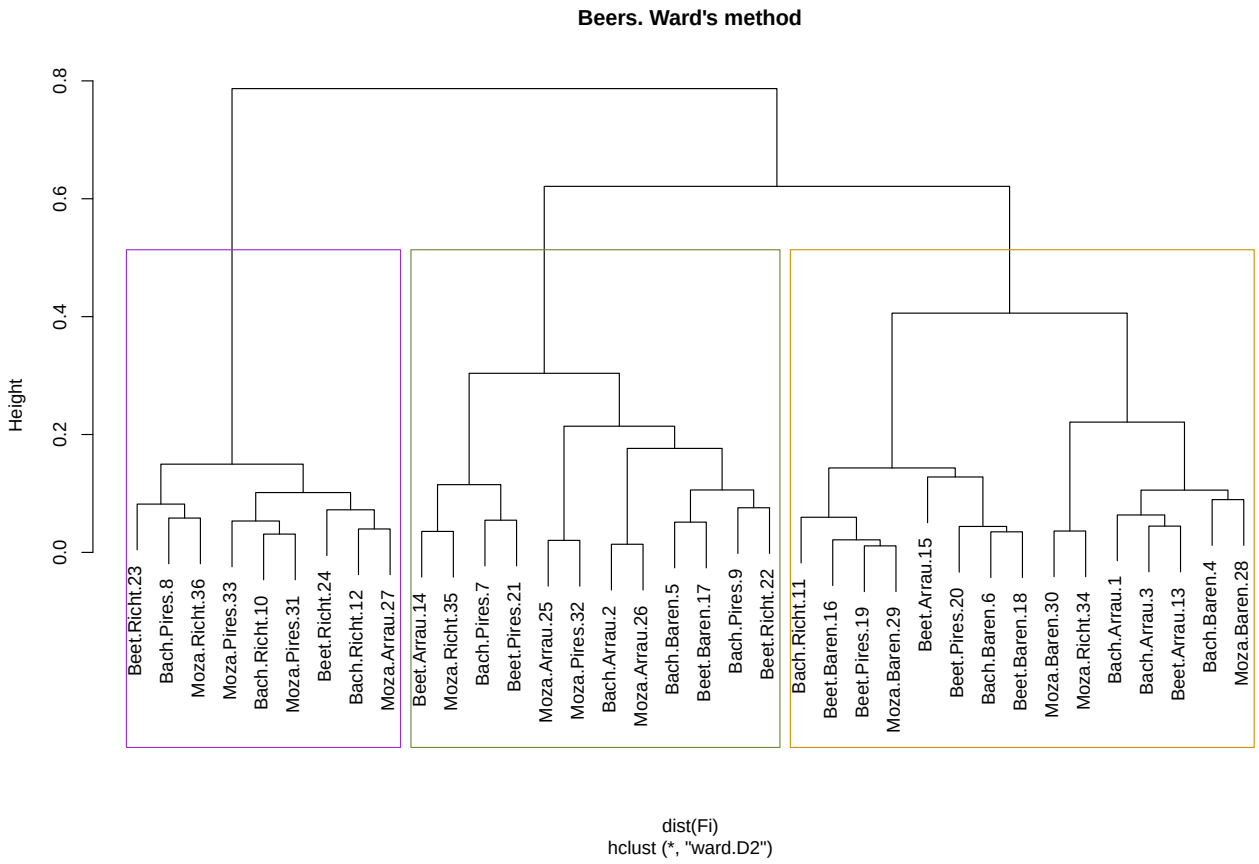

Moza.Arrau.25
Moza.Pires.32
Bach.Pires.9
Beet.Richt.22
Bach.Baren.5
Bach.Pires.7
Bach.Arrau.2
Beet.Baren.17
Beet.Richt.23
Moza.Arrau.26
Beet.Pires.21
Moza.Pires.33
Moza.Richt.35
Moza.Pires.31
Beet.Arrau.14
Moza.Richt.36
Bach.Arrau.1
Bach.Pires.8
Bach.Richt.10
Bach.Richt.12
Moza.Baren.28
Moza.Arrau.27
Bach.Baren.4
Beet.Arrau.13
Beet.Richt.24
Bach.Richt.11
Beet.Baren.16
Bach.Arrau.3
Moza.Richt.34
Moza.Baren.29
Moza.Baren.30
Beet.Pires.20
Beet.Pires.19
Beet.Arrau.15
Bach.Baren.6
Beet.Baren.18
0.2
0.1
0.0
−0.1
'Factor 2'
−0.1
0.0
0.1
0.2
'Factor 1'


Moza.Arrau.25
Moza.Pires.32
Beet.Richt.22
Bach.Pires.9
Bach.Baren.5
Bach.Pires.7
Bach.Arrau.2
Beet.Baren.17
Moza.Arrau.26
Beet.Richt.23
Moza.Pires.33
Beet.Pires.21
Moza.Richt.35
Moza.Pires.31
Beet.Arrau.14
Moza.Richt.36
Bach.Pires.8
Bach.Richt.10
Bach.Richt.12
Moza.Baren.28
Bach.Arrau.1
Moza.Arrau.27
Bach.Baren.4
Beet.Arrau.13
Beet.Richt.24
Bach.Richt.11
Beet.Baren.16
Bach.Arrau.3
Moza.Richt.34
Moza.Baren.29
Moza.Baren.30
Beet.Pires.20
Beet.Pires.19
Bach.Baren.6
Beet.Arrau.15
Beet.Baren.18
0.2
0.1
0.0
−0.1
−0.1
0.0
0.1
0.2
'Factor 2'
'Factor 1'

## 12.7 Cluster Analysis (K-Means)

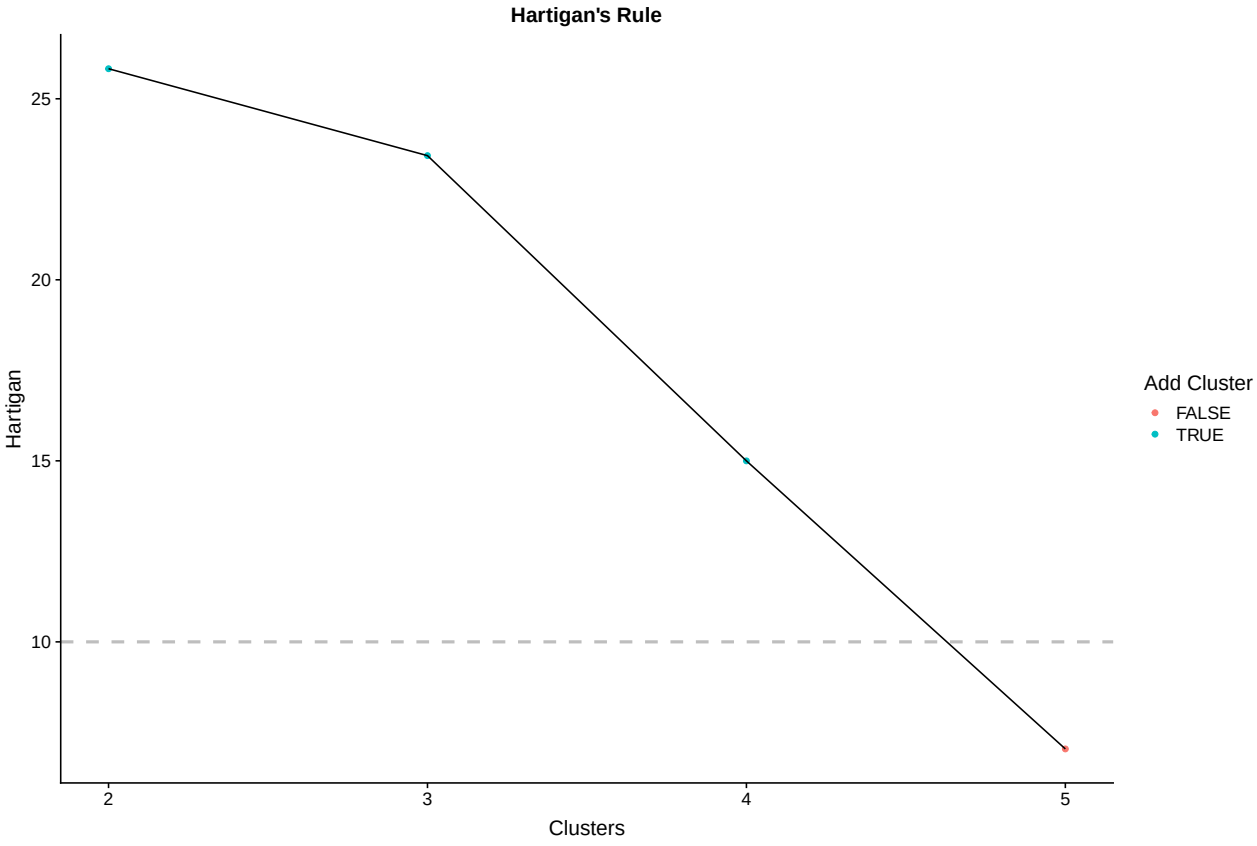


## 12.8 Cluster Analysis (hclust)




## 12.9 Cluster Analysis (Hartigan's Rule)

```
##   Clusters  Hartigan AddCluster
## 1        2 25.830051       TRUE
## 2        3 23.429672       TRUE
## 3        4 14.998705       TRUE
## 4        5  7.042848      FALSE
```


Hartigan's Rule
Hartigan
25
20
15
10
2
3
4
5
Clusters
Add Cluster
FALSE
TRUE