Multivariate Statisitcs using R

Athul Sudheesh

Contents

Chapter 1

Introduction

1.1 What is Multivariate Statisitcs ?

Multivariate Statistics is the area of statisitcs that deals with the analysis of data sets with more than one variable. The main objective is to study how the variables are related to each other, and how they work in combination to discriminate between the groups on whcih the observations are made.

1.2 Types of Analysis

There are many different models, each with it's own type of analysis. In this textbook we will be dealing mostly with the following ones (Abdi, 2003):

1. **Principal Component Analysis (PCA)** decompose the data table with correlated measurements into a new set of orthogonal variables.
2. **Correspondence Analysis (CA)** is a generalization of PCA to contigency table.
3. **Multiple Correspondence Analysis (MCA)** is a generalization of CA when several nominal variables are analyzed.
4. **Partial Least Squares (PLS)** methods relate the information present in two data tables that collect measurements on the same set of observations.
5. **Discriminant Analysis (DA)** is used when a set of independent variables are used to predict the group to which a given unit belongs.
6. **Multiple Factor Analysis (MFA)** combines several data tables into one single analysis
7. **Multi-dimensional Scaling (MDS)** is used to represent the units as points on a map such that their Euclidean distances on the map approximate the original similarities.
8. **Cluster Analysis (CA)** assign objects into groups so that objects from the same cluster are more similar to each other than objects from different groups.

1.3 Resampling

Resampling is a variety of methods for doing one of the following:

1. Estimating the precision of sample statisitcs by using subsets of available data (**jackknifing**) or drawing randomly with replacement from a set of data points (**bootstraping**)
2. Exchanging labels on data points when performing significance tests(**permutation tests**)
3. Validating models using random subsets (**cross validation**)

1.3.1 Jackknife

Jackknife was initially developed as a cross validation technique and was later extended to include variance estimation. In jackknife we start with estimating the parameter of interest from the whole sample. Then each variable is dropped from the sample and the parameter of interest is estimated from this smaller sample. This new estimates are called *partial estimates*. A *pseudo-value* is computed by finding the difference between *partial estimates* and whole sample estimates. These pseudo-values are used instead of the original values to make the estimate of the parameter of interest and their standard deviation is used to estimate the parameter standard error for computing confidence intervals.

1.3.2 Bootstrap

Bootstrap is a statistical method for inferences - standard error and bias estimates, confidence intervals, and hypothesis tests - without assumptions such as nomral distributions or equal variances.

1.3.3 Permutation Tests

A permutation test gives a simple way to compute the sampling distribution for any test statistic, under the strong null hypothesis that a set of genetic variants has absolutely no effect on the outcome. To estimate the sampling distribution of the test statistic we need many samples generated under the strong null hypothesis. If the null hypothesis is true, changing the exposure would have no effect on the outcome. By randomly shuffling the exposures we can make up as many data sets as we like. If the null hypothesis is true the shuffled data sets should look like the real data, otherwise they should look different from the real data. The ranking of the real test statistic among the shuffled test statistics gives a p-value.

1.4 Datasets

`World Health` Data

This dataset was downloaded from https://www.gapminder.org/data/ and was curated from WHO, UNICEF Child Info and MRI-HPA Center for Environment & Health. The data was preprocessed to isolate countries that were not missing data across a number of variables. It measures 175 countries (observations) on 16 quantitative variables(refer Table 1.1). The data was collected in 2002.

Correlation Plotting is one of the best first step processes in data analysis, for getting a sense of how different variables might be interacting with each other. Figure 1.1 shows the correlation plot of `World Health` data.

Table 1.1: Variables from World Health dataset

Variable.Names	Description
Fires	Fires age adjusted morality per 100
Drownings	Drownings age adjusted morality per 100
Poisonings	Poisonings age adjusted morality per 100
Falls	Falls age adjusted morality per 100
T.Chole_F	Total Cholesterol..mmol.L (Female)
T.Chole_M	Total Cholesterol..mmol.L (Male)
BP_F	Average Systolic Blood Pressure(Female)..mm.Hg.
BP_M	Average Systolic Blood Pressure(Male)..mm.Hg.
BMI_F	Body Mass Index (Female)
BMI_M	Body Mass Index (Male)
GenGovExp	General government expenditure on health as % of total government expenditure
TotalExp_GDP	Total expenditure on health as % of gross domestic product
PerCapitaExp_XchangeRate	Per capita total expenditure on health at average exchange rate..US
PerCapita_PPP	Per capita total expenditure on.health..PPP.int.
IMM_MCV	One year olds immunized with MCV
IMM_DTP3	One year olds immunized with three doses of DTP

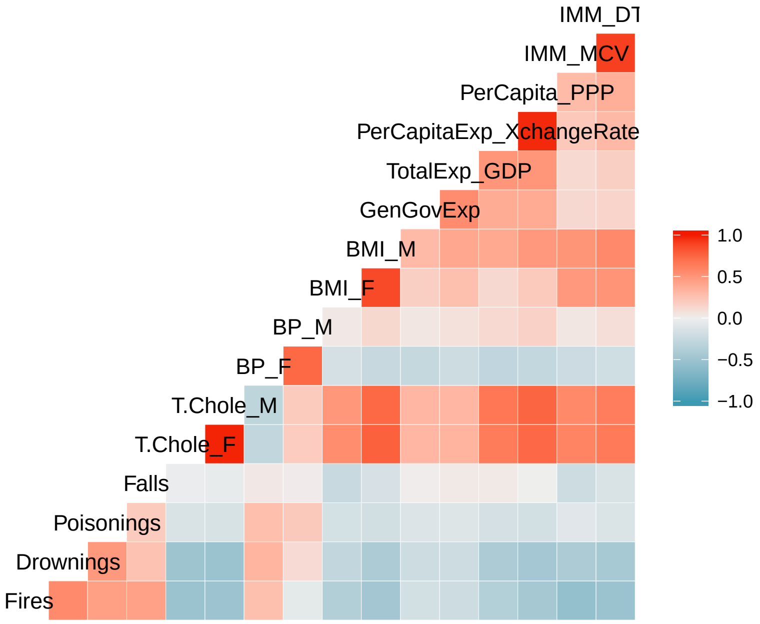


Figure 1.1: Correlation Plot of World Health Dataset

For most part of this textbook we will be only using a subsection of this dataset namely `WorldHealth_Risk` with the variables `T.Chole_F`, `T.Chole_M`, `BP_F`, `BP_M`, `BMI_F` and `BMI_M`.

Orange Juice Rating Data

This is the dataset from an experiment where a set of 10 orange juice were rated on 22 descriptors. The number at the intersection of a row and a column indicates the number of participants who rated the column(variable) relevant for the row(observations). Here the dataset is a contigency table (frequency count) and hence a correlation plot doesn't make much sense. So we do a heatmap of the given table(Ref. Figure 1.2).

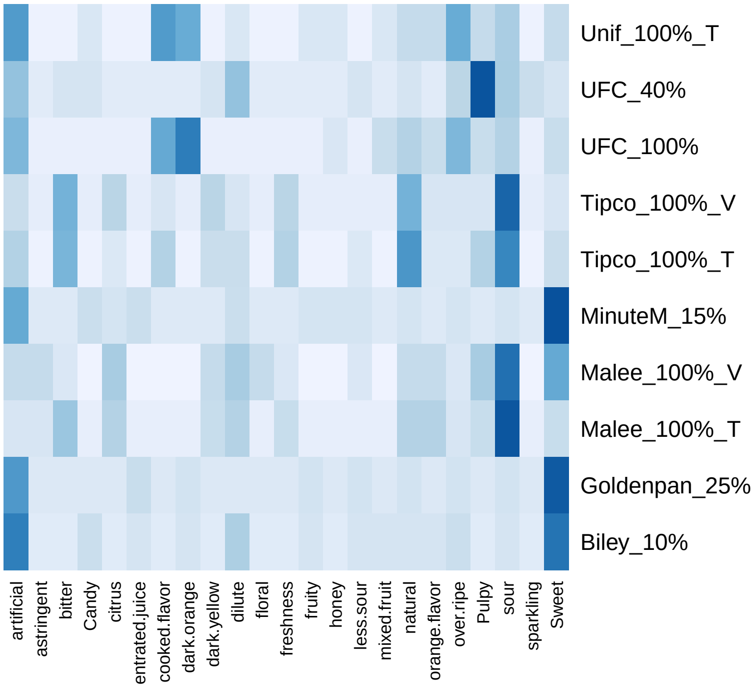


Figure 1.2: Heatmap of Orange Juice Rating

Musical Sorting Data

These are data from a sorting experiemnt conducted at The University of Texas at Dallas in Dr. Dowling's lab on 36 clips of classical music, each composed by 1 of 3 composers - Bach, Beethoven and Mozart. The experiment had 37 subjects to do the sorting.

Part I

One-Table Analysis

Chapter 2

Principal Component Analysis

Principal Component Analysis (PCA) is one of the most popular as well as the oldest multivariate statisitcs technique. It is commonly used for extracting the most important information from the data table and compressing the size of the dataset by keeping only this important information. This extracted infromation is expressed as a set of orthogonal variables called *principal components*. The first principal component is the line that maximizes the inertia of the cloud of the data points and subsequent components are defined as orthogonal to previous components that maximizes the remaining inertia. In this new representation, observations are represented as *factor scores* and variables as *loadings*. PCA works the best when the dataset is quantitative in nature(Herve Abdi, 2010b).

Suppose the data table to be analyzed is X and have the shape I (Observations) $\times$ J (Variables), then X has the following singular value decomposition [SVD]:

$$X = P\Delta Q^T$$

where $\mathbf{P}$ is the $I \times L$ matrix of left singular vectors, $\mathbf{Q}$ is the $J \times L$ matrix of right singular vectors, and Δ is the diagonal matrix of singular values.

2.1 Finding the components

In PCA, the components are obtained by the singular value decomposition of the given data table X. The factor score are computed as

$$\mathbf{F} = \mathbf{P}\boldsymbol{\Delta}$$

and the loadings $\mathbf{Q}$ are the coefficients of the linear combinations used to compute these factor scores.

$$\mathbf{F} = \mathbf{P}\boldsymbol{\Delta} = \mathbf{X}\mathbf{Q}$$

The $\mathbf{Q}$ can also be used to project new observations onto the components, called supplementary projections. The factor scores for supplementary observations are compted as

$$\mathbf{f}_{sup}^T = \mathbf{x}_{sup}^T\mathbf{Q}$$

.

Interpreting PCA

Let’s understand how to interpret the results of PCA using `WorldHealth_Risk` data as an example.

2.2 Inertia Explained by a Component

The given dataset had six variables and from Figure 2.1, we can see that PCA has generated six components. Figure 2.1 is called a scree plot and it helps us to understand how many principal components are there and how much variance each of them explain. The percentage of inertia is computed as

$$\tau = \frac{\lambda_i}{\sum_i \lambda_i} 100$$

where λ_i is the eigenvalue of the i^{th} component.

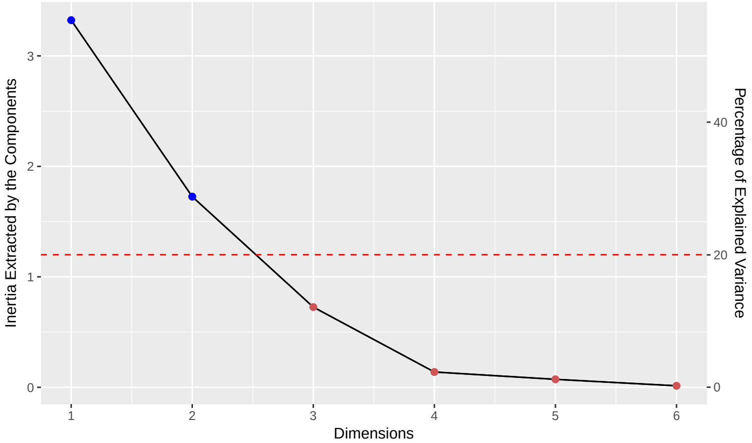


Figure 2.1: Scree Plot for Inference PCA

In our example the first two components explain most of the infomration, around 84% and by applying elbow rule only the first two components are important. We have also used inference analysis with our model to compute the quality of the model. The blue dots in the scree represent significant components and red dots represent insignificant ones.

2.3 Contribution of an Observation to a Component

The importance of an observation for a component can be obtained by the ratio of the squared factor score of this observation by the eigenvalue associated with that component. This ratio is called the contribution of the observation to the component.

$$ctr_{i,l} = \frac{f_{i,l}^2}{\lambda_l}$$

The contribution of all observations to first component is shown in figure 2.2

2.4 Factor Scores

Factor scores are the observations in the new representation space. Figure 2.3 shows the factor map for PCA of `World_Health Risk`. The oservations were also color-coded by their geographic region for better understanding of the data.

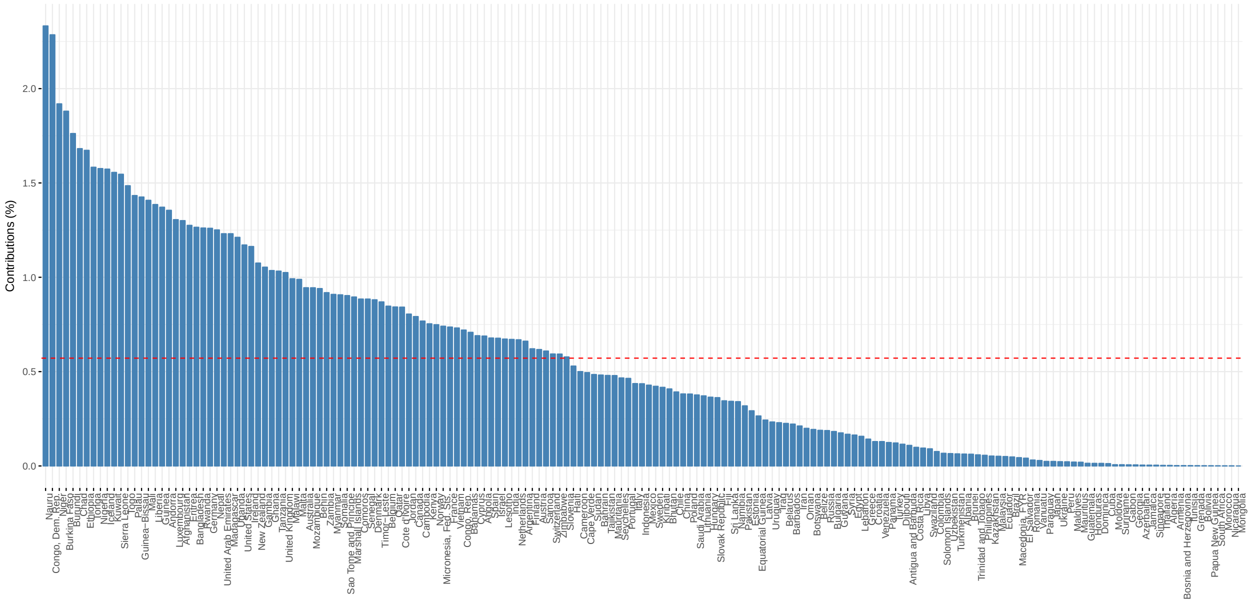


Figure 2.2: Contribution of observations to first component

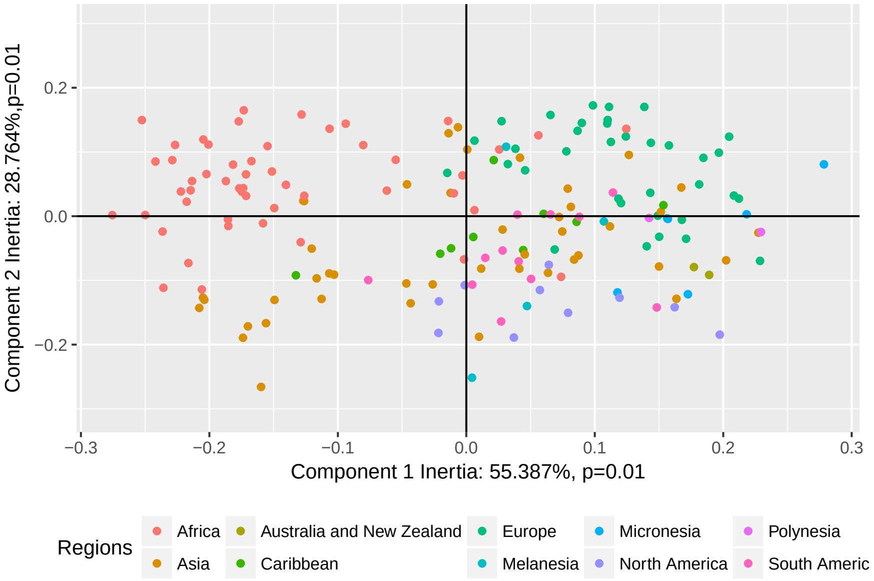


Figure 2.3: PCA World Health Characteristics. Factor scores of the observations plotted on the first two components

From figure 2.3 we can see that component 1 is seperating Africa and Asia from America and Europe, while component 2 is seperating Africa and Europe from Asia and America. To find the variables that account for these differences, we examine the loadings of the variables on the first two components(figure 2.5).

2.5 Contribution of Variables to Components

Examining the contribution of variables to the components helps us to understand the variables explained by each component. Figure 2.4 shows the contribution of variables to the first two components in our example.

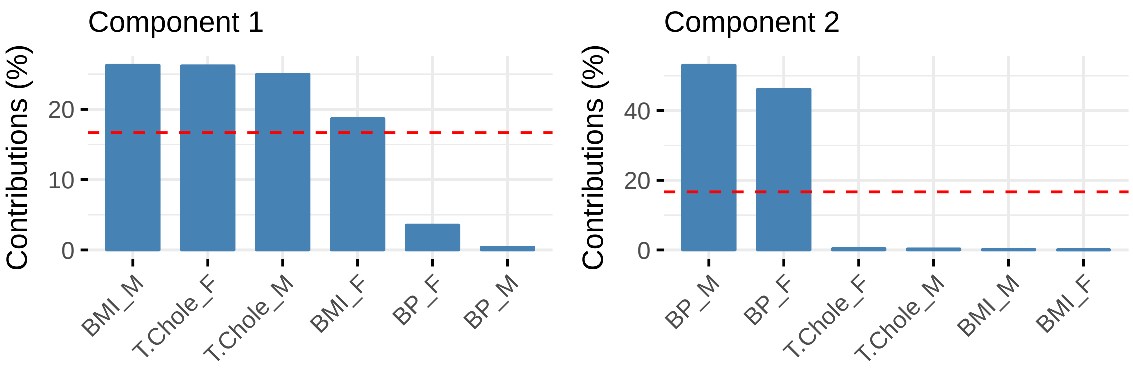


Figure 2.4: Contribution of variables to first two components

From figure (fig 2.4), we can see that component 1 explains most of BMI measure and Total Cholesterol measure whereas component 2 explains most of Blood Pressure measures

2.6 Loadings: Correlation of a Component and a Variable

The correlation between a component and a variable estimates the information they share. In the PCA framework, this correlation is called loadings.

The sum of squared coefficients of correlation between a variable and all the components is equal to one and hence by the property of a circle, if the data can be perfectly explained by the first two components then the loadings will be positioned on a circle, which is called the circle of correlations. When there are more than two components, the variables are positioned inside the circle of correlation. The closer a variable is to the circle of correlations, the better we can construct this variable from the first two components.

The circle of correlation for PCA on `WorldHealth_Risk` is shown in figure 2.5

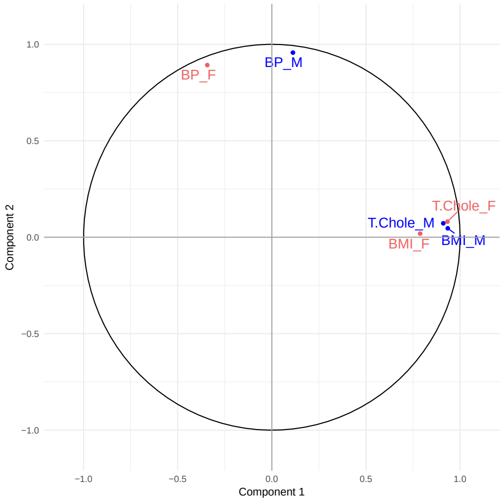


Figure 2.5: Correlation (and circle of correlations) of the variables with component 1 and 2

2.7 Bootstrap Ratios

Bootstrap ratios tell us the significance of the variables in our model. From figure 2.6, we can see that BP_F was significant for both the components. The bars with orchid color are the ones that are significant for component 1 alone and the bars with green are the ones that are significant for component 2 alone. The bars with gray color indicate insignificant variables.

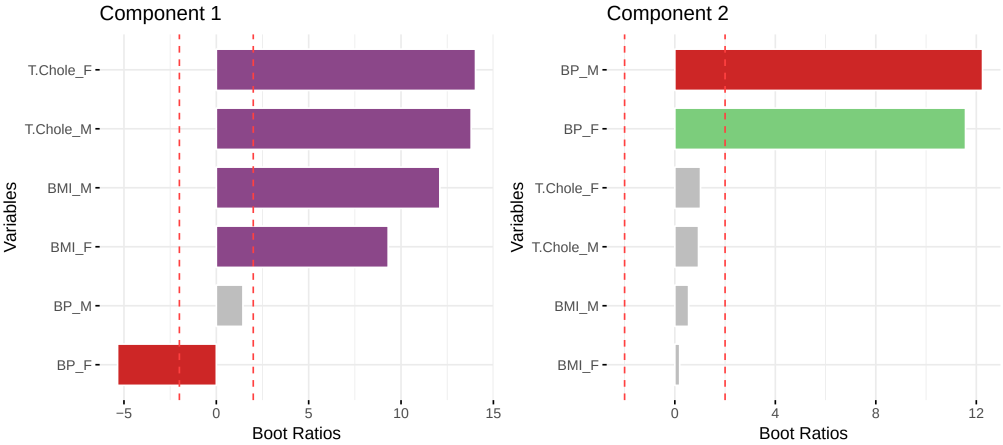


Figure 2.6: Bootstrap Ratios for Component 1 and 2

Chapter 3

Correspondence Analysis

Correspondence Analysis is a generalized principal component analysis tailored for the analysis of qualitative data. The goal of correspondence analysis is to transform a data table into two sets of factor scores: One for rows and one for columns. The factor scores give the best representation of the similarity structure of the rows and the columns of the table. In CA, the factor scores of the rows and columns have the same variance, and therefore both rows and columns can be represented in one single map called the biplot. The technique is also known by the names of *optimal scaling*, *dual-scaling* and *reciprocal averaging*(Herve Abdi, 2010a).

3.1 Computations

For this analysis we use the `Orange Juice Rating` dataset.

The first step of the analysis is to transform the data matrix into a probability matrix (denoted $\mathbf{Z}$)[The row totals of $\mathbf{Z}$ is denotes as $\mathbf{r}$ and column totals of $\mathbf{Z}$ is denotes as $\mathbf{c}$]. The probability matrix obtained in the first step is double centered by substracting $\mathbf{rc}^T$ from $\mathbf{Z}$. The heatmap of this matrix is shown in figure 3.1. The factor scores are obtained by the generalized signular value decomposition of this matrix. i.e

$$(\mathbf{Z} - \mathbf{rc}^T) = \mathbf{P\Delta Q^T}$$

From the GSVD, the row and column factor scores are obtained as:

$$\mathbf{F} = \mathbf{D_r^{-1} P \Delta}$$

and

$$\mathbf{G} = \mathbf{D_c^{-1} Q \Delta}$$

where $\mathbf{D_c} = diag\{\mathbf{c}\}$ and $\mathbf{D_r} = diag\{\mathbf{r}\}$

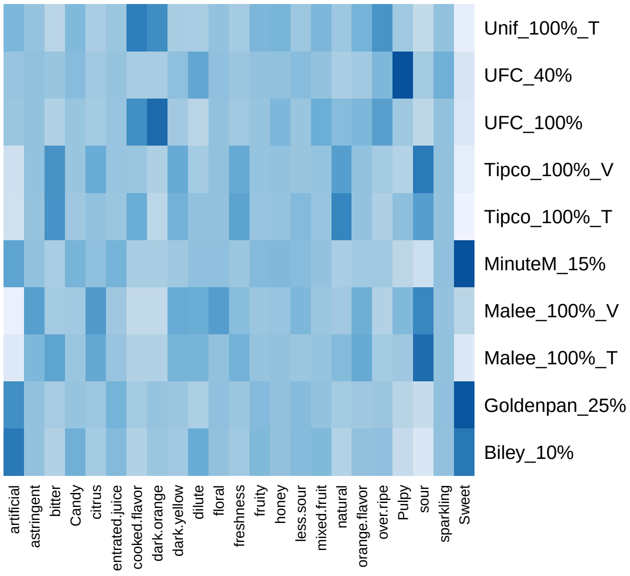


Figure 3.1: Heatmap of double centered probability matrix used for GSVD of Correspondence Analysis

3.2 Eigenvalues/Variances

As in PCA, here also we examine the eigenvalues to determine the number of axis to be considered in our interpretation. The Scree Plot of CA for `Orange Juice Rating` is shown in figure 3.2

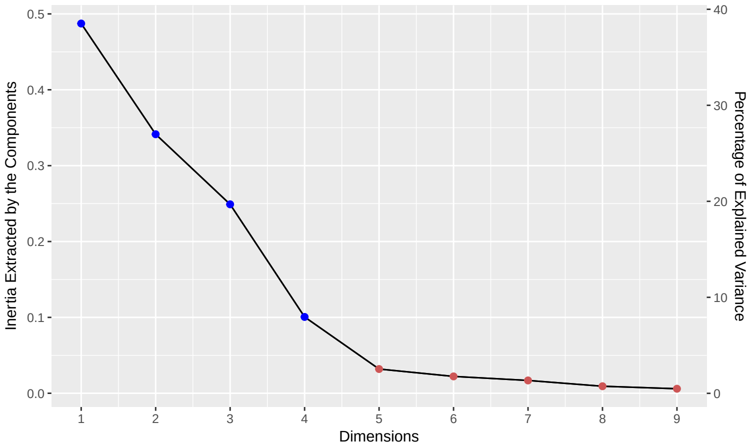


Figure 3.2: Scree Plot for Inference CA

From the above figure, we can see that most of the variance is explained by the first three factors (almost 85%). On inference analysis these three dimensions were aslo found to be significant and hence we consider these three dimensions for our further analysis.

3.2.1 Elements Important for a Factor

In CA, the rows and the columns of the table have similar role and hence we can use the same statistics to identify the rows and the columns important for a given dimension. To examine the importance of an element we look at its *contributions* which is the ratio of its squared factor scores to the eigenvalue of this factor. The contribution of columns to the first three components is shown in figure 3.3.

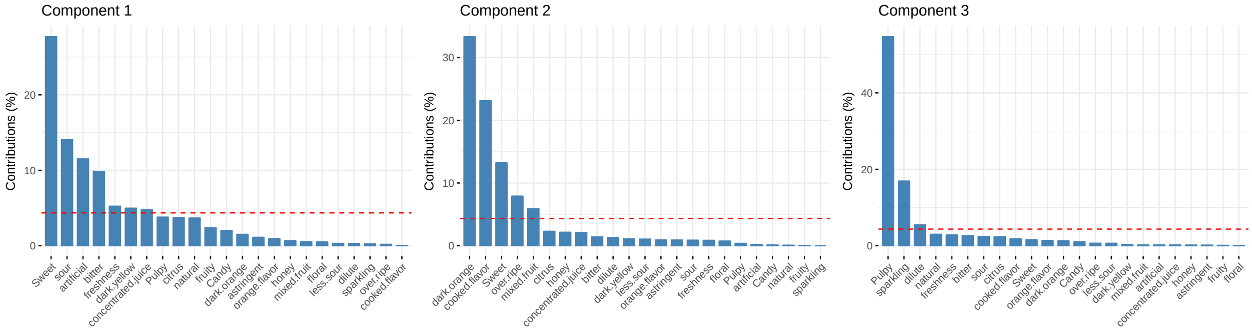


Figure 3.3: Contribution of Columns to Components 1,2 and 3

From the above figure we can see that component 1 explain the variables about sweet, sour, artifical and bitter, component 2 explain the variables on dark.orange, cooked flavor and sweetness, over.ripe and mixed.fruit and component 3 explain the variables on pulpy, sparkling and dilute.

3.3 Interpreting Factor Map

In a CA map when two row (respectively column) points are close to each other, this means that these points have similar profiles, and when two points have the same profile, they will be located exactly at the same place. In this plot the proximity between a row point and a column point cannot be interpreted. This map is called a symmetric plot. The symmetric plot of column elements is shown in figure 3.4

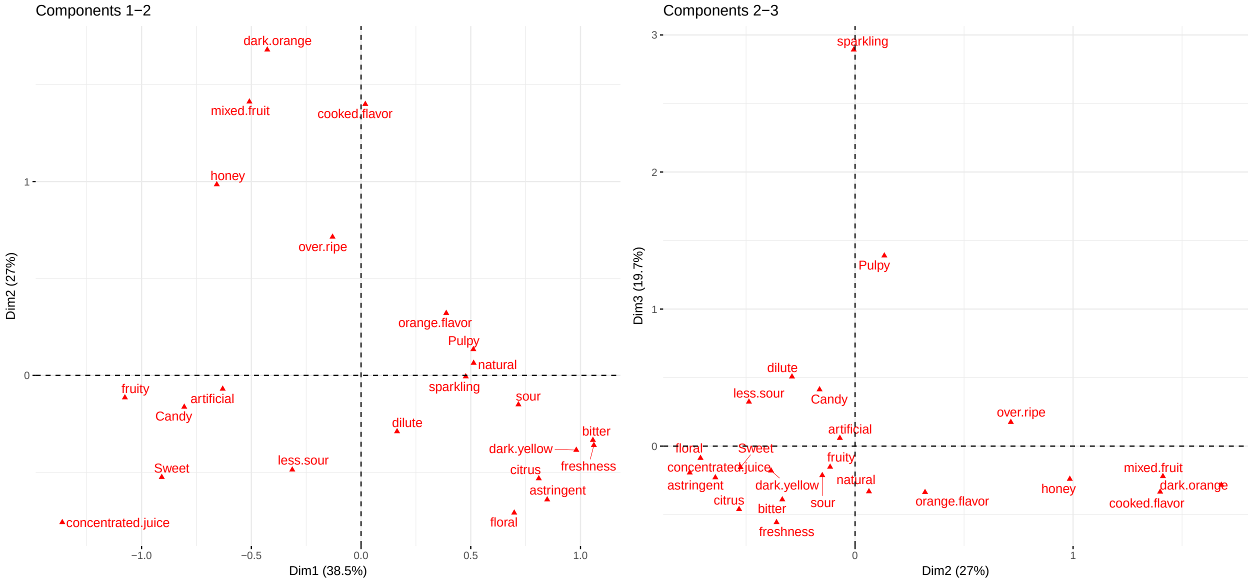


Figure 3.4: CA Symmetric Plot

By observing the above symmetric plot we can make the following conclusions:

- Component 1 contrasts the sweetness factor (sweet and candy vs sour and bitter), make (artificial vs natural) and smell (fruity vs floral) of the juice.

- Component 2 contrasts the colour (dark orange vs dark yellow), flavor (orange vs cooked)and taste (citrus vs honey) of the juice.
- Component 3 accounts for the concentration (dilute vs concentrated) and sparkling (sparkling vs natural) of the juice.

3.3.1 Interpreting Row and Column Proximity

The proximity of a row point and column point cannot be interpreted from a standard symmetric plot. To make it interpretable we normalize the column factor scores by the following formulae:

$$\hat{G} = D_c^{-1}Q$$

In tha asymmetric plot obtained with $\mathbf{F}$ and $\hat{\mathbf{G}}$, the distance from a row point to a column point reflects their association. An asymmetric biplot of our dataset is shown in figure 3.5

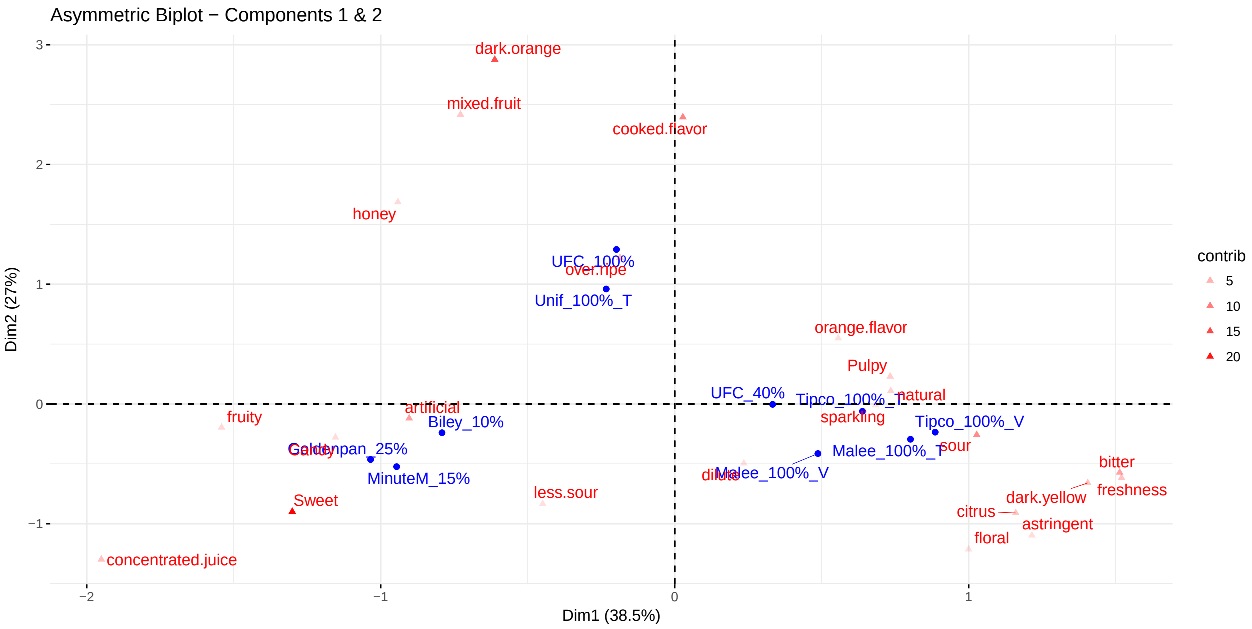


Figure 3.5: CA Asymmetric Biplot

From the above biplot the following conclusions can be made:

- Minute Made, Goldenpan and Biley are sweet and artifical and have fruity smell whereas Malee, Tipco and UFC are sour and natural and have floral smell.
- Malee, Tipco and some variations of UFC have dark yellow colour whereas Unif and some variations of UCF have dark orange colour.

3.4 Bootstrap Ratios

From figure 3.6 we can see that none of the column elements are significant for the first two components and only one element is significant for component 3. So the sample we have for our analysis is not a good estimate of the global population.

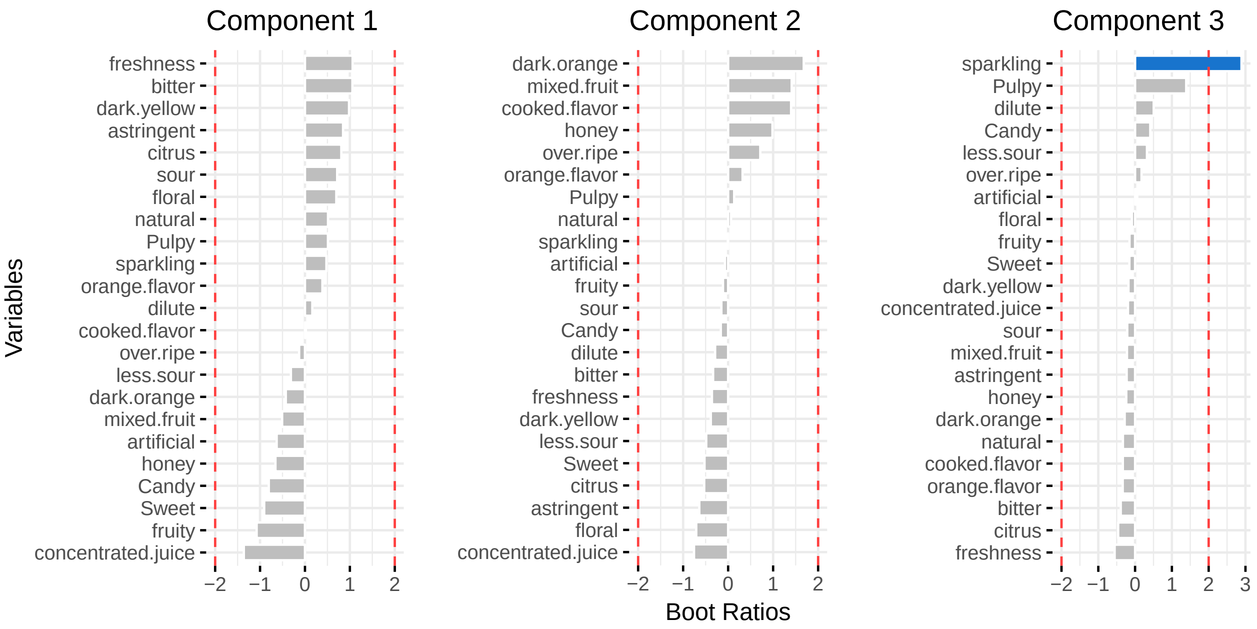


Figure 3.6: Bootstrap Ratios for Component 1,2 and 3

Chapter 4

Multiple Correspondence Analysis

Multiple Correspondence Analysis is an extention of correspondence analysis which allows one to analyze the pattern of relationships of several categorical dependent variables. It can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative(Herve Abdi, 2007b).

MCA can also accomodate quantitative variables by recording them as bins. For this analysis we will be using the `World Health Risk` data. Hence this is a quantitative data we convert it into a qualitaitve table by binning the variables based on the constrained explained in section 4.1

4.1 Binning

Histogram gives us the idea of how the data is distributed. It is also important to visualize the data as histogram to find the optimal number of bins we use to make the table a categorical one. The histogram and how binning is done to make our data a qualitative table is illustrated in figure 4.1. The binning was done based on the standard charts for Total Cholesterol, Systolic BP and BMI scores.

Total Cholesterol:

- <5.2: ideal
- 5.2 - 6.2: bordeline
- >6.2: high

Blood Pressure:

- <100 - low BP
- 100 - 130: ideal
- 130 - 150: pre-BP
- >150: high BP

Body Mass Index

- <19: underweight
- 19 - 25: healthy
- 25 - 30: overweight
- >40: obese

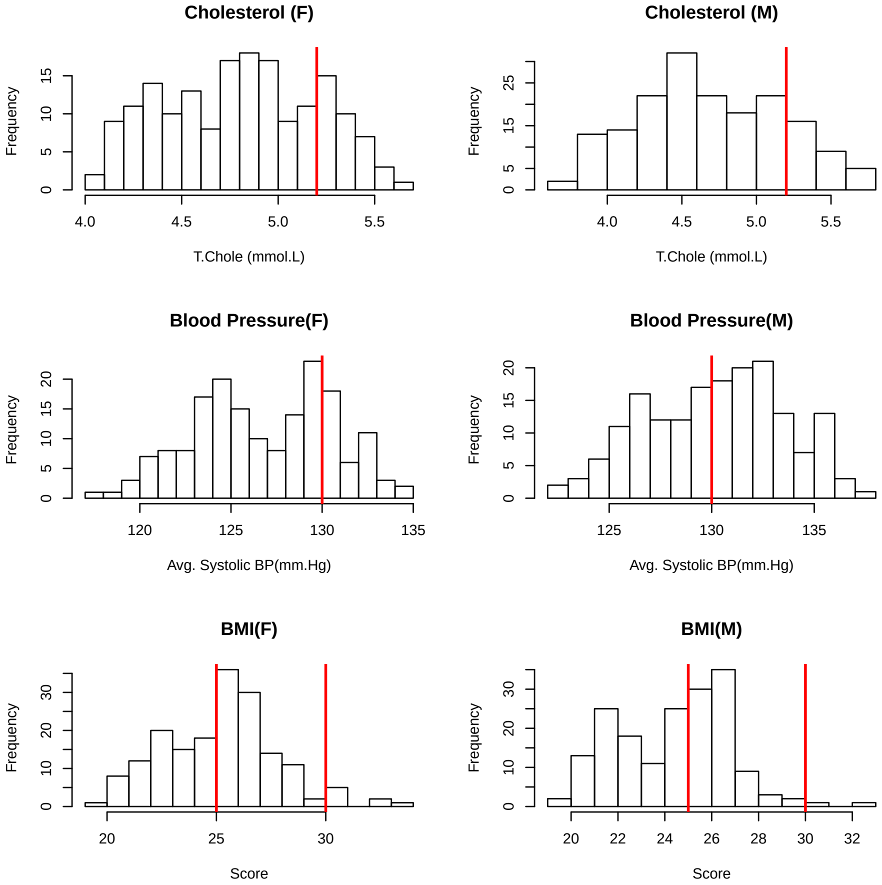


Figure 4.1: Histogram and Bins

4.2 Eigenvalues

MCA codes data by creating several binary columns for each variable with the constraint that one and only one of the columns get the value 1. As a consequence, the inertia (i.e. variance) of the solution space is artifically inflated and therefore the percentage of inertia explained by the first dimension is severly underestimated. This is corrected by the following formulae:

$$\hat{\lambda}_{\mathbf{l}} = ([\frac{\mathbf{K}}{\mathbf{K-1}})(\lambda_{\mathbf{l}} - \frac{\mathbf{1}}{\mathbf{K}})]^{\mathbf{2}}$$

if $\lambda_{\mathbf{l}} > \frac{\mathbf{1}}{\mathbf{K}}$ and

$$\hat{\lambda}_{\mathbf{l}} = \mathbf{0}$$

if $\lambda_{\mathbf{l}} \leq \frac{\mathbf{1}}{\mathbf{K}}$, where **K** is the number of variables.

The scree plot with the corrected eigen values is shown in figure 4.2

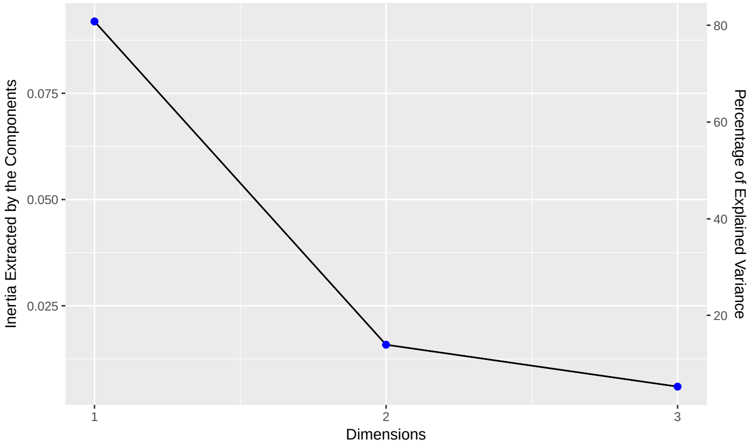


Figure 4.2: Scree Plot for Inference MCA

From the above plot we can see that there are 3 components and most of the variance is explained by the first component itself(almost 90%).

4.3 Interpreting Factor Map

As in CA, in MCA also proximities are meaningful only between points from the same set(i.e. rows with rows, columns with columns). When two row points are close to each other they tend to reflect the same level of nominal variables. The factor map of observations is shown in figure 4.3

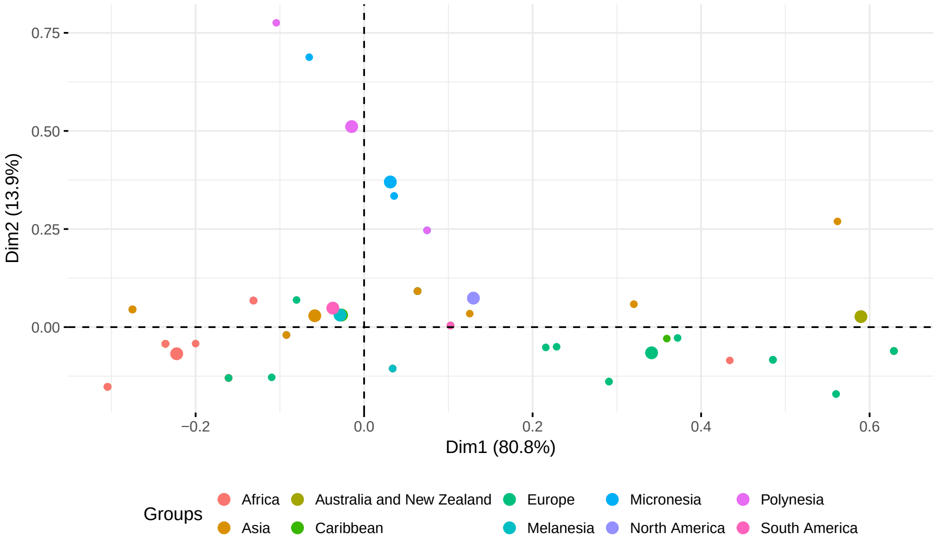


Figure 4.3: Factor Map of Observations

From the above map we can say that component 1 is seperating African countiries from European countires and component 2 is pulling Micronesian and Polynesian countires. To find the variables that account for these differences, we examine the factor map for variables.

For the proximity between variables we need to distinguish two cases. First, the proximity between levels of different nominal variables mean that these levels tend to appear together in the observations. Second, the proximity between levels mean that the groups of observations associated with these two levels are themselves similar. The factor map of variables is shown in figure 4.4

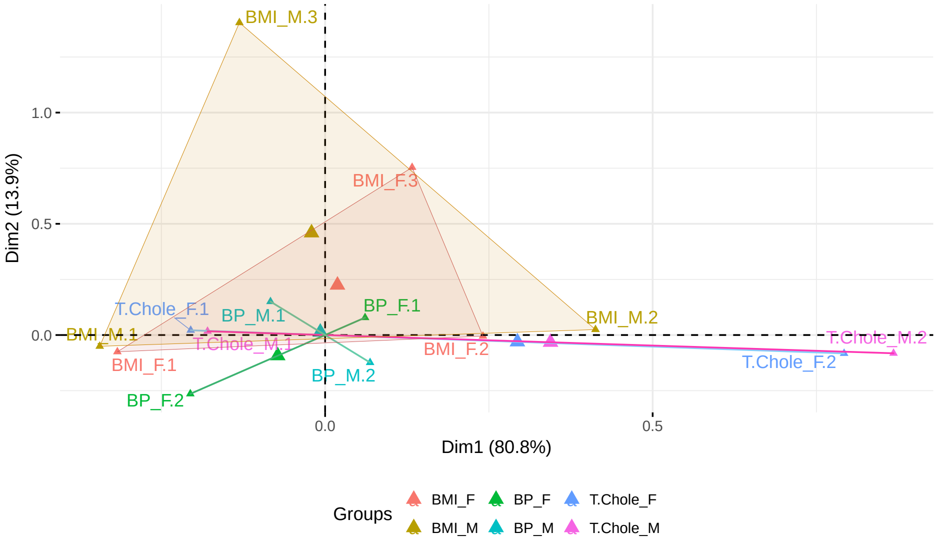


Figure 4.4: Factor Map of Variables

From the above map we can see that Total Cholesterol and BMI for both male and female have similar trends whereas BP has opposing trends for male and female. The huge variance for BMI along dimension 2 is due to the observations driving these factos. We can also infer that Total Cholesterol has more variance than BMI along dimension 1.

4.3.1 Interpreting Row and Column Proximity

Using the theory explained in section 3.3.1, an asymmetric biplot is generated as shown in figure 4.5

From figure 4.5 we can see that African and Asian countries have a healtheir population (Ideal Cholesterol levles and BMI scores) whereas American and European countries have a slightly unhealthy population (borderline cholesterol and overweight). Also males from Asian and African countries have normal BP whereas those from American and European countires have pre-BP levles. However the trend was reversed for females. The population of Polynesian, Micronesian and Melanesian countires were found to be obese.

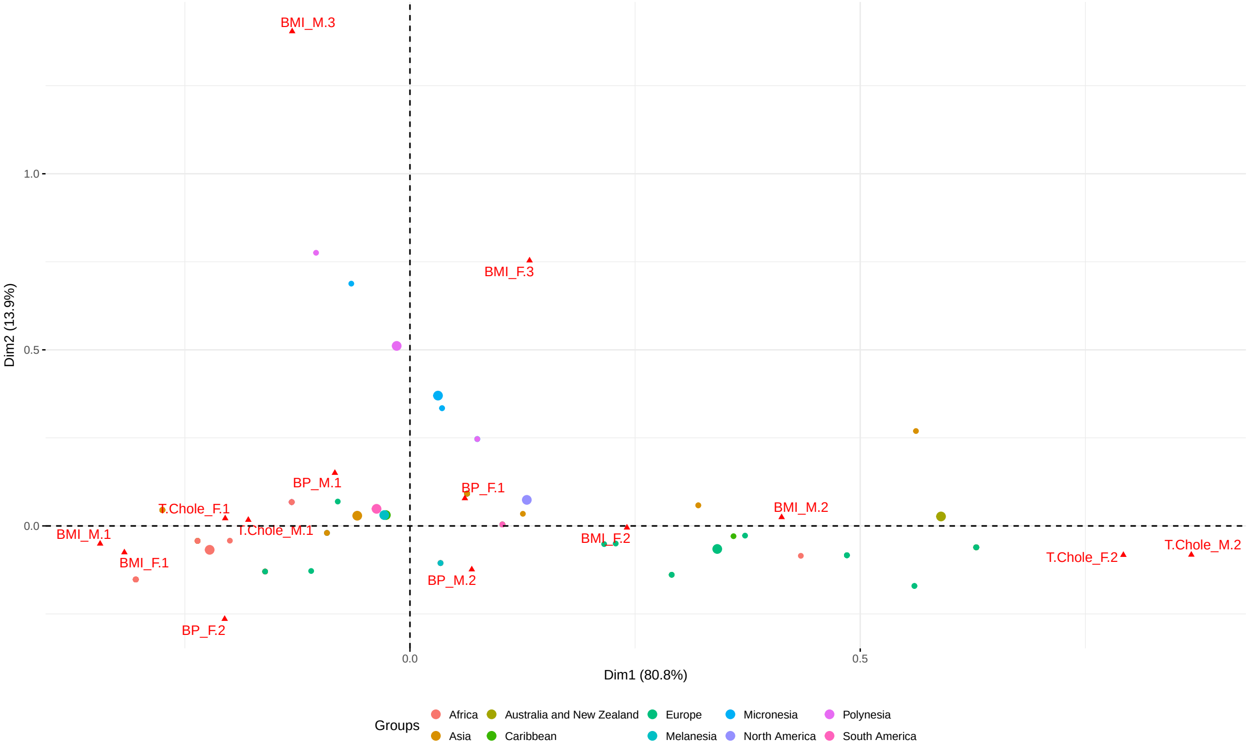


Figure 4.5: MCA Asymmetric Biplot

Part II

Two Table Analysis

Chapter 5

Partial Least Square Methods

Partial Least Square methods also sometimes called projection to latent structures relate the information present in two data tables that collect measurements on the same set of observations. PLS method proceed by deriving latent variables which are linear combinations of the variable of the data table. When the goal is to find the shared information between two tables, the approach is equivalent to correlation problem and the technique is called Partial Least Square Regression. In this case there are two sets of latent variables and these latent variables are required to have maximal covariance. The original variables are described by their saliences. By analogy with principal component analysis, the latent variables are akin to factor scores and the saliences are akin to loadings(Herve Abdi, 2013).

5.1 Computation

The main analytical tool for PLSC is the singular value decomposition of the matrix $\mathbf{R}$, where $\mathbf{R} = \mathbf{Z_Y^T Z_X}.Z_X$ and Z_Y are the rescaled versions of X and Y. The SVD of R decomposes it into three matrices:

$$\mathbf{R} = \mathbf{U \Delta V^T}$$

In PLSC vocabulary, the singular vectors are called saliences: so U is the matrix of Y-saliences and $V$is the matrix of X-saliences. The latent variables are obtained by projecting the original matrices onto their respective saliences. Specifically, we obtain the latent variables for X as:

$$\mathbf{L_X} = \mathbf{Z_X V}$$

and for Y as:

$$\mathbf{L_Y} = \mathbf{Z_Y U}$$

5.2 Interpreting PLSC

For this analysis we compare three tables from the `World Health` dataset. From the correlation plot of `World Health` dataset (Figure 1.1), we see that there is a strong correlation between `World Health Risk` and `Spendings` as well as between `World Health Risk` and `Immunization`. So we compare these two subtables in this analysis.

A heatmap of $X^T Y$ (X is the matrix of `Risk Factor` data and Y, the matrix of `Spendings` data) is shown in figure 5.1

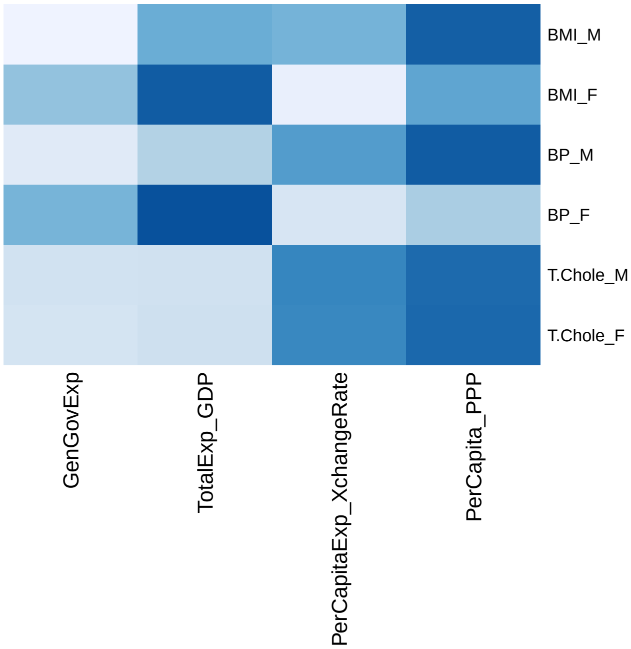


Figure 5.1: Heatmap of PLS

5.3 Eigenvalues/Variances

Scree plot of PLSC analysis between `World Health Risk` and `World Health Spending` is shown in figure 5.2.

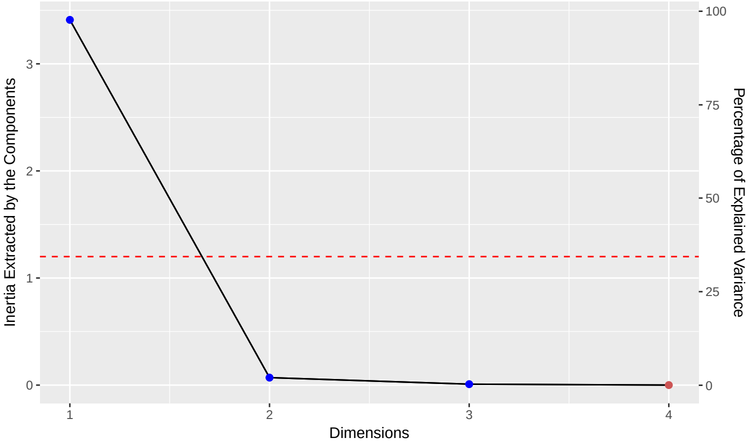


Figure 5.2: Scree Plot for PLS

The above scree plot reveals that there are four components and the first component explains more than 90% of the variance.

5.4 Interpreting Latent variables

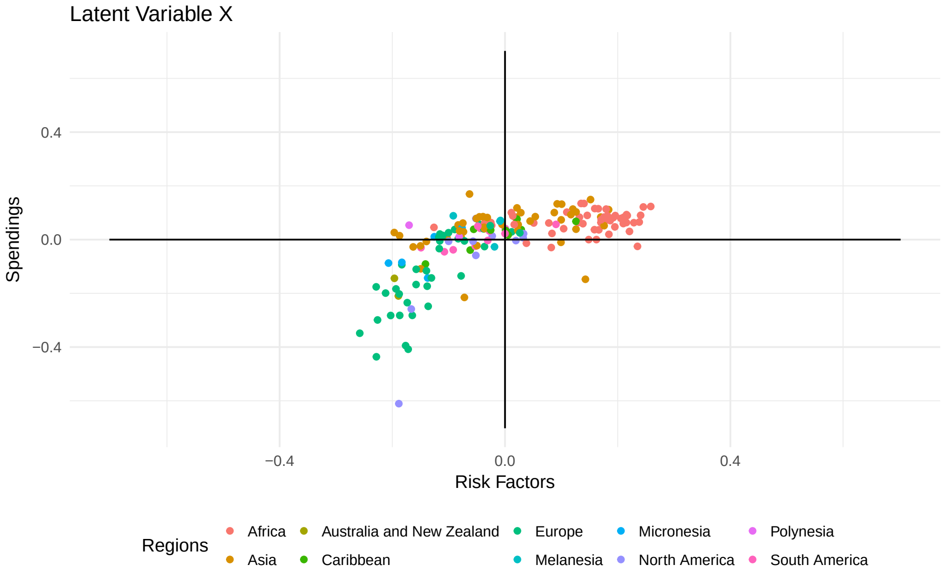


Figure 5.3: The X latent variable plot for PLSC between World Health Risk and Spendings

The above plot is a figure showing the latent variables of Risk Factors against Spendings for first component. Here we can see that PLSC was trying to maximize whatever covariance they were having. From the above plot we can conclude that there is some kind of linear trend for Asia and North America whereas variance of Europe is mostly explained by spendings and variance of Africa is explained by Risk Factors.

Now we do the analysis on the second set of tables - WorldHealth_Risk and WorldHealth_Immunization. The scree plot for the PLSC of these are shown in figure 5.4.

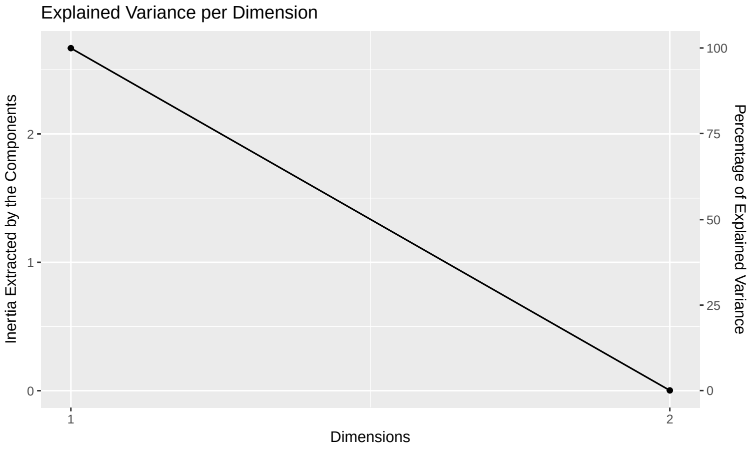


Figure 5.4: Scree Plot for PLSC WorldHealth Risk and Worldhealth Immunization

Here we can see that there's only two components and the first component alone explains all the variance.

Figure 5.5 shows the factor map of latent variables (X against Y) for the first component of Risk Factors vs Immunization anaysis.

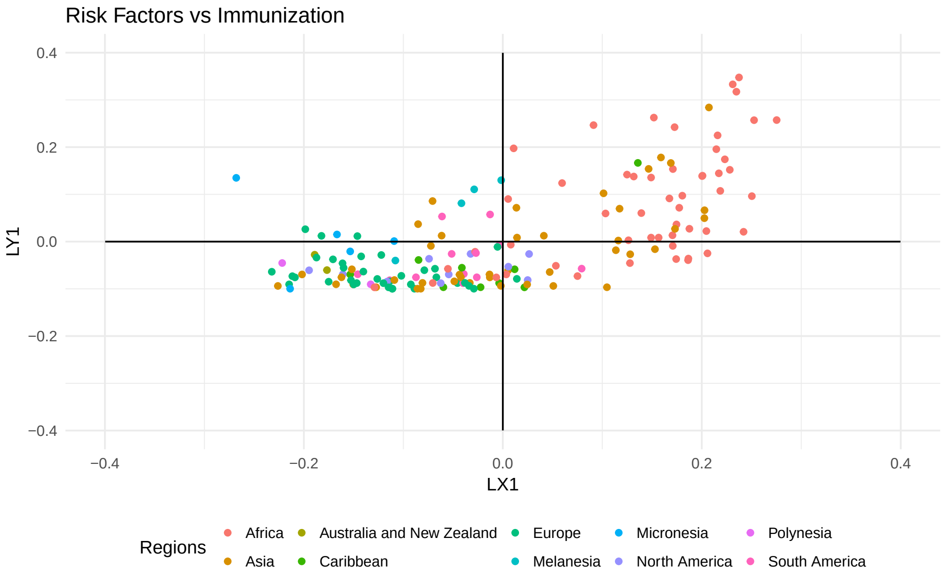


Figure 5.5: Factor Map of Observations

Here we can see that there isn't any relationship between the tables.

5.5 Saliences

Figure 5.6 shows the saliances for Risk Factors. From the figure we can see that Component one explains most of the variance in blood pressure whereas component two explains the variance of BMI and Total Cholesterol.

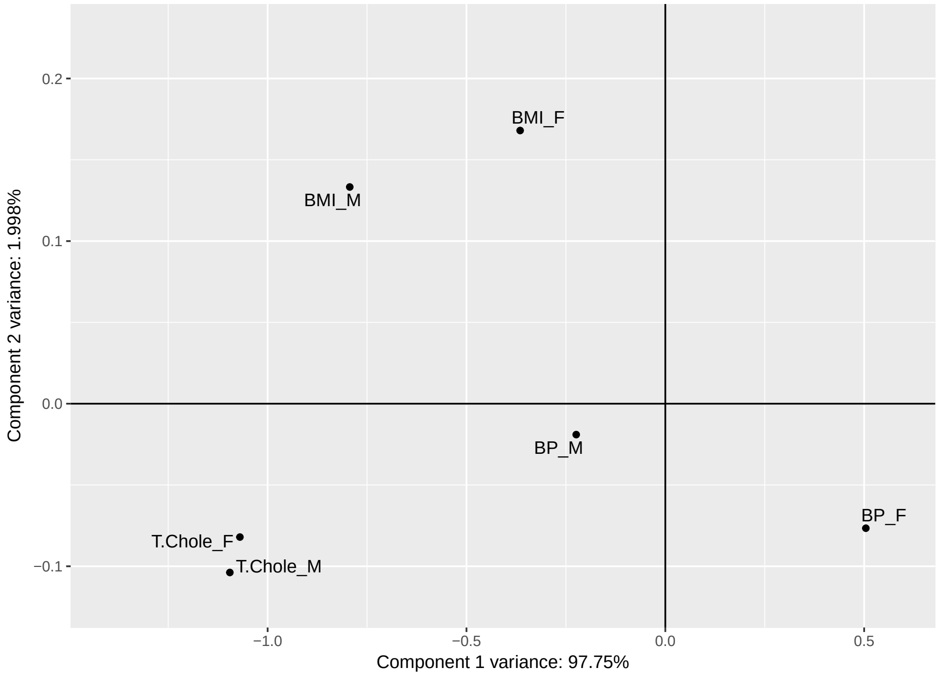


Figure 5.6: Salience for Risk Factors

Figure 5.7 shows the salience for Spendings. From the figure we can see that the general government expenditure and the total expenditure is almost orthogonal to the per capita expenditures.

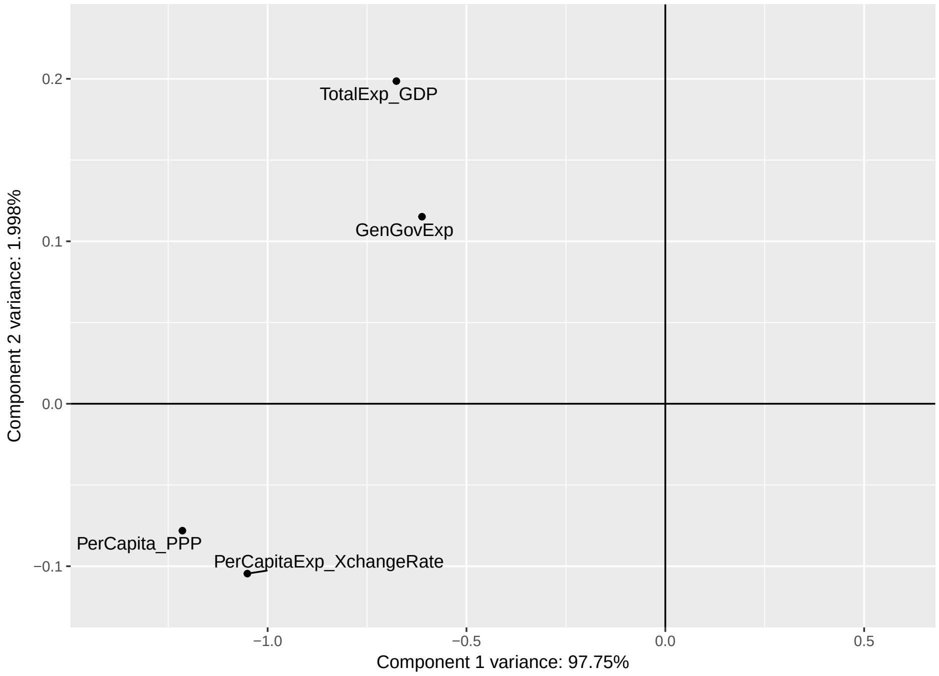


Figure 5.7: Saliences for Spending

5.6 Bootstrap Ratios

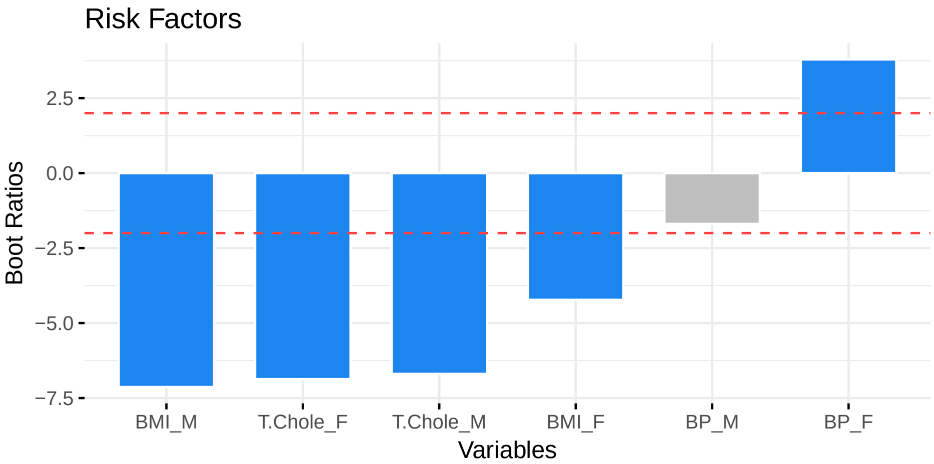


Figure 5.8: Bootstrap Ratios for Risk Factors

From the above figure we can see that all variables of the Risk Factor table are significant for component one except BP_M

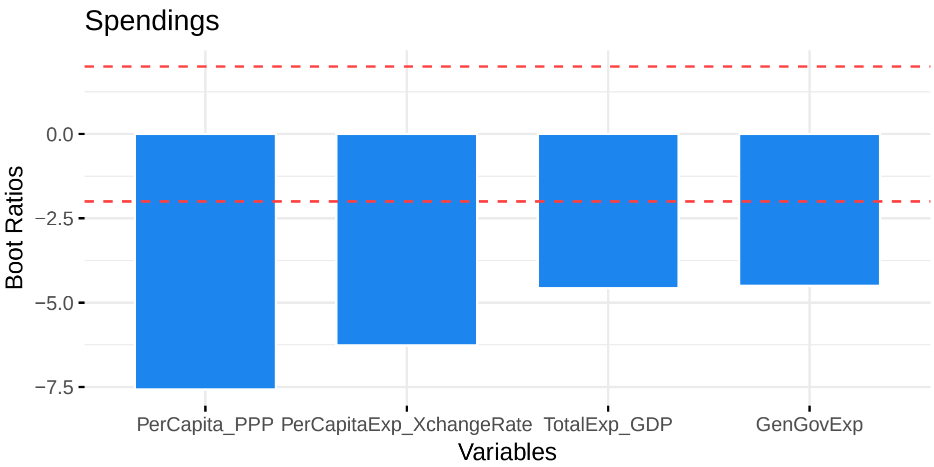


Figure 5.9: Bootstrap Ratios for Spendings

For Spendings table, all the variables were significant for component one.

Chapter 6

Barycenter Discriminant Analysis

Barycenter Discriminant Analysis is a robust version of Discriminant Analysis that is used like discriminant analysis - when multiple measurements describe a set of observations in which each observation belongs to one category from a set of priori defined categories. The goal of BADA is to combine the measurements to create new variables that best seperate the categories. These discriminant variables are also used to assign the original observations or new observations to the a-priori defined categories(Herve Abdi, 2018).

The first step of BADA is to compute the barycenter of each of the N categories describing the rows. The N by J matrix of barycenters is computed as:

$$\mathbf{R} = diag\{\mathbf{Y^T M 1}\}^{-1}\mathbf{Y^T M X}$$

where Y is the design matrix for the categories describing the rows of X

The matrix R is then analyzed using a GPCA under the constrints provided by the matrices B and W.

$$\mathbf{R} = \mathbf{P \Delta Q^T}$$

with $\mathbf{P^T B P} = \mathbf{Q^T W Q} = \mathbf{I}$. The N by L matrix of factor scores for the categories is obtained as:

$$\mathbf{F} = \mathbf{P\Delta} = \mathbf{RWQ}$$

The loadings describe the variables of the barycentric data matrix and are used to identify the variables important for the separation between the groups.

For this analysis we use the `World Health` dataset and the heatmap of designmatrix * data is shown in figure 6.1

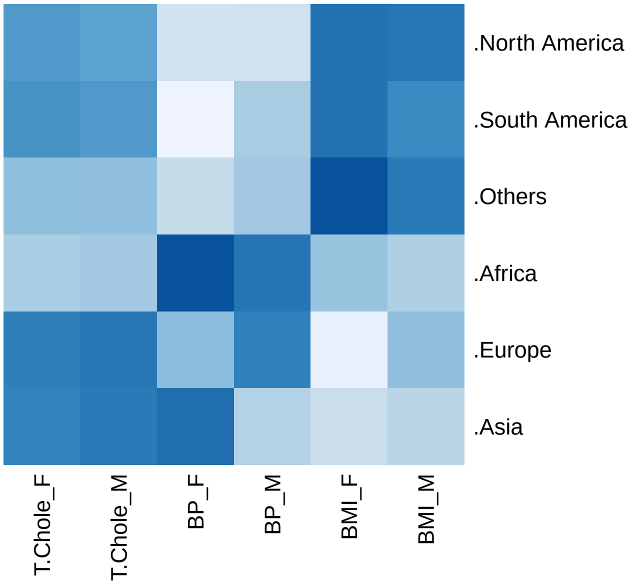


Figure 6.1: BADA Heatmap

6.1 Eigenvalues/Variance

The Scree plot for BADA is shown in figure 6.2. From the figure we can see that BADA generated five components and out of them only two of them are important. On inference analysis none of them were found to be significant.

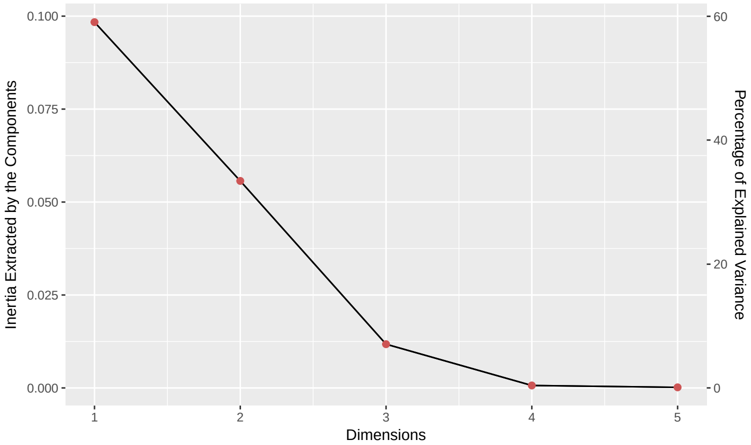


Figure 6.2: Scree Plot for BADA

6.2 Factor scores

IN BADA we can plot two factor map for our observations - one for individuals and other for groups (based on their mean). The factor map for groups is shown in figure 6.3

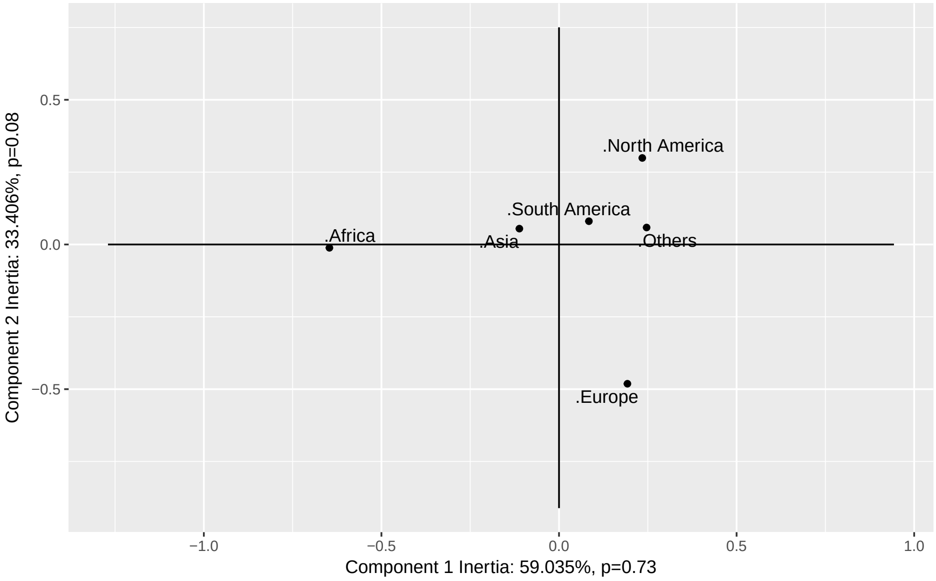


Figure 6.3: BADA Factor Scores (Group)

From tha above plot we can see that Africa and Europe is very much seperated from all other groups. To find the variables that account for these differences, we examine the loadings of the variables(figure 6.5).

Factor map of individuals with tolerance intervals is shown in figure 6.4.

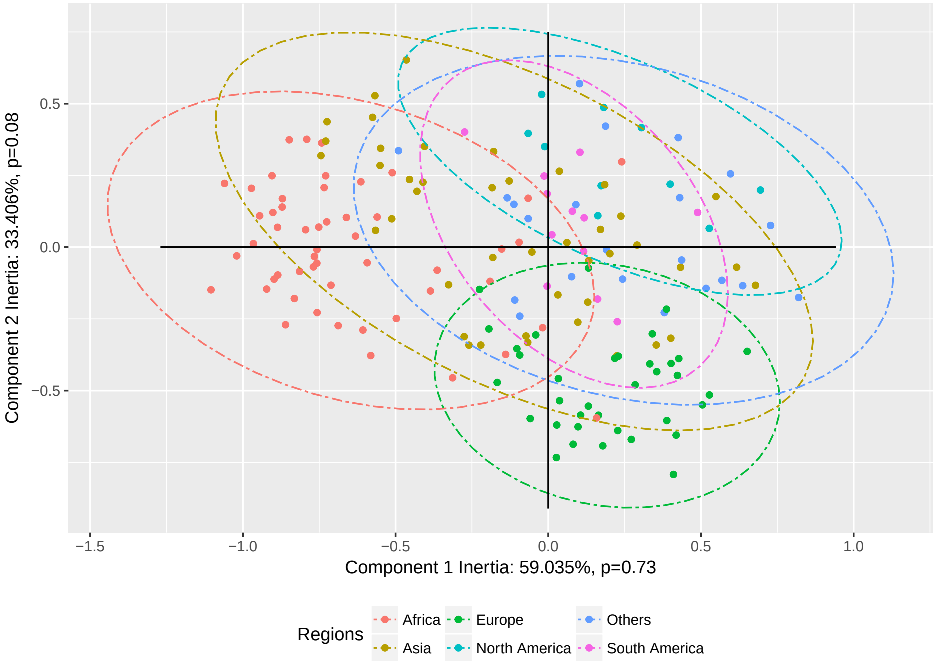


Figure 6.4: BADA Factor Scores (Individuals) with tolerance intervals

6.3 Loadings

Figure 6.5 shows the loadings of the BADA on World health risk. We can see that the loadings plot of BADA looks very much similar to that of the PCA Loadings plot.

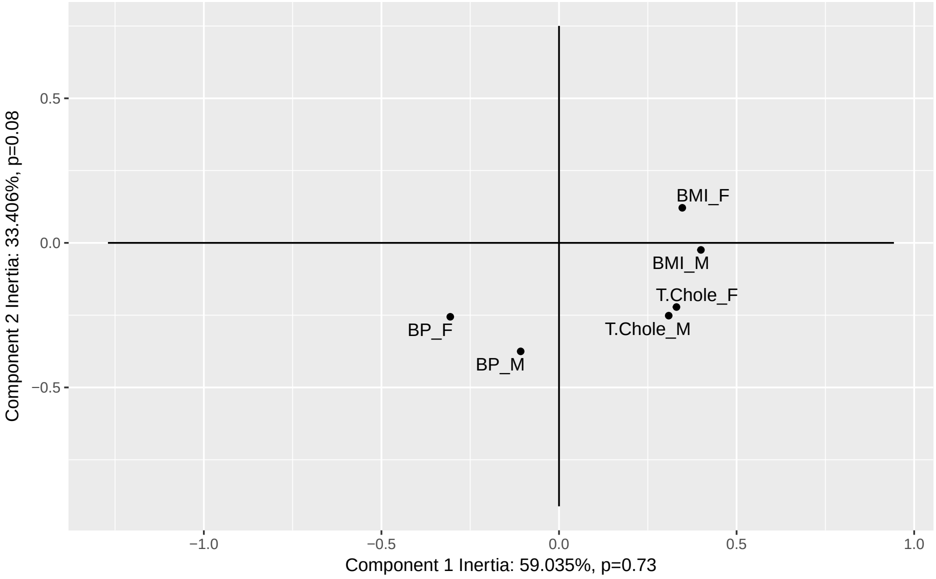
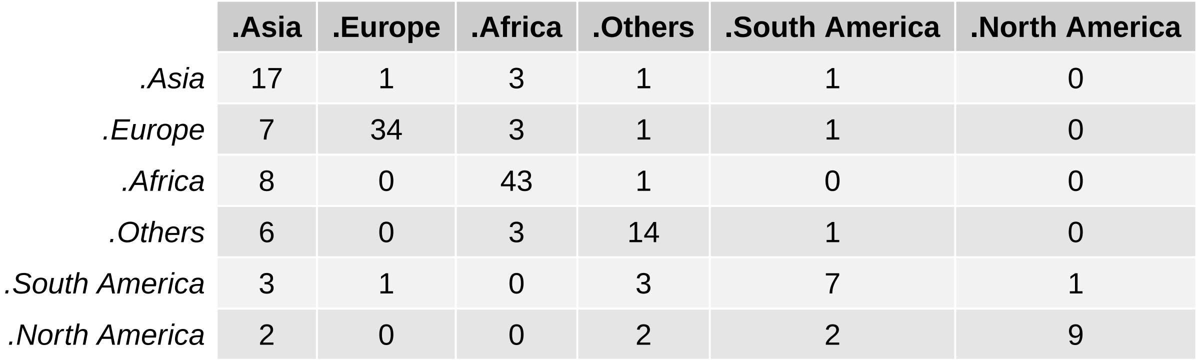


Figure 6.5: BADA Loadings

6.4 Confusion Matrix

A confusion matrix helps us to understand the quality of the model in making predictions. The confusion matrix (Fixed Model) for BADA is shown in figure 6.6.



	.Asia	.Europe	.Africa	.Others	.South America	.North America
.Asia	17	1	3	1	1	0
.Europe	7	34	3	1	1	0
.Africa	8	0	43	1	0	0
.Others	6	0	3	14	1	0
.South America	3	1	0	3	7	1
.North America	2	0	0	2	2	9

Figure 6.6: Confusion Matrix (Fixed Model)

6.5 Reliability and Stability of the Analysis

Figure 6.7 shows the factor map of BADA for groups with 95% bootstrap confidence intervals.

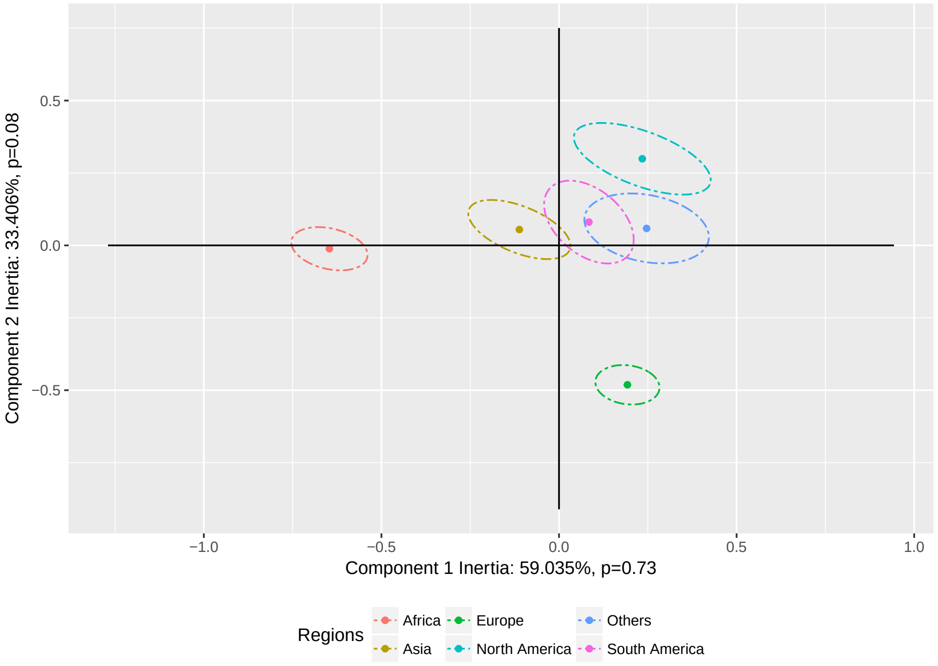


Figure 6.7: BADA Factor Scores (Group) with confidence intervals

6.5.1 Bootstrap Ratios

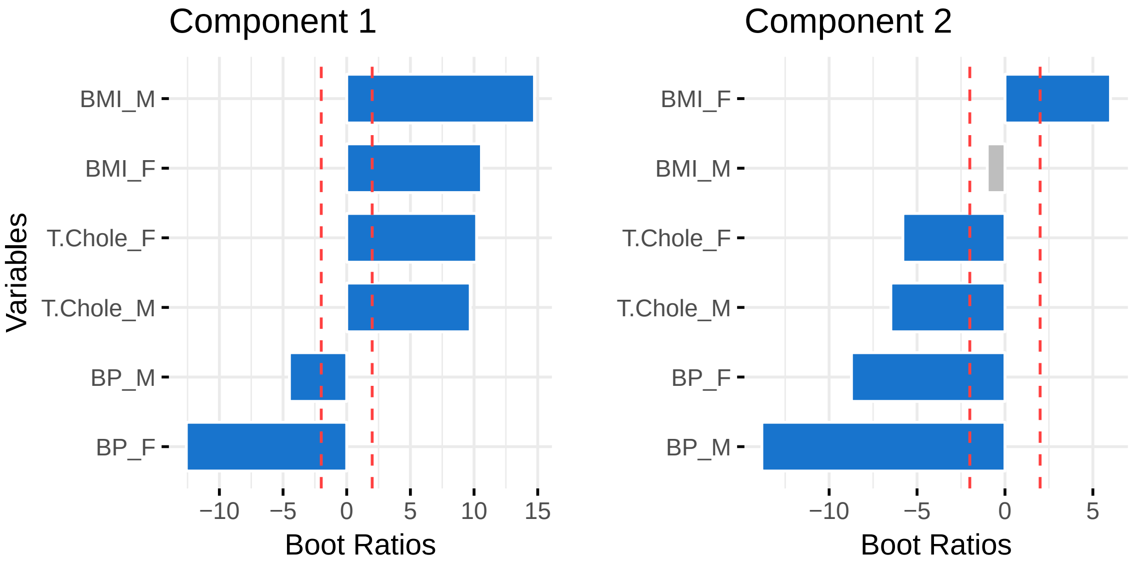
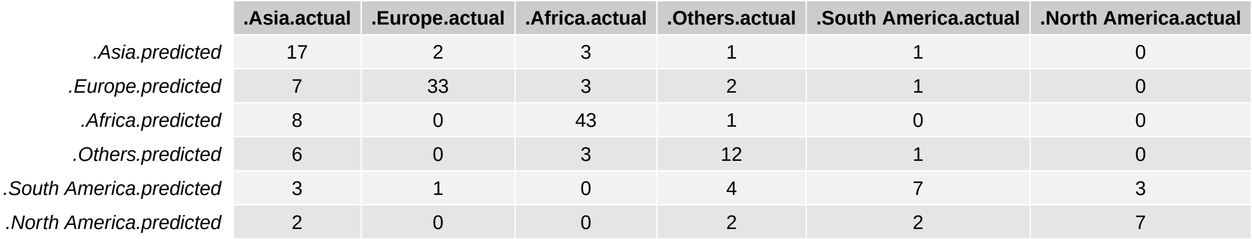


Figure 6.8: BADA Bootstrap Ratios

From the above figure we can see that all the variables were significant for component one whereas only BMI_Male was insignificant for component two.

6.5.2 Confusion Matrix (Random Model)

A jackknife procedure was used inorder to evaluate the generalization capacity of the analysis. This gave the following random effect confusion matrix:



	.Asia.actual	.Europe.actual	.Africa.actual	.Others.actual	.South America.actual	.North America.actual
.Asia.predicted	17	2	3	1	1	0
.Europe.predicted	7	33	3	2	1	0
.Africa.predicted	8	0	43	1	0	0
.Others.predicted	6	0	3	12	1	0
.South America.predicted	3	1	0	4	7	3
.North America.predicted	2	0	0	2	2	7

Figure 6.9: Confusion Matrix (Random Model)

Chapter 7

Discriminant Correspondence Analysis

Discriminant Correspondence Analysis is an extention of discriminant analysis and correspondence analysis. Like discriminant analysis, the goal of discriminant correspondnce analysis is to categorize observations in predefined groups, and like correspondence analysis, it is used with nominal variables(Abdi, 2007).

The main idea behind discriminant correspondence analysis is to represent each group by the sum of its observations and to perform a simple CA on the groups by variables matirx. The original observations are then projected as supplementary elements and each observation is assigned to the closest group.

7.1 Computation

DICA is obtained by the following singular value decomposition:

$$\mathbf{D_r}^{-\frac{1}{2}}(-\mathbf{rc^T})\mathbf{D_c}^{-\frac{1}{2}} = \mathbf{P\Delta Q^T}$$

where $\mathbf{D_c} = \mathbf{diag}\{\mathbf{c}\}$ and $\mathbf{D_r} = \mathbf{diag}\{\mathbf{r}\}$, $\mathbf{r}$ and $\mathbf{c}$ are vectors of row totals and column totals.

The row and column factor scores are obtained as

$$\mathbf{F} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{P\Delta}$$

and

$$\mathbf{G} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{Q\Delta}$$

For this analysis we use the `World Health` dataset and the heatmap is shown in figure 7.1

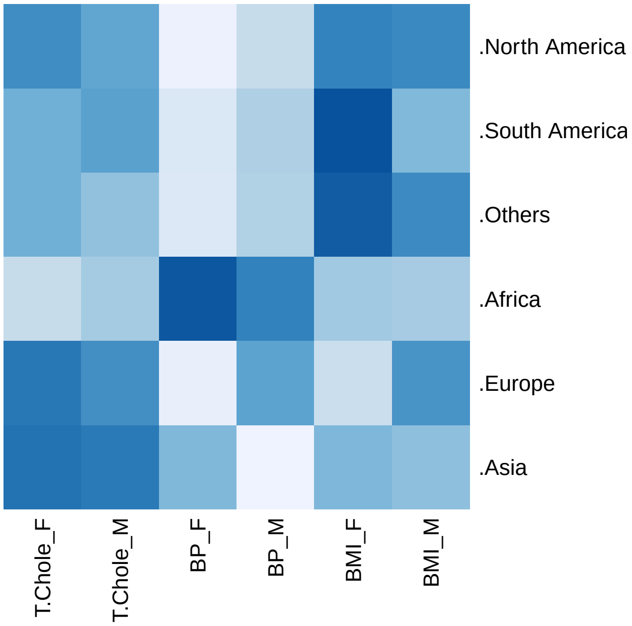


Figure 7.1: DICA Heatmap

7.2 Eigenvalues/ Variance

Figure 7.2 shows the Scree plot for DICA and we can see that DICA generated five components with one important component (explaining more than 75%).

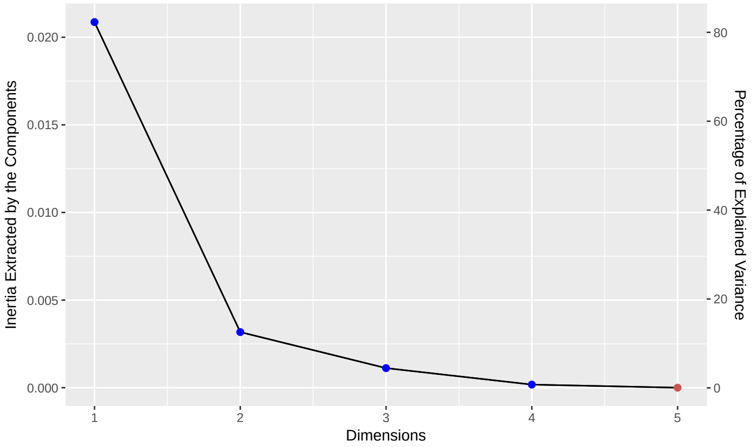


Figure 7.2: Scree Plot for BADA

7.3 Interpreting the Observations and Groups (Rows)

Interpreting the meaning of the factors is done in two stages: first by looking at the observations/groups and then by looking at the variables. The realtive position of the observations and groups is shown in figure 7.4 and figure 7.3.

As in MCA, the proximity between group points in the DICA maps represent their similariy, and the proximity between variable points represent their association. From figure 7.3 we can see that component 1 is contrasting Africa with Europe. To find the variables that account for these differences, we examine the loadings of the variables (figure **??**)

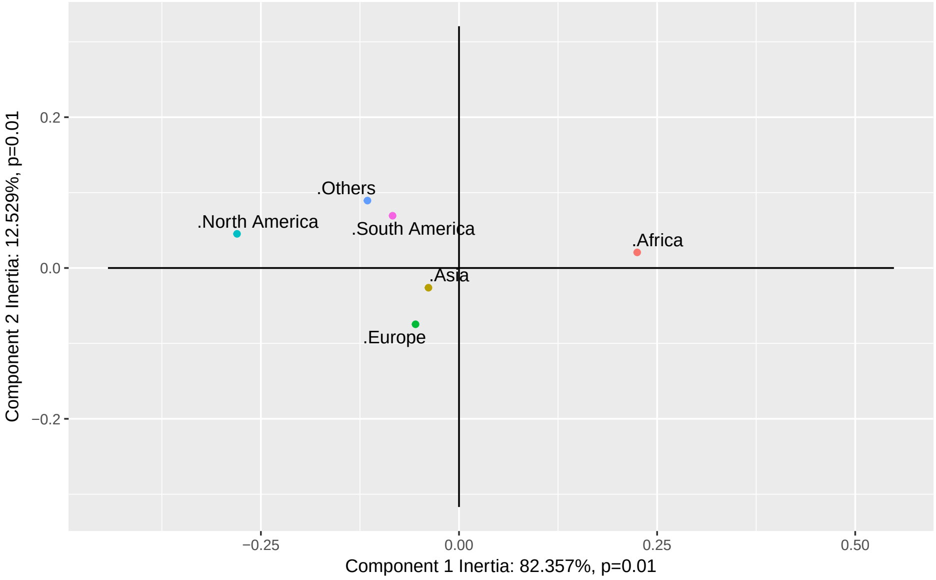


Figure 7.3: DICA Factor Map of Group Means

Figure 7.4 shows the factor map of the individuals (with convex hulls).

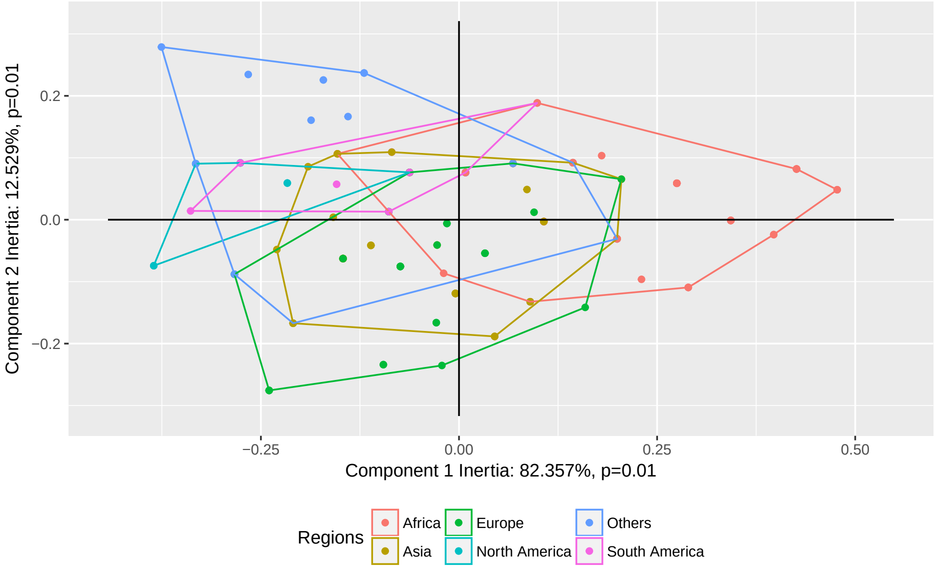
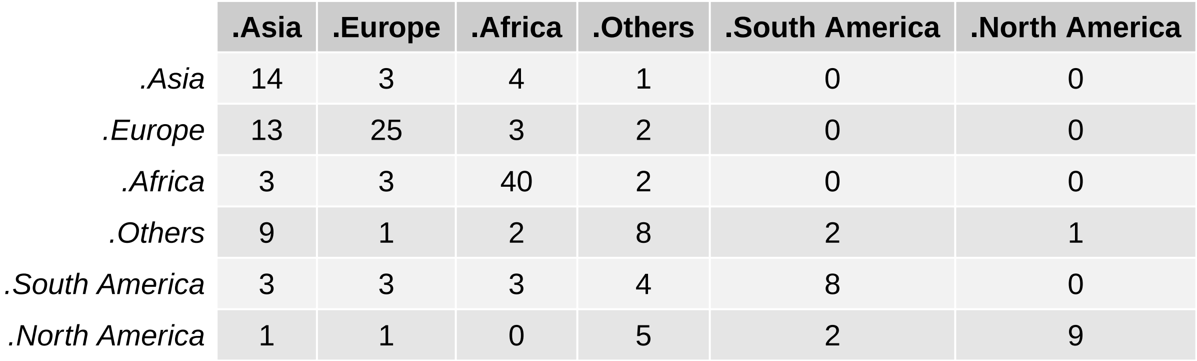


Figure 7.4: DICA Factor Map of Individuals

7.4 Confusion Matrix (Fixed Model)



	.Asia	.Europe	.Africa	.Others	.South America	.North America
.Asia	14	3	4	1	0	0
.Europe	13	25	3	2	0	0
.Africa	3	3	40	2	0	0
.Others	9	1	2	8	2	1
.South America	3	3	3	4	8	0
.North America	1	1	0	5	2	9

Figure 7.5: DICA Confusion Matrix (Fixed Model)

7.5 Inferences

Figure 7.6 shows the factor map of groups with 95% confidence interval

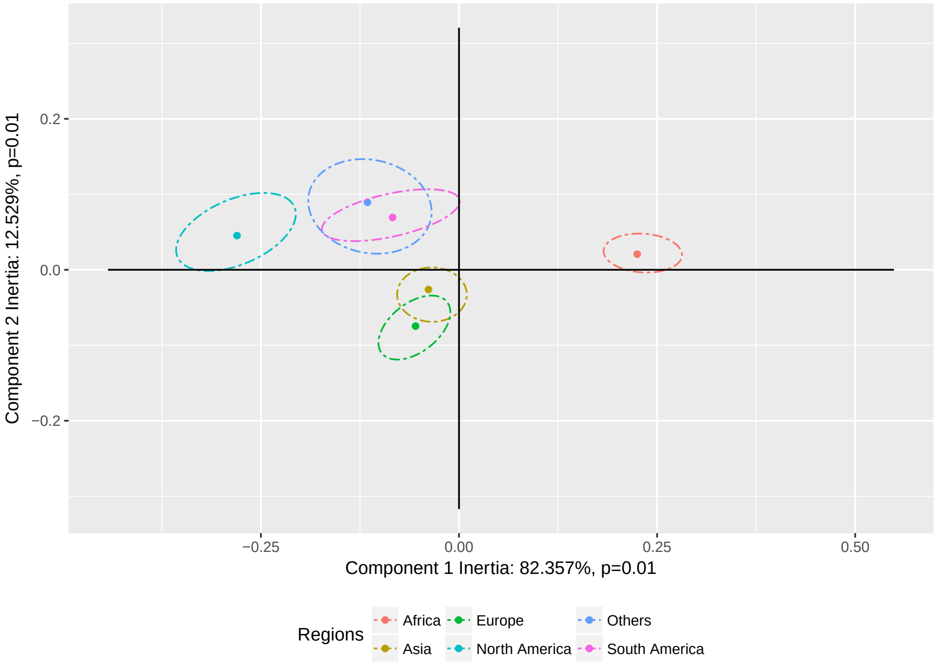


Figure 7.6: DICA Factor Scores (Group) with confidence intervals

7.5.1 Bootstrap Ratios

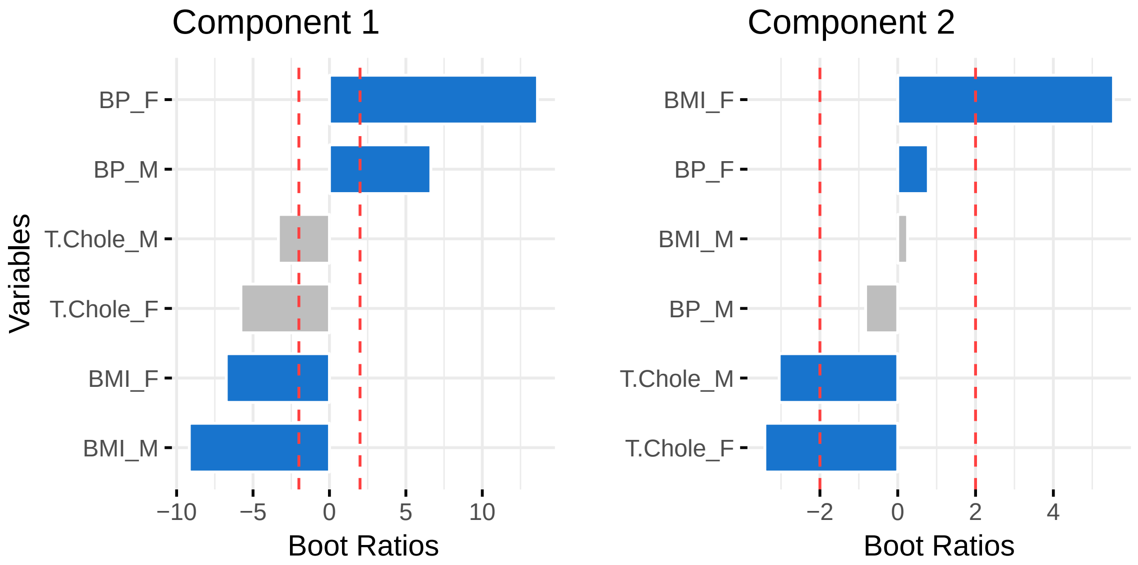
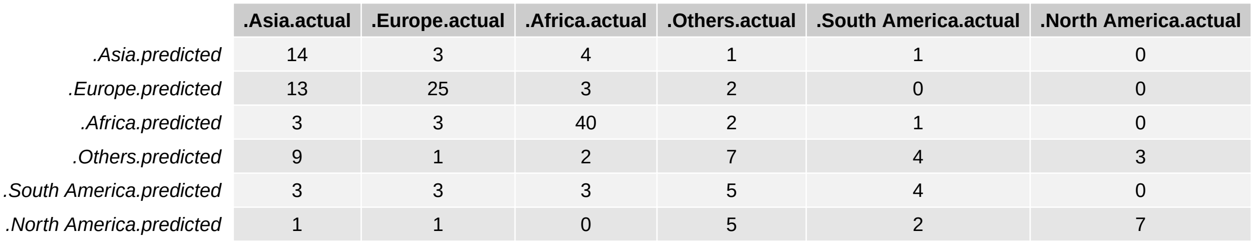


Figure 7.7: DICA Bootstrap Ratios

7.5.2 Confusion Matrix



	.Asia.actual	.Europe.actual	.Africa.actual	.Others.actual	.South America.actual	.North America.actual
.Asia.predicted	14	3	4	1	1	0
.Europe.predicted	13	25	3	2	0	0
.Africa.predicted	3	3	40	2	1	0
.Others.predicted	9	1	2	7	4	3
.South America.predicted	3	3	3	5	4	0
.North America.predicted	1	1	0	5	2	7

Figure 7.8: Confusion Matrix (Random Model)

Part III

Multi Table Analysis

Chapter 8

DiSTATIS

DISTATIS is a generalization of classical multidimensional scaling. Its goal is to analyze several distance matrices computed on the same set of objects. DISTATIS first evaluates the similarity between distance matrices. From this analysis, a compromise matrix is computed which represents the best aggrigate of the original matrices. The original distance matrices are then projected onto the compromise(Herve Abdi, 2007a).

To illustrate DISTATIS we will use the `Musical Sorting` Data. Three different "composers" are compared, each of them computing a distance matrix between the subjects. Heatmap of the distance cube for the given dataset is shown in figure 8.1

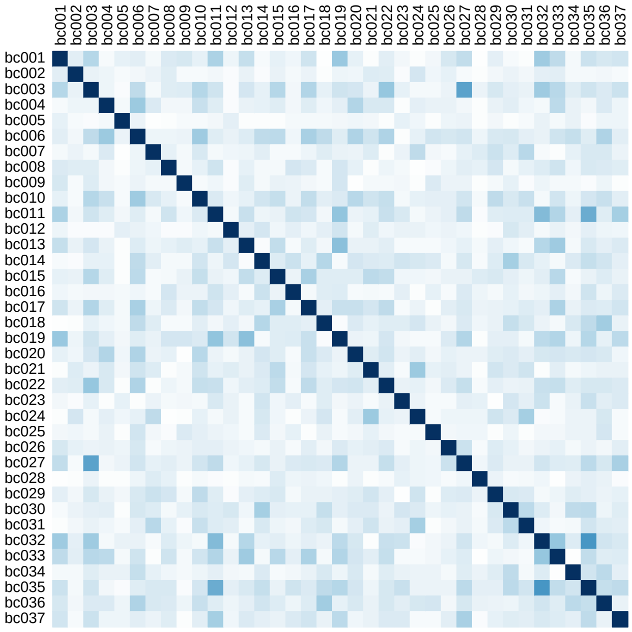


Figure 8.1: Heatmap of S Plus Matrix

8.1 Computation

The given dataset is first convereted into an indicator matrix L and this indicator matrix is transformed into a co-occurance matrix as:

$$\mathbf{R_{[t]}} = \mathbf{L_{[t]}L^T_{[t]}}$$

The co-occurance matrix is transformed then into a distance matrix by sustracting the co-occurance matrix from a conformable matrix filled with 1's.

$$\mathbf{D_{[t]}} = \mathbf{1} - \mathbf{R_{[t]}}$$

Distance matrices cannot be analyzed directly and need to be transformed. This step corresponds to MDS and transforms a distance matrix into a corss-product matrix.

The corss product matrix denoted by $\hat{\mathbf{S}}$ is obtained as

$$\hat{\mathbf{S}} = -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{\Xi D \Xi^T}$$

where $\mathbf{\Xi} = \mathbf{1} - \mathbf{1m^T}$, $\mathbf{m}$ is the vector of mass.

The compromise matrix is a cross product matrix that gives the best compromise of the studies. It is obtained as a weighted average of the study cross-product matrices. The weights are chosen so that studies agreeing the most with other studies will have the larger weights.

8.2 Comparing the studies

To analyze the similarity structure of the studies we start by creating a between study cosine matrix $\mathbf{C}$. This cosine, aslo known as the $\mathbf{R_V}$ coefficient is defined as

$$\mathbf{R_V} = \frac{trace\{\mathbf{S^T_{[t]}S_{[t']}}\}}{\sqrt{trace\{\mathbf{S^T_{[t]}S_{[t]}}\} \times trace\{\mathbf{S^T_{[t']}S_{[t']}}\}}}$$

8.3 PCA of the cosine matrix

The cosine matrix has the following eigendecomposition:

$$\mathbf{C} = \mathbf{P\Theta P^T}$$

with $\mathbf{P^TP} = \mathbf{I}$, where $\mathbf{P}$ is the matrix of eigenvectors and $\mathbf{\Theta}$ is the diagonal matrix of the eigenvalues of $\mathbf{C}$. The Scree Plot is shown in figure 8.2.

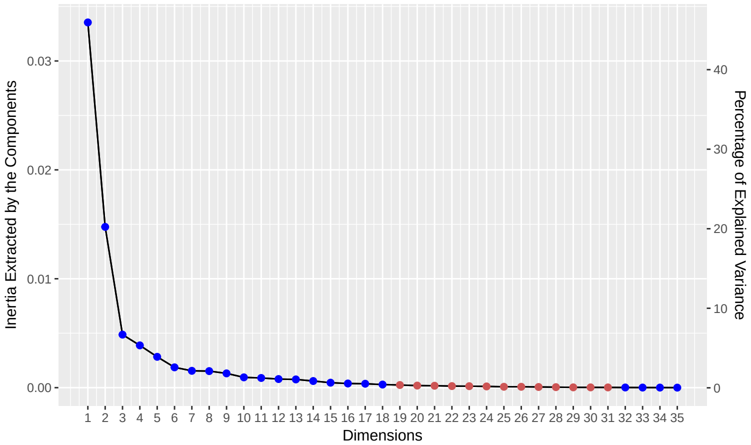


Figure 8.2: Scree Plot for DISTATIS

An element of a given eigenvector represents the projection of one study on this eigenvector. Thus the T studies can be represented as points in the eigenspace and their similarities analyzed visually. Figure 8.3 displays the projection of subjects onto the first and second component.

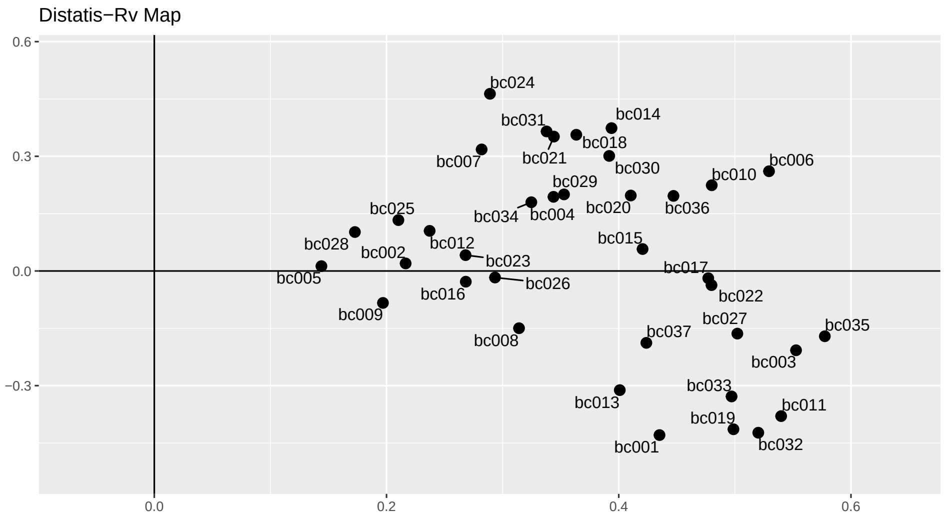


Figure 8.3: Plot of the between-composers space

8.4 Computing the compromise

As for STATIS the weights are obtained by dividing each element of $\mathbf{p_1}$ by their sum. The vector containing these weights is denoted α. With α_t denoting the weight for the t-th study, the compromise matrix, denoted

$$\mathbf{S}_{[+]} = \sum_{t}^{T} \alpha_t \mathbf{S}_{[t]}$$

The eigendecomposition of the compromise is:

$$\mathbf{S}_{[+]} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathbf{T}}$$

and the compromise factor scores for the observations are computed as:

$$\mathbf{F} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}$$

The compromise plot of the music is shown in figure 8.4. The music are color coded based on their composer.

```
## [1] "Making constraints"
```

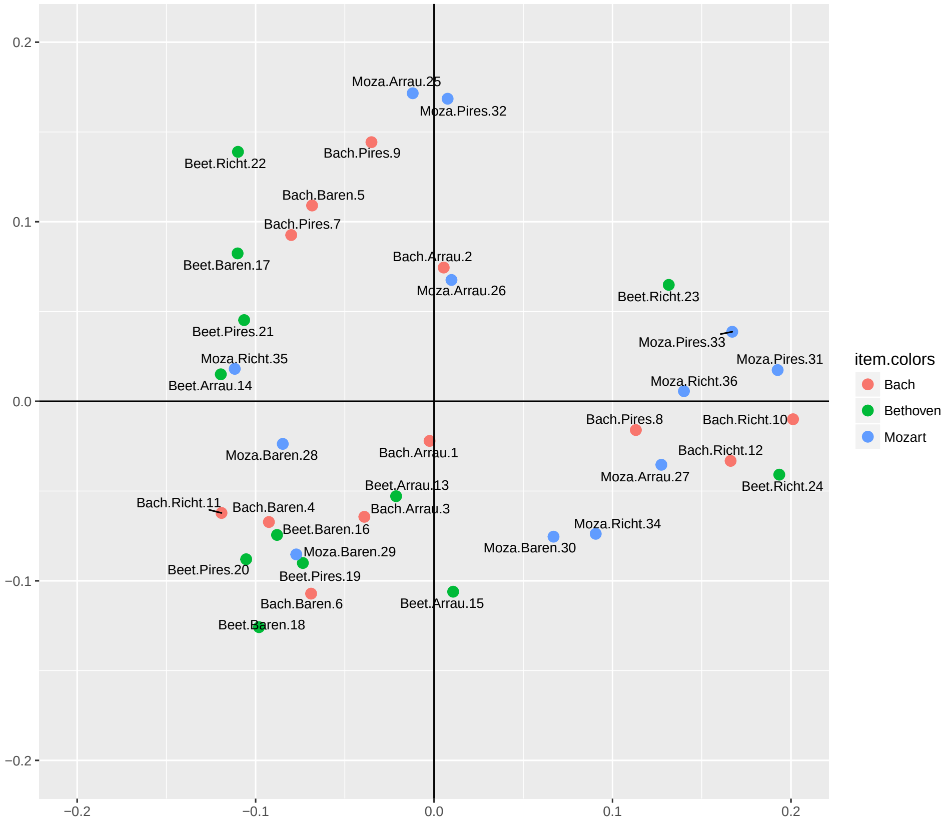


Figure 8.4: Analysis of the compromise: Plot of the subjects in the plane defined by the first two components of the compromise matrix

8.5 Bootstrap

A bootstrap sampling was conducted to compute the confidence intervals and the compromise plot with the bootstrap intervals is shown in figure 8.5

```
## [1] "Making constraints"
```

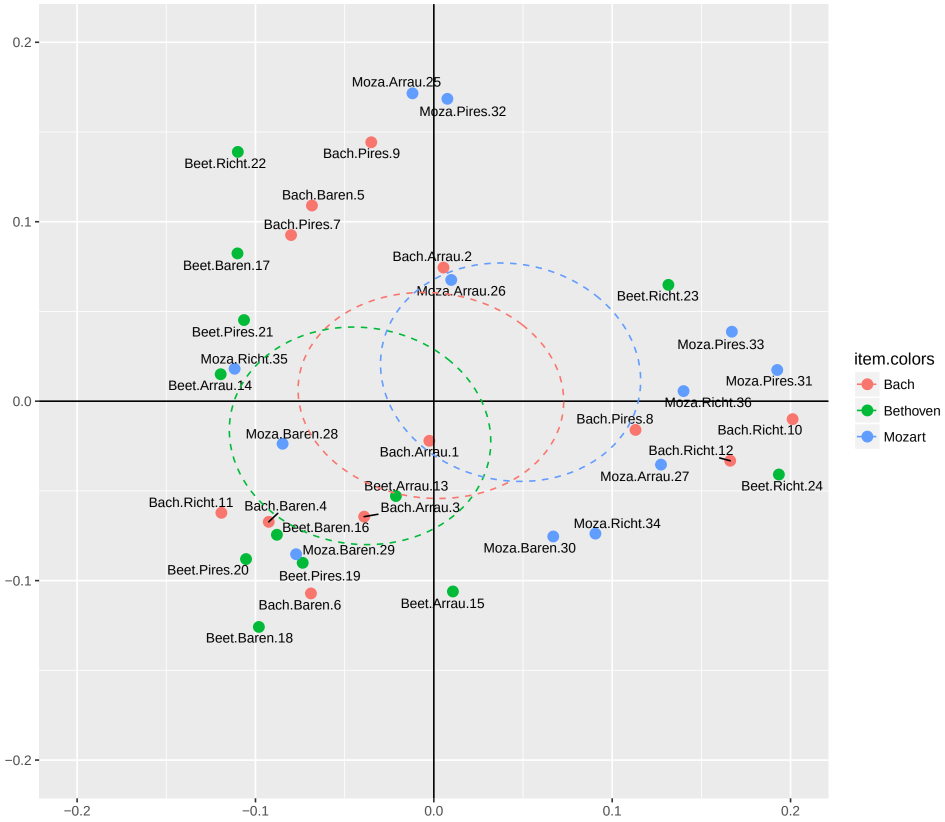


Figure 8.5: Compromise with Bootstraped Confidence Intervals

Chapter 9

MFA/STATIS

Multiple Factor Analysis (MFA) analyzes observations described by several blocks or sets of variables. MFA is performed in two steps. First a principal component analysis is performed on each data set which is then normalized by dividing all its elements by the square root of the first eigenvalue obtained from its PCA. Second, the normalized dataset are merged to form a unique matrix and a global PCA is performed on this matrix. The individual data sets are then projected onto the global analysis to analyze communalities and discrepancies(Herve Abdi, 2007c).

To illustrate MFA, we use the `World Health` dataset. The heatmap of the dataset is shown in figure 1.1

9.1 Computation

For our exaple we have four sub-tables. Let's call them $\mathbf{X}_{[1]}$, $\mathbf{X}_{[2]}$, $\mathbf{X}_{[3]}$ and $\mathbf{X}_{[4]}$. The first singular value is the normalizing factor used to divide the elements of the data table. For example, the PCA of the first group gives a first eigenvalue φ_1. This gives the first normalized data matrix $\mathbf{Z}_{[1]}$:

$$\mathbf{Z}_{[1]} = \varphi_1^{-1} \times \mathbf{X}_{[1]}$$

. Similarly matrices $\mathbf{Z}_{[2]}$, $\mathbf{Z}_{[3]}$ and $\mathbf{Z}_{[4]}$ are computed. The Scree plot of our given dataset is shown in figure 9.1

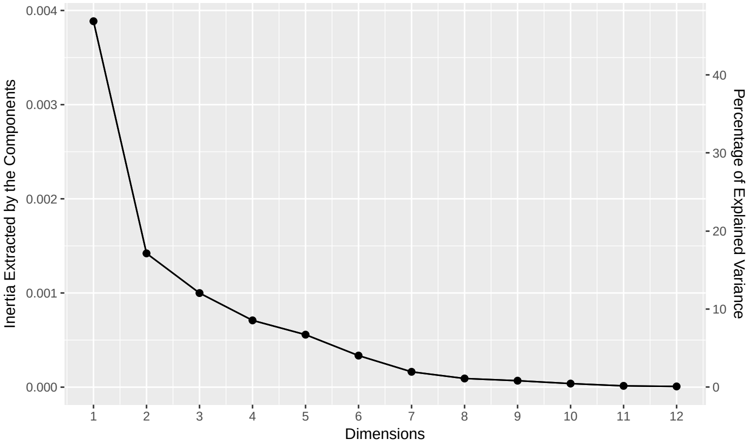


Figure 9.1: Scree Plot for MFA

These normalized matrices are concatenated into an $I \times T$ matrix called the global data matrix dentoed as $\mathbf{Z}$.

$$\mathbf{Z} = [\mathbf{Z}_{[\mathbf{1}]} \;\; \mathbf{Z}_{[\mathbf{2}]} \;\; \mathbf{Z}_{[3]} \;\; \mathbf{Z}_{[4]}]$$

9.2 Computing the global PCA

To analyze the global matrix, we use standard PCA. This amounts to computing the singular value decomposition of the global data matrix:

$$\mathbf{Z} = \mathbf{U\Delta V^T}$$

and the global factor scores are obtained as:

$$\mathbf{F} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U\Delta}$$

The factoe scores of the given data table is shown in figure **??** (Components 1 and 2) and 9.2 (Components 3 and 4)

```
## pdf
##   2
```

From the above figure we can see that component 1 is seperating Africa from Europe and component 2 is seperating Asia from Europe

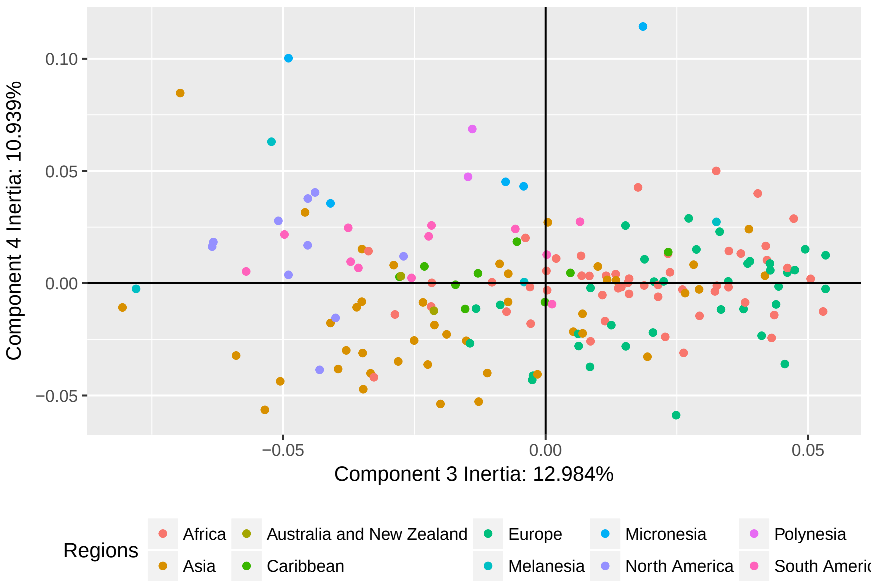


Figure 9.2: MFA World Health Characteristics. Factor scores of the observations plotted on components three and four

9.3 Partial Factor Scores

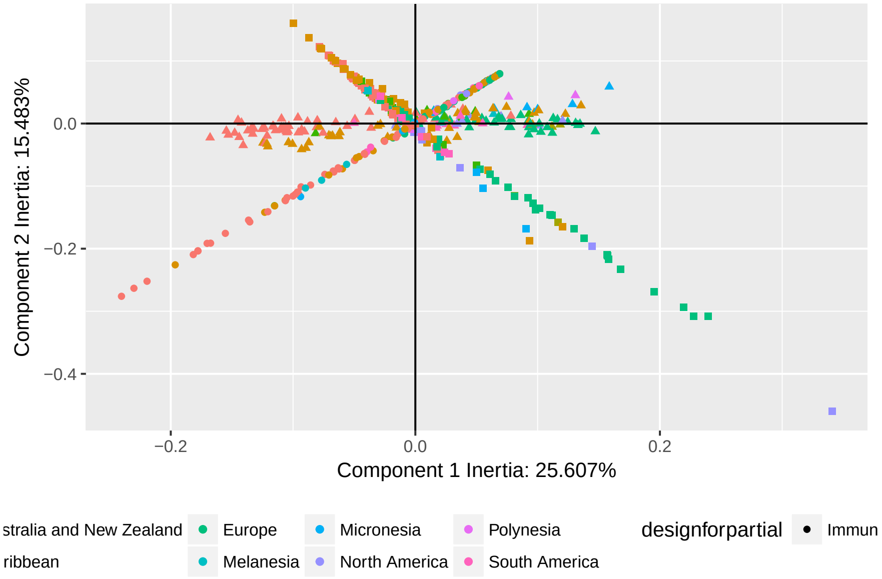


Figure 9.3: MFA World Health Characteristics. Partial Factor scores of the observations plotted on the first two components

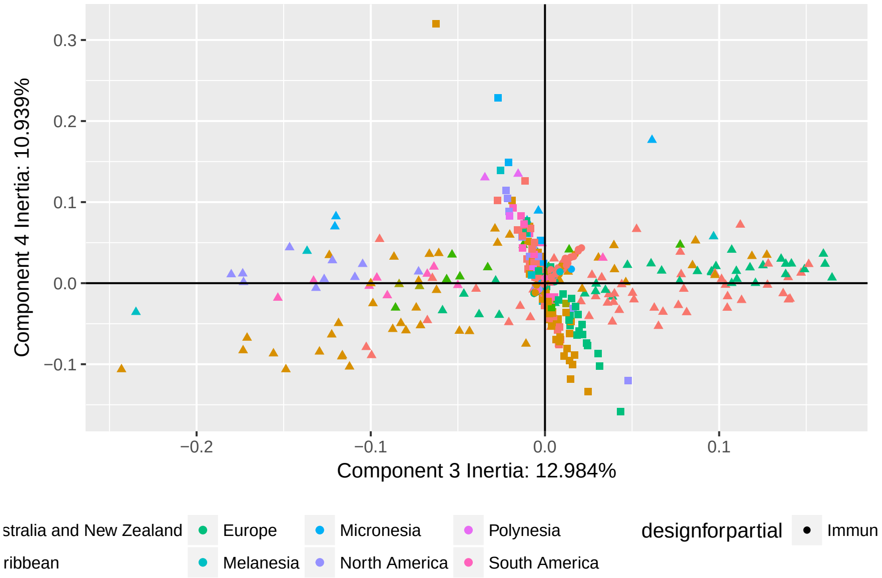


Figure 9.4: MFA World Health Characteristics. Partial Factor scores of the observations plotted on components three and four

9.4 Loadings

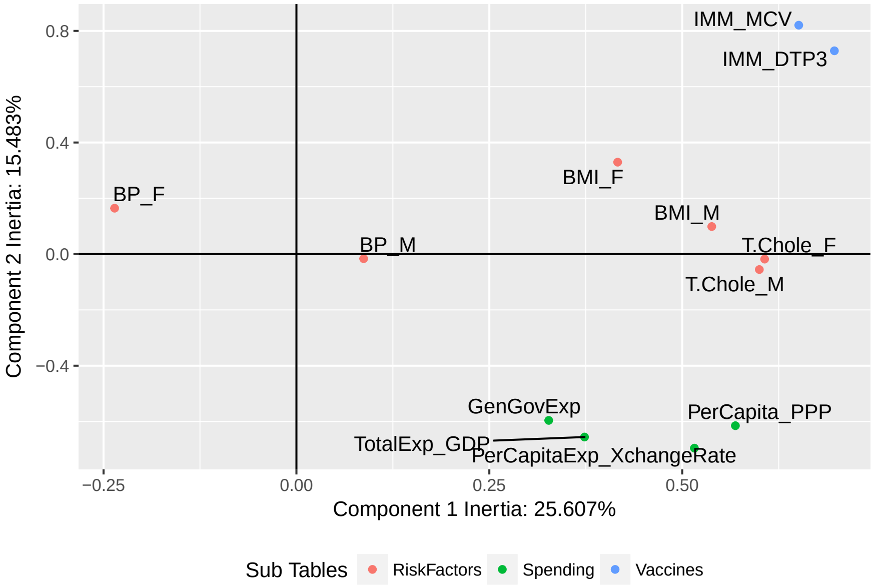


Figure 9.5: MFA World Health Characteristics. Loadings of the observations plotted on the first two components

From the above figure we can see that variables within each sub-table is very well correlated as they are very near to each other in the Loadings map.

Bibliography

Abdi, H. (2003). Multivariate analysis. In *Encyclopedia of Social Sciences Research Methods*. Sage.

Abdi, H. (2007). Discriminant correspondence analysis. In *Encyclopedia of Measurement and Statistics*. Sage.

Herve Abdi, D. V. (2007a). Distatis. In *Encyclopedia of Measurement and Statistics*. Sage.

Herve Abdi, D. V. (2007b). Multiple correspondence analysis. In *Encyclopedia of Measurement and Statistics*. Sage.

Herve Abdi, D. V. (2007c). Multiple factor analysis. In *Encyclopedia of Measurement and Statistics*. Sage.

Herve Abdi, L. J. W. (2010a). Correspondence analysis. In *Encyclopedia of Research Design*. Sage.

Herve Abdi, L. J. W. (2010b). Principal component analysis. In *Wiley Interdisciplinary Reviews: Computational Statistics*. Wiley.

Herve Abdi, L. J. W. (2013). Partial least square methods: Partial least squares correlation and partial least square regression. In *Computational Toxicology: Volume II*. Springer Science.

Herve Abdi, L. J. W. (2018). Barycentric discriminant analysis. In *Encyclopedia of Social Networks and Mining*. Springer.