Research Methods Cookbook

Sruti Jain

November 10, 2017

## Objective

The objective of this markdown book is analyzing the output of multiple data analysis techniques and selecting the most appropriate one given a dataset. Below are all the techniques covered in this cookbook that are used for this analysis:

1. **Principal Component Analysis (PCA)** is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation). The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.
2. **Correspondence Analysis (CA)** is an extension of principal component analysis suited to explore relationships among qualitative variables in the form of contingency tables. When analyzing a two-way contingency table, a typical question is whether certain row elements are associated with some elements of column elements. Correspondence analysis is a geometric approach for visualizing the rows and columns of a two-way contingency table as points in a low-dimensional space, such that the positions of the row and column points are consistent with their associations in the table. The aim is to have a global view of the data that is useful for interpretation.
3. **Multiple Correspondence Analysis (MCA)** is an extension of the simple correspondence analysis for summarizing and visualizing a data table containing more than two categorical variables & is obtained by using a standard correspondence analysis on an indicator matrix . It can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative. MCA can also accommodate quantitative variables by recoding them as "bins."
4. **Partial Least Square Analysis (PLS)** 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 methods proceed by deriving latent variables which are linear combinations of the variables of a data table.When the goal is to ???nd the shared information between two tables,the approach is equivalent to a correlation problem and the technique is then called partial least square correlation (PLSC). In this case there are two sets of latent variables (one set per table), and these latent variables are required to have maximal covariance. When the goal is to predict one data table the other one, the technique is then called partial least square regression.
5. **Barycentric Discriminant Analysis (BADA)** 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 a priori defined categories. The goal of BADA is to combine the measurements to create new variables (called components or discriminant variables) that best separate the categories. These discriminant variables are also used to assign the original observations or "new" observations to the a-priori defined categories.
6. **Discriminant Correspondence Analysis (DICA)** 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 apriori and the a posteriori classifications can be used to assess the quality of the discrimination.
7. **DiSTATIS** 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. The general idea behind DISTATIS is first to transform each distance matrix into a cross-product matrix as it is done for a standard MDS. Then, these cross-product matrices are aggregated to create a compromise cross-product matrix which represents their consensus. The compromise matrix is obtained as a weighted average of individual cross-product matrices. The PCA of the compromise gives the position of the objects in the compromise space. The position of the object for each study can be represented in the compromise space as supplementary points. Finally, as a byproduct of the weight computation, the studies can be represented as points in a multidimensional space.
8. **Multiple Factor Analysis (MFA/STATIS)** combines several data tables into one single analysis. The ???rst step is to perform a PCA of each table. Then each data table is normalized by dividing all the entries of the table by the ???rst eigenvalue of its PCA. This transformation-akin to the univariate Z-score-equalizes the weight of each table in the ???nal solution and therefore makes possible the simultaneous analysis of several heterogenous data tables.
9. **Inferential Methods** Used for testing the stability of above methods with 0.05 level of significance
   1. **Bootstrapping**: In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates.This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of resampling methods.
   2. **Permutation**: In simple terms, permutation is all the possible arrangements of a collection of things, where the order is important. In statistics, it is defined as the notion that relates to the act of arranging all the members of a set into some sequence or order, or if the set is already ordered, rearranging (reordering) its elements, a process called permuting.
   3. **Jackknife**: In statistics, the jackknife is a resampling technique especially useful for variance and bias estimation. The jackknife predates other common resampling methods such as the bootstrap. The jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations. Given a sample of size n, the jackknife estimate is found by aggregating the estimates of each n-1-sized sub-sample.

## function (f)   
## markdown:::.b64EncodeFile(f)  
## <environment: namespace:knitr>

## Dataset Information: SAM, BAMA, OSIQ, Judge, BFI

1. **The Survey of Autobiographical memory (SAM)** was designed to access trait mnemonics of naturalistic episodic autobiographical, semantic, spatial memory & future thinking. It records the data taken from a sample of 153 people (rows) to measure mnemonic abilities using 31 variables. There are two nominal variables (Active, Sex) & one factor variable (MemoryGroup). The rest of the variables are all integer with two (ID & Age) being a ratio-scale and all others interval variables.

*Research Hypothesis*: How the survey questionaire related to memory?

1. **BAMA Politics Dataset** was a statewide survey of 501 adult citizens collecting responses to questions on various topics including education, ratings and images of political parties. The contingency table was created for politics versus education and then CA analysis was run on the contingency table.

*Research Hypothesis*: Does the education of an individual influence the choice of political party?

1. **Object-Spatial Imagery Questionnaire (OSIQ)** The same participants of SAM data were now asked the Object-Spatial Imagery questions that consisted of two separate scales: an object imagery scale, with self-report items designed to assess object imagery preferences and experiences, and a spatial imagery scale, with self-report items designed to assess spatial imagery preferences and experiences. There were 14 Spatial Imagery questions & 16 Object Imagery questions that were rated on a scale of 1-5.

*Research Hypothesis*: How does OSIQ (Object-Spatial Imagery) survey relate to different types of memory (Episodic, Semantic, Spatial, Future) ?

1. **Judges Sort Data - Orange Juice Test**: The dataset consist of rating from 44 Judges to 10 different orange juice samples. We are interested in knowing how each judge's rating is comparable to other judges while also how these juices are related or similar to one another. So we compute Distance Matrix and then perform a MDS/DISTATIS to evaluates the similarity between Distance matrices. We then compute a compromise matrix which represents the best aggregate of the original matrices. The original distance matrices are then projected onto the compromise for analysis.

*Research Hypothesis*: How are the ratings of various judges comparable and which juices share a common taste ?

1. **BFI Dataset (BIG 5 personality questionnaire)**: Same participants of SAM & OSIQ data were now asked questions related to personality traits. The survey consisted of question divided into 5 segments (Ex\_: Extraversion questions, Ag\_: Agreeable questions, Co\_: Conscientiousness questions, Ne\_: Neuroticism questions, Op\_: Openness questions) all rated on scale of 1-5.

*Research Hypothesis*: How does memory (SAM & OSIQ) have an impact to personality traits (BFI) of an individual ?

## Functions used: PlotScree(), perm4PLSC(), Boot4PLSC()

## First Analysis : Principal Component Analysis (PCA) of SAM Dataset

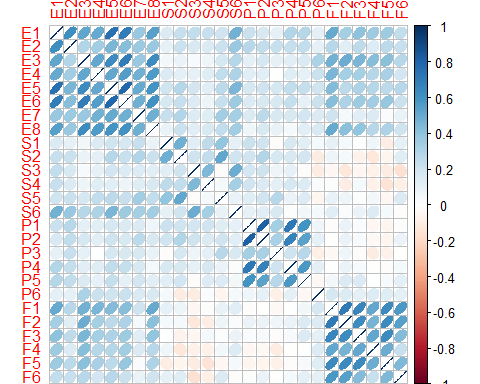
**Loading the SAM Dataset**

## ID memoryGroups Active Sex Age E1 E2 E3 E4 E5 E6 E7 E8 S1 S2 S3 S4 S5  
## 1 164 High TRUE 2 36 5 5 5 5 5 5 5 5 5 5 4 5 5  
## 2 220 High TRUE 1 63 4 4 4 3 4 4 4 3 4 4 4 4 4  
## 3 222 High TRUE 1 43 1 5 1 1 1 1 1 1 2 5 4 4 3  
## S6 P1 P2 P3 P4 P5 P6 F1 F2 F3 F4 F5 F6  
## 1 5 4 4 2 5 3 5 4 3 3 4 4 5  
## 2 3 4 5 3 5 4 4 4 4 4 4 4 4  
## 3 4 4 2 5 1 2 3 1 1 1 1 5 5

For more info, try: "summary(SAMdata)" & "str(SAMdata)"

**Preprocessing the dataset, creating design variables & exploratory analysis**

First we leave out the mystery group to be analyzed later on as supplementary elements. Next we separate the questionaire variables and design variables (Memory Groups, Gender, Age) while keep the row namas as participant IDs. We also do a correlation plot of the analysis variables to get some insights into the data.



Correlation heatmap plot indicates that variables within the questionaire that are testing for a particular memory type are strongly correlated to one another. That is to say that all variable testing episodic memory are highly correlated & this is applicable for variables refering to all memory groups. Also negative correlation was observed between variables testing for future & semantic memory types.

**Using InPosition Inference battery for PCA Analysis**

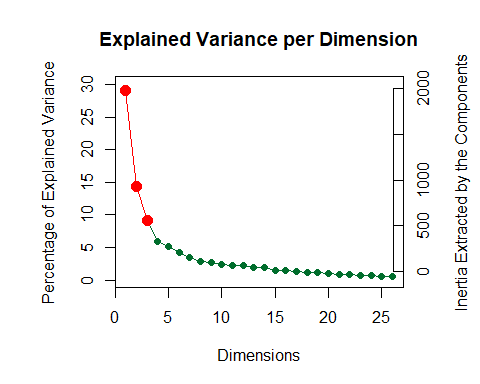
We use the function epPCA in InPosition library for this analysis and store the results in res\_pca. We center the variables as 0 is not a meaningful value for these variables (there is nothing like zero memory score). Also as we have a balanced scale across all variables (Scale: 1-5), I preferred not scaling these variables. My design matrix is already in the form of nominal matrix, the parameter "make\_design\_nominal" = FALSE. I have given 2000 iterations for my bootstrap results.

res\_pca <- epPCA.inference.battery(ModSamData, center = TRUE, scale = FALSE, graphs = FALSE, DESIGN = Des.mat1, make\_design\_nominal = FALSE, test.iters = 2000)

## [1] "It is estimated that your iterations will take 36.67 minutes."  
## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please take note of the progress bar."  
## ===========================================================================

\*\* Scree Plot \*\*

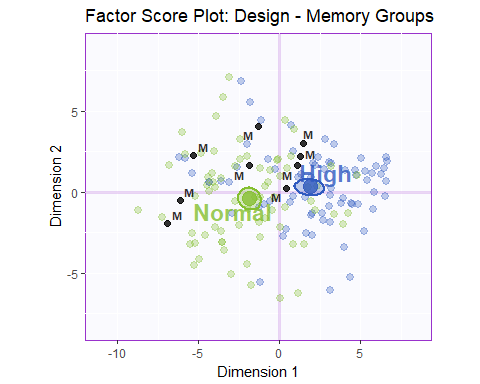
A Scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The number of components is min(nrow(DATA), ncol(DATA)). Here, 26 columns gives max of 26 components. The scree plot is used to determine how many of the components should be interpreted. The significant components are color coded using the bootstrap results.

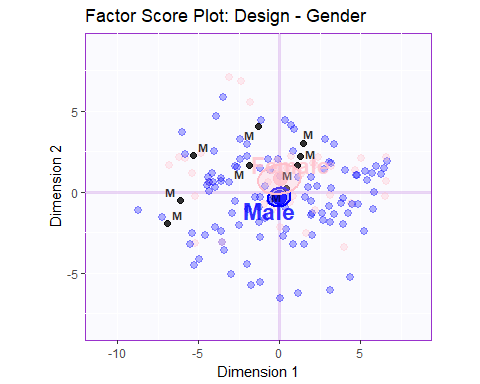


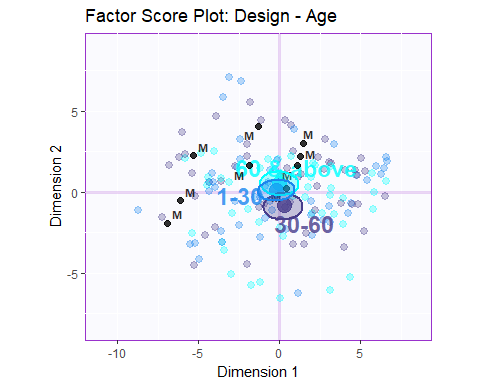
We have three significant components with the first component explaining roughly 27% of variance and second component explaining 15% of variance in the data. We will be analyzing just the first two component in this cookbook.

**Factor scores**:

Factor scores are the coordinates of the observations on the components. The distances between them show which individuals are most similar. Factor scores can be color-coded to help interpret the components.



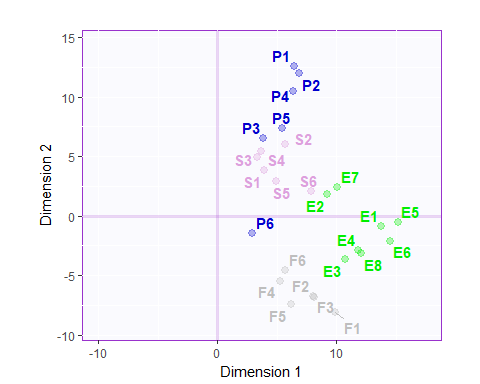




In the above factor score plots, we see that Component 1 mainly distinguish between subjects in "High" & "Normal" Memory Groups. Gender & Age have no effect on Memory as per the Factor Score plots for Gender & Age. There isn't anything that component 2 is separating and things are almost evenly balanced on Component 2. I have also plotted the supplementary elements to the plots but they don't tend to show any trend with respect to Memory Groups, Gender or Age.

**Loadings**

Loadings describe the similarity (angular distance) between the variables. Loadings show how the input variables relate to each other. Loadings also show which variables are important for (which components load on) a certain component.



1. Component 1: Seperates normal memory score versus high memory scores and all the varaibles test for the memory scores in positive terms.
2. Component 2: Mainly composed of Spatial memory variable. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated.

**Bootstrap ratios**

Here we are going to plot the bootstrap ratios for component1 & component2 and also plot the correlation circle

All the variables are significant for the first component as per the bootstrap results. Questions related to Episodic majorly contributes to this Component.

The second component mainly distinguishes between Future & Spatial memory questions and these two groups majorly contributes to the second component as seen in the graph.

From correlation circle plot above, it is concluded that all variables test for memory scores in a positive manner and Future and Spatial Memory questions are negatively correlated to one another as also seen in the correlation heatmap.

**Summary**

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

1. Component 1: The latent structure of the SAM data as revealed by PCA indicated that the first component characterized good versus poor memory. In other words,when people reported having high or low abilities for one category of memory,they tended to do the same for other categories.
2. People who are good at Spatial memory & Future thinking are indicated to have an overall good memory in other categories as well.
3. Component 2 is mainly composed of Spatial memory variable. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated.

## Second Analysis :Correspondence Analysis (CA) of BAMA Politics Dataset

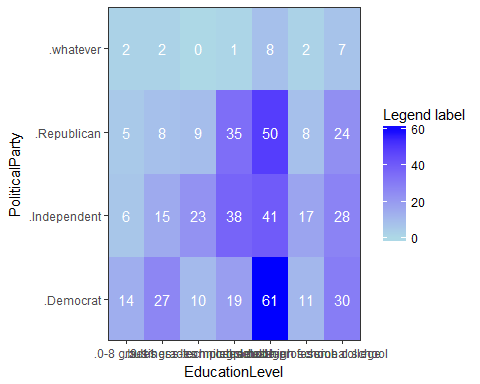
**Loading the BAMA Politics Dataset**

## q1 q2 q3 q4 q5 q6 q7 q8 q9.1  
## 1 excellent good poor favor legal as now favor favor favor positive  
## q9.2 q9.3 q9.4 q10 q11 q12.1 q12.2 q12.3 q12.4 q13.1 q13.2  
## 1 negative dk-na dk-na Bush dk-na dk-na dk-na dk-na dk-na Bush Bush  
## q13.3 q13.4 q13.5 q13.6 q14.1 q14.2 q14.3 q14.4 q14.5  
## 1 Bush Bush Bush Bush both equal both equal Bush Bush Bush  
## q14.6 q14.7 q14.8 q14.9 q14.10 q14.11 q14.12 q15  
## 1 both equal both equal both equal Bush Dukakis both equal Bush yes  
## q16 q17 q18 q19 q20 q21  
## 1 not asked both equal no difference dk-na better Republicans

For more info, try: "summary(BAMA)", "str(BAMA)" or "head(BAMA)"

**Preprocessing the dataset, creating design variables & contingency tables**

First, we separate the questionaire variables from the design variables. Next we want to club several responses in a variable that contain only a few participants(eg dk-na, none, na) to one category. We then compute a contingency table with politics & education as primary variables to be analyzed.



## Warning in chisq.test(Politics\_vs\_Education): Chi-squared approximation may  
## be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: Politics\_vs\_Education  
## X-squared = 44.047, df = 18, p-value = 0.0005681

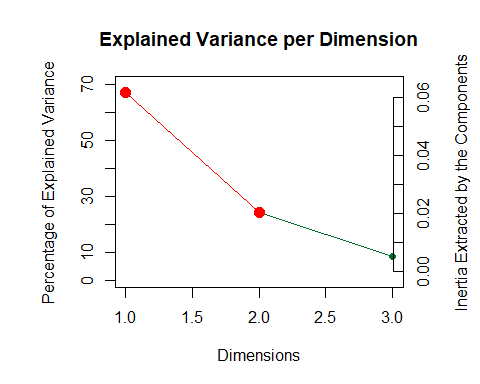
**Using InPosition Inference battery for CA Analysis**

We use the function epCA.inference.battery in InPosition library for this analysis and store the results in resinf\_ca. We are using Chi-Square distance for this analysis. I have given 2000 iterations for my bootstrap results. Also as my data is symmetric (changing the rows with columns will not change the interpretation of my analysis), I would like symmetric = TRUE.

# Then, we can move on to CA & Inferential CA  
resinf\_ca <- epCA.inference.battery(Politics\_vs\_Education, graphs = FALSE, test.iters = 2000, symmetric = TRUE)

## [1] "It is estimated that your iterations will take 5.67 minutes."  
## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please take note of the progress bar."  
## ===========================================================================

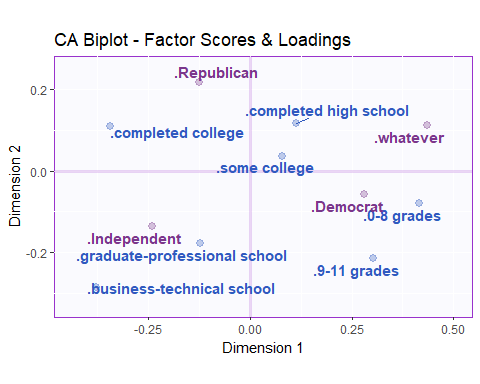
\*\* Scree plot for CA\*\*



As no of components is min(row,col)-1, we see that there are 3 components out of which the first two components are significant and they explain almost 90% variance in data. We will be analysing the first two component in this cookbook.

**Biplot - Factor Scores & loadings**

The graph below is called symetric biplot plot and shows a global pattern within the data. The distance between any row points or column points gives a measure of their similarity (or dissimilarity). Row points with similar profile are close on the factor map. The same holds true for column points.



In the above biplot we see that the component 1 characterized Republican & Independent versus Democrat parties. Also it indicates that educated people prefer Republican & Independent over Democrat. Component 2 mainly distinguishes people supporting the Republican versus Independent & Democrats.

**Bootstrap ratios**

Here we are going to plot the bootstrap ratios that we get from inference results for component1 & component2

From the bootstrap ratio plots we see that the major contributors to component1 are the Independent & Democrat parties that goes from people with high eductation to no education. So this implies that the choice of political parties depends on the level of education of an individual. Component 2 has just the Republican as a significant component overall segregating all the 3 choice of parties with different level of education amongst people.

**Summary**

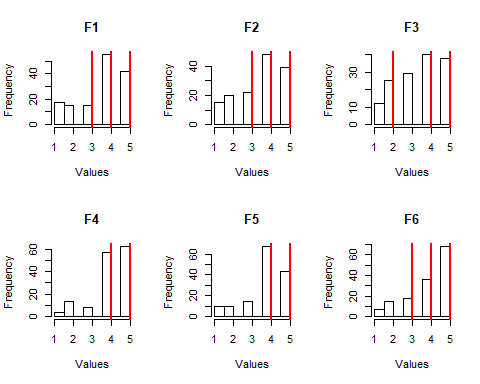
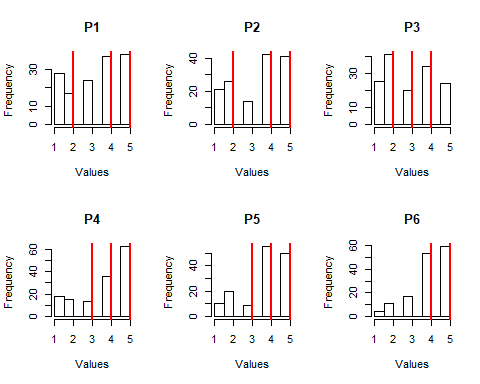
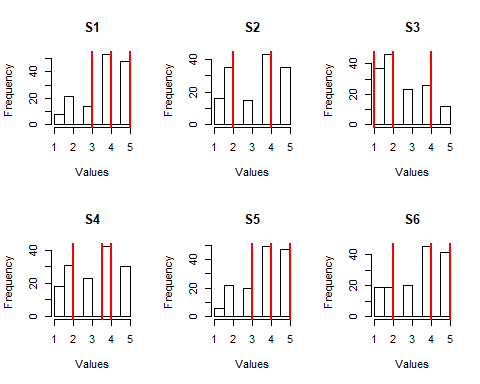
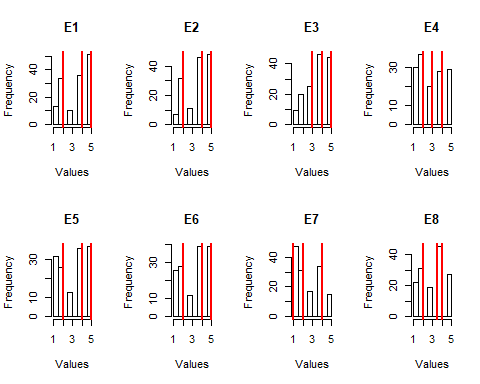
When we interpret the biplot and bootstrap ratio plots together, the CA revealed:

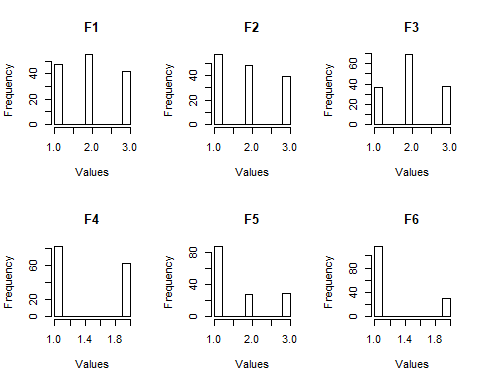
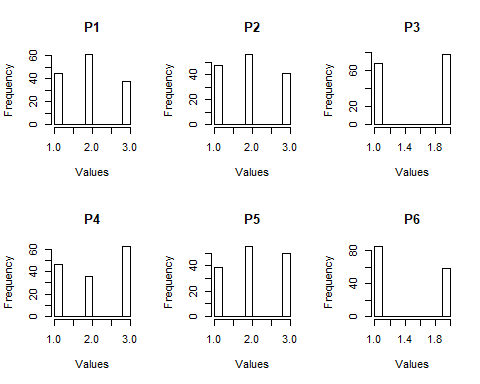
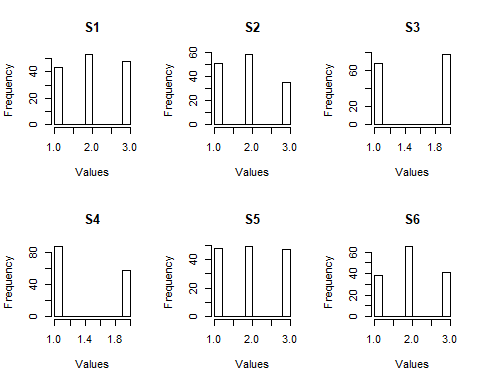
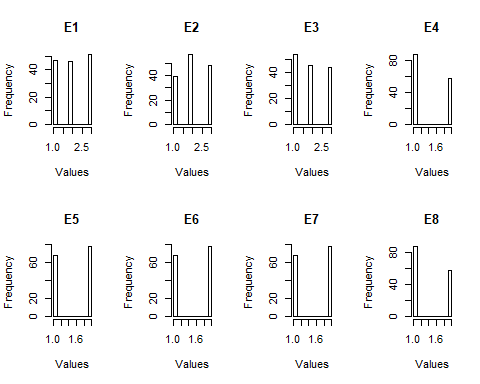
1. Component 1: The latent structure of the Bama Politics data as revealed by CA indicated that the first component characterized Republican & Independent versus Democrat. Also it indicates that educated people prefer Republican & Independent over Democrat.
2. Component 2: Mainly distinguishes people supporting the Republican versus Independent & Democrats.

## Third Analysis : Multiple Correspondence Analysis (MCA) of SAM Dataset

We will be using the same dataset as we used for PCA, ie SAM Dataset but only after binning all the quantitative variables.

**Plotting the hist, creating binned variables & performing exploratory analysis** We need to create binned (qualitative data) by using the cut function available in R but in a way that the data is balanced. So we first plot the hist of each of these variables, and then decide on the no of levels and their cutoff values.





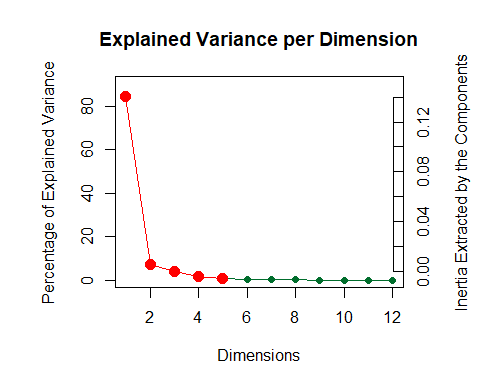
**Using InPosition Inference battery for MCA Analysis**

We use the function epMCA in InPosition library for this analysis and store the results in res\_mca. I have used a asymmetric method as changing the rows with columns will change the analysis of my design. My design matrix is already in the form of nominal matrix, the parameter "make\_design\_nominal" = FALSE. I have given 2000 iterations for my bootstrap results.

res\_mca <- epMCA.inference.battery(new\_data, graphs = FALSE, DESIGN = Des.mat1, make\_design\_nominal = FALSE, test.iters = 2000, symmetric = FALSE)

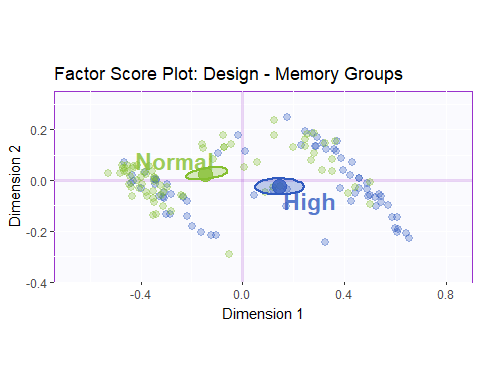
## [1] "It is estimated that your iterations will take 9 minutes."  
## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please take note of the progress bar."  
## ===========================================================================

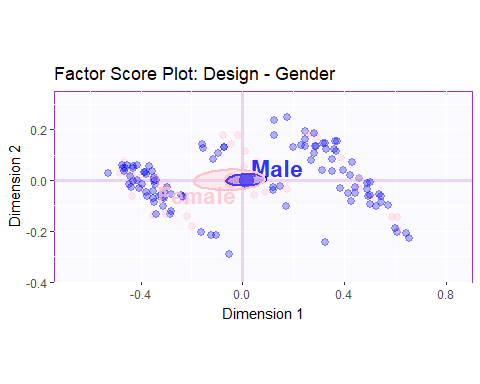
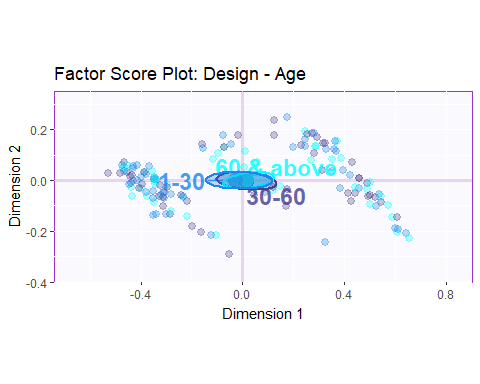
\*\* Scree Plot \*\* A Scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The number of components is min(nrow(DATA), ncol(DATA)). The scree plot is used to determine how many of the components should be interpreted. The significant components are color coded using the bootstrap results.



The Scree plot indicates that there are five significant components as per the inference results, however we will be analysing just the first two as these explain for around 92% of the variance in the data.

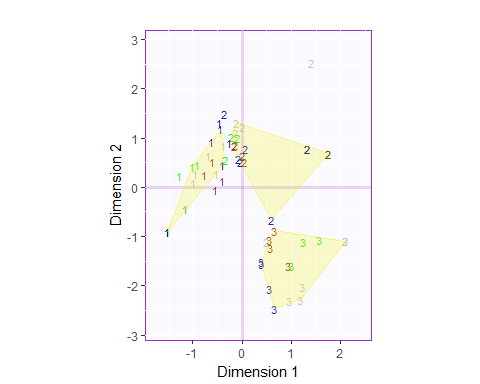
**Factor scores**: Factor scores are the coordinates of the observations on the components. The distances between them show which individuals are most similar. Factor scores can be color-coded to help interpret the components.



  In the above factor score plots, we see that Component 1 mainly distinguish between subjects in "High" & "Normal" Memory Groups. Gender & Age have no effect on Memory as per the Factor Score plots for Gender & Age. There isn't anything that component 2 is separating and things are almost evenly balanced on Component 2.

**Loadings**

Loadings describe the similarity (angular distance) between the variables. Loadings show how the input variables relate to each other. Loadings also show which variables are important for (which components load on) a certain component.



1. Component 1: Seperates normal memory score versus high memory scores and all the variables test for the memory scores in positive terms. The convex hulls are plotted with 60% tolerance levels for each score.
2. Component 2 is mainly composed of Spatial memory variable. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated.

**Bootstrap ratios**

Here we are going to plot the bootstrap ratios for component1 & component2 and also plot the correlation circle

The first component is composed of low versus high memory scores across all memory groups as seen in the graph. Questions related to Episodic & Future memory majorly contributes to this Component.

The second component mainly distinguishes between Future & Spatial memory questions and these two groups majorly contributes to the second component as seen in the graph.

**Summary**

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

1. Component 1: The latent structure of the SAM data as revealed by MCA indicated that the first component characterized good versus poor memory. In other words,when people reported having high or low abilities for one category of memory,they tended to do the same for other categories.
2. Component 2 is mainly composed of Spatial & Future memory variable. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated.

## Fourth Analysis : Partial Least Square - Correlation Analysis (PLSC) of SAM & OSIQ Dataset

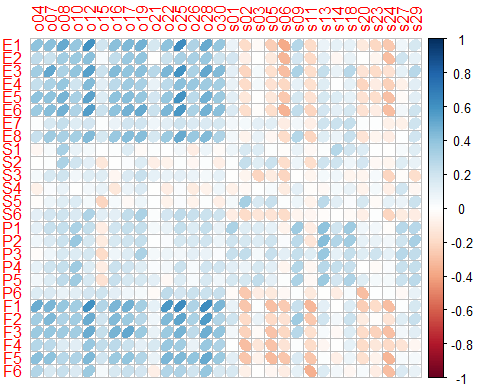
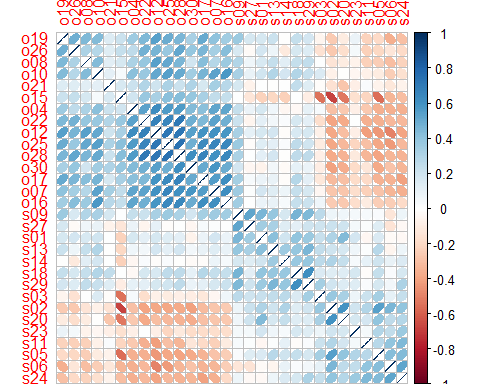
**Loading the OSIQ Dataset**

## ID memoryGroups Active Sex Age s01 s02 s03 o04 s05 s06 o07 o08 s09 o10  
## 1 164 High TRUE 2 36 4 3 3 4 2 1 4 2 1 5  
## 2 220 High TRUE 1 63 2 4 2 4 2 4 4 4 2 5  
## 3 222 High TRUE 1 43 5 5 1 2 5 1 1 1 1 1  
## 4 444 High TRUE 1 34 4 4 2 2 5 5 5 3 4 4  
## 5 586 High TRUE 2 62 3 2 3 4 2 2 4 2 5 5  
## 6 590 High TRUE 2 43 1 1 4 4 2 2 4 1 3 5  
## s11 o12 s13 s14 o15 o16 o17 s18 o19 s20 o21 o22 s23 s24 o25 o26 s27 o28  
## 1 2 3 3 4 3 2 3 1 4 4 2 3 2 2 5 1 5 2  
## 2 3 4 4 3 4 4 4 3 2 4 4 3 2 2 4 4 2 4  
## 3 1 2 3 1 1 1 2 1 4 5 1 1 5 1 1 5 5 1  
## 4 4 4 2 5 2 4 5 2 4 4 4 4 2 4 5 4 4 2  
## 5 3 5 5 3 5 5 5 3 5 2 3 3 1 1 5 3 5 4  
## 6 2 4 4 3 4 4 5 1 2 1 4 2 1 1 4 3 2 4  
## s29 o30  
## 1 1 3  
## 2 3 4  
## 3 1 1  
## 4 1 1  
## 5 3 4  
## 6 2 5

For more info, try: "summary(OSIQ)" & "str(OSIQ)"

**Preprocessing the dataset & performing exploratory analysis**

First we leave out the mystery group to be analyzed later on as supplementary elements. Next we separate the questionnaire variables and design variables (Memory Groups, Gender, Age) while keep the row names as participant IDs. We also do a correlation plot of the analysis variables to get some insights into the data.

 Correlation plot of OSIQ data indicates that variables within the questionaire that are testing for a particular memory type are strongly positively correlated to one another. The object and spatial memory variables are negatively correlated to one another. Also the correlation plot of the combined data represents that object memory questions of OSIQ dataset are positively correlated with SAM variables and spatial questions of OSIQ are negatively correlated with SAM Variables.

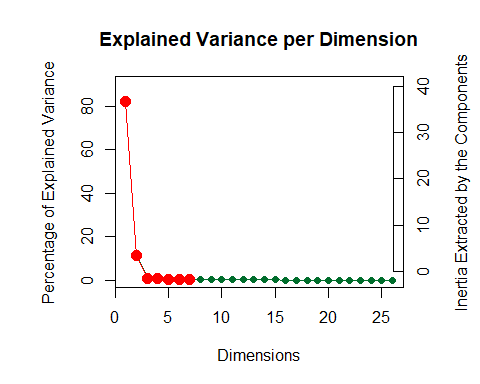
**Using TExPosition position for implementing PLSC Analysis**

We use the function tepPLS in TExPosition library for this analysis and store the results in pls.res. We give the two data matrix as input to the function and a design matrix.

pls.res <- tepPLS(ModSamData, ModOSIQData, graphs = FALSE, DESIGN = Des.mat1, make\_design\_nominal = FALSE)  
# Calling the functions perm4PLSC() and Boot4PLSC to compute inference results using permutation & bootstrap  
permres <- perm4PLSC(ModSamData, ModOSIQData)  
bootres <- Boot4PLSC(ModSamData, ModOSIQData)

## Warning in matrix(svd.S$d, nJ, nf2keep, byrow = TRUE): data length [26] is  
## not a sub-multiple or multiple of the number of rows [30]

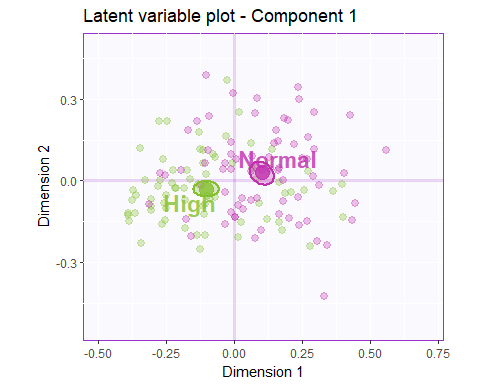
\*\* Scree Plot \*\* A Scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The number of components is min(nrow(DATA), ncol(DATA)). The scree plot is used to determine how many of the components should be interpreted. The significant components are color coded using the bootstrap results.

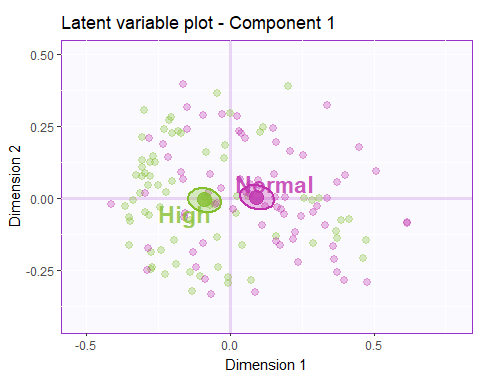


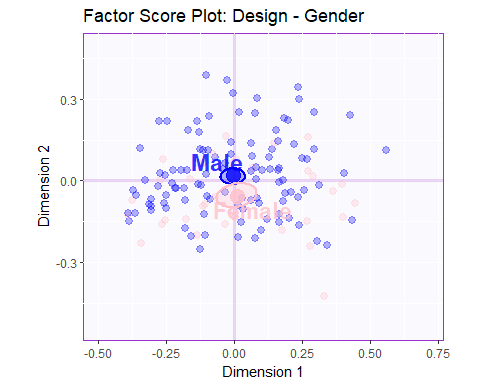
We have seven significant components with the first component explaining roughly 82% of variance and second component explaining 10% of variance in the data. We will be analyzing just the first two component in this cookbook as in total they explain for 92% of variance in the data.

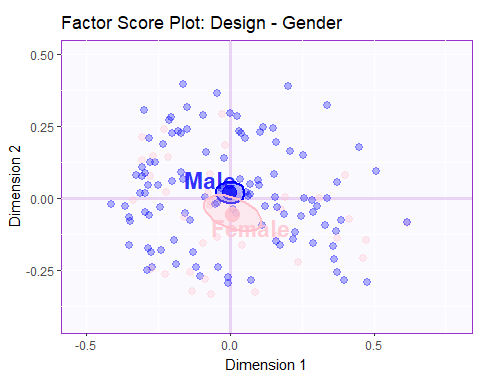
**Latent variables plot**:

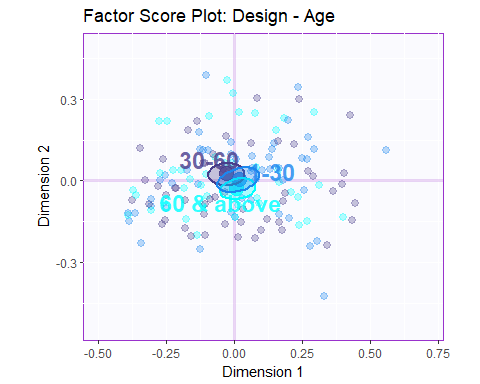
Latent variable are the coordinates of the observations on the components. The distances between them show which individuals are most similar. Latent variable scores can be color-coded to help interpret the components.

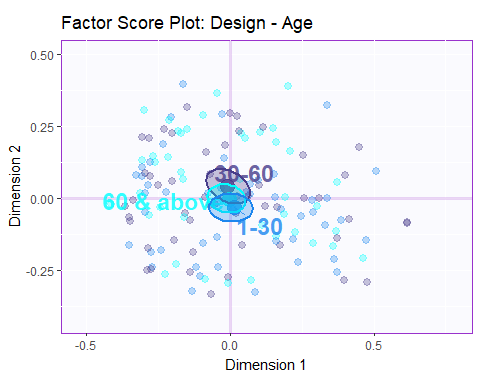






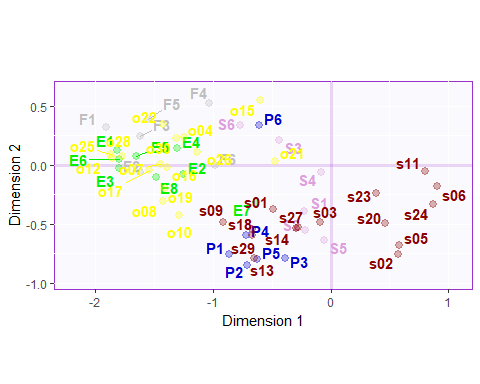






**Salience Plot**

Salience describe the similarity (angular distance) between the variables. Loadings show how the input variables relate to each other. Salience also show which variables are important for (which components load on) a certain component.



1. Component 1 & Component 2: Seperates normal memory score versus high memory scores and all the varaibles test for the memory scores in positive terms for both the SAM & OSIQ data.
2. Component 1 separates object questions from spatial ones. Component 2 separates mainly the future & spatial memory components.

**Bootstrap ratios**

Here we are going to plot the bootstrap ratios for component1 & component2 and also plot the correlation circle

```{{r bootstrap\_pls1, echo=FALSE}} # Storing the results in a variable for ease of use bootstrapC1 <- data.frame(bootresbootRatios.j[,c(1,2)]) color4row <- as.matrix(color4row)

# This prints the Bootstrap for Component 1 for both the SAM & OSIQ variables

bootstrap\_C1\_SAM <- prettyBars(bootres$bootRatios.i[,c(1,2)], axis = 1, threshold.line = TRUE , fg.col = color4col)

legend(x="topright", pch = 15, legend = c("Episodic", "Semantic", "Spatial", "Future"), col=c("green2", "plum", "mediumblue", "grey"))

bootstrap\_C1\_OSIQ <- prettyBars(bootres$bootRatios.j[,c(1,2)],axis = 1, threshold.line = TRUE, fg.col = color4row)

legend(x="topright", pch = 15, legend = c("Spatial", "Object"), col=c("darkred", "yellow"))

```

Component1 is mainly comprised of Episodic and Future memory components of SAM Data and it contains mostly object memory components of OSIQ.

Component2 is mainly comprised of Future versus spatial variables of the SAM Data. It also differentiate between the object & spatial memory components of OSIQ dataset and therefore contains both these variables.

**Summary**

When we interpret the Latent variable and Salience together, the PLS analysis revealed:

1. The first two component explains around 94% of variance in the data. Therefore analyzing just the first two components.
2. Component 1 & Component 2: Mainly distinguishes people with high versus normal memory group & high versus low memory score across all memory groups.
3. Compenent 1 also distinguishes between object imagery and spatial imagery. Component 2 seperates future versus the spatial memory.

## Fifth Analysis : Barycentric Discriminant Analysis (BADA) of SAM Dataset

We will use the same SAM dataset that we used for PCA analysis and see if there are any difference in the results:

**Creating design variable** We want to create another design variable that can take into account both the age and memory groups into account.

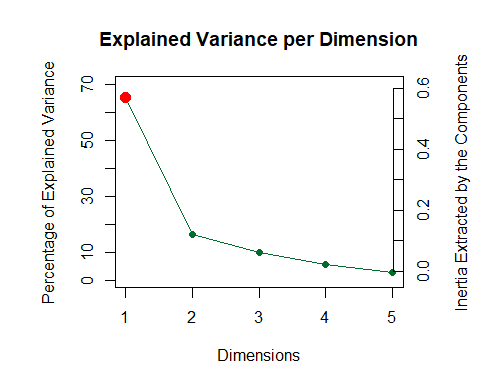
**Using TInPosition Inference battery for BADA Analysis**

We use the function tepBADA.inference.battery in TInPosition library for this analysis and store the results in resBADAInf. The design matrix we create is not in nominal format, the parameter "make\_design\_nominal" = TRUE.

resBADAInf <- tepBADA.inference.battery(ModSamData, DESIGN = DesBADA, make\_design\_nominal = TRUE, center=TRUE, graphs = FALSE )

## [1] "It is estimated that your iterations will take 0.38 minutes."  
## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please take note of the progress bar."  
## ===========================================================================

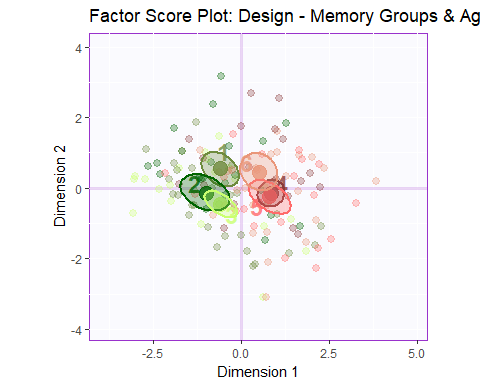
\*\* Scree Plot \*\* A Scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The number of components is min(nrow(DATA), ncol(DATA)). The scree plot is used to determine how many of the components should be interpreted. The significant components are color coded using the bootstrap results.



We have one significant components explaining roughly 75% of variance and second component explaining 15% of variance in the data. We will be analyzing just the first two component in this cookbook.

**Factor scores**:

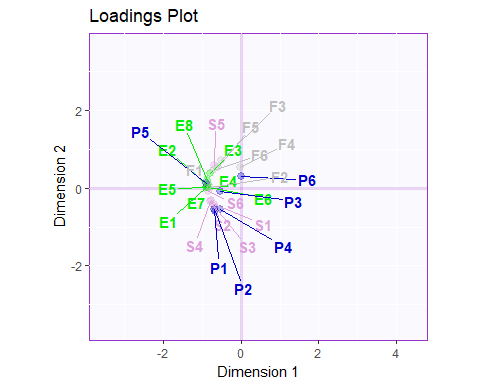
Factor scores are the coordinates of the observations on the components. The distances between them show which individuals are most similar. Factor scores can be color-coded to help interpret the components.



In the above factor score plots, we see that Component 1 mainly distinguish between subjects in "High" & "Normal" Memory Groups. Gender & Age have no effect on Memory as per the Factor Score plots for Gender & Age. There isn't anything that component 2 is separating and things are almost evenly balanced on Component 2.

**Loadings**

Loadings describe the similarity (angular distance) between the variables. Loadings show how the input variables relate to each other. Loadings also show which variables are important for (which components load on) a certain component.



Component 2 separates Spatial memory to future memory variables. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated. There isn't anything much that component 1 is defining in the above plot.

**Bootstrap ratios**

Here we are going to plot the bootstrap ratios for component1 & component2 and also plot the correlation circle

Questions related to Episodic majorly contributes to this Component. Also it clearly separates high versus normal memory groups irrespective of their age.

The second component mainly distinguishes between Future & Spatial memory questions and these two groups majorly contributes to the second component as seen in the graph. It also tries to separate the older age groups from younger ones

**Summary**

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

1. Component 1: The latent structure of the SAM data as revealed by BADA indicated that the first component characterized good versus poor memory. In other words,when people reported having high or low abilities for one category of memory,they tended to do the same for other categories.
2. Component 2 is mainly composed of Spatial memory variable. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated. It also separates observation of older age group from younger ones. Combining the loading with the factor score analysis we can infer that aged people are good at semantic & spatial memory and younger groups are good at episodic and future memory.

## Sixth Analysis : Discriminant Component Analysis (DICA) of SAM Dataset

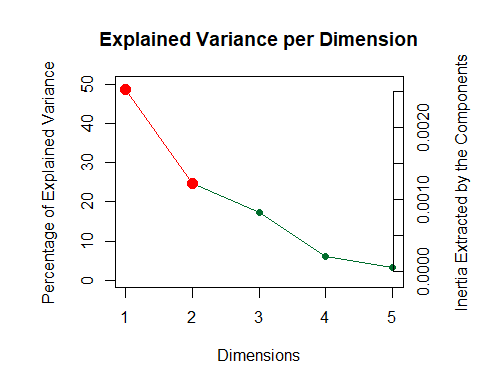
DICA is the same technique as BADA but it is used for qualitative data. As our data is interval scale on (1-5), we can bin the data so that it is qualitative and then run the analysis with the same design matrix that we used in BADA analysis and compare the results.

We use the function tepDICA.inference.battery in TInPosition library for this analysis and store the results in res\_dica. My design matrix is not in the form of nominal matrix, the parameter "make\_design\_nominal" = TRUE.

res\_dica <- tepDICA.inference.battery(new\_data, DESIGN = DesBADA, make\_design\_nominal = TRUE, graphs = FALSE)

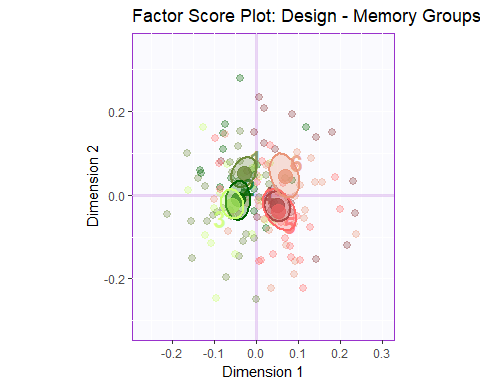
## [1] "It is estimated that your iterations will take 0.47 minutes."  
## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please take note of the progress bar."  
## ===========================================================================

\*\* Scree Plot \*\* A Scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The number of components is min(nrow(DATA), ncol(DATA)). The scree plot is used to determine how many of the components should be interpreted. The significant components are color coded using the bootstrap results.

 We have two significant components with the first component explaining roughly 49% of variance and second component explaining 22% of variance in the data. We will be analyzing just the first two component in this cookbook.

**Factor scores**:

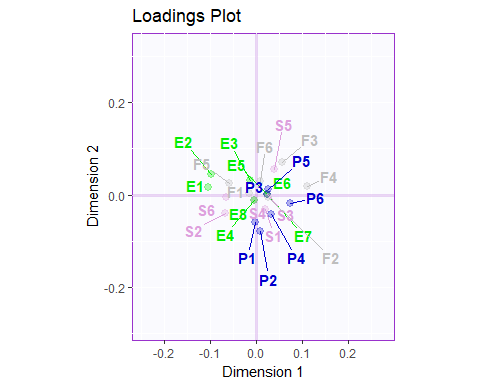
Factor scores are the coordinates of the observations on the components. The distances between them show which individuals are most similar. Factor scores can be color-coded to help interpret the components.



In the above factor score plots, we see that Component 1 mainly distinguish between subjects in "High" & "Normal" Memory Groups. Gender & Age have no effect on Memory as per the Factor Score plots for Gender & Age. There isn't anything that component 2 is separating and things are almost evenly balanced on Component 2.

**Loadings**

Loadings describe the similarity (angular distance) between the variables. Loadings show how the input variables relate to each other. Loadings also show which variables are important for (which components load on) a certain component.



Component 2 separates Spatial memory to future memory variables. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated. There isn't anything much that component 1 is defining in the above plot.

**Bootstrap ratios**

Here we are going to plot the bootstrap ratios for component1 & component2 and also plot the correlation circle

Questions related to Episodic & Future majorly contributes to this Component. Also it clearly separates high versus normal memory groups irrespective of their age.

The second component mainly distinguishes between Future & Spatial memory questions and these two groups majorly contributes to the second component as seen in the graph. It also tries to separate the older age groups from younger ones

**Summary**

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

1. Component 1: The latent structure of the SAM data as revealed by BADA indicated that the first component characterized good versus poor memory. In other words,when people reported having high or low abilities for one category of memory,they tended to do the same for other categories.
2. Component 2 is mainly composed of Spatial memory variable. People who are generally good at Spatial memory task have a poor future thinking ability as they both are orthogonal and negatively correlated. It also separates observation of older age group from younger ones. Combining the loading with the factor score analysis we can infer that aged people are good at semantic & spatial memory and younger groups are good at episodic and future memory.

## Seventh Analysis : DISTATIS analysis of Judge Sort Data (Orange Juices)

**Loading the Dataset**

## J1 J2 J3 J4 J5 J6 J7 J8 J9 J10 J11 J12 J13 J14 J16 J17 J18  
## UFC\_100% 1 1 5 2 1 2 4 2 3 1 1 1 1 3 1 1 1  
## Unif\_100%\_T 1 2 5 2 1 2 4 2 3 1 3 1 1 3 1 1 1  
## Malee\_100%\_T 2 4 2 3 3 5 3 3 1 3 3 1 3 4 2 3 2  
## Malee\_100%\_V 2 4 2 3 3 1 3 4 1 3 2 2 4 1 2 3 2  
## Tipco\_100%\_T 3 1 1 2 2 5 2 3 1 3 1 2 3 4 4 1 3  
## Tipco\_100%\_V 4 1 1 3 4 5 3 3 1 3 1 1 3 4 5 1 3  
## J19 J20 J22 J23 J24 J25 J26 J27 J28 J29 J30 J31 J32 J33 J34  
## UFC\_100% 2 4 1 1 3 3 2 2 2 3 2 1 1 1 2  
## Unif\_100%\_T 2 4 1 2 4 2 1 2 2 3 3 1 1 1 2  
## Malee\_100%\_T 3 1 3 3 5 1 4 3 3 1 3 2 3 2 1  
## Malee\_100%\_V 3 1 4 4 6 5 4 3 4 1 3 2 3 3 4  
## Tipco\_100%\_T 3 1 3 4 1 1 2 2 4 1 3 3 1 2 1  
## Tipco\_100%\_V 3 1 4 4 1 1 2 2 3 1 2 3 1 2 1  
## J35 J36 J37 J38 J39 J40 J41 J42 J43 J44  
## UFC\_100% 3 4 1 3 1 1 1 3 3 1  
## Unif\_100%\_T 1 4 1 4 1 1 1 2 3 1  
## Malee\_100%\_T 4 5 4 2 3 2 2 4 1 3  
## Malee\_100%\_V 4 5 2 2 3 2 2 3 1 3  
## Tipco\_100%\_T 1 1 4 4 3 3 2 1 2 4  
## Tipco\_100%\_V 1 5 4 4 3 3 2 4 2 6

For more info, try: "summary(OJ)" & "str(OJ)"

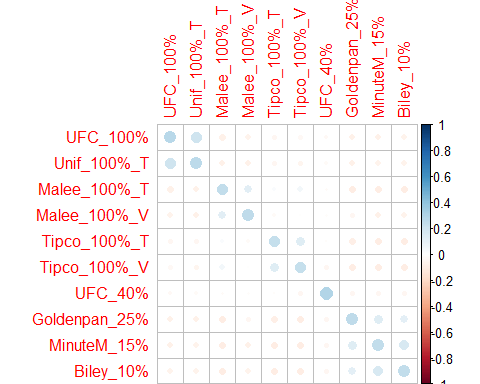
**Computing distance cube & performing DISTATIS Analysis**

We first compute the distance matrix from the input data and then we perform the DISTATIS analysis & compute the bootstrap results using the function "BootFactorScores()". We also view the results using the heatmap of the Splus & C matrix.

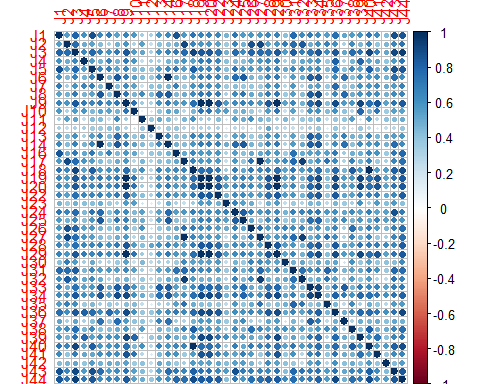
DistanceCube <- DistanceFromSort(OJ) # Computing distance matrix  
  
resDistatis <- distatis(DistanceCube) # Then, we can move on to DISTATIS   
BootF <- BootFactorScores(resDistatis$res4Splus$PartialF) # Performing bootstrap operation

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

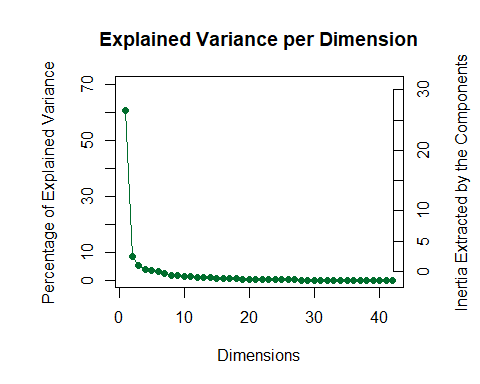
corrplot(resDistatis$res4Splus$Splus) # Analysing the output by visualizing the Splus and C matrix



corrplot(resDistatis$res4Cmat$C)

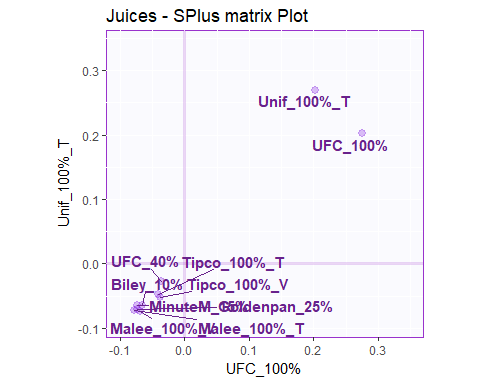
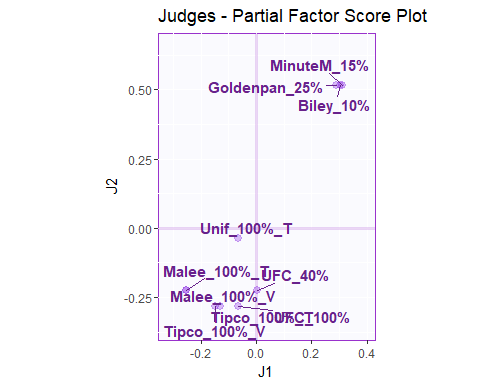
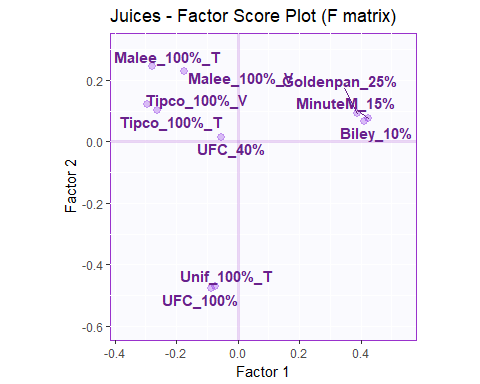
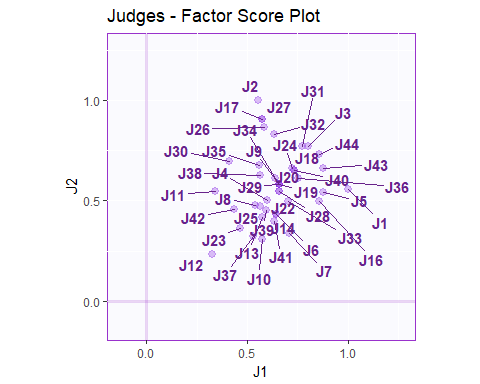


\*\* Scree Plot \*\* A Scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The number of components is min(nrow(DATA), ncol(DATA)). The scree plot is used to determine how many of the components should be interpreted. The significant components are color coded using the bootstrap results.



The first component is explaining roughly 61% of variance and second component explaining 11% of variance in the data. We will be analyzing just the first two component in this cookbook.

**Factor scores**: Factor scores are the coordinates of the observations on the components. The distances between them show which individuals are most similar. Factor scores can be color-coded to help interpret the components.

 As we do not have any design for the judges, we cant infer much with the factor score graphs of the judges. From the juices plot we can infer that more concentrated drinks are separated with less concentrated drinks by factor 1. Judges were able to make distinction between highly concentrated juices with less concentrated ones. Factor 2 separates UNIF\_100% & UFC\_100% from all other drinks indicating judges have rated these drinks quite similarly but different from all others.

**Bootstrap ratios**

Here we are going to plot the bootstrap ratios for Factor 1 & Factor 2 that we get from the BootFactorScores() function.

As seen in the bootstrap plots, Factor 1 is mainly comprised of highly concentrated drinks on one side with less concentrated drinks on the other side. Factor 2 is mainly UNIF\_100% & UFC\_100% on one side with all other drinks on the other side.

**Summary**

When we interpret the factor scores and bootstrap plots together, the DISTATIS Analysis revealed:

1. Factor 1 mainly seperates Concentrated orange juices (100%) with lesser concentrated drinks (40,25,10%). Judges were able to make distinction between highly concentrated juices with less concentrated ones.
2. Factor 2 separates UNIF\_100% & UFC\_100% from all other drinks indicating judges have rated these drinks quite similarly but different from all others.

## Eigth Analysis : Multiple Factor Analysis (MFA) of SAM, OSIQ & BFI Dataset

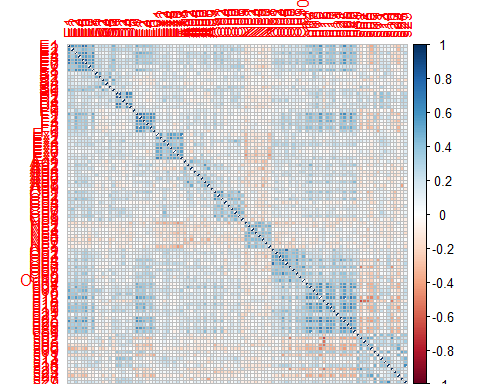
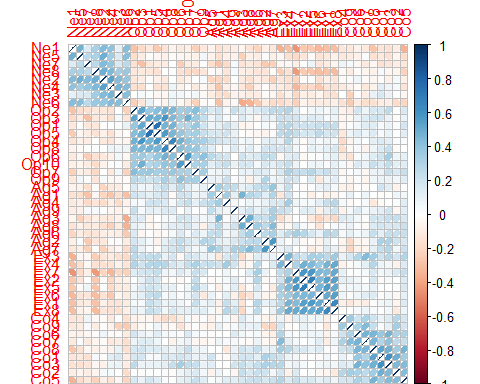
**Loading the SAM Dataset**

## ID memoryGroups Active Sex Age Ex1 Ex2 Ex3 Ex4 Ex5 Ex6 Ex7 Ex8 Ag1 Ag2  
## 1 164 High TRUE 2 36 5 2 4 4 4 3 3 3 2 4  
## 2 220 High TRUE 1 63 2 2 4 4 4 3 2 2 3 4  
## 3 222 High TRUE 1 43 3 2 4 3 2 2 2 5 1 1  
## 4 444 High TRUE 1 34 2 2 2 2 2 3 2 4 2 5  
## 5 586 High TRUE 2 62 4 3 2 3 3 3 4 4 2 4  
## 6 590 High TRUE 2 43 5 4 2 4 4 4 4 5 2 4  
## Ag3 Ag4 Ag5 Ag6 Ag7 Ag8 Ag9 Co1 Co2 Co3 Co4 Co5 Co6 Co7 Co8 Co9 Ne1 Ne2  
## 1 5 3 3 3 3 4 4 5 4 5 2 3 4 4 4 3 3 4  
## 2 4 4 4 4 4 4 5 3 2 4 3 4 4 4 4 4 2 2  
## 3 1 5 1 1 1 1 4 1 1 2 4 1 5 4 1 5 5 5  
## 4 5 3 3 3 4 4 4 3 4 5 2 2 4 4 4 5 5 3  
## 5 5 2 2 3 4 4 5 5 1 5 1 2 5 4 5 2 2 3  
## 6 5 4 4 4 2 2 4 4 4 5 5 2 4 4 5 2 4 4  
## Ne3 Ne4 Ne5 Ne6 Ne7 Ne8 Op1 Op2 Op3 Op4 Op5 Op6 Op7 Op8 Op9 Op10  
## 1 5 5 2 1 3 4 2 5 3 4 2 3 5 5 3 4  
## 2 4 3 2 2 2 2 5 5 5 5 4 4 4 5 4 4  
## 3 1 5 5 5 5 2 5 4 1 3 4 1 1 5 5 2  
## 4 2 4 3 2 2 3 4 5 5 5 3 4 2 5 4 5  
## 5 4 3 4 3 4 3 3 4 4 4 4 5 4 4 5 2  
## 6 4 5 2 4 1 2 4 4 4 2 2 5 4 2 5 4

For more info, try: "summary(BFIData)" & "str(BFIData)"

**Preprocessing the dataset, creating design variables & exploratory analysis**

First we leave out the mystery group to be analyzed later on as supplementary elements. Next we separate the questionaire variables and design variables (Memory Groups, Gender, Age) while keep the row namas as participant IDs. We also do a correlation plot of the analysis variables to get some insights into the data. We then combine the three datasets together, plot their correlation plot and create new design variables.



The correlation plot indicates that Neuroticism is negatively correlated with all other variables of memory & behavior. Also spatial memory component of OSIQ is negatively correlated with other variables.

**Using mpMFA function of MExPosition library for MFA**

We use the function mpMFA in MExPosition library for this analysis and store the results in resMFA & resMFA1 depending on the two design matrix we created in the previous step.

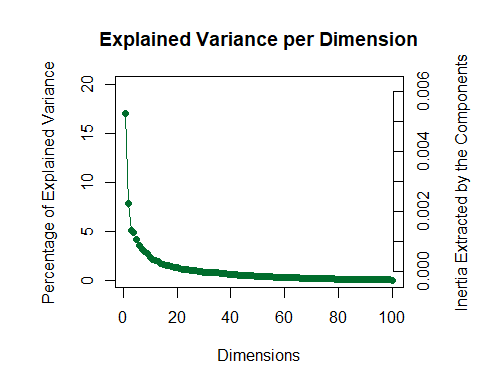
resMFA <- mpMFA(combinedData, column.design = colDes, make.columndesign.nominal = TRUE, graphs = TRUE, DESIGN = Des.mat1, make.design.nominal = FALSE)

## [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"

resMFA1 <- mpMFA(combinedData, column.design = colDesNew, make.columndesign.nominal = TRUE, graphs = TRUE, DESIGN = Des.mat1, make.design.nominal = FALSE)

## [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"

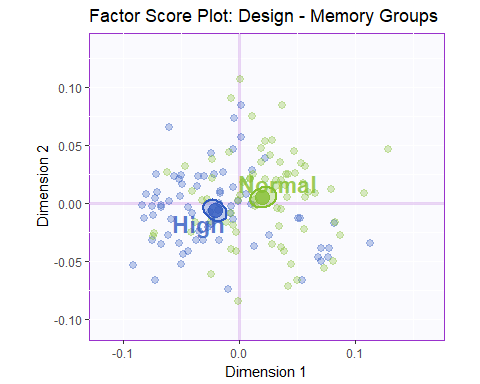
\*\* Scree Plot \*\* A Scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The number of components is min(nrow(DATA), ncol(DATA)). The scree plot is used to determine how many of the components should be interpreted. The significant components are color coded using the bootstrap results.

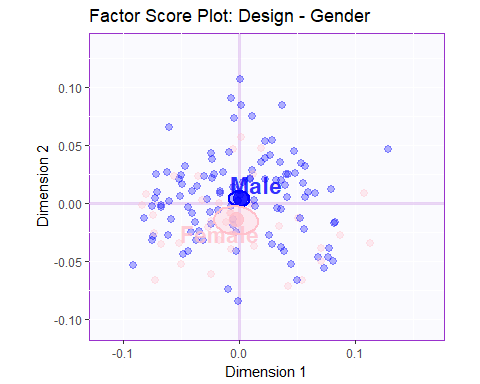


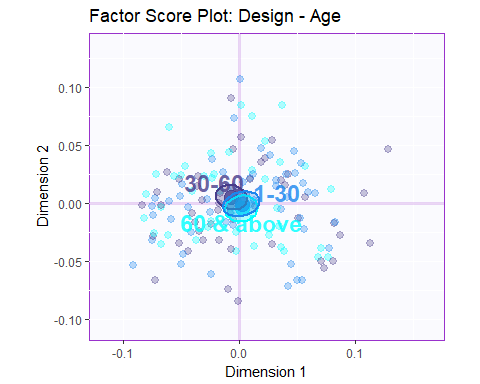
The first component explains roughly 17% of variance and second component explaining 8% of variance in the data. We will be analyzing just the first two component in this cookbook.

**Factor scores**:

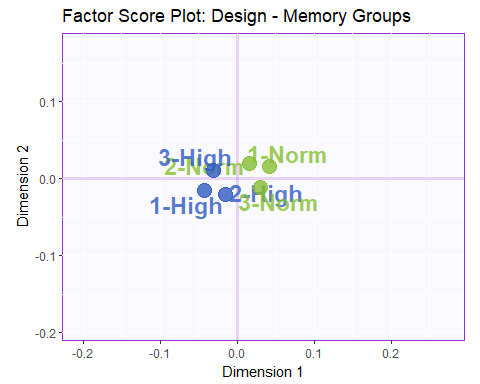
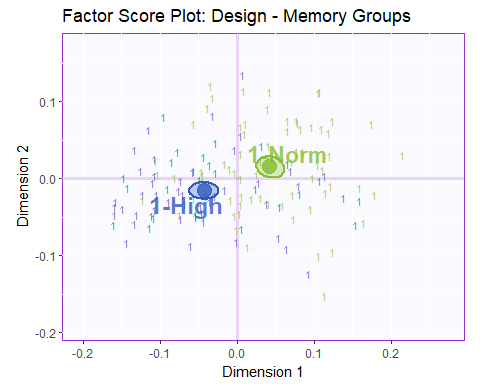
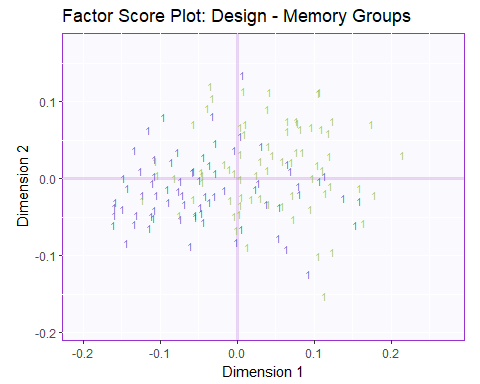
Factor scores are the coordinates of the observations on the components. The distances between them show which individuals are most similar. Factor scores can be color-coded to help interpret the components.



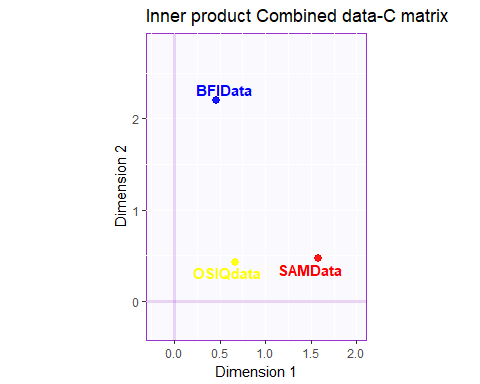




In the above factor score plots, we see that Component 1 mainly distinguish between subjects in "High" & "Normal" Memory Groups. Gender & Age have no effect on Memory as per the Factor Score plots for Gender & Age. There isn't anything that component 2 is separating and things are almost evenly balanced on Component 2.

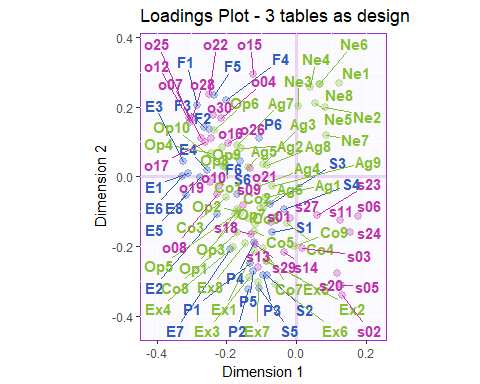
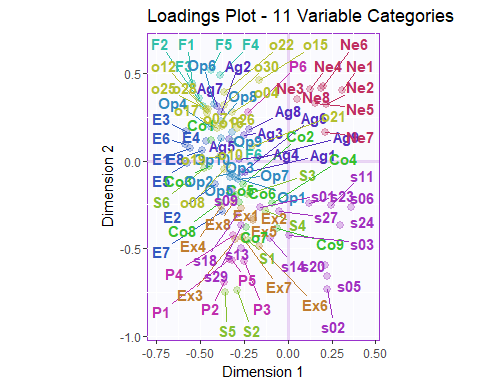


Let's plot the Inner product C matrix of combined data to find whihc table is similar to the other one and the relative standing of each of the data tables.

 It is seen that the two memory tables "SAM & OSIQ" are close to each other than behaviour BFI dataset.

**Loadings**

Loadings describe the similarity (angular distance) between the variables. Loadings show how the input variables relate to each other. Loadings also show which variables are important for (which components load on) a certain component.



**Summary** When we interpret the factor scores and loadings together, the MFA revealed:

Component 1 mainly distinguishes High versus Normal Memory Groups. It also separates Neuroticism questions from all others in the Combined Dataset. Also Inner Product Combined Data, we see that component 1 separates Memory versus behavior questions.

Component 2 mainly distinguishes between spatial and object memory and also future versus spatial memory components.