All Countries Environmental Data

## Method: BADA

## Dataset

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

str(country\_env\_df)

## 'data.frame': 137 obs. of 29 variables:  
## $ Country : Factor w/ 137 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Happiness\_Rank : Ord.factor w/ 3 levels "VH"<"H"<"U": 3 2 2 3 1 3 1 1 2 2 ...  
## $ accessibility\_to\_cities: num 317.7 73.8 1212.8 378.2 209.2 ...  
## $ elevation : num 1832 652 557 1061 683 ...  
## $ aspect : num 201 192 185 174 145 ...  
## $ slope : num 1.516 1.89 0.171 0.193 0.624 ...  
## $ cropland\_cover : num 9.51 23.35 3.69 2.79 21.96 ...  
## $ tree\_canopy\_cover : num 0.375 12.805 0.177 19.87 8.834 ...  
## $ isothermality : num 35.9 33.2 40.3 64.3 49.9 ...  
## $ rain\_coldestQuart : num 128.72 392.51 25.29 8.05 79.09 ...  
## $ rain\_driestMonth : num 1.722 40.088 0.935 0.26 17.183 ...  
## $ rain\_driestQuart : num 8.3 138.15 6.09 4.43 60.49 ...  
## $ rain\_mean\_annual : num 311.3 1151.1 79.5 1023.4 539.9 ...  
## $ rain\_seasonailty : num 91.6 38.5 67.1 91.5 48.3 ...  
## $ rain\_warmestQuart : num 12.69 138.33 9.51 318.54 183.14 ...  
## $ rain\_wettestMonth : num 67.8 159 13.4 202.2 79.2 ...  
## $ rain\_wettestQuart : num 175.8 435.9 33.3 524.3 211.7 ...  
## $ temp\_annual\_range : num 40.3 27.1 36.5 21.5 26.8 ...  
## $ temp\_coldestQuart : num -0.261 3.58 13.152 18.794 8.024 ...  
## $ temp\_diurnal\_range : num 14.72 9.11 14.87 13.85 13.46 ...  
## $ temp\_driestQuart : num 21.1 19.6 26.9 18.9 11.1 ...  
## $ temp\_max\_warmestMonth : num 32 26.3 41.5 31 28.2 ...  
## $ temp\_mean\_annual : num 11.5 11.5 23 21.6 14.2 ...  
## $ temp\_min\_coldestMonth : num -8.312 -0.806 5.058 9.549 1.443 ...  
## $ temp\_seasonality : num 88.2 62.7 75.1 18.5 47.6 ...  
## $ temp\_warmestQuart : num 22.7 19.6 32.5 23.3 20.2 ...  
## $ temp\_wettestQuart : num 3.95 5.27 20.81 22.76 16.48 ...  
## $ wind : num 3.43 2.47 4.03 2.16 4.27 ...  
## $ cloudiness : num 114.2 181.1 90.7 187.5 159 ...

* Research Question

How do the 137 countries differ on these variables?

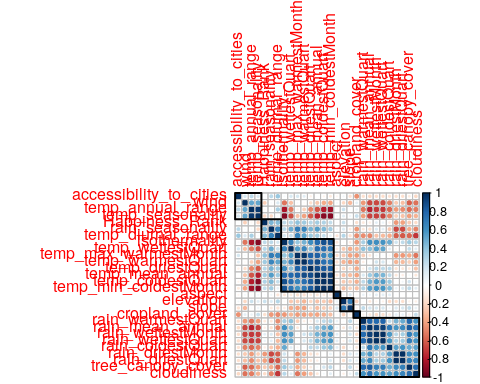
## Analysis

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

### Correlation Plot

Visually analyze multicollinearity in the system.

corr\_result = cor(country\_env\_df\_for\_corr)  
corrplot(corr\_result,order = 'hclust', addrect = 7)



### Identify Latent Components

#### PCA (with Inference)

country\_env\_pca <- tepBADA(DATA = country\_env\_df\_for\_pca, center = TRUE, scale = 'SS1', DESIGN = country\_env\_df$Happiness\_Rank, graphs = FALSE)  
country\_env\_pca\_inf <- tepBADA.inference.battery(DATA = country\_env\_df\_for\_pca, center = TRUE, scale = 'SS1', DESIGN = country\_env\_df$Happiness\_Rank, graphs = FALSE)

## [1] "It is estimated that your iterations will take 0.04 minutes."  
## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please take note of the progress bar."  
## ===========================================================================

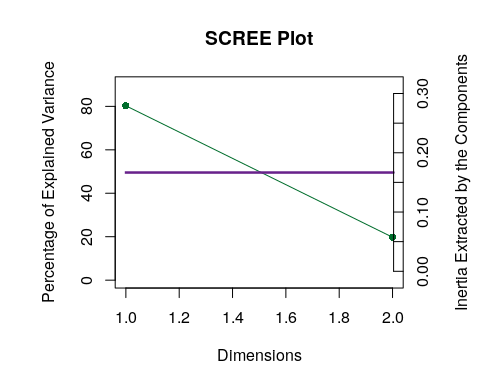
Now we have Factor scores and Loadings. \* Factor Scores are the new Data points w.r.t. new Components achieved with help of SVD. \* Loadings represent correlation between variables w.r.t the choosen Components. Can be interpreted in 3 ways + As slices of inertia of the contribution data table w.r.t.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

#### Scree Plot

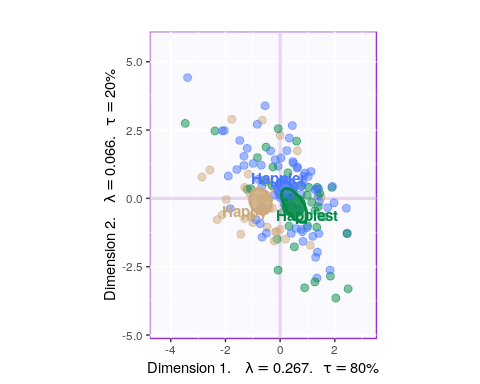
Gives amount of information explained by corresponding component. Gives an intuition to decide which components best represent data in order to answer the research question.

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

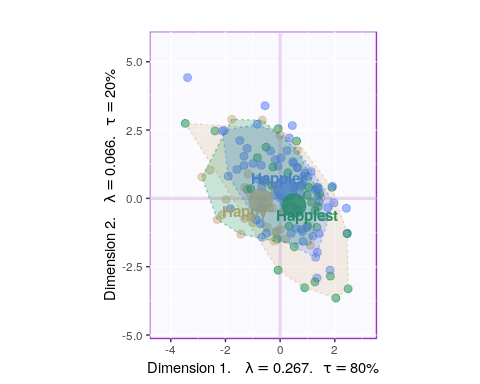
PTCA4CATA::PlotScree(ev = country\_env\_pca$TExPosition.Data$eigs,  
 #p.ev = country\_env\_pca\_inf$Inference.Data$components$p.vals,  
 title = 'SCREE Plot',  
 plotKaiser = TRUE  
)



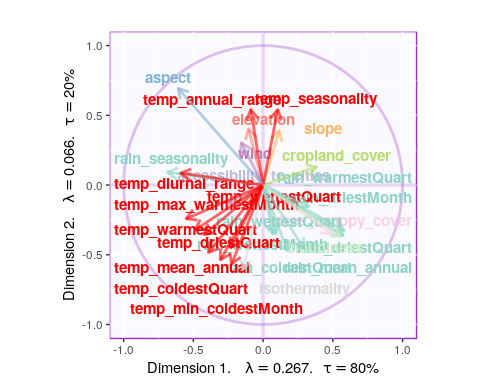
#### Factor Scores



* With Tolerance Interval



#### Loadings

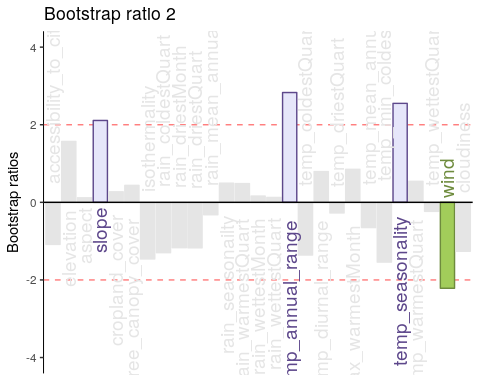
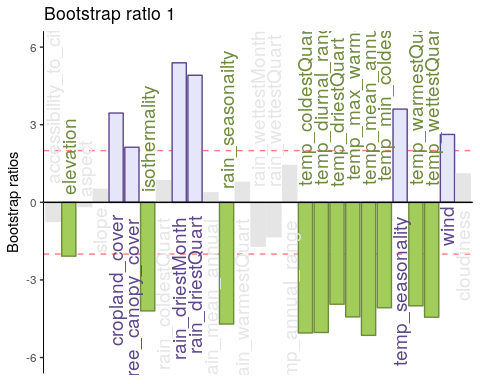


* Component 1:
  + Rows: Normal & Happy
  + Columns: Cloudiness & Rain vs Cropland, Aspect, Elevation
  + Interpret: People in countries with more Cloudiness, Trees and Rain tends to be happier.
* Component 7:
  + Rows: Happy & Unhappy
  + Columns: Temp and Rain vs Accessibility and Cropland
  + Interpret: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

#### Most Contributing Variables

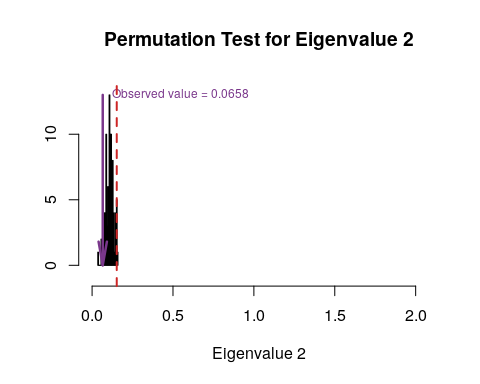
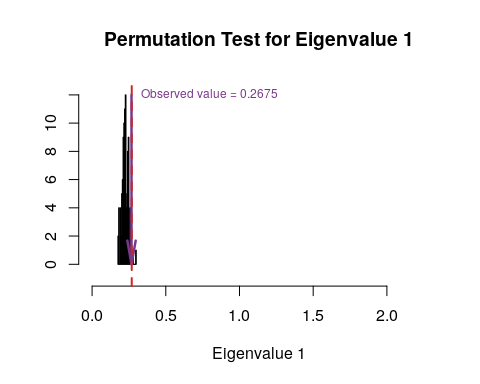
* With Bootstrap Ratio

BR <- country\_env\_pca\_inf$Inference.Data$boot.data$fj.boot.data$tests$boot.ratios  
  
for (i in c(1, 2)) {  
 laDim = i  
 ba001.BR1 <- PrettyBarPlot2(BR[,laDim],  
 threshold = 2,  
 font.size = 5,  
 #color4bar = gplots::col2hex(col4J), # we need hex code  
 main = paste0('Bootstrap ratio ',laDim),  
 ylab = 'Bootstrap ratios'  
 #ylim = c(1.2\*min(BR[,laDim]), 1.2\*max(BR[,laDim]))  
 )  
 print(ba001.BR1)  
}



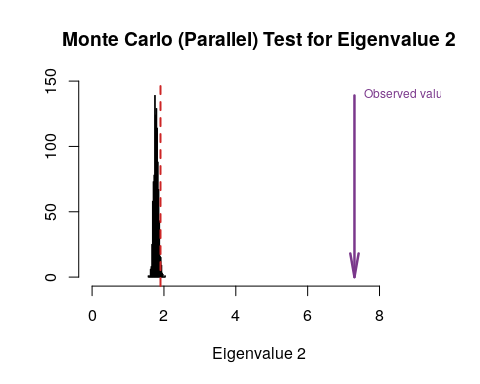
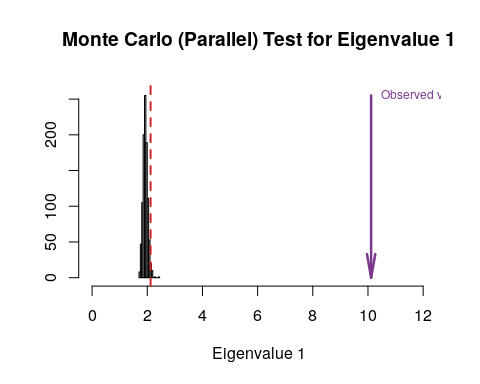
#### Permutation Test

for (i in c(1, 2)) {  
zeDim = i  
pH1 <- prettyHist(  
 distribution = country\_env\_pca\_inf$Inference.Data$components$eigs.perm[,zeDim],   
 observed = country\_env\_pca\_inf$Fixed.Data$TExPosition.Data$eigs[zeDim],   
 xlim = c(0, country\_env\_pca\_inf$Fixed.Data$TExPosition.Data$eigs[zeDim]+2), # needs to be set by hand  
 breaks = 20,  
 border = "black",   
 main = paste0("Permutation Test for Eigenvalue ",zeDim),  
 xlab = paste0("Eigenvalue ",zeDim),   
 ylab = "",   
 counts = FALSE,   
 cutoffs = c( 0.975))  
}



#### Parallet Test

country\_env\_pca\_mc <- data4PCCAR::monteCarlo.eigen(X = country\_env\_df\_for\_pca, nIter = 1000)  
for (i in c(1, 2)) {  
 zeDim = i  
 pH1.p <- prettyHist(country\_env\_pca\_mc$rand.eigs[,zeDim],   
 observed = country\_env\_pca\_mc$fixed.eigs[zeDim],   
 xlim = c(0, country\_env\_pca\_mc$fixed.eigs[zeDim]+2), # needs to set by hand  
 breaks = 20,  
 border = "black",   
 main = paste0("Monte Carlo (Parallel) Test for Eigenvalue ",zeDim),  
 xlab = paste0("Eigenvalue ",zeDim),   
 ylab = "",   
 counts = FALSE,   
 cutoffs = c( 0.975))  
}



#### Bootstrap Test

#country\_env\_pca\_br <- PTCA4CATA::Boot4Mean(country\_env\_pca$ExPosition.Data$fi, design = country\_env\_df$Happiness\_Rank, niter=100, suppressProgressBar = FALSE)  
country\_env\_pca\_bs <- data4PCCAR::boot.eigen(X = country\_env\_df\_for\_pca, nIter = 1000)  
  
for (i in c(1, 2)) {  
 zeDim = i  
 prettyHist(country\_env\_pca\_bs$boot.eigs[,zeDim],   
 observed = country\_env\_pca\_bs$fixed.eigs[zeDim],   
 xlim = c(0, country\_env\_pca\_bs$fixed.eigs[zeDim]+2), # needs to set by hand  
 breaks = 20,  
 border = "black",   
 main = paste0("Bootstrapped distribution for Eigenvalue ",zeDim),  
 xlab = paste0("Eigenvalue ",zeDim),   
 ylab = "",   
 counts = FALSE,   
 cutoffs = c(0.025, 0.975))  
}

