S&DS 363 Problem Set 1: Principal Component Analysis

Franklin Wu and Teresa Nguyen

2025-02-03

## Human Development Index Principal Component Analysis

*Principal Component Analysis (PCA)* is a statistical method used for dimensionality reduction, simplifying large data sets into smaller, more manageable sets. In this homework assignment, we will apply PCA to the 2021 Human Development Index and explain its effectiveness in our analysis.

### Import Functions from JDRS

#This chunk defines several helpful functions  
  
#n is the number of observations in the dataset  
#p is the number of variables in the dataset  
  
parallel<-function(n,p){  
   
 if (n > 1000 || p > 100) {  
 print ("Sorry, this only works for n<1000 and p<100")  
 stop()  
 }  
   
 coefs <- matrix(  
 c(0.0316, 0.7611, -0.0979, -0.3138, 0.9794, -.2059, .1226, 0, 0.1162,   
 0.8613, -0.1122, -0.9281, -0.3781, 0.0461, 0.0040, 1.0578, 0.1835,   
 0.9436, -0.1237, -1.4173, -0.3306, 0.0424, .0003, 1.0805 , 0.2578,   
 1.0636, -0.1388, -1.9976, -0.2795, 0.0364, -.0003, 1.0714, 0.3171,   
 1.1370, -0.1494, -2.4200, -0.2670, 0.0360, -.0024, 1.08994, 0.3809,   
 1.2213, -0.1619, -2.8644, -0.2632, 0.0368, -.0040, 1.1039, 0.4492,   
 1.3111, -0.1751, -3.3392, -0.2580, 0.0360, -.0039, 1.1173, 0.5309,   
 1.4265, -0.1925, -3.8950, -0.2544, 0.0373, -.0064, 1.1421, 0.5734,   
 1.4818, -0.1986, -4.2420, -0.2111, 0.0329, -.0079, 1.1229, 0.6460,   
 1.5802, -0.2134, -4.7384, -0.1964, 0.0310, -.0083, 1.1320),ncol=8, byrow=TRUE)  
   
 calclim <- p  
 if (p > 10) calclim <- 10  
 coefsred <- coefs[1:calclim, ]  
 temp <- c(p:1)  
 #stick <- sort(cumsum(1/temp), decreasing=TRUE)[1:calclim]  
 multipliers <- matrix(c(log(n),log(p),log(n)\*log(p),1), nrow=1)  
 longman <- exp(multipliers%\*%t(coefs[,1:4]))[1:calclim]  
 allen <- rep(NA, calclim)  
 leig0 <- 0  
 newlim <- calclim  
 if (calclim+2 < p) newlim <-newlim+2  
 for (i in 1:(newlim-2)){  
 leig1 <- coefsred[i,5:8]%\*%matrix(c(1,log(n-1),log((p-i-1)\*(p-i+2)/2), leig0))  
 leig0 <- leig1  
 allen[i] <- exp(leig1)  
 }  
 pcompnum <- c(1:calclim)  
 #data.frame(cbind(pcompnum,stick,longman,allen))  
 data.frame(cbind(pcompnum,longman,allen))   
}  
  
#########  
#this function makes a nice plot if given the input from a PCA analysis  
#created by prcomp()  
##  
#arguments are  
# n=number of observations  
  
parallelplot <- function(comp){  
 if (dim(comp$x)[1] > 1000 || length(comp$sdev) > 100) {  
 print ("Sorry, this only works for n < 1000 and p < 100")  
 stop()  
 }  
 #if (round(length(comp$sdev)) < round(sum(comp$sdev^2))) {  
 # print ("Sorry, this only works for analyses using the correlation matrix")  
 # stop()  
 # }  
   
 parallelanal <- parallel(dim(comp$x)[1], length(comp$sdev))  
 print(parallelanal)  
 calclim <- min(10, length(comp$sdev))  
 eigenvalues <- (comp$sdev^2)[1:calclim]  
 limits <- as.matrix(parallelanal[,2:3])  
 limits <- limits[complete.cases(limits)]  
 ymax <- range(c(eigenvalues),limits)  
 plot(parallelanal$pcompnum, eigenvalues, xlab="Principal Component Number",  
 ylim=c(ymax), ylab="Eigenvalues and Thresholds",  
 main="Scree Plot with Parallel Analysis Limits",type="b",pch=15,lwd=2, col="red")  
 #lines(parallelanal$pcompnum,parallelanal[,2], type="b",col="red",pch=16,lwd=2)  
 lines(parallelanal$pcompnum,parallelanal[,2], type="b",col="green",pch=17,lwd=2)  
 lines(parallelanal$pcompnum,parallelanal[,3], type="b",col="blue",pch=18,lwd=2)  
 #legend((calclim/2),ymax[2],legend=c("Eigenvalues","Stick Method","Longman Method",  
 # "Allen Method"), pch=c(15:18), col=c("black","red","green","blue"),lwd=2)  
 legend((calclim/2), ymax[2], legend=c("Eigenvalues","Longman Method","Allen Method"),   
 pch = c(16:18), col= c("red","green","blue"), lwd=2)  
}  
  
  
#make score plot with confidence ellipse.  
#arguments are output from prcomp, vector with components for plotting   
# (usually c(1,2) or c(1,3)  
#and a vector of names for the points  
  
ciscoreplot<-function(x, comps, namevec){  
 y1<-sqrt(5.99\*(x$sdev[comps[1]]^2))  
 ymod<-y1-y1%%.05  
 y1vec<-c(-y1,seq(-ymod,ymod,by=0.05),y1)  
 y2vecpos<-sqrt((5.99-(y1vec^2)/x$sdev[comps[1]]^2)\*x$sdev[comps[2]]^2)  
 y2vecneg<--sqrt((5.99-(y1vec^2)/x$sdev[comps[1]]^2)\*x$sdev[comps[2]]^2)  
 y2vecpos[1]<-0  
 y2vecneg[1]<-0  
 y2vecpos[length(y2vecpos)]<-0  
 y2vecneg[length(y2vecneg)]<-0  
   
 plot(x$x[,comps[1]],x$x[,comps[2]],   
 pch = 19,   
 cex = 1.2,  
 xlim = c(min(y1vec, x$x[, comps[1]]), max(y1vec, x$x[, comps[1]])),  
 ylim = c(min(y2vecneg, x$x[, comps[2]]), max(y2vecpos, x$x[, comps[2]])),  
 main = "PC Score Plot with 95% CI Ellipse",   
 xlab = paste("Scores for PC", comps[1], sep = " "),   
 ylab = paste("Scores for PC", comps[2], sep = " "))  
   
 lines(y1vec,y2vecpos,col="Red",lwd=2)  
 lines(y1vec,y2vecneg,col="Red",lwd=2)  
 outliers<-((x$x[,comps[1]]^2)/(x$sdev[comps[1]]^2)+(x$x[,comps[2]]^2)/(x$sdev[comps[2]]^2))>5.99  
   
 points(x$x[outliers, comps[1]], x$x[outliers, comps[2]], pch = 19, cex = 1.2, col = "Blue")  
   
 text(x$x[outliers, comps[1]],x$x[outliers, comps[2]], col = "Blue", lab = namevec[outliers])  
}

### Data Cleaning

Read in Data, choose the columns that best fit our needs, and rename the columns.

data <- read.csv("../../Documents/sds-363/Human Development Index - Full.csv")  
  
# Based on going through the Kaggle dataset and choosing the most relevant features  
data <- data[, c(  
 "Country",  
 "UNDP.Developing.Regions",  
 "HDI.Rank..2021.",  
 "Human.Development.Index..2021.",  
 "Life.Expectancy.at.Birth..2021.",  
 "Mean.Years.of.Schooling..2021.",  
 "Gross.National.Income.Per.Capita..2021.",  
 "Gender.Development.Index..2021.",  
 "Coefficient.of.human.inequality..2021.",  
 "Overall.loss......2021.",  
 "Gender.Inequality.Index..2021.",  
 "Maternal.Mortality.Ratio..deaths.per.100.000.live.births...2021.",  
 "Adolescent.Birth.Rate..births.per.1.000.women.ages.15.19...2021.",  
 "Labour.force.participation.rate..female....ages.15.and.older...2021.",  
 "Labour.force.participation.rate..male....ages.15.and.older...2021.",  
 "Carbon.dioxide.emissions.per.capita..production...tonnes...2021."  
)]  
  
# rename for better readability  
colnames(data) <- c(  
 "Country",  
 "Region",  
 "HDI\_Rank\_2021",  
 "HDI\_2021",  
 "Life\_Expectancy\_2021",  
 "Mean\_Years\_Schooling\_2021",  
 "GNI\_Per\_Capita\_2021",  
 "Gender\_Dev\_Index\_2021",  
 "Human\_Inequality\_Coeff\_2021",  
 "Overall\_Loss\_2021",  
 "Gender\_Inequality\_Index\_2021",  
 "Maternal\_Mortality\_Rate\_2021",  
 "Adolescent\_Birth\_Rate\_2021",  
 "Female\_Labour\_Force\_2021",  
 "Male\_Labour\_Force\_2021",  
 "CO2\_Emissions\_percapita\_2021"  
)  
  
# removes all rows in the data object that contain any missing (NA) values  
# our data frame should only include complete cases   
data <- data[complete.cases(data), ]

## Initial Analysis with Scatterplots and Heatmaps

We begin our initial analysis of the data using scatterplots and heatmaps to assess linearity, correlations, and whether the data appears suitable for PCA at first glance. Our goal is to use PCA to reduce redundant variance and capture shared variance. Additionally, we can check for outliers and non-linear patterns.

library(corrplot)

## corrplot 0.95 loaded

library(PerformanceAnalytics)

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':  
##   
## legend

corrplot.mixed(  
 cor(data[, -c(1:3)]),  
 lower.col = "black",  
 upper = "ellipse",  
 tl.col = "black",  
 number.cex = 0.3,  
 order = "hclust",  
 tl.pos = "lt",  
 tl.cex = 0.5  
)

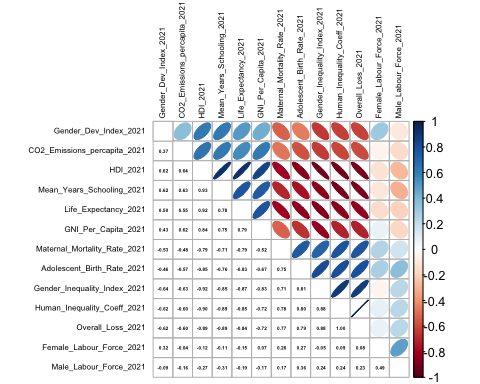
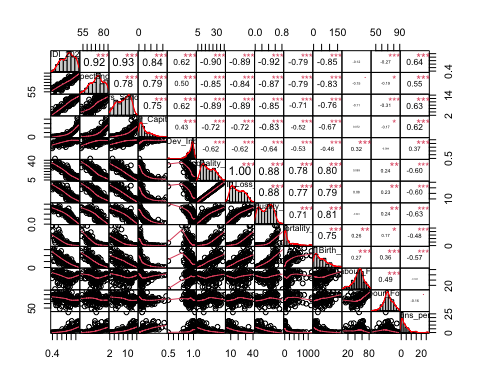


chart.Correlation(data[, -c(1:3)])



*Based on our plots showing the correlation between our variables, there appear to be decently strong positive and negative correlations between them, suggesting that PCA may be an effective strategy to utilize since we can reduce dimensionality without losing too much information. For example, variables such as HDI, life expectancy, and mean years of schooling have strong positive correlations that are over 0.75, making PCA suitable for dimensionality reduction, as shown by the strong correlations with certain variables, which could be redundant. There are also many other variables that are correlated with one another, which is a sign that PCA can capture the variance of the features. For instance, gender and human inequality indices show strong negative correlations with the mean years of schooling and the human development index. Intuitively, these relationships imply that these variables may be interconnected, making them suitable for PCA.*

*Moreover, the relationships exhibited between variables, for the most part, appear to be linear. However, looking at the scatter plots between our variables, there does seem to be some non-linearity (for instance, the relationship between GNI per capita and mortality rate), which may suggest the use of transformations.*

### Multivariate Normality Check and Transformations

We check for multivariate normality using the Mahalanobis distance and a Chi-square Q-Q plot. If the data is multivariate normal, the principal components will be uncorrelated and independent, making it easier to interpret each principal component. The Mahalanobis distance measures how far a data point is from the center of the data distribution and is primarily used to assess whether a data point is an outlier. A large Mahalanobis distance suggests the presence of an outlier. The Chi-square Q-Q plot provides a visual representation of the Mahalanobis distances.

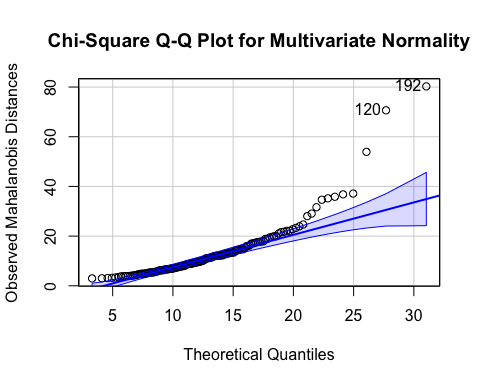
library(MASS)  
library(car)

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3

data1 <- as.matrix(data[, -c(1:3)])   
  
center <- colMeans(data1) # Mean vector  
cov\_matrix <- cov(data1) # Covariance matrix  
mahalanobis\_dist <- mahalanobis(data1, center, cov\_matrix)  
  
theoretical\_quantiles <- qchisq(ppoints(length(mahalanobis\_dist)), df = ncol(data1))  
  
qqPlot(mahalanobis\_dist, distribution = "chisq", df = ncol(data1),  
 main = "Chi-Square Q-Q Plot for Multivariate Normality",  
 xlab = "Theoretical Quantiles",  
 ylab = "Observed Mahalanobis Distances")

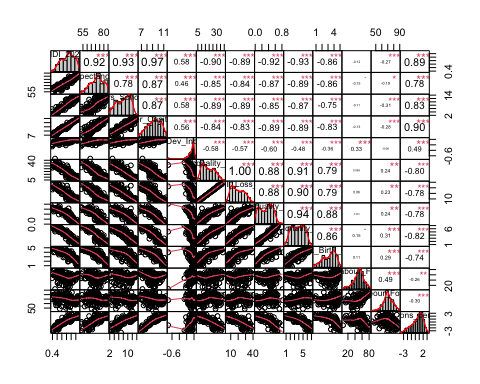


## 192 120   
## 147 92

# MAKE chi-square quantile plot  
#library(heplots)  
#cqplot(data[, -c(1:3)], main = "World Bank Data")

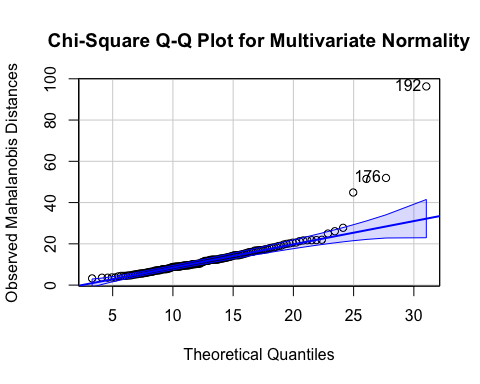
*Based on the Chi-square quantile plot, a large number of observations fall outside the boundary expected under multivariate normality and follow a non-linear pattern, with the latter half curving upwards. The non-linear tail suggests that the data may not be multivariate normal, which could impact our PCA results. To address this issue, we decided to apply a log transformation to specific columns (as shown below).*

# Apply log transformation to specific columns   
data[c(  
 "GNI\_Per\_Capita\_2021",   
 "Gender\_Dev\_Index\_2021",   
 "Maternal\_Mortality\_Rate\_2021",   
 "Adolescent\_Birth\_Rate\_2021",   
 "CO2\_Emissions\_percapita\_2021"  
)] <- lapply(  
 data[c(  
 "GNI\_Per\_Capita\_2021",   
 "Gender\_Dev\_Index\_2021",   
 "Maternal\_Mortality\_Rate\_2021",   
 "Adolescent\_Birth\_Rate\_2021",   
 "CO2\_Emissions\_percapita\_2021"  
 )],   
 log  
)  
  
chart.Correlation(data[, -c(1:3)])



*We used log transformations of the following variables: GNI Per Capita, Gender Dev Index, Maternal Mortality Rate, Adolescent Birth Rate, and CO2 Emissions per capita. These variables were not distributed normally and thus has non-linear relationships with our other variables, potentially impacting the effectiveness of PCA. Following the transformations of these variables, the distribution (using a histogram) appear to be normal and all of our relationships between the variables are mostly linear.*

data1 <- as.matrix(data[, -c(1:3)])   
  
center <- colMeans(data1) # Mean vector  
cov\_matrix <- cov(data1) # Covariance matrix  
mahalanobis\_dist <- mahalanobis(data1, center, cov\_matrix)  
  
theoretical\_quantiles <- qchisq(ppoints(length(mahalanobis\_dist)), df = ncol(data1))  
  
qqPlot(mahalanobis\_dist, distribution = "chisq", df = ncol(data1),  
 main = "Chi-Square Q-Q Plot for Multivariate Normality",  
 xlab = "Theoretical Quantiles",  
 ylab = "Observed Mahalanobis Distances")



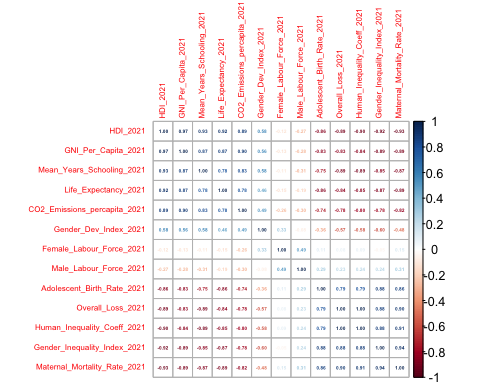
## 192 176   
## 147 136

*After making log adjustments of our variables with distributions that were not normal, our normal quantile plot is contained more within the bounds. There are still a few observations that are outliers and curve upwards, but our distribution appears to be much more similar to a multi-variate normal dsitribution than before we made the transformations. Based on the chi-square quantile plot, most of the data points fall within the diagonal line, showing that the squared Mahalanobis distances are consistent with the chi-square distribution, which means that the data is mostly mutlivariate normal. However, there are a couple of data points that are outliers such as at 176 and 192.*

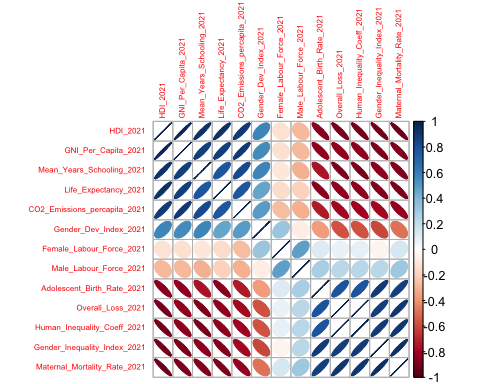
### Correlation Analysis and PCA Suitability

For our next step, we created correlation plots to determine correlation coefficients and correlations. We order it based on first principal component to identify the correlated groups and analyze whether PCA is a suitable method.

corrplot(cor(data[ ,-c(1:3)]), method="number", order="FPC", tl.cex = .5,   
 number.cex = .3)



corrplot(cor(data[ ,-c(1:3)]),method = "ellipse", order="FPC", tl.cex = .5)



dim(data)

## [1] 150 16

*As mentioned in question 1, there appears to be fairly strong correlations (both positive and negative) between our variables. Based on the corrplot, some variables with very strong positive correlations include overall loss and gender inequality (r = 0.88), adolescent birth rate and human inequality coefficient (r = 0.88), and gender inequality and adolescent birth rate (r = 0.81). Likewise, there are some decently strong negative correlation between our variables such as mean years of education and human inequality (r = -0.89), gender inequality and life expectancy (r = -0.85) and human inequality and life expectancy (-0.85). This suggests that PCA should work relatively well on our given data since strong correlations implies that a lot of the variance is being described by multiple variables and as such, we can capture a lot of the variability by using a set of principle components (thus reducing our dimensions overall). Our dataset is composed of 150 complete observations and 13 variables (there are 16 total variables but we excluded 3 of them from our analysis since they are categorical). Relative to the number of variables our sample size is sufficiently large enough (there are over 10 times the amount of complete oberservations as variables) and we can use a rule of them to ensure that our data set is large enough for PCA.*

summary.PCA.JDRS <- function(x){  
 sum\_JDRS <- summary(x)$importance  
 sum\_JDRS[1, ] <- sum\_JDRS[1, ]^2  
 attr(sum\_JDRS, "dimnames")[[1]][1] <- "Eigenvals (Variance)"  
 sum\_JDRS  
}  
  
pc1 <- prcomp(data[, -c(1:3)], scale. = TRUE)  
  
round(summary.PCA.JDRS(pc1),2)

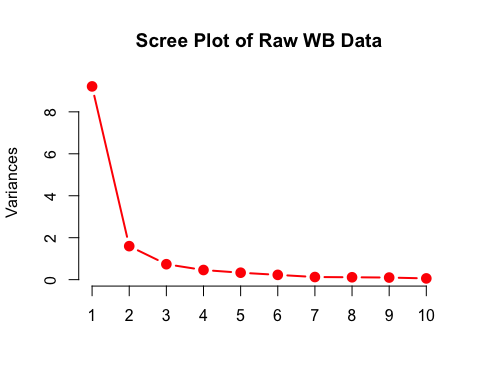
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11  
## Eigenvals (Variance) 9.21 1.60 0.74 0.46 0.33 0.23 0.13 0.11 0.10 0.06 0.03  
## Proportion of Variance 0.71 0.12 0.06 0.04 0.03 0.02 0.01 0.01 0.01 0.00 0.00  
## Cumulative Proportion 0.71 0.83 0.89 0.92 0.95 0.97 0.98 0.98 0.99 1.00 1.00  
## PC12 PC13  
## Eigenvals (Variance) 0.01 0  
## Proportion of Variance 0.00 0  
## Cumulative Proportion 1.00 1

*Using the total variance explained by a given number of principal components, we can use a total explained cutoff of 80%. Under this threshold, we would need two principal components to capture at least 80% of the total variance. In this case, the first principal component explains 71% of the total variance and 12% for the second principal component for a total of 83%. Using the eigenvalues > 1 rule, we would retain only the first two principal components, with eigenvalues of 9.21 and 1.60, respectively.*

### Scree plots

We will now apply scree plots to display the eigenvalues of each principal component. We will also apply parallel analysis to decide how many components to retain.

screeplot(pc1, type = "lines", col = "red", lwd = 2, pch = 19, cex = 1.2,   
 main = "Scree Plot of Raw WB Data")

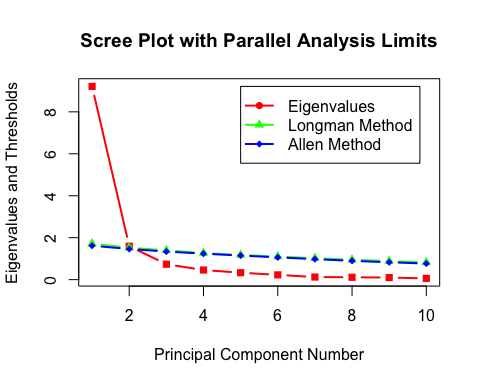


*Using the scree plot elbow method, there is an apparent elbow at the first principal component transitioning to the second principal component, which indicates that PC1 explains the majority of the variance in the data and variance in PC2 decreases significantly. Although the curve flattens after PC2, there appears to be a subtle change in slope near PC3, but it is not as distinct.*

*Using the second principle component as the cutoff, we would only retain one principle component according to the scree plot.*

source("http://reuningscherer.net/multivariate/r/parallel.R.txt")  
parallelplot(pc1)

## pcompnum longman allen  
## 1 1 1.7134446 1.6187237  
## 2 2 1.5239862 1.4604669  
## 3 3 1.3946482 1.3391165  
## 4 4 1.2691669 1.2390785  
## 5 5 1.1795432 1.1482711  
## 6 6 1.1007378 1.0622446  
## 7 7 1.0243853 0.9781679  
## 8 8 0.9512595 0.8959654  
## 9 9 0.8864132 0.8292922  
## 10 10 0.8261170 0.7660355



*Since parallel analysis assumes that the data is suitable for PCA, we can use it as a valid method to determine the number of principle components to keep. After transforming the data, we have linear relationships between all of our variables and also normality (although this isn’t a necessary condition for PCA). Based on parallel analysis, we should retain 2 principle components, since the cutoff is right below the second component.*

### Interpretation of PCA

# principal component loadings to indicate how much of each original var.  
# contributes to each principal component  
round(pc1$rotation,2)

## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8  
## HDI\_2021 0.32 -0.01 -0.04 -0.06 0.16 0.06 0.34 0.08  
## Life\_Expectancy\_2021 0.30 0.00 -0.25 0.08 0.11 -0.50 0.32 0.47  
## Mean\_Years\_Schooling\_2021 0.31 -0.01 0.10 -0.15 -0.18 0.47 0.53 -0.36  
## GNI\_Per\_Capita\_2021 0.31 0.01 -0.04 -0.09 0.39 0.14 0.20 0.10  
## Gender\_Dev\_Index\_2021 0.20 -0.39 0.62 -0.34 0.13 -0.46 -0.06 -0.23  
## Human\_Inequality\_Coeff\_2021 -0.31 0.05 0.02 0.08 0.51 0.01 0.17 -0.10  
## Overall\_Loss\_2021 -0.31 0.05 0.02 0.07 0.54 0.00 0.17 -0.09  
## Gender\_Inequality\_Index\_2021 -0.31 0.11 0.01 -0.26 -0.04 0.07 0.18 0.12  
## Maternal\_Mortality\_Rate\_2021 -0.32 -0.04 0.10 -0.15 0.11 0.01 0.13 -0.14  
## Adolescent\_Birth\_Rate\_2021 -0.29 -0.05 0.22 -0.47 -0.20 0.14 0.19 0.60  
## Female\_Labour\_Force\_2021 -0.04 -0.72 0.15 0.46 0.06 0.35 0.00 0.29  
## Male\_Labour\_Force\_2021 -0.10 -0.55 -0.67 -0.42 0.00 -0.07 -0.02 -0.21  
## CO2\_Emissions\_percapita\_2021 0.29 0.10 -0.02 -0.35 0.39 0.37 -0.56 0.20  
## PC9 PC10 PC11 PC12 PC13  
## HDI\_2021 -0.06 -0.12 0.03 0.85 0.00  
## Life\_Expectancy\_2021 -0.18 0.39 -0.13 -0.24 -0.02  
## Mean\_Years\_Schooling\_2021 0.05 0.38 -0.07 -0.24 -0.01  
## GNI\_Per\_Capita\_2021 -0.02 -0.70 0.10 -0.41 0.01  
## Gender\_Dev\_Index\_2021 0.02 0.05 0.12 -0.01 0.00  
## Human\_Inequality\_Coeff\_2021 0.22 0.16 0.04 0.01 -0.72  
## Overall\_Loss\_2021 0.23 0.19 0.01 0.00 0.69  
## Gender\_Inequality\_Index\_2021 -0.51 0.09 0.71 -0.03 0.01  
## Maternal\_Mortality\_Rate\_2021 -0.64 -0.12 -0.63 0.02 0.00  
## Adolescent\_Birth\_Rate\_2021 0.38 -0.06 -0.20 0.00 -0.01  
## Female\_Labour\_Force\_2021 -0.15 0.07 0.07 -0.02 0.01  
## Male\_Labour\_Force\_2021 0.10 -0.02 0.02 0.01 0.00  
## CO2\_Emissions\_percapita\_2021 -0.15 0.33 -0.07 0.01 -0.01

*Based on our analysis, we decided to retain only the first 2 principal components. The first component appears to be a general measure of overall well-being and development. Significant positive loadings for this component include variables such as HDI, life expectancy, and mean years of schooling, all of which are important benchmarks for assessing a country’s development. On the other hand, the loadings that are equally large in magnitude but negative include the inequality level, overall loss, gender inequality and the mortality rate. These negative loadings are indicative of lesser developed countries, which is why they have opposite signs to the positive indicators of development.*

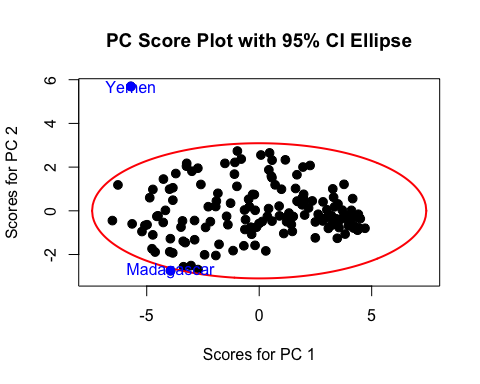
*Our second principle component describes gender equality. It is characterized by large loadings for three key variables: the gender development index, participation in the female labor force and participation in the male labor force. These variables share a common relationship with gender equality and provide a holistic measure when considered alongside the gender inequality index and the differences in employment levels between males and females.*

### Visualizations and Validations of PCA Results

We want to see how observations of countries are distributed among the two principal components and identify outliers, which deviate from multivariate normal distributions of scores.

We also create a biplot to help understand how the observations and variables interact in a reduced dimensional space.

source("http://reuningscherer.net/multivariate/r/ciscoreplot.R.txt")  
  
ciscoreplot(pc1, c(1, 2), data[, 1])

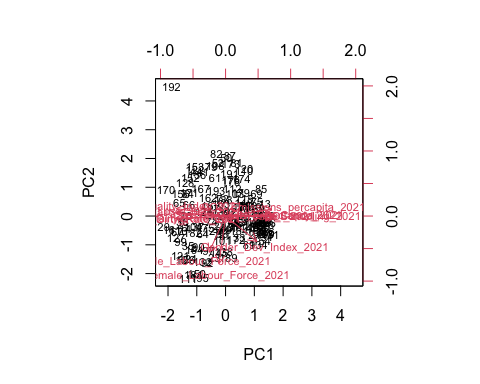


*The score plot visualizes the distribution of observations based on the first and second principal components. It includes a 95% confidence ellipse to identify outliers and confirm the distribution of points. There don’t appear to be any clear trends/groupings in our score plot of the first and second principle component.*

*There are 2 outliers in our 95% confidence ellipse of our 2 retained components: Yemen and Madagascar. Madagascar appears to fall outside of the ellipse in the direction of the second component only slightly (this is in large part due to it’s orientation in relation to the first principle component). In contrast, Yemen lies outside the ellipse in the direction of the second principle component by a good margin, indicating a more significant deviation in the factors captured by PC2, possibly related to gender equality.*

*Our interpretations of the 95% confidence ellipse are valid given we have a multivariate normal distribution in our dataset as shown in the chi-square quantile plot and is a necessary condition for interpreting the ellipse.*

biplot(pc1, choices = c(1, 2), pc.biplot = TRUE, cex = 0.7)



*The biplot shows the distribution of countries based on their principal component scores for the first and second principle components. The arrows indicate the strength of a variable’s contribution (magnitude) while the direction underscores how the variable contributed to the components. Similar to the score plot above, there doesn’t appear to be many distinct trends between the principal component scores of the countries. However, the arrows reinforce our interpretation of the principle components. The second principal component is primarily influenced by variables such as the female/male labor force and gender development index (the arrows point up). On the other hand, the first component is influenced by development variables such as GNI per capita and the mortality rate, which point in opposite directions since they are negatively correlated. Likewise, when the arrows are perpendicular, this reflects smaller loadings for those variables in defining that particular principal component, as observed in the data.*

*Based on our findings, we were able to utilize PCA because our variables had strong correlations, linear relationships, and were normally distributed (although this point itself isn’t necessary for PCA it allowed us to use methods like parallel anaylsis). We were able to capture approximately 78% of the total variance using two principal components, derived through various methods including a scree plot, parallel analysis, and the eigenvalue > 1 criteria. The first principle component captures the overall development of a country, with development-related variables showing large positive loading, while variables indicative of lower development had large negative loadings. In contrast, the second principal component focuses solely on gender equality, driven by variables relating to female/male labor employment and the gender inequality index.*

*Finally, we created a score plot (with 95% confidence interval ellipse) and a biplot to analyze our results. In the 95% confidence interval ellipse, there were 2 outliers which is unsurprising given we had over 150 observations and the deviation for one of those outliers was negligable in the direction of the second component. The biplot validated our interpretations of the principal components and was appropiate for analysis, given that we verified the multivariate normality of the data’s distribution.*