PCA Inference

Richa Singh

September 22, 2018

## PCA

Principal Component Analysis is the analysis of data to identify patterns and finding patterns to reduce the dimensions of the dataset with minimal loss of information. Here, our desired outcome of the principal component analysis is to project a feature space (our dataset consisting of n d-dimensional samples) onto a smaller subspace that represents our data well. PCA gives one map for the rows (called factor scores), and one map for the columns (called loadings).These 2 maps are related, because they both are described by the same components. However, these 2 maps project different kinds of information onto the components, and so they are *interpreted differently*. Factor scores are the coordinates of the row observations. They are interpreted by the distances between them, and their distance from the origin. Loadings describe the column variables. Loadings are interpreted by the angle between them, and their distance from the origin. The distance from the origin is important in both maps, because squared distance from the mean is inertia. Because of the Pythagorean Theorem, the total information contributed by a data point is also equal to the sum of its squared factor scores.

## Dataset : IBM-HR-Employee-NoAttrition

The dataset consists of 1233 observations and 32 variables describing the HR-IBM Employees. The variables Sub,Age,Monthly Income,Daily Rate,Hourly Rate, MonthlyRate,DistanceFromeHome,PerformanceRating, Education, JobInvolvement, Joblevel,StockOptionLevel,PercentSalaryHike,TotalWorkingYears,TrainingTimesLastYear,WorkLifeBalance,WorkLifeBalance,YearsAtCompany,YearsInCurrentRole,YearsSinceLastPromotion,YearsWithCurrManager,EnvironmentSatisfaction,JobSatisfaction,RelationshipSatisfaction,Attrition,BusinessTravel,EducationField,Gender,JobRole,MaritalStatus,OverTime are Qualitative variable.

my\_data <- read.csv("IBM-HR-Emplyee-NoAttrition.csv")  
rownames(my\_data) <- my\_data$Subj  
my\_data1 <- my\_data[sapply(my\_data, function(x) !is.factor(x))]  
my\_data1 <- subset(my\_data1, select=-Subj)

## Results

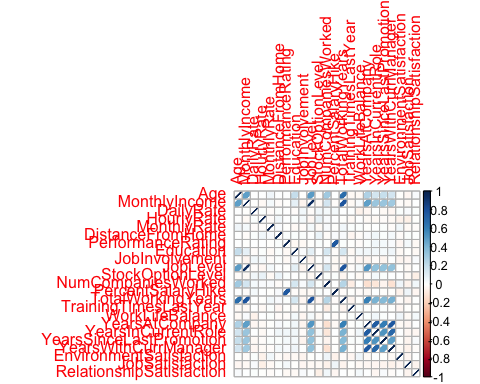
## Correlation Plot

Let’s analyze the correlation plot and find the prinicipal components.

library(corrplot)

## corrplot 0.84 loaded

cor.res <- cor(my\_data1)  
corrplot(cor.res, method = "ellipse")



DESIGN variables

#Create an empty list, called DESIGN  
library(ExPosition)

## Loading required package: prettyGraphs

DESIGN <- list()  
DESIGN$rows$Region$labels <- unique(my\_data$Department)   
DESIGN$rows$Region$vec <- my\_data$Department  
  
#Convert the vector to a matrix  
DESIGN$rows$Region$mat <- makeNominalData(as.matrix(DESIGN$rows$Region$vec))  
automatic\_colors <- createColorVectorsByDesign(DESIGN$rows$Region$mat)  
DESIGN$rows$Region$color\_groups <- c("yellow","red","green")  
DESIGN$rows$Region$color\_observ <- as.matrix(DESIGN$rows$Region$vec)   
  
DESIGN$rows$Region$color\_observ[which(DESIGN$rows$Region$vec=="Research & Development")] <- DESIGN$rows$Region$color\_groups[1]  
DESIGN$rows$Region$color\_observ[which(DESIGN$rows$Region$vec=="Human Resources")] <- DESIGN$rows$Region$color\_groups[2]  
DESIGN$rows$Region$color\_observ[which(DESIGN$rows$Region$vec=="Sales")] <- DESIGN$rows$Region$color\_groups[3]

Because each variable is measured on different units, I choose to center and scale the columns. The rows are color-coded by the DESIGN variable, state.division. \* center = TRUE: substracts the mean from each column \* scale = TRUE: after centering (or not), scales each column (see the help for different scaling options) \* DESIGN: colors the observations (rows)

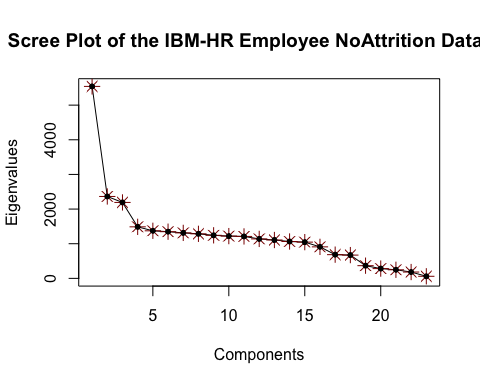
res\_pca <- epPCA(my\_data1, center = TRUE, scale = TRUE , graphs = FALSE)

## Scree Plot

A Scree Plot is a simple line segment plot that shows the fraction of total variance in the data as explained or represented by each PC.(In the PCA literature, the plot is called a 'Scree' Plot because it often looks like a 'scree' slope, where rocks have fallen down and accumulated on the side of a mountain.)The scree plot shows the eigenvalues, the amount of information on each component. The number of components (the dimensionality of the factor space) is min(nrow(DATA), ncol(DATA)) minus 1. Here, min(1233,31)-1 give 30 components. The scree plot is used to determine how many of the components should be interpreted.

* plot draws the line that connects all data points by type = "l"
* The first points function draws round purple dots.
* The second points function draws black circles around the dots (just to make it prettier).

name\_the\_scree <- plot(res\_pca$ExPosition.Data$eigs,  
 ylab = "Eigenvalues",  
 xlab = "Components",  
 type = "l",  
 main = "Scree Plot of the IBM-HR Employee NoAttrition Dataset",  
 ylim = c(-1, max(res\_pca$ExPosition.Data$eigs)))  
points(res\_pca$ExPosition.Data$eigs, cex = 1.5, pch = 8, col = "dark red")  
points(res\_pca$ExPosition.Data$eigs, cex = 1, pch = 20, col = "black")

 The top 4 compnents can be analyzed.

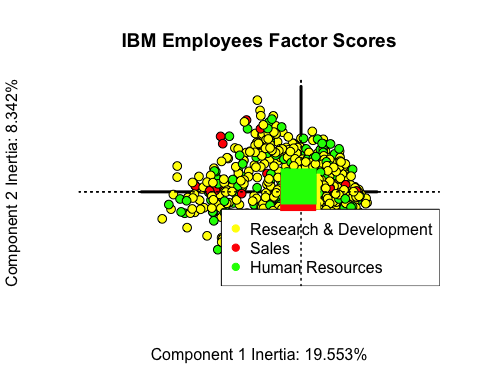
## Factor Scores for Department Variable

## Component 1 Vs Component 2

Factor scores are the coordinates of the 1233 rows of the employees on the components. The distances between them show which all employees are the most similar ones. Factor scores (employees) can be color-coded to help interpret the components.

* prettyPlot helps plot the factor scores. In order to print the result in an Rmd, dev.new needs to be FALSE.

#FACTOR SCORES  
OriginAsDesignMat <- makeNominalData(as.matrix(my\_data$Department))  
Fi2plot <- res\_pca$ExPosition.Data$fi  
fi\_all.sum <- t(Fi2plot) %\*% OriginAsDesignMat  
fi\_all.count <- colSums(OriginAsDesignMat)  
fi\_all.mean <- t(fi\_all.sum)/fi\_all.count  
  
#pretty plot between component1 and 2   
name\_the\_plot <- prettyPlot(data\_matrix = res\_pca$ExPosition.Data$fi,   
 dev.new=FALSE,  
 main = "IBM Employees Factor Scores",  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE, contributions = res\_pca$ExPosition.Data$ci,   
 display\_points = TRUE, pch = 21, cex = 1.2, col = DESIGN$rows$Region$color\_observ,   
 display\_names = FALSE,   
 xlab = paste0("Component 1 Inertia: ", round(res\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(res\_pca$ExPosition.Data$t[2],3), "%")  
)  
legend(x="bottomright", pch = 19, legend = DESIGN$rows$Region$labels, col=DESIGN$rows$Region$color\_groups)  
prettyPlot(fi\_all.mean,   
 col = DESIGN$rows$Region$color\_groups,   
 display\_names = FALSE,  
 pch = 15,   
 cex = 5.0,  
 dev.new = FALSE,  
 new.plot = FALSE)

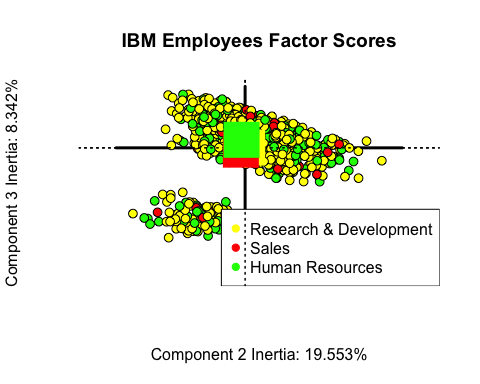


## $col  
## [,1]   
## [1,] "yellow"  
## [2,] "red"   
## [3,] "green"   
##   
## $pch  
## [,1]  
## [1,] 15  
## [2,] 15  
## [3,] 15  
##   
## $constraints  
## $constraints$minx  
## [1] -0.1831883  
##   
## $constraints$miny  
## [1] -0.0769787  
##   
## $constraints$maxx  
## [1] 0.08769324  
##   
## $constraints$maxy  
## [1] 0.3439835

* Component 1: May be Sales vs Human Resources
* Component 2: No significant inference could be observed

## Component 2 Vs Component 3

name\_the\_plot <- prettyPlot(data\_matrix = res\_pca$ExPosition.Data$fi,   
 dev.new=FALSE,  
 main = "IBM Employees Factor Scores",  
 x\_axis = 2, y\_axis = 3,   
 contributionCircles = FALSE, contributions = res\_pca$ExPosition.Data$ci,   
 display\_points = TRUE, pch = 21, cex = 1.2, col = DESIGN$rows$Region$color\_observ,   
 display\_names = FALSE,   
 xlab = paste0("Component 2 Inertia: ", round(res\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 3 Inertia: ", round(res\_pca$ExPosition.Data$t[2],3), "%")  
)  
legend(x="bottomright", pch = 19, legend = DESIGN$rows$Region$labels, col=DESIGN$rows$Region$color\_groups)  
prettyPlot(fi\_all.mean,   
 col = DESIGN$rows$Region$color\_groups,   
 display\_names = FALSE,  
 pch = 15,   
 cex = 5.0,  
 dev.new = FALSE,  
 new.plot = FALSE)



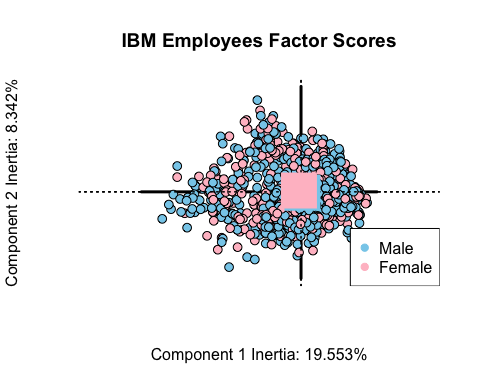
## $col  
## [,1]   
## [1,] "yellow"  
## [2,] "red"   
## [3,] "green"   
##   
## $pch  
## [,1]  
## [1,] 15  
## [2,] 15  
## [3,] 15  
##   
## $constraints  
## $constraints$minx  
## [1] -0.1831883  
##   
## $constraints$miny  
## [1] -0.0769787  
##   
## $constraints$maxx  
## [1] 0.08769324  
##   
## $constraints$maxy  
## [1] 0.3439835

* Component 2: Definitely Sales vs Human Resources
* Component 3: Maybe divides Sales & Human Rescources VS Research and Development

## Factor Scores for Gender Variable

## Component 1 Vs Component 2

#Create an empty list, called DESIGN  
library(ExPosition)  
DESIGN1 <- list()  
DESIGN1$rows$Region$labels <- unique(my\_data$Gender)   
DESIGN1$rows$Region$vec <- my\_data$Gender  
  
#Convert the vector to a matrix  
DESIGN1$rows$Region$mat <- makeNominalData(as.matrix(DESIGN1$rows$Region$vec))  
automatic\_colors <- createColorVectorsByDesign(DESIGN1$rows$Region$mat)  
DESIGN1$rows$Region$color\_groups <- c("skyblue","pink")  
DESIGN1$rows$Region$color\_observ <- as.matrix(DESIGN1$rows$Region$vec)   
  
DESIGN1$rows$Region$color\_observ[which(DESIGN1$rows$Region$vec=="Male")] <- DESIGN1$rows$Region$color\_groups[1]  
DESIGN1$rows$Region$color\_observ[which(DESIGN1$rows$Region$vec=="Female")] <- DESIGN1$rows$Region$color\_groups[2]  
OriginAsDesignMat1 <- makeNominalData(as.matrix(my\_data$Gender))  
Fi2plot1 <- res\_pca$ExPosition.Data$fi  
fi\_all.sum1 <- t(Fi2plot1) %\*% OriginAsDesignMat1  
fi\_all.count1 <- colSums(OriginAsDesignMat1)  
fi\_all.mean1 <- t(fi\_all.sum1)/fi\_all.count1  
  
#pretty plot between component1 and 2   
name\_the\_plot1 <- prettyPlot(data\_matrix = res\_pca$ExPosition.Data$fi,   
 dev.new=FALSE,  
 main = "IBM Employees Factor Scores",  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE, contributions = res\_pca$ExPosition.Data$ci,   
 display\_points = TRUE, pch = 21, cex = 1.2, col = DESIGN1$rows$Region$color\_observ,   
 display\_names = FALSE,   
 xlab = paste0("Component 1 Inertia: ", round(res\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(res\_pca$ExPosition.Data$t[2],3), "%")  
)  
legend(x="bottomright", pch = 19, legend = DESIGN1$rows$Region$labels, col=DESIGN1$rows$Region$color\_groups)  
prettyPlot(fi\_all.mean1,   
 col = DESIGN1$rows$Region$color\_groups,   
 display\_names = FALSE,  
 pch = 15,   
 cex = 5.0,  
 dev.new = FALSE,  
 new.plot = FALSE)

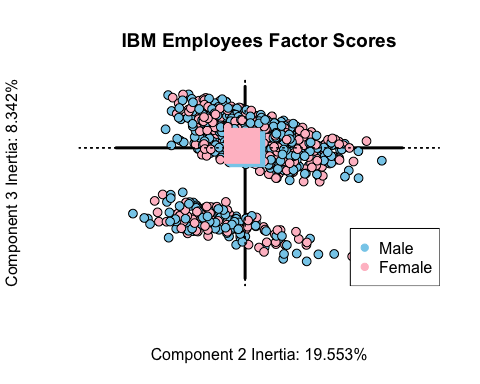


## $col  
## [,1]   
## [1,] "skyblue"  
## [2,] "pink"   
##   
## $pch  
## [,1]  
## [1,] 15  
## [2,] 15  
##   
## $constraints  
## $constraints$minx  
## [1] -0.1331817  
##   
## $constraints$miny  
## [1] -0.05849934  
##   
## $constraints$maxx  
## [1] 0.09115308  
##   
## $constraints$maxy  
## [1] 0.08547208

* Both the means lie almost on the origin so can’t make any significant inference for both the Components.

## Component 2 Vs Component 3

name\_the\_plot1 <- prettyPlot(data\_matrix = res\_pca$ExPosition.Data$fi,   
 dev.new=FALSE,  
 main = "IBM Employees Factor Scores",  
 x\_axis = 2, y\_axis = 3,   
 contributionCircles = FALSE, contributions = res\_pca$ExPosition.Data$ci,   
 display\_points = TRUE, pch = 21, cex = 1.2, col = DESIGN1$rows$Region$color\_observ,   
 display\_names = FALSE,   
 xlab = paste0("Component 2 Inertia: ", round(res\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 3 Inertia: ", round(res\_pca$ExPosition.Data$t[2],3), "%")  
)  
legend(x="bottomright", pch = 19, legend = DESIGN1$rows$Region$labels, col=DESIGN1$rows$Region$color\_groups)  
prettyPlot(fi\_all.mean1,   
 col = DESIGN1$rows$Region$color\_groups,   
 display\_names = FALSE,  
 pch = 15,   
 cex = 5.0,  
 dev.new = FALSE,  
 new.plot = FALSE)

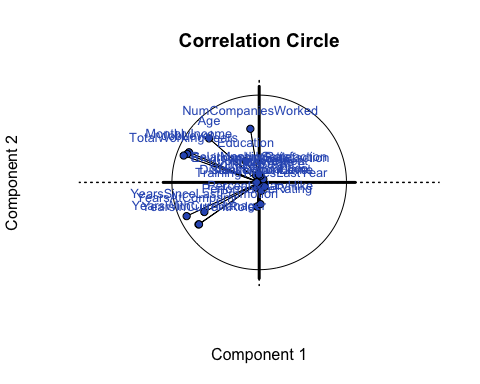


## $col  
## [,1]   
## [1,] "skyblue"  
## [2,] "pink"   
##   
## $pch  
## [,1]  
## [1,] 15  
## [2,] 15  
##   
## $constraints  
## $constraints$minx  
## [1] -0.1331817  
##   
## $constraints$miny  
## [1] -0.05849934  
##   
## $constraints$maxx  
## [1] 0.09115308  
##   
## $constraints$maxy  
## [1] 0.08547208

* Component 2: Maybe we can say that Component 2 divides Female vs Male
* Component 3: No significant inference

## Correlation Circle

correlationCircle <- correlationPlotter(data\_matrix = my\_data1 , factor\_scores =res\_pca$ExPosition.Data$fi , x\_axis = 1, y\_axis = 2, col = NULL ,pch = NULL, xlab = NULL, ylab = NULL, main = "Correlation Circle", asp = 1, dev.new = FALSE)

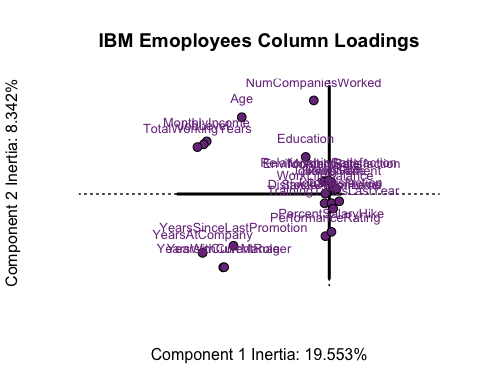


## Loadings

## Component 1 Vs Component 2

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.

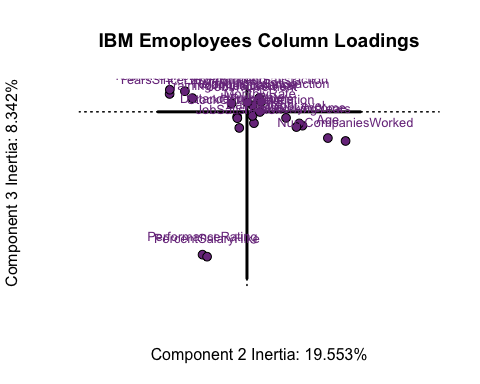
name\_another\_plot <- prettyPlot(data\_matrix = res\_pca$ExPosition.Data$fj,   
 dev.new=FALSE,  
 main = "IBM Emoployees Column Loadings",  
 x\_axis = 1, y\_axis = 2,   
 contributionCircles = FALSE, contributions = res\_pca$ExPosition.Data$cj,   
 display\_points = TRUE, pch = 21, cex = 1.2, col = res\_pca$Plotting.Data$fj.col ,   
 display\_names = TRUE,   
 xlab = paste0("Component 1 Inertia: ", round(res\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 2 Inertia: ", round(res\_pca$ExPosition.Data$t[2],3), "%")  
)



* Component 1: YearsatCompany & JobLevel & MonthlyIncome
* Component 2: Performance Rating VS NumCompaniesWorked

## Component 2 Vs Component 3

name\_another\_plot <- prettyPlot(data\_matrix = res\_pca$ExPosition.Data$fj,   
 dev.new=FALSE,  
 main = "IBM Emoployees Column Loadings",  
 x\_axis = 2, y\_axis = 3,   
 contributionCircles = FALSE, contributions = res\_pca$ExPosition.Data$cj,   
 display\_points = TRUE, pch = 21, cex = 1.2, col = res\_pca$Plotting.Data$fj.col,   
 display\_names = TRUE,   
 xlab = paste0("Component 2 Inertia: ", round(res\_pca$ExPosition.Data$t[1],3), "%"),  
 ylab = paste0("Component 3 Inertia: ", round(res\_pca$ExPosition.Data$t[2],3), "%")  
)



* Component 2: YearsatCompany VS NumCompaniesWorked
* Component 3: Performance Rating & PercentSalaryHike

## Summary

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

* Component 1: Usually the Research and Development people have more number of years at Company & Monthly Income
* Component 2: Human Resources people have high Performance Rating
* Component 3: Males have worked in more number of Companies