HW4p5

Sagar Ganapaneni

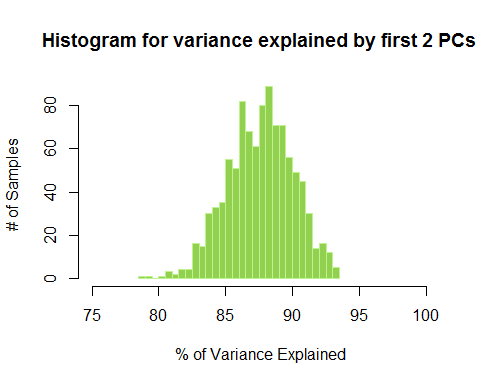
October 27, 2016

1. Using the USArrests dataset, plot a histogram of the proportion of variance explained by the first 2 principal components in 1000 Bootstrap resamplings of the data.

## Load data  
library(ISLR)

## Warning: package 'ISLR' was built under R version 3.2.5

attach(USArrests)  
set.seed(12)  
PC12\_prop <- numeric(1000)  
for(i in 1:1000){  
 ## sample USArrest  
 split = sample(1:nrow(USArrests), size=nrow(USArrests)/2)  
 USArrest\_boot <- USArrests[split, ]  
 ## do PCA   
 PCA\_USArrests <- prcomp(USArrest\_boot,center = TRUE,scale. = TRUE)  
  
 ## Explnation by first two PCS  
 PC\_variance<- PCA\_USArrests$sdev^2  
 PC\_var\_prop <- PC\_variance/sum(PC\_variance)  
 PC12\_prop[i] <- (PC\_var\_prop[1]+PC\_var\_prop[2])\*100  
}  
  
hist(PC12\_prop,   
 main="Histogram for variance explained by first 2 PCs",  
 ylab="# of Samples ",   
 xlab=" % of Variance Explained",   
 border="#c0f292",   
 col="#92d050",  
 xlim=c(75,100),  
 breaks=25)

 (b) Estimate a standard error and 95% confidence intervals for the proportion of variance explained by the first 2 principal components.

# mean  
pc\_mean<-mean(PC12\_prop)  
# sd  
pc\_sd<-sd(PC12\_prop)  
# confidance interval  
pc\_CI <- c(pc\_mean-2\*pc\_sd,pc\_mean+2\*pc\_sd)  
pc\_CI

## [1] 82.73915 92.49213

(c)Suppose we compute the first principal component from each of 1000 Bootstrap resamplings of the data. Using the resulting 1000 vectors, we estimate the standard error of each entry or loading using Eq. 5.8 in the textbook. Explain why this would be problematic.

## if the interested statistic is Principal component, then it would be a vector, it would be difficult measure standard devition oa vector component in multi dimensional space.

1. There is a way around the problem alluded to in part 3. Write a function in R which, given a data.frame:

Computes the vector of loadings for the first principal component and defines i to be the index of the element with highest absolute value. For each of 1000 bootstrap resamplings of the data.frame, computes the vector of loadings for the first principal component and multiplies it by the sign of its ith element to generate signed loadings. Plots a boxplot of the signed loadings in the boostrap samples.

library(ISLR)  
attach(USArrests)

## The following objects are masked from USArrests (pos = 3):  
##   
## Assault, Murder, Rape, UrbanPop

set.seed(12)  
data<-USArrests  
  
signed\_loadings\_generator<- function(data){  
   
 ## initialize output dataframe  
  
 ## loop 1000 times  
 for (n in 1:1000){  
   
 ## sample data  
 split = sample(1:nrow(data), size=nrow(data)/2)  
 data\_boot <- USArrests[split, ]  
 ## do PCA   
 PCA\_data <- prcomp(data\_boot,center = TRUE,scale. = TRUE)  
  
 ## PC1  
 pc\_1<-PCA\_data$rotation[,1]  
   
 ## index   
 i <- which.max(abs(pc\_1))  
   
 ## sign loading  
 if (pc\_1[i]<0) pc\_1<- (-pc\_1)  
   
 ## rbind  
 pc\_1<-t(as.data.frame(pc\_1))  
 if (exists("pc1\_loadings")) pc1\_loadings<-rbind(pc1\_loadings,pc\_1)  
 else pc1\_loadings<-pc\_1  
 }  
 ## plot   
boxplot(pc1\_loadings, main= "boxplot of the signed loadings in the boostrap samples" , col=(c("gold","darkgreen","red","blue")),ylab="signed value" , xlab="elements")  
}

1. Apply the function to USArrests

signed\_loadings\_generator(USArrests)

