**Exercise PCA**

Jitesh Rawat

3/22/2021

**library**(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.4 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

**library**(broom)  
**library**(knitr)  
*#install.packages("ggfortify")*  
**library**(ggfortify)

## Warning: package 'ggfortify' was built under R version 4.0.4

data <- datasets**::**swiss  
**head**(data)

## Fertility Agriculture Examination Education Catholic  
## Courtelary 80.2 17.0 15 12 9.96  
## Delemont 83.1 45.1 6 9 84.84  
## Franches-Mnt 92.5 39.7 5 5 93.40  
## Moutier 85.8 36.5 12 7 33.77  
## Neuveville 76.9 43.5 17 15 5.16  
## Porrentruy 76.1 35.3 9 7 90.57  
## Infant.Mortality  
## Courtelary 22.2  
## Delemont 22.2  
## Franches-Mnt 20.2  
## Moutier 20.3  
## Neuveville 20.6  
## Porrentruy 26.6

*#Compute Smaple Covariance matrix S*  
s <- **cov**(data)  
**head**(s)

## Fertility Agriculture Examination Education Catholic  
## Fertility 156.04250 100.169149 -64.366929 -79.729510 241.56320  
## Agriculture 100.16915 515.799417 -124.392831 -139.657401 379.90438  
## Examination -64.36693 -124.392831 63.646623 53.575856 -190.56061  
## Education -79.72951 -139.657401 53.575856 92.456059 -61.69883  
## Catholic 241.56320 379.904376 -190.560611 -61.698830 1739.29454  
## Infant.Mortality 15.15619 -4.025851 -2.649537 -2.781684 21.31812  
## Infant.Mortality  
## Fertility 15.156193  
## Agriculture -4.025851  
## Examination -2.649537  
## Education -2.781684  
## Catholic 21.318116  
## Infant.Mortality 8.483802

*# Computing Eigen Values and Eigen Vectors of S*  
eig\_val <- **eigen**(s)  
eig\_val**$**values

## [1] 1921.562488 466.657132 145.284829 22.649607 13.377631 6.191249

eig\_val**$**vectors

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] -0.15163143 -0.14270789 0.81454413 -0.49552828 0.12247267 -0.1805890073  
## [2,] -0.28121756 -0.85914886 -0.35256541 -0.24078519 0.02235164 0.0006755062  
## [3,] 0.12207834 0.17688621 -0.18767793 -0.57042350 -0.76887882 -0.0450281625  
## [4,] 0.06329733 0.32260928 -0.40096045 -0.58075837 0.61904343 -0.1032102386  
## [5,] -0.93748965 0.32543441 -0.07870742 0.04104832 -0.08547871 0.0043788222  
## [6,] -0.01131739 0.01498883 0.10014161 -0.17923725 0.05296062 0.9770814149

*#Perform PCA*  
PC <- **as.matrix**(data) **%\*%** eig\_val**$**vectors  
**ggplot**(PC,**aes**(PC[,1],PC[,3]))**+**  
 **geom\_point**()



x <- **c**(1**:**6)  
y <- eig\_val**$**values  
**plot**(x,y,type="l", main="Scree Plot")



**var**(PC[,1])**/sum**(eig\_val**$**values)

## [1] 0.7460284

**var**(PC[,2])**/sum**(eig\_val**$**values)

## [1] 0.1811752

**var**(PC[,3])**/sum**(eig\_val**$**values)

## [1] 0.05640546

**var**(PC[,4])**/sum**(eig\_val**$**values)

## [1] 0.008793495

**var**(PC[,5])**/sum**(eig\_val**$**values)

## [1] 0.005193739

**var**(PC[,6])**/sum**(eig\_val**$**values)

## [1] 0.002403694

eig\_val**$**vectors

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] -0.15163143 -0.14270789 0.81454413 -0.49552828 0.12247267 -0.1805890073  
## [2,] -0.28121756 -0.85914886 -0.35256541 -0.24078519 0.02235164 0.0006755062  
## [3,] 0.12207834 0.17688621 -0.18767793 -0.57042350 -0.76887882 -0.0450281625  
## [4,] 0.06329733 0.32260928 -0.40096045 -0.58075837 0.61904343 -0.1032102386  
## [5,] -0.93748965 0.32543441 -0.07870742 0.04104832 -0.08547871 0.0043788222  
## [6,] -0.01131739 0.01498883 0.10014161 -0.17923725 0.05296062 0.9770814149

**names**(data)

## [1] "Fertility" "Agriculture" "Examination" "Education"   
## [5] "Catholic" "Infant.Mortality"

*#From below PC values we see that 3 components actually suffice*

**eigen**(**cor**(data))**$**vectors

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] -0.4569876 0.3220284 0.17376638 0.53555794 0.38308893 0.47295441  
## [2,] -0.4242141 -0.4115132 -0.03834472 -0.64291822 0.37495215 0.30870058  
## [3,] 0.5097327 0.1250167 0.09123696 -0.05446158 0.81429082 -0.22401686  
## [4,] 0.4543119 0.1790495 -0.53239316 -0.09738818 -0.07144564 0.68081610  
## [5,] -0.3501111 0.1458730 -0.80680494 0.09947244 0.18317236 -0.40219666  
## [6,] -0.1496668 0.8111645 0.16010636 -0.52677184 -0.10453530 -0.07457754

comp <- **prcomp**(data)  
**summary**(comp)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 43.836 21.6022 12.05342 4.75916 3.65754 2.4882  
## Proportion of Variance 0.746 0.1812 0.05641 0.00879 0.00519 0.0024  
## Cumulative Proportion 0.746 0.9272 0.98361 0.99240 0.99760 1.0000

y <- **eigen**(**cor**(data))**$**values  
**plot**(x,y,type="l",main="Scree plot")



*#From the above scree plot we can see that first 4 principal components are enough*

data **%>%** **head**(10)

## Fertility Agriculture Examination Education Catholic  
## Courtelary 80.2 17.0 15 12 9.96  
## Delemont 83.1 45.1 6 9 84.84  
## Franches-Mnt 92.5 39.7 5 5 93.40  
## Moutier 85.8 36.5 12 7 33.77  
## Neuveville 76.9 43.5 17 15 5.16  
## Porrentruy 76.1 35.3 9 7 90.57  
## Broye 83.8 70.2 16 7 92.85  
## Glane 92.4 67.8 14 8 97.16  
## Gruyere 82.4 53.3 12 7 97.67  
## Sarine 82.9 45.2 16 13 91.38  
## Infant.Mortality  
## Courtelary 22.2  
## Delemont 22.2  
## Franches-Mnt 20.2  
## Moutier 20.3  
## Neuveville 20.6  
## Porrentruy 26.6  
## Broye 23.6  
## Glane 24.9  
## Gruyere 21.0  
## Sarine 24.4

**head**(comp**$**x)

## PC1 PC2 PC3 PC4 PC5 PC6  
## Courtelary 37.032433 -17.434879 22.609928 1.6927673 5.0418785 -0.1918163  
## Delemont -42.797334 -14.687668 12.063389 3.4394561 4.6872632 -0.7298603  
## Franches-Mnt -51.081639 -19.274036 22.541458 3.6850353 3.1728935 2.4301351  
## Moutier 7.716707 -5.458722 20.799893 0.1554713 3.2391283 1.9073690  
## Neuveville 35.032658 5.126097 9.218281 -5.8461713 5.8749706 1.1783747  
## Porrentruy -44.161953 -25.922412 10.045238 8.1646581 -0.1905812 -6.3829483

PC1=**-**0.4569876**\***0.75 -0.4242141**\***0.18**+** 0.5097327**\***0.056**+** 0.4543119**\*** -0.3501111**\***0.0088 -0.1496668**\***0.002403694

PC <- **as.matrix**(data) **%\*%** eig\_val**$**vectors  
**ggplot**(PC,**aes**(PC[,1],PC[,2]))**+**  
 **geom\_point**()



*#we see that PC 1 and PC 2 are not much correlated*

**ggplot**(PC,**aes**(PC[,2],PC[,3]))**+geom\_point**()



*#We see that there is some positive correlation here*

**ggplot**(PC,**aes**(PC[,1],PC[,3]))**+** **geom\_point**()



*#PC1 and PC3 are not much correlated*