project

Benjan

11/18/2022

Lets read the data first.

cancer <- read.csv("C:/Users/benja/OneDrive/Desktop/data.csv")

# 3. Dimensionality Reduction using PCA

Lets examine the data first.

summary(cancer)

## id diagnosis radius\_mean texture\_mean   
## Min. : 8670 Length:569 Min. : 6.981 Min. : 9.71   
## 1st Qu.: 869218 Class :character 1st Qu.:11.700 1st Qu.:16.17   
## Median : 906024 Mode :character Median :13.370 Median :18.84   
## Mean : 30371831 Mean :14.127 Mean :19.29   
## 3rd Qu.: 8813129 3rd Qu.:15.780 3rd Qu.:21.80   
## Max. :911320502 Max. :28.110 Max. :39.28   
## perimeter\_mean area\_mean smoothness\_mean compactness\_mean   
## Min. : 43.79 Min. : 143.5 Min. :0.05263 Min. :0.01938   
## 1st Qu.: 75.17 1st Qu.: 420.3 1st Qu.:0.08637 1st Qu.:0.06492   
## Median : 86.24 Median : 551.1 Median :0.09587 Median :0.09263   
## Mean : 91.97 Mean : 654.9 Mean :0.09636 Mean :0.10434   
## 3rd Qu.:104.10 3rd Qu.: 782.7 3rd Qu.:0.10530 3rd Qu.:0.13040   
## Max. :188.50 Max. :2501.0 Max. :0.16340 Max. :0.34540   
## concavity\_mean concave.points\_mean symmetry\_mean fractal\_dimension\_mean  
## Min. :0.00000 Min. :0.00000 Min. :0.1060 Min. :0.04996   
## 1st Qu.:0.02956 1st Qu.:0.02031 1st Qu.:0.1619 1st Qu.:0.05770   
## Median :0.06154 Median :0.03350 Median :0.1792 Median :0.06154   
## Mean :0.08880 Mean :0.04892 Mean :0.1812 Mean :0.06280   
## 3rd Qu.:0.13070 3rd Qu.:0.07400 3rd Qu.:0.1957 3rd Qu.:0.06612   
## Max. :0.42680 Max. :0.20120 Max. :0.3040 Max. :0.09744   
## radius\_se texture\_se perimeter\_se area\_se   
## Min. :0.1115 Min. :0.3602 Min. : 0.757 Min. : 6.802   
## 1st Qu.:0.2324 1st Qu.:0.8339 1st Qu.: 1.606 1st Qu.: 17.850   
## Median :0.3242 Median :1.1080 Median : 2.287 Median : 24.530   
## Mean :0.4052 Mean :1.2169 Mean : 2.866 Mean : 40.337   
## 3rd Qu.:0.4789 3rd Qu.:1.4740 3rd Qu.: 3.357 3rd Qu.: 45.190   
## Max. :2.8730 Max. :4.8850 Max. :21.980 Max. :542.200   
## smoothness\_se compactness\_se concavity\_se concave.points\_se   
## Min. :0.001713 Min. :0.002252 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.005169 1st Qu.:0.013080 1st Qu.:0.01509 1st Qu.:0.007638   
## Median :0.006380 Median :0.020450 Median :0.02589 Median :0.010930   
## Mean :0.007041 Mean :0.025478 Mean :0.03189 Mean :0.011796   
## 3rd Qu.:0.008146 3rd Qu.:0.032450 3rd Qu.:0.04205 3rd Qu.:0.014710   
## Max. :0.031130 Max. :0.135400 Max. :0.39600 Max. :0.052790   
## symmetry\_se fractal\_dimension\_se radius\_worst texture\_worst   
## Min. :0.007882 Min. :0.0008948 Min. : 7.93 Min. :12.02   
## 1st Qu.:0.015160 1st Qu.:0.0022480 1st Qu.:13.01 1st Qu.:21.08   
## Median :0.018730 Median :0.0031870 Median :14.97 Median :25.41   
## Mean :0.020542 Mean :0.0037949 Mean :16.27 Mean :25.68   
## 3rd Qu.:0.023480 3rd Qu.:0.0045580 3rd Qu.:18.79 3rd Qu.:29.72   
## Max. :0.078950 Max. :0.0298400 Max. :36.04 Max. :49.54   
## perimeter\_worst area\_worst smoothness\_worst compactness\_worst  
## Min. : 50.41 Min. : 185.2 Min. :0.07117 Min. :0.02729   
## 1st Qu.: 84.11 1st Qu.: 515.3 1st Qu.:0.11660 1st Qu.:0.14720   
## Median : 97.66 Median : 686.5 Median :0.13130 Median :0.21190   
## Mean :107.26 Mean : 880.6 Mean :0.13237 Mean :0.25427   
## 3rd Qu.:125.40 3rd Qu.:1084.0 3rd Qu.:0.14600 3rd Qu.:0.33910   
## Max. :251.20 Max. :4254.0 Max. :0.22260 Max. :1.05800   
## concavity\_worst concave.points\_worst symmetry\_worst fractal\_dimension\_worst  
## Min. :0.0000 Min. :0.00000 Min. :0.1565 Min. :0.05504   
## 1st Qu.:0.1145 1st Qu.:0.06493 1st Qu.:0.2504 1st Qu.:0.07146   
## Median :0.2267 Median :0.09993 Median :0.2822 Median :0.08004   
## Mean :0.2722 Mean :0.11461 Mean :0.2901 Mean :0.08395   
## 3rd Qu.:0.3829 3rd Qu.:0.16140 3rd Qu.:0.3179 3rd Qu.:0.09208   
## Max. :1.2520 Max. :0.29100 Max. :0.6638 Max. :0.20750   
## X   
## Mode:logical   
## NA's:569   
##   
##   
##   
##

As we can see that the mean of the columns are very different, if we use the unscaled data then only few of the variable will carry major weight in determining the principal components. Thus, we will use the correlation matrix of the numerical columns of the cancer data.

cancer\_pca <- princomp(covmat=cor(cancer[3:32]))

To look at the how significant the components are, we can use the cumulative proportion of variance explained by the components.

summary(cancer\_pca)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 3.6443940 2.3856560 1.67867477 1.40735229 1.28402903  
## Proportion of Variance 0.4427203 0.1897118 0.09393163 0.06602135 0.05495768  
## Cumulative Proportion 0.4427203 0.6324321 0.72636371 0.79238506 0.84734274  
## Comp.6 Comp.7 Comp.8 Comp.9 Comp.10  
## Standard deviation 1.09879780 0.82171778 0.69037464 0.64567392 0.59219377  
## Proportion of Variance 0.04024522 0.02250734 0.01588724 0.01389649 0.01168978  
## Cumulative Proportion 0.88758796 0.91009530 0.92598254 0.93987903 0.95156881  
## Comp.11 Comp.12 Comp.13 Comp.14  
## Standard deviation 0.54213992 0.511039500 0.49128148 0.396244525  
## Proportion of Variance 0.00979719 0.008705379 0.00804525 0.005233657  
## Cumulative Proportion 0.96136600 0.970071383 0.97811663 0.983350291  
## Comp.15 Comp.16 Comp.17 Comp.18  
## Standard deviation 0.306814219 0.282600072 0.243719178 0.229387845  
## Proportion of Variance 0.003137832 0.002662093 0.001979968 0.001753959  
## Cumulative Proportion 0.986488123 0.989150216 0.991130184 0.992884143  
## Comp.19 Comp.20 Comp.21 Comp.22  
## Standard deviation 0.222435590 0.176520261 0.1731268145 0.1656484305  
## Proportion of Variance 0.001649253 0.001038647 0.0009990965 0.0009146468  
## Cumulative Proportion 0.994533397 0.995572043 0.9965711397 0.9974857865  
## Comp.23 Comp.24 Comp.25 Comp.26  
## Standard deviation 0.1560155049 0.1343689213 0.1244237573 0.090430304  
## Proportion of Variance 0.0008113613 0.0006018336 0.0005160424 0.000272588  
## Cumulative Proportion 0.9982971477 0.9988989813 0.9994150237 0.999687612  
## Comp.27 Comp.28 Comp.29 Comp.30  
## Standard deviation 0.0830690308 3.986650e-02 0.0273642668 1.153451e-02  
## Proportion of Variance 0.0002300155 5.297793e-05 0.0000249601 4.434827e-06  
## Cumulative Proportion 0.9999176271 9.999706e-01 0.9999955652 1.000000e+00

The first two components explain 64 percent of the total variance and thus we will use these two as significant components. This reduces the original dataframe from 30 numerical dimensions to only 2 while preserving majority of the information.

Lets look more into these two principal components.

cancer\_pca$loadings[,1:2]

## Comp.1 Comp.2  
## radius\_mean 0.21890244 0.233857132  
## texture\_mean 0.10372458 0.059706088  
## perimeter\_mean 0.22753729 0.215181361  
## area\_mean 0.22099499 0.231076711  
## smoothness\_mean 0.14258969 -0.186113023  
## compactness\_mean 0.23928535 -0.151891610  
## concavity\_mean 0.25840048 -0.060165363  
## concave.points\_mean 0.26085376 0.034767500  
## symmetry\_mean 0.13816696 -0.190348770  
## fractal\_dimension\_mean 0.06436335 -0.366575471  
## radius\_se 0.20597878 0.105552152  
## texture\_se 0.01742803 -0.089979682  
## perimeter\_se 0.21132592 0.089457234  
## area\_se 0.20286964 0.152292628  
## smoothness\_se 0.01453145 -0.204430453  
## compactness\_se 0.17039345 -0.232715896  
## concavity\_se 0.15358979 -0.197207283  
## concave.points\_se 0.18341740 -0.130321560  
## symmetry\_se 0.04249842 -0.183848000  
## fractal\_dimension\_se 0.10256832 -0.280092027  
## radius\_worst 0.22799663 0.219866379  
## texture\_worst 0.10446933 0.045467298  
## perimeter\_worst 0.23663968 0.199878428  
## area\_worst 0.22487053 0.219351858  
## smoothness\_worst 0.12795256 -0.172304352  
## compactness\_worst 0.21009588 -0.143593173  
## concavity\_worst 0.22876753 -0.097964114  
## concave.points\_worst 0.25088597 0.008257235  
## symmetry\_worst 0.12290456 -0.141883349  
## fractal\_dimension\_worst 0.13178394 -0.275339469

The first component seems to be a linear combination of all variables expect *symmetry-se* , *texture-se*, *fractal-dimension-mean*, *smoothness-se*.

The second component is linear combination of all except *concave.poin-mean*, *texture\_worst*, *concave.point\_worst* and here almost half of the variables are contributing in the opposite direction.

# 5. Factor Analysis

Lets do EFA first. We will check the number of factors required.

sapply(1:6, function(f) factanal(cancer[3:32], factors = f,lower=0.1)$PVAL)

## objective objective objective objective objective objective   
## 0 0 0 0 0 0

We cannot determine the appropriate number of factors. So, we will choose 2 factors as we believe that 2 is adequate.

final\_factors <- factanal(covmat = cor(cancer[3:32]), factors = 2,lower=0.1)  
print(final\_factors$loadings,cut =.3)

##   
## Loadings:  
## Factor1 Factor2  
## radius\_mean 0.980   
## texture\_mean 0.347   
## perimeter\_mean 0.982   
## area\_mean 0.980   
## smoothness\_mean 0.629   
## compactness\_mean 0.545 0.788   
## concavity\_mean 0.718 0.629   
## concave.points\_mean 0.853 0.442   
## symmetry\_mean 0.595   
## fractal\_dimension\_mean 0.855   
## radius\_se 0.737   
## texture\_se   
## perimeter\_se 0.731   
## area\_se 0.787   
## smoothness\_se 0.326   
## compactness\_se 0.759   
## concavity\_se 0.651   
## concave.points\_se 0.399 0.560   
## symmetry\_se 0.336   
## fractal\_dimension\_se 0.696   
## radius\_worst 0.982   
## texture\_worst 0.333   
## perimeter\_worst 0.981   
## area\_worst 0.969   
## smoothness\_worst 0.594   
## compactness\_worst 0.456 0.744   
## concavity\_worst 0.572 0.680   
## concave.points\_worst 0.782 0.500   
## symmetry\_worst 0.508   
## fractal\_dimension\_worst 0.842   
##   
## Factor1 Factor2  
## SS loadings 10.907 7.392  
## Proportion Var 0.364 0.246  
## Cumulative Var 0.364 0.610

The larger loadings in the first factor are of radius, area and perimeter while the larger loadings in the second factor are related to smoothness, compactness, symmetry and fractal-dimension. Radius, area and perimeter are things that are easy to measure and have fixed method of measurement, so we said that the first factor was *quantitative* measures. Symmetry, smoothness are things which are difficult to measure as their interpretation of measurement varies, it is not easy say what things are smooth and how to quantify it. So, we said the second factor is *qualitative* measures.