```
> #PRINCIPAL COMPONENTS ANALYSIS (PCA)
>
> library(faraway)
> library(MASS)
>
> #write output to a file, append or overwrite, split to file and terminal
>  sink("C:/Users/jmard/OneDrive/Desktop/Computing and Graphics in Applied
Statistics2020/Output/PCAexample_Out.txt",append=FALSE,split=TRUE)
>
> #save graph(s) in pdf
> windows(7,7)
>  pdf(file="C:/Users/jmard/OneDrive/Desktop/Computing and Graphics in Applied Statistics2020/Output/PCAexample_Figure.pdf")
>
> if (FALSE)
+ {"
+ R code from page 161 of Faraway Linear Models with R (2nd edition) used
+ Data set - fat:  dimensions of the human body as measured in a study of 252 men.
+ "}
>
> data(fat,package="faraway")
> head(fat,1L)
  brozek siri density age weight height adipos  free neck chest abdom  hip thigh knee ankle biceps forearm wrist
1   12.6 12.3  1.0708  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2 94.5    59 37.3  21.9     32    27.4  17.1
> nrow(fat)
[1] 252
>
> pairs( ~ neck + chest + abdom + hip + thigh, data=fat, main="Simple Scatterplot Matrix")
>
> pairs( ~ knee + ankle + biceps + forearm + wrist, data=fat,main="Simple Scatterplot Matrix")
>
> #We see these measures are highly correlated.
> #Question - there are 13 predictors in the data set (includes age,weight,height) of which 10 are circumference measurements.
> #We have 10 highly correlated predictors - there may be less information to be extracted than the number of predictors might suggest.
> #Can we reduce the dimensionality of the 10 variables?
> #PCA aims to discover this lower dimension of variability in higher dimensional data
>
> cfat <- fat[,9:18]
> head(cfat,1L)
  neck chest abdom  hip thigh knee ankle biceps forearm wrist
1 36.2  93.1  85.2 94.5    59 37.3  21.9     32    27.4  17.1
>
> if (FALSE)
+ {"
+ Suppose all 10 variables are centered - that is neck - mean(neck) , chest - mean(chest, and so on.
+ 1. Find the u1 such that var (u1'X) is maximized subject to u1'u1 = 1.
+                                  X is n by k with the 10 centered variables forming the column vectors
+ 2. Find u2 such that var (u2'X) is maximized subject to u1'u2 = 0 and u2'u2 = 1.  u1'u2=0 => orthogonal (independence)
+ keep going for u3, u4, . . ., u10 (in this case k=10)
+ for high dimensional data we can stop when the remaining variation is negligible.
+ "}
>
> #now perform Principal Component Analysis
>
> prfat <- prcomp(cfat)
```

```
> dim(prfat$rot) # $rot contains the rotation matrix
[1] 10 10
> prfat$rot
               PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8         PC9          PC10
neck    0.12247857 -0.02419605  0.19705886 -0.24294170  0.26441810 -0.06697055 -0.61977866 -0.62290330 -0.03465366  0.1970373708
chest   0.50161641  0.38414429  0.63962671  0.36449079 -0.23977622 -0.00628437  0.02682772  0.01709213 -0.01715424  0.0001180099
abdom   0.65808293  0.38431916 -0.54918684 -0.32733587  0.03614857 -0.05798794  0.04237543  0.04519991  0.04429945 -0.0004647401
hip     0.41956706 -0.50864092 -0.17543589  0.52422020  0.37503871  0.33497539 -0.06370526  0.03017756  0.04125882  0.0113777686
thigh   0.27968753 -0.59999647  0.01776404 -0.21848860 -0.67577523 -0.18155272 -0.10850160 -0.01559684  0.08899493 -0.0644730639
knee    0.12148556 -0.17468550  0.04381024  0.01090359  0.20196703 -0.52217519  0.17184353  0.08891511 -0.76781015  0.1107333685
ankle   0.05596265 -0.11532160  0.10008929  0.05672153  0.28765471 -0.60340059  0.35115626 -0.13471329  0.60725092  0.1152577102
biceps  0.14540629 -0.18341990  0.33968251 -0.51136755  0.18245861  0.44115762  0.55435680 -0.16739424 -0.06692188  0.0247215390
forearm 0.07391475 -0.08818365  0.29297576 -0.33309611  0.29020582 -0.03332036 -0.36513174  0.73825438  0.15389409  0.0511729735
wrist   0.03934804 -0.01420681  0.07867510 -0.05144840  0.18141400 -0.12022251 -0.06367487 -0.09699713 -0.02172551 -0.9633866143
> dim(prfat$x)  # principal components are found in prfat$x
[1] 252  10
> summary(prfat)  #the first principal component explains 0.867 of the variation in the predictor data
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     15.990 4.06584 2.96596 2.00044 1.69408 1.49881 1.30322 1.25478 1.10955 0.52737
Proportion of Variance  0.867 0.05605 0.02983 0.01357 0.00973 0.00762 0.00576 0.00534 0.00417 0.00094
Cumulative Proportion   0.867 0.92304 0.95287 0.96644 0.97617 0.98378 0.98954 0.99488 0.99906 1.00000
>
> round(prfat$rot[,1],2)  #linear combination describing the first principal component
   neck   chest   abdom     hip   thigh    knee   ankle  biceps forearm   wrist
   0.12    0.50    0.66    0.42    0.28    0.12    0.06    0.15    0.07    0.04
> #chest, abdomen, hip, and thigh measures dominate the first principal component
> #however, the data are not normalized so the result may be due to larger measures for these variables
>
> #should center and scale prior to performing PCA  - subtract out the mean and divide by the std dev
> prfatc <- prcomp(cfat,scale=TRUE)
> summary(prfatc)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9   PC10
Standard deviation     2.6498 0.85301 0.81909 0.70114 0.54708 0.52831 0.45196 0.40539 0.27827 0.2530
Proportion of Variance 0.7021 0.07276 0.06709 0.04916 0.02993 0.02791 0.02043 0.01643 0.00774 0.0064
Cumulative Proportion  0.7021 0.77490 0.84199 0.89115 0.92108 0.94899 0.96942 0.98586 0.99360 1.0000
>
> round(prfatc$rot[,1],2) #describe what this principal component measures
   neck   chest   abdom     hip   thigh    knee   ankle  biceps forearm   wrist
   0.33    0.34    0.33    0.35    0.33    0.33    0.25    0.32    0.27    0.30
>
> round(prfatc$rot[,2],2) #describe what this principal component measures
   neck   chest   abdom     hip   thigh    knee   ankle  biceps forearm   wrist
   0.00   -0.27   -0.40   -0.25   -0.19    0.02    0.62    0.02    0.36    0.38
>
> #note that the first principal component is orthogonal to the second principal component
> t(prfatc$rot[,1]) %*% prfatc$rot[,2]
              [,1]
[1,] -9.714451e-17
>
> #principal components analysis can be very sensitive to outliers
> #so need to check for outliers - use Mahalanobis which is a measure of the distance of a point
> #from the mean that adjusts for the correlation in the data.
```

```
>
> #Mahalanobis distance di=sqrt[(x-mu)'(sigma-1)(x-mu)]  , sigma is a measure of covariance
> #Use robust measures of center and covariance – these are provided by the cov.rob() function from the MASS package
> #If the data are multivariate normal with dimension m, then we expect d2 to follow a chi2 distribution with m df
> #Remove outliers or use robust PCA methods
>
> robfat <- cov.rob(cfat)
> md <- mahalanobis(cfat, center=robfat$center, cov=robfat$cov)
> n <- nrow(cfat);p <- ncol(cfat)
> plot(qchisq(1:n/(n+1),p), sort(md), xlab=expression(paste(chi^2,"quantiles")), ylab="Sorted Mahalanobis distances")
> abline(0,1)
>
> #now link the predictors to the response in a regression model using PCA
> #instead of using the predictors in their original form use the principal components - known as Principal Component Regression or PCR
>
> #model the percentage of body fat that is described by the variable Brozek.
> lmoda <- lm(fat$brozek ~ ., data=cfat)  #the '.' after '~' indicates to include all 10 predictors in the data set
> summary(lmoda)

Call:
lm(formula = fat$brozek ~ ., data = cfat)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.3159 -2.7435 -0.1584  2.8388 10.5150 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.228749   6.214309   1.163  0.24588    
neck        -0.581947   0.208580  -2.790  0.00569 ** 
chest       -0.090847   0.085430  -1.063  0.28866    
abdom        0.960229   0.071582  13.414  < 2e-16 ***
hip         -0.391355   0.112686  -3.473  0.00061 ***
thigh        0.133708   0.124922   1.070  0.28554    
knee        -0.094055   0.212394  -0.443  0.65828    
ankle        0.004222   0.203175   0.021  0.98344    
biceps       0.111196   0.159118   0.699  0.48533    
forearm      0.344536   0.185511   1.857  0.06450 .  
wrist       -1.353472   0.471410  -2.871  0.00445 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.071 on 241 degrees of freedom
Multiple R-squared:  0.7351,    Adjusted R-squared:  0.7241 
F-statistic: 66.87 on 10 and 241 DF,  p-value: < 2.2e-16

>
> #these regression results are hard to interpret because there is indication of collinearity.
> vif(lmoda)
    neck    chest    abdom      hip    thigh     knee    ankle   biceps  forearm    wrist 
3.893022 7.854662 9.021878 9.868755 6.513178 3.973479 1.795683 3.499619 2.127861 2.932970 
> #abdomen circumference has a positive effect while hip circumference has a negative effect??
>
> #now regress on the first two principal components
```

```
> lmodpcr <- lm(fat$brozek ~ prfatc$x[,1:2])
> summary(lmodpcr)

Call:
lm(formula = fat$brozek ~ prfatc$x[, 1:2])

Residuals:
     Min       1Q   Median       3Q      Max 
-17.6966  -3.6115  -0.1938   3.4381  20.8732 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         18.9385     0.3291  57.542   <2e-16 ***
prfatc$x[, 1:2]PC1   1.8420     0.1245  14.800   <2e-16 ***
prfatc$x[, 1:2]PC2  -3.5505     0.3866  -9.184   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.225 on 249 degrees of freedom
Multiple R-squared:  0.5492,	Adjusted R-squared:  0.5456 
F-statistic: 151.7 on 2 and 249 DF,  p-value: < 2.2e-16

> #R2 is lower so lose some predictive power
> #two PCs are orthogonal.
> 
> #recall the first two PCs
> round(prfat$rot[,1],2)
   neck   chest   abdom     hip   thigh    knee   ankle  biceps forearm   wrist 
   0.12    0.50    0.66    0.42    0.28    0.12    0.06    0.15    0.07    0.04 
> round(prfat$rot[,2],2)
   neck   chest   abdom     hip   thigh    knee   ankle  biceps forearm   wrist 
  -0.02    0.38    0.38   -0.51   -0.60   -0.17   -0.12   -0.18   -0.09   -0.01 
> 
> #since the first two PCs still require measuring all 10 circumference variables, examine the first two PCs for possible variables
> 
> lmodr <- lm(fat$brozek ~ scale(abdom) + I(scale(ankle)-scale(abdom)), data=cfat)
> summary(lmodr)

Call:
lm(formula = fat$brozek ~ scale(abdom) + I(scale(ankle) - scale(abdom)), 
    data = cfat)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.134  -3.390  -0.074   3.107  14.873 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    18.9385     0.2794  67.789  < 2e-16 ***
scale(abdom)                    5.7629     0.3284  17.548  < 2e-16 ***
I(scale(ankle) - scale(abdom))  -0.9950     0.3140  -3.169  0.00172 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.435 on 249 degrees of freedom
Multiple R-squared:  0.6752,    Adjusted R-squared:  0.6726
F-statistic: 258.8 on 2 and 249 DF,  p-value: < 2.2e-16

>
> #We have a simple model that fits almost as well as the ten-predictor model
>
> #Adjusted R-squared:  0.7241  all 10 circumference measures
> #Adjusted R-squared:  0.6726  abdom, ankle
>
> ##-------------------------THE END-----------------------##
>
>
> dev.off()
windows
      2
>
```