#Biostatistics with R, BIOL597
#John Armitage Graduate Project, Spring 2014

#Prototyping Principal Components Analysis (PCA)

#PCA is a multivariate method of ordination. Ordination is the process of examining data trends and groupings, as opposed to hypothesis testing. Imagine plotting all measured variables onto a multi-dimensional grid, and rotating the grid so that the most "influential" variables (those which best explain variance) are plotted against one another, in order of decreasing influence. Typically, the two most influential variables are primarily dealt with. The axes along which these variables are plotted are called eigenvectors, and the variance describe by each eigenvector is an eigenvalue.

#I have chosen to prototype using the iris dataset, included with R

```
>iris
```

#calling the variable to be sure data displays correctly

```
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1           5.1         3.5          1.4         0.2     setosa
2           4.9         3.0          1.4         0.2     setosa
3           4.7         3.2          1.3         0.2     setosa
4           4.6         3.1          1.5         0.2     setosa
5           5.0         3.6          1.4         0.2     setosa
...
150         5.9         3.0          5.1         1.8  virginica
```

#actual data reduced here to save space

#we must remove the factor variable Species, and create an object to which we can apply functions

```
>irisPCA=iris[-5]
>attach(irisPCA)
>irisPCA
   Sepal.Length Sepal.Width Petal.Length Petal.Width
1           5.1         3.5          1.4         0.2
2           4.9         3.0          1.4         0.2
3           4.7         3.2          1.3         0.2
4           4.6         3.1          1.5         0.2
```

5 5.0 3.6 1.4 0.2

…

150 5.9 3.0 5.1 1.8

#the data displayed correctly, with Species excluded. This means we are now dealing with only 4 variables.

#next, calculate a covariance matrix, looking at the covariance between each variable, NOTE: it does not matter here if you use the raw data or mean centered data, the variance is the same

```
> cov(irisPCA,irisPCA)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

#looks good, let's create a variable we can call

```
> covmatrix=cov(irisPCA,irisPCA)
```

#Now we need to find the eignevalues and eigenvectors from this matrix

```
> eigen(covmatrix)
$values
[1] 4.22824171 0.24267075 0.07820950 0.02383509

$vectors
            [,1]       [,2]       [,3]      [,4]
[1,]  0.36138659 -0.65658877 -0.58202985  0.3154872
[2,] -0.08452251 -0.73016143  0.59791083 -0.3197231
[3,]  0.85667061  0.17337266  0.07623608 -0.4798390
[4,]  0.35828920  0.07548102  0.54583143  0.7536574
```

#Now we need to order eigenvalues from highest to lowest. By chance, they are already in descending order in the output

#Choose how many eigenvectors you want to keep. I have chosen to keep all eigenvectors, for the most accurate results

```
> eigen(covmatrix)$vectors
           [,1]       [,2]       [,3]      [,4]
[1,]  0.36138659 -0.65658877 -0.58202985  0.3154872
[2,] -0.08452251 -0.73016143  0.59791083 -0.3197231
[3,]  0.85667061  0.17337266  0.07623608 -0.4798390
[4,]  0.35828920  0.07548102  0.54583143  0.7536574
```

#let's set this to a variable we can call

```
> eigenvec=eigen(covmatrix)$vectors
```

#for later, we can also create a variable which will only deal with eigenvalues

```
> eigenval=eigen(covmatrix)$values
```

#Now we calculate the principal component scores by multiplying the data by the transposed ranked eigenvectors. In order to transpose a matrix (which means switching the columns and rows), use the function t()

```
> PCS=t(eigenvec)*irisPCA
> PCS
    Sepal.Length Sepal.Width Petal.Length  Petal.Width
1     1.8430716   1.2540122   1.19933885 -0.016904503
2    -3.2172850   0.2264431   0.24272173 -0.146032287
3    -0.3972558   1.1564371   0.46577596  0.171334121
4    -3.3587426  -2.0354252   0.11322153  0.034674533
5     4.2833530  -0.3042811   0.50594123  0.071657839
…
150   1.0228987  -2.1904843  -3.34860273  0.135865836
```

#Next, calculate the correlation of the original variables with the principal components (called the componenet scores)

```
> compscores=((eigenvec*sqrt(eigenval))/(sqrt(irisPCA)))
> compscores
    Sepal.Length  Sepal.Width Petal.Length  Petal.Width
1   0.329053754  0.025916500 -1.141062246  0.535708605
2  -0.018809752  0.006727993 -0.303993003  0.123687879
```

3 0.110508247 0.415410002 0.042524449 -3.018966934

4 0.025790706 -0.023648319 0.009514820 -0.804289886

5 -0.603793387 0.126267729 0.628040890 0.108416498

…

150 -0.148081771 0.031936073 -0.018437245 0.008685802

#determining the amount of variance described by each eigenvalue

> FE=eigenval/sum(eigenval)

> FE

[1] 0.924618723 0.053066483 0.017102610 0.005212184

#This means that about 92% of variation is explained by the first eigenvalue, which is further illustrated by a Scree Plot

#

#The entire process can be streamlined by running the function prcomp()….

> PCS=prcomp(irisPCA,retx=TRUE,center=TRUE,scale.=FALSE)

> PCS

Standard deviations:

[1] 2.0562689 0.4926162 0.2796596 0.1543862

Rotation:

	PC1	PC2	PC3	PC4
Sepal.Length	0.36138659	-0.65658877	0.58202985	0.3154872
Sepal.Width	-0.08452251	-0.73016143	-0.59791083	-0.3197231
Petal.Length	0.85667061	0.17337266	-0.07623608	-0.4798390
Petal.Width	0.35828920	0.07548102	-0.54583143	0.7536574

> PCS$x

	PC1	PC2	PC3	PC4
[1,]	-2.684125626	-0.319397247	0.027914828	0.0022624371
[2,]	-2.714141687	0.177001225	0.210464272	0.0990265503

[3,] -2.888990569 0.144949426 -0.017900256 0.0199683897

[4,] -2.745342856 0.318298979 -0.031559374 -0.0755758166

[5,] -2.728716537 -0.326754513 -0.090079241 -0.0612585926

…

[150,] 1.390188862 0.282660938 -0.362909648 -0.1550386

> PCS$center

Sepal.Length Sepal.Width Petal.Length Petal.Width

5.843333 3.057333 3.758000 1.199333

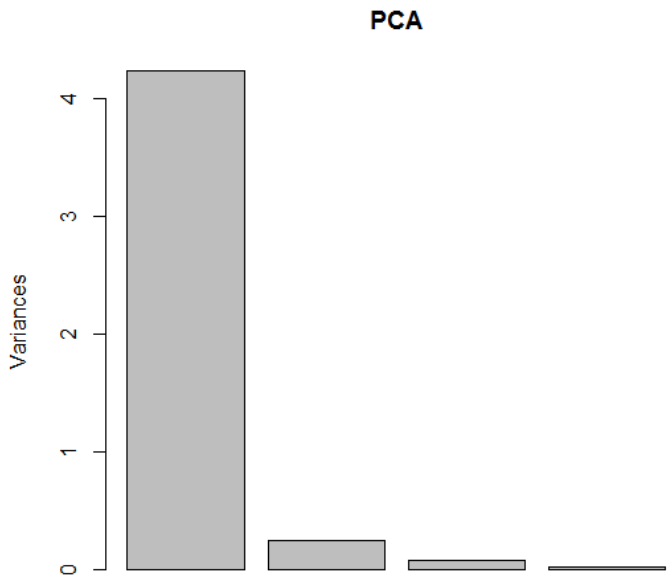
#Unfortunately my calculated principal components scores do not match those of the prcomp() function, indicating an error on at least one party's computations

##########

#INTERPRETATION OF PCA

#The first type of plot to examine is the scree plot, which will graphically demonstrate which will graphically illustrate the variance explained by eigenvectors as a percentage

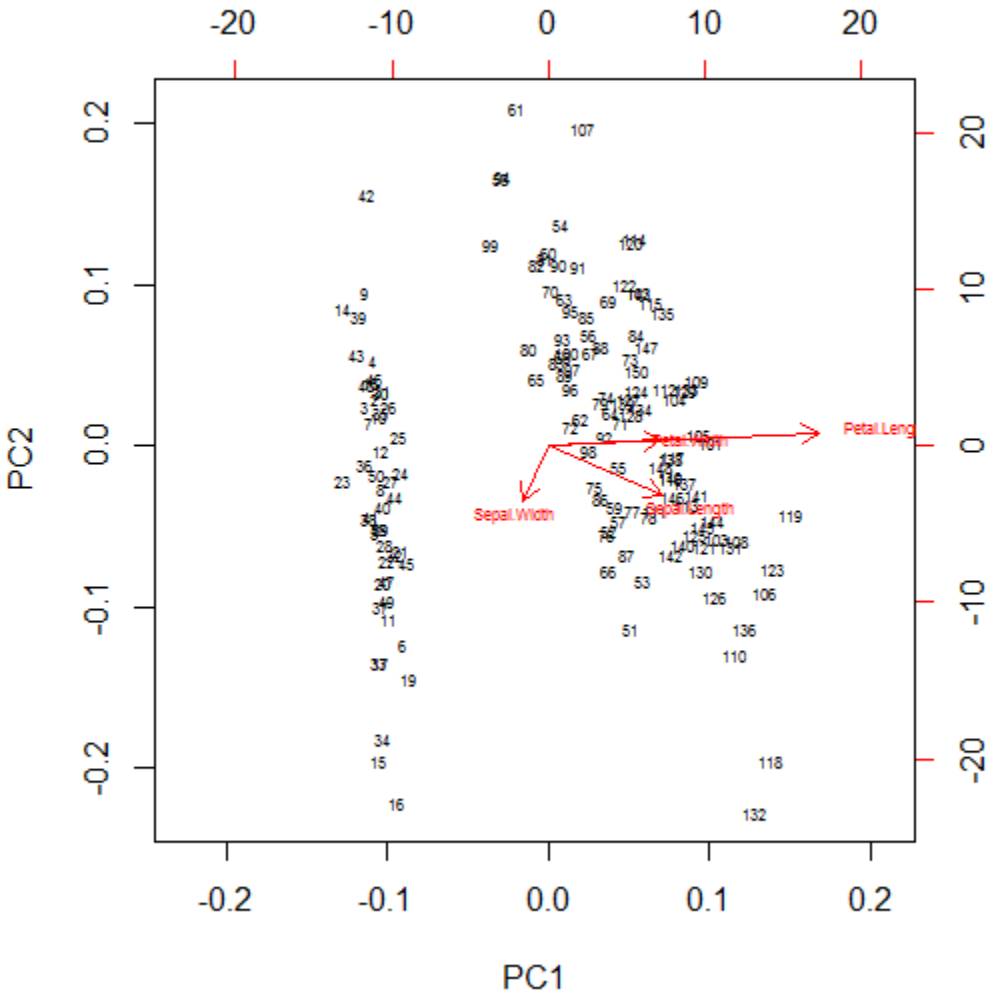
>plot(PCA)



#Clearly one eigenvector dominates

#Next is the biplot…

> biplot(PCS,choices=c(1,2),cex=0.5)



#This specific plot uses principal component 1 (PC1) and principal component 2 (PC2) as axes. This data has been mean-centered (but not scaled), which is why the vectors begin at zero. This is called a biplot because the variables and the individual data points are both plotted.

#The axes represent recombinations of the original variables-these are the principal components. The 4-dimensional space in which sepal width, sepal length, petal width, and petal length existed has been restructured in way that best shows their relationships. The principal components represent a warped space of orthogonal eigenvectors. You will notice that the lowest values of data are clustered to the left, at -0.1 along the PC1 axis. These values correspond the species *Iris Setosa*, indicating that this species is quite distinct from the other two. There is very little clustering occurring along the vertical PC2 axis. This makes sense, as our scree plot indicated that PC1 explained nearly all variance between individuals.

#The red arrows represent the variables as vectors. The length of the vector is proportional to its influence on variability within data. Petal length has the longest vector, meaning it contributes most to variability. The direction of the vector indicates the variable's influence of principal components. As you can see, petal length and petal width have almost no vertical aspect to their slope, indicating they primarily comprise PC1. Sepal length and sepal width have short lengths, meaning they have little impact on variability between specimens. Additionally, they have some vertical aspect, meaning they help to define PC2 more than petal length and petal width.

#Special thanks to:

#Dr. Stein for providing the following link
http://biotoolbox.binghamton.edu/Multivariate%20Methods/Multivariate%20Ordinations/pdf%20files/010-2012%20MVO%20010.pdf

#Lindsay I Smith for uploading the following free PCA tutorial
http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

#Brian S Everitt and Torsten Hothorn for uploading the following free PCA tutorial http://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_principal_components_analysis.pdf

#and the R community for their many posts, questions, answers, and more found online