Stat 428/528: Advanced Data Analysis 2

Chapters 13: Principal Component Analysis (PCA)

April 17, 2019

THE UNIVERSITY OF NEW MEXICO

PCA

Principal component analysis (PCA) is a multivariate technique for understanding variation, and for summarizing measurement data possibly through variable reduction.

- Principal components (the variables created in PCA) are sometimes used in addition to, or in place of, the original variables in certain analyses.
- Given data on p variables or features $X_1, X_2, \cdots, X_p$, PCA uses a rotation of the original coordinate axes to produce a new set of p uncorrelated variables, called principal components,
  ——-that are unit-length linear combinations of the original variables.
  —— A unit-length linear combination $a_1X_1 + a_2X_2 + \cdots + a_pX_p$ has $a_1^2 + a_2^2 + \cdots + a_p^2 = 1$.

The idea of principal components is to find linear combinations of variables that explain variation in the data.

Typically, we have a single sample and many variables, all of which are considered random.

Principal components is generally used more to describe the data rather than doing inference, and so doesnt assume that the data are multivariate normal, although the ideas are easier to visualize when the data is multivariate normal and 2-dimensional.

A crude example for the chile data, is that if we just looked at length and width, we might construct two new variables:
size = length+width
shape = length - width

- we've transformed the two variables of length and width into two new variables, size and shape.
  ——This doesnt reduce the dimensions of the data,
  ——but these two new variables might give a nice way to interpret the variation in the data,
  ——and they dont lose any of the information in the original data.
- Typically, with principal components, we transform *n* observations of *p* variables into *n* observations of a new set of *p* variables, where the new variables are linear combinations of the old variables.

The principal components have the following properties.

- The first principal component

$$PRIN1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

  has the largest variability among all unit-length linear combinations of the original variables.

- The second principal component

$$PRIN2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

  has the largest variability among all unit-length linear combinations of $X_1, X_2, \cdots, X_p$ that are uncorrelated with PRIN1

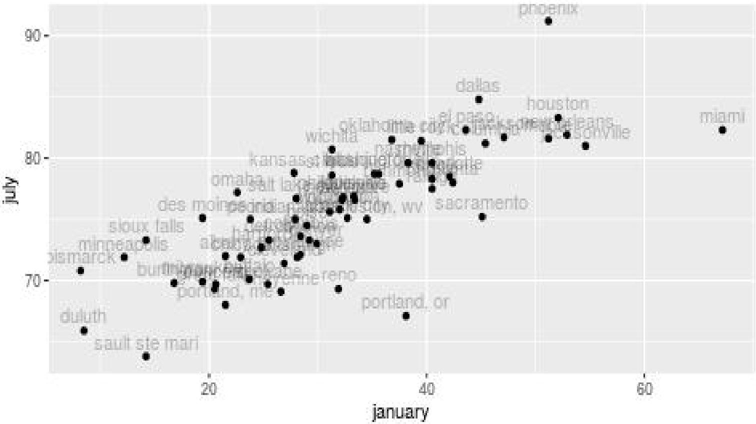
- In general, the j th principal component PRINj for $j = 1, 2, \cdots, p$, has the largest variability among all unit- length linear combinations of the features that are uncorrelated with PRIN1, PRIN2, . . . , PRIN(j-1).
- The last or p th principal component PRINp has the smallest variability among all unit-length linear combinations of the features.

Example: Temperature Data

The following temperature example includes mean monthly temperatures in January and July for 64 U.S. cities.

```
> head(temp)
         city january july id
1      mobile    51.2 81.6  1
2     phoenix    51.2 91.2  2
3 little rock    39.5 81.4  3
4  sacramento    45.1 75.2  4
5      denver    29.9 73.0  5
6    hartford    24.8 72.7  6
```

Mean temperature in Jan and July for selected cities



The princomp() procedure is used for PCA.

- By default the principal components are computed based on the covariance matrix.
- The correlation matrix may also be used with the cor = TRUE option.
- The principal component scores are the values of the principal components across cases.
- The principal component scores PRIN1, PRIN2, . . . , PRINp are centered to have mean zero.

```
> # perform PCA on covariance matrix
> temp.pca <- princomp( ~ january + july, data = temp)
> # standard deviation and proportion of variation for
each component
> summary(temp.pca)
Importance of components:
                          Comp.1    Comp.2
Standard deviation     12.3217642 3.0004557
Proportion of Variance  0.9440228 0.0559772
Cumulative Proportion   0.9440228 1.0000000
> # coefficients for PCs
> loadings(temp.pca)

Loadings:
        Comp.1 Comp.2
january -0.939  0.343
july    -0.343 -0.939
```

```
               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
> # scores are coordinates of each observation on PC scale
> head(temp.pca$scores)
      Comp.1     Comp.2
1 -20.000106  0.9239612
2 -23.291460 -8.0941867
3  -8.940669 -2.8994977
4 -12.075589  4.8446790
5   2.957414  1.7000283
6   7.851160  0.2333138
```

PCA is effectively doing a location shift (to the origin, zero) and a rotation of the data.

When the correlation is used for PCA (instead of the covariance), it also performs a scaling so that the resulting PC scores have unit-variance in all directions.

Figure: Plot original data with PCA vectors overlayed

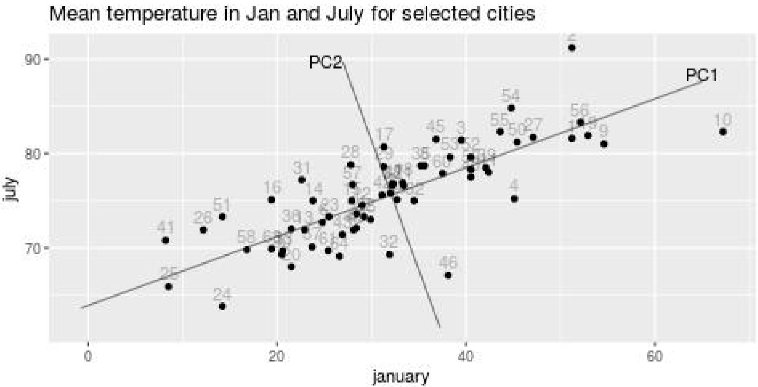
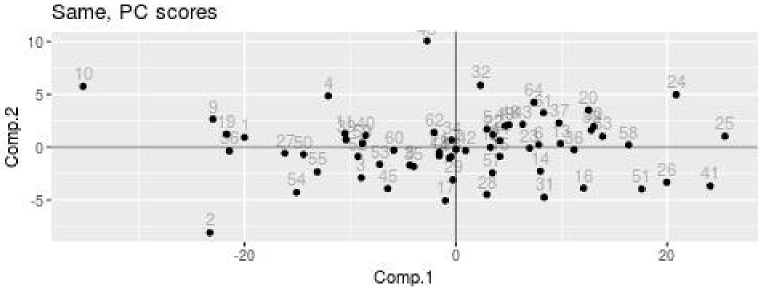


Figure: plot PCA scores (data on PC-scale centered at 0)



Some comments on the output:

1. You can visualize PCA when $p = 2$.
   ——-In the temperature plot, the direction of maximal variability corresponds to the first PC axis.
   —– The PRIN1 score for each city is obtained by projecting the temperature pairs perpendicularly onto this axis.
   ——The direction of minimum variation corresponds to the second PC axis, which is perpendicular to the first PC axis.
   ——The PRIN2 score for each city is obtained by projecting the temperature pairs onto this axis.

2. The total variance is the sum of variances for the monthly temperatures: $163.38 = 137.18 + 26.20$.

```
> # variance of data (on diagonals,covariance of off-diags
> var(temp[,c("january","july")])
         january     july
january 137.1811 46.72910
july     46.7291 26.20035
> # sum of variance
> sum(diag(var(temp[,c("january","july")])))
[1] 163.3814
> # variance of PC scores
> var(temp.pca$scores)
             Comp.1       Comp.2
Comp.1 1.542358e+02 1.831125e-15
Comp.2 1.831125e-15 9.145635e+00
> # sum is same as original data
> sum(diag(var(temp.pca$scores)))
[1] 163.3814
```

```
> eigen(var(temp[,c("january","july")]))
eigen() decomposition
$values
[1] 154.235808   9.145635

$vectors
           [,1]       [,2]
[1,] -0.9393904  0.3428493
[2,] -0.3428493 -0.9393904
```

3. The eigenvalues of the covariance matrix are variances for the PCs. The variability of
   $\text{PRIN1} = +0.939 \text{ JAN} + 0.343 \text{ JULY}$
   is 154.236.

   The variability of
   $\text{PRIN2} =$-0.343 JAN + 0.939 JULY
   is 9.146.

   The proportion of the total variability due to PRIN1 is $0.944 = 154.23/163.38$.

   The proportion of the total variability due to PRIN2 is $0.056 = 9.146/163.38$.
4. Almost all of the variability (94.4%) in the original temperatures is captured by the first PC. The second PC accounts for the remaining 5.6% of the total variability.

5. PRIN1 weights the January temperature about three times the July temperature. This is sensible because PRIN1 maximizes variation among linear combinations of the January and July temperatures. January temperatures are more variable, so they are weighted heavier in this linear combination.
6. The PCs PRIN1 and PRIN2 are standardized to have mean zero. This explains why some PRIN1 scores are negative, even though PRIN1 is a weighted average of the January and July temperatures, each of which is non-negative.

PCA on Correlation Matrix

The features are standardized to have mean zero and variance one by using the Z-score transformation: (Obs Mean)/Std Dev. The PCA is then performed on the standardized data.

```
temp.z <- temp
# manual z-score
temp.z$january <- (temp.z$january - mean(temp.z$january))
 / sd(temp.z$january)
# z-score using R function scale()
temp.z$july <- scale(temp.z$july)
# the manual z-score and scale() match
all.equal(temp.z$january, as.vector(scale(temp.z$january)))

# scale() includes attributes for the mean() and sd() used
 for z-scoring
str(temp.z)
head(temp.z)
```

```
> head(temp.z)
         city    january        july id
1      mobile  1.6311459  1.17005226  1
2     phoenix  1.6311459  3.04555476  2
3 little rock  0.6322075  1.13097929  3
4  sacramento  1.1103319 -0.08028274  4
5      denver -0.1874344 -0.51008540  5
6    hartford -0.6228691 -0.56869485  6
```

```
> # perform PCA on correlation matrix
> temp.pca2 <- princomp( ~ january + july, data = temp,
cor = TRUE)
> # standard deviation and proportion of variation for each
> summary(temp.pca2)
Importance of components:
                          Comp.1    Comp.2
Standard deviation     1.3339592 0.4696305
Proportion of Variance 0.8897236 0.1102764
Cumulative Proportion  0.8897236 1.0000000
> # coefficients for PCs
> loadings(temp.pca2)

Loadings:
        Comp.1 Comp.2
january  0.707 -0.707
july     0.707  0.707
```

```
               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
> # scores are coordinates of each observation on
PC scale
> head(temp.pca2$scores)
      Comp.1     Comp.2
1  1.9964045 -0.3286199
2  3.3330689  1.0080444
3  1.2566173  0.3554730
4  0.7341125 -0.8485470
5 -0.4971200 -0.2299523
6 -0.8492236  0.0386098
```

The standardized features are dimensionless, so the PCs are not influenced by the original units of measure, nor are they affected by the variability in the features.

The only important factor is the correlation between the features, which is not changed by standardization.

The PCs from the correlation matrix are
$\text{PRIN1} = +0.707\ \text{JAN} + 0.707\ \text{JULY}$
and
$\text{PRIN2} =$ - $0.707\ \text{JAN} + 0.707\ \text{JULY}$.

PCA is an exploratory tool, so neither a PCA on the covariance matrix nor a PCA on the correlation matrix is always the “right” method. You can do both and see which analysis is more informative.

Interpreting Principal Components

The coefficients or loadings in a principal component reflect the relative contribution of the features to the linear combination.
——Most researchers focus more on the signs of the coefficients than on the magnitude of the coefficients.
——The principal components are then interpreted as weighted averages or comparisons of weighted averages.

Example:
The difference $Z = X - Y$ is a comparison of X and Y .

- The sign and magnitude of Z indicates which of X and Y is larger, and by how much.
- $Z = 0$ if and only if $X = Y$, whereas $Z < 0$ when $X < Y$ and $Z > 0$ when $X > Y$.

In the temperature data, PRIN1 is a weighted average of January and July temperatures:
PRIN1 = +0.94 JAN + 0.34 JULY.
PRIN2 is a comparison of January and July temperatures:
PRIN2 = - 0.34 JAN + 0.94 JULY.

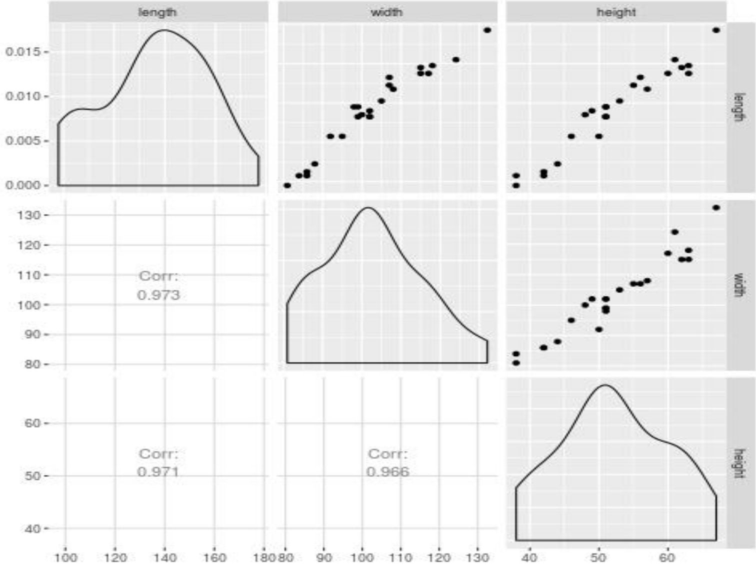
- Principal components often have positive and negative loadings when $p \geq 3$.
- To interpret the components, group the features with + and - signs together and then interpret the linear combination as a comparison of weighted averages.
- You can often simplify the interpretation of principal components by mentally eliminating features from the linear combination that have relatively small (in magnitude) loadings or coefficients. This strategy does not carry over to all multivariate analyses, so I will be careful about this issue when necessary.

Example: Painted turtle shells

Jolicouer and Mosimann gave the length, width, and height in mm of the carapace (shell) for a sample of 24 female painted turtles.

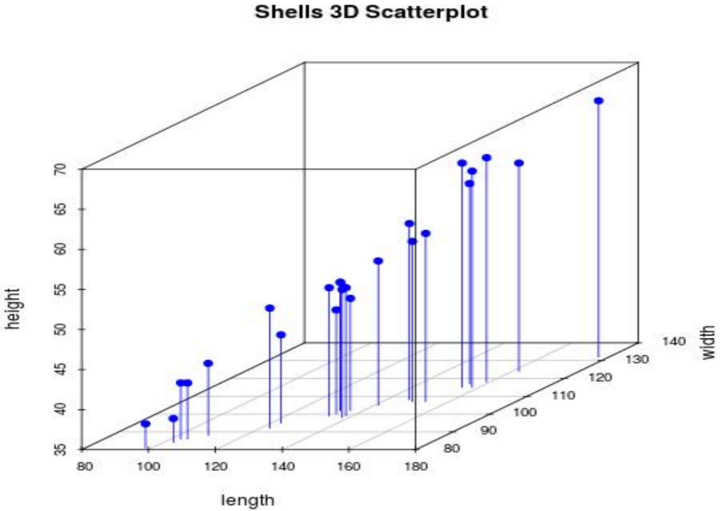
```
> head(shells)
  length width height
1     98    81     38
2    103    84     38
3    103    86     42
4    105    86     42
5    109    88     44
6    123    92     50
```

Figure: Scatterplot of painted turtle shells)



The plots show that the shell measurements are strongly positively correlated, which is not surprising.

Figure: Scatterplot of painted turtle shells)



PCA on shells using covariance matrix

```
> # perform PCA on covariance matrix
> shells.pca <- princomp( ~ length + width + height,
data = shells)
> # standard deviation and proportion of variation for each
component
> summary(shells.pca)
Importance of components:
                           Comp.1      Comp.2      Comp.3
Standard deviation     25.4970668 2.547081962 1.653745717
Proportion of Variance  0.9860122 0.009839832 0.004148005
Cumulative Proportion   0.9860122 0.995851995 1.000000000
> # coefficients for PCs
> loadings(shells.pca)

Loadings:
       Comp.1 Comp.2 Comp.3
length  0.814  0.555 -0.172
width   0.496 -0.818 -0.291
height  0.302 -0.151  0.941

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```

The three principal components from the raw data are given below.

- PRIN1 = 0.81 Length + 0.50 Width + 0.30 Height
  ——-a weighted average of the carapace measurements, and can be viewed as an overall measure of shell size.
  ——As PC1 increases, length, width and height increase
- PRIN2 = -0.55 Length + (0.82 Width + 0.15 Height)
  ——measures of shape, a comparison of length with an average of width and height
  ——As PC2 increases, width and height increase, while length decreases.
- PRIN3 = -(0.17 Length + 0.29 Width) + 0.94 Height
  ——- measures of shape, a comparison of height with length and width
  ——-As PC3 increases, height increase, while length and width decrease.

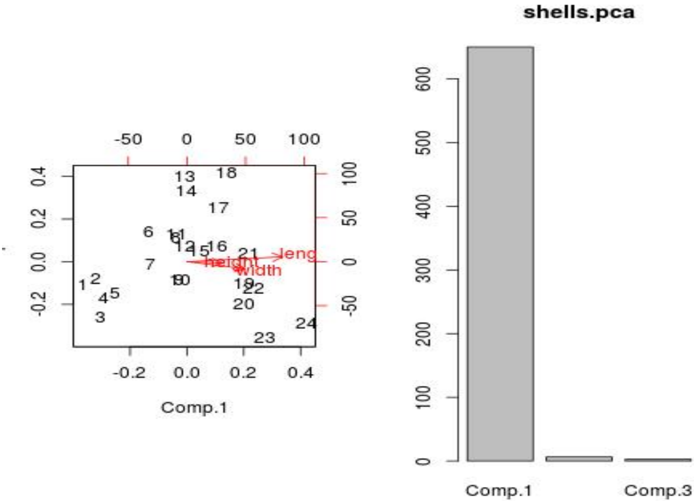
- Jolicouer and Mosimann argue that the size of female turtle shells can be characterized by PRIN1 with little loss of information
  —— because this linear com- bination accounts for 98.6% of the total variability in the measurements.
- The carapace measurements are positively correlated with each other, so larger lengths tend to occur with larger widths and heights.
  ——The primary way the shells vary is with regards to their overall size, as measured by a weighted average of length, width, and height.

How many components to keep

There are four typical criteria

- Keep enough principal components so that the proportion of variance explained meets a threshold, e.g., 80% or 90%.
- Keep components with larger than average eigenvalues, $\bar{\lambda} = \frac{1}{p} \sum_i \lambda_i$
- Use a scree graph, plotting λ_i against i, and see where there is a large break in the eigenvalues
- Test for significance of larger components

Figure: Scatterplot of painted turtle shells)

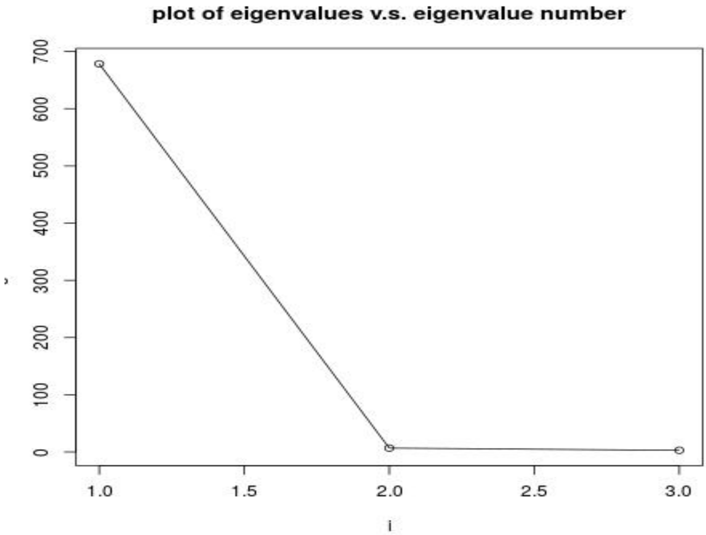


PRIN1 dominate the contribution in explaining variabilities.

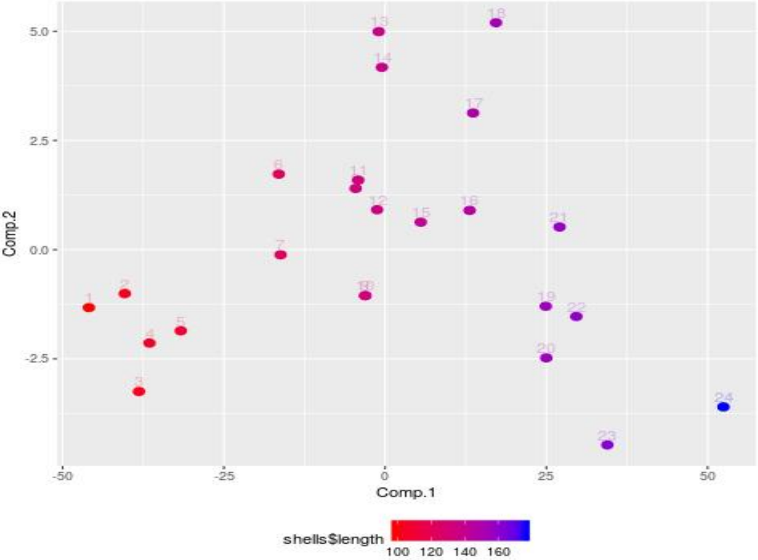
```
# eigenvalues and eigenvectors of covariance matrix
give PC variance and loadings
>eigen(var(shells[,c("length","width","height")]))
eigen() decomposition
$values
[1] 678.365651   6.769697   2.853783

$vectors
          [,1]       [,2]       [,3]
[1,] 0.8138808  0.5548963 -0.1723025
[2,] 0.4961059 -0.8180268 -0.2910518
[3,] 0.3024516 -0.1514012  0.9410636
```

$\bar{\lambda} = (678.37 + 6.77 + 2.85)/3 = 229.33$, only $\lambda_1 = 678.37 > \bar{\lambda}$, suggest keeping the first component.

Large break occurs at $i = 1$, keep one component.



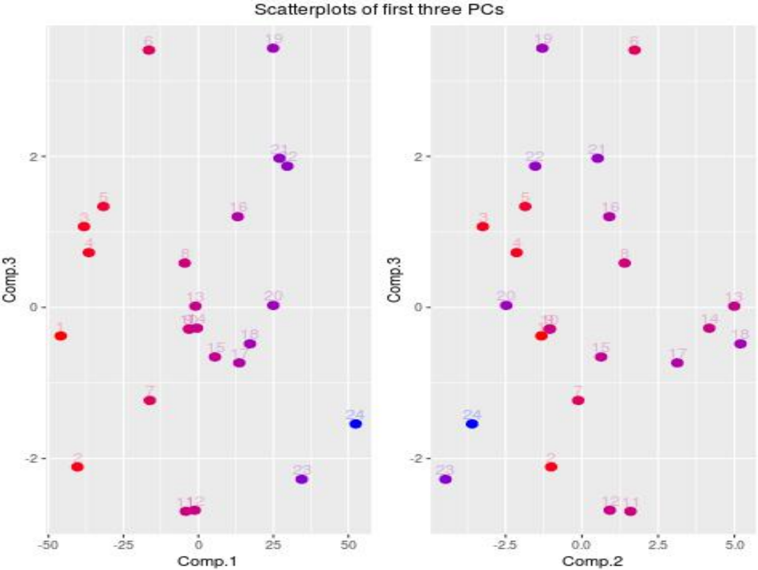
Two-dimensional plots of PC1 against PC2



- No. 24 obsn has the largest PC1, which means overall measure of shell size is large.
- No. 24 obsn has small PC2, indicating when length is large, average of width and height is small
- With increase of length, PC1 increase, while PC2 increase and then decrease.

```
> shells[24,]
   length width height id
24    177   132     67 24
> cbind(mean(shells$length), mean(shells$width),
mean(shells$height))
         [,1]     [,2]     [,3]
[1,] 136.0417 102.5833 52.04167
```

Two-dimensional plots of PC1 against PC3, and PC2 against PC3


Scatterplots of first three PCs
Comp.3
Comp.1
Comp.2