**4 - Principal Components Analysis (PCA)**

Principal Component Analysis is a dimension reduction technique that has many useful applications in data analysis/statistics. Dimension reduction is achieved by forming new variables which are linear combinations of the variables in the original data set. These linear combinations are chosen to account for as much of the original variance-covariance/correlation structure in the original variables as possible. The new variables formed by these linear combinations are uncorrelated which is desirable property.

**4.1 - Terminology and Concepts**The principal component is a linear combination of the original of the *p* variables in the data set.

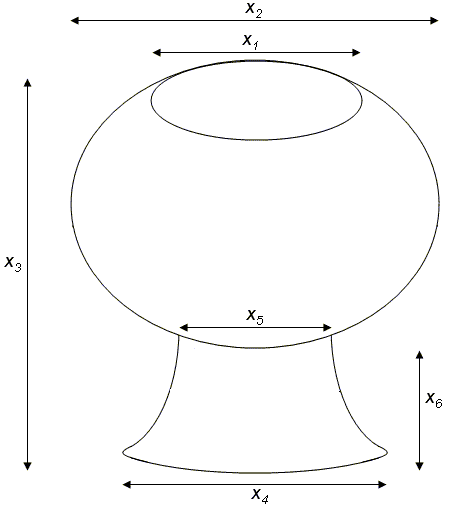
The coefficients are called the loadings for the principal component . Interpretation of the loadings can be an important aspect of the PCA. First we can gain important insight into how the original variables relate to one another from them. Also examining the loadings leads to our understanding what the scores (values) for the principal component are measuring.   
  
We hopefully can capture much of the information in the original variables with a much smaller number of principal components where *k << p*.

We can use the principal components in a number of ways:

* Visualize our data in a lower dimensional space
* Look for multivariate outliers. (There are variations of PCA that are specifically designed to help identify outliers in high dimensions).
* Look for clusters of similar observations
* Understand underlying structure in our data set.
* Use the principal components in subsequent analyses, e.g. regression.
* PCA can also be used as part of a process to impute (fill in) missing values in multivariate data set.

Example 4.1: Prehistoric Goblets

Measurements were made on 25 prehistoric goblets from Thailand (Professor C.F.W. Higham, University of Otago, as taken from Manly, B.F.J. 1986. *Multivariate Statistical Methods: A Primer.* Chapman and Hall, London, 159pp.) You have been asked to help organize these goblets according to their similarities. It is believed that different cultures will likely produce pottery with different to very different characteristics. The scientist has measured the mouth width (), total width (), total height (), base width (), stem width (), and stem height () on each of the 25 goblets.



> Goblets=read.table(file.choose(),header=T)

> Goblets

MouthWidth TotalWidth TotalHeight BaseWidth StemWidth StemHeight

1 13 21 23 14 7 8

2 14 14 24 19 5 9

3 19 23 24 20 6 12

4 17 18 16 16 11 8

5 19 20 16 16 10 7

6 12 20 24 17 6 9

7 12 19 22 16 6 10

8 12 22 25 15 7 7

9 11 15 17 11 6 5

10 11 13 14 11 7 4

11 12 20 25 18 5 12

12 13 21 23 15 9 8

13 12 15 19 12 5 6

14 13 22 26 17 7 10

15 14 22 26 15 7 9

16 14 19 20 17 5 10

17 15 16 15 15 9 7

18 19 21 20 16 9 10

19 12 20 26 16 7 10

20 17 20 27 18 6 14

21 13 20 27 17 6 9

22 9 9 10 7 4 3

23 8 8 7 5 2 2

24 9 9 8 4 2 2

25 12 19 27 18 5 12

Question: Do we need all six variables/dimensions to sufficiently describe the different characteristics of these goblets?

Answer: If no, then you have an opportunity to use principal components to reduce the dimension in the data while maintaining important characteristics in the data.

**4.2 - Mathematical/Statistical Concepts**

The principal components are particular linear combinations of the columns of your data matrix. Principal components depend solely on the variance/covariance matrix (or more generally on the correlation matrix) of the original *p* variables .

A principal component analysis does not necessarily rely on the multivariate normal distribution; however, if this is assumed a PCA definitely performs better.

Consider the following linear combinations of (the data matrix) that has variance/covariance matrix 



The variance/covariance term for each linear combination is given by



The principal components are defined to be the uncorrelated linear combinations that achieve maximum variances for . In particular,

|  |  |
| --- | --- |
| First Principal Component: | The linear combination that maximizes subject to the constraint of ,  i.e. the length of is 1. |
| Second Principal Component: | The linear combination  that maximizes subject to the constraint of  AND . |
| Third Principal Component: | The linear combination that maximizes subject to the constraint of   AND AND   etc.. |

Result 1:

Let be the sample variance/covariance matrix associated with (the data matrix). Let be the eigenvalues / eigenvectors of where it is assumed , then the principal components are given by

|  |  |  |
| --- | --- | --- |
|  | and |  |

and because the eigenvectors are orthogonal we have the desired covariance

result, namely for all . Also if we standardize the

variables first then , the sample correlation matrix.

In R

> gob.cor = cor(Goblets)

> gob.cor

MouthWidth TotalWidth TotalHeight BaseWidth StemWidth StemHeight

MouthWidth 1.0000000 0.6234051 0.3464089 0.6748429 0.6901040 0.5875703

TotalWidth 0.6234051 1.0000000 0.8392292 0.8287898 0.5807725 0.7970192

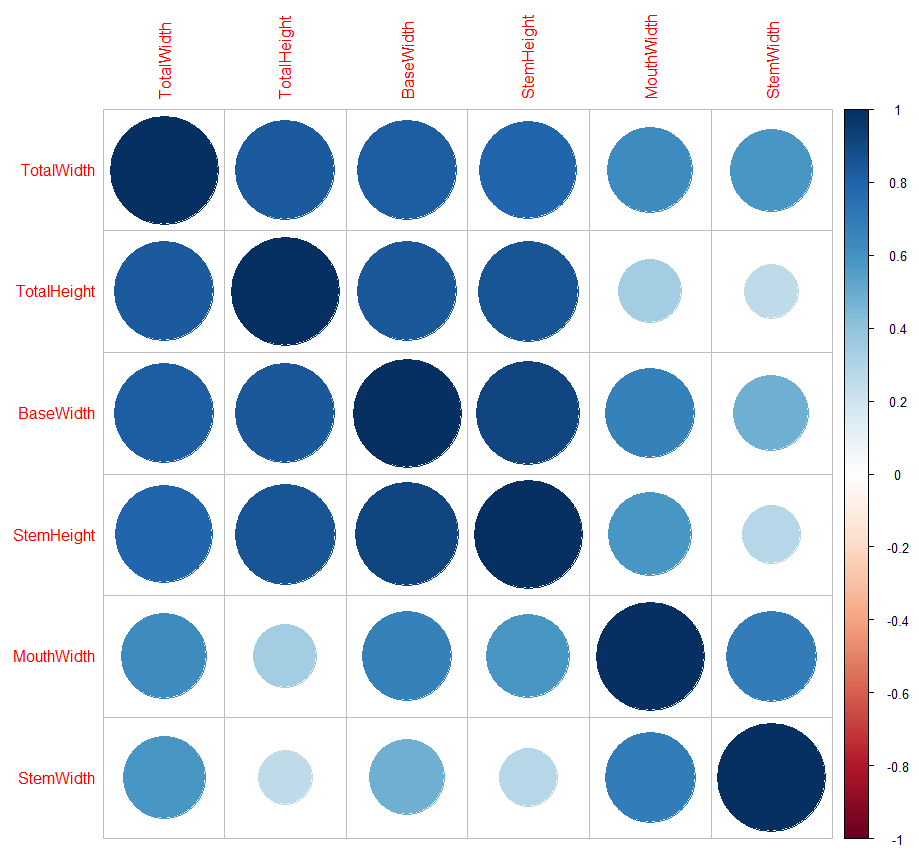
TotalHeight 0.3464089 0.8392292 1.0000000 0.8430518 0.2511584 0.8575089

BaseWidth 0.6748429 0.8287898 0.8430518 1.0000000 0.4874610 0.9101886

StemWidth 0.6901040 0.5807725 0.2511584 0.4874610 1.0000000 0.2885165

StemHeight 0.5875703 0.7970192 0.8575089 0.9101886 0.2885165 1.0000000

> corrplot(gob.corr,order=”hclust”)



Stem width and mouth width are moderately correlated and less correlated with the other four features.

Total width, total height, base width, and stem height are all moderately to strongly correlated with each other. Highlighted in yellow.

> gob.scale = scale(Goblets)

> var(Goblets)

MouthWidth TotalWidth TotalHeight BaseWidth StemWidth StemHeight

MouthWidth 9.043333 8.13000 6.330000 8.408333 4.478333 5.548333

TotalWidth 8.130000 18.80667 22.115000 14.891667 5.435000 10.853333

TotalHeight 6.330000 22.11500 36.923333 21.225000 3.293333 16.361667

BaseWidth 8.408333 14.89167 21.225000 17.166667 4.358333 11.841667

StemWidth 4.478333 5.43500 3.293333 4.358333 4.656667 1.955000

StemHeight 5.548333 10.85333 16.361667 11.841667 1.955000 9.860000

> var(gob.scale)

> var(gob.scale)

MouthWidth TotalWidth TotalHeight BaseWidth StemWidth StemHeight

MouthWidth 1.0000000 0.6234051 0.3464089 0.6748429 0.6901040 0.5875703

TotalWidth 0.6234051 1.0000000 0.8392292 0.8287898 0.5807725 0.7970192

TotalHeight 0.3464089 0.8392292 1.0000000 0.8430518 0.2511584 0.8575089

BaseWidth 0.6748429 0.8287898 0.8430518 1.0000000 0.4874610 0.9101886

StemWidth 0.6901040 0.5807725 0.2511584 0.4874610 1.0000000 0.2885165

StemHeight 0.5875703 0.7970192 0.8575089 0.9101886 0.2885165 1.0000000

> eigen.gob = eigen(gob.cor)

> eigen.gob

$values

[1] 4.27177774 1.09217742 0.38474775 0.14239633 0.06529366 0.04360709

$vectors

[,1] [,2] [,3] [,4] [,5] [,6]

[1,] -0.3660233 0.48592912 -0.6179335 0.32436829 -0.27835629 0.2556581

[2,] -0.4515367 -0.03412653 0.3752732 0.67427405 0.08391876 -0.4386709

[3,] -0.4111609 -0.44135161 0.3163501 -0.02019451 -0.38254463 0.6239630

[4,] -0.4618586 -0.11457532 -0.1588367 -0.54119094 -0.38182563 -0.5564635

[5,] -0.2963653 0.68277080 0.4914536 -0.35921044 0.22136144 0.1625790

[6,] -0.4381125 -0.29768029 -0.3324080 -0.13346207 0.75785442 0.1295892

> e1 = eigen.gob$vectors[,1]

> e1%\*%gob.cor%\*%e1

[,1]

[1,] 4.271778

> e2 = eigen.gob$vectors[,2]

> e2%\*%gob.cor%\*%e2

[,1]

[1,] 1.092177

> e1%\*%gob.cor%\*%e2

[,1]

[1,] 1.855039e-15

Etc…

Continuing with Example 4.1

Consider only the TotalWidth and TotalHeight from the goblet data.

|  |  |
| --- | --- |
|  |  |

Getting the variance/covariance matrix in R…

> X = cbind(TotalWidth,TotalHeight)

> var(X)

TotalWidth TotalHeight

TotalWidth 18.80667 22.11500

TotalHeight 22.11500 36.92333

Getting the eigenvalues/eigenvectors of the variance /covariance matrix in R…

> eigen(var(X))

$values

[1] 51.763256 3.966744

$vectors

[,1] [,2]

[1,] 0.5572085 -0.8303726

[2,] 0.8303726 0.5572085

|  |  |
| --- | --- |
| **Eigenvalues & Eigenvectors of** |  |

Result 2:

Let be the variance/covariance matrix associated with (the data matrix) and let be the eigenvalues / eigenvectors of where it is assumed .



Result 2 implies that  can be interpreted as the contribution to the total variance that is due to the 1st principal component (i.e. the linear combination of the original variables with maximal variance),  can be interpreted as the contribution to the total variance that is due to the 2nd principal component, etc.

Getting the linear combinations (i.e. scores) for Example 4.1.

> e1=eigen(var(X))$vectors[,1]

> e1

[1] 0.5572085 0.8303726

> e2=eigen(var(X))$vectors[,2]

> e2

[1] -0.8303726 0.5572085

> score1=X%\*%e1 🡨

> score2=X%\*%e2 🡨

|  |  |
| --- | --- |
| > cbind(score1,score2)  [,1] [,2]  [1,] 30.79995 -4.6220304  [2,] 27.72986 1.7477866  [3,] 32.74474 -5.7255671  [4,] 23.31571 -6.0313718  [5,] 24.43013 -7.6921171  [6,] 31.07311 -3.2344492  [7,] 28.85516 -3.5184936  [8,] 33.01790 -4.3379860  [9,] 22.47446 -2.9830454  [10,] 18.86893 -2.9939256  [11,] 31.90349 -2.6772408  [12,] 30.79995 -4.6220304  [13,] 24.13521 -1.8686285  [14,] 33.84828 -3.7807776  [15,] 33.84828 -3.7807776  [16,] 27.19441 -4.6329105  [17,] 21.37093 -4.9278350  [18,] 28.30883 -6.2936558  [19,] 32.73386 -2.1200323  [20,] 33.56423 -1.5628238  [21,] 33.56423 -1.5628238  [22,] 13.31860 -1.9012689  [23,] 10.27028 -2.7425217  [24,] 11.65786 -3.0156859  [25,] 33.00702 -0.7324512 |  |

These principal components are orthogonal (i.e. covariance = 0) and the total variance of these linear combinations is the same as the original data.

> round(var(cbind(score1,score2)),4)

[,1] [,2]

[1,] 51.7633 0.0000

[2,] 0.0000 3.9667

The proportion/percentage of the total variation due to the 1st principal component is . The remaining 7% is due to the 2nd principal component.

**Getting this done in R…**

There are several functions between base R and R packages that perform principal component analysis. We will first use the princomp function in base R.

> pca = princomp(X)

R returns the following…

> pca

Call:

princomp(x = X)

Standard deviations:

Comp.1 Comp.2

7.049307 1.951429

2 variables and 25 observations.

Additional output/results are accessible, but you need to know the names…



> names(pca) (or use attributes(pca))

[1] "sdev" "loadings" "center" "scale" "n.obs" "scores"

[7] "call"

The eigenvectors can be found under the loadings and can be accessed as follows:

> pca$loadings

Loadings:

Comp.1 Comp.2

TotalWidth 0.557 -0.830

TotalHeight 0.830 0.557

Comp.1 Comp.2

SS loadings 1.0 1.0

Proportion Var 0.5 0.5

Cumulative Var 0.5 1.0

The summary() function computes the proportion of variance due to each principal component automatically.

> summary(pca)

Importance of components:

Comp.1 Comp.2

Standard deviation 7.0493067 1.9514288

Proportion of Variance 0.9288221 0.0711779

Cumulative Proportion 0.9288221 1.0000000

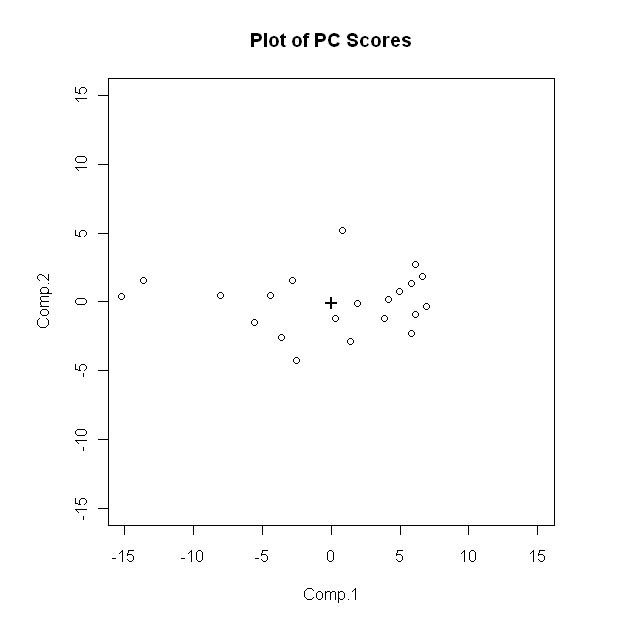
The actual linear combination values (i.e. PC scores) for the ***centered*** data can be obtained by asking for the scores…

|  |  |
| --- | --- |
| > pca$scores  Comp.1 Comp.2  [1,] 3.8865327 -1.19752382  [2,] 0.8164460 5.17229310  [3,] 5.8313223 -2.30106062  [4,] -3.5977011 -2.60686526  [5,] -2.4832842 -4.26761053  [6,] 4.1596969 0.19005729  [7,] 1.9417431 -0.09398703  [8,] 6.1044865 -0.91347950  [9,] -4.4389539 0.44146112  [10,] -8.0444888 0.43058096  [11,] 4.9900695 0.74726577  [12,] 3.8865327 -1.19752382  [13,] -2.7782087 1.55587808  [14,] 6.9348591 -0.35627103  [15,] 6.9348591 -0.35627103  [16,] 0.2809979 -1.20840399  [17,] -5.5424907 -1.50332847  [18,] 1.3954148 -2.86914926  [19,] 5.8204422 1.30447424  [20,] 6.6508148 1.86168272  [21,] 6.6508148 1.86168272  [22,] -13.5948132 1.52323759  [23,] -16.6431396 0.68198479  [24,] -15.2555585 0.40882063  [25,] 6.0936063 2.69205535 | R *centers* the data before running a PCA, which has no effect on the variance/covariance structure…    Thus, the scores are computed as follows…  1st observation scores for :  3.89 = 0.5572\*(21-17.84) + 0.8304\*(23-20.44)  = 0.5572\*3.16 + 0.8304\*2.56    -1.19 = -0.8304\*(21-17.84) + 0.5572\*(23-20.44)  = -0.8304\*3.16 + 0.5572\*2.56  2nd observation scores for :  0.82 = 0.5572\*(14-17.84) + 0.8304\*(24-20.44)  = 0.5572\*-3.84 + 0.8304\*3.56    5.17 = -0.8304\*(14-17.84) + 0.5572\*(24-20.44)  = -0.8304\*-3.84 + 0.5572\*3.56  etc… |

> plot(pca$scores,xlim=c(-15,15),ylim=c(-15,15))

> points(0,0,pch="+",cex=1.5)

> title("Plot of PC Scores")



The variance/covariance matrix of these linear combinations matches what we obtained above.

> round(var(pca$scores),4)

Comp.1 Comp.2

Comp.1 51.7633 0.0000

Comp.2 0.0000 3.9667