12 Introduction to Principal Component Analysis

PCA is a multivariate technique for understanding variation and for summarizing measurement data. It is frequently used for variable reduction.

Given data on p measurement variables $X_1, X_2, \cdots, X_p$, PCA produces a new set of p uncorrelated variables (the principal components) that are unit-length linear combinations of the original variables. That is,

$$\begin{aligned} PRIN1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ PRIN2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ PRINp &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

where unit length means $a_{j1}^2 + a_{j2}^2 + \cdots a_{jp}^2 = 1$ for all $j = 1, \cdots, p$.

The first principal component has the largest variability among all such possible linear combinations. The second has the largest variability among all such linear combinations which are uncorrelated with $PRIN1$. The third principal component has the largest variability among all such linear combinations that are uncorrelated with $PRIN1$ and $PRIN2$ and so forth down to $PRINp$. Thus, the ordered principal components are uncorrelated variables with progressively less variation (from $PRIN1$ to $PRINp$).

Example 1
Jolicouer and Moismann provided data on the height, length, and width of the shell for a sample of female painted turtles. Principal component analysis is used to identify the linear combinations of the measurements that account for the most of the variation in the size and shape of the shells.

```
> turtlesF = read.table("turtlesF.txt",header=T)
> turtlesF
   Length Width Height
1      98    81     38
2     103    84     38
3     103    86     42
4     105    86     40
5     109    88     44
6     123    92     50
7     123    95     46
8     133    99     51
9     133   102     51
10    133   102     51
11    134   100     48
12    136   102     49
13    137    98     51
14    138    99     51
15    141   105     53
16    147   108     57
17    149   107     55
18    153   107     56
19    155   115     63
20    155   117     60
21    158   115     62
22    159   118     63
23    162   124     61
24    177   132     67

> ss.pr1  = princomp(as.matrix(turtlesF), cor=T)
> names(ss.pr1)
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
[7] "call"
> ss.pr1$loadings[,1:3]
           Comp.1      Comp.2     Comp.3
Length -0.5783865 -0.06171004  0.8134254
Width  -0.5769696 -0.67396612 -0.4613846
Height -0.5766932  0.73618037 -0.3542081

> ss.pr1$sdev^2/sum(ss.pr1$sdev^2)
     Comp.1      Comp.2      Comp.3
0.980439578 0.011547360 0.008013062
```

```
> ss.pr1$scores[,1:3]
          Comp.1       Comp.2       Comp.3
 [1,]  3.0346502 -0.039113234 -0.091287510
 [2,]  2.7606815 -0.211555370 -0.003634567
 [3,]  2.3820993  0.051828467 -0.252844118
 [4,]  2.4707979 -0.138333171 -0.085985178
 [5,]  1.9809803  0.113182572 -0.178756523
 [6,]  0.9788082  0.414184328 -0.041002066
 [7,]  1.1325163 -0.111877875  0.028382794
 [8,]  0.3137378  0.108878178  0.054018963
 [9,]  0.1788134 -0.048728842 -0.053875852
[10,]  0.1788134 -0.048728842 -0.053875852
[11,]  0.4574286 -0.222965906  0.190123332
[12,]  0.2397031 -0.241857503  0.152092640
[13,]  0.2474771  0.149545757  0.246422107
[14,]  0.1746935  0.094043061  0.249566720
[15,] -0.3228981 -0.045844456  0.062465908
[16,] -0.9133083  0.147201568  0.011948728
[17,] -0.7796349  0.009575602  0.214772607
[18,] -0.9630285  0.089821304  0.326890894
[19,] -1.8835515  0.308398444 -0.192848939
[20,] -1.7570267 -0.073014290 -0.131819059
[21,] -1.8948200  0.207383578 -0.031200366
[22,] -2.1297114  0.138923331 -0.144305547
[23,] -2.3386705 -0.369419370 -0.154126685
[24,] -3.5485506 -0.281527330 -0.121122429
```

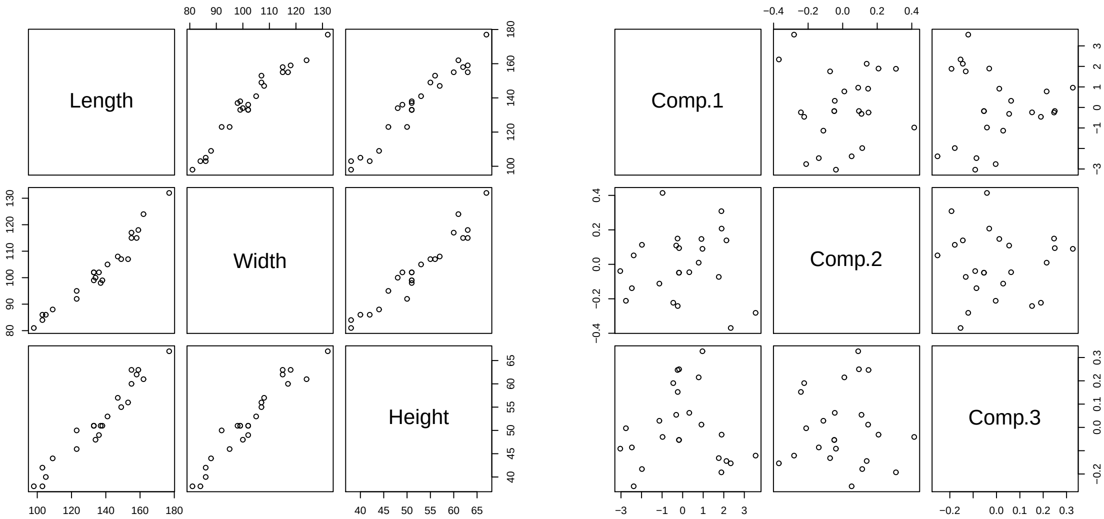


Figure 61: Turtle shell measurements data (left) and Principal Component Scores (right)

Example 2
The data file `socsupport` in the `DAAG` package consists of support measures, demographic information, and the Beck depression index (`BDI`) on a sample of healthy individuals. The variables in the data set are:

- `gender`: male or female
- `age`: 18-20, 21-24, 25-30, 31-40, 40+
- `country`: Australia, other
- `marital`: married, single, other
- `livewith`: alone, friends, parents, partner, residences, other
- `employment`: full-time, part-time, govt assistance, parental support, other
- `firstyr`: first year, other
- `enrolment`: full-time, part-time, blank
- `emotional`: availability of emotional support (5 questions)
- `emotionalsat`: satisfaction associated with available emotional support (5 questions)
- `tangible`: availability of tangible support (4 questions)
- `tangiblesat` associated satisfaction with tangible support (4 questions)
- `affect`: availability of affectionate support sources (3 questions)
- `affectsat`: associated satisfaction (3 questions)
- `psi`: availability of positive social interaction (3 questions)
- `psisat`: associated satisfaction (3 questions)
- `esupport`: extent of emotional support sources (4 questions)
- `psupport`: extent of practical support sources (4 questions)
- `supsources` extent of social support sources (4 questions)
- `BDI`: Score on the Beck depression index (total over 21 questions)

One study goal was to examine how the support measures (variables 9 - 19) may impact `BDI`. Other goals including looking at the impact of the demographic information in variables 1 - 8. We will focus on the support measures for this analysis.

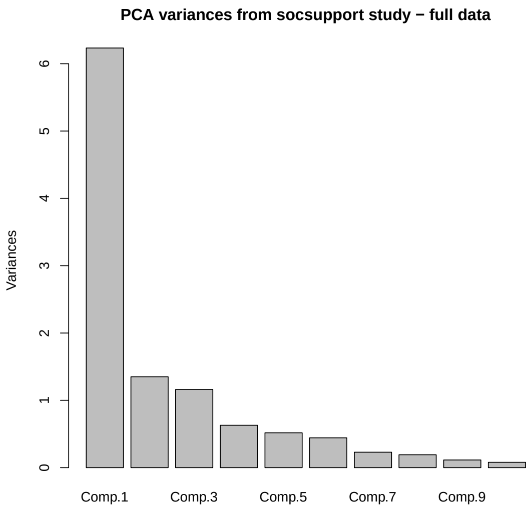
```
> library(DAAG)
> data(socsupport)
> not.na = complete.cases(socsupport)
> fullobs = socsupport[not.na,]
> fullobs[1:10,]
   gender   age   country marital   livewith         employment    firstyr
1    male 21-24 australia   other    partner employed part-time      other
2  female 21-24 australia  single    partner   parental support      other
3    male 21-24 australia  single residences employed part-time      other
4    male 18-20 australia  single    parents employed part-time first year
5  female 21-24 australia  single    friends employed part-time      other
6  female 21-24 australia  single    friends    govt assistance      other
7  female 25-30 australia married    partner employed part-time      other
8  female 25-30 australia married    partner employed part-time      other
10   male   40+ australia   other      alone employed part-time      other
11 female 21-24 australia  single    parents employed part-time      other
   enrolment emotional emotionalsat tangible tangiblesat affect affectsat psi
1  full-time        22           23       17          18     15        15  12
2  full-time        21           20       12          10     10         6   9
3  full-time        21           18       16          16     15        15  13
4  full-time        19           19       20          17     11        11  13
5  full-time        16           19       11          15      6        10  11
6  full-time        20           17       16          15     12        14  12
7  full-time        20           23       20          20     14        15  15
8  part-time        20           20       16          16     12        12  12
10 full-time        13           18        6          14      6        12   6
11 full-time        20           18       13          13     13        14  11
   psisat esupport psupport supsources BDI
1      13       13       11         13   5
2       6       12        7         10   8
3      12       14       13         14  16
4      12       15       15         15   0
5      12        9        7          9   9
6      11       13       12         13   0
7      15       15       10         13   1
8      12       13       11         11  14
10     11       10        8          9  20
11     12       12        8         14  13
```

```
> # correlations between emotional support measures
> cor(fullobs[,9:19])
             emotional emotionalsat  tangible tangiblesat    affect affectsat
emotional    1.0000000    0.8404097 0.4184066   0.4466863 0.6327119 0.5598617
emotionalsat 0.8404097    1.0000000 0.3005215   0.4700119 0.5551086 0.5916673
tangible     0.4184066    0.3005215 1.0000000   0.8457784 0.5244751 0.3417302
tangiblesat  0.4466863    0.4700119 0.8457784   1.0000000 0.5570396 0.4887023
affect       0.6327119    0.5551086 0.5244751   0.5570396 1.0000000 0.8590008
affectsat    0.5598617    0.5916673 0.3417302   0.4887023 0.8590008 1.0000000
psi          0.6522592    0.6269751 0.4723045   0.5451902 0.6150706 0.5769021
psisat       0.5808499    0.6544473 0.2950583   0.4784861 0.4835865 0.6241824
esupport     0.5648627    0.4207017 0.4115311   0.3890842 0.4179985 0.3122408
psupport     0.4116978    0.3389021 0.5248865   0.5141296 0.3175150 0.2874456
supsources   0.4666297    0.3913016 0.3779575   0.3977396 0.3537538 0.3704146
                   psi    psisat  esupport  psupport supsources
emotional    0.6522592 0.5808499 0.5648627 0.4116978  0.4666297
emotionalsat 0.6269751 0.6544473 0.4207017 0.3389021  0.3913016
tangible     0.4723045 0.2950583 0.4115311 0.5248865  0.3779575
tangiblesat  0.5451902 0.4784861 0.3890842 0.5141296  0.3977396
affect       0.6150706 0.4835865 0.4179985 0.3175150  0.3537538
affectsat    0.5769021 0.6241824 0.3122408 0.2874456  0.3704146
psi          1.0000000 0.8503953 0.6547506 0.5815234  0.5960882
psisat       0.8503953 1.0000000 0.5669334 0.5115678  0.5453711
esupport     0.6547506 0.5669334 1.0000000 0.5853292  0.6465774
psupport     0.5815234 0.5115678 0.5853292 1.0000000  0.7548660
supsources   0.5960882 0.5453711 0.6465774 0.7548660  1.0000000
```

```
> fit = lm(BDI~emotional + emotionalsat + tangible + tangiblesat + affect + affectsat
+ + psi + psisat + esupport + psupport + supsources,data=fullobs)
> summary(fit)

Call:
lm(formula = BDI ~ emotional + emotionalsat + tangible + tangiblesat +
    affect + affectsat + psi + psisat + esupport + psupport +
    supsources, data = fullobs)

Residuals:
    Min      1Q  Median      3Q     Max
-15.915  -5.678  -1.074   4.446  31.681

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.82174    7.22039   5.654 2.47e-07 ***
emotional     0.71521    0.58350   1.226   0.2240
emotionalsat -0.56894    0.65352  -0.871   0.3867
tangible     -0.24146    0.55588  -0.434   0.6652
tangiblesat   0.08959    0.72561   0.123   0.9021
affect       -1.37376    0.87331  -1.573   0.1198
affectsat     1.12655    0.88484   1.273   0.2067
psi          -0.36446    0.96499  -0.378   0.7067
psisat       -1.82040    1.02679  -1.773   0.0801 .
esupport      0.23530    0.57631   0.408   0.6842
psupport      0.77274    0.48958   1.578   0.1185
supsources   -1.09534    0.66225  -1.654   0.1022
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.66 on 78 degrees of freedom
Multiple R-squared: 0.3075,     Adjusted R-squared: 0.2098
F-statistic: 3.148 on 11 and 78 DF,  p-value: 0.001454
```

```
> ## Principal components regression
> ss.pr1 = princomp(fullobs[,9:19], cor=TRUE)
> ss.pr1$loadings

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
emotional    -0.321 -0.217 -0.153  0.508  0.299         0.208  0.147 -0.460
emotionalsat -0.304 -0.319 -0.192  0.487         0.318 -0.164 -0.152  0.414
tangible     -0.262  0.209  0.615  0.184        -0.121         0.222 -0.334
tangiblesat  -0.294         0.530  0.122 -0.306        -0.385 -0.147  0.229
affect       -0.307 -0.354  0.222 -0.312  0.393 -0.126  0.224         0.299
affectsat    -0.294 -0.427        -0.500  0.106  0.148 -0.165 -0.233 -0.252
psi          -0.351        -0.168        -0.333 -0.299  0.388  0.455  0.348
psisat       -0.324        -0.297 -0.168 -0.592 -0.114               -0.416
esupport     -0.289  0.294 -0.237         0.336 -0.655 -0.263 -0.372
psupport     -0.279  0.492        -0.114         0.408  0.524 -0.472
supsources   -0.284  0.401 -0.230 -0.221  0.259  0.373 -0.440  0.501
             Comp.10 Comp.11
emotional     0.437
emotionalsat -0.448
tangible     -0.540
tangiblesat   0.540
affect               -0.564
affectsat    -0.106   0.536
psi                   0.396
psisat               -0.469
esupport
psupport
supsources

> vars = (ss.pr1$sd^2)
> por.vars = vars/(sum(ss.pr1$sd^2))
> data.frame(variance = vars, portion = por.vars, cum.var = cumsum(por.vars))
          variance     portion   cum.var
Comp.1  6.23353658 0.566685144 0.5666851
Comp.2  1.35041288 0.122764808 0.6894500
Comp.3  1.16131376 0.105573978 0.7950239
Comp.4  0.62929371 0.057208519 0.8522324
Comp.5  0.51807128 0.047097389 0.8993298
Comp.6  0.44343136 0.040311942 0.9396418
Comp.7  0.22937251 0.020852046 0.9604938
Comp.8  0.19190191 0.017445628 0.9779395
```

```
Comp.9  0.11405360 0.010368509 0.9883080
Comp.10 0.07926632 0.007206030 0.9955140
Comp.11 0.04934609 0.004486008 1.0000000

> ss.pr1$loadings[,1] # all of the loadings of the first principal component
   emotional emotionalsat     tangible  tangiblesat       affect    affectsat
  -0.3213099   -0.3037267   -0.2615739   -0.2936168   -0.3070059   -0.2935072
         psi       psisat     esupport     psupport   supsources
  -0.3510474   -0.3237180   -0.2887159   -0.2786228   -0.2836617


> ss.pr1$loadings[,2] # all of the loadings of the second principal component
   emotional emotionalsat     tangible  tangiblesat       affect    affectsat
 -0.21706010  -0.31946007   0.20861591   0.08448239  -0.35380957  -0.42660133
         psi       psisat     esupport     psupport   supsources
  0.01462971  -0.06040658   0.29403945   0.49214145   0.40059581


> plot(ss.pr1,main="PCA variances from socsupport study - full data")
> pairs(ss.pr1$scores[, 1:3],main="full data")
```



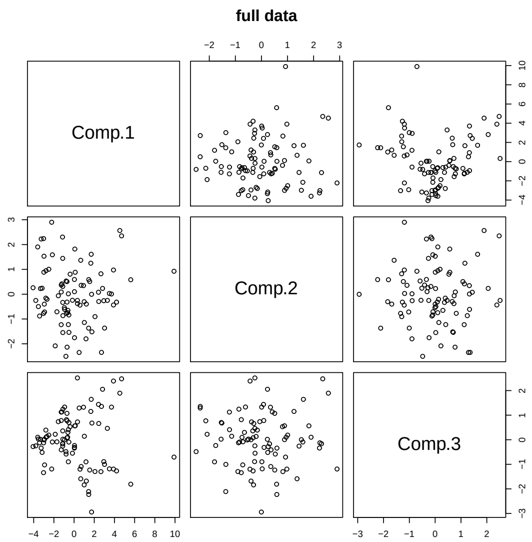


Figure 62: Scree plot for PCA of support measures (left); Plots of first three principal components (right) for suppport data.

```
> ss.lm = lm(BDI ~ ss.pr1$scores[, 1:6], data=fullobs)
> summary(ss.lm)

Call:
lm(formula = BDI ~ ss.pr1$scores[, 1:6], data = fullobs)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.559  -5.287  -0.200   3.689  34.900 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 10.8556     0.9205  11.793  < 2e-16 ***
ss.pr1$scores[, 1:6]Comp.1   1.7585     0.3687   4.770  7.8e-06 ***
ss.pr1$scores[, 1:6]Comp.2   0.5262     0.7921   0.664    0.508    
ss.pr1$scores[, 1:6]Comp.3   0.6218     0.8542   0.728    0.469    
ss.pr1$scores[, 1:6]Comp.4   1.1348     1.1604   0.978    0.331    
ss.pr1$scores[, 1:6]Comp.5   1.9708     1.2789   1.541    0.127    
ss.pr1$scores[, 1:6]Comp.6   1.1676     1.3824   0.845    0.401    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 8.733 on 83 degrees of freedom
Multiple R-squared: 0.2507,     Adjusted R-squared: 0.1965 
F-statistic: 4.627 on 6 and 83 DF,  p-value: 0.0004208
```

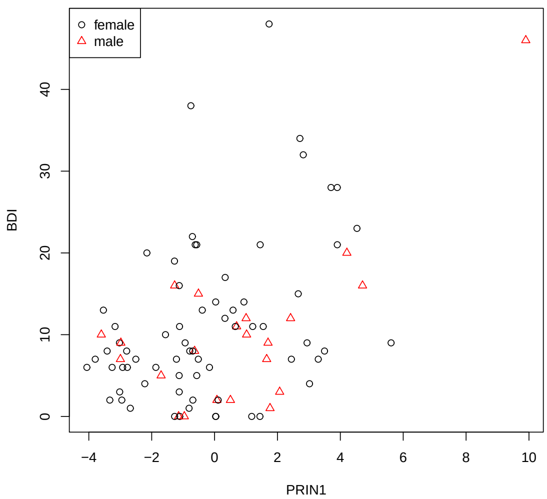


Figure 63: Plot of BDI vs. PRIN1 (bottom) for suppport data.