Homework 8 Solutions
401 Fall 2001
Brian Junker and Kimberly Sellers

Thursday December 6, 2001

1. RWG pp284-285.

   - Problem 4 page 285. Extract principal components, and construct a scree graph. How many components should we retain?

     Note: For this problem, I took the log of the first three variables and kept the last four variables as they were in the data file from the web-site. Next, generate the principal components on the standardized variables.

```
> streams.dat_read.table("../HW8/streams.dat", header=T)
> pc.streams_princomp(streams.dat, cor=T)
> streams2.dat_cbind(log(streams.dat[,1:3]), streams.dat[,4:7])
> summary(pc.streams)
Importance of components:
                          Comp. 1   Comp. 2   Comp. 3   Comp. 4   Comp. 5
    Standard deviation 1.6652019 1.4491818 0.9161872 0.8056163 0.6635273
Proportion of Variance 0.3961282 0.3000183 0.1199141 0.0927168 0.0628955
 Cumulative Proportion 0.3961282 0.6961465 0.8160606 0.9087774 0.9716729

                           Comp. 6     Comp. 7
    Standard deviation 0.36185086 0.259525523
Proportion of Variance 0.01870515 0.009621928
 Cumulative Proportion 0.99037807 1.000000000
```

     The above summary also verifies that the components are listed in order of descending variance, i.e. the first principal component explains the most variance, ... We can see the cumulative proportion from the summary of the princomp object above. We can also view the same information in a screeplot, see figure 1.

     From the above data, there are two possible cutoff points. One is at the bend of the elbow, which would include two principal components. The other is at the 90% mark, which would include four principal components. Because the elbow is relatively sharp, I will choose to use the first two principal components.

     The loadings are displayed below. I used the cutoff of 0.3 because it appeared to eliminate variables in each column while retaining variables of similar importance. When many of the loadings are around 0.5 (as is the case here), the cutoff point is more subjective.

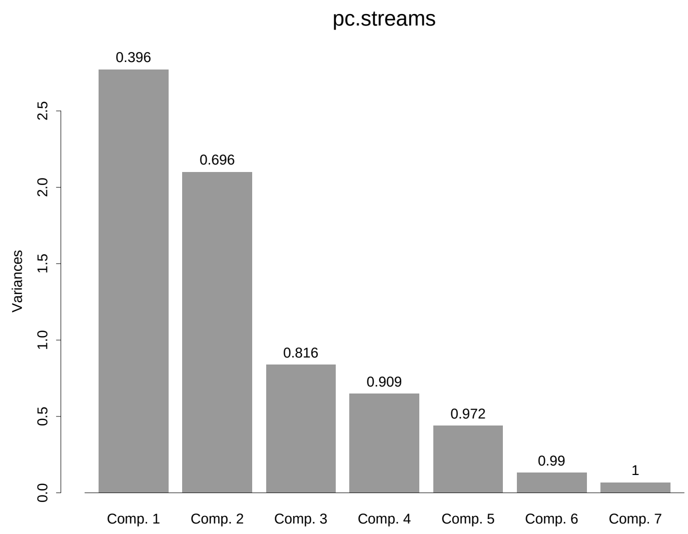

Figure 1: Screeplot for Streams data

```
> print(loadings(pc.streams), cutoff=0.3)
       Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6 Comp. 7
mdepth         -0.605           0.367           0.589
mwidth         -0.521   0.460           0.358  -0.453  -0.333
 mvelo  0.524                                           0.687
  silt -0.504                   0.460   0.423           0.479
leaves -0.435           0.631          -0.529
  sand  0.452                   0.552  -0.469  -0.439
   mud         -0.532  -0.482  -0.502          -0.362
```

From the above loadings, we see that the first principal component places negative loadings on % silt at the bottom and %leaves on the bottom, and positive loadings on mean velocity of the stream and % sand on the bottom. This component may be some measure of the velocity of the stream (putting positive loadings on faster streams). The second principal component places negative loadings on mean depth, mean width and % mud at the bottom of the stream. As is, this components is a bit difficult interpret.

- Problem 8.6 page 285. Perform a varimax rotation of the retained components from Exercise 4. How would you characterize a stream with high score on rotated component 1? A low score on rotated component 1. A high score or a low score on rotated component 2?

```
> pc2.streams_princomp(streams.dat, cor=T, scores=2)
> print(loadings(rotate(pc2.streams, rotation="varimax")), cutoff=0.3)
       Comp. 1 Comp. 2
mdepth         -0.624
mwidth         -0.560
 mvelo  0.551
  silt -0.509
leaves -0.423
  sand  0.486
```

```
    mud          -0.522
```

We can see that the varimax rotation provides us with the same types of factor loadings on the same variables. Thus the interpretation above is still the same. A high score on component one would represent a faster stream with less "obstructions". A lower score on component one would represent a slower stream with more obstructions. A high score on component two would represent a shallow, narrow stream without a muddy bottom. A low score on component two would represent a deep, wide stream with a muddy bottom.

2. Problem 7 parts b,c,d,e page 285-286. Factor analyze the Shetland and Orkney Youth Survey data:

   - Use principal factoring with iteration, specifying two factors. Why is it important now to specify the number of factors?

     The principal factoring can be obtained with the `factanal()` command.

```
> shetland.fa_factanal(shetland.dat, factors=2, method="principal",
    rotation="none")
> summary(shetland.fa)
Importance of factors:
                 Factor1   Factor2
   SS loadings 2.0414201 0.8640088
Proportion Var 0.3402367 0.1440015
Cumulative Var 0.3402367 0.4842382

The degrees of freedom for the model is 4.

Uniquenesses:
   famconf    differ  critical      home   badname    famsac
 0.7698657 0.3767407 0.4873273 0.6779488 0.3497462 0.4329423

Loadings:
         Factor1 Factor2
 famconf -0.306   0.369
  differ  0.777   0.142
critical  0.551   0.457
    home -0.460   0.332
 badname  0.771   0.237
  famsac -0.485   0.576
```

     We need to specify the number of factors, because the iterative technique requires re-estimating the uniquenesses at each iteration. This requires knowledge of the number of factors needed so that the uniquenesses can be estimated. (Recall that the uniqueness of a given data point is the sum of the squared factor values for that data point).

   - Apply a varimax (orthogonal) rotation to the two factors obtained form the problem above. Circle the highest loading in each row (variable). Interpret and suggest names for the factors?

```
> shetland.fa.var_factanal(shetland.dat, factors=2,
          method="principal", rotation="varimax")
> summary(shetland.fa.var)
```

```
Importance of factors:
                 Factor1   Factor2
   SS loadings 1.6674253 1.2380036
Proportion Var 0.2779042 0.2063339
Cumulative Var 0.2779042 0.4842382

The degrees of freedom for the model is 4.

Uniquenesses:
   famconf    differ  critical      home   badname    famsac
 0.7698657 0.3767407 0.4873273 0.6779488 0.3497462 0.4329423

Loadings:
         Factor1 Factor2
 famconf          0.478
                  -----
  differ  0.722  -0.320
          -----
critical  0.713
          -----
    home -0.193   0.534
                  -----
 badname  0.770  -0.238
          -----
  famsac          0.749
                  -----
```

The largest factor loadings in each row are underlined above. It appears that factor 1 puts large positive loadings on differ, badname and critical. This factor appears to represent the communities willingness to change.

Factor two puts moderate positive loading on `famconf,` `home` and `badname`. This factor appears to place positive loading on strong family values.

- Obliquely rotate the two factors from 7b. Again circle the highest loading in each row. How do the results compare with your orthogonal-rotation results of 7c?

```
> shetland.fa.obl_factanal(shetland.dat, factors=2,
              method="principal",  rotation="promax")
> summary(shetland.fa.obl)
Importance of factors:
                     [,1]      [,2]
   SS loadings 1.7781208 1.1273081
Proportion Var 0.2963535 0.1878847
Cumulative Var 0.2963535 0.4842382

The degrees of freedom for the model is 4.
```

```
Uniquenesses:
   famconf    differ  critical      home   badname    famsac
 0.7698657 0.3767407 0.4873273 0.6779488 0.3497462 0.4329423

Loadings:
           [,1]   [,2]
 famconf         0.470
                 -----
  differ  0.751 -0.242
          -----
critical  0.702  0.142
          -----
    home -0.249  0.510
                 -----
 badname  0.791 -0.155
          -----
  famsac -0.155  0.737
                 -----
```

The highest loadings in each row correspond to the same variables as the the varimax rotations. Thus, the factor interpretation would be the same as above.

- Calculate the factor scores for the two orthogonal factors of 7c an the two oblique factors of 7d. Obtain a correlation matrix for these four variables and discuss what it shows.

```
> scores_cbind(shetland.fa.var$scores, shetland.fa.obl$scores)
> cor(scores)
             var.Factor1   var.Factor2 obl.factor1 obl.factor2
var.Factor1  1.0000000000 -0.1142787  0.9953455  0.0009524839
var.Factor2 -0.1142787039  1.0000000 -0.2094861  0.9933394790
obl.Factor1  0.9953454733 -0.2094861  1.0000000 -0.0954226032
obl.Factor2  0.0009524839  0.9933395 -0.0954226  1.0000001192
```

The above correlation shows that the first factors from each rotation are highly correlated (0.995) and the second factors are also relatively correlated (0.993). Thus, the two different rotations are providing very similar information.

Note that there appears to be some instability in the Splus routines. The varimax rotation should produce factors which are orthogonal (i.e. their correlation should be 0). We see above that the correlation between the two varimax factors is -0.114. This is too large to be attributed to round-off error.