Module 7 R Activity

Zahlen Zbinden

2023-11-18

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(psych)
```

```
Warning: package 'psych' was built under R version 4.3.2
```

```
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

1. Read in Seeds3Data.csv, and display the first six rows of data

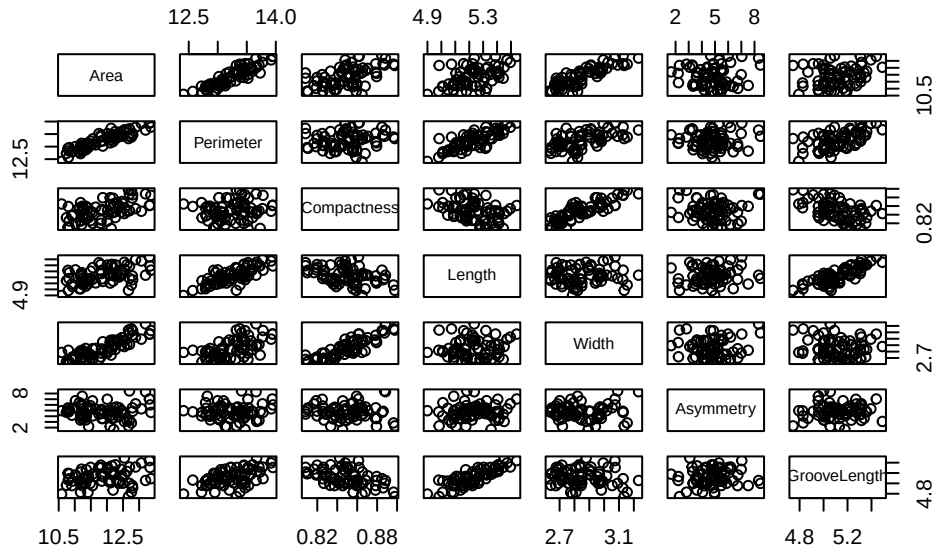
```
seeds <- read_csv("D:/RepoMan/osu/data/Seeds3Data.csv")
head(seeds)
```

```
# A tibble: 6 x 7
   Area Perimeter Compactness Length Width Asymmetry GrooveLength
  <dbl>     <dbl>       <dbl>  <dbl> <dbl>     <dbl>        <dbl>
1  13.1      13.9       0.848   5.47  2.99      5.30         5.40
2  13.3      13.9       0.861   5.54  3.07      7.04         5.44
3  13.3      14.0       0.862   5.39  3.07      6.00         5.31
4  12.2      13.3       0.865   5.22  2.97      5.47         5.22
5  11.8      13.4       0.827   5.31  2.78      4.47         5.18
6  11.2      13.1       0.817   5.28  2.69      6.17         5.28
```

2. Make a pairs plot. Based on these plots, does it look like the assumption of multivariate normality is at least remotely reasonable?

A majority of the pairwise plots are eliptical or spherical, with a cluster of points in the middle, I think the assumption of multivariate normality is at least remotely reasonable.

```
pairs(seeds)
```



3. Calculate and report the loadings matrix for the principal components factor analysis solution using the correlation matrix of the seeds data. (That is, using the standardized data)

```r
seeds_cor <- cor(seeds)
seeds_eig <- eigen(seeds_cor)
load_seeds <- seeds_eig$vec %*% diag(sqrt(seeds_eig$val))
load_seeds
```

```
          [,1]       [,2]        [,3]        [,4]        [,5]         [,6]
[1,] 0.9660520 -0.2362852 -0.05436267  0.05918577 -0.05162193  0.041081175
[2,] 0.9610151  0.1746605 -0.08456224  0.16564007 -0.10460857  0.018573427
[3,] 0.3578997 -0.9077557  0.02416597 -0.18938449  0.07683555  0.074275346
[4,] 0.7024690  0.6784037 -0.03166096  0.10240053  0.18656843  0.003121665
[5,] 0.7485206 -0.6477141  0.05815779 -0.05790370  0.02029820 -0.114175600
[6,] 0.1218485  0.1114771  0.98604517  0.01594798 -0.01049231  0.008734369
[7,] 0.4798270  0.7919195 -0.03398270 -0.37328627 -0.04600244 -0.003987631
              [,7]
[1,]  1.010139e-02
[2,] -8.269052e-03
[3,] -4.058779e-03
[4,] -1.306307e-04
[5,] -3.576217e-04
[6,] -3.553188e-05
[7,]  9.660232e-06
```

4. Examine the first three columns of the loadings matrix, can you come up with an interpretation for these first three factors?

The first loading is all postive, compactness and asymmetry do not contribute as much as the rest.

```r
load_seeds[,1]
```

```
[1] 0.9660520 0.9610151 0.3578997 0.7024690 0.7485206 0.1218485 0.4798270
```

The seconds loading is split between positive and negative, this is a contrast between the area, Compactness, and width with the remaining variables.

```r
load_seeds[,2]
```

```
[1] -0.2362852  0.1746605 -0.9077557  0.6784037 -0.6477141  0.1114771  0.7919195
```

The third loading is a contrast between area, perimeter, length, and groovelength with the remaining variables.

```
load_seeds[,3]
```

```
[1] -0.05436267 -0.08456224  0.02416597 -0.03166096  0.05815779  0.98604517
[7] -0.03398270
```

5. Calculate the fitted corelation matrix for the three factor model, and find the residual matrix. Does it look like the principal components three-factor model captures most of the structure in the correlation matrix?

Most of the differences are reasonably small in magniture, so the fit is decent.

```
seeds_pca <- principal(r = cor(seeds), nfactors = 3, rotate = "none", scores = TRUE)
seeds_pca
```

```
Principal Components Analysis
Call: principal(r = cor(seeds), nfactors = 3, rotate = "none", scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
              PC1   PC2   PC3   h2      u2 com
Area         0.97  0.24 -0.05 0.99 0.00796 1.1
Perimeter    0.96 -0.17 -0.08 0.96 0.03879 1.1
Compactness  0.36  0.91  0.02 0.95 0.04730 1.3
Length       0.70 -0.68 -0.03 0.95 0.04530 2.0
Width        0.75  0.65  0.06 0.98 0.01680 2.0
Asymmetry    0.12 -0.11  0.99 1.00 0.00044 1.1
GrooveLength 0.48 -0.79 -0.03 0.86 0.14147 1.7

                       PC1  PC2  PC3
SS loadings           3.28 2.43 0.99
Proportion Var        0.47 0.35 0.14
Cumulative Var        0.47 0.82 0.96
Proportion Explained  0.49 0.36 0.15
Cumulative Proportion 0.49 0.85 1.00

Mean item complexity =  1.5
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.02

Fit based upon off diagonal values = 1
```

```r
m <- 3
uni_pcfa3 <- diag(seeds_cor - load_seeds[,1:m] %*% t(load_seeds[,1:m]))
uni_pcfa3
```

```
       Area    Perimeter  Compactness       Length        Width    Asymmetry
0.0079574798 0.0387929360 0.0473034852 0.0453034114 0.0168010512 0.0004407172
GrooveLength
0.1414747626
```

```r
fit_pcfa3 <- load_seeds[,1:m] %*% t(load_seeds[,1:m]) + diag(uni_pcfa3)
fit_pcfa3
```

```
          [,1]       [,2]        [,3]        [,4]        [,5]        [,6]
[1,] 1.00000000 0.89171791  0.55892519  0.52004601  0.87299347  0.03776754
[2,] 0.89171791 1.00000000  0.18335437  0.79625093  0.60129154  0.05318671
[3,] 0.55892519 0.18335437  1.00000000 -0.36517648  0.85726685 -0.03375572
[4,] 0.52004601 0.79625093 -0.36517648  1.00000000  0.08455953  0.13000214
[5,] 0.87299347 0.60129154  0.85726685  0.08455953  1.00000000  0.07634701
[6,] 0.03776754 0.05318671 -0.03375572  0.13000214  0.07634701  1.00000000
[7,] 0.27826635 0.60231168 -0.54796073  0.87538056 -0.15575343  0.11323863
          [,7]
[1,]  0.2782664
[2,]  0.6023117
[3,] -0.5479607
[4,]  0.8753806
[5,] -0.1557534
[6,]  0.1132386
[7,]  1.0000000
```

```r
res_pcfa3 <- seeds_cor - fit_pcfa3
res_pcfa3
```

```
                   Area    Perimeter  Compactness        Length        Width
Area          0.000000000  0.015883121 -0.012164947 -0.0034434456 -0.009168988
Perimeter     0.015883121  0.000000000 -0.037994206 -0.0024959663 -0.013832214
Compactness  -0.012164947 -0.037994206  0.000000000 -0.0048255913  0.004046705
Length       -0.003443446 -0.002495966 -0.004825591  0.0000000000 -0.002498738
Width        -0.009168988 -0.013832214  0.004046705 -0.0024987382  0.000000000
Asymmetry     0.001843986  0.003901732 -0.003177590 -0.0002971827 -0.002133661
```

```
GrooveLength -0.019882220 -0.057093057  0.066863783 -0.0468197653  0.021136177
                Asymmetry GrooveLength
Area          0.0018439862 -0.019882220
Perimeter     0.0039017315 -0.057093057
Compactness  -0.0031775902  0.066863783
Length       -0.0002971827 -0.046819765
Width        -0.0021336612  0.021136177
Asymmetry     0.0000000000 -0.005505319
GrooveLength -0.0055053195  0.000000000
```

6. Use the factanal() function to perform maximum likelihood factor analysis on the seeds data with m = 3 factors, with no rotation of the factors. Display the resulting object.

```
seeds_mlfa3 <- factanal(x = seeds, factors = 3, rotation = "none")
seeds_mlfa3
```

```
Call:
factanal(x = seeds, factors = 3, rotation = "none")

Uniquenesses:
        Area    Perimeter  Compactness       Length        Width    Asymmetry
       0.005        0.005        0.005        0.005        0.036        0.963
GrooveLength
       0.270

Loadings:
             Factor1 Factor2 Factor3
Area          0.956   0.283
Perimeter     0.981  -0.138  -0.123
Compactness   0.297   0.946   0.116
Length        0.740  -0.645   0.175
Width         0.701   0.687
Asymmetry                     0.163
GrooveLength  0.482  -0.683   0.179

               Factor1 Factor2 Factor3
SS loadings      3.242   2.353   0.122
Proportion Var   0.463   0.336   0.017
Cumulative Var   0.463   0.799   0.817

Test of the hypothesis that 3 factors are sufficient.
```

```
The chi square statistic is 159.61 on 3 degrees of freedom.
The p-value is 2.23e-34
```

7. What is the result a hypothesis test that m = 3 factors is sufficient to explain the correlation structure in this data? Would you reject or fail to reject this null hypothesis?

With a p-value ~= 0, i would rejet the null hypothesis that 3 factors are sufficient.

8. Examine three columns of the loadings matrix: can you come up with an interpretation for these first three factors? Do these factors seem similar to the factors you dound using the principal components as a solution?

The first factor has contribution from all of the factors except for asymmetry which is very similar to our first interpretation, the second shows contrast between the same variables as our first interpretation, and the third shows the same contrasts as our first interpretation again.

```
seeds_mlfa3$loadings
```

```
Loadings:
             Factor1 Factor2 Factor3
Area          0.956   0.283
Perimeter     0.981  -0.138  -0.123
Compactness   0.297   0.946   0.116
Length        0.740  -0.645   0.175
Width         0.701   0.687
Asymmetry                     0.163
GrooveLength  0.482  -0.683   0.179

               Factor1 Factor2 Factor3
SS loadings      3.242   2.353   0.122
Proportion Var   0.463   0.336   0.017
Cumulative Var   0.463   0.799   0.817
```

9. Calculate the fitted correlation matrix for the three-factor model, and find the residual matrix. Does it look like the maximum likelihood three-factor model captures most of the structure in the correlation matrix?

Yes the avlues are all very small, which would indicate that the three-factor model captures most of the structure in the correlation matrix.

```
fit_mlfa3 <- seeds_mlfa3$load %*% t(seeds_mlfa3$load) + diag(seeds_mlfa3$uni)
res_mlfa3 <- seeds_cor - fit_mlfa3
```

10. Does it look like the principal components three-factor solution or the maximum likelihood three-factor solution fits the observed correlation matrix better?

The three-factor solution seems to fin the observed correlation matrix better.

11. Use the factanal() function to perfom maximum likelihood factor analysis on the seeds data with 3 factors, with the varimax rotation of the factors. Display the resulting object, and compare teh factors with rotation to the factors without rotation, are the rotated factors substantially easier to interpret?

They are easier to interpret as there is more seperation between the variables. This helps to showcase the relative differences better.

```
seeds_mlfa3_rot <- factanal(x = seeds, factors = 3, rotation = "varimax")
```

12. Use the factanal() function, with the regression method.

Display the scores for the 10th observation, speculate…

We should see the variables that are part of factor 1 beiing negatively affected, the variables that are part of factor 2 will also be negatively affected but not at the same magnitude, and the variables of the final factor should be positive.

Take the 10th observations scores, and multiply them by the loadings for each of the variables, both numbers are negative, and I would expect the 10th observation to be less than average.

We can see that Compactness is less than average on the 10th observation, but GrooveLength is very slightly higher than average.

```
seeds_mlfa3_rot_reg <- factanal(x = seeds, factors = 3, rotation = "varimax", score = "reg
seeds_mlfa3_rot_reg
```

```
Call:
factanal(x = seeds, factors = 3, scores = "regression", rotation = "varimax")

Uniquenesses:
        Area    Perimeter  Compactness       Length        Width    Asymmetry
       0.005        0.005        0.005        0.005        0.036        0.963
GrooveLength
       0.270

Loadings:
             Factor1 Factor2 Factor3
```

```
Area          0.777   0.626
Perimeter     0.443   0.892
Compactness   0.948  -0.301
Length                0.890   0.445
Width         0.965   0.182
Asymmetry                     0.185
GrooveLength -0.252   0.702   0.417

               Factor1 Factor2 Factor3
SS loadings      2.699   2.598   0.420
Proportion Var   0.386   0.371   0.060
Cumulative Var   0.386   0.757   0.817

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 159.61 on 3 degrees of freedom.
The p-value is 2.23e-34
```

```r
seeds_mlfa3_rot_reg$scores[10,]
```

```
   Factor1    Factor2    Factor3
-1.7857337 -0.2419468  1.5554540
```

```r
scale(seeds, center = TRUE, scale = F)[1:10,]
```

```
            Area   Perimeter   Compactness       Length       Width Asymmetry
 [1,]  1.19614286  0.67214286 -0.0014085714  0.242485714  0.14022857    0.5156
 [2,]  1.44614286  0.69214286  0.0118914286  0.311485714  0.21922857    2.2466
 [3,]  1.46614286  0.70214286  0.0125914286  0.159485714  0.22022857    1.2066
 [4,]  0.34614286  0.07214286  0.0157914286 -0.005514286  0.11322857    0.6806
 [5,] -0.05385714  0.15214286 -0.0220085714  0.084485714 -0.07677143   -0.3174
 [6,] -0.66385714 -0.11785714 -0.0327085714  0.049485714 -0.16677143    1.3806
 [7,] -0.44385714 -0.11785714 -0.0159085714 -0.053514286 -0.13477143   -2.5674
 [8,]  0.61614286  0.21214286  0.0163914286  0.037485714  0.11322857   -0.3674
 [9,]  0.82614286  0.46214286 -0.0003085714  0.156485714  0.05722857   -1.5284
[10,] -1.08385714 -0.31785714 -0.0387085714  0.087485714 -0.20577143    0.6736
      GrooveLength
 [1,]       0.2786
 [2,]       0.3236
 [3,]       0.1906
 [4,]       0.1046
```

```
 [5,]  0.0616
 [6,]  0.1586
 [7,]  0.0156
 [8,] -0.1144
 [9,]  0.1996
[10,]  0.0776
```