Factor Analysis Report on Dry Bean Dataset

By:

Thivajini Ravichandran

S/18/824

Contents

1. Introduction

Factor analysis is a method for modelling observed variables and their covariance structure in terms of a smaller number of underlying unobservable (latent) factors. Here, we model the observed variables as the linear function of factors.

The main objective of factor analysis is reducing the dimensions into small number of dimensions and convert them into interpretable way.

For this project, I use dry bean dataset to perform factor analysis. I am also interested in doing confirmatory analysis to check whether the selected number of factors are adequately sufficient to represent the whole dataset.

2. Methodology

In this dataset seven different types of dry beans were used, taking into account the features such as form, shape, type, and structure by the market situation. A computer vision system was developed to distinguish seven different registered varieties of dry beans with similar features in order to obtain uniform seed classification. For the classification model, images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. Bean images obtained by computer vision system were subjected to segmentation and feature extraction stages, and a total of 16 features; 12 dimensions and 4 shape forms, were obtained from the grains.

16 variables are described below:

- Area (A): The area of a bean zone and the number of pixels within its boundaries.
- Perimeter (P): Bean circumference is defined as the length of its border.
- Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.
- Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
- Aspect ratio (K): Defines the relationship between L and l.
- Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.
- Convex area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
- Equivalent diameter (Ed): The diameter of a circle having the same area as a bean seed area.
- Extent (Ex): The ratio of the pixels in the bounding box to the bean area.
- Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
- Roundness (R): Calculated with the following formula: (4piA)/(P^2)
- Compactness (CO): Measures the roundness of an object: Ed/L
- ShapeFactor1 (SF1)

- ShapeFactor2 (SF2)
- ShapeFactor3 (SF3)
- ShapeFactor4 (SF4)

Use methods are,

  - Exploratory factor analysis
  - Confirmatory factor analysis

3. Results and Discussion

Model :

$$\underline{X} = \underline{\mu} + \underline{LF} + \underline{\varepsilon}$$

where, x – Vector of traits

μ - Vector of population mean

L – Loading matrix

ε - Specific factors

- In this dataset, all variables cannot measured in same unit. So, I standardized the dataset in order to put same weight for all variables.
- After standardizing we can get correlation matrix.
- Using that correlation matrix we can find eigen values and eigen vector which are necessary for the model.

```
dry_bean_eigen <- eigen(dry_bean_cor)

#Eigen values
dry_bean_eigen$values
```

```
##  [1] 8.874630e+00 4.228956e+00 1.281050e+00 8.182528e-01 4.382869e-01
##  [6] 1.839617e-01 1.116241e-01 5.201320e-02 8.260261e-03 1.453890e-03
## [11] 1.054189e-03 2.939829e-04 1.487946e-04 1.001027e-05 2.146113e-06
## [16] 1.784792e-06
```

From Kaiser (1961) method, we take factors where eigenvalue should be at least greater than one. For this dataset first three eigenvalues are relatively higher than 1. Therefore our factor model has only three factors.

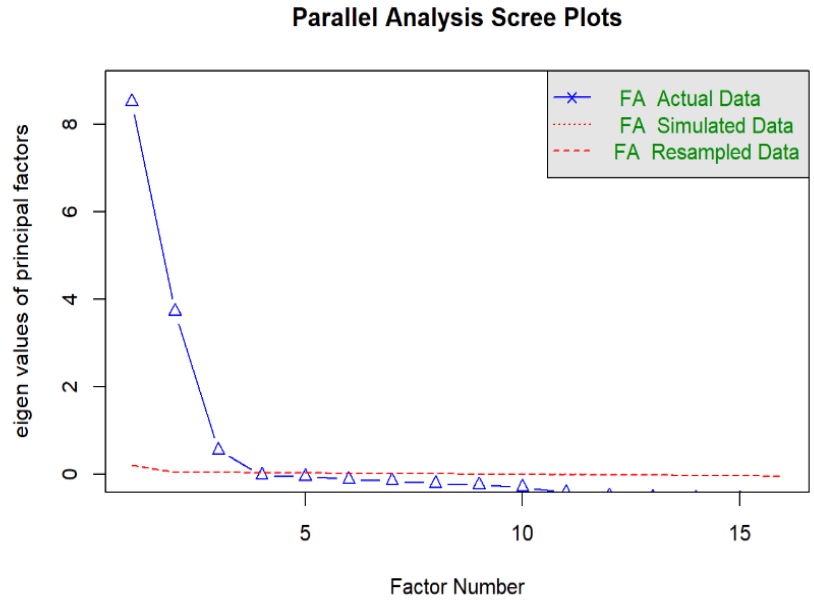
- Proportion of variation

```
prop_var <- dry_bean_eigen$values/sum(dry_bean_eigen$values)
prop_var
```

```
##  [1] 5.546644e-01 2.643097e-01 8.006564e-02 5.114080e-02 2.739293e-02
##  [6] 1.149761e-02 6.976507e-03 3.250825e-03 5.162663e-04 9.086812e-05
## [11] 6.588679e-05 1.837393e-05 9.299660e-06 6.256417e-07 1.341321e-07
## [16] 1.115495e-07
```

Cumulative proportion variation explained by first three factors = 0.91. Therefore we can conclude that factor model explains about 91% of total variation and this is a good model.

- Scree Plot and Parallel Analysis Scree Plots




```
## Parallel analysis suggests that the number of factors =  3  and the number of components =  NA
```

- Hypothesis Testing

  H 0: Three factors are sufficient.

  H a: More factors are needed

```
Mean item complexity =  1.7
Test of the hypothesis that 3 factors are sufficient.

df null model =  120  with the objective function =  76.92
df of  the model are 75  and the objective function was  87895294
```

To estimate the parameters of the above model we use 2 methods.

Such as, principal component method and maximum likelihood estimation method.

- PC Factor Loadings :

```
print(dry_bean_pc$loadings)
```

```
##
## Loadings:
##                   PA1    PA2    PA3
## Area               0.842  0.501
## Perimeter          0.930  0.367
## MajorAxisLength    0.974  0.203
## MinorAxisLength    0.708  0.709
## AspectRation       0.680 -0.684  0.184
## Eccentricity       0.684 -0.656  0.165
## ConvexArea         0.844  0.498
## EquivDiameter      0.890  0.458
## Extent            -0.161  0.368
## Solidity          -0.455  0.254  1.134
## roundness         -0.717  0.421  0.144
## Compactness       -0.710  0.688 -0.168
## ShapeFactor1      -0.653 -0.665
## ShapeFactor2      -0.935  0.272 -0.129
## ShapeFactor3      -0.712  0.685 -0.167
## ShapeFactor4      -0.556  0.188  0.341
##
##                    PA1   PA2   PA3
## SS loadings      8.826 4.156 1.575
## Proportion Var   0.552 0.260 0.098
## Cumulative Var   0.552 0.811 0.910
```

Principal component method explains 91 % of total variation in the dataset.

Factor loadings are not giving clear conclusion about the model because some variables are cross overlapped with more than one factor. Therefore we have to rotate them.

- ML factor loadings :

```
##
##                   ML1   ML2   ML3
## SS loadings     6.886 5.985 0.236
## Proportion Var  0.430 0.374 0.015
## Cumulative Var  0.430 0.804 0.819
```

From this result, we can clearly see that here also we get the same group of factors. But this method explains only 81.9% of total variation.

Therefore, among 2 methods principal component method is better for this dataset to estimate the parameters.

4. Conclusion and Recommendation

- According to the analysis we can get three factors.
- 3 factor model explains about 91% total variation of the dataset. So, this is enough to interpret the data.
- After rotating the factor loadings from “varimax” method :

```
print(dry_bean_pc_rotate$loadings,cutoff = 0.5,sort = TRUE)
```

```
##
## Loadings:
##                   PA1    PA2    PA3
## Area               0.969
## Perimeter          0.948
## MajorAxisLength    0.884
## MinorAxisLength    0.993
## ConvexArea         0.969
## EquivDiameter      0.979
## ShapeFactor1      -0.922
## AspectRation              -0.975
## Eccentricity              -0.951
## roundness                  0.720
## Compactness                0.992
## ShapeFactor2      -0.556   0.803
## ShapeFactor3               0.991
## Solidity                          1.227
## Extent
## ShapeFactor4
##
##                    PA1   PA2   PA3
## SS loadings      6.890 5.721 1.946
## Proportion Var 0.431 0.358 0.122
## Cumulative Var 0.431 0.788 0.910
```

This results clearly shows the factor. We can name the factors based on number of variables highly load by each factor. Factor 1 has high impact on area and perimeter of dry bean than shape factor. So, we can say that this is the contrast between shape factor and dimension measures. In factor 2, there is a contrast between aspect ration, eccentricity and roundness, compactness, shape factor. Factor 3 is only represents solidity. Therefore, we might call factor1 as “dimension measures”, factor 2 as “ratio measures” and factor 3 as “physical state measure”.

```
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.02
##
## Fit based upon off diagonal values = 1
```

The root means the square of residuals (RMSR) is 0.01. This is acceptable as this value should be closer to 0.

- Communalities :

```
##           Area      Perimeter MajorAxisLength MinorAxisLength    AspectRation
##     0.99500000     0.99124725     0.99500000     0.99500000     0.96843799
##   Eccentricity     ConvexArea   EquivDiameter         Extent       Solidity
##     0.94932692     0.99500000     0.99500000     0.15686590     0.09629877
##      roundness    Compactness    ShapeFactor1    ShapeFactor2    ShapeFactor3
##     0.66578990     0.99500000     0.99500000     0.98542909     0.99500000
##   ShapeFactor4
##     0.30702421
```

Except for extent, solidity and shape factor 4, the model explains above 90 % of the variation on other variables such as area, perimeter, major and minor axis length, aspect ration, eccentricity, convex area compactness and shape factors 1,2,3.

5.References

https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/

https://www.indeed.com/career-advice/career-development/confirmatory-factor-analysis

6.Appendices

Part of the dataset: https://archive.ics.uci.edu/dataset/602/dry+bean+dataset

Area <int>	Perimeter <dbl>	MajorAxisLength <dbl>	MinorAxisLength <dbl>	AspectRation <dbl>	Eccentricity <dbl>	ConvexArea <int>
28395	610.291	208.1781	173.8887	1.197191	0.5498122	28715
28734	638.018	200.5248	182.7344	1.097356	0.4117853	29172
29380	624.110	212.8261	175.9311	1.209713	0.5627273	29690
30008	645.884	210.5580	182.5165	1.153638	0.4986160	30724
30140	620.134	201.8479	190.2793	1.060798	0.3336797	30417
30279	634.927	212.5606	181.5102	1.171067	0.5204007	30600

EquivDiameter <dbl>	Extent <dbl>	Solidity <dbl>	roundness <dbl>	Compactness <dbl>	ShapeFactor1 <dbl>	ShapeFactor2 <dbl>
190.1411	0.7639225	0.9888560	0.9580271	0.9133578	0.007331506	0.003147289
191.2728	0.7839681	0.9849856	0.8870336	0.9538608	0.006978659	0.003563624
193.4109	0.7781132	0.9895588	0.9478495	0.9087742	0.007243912	0.003047733
195.4671	0.7826813	0.9766957	0.9039364	0.9283288	0.007016729	0.003214562
195.8965	0.7730980	0.9908932	0.9848771	0.9705155	0.006697010	0.003664972
196.3477	0.7756885	0.9895098	0.9438518	0.9237260	0.007020065	0.003152779

Codes:

Load the necessary packages

```{r}
library(tidyverse)
library(psych)
library(corrplot)
library(ggplot2)
```

Import the data set
This data set have 16 columns.I consider only numerical columns for the analysis

```{r}
dry_bean <- read_csv("../data/Dry_Bean_Dataset.csv",col_types = "idddddidddddddd")
head(dry_bean)
```

```{r}
glimpse(dry_bean)
```

##This dataset have 16 columns and 13611 rows.

```{r}
str(dry_bean)
```

Check for missing values

```{r}
sum(is.na(dry_bean))
```

Describe the data set

```{r}
describe(dry_bean)
```

Standardized the data set

```{r}
std_dry_bean <- scale(dry_bean)
std_dry_bean
```

Get the correlation matrix of standardized data

```{r}
dry_bean_cor <- cor(std_dry_bean)
dry_bean_cor
```

Find the eigen values and eigen vectors of above correlation matrix

```{r}
dry_bean_eigen <- eigen(dry_bean_cor)

#Eigen values
dry_bean_eigen$values

#Eigen vectors
dry_bean_eigen$vectors
```

##According to the eigen values we select number of factor as 3.
##Because first 3 eigen values are greater than 1.

Proportion of variance explained

```{r}
prop_var <- dry_bean_eigen$values/sum(dry_bean_eigen$values)
prop_var
```

Draw the scree plot

```{r}
scree(dry_bean)
```

##From the above scree plot after dimension 4 graph gradually bendsoff, so we select number of factor as 3
##That is adequately retained the dataset
Parallel analysis Scree Plots

```{r}
fa.parallel(dry_bean,fm="pa",fa="fa")
```

##From the scree plot we can see that number of factors to be 3

```{r}
dry_bean_pc <- fa(dry_bean_cor,nfactors = 3,rotate = FALSE,fm = "pa")
dry_bean_pc
```

Get un rotated loadings from PC method

```{r}
print(dry_bean_pc$loadings)
```

##From the above results we can see that some variables are cross loading with more principal components.
##So, to identify the factors clearly we need to rotate it inorder to get the clear interpretation.
##About 91 % of total variation is explained by the 3 factors.

Get unrotated PC method communalities

```{r}
dry_bean_pc$communalities
```

Rotate the PC method factor loadings using 'Varimax' method

```{r}
dry_bean_pc_rotate <- fa(dry_bean_cor,nfactors = 3,rotate = "varimax",fm = 'pa')
dry_bean_pc_rotate
```

```{r}
print(dry_bean_pc_rotate$loadings,cutoff = 0.5,sort = TRUE)
```

Get rotated communalities in PC method

```{r}
dry_bean_pc_rotate$communalities
```

b.) Factor Analysis from 'Maximum Likelihood Method'

```{r}
dry_bean_ml <- fa(dry_bean_cor,nfactors = 3,rotate = FALSE, fm = 'ml')
dry_bean_ml
```

Get unrotated ML loadings

```{r}
print(dry_bean_ml$loadings)
```

Get unrotated Communalities in ML method

```{r}
dry_bean_ml$communalities
```

Rotate the ML method factor loadings using 'Varimax' method

```{r}
dry_bean_ml_rotate <- fa(dry_bean_cor,nfactors = 3,rotate = "varimax",fm = 'ml')
dry_bean_ml_rotate
```

Get rotated ML loadings

```{r}
print(dry_bean_ml_rotate$loadings,cutoff = 0.5)
```

##From this about 81.9 % of total variation is explained by these factors.

```{r}
dry_bean_ml_rotate$communalities
```

```{r}
library(lavaan)
model <- '
DM =~Area+Perimeter+MajorAxisLength+MinorAxisLength+ConvexArea+EquivDiameter+ShapeFactor1
RM =~AspectRation+Eccentricity+roundness+Compactness++ShapeFactor2+ShapeFactor3
PM =~Solidity'

fit <- cfa(model,data = dry_bean)


summary(fit,fit.measures = TRUE,standardized = TRUE)
```