Factor Analysis - Seishu data

T Shapla

11/29/2023

# Example: Seishu data

For 30 brands of Japanese Seishu wine, Siotani et al. (1963) studied the relationship between = taste, = odor, and = pH, = acidity 1, = acidity 2, = sake meter, = direct reducing sugar, = total sugar, = alcohol, = formyl-nitrogen. The data ‘Seishu’ is given in Canvas.

We will carry out a factor analysis of the data to describe the covariance or correlation structure of these 10 variables using a smaller number of factors.

setwd("D:/LianZuo/Applied Statistics Course Materials/STAT 577 - Applied Multivariate Statistics/data-577")  
seishu=read.table("Seishu.dat")  
colnames(seishu)=c("taste", "odor", "pH", "acid1", "acid2", "sake-meter", "r.sugar", "t.sugar", "alcohol", "formyl.n")  
head(seishu, 3)

taste odor pH acid1 acid2 sake-meter r.sugar t.sugar alcohol formyl.n  
1 1.0 0.8 4.05 1.68 0.85 3.0 3.97 5.00 16.9 122  
2 0.1 0.2 3.81 1.39 0.30 0.6 3.62 4.52 15.8 62  
3 0.5 0.0 4.20 1.63 0.92 -2.3 3.48 4.46 15.8 139

dim(seishu)

[1] 30 10

Before we perform factor analysis, let us first do a preliminary analysis to check how many factors we may consider in the study where the number of observed variables is 10, .

Using the criteria: Choose m equal to the number of eigenvalues of the correlation matrix greater than the average eigenvalue, we see that first four eigenvalues are greater than the mean value of 1.

ev.seishu=eigen(cor(seishu))$values  
ev.seishu

[1] 3.17448729 2.56493022 1.43212322 1.27657967 0.54223281 0.47275791  
 [7] 0.25092430 0.11845314 0.10353588 0.06397556

mean(ev.seishu)

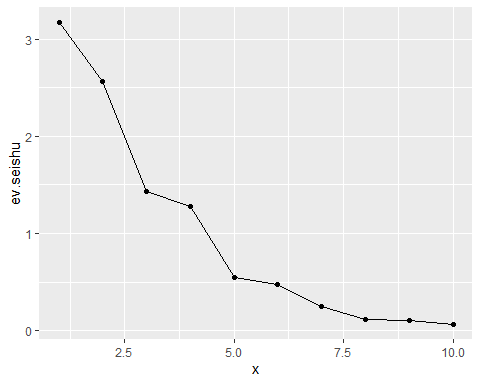
[1] 1

which(ev.seishu>mean(ev.seishu))

[1] 1 2 3 4

We also plot the eigenvalues of the correlation matrix R and see that a sharp fall occurs at 4th eigenvalue and afterwards the line has a very small slope and becomes more flat.

library(ggplot2)  
x=1:10  
ev.seishu.data=data.frame(cbind(x, ev.seishu))  
ggplot(ev.seishu.data, mapping = aes(x=x, y=ev.seishu))+geom\_point()+geom\_line()



Based on these two criteria above, we may consider 4 factors in our study.

In R, factor analysis is implemented by the factanal() function from build-in stats package. The function performs maximum-likelihood factor analysis on the correlation matrix calculated from the data matrix. The number of factors to be fitted is specified by the argument ‘factors’. At first, we will not use any rotation specified by rotation=“none” in factanal() function.

seishu.fac.none=factanal(seishu, factors=4, rotation = "none")  
seishu.fac.none

Call:  
factanal(x = seishu, factors = 4, rotation = "none")  
  
Uniquenesses:  
 taste odor pH acid1 acid2 sake-meter r.sugar   
 0.005 0.608 0.290 0.636 0.028 0.112 0.251   
 t.sugar alcohol formyl.n   
 0.020 0.232 0.206   
  
Loadings:  
 Factor1 Factor2 Factor3 Factor4  
taste 0.996   
odor 0.186 0.145 0.565 0.132   
pH -0.509 0.520 0.246 -0.346   
acid1 -0.138 0.440 0.126 0.368   
acid2 -0.446 0.843 0.245   
sake-meter -0.461 -0.618 0.538   
r.sugar 0.857 0.121   
t.sugar 0.952 0.267   
alcohol 0.304 0.500 0.650   
formyl.n -0.409 0.780 0.133   
  
 Factor1 Factor2 Factor3 Factor4  
SS loadings 2.624 2.509 1.489 0.991  
Proportion Var 0.262 0.251 0.149 0.099  
Cumulative Var 0.262 0.513 0.662 0.761  
  
Test of the hypothesis that 4 factors are sufficient.  
The chi square statistic is 11.38 on 11 degrees of freedom.  
The p-value is 0.412

The first chunk provides the estimated uniquenesses (variance of error terms), , which range from 0 to 1. The uniqueness, sometimes referred to as noise, specific variance or residual variance corresponds to the proportion of variability in a variable, which can not be explained by the underlying factors. A high uniqueness for a variable indicates that the factors do not account well for its variance.

The second chuck provides the estimated loadings for factors. Note that when a factor analysis is performed based on a correlation matrix, loading is the estimated correlation between ith variable and jth factor . For example, 0.565 is the sample correlation between odor and factor 3.

Using a cutoff value of 0.5 for factor loadings, we make the following observations:

* Factor 1 is primarily a measure of pH, reducing sugar and total sugar variables.
* Variables pH, acidity 2, sake-meter, alcohol and formyl-nitrogen can be associated with Factor 2.
* Variables taste and odor can be described by Factor 3.
* Factor 4 can be considered as a measure of sake-meter and alcohol variables.

Rest of the outputs will be discussed later part of this documentation.

It appears that there may be clear distinction between factors for this data. However, there are situations where loadings are mostly of moderate size and a factor rotation method is needed to be applied to make the description of factors more straightforward.

Let us now apply “varimax” rotation to seishu data.

seishu.fac.vm=factanal(seishu, factors=4, rotation = "varimax")  
seishu.fac.vm

Call:  
factanal(x = seishu, factors = 4, rotation = "varimax")  
  
Uniquenesses:  
 taste odor pH acid1 acid2 sake-meter r.sugar   
 0.005 0.608 0.290 0.636 0.028 0.112 0.251   
 t.sugar alcohol formyl.n   
 0.020 0.232 0.206   
  
Loadings:  
 Factor1 Factor2 Factor3 Factor4  
taste 0.110 0.987   
odor 0.168 0.575 0.169   
pH 0.814 -0.105 0.112 -0.153   
acid1 0.318 0.113 0.496   
acid2 0.912 0.131 0.345   
sake-meter -0.438 -0.824 0.129   
r.sugar -0.391 0.742 0.193   
t.sugar -0.273 0.903 0.101 0.283   
alcohol 0.234 0.844   
formyl.n 0.828 0.324   
  
 Factor1 Factor2 Factor3 Factor4  
SS loadings 2.719 2.149 1.397 1.347  
Proportion Var 0.272 0.215 0.140 0.135  
Cumulative Var 0.272 0.487 0.626 0.761  
  
Test of the hypothesis that 4 factors are sufficient.  
The chi square statistic is 11.38 on 11 degrees of freedom.  
The p-value is 0.412

# Explanation of Outputs

The first chunk provides the estimated uniquenesses (variance of error terms), , . Note that the uniquenesses remain the same with and without rotation technique.

The second chuck provides the estimated loadings for factors after applying the varimax rotation. Using a threshold value of 0.5, variables pH, acidity 2 and formyl-nitrogen can be described by factor 1; variables sake meter, reducing sugar and total sugar constitutes factor 2; variables taste and odor can be associated with factor 3, whereas acidity 1 and alcohol are aligned with factor 4.

The first row in the table beneath the loadings is ‘SS loadings’. ‘SS loadings’ gives the sum of squared loadings associated with a factor. For example, SS loadings of 2.719 is found by computing the sum of squares of loadings associated with factor 1:

.

A factor is worth keeping if the SS loading is greater than 1 (Kaiser’s rule). The SS loadings provides the total variance in original variables that is explained by a given factor.

The proportion of variation explained by each factor is given by the row ‘Proportion Var’, and is calculated as given below based on a correlation matrix :

, where , is the number of variables in the data, and is the sample correlation matrix.

The row ‘Cumulative Var’ gives the cumulative proportion of variance explained. These numbers range from 0 to 1.

The last section of the output shows the results of a hypothesis test. The null hypothesis, , is that the number of factors considered in the model is sufficient. We reject at 5% level of significance if the p-value is less than 0.05. Such a result indicates that the number of factors is too small and more factors are needed to include. In contrast, we do not reject if the p-value exceeds 0.05. Such a result indicates that there are likely enough (or more than enough) factors included in the model. The high p-value in our example above leads us to not reject the , and indicates that we have fitted an appropriate model. This hypothesis test is available only with the maximum likelihood estimations of laodings and specific variances. (For test statistic formula and degrees of freedom, please see the Chapter note).

Since p-value is found to be 0.412, we conclude that four factors considered are sufficient for the fit.

Based on the variables that belong to each factor in this analysis, we can name or label each factor. For example, factor 3 can be named ‘flavor’ as odor and taste variables are associated with it.

In order to print the rotation matrix , we extract ‘rotmat’ from factanal object ‘seishu.fac.rot’.

seishu.fac.vm$rotmat

[,1] [,2] [,3] [,4]  
[1,] 0.98579715 -0.003146453 -0.1526563 0.06992946  
[2,] 0.09031392 0.848817553 0.5020006 -0.13909554  
[3,] 0.08842375 -0.400958998 0.7887977 0.45739621  
[4,] -0.11058308 0.344573277 -0.3201400 0.87552899

Recall that communality is the part of the variance of the ith variable that is explained by all the factors considered in the model. To find communalities , we can use:

loads=seishu.fac.vm$loadings  
loads

Loadings:  
 Factor1 Factor2 Factor3 Factor4  
taste 0.110 0.987   
odor 0.168 0.575 0.169   
pH 0.814 -0.105 0.112 -0.153   
acid1 0.318 0.113 0.496   
acid2 0.912 0.131 0.345   
sake-meter -0.438 -0.824 0.129   
r.sugar -0.391 0.742 0.193   
t.sugar -0.273 0.903 0.101 0.283   
alcohol 0.234 0.844   
formyl.n 0.828 0.324   
  
 Factor1 Factor2 Factor3 Factor4  
SS loadings 2.719 2.149 1.397 1.347  
Proportion Var 0.272 0.215 0.140 0.135  
Cumulative Var 0.272 0.487 0.626 0.761

communality=apply(loads^2,1,sum) # communality  
communality

taste odor pH acid1 acid2 sake-meter r.sugar   
 0.9950036 0.3920669 0.7101515 0.3642198 0.9715226 0.8878609 0.7492818   
 t.sugar alcohol formyl.n   
 0.9797906 0.7678114 0.7941803

To find the sum of squares of the loadings associated with each factor,

ss.loadings=apply(loads^2,2,sum)  
ss.loadings

Factor1 Factor2 Factor3 Factor4   
2.718567 2.149057 1.396779 1.347486

Below we find the sum of communalities and the unique variances. Note that all the results very close to 1. Why?

communality+seishu.fac.vm$uniquenesses

taste odor pH acid1 acid2 sake-meter r.sugar   
 1.0000036 0.9999929 1.0000001 1.0000063 1.0000000 0.9999993 0.9999983   
 t.sugar alcohol formyl.n   
 0.9999999 1.0000002 0.9999995