Dimension Reduction - PCA and Exploratory Factor Analysis

Josh Jarvey

07/13/2020

Create a Word document from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word document.

## Exercise 1

A researcher was interested in learning what motivates international fans when they watch U.S. sports. A questionnaire was created in which respondents reported their score for 42 “importance factors” about fan motivation by circling the number that indicates why they watch, read, and/or discuss U.S. sports (5=Very High, 4=High, 3=Average, 2=Low, 1=Very Low).

The fans were categorized on issues such as gender, interest in U.S. sports, and the media source from which their information comes. Four hundred surveys were completed for the study.

The data is in the file ifanmot.rda and the survey is in the file IFM\_Survey.docx.

### Part 1a

Conduct Bartlett’s test for sphericity on the responses for the 42 survey questions found in columns 1 through 42 of the file ifanmot.rda. State the null and alternative hypothesis and report on the results. Note, in the R function, n represents the sample size of the data that was used to create the correlation/covariance matrix.

Is factor analysis warranted based on this measure?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

#loading the necessary packages, and the dataset  
library(DS705data)  
library(psych)

## Warning: package 'psych' was built under R version 4.0.2

data("ifanmot")  
  
 #create a correlation matrix of all 42 survey questions among each other  
matx = cor(ifanmot[1:42])  
 #complete a bartlett test to see if the correlation matrix is "spherical", which tests if the matrix is the identity matrix  
cortest.bartlett(matx,n=400)

## $chisq  
## [1] 10197.66  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 861

H\_0 = The correlation matrix is the identity matrix H\_a = The correlation matrix is not the identity matrix

Conclusion: At the 5% level of significance, there is enough evidence to suggest that the correlation matrix is not the identity matrix (p-value=0). We can continue with our factor analysis.

### Part 1b

Compute the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) for the responses for the 42 survey questions found in columns 1 through 42 of the file ifanmot.rda.

Is the overall MSA value acceptable for factor analysis?

Should any questionnaire items be dropped from the factor analysis because of low MSA values? If so which ones?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

#complete the KMO factor adequacy check - those variables that are lower than 0.50 should be removed. There are none in this example.  
 #note: capital K M O (lower case does not work)  
KMO(matx)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = matx)  
## Overall MSA = 0.93  
## MSA for each item =   
## Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16   
## 0.95 0.91 0.96 0.90 0.91 0.97 0.80 0.94 0.96 0.76 0.95 0.92 0.96 0.96 0.94 0.91   
## Q17 Q18 Q19 Q20 Q21 Q22 Q23 Q24 Q25 Q26 Q27 Q28 Q29 Q30 Q31 Q32   
## 0.90 0.96 0.94 0.95 0.97 0.87 0.87 0.96 0.96 0.93 0.94 0.96 0.96 0.89 0.89 0.89   
## Q33 Q34 Q35 Q36 Q37 Q38 Q39 Q40 Q41 Q42   
## 0.95 0.96 0.96 0.95 0.95 0.94 0.88 0.91 0.79 0.81

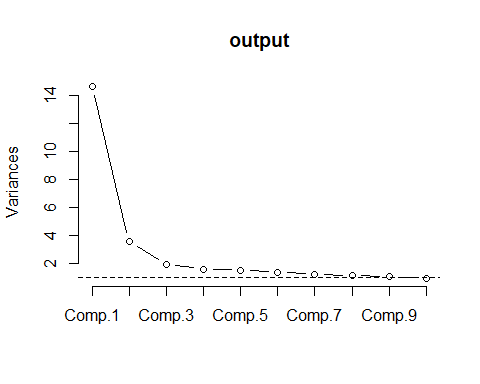
The overall MSA is “superb” with a score of 0.93. Additionally, there are no variables that should be dropped from this factor analysis, using the 0.50 value as a cut off threshold (those that would be below 0.50 would indicate they could be dropped and stand on their own).

### Part 1c

Use R to create a scree plot for the questionnaire items that you deemed to be appropriate for the factor analysis from the previous question. Use the scree plot to answer the questions below. (You may need to print the eigenvalues using code like *output$sdev^2* to see which ones are over 1)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

#complete a principal component analysis using all 42 questions to get the eigenvalues.   
output = princomp(ifanmot[,1:42], cor = TRUE)  
 #create the scree plot using the output from the above step. Where the line starts to flatten out, this is the number of factors to consider.  
plot(output, type = "lines")  
 #kaiser's rule is the dotted line, but this dataset has >30 variables, so we will focus on the knee of the scree plot to determine factors, 3  
abline(h=1, lty=2)



#extract the eigen values from the $sdev vector. Check for those that are greater than 1 (kaiser's rule) - there are 9.  
print(sum((output$sdev^2)>1))

## [1] 9

##### Where would you say the “knee” is in the scree plot?

The “knee” of the scree plot begins at the 3rd factor, as the line begins to “straighten out” after that (slopes are minimal).

##### How many factors does the knee in the scree plot suggest extracting?

The knee in the scree plot suggests 3 factors to explore.

##### How many components have eigenvalues (aka variances, latent roots) greater than 1?

There are 9 components that have an eigenvalue greater than 1.

##### How many factors does this suggest extracting?

If we were to follow Kaiser’s rule, then this would suggest 9 factors to be extracted.

##### Is the Kaiser Rule (eigenvalues >1) recommended for this data set? Explain.

We will not use Kaiser’s Rule because there are greater than 30 variables in this data set.

### Part 1d

Use a principal components extraction with the varimax rotation to extract 3 factors. Print the output with factor loadings under 0.5 suppressed and sort the loadings. (Note - in the video the columns were labeled PC1, PC2, etc., but the newer version of principal() uses RC1, RC2, etc.)

Answer the questions below.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

#extract the factors using the principal() function from the psych package  
 #per the above steps, we identified 3 factors we'd like to extract. We'll use varimax rotation as well.  
fa.out = principal(ifanmot[,1:42],nfactors = 3,rotate = "varimax")  
 #print.psych() gives more relevant ouput. We want to see all the variables and what their factor scores are.  
 #cut off the ones below 0.50 (since they are not strong enough), and sort them.  
print.psych(fa.out, cut=0.50, sort = TRUE)

## Principal Components Analysis  
## Call: principal(r = ifanmot[, 1:42], nfactors = 3, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## item RC1 RC2 RC3 h2 u2 com  
## Q13 13 0.69 0.55 0.45 1.3  
## Q3 3 0.68 0.53 0.47 1.3  
## Q5 5 0.66 0.44 0.56 1.0  
## Q6 6 0.66 0.50 0.50 1.3  
## Q9 9 0.66 0.55 0.45 1.5  
## Q15 15 0.65 0.47 0.53 1.2  
## Q16 16 0.65 0.48 0.52 1.2  
## Q37 37 0.64 0.53 0.47 1.6  
## Q38 38 0.62 0.42 0.58 1.2  
## Q28 28 0.61 0.54 0.46 1.8  
## Q26 26 0.59 0.58 0.42 1.9  
## Q8 8 0.58 0.46 0.54 1.7  
## Q4 4 0.58 0.35 0.65 1.1  
## Q36 36 0.58 0.51 0.61 0.39 2.0  
## Q17 17 0.58 0.38 0.62 1.3  
## Q1 1 0.57 0.46 0.54 1.8  
## Q20 20 0.54 0.58 0.42 2.6  
## Q40 40 0.53 0.39 0.61 1.6  
## Q27 27 0.53 0.52 0.56 0.44 2.1  
## Q29 29 0.52 0.49 0.51 2.4  
## Q39 39 0.38 0.62 1.9  
## Q12 12 0.24 0.76 1.1  
## Q14 14 0.42 0.58 2.7  
## Q19 19 0.42 0.58 2.9  
## Q10 10 0.75 0.58 0.42 1.0  
## Q7 7 0.72 0.52 0.48 1.0  
## Q41 41 0.66 0.44 0.56 1.0  
## Q22 22 0.64 0.52 0.48 1.5  
## Q42 42 0.64 0.41 0.59 1.0  
## Q2 2 0.58 0.36 0.64 1.1  
## Q23 23 0.55 0.45 0.55 1.9  
## Q35 35 0.53 0.56 0.44 2.2  
## Q18 18 0.50 0.38 0.62 2.0  
## Q11 11 0.42 0.58 2.3  
## Q34 34 0.49 0.51 2.8  
## Q21 21 0.44 0.56 2.8  
## Q32 32 0.82 0.71 0.29 1.1  
## Q31 31 0.78 0.67 0.33 1.2  
## Q30 30 0.73 0.55 0.45 1.1  
## Q33 33 0.64 0.50 0.50 1.5  
## Q25 25 0.42 0.58 2.4  
## Q24 24 0.38 0.62 2.9  
##   
## RC1 RC2 RC3  
## SS loadings 9.63 5.53 4.96  
## Proportion Var 0.23 0.13 0.12  
## Cumulative Var 0.23 0.36 0.48  
## Proportion Explained 0.48 0.27 0.25  
## Cumulative Proportion 0.48 0.75 1.00  
##   
## Mean item complexity = 1.7  
## Test of the hypothesis that 3 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.06   
## with the empirical chi square 2531.01 with prob < 1.2e-194   
##   
## Fit based upon off diagonal values = 0.97

##### What is the cumulative variance explained (as a percent)?

The cumulative variance explained is determined by adding up each’s factors “proportion var”. This then displayed in the last factor’s “cumulative var” column, which in this example is 0.48 - or 48%.

##### Is this considered an acceptable percent of total variation?

This would not be considered an acceptable percent of total variation since we use a 60% cutoff.

### Part 1e

Use a principal components extraction with the varimax rotation to extract 9 factors. Print the output with factor loadings under 0.5 suppressed and sort the loadings.

Answer the questions below.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

#extract the factors using the principal() function from the psych package  
 #per the above steps, we identified 9 factors we'd like to extract. We'll use varimax rotation as well.  
fa.out2 = principal(ifanmot[,1:42],nfactors = 9,rotate = "varimax")  
 #print.psych() gives more relevant ouput. We want to see all the variables and what their factor scores are.  
 #cut off the ones below 0.50 (since they are not strong enough), and sort them.  
print.psych(fa.out2,cut=0.50, sort = TRUE)

## Principal Components Analysis  
## Call: principal(r = ifanmot[, 1:42], nfactors = 9, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## item RC1 RC9 RC5 RC3 RC4 RC7 RC2 RC6 RC8 h2 u2 com  
## Q6 6 0.70 0.62 0.38 1.6  
## Q3 3 0.67 0.62 0.38 1.9  
## Q9 9 0.64 0.64 0.36 2.3  
## Q28 28 0.62 0.65 0.35 2.6  
## Q15 15 0.62 0.61 0.39 2.2  
## Q8 8 0.57 0.57 0.43 2.8  
## Q13 13 0.55 0.62 0.38 3.3  
## Q36 36 0.54 0.72 0.28 3.9  
## Q37 37 0.53 0.63 0.37 3.6  
## Q1 1 0.52 0.55 0.45 3.0  
## Q29 29 0.51 0.56 0.44 3.2  
## Q19 19 0.69 0.71 0.29 2.1  
## Q20 20 0.63 0.73 0.27 2.9  
## Q14 14 0.60 0.56 0.44 2.2  
## Q18 18 0.54 0.54 0.46 3.0  
## Q11 11 0.53 0.54 0.46 2.9  
## Q35 35 0.53 0.63 0.37 3.7  
## Q23 23 0.75 0.75 0.25 1.7  
## Q22 22 0.71 0.75 0.25 2.1  
## Q24 24 0.63 0.60 0.40 2.1  
## Q27 27 0.57 0.69 0.31 3.5  
## Q26 26 0.57 0.72 0.28 3.5  
## Q25 25 0.52 0.58 0.42 3.5  
## Q21 21 0.49 0.51 4.6  
## Q32 32 0.83 0.80 0.20 1.4  
## Q31 31 0.80 0.75 0.25 1.4  
## Q30 30 0.76 0.67 0.33 1.3  
## Q33 33 0.59 0.58 0.42 2.5  
## Q4 4 0.82 0.76 0.24 1.3  
## Q5 5 0.75 0.74 0.26 1.7  
## Q12 12 0.74 0.62 0.38 1.3  
## Q39 39 0.82 0.83 0.17 1.5  
## Q40 40 0.75 0.73 0.27 1.6  
## Q38 38 0.64 0.67 0.33 2.4  
## Q10 10 0.83 0.79 0.21 1.3  
## Q7 7 0.80 0.76 0.24 1.4  
## Q2 2 0.70 0.64 0.36 1.6  
## Q41 41 0.82 0.74 0.26 1.2  
## Q42 42 0.81 0.74 0.26 1.2  
## Q34 34 0.56 0.44 5.4  
## Q17 17 0.75 0.74 0.26 1.7  
## Q16 16 0.69 0.76 0.24 2.2  
##   
## RC1 RC9 RC5 RC3 RC4 RC7 RC2 RC6 RC8  
## SS loadings 5.65 3.67 3.66 3.31 2.49 2.45 2.38 2.26 2.11  
## Proportion Var 0.13 0.09 0.09 0.08 0.06 0.06 0.06 0.05 0.05  
## Cumulative Var 0.13 0.22 0.31 0.39 0.45 0.51 0.56 0.62 0.67  
## Proportion Explained 0.20 0.13 0.13 0.12 0.09 0.09 0.09 0.08 0.08  
## Cumulative Proportion 0.20 0.33 0.46 0.58 0.67 0.76 0.84 0.92 1.00  
##   
## Mean item complexity = 2.4  
## Test of the hypothesis that 9 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.04   
## with the empirical chi square 1014.38 with prob < 2.3e-34   
##   
## Fit based upon off diagonal values = 0.99

##### What is the cumulative variance explained (as a percent)?

When implementing Kaiser’s Rule and using 9 factors, the cumulative variance explained is now 0.67 - or 67%.

##### Is this considered an acceptable percent of total variation?

Now that we have a cumulative variance explained that is above 60%, we have an acceptable number of factors extracted.

### Part 1f

Read the questions in the survey (IFM Survey.docx) for the groups of items that load onto each factor and put a comprehensive label on each of the 9 factors from the most recent factor analysis (extracting 9 factors with principal components and a varimax rotation).

For consistency assign the following 9 labels to the most appropriate factors:

**Artistic, Boredom, Entertainment, Fun, Gambling, Identification, Loyalty, Patriotism, Social**

Factors 1 through 9 move from left to right in the sorted output (even though the columns labels PC1-PC9 in the output are not in order).

I have labeled the second factor (labeled PC9 in the R output) for you as “Social”. Look at the survey items that correspond to the second factor (Q11, Q14, Q18, Q19, Q20, and Q35) and see if that label makes sense.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

Factor 1: Fun

Factor 2: Social

Factor 3: Identification

Factor 4: Patriotism

Factor 5: Artistic

Factor 6: Loyalty

Factor 7: Gambling

Factor 8: Entertainment

Factor 9: Boredom

### Part 1g

Combine the factor scores produced by the 9-factor solution with the original data frame. Also, rename the factor scores using the labels you assigned previously. Some R code to begin this has been provided. Add to it to complete this request.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g -|-|-|-|-|-|-|-|-|-|-|-

require(psych)  
require(DS705data)  
data(ifanmot)  
fan <- principal(ifanmot[,1:42],nfactors=9,rotate="varimax")  
fulldata <- cbind(ifanmot,fan$scores)  
require(plyr)

## Loading required package: plyr

fulldata <- rename(fulldata,c("RC1"="Fun","RC9"="Social","RC5"="Identification","RC3"="Patriotism","RC4"="Artistic","RC7"="Loyalty","RC2"="Gambling","RC6"="Entertainment","RC8"="Boredom")) # REPLACE the ? with your factor labels  
  
head(fulldata[,43:54])

## Gender Interest Source Fun Social Identification Patriotism  
## 1 Female Weekly TV 0.4706228 0.9003270 -1.22035741 -1.7746931  
## 2 Female Annually TV -0.3430227 -2.2575302 -0.09179851 -1.5694131  
## 3 Male Daily Newspapers 0.6152768 0.1729964 -1.71745743 -0.3422020  
## 4 Female Monthly Internet 0.2222439 -0.8938861 -0.56654534 1.5159876  
## 5 Female Monthly TV 1.0165101 -0.7719318 0.59416186 -1.1133732  
## 6 Female Monthly Internet -0.2503532 0.3934291 -0.75419121 0.4576417  
## Artistic Loyalty Gambling Entertainment Boredom  
## 1 0.6008548 -2.2227639 -1.0921480 -0.76224986 1.3293399  
## 2 0.7383736 -0.1574072 -1.0375819 -0.91086255 -0.3673669  
## 3 0.9907986 1.1392077 -1.0889313 -0.25873903 -0.8007895  
## 4 -0.8286745 0.3396687 1.0650654 0.23861899 -0.1657931  
## 5 -0.5194569 -0.2845462 0.4059857 0.02974279 -0.2103151  
## 6 -0.8237031 -0.4570054 1.3240365 -0.88828765 0.5270300