Dimension Reduction - PCA and Exploratory Factor Analysis

Matt Allen

08/08/2018

Create a Word document from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word document via D2L Dropbox.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

mat <- cor(ifanmot[,1:42])  
cortest.bartlett(mat,n=400)

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

: The correlation matrix is the identity matrix

: The correlation matrix is not the identity matrix

At a 5% level of significance, we can reject the null hypothesis in favor of the alternative that the correlation matrix is not the identity matrix. In other words, correlations exist between the variables.

The p-value is reported as zero. The result is a strong indication that factor analysis is warranted.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

KMO(mat)

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

The overall MSA value would be considered superb by the guidelines set by Hutcheson and Sofroniou, and so it is acceptable for factor analysis.

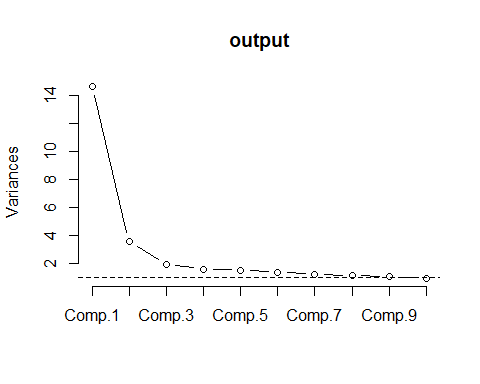
The lowest MSA value for a question is 0.76, which is still considered good by Hutcheson and Sofroniou, so no questionnaire items need to be dropped from the factor analysis because of low MSA values.

### 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.

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

output <- princomp(ifanmot[,1:42], cor=TRUE)  
{  
plot(output,type="lines") # scree plot   
abline(h=1,lty=2) # add horizonal dotted line at 1  
}



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

The knee is at Comp.3.

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

The number of factors to extract suggested by the knee in the scree plot is 3.

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

From the plot it, appears 9 components have eigenvalues greater than 1.

How many factors does this suggest extracting?

It would suggest extracting 9 components, but that is suggested for fewer than 30 variables, which is not the case in this example.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

What is the cumulative variance explained (as a percent)? The cumulative variance explained is 48%.

Is this considered an acceptable percent of total variation? This would not be considered acceptable, since a 60% minimum is considered acceptable.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

What is the cumulative variance explained (as a percent)? The cumulative variance explained is 67%.

Is this considered an acceptable percent of total variation? Since a 60% minimum is considered acceptable, 67% would be considered acceptable.

### 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: Boredom Factor 9: Entertainment

### 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)

## Loading required package: DS705data

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

## Loading required package: plyr

## Warning: package 'plyr' was built under R version 3.4.4

fulldata <- rename(fulldata,c("RC1"="Fun","RC9"="Social","RC5"="Identification","RC3"="Patriotism","RC4"="Artistic","RC7"="Loyalty","RC2"="Gambling","RC6"="Boredom","RC8"="Entertainment"))