Stat 407 Homework 3 Fall 2012
Due: Wednesday, October 11 in class Individual work is expected.

In this homework you will examine the results of factor analysis on data simulated to look like Spearman's original data. Spearman's original model for exams scores of school children was:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.83 & 0.78 & 0.70 & 0.66 & 0.63 \\ 0.83 & 1 & 0.67 & 0.67 & 0.65 & 0.57 \\ 0.78 & 0.67 & 1 & 0.64 & 0.54 & 0.51 \\ 0.70 & 0.67 & 0.64 & 1 & 0.45 & 0.51 \\ 0.66 & 0.65 & 0.54 & 0.45 & 1 & 0.40 \\ 0.63 & 0.57 & 0.51 & 0.51 & 0.40 & 1 \end{bmatrix}$$

$$\begin{aligned} \text{Classics} &= l_1 f + \varepsilon_1 \\ \text{French} &= l_2 f + \varepsilon_2 \\ \text{English} &= l_3 f + \varepsilon_3 \\ \text{Math} &= l_4 f + \varepsilon_4 \\ \text{Disc} &= l_5 f + \varepsilon_5 \\ \text{Music} &= l_6 f + \varepsilon_6 \end{aligned}$$

1. Simulate data using the following values $l_1 = 0.8, l_2 = 0.7, l_3 = 0.76, l_4 = 0.9, l_5 = 0.83, l_6 = 0.78$, $\psi_1 = 10, \psi_2 = 15, \psi_3 = 9, \psi_4 = 5, \psi_5 = 15, \psi_6 = 20$. Values for f, call it intelligence, need to be generated by a sample from a univariate normal distribution and scaled to take values between 0 and 100 (numbers approximating test scores).

   (a) Make a histogram of the variable intelligence. This is our simulated factor. It should be a bell-shaped distribution ranging from 0-100.

   (b) Compute the correlation matrix for the exam scores for Classics, French, English, Math, Disc, Music. It's not exactly the same as Spearman's original but similar.

   (c) Load the data into ggobi. Look at the exams scores Classics, French, English, Math, Disc, Music in a tour. Describe the shape.

   (d) Summarize your findings in a table like this below, so set this table up and begin by filling in the values for the true loadings.

Variable	True loadings	estimated loadings		uniquenesses	
		PCA	MLE	PCA	MLE
Classics	0.8				
French	0.7				
English					
Math					
Disc					
Music					
Variance	–				
Cum %	–				

   (e) Compute a principal component analysis. Draw a scree plot, hand this in. How much variation does the first principal component (factor) explain? Add the loadings to the table. How well do the estimated loadings match the true loadings? Compute uniquenesses and add these to the table.

   (f) Compute the factor analysis model using maximum likelihood, setting number of factors to be 1. How much variation does the first factor explain? Add the loadings and uniquenesses to the table.

   (g) The output from the MLE method includes a test for whether more factors are needed. Report the p-value form this test, and your answer to the question "Are more factors needed?".

   (h) From the uniquenesses, which exam (variable) has the **least** variance explained by the model?

2. To create a more complex example, simulate data with a second factor called "language" and from the following model:

$$\begin{aligned}
\text{Classics} &= l_{11}f_1 + l_{12}f_2 + \varepsilon_1 \\
\text{French} &= l_{21}f_1 + l_{22}f_2 + \varepsilon_2 \\
\text{English} &= l_{31}f_1 + l_{32}f_2 + \varepsilon_3 \\
\text{Math} &= l_{41}f_1 + l_{42}f_2 + \varepsilon_4 \\
\text{Disc} &= l_{51}f_1 + l_{52}f_2 + \varepsilon_5 \\
\text{Music} &= l_{61}f_1 + l_{62}f_2 + \varepsilon_6
\end{aligned} \qquad \mathbf{L} = \begin{bmatrix} 0.70 & 0.70 \\ 0.30 & 0.95 \\ 0.30 & 0.95 \\ 0.95 & 0.30 \\ 0.40 & 0.90 \\ 0.95 & 0.30 \end{bmatrix}$$

and $\psi_1 = 5, \psi_2 = 6, \psi_3 = 9, \psi_4 = 5, \psi_5 = 5, \psi_6 = 8$. Values for the two factors, f_1 = intelligence, and f_2 = language, are generated from two separate univariate normal distributions. Write out the data into a file called `spearman2.csv` and email this to the instructor. Then answer these questions about it.

(a) Load the data into ggobi. Describe the shape.

(b) Using the mle method estimate the parameters to a one factor model. Report the p-value from the test for whether one factor is sufficient. Do you need more than one factor, according to this test?

(c) Using the mle method estimate the parameters to a two factor model. Report the p-value from the test for whether two factor is sufficient. Do you need more than two factor, according to this test?

(d) Compute a principal component analysis. Draw a scree plot, hand this in. How much variation do the first two principal components (factors) explain?

(e) Use varimax rotation. Describe how this changes the the loadings for the f_1, f_2 of the model.

(f) Tabulate the information from your results as follows:

Variable	True load.		est. load.		rot. load.		uniqueness
	f_1	f_2	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$	
Classics	0.70	0.70					
French	0.30	0.95					
English							
Math							
Disc							
Music							
Variance	–	–					
Cum. %	–	–					

Sample code:

- Simulate data:

```
# Original correlation matrix
> spearman.cor <- matrix(c(1,0.83,0.78,0.70,0.66,0.63,
                   0.83,1,0.67,0.67,0.65,0.57,
                   0.78,0.67,1,0.64,0.54,0.51,
                   0.70,0.67,0.64,1,0.45,0.51,
                   0.66,0.65,0.54,0.45,1,0.40,
                   0.63,0.57,0.51,0.51,0.40,1),ncol=6,byrow=T)
> spearman.cor

# Sample from a univariate standard normal distribution
> x <- rnorm(100)
> x <- (x-min(x))/(max(x)-min(x))*100
# Check the sample
> summary(x)
```

```
> library(ggplot2)
> qplot(x, geom="histogram", binwidth=10)

> intelligence<-x # This is f, the factor
> classics<-0.8*intelligence+rnorm(100)*10
> french<-0.7*intelligence+rnorm(100)*15
> english<-0.76*intelligence+rnorm(100)*9
> math<-0.9*intelligence+rnorm(100)*5
> disc<-0.83*intelligence+rnorm(100)*15
> music<-0.78*intelligence+rnorm(100)*20
> spearman<-data.frame(intelligence, classics, french, english, math, disc, music)
> library(rggobi)
> ggobi(spearman)
```

- To compute PCA factor model, loadings, and uniquenesses for a single factor model are calculated as $L_1 = \lambda_1 * \mathbf{e}_1^2$, $\Psi = 1 - \lambda_1 * \mathbf{e}_1^2$:

```
> spearman.pca <- prcomp(spearman[,-1], scale=T, retx=T)
> spearman.pca
> screeplot(spearman.pca,type="lines")
> spearman.pca$rotation[,1]*spearman.pca$sdev[1]
> 1-spearman.pca$rotation[,1]^2*spearman.pca$sdev[1]^2
```

- To compute a factor analysis using maximum likelihood estimation:

```
> spearman.fa<-factanal(spearman[,-1],1)
> spearman.fa
```

- Simulate data for a two factor model:

```
> x<-rnorm(100)
> x<-(x-min(x))/(max(x)-min(x))*100
> summary(x)
> qplot(x, geom="histogram", binwidth=10)

> language<-x
> classics2<-0.7*intelligence+0.7*language+rnorm(100)*5
> french2<-0.3*intelligence+0.95*language+rnorm(100)*6
> english2<-0.3*intelligence+0.95*language+rnorm(100)*9
> math2<-0.95*intelligence+0.3*language+rnorm(100)*5
> disc2<-0.4*intelligence+0.9*language+rnorm(100)*5
> music2<-0.95*intelligence+0.3*language+rnorm(100)*8
> spearman2<-data.frame(intelligence, language, classics2, french2, english2,
  math2, disc2, music2)
> ggobi(spearman2)
```

- Fit factor model:

```
> spearman2.fa<-factanal(spearman2[,-c(1:2)],1,rotation="none")
> spearman2.fa
> spearman2.fa<-factanal(spearman2[,-c(1:2)],2,rotation="none")
> spearman2.fa
> varimax(spearman2.fa$loadings)
```