Psychology 312

Homework 6

1. (50 points). *Exploratory and Confirmatory Factor Analysis.* An article by Long and Perkins(2003) is online in the Statistics Handouts section. Long and Perkins reviewed a previously published measure that was never properly examined at the time of publication. The measure supposedly presented 12 items representing 4 constructs. Long and Perkins gathered data on these 12 variables, and performed two confirmatory factor analyses that they reported in their paper. One is a single factor model, and one is a 4 factor model with each factor loading on 3 variables. Here are descriptions of the 12 items:

I am going to read some things that people might say about their block. Each time I read one of these statements, please tell me if it is mostly true or mostly false about your block simply by saying "true" (2=MORE SOC) or "false" (1=LESS SOC).

SCI1 I think my block is a good place for me to live.
SCI2 People on this block do not share the same values. (reverse)
SCI3 My neighbors and I want the same things from the block.
SCI4 I can recognize most of the people who live on my block.
SCI5 I feel at home on this block.
SCI6 Very few of my neighbors know me. (reverse)
SCI7 I care about what my neighbors think of my actions.
SCI8 I have almost no influence over what this block is like. (reverse)
SCI9 If there is a problem on this block people who live here can get it solved.
SCI10 It is very important to me to live on this particular block.
SCI11 People on this block generally don't get along with each other. (reverse)
SCI12 I expect to live on this block for a long time.

NOTE: Items 2,6,8,11 are reverse coded! This means that the loading on this item may be negative on the designated factor.

Data for these items is available in the file *LongPerkinsTime1SCI.csv.*

Here is what I want you to do:

a. Using the **QuickCFA** function, replicate the two (one factor and 4 factor) confirmatory factor analyses of Long and Perkins, which they report in Table 2 of their article. You can load the data from the Internet and you should remove missing values, as below:

```
# create new dataset without missing data
data <-
read.csv('http://www.statpower.net/Content/312/Homework/LongPerkinsTime1SCI.csv')
### remove missing values
data <- na.omit(data)
```

b. Load the **AdvancedFactorFunctions** library.

```
source(
'http://www.statpower.net/Content/312/R Stuff/AdvancedFactorFunctionsV1.04.r')
```

c. Use the **FA.Stats** and **Scree.Plot** functions to perform exploratory factor analyses on **cor(data)** and report overall statistics for 1–5 factors. Examine the output and state what number of factors would be dictated by (1) the Scree test, (2) the Kaiser-Guttman rule, the old-fashioned "sequential chi-square test," and the RMSEA confidence interval plot. When does it say to stop adding factors? Do these methods agree or disagree on the number of factors?

d. We're going to try to find a good model with only 3 factors. Using the *exploratory-confirmatory approach* of Jöreskog (1978) via the **QuickJoreskog** function, produce a good-fitting model with 3 factors for the same 12 SCI variables, and provide your name and description for the constructs that you think each of these three factors are capturing. There is a problem with this solution, and you will find it — if you look carefully. Remember in class we talked about a "Heywood Case"? (*Hint: A variance can never be negative.*)

Try a 4 factor orthogonal and oblique solution. Try a bifactor model with 4 factors. See what you can find! On substantive and statistical grounds, try do decide which model is your favorite.

e. Long and Perkins did a follow-up analysis in which they added several variables to the analysis and derived their own factors, shown in their Table 3. You will have to study the table carefully, as it contains a lot of information in a compact and clever layout. You do not need to replicate their analysis, but simply compare your solution with theirs and evaluate the question: How do your "best solution" factors compare to theirs in terms of the substantive interpretation?

2. (20 points). Normally we think of the common factor model in terms of p, the number of variables, being substantially larger than m, the number of factors. However, Spearman actually believed that a number of mental tests could be explained in terms of only *one* factor, which he called "g" (for "general intelligence"). Spearman gathered data on a number of mental tests, and seemed to find that a factor analysis supported a single factor model. He therefore concluded that the existence of g had been verified. Suppose that you have $p = 6$ mental tests and that actually there are $m = 12$ factors underlying these six tests. Suppose moreover that these factors have loadings that do not have a nice, clean, simple structure, but are, rather, "all over the place" in an essentially random pattern, like this

$$\mathbf{F} = \begin{bmatrix} .121 & .064 & .194 & .228 & .050 & .087 & .284 & .215 & .161 & .321 & .352 & .046 \\ .109 & .013 & .211 & .303 & .218 & .331 & .256 & .102 & .127 & .329 & .278 & .129 \\ .043 & .258 & .135 & .332 & .014 & .207 & .318 & .205 & .269 & .019 & .112 & .178 \\ .230 & .009 & .366 & .344 & .436 & .081 & .058 & .221 & .283 & .154 & .193 & .067 \\ .241 & .296 & .013 & .097 & .221 & .001 & .058 & .304 & .337 & .474 & .398 & .049 \\ .076 & .384 & .006 & .152 & .105 & .312 & .370 & .370 & .270 & .176 & .199 & .312 \end{bmatrix}$$

Assume that the diagonal entries of $\mathbf{U}^2$ are values that, when added to $\mathbf{FF}'$, put 1's on the diagonal. In an effort to save you lots of computation time, I have put the **F** matrix online in a text file called *Question 2 F Matrix.txt* on the website. Use my "**`MakeFactorCorrelationMatrix`**" function (from my **`R Utility Functions`** library to create the correlation matrix exactly corresponding to the above **F.** Factor analyze this correlation matrix, using maximum likelihood factor analysis. Use a "dummy" N value of 100. Examine the scree plot and the eigenvalues of the correlation matrix. Use **`FA.Stats`** and **`Scree.Plot`** to help you decide.

a. What is the appropriate number of factors according to the statistical criteria?

b. Why does the program return the "wrong" number of factors?

c. Discuss what happened in terms of the general logic and philosophy of model fitting as a part of social science. (Incidentally, an example similar to this one was the source of some controversy early in the 20th century.)

3. (20 points) (From Rencher, *Methods of Multivariate Statistics,* p. 441). A manipulated sample of 20 voices were played to 30 judges, who rated them on 15 adjectives. Averaging over the judges produced a sample with $n = 20$ and $p = 30$. The correlation matrix is available online as *Voices.csv*.

a. Perform an exploratory factor analysis, starting from scratch.

b. Examine the statistical criteria, choose a number of factors.

c. Pick a nice rotation

d. Name the factors.

4. (10 points) At the end of Chapter 13 of his book, Rencher attempts to characterize the distinction between principal components and common factors.

Additional differences are that (3) principal component analysis requires essentially no assumptions, whereas factor analysis makes several key assumptions; (4) the principal components are unique (assuming distinct eigenvalues of **S**), whereas the factors are subject to an arbitrary rotation; and (5) if we change the number of factors, the (estimated) factors change. This does not happen in principal components. The ability to rotate to improve interpretability is one of the advantages of factor analysis over principal components. If finding and describing some underlying factors is the goal, factor analysis may prove more useful than principal components; we would prefer factor analysis if the factor model fits the data well and we like the interpretation of the rotated factors. On the other hand, if we wish to define a smaller number of variables for input into another analysis, we would ordinarily prefer principal components, although this can sometimes be accomplished with factor scores.

Rencher seems to believe that principal components cannot be rotated! In what (very limited) sense is he right, and in what (much more significant) sense is he completely wrong? (Hints: Suppose you have the first two principal components. If you rotated them, are the individual variables still principal components? Is the vectorspace spanned by the two components changed from before they were rotated? Does the amount of variability in the observed variables explained by the two components change after you rotate them?)