

## Homework 4

Psychology 312

1. (20 points) Consider the following factor pattern

$$\mathbf{F} = \begin{bmatrix} .6 & 0 \\ .5 & 0 \\ .6 & 0 \\ .5 & 0 \\ .7 & 0 \\ 0 & .8 \\ 0 & .6 \\ 0 & .4 \\ 0 & .4 \end{bmatrix}$$

Suppose this factor pattern fits a population correlation matrix *perfectly*. If this is the factor pattern for the population *correlation matrix*, and if the factors are uncorrelated, then (hint: remember the “fundamental theorem of factor analysis.”)

- a) (10 points). What is the matrix  $\mathbf{U}^2$ , the variance-covariance matrix of the factor residuals?
  - b) (5 points) What is the population correlation matrix  $\mathbf{R}$  implied by  $\mathbf{F}$ ? (Hint: Use the `R` library function provided on the course website!)
  - c) (5 points) Perform a maximum likelihood factor analysis using `factanal`, and rotate to varimax simple structure. (Hint: use the `covmat` option. If you are not sure what that is, ask for help with `?factanal`) Do you get the “right” result, i.e., does the program recover the  $\mathbf{F}$  above, or does it give you another  $\mathbf{F}$ ?
2. (25 points). Given the data in the file `confirm.sav`, use `sem` to test the hypothesis that the factor pattern has precisely the form in  $\mathbf{F}$  below, with all factors uncorrelated, using maximum likelihood.

$$\mathbf{F} = \begin{bmatrix} \theta_1 & 0 & 0 \\ \theta_2 & 0 & 0 \\ \theta_3 & 0 & 0 \\ 0 & \theta_4 & 0 \\ 0 & \theta_5 & 0 \\ 0 & \theta_6 & 0 \\ 0 & 0 & \theta_7 \\ 0 & 0 & \theta_8 \\ 0 & 0 & \theta_9 \end{bmatrix}$$

(a) (15 points). Report the chi-square statistic,  $p$ -value, and degrees of freedom. What is the implication of this statistic?

(b) (10 points) You can improve the fit of the model to the data substantially by “freeing up” one or more of the values in the factor pattern and making it a free parameter --- but which one(s)? Use a technique discussed in the *confirmatory factor analysis* handout to identify a candidate loading and modify the model. Test the new improved model and report the new results.

3. (35 points) Karl Jöreskog, in his 1978 Presidential Address paper, outlined a method (that I call “exploratory-confirmatory factor analysis”) for obtaining improved simple structure in non-trivial realistic modeling situations. Jöreskog deals with the correlated factor model, and uses the notation

$$\Sigma = \mathbf{\Lambda}\Phi\mathbf{\Lambda}' + \Psi^2$$

where  $\mathbf{\Lambda}$  is the factor pattern, and  $\Phi$  is the matrix of intercorrelations between the factors. This method consists of several steps, several of which are actually unnecessary, as we shall see. Appended to the end of this test is an excerpt from page 456 of the original paper, with my comments interspersed. . I want you to replicate the analysis as performed by Jöreskog, but with a few important differences and additions to test your problem-solving skills and knowledge. The R data file is called **Harman74.cor** and is actually available from within R in the **datasets** package. The actual correlation matrix is **Harman74.cor\$cov**. To see a description of this file, type **?factanal** and scroll to the bottom of the Help page, then click on the link.

a. First, let's retrace the steps of the original analysis. You recall that the first step was to do an exploratory factor analysis, in order to find "reference variables" for each factor. The reference variable for each factor is the variable having the highest loading on each factor. In the real world, of course, you would not necessarily be replicating someone else's analysis. Rather, you would be quite possibly starting from scratch. In that case, your first goal would be to determine the correct number of factors. Use the **eigen** command to examine the eigenvalues for each factor. Above average factors (in terms of their variance accounted for) have eigenvalues greater than 1. There are two common rules of thumb. One called the "Kaiser-Guttman rule," says retain factors corresponding to eigenvalues greater than 1.0. Another rule, called the "scree test," involves looking for breaks, or "elbows," in the plot of eigenvalues. These "elbows" show where there is a sudden relative dip in variance accounted for. You can see in the scree plot that there are two elbows, showing a very powerful first factor, followed by 3 somewhat less powerful factors, followed by a whole bunch of relatively equal, weaker factors. You can see that the 5th eigenvalue is 1.02, markedly lower than the corresponding eigenvalues for factors 2,3,4. So the scree test would suggest keeping 4 factors, while the Kaiser-Guttman rule, interpreted literally, would suggest retaining 5. See if you can get R to make a scree plot, and include this scree plot and table of eigenvalues in your output, which you should save periodically (in case it gets obliterated by a machine crash).

b. We'll retain 4 factors. Use **factanal** to extract 4 factors with a varimax rotation. If you attain a successful solution, it should match very closely the results in Table 4a, page 458, of the Jöreskog (1978) paper. Make sure to save your rotated varimax pattern to hand in.

c. Next, we will use **sem** to set up a reference factor solution, as described in the *Confirmatory Factor Analysis with R* handout. This solution involves choosing a reference variable for each factor. This is the variable that has the highest loading for that factor. These variables are underlined in Table 4a of the Jöreskog article. Other factors for the reference variable have their loadings constrained to zero. Go to each row in the factor pattern, and circle the values in the other columns for the reference variable. For example, for Visual Perception, circle the .160 in column 1, the .187 in

column 2, and the .160 in column 3. These are to be constrained to zero in the reference factor solution. Set up the model, and *be careful!*

d. Set up the model, and run sem on it. Ultimately, you should get convergence, with a chi-square of 245.358, and 186 degrees of freedom. Click on the model summary, so it is added to your workbook. Compare your solution to Table 4b in Jöreskog. You should have virtually identical results.

f. Alter your model by eliminating all the “nonsignificant” parameters that have p-values less than .05. You can simply use ## characters to make the program skip over the paths you want to eliminate. Rerun the model. If, in the previous step, you eliminated all the paths that were not significant at the .05 level, you would discover that Jöreskog did not eliminate the same paths that you did.

Try to reconcile the discrepancies. Did his “rule” on page 457 differ slightly from yours? This may explain why he did not eliminate some paths you did eliminate. However, perhaps there is another explanation. Did he make any outright gross mistakes, i.e., did he fail to eliminate any paths he definitely should have, using *his* rule? Which one? Now reassess the fit of the model, examining the chi-square statistic, and the RMSEA indices.

g. In view of your general knowledge of statistical testing, can you see any potential problems with what we have done in following the recommendations of the foremost figure in the history of structural equation modeling? Do we really believe that these coefficients are zero? Should we care whether they are exactly zero? What is the logical distinction between not being able to reject the hypothesis that a parameter is zero, and having reasonable evidence to say that it is within a certain tolerance of zero? Is it in fact possible in this case to obtain strong evidence that, say, a factor loading is less than 0.10 in absolute value?

4. (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  $\mathbf{F}$  matrix online in a text file called *Question 4 F Matrix.txt* on the website. Use my

“**MakeFactorCorrelationMatrix**” function to create the correlation matrix exactly corresponding to the above  $\mathbf{F}$ . 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. (You can make your own scree plot by using the **eigen** function.) Examine the chi-square fit statistic as well. How many factors does this information suggest that you retain?

a. Why does the program return the “wrong” number of factors?

b. 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.)

Once a sufficient number of restrictions has been imposed on the model to make it identified, standard errors for factor loadings can be estimated by the method described in Section 2. **(JS. That is, by using the confirmatory factor analysis program.)** The simplest way to achieve identification, assuming that  $\Phi$  is a correlation matrix with one's in the diagonal, is to set at least  $k - 1$  zeroes in each column of  $\Lambda$ . Sometimes one has enough knowledge about the factorial nature of the tests to be able to specify *a priori* that certain variables should not load on certain factors. If this is not possible, as in a completely exploratory analysis, one can proceed as follows:

- (i) rotate the factors orthogonally by the varimax method [Kaiser, 1958];
- (ii) raise the varimax factor loadings to some power 3 or 4 while retaining their signs. Use these numbers as a target and perform a promax [Hendrickson and White, 1964] procrustes rotation. This produces an oblique solution. **(JS. Our modern software can do this for us. We simply select a Promax rotation. Moreover, in the actual 24 variable example in the paper, Jöreskog performs a varimax rotation. Since we only have the correlation matrix for the example, we'll use perform a varimax rotation.)**
- (iii) in the promax **(JS. varimax)** factor matrix find the largest factor loading in each column and, assuming that these are in different rows, rotate the original unrotated factor matrix or the varimax factor matrix so that the other loadings in these rows become zero. This results in an oblique solution in which each factor vector passes through one test point.

The solution obtained in (iii) has  $k - 1$  zeroes in each column of  $\Lambda$  and has exactly the same fit to the data as any other rotated solution. A confirmatory analysis with the same fixed zeroes will yield the same non-zero factor loadings but standard errors for these can now be obtained. An inspection of the loadings in relation to their standard errors will usually reveal that a large number of the loadings are insignificant. By eliminating these, i.e. by setting them equal to zero, one can reduce a lot of noise in the model and estimate the really significant loadings more precisely. **(JS. Careful inspection of the above paragraph reveals that you can actually completely skip the rotation step,**

because it is always possible to rotate a factor pattern so that it satisfies the above conditions. Consequently, simply use the confirmatory factor analysis program to zero the appropriate loadings in each row and column, perform the analysis, and you will have precisely the same output as shown on page 459.)

To obtain the best-fitting simple structure, proceed with the next step:

(iv) set all insignificant loadings to zero. The resulting solution will not in general have a significantly worse fit than the original solution but will display a neat simple structure with many zero loadings. Since it is a possibility that the loadings that were set to zero in step (iii) are not zero, it is recommended as a safeguarding final step that the zero loadings are checked again as follows. **(JS. Class: you may skip this step for the purposes of the assignment.)** Find the largest absolute derivative of **(JS the discrepancy function)** with respect to the fixed zero loadings and relax this particular zero loading while keeping all other zeroes fixed. This loading will make the function decrease maximally and hence make the largest improvement in fit. **(JS. Now we would use modification indices.)** If the improvement in fit is significant this step should be repeated again. Otherwise the final solution is “best-fitting” in the following sense:

- (a) all non-zero loadings are significant;
- (b) all zero loadings are such that if they were relaxed they would not be significant.