STA380.18
Homework on Factor Analysis

Directions: *Be sure to show your work and explain your answer for each question, even if the question seems to require only a Yes or No answer. Your homework solutions are to be entirely your own effort. You may not communicate with anyone about the homework, except for the TA and/or the instructor. You may use the Canvas postings, in-class discussion, any of the recommended textbooks, and computer software, if necessary, but no other resources. In writing up your solutions, it is recommended to support your answers with cut-and-pasted output, provided your answers are clearly labeled and circled or highlighted. The grader will not search through unlabeled computer output to try to find your answers.*

For your reference:
Some important facts and formulas from statistics and probability:

- $Corr(X,Y) = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$, where *Corr* mean correlation and *Cov* means covariance.
- In a simple linear regression of Y on X, the slope is $Corr(Y,X)\dfrac{\sigma_Y}{\sigma_X}$.
- If X and Y are uncorrelated, then $Cov(X,Y) = Corr(X,Y) = slope = 0$.
- In a multiple regression, if the predictor variables are all uncorrelated with each other, then their slope coefficients are all the same as they would be in separate simple regressions.
- For any random variables *X, Y, U, V,* and constants *a, b, c, d,* we have $Cov(aX+bY, cU+dV) = acCov(X,U) + adCov(X,V) + bcCov(Y,U) + bdCov(Y,V)$.

Set-up for questions 1-4:
Suppose you have three manifest variables X_1, X_2, X_3. Consider the following two-factor model (in which the common and unique factors satisfy the usual assumptions as to uncorrelatedness and standardization):

$$\left\{\begin{array}{rcl} X_1 &=& 0.8\xi_1 + 0.4\xi_2 + \varepsilon_1 \\ X_2 &=& 0.6\xi_1 + 0.6\xi_2 + \varepsilon_2 \\ X_3 &=& 0.4\xi_1 + 0.8\xi_2 + \varepsilon_3 \end{array}\right\}.$$

1.
   a) Calculate the correlation matrix of X_1, X_2, X_3.
   b) Calculate the communalities of X_1, X_2, X_3.

2.
   a) How much of the total variance of X_1, X_2, X_3 is explained by each factor?
   b) Can this model be computationally equivalent to a principal components analysis? Explain.

3. Consider the orthonormal rotation matrix $M = \begin{bmatrix} .7071 & .7071 \\ .7071 & -.7071 \end{bmatrix}$. Let *F* denote the factor pattern matrix for the two-factor model of the set-up.

a) Compute the rotated factor pattern $F^* = FM$. Verify that the rotated factor pattern yields the same correlation matrix for X_1, X_2, X_3 as the original factor pattern matrix *F* in question 1(a).

b) Calculate the communalities of X_1, X_2, X_3 in the rotated two-factor model. Are they the same as in question 1(b)? Explain why or why not.

4. Consider the one-factor model $\left\{ \begin{matrix} X_1 & = & a\xi + \varepsilon_1 \\ X_2 & = & b\xi + \varepsilon_2 \\ X_3 & = & c\xi + \varepsilon_3 \end{matrix} \right\}$, in which the common and unique factors satisfy the usual assumptions as to uncorrelatedness and standardization. *a*, *b*, *c* are constants.

a) Is it possible for the one-factor model to produce the same correlation matrix for X_1, X_2, X_3 as the two-factor model produces? If so, provide values for *a*, *b*, *c* that achieve this goal. If not, then explain why not.

b) What is an important implication of your findings in questions 3(a) [in the preceding question] and 4(a) [in this question]?

Set-up for questions 5-10:

The text file called **EvaluateSupervisors.Data.txt** contains information on a random sample of 30 supervisory personnel from a large corporation. Each supervisory position was evaluated by employees on six different dimensions of performance. An overall rating was also obtained. The scale of the various ratings is unimportant for the purposes of this analysis. The variables are described in the file.

Create a SAS dataset called **WORK.EvaluateSupervisors** to contain all of the data on the 30 supervisors. Please name the columns **OVERALL, BEEFS, PRIVILEGE, NEWLEARN, RAISES, CRITICAL, ADVANCE** in the SAS dataset. Interest ultimately focuses on understanding the structure that may underlie those variables that are related to the overall rating. Throughout the analysis, exclude **OVERALL** from the set of variables for which you determine underlying factors. Note also that there are 11 lines of descriptive text in the file that must be excluded from the data to be read.

5. Run a "principal components" style of factor analysis to extract 6 factors.

a) Submit your SAS code.

b) How many factors would you retain? Why?

c) Try to interpret the first two factors.

6. How well do the six factors of question 5 explain the OVERALL supervisor rating – individually and collectively?

7. Run a principal factor analysis, with R-square type initial estimates of communalities, to extract six factors.

a) Submit your SAS code.
b) Try to interpret the first two factors.
c) Succinctly compare the factor analysis of Q7 with the factor analysis of Q5.

8. Run a principal factor analysis, with R-square type initial estimates of communalities, to extract two factors, apply a varimax rotation to the initial factor pattern, and add factor scores for all 30 supervisors to a dataset called **WORK.EvaluateSupervisors_scores** (including all original data as well).

a) Submit your SAS code.
b) Are the first two initial factors the same as the first two factors in the preceding question?
c) Verify that the varimax factor rotation matrix is orthonormal.
d) Try to interpret the varimax rotated factor pattern.

9. Calculate means, standard deviations, and correlation for the factor scores of the 30 supervisors of the varimax rotated factors that you stored in WORK.EvaluateSupervisors_scores in the preceding question. *[You may use* PROC CORR*, which includes* PROC MEANS *output by default.]*

a) Given the assumptions of factor analysis, do your summary statistics surprise you? Why or why not?
b) Manually compute the factor scores of the first supervisor in the dataset and verify that the SAS-calculated factor scores are correct for him/her.

10. Extract the default number of factors by the maximum likelihood method with R-square type initial estimates of communalities. You may need to use the ULTRAHEYWOOD or HEYWOOD option to deal with estimated communalities that exceed 1. You may assume that the assumptions of the maximum likelihood model apply.

a) Does the output provide support for the hypothesis that common factors exist?
b) Does the output provide support for the hypothesis that the default number of extracted factors is adequate?