Homework 5 (100 points maximum)
Psychology 312

1. (25 points). In his new book on *Constrained Principal Component Analysis and Related Techniques*, Yoshio Takane discusses an interesting example based on data from a 1978 article in *Biological Psychiatry* by Mezzich. The data were gathered by asking 11 psychiatrists to rate 4 "prototypical" cases on the basis of 17 rating scales. The ratings were on a 7 point scale ranging from 0 to 6, with 0 standing for "does not apply at all" and 6 standing for "applies very well." The 4 cases are labeled MDD (Manic-Depressive Depressed), MDM(Manic-Depressive Manic), SSP (Simple Schizophrenia), PSP (Paranoid Schizophrenia).

   The 17 rating scales are, for graphical simplicity, labeled A-Q,

A.	Somatic concern
B.	Anxiety
C.	Emotional withdrawal
D.	Conceptual disorganization
E.	Guilt feeling
F.	Tension
G.	Mannerism and posturing
H.	Grandiosity
I.	Depressive mood
J.	Hostility
K.	Suspiciousness
L.	Hallucinatory behavior
M.	Motor retardation
N.	Uncooperativeness
0.	Unusual thought content
P.	Blunted affect
Q.	Excitement

   The data file *mezzich.csv* contains data for the 17 rating variables, along with codes for the psychiatrist, and for the prototypical patient type being evaluated. There are also 4 binary indicator variables representing the 4 patient categories.

   (a) (10) Load in the data and isolate the 17 rating variables and the patient type variable. On this reduced data set, create a PCA data object called `res.pca` with the FactoMineR `PCA` command with the `patient` category as a qualitative supplementary variable. Then use FactoMineR `plot` command to display the variables factor map. What clinical dimensions do the 2 main dimensions represent?

   (b) (10) Examine the individuals factor map, displaying the 4 patient types on the plot in different colors with the command
   `plot(res.pca,choix="ind",habillage = 18)`.

Note that the average rating for each diagnostic type is displayed in the appropriate color on the plot. Does the positioning of the 4 patient types make sense in terms of your labeling of the diagnostic dimensions? Explain how. Which ratings seem most "out of line" with the overall diagnostic trends for the other pychiatrists' ratings of patients of the same type?

(c) (5) Replace the patient type with the psychiatrist ID as variable 18 in the file. Redo the individuals factor map, displaying the psychiatrists as a supplementary variable. Surveying the display of the average ratings produced by the psychiatrists, what important facts can you glean from the plot?

2. (35 points). The classic common factor analysis model can be fit to the Mezzich data we looked at in question 1. We will use the Advanced Factor Functions from the R Support Materials section of the website. A bug was recently fixed in this file, so be sure to download the latest version. You should work your way through the handout *Advanced Exploratory Factor Analysis with R.*

   (a) (10) Compute the correlation matrix for the 17 ratings, and use the `FA.Stats` function to draw an RMSEA plot for 1 to 5 factors. Also, re-examine the Scree Plot for this correlation matrix. Let's start with the Scree Plot. What do the Scree test and the Kaiser-Guttman rule suggest as the correct number of factors?

   (b) (5) What does the sequential χ^2 test suggest as the correct number of factors?

   (c) (10) Examine the RMSEA plot, and compare it with the comparable plot for the 24 Psychological Variables discussed in the Advanced Exploratory Factor Analysis handout. Do you see a noticeable difference between the plots? Is there some aspect of the plot that might make you reconsider going "all the way to 5 factors"?

   (d) (10) Using the `MLFA` function, generate 3- and 4-factor solutions. Examine the Varimax patterns for both solutions, and *name the factors.* (Sixteen of the 17 items are described in the original article by Overall and Gorham, 1962, which is posted at the website.) What is gained, conceptually, by going from a 3-factor to a 4-factor solution?

3. (20 points) Consider the following factor pattern

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

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). What is the matrix $\boldsymbol{U}^2$, the variance-covariance matrix of the factor residuals?

(b) (5) What is the population correlation matrix $\boldsymbol{R}$ implied by $\boldsymbol{F}$? (Hint: Use the R library function `MakeFactorCorrelationMatrix` provided in the *Steiger R Library Functions* on the course website! You can input a dummy value of n, like, say, 100.)

(c) (5) Perform a maximum likelihood factor analysis using `MLFA`, examine the varimax simple structure pattern. Do you get the "right" result, i.e., does the program recover the $\boldsymbol{F}$ above, or does it give you another $\boldsymbol{R}$?

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

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

Assume that the diagonal entries of $\boldsymbol{U}^2$ are values that, when added to $\boldsymbol{F}\boldsymbol{F}'$, put 1's on the diagonal. In an effort to save you lots of computation time, I have put the $\boldsymbol{F}$ matrix online in a text file called *Q04FMatrix.txt* on the website. Use the `MakeFactorCorrelationMatrix` function to create the correlation matrix exactly corresponding to the above $\boldsymbol{F}$.

(a) (10) 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. What values of m, the number of factors, do these analyses suggest.Examine the table of chi-square fit statistics and RMSEA confidence intervals as well. How many factors does this information suggest that you retain?

(b) (5) The true number of factors is 12. Why does the program return the "wrong" number of factors?

(c) (5) 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.)