STA 431s23 Assignment Ten[1]

This assignment is for the quiz on Monday April 10th, makeup day. For the quiz, please bring a printout of your full R input and output for Question 15. The other problems are not to be handed in. They are practice for the Quiz.

---

1. Let

$$\begin{aligned} D_1 &= \lambda_1 F_1 + e_1 \\ D_2 &= \lambda_2 F_2 + e_2 \\ D_3 &= \lambda_3 F_3 + e_3, \end{aligned}$$

where F_1, F_2, F_3, e_1, e_2 and e_3 are all independent with $F_j \sim N(0,1)$ and $e_j \sim N(0,\omega_j)$. All the expected values are zero. You can tell from the notation which variables are observable.

   (a) Give the variance-covariance matrix of the observable variables.

   (b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

   (c) Even though the parameters are not identifiable, the model itself is testable. That is, it implies a set of equality restrictions on the covariance matrix $\boldsymbol{\Sigma}$ that could be tested, and rejecting the null hypothesis would call the model into question. State the null hypothesis. Again, it is a statement about the $\sigma_{i,j}$ values.

2. Here is another factor analysis model. This one has a single underlying factor.

$$\begin{aligned} D_1 &= \lambda_1 F + e_1 \\ D_2 &= \lambda_2 F + e_2 \\ D_3 &= \lambda_3 F + e_3, \end{aligned}$$

where the factor and error terms are all independent, $F \sim N(0,1)$, $e_j \sim N(0,\omega_j)$, and λ_1, λ_2 and λ_3 are nonzero constants with $\lambda_1 > 0$.

   (a) Give the variance-covariance matrix of the observed variables.

   (b) Are the model parameters identifiable? Answer Yes or No and prove your answer. You are proving part of the 3-variable rule, so don't just cite it.

3. Suppose we added another variable to the model of Question 2. That is, we add

$$D_4 = \lambda_4 F + e_4,$$

with assumptions similar to the ones of Question 2. Now suppose that $\lambda_2 = 0$, while the other factor loadings are non-zero.

   (a) Is λ_2 identifiable? Justify your answer.

(b) Are the other factor loadings identifiable? Justify your answer.

4. Suppose we added a fifth variable to the model of Question 3. That is, we add

$$D_5 = \lambda_5 F + e_5,$$

with assumptions similar to the ones of Question 2. Now suppose that $\lambda_3 = \lambda_4 = 0$, while the other factor loadings are non-zero.

(a) Are λ_3 and λ_4 identifiable? Justify your answer.

(b) Are the other three factor loadings identifiable? Justify your answer.

(c) State the general pattern that is emerging here.

5. We now extend the model of Question 2 by adding a second factor. Let

$$\begin{aligned} D_1 &= \lambda_1 F_1 + e_1 \\ D_2 &= \lambda_2 F_1 + e_2 \\ D_3 &= \lambda_3 F_1 + e_3 \\ D_4 &= \lambda_4 F_2 + e_4 \\ D_5 &= \lambda_5 F_2 + e_5 \\ D_6 &= \lambda_6 F_2 + e_6, \end{aligned}$$

where all expected values are zero, $Var(e_i) = \omega_i$ for $i = 1, \ldots, 6$, $Var(F_1) = Var(F_2) = 1$, $Cov(F_1, F_2) = \phi_{12}$, the factors are independent of the error terms, and all the error terms are independent of each other. All the factor loadings are non-zero, and they might be positive or negative.

(a) Give the covariance matrix of the observable variables. Show the necessary work. A lot of the work has already been done.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

(c) Write the model in matrix form as $\mathbf{d} = \mathbf{\Lambda F} + \mathbf{e}$. That is give the matrices. For example, $\mathbf{d}$ is 6×1.

(d) Recall that a *rotation* matrix is any square matrix $\mathbf{R}$ satisfying $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$. Give a specific 2×2 rotation matrix $\mathbf{R}$ so that $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_2 = \mathbf{\Lambda R}$ yield the same $\mathbf{\Sigma} = cov(\mathbf{D})$. Hint: Use your answer to Question 5b.

(e) Suppose we add the conditions $\lambda_1 > 0$ and $\lambda_4 > 0$. Are the parameters identifiable now?

(f) In a goodness of fit test for this model, what are the degrees of freedom?

6. In Question 5, suppose we added just two variables along with the second factor. That is, we omit the equation for D_6, while keeping $\lambda_1 > 0$ and $\lambda_4 > 0$. Are the model parameters identifiable in this case? Answer Yes or No. Calculate or cite a rule.

7. Let's add a third factor to the model of Question 5. That is, we keep the equation for D_6 and add

$$\begin{aligned} D_7 &= \lambda_7 F_3 + e_7 \\ D_8 &= \lambda_8 F_3 + e_8 \\ D_9 &= \lambda_9 F_3 + e_9 \end{aligned}$$

with $\lambda_1 > 0$, $\lambda_4 > 0$, $\lambda_7 > 0$ and other assumptions similar to the ones we have been using. Are the model parameters identifiable? You don't have to do any calculations if you see the pattern.

8. In this factor analysis model, the observed variables are *not* standardized, and the factor loading for D_1 is set equal to one. Let

$$\begin{aligned} D_1 &= F + e_1 \\ D_2 &= \lambda_2 F + e_2 \\ D_3 &= \lambda_3 F + e_3, \end{aligned}$$

where $F \sim N(0, \phi)$, e_1, e_2 and e_3 are normal and independent of F and each other with expected value zero, $Var(e_1) = \omega_1, Var(e_2) = \omega_2, Var(e_3) = \omega_3$, and λ_2 and λ_3 are nonzero constants.

   (a) Calculate the variance-covariance matrix of the observed variables.

   (b) Are the model parameters identifiable? Answer Yes or No and prove your answer. You are proving another part of the 3-variable rule, so don't just cite it.

9. We now extend the preceding model by adding another factor. Let

$$\begin{aligned} D_1 &= F_1 + e_1 \\ D_2 &= \lambda_2 F_1 + e_2 \\ D_3 &= \lambda_3 F_1 + e_3 \\ D_4 &= F_2 + e_4 \\ D_5 &= \lambda_5 F_2 + e_5 \\ D_6 &= \lambda_6 F_2 + e_6, \end{aligned}$$

where all expected values are zero, $Var(e_i) = \omega_i$ for $i = 1, \ldots, 6$,

$$cov \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix},$$

and $\lambda_2, \lambda_3, \lambda_5$ and λ_6 are nonzero constants.

   (a) Give the covariance matrix of the observable variables. Show the necessary work. A lot of the work has already been done in Question 8.

   (b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

10. Let's add a third factor to the model of Question 9. That is, we add

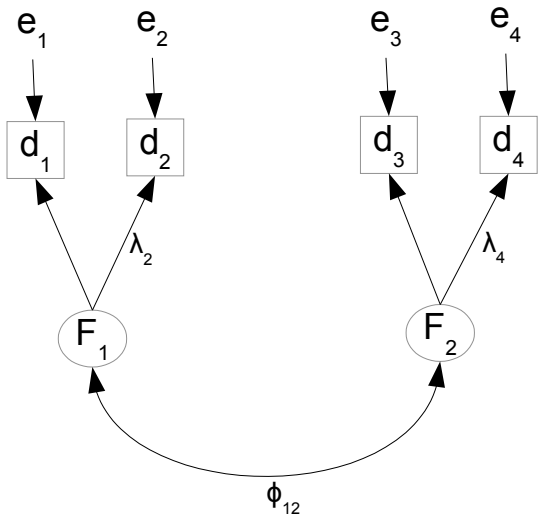
$$\begin{aligned} D_7 &= F_3 + e_7 \\ D_8 &= \lambda_8 F_3 + e_8 \\ D_9 &= \lambda_9 F_3 + e_9 \end{aligned}$$

and

$$cov \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{12} & \phi_{22} & \phi_{23} \\ \phi_{13} & \phi_{23} & \phi_{33} \end{pmatrix},$$

with $\lambda_8 \neq 0$, $\lambda_9 \neq 0$ and so on. Are the model parameters identifiable? You don't have to do any calculations if you see the pattern.

11. This question leads to the two-variable, two-factor rule. Consider the following path diagram.



(a) This is definitely a surrogate model. Give the equations of the original *uncentered* model.

(b) The ϕ_{12} in the path diagram is actually ϕ'_{12}. Express ϕ'_{12} in terms of the parameters of the original model.

(c) Give the covariance matrix for the surrogate model. Omit the primes from now on.

(d) Assuming λ_2, λ_4 and ϕ_{12} are all non-zero, show that all the parameters are identifiable. This is the 2-variable 2-factor rule, so don't just cite it.

(e) Counting parameters and covariance structure equations, how many equality constraints on the covariance matrix should be implied by the model?

(f) What is the equality constraint? Multiply through by denominators so that there are no fractions.

(g) Would this equality constraint hold even with zero values for some of λ_2, λ_4 and ϕ_{12}?

12. It's helpful to have a version of the Extra Variables Rule that does not depend on reference variables. That way, for example, we could freely add variables to the bi-factor model. Accordingly, let

$$\begin{aligned} \mathbf{d}_1 &= \mathbf{\Lambda}_1\mathbf{F} + \mathbf{e}_1 \\ \mathbf{d}_2 &= \mathbf{\Lambda}_2\mathbf{F} + \mathbf{e}_2 \\ \mathbf{d}_3 &= \mathbf{\Lambda}_3\mathbf{F} + \mathbf{e}_3 \end{aligned}$$

where the random vectors $\mathbf{d}_1$ and $\mathbf{F}$ are $p \times 1$, and the $p \times p$ matrix $\mathbf{\Lambda}_1$ has an inverse, which it certainly will if the elements of $\mathbf{d}_1$ are reference variables. The factors are independent of the error terms.

Suppose that $\mathbf{d}_1$ and $\mathbf{d}_2$ belong to a model whose parameters have already been identified somehow, and we want to add $\mathbf{d}_3$ to the model. That is, $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, $\mathbf{\Phi}$, $\mathbf{\Omega}_{11}$, $\mathbf{\Omega}_{12}$ and $\mathbf{\Omega}_{22}$ are identified, and we seek to identify $\mathbf{\Lambda}_3$, $\mathbf{\Omega}_{33}$ and $\mathbf{\Omega}_{23}$. It will be assumed that $\mathbf{\Omega}_{13} = \mathbf{O}$.

(a) Write $\mathbf{\Sigma} = cov\begin{pmatrix} \mathbf{d}_1 \\ \hline \mathbf{d}_2 \\ \hline \mathbf{d}_3 \end{pmatrix}$ as a partitioned matrix.

(b) Show that $\mathbf{\Lambda}_3$, $\mathbf{\Omega}_{33}$ and $\mathbf{\Omega}_{23}$ are identifiable.

13. Suppose that the parameters of factor analysis models for two non-overlapping sets of observable variables are identifiable, and we want to combine the two models. Suppose there are p_1 factors in model one and p_2 factors in model two, and the models can be written as

$$\begin{aligned} \mathbf{d}_1 &= \mathbf{\Lambda}_1\mathbf{F}_1 + \mathbf{e}_1 \\ \mathbf{d}_2 &= \mathbf{\Lambda}_2\mathbf{F}_1 + \mathbf{e}_2 \\ \mathbf{d}_3 &= \mathbf{\Lambda}_3\mathbf{F}_2 + \mathbf{e}_3 \\ \mathbf{d}_4 &= \mathbf{\Lambda}_4\mathbf{F}_2 + \mathbf{e}_4, \end{aligned}$$

where $\mathbf{d}_1$ is $p_1 \times 1$, $\mathbf{d}_3$ is $p_2 \times 1$, and the square matrices $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_3$ both have inverses. These conditions will definitely be satisfied if $\mathbf{d}_1$ contains reference variables for $\mathbf{F}_1$ and $\mathbf{d}_2$ contains reference variables for $\mathbf{F}_2$.

(a) The parameter matrices of the combined model are $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, $\mathbf{\Lambda}_3$, $\mathbf{\Lambda}_4$, $\mathbf{\Phi}_{11}$, $\mathbf{\Phi}_{12}$, $\mathbf{\Phi}_{22}$, $\mathbf{\Omega}_{11}$, $\mathbf{\Omega}_{12}$, $\mathbf{\Omega}_{13}$, $\mathbf{\Omega}_{14}$, $\mathbf{\Omega}_{22}$, $\mathbf{\Omega}_{23}$, $\mathbf{\Omega}_{24}$, $\mathbf{\Omega}_{33}$, $\mathbf{\Omega}_{34}$ and $\mathbf{\Omega}_{44}$ — except let's set $\mathbf{\Omega}_{13} = cov(\mathbf{e}_1, \mathbf{e}_3) = \mathbf{O}$. Assuming $\mathbf{\Omega}_{13} = \mathbf{O}$, what parameters need to be identified for the combined model to be identifiable?

(b) Show how $\mathbf{\Phi}_{12} = cov(\mathbf{F}_1, \mathbf{F}_2)$ can be identified.

(c) Show how $\mathbf{\Omega}_{14} = cov(\mathbf{e}_1, \mathbf{e}_4)$ can be identified.

(d) Show how $\mathbf{\Omega}_{23} = cov(\mathbf{e}_2, \mathbf{e}_3)$ can be identified.

(e) Show how $\mathbf{\Omega}_{24} = cov(\mathbf{e}_2, \mathbf{e}_4)$ can be identified.

14. All the models with identifiable parameters are surrogate models — all of them. Consider the model of Question 5.

(a) Write the equations of the *original* uncentered model. You don't have to give additional specifications; just write the equations.

(b) Noting that the equations of the centered original model look exactly like the ones in Question 5, show how the model of Question 5 arises from a re-parameterization of the centered original model by a change of variables — actually, two changes of variables. Do it this way.

i. Re-write the model equations, showing what happens to the factor loadings.

ii. Denote the variances and covariances of factors covariance under the original model by ϕ_{ij}, and the variances and covariances under the surrogate model as ϕ'_{ij} What is ϕ'_{12} in terms of the parameters of the original model?

iii. How are the ω_j affected by the re-parameterization?

(c) How much of $Var(D_2)$ is explained by F_1 under the original model?

(d) How much of $Var(D_2)$ is explained by F'_1 under the surrogate model? Make sure you put a prime on the parameter(s).

(e) Did re-parameterization affect the explained variance?

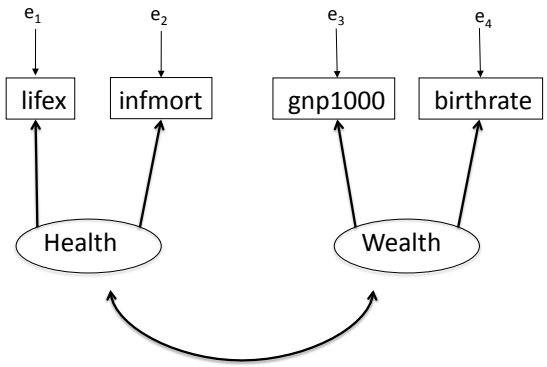
15. The R part of this assignment is based on the Poverty Data. The data are given in the file http://www.utstat.toronto.edu/brunner/data/illegal/poverty.data.txt. This data set contains information from a sample of 97 countries. In order, the variables include Live birth rate per 1,000 of population, Death rate per 1,000 of population, Infant deaths per 1,000 of population under 1 year old, Life expectancy at birth for males, Life expectancy at birth for females, and Gross National Product per capita in U.S. dollars. There is also a variable with numeric values representing continent, and finally the name of the country.

    When you read the data, use the `na.strings = "."` option on `read.table`. This is so that the SAS missing value code, a period, will be treated as `NA`.

    The poverty data set can be very challenging and frustrating to work with, because correlated measurement errors produce negative variance estimates and other numerical problems almost everywhere you turn. To make your job easier (possible), please confine your analyses to the following four variables:

    - Life Expectancy: Average of life expectancy for males and life expectancy for females.
    - Infant mortality rate.
    - Birth rate.
    - GNP/1000 = Gross national product in thousands of dollars. The re-scaling is a solution to numerical problems in fitting the model.

    Here is a picture of a factor analysis model with 2 factors.



    The reason for making birth rate an indicator of wealth is that birth control costs money.

    (a) Fit the model with `lavaan`. Don't bootstrap. My value of $\widehat{\omega}_1$ is 1.432. You can't use the original model; you'll have to re-parameterize. Which of the two standard re-parameterizations should you choose? Suppose we are interested in the correlation between Health and Wealth.

    (b) Does this model fit the data adequately? Answer Yes or No, and back up your answer with two numbers from the printout: the value of a test statistic, and a p-value.

    (c) Why does the goodness of fit test have one degree of freedom?

(d) What is the maximum likelihood estimate of the correlation between factors? The answer is a single number from the printout.

(e) Give a 95% confidence interval for the correlation between factors. Do it the easy way. What is funny about the confidence interval?

(f) Now fit a model with the other common re-parameterization. Request standardized output when you apply `summary`.

i. Compare the two likelihood ratio tests for model fit. What do you see?

ii. Compare the two $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ matrices (not part of summary; see lecture slides). What do you see?

iii. Give the maximum likelihood estimate of λ_2/λ_1 for the *original* model, based on output from the *first* surrogate model you fit. Can you find this number in the output from the second model?

iv. Based on the output from the second model, give the maximum likelihood estimate of the correlation between Health and Wealth. Can you find this number in the output from the first model?

v. Estimate the proportion of variance in infant mortality rate that is explained by the Health factor. You might need a calculator, but it's quick if you do it the easy way.

(g) Finally, the high estimated correlation between factors suggests that there might be just one underlying factor: wealth. Try a single-factor model and see if it fits. Locate the relevant chi-squared statistic, degrees of freedom and p-value. Do the estimated factor loadings make sense? What do you conclude? Do you like the one-factor model or the two-factor model?

Please bring a printout of your full R input and output for Question 15 to the quiz.