STA 2053 Assignment 5 (Mostly factor analysis, the general model and surrogate models)[1]

Please bring your complete R input and output for Question 5 to the quiz. The other questions are not to be handed in. They are practice for the quiz on November 28th.

1. The following model is centered, and has zero covariance between all pairs of exogenous variables including error terms. Only W_1, W_2, V_1 and V_2 are observable.

$$
\begin{array}{rcl}
Y_1 & = & \gamma_1 X_1 + \gamma_2 X_2 + \epsilon_1 \\
Y_2 & = & \beta Y_1 + \gamma_3 X_1 + \epsilon_2 \\
W_1 & = & \lambda_1 X_1 + e_1 \\
W_2 & = & \lambda_2 X_2 + e_2 \\
V_1 & = & \lambda_3 Y_1 + e_3 \\
V_2 & = & \lambda_4 Y_2 + e_4
\end{array}
$$

   (a) Make a path diagram.

   (b) Referring to the general two-stage structural equation model

$$
\begin{array}{rcl}
\mathbf{y}_i & = & \boldsymbol{\beta}\mathbf{y}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i & = & \left(\begin{array}{c} \mathbf{x}_i \\ \mathbf{y}_i \end{array} \right) \\
\mathbf{d}_i & = & \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i,
\end{array}
$$

   write the model equations in matrix form. This means put symbols from the model above in the matrices. Also give the matrices $\boldsymbol{\Phi}_x = cov(\mathbf{x}_i)$, $\boldsymbol{\Psi} = cov(\boldsymbol{\epsilon}_i)$, $\boldsymbol{\Omega} = cov(\mathbf{e}_i)$ and $\boldsymbol{\Phi} = cov(\mathbf{F}_i)$ in terms of the parameters of this specific model.

   (c) Is the entire parameter vector identifiable? How do you know?

2. Consider the general factor analysis model

$$
\mathbf{d}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i,
$$

   where $\boldsymbol{\Lambda}$ is a $k \times p$ matrix of factor loadings, the vector of factors $\mathbf{F}_i$ is a $p \times 1$ multivariate normal with expected value zero and covariance matrix $\boldsymbol{\Phi}$, and $\mathbf{e}_i$ is multivariate normal and independent of $\mathbf{F}_i$, with expected value zero and covariance matrix $\boldsymbol{\Omega}$. All covariance matrices are positive definite.

   (a) Calculate the matrix of covariances between the observable variables $\mathbf{d}_i$ and the underlying factors $\mathbf{F}_i$.

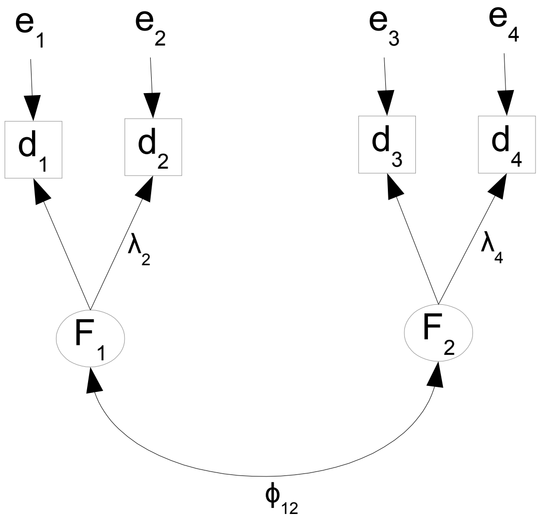
   (b) Give the covariance matrix of $\mathbf{d}_i$.

[1] This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/brunner/oldclass/2053f22

(c) Because $\mathbf{\Phi}$ symmetric and positive definite, it has a square root matrix that is also symmetric. Using this, show that the parameters of the general factor analysis model are not identifiable.

(d) In an attempt to obtain a model whose parameters can be successfully estimated, let $\mathbf{\Omega}$ be diagonal (errors are uncorrelated) and set $\mathbf{\Phi}$ to the identity matrix (standardizing the factors). Show that the parameters of this revised model are still not identifiable. Hint: An orthogonal matrix $\mathbf{R}$ (corresponding to a rotation) is one satisfying $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$.

3. 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.

(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}?

4. 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. If the errors of the first model have zero covariance with the errors of the second model, the only parameters to be identified are the covariances in $cov(\mathbf{F}_1, \mathbf{F}_2)$. Suppose there are p_1 factors in model one and p_2 factors in model two, and the models can be written as

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

   where $\mathbf{d}_1$ is $p_1 \times 1$, $\mathbf{d}_3$ is $p_2 \times 1$, and the square matrices $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_3$ 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 $\mathbf{F}_2$. Show how $\boldsymbol{\Phi}_{12} = cov(\mathbf{F}_1, \mathbf{F}_2)$ can be identified.

5. The Arthritis data are simulated, but engineered to reproduce sample statistics from the baseline period of an actual study. This way there are no issues of data ownership or copyright, and also the study design is better than the original.

   Your job is to fit a reasonable model (following the guidance below), and try to answer the main question of the study: How does exercise at time one affect pain at time two? Here are the details.

   In a study of exercice and arthritis pain, rheumatoid arthritis patients were clinically assessed for disease severity by a physician. Disease severity was also estimated by X-rays (based on joint erosion) and a blood test (based on elevated ESR and C-reactive protein, rheumatoid factor and anti-citrullinated protein antibody). The doctor made the clinical assessment before seeing the X-ray and blood test results.

   One week later, patients and their spouses came into the clinic again. They both filled out questionnaires and the patients had more tests. Pain was measured in two ways: self-report about pain during the preceding week, and an electroencephalograph (EEG, or brain wave) test. In the EEG test, electrodes on the patient's scalp measured electrical activity in the brain during a standard passive joint movement exam. Passive means the patient relaxes while a technician moves the joint gently through a moderate range of motion. In the absence of loud noises or emotionally arousing stimuli, general autonomic nervous system activation is a fairly dependable indication of subjective pain.

   Exercise/physical activity level during the preceding week was measured in three ways: self-report, spouse's report, and by accelerometer, a motion detector/fitness tracker that the patient had been wearing during the past week.

   One week after that, patients and their spouses came in again, and the same measurements of pain and exercise were collected a second time. Here are the observable variables.

```
clinical    = Disease severity based on clinical assessment
xray        = Disease severity based on x-ray
blood       = Disease severity based on blood test
selfpain1   = Self-reported pain at time one
```

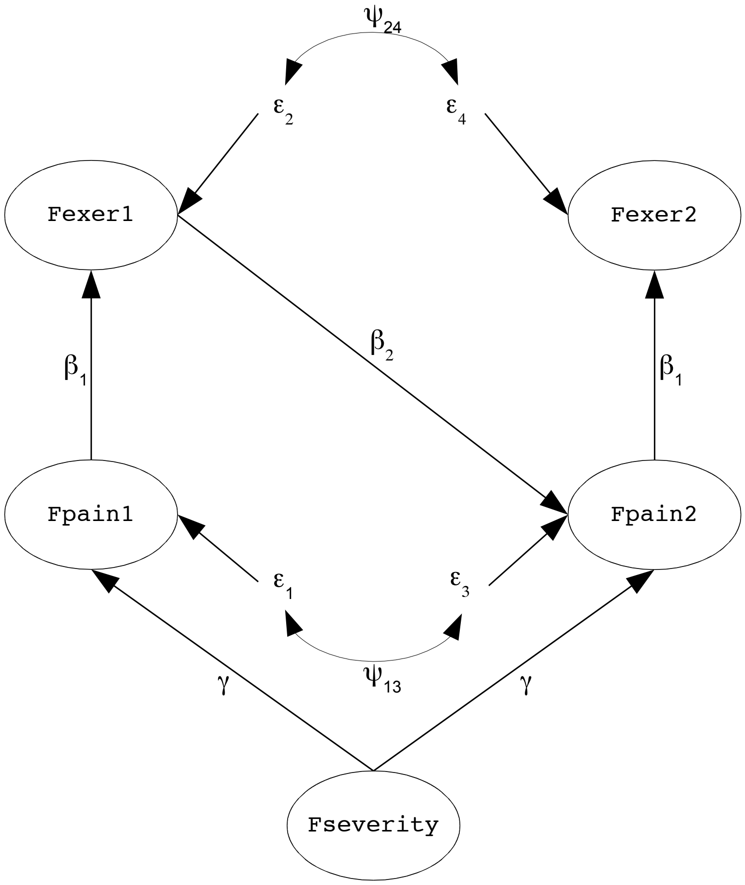
```
EEG1        = Pain assessed from brain waves at time one
selfexer1   = Self-reported exercise/physical activity at time one
spouseexer1 = Spouse report of exercise/physical activity at time one
acceler1    = Accelerometer (fitness tracker) data at time one
selfpain2   = Self-reported pain at time two
EEG2        = Pain assessed from brain waves at time two
selfexer2   = Self-reported exercise/physical activity at time two
spouseexer2 = Spouse report of exercise/physical activity at time two
acceler2    = Accelerometer (fitness tracker) data at time two
```

The raw data are in the form of a Microsoft Excel spreadsheet. They are available from

http://www.utstat.toronto.edu/∼brunner/data/legal/Arthritis1.xls .

I used the `readxl` package to get the data into R.

My (surrogate) measurement model is routine, but I doubt that you would come up with exactly my latent variable model. Like all models it's debatable, but this is what you should use. Here is a path diagram, followed by a bit of explanation.



Rheumatoid arthritis is an auto-immune disorder with no known cure. It tends to get worse very gradually over time. This is why there's only one latent disease severity. This is a fairly stable system, so severity affects pain in the same way at both time periods.

Similarly, pain affects exercise the same way at both time periods. Exercise at Time One may affect pain at Time Two; certainly the conventional wisdom is that it helps[2].

It might clarify thing to think in terms of individual patients. Patient One has a bad case of the disease, so she experiences more pain than average at both Time One and Time Two. Because of the pain at Time One, she exercises less than average at Time One. Patient One is below average in exercise at Time One, and above average in pain at Time Two.

Patient Two has the disease; it's not good, but it could be worse. She hurts less than average at both Time One and Time Two. Because the pain is not too bad at Time One, she exercises more than average. Thus she is above average in exercise at Time One and below average in pain at Time Two.

The curved arrows between error terms are what makes the model unusual. Pain really has momentum. Once it gets going it's harder to block, and it could be that pain at time one is directly contributing to pain at time two. But there are other things that would help produce a positive covariance between true pain at time one and true pain at time two, quite apart from disease severity. One omitted variable is the person's pain sensitivity, and there may be more. This is the reasoning behind the curved arrow connecting ϵ_1 and ϵ_3.

Exercise has momentum for reasons that are even more obvious. There's habit (established before the study began), social obligations to workout partners, New Year's resolutions, and just plain enjoyment of exercise (or the opposite). This explains the curved arrow between ϵ_2 and ϵ_4.

I hope you have been wondering about identifiability. If the curved arrows were replaced by straight arrows from Time One to Time Two, this model would satisfy the Acyclic Rule, with one variable in each set. The curved and straight arrows play the same role in the covariance matrix, and everything is okay. What would kill identifiability would be to have them both, because then they would be redundant.

The latent variable model is given, but more than one measurement model is acceptable. I think my measurement model is the most obvious one, but it's definitely debatable. My model fit, once I fixed the typos. Your job is to do something reasonable and be ready to interpret the output of `summary`, especially the tests for γ, β_1 and β_2 at the $\alpha = 0.05$ level.

Please bring your *complete* R printout from Question 5 to the quiz, showing all input and output. It may be handed in.

[2] I think this is really interesting. If you apply the `corr` function (this will not be on the quiz), you will see negative correlations between the exercise measurements at Time One and the pain measurements at Time Two. This seems to support the conventional wisdom, but not so fast! The counter-argument is this. Because of disease severity, the more pain at Time One, the more pain at Time Two. And because exercise hurts when you have this disease, the more pain at Time One the less exercise at Time One. Therefore, the less exercise at Time One the more pain at Time Two, even if there is no direct link.