STA 431s23 Assignment Nine[1]

For the Quiz on Friday March 31st, please bring a printout of your full R input and output for Question 6. The other problems are not to be handed in. They are practice for the Quiz.

---

1. Let $\mathbf{z} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma})$, where the elements of $\mathbf{z}$ are standardized, so that $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{D}\mathbf{C}^\top$ is a correlation matrix.

   (a) What is the distribution of $\mathbf{y} = \mathbf{C}^\top\mathbf{z}$? (The elements of $\mathbf{y}$ are the *principal components* of $\mathbf{z}$.)

   (b) What is the variance of the scalar random variable y_j, that is, element j of $\mathbf{y}$?

   (c) How do you know that the elements of $\mathbf{y}$ are independent?

   (d) Write $\mathbf{z}$ as a function of $\mathbf{y}$.

   (e) Using the notation $\mathbf{C} = [c_{ij}]$,

      i. Write the scalar random variable z_1 as a function of $y_1, \ldots, y_k$.

      ii. What is $Var(z_1)$?

      iii. What proportion of $Var(z_1)$ is "explained" by y_4?

   (f) Calculate $cov(\mathbf{z}, \mathbf{y})$. Simplify.

   (g) What is element i, j of the matrix $cov(\mathbf{z}, \mathbf{y})$? Give your answer in terms of the c_{ij} and λ_j.

   (h) What is the squared correlation between z_1 and y_4? Compare this to your answer to Question 1(e)iii.

   (i) To answer that last question, you needed to standardize the principal component y_4. The whole vector of principal components can be standardized with $\mathbf{y}_2 = \mathbf{D}^{-\frac{1}{2}}\mathbf{y}$. Verify that the standardization works by calculating $cov(\mathbf{y}_2)$.

   (j) Write $\mathbf{z}$ as a function of $\mathbf{y}_2$.

   (k) Calculate $cov(\mathbf{z}, \mathbf{y}_2)$. Because $\mathbf{z}$ and $\mathbf{y}_2$ are both standardized, this is a matrix of correlations.

   (l) Question 1h tells us that if you were to square the elements in column j of $cov(\mathbf{z}, \mathbf{y}_2)$ and add them up, you'd get the total amount of variance in all the z_i variables that is explained by principal component j. Calculate all these quantities at once with a matrix operation. What you want are the diagonal elements of a certain matrix product.

2. We usually don't retain all k principal components. Instead, we summarize the variables with a smaller set of p principal components that explain a good part of the total variance. Typically, components associated with eigenvalues greater than one are retained. This may be accomplished with a $p \times k$ *selection matrix* that will be denoted by $\mathbf{J}$. Each row of $\mathbf{J}$ has a one in the position of a component to be retained, and the rest zeros. For example, if there were five principal components, the first two would be selected as follows.

$$\mathbf{Jy} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

If $\mathbf{A}$ is any $k \times k$ matrix, then $\mathbf{JAJ}^\top$ is the $p \times p$ sub-matrix with rows and columns indicated by $\mathbf{J}$. A sub-matrix of the identity is another (smaller) identity matrix, so $\mathbf{JJ}^\top = \mathbf{I}_p$. Selection matrices are quite flexible and can even be used to re-order variables, but here they will just be used to select the first p principal components.

So, let $\mathbf{J}$ be the matrix that selects the first p elements of a vector. Let $\mathbf{f} = \mathbf{Jy}_2$. That's the first p standardized principal components.

(a) Show $cov(\mathbf{f}) = \mathbf{I}_p$.

(b) Let $\mathbf{L} = cov(\mathbf{z}, \mathbf{f})$. Calculate $\mathbf{L}$. This is the matrix of correlations between the z variables and the first p principal components. That is, it should be the first p columns of $cov(\mathbf{z}, \mathbf{y}_2) = \mathbf{CD}^{\frac{1}{2}}$. Indeed, post-multiplication by $\mathbf{J}^\top$ selects the first p columns.

(c) Let $\mathbf{f}' = \mathbf{Rf}$, where $\mathbf{R}$ is an orthogonal (rotation) matrix.

i. Calculate $cov(\mathbf{f}')$.

ii. Calculate $cov(\mathbf{z}, \mathbf{f}')$. Leave your answer in terms of $\mathbf{L}$. This is very quick. Why is $cov(\mathbf{z}, \mathbf{f}')$ a matrix of correlations, rather than just covariances?

iii. If you square the correlations in row j of $\mathbf{L}$ and add them up, you get the proportion of variance in z_j that is explained by the first p principal components. Show that this quantity is not affected by rotation of the components. Hint: calculate all the proportions of explained variation at once with a matrix operation. What you want are the diagonal elements of a certain matrix product. If you do the operation on $cov(\mathbf{z}, \mathbf{f})$ and $cov(\mathbf{z}, \mathbf{f}')$, you get the same answer.

3. The following is based on data from one of my classes.

```
> pc3 = prcomp(dat, scale = T, rank=3)
> L = cor(dat,pc3$x) # Correlations of variables with components
> round(L,3)
          PC1    PC2    PC3
Quiz 1  0.546 -0.396  0.377
Quiz 2  0.657 -0.156  0.237
Quiz 3  0.646 -0.287  0.133
Quiz 4  0.614  0.225  0.446
Quiz 5  0.606  0.417 -0.185
Quiz 6  0.511  0.616  0.106
Quiz 7  0.589  0.390 -0.257
Quiz 8  0.539 -0.239 -0.582
Quiz 9  0.730 -0.052 -0.319
Quiz 10 0.318 -0.351 -0.337
Final   0.763 -0.201  0.186

> vm3 = varimax(L); Rt = vm3$rotmat
> L2 = L %*% Rt # Also, L2 = vm3$loadings
> round(L2,3)
         [,1]   [,2]   [,3]
Quiz 1  0.763 -0.050 -0.115
Quiz 2  0.658  0.225 -0.168
Quiz 3  0.641  0.127 -0.303
Quiz 4  0.602  0.481  0.181
Quiz 5  0.138  0.711 -0.228
Quiz 6  0.168  0.779  0.130
Quiz 7  0.093  0.688 -0.289
Quiz 8  0.115  0.193 -0.798
Quiz 9  0.330  0.422 -0.593
Quiz 10 0.156 -0.056 -0.557
Final   0.718  0.257 -0.274
```

You will need a calculator for these questions. Please round your answers to three decimal places.

(a) What proportion of the variance in Quiz 9 score is explained by the first *unrotated* principal component?

(b) What proportion of the variance in Quiz 9 score is explained by the first *rotated* principal component?

(c) What proportion of the variance in Final Exam score is explained by the unrotated principal components?

(d) What proportion of the variance in Final Exam score is explained by the rotated principal components?

4. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} z_{i,1} &= \lambda_{11}F_{i,1} + \lambda_{12}F_{i,2} + e_{i,1} \\ z_{i,2} &= \lambda_{21}F_{i,1} + \lambda_{22}F_{i,2} + e_{i,2} \\ z_{i,3} &= \lambda_{31}F_{i,1} + \lambda_{32}F_{i,2} + e_{i,3} \\ z_{i,4} &= \lambda_{41}F_{i,1} + \lambda_{42}F_{i,2} + e_{i,4} \\ z_{i,5} &= \lambda_{51}F_{i,1} + \lambda_{52}F_{i,2} + e_{i,5}, \end{aligned}$$

where all expected values are zero, $Var(F_{i,1}) = Var(F_{i,2}) = 1$, and all the $F_{i,j}$ and $e_{i,j}$ are independent. As the notation suggests, the $z_{i,j}$ are standardized, so that $Var(z_{i,j}) = 1$ for all i and j. Only the $z_{i,j}$, are observable.

Please give Greek letter answers to the following. Be able to show your work if necessary.

(a) What is $Var(e_{i,2})$?

(b) What is the uniqueness of $z_{i,2}$?

(c) What is the communality of $z_{i,2}$?

(d) What is $Corr(z_{i,3}, F_{i,2})$?

(e) What is the reliability of $z_{i,3}$ as a measurement of $F_{i,2}$?

(f) What is the reliability of $s_i = z_{i,1} + z_{i,2} + z_{i,3} + z_{i,4} + z_{i,5}$ as a measurement of $F_{i,1}$? You can see that this is general.

(g) What proportion of the variance in $z_{i,4}$ is explained by the common factors?

(h) What proportion of the variance in the observable variables is explained by Factor One? You are being asked for a proportion, so the answer is between zero and one.

(i) What is $Cov(z_{i,2}, z_{i,5})$?

(j) What are the parameters of this model?

(k) All the factor loadings are correlations, so you might think that the parameter space is a hyper-cube, running from -1 to $+1$ in eight dimensions. However ..., what inequality constraint must λ_{21} and λ_{22} obey?

(l) What does this look like geometrically, in just the $(\lambda_{21}, \lambda_{22})$ plane?

(m) Does this model pass the test of the Parameter Count Rule? Answer Yes or No and give the numbers. Remember, $\boldsymbol{\Sigma}$ is a correlation matrix, so there are no equations corresponding to the diagonal elements.

5. Consider the general factor analysis model

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

where $\mathbf{\Lambda}$ 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 $\mathbf{\Phi}$, and $\mathbf{e}_i$ is multivariate normal and independent of $\mathbf{F}_i$, with expected value zero and covariance matrix $\mathbf{\Omega}$. 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 $\mathbf{d}_i$.

(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 an orthogonal rotation) is one satisfying $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$.

(e) As in Question 5d, suppose that $\mathbf{\Phi}$ is set to the identity matrix, standardizing the factors as well as making them uncorrelated. In addition, standardize the observable data to obtain $\mathbf{z}_i$. Write

$$\begin{aligned} \mathbf{z}_i &= \mathbf{\Lambda}\mathbf{F}_i + \mathbf{e}_i \\ &= \mathbf{\Lambda}\mathbf{R}^\top\mathbf{R}\mathbf{F}_i + \mathbf{e}_i \\ &= \mathbf{\Lambda}_2\mathbf{F}'_i + \mathbf{e}_i \end{aligned}$$

i. Calculate $cov(\mathbf{z}_i, \mathbf{F}_i)$ and $cov(\mathbf{z}_i, \mathbf{F}'_i)$. These are matrices of correlations.

ii. Show that the communalities of the variables are not affected by rotation. You want the diagonal of a certain matrix product.

6. The `statclass` data include marks on quizzes, computer assignments, a midterm test and the final exam. The column labelled `S` is sex, and the column labelled `E` is ethnic background. We will not use `S` or `E` on this assignment. They are both just guesses by the prof, and are are likely measured with error. We are just going to do an exploratory factor analysis on the other variables. The data are available at

   https://www.utstat.toronto.edu/brunner/openSEM/data/statclass.data.txt.

   Use the `header=TRUE` option on `read.table()`.

   (a) To help decide on the number of factors, calculate the eigenvalues of the correlation matrix and prepare a scree plot. Print hard copy of the scree plot and bring it to the quiz.

   (b) The number of eigenvalues greater than one and the scree plot point to different answers. Going with the smaller number, carry out a maximum likelihood factor analysis with a varimax rotation. Does it fit? (Use the $\alpha = 0.05$ significance level, of course.) Be able to give the chi-squared statistic, the degrees of freedom and the p-value.

   (c) Amazingly, the model fits. Try a smaller number of factors, and keep trying until the model no longer fits. For all the models you estimate, be able to give the chi-squared statistic, the degrees of freedom and the p-value.

   (d) I have a lot of trouble deciding between 3 and 4 factors. I finally decided to go with four, even though Factor 4 seems to be dominated by Computer Assignment 3, whatever that was. This is not really a question. I suppose the correct answer is "Okay."

   (e) For the 4-factor model, obtain communalities in two different ways. One way to get them is to extract the diagonal of a certain matrix product with the `diag()` function. The other way is more obvious. Use `cbind()` to display the two sets of estimates side by side.

   (f) What estimated proportion of the variance of Computer Assignment 3 is explained by the common factors?

   (g) What estimated proportion of the final exam's variance is explained by the common factors?

   (h) What is the estimated correlation between score on the midterm and Factor 1?

   (i) What is the estimated reliability of the final exam as a measure of Factor 1?

Please bring a printout of your full R input and output for Question 6 to the quiz.