ST 557: Applied Multivariate Analysis
Fall 2019

Homework 5

Due: Friday, November 15
110 Points

1. **(45 points)** Data on national track records for men for 55 different countries are given in the file TrackData.csv, available on Canvas. For each country, the national record for eight different distances is recorded. The first column of the data set contains the country name, and the second column contains a three-letter abbreviation. The next eight columns contain the national records for each of the eight distances: the first three distances (100m, 200m, 400m) are recorded in seconds, and the remaining five distances (800m, 1500m, 5000m, 10000m, Marathon) are recorded in minutes.

   (a) Obtain the sample covariance matrix $\mathbf{S}$ and the sample correlation matrix $\mathbf{R}$ for the distances based on this data. Which of these matrices would you find more interesting/appropriate to use for a principal component analysis of this data, and why?

   (b) Determine the eigenvalues and eigenvectors of $\mathbf{S}$.

   (c) Determine the eigenvalues and eigenvectors of $\mathbf{R}$.

   (d) Construct four plots of the loadings for the first four principal components of $\mathbf{S}$. Include the loadings for the corresponding principal component of $\mathbf{R}$ on the same plot, in a different color or plotting character.

   (e) How would you interpret the loadings for the first principal component found using $\mathbf{S}$?

   (f) How would you interpret the loadings for the first principal component found using $\mathbf{R}$?

   (g) What does the second principal component of $\mathbf{R}$ seem to represent?

   (h) Plot the scree plot and cumulative variance explained plot for the principal components of both $\mathbf{S}$ and $\mathbf{R}$.

   (i) How many principal components of $\mathbf{R}$ would you want to keep? Explain.

2. **(15 points)** The weekly rates of return for five stocks listed on the New York Stock Exchange are given in the file NYSEData.csv, available on Canvas. For each of the 103 weeks (two years), the return rates are reported for all five stocks. The first three stocks are for financial companies, and the last two are energy/petroleum stocks.

   (a) Construct the sample covariance matrix $\mathbf{S}$ for the five stocks, and find the sample principal components.

   (b) What proportion of the total sample variance is explained by the first three principal components?

   (c) How would you interpret the first three principal components (that is, interpret the loadings for each of these components on the original variables–what do each of these directions represent)?

3. **(20 points)** Let $\tilde{\mathbf{L}}$ be the $(p \times m)$ matrix of factor loadings $\tilde{l}_{ij}$ estimated for a sample covariance matrix $\mathbf{S}$ using principal components:

$$\tilde{\mathbf{L}} = \left[\hat{\mathbf{e}}_1\sqrt{\hat{\lambda}_1} \ : \ \hat{\mathbf{e}}_2\sqrt{\hat{\lambda}_2} \ : \ \ldots \ : \ \hat{\mathbf{e}}_m\sqrt{\hat{\lambda}_m} \right]$$

where $\hat{e}_j$ is the eigenvector corresponding to the jth largest eigenvalue of $\mathbf{S}$. Let $\tilde{\mathbf{\Psi}}$ be the diagonal matrix of the specific variances

$$\tilde{\mathbf{\Psi}} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \ldots & 0 \\ 0 & \tilde{\psi}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tilde{\psi}_1 \end{bmatrix}$$

where $\tilde{\psi}_j = s_{jj} - \sum_{k=1}^{m} \tilde{l}_{jk}^2 = s_j^2 - \sum_{k=1}^{m} \tilde{l}_{jk}^2$. We will establish the following inequality:

$$\text{Sum of Squared Entries of } (\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\mathbf{\Psi}})) \leq \hat{\lambda}_{m+1}^2 + \hat{\lambda}_{m+2}^2 + \ldots + \hat{\lambda}_p^2$$

(a) Explain why

$$\text{Sum of Squared Entries of } (\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\mathbf{\Psi}})) \leq \text{Sum of Squared Entries of } (\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T)$$

(b) Recall that $\mathbf{S}$ can be written as

$$\mathbf{S} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^T = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j^T.$$

Use this fact to express $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$ in terms of the $\hat{\lambda}_j$ and $\hat{\mathbf{e}}_j$.

(c) The sum of squared entries of a matrix $\mathbf{A}$ is given by

$$\text{Sum of Squared Entries of } \mathbf{A} = \text{tr}(\mathbf{A}\mathbf{A}^T)$$

(convince yourself of this fact using the definition of matrix multiplication). Use this fact and your result from part (b) to express the sum of squared entires of $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$ in terms of the $\hat{\lambda}_j$.

(d) Use your results from parts (a) and (c) to conclude that

$$\text{Sum of Squared Entries of } (\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\mathbf{\Psi}})) \leq \hat{\lambda}_{m+1}^2 + \hat{\lambda}_{m+2}^2 + \ldots + \hat{\lambda}_p^2$$

4. **(30 points)** A study of adults aged 65 years and older was conducted on a sample of generally healthy adults randomly selected from Medicare rolls. A number of physiological variables were measured for each subject:

- weight = Subject weight.
- height = Subject height.
- physact = Physical activity of the subject for the week prior to the study.
- ldl = A laboratory measure of a certain kind of cholesterol in the subject's blood. LDL (low density lipoprotein) is often referred to as ?bad cholesterol".
- alb = A laboratory measure of a certain kind of protein in the subject's blood. Albumin is made by the liver, and persons with poor liver function or poor nutritional status will have low levels of albumin.

- crt = A laboratory measure of creatinine in the subject's blood. Creatinine is a waste product of muscles that is excreted by the kidneys. High levels of creatinine are taken as indication of kidney disease.
- plt = A laboratory measure of the number of platelets circulating in the subject's blood. Low platelet levels are often an indication of chronic disease or infections.
- sbp = A measurement of the subject's systolic blood pressure.
- aai = The ratio of systolic blood pressure measured in the subject's ankle to the systolic blood pressure measured in the subject's arm. A low ankle to arm index suggests peripheral arterial disease.
- fev = A measure of forced expiratory volume in the subject. Normal FEV measurements depend upon the condition and size of the lungs (note that lung size is usually proportional to body size).
- dsst = A measure of cognitive function (ability to think) for the subject. High scores indicate better cognitive performance.
- atrophy = A measure of global brain atrophy detected by MRI. High values indicate severe atrophy. The correlation matrix for these variables is given in the file ?PhysioData.csv?, available on Canvas.

(a) Perform a principal components factor analysis based on the given correlation matrix, for m = 2 and m = 3 factors. Describe how you might interpret the resulting factors for each model: can you describe the underlying latent variables for these two models? Which variables contribute most to each factor?

(b) What is the residual matrix for the principal components factor analysis model with m = 2 factors? With m = 3 factors?

(c) Perform a maximum likelihood factor analysis based on the given correlation matrix, for m = 2 and m = 3 factors. [Note that the ?factanal()? function in R can operate on just a covariance or correlation matrix instead of a full dataset: you just have to instead provide the argument ?covmat = ?.] Describe how you might interpret the resulting factors for each model: can you describe the underlying latent variables for these two models? Which variables contribute most to each factor?

(d) What is the residual matrix for the maximum likelihood factor analysis model with m = 2 factors? With m = 3 factors?

(e) Which method (principal components or maximum likelihood) do you prefer for this data? Explain your choice.

(f) Are the factors resulting from the two methods similar for the m = 2 models? Are the factors from the two methods similar for the m = 3 models?