Stat ST465/665, Project 5

1. **(42 points)** The project is to conduct a principal components analysis of the paper mill response data (paper_mill_data_response.txt, *Aldrin, M. , "Moderate projection pursuit regression for multivariate response data", Computational Statistics and Data Analysis, 21 (1996), p. 501-531*).

   (a) Label the variables $r_1, \cdots, r_{13}$. Carry out an initial investigation. Do not remove outliers or transform the data. Indicate if you had to process the data file in anyway. Explain any conclusions drawn from the evidence and backup your conclusions.

   (b) Display the sample correlation matrix $\boldsymbol{R}$.

   (c) Perform a principal component analysis using $\boldsymbol{R}$.

      i. List the eigenvalues and describe the percent contributions to the variance.

      ii. Determine the number of principal components to retain and justify your answer by considering at least three methods. Note and comment if there is any disagreement between the methods.

      iii. Give the eigenvectors for the first two principal components and write out the principal components.

      iv. Considering the coefficients of the principal components, describe dependencies of the principal components on the variables.

      v. Display a scatter plot of the first two principal components. Make observations about the plots.

   (d) Include your code.

2. **(85 points)** The project is to conduct a principal components analysis of the flea beatle data (fleabeetledata.xlsx, *Lubischew, A., On the use of discriminant functions in taxonomy, Biometrics 18 (1962), 455-477.*). The data has two groups. You will conduct three principal component analysis, one for each individual group and one for the entire data set ignoring groups. You will use $\boldsymbol{S}$ for the PCA.

   (a) Carry out an initial investigation. Do not remove outliers or transform the data. Indicate if you had to process the data file in anyway. Explain any conclusions drawn from the evidence and backup your conclusions. *Hint: Pay attention to potential differences between the groups.*

   (b) For the Haltica oleracea group,

      i. Display the relevant sample covariance matrix $\boldsymbol{S}$.

      ii. List the eigenvalues and describe the percent contributions to the variance.

      iii. Determine the number of principal components to retain and justify your answer by considering at least three methods.

      iv. Give the eigenvectors for the principal components you retain.

      v. Considering the coefficients of the principal components, Describe dependencies of the principal components on the variables.

      vi. Using at least the first two principal components, display scatter plots of pairs of principal components. Make observations about the plots.

   (c) For the Haltica carduorum group,

      i. Display the relevant sample covariance matrix $\boldsymbol{S}$.

      ii. List the eigenvalues and describe the percent contributions to the variance.

      iii. Determine the number of principal components to retain and justify your answer by considering at least three methods.

      iv. Give the eigenvectors for the principal components you retain.

      v. Considering the coefficients of the principal components, Describe dependencies of the principal components on the variables.

      vi. Using at least the first two principal components, display scatter plots of pairs of principal components. Make observations about the plots.

   (d) For the entire data set (ignoring groups),

      i. Display the relevant sample covariance matrix $\boldsymbol{S}$.

      ii. List the eigenvalues and describe the percent contributions to the variance.

      iii. Determine the number of principal components to retain and justify your answer by considering at least three methods.

      iv. Give the eigenvectors for the principal components you retain.

      v. Considering the coefficients of the principal components, Describe dependencies of the principal components on the variables.

      vi. Using at least the first two principal components, display scatter plots of pairs of principal components. Make observations about the plots.

   (e) Compare the results for the three principal component analyses. Do you have any conclusions?

   (f) Show your code.

   Key for Flea Beetle Data

   x1 = distance of transverse groove from posteriori border of prothorax

   x2 = length of elytra

   x3 = length of second antennal joint

   x4 = length of third antennal joint

3. **(72 points)** The project is to conduct a principal components analysis of the Mali Farm data (malifarmdata.xlsx, *R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis, Pearson, New Jersey, 2019.*). You will use $\boldsymbol{S}$ for the PCA.

   (a) Store the data in matrix $\boldsymbol{X}$.

   (b) Carry out an initial investigation. Indicate if you had to process the data file in anyway. Do not transform the data. Explain any conclusions drawn from the evidence and backup your conclusions. *Hint: Pay attention to detection of outliers.*

      i. The data in rows 25, 34, 52, 57, 62, 69, 72 are outliers. Provide at least two indicators for each of these data that justify this claim.

      ii. Explain any other conclusions drawn from initial investigation.

   (c) Create a data matrix $\widetilde{\boldsymbol{X}}$ by removing the outliers.

   (d) Carry out principal component analyses on $\boldsymbol{X}$.

      i. Give the relevant sample covariance matrix $\boldsymbol{S}$.

      ii. List the eigenvalues and describe the percent contributions to the variance.

      iii. Determine the number of principal components to retain and justify your answer by considering at least three methods.

      iv. Give the eigenvectors for the principal components you retain.

      v. Considering the coefficients of the principal components, Describe dependencies of the principal components on the variables.

      vi. Using at least the first two principal components, display appropriate scatter plots of pairs of principal components. Make observations about the plots.

   (e) Carry out principal component analyses on $\widetilde{\boldsymbol{X}}$.

      i. Give the relevant sample covariance matrix $\boldsymbol{S}$.

      ii. List the eigenvalues and describe the percent contributions to the variance.

      iii. Determine the number of principal components to retain and justify your answer by considering at least three methods.

      iv. Give the eigenvectors for the principal components you retain.

      v. Considering the coefficients of the principal components, Describe dependencies of the principal components on the variables.

      vi. Using at least the first two principal components, display appropriate scatter plots of pairs of principal components. Make observations about the plots.

   (f) Compare the results for the two analyses. How much effect did the outliers have on the principal component analysis? Which result do you like more and why?

   (g) Include your code.

   Key for Mali farm data

   Family = number of people in the household

   DistRD = distance in kilometers to the nearest passable road

   Cotton = hectares of cotton planted in 2000

   Maize = hectares of maize planted in 2000

   Sorg = hectares of sorghum planted in 2000

   Millet = hectares of millet planted in 2000

   Bull = total number of bullocks

   Cattle = total number of cattle

   Goat = total number of goats