Stat ST465/665, Assignment 4

Problems

1. **(14 points)** Read in data matrix "assignment4_data1.txt" to create a data matrix $\boldsymbol{X}$. The assignment is to use the matrix scatter plot, a plot of the statistical distances to the sample mean, and the univariate q-q plots to detect outliers in the data set. *Hint: There are 3 or less outliers in the data.*

   (a) Compute and display the sample covariance matrix and mean vector $\boldsymbol{S}$ and $\underset{\sim}{\bar{x}}$.

   (b) Show a matrix scatter plot and univariate q-q plots.

   (c) Compute the statistical distance $\underset{\sim}{D}$ vector between the data points and the sample means where $\underset{\sim}{D}_i = (\underset{\sim}{x}_i^\top - \underset{\sim}{\mu})^\top \boldsymbol{S}^{-1}(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}})$ with $\underset{\sim}{x}_i^\top$ denoting the ith row vector of $\boldsymbol{X}$. Show a plot of the values versus index.

   (d) Use the graphs and $\underset{\sim}{D}$ to identify outliers. Explain your choices. Each outlier should have at least two indicators.

   (e) Remove the outliers to get a new data set, then compute and display sample covariance matrix and mean vector $\boldsymbol{S}$ and $\underset{\sim}{\bar{x}}$ for the cleaned data set. Describe the effect of removing outliers on the sample covariance and mean.

   (f) Show a matrix scatter plot and univariate q-q plots. Is this evidence consistent with a normal distribution? Explain.

2. **(14 points)** Read in data matrix "assignment4_data2.txt" to create a data matrix $\boldsymbol{X}$. The assignment is to use univariate Box-Cox transformations to try to improve the degree to which the components of the data fit a normal distribution.

   (a) Compute and display the sample covariance matrix and mean vector $\boldsymbol{S}$ and $\underset{\sim}{\bar{x}}$.

   (b) Show a matrix scatter plot and univariate q-q plots. Use the plots to discuss the degree to which the data are consistent with a normal distribution.

   (c) Define a new data set $\underset{\sim}{Y}$ by performing a Box-Cox transformation on each column of $\boldsymbol{X}$. List the parameter λ used in the Box-Cox transformations.

   (d) Show a matrix scatter plot and univariate q-q plots. Use these graphs to explain if the Box-Cox transformations improved the degree to which the data are consistent with a normal distribution.

   (e) Describe the effect of transforming the data on the sample covariance and mean.

3. **(9 points)** Read in data matrix "assignment4_data3.txt" to create a data matrix $\boldsymbol{X}$. The assignment is to evaluate if the data are consistent with a normal distribution both before and after using univariate Box-Cox transformations.

   (a) Show a matrix scatter plot and univariate q-q plots. Use the plots to discuss the degree to which the data are consistent with a normal distribution.

   (b) Define a new data set $\underset{\sim}{Y}$ by performing a Box-Cox transformation on each column of $\boldsymbol{X}$. List the parameter λ used in the Box-Cox transformations.

   (c) Show a matrix scatter plot and univariate q-q plots. Use these graphs to explain if the Box-Cox transformations improved the degree to which the data are consistent with a normal distribution.

   (d) Considering all the indicators, are the transformed data consistent with a normal distribution?