ST 557: Applied Multivariate Analysis
Fall 2019

Homework 6

Due: Wednesday, December 4
120 Points

1. **(15 points)** Data from a study of a comparison of non-diabetic and diabetic patients was obtained for three primary variables

$$\begin{aligned} X_1^{(1)} &= \text{ glucose intolerance} \\ X_2^{(1)} &= \text{ insulin response to oral glucose} \\ X_3^{(1)} &= \text{ insulin resistance} \end{aligned}$$

and two secondary variables

$$\begin{aligned} X_1^{(2)} &= \text{ relative weight} \\ X_2^{(2)} &= \text{ fasting plasma glucose} \end{aligned}$$

The data for $n = 46$ non-diabetic patients yields the **covariance matrix:**

$$\mathbf{S} = \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{21} & \mathbf{S}_{22} \end{array}\right] = \left[\begin{array}{ccc|cc} 1106.00 & 396.70 & 108.40 & 0.79 & 26.23 \\ 396.70 & 2382.00 & 1143.00 & -0.21 & -23.96 \\ 108.40 & 1143.00 & 2136.00 & 2.19 & -20.84 \\ \hline 0.79 & -0.21 & 2.19 & 0.02 & 0.22 \\ 26.23 & -23.96 & -20.84 & 0.22 & 70.56 \end{array}\right]$$

Determine the sample canonical variates and their correlations. Try to interpret these quantities.

2. **(15 points)** Crude oil samples were obtained from two zones of sandstone:

$$\begin{aligned} &\pi_1 : \text{Sub-Mulinia} \\ &\pi_2 : \text{Upper} \end{aligned}$$

The samples were analyzed for quantities of three trace elements and two measures of hydrocarbons:

$$\begin{aligned} X_1 &= \text{ vanadium} \\ X_2 &= \text{ iron} \\ X_3 &= \text{ beryllium} \\ X_4 &= \text{ saturated hydrocarbons} \\ X_5 &= \text{ aromatic hydrocarbons} \end{aligned}$$

Data for these variables for samples from each population are given in the file CrudeOilData.csv, available on Canvas. The first column of the file indicates the population from which the sample was taken.

(a) Obtain the Fisher's Discriminant Function rule for this data: what are the values for a new observation $\mathbf{x}_0$ for which the observation would be classified as coming from π_1?

(b) Construct the confusion matrix for the given data, comparing true population membership to the predicted classification based on Fisher's Discriminant Function. What is the apparent error rate (APER) for this classifier, based on the given data?

(c) If a new observation has $\mathbf{x} = [4.0, 17.0, 0.50, 5.54, 3.51]^T$, to which population would Fisher's Discriminant Function assign this observation?

3. **(15 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) Calculate the Euclidean distances between all of the pairs of countries. Using these distances as a measure of dissimilarity, cluster the countries using single linkage and complete linkage hierarchical clustering procedures. Plot the dendrograms and compare the results. Which linkage produces a 'better' clustering of this data, in your opinion? Explain your answer.

   (b) Perform a k-means clustering for the countries for $k = 2, 3$, and 4. Produce tables/lists indicating which countries are grouped together for each k-means clustering.

   (c) Compare the k-means clustering results from part (b) to the hierarchical clustering results from part (a). Do you prefer k-means clustering or hierarchical clustering for this data? Explain your answer.

4. **(15 points)** Nine different archaeological sites were studied, and the frequencies of different types of potsherds found at each site were recorded. Based on these frequencies, "distances" were computed between the sites. The computed distances are given in the file ArchaeoData.csv, available on Canvas. Given these distances, determine the coordinates of the sites in $q = 2$ and $q = 3$ dimensions using multidimensional scaling. Plot the coordinates for the $q = 2$ solution.

5. **(60 points)** For each of the following scenarios, indicate which of the following test procedures you would use to answer the specified questions:

   a. Univariate one-sample t-test
   b. Univariate one-sample t-test (paired differences)
   c. Univariate two-sample t-test
   d. Hotelling's one-sample T^2 test
   e. Hotelling's one-sample T^2 test (paired differences)
   f. Hotelling's one-sample T^2 test (repeated measures)
   g. Hotelling's two-sample T^2 test
   h. Bonferroni simultaneous tests
   i. ANOVA
   j. MANOVA
   k. Multivariate multiple regression
   l. Principal Components Analysis

m. Factor Analysis

n. Canonical Correlation Analysis

o. Discriminant Function Analysis/Linear Discriminant Analysis

p. Clustering

q. Multidimensional Scaling

(a) Twenty subjects were given each of three diets (in random order) and the subjects' blood pressures were measured at the end of each diet, so there were three blood pressure measurement associated with each subject.

Question: Did the different treatments affect the subjects' blood pressure differently?

(b) Two varieties of chickweed are difficult to distinguish. Measurements on four variables were obtained for chickweed plants whose variety was known.

Question: Use these observations to establish a rule for classifying a new candidate plant into one of the two varieties.

(c) Each of 50 eight-year-old girls and 50 eight-year-old boys were given a total of 10 tests. Five of these tests had to do with language and five had to do with mathematical reasoning.

Question: Do scores differ between boys and girls?

Question: Combining the boys and girls, what combination of the language tests is most associated with some combination of the math tests?

(d) Daily measurements of seven pollution-related variables were recorded over an extended period of time at a single location in Los Angeles.

Question: Find a low-dimensional representation for these variables that captures most of the variability.

Question: Test whether the pollution on weekends differed from that on weekdays.

(e) For each of a sample of 42 new microwaves made by a certain manufacturer, the amount of radiation emitted when the door of the microwave is closed and the amount of radiation emitted when the door of the microwave is opened are measured.

Question: Construct a confidence interval for the difference in amount of radiation emitted under these two conditions.

(f) A sample of 50 married couples was obtained. The wife and the husband each answered four questions regarding their relationship on a scale of 0 to 10.

Question: Do the wife's answers tend to be similar to the husband's answers, and in what way are they most similar? That is, what combination of the wife's answers is most similar to what combination of the husband's answers?

(g) The standardized scores for each of the ten events in the decathlon were obtained for each of 50 entrants.

Question: Can the variation in the scores be explained by three underlying athletic abilities, and how might these abilities be described?

(h) For 15 different species of predator fish, data were gathered on several aspects of their diet.

Question: How can these species of fish be grouped based on similarities in their diet?

(i) Calcite content was measured at 25 equally-spaced locations along the leg bone for each of seven Tyrannosaurus Rex skeletons and also for each of five skeletons of a newly-discovered type of dinosaur.

Question: Do the calcite concentrations at these locations differ between the two dinosaur species?

Question: Combining the dinosaur species, is calcite concentration the same at all of the measured locations in the leg bone?

Question: Based on these measurements, construct a rule for classifying a new bone as coming from a Tyrannosaurus Rex or from the newly-discovered species.

(j) Blood samples from 40 patients were obtained and each divided into six subsamples, which were sent to six different laboratories to have iron content measured.

Question: Do the six different laboratory results have the same means?

(k) Measurements on six accounting and financial variables were obtained from a sample of insurance companies that were distressed (close to bankrupt) and an independent sample of insurance companies that were solvent.

Question: Establish a rule for classifying future insurance companies as solvent or distressed based on these variables.

(l) DNA analysis was performed on hair specimens from each of 100 mummies taken from Egyptian pyramids. For each mummy, twenty variables concerning the DNA sequence were measured.

Question: Based on the measured variables, identify groups of mummies that are related to each other (have similar values of the variables).

Question: Based on the distances between these variables, construct a two-dimensional plot of the mummies to visualize the groupings.

(m) SAT subject test scores are obtained for a random sample of 100 12th graders who took Math, Biology, Literature, and World History subject tests.

Question: Test whether the average score for all four tests is 500.

Question: Test whether the average scores are equal for all four tests.

(n) A wildlife ecologist measured tail length and wing length for a sample of 45 female hook-billed kites and 45 male hook-billed kites.

Question: Are average tail length and wing length the same for female and male hook-billed kites?

(o) Several measurements were obtained on chief executive officers (CEO) of companies, regarding the degree to which the officers took risks. Several additional measurements were available on the success of the company under their leadership.

Question: What aspects of risk-taking propensity of the CEO are associated with which aspects of company success?

Question: What combination of risk-taking propensities displays the greatest variation between CEOs?

(p) The age, diameter, and height were measured for a sample of trees that contained eagle roost sites and for an independent sample of trees that did not contain eagle roost sites.

Question: Construct confidence intervals for the difference in age, difference in diameter, and difference in height between roosting trees and non-roosting trees.

Question: Determine a rule for classifying a new tree as a likely roosting site or unlikely roosting site, based on these three variables.

(q) For all of the NBA rookies who started in 2000, data were collected on their free-throw percentages each year for the first five years of their NBA careers.

Question: Does average free-throw percentage change over these five years?

(r) Twelve measurements were taken on fossilized skull measurements from 20 kinds of squirrels. The goal of the analysis was to order the 20 squirrels chronologically, on the basis of the similarities between the skull measurements for different squirrels.

Question: Find a one-dimensional representation of the 20 squirrels that best captures the differences between the measured variables.

(s) The protein, fat content, calories, and Vitamin A content were measured for each of ten brands of hot dogs.

Question: Group the brands of hot dogs based on their nutritional content.

Question: What combination of these nutritional measurements captures the greatest difference between the hot dog brands?

(t) Measurements were obtained on five pre-college predictor variables and four college performance variables for each of several hundred students.

Question: What combination of pre-college variables is most associated with a combination of college performance?

Question: Combining the two variable sets, are there a few underlying abilities that explain the pre-college and college performance?