STA 9705/70500: Multivariate Statistical Methods

Take-home Final Exam

Directions: You are not allowed to discuss this exam with anyone but your instructor. When answering a question, first write down your **complete** answer, and then show **relevant** `SAS` output (if any). You cannot only show `SAS` output without any illustration. Clearly label all the answers. Attach your full version of `SAS` code to the end. **Any submission not in such format will not be accepted**. The data sets are available on Blackboard. Good luck!

1. The data in the file `iris.txt` contain observations on X_1 = Sepal length, X_2 = Sepal width, X_3 = Petal length and X_4 = Petal width for samples from three species of iris (1 = Iris setosa; 2 = Iris versicolor; 3 = Iris virginica).

   (a) Is there any difference between the three species of iris in sepal and petal? Conduct an appropriate and complete test using $\alpha = 0.05$. List all **assumptions** for the chosen test and explain every **notation** you may use.

   (b) Conduct a discriminant analysis and write down the resulting discriminant functions. Plot the first discriminant function against the second one, and discuss the separation.

   (c) Do we need both discriminant functions for the separation? Conduct test(s) to explain.

   (d) Conduct an appropriate and complete **contrast test** to find out **specific** differences among the species, based on what you find in (b).

   (e) Which variable does contribute most to separating the species? Explain.

   (f) Among linear, quadratic and k nearest neighbor classification methods, which is the best method for the iris data based on their error rates? Show all necessary work.

   (g) Use the linear classification method to assign a new observation $\mathbf{x}_0' = (5.1, 3.5, 1.75, 0.3)$ into an appropriate species. Show all necessary work.

2. The weekly rates of return for five stocks: *JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell* and *Exxon Mobil*, listed on the New York Stock Exchange are given in file `stock_price.txt`.

   (a) Should we use covariance matrix or correlation matrix if we want to apply principal component analysis (PCA) to this data set? Explain.

   (b) Following your decision in part (a), show the **coefficients** of all principal components and the **variance** explained by each component.

   (c) Show principal components for the first five observations (Use SAS output only).

   (d) How many principal components should we keep? Explain.

   (e) What are the interpretations of the components retained in part (d)?

   (f) Is there any outlier in this data set? Explain.

3. A firm is attempting to evaluate the quality of its sales staff and is trying to find an examination or series of tests that may reveal the potential for good performance in sales. The firm has selected a random sample of 50 sales people and has evaluated each on 3 measures of performance: X_1 = growth of sales, X_2 = profitability of sales, and X_3 = new-account sales. Each of the 50 individuals took each of 4 tests, which purported to measure X_4 = creativity, X_5 = mechanical reasoning, X_6 = abstract reasoning, and X_7 = mathematical ability, respectively. The data are given in file `salesman.txt`.

   (a) Fit a factor model with 3 factors using **iterated principal factor** method. Show **initial** estimates of communalities, and **final** estimates of loadings, communalities, specific variances, and the proportion of variance explained by each factor.

   (b) Did you find any 'unusual' estimate in (a)? If so, explain why it happened.

   (c) Do an orthogonal rotation to the factors obtained in (a). Show the resulting loadings, communalities, specific variances, and the proportion of variance explained by each factor. Does any of those estimates not change after rotation? Explain.

   (d) On which factor does *profitability of sales* depend most? Why? Use rotated loadings.

   (e) Show the complexity of each variable using threshold 0.6, and interpret rotated factors.

   (f) Is the model fit in (a) sufficient for doing factor analysis to the data? If yes, explain why. If no, try to improve it.

4. The file `university.txt` gives the data on some universities for certain variables used to compare or rank major universities. These variables include X_1 = average SAT score of new freshmen, X_2 = percentage of new freshmen in top 10% of high school class, X_3 = percentage of applicants accepted, X_4 = student-faculty ratio, X_5 = estimated annual expenses and X_6 = graduation rate (%). Because SAT and Expenses are on a much different scale from that of the other variables, you need to **standardize** the data first.

   (a) Use hierarchical clustering with **average** linkage and answer the following questions.

      (i) Show the resulting dendrogram and cluster history.

      (ii) Decide the value of g (number of clusters) using both methods given in the lecture notes, and show the **resulting clusters**. Note that the two methods may give different results. (Let $k = 1.25$.)

   (b) Cluster the universities with K-means method. Use the first 3 observations that are at least $r = 1.5$ apart as seeds. Show the **seeds** and **resulting clusters**, and plot the first two **discriminant functions** against each other, labeled by the university's name. Do we have a reasonable clustering here? Explain.

   (c) Redo (b) using the 3 centroids given by the hierarchical clustering in (a) as seeds. Show the **seeds** and **resulting clusters**, and compare them to those given in (b). Which seeds do give us a better clustering, the centroids in (a) or those used in (b)? Explain.