Stat 501: Final Exam, Due 1:45 pm, May 6, 2021

The total number of points on this part of the exam is 100. However, scores will be truncated at 90. There is an additional 10 points in the quiz section which has to be filled in online and separately. Overall, therefore, *i.e.* there is a general 10% bonus in this exam.

General Instructions

This exam is open-book, open-notes. You are not allowed to communicate with anybody **other than the instructor** in any way, shape or form. This includes human-controlled bots.

You ***are required*** to provide explanations and also summarize the results into human-digestible form. Just providing R or SAS code or its output is *not acceptable as a solution to any of the problems.* Indeed, providing undigested code is not needed.

You ***are not required*** to turn in code, but if you do, it should be *annotated code* that works (or identify clearly if it doesn't).

For each question, I have indicated the maximum amount of space that I think is expected. You are not required to fill up the pages, and you will not be penalized for going over.

Please make sure that the final answer to each sub-part is clearly indicated. Finally, the file should be turned in as a document in the Portable Document Format (PDF).

The answers to Question 1 may be handwritten and scanned or photographed using your portable device. Please make sure that the scanning is legible enough for grading, and is turned in as a pdf or a png. Please also make sure that the document is loaded such that it is correct side up.

To account for potential technical issues, there will be a final 10-minute grace period.

Please note that only one submission is allowed.

All submissions have to be on Canvas: no submissions will be accepted via e-mail. This includes people who do not turn in the exam on time. If you e-mail me your exam, you will get a zero.

During the exam, the instructor will be checking e-mail when online (with, no major gaps). Where appropriate, the instructor may, occasionally, provide answers to questions asked during the exam in the place where the exam is posted. Please check the link occasionally. While I do not anticipate major announcements during the exam, if necessary, such will also be sent on e-mail.

To communicate with the instructor, please only use e-mail. Recall that Canvas provides e-mails to the instructor in a digest form at the end of the day, so do not necessarily expect a timely answer if you write to him on Canvas.

Question	Maximum Points	Points Awarded
Question 1	15	
Question 2	85	
Total	100	

Question 1

1. We consider the problem of classifying a multinomial observation vector $\boldsymbol{X}$ into one of two classes $Mult(n, \boldsymbol{p}_1)$ or $Mult(n, \boldsymbol{p}_2)$ with prior probabilities π and $(1-\pi)$, and equal misclassification costs. Show that the discriminant rule for this problem is linear: *i.e.* it can be reduced to classification based on whether $\boldsymbol{a}'\boldsymbol{x} + c$ is positive or negative. [*15 points*]

Question 2

The dataset for this problem is available in the file $wine.dat$ on Canvas. The data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The constituent attributes are

```
 1) Alcohol        2) Malic acid       3) Ash   4) Alkalinity of ash  5) Magnesium
 6) Total phenols   7) Flavanoids  8) Nonflavanoid phenols  9) Proanthocyanins
10)Color intensity    11)Hue          12)OD280/OD315 of diluted wines 13) Proline
```

Note that the first attribute in the file is the wine cultivar, and that the column delimiter is ",".

1. Is a multivariate normality assumption reasonable for the distribution of the attributes for each cultivar? [*10 points; 1 page*]
2. Please perform a detailed factor analysis for wines in the third cultivar (*i.e.* cultivar given by 3). Use BIC to determine the number of factors. [*15 points; 2 pages*]
3. Use a multivariate analysis of variance to investigate whether cultivars have an effect on the average hue, ash content, and color intensity of the wine. Set-up the model, and summarize the analysis and the results. In doing so, evaluate the necessary assumptions in your model. [*15 points; 1 page*]
4. Using `set.seed()` with seed given by the last four digits of your university ID, split the *original* dataset into a random training set of 128 observations and a test set of the remaining 50 observations. Call them `wine.train` and `wine.test`. Test classification rules obtained from the training set and tested on the test set using the following:
   (a) Quadratic discriminant analysis
   (b) k-Nearest-neighbor classification with a cross-validated choice of k. (Note that you may need to account for the randomness of tie-breakers in determining this.)
   (c) The best cross-validated classification tree

   Provide detailed summaries on each method, and on the performance on the test set. [*15 points; 4 pages*]
5. Ignoring the cultivar information, group the constituent attributes using
   (a) Hierarchical clustering with average linkage, and $K = 3$ groups.
   (b) K-means clustering with appropriate initialization and with $K = 3$.
   (c) Model-based clustering, with BIC to determine the number of groups and the variance-covariance structure.

   In each case, display and comment on the results and also compare the results with the available cultivar information by means of a cross-table of frequencies of classification. [15 points; *4 pages*]
6. Perform an **appropriate** principal components analysis to reduce the dimensionality of the variables in the dataset. How many principal components are enough to account for at least 80% of the variation? Provide an interpretation for the first few principal components, if possible. [*10 points; 2 pages*]
7. Carrying on with the "first few" principal components analysis, does reducing dimensionality of the dataset using principal components in this way improve the distinctiveness of the cultivars as explained by these new projections of the dataset? Answer this question descriptively (visually). [*5 points; 2 pages*]