FINAL (TAKE HOME) STA 135 DUE: MARCH 17TH, BETWEEN 4:10-4:25 PM

Instructions.
1. Please show sufficient intermediate steps in your work and solutions should be well organized. Do not use a pen that bleeds through the paper. Attach any R code in a well-organized appendix.
2. The exam is open-book (i.e., textbook Johnson & Wichern 6E and class materials), but no collaboration with others is permitted. Strict adherence to the code of academic conduct is expected and will be enforced.
3. If clarification on a problem is needed, you may consult the Professor on Monday 3/13 or the TA on Tuesday 03/14. In order to be fair to all students, you may not ask for direct or indirect assistance with examination questions.
4. Turn in a hard copy of the exam including this cover page to Christopher Conley between 4:10-4:25 pm in Math Science Building, room 1117 on March 17th. In order to be fair to all students, late and/or digital submissions are not permitted. Please plan accordingly so that you can complete the exam on time.
5. Do not distribute this exam in any form.
6. Total points: 130

Name:_____

1. (16 points) Consider an orthogonal factor model with arbitrary number of factors $m \leq p$.

   (a) Prove that the estimated covariance matrix is unchanged by a suitable rotation of the loadings matrix.

   (b) Consider the high dimensional setting of $p >> n$. Describe the structure of $\tilde{\mathbf{l}}_{\mathbf{n+1}}, \ldots, \tilde{\mathbf{l}}_{\mathbf{p}}$, where $\tilde{\mathbf{l}}_{\mathbf{j}}$ is the jth column vector of the estimated loading matrix under the principal component solution.

2. (30 points) A financial advising group is interested to see if there is any correlation in weekly stock returns between the major firms in two different sectors, biotechnology and energy. Acting as a consultant, the group hired you to perform a canonical correlation analysis to explore the potential correlation structure between three top biotech firms and two top energy firms sampled across 103 weeks ($n = 103$) on the New York Stock Exchange. Data is provided in "two-markets.DAT" on canvas under the "Sample Exams" folder and has already been standardized to have mean 0 and variance 1 for each firm. Your analysis should address the following steps:

   (a) Justify how many canonical correlations are significant through a hypothesis testing framework at $\alpha = 0.05$.

   (b) Report the first pair of canonical variables as mathematical expressions (i.e., do not just copy and paste R output). Comment on any notable structure of the coefficients of the canonical variates.

   (c) Report the sample correlations: $R_{U_1,Z^{(2)}}$, $R_{V_1,Z^{(1)}}$. Compare these additional correlations with $\hat{\rho}_1^*$ to make an argument about how much the top biotech and top energy firms correlate in stock returns.

Cont.

3. (35 points) An talent acquisition firm (i.e., job recruiter) is trying to classify job applicants as either short-tenure (population 1) or long-tenure employees (population 2) before placing them with employers. Depending on the role the firm is seeking to fill, they have estimated the cost of placing a short-tenure employee in a long-tenure position is double that of placing a long-tenure employee in a short-tenure position. Past experience suggests short-tenure employees are 3 times more prevalent than long-tenure employees. The firm has historical data on applicants' tenure (measured in months) at their past two positions and have population labels as short- or long-tenure after having placing them with an employer, who was the recruiter's client. The firm assumes the tenure length for the past two positions of applicants is multivariate normal with common covariance for both short- and long-tenure employees. Sample statistics are: $n_1 = 400, n_2 = 133$ and

$$\overline{x}_1 = \begin{bmatrix} 12 \\ 18 \end{bmatrix} \qquad \overline{x}_2 = \begin{bmatrix} 36 \\ 24 \end{bmatrix} \qquad S_{\text{pooled}} = \begin{bmatrix} 2.4 & 1.7 \\ 1.7 & 3.4 \end{bmatrix}$$

   (a) First test if the mean vectors of the applicants' past two positions differs between the short- and long-tenure labeled employees at $\alpha = 0.05$ to justify pursuing a minimum ECM rule?

   (b) Construct a minimum ECM rule for short- and long-tenure employee populations.

Cont.

(c) Based on your rule from part(b), assign a new applicant with tenure history $\mathbf{x_0} = (18, 24)^T$ to either short- or long-tenure population.

(d) Describe briefly how your analysis would change if the tenure length for the past two positions of applicants has a multivariate T distribution.

(e) Describe briefly the advantages/disadvantages of using logistic regression instead of discriminant analysis.

Cont.

4. (24 points)

(a) PCA can be sensitive to outliers. Describe a concise strategy to identify outliers in a PCA analysis (illustrations allowed).

(b) Factor Analysis can be rather subjective, describe a concise strategy to validate an orthogonal factor model when the data set is large.

(c) For discriminant analysis, prove formally that allocating an observation $\mathbf{x_0}$ as π_1 when $P(\pi_1|\mathbf{x_0})>$ $P(\pi_2|\mathbf{x_0})$ (i.e., under posterior probability) is equivalent to the minimum ECM rule when misclassification costs are equal. State the theorem formally and prove the result rigorously.

5. (25 points) Researchers were interested in whether viewing time (early evening vs. late evening) influenced netflix ratings within 4 major genres: drama, horror, action, comedy. The variables of interest are ratings by genre. 61 volunteers were recruited to watch 2 films from each genre, where the films within genre were selected for having historically similar ratings after averaging over early and late viewing times so as to not be biased. Each participant was randomly assigned either an early or late viewing time for each film within a genre. The response variable is the difference in early minus late viewing time ratings within genre. Assume other variables such as day of week are controlled for and that the ratings are approximately distributed multivariate normal. The sample statistics are:

mean for differences in ratings by genre $= \begin{bmatrix} 0.6 \\ -8.4 \\ -0.1 \\ -10.7 \end{bmatrix}$ and covariance matrix $= \begin{bmatrix} 5 & 0 & 1 & 0 \\ 0 & 2 & -1 & 1 \\ 1 & -1 & 4 & 1 \\ 0 & 1 & 1 & 4 \end{bmatrix}$

(a) State the null and alternative hypothesis at $\alpha = 0.05$

(b) Evaluate whether there is a difference in ratings due to viewing time through the appropriate statistical test. R can be used for computation, but clear mathematical expressions should be presented.

(c) Report only the genres that are influenced by viewing time. Justify each affected genre with a simultaneous confidence interval using a Bonferroni correction. You may use any required table from your textbook. Please give details of calculations.

Reference: Johnson & Wichern, Applied Multivariate Statistical Analysis 6E

The End.