STAT 4350 / 5350
APPLIED MULTIVARIATE ANALYSIS
(Fall 2006)

Lecture: TBA

Instructor: Michael Dowd
Office: Chase 116
Phone: 494-1048
e-mail: mdowd@mathstat.dal.ca
Office Hours: MW 1:00-3:00pm

Course Website:

`www.mathstat.dal.ca/~mdowd/stat4350/stat4350.html`

Copies of assignments, computer code, data sets and some lecture notes will be made available on course website. Announcements will also be posted here.

University Policies

Students with Disabilities:

Students with disabilities should register as quickly as possible at Student Accessibility Services if they want to receive academic (classroom/exam) accomodations. To do so please phone 494-2836, email *disabilities@dal.ca*, or drop in at the Killam G28.

`http://studentaccessibility.dal.ca/`

Intellectual Honesty

`http://plagiarism.dal.ca/policies/index.html`

Method of Evaluation:

- Assignments 20%
- Project 30%
- Midterm Exam 15%
- Final Exam 35%

(Note: no supplemental examination)

Grading Scheme:

A+	93 and above
A	85 up to but not including 93
A-	80 up to but not including 85
B+	75 up to but not including 80
B	70 up to but not including 75
B-	65 up to but not including 70
C+	60 up to but not including 65
C	55 up to but not including 60
C-	50 up to but not including 55
D	47 up to but not including 50
F	below 47

Formal Prerequisites:

1. STAT 3340 - Regression

2. MATH 2135/MATH 2040 - Linear Algebra

Skills Required:

- Basic concepts from introductory statistics including: probability, expectation, inference and hypothesis testing

- Familiarity with multiple regression (matrix formulation) and analysis of variance procedures.

- A knowledge of matrix algebra, and some basic calculus

Course Text: This is a required text and will be followed closely.

Applied Multivariate Statistical Analysis
(fifth edition)
by R.A. Johnson and D.W. Wichern.

(Available in Dal Bookstore)

Statistical Software:

Assignments and projects will require some use of software to carry out multivariate analyses. You are free to use software of your choice. Both Matlab and S-Plus will be used in the course.

Course Goals:

This class deals with the stochastic behaviour of several variables in systems where their interdependence is the object of interest. Emphasis is on both theory and application.

1. Develop an understanding of the theory underlying multivariate statistical methods (foundation)

2. Provide an understanding of how multivariate analysis procedures work, including the ability to apply methods to real data sets and to interpret the results.

3. Provide some basic experience in using modern statistical software packages for multivariate analysis.

Course Outline

- Background - matrix algebra/random vectors (Chapter 2)
- Sample Geometry and random sampling (Chapter 3)
- Multivariate normal distribution (Chapter 4)
- Inferences about a mean vector (Chapter 5)
- Comparisons of several multivariate means (Chapter 6)
- Multivariate linear regression (Chapter 7)
- Principal components (Chapter 8)
- Factor Analysis (Chapter 9)

Course Evalulation Details

Assignments:

There will be regular assignments worth 20% of the total mark. Late assignments will be given a zero mark unless a medical excuse is presented. Assignments will require both analytic work, as well as the use of statistical software.

Project:

There will be a project worth 30% of the total mark. The project will require written report. For graduate students it will also include a oral presentation . Possible topics could include analysis of a multivariate data set of interest, or treatment of a multivariate topic not covered in the course.

Exams:

The midterm will take place during regularly scheduled class time. The final will take place during the Dalhousie exam period as scheduled by the registrar. Further details for these exams will be announced in class

DALHOUSIE UNIVERSITY FACULTY OF SCIENCE
Department of Mathematics and Statistics
STATISTICS 5350 - GRADUATE STUDENTS - Final Examination
Date and Time: 1230-330, Tues, Dec 5, 2006

Name: _____ **Student ID:** _____

This final exam is for **Graduate Students**. There are 9 questions worth 5 points each (the exam is worth 35% of your final mark). The time available is 3 hours. You may refer to a 8.5 in by 11 in formula sheet (double sided) and you may use a calculator.

1. The $n \times p$ data matrix describes n multivariate observations on p variables and plays a central role in multivariate analysis. Remark on the consequences (if any) for multivariate analyses of the following situations.
   (a) Row sums are one
   (b) Column sums are one
   (c) More rows that columns
   (d) More columns than rows
   (e) Correlations between the columns
   (f) Correlations between the rows
   (g) Normality of the columns or rows.

2. Let $\mathbf{x}$ be a $p \times 1$ normally distributed random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Suppose $\mathbf{P}$ is the $p \times p$ matrix of eigenvectors of $\boldsymbol{\Sigma}$. Define the transform $\mathbf{z} = \mathbf{P}^{-1}(\mathbf{x} - \mathbf{d})$ where $\mathbf{d}$ is a vector of constants.
   (a) What is probability distribution of $\mathbf{z}$?
   (b) If $\boldsymbol{\mu} = \mathbf{d}$, what is the distribution for each of the components of the random vector $\mathbf{z}$?
   (c) What is the covariance between the components of $\mathbf{z}$?

3. Outline the major distinctions between factor analysis and principal components analysis. With reference to the method of principal factors, describe how factor analysis may be viewed of as a rotation of the principal components. Describe the ambiguity associated with this rotation, and how it is alleviated in practice.

4. The gas mileage of 4 different car models was tracked over a 1 year period using a sample of 100 cars for each model type. Let x_i denote the mileage of model i ($i = 1, \ldots, 4$). The following correlation matrix was obtained

$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.1 & 0.7 & 0.2 \\ 0.1 & 1.0 & 0.1 & 0.8 \\ 0.7 & 0.1 & 1.0 & 0.3 \\ 0.2 & 0.8 & 0.3 & 1.0 \end{bmatrix}$$

The associated matrix of eigenvalues and eigenvectors are

$$\mathbf{\Lambda} = \begin{bmatrix} 0.1650 & 0 & 0 & 0 \\ 0 & \bullet & 0 & 0 \\ 0 & 0 & 1.4077 & 0 \\ 0 & 0 & 0 & 2.1095 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} \bullet & 0.6834 & -0.5446 & 0.4508 \\ 0.6343 & 0.2680 & 0.5320 & 0.4929 \\ 0.3209 & -0.6374 & -0.5082 & 0.4821 \\ -0.6793 & -0.2343 & 0.4027 & 0.5670 \end{bmatrix}$$

(a) Fill in the 2 gaps (denoted by $\bullet$) found in the eigenvalues and eigenvectors in the above. Show your reasoning.

(b) Interpret the results from this principal components analysis.

(c) What might you expect a factor analysis to yield for the loadings for $m = 2$ factors? Why?

5. Consider the multivariate regression $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y}$ is $n \times m$ and $\boldsymbol{\beta}$ is $p \times m$, and conforming to the usual assumptions.

(a) What extra information is used in this multivariate regression that would not be incorporated if you simply carried out m separate multiple linear regressions? How does this affect the interpretation of the regression results?

(b) Derive the result that

$$\text{cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}$$

Show *all* steps of this derivation be sure to make reference to the necessary regression assumptions you are using.

(c) Suppose that $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{I}$. What can you say about your $n \times m$ matrix of predicted $\mathbf{Y}$ (or $\hat{\mathbf{Y}}$)? What is $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ for this case? Outline a situation where this would occur.

6. In an investigation of adult intelligent scores, a researcher reported that the mean verbal and quantitative scores for $n = 32$ subjects were $\bar{x}_v = 6$ and $\bar{x}_q = 3$ respectively. The associated sample covariance matrix was

$$\mathbf{S} = \begin{bmatrix} 2 & 1.2 \\ 1.2 & 2 \end{bmatrix}$$

   (a) Find the 90% confidence region for (μ_v, μ_q) and provide a graph of it indicating all relevant quantities.

   (b) Test the null hypothesis that the observations come from a population with $\mu_v = 5$ and $\mu_q = 4$. Use an significance level $\alpha = 0.01$.

7. The principal component solution to the factor model allows the total variance of the $p \times p$ sample covariance matrix $\mathbf{S}$ to be written

$$tr(\mathbf{S}) = tr(\mathbf{L}\mathbf{L}' + \mathbf{\Psi})$$

   Define the eigenvalue-eigenvector pairs of $\mathbf{S}$ as $(\lambda_i, \mathbf{e}_i)$, $i = 1, \ldots, p$. The loading matrix $\mathbf{L}$ is defined using the first m eigenvalue/eigenvector pairs as

$$\mathbf{L} = \left[\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_m}\mathbf{e}_m\right]$$

   (a) Suppose that $m = p$. What is $tr(\mathbf{\Psi})$?

   (b) Suppose that $m < p$. Show how the total variance $tr(\mathbf{S})$ can be partitioned into (i) a part, $\sum_{i=1}^{m} \lambda_i$, due to $\mathbf{L}\mathbf{L}'$, and (ii) a remainder, $\sum_{i=m+1}^{p} \lambda_i$, due to $\mathbf{\Psi}$.

8. Consider the zero-mean bivariate normal distribution $f(x, y)$ with equal variances for X and Y and a positive correlation.

   (a) Sketch contours of constant probability density for the joint distribution. Write the equation for the ellipse, centered at the origin, which encloses an probability of 0.9.

   (b) Explain and graphically illustrate the operation of determining a conditional distribution $f(y|x) = f(x, y)/f(x)$ for some value $x = x_0$.

   (c) Consider the marginal distribution $f_x(x) = \int f(x, y)dy$. Does the marginal distribution depend on the correlation? Why or why not?

   (d) Now suppose that X and Y are independent, or $f(x, y) = f_x(x)f_y(y)$. Draw the joint distribution and the marginals.

9. Assume $x_i \sim N(\gamma_i, \sigma^2)$ where $\sum_{i=1}^{p} \gamma_i = 0$ for $i = 1, \ldots, p$. Further, assume that the x_i are independent and jointly normal. Find

   (a) The joint distribution of the random vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$

   (b) The distribution of $\bar{\mathbf{x}}$

   (c) The distribution of $y = \sum_{i=1}^{p} \gamma_i x_i / (\sigma \sum_{i=1}^{p} \gamma_i^2)$

   (d) The covariance between $\bar{\mathbf{x}}$ and y (as defined above)

   Hint: Find matrix expressions for which expresses y and $\bar{\mathbf{x}}$ in terms of $\mathbf{x}$.