STA437H1S/2005H1S
Applied Multivariate Analysis

Instructor: K. Knight (office: 700 University Ave. rm 9083)
E-mail: keith.knight@utoronto.ca

Office hours: To be determined. Do not hesitate to contact me by e-mail as many problems you might encounter can be easily resolved this way. There will also be a Piazza site for this course where students can ask questions about the course material.

About the course: The main goal of this course is to provide students with the some of the tools necessary to analyze multivariate data. The focus of the course will primarily be on exploratory (graphical and computational), rather than inferential, methods; we will also consider some of the statistical theory behind these methods. In addition, we will make extensive use of linear algebra in the course (while trying to avoid making this a linear algebra course!).

Textbook: The required textbook is *An Introduction to Applied Multivariate Analysis using R* by Everitt and Hothorn (Springer). It is available online through the UofT Library system and a link is provided on Quercus (on the Syllabus page). The textbook will be supplemented by handouts on Quercus.

Computing: We will use the software package R extensively in this course both for data analysis as well as some numerical computation. R is free software and can be downloaded (for Windows, Mac, and Linux operating systems) from cran.utstat.utoronto.ca. Of interest to many of you will be RStudio, which provides a very nice environment for using R; information on RStudio (including downloads) can be found at www.rstudio.com.

A useful book that gives a good introduction to R programming is

> *A First Course in Statistical Programming with R* by Braun and Murdoch (Cambridge University Press)

The textbook for this course also provides a lot of examples of R code as will the handouts for the course.

Evaluation: The course grade will be based on four homework assignments (10% each for a total of 40%), a midterm exam (25%), and a final exam (35%).

- Homework assignments will involve both data analysis and theory problems. Two assignments will be handed in before the midterm and two after.
- Students enrolled in STA2005H1S will typically be required to do some additional work on the homework assignments as well as on the exams.

- The midterm exam will be an online exam given on Wednesday March 8; details will be provided later. (To facilitate this exam, there will be no lectures on March 8.) If the midterm exam is missed due to illness or any other circumstances (with appropriate documentation), the weight from the midterm will be carried over to the final exam.
- The final exam will be held during the April exam period at a date and time to be announced later.
- **Students should familiarize themselves with the University's policieson academic integrity, which can be found at** `https://tinyurl.com/tsqukhx` .

Syllabus

The following topics will be covered in the course:

Introduction. Review of linear algebra; covariance, correlation, and distance measures; the multivariate normal distribution; estimation of "dispersion" matrices; introduction to graphical models.

Simple multivariate visualization. pair-wise scatterplots; finding "interesting" projections (projection pursuit); "The Grand Tour".

Principal components analysis. Computation and interpretation; the biplot; multidimensional scaling; independent components analysis.

Cluster Analysis. hierarchical clustering; k-means clustering; model-based clustering.

Classification. Bayes classification; Fisher's linear discriminant; logistic regression; nonparametric methods.

Multivariate Regression. MANOVA, repeated measures designs, functional data analysis.