Syllabus: Statistics 206 Applied Multivariate Statistical Analysis

Instructor:	Yuting Wei, Sequoia 202; `ytwei AT stanford`
Place and time:	MW 3:00-4:20pm; McCullough 122
T.A.:	Zijun Gao; `zijungao AT stanford`
Office hours:	Yuting Wei: W 4:30-5:30pm, Location: Sequoia 202
	Zijun Gao: M 5:00-7:00pm, Location: Sequoia 105 (stats library)
Course website:	W19-STATS-206-01 on *Canvas*

Synopsis: The course will cover canonical methods for learning from multivariate data, plus a selection of more modern methods and topics, time permitting.

Tentative list of topics, with chapters of the main text in parentheses

1. **Getting started with multivariate**
   Introduction, sampling theory, linear algebra [Ch. 1-3]
   Multivariate normal distribution theory [Ch. 4]
2. **Inference about mean vectors**
   Inference on a single vector [Ch. 5]
   Inference on several mean vectors [Ch. 6]
   Multivariate response linear regression (brief) [Ch. 7]
3. **Analysis of covariance structure**
   Principal components [Ch. 8]
   Factor analysis/factor models [Ch. 9]
   Canonical correlations (brief) [Ch. 10]
4. **Classification and clustering**
   Discrimination and classification [Ch. 11]
   Clustering [Ch. 12]
   Multi-dimensional scaling [handouts]
5. **Special topics (time permitting)**
   Independent component analysis [handouts]
   Gaussian graphical models [handouts]

Required textbook: *Applied Multivariate Statistical Analysis*, 6th edition, Pearson/Prentice-Hall 2007, by Johnson Richard A. and Wichern, Dean W.

Additional references:

- Everitt, B. and Hothorn, T. *An Introduction to Applied Multivariate Analysis with R*. Springer 2011.

- Zelterman, D. *Applied Multivariate Statistics with R.* Springer 2015.
- Koch, I. *Analysis of multivariate and high-dimensional data.* Cambridge University Press, 2013.
- Mardia, K.V., Kent J.T., and Bibby, J.M. *Multivariate Analysis.* Academic Press. 1979 (Paperback reprint 2003).

The first three are available as e-books to the Stanford community.

Prerequisite: The *Stanford Bulletin* says STAT 200 (Intro to Statistical Inference), taken concurrently or previously.

- Multivariable calculus and linear algebra: There will be some of both, with a bit more of the latter. The use of calculus will be incidental, not major. The linear algebra (matrices, spectral decomposition etc.) is a bit more important. We'll review many of the results that you need to know.
- Probability and Statistics: Without at least two prior courses (that include probability, maximum likelihood, sampling theory, confidence intervals, testing, etc) this course may prove a bit much.

Homework:

- There will be six homework exercises in total. Homework exercises are usually posted on Fridays, and are due at the *beginning* of Wednesday classes (exact date to be announced). You must turn in a paper copy of the homework. No late homework will be accepted but the lowest one will be dropped for grading. Graded homeworks will be returned in class, or picked up from the TA, as announced.
- Homework is an essential part of the course. You may collaborate on homework problems, but your write-up must be your own. Applied problems (using `R`) *must* be written up into a human-readable document. LaTeXor R markdown is preferred, but you may use a word processing application if you prefer. Markdown must be compiled to PDF, *not html*. Printouts of graphics or tables that are not embedded into a document will not be accepted. Problems that involve no programming may be handwritten.

Final: In class, closed book with one (two-sided) page of notes allowed. *Tentative* date for exam is March 18th 3:30-6:30 p.m. Please do not make conflicting travel plans.

Grading: 40% homework + 60 % final. In addition, your worst homework score will be dropped, assuming you have completed all homeworks.

Software: We will be using the `R` language, freely available at `cran.r-project.org`. `R` use is required for homework assignments. I recommend use of the `RStudio` IDE, available for free at `rstudio.com`. If you are unfamiliar with the language, there are resources provided on our canvas folder and there will also be a TA session to get you started with `R`.

There will be time to get up to speed with `R` during the earlier part of the course before our use of it becomes more substantial. A helpful introductory book (also online via the Library) is Dalgaard, P, *Introductory Statistics with R*, Springer 2008. Finally, never underestimate the power of Googling things and finding code snippets on stack overflow.

Canvas folders: The folder architecture is fairly self-explanatory.
`Lectures` contains lecture notes for each class meeting, including source, from which you can pull code blocks.
`Homework` contains homework assignments
`Data` contains some datasets; please download the Johnson and Wichern datasets from `http://esminfo.prenhall.com/math/johnsonwichern/data.html`
`R Resources` contains the R tutorial and information.