Math 570 Applied Multivariate Analysis.
Spring 2016

- **Instructor:** Xingye Qiao
- **Email:** qiao@math.binghamton.edu
- **Phone number:** (607) 777-2593
- **Office:** WH-134

- **Meeting time & location:** MWF 10:50 – 11:50 at WH-100E.
- **Office hours:** MW 3–4 and F 9:45 –10:45
  If you need to reach me, please e-mail qiao@math.binghamton.edu.
  Please include [Math570] in the subject line of your email, or your email may not be read promptly.

Prerequisite

Math 501 and Math 502, or equivalent. **Graduate students from outside of the mathematical department and senior undergraduate students may take this course with some preparation (please consult the instructor prior to the semester).** One lecture session will be devoted to reviewing linear algebra materials that are useful in this course.

Learning Objectives

1. A review of the theoretical aspect of Multivariate Statistical Analysis, including: multivariate normal distributions, the multivariate Central Limit Theorem, quadratic forms, Wishart distributions, Hotelling's T square, inference about multivariate normal distributions.
2. Modern applied multivariate statistical methods, including: Principal Component Analysis, Canonical Correlation Analysis, Classification (Bayes rule, Linear and Quadratic discriminant analysis, cross-validation, and logistic regression etc.), factor analysis and Independent Component Analysis, clustering and multidimensional scaling.
3. Machine learning approaches, including Classification and Regression Trees, Support Vector Machine and other large margin classifiers, kernel methods, LASSO and sparsity methods, additive models, etc., if time permits.

Recommended Texts

The required texts are **Härdle & Simar 2012** and **Izenman 2013** (see below for details).

- **Elementary**
  - Johnson, Richard A & Wichern, Dean W. 2007. Applied multivariate statistical analysis. Upper Saddle

River, N.J: Pearson Prentice Hall. Amazon Link
  - Härdle, Wolfgang & Simar, Léopold. 2012. Applied multivariate statistical analysis. Berlin: Springer (also visit here for sample codes). Amazon Link

- **Advanced and applied**
  - Izenman, Alan Julian. 2013. Modern multivariate statistical techniques: Regression, classification, and manifold learning. New York: Springer New York. Amazon Link || Book Home Page (including R, S-plus and MATLAB code and data sets)
  - Hastie, Trevor, Tibshirani, Robert, and Friedman, J. H. 2009. The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer New York. Amazon Link
  - James, Witten, Hastie and Tibshirani, 2014. An Introduction to Statistical Learning with Applications in R. Book Home Page. The PDF file of the book can be downloaded for free. There is also a R library for this book.

- **Theoretical**
  - Anderson, T. W. 2003. An introduction to multivariate statistical analysis. Hoboken, N.J: Wiley-Interscience. Amazon Link
  - Muirhead, Robb J. 1982. Aspects of multivariate statistical theory. New York: Wiley. Amazon Link

- **Working with R or SAS**
  - Everitt, Brian, and Hothorn, Torsten. 2011. An introduction to applied multivariate analysis with R. New York: Springer. Amazon Link
  - Khattree, Ravindra, and Naik, Dayanand N. 1999. Applied multivariate statistics with SAS software. Cary, NC: SAS Institute. Amazon Link
  - Khattree, Ravindra, and Naik, Dayanand N. 2000. Multivariate data reduction and discrimination with SAS software. Cary, NC: SAS Institute. Amazon Link
  - A Little Book of R for Multivariate Analysis

Grading

- Homework (50%): there will be about four to five homework assignments.
- Midterm exam (20% or 15%): a midterm exam focusing on the theoretical part of the course will be administered in the middle of the semester. Students who are not in the Math PhD program will receive a slightly easier set of problems and a smaller weight for the midterm exam than Math PhD students will do.
- Presentation (10%): each student will choose a research topic (either original research or research conducted by other researchers) related to this course and give a 30-minute presentation. The presentation of each student shall be judged by peer students and the instructor.
- Course project (15% or 20%): a final project will be assigned to each student by the end of the semester. Students who are not in the Math PhD program will gain a greater weight for the course project than Math PhD students will do. The guidelines for the final project can be found here.
- Lecture attendance and participation (5%)

Software

There is no designated software for this course. You may use the software that makes the most sense for you. Many pharmaceutical companies use SAS for compliance with FDA regulations. Academic intuitions as well as labs often use R and python. Corporations often use MATLAB, Stata, Minitab, S, etc. because of the relatively high reliability despite the cost. However, it is expected that the student immerse herself with use of at least one software.

> Used to be expensive, SAS University Edition is now free for download and use.

Schedule

The course deck may be downloaded from the blackboard.

- Week 1 (Jan 25 - 29)
  - Multivariate Data Exploration, Reading: HS Ch. 1, Izenman Ch. 4.
  - Matrix algebra review, Reading: HS Ch. 2, Izenman Sec. 3.2.
- Week 2 (Feb 1 - 5)
  - Random vectors and multivariate normal distribution; The Wishart distribution
- Week 3 (Feb 8 - 12)
  - The Wishart distribution; Inference about multivariate normal distribution
- Week 4 (Feb 15 - 19)
  - PCA
- Week 5 (Feb 22 - 26)
  - CCA; Classification: Bayes rule
- Week 6 (Feb 29 - Mar 4)
  - LDA, QDA, Logistic regression
- Week 7 (Mar 7 - Mar 11)
  - Cross validation and other classifiers; FA;
- Week 8 (Mar 14 - Mar 18)
  - ICA; k-means; Hierarchical clustering
- Week 9 (Mar 21 - Mar 25)
  - Gaussian mixture/EM
- Week 10 (Mar 28 - Apr 1)
  - Spring Break
- Week 11 (Apr 4 - Apr 8)
  - MDS; SVM
- Week 12 (Apr 11 - Apr 15)
  - 4 presentations and midterm exam.
- Week 13 (Apr 18 - Apr 22)
  - LASSO
- Week 14 (Apr 25 - Apr 29)
  - Tree methods
- Week 15 (May 2 - May 6)

    - bootstrap, bagging, subsampling; boosting; RF
- Week 16 (May 9 - May 11)
    - final projects.

Student Presentations

- Feb 26: **Xu Chu** presents Zou et al. 2006, “Sparse Principal Component Analysis.”
- Feb 29: **Tianqi Zhang** presents Shen and Huang 2007, “Sparse principal component analysis via regularized low rank matrix approximation.”
- Mar 4: **Liping Gu** presents Tibshirani et al. 2002, “Diagnosis of multiple cancer types by shrunken centroids of gene expression.”
- Mar 7: **Xin Gu** presents Fan and Fan 2008, “High-dimensional classification using features annealed independence rules.”
- Mar 11: **Haomiao Meng** presents Bickel and Levina, 2004, “Some theory for Fisher's linear discriminant function, `naive Bayes', and some alternatives when there are many more variables than observations.”
- Mar 16: **Baiyang Qi** presents Ahn and Marron, 2008, “The maximal data piling direction for discrimination.”
- Mar 21: **Rui Gao** presents Witten and Tibshirani, 2011, “Penalized classification using Fisher's linear discriminant.”
- Mar 25: **Yinsong Chen** presents Mai and Zou 2013, “A Note On the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Method.”
- Apr 4: **Hao Xu** presents Zhu and Hastie, 2004, “Classification of gene microarrays by penalized logistic regression.”
- Apr 11: **Sreedhar Kumar** presents Sun and Wang, 2012, “Regularized k-means clustering of high-dimensional data and its asymptotic consistency.”
- Apr 11: **Yuan Fang** presents Witten and Tibshirani, 2010, “A framework for feature selection in clustering.”
- Apr 13: **Wenming Deng** presents Liu et al., 2008, “Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data.”
- Apr 13: **Liang Chen** presents Jung and Qiao 2014, “A statistical approach to set classification by feature selection with applications to classification of histopathology images.”
- Apr 22: **Junle Lu** presents Zhu and Hastie, 2005, “Kernel logistic regression and the import vector machine.”
- Apr 29: **Miaolin Fan** presents Wang, Nan, Rosset and Zhu, 2011, “Random lasso.”
- May 2: **Xiang Li** presents Qiao and Zhang 2015, “Distance-weighted Support Vector Machine.”