Hong Kong Baptist University
Faculty of Science
Department of Mathematics

Title (Units): **STAT 3710 MULTIVARIATE ANALYSIS AND DATA MINING (3,3,0)**

Course Aims: To provide an understanding of the classical multivariate analysis and modern techniques in data mining. Very often, observations in the social, life and natural sciences are multidimensional or very high dimensional. This kind of data sets can be analyzed by techniques in multivariate analysis and/or data mining. With the help of statistical package, such as Matlab, students will learn how to treat real multivariate problems.

Prerequisite: STAT 2110 Regression Analysis

Prepared by: H. Peng

Learning Outcomes (LOs):

Upon successful completion of this course, students should be:

No.	Learning Outcomes (LOs)
	Knowledge
1	Able to apply the basic graph techniques to find useful information from multivariate data
2	Able to understand statistical theory of multivariate normal distribution
3	Able to apply projection technique to analyze multivariate data
4	Able to apply classification technique to mining useful information from multivariate data.
	Skills
5	Able to manipulate the software Matlab
6	Able to figure graphs for multivariate data
7	Able to write Matlab program to calculate multivariate statistics
8	Able to design and implement innovative multivariate data processing system for mining special useful information from data.
	Attitudes
9	Able to work effectively in a team
10	Able to solve problems independently

Assessment:

No.	Assessment Methods	Weighting	Remarks
1	Continuous Assessment (assignments, and mini-project)	40%	Assignments are designed to measure students understanding of the theory of multivariate statistics and data mining. The mini-project is designed to achieve LOs 8-10 by facilitating students working in a team environment to creatively model to multivariate datasets, and mining some useful information from the multivariate statistical models.
2	Final Examination	60%	Final Examination is designed to see how far students have achieved their intended learning outcomes especially in the Knowledge domain. Students should have a thorough understanding of the knowledge and apply them correctly in different context to do well in the exam.

Learning Outcomes and Weighting:

Content	LO No.	Teaching (in hours)
I. Introduction and Matrix Algebra	2-3	6
II. Multivariate Normal Distribution and Its Sampling Theory	1-2, 5-6	8
III. Tests of Hypotheses on Means and Covariance Matrices	1-2, 5-7	6
IV. Multivariate Methods by projection	1,3, 5-6, 8	10
V. Classification	1,4, 5-6, 8	10

Textbook: R.A. Johnson and P.W. Wichern, Applied Multivariate Statistical Analysis, 5th Ed., Prentice-Hall International Book Company, 2002.

References: D.F. Morrison, Multivariate Statistical Methods, 3rd Ed., McGraw-Hill International Book Company, 1990.

T.W. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd Ed., Wiley, 2003.

T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical learning, data Mining, Inference, and Predeiction, Springer, 2001.

J. Han and M. Kamber, Data Mining: Concepts and Techniques, The Morgan Kaufmann Publishers, 2001.

Software: Matlab

Course Content in Outline:

	Topic	Hours
I.	Introduction and Matrix Algebra A. Introduction to multivariate analysis B. Basic statistics of a data set C. Data displays and graphical representations D. Matrix algebra E. Differentiation with vectors and matrices	6
II.	Multivariate Normal Distribution and Its Sampling Theory A. Random vector and its distribution B. Moments of multivariate distributions C. Multivariate normal distribution D. Matrix normal distribution E. Maximum likelihood estimation F. Properties of estimators	8
III.	Tests of Hypotheses on Means and Covariance Matrices A. From univariate to multivariate problems B. Tests of hypotheses on means and the T2-statistic C. Two samples problem D. Testing equality of several means E. Some tests on covariance matrices	6
IV.	Multivariate Methods by projection A. Principal component analysis B. Factor analysis C. Corresponding analysis D. Canonical correlation analysis	10
V.	Classification A. Discrimination analysis B. Clustering analysis C. Decision tree D. Boosting E. Support Vector Machines.	10