MUTIVARIATE DATA ANALYSIS

Significant advances have been made in sensing technology, computing resources, and communication. This has led to the acquisition of large amounts of data from diverse sources. The analysis of large amounts of such data to obtain useful information has become extremely important. The terms 'data mining' and 'big data' is often used to denote them as the next big challenges.

Multivariate statistical methods are the key tools used for data mining. Such techniques are now being used in industrial applications such as fault diagnosis, controller performance monitoring, process modeling etc. Commercial software packages are now available for readily deploying these techniques. This course will provide an introduction to some of the important multivariate methods, belonging to the Principal Components Analysis family. Applications in process industries, chemometrics, biomedical and speech signal processing using these techniques will be provided. Besides students of chemical engineering, this course will be of use to students in industrial mathematics, biotechnology, chemistry, computer science and electrical engineering.

Students are expected to be familiar with MATLAB or Python. A basic knowledge of linear algebra and probability & statistics will also be useful for better understanding of the course

A single text book will not be used as reference for this course. Instead, individual chapters from books, lecture notes available on Internet, and journal papers will be used as primary reference material.

Course Contents

Linear Regression in two variables: Ordinary Least Squares Regression (OLS), Total Least Squares Regression (TLS), and Weighted TLS (WTLS) for single input, single output (SISO) systems

Multilinear Regression: OLS based regression involving multi-input and single output (MISO) systems, Model quality assessment, Confidence Intervals for regression parameters. Extension to multi-input multi-output (MIMO) systems

Multivariate Regression: TLS based regression for multi-input multi-output (MIMO) systems using Principal Component Analysis (PCA)

Perspectives of PCA: PCA as a compression technique, PCA as a denoising technique and PCA as a model identification/regression technique

Iterative PCA: Simultaneous model identification and noise variance estimation

Applications of PCA: Image compression, Multivariate calibration, Model identification for flow processes, Fault Diagnosis

Nonlinear Regression: OLS based multivariate nonlinear regression given model structure, Kernel PCA and Kernel PCR

Dynamic model identification: Last PCA, Dynamic PCA and Dynamic Iterative PCA for SISO systems

Non-negative Matrix Factorization (NMF) for blind deconvolution.

Application of NMF in chemometrics and speech processing

GRADING POLICY

Assignments (4 or 5): 25 marks

Quizzes (2): 30 marks

Final Exam: 45 marks

Depending on the class strength, the final exam may either be a take home exam or an in-class computer based exam.