UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

# COURSE SYLLABUS

# Chemometrics

**2425-2-F7501Q047**

## Aims

The main objective of the course is to provide the student with the theoretical foundations and operational tools of the main chemometric techniques, necessary in modern chemistry for the adequate treatment of the information contained in experimental chemical data. The aim of the course is therefore to provide the student with the fundamental elements of multivariate analysis to treat complex systems of chemical, pharmacological and environmental interest. Knowledge of the principles and operating conditions of the main chemometric techniques will be developed together with the ability to choose and manage the investigation approaches most suited to the purposes of the analysis. The student will then be able to evaluate the characteristics of the basic chemistry approaches, the fields of application, the advantages and disadvantages of the individual techniques and will therefore be able to suggest the choice of the chemometric technique considered most suitable for a specific problem.

In particular, at the end of the course, the student must demonstrate that he/she is able to achieve the following formative objectives.

**Knowledge and understanding:**

o describe the main chemometric methods presented in the course
o describe the fundamental parameters for the evaluation of the results
o describe the advantages and disadvantages of the different chemometric algorithms

**Applying knowledge and understanding:**

o select the most suitable multivariate method to deal with a specific problem
o understand which kind of information will be possible to extract from the data under analysis
o evaluate alternative chemometric methods to face the problem
o concretely apply the selected chemometric methods and calculate the related statistical information parameters

**Making judgements:**

o acquire knowledge and skills to develop a critical understanding of the main chemometric methods
o select methods and parameters useful for extracting specific information from the data under analysis
o justify a critical discussion on the methods used and the information obtained from the analysis of the data

**Learning ability:**

o understand the different chemometric approaches and their methodological application in order to use them correctly when analysis a multivariate problem

## Contents

Introduction to chemometrics. The structure of multivariate data. Strategies for the rationalization of complex problems: Principal Components Analysis. Similarity and diversity. Cluster Analysis methods. The concept of bias and validation methods. Multivariate regression methods. Multivariate classification methods. Introduction to neural networks and selection of variables. Data fusion strategies. Analysis of the relationships between molecular structure, chemical-physical properties and biological activities (QSAR).

Practical experience in the laboratory to learn the tools and methods of analysis through the main chemometric techniques described in the course (analysis of the structure of chemical data, calibration of regression models and classification).

## Detailed program

Introduction to chemometrics: objectives, methods and applications of chemometrics for the analysis of complex chemical systems. The structure of multivariate data. Elements of matrix calculation. Elementary statistical parameters: position and dispersion indices, covariance and correlation. Data scaling and pre-treatment: centering, auto-scaling, range scaling, variance scaling.

Strategies for the rationalization of complex problems, the analysis of the structure and the exploration of chemical data related to complex systems; Principal component Analysis (PCA): objectives of the PCA, references to the algorithm of diagonalization, matrices of scores and loadings; eigenvalues ??and definition of significant components (rank analysis). Examples of PCA application on chemical data. Multivariate correlation.

Analysis of similarity and diversity in complex systems: the concepts of analogy, similarity, dissimilarity and distance. Distance and similarity measures for quantitative and binary data. Cluster Analysis: agglomerative hierarchical methods and non-hierarchical methods. Similarity analysis strategies. Examples of application of Cluster Analysis on chemical data.

The concept of bias and validation methods: statistical estimators; bias and variance. Descriptive and predictive models. Validation techniques of multivariate statistical models: cross-validation, bootstrap, leave-one-out, leave-many-out, and scrambling.

Multivariate regression methods: strategies based on quantitative models and regression parameters. Multiple regression analysis. The biased regression methods: methods reduces, selection of the best sub-models, regression with Principal Components, Partial Least Square regression. Genetic algorithms for the selection of variables. The Sequential Replacement method. Examples of application of multivariate regression on chemical data.

Multivariate classification methods: strategies based on classification and classification parameters. The methods

of local classification: k Nearest Neighbors (kNN), N3, BNN. The Bayesian probabilities and methods of linear and quadratic discriminant analysis. Tree classification methods (CART). Neural networks and Kohonen maps.

Consensus and data fusion methods: introduction to modern strategies for the concatenation of different sources of chemical information through consensus analysis and data fusion approaches; definition of data fusion levels.

Introduction to the relationships between molecular structure, chemical, physical, biological and environmental properties (QSAR): QSAR methodologies, molecular descriptors and their application.

There are three practical experiences on real data to acquire the tools and methods of analysis on the following chemometric themes: analysis of the chemical data structure through Principal Components Analysis, calibration of regression and classification models. Practical experiences are performed in the informatic laboratory using MATLAB software and specific graphical tooboxes provided by the teachers. The experiences include a brief introduction to the use of MATLAB software (data import and management, integration with multivariate statistical toolboxes provided by teachers).

## Prerequisites

Basic knowledge on the main elementary statistical indices, basic computer operating skills in practical laboratory experiences.

## Teaching form

The course is divided into a part of lectures and frontal exercises, in which the theoretical notions on chemometric themes are given. At the end of the course lectures, the students attend three different practical sessions in the infromatic laboratory to acquire the tools and operating methods for the analysis of the main chemometric techniques.

## Textbook and teaching resource

The reference textbook of the course is: R.Todeschini, Introduzione alla Chemiometria (Edises, Naples 1998). The book is also provided in pdf format on the e-learning page of the course. The slides of the lessons are also provided on the e-learning page of the course. In addition, the teachers provide on the e-learning platform a chm file including scientific articles for the study of the topics presented in the course. For each laboratory experience, through the e-learning platform, the introductory slides, the data and the toolboxes necessary for the development of the experiences through MATLAB software are provided.

## Semester

Fisr semester

## Assessment method

The exam consists of an oral examination, where topics presented in the lessons are discussed. In addition to the theoretical fundamentals given in the course, students' skills and aptitudes are also assessed to adapt the theoretical foundations of chemometrics to particular operative and practical conditions; the expositive ability and adequacy of the student's language is also assessed.

For the admission to the oral examination a multiple response test is provided in the informatic laboratory; each test includes 30 questions on the topics presented in the lessons of the course; students who obtain a positive result (at least 60% correct answers) can take the oral exam. To access the oral examination, it is mandatory to attend the lab experiences.

The oral test is usually consecutive to the computer test but the student can ask to have the oral examination in subsequent dates to the multiple response test. In the case of not passing the oral examination, there is no jump call and a new multiple response test is not required.

## Office hours

Teachers are always available to receive students in their offices upon an e-mail request.

## Sustainable Development Goals