PhD courses: Scienze Chimiche &

International Doctorate in Structural Biology

SYLLABUS

1 Lecturer information

Name and Surname: Hans Grahn

Affiliation: Sapheneia, Inc. (Colorado USA)
e-mail: hans.picea@gmail.com

Proposed by: Cristina Nativi

e-mail: cristina.nativi@unifi.it

2 Title of the course

"Chemometrics and Experimental Design in Chemistry"

3 Course program

Data files and formats

The starting point of all data analysis is numbers available in files. A good starting point is to learn how different file formats present the data and to go from one format to another one.

Getting a quick overview of the raw data by using different types of plots.

An easy way of getting a first overview of a large data sets is to visualize the whole set or parts of it. Different techniques can be used. A visualized data set is a map; reading and understanding this map is the only way use the data in a meaningful way.

Preprocessing of data

All data gains from preprocessing for easier interpretation and some basic techniques for preprocessing are fun-damental knowledge obtained here.

Visualizing statistical parameters from the raw data.

Once the raw data are mapped and visualized, statistical parameters can be visualized to create visual summar-ies of the data.

What is structure and what is noise?

All data contain different types of noise originating from different sources. Therefore, it is important to separate the noise form meaningful and useful structures. Noise can be visually described and interpreted.

Multivariate models for data exploration.

When data are correlated, a simplification can be made by calculating a multivariate model. A number of these models are introduced and their visual interpretation is explained and trained.

Visualization of the raw data to get an overview

Date file formats and how to extract an overview map. How to use this map to quickly get beforehand information on the data.

Data pre-processing

All data needs some form of preprocessing. Some preprocessing methods are statistically based and some are based on the physics of the data generating equipment.

Multivariate data analysis

The preprocessed data form a structure in multivariate space and there are methods (algorithms) to simplify this structure.

Interpretation of multivariate results as tables and figures.

Once a multivariate space model is made, it can be interpreted by looking at tables and figures. Making the cor-rect figures and tables and interpreting them is of the utmost importance.

Studying and removing noise components

Data analysis may become an iterative process where noise components are removed after a first model is made and then a new model is constructed from the modified data set.

4 Course content detailed per lesson of two hours (possibly with dates and room real and virtual)

Lesson 1: April, 15th 2024 - 11:00-13:00 - Introduction of the course, background, basic statistics

Lesson 2: April, 16th 2024 - 11:00-13:00 – Practical workshop on Multivariate Analysis

Lesson 3: April, 17th 2024 - 11:00-13:00 – Multivariate Regression and predictions

Lesson 3: April, 18th 2024 - 11:00-13:00 - Multivariate Image Analysis (MIA) for chemists

Lesson 4: April, 22nd 2024 - 11:00-13:00 - Exam check (questionnaire)

Lesson 5: April, 24th 2024 - 11:00 -13:00 - Questionnaire results

5 Suggested reading

=====

6 Learning Objectives

Many scientific studies and industrial processes are capable of producing huge amounts of data and one quickly loses overview. The course introduces methods to obtain insight in large to mega amounts of data and to regain control by easy visualization. All large data sets are only structures in multivariate space and there is an easy pathway from raw data (a file) to a multidimensional

visualization of this space giving a separation in meaningful structures and noise structures. Tables of statistical properties of the data and figures showing relationships inside the data can be made availa-ble for interpretations and decision making.

7 Knowledge and Skills to be acquired

The course participants will learn to handle data sets, data files, mapping of the data and a minimum of multivar-iate analysis and a deeper understanding of multivariate space and how to interpret the models made on it. Participants can bring their own data.

8 Prerequisites

Master degree in: Chemistry, or Medicinal Chemistry, or Biology or Physics

9 Teaching Methods

☐ MODE 1 - Pre-recorded lessons uploaded on the moodle platform (a meeting must be organized with PhD students in order to clarify eventual doubts)

X MODE 2 (preferred) - Lessons delivered in-person and in remote with simultaneous recording by the WEBEX platform

(The lessons must be recorded and available to all the students that cannot take part to the lessons in streaming. The Webex platform must be used. All course content should be uploaded to the Moodle platform on the Chemical Sciences PhD page "Courses and Seminars of the PhD in Chemical Sciences 2022-2023")

10 Further information

N/A

11 Type of Assessment

The final evaluations will have to be validated maximum 1 month after the end of the course

Written Test

12 Period

April, 22nd 2024 (11:00-13:00)