INTRODUCTION

- Chemometrics Introduction
  What is this and why we need it
- Some definitions
- Overview of methods
- Examples

Without equations!

THE USE OF MULTIVARIATE ANALYSIS IN THE DISCIPLINE X:

- Biometrics (used in biology)
- Technometrics (used in engineering)
- Psychrometrics (used in psychology)
- Chemometrics (used in chemistry)

Statistical, mathematical or graphical technique, considers multiple variables simultaneously

“Chemometrics is the chemical discipline that uses mathematics and statistics to design or select optimal experimental procedures, to provide maximum relevant chemical information by analyzing chemical data, and to obtain knowledge about chemical systems”

D. L. Massart
CHEMOMETRICS – INTRODUCTION

The scientific world today

- The data flood generated by modern analytical instrumentation produces large quantity of numbers to understand and quantify phenomena around us.
- The evolution of personal computers allows faster acquisition, processing and interpretation of chemical data.
- Every scientist uses software related to mathematical methods or to processing of knowledge.
- A deeper understanding of these methods and tools for viewing all data simultaneously are needed.

CHEMOMETRICS – INTRODUCTION

Useful at any point in an analysis, from the first conception of an experiment until the data is discarded.

- Use of mathematical and statistical methods for selecting optimal experiments
  - Statistical experimental design
  - Design of Experiments (DoE)...
- Extracting maximum amount of information when analysing multivariate (chemical) data
  - Classification
  - Process monitoring
  - Multivariate calibration...
Huge growth area in past 15 years

- Process Control and analysis
- Food and feed analysis
- Biology – metabolomics etc
- Environmental monitoring
- Analytical Chemistry

Linear algebra is the language of Chemometrics. One cannot expect to truly understand most chemometric techniques without a basic understanding of linear algebra (Wise and Gallagher, 1998)

Matrix and vector operations

- Samples are referred to as OBJECTS
- Measurement results (e.g. concentration, absorbances, …) are referred to as VARIABLES
- A data table of K variables and M objects is referred to as a DATA MATRIX OF SIZE M x K

If X has 3 rows and 5 columns
3 x 5 data matrix

CHEMOMETRICS: Extract meaningful information about the objects and the variables from data matrices
In summary...

\[
X = \begin{bmatrix}
X_{11} & X_{12} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
X_{m1} & X_{m2} & \cdots & X_{mp}
\end{bmatrix}
\]

CHEMOMETRICS – IMPORTANT DISCIPLINES

- Sampling, selection of objects and variables
- Clustering
- Multivariate regressions, calibrations and predictions
- Neural Networks
- Validation
- Graphical display and outlier detection

Sample Selection

- Calibration data set (X)
- Sample Selection

Population sources of variation:
- origins of samples
- processes
- varieties
- storage conditions
- sample preparations (t, particle size)
- residual moisture

Measurement factors:
- room temperature
- operator
- instrument setup

Source: Chemometrics – Introduction - Jens C. Frisød
Grouping of objects e.g. how similar is the behaviour of compounds, how similar are products…

**PCA**

Looking at relationships:

- Between samples: food samples / patients / people / spectra / …
- Between variables: elemental compositions / compound concentrations / spectral peaks / …

**Principal Component Analysis (PCA)**

PCA is a chemometric technique for visualizing high-dimensional data. PCA reduces the dimensionality (the number of variables) of a data set by maintaining as much variance as possible.

Scores on PC 1 (96.09%) Scores on PC 3 (0.34%)

Objectspace

Clusters

Outliers

Quantitative or qualitative estimation. Especially mixtures.

Univariate calibration: one measurement e.g. a peak height

Multivariate calibration: several measurements e.g. spectra

Statistical, mathematical or graphical technique, considers multiple variables simultaneously.

**Calibration**
Linking two sets of data together: peak height to concentration / spectra to concentrations / biological activity to structure / …

Multivariate Calibration

- Relate instrumental response to chemical concentration.
- Improper calibration yields improper results.
- Linear, non-linear and multivariate algorithms are available, depending on the instrument and the analysis involved.

- Infra-red spectroscopy (IR-Raman)
- Fourier Transform Infra-red spectroscopy (FTIR)
- Nuclear Magnetic Resonance (NMR)
- X-ray Fluorescence (XRF)
- X-ray Diffraction (XRD) spectroscopy
- Moisture (regression)
- Concentration of protein (regression)
- Country of origin (discrimination)
- PDO (discrimination)
- …

Calibration uses empirical data and prior knowledge for determining how to predict unknown quantitative information from available measurements via some mathematical transfer functions.
Multivariate analysis

Unsupervised

Principal Component Analysis (PCA)
Cluster Analysis (CA)

Supervised

Regression

Multiple Linear Regression (MLR)
Principal Component Regression (PCR)
Partial Least Squares (PLS)
Artificial Neural Networks (ANN)
LS-Support Vector Machines (LS-SVM)

Local techniques

PLS-Discriminant Analysis (PLS-DA)
SIMCA
K-Nearest Neighbours (K-NN)
Support Vector Machines (SVM)

Unsupervised - Pattern recognition

Supervised - Classification

It includes any techniques for model training where the focus is on the relationships between independent variables of interest and one or more dependent variables (spectra).
Find a set of predictor variables which gives a good fit, predicts the dependent value well and is as small as possible.

- **Backward elimination**: Start with all the predictor variables and potentially drop predictor variables in subsequent steps.

- **Forward selection**: Start with no predictor variables and potentially add predictor variables in subsequent steps.

- **Stepwise regression**: Combination of backward elimination and forward selection.

Reducing number of variables

Possibility of constructing fast spectrometers based on a reduced number of variables aiming to a reduction of training and utilization times to be used, for instance, in a conveyor belt. Moreover, the fact of grouping the different selected variables in regions allows an easy chemical interpretation of the spectra.

**References**


**CHEMOMETRICS – SOFTWARE**

**TABLE XXVI**

<table>
<thead>
<tr>
<th>Type</th>
<th>Company</th>
<th>Address</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declared</td>
<td>NIRS</td>
<td>Instrument, Inc.</td>
<td>7916 West Chevrolet Blvd., P.O. Box 53190, 814-762-6608 (USA)</td>
</tr>
<tr>
<td>NIRS</td>
<td>Foss</td>
<td>Silver Spring, MD, U.S.A.</td>
<td>301-256-0134 or 301-303-0425 (USA)</td>
</tr>
<tr>
<td>Vision</td>
<td>Foss</td>
<td>Silver Spring, MD, U.S.A.</td>
<td>301-256-0134 or 301-303-0425 (USA)</td>
</tr>
<tr>
<td>Basdee</td>
<td>Basdee</td>
<td>Buffalo, NY, U.S.A.</td>
<td>716-670-4832 (USA)</td>
</tr>
<tr>
<td>SpectraTech</td>
<td>LT Industries</td>
<td>Rochester, NY, U.S.A.</td>
<td>315-442-3205 (USA)</td>
</tr>
<tr>
<td>Analytic</td>
<td>Analytic</td>
<td>Cordova, Switzerland</td>
<td>+41 71 805 7050 (CH)</td>
</tr>
<tr>
<td>Delguex</td>
<td>Delguex</td>
<td>Boehringer Corporation</td>
<td>805-303-4434 (CH)</td>
</tr>
<tr>
<td>EAS</td>
<td>EAS</td>
<td>Biomedica, Chorley, U.K.</td>
<td>01254 751 207 (UK)</td>
</tr>
<tr>
<td>General</td>
<td>General</td>
<td>Galactic Industries</td>
<td>602-692-2288 (USA)</td>
</tr>
<tr>
<td>Draper</td>
<td>Draper</td>
<td>Woodland Hills, CA, U.S.A.</td>
<td>310-660-1000 (USA)</td>
</tr>
<tr>
<td>Unisyn</td>
<td>Unisyn</td>
<td>Cambridge, U.K.</td>
<td>510-701-1042 (CA)</td>
</tr>
<tr>
<td>U-Met</td>
<td>U-Met</td>
<td>The MetWorks, Inc.</td>
<td>501-624-7701 (USA)</td>
</tr>
</tbody>
</table>
SVM discrimination models

*NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed J.A. Fernández Pierna, Ph. Vermeulen, O. Amand, A. Tossens, P. Dardenne and V. Baeten. Special issue Chemometrics and Intelligent Laboratory Systems 137 (2015) 231-239*
EXAMPLE: LI E S - IMPURITIES

KNN - PCA models

EXAMPLE: LI E IL - QUALITY PARAMETERS

EXAMPLE: LI E IL - ADULTERATION

EXAMPLE: SUGAR BEET PLANTS - CUST

SVM discrimination models
Multivariate regression method comparison: PLS, ANN and LS-SVM

Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products

Feed: Ash, Fat, Fibre, Starch, Protein

Feed Ingredients: Ash, Fat, Fibre, Protein

Fresh Silages: Dry Matter, Fibre, Protein

Soils: CEC, COT_Sk, N_Kj

Pre-processing: SNV + detrend + First derivative
EXAMPLE: FEED PRODUCTS

EXAMPLE: FEED PRODUCTS

EXAMPLE: FEED PRODUCTS

EXAMPLE: TRANSFERT

EXAMPLE: TRANSFERT

http://www.cra.wallonie.be/
j.fernandez@cra.wallonie.be

QUESTIONS ABOUT CHEMOMETRICS?

Vibrational spectroscopy and Chemometrics

Training session
17-21 February 2014

Vibrational Spectroscopy
NIR/MIR/RAMAN
Theory and applications, experimental/sample plan

IR microscopy/Hyperspectral imaging/embedded IR/
MEMs/instrument standardization/networking

Chemometrics applied to vibrational data

Exploratory analysis:
Data visualization
Principal component analysis
Outlier detection
Uncertainty estimation

Quantification and classification:
Multivariate calibration:
Partial least squares PLS
Multiple linear regression MLR
Support vector machines SVM