Multivariate Analysis

Introduction
The Course...

• Multivariate Analysis (Chemometrics 1)
• Lectures and Lab
  – Room ...
  – Mon/Fri, 10:30 to 12:10 h
• My room: 209, IQ-1
• Softwares
  – R/Rstudio
  – Matlab, Octave, Spreadsheets, ...
• E-mail
  – Subject: [MA] ...
  – anselmo.disciplinas@gmail.com
• Course website
  – anselmo.quimica.ufg.br → Multivariate Analysis
    • Support materials
    • Syllabus
Final Course Exam

• Scientific report
  – Research project
  – Science articles
    • 2015...
    • QUALIS A1... B2

  – Deadline: 12/16
• **Who are you?**
  
  – Name
  – Background
  – Research Project
is not a *single* tool but a *range of methods*

– Basic Statistics
– Signal Processing
– Factorial Design
– Calibration
– Curve Fitting
– Detection
– Pattern Recognition
– Neural Networks
– Data Mining
– …
Exploratory Data Analysis

EDA can reveal *hidden patterns* in complex data by reducing the information to a more comprehensible form.
EDA can expose possible outliers and indicate whether there are patterns or trends in the data.

- In-Line Outlier:
  - Trend with only the In-Line Outlier:
    - $y = 0.97x + 1.21$
    - $R^2 = 0.93$
    - $n = 29$

- Cross-Trend Outlier:
  - Trend with only the Cross-Trend Outlier:
    - $y = 0.75x + 3.70$
    - $R^2 = 0.37$
    - $n = 29$

- Fringe Outliers:
  - Trend with only the Fringe Outliers:
    - $y = 0.77x + 2.08$
    - $R^2 = 0.56$
    - $n = 31$

- Trend with all Outliers:
  - $y = 0.81x + 2.57$
  - $R^2 = 0.52$
  - $n = 33$

- Trend without any Outliers:
  - $y = 0.93x + 1.52$
  - $R^2 = 0.84$
  - $n = 28$
EDA algorithms such as principal component analysis (PCA) are designed to reduce large complex data sets into a series of optimized and interpretable size.
The goal of *chemometric regression analysis* is to develop a model which correlates the information in the set of known measurements to the desired property.

---

Godinho et al., *Talanta* 2014, 129, 143
• Chemometric algorithms for performing regression include *partial least squares (PLS)* and principal component regressions (PCR).

• Chemometric regression is extensively used in making decisions relating to product quality in the on-line monitoring and *process control industry* where fast and expensive systems are needed to test.
Predicting soil texture using image analysis

Pedro Augusto de Oliveira Morais\textsuperscript{a}, Diego Mendes de Souza\textsuperscript{b}, Márcia Thaís de Melo Carvalho\textsuperscript{b}, Beata Emoke Madari\textsuperscript{b}, Anselmo Elcana de Oliveira\textsuperscript{a,\textdagger}

\textsuperscript{a} Laboratory of Theoretical and Computational Chemistry, Instituto de Química, UFG, Goiânia, GO 74690-900, Brazil

\textsuperscript{b} Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO 75375-000, Brazil

\textdagger} Corresponding author.

\textbf{Fig. 6.} Image-based method for clay and sand content estimates.
Fig. 8. Plot of the regression between PLS2-MIA predicted and pipette measured particle contents for the 18 samples of the validation set. The best regression model was built using fused data containing RGB, HSV, and Grayscale color systems (see Tables 1 and 2).
Classification Model

It is used to *predict a sample's class* by comparing the sample to a previously analyzed experience set, in which categories are already known.
Treatment
- Chemotherapy
- Radiotherapy
- Chemotherapy and Radiotherapy
- None

Cancer group: Red number
Healthy group: Blue number

Samples
- Carcinoma
- Leukemia
- Lymphoma
- Control

* Metastatic Cancer
**k-nearest neighbour (k-NN)** is primarily used in Chemometrics

- This can be thought as separating chromatographic data set from spectroscopic data set and doing analysis.
When these techniques are used to create a classification model, the answers provided are more reliable and include the ability to *reveal unusual samples* in the data.

- Therefore, Chemometrics helps in *standardizing* data.
Earwax: an innovative tool for assessment of tobacco use or exposure. A pilot study in young adults

Engy Shokry\textsuperscript{1} \cdot Anselmo Elcana de Oliveira\textsuperscript{2} \cdot Melissa Ameloti Gomes Avelino\textsuperscript{3} \cdot Mariana Moreira de Deus\textsuperscript{3} \cdot Naiara Zedes Pereira\textsuperscript{2} \cdot Nelson Roberto Antoniosi Filho\textsuperscript{1}

\begin{table}[h]
\centering
\caption{Characteristics of subjects enrolled in the study including non-smokers, passive and active smokers}
\begin{tabular}{lccc}
\hline
Characteristic & Non-smoker & Passive smoker & Active smoker \\
\hline
Number of subjects & 16 & 20 & 25 \\
Age range (years) & 18–35 & 18–25 & 18–35 \\
Sex & & & \\
   Males & 9 & 11 & 17 \\
   Females & 7 & 9 & 8 \\
Ethnicity & & & \\
   Hispanic/Latino & 16 & 20 & 25 \\
   Not Hispanic/Latino & 0 & 0 & 0 \\
Self-reported average number of cigarettes smoked/day & 0 & 0 & \geq7 \\
\hline
\end{tabular}
\end{table}
Fig. 4 Regression tree for predicting active smokers (Active S), non-smokers (Non S), and passive smokers (Passive S) based on concentration profiles of cotinine and anabasine in ear wax samples of 61 adults.
**Fig. 5** Support vector machine discrimination model based on concentration profiles of nicotine and its related compounds in earwax samples of three study groups: non-smokers (Non S) (*red filled diamond*), passive (Passive S) (*green filled square*), and active smokers (Active S) (*blue filled triangle*)
The Analytical Process

Tools: Exploratory data analysis
Data mining

Measurements:
- Voltages
- Currents
- Volumes...

Samples:

System:

Fonte: M.A. Sharaf; D.L. Illman; B.R. Kowalski, Chemical Analysis: Chemometrics
Challenges with Data Analytics

Aggregating data from multiple sources

Cleaning data

Choosing a model

Moving to production
automated analysis $\rightarrow$ large amount of data

chromatography and spectroscopy methods

one sample/analyte at a time $ightarrow$
many samples/analytes

many variables are measured on each run

Multivariate Data
Univariate × Multivariate Data

• Control of product humidity during a month
• Lunch time of an employee
• Calibration curve: peak intensity × analyte concentration
• Datasheet including all quality control data
• Historical information regarding employees’ productivity
• Multivariate calibration: spectrum × analyte concentration
covariance
• Artificial Intelligence
• Machine learning
• 2D and 3D pattern recognition:
  – human being x computer?
  – pattern recognition on measurement data tables/sheets containing lots of numbers and a huge amount of sample information.
• Handwriting recognition
Optical Character Recognition (OCR)
• Speech recognition
• **Speaker recognition (voice biometrics)**
- **Personal ID**
  - Fingerprint
  - Iris scanning
  - Walking
• Radars

An airborne image of an A-3 flight prior to automated motion compensation, image centering, and overlay fitting (a) and the image after automated processing (b)
• Electrocardiogram

NORMAL RATE AND RHYTHM

ABNORMALITIES IN RHYTHM

Extrasystole (premature beat)

Tachycardia

Ventricular fibrillation

CARDIAC MYOPATHIES

Myocardial infarction

Complete heart block
- Weather Forecasting
• **Stock Market**
• 217 males
diagnosed with bipolar disorder, major depressive disorder, schizophrenia and other psychiatric issues.
about 20 percent went from no suicidal thoughts to a high level of suicidal thoughts while they were being seen at a clinic at the university.
blood samples
RNA biomarkers that appeared to predict suicidal thinking.
it's unclear how well the biomarkers would work in the larger population due to the fact that the study was limited to high-risk males with psychiatric diagnoses, but that the app is ready to be deployed and tested on a wider group in real-world settings such as emergency rooms.
• **Diabetes (earwax)**

![Graph](image)

**Fig. 2.** Robust PCA scores plot from 110 VOCs analyzed by HS-GC/MS: a) healthy volunteers (◇); b) diabetic patients with insulin dependent diabetes mellitus (type 1) (▲); and c) patients with non-insulin dependent diabetes mellitus (type 2) (■).

Fig. 7. Decision tree plot using only methoxyacetone in discrimination between the healthy volunteers, patients with non-insulin dependent diabetes mellitus (type 2); and diabetic patients with insulin dependent diabetes mellitus (type 1).
Preprocessing techniques are designed to transform the data into the most informative representation in the context of the goal study.
FT-NIR spectra of 268 mineral oil samples:
(A) Raw data;
(B) Base-line correction;
(C) Multiple scattering correction;
(D) Savitzky-Golay filter.
Preprocessing

Smoothing → Baseline correction → Peak Picking
Unsupervised learning refers to methods that make no a priori assumptions about category-membership of the samples, but rather assist the analyst in uncovering intrinsic clusters or other patterns in the data.
In *supervised/machine learning* the computer “learns” to optimally classify the samples based on advance knowledge about their category membership.

- Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations.
- New data is classified based on the model built on training set known categories.

![Diagram of classification](image)
clustering
• Medical Big Data

Diagnosis procedure in CAD:
- Pre-processing
- Feature extraction and then
- Pattern recognition and classification

Automatic Disease diagnosis

Treatments, Therapies or Rehabilitations
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."