Selected Aspects of 40 Years Applied Chemometrics

VARMUZA Kurt
Vienna University of Technology, Austria
Institute of Statistics and Mathematical Methods in Economics
Laboratory for ChemoMetrics
www.lcm.tuwien.ac.at

Autumn School of Chemoinformatics
25 - 26 Nov 2015, Tokyo, Japan
26 Nov 2015, The University of Tokyo
Contents of Tutorial

1 Basics (history, strategies)
2 Empirical multivariate models (optimum complexity, evaluation)
3 One class classification

With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)

Supported by Austrian Science Fund (Project P26871-N20).
Collaboration with Peter Filzmoser, Irene Hoffmann, et al., and COSIMA team acknowledged.

This is an adjusted version of the lecture for presentation in web.
Rosetta mission (ESA) to COMET 67P/Churyumov-Gerasimenko - Chury

COSIMA

Launch 2 March 2004
Arrival 6 Aug 2014
Landing 12 Nov 2014

12 Aug 2015, ca perihelion (186.10^6 km from sun); 330 km from comet; OSIRIS camera; animation, 17 pictures (ca 21 hours, incl. big outburst at 17:35 GMT).

http://www.esa.int/spaceinimages/Images/2015/08/Approaching_perihelion_Animation
Where are Rosetta and the comet?
http://www.esa.int/Our_Activities/Space_Science/Rosetta

26 Nov 2015
COSIMA: TOF-SIMS instrument

primary ions $^{115}\text{In}^+$, 8 kV, 5 ns pulses

secondary ions

mass spectrum

collected by COSIMA, 25-31 Oct 2014; 10-20 km from comet (3.1 AU from sun); 1-10 m/s impact speed; a shattered dust particle with height ca 60 µm.

Comet Material

... most **pristine** material in our solar system in the form of ice, mixed with dust, **silicates**, and **refractory organic** material (probably many different species) ...

... aggregate of **pre-solar grains** (grains that existed prior to the formation of the Solar System), ...

... comet material (water/organic/inorganic) may have been the **seed of life on earth** ...

Contents of Tutorial

1 Basics (history, strategies)
2 Empirical multivariate models (optimum complexity, evaluation)
3 One class classification

With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)
Chemometrics

- Uses methods from statistics, mathematics, and informatics,
- to extract relevant information from chemical/physical data,
- and to select or optimize chemical processes and experiments.

Perhaps a part of Cheminformatics

Mostly using multivariate data
History

Kowalski

Svante Wold

Lappeenranta, 2007

Bruce Kowalski
(1942 - 2012)
Loen, 2011

40 years
Multivariate data analysis

UNSUPERVISED

Exploratory data analysis, Cluster analysis
Search for similar object or similar variables

PCA
HCA

objects

variables

1

m

Typical: collinear variables, and often $m > n$

SUPERVISED

Calibration, Classification
Mathematical models for $y = f(X)$, prediction!

PLS

KNN

LDA

property or class

1
- **Exploratory data analysis**
- **Multivariate calibration**
- **Multivariate classification**

... *data-driven* ...

... *empirical models* !
Multivariate data analysis

Some aspects for empirical models (in chemometrics)

- More variables than objects \((m > n)\)
- Multicollinearity
- Parsimonious (interpretation, understanding)
- Tested (for new cases, domain, performance)
- Robust (data distribution, outliers)

Trial and error ...
Contents of Tutorial

1 Basics (history, strategies)
2 Empirical multivariate models (optimum complexity, evaluation)
3 One class classification

With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)
Multivariate data analysis

Optimum model complexity

Optimum Model

Underfitting
model too simple

Overfitting
model too much adapted to calibration data

Error
mean squared error
\[ MSE = \frac{1}{n} \sum e_i^2 \]

\textit{MSEC,} calibration
\textit{MSEP,} prediction

Complexity of model
No. of PCA/PLS-components, no. of variables, non-linearities
Multivariate data analysis

Optimization and Evaluation (separated)

- Variables $X$ and $Y$ with $m$ objects and $n$ variables.
- Split into Calibration set and Test set.
- Model with optimized complexity.
- Performance for new cases, e.g., by repeated (random) double cross validation (rdCV).

Optimization of performance within calibration set; cross validation, bootstrap.
Estimation of model performance: CALIBRATION

\[ y_i \] reference ("true") value for object \( i \)
\[ \hat{y}_i \] calculated (predicted) value (test set !)
\[ e_i = y_i - \hat{y}_i \] prediction error for object (residual)
\[ i = 1 \ldots z \] \( z \) is the number of predictions

Specify:
\[ \triangleleft \] which data set (calibration set, test set)
\[ \triangleleft \] which strategy (cross validation, ...)

Distribution of prediction errors

\[ \text{bias} = \text{mean of prediction errors } e_i \]
\[ \text{SEP} = \text{standard deviation of prediction errors } e_i \]
\[ = \text{Standard Error of Prediction} \]
\[ \text{CI} = \text{confidence interval, } \text{CI}_{95\%} \approx \pm 2\times \text{SEP} \]

User friendly! All in units of \( y \)!
SEP (or any other performance criterion) must NOT be considered as a single number.

Depends on
- the used objects, and variables;
- the random split in a CV (or a bootstrap);
  repetitions highly recommended, e. g., rdCV.

It is an estimation.
It has a distribution (variation) - boxplots recommended.
Estimation of model performance: CALIBRATION

\( X \ n = 166 \) fermentation samples (cereals), centrifuged
\( m = 235 \) NIR absorbances, 1115 - 2285 nm (step 5 nm), 1\(^{\text{st}}\) deriv., (7 points, 2\(^{\text{nd}}\) order)

\( Y \) ethanol content, reference method HPLC; 21.7 - 88.1 g/L

PLS regression,
\([\text{ethanol}] = f(\text{NIR abs.})\)

dCV
30 repetitions,
4 and 7 segments

Comparison of variable selection methods.

### Class assignment table (binary classification)

<table>
<thead>
<tr>
<th></th>
<th>assigned class</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>true class 1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>true class 2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td>sum</td>
<td>$n_{\rightarrow 1}$</td>
<td>$n_{\rightarrow 2}$</td>
</tr>
</tbody>
</table>

### Predictive ability

- **class 1**: $P_1 = \frac{n_{11}}{n_1}$
- **class 2**: $P_2 = \frac{n_{22}}{n_2}$

### Average predictive ability

$$P = \frac{(P_1 + P_2)}{2}$$

### Avoid: Overall predictive ability

$$= \frac{(n_{11} + n_{22})}{n}$$
Extraterrestrial Material

- Collected in space and brought safely to Earth (Stardust [near comet], Hayabusa [asteroid surface])
- Measurements in space (Rosetta, Mars, Moon, ...)
- Coming autonomously (meteorites, ca 40,000 t/year)

*Finds* and *Falls* (witnessed, observed, samples)
Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

- ordinary chondrite
- carbonaceous chondrite
- Martian (from Mars surface)

Samples from Natural History Museum (NHM) Vienna: 10 meteorites
Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

- ordinary chondrite
- carbonaceous chondrite
- Martian (from Mars surface)

110 – 660 TOF-SIMS spectra per meteorite class
+ 280 spectra from substrate (Au)

\( n = 3372 \) spectra (objects)

\( m = 299 \) variables (peak heights at \( m/z \) 1 – 300, excl. 115; appr. only inorganic ions, \( m/\Delta m = 1200 \)), sum 100

10 + 1 = 11 classes

Samples from Natural History Museum (NHM) Vienna: 10 meteorites
KNN classification,
Euclidean distance,
rdCV strategy
(20 repetitions, 2 and 5 segments),
optimum no. of neighbors = 1

Predictive abilities
(mean of 20 repetitions) per meteorite class:
90 – 97 %
Total mean: 94 %

Class assignment frequencies

True class
Tissint
Tieschitz
Tamdakht
Substrate
Renazzo
Pultusk
Ochansk
Murchison
Mocs
Lance
Allende

Assigned class
Multivariate data analysis - EXAMPLE

Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

KNN classification, Euclidean distance, \textit{rdCV} strategy (20 repetitions, 2 and 5 segments), optimum no. of neighbors = 1

\textbf{Predictive abilities} (mean of 20 repetitions) per meteorite class: 90 – 97 %

Total mean: 94 %
Multivariate data analysis - EXAMPLE

Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

PCA

$X \ (3372 \times 299)$, row sum = 100

Classes

11 Tissint
10 Tieschitz
9 Tamdakht
8 Substrate
7 Renazzo
6 Pultusk
5 Ochansk
4 Murchison
3 Mocs
2 Lance
1 Allende
Contents of Tutorial

1 Basics (history, strategies)
2 Empirical multivariate models (optimum complexity, evaluation)
3 One class classification

With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)
**TOF-SIMS on Meteorite Grains**

**Meteorite grains prepared on a gold foil (10 mm x 10 mm)**

**Samples**
Christian Köberl, Franz Brandstätter, Ludovic Ferrière, Natural History Museum Vienna

**Preparation**
Cécile Engrand, Univ. Paris Sud (Orsay)

**TOF-SIMS (COSIMA twin)**
Martin Hilchenbach, Max Planck Institute for Solar System Research, Göttingen

**Ordinary Chondrites**
Tieschitz, Ochansk

**Carbonaceous Chondrite**
Renazzo

**Martian meteorite (Shergottite)**
Tissint
TOF-SIMS on Meteorite Grains

Photographic picture of a target with a meteorite grain.

TOF-SIMS measuring positions (155 query spectra)

63 background (Off grain) spectra

1000 μm
On/Off methods

Multi-class classification

Binary classification

One-class classification

Only one target class, all others (outlier)
On/Off methods

**Multi-class classification**

**Binary classification**

**One-class classification**

SIMCA (S. Wold): PCA model of each class

Only one *target* class, all others (outlier)
Recognition of potentially relevant spectra (TOF-SIMS)

Target class spectra
(background, off-grain)

Query spectra
off-grain (background)
(or not (potentially relevant, on-grain))
On-grain  –  Off-grain

Recognition of potentially relevant spectra (TOF-SIMS)

- Univariate; intensity of a selected ion (element, e.g., Fe, ...)

- Ratios of variables (or other 'simple' heuristic combinations)

- **One-class classification** (*target* class = off-grain)  
  1. PCA: orthogonal and score distance
  2. KNN distance distribution

- Weights from sparse and robust PLS-DA

- Cluster analysis

- Deconvolution

- NMF (nonnegative matrix factorization)
Recognition of potentially relevant spectra (TOF-SIMS)
One-class classification
Distances to PCA model made from Off-grain spectra

Demo scheme

Target class: $X_0$, $m = 3$ variables;

- PCA model with $A = 2$ components (scores $t_1$ and $t_2$);
- Projection of $X_0$-points into the PCA model (plane, defined by $t_1$ and $t_2$);
- Query points 1, 2, 3 in x-space
- Projections of query points into the PCA plane

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

Distances to PCA model made from Off-grain spectra

Demo scheme

**Target class:** $X_0$, $m = 3$ variables;

PCA model with $A = 2$ components (scores $t_1$ and $t_2$);

- Projection of $X_0$-points into the PCA model (plane, defined by $t_1$ and $t_2$)
- Query points 1, 2, 3 in x-space
- Projections of query points into the PCA plane

**Score distance (SD)**

$= \text{Mahalanobis distance from center, measured in the PCA space (plane).}$

Describes the distance to the center (of the background spectra) in PCA score space, considering the covariance structure of the $x$-variables.

Recognition of potentially relevant spectra (TOF-SIMS)
One-class classification
Distances to PCA model made from *Off-grain spectra*

Demo scheme

**Target class:** $X_0$, $m = 3$ variables;

PCA model with $A = 2$ components (scores $t_1$ and $t_2$);

- Projection of $X_0$-points into the PCA model (plane, defined by $t_1$ and $t_2$)
- Query points 1, 2, 3 in x-space
- Projections of query points into the PCA plane

**Orthogonal distance (OD)**
= Distance in x-space between point and its projection onto the PCA space.

Describes information loss by projecting into the $A$-dimensional PCA score space.

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

Distances to PCA model made from *Off-grain spectra*

---

**Demo scheme**

*Target class: \( X_0 \), \( m = 3 \) variables;*

PCA model with \( A = 2 \) components (scores \( t_1 \) and \( t_2 \)):

- Projection of \( X_0 \)-points into the PCA model (plane, defined by \( t_1 \) and \( t_2 \))
- Query points 1, 2, 3 in x-space
- Projections of query points into the PCA plane

---

**Large OD** - AND/OR large SD - indicate an *outlier;*

- a spectrum not belonging to the background group,
- a potentially relevant relevant spectrum.

---

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

Mean KNN distances within the Off-grain spectra

Target class (off-grain);

$n$ objects

E. g., consider $k = 3$ neighbors

$(k$ to be optimized)$

(1) mean distance to $k$ nearest neighbors of object $i$

(2) For all objects of target class, $i = 1 \ldots n$

(3) Distribution of the mean distances

(4) Cutoff value (quantile 0.99)
COSIMA: Au target with collected comet particles (grains)

Target 2D0 (Au black), [1 cm x 1 cm], collection Aug – Dec 2014 (94 days), ca 3 AU from sun; 50-100 km from comet. [ Yves Langevin (Paris, Orsay) ]

Grain DONIA (ca 200 µm diameter), collected 18-24 Aug 2014.

Original pictures not shown here

200 µm

4 x 3 TOF-SIMS spectra scanned (7 Sep 2014, 12:14 - 13:22)
Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

OD (Orthogonal Distance), KNN mean distance

OD

KNN \((k = 10)\)

Off-grain (target class), \(n_1 = 59\); Query spectra, \(n_2 = 96\) (plot); \(m = 3437\) variables (inorganic ions, sum 100)
Off-grain spectra (background)

On-grain spectra comet/meteorite

Chemist: Univariate (interpretation)

Maybe differences only multivariate

Inorganics: Na, Mg, Al, Si, Ca, Mn, Fe. silicates (olivine, pyroxene?), Fe-sulfide?

Organics: C, N, O, (S, Cl, F). No clear signature for specific organics; perhaps macromolecular with too low secondary ion yield, …
**Rosetta: Mass Spectra of Gas Phase**

**GC-MS instrument ROSINA (Orbiter)**

- $\text{N}_2$, $\text{O}_2$
- $\text{H}_2\text{O}$
- $\text{CO}$, $\text{CO}_2$, $\text{CH}_4$
- $\text{CH}_3\text{OH}$, $\text{CH}_2\text{O}$
- $\text{NH}_3$, $\text{HCN}$
- $\text{H}_2\text{S}$, $\text{CS}_2$, $\text{SO}_2$

$m/\Delta m$ 9000 (50% peak height); $m/z$ 12-150; double-focusing magnet ms.

Altwegg K. et al.: Nature 2014, 2015; University Bern
Rosetta: Mass Spectra of Gas Phase

GC-MS instrument COSAC (Philae Lander)

Fig. 1. Mass spectra taken by COSAC in “sniffing mode.” Top (green): spectrum taken 25 min after first touchdown; the m/z 18 peak reached a height of 330 counts, but the spectrum is truncated to show smaller peaks more clearly; middle (red): final spectrum, taken 2 days later at the current Philae position; bottom (blue): first spectrum, obtained in orbit 27 days before landing, from a distance of 10 km.

Goessmann F. et al.: Science, 349, issue 6247 (2015); MPS Göttingen
Rosetta: Mass Spectra of Gas Phase

GC-MS instrument COSAC (Philae Lander)

Table 1. The 16 molecules used to fit the COSAC mass spectrum.

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Molar mass (u)</th>
<th>MS fraction</th>
<th>Relative to water</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>H$_2$O</td>
<td>18</td>
<td>80.92</td>
<td>100</td>
</tr>
<tr>
<td>Methane</td>
<td>CH$_4$</td>
<td>16</td>
<td>0.70</td>
<td>0.5</td>
</tr>
<tr>
<td>Methanenitrile (hydrogen cyanide)</td>
<td>HCN</td>
<td>27</td>
<td>1.06</td>
<td>0.9</td>
</tr>
<tr>
<td>Carbon monoxide</td>
<td>CO</td>
<td>28</td>
<td>1.09</td>
<td>1.2</td>
</tr>
<tr>
<td>Methylamine</td>
<td>CH$_3$NH$_2$</td>
<td>31</td>
<td>1.19</td>
<td>0.6</td>
</tr>
<tr>
<td>Ethanenitrile (acetonitrile)</td>
<td>CH$_3$CN</td>
<td>41</td>
<td>0.55</td>
<td>0.3</td>
</tr>
<tr>
<td>Isocyanic acid</td>
<td>HNCO</td>
<td>43</td>
<td>0.47</td>
<td>0.3</td>
</tr>
<tr>
<td>Ethanal (acetaldehyde)</td>
<td>CH$_3$CHO</td>
<td>44</td>
<td>1.01</td>
<td>0.5</td>
</tr>
<tr>
<td>Methanamide (formamide)</td>
<td>HCONH$_2$</td>
<td>45</td>
<td>3.73</td>
<td>1.8</td>
</tr>
<tr>
<td>Ethylamine</td>
<td>C$_2$H$_5$NH$_2$</td>
<td>45</td>
<td>0.72</td>
<td>0.3</td>
</tr>
<tr>
<td>Isocyanomethane (methyl isocyanate)</td>
<td>CH$_3$NCO</td>
<td>57</td>
<td>3.13</td>
<td>1.3</td>
</tr>
<tr>
<td>Propanone (acetone)</td>
<td>CH$_3$COCH$_3$</td>
<td>58</td>
<td>1.02</td>
<td>0.3</td>
</tr>
<tr>
<td>Propanal (propionaldehyde)</td>
<td>C$_2$H$_5$CHO</td>
<td>58</td>
<td>0.44</td>
<td>0.1</td>
</tr>
<tr>
<td>Ethanamide (acetamide)</td>
<td>CH$_3$CONH$_2$</td>
<td>59</td>
<td>2.20</td>
<td>0.7</td>
</tr>
<tr>
<td>2-Hydroxyethanal (glycolaldehyde)</td>
<td>CH$_2$OHCHO</td>
<td>60</td>
<td>0.98</td>
<td>0.4</td>
</tr>
<tr>
<td>1,2-Ethandiol (ethylene glycol)</td>
<td>CH$_2$(OH)CH$_2$(OH)</td>
<td>62</td>
<td>0.79</td>
<td>0.2</td>
</tr>
</tbody>
</table>

91 isomers (MOLGEN)
201 ion formulae
C$_c$H$_h$N$_n$O$_o$ (m/z 1 – 62), potential fragment ions

Goessmann F. et al.: Science, 349, issue 6247 (2015); MPS Göttingen
Organics in Extraterrestrial Material

Fall 28 Sep 1969
Carbon-rich chondrite, total ca 100 kg, considered to be similar to comet material

High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall

30 mg freshly broken meteorite sample

extraction by methanol, acetonitril, toluene, ...

Fourier transform ion cyclotron resonance mass spectrometer (FTICR-MS); electrospray ionization (ESI), only quasi molecular ions (M+H)^+, (M-H)^- m/Δm ca 10^6

Peak list mass spectra
Suitable for highly complex organic mixtures (one peak per brutto formula)

Organics in Extraterrestrial Material

Identified molecular formulas
mass range 150 – 1000

<table>
<thead>
<tr>
<th>ESI(-), methanol extract</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHO</td>
</tr>
<tr>
<td>CHOS</td>
</tr>
<tr>
<td>CHNO</td>
</tr>
<tr>
<td>CHNOS</td>
</tr>
<tr>
<td>Sum</td>
</tr>
<tr>
<td>Mass peaks</td>
</tr>
</tbody>
</table>

Typical carbon-rich meteorites contain 3 – 5 % C.
70 % macromolecular
30 % soluble

Total estimated >50,000

Millions of different chemical compounds (isomers) with elements C, H, O, N, S, P.
Extraterrestrial chemodiversity very high

Selected Aspects of 40 Years Applied Chemometrics

Data in chemistry, ...

Sometimes CHEMOMETRICS helps, but not always.

Everything should be made as simple as possible, but not simpler.

Thank you for your interest

Autumn School of Chemoinformatics
25 - 26 Nov 2015, Tokyo, Japan
26 Nov 2015, The University of Tokyo