One-class classification
for the recognition of relevant measurements
applied to
mass spectra from cometary and meteoritic particles

Varmuza K.*[1], Filzmoser P.[1], Ortner I.[1], Hilchenbach M.[2], Kissel J.[2], Merouane S.[2], Paquette J.[2], Stenzel O.[2], Engrand C.[3], Cottin H.[4], Fray N.[4], Isnard R.[4], Briois C.[5], Thirkell L.[5], Baklouti D.[6], Bardyn A.[7], Siljeström S.[8], Schulz R.[9], Silen J.[10], Brandstätter F.[11], Ferrière L.[11], Koeberl C.[11,12]

[1] TU Wien - Vienna University of Technology (Austria)
Institute of Statistics and Mathematical Methods in Economics, Computational Statistics

kurt.varmuza@tuwien.ac.at | http://www.lcm.tuwien.ac.at/vk/ | https://institute.tuwien.ac.at/cstat/home/EN/

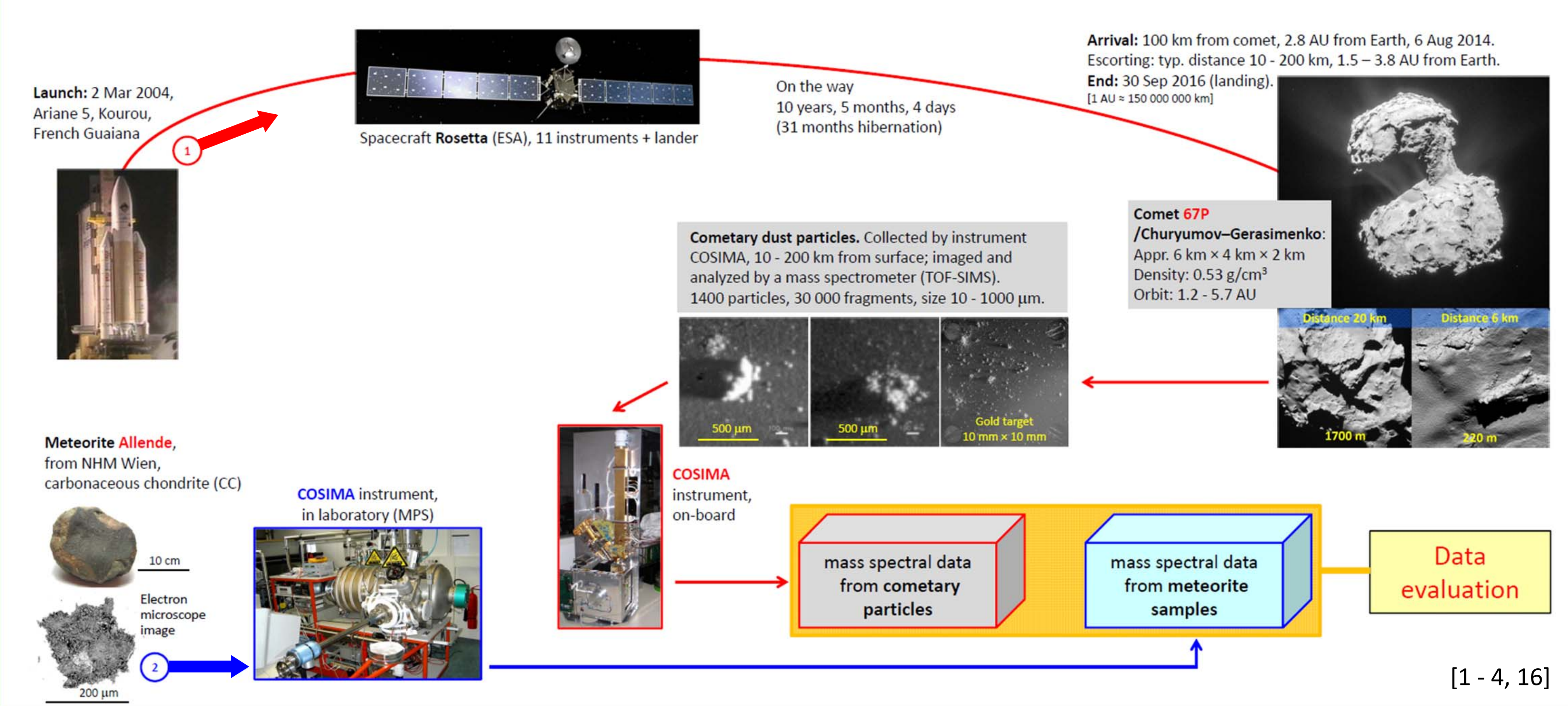
[2]Max Planck Inst. for Solar System Res., Göttingen (Germany); [3]CSNSM, CNRS/Univ. Paris Sud, Univ. Paris Saclay, Orsay (France); [4]Lab. Interuniversitaire des Systèmes Atmosphériques, Univ. Paris Est, Créteil (France); [5]Lab. de Physique et Chimie de l'Environnement et de l'Espace, Univ. d'Orléans (France); [6]IAS, CNRS/Univ. Paris Sud, Univ. Paris Saclay, Orsay (France); [7]Carnegie Institution of Washington, DC (USA); [8]Bioscience and Materials / Chemistry and Materials, Res. Inst. of Sweden, Stockholm (Sweden); [9]European Space Agency, Noordwijk (The Netherlands); [10]Finnish Meteorological Inst., Helsinki (Finland); [11]Natural History Museum (NHM), Vienna (Austria); [12]Dept. of Lithospheric Res., University of Vienna (Austria)

Poster at SSC16 – 16th Scandinavian Symposium on Chemometrics
June 17-20, 2019, Nesbru - Holmen / Oslo, Norway | https://ssc16.org/

title

(1) Motivation / Rosetta project / COSIMA instrument

Data from deep space (comet), and laboratory (meteorites) → Multivariate statistics Chemoinformatics → Information about chemical composition of samples



[1 - 4, 16]

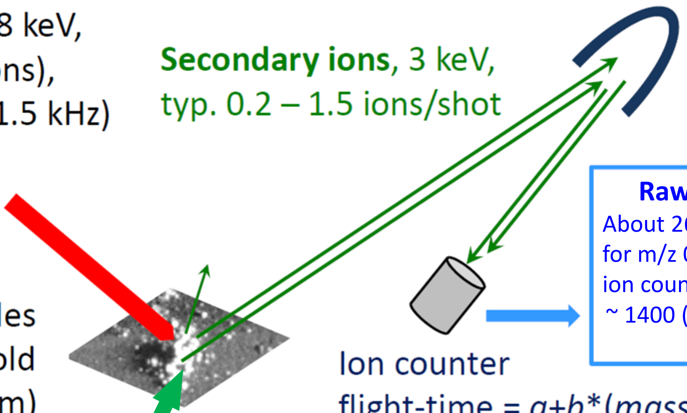
(2) Selection of potentially relevant spectra

measured on cometary particles or meteorite grains

COSIMA: TOF-SIMS / Time-Of-Flight Secondary Ion Mass Spectrometry

Primary ions, $^{115}In^{+}$, 8 keV,
3 ns shots (à 1000 ions),
typ. 225,000 shots (1.5 kHz)
per spectrum,
spot 30 µm x 50 µm

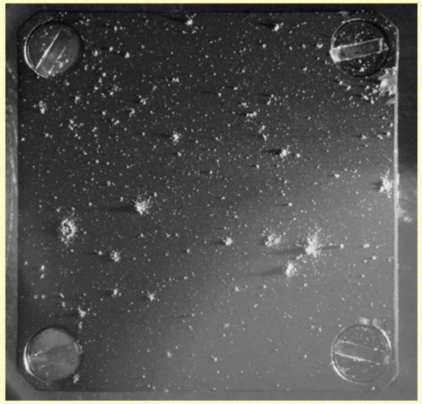
Secondary ions, 3 keV,
typ. 0.2 – 1.5 ions/shot



Raw mass spectrum
About 26,300 time (mass) bins for m/z 0 – 300 Dalton with ion counts. Mass resolution ~ 1400 (half peak height).

Cometary particles collected on a gold plate (1 cm x 1 cm)

Ion counter
flight-time = $a+b*(mass/z)^{0.5}$ [1]



Gold target (1 cm x 1 cm) with collected cometary particles.
Dec 2014 – Feb 2015, 20 – 140 km from comet (*COSIMA target 2CF*).

The position of the primary ion beam (~ 30 µm × 50 µm wide) has uncertainties up to ±70 µm. Therefore, an **evaluation of the spectra's origin** is necessary: From **background** (Au target material) or cometary **particle** (10 - 1000 µm size) ?

Strategies

- Ratios of selected ion counts, e.g., $C^{+}/CH_3^{+} > 1$
- **Multivariate methods are used here**

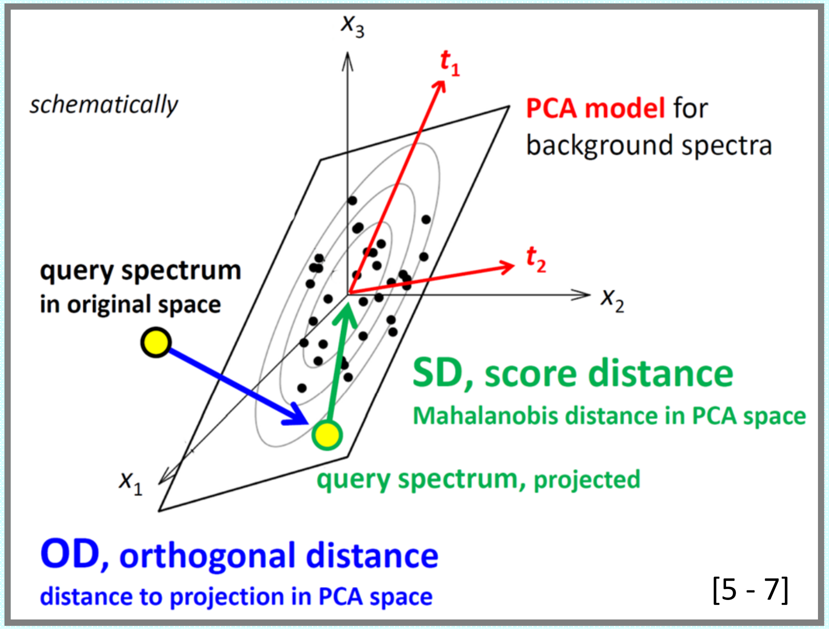
One-class classification

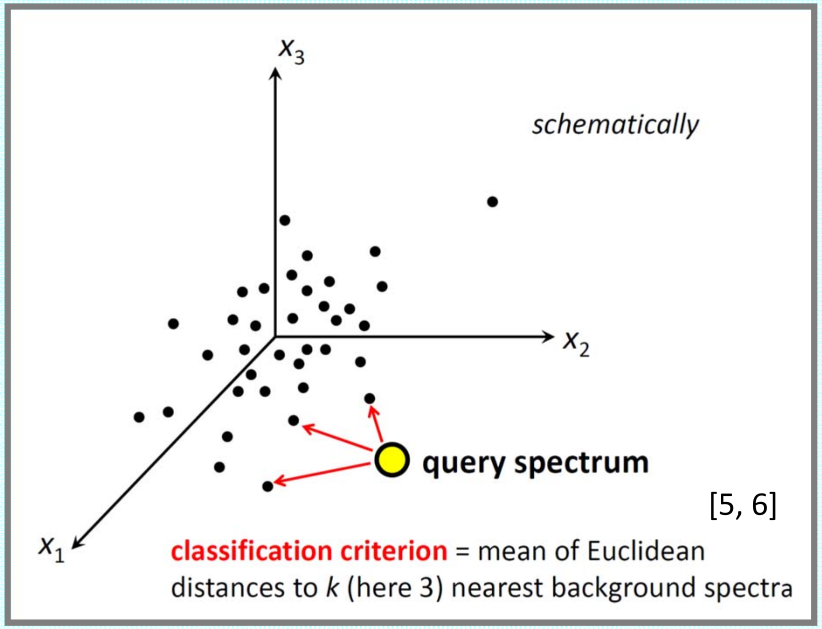
- ☐ Target class = background spectra
- ☐ Combination of
  - PCA approach (distances of query spectrum to PCA model)
  - KNN approach (mean distance of query spectrum to *k* background spectra)

SIMS data

(3) One-class classification

PCA approach combined with KNN approach





Classification

A query spectrum is NOT assigned to the background class, that means is considered potentially relevant if

- $OD > OD_{CUT}$ *. AND .* $SD > SD_{CUT}$
- *. AND .*
- mean KNN distance > KNN_{CUT}

CUToff values are typically 0.90 quantiles of empirical distributions + *safety addition*

(4) Data and Methods

Data

Variables. $m = 9$ mass spectral peak heights (ion counts) for C^+, CH^+, CH_2^+, CH_3^+, Mg^+, Al^+, K^+, Ca^+, Fe^+ (most abundant isotopes); for organics and inorganics.

Objects. $n = 1152$ spectra
55 from background for comet data (space),
121 from background for meteorite data (laboratory),
275 from 3 cometary particles (or neighborhood),
701 from 3 meteorites (Allende, Lancé, Murchison)

Preprocessing

Transformation (scaling). Because of the compositional data type (relative ion abundances are relevant) the **centered log-ratio** transformation (**clr**) has been applied (for PCA and KNN) [8].

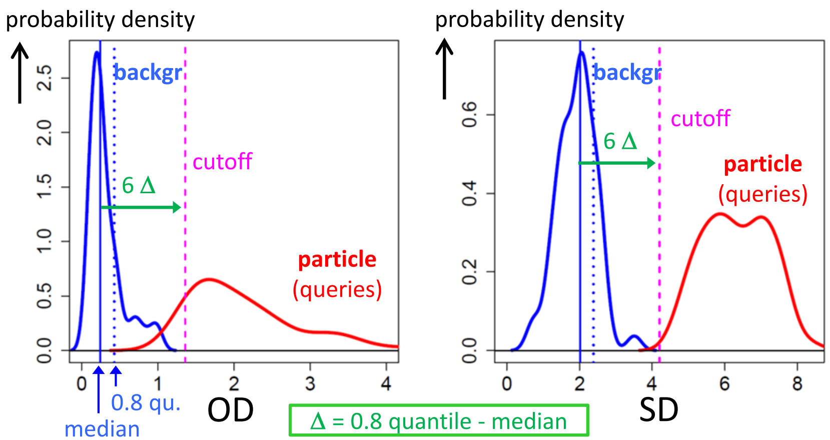
$$_{\text{CLR}}x_j = \ln[x_j / G(\boldsymbol{x})]$$

G, geometric mean of $x_1 \ldots x_m$; $j = 1 \ldots m$

PCA. Robust [9], minimum 90% variance preserved (typically 4 components).

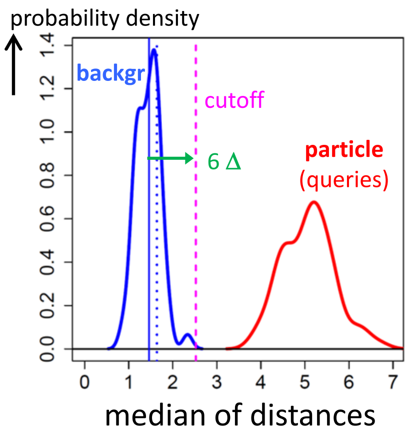
PCA approach (Example)

Distributions of OD (left) and SD (right) for background spectra (**blue**, 55 spectra) and spectra on/near the cometary particle *Kerttu* (**red**, 68 query spectra). Query spectra with distances > cutoff are considered as relevant (63 selected).



KNN approach (Example)

Distributions of median distances from query spectra to $k = 8$ nearest background spectra (for inscriptions and colors see left). Query spectra with median distances > cutoff are considered as relevant (all 68 selected).



Considering $k = 8$ nearest neighbors is a compromise between

- overfitting (instability) with a too small k, and
- underfitting (the bulk of 55 background spectra is taken) with a too big k.

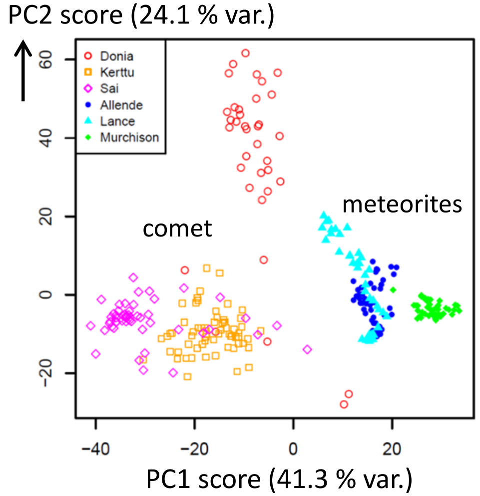
(5) Results

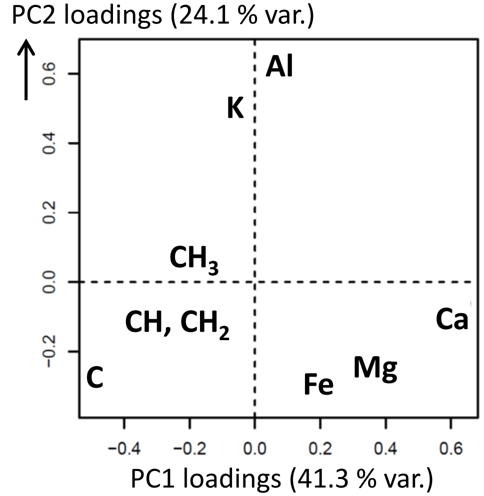
Selection of potentially relevant spectra

by 1-class classification with OD, SD and KNN

Sample particle class		Number of spectra (objects) Used	Selected by OD&SD	KNN	OD&SD & KNN
Comet	*Donia*	147	36	87	36
Comet	*Kerttu*	68	63	68	63
Comet	*Sai*	60	52	60	52
Meteorite	*Allende*	447	212	301	212
Meteorite	*Lancé*	121	105	116	105
Meteorite	*Murchison*	133	123	130	123
Sum		976	591	762	591

PCA of selected spectra



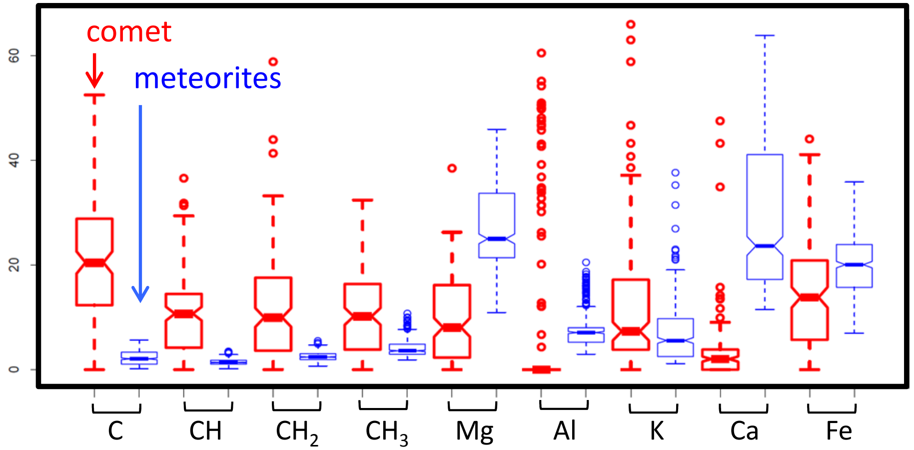


n = 301 spectra (50 randomly selected from each meteorite, 151 comet spectra) for better balanced data set.
m = 9 variables, sum 100 normalized for better interpretation; PDMS subtracted.

- **Carbon-containing ions prominent in comet data.**
- **Comet data more diverse than meteorite data.**

Comparison of comet and meteorite data

Distribution of sum-100 normalized ion counts (univariate)



All *n* = 591 selected spectra used.
Contamination of PDMS (polydimethylsiloxane) subtracted.
Normalized to sum 100 of *m* = 9 variables.

- **Comet material contains more carbon (based on CH_{0-3}^+ ions) than the considered meteorites (which are C-rich meteorites, so called *carbonaceous chondrites*).**
- **Ca^+ and Mg^+ are more prominent in meteorites than in comet.**

(6) Summary

One-class classification

based on orthogonal & score distances and a *k*-nearest neighbor approach

- Data from background (target) define the "one-class".
- Minimum assumptions; concepts from robust statistics and compositional data processing.
- Cutoff criteria solely derived from the "one-class data".
- Stable and reliable results with *difficult* TOF-SIMS data from space and with laboratory data.

Cometary/meteoritic material

TOF-SIMS data from space and lab, including results from the ***COSIMA team***

- Cometary particles appear diverse and different from CC meteorites (carbonaceous chondrites) [10].
- More (organic) carbon in comet than in CC meteorites.
- Organics: macromolecular [11].
- Ions $C_3H_{0\text{-}4}^+$, C_4^+, etc. indicate unsaturated organic compounds in cometary particles [12].
- Atomic ratios from SIMS data:
  C/Si ~ 5 [13] C/N ~ 30 [14]
  C/H ~ 1 [15]

summary

References

[1] Kissel J., et al.: *Space Sci. Rev*., **128**, 823 (2007)
[2] Langevin Y., et al.: *Icarus*, **271,** 76 (2016)
[3] Hornung K., et al.: *Planetary and Space Science*, **133**, 63 (2016)
[4] Hilchenbach M., et al.: *The Astrophysical Journal Letters*, **816**: L32 (2016)
[5] Brereton R. G.: Chemometrics for pattern recognition, Wiley, Chichester, UK (2009)
[6] Xu Y., Brereton R. G.: *J. Chem. Inf. Model.*, **45**, 1392 (2005)
[7] Pomerantsev A. L.: *J. Chemometrics*, **22**, 601 (2008)
[8] Filzmoser P., Hron K., Templ M.: Applied compositional data analysis, Springer Nature, Cham, Switzerland (2018)
[9] Hubert M., et al.: *Technometrics*, **47**, 64 (2005)
[10] Stenzel O., et al.: *MNRAS*, **469,** Suppl_2, S492 (2017)
[11] Fray N., et al.*: Nature*, **528**, 72 (2016)
[12] Varmuza K., et al.: *J. Chemometrics,* **32***,* e3001, 1-13 (2018)
[13] Bardyn A., et al.: *MNRAS,* **469,** Suppl_2**,** S712-S722 (2017)
[14] Fray N., et al.: *MNRAS,* **469,** S506-S516 (2017)
[15] Isnard R., et al.: *Astronomy & Astrophysics*, no. aa34797-18 (2019)
[16] http://www.esa.int/spaceinimages/Missions/Rosetta/

This work was supported by the Austrian Science Fund (FWF), project P 26871 - N20.

Acknowledgments. COSIMA was built by a consortium led by the Max-Planck-Institut für Extraterrestrische Physik, Garching, Germany, in collaboration with the Laboratoire de Physique et Chimie de l'Environnement et de l'Espace, Orléans, France, the Institut d'Astrophysique Spatiale, CNRS/Université Paris Sud, Orsay, France, the Finnish Meteorological Institute, Helsinki, Finland, the Universität Wuppertal, Wuppertal, Germany, von Hoerner und Sulger GmbH, Schwetzingen, Germany, the Universität der Bundeswehr, Neubiberg, Germany, the Institut für Physik, Forschungszentrum Seibersdorf, Seibersdorf, Austria, the Institut für Weltraumforschung, Österreichische Akademie der Wissenschaften, Graz, Austria and is led by the Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany. The support of the national funding agencies of Germany (DLR, grant 50QP1302), France (CNES), Austria, Finland and the ESA Technical Directorate is gratefully acknowledged. The authors thank the other members of the COSIMA team for their contributions.

One-class classification for the recognition of relevant measurements - applied to mass spectra from cometary and meteoritic particles

Varmuza K.[1], **Filzmoser P.**[1], **Ortner I.**[1], **Hilchenbach M.**[2], **Kissel J.**[2], **Merouane S.**[2], **Paquette J.**[2], **Stenzel O.**[2], **Engrand C.**[3], **Cottin H.**[4], **Fray N.**[4], **Isnard R.**[4], **Briois C.**[5], **Thirkell L.**[5], **Baklouti D.**[6], **Bardyn A.**[7], **Siljeström S.**[8], **Schulz R.**[9], **Silen J.**[10], **Brandstätter F.**[11], **Ferrière L.**[11], **Koeberl C.**[11,12]

[1] TU Wien - Vienna University of Technology (Austria), Institute of Statistics and Mathematical Methods in Economics (Computational Statistics); kurt.varmuza@tuwien.ac.at

Motivation. The mass spectrometer COSIMA on board of the ESA mission Rosetta to comet Churyumov-Gerasimenko (67P) collected particles (20 - 1000 μm diameter) at distances 10 - 1500 km from the comet and measured TOF-SIMS spectra at the particle surfaces. Because of the special conditions for these remote experiments, it is not trivial to assign the spectra either to particles or to the background (target). An objective classification of the spectra's origin (measuring spot 35 μm x 50 μm with position uncertainties up to 70 μm) has been developed by applying multivariate one-class classification strategies.

Method. The single class (target, background) for one-class classification is described by a set of multivariate objects (spectral data) measured on the target (gold). Two methods for modelling the target class are applied: robust PCA, and KNN. Criteria are defined for characterizing the dissimilarity (δ) between a query object and the target class: for robust PCA the orthogonal and the score distances from the median; for KNN the median of the distances to the k nearest neighbors. The cutoff values of δ for assigning a query object to the target class or not (the later indicates a potentially relevant object) are derived from the distributions of δ for the target objects, based on median, 0.8-quantile and an adjustable parameter (controlling the efficiency of classification). Because of the nature of the data, concepts for compositional data and robust methods have been preferred.

Application. The data used consist of 275 spectra measured on three cometary particles, and 701 spectra measured by a laboratory twin instrument of COSIMA on particles from three meteorites (carbonaceous chondrites, often considered having similar composition as comet material). A set of nine variables is derived from the measured ion counts at masses 12-15 (CH_{0-3}^+), 24 (Mg^+), 27 (Al^+), 39 (K^+), 40 (Ca^+), and 56 (Fe^+) characterizing minerals and presumed organics. Results show distinctive differences between the cometary and the meteoritic samples with considerably more carbon containing material in the comet particles.

Affiliations of coauthors. [2]Max Planck Inst. for Solar System Res., Göttingen (Germany); [3]CSNSM, CNRS/Univ. Paris Sud, Univ. Paris Saclay, Orsay (France); [4]Lab. Interuniversitaire des Systèmes Atmosphériques, Univ. Paris Est, Créteil (France); [5]Lab. de Physique et Chimie de l'Environnement et de l'Espace, Univ. d'Orléans (France); [6]IAS, CNRS/Univ. Paris Sud, Univ. Paris Saclay, Orsay (France); [7]Carnegie Institution of Washington, DC (USA); [8]Bioscience and Materials / Chemistry and Materials, Res. Inst. of Sweden, Stockholm (Sweden); [9]European Space Agency, Noordwijk (The Netherlands); [10]Finnish Meteorological Inst., Helsinki (Finland); [11]Natural History Museum, Vienna (Austria); [12]Dept. of Lithospheric Res., Univ. of Vienna (Austria).

Acknowledgement. Austrian Science Fund (FWF), project P26871-N20.

SSC16

NORWAY - June 17th - 20th 2019

The 16th Scandinavian Symposium on Chemometrics

Book of Abstracts

Contents

Session 5: Chemometrics in action. Chair: Jens Petter Wold

17. Johan Trygg – Perspective on the application of multivariate technologies in biopharmaceutical manufacturing
18. Gerjen H. Tinnevelt – A novel unbiased method links variability of co-expression between multiple proteins on single cells to a clinical phenotype
19. Lars Munck – Natural Computing expressed in irreducible barley spectra reveal the functional composition in diagnostic fingerprints without compression
20. Giorgio Tomasi – Optim2DCOW: an algorithm for automated 2D Correlation Optimized Warping for GC × GC – MS data

Session 6: PhD Projects. Chair: Ingrid Måge

21. Elise A. Kho – Characterization of Haemonchus contortus infections in sheep faeces by infrared spectroscopy
22. Raju Rimal – Simulation of multi-response linear model data and comparison of prediction methods
23. André van den Doel – Is river water out of control?
24. Silje S. Fuglerud – Aqueous glucose sensing by fiber-based near-infrared spectroscopy
25. George Stavropoulos – Data fusion strategies to improve prediction accuracy in Crohn's Disease
26. Anne Bech Risum – Multiway modelling of five-way protein fluorescence data; challenges and new approaches

Session 7: Path modelling, graphical modelling and causality. Chair: Jeroen Jansen

27. Rosaria Romano – University of Calabria, Italy: "Path modeling with multi-block regression method SO-PLS"

Session 8: Method Development. Chair: Age Smilde

28. José Camacho – Cross-Product Penalized Component Analysis: A new tool for Exploratory Data Analysis
29. Lennart Eriksson – Multiblock Orthogonal Component Analysis (MOCA) – A Novel Tool for Data Integration
30. Lars Erik Solberg – Consensus and distinct subspaces for blocks of distances
31. Kristian Hovde Liland – Fast "shortcut calculations" for cross validating Partial Least Squares prediction models
32. Raffaele Vitale – A novel procedure for the simultaneous optimisation of the complexity and significance level of SIMCA models in the presence of strong class overlap

33. Ryan Gosselin – A Novel Dynamic-PLS Algorithm for Meaningful and Robust Models
34. Erik Andries – Calibration Updating Using Unlabeled Secondary Samples

Session 9: Chemometrics in action. Chair: Barry Wise

35. Harald Martens – Big Data Cybernetics: Chemometrics and hybrid modelling for control theory
36. Federico Marini – A general SIMCA framework for single- and multi-block data
37. Chun Kiang Chua – Recent Development of Band-Target Entropy Minimization Algorithm for Hyphenated Techniques
38. Ingunn Berget – Sequential Clusterwise Rotations (SCR); a tool for clustering three-way data
39. Joan Borràs-Ferrís – Defining multivariate raw materials specifications via PLS model inversion
40. Anita Rácz – QSAR behind the curtains: best practices by multi-level comparisons
41. Jose M. González-Martinez – Energy Dispersive X-Ray Hyperspectral Imaging for Homogeneity Studies of Catalyst Extrudates

Posters

1. Lennart Eriksson – An OPLS®-based Multivariate Solver

2. Marian Kraus – Fast standoff investigation of chemical and biological samples using laser induced fluorescence signals, machine learning and an interactive interface

3. Andrei Barcaru – Chasing the interesting in the data with the Supervised Projection Pursuit

4. Ramin Nikzad-Langerodi – Domain Regularization in Partial Least Squares Regression: New Solutions for Old Problems

5. Dillen Augustijn – N-way Data Analysis of Protein Fluorescence in Formulation Screening

6. Kurt Varmuza – One-class classification for the recognition of relevant measurements – applied to mass spectra from cometary and meteoritic particles

7. Magnus Fransson – Applying Convolutional Neural Networks to Vibrational Spectroscopy Data

8. Rola Houhou – PCA – LDA in functional and discrete framework applied to Raman spectra

9. Alba González Cebrián – Dealing with outliers and missing data in PCA model building

10. José Camacho – Comparison of Sparse Principal Component Analysis for Data Interpretation

11. Robert van Vorstenbosch – The Detection of Colorectal Cancer using Exhaled Breath

12. Carl Emil Eskildsen – The cage of covariance: A consequence of regressing high dimensional response variables onto a lower dimensional subspace of explanatory variables

13. Tim Offermans – Improving process control of a dairy processing plant using a soft-sensor on parallel production data streams

14. Morten Arendt Rasmussen – One-Button Chemometrics

15. Agnese Brangule – Use of innovative FTIR spectroscopy sampling methods and chemometrics for authentication and differentiation of herbals

16. Johan Trygg – Data Fusion in metabolomics

17. Johan Trygg – Design of Experiments for data generation and data processing in ‘omics studies (genomics – metabolomics)

18. Johan Trygg – Multivariate patent analysis

19. Carlo G. Bertinetto – Effects of long distance walking analyzed by multidimentional flow cytometry analysis of neutrophils

20. Dávid Bajusz – Similarity metrics for binary data structures in cheminformatics, metabolomics and other fields

21. Veeramani Manokaran – Rapid identification of reaction systems using spectroscopic measurements and micro-reactors

22. Jacob Kræmer Hansen – Novel NIR analysis of Heterogeneous Powder

23. Mona Stefanakis – Infrared spectroscopy and multivariate data analysis for the labelfree early stage diagnosis and demarcation of head and neck cancer in a mouse model

24. Roel Bouman – Process pls: A new path modeling algorithm for high dimensional and multicollinear data

25. Gavin Rhys Lloyd – Getting more from the PLS model: application to metabolomics

26. Gavin Rhys Lloyd – Statistics in R Using Class Templates (StRUCT)

27. Mercedes Bertotto – Detection of High Fructose Corn Syrup in Honey by Fourier Transform Infrared Spectroscopy and Chemometrics

28. Sumana Narayana – Mid-infrared spectroscopy and multivariate analysis to characterize Lactobacillus acidophilus fermentation processes

29. Ellen Færgestad Mosleth – Gene expression in petroleum workers exposed to sub-ppm benzene levels

30. Barry M. Wise - A Comparison of ANNs, SVMs, and XGBoost in Challenging Classification Problems

31. Mats Josefson - Experiments with complex numbered multivariate data analysis