Drushbametrics Project

Eleventh Winter Symposium on Chemometrics

Modern Methods of Data Analysis

Russia, Saint Petersburg, February 26–March 2, 2018

Russian Chemometrics Society
Sensors System LLC
Semenov Institute of Chemical Physics RAS

Eleventh Winter Symposium on Chemometrics

Modern Methods of Data Analysis

The organizing committee

Co-chairmen

Dmitry Kirsanov
Alexej Skvortsov

Secretary

Irina Yaroshenko

Members

Sergey Kucheryavskiy
Andrey Legin
Federico Marini
Vitaly Panchuk
Alexey Pomerantsev
Oxana Rodionova

The symposium is included in the list of FASO Russia conferences

Address

Mendeleev center, Universitetskaya nab. 7/9 199034 Saint Petersburg, Russia

http://wsc.chemometrics.ru/wsc11 e-mail: wsc11@chemometrics.ru

Thanks

The WSC-11 organizers and participants wish to express greatest appreciation to the following conference sponsors for their valuable economic and friendly help:

CSort Ltd

Art photonics Gmbh

Bicasa

American Elements

INTERTECH Corporation

Finally, we are grateful to all the WSC-11 attendees, lecturers, accompanying persons, and visitors for their interest in the conference.

See you again at the next WSC-12 conference!

Useful information

Conference and activities

Conference sessions will be held in a conference hall (nr. 15 in the map below). In free time participants can try various winter activities: skiing, skating, sledge skating, horse riding, fishing. In addition, sauna and billiard are also available.

Meals

All meals will be served in the restaurant "Nautilus" (3). The banquet will take place in the banquet room of the conference hall.

Scores & Loadings

Traditional “Scores and Loadings” meetings will be held in coffee-break area (same building as the conference-hall).

Communication

The main Russian cellular networks MTS and Megafon have a proper coverage around the hotel. WiFi is available in all the buildings.

Money

You can exchange EUR and US dollars in Saint Petersburg banks, at the railway station or in the airport. The organizing committee could also exchange a reasonable amount of currency. There is also an ATM machine at the reception where you can get some cash in Russian rubles. Credit cards are accepted in the hotel.

Excursion

Friday morning all participants are invited to a Saint Petersburg city tour, where you will have the possibility to see the most impressive palaces and cathedrals, will learn the life stories of the Romanov Dynasty and will enjoy Russian architecture and cultural heritage.

Connection with Saint-Petersburg

The hotel is located about 80 km from Saint-Petersburg and about 20 km from the city of Zelenogorsk.

Besides the official conference transfer you can reach the hotel by shuttle van 827 from “Grazhdanskiy prospect” metro station (red line), the stop is “Aurora club”. Another option is the shuttle van 679 and 830 from “Parnas” metro station (blue line).

The coordinates of the resort are:

Latitude 60°18′7.01″ N (60.301946)

Longitude 29°17′25.42″ E (29.290395)

Miscellaneous

The official conference language is English.

Everyone is encouraged to have his/her badge attached, both during the symposium sessions and social activities.

Useful Phone Numbers

Irina Yaroshenko, *conference secretary*	+7 (952) 371-01-39
Dmitry Kirsanov, *local organizing committee*	+7 (921) 333-12-46
Aurora hotel reception	+7 (911) 271-67-50

Map of “Aurora club” resort

1. Entrance and security
2. Resort administration
3. Restaurant “Nautilus”
   bar “Captain Nemo”
4. Townhouse “Cape”
5. Apart hotel “Odysseus”
   Grocery store
   Leasing point
6. Playground for children
7. Beach
8. Leisure ground “Aurora”
9. Townhouse “Breeze”
10. Townhouse “Lagoon”
11. Townhouse “Poseidon”
12. Townhouse “Neptune”
13. Riding club “Aurora”
14. Leisure ground “Cape”
15. Conference-hall
16. Leisure ground “Summer theatre”
17. Hotel “Reef”
18. Sport ground
19. Hotel “Shell”
20. Mini-golf
21. Russian banya (sauna)
22. Diving club

Monday, February 26, 2018

11:00–13:00	**Registration**
13:00–14:00	**Lunch**
14:00–15:00	**Registration and opening**
Session 1	**Chair: Kim Esbensen**
15:00–15:30	**T1** *Yuri Kalambet* Points per peak and integration rules
15:30–16:00	**T2** *Andrey Samokhin* Decreasing false positive identification rate by predicting absence of the correct answer in mass spectral library
16:00–16:30	**Coffee break**
16:30–17:00	**T3** *Henning Schroeder* Kinetic modeling and the set of feasible reaction rates
17:00–17:30	**T4** *Alisa Rudnitskaya* Application of ComDim (Common Component and Specific Weight Analysis) to the interpretation of the electronic tongue and infrared spectral data
17:30–18:00	**T5** *Dmitry Kirsanov* PLS regression for spectral data smoothing
18:00–19:00	**Free time**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Tuesday, February 27, 2018

08:00–09:00	**Breakfast**
Session 2	**Chair: Paul Gemperline**
09:00–10:00	**L1** *Alexey Pomerantsev* Multi-class PLS-DA: soft and hard approaches
10:00–10:30	**T6** *Oxana Rodionova* Data Driven SIMCA – more than One-Class Classifier
10:30–11:00	**Coffee break**
11:00–11:30	**T7** *Sergey Kucheryavskiy* Using decision trees and their ensembles for analysis of NIR spectroscopic data
11:30–12:00	**T8** *Karoly Heberger* Bias-variance trade-off. Comparison of cross-validation and bootstrapping by sum of ranking differences
12:00–13:00	**Free time**
13:00–14:00	**Lunch**
Session 3	**Chair: Cyril Ruckebusch**
14:00–15:00	**L2** *Peter Harrington* Enhanced restricted Boltzmann machines for classification of analytical measurements
15:00–15:30	**T9** *Federico Marini* Non-linear predictive modeling using local approaches
15:30–16:00	**T10** *Dörgő Gyula* Multivariate statistical models and Bayes chain rule-based analysis of sequence to sequence deep learning models
16:00–16:30	**Coffee break**
Session 4	**Chair: Oxana Rodionova**
16:30–17:00	**T11** *Alon Mazafi* Smart multi-electrode array for simultaneous detection of multiple analytes in urine
17:00–17:30	**T12** *Ekaterina Oleneva* Distinction of cancer and healthy tissue using NIR-spectroscopy and multivariate data processing
17:30–18:00	**T13** *Yulia Monakhova* Breakthrough in heparin analysis: holistic control by NMR spectrometry and chemometrics
18:00–19:00	**Free time**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Wednesday, February 28, 2018

08:00–09:00	**Breakfast**
Session 6	**Chair: Alexey Pomerantsev**
09:00–10:00	**L3** *Søren Balling Engelsen* Single kernel spectroscopy. Motivations, practicality and new technology
10:00–10:30	**T14** *Andrey Bogomolov* Optical multisensor systems
10:30–11:00	**Coffee break**
11:00–11:30	**T15** *Anastasiia Melenteva* Multi-mode fiber spectroscopy for cancer diagnostics
11:30–12:00	**T16** *Aleksey Zarubin* The use of FT-IR/ATR spectroscopy and PLS-DA in checking the authenticity of the aspirin tablets
12:00–13:00	**Free time**
13:00–14:00	**Lunch**
Session 7	**Chair: Federico Marini**
14:00–15:00	**L4** *Cyril Ruckebusch* Image processing in chemical imaging
15:00–15:30	**T17** *Shuxia Guo* Multimodal image analysis for tissue diagnosis of skin melanoma
15:30–16:00	**T18** *Janos Elek* Near infrared study of hydration state of the human body through skin
16:00–16:30	**Coffee break**
Session 8	**Chair: Karoly Heberger**
16:30–19:00	**Poster Session**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Thursday, March 1, 2018

08:00–09:00	**Breakfast**
Session 9	**Chair: Søren Balling Engelsen**
09:00–10:00	**L5** *Paul Gemperline* Multivariate modeling and chemometric resolution of pure component profiles from mixture spectra of evolving systems
10:00–10:30	**T19** *Alexej Skvortsov* Estimating rotation ambiguities in MCR: a quest for speed and measure
10:30–11:00	**Coffee break**
11:00–11:30	**T20** *Valeria Belikova* Distance estimation between objects in spectral data analysis
11:30–12:00	**T21** *Dávid Bajusz* Large-scale comparison of similarity metrics for molecular and interaction fingerprints
12:00–13:00	**Free time**
13:00–14:00	**Lunch**
Session 10	**Chair: Dmitry Kirsanov**
14:00–15:00	**L6** *Kim Esbensen* MSPC is not enough for proper process monitoring and control — chemometrics can do better: TOS variography
15:00–15:30	**T22** *Cristina Malegori* How different water activities affect rice germ shelf life: an aquaphotomics approach
15:30–16:00	**T23** *Viacheslav Artyushenko* Multi-spectral fiber spectroscopy methods for guided diagnostics of abdominal cancer
16:00–16:30	**T24** *Ramin Nikzad-Langerodi* Domain-invariant Partial Least Squares (di-PLS) Regression: a novel method for unsupervised and semi-supervised calibration model adaptation
16:30–17:00	**Closing**
17:00–17:30	**Coffee break**
17:30–19:00	**Free time**
19:00–00:00	**Banquet**

Friday, March 2, 2018

08:00–09:00	**Breakfast**
09:00–10:00	**Check out**
10:00–15:00	**Bus to Saint Petersburg and excursion**

Abstracts

L01. Multi-class PLS-DA: soft and hard approaches

A.L. Pomerantsev[1,2], *O.Ye. Rodionova*[1,2]

[1]*NN Semenov Institute of Chemical Physics RAS, Moscow, Russia*

[2]*Branch of Institute of Natural and Technical Systems RAS, Sochi, Russia*

In this talk, we consider the multi-class version of PLS-DA, which, in fact, is not more complex than the conventional binary (two-class) PLS-DA [1]. Our method does not utilize the PLS scores, but is entirely based on the predicted dummy responses, $\mathbf{Y}_{\text{hat}}$. To get around the degeneracy of this matrix, we use the PCA scores, **T**. These scores are used for classification by the LDA approach. We also propose the novel soft version of the PLS-DA method, which is based on QDA applied to the **T** data. In this version, discrimination rule employs the Mahalanobis distances and a threshold, which is calculated for a given type I error. According to this rule, a sample can be simultaneously attributed to several classes, or it may be not allocated at all. It was demonstrated that the soft PLS-DA is able to avoid misclassification, in case a new object is not a member of any target class.

The principal measures of classification quality (sensitivity, specificity, and efficiency) are defined for the multi-class PLS-DA. It is also shown how these characteristics are used for the selection of the complexity of the model, which is the number of the PLS latent variables. A popular opinion that an equal number of objects in the training classes is preferred for a good PLS-DA model is analyzed and found to be wrong.

The comparison of the discriminant (PLS-DA) and the class-modeling (SIMCA) methods is conducted using the simulated and real-world examples. In particular, it is shown that SIMCA is better when one class is tight. On the contrary, PLS-DA is preferable in cases when we separate two classes with the same major components but different impurities. We considered a very popular PLS-DA strategy, when one target class is discriminated against a collection of all available alternative classes. It is demonstrated that this approach may be a reasonable method for authentication when the soft PLS-DA is applied, but not in the case of the hard one.

Finally, we can repeat our notion presented in [2], “The ‘best’ classification

method does not exist. Every task at hand requires an application of a pertinent chemometric method best suited to answer the posed question."

References

1. AL Pomerantsev, OY Rodionova, Multiclass partial least squares discriminant analysis: Taking the right way – A critical tutorial, *J. Chemom* (submitted, July 2017)

2. OY. Rodionova, AV Titova, AL Pomerantsev, *Trends Anal. Chem.*, **78** (4), 17-22 (2016)

L02. Enhanced Restricted Boltzmann Machines for Classification of Analytical Measurements

Peter de Boves Harrington
Ohio University, Center for Intelligent Chemical Instrumentation, Department of Chemistry & Biochemistry, Clippinger Laboratories, Athens, OH, USA

Deep learning networks have many successful applications; most notably face and feature recognition in photographs, automatic translation of speech and textual images, hand-written character recognition, and speech recognition for digital personal assistants. These networks comprise many layers and one innovation is the restricted Boltzmann machine (RBM) [1]. An RBM forms a single layer of a network and it may be trained individually so that layers of trained RBMs may be stacked together to form a deep learning network. The RBM was designed for binary or bit encoded images. Modifications have been made to the RBM to accept real encoded inputs that are amenable to analytical measurements. However, these RBMs have a notorious reputation to be very difficult to train and require fine-tuning of additional parameters.

A modified and simple version of the RBM has been developed that has much better convergence properties and can enhance classification when coupled to other chemometric classifiers. The new algorithm will be demonstrated with several reference classification datasets from the machine learning archives. Modifications to the algorithm include a scaling parameter that controls the hardness or slopes of the sigmoidal outputs. In addition, binary and bipolar output encoding will be compared. The encoded outputs are then used as inputs

into the super partial least squares-discriminant analysis that uses an internal bootstrap Latin partition (BLP) to optimize the number of latent variables [2]. In addition, a support vector machine classification tree also is used as a classifier that uses a fuzzy entropy function to encode the bipolar classes at each branch of the tree. BLP external validation3 demonstrates that the RBM can significantly improve classification results when the number of RBM features (i.e., encoded outputs) are increased beyond the number of initial variables of the data.

References

1. Hinton, G. E. In Learning distributed representations of concepts, Eighth Annual Conference of the Cognitive Science Society, Amherst, Mass, Amherst, Mass, 1986.

2. Harrington, P. D.; Kister, J.; Artaud, J.; Dupuy, N., Automated Principal Component-Based Orthogonal Signal Correction Applied to Fused Near Infrared-Mid-infrared Spectra of French Olive Oils. Anal Chem 2009, 81 (17), 7160-7169.

3. Harrington, P. B., Multiple Versus Single Set Validation of Multivariate Models to Avoid Mistakes. Crit Rev Anal Chem 2018, 48 (1), 33-46.

L03. Single kernel spectroscopy. Motivations, practicality and new technology

Weiwei Cheng, Klavs Martin Sørensen & Søren Balling Engelsen,
Department of Food Science, Faculty of Science, University of Copenhagen, Frederiksberg C, Denmark

Plant phenomics is considered as a key technology with which to solve the global challenges of food security, climate change and biofuel production. It is of primary interest to develop new varieties of cereals that can grow and produce efficiently under new climate conditions: typically more drought, salinity, occasional flooding, biotic stress, etc. Moreover, it is well known to biologists that, even in the case of a single variety grown under the same conditions, the range in crop traits is substantial. Major variations in, for example, protein content can be observed in the same cereal crop and such variations in quality traits between individual biological entities (seeds) represent a potential

economic, functional and quality reward if ultra- high throughput sorting methods can be devised.

Near infrared spectroscopy has proven as a very efficient method for plant phenomics as it is very fast, reproducible and provides an excellent mass overview of the bulk chemical composition of the crop. Single kernel NIR spectroscopy can be used for breeding of cereals, for quality control of cereals and for sorting of cereals according to different quality traits. While hyperspectral imaging and/or reflectance spectroscopy can be used in quality control for example to detect seeds with diseases and defects it has no relevance in high throughput seed sorting. In order to attain information about the chemical information of the endosperm it is necessary to use transmittance spectroscopy. This can in practice best be achieved in the energy rich short-wavelength NIR region (SW-NIR), but new instrument technology may soon change the game. This paper gives an update on how single seed sorting according to protein content can be achieved in the SW-NIR region (new chemometric approach using rPLS), how the sample presentation to the spectrometer influence the sorting accuracy (sampling theory), and how new technology such as supercontinuum lasers (experimental spectrometer) may open of for sorting of more complex traits such as beta-glucan content which best can be found in the long wavelength NIR region (LW-NIR).

L04. Image processing in chemical imaging

Cyril Ruckebusch
Université de Lille LASIR CNRS F-59000 Lille, France

Chemical imaging aims to create a mapping of the different compounds in a sample or process to provide more comprehensive insight compared to single point measurement. This can be achieved by simultaneous measurement of spatial, spectral and/or temporal information. On top of chemical analysis, image processing is required for operations such as denoising and deconvolution or to encourage some additional spatial properties. In this talk, we present our recent efforts to implement image processing methodologies in chemical imaging chemometrics and we will show how the results obtained can be relevant to different types of imaging modalities.

We will start with a very short introduction to sparsity, a well-known concept in statistics and chemometrics, related to simplicity. Sparse modeling also plays an important role for signal/image processing, for regularizing inverse problems or to encourage signal roughness.

The first part of the talk will deal with the analysis of hyperspectral images which provide a way to explore both spatial and spectral information. We recently proposed an adaptation of the Multivariate Curve Resolution – Alternating Least Squares framework that opens the possibility to implement any kind of spatial information as a constraint.[1] Examples are provided for the implementation of image processing tools as spatial constraints.[2,3] We show that more accurate chemical maps and spectral profiles of the image compounds are obtained.

The second part will be on the analysis of super-resolution fluorescence microscopy data (spatial-temporal or "time-lapse" data) which can reveal both structural information at the nanoscale and provide dynamic insight into biological processes occurring in live cell samples. The core of the proposed approach is sparse deconvolution with penalized regression for densely-labeled samples.[4,5] Sparsity of the fluorophore distribution in the spatial domain and continuity of the fluorophore localizations in the time mode are combined to achieve improved spatial resolution and faster temporal resolution.

References

1. On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis, S. Hugelier, O. Devos, C. Ruckebusch, Journal of Chemometrics, 2015

2. Application of a sparseness constraint in multivariate curve resolution–alternating least squares, Analytica Chimiva Acta, 2018

3. Edge-preserving image smoothing constraint in multivariate curve resolution–alternating least squares (MCR-ALS) of Hyperspectral Data, S. Hugelier, O. Devos, C. Ruckebusch, Applied Spectroscopy, 2017

4. Sparse deconvolution of high density super-resolution images, S. Hugelier, J. de Rooi, R. Bernex, S. Duwé, O. Devos, M. Sliwa, P. Dedecker, P. H. C. Eilers, C. Ruckebusch, Scientific Reports, 2016.

5. Improved super-resolution microscopy imaging by sparse deconvolution with an inter-frame penalty, S. Hugelier, P. H. C. Eilers, O. Devos, C. Ruckebusch, Journal of Chemometrics, 2017.

L05. Multivariate modeling and chemometric resolution of pure component profiles from mixture spectra of evolving systems

Paul J. Gemperline
Department of Chemistry, East Carolina University, Greenville, NC, USA

In chemometrics, two very different classes of mathematical tools, self-modeling methods and first-principles methods, have been developed to resolve pure component concentration profiles and spectra from mixture spectra recorded over time during evolving processes. This paper presents examples of each approach, as well as their advantages and disadvantages with applications from chromatograph to industrial batch processes. Examples are shown to illustrate the discovery of hidden behaviors in batch processes as well as modeling of reaction calorimetry, dissolution, and crystallization process in comprehensive first-principles models.

In self-modeling curve resolution (SMCR) methods, realistic constraints such as non-negativity of concentration profiles or equality constraints for known pure component spectra are imposed to produce solutions that obey the constraints. In many cases, SMCR is the only method available for resolving the pure component profiles; however, it is less widely appreciated that in most circumstances, SMCR techniques do not produce unique mathematical solutions, rather a range of feasible solutions that obey boundaries imposed by the constraints. In this presentation, SMCR with a method for computing the range of feasible solutions is illustrated and an algorithm for SMCR that yields improved results by use of soft constraints with penalty functions is also described.

Methods of fitting first-principles multivariate kinetic models are powerful alternatives to SMCR. Such modeling methods do not suffer from ambiguities in the resulting solutions. In process monitoring and control applications, numerical fitting of a comprehensive kinetic model makes use of dynamic information to estimate reaction rates, chemical equilibria, process states, end-points, deviations from optimal performance, and can provide mechanistic information for process adjustment and optimization.

L06. MSPC is not enough for proper Process Monitoring and Control — Chemometrics can do better: TOS variography

K.H. Esbensen
KHE Consulting (KHEC), Copenhagen, Denmark (www.kheconsult.com)
Geological Survey of Denmark and Greenland (GEUS), Copenhagen, Denmark.
Department of Chemistry and Bioscience, University of Aalborg, Denmark
Advanced Analytical & Analysis Associates (AAAA)

Process monitoring and control in technology and industry is incomplete without full understanding of all sources of variation. It is not enough to be in command of Multivariate Statistical Process Control (MSPC) if the process data are affected by significant errors which have not been adequately identified, quantified and reduced to below a relevant a priori acceptance threshold, i.e. process data are typically affected both by classical analytical measurement errors as well as sampling and/or PAT sensor acquisition errors. The latter categories can dominate over analytical errors by factors 10-20 or more when proper process sampling competence is not brought to bear in the design, maintenance and operation of the total process measurement system. These errors are often physical sampling errors associated with process sample- and/or reference sample extraction, but it is not sufficiently known in chemometrics that PAT signal acquisition gives rise to identical error types and magnitudes. All this is well understood and solutions abound in the Theory of Sampling (TOS).

In science, technology and industry, in materials processing and in goods production and manufacturing it is essential to eliminate, or reduce maximally, all unnecessary contributions to the Total Measurement Uncertainty (MUtotal)

in order to perform valid process multivariate monitoring and control (MSPC) in order to bring forth the process signals with maximum signal/noise enhancement. This task is standard in the Theory of Sampling (TOS), which forms the only reliable scientific framework from which to seek resolution. A variographic process description allows quantification of the sum-total effect of all unwanted, indeed unnecessary, adverse sampling/sensor acquisition errors. Continuous variographic process characterisation, with application-dependent updating, is able to issue warnings that a particular process measurement system is not, or is no longer, fit-for-purpose representative and must therefore be rectified. The variographic approach works directly on the routine process data coming off the process – i.e. no extra measurements are needed than the existing process data. Variographics is a real-time, on-line self-controlling plug-in facility of wide applicability, yet with a very simple usage. This presentation calls for a process monitoring paradigm shift in chemometrics; variographics constitutes a missing link in MSPC and PAT. Examples and case histories from selected industrial sectors will illustrate the above issues, a.o. self-contradictionary FDA sampling demands for critical pharmaceutical mixing processes, EU GMO monitoring and control, and minerals processing pathways in a major European mining and value-added processing company.

T01. Points per peak and integration rules

Yuri Kalambet[a], *Yuri Kozmin*[b], *Andrey Samokhin*[c]
[a]*Ampersand Ltd, Moscow, Russian Federation*
[b]*Shemyakin–Ovchinnikov Institute of bioorganic chemistry RAS, Moscow, Russian Federation*
[c]*Chemistry Department, Lomonosov Moscow State University, Moscow, Russian Federation*

Integration rules for the case of time series with constant data rate are discussed. All rules have extreme efficiency when applied to peaks, error drops down exponentially on data rate. The most efficient rule for peaks is Trapezoidal. Rules, based on Euler-Maclaurin formula are constructed. It is shown, that in general case integration rules, based on Euler-Maclaurin formula are the best choice for numerical integration. Composite rules, based on Newton-Cotes formulas, always have lower efficiency compared to Euler-Maclaurin rules of corresponding order. Elementary Newton-Cotes rules can be treated as implementation of Euler-Maclaurin rules to the regions with minimal number of points for appropriate rule order.

T02. Decreasing false positive identification rate by predicting absence of the correct answer in mass spectral library

Andrey Samokhin, *Ksenia Sotnezova, Igor Revelsky*
Division of Analytical Chemistry, Chemistry Department, Lomonosov Moscow State University, Moscow, Russia

Gas chromatography coupled with electron ionization mass spectrometry is widely used for tentative identification of components of complex mixtures. Such identification is usually based on commercial mass spectral libraries. The largest libraries (NIST and Wiley) contain spectra of hundreds of thousands of compounds. At the same time, the number of small molecules in chemical structure databases (such as CAS, PubChem or ChemSpider) reaches more than fifty million.

Library search algorithms available in the market cannot predict the absence of the correct answer in the library. In this work, we tried to implement such prediction by means of supervised classification. Multivariate model was built

using partial least squares discriminant analysis (PLS-DA). Similarity indexes of several top candidates (proposed by the library search algorithm) were used as initial variables. Training, validation and prediction set contained 3000, 1500 and 1500 objects respectively. The developed model was able to correctly predict absence of compound in the library in 30% of cases (only 1% of compounds actually presented in the library were wrongly classified).

The reported study was funded by The Russian Foundation for Basic Research, according to the research project No. 16-33-60169 mol_a_dk

T03. Kinetic modelling and the set of feasible reaction rates

Henning Schröder[a,b], *Mathias Sawall*[a], *Annekathrin Moog*[a], *Klaus Neymeyr*[a,b]
[a]*Department of Mathematics, University of Rostock, Rostock, Germany*
[b]*Leibniz Institute for Catalysis, Rostock, Germany*

A spectrum, taken from a chemical reaction system of s components, is a superposition of the s pure component spectra weighted by the corresponding concentrations. A series of spectra can be stored row-wise as a matrix D. The aim of multivariate curve resolution (MCR) methods is to recover the concentration profiles and pure component spectra from D. This requires the computation of a nonnegative matrix factorization $D=CS^T$.

In general, this task cannot be solved uniquely. Typically, there exist continua of feasible factorizations. However, only a single, namely the chemically true, solution is desired. Kinetic modeling can help to extract these factors. It is often believed that kinetic modeling is sufficient in order to guarantee a unique factorization. However, particularly for first-order reaction systems it can be shown that there still exist qualitatively and quantitatively different solutions [1,2].

In the first part of the talk, we present a kinetic hard-model approach to solve MCR problems. One can determine a pair of factors C and S by an optimization so that C is also consistent with the kinetic model. The reaction rate constants are computed simultaneously.

In the second part, a model data matrix D is analyzed which is consistent with the reversible kinetic model $X \rightleftharpoons Y$. Even under this assumption there exists a

set of feasible factorizations. We present a general analysis of the ambiguity of rate constants for arbitrary first-order kinetic models.

The results are applied to FTIR data sets on hydroformylation.

References

1. H. Schröder, M. Sawall, C. Kubis, D. Selent, D. Hess, R. Franke, A. Börner, K. Neymeyr, On the ambiguity of the reaction rate constants in multivariate curve resolution for first-order reaction systems, Anal. Chim. Acta 927, 21-34 (2016).

2. S. Vajda, H. Rabitz, Identifiability and Distinguishability of First-Order Reaction Systems, J. Phys. Chem., 92, 701-707 (1988).

T04. Application of ComDim analysis (Component and Specific Weight Analysis) to the interpretation of the electronic tongue and infrared spectral data

A. Rudnitskaya
CESAM and Chemistry Department, University of Aveiro, Aveiro, Portugal

Electronic tongue multisensor systems (ETs) and infrared spectroscopy became popular analytical tools due to the simplicity of the measuring procedures, high speed and low cost of analysis. Both instruments produce complex non-selective signals in the multicomponent samples. These signals can be related to the property or concentration of the component of interest using chemometric techniques, but are often difficult to interpret. Some insights can be gained from studying response to the target compounds in the individual solutions and by standard addition to the analyzed samples, but it is not always feasible due to the nature and number of the compounds or properties of interest. Alternative approach may consist in comparing the instruments' response with the reference data. Such comparison can be carried out using multi-block analysis methods, application of one of which, ComDim or Common Components and Specific Weight Analysis, will be discussed.

ComDim analysis was developed to enable description and comparison of several data tables measured on the same samples [1]. The ComDim analysis is based on the assumption that a set of data tables contains a common structure, which is revealed by calculating a set of common components or common

dimensions that recover the maximum total variance of each of the data tables. The ComDim analysis is, thus, allows to visualize samples on the basis of the extracted common components, i.e. on the basis of the variance that is common to all data tables and to investigate the relationships among various data tables. The ComDim analysis was first applied to the sensory profiling analysis [1] and later on to the coupling of different kinds of measurements made on food products [2].

This study discusses ComDim analysis applications to the interpretation of the response of the ET based on potentiometric chemical sensors and Fourier Transform Mid-Infrared spectra using sets of chemical parameters and sensory attributes measured in wine, coffee and oak wood extracts as examples.

References

1. E.M. Qannari, I. Wakeling, P. Courcoux, H.J.H. MacFie, Food Qual. Prefer. 12 (2001) 365-368.

2. M. Hanafi, G. Mazerolles, E. Dufour, E. M. Qannari, J. Chemometrics 20 (2006) 172–183.

Acknowledgements

Financial support of this work by CESAM (UID/AMB/50017) and by FCT/MEC through national funds and the co-funding by the FEDER, within the PT2020 Partnership Agreement and Compete 2020, and fellowship SFRH/BPD/104265/2014 is kindly acknowledged.

T05. PLS regression for spectral data smoothing

V.V.Panchuk[1,2], V.G.Semenov[1,2], A.V.Legin[1,3], D.O.Kirsanov[1,3]

[1] Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia

[2] Institute for Analytical Instrumentation RAS, St. Petersburg, Russia

[3] Laboratory of artificial sensory systems, ITMO University, St. Petersburg, Russia

Smoothing of instrumental signals is important prerequisite in data processing. Various smoothing methods were suggested through the last decades each having their own benefits and drawbacks. Numerous smoothing methods are described in literature based on three main concepts: 1) averaging in a certain spectral window (like e.g. Savitzky-Golay filter); 2) frequency domain

representation (Fourier filter); 3) signal approximation with appropriate mathematical function (e.g. ALS procedure by Paul Eilers). In this paper we suggest a novel approach to the fundamental issue in analytical chemistry – signal smoothing. The overall idea behind this type of filtering is in sorting the signals according to the variance they hold. This can be implemented through signal reconstruction with the help of PLS regression – a basic tool of chemometrics. The presentation will show that PLS smoothing allows for significant signal-to-noise ratio improvement without serious distortion of line parameters (width, position, amplitude). Moreover, the PLS smoothing allows for spectral resolution improvement. The examples from Mössbauer and EDX spectrometry illustrate the process of filtering parameters selection and show that the suggested procedure can be successfully applied for spectral preprocessing both in quantitative and qualitative analysis. The basis line employed for PLS decomposition can be of any complexity (e.g. triplet), thus the smoother performance can be adapted to the wide variety of real world applications.

T06. Data Driven SIMCA – more than One-Class Classifier

O.Ye. Rodionova[a,b], *A.L.Pomerantsev*[a,b]
[a]*Semenov Institute of Chemical Physics RAS, Moscow, Russia*
[b]*Branch of Institute of Natural and Technical Systems RAS, Sochi, Russia*

The main purpose of the SIMCA method is to solve one-class classification (OCC) problems, employing the decision rule developed on the base of the target class objects. Data driven SIMCA (DD-SIMCA) is an advanced OCC method, which computes the misclassification errors theoretically. The data driven approach adds a possibility to estimate parameters for the *score distance* (SD) and the *orthogonal distance* (OD) distributions. The fact that both the SD and OD follow the scaled chi-squared distribution provides a possibility to introduce the *total distance* (TD) that is a new statistic calculated as a weighted sum of the SD and OD variables. Using TD, we can develop an acceptance area /decision rule for a given value of the type I error, α [1, 2].

Analysis of the parameters of the SD and the OD distributions provides additional information that can be important for understanding the data structure. We are reporting the following possibilities:

1. changes in the numbers of degrees of freedom for SD and OD against the number of principal components reveal the structure of the training set;
2. comparison the estimates of parameters calculated by the classical and the robust methods helps to reveal outliers and to clean the training set;
3. the behavior of the test samples in the Extreme plot [2] helps to understand to what extent the test set is similar to the training set.

All the above-mentioned issues are illustrated using simulated data and real-world examples.

These are only several features that are provided by the data driven approach and we consider that potentially it has more possibilities and practical applications.

Acknowledgements: We acknowledge partly funding from the IAEA in the frame of projects D5240 and G42007.

References

1. Pomerantsev A. Acceptance areas for multivariate classification derived by projection methods. *J Chemometrics*. 2008;22(11-12):601-609.
2. Pomerantsev AL, Rodionova OY. Concept and role of extreme objects in PCA/SIMCA. *J Chemometrics.* 2013;28(5):429-438.

T07. Using decision trees and ensembles for analysis of NIR spectroscopic data

Sergey Kucheryavskiy
Department of Chemistry and Bioscience, Aalborg University, Esbjerg, Denmark

Advanced machine learning methods, like convolutional neural networks and decision trees, became extremely popular in the last decade. This, first of all, is directly related to the current boom in Big data analysis, where traditional statistical methods are not efficient. According to the kaggle.com — the most popular online resource for Big data problems and solutions — methods based

on decision trees and their ensembles are most widely used for solving the problems.

It can be noted that the decision trees and convolutional neural networks are not very popular in Chemometrics. One of the reasons for that is the landscape of the data matrix: the modern machine learning methods need number of measurements much larger than the number of variables to avoid overfitting, which is opposite to the layout of the data we usually deal with. Another drawback is a lack of interactive instruments for exploring and interpretation of the models.

In this presentation, we are going to discuss an applicability of decision trees based methods (including gradient boosting) for solving classification and regression tasks with NIR spectra as predictors. We will cover such aspects as evaluation, optimization and validation of models, sensitivity to outliers and selection of most important variables.

T08. Bias-variance trade-off. Comparison of cross-validation and bootstrapping by sum of ranking differences

Károly Héberger[1] and Klára Kollár-Hunek[2]

[1]Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

[2]Budapest University of Technology and Economics, Department of Inorganic and Analytical Chemistry Budapest, Hungary

It is well-known that "five- or tenfold cross-validation will overestimate the true prediction error. On the other hand, leave-one-out cross-validation has low bias but can have high variance." [1]. However, the characteristics of bootstrapping (randomly drawn datasets with replacements) are not so well-known, especially when comparing it to the variants of cross-validation (stratified, venetian blind, random split and leave-one-out).

Sum of (absolute) Ranking Differences (SRD) provides an easily perceivable method- and model comparison procedure [2,3]. The procedure incorporates two types of validation: i) Comparison of Ranks with Random Numbers (CRRN) i.e. a permutation test and ii) a stratified sevenfold cross-validation [3]. Ties

(equal numbers in the input matrix) can deteriorate the ranking; hence, the procedure has been extended for such cases [4].

A detailed comparison of cross-validation and bootstrapping has been carried out and illustrated with two case studies: i) comparison of classifiers using performance parameters (accuracy, specificity, sensitivity and Cohen's Kappa) and ii) Comparison of quantitative structure-retention relationship models in prediction of retention indices of polycyclic aromatic hydrocarbons.

Scaling (standardization, rank transformation, normalization, etc.) of the input data is a prerequisite for SRD calculations. As SRD is normalized for the same scale, analysis of variance (ANOVA) is able to discriminate the effects of various factors: i) cross validation type (stratified and bootstrap), ii) leave-many-out (from five to ten), iii) classifier types, etc. Computer codes are freely available from the authors upon request and/or from our homepage: http://aki.ttk.mta.hu/srd

Acknowledgement

The authors thank the support of the National Research, Development and Innovation Office of Hungary (OTKA, contracts Nos K 119269 and KH-17 125608).

References

1. Hastie, T.; Tibshirani, R.; and Friedman, J. H. Cross-Validation. Chapter 7.10 in: The elements of statistical learning: Data mining, inference, and prediction, 2nd ed., pp. 241-249. New York, Springer (2009).

T09. Non-linear predictive modelling using local approaches

Alessandra Biancolillo, Federico Marini
Department of Chemistry, University of Rome "La Sapienza", Rome, Italy

While the development of modern analytical instrumentation has made possible, on one hand, the characterization of increasingly complex matrices, making also possible their high throughput analysis, on the other hand has resulted in the possibility that signals be affected by sources of variations other than those one may be interested in. As a consequences, non-linearities can be introduced in the relations between the collected signals and the responses to be predicted.

In such situations, the possibility of using non-linear modeling approaches should allow more accurate predictions but quite often, apart from being more prone to overfitting, such models require a higher number of samples and their optimization may not be straightforward.

Local approaches, i.e. those techniques which operate by approximating the globally non-linear problem into a composition of smaller locally linear models, are a family of versatile predictive models, which allow to couple the advantages of the linear predictive techniques with the possibility of modeling global relations with a tunable complexity.

The present communication aims at discussing the basics of locally linear approaches for regression and classification and to introduce some novel implementations for nonlinear modeling of multi-way arrays or in the context of data fusion. Moreover, some possible ways of interpreting on the global basis the results of the individual local models will also be presented.

T10. Multivariate statistical models and Bayes' chain rule-based analysis of sequence to sequence deep learning models

Gy. Dörgő[1,2], P. Pigler[1,2], J. Abonyi[1,2]
[1]MTA-PE Lendület Complex Systems Monitoring Research Group,
[2]Department of Process Engineering, University of Pannonia, Veszprém, Hungary

In complex chemical plants, an enormous amount of discrete data is recorded every day by the process control system in the form of alarms and warnings. These datasets can carry a high amount of process relevant information giving the opportunity for the development of data-driven models for event prediction or alarm suppression purposes by exploring information from the historical log files of alarm systems. Our work aims the development of sequence-based Bayes model and recurrent neural network based solutions for the determination of frequently occurring alarm patterns, and how these patterns can be applied for the prediction of future incoming sequences.

The proposed recurrent neural network model applies an encoder layer of Long Short-Term Memory (LSTM) units to map the sequences of events have presently happened in our process into a vector of fixed dimensionality, and a decoder LSTM layer to predict the sequence of future events based on the

information contained by the internal state vector of the encoder layer. We demonstrate how this model can predict the probability of a future event and show how the events above a certain probability threshold can represent frequently occurring operating patterns. The extracted sequences are compared to the ones obtained by the multi-temporal sequence based Bayes model.

The determination of the optimal structure of the proposed recurrent neural network based approach is usually carried out by trial and error type technics, without the demand of deeper understanding of the causes and consequences. We use multivariate chemometric models for the visualization of the layer activities and weights to support the determination of the optimal model structure and extract prediction-relevant alarm signals.

The methodology is motivated by the revision of the alarm management system of an industrial delayed coker unit. We show illustrative examples related to this industrial dataset and present a reproducible benchmark example on the simulator of a vinyl acetate [1] production technology to ensure the comparability of the results.

The research has been supported by the National Research, Development and Innovation Office NKFIH, through the project OTKA 116674 (Process mining and deep learning in the natural sciences and process development). Gyula Dörgő was supported by the ÚNKP-17-3-II. New National Excellence Program of the Ministry of Human Capacities.

References

1. R. Chen, K. Dave, T. J. McAvoy, "A Nonlinear Dynamic Model of a Vinyl Acetate Process" in Industrial & Engineering Chemistry Research, vol. 20., pp 4478–4487, Mar. 2003.

T11. Smart multi-electrode array for simultaneous detection of multiple analytes in urine

Alon Mazafi and Hadar Ben-Yoav
Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Smart electrochemical sensors — arrays of partially-selective electrodes that generate cross-reactive signals from multiple redox-active analytes that can be differentiated using chemometrics — have been used previously to differentiate between various analytes in food and the environment. However, their use to analyze biofluids is hindered by the abundance of redox molecules that biofluids contain and that generate overlapping and masking electrochemical signals. For example, these interfering signals make it difficult to distinguish between redox profiles of neurotransmitters biomolecules (dopamine and norepinephrine) in urine—fundamentals biomarkers for depression detection or schizophrenia treatment efficacy monitoring.

Here, we present a novel smart multi-electrode array for dopamine and norepinephrine simultaneous detection in urine. The main innovation in the presented array is our use of charged polymers and electrocatalytic materials to modify the surface of the electrodes and to produce a set of cross-reactive signals with the subsequent calculation of the cross-reactivity index for electrode combinations, predicting the optimal array for the chemometric model. Our methodology based on measuring sets of electrochemical signals from dopamine, norepinephrine, and uric acid (i.e., main interfering molecule) combinatorial mixtures that were processed with asymmetric least-square spline regression method to subtract the baseline signals. The baseline-subtracted signals of the entire electrodes array were combined to a super column vector that represents the array response, and were analyzed by using a leave one out cross validation to evaluate various electrode configurations and number of latent variables. The relationship between the electrode configuration cross validation results and the multiple analytes levels prediction was evaluated using a partial least square regression (PLSR) model. Using our novel methodology, we utilized the optimal electrode configuration of bare and chitosan polymer-modified electrodes to differentiate between dopamine and

norepinephrine levels in buffer solution at root mean square error values of 2.90 μM for dopamine and 2.37 μM for norepinephrine. Importantly, we utilized an array of bare and chitosan encapsulating electrocatalytic carbon nanotubes – modified electrodes to successfully differentiate dopamine and norepinephrine spiked in undiluted urine samples of three healthy volunteers (Pearson correlation coefficient of 0.93 for dopamine and 0.78 to norepinephrine between the known and the estimated concentrations). By further expanding the available electrochemical datasets to other neurotransmitters in urine, we will achieve improved model training in biological fluids, and will enable rapid and simultaneous detection of a spectrum of neurotransmitters in a single step.

T12. Distinction of cancer and healthy tissue using NIR-spectroscopy and multivariate data processing

Ekaterina Oleneva[1], Dmitry Kirsanov[1, 2], Andrey Panchenko[3], Ekaterina Gubareva[3], Viacheslav Artyushenko[4], Andrey Legin[1, 2]

[1] Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia

[2] Laboratory of artificial sensory systems, ITMO University, St. Petersburg, Russia

[3] Laboratory of carcinogenesis and gerontology, FSBI «Petrov Research Institute of Oncology»
of the Ministry of Healthcare of the Russian Federation, St. Petersburg, Russia

[4] art photonics GmbH, Berlin, Germany

In recent decades, cancer has become a dangerous threat for all age groups of people, especially for patients under 40 years old. This ominous trend poses a challenge to oncologists to cure cancer with minimal side effects, which can appear even in several years after the treatment (secondary tumors, heart diseases, infertility, digestion problems and stress disorders). Surgery is a common method of primary solid tumors treatment, and the volume of removed tissue defines the frequency and severity of side effects. A surgeon has to make a choice between radical removal of an affected organ (i.e., mastectomy) and organ-conservation surgery, which allows for better patient's recovery. In the latter case, the problem of dissection margins assessment has the highest priority. Nowadays, such type of surgery requires the intraoperative consultation of pathologist, who histologically evaluates the specimen (frozen sections) provided by a surgeon during the operation. Evidently, this method of

surgical margins evaluation is time-consuming and, in some cases (i.e. borderline tumors), less accurate.

The application of various spectroscopic methods for such clinical tasks has been extensively studied during the past decades [1]. The difference in cancer cells and normal cells metabolism allows for reliable tumor margins determination, for example, in near infra-red spectral region. In current work, we studied the possibility of tumor/healthy tissue distinction by their NIR spectra (939–1796 nm), processed by various multivariate classification methods. NIR measurements were performed on mice and rats tissue specimens, obtained after the surgery and stored in paraffin blocks, by the following steps:

1. Erlich carcinoma, inoculated on the skin/normal skin distinction (10 mice, 60 spectra). A sample has been attributed to malignant or normal one by visual evaluation.
2. Glioma/normal brain tissues distinction (56 rats, 238 spectra). Due to impossibility of visual tumor margins assessment, histological assessment of all 56 paraffin blocks has been performed.
3. Erlich carcinoma/hyperplasia or inflammation/normal skin distinction. Three-level classification of the data, obtained at the first step, and additional data set (59 mice, 228 spectra), included only benign transformations of normal skin tissue.

Among the variety of multivariate classification methods suitable for the task of linearly separated classes, linear support vector machines (SVM) method was the most efficient one. Independent validation set has been chosen for the models evaluation. For the first data set, the SVM model accuracy, sensitivity and selectivity attained the level of 100%. Other results will be reported during oral presentation on the conference. The obtained results show, that NIR spectroscopy is a powerful and simple method for reliable cancer/normal tissue intraoperative differentiation.

References:

1. V.R. Kondepati et al. Recent applications of near-infrared spectroscopy in cancer diagnosis and therapy, Anal Bioanal Chem 390 (2008) 125–139.

T13. Breakthrough in heparin analysis: holistic control by NMR spectrometry and chemometrics

Yulia B. Monakhova[a,b,c], *Bernd Diehl*[b]

[a]*Spectral Service AG, Cologne, Germany*

[b]*Institute of Chemistry, Saratov State University, Saratov, Russia*

[c]*Institute of Chemistry, Saint Petersburg State University, St Petersburg Russia*

Heparin and low molecular weight heparins (LMWHs) are the most widely used anticoagulant drugs during surgeries as well as for the treatment of thrombotic and cardiovascular disorders. The focus of the current study was to develop an analytical procedure based on high resolution (600 MHz) nuclear magnetic resonance (NMR) method combined with chemometric tools suitable for holistic control of qualitative and quantitative heparin parameters for pharmaceutical control.

First, NMR spectroscopy was used to distinguish heparin and LMWHs produced from porcine, bovine and ovine mucosal tissues as well as their blends. For multivariate analysis several statistical methods such as PCA, FDA, PLS-DA, LDA were utilized for the modeling of NMR data of more than 100 authentic samples. Moreover, the intra-species differences within the bovine heparin group were examined. Significant improvement of chemometric models was achieved by switching to 2D NMR experiments (heteronuclear multiple-quantum correlation (HMQC). The classification models were validated using representative independent test sets.

The full characterization of heparin and low molecular weight heparin (LMWH) also requires the determination of average molecular weight. To determine this characteristic, PLS was utilized for modeling of diffused-ordered spectroscopy NMR data (DOSY) with root mean square error of prediction of 498 Da and 179 Da for heparin and LMWH, respectively.

Moreover, retrospective multivariate analysis was performed on a big dataset of 990 NMR heparin spectra recorded over six years (2012-2017) in our laboratory. Several steps of statistical analysis of accumulated data were used to differentiate samples according to animal origin (bovine, porcine and ovine heparin), purity grade (crude and purified heparin), distributing company as well as to estimate their closeness to the heparin reference sample (SST)

provided by US Pharmacopeia. The projection of new samples on these models can automatically forecast of all mentioned qualitative heparin properties within one minute.

Holistic control with only one sample preparation according to Pharmacopeia using just four sequential NMR experiments (total measurement time of 20 min) combined with chemometrics provides purity, assay and provenience of heparin samples as well as enables correlation with biological activities and macroscopic values.

T14. Optical multisensor systems

Andrey Bogomolov
Blue Ocean Nova AG, Anton-Huber-Straße 20, 73430 Aalen, Germany
Samara State Technical University, Molodogvardeyskaya Street 244, 443100 Samara, Russia

Optical multisensor systems is a special class of analytical devices taking an intermediate position between the traditional single-channel sensing and full-scale spectrophotometry, thus combining the features of both. This presentation reviews and analyzes the development of multisensor systems over the last two decades, including both their technical achievements and applications. A close consideration is given to multivariate data analysis and specialized software being two important components of the multisensor analysis.

T15. Multi-mode fiber spectroscopy for cancer diagnostics

Anastasiia Melenteva[1], Valeria Belikova[1], Olga Bibikova[2], Urszula Zabarylo[2], Viacheslav Artyushenko[2], Andrey Bogomolov[1,3], Thaddäus Hocotz[4]
[1] Samara State Technical University, Samara, Russia
[2] art photonics GmbH, Berlin, Germany
[3] Blue Ocean Nova AG, Aalen, Germany
[4] Charité - Universitätsmedizin Berlin, Germany

Oncological diseases are among of the leading causes of death in the world. One of the most common cancers for several decades is tumors of the gastrointestinal tract. According to the data reported by the World Health Organization in 2015, the mortality from stomach cancer was 754,000 people in the year, from colorectal cancer – 774,000 people.

A characteristic sign of cancer is the rapid formation of abnormal cells that grow beyond their usual boundaries. To diagnose stomach and colorectal cancers, various modern methods of analysis are used, such as gastroscopy, fluoroscopy, ultrasound, computed tomography, colonoscopy and others. However, these methods are time-consuming, very expensive and their success depends on the correct diagnostics. Therefore, novel, time-saving, non-invasive and simple methods is a pressing need in the modern cancer diagnostics.

Optical spectroscopy can be successfully used for cancer diagnostics. Using fiber optic probes in fluorescence, Raman, middle (MIR) and near (NIR) IR regions is a very powerful and flexible method for non-invasive in vivo applications.

The present study is a part of a larger research project, the main purpose of which is the development of new optical techniques for tumor margin identification. The performance of fluorescence, MIR, NIR and Raman spectroscopy in the analysis of colorectal and stomach tissue for cancer lesion identification was assessed. Principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) were used to rank different spectroscopic methods in their ability to recognize health and tumor tissues.

T16. The use of FT-IR/ATR spectroscopy and PLS-DA in checking the authenticity of the aspirin tablets

Aleksey G. Zarubin[a,b], *Daniil I. Terekhov*[b]

[a]*Department of Oil and Gas Storage and Transportation, School of Earth Sciences & Engineering, Tomsk Polytechnic University, Tomsk, Russia*

[b]*Department of Analytical Chemistry, Faculty of Chemistry, Tomsk State University, Tomsk, Russia*

The one of the common drug is aspirin due to its effects on human body while treat fever, pain, and inflammation. The significant intake of aspirin tablets leads to evolution of its authentication and verification methods. The development of drugs authentication and verification for the anti-counterfeiting systems is of importance. The usage of NIR spectroscopy in qualitative analysis of the aspirin manufacturer was established previously [1].

The purpose of this study was to examine the possibility of using the method of IR spectroscopy in a limited middle range for qualitative authenticity of aspirin manufacturer.

In this work 72 samples of aspirin tablets of two manufacturers for 36 for each were examined. The FTIR spectrometer Nicolet iS10 (Thermo Scientific) was used as a measuring tool in mode ATR with diamond crystal. The spectral recording parameters were as follows: spectral range from 4000 to 600 cm-1, resolution – 4 cm-1, number of scans – 128, ATR-baseline correction. The spectral data of each manufacturer's aspirin was divided into a calibration set (24 samples) and a test set (12 samples). The range of the spectrum was limited to the interval from 2920 to 2800 cm-1, which characterizes the vibrations of the CH groups, for the convenience of processing. The PCA method was applied to a limited data interval for dimension reduction. From the diagram PC1–PC2 it was established that the classification of tablets of aspirin by manufacturer is possible. Further on the calibration set a PLS-DA model was built for the classification in the Microsoft Excel using the CHEMOMETRICS library [2]. The test showed single cases of errors of the first and second kind when using two latent variables and the absence of errors when using three latent variables.

The studies revealed that IR-spectroscopy in a limited middle range with the PLS-DA method is suitable for qualitative authenticity of aspirin manufacturer.

References

1. Balykova, K. S., et al. "Analysis of tablets of acetylsalicylic acid by near-infrared spectroscopy." Bulletin of Roszdravnadzor 2 (2013).

2. Pomerantsev A. L. Chemometrics in Excel, John Wiley & Sons, 2014.

T17. Multimodal image analysis for tissue diagnosis of skin melanoma

S. Guo[1, 2], S. Pfeifenbring[1, 2], T. Meyer[1, 2], G. Ernst[2, 3], F. von Eggeling[1, 2, 3], V. Maio[4], D. Massi[4], R. Cicchi[5, 6], F. S. Pavone[6, 7], J. Popp[1, 2], T. Bocklitz[1, 2]

[1]Institute of Physical Chemistry and Abbe Centre of Photonics, University Jena, Germany

[2]Leibniz Institute of Photonic Technology, Jena, Germany

[3]Department of Otorhinolaryngology, Jena University Hospital, Jena, Germany

[4]Department of Medical and Surgical Critical Care, University of Florence, Florence, Italy

[5]National Institute of Optics-National Research Council (INO-CNR), Sesto Fiorentino, Italy

[6]European Laboratory for Non-Linear Spectroscopy (LENS), University of Florence, Sesto Fiorentino, Italy

[7]Department of Physics, University of Florence, Sesto Fiorentino, Italy

Melanoma has become a common cancer in the western world, with an average increase of 4.6% from 1975 to 1985 and 2.7% from 1986 to 2007 for annual incidence [1]. Histopathological inspection has remained to be the 'gold standard' method for the diagnostics of melanoma. To do so, the skin tissue has to be sectioned, stained, and imaged via microscopy. Thereafter a pathologist can conduct the diagnosis based on the microscopic images and the visual appearance of the tissue. However, the requirement of skin biopsy makes it unsuitable for in vivo diagnosis. The cumbersome procedure also leads to long time-delay from the resection to diagnosis. More importantly, the diagnosis is dependent on the pathologist's experience and different pathologists may give controversial diagnoses for the same skin section. An alternative diagnostic approach is multimodal imaging, the combination of two-photon excited fluorescence (TPEF) and second harmonic generation (SHG). It is label-free, noninvasive, and provides molecular contrast, which makes it ideally suitable for in vivo measurements and clinical cancer study [2]. However, the bottleneck to apply this technology in clinical practice is to translate the optical signal into high-level diagnostic relevant information.

Here we present a data pipeline translating multimodal image into information relevant for diagnostics of skin melanoma. As a first step, the images were pre-processed with the combination of wavelet and Fourier transform to remove the mosaicking artifacts caused by uneven illumination during the measurement. Background exclusion and intensity standardization were performed in the next step. Thereafter, texture features based on first-order histogram features and the local gray-level co-occurrence matrix (GLCM) features were extracted. The calculation was performed in local mode using moving window with different window sizes. The extracted features characterized the morphology of the skin tissue in multiple scales. Based on these features, three tissue types (normal epithelium, melanoma, and other tissues) were classified with a hierarchical statistical model, which was built in local mode and its performance was verified with leave-one-image-out cross-validation. The satisfactory tissue type prediction demonstrated the great potential of multimodal imaging combined with image analysis to differentiate tissue types. In addition, the global-mode classification was constructed to relate the whole image with the clinical diagnosis, of which the prediction indicated a fair agreement with the clinical diagnosis. To summarize our study, we compared the performance of the first-order histogram and GLCM based texture features on the basis of the Fisher's discriminant ratio (FDR) as well as the results of the classification, which illustrated the superiority of the first-order histogram features to the GLCM based texture features.

Acknowledgements

Financial supports of the EU, BMBF, German Science Foundation (BO 4700/1-1, PO 563/30-1, STA 295/11-1), the Fonds der Chemischen Industrie, the Carl-Zeiss Foundation and Leibniz association via the ScienceCampus 'InfectoOptics' for the project "BLOODi", the scholarship for S.G. from the "China Scholarship Council" (CSC), and the grant for T.B from Raman4Clinics are greatly appreciated. This work was partially supported by Italian Ministry for Education, University and Research in the framework of the Flagship Project NANOMAX, by Italian Ministry of Health (GR-2011-02349626), by EC Horizon 2020 research and innovation programme under grant agreement No 654148 Laserlab-Europe and by Ente Cassa di Risparmio di Firenze.

References

1. Rigel, D. S., Russak J., Friedman R., CA: a cancer journal for clinicians, 2010, 60(5): 301-316.

2. Vogler, N., et al., Annual Review of Analytical Chemistry, 2015. 8: 359-387.

T18. Near infrared study of hydration state of the human body through skin

J. Elek[1], E. Markovics[2], Gy. Kovács[3]

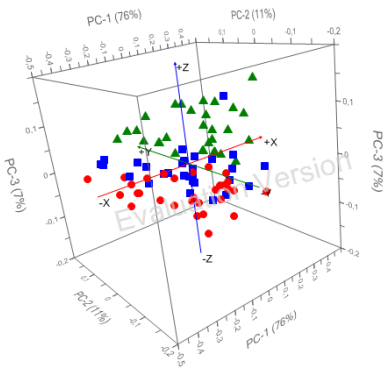
[1] Sceince Port Kft, Hungary

[2] University of Debrecen

[3] CTS Kft,

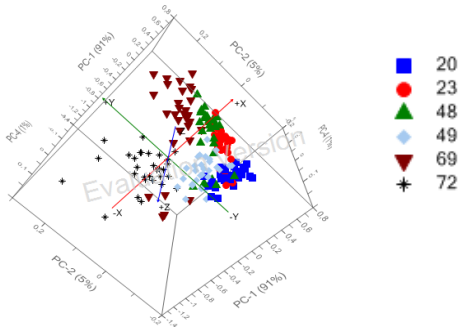
The daily water intake is a very important in all ages, but has a pronounced importance in case of children and elderly persons. This study is very far yet from any medical applications, but the basic idea is to monitor the hydration level of the human body by near infrared examination of the human skin.

Therefore spectra from the arm of male and female volunteers of different ages were recorded and evaluated. The volunteers were asked not to drink for at least 2 hours before recording the first spectrum then a glass of water (200 ml) was consumed. After 10 minutes the second spectrum was recoded and the patients drunk one more glass of water. After 10 minutes the third spectrum was also recorded.



As the figure above shows the effect of drinking is pretty much visible: the green triangles represent the thirsty state, blue squares represent the first drink while the red dots represent the state after the second drink. These groups form independently from the sex and the age of the persons, but further evaluation of

the dataset reveals interesting facts. By plotting the appropriate principal components the spectra of the volunteers show grouping either based on sex, but more interestingly based on age too.



T19. Estimating rotational ambiguities in MCR: a quest for quantitative measures

Alexey N. Skvortsov
Biophysics department, Peter the Great St.-Petersburg Polytechnic University, St-Petersburg, Russia
Laboratory of cell Signaling and transport, Research Institute of Influenza, St-Petersburg, Russia

Multivariate curve resolution (MCR) is a set of techniques, which are widely used in modern chemometrics for decomposition of data matrices into chemically significant factors, e.g. spectra and concentrations of individual components in chromatography. The decomposition is searched to satisfy a set of physicochemical constraints. If the constraints are not sufficiently strong, or the problem is close to rank-deficient one, the solution may be far from being unique. This is a well known issue of MCR, known as the residual rotational ambiguity (RA). In case of large RA, a single MCR solution has little practical value, and it may even lead to overconfident conclusions. So good MCR techniques are to numerically evaluate the extent of RA and report it in comprehensible format.

For comparing different models and estimation of the constraint strength some quantitative measure of RA is of great interest. One of the oldest approaches for estimating RA is the geometric construction of areas of feasible solutions (AFS). For 3-component nonnegative factorization it has begun from the analytical approach, proposed by Borgen and Kowalski. In the last decade, many effective

methods have been proposed, which extend AFS approach to noisy data, other constraints, and more components (simplex tracking, polygon inflation, bounding rectangles etc). Other approaches for estimating RA also exist (MCR-BANDS, MCR-FMIN etc). While AFS provide a good graphical display of RA, they have some disadvantages. First, the distance between the points in AFS-space is not simply related to the similarity of spectra/concentrations. Second, the size and shape of AFS depends on the selection of norms and, for noisy data, on formulation of constraint strengths. Third, AFS boundaries are not themselves solutions. Consequently, geometric properties of AFS are poor candidates for RA measurement.

In the present work, we tested some combinations of approaches for semi-quantitative measurement of RA, which would be comparable to AFS by graphical visibility but would be less sensitive to norm selection and realizations of constraints. For that we used (1) previously developed MCR based on charged particle swarm optimization (cPSO-MCR), and (2) spherical coordinates for generating candidate solutions. Particle motion equations of PSO were modified to describe the motion of particles on the sphere. The selection of spherical coordinates, which is probably the most natural way to describe rotational ambiguity, was found to be very beneficial to cPSO-MCR method, as the search space has become full, continuous and finite (plane cPSO had problems with runaway particles). cPSO-MCR maximizes the dissimilarity of a set of candidate solutions, which is converted to the repulsion force of particles. A statistic of the dissimilarity in the stabilized cPSO swarm (e.g. mean or maximum) may then be viewed as some measure of RA. Previously we used Euclidean distance between normalized rotation matrices as a measure of dissimilarity. It was found to be poor selection for RA quantification. In spherical coordinates the principal angles may be utilized. However, the dissimilarity may be calculated directly from pairs of candidate spectra and concentrations. Both approaches were tested on simulated and real data. The approaches themselves were able to work, and produced some interesting results when compared to AFS. Still the overall success was limited. It was found that the dissimilarity measure influenced the results even more than the constraint efficiency. That led us to detailed re-evaluation of what is the dissimilarity of spectra and concentrations,

what is constraint power, and how they could be specified in the presence of scale ambiguity. The results of this analysis are briefly reviewed.

Several developments of significance-based variable selection, which could reduce the calculation burden for large noisy datasets, are also discussed.

T20. Distance estimation between objects in spectral data analysis

V. Belikova[1], A. Bogomolov[1,2]

[1] Samara State Technical University, Samara, Russia

[2] Blue Ocean Nova AG, Aalen, German

Spectrum comparison is an important task of the spectral data analysis. Mathematically similarity can be defined as a distance in a chosen space that can the raw space of the spectral variables or a factors space, e.g. in principal component analysis (PCA). There are several methods typically used for the distance calculation and the result depends on the chosen function. When the data includes classes, their analysis is comparing two or more ensembles of objects (i.e. spectra) and their mutual distances.

Pairwise comparison of spectra is required in different applications, such as spectral database search, comparison of curve resolution methods, search of neighbors (e.g. in KNN), distance matrix calculation in some clustering methods, and for the outliers detection. The classical distance functions typically used in practice are Euclidean or Mahalanobis distances and cosine. For the correct work of algorithms, the function should take relevant features of the data into account. For example, different background intensity and the variable correlation within peaks are the features typical for spectral data. They allow representing a spectrum as a curve, which is characterized by the shape and scale; both of them may contain relevant information. In the present report, we are going to suggest a flexible distance function that allows tuning the measured distance values toward the shape or scale depending on an application.

Dimensional reduction used by different methods, such as PCA and partial least-squares (PLS) regression can be very helpful for the analysis of spectral ensembles. In this case, the groups of points on the score plots of the first two or three principal components are considered. One of the most common practical

problems of this type is to describe the "separation degree" of two groups of points corresponding to different object classes. To solve this problem, any measure of clustering is typically used, e.g., the distance between the group centers (i.e., Davies-Bouldin index). However, for a quantitative model comparison, it is also necessary that the class separation measure reflects both the distance and the class intersection degree. An appropriate measure of class separation will be proposed and discussed.

T21. Large-scale comparison of similarity metrics for molecular and interaction fingerprints

Dávid Bajusz, Anita Rácz, Károly Héberger
Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

Molecular fingerprints are ubiquitously applied to represent molecular structure in a wide range of applications in cheminformatics, computational drug discovery and related fields [1]. Their greatest advantages are their compactness and machine-readability: the latter enables the fast comparison of molecular structure and the quantification of their similarity. Similarly, interaction fingerprints encode information about protein-ligand complexes (highly relevant in drug discovery) in an equally compact manner [2].

However, there are a great number of methods for calculating the similarity of two such binary data structures and – despite the general preference of the cheminformatics community towards the Tanimoto coefficient – the choice of similarity metric is not trivial.

Recently, we have shown with robust statistical methods on a large dataset that the Tanimoto coefficient is a justified choice from a small pool of commonly known and easily available similarity metrics – although other, equally consistent metrics could be identified as well [3]. In 2012, Todeschini and coworkers have compared a greater number of similarity metrics using a different methodology on simulated and real virtual screening datasets [4].

Currently, we are extending our methodology to a larger pool of similarity metrics [1,4] and molecular, as well as interaction fingerprints, implemented in various cheminformatics and modeling packages. Similarity metrics will be

compared and ranked based on their consistency with a suitable reference method (data fusion). We also plan on releasing an open-source Python module implementing the studied similarity metrics, which will be readily applicable with the Cinfony toolkit, the open-source aggregator of cheminformatics software [5].

Acknowledgement

The authors thank the support of the National Research, Development and Innovation Office of Hungary (OTKA contracts K 119269 and KH-17 125608).

References

1. Bajusz, D.; Rácz, A.; Héberger, K. Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In: Comprehensive Medicinal Chemistry III; Chackalamannil, S.; Rotella, D.P.; Ward, S.E., Eds.; Elsevier: Oxford, 2017; pp. 329–378.

2. Vass, M.; Kooistra, A.J.; Ritschel, T.; Leurs, R.; de Esch, I.J.; de Graaf, C. Molecular Interaction Fingerprint Approaches for GPCR Drug Discovery. Curr. Opin. Pharmacol., 2016, 30, 59–68.

3. Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? J. Cheminform., 2015, 7, 20.

4. Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. J. Chem. Inf. Model., 2012, 52, 2884–2901.

5. O'Boyle, N.M.; Hutchison, G.R. Cinfony – Combining Open Source Cheminformatics Toolkits behind a Common Interface. Chem. Cent. J., 2008, 2, 24.

T22. How different water activities affect rice germ shelf life: an aquaphotomics approach

Cristina Malegori[a], Paolo Oliveri[a], Roumiana Tsenkova[b], Carola Cappa[c], Mara Lucisano[c]

[a]DIFAR Department of Pharmacy, University of Genova, Genova, Italy

[b]Bio Measurement Technology Lab, Kobe University, Kobe, Japan

[c]DeFENS Department of Food, Environmental and Nutritional Sciences, Università degli Studi di Milano, Milano, Italy

The aim of this study is to investigate how different water activities affect rice germ shelf life. In fact, this matrix (a by-product of rice milling process) could be interesting for human nutrition but, for its composition characterized by unsaturated fatty acids, it undergoes rancidity during storage. Dried samples at different water activities (0.55, 0.45 and 0.36) were packed in air and stored at 27°C for 320 days (for a total of 7 sampling points). All the samples were analysed by FT-NIR spectroscopy in reflectance (in the 800 – 2780 nm spectral range) with a rotating sample holder, as a non-targeted analytical approach.

First of all, focusing on the Aquaphotomics approach, spectral analysis was done in the water first overtone range, 1300 and 1600 nm; then an exploratory principal component analysis (PCA) was performed, followed by a partial least squares regression method (PLSR) to understand the NIR spectral changes that characterize the differences caused by different water activity levels in rice germ during storage. In more depth, four wavelengths, essential for the description of this system, were found to define the so called water matrix coordinates (WAMACS): 1343 nm, associated with protonated water, 1392 nm, typical absorbance of trapped water, 1410 nm, the well-known band of free water and 1436 nm, the Zundel cation band (H_5O_2+) and the respective WAter Spectral Patterns, WASP. Radial graphics of the normalised WAMACS absorbance values were built: such Aquagrams allow to understand the modification of different water molecular the structures along time, for each water activity under investigation.

Thanks to this state-of-the-art approach, the water molecular conformation changes related to different water activities in rice germ along the storage were

discovered. These findings will open the venue of understanding the water molecular structure behind water activity in general.

Acknowledgment: the authors thank 'Rondolino - Società Cooperativa Agricola' for samples supply and economic support.

T23. Multi-spectral fiber spectroscopy methods for guided diagnostics of abdominal cancer

Olga Bibikova[1], Valeria Belikova[2], Anastasia Melenteva[2], Urszula Zabarylo[3], Iskander Usenov[1,4], Tatiana Sakharova[1], Andrey Bogomolov[2,5], Viacheslav Artyushenko[1]

[1] art photonics GmbH, Berlin, Germany

[2] Samara State Technical University, Samara, Russia

[3] Charité-Universitätsmedizin, Berlin, Germany

[4] Technical University of Berlin, Berlin, Germany

[5] Blue Ocean Nova AG, Aalen, Germany

Early diagnosis and treatment are currently the recommended management strategy for cancer therapy: however, the current "gold standard" for diagnosis (clinical examination of the suspicious tissue, followed by biopsy and histopathology) is invasive, time consuming and strongly depends on "human factor". Therefore, noninvasive spectroscopic investigation becomes a perspective alternative for label-free cancer specification. Moreover, opportunities of fiber optics, including non-destructiveness and non-invasiveness, provide a significant diagnostic potential for various types of the human cancer ex-vivo and in-vivo.

Our development of Multi-Spectral Fiber (MSF-) system enables to test various single and combined fiber probes for four key spectroscopy methods: Raman scattering, FTIR-absorbance, diffuse NIR-reflection, and fluorescence – to select the best method or their best combination for a detection of abdominal cancer tissues. Detailed comparison of accuracy, sensitivity and specificity for tissues determination were done by multivariate data analysis of spectroscopic data to build prediction models for each method. In addition, diagnostic capabilities of fluorescence and FTIR spectroscopy combination have been investigated using matching pairs of normal and malignant biopsy samples of patients with kidney

and colon cancer. Based on synergic gain of fiber-based FTIR-ATR and fluorescence spectroscopy, combined fiber probe was developed. The probe allows to contribute to the efficient detection and monitoring of tumor spread, determine successful diagnosis and realistic cancer prognosis.

T24. Domain-invariant Partial Least Squares (di-PLS) Regression: a novel method for unsupervised and semi-supervised calibration model adaptation

Ramin Nikzad-Langerodi[1], Werner Zellinger[1], Edwin Lughofer[1], Thomas Reischer[2], Susanne Saminger-Platz[1]

[1]Department of Knowledge-Based Mathematical Systems, Johannes Kepler University, Linz, Austria

[2]Metadynea GmbH, Krems, Austria

Multivariate calibration models often fail to extrapolate beyond the calibration samples due to (small) changes associated with the instrumental response, environmental condition or sample matrix. Different methods have been developed in the past to adapt a source calibration model to a target domain, mostly in the context of calibration transfer from one instrument to another [1,2]. However, most of these methods are tailored to solve one particular task related to model adaptation, while a unified model adaptation framework is largely missing.

To fill this gap, we here introduce domain-invariant partial least squares (di-PLS) regression. In particular, we introduce a domain regularizer into the PLS objective in order to align first and second order statistics of source and target domain data while constructing the latent variable subspace (Figure 1). We show that a domain-invariant weight vector can be derived in closed-form, which facilitates implementation and keeps computational costs low. Furthermore, our method is flexible in the sense that it can cope with (partially) labeled data in the source and target domain as well as fully unlabeled data in the latter underpinning its uniqueness among model adaptation techniques so far proposed in chemometrics. We apply di-PLS on a simulated data set where the aim is to desensitize a source calibration model to an unknown interferrent in the target domain (i.e. unsupervised model adaptation). In addition, we

investigate unsupervised and semi-. supervised model adaptation on a real-world FT-NIR data set from a melamine formaldehyde condensation process.

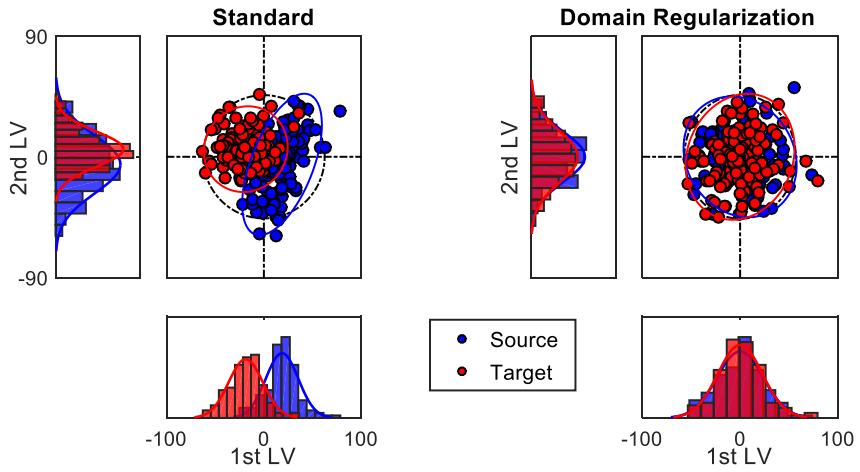


Figure 1. Distribution of the projections (scores) from source (blue) and target (red) data in a two-dimensional latent variable (LV) subspace without (left) and with (right) domain regularization.

Acknowledgement

This work was funded by the Austrian research funding association (FFG) under the scope of the COMET program within the research network "Industrial Methods for Process Analytical Chemistry" (imPACts) (contract #843546). This publication reflects only the authors' views.

References

1. Yongdong. Wang, David J. Veltkamp, and Bruce R. Kowalski**,** *Analytical Chemistry* **1991** *63* (23), 2750-2756

2. John H. Kalivas, Gabriel G. Siano, Erik Andries, Hector C. Goicoechea, *Applied Spectroscopy* **2009,** 63 (7), 800 - 809

P01. Big data analysis and comprehensive analytical control of fertilizers

Yunovidov Dmitry, Sokolov Valery and Bahvalov Alexey
Research Institute for Fertilizers and Insectofungicides, Cherepovets, Russia

Modern industry is a complex and multivariate process, where is no universal and high quality analytical control without taking into account of many physical and chemical properties of studied samples [1, 2]. In the proposed research, big data mining and exploration for industrial produced mineral fertilizers was described. The technique was developed using optical logger and energy disperse X-ray fluorescence spectrometer.

About 7 grades and 600 samples of industrially produced complex mineral fertilizer was processed. Each object was prepared by pressing of three types of samples: granules, powder with fraction less than 500 μm and powder with fraction less than 100 μm. Moreover, algorithms of classification and regression in Python 2.7 programming language was constructed and quality metrics for classification (F-metric – harmonic mean of precision and recall) and regression (coefficient of determination) was calculated (table 1).

Table 1. Big data analysis of allocated fertilizers properties

	nitrogen	**phosphorus**	**potassium**	**sulfur**	**fraction**	**is dry**
values	[0; 16]	[15; 52]	[0; 20]	[0, 20]	[granules, 500 μm, 100 μm]	[yes, no]
Classification, F-metric (%)						
linear	99,31	99,78	99,59	99,56	92,40	72,94
linear with L1*	99,65	99,78	99,57	98,87	92,51	73,08
linear with L2*	99,65	99,78	100,0	98,99	91,33	68,46
random forest	100,0	100,0	100,0	98,99	98,40	73,37
Regression, R2 (%)						
linear	98,78	98,00	99,61	98,17	-	-
linear with L1*	97,04	97,10	99,26	95,79	-	-
linear with L2*	98,78	97,70	99,61	98,17	-	-

* Type of regularization

With obtained results, the possibility of indirect determination of nitrogen content with ED XRF was showed. Furthermore, high accuracy of comprehensive

analytical control (generally more than 98% for selected quality metrics) was achieved.

References

1. Hasikova J. et al. On-Line XRF analysis of phosphate materials at various stages of processing // Procedia Eng., 2014. Vol. 83. P. 455–461.

2. Yunovidov D.V., Sokolov V.V., Bakhvalov A.S. The Use of the Sample Spectrum for Assessing the Impact of Different Stages of the NPKS Fertilizer Preparation on the Results of X-Ray Fluorescence Analysis // Zavodskaya Laboratoriya. Diagnostika Materialov, 2017. Vol. 83, № 9. P. 15-21.

P02. Revealing of counterfeit tablets among antihistamine medicines. NIR-based approach and other techniques

A.V. Titova[1,2], K.S. Balyklova[1,3], O.Ye. Rodionova[1,4], A.L. Pomerantsev[4,5]

[1]Information and Methodological Center for Expertise, Stocktaking and Analysis of Circulation of Medical Products, Roszdravnadzor, Moscow, Russia

[2]Pirogov Russian National Research Medical University, Moscow, Russia

[3]I.M. Sechenov First Moscow State Medical University, Moscow, Russia

[4]N.N.Semenov Institute of Chemical Physics RAS, Moscow, Russia

[5] Branch of Institute of Natural and Technical Systems RAS, Sochi, Russia

An exemplary analysis of suspected drugs which have the same designation as the brand of a widely used medication for treating allergies is presented. For a rapid testing the Near Infrared (NIR) measurements accompanied with chemometric data processing, that is the NIR-based analysis [1], is applied. A model that had previously been developed and stored in a library for an everyday monitoring in drugstores is used for recognition of the counterfeits.

We also discuss the procedure of the library model development that involves comparison of a target medicine with its analogues produced by the other manufacturers. This study demonstrates the importance of the model validation against similar but still alien objects [2], because this procedure trains the model for recognition of counterfeits of various grades, not only rough and evident, but also ‘the high quality’ ones. It is shown that in some cases the NIR-based analysis was more specific than laboratory tests in revealing counterfeits.

Additionally, a new instrument, VisCam, which is applied for a visual analysis of the primary and secondary packages is presented. It is shown that VisCam helps revealing hidden violations in the primary and secondary packages.

References

1. O.Ye. Rodionova, A.L. Pomerantsev, NIR based approach to counterfeit-drug detection, Tr. Anal. Chem., 29, 781-938 (2010).

2. O.Ye. Rodionova, K.S. Balyklova, A.V. Titova, A.L. Pomerantsev, Quantitative risk assessment in classification of drugs with identical API content, J/Pharm. Biomed. Anal.98 (2014), 186-192.

P03. Software implementation of the Hard and Soft Partial Least Squares Discriminant Analysis

Y.V. Zontov[a], S.V. Kucheryavskiy[b], O.Ye. Rodionova[c,d], A.L.Pomerantsev[c,d]
[a]National Research University Higher School of Economics (HSE), Moscow, Russia
[b]Depatment of Chemistry and Bioscience, Aalborg University, Esbjerg, Denmark
[c]Semenov Institute of Chemical Physics RAS, Moscow, Russia
[d]Branch of Institute of Natural and Technical Systems RAS, Sochi, Russia

We present a software implementation of hard and soft approaches to Partial Least Squares Discriminant Analysis (PLS-DA) [1]. The Soft PLS-DA is based on Quadratic Discriminant Analysis applied to the super-score matrix T. It can simultaneously attribute a sample to several classes. It also allows to detect samples, which are not members of any training classes, and, therefore, reduce a number of false positives e.g. in the presence of outliers. The Hard PLS-DA represents mostly the conventional PLS-DA classification. Thus the software can be used for both multi-class as well as two-class classification.

Both methods are implemented as MATLAB toolbox, PLS-DA Tool, using object oriented programming. The toolbox provides instruments for data pre-processing as well as for interpretation, validation and visualization of classification models. The main class, *PLSDAModel*, is responsible for the logic and contains implementation of both methods and auxiliary algorithms. The instance of this class has fields, which represent the actual model, and methods for data visualization and statistics. The *PLSDAGUI* class provides graphical user interface, where a user can create and manipulate datasets, calibrate, validate

and explore models interactively. This class encapsulates all the necessary data such as training or validation sets and labels of the samples and variables, and uses these elements for the model development and validation. The PLS-DA Tool has its own implementation of all necessary statistical functions and does not require the MATLAB Statistics Toolbox.

References

1. A. Pomerantsev, O. Rodionova, Partial least squares discriminant analysis: Taking the right way – A critical tutorial. J. Chemometrics, 2017, accepted.

P04. Integral quality parameters determination in plastics with XRF spectrometry

V.V.Panchuk[1,2], D.O.Kirsanov[1,3], A.V.Legin[1,3], V.G.Semenov[1,2]

[1]Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia

[2]Institute for Analytical Instrumentation RAS, St. Petersburg, Russia

[3]Laboratory of artificial sensory systems, ITMO University, St. Petersburg, Russia

Energy-dispersive X-ray fluorescence spectrometry (EDX) is widely employed modern elemental analysis method. Simple sample preparation, non-destructivity, multielemental capabilities and short analysis time are advantages of EDX. These circumstances led to a broad application of the method in material and environmental science, biology, chemistry, industrial process monitoring, etc. However, most of the serial production EDX spectrometers cannot allow for determination of elements with atomic numbers below 11 (sodium). Thus EDX can hardly be applied for analysis of organic materials. The spectrum contains fluorescence lines of the elements composing the sample, but also contains the signal from reflected X-ray tube radiation which consists of deceleration radiation and elastic/inelastic scattering of cathode radiation. The intensity of scattered radiation (especially for inelastic scattering) depends on average mass absorption coefficient of the sample, which in turn depends on elemental composition. Thus, scattered radiation can be employed as a source of information on integral properties of the samples as determined by its' "average" molecular mass. A general feasibility of studying various light elements through the analysis of scattered radiation was already shown in the

literature [1], although lack of selectivity was the reason for significant imprecision.

The present study deals with classical chemometric tool – PLS regression, in order to address this problem. The quantification of light elements (O, H, C) and integral physical and chemical properties of various plastics will be shown as a feasibility example of the approach.

References

1. Kalinin B.D., Plotnikov R.I., Rechinsky A.A. On the possibility of determining the composition of organic compounds by the intensity of scattered X-ray radiation // Analitika i Kontrol. 2011(2), p. 163-169 [In Russian]

P05. Similarity metrics under the magnifying glass: comparative study on metabolomic fingerprints

Anita Rácz[1], Filip Andrić[2], Dávid Bajusz[3], Károly Héberger[1]
[1]Plasma Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Hungary
[2]Faculty of Chemistry, University of Belgrade, Belgrade, Serbia
[3]Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Hungary

Contemporary metabolomic fingerprinting is based on multiple spectrometric and chromatographic signals, used either alone or combined with structural and chemical information of metabolic markers at the semiqualitative level. However, signal shifting, convolution, and matrix effects may compromise metabolomic patterns. Recent increase in the use of qualitative metabolomic data, described by the presence (1) or absence (0) of particular metabolites, demonstrates great potential in the field of metabolomic profiling and fingerprint analysis.

The aim of this study is a comprehensive evaluation of binary similarity measures for the elucidation of patterns among samples of different botanical origin and various metabolomic profiles.

Nine qualitative metabolomic data sets covering a wide range of natural products and metabolomic profiles were applied to assess 44 binary similarity measures for the fingerprinting of plant extracts and natural products. The

measures were analyzed by the novel Sum of Ranking Differences method (SRD), searching for the most promising candidates.

SRD and analysis of variance (ANOVA) revealed that Baroni-Urbani-Buser (BUB) and Hawkins-Dotson (HD) similarity coefficients were the most consistent measures while Dice (Di1), Yule (Yu), Russel-Rao (RR), and Consonni-Todeschini 3 (CT3) were ranked the worst. ANOVA revealed that concordantly and intermediately symmetric similarity coefficients are better candidates for metabolomic fingerprinting than the asymmetric and correlation based ones. The fingerprint analysis based on the BUB and HD coefficients and qualitative metabolomic data performed equally well as the quantitative metabolomic profile analysis.

Fingerprint analysis based on the qualitative metabolomic profiles and binary similarity measures proved to be a reliable way in finding patterns in metabolomic data.

Acknowledgement

The project was supported by the National Research, Development and Innovation Office of Hungary under grant numbers K 119269 and KH_17 125608.

P06. Development of methods for solving some problems in hydrology and hydrochemistry using PCA-modeling

Tatiana Gubareva and Boris Gartsman
Water Problems Institute, Russian Academy of Sciences, Moscow

The aim of the project is to develop the general principles and methods for solving the problems of hydrology and hydrochemistry based on the PCA modeling using the example of the following:

1. the identification of water sources of river flow (end-members) and drainage mechanisms in small watersheds on the basis of the mixing model with application of hydrochemical tracers, quantitative assessment of sources;
2. the development of an effective scheme for the classification of river basins for landscape-hydrological zoning of territories;

3. the detection of types (classification) of dissolved organic matter (DOM) and their quantitative evaluation based on optical indices in natural waters from fluorescence spectroscopy data.

As a result of the implementation of the project, the accumulation of experience in the application of PCA and the development of methodological recommendations for ensuring more active use of PCA-modeling in the field of hydrology, hydrochemistry and related fields of Earth sciences are expected.

The target results for each of the tasks are:

For 1 task. Adaptation and experience of application of modern methods of analysis of natural tracers and determination of the number of water sources. Development and testing of a geochemical model for mixing four sources (End-Member Mixing Analysis model) for quantitative estimation of runoff components.

For 2 tasks. Development of an effective scheme for the classification of river basins according to a set of morphometric, structural-hydrographic, landscape characteristics for the identification of objectively existing regions similar in physicogeographical and hydrological characteristics. An effective scheme is understood as an accessible set of input information, an adequate set of methods of mathematical analysis, a clear algorithm for their implementation.

For 3 tasks. Adaptation and testing of the model of "parallel factor analysis" (PARAFAC model) on the data of fluorescence spectroscopy of water samples collected in a river basin.

The presentation of this project will reflect the results for each task.

P07. Chemometrics assisted spectroscopic determination of aloin in Aloe Vera extract

Natalia A. Burmistrova, Olga A. Krivec, Yulia B. Monakhova
Institute of Chemistry, Saratov State University, Saratov, Russia

Aloin is a natural bioactive compound which is the main useful component of Aloe Vera and is widely used in medical drugs and alcoholic beverages production. The using in practice natural aloin extract is characterized by unstable aloin concentration and complex multicomponent biological matrix

(included organic acids, saccharides, volatile, flavonoids, minerals etc.). Thus the high performance and selective methods for aloin monitoring in nature materials must be developed. Atomic spectroscopy and chromatography are the most commonly used laboratory methods for this purpose. Nevertheless, the application of electronic spectroscopy to aloin determination can be alternative approach in terms of low cost screening. However, the overlapping of the matrix components bands in electronic spectra prevents the aloin determination without prior separation. Obviously, a multivariate data processing of spectral data is a particular interesting to solve this problem.

The successful application of Mutual Information Least Dependent Component (MILCA) algorithm to decomposition of spectra as well as to determination of aloin concentration in two- and three-component solutions contained organic acid choose as model systems has been demonstrated.

The spectroscopy studying of Aloe Vera extracts from different plant parts which contained aloin at concentration range 0-150 ppm was carried out. First the Principal component analysis (PCA) of spectral dataset was performed and the separation of samples for groups in score plot corresponding of type samples was observed.

Further MILCA algorithm which allow to perform “blind” source separation and have the advantage of `independent' component analysis was used for determination of aloin concentration in Aloe Vera extracts and crucial importance of calibration system optimization to minimize the influence of matrix components and to successful decomposition of the system has been shown.

The work was supported by the Russian Ministry of Science and Education (project 4.1063.2017).

P08. Sorting for special cereal-based products needs new quality analysis standards

Kapustin D., Zhilin S.
Altai State University, Barnaul, Russia

Making cereal-based products with special characteristics and ultimate quality often requires a raw material with special properties. For instance, one needs

high vitreousity wheat to produce premium pasta. In some cases it even demands to discard grains with relatively low vitreousity in high-vitreousity sorts of durum wheat. Modern optical sorters enable a user to precisely separate grains on light transmittance level. But a paradox consists in the impossibility of precise evaluation of sorting quality (wheat vitreousity for obtained fractions) using official techniques stated in Russian national standards (GOST).

Standard technique is based on grain color evaluation by an expert for sample consisting of 100 grains. An expert is prescribed a) to divide sample grains into three categories: vitreous, semi-vitreous and non-vitreous, b) to count grains in categories and c) to compute vitreousity index by averaging these counts with weights 1, 0.5 and 0 correspondingly. According to our double blind experiments the difference of vitreousity index values obtained by qualified experts can reach about 20% for the same sample and same person in different attempts and up to 45% for different experts and the same sample.

The replacement of subjective expert evaluations by instrumental measurements seems to be a natural way to overcome these difficulties. Towards this end we developed an experimental stand for massive grains imaging and single kernel analysis. Imaging of grains can be performed both for reflected and transmitted light. It is possible to vary sample size from single kernel up to hundreds of thousands. Different properties of grains can be studied by analyzing of obtained images: colorimetric, granulometric, etc.

Series of experiments performed using this imaging setup show that image-based vitreousity measurements are highly reproducible and after calibration and standardization may provide reliable and rapid technique for vitreousity (and some other quality characteristics) evaluation. It is important that such measured properties can be directly employed for fine and flexible tuning of raw material sorting to manage quality and features of special products.

P09. Semi-empirical formula for mean excitation energy: interval approach

Vladimir A. Smolar[a], Ilya I. Maglevanny[b], Sergei I. Zhilin[c]
[a]Volgograd State Technical University, Volgograd, Russia
[b]Volgograd State Social Pedagogical University, Volgograd, Russia
[c]Altai State University, Barnaul, Russia

It is shown that the electronic configuration of atoms affects significantly the dependence of the mean excitation energy of substances I from its atomic number Z, and this dependence corresponds to the extended periodic table of elements. Assessment of this dependence using the method of Thomas-Fermi offers $I \approx 1.58Z^{4/3}$ eV. This evaluation is in agreement with NIST data [1, 2] and take place for atoms of *s*, *g*, *f*, *d* blocks starting from the fourth period. For atoms of *p* blocks, when filled outer shells, the dependence of I on Z takes the form $I \propto Z^{1/2}$. Semi-empirical formula $I = C_n Z^{\alpha}$, $\alpha = 4/3, 1/2$, where C_n is constant for n-th period in the periodic table of elements, is proposed for Z in interval from 5 to 126. Interval estimates of coefficients C_n are constructed using interval approach [3]. Both subsequences of obtained interval estimates of C_n for $\alpha = 4/3$ and $\alpha = 1/2$ are extrapolated for $Z > 100$ using interval approach as well. The proposed formula gives the values of average excitation energy recommended by NIST with high accuracy, and allows to predict values of average excitation energies for the elements with $Z > 100$, whose electronic configuration is calculated but cannot be measured now.

References

1. Seltzer S.M., Berger M. J. // Int J Appl Radiat Isot 33(11), 1982, p 1189.

2. Berger, M.J., Coursey, J.S., Zucker, M.A., and Chang, J. ESTAR, PSTAR, and ASTAR: Computer Programs for Calculating Stopping-Power and Range Tables for Electrons, Protons, and Helium Ions (version 1.2.3), http://physics.nist.gov/Star. NIST, 2005.

3. Zhilin S.I. // Chemometr Intell Lab Syst 88(1), 2007, pp 60-68.

P10. Express evaluation of three quality parameters in vegetable oils using potentiometric multisensor system

V. Semenov[1,2], S. Volkov[3], M. Khaydukova[1,2], A. Fedorov[2,3], I. Lisitsyna[3], D. Kirsanov[1,2], A. Legin[1,2]

[1]Laboratory of chemical sensors, Institute of Chemistry, St. Petersburg State University, Russia

[2]ITMO University, St. Petersburg, Russia

[3]All-Russian Research Institute of Fats (ARRIF), St. Petersburg, Russia

Nutritional value and retail price of edible vegetable oils are strongly correlated with their quality. According to regulatory documents there are certain standards introduced for production process, proper labeling and quantification of the quality parameters (peroxide value, anisidine value, fatty acid composition, etc.). Most of the analytical methods endorsed for the vegetable oil analysis are time consuming; require complex laboratory equipment and several experiments for one sample, since each test is targeting only one specific quality parameter. These limitations exclude a possibility of express evaluation of the edible oil quality on-the-spot. That is why the development of the simply handled technique which allow for simultaneously estimate of the several oil quality parameters is demanded.

In this study, potentiometric multisensor system was applied for express evaluation of three quality parameters in vegetable oils. Sample set consisted of 11 refined sunflower oils was analyzed by standard analytical methods and peroxide value (PV), anisidine value (AV), total tocopherol content (TT) were estimated. Potentials of the 12 solid state sensors were registered in oil-water-alcohol emulsions, 4 repetitions for each sample. Obtained potentiometric data were related with the numerical values of quality parameters by PLS regression. Application of multivariate regression tools yielded root mean square errors of prediction: 0.5 mmol ½O/kg for PV; 0.8 arbitrary units for AV; 10 mg/100 g for TT. The parameters under study have the following ranges: PV (0.0 – 4.0); p-AV (0.4 – 3.8); TT (37.0 – 100.7).

It is shown that important edible oil quality indicators can be predicted from potentiometric multisensor system response using multivariate data processing and suggested approach seems rather promising for development of fast and

simple method for express oil quality evaluation. The precision of the obtained models in prediction is not very high, however the practicability of the suggested method can be justified by its simplicity, “green” character, relatively low cost, very short analysis time.

Acknowledgement

Authors acknowledge the financial support provided by RSF project #17-73-10284.

P11. Raman transduction for ISE polymeric membranes: feasibility study and preprocessing optimization

Yulia Ashina, Maria Khaydukova

Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia

The applicability of Raman spectroscopy for metal cations aqueous solutions analysis by ISE polymeric sensor membranes was examined. The idea of the developed approach consisted in employing the optical transduction of analytical signal derived from ISE membrane to yield the information about metal content in a sample solution. As an example, changes in the spectral signature of N^2,N^2,N^9,N^9-tetrabutyl-1,10-phenanthroline-2,9-dicarboxamide (PDAM) [1], caused by binding Cd^{2+} cations in the PVC-plasticized polymeric membrane as a matrix are discussed. Quantification of Cd^{2+} concentration was performed with partial least squares regression (PLSR) [2]. To maximize the performance of the proposed method, the preprocessing optimization of the Raman spectral data was performed.

The proposed approach demonstrated sufficient reproducibility and accuracy of Cd^{2+} detection in aqueous solutions down to 10^{-5} mol/L.

Authors appreciate financial support provided by RSF project #17-73-10284

References

1. Alyapyshev, M., et al., *1,10-Phenanthroline-2,9-dicarboxamides as ligands for separation and sensing of hazardous metals.* RSC Advances, 2016. **6**(73): p. 68642-68652.

2. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 109-130.

P12. PCA-analysis of voltammetric time rows of temporal multisensory systems

A.V. Sidelnikov, E.I. Maksyutova, A.A. Tikhonova
Bashkir state university, Ufa, Russia

Modern instrumental methods of analytical chemistry have ample opportunities for solving problems of qualitative and quantitative analysis of a large range of different nature compounds. The development of chemometric methods for processing various signals, including multivariate data, in recent decades has contributed to the development of multisensory chemical analysis methods, which are directly related to the need to process large data arrays. For example, multisensory systems of the "electronic tongue" type and the "electronic nose" type are being developed. They allow to identify several components at their simultaneous presence in the analyzed solution or to solve more complex problems of nonparametric estimation of the solutions properties without detailed quantitative analysis.

If a sufficiently large number of methods and sensors are proposed to detect individual compounds in a complex mixture, then for the recognition of multicomponent mixtures (solutions), it remains relevant to solve the problem of creating multisensor methods for extracting chemical information. One of the ways to solve this problem is creating of new methods for electrochemical data registration, including activation of the electrode surface in conditions of continuous functioning of the sensor system. The accumulation of chemical information about the nature of the depolarizer, the formation of an analytical signals array specifically associated with a small diversity in substances are possible not only in the presence of a sensors array, as is customary in the practice of multisensory analysis, but also in the conditions of repeated information accumulation in the conditions of continuous functioning of even one (initially non-selective) working electrode.

Multisensory method is the basis of the functioning of most known “electronic tongues”, “electronic noses” and an important factor in the accumulation of information about the difference between similar nature compounds. The developed method allows to create multisensory formation on the one electrode during the whole time of the electrode functioning. Each electrode activation cycle changes the sensor sensitivity (and its response accordingly) in such a way that, in the presence of a large number of measurements at one point the electrode surface changes qualitatively. Thus, we create the conditions for the sensory system functioning as if we use a large number of sensors with different sensitivities and obtain multivariate information in time about the solution and its components. The proposed multisensory system can be used to recognize mixtures of biologically active compounds, including differentiating enantiomers from each other.

This work was supported by the Russian Foundation for Basic Research, project № 17-43-020232 r-Povolzh'ye-a.

P13. Systematic comparison of smoothing methods

Y.A. Kalambet[1], *Y.P.Kozmin*[2], *A.S.Samokhin*[3]

[1]*Ampersand Ltd., Moscow Russian Federation*

[2]*Shemyakin–Ovchinnikov Institute of bioorganic chemistry RAS, Moscow, Russian Federation*

[3]*Chemistry Department, Lomonosov Moscow State University, Moscow, Russian Federation*

Technology of comparison of smoothing methods is proposed. The technology is based on modelling of the signal and noise, smoothing of noisy signal, subtraction of original signal and study of the residuals repeated many times. Both systematic and random errors of smoothing can be revealed, local and integral. Comparison of moving average, median, Gaussian, Savitzky-Golay smoothing methods and Adaptive smoothing algorithm with different basic function sets (Legendre, Cosine, Fourier, Hermitian) is performed. Criteria of smoothing method selection for chromatographic data processing are discussed.

P14. A new diagnostic method for the solid electrodes surface using dynamic impedance spectroscopy, voltammetry and the principal components analysis

E.I. Maksyutova, A.V. Sidelnikov, D.I. Dubrovskiy
Bashkir state university, Ufa, Russia

The electrode / solution interphase boundary has long attracted the attention of researchers as a kind of electrochemical microreactor, in which processes under controlled conditions give useful information about the electrode material, the structure of its surface, the components of the solution both in depth and in the near-electrode layer, etc. The nature of the analytical signal (or signals - in the case of a set of electrodes), which is generated in this process, contains all the completeness of information.

New materials and new technologies for the formation of surface structures from materials such as nanomaterials, composites based on them, supramolecular structures, etc., require new methods for controlling the quality of surface the created electrodes-sensors. This includes monitoring the surface morphology, the degree of its homogeneity, purity, etc., both at the final stage of their manufacture, and in the process of the external action during the signal measurement. Here, the need for investigation and parametrization of the process of structural change of the electrode surface during the continuous electrode-sensor functioning moves to the forefront.

The problem is directly related to solving the problems of sensory systems creating, when, in the long-term operation of the sensor, its sensitivity to the solution components under changes with time. Using the proposed hybrid method of diagnostics, it is established that at certain intervals of the sensor's functioning, quasi-stable states of its surface can be reached, when an analytical signal can be obtained with sufficient precision values.

The solution of the problem of monitoring and parameterization of the sensor "aging" process in the conditions of system continuous functioning is directly related to solving the problem of reliable detection of trace components. It is known that the smaller the dispersion of background signals (rather than the absolute value of them), the lower the detection limit of components. The sensor

self-modifying as a way of increasing the response sensitivity to a change in nature and concentration of the analyte using multiple oxidation-reduction cycles simultaneously forms a multivariate array of signals, which is an important data bank for diagnostic of the sensor surface and controlling the process of modifying the sensor.

This work was supported by the Russian Foundation for Basic Research, project № 17-43-020232 r-Povolzh'ye-a.

P15. Identifying fluorophores in Arctic shelf seas by PARAFAC analysis of excitation-emission fluorescence spectra of seawater

Krylov Ivan
Moscow State University, Moscow, Russia

Dissolved organic matter (DOM) plays an important role in the environment by supporting growth of marine biota and participating in flocculation of colloid clay particles in estuarine zones; it is also an indicator of organic loadings in streams and terrestrial processing of organic matter.

Studying carbon transport requires the ability to distinguish its different sources. In seawater, carbon sources include autochthonous (of biological or aquatic bacterial origin) and allochthonous (terrestrial runoff) ones. DOM in seawater mostly consists of humic substances which have heterogeneous molecular structure. Detailed characterization of DOM is possible by means of high resolution mass-spectroscopy, which requires large sample volumes (5-10 l) to concentrate and is only possible under laboratory conditions. Different optical methods are also used to study DOM, especially spectrofluorometry. It is impossible to discern separate emission lines corresponding to individual compounds in DOM fluorescence spectra, so instead, a small number of fluorophores with defined optical characteristics is described and correlated to allochthonous protein-like compounds or terrigenous humic substances. Two-dimensional spectrofluorometry, with both excitation and emission wavelengths as independent variables, producing an excitation-emission matrix (EEM), is one of the most informative optical methods for DOM research.

Tensor rank decomposition methods (e.g. PARAFAC) are successfully employed to find the independent components (considered to be correlated to

fluorophores) comprising each individual EEM in the dataset. There is still only fragmentary information on PARAFAC components of EEMs of Arctic seawater. In this work, 80 samples of DOM from shelf seas were collected during the cruises to the Kara Laptev, White and East Siberian seas in autumn (2015-2017) and spring (2016). A few samples were taken from freshwater ponds of Novaya Zemlya archipelago. EEMs were recorded in wide ranges of excitation (230–550 nm) and emission (240–650 nm) wavelengths. Spectra were normalized to the area of Raman water peak at 350 nm excitation wavelength. Due to hardware limitations, the data contained unfiltered scattering peaks. A PARAFAC model built from raw spectra did not provide an accurate description of fluorescence and the resulting components only correlated with scattering bands, therefore different techniques were tried to minimize the interference. Zeroing out the wavelength regions where scattering information was present caused step-like artefacts in obtained components, but most of fluorescence information was recovered. Replacing the zeroed out parts of the spectra with smoothly interpolated surfaces resulted in more interpretable components. Subtracting normalized spectrum of clean water did not get rid of scattering signal; resulting components contained both unneeded scattering information and artefacts caused by attempts to remove them. Decompositions into 5 and 6 factors were obtained, which allowed evaluating the relative distribution of marine and terrigenous DOM in surface waters of the studied areas.

This work is supported by the Russian Foundation for Basic Research, grant No 16-35-60032.

P16. The condition diagnosis of the technological process of olefins production

Osipenko U.Y., Rusinov L.A.
Saint-Petersburg State Institute of Technology

The obtaining olefins from normal paraffins is one of the main steps of linear alkyl benzene (LAB) production. LAB is the important industrial product, which is used as a component in the manufacture of synthetic detergents.

A crucial task is an early detection and prevention of dangerous abnormal situations before an emergency shutdown system is actuated. In this case, an

immediate problem is the developing of a diagnostic system for detecting abnormal situations.

Analysis of the olefins obtaining process has been executed to elaborate a diagnostic model of the process. Localization of abnormal situations advent has permitted to isolate 10 places, which unite the troubles related to the processes occurring in the main equipment of the olefin production unit.

It proposes to realize process monitoring with the aid of principal component analysis (PCA). As PCA makes it possible to reduce the dimension of the model describing of the technological process current state and to remove redundancy of source data.

The frame-production multilayer diagnostic model has been selected to further identification of cause of abnormal situation detected while process monitoring. In this model, abnormal situations form into groups by some kind of feature. Finding the cause of trouble is achieved by using the production rules connecting symptoms with the cause inducing of an abnormal situation.

Analysis of the olefins obtaining process has shown the necessity in development a fuzzy diagnostic model, because fuzziness of threshold values determining evolution of abnormal situations has been found out for some process parameters. Moreover, for diagnostics of the catalyst activity critical fall, there are no any exact values of process parameters indicating to necessity of exigent catalyst change-out.

P17. Multivariate calibration for chlorine determination in concrete by Laser-induced breakdown spectroscopy of CaCl molecule

Labutin T.A., Zakuskin A.S. Popov A.M., Zaytsev S.M.
Lomonosov Moscow State University, Moscow, Russia

Concrete degradation is still a serious problem in the construction industry. Interaction with an atmosphere, water, or ground can modify the chemical composition of concrete reinforced structures. The main forms of chemical attack on concrete reinforced structures are chloride corrosion of steel reinforcement, carbonation, and sulfate attack. It was shown that laser-induced breakdown spectroscopy (LIBS) can be used for the elemental analysis of

building structures [1]. The main advantages of LIBS for diagnostics of building materials are: relatively simple apparatus, which allows miniaturization and automation of the analytical measurements in the field; high throughput and local analysis. At the same time known techniques of determination of chlorine and sulfur by LIBS, which provide needed level of sensitivity, can be hardly realized in portable arrangement for field measurements. Thus, it is still important the development of approaches, which can help in miniaturization of instrument with appropriate sensitivity of non-metals determination in concrete.

The use of emission of diatomic molecules instead of atomic lines to determine hardly excited species like halogens is well-known approach in atomic emission spectroscopy. The advantages of the molecular bands are related to a sufficiently high excitation potential chlorine line (~ 10.5 eV), and a low degree of its atomization in plume. The molecular bands of MgCl, CaCl and AlCl were evaluated for chlorine detection. We have also applied double-pulse LIBS to use the atomic lines Cl I 833.7 nm for the comparison.

It was found that only the orange system of CaCl molecule was sufficiently intense for chlorine determination in concrete by common LIBS at the late times of plasma existence (30 μs after laser pulse). The analytical signal has depended linearly on chlorine content in concrete. However, the complex structure of the vibrational levels of the orange system results in the spectrum with a wide asymmetric band accomplished by numerous interferences from the atomic lines. This has prevented determination of low chlorine content, which is important to reveal the dangerous level of chlorine accumulation in concrete.

In such a situation, the most appropriate and simple solution is the use of multivariate calibrations to provide accurate analysis. We have prepared 20 samples simulating the concrete samples with different chlorine contents. The application of PCR provided a calibration model appropriate for reliable chlorine determination at the threshold level. Thus the multivariate calibration with the use of CaCl emission measured with the common LIBS setup as analytical signal improved the sensitivity.

References

1. G. Wilsch, F. Weritz, D. Schaurich, H. Wiggenhauser, Determination of chloride content in concrete structures with laser-induced breakdown spectroscopy, Constr. Build. Mater. 19 (2005) 724-730.

P18. Impact of various ratios of apple varieties used for apple juice production on its taste qualities

V. Gilemkhanova, F. Conradi, V. Kindsvater, J. Schneider, M. Pein-Hackelbusch
Hochschule Ostwestfalen-Lippe, Lemgo, Germany

Apple is the most widely grown fruit on earth and this fact is the reason for a high popularity of apple juice production. Depending on season of a year, climate, and storage conditions, its taste qualities vary significantly [1], but for producers and consumers it's important for any apple juice to have the same taste throughout a year. In both, eastern and western juice manufacturing facilities, the juice product is usually a blend of the juice from two or more varieties. This blending procedure allows for a more uniform product throughout the season and from season to season [2]. Producer must keep in mind that using distinct varieties results in different taste qualities, which in its turn leads to changing the taste of the whole product.

A promising technique for quality control of apple juice can be represented by a potentiometric multisensory system, or "electronic tongue". Such systems are well-known tools for taste evaluation in both food and pharmaceutical industries [3]. Common advantages include rapidness, low cost, and a possibility to use the system in-line.

In this study, the applicability of a potentiometric multisensor system (so called e-tongue) was applied to apple juices of various composition in order to distinguish 100% apple juice of one apple variety and its mixtures when up to 5% of whole volume consists of other apple juice was evaluated. Therefore, in-house produced apple juices were analyzed in various compositions. Reliability of the e-tongue data was proven by using NIR.

References:

1. T.A. Eisele, S.R. Drake, The partial compositional characteristics of apple juice from 175 apple varieties. Journal of Food Composition and Analysis 18 (2005) 213-221

2. William H. Root, Diane M. Barrett, Chapter 18 Apples and Apples Processing 2005

3. E.A. Baldwin, J. Bai, A. Plotto, S. Dea, Electronic Noses and Tongues: Applications for the Food and Pharmaceutical Industries, Sensors 11 (2011) 4744-4766

P19. The classification of vegetation types at the territory of the Lake Baikal basin based on satellite images and chemometrics

Elena P. Yankovich[1], Ksenia S. Yankovich[2, 3], Pavel O. Dedeev[4]

[1]National Research Tomsk Polytechnic University

[2]Saint Petersburg National Research University of Information Technologies, Mechanics and Optics

[3]University of Paris 7 (Paris Diderot University)

[4]University of Paris-Saclay (ENSTA ParisTech)

The problem of forest fire affects many regions and the detailed approach is needed in case of protected nature areas. The research areas are two forest districts located in the territory of the Lake Baikal basin. One of the main indicators is the vegetation type and its phenology. These data are necessary to estimate the level of a forest fire danger. Needed information was extracted from images after their interpretation, photogrammetric plotting, photometric and computer (automated, digital) processing. Two types of classification algorithms were applied for the purpose of the research (supervised classification and unsupervised classification (clustering analysis). The images were used together with information from other sources.

The spectral response of vegetation varied in reference with stages of its development and seasonal variations at the studied range of wavelengths. Each class had the spectral characteristics, based on which further evaluation of the physicochemical condition of the observed surface was conducted. The knowledge of these characteristics, the possibility of their remote capture and the usage of well-known spectral indexes allowed interpreting received images with the high precision. The classification inside the class was carried in the research. The level of seasonal activity, the development stage or level of defoliation were determined for separate sections (inside one vegetation type).

The surface distribution of spectral characteristics was studied as well. The use of chemometrics for multidimensional data interpreting allowed differentiation of the vegetation types.

Distribution map of different vegetation types was obtained as a result of the analysis of studied areas. Also the boundaries of the forest fire danger areas were determined. Moreover, the primary model was constructed, aimed at the automated classification and at the differentiation of vegetation types at the territory of the Lake Baikal basin.

The research was implemented with the financial support of the Russian Foundation for Basic Researches № 17-29-05093.

P20. Partial least squares density modelling (PLS-DM) – an efficient approach to one-class classification

Paolo Oliveri, Cristina Malegori, Monica Casale
DIFAR Department of Pharmacy, University of Genova, Genova, Italy

In the present work, a novel class-modelling method, called partial least squares density modelling (PLS-DM), is presented. Class-modelling (also referred to as one-class classifiers) is one of the two families in which it is possible to divide the qualitative data modelling and it is the choice to be preferred when the focus is on a single class – a typical case for the verification of authenticity claims.

PLS-DM is based on a partial least squares (PLS) regression, in which a distance-based sample density measurement is used as the response variable. A kernel approach involving a potential-function estimation of probability density is subsequently applied on PLS scores, and is considered, jointly with residual *Q* statistics, to develop efficient class models effectively capable to deal with data characterised by non-normal and non-uniform distributions.

The work critically discusses the influence of adjustable model parameters (e.g., the pre-processing, the number of latent variables, and the smoothing coefficient of potential functions) on the resulting performances by means of evaluation of sensitivity and specificity within a cross-validation cycle and by application of the Pareto optimality criterion; moreover, performances of the optimal model are evaluated on an external test set.

The potential of PLS-DM is illustrated presenting ad-hoc case studies related to verification of food authenticity claims, including a critical comparison with well-established class-modelling methods, such as soft independent modelling of class analogy (SIMCA) and unequal dispersed classes (UNEQ).

Acknowledgments

Financial support by the Italian Ministry of Education, Universities and Research (MIUR) is acknowledged – Research Project SIR 2014 "Advanced strategies in near infrared spectroscopy and multivariate data analysis for food safety and authentication", RBSI14CJHJ (CUP: D32I15000150008).

P21. Determining the age of 20th century paintings using FTIR spectroscopy and PLS regression

Andrey Samokhin[1], *Vilena Kireeva*[2], *Andrey Borisov*[2], *Alexander Revelsky*[1]
[1]*Division of Analytical Chemistry, Chemistry Department, Lomonosov Moscow State University, Moscow, Russia*
[2]*ARTconsulting, Moscow, Russia*

Determining the age of 20th century paintings still remains difficult, despite the existence of a large number of modern analytical methods (including infrared spectroscopy, X-ray diffraction, scanning electron microscopy, pyrolysis gas chromatography/mass spectrometry and etc.). Chemical composition of oil paints changes during natural aging (due to polymerization, oxidation and some other processes). Previously it has been shown that there is correlation between intensities of particular peaks in FTIR spectra and the age of paintings [1–3]. Nevertheless, developed one-dimensional regression models [1,3] have limited applicability due to low accuracy of prediction.

In the present work more than 200 samples of zinc white paints (collected from 54 paintings) were analyzed. Each sample was scraped off from the paint layer. Measurements were performed using ATR-FTIR spectrometer in the region 600–4000 cm^{-1}. A number of pre-processing techniques were compared (multiplicative scatter correction, vector normalization, first or second differentiation). Multi-dimensional regression model was built using PLS

technique. Calibration set (containing samples collected from 34 paintings) was used to construct and optimize (cross-validation) model. Difference between predicted and actual ages obtained for validation set (containing 82 samples of 20 paintings) was less than 14 years for 95% of samples.

References

1. *Zyablov E.M., Lukashin D.E., Borisov A.N., Gerasimov V.K., Kostina J.V., Kireeva V.N.* Russian Federation patent RU 2386119. 2008 Nov 17. (*In Russian*)

2. *Balakhnina I.A., Brandt N.N., Chikishev A.Y., Grenberg Y.I., Grigorieva I.A., Kadikova I.F., Pisareva S.A.* Fourier transform Infrared (FT-IR) Microspectroscopy of 20th Century Russian Oil Paintings: Problem of Dating. // Applied Spectroscopy. 2016. Vol. 70. N 7. P. 1150–1156.

3. *Balakhnina I.A., Brandt N.N., Valenti D., Grigorieva I.A., Spagnolo B., Chikishev A.Y.* Statistical Approximation of Fourier Transform-IR Spectroscopy Data for Zinc White Pigment from Twentieth-Century Russian Paintings. // Journal of Applied Spectroscopy. 2017. Vol. 84. N 3. P. 484–489.

P22. Internal standard calculations for non-linear detectors

Yuri Kalambet, Yuri Kozmin
Ampersand Ltd, Moscow, Russia
Shemyakin Institute of Bioorganic Chemistry, Moscow, Russia.

Internal Standard (ISTD) is a well-known chromatographic technique, where known amount of a component, called internal standard is added to both standard and unknown samples. The “traditional” Internal Standard quantification scheme plots the response ratio (analyte to standard) versus amount ratio (again analyte to standard). Internal standard component itself does not have any calibration curve within this scheme. Quantification procedure uses ratio plot to get concentration ratio from response ratio. Our example demonstrates, that this approach may cause systematic errors if external standard – type calibration function of either analyte or standard is not directly proportional (linear going through origin).

We offered alternative calculation scheme that allows wide variations of standard and analyte concentrations. In the case of non-directly proportional dependencies, it requires that External standard calibration functions of both internal standard component and analyte are explicitly measured. Internal standard calculation consists of two parts:

1. Calculation of Relative concentration, i.e. concentration of analyte, provided concentration of Internal Standard is known.
2. Improvement of calibration curves (Relative calibration) of the analytes.

The calculation scheme can be further extended to the case of directly proportional dependencies and can successfully replace "traditional" calculation scheme. The described calculation scheme is successfully implemented for Internal Standard calculations in commercial chromatographic software.

P23. Confocal Raman spectroscopy and multivariate data analysis in evaluation of spermatozoa with normal and abnormal morphology

A.V.Irzhak[1], R.V.Nazarenko[2], O.Ye.Rodionova[3] and A.L. Pomerantsev[3]
[1]Institute of Microelectronics Technology and High Purity Materials RAS, Chernogolovka, Russia
[2]EKO Infertility treatment centre, Moscow, Russia
[3]Semenov Institute of Chemical Physics RAS, Moscow, Russia.

The aim of this study is to assess feasibility of the confocal Raman spectroscopy and multivariate analysis methods in studying the sperm nuclear DNA. Additional goal is to perform a comparative analysis of the Raman spectra (RS) obtained from the morphologically normal and abnormal spermatozoa.

We performed analysis of semen taken from healthy donors and assembled two sets of samples: morphologically normal sperm (N group, 125 samples) and morphologically abnormal sperm (A group, 36 samples). The Raman spectral analysis of the sperm nucleus was carried out at the 532 nm laser excitation wavelength and the 10 mW power in the range of 280 – 1730 cm^{-1} at the resolution of 3 - 5 cm^{-1}. Principal component analysis (PCA) performed by

specially designed Chemometrics Add- In for the Microsoft Excel after baseline correction and normalizing.

In the result, the N group (125 samples) was divided into two classes: the NN class (102 samples) and the NA class (23 samples). The NN class was used as the target class. It comprises the samples recognized as the normal both by morphology and by the results of spectral data analysis. The NA set comprises the samples considered normal according to the morphology analysis, but treated as extraneous samples relative to the NN target class. The final acceptance area was designed at 3 PCs using samples of the NN target class (102 samples).

The spectra of the second A group were analyzed using the model developed for the NN target class. A group was divided into two classes: the AA class (19 samples), morphologically and spectrally anomalous, and the AN class (17 samples), morphologically abnormal but spectrally normal.

Studying the group of morphologically abnormal spermatozoa, we noted the predominance of abnormal spectra with a high peak of 1045 cm^{-1}. However, we did not succeed in isolating the characteristic features inherent in the whole class of anomalous spectra due to its heterogeneity. Apparently, this can be explained by the fact that abnormality is an internal property determined by the multiple damages of the nuclear DNA, which manifests itself not in the specific absorption bands but in the entire spectral range. Thus, the separation of samples into normal and abnormal is a non-trivial task that can only be solved by proper analysis of spectral data.

P24. ToF-SIMS for molecular analysis of lipid components in model systems, cell structures and tissue sections

N.A. Trankova[1,2], A.A. Gulin[2,3], A.E. Solodina[2], S.K. Gularyan[2]

[1]Moscow Institute of Physics and Technology (state university), Dolgoprudny, Russia

[2]N. N. Semenov Institute of Chemical Physics RAS, Moscow, Russia

[3]Moscow State University, Department of Chemistry, Moscow, Russia

Lipid composition of living systems is known to be sensitive to appearance of different pathologies. Time-of-flight secondary ion mass spectrometry (ToF-SIMS) is a surface sensitive technique allowing to study surface composition and distribution of lipids. In this study we identify characteristic ions of the most common lipid species by investigation of model lipid. We found correlation between the lipid concentration in the alcohol solution used for film formation and the signal of lipid secondary ions. Identified ions were detected in glioblastoma culture cell membrane. Furthermore, the variation coefficient of signal ratio of these lipids in cell membrane is within 20-30% for different areas of cell culture. The further comparison of lipid composition in healthy and affected tissue areas (brain tissue sections of a mouse affected by a neuroglioma were used) shows significant differences for some lipid species. For instance, in affected area the signal of cholesterol is 40-50% lower than in health area, but the level of myristic fatty acid is 3 times higher in affected area. Studying of the differences in chemical distribution between healthy and affected tissues is valuable for understanding fundamental processes of tumor formation and treatment therapy.

The work was supported by the Russian Foundation for Basic Research (project 17-53-45080).

Participants

Artyushenko Viacheslav
art photonics GmbH
President
Berlin, Germany
sa@artphotonics.com

Ashina Yulia
Saint Petersburg State University
Researcher
Saint Petersburg, Russia
ashina.julia91@gmail.com

Bajusz Dávid

Budapest, Hungary
bajusz.david@ttk.mta.hu

Engelsen Søren Balling
University of Copenhagen
Professor
Copenhagen, Denmark
se@food.ku.dk

Belikova Valeria
Samara State Technical University
PhD student
Samara, Russia
valerya.belickova@yandex.ru

Bogomolov Andrey
Blue Ocean Nova
Chemometrician
Aalen, Germany
ab@globalmodelling.com

Burmistrova Natalia
Saratov State University
Professor
Saratov, Russia
naburmistrova@mail.ru

Dörgő Gyula
University of Pannonia
PhD student
Veszprém, Hungary
gydorgo@gmail.com

Esbensen Kim
KHE Consulting
Professor
Copenhagen, Denmark
khe.consult@gmail.com

Galkin Evgenii
CSort, Ltd.
Head of R&D department
Barnaul, Russia
tech9@csort.ru

Gemperline Paul J.
East Carolina University
Professor
Greenville, USA
gemperlinep@ecu.edu

Gilemkhanova Venera
Hochschule Ostwestfalen-Lippe
Research assistant
Detmold, Germany
v.gilemkhanova@gmail.com

Gubareva Tatiana
Institute of Water Problems RAS
Senior researcher
Moscow, Russia
Tgubareva@bk.ru

Guo Shuxia
University of Jena

Jena, Germany
shuxia.guo@uni-jena.de

Harrington Peter
Ohio University
Professor
Athens, USA
peter.harrington@ohio.edu

Heberger Karoly
Research Centre for Natural Sciences
Scientific advisor
Budapest, Hungary
heberger@chemres.hu

Hohmann Monika

Renningen, Germany
monika.hohmann@de.bosch.com

Istomin Andrey
CSort, Ltd.

Barnaul, Russia
istomin@csort.ru

Elek Janos
Science Port Ltd.
Manager
Debrecen, Hungary
elek@scienceport.hu

Kalambet Yuri
Ampersand Ltd.
General director
Moscow, Russia
kalambet@ampersand.ru

Kapustin Denis
Altai State University
Student
Barnaul, Russia
a19452000@yandex.ru

Khaydukova Maria
Saint Petersburg State University
Researcher
Saint Petersburg, Russia
khaydukova.m@gmail.com

Kirsanov Dmitry
Saint Petersburg State University
Professor
Saint Petersburg, Russia
d.kirsanov@gmail.com

Krylov Ivan
Moscow State University
Undergraduate student
Moscow, Russia
krylov.ivan@gmail.com

Kucheryavskiy Sergey
Aalborg University
Associate professor
Esbjerg, Denmark
svkucheryavski@gmail.com

Labutin Timur
Moscow State University
Assistant professor
Moscow, Russia
timurla@laser.chem.msu.ru

Maksyutova Elza
Bashkir State University
PhD student
Ufa, Russia
elzesha@gmail.com

Malegori Cristina
University of Genova
Postdoc
Genova, Italy
malegori@difar.unige.it

Marini Federico
Sapienza University of Rome
Professor
Rome, Italy
federico.marini@uniroma1.it

Mazafi Alon
Ben Gurion University of the Negev
Student
Beer Sheva, Israel
alonmat@post.bgu.ac.il

Melenteva Anastasiia
Samara State Technical University
Researcher
Samara, Russia
melenteva-anastasija@rambler.ru

Monakhova Yulia
Saratov State University
Researcher
Saratov, Russia
yul-monakhova@mail.ru

Nikzad-Langerodi Ramin
Johannes Kepler University
Research assistant
Linz, Austria
ramin.nikzad-langerodi@jku.at

Oleneva Ekaterina
Saint Petersburg State University
Researcher
Saint Petersburg, Russia
ekaterina.oleneva@inbox.ru

Oliveri Paolo
University of Genova
Assistant professor
Genova, Italy
oliveri@difar.unige.it

Osipenko Uliana
Saint Petersburg State Institute of Technology
Lecturer
Saint Petersburg, Russia
osipenko.u@gmail.com

Panchuk Vitaly
Saint Petersburg State University
Associate professor
Saint Petersburg, Russia
vitpan@mail.ru

Pomerantsev Alexey
Institute of Chemical Physics RAS
Principal researcher
Moscow, Russia
alexey.pomerantsev@gmail.com

Pyatak Tetyana
Blue Ocean Nova
Application engineer
Aalen, Germany
tpyatak@blueoceannova.com

Rácz Anita

Budapest, Hungary
racz.anita@ttk.mta.hu

Rodionova Oxana
Institute of Chemical Physics RAS
Leading researcher
Moscow, Russia
oxana.rodionova@gmail.com

Ruckebusch Cyril
Lille1 University
Professor
Lille, France
cyril.ruckebusch@univ-lille1.fr

Rudnitskaya Alisa
University of Aveiro
Researcher
Aveiro, Portugal
alisa@ua.pt

Samokhin Andrey
Moscow State University
Research associate
Moscow, Russia
andrey.s.samokhin@gmail.com

Schroeder Henning
University of Rostock
PhD student
Rostock, Germany
henning.schroeder2@uni-rostock.de

Sidelnikov Artem
Bashkir State University
Professor
Ufa, Russia
artsid2000@mail.ru

Skvortsov Alexej
Saint Petersburg State Polytechnical University
Head of department
Saint Petersburg, Russia
colbug@mail.ru

Titova Anna
Pirogov Russian National Research Medical University
Professor
Moscow, Russia
titova1701@yandex.ru

Trankova Natalia
Institute of Chemical Physics RAS
Junior researcher
Moscow, Russia
nattrankova@gmail.com

Yankovich Elena
Tomsk Polytechnical University
Senior lecturer
Tomsk, Russia
yankovich@tpu.ru

Yaroshenko Irina

Saint Petersburg, Russia
irina.s.yaroshenko@gmail.com

Yunovidov Dmitry
Research Institute For Fertilizers And Insectofungicides
Researcher
Cherepovets, Russia
dm.yunovidov@gmail.com

Zarubin Aleksey
Tomsk Polytechnical University
Assistant professor
Tomsk, Russia
zagtpuru@gmail.com

Zhilin Sergey
Altai State University

Barnaul, Russia
szhilin@gmail.com

Zontov Yury

Moscow, Russia
yury.zontov@gmail.com