Drushbametrics Project

Eighth Winter Symposium on Chemometrics

Modern Methods of Data Analysis

Russia, Drakino, February 27–March 2, 2012

Drushbametrics Project

Eighth Winter Symposium on Chemometrics

Modern Methods of Data Analysis

Russia, Drakino, February 27–March 2, 2012

Russian Chemometrics Society
Scientific Consul for Analytical Chemistry RAS
Semenov Institute of Chemical Physics RAS
Aalborg University
Vernadsky Institute of Geochemistry and Analytical Chemistry RAS
Polycert Ltd
with support of
Russian Foundation for Basic Research
CAMO Software AS
Intertech Trading Corporation
Lumex Instruments
CAC 2012 Organizing Committee

Drushbametrics
WSC
8
project

Eighth Winter Symposium on Chemometrics

Modern Methods of Data Analysis

The organizing committee

Co-chairpersons
- Lev Gribov
- Oxana Rodionova

Secretary
- Alexey Pomerantsev

Members
- Michal Daszykowski
- Kim Esbensen
- Sergey Kucheryavskiy
- Federico Marini
- Leon Rusinov

ICP RAS, Kosygin St. 4, 19991 Moscow, Russia
http://wsc.chemometrics.ru/wsc8 e-mail: wsc8@chemometrics.ru

Thanks

The WSC-8 organizers and participants wish to express greatest appreciation to the following conference sponsors for their valuable economic and friendly help:

Russian Foundation for Basic Research (RFBR)

CAMO Software AS

Intertech Trading Corporation

Lumex Instruments

CAC 2012 Organizing Committee

Finally, we are grateful to all the WSC-8 attendees, lecturers, accompanying persons, and visitors for their interest to the conference.

See you again at the next WSC-9 conference!

Useful information

Conference and activities

Symposium sessions will be held at the conference room (#1in the park map). Skis, skates, sauna, etc are available for rent at #6.…

Meals

All meals (buffet) will be served in “The Wright brothers” restaurant (#8). The banquet will take place in the new restaurant (#19)

Scores & Loadings

The “Scores and Loadings” meeting will be held in “The Wright brothers” restaurant with the drinks at reasonable prices: 3$ for a glass of wine, or a bottle of beer, or a shot of whisky, etc. Tea, coffee, cookies are free.

Communications

The main three Russian cellular networks, Beeline, MTS and Megafon, have a proper coverage around the hotel. Internet WiFi is also available.

Money

ATM machine that accepts Master and Visa cards is in hotel Polet lobby. The hotel accepts the same types of card. The organizing committee could also change a reasonable amount of currency.

Transportation and excursion

Two buses from Drakino to metro station Domodedovskaya are scheduled for March 2.

Excursion bus B1 leaves at 10:00. It is combined with excursion to Chekhov's museum in Melikhovo. Expected arrival time is 14:30.

Bus B2 leaves at 9:00. This is a direct trip. Expected arrival time is 11:30.

Connection with Moscow

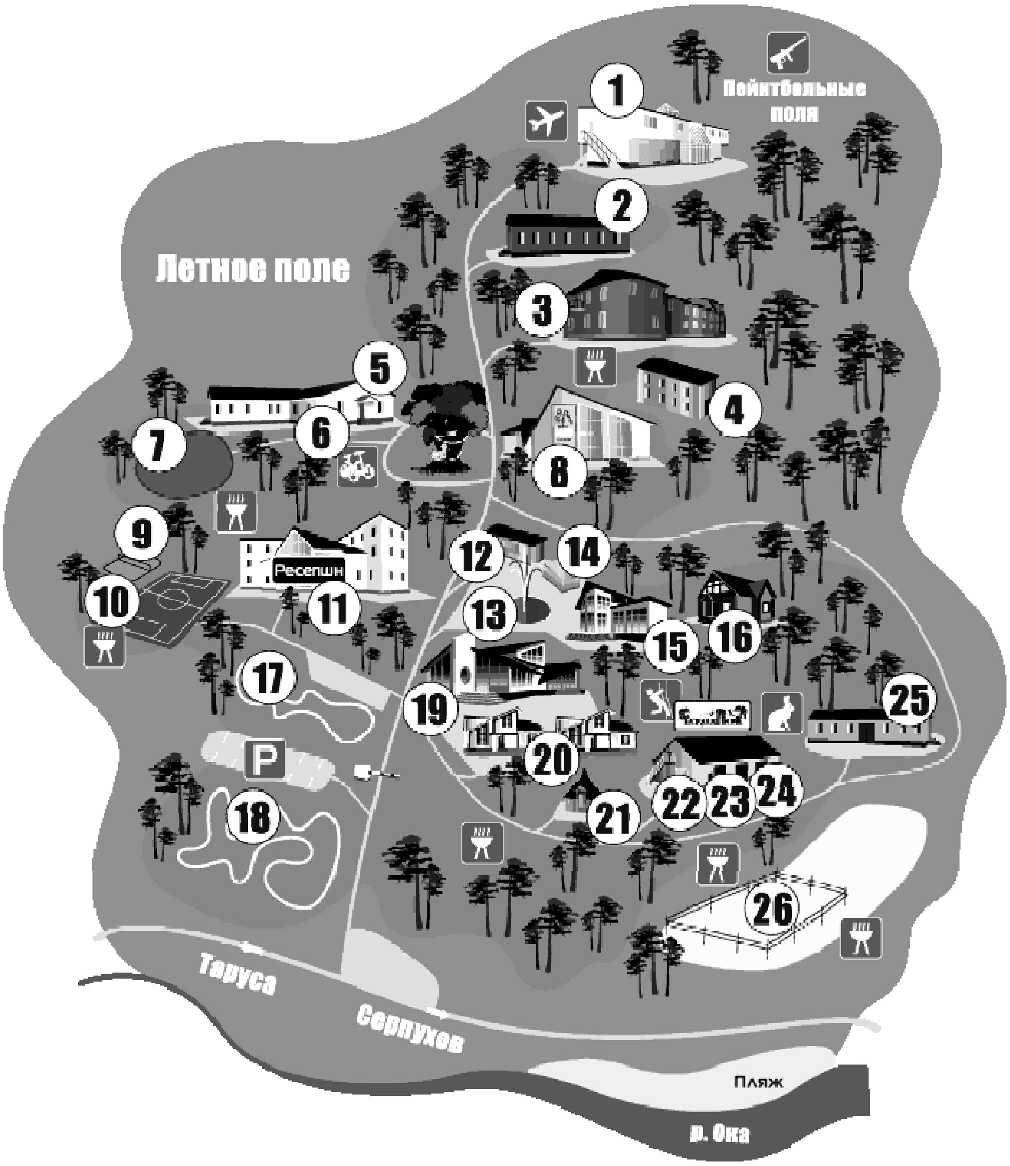
The hotel is located about 100 km from Moscow and about 20 km from train station Serpukhov, which is available by taxi (400 rub), or by the public buses.

Miscellaneous

The conference official language is English.

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

Map of park-hotel Drakino



1. Conference room

2. Paintball club

3. Hotel Icarus

4. Small conference room

5. Administration

6. Rent skis, skates, etc

7. Skating ring

8. Restaurant "The Wright brothers", all meals, Scores & Loadings

9. Volleyball place

10. Football field

11. Hotel Polet/reception

12. Cafe (not in winter)

13. Fountain square

14. Scene

15. SPA complex

16. SPA villa

17. Carting

18. "Off-Road" route

19. New Restaurant. Banquet

20. Cottages "Tirol Village"

21. Grill house

22. Sauna

23. All-terrain and snow vehicles

24. Cafe (not in winter)

25. Riding club

26. Riding

Useful Phone Numbers

Alexey Pomerantsev, symposium secretary	+7 (903) 5618227 (mobile)
Drakino reception	+7 (4967) 310 444 (local)

Monday, February 27, 2012

09:00–12:30	**Registration**
13:00–14:00	**Lunch**
14:00–14:30	**Opening**
Session 1	**Chair: Oxana Rodionova**
14:30–15:00	**T1** *Kim Esbensen* Replication — ending decades of confusion
15:00–15:30	**T2** *Sergey Kucheryavskiy* Classification of objects on hyper-spectral images
15:30–16:00	**T3** *Elena Chernetsova* Characterization of propolis using planar chromatography, direct analysis in real time mass spectrometry and multivariate data analysis
16:00–16:30	**Coffee break**
Session 2	**Chair: Federico Marini**
16:30–17:00	**T4** *Dmitri Kirsanov* 3-way PLS modeling of potentiometric multisensor system response for quantification of bitter taste in pharmaceuticals
17:00–17:30	**T5** *Elek Janos* Principal Component Analysis of an XI-- XV century silver coins' XRF dataset and verification by additional multilinear methods
17:30–18:00	**T6** *Veli-Matti Taavitsainen* Rational function ridge regression
18:00–18:30	**T7** *Yuri Kalambet* The best method of noise filtering

18:30–19:00	**Free time**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Tuesday, February 28, 2012

08:00–09:30	**Breakfast**
Session 3	**Chair: John Kalivas**
09:30–10:30	**L1** *Federico Marini* Coupling NIR and chemometrics for food and drug analysis
10:30–11:00	**T8** *Gerda Edelman* Identification and age estimation of blood stains on coloured backgrounds by near infrared spectroscopy
11:00–11:30	**Coffee break**
11:30–13:00	**Free time**
13:00–14:00	**Lunch**
Session 4	**Chair: Age Smilde**
14:00–15:00	**L2** *Andrey Bogomolov* Multivariate process trajectories
15:00–15:30	**T9** *Zdravko Stefanov* Process chemometrics in the Dow Chemical company
15:30–16:00	**T10** *Leon Rusinov* Fault diagnosis in chemical processes and equipment with feedback
16:00–16:30	**Coffee break**
Session 5	**Chair: Beata Walczak**
16:30–17:00	**T11** *Andrey Samokhin* Identification of unknown organic compounds based on comparison of electron ionization mass spectra
17:00–17:30	**T12** *Cyril Ruckebusch* Chemometrics for the resolution of time-resolved spectra in photochemistry
17:30–18:00	**T13** *Semen Spivak* The informativeness of the experiment in inverse problems of chemical kinetics and health insurance
18:00–18:30	**T14** *Alexey Pomerantsev* Successive Bayesian Estimation as a tool for chemometric modelling of kinetic data
18:30–19:00	**Free time**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Wednesday, February 29, 2012

08:00–09:30	**Breakfast**
Session 6	**Chair: Leon Rusinov**
09:30–10:30	**L3** *John Kalivas* Consensus multivariate calibration or maintenance without reference samples using Tikhonov type regularization approaches
10:30–11:00	**T15** *Sergey Kryvenko* The thermospectroscopic noninvasive glucometer
11:00–11:30	**Coffee break**
11:00–13:00	**Free time**
13:00–14:00	**Lunch**
Session 7	**Chair: Lutgarde Buydens**
14:00–15:00	**L4** *Beata Walczak* Dealing with nonlinearities
15:00–15:30	**T16** *Oxana Rodionova* Robust SIMCA bearing on non-robust PCA
15:30–16:00	**T17** *Sergey Tikhomirov* Practical Implementation of Chemometrics in OMNIC Program Set
16:00–16:30	**Coffee break**
Session 8	**Chair: Alexey Pomerantsev**
16:30–18:30	**Poster Session**
18:30–19:00	**Free time**
19:00–20:00	**Dinner**
20:00–00:00	**Scores & Loadings**

Thursday, March 1, 2012

08:00–09:30	**Breakfast**
Session 9	**Chair: Semen Spivak**
09:30–10:30	**L5** *Lutgarde Buydens* Strategies for the fusion of complex data
10:30–11:00	**T18** *Natalia Kireeva* Generative Topographic Maps (GTM) as a universal tool for data visualization, applicability domain analysis, structure – activity modeling and comparison of databases
11:00–11:30	**Coffee break**
11:00–13:00	**Free time**
13:00–14:00	**Lunch**
Session 10	**Chair: Veli-Matti Taavitsainen**
14:00–15:00	**L6** *Age Smilde* Metabolomics based systems biology
15:00–15:30	**T19** *Ekaterina Khrameeva* The impact of interchromosomal associations on the functional state of the human genome
15:30–16:00	**T20** *Pavel Mazin* Age changes and tissue differences of alternative splicing in the primate brain
16:00–16:30	**Coffee break**
Session 11	**Chair: Kim Esbensen**
16:30–17:00	**T21** *Yulia Monakhova* NMR and Chemometrics: a powerful combination for food analysis
17:00–17:30	**T22** *Maria Sushkova* PARCEL software as an instrument for InfraLUM type spectrometers calibration
17:30–18:00	**Closing**
18:00–20:00	**Free time**
20:00–00:00	**Banquet**

Friday, March 2, 2012

08:00–09:30	**Breakfast**
10:00–14:30	**Bus B1. Excursion**
09:00–11:30	**Bus B2**

Abstracts

L01. Coupling NIR and chemometrics for food and drug analysis

Federico Marini

Department of Chemistry, University of Rome "La Sapienza", Rome, Italy

The methods of vibrational spectroscopy, such as NIR and FT-IR, are fast and easy to use fingerprinting techniques, that can be used without (or with minimum) sample preparation, are not destructive, and have shown great effectiveness for the solution of many different kind of problems. However, it must be stressed that these techniques, in order to provide accurate and reliable answers, need to be coupled to a proper chemometric data treatment.

In this communication, the successful coupling of chemometric regression and classification techniques to experimental data collection in the near infrared region of the electromagnetic spectrum for the solution of problems related to the authentication of the quality of food and drugs will be presented.

In particular, in the framework of food quality control, the possibility of using spectral fingerprinting coupled to chemometric pattern recognition for tracing the origin of foodstuff will be shown. Attention will be focused on two high value-added products: olive oil and pistachio nuts. Additionally, the possibility of quantifying important nutritional parameters of cereals and powdered milk, and to evidence the presence of toxins will be shown.

As far as drug analysis in concerned, attention will be focused on the possibility of predicting the enantiometric excess of APIs in different formulations and commercial drugs in a fast and accurate way.

In all these frameworks, the importance of selecting the relevant spectral features on the quality of the final results will be stressed by means of representative examples.

L02. Multivariate process trajectories

Andrey Bogomolov

J&M Analytik AG, Essingen, Germany

Process analysis has been a favored chemometrics application area for decades. It covers a wide range of processes, from lab syntheses and metabolic pathways to the industrial production and environmental monitoring. Modern instrumentation and mathematics have reached a point where process analytical methods are rapidly growing in number and range of applications. As a consequence, methodological studies are starting to gain momentum in the field of process analysis.

The present lecture introduces a methodological concept of multivariate process trajectory. Being a common attribute of all kinds of processes, the trajectory provides a useful point of view for the systematic consideration of various analytical issues. Here, the process analysis workflow is presented in a systematic way, in relation to the objectives and methods of modern process analytical chemistry and technology. Several practical examples of industrial process analysis are presented and discussed in terms of capture, resolution and analysis of trajectories in a designed analytical space.

1. A. Bogomolov, *Chemom. Intell. Lab. Syst.* **108** ,49–63 (2011)

L03. Consensus multivariate calibration or maintenance without reference samples using Tikhonov type regularization approaches

J.H. Kalivas, J. Ottaway, J.A. Farrell, P. Shahbazikah
Department of Chemistry, Idaho State University, Pocatello, Idaho, USA

A majority of the work that goes into forming a multivariate calibration model comes in the collection and analysis of reference samples. Analysis of reference samples typically involves meticulous laboratory methods that are often time consuming and costly. The purpose of this talk is to report on a method that forms a calibration model without the need to collect reference samples (or very few). In the approach presented, a pure component analyte spectrum is used in conjunction with pure component interferent spectra, blanks, and/or constant analyte samples to form the model. For constant analyte samples, actual concentrations of the analyte are not needed; samples must just be constant within a certain tolerance. The methods are based on Tikhonov type regularization approaches in two norm and/or one norm (for sparse modeling, i.e., variable selection). Two approaches are studied for selection of tuning parameters; selecting values for 'a' model and using a consensus approach to select a family of models. With proper selection of tuning parameters, validation prediction errors are comparable to those from a model built with reference samples. Since no reference samples are needed, the approach is valuable for reducing both the time and cost of model building. The method is also applicable to updating a model from one set of experimental conditions to another set of experimental conditions. Results are presented for near infrared, ultraviolet-visible, and fluorescence spectral datasets.

L04. Dealing with nonlinearities

Beata Walczak

Institute of Chemistry, University of Silesia, Katowice, Poland

Increasingly more often we have to deal with non-linear systems. This is especially true, when studying really complex biological data. Neural Networks were the first efficient tool for modeling non-linear multivariate data, but they did not guarantee the optimal solution. From 1995, NN were systematically replaced by Support Vectors Machines, the linear method leading to optimal solution according to statistical learning theory, and able to deal with non-linearities due to the so-called 'kernel trick'. This 'kernel trick' can be applied to other methods (e.g., to RBF-PLS [1]). Nowadays, a long list of possible kernels exists. However, typical chemometric methods (such, as PLS, LDA, and PCR) can be applied with different dissimilarity/similarity matrices, as well as offering new and efficient tools for non-linear modeling [2]. In the arsenal of chemometric methods, there is also a method called nominal-PLS [3], well suited for the studies of dynamic biological systems.

The performance of the so-called dissimilarity-PLS and nominal PLS is going to be illustrated for simulated and real data sets.

1. B. Walczak, D.L. Massart, *Anal. Chim. Acta*, **331** (1996) 187-193

2. P. Zerzucha, M. Daszykowski, B. Walczak, *Chemom. Intell. Lab. Syst.*, (2011), doi:10.1016/j.chemolab.2011.11.005

3. H. Martens, Proceedings of the 6th International Conference on Partial Least Squares and Related Methods, Beijing, 4.7.2009, p.139-144 (E.V.Vinzi, M.Tenenhaus and R.Guan, eds.), Publishing House of Electronics Industry, http://www.phei.com.cn

L05. Strategies for the fusion of complex data

L. Buydens
Radboud University Nijmegen, Institute for Molecules and Materials, the Netherlands

Due to the ever-increasing production of complex data by a large variety of analytical technologies, chemometric data analysis and data mining have become crucial tools in modern analytical laboratories. The omics field is a clear example of this. In the last decade data fusion has become widespread in the omics field. It is common to perform linear data fusion. However, many biological or chemical data display complex e.g. non-linear parameter dependences. The linear methods are bound to fail in such situations. In this lecture two alternative approaches will be presented. The first one is based on the hierarchical fusion of mid-level fusion models. The second method is a nonlinear kernel fusion model in order to cope specifically with nonlinearities (1). We use our pseudo-sample approach (2,3) to reveal the contribution of the individual variables.

As example problem we use metabolomics datasets (GC-MS and NMR) of the cerebrospinal Fluid and plasma in a Multiple Sclerosis study.

1. S. Yu et al., Kernel-based Data Fusion for Machine Learning. Methods and applications in Bioinformatics and Text mining. Springer: Berlin 2011.
2. P. Krooshof et al., *Analytical Chemistry* **82** (2010) 7000-7007
3. Postma et al. *Analytica Chimica Acta* **705** (2011) 123-134

L06. Metabolomics based systems biology

Age K. Smilde
Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, the Netherlands

Metabolomics is the new branch on the tree of functional genomics tools. Broadly speaking, metabolomics aims at measuring comprehensively the metabolites of a biological system. Metabolomics can play a role in systems biology by probing systems biology concepts. Such concepts can be regulation, resilience, networks or phenotyping.

Although the field of systems biology is rapidly evolving, some general trends on combining metabolomics with systems biology models can be seen illustrating the interplay between systems biology and metabolomics. Examples are metabolic network inference for microbial systems, regulatory networks, challenge or resilience tests and nutrikinetics in human systems biology.

An introduction into metabolomics will be given and the way this tool can be used for exploring the above mentioned systems biology concepts will be illustrated with examples of real-life problems and data.

T01. Replication — ending decades of confusion

Kim H. Esbensen
Geological Survey of Denmark and Greenland, Copenhagen, Denmark

This presentation gives an overview of an issue, which has not always received close enough attention, the issue of “replication”, as in following three examples: i) replicate *samples*; ii) replicate *measurements*; iii) replicate *analysis*. What is really meant here — are all three statements identical? It is a general, but many times only tacitly implied connotation that a beneficial averaging is carried out by ”replication”. There are however many uncertainties and imprecise assumptions involved, which need careful analysis. It is here shown how incorporation of sampling and sample handling/preparation errors (Theory of Sampling, TOS) allows a unified understanding of the process of “replication”, resolving a decade long confusion with implications for data analysis, modeling and validation.

T02. Classification of objects on hyperspectral images

Sergey Kucheryavskiy
Department of Biotechnology, Chemistry and Environmental Engineering
Aalborg University, Esbjerg, Denmark

Hyperspectral imaging is a modern analytical technique combining benefits of digital imaging and vibrational spectroscopy. It allows to reveal and visualise spatial distribution of various chemical components.

In a hyperspectral image every pixel is a spectrum (usually VNIR, SWIR or Raman) of a depicted area. Such image can be represented as a cube or a set of 2D “slices” — one slice for each spectral band. It contains large amount of data and to reveal useful information proper methods for processing and analysis are needed.

Multivariate image analysis (MIA) is one of such methods widely spread among chemometicians. In most of the cases MIA treats pixels as objects, so an image cube has to be unfolded into a matrix, where rows represent pixels and columns — wavelengths. So in fact, multivariate image analysis works with an image as with a large set of spectra, without taking into account information about spatial relations of the pixels. This works well in general, especially for exploratory analysis or multivariate curve resolution, but for some specific tasks it is not beneficial at all.

One of such tasks is classification or clustering of objects on hyperspectral images. An object here means a set of connected pixels, fully or partly separated from other objects. That could be, for example, tablets, cereals, biological cells, etc. If objects from opposite classes are not absolutely different (e.g. there are similar pixels) it can lead to a problem. For example, if two different tablets have the same or similar excipient and different active ingredients, some of the pixels chemically will be identical. But these similar pixels will be associated with different classes when a classification model is being calibrated. This can give unstable model and poor classification results.

In the present work a classification method that combines classic image classification approach and MIA is proposed. The basic idea is to group all pixels and calculate spectral properties of the pixel group to be used further as a vector of predictors for calibration and class prediction. The grouping can be done with mathematical morphology methods applied to a score image where objects are well separated. In the case of small overlapping a watershed transformation can be applied to disjoint the objects.

The method has been tested on several simulated and real cases and showed good results and significant improvements in comparison with a standard MIA approach. The results as well as method details will be reported.

T03. Characterization of propolis using planar chromatography, direct analysis in real time mass spectrometry and multivariate data analysis

Elena S. Chernetsova[1,2,3], *Petar Ristivojevic*[4], *Gertrud E. Morlock*[1,5]
[1]*Institute of Food Chemistry, University of Hohenheim, Stuttgart, Germany*
[2]*National Research Center "Kurchatov Institute", Moscow, Russia*
[3]*People's Friendship University, Moscow, Russia*
[4]*Faculty of Chemistry, University of Belgrade, Belgrade, Serbia*
[5]*Institute of Nutritional Science, Justus-Liebig-University of Gießen, Germany*

Sophisticated statistical tools are required to extract the full analytical power from high-performance thin-layer chromatography (HPTLC). Especially the combination of HPTLC separations or fingerprints with chemometrics is rarely used and this utility has not been harnessed so far. Also, the newly developed prompt mass spectrometric method, employing Direct Analysis in Real Time ionization, is perspective for sample characterization and differentiation by means of multivariate data analysis, but needs studies due to the methods novelty.

This is a first study on propolis differentiation using HPTLC, Direct Analysis in Real Time mass spectrometry (DART-MS) and multivariate data analysis. The results obtained by the chemometric evaluation of HPTLC and DART-MS data provide complementary information. The complexity, expense, and analysis time are significantly reduced due to the use of statistical tools for evaluation of fingerprints. The developed approaches allowed to categorize 91 propolis sample from Germany or other locations using their flavonoid and phenolic acid profile. The higher level of confidence can be obtained when combining these 2 different approaches (HPTLC and DART-MS) for ultrafast sample characterization. HPTLC with selective post-chromatographic derivatization provides the information on the functional groups in marker compounds, while DART-MS helps getting the information on possible elemental formulae of the principal components (flavonoid and phenolic markers).

This work was financially supported by the joint program of DAAD and the Ministry of Education and Science of the Russian Federation 'Mikhail Lomonosov' – 'The development of a scientific potential of a higher school (2009-2011)' (project #2.2.2.3/15112) and by the program Erasmus Mundus Action 2 "IAMONET-RU".

T04. 3wayPLS modeling of potentiometric multisensor system response for quantification of bitter taste in pharmaceuticals

Dmitry Kirsanov[1,3], Yulia Blinova[1], Alisa Rudnitskaya[2,3], David Clapham[4], Andrey Legin[1,3]
[1]Chemistry Department, St. Petersburg State University, St. Petersburg, Russia
[2]CESAM/Chemistry Department, Aveiro University, Aveiro, Portugal
[3]Sensor Systems LLC, St. Petersburg, Russia
[4]GSK Pharmaceuticals, Park Road, Ware, UK

In the present work the application of the artificial sensory system (Electronic Tongue, ET) for quantification of the bitter taste of various chemically dissimilar substances is reported. This approach to artificial quantification of bitterness is practically feasible and may be particularly useful on the early stages of development of novel API (active pharmaceutical ingredients) in pharmaceutical research and for flavour control of various pharmaceutical compositions, healthcare products and food ingredients [1].

The measurements were performed in a set of bitter substances provided by GlaxoSmithKline and assessed by human and rat sensory panels to produce bitterness intensity scores for each substance in different concentrations. The set consisted of 8 substances mostly used as API of various chemical nature, both inorganic and organic – potassium nitrate, quinine, caffeine, paracetamol, chlorhexidine, ibuprofen, etc. It must be pointed out that most of the substances studied belong to different classes and has got little in common both in composition and in properties of solution when dissolved. Furthermore, their behavior in water solutions is very different, some of them have acidic properties, and some of them are not ionized at neutral pH values. The latter issue is of crucial importance for potentiometric sensors as they respond to ionic species only. To circumvent this limitation we designed the experiment where all substances in all concentrations where studied at different pH levels ranging from 2 to 10, thus promoting their ionization. This experiment yielded 3 way data array (sensors X samples X pH). 3wayPLS regression models where constructed with reference data from rat sensory panel (% of inhibition values). These models where validated by various methods and were further applied for prediction of bitterness of unknown substances employed neither for calibration nor for validation. Predicted values were compared with rat panel estimates for these substances.

The details of experiment as well as results will be discussed in the presentation.

1. A. Legin, A. Rudnitskaya, D. Clapham, B. Seleznev, K. Lord, Yu. Vlasov, *Anal. Bioanal. Chem.*, **380**, 36-45, (2004)

T05. Principal Component Analysis of an XI-XV century silver coins' XRF dataset and verification by additional multilinear methods

Anita Rácz[1], János Elek[1], Károly Héberger[2], Róbert Rajkó[3], Adrienn Lengyel[1]
[1]Analab Ltd, University of Debrecen
[2]Chemical Research Center Hungarian Academy of Sciences
[3]University of Szeged

The advantages of the X-ray fluorescent elemental analysis such as quickness, negligible sample preparation demand and most of all non destructivity make the method a very promising additional data source to support the work of archaeologists. With use of the current handheld instruments the examination of sensitive artefacts is available on site without sampling or taking the subjects to a laboratory.

Though the results of the elemental composition are readily accessible the recognition of the ulterior coherences in the produced dataset needs multilinear evaluation methods. The primary output of an XRF spectrometer is the matrix of the excitation voltages and the corresponding fluorescent intensities. The secondary output is the elemental composition derived from the primary matrix.

In this study 248 silver coins of various Hungarian regal dynasties from XI-XV century were analysed in order to see if any sort of grouping can be observed among the coins. The first question of the statistical processing is whether compositional information is sufficient or the whole spectral information is necessary to find the coherent pieces.

Data evaluation was made initially by principal component analysis. The results show two big, historically non overlapping sets of the coins. Ranking was tested with additional chemometric methods such as linear discriminant analysis (LDA), classification and regression tree (CART) and partial least squares (PLS).

There are many collections worldwide that contain pieces with uncertain origin. Our work shows, that ranking of these pieces becomes possible by selection of the appropriate test method and constructing a didactic set of minted coins with known origin and an analyzing algorithm.

T06. Rational function ridge regression

Veli-Matti Taavitsainen
Helsinki Metropolia University of Applied Sciences, Finland

Quite often linear methods fail to give good enough fit in chemometric problems. However, many of the existing nonlinear methods are tedious or slow to use in highly multilinear cases. Recently, a simple idea of combining rational functions with ridge regression by linearization has shown to be a potential method in solving chemometric problems. This presentation gives the basics of the method and some examples in different fields of application.

Taavitsainen, V-M. (2010), *J. Chemometrics*, **24**, 665-673.

T07. The best method of noise filtering

Yuri Kalambet[1]*, Yuri Kozmin*[2]*, Sergey Maltsev*[1]
[1]*Ampersand Ltd., Moscow, Russia*
[2]*Institute of Bioorganic Chemistry RAS, Moscow, Russia*

The ultimate solution of the problem “Which smoothing method is better” is offered. A method of noise filtering based on confidence interval evaluation is described. In the case of the approximation of a function by least-squares method, the minimal confidence interval is used as a criterion for the selection of the proper parameters of the approximating function. In the case of the polynomial approximation optimized parameters include the degree of the polynomial, the number of points (window) used for approximation, and the position of the window centre with respect to the approximated point. Special considerations on confidence interval evaluation and quality of polynomial fit using noise properties of the data array are discussed. The Method provides the lowest possible confidence interval for every data point.

The Method is demonstrated using generated and measured data. Improvement of noise reduction compared to competing methods can vary depending on the input data, but always exists. Excellent noise reduction properties are combined with conserved object shape (e.g. chromatographic peak, photographic object) without artefacts. The method requires extra computations, which can be easily paralleled.

Improvement of signal to noise ratio compared to the best alternative methods is seldom less than 2 and, depending on the nature of the data, can be as high as 15. Peak shape does not change, and baseline noise decreases drastically.

T08. Identification and age estimation of blood stains on coloured backgrounds by near infrared spectroscopy

Gerda Edelman
Academisch Medisch Centrum, Amsterdam, the Netherlands

Blood stains are an important source of information in forensic casework. After a non-destructive identification of blood stains, DNA can provide information about who was involved and the patterns of the blood stains reveal insight in what might have happened. An interesting question that remains is when the blood left the human body which would help crime scene investigators to determine the temporal aspects of a crime.

When blood exits the human body, oxyhemoglobin auto-oxidizes into methemoglobin, which in turn denatures into hemichrome. These reactions cause a colour change from red to brown. We previously showed that visible reflectance spectra can be used to measure this colour change quantitatively and subsequently estimate the age of blood stains. We successfully captured an entire crime scene and automatically localized fresh and old blood stains within the scene using hyperspectral imaging. However, the use of visible reflectance spectra is severely hampered when blood stains are deposited on dark backgrounds which absorb most of the visible light.

In the present study, we aimed to determine whether Near Infrared (NIR) reflectance spectroscopy can be used to distinguish blood from other substances on coloured backgrounds, and to estimate the age of blood stains on coloured backgrounds. To achieve these objectives, the following questions were addressed:

1. Do blood stains have specific spectral features in the NIR region, which are useful for blood stain identification and age estimation?

2. What is the sensitivity and specificity of blood stain identification on coloured backgrounds using NIR reflectance spectroscopy?

3. What is the prediction error if we estimate the age of blood stains on coloured backgrounds using NIR reflectance spectra?

NIR reflectance measurements were performed repeatedly on blood stains and other substances on several coloured backgrounds. To minimize the spectral variation caused by non-chemical properties (like non uniform light scattering and different sample thicknesses), several pre-processing methods were applied to the spectra. To find spectral features useful for the identification of blood stains, we calculated Pearson's correlation coefficient between all blood stains and the correlation between blood stains and other samples in several wavelength regions. For the age estimation task, a partial least squares regression model was created and tested on blood stains on black cotton.

We have identified spectral features in the NIR region which can be used for blood stain identification and age estimation even when blood is deposited on black backgrounds. The sensitivity and specificity of blood stain identification using NIR reflectance spectroscopy were both 100 %. The root mean squared error of prediction of the age of blood stains on black cotton using NIR reflectance spectra was 8.9 %.

This finding is a step toward the implementation of blood stain localization, identification and age determination in forensic casework, where a large variety of backgrounds can be encountered. Because NIR spectroscopy is non-destructive, the blood stain is preserved for subsequent analysis in the forensic lab e.g. for DNA analysis.

T09. Process chemometrics in the Dow Chemical company

Zdravko Stefanov and Leo Chiang
Dow Chemical, Freeport, USA

Process chemometrics (chemometrics methods applied to chemical processes) is a technology that has gained increasing significance in Dow Chemical's plant troubleshooting efforts. It has proved to be highly efficient in the discovery of root causes for a multitude of plant problems. Solving these problems has lead to multimillion dollar savings for the company. This presentation will show several examples of successful application of process chemometrics for plant troubleshooting. The examples come from batch and continuous processes and show the versatility of the technology. The importance of the online application of the technology will be discussed as well.

T10. Fault diagnosis in chemical processes and equipment with feedback

Rusinov L.A., Vorobjev N.V., Kurkina V.V.
Saint-Petersburg State Technological Institute, Saint-Petersburg, Russia

Potentially dangerous processes (PDP) with the repeated cycles (recycles), and the control loops (sensors, actuators, valves, etc.) represent special complexity, because the loops mask the feedback. As a result, the diagnostic system learns about a fault when an abnormal situation has developed so far that the recycle agitator (or controller) resources are not suffice for regime maintenance. Therefore, it is necessary to build diagnostic model (DM) of the process section in the closed loop formed by feedback (further - object). However, it is not possible to identify the fault cause using deviations from this model. A bank of models accounting for the possible abnormal situations is necessary.

If the residuals $\boldsymbol{\varepsilon}$ between object output variables and the model corresponding to process's normal functioning exceed some threshold, a fault is considered to be detected. It leads to initiation of the models describing object faults. The fault described by the model which at this moment displays the smallest residual $\boldsymbol{\varepsilon}_i$, will obviously be the most probable. Thus, the method requires the presence of diagnostic models, simple in implementation in real time. This requirement can be a severe limitation for its usage. The second limitation is the necessary knowledge of all possible faults of the object under control. If the model of the arising fault is not in the bank, it would be detected, but not correctly identified.

DM of objects with feedback can be constructed by the same principles, as DM PDP. They should be as simple as possible for facilitation of operation in real time. Besides, some objects have their own microprocessor control systems. Usually these systems carry out diagnostics only on the state of object electronics. Therefore, it is important also to have the possibility of implementation of DM directly in microprocessor object systems.

From possible approaches to on-line diagnostic of faults in objects of the class under consideration one of the most promising is fuzzy DM that includes the specification of each fault in the form of fuzzy rules. The result of model operation is defined by the total contribution of all rules included in the model. The number of rules in the model is determined by clusterisation of an array of the historical data received from the object during its normal operation and in the presence of faults. Each cluster introduces some operational field of the system connected to a certain fault. Then the number of clusters c, discovered in a data array, will be equal to number of fuzzy rules. Normally it is not known a priori and must be determined in the course of clustering.

In the case of usage of Kalman filters the object DM is formed in a state space. The Kalman filter actually searches for an optimal estimation by least squares method. Since the majority of PDP are essentially nonlinear, extended modifications of Kalman filters are used: EKF and UKF for DM.

These methods were explored by constructing modules for control and fault diagnosis for electropneumatic valves with the positioners, located in control loops, and faults in chemical technological process with recycles.

T11. Identification of unknown organic compounds based on comparison of electron ionization mass spectra

A.S. Samokhin, *I.A. Revelsky*
Lomonosov Moscow State University, Chemistry Department, Moscow, Russia

Gas chromatography/electron ionization mass spectrometry is widely used for identification of semi-volatile organic compounds. Usually identification is based on searching mass spectra of unknown compound against a large mass spectral database. Experimental and library mass spectra generally have some differences caused by dissimilarity of registration conditions (type and model of mass spectrometer, method of instrument calibration, ion source temperature and etc.). The existence of such differences very often makes impossible caring out reliable identification using only search against commercial mass spectral database. However the list of possible candidates may be obtained by this way. For the most reliable identification, mass spectra of unknown and reference compounds should be registered under identical experimental conditions and compared. The main problem is how to compare mass spectra containing hundreds of peaks. Visual comparison is very subjective and does not allow quantitative estimation of similarity between mass spectra. Application of mathematical algorithms used in library search programs requires carrying out a large number of additional experiments (which are needed to determine critical value of similarity index provided reliable identification). Comparison of intensities of individual mass spectral peaks is used in real practice, however only several (usually three) peaks are considered. Moreover tolerance for intensities of mass peaks may be quite large and may theoretically result in false positive identification.

We have investigated possibility of comparison of both “reduced” and “full” electron ionization mass spectra using principal component analysis. Several organic compounds referred to different classes were considered.

T12. Chemometrics for the resolution of time-resolved spectra in photochemistry

C. Ruckebusch, O. Devos, M. Sliwa
LASIR CNRS Université Lille Nord de France

The extraction of detailed photochemical information is expected from multivariate analysis of transient absorption spectra, e.g. number of transient species, connectivity and kinetics of these species. Specific chemometric procedures are often required since instrument characteristics strongly alter the data structure. For instance, ultrafast measurements are affected by convolution of the signal of the photoinduced species with the instantaneous response function of the instrument. In addition, the optical characteristics of ultrafast pump-probe experiments result in the presence of intrinsic artifacts superposing onto the weak difference absorption spectra of target molecules leading to strong data distortions. Thus, chemometric analysis of femtosecond spectroscopy data is a topic in itself. This presentation will provide an overview of the strategies developed to unravel the complex underlying information including data pre-processing, data exploration, data modeling and kinetic fitting [1]. The examples will focus on the characterization of the ultrafast photodynamics of photoswitchable molecules, of interest for applications in high-resolution imaging or for the design of new photochromic materials. In particular, how some of the issues can be tackled using multivariate curve resolution will be detailed [2].

1. C. Ruckebusch, M. Sliwa, P. Pernot, A. de Juan, R. Tauler, J. Photochem. Photobiology C: Photochemistry Reviews doi:10.1016/j.jphotochemrev.2011.10.0022.

2. N. Mouton, A. De Juan, M. Sliwa, C. Ruckebusch, *Chemom. Intell. Lab. Syst.*, 2011, **105**, 74–82.

T13. The informativeness of the experiment in inverse problems of chemical kinetics and health insurance

S.I. Spivak
Bashkir State University, Ufa, Russia
The Institute of Petrochemistry and Catalysis RAS, Ufa, Russia

Modern technology allows parallel calculations in interpreting experimental research to address such classes of problems whose solution by traditional methods is extremely difficult. In the first place it is important to reduce computing time for solving inverse problems. By parallelism we mean a partition of the whole problem into subproblems of smaller dimension, the analysis can increase substantially the level reliability of the solution of inverse problems. Possibility of providing such "internal parallelism" should follow from the physical content of the tasks. The purpose of this paper is the selection of the internal division in the solution of inverse problems of chemical kinetics and health insurance.

T14. Successive Bayesian Estimation as a tool for chemometric modelling of kinetic data

A.L. Pomerantsev
Semenov Institute of Chemical Physics RAS, Moscow, Russia
State South Research & Testing Site RAS, Sochi, Russia

Spectroscopy was and still is a principal source of data in chemometrics. However, a new pronounced trend in today’s chemometrics is a frequent utilization of kinetic data. Hard modelling approach in multivariate curve resolution (MCR) is the main example, which combines kinetics (chromatography, in a broad sense, is a kinetic process) with spectroscopy. Besides this, other instances are known, such as Differential Scanning Calorimetry (DSC) [1], Thermo-Mechanical Analysis (TMA) [2], Dissolution Profiles (DP) [3], etc. Unlike MCR, these techniques utilize kinetic data *per se*, as a primary data source.

The presentation aims at generalization of the experience in a specific approach to kinetic modelling. This approach extracts relevant information from the kinetic data, and then utilizes it in the consequential chemometric analysis. Formal kinetic (*aka* hard) modelling, e.g. the direct and inverse kinetic problems, is beyond the scope.

Kinetic data are often presented in blocks that can be processed separately, but in relation to other blocks. The relations manifests themselves via kinetic parameters, which can be either block-partial, or block-common, i.e. shared between several blocks. In typical real-world cases, the partial parameters are counted in hundreds, while the common ones in dozens. For this reason, simultaneous estimation of all parameters is a difficult task.

Successive Bayesian Estimation method is rather straightforward. Blocks are processed not simultaneously but successively, one by one in the block-wise steps. During such a fitting the estimates that have been found at the previous step are further used as a priori information for the processing of the next block. At step (block) b, parameters $\mathbf{a}$ are obtained by minimization of objective function

$$Q_b(\mathbf{a}) = S_b(\mathbf{a}) + B_{b-1}(\mathbf{a}),$$

where S_b is a least squares term, calculated for current block b

$$S_b(\mathbf{a}) = \sum \left(y_{ib} - f_b(t_{ib}, \mathbf{a})\right)^2$$

and B_{b-1} is a Bayesian term

$$B_{b-1}(\mathbf{a}) = s_{b-1}^2 \left[N_{b-1} + (\mathbf{a} - \hat{\mathbf{a}}_{b-1})^{\mathrm{t}} \mathbf{H}_{b-1} (\mathbf{a} - \hat{\mathbf{a}}_{b-1})\right].$$

Values s_{b-1}^2 (variance), N_{b-1} (DoF), $\hat{\mathbf{a}}_{b-1}$ (estimates), and $\mathbf{H}_{b-1}$ (*a priori* matrix) keep track of the posterior information acquired at previous block b–1. Matrix

$\mathbf{H}_{b-1}$ accounts for the common parameters only, that is why it is recalculated from posterior matrix $\mathbf{A}_{b-1}$ with elimination of all information on the particular parameters.

Several real-world examples illustrate the SBE implementation.

1. A.L. Pomerantsev, O.Ye. Rodionova, *Chemom. Intell. Lab.Syst.*, **79**, 73 (2005)
2. E.V. Bystritskaya, A.L. Pomerantsev, O.Ye. Rodionova, *J. Chemometrics*, **14**, 667 (2000).
3. A.L. Pomerantsev, O.Ye. Rodionova, et al *Analyst,* **136**, 4830 (2011)
4. A.L Pomerantsev, *Chemom. Intell. Lab. Syst.*, **66**, 127 (2003)

T15. The thermospectroscopic noninvasive glucometer

Kryvenko S.S., Pulavskyi A.A.
Biopromin LTD, Kharkov, Ukraine

Diabetes mellitus is a chronic condition, which occurs if the pancreas cannot produce sufficient insulin or if the body cannot effectively use the insulin produced by it. Hyperglycemia, hypoglycemia, and related metabolic disorders can inflict serious harm, even lethal outcomes, on practically all the body systems, especially the nervous and blood-vascular systems. The control of the glucose concentration in blood is the top-priority task to prevent complications and consequences of diabetes mellitus. The methods of control, which are available in the market, are invasive, i.e. require blood sampling. A noninvasive method for quick, painless, safe, and convenient determination and control of the blood glucose concentration is desirable.

The purpose of this work is to develop a simple-to-operate noninvasive gauge of the glucose concentration based on optical and infrared spectral analysis of soft tissues at bifurcation points.

The noninvasive analysis of the whole-blood glucose concentration can be divided into several groups. Of particular interest is epidermal analysis which can be conducted by optical and infrared spectroscopy. After blood sampling, measurement of a thermally stabilized spectrum is carried out. An advantage of such measurements is the relative insensitivity to the location of the probe.

Nonlinear relationships between the input and output variables complicates the calibration task. One of the most effective methods of the removal of the nonlinearity in correlated data is neural networks and modern chemometric methods. The methods of RBF-PLS (radial basis function - partial least squares) and multilayer perceptron were used. With the help of the heuristic method we have determined that the best nonlinear function, which interpolates the present dependencies in a good way, is the Fermi function (exponential sigmoid) which is the function of activating the hidden layer of the neural network and the replacement of the radial function in RBF-PLS.

The experiment involved a total of 238 male sufferers of type 1 and 2 diabetes as well as healthy patients. The age of the patients varied from 16 to 77 years with an average age of 43 years. It is important to note that during the experiment hematocrit was not measured. The concentration of glucose varied from 2 mmol/l to 29 mmol/l. The average value of the concentration of glucose equalled 7.1mmol/l. The patients were divided into two groups – calibration and test – with the help of the Kennard-Stone method. The calibration group consisted of 166 patients and the test group consisted of 72 patients. The mean-square error of prediction (RMSEP) equalled 1.34 mmol/l. The Clarke and Parks error grids were used for the analysis of the clinical accuracy of the obtained results. All of results are within the allowed limits of the generally accepted error grids, 87.5% from the required 95% correspond to the requirement of the standard ISO 15197:2003.

T16. Robust SIMCA bearing on non-robust PCA

O.Ye. Rodionova[1], *A. L. Pomerantsev*[1, 2]
[1]*Institute of Chemical Physics RAS, Moscow, Russia*
[2]*State South Research & Testing Site RAS, Sochi, Russia*

It is well known that outliers strongly influence the performance of the projection methods (PCA) and classification techniques (SIMCA). To lessen the influence of outliers various robust procedures have been developed.

The main steps of the SIMCA method are as follows:

(1) Data pre-processing;
(2) PCA;
(3) Construction of the classification rules.

Spherical PCA (S-PCA) firstly projects all data objects onto the hyper sphere, afterwards uses regular PCA. Thus, robust procedure is applied on the first step, pre-processing. Another approach is to utilize various robust PCA algorithms, such as Projection Pursuit (PP). Application of the robust procedure on the second step is the most straightforward approach. Leverages and residuals are also 'robustified' by normalization with the help of robust estimators. Thus, ordinarily robustification is employed on the first or the second steps and application of a robust classification rule.

We propose to robustify SIMCA only on the third step. This is done by robust, data-driven, estimations of the degrees of freedom for the orthogonal and score distances distributions and therefore application of a special procedure for calculation of the robust acceptance area. Moreover, the acceptance area can be calculated for various significance levels. As a result the classification problem is solved with respect to a specified value of the probability of type I error α.

The procedure is demonstrated on the real world example of packed raw materials. Results are compared with robust SIMCA based on PP.

T17. Practical Implementation of Chemometrics in OMNIC Program Set

S. Tikhomirov
Intertech Trading Corporation, Moscow, Russia

Practical application of chemometrics is one of the most important trends in modern analytical chemistry in general and molecular spectroscopy in particular. Building a complete program set for such application is a real challenge for any company since effectiveness of such program set is determined by it’s practical implementation. Building of easy, clear and intuitive user interface for creation, testing and modification of chemometric models is a central task here.

Our report describes big program complex manufactured by Thermo Scientific (USA) especially for molecular spectroscopy: FT-IR, FT-NIR and Raman. All parts of this system are tightly integrated: OMNIC program controlling an analytical instrument can call chemometric method developed by TQ Analyst program. This last program can, in its turn, take control over the spectrometer for standard spectra measurement. All TQ methods can be used for quantification of separate spectra as well as for spectra arrays: maps, kinetic data, etc. OMNIC program complex can be easily automated for routine and research measurements with MACROS BASIC module.

Some application examples from different labs are discussed.

T18. Generative Topographic Maps (GTM) as a universal tool for data visualization, applicability domain analysis, structure-activity modeling and comparison of databases

N. Kireeva[1,2]*, I.I. Baskin*[1,3]*, S.L. Kuznetsov*[2]*, E.P.Dyachkov*[4] *and A. Varnek*[1]
[1]*Laboratoire d'Infochimie, Strasbourg, France*
[2]*Frumkin Institute of Physical Chemistry and Electrochemistry RAS, Moscow, Russia*
[3]*Department of Chemistry, Moscow State University, Moscow, Russia*
[4]*Kurnakov Institute of General and Inorganic Chemistry RAS, Moscow, Russia*

This presentation concerns Generative Topographic Maps (GTM) [1], which is a universal tool to visualize the chemical space, to develop classification models, to estimate the applicability domain of models and to compare databases of chemical compounds. The model calculations performed on the DUD [2] and HERG (human ether-a-go-go related gene) channel inhibitors datasets using different types of descriptors to illustrate the utility of GTM.

As a tool for visualization GTM overcomes most of drawbacks of such popular approaches as Principal Component Analysis (PCA), Sammon Mapping (SM) and Self-Organizing Maps (SOM). The probabilistic character of the GTM approach appears to offer additional advantages, which directly stem from the rigorous probabilistic character of the GTM approach. Thus, GTM can also be used to assess an overlap between the datasets, to develop classification models if information about activities of compounds is available and to define the applicability domain of models.

In this study, we consider some aspects of GTM models for data visualization, the chemical interpretation of the maps and application of GTM as a tool for structure-property modelling, applicability domain estimation and data distribution comparison. Our ultimate goal is to present GTM as a unique universal approach for data processing in chemoinformatics.

This work has been supported by RFBR (grant 11-03-00161-a).

1. Bishop, C.M. and M. Svensen, *Neural Computation* **10** (1998) 215-234.
2. Huang, N.; B.K. Shoichet, and J.J. Irwin, *J. Med. Chem.* **49** (2006) 6789-6801.

T19. The impact of interchromosomal associations on the functional state of the human genome

Ekaterina E. Khrameeva[1,2]*, Gennady G. Fedonin*[1]*, and Mikhail S. Gelfand*[1,2]
[1]*Institute for Information Transmission Problems RAS, Moscow, Russia*
[2]*Moscow State University, Moscow, Russia*

The chromatin state is one of the main determinants of transcription rate in eukaryotes. Dynamic chromatin movements and interactions play a crucial role in gene regulation. Adjacent genes can be packed together in distinct chromatin domains that ensure coordinated gene expression. Evidence is emerging that distant genes located on different chromosomes also can interact, and their proximity might be essential for coordinated regulation. However, it has been unclear whether these interchromosomal associations are exceptional, or occur frequently.

We systematically analyzed genome-wide interchromosomal interactions in the nuclei of human cells. 3D data from [1] were associated with the results of several high-throughput studies of the chromatin functional state [2]. All pairs of regions from different chromosomes were divided into groups according to their proximity, and the distribution of various chromatin marks was calculated within these groups and then compared between the groups.

The results show that, indeed, gene regions that are spatially close tend to have similar patterns of histone modifications, methylation state, open or closed chromatin state, and expression level. Spatially close genome domains tend to have similar chromatin state and to be coregulated and coexpressed.

Moreover, we found that interacting domains may produce transcripts composed of sequence segments coming from two different chromosomes. We analyzed chimeric transcripts as determined by genome mapping of paired-read RNA-Seq data [3,4] and observed that the frequency of pairs mapping to two different genome loci is higher among spatially proximal regions. We suggest that these transcripts might be formed by trans-splicing.

In summary, interchromosomal associations seem to be much more common than previously believed, and they likely play important roles in the regulation of gene expression.

This research has been supported by State contracts, grants of the Russian Foundation of Basic Research, programs "Molecular and Cellular Biology" and "Basic Science for Medicine" of the Russian Academy of Sciences.

1. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, et al. *Science* **326**, 289–293 (2009).
2. *ENCODE Project Consortium Science* **306**, 636–640 (2004).
3. A. G. Xu, L. He, Z. Li, Y. Xu, M. Li, et al. *PLoS Comp. Biol.*, **6**(7), e1000843 (2010).
4. M.F. Berger, J.Z. Levin, K. Vijayendran, A. Sivachenko, X. Adiconis, et al. *Genome Res.* **20**, 413–427 (2010).

T20. Age changes and tissue differences of alternative splicing in the primate brain

Pavel Mazin[1,2], Mikhail Gelfand[1,2] and Philipp Khaitovich[3]
[1]Moscow State University, Moscow, Russia
[2]Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia
[3]MPG-CAS Partner Institute of Computational Biology, Shanghai, China

Humans differ strikingly from their close relatives, such as chimpanzees, in terms of anatomy, behavior and cognitive abilities [1]. Genetically, however, the human and the chimpanzee genomes are extremely similar. Recent studies have shown that gene expression differs substantially between humans and other primates and these differences could be linked with the accelerated brain evolution in the human lineage [2]. Alternative splicing allows a single gene to produce multiple transcripts and, consequently, multiple proteins, by utilizing different splice sites in pre-mRNA. The microarrays-based analysis has shown that at least 4% of expressed genes splice differently between humans and chimpanzees [3].

Here, we used next-generation sequencing technology to investigate splicing differences between humans, chimpanzees and rhesus macaques in two brain regions, prefrontal cortex and cerebellar cortex, in newborns and adults. We sequenced poly-A+ transcriptome in 30 individuals: 10 humans, 10 chimpanzees and 10 rhesus macaques, resulting in more than 15 millions reads. To reduce individual variance, we pooled RNA from 5 individuals with similar age resulting in 12 samples distributed among the three species, two brain regions and two ages. Our results show that most of the splicing variance in our data relates to interspecific differences (55%) and relatively few corresponds to brain region and age (8% and 7% respectively). But on the gene expression level, brain region and age are responsible for 40% and 28% variance, respectively, while interspecific differences take only 12% of the total variance. So, regulation of gene expression and splicing evolve differently. To investigate the dynamics of splicing changes during the whole lifespan, we sequenced prefrontal cortex from 13 humans, 15 chimpanzees and 15 rhesuses with ages ranging from birth to old age. In this sample, 5473 (52% of expressed) genes splice differently between species (q-value = 0.05) and 1657 genes change splicing with age in at least one species (1237, 507 and 421 in human, chimpanzee and rhesus respectively), 230 of these genes have significantly different age-related splicing patterns between species. Intriguingly, most of them (149 genes) change significantly with age only in human but not in chimpanzee or rhesus. Thus, splicing differences between species are not restricted to average differences in the exon inclusion rates, but also encompass regulation of splicing during development and aging and may contribute to human-specific features of the brain development.

1. G. Klein (1989) The Human Career: Human Biological and Cultural Origins, The Univ. of Chicago Press, Chicago, 1989.

2. W. Enard et al (2002), *Science* **296**:340-343

3. J.A. Calarco et al (2007), *Genes Dev.* **21**:2963-2975

T21. NMR and Chemometrics: a powerful combination for food analysis

Yulia B. Monakhova[1,2,3]*, Hartmut Schäfer*[2]*, Eberhard Humpfer*[2]*, Manfred Spraul*[2]*, Thomas Kuballa*[1]*, Dirk W. Lachenmeier*[1]
[1]*Chemisches und Veterinäruntersuchungsamt (CVUA) Karlsruhe, Germany*
[2]*Bruker Biospin GmbH, Rheinstetten, Germany*
[3]*Department of Chemistry, Saratov State University, Saratov, Russia*

Nuclear magnetic resonance (NMR) spectroscopy is gaining more and more importance in mixture analysis with one large application field in food analysis. In this study, 400 MHz 1H nuclear magnetic resonance (NMR) spectroscopy coupled with different chemometric approaches (principal component analysis, PCA; partial least squares regression, PLS; soft independent modeling of class analogy, SIMCA) is used in the context of different projects in food analysis.

Our first example is the classification of milk and milk substitutes based on their NMR spectra. The best PCA model obtained in the 6 - 3 ppm region suggests that the samples can be separated into five groups: lactose-containing and lactose-free milk, as well as oat, rice and soy substitutes. Using SIMCA analysis, a qualitative classification into these five groups was possible with 100% accuracy. Furthermore, quantitative data regarding nutrition labeling can be predicted from the same spectra using PLS regression.

The second example deals with recent appearance of a phenomenon related to consumption of pine nuts leading to taste disturbance (so-called "pine mouth" or pine nut syndrome (PNS)). Our aim was to develop a nontargeted NMR screening method that would allow for rapid selection of suspect pine nut samples. PCA shows that the samples can be separated according to their botanical species into three (1H NMR) or two (13C NMR) groups. Botanical species *Pinus armandii*, *P. koraiensis*, *P. massoniana* cultivated in Asian countries (China, Japan and Korea) are in one group, whereas two separate clusters of *P. pinea* (harvested in Mediterranean Europe and Near East) and *P. gerardiana* (common in Afghanistan, Pakistan and India) are also observed. The PNS species were only found in the sample group with a Chinese origin.

Our method is a good alternative to macroscopic analysis or to time-consuming GC-MS analysis.

NMR spectroscopy coupled with several chemometric techniques is also introduced for the analysis of alcoholic beverages in the context of governmental health and safety control. To improve overall sensitivity of NMR measurements, suppression of the eight 1H NMR frequencies of water and ethanol (the OH singlet of both water and ethanol, as well as the CH2 quartet, and CH3 triplet of ethanol) was applied. At the first step, to detect potentially unsafe samples especially in the context of unrecorded alcohol, a nontargeted approach based on principal component analysis (PCA) was applied. The PCA scores plot of 304 samples clearly shows 7 conspicuous samples with highly divergent scores from the rest of the samples. These samples contained toxic substances such as diethyl phthalate (DEP) or polyhexamethyleneguanidine (PHMG), as well as extremely high concentrations of methanol or ethyl carbamate. Multivariate curve resolution and spectra computation methods were used to confirm the presence of PHMG and DEP in the investigated beverages. In addition to the nontargeted approach, partial least squares (PLS) regression allowed to quantify 13 parameters with high correlation to gas chromatographic reference analysis (R2 = 0.92–0.98). Furthermore, the genotoxic carcinogen ethyl carbamate can be quantified based on PLS calibration. The accuracy of the PLS model evaluated by test set validation was found sufficient to providing reliable ethyl carbamate quantification below the level of 1 mg/L recommended by the European Commission.

In conclusion, the combination of high-throughput NMR and statistical analysis is a powerful tool in analysis of complex food matrixes, including milk and its substitutes, pine nuts and alcoholic beverages. The main advantage of the chemometric models is that they can be successfully used for quantification of substances whose resonances overlap with signals of other compounds. Therefore, NMR spectroscopy is judged as suitable for the rapid routine analysis of the mentioned products and the application range will be extended to further matrices in the future.

T22. PARCEL software as an instrument for InfraLUM type spectrometers calibration

<u>Sushkova M.V.</u>, Zharinov K.A.
Lumex Ltd., Saint Petersburg, Russia

The PARCEL software being developed by the Lumex company allows to calculate the optimal calibration models for determining the properties of the analyzed objects based on their absorption spectra. The program works with data obtained on near infrared spectrometers InfraLUM. The operating principle of the devices is based on recording the absorption spectra. This equipment is used mainly in the agricultural field to determine the key properties of various crops. Under the calibration model the relation connecting the determined properties of samples with spectral data is understood.

In the PARCEL software, as in similar programs, the possibility of choosing the optimal parameters of the calibration model by “trial-and-error” method is realized, e.g. by iterating of different combinations of actions on the data set the model is constructed for.

The different evaluation criteria for models are available in the PARCEL software. For visual evaluation of models a graphical representation of the distribution of samples is available: by the graphs of Scores and Loadings, by the graphs of the reference values and calculated values obtained by analyzing RMSEC, RMSECV, RMSEP. In addition, the analysis of the quality of the model can be performed on the stability criterion F, Student's t distribution, the Mahalanobis distance, etc.

For the calculation of calibration models in the PARCEL software the following algorithms can be selected: the Principal Components Regression (PCR), the Partial Least Squares Regression (PLS), the Quadratic Optimization in Hypercube (HQO).

To improve the quality of the model, the different preprocessing options for initial data are suggested that can be used either separately or in various combinations. The opportunity to adjust the input array is provided. For example, it is possible to remove the spectral points in the calibration sample set or remove the separate samples (outliers).

The feature of the PARCEL software is that program is adapted under the routine analysis: the options are implemented to simplify the operator's work.

Not only possibility of the manual selection is realized, but also step-type optimization of certain parameters. At any step the configuration parameters can be varied in the intervals specified by the operator. The sequence of steps is stored, with the ability to return to any previous step. Each step narrows the search range of optimal parameters.

Step-type search algorithm appeared very effective in practice of the Lumex company. Experienced operators can find an optimal model within 3-4 steps.

If the analysis is routine, an experienced operator using step-type optimization can create a template that can be applied by less experienced operator. In the generated template the settings for each step are saved and the criterion of selection of optimal models for each step is stored.

We hope that the software Parcel can be useful not only for the processing of spectral information obtained from InfraLUM devices, but also for the analysis of any multivariate data sets.

P01. Areas of uncertainty in solution of inverse problems of chemical kinetics

F.T. Badtrdinova
Birsk State Social Pedagogical Academy, Birsk, Russia

The inverse problem of chemical kinetics is a problem of identifying reactions and the rate constants, other kinetic parameters associated with these reactions. Solution of this problem is often obstructed with ambiguity upon the estimation of specific kinetic parameters. The ambiguity reflects the nature of a kinetic model that describes only some features of chemical reactions in a certain area of the reaction factors. Therefore, it is necessary to evaluate the uncertainty limits for the kinetic parameter estimates. For this purpose, we suggest using a method that is based on the Kantorovich idea [1].

In this method only knowledge of a maximum experimental error is used. Each measured value is considered to be an interval (segment) [x] that is a set of all possible x values limited by inequalities $x^{-} \leq x \leq x^{+}$ [2]. Under this assumption each kinetic parameter can be estimated by a region, such that every its point is a result of a numerical simulation of the reaction. Considering all these regions together we obtain a multidimensional area that consists of the points that represent a valid set of the kinetic parameters.

1. Kantorovich, LV , *Siberian mathematical journal*, **3**, 701-709. (1962)
2. Khlebnikov A.I., *J. Analyt. Chem*., 31, 321-322, (1996)

P02. The modification of carbon-paste electrodes using motor oils for their identification

Bikmeev D.M., Sidelnikov A.V., Kudasheva F.Kh., Maistrenko V.N.
Bashkir State University, Ufa, Russia

Motor oils are an important part of lubricant materials, which are used in different technologies. Topical problems in the use of motor oils are the control of their quality and also the detection of counterfeit products. Multisensor systems such as “electronic tongues” and “electronic noses” are of great interest in solving these problems. In our work a three-electrode system with carbon-paste electrodes (CPE) has been proposed. As a binder in such CPE, the analyte itself, i.e. a viscous organic liquid (motor oil) has been used. Multivariate signals from the analyte have been formed using the voltammograms of markers, which are able to selectively interact and accumulate on the surface of the working electrode, depending on the nature of the binder. To evaluate the similarities and differences in the obtained analytical signals PCA has been carried out. The results of PCA have shown that the data are grouped into clusters depending on different types of motor oils. The identification of motor oils has been carried out using the PLS-DA, which showed the possibility of their recognition with RMSEC equal to 0.2.

P03. Modeling of potentiometric multisensor array dynamic response in flow-through system with 3wayPLS regression

Yulia Blinova[1], *Dmitry Kirsanov*[1], *Vasily Babain*[2], *Andrey Legin*[1]
[1]*Chemistry Department, St. Petersburg State University, St. Petersburg, Russia*
[2]*Khlopin Radium Institute, St. Petersburg, Russia*

There is a strong demand for quantifying rare earth metals (RE) in different industrial solutions with the help of relatively inexpensive means of chemical analysis that can be fully automated and performed remotely. Ion-selective electrodes could be a nice alternative for “heavy” ICP- based methods, however all RE have got very similar chemical properties, ionic radii, oxidation states, etc. and highly selective sensors for each particular lanthanide can hardly be developed. Lanthanide selective potentiometric sensors widely reported in recent literature [1, 2] are of questionable use in most of the practical tasks since they have pH working range of 4-8, while the presence of Me3+ species in solutions with pH 4 and higher is negligible and most industrial solutions containing RE are strongly acidic (pH<2).

The use of different organic ligands suggested for the liquid extraction of actinides and lanthanides allows for development of polymeric sensors with pronounced sensitivities towards RE in acidic media [3]. These sensors do not exhibit sharp selectivity for any particular RE, but possess cross-sensitivity to all lanthanides. Being placed in the mixture of several RE a set of such sensors based on different ligands yields a kind of “unresolved spectrum” signal, which contains information on most of (ultimately – all) RE in the mixture. This unresolved but reproducible potentiometric response can be processed with classic chemometric techniques such as PLS to extract qualitative and quantitative data about multiple components in the analyzed media. These sensor arrays typically produce rather compact data sets which can be handled as a matrix (samples X sensors) A very promising way for sensor systems data treatment is considering their response also in a time domain, i.e. taking into account the information hidden in the kinetics of sensor system response. This will lead to the appearance of the third dimension in a multivariate data structure (samples X sensors X time), and in spite of being more complex and more computationally demanding, such 3D models could offer extra analytical information exploited from the chemical kinetics. In this study we have constructed special flow-through system with 9 potentiometric sensors based on different ligands. This system was applied for analysis of model mixtures of several RE metals in nitric acid (pH=1.6) and dynamic response of the multisensor system was modeled with 3wayPLS regression to quantify each particular constituent of the mixture. Experimental details and prediction performance of these models will be discussed in the presentation.

1. M. Shamsipur, M. Hosseini, K. Alizadeh, M. F. Mousavi, A. Garau, V. Lippolis, A. Yari, *Anal. Chem.* 2005, **77**, 276.

2. H. A. Zamani, G. Rajabzadeh, M. R. Ganjali, P. Norouzi, *Anal. Chim. Acta* 2007, **598**, 51

3. V. Babain, A. Legin, D. Kirsanov, A. Rudnitskaya, Y. Tatuev, V. Baulin, *Radiochim. Acta* 2009, **97**, 479

P04. Phenomenological modelling of complexonometric titration curves with photometric end–point detection

E. A. Cherednik[1]*, S.V. Romanenko*[1]*, A. S. Salnikov*[2]*, A.G. Zarubin*[3]

[1] *Tomsk Polytechnic University, Institute non–destructive testing, Russia*

[2]*Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia*

[3]*Tomsk State University, Chemistry Department, Tomsk, Russia*

At present great attention in analytical chemistry is paid to charge-coupled devices as colour-change detectors (e.g., WEB-cam). In the proposed WEB–cam application, end–point detection of complexonometric titrations is described. Solutions of Ca^{2+}, Mg^{2+}, Ni^{2+} and Zn^{2+} in concentrations from 0.1 to 10^{-5} were examined. The analytical signals corresponded to s-shaped curves that were fit with the Boltzmann sigmoid function.

Examined elements were chosen due to the different colour transitions, which occurred in the neighbourhood of the end points. Thus the colour transitions of Mg^{2+} and Ni^{2+} complexonomteric titrations (red-blue and green-violet, accordingly) may be defined as high-contrast colour transitions, and the cases of Zn^{2+} and Ca^{2+} (yellow-pink and pink-violet, accordingly) – as low-contrast colour transitions.

High-contrast colour transitions were treated with the RGB system, and low–contrast colour transitions with the HSV system. In the first case the analytical signal was the value of the more variable colour channel, in the second it was the hue value of the examined solution.

In order to obtain the concentration value a phenomenological approach to the analytical signal description was proposed. The titration curve was described with the Boltzman sigmoid function

$$I = A_2 + \frac{A_1 - A_2}{1 + \exp\frac{(V - V_{ep})}{dx}}$$

preliminary modified for this purpose. Thereby all its coefficients were endowed with physical meaning: I is the current value of colour characteristic of examined solution; A_1 and A_2 are the colour characteristic values of examined solution before and after its titration, accordingly; V and V_{ep} are the currently added titrant's volume and its volume at the end point, accordingly; dx is the rate of colour's change.

The presented dependence resulted that coordinates (x_0, A_1+A_2/2) determine the end point position. Experimental data modelling with the proposed approach gave correlation coefficients to the Boltzmann sigmoid function was no less than 0.995 over the entire concentration range. Relative error in the estimation of examined elements' concentration over the whole concentration range did not exceed 3 relative percent.

The proposed method was approved for complexonometric determination of general water hardness for *Chyornoe* and *Barochnoe* lake water in the Tomsk region. The concentration results were consistent with ones obtained by ICP – atomic emission spectroscopy.

P05. Process fault detection and diagnosis with use of discriminant analysis

Galiaskarov M.R., Rusinov L.A.
Saint-Petersburg State Technological Institute, Saint Petersburg, Russia

The majority of technological processes in the chemical industry are potentially dangerous. Therefore, systems of early diagnostics of faults and abnormal situations are important tools to increase their efficiency. At the same time, such processes frequently have no mathematical descriptions connecting observable parameters of process with the causes of faults — the diagnostic models. One solution for this problem is the use of statistical information for forming models of abnormal situations during system training. Thereafter, abnormal situations are detected in the process data with the help of the discriminant analysis.

The basic idea of the discriminant analysis consists in determining, whether the sets differ by the average of any variable or of linear variable combination, and then to use this variable for predicting for new samples their belonging to this or that set. In this case, the technique of the discriminant analysis of Fisher (FDA) is usually used.

FDA is the technique of constructing linear boundaries between classes such as the distance between classes is maximized and the scatter inside each class is minimized.

The basic problems of the discriminant analysis are, first, definition of a set of the discriminant variables and, secondly, choice of a type of discriminant functions. There are various criteria of the consecutive selection of the variables for receiving the best discrimination of classes. More often in practice the linear discriminant function is applied — the linear discriminant analysis.

However, linear discriminant function does not frequently provide a good separation of classes, since technological processes in the chemical industry are, as a rule, essentially nonlinear. In this case, it is expedient to use the discriminant analysis with nonlinear discriminant function, for example — a kernel Fisher's discriminant analysis (KFDA).

KFDA is practically a linear discriminant analysis with the nonlinear transformation of the initial data executed before its carrying out with the help of kern — functions.

Results of experiments showed, that KFDA has confidently detected abnormal situations in technological process at early stages of their development, having shown considerably higher characteristics in comparison with FDA.

P06. Application of scanning probe spectroscopy in monitoring of oil pollution

A.O. Gyruanova[1]*, V.V. Ermakov*[1]*, A.Y. Bogomolov*[2] *and D.E. Bykov*[1]
[1]*Samara State Technical University, Samara, Russia*
[2]*J&M Analytik AG, Essingen, Germany*

Traditional scanning of oil pollution in soils requires lengthy sample preparation and gives no information about how well the sample properties are represented by the measured values. The use of pattern recognition chemometric procedures with spectroscopy offers a simple and effective solution for the identification of sludge by hydrocarbons derived from oil.

A transition from traditional to hyperspectral measurements not only provides a completely new qualitative nature of the data, but also increases the quantity of the information. The challenge is to reduce the dimensionality of the data while retaining important spectral information with the power to classify areas of a sample effectively. Such complex signals require advanced chemometric procedures to extract the relevant information especially because the measurements are performed directly in the process.

Applying an immersion fiber probe from J&M Analytic AG, five types of sludge with different characteristics and of different origins were studied: treated contaminated soil, post-processed bottom barn sludge, barn flotsam, bottom barn sludge, and sewage treatment refinery sludge. Data collection was performed from the near infrared spectral range to UV-region.

Principal component analysis with autoscaling revealed that sample differentiation was possible using the wavelength range from 200 to 600 nm (UV-Vis). Different types of sludge are grouped in separate clusters. In this case there is a clear division into three highly contaminated sludges on one side, and two relatively clean sludges on the other side. That suggests the greatest effect on the position of the first principal component of oil products in sludge. The first principal component describes 98% of the variance of the samples in the multidimensional space. No relationship between the second and higher principal components and any of the measured properties of the samples was found.

The results obtained show clearly differentiated clusters for the different of types of sludge and correct predictions, thus allowing the differentiation contaminating soils in a reliable manner. Additionally, no prior treatment of the samples is required, thus minimizing the important errors associated with this step of the analytical process. As for the future development, the widespread use of hyperspectral surveys in the cosmic space, air and ground surveillance, expands the boundaries of the proposed methodology, and will make the monitoring of oil pollution in the soil even more effective and faster.

P07. The differentiation of oil sludge depositories depending on their resource potential

E. V. Ilyushina[1], A. O. Guryanova[1], V.V. Ermakov[1], A. Bogomolov[2] and D.E. Bykov[1]
[1]Samara State Technical University, Samara, Russia
[2]J&M Analytik AG, Essingen, Germany

The differentiation of oil sludge depositories is one of the principle tasks for the petroleum companies while making an inventory and entering oil-contaminated waste facilities in the company's books. The largest and most essential projects are primarily utilized allowing reducing the payback period of large-scale ecological projects of the companies.

This requires the development of criteria for stratification of oil sludge depository and building differentiated models of the depository condition due to the size of sludge storage pits resource potential. First of all the physical and chemical composition of oil sludge is carried out taking into account its consumer appeal.

For determining the cost of oil sludge depository the oil companies use the traditional formula. There are several authors who offer another formula. The basis of these formulas is the dependence of the oil sludge depositories cost on the waste quantity and the size of the depository, but none of them takes into account the resource potential of the waste, as a mixture of substances. In our work we try to solve this problem using the methods of multivariate data analysis.

P08. Detection of amino acids in solutions by an «electronic eye» system

A.V. Kalach, *O.B. Rudakov*

The State Architectural-Building University, Voronezh, Russia

Volume effects (counteraction, swelling) in 12 polymeric, ionogenic, and non-ionic sorbents in the form of spherical granules in aqueous amino acids (glycine, α–, β–alanines, isoleucine), dipeptide glycil-glycine and the protein lysozyme are studied by digital microscopy. Granules of polymers are used, for the first time, as sensitive elements of sensor controls. A prototype for an optical multisensor system, an «electronic eye», based on digital video recording of volume effects of several granules in an analyzed solution and the supplied system of processing of the multidimensional data is created.

The choice of granules for use as sensitive elements of multitouch system is theoretically proved and experimentally validated in the experiments with the amino acids and water-soluble medical products, which has been identified and measured in the water solutions.

The possibility of quantitative identification of the general fiber in isotonic solutions and a plaintive liquid by digital microphotographic measurement of effects of the counteraction of granules cation-exchange pitch C 120 E is shown.

Techniques are approved by “EcoWater” Enterprise, in the Laboratory of Mass Analyses of the Voronezh State Architectural-Building University, by the Chair of Pharmaceutical Chemistry and Clinical Pharmacy of the Voronezh State Medical Academy.

P09. Different validation modes for PLS models in prediction of rare earth metals in complex mixtures by potentiometric multisensor system

<u>M.M. Khaydukova</u>[1], D.O. Kirsanov[1], V.A. Babain[2], A. Legin[1]
[1]Chemistry Department, St. Petersburg State University, St. Petersburg, Russia
[2]Khlopin Radium Institute, St. Petersburg, Russia

PLS regression is the most widely used method to obtain calibration models for numerical predictions of various parameters in chemometrics. Predictive ability of PLS regression models should be properly evaluated before any kind of real life application of such PLS models can be considered. The results available in the field of electronic tongue (ET) indicate that the researchers usually do not pay serious attention to the realistic estimation of the predictive ability of PLS models, moreover they often do not perform this important step at all. In the vast majority of the papers numerical parameters of the regressions (such as RMSEP, offset, R2, slope) are reported for the validation procedure based on a full cross-validation or a single random split test set validation. However, these parameters often do not suggest a realistic estimate of the further predictive power of the model, since the same objects (samples) are used for the development and validation of the model. These issues are widely described in chemometric literature [1-3]. Cross-validation is widely known to produce over-optimistic results and can serve only as a rough estimation of a model performance. Test set validation is a more preferable option, but it requires a large number of samples for training, optimization and evaluation of predictive ability of a model. Large number of different samples for the ET is rarely available in research, because all of the samples should be evaluated with various reference techniques (instrumental, sensory panel, etc) and this could be fairly expensive if doable at all. This is typical not only for ET applications but for many other areas as well.

In this study we compare several validation approaches (full cross-validation, single random split test set, leave-one-object-completely-out (LOCO)) for PLS regression. Data set was obtained from potentiometric measurements with multisensor system based on 14 different types of polymeric membranes with high cross-sensitivity towards rare earth metals (RE). Measurements were performed in 39 model mixtures, containing yttrium, lanthanum and gadolinium. Concentration of each metal was varied over 5 different levels in the range 10-5–10-3 M. The data from multisensor system were used to build PLS regression models for prediction of each particular RE in a mixture. The results of comparison of three different validation modes in this application will be presented.

1. K. H. Esbensen, P. Geladi, *J. Chemometrics* **24** (2010) 168.
2. R. G. Brereton, *TrAC* **25** (2006) 1103.
3. E. Anderssen, K.Dyrstad, F. Westad, H. Martens, *Chemom. Intel. Lab. Syst.* **84** (2006) 69

P10. A study of spatial structure of ^{137}Cs contamination in natural conjugated landscapes: fractal stability and dynamics

Elena Korobova
Vernadsky Institute of Geochemistry and Analytical Chemistry, Moscow, Russia

^{137}Cs was used as a tracer to test the hypothesis of the regular and fractal character of distribution of chemical elements in the environment [1]. Field gamma-spectrometry was used to measure Cs-137 activity in soil cover and overground vegetation in a regular grid (5, 2, 1, 0.5, 0.2 m) and along cross-sections. Vertical distribution of radiocesium was studied by core sampling to the depth of 40 cm. Moss and litter samples were collected from the area 15x15 cm in 10 and 5 m grid. Laboratory determination of ^{137}Cs in soil cores (in increments of 2, 5 and 10 cm), moss and litter samples was performed with the help of Canberra gamma-spectrometer (HP Ge detector).

The revealed system of ^{137}Cs polycentric anomalies both in soils and moss cover was related to meso- and microrelief features therefore exhibiting different levels of spatial organization. Repeated measurements proved stability of the observed Cs-137 structure on one hand and its slow relief-dependent dynamics on the other. This dependence allowed suggesting Cs-137 water migration. Cs-137 spatial distribution and its dynamics were reflected in the form of frequency histograms which seemed to be helpful in revealing the character of spatial structure of geofields of chemical elements.

The author is grateful to Dr. S.L. Romanov who inspired this study and participated in field works and interpretation of the obtained data, to Drs F.V. Moiseenko, V.L.. Samsonov and

V. Yu. Beryozkin for participation in field studies in different years, to S. Kirov and Dr. N.N. Dogadkin for Cs-137 determination in the collected samples.

1. Romanov S.L., E.M. Korobova. *Mineralogical Magazine*, **75** (3), 2011, 1746.

P11. Spectroscopy for secondary students: exploring the limits of analysis at home

Christopher A. Marks, James M. Quagliano
St. Christopher's School, Richmond, VA USA

A visible spectrometer has been constructed from materials readily available to many secondary students. The heart of the system consists of a digital photo frame (displaying a pure white image) source, plain glass salt shaker cuvettes, and a digital camera detector. Black plastic crates and DVD cases have been modified to maintain the system geometry. Black construction paper, and fabric were used to minimize stray light and reflections.

A sample and blank cuvette are photographed side-by-side in front of the source. The images are analyzed using the GNU Image Manipulation Program (GIMP 2.6.X). Red, green, and blue (RGB) intensities (I) are averaged from appropriate image regions (containing thousands of pixels) for the blank and the sample. The sample absorbances are calculated as $-\log10(I_{\text{sample}}/I_{\text{blank}})$ for each of the red, green, and blue channels.

The system has been used to explore: Beer's law using food colors (including mixtures); the pH properties of household products using a red cabbage indicator; and the classification of vegetable oils. All quantitative measurements were made using common kitchen or medicine cabinet equipment.

The three channel (RGB) limitation of the present set-up gives less than satisfactory results in some of the above cases. However the pedagogical value of this simple system, including the demonstration of its limits, is high. Many important concepts can be introduced. Future work will involve modification of the spectrometer with a continuous source and the addition of an inexpensive diffraction grating film (available in hobby shops and online). The superiority of multivariate methods can then be demonstrated. A thorough investigation of Mountain Dew is planned.

P12. Quantitative Analysis of Milk with Vis/SW-NIR Spectroscopy

Anastasiia Melenteva[1]*, Andrey Bogomolov*[1,2]
[1]*Samara State Technical University, Samara, Russia*
[2]*J&M Analytik AG, Essingen, Germany*

Cow's milk is a mass market product that requires a routine quality control at different stages of its production and sale. Concentrations of the main nutrients, i.e. fat, total protein and lactose are the most important quality assessment parameters of the commercially produced milk and its derivatives. Growing quality requirements in the pharmaceutical and food industries promote the dissemination of spectral analysis enabling instant i.e. in-, on- or at-line monitoring of production processes and products. Optical spectroscopy is one of the most favoured techniques for the industrial milk analysis due to its wide availability, undemanding sampling, quick measurement and the applicability of fibre optics and remote probes.

Spectroscopic analysis of homogenized and natural raw milk is typically performed by infrared (IR) or near infrared (NIR) spectroscopy. The spectral analysis in the visible (Vis) region is economically very attractive, but it is complicated by the high light scatter by suspended fat and protein particles. For a successful quantitative analysis the unknown and variable particle size distribution should be taken into account during the data collection and modelling.

The present work reports on the development of a milk analyser capable of accurate determining the raw milk nutrient contents from spectra in Vis and adjacent short wavelength NIR (SW-NIR) region (400–1000 nm). The method is conceptually based on measuring the scattered light intensities at different wavelengths and at different positions with respect to the light source (space resolved measurement). The resulting spectra enable quantitative analysis of the milk components in a wide concentration range. The modelling was built and validated on a representative set of designed milk samples with controllably varying component concentrations and fat globule sizes.

The methods of data analysis have been discussed.

P13. Multivariate Analysis in Combination with Electrospray Ionization Tandem Mass Spectrometry as a Model for Prediction of 3-Ketosteroids' and Corticosteroids' Structures

S. A. Appolonova, N.V. Mesonzhnik and G.M. Rodchenkov
Antidoping Centre, Moscow, Russia

Mass-spectrometry became a common method for screening and confirmation procedures used by World Anti-Doping (WADA) accredited laboratories. Technological progress made it possible to adapt new analytical instruments and strategies for the purposes of doping control. Nevertheless, MRM mode of HPLC-MS/MS methods does not allow us to make non-target analysis and detect new compounds that are not included into target list. Therefore, it is necessary to develop new approaches allowing not only finding unknown substances in biological matrixes, but also to propose and predict their chemical structures.

The main objective of this study was to investigate the possibility of discrimination and structural elucidation among (cortico)steroids on the basis of their LC-(ESI)-MS/MS mass spectra using multivariate statistical techniques.

The principle of our study was, in a first stage, to collect replicates of the mass spectra of each compound, then, in a second step, to analyze these raw data considering each ion in the spectra as a statistical variable and each mass spectrum as an observation. After selection of discriminant ions using principal component analysis (PCA), a partial least square regression analysis (PLS) allowed us to set up a predictive model for further structural determinations.

Tandem mass spectrometry measurements were used at the collision energy of 20eV to obtain product ions for each analyte. Experimental design and multivariate analysis were performed in SIMCA-P+ (version 12.0.1, Umetrix, Sweden). All mass spectra data were converted into text files and then imported into SIMCA-P+ as an array of m/z and intensity values. For all analytes the spectral data were referenced by finding out the common mass shifts due to the $[M+H]^+$ ion. PCA and PLS analysis were applied to all previously referenced data transferred into SIMCA-P+ matrix for the structural model evaluation. Analysis of the obtained model showed that the compounds were divided into different groups corresponding to the numbers and position of double bonds and the presence of functional groups.

Thus, the obtained models have the ability to predict new compounds that share the common structural elements. The model was verified on the (cortico)steroids analytes that were not previously used in model evaluation.

In this study we have shown the possibility of discrimination and structural elucidation for the structural elucidation of expected (targeted) and unexpected (non-target) metabolites and designer modifications of 3-ketosteroids and corticosteroids.

This study was financially supported by the Partnership for Clean Competition (PCC) http://www.cleancompetition.org/funded-research.aspx

P14. Detecting unexpected drugs of abuse using UPLC-HRMS in combination with multivariate analysis

N.V. Mesonzhnik, S. A. Appolonova and G.M. Rodchenkov
Antidoping Centre, Moscow, Russia

Ultra performance liquid chromatography with high resolution mass spectrometry (UPLC-HRMS) is an analytical technique that, due to its high sensitivity, has become increasingly popular for the detection of controlled and banned substances in urine samples and for the searching of unknown compounds. In spite of the progress that has been made, the ability to do non-target screening on a routine basis is highly time-consuming because the matrix for analysis is complex and it becomes difficult to pick out individual ions and it is not possible to discriminate unexpected compounds present in the sample from endogenous compounds and background ions.

The aim of the present work was to develop a screening strategy for the unexpected drugs (non-target screening) search based on application of UPLC-HRMS data in combination with multivariate analysis (MVA) in the anti-doping field and establishment of the criteria's for the method. The result was evaluated on example of "designer" steroids and/or their metabolites. The main idea was to create robust and interpretable multivariate models based on metabolomic strategy not for search of biomarkers but for the determination of unexpected drugs under specified conditions based on any a priori knowledge.

Tree types of sample mixtures were analyzed using UPLC-HRMS (Exactive, Thermo Finnigan) in combination with MZmine 2.4 and SIMCA-P+ (v 12.0.1, Umetrix, Sweden) software: 1) only endogenous steroids; 2) endogenous and target exogenous steroids and/or their metabolites; 3) endogenous and non-target compounds and/or their metabolites. Generic sample preparation protocols were employed in all cases. Before data analysis, the raw mass spectra were pre-processed in the MZmine 2.4 as follow: first, mass values are detected within each spectrum. In the second step, a chromatogram is constructed for each of the mass values which span over certain time range. Finally, deconvolution algorithms are applied to each chromatogram to recognize the actual chromatographic peaks. These parameters were set as follows for peak detection: m/z bin size at 0.250, noise level (absolute value) at 250, minimum peak eight (absolute value) at 15, minimum peak duration at 3 s, tolerance for m/z variation at 0.2 Da and tolerance for intensity variation at 20%. Then, several modules are used for further processing of peak detection results, including deisotoping, filtering and alignment. In this way, aligned peak tables were created according to specified peak finding and alignment parameters: balance between m/z and RT at 10.0; m/z tolerance size at 0.2; RT tolerance size at 1%. Then SIMCA-P+ software was used for multivariate data analysis. The results showed that the evaluation of the extracted information data using partial least square regression (PLS) provided a robust, predictive and transparent model for the compound differences between the tree types of urine samples. The characteristics of this multivariate model were the following: R2(Y)=0.910 and Q2(Y)=0.812.

The obtained models have allowed to search unexpected compounds under specified conditions and this research has shown proof of principle but further studies are needed to validate the methodology before routine application.

This study was financially supported by the Partnership for Clean Competition (PCC) http://www.cleancompetition.org/funded-research.aspx

P15. Data mining adequate visibility model for “black box”

V.M. Mordashev

Russian Scientific Centre “Kurchatov Institute”, Moscow, Russia

The main task of approximation in its classical formulation is formulated as follows: “At some point sets M in the space of an arbitrary number of measurements are given two functions: $\varphi(P)$ and $F(P, A_1, A_2, \ldots, A_m)$ of the point $P \in M$, the second of which also depends on a number of parameters $A_1, A_2, \ldots, A_m$, these parameters are required to determine, so that deviations in M the function $F(P, A_1, A_2, \ldots, A_m)$ from the function $\varphi(P)$ was the smallest. This, of course, must be specified what is meant by the deviation”

Now, the computational methods and tools allow almost always solving this problem for the approximation almost of any set of initial data (values $\varphi(P)$).

If we study the dependence of one or two variables, the type of model can "spy" on point graphs. For three variables, this approach is useless for any experiments. We live in the three-dimensional world! If we could somehow separate the variables in the data into separate groups so that their point graphs could be examined separately by using visualization!

Nonlinear multifactor analysis (NLMFA) created by the author [1, 2] allows such separating in multifactor tables. As the result is nonlinear polynomial transformation (new scale of measurement) of the original table, this by the best way (in the sense of minimum least squares) is approached by the sum of the tables, which depended on fewer variables (groups of variables).

The NLMFA was successfully applied to nearly two hundred problems from a variety of areas, of which only one task has not been solved adequately

The NLMFA requires data in a table (type matrices), i.e. applied only to data of the active experiment. At the same time, most research active experiment is not possible; available are data only the passive experiments, which usually are the disordered massive, the "cloud data".

The way out is obvious. It should be by interpolation and extrapolation or by using neural networks out from the “cloud data” get the “quasi-truthful” table like the active experiment (the model “black box” as a “black box”); afterward using the NLMFA find the kind of the visual model; and, finally, to make the model by means of the original "cloud data".

1. Mordashev V.M., *Soviet Math. Dokl.*, **12** (1971),782-783.
2. Mordashev V.M, *Physics of Atomic Nuclei*, 2010, iss.13, 2167-2179.

P16. Calculating Principle Components Regression in MapReduce architecture

P. Nuzhdin, *S. Zhilin*
Altai State University, Barnaul, Russia

In an increasing number of scientific disciplines advanced experimental facilities produce huge amount of data which are subject of suitable processing and analytical procedures of different kind. In such application areas as microarrays, genomics, proteomics and metabolomics, medical imaging, remote sensing, homeland defense (e.g. face recognition, epidemiology) sizes of datasets often reach and exceed terabyte scale. Working with large datasets of such size requires a special computational infrastructure and algorithms.

MapReduce [1] is a programming model proposed by Google for distributed processing of large datasets on massively parallel systems. An algorithm of data processing must be expressed in terms of a map and a reduce function. These functions are used for an automatic parallelization of a task by a runtime system across computational nodes. Some of nodes act as masters which partition the problem into smaller sub-problems and distribute them to worker nodes while workers solve their sub-problems using function map and pass the answer back to their masters. The master aggregates partial answers into a final solution of the original problem using reduce function. Runtime system is responsible for machine failures handling and inter-node communication scheduling to make efficient use of the network and disks. So it is not necessary for a user to care about technical details and he could afford to concentrate on algorithmic essence of the problem. Apache Hadoop is an open source implementation of MapReduce extensively used for a processing of datasets of tera- and even petabyte scale.

The poster presentation is devoted to an algorithm for calculation of principal components regression (PCR) in MapReduce model. The algorithm consists of the following main steps: centering of input matrix of regressors, optional regressors scaling, principal components decomposition of the preprocessed input matrix, PCR parameters calculation, regression quality evaluation and calculation of prediction for a given set of samples. All but one of these steps are implemented in terms of map-reduce functions and could be parallelized and scheduled by Hadoop. Principal components decomposition is the only computational step realized in non-parallel manner because sequential implicit QL-algorithm [2] used in this step solves eigenvalue problem for up to 1000 regressors faster than Hadoop starts and warms.

The algorithm was tested on experimental 2-node Hadoop cluster for synthetic datasets of the dimension 1,000,000x500 and demonstrated speedup factor of 1.8.

1. Dean J., Ghemawat S. Simplified data processing on large clusters. Operating Systems Design and Implementation, 2004, pp. 137-149.
2. Demmel J.W. Applied Numerical Linear Algebra, SIAM, 1997.

P17. Calibration set design for MLR methods in spectrophotometric analysis of multicomponent mixtures

I.V. Vlasova, A.S. Shelpakova, A.A. Nagaev
Omsk State University, Omsk, Russia

Use of regression coefficients calculated from multiple linear regression (MLR) of the calibration set spectra instead of the absorption coefficients allows improved accuracy of analysis. The design of a well-conditioned concentration matrix for use as a calibration set for mixtures is the main problem here. An incorrect sampling of calibration set concentrations can impair the results of analysis. In this paper a new algorithm for the calibration set design in MLR methods for spectrophotometric analysis of multicomponent mixtures is proposed. A minimum of random errors functional is suggested for use as an optimality criterion for the quantitative compositions of calibration sets. The theoretical justification is carried out by using the methods of matrix analysis.

A MATLAB package, «Optic-MLR», is developed based on the above mentioned algorithm. The package consists of three modules: 1) GENERATE, 2) KOEFF_COUNTER and 3) CONCENTR_COUNTER. The design matrix concentrations for the calibration set is the purpose of the first module. The input data are the number of mixture components (n), minimum and maximum concentration of each component in the calibration set, and the numbers of calibration mixtures from m = n to a user-specified upper limit. There are two kinds of initial data input: a normal method (no weighted data) and a special method using weighted coefficients that take into account the contribution of the components to the absorbance of the mixture. The smaller the value of weighted coefficients, the bigger is the contribution of component to the absorbance. Numerical values of the weighted coefficients are specified by the user. The second module is used to calculate the regression coefficients from the spectra of calibration set mixtures. The calculation of the component's concentrations in the test sets is the purpose of the third module.

Test mixtures containing additive and nonadditive absorbances and from 2 to 6 components were analyzed using «Optic-MLR». For each type of mixture, we also varied the calibration set size. The proposed algorithm, taking into account the level of random errors, allowed as few as 2*n* samples, where *n* is a number of components, to be used for calibration. The calculated regression coefficients allowed determination of the components with a relative error not exceeding 3-5% even with nonadditive absorbance and in the presence of other components.

P18. Spectrophotometric analysis of mixtures containing microcomponents with PLS

I.V. Vlasova, A.S. Shelpakova, P.A. Katunina
Omsk State University, Omsk, Russia

Spectrophotometric analysis of multicomponent systems is difficult when individual analyte's contributions to the absorbance differ greatly. Medicines, which contain an active component at the microquantity level, often fall into this category. The contribution of microcomponents to the total absorbance of mixture is significant only in peak region of absorbance and is less than 10%. The absorbance of microcomponents in other regions of the spectrum exists on background level.

We have previously shown that the optimal calibration set size for analysis of mixtures using PLS should be not less than $2n+1$, where n is the number of components, when the contributions of the components to the absorbance is roughly comparable. The presence of microcomponents requires additional conditions. More principal components are needed in the model, and the calibration set size must be increased to $3n+1$. These conditions allowed determination of microcomponents in 4-5-component mixtures with a relative error not exceeding 6%. The relative standard deviation did not exceed 0.04 in all cases.

The constructed calibration is applied to the analysis of drugs. In this study we investigated the following medicines: "Teodibaverin", "Tetralgin", "Pentalgin" and "Andipal." Dibazole is the microcomponent in Teodibaverin, Phenobarbital is the microcomponent in Tetralgin and Pentalgin, while Andipal contains both Dibazole and Phenobarbital. The microcomponent concentrations are 10-30 times smaller than those of the other active ingredients in the drugs. The contents of all active components in the drugs correspond to the nominal composition. The errors in the microcomponents' determination ranged from 1 to 7%, and the relative standard deviation was about 0.12-0.16 for all the microcomponents.

P19. Path modeling of complex chemical reactions using graph theory

S.I.Spivak[1,2], *A.S.Ismagilova*[1]
[1]*Bashkir State University, Ufa, Russia*
[2]*Institute of Petrochemistry and Catalysis RAS, Ufa, Russia*

Graph theory is applied for solution of the chemical kinetics problems. There is a one to one correspondence between the Volpert graph [1] and the mechanism, i.e. a stoichiometric matrix of complex chemical reactions [2]. Each subgraph corresponds to the reaction path. The main properties of reaction systems have a graphical interpretation, and thus they can be identified using the graph of the system. Combining the methods for identification of the complex reaction components [2] with algorithms for theoretical and graphical selection of the chemical reaction paths [3], one can divide a complex scheme into a set of lower dimension subsystems. Each a subsystem has an individual chemical interpretation. This approach allows to simplify the study of informativeness of complex kinetic models. An important feature is a clear physicochemical interpretation.

The main result is as follows. A set of the chemical reaction stages can be divided into the subsystems. The number of such subsystems is equal to the number of independent reaction paths. The subsystem parametric functions bases uniquely identify the basis of the initial system.

1. Volpert A.I., Hud'jaev S.I. Analysis in classes of discontinuous functions and equations of mathematical physics, Springer, 1985.
2. Gorskii V.G., Spivak S.I. *J. Struct.Chem*, **29**, 924-930 (1988).
3. Spivak S.I., Ismagilova A.S., Khamitova I.A. *Dokl. Phys.Chem.*, **434**, 169-171 (2010).

P20. Application of Chemometric tools to Identification of Ignitable Liquids in Simulated Arson Debris

Nikolai A. Sinkov[1], P. Mark L. Sandercock[2], and James J. Harynuk[1]
[1]Department of Chemistry, University of Alberta, Edmonton, Canada
[2]Trace Evidence Services, National Centre for Forensic Services – Alberta, Royal Canadian Mounted Police, Edmonton, Canada

Detection and identification of ignitable liquids (ILs) in fire debris is an important part of arson investigation. It is often a complicated process since most ILs used are complex mixtures containing hundreds of individual compounds that are weathered to varying degrees by fire and mixed with a highly variable and potentially complex matrix that may contain a variety of pyrolysis products that may interfere with the analysis [1]. A commonly used method for IL identification is headspace extraction of volatile components from arson debris followed by GC-MS analysis, such as ASTM E-1618 protocol used in the Royal Canadian Mounted Police forensic laboratories. Data interpretation is a costly bottleneck in that analysis, which we seek to address through application of chemometric techniques.

Before a chemometric technique such as PLS-DA is applied to a datasets containing GC-MS chromatograms, several challenges should be addressed, such as data alignment and variable selection. Our research has been focused on development of tools to address those problems. In previous works, we have demonstrated the use of a deuterated n-alkane ladder for retention time alignment prior to construction of PLS-DA model [2]. Variable selection was performed using Discriminating Variable test-based variable ranking [3] followed by use of Cluster Resolution metric [4] to determine the optimal number of variables to use.

In this study, we apply this approach to PLS-DA classification of ignitable liquids containing greater variety matrices using further improved variable selection algorithms.

1. E. S. Bodle, J.K. Hardy, *Anal. Chim. Acta* **589** (2007) 247-254
2. N. A. Sinkov, B. M. Johnston, P. M. L. Sandercock, J. J. Harynuk, *Anal. Chim. Acta* **697** (2011) 8-15
3. T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, *Anal. Chem.* **81** (2009) 2581-2590
4. N. A. Sinkov, J. J. Harynuk, *Talanta 83* (2011) 1079–1087

P21. Handling Rank Deficiency in UV/Vis Spectra from Kinetic Studies by Various Multivariate Curve Resolution Techniques

Skvortsov A.N.
St. Petersburg State Polytechnical University, St. Petersburg, Russia

Monitoring charge transfer (CT) and *d*-transition UV/vis absorption bands by optical methods is one of the oldest instruments for studying ligand exchange in coordination chemistry. It is still widely used as it combines versatility, high sensitivity, low response time, and it is relatively simple. UV/vis bands are strongly overlapped, so appropriate numerical handling is required to quantitatively analyze mixtures with more than two absorbing species. Multivariate curve resolution (MCR) techniques are usually employed for this purpose. These include soft approaches, which base only on general constrains (e.g. concentration non-negativity), as well as hard and mixed hard-soft approaches which rely on a physicochemical model. Most of MCR methods base on linear least-square decomposition of experimental data matrix over known spectra or concentrations. The decomposition is valid if the efficient rank of experimental matrix is no less than the number of analyzed components, otherwise a rank-deficiency problem arises. Generally all spectra of the components are linear independent, and the rank of data matrix yielding from kinetic or titration studies is equal to the number of independent concentrations. Yet, casual linear dependence of spectra is not so improbable, as it may seem from general assumptions. In many coordination chemistry experiments the studied process is a sequential substitution of similar ligands (e.g. three-step hydrolysis of thichloride complex). In each step the changes in CT-bands are very similar, as they mostly correspond to change of electronic structure of leaving and entering ligands, which are the same for the steps. Although the spectra of the substances are not exactly degenerate, they may easily become so when noise intervenes in. Thus the ability of MCR approaches to handle this situation is of significant interest.

In the present work we used the experimental data on two-step acid hydrolysis of antineoplastic complex cis-dichlorodiamminoplatinum (cisplatin), monitored by its UV-absorption in CT bands. The data were shown to have additional rank deficiency due to “casual” degeneracy of spectra of cisplatin and its two hydrolysis products. Hydrolysis kinetics was studied at different temperatures and chloride concentrations, so augmented matrices were available for three-way analysis. The data were analyzed using MCR with alternating least squares (MCR-ALS) with general constrains (soft approach), MCR-ALS constrained with hydrolysis kinetic model (hard-soft or “grey” approach), and full-matrix nonlinear regression with hydrolysis model (hard approach). Convergence, stability and rotation ambiguity (for soft approach) were assessed.

Soft approaches were unable to handle the data correctly, the rank of the problem had to be reduced to achieve stable resolution, which had little in common with the studied process. Mixed hard-soft and hard approaches gave almost similar results, although the stability of the solution was strongly degraded by rank-deficiency. This is an anticipated behavior. As the data matrix is degenerate, the linear least-square steps are unable to reconstruct spectra and concentrations of components unambiguously. So the resolution of significant principal components of experimental data into true spectra is driven by the kinetic model alone. The kinetic model relies on the nonlinearity of the kinetic curves, so the complete resolution is still possible. Yet it is loosely supported by the data from experimental matrix and hence strongly depends on *a priori* kinetic model selection. Quite expectedly the tested algorithms displayed slower convergence when used for the studied experimental data, as compared to test examples with non-degenerate spectra. The methods of overcoming of slow convergence are discussed.

P22. On-Line Database of Dyes, "UDD"

O. Sovyn, Yu. Savchyn, I. Patsay
Ivan Franko National University of L'viv , Lviv, Ukraine

Some of the most important information obtained from spectrophotometric research of dyes are the values of acid constants (in the range 10^{-2}–10^{-12}) and the pure spectra of dyes (molar absorption coefficients).

For monobasic dyes this data can be simply calculated. For polybasic dyes it can be obtained from spectrophotometrical experiments [1] with using SpectroCalc-H_5R [2] or similar software.

In our opinion information about the acidic properties and spectrophotometric profiles of protolyzed products are badly needed for any research where dyes are used. So the creation of an on-line Database of Dyes, "UDD" was proposed. It has options for: adding your own experimental data of spectrophotometrical dyes, finding the data of others, dye identification by wavelength maxima in the molar absorption coefficient spectra, and determination of the concentration of dye in solution.

The interface of the database is similar to CMS Wordpress and other freeware templates. Most searching is done using php-programming. Open Access of the database to anyone who needs research information about dyes, and any registered user's ability to add data confirms the service's adherence to the principles of Web 2.0. Part of the data is now in beta-testing, and it will be fully and freely available on www.chem.org.ua/udd/ beginning in April 2012.

1. O. Yatskiv, I. Patsay, O. Sovyn, J. Solution Chem., **39** (4), 2010, 566-574.
2. SpectroCalc-H5A software
http://www.franko.lviv.ua/faculty/Chem/spectrocalc/index.htm

P23. Chemometric modelling of apple quality with spectroscopy

Sylvia Travers
Department of Food Science, Aarhus University, Aarslev, Denmark

Fruit quality analysis has traditionally depended on the measurement of parameters such as dry matter, sugar, starch, acidity and firmness. Dry matter (DM) is considered a reliable physiological indicator of quality because of its strong relationship with sugar content. However measuring DM using traditional approaches is time consuming and requires destructive sampling.

The present work evaluated the potential of near infrared spectroscopy and chemometrics to predict apple quality. Harvest quality analysis was performed on 150 'Elshof' apples picked from an orchard at the Department of Food Science, University of Aarhus in September 2011. Two dry matter and sugar measurements were taken at the equator on opposite sides of each fruit. Reflectance spectra were obtained using an AgriQuant FT-NIR Analyser from the same locations. The AgriQuant was fitted with an InGaAs detector covering a spectral range of 1000-2500nm.

Calibration models were constructed using quality analysis data and spectra. Spectra below 1000nm and above 2300nm were removed due to excessive noise. The level of prediction of dry matter and sugar varied according to the preprocessing methods used. The results demonstrate a correlation between the two quality parameters and spectra.

P24. ICA algorithms in IR spectroscopic analysis of complex mixtures

Tsikin A.M., Monakhova Y.B., Mushtakova S.P.
Saratov State University, Saratov, Russia

Spectral analysis in the IR region allows simultaneous identification and quantification a number of substances in complex matrices (e.g., pollutants in petroleum products). The most toxic components among these are benzene, toluene, xylenes, and PAHs.

However, despite the relatively high selectivity of the method (for example, compared with the electron spectroscopy), the problem of separating overlapping signals is still acute. Therefore, testing of mathematical methods for spectra decomposition in the IR spectral range has a certain practical interest.

Independent component analysis (ICA, including the algorithms MILCA — Mutual Information Least Dependent Component Analysis and SNICA — Stochastic Non-negative Independent Component Analysis) was used to separate the spectra of mixtures containing up to six components in various states of aggregation. The IR spectra of individual compounds were obtained from the NIST database (USA). The best results for spectra with a high degree of overlap (R>0.8) were showed by the MILCA algorithm. The correlation coefficients between the pure and the estimated spectra (R) were in the range of 0.95–1.00 for solid compounds, 0.80-1.00 for liquids and 0.70–1.00 for gases. The Amari index (amari) was 0.0009–0.0958 for solids, 0.0002–0.1366 for liquids and 0.0002–0.1025 for gaseous substances. A general rule is that qualitative and quantitative analysis errors increase with the number of components in the mixture and their degree spectral overlap, but even for a 6-component system with a high degree of spectral overlap (R>0.9) the relative error of qualitative and quantitative analysis does not exceed 15%.

Furthermore, more than 30 prepared mixtures of hydrocarbons containing up to 4 components (such as benzene, toluene, isooctane, and PAHs) were also analyzed in various states of aggregation by different ICA methods: MILCA, SNICA, SIMPLISMA, JADE, RADICAL, FastICA. The best results were obtained by MILCA (R=0.92–1.00, amari 0.0369–0.3042) and SNICA (R=0.94–1.00, amari 0.0042–0.3180) algorithms.

Many different smoothing algorithms (spline interpolation, moving average, locally weighted scatter plot smoothing using linear least squares fitting, and first- and second-degree polynomials that is resistant and not resistant to outliers, and Savitzky-Golay smoothing) were used to increase signal to noise ratio of the mixture spectra. Both MILCA and SNICA methods showed up to a five fold improvement in the quantitative analysis by using the method of moving average for the benzene-toluene-isooctane system (liquid aggregate state). For the anthracene-pyrene-phenanthrene-fluoranthene system (solid state of aggregation) the relative error decreased up to 20 times when the method of locally weighted scatter plot smoothing using linear least squares fitting and a first-degree polynomial that is resistant to outliers was applied. In case of a small number of samples over the wavelength spectra smoothing should be used. Herewith the method of moving average application is preferred due to high level and stability of its results.

Based on the results we can conclude the new ICA methods (MILCA and SNICA) are applicable for qualitative and quantitative spectroscopic mixture analysis in the IR region.

P25. Process fault detection and diagnosis with use of Kalman filters

Vorobiev N.V., Rusinov L.A.
Saint-Petersburg State Technological Institute, Saint Petersburg, Russia

Most processes in the chemical industry are classified as potentially dangerous (PDP). So there is a need for continuous monitoring and diagnostics their state. However, many processes have recycles, which forms circuits with feedback. In addition, most are automated and have control systems with feedback too. The presence of feedback makes it difficult to diagnose because they create masking effects: a fault cannot often be detected until resource of recycle agitator or controller is sufficient. In this case, use of a Kalman filter is effective; it allows a continuous estimation of parameters from noisy data.

Most PDP are assumed to be stationary, in particular, with absence of drift, and presence of noise with zero mean value. Then the residuals of the output of Kalman filter, tuned to the normal process flow in the absence of faults, will be close to zero. Faults will shift the residuals away from zero. If the noise at the filter output is high enough, it is then advisable to use additional smoothing of the Kalman filter output.

Fault detection is produced by the residuals exceeding some threshold. The threshold value is determined, by the requirements of the fault detector operation: the higher the threshold, the less the probability of false detections, but the fault will be detected later. And on the contrary, the lower the threshold, the fault will be detected earlier, but the probability of false detections will increase. The compromise can be reached analogously to threshold selecting when monitoring PDP by means of the cumulative sums cards.

To diagnose the causes of the faults, the use of a bank of Kalman filters is proposed. Each filter is adjusted for a specific cause, which has resulted in a fault in PDP under control. Then the most probable cause of the fault will be that corresponding to the filter with the minimum residual at the moment of fault detection.

Considering that the majority of PDP are essentially nonlinear, for high-quality work of diagnostic system using Kalman filters, it is necessary to use their modifications – the expanded Kalman filter (EKF), or unscented Kalman filter (UKF).

The comparative analisys of these filters was carried out with the example of diagnostic of operation of control of electro-pneumatic valves with positioners, often used in control loops in modern PDP. Besides, as the second object the Tennessee Eastman Process often used for comparison technological, algorithmic, and other solutions on automation was selected. The study has shown approximate equivalence of diagnostic results at usage of both filters. At the same time, application of filter EKF requires less computing expenditures so it is preferable. The UKF can work with more nonlinear models than EKF. The choice of a certain filter modification depends on the complexity and nonlinearity of the PDP under control.

P26. Use of spectrometry in near-infrared region and chemometrics for organic material quality control

<u>E.V. Vostroknutova</u>, M.A. Yaburov, V.M. Golik, A.V. Saprygin
Joint Stock Company Ural Electrochemical Integrated Plant, Analytical Centre, Novouralsk, Russia

Incoming quality control of organic material used in manufacture of gas centrifuges for uranium isotope separation is carried out by standard laboratory methods, which are generally associated with high labor cost. At present, there is a possibility to develop express methods based on spectrometry in near-infrared region (NIR), which enables to determine certain parameters of material quality on the basis of spectral-response characteristics by means of parameter value and reference material spectrum comparison. In this case, experimental data are processed using chemometrics.

Near-infrared spectra of test objects, such as epoxy resins (ED-16, ED-20, DEG-1) and amine hardeners (triethanolamine, polyethylenepolyamine) were obtained by using Thermo Fischer Scientific Antaris II near-infrared spectrometer.

The quality parameters of selected material were determined: mass fraction of epoxy groups for epoxy resins, mass fraction of water for hardeners. Calibration models were developed by mathematical treatment of spectra produced by a training set of reference materials using projection of latent structure (PLS). Estimate of quality parameters for the developed model (root-mean-square errors of calibration (RMSEC) and prediction (RMSEP), correlation factor and basic component quantity proved their acceptability for practical use.

Thus, calibration models were developed for determination of epoxy groups' mass fraction in ED-16, ED-20, DEG-1 epoxy resins, and water mass fraction in triethanolamine and polyethylenepolyamine. It is the authors' opinion that the results obtained could be used to develop the express methods of incoming quality control for epoxy resins and amine hardeners.

P27. Implementation of the web-based flow-oriented approach to the process control and optimization

Yury V. Zontov[1]*, Oxana Ye. Rodionova*[2]*, Alexey L. Pomerantsev*[2,3]
[1]*Moscow State Institute of Electronics and Mathematics, Moscow, Russia*
[2]*Semenov Institute of Chemical Physics RAS, Moscow, Russia*
[3]*State South Research & Testing Site RAS, Sochi, Russia*

We present CSDesign, a modular web-based system [1] providing a flow-oriented style environment for the synthesis and analysis of various processes.

The main features of the system are as follows:

- *Extensibility.* This means that the functionality and the range of tasks solved by the system may be expanded by addition of a new task-specific software module.
- The programming framework developed as a CSDesign part provides a number of interfaces, which a new software module should implement.
- *Intuitiveness.* This means a specific problem development in the form of a flow-chart drawing including such regular actions as drag-n-drop of the solutions' components, and their interconnection with links.
- *Server-side calculations.* Most of all complex and resource-consuming calculations are processed remotely and the results are transferred back to a client asynchronously using the Ajax technology. The rich user interface is used primarily for visualization and input.

The functionality of the system is illustrated with a real-world example of multivariate statistical process control (MSPC) and optimization (MSPO) described in [2]. A process under consideration is a 7-stage food process that is represented by 25 process X-variables and one output y-variable, which is the end-product quality.

Two mathematical methods are implemented for the process control and optimization. They are the PLS regression and the SIC-method.

To accomplish the process modeling several software components such as matrix input module, autoscaling module, and PLS module are implemented. The SIC method for Matlab [3] is adopted and incorporated into the SIC and the Object Status Plot drawing modules. Each module consists of a GUI-part written in HTML, a DLL-library, and several Matlab m-files.

1. Y.V. Zontov, *Quality Journal*, 2010, **59**, 2-6 (in Russian)

2. A.L.Pomerantsev, O.Ye. Rodionova, A. Höskuldsson, *Chemom.Intel.Lab. Systems* **81** (2006), 165–179

3. Y.V. Zontov, O.Ye. Rodionova, SSC-10 proceedings, Lappeenranta, Finland, 2007, PO 5, p. 67

P28. A new use of target factor analysis (TFA)

K. Higgins, J.H. Kalivas

Department of Chemistry, Idaho State University, Pocatello, Idaho, USA

Classifying samples into known categories is a common problem in analytical chemistry and other fields. Numerous mathematical methods based upon measurements of representative samples are available to accomplish classification goals. With spectroscopic data, samples are measured and the corresponding spectra are compared to existing spectral data sets of known classification (library sets) to determine the appropriate classification. Presented in this poster is a study using target factor analysis (TFA) as a classifier. To the authors' knowledge, TFA has not been previously utilized for such a purpose. Specifically, the library set with the smallest residual relative to a new sample spectrum projection is the classification. From the geometry of the TFA projection, an angle merit can also be used for classification purposes. The TFA merit is compared to the k nearest neighbors (KNN) and Mahalanobis distance based on the covariance structure of each library set. Results from four unique data sets are presented. Three are spectroscopic including identification of plastic type (Raman), gasoil plant source (ultraviolet), and extra-virgin olive oil adulterant (fluorescence). The fourth data set is an archeological benchmark data set. The TFA approach performs best for most data sets. Results from this study indicate that TFA is also useful as a classification tool.

P29. Rapid identification of imitation cheese and imitation ice cream based on vegetable fat using NMR spectroscopy and chemometrics

Yulia B. Monakhova[1,2,3], *Rolf Godelmann*[1], *Thomas Kuballa*[1], *Claudia Andlauer*[1], *Dirk W. Lachenmeier*[1]

[1] *Chemisches und Veterinäruntersuchungsamt (CVUA) Karlsruhe, Germany*

[2] *Bruker Biospin GmbH, Rheinstetten, Germany*

[3] *Department of Chemistry, Saratov State University, Saratov, Russia*

Vegetable oils and fats may be used as cheap substitute for milk fat to manufacture imitation cheese or imitation ice cream. The consumer may be deceived if such imitate products are marketed without adequate labeling. In this study, 400 MHz ^1H and ^{13}C nuclear magnetic resonance (NMR) spectroscopy is used in the context of food surveillance to validate the labeling of milk products. Using principal component analysis (PCA), imitate products can be easily detected. In both cheese and ice cream, a classification according to the type of raw material (milk fat, plant fat) was possible. The loadings plot shows that imitation products are distinguishable by differences in their fatty acid ratios.

Furthermore, a classification according to the types of cheese (Edamer, Gouda, Emmentaler, Feta) was possible. Quantitative data regarding composition of the investigated products can also be predicted from the same spectra using partial least squares (PLS) regression. The models obtained for 13 parameters in cheese (R^2 0.75–0.95) and 17 parameters in ice cream (R^2 0.83–0.99) (e.g., fatty acids and esters) are suitable for a screening analysis. NMR spectroscopy was judged as suitable for the rapid routine analysis of dairy products based on milk or on plant fat substitutes.

P30. Mathematical programming method as a means of NIR analyzers calibration

Skutin I.V., Zharinov K.A., Sushkova M.V.
Lumex Ltd, Saint Petersburg, Russia

NIR-analysis is a modern method of determining the composition of organic samples (food, feed, pharmaceuticals, etc.). Nowadays it represents the only alternative to the classical method. The idea of NIR-analysis method consists in determining the composition of the sample by its spectrum without separating the components. The spectra for each sample are unique as the different components of organic samples (protein, fat, moisture, fiber, etc.) selectively absorb light at different wavelengths, and the intensity of the absorption is directly related to the amount of the absorbing component in the sample. Thus, the intensity of the absorption band may be related by the calibration with the amount of substance of interest in the sample. However the quantitative relationships used in NIR-spectroscopy are often too complex. This occurs because the spectra in the NIR-area contain many broad overlapping absorption bands superimposed on the band of the test component. It prevents to construct a simple calibration between the intensities of the characteristic band with the content of the substance of interest. In this case, a multivariate calibration model can be created linking the entire array of spectral information containing the component of interest.

Nowadays there exist many methods for multivariate calibration models constructing, but most applied of them are: the method of Principal Components Regression (PCR) and the method of Partial Least Squares Regression (PLS). Using of these methods permit to construct the calibrations allowing to determine the composition of organic samples with a high accuracy. Using the analyzer InfraLUM FT-40, developed by the Lumex company, with the calibration calculated using the PLS method, it is possible to determine, for example, the amount of protein in wheat, up to 0.3% or the amount of moisture in samples of corn, up to 0.7 %. In conditions of production, when the huge quantities of raw material are being analysed, every tenth of a percent have a significantly affects on the quality of the final product. Therefore, the improving of quality of the calibration model and, as a consequence, of the analyzer in general is an actual problem.

To improve the quality of calibration model in Lumex company it was proposed to use the mathematical programming at the model calculating. This method is known from optimization theory and applied to finding the extremum of the objective function at a given area of a vector space. The problem of computing a calibration model can be represented as a problem of minimizing the square of the difference of spectral data and the required properties of a sample, linked by calibration coefficients. Since the objective function is quadratic, then the solution of the problem of minimizing the objective function the Lemke method was used known from the theory of mathematical programming as a solution to the problem of complementarity. To test the resulting calibration model the InfraLUM analyzer FT-40 was used, with the calibration calculated applying the method of mathematical programming, and samples of wheat and corn that have been analyzed previously. As a result, it became possible to determine the amount of protein and moisture in the samples of wheat and corn up to 0.2% and 0.4% respectively, that is significantly higher than the results obtained using the calibration model calculated applying the PLS method.

1. T. Naes, T. Isaksson, T. Fearn, T. Davies. Multivariate Calibration and Classification: NIR Publications, 2002. 335p.

2. Handbook of Near-Infrared Analysis / edited by Donald A. Burns and Emil W. Ciurczak: Practical spectroscopy v.13. p. 247-280

Participants

Adrienn Lengyel
Analab Ltd.
Laboratory manager
Debrecen, Hungary
lengyeladrienn@analab.hu

Badrtdinova Fairuza
Birsk State Social Pedagogical Academy
Lecturer
Birsk, Russia
fairusa85@mail.ru

Bikmeev Denis
Bashkir State University
PhD student
Ufa, Russia
limm@list.ru

Blinova Yulia
Saint Petersburg State University
Student
Saint Petersburg, Russia
juljascka@mail.ru

Bogomolov Andrey
Essingen, Germany
a.bogomolov@mail.ru

Buydens Lutgarde
Radboud University Nijmegen
Professor
Nijmegen, Netherlands
L.Buydens@science.ru.nl

Cherednik Ekaterina
Tomsk Polytechnical University
PhD student
Tomsk, Russia
kinder_che@mail.ru

Chernetsova Elena
Giessen, Germany
chern_es@mail.ru

Edelman Gerda
Academisch Medisch Centrum
PhD student
Amsterdam, Netherlands
gerdaedelman@gmail.com

Esbensen Kim
Geological Survey of Denmark and Greenland (GEUS)
Research professor
Copenhagen, Denmark
ke@geus.dk

Galiaskarov Marat
S Saint-Petersburg State Technological Institute
PhD student
Saint Petersburg, Russia
gm.iamdeamon@gmail.com

Gelfand Mikhail
Institute of Information Transmission Problems RAS
Vice-director for science
Moscow, Russia
gelfand@iitp.ru

Guryanova Alena
Samara State Technical University
PhD student
Samara, Russia
belki.da@gmail.com

Ilyushina Ekaterina
Samara State Technical University

Samara, Russia
iljushinakatja@rambler.ru

Janos Elek
Analab Ltd.
Scientific advisor
Debrecen, Hungary
elekjanos@analab.hu

Janos Kozak

Debrecen, Hungary
kozakjanos@nnk.hu

Kalach Andrew
Voronezh State University of Architecture and Civil Engineering
Docent
Voronezh, Russia
a_kalach@mail.ru

Kalambet Yuri
Ampersand Ltd.

General director
Moscow, Russia
kalambet@ampersand.ru

Kalivas John
Idaho State Univeristy
Professor
Pocatello, USA
kalijohn@isu.edu

Khaydukova Maria
Saint Petersburg State University
Student
Saint Petersburg, Russia
lmspb@yandex.ru

Khrameeva Ekaterina
Institute of Information Transmission Problems RAS
Researcher
Moscow, Russia
ekhrameeva@gmail.com

Kimstach Tatiana
INTERTECH Corporation
Sales deputy director
Moscow, Russia
tbk@intertech-corp.ru

Kireeva Natalia
Institute of Physical Chemistry and Electrochemistry
Research associate
Moscow, Russia
nkireeva@gmail.com

Kirsanov Dmitri
Saint Petersburg State University
Senior researcher
Saint Petersburg, Russia
d.kirsanov@gmail.com

Kokoulina Elena
Infraspek Ltd.
Chemist
Saint Petersburg, Russia
elena-kokoulina@yandex.ru

Korobova Elena
Vernadsky Institute of Geochemistry and Analytical Chemistry RAS
Science secretary
Moscow, Russia
Korobova@geokhi.ru

Kryvenko Sergey
BioPromin Ltd.
Engineer
Kharkov, Ukraine
krivenkos@inbox.ru

Kucheryavskaya Elena
Esbjerg, Denmark
menhelen@mail.ru

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

Kurkina Victoria
St-Petersburg State Technological Institute
Associate professor
Saint Petersburg, Russia
victoria.kurkina@gmail.com

Marini Federico
Sapienza University of Rome
Researcher
Rome, Italy
fmmonet@hotmail.com

Marks Christofer
Chemistry teacher
Richmond, USA
camarks1@gmail.com

Mazin Pavel
Institute of Information Transmission Problems RAS
Researcher
Moscow, Russia
iaa.aka@gmail.com

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

Mesonzhnik Natalia
Moscow Antidoping Center
Senior research assistant
Moscow, Russia
n_mesonzhnik@mail.ru

Monakhova Yulia
Karlsruhe, Germany
yul-monakhova@mail.ru

Mordashev Vladimir
Russian Scientific Centre “Kurchatov Institute”
Senior research assistant
Moscow, Russia
mrdshv@rambler.ru

Nuzhdin Pavel
Altai State University
Student
Barnaul, Russia
pnzhdin@gmail.com

Pomerantsev Alexey
Institute of Chemical Physics RAS
Leading researcher
Moscow, Russia
forecast@chph.ras.ru

Rodionova Oxana
Institute of Chemical Physics RAS
Leading researcher
Moscow, Russia
rcs@chph.ras.ru

Ruckebusch Cyril
Lille1 University

Associate professor
Lille, France
cyril.ruckebusch@univ-lille1.fr

Rusinov Leon
Saint-Petersburg State Technological Institute
Professor
Saint Petersburg, Russia
lrusinov@yandex.ru

Samokhin Andrey
Moscow State University
PhD student
Moscow, Russia
andrey.s.samokhin@gmail.com

Shelpakova Anna
Omsk State University
Lecturer
Omsk, Russia
shelik85@mail.ru

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

Sinkov Nikolai
University of Alberta
Graduate Student
Alberta, Canada
sinkov@ualberta.ca

Skvortsov Alexej
Saint Petersburg State Polytechnical University
Associate professor
Saint Petersburg, Russia
colbug@mail.ru

Smilde Age
University of Amsterdam

Professor
Amsterdam, Netherlands
a.k.smilde@uva.nl

Sovyn Olena
Ivan Franko National University of Lviv
PhD student
Lviv, Ukraine
olenasovyn@gmail.com

Spivak Semen
Bashkir State University
Professor

Ufa, Russia
s.spivak@bashnet.ru

Stefanov Zdravko
Dow Chemical
Analytical manager
Freeport, USA
zstefanov@dow.com

Sushkova Maria
Lumex, Ltd.
Research engineer
Saint Petersburg, Russia
maria.v.sushkova@gmail.com

Taavitsainen Veli-Matti
Helsinki Metropolia University of Applied Sciences
Principal lecturer
Helsinki, Finland
Veli-Matti.Taavitsainen@metropolia.fi

Tikhomirov Sergey
INTERTECH Corporation
Technical specialist
Moscow, Russia
svt@intertech-corp.ru

Travers Sylvia
Aarhus University
PhD student
Aarslev, Denmark
sylvia.travers@agrsci.dk

Tsikin Alexey
Saratov State University
PhD student
Saratov, Russia
tsikinalexey@mail.ru

Vorobiev Nicolay
Saint-Petersburg State Technological Institute
PhD student
Saint Petersburg, Russia
enicolas@yandex.ru

Vostroknutova Elena
Joint-Stock Company Ural Electrochemical Integrated Plant
Chemical engineer
Novouralsk, Russia
seminar@ueip.ru

Walczak Beata
University of Silesia
Professor
Katowice, Poland
beata@us.edu.pl

Zhilin Sergey
Altai State University
Docent
Barnaul, Russia
szhilin@gmail.com

Zontov Yuri
Moscow State Institute of Electronics
and Mathematics
Mathematician
Moscow, Russia
Yury.zontov@gmail.com

Notes